

Implementación de una versión del juego “¿Quién es Quién?” mediante CLIP (Contrastive Language-Image Pretraining)

TITULACIÓN:

Máster en Inteligencia Artificial

Curso académico:

2020-2021

Lugar de residencia, mes y año:

Manresa (Catalunya), 11 / 2021

Alumno/a:

Arnau Martí Sarri

D.N.I:

39384366V

Director:

Rodríguez Fernández, Víctor

Convocatoria:

Primera

Orientación:

Créditos:

viu

Universidad
Internacional
de Valencia

Resumen

Este trabajo describe el proceso seguido para crear una aplicación web basada en el juego *¿Quién es Quién?*. Al ser un juego de preguntas con respuestas de *Sí* o *No*, será necesario un sistema capaz de interpretar las preguntas de los usuarios para responder de forma binaria. Se usará CLIP, *Contrastive Language-Image Pretraining*, un modelo presentando recientemente que será el motor del juego y con el que se efectuarán diferentes experimentos. Todo el proyecto se realiza con Python, desde la manipulación de CLIP hasta el desarrollo *Front-End*, este último mediante la plataforma *streamlit*.

Índice

1.	Objetivos	5
2.	Estado del arte	6
3.	Desarrollo del Proyecto	8
3.1.	Conjunto de datos “Celeba”	9
3.2.	CLIP	15
3.2.1.	Método de clasificación “Zero-Shot”	15
3.2.2.	Método para aplicar “Fine-Tuning” a CLIP	17
3.3.	Streamlit	19
3.4.	Descripción del juego.....	20
3.4.1.	¿Cómo jugar?	22
3.4.2.	Puntuación	26
3.5.	Experimentos.....	28
4.	Resultados	31
4.1.	Frase neutral vs Descripción.....	32
4.2.	Frase neutral vs Descripción opuesta	37
4.3.	Descripción vs Descripción opuesta	40
4.4.	Múltiples descripciones.....	44
4.5.	Aplicar “Fine Tuning”	49
5.	Valoración y discusión de los resultados	53
6.	Conclusiones y desarrollos futuros.....	56

7.	Referencias bibliográficas	59
8.	Anexos	61
8.1.	Recursos	61
8.1.1.	Conjunto de datos “Celeba”	61
8.1.2.	CLIP - Github.....	61
8.1.3.	CLIP - web.....	61
8.1.4.	Streamlit	61
8.1.5.	Quien es Quien - Github	61
8.1.6.	Quien es Quien - archivo “.pdf” de los experimentos	61
8.1.7.	Quien es Quien - código Python de los experimentos.....	61
8.1.8.	Quien es Quien - código Python de la aplicación web	62
8.1.9.	Quien es Quien - aplicación web	62
8.2.	Artículo publicado	63

1. Objetivos

El objetivo principal de este Trabajo de Fin de Máster es crear una aplicación basada en el juego “Quién es Quién”, para un jugador, utilizando los conocimientos aprendidos durante el curso y las últimas novedades del mundo de la inteligencia artificial.

El juego consiste en encontrar una imagen ganadora seleccionada de un conjunto de imágenes posibles. Para ello el jugador puede realizar preguntas que tengan dos posibles respuestas, “Sí” o “No”. Cada vez que realiza una pregunta se descartan las imágenes cuya respuesta a la pregunta no coincide con la respuesta de la ganadora y el juego finaliza cuando el jugador descubre cual es la imagen ganadora.

Teniendo en cuenta el desarrollo del juego, para realizar una versión de éste que proponga un reto a los diseños basados en Inteligencia Artificial, se persigue el objetivo de permitir al jugador realizar su propia pregunta. Esto significa que se necesitará un método para dividir un conjunto de imágenes en dos grupos a partir de la pregunta introducida por el jugador, de forma que se puedan descartar las imágenes del grupo que no incluye la ganadora. También se pretende que el jugador pueda usar su propio conjunto de imágenes de personas para jugar, con lo que la aplicación debe ser capaz de lidiar con imágenes nuevas. Por lo tanto, será necesario idear un sistema “Zero Shot” capaz de clasificar imágenes a partir de uno o varios textos. Es decir, que permita al usuario introducir nuevas preguntas o nuevos conceptos de una forma que no sea necesario entrenar previamente ningún modelo y que se puedan clasificar las imágenes razonablemente bien usando la nueva información.

Finalmente, con el objetivo de trabajar con las últimas tendencia y los últimos recursos disponibles públicamente, se analiza la opción de utilizar el modelo CLIP (“Contrastive Language-Image Pretraining”, detallado en el **Apartado 2**, un modelo entrenado recientemente con imágenes y texto, que ha demostrado ofrecer ciertas propiedades muy útiles para trabajar juntamente con dichos formatos.

2. Estado del arte

Uno de los factores que ha propiciado la “Revolución 4.0” es la posibilidad de relacionar datos de muchas fuentes distintas. En concreto, la capacidad de relacionar la información contenida en imágenes con la obtenida en otros formatos como pueden ser series de datos temporales, estadísticas, etiquetas, e incluso texto, audio, vídeo o cualquier otro, es de gran importancia en las nuevas aplicaciones del llamado “Internet de las Cosas” [6].

En este proyecto es necesario resolver un problema de clasificación de imágenes con caras de personas. Se podría pensar en una solución mediante redes neuronales de aprendizaje profundo [7], usando varios modelos entrenados para clasificar imágenes según diferentes características, y así poder responder a las posibles preguntas de los jugadores. Pero obtener modelos entrenados para clasificar específicamente ciertas características no resolvería el problema de permitir al usuario introducir su propio texto o pregunta, puesto que es imposible que sin saber lo que va a preguntar se puedan tener todas las posibilidades cubiertas. Al dar libertad al jugador para preguntar lo que desee se requiere un sistema “Zero-Shot” [1, 8] capaz de relacionar el texto con la imagen y responder a la pregunta.

La idea de aprovechar la información contenida en los textos, que tanto usamos para comunicarnos, y en las imágenes, que con el auge de los móviles están al alcance de cualquiera en cualquier momento, no es para nada nueva. De hecho, se pueden encontrar diferentes maneras de sacar provecho a la combinación de procesamiento de lenguaje natural con visión por computador, como por ejemplo la navegación de robots mediante instrucciones textuales [13], la obtención de descripciones de imágenes [3], utilizar texto para mejorar la clasificación de imágenes [10], etc.

Para abordar la solución al problema planteado combinando texto e imágenes, se han focalizado los experimentos a un modelo que fue presentado por Open AI este año llamado CLIP (Contrastive Language-Image Pretraining). Es una DNN (Deep Neural Network), entrenada mediante aprendizaje profundo a partir de una gran cantidad de imágenes emparejadas con texto natural, obtenidas de Internet. Se entrenó con el objetivo

de aprender a discriminar qué grupo de textos de un conjunto de miles de textos aleatorios, está relacionado con una imagen concreta. Para conseguirlo usaron un codificador de texto creado con un “Tranformer” [14], un codificador de imágenes creado con un “Vision Transformer” [4] y un conector que permite contrastar las dos codificaciones realizadas [15].

El resultado es una modelo preentrenado que permite relacionar texto e imagen de forma generalizada y que ofrece muchas posibilidades para realizar aplicaciones “Zero Shot”. Se pueden destacar algunos ejemplos donde CLIP ha obtenidos buenos resultados como en la clasificación de imágenes de arte [2] o la recuperación de textos y video [5]. También es muy interesante su contribución en DALL-E [12], una aplicación para crear imágenes a partir de una descripción textual que obtiene resultados sorprendentes.

Además, CLIP ha demostrado tener una gran eficiencia realizando clasificación de imágenes, sin necesidad de entrenamiento previo, ni siquiera un pequeño conjunto de datos para realizar “Fine-Tuning”. Se puede aplicar el modelo entrenado de CLIP directamente a un conjunto de imágenes a clasificar, codificando un texto descriptivo para cada clase. Al introducir una imagen y cierta cantidad de textos o frases a CLIP, retorna un valor para cada texto o frase introducidos, que de alguna forma es proporcional a la relación entre dicho texto y la imagen. Por ejemplo, si se trata de clasificar animales, objetos y plantas, se podrían codificar los textos “la foto de un animal”, “la foto de un objeto” y la “foto de una planta”. Con ello, al introducir una imagen de un objeto a CLIP, en general, se obtendría un valor más grande con el segundo texto. Al utilizar CLIP de esta forma, para que funcione lo mejor posible, la complejidad se traslada a la redacción de los textos a usar, siendo necesario chequear las mejores descripciones textuales para cada caso, aunque sea mediante prueba y error [11]. Esto da lugar a un nuevo concepto, el “prompt engineering”, que consiste en encontrar la mejor manera de introducir los datos a grandes modelos que se han entrenado de forma genérica, para adaptarlos a casos concretos eficazmente.

3. Desarrollo del Proyecto

A continuación, se describen los diferentes recursos y herramientas usadas en este trabajo, las normas del juego que se ha creado y los experimentos llevados a cabo.

En el **Anexo 8.1** se pueden encontrar los enlaces más interesantes en relación con este proyecto, como por ejemplo los enlaces al repositorio del juego y a la aplicación web.

En el **Anexo 8.2** se encuentra un artículo que se ha escrito durante el transcurso de este trabajo de fin de Máster titulado “An implementation of the ‘Guess who?’ game using CLIP”, que ha sido aceptado en el congreso IDEAL 2021 (22nd International Conference on Intelligent Data Engineering and Automated Learning, Manchester 25-27 November, <https://ideal-conf.com/>).

3.1. Conjunto de datos “Celeba”

Para el juego “Quien es Quien”, serán necesarias imágenes de caras de personas, y para poder realizar pruebas de las soluciones que se planteen sería interesante disponer de un gran conjunto de imágenes de caras de personas etiquetadas con características que encajen con el juego, como pueden ser el género, la franja de edad, el color del pelo, llevar gafas, sombrero u otros complementos, etc.

Gracias al MMLAB (“Multimedia Laboratory”) de la “Chinese University of Hong Kong” se dispone públicamente de un conjunto de datos [9] que cumple muy bien con las expectativas, puesto que tiene 200599 imágenes marcadas con 40 etiquetas binarias de características diferente, por ejemplo: gafas, sombrero, flequillo, pelo ondulado, nariz puntiaguda, bigote, cara ovalada, sonrisa, etc.

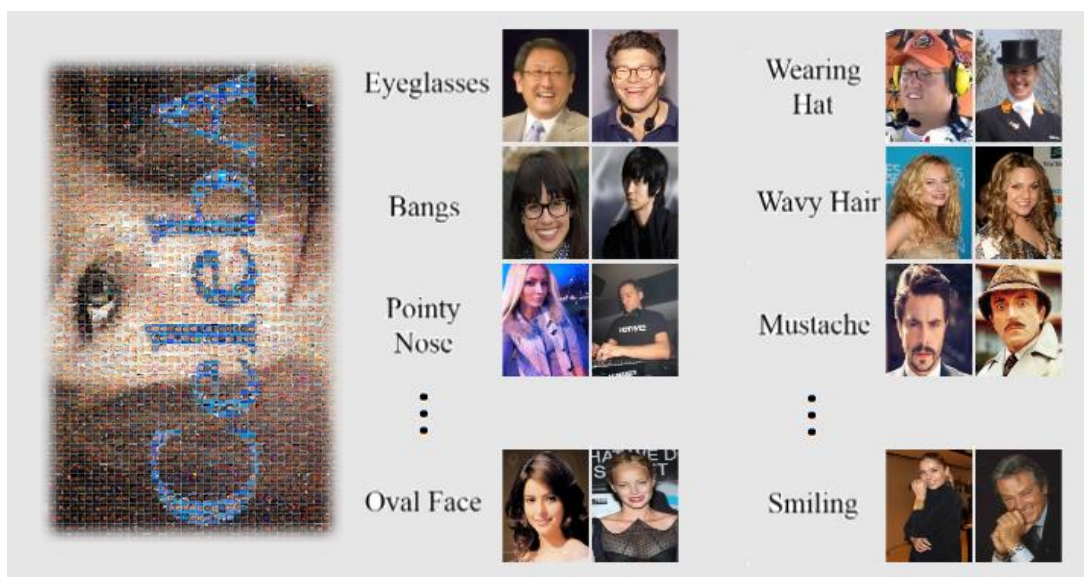


Figura 1: ejemplos de imágenes disponibles en el conjunto de datos “Celeba”

Se utilizará este conjunto de imágenes y sus etiquetas al realizar los experimentos para comprobar cuales son las mejores estrategias con CLIP como clasificador, y así obtener respuestas lo más correctas posible a las preguntas del tipo “Verdadero” o “Falso” que deben hacer los usuarios.

Se han contabilizado las etiquetas del conjunto de datos para ver cuántos datos disponibles hay de cada una.

	Etiqueta en inglés	Traducción	Casos verdaderos
1	5 o Clock Shadow	Tiene la sombra de las 5 en punto	22516
2	Arched Eyebrows	Tiene las cejas arqueadas	54090
3	Attractive	Es atractivo	103833
4	Bags Under Eyes	Tiene bolsas debajo de los ojos	41446
5	Bald	Es calvo	4547
6	Bangs	Tiene flequillo	30709
7	Big Lips	Tiene los labios grandes	48785
8	Big Nose	Tiene la nariz grande	47516
9	Black Hair	Tiene el pelo negro	48472
10	Blond Hair	Tiene el pelo rubio	29983
11	Blurry	Imagen Borrosa	10312
12	Brown Hair	Tiene el pelo castaño	41572
13	Bushy Eyebrows	Tiene las cejas espesas	28803
14	Chubby	Está regordete	11663
15	Double Chin	Tiene papada	9459
16	Eyeglasses	Lleva anteojos	13193
17	Goatee	Tiene perilla	12716
18	Gray Hair	Tiene el pelo canoso	8499
19	Heavy Makeup	Lleva maquillaje pesado	78390
20	High Cheekbones	Tiene los pómulos altos	92189
21	Male	Es un hombre	84434
22	Mouth Slightly Open	Tiene la boca ligeramente abierta	97942
23	Mustache	Tiene bigote	8417
24	Narrow Eyes	Tiene los ojos estrechos	23329
25	No Beard	No tiene barba	169158
26	Oval Face	Cara ovalada	57567
27	Pale Skin	Piel pálida	8701
28	Pointy Nose	Nariz puntiaguda	56210
29	Receding Hairline	Cada vez más calvo	16163
30	Rosy Cheeks	Mejillas sonrosadas	13315
31	Sideburns	Patillas	11449
32	Smiling	Está sonriendo	97669
33	Straight Hair	Tiene el pelo lacio o estirado	42222
34	Wavy Hair	Tiene el pelo ondulado	64744
35	Wearing Earrings	Lleva aretes	38276

36	Wearing Hat	Lleva sombrero	9818
37	Wearing Lipstick	Usa lápiz labial	95715
38	Wearing Necklace	Lleva collar de uso	24913
39	Wearing Necktie	Lleva corbata	14732
40	Young	Es joven	156734

Tabla 1: etiquetas del conjunto de datos “Celeba” con la cantidad de casos verdaderos

Se han detectado las siguientes peculiaridades al analizar el conjunto de imágenes y sus respectivas etiquetas.

Las únicas etiquetas excluyentes, es decir, que no pueden ser verdad a la vez en ningún caso, son “Bald” y “Bangs” (“Es calvo” y “Tiene flequillo”).



Es calvo y no tiene flequillo



Es calvo y no tiene flequillo



Tiene flequillo y no es calvo



Tiene flequillo y no es calvo



No es calvo y no tiene flequillo



No es calvo y no tiene flequillo

Figura 2: imágenes combinando las etiquetas “Es calvo” y “Tiene flequillo”

Si las únicas etiquetas excluyentes son las anteriores, significa que las etiquetas “Straight Hair” y “Wavy Hair” (“Tiene el pelo lacio o estirado” y “Tiene el pelo ondulado”) no son excluyentes, es decir, hay imágenes que cumplen las dos condiciones.



Pelo estirado y no ondulado



Pelo estirado y no ondulado



Pelo ondulado y no estirado



Pelo ondulado y no estirado



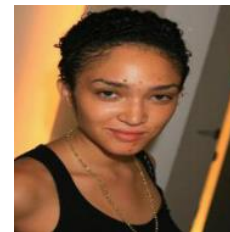
Pelo estirado y ondulado



Pelo estirado y ondulado



Pelo ni estirado ni ondulado



Pelo ni estirado ni ondulado

Figura 3: imágenes combinando las etiquetas “Straight hair” y “Wavy hair”

Lo mismo pasa con las etiquetas referentes al color del pelo y la etiqueta calvo. Es decir, hay imágenes de calvos que tienen un color de pelo asociado.



59 casos, calvo con pelo negro



1 caso, calvo con pelo rubio



3 casos, calvo con pelo castaño



1101 casos, calvo con pelo gris

Figura 4: imágenes que combinan la etiqueta “Calvo” con las de color de pelo (“Negro”, “Rubio”, “Castaño” y “Gris”)

También hay imágenes con muchas de las combinaciones de color de pelo posibles entre las 4 etiquetas que hay, negro, rubio, castaño y gris. Con lo que se pueden encontrar imágenes que combinan dichas características.



