Descripción automática de imágenes

Jose Arias Moncho
Victoria Beltrán Domínguez

Índice

1. Introducción

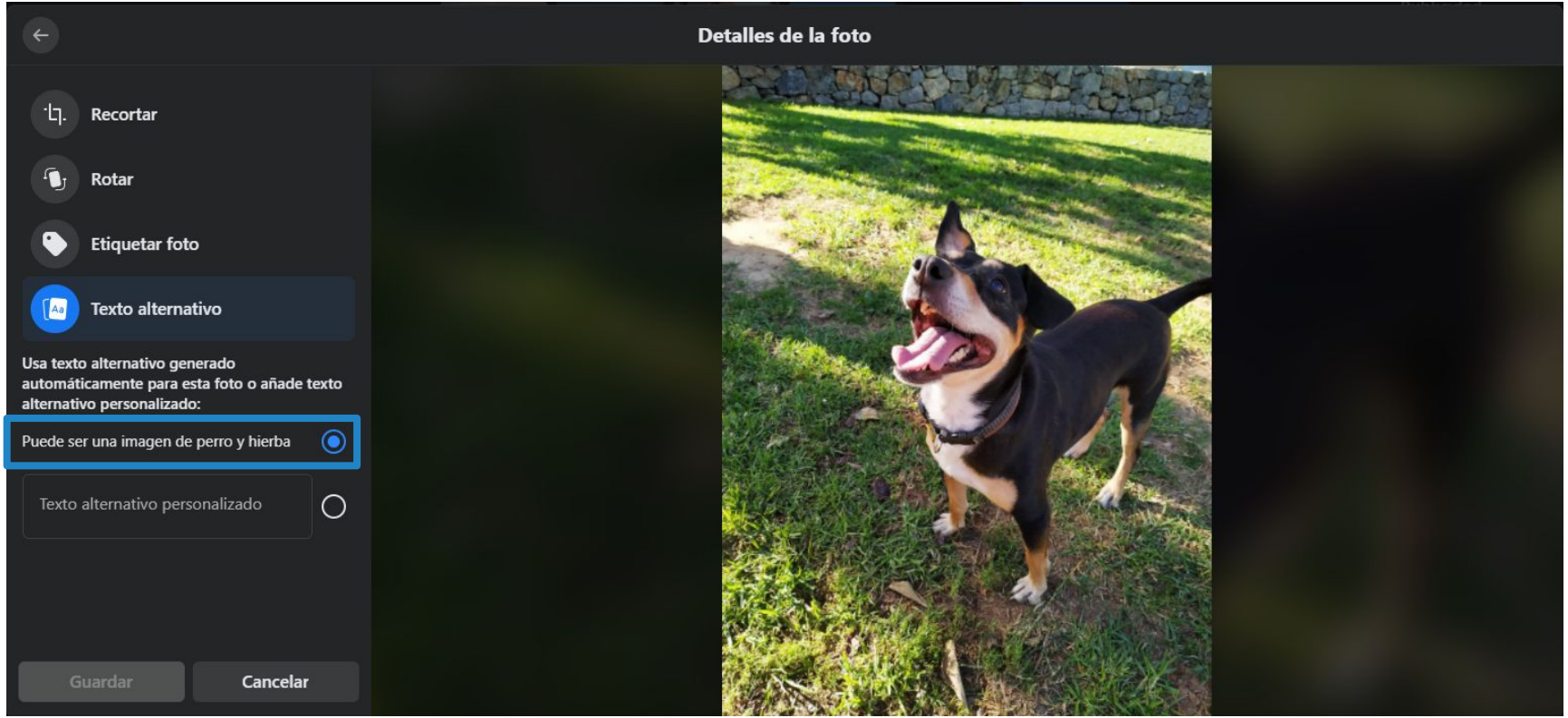
Describir el contenido de una imagen en palabras.

Necesitamos identificar qué hay (visión por computador) y saber expresarlo en lenguaje natural (procesamiento de lenguaje natural).

¿Para qué podría interesar?

- Anotaciones para personas invidentes
- Clasificación de imágenes para el comercio online
- Redes sociales
- Asistentes virtuales
- ...

