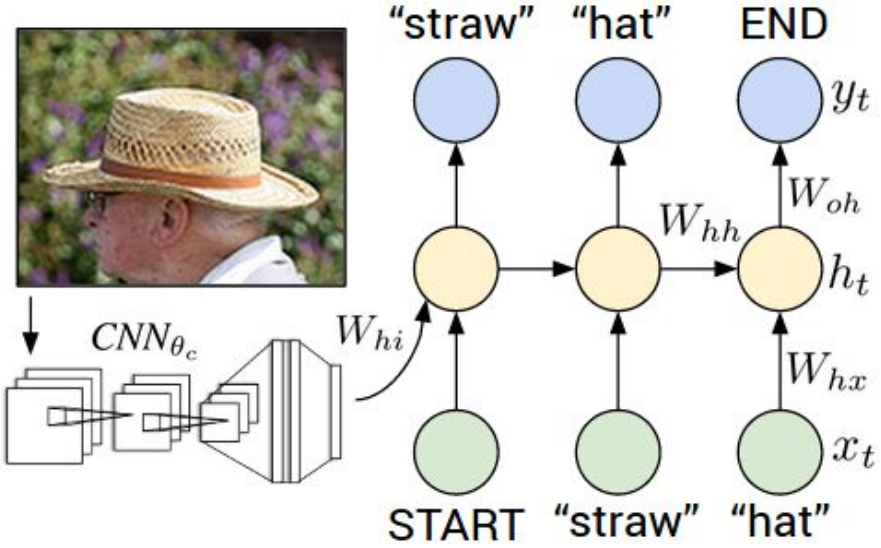
Advanced Image Captioning



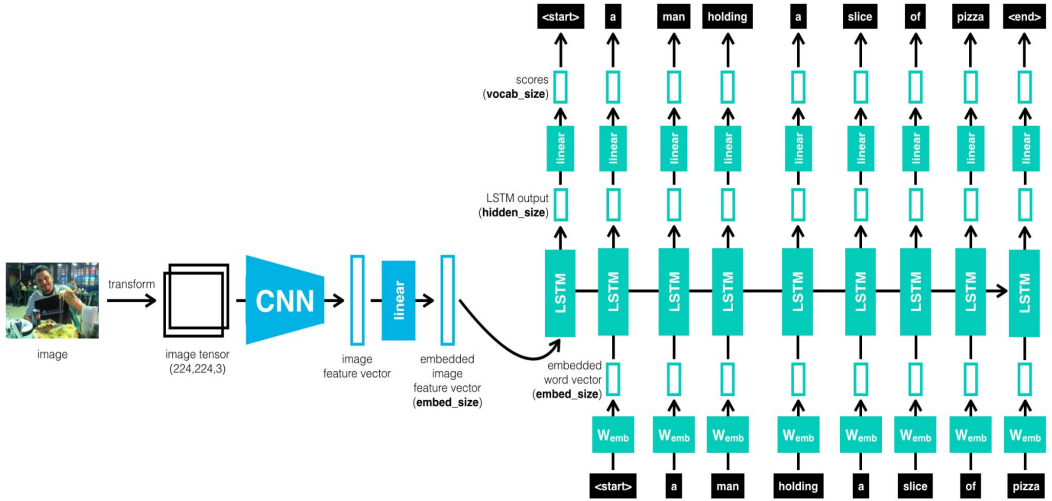
By -
Parth Maheshwari
Vishal Kumar Panda

Introduction

- **Images can have multiple interpretations and captions, depending on the context or focus. Identifying the most appropriate caption is challenging.**
- **A classic method involves encoding image features using a pretrained model(VGG,ResNet,Xception,Inception) and decoding the hidden state produced by the model to generate captions by using a decoder such RNN,LSTM,GRU etc.**
- **Classic models face issues with capturing complex information and lack the essence in the generates captions.**

Basic Model Architecture

- A pre-trained CNN model Xception is used to extract high-level features from the input images.
- A word embedding layer is used to capture the semantic meaning of the words.
- An LSTM (Long Short-Term Memory) layer is used to process the embedded text tokens and learn the sequential relationship between the words in the image captions.
- A layer with a softmax activation function is used to generate the final output, predicting the most likely next word in the caption based on the image features and the LSTM-generated context.



Dataset & Evaluation

Dataset:

- Flickr8k_Dataset: Contains a total of 8000 images in JPEG format with different shapes and sizes. Of which 6000 are used for training, 1000 for test and 1000 for development.
- Flickr8k_text: Contains text files describing train_set, test_set. Flickr8k.token.txt contains 5 captions for each image i.e. a total of 40460 captions.