Pelo negro y rubio



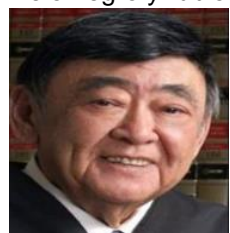
Pelo negro y rubio



Pelo negro y castaño



Pelo negro y castaño



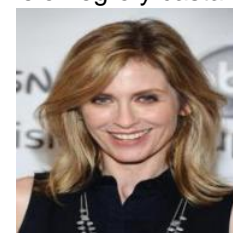
Pelo negro y gris



Pelo rubio, castaño y gris



Pelo rubio y castaño



Pelo rubio y castaño



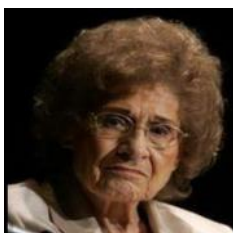
Pelo rubio y gris



Pelo rubio y gris



Pelo castaño y gris



Pelo castaño y gris

Figura 5: imágenes combinando las etiquetas “Pelo negro”, “Pelo rubio”, “Pelo castaño” y “Pelo gris”

Finalmente, se observan las etiquetas relativas a la barba, que son tres, “5 o’clock shadow” (sombra de las cinco en punto), “goatee” (perilla) y “no beard” (sin barba). Se puede encontrar algún error, como una mujer sin barba etiquetada con barba o con sombra de las cinco en punto.



Con sombra y perilla y sin barba



Con sombra y perilla y sin barba



Con barba, sin sombra ni perilla



Con barba, sin sombra ni perilla



Con sombra, perilla y
barba



Con sombra, perilla y
barba



Con sombra, sin
perilla ni barba



Con sombra, sin perilla
ni barba



Sin sombra ni barba,
pero con perilla



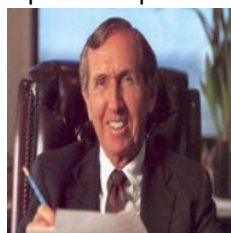
Sin sombra ni barba,
pero con perilla



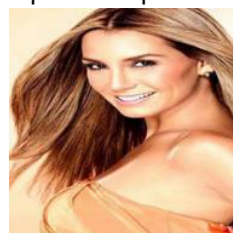
Sin sombra, pero con
perilla y barba



Sin sombra, pero con
perilla y barba



Sin sombra ni perilla
ni barba



Sin sombra ni perilla
ni barba

Figura 6: imágenes combinando las etiquetas “sombra de las 5 en punto”, “Perilla” y “Sin barba”

3.2. CLIP

3.2.1. Método de clasificación “Zero-Shot”

Como ya se ha comentado, CLIP se puede utilizar para realizar clasificación de imágenes “Zero-Shot” a partir de una lista de etiquetas en formato de texto natural (en inglés).

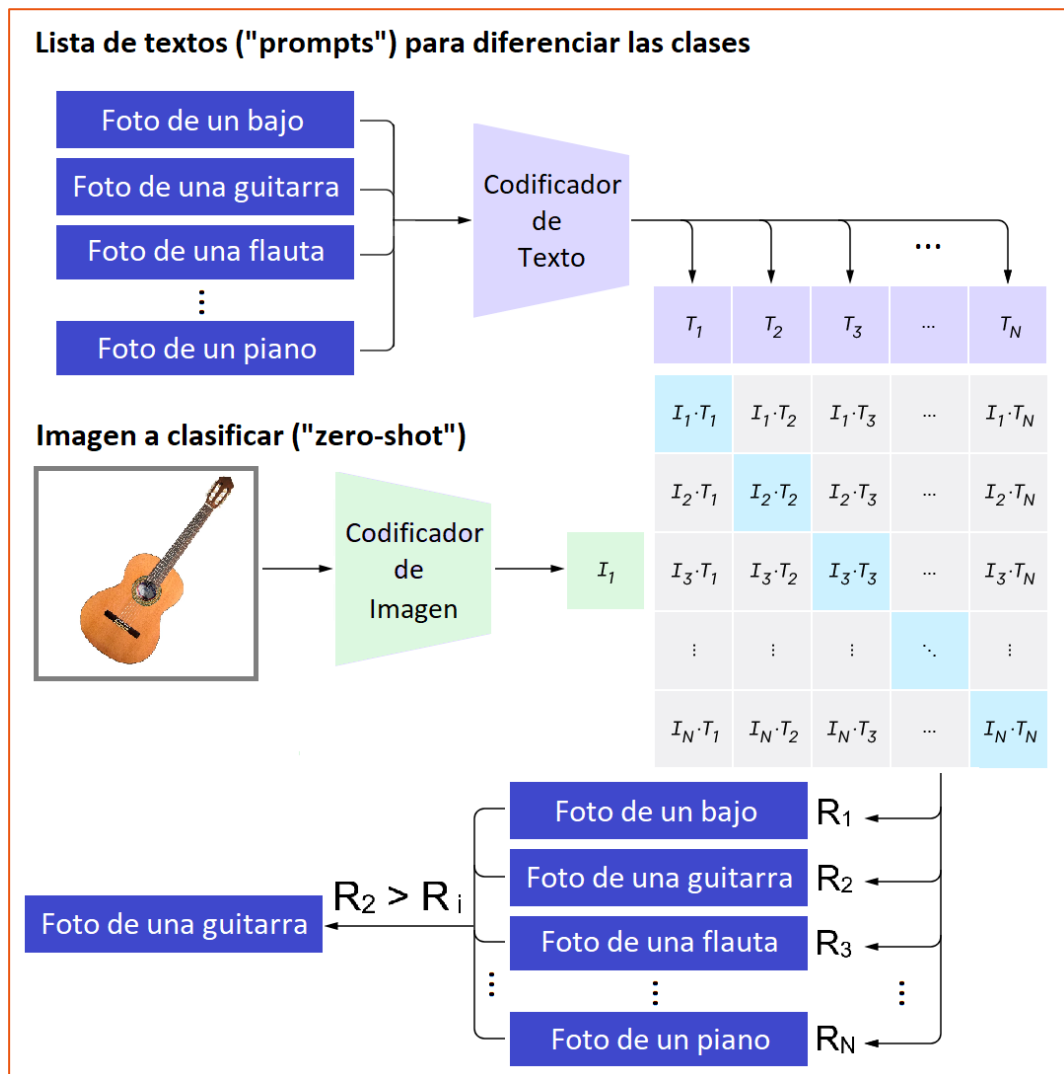


Figura 7: diagrama del uso de CLIP como clasificador “Zero-Shot”

El ejemplo que se muestra en la **Figura 7** se trataría de un clasificador de instrumentos musicales, aunque a la práctica sea necesario escribir los “prompts” en inglés, puesto que es el idioma con el que se entrenó CLIP.

Para cada una de las N clases que se desee diferenciar, se debe escribir un texto descriptivo, llamado “prompt”. Cada “prompt” se introduce en el “Codificador de Texto” para obtener 512 parámetros (T_0, T_1, \dots, T_N), obteniendo una matriz de $512 \times N$. Para clasificar una imagen en función de los “prompts” creados primero se debe introducir la imagen en el “Codificador de Imagen”, y obtener otros 512 parámetros (I_1). Finalmente se combinan los parámetros de los “prompts” con los de la imagen para conseguir N resultados (R_i), uno para cada “prompt”, cuyo valor es proporcional a la relación entre el texto y la imagen, de forma que cuanto más relación hay entre ellos, más grande es dicho valor.

3.2.2. Método para aplicar “Fine-Tuning” a CLIP

Al usar CLIP de esta forma, se aprovecha solo el proceso de codificación de CLIP que obtiene las características de cada imagen, descartando el codificador de textos. Los datos que se extraen de esta forma se obtienen en un formato 1x512, y se introducen a una pequeña red neuronal que se entrena para responder a la pregunta binaria en cuestión (en el ejemplo hombre o mujer).

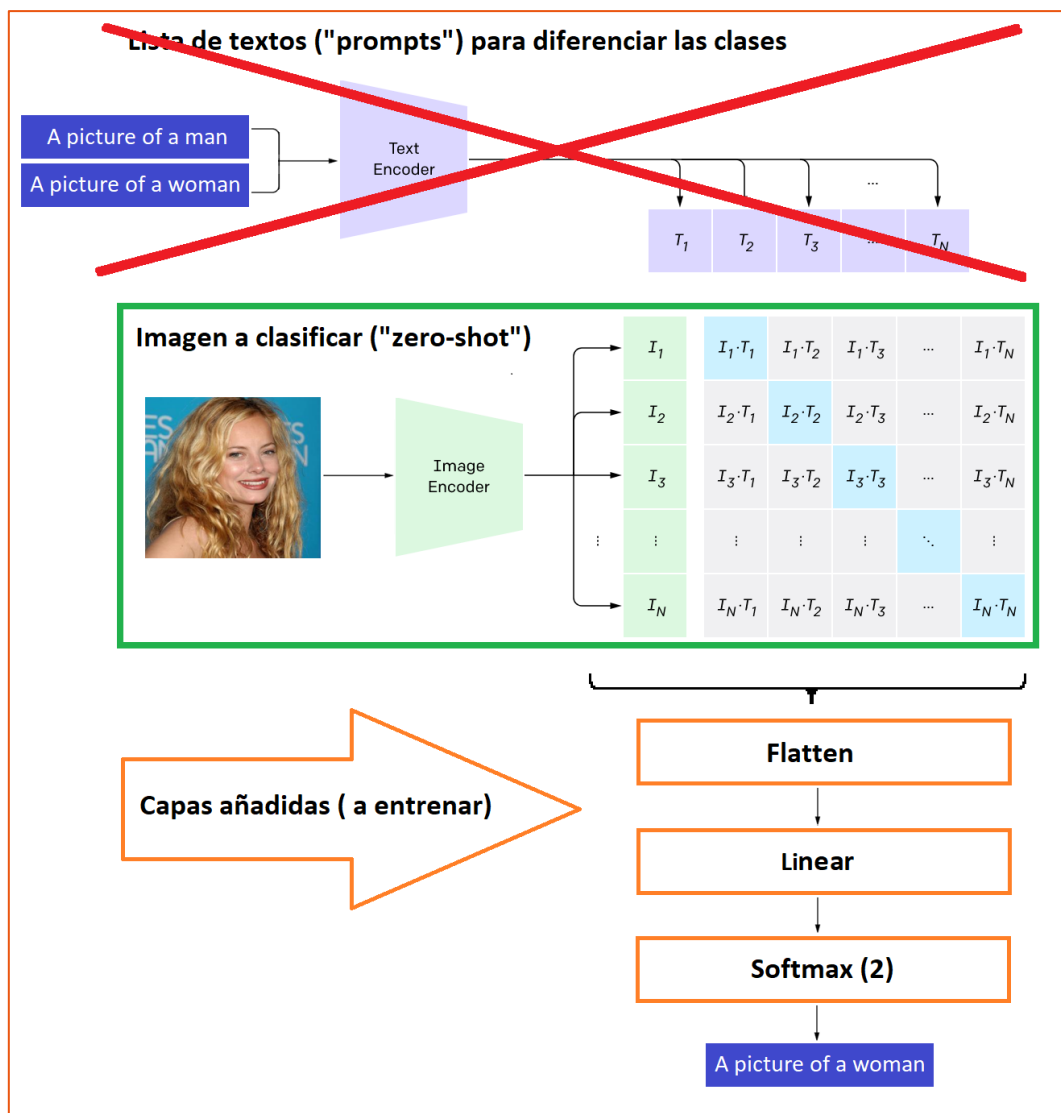


Figura 8: diagrama del uso de la codificación de imágenes de CLIP como datos de entrada para una Red Neuronal

El ejemplo que se muestra en la **Figura 8** se trataría de un clasificador de género que permitiría clasificar imágenes de hombres y mujeres, por ejemplo, para el juego “¿Quién es Quién?”.

Con este método se pretende mejorar los resultados de CLIP al clasificar una característica concreta, entrenando una capa final para reconocer los patrones de las características extraídas por el codificador de CLIP que sean más relevantes en cada clase. La idea es conseguirlo de una forma mejor o más efectiva de lo que se puede conseguir mediante la elección o creación de los “prompts” o textos descriptivos. Hay que tener en cuenta que, aunque es posible que con esta estrategia se consiga mejorar el rendimiento en comparación con la de los “prompts”, se pierde la flexibilidad que da el uso de lenguaje natural y no se puede adaptar para permitir que el usuario realice sus propias preguntas.

3.3. Streamlit

Para poder crear una aplicación a la que se pueda acceder fácilmente vía web se ha optado por la opción de “Streamlit” (<https://streamlit.io/>). Es una herramienta que consigue facilitar mucho la creación de aplicaciones web para mostrar públicamente los resultados de un programa realizado en Python. Esta API permite crear y lanzar aplicaciones de una forma rápida, sencilla y gratuita, sin tener mucha experiencia en “Front-End”. En poco tiempo se puede aprender a utilizar si se dispone de conocimientos básicos de Python puesto que no hace falta saber ningún otro tipo de lenguaje de programación como HTML, CSS, JavaScript, etc.

Después de realizar la aplicación se ha detectado que debido a la limitación de los recursos que ofrece “Streamlit”, el consumo de memoria RAM durante el proceso realizado con el modelo CLIP puede llegar a ser excesivo y producir un fallo de la aplicación web. No se han hecho pruebas de despliegue de la aplicación en una máquina con GPU, cosa que podría aumentar el rendimiento en la inferencia.

El código del juego, basado en “Streamlit”, ha sido liberado en Github, <https://github.com/ArnauDIMAI/CLIP-GuessWho>.

3.4. Descripción del juego

La versión original del juego “¿Quién es Quién?” es para dos jugadores. Se dispone de un panel con distintas imágenes de caras de personas y consiste básicamente en averiguar la imagen seleccionada por el jugador adversario. Para ello, se pueden realizar preguntas con respuesta del tipo “Verdadero” o “Falso” para descartar imágenes. Al empezar el juego, los jugadores eligen una imagen de todas las posibles, que el otro jugador no debe conocer, y seguidamente se van alternando un turno cada uno. En cada turno los jugadores pueden decidir entre realizar una pregunta o intentar descubrir la imagen del otro, pero solo una de las dos cosas. Si un jugador intenta descubrir la imagen del otro y acierta, gana la partida, pero si falla, la pierde.



Figura 9: imágenes del juego real de “¿Quién es Quién?”

En la versión propuesta para un solo jugador el objetivo es descubrir la imagen ganadora lo más rápido posible. En cada turno el jugador puede realizar una pregunta, que le permite descartar imágenes. Se consigue la máxima puntuación acertando la imagen ganadora en el primer turno, es decir, descartando el resto de las imágenes con una sola pregunta. Con cada nuevo turno se permite una nueva pregunta, pero también se produce una penalización.

Uno de los objetivos planteados es permitir a los usuarios jugar con cualquier conjunto de imágenes de caras de personas, pero esto puede suponer un problema en el momento del descarte de imágenes. En el juego original es importante que los dos jugadores respondan lo mismo al realizar las preguntas sobre las características de las imágenes, para evitar que un jugador descarte una imagen que no es, al interpretar de diferente manera una

descripción. De hecho, en el juego de tablero, el conjunto de imágenes de caras de personas son dibujos que presentan unas características muy concretas, de forma que no pueden conducir a confusión al realizar preguntas del tipo “Sí” o “No” entre jugadores. Es decir, al realizar una pregunta sobre el pelo, como puede ser “¿Tienes el pelo rubio?”, no se puede producir ningún error en el descarte porque todas las imágenes con el pelo rubio tienen exactamente el mismo color, así como todas las de pelo negro, blanco, castaño o peli-rojo. Pero al usar imágenes reales es más fácil que se pueda producir alguna confusión, como por ejemplo que un jugador responda que su imagen tiene el pelo rubio, pero el otro jugador la descarte por considerar que es castaño claro. Para solucionar este problema se decide realizar el descarte de imágenes de forma automática por el juego, al seguir siempre el mismo criterio no se pueden producir confusiones y simplemente se descartan siempre las imágenes que no coinciden con la ganadora. Para hacerlo se utiliza CLIP (**Apartado 3.2.1**), al realizar una pregunta descriptiva sobre las caras de las imágenes CLIP determina cuales son las que responden “Verdadero”, con lo que siguen en el juego, y cuales las que responden “Falso”, con lo que son descartadas. A continuación, se muestran un esquema de la propuesta descrita y un ejemplo de una clasificación de imágenes realizada por CLIP remarcando en rojo las descartadas.

