

IMAGE CAPTIONING

ANIOL PETIT & JAN AGUILÓ

TABLE OF CONTENTS

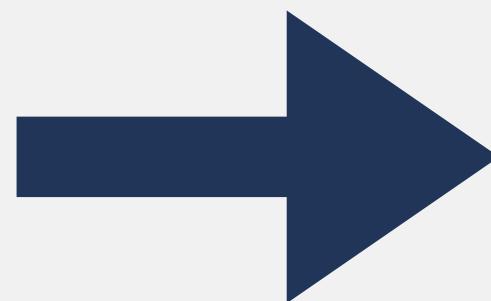
- 01** Problem Description
- 02** State-of-the-Art
- 03** Methodology: Data Analysis
- 04** Methodology: Model Architecture
- 05** Experiments
- 06** Results + Discussion
- 07** Conclusions
- 08** References

PROBLEM DEFINITION

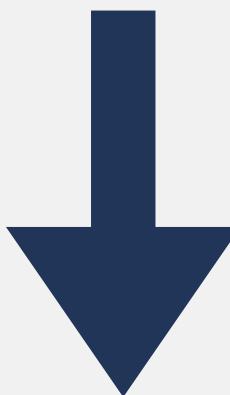


**HOW CAN WE AUTOMATICALLY GENERATE A
MEANINGFUL AND GRAMMATICALLY CORRECT
SENTENCE THAT DESCRIBES AN IMAGE?**

PROBLEM DEFINITION



MODEL



<START> A GLASS VASE FILLED WITH WINE BOTTLES AND WINE . <END>

STATE-OF-THE-ART

Traditional CNN-RNN Model

SIMPLE

STRUGGLES WITH COMPLEX IMAGES (FIXED-SIZE ENCODING)

[HTTPS://ARXIV.ORG/ABS/2404.18062](https://arxiv.org/abs/2404.18062)

STATE-OF-THE-ART

Traditional CNN-RNN Model

SIMPLE

STRUGGLES WITH COMPLEX IMAGES (FIXED-SIZE ENCODING)

[HTTPS://ARXIV.ORG/ABS/2404.18062](https://arxiv.org/abs/2404.18062)

Attention mechanisms

IMPROVES ACCURACY AND INTERPRETABILITY

DEPENDS ON RNN'S (HARD TO PARALLELIZE)

[HTTPS://ARXIV.ORG/ABS/1502.03044](https://arxiv.org/abs/1502.03044)

STATE-OF-THE-ART

Traditional CNN-RNN Model

SIMPLE

STRUGGLES WITH COMPLEX IMAGES (FIXED-SIZE ENCODING)

[HTTPS://ARXIV.ORG/ABS/2404.18062](https://arxiv.org/abs/2404.18062)

Attention mechanisms

IMPROVES ACCURACY AND INTERPRETABILITY

DEPENDS ON RNN'S (HARD TO PARALLELIZE)

[HTTPS://ARXIV.ORG/ABS/1502.03044](https://arxiv.org/abs/1502.03044)

Transformer-based Models

BEST RESULTS AND BETTER SCALABILITY

MASSIVE DATASETS AND COMPLEX TRAINING MODELS

[HTTPS://ARXIV.ORG/ABS/2203.15350](https://arxiv.org/abs/2203.15350)

STATE-OF-THE-ART

Traditional CNN-RNN Model

SIMPLE

STRUGGLES WITH COMPLEX IMAGES (FIXED-SIZE ENCODING)

[HTTPS://ARXIV.ORG/ABS/2404.18062](https://arxiv.org/abs/2404.18062)

Attention mechanisms

IMPROVES ACCURACY AND INTERPRETABILITY

DEPENDS ON RNN'S (HARD TO PARALLELIZE)

[HTTPS://ARXIV.ORG/ABS/1502.03044](https://arxiv.org/abs/1502.03044)

