DEEP LEARNING FOR IMAGE CAPTIONING

Drithi Davuluri(B21AI055)
Nishtha Karki(B21AI051)
Nandi Venkata Durga Hima Varshitha (B21CS049)

PROBLEM STATEMENT

Image captioning is the task of generating textual descriptions of images. It requires both computer vision techniques to understand the content of the image, and natural language processing techniques to generate the captions. Image captioning has many applications such as assisting visually impaired users, automatic image tagging, search engine optimization, and human-robot interactions.

SOLUTION STRATEGY

Our idea is to implement image captioning with an encoder-decoder model (which is a generalized term). By studying all the literature , we plan to utilize two main components: an encoder and a decoder. The encoder will process the input image, extracting its features and encoding them into a vector representation. We will then feed these vectors into the decoder, which will sequentially generate words to form a caption. While training, our model will learn to align the generated words with the visual features extracted by the encoder. This approach will help our model to automatically produce descriptive captions for images.

Model-1:

The encoder uses a pre-trained ResNet-50 model as a feature extractor. In this implementation, the ResNet-50 model is truncated by removing the last two layers, which were responsible for classification. The remaining layers are used to extract rich visual features from the input images.

The decoder is responsible for generating the textual caption for the input image features. It consists of the following components:

- Word Embedding Layer
- LSTM Cell
- Fully Connected Layer
- Dropout Layer

The decoder has a generate_caption method for inference. It generates captions word-by-word in an auto-regressive manner.
During training, the combined model takes an image and ground truth caption. The encoder extracts visual features from the image, which are fed into the decoder along with the caption. The decoder computes word probabilities, and loss is calculated against the ground truth. During inference, the model generates captions by first encoding the image with encoder , then passing the features to decoder’s generate_caption method.

Model-2 :
In this model, we are using a pre-trained vision encoder-decoder model called "vit-gpt2-image-captioning" from the "nlpconnect" repository for the task of image captioning. The model is based on the combination of a Vision Transformer (ViT) as the encoder and a GPT-2 language model as the decoder. The encoder (ViT) is responsible for extracting visual features from the input image, while the decoder (GPT-2) generates the textual caption based on those visual features.
We implemented this model so as to compare the performance of our implemented model with a pretrained model.

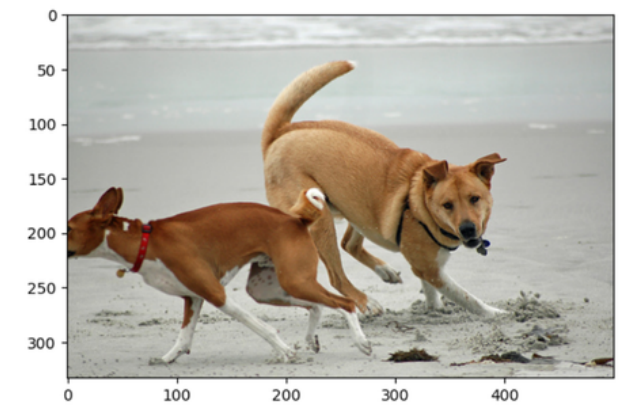
DATASET

Flickr8K is a dataset with around 8,000 good-quality photos from Flickr, each with five captions written by people. These captions show different ways to describe what's in the photos. This dataset is useful for training our model to write captions for images. The following is the representation of the dataset along with the example of one image with its corresponding 5 captions-

There are 40455 image to captions

	image	caption
0	1000268201_693b08cb0e.jpg	A child in a pink dress is climbing up a set o...
1	1000268201_693b08cb0e.jpg	A girl going into a wooden building .
2	1000268201_693b08cb0e.jpg	A little girl climbing into a wooden playhouse .
3	1000268201_693b08cb0e.jpg	A little girl climbing the stairs to her playh...
4	1000268201_693b08cb0e.jpg	A little girl in a pink dress going into a woo...
...	...	...
40450	997722733_0cb5439472.jpg	A man in a pink shirt climbs a rock face
40451	997722733_0cb5439472.jpg	A man is rock climbing high in the air .
40452	997722733_0cb5439472.jpg	A person in a red shirt climbing up a rock fac...
40453	997722733_0cb5439472.jpg	A rock climber in a red shirt .
40454	997722733_0cb5439472.jpg	A rock climber practices on a rock climbing wa...

