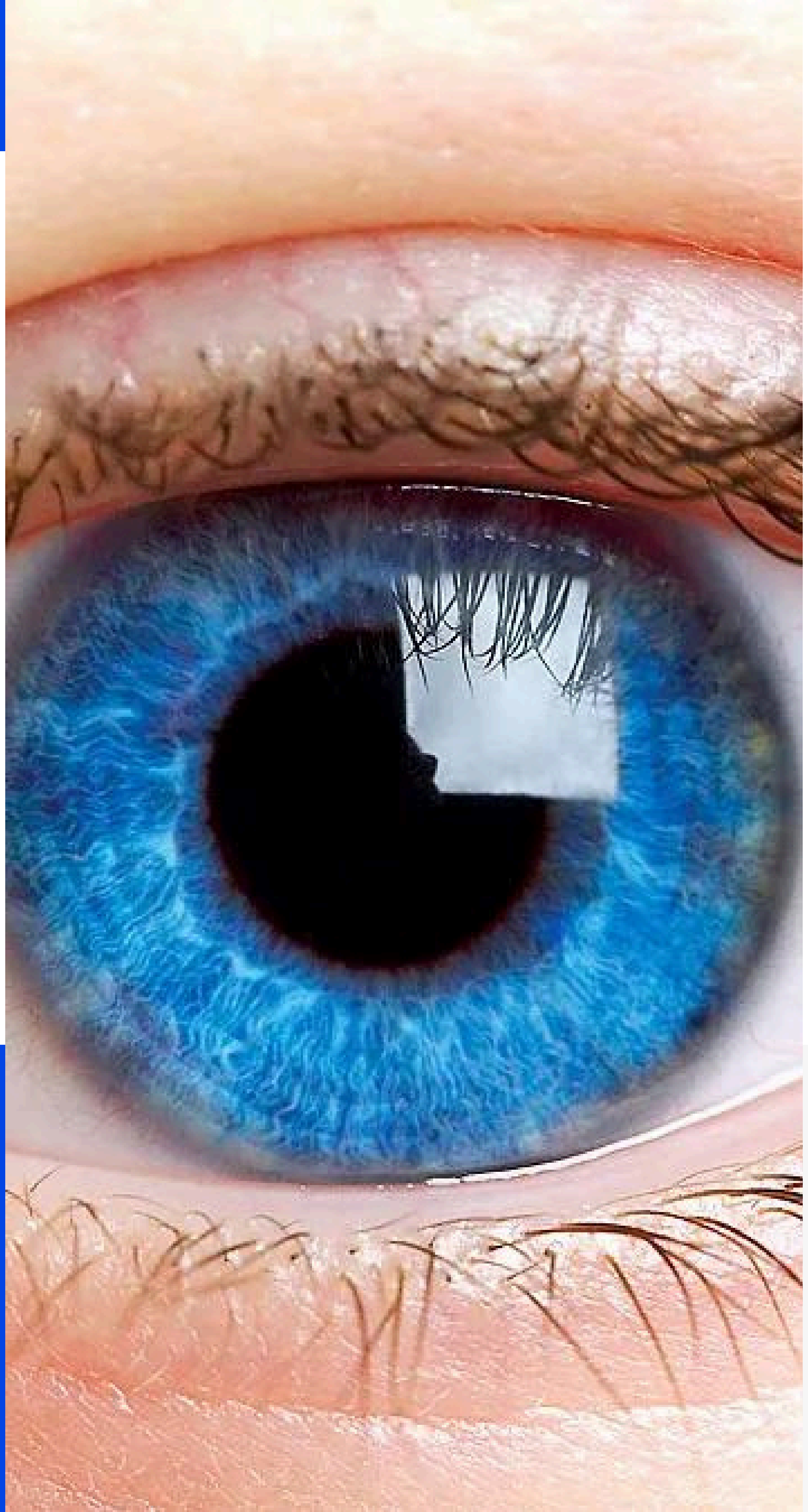
LucidLens

LUCID-LENS



LucidLens

CONTENT

01 Problem Statement and Why

02 Potential Application and Impact

03 Literature Survey

04 Dataset and Preprocessing

05 ML Methodology

06 Performance Evaluation

07 Challenges and References

PROBLEM

LucidLens

01 **Challenges in Object Identification:**

Visually impaired individuals face challenges in identifying objects and understanding their surroundings due to the lack of visual cues.

02 **Limitations of Traditional Methods:**

Traditional methods of object identification through touch or assistance from others are often time-consuming and not always feasible.

03 **The Need for Swift:**

There's a need for a solution that can swiftly and accurately identify objects and provide auditory feedback to visually impaired individuals.

WHY THIS????

LucidLens

Empowering Independence

This project aims to enhance independence and autonomy for visually impaired individuals by providing real-time auditory descriptions of their environment.

Fostering Inclusivity

By converting visual information into speech, we can bridge the gap between the sighted and visually impaired communities, fostering inclusivity and accessibility.

Enhancing Quality of Life

The implementation of this project can significantly improve the quality of life for visually impaired individuals, empowering them to navigate their surroundings with confidence and ease.

POTENTIAL APPLICATION AND IMPACT

LucidLens

For visually impaired

The model will describe the surroundings and convert it to speech. This will give them a sense of independence and dignity.

Guide for Visitors

We will collect the dataset of our university and then this model can be used to guide the visitors in the campus. This can give them information about specific locations and objects in the campus.

Automatic alternate Text Generation

Generate alt text for images on websites, improving SEO and accessibility. Alt text provides a brief description of an image for those who cannot see it.

LITERATURE SURVEY

Title: - Indoor object detection and recognition for an ICT mobility assistance of visually impaired people

Method: - YOLO v3, DarkNet-53,Flickr8k, 16 indoor classes

Cons: -73.19% accuracy, and it is only focused on indoor navigation. Used pretrained model and trained on the new dataset

Reference: - Afif, M., Ayachi, R., Said, Y., Pissaloux, E., Atri, M., 2020b. An evaluation of retinanet on indoor object detection for blind and visually impaired persons assistance navigation. Neural Processing Letters , 1–15.

LITERATURE SURVEY

LucidLens

Title: - Object detection and Narrator for Visually Impaired people

Method: - Used YOLO, trained on Imagenet dataset

Cons: Accuracy is 62.5% for normal phones and 75% for iphones and Samsung. The results are camera dependent.

Reference: - Nasreen, J., Arif, W., Shaikh, A.A., Muhammad, Y., Abdullah, M., 2019. Object detection and narrator for visually impaired people, in: 2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS), IEEE. pp. 1–4.

LITERATURE SURVEY

LucidLens

Title: - Building A Voice Based Image Caption Generator with Deep Learning

Method: - NLP ,CNN, LSTM (Long short term memory), RNN (recurrent neural network) flicker8k dataset

Cons: Accuracy is 90% but the dataset is small. Big datasets could be used. According to current trends, it’s not sufficient.

We are working to overcome these shortcoming with our model.

