UNIVERSITY OF CALGARY

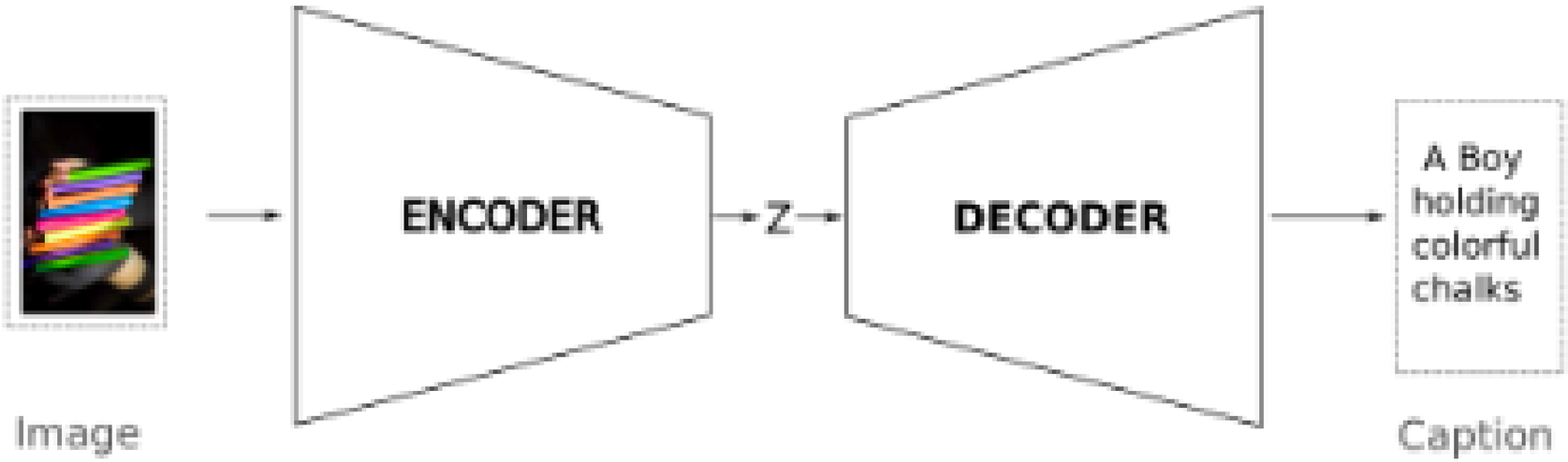
Model Architectures for Image Captioning: CNN-RNN vs CNN-GPT

GROUP 25

- Jacob Idoko
- Gabriel Gabari
- Aakash Sorathiya
- Jagrit Acharya
- Chioma Ukaegbu
- Emmanuel Alafonye

