IMAGE CAPTIONING

ANIOL PETIT & JAN AGUILÓ

TABLE OF CONTENTS

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

STATE-OF-THE-ART

Traditional CNN-RNN Model

SIMPLE

STRUGGLES WITH COMPLEX IMAGES (FIXED-SIZE ENCODING)

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

STATE-OF-THE-ART

Traditional CNN-RNN Model

SIMPLE

STRUGGLES WITH COMPLEX IMAGES (FIXED-SIZE ENCODING)

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

Transformer-based Models

BEST RESULTS AND BETTER SCALABILITY

MASSIVE DATASETS AND COMPLEX TRAINING MODELS

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

Attention mechanisms

IMPROVES ACCURACY AND INTERPRETABILITY

DEPENDS ON RNN'S (HARD TO PARALLELIZE)

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

Our model:
CNN - LSTM + attention

METHODOLOGY: DATA ANALYSIS

THE DATASET

+**82.000** TRAINING IMAGES

+40.000 VALIDATION IMAGES

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

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

MSCOCO

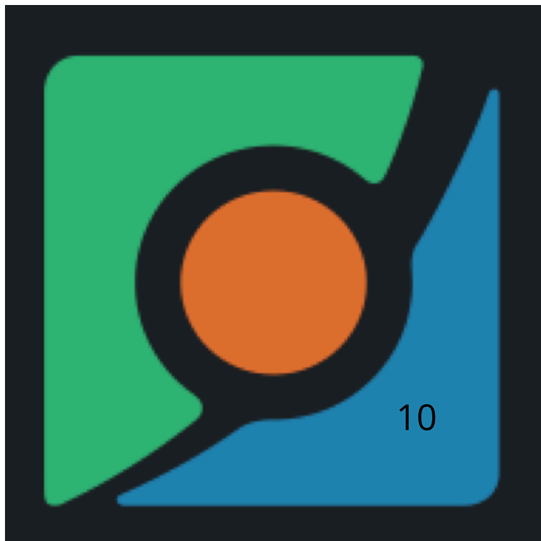
METHODOLOGY: DATA ANALYSIS

PREPROCESSING

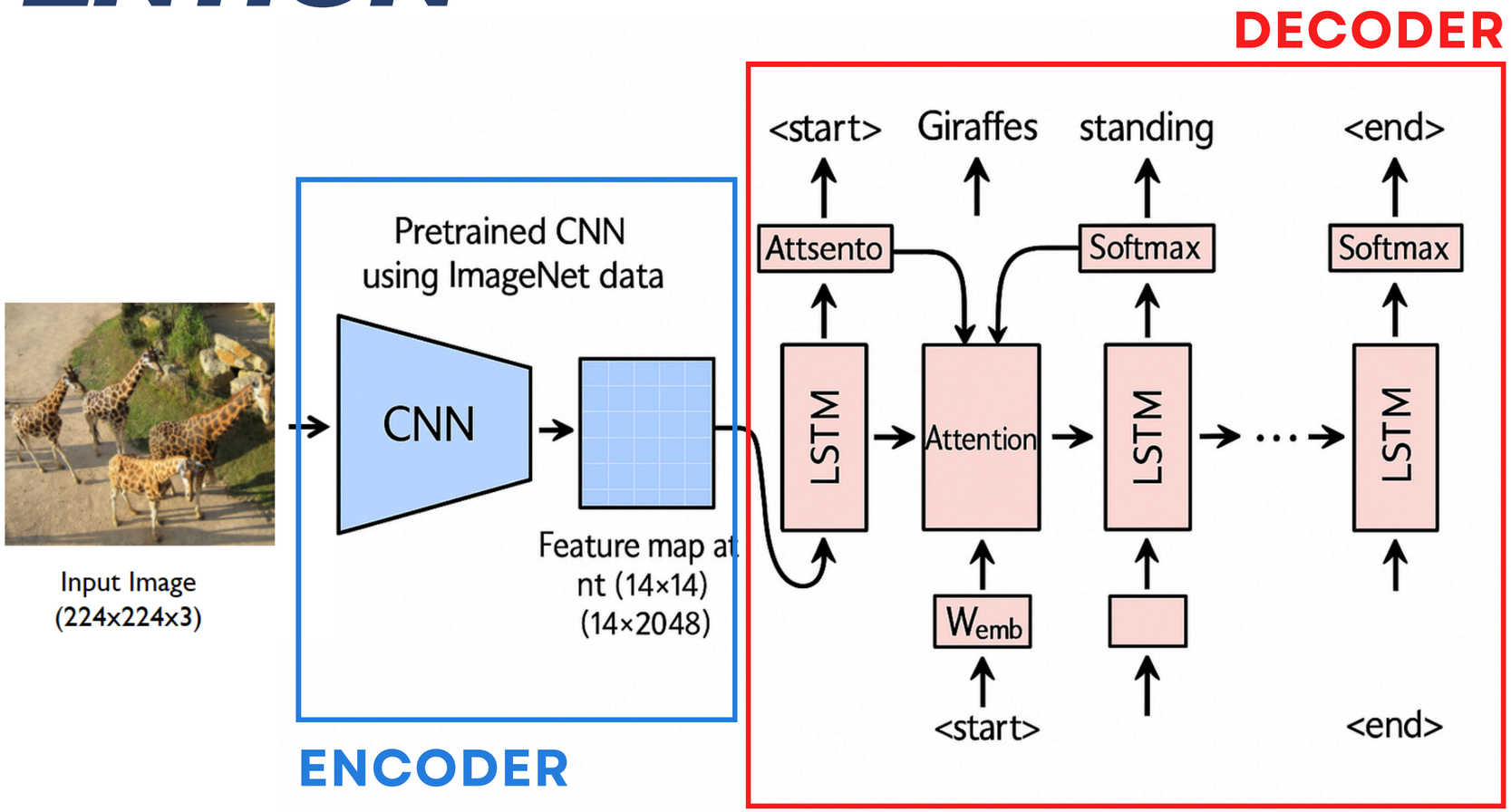
RESIZE

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

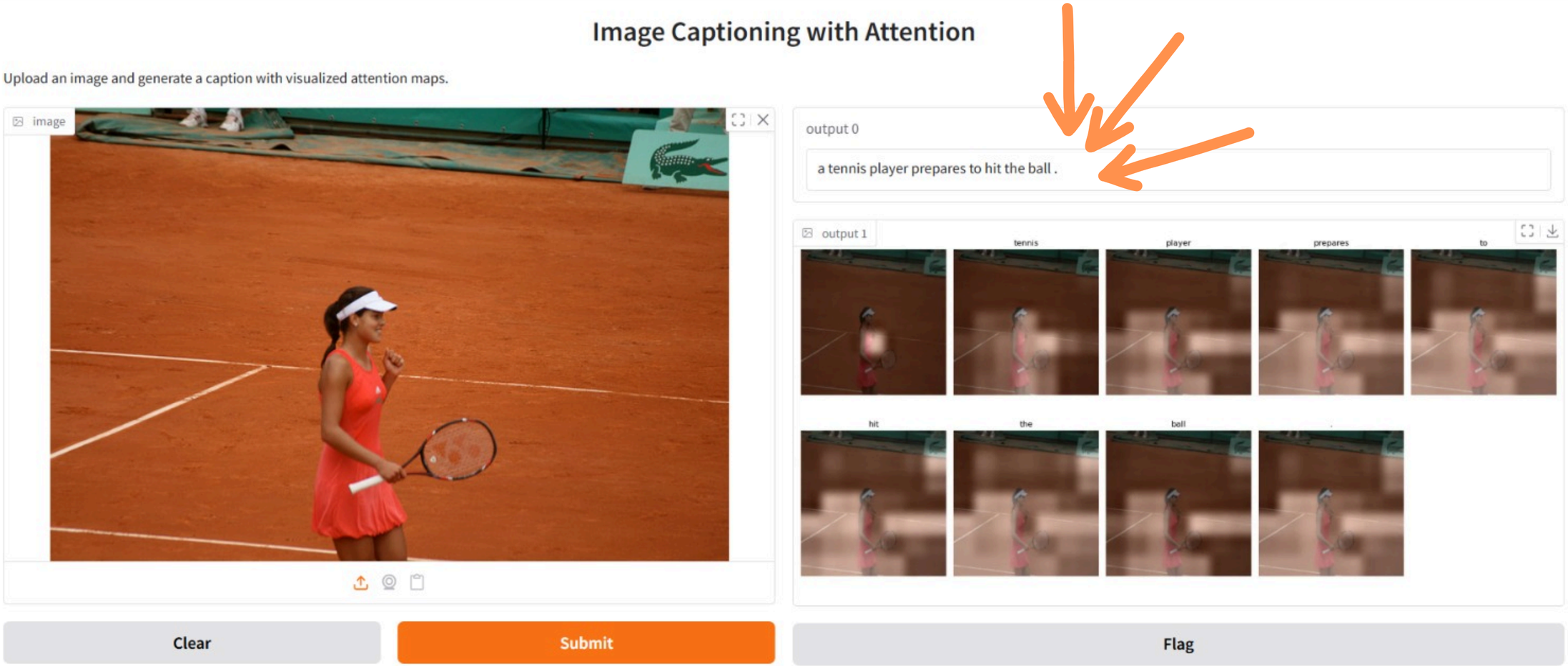
INCREASED ENCODER'S CAPACITY

RESULTS (WITHOUT ATTENTION)

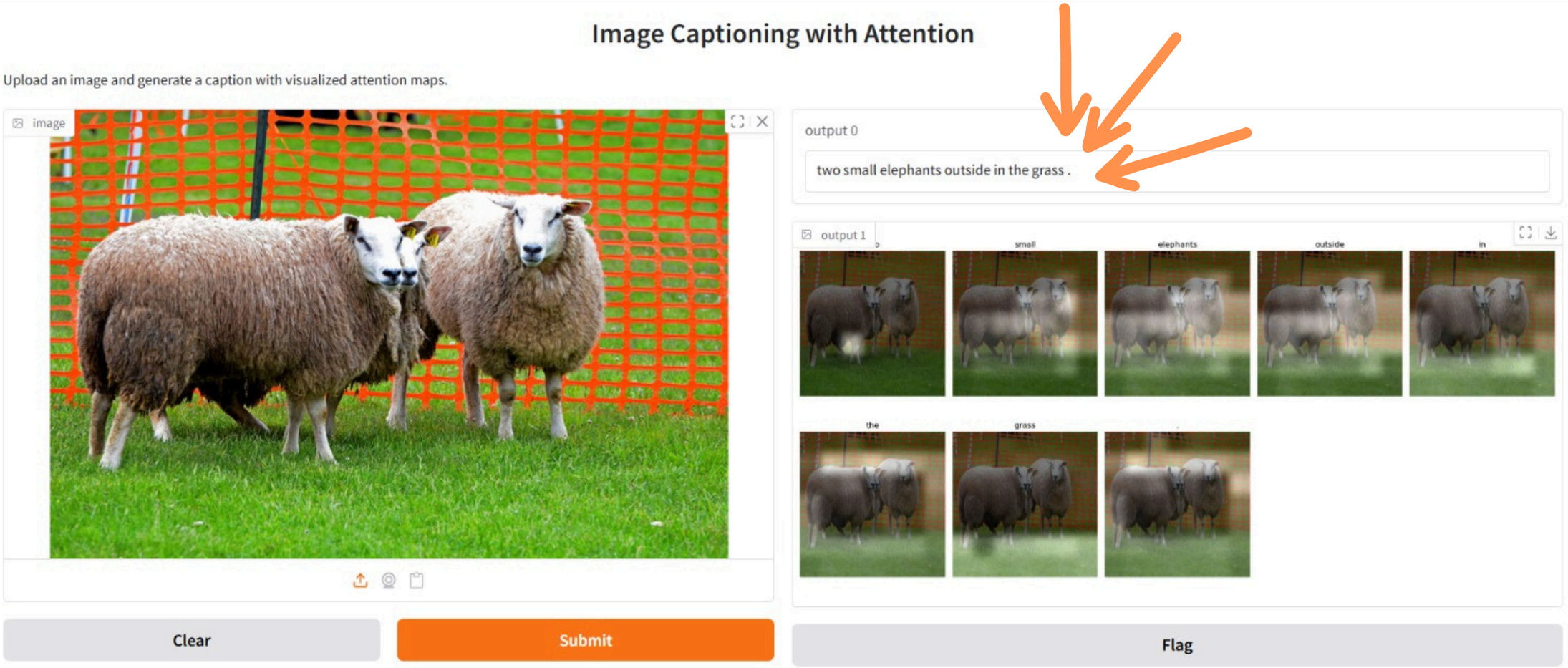
<start> a man in a kitchen with a blender and a bowl of food . <end>