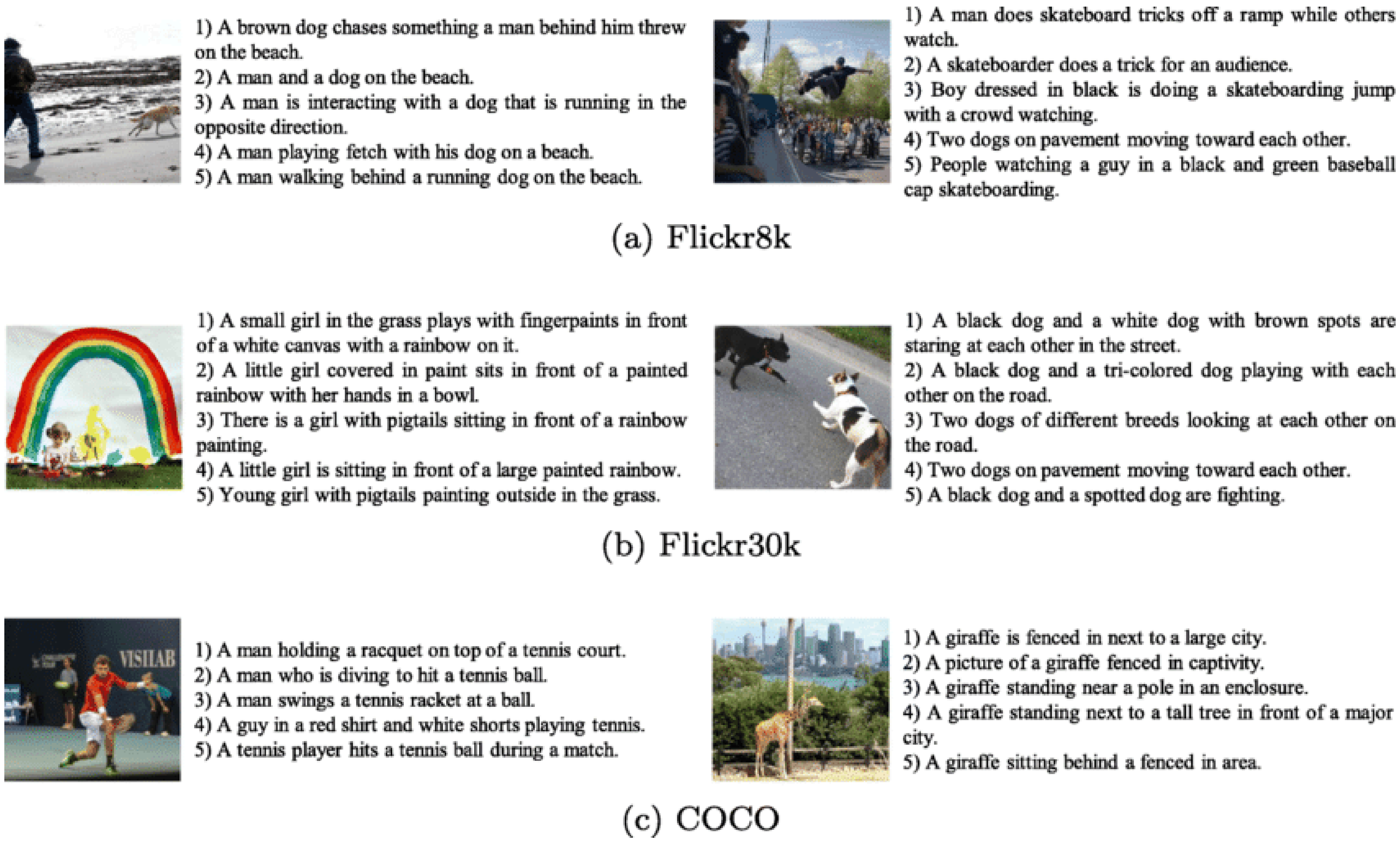
Reference: -Anu, M., Divya, S., et al., 2021. Building a voice based image caption generator with deep learning, in: 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE. pp. 943–948

DATASET

Flickr 8k
8091 images

Flickr 30k
31,783 images

MS COCO
164K images



1) A brown dog chases something a man behind him threw on the beach.
2) A man and a dog on the beach.
3) A man is interacting with a dog that is running in the opposite direction.
4) A man playing fetch with his dog on a beach.
5) A man walking behind a running dog on the beach.

1) A man does skateboard tricks off a ramp while others watch.
2) A skateboarder does a trick for an audience.
3) Boy dressed in black is doing a skateboarding jump with a crowd watching.
4) Two dogs on pavement moving toward each other.
5) People watching a guy in a black and green baseball cap skateboarding.

(a) Flickr8k

1) A small girl in the grass plays with fingerpaints in front of a white canvas with a rainbow on it.
2) A little girl covered in paint sits in front of a painted rainbow with her hands in a bowl.
3) There is a girl with pigtails sitting in front of a rainbow painting.
4) A little girl is sitting in front of a large painted rainbow.
5) Young girl with pigtails painting outside in the grass.

1) A black dog and a white dog with brown spots are staring at each other in the street.
2) A black dog and a tri-colored dog playing with each other on the road.
3) Two dogs of different breeds looking at each other on the road.
4) Two dogs on pavement moving toward each other.
5) A black dog and a spotted dog are fighting.

(b) Flickr30k

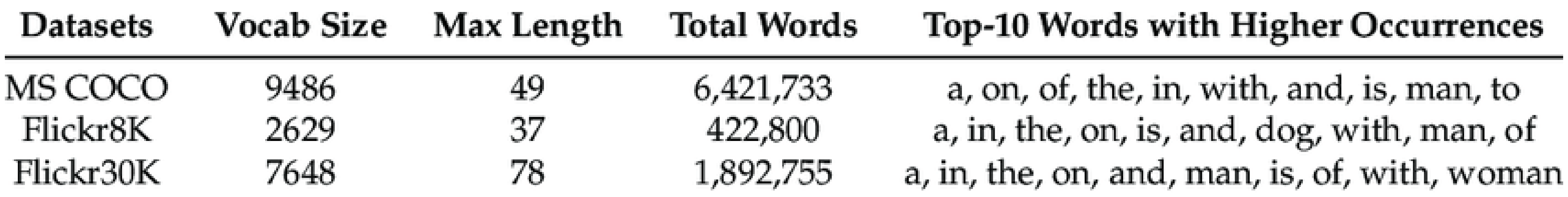
1) A man holding a racquet on top of a tennis court.
2) A man who is diving to hit a tennis ball.
3) A man swings a tennis racket at a ball.
4) A guy in a red shirt and white shorts playing tennis.
5) A tennis player hits a tennis ball during a match.

1) A giraffe is fenced in next to a large city.
2) A picture of a giraffe fenced in captivity.
3) A giraffe standing near a pole in an enclosure.
4) A giraffe standing next to a tall tree in front of a major city.
5) A giraffe sitting behind a fenced in area.

(c) COCO

DATASET

LucidLens



Datasets	Vocab Size	Max Length	Total Words	Top-10 Words with Higher Occurrences
MS COCO	9486	49	6,421,733	a, on, of, the, in, with, and, is, man, to
Flickr8K	2629	37	422,800	a, in, the, on, is, and, dog, with, man, of
Flickr30K	7648	78	1,892,755	a, in, the, on, and, man, is, of, with, woman

Why Flickr 30k

- A standard Benchmark for sentence based description of images
- Good long Captions
- Descent Vocab Size
- Large and Diverse
- Freely available and widely researched

Data Collection

- Images are collected from flickr platform
- Annoatated by humans
- Included Criteria such as human and animals
- In accordance with Flickr's terms of service and by anonymizing any personal information that could be identified in the captions.

