Automated Image Captioning: Exploring the Potential of Microsoft Computer Vision for English and Spanish

Martinez Gutierrez, Maria Fernanda

UNIVERSITY OF GENEVA

MASTER THESIS

Automated Image Captioning: Exploring the Potential of Microsoft Computer Vision for English and Spanish

---

Author: María Fernanda Martínez Gutiérrez

Project Advisor: Max De Wilde

A thesis submitted in fulfillment of the requirements for the degree of Master in Multilingual Communication Technology

in the

Faculty of Translation and Interpreting
Department of Translation Technology

December 19, 2019

Declaration of Authorship

I, María Fernanda MARTÍNEZ GUTIÉRREZ, declare that this thesis titled, "Automated Image Captioning: Exploring the Potential of Microsoft Computer Vision for English and Spanish" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:
María Fernanda Martínez Gutiérrez

Date:
18-11-2019

"Aim for simplicity in Data Science. Real creativity won't make things more complex. Instead, it will simplify them."

Damian Duffy Mingle

UNIVERSITY OF GENEVA

Abstract

Faculty of Translation and Interpreting

Department of Translation Technology

Master in Multilingual Communication Technology

Automated Image Captioning: Exploring the Potential of Microsoft Computer Vision for English and Spanish

by María Fernanda MARTÍNEZ GUTIÉRREZ

With the rise of deep learning, reflected by the creation of architectures such as Convolutional Neural Networks (CNNs), researchers are becoming increasingly interested in the utility and relevance of machines that can properly generate information about images in different languages. This thesis focuses on Microsoft Azure's Computer Vision API, specifically its functionality for image description in both English and Spanish applied to a corpus of flora pictures. To assess the accuracy of the API's captions, a combination of human and machine evaluation was used. Although the initial hypothesis was that the CNNs of the API were robust enough to generate pertinent captions in both languages, the evaluations seemed to indicate that the technology is not yet mature enough to accomplish this task. This exploratory study therefore serves as a reflection on the use of automated image captioning for multilingual purposes and of the potential and limits of this technology.

Acknowledgements

First and foremost, I would like to thank my project advisor Max De Wilde for his support and encouraging words throughout the process of writing this research. You were present even through the distance. Thank you for your patience, time and willingness to help.

I am grateful to Pierrette Bouillon, the reader, and jury of this project, whose perspectives and lessons during the course of the Master were quite insightful. I am grateful for your flexibility when choosing the date of the defense and for your kindness.

Furthermore, the lessons taught by Violeta Seretan in her *Seminaire de Recherche* class were key for choosing the metrics for this project. I am grateful to her because I learned so much from her about statistics and how to justify qualitative data in a research project.

I would like to give special thanks to my statistical friend Paula Torres for revising all my calculations and giving me advice on how to interpret the results. I am extremely thankful for the time she dedicated to helping me even in the midst of her very busy life.

To my mom, I owe great gratitude for teaching me the value of education and of following my dreams, and for her support and unconditional love that taught me to be strong and driven.

Finally, to all the friends I made in Geneva who were there for me when I needed it the most. Thank you to Nathalia, for involving me in all your projects; to Daphne, for being my proofreader and friend; to Siofra, for offering me shelter when I had no place to go; to Fleur, for believing in me and my work; and to Alex and Myriam, for all your love and support.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	**A**rtificial **I**ntelligence
ANN	**A**rtificial **N**eural **N**etwork
API	**A**pplication **P**rogramming **I**nterface
BLEU	**B**ilingual **E**valuation **U**nderstudy
CNN	**C**onvolutional **N**eural **N**etwork
DL	**D**eep **L**earning
GRU	**G**ated **R**ecurrent **U**nits
LSTM	**L**ong **S**hort **T**erm **M**emory
MB	**M**ega **B**ytes
ML	**M**achine **L**earning
MT	**M**achine **T**ranslation
NLP	**N**atural **L**anguage **P**rocessing
ReLU	**R**ectified **L**inear **U**nits
UI	**U**ser **I**nterface

Chapter 1

Introduction

Humans have been exploring the idea of creating a machine that can think since Antiquity. "Today, artificial intelligence (AI) is a thriving field with many practical applications and active research topics" (Bengio, Courville, and Goodfellow, 2016). Within the Machine Learning (ML) branch of AI, the Computer Vision application represents a quantum leap forward, especially in the development of Automated Image Captioning.

Architectures related to the Computer Vision field of study, such as Artificial Neural Networks (ANN), and most notably Convolutional Neural Networks (CNN) have allowed researchers to classify representations from large collections of data. "These techniques are better known under the umbrella term deep learning and have achieved a breakthrough in performance in a wide range of image analysis applications such as image classification, segmentation, and annotation" (Boscaini, 2017).

There are various image captioning systems currently on the market that seek to offer solutions for companies or developers in the Cognitive Services community. Nevertheless, most of these systems only describe generic visual content without identifying key entities (Tran et al., 2016). In other words, the systems manage to identify certain elements of a picture but are unable to recognize more specific content such as landmarks, public personalities or other aspects that would most likely add another dimension to the description of the image.

A quite popular system in the industry of Machine Learning has been the Microsoft Azure cloud which advertises an application called *Computer Vision API* that is said to be capable of reading text within an image, among

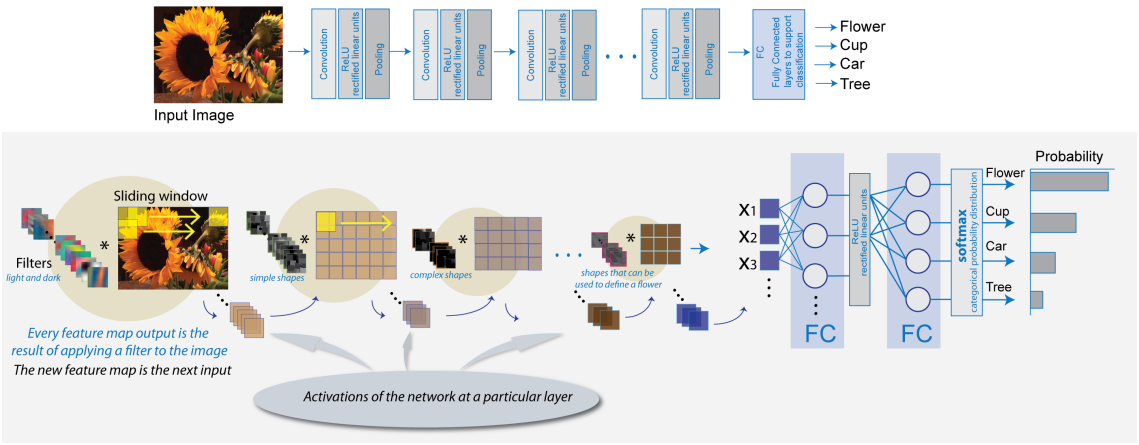

FIGURE 1.1: CNN processing of a sunflower's picture. **Source:** (MathWorks, 2018)

other functionalities. According to the Microsoft Azure web site[1], this product is supposed to "extract rich information from images to categorize and process visual data -and perform machine-assisted moderation of images" (Microsoft, 2018). This caught my attention since I was interested in classifying a set of images containing different types of flora with their respective terminology in both English and Spanish.

In essence, this research project examines whether CNNs present in an image captioning system such as Microsoft Azure's *Computer Vision API* is capable of adapting a given corpus of images from one language to another or whether the system needs to be trained with labeled data in order to obtain the desired results. I chose to focus on the adaptation of a corpus from English to Spanish due to the fact that the program's default language is English and my mother tongue is Spanish.

This thesis is divided into five chapters. The first chapter comprises a brief introduction to the subject of Machine Learning and Computer Vision. In Section 1.1 I state the research objectives of this project. In Section 1.2, I discuss my personal motivation for working in this field of studies and using a corpus of flora images. Subsequently, in Section 1.3 I address the principal and secondary research questions, before finally discussing in Section 1.4 the null and alternative hypothesis generated for this research project.

[1]For further information about Microsoft Azure and its Computer Vision API product, please visit this link: https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/

Chapter 2 focuses on the literary review carried out to guide the reader through this topic, including a description of each aspect related to this investigation. Chapter 3 describes the research methodology, paying particular attention to the assessment tools, which included both human and machine evaluation. Chapter 4 is dedicated to the results achieved. Finally, Chapter 5 is composed of reflections on further considerations as well as on the project's limitations, advantages, disadvantages and conclusions.

1.1 Research Objectives

This section presents the main objective and the several secondary objectives of this project. In this section, I cover all of them to give the reader insight into what this research seeks to analyze.

1.1.1 Main Objective

The main objective of this research is to evaluate whether a system such as the *Computer Vision API* presented by Microsoft Azure has sufficiently robust CNNs to transfer entities from one language to another without having to label each of the entities into the target language.

1.1.2 Secondary Objectives

The study's secondary objectives include:

- To evaluate how precisely the API tags and describes the images in both languages
- To examine if the system can correctly read a specialized corpus of images such as a corpus built on flora pictures
- To consider the limitations, advantages and disadvantages of using this system.

1.2 Personal Motivation

There are two aspects that I would like to address when referring to my motivation. For this reason, I have divided this section into two parts. In the first part, I discuss my motivation regarding the choice of the field of study. In the second part, I discuss the reasons behind choosing a corpus based on flora images.

1.2.1 For Automated Image Captioning

Automated Image Captioning is a topic that has gained relevance in the academic community in the past few years. Its appeal lies in the fact that it is helpful and has multiple applications for everyday tasks, for instance, "semantic image search, bringing visual intelligence to chatbots, or helping visually-impaired people to see the world around them" (Tran et al., 2016).

I therefore believe that research on this topic could be useful, particularly as regards the application of machine learning to languages, since it could be beneficial for translators and professionals in this area. Moreover, this research would complement my master's studies and would help me to better understand the market I am facing as a graduate. Additionally, I would be gaining experience in the Computer Vision field, which is on the rise.

1.2.2 For a Botanical Corpus

I would further like to mention that Automated Image Captioning applied to translating a botanical corpus from one language to another could also be practical since, as stated by Mäder and Wälden (2017), species knowledge is essential for protecting biodiversity, and since the identification of plants through conventional means is complex and time-consuming due to the use of specific botanical terms that can be frustrating for non-experts.

1.3 Research Questions

In this section, I state the two main research questions of this investigation, as well as the secondary questions that emerge from them.

1.3.1 Main Questions

This project seeks to answer the following two questions:

Does the system *Computer Vision API* proposed by Microsoft Azure have sufficiently robust CNNs in order to adapt the data labels of a specialized corpus from English to Spanish? And if so, does the API provide pertinent captions in both languages?

1.3.2 Secondary Questions

To further explore the implications of this question, it is necessary to pose other questions such as:

- What are the advantages and disadvantages of using the *Computer Vision API* proposed by Microsoft Azure?
- Is it necessary to train this system from scratch using labeled Spanish data in order to obtain a desirable translation?

1.4 Hypotheses

In this section, I address both the null and alternative hypothesis for this research project to further rule out chance (sampling error) as a plausible explanation for the results of this investigation.

1.4.1 Null Hypothesis

Hypothesis H_0 (Null hypothesis)**:** In this scenario, the sample of appraisers is not able to reach an agreement or their agreement is too weak to prove any statistical significance. If the results of the calculations for the Fleiss' Kappa, Kendall's Coefficient and BLEU metrics are lower than or equal to 0.50, it is proven beyond a reasonable doubt that the null hypothesis cannot be rejected.

In addition, should the scores given by the appraisers be less or equal to 50%, then the null hypothesis cannot be rejected either.

1.4.2 Alternative Hypothesis

Hypothesis H_1 (Alternative hypothesis)**:** In an alternative case, the expected results for this investigation are that the system will be able to transfer the data from one language to another, properly identify the elements in the images in both languages and generate satisfactory captions, though not perfect ones. Hence, my hypothesis is that the performance of API will receive a majority of positive ratings and get an approval rate of more than 50% from the appraisers.

This hypothesis is based on the assumption that using a scalable tool such as Microsoft Azure will provide the user with various solutions for projects concerning adaptation from one language to another.

In fact, the Microsoft Corporation is known for their services in terminology and languages. An example of this is their language portal, which offers language-related solutions including various style guides, terminology and UI translations[2].

To sum up the previous formulations, the following table shows the expected results for both the null hypothesis and the alternative hypothesis:

	Kappa & Kendall	**Percentages**	**BLEU**
$\mathbf{H_0}$	≤ 0.50	$\leq 50\%$	≤ 0.50
$\mathbf{H_1}$	>0.50	>50%	>0.50

TABLE 1.1: Null and Alternative Hypotheses

This concludes the introduction. In the following chapter, I will address the current state of research on Computer Vision and applications of this technology.

[2]For further information about Microsoft's Language Portal, please visit this link: https://www.microsoft.com/en-us/language

Chapter 2

Literature Review

> "Are there imaginable digital computers which would do well in the imitation game?"
>
> *Alan Turing*

Is it even plausible that a machine could be capable of *learning* from previous *experience*, or at least pretending to do so? Or developing a sense of what *it* was supposed to do in certain situations, just like a human would? Machine learning might be a considerable step forward in our time.

In this chapter, in Section 2.1, I review the current state and importance of Machine Learning by analyzing its etymology and development in past years. Later, in Section 2.2 I talk about the applications of machine learning with a specific focus on Computer Vision, and trace a brief chronology of the field to better understand the processes that the discipline went through.

Next, in Section 2.3 I address Machine Learning Methods highlighting Deep Learning and its architectures such as Artificial Neural Networks (ANN) and Convolutional Neural Networks (CNN). Furthermore, in Section 2.4 I delve into various types of Deep Learning such as supervised learning, semi-supervised learning and unsupervised learning.

In Section 2.5 I talk about different machine learning service platforms before focusing on the platform I chose for this project, Microsoft Azure. Finally, in Section 2.6 I briefly discuss the state of the art of Automated Plant Species Identification as well as current apps in the market created for this purpose.

2.1 Machine Learning

Machine learning (ML) is closely intertwined with the discipline of artificial intelligence. As such, it attempts through statistical techniques to give computer systems the ability to *learn*, for instance, by "progressively improving their performance on a specific task from a given data without being explicitly programmed" to do so (Korza et al., 1996).

In other words, what makes machine-learning systems one of a kind is that their behavior is not determined by an explicit programming process. Instead of using a specific and unambiguous set of rules to describe a program's possible behaviors, "a machine learning system seeks for patterns within a set of example behaviors in order to produce an approximate representation of the rules themselves" (Hebron, 2016). One could even go so far as to say that the system learns from *experience*.

Actually, this process is fairly similar to our own mental processes for learning about the world around us. "Long before we encounter any formal description of the "laws" of physics, we learn to operate within them by observing the outcomes of our interactions with the physical world" (Hebron, 2016).

The term *machine learning* was conceived in 1959 by Arthur Samuel. In fact, the beginning of the field of machine learning dates at least to the middle of the last century. Nonetheless, "it was only in the early 1990s that the field began to have widespread practical impact" (Bishop, 2013). During the last decade, we have witnessed tremendous growth in the number of successful applications, in areas ranging from web search to autonomous vehicles, and even to medical imaging and speech recognition (Bishop, 2013).

Indeed, inexpensive computers have been made available at the same time as there has been a rapid development in improved machine-learning algorithms that has led to a greater interest from both researchers and the commercial sector. This has led most notably to "the *data deluge* characterized by an exponentially increasing quantity of data being gathered and stored on the world's computers" (Bishop, 2013).

A considerable amount of machine learning techniques have already been developed, including logistic regression, neural networks, decision trees, support vector machines and Kaiman filters, to name but a few (Bishop, 2013). Further contributions to this "multi-disciplinary effort have come from the

fields of statistics, artificial intelligence, optimization, signal processing, speech, vision and control theory, as well as from the machine learning community itself" (Bishop, 2013).

Machine learning systems often require a very large number of examples to produce a strong intuition for the behaviors of a complex system (Hebron, 2016). For instance, if one would like to work with an automated captioning system, one would have to train it with thousands of millions of images in a certain category so the system would *learn*, for example, that an apple is an apple regardless of its position in the frame.

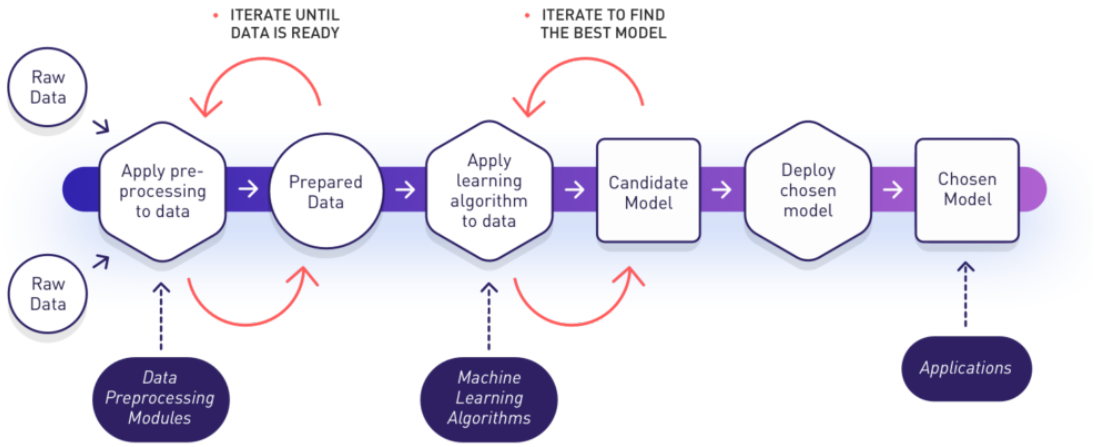

FIGURE 2.1: What a Machine Learning Project usually embodies. **Source:** (Central, 2018)

2.2 Applications of Machine Learning

Machine learning is employed in a wide range of computing tasks where designing and programming explicit algorithms with good performance would be a complex task (DeepAI, 2018b). To better understand ML, one could look at some examples of its applications, which include email filtering, detection of network intruders, Natural Language Processing, Computational Linguistics, Automated Theorem Proving, Automated Medical Diagnosis, Affective Computing and Computer Vision.