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

FINAL RESULTS W/ATTENTION



FINAL RESULTS W/ATTENTION

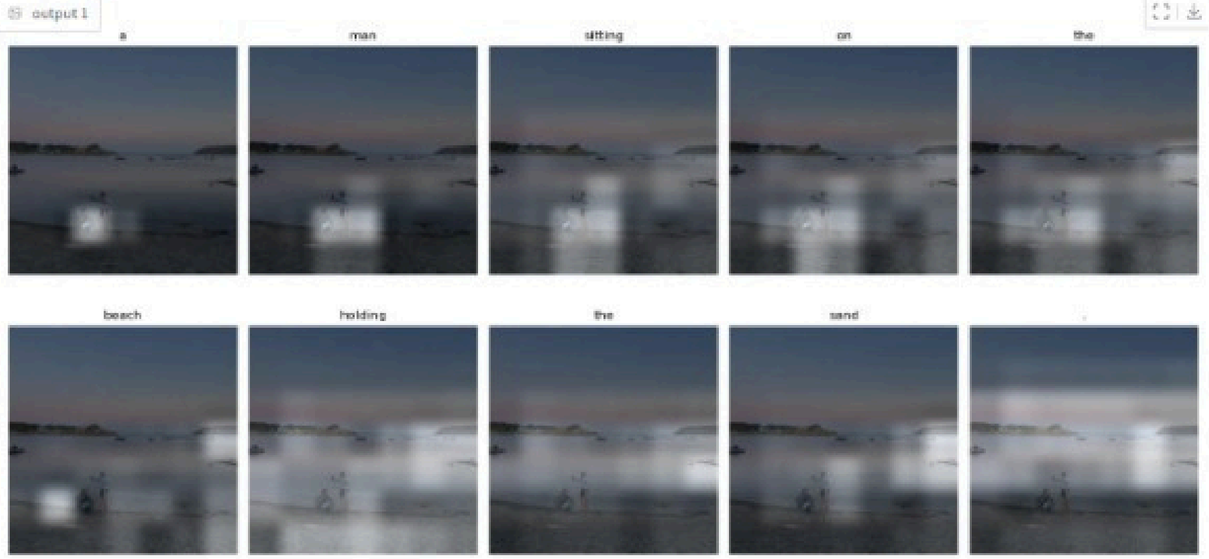


FINAL RESULTS W/ATTENTION

output 0

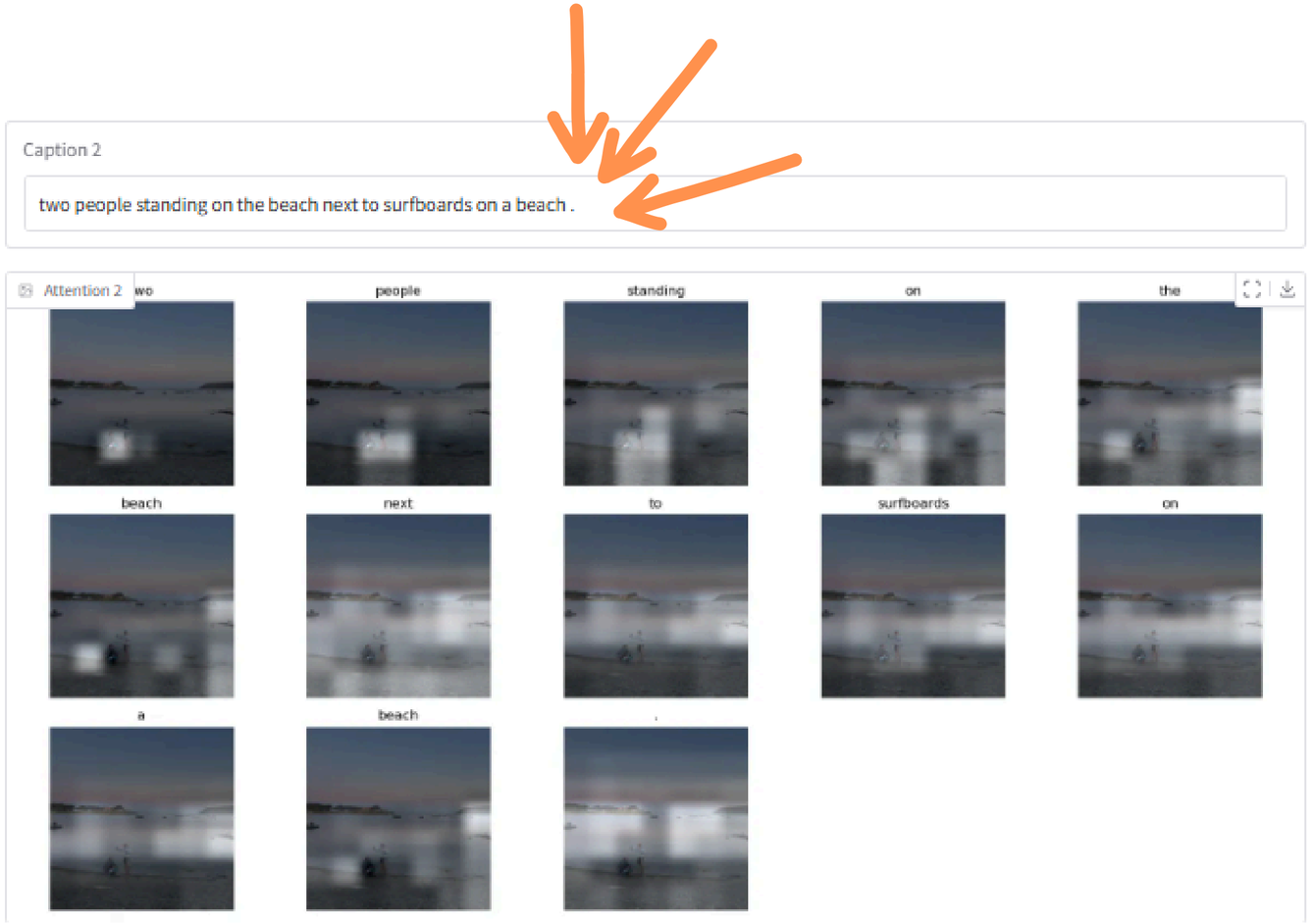
a man sitting on the beach holding the sand .