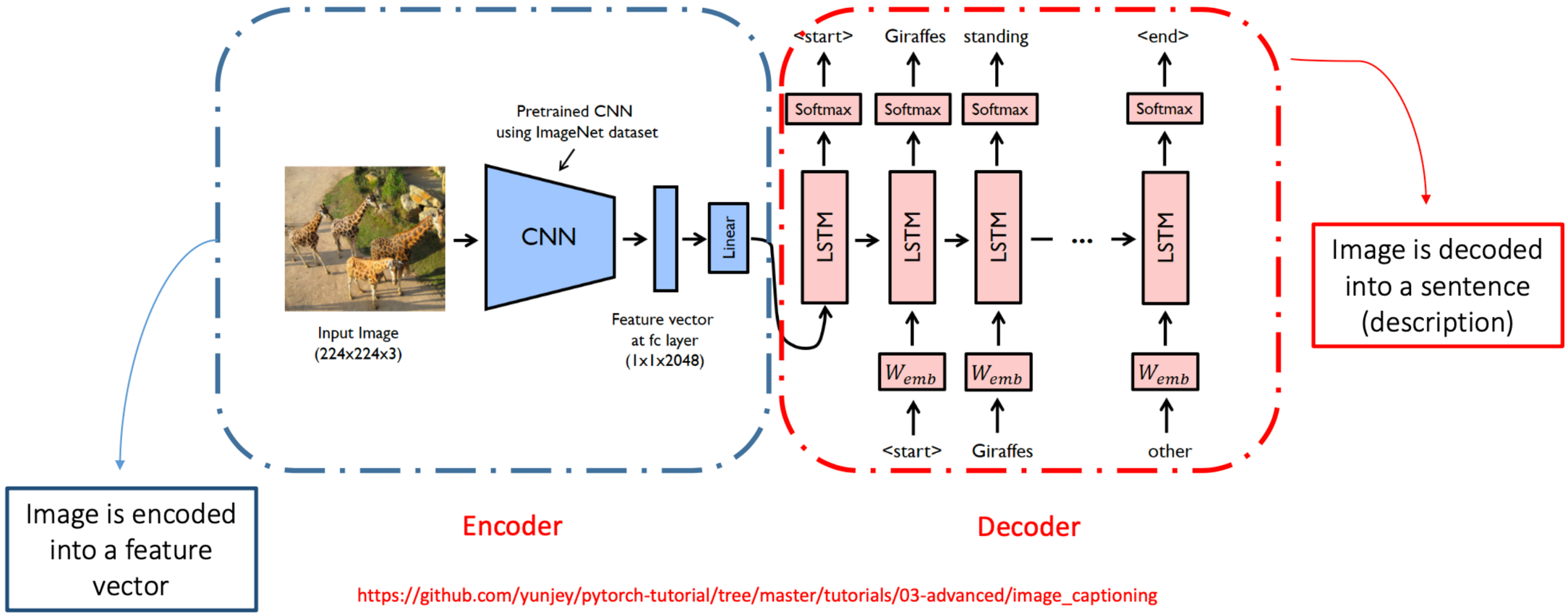
INTRODUCTION - WHAT IS IMAGE CAPTIONING?

- Generation of natural language descriptions for visual content
- Bridges the gap between computer vision and natural language processing (NLP)
- Has a wide range of applications such as assisting visually impaired individuals by providing audio descriptions of images



An illustration of an encoder-decoder model for image captioning

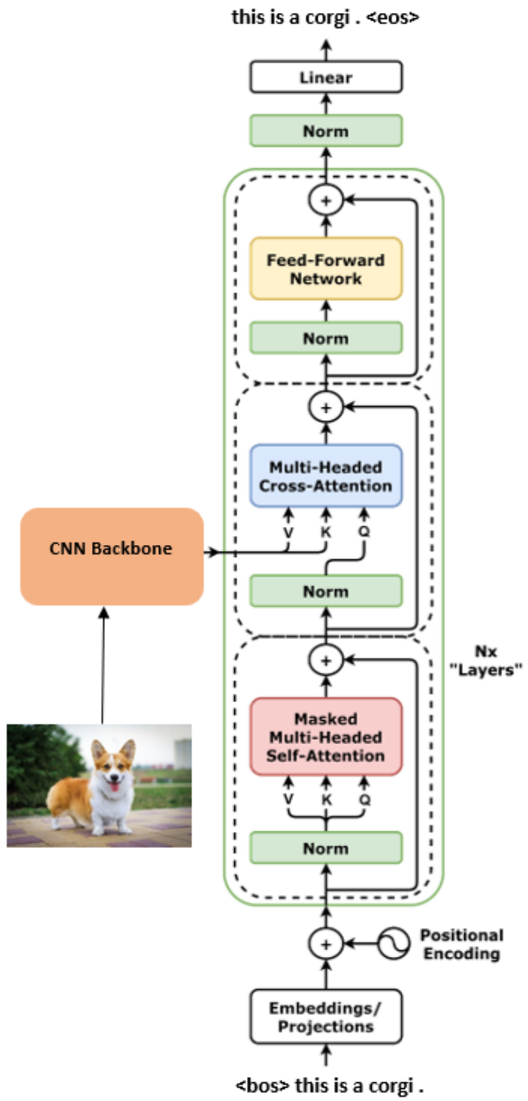
CNN-RNN



https://github.com/yunjey/pytorch-tutorial/tree/master/tutorials/03-advanced/image_captioning

