



UNIVERSIDAD DE BURGOS  
ESCUELA POLITÉCNICA SUPERIOR  
Grado en Ingeniería en Informática



## **TFG del Grado en Ingeniería Informática**

**Estudio de herramientas de  
reconocimiento de imágenes con  
aplicación prototipo**



Presentado por Bryan Reinoso Cevallos  
en Universidad de Burgos — Julio de 2015  
Tutores: Dr. José Francisco Díez Pastor,  
Dr. César I. García Osorio



## **Resumen**

Este es un proyecto que trata el estudio de herramientas y técnicas usadas en algunos artículos sobre el reconocimiento de imágenes y, también, la implementación de un prototipo de aplicación que use los métodos que en los artículos se han leído.

En el proyecto se han estudiado con detenimiento las técnicas que se han usado en el reconocimiento de imágenes y, más específicamente, en el de la generación de descripciones a partir de una imagen. Se ha visto a través de este estudio que cualquier tarea que esté relacionada con el reconocimiento de imágenes, es una tarea laboriosa y compleja. También se ha concluido con el estudio, la necesidad de modelos grandes y potentes para poder soportar la carga de computo de esta tarea, además de requerir un soporte físico con los suficientes recursos para soportar el modelo de manera efectiva y que permita obtener un rendimiento aceptable.

Este proyecto también cuenta con una aplicación prototipo, cuyo objetivo es mostrar los posibles usos de este tipo de herramientas y de su funcionamiento. Esta aplicación está construida con la idea de que sea usada por una persona invidente, por lo tanto su interfaz gráfica es dirigida a este tipo de usuario. Esto se ha realizado para dar una visión de las posibilidades que tiene el desarrollo de este tipo de herramientas y modelos.

En conclusión, se ha podido ver que este tipo de herramientas aún tienen un camino largo por desarrollar, pero que han dado un paso más al conseguir evolucionar hacia la tarea de la generación de frases a partir de imágenes. Por lo tanto, estas herramientas se cree que irán evolucionando para abrir un nuevo abanico de posibilidades al desarrollo de aplicaciones para personas con ciertas discapacidades físicas. A pesar de esto, se ha conseguido desarrollar una aplicación prototipo que aporta unos resultados bastante satisfactorios y que inspira a seguir trabajando en esta línea de investigación.

## **Descriptores**

aprendizaje profundo, descripción automática de imágenes, lectura automática de texto, apoyo a personas con discapacidad



## **Abstract**

This project is about the study of tools and techniques used in some image recognition articles and the development of a prototype application using the methods discussed on the articles read.

In this project we have studied in detail the techniques used on image recognition and more specifically in the automatic generation of text descriptions for images.. We have seen with this study that any task related with image recognition is laborious and complex. Further, we have concluded with this study that the existence of large models, who can assume the processing of this kind of task, is needed. Also, hardware with enough resources, who can assume the model in effective way and permit obtain an acceptable performance, is required.

This project also has a prototype application whose objective is to show the possible uses of this kind of tools. Further, it shows how these tools work. This application is built with the idea to be used by a blind person. Thus its user interface is aimed at this kind of user. This has been realized with the objective of giving a view of the possibilities that the development of this kind of tools and models have.

In conclusion, we have seen that these kind of tools still have a large way to go, but they have experimented a great advance on image recognition and now they can produce descriptions from an image. Thus we believe that these tools will evolve to give us a new set of possibilities to develop applications and help disabled persons. Despite this , we managed to develop that has pretty good results and inspired to continue working on this line of research

## **Keywords**

deep learning, automatic image description, text to speech, support to people with disabilities



# Índice general

<b>1. Introducción</b>	<b>1</b>
<b>2. Objetivos del proyecto</b>	<b>3</b>
2.1. Objetivos funcionales . . . . .	3
2.1.1. Objetivos funcionales en el cliente Android . . . . .	3
2.1.2. Objetivos funcionales en el servidor . . . . .	3
2.2. Objetivos a nivel teórico . . . . .	4
<b>3. Conceptos teóricos</b>	<b>5</b>
3.1. Conceptos teóricos en el lado del Servidor . . . . .	5
3.1.1. Machine Learning . . . . .	5
3.1.2. Deep Learning . . . . .	9
3.1.3. Servicio Web . . . . .	9
3.2. Conceptos teóricos en el lado del Cliente . . . . .	10
3.2.1. Funcionamiento de las librerías Text2Speech . . . . .	10
3.3. Conceptos teóricos asociados a los artículos de investigación . . . . .	11
3.3.1. Convolutional Neural Networks o Redes Neuronales Convolucionales . . . . .	11
3.3.2. Recurrent Neural Networks o Redes Neuronales Recurrentes . . . . .	13
<b>4. Técnicas y herramientas</b>	<b>15</b>
4.1. Técnicas de desarrollo . . . . .	15
4.1.1. Metodología de desarrollo ágil . . . . .	15
4.1.2. Desarrollo software con control de versiones . . . . .	16
4.2. Herramientas utilizadas . . . . .	16
4.2.1. Gestor de Tareas: VersionOne . . . . .	16
4.2.2. Gestor de Versiones: GitHub . . . . .	17
4.2.3. IDE de desarrollo: Android Studio . . . . .	17
4.2.4. Herramienta de Generación de Documentación: TeXMaker . . . . .	18
4.2.5. Herramientas de Deep Learning: NeuralTalk . . . . .	18
4.2.6. Herramientas de Deep Learning: Caffe . . . . .	19
4.2.7. Herramientas de programación: MATLAB . . . . .	19
4.2.8. Herramientas de desarrollo de servidores: Flask . . . . .	19
4.2.9. Biblioteca Text to Speech para Android . . . . .	20
4.2.10. Biblioteca de Apache para conexiones Http para Android . . . . .	20
4.2.11. Biblioteca de traducción de texto . . . . .	21
<b>5. Estado del arte</b>	<b>23</b>
5.1. Artículo del DeepBelief SDK . . . . .	23
5.1.1. Introducción . . . . .	23
5.1.2. El dataset o conjunto de datos . . . . .	24
5.1.3. Arquitectura . . . . .	25
5.2. Artículo del NeuralTalk . . . . .	28
5.2.1. Introducción . . . . .	28

5.2.2. El modelo . . . . .	29
5.3. Artículo del Arctic Caption . . . . .	32
5.3.1. Introducción . . . . .	32
5.3.2. Generación de Frases de Imágenes con un Mecanismo de Atención . . . . .	33
5.3.3. Conclusiones . . . . .	34
5.4. Aplicaciones existentes relacionadas con el proyecto . . . . .	34
5.4.1. TapTapSee . . . . .	34
5.4.2. CamFind . . . . .	35
5.4.3. Talking Goggles . . . . .	35
<b>6. Aspectos relevantes del desarrollo del proyecto</b>	<b>37</b>
6.1. Dificultades encontradas . . . . .	37
6.1.1. Dificultades con DeepBeliefSDK . . . . .	37
6.1.2. Dificultades con GSOAP y Apache . . . . .	37
6.1.3. Dificultades en la instalación de Herramientas . . . . .	39
6.1.4. Instalando NeuralTalk . . . . .	39
6.1.5. Dificultades con NeuralTalk . . . . .	40
6.1.6. Dificultades con Android . . . . .	41
6.1.7. Dificultades con Librería de Traducción en el servidor . . . . .	42
<b>7. Conclusiones y líneas de trabajo futuras</b>	<b>43</b>
7.1. Conclusiones del Proyecto . . . . .	43
7.2. Líneas de trabajo Futura . . . . .	44
7.2.1. Líneas de trabajo futuras en el aspecto teórico . . . . .	44
<b>Anexos</b>	<b>44</b>
<b>I. Plan del proyecto software</b>	<b>47</b>
I.1. Introducción . . . . .	47
I.1.1. Problemas encontrados . . . . .	47
I.2. Planificación temporal del proyecto . . . . .	47
I.2.1. Sprint 1:23 de Diciembre al 12 de Febrero . . . . .	47
I.2.2. Sprint 2:12 de Febrero al 26 de Febrero . . . . .	48
I.2.3. Sprint 3:26 de Febrero al 6 de Marzo . . . . .	48
I.2.4. Sprint 4:6 de Marzo al 13 de Marzo . . . . .	49
I.2.5. Sprint 5:13 de Marzo al 27 de Marzo . . . . .	50
I.2.6. Sprint 6: 28 de Marzo al 22 de Abril . . . . .	51
I.2.7. Sprint 7: 22 de Abril al 27 de Abril . . . . .	53
I.2.8. Sprint 8: 27 de Abril al 8 de Mayo . . . . .	54
I.2.9. Sprint 9: 9 de Mayo al 15 de Mayo . . . . .	56
I.2.10. Sprint 10: 15 de Mayo al 5 de Junio . . . . .	57
I.2.11. Sprint 11: 6 de Junio al 20 de Junio . . . . .	58
I.2.12. Sprint 12: 21 de Junio a 30 de Junio . . . . .	58
<b>II. Especificación de requisitos</b>	<b>63</b>
II.1. Introducción . . . . .	63
II.2. Requisitos Funcionales . . . . .	63
II.2.1. Requisitos Funcionales en el Servidor . . . . .	63
II.2.2. Requisitos Funcionales en el Cliente . . . . .	63
II.3. Diagrama de casos de uso . . . . .	64

<b>III. Especificación de diseño</b>	<b>73</b>
III.1. Introducción . . . . .	73
III.2. Diseño en el Servidor . . . . .	73
III.2.1. API RESTful . . . . .	73
III.3. Diagrama de despliegue . . . . .	74
III.4. Diagrama de clases . . . . .	75
III.5. Diagrama de secuencia del sistema . . . . .	75
<b>IV. Manual del programador</b>	<b>79</b>
IV.1. Instalación del JDK . . . . .	79
IV.2. Instalación de Android Studio . . . . .	79
IV.3. SDK Manager, instalando herramientas . . . . .	83
IV.4. Importando Cliente en Android Studio . . . . .	85
IV.5. Instalación de Matlab . . . . .	87
IV.5.1. Activar MATLAB . . . . .	88
IV.6. Instalación de CUDA . . . . .	90
IV.7. Instalando Servidor y sus Dependencias . . . . .	92
IV.8. Configuración del Servidor . . . . .	94
<b>V. Manual del usuario</b>	<b>101</b>
V.1. Introducción . . . . .	101
V.1.1. Instalar Aplicación . . . . .	101
V.2. Uso de la aplicación . . . . .	101
<b>Bibliografía</b>	<b>105</b>



# Índice de figuras

3.1. Ejemplo de clasificación . . . . .	6
3.2. Ejemplo de regresión lineal. . . . .	6
3.3. Cambio de dimensionalidad en máquinas de vectores soporte. . . . .	8
3.4. Clasificación con árbol de decisión [20]. . . . .	8
3.5. Gráfico de librería <i>Text to Speech</i> [1] . . . . .	10
3.6. Gráfico de una CNN [11] . . . . .	12
3.7. Gráfico de conexiones recurrentes en una RNN . . . . .	14
5.1. Gráfico demostración de Saturación vs No Saturación . . . . .	25
5.2. Overlapping Pooling . . . . .	27
5.3. Asociación de imágenes a regiones del espacio . . . . .	29
5.4. Generación de descripciones de regiones de la imagen . . . . .	29
5.5. Diagrama del modelo NeuralTalk . . . . .	31
5.6. Diagrama de la RNN del modelo NeuralTalk . . . . .	32
I.1. Burn-down del sprint 3 . . . . .	49
I.2. Burn-down del sprint 4 . . . . .	50
I.3. Burn-down del sprint 5 . . . . .	51
I.4. Burn-down del sprint 6 . . . . .	53
I.5. Burn-down del sprint 7 . . . . .	54
I.6. Burn-down del sprint 8 . . . . .	56
I.7. Burn-down del sprint 12 . . . . .	58
II.1. Diagrama de caso de uso del cliente . . . . .	65
II.2. Diagrama de caso de uso del servidor . . . . .	66
III.1. Diagrama de despliegue . . . . .	74
III.2. Diagrama de clases . . . . .	75
III.3. Diagrama de secuencia . . . . .	76
IV.1. Página de descarga del JDK de Java . . . . .	80
IV.2. Instalación del JDK, paso 1. . . . .	80
IV.3. Instalación del JDK, paso 2. . . . .	80
IV.4. Página de descarga de Android Studio . . . . .	81
IV.5. Instalación de Android Studio, paso 1. . . . .	81
IV.6. Instalación de Android Studio, paso 2. . . . .	81
IV.7. Instalación de Android Studio, paso 3. . . . .	82
IV.8. Instalación de Android Studio, paso 4. . . . .	82
IV.9. Instalación de Android Studio, paso 5. . . . .	82
IV.10. SDK Manager paso 1. . . . .	84
IV.11. SDK Manager paso 2. . . . .	84
IV.12. Importando proyecto en Android Studio . . . . .	85
IV.13. Importando proyecto en Android Studio . . . . .	86

IV.14 Importando proyecto en Android Studio . . . . .	86
IV.15 Importando proyecto en Android Studio . . . . .	87
IV.16 Instalación de Matlab paso 1. . . . .	88
IV.17 Instalación de Matlab paso 2. . . . .	88
IV.18 Instalación de Matlab paso 3. . . . .	89
IV.19 Instalación de Matlab paso 4. . . . .	89
IV.20 Instalación de Matlab paso 5 . . . . .	89
IV.21 Activación de Matlab paso 1. . . . .	90
IV.22 Activación de Matlab paso 2. . . . .	90
IV.23 Comprobando directorio principal de instalación . . . . .	95
V.1. Instalando aplicación paso 1. . . . .	102
V.2. Instalando aplicación paso 2. . . . .	102
V.3. Instalando aplicación paso 3 . . . . .	103

# Índice de tablas

II.1. Caso de uso: RF1 Recibir imagen . . . . .	65
II.2. Caso de uso: RF2 Procesar imagen . . . . .	67
II.3. Caso de uso: RF3 Devolver predicción . . . . .	67
II.4. Caso de uso: RF3.1 Traducir predicción . . . . .	68
II.5. Caso de uso: RF4 Solicitar predicción . . . . .	68
II.6. Caso de uso: RF4.1 Tomar Foto . . . . .	69
II.7. Caso de uso: RF4.2 Mandar Foto . . . . .	69
II.8. Caso de uso: RF4.3 Leer predicción . . . . .	70



# 1. INTRODUCCIÓN

---

Este proyecto tiene como objetivo el estudio de herramientas y tecnologías de *Machine Learning*. Las herramientas que se quieren estudiar tendrán como funcionalidad el tratamiento de imágenes y la extracción de una predicción a partir de estas.

El estudio de este tipo de herramientas implica la lectura de la documentación y los conceptos teóricos adheridos a estas. Al finalizar este estudio, se deberá conocer qué tipo de técnicas se han usado para construir este tipo de modelos y por qué se han utilizado.

Como segundo objetivo, y para que el proyecto tenga también una dimensión práctica, se desarrollará un prototipo que utilice las técnicas y herramientas estudiadas. De este modo, además de la parte teórica, se aborda también la parte tecnológica de este interesante área del aprendizaje automático.

La aplicación de soporte para los usuarios será desarrollada en Android y constará de una interfaz y funcionamiento orientada al uso por personas con dificultades de visión, además se guiará al usuario a través de la aplicación con un asistente de voz que irá diciendo al usuario qué debe hacer y en qué punto de la aplicación se encuentra.

La aplicación tendrá como objetivo inicial mandar una foto tomada por el usuario a un servidor que la procesará y extraerá una predicción que le dirá al usuario lo que se puede observar en la imagen que él ha elegido. La predicción será una frase que describa la imagen y está será leída por la aplicación.

Por otra parte en el lado del servidor usaremos algoritmos de *Machine Learning* para el procesado de la imagen y la extracción de la predicción. El servidor recibirá una petición de tipo POST y cargará la imagen, la cuál será procesada y en respuesta devolverá una frase en español que describa la imagen.

Lo que busca este segundo objetivo del proyecto es conseguir crear un prototipo de esta aplicación que tenga un correcto funcionamiento y que sirva como base para desarrollar sobre él una aplicación más optimizada y de aún mejor funcionamiento.



## 2. OBJETIVOS DEL PROYECTO

---

### 2.1 Objetivos funcionales

Como se ha comentado anteriormente, el espíritu principal del proyecto es el estudio de las técnicas y herramientas de *Deep Learning*, pero desde el punto de vista funcional, el principal objetivo del proyecto es tener un prototipo de la aplicación que funcione de manera adecuada, para esto tenemos que tener en cuenta tanto los objetivos del lado del servidor como del lado del cliente.

#### 2.1.1 Objetivos funcionales en el cliente Android

El principal objetivo que podemos encontrarnos en el lado del cliente es el disponer de un prototipo de aplicación Android con la que el usuario pueda mandar de manera intuitiva una imagen al servidor y recibir de él la predicción esperada. Para que esto funcione de manera adecuada, se han propuesto una serie de requisitos que deben cumplirse:

- Para poder enviar la imagen al servidor y recibir de él la predicción, se debe establecer una conexión con él desde la aplicación Android y hacer un correcta petición de tipo POST, la cual subirá la imagen al servidor para ser procesada. Para la realización de esto se procederá al estudio de la documentación Android, buscando la manera más sencilla y adecuada para nuestro caso de uso.
- Para que la persona que lo utilice no tenga por qué saber la estructura de la interfaz gráfica de la aplicación y poder usarla sin problemas, nos apoyaremos sobre los eventos *ontouch* de Android, evitando de esta manera la dependencia de la interfaz gráfica para que la aplicación pueda llegar a ser usada por personas con dificultades visuales.
- Para ofrecer una guía fácil y útil para el usuario a lo largo de su experiencia usando la aplicación, usaremos la librería Text2Speech para ir guiando al usuario con instrucciones en voz alta, explicándole lo que debe hacer en cada momento. Finalmente, le devolveremos la predicción y se le dirá cuáles son sus opciones.
- Para poder tomar la imagen se requerirá de un dispositivo con cámara de fotos y se tendrá que establecer los permisos sobre la aplicación para el uso de la misma.

#### 2.1.2 Objetivos funcionales en el servidor

En el lado del servidor tenemos un objetivo claro y es el de recibir una imagen a través de una petición POST desde un cliente, procesarla y, finalmente, devolver la predicción en español. Para conseguir todo esto, tenemos una serie de requisitos y objetivos que tenemos que cumplir:

- Para poder recibir las peticiones y mandar respuestas, tenemos que tener un servidor con la capacidad de gestionar este tipo de operaciones, para ello, se deberá proceder al estudio

## 2. OBJETIVOS DEL PROYECTO

de las herramientas disponibles de predicción y buscar un *framework* de programación de servicios web, que se adapte mejor a nuestro caso de uso.

- Para procesar la imagen, que es un objetivo principal en el servidor. Se estudiará el funcionamiento de distintas herramientas de *deep learning*, eligiendo las que más interesantes resulten y que se puedan usar en el servidor.
- Finalmente, la predicción, que hayamos obtenido tras el procesado de la imagen, será devuelta como respuesta a la petición POST.

### 2.2 Objetivos a nivel teórico

Este proyecto conlleva una carga bastante grande en cuanto al estudio teórico de las herramientas que se van a proceder a usar y, con ello, una serie de objetivos a nivel teórico a tener en cuenta.

Como primer objetivo básico del proyecto se tiene el correcto estudio y entendimiento de las herramientas y protocolos se se van a usar. Esto implica un estudio bastante concienzudo de toda la información que aporta cada herramienta estudiada y su funcionamiento interno. El alumno deberá ser capaz, al finalizar el proyecto, de aportar una explicación coherente y suficientemente razonada de cómo y por qué funcionan las herramientas estudiadas.

En segundo lugar se pretende que el alumno sea capaz de interpretar y entender artículos de investigación que estén relacionados con este tema. Estos artículos deberán ser posteriormente debidamente explicados en la documentación del proyecto. Aunque el entendimiento de estos artículos no será profundo, sí será lo suficientemente riguroso como para poder realizar una presentación del funcionamiento de las herramientas.

En último lugar se pedirá una comparación crítica de las herramientas estudiadas y por qué unas son mejores o se entiende que funcionan mejor que las otras. O hacer una comparación de cómo está construida cada una y lo que diferencia sus arquitecturas.

## 3. CONCEPTOS TEÓRICOS

---

En este capítulo se profundizará en los conceptos teóricos con los que se ha trabajado a lo largo de todo el proyecto.

### 3.1 Conceptos teóricos en el lado del Servidor

En este apartado se pretende explicar todos los conceptos teóricos que se utilizan en el lado del servidor, tanto directamente usados por el alumno, como los que se usan en proyectos o librerías de apoyo que se usan en el proyecto.

#### 3.1.1 Machine Learning

El *Machine Learning* o también conocido como aprendizaje computacional, entre otros nombres, es una rama de la Inteligencia Artificial que pretende conseguir el objetivo de que los computadores puedan aprender de manera automática. De forma más general, se podría decir que el *Machine Learning* pretende crear programas informáticos capaces de generalizar comportamientos a partir de una información no estructurada que suele estar suministrada en forma de ejemplos.

El *Machine Learning* tiene muchísimas aplicaciones y es ahora mismo un campo de la Inteligencia Artificial que se encuentra en activo y del que surgen bastantes proyectos.

#### ■ Tipos de algoritmos

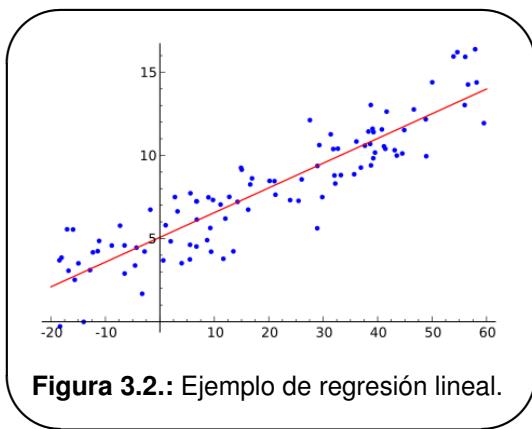
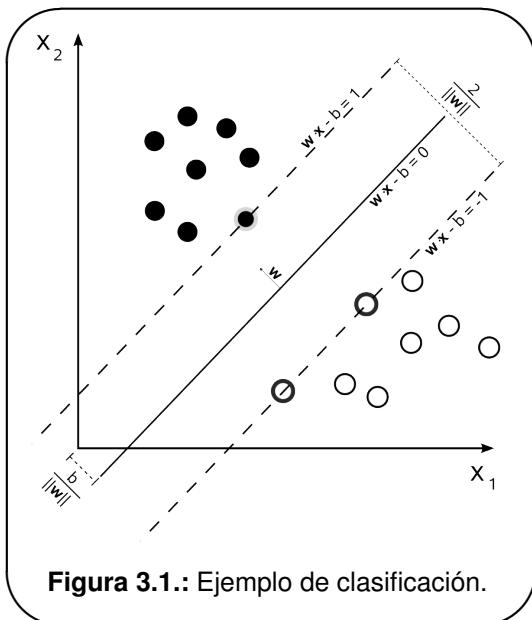
Aunque existen otras, la primera gran clasificación de los algoritmos de aprendizaje automático es en las tres siguientes categorías:

- **Aprendizaje Supervisado:** Este tipo de algoritmos crea una función que relaciona las entradas al sistema con las salidas del mismo. Aquí tendremos dos tipos de problemas a resolver, los cuáles son:

- Clasificación: El problema de clasificación se trata de que a través de un conjunto de datos, entrenamos nuestro modelo para que este devuelva como resultado una clase que clasifique al valor de entrada. Internamente estamos creando una función que nos devolverá la clase perteneciente a cada ejemplo con el menor error posible. En la Figura 3.1 se puede ver un ejemplo de clasificación en el que se quiere distinguir dos clases y donde la frontera de decisión es lineal [21].

Ejemplo: Tenemos un conjunto de datos con una estructura del tipo (altitud, presiónALTA, BAJA). El modelo será entrenado para recibir un valor cualquiera de altitud, y este devolverá o bien, clase0 = Presión baja o, clase1= Presión alta. Lo que el modelo está preparado para devolver es una clase.

### 3. CONCEPTOS TEÓRICOS



- Regresión: Este problema es bastante similar al de clasificación, pudiendo llegar a considerarse al de clasificación como un tipo de regresión. La principal diferencia de este con el de clasificación es que no esperamos una clase, sino un valor devuelto por la función construida, donde dicho valor será la predicción correspondiente al ejemplo. Internamente también creamos una función, pero esta está diseñada para devolver un valor numérico intentando predecir el estado del ejemplo, en función de los parámetros de entrada. Osea, la regresión nos devolverá un número como resultado, mientras que la clasificación devolverá una variable categórica. En la figura 3.2 podemos ver un ejemplo de regresión lineal a partir de una serie de ejemplos [23].

Ejemplo: Tenemos un conjunto de datos del tipo (Altitud,Presión), en este caso tanto la altitud como la presión toman valores numéricos reales. El modelo se preparará para buscar una función que se ajuste mejor a los datos de entrenamiento. Ante un ejemplo para predecir, el modelo devolverá un número real, que será el valor esperado de la presión a esa altitud y no una clase.

- **Aprendizaje no Supervisado:** En este tipo de algoritmo se lleva a cabo el modelado con una serie ejemplos que tan sólo constan de sus valores de entrada, mientras que el sistema desconoce a qué clase o qué salida le corresponde a cada ejemplo. Esto obliga al algoritmo a tener la capacidad de reconocimiento de patrones y ser capaz de diferenciar entre los ejemplos y dividirlos en grupos, en el que el resultado de la ejecución del algoritmo será el grupo al que pertenece cada ejemplo.
- **Aprendizaje semisupervisado:** Este tipo de algoritmos son una combinación de los dos anteriores, tiene en cuenta tanto los ejemplos etiquetados como los no etiquetados.

Dentro de este tipo de algoritmos, que son los que serán usados en el proyecto; podemos identificar los más significativos y relacionados con el proyecto:

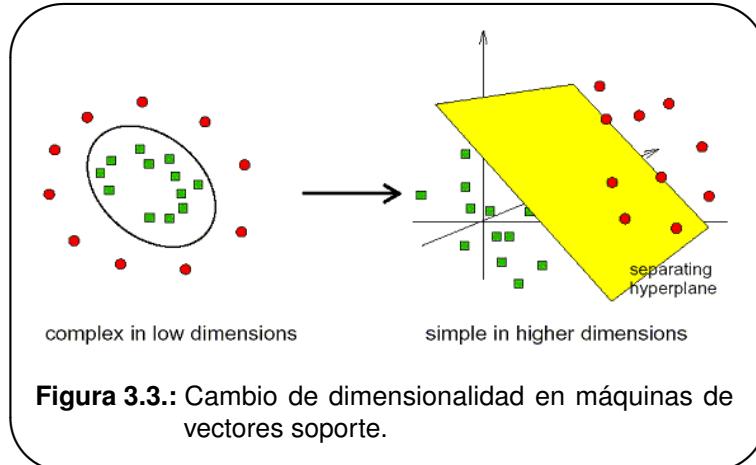
- Máquinas de vectores soporte: Se trata de un conjunto de algoritmos de aprendizaje supervisado, su principal particularidad es que a la hora de aprender la frontera de decisión, intenta encontrar aquella frontera que esté lo más alejada posible de ambas clases. Se dice que intentan maximizar el “margen”, la distancia entre la frontera y las instancias más cercanas a la misma. Estas instancias más cercanas reciben el nombre de vectores soporte, de ahí el nombre del método.

Aunque inicialmente están diseñados para resolver problemas de clasificación, se han publicado variantes capaces de resolver problemas de regresión.

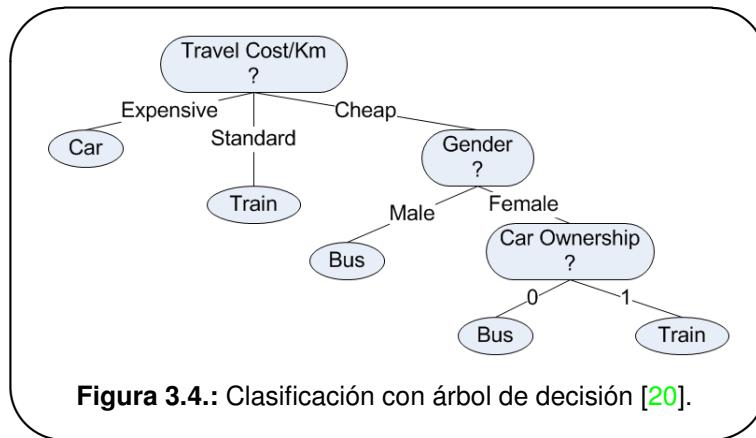
Aunque en el modelo básico la frontera que se aprende es lineal, si se añaden nuevas características basadas en las existentes, lo que en la práctica incrementa la dimensión espacial del conjunto de datos, la frontera aprendida puede ser más compleja que la lineal. Por ejemplo, un conjunto de datos que consista en las coordenadas  $x,y$ , y donde una de las clases se encuentre en el origen y la otra en un anillo en torno a la primera, no es linealmente separable (no hay una manera de trazar una línea recta que separe ambas clases, dejando cada una a un lado de la línea). Pero si se añade como característica el producto de las coordenadas  $x,y$ , se obtiene un nuevo conjunto de datos de tres dimensiones que sí es linealmente separable. Ahora una de las clases sigue siendo una nube en torno al origen, pero la otra clase se aleja de este, permitiendo que un plano separe a ambas. Para clarificar este ejemplo podemos ver la Figura 3.3 [19].

- Árboles: Los árboles de decisión son unos de los primeros métodos del *machine learning*. Estos árboles están compuestos por una serie de nodos internos de decisión y unos nodos

### 3. CONCEPTOS TEÓRICOS



**Figura 3.3.:** Cambio de dimensionalidad en máquinas de vectores soporte.



**Figura 3.4.:** Clasificación con árbol de decisión [20].

hojas, que se corresponden con la predicción o el resultado del modelo [7]. Cuando el árbol está construido, su uso es simple. Una nueva instancia se deriva al subárbol izquierdo o derecho del nodo raíz dependiendo de cuál sea el valor de uno de sus atributos. Ahora se considera el nodo raíz del nuevo árbol y se evalúa una nueva decisión que se basará en otro atributo, o en el mismo, aunque considerando un valor de división distinto. Así sucesivamente hasta llegar a un nodo hoja, momento en el que se predecirá la clase de la instancia de acuerdo a la etiqueta asignada al nodo hoja. Esto puede verse más claro en la Figura 3.4. El proceso de aprendizaje consiste en determinar los atributos y valores de división de cada nodo.

- Redes neuronales: Es un conjunto de algoritmos de aprendizaje tanto supervisado como no supervisado, los cuales usan el concepto biológico de la neurona y de las interconexiones neuronales para crear un modelo de neurona artificial y las redes neuronales. La neurona artificial tendrá una serie de entradas, que son procesadas por la función de activación de la neurona y devuelve una salida que será parte o el resultado del modelo, o bien podría ser la entrada a otra neurona, formando con ello grandes redes de neuronas artificiales. Las entradas a las neuronas procedentes de otras tienen un peso que indica cuánto influye esa entrada en el valor que genera la neurona. El proceso de aprendizaje consiste en determinar estos pesos.

### 3.1.2 Deep Learning

Se trata de un conjunto de algoritmos cuyo objetivo es intentar modelar abstracciones de alto nivel sobre los datos, usando para ello arquitecturas compuestas. Estas arquitecturas son de transformaciones no lineales y múltiples.

La definición de aprendizaje profundo o *Machine Learning* no está muy clara, ya que existe más de una definición. Por norma general, hace referencia a algoritmos centrados en el aprendizaje de manera automática. Aún teniendo este punto en común, podemos encontrar diferentes algoritmos cuyas características son las siguientes:

- Usar un conjunto de capas en forma de cascada, o puestas de manera consecutiva una tras de otra, lo que significa que la salida de una capa será la entrada de la capa posterior. Los algoritmos de este tipo pueden ser tanto de aprendizaje supervisado como de no supervisado, esto implica la necesidad de que algunos algoritmos tengan la capacidad de detectar patrones.
- Deben estar basados en el aprendizaje no supervisado de varios niveles de características o representaciones de datos. Las características forman una representación jerárquica, en las que las características de bajo nivel derivan a las de alto nivel.
- Aprender en varios niveles de abstracción que se corresponden a distintos niveles de representación, que forman una jerarquía de conceptos [22].

