Deep Learning Neural Networks to Detect Anomalies in Video Sequences

UNIVERSIDAD DE MÁLAGA

PhD. THESIS

Jorge García González

Department of Computer Languages and Computer Science
Escuela Técnica Superior de Ingeniería Informática
Universidad de Málaga

November 2023

Documento maquetado con TEXIS v.1.0+.

Deep Learning Neural Networks to Detect Anomalies in Video Sequences

Memorandum for obtaining the PhD. degree by the University of Málaga presented by

Jorge García González

Directed by

Rafael Marcos Luque Baena PhD. and Juan Miguel Ortiz de Lazcano Lobato PhD.

Department of Computer Languages and Computer Science
Escuela Técnica Superior de Ingeniería Informática
Universidad de Málaga

November 2023

AUTOR: Jorge García González

https://orcid.org/0000-0001-8610-3462

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA PARA OBTENER EL TÍTULO DE DOCTOR

D./Dña JORGE GARCÍA GONZÁLEZ
Estudiante del programa de doctorado TECNOLOGÍAS INFORMÁTICAS de la Universidad de Málaga, autor/a de la tesis, presentada para la obtención del título de doctor por la Universidad de Málaga, titulada: DEEP LEARNING NEURAL NETWORKS TO DETECT ANOMALIES IN VIDEO SEQUENCES.

Realizada bajo la tutorización de RAFAEL MARCOS LUQUE BAENA y dirección de RAFAEL MARCOS LUQUE BAENA Y JUAN MIGUEL ORTIZ DE LAZCANO LOBATO (si tuviera varios directores deberá hacer constar el nombre de todos)

DECLARO QUE:

La tesis presentada es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, conforme al ordenamiento jurídico vigente (Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), modificado por la Ley 2/2019, de 1 de marzo.

Igualmente asumo, ante a la Universidad de Málaga y ante cualquier otra instancia, la responsabilidad que pudiera derivarse en caso de plagio de contenidos en la tesis presentada, conforme al ordenamiento jurídico vigente.

En Málaga, a 21 de NOVIEMBRE de 2023

Fdo.: JORGE GARCÍA GONZÁLEZ Doctorando/a	Fdo.: RAFAEL MARCOS LUQUE BAENA Tutor/a

Edificio Pabellón de Gobierno. Campus El Ejido. 29071
Tel.: 952 13 10 28 / 952 13 14 61 / 952 13 71 10
E-mail: doctorado@uma.es

Fdo.: RAFAEL MARCOS LUQUE BAENA Y JUAN MIGUEL ORTIZ DE LAZCANO LOBATO
Director/es de tesis

Edificio Pabellón de Gobierno. Campus El Ejido. 29071
Tel.: 952 13 10 28 / 952 13 14 61 / 952 13 71 10
E-mail: doctorado@uma.es

UNIVERSIDAD DE MÁLAGA

ANDALUCÍA TECH
Campus de Excelencia Internacional

AUTORIZACIÓN PARA LA LECTURA E INFORME SOBRE LA TESIS DE D. JORGE GARCÍA GONZÁLEZ

Rafael Marcos Luque Baena, Profesor Titular del departamento de Lenguajes y Ciencias de la Computación de la Universidad de Málaga, en calidad de tutor y director de la tesis doctoral de D. Jorge García González titulada **Deep Learning Neural Networks to Detect Anomalies in Video Sequences**; y Juan Miguel Ortiz de Lazcano Lobato, Profesor Titular del departamento de Lenguajes y Ciencias de la Computación de la Universidad de Málaga, en calidad de director de dicha tesis, AUTORIZAN su lectura.

Asimismo, Rafael Marcos Luque Baena y Juan Miguel Ortiz de Lazcano Lobato, en calidad de tutor y directores de la mencionada tesis, INFORMAN que las publicaciones que avalan la tesis no han sido utilizadas en tesis anteriores. También señalan la idoneidad de presentar la tesis por compendio de artículos dada la gran cantidad de producción científica de alta calidad en problemas iguales o relacionados.

Málaga a 21 de Noviembre de 2023.

Fdo: Rafael Marcos Luque Baena

Fdo: Juan Miguel Ortiz de Lazcano Lobato

A mi ansiedad, compañera inseparable de viaje.

Any sufficiently advanced technology is indistinguishable from magic.

Third Clarke's Law, Arthur C. Clarke.

Agradecimientos

El Kárate empieza y termina con Cortesía.

Gichin Funakoshi

Como corresponde, empiezo por la familia. Mi primer agradecimiento es a mis padres, Paco y Asunción (a los que quizás me refiero por sus nombres por primera vez en mi vida). Sin ellos yo no estaría aquí, tanto en sentido figurado como literal. En mi etapa como estudiante les di sobradas razones para dudar de mí, pero tuvieron paciencia y gracias a eso yo estoy presentando una tesis doctoral. También incluyo en los agradecimientos a mi hermano Fran, que me ha aguantado como hermano mayor durante toda su vida y aún se ríe de los comentarios absurdos que solo comparto con él. Voy a acordarme de Zarpa, Atila y Mancha. Mis perros no han hecho nada en particular, pero son adorables y eso me parece suficiente para que aparezcan en este documento (figuras 2.11 y 2.19).

Continúo con mis compañeros del ICAI. Estoy agradecido a Rafa y Juanmi por dirigir esta tesis. Sin ellos todavía estaría dándole vueltas a publicar el primer artículo. A Ezequiel, que es una fuente de conocimiento e inspiración además de una gran persona. A mis compañeros de grupo Enrique, Esteban, Miguel ángel, Karl, Jesús, Rosa, Iván y José David. Parece mentira, pero tras tanto chiste voy llevando mejor lo de volar. Por último, a Paco Vico, que se acordó de mí cuando surgió una oportunidad de trabajar en el grupo de investigación y con ese simple gesto definió lo que serían los siguientes años de mi vida.

Agradezco también a mis amigos, que aunque no me han ayudado directamente, me ayudan indirectamente a afrontar todo lo que hago. Emilio: eres como un hermano para mí; Dani: me encanta que mi estancia por esta tesis (y la adicción a las Magic) nos haya vuelto a unir más; Alfonso: espero que siempre haya una Caricia de Satele; José Carlos: deja de hacernos quedar de feos a los demás porque las comparaciones empiezan a ser odiosas; Paco y Carlos: Pyrestone murió, pero no llegamos al extremo de matarnos entre nosotros así que no salió tan mal; David: eres un buen ejemplo en tantas cosas que no voy a decir ninguna por no olvidarme otras; Victor: tendremos que jugarnos quién es el Doctor García del grupo; Alejandro: maduramos en el momento en el que empezamos a admirar a Krilin, pero en parte siempre seremos críos; Pablo: es totalmente normal que me ames, pero dejemos de lado lo personal y montemos un garito de juegos de mesa; Adolfo: el garito de juegos de mesa necesitará a alguien que le dé criterio y estilo. Aún más que a ellos quiero agradecer a Lina, Alejandra, Irene y Marisa. Que me soportasteis desde el principio con buen humor y una sonrisa. Vuestros novios tenían un vínculo de

amistad que les impelía a aguantarme, pero vosotras no y aun así no lo hicisteis. Ellos ganaron unas parejas estupendas y yo unas amigas con las que criticarlos. Termino este párrafo añadiendo dos nombres de última hora a la lista: Maria y Beatriz. Me alegro mucho de poder añadirlas.

Quiero mencionar a mi *sensei* Paco Villanueva, mis *senpais* Juan y Oscar y mis compañeros de entrenamientos Mario y Emilio (otra vez). El kárate que he aprendido con vosotros me ha permitido mantener la cabeza en su sitio durante estos años.

Para terminar le agradezco especialmente a Mónica haberme acompañado cuando la ansiedad y la frustración me superaban.

Contents

List of Figures

Acronyms

AE	Autoencoder
AI	Artificial Intelligence
AIRe	Artificial Intelligence Review
ANN	Artificial Neural Networks
ASoC	Applied Soft Computing
BS	Background Subtraction
BBOX	Bounding Box
CNN	Convolutional Neural Network
CUB	Bicubic Interpolation
CV	Computer Vision
DA	Denoising Autoencoder
DL	Deep Learning
ECAI	European Conference on Artificial Intelligence
FNN	Feedforward Neural Network
FS	Foreground Segmentation
GAN	Generative Adversarial Network
GPU	Graphic Processing Unit
ICAE	Integrated Computer-Aided Engineering
ICIP	IEEE International Conference on Image Processing
IC	Image Classification
IS	Instance Segmentation
IWINAC	International Work-Conference on the Interplay Between Natural and Artificial Computation
JCR	Journal Citation Report
LiDAR	Light Detection and Ranging
LIN	Linear Interpolation

ML ..... Machine Learning
MLP ..... Multilayer Perceptron
MP ..... Max Pooling
MSE ..... Mean Squared Error
ND ..... Novelty Detection
NN ..... Nearest Neighbor
OD ..... Object Detection
OR ..... Object Recognition
OOD ..... Out-of-Distribution Detection
OSR ..... Open-Set Recognition
PMDAPF ..... Probabilistic Mixture of Deeply Autoencoded Patch Features
RANSAC ..... Random Sample Consensus
RGB ..... Red-Green-Blue
SAA ..... Small-Angle Approximation
SDA ..... Stacked Denoising Autoencoder
SOCO ..... International Conference on Soft Computing Models in Industrial and Environmental Applications
SR ..... Super-Resolution
SS ..... Semantic Segmentation
SSD ..... Single-Shot Detector
SVM ..... Support Vector Machine
VAE ..... Variational Autoencoder
AVG ..... Window Averaging
YOLO ..... You Only Look Once

Abstract

It's a dangerous business, Frodo, going out your door. You step onto the road, and if you don't keep your feet, there's no knowing where you might be swept off to.

The Lord of the Rings, J. R. R. Tolkien.

The advances in Artificial Intelligence and Computer Vision over the last decade have conveniently overlapped with the explosive increase in multimedia content generation by individuals and businesses. This massive amount of data represents a great opportunity, but also a great challenge for its analysis. The option of doing it by traditional human methods is unfeasible, so computers are expected to perform all or some part of the task of processing an ever-increasing amount of video. In this context, Computer Vision based on Artificial Neural Networks and particularly Deep Learning based tools are expected to be the answer to this need, and for this scientists and engineers have to overcome their current limitations.

The main problem addressed in this PhD thesis is the detection of foreground anomalies in generic video sequences by using Computer Vision techniques especially focused on being robust to noise in the images. As a derivative problem, the analysis of traffic sequences using the same tools is addressed.

As a witness to these objectives, a series of research works carried out over four years are included in this document. The first of these was published in the journal Artificial Intelligence Review in 2020 on the acceleration of Foreground Segmentation methods by using image pre-processing methods to alter the size of images while minimising the loss of quality in the final segmentation. The second paper was published in the journal Integrated Computer-Aided Engineering in 2020 and deals with the application of Stacked Denoising Autoencoder networks to overlapping patches of the image to obtain a latent representation that is then analysed using a probabilistic model. The third work was presented at the 2020 European Conference on Artificial Intelligence and proposes a method based on pixel-level semantically segmenting the objects in the images to analyse their movement and thus identify the foreground. The fourth work was presented IEEE International Conference on Image Processing in 2020 and studies how the type and training of Stacked Denoising Autoencoder networks affect patch-based Foreground Segmentation models. The fifth paper, presented at the International Conference on Soft Computing Models in Industrial and Environmental Applications 2021 conference, studies the use of classical filters and Stacked Denoising Autoencoder networks to clean up images as a step before using classical Foreground Segmentation methods

and thus increase the system's robustness to noise. The sixth paper proposes a vehicle speed analysis method to estimate the pollution generated by vehicles based only on images obtained from traffic cameras and was published in the journal Applied Soft Computing in 2021. The seventh work was also presented at the International Conference on Soft Computing Models in Industrial and Environmental Applications congress in 2021 and proposes an estimation of relative speeds for other vehicles from the images recorded by an onboard camera. Finally, the eighth work comprising this thesis was presented at the 2022 IEEE International Conference on Image Processing congress and refines the use of Stacked Denoising Autoencoder networks for Foreground Segmentation from previous work by removing the need to divide the image into patches.

These eight works constitute the memorandum of this PhD thesis and the author presents the results of these four years of research.

Chapter 1

Introduction

Fear tends to come from ignorance.
Once I knew what the problem was, it
was just a problem, nothing to fear.
The Name of the Wind, Patrick Rothfuss.

Abstract: This first chapter serves as a general introduction to the context, motivation, problem to work on and methodology of this thesis. A structure of the document is also provided to facilitate its follow-up.

1.1 Context

If the last two centuries have seen a continuous series of technological advances that have changed the way human beings live, the last few decades have seen this process of change accelerate. The way of life of a young European citizen at the beginning of the 21st century now bears little resemblance to how his or her parents or grandparents lived the same stage of life. Some of these changes, such as the advent of the internet, the ease of access to technology and its miniaturisation, have had a radical transformative effect on one particular point: whereas before one could at most aspire to be a consumer of multimedia content distributed via television, now anyone with a phone or computer can be a generator of such content.

The existence of video publishing and viewing platforms such as Youtube or Twitch has blended with social networks. Not only is multimedia content published on Facebook, Twitter or Instagram, but there are also specific networks such as TikTok whose form of communication is multimedia content. In Spain there are people who remember the beginning of television broadcasts in the middle of the last century and now live in a society where every citizen can offer their own variety of content in half a dozen different platforms with a device that fits in the palm of a hand. We now live in a multimedia world and, barring an apocalypse with its associated fall of civilisation as we know it, that is not going to change.

(a) Example of image from Twitch channel Chiclana & Friends[a].

[a] https://www.twitch.tv/chiclanafriends

(b) Example of traffic camera footage from scene Street Light of ChangeDetection.net dataset Goyette et al. (2012).

Figure 1.1: Examples of very different footage can be found on the Internet.

In addition to data generated by private citizens, public administrations and companies also have multimedia data recording and storage devices for their own purposes. It is common in our daily life to find traffic cameras owned by the city council or video cameras of some security companies monitoring certain areas. Although often overlooked, these devices are generating their own content, not for distribution for entertainment or informational reasons, but for more specific purposes.

The explosion in multimedia content generation offers not only great opportunities, but a multitude of problems, challenges and questions. How do we deal with these massive amounts of data? If a single camera is recording for 24 hours, do we need 3 people doing 8-hour shifts to watch its content permanently? As the saying goes: what goes on the Internet, never leaves the Internet. How can a platform control that its users do not violate its content rules and use it maliciously? It is their platform and they are partly responsible for its use. Can they rely on reports from other users or their own moderators? That is a reactive strategy, but not a preventive one. It does not serve to prevent the publication of content, only to chase down what is already published.

There is too much data being generated all the time to make a purely human system control it: we need automated tools to analyse and extract information from that multimedia content. We need computers to be able to perform or at least simplify the analysis of all those hours of video or the amount of data will overwhelm us. In some cases that data is a mere tool that, if not analyzed, will be wasted, but in other cases, content analysis is a requirement if multimedia networks are not to be chaotic and harmful.

In parallel to this, developments in AI (Artificial Intelligence) have made it possible to generate new, increasingly precise and flexible analysis systems. Specifically, over the last few decades, the CV has developed methods to analyse images and videos for multiple purposes, and the development of DL has allowed the capacity of computers to process this kind of data to increase exponentially in the last few years. Whereas before it was necessary to manually set a series of filters

to detect specific objects in an image, we can now train models that automatically learn more, more varied and better-adapted filters than any human could design for the objects we want. Every day we can automate more and more media processing tasks, but the more we get, the more we want to do.

1.2 Problem and Goal

In the context of working to improve the analysis of video sequences by computers and with deep learning as an established tool with more than proven results, this thesis has focused on the proposal of methods for analysing anomalies in video sequences using deep learning tools.

Of course, many different types of anomalies can be analyzed in a video sequence depending on the context of those anomalies. We focus on two cases.

The main focus of this thesis has been done by proposing new methods for the analysis and identification of foreground anomalies or objects in motion in relation to the background. The segmentation of the foreground with respect to the background is a process that traditional computer vision has used as a previous step to perform subsequent analysis of the foreground. Our approach to this problem has mainly consisted of working on noise robustness in static generic cameras using autoencoders or semantic segmentation networks in combination with probabilistic models. Meta-algorithms have also been proposed to reuse classical foreground segmentation methods, either by altering the size of the images or by pre-processing them to reduce their noise.

A second branch of work has been based on using object detection networks in the context of traffic to analyse vehicle behaviour. In this second line, we have analysed both dangerous overtaking on roads and fuel consumption based solely on video images.

This can be summarized into two main goals followed in this thesis:

1. Create methods for foreground detection that are robust to noise, both intrinsic to the recording and to artificial noise that may be present in the video sequence.
2. Explore the possibility of estimating object velocities from video images to analyze anomalous behavior.

To achieve these objectives, some general constraints guide and condition the research:

- Use only generic neural network models that are pre-trained or trained by unsupervised training to not rely on labeled data.
- Not to limit ourselves to using artificial neural network models, but to integrate them into pipelines that use other analysis tools.

Thus different work lines are established:

- Seek to increase the resolution of patch-level segmentation strategy so its benefits are maintained with a resolution similar to that of pixel-level methods.

- Study the reuse of classical foreground segmentation algorithms in combination with preprocessing techniques to add robustness to the system.
- Estimate the speed and pollution generated by a vehicle from video images.
- Estimate the speed of other cars on the road from images obtained using an onboard camera.

1.3 Methodology

In order to tackle the problems posed in this thesis, some general methodological principles have been followed:

- **Scientific Method**: All our projects adhere to the principles of reproducibility (setting up experiments that can be reproduced by other researchers). We also approach each problem under the classic scheme of Observing, Generating a Working Hypothesis, Experimenting to test it and Proving or Disproving the Hypothesis to obtain Conclusions.
- **Incremental methodology**: The objectives that have emerged in the thesis are based on the results obtained previously and these have guided the following experiments and tests to improve the system or to face other problems observed during a previous experiment.
- **Simple Implementation**: The implementations have been designed to be as simple, modifiable and understandable as possible.
- **External Evaluation Criteria**: As far as possible, established measures and datasets have been used so that the results can be easily understood and compared with those of other articles.
- **Comparison with other methods**: Whenever possible, the results of other methods for the same experiment have been included to provide a direct and objective reference to identify the advantages and disadvantages of our proposals.

1.4 Structure on this Thesis

This PhD thesis can be divided into three distinct parts, not including the appendices. The first part is Chapter 2 and includes the theoretical background of the thesis. This foundation underpins the works included in the thesis, which run from chapters 3 to 10 and make up the second part. Finally, the third part includes the conclusions and the most promising lines of future work.

The theoretical foundations include basic concepts of Probability Theory, the Anomaly Detection problem in its broadest sense and the specific problem of Foreground Segmentation to which most of the work in this thesis is devoted. The fundamentals of artificial neural networks are also described, including a formalisation of the layer concept (we are not aware of any other such formalisation) as well as the description and a brief explanation of the solutions to the problems of

Object Recognition, Object Detection, Semantic Segmentation and Instance Segmentation in images. It ends with an overview of the relationship between advances in hardware and advances in Deep Learning.

The second part includes the papers that support this thesis and each one corresponds to a chapter.

The first of the eight papers presented constitutes chapter 3 of this thesis and is entitled *The effect of downsampling-upsampling strategy on foreground detection algorithms* and was published in 2020 in the journal AIRe (Artificial Intelligence Review), which said year occupies Q1 position (14/139) of the JCR (Journal Citation Report) ranking in Artificial Intelligence category. The aforementioned work is framed within the study of obtaining the foreground by using previously published methods within a method that increases its performance. To this end, a method based on reducing the size of the frames before being processed is proposed along with and a study with extensive experiments to see its results.

The second of the papers presented is chapter 4 of this thesis and is entitled *Background subtraction by probabilistic modelling of patch features learned by deep autoencoders*. It was published in 2020 in the journal ICAE (Integrated Computer-Aided Engineering). The aforementioned journal was ranked Q2 in the JCR (Journal Citation Report) ranking in the Artificial Intelligence category (37/139). This work continues with the study of the detection of foreground anomalies through the proposed method specially designed to be resistant to noise based on the use of overlapping patches that are processed with a SDA (Stacked Denoising Autoencoder) before being analysed with a probabilistic model.

The chapter 5 and third work presented was entitled *Underground detection by probabilistic mixture models using semantic information from deep networks* and was presented at the 2020 international conference ECAI (European Conference on Artificial Intelligence). The conference was rated A in the CORE classification and A- in the GGS classification. The paper also deals with the FS problem by applying a probabilistic model to the information obtained from a Semantic Segmentation network.

The fourth work included and chapter 6 in this thesis is entitled *Deep autoencoder architectures for foreground object detection in video sequences based on probabilistic mixture models* and was presented at the ICIP (IEEE International Conference on Image Processing) 2020 congress, which had category B in the CORE classification and A- in the GGS classification. In this case, it consists of a study of the application of SDA (Stacked Denoising Autoencoder) neural networks and how their architecture and training conditions affect their application to the FS problem.

Chapter 7 is the fifth work and named *Foreground segmentation improvement by image denoising preprocessing applied to noisy video sequences*. It was presented at the 2021 SOCO (International Conference on Soft Computing Models in Industrial and Environmental Applications) conference, classified as *Work in Progress*. The work consists of a study on the use of various image preprocessing techniques in the context of a metamethod that provides noise robustness to other Foreground Segmentation methods. The idea is to apply classical filters or SDA networks to obtain lower noise versions of sequences that are processed with classical and very fast Foreground Segmentation methods.

The sixth work and chapter 8 of this thesis is entitled *Road pollution estimation from vehicle tracking in surveillance videos by deep convolutional neural networks.*

The paper was published in the journal ASoC (Applied Soft Computing) in 2021 and in that year the journal was ranked Q1 (23/145) in the AI category of the JCR classification. This work, unlike the previous ones, applies convolutional neural networks for traffic analysis. The proposed method uses vehicle information detected in video sequences by an Object Detection model to estimate speeds and pollution generated per time instant.

Chapter 9 and seventh work is entitled *Vehicle overtaking hazard detection over onboard cameras using deep convolutional networks* and was presented at the SOCO 2022 conference, again classified as *Work in progress*. The work continues the line of the previous one on estimating vehicle speeds from video sequences, but this time under the premise that the camera is in the vehicle and the aim is to identify abnormally high relative speeds to identify possible dangerous overtaking.

The last work and chapter 10 of the thesis is entitled *Moving object detection in noisy video sequences using deep convolutional disentangled representations* and was presented at the 2022 ICIP (IEEE International Conference on Image Processing) conference, classified as A- in the GGS classification and B in the CORE classification. This work closes the thesis with a further contribution to the problem of noise-robust Foreground Segmentation using SDA networks. This work is focused on solving the slow-resolution problems derived from patch-level processing. Through the exclusive use of convolutional layers and the application of probabilistic analysis to the depth of the channels, the advantages of applying an SDA are achieved without the restriction of the patches or the need to overlap them with the consequent increase in computation requirements.

The third and final part of this thesis recapitulates the conclusions drawn from all the work presented and presents the most promising lines of future work.

Chapter 2

Fundamentals and State of the Art

-Those analysis droids only focus on symbols. Huh! I should think that you Jedi would have more respect for the difference between knowledge and... wisdom.
-Well, if droids could think, there'd be non of us here, would there?

Star Wars Episode II, The Attack of the Clones.

Abstract: This chapter introduces information about the technical and scientific context of this PhD thesis. It includes the basis of Gaussian Models, foreground anomaly detection and ANN (Artificial Neural Networks) fundamentals.

