Image Caption Generator

Final Project Report
Atharv Shendage (12040360)
Ashemesh Dawande (12040340)

1 Project Overview

1.1 Final Deliverables

1. We used pre-trained transformer based model ViT-GPT2-COCO-EN , which uses Vision transformer which relies on self-attention mechanism and GPT-2 is a transformer-based language model developed by OpenAI trained on Microsoft COCO dataset.
2. We also used ROUGE evaluation metric with BELU.
3. We also trained our model on flicker 30k dataset.
4. We also completed the pipeline.

1.2 ML models :

We trained 3 different models and also used one pretrained model vit-gpt2. Our first model was trained on Flickr8k dataset with vgg16 and LSTM model , then we increased the no of epochs , we also then used flickr30k data set.

1.3 Model Comparisons

Model	Complexity	Ease of Interpretation
Flickr8k + VGG16 + LSTM (5 Epochs)	Moderate complexity	Relatively easy to interpret
Flickr8k + VGG16 + LSTM (20 Epochs)	Increased complexity with longer training	Potential decrease in ease of interpretation
Flickr30k + VGG16 + LSTM (1 epoch)	Higher complexity with larger dataset	Interpretability depends on dataset quality
ViT-GPT2-COCO-EN	High complexity due to pretrained architecture	Easier to use, may be less interpretable

Model	BELU-1 Score	Rouge1 score
Flickr8k + VGG16 + LSTM (5 Epochs)	0.51	0.41
Flickr8k + VGG16 + LSTM (20 Epochs)	0.54	0.45
Flickr30k + VGG16 + LSTM (1 epoch)	0.49	0.40
ViT-GPT2-COCO-EN	0.68	0.53

Model	Training Time
Flickr8k + VGG16 + LSTM (5 Epochs)	55 mins
Flickr8k + VGG16 + LSTM (20 Epochs)	220 mins
Flickr30k + VGG16 + LSTM (1 Epoch)	40 mins
ViT-GPT2-COCO-EN	–

1.4 Comparing Results With Research Paper

The results in the research paper : `https://arxiv.org/pdf/1502.03044v3.pdf` were as follows :

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
	Log Bilinear (Kiros et al., 2014a) °	65.6	42.4	27.7	17.7	17.31
Flickr8k	Soft-Attention	67	44.8	29.9	19.5	18.93
	Hard-Attention	67	45.7	31.4	21.3	20.30
	Google NIC $^{\dagger\circ\Sigma}$	66.3	42.3	27.7	18.3	-
Flickr30k	Log Bilinear	60.0	38	25.4	17.1	16.88
	Soft-Attention	66.7	43.4	28.8	19.1	18.49
	Hard-Attention	66.9	43.9	29.6	19.9	18.46

Comparing these results with our models we can see where improvements are needed , In language part , We can use attention based models to improve accuracy , In image part we can focus on improvement in object detection and not just use feature map of images. We can also add hyper-parameter which favors objects with high confidence over those with low confidence.