Los algoritmos de aprendizaje profundo contrastan con los algoritmos de aprendizaje poco profundo por el número de transformaciones aplicadas a la entrada mientras se propaga desde la capa de entrada a la capa de salida. Cada una de estas transformaciones incluye parámetros que se pueden entrenar como pesos y umbrales. No existe un estándar de facto para el número de transformaciones (o capas) que convierte a un algoritmo en profundo, pero la mayoría de investigadores en el campo considera que aprendizaje profundo implica más de dos transformaciones intermedias.

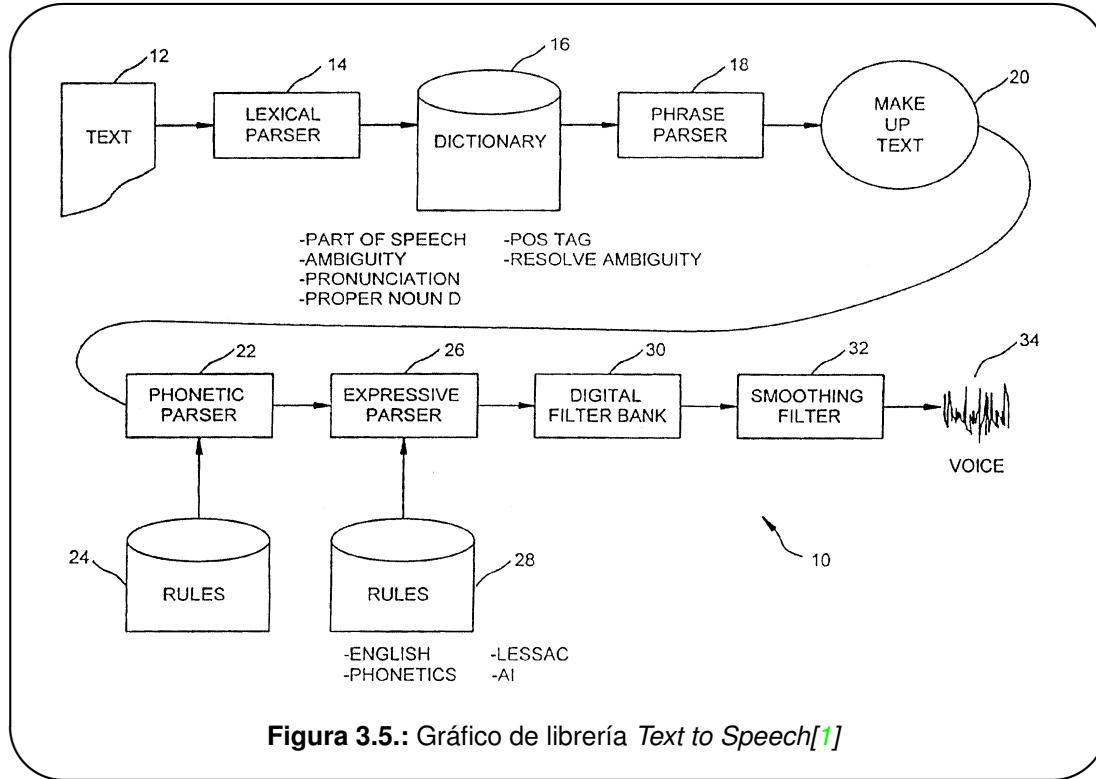
### 3.1.3 Servicio Web

Se trata de una tecnología que nos permite el correcto intercambio de datos entre distintas aplicaciones, para esto usa una serie de protocolos y estándares. El objetivo principal al usar este tipo de tecnología es conseguir que aplicaciones que trabajan en con distintos software, distinto idioma de programación, distinta localización e incluso con distinta plataforma de instalación; puedan comunicarse de manera adecuada y correcta consiguiendo que el paso de datos de una a otra sea posible, no sólo de manera adecuada sino, también, correcta. Para conseguir este objetivo, los servicios web utilizan estándares abiertos, osea, que es accesible para todos [24].

Entre estos estándares los más relevantes en este proyecto y su desarrollo han sido:

- XML: El formato de este tipo de documentos lo ha llevado a ser muy usado, incluso llega a ser la base para la definición de otros estándares. Este estándar se uso en la primera parte del proyecto, para una explicación más detallada ver el Capítulo 6.
- SOAP: Protocolos para establecer el intercambio de datos entre distintas aplicaciones. Hace referencia al tipo de dato. Estos protocolos se uso en la primera parte del proyecto, para una explicación más detallada ver el Capítulo 6.
- WSDL: Lenguaje para los servicios web con el que establecemos la comunicación con estos, en el se especifican los datos del servidor y los servicios que este ofrece, determinando los

### 3. CONCEPTOS TEÓRICOS



tipos de datos de respuesta y petición, los cuáles están establecidos en la especificación SOAP. Este lenguaje esta basado en XML. Este lenguaje se uso en la primera parte del proyecto, para una explicación más detallada ver el Capítulo 6.

- REST: Protocolo que usa Http para establecer la conexión con el servidor y que, gracias a los distintos tipos de peticiones que posee, puede realizar las distintas operaciones en función de lo servicios que aporte el servicio web.

## 3.2 Conceptos teóricos en el lado del Cliente

### 3.2.1 Funcionamiento de las librerías Text2Speech

Se usó una biblioteca para que el dispositivo lea las frases deseadas en voz alta, para esto se usa una librería del tipo *text to speech*. Este tipo de bibliotecas lo que hace es procesar los datos que se quieren leer y convertirlos en un clip de audio, en el que podemos escuchar una voz que lee lo deseado.

Para entender el funcionamiento de este tipo de bibliotecas nos vamos a ayudar del gráfico que se muestra en la Figura 3.5, además se procede a poner una serie de pasos a través de los cuáles tienen que pasar los datos para llegar al clip final:

- En primer lugar se almacena el texto en la memoria del dispositivo en el que se va a procesar con esta biblioteca.
- Se procede a aplicar una serie de reglas de análisis léxico para convertir el texto en un conjunto de componentes de pluralidad.

- Se asocia la información de pronunciado y de significado a esta serie de componentes.
- El analizador fonético enmarca el texto usando las reglas de análisis fonético.
- Guardamos el conjunto de sonidos en memoria, cada sonido guardado es asociado con alguna información de pronunciación.
- Se recogen los sonidos asociados a los componentes para generar una fila o conjunto de sonidos que se asocian con el analizador fonético y con las reglas de análisis expresivo para generar la salida final.

### **3.3 Conceptos teóricos asociados a los artículos de investigación**

En apartados posteriores en la documentación (ver Capítulo 5) se explicará de manera más o menos detallada los artículos de investigación asociados a las herramientas que se han utilizado, dentro de estos artículos tenemos conceptos teóricos algo avanzados, que se considera oportuno explicar en esta apartado para el posterior entendimiento de dichos artículos.

#### **3.3.1 Convolutional Neural Networks o Redes Neuronales Convolucionales**

En este apartado de los conceptos teóricos se procederá a explicar las Redes Neuronales Convolucionales[11], las cuáles son muy importantes debido a que en el tratamiento de imágenes es una herramienta básica para el procesado de las mismas.

Las Redes Neuronales Convolucionales, CNNs a partir de ahora, son muy similares a las Redes Neuronales Multicapa. Poseen varias capas conectadas totalmente entre si, están formadas por un conjunto de neuronas artificiales, poseen pesos y umbrales que cambian en el entrenamiento de la red. Cada neurona recibe algunas entradas y obtiene el producto escalar de estas. La red entera posee una función objetivo. También posee una función de pérdida y se aplican todos los pasos normales que posee una Red Neuronal Multicapa común.

Parece que las CNNs no tienen nada nuevo en principio, pero la diferencia está en que la CNN lo que hace es dar por hecho que el input de la red es una imagen. Dar por hecho que la entrada de la red es una imagen nos aporta grandes beneficios, puesto que podemos configurar el resto de la red para tratar exclusivamente con este tipo de dato.

#### **■ Visión Global de la arquitectura de las CNNs**

Las CNNs toma ventaja de que el dato de entrada son imágenes y no otra cosa, lo que hace que la arquitectura de este tipo de redes sean restringidas de una manera más sensata. En particular, a diferencia de una red neuronal normal, las CNNs tienen sus neuronas organizadas en capas que conforman tres dimensiones: Altura, Anchura y Profundidad (la profundidad hace referencia al número de canales que tiene la imagen de entrada, pudiendo ser inclusive uno).

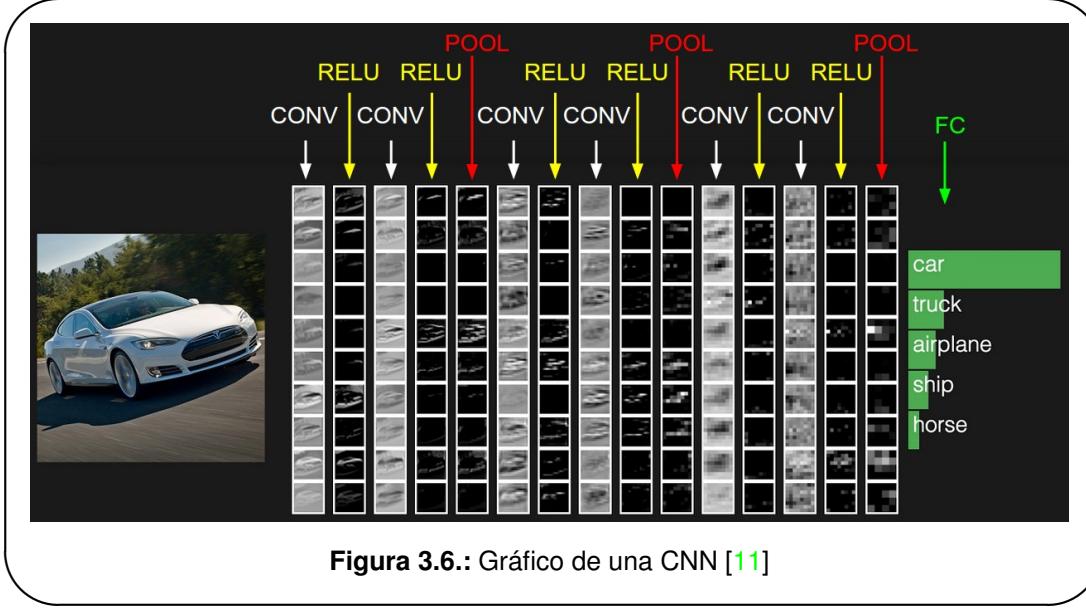


Figura 3.6.: Gráfico de una CNN [11]

### ■ Ejemplo de una CNN

En cada capa de una CNN un volumen de activaciones (la entrada de la capa, que pasa a través de las neuronas y su función de activación) es transformado en otro volumen con el uso de una función diferencial. Las CNNs tienen tres tipos principales de capas: **Capa Convolucional**, **Capa de puesta en común o de Pooling** y **Capa Completamente Conectada**. Si juntamos estas tres capas, entonces formamos una CNN completa.

Vamos a ver un pequeño ejemplo, aunque se detallará la explicación de las capas más adelante. Este ejemplo será para la clasificación de una imagen y tiene una arquitectura [INPUT-CONV-RELU-POOL-FC] y unas 10 clases. La explicación de las capas será:

- **INPUT[32X32X3]:** Esta capa es la de entrada y contendrá los píxeles de la imagen que se vaya a procesar, la imagen tendrá una resolución de 32 de altura por 32 de anchura; además podemos ver que tendrá 3 canales, que podrían ser RGB.
- **CONV:** Esta capa computará las salidas de cada neurona que están conectadas a regiones locales de la entrada, cada computación realiza un producto escalar entre los pesos y la región a la que están conectadas en el volumen de entrada. El resultado dará una matriz de tamaño [32X32X12].
- **RELU:** Esta capa aplicará una función de activación, que podría ser como la función  $\max(0,x)$  con umbral cero. Esta capa no afectará al tamaño de la matriz resultante.
- **POOL:** Esta capa hará un cómputo que reducirá la resolución de la matriz, lo que deja la matriz con tamaño [16X16X12].
- **FC:** (*fully-connected* o completamente conectada) Esta capa los resultados de clases, lo que provoca una matriz de tamaño [1X1X10], donde cada uno de los 10 números se corresponde con el resultado para cada clase.

Por lo tanto, la CNN transforma la imagen original en una matriz de resultados para cada clase, que contiene la probabilidad de que la imagen de entrada pertenezca a esa clase. Para verlo de manera gráfica podemos ver la Figura 3.6.

### 3.3.2 Recurrent Neural Networks o Redes Neuronales Recurrentes

En el siguiente apartado se procede a explicar las Redes Neuronales Recurrentes [12], RNN a partir de ahora. Esta clase de redes se usan en varios de los artículos que se procederá a explicar en el apartado de Estado del Arte (5).

Las RNNs se consideran un caso especial de redes neuronales también, esto se debe a que en general las redes neuronales tienen un conexión entre las capas de tal manera que la comunicación entre capas se hace sólo hacia delante. Por otro lado, las RNNs permiten conexiones recurrentes entre las distintas capas, esto significa que la información ya no viaja sólo en dirección hacia delante sino que existe una propagación de esta información hacia atrás (*back-propagation*). El hecho de permitir que exista este tipo de conexiones entre las distintas capas de la red neuronal añade un elemento temporal a la red, entonces esta puede predecir eventos adelante en el tiempo.

Cuando hablamos de una RNN, entonces estamos hablando de un tipo de red cuyo dato de entrada será siempre un vector, osea una secuencia de datos. La salida de esta red es también un vector de datos. En principio pudiera llegar a parecer extraño que tanto los datos de entrada como los de salida sean tratados como vectores o secuencias de datos, pero esto nos permite que la red procese los datos de una manera secuencial.

#### ■ Visión global de la arquitectura de las RNNs

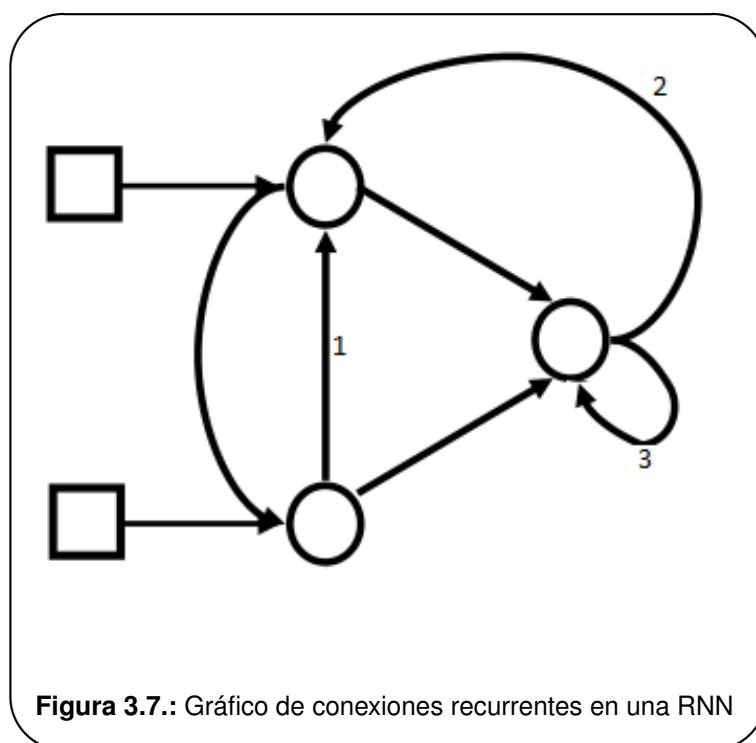
La arquitectura de las RNNs no tiene diferencia, en cuanto a capas se refiere, con las redes neuronales normales. La redes recurrentes poseen las mismas características que estas, estas tienen una función de activación dentro de la neurona, tienen capa de entrada, capas ocultas y capa de salida, poseen una serie de pesos que van cambiando en función del entrenamiento de la red.

La diferencia de una RNN respecto de una red neuronal común o básica se encuentra en el modo en el que están conectadas las neuronas con respecto al resto de neuronas. En las redes básicas las neuronas están conectadas solamente con las capas posteriores, lo que significa que la propagación de la información se hace únicamente hacia a delante sin que exista ningún tipo de retroalimentación. En cambio, las neuronas de una RNN posee también conexiones recurrentes, las conexiones recurrentes pueden ser de tres tipos, se puede ver de forma gráfica en la imagen 3.7:

1. Una conexión de una neurona de una capa A con otra neurona de la misma capa A
2. Una conexión de una neurona de una capa A con otra neurona de otra capa B, pero que se encuentra en un instante de tiempo anterior a esta, osea, se produce una propagación hacia atrás de los datos.
3. Una conexión de una neurona de cualquier capa con ella misma, una conexión a ella misma.

El hecho de añadir todos estas posibles conexiones también provoca que se añadan más pesos para formalizar la red y se necesite el uso de más memoria, además de que los pesos de las capas en una conexión recurrente son guardados en un estado intermedio al cuál acceden las neuronas objetivo de la conexión recurrente para producir predicciones temporales.

### 3. CONCEPTOS TEÓRICOS



## 4. TÉCNICAS Y HERRAMIENTAS

---

En este capítulo se indicarán las técnicas y herramientas utilizadas durante la realización del proyecto.

### 4.1 Técnicas de desarrollo

En esta sección se indicarán las técnicas de desarrollo utilizadas a lo largo del proyecto.

#### 4.1.1 Metodología de desarrollo ágil

Para llevar a cabo este proyecto hemos seguido una metodología ágil de desarrollo de proyectos.

El uso de una metodología ágil de desarrollo viene justificada por el hecho de que los datos respecto a los últimos años, la inmensa mayoría de los proyectos informáticos que se desarrollaron con una metodología clásica de gestión de proyectos han fracasado o no se han llegado a terminar. Ante este problema, se ha ideado una nueva e innovadora metodología de gestión de proyectos, esta es llamada metodología de desarrollo ágil.

La metodología de desarrollo ágil tiene como objetivo el evitar el fracaso de los proyectos, para ello se pretende que los proyectos sigan un nuevo método de desarrollo llamado *Scrum*. Las principales características del *Scrum* [18] son:

- Adopts a strategy of incremental development in contrast to a full execution of the product as is done in classical project management.
- Focuses more on the knowledge of people in self-organized teams than on the quality of the processes used throughout the project.
- In agile development we can observe a clear overlap of development phases, while in classical management these occur sequentially or in cascade.

Al usarse la metodología de desarrollo *Scrum* podemos dividir el proceso en varias partes:

- In the first place, development has been guided through iterations, which have been delimited as *sprints*. These *sprints* have a duration of 1 or 2 weeks.
- In each *sprint* a series of tasks and objectives are proposed to be completed before the end of the same.
- At the end of each *sprint*, a meeting has been held where the objectives achieved in that *sprint* were established, this has been reflected in the project planning as *Retrospective Meeting*; subsequently, new objectives were proposed for the next *sprint*, this is represented in the project planning as *sprint planning*.

Creemos que gracias al uso de este tipo de metodología el proyecto tiene mayores posibilidades de terminar de manera exitosa.

### 4.1.2 Desarrollo software con control de versiones

El control de versiones en desarrollo software se puede describir como la gestión de los cambios que se producen en nuestro software o en la configuración del mismo. Entendiéndose como versión el estado en el que se encuentra el software en un determinado momento.

En el proyecto se ha usado el control de versiones sobre el software a programar por parte del alumno. Para realizar el control de versiones se ha creado un repositorio público en GitHub<sup>1</sup>.

En GitHub las versiones se van determinando a través de *commits*. Lo *commits* son una actualización del estado actual del proyecto, estas actualizaciones llevan un título y un comentario para poder identificar los cambios del software y qué documentos han sido añadidos en qué *commit*.

En el proyecto la asiduidad de los *commits* no sigue ninguna pauta, sino que simplemente cuando se determina que se ha producido un cambio en el software lo suficientemente importante, se realizaría un *commit* con un comentario descriptivo del cambio que se ha producido.

Lo interesante del uso de este tipo de metodología, junto con la herramienta, es que se pueden extraer gráficos y datos que nos aportan información bastante representativa de cómo ha ido evolucionando el proyecto en el aspecto software. También nos asegura que los cambios realizados no se pierden en caso de que la máquina falle, o en caso de habernos equivocado al hacer un *commit* este puede ser deshecho.

En conclusión el uso de control de versiones aporta muchas ventajas y datos con los que posteriormente podremos analizar más en profundidad la evolución del proyecto y usarlo como *feedback* para posteriores proyectos que se vayan a realizar.

## 4.2 Herramientas utilizadas

En este apartado se mostrarán las distintas herramientas utilizadas para el desarrollo del proyecto.

### 4.2.1 Gestor de Tareas: VersionOne

Se han estudiado varias posibles herramientas, entre ellas están:

- **PivotalTracker**<sup>2</sup>
- **FogBugz**<sup>3</sup>
- **VersionOne**<sup>4</sup>

<sup>1</sup><https://github.com/garfio1/Proyecto-Fin-de-Grado>

<sup>2</sup><http://www.pivotaltracker.com/>

<sup>3</sup><https://www.fogcreek.com/FogBugz/>

<sup>4</sup><http://www.versionone.com/>

Se ha optó por la herramienta VersionOne<sup>2</sup>, que a priori ofrecía unas condiciones notablemente mejores a las otras en su versión gratuita y además resulta bastante intuitiva y fácil de usar.

Con esta herramienta nos encargaremos de generar los Sprints del proyecto y se gestionarán las tareas dentro de cada uno. Esta herramienta es usada en el desarrollo ágil, que es el tipo de desarrollo que se llevará acabo en el proyecto, además de que con la herramienta podemos, posteriormente, generar una serie de gráficos e informes que nos serán de gran ayuda a la hora de documentar el proyecto y de gestionar el avance del mismo.

#### 4.2.2 Gestor de Versiones: GitHub

Se han estudiado varias posibles herramientas, entre ellas están:

- **GitHub**<sup>1</sup>
- **Bitbucket**<sup>5</sup>
- **Sourceforge**<sup>6</sup>

Finalmente se decidió que se iba a usar la herramienta GitHub<sup>1</sup> porque se tenía experiencia previa en el uso de la misma, ofrece unas condiciones bastante razonables en su versión gratuita y se puede hacer un buen seguimiento del proyecto con ella. Además de que es bastante sencilla la generación de commits en Windows, porque dispone de una aplicación con la que *commits* se realizan de manera sencilla. Se puede ver de una manera bastante clara cómo ha ido avanzando el proyecto y se pueden extraer unos gráficos muy útiles a la hora de documentar y determinar cómo ha avanzado el proyecto. Además al estar en su versión gratuita, los miembros del jurado y cualquier persona que desee el acceso al proyecto sólo necesitará el link del mismo para poder verlo y comprobar la asiduidad de los *commits* y cómo ha sido la evolución del proyecto (<https://github.com/garfio1/Proyecto-Fin-de-Grado>).

#### 4.2.3 IDE de desarrollo: Android Studio

Se han estudiado varias posibles herramientas, entre ellas están:

- **Eclipse**<sup>7</sup>
- **Android Studio**<sup>8</sup>

La elección de Android Studio<sup>3</sup> ha sido porque no sólo es una herramienta exclusivamente dedicada a aplicaciones Android, sino que resultaba más prometedora que Eclipse; la cuál pensamos que puede quedar obsoleta para este tipo de aplicaciones. También vemos que Android Studio<sup>3</sup> ofrece un gestor de instalación de paquetes de *kit* de desarrollo y *plugins*, lo cuál puede resultar muy útil a la hora de la configuración del entorno de desarrollo. Además, en Android Studio<sup>3</sup> no puede sólo usarse máquinas virtuales de móviles, sino que se puede conectar un dispositivo móvil

---

<sup>5</sup> <https://bitbucket.org/>

<sup>6</sup> <http://sourceforge.net/>

<sup>7</sup> <https://eclipse.org/>

<sup>8</sup> <http://developer.android.com/sdk/index.html>

al ordenador e ir ejecutando tu aplicación sobre el contando con un depurador y un *logcat* para errores, lo cual facilita en gran medida la programación.

### 4.2.4 Herramienta de Generación de Documentación: **T<sub>E</sub>XMaker**

Esta herramienta ha sido elegida por su interfaz gráfica, la cual es muy intuitiva y fácil de usar. Además de que es muy fácil de instalar y tiene ayuda en todo momento, pues autocompleta las etiquetas que se quiera usar y las referencias. Además, su consola de errores ayuda en gran medida a la hora de encontrar errores en los documentos. También permite el uso de proyectos con más de un documento de tipo **T<sub>E</sub>X**, dando la opción de marcar uno de ellos como documento maestro.

Por debajo esta herramienta esta usando **L<sup>A</sup>T<sub>E</sub>X**, que es un sistema para preparar documentos de manera sencilla y fácil. El objetivo de este sistema es que nosotros nos concentremos en escribir el contenido de manera adecuada y que no se nos olvide ningún detalle, dejando a **L<sup>A</sup>T<sub>E</sub>X** que se encargue de optimizar el documento para que tenga la presentación más óptima y adecuada.

En conclusión, sin duda este tipo de sistema y herramienta es uno de los más adecuados para el trabajo que vamos a realizar.

### 4.2.5 Herramientas de Deep Learning: **NeuralTalk**

Se han estudiado varias posibles herramientas, entre ellas están:

- **Lib CCV**<sup>9</sup>
- **Overfeat**<sup>10</sup>
- **Deep Belief SDK**<sup>11</sup>

Esta herramienta, NeuralTalk, ha sido elegida porque venía en el mismo idioma en el que se iba a programar el servidor, además de que su aplicación en nuestro proyecto era mucho mas práctica, pues devuelve una frase descriptiva de lo que en una imagen hay. Lamentablemente, su funcionamiento e instalación son algo enrevesadas y requiere de una configuración que depende mucho de la máquina en la que se trabaje y de la estructura de directorios que esta tenga, pero una vez configurada su funcionamiento es el esperado.

Aunque cabe destacar que se uso la herramienta DeepBeliefSDK<sup>4</sup> en gran parte del proyecto, puesto que sus pruebas fueron correctas y funcionaba de manera bastante adecuada. Pero finalmente desistimos de su uso como herramienta principal debido a que al intentar combinarlo con la herramienta GSOAP<sup>12</sup>, acabo no teniendo un correcto funcionamiento debido a que el módulo de GSOAP<sup>5</sup> para Apache no era bueno y su documentación bastante floja. Esto está explicado de forma mucho más detallada en el apartado de Aspectos Relevantes (6.1.2).

---

<sup>9</sup><http://libccv.org/post/with-a-sub-10-image-classifier-a-decent-face-detector-here-comes-ccv-0.7/>

<sup>10</sup><http://cs.stanford.edu/people/karpathy/rnn/>

<sup>11</sup><https://github.com/jetpacapp/DeepBeliefSDK>

<sup>12</sup><http://www.cs.fsu.edu/~engelen/soap.html>

#### 4.2.6 Herramientas de Deep Learning: Caffe

Esta herramienta [9] es usada en nuestro proyecto debido a que la herramienta NeuralTalk tiene dependencia de este proyecto para extraer características de las imágenes y luego poder realizar una predicción con ellas. Además este proyecto está bien estructurado y, si se sigue correctamente su documentación, es relativamente fácil de instalar.

Caffe es un arquitectura para el *Deep Learning* que está programada en C++ puro y en CUDA, pero también se puede usar en línea de comandos en un sistema Unix y cuenta con *wrappers* para Python y MATLAB. Sus principales características que lo convierten en un proyecto que destaca sobre los demás son:

- Es una arquitectura que funciona de manera especialmente rápida.
- Su código está bien testeado.
- Tiene buenas herramientas, modelos de referencia, demos y repositorios.
- Posibilidad de ejecutarlo tanto en CPU como en GPU

#### ■ Anatomía de un modelo Caffe

Las *deep networks* o redes profundas son modelos compositivos que se representan de forma natural como un conjunto o una colección de capas interconectadas que trabajan sobre fragmentos de datos. Caffe define una red capa a capa en su propio esquema de modelo. La red define el modelo entero desde abajo hasta arriba a partir de unos datos de entrada. Como los datos y sus derivados fluyen a través de la red en el *Forward and Backward*, Caffe guarda, comunica y manipula la información como burbujas: la burbuja es una matriz estándar y una interfaz de memoria unificada para el *framework*. La capa que sigue es tratada como la base tanto del modelo como de la computación. La red se considera como un conjunto de capas y de sus interconexiones. Los detalles dentro de cada burbuja describen como la información será guardada y cómo esta será enviada a través de las distintas capas y redes.

La forma en que se resuelve la predicción es configurada a parte para desacoplar el modelo de la optimización del mismo.

#### 4.2.7 Herramientas de programación: MATLAB

Tuvimos que instalar MATLAB porque NeuralTalk usa un script de MATLAB para poder preparar las imágenes para extraerles las características, además se usa el wrapper de caffe para MATLAB.

#### 4.2.8 Herramientas de desarrollo de servidores: Flask

Se han estudiado varias posibles herramientas, entre ellas están:

- Axis2/C<sup>13</sup>

---

<sup>13</sup><http://axis.apache.org/axis2/c/core/>

- **GSOAP<sup>5</sup>**
- **Tomcat<sup>14</sup>**

Se ha escogido Flask en concreto porque sobre todas las demás su funcionamiento era muy inmediato y además se escribe en Python, que es un lenguaje de programación muy versátil y fácil de usar. El hecho de que esta herramienta tenga un funcionamiento y programación tan sencilla la hace una herramienta que, a nuestro parecer, destaca sobre el resto y es interesante trabajar con ella. Además tiene una documentación sencilla, repleta de ejemplos y explicaciones para poder realizar tu servidor. Sus ejemplos son muy útiles y funcionan a la primera, sin tener que configurar casi nada. Tiene una forma de establecer los URLs de forma muy sencilla y es fácil estructurar tu servidor con esta API. A pesar de ser una API tan fácil de usar tiene una gran potencia y se pueden construir con ella servidores bastante grandes y fácilmente escalables, por tanto se convierte en la herramienta perfecta para la programación de nuestro servidor.

### 4.2.9 Biblioteca Text to Speech para Android

En nuestra aplicación cliente hemos usado la biblioteca *Text to Speech* que nos ofrece Android para realizar la lectura de mensajes para el cliente.

El funcionamiento de este tipo de bibliotecas es parecido, se trata de un problema de procesadores de lenguajes. Primero se analiza la entrada a través de una serie de reglas para acabar asociando estas a unos sonidos de salida, estos acabarán constituyendo el clip de audio con la frase esperada. Esto se puede ver explicado de forma más detallado en el apartado de Conceptos Teóricos (ver 3.2.1).

### 4.2.10 Biblioteca de Apache para conexiones Http para Android

Para establecer la conexión con el Servidor se ha usado la biblioteca de Apache, que nos permite establecer una conexión con un servidor, a través del protocolo HTTP, de manera muy sencilla.

Se ha usado este tipo de biblioteca debido a que el servicio web que se va a programar usa el estándar REST para aportar los servicios y realizar el intercambio de datos con los distintos clientes que se conecten a este. Este tipo de conexión y la explicación de servicio web se puede ver más detallada en el apartado de conceptos teóricos (ver 3.1.3).

Esta biblioteca tiene una serie de elementos que se utilizan dentro de la aplicación Android para que todo funcione de manera correcta, los elementos son los siguientes:

- **HttpClient**: Este objeto se usa para establecer el cliente que ejecutará la petición. Es un objeto que te representa como cliente que se conecta al servidor.
- **HttpPost**: Este objeto está hecho para determinar el tipo de operación que se va a solicitar al servicio web, en este caso es del tipo POST. Este objeto llevará asociada una cabecera con la que se establece la conexión.
- **MultipartEntityBuilder**: Este objeto maneja los datos que queremos enviar con nuestra petición, en nuestro caso será una imagen. Después de haberle añadido los datos a este objeto, podemos crear a partir de él un objeto de tipo **HttpEntity**.

---

<sup>14</sup><http://tomcat.apache.org/>

- `HttpEntity`: Este objeto es una entidad que lo que realmente tiene en su interior es la parte de la cabecera de la petición donde van definidos los datos, osea los metadatos de la cabecera de una petición HTTP.
- `HttpResponse`: En este objeto recibimos la respuesta que el servidor nos da tras ejecutar la operación que hayamos definido.