Preprocessing:

- The images are resized to a fixed size of (299, 299) to ensure consistent input dimensions for the Xception model which is used to extract image features with a results vector of size (2048,).
- All words were converted to lowercase, and special characters and numbers were removed,start and end tokens were added and captions were tokenized and converted to numerical tokens .
- The Glove embeddings weights are utilized for the Embedding layer in the model to provide a better semantic understanding of the words.

Evaluation:

- BLEU(Bilingual Evaluation Understudy Score) for evaluating a generated captions to a reference test captions
- ROGUE(Recall-Oriented Understudy for Gisting Evaluation) which compares automatically produced captions against a reference set of captions.

Results

	Vanilla CNN-LSTM	With GloVe Embeddings	With Bahdnau Attention	With Bahdnau Attention and GloVe Embeddings
BLEU-2	0.16	0.13	0.19	0.16
BLEU-3	0.05	0.037	0.11	0.06
BLEU-4	0.014	0.01	0.07	0.02
Rouge-Precision	47.10	44.70	47.22	47.07
Rouge - Recall	47.50	43.42	47.02	47.21
Rouge - F1 Score	44.62	41.43	44.46	44.52

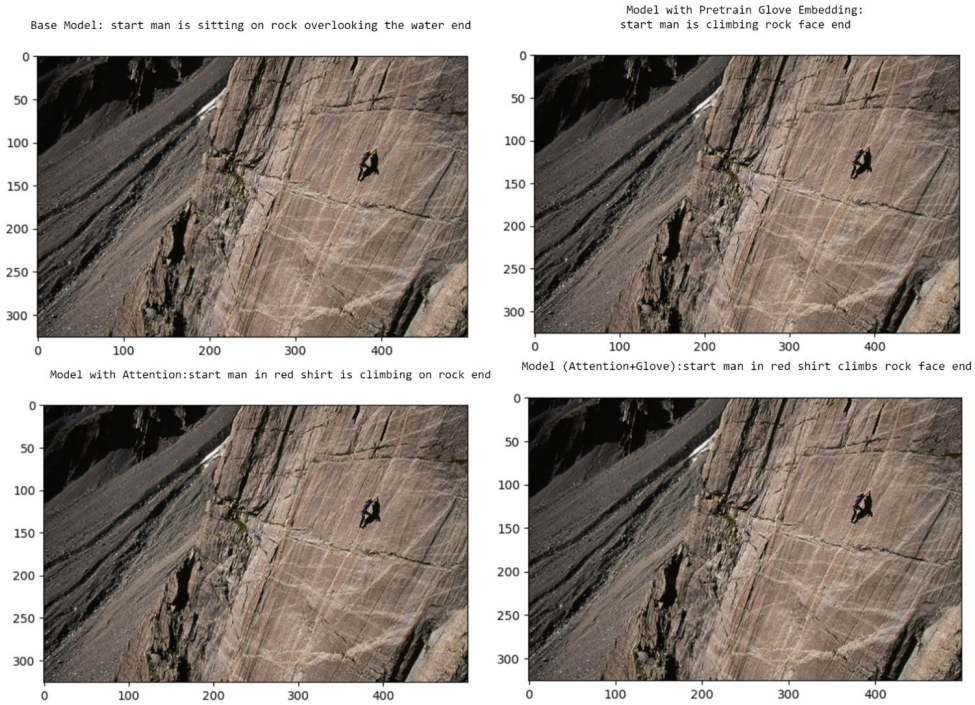
Critiques

- The base model captures the presence of a man and some context like sitting on a rock overlooking water, but it doesn't accurately describe the image.

- Model with Pretrained GloVe Embedding has a semantic understanding of words but generates a caption that doesn't show significant improvement compared to the base model

- Model with Attention focuses on different parts of the image, capturing more accurate details like the man in a red shirt climbing on a rock. The attention mechanism has improved the model's ability to generate better captions.

- Model (Attention+GloVe): This model combines attention and GloVe embeddings, generating a caption where the man is sitting on a rock. It captures the same general meaning but phrases it differently.



Conclusion

- The image captioning model with both Attention and Glove embeddings has significant improvement compared to the base model.
- The generated captions are moderately accurate as indicated by the metrics with room for more improvement
- Fine-tuning the model with a larger and more diverse dataset could possible increase the model performance.
- Experimenting with different architectures, such as using bidirectional LSTMs, GRUs, or Transformers would provide better context understanding and improve caption generation.

THANK YOU