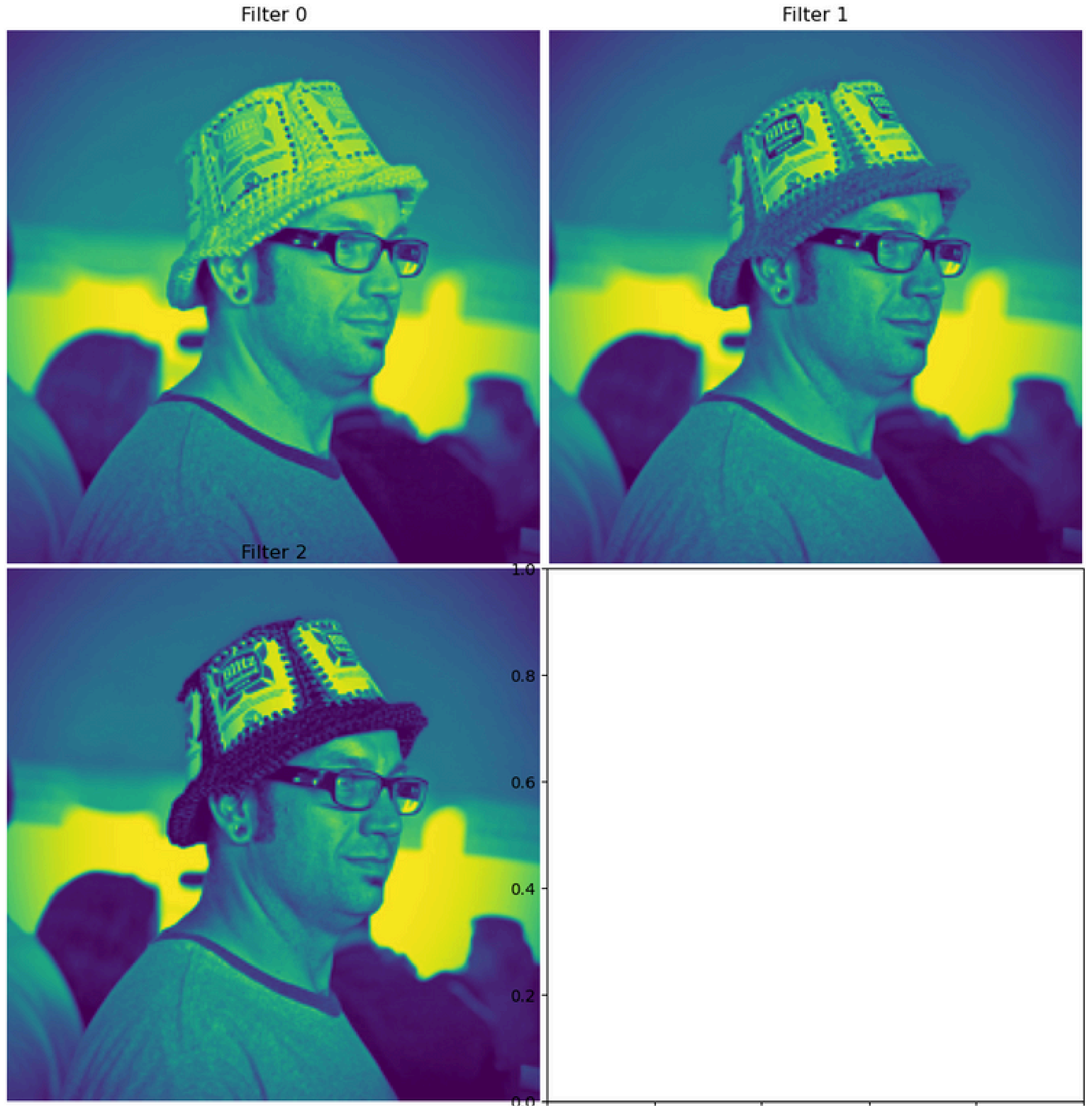
DataSet Details

- 31783 Images
- 158915 Captions
- No null Values

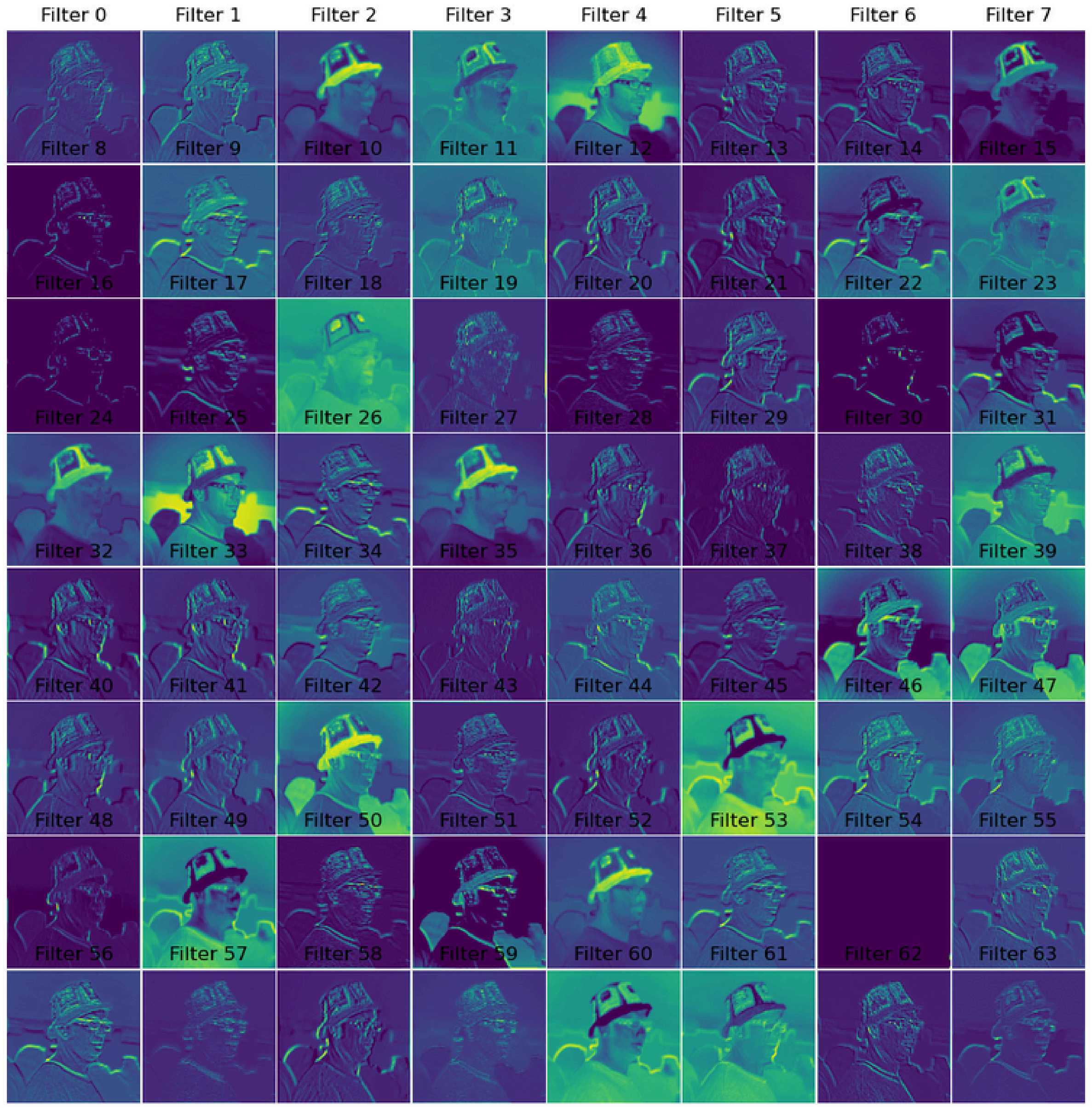
DATASET

LucidLens

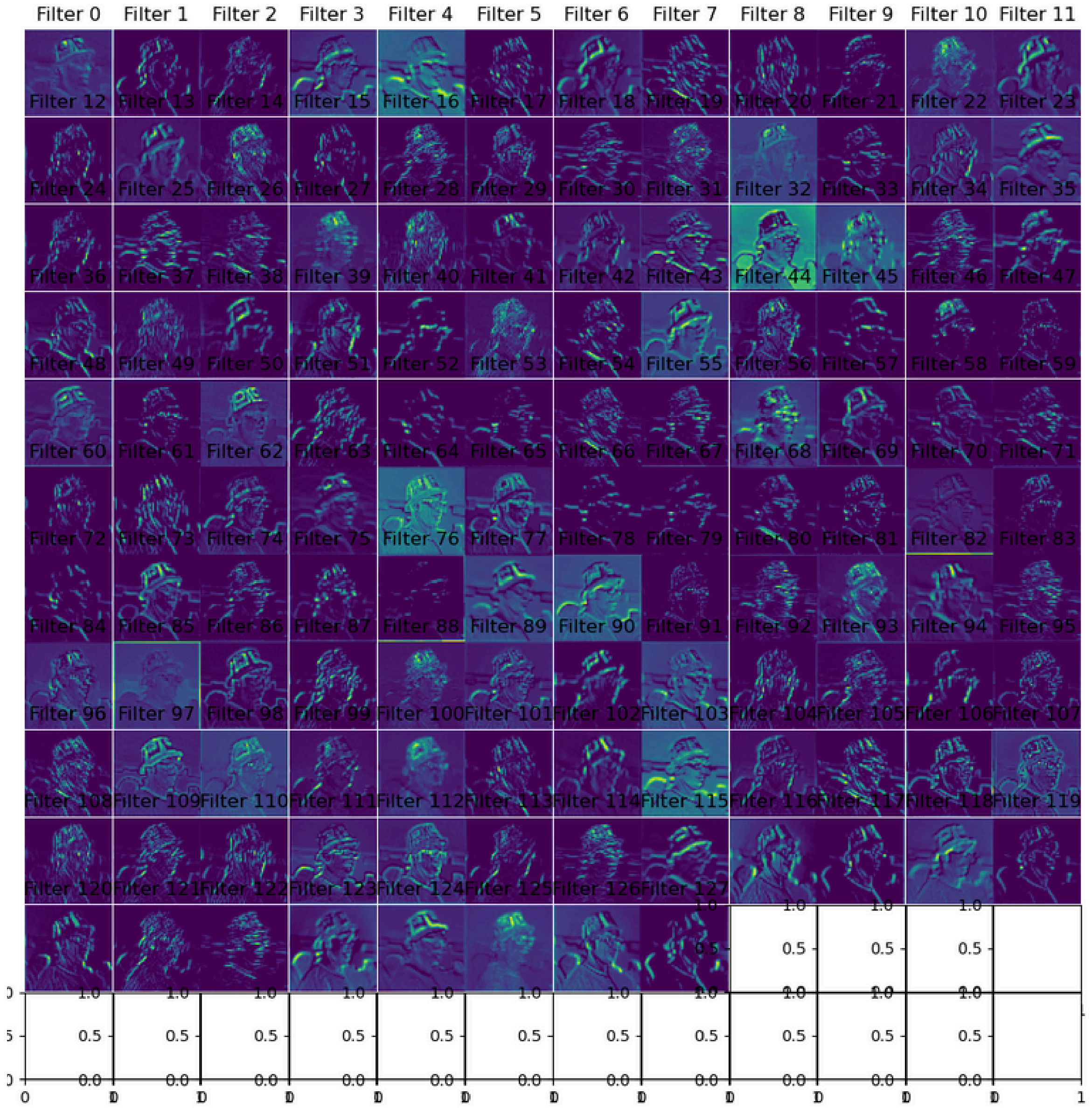
Visualization of VGG16 CNN Layers



Layer input_layer Activations



Layer block1_pool Activations

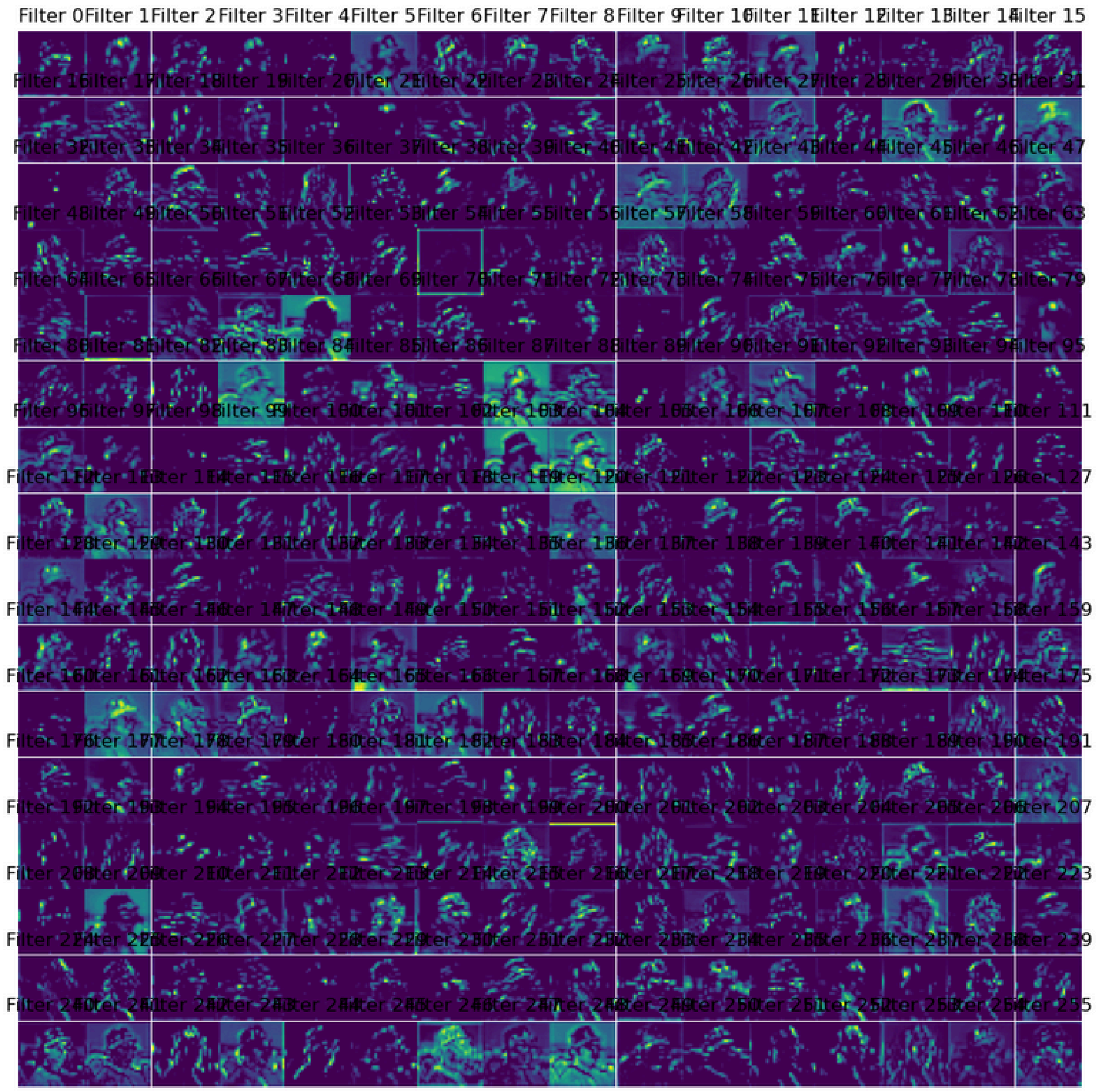


Layer block2_pool Activations

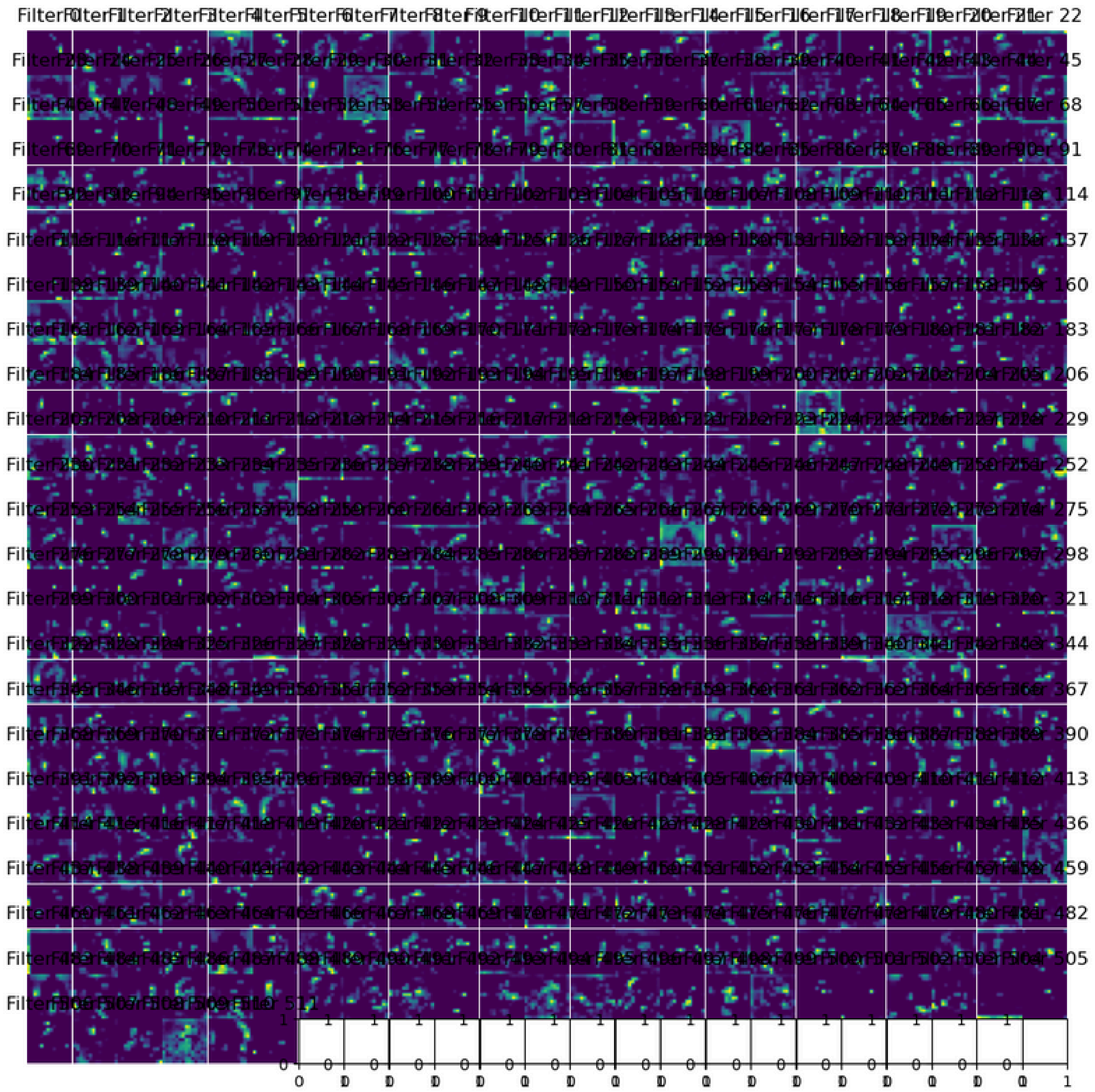