Como nota final cabe destacar que el procesamiento de todo esto debe hacerse en un hilo separado del hilo principal de ejecución de la aplicación debido a que los estándares de Android así lo establecen.

#### 4.2.11 Biblioteca de traducción de texto

En el lado del servidor se procede al uso de una biblioteca de traducción porque la herramientas nos devuelven las cadenas en inglés, pero nosotros las queremos en español.

En un principio se optó por intentar usar la biblioteca de Google para la traducción de texto, pero nos encontramos con el inconveniente de que esta era de pago. Entonces, ante la búsqueda de una solución, se procedió a usar un proceso algo más rudimentario pero igualmente válido. Como no se tenía acceso a una API de traducción se usaron los conceptos aprendidos de servicio web y de sus peticiones para simular una conexión a la página de traducción de Google y recibir de ella la cadena traducida.

El resultado se obtiene en un tipo de dato json pero este es fácilmente convertible a texto y de ahí se extrae la cadena ya traducida.



## 5. ESTADO DEL ARTE

---

En este capítulo se procede a la presentación de una serie de artículos que han conformado el estudio teórico que conlleva este proyecto, el cuál tiene un gran peso en el mismo; pues el estudio de cada una de las herramientas que se han tenido en cuenta tiene por detrás un concienzudo estudio por parte del alumno, que ha facilitado al alumno la compresión del funcionamiento de estas.

Para realizar la explicación de los artículos, los contenidos de estos se presentan de manera resumida, ordenada y eliminando los aspectos que sean demasiado avanzados o que no se consideren del todo necesaria su explicación en este proyecto.

### 5.1 Artículo del DeepBelief SDK

El artículo de investigación que se va a explicar es el que se ha usado para implementar la herramienta *Deep Belief SDK*<sup>1</sup> y se puede encontrar con el nombre de *ImageNet Classification with Deep Convolutional Neural Networks* [14]. Lo siguiente pretende ser un resumen explicativo del artículo, dónde se tratarán los temas más interesantes del mismo.

#### 5.1.1 Introducción

Los enfoques actuales sobre el reconocimiento de imágenes ha hecho esencial el uso de técnicas de *Machinne Learning* para la resolución de este tipo de problemas. Para mejorar los rendimientos que actualmente se pueden encontrar, se podrían recolectar conjuntos de datos más grandes, entrenar modelos más potentes y usar mejores técnicas para evitar el sobreajuste del modelo. Hasta hace poco se contaban con conjuntos de datos (*datasets*) relativamente pequeños, del orden de diez mil imágenes. Las tareas de reconocimiento simples pueden ser resueltas lo suficientemente bien con *datasets* de este tamaño. Pero ahora existen tareas más complejas y la posibilidad de trabajar con *datasets* bastante más grandes. El *datasets* nuevo más largo está incluido en LabelMe [17], el cuál consiste en un conjunto de imágenes de alta resolución con sus predicciones (*labels*) y clasificadas en más de veintidós mil categorías.

Para poder entrenar tal cantidad de datos es necesario un modelo lo suficientemente grande y capaz de predecir esa cantidad de categorías. De todas formas, la inmensa complejidad del reconocimiento de objetos significa que este problema no puede ser especificado incluso por un *dataset* tan largo como lo es ImageNet. Esto significa que los autores del proyecto tuvieron que contar con conocimiento a priori para compensar todo los datos que no tenían disponibles dentro del *dataset*. Un modelo que encaje en este tipo de descripción es, sin lugar a dudas, una Red Neuronal Convolutacional (CNN) (ver sección 3.3). Su capacidad puede ser controlada variando su amplitud y su profundidad y, además, crean fuertes y, en su mayoría, supuestos correctos sobre la naturaleza de las imágenes.

---

<sup>1</sup><https://github.com/jetpacapp/DeepBeliefSDK>

A pesar de las cualidades tan atractivas de las CNNs, estas no trabajan bien con imágenes de alta resolución porque resulta demasiado cara este tipo de ampliación. Pero pudieron resolver este problema debido a que las GPUs (tarjetas gráficas) actuales vienen con una implementación altamente optimizada de convolución 2D, que son lo suficientemente potentes para facilitar el entrenamiento de CNNs particularmente grandes.

La red está entrenada con un subconjunto de datos de ImageNet usados en las competiciones ILSVRC-2010 y ILSVRC-2012 [2] y se han logrado resultados bastante mejores que los que han sido obtenidos hasta entonces sobre ese conjunto de datos. El modelo del artículo posee una implementación altamente optimizada para GPU de convolución 2D y todo el resto de operaciones que internamente se necesitan con las CNNs. La red contiene algunas características inusuales que mejoran el rendimiento y reducen el tiempo de entrenamiento. El tamaño de la red convierte al sobreajuste en un problema bastante serio, para solucionar esto se usan métodos para prevenir el sobreajuste. La red final contiene cinco CNNs y tres capas totalmente conectadas, esta profundidad parece ser importante ya que si se aumenta o disminuye el número de CNNs, entonces el rendimiento se reduce.

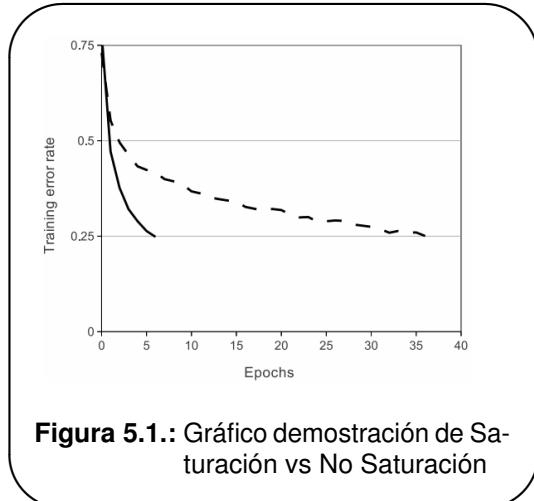
En el artículo se cuenta que un gran inconveniente se encuentra en que el tamaño de la red está limitado por la cantidad de memoria disponible en las GPUs actuales y por la cantidad de tiempo de entrenamiento se están dispuestos a tolerar. La red tarda de cinco a seis días para ser entrenada sobre dos tarjetas gráficas GTX 580 3GB. Todos los experimentos que realizaron apuntan a que los resultados pueden ser mejorados con la mejora de las GPUs, haciéndolas más rápidas, y con *datasets* más grandes.

### 5.1.2 El dataset o conjunto de datos

Se usa el un subconjunto del *dataset* ImageNet, que contiene sobre los quince millones de imágenes de alta resolución y están clasificadas con alrededor de veintidós mil categorías. El subconjunto que se usa es el que se utiliza en la competición llamada *ImageNet Large-Scale Visual Recognition Challenge* (ILSVRC), y este contiene uno coma dos millones de imágenes de entrenamiento, cincuenta mil imágenes de validación y ciento cincuenta mil imágenes de *test* o prueba.

ILSVRC-2010 es la única versión del ILSVRC que tiene disponibles las etiquetas de las imágenes, por lo tanto este es el conjunto de datos sobre el que se han realizado la mayoría de los experimentos. En ImageNet es posible mostrar los errores de dos formas: top-1 y top-5, dónde top-5 es el error sobre las imágenes de test en el cual la predicción correcta no se encuentra dentro de las cinco clases más probables consideradas por el modelo.

ImageNet consiste en un conjunto de imágenes, cuyo tamaño es variable. Sin embargo, para el modelo, se necesitan imágenes con un tamaño fijo, osea, que todas las imágenes tengan el mismo tamaño (ver sección 3.3). Por lo tanto, para solucionar este problema, cambiaron el tamaño de las imágenes a un tamaño común, 256X256. Las imágenes no se preprocesan de ninguna otra manera, excepto para extraer la actividad principal sobre el conjunto de entrenamiento a partir de cada pixel, por lo tanto, las imágenes son tratadas con los valores de sus filas RGB, se trabaja con los tres canales de color.



### 5.1.3 Arquitectura

Contiene en total ocho capas de aprendizaje, de las cuales cinco son convolucionales y tres son capas totalmente conectadas (*fully-connected*).

#### ■ ReLU de no linealidad

El método normal para modelar la salida de una neurona como función aplicada sobre la entrada es:

$$\text{salida} = f(\text{entrada}) \quad (5.1)$$

donde la función de activación utilizada es con la función tangente.

$$\text{salida} = \text{tangente}(\text{entrada}) \quad (5.2)$$

En términos de tiempo de entrenamiento con gradiente descendente, estas saturaciones no lineales son mucho más lentas que si no usáramos saturamiento, con la función máximo.

$$\text{salida} = \text{máximo}(0, \text{entrada}) \quad (5.3)$$

Las neuronas con esta no linearidad se conocen como *Rectified Linear Units* (ReLUs), tal y como se puede ver en Nair and Hinton [15]. Las Redes Neuronales Convolucionales Profundas se entranan en un tiempo considerablemente menor que las que usan la función tangente. Esto se demuestra en la imagen 5.1, donde se ve que entrenando una red pequeña, se necesita un número de iteraciones menor para llegar al 25 % de error de entrenamiento si no usamos el modelo con neuronas con saturación.

El trabajo del artículo no es el primero en considerar el uso de modelos de neuronas diferentes a los tradicionales en las CNNs. Pero el objetivo de estos otros trabajos era distinto y el principal objetivo de este conjunto de datos es prevenir el sobreajuste, el objetivo era distinto al de hacer que la red se entrene de manera más rápida, lo que se pretende en el artículo con sus ReLUs. El aprendizaje rápido tiene una buena influencia sobre el rendimiento sobre el entrenamiento de grandes modelos con grandes conjuntos de datos.

## ■ Entrenamiento en múltiples GPUs

El uso de una tarjeta gráfica GTX 580, tal y como se comenta en el artículo, que tiene sólo 3GB de memoria, limita el tamaño máximo de las redes que pueden ser entrenadas sobre esta. Si se añade el problema de que trabajaron con conjunto de datos con uno coma dos millones de ejemplos de entrenamiento, los cuales son suficientes para que la red resultante sea demasiado grande como para que esta pueda ser entrenada sobre una sola GPU, entonces el modelo era inabordable por una sola tarjeta gráfica. Por lo tanto, el modelo tuvo que ser entrenado sobre dos GPUs. Las GPUs actuales están bien preparadas para la paralelización, puesto que estas pueden acceder a la memoria de otra sin necesidad de pasar por la memoria principal del sistema. La paralelización usada en el modelo, básicamente entrena la mitad de neuronas en cada GPU, pero estas solo pueden comunicarse con capas específicas de la otra GPU. Por lo que la decisión de qué capa se comunicaba con qué otra capa, se convirtió en un problema a resolver, pero esto permitía reducir la carga computacional de la comunicación hasta un valor aceptable.

Como resultado, la arquitectura que obtuvieron es una arquitectura similar la CNN “columnar” empleada por Cireşan [3], la cual tenía en su estructura una capa opcional de preprocesado de imagen, una capa convolucional, una capa de tipo *Max-Pooling* y una capa de clasificación; pero la diferencia es que las columnas del modelo presentado en el artículo no son independientes. Esto reduce el error top-1 en 1,7 % y el error top-5 en 1,2 %.

## ■ Respuesta local a la Normalización

Lo más destacable de este apartado es que al utilizar neuronas de tipo ReLU en el modelo del artículo, la normalización no es necesaria para evitar el saturamiento de las neuronas.

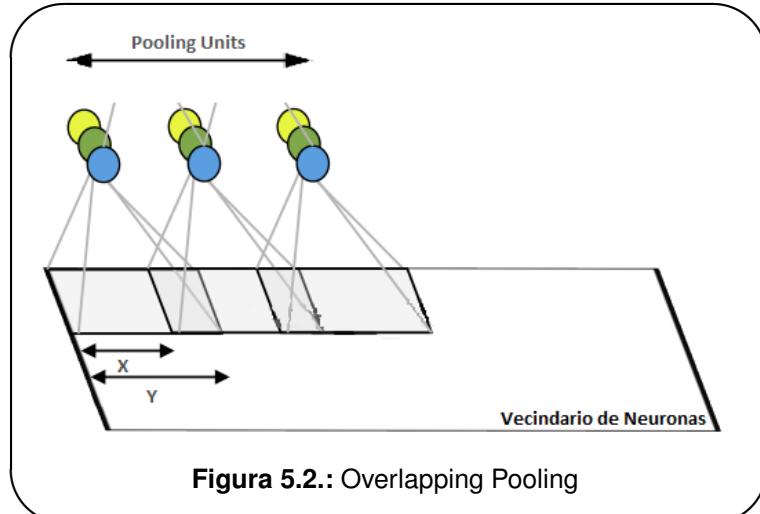
## ■ Overlapping Pooling o Puesta en común superpuesta

El trabajo de las capas de *pooling*, o puesta en común, es la de resumir las salidas de los grupos de neuronas vecinas que le corresponde. Tradicionalmente los vecindarios de neuronas al ser resumidos no se superponían. Para que quede más claro, una capa de *pooling* está formada por una cuadrícula de unidades de *pooling* que están separadas entre sí por un número  $X$  de píxeles, y cada resumen realizado se hace sobre un vecindario de tamaño  $Y \times Y$ . Si la separación entre unidades de *pooling* es igual que  $Y$ , entonces la capa de puesta en común es la tradicional; pero si nos encontramos con que la separación es menor que el valor  $Y$ , entonces las unidades de *pooling* se superponen al resumir vecindarios y así obtenemos una capa de puesta en común superpuesta (ver 5.2).

En conclusión, se obtiene que el error en top-1 y top-5 se redujo en 0.4 % y 0.3 %, respectivamente. Por eso los autores del artículo decidieron usar unos valores de dos para  $X$  y de tres para  $Y$ .

## ■ Arquitectura

Como se ha dicho antes, la arquitectura de la red está formada por ocho capas, de las cuales las cinco primeras con convolucionales y las tres restantes son tres capas completamente conectadas. La salida de la última capa completamente conectada es una matriz de 1000 datos, que se corresponde con las 1000 *labels* de clase que tenemos.



Las capas dos, cuatro y cinco convolucionales están únicamente conectadas con la capa anterior, la cual está situada en la misma GPU. La capa 3, por el contrario, está completamente conectada con la segunda capa. Capas de normalización están después de la primera y segunda capa convolucional. Capas de *pooling* están después de las de normalización y después de la quinta capa convolucional. La ReLU, función de no linealidad, se aplica en todas las capas, tanto convolucionales como las completamente conectadas.

## ■ Reduciendo el sobreajuste

Debido a la cantidad de parámetros de la red neuronal, unos sesenta millones, y a la cantidad de clases que contiene el conjunto de datos, se hace imposible entrenar una red sin tener en cuenta la probabilidad de que se produzca un sobreajuste.

Para evitar el sobreajuste de la red hemos usado dos métodos:

- El primero es aumentar el conjunto de datos original, esto se hace creando nuevas imágenes a partir de las imágenes originales pero conservando su etiqueta. Las dos formas para hacer esto que se ha usado en este artículos son:
  - Hacer traslaciones en la imagen y reflexiones horizontales a través de los cuáles obtenemos una nueva imagen.
  - Se ha aumentado la intensidad de los canales RGB de la imagen para obtener otra imagen que sea distinta para el modelo.
- Se usa un nuevo método llamado *dropout* el que consiste en poner a cero la salida de las neuronas de la capa oculta con una probabilidad de 0.5. Las neuronas que sean abandonadas en este proceso, entonces no participaran en la propagación hacia atrás.

## ■ Detalles del aprendizaje

En el artículo se explica que se inicializan los pesos de cada capa a partir de una distribución Gaussiana con una desviación estándar de 0.01. Los sesgos de las neuronas en las capas dos, cuatro

y cinco los determinaron con un valor constante de 1. Esta inicialización lo que hace, según se comenta en el artículo, es acelerar el proceso de aprendizaje en los pasos tempranos dando a las ReLUs entradas positivas. En el resto de capas los sesgos se inicializan a cero.

Los autores han usado el ratio de aprendizaje igual para todas las capas y este es ajustado de forma manual. La heurística que siguieron es que el ratio de aprendizaje es dividido entre diez cuanto el ratio del error de validación deja de mejorar con el ratio de aprendizaje actual. El ratio de entrenamiento se inicializa a 0.01. La red la entrenaron en 90 ciclos con un conjunto de imágenes de 1.2 millones de tamaño, el cuál tomó 5 o 6 días en finalizar sobre las dos tarjetas gráficas ya mencionadas.

## ■ Resultados

El resultado sobre el dataset ILSVRC-2010 ha conseguido, sobre el conjunto de test, unos ratio de error top-1 y top-5 de 37,5 % y 17 %. Teniendo en cuenta que en la competición de 2010 se consiguió el mejor rendimiento como un 47,1 % y de 28.2 %, se puede decir que este modelo tiene bastante mejor resultados.

## ■ Conclusión

Las conclusiones más importantes que se pueden sacar del artículo es que una red neuronal convolucional profunda, siempre que sea lo suficientemente grande, es capaz de romper los récords sobre los resultados ya existentes. También es interesante el hecho de que si alguna de las capas convolucionales es quitada o añadida a este modelo, entonces su error ratio se incrementa; por alguna razón la profundidad de la red es importante a la hora de mejorar los ratio de error.

## 5.2 Artículo del NeuralTalk

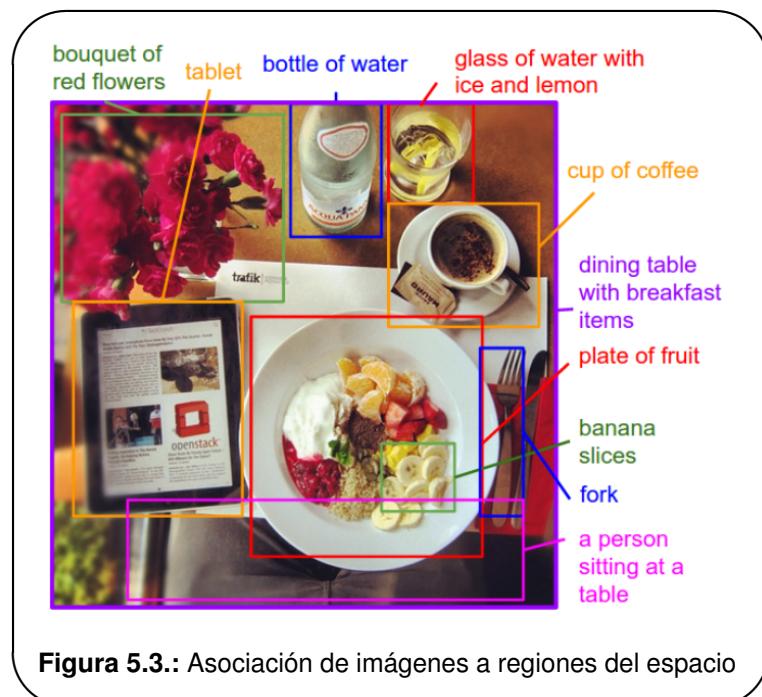
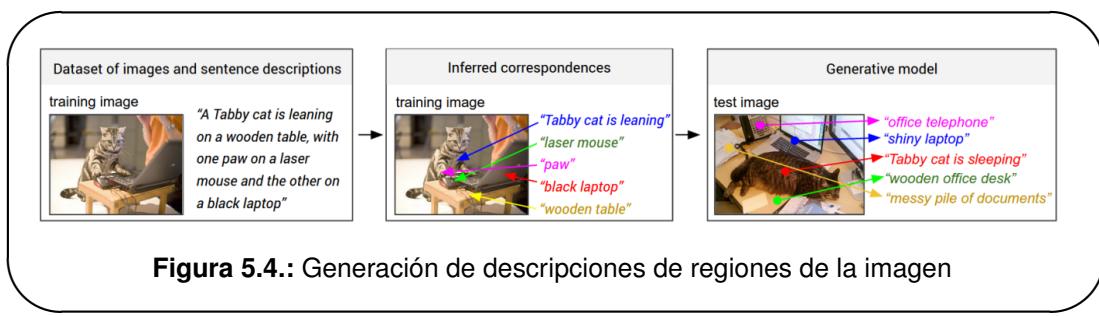
En este apartado se procederá a resumir el artículo asociado a la herramienta NeuralTalk [13].

### 5.2.1 Introducción

Una vistazo rápido es suficiente para el humano para poder extraer una inmensa cantidad de detalles de la escena que este presenciando [8]. Sin embargo esta tarea es muy compleja para nuestros modelos de reconocimiento visual. Hasta ahora los esfuerzos en el reconocimiento de imágenes ha sido el de etiquetar imágenes a través de un conjunto fijo de categorías, el cuál ha progresado bastante. Sin embargo, estos métodos tienen un vocabulario muy restrictivo en comparación con el vocabulario descriptivo que tiene el ser humano.

Algunos pioneros se acercan a resolver el reto de generar descripciones de imágenes. Sin embargo, estos modelos usualmente tienen ciertas deficiencias que les impide alcanzar una gran variedad en el reconocimiento. Por otra parte, el tema central de estos trabajos ha sido reducir la complejidad de las imágenes en una sola sentencia.

El trabajo presentado en el artículo tiene el objetivo de generar descripciones complejas (ver Figura 5.3) a partir de una imagen. El principal reto del trabajo fue diseñar un modelo lo suficientemente rico para a la vez razonar sobre el contenido de la imagen y su representación en lenguaje

**Figura 5.3.:** Asociación de imágenes a regiones del espacio

natural. El segundo reto fue el de encontrar un *dataset* lo suficientemente grande sobre el que trabajar.

La idea central que se propone en el artículo es aprovechar estos grandes conjuntos de datos mediante el tratamiento de las frases como etiquetas débiles, en la que los segmentos contiguos de palabras corresponden a algunos en particular, pero la ubicación es desconocida en la imagen. Para esto, los autores tuvieron que realizar dos aportaciones al modelo, que son las siguientes:

- Desarrollaron un modelo de red neuronal que es capaz de relacionar los segmentos de frases con la región de la imagen que esta describiendo.
- Introdujeron una arquitectura de red neuronal recurrente multimodal que es capaz de generar una descripción en texto a partir de una imagen que toma como entrada.

### 5.2.2 El modelo

El objetivo final del modelo presentado en el artículo es generar descripciones de regiones de imágenes. Durante el entrenamiento, la entrada del modelo es un conjunto de imágenes y sus

correspondientes etiquetas, que son frases en lenguaje natural (ver Figura 5.4). En primer lugar se presenta un modelo que asocia fragmentos de oraciones a regiones de la imagen. A continuación, se tratan estas asociaciones como datos de entrenamiento para una segunda red, que aprende a generar los fragmentos de las oraciones.

## ■ Aprendiendo a relacionar datos visuales con datos de lenguaje

El modelo de alineación necesita una entrada de un conjunto de imágenes con sus respectivas frases descriptivas. La idea principal de este modelo es que las personas puede escribir frases referentes a la imagen, pero no sabemos a que parte de la imagen se refieren estas. Asumieron, entonces, que existe una relación entre la frase y el objeto que se describe de la imagen; por tanto, procedieron a intentar encontrar estas relaciones para posteriormente aprender a generar estos fragmentos de imágenes a los que las oraciones se refieren.

Primero se creó una red neuronal que asocia las palabras con las regiones de la imagen. Entonces, el siguiente objetivo que se propusieron fue relacionar semánticamente estas palabras.

## ■ Representando imágenes

Los autores observaron que las frases hacen referencias frecuentes a los atributos de los objetos que están describiendo. Así, siguiendo el método de Girshick [6] para la detección de objetos en todas las imágenes, usaron una Red Neuronal Convolutacional. La CNN, tal y como se comenta en el artículo, fue pre-entrenada con el *dataset* de ImageNet [4], y afinada sobre las 200 clases del ImageNet Detection Challenge [16].

## ■ Representando frases

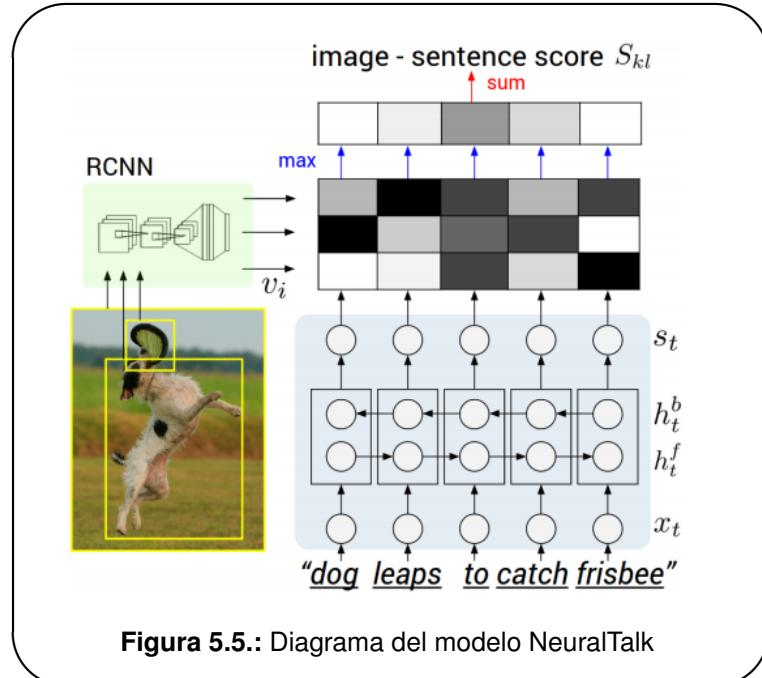
Para establecer las relaciones inter-modales, pensaron que sería conveniente representar las palabras de las frases en el mismo espacio dimensional que ocupan en la región de la imagen. Comentan en el artículo que el enfoque más simple podría ser proyectar cada palabra en este espacio. Debido a que lo anterior presenta ciertos defectos, determinaron que se podría usar una extensión de este método, que fue el uso de bigramas de palabras o el uso de relaciones de dependencia. Sin embargo, esto sigue imponiendo un tamaño máximo arbitrario de la imagen y requiere el uso de Árboles de Dependencia, que debe ser entrenada con texto no relacionado (ver Figura 5.5).

Para conseguir este propósito en el modelo del artículo, se uso una Red Neuronal Recurrente Bidireccional (ver sección 3.3.2) para computar las representaciones de las palabras.

- Objetivo de alineamiento:

Se ha descrito las transformaciones que asocian a cada imagen con la frase en un conjunto de vectores en común, dentro de un mismo espacio dimensional. Como la extracción y la predicción se realizan sobre la imagen y la frase entera, se creó una puntuación imagen-frase como una función de las puntuaciones región-palabra individuales. Intuitivamente, una predicción del tipo imagen-frase, tendrá una puntuación elevada si sus puntuaciones región-palabra son elevadas, osea se relacionan bien y tienen buen soporte.

- Decodificando segmentos de texto alineados a imágenes:



**Figura 5.5.: Diagrama del modelo NeuralTalk**

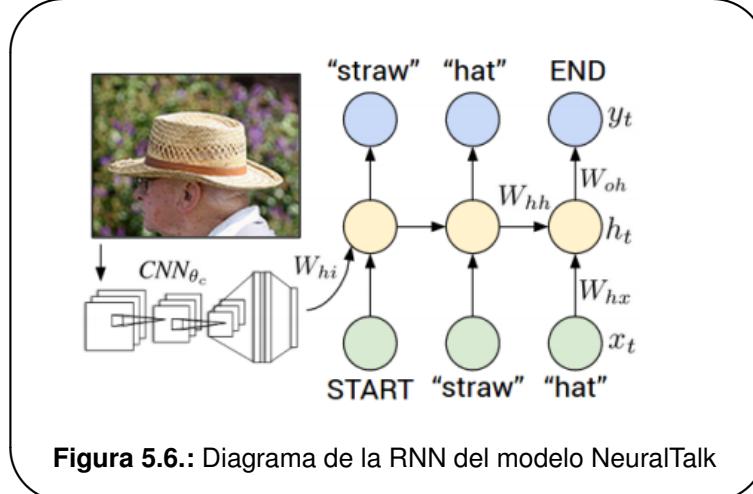
Lo que se pretende es generar secuencias de palabras que estén asociadas a una imagen, no una palabra suelta asociada a la misma; además estas deben estar relacionadas semánticamente entre ellas, por lo tanto se meten en una «caja» para posteriormente generar la secuencia con esa relación semántica exigida.

### ■ Red Neuronal Recurrente Multimodal para la generación de descripciones

En esta sección se toma en cuenta que la entrada sera un conjunto de imágenes y sus correspondientes descripciones. Estas pueden ser imágenes y su frase descriptiva o regiones y fragmentos de texto, como se ha comentado anteriormente. El desafío clave es diseñar un modelo que pueda sacar como predicción a través de una imagen, que consiste en una secuencia variable (la frase que describa la imagen). En anteriores trabajos basados en RNN, esto se conseguía obteniendo una probabilidad de cuál sería la siguiente palabra en una secuencia, teniendo disponible la palabra actual y el contexto anterior (dado por las conexiones recurrentes). En el modelo del artículo se usa una pequeña extensión de estas redes.

- Entrenando la RNN:

LA RNN está entrenada para predecir la siguiente palabra usando la palabra actual y el contexto en el tiempo anterior. Se condiciona las predicciones de la RNN a través de la interacción con su umbral en los primeros pasos de la predicción. El entrenamiento funciona de la siguiente manera: se empieza con una palabra especial para determinar el comienzo de la predicción, y la primera predicción obtenida será la primera palabra de la frase a obtener. La siguiente palabra se traduce tomando la palabra actual como la recién predicha y el contexto anterior, esperando que el modelo prediga la siguiente palabra como la segunda; y así sucesivamente. Cuando se llega a la última palabra, la palabra predicha será una palabra reservada para determinar el fin de la predicción. Para aclarar esto no ayudaremos de la Figura 5.6.



- Testeando la RNN:

Para predecir una frase, se computa la representación de la imagen y se empieza con la palabra reservada que determina el comienzo de la predicción. Se muestra una palabra de la distribución, que se toma como palabra actual y se procede a predecir la siguiente palabra; esto se repite hasta obtener la palabra reservada que determina el fin de la frase.

## ■ Conclusiones

Crearon un modelo capaz de generar frases a partir de imágenes, pero tiene la gran limitación de que estas sólo pueden tener una resolución fija. Por último, hay que tener en cuenta que este modelo en realidad consiste en dos modelos, uno que preprocesa las imágenes para extraer las palabras asociadas a regiones de la imagen y el segundo es la RNN que predice las frases.

## 5.3 Artículo del Arctic Caption

En este apartado se hablará del proyecto Arctic Caption, para ello se va a proceder a resumir el artículo que viene asociado a este [26].

### 5.3.1 Introducción

La generación de frases automáticamente a través de una imagen dada es una tarea muy cercana al corazón del entendimiento de una escena, que es uno de las primeras metas de la visión por computadora. No sólo los modelos que realizan estas operaciones deben ser lo suficientemente potentes para enfrentar la carga computacional que lleva el hecho de detectar qué objetos hay en una imagen, sino que también deben ser capaz de relacionar estos objetos a través del lenguaje natural. Por estas razones la generación de frases ha sido considerado como un problema de extrema dificultad. Es uno de los retos más importantes del *Machine Learning*, el imitar la capacidad humana de comprender grandes cantidades de información visual y transcribirla a lenguaje natural.

A pesar de la complicación que esta tarea conlleva, ha habido un reciente aumento de estudios que están tratando de resolver este tipo de problema. Esto se debe a la mejora en los algoritmos

de aprendizaje máquina y el aumento de conjuntos de datos sobre los que poder trabajar. Trabajos recientes han mejorado la calidad de la generación de frases a través de la combinación de CNNs, que obtienen una representación vectorial de la imagen, y de RNNs, que decodifican las representaciones de las CNN y las convierten en frases en lenguaje natural.

El humano tiene la capacidad de centrar la atención sólo en ciertas partes de una imagen, gracias a esto es capaz de trabajar incluso con imágenes muy desordenadas. En trabajos anteriores se ha usado las CNNs para la extracción de estas imágenes y con esto formar la frase, lo que ha proporcionado un resultado bastante satisfactorio. Pero, esto tiene una desventaja y es la pérdida de información que pudiera llegar a ser necesaria para formar frases aún más descriptivas. Para evitar esta pérdida de información se podrían usar un mayor número de representaciones a bajo nivel. Sin embargo, hay que encontrar un modelo y un mecanismo que sea capaz de trabajar con todas estas características.