CNN-RNN model

CNN-GPT



Decoder Block

3 Main Sub-blocks

- Masked Multi-Headed Self-Attention
- Multi-Head Attention
- Feed-Forward Network

Takes 2 Inputs

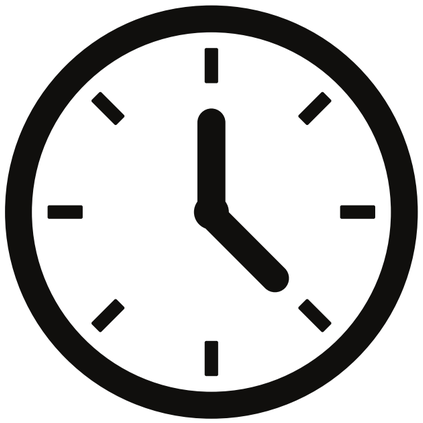
- Previously Generated Words
- Image Feature from CNN

Parameters

- 6 Decoder Layers
- 16 Attention Heads

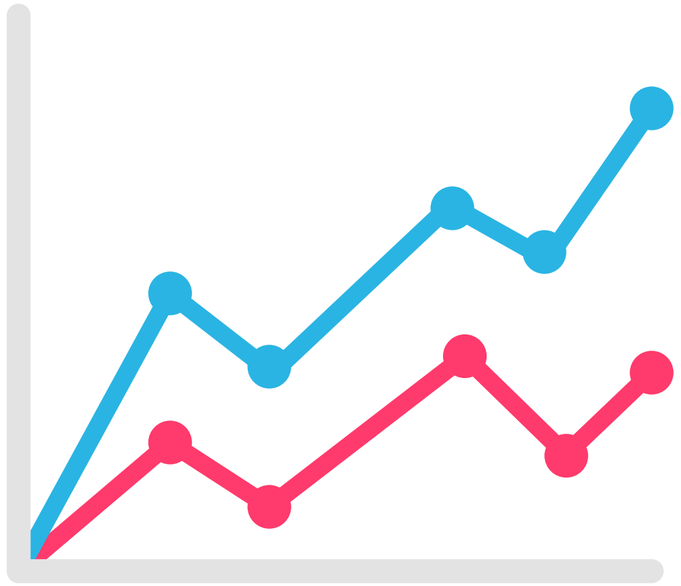
Training Detail

Training Time on NVD V-100



- CNN-RNN (~5 Hrs)
- CNN-GPT (~11 Hrs)