2.1 Basic Probability

Dealing with uncertainty is a necessity when working with most real-world problems. It may arise due to inaccurate sensors involved in the problem or unknown information that needs to be modelled in order to take it into account and draw conclusions based on the data we observe.

A key element in dealing with this uncertainty is Probability Theory. It is this that provides a consistent framework for manipulating and measuring uncertainty, as well as the forms in which it presents itself.

In this work, we will only present some basic concepts of probability following Bishop (2006) and used in the works later presented. To begin with, we will go to the most basic formal element following: we say $p(X = x_i) = a$ when a random variable X which could take values $x \in \{x_1, ..., x_m\}$ has a probability of a to take

value x_i with $a \in [0,1]$. The value a would correspond to the number c_i of times X has taken value x_1 of M attempts and would be calculated as $a = \frac{c_i}{M}$ with $M \to \infty$.

Likewise we can state that, given two random variables X and Y, which respectively can take on values $x \in \{x_1, ..., x_m\}$ and $y \in \{y_1, ..., y_n\}$, $p(X = x_i, Y = y_j)$ states the probability of X take value x_i and Y take value y_i simultaneously. We call it the joint probability of $X = x_i$ and $Y = y_i$ and is simmetric so $p(X = x_i, Y = y_j) = p(Y = y_j, X = x_i)$

We know the sum rule of probability:

$$p(X = x_i) = \sum_{j=1}^{n} p(X = x_i, Y = y_j) \tag{2.1}$$

When we state the probability of $X = x_i$ only when $Y = y_i$ we call it the conditional probability of $X = x_i$ given $Y = y_i$ and we denote it $p(X = x_i \mid Y = y_i)$.

The following well-known relationship is called the product rule of probability:

$$p(X = x_i, Y = y_i) = p(X = x_i \mid Y = y_i)p(Y = y_i) \tag{2.2}$$

For the sake of simplicity, we will simplify the notation from now on from $p(X = x_i)$ to simply $p(X)$ to denote the distribution over random variable X or $p(x_i)$ to denote the probability for value x_i.

2.1.1 Bayes' Theorem

From joint probability symmetry and equation 2.2, the following relationship between conditional probabilities can be obtained:

$$p(Y \mid X) = \frac{p(X \mid Y)p(Y)}{p(X)} \tag{2.3}$$

This relation is called the Bayes' Theorem (Bayes (1763)) and plays a very important role in Machine Learning in the works that will be presented in this thesis since it allows to obtain the conditional probability of Y given X from the conditional probability of X given Y and the probability of X and Y. It, therefore, allows us to deduce unknown probabilities from other known probabilities.

Given a model m_θ with parameters σ to fit some set of data $D = d_1, ..., d_T$. Even if we can obtain statistical and probability information from D, it would be also desirable to address and measure the uncertainty about the choice of the model parameters θ. Some assumptions usually can be made to obtain a *prior* probability distribution $p(m_\theta)$ before exposing the model m_θ to D. After observing the data D, we can estimate how likely is to observe D from m_θ, so we have $p(D \mid m_\theta)$ and this is often called *likelihood*. If we can measure the probability distribution over D, we also have $p(D)$, so we can apply Bayes' theorem:

$$p(m_\theta \mid D) = \frac{p(D \mid m_\theta)p(m_\theta)}{p(D)} \tag{2.4}$$

and obtain how good is m_σ after observing data D. We usually call $p(m_\theta \mid D)$ *posterior* probability.

2.1.2 The Gaussian Distribution

For a single continuous variable x, the *Normal* or *Gaussian* distribution is defined as follows:

$$N(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \tag{2.5}$$

which is defined by parameters *mean* μ and *variance* σ^2 and σ is called *standard deviation*.

For a D-dimensional vector $\mathbf{x}$ of continuous variables, we can also define the Gaussian distribution as follows:

$$N(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{\mid \boldsymbol{\Sigma} \mid^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \tag{2.6}$$

with D-dimensional vector $\boldsymbol{\sigma}$ called mean, covariance $\boldsymbol{\Sigma}$ is a $D \times D$ matrix and $\mid \boldsymbol{\Sigma} \mid$ is the determinant of $\boldsymbol{\Sigma}$.

2.2 Anomaly Detection in Images and Videos

2.2.1 Generic Anomaly Detection

Anomaly detection in the generic sense consists of, given a set of data, identifying those data that do not match the pattern or that represent an unusual event. More formally we could say that given a data set D described by a probability distribution $p(D)$, the problem is to decide whether a data d fits the distribution $p(D)$ or not.

A huge variety of problems of all areas and levels of complexity can fit into this general definition. It may range from identifying unusual banking transactions for detecting crime to identifying anomalous behaviour on the road in order to prevent accidents. This PhD thesis is mainly focused on identifying foreground anomalies, which we will discuss specifically in the section 2.2.2, and includes some works focused on traffic, but in this section we will give a general overview of the problem.

The criteria for deciding what is and what is not an anomaly is fuzzy and depends very much on the problem. There is no general definition of an anomaly and any approach to solving the problem involves deciding that criterion.

It is important to note that the distribution $p(D)$ over the data can be arbitrarily complex depending on the problem and is perhaps unknown a priori. Therefore, the distribution $p(D)$ must first be identified and several scenarios are possible:

- Data D does not include unidentified outliers. This case allows us to analyze D to obtain $p(D)$, but with no outliers reference it's very difficult to create criteria to define what is an outlier.
- Data D includes identified outliers. This case is the traditional supervised learning and allows us to obtain $p(D')$ with $D' \subset D$ with no outliers.
- Data D includes unidentified outliers. This case will make it difficult to obtain $p(D)$ and maybe criteria to estimate which elements are likely to be outliers should be applied to transform it to the previous scenario.

According to Salehi et al. (2021), Generic Anomaly Detection can be divided into several categories:

- ND (Novelty Detection) is a name often used for Anomaly Detection and in the literature their use is usually interchangeably. The premise to face this problem is to classify between normal and outlier samples. The system to perform the task can be trained in a supervised or unsupervised way but, in the latter case, some process to select normal samples needs to be applied during sampling.
- OSR (Open-Set Recognition), unlike ND, is not a binary classification problem, but a multi-class classification problem typically approached as a machine learning problem. In this problem, there are K classes regarded as normal that are provided to the system at training time. Once trained, the system is exposed to N classes, of which $N - K$ are unknown and K are known classes. The objective is to classify known classes while identifying the unknown ones. Different subtypes of this problem can be considered depending on what labeled information is provided to the system.
- OOD (Out-of-Distribution Detection), like OSR, is a multi-class classification problem but in this case, it seeks to identify classes that are semantically different (e.g. two different datasets) from the categories used during training.

2.2.1.1 Some Machine Learning approaches to Anomaly Detection

The following ML (Machine Learning) approaches are options to approach an anomaly detection problem.

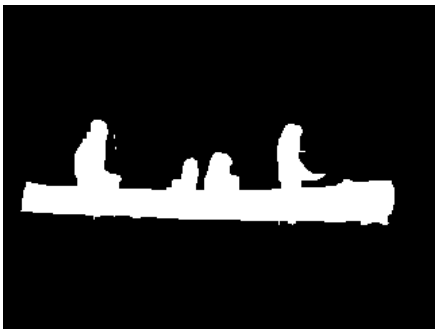
- Isolation Forests: This approach assumes outliers will be more easily isolated than normal data. The idea is to create a Decision Tree and use the depth of the tree as criteria to identify anomalies. The fewer splits needed to classify the data, the more anomalous (e.g. Liu et al. (2008)).
- One-Class Support Vector Machine: The approach is to use a SVM to learn the decision boundary needed to encapsulate normal data in high-dimensional space. Incoming instances falling out of the boundary are classified as an anomaly (e.g. Amer et al. (2013)).
- Distance-based methods: These methods measure the distance between instances to create clusters. These approaches use some predefined number of clusters or some other parameters to decide how many clusters and which elements are included in them. That information will be later used to decide if the incoming data fits inside the clusters or if it's an anomaly (e.g. Mazarbhuiya y Shenify (2023))
- Artificial Neural Networks: if the data is labelled, ANN can be used to classify between anomaly or normal as with any other binary classification problem. If the data is not labelled or an unsupervised approach is desired, ANN can still be used to process data within a pipeline. This approach is used in various works within this thesis (e.g. Pang et al. (2022)).

2.2.2 Background Subtraction

2.2.2.1 The Problem

BS (Background Subtraction), also called FS (Foreground Segmentation) (in this work both names will be used interchangeably) is the problem of classifying each pixel of each frame in a video sequence as belonging to a genuine moving object or belonging to a static object. We call the former pixels foreground in the sense that we consider them to be the relevant ones in the image and the latter pixels background in the sense that we consider them to be the irrelevant ones. The objective is to know where in the image there is movement in order to focus computational resources on analysing those places since typically moving objects are the ones that require the most attention. For each incoming image, the process should output information about each pixel we typically show as a binary segmented image.

(a) Example of an incoming image to obtain the foreground segmentation.



(b) Example of foreground segmentation image.

Figure 2.1: FS example using image from Goyette et al. (2012).

Moving objects will depend purely on the context. They can be people in the context of a video surveillance camera in a bank, vehicles in the context of a traffic camera on a motorway, both in the context of a surveillance camera on a street or in general any other element whose moving behaviour is of interest for computational analysis.

When we defined the problem above, we specified *genuine* motion. What do we mean by genuine? Although we usually don't notice it because we don't pay attention to it, in everyday scenarios there are many elements in constant motion. A simple example is the leaf of a tree swaying in the breeze. Strictly speaking, it is moving, but it is a movement that we are not usually interested in. This uninteresting movement is often referred to as noise along with other effects that can be found in videos.

2.2.2.2 Solution

In order to face the problem of foreground segmentation, the usual approach is to make a characteristic assumption of the problem: the background is what we see most often in the video sequence. Based on this assumption, the methods are based on modelling the background as they see appropriate and handling the foreground

as the anomalies in relation to that background.

Following Garcia-Garcia et al. (2020), any proposed method to solve the BS problem can be summarised by answering the following four questions:

- How do we model the background?

  The fundamental question in any background subtraction method is what strategy it follows to model the background. There is likely to be a different strategy for each method, as this is the key element that defines them. Potentially any approach applicable to a binary classification problem can be applied, ranging from explicit strategies based on analysing the data by purely statistical tools (e.g. Wren et al. (1997) and Stauffer y Grimson (1999)) to implicit strategies based on training neural networks specifically to perform the classification (e.g. Zhang et al. (2015) and Zheng et al. (2020)). The modelling strategy of the background is one of the most relevant aspects of how the method will be able to deal with the problems discussed in the section 2.2.2.3.

- How do we initialize the first background model?

  Once we have established that a good modelling of the background is fundamental to be able to decide what is the background and what is foreground, it is necessary to obtain a first initialisation of this background model. Normally at this point, we continue working with the premise that the background is the most frequent thing in the sequence. Based on this, it is reasonable to take a temporal segment of the video sequence and obtain statistical values for each pixel or region. The mean or median of the values for a pixel is likely to be the background value for that pixel. A good initialisation of the background model is key to the correct functioning of the method. A background model that is far from reality makes it impossible for the method to work properly.

- How do we use the background model to classify each pixel into background or foreground?

  This point is closely related to how we model the background. Once we have a background model and a new image comes in, what strategy do we follow to classify each pixel? The strategy we follow will define how resilient our method is to change elements of the image such as noise or shadows.

- How do we update the background model?

  Ideally the background model, once initialised, would remain as a static reference for the rest of the processing. However, in no real case can we assume that this is the case. The background is likely to change over time due to changes in lighting, moving objects, etc. The method should therefore include a strategy to update its background model over time. Typically, once it is decided that the pixel of a new image belongs to the background, the background model will be updated to incorporate this new information to adapt progressively. This updating strategy and how sensitive it is can cause the background model to become corrupted, either because it is too far from the actual background or because it is so unspecific that any possible foreground is classified as background.

2.2.2.3 Common problems to face

(a) A group of single-jet fountains.

(b) A fountain with several jets pointing to different places.

Figure 2.2: Two Dynamic Background examples from Goyette et al. (2012). Both sequences contain water movements that must be classified as background.

Molina-Cabello (2018) explains some of the problems these methods face. We focus only on the problems in the static camera FS methods, as they are the only ones addressed in this thesis.

- **Shadows**: Due to ambient lighting it is extremely common for shadows to appear in a video sequence. This implies that any moving object is likely to be casting a shadow that is moving with it. These shadows, of course, are not considered elements of interest and should therefore be considered background unless they are cast on an object of interest (in which case the interesting thing is still not the shadow). Of course, the existence of shadows implies a change in the colours of the pixels and, therefore, they are likely to be mistaken by some models for genuine moving objects. This is especially the case if the lighting in the scene casts very dark shadows that cause a drastic colour change in the pixels. Moreover, as the colour change is not an isolated pixel but in sets of pixels that move consistently (because they mimic the movement of a real object), the task of correctly classifying a shadow can become even more complicated.

- **Camouflage** When an animal or insect has a similar colouring or pattern to its surroundings and thus makes it difficult to be seen, we call this camouflage. BS methods face exactly the same problem. When the foreground and background look alike, it is easy for the methods to confuse one with the other. This need not be due solely to the existence of patterns intentionally designed to confuse but to simple coincidences such as the appearance of a dark green car against a background of trees of a similar colour. Although to the human eye, the difference is noticeable, the closeness of colour may be enough to confuse some methods.

- **Illumination changes** As mentioned above, FS methods are based on the generation of a background model which is used as a reference to decide whether pixels in the following images are foreground. An abrupt change

in lighting (such as turning on a street lamp in a dark street or turning off a lamp in a house) can cause a radical change in the background such that the previously available model is no longer an appropriate reference and many pixels that should be classified as background are now classified as foreground. Gradual lighting changes are easier to cope with if the method has an appropriate means of updating its background model, but a radical change can lead to a background model so different from the real one that the method can't adapt to it again.

- **Dynamic Background** Previously, we refer to the movement we are interested in as genuine or relevant movement, distinguishing it from movements that, although they exist, are not of interest to us. There are quite frequent cases in which there is punctual or constant movement in the background that typically should not be classified as foreground. The movement of leaves and tree branches, and the ripples of seawater, rain and snow are examples of such movements that are not relevant to us, thus they can be considered part of the background. This is especially problematic when it occurs in large quantities, for example, when the method is confronted with adverse weather conditions. It can result in noise in the segmented image with foreground where it should not be or may lead the methods to generate or adapt their background model to a too general one that does not allow to distinguish anything as foreground, blinding it. Figure 2.2 shows examples.

- **Movement in background** Let's assume a situation where a vehicle is parked for a long time in our video sequence. Perhaps even from the beginning. This vehicle would typically be classified as part of the background and so will appear in the background model. Now let's assume that the owner of the vehicle gets in and drives off. In this situation, all the pixels occupied by the vehicle have suddenly changed and the gap left by the vehicle, although it should be classified as background, is often classified as foreground because it looks nothing like the background model that the method has learned. Figure 2.3 is an example of this problem.

- **Static Foreground** Let us now assume the situation is symmetrical to the previous point. There is a free parking space on the street recorded in our video sequence and a car park. Once parked for a while, common sense tells us that this car becomes part of the background, at least as long as it remains parked. However, this abrupt change in the background from what the method had previously learned is also a problem that many methods fail to recover from, and the car may remain in the foreground indefinitely.

- **Noise in the sequence** Of course, a video sequence is susceptible to intrinsic noise, not from what has been recorded but from the media on which it was recorded. A defective camera or excessive compression when saving or transmitting the sequence file can generate artefacts or artificial noise that alters pixel values. As with any colour change, this can lead to errors in classification. When creating or choosing a method applicable in real cases, this problem must also be taken into account.

(a) Original situation with two parked cars.

(b) End of sequence situation with a car driving away.

Figure 2.3: Example from Goyette et al. (2012) where two parked car stays for most of the sequence until one drives away. The background model must adapt to the new background without the missing car.

2.2.2.4 Applications

Several applications need or benefit from the identification of moving objects as a first step to work efficiently. Some examples:

- **Human Video Surveillance**: Identification and tracking of persons or human-made objects in urban or inter-urban environments. In this kind of application, traffic monitoring is very common, for which identifying cars and distinguishing whether they are moving or stationary is a key task to identify incidents, congestion or unwanted parking. It also has security applications in the context of train stations or airports, where identifying abandoned luggage is extremely important to prevent terrorist attacks.
- **Observation of Wildlife Behaviour**: The specific characteristics of this type of surveillance require non-invasive systems. The first step before being able to track an animal is to identify it in motion by distinguishing it from the background, which in contexts with many irregularly shaped elements such as plants or rough terrain may not be trivial for some animals.
- **Human-Machine Interaction**: Many applications that require the interaction of a human and a machine need a reliable separation between the human and the background. In this field, there are many examples of play, but also of interaction with robots.
- **Gesture Recognition**: The detection of gestures as sign language is a task of extreme relevance for many people and in a first step, like so many others we are mentioning, it involves identifying the hands in the image.
- **Video Encoding**: Many encoding systems for video transmission for video conferencing or streaming systems are based on transmitting as little information as possible in order to save bandwidth and allow for faster data transmission. This is often achieved by not sending repeated information, such as a static background in an image. Identifying the foreground accurately and quickly is necessary to send only the information of those pixels.

- **Background Substitution**: Everyday applications such as video conferencing often allow us to replace our background with another one. This simple change is based on effectively identifying the foreground. The same principle is used in cinematographic applications to replace the background of a scene. In this second case, a uniform background colour is usually used to easily identify the elements to be replaced, but it is not always possible to have such a perfect uniform background or to dispense with the appearance of certain elements that we do not want to appear on the screen.

2.2.2.5 Some previous proposals

Traditional FS proposals are based on direct applications of probabilistic distributions for modelling the background. Wren (Wren et al. (1997)) assumes a static pixel-level background model obtained from the first frames of the sequence and uses analysis of pixel variance over time to determine which pixels belong to the foreground. Grimson (Stauffer y Grimson (1999)) proposes a more robust model by using a set of Gaussian distributions to model the background and detect pixel-level changes over time. KDE (Elgammal et al. (2000)) uses kernel density estimation to model the background. A sliding window is used to estimate each pixel density in the current frame to be compared to previously estimated background density. Zivkovic (Zivkovic (2004)) combines moving average and variance to model background to detect changes in each pixel. The method includes an adaptative threshold to determine whether a pixel belongs to the background or not.

Later approaches based on Self-Organizing Maps (Kohonen (1982)) were proposed. SOBS (Maddalena y Petrosino (2008)) handle scenes with moving backgrounds, camouflage and gradual illumination changes. It is later improved by SOBS_CF (Maddalena y Petrosino (2010)), which is also studied in Maddalena y Petrosino (2012) to enhance its robustness by using a fuzzy model to deal with decision problems and introducing a spatial coherence mechanism into the background update algorithm. FSOM (López-Rubio et al. (2011)) also follows this general approach and measures correlation statistically in order to obtain similarity among nearby pixels thus it can be used to provide feedback to the process and improve detection performance.

Another important line of work is spatial-based analysis methods. They use comparisons regions-by-regions still obtaining pixel-level segmentation. LOBSTER (St-Charles y Bilodeau (2014)) proposes a highly efficient adaptative spatial-based approach by using a spatiotemporal binary similarity descriptor as a core component instead of relying on comparisons between pixel-level intensities. SuBSENSE (St-Charles et al. (2014) and St-Charles et al. (2015)) uses a non-parametric balanced model that includes pixel-level feedback loops and spatiotemporal feature descriptors as the previous one. PAWCS (St-Charles et al. (2016)) seeks to offer a more consistent method based on using word dictionaries from background samples used to build smaller pixel models to enable a more stable long-term detection in combination with frame-level dictionaries and local feedback mechanism. MFBM (López-Rubio y López-Rubio (2015)) proposes a probabilistic model able to handle a variable number of features and proposes to study RGB (Red-Green-Blue) along with a set of region-based features for each pixel to classify them.

Patch-level foreground segmentation follows a similar logic to spatial analysis-based methods to increase noise robustness: the information of the specific pixels is

not as important as that of the region they are in. In this case, the use of patches restricts the definition of the segmentation but is a necessary condition for the use of autoencoders to be able to analyse the area as is done in Zhang et al. (2015), García-González et al. (2018), García-González et al. (2019a) and García-González et al. (2019b). This limitation will be addressed later in this thesis (chapter 10).

Other ML approaches like Competitive Learning has been also used. CL-VID (López-Rubio et al. (2018)) proposes a dual learning mechanism to manage input distribution with progressive change over time in order to adapt to a background that changes slowly. No supervised ML method is here reviewed since is not the main approach followed in this thesis.

2.3 Fundamentals of Neural Networks and Deep Learning

2.3.1 Artificial Neurons

ANN are biologically inspired mathematical models remotely mimicking the workings of the brain (McCulloch y Pitts (1943)). Figure 2.4 shows a comparison between both, biological and artificial, models. Due to the learning capacity, it is included within ML field.

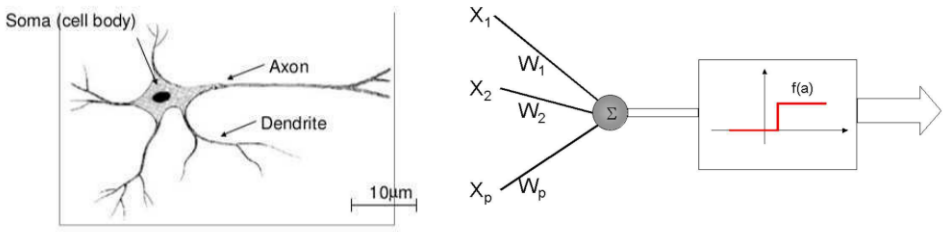


Figure 2.4: Comparison between natural and artificial neuron schemes from Rosa (2013).

In favour of generalisation, we will consider as an artificial neuron any unit of computation given an input x, returning an output y, i.e. an artificial neuron can be any function $y = f(x)$. We will distinguish artificial neurons whose computation is fixed and pre-established so that y depends only on its input x; from artificial neurons whose computation implies some learned values θ, which we will denote f_θ. Following the established convention, we call θ as weight, but f_θ will be referred to specifically as *learning artificial neurons*. For the sake of simplicity, we will omit *artificial* since there is no natural neuron-related work in this thesis and thus, no confusion is possible. Therefore, our explanations in this chapter will be about *neurons* and *learning neurons*, although we are aware that the standard convention is to assume that the term *neuron* implies a *learning neuron*.

Due to the critical relation with this thesis research, we will focus on FNNs (Feedforward Neural Network), MLPs (Multilayer Perceptron) and CNNs (Convolutional Neural Network).

2.3.1.1 Classical Learning Neuron

Following Cabello (2022), the classical learning neuron (u_θ) is a simple weighted sum and a bias b followed by an activation function f. Given an input vector $\mathbf{x}$, $u_\theta : \mathbb{R}^m \to \mathbb{R}$:

$$y = u_\theta(\mathbf{x}) = f(\begin{bmatrix}\omega_1, \cdots, \omega_n\end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} - b) = f((\sum_{i=1}^{n} \omega_i x_i) - b) \tag{2.7}$$

with $\theta = \begin{bmatrix}\omega_1, \cdots, \omega_n\end{bmatrix}$.

By itself, the processing power of a single learning neuron is very limited, not able to reproduce a simple logic function such as XOR (Minsky y Papert (1969)). On the other hand, the use of several learning neurons sequentially connected allows the creation of a universal function approximator (Hornik et al. (1989)). Therefore, the number and connections between neurons are key elements to facing hard problems. This order is called *network architecture* and is, along with the *learning process*, what will define each neuron role and overall network working.