En este artículo se explica un intento de insertar una forma de mecanismo de atención en el reconocimiento de las imágenes, este mecanismo tiene dos variantes: un mecanismo de atención «fuerte» y un mecanismo de atención «débil».

Las aportaciones de este artículo son las siguientes:

- Se introduce dos generadores de frase basados en la atención hacia la imagen en un mismo *framework*:
  - Un mecanismo de atención determinista «débil», que puede ser entrenado con técnicas normales de propagación hacia atrás.
  - Un método de atención estocástico «fuerte», que puede entrenarse con el método descrito en el artículo de Ronald J. Williams [25], que se conoce por el nombre de REINFORCE.
- Se muestra cómo interpretar los resultados, visualizando dónde y en qué se centro la atención.
- Finalmente se valida cuantitativamente la utilidad de la atención en la generación de frases. Se hace a través de tres conjuntos de datos.

### 5.3.2 Generación de Frases de Imágenes con un Mecanismo de Atención

#### ■ Detalles del modelo

En esta sección se describe las dos variantes del modelo de atención de las que se ha hablado previamente. La principal diferencia se encuentra en la función de activación de las neuronas.

Para detallar el modelo, se ha dividido la sección en dos partes:

- 1- Codificador: Características convolucionales

El modelo toma una imagen en forma de vector y genera una frase  $Y$  codificada como una secuencia de 1 a  $K$  palabras codificadas.

$$y = \{y_1, \dots, y_c\}, \forall y_i \in \mathbb{R}^K \quad (5.4)$$

En la ecuación,  $K$  es el tamaño del lenguaje, el número de posibles palabras.  $C$  es la longitud de la frase.  $Y$  es la frase formada por una serie de palabras.

En el artículo se ha usado una CNN para extraer el conjunto de vectores de características de la imagen. El extracto de características produce  $L$  número de vectores, cada uno de los cuales es una representación  $D$ -dimensional correspondiente a una parte de la imagen.

$$a = \{a_1, \dots, a_L\}, \forall a_i \in \mathbb{R}^D \quad (5.5)$$

Con la intención de obtener la correspondencia de los vectores de características y las porciones de la imagen, se extrae las características de la capa convolucional más baja a diferencia de los trabajos previos, que usaban una capa totalmente conectada. Esto permite al decodificador centrarse de forma selectiva en ciertas partes de la imagen seleccionando un subconjunto de todos los vectores de características, osea un subvector del vector  $a$  de la ecuación anterior.

- 2- Decodificador: Red de Memoria Largo Corto Plazo (LSTM)

El artículo explica que el modelo usa una LTSM, que produce un frase a partir de un vector de contexto, el estado de la capa oculta anterior y las palabras generadas anteriormente. La LTSM trabaja de manera que, para cada región de la imagen genera un peso positivo que puede ser interpretado como la probabilidad de que la región en la que nos estamos centrándon sea la correcta para generar la siguiente palabra. Este peso es computado por el modelo de atención, para el que se ha usado un Perceptrón Multicapa que está condicionado por el estado de su capa previa. Hay que destacar que el estado de la capa previa varía en función de las palabras que ya han sido generadas.

### 5.3.3 Conclusiones

Se ha propuesto un modelo de generación de frases a partir de imágenes que está basado en un método de atención, el cual puede ser fuerte o débil. Cabe destacar que los autores del artículo esperan que este tipo de trabajos incentiven a futuros trabajos a utilizar una metodología basada en la atención en las imágenes, la cuál está basada en la capacidad de el ser humano de prestar atención a ciertas partes de la imagen.

## 5.4 Aplicaciones existentes relacionadas con el proyecto

Como ya hemos comentado en la introducción (ver Capítulo 1), este proyecto tiene una dimensión teórica y otra práctica, ahora se procederá a comentar aplicaciones similares a la desarrollada en este proyecto para cubrir el estado del arte de la dimensión práctica [5].

### 5.4.1 TapTapSee

TapTapSee es una aplicación gratuita para IOS, la cuál posee una interfaz muy simple. La aplicación cuenta con tres botones, uno en la parte superior derecha de la pantalla, otro en la superior izquierda y el otro botón ocupa el resto de la pantalla, pues es el más importante.

Esta aplicación te permite reconocer objetos. Debes apuntar la cámara en la dirección del objeto a reconocer y tocar dos veces la pantalla, seguidamente una voz te informará de que se ha tomado una foto. La foto se envía a un servidor, donde se busca una coincidencia y, cuando esta es encontrada, se te comunica en voz alta qué objeto se encuentra en la foto.

Es importante destacar que no tienes por qué esperar a que la aplicación termine de reconocer el objeto para tomar otra foto, pues esta te permite tomar hasta cinco fotos de manera consecutiva sin que la primera haya sido reconocida aún. Esta es una utilidad muy destacable, pues convierte a la aplicación en una herramienta más potente.

Esta aplicación usa una combinación de una base de datos de imágenes y humanos para realizar el reconocimiento. Además, la misma foto puede devolver resultados diferentes en distintas ejecuciones del reconocimiento.

En conclusión, esta aplicación es una herramienta útil, pero no es comparable al objetivo que nuestra aplicación prototipo tiene, pues no se trata de un simple reconocimiento de imágenes, sino de la generación de una descripción basada en la imagen que se ha tomado como referencia.

#### **5.4.2 CamFind**

CamFind está desarrollada por la misma empresa que desarrolló la aplicación TapTapSee. Esto es debido a la experiencia que ganaron los desarrolladores al crear TapTapSee y a la cantidad de comentarios que recibieron de la comunidad, que sirvieron como retroalimentación para el desarrollo de CamFind.

CamFind no posee una interfaz tan sencilla como TapTapSee, pero tiene muchas más funcionalidades. CamFind usa el mismo sistema de reconocimiento que TapTapSee, por lo que sus resultados tienden a ser similares.

La utilidad más destacable de CamFind es que a partir de una foto, no sólo reconoce el objeto que se ha tomado, sino que, además, te ofrece información sobre el precio de dicho producto en Internet, comparación de precios, objetos relacionados, etc... Esta utilidad la convierte en una herramienta muy útil y potente.

Este tipo de herramienta es bastante útil, pero ha perdido el objetivo de apoyo a personas con dificultades de visión porque, al tener tantas funcionalidades, su interfaz se hace demasiada completa para que una persona con dificultades de visión pueda usarla con asiduidad y facilidad. Además, sigue sin acercarse al objetivo de nuestra aplicación prototipo, cuya tarea es mucho más compleja que la de este tipo de aplicaciones.

#### **5.4.3 Talking Goggles**

Talking Goggles usa la base de datos de imágenes Goggles de Google para hacer su reconocimiento.

Lo más interesante de Talking Goggles es que puedes poner el modo cámara de vídeo y ejecutar predicciones en tiempo real, aunque la aplicación sólo detectará unos pocos objetos. Además, el nivel de error de Talking Goggles es aún bastante alto, pues hay bastantes objetos que reconoce de manera incorrecta o que, simplemente, no los reconoce.

Cabe destacar que posee una interfaz muy sencilla, que consta de cuatro botones, que dividen la pantalla en cuatro trozos iguales.

Esta aplicación también usa una guía por voz que, en el caso de esta aplicación, es la propia, ha sido desarrollada específicamente para su uso en Talking Goggles.

## *5. ESTADO DEL ARTE*

En conclusión, podemos ver en Talking Goggles una aplicación potente y competitiva con el mercado actual de aplicaciones de este tipo pero, nuestra aplicación renueva todo lo que se encuentra en el mercado actual, ya que ninguna aplicación puede generar descripciones a partir de imágenes. Otro punto importante es que las técnicas y herramientas usadas en la aplicación prototipo apenas pueden llegar a superar el año de edad, por lo tanto, es una aplicación bastante innovadora.

# 6. ASPECTOS RELEVANTES DEL DESARROLLO DEL PROYECTO

---

En este capítulo se introducen los aspectos más relevantes del proyecto.

## 6.1 Dificultades encontradas

Durante el desarrollo del proyecto nos hemos encontrado varias dificultades que han hecho que el proyecto se retrase considerablemente y su avance no haya sido ni fácil, ni rápido.

### 6.1.1 Dificultades con DeepBeliefSDK

En principio se intentó usar DeepBeliefSDK directamente en Android, pero se encontró que fue excesivamente difícil para el alumno usarlo en Android. Aunque viene un ejemplo en Android, este se intentó hacer funcionar a través de la herramienta Android Studio, pero no funcionaba correctamente.

Se intentó en un principio, siguiendo los ejemplos aportados por el autor, hacer una aplicación lo más sencilla posible para un dispositivo Android. No se obtuvo un resultado satisfactorio, pues no se consiguió que se compilara de manera correcta la aplicación. Entonces se procedió al intento de compilar el ejemplo ofrecido en el proyecto, el cuál resultó que tampoco compilaba y era bastante más complejo que la aplicación de prueba inicial como para poder arreglar los errores.

En conclusión vimos que la ejecución en Android de esta aplicación iba a dar muchos problemas y determinamos que se programaría un servidor para que el cliente subiera ahí la foto y la librería se ejecutaría en el lado del servidor. Cabe destacar que después de un par de intentos de instalación del proyecto, los ejemplos, que estaban en la máquina Linux donde se alojaría el servidor, funcionaron de manera adecuada.

### 6.1.2 Dificultades con GSOAP y Apache

Los problemas con GSOAP y Apache son los que más han retrasado al proyecto y han supuesto una dificultad enorme a la hora de llevar a cabo el mismo.

Empezamos con que para usar GSOAP<sup>1</sup> se tuvo que estudiar una serie de cosas para poder adquirir los conocimientos necesarios para usar la herramienta, dichos conocimientos serán listados aquí:

---

<sup>1</sup><http://www.cs.fsu.edu/~engelen/soap.html>

- **XML:**<sup>2</sup> Se tuvo que coger un nivel adecuado en el uso de XML ya que la herramienta GSOAP se basa en el uso de este tipo de archivos como medio de comunicación en las distintas peticiones y respuestas que procesa. Además el XML también es necesario para comprender el funcionamiento de SOAP y de WSDL, los cuales son completamente necesarios para entender el funcionamiento de la herramienta GSOAP.
- **SOAP:**<sup>3</sup> Esta especificación se tuvo que estudiar para comprender el funcionamiento de la herramienta GSOAP y en qué se basaba su funcionamiento, entender el porqué debía funcionar la herramienta y como se realiza la comunicación gracias a ella. Aunque el SOAP no es usado directamente cuando usas GSOAP es necesario conocer esta especificación ya que GSOAP sí que usa WSDL, para el cual tenemos que tener un conocimiento básico, al menos, de SOAP para poder usarlo.
- **WSDL:**<sup>4</sup> Esta otra especificación sí que se usa directamente en la herramienta GSOAP y básicamente con ella vertebras toda la aplicación que vas a hacer, de hecho tienes dos opciones:
  - La primera es usar un archivo WSDL donde especificas las operaciones que el servidor va a realizar, después con la herramienta GSOAP generas todos los *stubs* y documentos necesarios para hacer tu aplicación.
  - La segunda sería a través de un documento de tipo .h o una cabecera de C. Con el cuál también generas los stubs y documentos necesarios para programar tu servidor, entre dichos documentos se encontrará un archivo WSDL que contendrá la especificación de las operaciones que hay dentro del fichero cabecera que hayas usado. Pero incluso en esta opción necesitas entender SOAP y WSDL porque a través de comentarios tienes que especificar características que irán directamente al fichero WSDL, y que serán necesarios para el correcto funcionamiento del servidor.

Una vez se ha estudiado lo anterior, se pasó al estudio de la documentación de la herramienta GSOAP, además del intento de hacer que funcionen sus ejemplos. Cuando se consiguió que funcionarán sus ejemplos se pasó a la programación de un servidor propio, una vez se programó y se hicieron las pruebas de que estaba bien programado, se procedió a intentar que este funcionaría desde un cliente GSOAP. Para que funcionaría con el cliente GSOAP, se hizo una investigación de cómo hacer que el servidor funcionaría en *localhost* y, siguiendo la recomendación que en la documentación de GSOAP encontramos, se instaló Apache y se intentó usar el módulo de Apache para su funcionamiento con GSOAP.

Como conclusión sacamos que, tras una larga investigación y mucho tiempo dedicada a esta herramienta, esta herramienta no tenía la documentación suficiente como para poder hacerla funcionar con Apache y, a pesar de haberlo intentado muchas veces, no conseguimos que el servidor GSOAP programado por nosotros devolviera alguna vez un resultado coherente al cliente. Finalmente, desecharmos la opción de trabajar con esta herramienta y le dimos un giro al proyecto con el que esperábamos tener avances más rápidos y mejores, optamos por la programación de un servidor en Flask.

<sup>2</sup><http://www.w3schools.com/xml/>

<sup>3</sup><http://www.cs.fsu.edu/~engelen/soap.html>

<sup>4</sup>[http://www.w3schools.com/webservices/ws\\_wsdl\\_documents.asp](http://www.w3schools.com/webservices/ws_wsdl_documents.asp)

### 6.1.3 Dificultades en la instalación de Herramientas

En este proyecto nos encontramos con que la instalación de las herramientas que se van a usar en el lado del servidor conllevan una carga de trabajo bastante grande. Un ejemplo claro es, que descartamos la herramienta Caffe por la cantidad de dependencias que esta poseía y su complejidad y, finalmente, la herramienta elegida, NeuralTalk, dependía a su vez de Caffe. Por tanto, la instalación de NeuralTalk conllevaba la instalación de Caffe y su cantidad de dependencias.

#### ■ Instalando Caffe

En un primer intento de instalar Caffe, se fueron instalando una a una las dependencias de este según íbamos encontrando los errores en instalación. Esto llevó mucho tiempo ya que cada dependencia tenía a su vez su manera de instalarse y no siempre era de manera directa y limpia, un claro ejemplo fue OpenCV, el cuál instalamos de manera manual pero a su vez requería otra instalación y la instalación final resultó no ser adecuada.

El problema de la instalación de OpenCV llevo a que la instalación completa de Caffe fallara. Lo que nos obligó a limpiar la máquina de toda dependencia de Caffe instalada y empezar desde cero. Afortunadamente nos topamos con un tutorial de cómo instalar PyCaffe en una máquina virtual, el cuál era lo suficientemente útil como para que nos sirva a nosotros; que realmente necesitábamos instalar MatCaffe (el *wrapper* de Caffe para MATLAB)<sup>5</sup>.

Gracias a que seguimos esos pasos se resolvió el error de instalación del OpenCV, cabe destacar que llegar a este punto lleva un proceso que puede superar fácilmente la hora por la cantidad de dependencias que se deben instalar aquí. El siguiente problema a tener en cuenta es que debíamos instalar MATLAB, por suerte se contaba con una licencia que la Universidad de Burgos había facilitado a los alumnos, de modo que la instalación de esta herramienta no fue realmente muy compleja, aunque puede tardar sobre otra media hora o incluso más.

Una vez se ha pasado el calvario de la instalación de las dependencias llega el proceso de configurar el makefile de Caffe para instalarlo definitivamente en tu máquina. El tutorial anteriormente mencionado nos serviría para algunos aspectos pero la definición de MATLAB la teníamos que hacer nosotros. Se debía definir el directorio en el que se encontraba el ejecutable de MATLAB, o eso ponía en la documentación. La definición del directorio tal y como ponía en la documentación falló, y se tuvo que intentar con distintos directorios hasta que se llegó a la solución correcta. Hay que tener en cuenta que cada intento rondaba los 15 minutos porque había que limpiar la anterior instalación del makefile y rehacer todas las comprobaciones del mismo.

Una vez se ha superado todos estos contratiempos, se puede decir que tienes Caffe instalado en tu máquina y que está listo para ser usado. La dificultad de esto es que no hay una guía paso a paso de todo lo que hay que hacer y, en muchos casos, tienes que investigar por tu cuenta cómo instalar cada cosa, lo que conlleva una pérdida de tiempo enorme. El proceso de instalación está claramente explicado en el anexo Manual del Programador (IV).

### 6.1.4 Instalando NeuralTalk

La instalación de NeuralTalk lleva consigo la carga de instalar Caffe por detrás. Pero una vez instalado no puedes realizar ningún tipo de prueba debido a que no tienes entrenada la red. El

---

<sup>5</sup><https://github.com/BVLC/caffe/wiki/Ubuntu-14.04-VirtualBox-VM>

entrenamiento de la red es inmediato porque la documentación es clara en cuanto a esto, pero el entrenamiento de la red puede tardar bastante tiempo.

Durante el proyecto se intento entrenar la red por nuestra cuenta, pero las horas que esto conllevaban podían llegar a sumar días, incluso llegar a superar la semana. De modo que se tuvo que buscar alguna red pre-entrenada. Por suerte la solución también fue encontrada en la documentación de la herramienta y se descargó una red pre-entrenada<sup>6</sup> para proceder a hacer pruebas con la herramienta.

### 6.1.5 Dificultades con NeuralTalk

El trabajo que se ahorro en la instalación de NeuralTalk lo tuvimos que invertir en hacer que este funcione de manera que nosotros queremos. Puesto que NeuralTalk inicialmente sólo viene predisposto para ser probado con las imágenes que ya tiene y devuelve un fichero HTML como resultado. No había la opción de trabajar sobre imágenes propias.

Mientras que la documentación de instalación de NeuralTalk era sencilla, la documentación para la configuración de este era muy precaria y sólo se podía encontrar a modo de comentario en los ficheros del proyecto. Además de que no encontrabas todo documentado en un sólo fichero, sino que un fichero documentaba una parte y hacía referencia a documentación que se encontraba en un comentario de otro fichero. Esto hizo que el proceso de configuración de NeuralTalk fuera lento y pesado.

En primer lugar se leyó la documentación del fichero que hacía las predicciones. En el ponía que se usaban *scripts* de MATLAB para la preparación de las imágenes, este script cogía el nombre de las imágenes de un fichero de texto llamado `tasks.txt`. El primer problema es que el script dependía de la cómo estaba estructurado los ficheros de NeuralTalk, osea que teníamos que adaptarlo a nuestro caso de uso; para ello se tuvo que leer el código y diferenciar dónde debíamos y qué debíamos cambiar para que se adaptara a nuestro caso de eso. Este proceso fue lento y pesado, porque los comentarios no eran lo suficientemente claros como para poder interpretarlo de manera inmediata, sino que se tuvo que entender el código casi en su totalidad para poder entender qué se debía modificar y por qué. Además, tuvimos que añadirle líneas para que encontrara a Caffe dentro de nuestro sistema, sino, no funcionaba. En este script nos encontramos además que la definición del objeto de Caffe está mal hecha o no funcionaba en nuestra máquina, de modo que se tuvo que recurrir a la documentación de Caffe. Otra vez se tuvo que indagar en los comentarios dentro del código, porque esto no estaba explicado dentro de la documentación que se encontró, y Caffe es un proyecto más extenso y complejo por lo que llevó mucho tiempo llegar a la solución.

Resuelto el problema anteriormente descrito, se procede a la programación del servidor, este tiene que:

- Escribir en el fichero `tasks.txt` el nombre de la imagen a procesar: Esto fue sencillo, ya que trabajar con ficheros en Python es bastante fácil.
- Ejecutar el *script* de Matlab: El script de Matlab no predecía, sino que extrae características de la imagen para poder usarla con NeuralTalk. El problema que surgió aquí es que Matlab nos daba un error de licencia ya que el servidor se ejecutaba como superusuario, pero Matlab detectaba la licencia a nombre del usuario normal. Además se tuvo que investigar la secuencia

---

<sup>6</sup><http://cs.stanford.edu/people/karpathy/neuraltalk/>

de comandos correcta para que el *script* se ejecutara, una vez más la documentación no decía exactamente cómo se debía ejecutar este.

Finalmente, cuando se descubrió cómo ejecutar el *script*, se tuvo que buscar una solución al problema con la licencia. Se intentó cambiar la licencia o reinstalar Matlab, pero no dio resultado. Al final se optó a hacer un script bash y ejecutarlo desde el servidor usando el nombre de usuario que sí aceptaba Matlab en su licencia.

- Seguidamente se tenía que coger el fichero HTML que devolvía NeuralTalk y procesarlo para convertir todos esos datos en la cadena que queríamos, esto no llevó demasiado tiempo porque con la función *split* de cadenas de texto y la facilidad de Python para trabajar con ficheros, se convirtió en una tarea bastante sencilla.

Finalmente se consiguió integrar así el servidor con NeuralTalk, pero la carga de trabajo que esto llevó fue bastante grande.

### 6.1.6 Dificultades con Android

Con Android encontramos los problemas más destacables a la hora de realizar la conexión con el servidor y enviarle los datos. Estos son fácilmente numerables por lo que vamos a tratarlos por puntos:

#### ■ Usando la librería de Apache para HTTP

El primer problema de todos es que para usar esta librería era necesario instalarla de manera externa y en la documentación de Apache o Android no se encontraba la solución de manera clara. Se tuvo que investigar durante bastante tiempo hasta dar la solución en un foro de Internet<sup>7</sup>.

Después se procedió a programar la conexión y se hizo con la documentación que se encontró de manera más usual, pero esta resultó estar obsoleta y, por lo tanto, no funcionaba al compilar y ejecutar la aplicación. Se tuvo que buscar información de manera más exhaustiva para dar con la solución correcta y que no fuera obsoleta.

Una vez se ha cambiado el código, nos encontramos con el problema de que nuestra imagen daba incompatibilidades con el tipo de dato que se tenía que definir en la petición de tipo POST, de manera que se recurrió otra vez a una búsqueda bastante exhaustiva hasta dar con la solución de cómo se tenía que tratar la imagen para mandarla dentro de una petición POST a un servidor.

Finalmente descubrimos que, debido a estándares establecidos por Android, lo que habíamos programado no se podía ejecutar porque esto no debía hacerse en el hilo principal de ejecución, sino que debía ser lanzado en un hilo diferente. De manera que se tuvo que reestructurar todo el código para adaptarlo a esta especificación, una vez más se tuvo que realizar una investigación de cómo se realizaba esto y por qué.

---

<sup>7</sup><http://stackoverflow.com/questions/28538078/java-lang-nosuchfielderror-org-apache-http-message-basicheadervalueformatter-in>

### ■ Localizando la imagen

Después de haber programado la conexión de manera correcta, la petición hacia el servidor se realizaba de manera incorrecta devolviendo un código de error. Descubrimos el origen de este error ayudándonos de los mensajes de error que devolvían las Excepciones que se producían.

Descubrimos que el error se encontraba al pasar los datos de la imagen desde un hilo a otro, de manera que tuvimos que investigar la manera de conseguir solucionar esto. Resulta que este tipo de problemas es bastante complejo de solucionar porque al intercambiar los datos de un hilo a otro se pierde la dirección en la que se encuentra la imagen. Para solucionar esto se ha puesto la imagen en memoria compartida y se crea la imagen con la dirección de la misma, para extraer esta dirección se tiene que crear una función que extraiga la dirección de la imagen. Al trabajar con la dirección y no con la cabecera de la imagen, no se pierde la imagen durante el proceso.

#### 6.1.7 Dificultades con Librería de Traducción en el servidor

El problema que se encontró es que el API de traducción de Google resultó ser de pago, por lo tanto no era del todo adecuado para su uso en este proyecto. Pero se encontró una solución alternativa.

La solución alternativa consistía en simular una conexión a través del navegador a la página de traducción de Google y recibir como respuesta la cadena ya traducida. Esta solución es bastante interesante ya que usa conceptos de los servicios web y más concretamente de los servicios web con una especificación de tipo REST.

Se mandará una petición a la página de Google a través de un url, como sabemos la dificultad está en cómo decir a la página qué queremos traducir y de qué a qué idioma lo queremos traducir. Pues bien, esto se hace definiendo un url. En este url se tiene que hacer uso de la construcción de url, porque en el url es donde van definido los parámetros que se pasan a la página web. Una vez se construye el url, se procede a lanzar la petición con los parámetros establecidos.

Finalmente, se recibe una respuesta que es recibida en un dato de tipo json, este tipo de dato es fácilmente convertido a un dato de cadena de caracteres en Python y, finalmente, se extrae de ese conjunto de caracteres la cadena que contiene la traducción del texto.

## 7. CONCLUSIONES Y LÍNEAS DE TRABAJO FUTURAS

---

En este apartado se presentarán las conclusiones y las posibles líneas de futuro que saquemos del proyecto.

### 7.1 Conclusiones del Proyecto

Se han estudiado distintas herramientas de *Machine Learning* en las que se han visto las técnicas más novedosas y también las más utilizadas en el apartado del reconocimiento de imágenes.

En primer lugar hemos visto que para el reconocimiento de imágenes se considera necesario el uso de las redes neuronales convolucionales, puesto que estas son una herramienta muy potente porque están dedicadas al uso exclusivo del procesamiento de imágenes. Contando con la premisa de que la entrada será siempre una imagen, la red puede dedicar la configuración de sus capas y la forma de procesar los datos de entrada de manera más específica, obteniendo resultados bastante mejores que con cualquier otra red. Por lo tanto, se considera de uso obligatorio una red de tipo convolucional para el procesamiento de imágenes, obteniendo una extracción de características de mejor calidad que con el resto de redes.

En segundo lugar, situándonos en el plano de la generación de frases a partir de imágenes, se ha visto que el uso de las redes convolucionales sigue estando presente; pero que para el procesado de la frase se usa un red neuronal recurrente. Por tanto, hemos llegado a la conclusión de que una red neuronal recurrente es muy útil en este tipo de aspectos gracias a su propagación hacia atrás y a el añadido temporal en sus capas, pudiendo usar la predicción anterior como condicionante de la predicción actual.

En tercer lugar, hemos visto que la diferencia principal en los modelos, al menos los estudiados, para la generación de frases está situada en detalles muy concretos del modelo. Estos detalles son tan pequeños que nos vamos a trabajar con la función de activación de las neuronas y con el número de capas. Por tanto, la generación de frases parece ser un problema mejorable aún y que conlleva mucho estudio para llegar a la configuración con mejor funcionamiento y la técnica a usar puede ser una de las existentes o quizás se encuentren mejores.

En conclusión, cualquier tarea que conlleve el reconocimiento de imágenes y el procesado de las mismas se convierte en una labor muy complicada, y tiene muchas dificultades para su realización como el uso de modelos suficientemente grandes para poder procesar toda la información y abarcar la cantidad de clases que existen. Además, la tecnología actual no permite poder ejecutar este tipo de modelos sobre cualquier dispositivo, sino que debe tener un mínimo de potencia para que el rendimiento obtenido sea aceptable.

## 7.2 Líneas de trabajo Futura

En este apartado se procederá a mostrar los posibles trabajos de cara al futuro. Para determinar las líneas de trabajo futuras lo dividiremos en dos partes, desde el punto de vista teórico, porque se ha realizado un estudio teórico bastante grande en el proyecto y representa una parte importante del mismo, y desde el punto de vista del desarrollo de la aplicación de prueba, la cuál se podría desarrollar para darle un punto más práctico al proyecto en futuras líneas de trabajo.

### 7.2.1 Líneas de trabajo futuras en el aspecto teórico

Se ha estudiado una serie de modelos que trabajan en la tarea del reconocimiento de imágenes y, con ello, también se han estudiado las técnicas más usadas y algunas que podrían considerarse innovadoras. Para aprovechar todo este estudio y aplicarlo a líneas de trabajo futuro se podría formular un modelo propio, aunque no tiene por qué ser mejor, de los ya existentes y estudiados para realizar pruebas con lo que se ha aprendido a lo largo del proyecto y de su estudio teórico.

La construcción de un modelo es un gran trabajo para una línea de trabajo futura ya que permitirá al alumno probar las técnicas que ha estudiado e, incluso, tener la satisfacción de realizar pruebas con un modelo propio.

Como conclusión se podría decir que la construcción y postulación de los métodos propios del alumno significa un avance muy grande en sus conocimientos, ya que no se trata de modelos simples, sino que requieren un trabajo y entendimiento de la materia sobre la que se está trabajando.

#### ■ Líneas de trabajo futuras en el aspecto teórico

Este aspecto es, quizás, el más sencillo de trabajar como líneas de trabajo futuras. Pues en el desarrollo de la aplicación de prueba que se muestra junto con este proyecto se puede hacer muchas mejoras aún.

Las posibles mejoras que se pueden trabajar de cara a futuro con la aplicación de prueba son:

- Se puede mejorar el servidor para que gestione peticiones de muchos clientes al mismo tiempo.
- Se puede intentar mejorar el uso de MATLAB y los sus *scripts* para que la ejecución sea más rápida.
- Mejorar la configuración de las herramientas usadas para poder usarlo en varias GPU y mejorar su rendimiento.
- Desarrollar la aplicación para que en vez de trabajar con fotos lo haga con vídeos. Esto además requeriría de la modificación del servidor para gestionar esos datos entrantes de alguna manera.
- Aumentar el conjunto de imágenes de entrenamiento o crear unas propias para, posteriormente, entrenar con dicho conjunto la red y mejorar su funcionamiento.

Los objetivos enumerados anteriormente podrían ser trabajos futuros con la aplicación, pudiendo hacer con ellos la aplicación mucho más útil y funcional, además de que mejoraría su rendimiento.

**Universidad de Burgos**

Escuela Politécnica Superior

**Ingeniería Informática**

**Área de Lenguajes y Sistemas Informáticos**



**Anexo I - Plan del proyecto software**

**Estudio de herramientas de reconocimiento de  
imágenes con aplicación prototipo**

**Bryan Reinoso Cevallos**

**Tutores: Dr. José Francisco Díez Pastor,  
Dr. César I. García Osorio**



# I. PLAN DEL PROYECTO SOFTWARE

---

Este apartado está dedicado al plan de proyecto software donde se comenta todo el proceso de planificación del proyecto.

## I.1 Introducción

La planificación del proyecto se lleva a cabo de forma ágil, con la metodología SCRUM. Y la plataforma sobre la que se ha trabajado para realizar la planificación es VersionOne<sup>1</sup>. La planificación se ha dividido en *sprints* o iteraciones, las cuales tienen la mayoría una duración de una semana, aunque por problemas algunas han durado más.

### I.1.1 Problemas encontrados

Debido a que las primeras dos iteraciones se realizan de forma simulada, osea después del tiempo previsto, por eso no se pueden sacar sus gráficos *burn-down*. Además de que al día 30 de uso del VersionOne<sup>1</sup> se caducó la prueba y perdí todos los datos del proyecto, pero contactando con el soporte de VersionOne<sup>1</sup> estos me reabrieron la prueba por una semana para extraer los datos necesarios, pero debido a la tardanza en su respuesta hay un *sprint* que dura 3 semanas y no podemos extraer su gráfico *burn-down* tampoco.

## I.2 Planificación temporal del proyecto

En este apartado se presentan los distintos *sprints* y las tareas de cada uno.

### I.2.1 Sprint 1:23 de Diciembre al 12 de Febrero

Este *sprint* es de una duración bastante mayor al resto porque se empezó con mucha antelación y su objetivo era el de documentarse. Se estudió los distintos aspectos del proyecto y se tomó contacto con el mismo.

Este *sprint* es importante porque es donde se asientan las bases del proyecto y se empieza a tomar contacto con el proyecto. Sin este *sprint* se hubiera comenzado sin ninguna base y el comienzo del proyecto hubiera sido más tarde y, por tanto, su avance más lento.

Documentarse consistió en la lectura, en primer lugar, de historia de *machine learning* y de sus avances para situarse en el contexto de lo que se iba a trabajar en este proyecto. Además, se tuvo que leer las definiciones y las herramientas más usadas en el aprendizaje máquina. Se leyó, sobretodo,

---

<sup>1</sup><http://www.versionone.com/>

información en Internet a través de páginas que hablan del tema y que tienen bien estructurada su información siendo, además, fiable.

También se buscó en este *sprint* una serie de herramientas con las que se iban a poder trabajar, situándonos un poco en el panorama del proyecto; para hacernos una idea de cómo se iba a realizar el proyecto.

Por lo tanto, el objetivo clave del proyecto era llegar a la segunda iteración con una idea base de cómo realizar el proyecto y con qué tipo de herramientas se podía acabar trabajando, para evitar que la carga inicial sea más grande.

### I.2.2 Sprint 2:12 de Febrero al 26 de Febrero

Esta iteración se dedicó exclusivamente a la decisión de las herramientas que se van a usar en el proyecto y a la familiarización con las mismas.

