Flickr Image Captioning with VGG19 and Transformer

IST 691 Deep Learning in Practice | Fall 2023

Group 5
Haotian Shen
Lu Guo
Chuan Tse Tsai
Ximeng Deng

Table of Contents

1. Project Overview
2. Data Loading
3. Data Preparation
4. Model Building
5. Custom Training
6. Validation and BLEU Scores
7. Testing and Future Improvements

1.Project Overview

The primary goal of this project is to create a deep learning model that can effectively caption images, a process that entails both recognizing and describing the content of an image in natural language.

Our methodology integrates Convolutional Neural Networks (CNNs) and Transformer, utilizing the VGG19 model for feature extraction and Transformer networks for generating word sequences in captions.

Image Captioning is beneficial for web accessibility. Alternative text enables screen readers to convey visual content to users with visual impairments. Besides, providing precise alt text improves search efficiency and user interaction within image-based content systems.

2. Data Loading

Our dataset is an image caption corpus, consisting of **40,455 captions describing 8,091 images,** aka the Flickr8k Dataset. The images and captions focus on people involved in everyday activities and events.

The dataset was originally published and hosted by Illinois.edu, now available on Jason Brownlee’s GitHub. For our work, we imported this dataset into Google Colab.

3. Data Preparation

A. Load the Captions

B. Caption Standardization and Tokenization

C. Image Feature Extraction

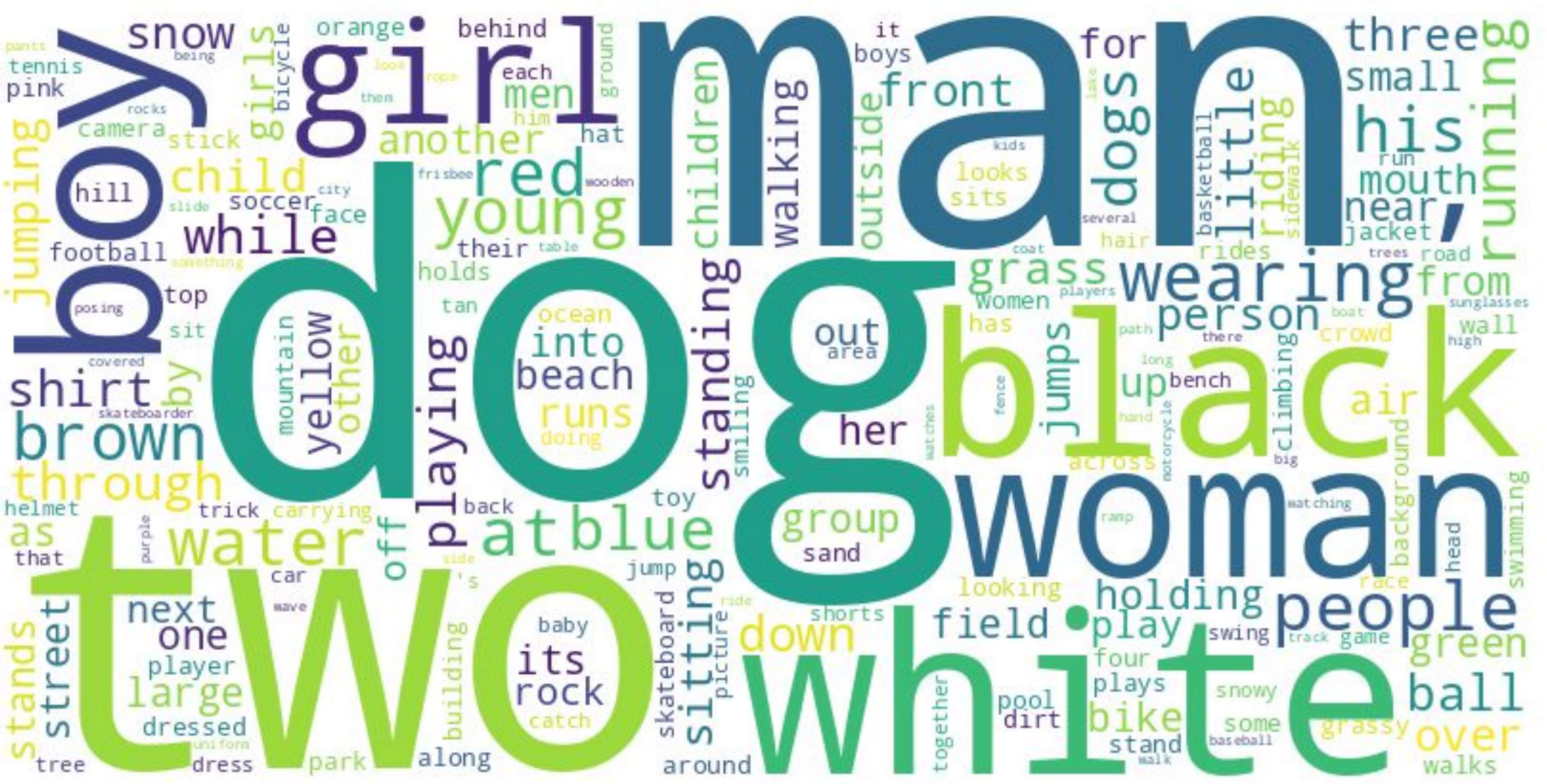
D. Data Splitting to Training and Validation set

E. Optimized TensorFlow Data Preparation Process

Caption Preparation

- Load the ‘Flickr8k.token.txt’ file, iterate over each line in the captions list.
- Split the image ID and the caption text, check for emptiness and incorrect formatting.
- Add [start] and [end] to each caption.
- Standardization: converting to lowercase and removing punctuation.
- The length of the longest caption (number of words) is 38.
  The actual number of unique tokens (words) in our dataset is 8,922.
  Parameters for Tokenizer, MAX_OUTPUT_SEQ_LENGTH = 50, VOCAB_SIZE = 7,000.
- Convert captions to vector sequences.
  Create mappings from words to indices and vice versa.

Captions Word Cloud

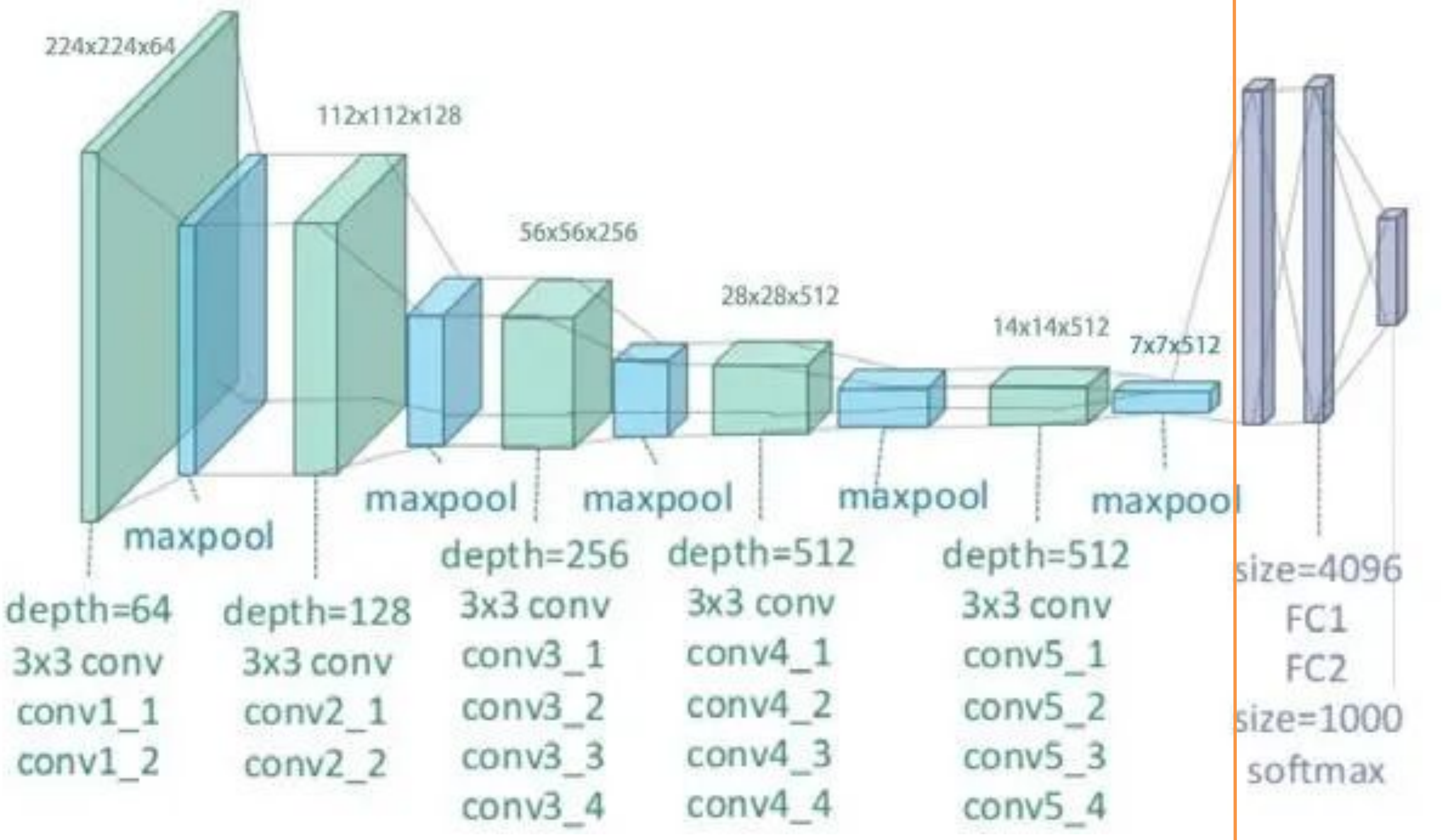


This frequency word cloud provides us the context for subsequent training and testing, better understand what image content we are or are not good at predicting.