If we observe equation 2.7, a neuron u_θ is a linear model followed by a function f. In order to approximate nonlinear functions, instead of applying u_{θ_1} to $\mathbf{x}$ we could apply u_{θ_1} to a transformed input $\phi(\mathbf{x})$. The ANN strategy is to learn ϕ transformation using a second u_{θ_2}. This is the insight behind the sequential application of neurons and its greater effectiveness in approximating functions (Mhaskar y Poggio (2016)) is the basis for the use of DL (Deep Learning).

2.3.1.2 Convolutional Learning Neuron

Given an input $\mathbf{x} \in \mathbb{R}^{h \times w \times d}$, a 2-dimensional convolutional neural unit $c_\theta : \mathbb{R}^{h \times w \times d} \to \mathbb{R}^{h' \times w'}$ with

$$h' = \left\lfloor \frac{h - k_h + 2p_h}{s_h} + 1 \right\rfloor \qquad \left\lfloor w' = \frac{w - k_w + 2p_w}{s_w} + 1 \right\rfloor \tag{2.8}$$

is defined as follows:

$$y = c_\theta(\mathbf{x}) = f((\sum \theta \odot x_i) - b) \tag{2.9}$$

with weights $\theta = \begin{bmatrix} \begin{bmatrix} \omega_{1,1,1} & \cdots & \omega_{1,k_w,1} \\ \vdots & \ddots & \vdots \\ \omega_{k_h,1,1} & \cdots & \omega_{k_h,k_w,1} \end{bmatrix} & \cdots & \begin{bmatrix} \omega_{1,1,d} & \cdots & \omega_{1,k_w,d} \\ \vdots & \ddots & \vdots \\ \omega_{k_h,1,d} & \cdots & \omega_{k_h,k_w,d} \end{bmatrix} \end{bmatrix}$
and $\odot$ as 2-dimensional convolutional function with kernel sizes k_h and k_w, padding values p_h and p_w and stride values s_h and s_w. Usually kernel is square so $k_h = k_w$, $p_h = p_w$ and $s_h = s_w$.

2.3.1.3 Max Pooling

Given an input $\mathbf{x} \in \mathbb{R}^{h \times w}$, a 2-dimensional MP (Max Pooling) neuron $MP : \mathbb{R}^{h \times w} \to \mathbb{R}^{h' \times w'}$ with h' and w' following equation 2.8:

$$\mathrm{MP}(\mathbf{x}) = \begin{bmatrix} max(NB(\mathbf{x}, k'_h, k'_w)) & \cdots & max(NB(\mathbf{x}, k'_h, w - k'_w)) \\ \vdots & \ddots & \vdots \\ max(NB(\mathbf{x}, h - k'_h, k'_w)) & \cdots & max(NB(\mathbf{x}, h - k'_h, w - k'_w)) \end{bmatrix} \tag{2.10}$$

$$\text{with } NB(\mathbf{x}, i, j) = \begin{bmatrix} x_{i-k'_h, j-k'_w} & \cdots & x_{i-k'_h, j+k'_w} \\ \vdots & \ddots & \vdots \\ x_{i+k'_h, j-k'_w} & \cdots & x_{i+k'_h, j+k'_w} \end{bmatrix}, k'_w = \left\lfloor \frac{k_w}{2} \right\rfloor \text{ and } k'_h = \left\lfloor \frac{k_h}{2} \right\rfloor.$$

MP (Max Pooling) layers are usually applied after a Convolutional Layer so in the following layers only the greater values activations coming from the Convolutional layer are used.

2.3.2 Layers

Even if it is possible to define ANN architectures based only on neurons units (Cabello (2022); Schmidhuber (2015)), most modern neural networks are not defined using just neurons due to the necessity of abstraction to reduce the complexity. Instead, the *layer* is the basic element defining most of ANN architectures nowadays. François Chollet states *"The fundamental data structure in neural networks is the layer"* and calls the layer *"The building blocks of deep learning"* (Chollet (2017)). Even with that importance, it is usually overlooked when approaching formal analysis. So in order to define ANN architectures formally, we will first define the key concept *layer* to allow a homogeneous interpretation of the research work in the following chapters.

A *layer* is an abstraction over the neural unit and neurons are functions so we will define *layer* l as an ordered set of functions f:

$$l = [f_1, f_2, ..., f_n] \tag{2.11}$$

with $f_i : \mathbb{R}^m \to \mathbb{R}^o$ and the same *potential* inputs $\mathbf{x} \in \mathbb{R}^m$ so $l : \mathbb{R}^m \to \mathbb{R}^{no}$. A restriction over layer l is all f_i must share the same kind of neuron.

We use the notation *potential* input because not all neurons in the layer need to share the same inputs, but all their inputs must be subsets of the same set $\mathbf{x}$ as follows:

$$l(\mathbf{x}) = [f_1(\mathbf{x_1}), f_2(\mathbf{x_2}), ..., f_n(\mathbf{x_n})] \tag{2.12}$$

where $\mathbf{x} = [x_1, x_2, ..., x_m]$ and for all $i \in \{1, ..., n\}$, $\mathbf{x_i} = [x_{j_i}, ..., x_{k_i}]$ with $1 \leq j_i < k_i \leq m$.

Same way, we can name *learning layer* l_θ as one following equations 2.11 and 2.12 with an ordered set of n learning neurons (f_{θ_i}):

$$l_\theta = [f_{\theta_1}, f_{\theta_2}, ..., f_{\theta_n}] \tag{2.13}$$

so given input $\mathbf{x} \in \mathbb{R}^m$,

$$l_\theta(\mathbf{x}) = [f_{\theta_1}(\mathbf{x_1}), f_{\theta_2}(\mathbf{x_2}), ..., f_{\theta_n}(\mathbf{x_n})] \tag{2.14}$$

where $\mathbf{x} = [x_1, x_2, ..., x_m]$ and for all $i \in \{1, ..., n\}$, $\mathbf{x_i} = [x_{j_i}, ..., x_{k_i}]$ with $1 \leq j_i < k_i \leq m$.

Therefore layer l_θ weights are $\theta = [\theta_1, \theta_2, ..., \theta_n]$.

Both the topology of connections with previous layers and the kind of neurons will define what kind of *layer* l is.

2.3.2.1 Dense Layer

We call dense layer $D_\theta : \mathbb{R}^m \to \mathbb{R}^n$ a learning layer whose learning neurons are classical u_θ and all of them share all inputs so it is defined by a restricted version of equation 2.14:

$$D_\theta(\mathbf{x}) = [u_{\theta_1}(\mathbf{x}), u_{\theta_2}(\mathbf{x}), ..., u_{\theta_n}(\mathbf{x})] \tag{2.15}$$

with input $\mathbf{x} \in \mathbb{R}^m$.

2.3.2.2 Convolutional Layer

Analogue to dense layer, we call convolutional layer $C_\theta : \mathbb{R}^{h \times w \times d} \to \mathbb{R}^{h' \times w' \times n}$ a learning layer whose learning neurons are convolutional c_θ all with the same input $\mathbf{x} \in \mathbb{R}^{h \times w \times d}$:

$$C_\theta(\mathbf{x}) = [c_{\theta_1}(\mathbf{x}), c_{\theta_2}(\mathbf{x}), ..., c_{\theta_n}(\mathbf{x})] \tag{2.16}$$

with and h' and w' defined by equations 2.8.

An ANN architecture involving convolutional layers is what we often denote as CNN (Convolutional Neural Network).

2.3.3 Feedforward Neural Network

According to Ian Goodfellow, FNN (Feedforward Neural Network) are those in which the information flows from input to output through intermediate computations used to define f with no feedback connections so, no output is fed back into itself (Goodfellow et al. (2016)).

Given a set of layers $L = \{l_1 : \mathbb{R}^{m_1} \to \mathbb{R}^{n_1}, ..., l_o : \mathbb{R}^{m_o} \to \mathbb{R}^{n_o}\}$ and a set of input vectors $I = \{\mathbf{x_1} \in \mathbb{R}^{s_1}, ..., \mathbf{x}_p \in \mathbb{R}^{s_p}\}$, an ANN architecture A could be defined as a directed graph $A = \{V, E\}$ with $V = L \cup I$ as vertices and E a set of paired vertices $E \subseteq \{(v_i, l_j) \mid (v_i, l_j) \in V \times L\}$ so given the following function $size : V \to \mathbb{N}$:

$$size(v_i) = \begin{cases} n_i & v \in L \\ s_i & v \in I \end{cases} \tag{2.17}$$

the following must be true:

$$\forall l_j, \sum_{(v_x, l_j) \in E} size(v_x) = m_j \tag{2.18}$$

so $\forall x\{(v_x, l_j)\} \subseteq E$ implies l_j takes as input concatenation of all v_x outputs if $v_x \in L$ or directly v_x if $v_x \in I$.

A FNN is an ANN whose architecture graph contains no cycle (Schmidhuber (2015)) so the following must be true for any walk $w = [v_1, ..., v_p] \mid v_i \in V$ in G:

$$\forall v_i, v_j \in w, i \neq j \rightarrow v_i \neq v_j \tag{2.19}$$

It is equivalent to state E is a strict partially ordered relation on V.

2.3.3.1 Multilayer Perceptron

The most naive approach to build FNN, once the dense layer is defined, is to stack them sequentially in order to create a MLP (Multilayer Perceptron). A MLP is an FNN with an architecture defined by:

- An input *layer* l_{θ_I}: layer whose inputs are the inputs of the network.
- n sequential hidden *dense layers* $l_{\theta_1}, ..., l_{\theta_n}$.
- A *dense* output *layer* l_{θ_O}: layer whose outputs are the outputs of the whole network.

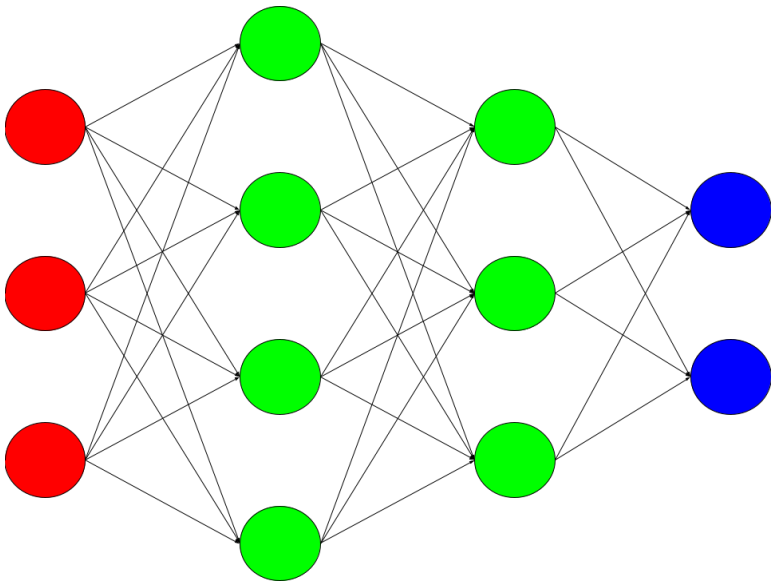


Figure 2.5: MLP example scheme. Red circles represent input *layer* neurons, green circles represent two hidden *layers* and blue circles represent output *layer*. Arrows represent one-direction (left to right) connections between networks.

This way, a MLP will implement a function $f : \mathbb{R}^m \rightarrow \mathbb{R}^s$ with $s = size(l_{\theta_O})$ and input $\mathbf{x} \in \mathbb{R}^m$ as the following:

$$f(\mathbf{x}) = l_{\theta_O}(l_{\theta_n}(...(l_{\theta_1}(l_{\theta_I}(\mathbf{x})))...)) \tag{2.20}$$

2.3.4 Activation Function

On both equations 2.7 on page 18 and 2.9 on page 18, there is a f element we simply named as activation function.

Some common activation functions are:

- Linear $f : \mathbb{R} \to \mathbb{R}$

$$f(x) = x \tag{2.21}$$

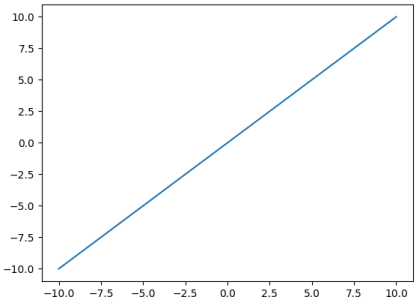


Figure 2.6: Linear function plot.

- Rectified Linear Unit (ReLU) $f : \mathbb{R} \to [0, \infty]$

$$f(x) = max(x, 0) \tag{2.22}$$

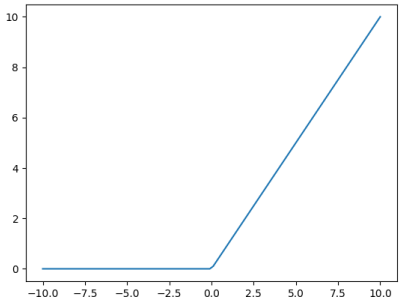


Figure 2.7: ReLU function plot..

- Sigmoid (logisctic) $f : \mathbb{R} \to [0, 1]$

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2.23}$$

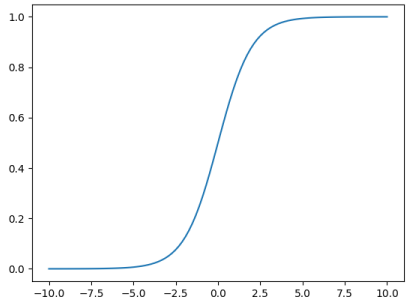


Figure 2.8: Sigmoid function plot.

- Hyperbolic Tangent $f : \mathbb{R} \to [-1, 1]$

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{2.24}$$

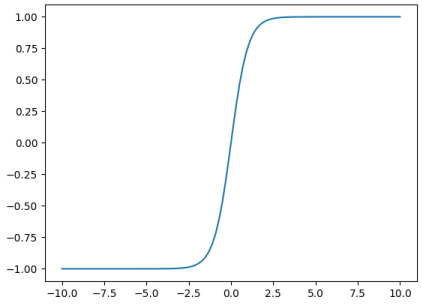


Figure 2.9: Hyperbolic tangent function plot.

2.3.5 Loss Functions

In order to perform the training process, it is necessary to measure the amount of error a ML model makes in predicting on a set of data. For this purpose, error functions are used which, in the case of ANN models, require that the gradient can be traced through the function (which is differentiable for the general case). Some of the most frequent given n samples to compute the loss are the following:

- Mean Square Error

  Measures the average squared difference between the reference $\mathbf{y}$ and the prediction $\mathbf{x}$. It is a loss commonly used on unsupervised problems as autoencoders since $\mathbf{y}$ can be a modified version of $\mathbf{x}$. It is also commonly used in regression problems.

$$MSE(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=0}^{n}(x_i - y_i)^2}{n} \tag{2.25}$$

- Mean Absolute Error

  As the previous one, this function is commonly used on regression problems. The function itself is not differentiable when the predicted value and the objective value are exactly the same but software implementations assume that is a rare case and includes ways to handle these exceptions.

$$MAE(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=0}^{n}|x_i - y_i|}{n} \tag{2.26}$$

- Binary Cross Entropy Loss

  The usual function to use when working on a binary classification problem. The value increases as the predicted probability $\mathbf{x}$ and the label $\mathbf{y}$ diverges.

$$BCEL(\mathbf{x}, \mathbf{y}) = -\frac{\sum_{i=0}^{n}(x_i log(y_i)) + (1 - x_i) log(1 - y_i))}{n} \tag{2.27}$$

- Cross Entropy Loss

  The usual function to use when working on classification problems with c classes.

$$CEL(\mathbf{x}, \mathbf{y}) = -\frac{\sum_{i=0}^{n}\sum_{j=0}^{c}(x_j log(y_i))}{n} \tag{2.28}$$

2.3.6 Training Strategies

ANN, as any other ML method, learns from data to approximate a function f. It's a natural question to ask what data is needed in order to perform the training.

Although the term *training strategy* could be used with a multitude of interpretations in the context of ML, in this case, we are going to refer to the main general strategies regarding the use of data and their needs. We will not go into other aspects such as the amount of data needed, balancing strategies, etc.

Once we have established that the learning process of an ANN (i.e. the execution of the algorithm that allows obtaining the necessary parameters θ to approximate the desired function f) requires a set of data, usually called *dataset*, it is natural to ask: What does a *dataset* consist of? Does the data have to meet any requirements?

The previous subsection has described the most common loss functions used during neural network training and while some may only need raw or procedurally generable data (e.g. MSE), others need additional information for that data (e.g. Categorical Cross Entropy). These needs depend on the problem we want to solve and how we want to solve it. Ideally, as little data as possible should be required and the data should not demand additional information given by any human. Unfortunately, although many problems and approaches can satisfy this condition, in other problems no approach has been found or no desirable results have been achieved.

Based on this need for data we can divide training strategies into three categories: *supervised training*, *unsupervised training* and *semi-supervised training*.

2.3.6.1 Supervised Learning

We call *supervised learning* any ML task that requires not only the data obtained by some automatic method (images from a camera, points from a LIDAR, signals from an electrocardiogram, etc.) but also some information associated with that data that requires human intervention to be obtained. Hence the use of the term *supervised*.

This associated information can be quite varied. In problems involving classification, we usually refer to the associated information as the *label* that the system has to learn, although *labelled data* has become a generic term for data with associated information.

Obtaining this information is often a major concern when tackling a problem. There are many problems that we speculate would be tractable if we had the necessary annotated data, but it does not exist or does not exist in enough quantity. Some leading researchers, such as Andrew Ng, have already expressed the view that AI should move away from focusing on generating new models to focusing on obtaining, cleaning and annotating data, which they see as much more meaningful in improving existing solutions. This implies a shift from the model-centric paradigm to the data-centric paradigm[1].

Generating this annotated data can require a lot of human labour, it has required in some cases ingenious solutions (the CAPTCHAs we often do on the internet to prove we are human, for example) and from it, a whole industry has been created with a multitude of software tools (e.g. CVAT[2]) and companies dedicated solely to data annotation. It has also brought to the table some ethical dilemmas linked to AI. For example: in order to annotate videos and classify them into different types of violence, people need to spend hours systematically watching violent videos to annotate them, with the following psychological damage that such exposure could cause.

The biggest practical problem, however, is that manual data annotation is not scalable. If annotating a piece of data costs on average x time (and therefore y money), annotating z pieces of data costs on average zx time (and consequently zy money). In a scenario where learning systems can consume huge amounts of data, annotation becomes one of the bottlenecks and a prohibitive constraint for most companies and researchers, who have to limit themselves to using public datasets or working with small amounts of annotated data.

Supervised learning is generally applied to problems of image and non-image classification, object detection and semantic or instance segmentation.

2.3.6.2 Unsupervised Learning

Although *supervised learning* can be applied to any problem for which we have labelled data, as mentioned above this is not an ideally desirable solution for several practical and ethical reasons. The opposite *training strategy* is to devise an approach to the problem that requires only the data without annotations, which we call *unsupervised learning*.

Unsupervised learning is the ideal learning paradigm during training, as it *theoretically* allows for a fully automatable workflow. We say *theoretically* because

[1] `https://datacentricai.org/blog/opening-remarks/`
[2] https://github.com/opencv/cvat

the data usually requires some pre-processing (cleaning, rescaling, etc.), but this, once identified, usually is programmed and applied to all data in a scalable way. Although this is the desirable paradigm, in many cases it is not applicable because the unlabelled data does not contain the necessary information to approximate the f function or we simply have not devised a process to learn it without human intervention.

Still, there are many other cases where unlabelled data is more than enough. For example, Stacked Denoising Autoencoder training (we will talk about them in the following subsection 2.3.7) is usually performed using the original data and noise-modified versions by an automatic process (adding Gaussian or uniform noise, for example). This process does not require any human intervention and it is therefore completely unsupervised.

Another field of study in which learning is fully unsupervised is neural rendering. Derived from NeRF (Mildenhall et al. (2022)), neural rendering is a problem in which knowing the images and the (sometimes approximate) position of the cameras with which they were obtained is enough to learn a 3D scene since it compares the rays with the RGBs values, both obtainable in a fully automatic way.

2.3.6.3 Semi-Supervised Learning

We have established the need for data, in many cases labelled, and some of the problems involved in getting these annotations, which are nevertheless necessary to tackle many machine learning problems. As we have stated, *unsupervised learning* is the desirable strategy, but it is not always possible to use it. Therefore what happens when we want to tackle a problem without an unsupervised training approach and we do not have access to enough labelled data?

There is an increasingly popular intermediate approach: semi-supervised learning. This approach assumes the existence of an annotated subset of the data. Typically the difference between the size of the subset and the global dataset should be significant, but this is not always feasible. *Semi-supervised learning* uses the information obtained from the labelled subset to obtain more labelled data and thus make it feasible to apply *supervised learning* strategies based on methods to make the labelling process scalable.

In practice, this translates into many different approaches that are under development but already applicable. The most basic is data augmentation: applying transformations to existing labelled data through simple processes such as rotating images, adding objects, noise or fake weather effects to multiply the amount of labelled data. All these transformations should either not affect the associated information or have a corresponding process to alter it accordingly (such as repositioning the bounding boxes to match a rotated image).

There are more advanced processes that rely on having an already trained model m that works reasonably well (the data it was trained on would be the already labelled subset) and using it to obtain new labelled data. Of course, if this process just worked, it would imply that the model m already perfectly approximates the desired function f and training another one would not be of interest. However, we can have a model m that works reasonably well in some cases (e.g. detecting objects when they are big enough) and transform data where it would fail into simplified versions where it gets it right to obtain the correct labelling using m. A good example of this case is García-Aguilar et al. (2023a), which uses Super-

Resolution to obtain labels for small objects that object detection models normally fail to detect.

Another approach is to generate fully synthetic data with their labels already associated using generative models such as GANs (Zhang et al. (2021)) or Diffusion Models (Trabucco et al. (2023)). In these cases, the labelled subsets are the data with which these models have been initially trained. These methodologies are still in their early stages of development, but their refinement has the potential to drastically reduce the existing bottleneck in data collection and labelling.

2.3.7 Autoencoders

Let an AE (Autoencoder) be an Artificial Neural Network that approximates the following function 2.29:

$$\tilde{\mathbf{x}} = g\left(f\left(\mathbf{x}\right)\right) \tag{2.29}$$

with:

$$f\ :\ \mathbb{R}^H \rightarrow \mathbb{R}^L \qquad g\ :\ \mathbb{R}^L \rightarrow \mathbb{R}^H \tag{2.30}$$

where $\tilde{\mathbf{x}} \in \mathbb{R}^H$ is a reconstructed version of $\mathbf{x} \in \mathbb{R}^H$ and f and g are the *encoder* and *decoder* section of the AE respectively. Therefore, an AE approximates the Identity function and, with that goal, it is usually trained in order to minimize the following reconstruction error $\mathcal{E}$:

$$\mathcal{E} = \sum_{\mathbf{x} \in R} D(\mathbf{x}, \tilde{\mathbf{x}}) \tag{2.31}$$

with $R \subset \mathbb{R}^H$ as training set and D a differentiable distance measurement function between two arbitrary elements from $\mathbb{R}^H$ such as MSE (Mean Squared Error).

According to Goodfellow et al. (2016) and Schmidhuber (2015), AE have been involved in the field of unsupervised learning artificial neural networks since the 1980s with Ballard (1987).

The most common approach is to define $L < H$ in order to force a dimensionality reduction to get a latent vector $\mathbf{z} = f(\mathbf{x})$ containing $\mathbf{x}$ most important features from the last *encoder* layer.