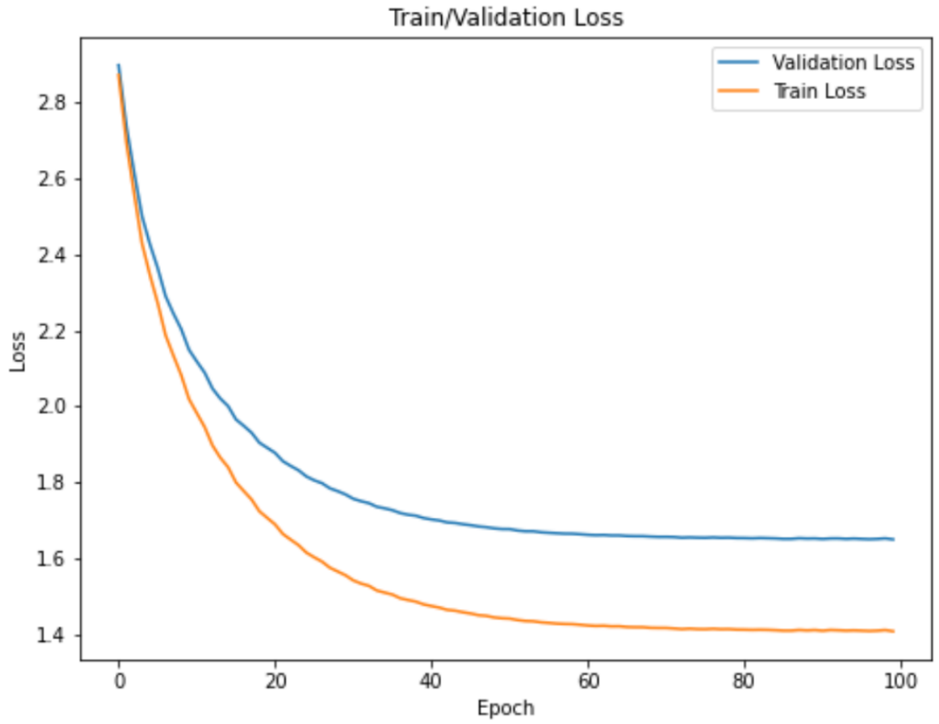
Evaluation Metrics



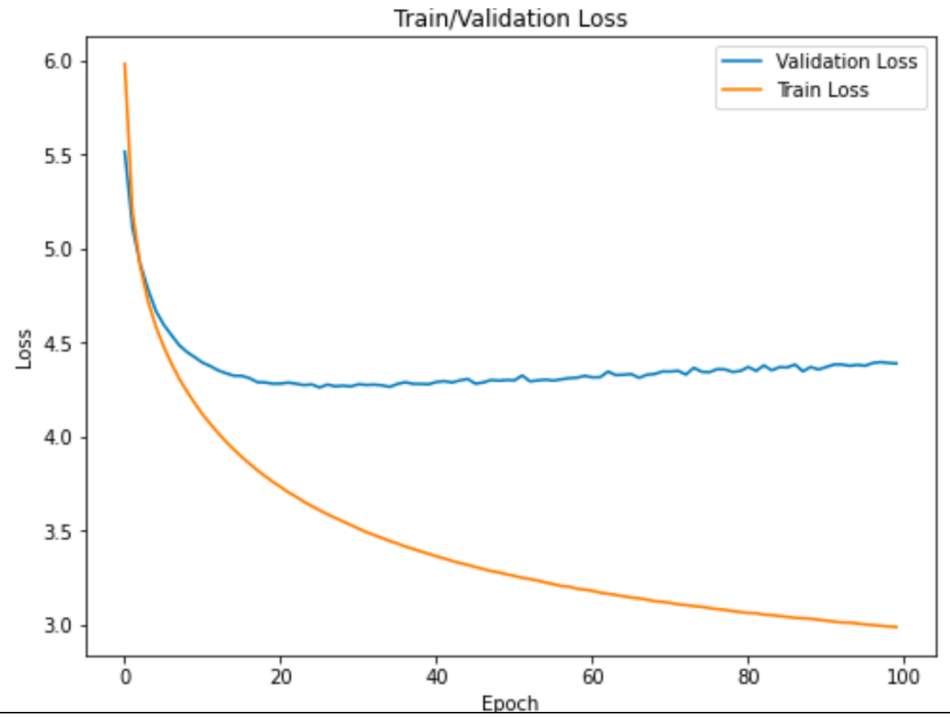
- BLEU Scores

Training Detail

Average epoch loss for train/validation set (CNN-RNN)