Customized stop words include {'[start]', '[end]', '.', 'and', 'a', 'in', 'on', 'the', 'with', 'is', 'are', 'to', 'an', 'of'}

Dog, man, women, boy, girl are the most frequent subject or object nouns. White and black are very common color adjectives.

VGG19



https://www.researchgate.net/figure/llustration-of-the-network-architecture-of-VGG-19-model-conv-means-convolution-FC-means_fig2_325137356

We're using the pre-trained VGG19 model for feature extraction.

Built as a deep CNN, VGG19 is an object-recognition model that supports up to 19 layers (16 conv., 3 fully-connected). It takes a 3-channel color image as input, scales it to 224×224×3 for training and testing.

We set **include_top=False,** the output of the last convolutional block ('block5_pool') is the output for the model. We also did **batch processing** with TensorFlow Dataset for efficiency.

This provided a rich and generalized feature representation of the images.

Other Preparation

- Create training and validation sets using an 80-20 split randomly.

- Optimized TensorFlow Data Preparation Process
  - Load pre-extracted features, and pair them with respective captions.
  - **Shuffling**, Buffer size 1000, larger BUFFER_SIZE means better shuffling but higher memory usage.
  - **Batching**, Batch size 64, image-caption pairs are processed together in one pass.
  - **Pre-fetching**, load data for the next batch while the current batch is being processed, improve efficiency and training speed.

A. Mask
B. Positional Encoding
C. Point-wise Feed Forward Network
D. Encoder
E. Decoder
F. Transformer Model
G. Set Parameters, Initialize model

4.Model Building

Mask

We created the padding mask and the look-ahead mask to ignore the padding tokens in sequences, and to prevent the model from peeking at future tokens in the sequence during training.

Positional Encoding

Provide the model with information about the position of each word in a sequence. For each position, the encoding includes a sine wave and a cosine wave relative to the dimension, where wavelengths decrease geometrically across different dimensions. Then, the generated positional encoding is added to the word embeddings.

Point-wise Feed Forward Network

A key component within the transformer attention blocks, it consists of two dense layers.

Encoder, Decoder, Multi-head Attention Layer

The Multi-Head Attention Layer allows us to simultaneously process different parts of the input sequence to capture various contextual relationships. Each 'head' in this layer independently computes scaled dot-product attention, generating queries, keys, and values from the input to dynamically prioritize different parts of the sequence. The encoder processes the image features and the decoder generates the corresponding captions.