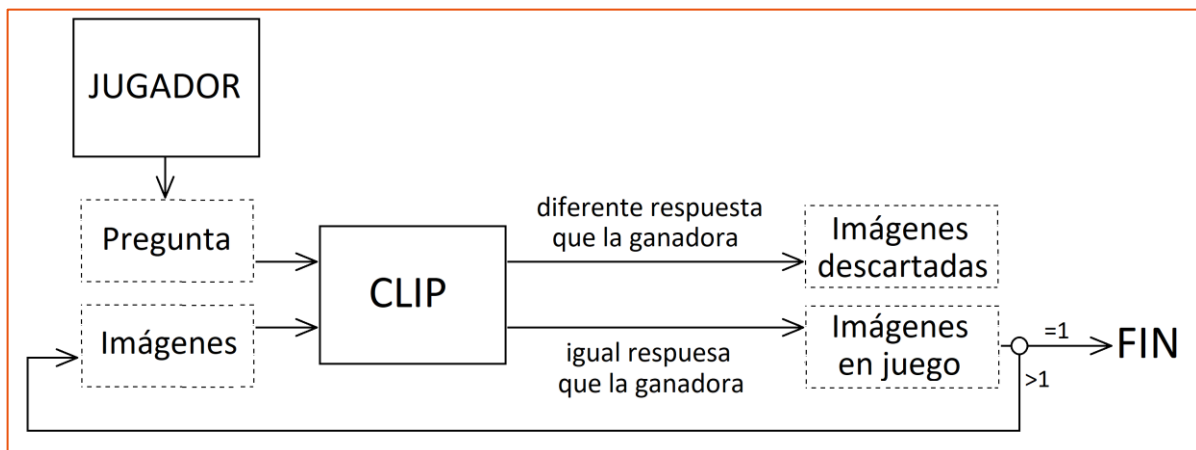


Figura 10: esquema del funcionamiento de CLIP en el juego

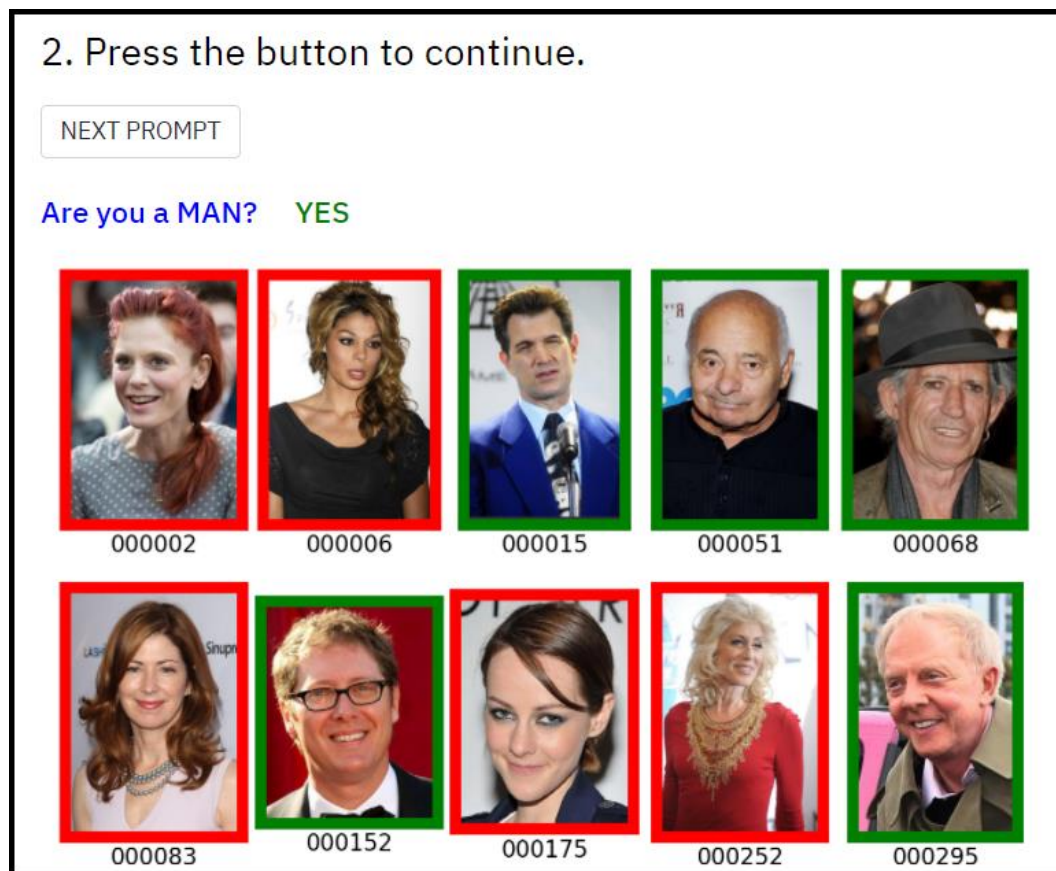


Figura 11: ejemplo de la respuesta de CLIP a la pregunta “¿Eres un hombre?”

3.4.1. ¿Cómo jugar?

Se ha creado una versión para un solo usuario del juego “Quién es Quién” y a continuación se describe como se lleva a cabo el desarrollo del juego.

Antes de empezar se permite al usuario cambiar las imágenes con las que jugar. Puede elegir usar sus propias imágenes o las del conjunto de datos “Celeba” (**Figura 12**).

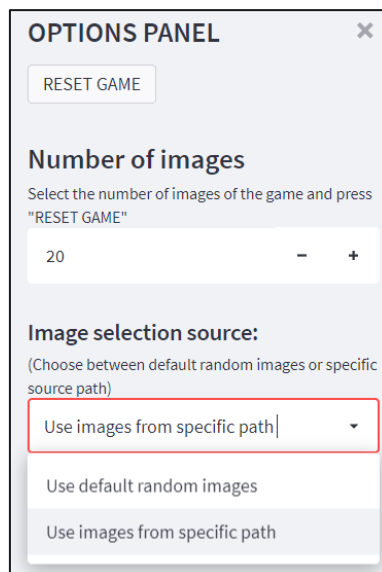


Figura 12: panel de opciones del juego

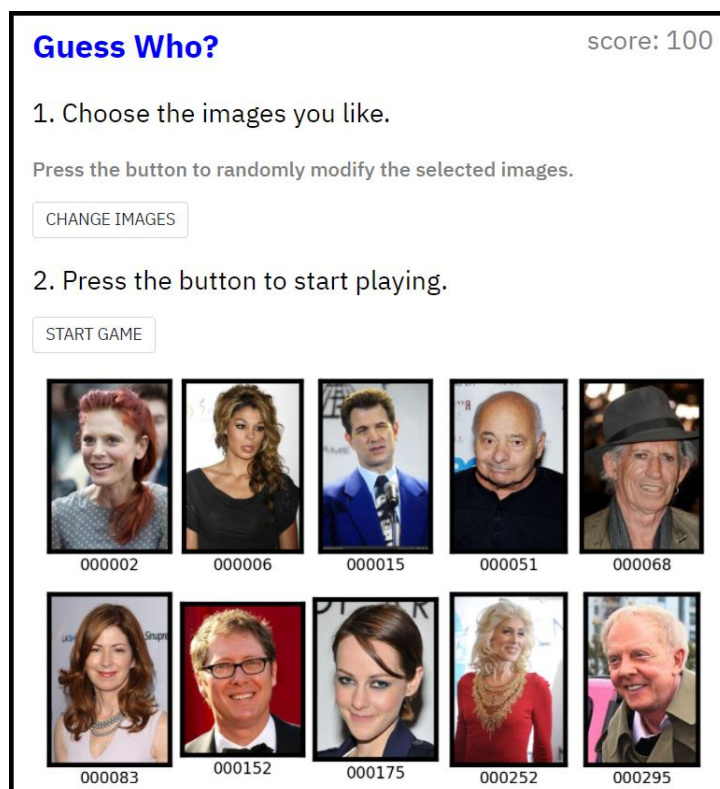


Figura 13: primera pantalla del juego, para seleccionar las imágenes e iniciar el juego

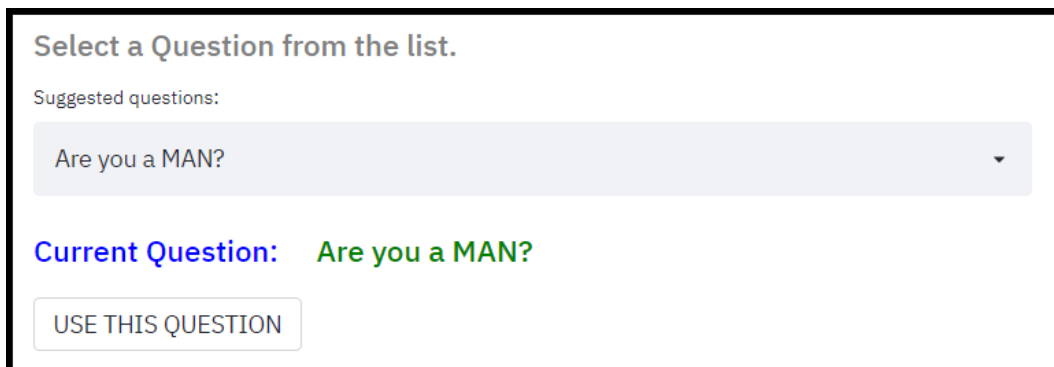
Al principio aparecen un conjunto de imágenes seleccionadas aleatoriamente, pero el jugador puede realizar una nueva selección aleatoria de imágenes pulsando un botón si no le gustan las que aparecen, tantas veces como desee (**Figura 13**).

Una vez se dispone de las imágenes con las que jugar, se debe elegir la modalidad de pregunta a realizar de 4 opciones posibles.

1. Seleccionar una pregunta de una lista.

Con esta opción el usuario puede elegir una pregunta de una lista desplegable. En este caso se utilizan los “prompts” para configurar CLIP que han resultado más eficaces durante las pruebas y experimentos realizados en el desarrollo del proyecto, para cada una de las preguntas disponibles (**Apartado 4**). Estas preguntas están basadas en las descripciones que se encuentran etiquetadas del conjunto de datos “Celeba” (**Tabla 1**).

Por ejemplo, en el caso de la pregunta “¿Eres un hombre?”, se pueden usar dos “prompts” como: “A picture of a man” vs “A picture of a woman” para configurar CLIP.



Select a Question from the list.

Suggested questions:

Are you a MAN?

Current Question: Are you a MAN?

USE THIS QUESTION

Figura 14: pantalla de selección de preguntas predefinidas

2. Crear una frase descriptiva en inglés

Esta opción permite al usuario introducir su propia descripción de una característica, en inglés. En este modo la estrategia que se sigue es utilizar dos “prompts” para configurar CLIP. El primero es una frase neutral o de referencia, “A picture of a person”, que se confronta con la descripción o el texto introducido por el usuario. Se recomienda al jugador usar una frase descriptiva parecida a la de referencia, como pueden ser “A picture of a (description)

person", o "A picture of a person with... (objeto, o característica)". De esta forma las imágenes que tienen relación con la descripción añadida por el usuario suelen obtener más peso con el segundo "prompt" que con el primero, el de referencia. Por el contrario, si el texto introducido por el usuario no aporta nada con relación a la imagen, el segundo "prompt" que incluye dicho texto innecesario o inútil, obtiene menos peso que el "prompt" con la frase neutral. Con este método, si el texto introducido es muy distinto al de referencia, se puede obtener como resultado que todas las imágenes pertenecen al mismo grupo, con lo que no se producen descartes.



Write your own prompt and press the button.

It is recommended to use a text like: "A picture of a ... person" or "A picture of a person ..." (CLIP will check -> "Your prompt" vs "A picture of a person")

A picture of a man

USE MY PROMPT: A picture of a man

Figura 15: pantalla de introducción de un "prompt" creado por el jugador

3. Crear dos frases descriptivas opuestas

Esta opción permite al usuario introducir dos frases descriptivas para diferenciar "Verdadero" y "Falso". En este caso se sigue una estrategia parecida a la anterior, también se usan dos "prompts" para configurar CLIP, pero los dos son introducidos por el usuario. La única restricción es que deben ser distintos. Se recomienda al jugador usar dos frases parecidas pero opuestas, como las que aparecen de ejemplo "A picture of a man" vs "A picture of a woman". De esta forma se puede amplificar el efecto que tienen los dos textos introducidos en los pesos que obtienen las imágenes al configurar CLIP con ellos, ya que, si las descripciones son opuestas, las imágenes que no tengan ninguna relación con uno de los textos es probable que si tengan relación con el otro texto, al ser un concepto opuesto. Igual que antes, si los textos introducidos son muy distintos, se puede obtener como resultado que todas las imágenes pertenecen al mismo grupo y no hay descartes.

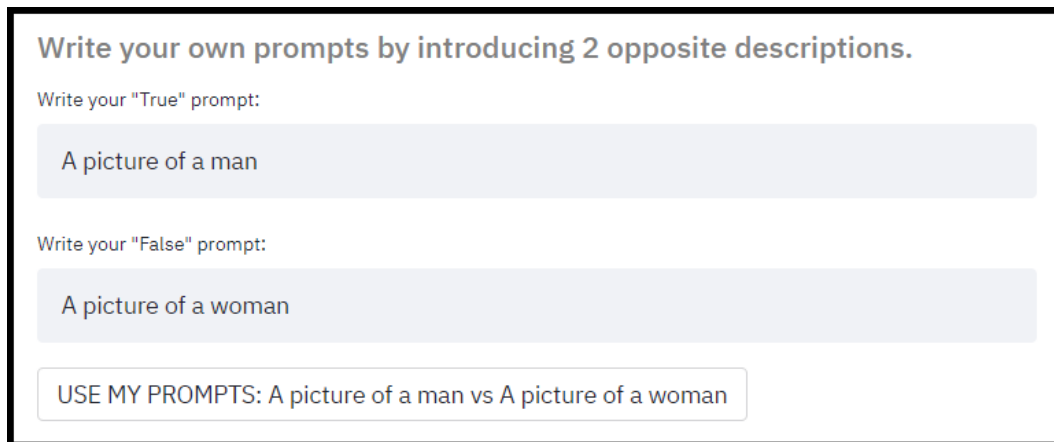


Figura 16: pantalla de introducción de dos “prompts” creados por el jugador

4. Elegir una imagen como ganadora

Finalmente, también se permite al usuario seleccionar directamente una imagen concreta como la imagen ganadora, siguiendo su propio instinto o probando suerte.

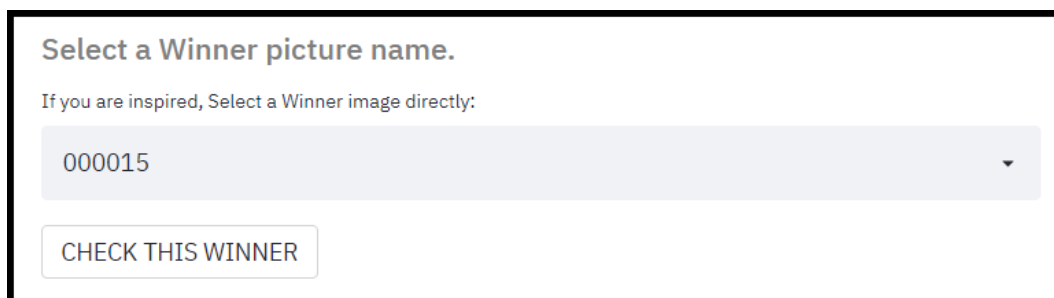


Figura 17: pantalla de selección de imagen ganadora según su identificador

3.4.2. Puntuación

Para motivar a los jugadores a encontrar la imagen ganadora con el mínimo número de turnos posibles, se establece un sistema de puntuación que empieza con un valor máximo de 100 y va disminuyendo a medida que se van haciendo preguntas. El número de puntos perdidos en cada turno es el número de imágenes que no se descartan y permanecen en el juego, sin tener en cuenta la ganadora.

En caso de seleccionar la cuarta opción, que permite elegir directamente una imagen ganadora, se realiza una penalización extra, puesto que con este modo siempre se descarta al menos una imagen y permite conseguir acertar a la primera con una probabilidad proporcional al número de imágenes total con el que se juega. Es decir, si por ejemplo se juega con 20 imágenes, con este método hay una probabilidad del 5% de dar con la ganadora sin introducir descripciones ni hacer preguntas.

También se realiza una penalización extra de varios puntos si al consumir un turno no se consigue realizar ningún descarte.

3.5. Experimentos

Los experimentos realizados se basan en utilizar CLIP con la metodología descrita en el **Apartado 3.2.1**. El reto que propone esta metodología es conseguir los “prompts” que permitan diferenciar las clases o características concretas que se requiera en cada caso. Para comprobar la eficacia de CLIP se aprovecha el conjunto de datos “Celeba” con todas sus etiquetas.

El juego “Quien es Quien” requiere, básicamente, realizar preguntas binarias, es decir, clasificar imágenes en dos grupos. La primera estrategia propuesta para realizar la clasificación de imágenes en dos opciones utiliza dos textos o “prompts” y consiste en contraponer una frase descriptiva contra una frase neutral. Es decir, para cada una de las etiquetas binarias se utilizan dos textos, el primero es el mismo para todas, el llamado “prompt” de referencia, el segundo es un texto parecido al primero pero que incluye la etiqueta del conjunto de datos “Celeba”. La frase de referencia debe ser neutral, y después de un experimento para probar varias opciones como “A image of a ...”, “A photo of a ...”, etc, se elige “A picture of a person. La frase descriptiva se construye modificando ligeramente la frase de referencia, usando la etiqueta “Celeba”. Por ejemplo, la etiqueta que permite diferenciar imágenes de hombres y de mujeres tiene asociada la palabra “male”, con lo que la segunda frase para el caso del género puede ser “A picture of a male person”. Sencillamente se añade la palabra “male” delante de “person”. De esta forma se espera que en las imágenes donde aparezca un hombre el segundo “prompt” obtenga más valor que el primero al existir cierta relación entre la imagen y la palabra extra “male”. Por lo contrario, en las imágenes de mujeres se espera que esta palabra no contribuya a dar más valor al “prompt” sino que lo perjudique, por no tener relación.

En una segunda estrategia, muy parecida a la anterior, se prueban algunas modificaciones de las frases propuestas buscando un sentido inverso de la descripción, para tratar de encontrar los casos que no cumplen la etiqueta mediante el “prompt” descriptivo, en vez de con la referencia. Por ejemplo, para el caso del género, donde se usa la palabra “male” en el conjunto de datos “Celeba”, se puede cambiar por la palabra “female” dando un sentido opuesto a la frase.

La tercera estrategia consiste en una mezcla de las anteriores, eliminando la frase neutral o de referencia para cambiar-la por la que tiene un sentido opuesto a la frase descriptiva. De esta forma cada una de las frase puede contribuir a dar más o más peso al “prompt” según si la imagen coincide o no con las descripciones puesto que las dos frases describen la etiqueta en un sentido u otro. Por ejemplo, con la etiqueta de género, “male”, se obtiene muy buen resultado utilizando esta estrategia con las frases “A picture of a male person” y “A picture of a female person”.

