


CSCI 1430: Attention Please! Deep Dive into Image Captioning Model Efficacy

by Kevin Kang (kkang11)

Abstract

This project investigates the performance of transformer-based models in image captioning tasks, critically examining the limitations of generalized models like VGG19, ResNet, and InceptionX. We compare the performance of un-tuned models with a fine-tuned model on a specific dataset, and create a baseline model using TensorFlow for further comparison. Our twin objectives are to evaluate the state-of-the-art in image captioning and assess the adaptability of these models to specialized datasets, a crucial aspect for applications like medical imaging. This research aims to improve the granularity and specificity of AI image captioning.


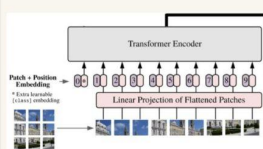
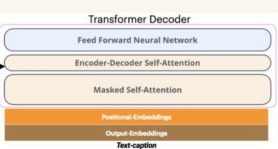
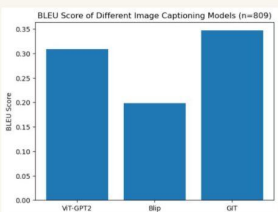

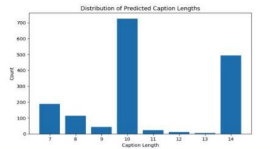
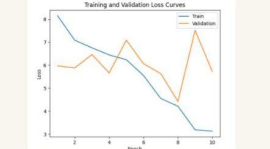
Poster



Attention Please! Deep-dive into Image Captioning Efficacy

Kevin Kang, Ethan Polley, Mithun Ramesh, Siddharth Somasi

CSCI 1430

Motivation	Problem	Goal							
<p>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.</p> <p>Image captioning is a challenging task in computer vision that requires a deep understanding of both visual and linguistic content. Our group aims to investigate and compare the performance of different transformer-based models for image captioning.</p>	<p>Foundational Models: Too Big to Fail?</p> <p>While image extractors like VGG19, ResNet, and InceptionX have achieved impressive results on a wide range of image recognition tasks, <u>there is a growing concern that their generalized nature may lead to blind spots in the AI</u>. 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.</p>	<p>Our Action Items</p> <ol style="list-style-type: none"> 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 Salesforce's BLIP 3. Fine-tune the GIT base models, retraining on niche datasets, to investigate model efficacy prior and after fine-tuning 							
<p style="text-align: center; background-color: #f00; color: white; padding: 2px;">Basic model architecture</p> <div style="display: flex; align-items: center; justify-content: center; gap: 10px;"> <div style="border: 1px solid #ccc; padding: 5px; text-align: center;">  Linear Projection + Position Embedding </div> <div style="font-size: 24px;">+</div> <div style="border: 1px solid #ccc; padding: 5px; text-align: center;"> <div style="border: 1px solid #000; padding: 2px; margin-bottom: 2px;">Decoder Block</div> <div style="border: 1px solid #000; padding: 2px; margin-bottom: 2px;">FF Neural Network</div> </div> <div style="font-size: 24px;">→</div> <div style="border: 1px solid #ccc; padding: 5px; text-align: center;"> Feed Forward Neural Network </div> <div style="font-size: 24px;">→</div> <div style="border: 1px solid #ccc; padding: 5px; text-align: center; background-color: #90EE90;"> Image caption </div> </div> <ul style="list-style-type: none"> 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 	<p style="text-align: center; background-color: #f00; color: white; padding: 2px;">ViT-GPT2 Model Architecture</p> <div style="display: flex; justify-content: space-around; align-items: center;"> <div style="text-align: center;">  <p>Transformer Encoder</p> </div> <div style="text-align: center;">  <p>Transformer Decoder</p> </div> </div>								
<p style="text-align: center; background-color: #f00; color: white; padding: 2px;">Results</p> <div style="display: flex; align-items: center;"> <div style="flex: 1;">  <table border="1" style="margin-top: 5px; font-size: 8px;"> <caption>BLEU Score of Different Image Captioning Models (n=809)</caption> <thead> <tr><th>Model</th><th>BLEU Score</th></tr> </thead> <tbody> <tr><td>ViT-GPT2</td><td>~0.30</td></tr> <tr><td>Blip</td><td>~0.20</td></tr> <tr><td>GIT</td><td>~0.33</td></tr> </tbody> </table> </div> <div style="flex: 1; padding-left: 10px;"> <p>'a stuffed animal with a flower in it'</p> <p style="color: red; font-weight: bold; font-size: 18px;">↓</p> <p>'a green and red toy with red eyes'</p>  </div> </div>	Model	BLEU Score	ViT-GPT2	~0.30	Blip	~0.20	GIT	~0.33	<p style="text-align: center; background-color: #f00; color: white; padding: 2px;">More results and Limitations</p> <div style="display: flex; justify-content: space-between;"> <div style="width: 45%;">  <p style="font-size: 8px;">Distribution of Predicted Caption Lengths</p> </div> <div style="width: 45%;">  <p style="font-size: 8px;">Training and Validation Loss Curves</p> </div> </div> <ul style="list-style-type: none"> For the basic model, lack of computational resources available became a large issue in our model's weaknesses <ul style="list-style-type: none"> 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
Model	BLEU Score								
ViT-GPT2	~0.30								
Blip	~0.20								
GIT	~0.33								
<p style="text-align: center; background-color: #f00; color: white; padding: 2px;">References</p> <p style="font-size: 8px;">Vaswani, Ashish, et al. "Attention is All You Need." Advances in Neural Information Processing Systems, vol. 30, 2017 Lambada. "Pokemon-Blip-Captions." Huggingface.co, 13 Dec. 2022 adityajn105. "Flickr 8k Dataset." Kaggle.com, 2020, www.kaggle.com/datasets/adityajn105/flickr8k</p>	<p style="text-align: center; background-color: #f00; color: white; padding: 2px;">Acknowledgements</p> <p>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!</p>								