Transformer Model Building

Model Parameters

VOCAB_SIZE = 7000

MAX_LENGTH = 50

num_layer = 2

emb_dim = 256

fc_dim = 1024

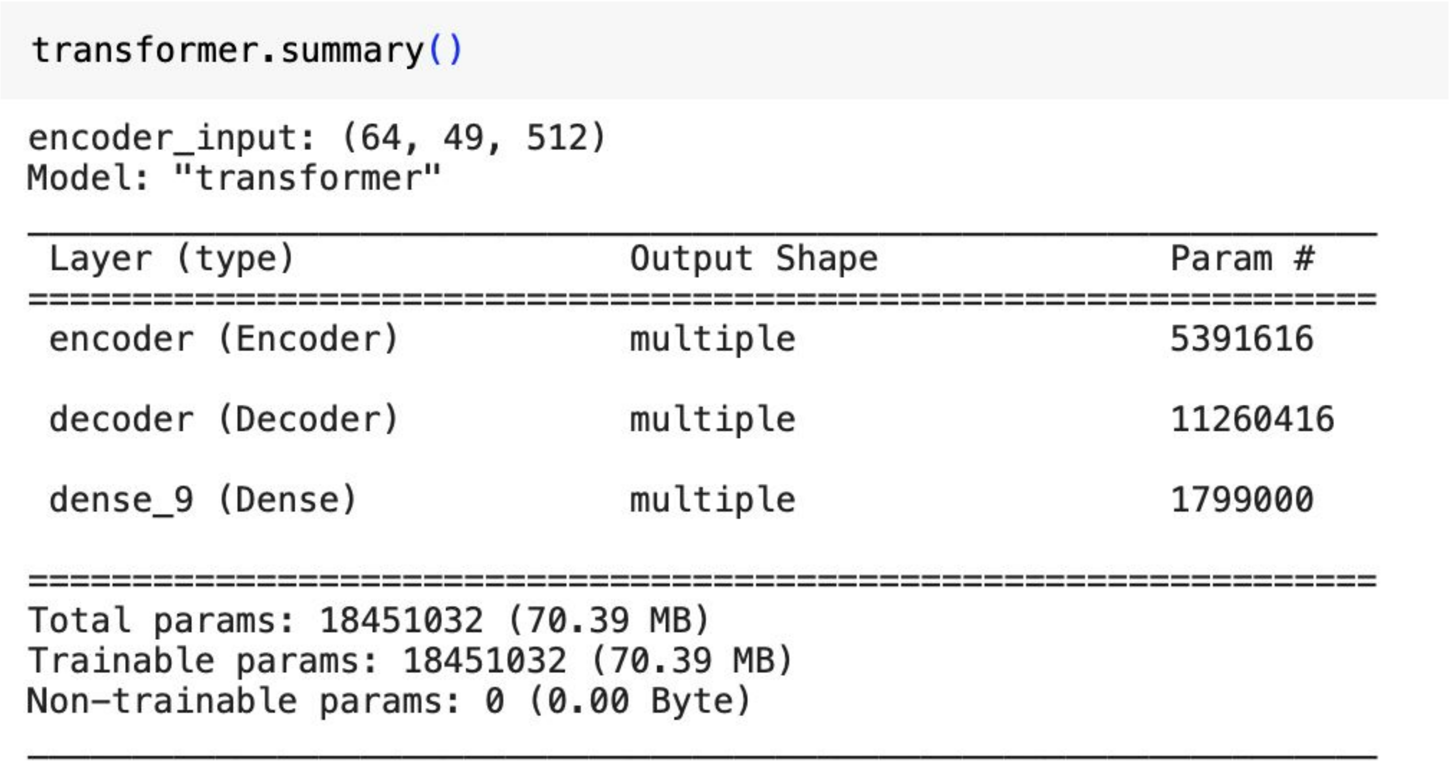
num_heads = 8

row_size = 7

col_size = 7

target_vocab_size = VOCAB_SIZE

dropout_rate = 0.1



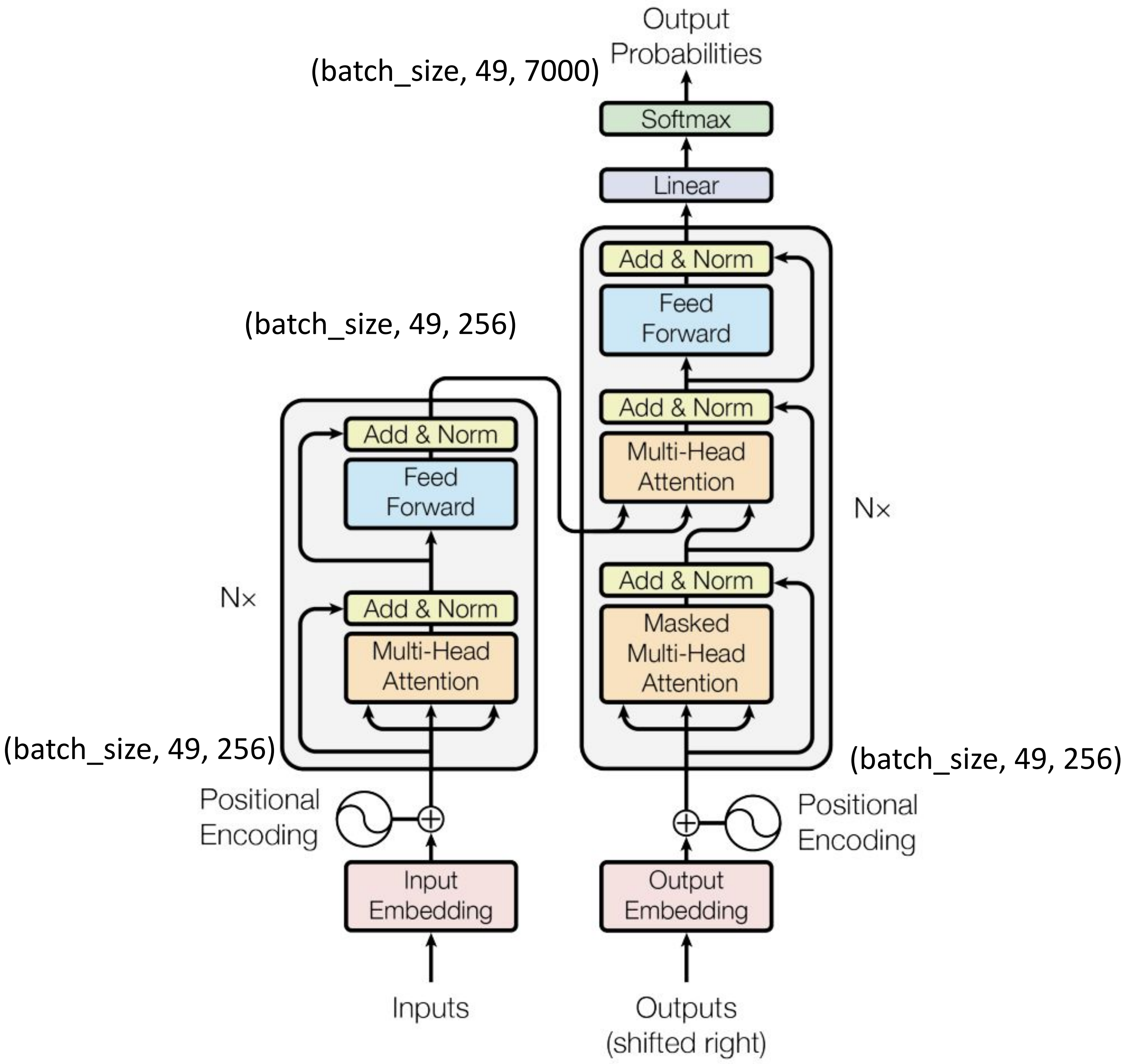
```
transformer.summary()
```

```
encoder_input: (64, 49, 512)
Model: "transformer"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 encoder (Encoder)           multiple                  5391616

 decoder (Decoder)           multiple                  11260416

 dense_9 (Dense)             multiple                  1799000

=================================================================
Total params: 18451032 (70.39 MB)
Trainable params: 18451032 (70.39 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

5. Custom Training

Custom Learning Rate Schedule and Optimizer
Based on the number of training steps, the schedule increases the learning rate linearly for the first warmup_steps training steps, and then decreases proportionally to the inverse square root of the step number. We also sets up the Adam optimizer.

Loss Function