En una cuarta estrategia, se mejoran los resultados de algunas etiquetas añadiendo más “prompts” a la lista. En este caso se pueden añadir textos para discriminar más de una característica o para tener en cuenta varias opciones. Esta estrategia es útil en casos donde, aunque la pregunta sea binaria, la característica a tener en cuenta no lo es, como el color del pelo, que realmente no tiene solo dos clases, sino que hay muchas más. En este caso, por ejemplo, está claro que el pelo puede ser rubio, moreno, marrón, blanco, etc, e incluso uno puede ser calvo. Con lo que tener todas los “prompts” codificados, permite gestionar los casos intermedios, como el pelo marrón que en cierto modo se puede considerar entre negro y rubio. Para el caso del color del pelo se realiza un experimento extra descartando las imágenes de “Celeba” etiquetadas con dos colores de pelo distintos porque esto no encaja con las normas del juego.

Finalmente se realizan algunos experimentos con la metodología descrita en el **Apartado 3.2.2**, en el que se trata de mejorar los clasificadores obtenidos hasta el momento realizando un “Fine-Tuning” de CLIP y también de un par de redes neuronales, VGG19 y Xception, para comparar resultados. Se elige VGG19 porque obtuvo buenos resultados en clasificación de caras en el estudio “Face recognition using popular deep net architectures: a brief comparative study” [7], y Xception porque obtuvo muy buenos resultados en una de las actividades realizadas durante este Máster sobre clasificación de imágenes.

Para adaptar VGG19, Xception y CLIP se añade una capa final “Linear + Softmax” binaria que se reentrena para responder a las preguntas del juego. En ambos casos se reduce el tamaño de las imágenes entradas a 112x112 píxeles para facilitar el cómputo y que los entrenamientos tengan una duración aceptable. Además, en los casos de VGG19 y Xception, se usa el preprocesado de imágenes original de cada una de las redes, puesto que

en las actividades realizadas durante el Máster se comprobó que se conseguían mejores resultados de esta forma. En el **Apartado 3.2.2** se muestra con un esquema como se realiza esta adaptación en CLIP.

4. Resultados

A continuación, se muestran los resultados de los distintos experimentos realizados con CLIP durante el desarrollo del proyecto. Aunque se dispone de una gran cantidad de imágenes, hay que tener en cuenta que realizar las pruebas requiere tiempo y recursos, la estrategia usada para realizar experimentos de manera eficiente consiste en elegir siempre los 4000 primeros casos de cada etiqueta y para cada respuesta posible. De esta forma, para cada una de las 40 etiquetas disponibles, siempre se tienen en cuenta 8000 ejemplos, 4000 verdaderos y 4000 falsos. En las tablas de resultados se muestra la Precisión Total y los Ratios de Verdaderos Positivos (RVP) y Negativos (RVN) obtenidos en los experimentos, entendiendo como casos positivos los que responden si a la pregunta, es decir, se considera cierta o se cumple la etiqueta “Celeba”, y negativos los que responden no o no cumplen con la etiqueta.

Precisión Total = $100 * \text{Aciertos Totales} / \text{Total Casos}$

Ratio Positivos Acertados: $\text{RPA} = \text{Verdaderos Acertados} / \text{Total Casos Verdaderos}$

Ratio Negativos Acertados: $\text{RNA} = \text{Falsos Acertados} / \text{Total Casos Falsos}$

Ratio Positivos Falsos: $\text{RPF} = 1 - \text{RPA}$

Ratio Negativos Falsos: $\text{RNF} = 1 - \text{RNA}$

Ratio Verdaderos Positivos: **RVP** = $100 * \text{RPA} / (\text{RPA} + \text{RNF})$

Ratio Verdaderos Negativos: **RVN** = $100 * \text{RNA} / (\text{RNA} + \text{RPF})$

En algunas tablas se muestra la **Mejora** conseguida respecto a los resultados del primer experimento (4.1). Hay que tener en cuenta que el peor resultado posible es una Precisión Total del 50%, lo que significa que la respuesta es aleatoria. Si se obtienen resultados de menos del 50%, 15% por ejemplo, aunque se responde prácticamente siempre erróneamente, no es un mal resultado. Es sencillo dar la vuelta al resultado, invirtiendo siempre la respuesta, y consiguiendo un 85% de aciertos. En estos casos, en las tablas se muestran los valores obtenidos realmente para mostrar que la frase actúa de forma opuesta a lo previsto, pero si se calcula la mejora obtenida respecto a otro método, se hace referencia al mejor caso, es decir, a invertir el resultado de la pregunta.

4.1. Frase neutral vs Descripción

En estos experimentos se utilizan dos “prompts” para determinar cada una de las preguntas de respuesta binaria que permiten hacer las etiquetas del conjunto de datos “Celeba”. La idea es elegir un “prompt” de referencia y otro descriptivo, de forma que este “prompt” descriptivo podría ser escrito por el usuario y el de referencia siempre sería el mismo. Con este método es relativamente sencillo que un usuario pueda introducir la descripción que quiera, sin necesidad que escriba las dos frases necesarias para realizar la clasificación binaria con CLIP.

En un primer experimento, para elegir la frase neutral o de referencia, se realizan pruebas con una de las etiquetas mas importantes para el juego, la etiqueta “Hombre / Mujer”, puesto que permite realizar una de las preguntas más comunes, “Eres un hombre?” o “Eres una mujer?”. Se chequean varias formas de crear los “prompts” como por ejemplo modificar la estructura de la frase, usar sinónimos para referirse a la imagen o usar solo palabras sin conectores. En el conjunto de datos “Celeba” usan la palabra “male” para referirse al género masculino, así que se usa esta palabra, junto a la palabra “female”, para crear los dos “prompts”. Se introducen diferentes parejas de “prompts” a CLIP para ver cuál es la que obtiene mejores resultados.

"Prompts" de referencia	"Prompts" descriptivos	RVP	RVN	PT
A picture of a person	A picture of a female person	98,96	97,33	98,12
A photo of a person	A photo of a female person	97,53	97,81	97,68
A image of a person	A image of a female person	97,32	97,99	97,65
A picture of a person	A picture of a male person	98,14	96,18	97,14
A snapshot of a person	A snapshot of a female person	99,02	93,38	96,02
A image of a person	A image of a male person	98,9	92,39	95,41
A photo of a person	A photo of a male person	98,97	92,2	95,34
A snapshot of a person	A snapshot of a male person	91,91	98,49	94,96
A image of a person	A image of a person who is a female	95,91	88,36	91,8
A image of a person	A image of a person who is female	95,2	88,18	91,39
A picture of a person	A picture of a person who is a female	89,47	90,62	90,04
A picture of a person	A picture of a person who is female	89,27	90,53	89,89
A photo of a person	A photo of a person who is a female	85,81	89,44	87,54
A photo of a person	A photo of a person who is female	85,14	90,16	87,49
A person	A Female	75,66	96,65	83,11

A snapshot of a person	A snapshot of a person who is female	95,96	71,58	79,38
A person	A Male	95,25	66,85	74,56
A snapshot of a person	A snapshot of a person who is a female	94,67	66,55	74,15
Person	Male	93,88	60,7	67,2
A image of a person	A image of a person who is a male	56,96	71,11	60,46
A snapshot of a person	A snapshot of a person who is a male	54,99	92,73	58,92
A photo of a person	A photo of a person who is a male	56,31	59,55	57,6
A snapshot of a person	A snapshot of a person who is male	54,11	89,42	57,45
A picture of a person	A picture of a person who is a male	55,92	59,54	57,3
A image of a person	A image of a person who is male	54,25	60,93	56,12
A picture of a person	A picture of a person who is male	54,11	56,15	54,92
Person	Female	52,59	93,91	54,9
A photo of a person	A photo of a person who is male	54,46	55,22	54,81

Tabla 2: resultados de la clasificación por género (hombre o mujer) con las palabras “male” y “female” y diferentes estructuras textuales para buscar una buena “Frase neutral”

En esta tabla se muestran los dos “prompts” chequeados para separar la clase género del conjunto de datos “Celeba” ordenados según la precisión total obtenida. En las primeras columnas , el de referencia y el descriptivo. También se muestran los resultados obtenidos de RVP, RVN y TP.

Al usar las etiquetas “male/female” en diversas variantes de “prompts”, se obtienen algunos muy buenos resultados por encima del 90%, incluso llegando al 98% con los prompts “A picture of a person / A picture of a female person”. Para las dos palabras clave, “male” como “female”, los mejores resultados son con la frase “A picture of a ... person”. Quizás, a parte de lo buenos que son algunos de los resultados con diferentes sinónimos para la palabra “imagen”, usando estructuras textuales diferentes e incluso quitándolas (“A person / A female” obtiene un 83%), lo más destacable es el caso de usar los prompts hechos con las etiquetas sueltas “Person / Male” o “Person / Female”. Sobre todo, si se compara su resultado con el obtenido simplemente añadiendo “A” a las etiquetas y usar “A person / A male” o “A person / A female”, con más o menos una mejora del 7 y del 28% de la precisión total respectivamente.

En un segundo experimento se usa la estructura con mejores resultados en el experimento anterior para seguir la estrategia de usar un “prompt” neutral y otro descriptivo. Por lo tanto, uno de los “prompts” siempre es el mismo y la frase de referencia elegida es “A picture of a person”. El otro “prompt” se crea usando la etiqueta “Celeba” para modificar ligeramente la frase anterior. Se han utilizado las siguientes variantes:

- añadir la etiqueta “Celeba” en algún lugar de la frase de referencia (por ejemplo, delante de “person” si es un adjetivo)
- añadir la palabra “with ...” junto a la etiqueta “Celeba” (por ejemplo, después de “person” si es un objeto)
- en un caso (etiqueta “smiling”) añadir la frase “who is ...” en vez de “with ...”.

La estrategia consiste en introducir a CLIP dos “prompts”, uno de referencia que siempre es el mismo y otro descriptivo creado a partir de modificar la referencia con la etiqueta “Celeba”. Se clasifican todas las etiquetas binarias del conjunto de datos “Celeba” mediante esta estrategia para observar los resultados. La frase neutral usada es “A picture of a person”, que obtuvo los mejores resultados en el primer experimento.

Etiquetas “Celeba”	“Prompts” descriptivos	RVP	RVN	PT
male	A picture of a male person	98,14	96,13	97,11
bald	A picture of a bald person	96,63	73,86	81,58
young	A picture of a young person	63,01	62,63	62,81
chubby	A picture of a chubby person	66,01	53,36	55,56
blurry	A blurry picture of a person	82,48	52,82	55,19
attractive	A picture of an attractive person	54,34	50,88	51,46
wearing hat	A picture of a person with hat	97,29	83,67	89,34
goatee	A picture of a person with goatee	91,6	77,05	82,78
blond hair	A picture of a person with blond hair	74,14	97,84	82,09
bangs	A picture of a person with bangs	88	77,71	82,05
eyeglasses	A picture of a person with eyeglasses	87,45	77,59	81,78
wearing necktie	A picture of a person with necktie	77,91	81,36	79,54
gray hair	A picture of a person with gray hair	83,66	74,04	78,05
black hair	A picture of a person with black hair	69,6	81,88	74,28
wavy hair	A picture of a person with wavy hair	72,71	73,94	73,31
pale skin	A picture of a person with pale skin	68,85	67,12	67,94
mustache	A picture of a person with mustache	91,78	61,38	67,89

receding hairline	A picture of a person with receding hairline	72,22	64,85	67,8
wearing necklace	A picture of a person with necklace	71,27	61,44	64,88
sideburns	A picture of a person with sideburns	67,08	62,46	64,41
wearing earrings	A picture of a person with earrings	75,03	60,05	64,34
bushy eyebrows	A picture of a person with bushy eyebrows	72,99	60,27	64,2
oval face	A picture of a person with oval face	59,92	64,57	61,8
arched eyebrows	A picture of a person with arched eyebrows	61,12	59,64	60,32
brown hair	A picture of a person with brown hair	55,56	85,63	59,62
5 o'clock shadow	A picture of a person with 5 o'clock shadow	67,23	54,51	57,15
mouth slightly open	A picture of a person with mouth slightly open	68,85	54,1	56,72
straight hair	A picture of a person with straight hair	56,27	55,64	55,94
heavy makeup	A picture of a person with heavy makeup	86,47	53,07	55,66
double chin	A picture of a person with double chin	65,36	52,71	54,61
big lips	A picture of a person with big lips	64,64	51,75	53,12
wearing lipstick	A picture of a person with lipstick	85,34	51,53	52,94
pointy nose	A picture of a person with pointy nose	52,74	52,16	52,41
big nose	A picture of a person with big nose	57,63	51,09	51,91
rosy cheeks	A picture of a person with rosy cheeks	49,36	49,76	49,65
high cheekbones	A picture of a person with high cheekbones	47,33	49,23	48,8
bags under eyes	A picture of a person with bags under eyes	47,7	49,05	48,66
narrow eyes	A picture of a person with narrow eyes	40,67	45,36	43,8
no beard	A picture of a person with no beard	16,93	30,02	25,09
Smiling	A picture of a person who is smiling	89,07	76,75	81,76

Tabla 3: resultados del experimento “Frase neutral vs Descripción Celeba” usando cómo “prompt” de referencia “A picture of a person”

En esta tabla se muestra la aplicación del método “Frase neutral vs Descripción” con todas las etiquetas del conjunto de datos “Celeba”. En la primera columna encontramos la etiqueta, en la segunda el “prompt” descriptivo creado al modificar la referencia con la etiqueta, y en las demás los resultados obtenidos de RVP, RVN y PT.

En estos experimentos es importante tener en cuenta que se trata de clasificación “Zero Shot”, sin tener mucha experiencia aún en cómo introducir los “prompts” correctamente para aprovechar el potencial de CLIP. Aun así, se obtienen algunos resultados muy buenos, a parte del que ya se ha visto en el primer experimento con la etiqueta “male” (con un 97% de precisión total), la etiqueta “Wearing Hat” consigue más de un 89% usando el conector “with”. Hay varias etiquetas por encima del 80%, “bald”, “goatee”, “blond hair”, “bangs”, “eyeglasses” y “smiling”, que no está nada mal para un sistema “Zero-Shot”. Cerca de estos

casos se encuentran “wearing necktie”, “gray hair”, “black hair”, “wavy hair”, “eyeglasses” y “No beard”, esto significa un total de 13 etiquetas por encima del 70%, lo que es casi la tercera parte de la etiquetas disponibles en “Celeba” (40 en total). Por debajo del 60% de precisión, que serían muy malos resultados, se encuentran 14 etiquetas, lo que representa un 35% del total. Se puede remarcar un caso curioso, el de la etiqueta “No beard”, en la que se obtiene un 25% de precisión. Como se ha comentado este resultado es equivalente a un 75%, si se invierte la respuesta.

4.2. Frase neutral vs Descripción opuesta

En este experimento se modifican los “prompts” descriptivos para tratar de detectar las imágenes que no cumplen con la condición de la etiqueta, en vez de las que si la cumplen. Se sigue usando la misma referencia “A picture of a person”, y se modifica la frase descriptiva para invertir su sentido o significado. Por ejemplo, en el caso de la etiqueta “Joven” se podría usar la descripción “Anciano” para tratar de describir lo contrario.

Etiquetas “Celeba”	"Prompts" descriptivos modificados	RVP	RVN	PT	Mejora
5 o'Clock Shadow	A picture of a person without 5 o'clock shadow	65,58	53,45	55,71	-1,44
Arched Eyebrows	A picture of a person with straight eyebrows	40,36	38,88	39,67	+0,01
Attractive	A picture of an unattractive person	47,18	43,23	46,02	+1,04
Bags Under Eyes	A picture of a person without bags under eyes	50,52	51,47	50,78	-0,56
Bald	A picture of a haired person	67,85	75,11	70,88	-10,7
Bangs	A picture of a person without bangs	63,75	71,42	66,75	-15,3
Big Lips	A picture of a person with small lips	49,36	47,44	48,98	-0,44
Big Nose	A picture of a person with small nose	48,18	45,06	47,35	+0,74
Blurry	A clear picture of a person	47,2	40,79	45,7	-0,89
Bushy Eyebrows	A picture of a person with wispy eyebrows	46,55	44,74	45,82	-10,02
Chubby	A picture of a slender person	51,39	81,37	52,66	-2,9
Double Chin	A picture of a person without double chin	47,94	36,44	46,42	-1,03
Eyeglasses	A picture of a person without eyeglasses	27,58	30,48	29,12	-10,9
Goatee	A picture of a person without goatee	14,99	23,54	19,86	-2,64
Heavy Makeup	A picture of a person without heavy makeup	40,85	25,73	36,71	+7,63
High Cheekbones	A picture of a person without high cheekbones	50,9	52,96	51,38	+0,18
High Cheekbones	A picture of a person with low cheekbones	51,11	50,6	50,78	-0,42
Male	A picture of a female person	76,04	98,36	83,85	-13,26
Mustache	A picture of a person without mustache	23,58	34,38	30,38	+1,73
Narrow Eyes	A picture of a person with separated eyes	52,2	58,89	53,52	-2,68
No Beard	A picture of a person with beard	55,21	98,96	59,41	-15,5
Pale Skin	A picture of a person with tanned skin	60,02	72,61	63,89	-4,05
Pointy Nose	A picture of a person with rounded nose	48,03	47,39	47,75	-0,16
Receding Hairline	A picture of a person without receding hairline	35,49	31,36	33,69	-1,49
Rosy Cheeks	A picture of a person with pale cheeks	47,46	44,28	46,49	+3,16
Sideburns	A picture of a person without sideburns	27,58	35,15	32,12	+3,47
Smiling	A picture of a person who is serious	56,84	83,7	61,38	-20,38
Wearing Earrings	A picture of a person without earrings	34,6	33,59	34,11	+1,55
Wearing Hat	A picture of a person without hat	21,41	10,97	17	-6,34

Wearing Hat	A picture of a haired person	67,35	63,6	65,25	-24,09
Wearing Lipstick	A picture of a person without lipstick	47,82	27,48	46,02	+1,04
Wearing Necklace	A picture of a person without necklace	27,88	34,58	31,82	-6,1
Wearing Necktie	A picture of a person without necktie	11,23	30,25	23,82	-3,36
Young	A picture of an aged person	62,86	82,34	68,4	+6,6
Young	A picture of an old person	53,22	97,63	56,04	-5,76

Tabla 4: resultados del experimento “Frase neutral vs Descripciones opuesta”

En esta tabla se muestra la aplicación del método “Frase neutral vs Descripción opuesta” con la mayoría de las etiquetas del conjunto de datos “Celeba”, algunas probando varias opciones para dar un sentido inverso a la frase descriptiva. En la primera columna encontramos la etiqueta, en la segunda el “prompt” descriptivo modificado para no cumplir con la etiqueta “Celeba”, en las siguientes los resultados obtenidos de RVP, RVN, TP, y en la última la mejora obtenida respecto al experimento del **Apartado 4.1 (Tabla 3)**.