Transformer-based Models

BEST RESULTS AND BETTER SCALABILITY

MASSIVE DATASETS AND COMPLEX TRAINING MODELS

[HTTPS://ARXIV.ORG/ABS/2203.15350](https://arxiv.org/abs/2203.15350)

**Our model:
CNN - LSTM + attention**

METHODOLOGY: DATA ANALYSIS

THE DATASET

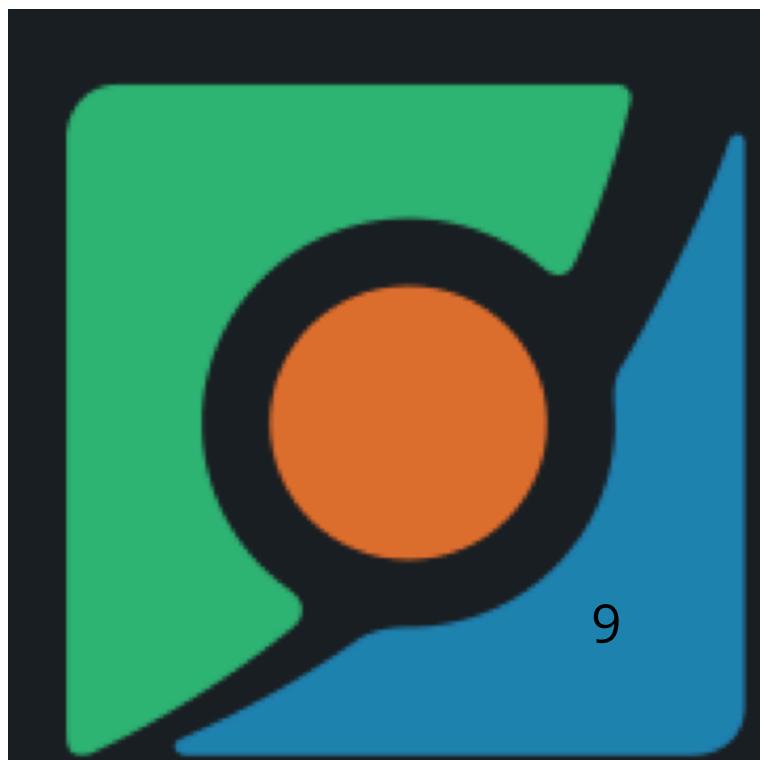
+82.000 TRAINING IMAGES

EACH WITH **5-HUMAN
WRITTEN CAPTIONS**
(VOCABULARY)

