

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



**Universidad**  
Internacional  
de Valencia

Trabajo de Fin de Máster

**Estudiante:** Arnau Martí Sarri

**Tutor:** Víctor Rodríguez Fernández

De:  
 Planeta Formación y Universidades

# ÍNDICE

Introducción

Objetivos

Estado del Arte

Metodología

Resultados

Conclusiones



# INTRODUCCIÓN

## ÍNDICE

Introducción

Objetivos

Estado del Arte

Metodología

Resultados

Conclusiones

# INTRODUCCIÓN

El proyecto consiste en crear una aplicación web del juego ¿Quién es Quién?

Se basa en que las preguntas las responda un sistema con Inteligencia Artificial

Se utilizan los recursos adquiridos durante el Máster





# OBJETIVOS

## ÍNDICE

Introducción

Objetivos

Estado del Arte

Metodología

Resultados

Conclusiones

## OBJETIVOS

Dar libertad al jugador para realizar las preguntas que se le ocurran

Permitir al jugador usar sus propias imágenes

Clasificación de imágenes a partir de un texto introducido

Comparación con otros algoritmos

# ÍNDICE

Introducción

Objetivos

Estado del Arte

Metodología

Resultados

Conclusiones

# ESTADO DEL ARTE

## Problemas a encarar:

### Clasificación de imágenes de caras de personas

- Para realizar muchas preguntas, deberemos entrenar muchos modelos?
- Hay **Datasets** disponibles para entrenar o verificar?

### Clasificación de imágenes “Zero-Shot”

- Sistema “Few-Shot” con soporte textual, razonablemente rápido?

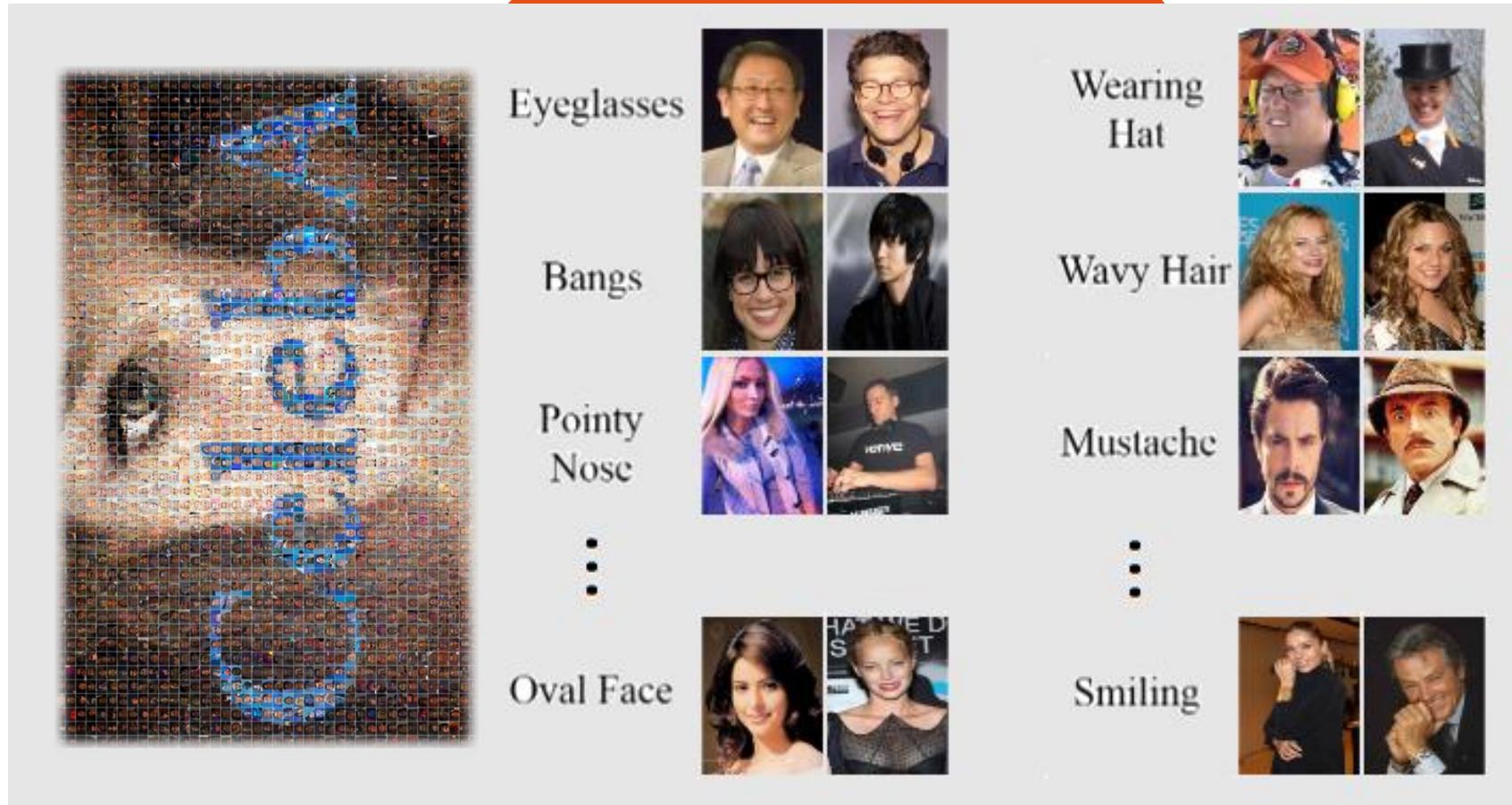
\* Crear App -> streamlit (Python)