Sparse Categorical Cross Entropy

Checkpoints
Allow the model's state (weights) and optimizer's state to be saved periodically.

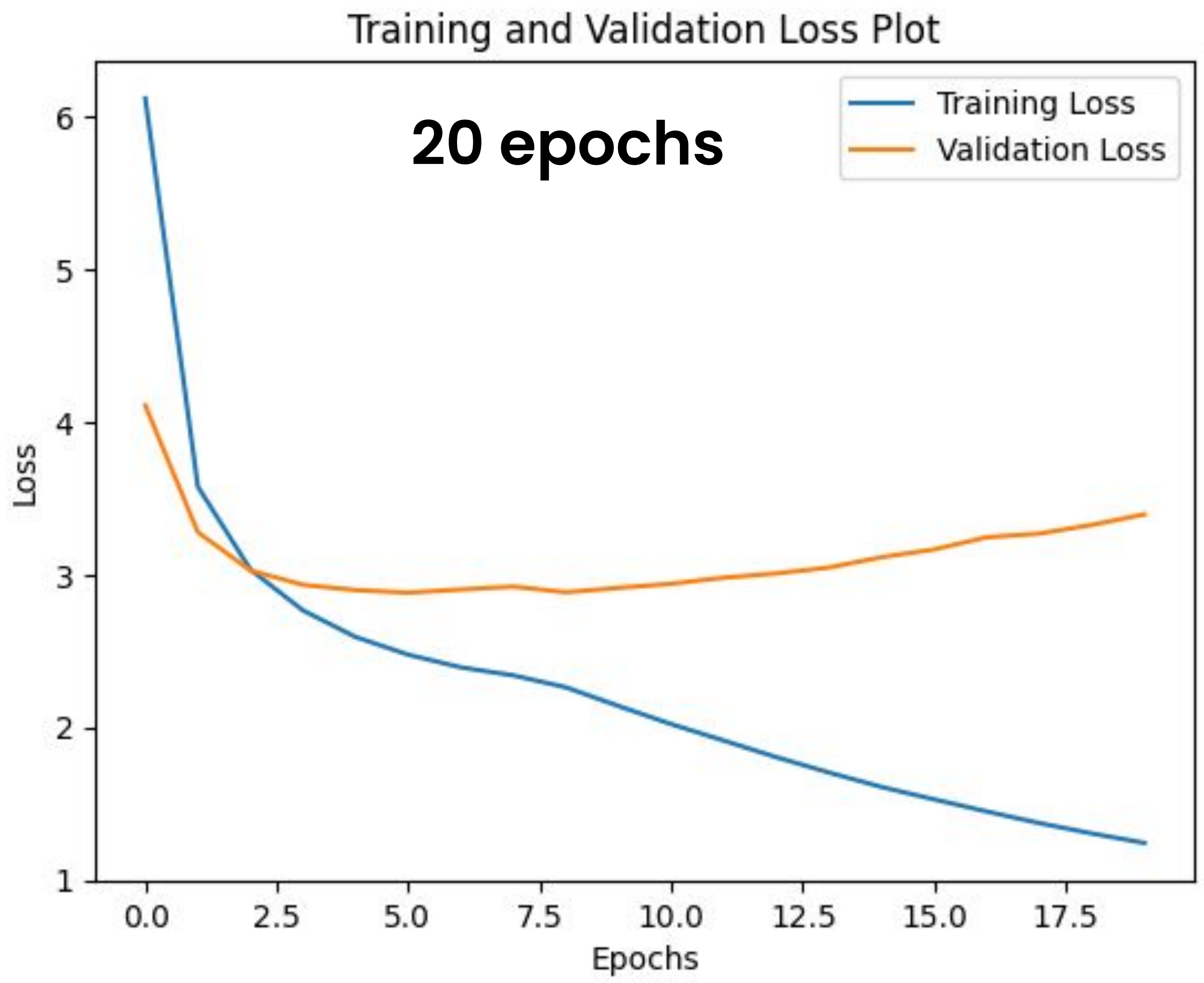
Training and Validation Steps
Iterates for 10 epochs, performs training and validation steps and prints out the losses over time.

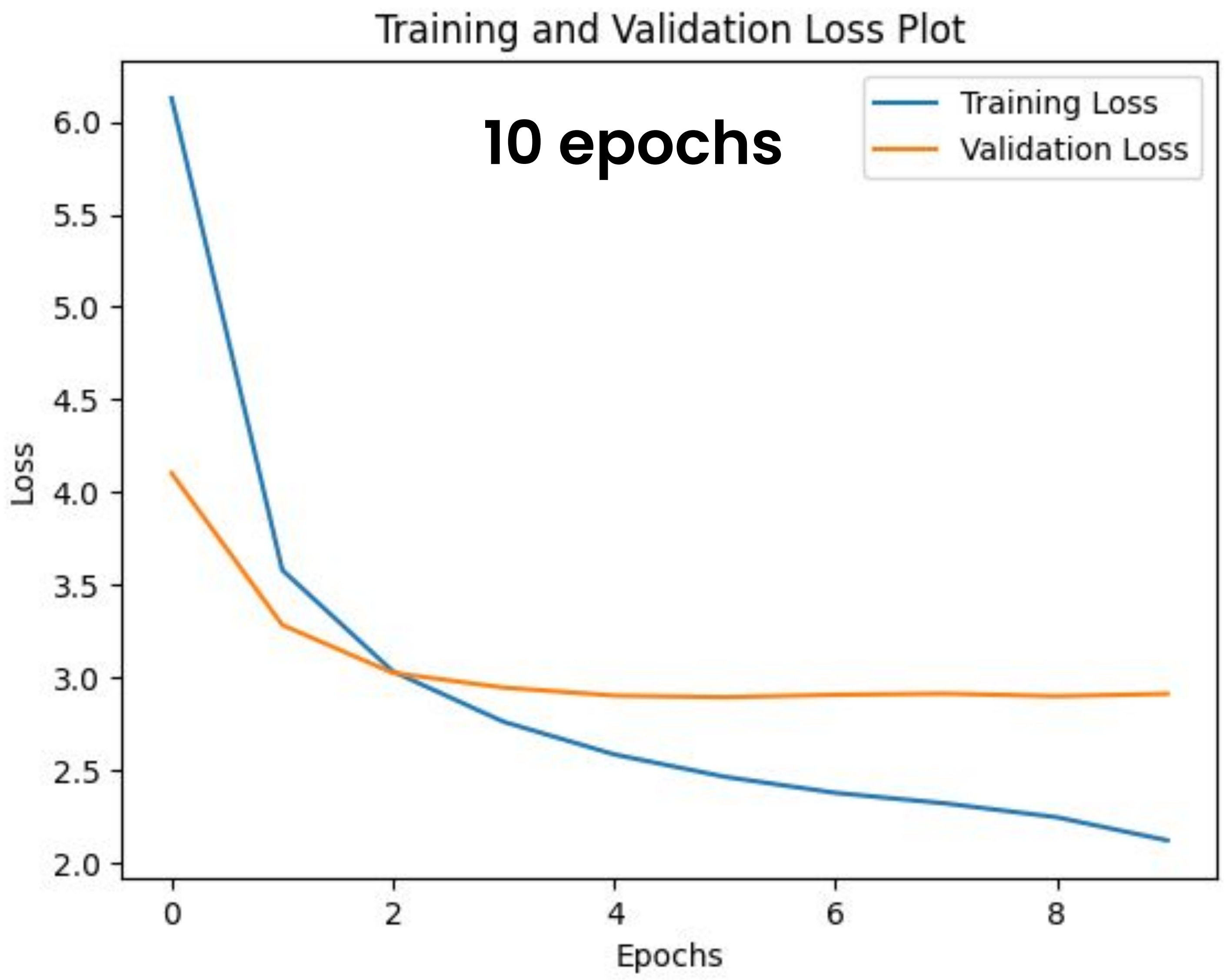
Problems Encountered and Modification

For Positional Encoding, initially, we created only the 1D positional encoding, which is suitable for text sequence. But this may be one of the reasons our model didn't work well at first.

In the context of Transformers applied to images or 2D data, we need 2 version of positional encoding, 1 for Encoder (spatial features) and 1 for Decoder (generate sequence). This dual encoding approach enhances the model's capability to handle the context of both image and text data.

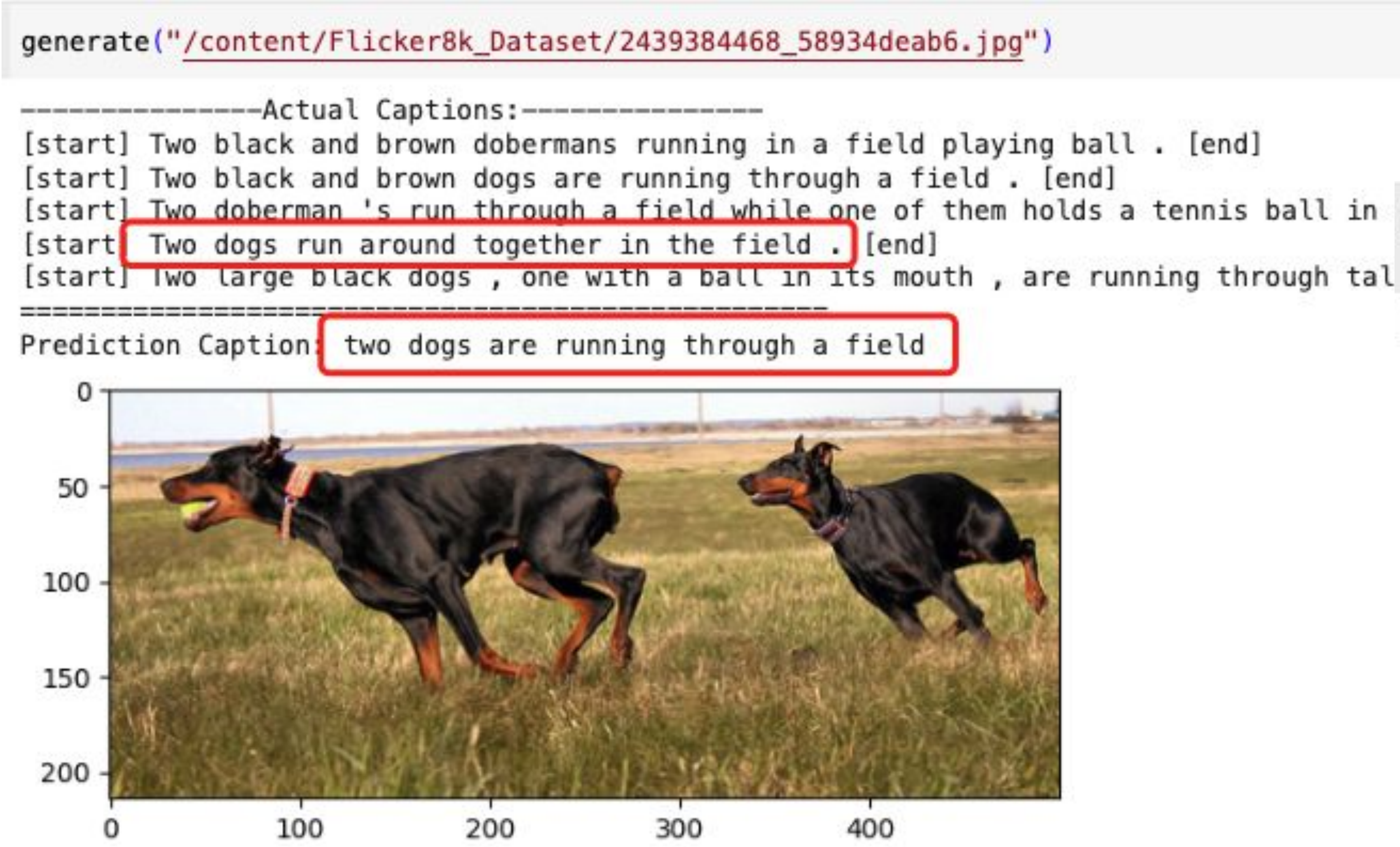
Train and Val Loss





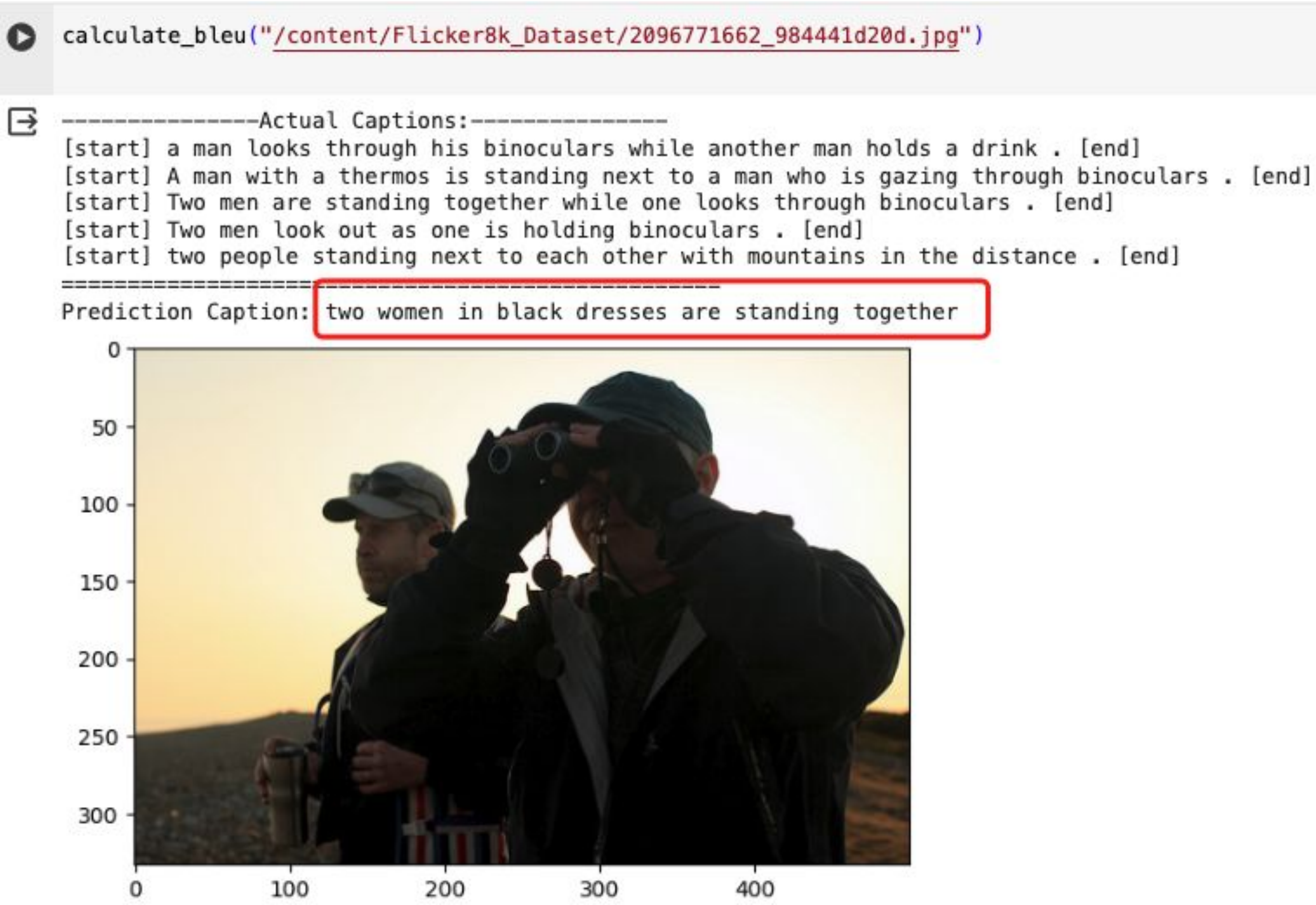
As training begins, the train loss drops quickly.
We decided to **train 10 epochs** because after this point, although the Train loss is still decreasing, the Validation loss starts to increase.

6.Validation and BLEU