This investigation focuses on the machine-learning application of Computer Vision to further understand how Automated Image Captioning systems work in a multilingual environment.

2.2.1 Computer Vision

Computer vision is a term that covers a wide range of aspects related to the analysis of visual input by computers: "it includes reliable industrial systems, academic research systems, and theoretical studies. Computers have provided fresh metaphors and models that are of increasing influence in the cognitive sciences such as neuroscience, philosophy, and linguistics" (Brown, 1984).

In principle, it spans all tasks carried out by biological vision systems, "including *seeing* or sensing a visual stimulus, understanding what is being seen, and extracting complex information into a form that can be used in other processes" (DeepAI, 2018b) such as captioning the elements present in an image. What seems like an innate task for us requires a great deal of preparation for a computer-vision system. This preparation usually involves training the system with a corpora of images by categorizing and labeling the elements in the images and *teaching* the system what they are.

Chronology of the field

To better understand the historical context of Computer Vision practices, I review below the chronology of notable developments in this field of study[1]:

- **1955-1970**: The earliest work in the field related to the analysis of single images of static scenes. Research focused on two-dimensional (2D) images such as documents, micrographs, and images of the earth's surface taken from high altitudes. "The interpretation of such images is usually called pattern recognition, not computer vision" (Brown, 1984).
- **1960**: Work on robot vision began during this period. Originally, it concentrated on the blocks world, meaning that the image was assumed to consist of a set of polyhedrons, i.e. three-dimensional geometrical figures with flat polygonal faces, straight edges and sharp corners or vertices.

  Due to the fact that the image is a perspective projection of the scene, "geometric analysis yields useful relationships between parameters of the block edges in the image and the 3D structure of the blocks. Such

[1] I followed the structure of Brown (1984) in order to create this chronology.

relationships can be used to recover geometric properties, such as the concavity or convexity of an edge" (Aloimonos and Rosenfeld, 1991).

In addition to this, a great amount of effort was devoted to what is called low-level processing of images[2], mostly focusing on the extraction of relevant intensity changes (edges) in an image. For polyhedrons, these changes "should correspond to depth and slope discontinuities or to shadow boundaries. Edge detection was usually achieved by convolving images with local operators and thresholding the results" (Brown, 1984).

"Finding homogeneous or smooth regions, which is essentially complementary to edge finding, was thought to have the potential of isolating image regions that were the images of surface patches with some physical significance" (Aloimonos and Rosenfeld, 1991). However, it was quickly realized that the physically significant parts of images could not be identified solely by examining the gray-level intensities in the picture. By the early 1970s, it had become clear that, unfortunately, low-level vision generally could not derive practical scene descriptions from a single image, because even seemingly simple problems such as edge detection were actually quite intricate (Brown, 1984).

During that time, which was generally a fruitful period for progress in the development of artificial intelligence, it was suggested that "high-level knowledge about the scene could be used in conjunction with low-level visual processing to introduce additional constraints" (Aloimonos and Rosenfeld, 1991). To experiment with such ideas, "full" vision systems were built (Tsotsos, 1987) that used information at all levels, including both general knowledge about the imaging process as well as domain-specific information. Nevertheless, the performance of these systems was not impressive. Many researchers therefore abandoned "the system building approach and concentrated on the study of specific visual abilities, possibly corresponding to identifiable modules in the human visual system" (Brown, 1984).

- **1970-1985**: During the 1970s the field of computer vision became more mathematically advanced (Brown, 1984). David Marr (1982) proposed a paradigm in which a vision system is conceptualized as a collection

[2]Mainly concerned with extracting descriptions from images. The analysis usually does not say anything about what the objects are in the scene, nor where the scene is relative to the observer. It only makes explicit the colors, where the edges of the object are, etc.

of individual autonomous components, or modules, each of which performs a different computational task (Marr, 1982).

The low-level modules engage directly with the image data in order to generate useful 2D descriptions, while the middle-level modules use these descriptions to perform 3D recovery and the high-level modules use the results of that recovery to reason about the world. "Low-level vision modules are devoted to extracting "simple" representations of the image intensity array that have some general physical significance" (Brown, 1984).

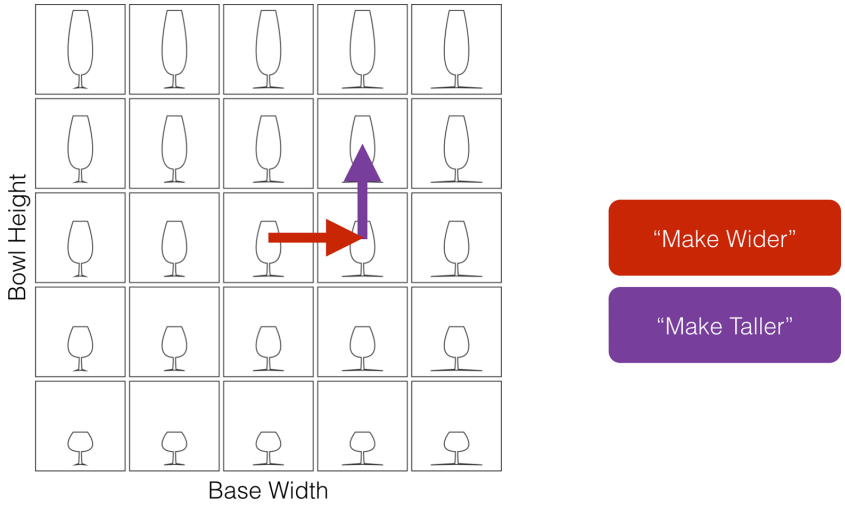

FIGURE 2.2: A sequence of low-level features related to the design of a wine glass. **Source:** (Hebron, 2016)

Tasks of particular interest at the lowest level included image restoration[3], edge detection, segmentation into homogeneous regions and texture representation (Brown, 1984). Low-level modules evaluated the image intensities without truly understanding the scene. Few trials were made to include such knowledge "in edge detection and segmentation processes by treating them as image labeling processes and making use of local consistency constraints on the labels" (Davis and Rosenfeld, 1981).

However, this procedure did not provide sufficient adaptable means of representing and integrating general knowledge. The middle-level modules used the results of the low-level modules as well as the image itself to recover the shapes, colors, spatial locations, and motions of objects in the scene (Brown, 1984).

[3]It consists in an estimation of the true intensities in a degraded image.

These modules analyzed diverse information in the image, such as contours, shading, texture and motion. During this period various mathematical techniques were developed for describing object geometry (Koenderink, 1990) and computing scene properties on the basis of various types of information present in images (Ullman, 1979).

- **1990-2000**: By the 1990s, some of the research topics of the previous decades had become more developed than others. Research in projective 3D reconstructions led to better understanding of camera calibration (Szeliski, 2010). At the same time, variations of graph cut were used to solve image segmentation. This decade also marked the first time statistical learning techniques were used to train machines to recognize faces in images (Szeliski, 2010).

  Toward the end of the 1990s, a significant change occurred that led to increased interaction between the fields of computer graphics and computer vision. This included image-based rendering, image morphing, view interpolation, panoramic image stitching and early light-field rendering (Szeliski, 2010).

"Recent work has seen the resurgence of feature-based methods, used in conjunction with machine learning techniques and complex optimization frameworks" (Sebe et al., 2005). A few examples of some of the utilities of ML techniques combined with computer vision could be found in the following categories:

- Automatic inspection (e.g.: manufacturing applications)
- Assisting humans in identification tasks (e.g.: a species identification system[4])
- Controlling processes (e.g.: an industrial robot)
- Detecting events (e.g.: visual surveillance or people counting)
- Interaction (e.g.: the input to a device for human-computer interaction)
- Modeling objects or environments (e.g.: medical image analysis or topographical modeling)
- Navigation (e.g.: by an autonomous vehicle or mobile robot)
- Organizing information (e.g.: indexing databases of images and image sequences)

[4]What I will be doing in this research.

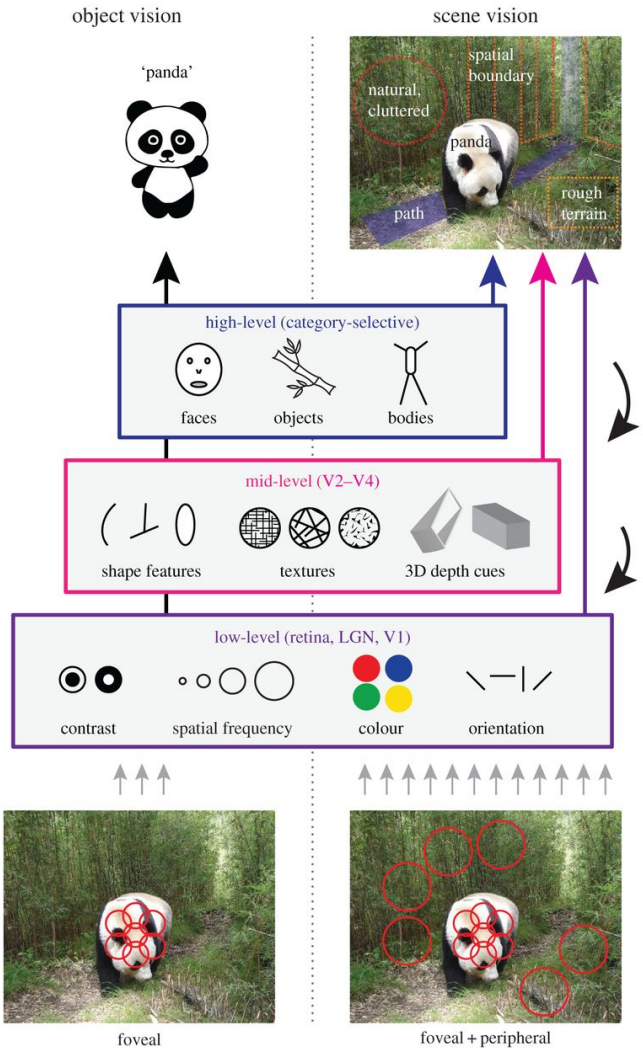

FIGURE 2.3: Hierarchical framework of visual perception. Low, medium and high-level processing of an image. **Source:** (Groen, Silson, and Baker, 2017)

In this research, the focus will be on the applications used in computer vision to develop automated image captioning systems. "Currently, the best algorithms for such tasks are based on convolutional neural networks"[5] (Russakovsky et al., 2015). An example of the capabilities of such systems is the ImageNet Large Scale Visual Recognition Challenge[6].

[5]Will be discussed in section 2.3, Machine Learning Methods.

[6]A benchmark in object classification and detection, with millions of images and hundreds of object classes. Performance of convolutional neural networks, on the ImageNet tests, is now close to that of humans.

2.3 Machine Learning Methods

Among current machine-learning methods, there is one that has enjoyed considerable success in recent years: Deep Learning. In this section I will discuss matters related to deep learning and some of its architectures, such as Artificial Neural Networks and Convolutional Neural Networks.

2.3.1 Deep Learning

Deep Learning (DL) is a machine learning technique that constructs artificial neural networks to mimic the structure and function of the human brain (DeepAI, 2018a). In theory, it allows computational models composed of multiple processing layers to learn representations of data with multiple levels of abstraction (LeCun, Bengio, and Hinton, 2015). The neural approach used in deep learning is motivated by two main ideas. One is that "the brain provides a proof by example that intelligent behavior is possible, and a conceptually straightforward path to building intelligence is to reverse engineer the computational principles behind the brain and duplicate its functionality" (Bengio, Courville, and Goodfellow, 2016).

Another is that it would be extremely interesting to understand the brain and the principles that underlie human intelligence, so ML models that shed a light on these basic scientific questions are of great utility apart from their ability to engineer applications (Bengio, Courville, and Goodfellow, 2016). Nevertheless, although neuroscience was an original core source of inspiration for DL researchers, it is no longer used as the main guide for this field since we do not currently possess enough information about our brains (Bengio, Courville, and Goodfellow, 2016). Much is still unknown about neural processes and how our brains work.

A key difference between traditional ML and DL is in how features are extracted. Traditional ML approaches use handmade features by applying several feature extraction algorithms. On the other hand, in the case of DL, the features are learned automatically and are represented hierarchically in multiple levels (Alom et al., 2017). This is the advantage of deep learning compared to traditional machine-learning approaches. The following table shows the different feature-based learning approaches with different learning steps[7].

[7] Taken from (Alom et al., 2017)

Approaches in ML	Learning Steps				
Rule Based	Input	Hand-design features	Output		
Traditional ML	Input	Hand-design features	Mapping from features	Output	
Representation Learning	Input	Features	Mapping from features	Output	
Deep Learning	Input	Simple features	Complex features	Mapping from features	Output

TABLE 2.1: Different feature ML approaches

Deep learning architectures such as deep convolutional neural networks, deep belief networks and recurrent neural networks have been applied to fields including computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics, drug design and board game programs, "where they have produced results comparable to and in some cases superior to human experts" (DeepAI, 2018b).

Challenges of Deep Learning

As mentioned by Alom et al. (2017), although DL is quite an advanced method of ML, it presents several challenges, such as:

- Multi-task and transfer learning (generalization) or multi-module learning. This means learning from different domains or with different models together[8]
- Dealing with causality in learning.
- Big data analytics using Deep Learning
- Scalability of DL approaches
- Ability to generate data which is important where data is not available for learning the system (especially for computer vision task such as inverse graphics).
- Energy efficient techniques for special purpose devices including mobile intelligence among others.

[8]What we will be doing in this research.

2.3.2 Architectures of Deep Learning

Artificial Neural Network (ANN)

One might briefly describe ANN as a major category of machine learning in which the algorithms focus on imitating biological learning systems, hence the term Artificial Neural Networks (Hebron, 2016). One could then wonder if there is a difference between DL and ANN. It is important to clarify that DL is the generalized term for training a system capable of learning and imitating human behavior, while ANNs are the architectures that help implement DL.

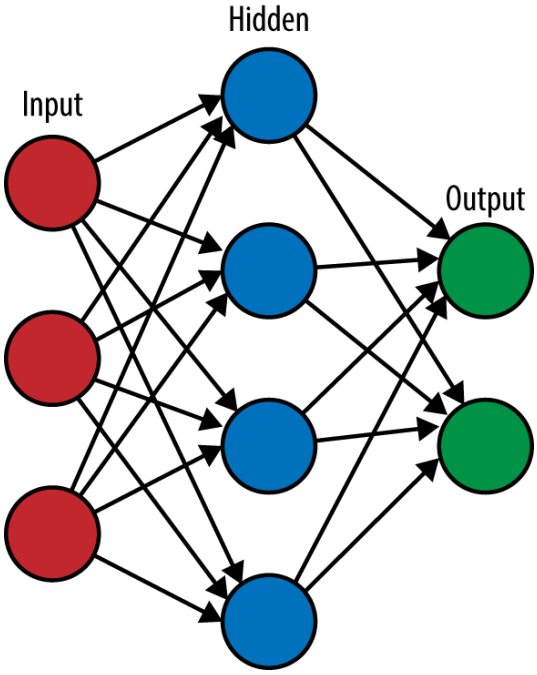

FIGURE 2.4: The simulated neurons of an ANN. **Source:** (Hebron, 2016)

As such, ANNs employ traditional computer circuitry and code to produce simplified mathematical models of neural architecture and activity[9] (Hebron, 2016). They consist of three layers (input, hidden and output), each of which is comprised of neurons/nodes that perform numerical computations and other operations. Each node in a layer is interconnected with other nodes present in consecutive layers. There are weights assigned to each interconnection and a bias assigned to each layer. These weights and biases are referred to as the parameters of the network (Sharma, 2018).

[9]Please refer to the Figure 2.7 to see some of the current Neural Networks.

There are several types of ANN architectures, including Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Units), Autoencoders, and Deep Belief Networks (Kazimipour, 2018). This thesis will focus on CNNs.

Convolutional Neural Network (CNN)

This network structure was first proposed by Fukushima in 1988 (Alom et al., 2017). "Convolutional neural networks (also known as ConvNets) are specifically suitable for images as inputs, although they are also used for other applications such as text, signals, and other continuous responses" (MathWorks, 2018).

Just like ANNs, "Convolutional neural networks are inspired from the biological structure of a visual cortex, which contains arrangements of simple and complex cells" (Hubel and Wiesel, 1959). As a whole, CNNs are neural networks with architectural constraints to reduce computational complexity and ensure translational invariance. In other words, the network interprets input patterns the same way regardless of translation. In terms of image recognition, for instance, an apple is an apple regardless of where it is in the image (Waschura, 2018).

As LeCun (2013) stated: "Convolutional Neural Networks are designed to recognize visual patterns directly from pixel images with minimal preprocessing. They can recognize patterns with extreme variability (such as handwritten characters), and with robustness to distortions and simple geometric transformations".