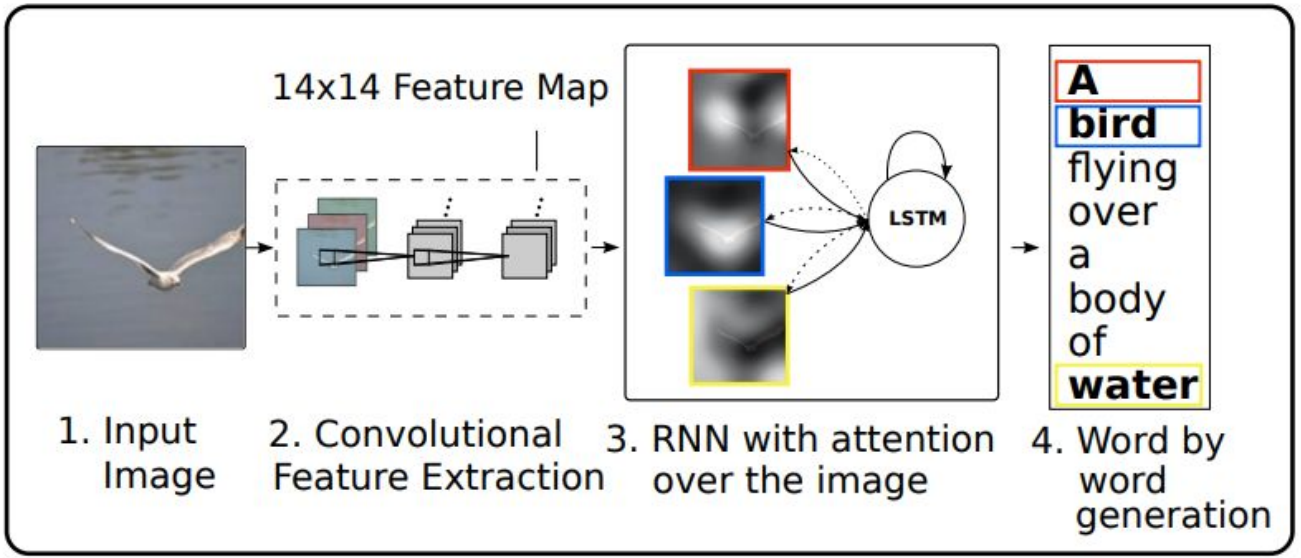
1. Introducción: Facebook's automatic Alt-Text tool



2. Estado del arte

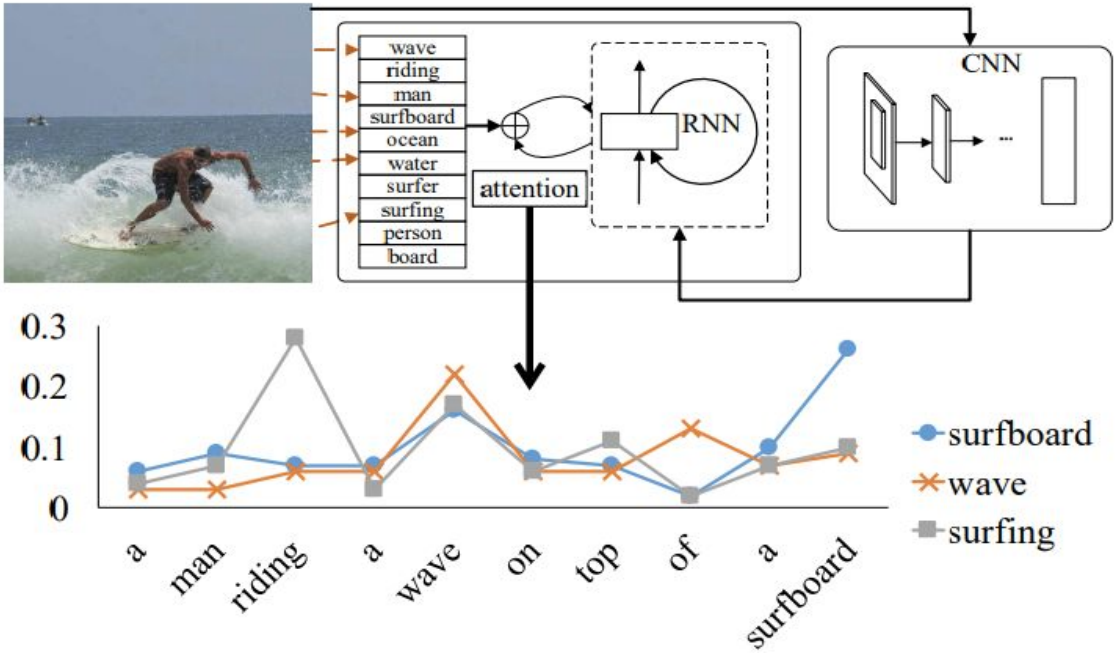
Existen diferentes enfoques que están en proceso de investigación:

1. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention (2015)



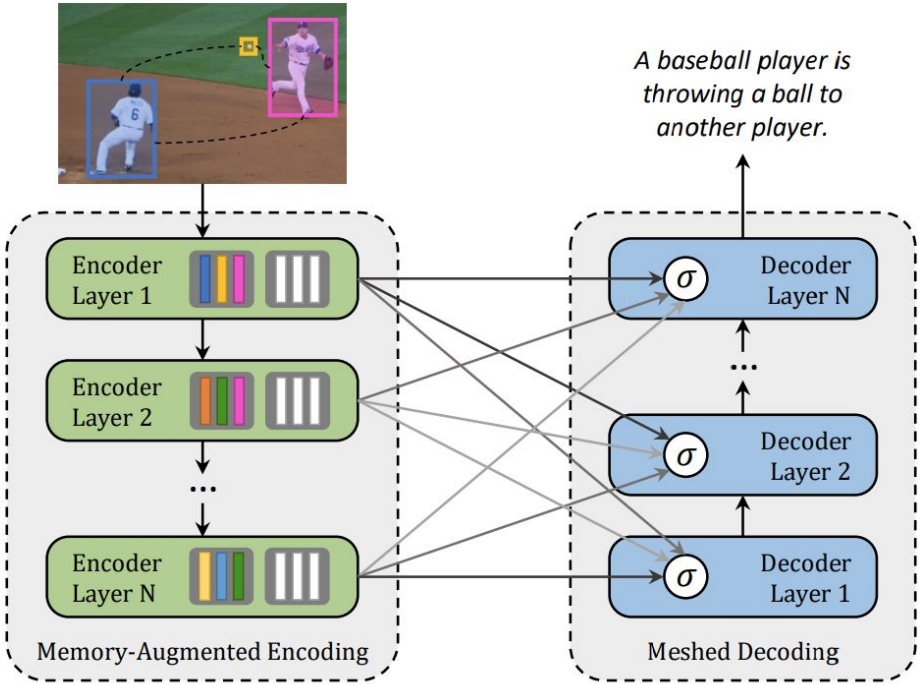
2. Estado del arte

2. Image Captioning with Semantic Attention (2016)



2. Estado del arte

3. Meshed-Memory Transformer for Image Captioning (2020)



3. Descripción de la tarea

Objetivo: dada una imagen, generar una descripción automática de esta.

Conjunto de datos: Flickr30k => contiene por cada imagen 5 descripciones en inglés de esta. Estándar para la tarea.

Tamaño del dataset: 31.014 imágenes con 5 descripciones cada una (155.070 descripciones).

Tamaño del vocabulario: 19.698 palabras únicas*

Tamaño de la secuencia más larga a predecir: 78 palabras

Imagen:

Descripción: A man wearing a helmet , red pants with white stripes going down the sides and a white and red shirt is on a small bicycle using only his hands while his legs are up in the air , while another man wearing a light blue shirt with dark blue trim and black pants with red stripes going up the sides is standing nearby , gesturing toward the first man and holding a small figurine of one of the seven dwarves .

4. Preproceso de datos

“A kid with 2 ice-creams is on the park”

Paso 1: pasarlo a minúscula: *“a kid with 2 ice-creams is on the park”*

Paso 2: limpiar cadena (!"#$%&'()*+, -./:;<=>?@[\]^_`{|}~):

“a kid with 2 icecreams is on the park”

Paso 3: filtrar valores que no sean palabras:

“a kid with icecreams is on the park”

Paso 4: añadir tokens de principio y fin en cada descripción:

“aaprincipioaa a kid with icecreams is on the park zzfinzz”

4. Preproceso de datos

Paso 5: calcular el vocabulario.

Paso 6: reducir el vocabulario por número de ocurrencias (mínimo 3).