40455 rows × 2 columns



Caption: Two different breeds of brown and white dogs play on the beach .
Caption: Two dogs are making a turn on a soft sand beach .
Caption: Two dogs playing in the sand at the beach .
Caption: Two dogs playing together on a beach .
Caption: Two large tan dogs play along a sandy beach .

INNOVATIONS

Attention
In addition to the base encoder-decoder architecture, we incorporated an attention mechanism to further enhance the image captioning capabilities. The inclusion of the attention mechanism in the image captioning model is an innovative approach that allows the model to dynamically focus on the most relevant visual features when generating each word of the caption, leading to improved results compared to standard encoder-decoder architectures. The attention module is implemented as a separate component, which is integrated into the decoder.
The attention module consists of the following components:

- Attention Projection Layers
- Attention Scoring Layer
- Attention Weighting

The attended visual context is then concatenated with the word embedding and fed into the LSTM cell of the decoder. This allows the decoder to selectively focus on the most relevant visual features when predicting the next word in the caption.
By incorporating the attention mechanism, the proposed model can better capture the local and global relationships.

Interpretations:
To provide better interpretability and understand the model's decision-making process, we have implemented a visualization technique that highlights the regions of the input image that the model attends to while generating each word in the caption. This visualization is achieved through the `plot_attention` function, which takes the input image, the generated caption, and the attention weights as inputs.
This function creates a figure with multiple subplots, where each subplot corresponds to a word in the generated caption. In each subplot, the function overlays the attention weights for that word onto the input image, using a grayscale color map and adjustable transparency. This visual representation allows us to observe which image regions the model focuses on when generating each word, providing valuable insights into the model's decision-making process.
The attention visualization technique provides a visual representation of the model's attention weights, allowing for direct interpretability of the image regions the model focuses on while generating each word in the caption, which was previously difficult to achieve in black-box neural network models.

CLIP based evaluation
In this approach, we incorporated an innovative technique for evaluating the similarity between the generated captions and the original captions. Instead of using model-specific metrics or techniques, we used the CLIP (Contrastive Language-Image Pre-training) model to generate text embeddings for both the original and generated captions.
Advantages of using CLIP Embeddings:

- The CLIP-based similarity scores are not specific to any particular captioning model, making the evaluation more robust and applicable to a wider range of models.
- The CLIP embeddings provide a meaningful and interpretable representation of the textual content, allowing for a more intuitive understanding of the similarities and differences between the captions.

RESULTS AND ANALYSIS

These are the images and the generated captions of the Encoder Decoder Model with Attention that have the highest similarity score calculated using clip embeddings. As we can observe , the model produces highly accurate results and the high similarity score justifies it.

Actual Caption: A man is climbing a rock wall .
Generated Caption (Model): a man climbing a rock wall . <EOS>
Generated Caption (Pretrained): a man sitting on top of a rock wall
Similarity Score (Model): 0.9580078125

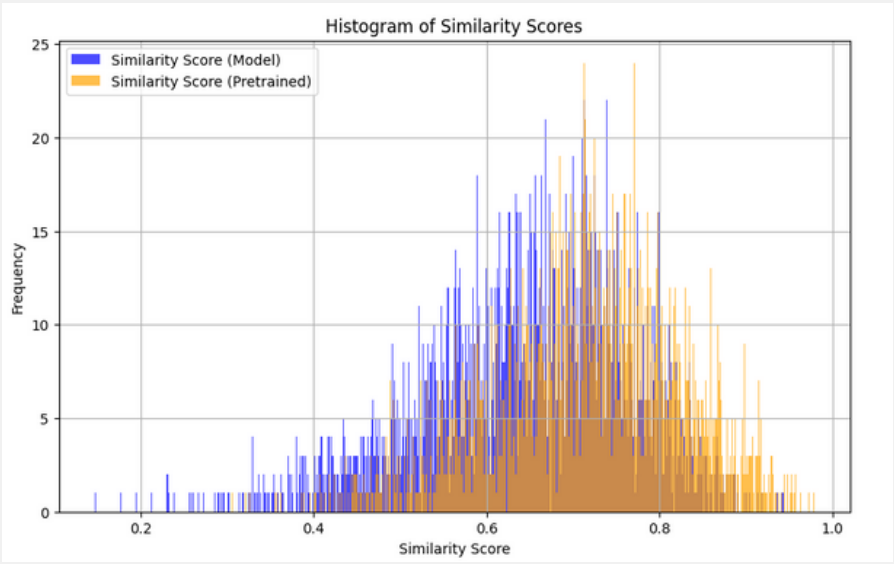
Actual Caption: A large bird is flying over water .
Generated Caption (Model): a large bird flying through the water . <EOS>
Generated Caption (Pretrained): a bird standing on top of a body of water
Similarity Score (Model): 0.93896484375