We call DA (Denoising Autoencoder) an *autoencoder* trained to reconstruct a denoised version of the input (Zhang et al. (2015)). A DA is trained with corrupted inputs $\hat{\mathbf{x}}$ instead of $\mathbf{x}$ in order to reduce error from equation 2.32:

$$\mathcal{E} = \sum_{\mathbf{x} \in R} D(\mathbf{x}, g(f(\hat{\mathbf{x}}))) \tag{2.32}$$

Due to the dimensionality reduction, less relevant information should be ignored while most relevant information should be included in latent code $\mathbf{z} = f(\hat{\mathbf{x}})$ in order to be later reconstructed by g into a cleaner approximation of $\mathbf{x}$.

It is important to note that the term stacked is usually used to define an AE with *encoder* and *decoder* composed by more than a single layer, so it is common to read about SDA.

2.3.7.1 Variational Autoencoders

A VAE (Variational Autoencoder) is a generative model capable of generating new data not included in *dataset* R but similar to it in a consistent way.

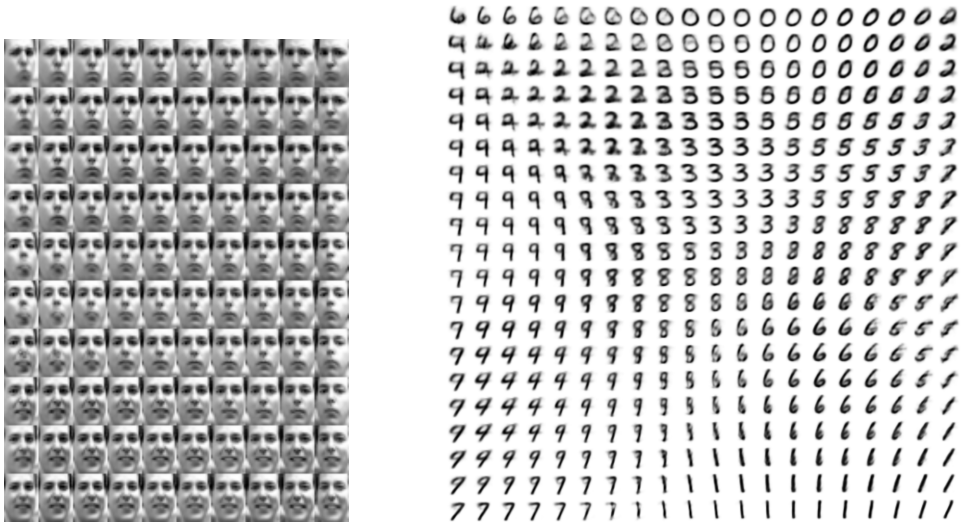
This feature could be assumed in the previously described autoencoders natively, since after all they learn a latent space between the *encoder* and the *decoder*, so it would be theoretically plausible to provide a random vector $\mathbf{z} \in \mathbb{R}^L$ to the *decoder* and expect the result $g(\mathbf{z})$ to be consistent with the *dataset* R it has been trained on. Unfortunately, this is not the case. If we try, for example, to interpolate the latent vectors $\mathbf{z_1}$ and $\mathbf{z_2}$ from two entries of the *dataset* R, we would expect to obtain intermediate versions of $g(\mathbf{z_1})$ and $g(\mathbf{z_2})$, but the discontinuities in the latent space lead to rough transitions and the output often is entirely different from the expected.

Kingma y Welling (2013) first defined VAE as an AE which learned latent vectors $\mathbf{z}$ are *Gaussian* distributions so *encoder* f (equation 2.29) outputs from input $\mathbf{x}$ should match the following:

$$f(\mathbf{x}) = \mathbf{z} = (\boldsymbol{\mu}, \boldsymbol{\sigma}) \tag{2.33}$$

And *decoder* g takes as input a sampled value from the *Gaussian* distribution obtained from the encoder.

$$\hat{\mathbf{x}} = g(\mathbf{s}) \qquad \mathbf{s} \sim N(\boldsymbol{\mu}, \boldsymbol{\sigma}) \tag{2.34}$$



(a) Learned Frey Face manifold.

(b) Learned MNIST manifold.

Figure 2.10: VAE learned manifolds visualization for two datasets from Kingma y Welling (2013)

This way, a VAE learns a latent space of *Gaussian* distributions, but still this not ensures any smoothness since learned distributions parameters could be afar and, in practice, work the same way as an usual AE. In order to penalize distributions to be very different, Kullback-Leibler divergence D_{KL} (Kullback y Leibler (1951)) is added to the usual loss function 2.31.

$$D_{KL}(P \parallel Q) = \sum_{i}^{R} = 1P(i)log(\frac{P(i)}{Q(i)}) \tag{2.35}$$

Usually, distribution $N(0,1)$ is taken as reference distribution when applying regularization using D_{KL} so the learned parameters of the *Gaussian* distributions do not grow too large or go too small and neither be too afar. This way a kind of overlapping between distributions is generated and the smoothness in the latent space is obtained so two similar inputs $\mathbf{x_1}$ and $\mathbf{x_2}$ should obtain similar latent representations $\mathbf{z_1}$ and $\mathbf{z_2}$.

In figure 2.10 two examples of this smoothness can be observed.

The proposed approach implies a problem. In order to train two neural models as encoder and decoder, equation 2.34 includes a sampling from an arbitrary *Gaussian* distribution that would lead to a non-differentiable computational flow. To avoid this, Kingma y Welling (2013) proposed the usually called *reparametrization trick* to deal with the distributions as if they were usual errors from Normal distribution.

$$\mathbf{s} \sim N(\boldsymbol{\mu}, \boldsymbol{\sigma}) \longrightarrow \mathbf{s} = \boldsymbol{\mu} + \boldsymbol{\sigma}\epsilon \qquad \epsilon \sim N(0,1) \tag{2.36}$$

Usually, GAN (Generative Adversarial Network) (Goodfellow et al. (2020)) or even more modern Diffusion Models (Ho et al. (2020)) have been preferred to VAE as a generative model, but deep VAE is still a field of research as a generative model (Child (2021)).

2.3.8 Object Recognition

Object Recognition, also known as Image Classification, is a classical CV classification problem consisting of the identification of the object appearing in an image. Typically a single object will appear in the front and classification will be performed among a set of predefined classes. Image 2.11 shows a perfect example of object recognition. As can be seen, the image is clear and usually there is only one object well-centred in it.

Although at first sight, it may seem the most obvious and straightforward CV problem, until the last decade there was no effective approach to solve it generically. The most classical approach (e.g. Bay et al. (2008)) was based on manually writing features detection filters to multiply the image with. Each filter should identify features in the image which, when combined, would allow for classification. With the advances in ML and its usefulness in solving classification problems, different ML approaches were developed to automatically learn how to identify the image and with them came large annotated images datasets as PASCAL[3] (Everingham et al. (2010)) or ImageNet[4] (Deng et al. (2009)) and their associated competitions (they exist since 2005 and 2010 respectively). According to Mitchell (2019), SVM (Support Vector Machine) approaches were the leading ones in ImageNet competitions during 2010 and 2011. In 2012, Alex Krizhevsky, Ilya Sutskever and Geoffrey Hinton proposed AlexNet (Krizhevsky et al. (2017)) to face Imagenet competition and changed the field forever by getting almost 10% better result than the second-best

[3]http://host.robots.ox.ac.uk/pascal/VOC/
[4]https://www.image-net.org/

Figure 2.11: This photo of my dog is an example of an image usually used when recognizing objects.

method. From then until nowadays, ANN is the default approach to face any CV problem.

AlexNet was a CNN (Convolutional Neural Network)-based model designed after Yann LeCun's LeNet LeCun et al. (1989) but much deeper (more layers and filters by convolutional layer). The insight behind the idea is to use stacked CNN layers to identify features of increasing complexity and abstraction. MP layers are typically used after CNN layer to select the features with greater activation so only the best matching features flow deeper into the network. Finally a set of dense layers, often called *classification head* takes the higher level features as information to classify the image into some class. Figure 2.12 shows a section of AlexNet structure to illustrate it. The combination of several levels of CNN employing MP and dense layers is maybe the most influential approach in ANN applied to CV and it is the base of the approaches of more complex problems we will discuss in following subsections 2.3.9 and 2.3.10 as well it is one of the first DL success based on very heavy GPU computation.

2.3.9 Object Detection

Object Detection is the natural evolution from the problem discussed in the previous subsection (Object Recognition). Once the identification of a single object in an image is taken as practically solved, the next natural step is to identify more than one object in an image. The evolution is natural but not trivial, as the problem now involves knowing which objects are in the image and *where* they are. Given an image, Object Detection problem consists of obtaining a list of BBOX (Bounding Box) with objects positions along with lists of matching classes and scores thus for

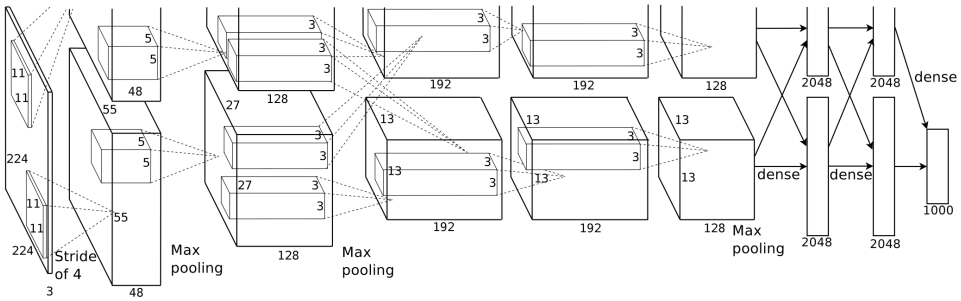


Figure 2.12: ImageNet structure example obtained from (Krizhevsky et al. (2017)). The basic idea of processing images using Max Pooling and Convolutional Neural Network layers before using dense layers to obtain final classification can be observed.

each position we know what class it is and the obtained score. This problem is one of the most widely researched problems in CV due to the very different situations where is involved.

There are two main approaches to solving this problem: Multi-stage and One-stage. We will discuss briefly their main features and their main examples.

2.3.9.1 Multi-stage Object Detection

We call Multi-stage OD (Object Detection) approach to models that, following R-CNN (Girshick et al. (2014)), use more than one model to perform the object detection task. Even if we call this approach *multi-stage* it could be also called *two-stage* since usually there are only two main stages:

- Region proposal extraction: the application of some model to obtain candidate regions to contain an object (point 2 from the pipeline in figure 2.13).
- Object recognition: to each candidate region an object recognition model is applied in order to obtain which object is within the region (points 3 and 4 from the pipeline in figure 2.13).

Some important examples of this approach are R-CNN Girshick et al. (2014), Fast R-CNN (Girshick (2015)) and Faster R-CNN (Ren et al. (2017)).

The main problem with this approach is the speed of processing. Since the approach is divided into two distinct phases, these models usually work in practice as if the data were processed first with one model and then with the other, even though part of the feature extraction should be common to both phases and therefore redundant if done twice.

2.3.9.2 One-stage Object Detection

Object detection underwent a major paradigm shift when the first single-stage OD methods were proposed. Unlike those described above, these methods do not separate object detection into a first proposal phase and a second recognition phase.

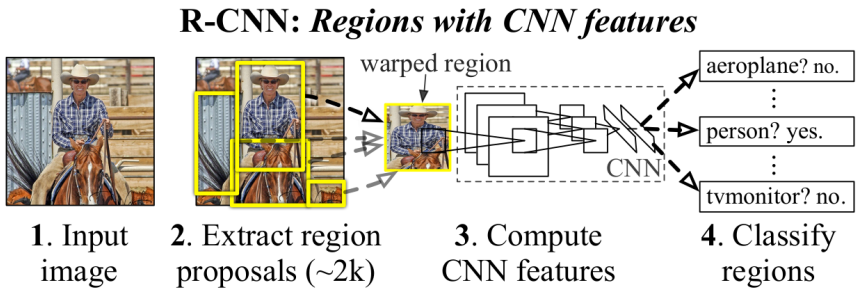


Figure 2.13: R-CNN scheme from Girshick et al. (2014).

Instead, they take advantage of the properties of CNN to, in a single step through the network, detect both objects' classes and where they are.

The key idea is simple but original. If we take any OR (Object Recognition) neural model as Krizhevsky et al. (2017), we can discard the last dense layers and perhaps some convolutional layers to keep the result of a convolutional layer of a size that suits us. From the activations of that layer, we can try to deduce not only what objects are in the image but also where they are. Instead of grouping all the information in a dense layer assuming that it is simply a classification problem, we can use convolutional layers to obtain the confidence in each class and the four values of the bounding box together with the probability that there is an object in that grid. Figure 2.14 shows the scheme of how a YOLO (You Only Look Once) model works.

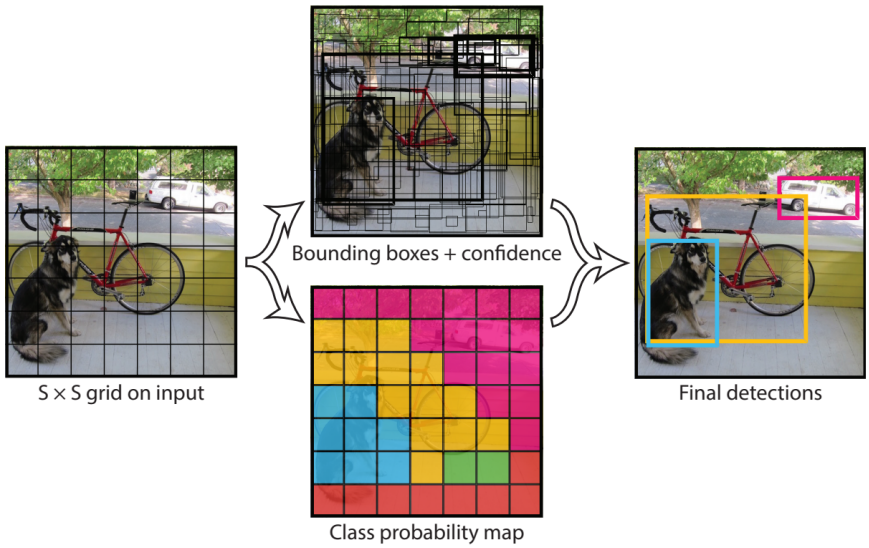


Figure 2.14: YOLO scheme from Redmon et al. (2016). The scheme illustrates how from a $S \times S$ grid bounding boxes, confidence and class probability is obtained to later be combined into final detections.

SSD (Single-Shot Detector) (Liu et al. (2016)) and YOLO (Redmon et al.

(2016)) models are the most important One-stage OD models. Even if initially they weren't as trustworthy as other multi-stage OD models, their speed allows them to work in real-time even with more restricted hardware. YOLO has been developing different updated versions and nowadays is both fast and trustworthy thus it is a good option when dealing with OD.

2.3.10 Pixel-Level Segmentation

Once OR and OD problems are both solved, there is still a natural step. Even if OD returns information about what and where are the objects, the information obtained is still quite imprecise about the space occupied by the object. The bounding box delimits a square within which the object is located, but the vast majority of objects are not square-shaped.

Therefore, the next step is to obtain pixel-level information on classification. In this respect, two problems arise whose terms are often used as analogues due to the similarity between them: Semantic Segmentation and Instance Segmentation.

2.3.10.1 Semantic Segmentation

Given an image, obtaining the class c for each pixel is called SS (Semantic Segmentation) so the output of any SS model would be c masks with the pixels belonging to that class in the image. Figure 2.18 shows an example of SS with three classes.

Figure 2.15: Semantic segmentation example from Deeplab (Chen et al. (2018)). As can be observed, the three people share the same mask as they are in the same class.

It is a good opportunity to show the U-net architecture (Ronneberger et al. (2015)) that we can see in figure 2.16. This structure was first designed to perform SS for medical purposes, but later has been used in many other different problems as Diffusions Models (Ho et al. (2020)). U-net is a good example of how this models usually works by leveraging an encoder-decoder structure to accumulate the salient features of each pixel region during encoding and then in successive decoding steps combine those features with those of the previous encoding steps to finally decide for each pixel which class it belongs to without requiring dense layers. Another good

example of SS is SegNet (Badrinarayanan et al. (2017)), which works similarly to U-Net but reduces the number of parameters needed in the model as it does not need to learn how to upsample.

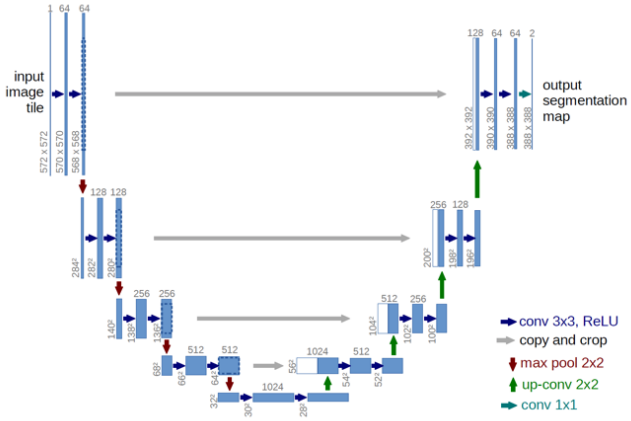


Figure 2.16: U-net architecture from Ronneberger et al. (2015).

2.3.10.2 Instance Segmentation

IS (Instance Segmentation) could be called the combination between OD and SS since it implies obtaining pixel-level masks but also detecting different instances for each class. Figure 2.17 includes some examples. Maybe the most important example is Mask R-CNN (He et al. (2017)), an evolution from the Multi-stage OD Faster R-CNN (Ren et al. (2017)) model.

(a) Example of different trucks with well-defined different masks.

(b) Example of different persons with well-defined different masks.

Figure 2.17: Examples of instance segmented images from He et al. (2017).

The Mask R-CNN model works as a brute-force approach to the problem. Its strategy is to first detect the objects and then apply a network similar to U-net and SegNet to obtain a segmentation mask for each of them.

Figure 2.19 shows a clear comparison between different CV problems.

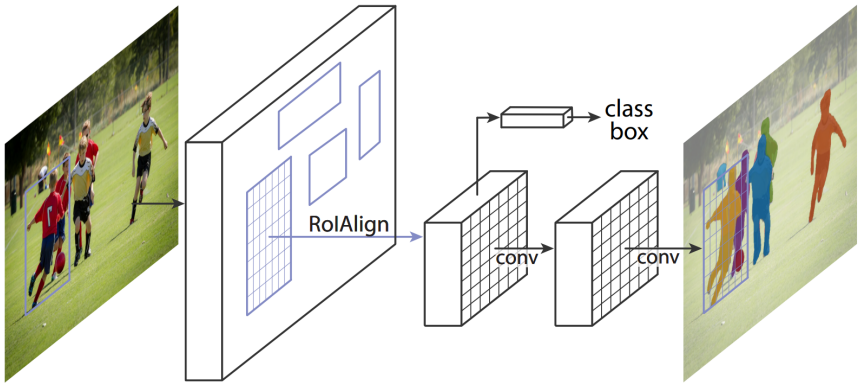


Figure 2.18: Mask R-CNN framework for instance segmentation from He et al. (2017).

2.3.11 Deep Learning and Graphic Processing Unit Acceleration

AlexNet depth was a key element in his high performance according to the authors (Krizhevsky et al. (2017)), allowing a higher amount of features and more levels of abstraction. The main problem with this approach was its high computational cost so in order to be trained, code was written with CUDA Toolkit[5] to run on GPU during training.

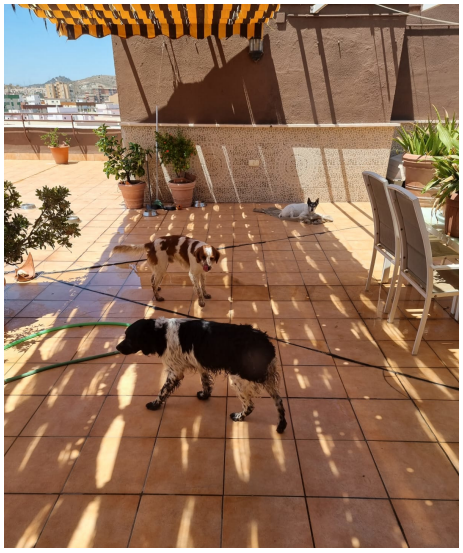
Although the technical support used to implement the code may seem trivial, it is not at all. The basic theoretical elements necessary for the DL, as we have mentioned previously, have been around since almost the 1980s and large datasets were already available in 2005, yet the paradigm did not start to work until the following decade. This time lag was due to a lack of hardware powerful enough to train the ANN in a reasonable amount of time with so much data (Dally et al. (2021)).

The first so-called GPU hardware dates back to 1999 with the NVIDIA Geforce 256 chip. Since its creation, the floating-point computing capabilities of its successors quickly made scientists want to perform their calculations on that specific hardware, even if they had to transform those calculations into vector and shader operations for the hardware to run them. Over time, GPU would become more accessible with the availability of more general-purpose programming languages, Moore's law would ensure that their power would increase exponentially and today it is easy to obtain a GPU, which democratises access to DL to a certain extent.

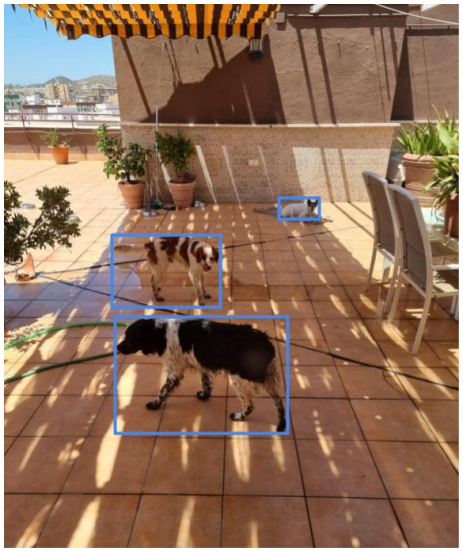
AlexNet was trained on a pair of GeForce 580s, which were high-end hardware but relatively common among gaming enthusiasts and affordable for individuals and research labs. It took two weeks to train. In 2010, Google Brain engineers trained one to find cats on the Internet using 16,000 CPUs. The experiment was repeated in 2012 using only 48 GPU. Today it is a trivial exercise.

From AlexNet to the present day, the complexity of computing models has been closely related to the computational power of the GPU. Figure 2.20 shows the computational power required by some of the famous computing models of the last

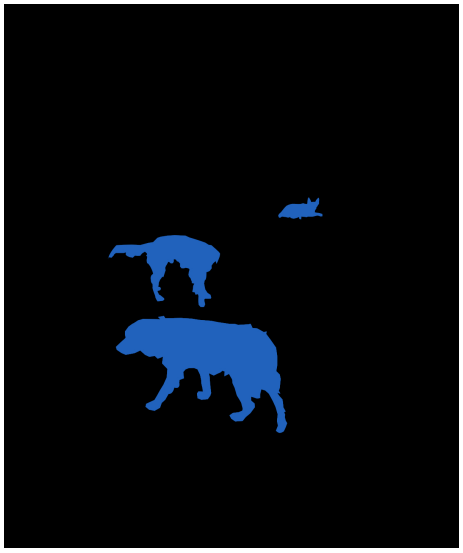
[5]https://docs.nvidia.com/cuda/



(a) Original Image.



(b) Object Detection with three dogs detected each one with a BBOX in the same colour since they belong to the same class.



(c) Object Detection with three dogs detected each one with a pixel-level mask in the same colour since they belong to the same class. A black mask as background since it semantically belongs to another class.

(d) Instance Segmentation with three dogs each one with one pixel-level mask in different colours since even if they belong to the same class, they are not the same instance.

Figure 2.19: Comparison between different CV problems.