Aquí se decidió:

- Gestor de versiones: Github<sup>2</sup>
- Gestor de Tareas: VersionOne<sup>1</sup>
- Entorno de desarrollo: Android Studio<sup>3</sup>
- Herramienta de ofimática: LATEXcon su editor TEXMaker

### I.2.3 Sprint 3:26 de Febrero al 6 de Marzo

En este *sprint* se instalaron y se crearon cuentas de las herramientas del anterior *sprint*. Se dividió en distintas tareas:

- Crear cuenta VersionOne<sup>1</sup>
- Crear cuenta en GitHub<sup>2</sup>
- Instalar Android Studio<sup>3</sup>
- Instalar librería DeepBeliefSDK<sup>4</sup>

Además se procedió a invitar a los tutores a VersionOne<sup>1</sup>, realizar el primer *commit* en GitHub<sup>2</sup>, probar ejemplos de Android y del DeepBeliefSDK<sup>4</sup> y, finalmente, se generó la documentación asociada a este *sprint*.

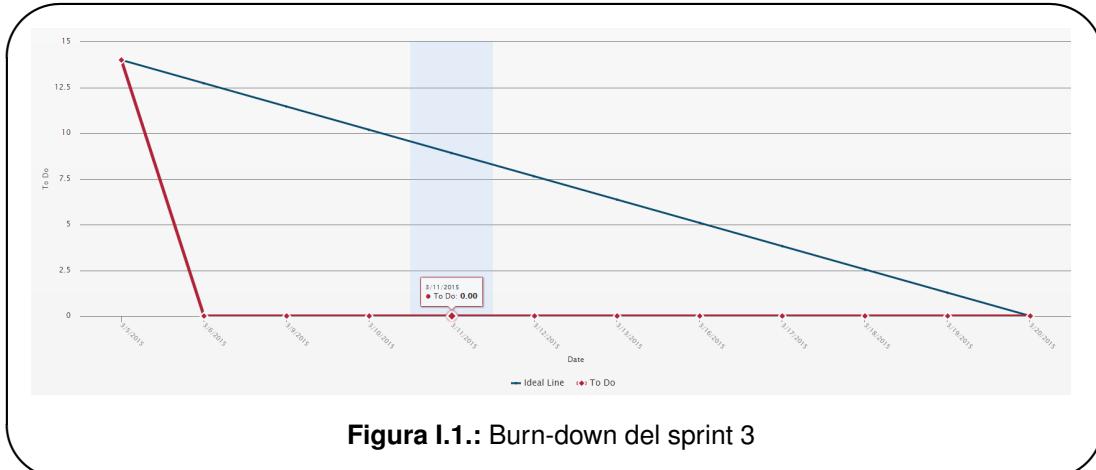
Podemos ver el gráfico *burn-down* de este sprint en la imagen I.1.

---

<sup>2</sup>[//github.com/](https://github.com/)

<sup>3</sup><http://developer.android.com/sdk/index.html>

<sup>4</sup><https://github.com/jetpacapp/DeepBeliefSDK>



#### I.2.4 Sprint 4:6 de Marzo al 13 de Marzo

En esta iteración se profundiza, sobretodo, en la utilización de la librería DeepBeliefSDK<sup>4</sup>, se probarán ejemplos y trabajarán sobre ellos. Además se generará la documentación asociada al *sprint*.

Se identifican las tareas:

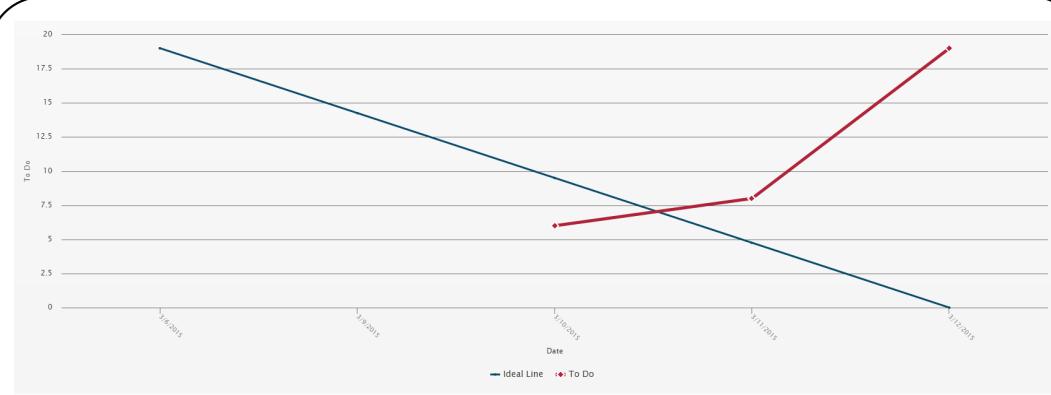
- Consulta de la documentación de Android: se pasa a avanzar en la programación de Android y se empiezan a realizar las primeras aplicaciones de prueba con ayuda de la documentación de Android.
- Instalación Linux: Se procede a instalar un Linux con su distribución Ubuntu en una máquina virtual sobre la que trabajar.
- Impresiones de Pantalla: Se hacen capturas de pantalla del proceso de instalación de las herramientas para la futura documentación en la sección: Manual del programador (IV)

Este fue el primer *sprint* en el que añadimos el *Retrospective Meeting* y el *Sprint Planning* al VersionOne<sup>1</sup>, pese a que sí que habíamos hecho estas reuniones.

En el *Retrospective Meeting* se determinó que:

- Se estudió el funcionamiento de la librería DeepBeliefSDK<sup>4</sup>
- Se hicieron las impresiones de pantalla para la documentación
- Se hicieron más pruebas con Android.
- Se envió invitaciones a los profesores a VersionOne<sup>1</sup> y Github<sup>2</sup>

El *Sprint Planning* simplemente sirvió para determinar las tareas a realizar en esta iteración, las cuales ya han sido comentadas antes.



**Figura I.2.: Burn-down del sprint 4**

Además esta iteración viene acompañada con un gráfico *burn-downs* el cuál puede verse en la imagen I.2

### I.2.5 Sprint 5:13 de Marzo al 27 de Marzo

Este *sprint* está dedicado en su mayoría al aprendizaje de GSOAP<sup>5</sup>, que es un *framework* para poder programar servidores en código C y C++. El estudio de esta herramienta llevó bastante tiempo debido a que su documentación, no solo estaba puramente en inglés, sino que además no era lo suficientemente precisa como para alguien que estuviera empezando a programar servicios WEB.

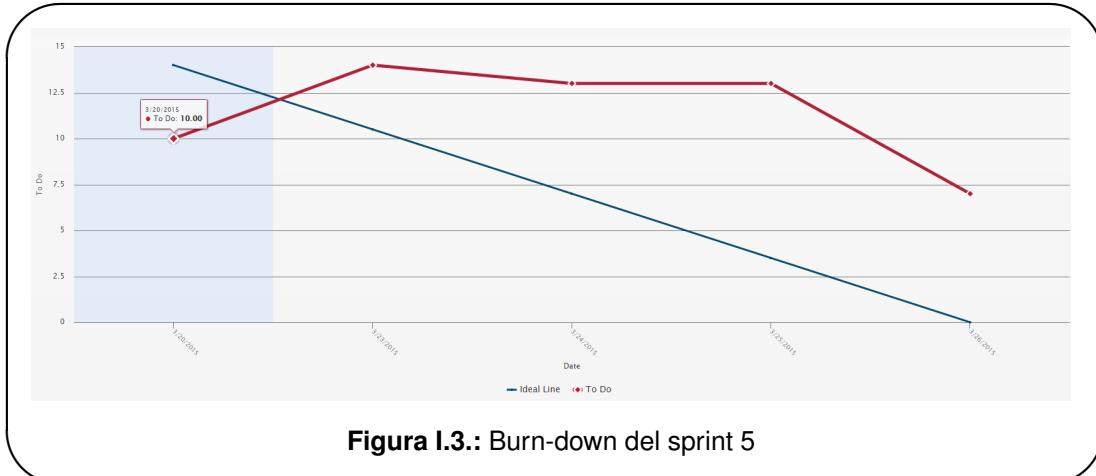
En el *Sprint Planning* se propuso como objetivos a tener en el siguiente *sprint*:

- Evaluar si usar Axiss2/c o GSOAP<sup>5</sup> para el servidor
- Continuar con la documentación
- Ejemplos de Android

En este *sprint* podemos identificar las siguientes tareas que finalmente se definieron en la herramienta VersionOne<sup>1</sup>:

- Pruebas en Linux: Se hacen pruebas de la librería GSOAP<sup>5</sup> en la máquina virtual de Linux, en su distribución Ubuntu.
- Pruebas en Android: Se hacen pruebas de distintos códigos y aplicaciones en Android que posteriormente nos puedan servir
- Primera versión del Cliente: Se genera la primera versión del cliente Android para el proyecto. Esta sólo es una interfaz sencilla que toma una foto con el móvil y la guarda.
- Estudio de WSDL y SOAP: Se estudia estas dos especificaciones de XML, junto con un pequeño repaso de XML. Estas dos especificaciones son totalmente necesarias para el desarrollo de la aplicación, aunque al final puede que no se lleguen a usar directamente, el

<sup>5</sup><http://www.cs.fsu.edu/~engelen/soap.html>



conocimiento de las mismas es requisito indispensable para conocer cómo trabaja el servidor internamente.

- Descargar materiales: Para el estudio y trabajo con la librería GSOAP<sup>5</sup> es necesario una serie de archivos, esta tarea es en la que nos descargamos todos los necesarios.
- Primera versión del fichero WSDL: Una versión inicial de un fichero WSDL con el que generar archivos *stubs* con la herramienta GSOAP<sup>5</sup>. Este fichero contiene las especificaciones de los servicios que dará el servidor al cliente y qué tipo de datos admite y devuelve.
- Generar documentación: Como en todas las iteraciones se procura generar la documentación asociada.

En el *Retrospective Meeting* se determina qué tareas del *sprint* han sido terminadas, cuáles no y el avance de la iteración. En este caso se identificó lo siguiente:

- Se decidió usar GSOAP<sup>5</sup> porque era más sencillo. Esta herramienta era para programar un servidor y poder ejecutar DeepBeliefSDK<sup>4</sup> en el lado del servidor y que el cliente sólo mande una foto.
- Se estudió GSOAP<sup>5</sup> y se consiguió implementar un cliente en Linux de un ejemplo básico de calculadora.
- Se continuó con ejemplos de Android creando el proyecto del cliente en su primera versión, aunque este no hacía nada más que mostrar una imagen por pantalla.

Esta iteración también tiene un gráfico de tipo *burn-down* asociado y se puede ver en la imagen I.3

### I.2.6 Sprint 6: 28 de Marzo al 22 de Abril

En esta iteración se procedió a crear prototipos, tanto de cliente como de servidor, del proyecto para poder empezar con el desarrollo software del mismo.

En la planificación dentro del *Sprint Planning* se determinó:

- Tratar de implementar el ejemplo básico de servidor, con el que podamos hacer las primeras pruebas.
- Tratar de hacer el cliente en Android, que se conectara al ejemplo básico de servido. Con esto empezaremos a establecer cómo se realizará la conexión entre cliente y servidor.
- Tratar de hacer el prototipo con DeepBeliefSDK<sup>4</sup> para empezar a intentar conectarlo con el servidor.
- Continuar con la documentación.

Esos fueron los objetivos fijados para este *sprint*, que es inusualmente largo debido a la serie de problemas que se encontraron en el mismo y que son explicados en el *Retrospective Meeting*.

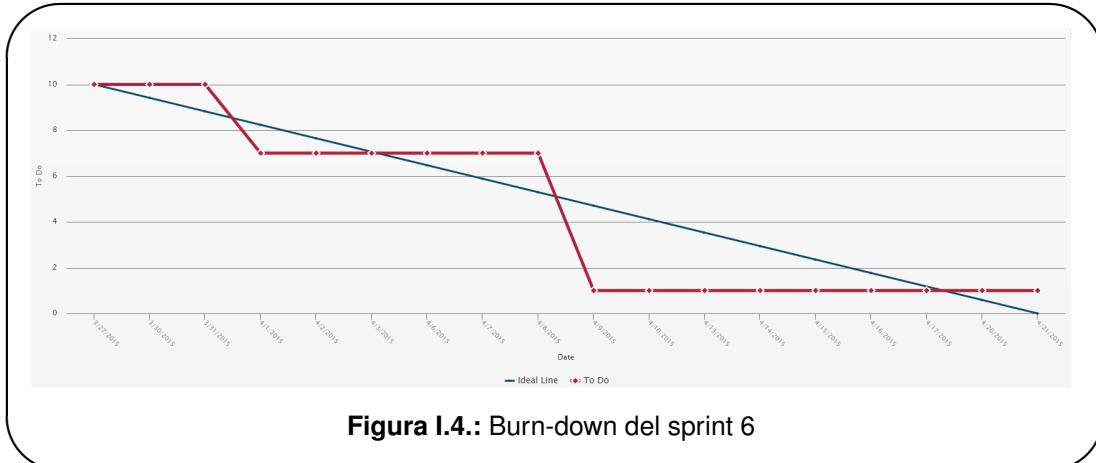
Mientras que en VersionOne lo que tenemos definido es:

- Instalar Apache y demás herramientas: Nos encontramos con que el servidor de GSOAP necesitaba de Apache para poder ejecutarlo en localhost y así establecer la conexión con el cliente Android, así que procedimos a la instalación de las herramientas necesarias.
- Construcción del servidor: Se montó el servidor con un ejemplo de calculadora básica y se comprobó con las propias herramientas de GSOAP, que este estaba bien definido.
- Construcción del cliente: Se montó un cliente GSOAP con el que también se probó el ejemplo básico de servidor, este y el servidor funcionaban de manera adecuada.
- Pruebas con distintos servicios en Linux: Se remontaron varias veces el cliente y el servidor, pero con distintas especificaciones SOAP, para probar varios tipos de datos en el envío.
- Primera versión del Servidor: Después de construir el ejemplo básico se pasó a conectar el DeepBeliefSDK con el servidor y este con Apache. Se empezó por probar la conexión con Apache, la cuál resultó un rotundo fracaso.
- Escritura de la documentación: Se genera, como siempre, la documentación asociada a la iteración.

Finalmente tenemos nuestro *Retrospective Meeting*, en el que determinamos lo siguiente:

- No se consiguió conectar GSOAP con Apache
- La documentación para el proceso de conectar Apache y GSOAP es escasa y la poca que hay no nos enseña un procedimiento correcto para hacer funcionar esto.
- Se decide abandonar la programación en C y la librería DeepBeliefSDK<sup>4</sup>
- Conclusión: No se consiguieron los objetivos y se decide tomar un nuevo rumbo con el proyecto.

Podemos observar el gráfico de la imagen I.2, en el que observamos el gráfico *burn-down* del *sprint*.



### 1.2.7 Sprint 7: 22 de Abril al 27 de Abril

Después del anterior *sprint* se procedió a cambiar totalmente el rumbo del proyecto, para ello se ha tomado la decisión de continuar el proyecto en Python, con una librería nueva y con un *framework* nuevo.

En el *Sprint Planning* se determinó los siguientes objetivos:

- Crear prototipo inicial de servidor con Python y Flask
- Estudio del funcionamiento de la librería NeuralTalk
- Prototipo del cliente en Android

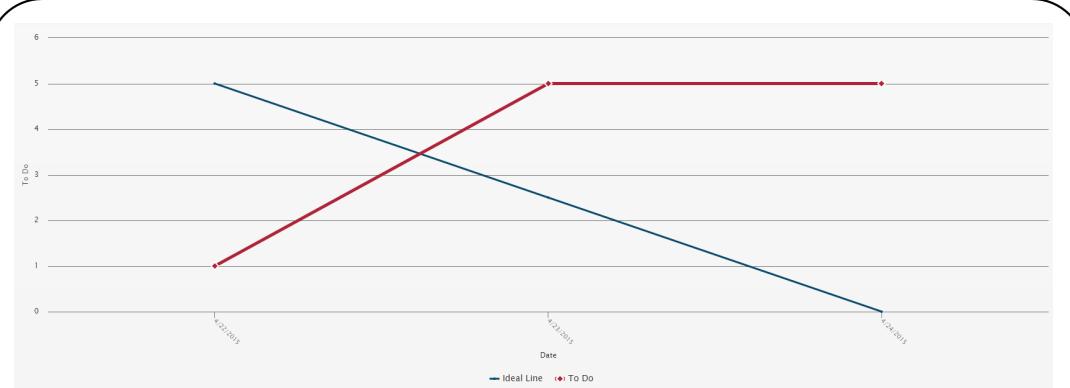
En VersionOne se definieron una serie de tareas para llevar a cabo estos objetivos, las tareas definidas dentro de la herramienta de gestión de tareas son:

- Instalación de NeuralTalk y sus dependencias: Se procede a la instalación del proyecto de GitHub Neuraltalk, además de empezar a hacer pruebas con él y leer la documentación.
- Instalación de Flask y Python: Para la programación del servidor es necesario la instalación de Flask y de Python.
- Primera versión del servidor: Se procede a la creación de una primera versión del servidor que reciba una imagen y la guarde en el sistema, posteriormente esta será procesada por la herramienta de predicción que se haya elegido.

Una vez se han definido las tareas estas se realizaron a lo largo de la semana, aunque como veremos en el gráfico *burn-down* producido que al final, debido a problemas a la hora de trabajar con NeuralTalk, se ve aumentado el trabajo por hacer ya que surgen tareas no planificadas, las cuales pasarán a ser resueltas en el siguiente *sprint*.

Finalmente, en el *Retrospective Meeting* se determinó lo siguiente:

- Se instalaron NeuralTalk y sus dependencias, lo cuál resultó bastante costoso ya que al final este proyecto era dependiente de otro que estudiamos al principio, Caffe.
- Se instaló correctamente Python, Flask y todas sus dependencias.



**Figura I.5.: Burn-down del sprint 7**

- Se programó la primera versión del servidor tal y como se había planificado.
- En el trabajo con la herramienta NeuralTalk se vio que esta era más compleja de lo esperado, pues no se podía simplemente predecir la imagen que uno quisiera sino que tendría que modificarse lo que en el proyecto original había para el caso de uso en el que nos encontramos.

Esto se debió a que la herramienta venía, en principio, preparada sólo para poder ser usada con las imágenes que ya venían en sus ejemplos y no se podían procesar por separado, sino que había que procesarlas juntas porque la herramienta no lo permitía de otra manera. Además de que la devolución se hacía en un documento HTML y no como nosotros lo necesitábamos.

Finalmente vamos a ver un gráfico de tipo *burn-down* asociado a esta iteración y generado por la herramienta VersionOne, el gráfico se puede observar en la imagen I.5.

### I.2.8 Sprint 8: 27 de Abril al 8 de Mayo

En este *sprint* se procedió a la creación de prototipos de cliente y servidor como principal objetivo. Por lo que en el *Sprint Planning* se determinaron los siguientes objetivos:

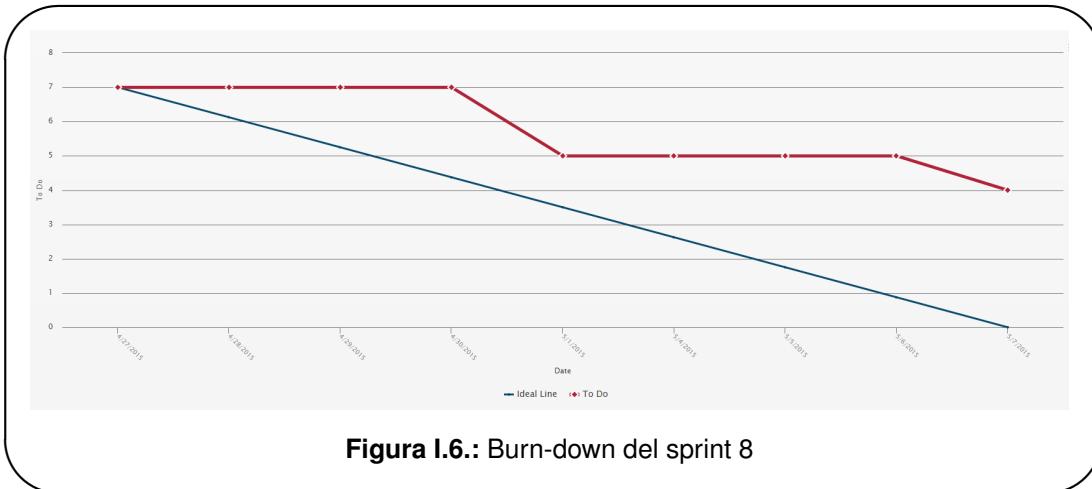
- Hacer el prototipo que permita subir la imagen al servidor, lanzar el reconocimiento y devolver la cadena de descripción de la imagen.
- Probar el API de traducción de Google para que la cadena devuelta, en vez de inglés, sea en castellano.
- Probar el Text2Speech para que la aplicación móvil, en vez de sólo mostrar la cadena devuelta, la lea.
- Buscar información sobre bibliotecas de prueba, tanto para el servidor como para el cliente, para empezar a elaborar el conjunto de pruebas de la aplicación.
- Actualizar la documentación con todas las nuevas cosas.
- Comprobar si se puede entrenar la red neuronal con imágenes propias.

Para la realización de estos objetivos se tuvo que determinar una serie de distintas tareas con las que ir trabajando. Esto se hace en VersionOne, en él se definieron las siguientes tareas:

- Instalando Caffe y sus dependencias: Nos encontramos ante la situación de que NeuralTalk usa el *wrapper* para Matlab de Caffe, por lo que debemos instalar Caffe y comprobar que este funciona.
- Entrenamiento de la red: Se comprobó que la red podía ser entrenada con distintos *datasets*, incluido uno propio. Como el entrenamiento completo de la red neuronal podía llegar a tardar varios días en un ordenador normal, se procedió a descargar una red ya entrenada.
- Pruebas con imágenes de ejemplo: Una vez descargada la red entrenada se procedió a probar esta con las imágenes que venían de ejemplo en la librería, comprobando que estas funcionaban pero que la ejecución de la misma sobre imágenes propias conllevaría algunas dificultades.
- Combinación de NeuralTalk con el servidor: Se combina NeuralTalk con el prototipo de servidor ya creado anteriormente. Aquí tenemos que resolver todos los problemas para lanzar la predicción sobre nuestras imágenes propias. Esto está detalladamente explicado en la sección de Aspectos Relevantes (6), más concretamente en la sección de Dificultades con NeuralTalk (6.1.5).
- Añadiendo librería de traducción: Se intentó añadir el API de Google para la traducción de cadenas de texto, pero esta resultó ser de pago y tuvimos que optar por otra solución, la cual es detallada en la sección de Aspectos Relevantes del Proyecto, más concretamente en el apartado de Dificultades con la librería de Traducción en el servidor (6.1.7).
- Conectando con el servidor : Se modifica el prototipo de cliente Android que se tenía para conectarlo con el nuevo servidor y mandarle una imagen, para después recibir la predicción.
- Estudiando las estructuras de conexión: Como surgió un problema al conectar la aplicación con el servidor, ya que nos devolvía como respuesta que hemos hecho una petición errónea, se procedió a estudiar la forma en qué se conectaba la aplicación con el servidor, aunque aquí no resultó estar el problema de la petición.
- Añadiendo librería de habla: Se usa la librería Text2Speech de Android para hacer que la aplicación de instrucciones guiadas por voz y que la predicción se lea en voz alta.

Como se puede observar hay una cantidad de tareas bastante elevadas y estas conllevan bastante trabajo, por lo que gran carga del desarrollo software se encuentra en este *sprint*. Como Resultado del *sprint* obtenemos lo siguiente en el *Retrospective Meeting*:

- Hacer el prototipo que permita subir la imagen al servidor, lanzar el reconocimiento y devolver la cadena de descripción de la imagen: **Hecho**. A pesar de la cantidad de problemas con esta tarea se logró hacer en esta iteración, pero la carga de trabajo que requirió fue bastante elevado.
- Probar el API de traducción de Google para que la cadena devuelta, en vez de inglés, sea en castellano: **Hecho**. Se tuvo problemas con el API de Google pero se encontró una solución para resolverlo y se añadió correctamente al servidor.
- Probar el Text2Speech para que la aplicación móvil, en vez de sólo mostrar la cadena devuelta, la lea: **Hecho**. Se añadió correctamente esta funcionalidad al prototipo de aplicación.



**Figura I.6.: Burn-down del sprint 8**

- Buscar información sobre bibliotecas de prueba, tanto para el servidor como para el cliente, para empezar a elaborar el conjunto de pruebas de la aplicación:**No hecho**. No se buscó información porque no hubo tiempo.
- Actualizar la documentación con todas las nuevas cosas: **No hecho**. No se tuvo tiempo de trabajar sobre la documentación en esta iteración.
- Si te quedara tiempo, intentar ver si se puede entrenar la red neuronal con imágenes propias **No hecho**. Se considera no hecho porque no se llegó a probar esta utilidad a pesar de que pareciera que sí es posible.

De este *sprint* también se tiene un gráfico *burn-down* asociado, el cuál puede verse en la imagen I.6.

## I.2.9 Sprint 9: 9 de Mayo al 15 de Mayo

En el anterior *sprint* se tuvo un gran avance en el desarrollo de la aplicación de muestra, pero no se avanzó en el desarrollo de la documentación del proyecto y, por tanto, se procedió a centrarse en este aspecto en esta iteración. En la reunión de *Sprint Planning* se determinaron los siguientes objetivos:

- Hacer documentación del proyecto: Se ha escrito la documentación pendiente de lo que se ha hecho y comenzado nuevos apartados que aún no se habían escrito.
- Instalar Caffe con octave: Se intentó instalar Caffe con octave porque daba problemas el Caffe, estos están explicados detalladamente en la sección: [6.1.5](#).
- Probar servidor con GPU: Se pretendía instalar el software desarrollado en una máquina anfitriona para poder ejecutar la predicción en GPU y reducir así su tiempo de ejecución.
- Buscar herramientas de pruebas: Se buscarán herramientas con la que hacer teses sobre la aplicación software para asegurar sus buen funcionamiento.

Las tareas VersionOne<sup>1</sup> y las tareas del *Sprint Planning* coinciden en este caso, así que no se volverán a escribir.

Y, en el *Retrospective Meeting*, se definió que:

- Hacer documentación del Proyecto: **Hecho**. Aunque no se terminó la documentación entera del proyecto, sí que se hizo las partes que se tenían planificadas por hacer.
- Instalar Caffe con Octave: **Fallo**. Se comprobó que no se podía instalar Caffe con Octave porque la instalación misma de MatCaffe, que es el *wrapper* para Matlab de Caffe, necesita del compilador mex de Matlab.
- Buscar Herramientas de prueba: **No hecho**. Debido a que se dedicó mucho tiempo a la documentación y a que la prueba de VersionOne<sup>1</sup> llegó a su fin, esta tarea quedó aplazada para futuros *sprints*.
- Probar servidor con GPU: **No hecho**. No se terminó la instalación del proyecto en una máquina anfitriona porque hubo problemas hardware, se perdieron todos los datos del disco duro y no se pudo trabajar sobre esto en este *sprint*.

Esta iteración no lleva asociado ningún gráfico ya que, debido a que se terminó la prueba en VersionOne, no se pudo actualizar día a día el progreso del *sprint* y, por tanto, no se genera ningún gráfico útil.

### I.2.10 Sprint 10: 15 de Mayo al 5 de Junio

En este *sprint* se tuvieron varios problemas, entre ellos que se seguía sin saber por qué no funcionaba VersionOne y se descubrió a mitad de la iteración. Por lo que este *sprint* no dispone de ningún gráfico representativo, ni tampoco de una planificación muy correcta en VersionOne<sup>1</sup>.

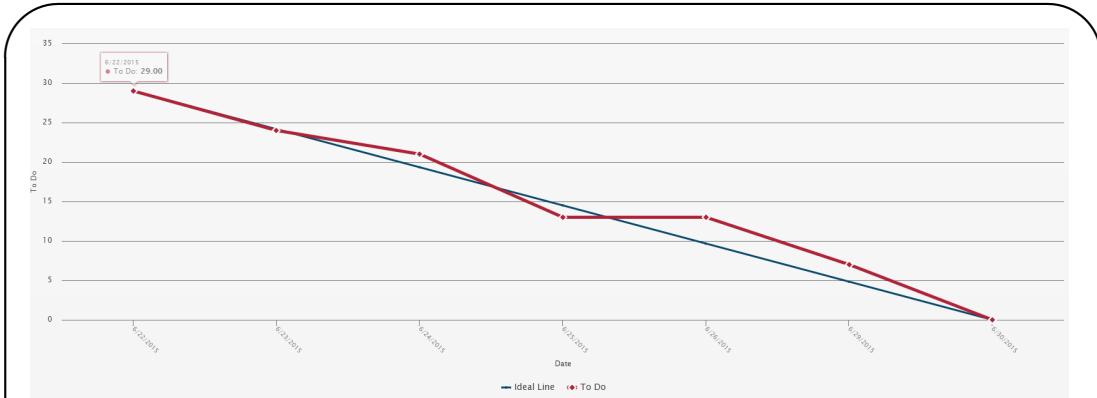
En el *Sprint Planning* se determinó que los objetivos de la iteración son:

- Acabar la documentación
- Seguir intentando configurar el proyecto en una máquina anfitriona y hacer pruebas con GPU, si es posible.
- Tratar de recuperar los datos perdidos del VersionOne, poniéndose en contacto con el equipo de soporte de VersionOne.

En VersionOne no tenemos una buena planificación de tareas porque los datos y la instancia de este programa se recuperó a mediados de la iteración, por lo que no fue posible recoger los datos ni el avance de la iteración de manera correcta.

En el *Retrospective Meeting* se determinó lo siguiente:

- Se avanzó bastante en la documentación, sobretodo en la parte de Plan de Proyecto Software (I).
- Se instaló lo máximo posible de la herramienta en anfitrión, pero se detectó la necesidad de modificar el cliente Android y el servidor Flask para facilitar su instalación y portabilidad.
- Se descubrió que VersionOne dejó de funcionar por haber terminado el período de prueba de este, así que se pidió que lo extendieran un poco; consiguiendo dos semanas más. Por lo que se tiene que sacar todos los datos que se pueda de VersionOne<sup>1</sup> y, si es posible, migrar el proyecto a una nueva versión de prueba de este.



**Figura I.7.: Burn-down del sprint 12**

### I.2.11 Sprint 11: 6 de Junio al 20 de Junio

Este *sprint* se encuentra situado en el proceso de recuperar los datos de VersionOne, por eso no se ha podido generar un gráfico de tipo *burn-down* a partir de esta iteración.

En la reunión del *Sprint Planning* se definieron los siguientes objetivos:

- Intentar integrar el trabajo de Bengio: se iba a intentar añadir una nueva librería sobre el servidor para probar su funcionamiento y compararla con la que ya se tenía.
- Documentar la instalación de ambas bibliotecas (Karpaty y Bengio): Generar la documentación del Manual del Programador del proceso de instalación de estas herramientas.
- En técnicas y herramientas se habla de Caffe, NeuralTalk, Deep Belief SDK y del código de Bengio: se añadiría las explicaciones de estas herramientas al apartado de Técnicas y Herramientas de la documentación.
- *Script* de instalación de NeuralTalk: Se generaría un *script* de instalación de la herramienta NeuralTalk.

Debido a que esta iteración coincidió con los exámenes y entrega de trabajos, no se avanza mucho. Por lo que en el *Retrospective Meeting* se definió lo siguiente:

- En técnicas y herramientas se habla de Caffe, NeuralTalk, Deep Belief SDK y del código de Bengio. DONE
- Intentar integrar el trabajo de Bengio. Se intentó pero no se logró.

El resto de tareas pasaron al siguiente *sprint*.

### I.2.12 Sprint 12: 21 de Junio a 30 de Junio

Este *sprint* se determinó que se debía hacer un trabajo algo más intensivo y se propusieron objetivos bastante exigentes.

En la reunión del *Sprint Planning* se definieron los siguientes objetivos:

- Terminar lo que no se terminó en la anterior iteración.
- Se propone finalizar la documentación para proceder a ser revisada por los tutores.
- Se propone terminar las aplicaciones de manera que queden entregables para el jurado.