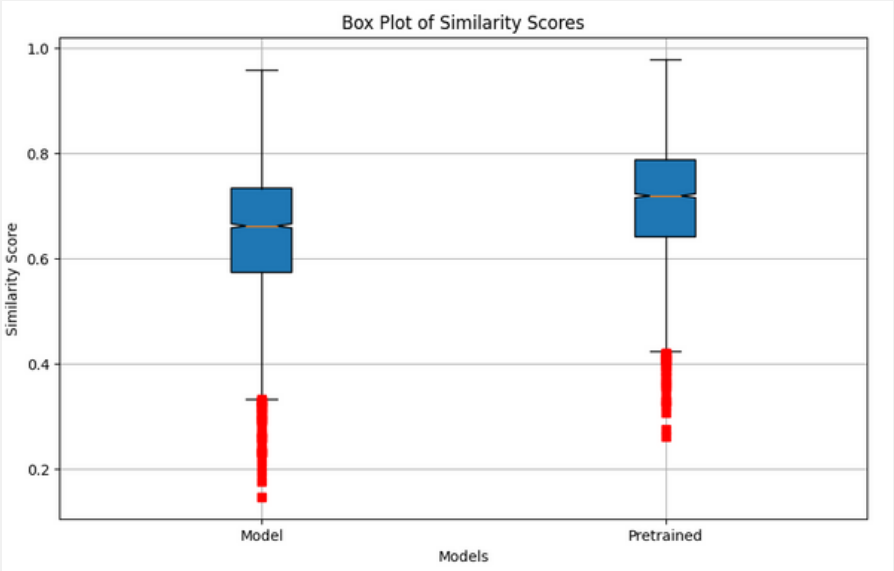
These are the images and the generated captions of the pretrained model that have the highest similarity score. As we can see, the scores have not drastically increased from the ones obtained in our model.

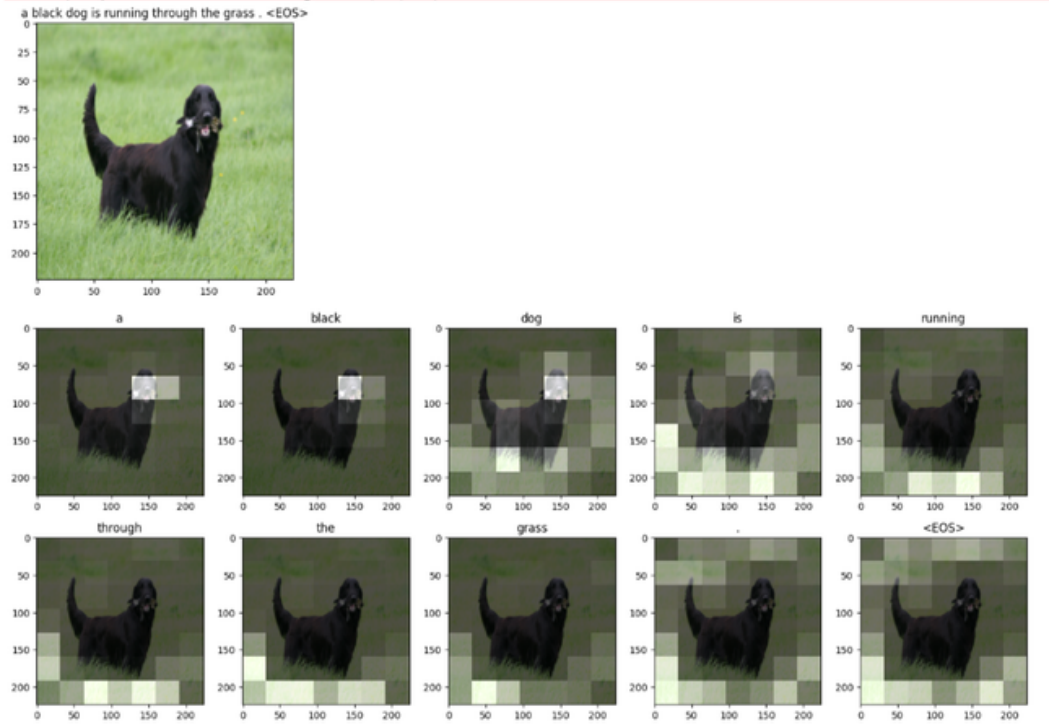
```
Actual Caption: A man is standing outside of a yellow building .
Generated Caption (Model): a man in a yellow shirt is standing on a yellow rail . <EOS>
Generated Caption (Pretrained): a man standing in front of a yellow building
Similarity Score (Model): 0.97265625
```

```
Actual Caption: A man and a woman on a bench .
Generated Caption (Model): a man in a black coat and a hat sits on a bench . <EOS>
Generated Caption (Pretrained): a man and a woman sitting on a bench
Similarity Score (Model): 0.96826171875
```