```
generate("/content/Flicker8k_Dataset/2439384468_58934deab6.jpg")
```

```
---------------Actual Captions:---------------
[start] Two black and brown dobermans running in a field playing ball . [end]
[start] Two black and brown dogs are running through a field . [end]
[start] Two doberman 's run through a field while one of them holds a tennis ball in
[start] Two dogs run around together in the field . [end]
[start] Two large black dogs , one with a ball in its mouth , are running through tal
==================================================
Prediction Caption: two dogs are running through a field
```



```
calculate_bleu("/content/Flicker8k_Dataset/2096771662_984441d20d.jpg")
```

```
---------------Actual Captions:---------------
[start] a man looks through his binoculars while another man holds a drink . [end]
[start] A man with a thermos is standing next to a man who is gazing through binoculars . [end]
[start] Two men are standing together while one looks through binoculars . [end]
[start] Two men look out as one is holding binoculars . [end]
[start] two people standing next to each other with mountains in the distance . [end]
==================================================
Prediction Caption: two women in black dresses are standing together
```

About BLEU

The BLEU score is an algorithm that evaluates the quality of machine-translated text (Marie, 2022).

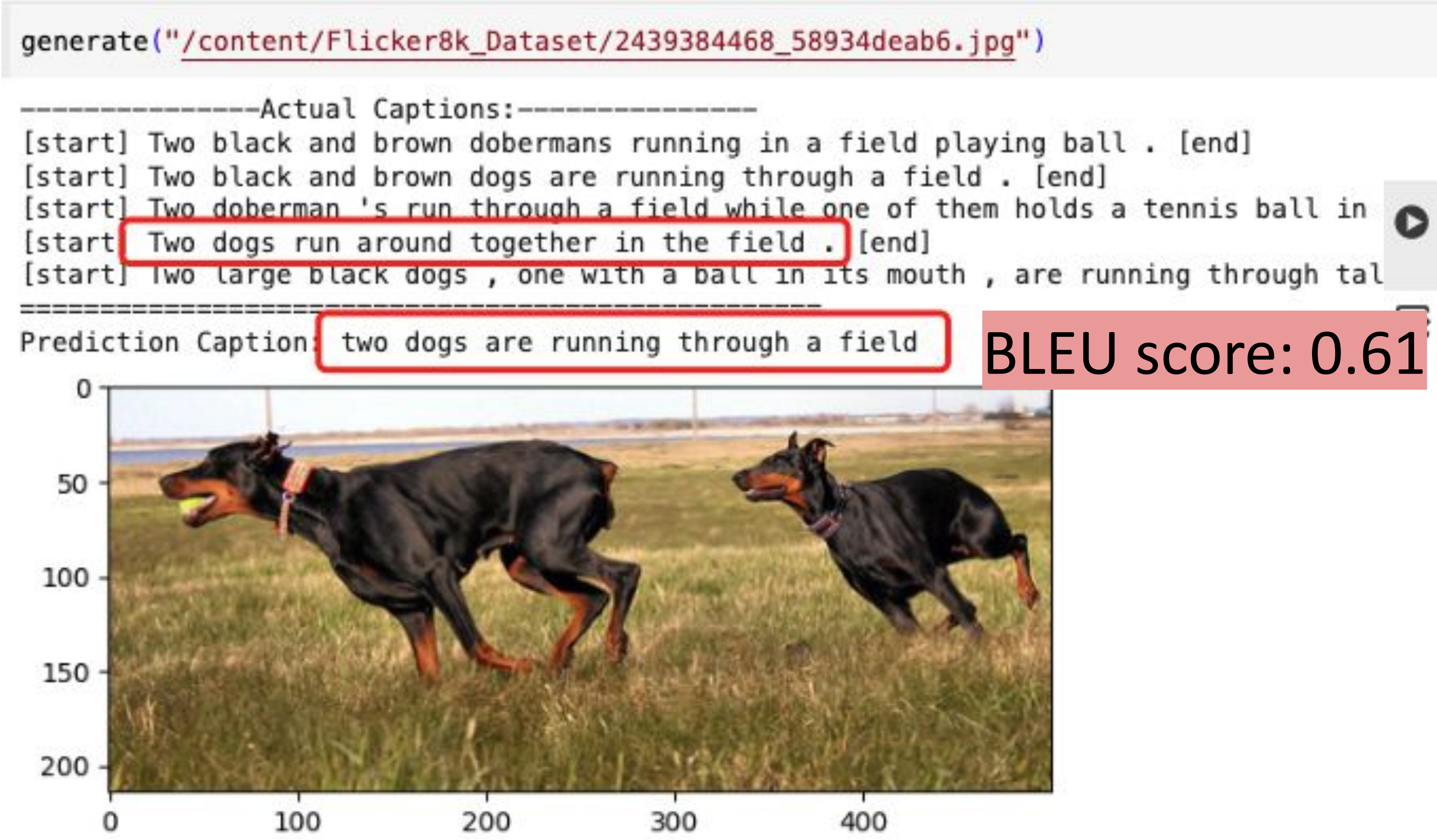
The closer, the better.

It compares one machine generated sentence with several candidate sentences.

Range from 0 to 1. the bigger, the better.

Marie, B. (2022, November 5). BLEU: A Misunderstood Metric from Another Age. Medium. https://towardsdatascience.com/bleu-a-misunderstood-metric-from-another-age-d434e18f1b37.

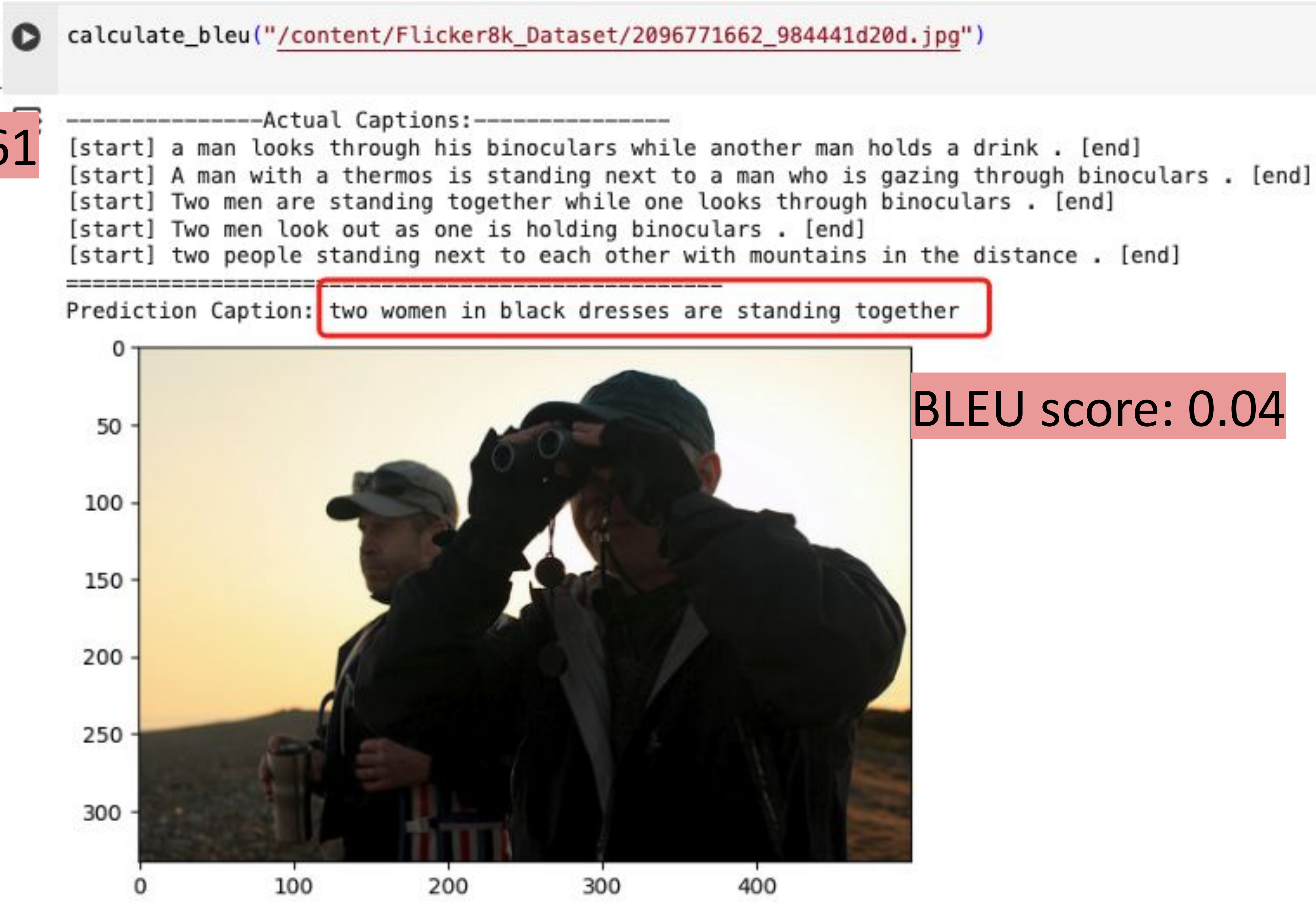
Validation and BLEU



```
generate("/content/Flicker8k_Dataset/2439384468_58934deab6.jpg")
```

```
---------------Actual Captions:---------------
[start] Two black and brown dobermans running in a field playing ball . [end]
[start] Two black and brown dogs are running through a field . [end]
[start] Two doberman 's run through a field while one of them holds a tennis ball in
[start] Two dogs run around together in the field . [end]
[start] Two large black dogs , one with a ball in its mouth , are running through tal
==================================================
Prediction Caption: two dogs are running through a field
```

BLEU score: 0.61