DATASET

LucidLens

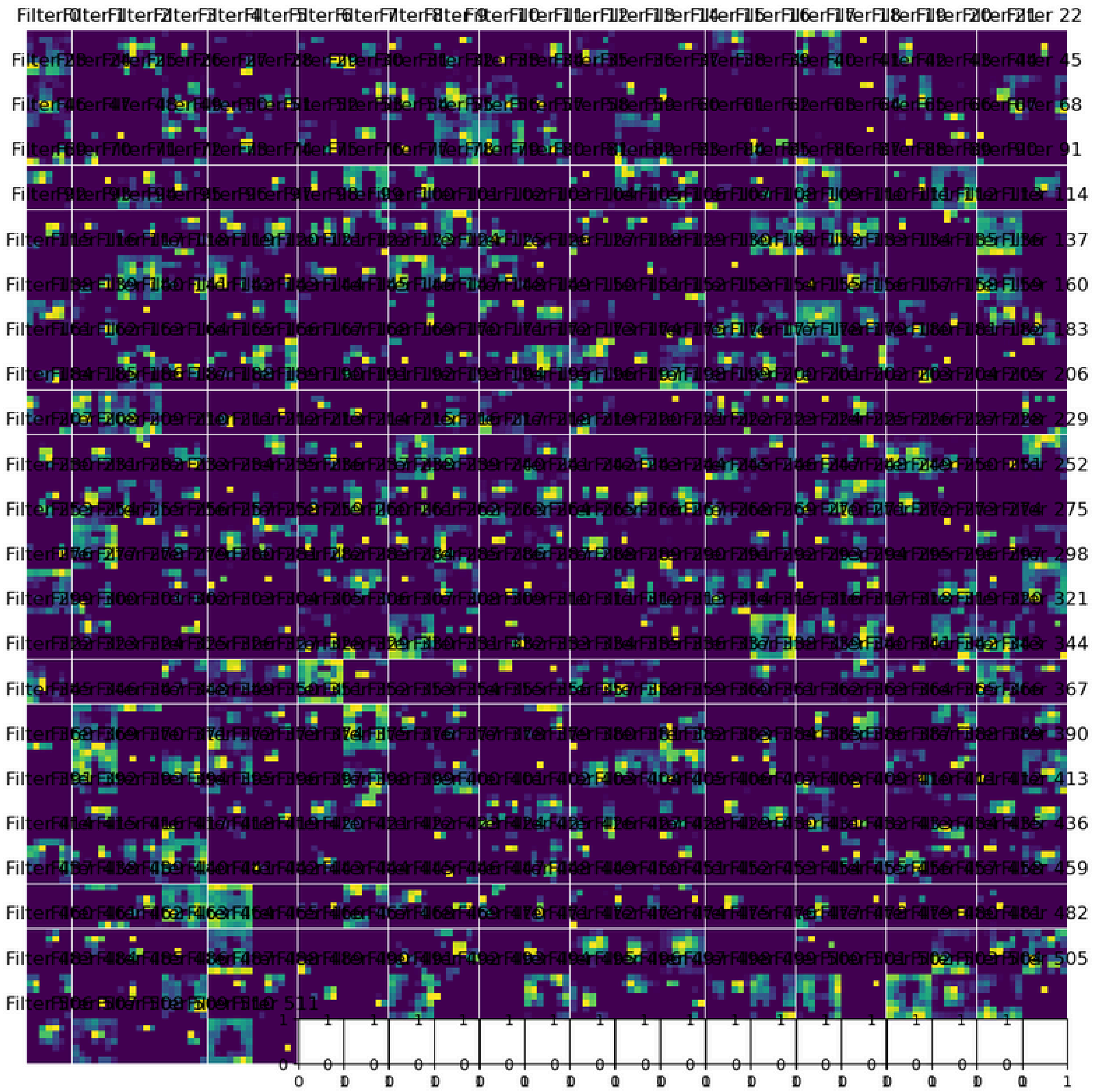
Layer block3_pool Activations



Layer block4_pool Activations



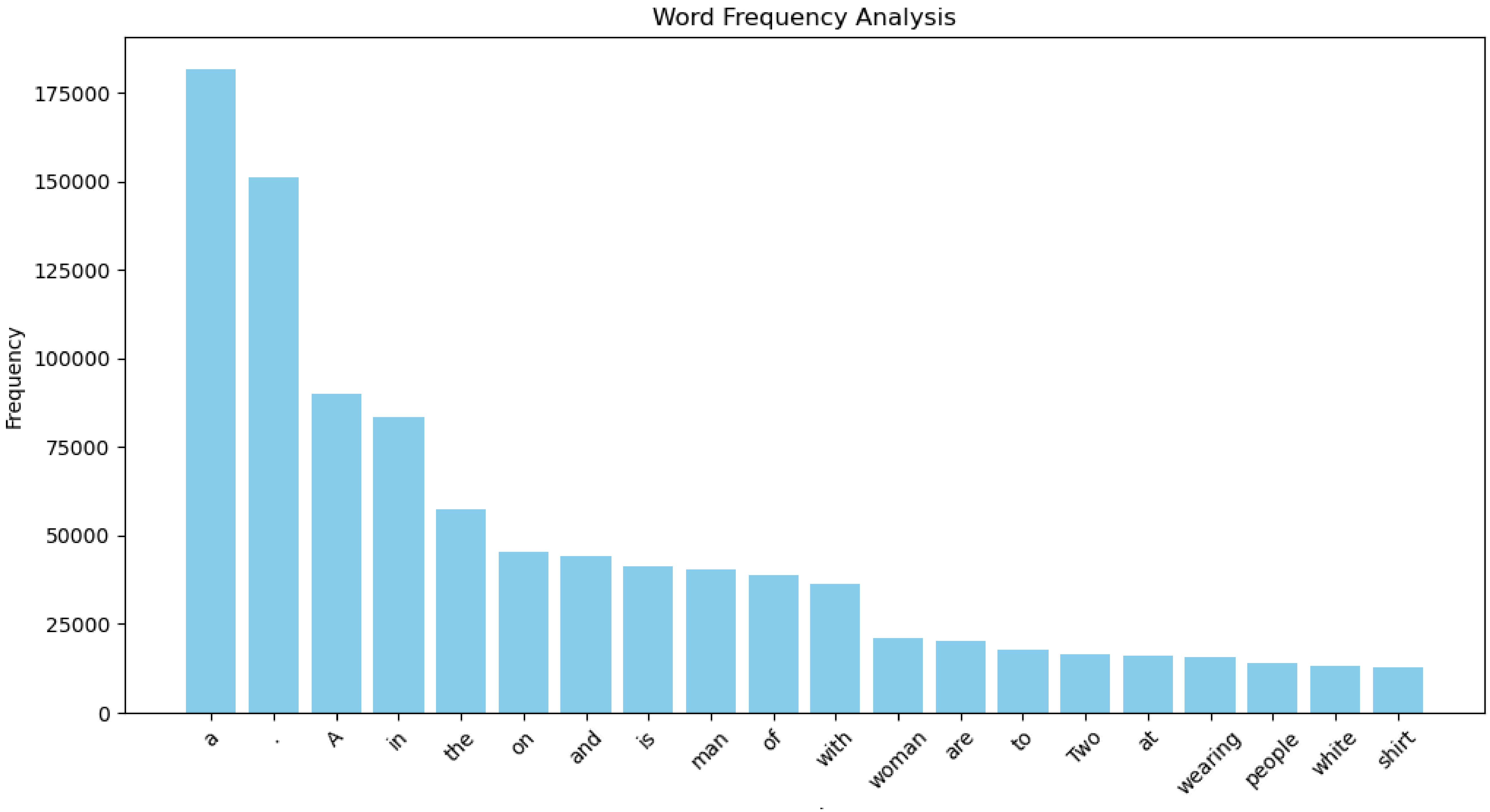
Layer block5_pool Activations



DATASET

LucidLens

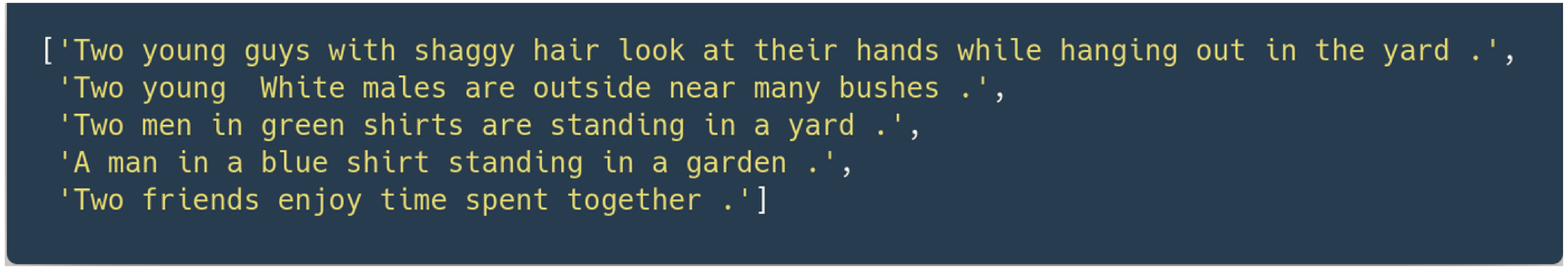
Visualization



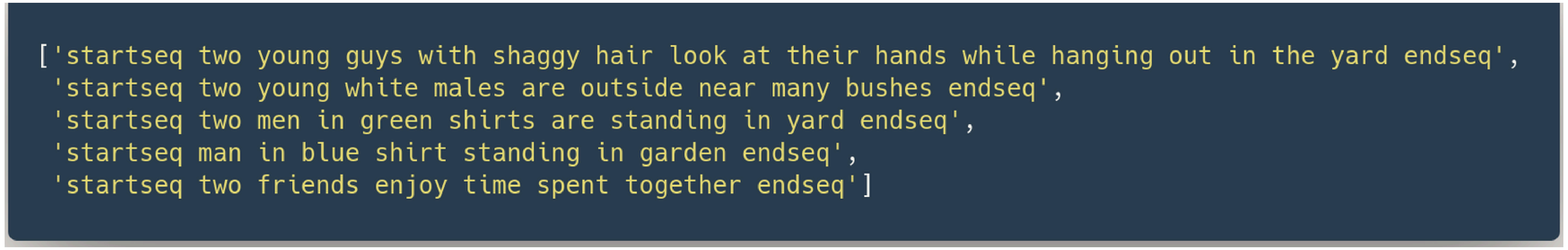
DATASET

Preprocessing

- Mapping Between Captions and image_name
- convert Uppercase to lowercase