202.599 imágenes

40 etiquetas



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

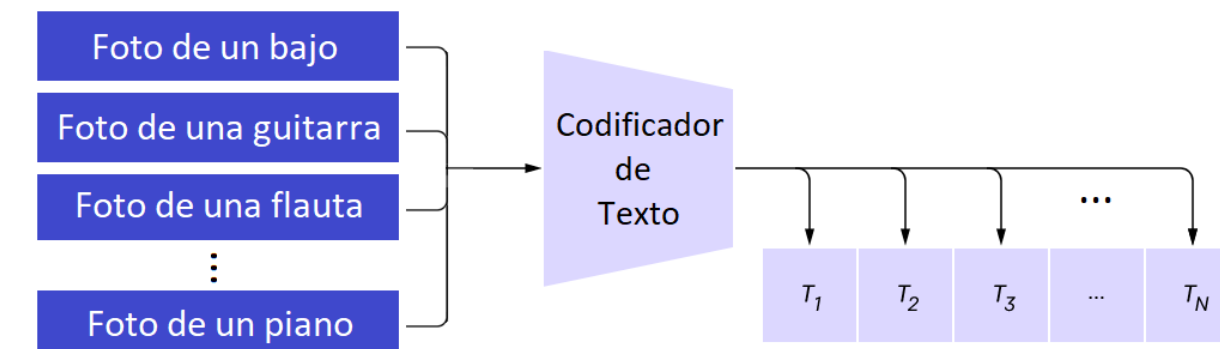
## ESTADO DEL ARTE

# Clasificar imágenes mediante CLIP

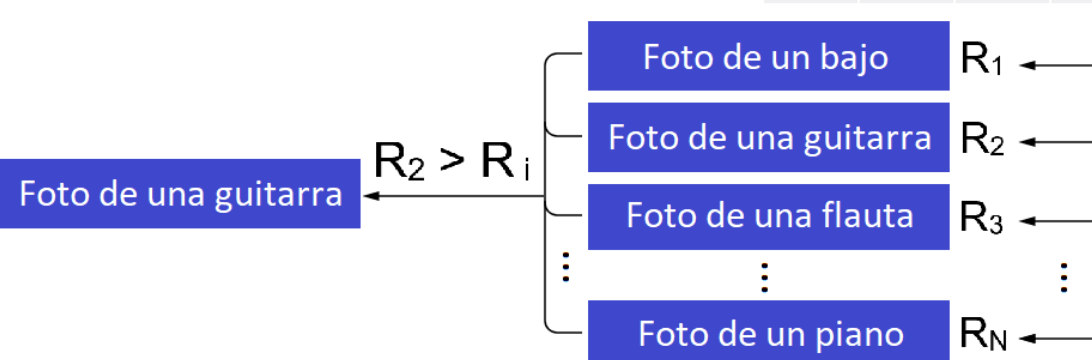
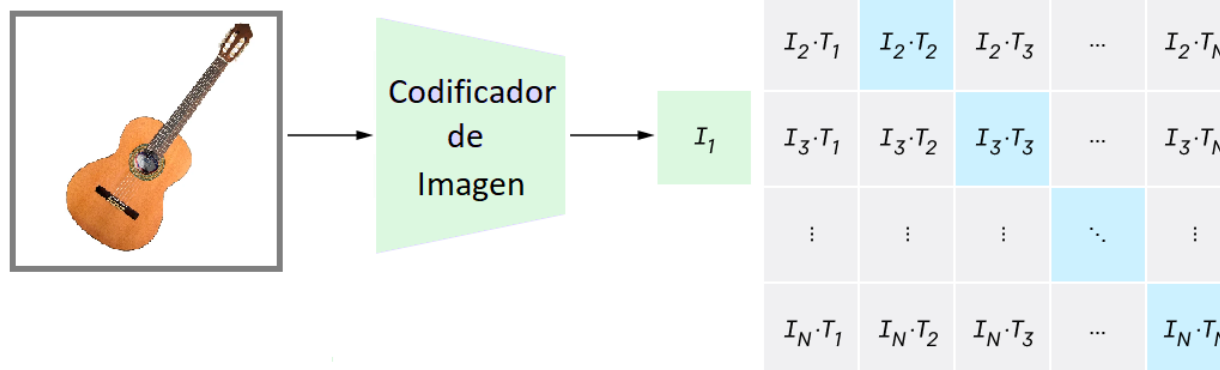
(Contrastive Language-Image Pretraining)

-> DNN entrenada con imágenes y texto en lenguaje natural obtenidos de internet