```
calculate_bleu("/content/Flicker8k_Dataset/2096771662_984441d20d.jpg")
```

```
---------------Actual Captions:---------------
[start] a man looks through his binoculars while another man holds a drink . [end]
[start] A man with a thermos is standing next to a man who is gazing through binoculars . [end]
[start] Two men are standing together while one looks through binoculars . [end]
[start] Two men look out as one is holding binoculars . [end]
[start] two people standing next to each other with mountains in the distance . [end]
==================================================
Prediction Caption: two women in black dresses are standing together
```

BLEU score: 0.04

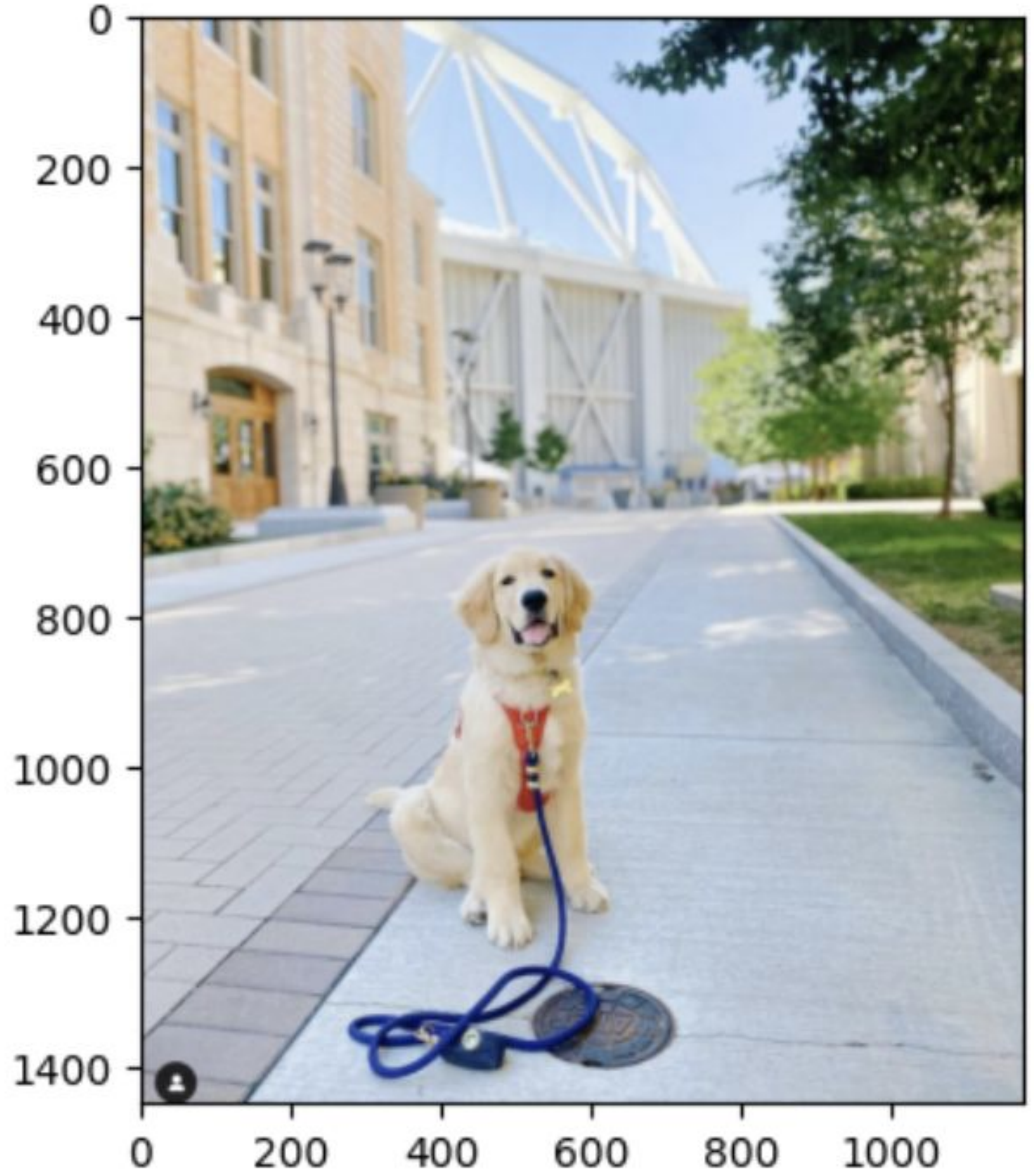
BLEU scores of validation set:

- Max BLEU score: 0.84
- Min BLEU score: 0
- Average BLEU score: 0.11

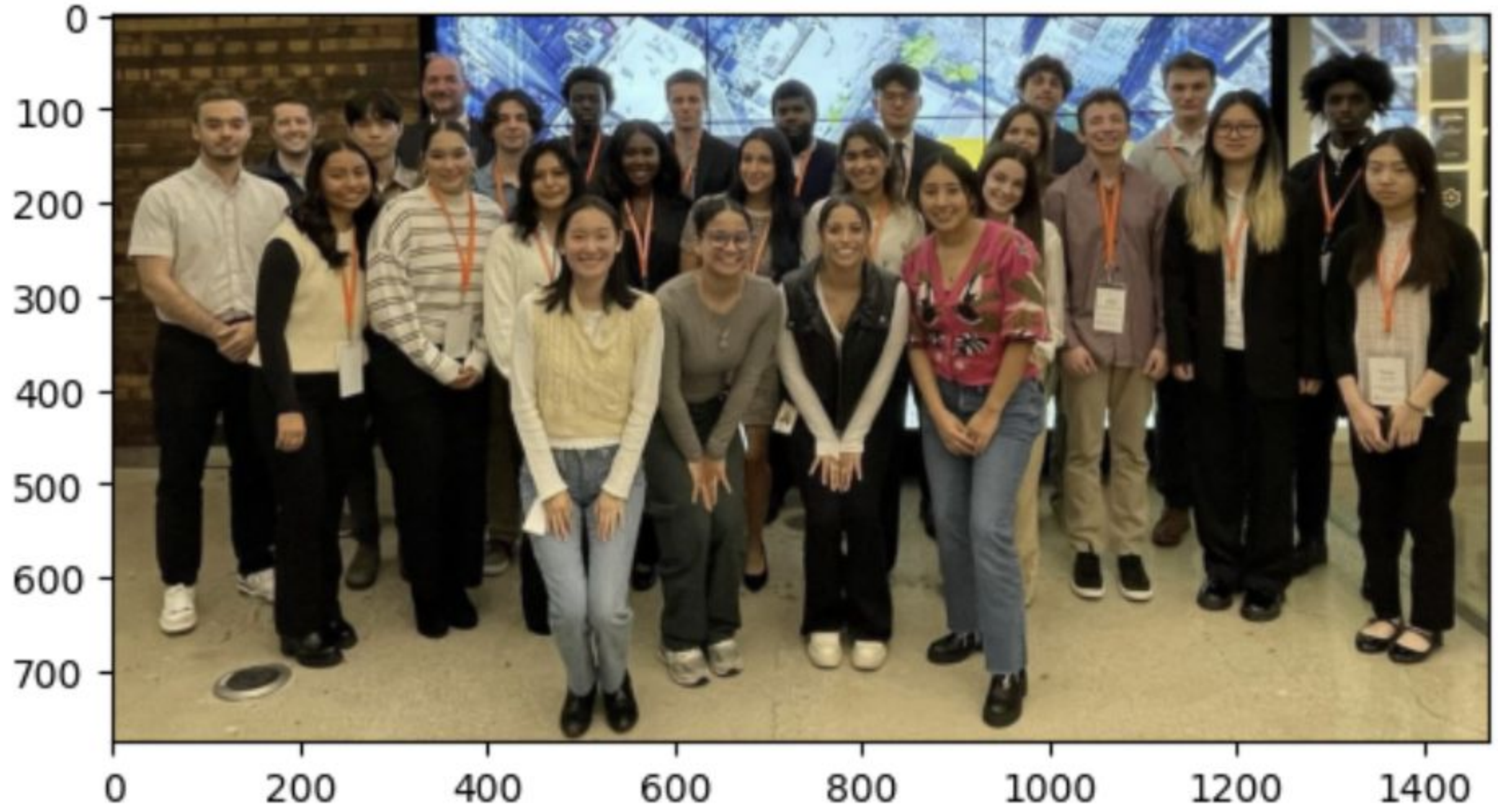
Prediction

images from Instagram
@syracuseu and @ischoolsu

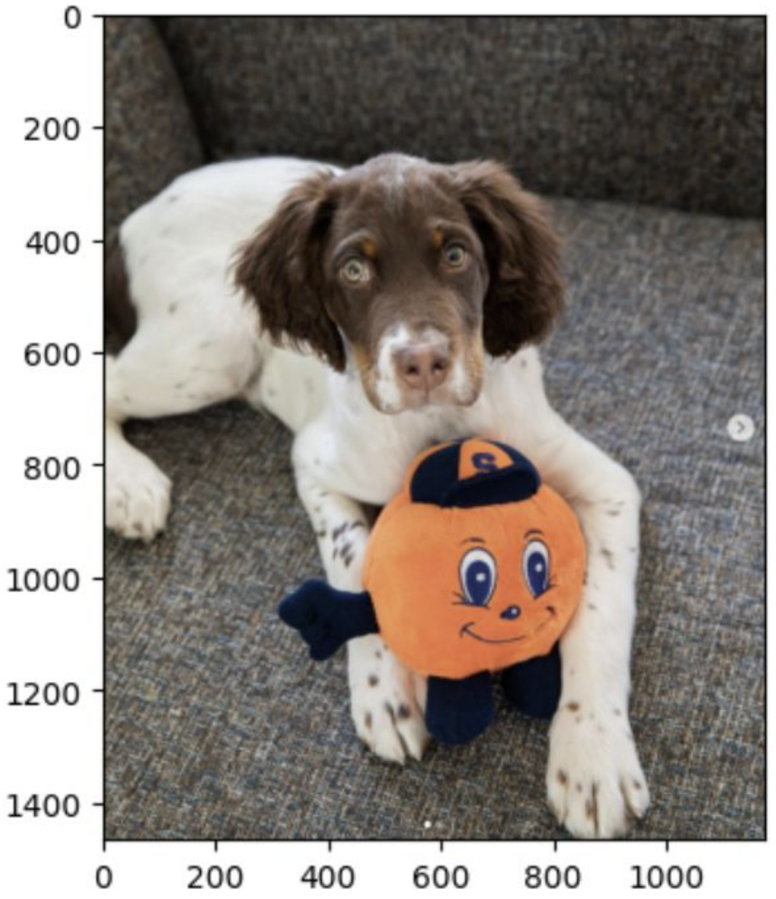
```