- Remove Special characters and full stops
- Tokenization



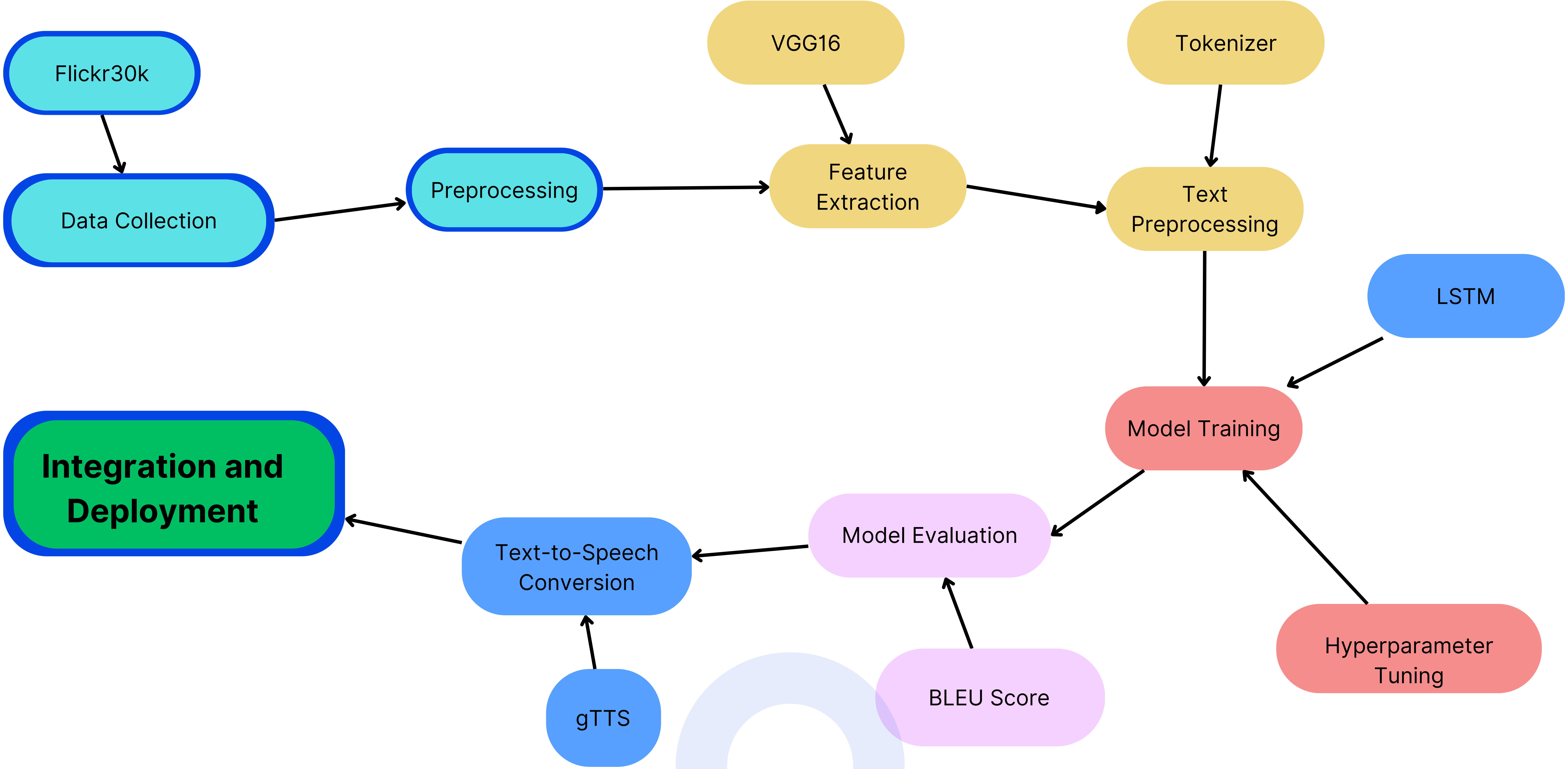
```
['Two young guys with shaggy hair look at their hands while hanging out in the yard .',
 'Two young  White males are outside near many bushes .',
 'Two men in green shirts are standing in a yard .',
 'A man in a blue shirt standing in a garden .',
 'Two friends enjoy time spent together .']
```



```