### Lista de textos ("prompts") para diferenciar las clases



### Imagen a clasificar ("zero-shot")



# ÍNDICE

Introducción

Objetivos

Estado del Arte

Metodología

Resultados

Conclusiones

# METODOLOGÍA

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

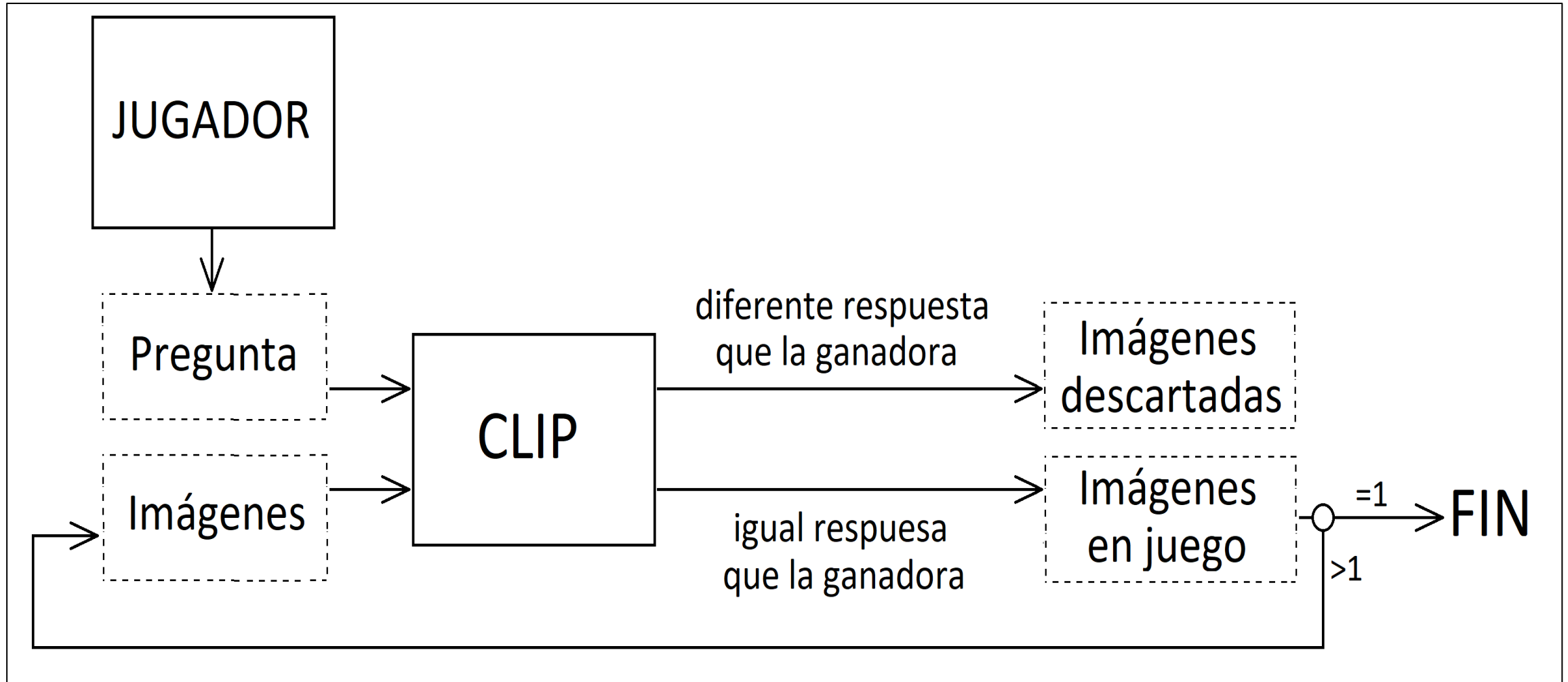
## Herramientas:

Dataset -> conjunto de datos “Celeba”

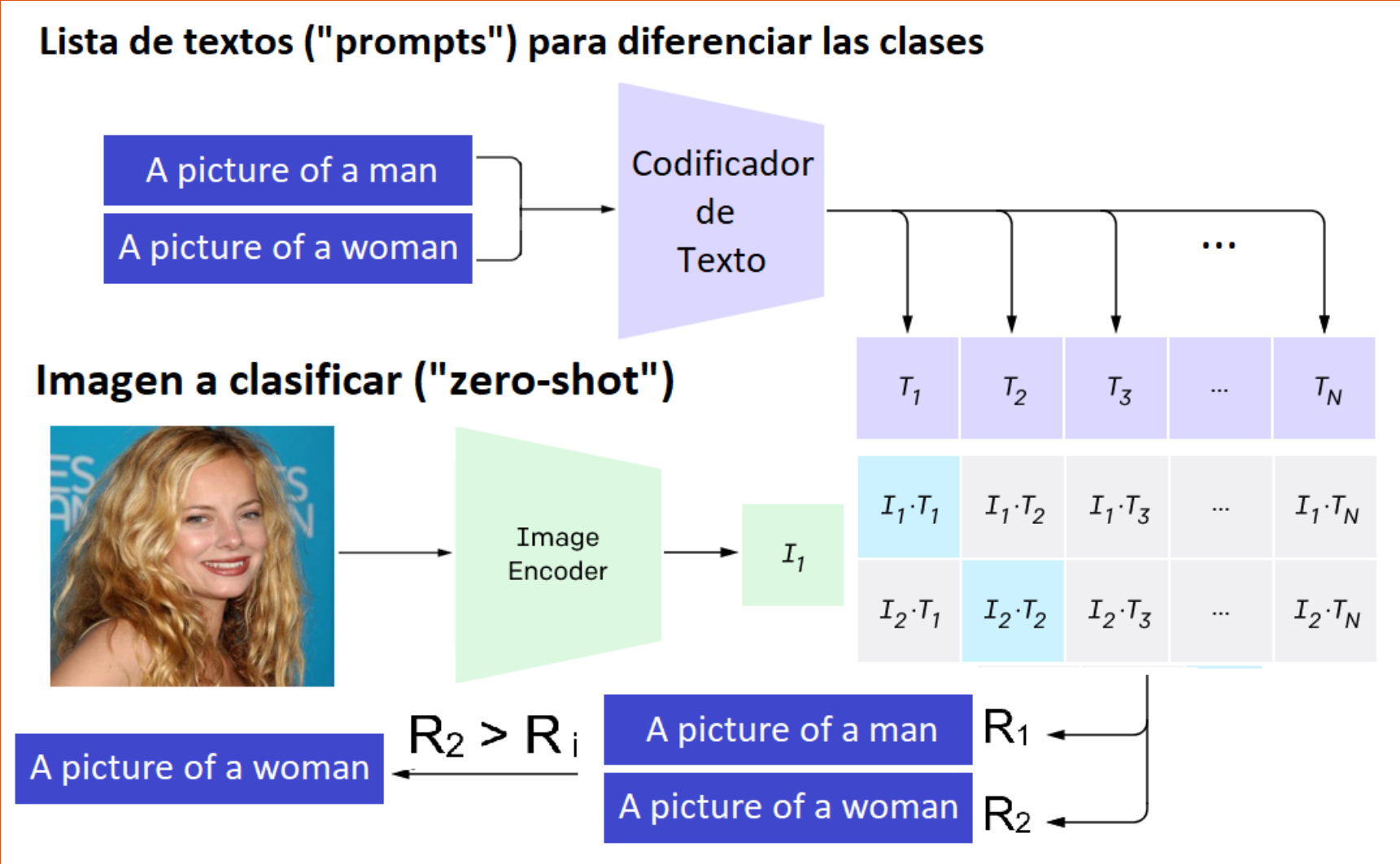
App -> streamlit (Python)