==================================================
Prediction Caption: a small dog is running on a sidewalk
```



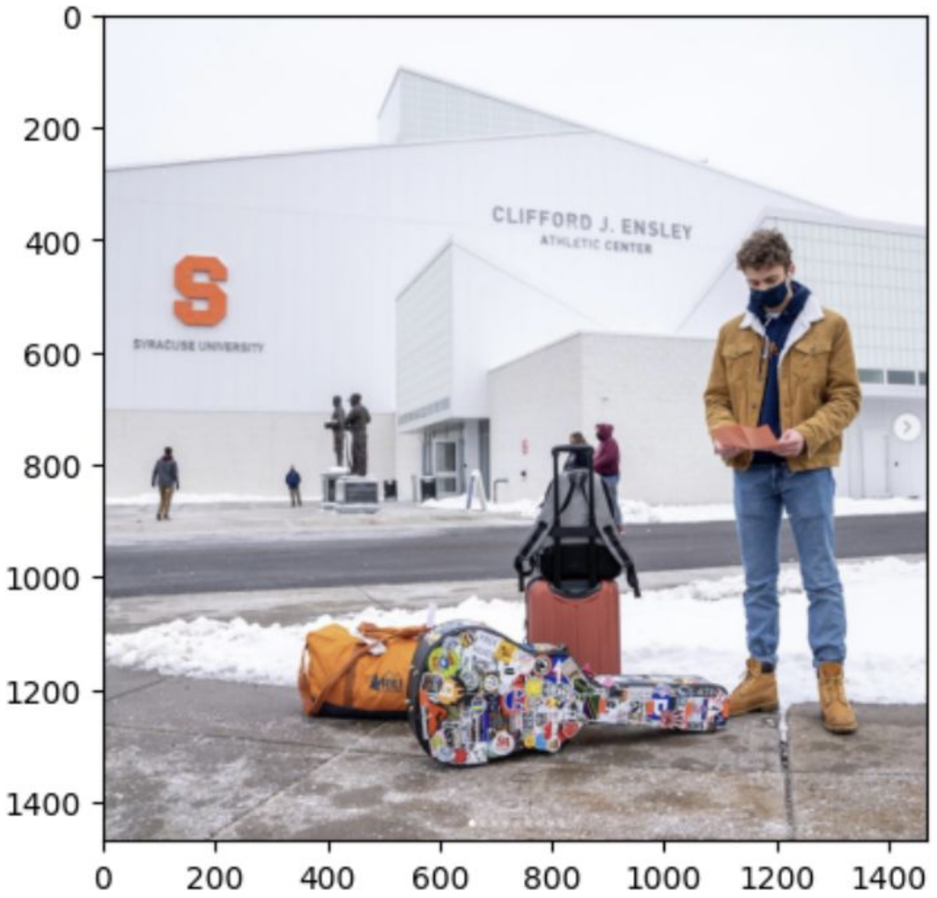
```
==================================================
Prediction Caption: a group of people are gathered around a large group of people
```



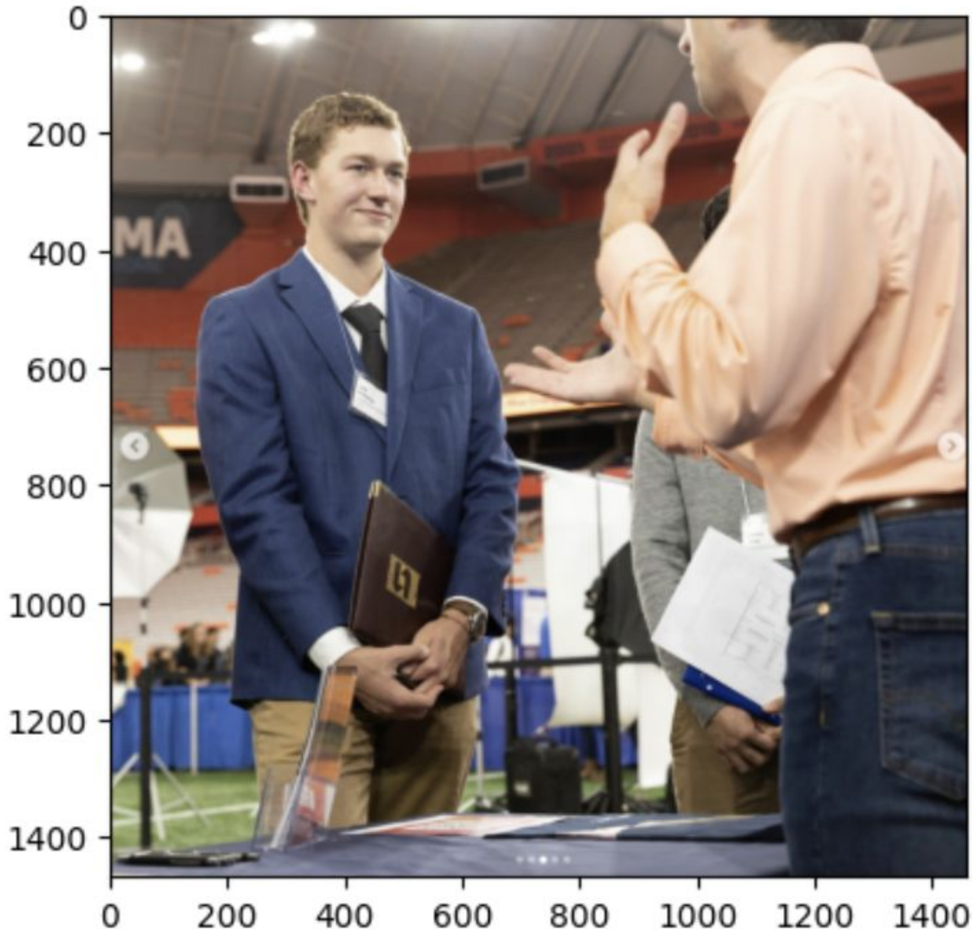
```
==================================================
Prediction Caption: a dog is running with a red ball in its mouth
```



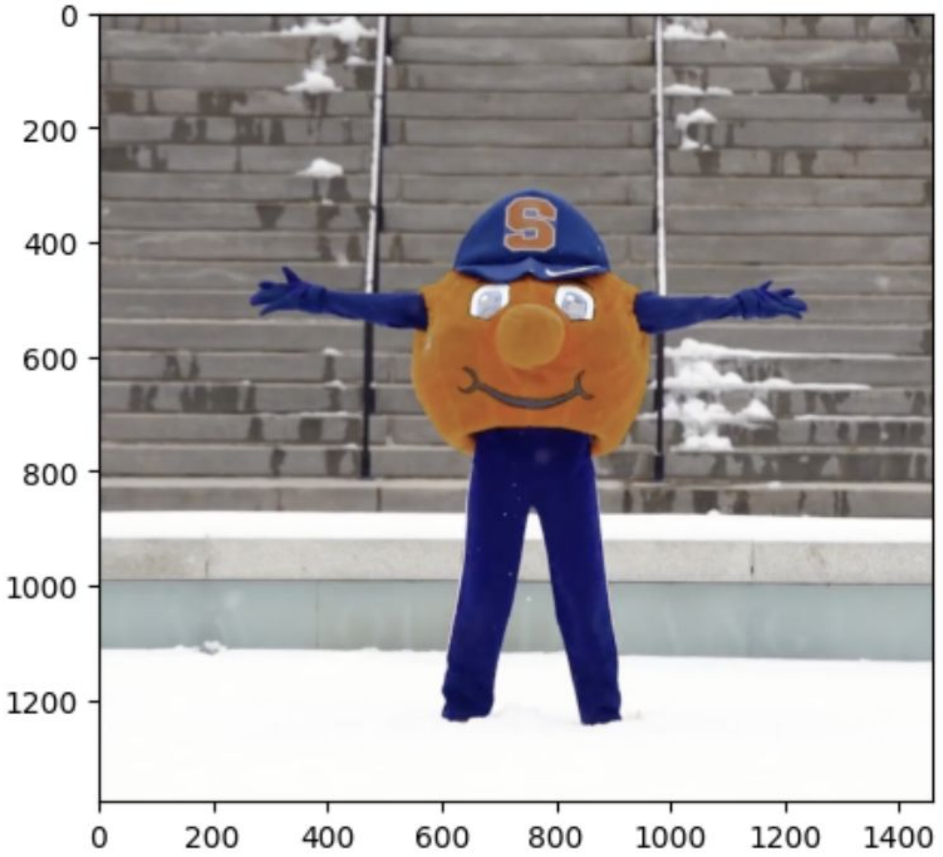
```
==================================================
Prediction Caption: two men are walking on a snowy road
```



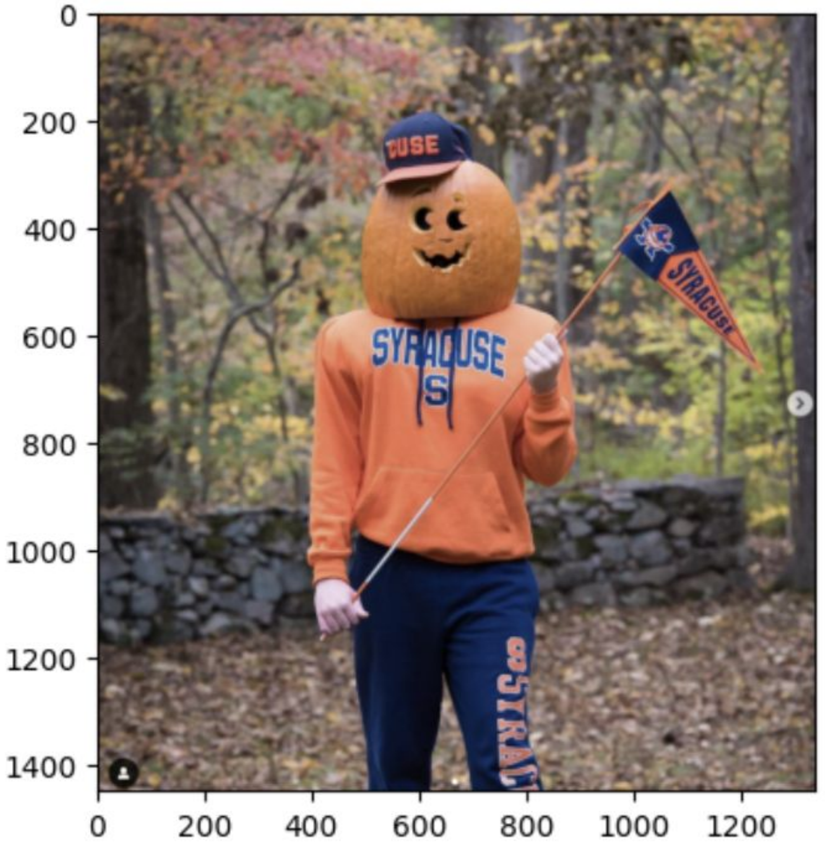
```
==================================================
Prediction Caption: two men are standing in a crowded area
```



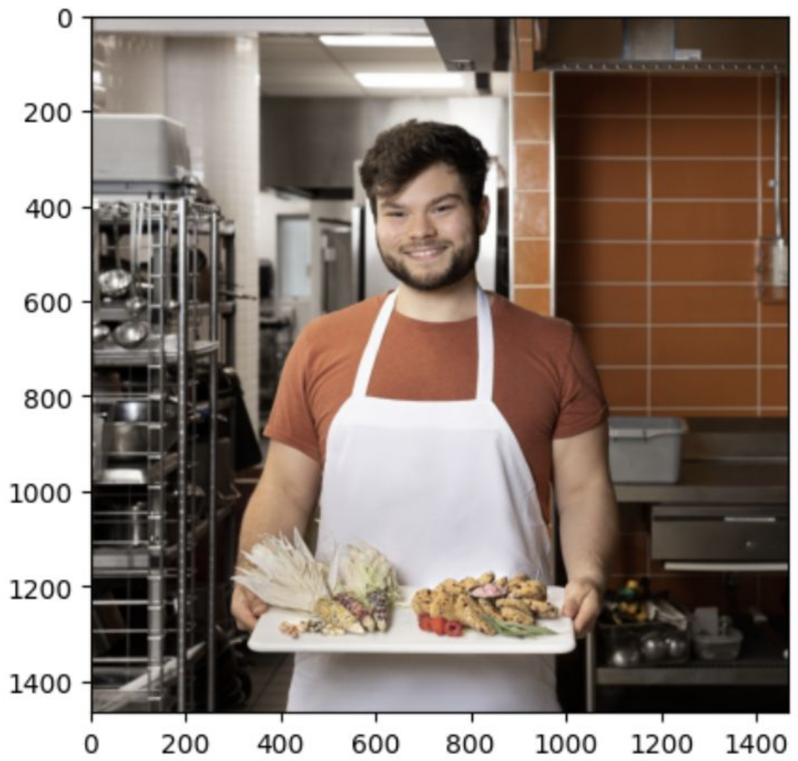
```
==================================================
Prediction Caption: a boy in a red jacket is standing in a snow covered area
```