Specifically, a CNN has one or more layers of convolution units (fairly similar to the ones in ANNs). A convolution unit receives its input from multiple units from the previous layer, which together create a proximity[10]. Therefore, the input units share their weights (Waschura, 2018).

CNNs Features and Notions

A CNN consists of an input and an output layer, as well as multiple hidden layers. The hidden layers of a CNN typically consist of convolutional layers, pooling layers, fully connected layers and normalization layers.

[10]To use an analogy, they constitute "a small neighborhood".

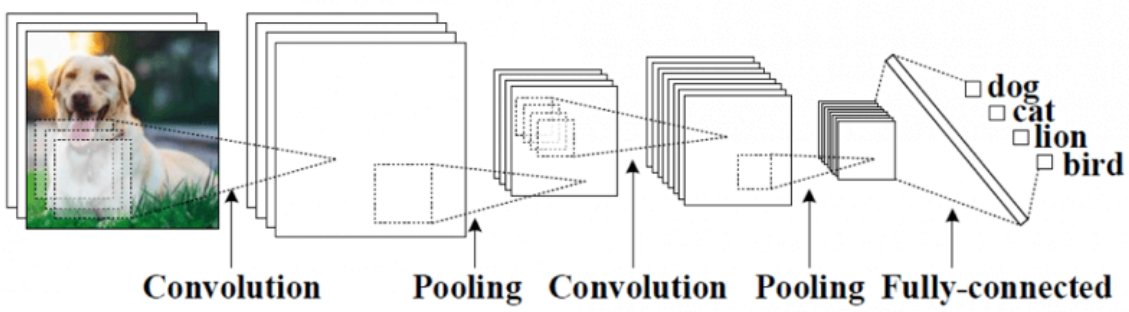

FIGURE 2.5: Visual representation of how a CNN works.
Source: (Carlsson, 2018)

Each layer has specific functions which I will briefly addressi in this section:

- **Convolutional:** "Convolutional layers parameters consist of a set of learnable filters" (Karpathy, 2018). Every filter is small in 2D space (in terms of width and height), but extends through the entire depth of the input volume. For instance, a typical filter on a first layer of a CNN might be 5x5x3 (5 pixels wide, 5 pixels high, and 3 layers deep because images have depth 3, the color channels) (Karpathy, 2018).

  During the process, each filter slides (convolves) across the width and height of the input volume and compute dot products between the entries of the filter and the input at any position. As the layers slide over the width and height of the input volume it will produce a 2D activation map "that gives the responses of that filter at every spatial position. Intuitively, the network will learn filters that activate when they see some type of visual feature" (Karpathy, 2018).

- **Pooling:** It is common to periodically insert a pooling layer between successive CNN layers. Its function is to progressively reduce the spatial size of the representation so there are fewer parameters and computation in the network, thereby controlling overfitting (Karpathy, 2018).

- **Normalization Layer:** Many types of normalization layers have been proposed for use in CNN architectures, "sometimes with the intentions of implementing inhibition schemes observed in the biological brain" (Karpathy, 2018). Nonetheless, these layers have fallen out of favor because their contribution has been shown to be minimal (Karpathy, 2018).

- **Fully connected:** Fully connected layers connect every neuron in one layer to every neuron in another layer. Neurons in a fully connected

layer have full connections to all activations in the previous layer (Karpathy, 2018).

- **Receptive field:** The input area of a neuron is called its receptive field. Hence, in a fully connected layer, the receptive field is the entire previous layer. In a convolutional layer, the receptive area is smaller than the entire previous layer.

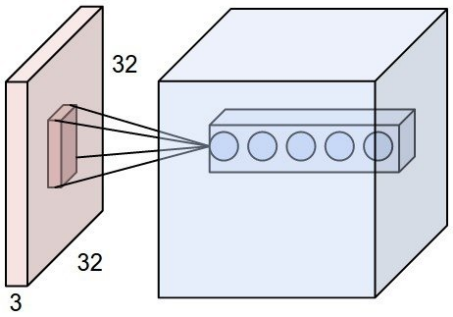

FIGURE 2.6: Visual representation of a Receptive field. **Source:** (Karpathy, 2018)

- **Weights:** Each neuron in a neural network computes an output value by applying function to the input values coming from the receptive field in the previous layer. "The function that is applied to the input values is specified by a vector of weights and a bias (typically real numbers). Learning in a neural network progresses by making incremental adjustments to the biases and weights" (LeCun, 2013).

All in all, the convolution units, as well as pooling units are especially beneficial (Kazimipour, 2018) because:

- They reduce the number of units in the network (since they consist of many-to-one mappings). In other words, there are fewer parameters to learn, which reduces the chance of overfitting, as the model is less complex than a fully connected network (Kazimipour, 2018).
- They consider the context/shared information in the form of "small neighborhoods". This is very important in many applications such as image, video, text, and speech processing/mining, since neighboring inputs (e.g. pixels, frames, words) usually carry related information (Kazimipour, 2018).

2.4 Types of Deep Learning

This section discusses the different ways a deep learning system can process feedback after data analysis is completed (DeepAI, 2018b). The terms presented in this section are mostly used to describe some of the key differences in how various models and algorithms learn and what types of information they can learn (Hebron, 2016).

Supervised Learning

This is a function that aims to map an input to an output based on example input-output pairs (Russell and Norvig, 2009). It infers a function from labeled training data consisting of a set of training examples. In a supervised learning environment, each example pair consists of an input object (typically a vector) and a desired output value (also called the supervisory signal) (Mohri, Rostamizadeh, and Talwalkar, 2012).

Semi-supervised Learning

Semi-supervised learning allows neural networks to mimic human inductive logic and sort unknown information quickly and accurately without human intervention. In other words, the data is labeled for reference and the program must infer the correct answer from the data(DeepAI, 2018b).

Unsupervised Learning

This branch of machine learning learns from test data that has not been labeled, classified or categorized. Instead of responding to feedback, unsupervised learning systems identify commonalities in the data and react to the presence or absence of such commonalities in each new piece of data (DeepAI, 2018b).

2.5 Machine Learning Service Platforms

Finally, several large technology companies and startups currently offer top-notch machine learning platforms "which provide designers with straightforward access to turnkey solutions or customized training on designer-provided data. The list of Machine Learning as a Service platforms is growing quickly" (Hebron, 2016). Some of the most popular platforms include Amazon Machine Learning, Google Prediction API, IBM Watson, Microsoft Azure, ClarifAI, and BigML.

2.5.1 Microsoft Azure

In this research I work closely with the Microsoft Azure cloud, which, according to its website, aims "to help your organization meet your business challenges. It's the freedom to build, manage, and deploy applications on a massive, global network using your favorite tools and frameworks" (Astala and Hamilton, 2017).

In fact, Microsoft Azure can perform a variety of tasks. Besides supporting both Linux and Windows, it comes with a wide selection of ready-to-run server applications and languages. For instance, "you can easily set up a website with .NET, Java, PHP, Node.js, and/or Python support using prefabricated images and configurations" (Vaughan-Nichols, 2017).

Several individuals have provided testimonials listing the advantages of Microsoft Azure's products. For example, the president of Snow Leopard Trust, Rhetic Sengupta, writes: "with Azure Machine Learning, we can automate the image classification process to feed our scientific models. The data preparation tool helps us combine years of data into an organized and complete source. This frees up our scientists to spend more time on snow leopard conservation" (Astala and Hamilton, 2017).

In this research, I analyze Microsoft Azure's *Computer Vision API* by using it to generate captions for a corpus of images made up of predominately of flora and testing the capacity of the system to transfer entities from one language to another without having to label each of the entities in the target language. The methodology I used to carry out this research will be discussed in the next chapter.

To conclude this chapter, I would like to briefly address the importance of using a flora corpus and why this topic is relevant today.

2.6 Automated Plant Species Identification

Owing to the current global warming crisis, it has become increasingly important to create a compendium of the identity and geographic distribution of plants for the purposes of future biodiversity conservation. It does not come as a surprise, then, that many botanists and scientists have dedicated great effort to this task. Ultimately, "rapid and accurate plant identification is essential for effective study and management of biodiversity" (Wäldchen and Mäder, 2017).

Due to time constraints and the overall tedium of classifying hundreds of thousands of plant species in an efficient way, there is increasing interest in automating the process of species identification. Currently, "the availability and ubiquity of relevant technologies, such as digital cameras and mobile devices, remote access to databases, new techniques in image processing and pattern recognition let the idea of automated species identification become reality" (Wäldchen and Mäder, 2017).

For a system to successfully identify plants in images,, it usually has to conduct feature extraction. As pointed out in the literature review, feature extraction is a key task in content-based image classification. It typically follows the pre-processing step in the classification process (Wäldchen and Mäder, 2017).

"The general purpose of feature extraction is to reduce the dimensionality of this information by extracting characteristic patterns" (Wäldchen and Mäder, 2017). These patterns can be found in colors, textures and shapes (Gonzalez and Woods, 2007) of the flower or leaf and play a relevant role in the proper distinction of species. In order for the machine to recognize different species, it has to be familiarized by its textures and patterns [11].

[11] It is important to note that the majority of plant classification approaches only focus on intact plant organs and are not applicable to degraded organs (e.g., deformed, partial, or overlapped) which often exist in nature (Wäldchen and Mäder, 2017).

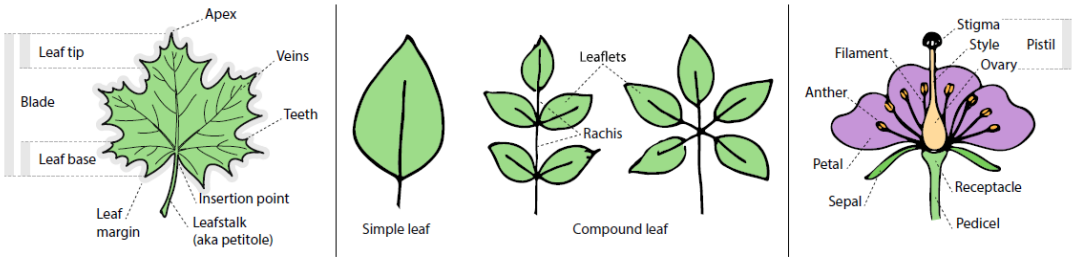

FIGURE 2.7: Leaves and flowers common features. **Source:** (Wäldchen and Mäder, 2017)

2.6.1 Examples of Plant Species Identification Apps

Some examples of botanic and image identification projects can be found in apps such as PlantNet and ArbolApp that seek to identify certain species of plants in different regions of the world. Unfortunately, these apps lack the "automated" part of the process, which is precisely what we aim to investigate in this research project: whether it is possible to train a system effectively to immediately recognize a specific plant in a given picture.

PlantNet

PlantNet is an example of a collaborative project that has been created to identify plants by taking pictures of them and sharing them with the community in order to properly identify the plant. With a database of 16.675 species of plants and over 709.411 images (PlantNet, 2019) this app has thousands of users who upload pictures of the plants they find across their regions.

In contrast to PlantNet, there are projects made exclusively by experts and that are not based on crowdsourcing. ArbolApp is one of them.

ArbolApp

ArbolApp is a guide to the wild trees that can be found in the Iberian Peninsula and the Baleares Islands. The project was developed by the Royals Botanic Garden and the Scientific Culture Area of the CSIC and contains a glossary of over 90 terms, 143 species, 500 photos and 370 illustrations (ArbolApp, 2019).

FIGURE 2.8: PlantNet allows you to take your own pictures and share them with others to identify the plants you see. **Source:** (PlantNet, 2019)

FIGURE 2.9: Screenshot of ArbolApp's features. **Source:** (ArbolApp, 2019)

Although ArbolApp is a useful compendium, the fact that only experts compile its flora corpus has a significant impact on the number of datasets available to classify. This could potentially indicate that a crowdsourcing model might be more useful. PlantNet has a large corpus, but one could certainly ask how the creators of the app can eliminate the "noise", or duplicates, when so many images are submitted every day. One could say that ArbolApp is

a curated compendium with a small corpus while PlantNet has a powerful corpus but has difficulty classifying all of its data. What if there were a system capable of classifying all the image data effectively yet quickly? Could Microsoft's API be the solution to this problem?

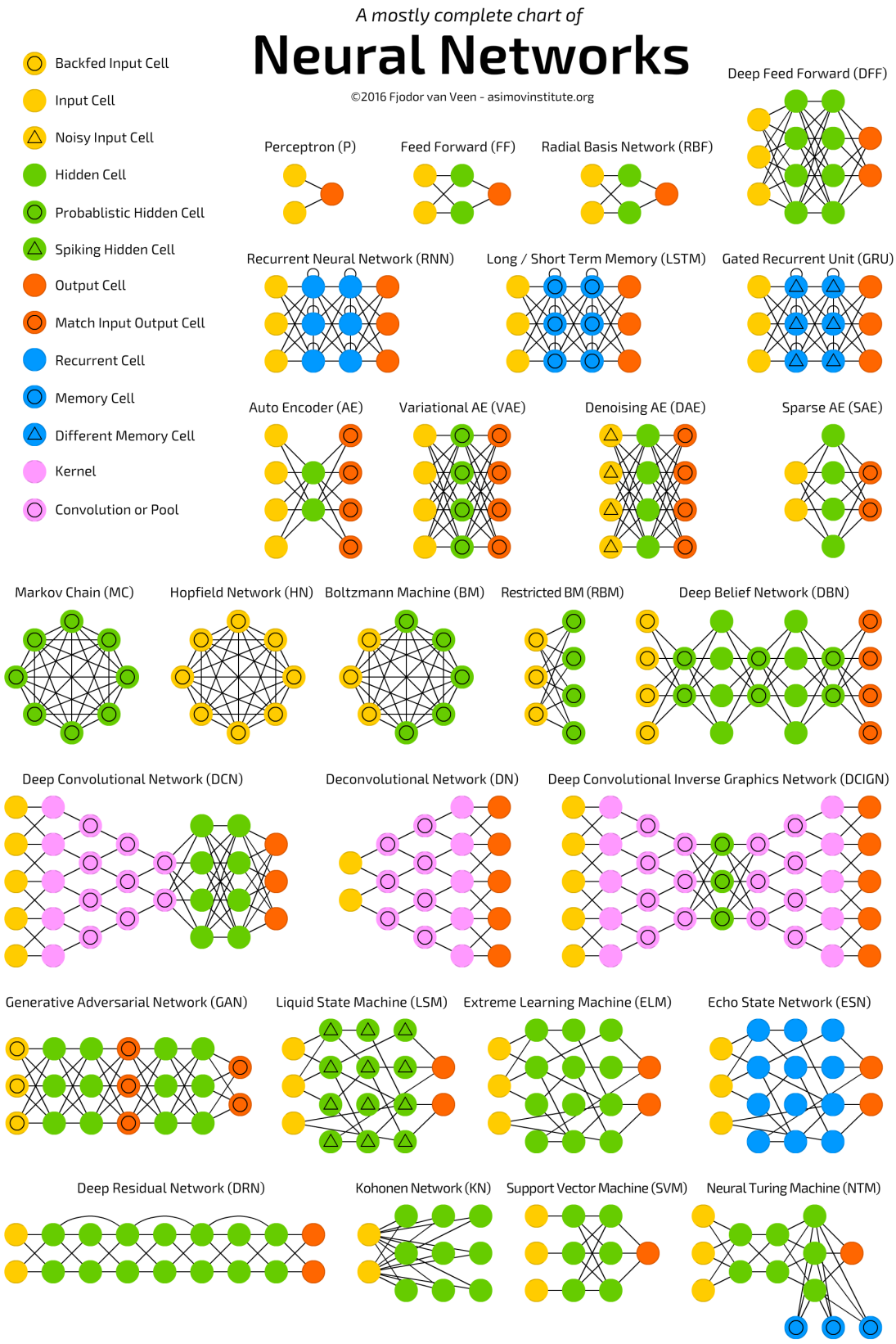

FIGURE 2.10: A mostly complete chart of Neural Networks.
Source: (Veen, 2016)

Chapter 3

Research Methodology

In this chapter, I will discuss the research methodology I used to examine to what extent Microsoft Azure's Computer Vision API captions are pertinent when applied to a corpus consisting of flora pictures. Section 3.1 will present the general methodology as well as the type of sampling used to choose the subjects.

Subsequently, in Section 3.1.1 there will be a discussion of the material (i.e. the questionnaires) created for the human participants. This section will focus on participants' guidelines and specifications (such as time constraints), the questions themselves and the randomness of the sample pictures.

I will put particular emphasis on the types of machine and human evaluation used and on the considerations that went into designing the evaluations. Consequently, there will be a description of the evaluation metrics that were used to assess the overall performance of the API in Section 3.1.3.

Lastly, I will describe how I used machine evaluation as a complement to human evaluation in Section 3.1.4, and I will explain why I decided to work with the BLEU metric instead of other metrics in Section 3.1.4.1.

Nevertheless, because the BLEU metric presents some limitations, in Section 3.1.4.2 I address possible concerns and further explain why the BLEU metric was a useful tool or an indicator for analyzing the results of this particular study.