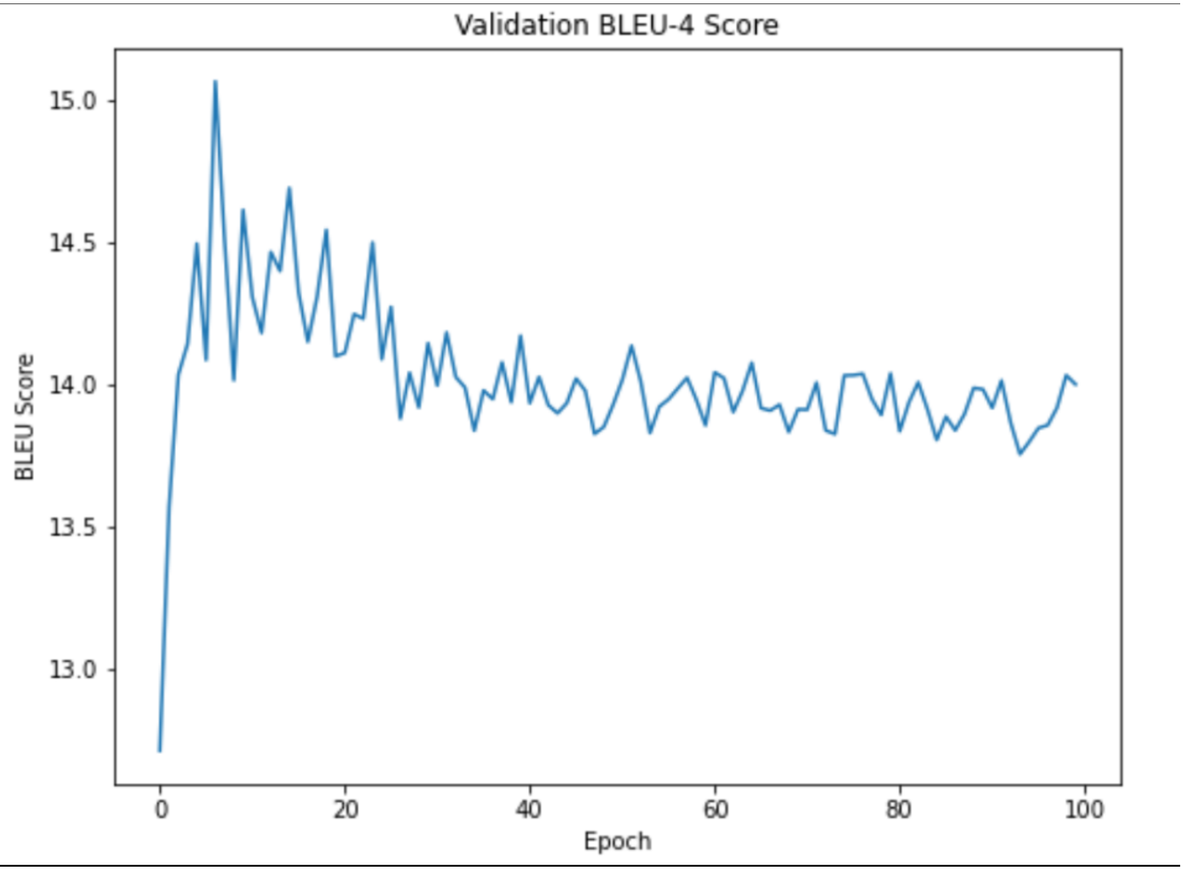


Average epoch loss for train/validation set (CNN-GPT)