Responder preguntas -> CLIP

## Integración de CLIP en el juego

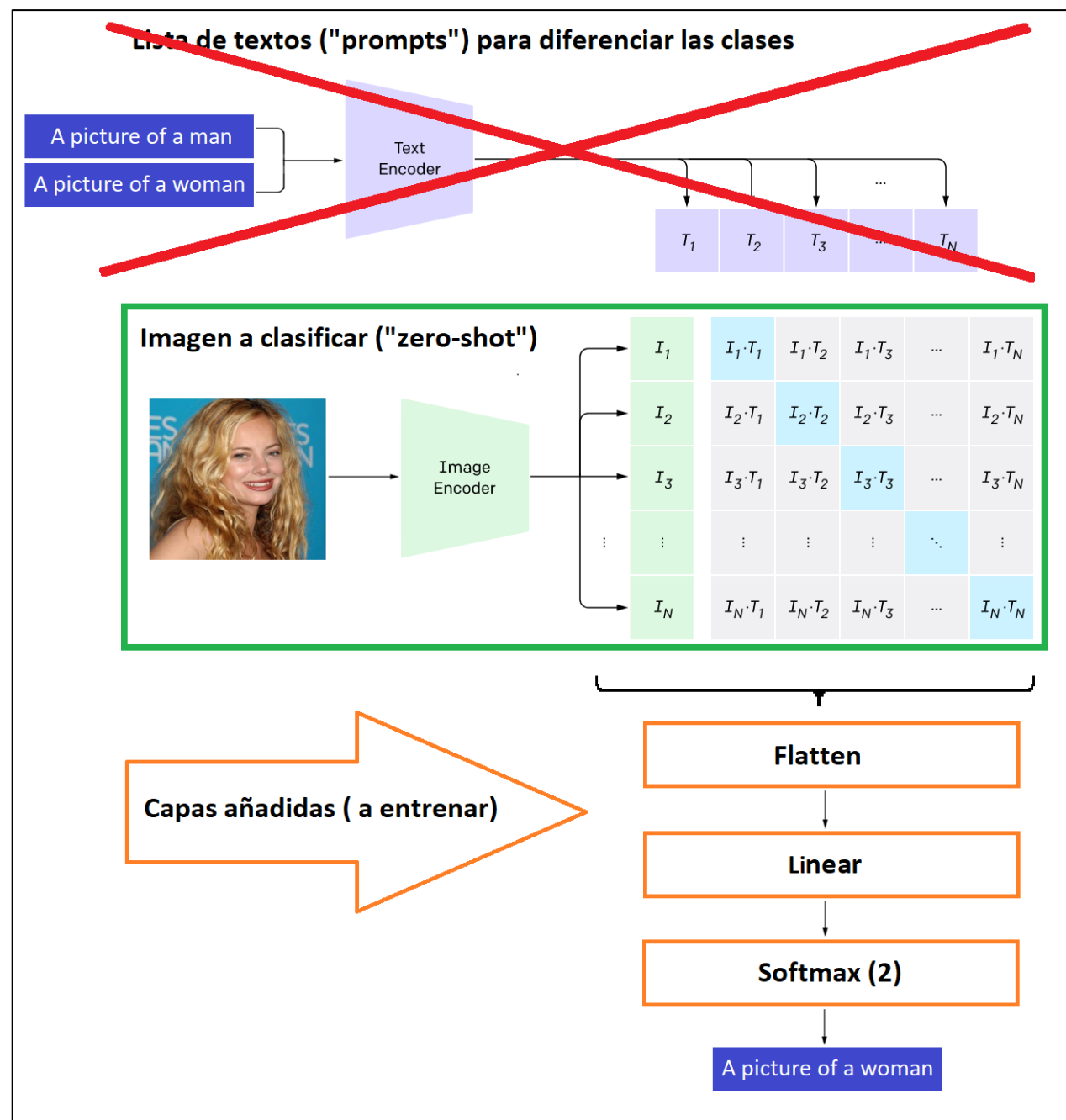


## Adaptación de CLIP para responder a una pregunta de “Si” o “No”



# METODOLOGÍA

“Fine Tuning”  
de CLIP para  
responder a  
una pregunta  
de “Si” o “No”



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

# ÍNDICE

Introducción

Objetivos

Estado del Arte

Metodología

Resultados

Conclusiones

# RESULTADOS



## RESULTADOS

# APLICACIÓN

## Pantallas de selección de imágenes

×

OPTIONS PANEL

RESET GAME

Number of images

Select the number of images of the game and press "RESET GAME"

20

–

+

Image selection source:

(Choose between default random images or specific source path)

Use images from specific path

Use default random images

Use images from specific path

Guess Who?

score: 100


1. Choose the images you like.

Press the button to randomly modify the selected images.


CHANGE IMAGES

2. Press the button to start playing.


START GAME




000002




000006




000015




000051




000068




000083




000152



000175



000252



000295

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

## RESULTADOS

# APLICACIÓN

Pantallas para que el jugador introduzca la pregunta a realizar