+40.000 VALIDATION IMAGES

WE WORKED WITH SUBSETS:
1.5K, 4K, 8K, 25K

MSCOCO



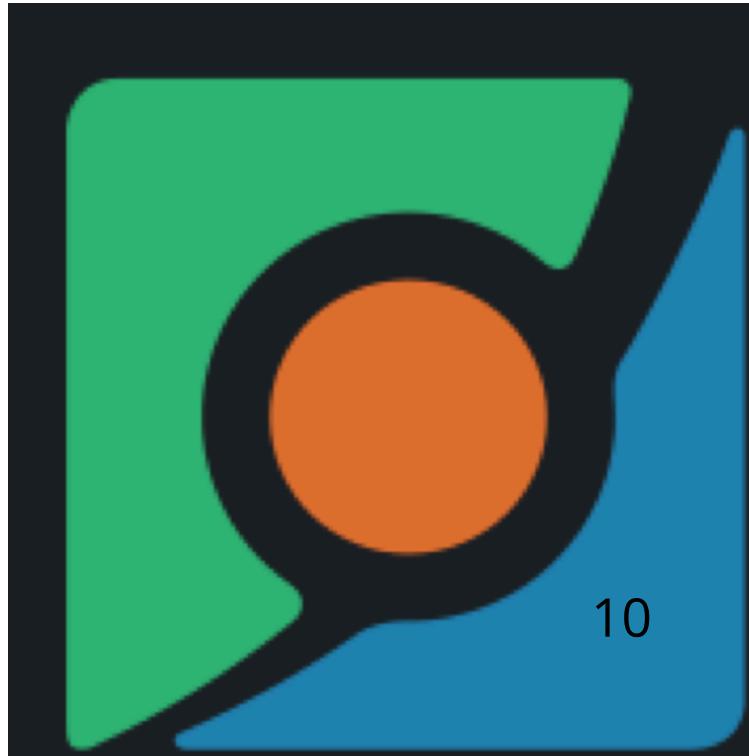
METHODOLOGY: DATA ANALYSIS

PREPROCESSING

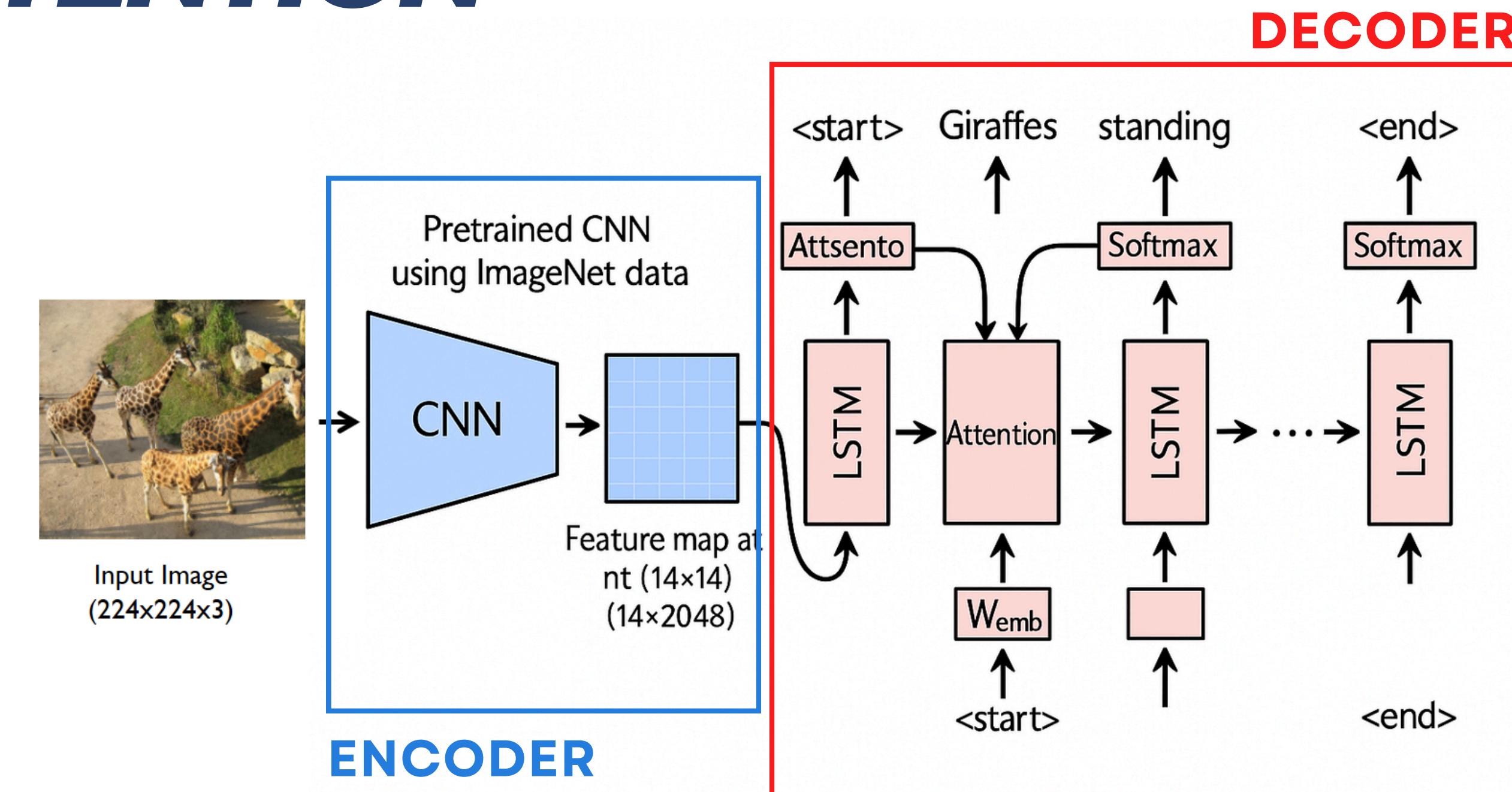
RESIZE

DATA LOADER

TRANSFORMATIONS



METHODOLOGY: CNN-LSTM W/ATTENTION



Source: https://github.com/yunjey/pytorch-tutorial/tree/master/tutorials/03-advanced/image_captioning + CHAT GPT

EXPERIMENTS

TEST PRETRAINED MODEL

ADDED ATTENTION +
TOP-P SAMPLING &
TESTED WITH SUBSETS

USED FULL DATASET

INCREASED ENCODER'S
CAPACITY

RESULTS (WITHOUT ATTENTION)



<start> a man in a kitchen with a blender and a bowl of food .

<end>



<start> a man in a suit and tie with a tie . <end>

FINAL RESULTS W/ATTENTION

Image Captioning with Attention

Upload an image and generate a caption with visualized attention maps.