3.1 Evaluation Metrics

This investigation uses methodological triangulation, which is a "method of cross-checking data from multiple sources to search for regularities in the

Scenario	Description
Analyze Image	Uses the Analyze Image operation to analyze a local or remote image. You can choose the visual features and language for the analysis and see both the image and the results.
Analyze Image with Domain Model	Uses the List Domain Specific Models operation to list the domain models from which you can select, and the Recognize Domain Specific Content operation to analyze a local or remote image using the selected domain model. You can also choose the language for the analysis.
Describe Image	Uses the Describe Image operation to create a human-readable description of a local or remote image. You can also choose the language for the description.
Generate Tags	Uses the Tag Image operation to tag the visual features of a local or remote image. You can also choose the language used for the tags.
Recognize Text (OCR)	Uses the OCR operation to recognize and extract printed text from an image. You can either choose the language to use, or let Computer Vision auto-detect the language.
Recognize Text V2 (English)	Uses the Recognize Text and Get Recognized Text Operation Result operations to asynchronously recognize and extract printed or handwritten text from an image.
Get Thumbnail	Uses the Get Thumbnail operation to generate a thumbnail for a local or remote image.

TABLE 3.1: The functionalities of the Computer Vision API

research data" (O'Donoghue and Punch, 2003). In particular, it refers to using a combination of several research methods to study the same phenomenon (Bogdan and Biklen, 2006).

When running the API, the user has several options, or scenarios, for what to do with the image, including "describe image", "generate tags", "recognize text" and "analyze image", as shown in Table 3.1.

In this research, I wanted to explore the *Describe Image* scenario. I therefore used two evaluations per language to assess this scenario: a human evaluation using surveys and a machine evaluation using the BLEU metric.

In other words, the feedback generated for this research project was gathered through both human and machine evaluation by means of surveys and experiments to later be assessed and cross-checked.

Finally, the participants for the surveys were chosen using quota sampling.

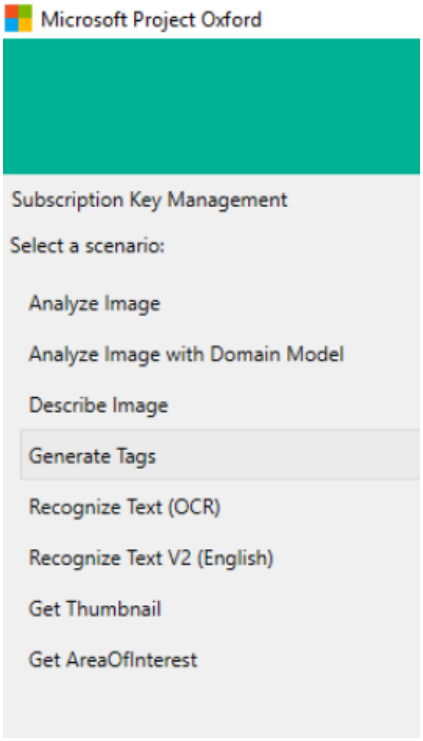

FIGURE 3.1: Computer Vision API Applications

To be more precise, I used a controlled quota sampling since the sampling group had two restrictions based on the following criteria:

- The linguistic background of the two groups had to either be English or Spanish.
- The surveys had a time limit of one week.

In total, 28 subjects for each linguistic group filled out the form in time. In the following section, I address aspects of the human evaluation process.

3.1.1 Human Evaluation

In this section, I introduce the participants that took part in this study and describe their role and requirements. I mention how the information was gathered by means of questionnaires. In addition, this section explores the purpose of the surveys and their relevance for the study.

Participants

The participants involved in this study were, as previously mentioned, a product of quota sampling. First, I separated the participants into two groups based on their linguistic background. The first group had to answer Survey A, which was composed of 10 images portraying flowers and the corresponding English captions generated by the API. Group A was given the task of rating the captions by assessing how pertinent they were in relation to the images.

Survey B used the same structure as Survey A, but was carried out in Spanish and asked the participants to rate the Spanish captions generated by the API. Group B also had to rank the captions from 1 to 5, 1 being the lower score and 5 the highest, based on how well the caption matched the image[1].

Due to the fact that the surveys had a time constraint, both Group A and Group B each comprised 28 people, as this was the number of the people who replied on time for each group.

Participants' names remained anonymous and their answers and data were used for learning purposes only.

Score		Criteria
1	Bad	The caption does not properly describe the image: when the caption describes a different scene than the one presented in the image.
2	Mediocre	There are major inconsistencies in the caption: when the caption partially describes the elements of the scene
3	Okay	The caption is good enough: when the caption manages to identify most of the key elements of the scene.
4	Good	Good caption: when the caption manages to identify all of the key elements of the scene.
5	Excellent	Perfect caption: when the caption describes every single element of the scene.

TABLE 3.2: Criteria for each score

3.1.2 From English to Spanish

The Computer Vision API "allows classifying the image content by providing a comprehensive list of tags and attempting to build a natural language description of the scene" (Bobriakov, 2018). Once you run the API, you have an interface in which you can choose an image to upload or the URL of an image you want to identify.

[1]Please refer to table 3.2 for more information about the criteria of the score.

The upper left corner of Figure 3.2 displays a language drop-down list that allows you to generate tags or descriptions in different languages. At the time this study was conducted, the API supported the following languages: English, Spanish, Japanese, Portuguese and Chinese.

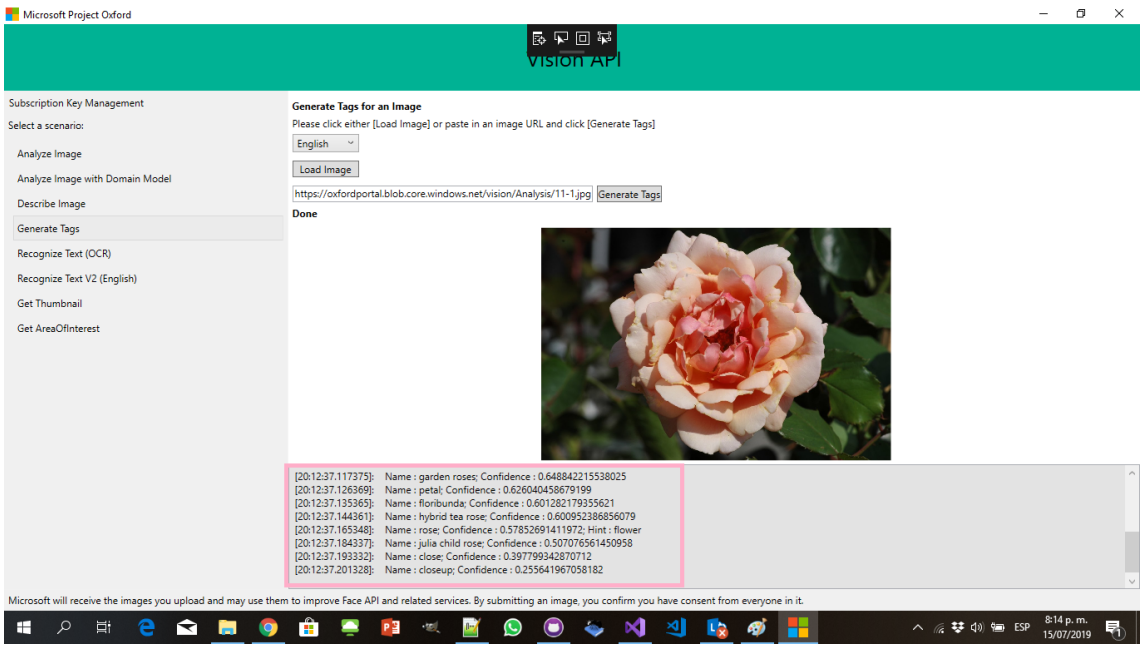

FIGURE 3.2: Languages and Tag function in the Computer Vision API

I chose ten different sets of flora images to be analyzed by the system in both English and Spanish using the "describe image" functionality. With this functionality, the API generates several captions followed by a score indicating "degree of confidence".

For this experiment, I always chose the first option, or in other words, the option that had the highest confidence score (between 50% and 99%) as shown in Table 3.3. Afterwards, I created the two surveys to assess the pertinence of the captions with regards to the image in both Spanish and English.

Questionnaires

The surveys were created via Google forms and consisted of 10 images with captions either in English (Survey A) or in Spanish (Survey B) to be rated. The link to the questionnaire was sent to the participants and they were given a period of one week to fill out the form.

Degree of Confidence			
Captions in English		**Captions in Spanish**	
A close up of a red flower hanging from a tree.	0.95	Una flor rosa en una rama.	0.69
A red flower in a field.	0.96	Flor de color rojo en el pasto.	0.64
A bunch of purple flowers.	0.97	Planta con flores moradas.	0.70
A tree in the middle of a lush green field.	0.93	Un árbol en un campo.	0.55
A close up of a flower.	0.99	Planta con flores rosas.	0.73
A pink flower on a plant.	0.91	Una flor rosa.	0.70
A white flower on a plant.	0.95	Flor amarilla en el pasto.	0.69
A yellow flower in the grass.	0.98	Flor de color amarillo en el pasto.	0.68
A close up of a flower.	0.96	Planta con flores amarillas.	0.71
A close up of a rose.	0.95	Una flor rosa.	0.71

TABLE 3.3: Degrees of Confidence of the captions per Language

Annotation guidelines were designed to accompany the surveys and provide instructions to the participants [2]. The English version of this document can be found in Appendix A and the Spanish version can be found in Appendix B.

To ensure the clean output of the captions, I used pictures I had taken myself that had never been posted online and thus had never been seen before.

After filling out the questionnaire, participants in both surveys were presented with the following disclaimer text concerning the authorization of personal data treatment: "I hereby authorize the processing of my personal data solely for educational purposes". The participants could then choose whether or not to give their consent.

Image-based corpus

The images used for the questionnaires comprised ten different types of flora commonly found in Switzerland. The pictures chosen for the survey were taken by myself and were never used to train the API, so both the machine and the human appraisers were seeing these images for the first time during this study.

All the images used were in JPEG format and their size was under 4 Megabytes. For this experiment, I did not take into account the specific name of the flowers, since it proved to be extremely difficult to train the system with specific flora terminology. Hence, I focused on determining whether the captions were pertinent and matched the scene rather than on the terminology of the plants.

[2]The guidelines shown in the appendices are detailed transcriptions of the instructions given orally.

Survey A - English

The survey comprised ten questions and ten images. Each question related to the caption generated by the system for one image. In each case, the appraiser was asked to match on a Likert scale the degree of accuracy of the caption with regard to the images.

Figure 3.3 shows part of the questionnaire in which the participant is asked to rate the caption given by the API from 1 to 5, with 5 being a perfect match and 1 being a complete mismatch.

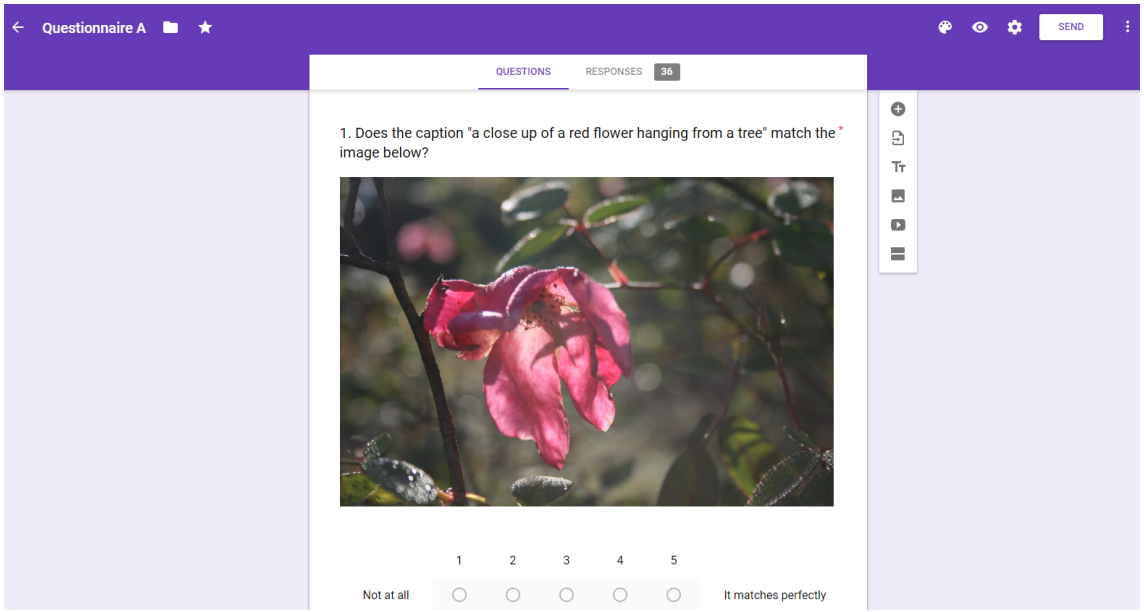

FIGURE 3.3: Survey A

Survey B - Spanish

In the same way, survey B was designed to review and rate the captions made by the machine. The participant saw the same images of flora presented in Survey A but with Spanish captions. Figure 3.4 depicts a snippet of the questionnaire.

The criteria taken into account to rate the captions was the same used in survey A and explained in Table 3.2.

Since both surveys were used to measure preference or opinion, and since data was ordinal, I decided to use non-parametric statistics in order to study the ratings given by the subjects. The methods used to assess the human evaluation were Fleiss' Kappa and the Kendall Rank Correlation Coefficient.

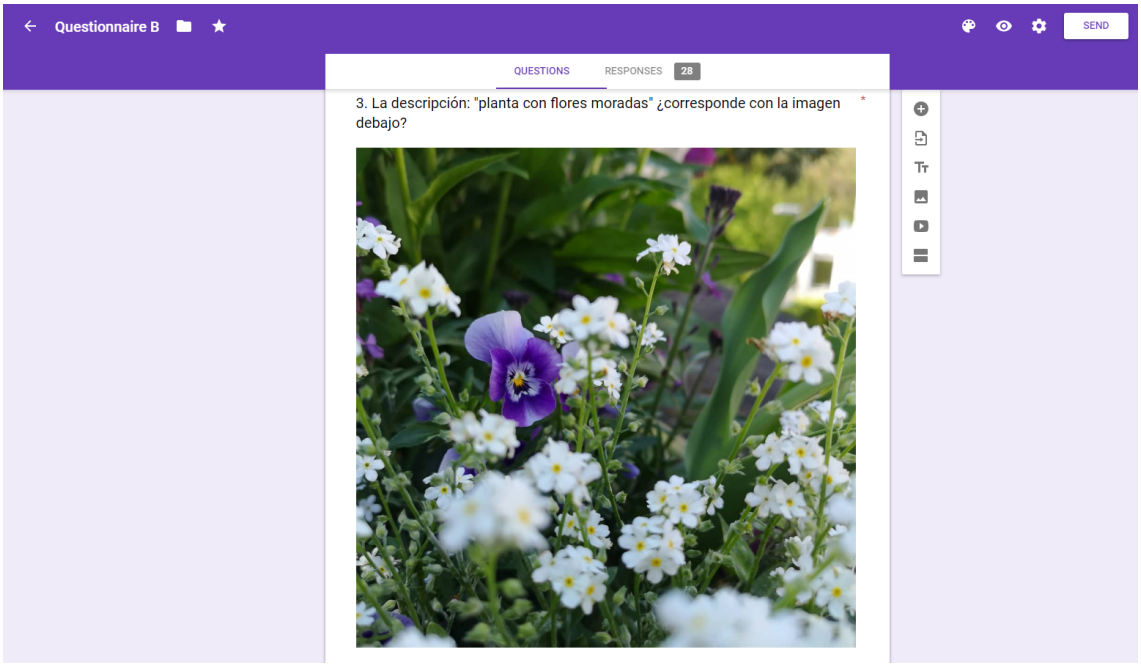

FIGURE 3.4: Survey B

3.1.3 Statistical Calculations

Below I will address the statistical calculation methods that I used for this research. I will briefly introduce them and mention why these methods are the most adequate for this study in particular. I used both Fleiss' Kappa and Kendall Rank Correlation Coefficient to analyze the data I gathered from the forms from two different perspectives.

Fleiss' Kappa

Since in both Survey A and B I used a Likert type of scale that goes from 1 to 5, and since I had different appraisers rating the captions, I used Fleiss' Kappa to measure the degree of agreement between appraisers to be able to better assess and interpret the results for each caption.

The kappa statistic measure of agreement is scaled to be 0 when the amount of agreement is what would be expected to be observed by chance and 1 when there is perfect agreement. For intermediate values, Landis and Koch (1977a, 165) suggest the interpretations presented in Table 3.4.

However, when you have ordinal ratings, "such as defect severity ratings on a scale of 1–5 like in the case of the forms I created, Kendall's coefficients,

Below 0.0	Poor
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost Perfect

TABLE 3.4: Kappa Score Agreement

which account for ordering, are usually more appropriate statistics to determine association than kappa alone" (MiniTab, 2019). Therefore, I also calculated Kendall's Coefficient.

Kendall Rank Correlation Coefficient

Kendall's coefficient of concordance indicates the degree of association of ordinal assessments made by multiple appraisers when assessing the same samples. Kendall's coefficient is commonly used in attribute agreement analysis.

Kappa statistics represent absolute agreement between ratings while Kendall's coefficients measure the associations between ratings. Therefore, kappa statistics treat all misclassifications equally, but Kendall's coefficients do not treat all misclassifications equally.

In the same way as Kappa, Kendall's coefficient is scaled from 0 to 1. "The higher the Kendall's coefficient value, the stronger the association. Usually Kendall's coefficients of 0.9 or higher are considered very good. A high or significant Kendall's coefficient means that the appraisers are applying essentially the same standard when assessing the samples" (MiniTab, 2019).

Now that all the methods related to the human evaluation section have been addressed, the next section focuses on the machine evaluation method I chose to work with.

