Attention Please! Deep-dive into Image Captioning Efficacy

Kevin Kang, Ethan Polley, Mithun Ramesh, Siddharth Somasi

BROWN

CSCI 1430

Motivation

The success of seminal research paper “Attention is All You Need” has highlighted the power of self-attention mechanisms in language modeling tasks, and has spurred a wave of research into transformer-based models. As companies like OpenAI continue to push the boundaries of what's possible with generative text models, it's clear that the race is far from over.

Image captioning is a challenging task in computer vision that requires a deep understanding of both visual and linguistic content. Our group aims to we aim to investigate and compare the performance of different transformer-based models for image captioning.

Problem

Foundational Models: *Too Big to Fail?*

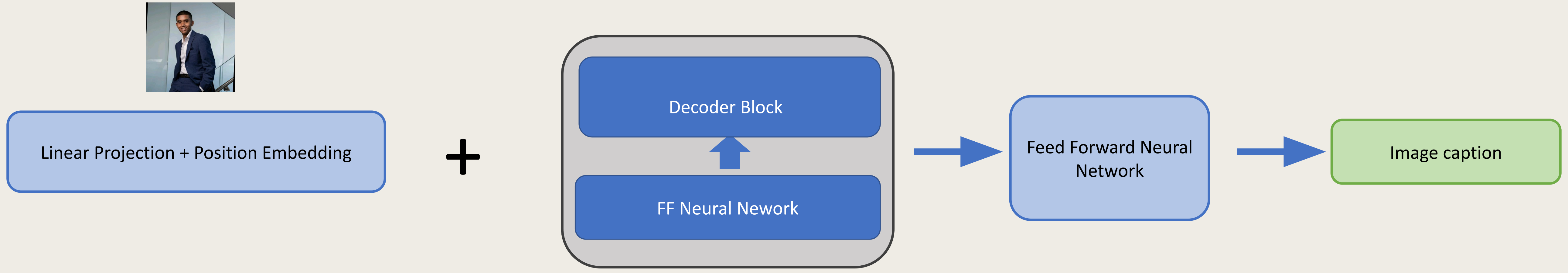
While image extractors like VGG19, ResNet, and InceptionX have achieved impressive results on a wide range of image recognition tasks, there is a growing concern that their generalized nature may lead to blind spots in the AI. This lack of specificity poses a significant problem for image captioning, as it can limit the model's ability to accurately describe and contextualize images. Our group aim to investigate this issue by comparing efficacy of un-tuned base image captioning models relative to a fine-tuned an image captioning model on a specific dataset.

Goal

Our Action Items

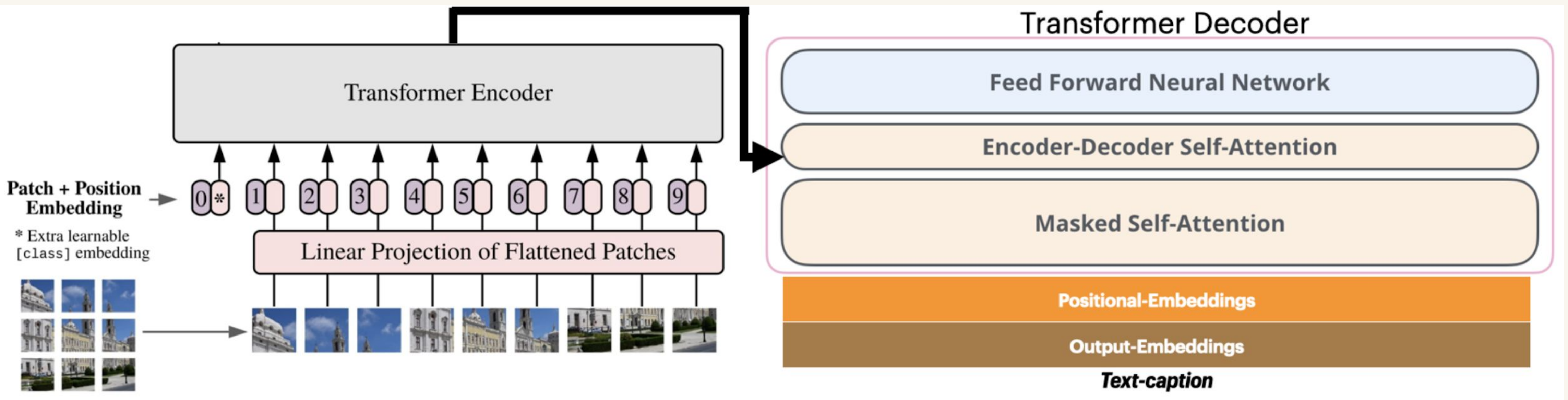
1. Implement a simple image captioning model by hand that utilized seq2seq architecture. Use different datasets and model architectures
2. Compare the efficacy of pre-existing models such as opensource ViT-GPT2, Microsoft Research Team’s GIT, and Saleforce’s BLIP
3. Fine-tune the GIT base models, retraining on niche datasets, to investigate model efficacy prior and after fine-tuning