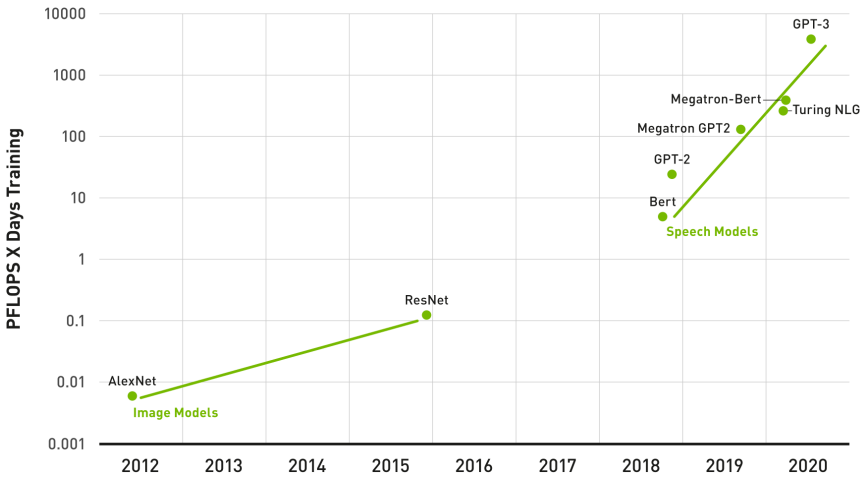


Figure 2.20: Exponential scaling of computational power required to train DL models from Dally et al. (2021).

few years. From the BERT model in 2018 (Devlin et al. (2019)) to the GPT-3 model in 2020 (Brown et al. (2020)), the computational need has become a thousand times greater in a span of only two years. In the figure 2.21 we can observe how the power of processors has evolved exponentially as well.

This leads to a simple conclusion. Much of the progress made in DL is due to sheer brute force. Simple increases in power mean that what was not possible a few years ago is now possible, and so to ignore the link between hardware evolution and the evolution of the field's capabilities is to omit a fundamental element.

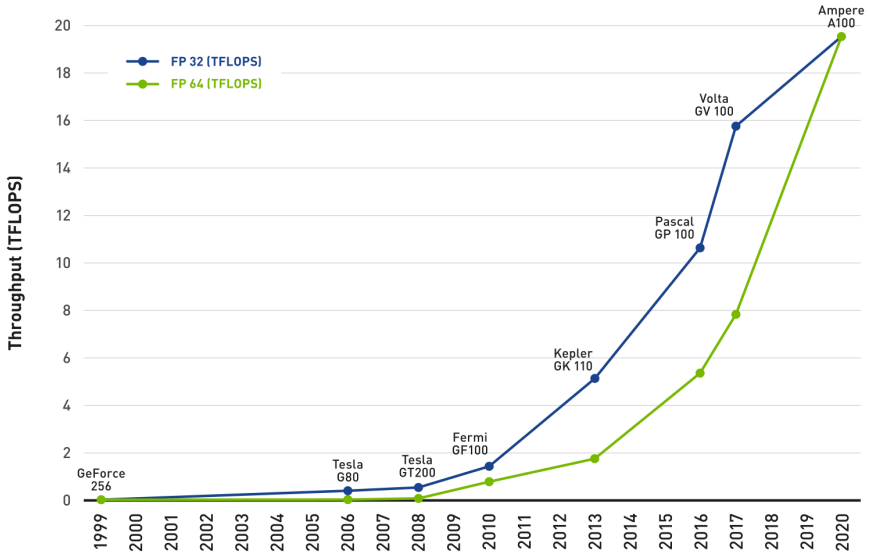


Figure 2.21: Exponential increase on GPU computational power since 2012 from Dally et al. (2021).

Chapter 3

The effect of downsampling-upsampling strategy on foreground detection algorithms

Don't do what you can't undo, until you've considered what you can't do once you've done it.

Assasin's Apprentice, Robin Hobb.

Abstract: This first chapter of the research work presented in this PhD Thesis includes a work published in the journal AIRe (Artificial Intelligence Review) in the year 2020. The work describes how many modern FS methods applied to surveillance systems with higher resolutions (from 640x480 to 1920x1080) could not be performed in real-time without high computational power, so a methodology to reduce computational requirements of FS methods is proposed. The approach is based on reducing the input frame by downsampling the original image before applying the FS method to later use an upsampling strategy to interpolate FS mask to recover the original frame size. Several FS and downsampling methods were tested in order to measure how the proposal affects segmentation quality and time. Experiments show how most tested FS methods maintain segmentation quality with resizing factors up to 0.5 while decreasing drastically overall execution time.

Título	The effect of downsampling-upsampling strategy on foreground detection algorithms
Autores	Miguel A. Molina-Cabello, Jorge García-González, Rafael M. Luque-Baena, Ezequiel López-Rubio
Revista	Artificial Intelligence Review
Año	2020
Factor de impacto	7.563
Categorías JCR	COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE (14/139) (Q1))
Estado	Publicado
DOI	https://doi.org/10.1007/s10462-020-09811-y
Referencia	Molina-Cabello et al. 2020

Chapter 4

Background subtraction by probabilistic modeling of patch features learned by deep autoencoders

We live not to forget our past, but to learn from it!

Final Fantasy IX.

Abstract: Previously to the beginning of this PhD Thesis, some works were published following the strategy stated in Zhang et al. (2015) to approach FS problem on video sequences with high noise levels. The key idea was to split images into patches to later use a methodology based on SDA to classify each patch as foreground or background. These initial patch-wise FS methods were García-González et al. (2018) and García-González et al. (2019b) and state a problem of the patch-wise FS approach: the size of the patches must be big enough to provide robustness to noise, but that implies a segmentation with resolution far lower from the pixel-level strategies. In order to face this problem, García-González et al. (2019a) was presented in the 8th IWINAC (International Work-Conference on the Interplay Between Natural and Artificial Computation) held in Almería, Spain in 2019. The idea was to use a patch overlapping strategy to increase resolution while maintaining the robustness to noise.

That work using overlapping was refined and extended to be published in ICAE (Integrated Computer-Aided Engineering) during 2020 with a more formal approach to the classification of each patch as foreground or background as well as with wider experiments to test the approach with a multitude of scenes and a great diversity of added noises. This chapter includes such an extension article. The general conclusion is that the approach is valid for a wide variety of noises. However,

when dealing with noise-free images more classical approaches such as St-Charles et al. (2015) or St-Charles et al. (2016) are often preferable due to better overall performance and lower computational cost.

The disadvantage of the computational cost of overlapping by requiring to process many more patches (of the order of at least 4 times more) will be partially solved later in chapter 10.

Título	Background subtraction by probabilistic modeling of patch features learned by deep autoencoders
Autores	Jorge García-Gonzáleza, Juan M. Ortiz-de-Lazcano-Lobato, Rafael M. Luque-Baena, Ezequiel López-Rubio
Revista	Integrated Computer-Aided Engineering
Año	2020
Factor de impacto	3.673
Categorías JCR	COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE (37/139) (Q2)) COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS (28/111) (Q2)) ENGINEERING, MULTIDISCIPLINARY (13/90) (Q1))
Estado	Publicado
DOI	https://doi.org/10.3233/ICA-200621
Referencia	García-González et al. 2020a

Chapter 5

Foreground detection by probabilistic mixture models using semantic information from deep networks

To succeed, planning alone is insufficient. One must improvise as well.

Foundation, Isaac Asimov.

Abstract: This chapter introduces our work presented at the 24th ECAI (European Conference on Artificial Intelligence) in 2020. In this work, we propose a FS method for static video surveillance cameras based on semantic information obtained from a SS method such as Mask R-CNN. The key idea is to combine semantic classifications for each pixel with the pixel colour using a probabilistic model and compare it with a background model to decide whether the pixel belongs to the foreground or not.

Since this proposal is a meta-algorithm not dependent on the used SS method and some of their errors are due to an erratic semantic classification, the proposal application should be easily upgradable by using new less error-prone SS methods without changes to our algorithm.

Título	Foreground detection by probabilistic mixture models using semantic information from deep networks
Autores	Jorge García-González, Juan M. Ortiz-de-Lazcano-Lobato, Rafael M. Luque-Baena, Ezequiel López-Rubio
Congreso	24th European Conference on Artificial Intelligence (ECAI)
Año	2020
GGS Rating	A-
CORE Rating	A
Estado	Publicado
DOI	https://doi.org/10.3233/FAIA200408
Referencia	García-González et al. 2020b

Chapter 6

Deep autoencoder architectures for foreground object detection in video sequences based on probabilistic mixture models

-The first lesson on Roke, and the last is, Do what is needful! And no more.
-The lessons in between, then, must consist in learning what is needful.
-They do.

The Farthest Shore, Ursula K. Le Guin.

Abstract:

This chapter presents a paper presented at the ICIP (IEEE International Conference on Image Processing) in 2020. In this case, it consists of a study of the application of autoencoder neural networks to the SPP problem.

In this work, we study how the architecture, pre-training, the type of AE (Autoencoder) and the activation function affect patch-based FS methods such as the one in chapter 4. In the experiments six different architectures are proposed, two of them based exclusively on dense layers, two of them composed exclusively of convolutional layers and two of them composed of a mixture of both. Two of each type are proposed to take into account different depths, thus changing the number of layers. As for the type of AE, the use of classical SDA (Stacked Denoising Autoencoder) and the use of VAE (Variational Autoencoder) are proposed. The latter are often used as generative models and in theory, offer a better regularised latent space. Regarding pre-training, pre-training with and without added Gaussian noise is tested to improve the model's ability to clean the noise from the patches. Two different activation functions are also tested for the innermost layer:

Hyperbolic Tangent and Sigmoid. A total of 48 ANN are trained.

The experiments are performed using three types of noise: Gaussian, Uniform and Salt and Pepper. In order to evaluate them, special attention has been paid to the False Positive and False Negative ratios. In the tests, a generalised resilience to noise is observed, which is increased with appropriate choices. A Pareto front is observed in the ratio of False Positives and False Negatives related to AE type.

Título	Deep Autoencoder Architectures For Foreground Object Detection In Video Sequences Based On Probabilistic Mixture Models
Autores	Jorge García-González, Miguel A. Molina-Cabello, Rafael M. Luque-Baena, Juan M. Ortiz-de-Lazcano-Lobato, Ezequiel López-Rubio
Congreso	2020 IEEE International Conference on Image Processing (ICIP)
Año	2020
GGS Rating	A-
CORE Rating	B
Estado	Publicado
DOI	https://doi.org/10.1109/icip40778.2020.9190834
Referencia	Garcia-Gonzalez et al. 2020

Chapter 7

Foreground Segmentation Improvement by Image Denoising Preprocessing Applied to Noisy Video Sequences

Have you ever considered that too many answers are the same as no answer at all?

A Clash of Kings, George R. R. Martin.

Abstract: This work was presented in 16th SOCO (International Conference on Soft Computing Models in Industrial and Environmental Applications) in 2021 celebrated in Bilbao, Spain. This work includes an approach to FS problem different from other proposals in this PhD Thesis and focused on image preprocessing in order to simplify the FS. Instead of proposing a new and complex FS algorithm robust to noise, this proposal aims to use DA to preprocess each image from the video sequence before being processed by a fast classic FS algorithm. Experiments with several combinations of DA architectures and FS algorithms are provided and tested against classical denoising methods. Results confirm the approach is effective when dealing with a high amount of noise with the right combination of DA and FS algorithm.

Título	Foreground Segmentation Improvement by Image Denoising Preprocessing Applied to Noisy Video Sequences
Autores	Jorge García-González, Juan Miguel Ortiz-de-Lazcano-Lobato, Rafael Marcos Luque-Baena, Ezequiel López-Rubio
Congreso	16th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO)
Año	2020
GGS Rating	Work in Progress
CORE Rating	-
Estado	Publicado
DOI	https://doi.org/10.1007/978-3-030-87869-6_37
Referencia	García-González et al. 2021a

Chapter 8

Road pollution estimation from vehicle tracking in surveillance videos by deep convolutional neural networks

There is never a second opportunity to make a first impression.
Sword of Destiny, Andrzej Sapkowski.

Abstract: This work was published in ASoC (Applied Soft Computing) journal in 2021. This research branch diverges from the main FS research presented in this PhD so far as will be following chapter 9. This time the work is focused on vehicle behaviour analysis. The aim is to estimate the velocity for each car and each instant using road video surveillance cameras in order to approximate pollution. The following strategy is to apply Homography to correct the traffic camera's image perspective. The Homography is obtained by using a satellite image from the same place obtained from any map application (e.g. Google Maps). CNN-based OD method is required to get vehicles' positions and perform their tracking. The paper includes a methodology to obtain per-frame speed and pollution from vehicles' paths once the camera image perspective has been corrected.

Although we consider the paper theoretically sound, the experimental section is very limited due to the lack of proper data to perform the study since no datasets with pollution emissions per frame were found. This is a pity since approaches like this one could be a great ally to knowing real-time vehicle-generated pollution.

Título	Road pollution estimation from vehicle tracking in surveillance videos by deep convolutional neural networks
Autores	Jorge García-González, Miguel A. Molina-Cabello, Rafael M. Luque-Baena, Juan M. Ortiz-de-Lazcano-Lobato, Ezequiel López-Rubio
Revista	Applied Soft Computing
Año	2021
Factor de impacto	7.388
Categorías JCR	COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE (23/145) (Q1)) COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS (11/111) (Q1))
Estado	Publicado
DOI	https://doi.org/10.1016/j.asoc.2021.107950
Referencia	García-González et al. 2021b

Chapter 9

Vehicle overtaking hazard detection over onboard cameras using deep convolutional networks

The purpose of a storyteller is not to tell you how to think, but to give you questions to think upon.

The Way of Kings, Brandon Sanderson.

Abstract:

As the work from chapter 7, this work was presented in SOCO, but the 17th edition in 2022 was held in Salamanca, Spain. This work follows the research branch focused on vehicle behaviour surveillance. This time we worked on the detection of dangerous overtaking from an onboard camera using only vehicle detections from a CNN-based OD model. The key is to use the centre of the BBOX to track the vehicles and BBOX changes in size to estimate the relative speed using SAA (Small-Angle Approximation) and RANSAC (Random Sample Consensus) algorithm. Even with a quite simple experimentation setup, results show a promising approach to estimating relative speed which only relies on video camera images and no other hardware.

Título	Vehicle Overtaking Hazard Detection over Onboard Cameras Using Deep Convolutional Networks
Autores	Jorge García-González, Iván García-Aguilar, Daniel Medina, Rafael Marcos Luque-Baena, Ezequiel López-Rubio, Enrique Domínguez
Congreso	17th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO)
Año	2022
GGS Rating	Work in Progress
CORE Rating	-
Estado	Publicado
DOI	https://doi.org/10.1007/978-3-031-18050-7_32
Referencia	García-González et al. 2021a

Chapter 10

Moving Object Detection in Noisy Video Squences using Deep Convolutional Disentangled Representations

Oh, but history moved in such vicious circles.

The Burning God, R. F. Kuang.

Abstract: The last research work chapter in this PhD Thesis was presented at the 29th ICIP in 2022, held in Bordeaux, France. This work follows directly works from chapters 4 and 6 on the proposal of a FS method based on DA and probabilistic model to detect genuine movement on noisy video sequences. The main difference is in this proposal DA includes only CNN layers in order to obtain an intermediated disentangled representation of the image. That representation allows us to obtain almost pixel-level foreground classification, dealing with the patch-level segmentation limitation presented in the previous chapters. As with chapter 6, a study with several training strategies and DA depths was done to test the proposal under very different training assumptions. Following the methodology, higher resolution FS masks were obtained and, as expected, results show our proposal is robust to a wide variety of noises, both Gaussian and uniform. Nevertheless, it was unexpected to obtain better results with non-ideal training assumptions than with ideal ones.

Título	Moving Object Detection in Noisy Video Sequences Using Deep Convolutional Disentangled Representations
Autores	Jorge García-González, Rafael M. Luque-Baena, Juan M. Ortiz-de-Lazcano-Lobato, Ezequiel López-Rubio
Congreso	2022 IEEE International Conference on Image Processing (ICIP)
Año	2022
GGS Rating	A-
CORE Rating	B
Estado	Publicado
DOI	https://doi.org/10.1109/ICIP46576.2022.9897305
Referencia	Garcia-Gonzalez et al. 2022

Chapter 11

Conclusions and Future Research Lines

But then it's difficult, isn't it, to make a passionate argument for what you already have? So boring. Whereas the delightful alternative? A bouquet of promises! A sackful of dreams! A glorious ship of fantasies, undamaged by collision with actually getting anything done.

The Problem with Peace, Joe Abercrombie.

Abstract:

In conclusion, the results obtained after the years of research dedicated to this PhD are presented, as well as some of the approaches for future work.

11.1 Conclusions

The main objective of this PhD thesis has been the Foreground Segmentation problem, a classic problem of Computer Vision that has been dealt with during the last decades and that, consequently, had already been worked on from many approaches with very good results before this research began. Therefore, it is normal that it is particularly difficult to find a niche in such an elaborate problem where we can make truly significant contributions. Our approach based on combining image processing by pre-trained networks or networks with unsupervised training with probabilistic models provided that small gap where most of the contributions could be made. Along the way, as is normal when one is learning about different problems and their solutions, ideas for applying networks to traffic problems emerged and this constitutes the secondary leg of this PhD thesis.

In support of this research, eight works have been included in the thesis, of which the author of the thesis is the first author of all but the first. Of these eight papers, three have been published in high-impact journals, three have been published in well-ranked international conferences and two smaller works testing tentative ideas have been published in lower-ranked conferences.

It is important to note that most of the works are closely related to each other, thus, the work published in AIRe (chapter 3) is an approach to the FS problem focusing on improving the result of other methods by specific image pre-processing. This approach is used again in the work published in the 2021 SOCO conference (chapter 7). In the first case, preprocessing was focused on altering the image size to study mainly how it affects performance in terms of speed with very positive conclusions indicating that with the right adjustment, a noticeable speed improvement could be achieved without losing segmentation quality. Meanwhile, in the second work, classical filters and filtering based on autoencoders networks are used to clean the images from noise and thus achieve results robust to this kind of image alterations with varied results that allude to the need to properly combine the type of preliminary filtering with the FS method that will be used afterwards.

The work published in ICAE (chapter 4) is also based on using autoencoder networks. In this case, the networks serve to transform image patches into data vectors that should contain the main information (but not the noise) of the patches as if they were a compressor. This data is then evaluated using a probabilistic model to determine whether or not that image patch belongs to the foreground. Because of the resolution constraints posed by a patch-level approach, a patch overlap strategy is used in this work to increase resolution at the cost of increased computational requirements. Extensive experiments show that a system capable of segmenting the foreground with significant robustness to image noise is achieved. In the work published in the ICIP 2020 conference (chapter 6), a study is made about how the type of autoencoder network and its training affects this kind of method. In this work, variants of classical autoencoder networks such as VAE networks are used, observing that the latter are especially effective in contexts in which it is desired to reduce False Negatives while increasing False Positives. This line of work culminates in the publication in the ICIP 2022 conference (chapter 10) with a method that uses the same concepts of autoencoders networks as a preliminary step to probabilistic models but eliminates the need to divide the image into patches by working on the depth of the convolution filters. In this way, noise resilience is achieved with virtually no loss of segmentation resolution and no need for an overlapping strategy that substantially increases the computational burden.

In parallel, but closely related, a meta-method that analyses with a probabilistic model the pixel-level segmentation masks of an IS (Instance Segmentation) model (He et al. (2017)) was proposed at the 2020 ECAI conference (chapter 5). The results, in this case, were very positive since by grounding the system on a model for such a general task as IS, the proposed meta-method will improve as the problems of such a model are solved or a similar one with better results appears.

All these works are focused on objective 1 described in section 1.2 at the beginning of this thesis. In them, mainly approaches to objective 1 are made using patches and autoencoders since they proved to be a promising way to deal with noise in FS. However, two studies reusing classical algorithms and one work based on leveraging pixel-level segmentation algorithms are also included. Overall, although the problem is far from being solved, we have made significant contributions

to the identification of foreground anomalies and, in particular, to doing so in noisy situations.

The second aspect of this thesis is the two methods proposed to perform traffic analysis based on applying Deep Learning to images. In this case, the two methods are based on estimating the speed of vehicles from completely different approaches. In the work published in ASoC in 2021 (chapter 8), a method is proposed to estimate the speed and from it, the pollution generated by vehicles using a homographic transformation and the images from a static traffic camera. In the work published in the 2022 SOCO (chapter 9), the perspective of the camera is changed and we work with onboard cameras and, based on the change in the BBOX size of each vehicle, the relative speed of the other drivers are estimated in order to try to warn of dangerous overtaking. These two works have involved tackling different problems but working with similar tools and paradigms. Both have been challenging because the problems they attempt to solve have rarely been tackled using only images and without relying on specific hardware, so the means of assessing the reliability of the models are still poor and underdeveloped, but for that reason too they are particularly interesting for the future.

These two papers focus on objective 2 described in section 1.2 at the beginning of this thesis. In the first, although an anomaly analysis is not carried out, work is done in the previous step of inferring speeds only from static cameras. In the second, we work with moving cameras and the application of speed estimation is directly to identify anomalous overtaking. Both articles, rather than achieving objective 2, demonstrate that this is a promising line of work.

11.2 Future Work

The completion of a doctoral thesis must have an end, although in principle it should only be the beginning of the doctoral student's research career. It is natural, therefore, that it should lead to conclusions and lines of work in which to continue what has been done, either by refining the proposals made, making other proposals for the same problem or tackling different problems with the approaches studied during the research period.

As a conclusion to the thesis, it is important to add a general one on the problem of Foreground Segmentation for the future. Although in the past it was an extremely relevant task as a previous step to solve other Computer Vision problems, Deep Learning has made this previous step unnecessary in many cases. Since the problem is so well worked and there are so many different proposals, it is not particularly attractive to continue generically working on Foreground Segmentation if it also loses relevance.

Based on the work carried out, the future research resulting from this thesis is divided into four main lines of research:

- To continue working on the proposed foreground detection models to try to improve their accuracy in detecting motion.
- To work on the acceleration of the foreground segmentation line based on the use of AE.
- To extend anomaly analysis to trajectory anomalies in video sequences using a similar approach to that used in the thesis work.

- To deepen image-based velocity analysis without the use of specific hardware.

11.2.1 Improving Foreground Segmentation

The most obvious line of work in the short term is to continue iterating on the models that have been proposed throughout the thesis in order to refine their shortcomings in terms of foreground segmentation quality. In that sense, there are three points where improvements could be made:

- Improve the background update criteria by using Continuous Learning techniques to identify when the background models we are handling have moved too far away from the actual background distribution before they become completely corrupted and they are performing poorly.
- Replace the Probabilistic Model criterion with some Supervised Learning technique such as a MLP that classifies the latent space representations resulting from the AE networks.
- Incorporate Data Augmentation techniques when training the AE networks so that they learn a more varied probability distribution in accordance with reality.

11.2.2 Acceleration

As previously mentioned, the role of Foreground Segmentation in the typical working pipeline of a Computer Vision system has been displaced among other things by the fast improvement of Object Detection systems. Since Deep Learning based Object Detection systems directly apply automatically learned feature detectors to the whole image, it does not make much sense to detect in advance in which area of the image there are objects of interest based on motion. On the other hand, if you want to detect motion, you can combine an object detector with a tracking system as was done in the works in chapters 8 and 9 to do so.