output 1

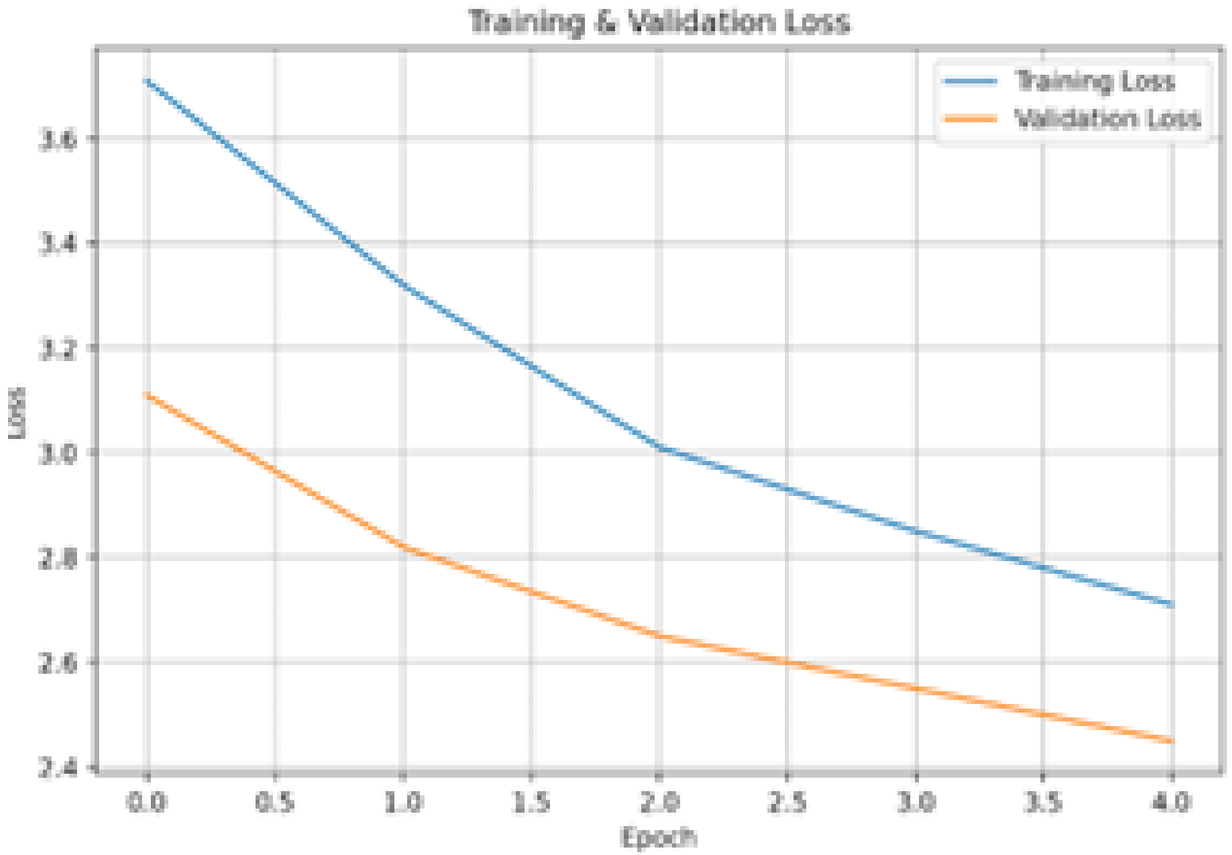


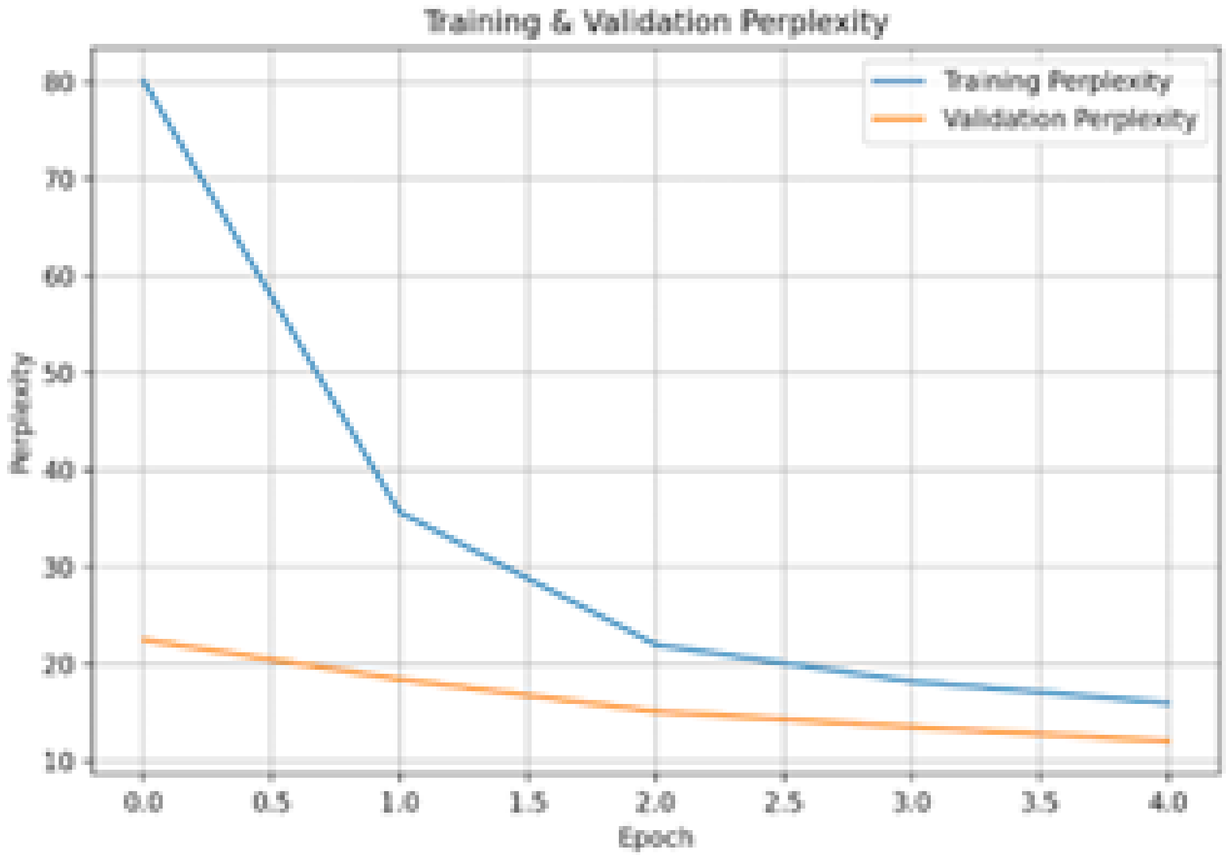
Flag

FINAL RESULTS W/ATTENTION



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 ressources 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)