Este conjunto de objetivos, los cuáles son bastante ambiciosos, se tuvieron que especificar en VersionOne a través de un conjunto bastante grande de tareas que resumiremos para evitar demasiadas líneas, que son las siguientes:

- Escribir documentación: Esta tarea está dirigida a escribir la documentación del proyecto, excluyendo los anexos.
- Lectura de artículos: En esta tarea se leen los artículos de investigación que formaran parte de la teoría del proyecto, concretamente serán situados en el apartado de Estado del Arte.
- Escribir anexos: Tarea en la que se recoge el trabajo que conlleva escribir los anexos del proyecto.
- Estructurar y acabar cliente: Se acabará de desarrollar el cliente y se dejará listo para la entrega final.
- Estructurar Servidor: Se terminará el servidor y se dejará listo para la entrega final.
- Construir *script* de instalación: Se hará un programa para la instalación del proyecto y, además, se harán los cambios pertinentes para facilitar la instalación del servidor. Esto estará explicado en el anexo del Manual del Programador.
- Hacer resúmenes: Se escribirán los resúmenes del proyecto, su versión en español y su versión en inglés.

Esta iteración tiene asociada un gráfico de tipo *burn-down*, que se puede ver en la imagen [I.7](#).



**Universidad de Burgos**

Escuela Politécnica Superior

**Ingeniería Informática**

**Área de Lenguajes y Sistemas Informáticos**



**Anexo II - Especificación de requisitos**

**Estudio de herramientas de reconocimiento de imágenes con aplicación prototipo**

**Bryan Reinoso Cevallos**

**Tutores: Dr. José Francisco Díez Pastor,  
Dr. César I. García Osorio**



## II. ESPECIFICACIÓN DE REQUISITOS

---

### II.1 Introducción

En este anexo se presentarán los distintos requisitos funcionales de la aplicación de ejemplo que se ha desarrollado. Se expondrán tanto los requisitos como los diagramas de caso de uso, plantillas de caso de uso.

### II.2 Requisitos Funcionales

Los requisitos que se establecen son las funcionalidades mínimas que se han propuesto como objetivos en la aplicación.

#### II.2.1 Requisitos Funcionales en el Servidor

Estos son los requisitos que se han establecido para que el servidor se considere que cumple con la funcionalidad exigida.

- **RF1 Recibir Imagen:** El servidor deberá ser capaz de, a través de una petición de tipo POST, recibir una imagen correctamente y guardarla en el sistema. Cuando este reciba una petición de tipo POST procederá a recibir los datos, seguidamente se guardarán los datos en el sistema de almacenamiento de la máquina en la que este alojado el servidor. Además deberá comprobar que lo que está recibiendo es una imagen para evitar que se nos envíe archivos maliciosos con extensiones no permitidas.
- **RF2 Procesar Imagen:** El servidor ejecutará la herramienta de predicción sobre los datos recibidos previamente. Deberá tener la configuración adecuada para que la herramienta funcione de manera adecuada y que ejecute una predicción sobre la imagen. La dificultad se centra en la configuración de una herramienta tan compleja como puede ser las de tratamiento de imágenes en *machine learning*.
- **RF3 Devolver Predicción:** El servidor deberá ser capaz de devolver una frase en español que sea el resultante de la predicción sobre la imagen. Para ello tiene otro requisito funcional subyacente a él.
  - **RF3.1 Traducir Predicción:** El servidor deberá ser capaz de traducir la predicción del inglés al español, ya que las herramientas que se pueden utilizar para esta aplicación trabajan todas en inglés.

#### II.2.2 Requisitos Funcionales en el Cliente

Se presentan ahora los requisitos que se han establecido en el cliente.

- **RF4 Solicitar Predicción:** El cliente Android deberá permitir al usuario tomar una foto. Teniendo en cuenta que la persona a la que va dirigida la aplicación es una persona con dificultades de visión o invidente, esta deberá permitir al usuario tomar la foto sin que este deba tener conocimiento alguno de la interfaz o de si hay un botón que este deba pulsar. Por tanto, la aplicación mostrará una pantalla sin interfaz que está a la espera de que el usuario simplemente toque la pantalla, entonces la aplicación tomará automáticamente la foto.
  - **RF4.1 Tomar Foto:** La aplicación tomará un foto de manera automática sin necesidad de que el usuario tenga que interactuar con la interfaz de manera específica, sino que podrá hacerlo sin saber qué se encuentra en la pantalla, tan sólo deberá tocar cualquier punto de la pantalla.
  - **RF4.2 Mandar Foto:** La aplicación cliente deberá ejecutar una petición de tipo POST hacia un servidor y mandar correctamente la imagen para que esta sea procesada por el servidor. Mientras el servidor procesa y recibe la imagen, la aplicación mostrará una ventana que informe de que la imagen se está procesando, esta ventana no es con el objetivo de que el usuario la vea, sino de que la aplicación esté a la espera de que la imagen sea procesada y que el usuario no tenga la capacidad de realizar nada en ese período. Esto asegurará que la aplicación no falle mientras se procesa la imagen.
  - **RF4.3 Leer Predicción:** Una vez se ha procesado la imagen la aplicación deberá ser capaz de recibir la predicción del servidor y, debido a que la persona que la use no tendrá la capacidad de leerla, la aplicación leerá la aplicación en alto a través de una librería *text to speech*.

### II.3 Diagrama de casos de uso

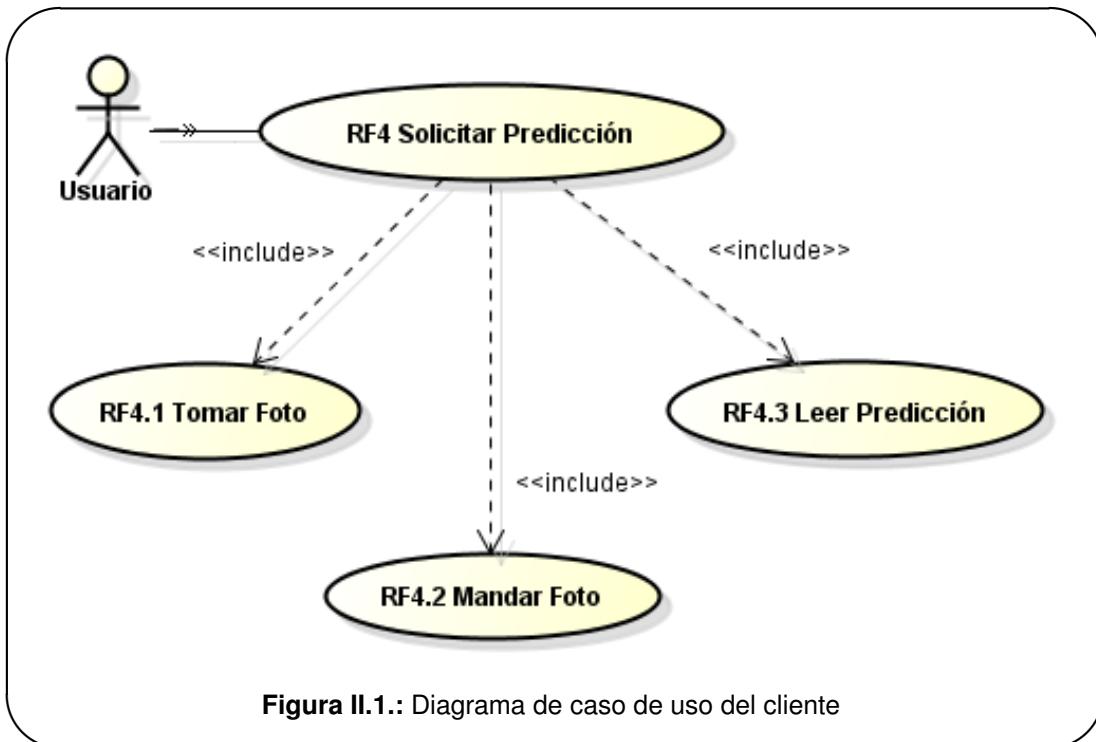
En este apartado se procede a presentar el diagrama o diagramas de caso de uso de la aplicación, además de las correspondientes tablas asociadas a cada uno de los requisitos funcionales del sistema.

El diagrama de casos de uso del cliente se puede ver en la imagen II.1.

El diagrama de casos de uso del servidor se puede ver en la imagen II.2.

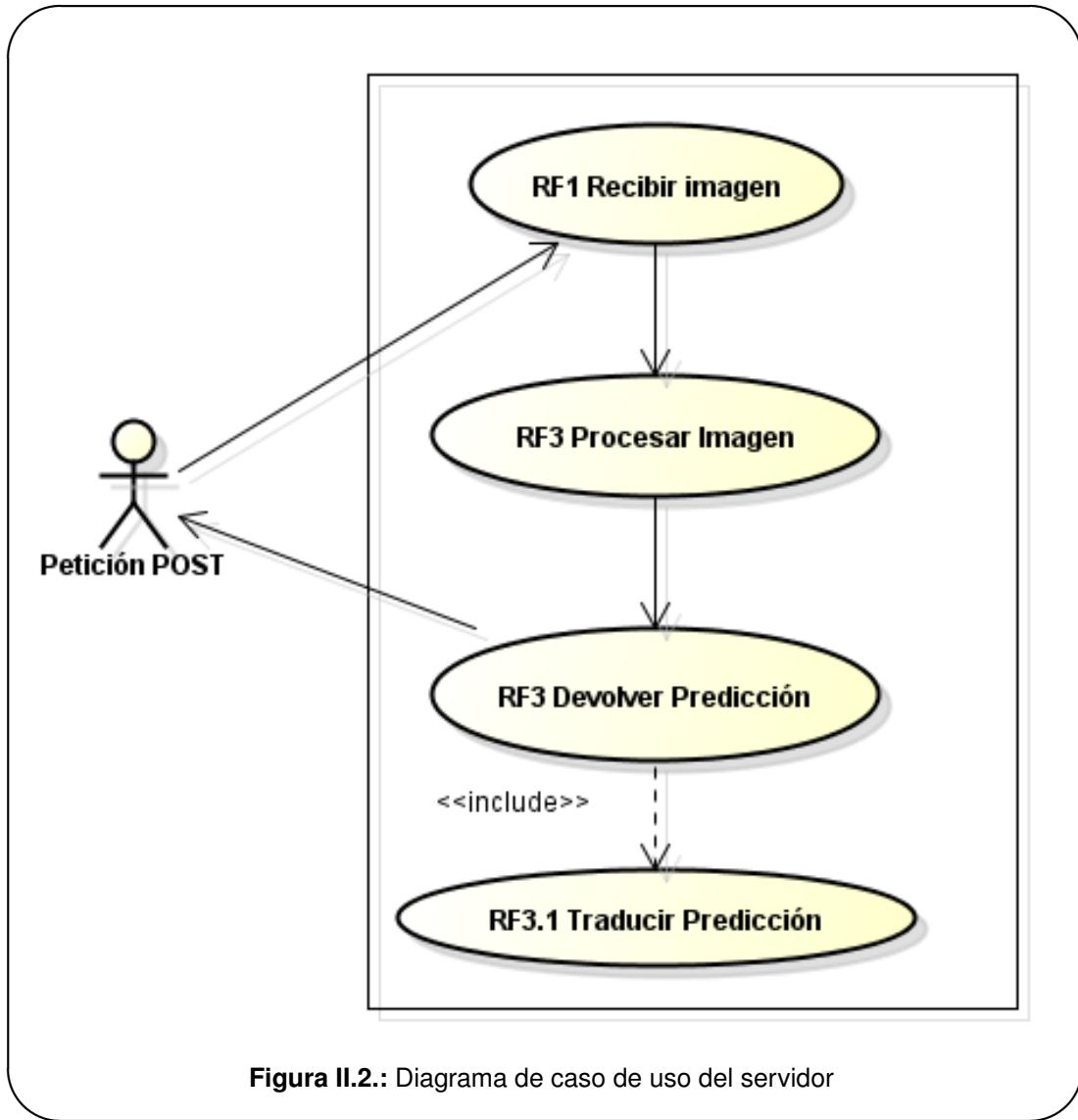
Mientras que las tablas de cada uno de los requisitos funcionales se pueden ver en:

- **RF1 Recibir Imagen:** Se puede ver en la tabla II.1.
- **RF2 Procesar Imagen:** Se puede ver en la tabla II.2.
- **RF3 Devolver Predicción:** Se puede ver en la tabla II.3.
  - **RF3.1 Traducir Predicción:** Se puede ver en la tabla II.4.
- **RF4 Solicitar Predicción:** Se puede ver en la tabla II.5.
  - **RF4.1 Tomar Foto:** Se puede ver en la tabla II.6.
  - **RF4.2 Mandar Foto:** Se puede ver en la tabla II.7.
  - **RF4.3 Leer Predicción:** Se puede ver en la tabla II.8.



RF1 Recibir imagen									
Descripción	Se recibirá una imagen enviada desde un cliente y se guardará.								
Precondiciones	Deberá haberse recibido una petición de tipo POST								
Secuencia normal	<table> <thead> <tr> <th>Paso</th><th>Acción</th></tr> </thead> <tbody> <tr> <td>1</td><td>Se detecta una petición POST</td></tr> <tr> <td>2</td><td>Se comprueba que el fichero recibido es una imagen</td></tr> <tr> <td>3</td><td>Si es imagen, se guarda, sino, se descarta.</td></tr> </tbody> </table>	Paso	Acción	1	Se detecta una petición POST	2	Se comprueba que el fichero recibido es una imagen	3	Si es imagen, se guarda, sino, se descarta.
Paso	Acción								
1	Se detecta una petición POST								
2	Se comprueba que el fichero recibido es una imagen								
3	Si es imagen, se guarda, sino, se descarta.								
Postcondiciones	Si se ha recibido una imagen, esta deberá estar guardada								
Excepciones	<table> <thead> <tr> <th>Paso</th><th>Acción</th></tr> </thead> </table>	Paso	Acción						
Paso	Acción								
Rendimiento									
Frecuencia	Alta, Media, Baja								
Importancia	Alta, Media, Baja								
Urgencia	Alta, Media, Baja								
Comentarios	Sólo se reciben imágenes tipo: jpg, jpeg y png								

**Tabla II.1.: Caso de uso: RF1 Recibir imagen**



**Figura II.2.:** Diagrama de caso de uso del servidor

RF2 Procesar imagen		
Descripción	La imagen será procesada a través de la herramienta elegida.	
Precondiciones	Se deberá haber guardado una imagen recibida por el usuario	
Paso	Acción	
Secuencia normal	1	Se escribe nombre de la imagen en fichero de procesado “tasks.txt”
	2	Se ejecuta <i>script</i> de extracción de características de la imagen
	3	Se lanza la extracción de frase.
Postcondiciones	Se tiene un documento “result.html” con la predicción	
Excepciones	Paso	Acción
Rendimiento		
Frecuencia	Alta, Media, Baja	
Importancia	Alta, Media, Baja	
Urgencia	Alta, Media, Baja	
Comentarios	El fichero tiene dentro la frase que se quiere obtener	

**Tabla II.2.:** Caso de uso: RF2 Procesar imagen

RF3 Devolver predicción		
Descripción	Se recogerá la predicción sobre la imagen y como respuesta a la petición esta será devuelta	
Precondiciones	1 Se deberá haber guardado una imagen recibida por el usuario 2 Deberá existir el documento “result.html” 3 Deberá traducirse la cadena al español	
Secuencia normal	Paso Acción 1 Se abre el fichero “result.html” si existe. 2 Se extrae la cadena en inglés. 3 Se Traduce la cadena al español 4 Se devuelve la cadena como resultado de la petición POST	
Postcondiciones	Se ha devuelto una cadena de texto, correspondiente a la predicción.	
Excepciones	Paso	Acción
	1	Si el fichero “result.html” no existe se genera un error.
Rendimiento		
Frecuencia	Alta, Media, Baja	
Importancia	Alta, Media, Baja	
Urgencia	Alta, Media, Baja	
Comentarios		

**Tabla II.3.:** Caso de uso: RF3 Devolver predicción

## II. ESPECIFICACIÓN DE REQUISITOS

### RF3.1 Traducir predicción

Descripción	A partir de una cadena de caracteres en inglés, se traducirá al español.	
Precondiciones	1	La cadena no deberá ser nula
	2	La cadena deberá estar en inglés
Secuencia normal	Paso	Acción
	1	Se crea url con cabecera para la traducción.
	2	Se manda petición a la página de Google traductor.
	3	Se reciben los datos en un json
	4	Se extrae la cadena del json
	5	Se devuelve la cadena traducida
Postcondiciones	Se ha devuelto una cadena de texto, correspondiente a la predicción.	
Excepciones	Paso	Acción
Rendimiento		
Frecuencia	Alta, Media, Baja	
Importancia	Alta, Media, Baja	
Urgencia	Alta, Media, Baja	
Comentarios	Se simula un conexión vía navegador para la traducción	

**Tabla II.4.: Caso de uso: RF3.1 Traducir predicción**

### RF4 Solicitar predicción

Descripción	El usuario solicitará una predicción cuando toque la pantalla.	
Precondiciones	Paso	Acción
	1	Se muestra pantalla en blanco a la espera de que el usuario la toque para tomar la foto.
Secuencia normal	2	Cuando el usuario toca la pantalla se toma automáticamente una foto.
	3	Se guarda la imagen en el dispositivo
	4	Se crea la petición POST y se manda
	5	Se recibe la predicción como respuesta a la petición POST
	6	Se muestra la predicción en pantalla junto con la imagen
	7	Se lee la predicción con la librería <i>Text to Speech</i>
	Postcondiciones	Se ha leído en voz alta el resultado de la ejecución.
Excepciones	Paso	Acción
	1	Si no se puede conectar con el servidor, se lee un mensaje de error en voz alta.
Rendimiento		
Frecuencia	Alta, Media, Baja	
Importancia	Alta, Media, Baja	
Urgencia	Alta, Media, Baja	
Comentarios	Este requisito es con el que interactúa el usuario.	

**Tabla II.5.: Caso de uso: RF4 Solicitar predicción**

## RF4.1 Tomar Foto

Descripción	Cuando el usuario ha solicitado la predicción, el sistema toma una foto de manera automática.	
Precondiciones	El usuario debe haber tocado la pantalla	
	Paso	Acción
	1	Se inicia la previsualización de la cámara.
Secuencia normal	2	Se toma la foto.
	3	Se trata los datos recibidos para guardarlos en el dispositivo
	4	Se crea la petición POST y se manda
	5	Se cierra la previsualización
Postcondiciones	Se ha guardado la imagen en el sistema.	
Excepciones	Paso	Acción
	1	Si la cámara no está disponible se genera un error.
Rendimiento		
Frecuencia	Alta, Media, Baja	
Importancia	Alta, Media, Baja	
Urgencia	Alta, Media, Baja	
Comentarios	Se usará la API Camera de Android para poder hacer la foto de manera automática.	

**Tabla II.6.: Caso de uso: RF4.1 Tomar Foto**

## RF4.2 Mandar Foto

Descripción	Se manda la foto al servidor a través de una petición de tipo POST.	
Precondiciones	Se ha tomado la foto correctamente	
	Paso	Acción
Secuencia normal	1	Se construye la petición POST y todos los objetos necesarios para poder ejecutar dicha petición.
	2	Antes de ejecutarla se crea una ventana para privar al usuario de hacer acciones innecesarias sobre la aplicación, pero esta no se muestra aún.
	3	Se pone en ejecución la petición POST
	4	Se muestra la ventana para que el usuario no pueda hacer acciones innecesarias.
	5	Una vez se ha terminado la petición, se recibe el resultado y se almacena.
	6	Se quita la ventana de privación de acciones.
Postcondiciones	Se ha obtenido el resultado de la predicción.	
Excepciones	Paso	Acción
	1	Si no se conecta con el servidor se lee un mensaje de error en vez de la predicción.
Rendimiento		
Frecuencia	Alta, Media, Baja	
Importancia	Alta, Media, Baja	
Urgencia	Alta, Media, Baja	
Comentarios	Se usará la API Http de Apache para poder hacer la petición POST de manera adecuada.	

**Tabla II.7.: Caso de uso: RF4.2 Mandar Foto**

## II. ESPECIFICACIÓN DE REQUISITOS

### RF4.3 Leer predicción

Descripción	Se lee en voz alta la predicción que se ha obtenido.	
Precondiciones	Se tiene una cadena para leer	
	Paso	Acción
Secuencia normal	1	Se llama al objeto de lectura con la cadena a leer.
	2	Se lee en voz alta.
Postcondiciones	Se ha leído el mensaje disponible para lectura.	
Excepciones	Paso	Acción
Rendimiento		
Frecuencia	Alta, Media, Baja	
Importancia	Alta, Media, Baja	
Urgencia	Alta, Media, Baja	
Comentarios	Se usará la API <i>Text to Speech</i> de Android para poder leer el mensaje.	

**Tabla II.8.: Caso de uso: RF4.3 Leer predicción**

**Universidad de Burgos**

**Escuela Politécnica Superior**

**Ingeniería Informática**

**Área de Lenguajes y Sistemas Informáticos**



**Anexo III - Especificación de diseño**

**Estudio de herramientas de reconocimiento de imágenes con aplicación prototipo**

**Bryan Reinoso Cevallos**

**Tutores: Dr. José Francisco Díez Pastor,  
Dr. César I. García Osorio**



## III. ESPECIFICACIÓN DE DISEÑO

---

### III.1 Introducción

En este apartado se procederá a explicar las especificaciones de diseño que se han ido utilizando para el desarrollo de la aplicación.

### III.2 Diseño en el Servidor

En primer lugar se introducirá las especificaciones que se han seguido para desarrollar el servidor.

#### III.2.1 API RESTful

Se ha construido el servidor de manera que siga, en la medida de lo posible, las especificaciones de la API RESTful. Esto se puede observar en el hecho de que se ha usado Flask para programar el servidor, que es un *framework* que permite la programación de peticiones de la API RESTful de manera muy sencilla. Además, tiene soporte para el uso de urls a la hora de trabajar con peticiones y funciones.

En el servidor se ha trabajado sólo con una url, la url principal del servidor. En la imagen (III.1) se puede observar cómo se ha definido la url base para el servidor. Además podemos ver que esta soporta dos tipos de métodos, o bien la petición de tipo GET, o bien la petición de tipo POST.

```

1 @app.route("/", methods=['GET', 'POST'])
2 def index():

```

**Listado III.1:** Url única y principal del servidor

Si nos fijamos, tenemos una función justo debajo de la notación “app.route”, esto implica que al acceder a esa url, se ejecuta esta función y es ahí donde se determina qué hace el servidor en cada caso.

En el servidor, la función trata de manera predeterminada la petición como una petición de tipo GET, así que habrá que tratar el método POST de alguna manera (III.2).

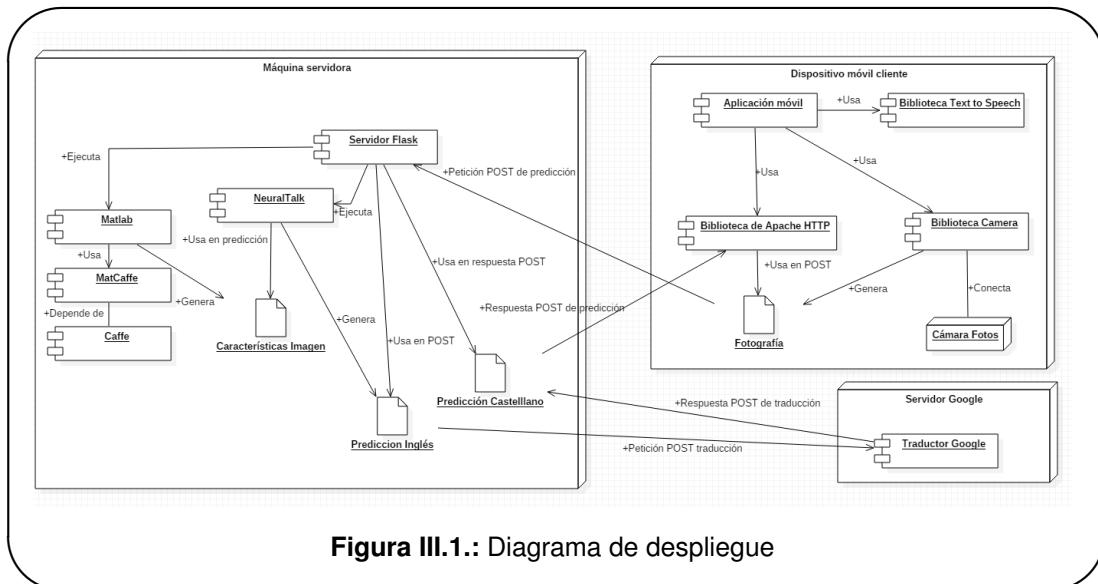
```

1 if request.method == 'POST':
2     #Aquí lo que se haría si es método POST
3     return "la respuesta de la operación"
4 #Aquí lo que se haría si es método GET
5 return "respuesta de un método GET"

```

**Listado III.2:** Tratando el método POST

### III. ESPECIFICACIÓN DE DISEÑO



EL objeto “request” es un objeto que usa Flask internamente y representa la petición que se está recibiendo. Para acceder a qué tipo de petición es, accedemos a su variable “method” en la que obtendremos la respuesta y con este valor se trabaja la petición POST.

Para asegurar la seguridad del sistema se ha usado una lista ([III.3](#)) de tipo de archivos admisibles por el servidor, lo que impide que se nos mande archivos maliciosos con extensiones extrañas, pues si el archivo tiene una extensión incorrecta este descarta la petición.

```
1 ALLOWED_EXTENSIONS = set(['jpg', 'JPG', 'jpeg', 'JPEG'])
```

**Listado III.3:** Tipos admitidos por el servidor

Posteriormente se define una función que se encarga de comprobar si el tipo es uno de los permitidos y si se cumple la condición, entonces se continúa el proceso, de lo contrario se descarta el archivo recibido.

```
1 def allowed_file (filename):
2     return '.' in filename and \
3         filename . rsplit ('.', 1)[1] in ALLOWED_EXTENSIONS
```

**Listado III.4:** Comprobando tipos

### III.3 Diagrama de despliegue

En esta sección se procede a presentar el diagrama de despliegue de la aplicación. Este diagrama se puede ver en la Figura [III.1](#).

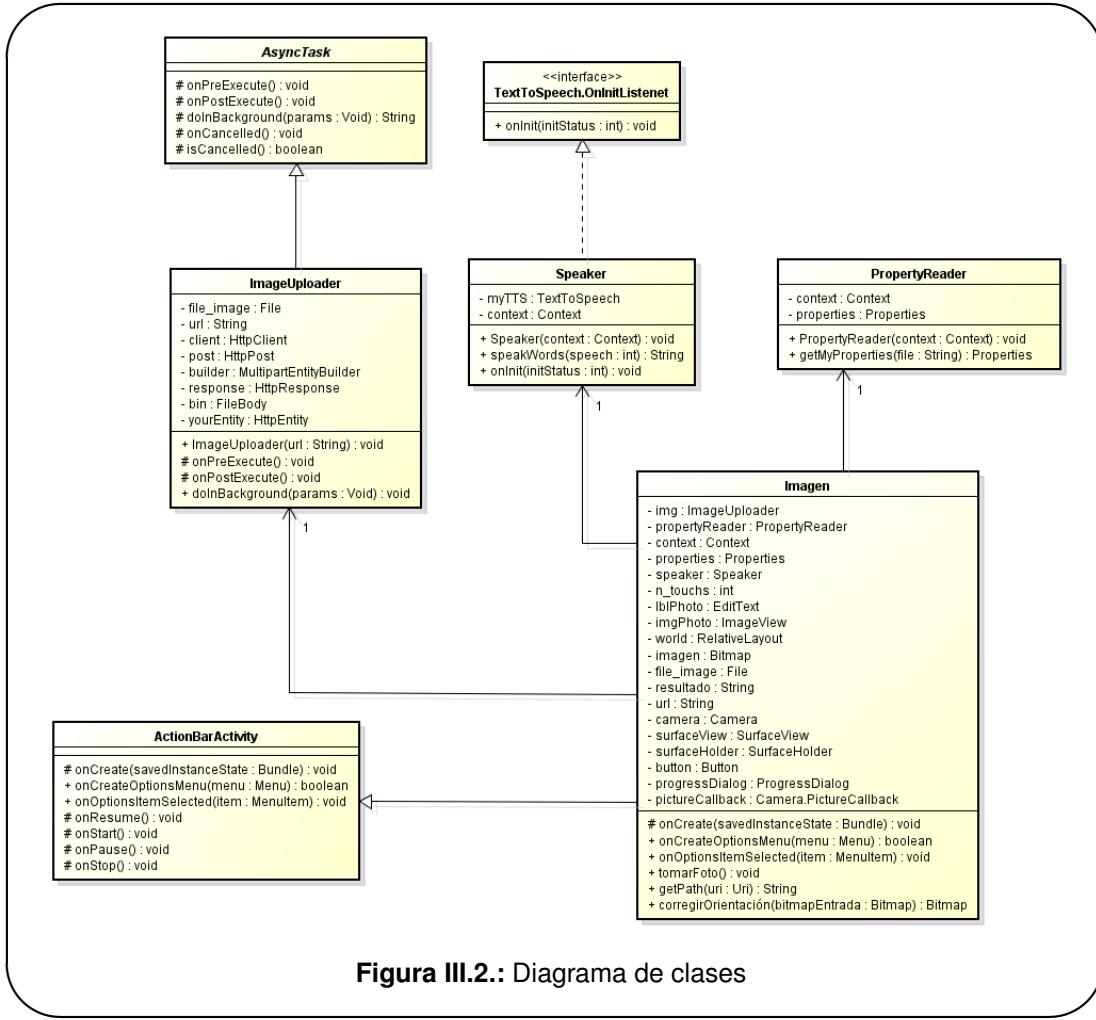


Figura III.2.: Diagrama de clases

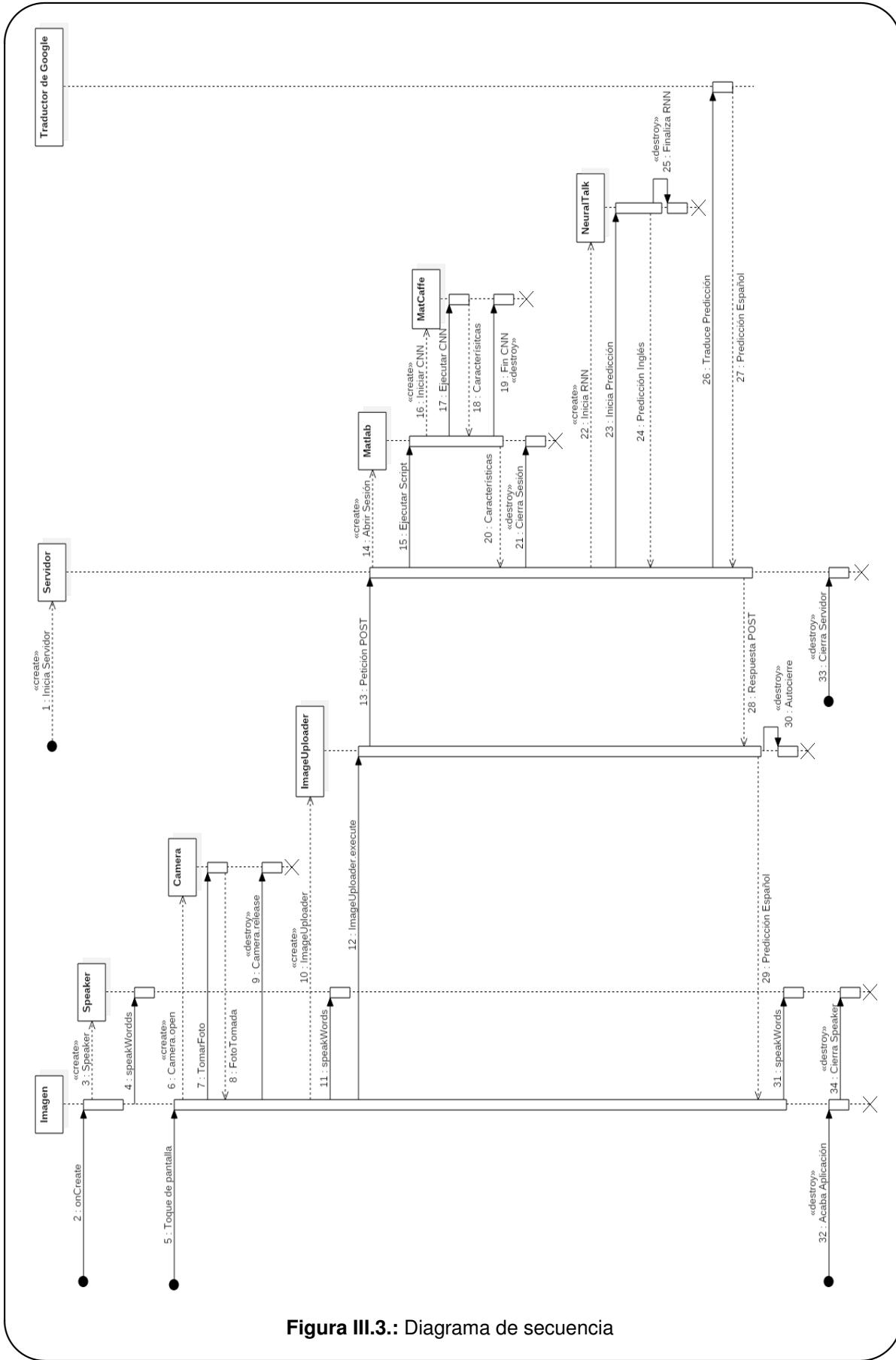
### III.4 Diagrama de clases

En esta sección se presenta el diagrama de clases de la aplicación cliente, que es la más interesante ya que el servidor no está definido en clases. El diagrama de clases se puede ver presentado en la Figura III.2.