When might it then make sense to apply a motion detection pre-process? When moving objects are of interest and there is a substantial time difference. Both Object Detection systems and pixel-level segmentation systems are still deficient when the number of objects in the image is very high and their size very small. In this sense, there are proposals for detecting (García-Aguilar et al. (2023a), García-Aguilar et al. (2022b)) and segmenting (García-Aguilar et al. (2023b), García-Aguilar et al. (2022a)) objects that are focused on improving the detection of small objects by using Super-Resolution. It might make sense, for example, to make a similar system that instead of relying on a detection pass to know which region to study in more detail, detects the areas where to apply Super-Resolution based on the use of Foreground Segmentation.

This can have two advantages:

- Increase speed because Foreground Segmentation is faster than applying the Object Recognition model.
- Increase reliability because Foreground Segmentation does not miss small objects that the detection model does miss.

Therefore, for the use of Foreground Segmentation to make sense, it must more than ever be fast and accurate. In that sense, a possible line of future work would be to focus efforts on optimising the work presented in chapter 10).

11.2.3 Anomalous Trajectory Detection

Although the research work in this thesis has focused on the identification of foreground anomalies based on motion, there are a large number of analysable anomalies in video sequences. Analogous to the use we have made of the SDA encoder in our work while ignoring the decoder, we consider the possible use of GAN networks to detect anomalies in motion trajectories using not the generator network, but the discriminator that is trained with it.

The approach is as follows:

1. Starting from a sequence with usual behaviour, obtain some visual representation of the flow of the movement such as Optical Flow.
2. Train a network with these images. This will result in a generator network to create images with similar Optical Flow and a discriminator network to distinguish which images have similar Optical Flow.
3. Use the discriminator network on other Optical Flow images to distinguish anomalous ones.

11.2.4 Speed Analysis

Two of the works presented in this doctoral thesis are based on the analysis of velocities and distances based solely on images. The analysis of velocities is a problem that can be trivial when specific hardware such as LiDAR or RADAR is available. These are specifically created to obtain the distance to other objects, but leaving aside the intrinsic problems with the way each one works, there is a huge disadvantage to cameras: there are many, many more cameras and they serve many more purposes.

It is much more plausible to have the presence of a camera than the presence of other hardware, and that makes it particularly interesting to pursue a line of work in which velocities can be obtained using one. This is a line of work that needs to be solidified, and for this, we believe that the creation of datasets with the necessary information to evaluate the proposed methods properly is crucial. Therefore, the next step in this line of work would be to create a dataset to use and share with the scientific community that associates in a simple way the speed and relative distances of two vehicles.

Appendix A

Resumen de publicaciones obtenidas

Resumen: En esta sección se muestran unas tablas que resumen la información asociada a las publicaciones obtenidas con los trabajos presentados en esta tesis. Para las revistas se ha utilizado el ranking *Journal Citation Reports (JCR)* del año correspondiente a la publicación, mientras que para los congresos se indica el ranking CORE [1] del año correspondiente (o el último en el caso de los publicados en 2022 ya que el ranking de ese año no ha sido publicado) y *GII-GRIN-SCIE (GGS) Conference Rating*[2] del año 2021. El factor de impacto se indica sin autocitas.

Título	The effect of downsampling-upsampling strategy on foreground detection algorithms
Autores	Miguel A. Molina-Cabello, Jorge García-González, Rafael M. Luque-Baena, Ezequiel López-Rubio
Revista	Artificial Intelligence Review
Año	2020
Factor de impacto	7.563
Categorías JCR	COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE (14/139) (Q1))
Estado	Publicado
DOI	`https://doi.org/10.1007/s10462-020-09811-y`
Referencia	Molina-Cabello et al. 2020

[1] `https://www.core.edu.au/`
[2] `https://scie.lcc.uma.es:8443/`

Título	Background subtraction by probabilistic modeling of patch features learned by deep autoencoders
Autores	Jorge García-Gonzáleza, Juan M. Ortiz-de-Lazcano-Lobato, Rafael M. Luque-Baena, Ezequiel López-Rubio
Revista	Integrated Computer-Aided Engineering
Año	2020
Factor de impacto	3.673
Categorías JCR	COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE (37/139) (Q2)) COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS (28/111) (Q2)) ENGINEERING, MULTIDISCIPLINARY (13/90) (Q1))
Estado	Publicado
DOI	https://doi.org/10.3233/ICA-200621
Referencia	García-González et al. 2020a

Título	Foreground detection by probabilistic mixture models using semantic information from deep networks
Autores	Jorge García-González, Juan M. Ortiz-de-Lazcano-Lobato, Rafael M. Luque-Baena, Ezequiel López-Rubio
Congreso	24th European Conference on Artificial Intelligence (ECAI)
Año	2020
GGS Rating	A-
CORE Rating	A
Estado	Publicado
DOI	https://doi.org/10.3233/FAIA200408
Referencia	García-González et al. 2020b

Título	Deep Autoencoder Architectures For Foreground Object Detection In Video Sequences Based On Probabilistic Mixture Models
Autores	Jorge García-González, Miguel A. Molina-Cabello, Rafael M. Luque-Baena, Juan M. Ortiz-de-Lazcano-Lobato, Ezequiel López-Rubio
Congreso	2020 IEEE International Conference on Image Processing (ICIP)
Año	2020
GGS Rating	A-
CORE Rating	B
Estado	Publicado
DOI	https://doi.org/10.1109/icip40778.2020.9190834
Referencia	Garcia-Gonzalez et al. 2020

Título	Foreground Segmentation Improvement by Image Denoising Preprocessing Applied to Noisy Video Sequences
Autores	Jorge García-Gozález, Juan Miguel Ortiz-de-Lazcano-Lobato, Rafael Marcos Luque-Baena, Ezequiel López-Rubio
Congreso	16th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO)
Año	2020
GGS Rating	Work in Progress
CORE Rating	-
Estado	Publicado
DOI	https://doi.org/10.1007/978-3-030-87869-6_37
Referencia	García-González et al. 2021a

Título	Road pollution estimation from vehicle tracking in surveillance videos by deep convolutional neural networks
Autores	Jorge García-González, Miguel A. Molina-Cabello, Rafael M. Luque-Baena, Juan M. Ortiz-de-Lazcano-Lobato, Ezequiel López-Rubio
Revista	Applied Soft Computing
Año	2021
Factor de impacto	7.388
Categorías JCR	COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE (23/145) (Q1)) COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS (11/111) (Q1))
Estado	Publicado
DOI	https://doi.org/10.1016/j.asoc.2021.107950
Referencia	García-González et al. 2021b

Título	Vehicle Overtaking Hazard Detection over Onboard Cameras Using Deep Convolutional Networks
Autores	Jorge García-González, Iván García-Aguilar, Daniel Medina, Rafael Marcos Luque-Baena, Ezequiel López-Rubio, Enrique Domínguez
Congreso	17th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO)
Año	2022
GGS Rating	Work in Progress
CORE Rating	-
Estado	Publicado
DOI	https://doi.org/10.1007/978-3-031-18050-7_32
Referencia	García-González et al. 2021a

Título	Moving Object Detection in Noisy Video Sequences Using Deep Convolutional Disentangled Representations
Autores	Jorge García-González, Rafael M. Luque-Baena, Juan M. Ortiz-de-Lazcano-Lobato, Ezequiel López-Rubio
Congreso	2022 IEEE International Conference on Image Processing (ICIP)
Año	2022
GGS Rating	A-
CORE Rating	B
Estado	Publicado
DOI	https://doi.org/10.1109/ICIP46576.2022.9897305
Referencia	Garcia-Gonzalez et al. 2022

Appendix B

Resumen en Español

Los avances en Inteligencia Artificial y Visión por Computador de la última década convenientemente han coincidido con el desbordante incremento en generación de contenido multimedia por parte de particulares y empresas. Esta masiva cantidad de datos representa una gran oportunidad, pero también un gran desafío para su análisis. La opción de hacerlo por medios humanos tradicionales es inviable, así que se espera de los ordenadores que realicen todo o parte de la tarea del procesado de una cantidad de videos que no hace más que crecer. En ese contexto, se espera que herramientas de Visión por Computador basadas en Redes Neuronales Artificiales y específicamente Aprendizaje Profundo sean la respuesta a dicha necesidad y para ello científicos e ingenieros tienen que superar sus limitaciones actuales.

El principal problema abordado en esta tesis doctoral es la detección de anomalías de primer plano en secuencias de video genéricas mediante el uso de técnicas de Aprendizaje Profundo especialmente enfocadas a ser robustas al ruido en las imágenes. Como problema derivado, se trata el análisis de secuencias de tráfico utilizando las mismas herramientas.

Como testigo de dichos objetivos se incluyen en este documento una serie de trabajos de investigación realizados a lo largo de cuatro años. El primero de ellos publicado en la revista Artificial Intelligence Review en 2020 sobre la aceleración de métodos de Segmentación de Primer Plano mediante el preprocesamiento de las imágenes para alterar su tamaño minimizando la pérdida de calidad en la segmentación final. El segundo trabajo fue publicado en la revista Integrated Computer-Aided Engineering en 2020 y versa sobre la aplicación de redes auto-codificadoras a parches solapados de la imagen para obtener una codificación que luego es analizada mediante un modelo probabilístico. El tercer trabajo también fue presentado en 2020 en el congreso European Conference on Artificial Intelligence y propone un método basado en segmentar semánticamente a nivel de píxel los objetos de las imágenes para analizar su movimiento y así identificar el primer plano. El cuarto trabajo fue presentado en el congreso IEEE International Conference on Image Processing de 2020 y realiza un estudio sobre cómo el tipo y entrenamiento de las redes autocodificadoras afectan a los modelos de Segmentación de Primer Plano basados en parches. El quinto trabajo, presentado en el congreso International Conference on Soft Computing Models in Industrial and Environmental Applications en 2021, estudia el uso de filtros clásicos y redes auto-codificadoras para limpiar las imágenes como paso previo a utilizar métodos clásicos de Segmentación de Primer Plano y

así incrementar la resistencia del sistema al ruido. El sexto propone un método de análisis de velocidades de vehículos para estimar la polución generada por estos basándose únicamente en imágenes conseguidas mediante cámaras de tráfico y fue publicado en la revista Applied Soft Computing en 2021. El séptimo trabajo se presentó también en el congreso International Conference on Soft Computing Models in Industrial and Environmental Applications de 2021 y propone una estimación de velocidades relativas de otros vehículos a partir de las imágenes grabadas por una cámara incorporada al propio vehículo. Por último, el octavo trabajo que compone esta tesis fue presentado en el congreso IEEE International Conference on Image Processing de 2022 y refina el uso de redes auto-codificadoras para Segmentación de Primer Plano de trabajos anteriores eliminando la necesidad de dividir la imagen en parches.

Con esos ocho trabajos se constituye el memorándum de esta tesis doctoral y el autor presenta los resultados de estos cuatro años de investigación.

B.1 Introducción

Si los dos últimos siglos han visto una continua serie de avances tecnológicos que han cambiado la forma en la que viven los seres humanos, las últimas décadas han visto acelerar el proceso del cambio. El modo de vida de un joven ciudadano europeo de principios del siglo XXI tiene poco en común con como sus padres o abuelos vivieron la misma etapa de su vida. Algunos de estos cambios, como la llegada de Internet, el fácil acceso a la tecnología y la miniaturización de esta, han tenido un efecto transformador radical en un punto particular de la vida: mientras que antes uno podía a lo sumo aspirar a ser un consumidor de contenido multimedia distribuido por medio de la televisión, ahora cualquiera con un teléfono móvil u ordenador puede ser un generador de dicho contenido.

La existencia de plataformas para publicar y visualizar videos como Youtube o Twitch se ha mezclado con las redes sociales. No solo se publica contenido multimedia en Facebook, Twitter e Instagram, sino que existen también redes sociales específicas como TikTok en las que la forma fundamental de comunicación es el intercambio de contenido multimedia. En España, hay gente que recuerda el comienzo de la emisión televisiva a mediados del siglo pasado y que ahora vive en una sociedad donde cada ciudadano puede ofrecer su propia variedad de contenido en media docena de plataformas diferentes con un dispositivo que cabe en la palma de una mano. Ahora vivimos en un mundo multimedia y, salvo apocalipsis con su consecuente caída de la civilización como la conocemos, eso no va a cambiar.

Además de los datos generados por los ciudadanos particulares, las administraciones públicas y las empresas también disponen de dispositivos de grabación y almacenamiento de datos multimedia para sus propios propósitos. Es común en nuestra vida diaria encontrar una cámara de tráfico del ayuntamiento o cámaras de video de alguna compañía de seguridad monitorizando ciertas áreas. Aunque frecuentemente los pasamos por alto, estos dispositivos están generando su propio contenido, no para distribuirlo por razones de entretenimiento o informativas, sino con propósitos más específicos.

La explosión en lo que a generación de contenido multimedia se refiere no ofrece solo grandes oportunidades, sino una multitud de problemas, desafíos y preguntas. ¿Cómo lidiamos con esas cantidades masivas de datos? Si una única cámara está

grabado 24 horas al día, ¿necesitamos tres personas haciendo turnos de 8 horas para ver su contenido permanentemente? Como reza el dicho: lo que acaba en Internet, nunca abandona Internet. ¿Cómo puede una plataforma controlar que sus usuarios no violen sus normas de contenido y la usen maliciosamente? Es su plataforma y son en parte responsables de su uso. ¿Pueden confiar en el reporte de otros usuarios o de sus propios moderadores? Esa estrategia es reactiva, pero no preventiva. No sirve para prevenir la publicación de contenido, solo para bloquear el contenido que ya ha sido publicado.

Hay demasiados datos siendo generados todo el tiempo para usar sistemas puramente humanos para controlarlos: necesitamos herramientas automáticas para analizar y extraer información de ese contenido multimedia. Necesitamos que los ordenadores sean capaces de realizar o al menos simplificar el análisis de todas esas horas de video o la cantidad de datos nos sobrepasará. En algunos casos esos datos son una mera herramienta que, si no se analizan, se desperdiciará, pero en otros casos el análisis de contenido es un requisito si no queremos que las redes multimedia sean caóticas y peligrosas.

En paralelo a esa circunstancia, los desarrollos en Inteligencia Artificial han permitido generar nuevos sistemas de análisis cada vez más precisos y flexibles. Concretamente la Visión por Computador ha desarrollado durante las últimas décadas métodos para analizar imágenes y vídeos con múltiples propósitos y el desarrollo del Aprendizaje Computacional ha permitido en los últimos años que la capacidad de los ordenadores para procesar esa clase de datos se incremente de manera exponencial. Mientras que antes era necesario establecer una serie de filtros manualmente para detectar según qué objetos en una imagen, ahora podemos entrenar modelos que aprendan automáticamente más filtros, más variados y mejor adaptados que los que ningún humano podría diseñar para los objetos que deseemos. Cada día podemos automatizar más tareas de procesamiento de contenido multimedia, pero cuanto más conseguimos, más queremos hacer.

B.2 Estado del Arte

El capítulo 2 está dedicado a los fundamentos teóricos sobre los que se sustentan los trabajos presentados en esta tesis así como el problema de la Segmentación de Primer Plano y otras aproximaciones para resolverlo y los problemas de Reconocimiento, Detección y Segmentación de objetos en imágenes.

Para empezar, la sección 2.1 expone brevemente los fundamentos básicos de la teoría de la probabilidad siguiendo el libro Bishop (2006) desde la notación hasta la distribución Gaussiana pasando por el Teorema de Bayes (Bayes (1763)) y su relevancia en los problemas de Aprendizaje Computacional.

Seguidamente, la sección 2.2.1 presenta el problema de la Detección de Anomalías en su sentido más amplío y general junto con algunos ejemplos, planteamientos posibles y algunas posibles aproximaciones basadas en árboles de Decisión (Liu et al. (2008)), Máquinas de Soporte Vectorial (Amer et al. (2013)), Agrupamiento (Mazarbhuiya y Shenify (2023)) y Redes Neuronales Artificiales (Pang et al. (2022)).

Tras esa presentación general se dedica la sección 2.2.2 al problema que se aborda directamente en la mayoría de los trabajos de esta tesis: la Segmentación de Primer Plano en secuencias de video (también llamada Sustracción de Fondo y *Foreground Segmentation* o *Background Subtraction* en la bibliografía en inglés).

Tras presentar en qué consiste detectar los objetos de primer plano y su relación con el movimiento, se describe un esquema general de solución para dicho problema (Garcia-Garcia et al. (2020)) basado en resolver cuatro cuestiones: cómo modelar el fondo de la secuencia, cómo inicializarlo, cómo utilizar dicho modelo para clasificar los píxeles en fondo o primer plano y cómo actualizar el modelo de fondo. Se incluye también una serie de desafíos que comúnmente hay que enfrentar a la hora de proponer una solución satisfactoria (Molina-Cabello (2018)) así como aplicaciones.

Dado que es el problema principal de la tesis, se describen propuestas previas para solucionarlo dividiéndolas principalmente en cinco grupos: los planteamientos clásicos basados en análisis estadísticos y probabilísticos a nivel de píxel (Wren et al. (1997),Stauffer y Grimson (1999), Elgammal et al. (2000), Zivkovic (2004)); los planteamientos que se basan en Mapas Autoorganizados (Kohonen (1982)) para resolver el problema (Maddalena y Petrosino (2008), Maddalena y Petrosino (2010), Maddalena y Petrosino (2012), López-Rubio et al. (2011)); los planteamientos basados en el análisis de regiones que consiguen obtener segmentaciones a nivel de píxel (St-Charles y Bilodeau (2014), St-Charles et al. (2014), St-Charles et al. (2015), St-Charles et al. (2016), López-Rubio y López-Rubio (2015)); los planteamientos basados en análisis a nivel de regiones que consiguen una segmentación a nivel de regiones (parches o *patches* en ingles) usando redes autocodificadoras para incrementar la resistencia a ruido (Zhang et al. (2015), García-González et al. (2018), García-González et al. (2019a), García-González et al. (2019b)) que es la línea que se sigue principalmente en esta tesis doctoral; y otros planteamientos basados en Aprendizaje Computacional como López-Rubio et al. (2018). Todos los planteamientos mencionados se basan exclusivamente en aprendizaje no supervisado, ya que esta tesis no enfrenta el problema por la vía supervisada.

En la sección 2.3 se muestran los fundamentos de las Redes Neuronales Artificiales y el Aprendizaje Profundo. Se parte del concepto más básico de neurona artificial (Cabello (2022)), con sus limitaciones (Minsky y Papert (1969)) y potencial (Hornik et al. (1989)) así como su dependencia de la arquitectura en la que estén incluidas (Mhaskar y Poggio (2016)). Se exponen también las neuronas convolucionales y se plantea la importancia de las capas a la hora de definir una arquitectura (Chollet (2017)). De cara a una formalización rigurosa, se definen las capas densas y convolucionales. Esto lo consideramos particularmente relevante porque no hemos sido capaces de encontrar una formalización que incluya el concepto de *capa* (*layer* en inglés) pese a que es el bloque básico de construcción de arquitecturas más utilizado en la práctica. A partir de ahí se definen las arquitecturas sin retroalimentación (*Feedforward Neural Network*) (Goodfellow et al. (2016)), que son las únicas utilizadas en los trabajos de esta tesis. Se muestran conceptos como el Perceptrón, las funciones de activación, las funciones de pérdida y las estrategias de entrenamiento.

La sección continúa ofreciendo una exposición de las Redes Autocodificadoras (*Autoencoder* (Ballard (1987)), sus versiones utilizadas para eliminar ruido (*Denoising Autoencoder* y *Stacked Denoising Autoencoder*) (Zhang et al. (2015)) y su versión Variacional (*Variational Autoencoder* con propiedades generativas (Kingma y Welling (2013)) que incluye la definición de la Divergencia KL (Kullback y Leibler (1951)).

Las siguientes subsecciones incluyen la descripción y principales soluciones de los problemas de Reconocimiento de Objetos (LeCun et al. (1989), Krizhevsky et al. (2017)); Detección de Objetos incluyendo su enfoque de múltiples fases (*multi-*

stage) (Girshick et al. (2014), Girshick (2015), Ren et al. (2017)) y soluciones de una única pasada (*single-stage*) (Liu et al. (2016), Redmon et al. (2016)); Segmentación Semántica (Chen et al. (2018)) y Segmentación de Instancias (He et al. (2017)).

Para terminar el Estado del Arte se ofrece una pequeña descripción de la relación entre los avances en los dispositivos físicos (*hardware*) y los avances en Aprendizaje Profundo (Dally et al. (2021)).

B.3 Trabajos que apoyan esta Tesis

El **primero** de los ocho trabajos presentados en esta tesis se denomina *The effect of downsampling-upsampling strategy on foreground detection algorithms* (*El efecto de la estrategia de reducir y aumentar el tamaño de la imagen en los algoritmos de detección de primer plano* en castellano) y fue publicado en el año 2020 en la revista AIRe (Artificial Intelligence Review), que dicho año ocupó posición en el primer cuartil (14/139) de la clasificación JCR en categoría Inteligencia Artificial. El mencionado trabajo se enmarca en el estudio de la obtención del primer plano mediante el uso de métodos publicados anteriormente dentro de un metamétodo que aumente su rendimiento. Para ello se plantea un metamétodo y un estudio con amplios experimentos para ver sus resultados.

El metamétodo propuesto consiste en, dada una secuencia de vídeo con resolución $N \times M$ y un método de Segmentación de Primer Plano(FS) m, aplicar un proceso de reducción de la resolución a las imágenes que componen la secuencia para trabajar sobre imágenes con resolución $N' \times M'$, siendo $N' < N$ y $M' < M$. El método m se aplica entonces a la secuencia de menor resolución para obtener imágenes segmentadas de tamaño $N' \times M'$ que luego se devuelven al tamaño original $N \times M$. Como métodos para reducir la resolución se aplican el Vecino Más Cercano (NN (Nearest Neighbor)), la Media por Ventanas (AVG (Window Averaging)), Interpolación Lineal (LIN (Linear Interpolation)) e Interpolación Bicúbica (CUB (Bicubic Interpolation)). Como método para incrementar el tamaño de las imágenes segmentadas a la resolución original se usa siempre la Interpolación Bicúbica (CUB). Como métodos de Segmentación de Primer Plano se usan MFBM (López-Rubio y López-Rubio (2015)), Wren (Wren et al. (1997)), Grimson (Stauffer y Grimson (1999)), Zivkovic (Zivkovic (2004)), SOBS (Maddalena y Petrosino (2008)), SOBS_CF (Maddalena y Petrosino (2010)), SuBSENSE (**?**), LOBSTER (St-Charles y Bilodeau (2014)) y PAWCS (St-Charles et al. (2016)).

Los experimentos implican el uso de los conjuntos de datos *ChangeDetection* (Goyette et al. (2012)) con 31 vídeos y CAMO_UOW (Li et al. (2018)) con 10 vídeos de alta resolución. Además de las originales, las secuencias se analizan aplicándoles reducciones que multiplican su resolución por 0.125, 0.25, 0.375, 0.5, 0.625, 0.75 y 0.875 de cara a estudiar cómo afectan los distintos métodos de reducción de tamaño así como los métodos de Segmentación de Primer Plano. Se analizan no solo la calidad de los resultados sino la velocidad de procesamiento (incluyendo el tiempo que toma reducir el tamaño de las imágenes de entrada y volver a aumentar el tamaño de las segmentaciones resultantes).

El estudio muestra que las medidas de calidad se mantienen estables hasta un factor de reducción de 0.75 de manera general y según el método de reducción utiliado hasta con un factor de 0.5. En la mayoría de los métodos de Segmentación de Primer Plano se observa un aumento notable del rendimiento en cuanto a veloci-

dad al utilizar el metamétodo y con las secuencias de alta resolución puede llegar a reducir entre 3 y 10 veces el tiempo total de ejecución sin resentir la calidad.