['startseq two young guys with shaggy hair look at their hands while hanging out in the yard endseq',
 'startseq two young white males are outside near many bushes endseq',
 'startseq two men in green shirts are standing in yard endseq',
 'startseq man in blue shirt standing in garden endseq',
 'startseq two friends enjoy time spent together endseq']
```

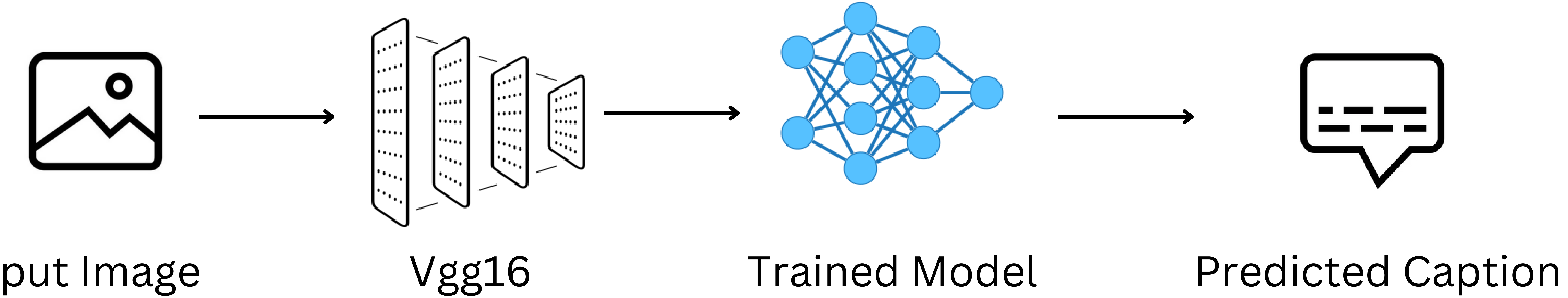
ML METHODOLOGY:

LucidLens



CAPTION PREDICTION

LucidLens



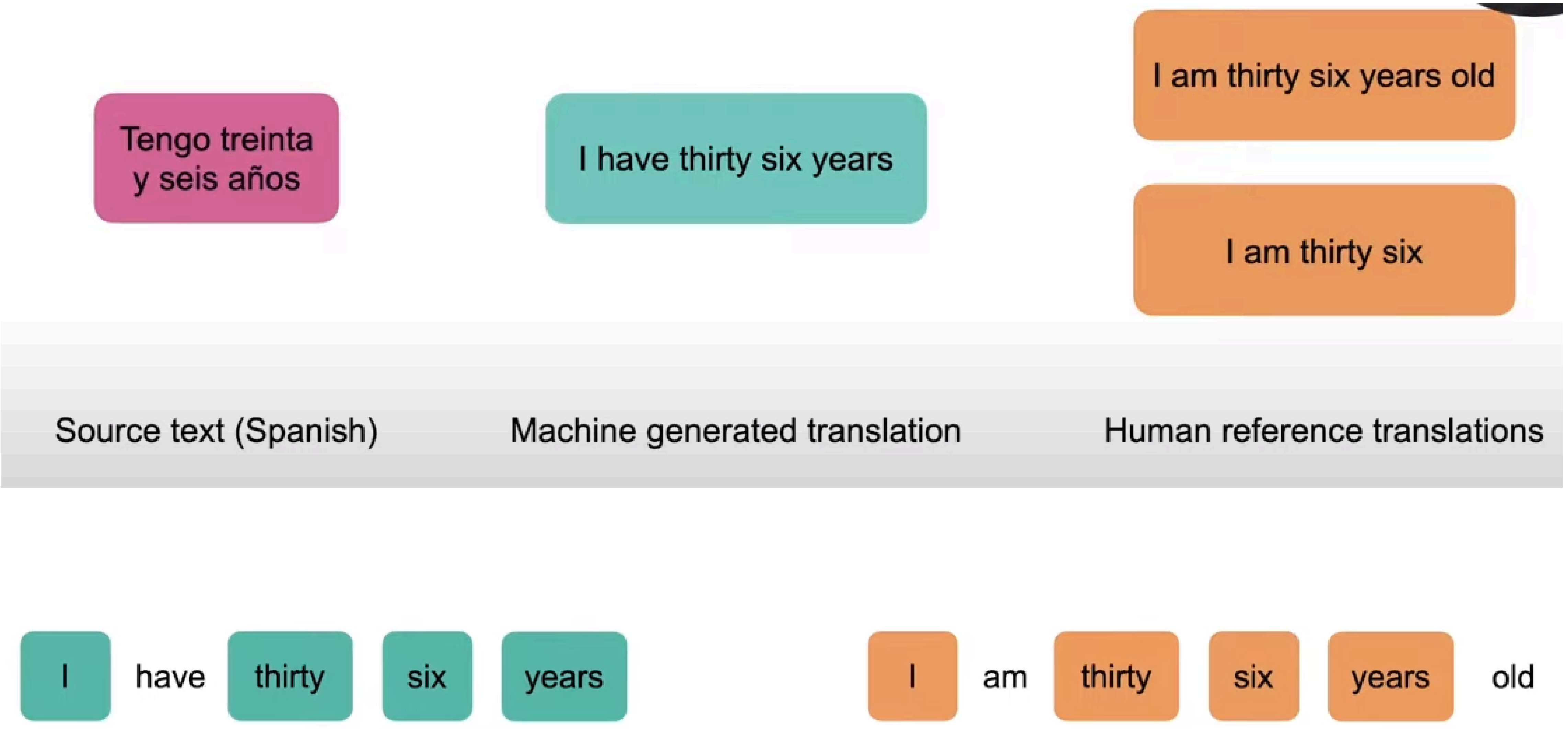
Input Image

Vgg16

Trained Model

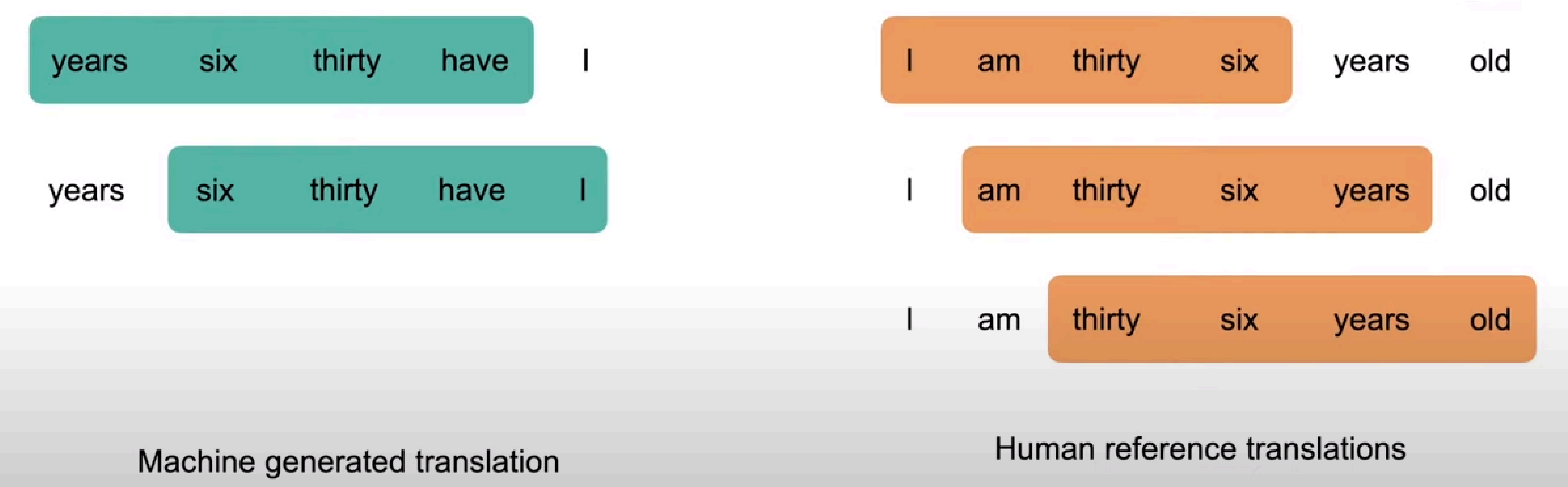
Predicted Caption

WHAT IS BLEU SCORE?


Tengo treinta y seis años
I have thirty six years
I am thirty six years old
I am thirty six
Source text (Spanish)
Machine generated translation
Human reference translations
I
have
thirty
six
years
I
am
thirty
six
years
old

BLEU1 = $\text{Unigram precision} = \dfrac{\text{Num word matches}}{\text{Num words in generation}} = \dfrac{4}{5}$

BLEU SCORE



years six thirty have I
years six thirty have I
I am thirty six years old
I am thirty six years old
I am thirty six years old