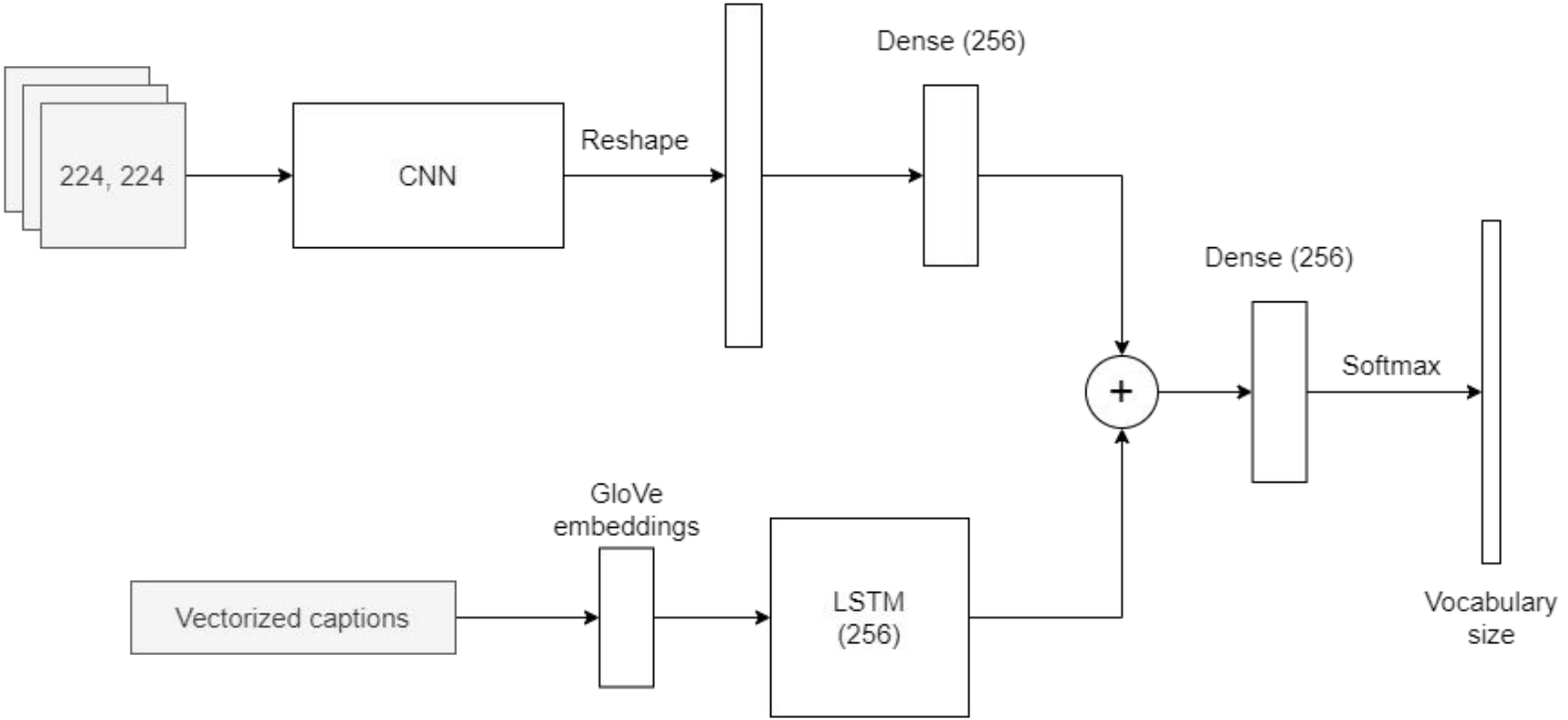
Paso 7: vectorizar con todo el vocabulario.