Percentages

In order to go further and explore other aspects of the human evaluation, I decided to calculate the percentages of positive and negative scores between raters. I did so by grouping the scores from 1 to 2 as negative scores and

grouping the scores from 3 to 5 as positive scores. Although having a "positive" and "negative" grouping of the scores could be deemed subjective, I still was very eager to see what the results would be and if they represented a significant change or difference amongst the results obtained by exploring the Fleiss' Kappa and Kendall's Coefficient metric.

3.1.4 Machine Evaluation

With regards to the machine evaluation part of this research, I chose the BLEU score metric to assess the overall performance of the API and the quality of the captions, which allowed me to compare the captions generated by the machine versus the captions generated by a native speaker.

I asked a native speaker in both English and Spanish to look at the images presented in the surveys and provide a brief description of the fundamentals of the scene. Subsequently, using the Interactive BLEU score evaluator by Tilde, I submitted both the machine caption as well as the human one for comparison.

BLEU

BLEU (an abbreviation of Bilingual evaluation understudy) is a metric used to measure the quality of machine-generated text. Individual text segments are compared with a set of reference texts and scores are computed for each one of them (Hossain et al., 2018). To date, it is the most commonly used metric in the image description literature (Vinyals et al., 2015).

BLEU is popular because it was one of the first methods in automatic evaluation of machine translated text and has a reasonable correlation with human judgements of quality (Callison-Bursch, Osborne, and Koehn, 2006). In this research I used the BLEU algorithm[3] to make a cross reference between the captions given by the native speaker and the machine after processing the survey images. Nevertheless, this metric has certain limitations that are worth mentioning before proceeding to the next chapter.

[3]I used the interactive BLEU score evaluator provided by Tilde Custom Machine Translation: https://www.letsmt.eu/Bleu.aspx

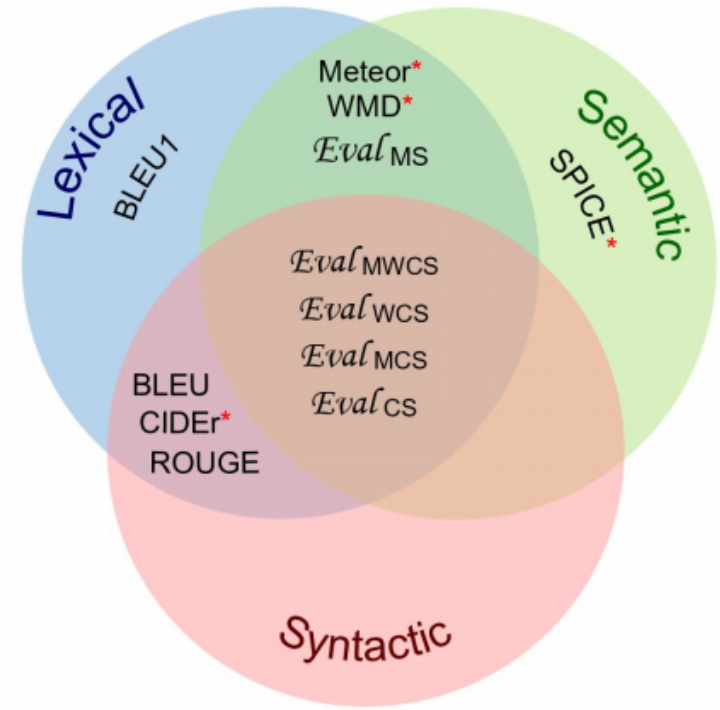

FIGURE 3.5: Diagram of different Machine Evaluation metrics

Limitations of the BLEU metric

In recent literature the BLEU metric has been said to present certain limitations. For example, its scores are good only if the generated text is short (Hossain et al., 2018) which can be a considerable issue in studies that seek to analyze Natural Language Processing (NLP). This is due to the fact that the BLEU metric does not recognize context, and is therefore unable to recognize patterns in sentences or synonyms that could potentially be useful.

Figure 3.5 shows a comparison between BLEU and other machine evaluation metrics. BLEU uses both lexical and syntactic features to assign a score. For in-depth studies, however, metrics that measure semantics are extremely practical (e.g. the SPICE metric). The fact that BLEU fails most of the time to recognize patterns or synonyms can be a significant limitation.

Nevertheless, in this research I chose to work with BLEU because I was analyzing only short captions consisting of six to eleven words. For a more accurate evaluation metric for both semantic and syntactic structures it would be recommended to use either CIDEr (Consensus-based Image Description Evaluation) or SPICE (Semantic Propositional Image Caption Evaluation) or

a combination of both (SPIDEr) as suggested by Hossain et al., 2018 so that the machine would be more sensitive to the use of synonyms and different patterns that could be formed between sentences.

Chapter 4

Results

In this chapter, there will be a brief account of the steps I followed to carry out the experiment. In section Section 4.1 I will discuss the process of installing and running the API, cloning the repository for the image dataset and other issues related to this process.

Section 4.2 presents a detailed summary of the data classification using tables and charts. It also presents the results obtained from the surveys and analyzes them in context.

Section 4.3 moves on with a discussion of the human evaluation part of the experiment, addressing the use of Fleiss' Kappa, Kendall Rank Correlation Coefficient as well as grouped percentages. The results of each metric will be commented and explained.

Section 4.4 contains an analysis of the machine evaluation part of the experiment and of the results given by the BLEU score. Finally, in Section 4.5 I offer some practical considerations before proceeding to the fifth and final chapter of this project.

4.1 Microsoft Azure Computer Vision API

Azure's Computer Vision service provides developers with access to advanced algorithms that process images and return information. One can either upload an image or specify an image URL to analyze. The image-processing algorithms can analyze the content of the picture in several different ways, depending on the visual features the user is interested in.

Before running the API, there are some compulsory requirements as well as some instructions to follow. Microsoft provides you with a set of indications

to start working with the API and also to orient the user on the folders and the code in general. In fact, an entire web page is dedicated to instructions on how to get started with the API[1]

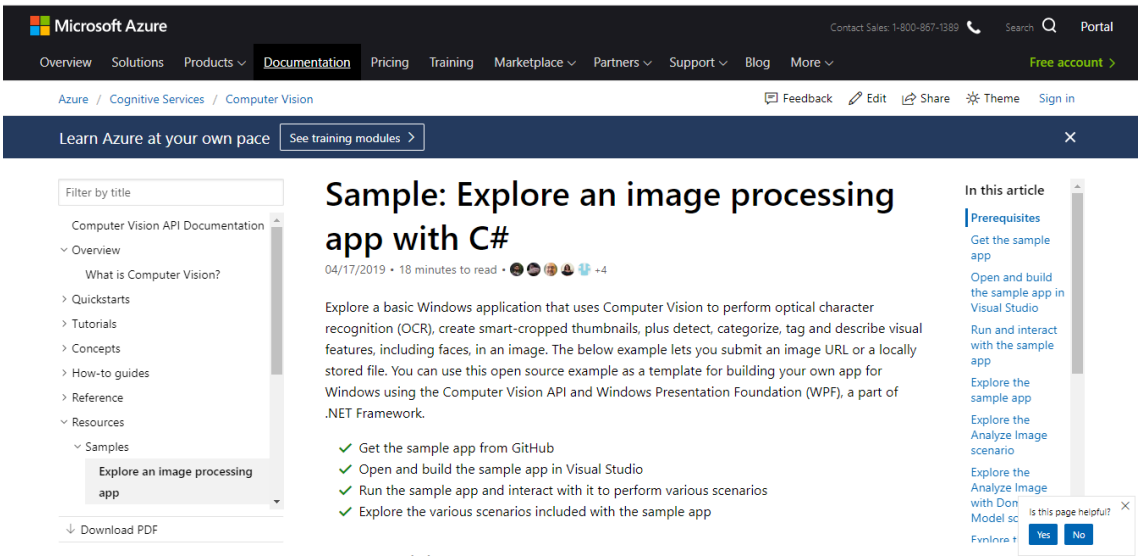

FIGURE 4.1: Instructions on how to start working with the API

In the next subsection, the requirements related to the API will be addressed.

4.1.1 Previous Requirements

There are certain requirements that need to be taken into account in order to work with the API. Some of them are more technical, while others are related to the image format, the installation of other programs or the compliance and privacy policy. The following sections will describe the process I followed before, during and after working with the API.

Technical skills and Image Parsing Datasets

It is important to have at least some basic knowledge of C# to understand the code and how to manipulate it, in case the user wants to add to the code or further train the API.

Depending on the corpus of images one wants to use to train the system, it is possible to use the images proposed by Microsoft[2] by cloning their image

[1] Please find these instructions here: https://bit.ly/2n99ej9
[2] Please note that these images are mostly used to recognize facial features

repository on GitHub, to upload one's own dataset or to use another previously existing set of images. The API uses advanced algorithms to process the image and usually, even without using the image repository of Microsoft on GitHub, you will receive descriptions for the images.

Furthermore, once the API is run, one has to use the subscription key and an endpoint to be able to deploy the app. These are given by Microsoft as soon as one starts the trial period or subscription to Azure as shown in Figure 4.2.

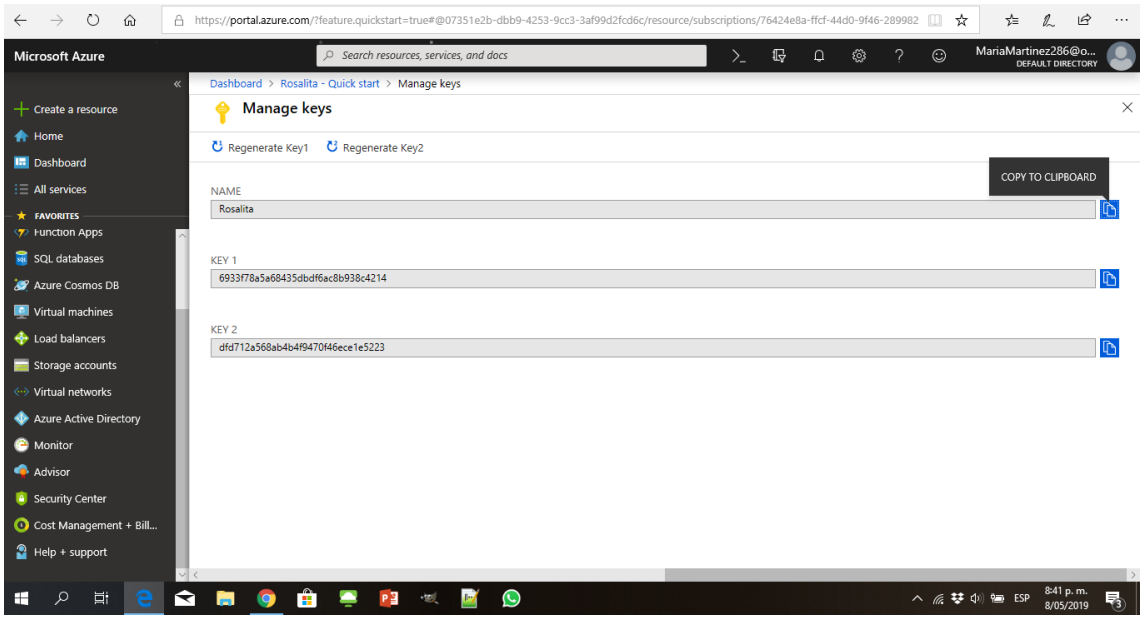

FIGURE 4.2: Subscription Key and Endpoint

Space & Software requirements

Before using the API, it should also be taken into account that one has to download Visual Studio 2015 or a later version in order to modify the code and run the app. This is relevant because the program is quite heavy, taking up between 2.3 GB and 60 GB of available hard disk space depending on installed features[3].

In Figure 4.3 you can see what the API looks like in Visual Studio. To start using the code and deploy the API, the user has to make use of GitHub and download Microsoft's repository and some sub-modules that will be used later to run the app.

[3]For further details on how to download Visual Studio 2017: https://docs.microsoft.com/en-us/visualstudio/productinfo/vs2017-system-requirements-vs

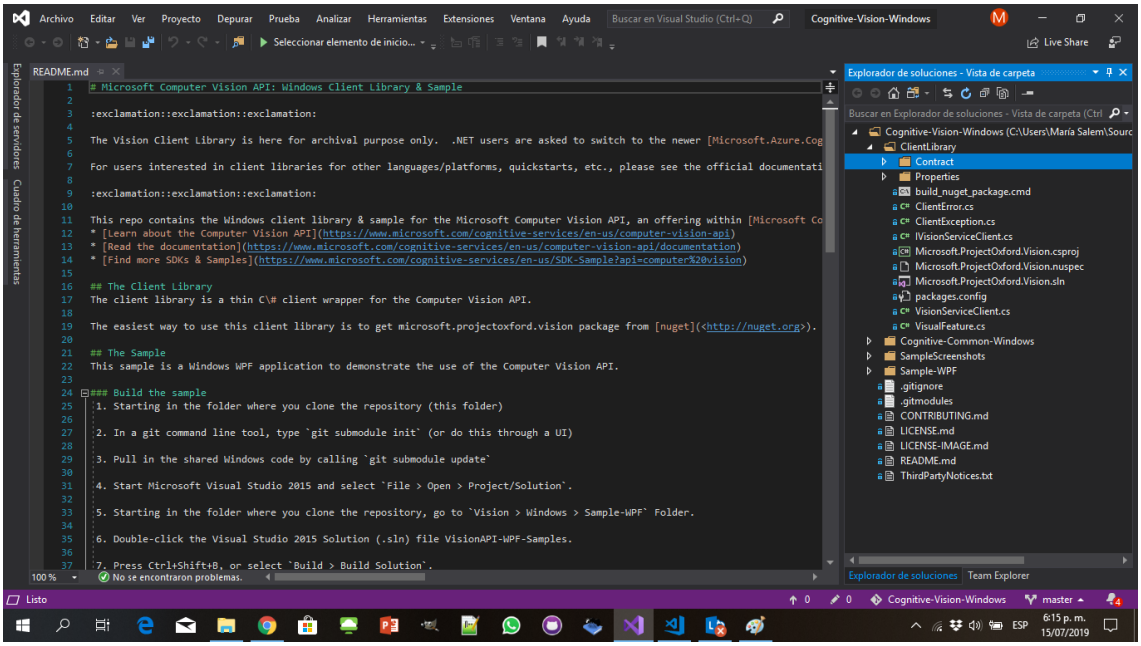

FIGURE 4.3: API as seen in Visual Studio

Image requirements

Another requirement is the criteria for the type of images you can analyze. The Computer Vision API analyzes images that meet the following criteria:

- The image must be in either JPEG, PNG, GIF, or BMP format
- The file size of the image must be less than 4 megabytes (MB)
- The dimensions of the image must be greater than 50 x 50 pixels.

Data privacy

One pertinent detail is that using the Computer Vision API constitutes implicit agreement to Microsoft's terms and conditions[4]. When you are using an image for analysis, the following notice is always displayed at the bottom of the app: "Microsoft receives the images you upload and may use them to improve Face API and related services. By submitting an image, you confirm you have consent from everyone in it."

[4]For further information on Microsoft's policies on this matter, please refer to this link: https://azure.microsoft.com/en-us/support/legal/cognitive-services-compliance-and-privacy/

Microsoft's compliance and privacy web page indicates that there is a possibility to control and delete the data after it is processed. However, I could not find a way to do so.

With this consideration, I conclude this section. In the following section I will address the ratings of captions in both English and Spanish to further examine the results of the surveys.

4.2 Result of Calculations & Criteria Weighting

This section is dedicated to the analysis of the results obtained from both the human and machine evaluations. Each subsection will discuss separately the results given by the English and Spanish appraisers and compare the scores of Group A and Group B. Both groups' responses are summarized in a table that displays the ten categories (image-caption pairs) used for the surveys and the scores given by the appraisers.

For instance, referring to Table 4.1, in category one, 5 appraisers rated the caption as bad (1), 8 as mediocre (2), 8 as okay (3), 4 as Good (4) and finally, 3 as excellent (5).

4.2.1 English Appraisers

After compiling the responses for Group A (shown in Table 4.1), I quickly noticed that the distribution of answers seemed to be very different for each image. The English captions seemed to have received predominantly low scores.

	Bad (1)	Mediocre (2)	Okay (3)	Good (4)	Excellent (5)
C1	5	8	8	4	3
C2	4	7	7	4	6
C3	13	9	5	1	0
C4	0	3	6	10	9
C5	2	3	6	10	7
C6	18	7	3	0	0
C7	5	7	4	9	3
C8	10	12	4	1	1
C9	9	9	3	4	3
C10	3	7	5	7	6

TABLE 4.1: Data gathered from Survey A - English Appraisers

In fact, on closer analysis and with the help of the bar chart in Figure 4.4, it becomes clear that the most prominent scores for the captions are either "bad" or "mediocre". However, one can also notice that in some categories the scores were more or less evenly distributed. For instance, in category 10, almost all the scores are represented in the table and there seems to be no clear agreement between appraisers.