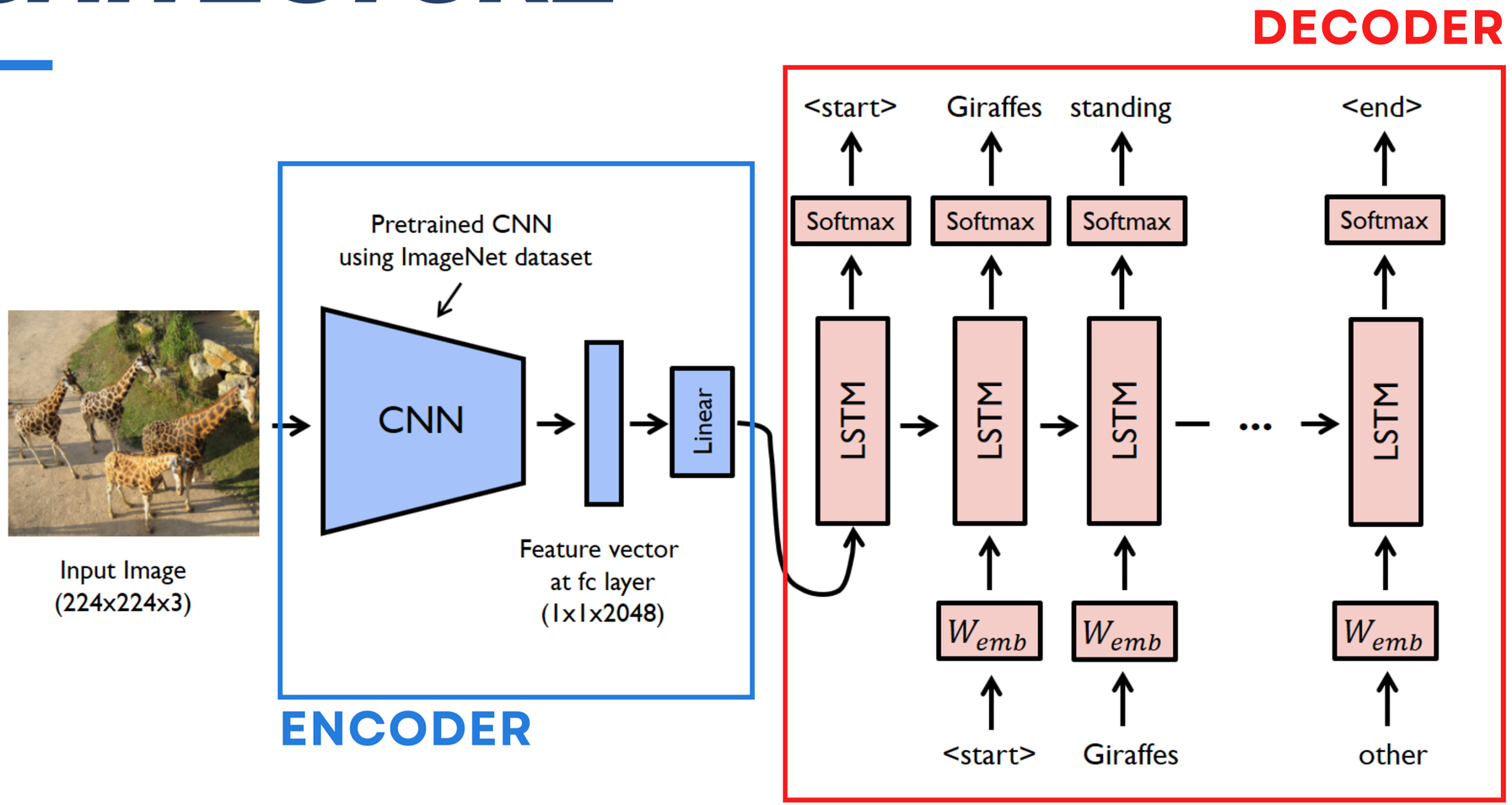
- https://ieeexplore.ieee.org/abstract/document/7508408
- https://hackernoon.com/understanding-architecture-of-lstm-cell-from-scratch-with-code-8da40f0b71f4

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 .