- TextVectorization
- Python mapping dictionary

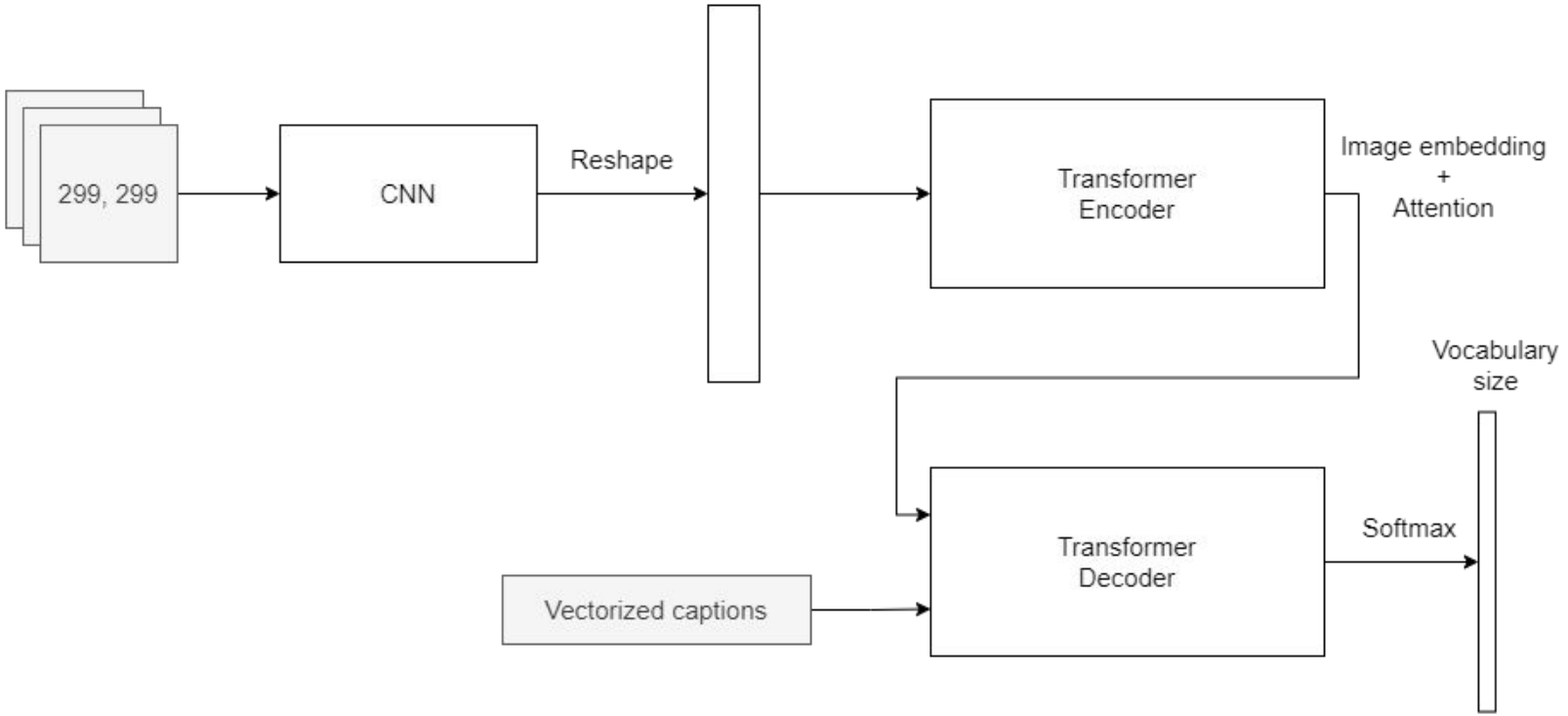
Paso 8: separar las imágenes utilizando las particiones de referencia para Flickr30k.

- Train: 29.000
- Val: 1.014
- Test: 1000 imágenes

5. Experimentación: LSTM



5. Experimentación: Transformer



6. Resultados obtenidos

- Medidas de precisión: BLEU (nltk.translate.bleu_score.corpus_bleu)

Referencias:

1. Un patito anda por la hierba
2. Un pato muy pequeño está de pie en la hierba

...

Predecido:

Un pato en la hierba

6. Resultados obtenidos

Modelo	Epochs	MAX TAMAÑO FRASE	TAMAÑO VOCAB	BLEU-3	BLEU-4
LSTM + GLOVE6B	60	80	9936	18.7	11.9
TRANSFORMER	60	80	9936	27.9	18.7
Show, Attend and Tell (2015)	-	-	-	29.6	19.9
Image Captioning with Semantic Attention (2016)	-	-	-	53.4	41.2

7. Discusión

LSTM

- Gran cantidad de recursos necesarios.
- Falta de mecanismos de atención.
- Falta de potencia.

TRANSFORMER

- Resultados decentes.
- Eficaz y rápido.
- Mecanismos de atención.