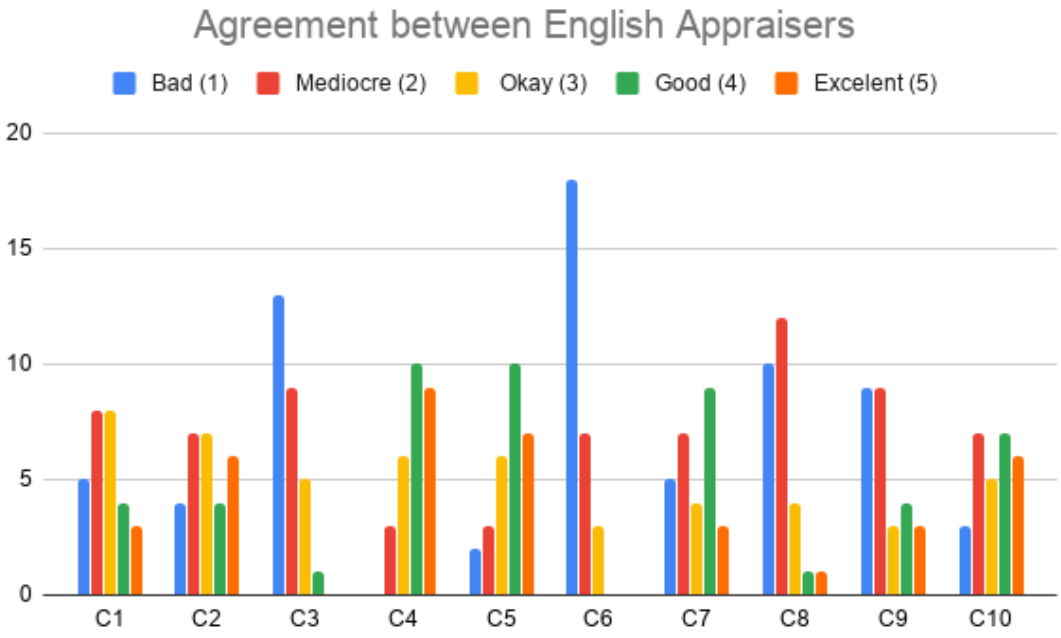

FIGURE 4.4: Agreement between English Appraisers

In Figure 4.5 shows another way to represent the data that portrays the fluctuations of the scores per category. For category 1, the appraisers seemed not to agree, with the scores evenly spread across the range from bad (1) to excellent (5). Category 2 had the same result, while category 3 saw most appraisers vote for a bad (1) score. Category 4 and 5 skewed toward a mostly good (4) score.

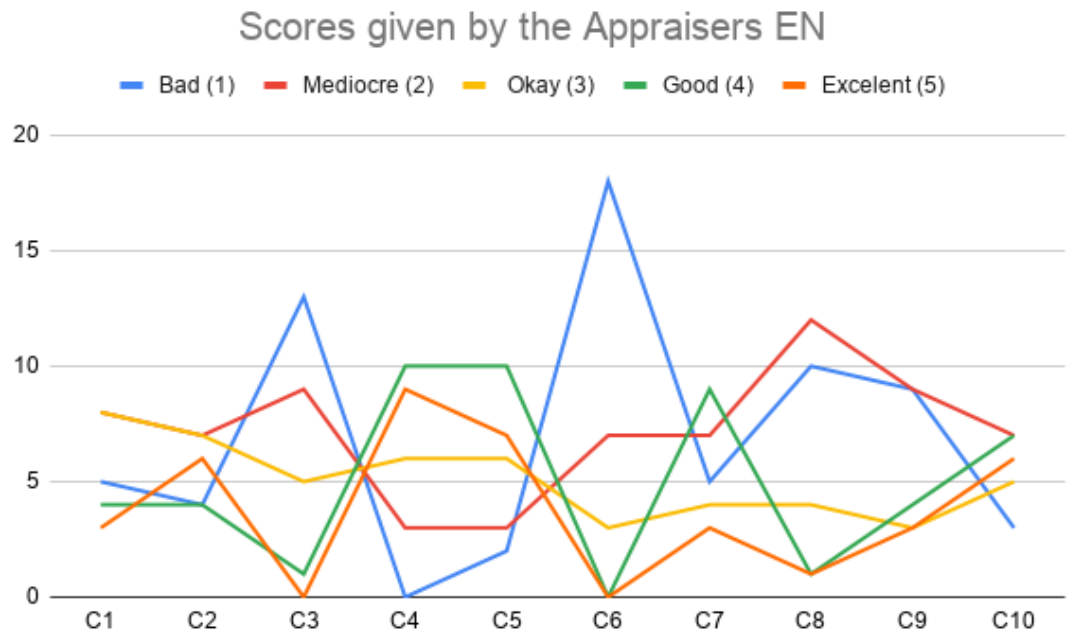

FIGURE 4.5: Scores given by the English Appraisers

Category 6 was the most noticeable because the majority of the votes were concentrated in a single score. There was almost unanimous agreement among raters in expressing that the caption was "bad (1)", with zero good (4) or excellent (5) scores recorded. Category 8 recorded a preference for the mediocre (2) score.

Further, category 9 favored the bad (1) and mediocre (2) scores equally, but there were a small number of votes for the rest of scores. Lastly, category 10 presented the same results as category 1 and 2; there was no overall preference among appraisers, who voted for all different scores.

Having discussed the scores given by the English appraisers, I will now address the scores given in Spanish. I will then focus on my calculation of Fleiss' Kappa and Kendall's Coefficient.

4.2.2 Spanish Appraisers

The scores of the Spanish appraisers can be found in Table 4.2 in the same format as the table for the English scores. At first glance, it is apparent that the Spanish appraisers agreed more in their ratings than the English appraisers. The table suggests that the appraisers mostly voted for the excellent (5) score.

	Bad (1)	Mediocre (2)	Okay (3)	Good (4)	Excellent (5)
C1	2	0	5	5	16
C2	4	3	4	5	12
C3	1	0	8	8	11
C4	0	1	7	4	16
C5	1	0	1	4	22
C6	9	6	3	6	4
C7	3	9	7	6	3
C8	9	8	7	3	1
C9	5	6	8	3	6
C10	1	0	4	7	16

TABLE 4.2: Data gathered from Survey B - Spanish Appraisers

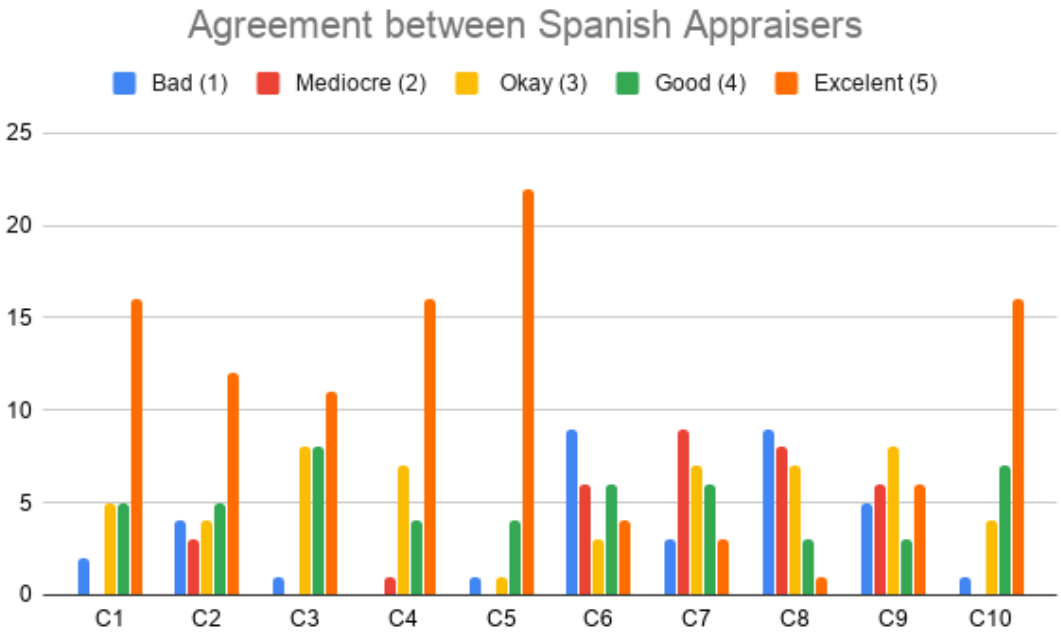

FIGURE 4.6: Agreement between Spanish Appraisers

Moreover, both Figure 4.6 and Figure 4.7 show a preference for the excellent (5) score. Categories 1, 2, 3, 4, 5 and 10 saw the most appraisers choose the excellent (5) score. Nevertheless, the ratings for categories 6, 7, 8 and 9 were more evenly distributed, ranging from poor (1) all the way to excellent (5).

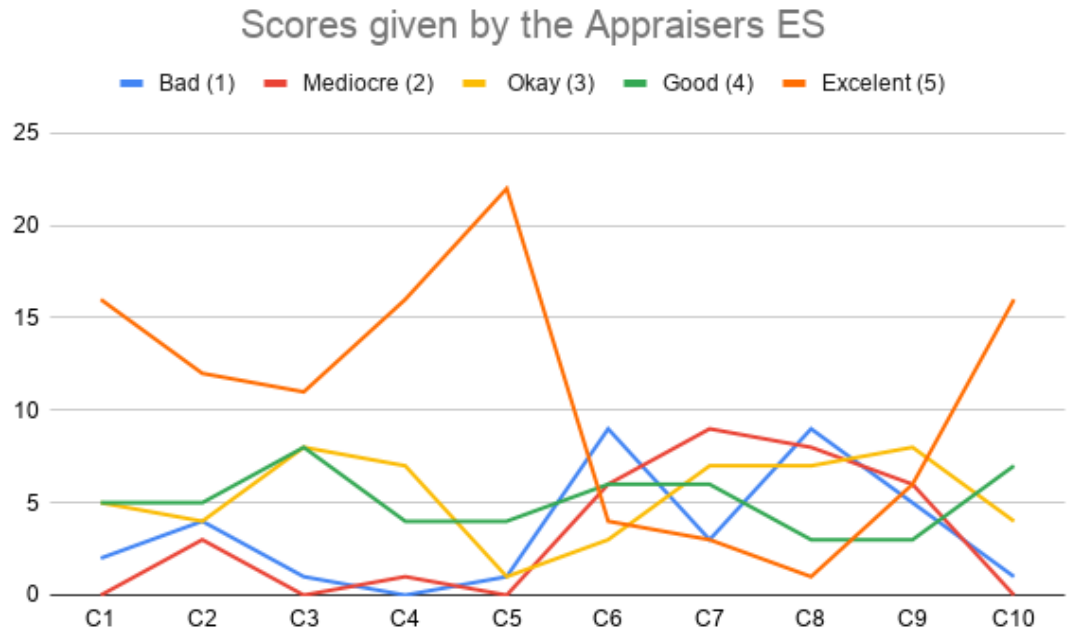

FIGURE 4.7: Scores given by the Spanish Appraisers

All in all, when looking at the scores, there appears to have been less consensus among the appraisers in English than among the Spanish appraisers, who gave the captions mostly positive scores. 6 out of 10 categories were rated as excellent (5) by many Spanish appraisers while the 4 remaining categories showed a lack of agreement amongst raters.

The next section will discuss the calculations made in order to assess the agreement between appraisers (Fleiss' Kappa) as well as the degree of association among raters (Kendall's Coefficient).

4.3 Human Evaluation

As mentioned in the methodology section of this paper, I decided to work with both Fleiss' Kappa and Kendall's Coefficient to better understand the agreement and association between appraisers and because this study does not have a regular distribution of data. In the following sections, I will discuss the calculations I made and findings I obtained through these metrics.

4.3.1 Fleiss' Kappa

To obtain the Fleiss' Kappa measurement, I followed the tutorial proposed by Zaiontz, 2019 and worked with Excel[5].

Since there were certain calculations to be made while generating the Kappa score, I have created a list of variables that the reader will encounter in this chapter relating to the calculations made when working with Fleiss' Kappa. This list is accompanied by the definition of each variable and why they are relevant for this study.

- **m**: number of observations; here, this is the number of appraisers.
- **n**: number of items; here, this is the number of questions posed to the appraisers.
- **Pe**: the expected proportion that agree, in other words, the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly seeing each category.
- **s.e**: the standard error of the Kappa coefficient for testing whether Kappa = 0.
- **p-value**: used to determine whether to reject or not to reject the null hypothesis. However, for Kappa this is rarely reported, probably because even relatively low values of Kappa can be significantly different from zero but not of sufficient magnitude to satisfy researchers. The p-value will be very important when we address Kendall's Coefficient.
- **z**: serves to test whether Kappa is statistically significant or not.

In the following subsections I will discuss the scores for both the English and Spanish appraisers and what these scores mean for my research.

English Appraisers

For the calculations, I used the same table that condensed the English results, in other words, Table 4.1. Furthermore, as seen in Table 4.3, the overall Kappa score for the English appraisers is 0.060, which is extremely low. This score shows that there is a lack of agreement between raters, to the degree that it could be argued that there is no agreement beyond that which could be explained by chance.

[5]Please refer to the Annexes to see all the calculations.

m	28
n	10
Pa	0.2582010582
Pe	0.2103316327
Kappa	0.060619657
s.e	0.008294383
z	7.308519117
p-value	0

TABLE 4.3: Fleiss' Kappa Score for English Appraisers

This raises more questions than answers, with regard not only to the performance of the API but also to what criteria the appraisers took used to rate the captions. The appraisers were asked to "assess how accurate the caption was with regards to the image", but based on the results, it appears that each appraiser applied a different set of criteria based upon their personal standard of accuracy. One could also theorize that the captions depict the content of the image to a certain extent but fail to fully describe all the elements, leaving the rater with the feeling that something is missing.

For instance, it is evident that the API failed to recognize certain colors for some images, mixing up, for instance, red and pink. This is what occurred to the first image in the Survey A, where the caption was "a close up of a red flower hanging from a tree" when in reality, the flower portrayed was pink. Inaccuracies like this make it difficult for the appraiser to give an objective score because even if the grammar of the sentence and most of the context is correct, a key detail about the colors is incorrect.

Although details like these might seem trivial, they could pose a problem if an investigator were to work with a system like this to identify different species of plants. In fact, for species of plants that are relatively similar, the colors may serve to indicate the use of the plant or even that the plant belongs to another type of species. If the system makes mistakes in identifying these species, then the database would be unusable.

Finally, another issue that could have confused appraisers is that at times the API mixed up singular and plural. For instance, image number 2 shows multiple red flowers, but the caption only states "a red flower in a field". This combined with the color confusion could have affected the appraisers' scores in English.

Spanish Appraisers

In the calculations of Fleiss' Kappa in Spanish we see that, as shown in Figure 4.9, the overall Kappa score for the Spanish appraisers is 0.089. This is slightly better than the score obtained for the English captions but still quite low. In the same way as in English, the captions in Spanish presented a few mistakes ranging from the use of plural and singular to wrong colors.

m	28
n	10
Pa	0.3137566138
Pe	0.2459183673
Kappa	0.08996140931
s.e	0.008659228159
z	10.389079450
p-value	0

TABLE 4.4: Fleiss' Kappa Score for Spanish Appraisers

For instance, in image 6, the caption "una flor rosa" states that there is a pink flower, but the flower in the picture is actually violet. Moreover, image 7 shows a bunch of flowers but the API fails to recognize this as well by using the singular form in the caption "flor amarilla en el pasto".

Moreover, the caption for image 8 is "flor de color amarillo en el pasto", but this is a mistake: the API seems to have confused the anthers of the flowers with their petals. This is particularly concerning given that the original purpose of using the API was to determine whether the system was hypothetically capable of properly identifying flora.

Ultimately, although the captions present similar problems in both in Spanish and in English, the appraisers in Spanish show more agreement. Notwithstanding, the Kappa score for both groups is quite low and the interrater reliability seems to be poor. Since this raises some concerns, I decided to also use the metric of Kendall Rank Correlation Coefficient.

4.3.2 Kendall Rank Correlation Coefficient

To go further and test the degree of association between appraisers, I used the Kendall Rank Correlation Coefficient. In this section I discuss the scores and compare them to the ones obtained by Fleiss' Kappa in order to see whether there is a correlation among them.

As mentioned in the previous chapter, the Kappa statistics represent absolute agreement between ratings, while Kendall's coefficients measure associations between ratings. Therefore, kappa statistics treat all misclassifications equally, but Kendall's coefficients do not treat all misclassifications equally.

I was curious to see the difference this made in the scores. As in previous sections, I will divide this section between English and Spanish appraisers.

English Appraisers

As can be seen in Figure 4.10, the overall score for Kendall for English appraisers is 0.012. Kendall's Coefficient, like Kappa, is measured from 0 to 1 with 1 being a perfect association. Therefore the score obtained from Kendall's calculations was extremely low, even more so than the score obtained from Fleiss' Kappa. This proves that the association between appraisers is poor.

Moreover, when calculating Kendall, I had to also carry out the Chi-Square test so that I could determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories to see whether the null hypothesis is true. Given that the score for the Chi-Square test is 3.15, I proceeded to calculate the p-value.

I found the p-value to be 0.95[6], which proves that there is strong evidence for a null hypothesis – meaning that there there no statistical significance in the tests I carried out.

Kendall W	0.01252319109
Chi-Square	3.155844156
Degrees of Freedom	9
p-value	0.957794174

TABLE 4.5: Kendall Score for English Appraisers

Bearing this in mind, I next calculated Kendall's Coefficient for the Spanish appraisers to see whether the null hypothesis were also true for that group.

Spanish Appraisers

As can be seen in Figure 4.11, the overall score for Kendall for Spanish appraisers is 0.056. As expected, this is higher than the English score. However,

[6]Please notice that a p-value that is less than or equal to 0.05 is usually used to indicate whether there is strong evidence against the null hypothesis.

it presents the same issue as did Kappa: the score is extremely low, meaning that the level of association is poor for this group as well.

I again carried out the Chi-Square test to find the p-value and prove whether the null hypothesis was true. Considering that in this case the p-value is 0.11, I found that the null hypothesis cannot be rejected.