Select a Winner picture name.

If you are inspired, Select a Winner image directly:

000015

CHECK THIS WINNER

Select a Question from the list.

Suggested questions:

Are you a MAN?

Current Question: Are you a MAN?

USE THIS QUESTION

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

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

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

## RESULTADOS

# APLICACIÓN

## Pantallas para mostrar los resultados



# Selección del “prompt” de referencia:

"Prompts" de referencia	"Prompts" descriptivos	RVP	RVN	PT
A picture of a person	A picture of a female person	98,96	97,33	<b>98,12</b>
A photo of a person	A photo of a female person	97,53	97,81	<b>97,68</b>
<u>A</u> image of a person	<u>A</u> image of a female person	97,32	97,99	<b>97,65</b>
A picture of a person	A picture of a male person	98,14	96,18	<b>97,14</b>
A snapshot of a person	A snapshot of a female person	99,02	93,38	<b>96,02</b>
...	...	...	...	...
A person	A Male	95,25	66,85	<b>74,56</b>
A snapshot of a person	A snapshot of a person who is a female	94,67	66,55	<b>74,15</b>
Person	Male	93,88	60,7	<b>67,2</b>
Person	Female	52,59	93,91	<b>54,9</b>
A photo of a person	A photo of a person who is male	54,46	55,22	<b>54,81</b>