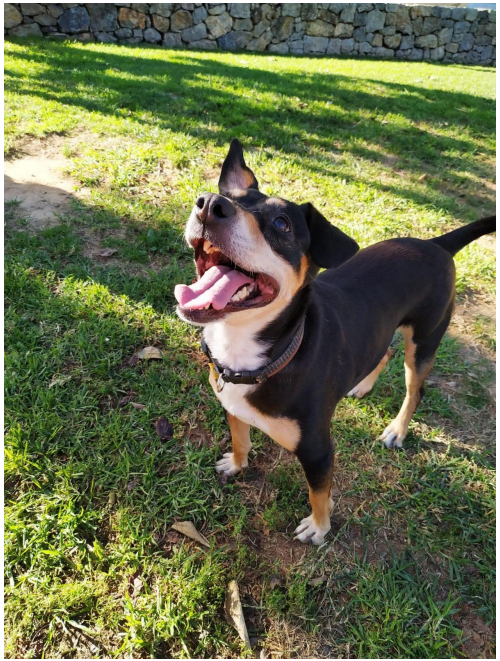
8. Conclusiones

- Image captioning: tarea multidominio.
- Flickr30k.
- Dos arquitecturas distintas: LSTM vs Transformer.
- Resultados decentes pero falta de recursos.

9. Trabajo futuro

- Utilizar features maps de las imágenes.
- Probar bottom-up approaches.
- Mecanismos de atención visual.

10. Demo:



Transformer prediction: a dog is running through a grassy area

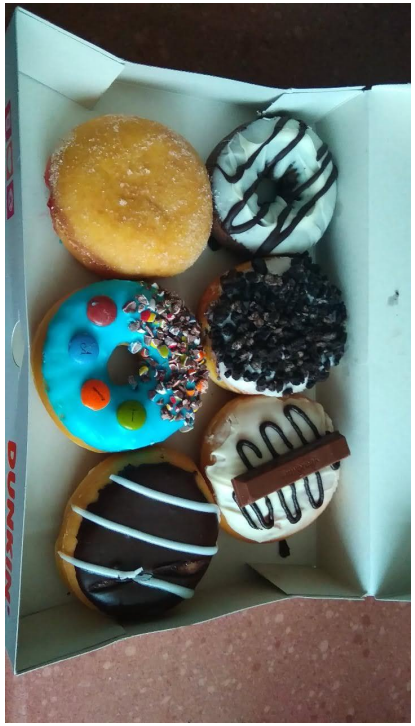
LSTM prediction: a man in a black shirt and jeans is plowing a lawn

10. Demo:

Transformer prediction: a group of people are standing in a room

LSTM prediction: a man in a red shirt is singing into a microphone

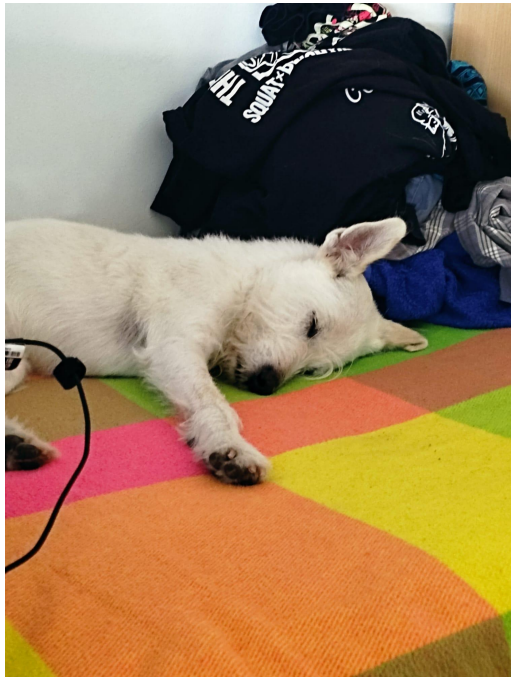
10. Demo:



Transformer prediction: a man is standing in front of a large white building

LSTM prediction: a man in a blue shirt is sitting on a bench

10. Demo:



Transformer prediction: a white dog is jumping over a red and white dog

LSTM prediction : a man in a black shirt is playing a guitar

¡Muchas gracias!

Jose Arias Moncho
Victoria Beltrán Domínguez