Basic model architecture

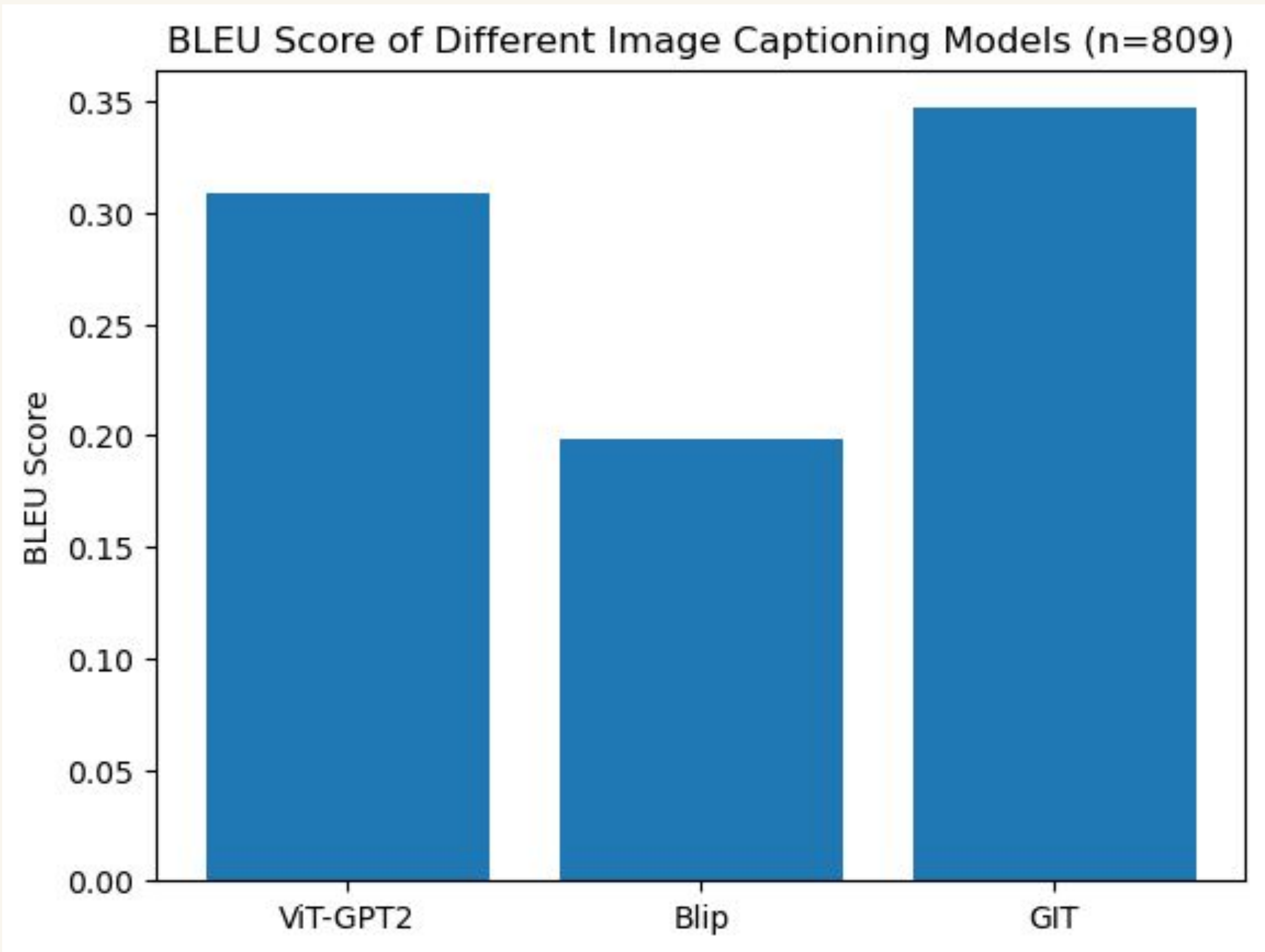

Linear Projection + Position Embedding
+
Decoder Block
FF Neural Nework
Feed Forward Neural Network
Image caption