## Estrategia con el “prompt” de referencia:

Etiquetas “Celeba”	"Prompts" descriptivos	RVP	RVN	PT
male	A picture of a male person	98,14	96,13	<b>97,11</b>
bald	A picture of a bald person	96,63	73,86	<b>81,58</b>
wearing hat	A picture of a person with hat	97,29	83,67	<b>89,34</b>
goatee	A picture of a person with goatee	91,6	77,05	<b>82,78</b>
blond hair	A picture of a person with blond hair	74,14	97,84	<b>82,09</b>
bangs	A picture of a person with bangs	88	77,71	<b>82,05</b>
eyeglasses	A picture of a person with eyeglasses	87,45	77,59	<b>81,78</b>
Smiling	A picture of a person who is smiling	89,07	76,75	<b>81,76</b>
wearing necktie	A picture of a person with necktie	77,91	81,36	<b>79,54</b>
gray hair	A picture of a person with gray hair	83,66	74,04	<b>78,05</b>
black hair	A picture of a person with black hair	69,6	81,88	<b>74,28</b>
wavy hair	A picture of a person with wavy hair	72,71	73,94	<b>73,31</b>
no beard	A picture of a person with no beard	16,93	30,02	<b>25,09</b>



## Estrategia con el “prompt” opuesto:

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	<b>1,59</b>
Wearing Hat	A picture of a person with hat	A picture of a person with hair	92,2	89,54	90,82	<b>1,48</b>
Bald	A picture of a bald person	A picture of a haired person	96,08	80,43	86,65	<b>5,07</b>
Smiling	A picture of a person who is smiling	A picture of a person who is serious	80,05	89,88	84,28	<b>2,52</b>
Pale Skin	A picture of a person with pale skin	A picture of a person with tanned skin	68,12	75,8	71,29	<b>3,35</b>
Wearing Necktie	A picture of a person with necktie	A picture of a person without necktie	69,12	55,01	57,94	<b>-21,6</b>
Mustache	A picture of a person with mustache	A picture of a person without mustache	74,07	53,33	55,85	<b>-12,04</b>
No Beard	A picture of a person with no beard	A picture of a person with beard	52,43	74,87	54,44	<b>-20,47</b>
Goatee	A picture of a person with goatee	A picture of a person without goatee	50,1	50,08	50,09	<b>-32,69</b>
Wearing Necklace	A picture of a person with necklace	A picture of a person without necklace	31,77	46,53	44,18	<b>-18,46</b>

# Estrategia con múltiples “prompts”:

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

Pruebas con la estrategia inicial usando el “prompt” de referencia

Etiquetas Celeba	“Prompts” descriptivos	RVP	RVN	PT	PT media
Black Hair	A picture of a black-haired person	90,42	78,97	87,88	86,48 (Mejora: +17.23)
Blond Hair	A picture of a blond-haired person	88,36	88,79	88,44	
Brown Hair	A picture of a tawny-haired person	82,06	77,99	81,58	
Gray Hair	A picture of a gray-haired person	98,69	68,41	88	

Pruebas modificando los “prompts”

## “Fine Tuning” con CLIP, VGG19 y Xception:

Etiquetas “Celeba”	Red	RVP	RVN	PT	Mejora
bald	CLIP	97,25	95,97	96,6	<b>+15,02</b>
	VGG19	95,52	93,93	94,71	<b>+13,13</b>
	Xception	93,26	91,78	92,51	<b>+10,93</b>
big nose	CLIP	76,72	76,62	76,67	<b>+24,76</b>
	VGG19	69,86	68,75	69,25	<b>+17,34</b>
	Xception	68,31	68,23	68,27	<b>+16,36</b>
high cheekbones	CLIP	83,21	83,64	83,42	<b>+32,22</b>
	VGG19	83,34	76,49	79,52	<b>+28,32</b>
	Xception	71,83	71,22	71,52	<b>+20,32</b>
male	CLIP	98,48	99,11	98,8	<b>+1,69</b>
	VGG19	93,87	92,79	92,7	<b>-4,41</b>
	Xception	91,2	88,45	89,77	<b>-7,34</b>