```
==================================================
Prediction Caption: a man in a red jacket is holding a baseball bat
```



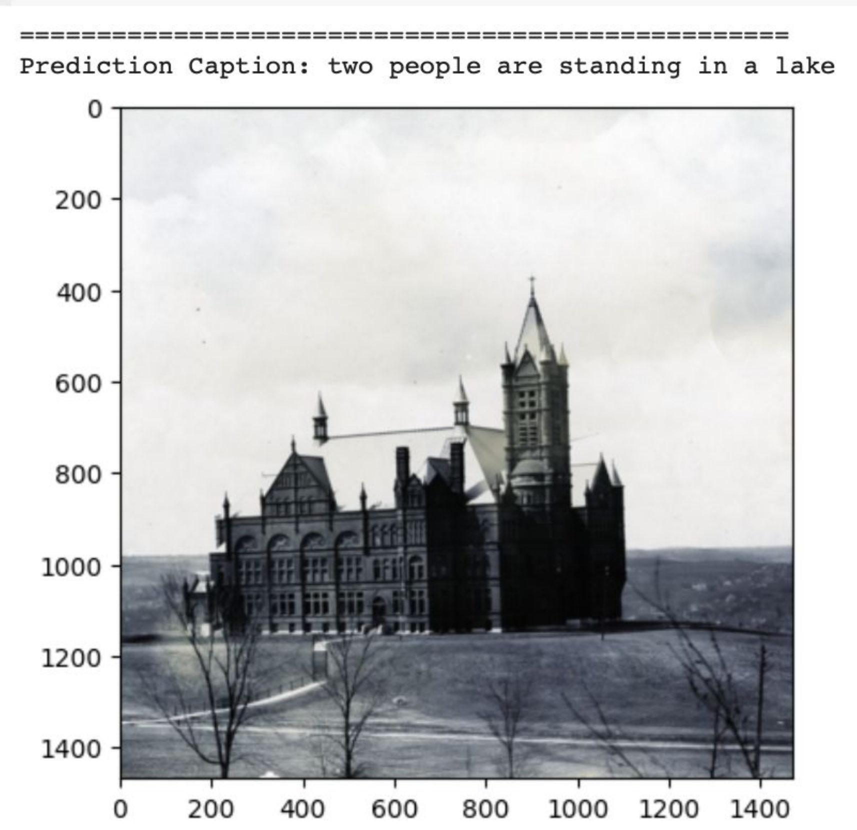
```
==================================================
Prediction Caption: a man in a black shirt is sitting in a restaurant
```



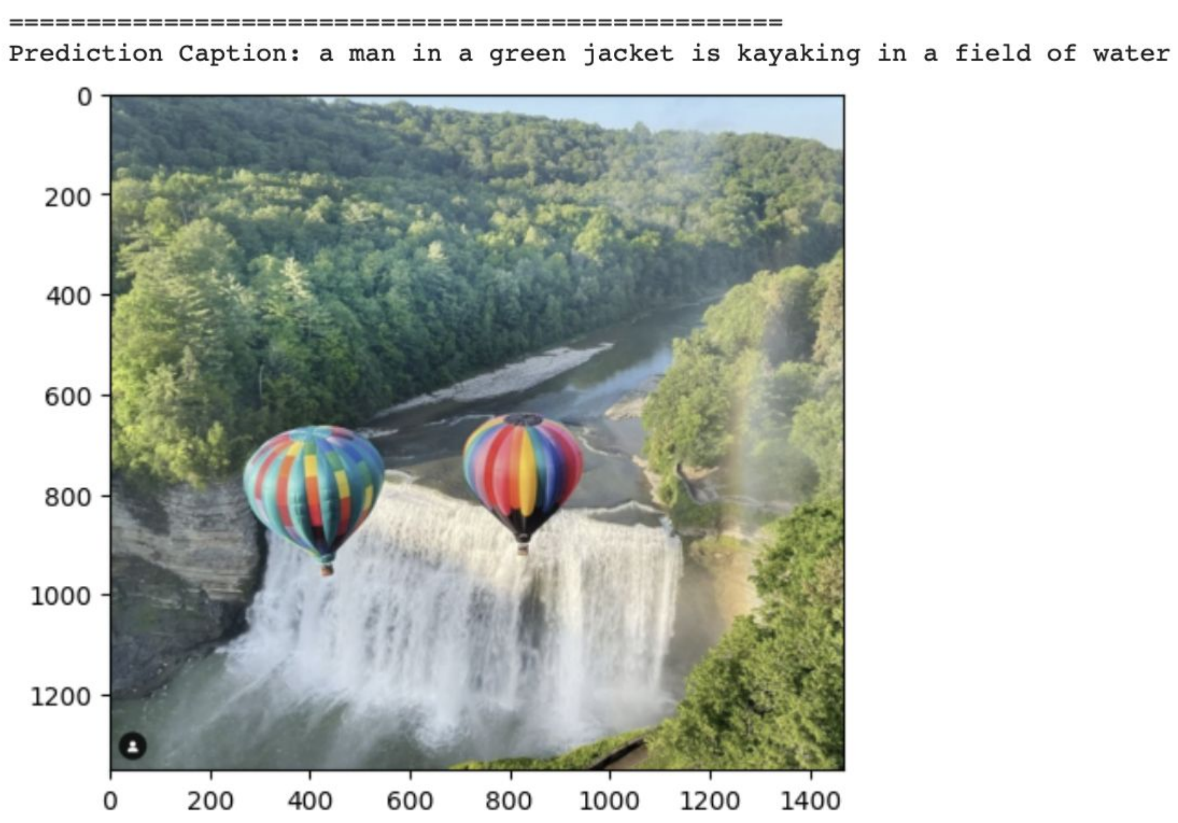
7. Future Improvements

Our model is not good at predicting architecture, landscape, or images without human or animals. At the same time, our current model tends to misidentify animals such as cats or horses as dogs.

Our project's current training set, consisting of ~8,000 images and ~40,000 sentences, is fairly limited in scope. To elevate the model's accuracy, future training projects could benefit from incorporating a larger dataset, with a wider variety of objects. It will necessitate increased computational capacity.



```
===================================================
Prediction Caption: two people are standing in a lake
```



```
===================================================
Prediction Caption: a man in a green jacket is kayaking in a field of water
```

Thanks !

Any Questions