El **segundo** de los trabajos presentados se denomina *Background subtraction by probabilistic modeling of patch features learned by deep autoencoders* (*Sustracción de fondo mediante modelado probabilístico de características de parches aprendidas por autocodificadores profundos*) y fue publicado en el año 2020 en la revista ICAE (Integrated Computer-Aided Engineering). La mencionada revista ocupó en el año 2020 posición en el segundo cuartil pero primer tercil de la clasificación JCR en categoría Inteligencia Artificial (37/139). Dicho trabajo continúa con el estudio de la detección de anomalías de primer plano mediante la propuesta del método PMDAPF (Probabilistic Mixture of Deeply Autoencoded Patch Features) especialmente diseñado para ser resistente al ruido.

PMDAPF (Probabilistic Mixture of Deeply Autoencoded Patch Features) es un método de Segmentación de Primer Plano basado en el análisis de secciones de la imágen de tamaño $N \times N$ denominados parches. Estos parches suelen ser de tamaño $16x16$ y sirven de entrada al codificador de una red neuronal autocodificadora (SDA (Stacked Denoising Autoencoder)) previamente entrenada para obtener una versión codificada en un vector de tamaño $L \ll N * N$ libre de ruido que debe contener las características principales del parche. Se usan las versiones codificadas para generar un modelo de fondo para cada región de tamaño $N \times N$. Una vez con un modelo de fondo, las siguientes imágenes se subdividen en los parches con las mismas secciones, se codifican con la misma red neuronal y el resultado se compara con el modelo de fondo para obtener la probabilidad de que se trate de primer plano. Después se realiza la actualización del modelo de cada parche ponderada con la probabilidad resultante de que pertenezca al fondo.

El planteamiento basado en parches tiene un problema intrínseco respecto a la resolución. La clasificación de cada parche es uniforme para todo el parche, por lo que la segmentación de primer plano se hace por secciones de tamaño $N \times N$. Por ello se introduce una estrategia de solapamiento entre parches que permite incrementar la resolución a costa de incrementar el coste computacional.

Los experimentos se realizan sobre 26 secuencias de cinco categorías del conjunto de datos *ChangeDetection* a los que se aplica nueve tipos de ruido distintos: cuatro niveles de ruido Gaussiano, ruido Sal y Pimienta (puntos blancos y negros), dos ruidos Uniformes y dos niveles de ruidos provocado por Compresión en la imagen. En total, junto con las secuencias originales, se usan 260 vídeos diferentes. La evaluación respecto a otros métodos del estado del arte muestra que PMDAPF ofrece mayor robustez a gran variedad de ruido que los competidores aunque con secuencias libres de dicho ruido su peor resolución lo haga ser una peor opción.

El **tercer** trabajo presentado lleva por título *Foreground detection by probabilistic mixture models using semantic information from deep networks* (*Detección de primer plano mediante modelos probabilísticos que usan información semántica de redes profundas*) y fue presentado en el congreso internacional ECAI (European Conference on Artificial Intelligence) del año 2020. El mencionado congreso tuvo una clasificación A en la clasificación CORE y A- en la clasificación GGS. El trabajo trata también el problema Segmentación de Primer Plano a partir de una red de segmentación semántica.

El método propuesto se basa en crear un modelo de fondo a partir de las máscaras a nivel de píxel proporcionadas por una red neuronal preentrenada de segmentación semántica o de instancias. Esta información es utilizada para crear el

modelo de fondo y posteriormente se proporciona a un modelo probabilístico de cara a obtener una segmentación de primer plano a nivel de píxel. Estás segmentaciones seguidamente son procesadas mediante operadores morfológicos para limpiar la imagen.

La propuesta tiene la ventaja de construirse como un metamétodo sobre un modelo de segmentación semántica: el modelo de segmentación semántica es fácilmente sustituible y los avances en dicho problema deberían implicar una mejora en el funcionamiento del modelo propuesto. Por otro lado, las carencias en el modelo de segmentación semántica implican carencias en el método de segmentación de primer plano. Por ejemplo, objetos que no sean detectados por el modelo de segmentación semántica, no podrán ser detectados como objetos en movimiento.

Los experimentos se realizan sobre cuatro categorías del conjunto de datos *ChangeDetection* y usando siete métodos del estado del arte como referencia. Las pruebas muestran que la propuesta es efectiva siempre que las segmentaciones a nivel de píxel en las que se basa funcionen correctamente. Como ventaja adicional, el método es inmune al ruido intrínseco de la escena (hojas en movimiento por el viento u ondulación de agua) porque no son objetos detectables por el modelo de segmentación semántica.

El **cuarto** trabajo incluído en esta tesis tiene por título *Deep autoencoder architectures for foreground object detection in vídeo sequences based on probabilistic mixture models* (*Arquitecturas para autocodificadores profundos para la detección de objetos de primer plano en secuencias de vídeos basada en modelos de mixturas probabilísticas*) y se presentó en el congreso ICIP (IEEE International Conference on Image Processing) del año 2020 que tuvo categoría B en la clasificación CORE y A- en la clasificación GGS. En este caso consiste en un estudio de la aplicación de redes neuronales autocodificadoras al problema de Segmentación de Primer Plano.

En este trabajo se realiza un estudio sobre cómo afectan la arquitectura, el pre-entrenamiento, el tipo de Redes Autocodificadoras y la función de activación a modelos de Segmentación de Primer Plano basados en parches como el PMDAPF anteriormente descrito. En el estudio se proponen seis arquitecturas distintas, dos de ellas basadas exclusivamente en capas densas, dos de ellas compuestas exclusivamente por capas convolucionales y dos de ellas compuestas por una mezcla de ambas. Se proponen dos de cada tipo para tener en cuenta distintas profundidades, por lo que cambian el número de capas.

En cuanto al tipo de Red Autocodificadora, se plantea el uso de las Redes Autocodificadoras clásicas (SDA (Stacked Denoising Autoencoder)) y el uso de las Redes Autocodificadoras Variacionales (VAE (Variational Autoencoder)). Estas últimas son frecuentemente utilizadas como modelos generativos y en teoría ofrecen un espacio latente mejor regularizado. Respecto al pre-entrenamiento, se plantean pre-entrenamientos con y sin ruido Gaussiano añadido para mejorar la capacidad del modelo del limpiar el ruido de los parches. También se prueban dos funciones de activación distintas para la capa más intensa: Tangente Hiperbólica y Sigmoide. En total se entrenan 48 redes neuronales.

Los experimentos que se plantean usan la categoría *baseline* de *ChangeDetection* con tres tipos de ruido añadidos: Gaussiano, Uniforme y Sal y Pimienta. Para evaluarlo se han atendido especialmente los ratios de Falsos Positivos y Falsos Negativos. En las pruebas se observa una resistencia generalizada al ruido que se ve incrementada con las elecciones adecuadas. Se observa un frente de Pareto en la relación entre los Falsos Positivos y Falsos Negativos obtenidos en función a la con-

figuración de entrenamiento. Los VAE tienden a un bajo ratio de Falsos Negativos a costa de aumentar el ratio de Falsos Positivos para todos los ruidos con todas sus configuraciones, mientras que los SDA muestran el comportamiento contrario. De media se observa una mejora notable de rendimiento si las redes se entrenan con ruido mientras que la función de activación no parece un elemento determinante, aunque los resultados con la Sigmoide son más constantes. El tipo de capas parece favorecer el uso de capas mixtas.

El **quinto** trabajo referido es un estudio llamado *Foreground segmentation improvement by image denoising preprocessing applied to noisy vídeo sequences* (*Mejora de la segmentación de primer plano mediante el uso de preprocesamientos de eliminación de ruido en imágenes aplicado a secuencias de vídeo*). Fue presentado en el congreso SOCO de 2021, clasificado como *Work in Progress*. El trabajo consiste en un estudio sobre el uso de diversas técnicas de preprocesado de imágenes en el contexto de un metamétodo que provea robustez al ruido a otros métodos de Segmentación de Primer Plano.

El metamétodo es sencillo: aplicar un método que limpie el ruido de una secuencia de vídeo antes de extraer la máscara de primer plano de ella. La idea clave es estudiar si es factible darle uso hoy día a los métodos de Segmentación de Primer Plano más antiguos que son generalmente los más rápidos pero también los menos resistentes al ruido. De cara a probar si es factible, se plantea el uso de dos Redes Autocodificadoras entrenadas para eliminar ruido (SDA), una con capas densas y otras con capas convolucionales; y dos filtros clásicos: el de la mediana reemplaza el valor de cada píxel por el valor mediano de la ventana que lo rodea y el de la media hace lo mismo con el valor medio. Los filtros se aplican con ventanas de 5×5. Como métodos de segmentación de primer plano se usan Wren, Zivkovic y KDE (Elgammal et al. (2000)). El entrenamiento de las redes neuronales se realiza utilizando el conjunto de datos *Tiny Images* (Torralba et al. (2008)) en el caso de la red de capas densas y el conjunto *ChangeDetection* para la red de capas convolucionales.

Como experimentos para poner a prueba el metamétodo se toma la categoría *dynamicBackground* de *ChangeDetection*, que incluyen ruido intrínseco en el fondo de la escena y se obtienen además versiones modificadas de las escenas añadiendo ruido Gaussiano de dos niveles de agresividad: medio ($\mu = 0$ y $\sigma = 0.2$) y muy agresivo ($\mu = 0$ y $\sigma = 0.4$). Se comparan los resultados con los obtenidos usando el método PAWCS, que es mucho más moderno y ofrece por sí mismo notable resistencia al ruido para tener una referencia del comportamiento de un método reciente.

Los resultados indican que el metamétodo es una opción viable si se hace una combinación adecuada entre el método de preprocesamiento y el método de segmentación. KDE, por ejemplo, tiende a ser sensible al ruido aún con las secuencias preprocesadas y solo tiene resultados mínimamente razonables con secuencias filtradas mediante un SDA con capas densas. Zikvovik tiene una gran mejora en sus resultados con los cuatro métodos de filtrado propuestos y Wren también. PAWCS por su parte muestra incapacidad para procesar correctamente las secuencias con ruido gaussiano medio y muy agresivo, obteniendo resultados mucho peores que métodos mucho más antiguos y rápidos en combinación con el preprocesado, por lo que el método se considera una posible aproximación apropiada al problema.

El **sexto** trabajo de esta tesis se denomina *Road pollution estimation from vehicle tracking in surveillance vídeos by deep convolucional neural networks* (*Es-*

timación de polución en carreteras mediante el seguimiento de vehículos en vídeos de videovigilancia usando redes neuronales convolucionales profundas). El artículo se publicó en la revista ASoC (Applied Soft Computing) en 2021 y en dicho año la revista ocupaba posición en el primer cuartil (23/145) en la categoría IA de la clasificación JCR. Este trabajo, a diferencia de los anteriores, aplica las redes neuronales convolucionales para el análisis del tráfico.

El trabajo incluye una propuesta consistente en estimar la polución en base a una estimación de la velocidad. La propuesta parte de secuencias de cámaras de tráfico para las que se busca la localización exacta en algún servicio de imágenes satelitales. A partir de las dos imágenes se obtiene una transformación homográfica para corregir la perspectiva de la cámara y poder usar la referencia de distancia de la imagen por satélite. Con esa información, se usa un modelo pre-entrenado de detección de objetos para seguir a los vehículos y el movimiento del centroide de sus detecciones para obtener las distancias recorridas. Así se puede obtener una estimación de la distancia recorrida entre cada par de imágenes y, dado que es conocido el tiempo entre una y otra, se puede estimar la velocidad. Una vez con la velocidad, se estima la polución a partir del Factor de Emisión. Este factor se establece en unidades de litro por cada 100 km por lo que se realizan las transformaciones necesarias para obtenerlo en gramos por *frame* de la secuencia.

La propuesta del artículo es muy novedosa y en la bibliografía no se encontró nada similar que permitiera estimar poluciones a partir de imágenes sin necesidad de *hardware* específico. Por desgracia esto implica que tampoco se encontraron disponibles conjuntos de datos específicos para poner a prueba la solución. Los experimentos se basan en estimar las poluciones para un par de escenas con dos modelos de detección de objetos distintos y con detección manual para observar si los resultados eran razonables y cómo afecta la detección de objetos en el sistema. Los resultados muestran que es factible, al menos en teoría, realizar la estimación de la polución en tiempo real en base a imágenes, pero requiere la generación de conjuntos de datos de estudio para valorar en detalle la precisión del sistema.

El **séptimo** trabajo lleva por título *Vehicle overtaking hazard detection over onboard cameras using deep convolutional networks* (*Detección de adelantamientos peligrosos de vehículos sobre cámaras integradas en coches usando redes convolucionales profundas*) y fue presentado en el congreso SOCO (International Conference on Soft Computing Models in Industrial and Environmental Applications) de 2022. El trabajo continúa la línea del anterior de estimar velocidades, pero esta vez lo hace partiendo de cámaras incorporadas al vehículo.

La propuesta en este caso se basa en utilizar la evolución del tamaño de las Cajas de Detección (BBOX) para estimar la velocidad relativa de otros vehículos. Para hacerlo, se aplica la Aproximación del ángulo pequeño y se obtiene una relación lineal de la inversa del diámetro aparente de los vehículos (su tamaño) respecto al tiempo asumiendo que el vehículo se mueva a velocidad constante. Una vez se establece que la inversa de los tamaños deben poder representarse en una recta, la velocidad relativa del otro vehículo (la velocidad a la que cambia su tamaño) será la pendiente de esa recta. Dos vehículos cuya velocidad relativa es cero mostrar una recta con inclinación cero, mientras que a mayor sea esa velocidad relativa, más inclinación tendrá la recta. A partir de ahí, establecemos un umbral y consideramos que cualquier vehículo que se acerque o aleje a una velocidad mayor que ese umbral, implica un adelantamiento peligroso.

Como con el caso anterior, la falta de datos experimentales dificulta poner a

prueba la teoría, pero en este caso se genera un conjunto de datos consistente en cuatro vídeos con 23 adelantamientos que anotamos manualmente de cara a tener una referencia a la hora de hacer las pruebas. Los experimentos, aunque básicos, muestran una posibilidad novedosa en cuanto al uso de cámaras incorporadas a los coches para estimar las velocidades relativas de otros coches sin necesidad de recurrir a *hardware* específico como LiDAR (Light Detection and Ranging). Consideramos por tanto que es una buena línea de trabajo a explorar con conjuntos de datos más desarrollados y metodologías más pulidas.

Como se ha comentado previamente al presentar el método PMDAPF, la necesidad de subdividir una imagen en parches de $N \times N$ para aplicar la codificación implica que la resolución de la imagen segmentada sea menor o aplicar alguna estrategia de solapamiento como hace PMDAPF que aumenta sensiblemente la necesidad de cómputo para paliar el problema. Para solucionarlo se propone una alternativa consistente en utilizar una Red Autocodificadora que tiene únicamente capas convolucionales. Al aplicar una capa convolucional con K filtros a una imagen RGB de tamaño $G \times H \times 3$, se obtiene (sin añadir relleno a la imagen), un cubo de datos con tamaño $(G-2) \times (H-2) \times K$. Hacerlo sucesivamente en la Red Autocodificadora permite obtener en el centro de esta una representación de tamaño $G' \times H' \times K'$. Si se aplica un número moderado de capas (los experimentos se hacen con 4, 5 y 6 capas de codificación y otras tantas de decodificación), la distancia entre cada valor del par (G, H) y su valor correspondiente en el par (G', H') es pequeña (12 a lo sumo) y la pérdida de resolución es mínima. Sin embargo, por la naturaleza de la capa convolucional, cada vector de tamaño K' en la posición (i, j) a lo largo de la profundidad del cubo va a contener información no solo de la posición (i, j) en los datos de entrada a la capa sino de los vecinos de dicha posición. En definitiva, cada capa convolucional hace que en el vector de profundidad de posición (i, j) se acumule información de toda la región que rodea esa posición, con lo que se obtiene una representación codificada de las regiones sin haber dividido la imagen en parches. Estos vectores de tamaño K' se utilizan para generar el modelo de fondo de cada posición y con un modelo probabilístico se van clasificando esas posiciones en las siguientes imágenes. Una vez se obtiene una segmentación de tamaño $G' \times H'$ se aplica un método para escalar la imagen al tamaño original $G \times H$ para hacerla coincidir con los datos originales.

Esta aproximación se presenta junto a un estudio que explora la profundidad de la Red Autocodificadora (8, 10 u 12 capas de convolución en total) y su proceso de entrenamiento. En cuanto al entrenamiento, se plantean tres opciones: la opción genérica es tener una única red pre-entrenada con datos sacados del conjunto de datos *ImageNet* Deng et al. (2009) con ruido genérico (Gaussiano $\mu = 0$ y $\sigma = 0.2$); la segunda opción es la realista, en la que se usan trozos del principio de la propia secuencia que se va a procesar sin añadir más ruido durante el entrenamiento; la tercera es la opción hipotéticamente ideal, en la que entrena una red para cada secuencia y se usa la secuencia limpia y se añade el apropiado ruido durante el entrenamiento.

Para experimentar se usan diez secuencias de las categorías *baseline* y *dynamicBackground* de *changeDetection* y se generan versiones con ruido usando ruido Gaussiano bajo ($\mu = 0$ y $\sigma = 0.1$), medio ($\mu = 0$ y $\sigma = 0.2$) y alto ($\mu = 0$ y $\sigma = 0.3$) además de ruido Uniforme. No se usan las versiones originales de las secuencias así que en total se estudian 40 secuencias distintas y se entrenan $3 \times (1+40+40) = 243$ redes neuronales.

Las conclusiones validan la aproximación, demostrando que se consigue gran resistencia al ruido con cualquiera de las opciones de entrenamiento elegidas. Además os muestra que las redes entrenadas con el planteamiento genérico obtienen los mejores resultados contra lo esperado. Esto es doblemente bueno, a parte de ser un planteamiento que asume muy poca información previa, es el que requiere menos capacidad de cómputo porque requiere entrenar solo una red para todas las secuencias. En cuanto a la profundidad de la red, parece premiarse que sea poco profunda.

B.4 Conclusiones y Trabajo Futuro

Como conclusiones se presentan los resultados obtenidos tras los años de investigación dedicados a esta tesis doctoral en Inteligencia Artificial así como algunos de los planteamientos para el trabajo futuro.

B.4.1 Conclusiones

El objetivo principal de esta tesis doctoral ha sido la SPP (Segmentación de Primer Plano), un problema clásico de VC (Visión por Computador) que se ha tratado durante las últimas décadas y que, en consecuencia, había sido ya trabajado desde multitud de aproximaciones con muy buenos resultados antes de que esta investigación comenzara. Por lo tanto, en un problema tan trabajado es normal que resulte especialmente complicado encontrar un hueco donde poder hacer aportaciones realmente significativas. Nuestro enfoque basado en combinar el procesamiento de la imagen por parte de Redes Neuronales Artificiales preentrenadas o redes con entrenamiento no supervisado con modelos probabilísticos ofreció ese pequeño hueco donde se pudieron realizar la mayoría de las contribuciones. Por el camino, como es normal cuando uno está aprendiendo sobre distintos problemas y sus soluciones, surgieron ideas para aplicar las Redes Neuronales Artificiales a problemas de tráfico y ello constituye la pata secundaria de esta tesis doctoral.

Para apoyar dicha investigación se han incluido en la tesis ocho trabajos, de los cuales el autor de la tesis es el primer autor de todos excepto el primero. De esos ocho trabajos, tres han sido publicados en revistas de alto impacto, otros tres fueron publicados en congresos internacionales bien posicionados y dos trabajos menores en los que se ponían a prueba ideas tentativas fueron publicados en congresos de menor categoría.

Es importante señalar que la mayoría de los trabajos están estrechamente relacionados entre sí, de esta manera, el trabajo publicado en AIRe (capítulo 3) es una aproximación al problema de la SPP centrándose en mejorar el resultado de otros métodos mediante un preprocesamiento específico de la imagen. Este planteamiento se vuelve a utilizar en el trabajo publicado en el SOCO del año 2021 (capítulo 7). En el primer caso el preprocesamiento se centró en alterar el tamaño de la imagen para estudiar principalmente cómo afecta al rendimiento en términos de velocidad con conclusiones muy positivas que indican que con el ajuste adecuado se puede conseguir una mejora de velocidad notable sin perder calidad en la segmentación. Mientras, en el segundo trabajo, se utilizan filtros clásicos y filtrado basado en redes auto-codificadoras para limpiar las imágenes de ruido y así conseguir resultados robustos a esta clase de alteraciones en las imágenes con resultados variados que

aluden a la necesidad de combinar apropiadamente el tipo de filtrado previo con el método de SPP que se usará después.

El trabajo publicado en ICAE (capítulo 4) se basa también en utilizar redes auto-codificadoras. En este caso las redes sirven para transformar parches de la imagen en vectores de datos que deben contener la información principal (pero no el ruido) de los parches como si de un compresor se tratara. Estos datos son luego evaluados usando un modelo probabilístico para determinar si ese parche de imagen pertenece o no a primer plano. Debido a las restricciones a la resolución que plantea una aproximación a nivel de parche, en dicho trabajo se utiliza una estrategia de solapamiento entre parches para aumentar la resolución a coste de incrementar los requisitos computacionales. Los extensos experimentos muestran que se consigue un sistema capaz de segmentar el primer plano con una importante robustez al ruido en la imagen. En el trabajo publicado en el congreso ICIP de 2020 (capítulo 6) se hace un estudio de cómo el tipo de red auto-codificadora y su entrenamiento afecta a este tipo de modelos. En dicho trabajo se llegan a utilizar variantes a las redes auto-codificadoras clásicas como las redes auto-codificadoras variacionales, observando que estas últimas son especialmente efectivas en contextos en los que se deseen reducir los Falsos Negativos aun aumentando los Falsos Positivos. Esta línea de trabajo culmina con la publicación en el ICIP de 2022 (capítulo 10) con un método que utiliza los mismos conceptos de redes auto-codificadoras como paso previo a modelos probabilísticos, pero elimina la necesidad de dividir la imagen en parches al trabajar sobre la profundidad de los filtros de convolución. De esta manera se consigue resistencia a ruido sin perder prácticamente resolución en la segmentación ni necesitar una estrategia de solapamiento que aumente sustancialmente los cálculos que realizar.

En paralelo, pero estrechamente relacionado, se propuso en el congreso ECAI de 2020 (capítulo 5) un meta-método que analiza con un modelo probabilístico las máscaras de segmentación a nivel de píxel de un modelo de Segmentación de Instancias (He et al. (2017)). Los resultados en este caso fueron muy positivos ya que al fundamentar el sistema sobre un modelo para una tarea tan general como la Segmentación de Instancias, el meta-método propuesto mejorará conforme los problemas de dicho modelo se solucionen o aparezca uno similar con mejores resultados.

Todos estos trabajos se centran en el objetivo 1 descrito en la sección 1.2 al principio de esta tesis. En ellos, las aproximaciones al objetivo 1 se realizan principalmente utilizando parches y redes autocodificadoras ya que demostraron ser una forma prometedora de tratar el ruido en la segmentación de primer plano. Sin embargo, también se incluyen dos estudios que reutilizan algoritmos clásicos y un trabajo basado en el aprovechamiento de algoritmos de segmentación a nivel de píxel. En conjunto, aunque el problema dista mucho de estar resuelto, hemos realizado importantes contribuciones a la identificación de anomalías de primer plano y, en particular, a hacerlo en situaciones ruidosas.