image



Clear

Submit

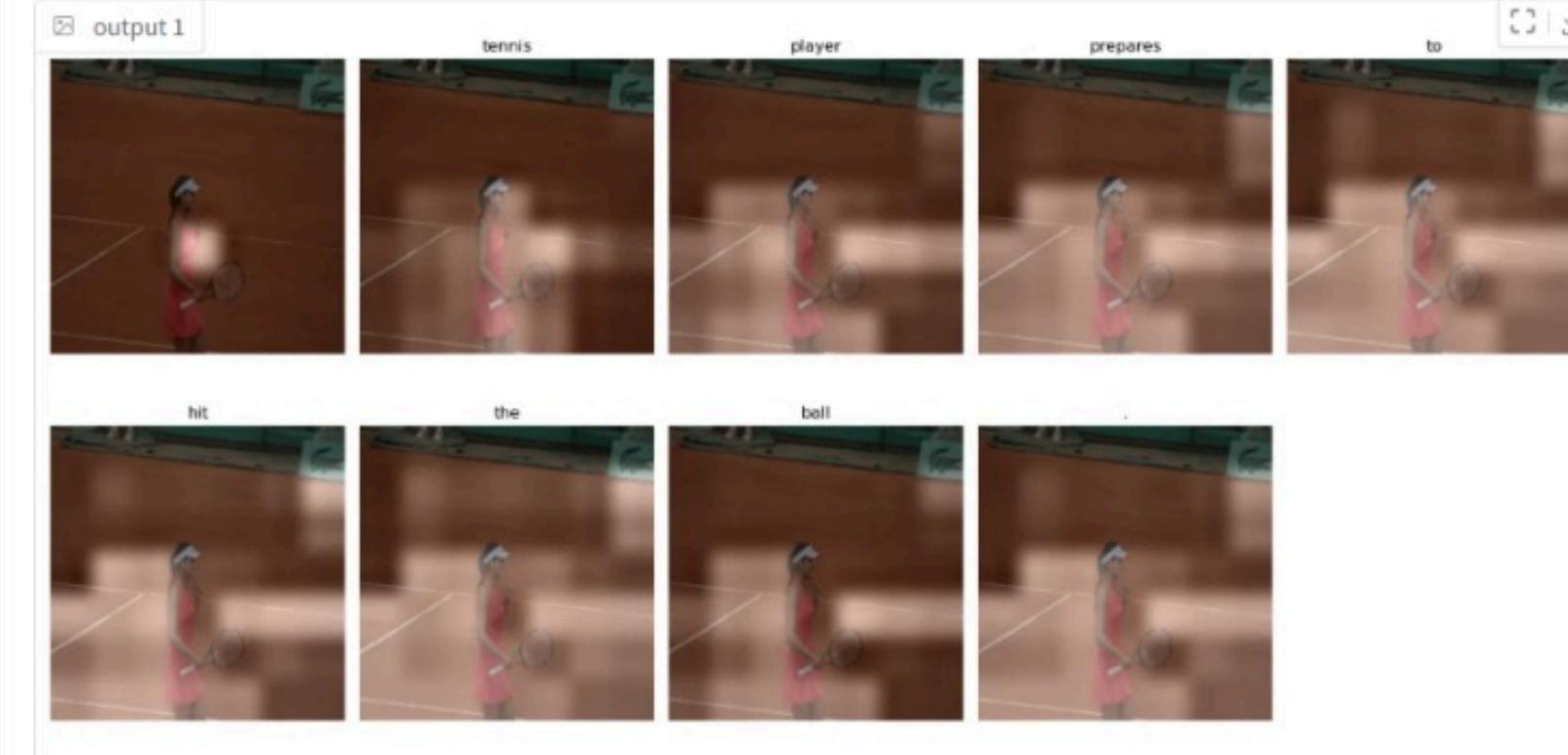
Flag

output 0

a tennis player prepares to hit the ball .

output 1

hit the ball .



Three orange arrows point from the word 'ball' in the caption to the eighth image in the grid, which shows the player's racket hitting the ball.

FINAL RESULTS W/ATTENTION

Image Captioning with Attention

Upload an image and generate a caption with visualized attention maps.

image



two small elephants outside in the grass .

output 0

output 1

small

elephants

outside

in

the

grass

Flag

Clear

Submit

FINAL RESULTS W/ATTENTION



output 0

a man sitting on the beach holding the sand .

Flag

output 1

man

sitting

on

the

beach

holding

the

sand

The diagram illustrates a visual captioning model's attention mechanism. It shows a photograph of a beach scene and a corresponding text caption "a man sitting on the beach holding the sand." Below the caption is a grid of 16 smaller images, each showing a different part of the scene highlighted by a white rectangular mask. Orange arrows point from the words "man," "sitting," "on," and "the" in the caption to their respective highlighted regions in the grid. The word "beach" points to the water, "holding" points to the person on the left, and "sand" points to the person on the right. A "Flag" button is located at the bottom of the interface.