En este experimento, en el que se modifican las descripciones con el objetivo de detectar los casos falsos, en general, no se consiguen buenos resultados. En la mayoría de los casos, o bien ya se tenían un resultado malo, que no mejora mucho, o simplemente se reduce la precisión. Con los casos cercanos al 50%, tanto se puede mejorar como empeorar, pero no son relevantes puesto que prácticamente son azar. Parece que al añadir palabras como “no” o “without”, CLIP no siempre es capaz de interpretar-las correctamente. En casos como las etiquetas “Goatee”, “Wearing Hat”, o “Wearing Necktie” donde se obtenía cerca de un 80% de precisión, se aprecia muy bien este fenómeno, puesto que al añadir “without” obtienen cerca de un 20% de precisión, es decir, prácticamente el mismo resultado, aunque requiere invertir la etiqueta porque conseguimos valores por debajo del 50%.

Hay algunos casos peculiares, como la etiqueta “Heavy Makeup”, en la que al añadir la palabra “without” mejoran los resultados en aproximadamente un 7%, consiguiendo casi un 64% de precisión, pero que sigue detectando los casos verdaderos y no los falsos. Otro caso raro es el de la etiqueta “No beard” que pierde eficacia al quitar la palabra “No” del “prompt”.

La única etiqueta con cierto éxito en este experimento es la etiqueta “Young”, en la que, usando la palabra “aged”, se consigue mejorar los resultados en más de un 5%

invirtiendo realmente la clasificación, puesto que se consiguen detectar las personas mayores. Aun así, simplemente se pasa de 62% a 68% de precisión, que no son valores buenos. Hay otras etiquetas, como “Male”, “Bald”, “Pale Skin” o “Smiling”, en las que también se consigue invertir realmente los resultados, pero a costa de perder mucha precisión.

4.3. Descripción vs Descripción opuesta

En este experimento se utiliza una estrategia diferente, aunque se siguen usando 2 “prompts”, ahora se deshecha la frase de referencia y se usan dos frases con sentido opuesto. Para hacerlo, en general se aprovechan los “prompts” opuestos de los **Tablas 3 y 4**. La idea es que, al usar la referencia, recaía todo el peso de conseguir buenos resultados en la frase descriptiva, de esta forma los dos “prompts” contribuyen a realizar la clasificación, permitiendo aprovechar un poco más el potencial de CLIP y alcanzar así mejores resultados. Obviamente, a costa de poner más dedicación a construir los “prompts” de cada etiqueta y también causando que, para adaptar este método al usuario, éste debería introducir las dos descripciones opuestas.

Etiquetas “Celeba”	"Prompts" descriptivos	"Prompts" opuestos	RVP	RVN	PT	Mejora
Male	A picture of a male person	A picture of a female person	99,24	98,17	98,7	1,59
Male	A snapshot of a male person	A snapshot of a female person	99,11	98,27	98,69	1,58
Male	A image of a male person	A image of a female person	99,24	98,07	98,65	1,54
Male	A photo of a male person	A photo of a female person	99,27	98,03	98,64	1,53
Male	A picture of a man	A picture of a woman	99,39	97,72	98,54	1,43
Male	A picture of a male	A picture of a female	99,24	97,52	98,36	1,25
Male	A Male	A Female	99,14	97,54	98,32	1,21
Male	A picture of a male or a man	A picture of female or a woman	99,34	97,18	98,24	1,13
Male	A image of a person who is a male	A image of a person who is a female	98,61	97,72	98,16	1,05
Male	A snapshot of a person who is a male	A snapshot of a person who is a female	97,79	98,17	97,98	0,87
Male	A image of a person who is male	A image of a person who is female	98,01	97,63	97,82	0,71
Male	A picture of a person who is male	A picture of a person who is female	98,23	97,34	97,79	0,68
Male	A picture of a person who is a male	A picture of a person who is a female	98,28	97,23	97,75	0,64
Male	A photo of a person who is male	A photo of a person who is female	97,91	97,41	97,66	0,55
Male	A snapshot of a person who is male	A snapshot of a person who is female	96,99	98,26	97,61	0,5

Male	A photo of a person who is a male	A photo of a person who is a female	97,4	97,55	97,48	0,37
Wearing Hat	A picture of a person with hat	A picture of a person with hair	92,2	89,54	90,82	1,48
Wearing Hat	A picture of a person with hat	A picture of a person with hair	92,2	89,54	90,82	1,48
Male	Male	Female	84,85	98,98	90,72	-6,39
Bald	A picture of a bald person	A picture of a haired person	96,08	80,43	86,65	5,07
Male	A picture of a person who is a male	A picture of a person who is not a male	78,76	97,87	85,94	-11,17
Smiling	A picture of a person who is smiling	A picture of a person who is serious	80,05	89,88	84,28	2,52
Wearing Hat	A picture of a hatted person	A picture of a haired person	82,22	82,34	82,28	-7,06
Eyeglasses	A picture of a person with eyeglasses	A picture of a person without eyeglasses	90,45	70,98	77,62	-4,16
Bangs	A picture of a person with bangs	A picture of a person without bangs	89,74	70,59	77,12	-4,93
Wearing Hat	A picture of a person with hat	A picture of a person without hat	78,11	74,08	75,94	-13,4
Pale Skin	A picture of a person with pale skin	A picture of a person with tanned skin	68,12	75,8	71,29	3,35
Wavy Hair	A picture of a person with wavy hair	A picture of a person with straight hair	66,6	68	67,28	-6,03
Bushy Eyebrows	A picture of a person with bushy eyebrows	A picture of a person with wispy eyebrows	69,32	62,03	64,82	0,62
Receding Hairline	A picture of a person with receding hairline	A picture of a person without receding hairline	70,76	60,7	64,12	-3,68
Straight Hair	A picture of a person with straight hair	A picture of a person with wavy hair	60,38	67,04	62,9	6,96
Mouth Slightly Open	A picture of a person with open mouth	A picture of a person with closed mouth	61,03	57,44	58,89	2,17
Double Chin	A picture of a person with double chin	A picture of a person without double chin	60,86	56,72	58,3	3,69
Wearing Necktie	A picture of a person with necktie	A picture of a person without necktie	69,12	55,01	57,94	-21,6
Chubby	A picture of a chubby person	A picture of a slender person	55,49	63,62	57,82	2,26
Big Lips	A picture of a person with big lips	A picture of a person with small lips	64,79	54,62	57,04	5,58
Mustache	A picture of a person with mustache	A picture of a person without mustache	74,07	53,33	55,85	-12,04
Wearing Earrings	A picture of a person with earrings	A picture of a person without earrings	67,42	53,46	55,78	-8,56
Blurry	A blurry picture of a person	A clear picture of a person	81,26	53	55,48	0,29

Wearing Lipstick	A picture of a person with lipstick	A picture of a person without lipstick	71,7	52,54	54,54	1,6
Heavy Makeup	A picture of a person with heavy makeup	A picture of a person without heavy makeup	66,48	52,61	54,5	-1,16
No Beard	A picture of a person with no beard	A picture of a person with beard	52,43	74,87	54,44	-20,47
Sideburns	A picture of a person with sideburns	A picture of a person without sideburns	58,75	51,67	52,8	-11,61
Pointy Nose	A picture of a person with pointy nose	A picture of a person with nose	51,36	68,23	52,52	0,11
Pointy Nose	A picture of a person with pointy nose	A picture of a person with rounded nose	51,68	51,1	51,32	-1,09
Big Nose	A picture of a person with big nose	A picture of a person with small nose	53,56	50,77	51,26	-0,65
High Cheekbones	A picture of a person with high cheekbones	A picture of a person without high cheekbones	50,43	50,24	50,31	-0,89
Attractive	A picture of an attractive person	A picture of an unattractive person	50,44	50,13	50,2	-2,74
Goatee	A picture of a person with goatee	A picture of a person without goatee	50,1	50,08	50,09	-32,69
High Cheekbones	A picture of a person with high cheekbones	A picture of a person with low cheekbones	43,54	49,67	49,38	-0,58
Narrow Eyes	A picture of a person with narrow eyes	A picture of a person with separated eyes	49,13	47,5	48,71	-4,91
Pointy Nose	A picture of a person with rounded nose	A picture of a person with nose	28,34	48,58	47,34	0,25
Rosy Cheeks	A picture of a person with rosy cheeks	A picture of a person with pale cheeks	46,22	46,73	46,49	3,16
Bags Under Eyes	A picture of a person with bags under eyes	A picture of a person without bags under eyes	46,19	46,13	46,16	2,5
Wearing Necklace	A picture of a person with necklace	A picture of a person without necklace	31,77	46,53	44,18	-18,46
Arched Eyebrows	A picture of a person with arched eyebrows	A picture of a person with straight eyebrows	29,91	41,34	37,9	1,78

Tabla 5: resultados de “Descripción vs Descripción Opuesta”

En esta tabla se muestra la aplicación del método “Descripción vs Descripción opuesta” probando varias opciones para diferentes etiquetas. En la primera columna encontramos la etiqueta, en la segunda el “prompt” descriptivo que cumple con la característica de la etiqueta y en la tercera el “prompt” modificado para no cumplir con la

etiqueta. En las últimas columnas se encuentran los resultados obtenidos de RVP, RVN, TP y la mejora obtenida respecto al experimento del **Apartado 4.1 (Tabla 3)**. Se muestran los datos ordenados por precisión, puesto que los resultados de precisión total que están por debajo del 60%, pueden presentar mejoras relativamente grandes, pero poco relevantes realmente, puesto que el 50% sería puro azar, y el 55%, aunque presenta un 5% más, no es mucho mejor.

Aplicar la estrategia de usar un antónimo para crear la descripción opuesta, suele tener un efecto positivo. Si los resultados con la referencia neutral ya eran correctos, por encima, o al menos cerca, del 70%, se pueden mejorar ligeramente. Son ejemplos de ello los resultados de las etiquetas “bald”, “male”, “pale skinned”, “smiling” y “wearing hat”, al contraponer las palabras clave “bald / hair”, “male / female” o “man / woman”, “pale / tanned”, “smiling / serious” y “hat / hair” respectivamente. Si los resultados con la referencia neutral eran malos, lo que sería aproximadamente estar por debajo del 60%, es difícil conseguir mejorarlos con este método, seguramente porque la palabra o estructura usada, de entrada, ya no consigue una buena interpretación o modelado por parte de CLIP.

Al aplicar esta técnica usando conectores de negación como “not” o “without”, en general no se obtienen buenos resultados ni mejoras. De hecho, con etiquetas como “goatee”, “mustache”, “no beard”, “sideburns”, “wearing necklace” o “wearing necktie” se reduce su precisión prácticamente al 50% o azar al usar este método. Hay alguna excepción, por ejemplo, la etiqueta “male”, con los “prompts” “A picture of a person who is a male / A picture of a person who is not a male”, que consigue un 85%, aunque sigue empeorando los resultados al compararlos con la estrategia de la referencia neutral.

4.4. Múltiples descripciones

Este experimento se centra en las etiquetas sobre el color de pelo, que son “Black hair”, “Brown Hair”, “Blond hair” y “Gray hair”. Estas etiquetas del conjunto de datos “Celeba” contemplan la posibilidad de que una persona tenga el pelo de dos colores, con lo que podemos encontrar que una imagen está etiquetada, por ejemplo, con pelo rubio y con pelo castaño. Esto no encaja bien con las normas del juego, puesto que se deben clasificar las imágenes en grupos sin generar confusión. Por ello, en este caso se eligen 8000 imágenes, las 2000 primeras de cada clase, forzando que sean excluyentes, es decir, sin seleccionar nunca imágenes con dos etiquetas verdaderas. Se han creado frases descriptivas con tres estructuras distintas para comparar resultados. La usada en el primer experimento “A picture of a person with ... hair”, y dos variantes distintas, “A picture of a ...-haired person” y “A picture of a person who is ...-haired”. Teniendo en cuenta que en el primer experimento “Brown hair” obtuvo menos de un 60% de precisión, se ha elegido un sinónimo de “Brown”, “Tawny”, que se ha incluido en estas pruebas para comparar resultados entre las dos opciones.

Al usar imágenes diferentes, para poder comparar correctamente las precisiones de las estrategias usadas, en la siguiente tabla se han calculado los resultados al utilizar la estrategia inicial, con dos “prompts” para cada etiqueta y la referencia “A picture of a person”, en el nuevo conjunto de imágenes. De esta forma se podrá calcular la mejora obtenida al aplicar diferentes estrategias.

Etiquetas “Celeba”	“Prompts” descriptivos	RVP	RVN	PT	PT media
Gray Hair	A picture of a person with gray hair	86,27	78,35	84,89	69,25
Blond Hair	A picture of a person with blond hair	71,96	98,32	70,82	
Black Hair	A picture of a person with black hair	69,17	81,79	67,72	
Brown Hair	A picture of a person with brown hair	57,86	89,34	46,42	

Tabla 6: resultados de la estrategia del primer experimento con las nuevas imágenes elegidas para este caso (con un único color de pelo etiquetado)

En esta tabla se muestran los resultados obtenidos al aplicar la estrategia de “Frase neutral vs Descripción”, con los nuevos datos sin etiquetas con color de pelo duplicado, para las 4 etiquetas referentes a color de pelo. Se observa que los resultados son un poco

diferentes. Con el nuevo conjunto de imágenes la etiqueta “gray hair” es la única que muestra una precisión ligeramente mayor, pasando de 78% a 84%. Las demás etiquetas, “black hair”, “blond hair” y “brown hair” reducen su porcentaje. La etiqueta “blond hair” pasa de 82% a 70% y “black hair” de 74% a 67%. La de “brown hair” reduce su porcentaje aproximadamente al 50%, hasta el punto en que se detectan más casos de pelo marrón con la referencia que con la frase descriptiva.

En el siguiente experimento se calcula la mejora obtenida al aplicar distintas modificaciones a los “prompts”, como cambiar la estructura de la frase o usar algún sinónimo. Básicamente se modifica la estructura de la frase usando “...-haired” o se usa el sinónimo “tawny” para pelo marrón. La estrategia sigue siendo comparar la descripción con la referencia neutral.

Etiquetas “Celeba”	"Prompts" descriptivos	RVP	RVN	PT	Mejora
Black Hair	A picture of a black-haired person	75,16	76,35	75,16	+7,44
Black Hair	A picture of a person who is black-haired	66,48	83,88	63,54	-4,18
Blond Hair	A picture of a blond-haired person	76,41	98,44	76,82	+6
Blond Hair	A picture of a person who is blond-haired	64,11	98,88	58,1	-12,72
Brown Hair	A picture of a person with tawny hair	59,32	77,06	51,68	-1,9
Brown Hair	A picture of a tawny-haired person	59,13	78,28	51	-2,58
Brown Hair	A picture of a brown-haired person	58,67	84,91	48,82	-4,76
Brown Hair	A picture of a person who is tawny-haired	56,75	80	45,21	-8,37
Brown Hair	A picture of a person who is brown-haired	54,26	88,87	37,62	-15,96
Gray Hair	A picture of a gray-haired person	90,77	76,86	87,52	+2,63
Gray Hair	A picture of a person who is gray-haired	74,96	83,9	74,95	-9,94

Tabla 7: resultados de la estrategia del primer experimento con las nuevas imágenes y las modificaciones propuestas

En esta tabla se muestran los resultados al usar la combinación “color-haired” de diferentes maneras para las 4 etiquetas referentes al color de pelo. Con la etiqueta “Brown Hair”, además, se prueba el sinónimo “tawny” como sustituto de “brown”. En este segundo experimento, en el que aún no se modifica la estrategia con los “prompts” y se sigue usando la misma que la del experimento anterior, simplemente modificando los “prompts” usados, se observa que usar frases modificadas con la combinación “color-haired” tiene un efecto positivo en algunas de las etiquetas analizadas, consiguiendo mejorar las precisiones totales de las referentes a pelo negro y pelo rubio cerca de un 6%.