La segunda vertiente de esta tesis son los dos métodos propuestos para hacer análisis del tráfico basándonos en aplicar Aprendizaje Profundo a imágenes. En este caso, los dos métodos se basan en la estimación de la velocidad de los vehículos partiendo desde planteamientos totalmente distintos. En el trabajo publicado en ASoC en 2021 (capítulo 8), se propone un método para, a partir de una transformación homográfica y las imágenes de una cámara de tráfico estática, estimar la velocidad y de ella la polución generada por los vehículos. En el trabajo publicado

en el SOCO de 2022 (capítulo 9) se cambia la perspectiva de la cámara y trabajamos con cámaras incorporadas al coche y a partir del cambio en el tamaño del BBOX de cada vehículo se estima la velocidad relativa de los otros conductores para así intentar advertir de adelantamientos peligrosos. Estos dos trabajos han supuesto el enfrentamiento a problemas distintos, pero trabajando con herramientas y paradigmas similares. Ambos han supuesto un desafío porque los problemas que intentan resolver se han tratado muy poco utilizando únicamente imágenes y sin depender de hardware específico, por ello los medios para evaluar la fiabilidad de los modelos aún son pobres y están poco trabajados, pero también por esa misma razón resultan especialmente interesantes de cara al futuro.

Estos dos trabajos se centran en el objetivo 2 descrito en el apartado 1.2 al principio de esta tesis. En el primero, aunque no se realiza un análisis de anomalías, se trabaja en el paso previo de inferir velocidades solo a partir de cámaras estáticas. En el segundo, se trabaja con cámaras en movimiento y la aplicación de la estimación de velocidad es directamente identificar adelantamientos anómalos. Ambos artículos, más que alcanzar el objetivo 2, demuestran que se trata de una línea de trabajo prometedora.

B.4.2 Trabajo Futuro

La realización de una tesis doctoral tiene que tener un fin, aunque en principio debe ser solo el inicio de la carrera investigadora del doctorando. Es natural, por tanto, que de ella salgan además de unas conclusiones, líneas de trabajo en las que se continúe lo realizado, ya sea refinando las propuestas hechas, haciendo otras para el mismo problema o abordando problemas distintos con los planteamientos estudiados durante el periodo de investigación.

Como conclusión a la tesis es importante añadir una general sobre el problema de la Segmentación de Primer Plano de cara a futuro. Pese a que en el pasado resultó una tarea extremadamente relevante como paso previo a resolver otros problemas de Visión por Computador, el Aprendizaje Profundo ha hecho que en muchos casos ese paso previo sea innecesario. Siendo un problema tan trabajado y con tantas propuestas distintas, que además pierda relevancia hace que no sea especialmente atractivo continuar trabajando en él de manera genérica.

A partir de los trabajos realizados, la investigación futura fruto de esta tesis se divide en cuatro líneas principales:

- Continuar trabajando en los modelos de primer plano propuestos para intentar mejorar su precisión al detectar movimiento.
- Trabajar en la aceleración de la línea de segmentación de primer plano basado en el uso de Redes Autocodificadoras.
- Extender el análisis de anomalías a anomalías de trayectorias en secuencias de vídeo utilizando un planteamiento similar al utilizado en los trabajos de la tesis.
- Profundizar en el análisis de velocidades basado en imágenes sin utilizar hardware específico.

B.4.2.1 Mejora de Detección de Primer Plano

La línea más obvia de trabajo a corto plazo es seguir iterando sobre los modelos que se han propuesto a lo largo de la tesis para pulir sus defectos en cuanto a calidad de la segmentación de primer plano. En ese sentido hay tres puntos donde se podrían hacer mejoras:

- Mejorar el criterio de actualización del fondo mediante el uso de técnicas de Aprendizaje Continuado de cara a identificar cuando los modelos de fondo que estamos manejando se han alejado demasiado de la distribución real del fondo antes de que se corrompan por completo.
- Sustituir el criterio del Modelo Probabilístico por alguna técnica de Aprendizaje Supervisado como un Perceptrón Multicapa que clasifique las representaciones de espacio latente resultante de las redes autocodificadoras.
- Incorporar técnicas de Aumentación de Datos a la hora de entrenar las redes autocodificadoras de cara a que aprendan una distribución de probabilidad más variada y acorde con la realidad.

B.4.2.2 Aceleración

Como hemos mencionado previamente, el papel que realizaba la Segmentación de Primer Plano en el esquema de trabajo típico de un sistema de Visión de Computador ha sido desplazado entre otras cosas por la rápida mejora de los sistemas de Detección de Objetos. Dado que los sistemas de Detección de Objetos basados en Aprendizaje Profundo aplican directamente detectores de características aprendidos automáticamente a toda la imagen, no tiene mucho sentido detectar previamente en qué zona de la imagen hay objetos de interés en base al movimiento. Por otro lado, si lo que se desea es detectar el movimiento, se puede combinar un detector de objetos con un sistema de seguimiento como se hizo en los trabajos de los capítulos 8 y 9 para hacerlo.

¿Cuando podría tener sentido entonces aplicar un proceso previo de detección de movimiento? Cuando interesen objetos que se estén moviendo y suponga una diferencia de tiempo sustancial. Tanto los sistemas de Detección de Objetos como los que segmentan la imagen a nivel de píxel son todavía deficientes cuando la cantidad de objetos en la imagen es muy alta y su tamaño muy pequeño. En ese sentido existen propuestas tanto para detectar (García-Aguilar et al. (2023a, 2022b)) como para segmentar (García-Aguilar et al. (2023b, 2022a)) objetos que van enfocadas a mejorar la detección de pequeños objetos mediante el uso de Super-Resolución. Podría tener sentido, por ejemplo, hacer un sistema similar que en vez de basarse en una pasada de detección para saber qué región estudiar con más detalle, detecte las zonas donde aplicar Super-Resolución en base al uso Segmentación de Primer Plano.

Esto puede tener dos ventajas:

- Aumentar la velocidad porque la Segmentación de Primer Plano sea más rápida que aplicar el modelo de Detección de Objetos.
- Aumentar la fiabilidad porque la Segmentación de Primer Plano no pase por alto objetos pequeños que el modelo de detección sí omita.

Por lo tanto, para que el uso de Segmentación de Primer Plano tenga sentido debe más que nunca ser rápido y preciso. En ese sentido, una posible línea de trabajo futuro consistiría en centrar los esfuerzos en optimizar a nivel de tiempo el trabajo presentado en el capítulo 10).

B.4.2.3 Detección de Trayectorias Anómalas

Aunque los trabajos de investigación de esta tesis se han centrado en la identificación de anomalías de primer plano en base al movimiento, existen gran cantidad de anomalías analizables en secuencias de vídeo. De manera análoga al uso que hemos hecho del codificador del SDA que en nuestros trabajos mientras ignorábamos el decodificador, planteamos el posible uso de redes GAN para detectar anomalías en las trayectorias de movimiento utilizando no la red generadora, sino el discriminador que se entrena con ella.

El planteamiento es como sigue:

1. Partiendo de una secuencia con comportamiento usual, obtener alguna representación visual del flujo del movimiento como el Flujo óptico.
2. Entrenar una red GAN con dichas imágenes. Esto tendrá como resultado una red generadora para crear imágenes con un flujo óptico similar y una red discriminadora que distinga qué imágenes tienen un flujo óptico similar.
3. Usar la red discriminadora sobre otras imágenes de flujo óptico para distinguir los flujos ópticos anómalos.

B.4.2.4 Análisis de Velocidades

Dos de los trabajos presentados en esta tesis doctoral se basan en el análisis de velocidades y distancias en base únicamente a imágenes. El análisis de velocidades es un problema que puede resultar trivial cuando se dispone de *hardware* específico como puede ser un LiDAR o un RADAR. Esos artefactos están creados específicamente para obtener la distancia a otros objetos, pero dejando de lado los problemas intrínsecos que tiene la forma de funcionar de cada uno, hay una desventaja enorme respecto a las cámaras: hay muchísimas más cámaras y tienen más propósitos.

Es mucho más plausible contar con la presencia de una cámara que la presencia de otro *hardware* y eso hace especialmente interesante continuar una línea de trabajo en la que se puedan obtener velocidades usando una. Esta es una línea de trabajo que hay que solidificar y para ello pensamos que es crucial la creación de conjuntos de datos con la información necesaria para evaluar los métodos propuestos debidamente. Por ello, el siguiente paso en esta línea de trabajo sería crear un conjunto de datos que usar y compartir con la comunidad científica que asocie de forma sencilla la velocidad y distancias relativas de dos vehículos.

Bibliography

Nadie puede ser sabio sin haber leído por lo menos una hora al día, sin tener biblioteca por modesta que sea, sin maestros a los que respetar, sin ser lo bastante humilde para formular preguntas y atender con provecho las respuestas...

Hombres Buenos, Arturo Pérez-Reverte.

Amer, M., Goldstein, M. y Abdennadher, S. Enhancing one-class support vector machines for unsupervised anomaly detection. páginas 8–15. ACM, 2013. ISBN 9781450323352.

Badrinarayanan, V., Kendall, A. y Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, páginas 2481–2495, 2017. ISSN 0162-8828.

Ballard, D. H. Modular learning in neural networks. 1987.

Bay, H., Ess, A., Tuytelaars, T. y Van Gool, L. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, vol. 110(3), páginas 346–359, 2008. ISSN 1077-3142. Similarity Matching in Computer Vision and Multimedia.

Bayes, T. Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, f. r. s. communicated by mr. price, in a letter to john canton, a. m. f. r. s. *Philosophical Transactions of the Royal Society of London*, vol. 53, páginas 370–418, 1763. ISSN 0261-0523.

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. y Amodei, D. Language models are few-shot learners. 2020.

Cabello, J. G. Mathematical neural networks. *Axioms*, vol. 11, página 80, 2022. ISSN 2075-1680.

CHEN, L.-C., PAPANDREOU, G., KOKKINOS, I., MURPHY, K. y YUILLE, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, páginas 834–848, 2018. ISSN 0162-8828.

CHILD, R. Very deep vaes generalize autoregressive models and can outperform them on images. En *International Conference on Learning Representations*. 2021.

CHOLLET, F. *Deep Learning with Python*. Manning Publications, 2017. ISBN 978-1-61729-443-3.

DALLY, W. J., KECKLER, S. W. y KIRK, D. B. Evolution of the graphics processing unit (gpu). *IEEE Micro*, vol. 41, páginas 42–51, 2021. ISSN 0272-1732.

DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K. y FEI-FEI, L. Imagenet: A large-scale hierarchical image database. páginas 248–255. IEEE, 2009. ISBN 978-1-4244-3992-8.

DEVLIN, J., CHANG, M.-W., LEE, K. y TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. páginas 4171–4186. Association for Computational Linguistics, 2019.

ELGAMMAL, A., HARWOOD, D. y DAVIS, L. Non-parametric model for background subtraction. páginas 751–767. Springer, 2000.

EVERINGHAM, M., GOOL, L. V., WILLIAMS, C. K. I., WINN, J. y ZISSERMAN, A. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, vol. 88, páginas 303–338, 2010. ISSN 0920-5691.

GARCÍA-AGUILAR, I., GARCÍA-GONZÁLEZ, J., LUQUE-BAENA, R. M. y LÓPEZ-RUBIO, E. Automated labeling of training data for improved object detection in traffic videos by fine-tuned deep convolutional neural networks. *Pattern Recognition Letters*, vol. 167, páginas 45–52, 2023a. ISSN 01678655.

GARCÍA-AGUILAR, I., GARCÍA-GONZÁLEZ, J., LUQUE-BAENA, R. M., LÓPEZ-RUBIO, E. y DOMÍNGUEZ, E. Optimized instance segmentation by super-resolution and maximal clique generation. *Integrated Computer-Aided Engineering*, páginas 1–14, 2023b. ISSN 10692509.

GARCÍA-AGUILAR, I., GARCÍA-GONZÁLEZ, J., LUQUE-BAENA, R. M., LÓPEZ-RUBIO, E. y DOMÍNGUEZ-MERINO, E. Enhanced image segmentation byÂ aÂ novel test time augmentation andÂ super-resolution. páginas 153–162. Springer International Publishing, 2022a. ISBN 978-3-031-06527-9.

GARCÍA-AGUILAR, I., LUQUE-BAENA, R. M. y LÓPEZ-RUBIO, E. Improved detection of small objects in road network sequences using <scp>cnn</scp> and super resolution. *Expert Systems*, vol. 39, 2022b. ISSN 0266-4720.

GARCIA-GARCIA, B., BOUWMANS, T. y SILVA, A. J. R. Background subtraction in real applications: Challenges, current models and future directions. *Computer Science Review*, vol. 35, página 100204, 2020. ISSN 1574-0137.

García-González, J., de Lazcano-Lobato, J. M. O., Luque-Baena, R. M. y López-Rubio, E. Background modeling by shifted tilings of stacked denoising autoencoders. vol. 11487 LNCS, páginas 307–316. 2019a.

García-González, J., de Lazcano-Lobato, J. M. O., Luque-Baena, R. M. y López-Rubio, E. Background subtraction by probabilistic modeling of patch features learned by deep autoencoders. *Integrated Computer-Aided Engineering*, vol. 27, páginas 253–265, 2020a.

García-González, J., de Lazcano-Lobato, J. M. O., Luque-Baena, R. M. y López-Rubio, E. Foreground detection by probabilistic mixture models using semantic information from deep networks. páginas 2696–2703. IOS Press, 2020b.

García-González, J., de Lazcano-Lobato, J. M. O., Luque-Baena, R. M. y López-Rubio, E. Foreground segmentation improvement by image denoising preprocessing applied to noisy video sequences. páginas 388–397. Springer International Publishing, 2021a.

García-González, J., de Lazcano-Lobato, J. M. O., Luque-Baena, R. M., ángel Molina-Cabello, M. y López-Rubio, E. Foreground detection by probabilistic modeling of the features discovered by stacked denoising autoencoders in noisy video sequences. *Pattern Recognition Letters*, vol. 125, páginas 481–487, 2019b.

García-González, J., de Lazcano-Lobato, J. M. O., Luque-Baena, R. M., Molina-Cabello, M. A. y López-Rubio, E. Background modeling for video sequences by stacked denoising autoencoders. vol. 11160 LNAI, páginas 341–350. 2018.

Garcia-Gonzalez, J., Luque-Baena, R. M., de Lazcano-Lobato, J. M. O. y Lopez-Rubio, E. Moving object detection in noisy video sequences using deep convolutional disentangled representations. páginas 1376–1380. IEEE, 2022. ISBN 978-1-6654-9620-9.

Garcia-Gonzalez, J., Molina-Cabello, M. A., Luque-Baena, R. M., de Lazcano-Lobato, J. M. O. y Lopez-Rubio, E. Deep autoencoder architectures for foreground object detection in video sequences based on probabilistic mixture models. IEEE, 2020.

García-González, J., Molina-Cabello, M. A., Luque-Baena, R. M., de Lazcano-Lobato, J. M. O. y López-Rubio, E. Road pollution estimation from vehicle tracking in surveillance videos by deep convolutional neural networks. *Applied Soft Computing*, página 107950, 2021b.

Girshick, R. Fast r-cnn. páginas 1440–1448. 2015.

Girshick, R., Donahue, J., Darrell, T. y Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. páginas 580–587. IEEE, 2014. ISBN 978-1-4799-5118-5.

Goodfellow, I., Bengio, Y. y Courville, A. *Deep Learning*. MIT Press, 2016.

GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A. y BENGIO, Y. Generative adversarial networks. *Communications of the ACM*, vol. 63, páginas 139–144, 2020. ISSN 0001-0782.

GOYETTE, N., JODOIN, P.-M., PORIKLI, F., KONRAD, J., ISHWAR, P. ET AL. Changedetection. net: A new change detection benchmark dataset. páginas 1–8. 2012.

HE, K., GKIOXARI, G., DOLLAR, P. y GIRSHICK, R. Mask r-cnn. páginas 2980–2988. IEEE, 2017. ISBN 978-1-5386-1032-9.

HO, J., JAIN, A. y ABBEEL, P. Denoising diffusion probabilistic models. páginas 6840–6851. Curran Associates, Inc., 2020.

HORNIK, K., STINCHCOMBE, M. y WHITE, H. Multilayer feedforward networks are universal approximators. *Neural Networks*, vol. 2, páginas 359–366, 1989. ISSN 0893-6080.

KINGMA, D. P. y WELLING, M. Auto-encoding variational bayes. 2013.

KOHONEN, T. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, vol. 43, páginas 59–69, 1982. ISSN 0340-1200.

KRIZHEVSKY, A., SUTSKEVER, I. y HINTON, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, vol. 60, páginas 84–90, 2017. ISSN 0001-0782.

KULLBACK, S. y LEIBLER, R. A. On information and sufficiency. *The Annals of Mathematical Statistics*, vol. 22, páginas 79–86, 1951. ISSN 0003-4851.

LECUN, Y., BOSER, B., DENKER, J. S., HENDERSON, D., HOWARD, R. E., HUBBARD, W. y JACKEL, L. D. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, vol. 1, páginas 541–551, 1989. ISSN 0899-7667.

LI, S., FLORENCIO, D., LI, W., ZHAO, Y. y COOK, C. A fusion framework for camouflaged moving foreground detection in the wavelet domain. *IEEE Transactions on Image Processing*, vol. 27, páginas 3918–3930, 2018.

LIU, F. T., TING, K. M. y ZHOU, Z.-H. Isolation forest. páginas 413–422. IEEE, 2008. ISBN 978-0-7695-3502-9.

LIU, W., ANGUELOV, D., ERHAN, D., SZEGEDY, C., REED, S., FU, C.-Y. y BERG, A. C. Ssd: Single shot multibox detector. *Lecture Notes in Computer Science*, páginas 21–37, 2016. ISSN 1611-3349.

LÓPEZ-RUBIO, E., LUQUE-BAENA, R. M. y DOMÍNGUEZ, E. Foreground detection in video sequences with probabilistic self-organizing maps. *International Journal of Neural Systems*, vol. 21, páginas 225–246, 2011. ISSN 0129-0657.

LÓPEZ-RUBIO, E., MOLINA-CABELLO, M. A., LUQUE-BAENA, R. M. y DOMÍNGUEZ, E. Foreground detection by competitive learning for varying input distributions. *International Journal of Neural Systems*, vol. 28, página 1750056, 2018. ISSN 0129-0657.

LÓPEZ-RUBIO, F. J. y LÓPEZ-RUBIO, E. Features for stochastic approximation based foreground detection. *Computer Vision and Image Understanding*, vol. 133, páginas 30–50, 2015.

MADDALENA, L. y PETROSINO, A. A self-organizing approach to background subtraction for visual surveillance applications. *Trans. Img. Proc.*, vol. 17, páginas 1168–1177, 2008. ISSN 1057-7149.

MADDALENA, L. y PETROSINO, A. A fuzzy spatial coherence-based approach to background/foreground separation for moving object detection. *Neural Computing and Applications*, vol. 19, páginas 179–186, 2010.

MADDALENA, L. y PETROSINO, A. The sobs algorithm: What are the limits? páginas 21–26. 2012.

MAZARBHUIYA, F. A. y SHENIFY, M. A mixed clustering approach for real-time anomaly detection. *Applied Sciences*, vol. 13, página 4151, 2023. ISSN 2076-3417.

MCCULLOCH, W. S. y PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, vol. 5, páginas 115–133, 1943.

MHASKAR, H. N. y POGGIO, T. Deep vs. shallow networks: An approximation theory perspective. *Analysis and Applications*, vol. 14, páginas 829–848, 2016. ISSN 0219-5305.

MILDENHALL, B., SRINIVASAN, P. P., TANCIK, M., BARRON, J. T., RAMAMOORTHI, R. y NG, R. Nerf. *Communications of the ACM*, vol. 65, páginas 99–106, 2022. ISSN 0001-0782.

MINSKY, M. y PAPERT, S. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, 1969.

MITCHELL, M. *Artificial Intelligence: A Guide for Thinking Humans*. Pelican, 2019.

MOLINA-CABELLO, M. A. Segmentación y detección de objetos en imágenes y vídeo mediante inteligencia computacional. 2018.

MOLINA-CABELLO, M. A., GARCÍA-GONZÁLEZ, J., LUQUE-BAENA, R. M. y LÓPEZ-RUBIO, E. The effect of downsampling-upsampling strategy on foreground detection algorithms. *Artificial Intelligence Review*, vol. 53, páginas 4935–4965, 2020. ISSN 15737462.

PANG, G., SHEN, C., CAO, L. y HENGEL, A. V. D. Deep learning for anomaly detection. *ACM Computing Surveys*, vol. 54, páginas 1–38, 2022. ISSN 0360-0300.

REDMON, J., DIVVALA, S., GIRSHICK, R. y FARHADI, A. You only look once: Unified, real-time object detection. páginas 779–788. IEEE, 2016. ISBN 978-1-4673-8851-1.

REN, S., HE, K., GIRSHICK, R. y SUN, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, páginas 1137–1149, 2017. ISSN 0162-8828.

RONNEBERGER, O., FISCHER, P. y BROX, T. U-net: Convolutional networks for biomedical image segmentation. 2015.

ROSA, J. L. G. Biologically plausible artificial neural networks. 2013.

SALEHI, M., MIRZAEI, H., HENDRYCKS, D., LI, Y., ROHBAN, M. H. y SABOKROU, M. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. *Trans. Mach. Learn. Res.*, vol. 2022, 2021.

SCHMIDHUBER, J. Deep learning in neural networks: An overview. *Neural networks : the official journal of the International Neural Network Society*, vol. 61, páginas 85–117, 2015.

ST-CHARLES, P.-L. y BILODEAU, G.-A. Improving background subtraction using local binary similarity patterns. páginas 509–515. 2014. ISSN 1550-5790.

ST-CHARLES, P.-L., BILODEAU, G.-A. y BERGEVIN, R. Flexible background subtraction with self-balanced local sensitivity. páginas 414–419. IEEE, 2014. ISBN 978-1-4799-4308-1.

ST-CHARLES, P.-L., BILODEAU, G.-A. y BERGEVIN, R. Subsense: A universal change detection method with local adaptive sensitivity. *IEEE Transactions on Image Processing*, vol. 24, páginas 359–373, 2015.

ST-CHARLES, P.-L., BILODEAU, G.-A. y BERGEVIN, R. Universal background subtraction using word consensus models. *IEEE Transactions on Image Processing*, vol. 25, páginas 4768–4781, 2016. ISSN 1057-7149.

STAUFFER, C. y GRIMSON, W. Adaptive background mixture models for real-time tracking. páginas 246–252. IEEE Comput. Soc, 1999. ISBN 0-7695-0149-4.

TORRALBA, A., FERGUS, R. y FREEMAN, W. T. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, páginas 1958–1970, 2008.

TRABUCCO, B., DOHERTY, K., GURINAS, M. y SALAKHUTDINOV, R. Effective data augmentation with diffusion models. 2023.

WREN, C., AZARBAYEJANI, A., DARRELL, T. y PENTLAND, A. Pfinder: real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, páginas 780–785, 1997. ISSN 01628828.

ZHANG, Y., LI, X., ZHANG, Z., WU, F. y ZHAO, L. Deep learning driven blockwise moving object detection with binary scene modeling. *Neurocomputing*, vol. 168, páginas 454 – 463, 2015.

Zhang, Y., Ling, H., Gao, J., Yin, K., Lafleche, J.-F., Barriuso, A., Torralba, A. y Fidler, S. Datasetgan: Efficient labeled data factory with minimal human effort. En *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 10145–10155. 2021.

Zheng, W., Wang, K. y Wang, F.-Y. A novel background subtraction algorithm based on parallel vision and bayesian gans. *Neurocomputing*, vol. 394, páginas 178–200, 2020. ISSN 09252312.

Zivkovic, Z. Improved adaptive gaussian mixture model for background subtraction. vol. 2, páginas 28–31. 2004.

Any technology distinguishable from magic is insufficiently advanced.

Gehm's corollary, Barry Gehm.