FINAL RESULTS W/ATTENTION



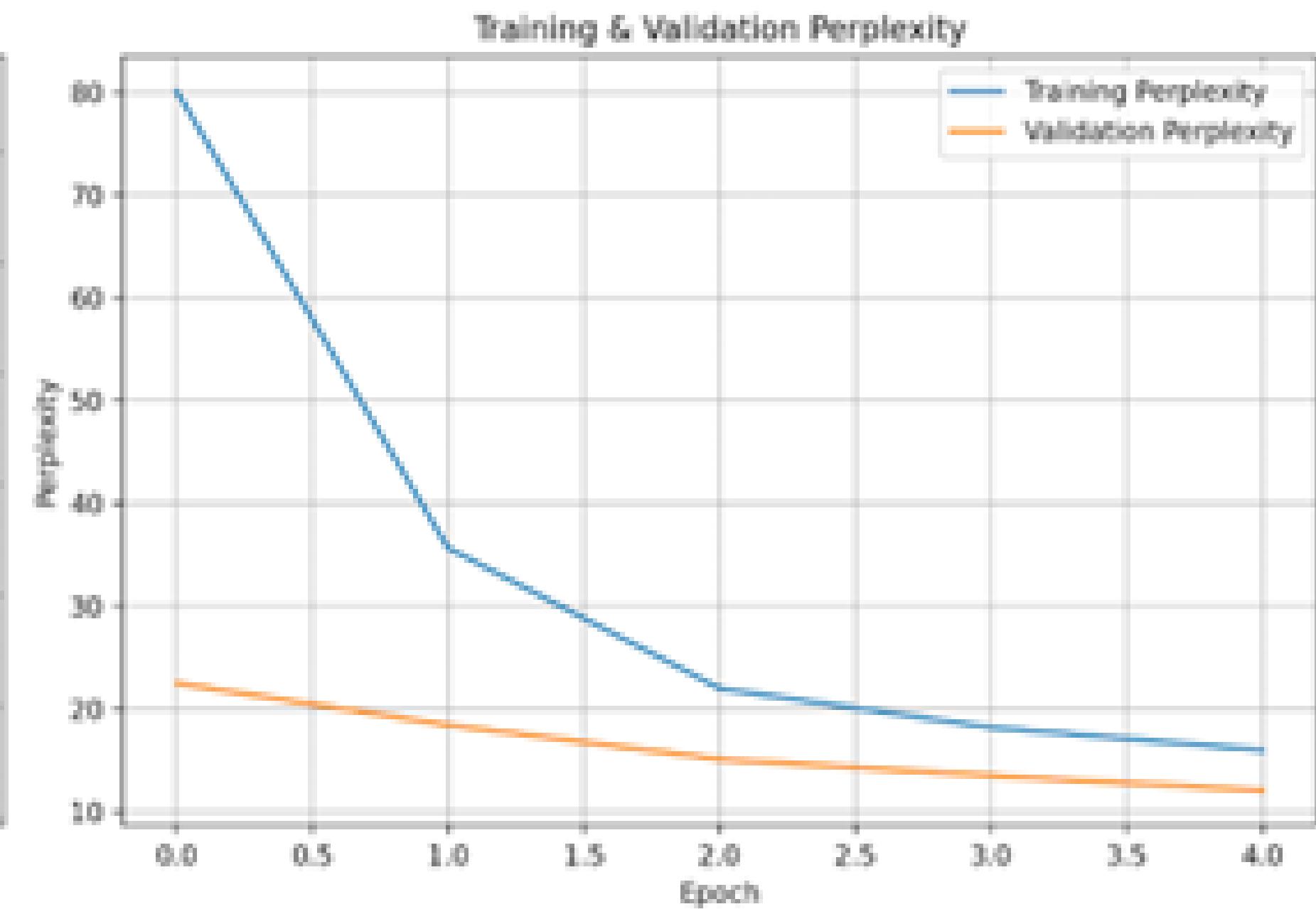