### III.5 Diagrama de secuencia del sistema

En esta sección se presenta el diagrama de secuencia del sistema, que se puede ver en la Figura III.3.

### III. ESPECIFICACIÓN DE DISEÑO



**Figura III.3.: Diagrama de secuencia**

**Universidad de Burgos**

Escuela Politécnica Superior

**Ingeniería Informática**

**Área de Lenguajes y Sistemas Informáticos**



**Anexo IV - Manual del programador**

**Estudio de herramientas de reconocimiento de  
imágenes con aplicación prototipo**

**Bryan Reinoso Cevallos**

**Tutores: Dr. José Francisco Díez Pastor,  
Dr. César I. García Osorio**



## IV. MANUAL DEL PROGRAMADOR

---

En esta sección se procederá a la explicación detallada de cómo instalar las herramientas necesarias y qué herramientas son necesarias para trabajar sobre este proyecto.

### IV.1 Instalación del JDK

La primera, y más esencial de las herramientas, es el JDK de java, que es el set o conjunto de herramientas y librerías para los desarrolladores de java.

Primero deberemos ir a la página de Oracle<sup>1</sup> en la que descargaremos el jdk, la página debería tener un aspecto más o menos como el de la imagen IV.1 En dicha página tendremos que aceptar la licencia y posteriormente descargar el JDK que sirva para nuestra máquina. Una vez hemos descargado dicho archivo, lo ejecutamos. Una vez ejecutado seguimos los siguientes pasos para su instalación.

En la imagen IV.2 vemos el primer paso para la instalación del *JDK*, el cual consta de dos pantallas. La primera será solamente una pantalla de bienvenida, por lo que debemos pulsar a *next* o siguiente. La segunda pantalla nos muestra los elementos que van a ser instalados, esto no lo tocamos; además nos muestra también el directorio en el que queremos instalar el *JDK*, esto lo podemos dejar como está o podemos, si queremos, configurarlo en un directorio personal que queramos. Lo único es que habrá que tener cuidado con el *PATH* del equipo para que luego *Android Studio* pueda encontrar la distribución de *JDK* que tengamos en nuestra máquina.

En la imagen IV.3 nos encontramos el paso 2 para la instalación de nuestro *JDK*, este también consta de dos pantallas. La primera es el estado de la instalación, veremos una barra que se irá llenando en función de que la instalación vaya avanzando. Finalmente se nos mostrará la pantalla que nos indicará que la instalación se ha realizado correctamente y nos da la opción de pulsar el botón *Next Steps* que nos llevará a la *API de Java* o simplemente finalizar la instalación.

### IV.2 Instalación de Android Studio

Para la programación del cliente se usará *Android Studio*, una herramienta ideada para programar en *Android* exclusivamente. Lo primero que debemos hacer es dirigirnos a la página web donde descargaremos *Android Studio*<sup>2</sup>, la cual tendrá un aspecto similar al de la IV.4. En ella haremos *click* sobre el botón de descarga (*Download Android Studio*) y procederemos a la instalación ejecutando el fichero que hemos descargado, todo este proceso es para un sistema operativo *Windows*.

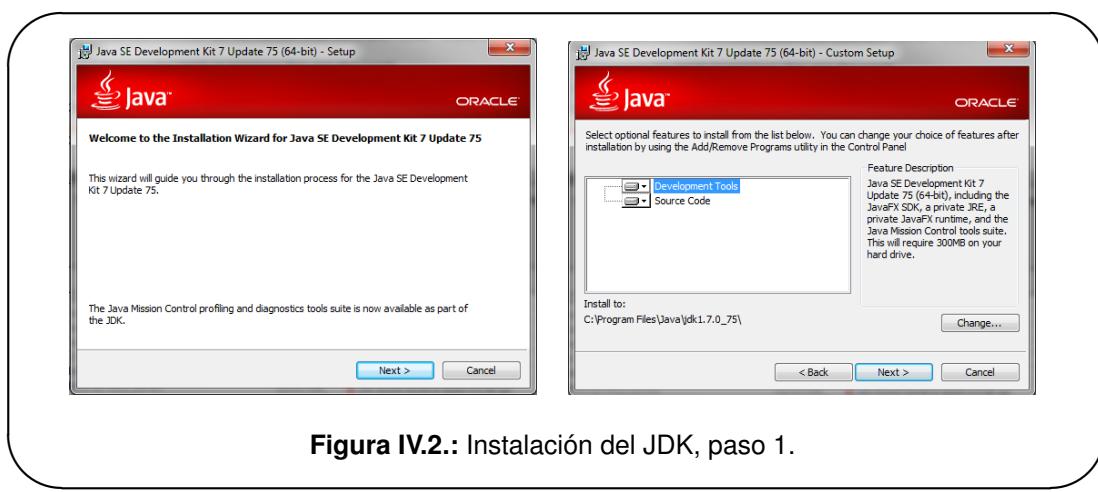
<sup>1</sup><http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html?ssSourceSiteId=otnes>

<sup>2</sup><http://developer.android.com/sdk/index.html>

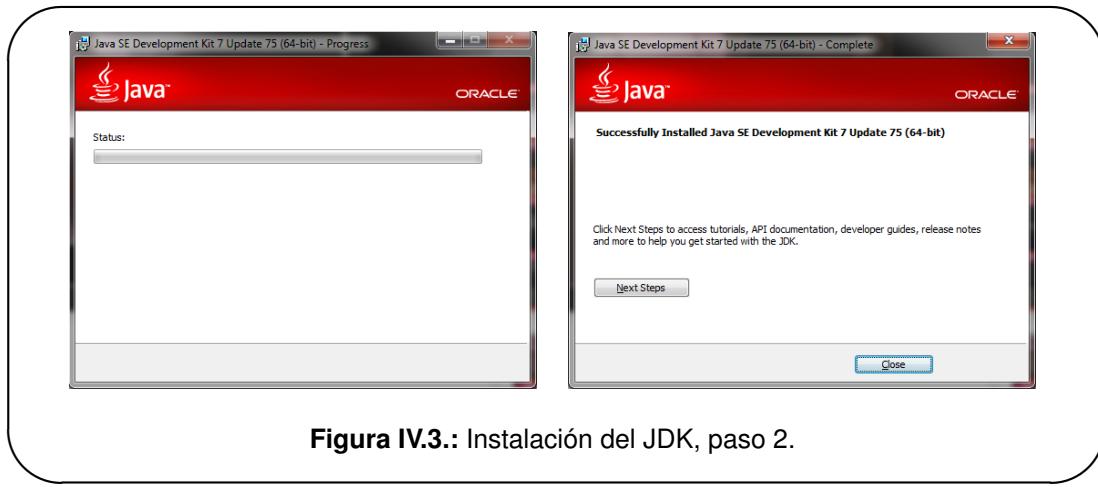
#### IV. MANUAL DEL PROGRAMADOR



**Figura IV.1.: Página de descarga del JDK de Java**



**Figura IV.2.: Instalación del JDK, paso 1.**



**Figura IV.3.: Instalación del JDK, paso 2.**

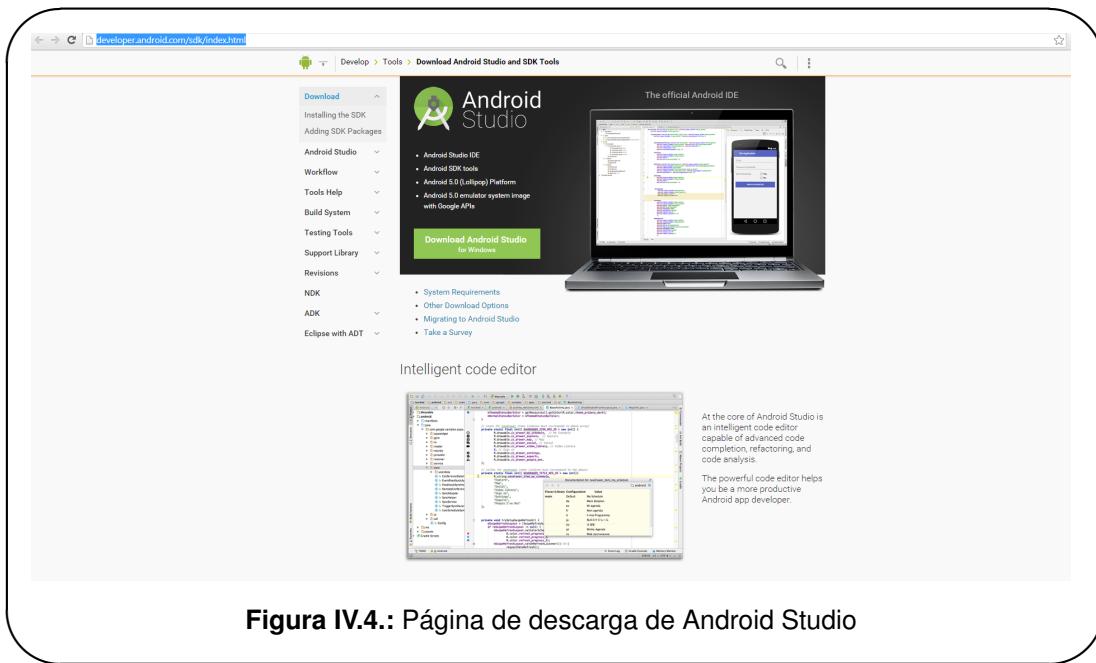


Figura IV.4.: Página de descarga de Android Studio



Figura IV.5.: Instalación de Android Studio, paso 1.



Figura IV.6.: Instalación de Android Studio, paso 2.

#### IV. MANUAL DEL PROGRAMADOR

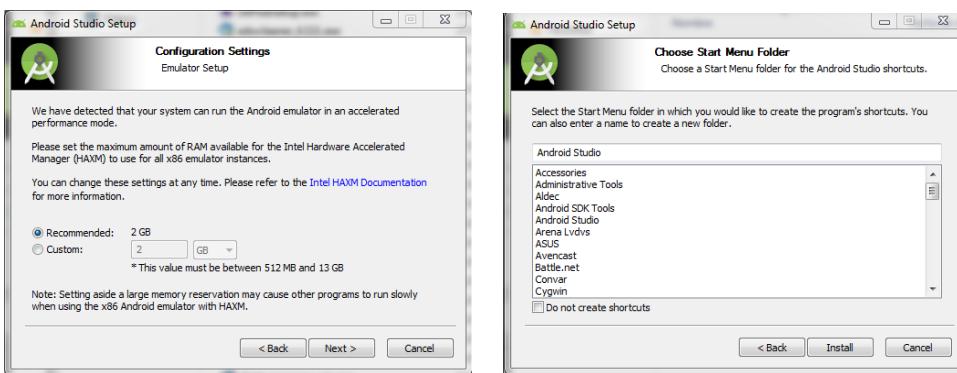


Figura IV.7.: Instalación de Android Studio, paso 3.

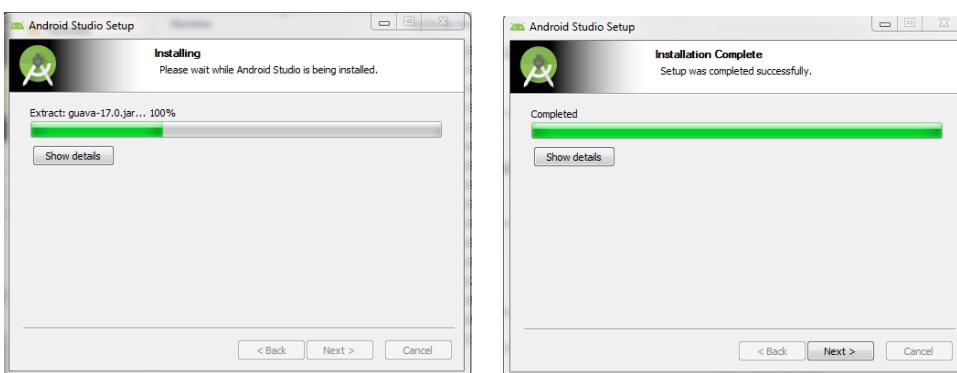


Figura IV.8.: Instalación de Android Studio, paso 4.

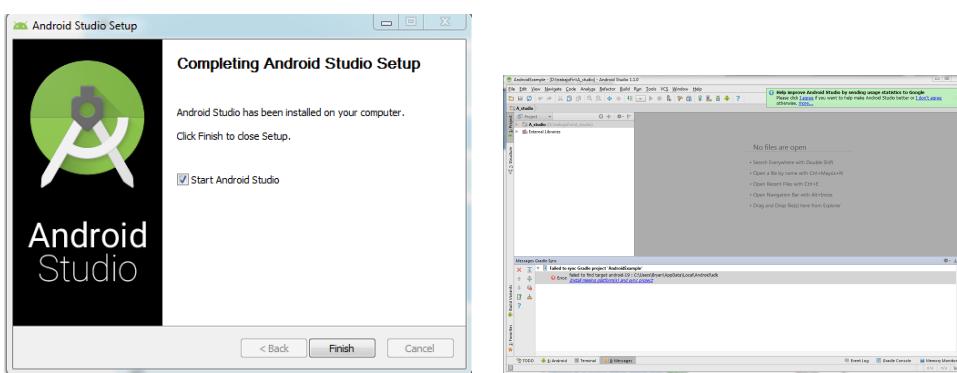


Figura IV.9.: Instalación de Android Studio, paso 5.

En el primer paso de la instalación, que se puede observar en la imagen IV.5, nos encontraremos con dos pantallas. La primera es simplemente una pantalla de bienvenida a la instalación de nuestro *Android Studio*, debemos pulsar al botón *next* para empezar nuestra instalación. Entonces pasamos a la siguiente pantalla y nos da la opción de elegir qué elementos instalar, nosotros hemos dejado todos marcados pero puede que algunos no sean obligatoriamente necesarios. Una vez seleccionados los componentes de nuestra instalación pasamos a dar el botón *next* de la pantalla, entonces saltamos al paso 2.

El segundo paso de nuestra instalación lo podemos observar en la imagen IV.6. Una vez hemos pasado la selección de componentes en nuestro *Android Studio*, tenemos que aceptar la licencia del software, pasaremos a leer la licencia, si es de nuestro interés, y cuando estemos de acuerdo con ella, pulsaremos el botón *I Agree*. Esto no habrá llevado a la siguiente pantalla de la instalación, en la que podremos elegir los directorios en los que instalaremos nuestros componentes, nosotros los hemos dejado por defecto aunque se pueden personalizar en función de las necesidades del usuario. Una vez establecida la configuración de directorios en la instalación, pulsaremos el botón *next*.

Pasaremos entonces al tercer paso de la instalación, la cual puede ser vista en la imagen IV.7. En la primera pantalla de este paso tendremos que elegir la cantidad de memoria *RAM* que asignaremos a nuestro *Android Studio*, podemos dejar la cantidad recomendada o podemos asignarle más cantidad, si disponemos de ella, para que tenga un funcionamiento más fluido. Una vez establecido esto, se pulsara el botón *next*. En nuestra segunda pantalla nos pregunta el directorio en el que se iniciará la aplicación y puedes poner la opción de no crear un acceso directo, nosotros hemos dejado la configuración por defecto. Finalmente pulsamos el botón *Install*.

Ahora en el paso 4, el cual está representado en la imagen IV.8, solamente tendremos que observar cómo avanza la instalación, esto se nos irá mostrando a través de una barra de progreso. Una vez se la barra se ha llenado podremos hacer *click* en el botón *Next* para pasar al último paso de nuestra instalación.

En nuestro quinto y último paso, el cual podemos ver en la imagen IV.9, se nos informará de que la instalación ha terminado y tendremos la opción de iniciar nuestro *Android Studio* nada más acabar la instalación. Entonces pulsamos al botón *Finish*, se nos abrirá nuestro *Android Studio* y con esto la instalación se considera terminada.

### IV.3 SDK Manager, instalando herramientas

Para que la futura importación del proyecto se haga de manera correcta hay que instalar una serie de herramientas, estas pueden ser instaladas a través del *SDK Manager* del *Android Studio*. Para abrir el *SDK Manager*, debemos hacer click sobre el botón resaltado en la primera imagen (IV.10). Seguidamente se nos abrirá una ventana gestora de paquetes de instalación, que se puede ver en la segunda imagen (IV.10). Entonces se seleccionan los paquetes a instalar, se abrirá una nueva ventana (IV.11) para aceptar las licencias, las aceptas y le das a instalar.

Los paquetes que necesitaremos serán:

- Android SDK Tools

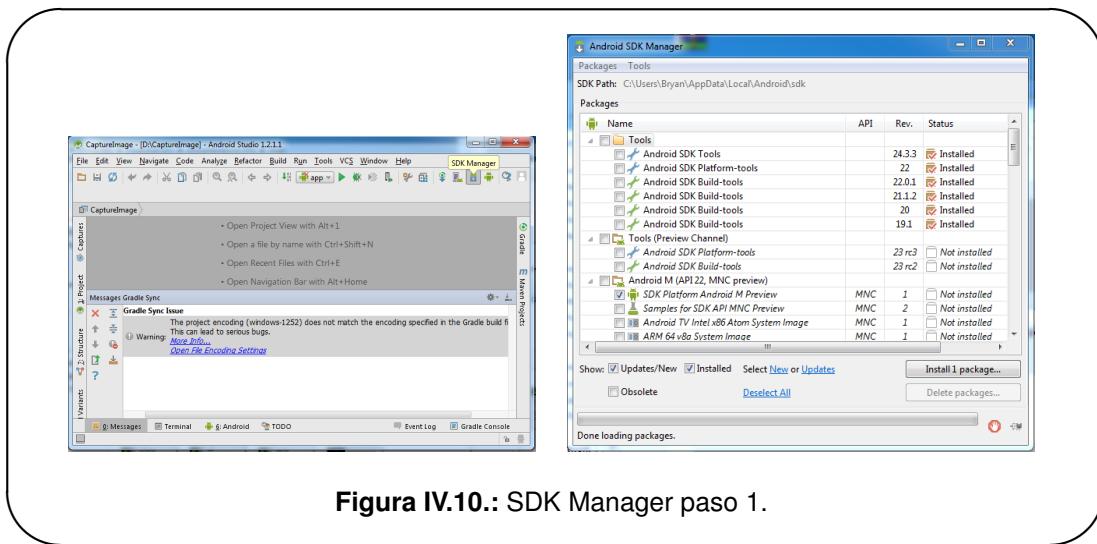


Figura IV.10.: SDK Manager paso 1.

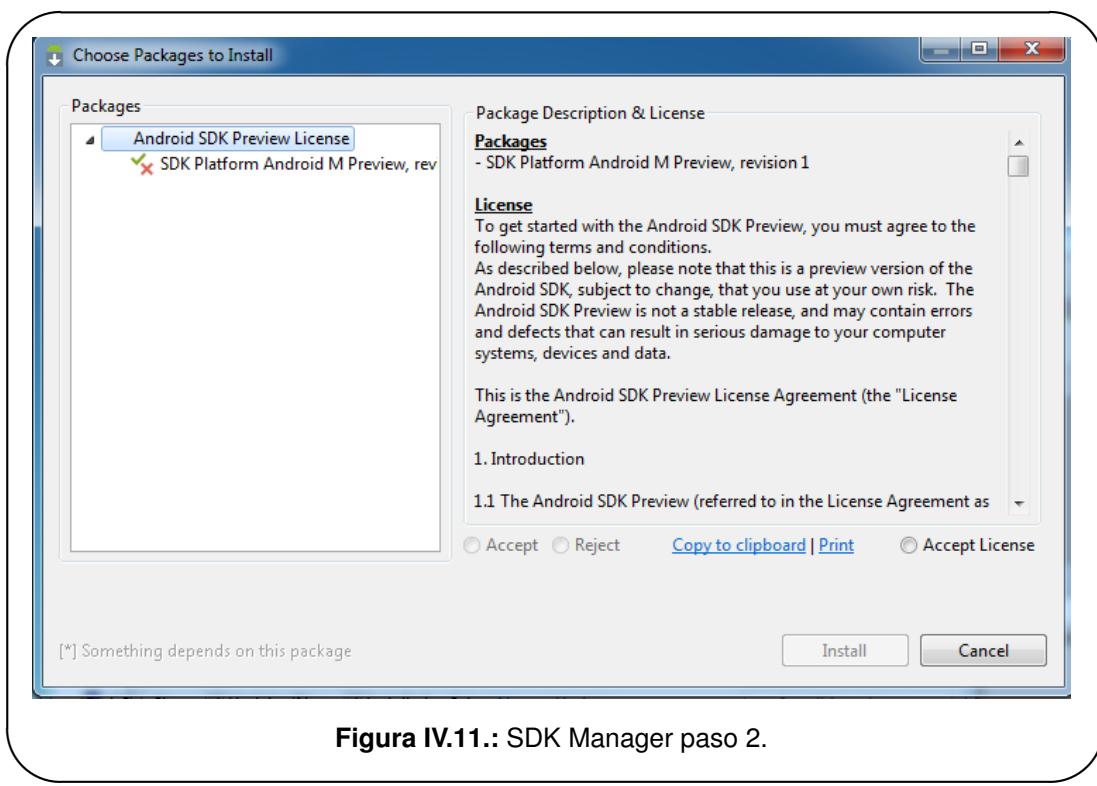
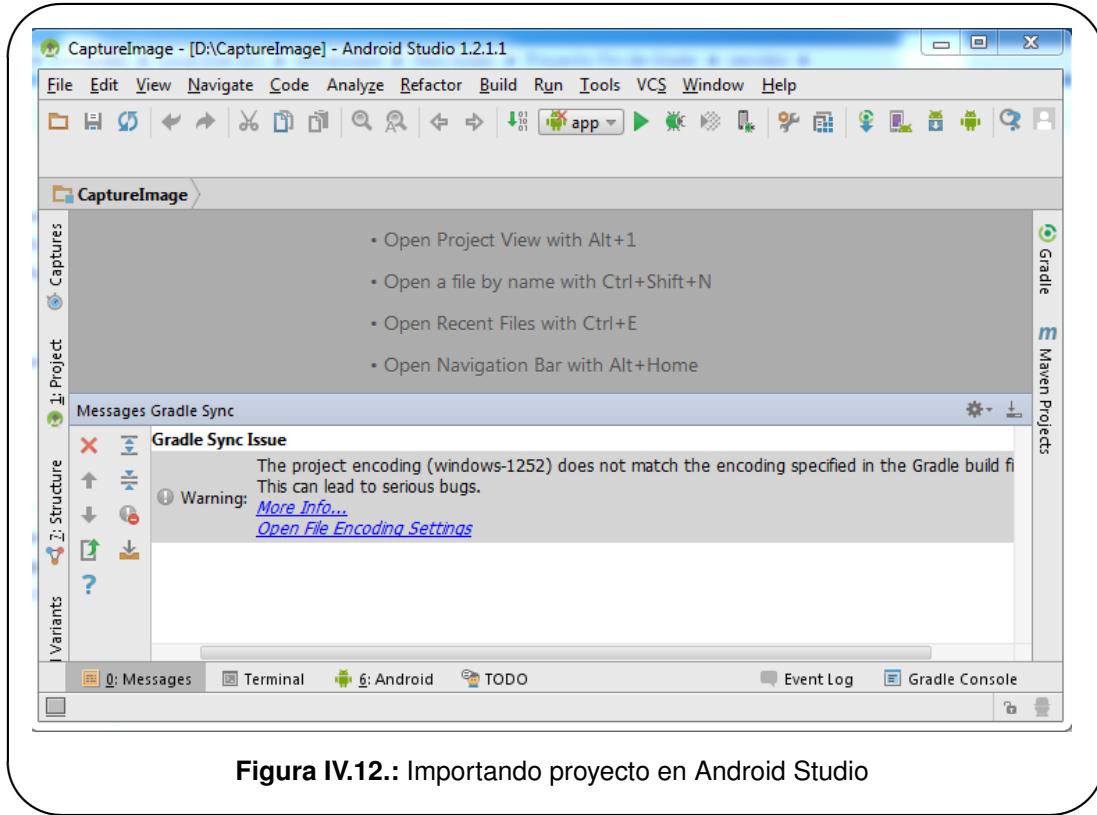


Figura IV.11.: SDK Manager paso 2.



**Figura IV.12.: Importando proyecto en Android Studio**

- Android SDK Platform-tools
- Android SDK Build-tools
- Todo el paquete Android 5.1.1(API 22)
- Android Support Repository
- Android Support Library
- Google Repository
- Google USB Driver
- Google Web Driver

## IV.4 Importando Cliente en Android Studio

Se procede a explicar la manera de importar el proyecto en Android Studio, que es la aplicación con la que se ha trabajado y con la que mejor se puede trabajar sobre el proyecto. Primero deberemos iniciar el Android Studio (IV.12), seguidamente le damos a *File* o Archivo ->*Open...* o *Abrir...* (IV.13). Entonces se nos abrirá una ventana (IV.14) en la que deberemos navegar hasta la carpeta en la que tenemos almacenado el proyecto, seleccionamos el archivo con nombre Imagen y que tiene el icono de Android Studio. Entonces le damos a aceptar y el programa procederá a importar el sólo todo el proyecto (IV.15), al finalizar podremos ponernos enseguida a trabajar.

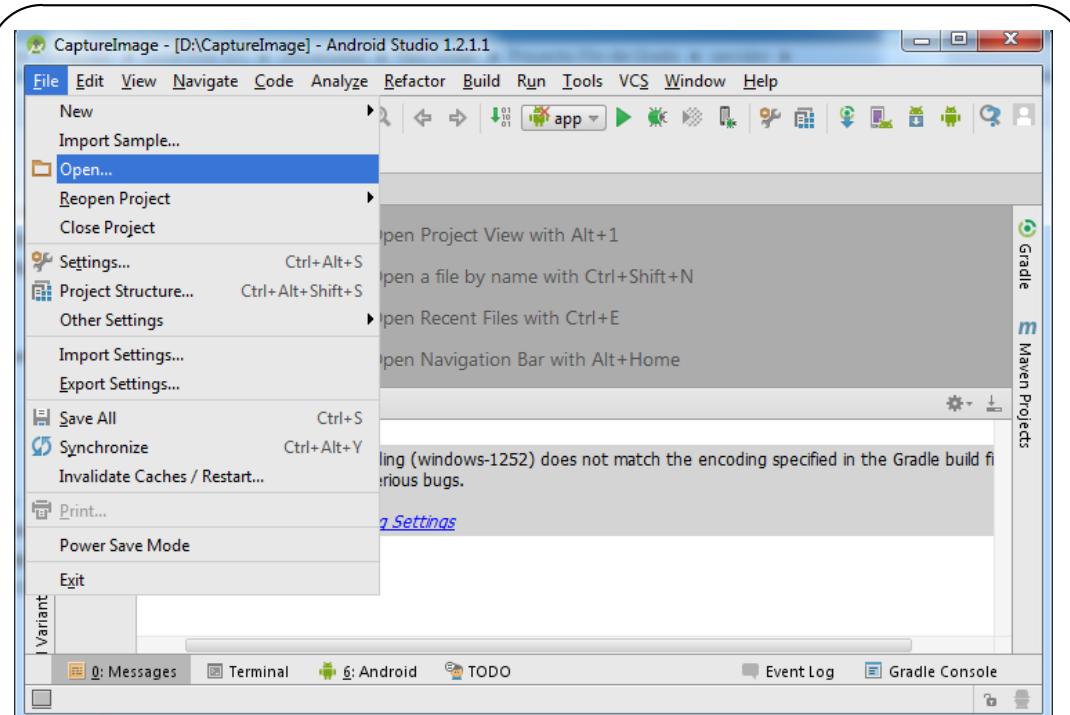


Figura IV.13.: Importando proyecto en Android Studio

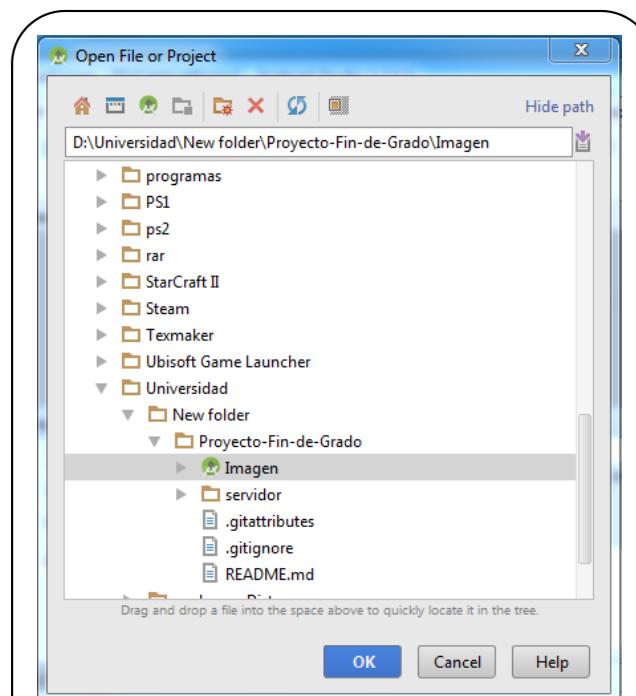
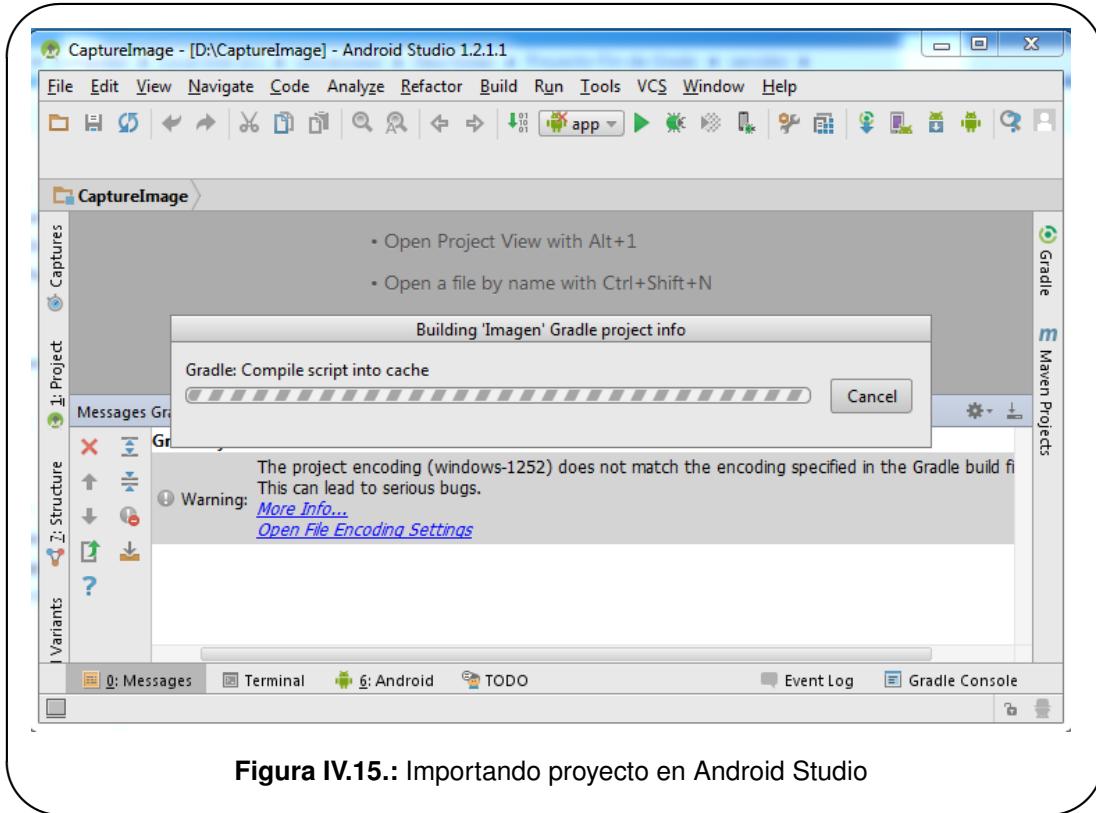


Figura IV.14.: Importando proyecto en Android Studio



## IV.5 Instalación de Matlab

Para poder ejecutar el *script* de instalación del servidor, el cuál se instala en Ubuntu, debemos tener como pre-instalado Matlab. Para ello necesitaremos una licencia, en este proyecto se ha trabajado con la licencia de estudiante de la UBU.