Kendall W	0.05638218924
Chi-Square	14.20831169
Degrees of Freedom	9
p-value	0.1151057349

TABLE 4.6: Kendall Score for Spanish Appraisers

In the light of these results, it can be stated that for both the Spanish and English appraisers the null hypothesis cannot be rejected due to the lack of statistical significance.

4.3.3 Percentages

To conclude the section on human evaluation, I thought it pertinent to generate a percentage of agreement based on the scores per appraiser to further prove the null hypothesis. In the same way that I have done throughout this chapter, I will discuss the percentages for each language separately.

These percentages were taken by grouping the scores[7] in the following way: scores that ranged from 1 to 2 were grouped together and scores ranging from 3 to 5 were grouped together.

Although this grouping was arbitrary, it was based the metric I designed for the appraisers: 1 and 2 were “poor” and “mediocre” whilst 3, 4 and 5 are “okay”, “good” and “excellent”. It seemed appropriate to divide this scale into "low" and "high scores" for the final calculation of percentages; however, I am aware of how subjective this might be.

Subsequently, the ratio was calculated as well as the percentage per group in each language. The percentages obtained will be portrayed in both Table 4.8 and Table 4.10 in English and Spanish respectively.

[7]This is very important since the scores were not grouped for the other calculations.

English Appraisers

For the English appraisers, once the categories were grouped and the percentages calculated, it can be seen that the percentage for the lower-score group (1-2) was 50.36% while the percentage for the higher-score group (3, 4 and 5) was 49.64%.

This indicates that again the null hypothesis cannot be rejected. My hypothesis was that the percentage for the higher scores would be higher than or equal to 50%, and this was not achieved, although the percentage was quite close.

Survey Questions EN										**Survey Scores**	**Total**
Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10		
13	11	22	3	5	25	12	22	18	10	1 and 2	141
15	17	6	25	23	3	16	6	10	18	3, 4 and 5	139

TABLE 4.7: Grouped Categories in the English Survey

Ratio EN										**Survey Scores**	**Percentages**
Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10		
46.43	39.29	78.57	10.71	17.86	89.29	42.86	78.57	64.29	35.71	1 and 2	50.36
53.57	60.71	21.43	89.29	82.14	10.71	57.14	21.43	35.71	64.29	3, 4 and 5	49.64

TABLE 4.8: Percentages of Agreement in English

Spanish Appraisers

Finally, by examining the grouped percentages in Spanish we obtain the following: the first group that contained the lower scores (1 and 2) has a percentage of 23.93% while the higher-scores group (3, 4 and 5) had a percentage of 76.07%.

If only the grouped scores were taken into account, it could be said that for the Spanish appraisers the null hypothesis could be rejected. However, of all the tests and calculations made, this was the only one to give this result. Moreover, the result is largely related to the fact that the categories of scores were grouped.

Survey Questions ES										Survey Scores	Total
Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10		
2	6	1	1	1	15	12	17	11	1	1 and 2	67
26	22	27	27	27	13	16	11	17	27	3, 4 and 5	213

TABLE 4.9: Grouped Categories in the Spanish Survey

Ratio ES										Survey Scores	Percentage
Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10		
7.14	21.43	3.57	3.57	3.57	53.57	42.86	60.71	39.29	3.57	1 and 2	23.93
92.86	78.57	96.43	96.43	96.43	46.43	57.14	39.29	60.71	96.43	3, 4 and 5	76.07

TABLE 4.10: Percentages of Agreement in Spanish

Generally speaking, it appears that the English appraisers were more critical than the Spanish appraisers with regards to the captions generated by the API. It could be said that there is perhaps a cultural bias, since it would seem that Spanish appraisers tend to give positive scores more often than the English speakers. One could ask oneself in this case whether the captions in Spanish were indeed slightly better or the judgment of the Spanish appraisers was more lenient.

In effect, it appears that the Spanish appraisers tend to give better scores regardless of the quality of the captions themselves. With this in mind, the section on human evaluation comes to an end. The next section is dedicated to machine evaluation and the results obtained from the BLEU score metric.

4.4 Machine Evaluation

In this section I address the results obtained using the BLEU score metric by working with the online tool Tilde Custom Machine Translation. This evaluation was made with the help of two native speakers in their respective languages that took the time to write captions for the same images shown to the machine[8].

The tool compared the captions generated by the native speaker to the captions generated by the machine and gave a score measuring how similar they

[8]Please note that in Tilde's Custom Machine Translation, the BLEU score only allows for one written human file to be selected. This is one of the reasons why I only picked one native speaker for each language

were. As it has been done in previous sections, the results for each language are considered separately.

4.4.1 BLEU Score

The BLEU score is a string-matching algorithm that provides basic quality metrics for MT researchers and developers. It is likely the most widely used MT quality assessment metric over the last 15 years. While it is largely understood that the BLEU metric has many flaws, it continues to be the primary metric used to measure MT system output, even with the rise of Neural MT (Vashee, 2019).

The main idea behind BLEU is that the closer a machine translation is to a professional human translation, the better it is. In this study, the “human translation" standard was provided by the captions written by the native speakers. The interactive BLEU score evaluator assesses the similarities by strings.

Although I am aware that using just one native speaker per language could potentially be a drawback, since there are numerous synonyms and paraphrases possible for each caption, the machine evaluation part of this study serves only as an indicator and a complement to the human one and should not to be seen as a standalone element.

English

Figure 4.8 shows an overall BLEU score of 8.49, which is significantly low. There is also a histogram that shows how close the machine-generated caption was to the human input. The captions for categories 1, 3, 4, 5 and 7 obtained lower scores, while categories 2, 6, 8, 9 and 10 obtained a better correlation.

The fact the scores obtained through machine evaluation are also low serves as a confirmation that the captions generated by the machine in English lacked the necessary richness or descriptive precision to be deemed accurate.

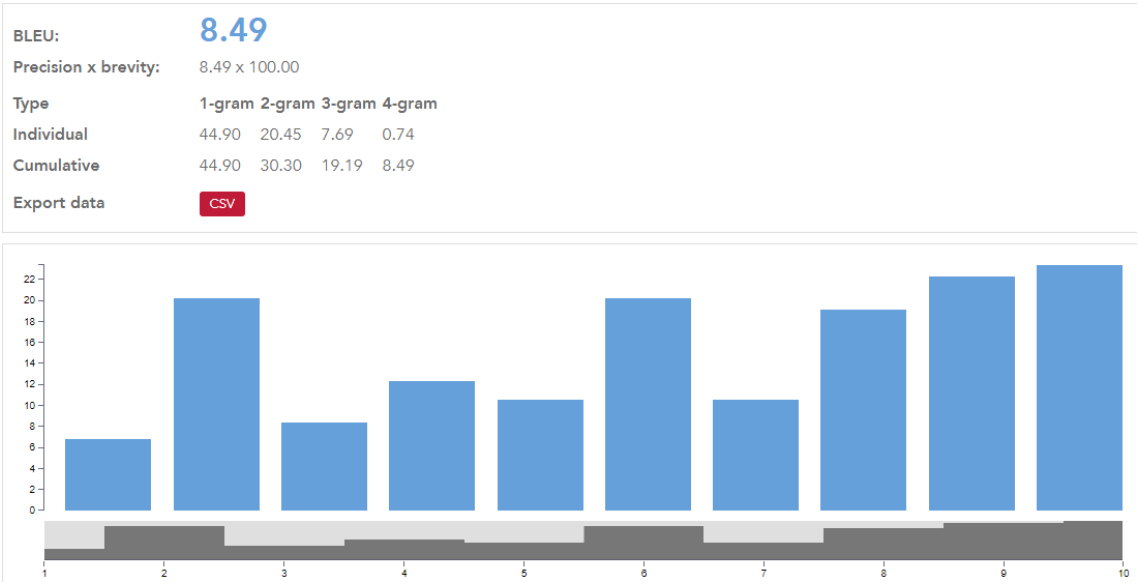

FIGURE 4.8: BLEU scores for English captions

Sentence	Human translated sentence	Machine translated sentence	BLEU score
1	pink flower with sunlight .	a close up of a red flower hanging from a tree .	6.754.313
2	a bunch of red flowers .	a red flower in a field .	20.164.946
3	blurry white flowers and purple flower in the middle .	a bunch of purple flowers .	8.423.556
4	a weeping willow .	a tree in the middle of a lush green field .	12.356.221
5	some pink flowers .	a close up of a flower .	10.552.670
6	a violet flower .	a pink flower on a plant .	20.164.946
7	some yellow flowers .	a white flower on a plant .	10.552.670
8	a bunch of white flowers .	a yellow flower in the grass .	19.070.828
9	a tiny yellow flower .	a close up of a flower .	22.316.181
10	a rose in a tree .	a close up of a rose .	23.356.899

TABLE 4.11: Sentence Comparison Machine vs Human in English

Spanish

For the Spanish captions, Figure 4.9 shows a BLEU score of 21.09, which is slightly higher than the English score. However, it is also a low score for BLEU, meaning that the correlation between the machine-generated captions and the human input is quite weak.

Category 2, 3, 5 and 9 showed weaker correlation, while the rest of the categories shared a higher correlation.

These results further confirm what was seen with the human evaluation scores. The weak correlation indicates that the captions lacked depth and failed to generate an average level of satisfaction among appraisers.

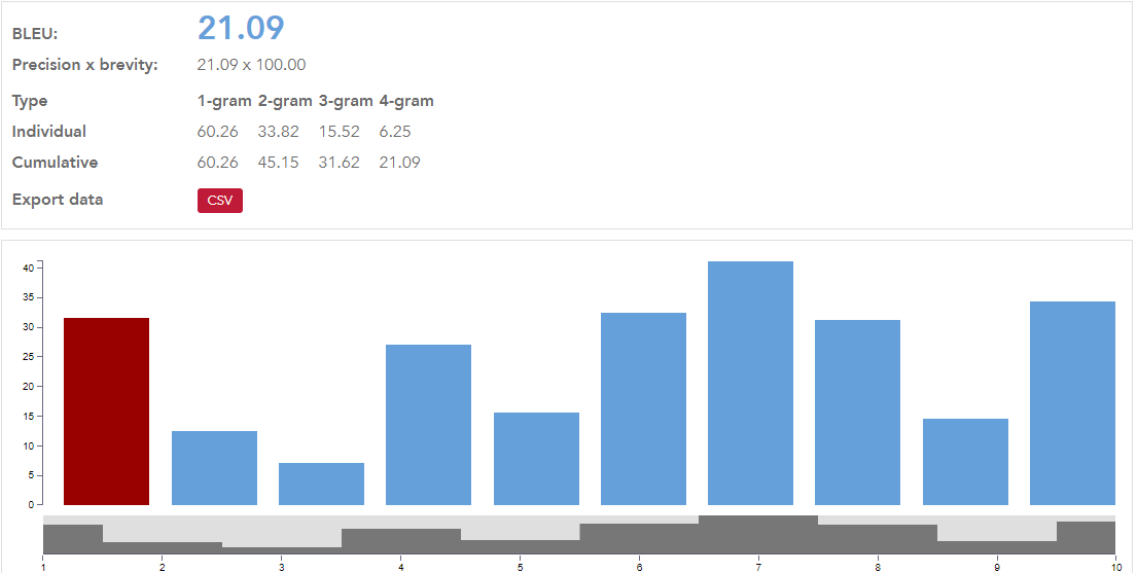

FIGURE 4.9: BLEU score for Spanish captions

Sentence	Human translated sentence	Machine translated sentence	BLEU score
1	una flor rosada con rocío .	una flor rosa en una rama .	31.559.845
2	flores rojas en el campo .	flor de color rojo en el pasto .	12.549.311
3	florecitas blancas y una flor morada en el centro .	planta con flores moradas .	7.115.864
4	un sauce en el campo .	un árbol en un campo .	27.054.113
5	flores rosadas .	planta con flores rosas .	15.619.700
6	una florecita violeta .	una flor rosa .	32.466.792
7	flores amarillas en el pasto .	flor amarilla en el pasto .	41.113.362
8	florecitas blancas en el pasto .	flor de color amarillo en el pasto .	31.239.399
9	una florecita amarilla .	planta con flores amarillas .	14.535.768
10	una rosa rosada .	una flor rosa .	34.329.452

TABLE 4.12: Sentence Comparison Machine vs Human in Spanish

4.5 Final Considerations

In this section, I would like to discuss two aspects related to the results and the system itself. I address the overwhelming negative scores and analyze issues that could have come into play in this study apart from those previously mentioned in the chapter. In essence, I would like to consider whether the reason the captions did not work was because of a fluency problem or an accuracy one.

In addition, I discuss the data protection issue that is intrinsically tied to this type of API: each time you use images or work with the system, you are agreeing to submit your data to the Microsoft Corporation. I believe this is an important question to consider when choosing whether or not to work with a pre-existing architecture.

4.5.1 Fluency vs Accuracy

When analyzing the considerable amount of negative scores, I wondered whether the captions had a fluency or accuracy problem. In other words, was this a linguistic issue or a semantic one?

When referring to **fluency**, in this particular case, what should be observed is the syntax of the captions and whether these are grammatically correct. **Accuracy**, on the other hand, demonstrates that there is coherence between the representations, objects, placements of the scene and what is said in the caption. In other words, an accurate caption properly describes the relevant aspects of the scene without adding or subtracting important elements.

When reflecting on these two concepts, I realized that the captions mainly had an accuracy problem rather than a fluency one. The syntax was not an issue, as the captions in both languages were grammatically decent. Most of the underlying problems related to the lack of coherence between what was portrayed in the scene and what was stated in the captions. Often, the captions generated would augment the number of objects by saying that there were "flowers" in plural when the picture showed only a single flower.

Similarly, some of the captions showed problems with colors, stating, for example, that a pink flower was purple and vice versa. Finally, it became apparent that there was an issue with regards to the placement of certain objects. If the flower was surrounded by other flowers, above other flowers or simply surrounded by grass, sometimes the captions would be vague or not mention the placement or create a semantically strange sentence such as "a pink flower on a plant".

4.5.2 Data Protection

Apart from the scores obtained through this study, I would like to mention that there are other factors to be considered when deciding whether to work with the API. These include data protection issues, previous requirements and the funds that must be invested to work with this tool. Although these factors are not quantitatively assessed, they play a key role in the decision to use this type of commercial tool or to create an architecture yourself.

These considerations are for users to ponder. It should be noted that Microsoft's API has a trial period of a month, which allows the user to get familiar with the API and see whether it matches their expectations. After the first month, you must pay per use. In other words, you are charged according to how much you use and work around the API; if you don't use it, Microsoft will not charge you.

This concludes the fourth chapter of this study. The fifth and final chapter presents the conclusions drawn from the study and briefly summarizes the results discussed here.

Chapter 5

Summary and Conclusions

This is the last chapter of the research project. In Section 5.1 I will summarize the results of the study and highlight the limitations, advantages and disadvantages of the API. Lastly, I will present the conclusions of this thesis in Section 5.2 and offer some reflections on the outlook for this type of API and on further research on this topic.

5.1 Summary of the Results

In total, four surveys were administered to English and Spanish native speakers to measure the accuracy of the captions generated by the API. The scores given by the appraisers were then analyzed using three tests: Fleiss' Kappa, the Kendall Rank Correlation Coefficient and percentage of agreement when scores were grouped into two categories – "low scores" (1–2) and "high to medium scores" (3–5). The API-generated captions were then assessed a second time using the BLEU score metric, a machine-evaluation tool.

After the calculations were obtained, the next step was to see whether the results confirmed or disproved the hypothesis proposed in this research, which was that the API would receive equal to or higher than 50% positive scores for both English and Spanish captions on both machine and human evaluation. In order to do so, I wanted to first see whether there was any statistical significance in the results, or in other words, to see whether the null hypothesis was rejected or not.

In fact, to make the findings clearer, I created a table of the results obtained from both the human and machine evaluation tests. In the right column per

appraiser there are two values, either "rejected" or "not rejected". These refer to the possibility of rejecting the null hypothesis or not according to the results obtained.

In this case, the null hypothesis represents that no statistical significance exists in a set of given observations. Hence, if the value is "not rejected", this means that the scores were not higher than 50% and therefore that there is not enough evidence to prove satisfaction amongst appraisers with regard to the captions.

On the opposite side of the spectrum, the "rejected" category is used when the score obtained by the test was high enough to prove satisfaction among appraisers.

Human Evaluation		
Tests	**English Appraisers**	**Spanish Appraisers**
Fleiss' Kappa	Not Rejected	Not Rejected
Kendall's Coefficient	Not Rejected	Not Rejected
Percentages	Not Rejected	Rejected
Machine Evaluation		
BLEU	Not Rejected	Not Rejected

TABLE 5.1: Results' Summary taking into account the Null Hypothesis

As can be seen, there was only enough evidence to reject the null hypothesis in the case of the percentages test. However, this is due to the fact that the scores were grouped into "high to medium" and "low" scores, making it easier to have a statistically significant amount of positive scores. I am aware that grouping the data in this way could be seen as an arbitrary measure, and so this result cannot outweigh the results obtained from the other tests.

Overall, the findings of this study show that there is not enough evidence to reject the null hypothesis in either the human evaluation or the machine evaluation.

The prevalence of low scores rating and the impossibility of rejecting the null hypothesis may be partially due to certain factors. For instance, a few errors were recurrent in the machine-generated captions, such as:

- Confusion between one color and another which could be a problem for identifying different species of plants.