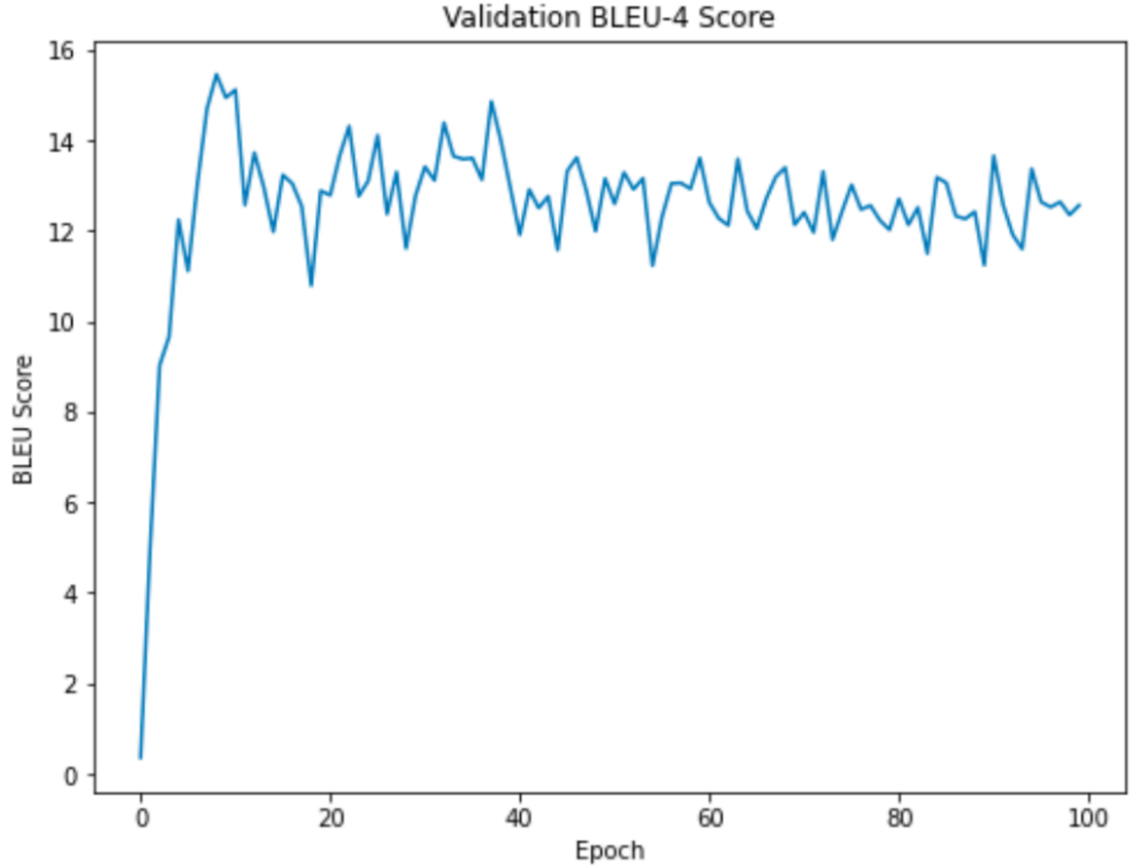


Training Detail

Average epoch BLEU score for train/validation set (CNN-RNN)



Average epoch BLEU score for train/validation set (CNN-GPT)



Evaluation on Test

Table 2. BLEU Scores on Flickr8k Test Set

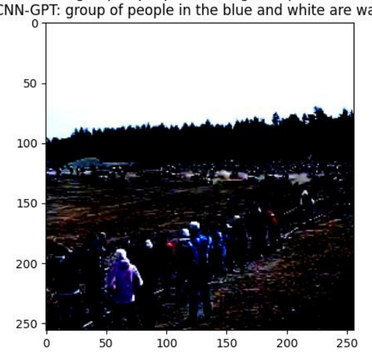
Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
CNN-RNN	64.25	40.98	24.81	14.63
CNN-GPT	52.65	33.6	20.95	12.27

Details

- Evaluate the performance of CNN-RNN & CNN-GPT
- Test unseen data- Computed BLEU Score
- BLEU - Similarity with generated and human hand-written
- A Higher Score indicates a better caption
- Sequence of n = 1 - 4 word sequence

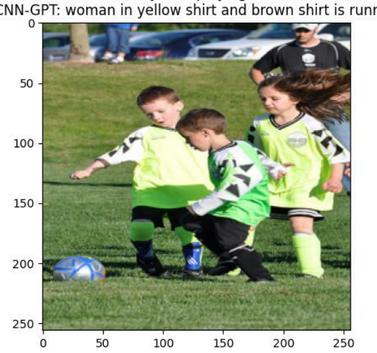
CNN-RNN: a brown dog is walking through the snow
CNN-GPT: dog runs through the snow

CNN-RNN: a group of people standing on a path in the desert
CNN-GPT: group of people in the blue and white are walking

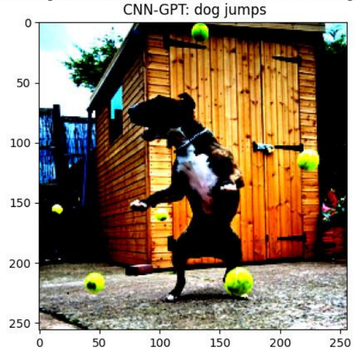


CNN-RNN: a man is surfing in the water
CNN-GPT: person is doing tricks in the back of the air