Etiquetas “Celeba”	Red	RVP	RVN	PT	Mejora
rosy cheeks	CLIP	84,76	88,92	86,72	<b>+36,37</b>
	VGG19	84,5	85,43	84,96	<b>+34,61</b>
	Xception	78,56	85,07	81,49	<b>+31,14</b>
smiling	CLIP	88,95	88,49	88,72	<b>+6,96</b>
	VGG19	81,83	84,06	82,91	<b>+1,15</b>
	Xception	71,26	70,48	70,86	<b>-10,9</b>
young	CLIP	85,29	83,72	84,49	<b>+21,68</b>
	VGG19	75,2	75,93	75,56	<b>+12,75</b>
	Xception	73,56	73,05	73,06	<b>+10,25</b>
wearing hat	CLIP	96,79	96,63	96,71	<b>+7,37</b>
	VGG19	95,37	94,36	94,86	<b>+5,52</b>
	Xception	91,22	93,29	92,22	<b>+2,88</b>



# ÍNDICE

Introducción

Objetivos

Estado del Arte

Metodología

Resultados

Conclusiones

# CONCLUSIONES

### Objetivos cumplidos:

Se ha creado una aplicación que respeta la filosofía del juego, aunque para un solo jugador

“Streamlit” ha resultado ser una buena opción para crear una “App web” con Python

Se ha mostrado el potencial de CLIP como clasificador de imágenes “Zero-Shot”, mediante descripciones con lenguaje natural

Para los casos mas complejos y para la optimización de resultados con CLIP es necesario realizar “prompt engineering”

## CONCLUSIONES

# “Prompt engineering” con CLIP:

## Clasificación binaria

- Referencia vs Descripción: buena estrategia para un estudio previo
- Frases opuestas sin usar negación suele dar buenos resultados
- Para preguntas binarias con características multimodales, es buena opción usar más de 2 “prompts” (uno por clase)

## “Fine Tuning” de CLIP

- No permite introducir preguntas nuevas, pero se obtienen muy buenos resultados (mejores que VGG19 y Xception)

### Propuestas de futuro:

Con el conjunto de datos “Celeba”, hacer más “prompt engineering” hasta lograr resultados razonablemente buenos con las 40 etiquetas

Profundizar en el estudio de las etiquetas que se pueden combinar como “No beard/5 o’clock shadow” o “wavy hair/straight hair”

Con el Crear la versión del juego para dos jugadores

Estudiar el lenguaje de Internet para mejorar los “prompts” de CLIP



# CONCLUSIONES

## Guess Who?

# Vídeo de ejemplo:

score: 100

1. Choose the images you like.

Press the button to randomly modify the selected images.

CHANGE IMAGES

2. Press the button to start the game.

START GAME



000557



000975



001993



002776



002267



ENLACE DEL VIDEO: [https://github.com/ArnauDIMAI/CLIP-GuessWho/blob/main/Video\\_ejemplo\\_QesQ.mp4](https://github.com/ArnauDIMAI/CLIP-GuessWho/blob/main/Video_ejemplo_QesQ.mp4)

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

**Enlace a la aplicación web:**

[https://share.streamlit.io/arnaudimai/clip-guesswho/main/app\\_qesq.py](https://share.streamlit.io/arnaudimai/clip-guesswho/main/app_qesq.py)

# Gracias por vuestra atención

# BIBLIOGRAFIA

- Zero-data Learning of New Tasks. <https://www.aaai.org/Papers/AAAI/2008/AAAI08-103.pdf>
- Predicting Visual Exemplars of Unseen Classes for Zero-Shot Learning.  
<https://arxiv.org/abs/1605.08151>
- Clasificación de imágenes con redes neuronales profundas mediante conjuntos de entrenamiento reducidos y aprendizaje “Few-Shot”. <https://repositorio.uam.es/handle/10486/693963>
- CLIP: Connecting Text and Images. <https://openai.com/blog/clip/>,  
<https://arxiv.org/pdf/2103.00020v1.pdf>
- Large-scale CelebFaces Attributes (CelebA) Dataset. <http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html>