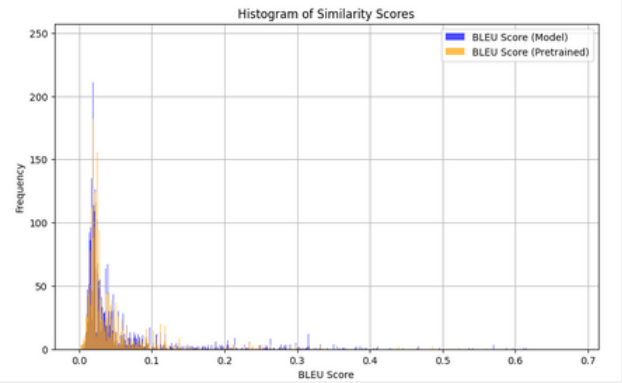
We can now analyze the distribution of the similarity scores to evaluate our model's performance relative to the pretrained one.
The results indicate that our model performs comparably to the pretrained model.







The above are the results of our implementation of interpretation of attention weights. We can observe that it highlights the relevant regions for words generation.



Here, we have implemented a standard accuracy metric for text i.e. BLEU Score. We can see that our metric of similarity score using CLIP has better results compared to the traditional BLEU score metric.The clip encodings used to find similarity score are performing better than BLEU as they inculcate context.

POSSIBLE WEAKNESSES

We have implemented Encoder Decoder model with Attention which is capable of producing results comparable to a pretrained model using vit-gpt2-image-captioning.
For further improvements:

- We could fine-tune the pretrained model used.
- We could calculate the loss as the avg of our loss and the loss calculated using MSE of clip embeddings.
- We could use the embeddings produced by clip to backpropagate in our model.

CONCLUSION

Our implementation of the Encoder-Decoder model with Attention for image captioning has achieved promising results comparable to a pre-trained model. The incorporation of an attention mechanism and innovative techniques like attention visualization and CLIP-based evaluation have significantly enhanced the model's performance and interpretability. While there is still room for improvement through fine-tuning and leveraging CLIP embeddings more effectively, our solution demonstrates the potential of attention-based architectures in generating accurate and descriptive captions. The insights gained from this project contribute to the advancement of image captioning technologies, paving the way for more intelligent and interpretable AI systems in various applications.

CONTRIBUTIONS:
We worked together on researching the different available methods and then through trial and error, we discovered various approaches possible out of which we implemented the above.

REFERENCES

- Understanding Deep Learning: DNN, RNN, LSTM, CNN and R-CNN
- Encoder-Decoder Model for image Captioning, Machine Translation
- Attention is all you need
- Significance of Interpretability in Machine Learning: Unveiling the 'Why' Behind Predictions
- CLIP from OpenAI: what is it and how you can try it out yourself

Link to Colab File