- Misuse of plural and singular forms: often, the captions used a singular form (“a flower”) when there were several flowers in the image.
- Wrong description of the placement of the main object in the image: for example, "a flower on a plant" instead of "a flower on a branch".

The fact that the size of the sample of appraisers was quite small could have also been a factor in the results. Perhaps with a larger sample the results could have been more statistically significant. In addition, the fact that the annotation guidelines were given orally may have also influenced the perception of the appraisers, especially if they were not paying attention or forgot about the instructions.

Moreover, asking appraisers to rank the captions based on the subjective criteria of being "good" or "bad" might have been not the best way to assess this type of API. Perhaps the results would have been different had the appraisers been asked more specific questions, such as if the colors corresponded to the image or if the sentence used plural and singular forms properly.

Although I am aware that there is always room for improvement, I did try to be as specific as possible with the annotation guidelines so that the appraisers would know what to look for when rating the image and would not simply make an arbitrary decision.

5.1.1 Limitations

A significant limitation of the API was that it takes time for the user to become familiar with the code. It is one thing to work with your own code; working with a code created by another person or a corporation is quite another.

Another limitation was that systems that use clouds such as Azure’s tend to require the user’s device to have internet connectivity in order to query the remotely hosted model. This may be limiting in some scenarios, and may incur data usage costs for the user (Hebron, 2016).

Whenever I wanted to test the system I needed to have an internet connection. In theory, most of the time you are connected to the WiFi and it is quite easy to find an internet connection. However, it would pose a problem if, for example, you wanted to work on your code while on the plane or you did

not want to share your data (since each time you work with the cloud you are sending your data to a remote server).

5.1.2 Advantages

The API offers several clear advantages, such as the convenience of having a preliminary code you can work with instead of building the API from scratch, and a support service you can count on if you run into problems. In addition, the advantage of using a cloud is that your data is saved automatically (although this could arguably be a disadvantage in terms of data protection).

Finally, there are many instructions, tutorials and step-by-step guides on the internet to help the developer understand the architecture and how to deploy, modify and explore the code to make the best of the API. Microsoft also provides the user with various image repositories in GitHub to train the system.

5.1.3 Disadvantages

Despite the many advantages of a system like this, there are several important downsides that should be taken into account when deciding whether to use a commercial API or build your own architecture.

First, using a system such as this one represents a recurring cost and vendor lock-in. Although the cost of an individual query is generally quite low and bulk rates are available, the costs can become prohibitive if the system is used often enough without a viable revenue model (Hebron, 2016). Additionally, at the moment, Microsoft Azure does not offer a solution for migrating your data to other platforms or systems. In fact, there is a huge switching cost to using a different operating system instead (Contorer, 2004).

Furthermore, training the system to recognize a specific terminology of flora proved to be extremely difficult, and the captions produced by the machine tended to be quite generic. The most difficult the terminology is, the lower the degree of confidence it has.

Having listed the limitation, advantages, and disadvantages of this API, I will proceed to conclude this chapter by presenting the conclusions of this research as well as an outlook for this technology.

5.2 Conclusions and Outlook

The main objective for this project was to assess the accuracy of the Computer Vision API by using human evaluation and machine evaluation. Hence, my hypothesis was that the appraisers would give the machine's captions an overall positive score of more than 50%, which would prove that the CNNs of the API are sufficiently robust in both English and Spanish.

However, having carried out all the necessary tests, I have reached the conclusion that, In the light of the data gathered and the results obtained through different tests, there is not enough evidence to be able to reject the null hypothesis due to the lack of statistical significance in the tests and hence, the alternative hypothesis cannot be proved either.

The appraisers did not give a score higher that the 50% for the captions either in English or in Spanish. I wonder whether this result may be influenced by the fact that the sample of appraisers was too small. I believe this leaves the question open as to whether having a vast number of appraisers would have changed the results of the study or if the results would have stayed relatively similar.

Regardless of the answer to that question, however, it is fair to say that there were still persistent problems with the generation of the captions, such as the misuse of plural and singular forms and the confusion between colors. This shows that even though the API can generate captions quickly, there are still unfortunately some adjustments to be made in order for the system to give a more detailed and precise result.

I believe that the confidence scores help the user to easily find the most suitable result for his or her query. However, despite this functionality, the issues regarding the colors, and the plural and singular forms were still present in many of the captions.

Moreover, in the spirit of being critical, I am aware that using human evaluation and non-parametric statistics could be deemed subjective. It is true that trying to measure opinions can be a rather complex task. Nevertheless, the high number of negative scores obtained indicates that there are aspects of the captions that need to be improved.

In this particular case, I was interested in seeing how the API might be adapted for an specific corpus such as one comprising flora pictures. Although the

system was not as effective as I was hoping it would be, I believe that working with this type of APIs and technologies can be profoundly helpful in several disciplines, especially when creating a compendium or an illustrated glossary in different languages.

In consideration of the foregoing, the only remaining issue that I think would be pertinent to address is whether to make use of already built platforms and/or clouds. The question remains: should you submit your images to a corporation or be independent and develop your own architecture?

The answer depends on several factors, including the scale of your project, your budget and the resources available to you. In any case, as mentioned through the results chapter, there are prerequisites for using the API and it is important to have a certain familiarity with programming in order to run and modify the API. Although Microsoft offers the advantage of having a code already in place, perhaps the opportunity of creating an architecture of your own could seem quite attractive as well.

All in all, I hope that this research makes a contribution, sheds some light on possibilities for the use of this type of API for multilingual purposes and encourages other researchers to explore this subject.

Appendix A

Annotation Guidelines

You will find in front of you a survey that aims to assess the quality of captions generated by a machine.

Before starting the survey, some data will be collected such as your age, the languages you speak and whether the language of the survey is your mother tongue or not. This information is collected in order to explore the randomness of the sample.

Once this part of the survey has been filled out, you will be faced with 10 questions that comprise an image, a caption and a scale from 1 to 5. Number 5 represents the best score, while number 1 the worst score.

The question asked in each case will always be formulated in the same way: "Does the caption '...' match the image below?", where the ellipsis will be replaced by the description of the scene. Your task will be to assign a number on the scale based on the degree to which the caption given by the machine matches the image.

Please bear in mind the following criteria when assessing the accuracy of the captions in relation to the image:

Score		Criteria
1	Bad	The caption does not properly describe the image: when the caption describes a different scene than the one presented in the image.
2	Mediocre	There are major inconsistencies in the caption: when the caption partially describes the elements of the scene
3	Okay	The caption is good enough: when the caption manages to identify most of the key elements of the scene.
4	Good	Good caption: when the caption manages to identify all of the key elements of the scene.
5	Excellent	Perfect caption: when the caption describes every single element of the scene.

TABLE A.1: Criteria for the English scores

At the end of the questionnaire you will find a disclaimer concerning the authorization of personal data treatment that reads the following: "I hereby authorize the processing of my personal data solely for educational purposes". You have the possibility to accept or reject this clause.

Please notice that all the information submitted will be confidential and for research purposes only.

Thank you for taking the time to fill in this questionnaire. Your contribution will be extremely helpful for this research project.

Appendix B

Anotación de las directrices

A continuación encontrará un cuestionario que tiene como objetivo evaluar la calidad de los subtítulos generados por una máquina.

Antes de empezar el cuestionario, será necesario que por favor conteste algunas preguntas relacionadas con su edad, las lenguas que habla, y si la lengua en la que está escrito el cuestionario corresponde a su lengua materna. Esta información es recolectada para poder evaluar la aleatoriedad de la muestra.

Una vez haya llenado esta parte, usted encontrará 10 preguntas que comprenden una imagen, un subtítulo y una escala que va de 1 a 5. En cada caso, la pregunta siempre será formulada de la misma manera: "Usted considera que la descripción: "..." ¿corresponde con la imagen debajo?". Los puntos suspensivos representan los subtítulos de cada imagen.

Su tarea consistirá en evaluar, teniendo en cuenta la escala, cuán preciso es el subtítulo en relación con la imagen. Por favor tenga en cuenta que los puntajes son asignados basados en los siguientes criterios:

Puntaje		**Criterio**
1	Malo	El subtítulo no describe de manera correcta la imagen: cuando el subtítulo no corresponde para nada con lo que se ve en la imagen.
2	Mediocre	Se presentan varias inconsistencias en el subtítulo: cuando en la descripción faltan los elementos principales de la imagen.
3	Regular	El subtítulo está bien: cuando la mayoría de elementos presentes en la imagen son descritos de manera correcta.
4	Bueno	Es un subtítulo sobresaliente: cuando la imagen ha sido descrita de manera casi perfecta, pero hay algunos elementos principales que no aparecen en la descripción.
5	Excelente	El subtítulo es perfecto: cuando la escena ha sido descrita a la perfección.

TABLE B.1: Criterio utilizado para los puntajes en español

Al finalizar el cuestionario, se encuentra una cláusula en la que se puede leer lo siguiente: "Autorizo el uso de mis datos personales exclusivamente para propósitos educativos". Usted tiene la posibilidad ya sea de aceptar esta cláusula o de negarla.

Toda la información aquí recopilada es confidencial y sólo será tratada para fines educativos.

Muchas gracias por tomarse el tiempo de resolver este cuestionario. Su ayuda es de gran importancia para el desarrollo de esta investigación.

Bibliography

Aloimonos, Y. and A. Rosenfeld (1991). "Computer Vision". In: *Science* 253.5025. URL: https://www.jstor.org/stable/2879170.

Alom, M. Z. et al. (2017). "The History Started from AlexNet: A Comprehensive Survey on Deep Learning Approaches". In: p. 8. URL: https://arxiv.org/ftp/arxiv/papers/1803/1803.01164.pdf.

ArbolApp (2019). URL: http://www.arbolapp.es/.

Astala, R. and M. Hamilton (2017). *Saving snow leopards with deep learning and computer vision on Spark*. URL: https://customers.microsoft.com/es-es/story/snow-leopard-trust-nonprofit-azure.

Bengio, Y., A. Courville, and I. Goodfellow (2016). *Deep Learning*. The MIT Press, p. 1. URL: http://www.freetechbooks.com/deep-learning-t935.html.

Bishop, C. (2013). "Model-based machine learning". In: 371.1984, pp. 1–2. URL: https://www.jstor.org/stable/41739971.

Bobriakov, Igor (2018). "Kappa statistics and Kendall's coefficients". In: URL: https://medium.com/activewizards-machine-learning-company/comparison-of-top-6-cloud-apis-for-computer-vision-ebf2d299be73.

Bogdan, R. and S. Biklen (2006). *Qualitative research in education: An introduction to theory and methods*. Allyn Bacon.

Boscaini, D. (2017). *Geometric Deep Learning for Shape Analysis - Extending deep learning techniques to non-Euclidean manifolds*, p. 7.

Brown, C. M. (1984). "Computer Vision and Natural Constraints". In: *Science* 224.4655, p. 1299. URL: https://www.jstor.org/stable/1692735.

Callison-Bursch, C., R. Osborne, and P. Koehn (2006). *Re-evaluation the Role of Bleu in Machine Translation Research*. Vol. 6. EACL.

Carlsson, G. (2018). *Using Topological Data Analysis to Understand the Behavior of Convolutional Neural Networks*. URL: https://www.ayasdi.com/blog/artificial-intelligence/using-topological-data-analysis-understand-behavior-convolutional-neural-networks/.

Central, Data Science (2018). *Soccer and Machine Learning*. URL: https://www.datasciencecentral.com/profiles/blogs/soccer-and-machine-learning-2-hot-topics-for-2018.

Contorer, Aaron (2004). "EU report takes Microsoft to task". In: URL: https://www.cnet.com/news/eu-report-takes-microsoft-to-task/.

Davis, L. S. and D. Rosenfeld (1981). *Artificial Intelligence*. Vol. 245. Wiley, New York, p. 17.

DeepAI (2018a). *Deep Learning*. URL: https://deepai.org/machine-learning-glossary-and-terms/deep-learning.

– (2018b). *Semi-supervised Learning*. URL: https://deepai.org/machine-learning-glossary-and-terms/semi-supervised-learning.

Gonzalez, R. and R. Woods (2007). *Digital image processing*. Vol. 3. Pearson Prentice-Hall Inc.

Groen, I. I., H. E. Silson, and C. Baker (2017). "Contributions of low- and high-level properties to neural processing of visual scenes in the human brain". In: p. 3. URL: http://rstb.royalsocietypublishing.org/content/372/1714/20160102.

Hebron, P. (2016). *Machine Learning for designers*. Ed. by O'Reilly Media. URL: https://www.oreilly.com/learning/machine-learning-for-designers.

Hossain, Z. et al. (2018). "A Comprehensive Survey of Deep Learning for Image Captioning". In: *ACM Comput. Surv* 0, p. 36. URL: https://arxiv.org/pdf/1810.04020.pdf.

Hubel, H. D. and T. N. Wiesel (1959). "Receptive Fields of Single neurons in the Cat's Striate Cortex". In: *Journal of Physiology* 148, pp. 574–591. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1363130/.

Karpathy, A. (2018). *CS231n Convolutional Neural Networks for Visual Recognition*. URL: http://cs231n.github.io/convolutional-networks/#conv.

Kazimipour, B. (2018). *What is the difference between a Convolutional Neural Network and a regular Neural Network?* Ed. by StackExchange. URL: https://ai.stackexchange.com/questions/5546/what-is-the-difference-between-a-convolutional-neural-network-and-a-regular-neur.

Koenderink, J.J. (1990). *Solid Shape*. MIT Press Cambridge.

Korza, J. R. et al. (1996). *Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming*. Springer, Dordrecht, pp. 151–170. URL: https://link.springer.com/chapter/10.1007/978-94-009-0279-4_9.

LeCun, Y. (2013). *LeNet-5, Convolutional Neural Networks*. URL: http://yann.lecun.com/exdb/lenet/.

LeCun, Y., J. Bengio, and G. Hinton (2015). *Deep Learning*. Vol. 521. Nature, p. 436.

Marr, D. (1982). *A Computational Investigation into the Human Representation and Processing of Visual Information*. Freeman, San Francisco.

MathWorks (2018). *Learn About Convolutional Neural Networks*. URL: https://www.mathworks.com/help/deeplearning/ug/introduction-to-convolutional-neural-networks.html.

Microsoft (2018). *Microsoft Azure - Computer Vision API*. URL: https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/.

MiniTab (2019). "Kappa statistics and Kendall's coefficients". In: URL: https://support.minitab.com/en-us/minitab/18/help-and-how-to/quality-and-process-improvement/measurement-system-analysis/supporting-topics/attribute-agreement-analysis/kappa-statistics-and-kendall-s-coefficients/.

Mohri, M., A. Rostamizadeh, and A. Talwalkar (2012). *Foundations of Machine Learning*. The MIT Press.

O'Donoghue, T. and K. Punch (2003). *Qualitative Educational Research in Action: Doing and Reflecting*. Routledge.

PlantNet (2019). URL: https://plantnet.org/en/.

Russakovsky, O. et al. (2015). "ImageNet Large Scale Visual Recognition Challenge". In: URL: https://arxiv.org/pdf/1409.0575.pdf.

Russell, S. and P. Norvig (2009). *Artificial Intelligence: A Modern Approach*. Prentice Hall.

Sebe, N. et al. (2005). *Machine Learning in Computer Vision*. Springer Science and Business Media, pp. 10–16. URL: https://books.google.ch/books?id=lemw2Rhr_PEC&redir_esc=y.

Sharma, A. (2018). *What is the difference between Neural Networks and Deep Learning?* URL: https://www.quora.com/What-is-the-difference-between-Neural-Networks-and-Deep-Learning.

Szeliski, R. (2010). *Computer Vision Algorithms and Applications*. Springer Science and Business Media, pp. 10–16. URL: https://books.google.ch/books?id=bXzAlkODwa8C&redir_esc=y.

Tran, K. et al. (2016). "Rich Image Captioning in the Wild". In: *Microsoft Research*, p. 1. URL: http://openaccess.thecvf.com/content_cvpr_2016_workshops/w12/papers/Tran_Rich_Image_Captioning_CVPR_2016_paper.pdf.

Tsotsos, J. K. (1987). *Encyclopedia of Artificial Intelligence*. Wiley, New York.

Ullman, S. (1979). *The interpretation of Visual Motion*. MIT Press Cambridge.

Vashee, Kirti (2019). "Understanding MT Quality: BLEU scores". In: URL: https://www.sdl.com/blog/understanding-mt-quality-bleu-scores.html.

Vaughan-Nichols, S. (2017). *Microsoft Azure*. URL: https://uk.pcmag.com/cloud-services/73781/microsoft-azure.

Veen, F. van (2016). *A mostly complete chart of Neural Networks*. URL: https://asimovinstitute.org.

Vinyals, O. et al. (2015). "Show and Tell: A Neural Image Caption Generator". In: p. 4. URL: https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Vinyals_Show_and_Tell_2015_CVPR_paper.pdf.

Waschura, J. (2018). *What is the difference between a Convolutional Neural Network and a regular Neural Network?* Ed. by StackExchange. URL: https://ai.stackexchange.com/questions/5546/what-is-the-difference-between-a-convolutional-neural-network-and-a-regular-neur.

Wäldchen, J. and P. Mäder (2017). "Plant Species Identification Using Computer Vision Techniques: A Systematic Literature Review". In: URL: https://link.springer.com/content/pdf/10.1007%2Fs11831-016-9206-z.pdf.

Zaiontz, Charles (2019). "Fleiss' Kappa". In: URL: http://www.real-statistics.com/reliability/interrater-reliability/fleiss-kappa/.