- The basic model was fairly straightforward: It utilized an image-feature extractor (experimented with VGG16, ResNet, InceptionV3) and Spacy-processed captions and was fed to a Seq2Seq model that predicted the next token (only one that predicted sensible tokens). We utilized GLOVE embeddings
- Experimented with RNN/LSTM/GRU and a TransformerBlock architecture on smaller subset of Flickr8k Dataset.
- Experimented with Greedy Search, which seemed to have limitations, and implemented Beam Search
- Had Flickr30k and COCO120K data created, but could not find resources to test: Possibly would improve performance

ViT-GPT2 Model Architecture


Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0 *
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches
Transformer Decoder
Feed Forward Neural Network
Encoder-Decoder Self-Attention
Masked Self-Attention
Positional-Embeddings
Output-Embeddings
Text-caption

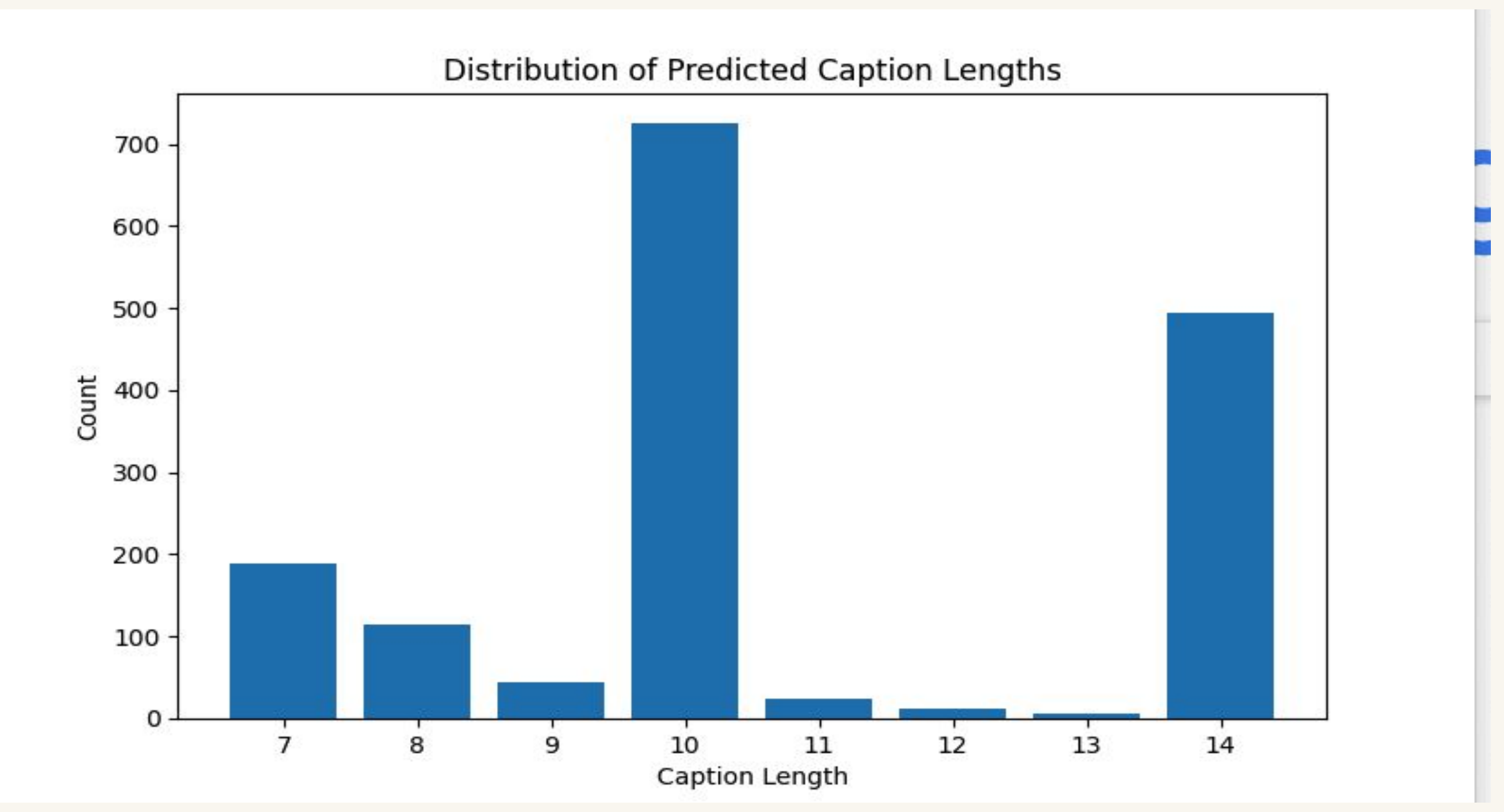
Results

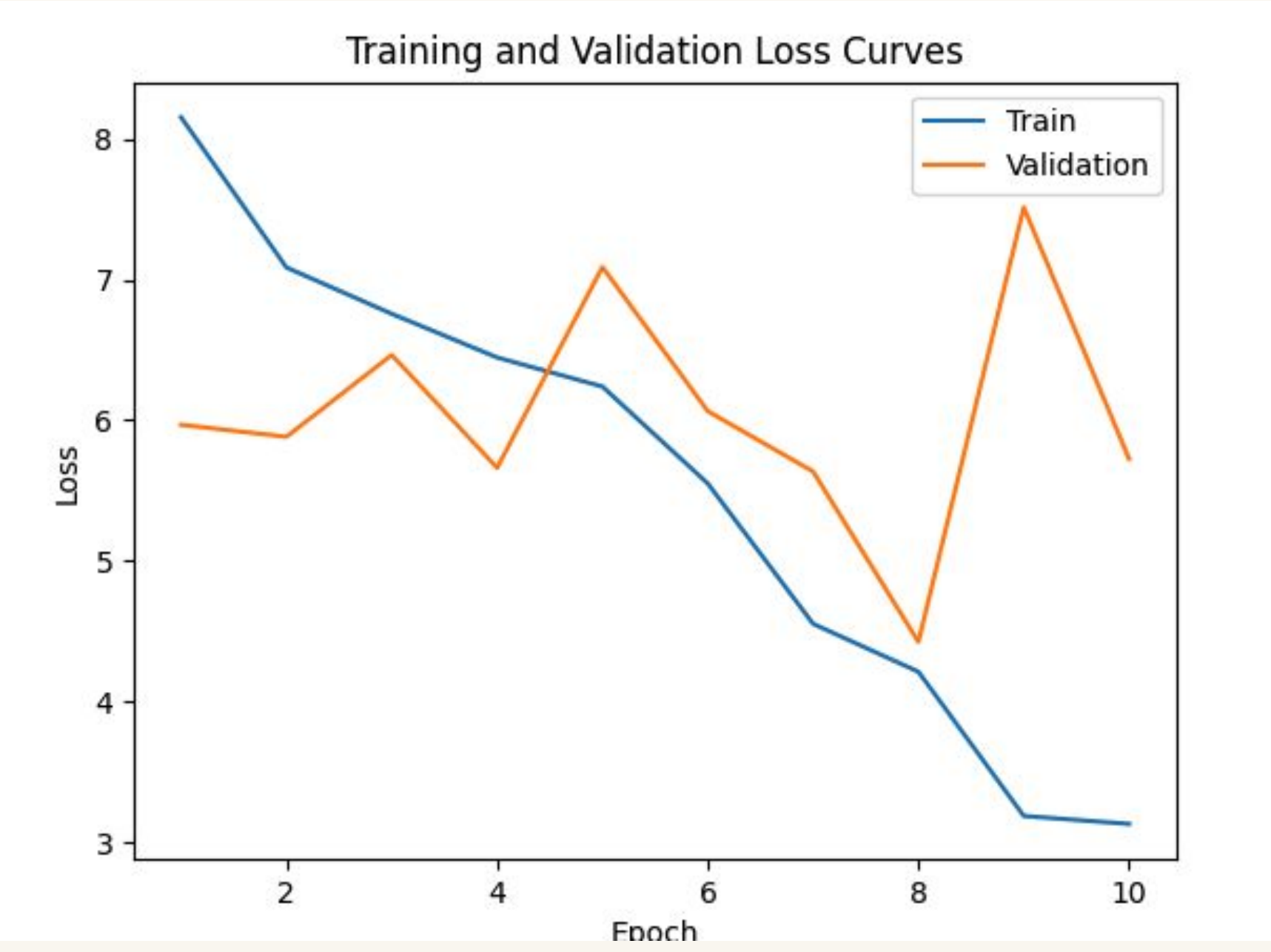


'a stuffed animal with a flower in it'

'a green and red toy with red eyes'

More results and Limitations





- For the basic model, lack of computational resources available became a large issue in our model’s weaknesses
  - Best BLEU score: 0.05. Not very strong, but we believe the use of Greedy Search + low data limits affected our performance. Greedy Search favored very specific tokens and required careful engineering. We implemented Beam Search but it wasn’t computationally efficient.
  - Performance on fine-tuned obviously did not go as plan given lack of generality on general data

References

Vaswani, Ashish, et al. “Attention Is All You Need.” Advances in Neural Information Processing Systems, vol. 30, 2017
Lambda. “Pokemon-Blip-Captions.” Huggingface.co, 13 Dec. 2022
adityajn105. “Flickr 8k Dataset.” Kaggle.com, 2020, www.kaggle.com/datasets/adityajn105/flickr8k

Acknowledgements

As senior capstoners, we would like to thank all those — whether Professors, fellow students, or close friends — that have helped us along the way to the finish line!