Finalmente se realiza un experimento clasificando directamente las 4 clases de pelo mediante CLIP, es decir, en vez de usar dos “prompts”, el de referencia y la descripción, para determinar si una imagen tiene o no el pelo de un determinado color, se usan cuatro “prompts” para determinar de qué color tiene el pelo la imagen introducida. En este caso, por lo tanto, no se usa el “prompt” de referencia. En la siguiente tabla se muestran los resultados usando esta estrategia con las diferentes variantes propuestas.

Etiquetas Celeba	"Prompts" descriptivos	RVP	RVN	PT	Mejora	PT media
Black Hair	A picture of a black-haired person	90,42	78,97	87,88	+20,16	86,48 (Mejora: +17,23)
Blond Hair	A picture of a blond-haired person	88,36	88,79	88,44	+17,62	
Brown Hair	A picture of a tawny-haired person	82,06	77,99	81,58	+28	
Gray Hair	A picture of a gray-haired person	98,69	68,41	88	+3,11	
Black Hair	A picture of a black-haired person	94,19	71,06	87,39	+19,67	85,85 (Mejora: +16,6)
Blond Hair	A picture of a blond-haired person	91,22	93,86	91,72	+20,9	
Brown Hair	A picture of a brown-haired person	77,54	83,1	77,78	+24,2	
Gray Hair	A picture of a gray-haired person	99,1	65,37	86,49	+1,6	
Black Hair	A picture of a person who is black-haired	91,31	79,95	88,71	+20,99	85,97 (Mejora: 16,72)
Blond Hair	A picture of a person who is blond-haired	82,75	91,19	83,64	+12,82	
Brown Hair	A picture of a person who is tawny-haired	84,25	72,97	82,42	+28,84	
Gray Hair	A picture of a person who is gray-haired	97,31	71,88	89,11	+4,22	
Black Hair	A picture of a person who is black-haired	94,72	67,99	86,31	+18,59	85,77 (Mejora: +16,52)
Blond Hair	A picture of a person who is blond-haired	90,5	94,69	91,26	+20,44	
Brown Hair	A picture of a person who is brown-haired	77,32	81,12	77,47	+23,89	
Gray Hair	A picture of a person who is gray-haired	98,55	68,56	88,02	+3,13	
Black Hair	A picture of a person with black hair	90,9	75,28	87,15	+19,43	86,05 (Mejora: +16,8)
Blond Hair	A picture of a person with blond hair	85,48	92,83	86,46	+15,64	
Brown Hair	A picture of a person with tawny hair	83,03	76,65	82,18	+28,6	
Gray Hair	A picture of a person with gray hair	98,39	69,44	88,41	+3,52	
Black Hair	A picture of a person with black hair	93,07	69,18	86,14	+18,42	86,04 (Mejora: +16,79)
Blond Hair	A picture of a person with blond hair	90,09	93,97	90,79	+19,97	
Brown Hair	A picture of a person with brown hair	79,36	83,38	79,64	+26,06	
Gray Hair	A picture of a person with gray hair	98,83	67,57	87,61	+2,72	

Tabla 8: resultados de “Descripción vs Múltiples Descripciones”

En la tabla se pueden observar las diferencias al usar la palabra “brown” o la palabra “tawny” para referirse al pelo marrón, porque cambia el color de fondo en cada uno de los

casos. Cada grupo de 4 filas usa la misma estructura textual para crear los “prompts”, que se va modificando ya sea con el sinónimo “tawny” o modificándola, por ejemplo con la palabra “with”. En la última columna se calcula la precisión media de acierto de las 4 clases de color de pelo.

Finalmente, en este tercer experimento se modifica la estrategia para usar múltiples descripciones. Se crean más de dos “prompts” para diferenciar las clases y se consiguen resultados muy buenos, con una precisión media de alrededor del 85-86% en todos los casos o variaciones probadas. Uno de los efectos que se produce al usar este método, es que al modificar la estructura de la frase o usar algún sinónimo, los resultados se mantienen muy estables en media. Es decir, si todos los “prompts” se modifican de la misma manera, con lo que lo único que no cambia entre los “prompts” introducidos a CLIP para diferenciar las clases, es palabra referente al color de pelo (“negro”, “rubio”, “marrón” o “gris”, en el ejemplo), se aprecian variaciones ligeras, de aproximadamente un 5% entre porcentajes de aciertos verdaderos y falsos.

Se demuestra que en casos donde es posible aplicar multiclase, aunque las etiquetas sean binarias, CLIP puede ser mucho más efectivo si dichas clases se tienen en cuenta. Con el ejemplo del color de pelo, se ha visto como todas las etiquetas resultan beneficiadas al usar esta estrategia, con incrementos de precisión de hasta un 28%. Etiquetas que ya tenían buenos resultados de precisión, como las de pelo gris o rubio cerca del 80%, pueden subir hasta cerca del 90%, y las que estaban a 60-74%, suben hasta valores entre 77-87%.

Uno de los beneficios que conlleva esta estrategia en el contexto del juego, es que, si una imagen se clasifica con “pelo marrón” pero realmente es “pelo negro”, este error no afecta o no es percibido por el jugador cuando realiza las preguntas sobre “pelo rubio” o “pelo gris”, ya que la respuesta que recibe es “No”, que es correcta. En cambio, al realizar la clasificación binaria con dos “prompts” el error siempre es un problema. Lo malo de esta estrategia es que no es posible adaptarla para que el usuario final la aplique, puesto que se le debería hacer introducir un “prompt” por cada clase y no es viable adaptar este detalle a un juego de preguntas Si o No. Por lo tanto, esta estrategia solo se podrá usar en el caso de jugar con la modalidad de preguntas predefinidas.

4.5. Aplicar “Fine Tuning”

En este último experimento se trata de aprovechar el Conjunto de datos “Celeba” para reentrenar algunos modelos de redes neuronales profundas públicas y verificar si se pueden mejorar los resultados obtenidos hasta el momento. Se ha usado el propio modelo de CLIP, adaptado como se explica en el **Apartado 3.2.2**, para reentrenarlo tratando de afinar sus resultados, y los modelos VGG19 y Xception, entrenados desde cero para la tarea de clasificar características de las imágenes del conjunto de datos “Celeba”.

Para seleccionar los datos de entrenamiento se utiliza una estrategia parecida a la usada para realizar las pruebas iniciales. Para los datos de test, se siguen usando las primeras 4000 imágenes con etiquetas falsas y las primeras 4000 con etiquetas verdaderas, para cada etiqueta. De forma que se realiza el cálculo de la precisión de los modelos siempre con los mismos datos, que son los mismos con los que se comprobó la eficacia de las estrategias creando “prompts” al usar CLIP en “Zero Shot” en los experimentos anteriores. Para los datos de entrenamiento, se utilizan las 4000 imágenes siguientes, con etiquetas verdaderas y falsas, de forma que se usan 8000 imágenes para entrenamiento. Con estos valores se realiza un reparto de 50% de datos de entrenamiento y 50% de test, con 16000 imágenes en total 8000 falsas y 8000 verdaderas. Hay que tener en cuenta que, debido a los tiempos de proceso y a los recursos disponibles, trabajar con más de 8000 imágenes de entrenamiento resulta muy costoso, también que realmente se dispone de imágenes de sobra y que tener muchas imágenes de test penaliza poco porque chequear un modelo no es tan costoso como entrenarlo. De los 8000 datos de entrenamiento un 25% se usan para la validación, siempre repartiendo al 50% los casos verdaderos y falsos.

Como se ha comentado se eligen solo tres modelos, CLIP, VGG y Xception, para realizar estas pruebas y comparar resultados, puesto que la duración de los entrenamientos y el procesado de datos en general limitan el tiempo disponible. Para adaptar estos modelos se añade un “Top Model” que consiste en una capa “Flatten” para redimensionar la entrada, una lineal y una con activación “softmax” binaria para obtener la respuesta a la pregunta. En la **Figura 8** se muestra con un esquema de cómo se realiza esta adaptación en CLIP.

En la siguiente tabla se muestran los resultados obtenidos al realizar el entrenamiento de las redes CLIP, VGG19 y Xception con las imágenes del conjunto de datos “Celeba”, para varias etiquetas que obtuvieron buenos y malos resultados en el experimento del **Apartado 4.1**. En la primera columna se muestra la etiqueta chequeada, en la segunda el modelo usado y en las siguientes los resultados obtenidos.

Etiquetas “Celeba”	Red	RVP	RVN	PT	Mejora
bald	CLIP	97,25	95,97	96,6	+15,02
	VGG19	95,52	93,93	94,71	+13,13
	Xception	93,26	91,78	92,51	+10,93
big nose	CLIP	76,72	76,62	76,67	+24,76
	VGG19	69,86	68,75	69,25	+17,34
	Xception	68,31	68,23	68,27	+16,36
high cheekbones	CLIP	83,21	83,64	83,42	+32,22
	VGG19	83,34	76,49	79,52	+28,32
	Xception	71,83	71,22	71,52	+20,32
male	CLIP	98,48	99,11	98,8	+1,69
	VGG19	93,87	92,79	92,7	-4,41
	Xception	91,2	88,45	89,77	-7,34
rosy cheeks	CLIP	84,76	88,92	86,72	+36,37
	VGG19	84,5	85,43	84,96	+34,61
	Xception	78,56	85,07	81,49	+31,14
smiling	CLIP	88,95	88,49	88,72	+6,96
	VGG19	81,83	84,06	82,91	+1,15
	Xception	71,26	70,48	70,86	-10,9
young	CLIP	85,29	83,72	84,49	+21,68
	VGG19	75,2	75,93	75,56	+12,75
	Xception	73,56	73,05	73,06	+10,25
wearing hat	CLIP	96,79	96,63	96,71	+7,37
	VGG19	95,37	94,36	94,86	+5,52
	Xception	91,22	93,29	92,22	+2,88

Tabla 9: resultados reentrenando modelos CLIP, VGG19 y Xception, con las etiquetas que tuvieron los peores y mejores resultados en el primer experimento

Para el caso de las etiquetas relativas a “Color de pelo” se usan los datos del experimento anterior (**Apartado 4.4**), y así se pueden entrenar los modelos mediante imágenes etiquetadas con un único color de pelo. También la mejora se calcula en referencia

a estos datos, para que la comparación de resultados entre estrategias sea con los mismos datos de test. Se calculan los datos como si las etiquetas fueran binarias, de forma que la predicción relativa a los aciertos de que el pelo no es negro (RVN) no requieren acertar si el pelo es rubio, marrón o gris, solamente es necesario que dicha predicción no sea pelo negro. En la columna “PT media”, se muestra la precisión relativa a acertar todas las clases.

Etiquetas “Celeba”	Red	RVP	RVN	PT	Mejora	PT media
black hair	CLIP	87,15	90,63	95,2	27,48	90,35 (Mejora: +21.1)
blond hair	CLIP	97,26	91,2	95,74	24,92	
brown hair	CLIP	93,87	85,91	91,99	38,41	
gray hair	CLIP	98,57	95,4	97,77	12,88	
black hair	VGG19	95,29	90,27	94,1	26,38	87.4 (Mejora: +18,15)
blond hair	VGG19	96,04	89,83	94,52	23,7	
brown hair	VGG19	93,33	83,07	90,87	37,29	
gray hair	VGG19	96,84	90,66	95,3	10,41	
black hair	Xception	93,28	88,23	92,15	24,43	82.76 (Mejora: +13.51)
blond hair	Xception	93,55	86,69	91,96	21,14	
brown hair	Xception	91,42	75,99	87,55	33,97	
gray hair	Xception	95,47	88,83	93,86	8,97	

Tabla 10: resultados reentrenando modelos CLIP, VGG19 y Xception para las etiquetas referentes al color del pelo.

En este experimento, en el que se aprovechan algunas redes neuronales profundas populares, adaptándolas para responder las preguntas del juego, y se comparan las precisiones conseguidas con las obtenidas mediante el método de “Zero-Shot” con CLIP, se han logrado muy buenos resultados con los tres modelos probados. Cabe destacar que, en todas las pruebas realizadas con diferentes etiquetas, CLIP siempre obtiene los mejores resultados. Es así hasta el punto de que la etiqueta “male”, que ya tenía unos resultados sorprendentes sin realizar “Fine Tuning” cerca del 98% de precisión, solo mejora en aproximadamente un 1% al aplicar “Fine Tuning” a CLIP, es decir, ni VGG19 ni Xception consiguen mejorar dichos resultados. Con lo que CLIP en “Zero-Shot” supera las dos en este caso concreto. Con la etiqueta “smiling” sucede algo parecido, aunque VGG19 si consigue superar ligeramente los resultados que se obtienen con CLIP en “Zero-Shot”.

Otro caso muy interesante sobre el potencial de CLIP en “Zero-Shot” es el experimento realizado con las etiquetas de color de pelo, donde se consigue una media de más del 85% de precisión (**Tabla 8**). Es cierto que al realizar “Fine-Tuning” a CLIP se logra subir hasta un 90% (**Tabla 10**), que no es poco, pero también es importante comentar que Xception no consigue superar los resultados de CLIP en “Zero-Shot”, en este caso concreto, y que VGG19, aunque lo supera, es por muy poco, llegando más o menos al 87%.

Además, hay que tener en cuenta que el tiempo de entrenamiento del modelo CLIP es mucho menor que el de VGG19 o el de Xception. En concreto con la etiqueta “male”, en los experimentos realizados con la plataforma Google Colab”, que permite usar GPU durante un cierto tiempo, la épocas de entrenamiento tardan menos de 1 segundo (puesto que solo se deben entrenar 1.026 parámetros)., mientras que VGG19 (20.033.602 parámetros) requiere unos 50 segundos y Xception (20.872.490 parámetros) unos 60. Estos resultados se pueden encontrar en el **Anexo 8.2.6 - Quien es Quien - archivo “pdf” de los experimentos**, en el apartado **Etiqueta: “male” - Modelos**).

5. Valoración y discusión de los resultados

A continuación, se comentan algunos de los resultados más destacables obtenidos en los experimentos del apartado anterior.

En primer lugar, cabe señalar la importancia de usar lenguaje natural al crear los “prompts” con la metodología de CLIP en “Zero-Shot” descrita en el **Apartado 3.2.1**, usada en la mayoría de los experimentos. En el experimento de la **Tabla 2**, donde se busca una buena frase de referencia, se realiza la prueba de usar la etiqueta “Celeba” como “prompt” directamente, es decir, para el caso de la etiqueta de género usar las combinaciones “person / male” o “person / female” para codificar los “prompts” a CLIP. En este ejemplo es muy impactante como al añadir la “A”, lo que sería “Un / una”, a la etiqueta “male”, creando los “prompts” nuevos “A person / A male” o “A person / A female”, como haríamos al hablar para referirnos a algo, se consigue en ambos casos una buena mejora, sobre todo en el caso de usar “female” con un 28% de mejora en la precisión total. Lo más curioso de este experimento es que la diferencia entre las frases usadas siempre es la misma, lo que parece confirmar que CLIP interpreta las estructuras textuales, con lo que responde mejor si los “prompts” usan lenguaje natural que si son simples etiquetas.

Otro caso curioso de los primeros experimentos, en la **Tabla 3**, es el de la etiqueta “No beard”, que incluye una negación con la palabra “No”. Como se ha comentado parece que CLIP no gestiona siempre bien algunas palabras o conceptos como la negación. En este caso parece que al enfrentar los dos “prompts”, el de referencia “A picture of a person” contra el descriptivo “A picture of a person with no beard”, la palabra “no” pierde su efecto, mientras que la palabra “beard” lo mantiene. De cierta manera, CLIP interpretara que la frase usada es “A picture of a person with beard” en vez de “... with no beard” (“con barba” en vez de “sin barba”), de forma que lo que se discrimina con la descripción (aunque esta incluya la palabra “no”) son las personas que si tienen barba y con la referencia los que no tienen barba. Dando como resultado un valor de precisión total de 25%, o sea por debajo de 50%, aunque no es tan malo porque al invertir la respuesta significa un 75%. Lo más raro de este ejemplo es que

en el experimento del **Apartado 4.2**, en el que se usa una frase opuesta, al usar la frase “A picture of a person with beard”, que viendo los resultados anteriores nos llevaría a pensar en conseguir un 75% o más, se consigue solamente un 60%, como si de alguna forma la palabra “no”, haya contribuido a que CLIP clasifique mejor las personas que si llevan barba. Es un caso que quizás merecería un estudio específico.

En el **Apartado 4.3**, cuando se aplica la estrategia de usar dos descripciones opuestas para crear dos “prompts”, hay un par de ejemplos que merecen especial atención, el de las etiquetas “straight hair” y “wavy hair”. Una de ellas mejora y la otra empeora al usar la estrategia de contraponer precisamente las palabras “straight” y “wavy” (lo que sería pelo estirado y ondulado). Al analizar estos resultados es importante recordar que solamente las etiquetas “Bald” y “Bangs” (“Es calvo” y “Tiene flequillo”) son excluyentes en este conjunto de imágenes. Esto implica que “wavy hair” y “straight hair” no se deben considerar características opuestas porque hay imágenes etiquetadas con pelo ondulado y estirado. Cualquier imagen etiquetada con las dos características se convierte inmediatamente en acierto para un caso, por ejemplo, para la etiqueta “straight”, pero fallo para el otro. Puesto que si se acierta “straight” se falla “wavy”. Para analizar el comportamiento de CLIP con las imágenes de pelo estirado y ondulado se debería usar otro conjunto de imágenes, o al menos descartar las imágenes con doble etiqueta.