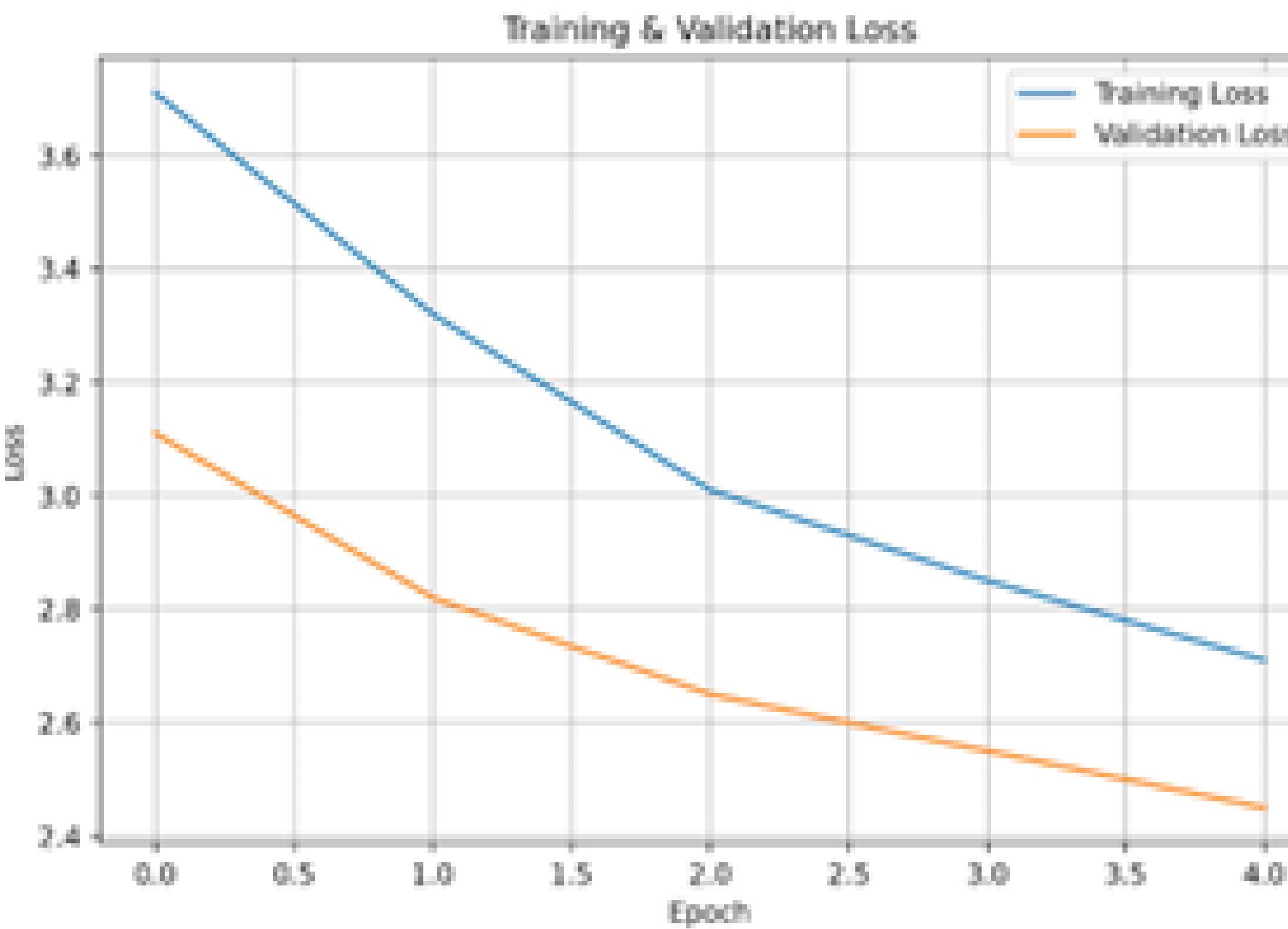
Caption 2