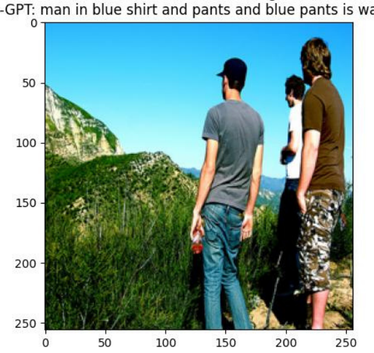
CNN-RNN: two boys are playing soccer on a field
CNN-GPT: woman in yellow shirt and brown shirt is running



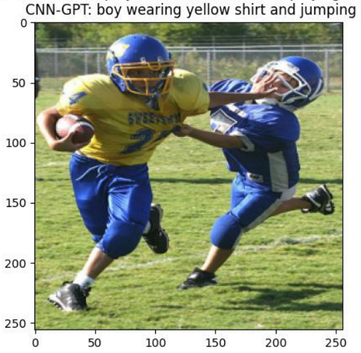
CNN-RNN: a dog with a tennis ball with its mouth is chasing a tennis ball
CNN-GPT: dog jumps



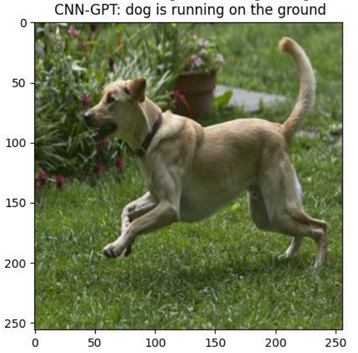
CNN-RNN: a man and a woman are standing in front of a rock wall
CNN-GPT: man in blue shirt and pants and blue pants is walking on



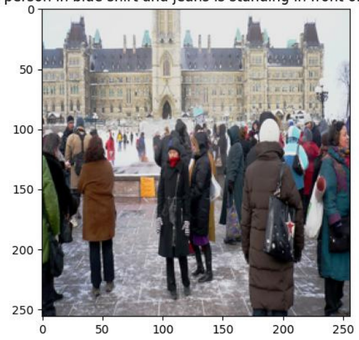
CNN-RNN: a football player in a blue uniform is playing with a ball
CNN-GPT: boy wearing yellow shirt and jumping



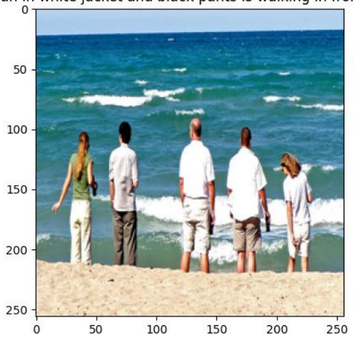
CNN-RNN: a tan dog runs through the grass
CNN-GPT: dog is running on the ground



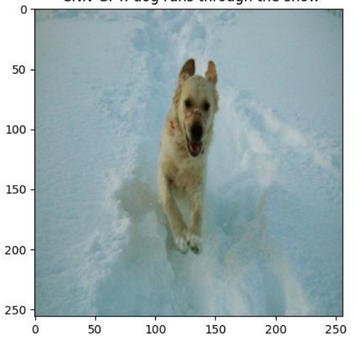
CNN-RNN: a group of people are standing in front of a group of people
CNN-GPT: person in blue shirt and jeans is standing in front of some buildin



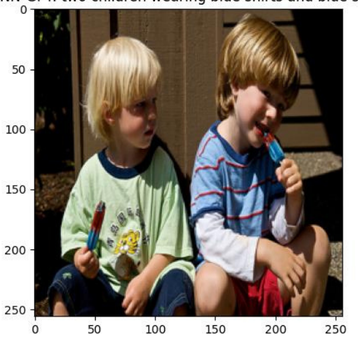
CNN-RNN: a group of people are standing on a beach
CNN-GPT: man in white jacket and black pants is walking in front of some wa



CNN-RNN: a white dog is running through the snow
CNN-GPT: dog runs through the snow



CNN-RNN: two boys play with a toy
CNN-GPT: two children wearing blue shirts and blue shirts



Visualizations of some model generated captions

Thank You!