En el experimento con “Fine Tuning” del **Apartado 4.5** es muy remarcable que, en todos los ejemplos, al comparar los tres modelos probados, CLIP, VGG19 y Xception, se mantiene siempre el mismo orden en la precisión obtenida. CLIP es el mejor, seguido de cerca por VGG19 y finalmente Xception. Mediante CLIP se consiguen resultados siempre por encima del 85% y los tiempos de entrenamiento con la metodología propuesta son muy pequeños. Es también digno de mención el ejemplo con la etiqueta “Rosy cheeks”, que se eligió por obtener los peores resultados al usar CLIP en modo “Zero-Shot” (prácticamente 50%), y con el “Fine-Tuning” de CLIP se logra subir hasta el 86% de precisión.

En general, en los experimentos realizados con la metodología “Zero-Shot” de CLIP, se aprecia claramente el potencial que tiene, puesto que todos los resultados se obtienen sin necesidad de entrenamiento, ni de “Fine Tuning”, y sin la necesidad de ser experto en el

lenguaje inglés, idioma mediante el cual se deben crear los “prompts” que permiten diferenciar la clases.

6. Conclusiones y desarrollos futuros

Los experimentos realizados han demostrado el gran potencial de CLIP trabajando como clasificador de imágenes “Zero-Shot” a partir de descripciones textuales en forma de lenguaje natural. Aunque para casos complejos o para optimizar resultados, es necesario realizar “prompt engineering”. En este sentido, para casos de clasificación binaria, usando dos “prompts”, la estrategia de usar una frase de referencia y varias descripciones modificadas puede ser un buen recurso para hacer un estudio previo de los “prompts” que pueden ser más eficaces en un contexto determinado. Por otra parte, usar frases con negaciones no suele tener buenos resultados, al menos con las pruebas realizadas, centradas en descripciones de caras de personas. La estrategia de usar descripciones parecidas, pero con algún antónimo o que las hace opuestas, es muy buena opción para hacer “Zero-Shot” con CLIP. Para casos que realmente no son binarios, es decir, que se podrían distinguir varias clases distintas claramente, en vez de solamente dos, es una buena estrategia crear un “prompt” para cada clase, de forma que sean muy parecidos y cambien sutilmente para describir cada una de las posibles clases, como con las etiquetas “color de pelo” en el **Apartado 4.4**. Aunque se han realizado pocas pruebas con esta estrategia, los resultados han sido buenos.

Las pruebas con “Fine Tuning” han dado buenos resultados. De hecho, con CLIP han sido muy buenos, siempre han conseguido los mejores valores de precisión total respecto a las redes VGG19 y Xception. Es realmente impresionante como CLIP supera VGG19 y Xception en absolutamente todos los casos probados. Quizás esto indica, en parte, que un buen “prompt engineering” de CLIP en “Zero-Shot”, también lo podría lograr, puesto que los dos métodos se basan en la misma codificación de la imagen que realiza el modelo CLIP.

En cuanto al uso de “streamlit”, la plataforma para crear una “App web” de forma sencilla, se puede decir que verdaderamente es una herramienta útil para crear una aplicación final rápidamente y accesible al público, con el único requerimiento de poseer algunos conocimientos básicos de Python. El único problema encontrado es que sus

recursos gratuitos pueden ser un poco justos para usar CLIP. Se podría tratar de mejorar el rendimiento de la aplicación para lograr que la limitación de recursos de “streamlit” no sea ningún problema.

Se ha conseguido una aplicación jugable que respeta la filosofía del juego, aunque para un solo jugador. Se permite al usuario introducir sus imágenes o jugar con imágenes del conjunto de datos “Celeba”. Se han creado tres opciones para que el jugador pueda realizar las preguntas. La primera permite al jugador elegir entre un conjunto de preguntas predefinidas, aunque en realidad se usa CLIP con las mejores configuraciones de “prompts” encontradas durante la elaboración de este trabajo. Las otras dos utilizan CLIP en “Zero-Shot” configurado con dos “prompts”. En la segunda opción se usa un “prompt” de referencia y uno introducido por el jugador. Siempre se usa la misma referencia, como en el **Apartado 4.1** la frase es “A picture of a person”, y se pide al jugador que modifique dicha frase para crear el segundo “prompt”. En la tercera opción se pide al jugador que introduzca los dos “prompts”, con la idea de que el mismo pueda aplicar la estrategia del **Apartado 4.3** que consiste en crear dos frases opuestas. Las clasificaciones de imágenes realizadas con preguntas por defecto, usando lo aprendido sobre CLIP, consiguen resultados muy buenos, aunque de vez en cuando se pueden ver fallos como por ejemplo un hombre clasificado como mujer, puesto que no se ha conseguido ningún caso de 100% de precisión. En los modos con “prompts” introducidos por el usuario, si no hace caso del consejo de usar una frase parecida a “A picture of a person” para la opción de un solo “prompt”, o frase parecidas de sentido opuesto para la opción de dos “prompts”, los resultados pueden no ser buenos. Cuando los “prompts” difieren mucho es fácil que CLIP clasifique todas las imágenes en un mismo grupo y por lo tanto no se consiga avanzar en el juego, puesto que en tal caso no se descartan imágenes.

Solo se han incluido en el juego las preguntas por defecto que tenían resultados razonablemente buenos en los experimentos realizados. Es por ello que se podría aprovechar el conjunto de datos “Celeba”, para hacer más “prompt engineering” con CLIP, hasta poder incluir al menos las 40 preguntas que permiten hacer las etiquetas de este conjunto de datos. Se podría profundizar en el estudio de la etiqueta “No beard”, que seguramente se puede combinar con algunas otras como “5 o'clock shadow”, y en el estudio

de las etiquetas “wavy hair” y “straight hair”, con las que se podría realizar un experimento parecido al llevado a cabo con las etiquetas de “color de pelo”.

Una de las opciones que también se podrían hacer en un futuro, y que no debería ser muy complicado, sería crear la versión de dos jugadores, de forma que cada uno de los usuarios pudiera elegir una de las imágenes del panel y después empezaran los turnos de preguntas hasta que uno de los dos adivinara la imagen elegida por el otro.

Finalmente, la última propuesta de futuro para ampliar el conocimiento sobre CLIP podría consistir en estudiar el vocabulario de Internet para aplicar mejores “prompts” a CLIP. Teniendo en cuenta que CLIP se entrenó con lenguaje natural extraído de Internet, tratar de estudiar el lenguaje que CLIP usó en su entrenamiento puede ser una buena forma de profundizar en el “prompt engineering” de CLIP.

7. Referencias bibliográficas

Changpinyo S., Wei-Lun C., Sha F. *Predicting visual exemplars of unseen classes for Zero-Shot learning*. 20 Aug 2017. [arXiv:1605.08151v2](https://arxiv.org/abs/1605.08151v2).

Conde M.V., Turgutlu K.; *Clip-art. Contrastive pre-training for fine-grained art classification*. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 3956–3960. June 2021. <http://doi.org/10.1109/CVPRW53098.2021.00444>.

Cornia M., Stefanini M., Baraldi L., Cucchiara R. *Meshed-memory transformer for image captioning*. In. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10578–10587 (2020). [arXiv:1912.08226v2](https://arxiv.org/abs/1912.08226v2).

Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. *An image is worth 16x16 words. Transformers for image recognition at scale*. 2020. [arXiv:2010.11929](https://arxiv.org/abs/2010.11929).

Fang H., Xiong P., Xu, L., Chen Y. *CLIP2video. Mastering video-text retrieval via image CLIP*. 2021. [ArXiv abs/2106.11097](https://arxiv.org/abs/2106.11097).

Gao J., Li P., Chen Z., Zhang J. *A survey on deep learning for multimodal data fusion*. Neural Computation 32, 829–864 (2020). https://doi.org/10.1162/neco_a_01273.

Gwyn T., Roy K., Atay M. *Face recognition using popular deep net architectures. A Brief Comparative Study*. Future Internet 2021, 13 (7), 164 (25 June 2021). <https://doi.org/10.3390/fi13070164>.

Hugo Larochelle and Dumitru Erhan and Yoshua Bengio. *Zero-data Learning of New Tasks*. 2008. <https://www.aaai.org/Papers/AAAI/2008/AAAI08-103.pdf>.

Liu, Ziwei and Luo, Ping and Wang, Xiaogang and Tang, Xiaoou, *Deep learning face attributes in the wild*, Proceedings of International Conference on Computer Vision (ICCV). December 2015. <https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html>

Nawaz, S., Calefati, A., Caraffini, M., Landro, N., Gallo, I. *Are these birds similar. Learning branched networks for fine-grained representations*. In. 2019 International Conference on Image and Vision Computing New Zealand (IVCNZ). pp. 1–5 (2019). <https://doi.org/10.1109/IVCNZ48456.2019.8960960>.

Radford A., Wook-Kim J., Hallacy C., Ramesh A., Goh G., Agarwal S., Sastry G., Askell A., Mishkin P., Clark J., Krueger G., Sutskever I. *Learning transferable visual models from natural language supervision*. 26 Feb 2021. [arXiv:2103.00020v1](https://arxiv.org/abs/2103.00020v1).

Ramesh A., Pavlov M., Goh G., Gray S., Voss C., Radford A., Chen M., Sutskever I., *Zero-Shot Text-to-Image generation*. [arXiv:2102.12092v2](https://arxiv.org/abs/2102.12092v2). (DALL-E, <https://openai.com/blog/dall-e/>).

Tan H., Yu L., Bansal M. *Learning to navigate unseen environments. Back translation with environmental dropout*. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Human Language Technologies, Volume 1 (Long and Short Papers). pp. 2610–2621 (2019). <https://aclanthology.org/N19-1268>.

Vaswani A., Shazeer N., Parmar N., Uszkoreit, J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. *Attention is all you need*. 2017. [arXiv.1706.03762](https://arxiv.org/abs/1706.03762).

Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.. *Contrastive learning of medical visual representations from paired images and text*. 2020. [arXiv.2010.00747](https://arxiv.org/abs/2010.00747).

8. Anexos

8.1. Recursos

A continuación, se muestran los enlaces más relevantes. Los primeros 4 hacen referencia a los recursos externos usados en el proyecto. Los demás son los enlaces de Github y de la aplicación web del juego ¿Quién es Quién?, con enlaces concretos a los archivos Python del código de la aplicación y los experimentos (que también está disponible en formato “.pdf”).

8.1.1. Conjunto de datos “Celeba”

<https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html>

8.1.2. CLIP - Github

<https://github.com/openai/CLIP>

8.1.3. CLIP - web

<https://openai.com/blog/clip/>

8.1.4. Streamlit

<https://streamlit.io/>

8.1.5. Quien es Quien - Github

<https://github.com/ArnauDIMAI/CLIP-GuessWho>

8.1.6. Quien es Quien - archivo “.pdf” de los experimentos

<https://github.com/ArnauDIMAI/CLIP-GuessWho/blob/main/Experiments.pdf>

8.1.7. Quien es Quien - código Python de los experimentos

<https://github.com/ArnauDIMAI/CLIP-GuessWho/blob/main/Experiments.ipynb>

8.1.8. Quien es Quien - código Python de la aplicación web

https://github.com/ArnauDIMAI/CLIP-GuessWho/blob/main/app_qesq.py

8.1.9. Quien es Quien - aplicación web

https://share.streamlit.io/arnaudimai/clip-guesswho/main/app_qesq.py

8.2. Artículo publicado

A continuación, se muestra el artículo que se ha escrito durante el desarrollo de este trabajo. Se titula “An implementation of the ‘Guess who?’ game using CLIP” y fue aceptado en el congreso IDEAL 2021 (22nd International Conference on Intelligent Data Engineering and Automated Learning, Manchester 25-27 November, <https://ideal-conf.com/>).

An implementation of the “Guess who?” game using CLIP

Arnau Martí Sarri¹[0000-0001-8886-2139] and
Victor Rodriguez-Fernandez²[0000-0002-8589-6621]

¹ Valencian International University, Calle Pintor Sorolla 21, 46002 Valencia, Spain
Dimai S.L., Camí de la Font Calda 10, 08270 Navarces, Spain
a.marti@dimaisl.com

² School of Computer Systems Engineering, Universidad Politécnica de Madrid, Calle de Alan Turing, 28038 Madrid, Spain victor.rfernandez@upm.es

Abstract. CLIP (Contrastive Language-Image Pretraining) is an efficient method for learning computer vision tasks from natural language supervision that has powered a recent breakthrough in deep learning due to its zero-shot transfer capabilities. By training from image-text pairs available on the internet, the CLIP model transfers non-trivially to most tasks without the need for any data set specific training. In this work, we use CLIP to implement the engine of the popular game “Guess who?”, so that the player interacts with the game using natural language prompts and CLIP automatically decides whether an image in the game board fulfills that prompt or not. We study the performance of this approach by benchmarking on different ways of prompting the questions to CLIP, and show the limitations of its zero-shot capabilities.

Keywords: CLIP · Guess who · Zero-shot learning · Language-image models.

Introduction

The ability to learn at the same time from different data modalities (image, audio, text, tabular data...) is a trending topic in the field of machine learning in general, and deep learning in particular, with many domains of application such as self-driving cars, healthcare and the Internet of Things [5]. Among all the possibilities that multimodal deep learning provides, one of the most interesting ones is the connection of text and images in the same model.

This concept brings into play challenging tasks in the areas of computer vision and natural language processing, such as multimodal image and text classification [7], image captioning [2], and visual-language robot navigation [10]. All of these tasks have the core idea of learning visual perception from supervision contained in text, or vice-versa.

In January 2021, the company OpenAI made a great milestone in the field of language-image models with the presentation of CLIP (Contrastive Language-Image Pre-training) [8] ¹. CLIP is a deep neural network designed to perform zero-shot image classification, i.e., to generalize flawlessly to unseen image classification tasks in which the data and the labels can be different each time. The way CLIP does so is by training on a wide variety of (image, text) pairs that are abundantly available on the internet, instructing the model to predict the most relevant text snippet, given an image. Since the code and weights of CLIP were publicly released on GitHub ², many researchers have explored its zero-shot capabilities in different areas such as art classification [1], video-text retrieval [4] or text-to-image generation [9].

In this work, we present an application of the zero-shot classification capabilities of CLIP in the popular game “Guess who?”, in which the player asks yes/no questions to describe people in a game board and try to guess who the selected person is. Each time the player makes a new question, CLIP will analyze the images in the game board and decide automatically which images fulfill it. Although this could be also tackled with a multi-label image binary classification model with a fixed set of labels, the power behind using CLIP relies on the use of natural language to interact with the model, which gives freedom to the player to ask any question and tests CLIP’s zero-shot generalization capabilities. We release our code in a public Github repository ³.

In summary, the contributions of this paper are:

- The implementation of the game engine based on AI, which allows for the use of any set of images instead of having a fixed board. To the best of our knowledge, there is no other version of the game “Guess who?”, that uses an AI in a similar way.
- The use of CLIP as a zero-shot classifier based on textual prompts, which allows the player to interact with the game through natural language.

The rest of the paper is structured as follows: in Section 2 we give some background on CLIP, in Section 3 we present a description of the game and how CLIP is integrated in it, not as a player but as the engine of the game. Then, in Section 4 we show some experiments on how changing the way the game prompts CLIP about the characteristic of a person affects its classification performance, and finally, in Section 5 we outline the conclusions and provide future lines of research in this topic.

Backgrounds on CLIP

CLIP (Contrastive Language-Image Pre-training), by OpenAI, is based on a large amount of work in zero-shot transfer, natural language supervision and multi-modal learning, and shows that scaling a simple pre-training task is sufficient to achieve competitive zero-sample performance on various image classification data sets. CLIP uses a large number of available supervision sources: the text paired with images found on the Internet. This data is used to create the following agent training task for CLIP: Given an image, predict which of a set of 32,768 randomly sampled text fragments is actually paired with it in the data set. This is achieved by combining a text encoder, built as a Transformer [11], and an image encoder, built as a Vision

¹ The paper was accompanied with a blog post publication:
<https://openai.com/blog/clip/>

² <https://github.com/openai/CLIP>

³ <https://github.com/ArnauDIMAI/CLIP-GuessWho>

Transformer [3], under a contrastive objective that connects them [12]. To the best of our knowledge, there is no other publicly available pretrained model with the scale of CLIP that connects text and image data.

Once pre-trained, CLIP can then be applied to nearly arbitrary visual classification tasks. For instance, if the task of a data set is classifying photos of dogs vs cats, we will check, for each image, whether CLIP predicts that the caption “a picture of a dog” is more likely to be paired with it than “a picture of a cat”. In case the task does not have a fixed set of labels, one can still use CLIP for classification by specifying a text description of a target attribute and a corresponding neutral class. For example, when manipulating images of faces, the target attribute might be specified as “a picture of a blonde person”, in which case the corresponding neutral prompt might be “a picture of a person”. CLIP’s zero-shot classifiers can be sensitive to wording or phrasing, and sometimes require trial and error “prompt engineering” to perform well [8].

Game description

The aim of the game, as in the original one, is to find a specific image from a group of different images of a person’s face. To discover the image, the player must ask questions that can be answered with a binary response, such as “Yes and No”. After every question made by the player, the images that don’t share the same answer that the winning one are discarded automatically. The answer to the player’s questions, and thus, the process of discarding the images will be established by CLIP (See Fig. 1). When all the images but one have been discarded, the game is over.