two people standing on the beach next to surfboards on a beach .

Attention 2

wo	people	standing	on	the
				
beach	next	to	surfboards	on
				
a	beach	i		

FINAL RESULTS EVALUATION



CONCLUSIONS

- **Attention** effectively achieves spatial aware captions and improves quality
- **Top-p sampling**: different captions for the same image
- Further work
 - More **evaluation metrics** (BLEU)
 - **Hyperparameter** tuning
 - **Transformer-based** methods
- Needs time and powerful resources to train a good model

REFERENCES

ORIGINAL MODEL

- https://github.com/yunjey/pytorch-tutorial/tree/master/tutorials/03-advanced/image_captioning

CNN

- <https://cs231n.github.io/convolutional-networks/>

RNN (LSTM)

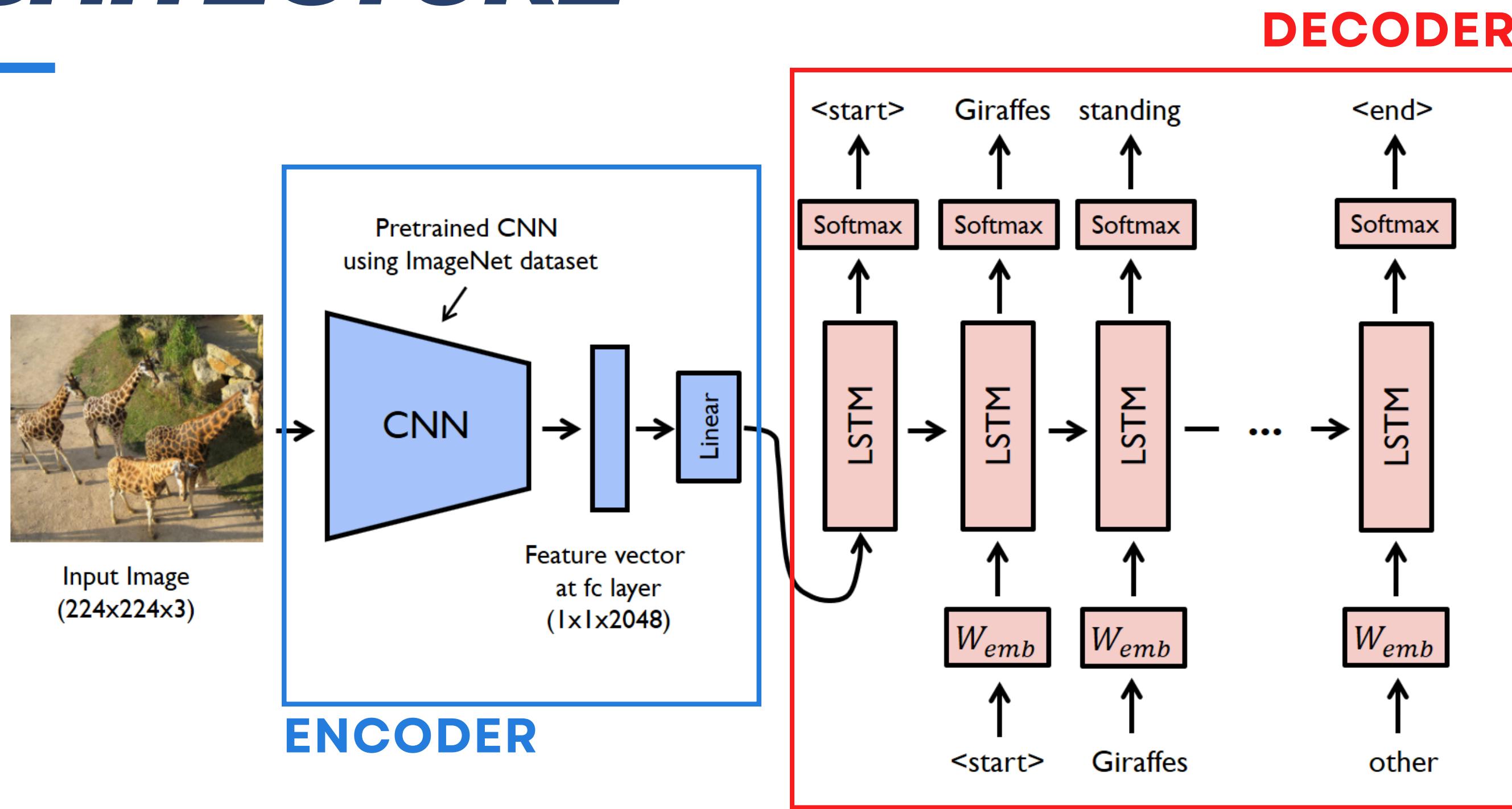
- <https://ieeexplore.ieee.org/abstract/document/7508408>
- <https://hackernoon.com/understanding-architecture-of-lstm-cell-from-scratch-with-code-8da40f0b71f4>

IMAGE CAPTIONING

- https://www.researchgate.net/publication/374492698_Automatic_Generation_of_Image_Caption_Based_on_Semantic_Relation_using_Deep_Visual_Attention_Prediction

**THANK
YOU**

METHODOLOGY: BASE MODEL ARCHITECTURE



Source: https://github.com/yunjey/pytorch-tutorial/tree/master/tutorials/03-advanced/image_captioning.

RESULTS (25K SAMPLES)



a desk with two books and other foods look .

a picture of a square style kitchen with toilet .



a gray and white cat sleeping on a toilet .

a large white bathroom sits next to a windows .