Machine generated translation

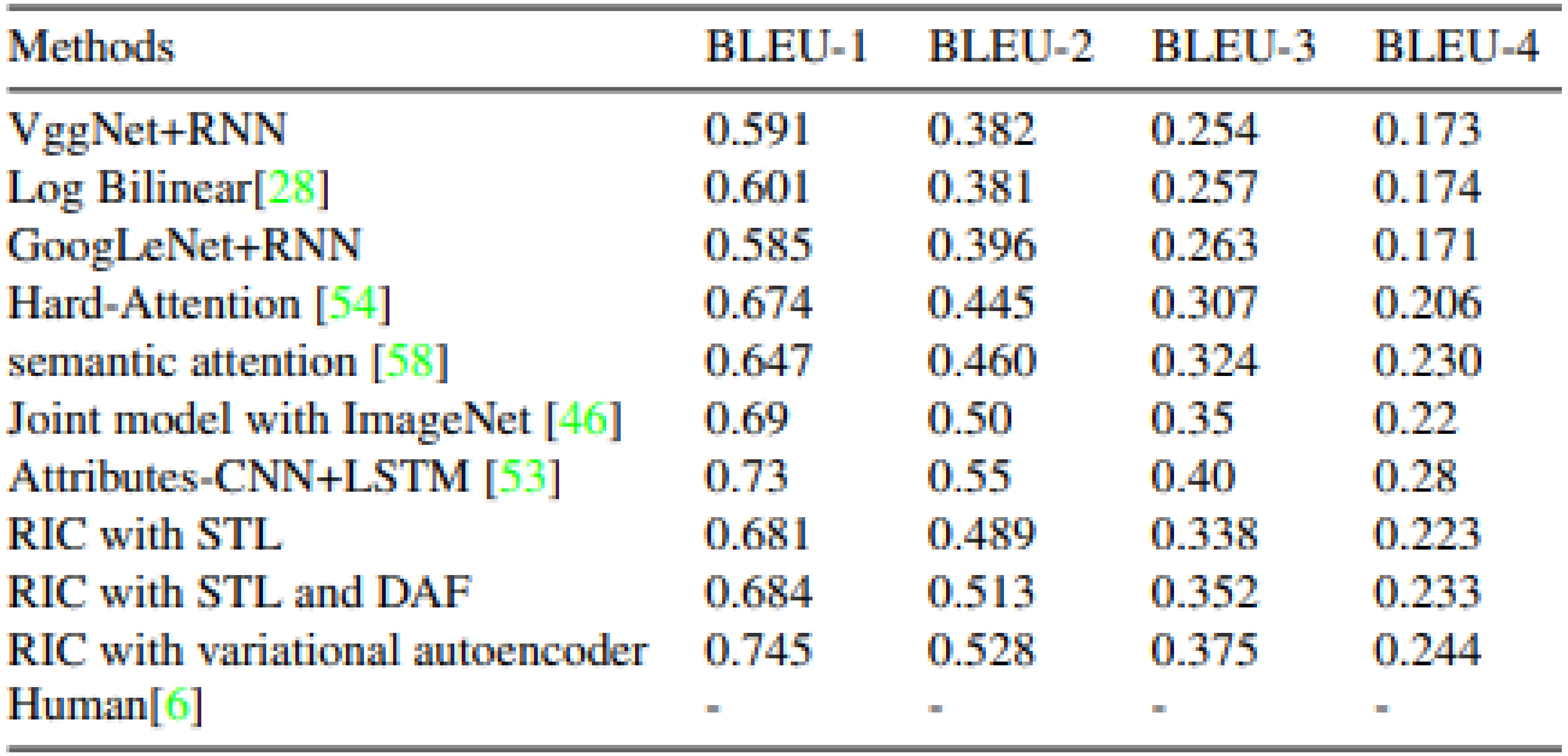
Human reference translations

$$4 - gram\ precision = \frac{Clip(Num\ 4 - gram\ matches)}{Num\ 4 - gram\ in\ generation}$$

COMPARISON WITH OTHERS

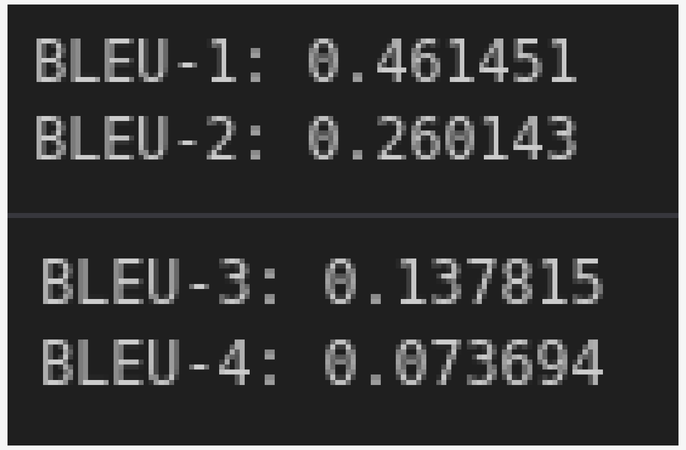
LucidLens

Others on Flickr30k

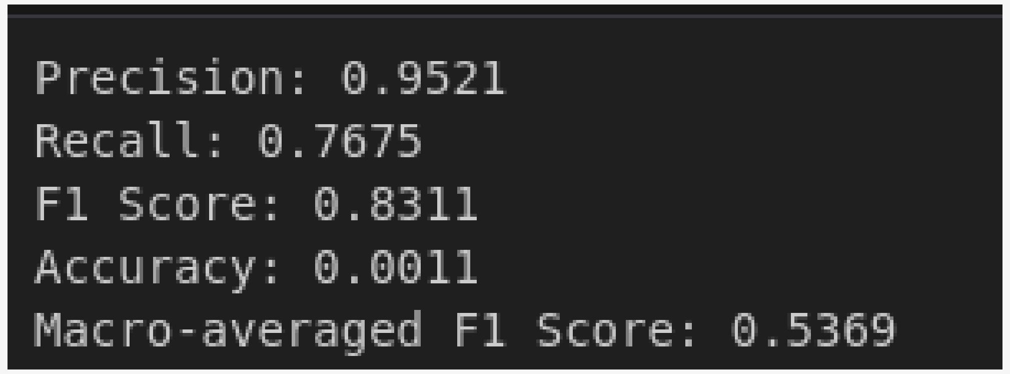


Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4
VggNet+RNN	0.591	0.382	0.254	0.173
Log Bilinear[28]	0.601	0.381	0.257	0.174
GoogLeNet+RNN	0.585	0.396	0.263	0.171
Hard-Attention [54]	0.674	0.445	0.307	0.206
semantic attention [58]	0.647	0.460	0.324	0.230
Joint model with ImageNet [46]	0.69	0.50	0.35	0.22
Attributes-CNN+LSTM [53]	0.73	0.55	0.40	0.28
RIC with STL	0.681	0.489	0.338	0.223
RIC with STL and DAF	0.684	0.513	0.352	0.233
RIC with variational autoencoder	0.745	0.528	0.375	0.244
Human[6]	-	-	-	-

Ours



```
BLEU-1: 0.461451
BLEU-2: 0.260143
BLEU-3: 0.137815
BLEU-4: 0.073694
```



```
Precision: 0.9521
Recall: 0.7675
F1 Score: 0.8311
Accuracy: 0.0011
Macro-averaged F1 Score: 0.5369
```

From Literature Review

Building A Voice Based Image Caption Generator with Deep Learning : Accuracy 90 %

Object detection and Narrator for Visually Impaired people: Accuracy 62-75 %

Indoor object detection and recognition for an ICT mobility : Accuracy 75 %

Reference:- Xie, Tian, Weiping Ding, Jinbao Zhang, Xusen Wan, and Jiehua Wang. 2023. "Bi-LS-AttM: A Bidirectional LSTM and Attention Mechanism Model for Improving Image Captioning" Applied Sciences 13, no. 13: 7916. https://doi.org/10.3390/app13137916

DEPLOYING AT PLAKSHA

LucidLens

Campus Tour Guide

Transfer Learning

- Obtain the dataset
- Preprocess the data
- Load Pre-trained model
- Fine-tune the model
- Train the model

Challenges

- Obtaining the dataset
- Low accuracy

CHALLENGES FACED:

1. **Data Quality and Diversity:** Ensuring the dataset covers a wide range of scenes, objects, and activities to generate diverse and meaningful captions.
2. **Hyperparameter Tuning:** Experimentation with learning rates, batch sizes, and model architectures to find the optimal configuration.
3. **Real-time Performance:** Ensuring the caption generation process is fast enough to provide real-time feedback for blind users.

DEMO

REFERENCES

LucidLens

1) Guan, Zhibin & Liu, Kang & Yan, Ma & Qian, Xu & Ji, Tongkai. (2018). Sequential Dual Attention: Coarse-to-Fine-Grained Hierarchical Generation for Image Captioning. Symmetry. 10. 626. 10.3390/sym10110626.

2) Yohannes, E., Lin, P., Lin, C.Y., Shih, T.K., 2020. Robot eye: Automatic object detection and recognition using deep attention network to assist blind people, in: 2020 International Conference on Pervasive Artificial Intelligence (ICPAI), IEEE. pp. 152–157

3) Tan, M., Pang, R., Le, Q.V., 2020. Efficientdet: Scalable and efficient object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10781–10790

4) Pardasani, A., Indi, P.N., Banerjee, S., Kamal, A., Garg, V., 2019. Smart assistive navigation devices for visually impaired people, in: 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), IEEE. pp. 725–729.

5) Mahendru, M., Dubey, S.K., 2021. Real time object detection with audio feedback using yolo vs. yolo v3, in: 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE. pp. 734–740.

6) Rajwani, R., Purswani, D., Kalinani, P., Ramchandani, D., Dokare, I., 2018. Proposed system on object detection for visually impaired people. International Journal of Information Technology (IJIT) 4, 1–6.

7) Nasreen, J., Arif, W., Shaikh, A.A., Muhammad, Y., Abdullah, M., 2019. Object detection and narrator for visually impaired people, in: 2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS), IEEE. pp. 1–4

8) Xie, Tian, Weiping Ding, Jinbao Zhang, Xusen Wan, and Jiehua Wang. 2023. "Bi-LS-AttM: A Bidirectional LSTM and Attention Mechanism Model for Improving Image Captioning" *Applied Sciences* 13, no. 13: 7916. https://doi.org/10.3390/app13137916

OUR TEAM

Govind (U20220037)

Prashant Mishra (U20220067)

Rahul Kumar (U20220071)

LucidLens

LucidLens

THANK YOU

FOR YOUR ATTENTION

May 2024