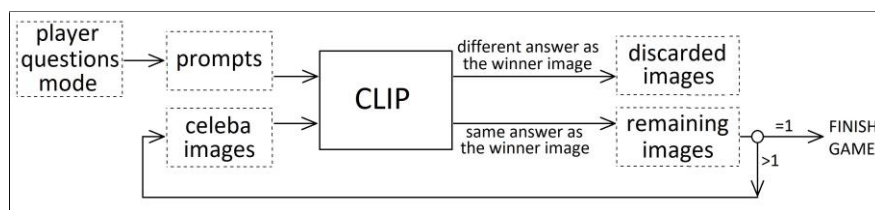


Fig.1: Diagram of how CLIP is integrated in the game.

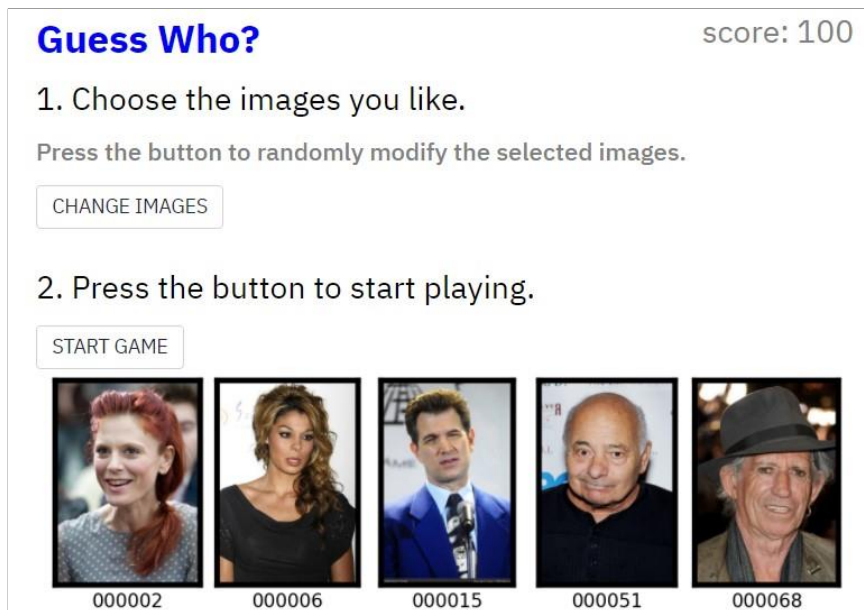


Fig.2: Screen of the game which allows the user selecting the images to play.

The first step of the game is to select the images to play (See Fig. 2). The player can press a button to randomly change the used images, which are taken from the *CelebA* (CelebFaces Attributes) data set [6]. This data set contains 202,599 face images of the size 178×218 from 10,177 celebrities, each annotated with 40 binary labels indicating facial attributes like hair color, gender and age.

Questions

The game will allow the player to ask the questions in 4 different ways:

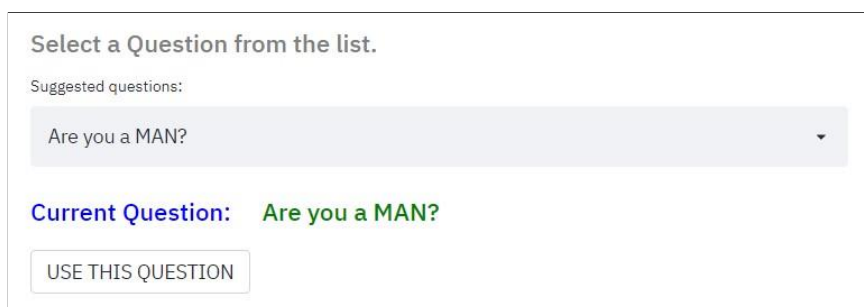


Fig.3: Game screen that allows the user to ask a default question.

1. Asking a question from a list (See Fig. 3). A drop-down list allows the player to select the question to be asked from a group of pre-set questions, taken from the set of binary labels of the Celeba data set. Under the hood, each question is translated into textual prompts for the CLIP model to allow for the

binary classification based on that question. When they are passed to CLIP along with an image, the model responds by giving a greater value to the prompt that is most related to the image.

Write your own prompt and press the button.

It is recommended to use a text like: "A picture of a ... person" or "A picture of a person ..." (CLIP will check -> "Your prompt" vs "A picture of a person")

A picture of a man

USE MY PROMPT: A picture of a man

Fig.4: Game screen that allows the user to create his own prompt using 1 text input.

2. Use one prompt (See Fig. 4). This option is used to allow the player introducing a textual prompt for CLIP with his/her own words. The player text will be then confronted with the neutral prompt, "A picture of a person", and the pair of prompts will be passed to CLIP as in the previous case.

Write your own prompts by introducing 2 opposite descriptions.

Write your "True" prompt:

A picture of a man

Write your "False" prompt:

A picture of a woman

USE MY PROMPTS: A picture of a man vs A picture of a woman

Fig.5: Game screen that allows the user to create his own prompt using 2 text inputs.

3. Use two prompts (See Fig. 5). In this case two text input are used to allow the player write two sentences. The player must use two opposite sentences, that is, with an opposite meaning.

Select a Winner picture name.

If you are inspired, Select a Winner image directly:

000015

CHECK THIS WINNER

Fig.6: Game screen that allows the user to select the winner image directly.

4. Select a winner (See Fig. 6). This option does not use the CLIP model to make decisions, the player can simply choose one of the images as the winner and if the player hits the winning image, the game is over.

Punctuation

To motivate the players in finding the winning image with the minimum number of questions, a scoring system is established so that it begins with a certain number of points (100 in the example) and decreases with each asked question. The score is decreased by subtracting the number of remaining images after each question. Furthermore, there are two extra penalties. The first is applied when the player uses the option “Select a winner”. This penalty depends on the number of remaining images, so that the fewer images are left, the bigger will be the penalty. Finally, the score is also decreased by two extra points if, after the player makes a question, no image can be discarded.

The “Guess Who?” game has a handicap when it uses real images, because it is necessary to always ensure that the same criteria are applied when the images are discarded. The original game uses images with characters that present simple and limited features like a short set of different types of hair colors, what makes it very easy to answer true or false when a user asks for a specific hair color. However, with real images it is possible to doubt about if a person is blond haired or brown haired, for example, and it is necessary to apply a method which ensures that the winning image is not discarded by mistake. To solve this problem, CLIP is used to discard the images that do not coincide with the winner image after each prompt. In this way, when the user asks a question, CLIP is used to classify the images in two groups: the set of images that continue because they have the same prediction than the winning image, and the discarded set that has the opposite prediction. Fig. 7 shows the screen that is prompted after calling CLIP on each image in the game board, where the discarded images are highlighted in red and the others in green.

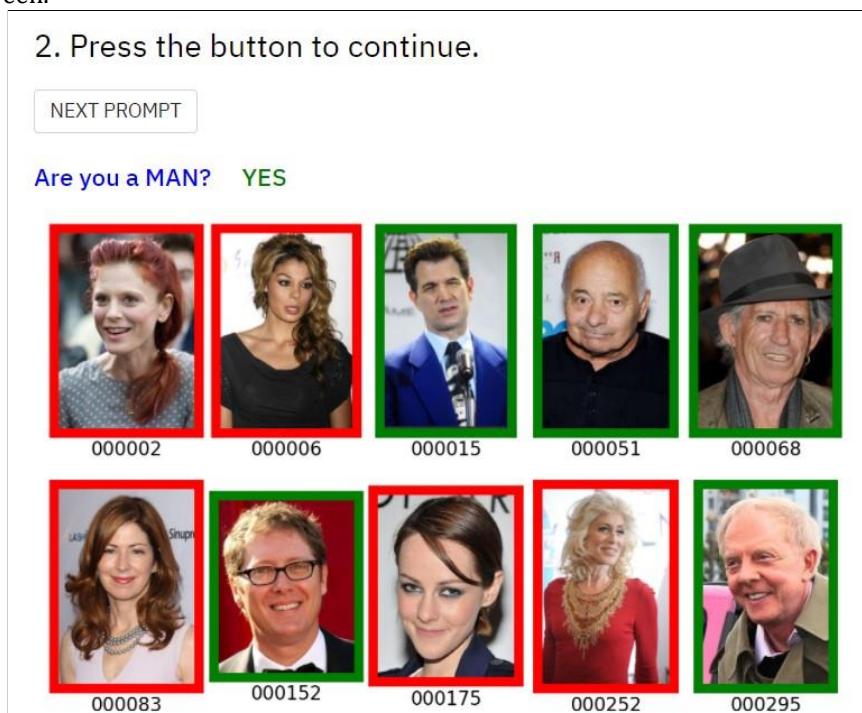


Fig.7: Game screen showing CLIP answer after a question.

Experiments and prompt analysis

To use CLIP as a zero-shot image classifier in the game, we create a pair of textual prompts for each class to address each player question as its own binary classification problem. Two basic prompting methods are proposed to create the textual descriptions:

1. **Target vs neutral.** This method consists in using a standard neutral prompt that fulfills all images, like “A picture of a person”, and another sentence, very similar, which changes only some words and is more specific for the target class, like “A picture of a person with eyeglasses”. In this way, when the additional information is added to the prompt, CLIP should return a bigger value for the second sentence than for the first for an image of a person with eyeglasses. And vice versa, i.e., when the extra information is not related to the image, CLIP should return a smaller likelihood value for this sentence. This is the method used to allow the user to introduce his own prompt.
2. **Target vs contrary.** This method consists in using two opposite sentences that represent opposite concepts, like “A picture of a man” and “A picture of a woman”. This method is only implemented in the game for the attributes included in the set of labels of the Celeba data set.

We take advantage of the labeled images from the Celeba data set to validate the performance of the textual prompts introduced to CLIP. The True Positive Rate (TPR), True Negative Rate (TNR), and the accuracy (average of TPR and TNR) are calculated to analyze the results. We used the first 4.000 true images and the first 4.000 false images for each of the 40 binary labels of the data set to calculate these rates.

Results table for the “Target vs neutral” method

Celeba label	Target prompt	TPR	TNR	Acc
male	A picture of a male person	98.14		
		96.13		
		97.11		
wearing hat	A picture of a person with hat	97.29		
		83.67		
		89.34		
goatee	A picture of a person with goatee	91.6		
		77.05		
		82.78		
blond hair	A picture of a person with blond hair	74.14		
		97.84		
		82.09		
bangs	A picture of a person with bangs	88	77.71	
		82.05		
eyeglasses	A picture of a person with eyeglasses	87.45		
		77.59		
		81.78		

smiling	A picture of a person who is smiling	89.07	
		76.75	
		81.76	
bald	A picture of a bald person	96.63	
		73.86	
		81.58	
wearing necktie	A picture of a person with necktie	77.91	
		81.36	
		79.54	
gray hair	A picture of a person with gray hair	83.66	
		74.04	
		78.05	
...
big lips	A picture of a person with big lips	64.64	...
		51.75	
		53.12	
wearing lipstick	A picture of a person with lipstick	85.34	
		51.53	
		52.94	
pointy nose	A picture of a person with pointy nose	52.74	
		52.16	
		52.41	
big nose	A picture of a person with big nose	57.63	
		51.09	
		51.91	
attractive	A picture of an attractive person	54.34	
		50.88	
		51.46	
rosy cheeks	A picture of a person with rosy cheeks	49.36	
		49.76	
		49.65	
high cheekbones	A picture of a person with high cheekbones	47.33	49.23
bags under eyes	A picture of a person with bags under eyes	47.7	49.05
narrow eyes	A picture of a person with narrow eyes	48.66	
no beard	A picture of a person with no beard	40.67	45.36
		43.8	16.93
			30.02
			25.09

Table 1: “Target vs neutral” prompting method applied on Celeba data set. True Positive Rate, True Negative Rate and Accuracy are shown in percentage.

Table 1 shows the top ten and bottom ten results sorted by accuracy, as well as the labels of the Celeba data set and the CLIP textual inputs used. In this experiment, we simply use the literal Celeba labels to create the target prompt, and the neutral prompt is kept as “A picture of a person”. With all this, approximately 25% of the target prompts obtained an accuracy above 70%. The ‘male’ and the ‘wearing hat’ features obtained remarkable accuracy results, 97% and 89% respectively.

In general, CLIP works sufficiently well when the label represents a physical object (e.g., hat or eyeglasses) or a common expression (e.g., smile), which will arguably be common in the data set in which CLIP has been trained on. The CLIP data set has millions of images and natural text from the Internet, so we must think about how the image descriptions in Internet often look like, in order to engineer good

prompts. For example, when talking about a specific person appearing in an image that contains several people, it is common to talk about “the one who wears a hat” or “the one who has a goatee”, but is unlikely to use a description like “the one who has narrow eyes” or “the one who has rosy cheeks”.

Another key for CLIP performance lies in the ambiguity of elements or concepts. Some labels present no doubt about whether they are true or false, but some others are susceptible to observer interpretation. For instance, asking if a person wears a hat is a very objective concept that raises no doubt, so practically everyone would respond the same. However, asking if a person has big lips, has a pointy nose, or is attractive, are relative or subjective questions whose answer will depend on the observer. In these cases, it seems that CLIP does not work so well.

Finally, a remarkable result in this experiment is the accuracy obtained with the feature “no beard”, which is related to a negation. In this case, the accuracy is 25%, but in a binary classification such as this, what CLIP really has reached is the 75 % of accuracy, if we invert its response. That means that CLIP ignored the “no” word in the prompt and classified the images of a person with beard with a 75% of accuracy. That indicates that CLIP probably is not able to deal with the concept of negation.

Results for the “Target vs contrary” method

Target prompt	Contrary prompt	TPR	TNR	Acc	Gain
A picture of a man	A picture of a woman	99.39	97.72	98.54	+1.43
A picture of a bald person	A picture of a haired person	96.08	80.43	86.65	+5.07
A picture of a person who is smiling	A picture of a person who is serious	80.05	89.88	84.28	+2.52
A picture of a person with pale skin	A picture of a person with tanned skin	68.12	75.8	71.29	+3.35
A picture of a young person	A picture of an aged person	65.17	78.2	69.72	+5.59
A picture of a person with straight hair	A picture of a person with wavy hair	60.38	67.04	62.9	+6.96
A picture of an attractive person	A picture of an unattractive person	50.44	50.13	50.2	-1.26

Table 2: “Target vs contrary” prompting method applied on Celeba data set. True Positive Rate, True Negative Rate, Accuracy and Gain are shown in percentage.

Table 2 shows some examples of the performance of the second proposed prompting method. The results of the first method, “target vs neutral” are also shown for easy comparison. This experiment shows how, in general, this method allows to improve the classification results by prompting “contrary words” to the target attribute, since. Even the “male” attribute, which obtained a 97 % of accuracy in the first experiment, can be improved with this method, reaching almost the 1,5% of accuracy improvement.

However, it must be remarked that it is difficult to find useful “contrary words”. As an example, the label “wearing a necktie”, “wearing lipstick” or “having rosy cheeks” only can be inverted using a negation, what does not work properly with CLIP. In cases where the antonym is very similar to the original word as “attractive” and “unattractive”, CLIP does not seem to work properly either, but more tests must be done to ensure it.

Conclusions and Future work

In this work, we present an implementation of the popular game “Guess Who?”, in which the part of the game engine that decides whether an image fulfills the player question or not is made by CLIP, a language-image model that estimates how good a text caption pairs with a given image. To do that, we take each player prompt describing a person attribute and confront it with another prompt that does not represent it. In this way, we can create a binary classifier for each attribute of interest just by getting the maximum likelihood of CLIP’s output to those two prompts. We have tried different prompting methods, such as confronting the target prompt with a neutral one (e.g., “A picture of a person”) or using a prompt that describes the contrary of the target attribute. Experiments have been made with the labelled images of the Celeba data set to analyze the performance of these two prompting methods, showing that, as long as there is a clear contrary word to the target attribute, using the contrary prompt method normally leads to a better zero-shot classification performance.

As future work, we will continue working on “prompt engineering” due to CLIP-based zero-shot classifiers can be sensitive to wording or phrasing. Moreover, we will deploy the game as a public web application where interested users can play with their own images and give feedback about its performance.

Acknowledgements

This work has been partially supported by the company Dimai S.L, the “Convenio Plurianual” with the Universidad Politécnica de Madrid in the actuation line of “Programa de Excelencia para el Profesorado Universitario” and by next research projects: FightDIS (PID2020-117263GB-I00), IBERIFIER (2020-EU-IA0252:29374659), and the CIVIC project (BBVA Foundation Grants For Scientific Research Teams SARS-CoV-2 and COVID-19).

References

1. Conde, M.V., Turgutlu, K.: Clip-art: Contrastive pre-training for fine-grained artclassification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 3956–3960 (June 2021)
2. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10578–10587 (2020)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. Fang, H., Xiong, P., Xu, L., Chen, Y.: Clip2video: Mastering video-text retrieval via image clip. ArXiv **abs/2106.11097** (2021)
5. Gao, J., Li, P., Chen, Z., Zhang, J.: A survey on deep learning for multimodal data fusion. *Neural Computation* **32**, 829–864 (2020)
6. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. pp. 3730–3738 (2015)
7. Nawaz, S., Calefati, A., Caraffini, M., Landro, N., Gallo, I.: Are these birds similar: Learning branched networks for fine-grained representations. In: 2019 International Conference on Image and Vision Computing New Zealand (IVCNZ). pp. 1–5 (2019). <https://doi.org/10.1109/IVCNZ48456.2019.8960960>
8. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Aspell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)

9. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. arXiv preprint arXiv:2102.12092 (2021)
10. Tan, H., Yu, L., Bansal, M.: Learning to navigate unseen environments: Backtranslation with environmental dropout. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 2610–2621 (2019)
11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
12. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. arXiv preprint arXiv:2010.00747 (2020)