Al trabajar en Ubuntu de 64 bit, hemos descargado el instalador para Linux de Matlab. Hay que recordar que Caffe sólo trabaja con ciertas versiones de Matlab a la hora de descargar el instalador (2014a/b, 2013a/b, and 2012b). Una vez se ha descargado el instalador se tiene que ejecutar el archivo *install* del conjunto de archivos que hemos descargado ([IV.9](#)).

```
1 sudo ./ install
```

**Listado IV.1:** Ejecutando *install* de Matlab

Entonces nos aparecerá una ventana a través de la cual guiaremos nuestra instalación.

Describiremos la instalación con el tipo de licencia que se ha usado en el proyecto.

Primero debemos elegir la primera opción de la primera imagen y después aceptar la licencia de la segunda imagen en el paso 1 ([IV.16](#)).

Seguidamente, en el segundo paso de instalación, deberemos aportar nuestras credenciales, nuestra cuenta de MathWorks, como se muestra en la primera imagen y después seleccionar el lugar de instalación, tal y como se muestra en la segunda imagen del paso 2 ([IV.17](#)). Se recomienda dejar la carpeta de destino de la instalación a la que viene por defecto, porque evitará futuros problemas de configuración cuando instalaremos el servidor y sus dependencias.



Figura IV.16.: Instalación de Matlab paso 1.



Figura IV.17.: Instalación de Matlab paso 2.

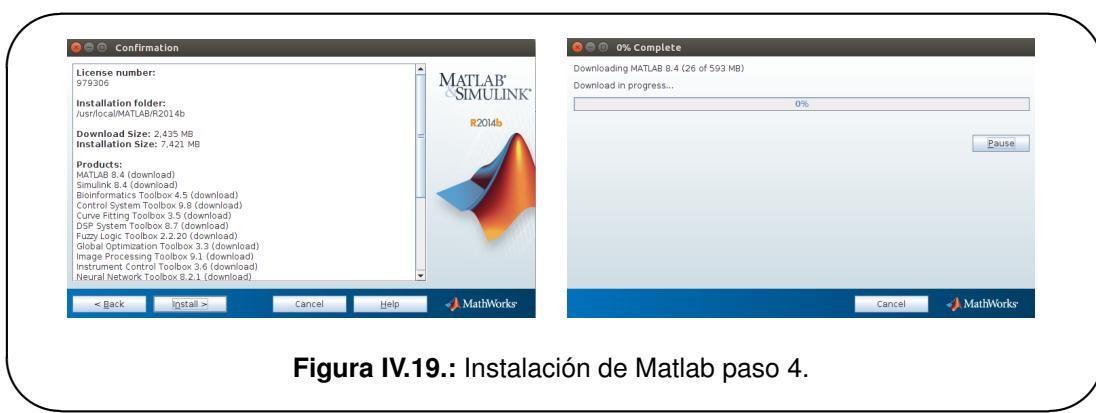
En el tercer paso de la instalación decidiremos si creamos un link para invocar a Matlab desde el terminal, esto se recomienda hacerlo y dejar el link en la ruta por defecto del instalador, como podemos ver en la primera imagen. Después se nos pedirá seleccionar los paquetes que queremos instalar a Matlab, seleccionaremos los básicos, como se ve en la imagen dos del tercer paso (IV.18).

En el cuarto paso se nos pide la confirmación de los paquetes que hemos solicitado instalar en el paso previo, esto puede verse en la primera imagen; comprobaremos que tenemos los paquetes que deseamos y le damos a confirmar. En la segunda imagen vemos la barra de progreso de la instalación que se nos muestra mientras Matlab es instalado (IV.19).

En el quinto y último paso se te muestra la ventana de instalación satisfactoria y te comunica que debes activar Matlab para poder usarlo. Dejamos la opción de “Activar MATLAB” marcada y le damos a siguiente. Así concluimos la instalación de Matlab, pero se tiene que activar antes (IV.20).

#### IV.5.1 Activar MATLAB

Después de la instalación de Matlab se debe haber recibido la solicitud de activación del mismo, por lo que se nos abre una ventana de activación como se puede ver en la primera imagen del paso 1(IV.21). Esta ventana es sólo informativa, por lo que daremos al botón de siguiente. En la segunda imagen del paso 1 podemos ver que nos pide el nombre de usuario con el que se va a querer activar Matlab, se aconseja poner el nombre de usuario normal, no root. Una vez decidido el nombre con el que se va a activar Matlab se pulsa al botón de siguiente.

**Figura IV.18.: Instalación de Matlab paso 3.****Figura IV.19.: Instalación de Matlab paso 4.****Figura IV.20.: Instalación de Matlab paso 5**



**Figura IV.21.: Activación de Matlab paso 1.**



**Figura IV.22.: Activación de Matlab paso 2.**

En el transcurso del primer paso hasta el segundo paso, el activador hará hechas algunas comprobaciones con nuestra licencia y nos muestra una ventana de confirmación, comprobamos que los datos son correctos y pulsamos el botón confirmar, esto se puede ver en la primera imagen del paso 2 (IV.22). Finalmente nos muestra una ventana informativa explicando que la activación ha sido un éxito, como vemos en la segunda imagen del paso 2.

## IV.6 Instalación de CUDA

Si no nos encontramos en una máquina virtual y queremos instalar los *drivers* de NVIDIA en nuestra máquina junto con CUDA, entonces tenemos que seguir unos pasos especiales para hacerlos, los cuales serán explicados en este apartado para el caso de querer instalar controladores de NVIDIA, de lo contrario, sólo se tendrá que descomentar las siguientes líneas del *script* de instalación del servidor y sus dependencias:

```

1 sudo apt-get install curl
2 cd ~/Downloads/
3 curl -O "http://developer.download.nvidia.com/compute/cuda/6_5/rel/installers/
   cuda_6.5.14_linux_64.run"
4 chmod +x cuda_6.5.14_linux_64.run
5 ./cuda_6.5.14_linux_64.run --kernel-source-path=/usr/src/linux-headers-'uname
   -r'/

```

**Listado IV.2: Instalación de CUDA en script**

En caso de querer instalar los controladores de NVIDIA, entonces tendremos que proceder a realizar una operación más compleja para realizar su instalación, debido a que para instalar los controladores no debemos estar en modo interfaz gráfica y tener desactivadas ciertos servicios.

Primero tendremos que desinstalar cualquier rastro de algún controlador de NVIDIA que tengamos previamente en nuestra máquina, lo hacemos con el siguiente comando:

```
1 sudo apt-get remove --purge nvidia*
```

#### **Listado IV.3: Instalación *driver* de NVIDIA paso 1**

Después de esto se deberá desactivar el controlador libre “nveau”, el cual lo desactivamos de la siguiente manera:

- Primero, se tiene que ejecutar el siguiente comando en tu shell:

```
1 sudo nano /etc/modprobe.d/blacklist.conf
```

#### **Listado IV.4: Instalación *driver* de NVIDIA paso 2**

- Segundo, debes añadir lo siguiente al final del fichero, que se ha abierto tras ejecutar el anterior comando, y guardar los cambios:

```
1 blacklist nouveau
```

#### **Listado IV.5: Instalación *driver* de NVIDIA paso 3**

- Tercero, una vez realizado esto tienes que ejecutar los siguientes comando para que los cambios hagan efecto, aunque si no lo hacen tendrás que reiniciar el ordenador:

```
1 echo options nouveau modeset=0 | sudo tee -a /etc/modprobe.d/nouveau-kms.conf
2 update-initramfs -u
```

#### **Listado IV.6: Instalación *driver* de NVIDIA paso 4**

Ahora que se ha desactivado el controlador libre “nveau”, procedemos a la instalación del *driver*, pero este debe ser sin interfaz gráfica. Para trabajar sin interfaz gráfica, primero, debemos pulsar la secuencia de teclas: CTRL+ALT+F1. Esto iniciará un shell sin interfaz gráfica.

Antes que nada debemos identificarnos ante el sistema, para identificarnos deberemos introducir nuestro nombre de usuario y la contraseña, la contraseña se te solicitará una vez hayas introducido el nombre de usuario.

Una vez en el shell se procede a detener el demonio, controlador interno de Unix, que se encarga de mantener en ejecución la interfaz gráfica, para ello tenemos que ejecutar la siguiente orden en el shell:

```
1 sudo service gdm stop
```

#### **Listado IV.7: Instalación *driver* de NVIDIA paso 5**

Ahora se procede a descargar el *driver* con la secuencia que antes hemos comentado sobre el *script* de instalación:

```
1 sudo apt-get install curl
2 cd ~/Downloads/
```

```

3 curl -O "http://developer.download.nvidia.com/compute/cuda/6_5/rel/installers / 
      cuda_6.5.14_linux_64.run"
4 chmod +x cuda_6.5.14_linux_64.run
5 ./cuda_6.5.14_linux_64.run --kernel-source-path=/usr/src/linux-headers-`uname 
      -r`/

```

**Listado IV.8:** Instalación *driver* de NVIDIA paso 6

Una vez ejecutado esto, se iniciará la instalación y deberás aceptar la licencia e instalar todo lo que se te propone, los controladores NVIDIA, CUDA y los ejemplos, todos deben ser instalados en sus directorios por defecto.

Cuando esto haya terminado tendrás ya instalado los *drivers* de NVIDIA, pero aún queda hacer un par de pasos, lanzar el demonio de la interfaz de nuevo y reiniciar el sistema, hazlo con los siguientes comandos:

```

1 sudo service gdm start
2 sudo reboot

```

**Listado IV.9:** Instalación *driver* de NVIDIA paso 7

Con esto ya tendremos el controlador de NVIDIA y CUDA instalados, siendo CUDA necesario para la instalación de las dependencias del servidor y las herramientas.

## IV.7 Instalando Servidor y sus Dependencias

Para instalar el servidor se ha usado en parte una pequeña guía<sup>3</sup>. Gracias a esta guía se instaló correctamente las dependencias, a la que podrás acceder mirando la bibliografía[10].

Para usar el *script* de instalación deberás haber seguido los dos anteriores pasos (IV.6) (IV.5). Ahora para iniciar el proceso de instalación deberás ejecutar el siguiente comando en el *shell*:

```
1 sudo . ./install .sh
```

**Listado IV.10:** Instalar Servidor

**IMPORTANTE:** Si te encuentras en una máquina virtual, deberás ir al *script* de instalación y modificarlo **antes de ejecutar la instalación**. Deberás cambiar la linea 47 de la siguiente manera:

```

1 #Esta es la línea original en el script :
2 cp ~/Proyecto-Fin-de-Grado/servidor/Install/Makefile.config ~/caffe
3
4 #Esto es lo que se debe poner:
5 cp ~/Proyecto-Fin-de-Grado/servidor/Install/MaquinaVirtual/Makefile.config ~/ 
  caffe

```

**Listado IV.11:** Cambio para máquinas virtuales

Si tienes algún problema con la instalación o con el *script*, se procede a explicar paso a paso qué hace el programa de instalación y qué deberías hacer tú para instalarlo sin el uso del *script*. Todo esto debe ser siempre realizado tras haber hecho los anteriores dos pasos (IV.6) (IV.5).

<sup>3</sup><https://github.com/BVLC/caffe/wiki/Ubuntu-14.04-VirtualBox-VM>

En primer lugar se va a un directorio por defecto para hacer la instalación y seguidamente se instala la herramienta “git” para clonar los proyectos desde Github.

```
1 cd ~
2 apt-get install git
```

**Listado IV.12:** Explicación Instalación 1

Vemos que hemos hecho la acción de irnos a un directorio por defecto, este directorio sera la base de toda la instalación y, por tanto, no se recomienda cambiarlo para que la configuración del servidor sea más sencilla.

Ahora se procede a instalar los paquetes esenciales de Linux, además de actualizar las cabeceras del *kernel* a su última versión para evitar errores futuros.

```
1 apt-get install build-essential
2 apt-get install linux-headers-`uname -r`
```

**Listado IV.13:** Explicación Instalación 2

En este paso clonaremos nuestro proyecto al directorio principal de instalación.

```
1 cd ~
2 git clone https://github.com/garfio1/Proyecto-Fin-de-Grado.git
```

**Listado IV.14:** Explicación Instalación 3

En el siguiente paso se ha copiado la línea de instalación de dependencias de la Wiki[10] de Github para instalar Caffe. Esto instalará todas las dependencias necesarias para Caffe.

```
1 apt-get install -y libprotobuf-dev libleveldb-dev libsnappy-dev libopencv-dev
    libboost-all-dev libhdf5-serial-dev protobuf-compiler gfortran libjpeg62
    libfreeimage-dev libatlas-base-dev git python-dev python-pip libgoogle-
    glog-dev libbz2-dev libxml2-dev libxslt-dev libffi-dev libssl-dev libgflags
    -dev liblmdb-dev python-yaml
2 easy_install pillow
```

**Listado IV.15:** Explicación Instalación 4

Ahora se procede a clonar Caffe[9] en nuestro directorio principal de instalación.

```
1 cd ~
2 git clone https://github.com/BVLC/caffe.git
```

**Listado IV.16:** Explicación Instalación 5

Ahora se instalan todos los requisitos para poder instalar PyCaffe, esto no debería ser necesario porque no usamos PyCaffe, pero la evolución del proyecto NeuralTalk apunta a que acabará usando esta librería de Caffe para Python, así que se instala aquí también para tener trabajo ya hecho.

```
1 cd caffe
2 cat python/requirements.txt | xargs -L 1 sudo pip install
```

**Listado IV.17:** Explicación Instalación 5

Ahora se procede a hacer unos enlaces con los nombres de las librerías Python para que PyCaffe funcione de manera adecuada y correcta.

```

1 sudo ln -s /usr/include/python2.7/ /usr/local/include/python2.7
2 sudo ln -s /usr/local/lib/python2.7/dist-packages/numpy/core/include/numpy/ /
    usr/local/include/python2.7/numpy

```

**Listado IV.18:** Explicación Instalación 6

En el siguiente paso el programa de instalación supone que has seguido todos los pasos hasta ahora como se han ido comentando y sustituye el Makefile.config de Caffe por el que viene en el proyecto. Puesto que si has instalado las cosas según se han ido comentando, este archivo debería servirte para poder hacer los *make* (instalación a través de un Makefile) de Caffe.

```

1 rm ~/caffe/Makefile.config
2 cp ~/Proyecto-Fin-de-Grado/servidor/Install/Makefile.config ~/caffe

```

**Listado IV.19:** Explicación Instalación 7

Finalmente se ejecuta la instalación de Caffe a través de la herramienta *make*, se instalan tanto MatCaffe como PyCaffe.

```

1 make pycaffe
2 make matcaffe
3 make all
4 make test

```

**Listado IV.20:** Explicación Instalación 8

En este punto nos falta el clonado del proyecto NeuralTalk en nuestra máquina, por tanto, es lo que se procede a hacer en el penúltimo paso de la instalación

```

1 cd ~
2 git clone https://github.com/karpathy/neuraltalk.git

```

**Listado IV.21:** Explicación Instalación 9

Finalmente nos resta únicamente instalar las librerías específicas usadas en el servidor para pasar a la configuración del proyecto para que funcione de manera adecuada.

```

1 pip install beautifulsoup
2 pip install flask

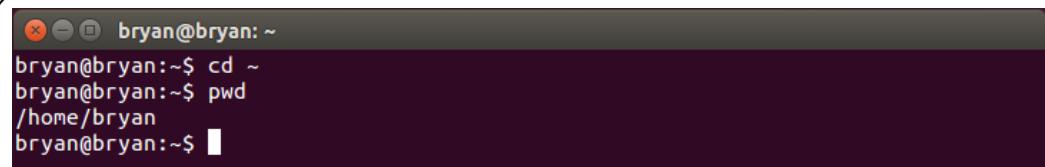
```

**Listado IV.22:** Explicación Instalación 10

Con esto tenemos instalado en nuestra máquina todas las herramientas necesarias para el funcionamiento del servidor. Ahora hay que pasar a la configuración del mismo. Como se ha explicado en el apartado de Aspectos Relevantes (6), la configuración no ha sido nada sencilla; pero se ha preparado para que ahora sea mucho más sencilla y rápida.

## IV.8 Configuración del Servidor

En este paso se procederá a explicar cómo se configura el servidor para que funcione, ahora resulta una tarea bastante sencilla porque se han preparado una serie de variables y ficheros para que esto sea así.



```
bryan@bryan:~$ cd ~
bryan@bryan:~$ pwd
/home/bryan
bryan@bryan:~$
```

**Figura IV.23.:** Comprobando directorio principal de instalación

En primer lugar, y lo más importante, es tener claro el directorio principal de la instalación, pues todas las herramientas deberían estar instaladas en él. Si se ha seguido correctamente los pasos y se ha mantenido el mismo directorio principal de instalación, deberías ser capaz de obtenerlo con la siguiente secuencia de comandos:

- 1 cd ~
- 2 pwd

**Listado IV.23:** Configurando el servidor 1

El resultante de ejecutar los anteriores comandos es una cadena de caracteres que te comunican cuál es el directorio principal de la instalación, por ejemplo a mí me devuelve “/home/bryan”, como se puede ver en la imagen IV.23.

Antes de continuar con la instalación, debemos cambiar los permisos de las carpetas NeuralTalk y Proyecto-Fin-de-Grado para poder hacer la configuración sobre ellas, esto se hace de la siguiente manera: lugar ejecutaremos el siguiente comando para empezar con la configuración:

- 1 chmod -R 777 ~/Proyecto-Fin-de-Grado
- 2 chmod -R 777 ~/neuraltalk

**Listado IV.24:** Configurando el servidor 2

Una vez se ha detectado este directorio, podemos empezar con la configuración del servidor. En primer lugar ejecutaremos el siguiente comando para empezar con la configuración:

- 1 cd ~/Proyecto-Fin-de-Grado/servidor

**Listado IV.25:** Configurando el servidor 3

Con el anterior comando nos situamos en la carpeta en la que tendremos el primer documento a modificar. Tendremos que abrir con un editor de texto el archivo “serv.py” que nos encontramos en dicha carpeta. Podemos hacerlo con el siguiente comando si queremos, aunque el procesador de texto es a elección del usuario:

- 1 vi serv.py

**Listado IV.26:** Configurando el servidor 4

Una vez tengamos abierto el archivo, deberemos localizar dos líneas en concreto. Las líneas a localizar son las que se define la ruta del directorio principal de instalación y en la que se define el usuario definido en la licencia de MATLAB. Podemos ver las líneas a continuación:

- 1 #Deberían estar en las líneas 12 y 24 respectivamente pero si no lo están sólo busca lo siguiente :

```

2 ROOT='/home/bryan'
3
4 USER='bryan'
```

**Listado IV.27:** Configurando el servidor 5

Una vez hemos localizado las líneas anteriores, en la constante “ROOT” deberás poner el directorio que obtuviste en IV.23 o tu directorio principal de instalación, el código está preparado para que, si has instalado las herramientas en el directorio correcto, no tengas que modificar más variables ni que estar picando el código en busca de problemas; y en la constante “USER”, deberás poner el nombre de usuario que utilizaste al realizar el paso de Activación de MATLAB(IV.5.1). Con estas modificaciones ya hemos configurado la mitad del servidor, sólo nos queda ahora hacer las modificaciones pertinentes para que la herramienta NeuralTalk funcione en nuestra máquina.

Para empezar, junto con el proyecto se facilita una copia del archivo “extract\_features.m” de NeuralTalk, sólo que este archivo está modificado para que la configuración sea muy sencilla. Lo que haremos será borrar el archivo que viene por defecto junto con la herramienta NeuralTalk y se copiará el que viene adjunto con el proyecto, lo haremos de la siguiente manera:

```

1 rm ~/neuraltalk / matlab_features_reference / extract_features .m
2 cp ~/Proyecto—Fin—de—Grado/servidor/Install/ extract_features .m ~/neuraltalk /
   matlab_features_reference /
```

**Listado IV.28:** Configurando el servidor 6

Ahora debemos dirigirnos a la carpeta en la que acabamos de copiar el fichero y abrirlo para proceder a modificarlo, esto se puede hacer vía comandos de shell o puede el usuario abrirlo con cualquier procesador de texto que el prefiera:

```

1 cd ~/neuraltalk / matlab_features_reference /
2 vi extract_features .m
```

**Listado IV.29:** Configurando el servidor 7

Una vez tenemos el fichero abierto para su modificación debemos buscar una línea si estamos en una máquina anfitriona, dos si estamos en una máquina virtual:

```

1 %Si estamos en máquina virtual buscamos también la siguiente línea , típicamente
   está en la línea 3:
2 use_gpu = 1;
3
4 %Debería estar en la línea 5
5 root='/home/bryan';
```

**Listado IV.30:** Configurando el servidor 8

Si nos encontramos en una máquina virtual, la variable “use\_gpu” deberíamos ponerla con el valor 0. Mientras que lo siguiente se hace tanto para máquina virtual como para máquina anfitriona, se cambiará la variable “root” por el directorio que se obtuvo en IV.23.

Con esto ya tenemos hecho las modificaciones pertinentes para que el servidor funcione correctamente, sólo nos queda descargar un fichero necesario para la ejecución del *script* de MATLAB. El fichero en cuestión será descargado de la siguiente manera:

```

1 cd ~/Proyecto-Fin-de-Grado/servidor/uploads
2 wget "http://www.robots.ox.ac.uk/~vgg/software/very_deep/caffe /
VGG_ILSVRC_16_layers.caffemodel"

```

**Listado IV.31:** Configurando el servidor 9

Ahora ya podemos ejecutar el servidor, para ello podremos ejecutar los siguientes comandos:

```

1 cd ~/Proyecto-Fin-de-Grado/servidor
2 sudo python serv.py

```

**Listado IV.32:** Ejecutando el servidor

Ahora hay que tener en cuenta que, si has realizado la instalación sobre una máquina anfitriona, podría darse un error al ejecutar el servidor, este error mostrará un mensaje al ejecutar Matlab, que será el siguiente:

```
1 Cannot open shared object file : No such file or directory
```

**Listado IV.33:** Error en CUDA

Para resolver este error solamente debemos ejecutar el siguiente comando:

```
1 sudo ldconfig /usr/local/cuda/lib64
```

**Listado IV.34:** Solución al error de CUDA

Como punto final cabe destacar que este tutorial de instalación funciona para las herramientas en la versión en la que se encontraban al realizar el proyecto, estas podrán ser encontradas en el CD entregado por si este tutorial no funcionara con las versiones actuales.



# **Universidad de Burgos**

**Escuela Politécnica Superior**

## **Ingeniería Informática**

**Área de Lenguajes y Sistemas Informáticos**



### **Anexo V - Manual del usuario**

### **Estudio de herramientas de reconocimiento de imágenes con aplicación prototipo**

**Bryan Reinoso Cevallos**

**Tutores: Dr. José Francisco Díez Pastor,  
Dr. César I. García Osorio**



# V. MANUAL DEL USUARIO

---

## V.1 Introducción

En este anexo se presentará el manual de usuario, dónde se explicará cómo instalar y usar la aplicación.

### V.1.1 Instalar Aplicación

Si queremos extraer la aplicación del proyecto, deberemos dirigirnos al directorio «/Proyecto-Fin-de-Grado/Imagen/app», en él encontraremos un archivo llamado `app-release.apk`, este archivo se guardará en el dispositivo en el que se vaya a instalar la aplicación. Hay que destacar que la instalación de la aplicación se debe hacer de forma manual, aún no se ha planificado ni modelado una manera de que una persona invidente pueda hacerlo, por tanto esta deberá tener alguien que le ayude a instalarla.

Para empezar la instalación, tenemos que buscar en los archivos del dispositivo el fichero de instalación con el nombre `app-release.apk`, una vez encontrado tocamos la pantalla sobre este y se abrirá una ventana de instalación. En la ventana de instalación se nos muestra los permisos que requiere la aplicación y nos da la opción de instalar, presionamos al botón de instalar para proceder con la instalación. Estos pasos están representados en la imagen ??.

Después de haber presionado el botón de instalar, se nos muestra una ventana que nos informa de que la aplicación se está instalando. Seguidamente, una vez se ha instalado la aplicación, se mostrará una ventana informándonos de que la aplicación se ha instalado correctamente y nos dará la opción de abrirla. Esto esta representado en la imagen ??.

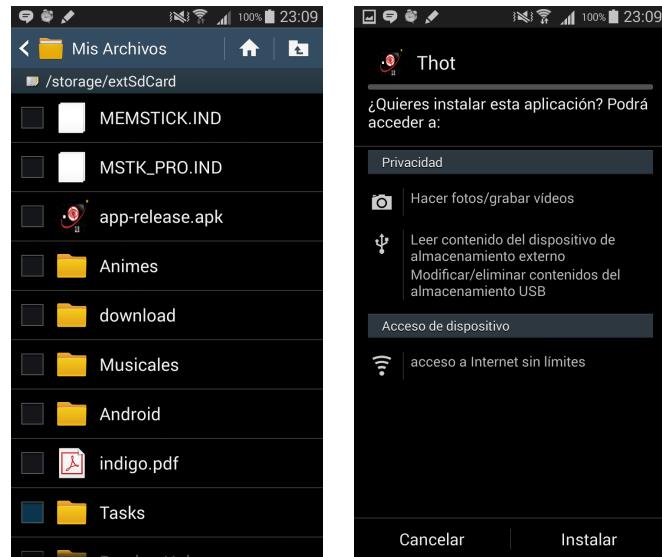
Si hemos abierto la aplicación, se nos mostrará la pantalla inicial (ver Figura V.3).

## V.2 Uso de la aplicación

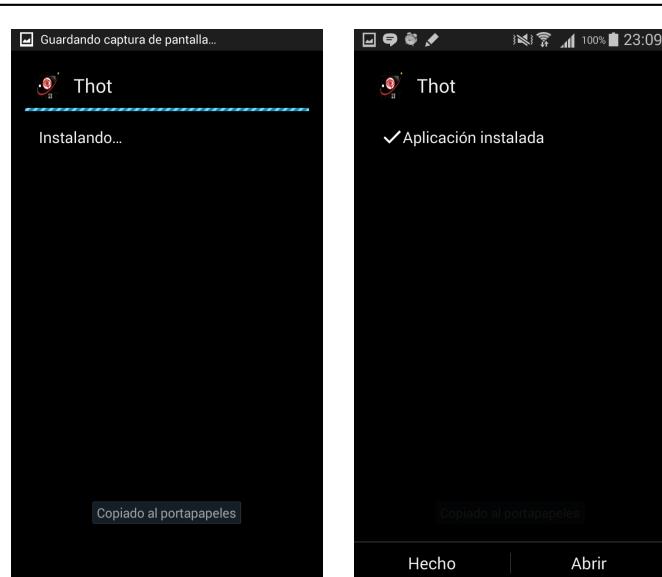
Cuando se ha iniciado la aplicación, y se nos muestra la pantalla inicial (ver Figura V.3), el usuario deberá apuntar con el móvil en la dirección que se quiera tomar la foto y tocar la pantalla. Entonces la aplicación se encargará de mandar la foto y recibir la predicción.

Si el usuario no sabe qué tiene que hacer cuando abre la aplicación, pasado un tiempo, se le lee un mensaje en voz alta con la librería *Text to Speech* diciéndole qué tiene que hacer.

Tras la ejecución de la petición al servidor y el procesado de la imagen, la aplicación leerá en voz alta la predicción. Y aquí termina la ejecución de la aplicación, si se quiere hacer otra predicción, se debe esperar a que la aplicación indique que podemos tomar otra foto.



**Figura V.1.: Instalando aplicación paso 1.**



**Figura V.2.: Instalando aplicación paso 2.**





# BIBLIOGRAFÍA

---

- [1] Edwin R Addison, H Donald Wilson, Gary Marple, Anthony H Handal, and Nancy Krebs. Text to speech, March 8 2005. US Patent 6,865,533.
- [2] Alex Berg, Jia Deng, and L Fei-Fei. Large scale visual recognition challenge 2010, 2010.
- [3] Dan C Cireşan, Ueli Meier, Jonathan Masci, Luca M Gambardella, and Jürgen Schmidhuber. High-performance neural networks for visual object classification. *arXiv preprint arXiv:1102.0183*, 2011.
- [4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.
- [5] American Foundation for the Blind. A review of the taptapsee, camfind, and talking goggles object identification apps for the iphone, 2012. URL: <http://www.afb.org/afbpress/pub.asp?DocID=aw140704>.
- [6] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jagannath Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 580–587. IEEE, 2014.
- [7] Ozan Irsoy. Deep learning with decision trees, 2014. URL: <https://regularizer.wordpress.com/2014/11/18/deep-learning-with-decision-trees/>.
- [8] Asha Iyer, Christof Koch, and Pietro Perona. What do we perceive in a glance of a real-world scene? In *J Vision*.
- [9] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.
- [10] Andrej Karpathy. GitHub Wiki ubuntu 14.04 virtualbox vm. <https://github.com/BVLC/caffe/wiki/Ubuntu-14.04-VirtualBox-VM>.
- [11] Andrej Karpathy. Cs231n: Convolutional neural networks for visual recognition, 2014-2015. URL: <http://cs231n.github.io/convolutional-networks/>.
- [12] Andrej Karpathy. The unreasonable effectiveness of recurrent neural networks, 2014-2015. URL: <http://karpathy.github.io/2015/05/21/rnn-effectiveness/>.
- [13] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *arXiv preprint arXiv:1412.2306*, 2014.
- [14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. URL: <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>

classification-with-deep-convolutional-neural-networks.pdf.

- [15] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814, 2010.
- [16] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, pages 1–42, 2014.
- [17] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3):157–173, 2008.
- [18] Ken Schwaber. *Agile project management with Scrum*. Microsoft Press, 2004.
- [19] Tad Slaff. How to trade the rsi: An analysis using a support vector machine, 2014. URL: <https://www.inovancetech.com/how-to-trade-rsi.html>.
- [20] Kardi Teknomo. How to use a decision tree?, 2009. URL: <http://people.revoledu.com/kardi/tutorial/DecisionTree/how-to-use-decision-tree.htm>.
- [21] Wikipedia. Aprendizaje automático — wikipedia, la enciclopedia libre, 2015. URL: [https://es.wikipedia.org/w/index.php?title=Aprendizaje\\_autom%C3%A1tico&oldid=83177301](https://es.wikipedia.org/w/index.php?title=Aprendizaje_autom%C3%A1tico&oldid=83177301).
- [22] Wikipedia. Aprendizaje profundo — wikipedia, la enciclopedia libre, 2015. URL: [https://es.wikipedia.org/w/index.php?title=Aprendizaje\\_profundo&oldid=83177869](https://es.wikipedia.org/w/index.php?title=Aprendizaje_profundo&oldid=83177869).
- [23] Wikipedia. Regresión lineal — wikipedia, la enciclopedia libre, 2015. URL: [https://es.wikipedia.org/w/index.php?title=Regresi%C3%B3n\\_lineal&oldid=82160733](https://es.wikipedia.org/w/index.php?title=Regresi%C3%B3n_lineal&oldid=82160733).
- [24] Wikipedia. Servicio web — wikipedia, la enciclopedia libre, 2015. URL: [https://es.wikipedia.org/w/index.php?title=Servicio\\_web&oldid=83082613](https://es.wikipedia.org/w/index.php?title=Servicio_web&oldid=83082613).
- [25] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.
- [26] Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2015.