Image Captioning

Problem Statement

The goal is to create a system that combines the power of RNN, CNN and FFNN. You will have a two stage DNN, wherein the first stage is a CNN processing an image and an RNN/Transformer processing the caption of the image. The FFNN will take outputs of CNN and RNN and will give the verdict as a value between 0 and 1 (both included), expressing the degree of consistency between the image and the caption (1- consistent, 0-inconsistent). For example, if the image is that of a tiger chasing a deer, the caption of "a peaceful scene of nature" is inconsistent with the picture. On the other hand, the picture of a long line of people can have many consistent captions- (a) Crowd eagerly waiting for a ticket to the cricket stadium, or (b) Hungry people in food-line during covid, or (c) Students waiting in queue for an admission form, but not (d) Snow-flakes falling from the sky.

The dataset that will be used for this assignment is MS-coco (https://arxiv.org/pdf/1405.0312.pdf,https://arxiv.org/pdf/1504.00325.pdf, https://cocodataset.org/#home).

Dataset Discussion

Details of Examples: positive and negative

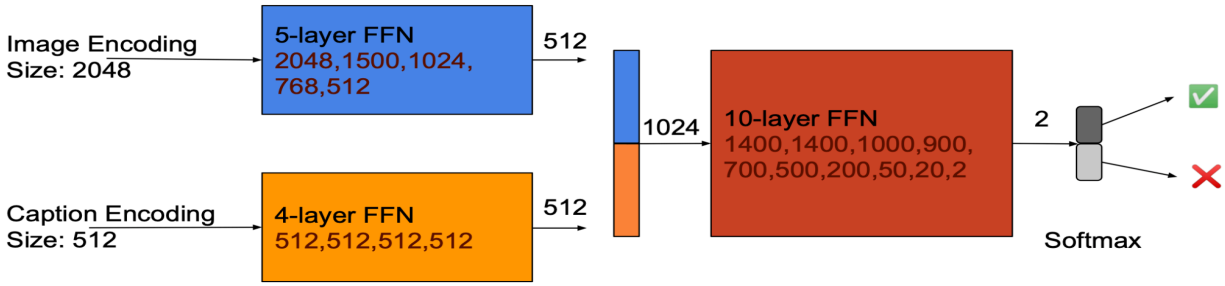
- How many positive examples (1 category): 25014
- How many negative examples (0 category): 25014

Dataset : COCO 2017 (Common Objects in Context)

Creation of Negative Samples:

Randomly chosen 10 captions from dataset excluding the current caption and checked cosine similarity between each caption embedding (sentence embeddings - BERT) and assigned the one with the lowest similarity as negative sample.

System implementation



Details of the FFNN N/W

- Layers: Two FFN of 5 and 4 layers followed by 10 layer FFN
- Different Hyper parameters: Learning rate: 10-4, Adam : 10-3
- Model diagram

Image Encoding Size: 2048
Caption Encoding Size: 512
Details of the RNN/Transformer N/W
CLIP ViT's Text Transformer
Hyper Parameters : [CLIP ViT-B/32]

- Learning rate : 5 × 10^-4
- Embedding dimension: 512
- Text Transformer:
  - Layers: 12
  - Width: 512
  - Heads: 8

Training details (hyper-parameters)

- How many epochs? 200 epochs
- What is the learning rate? 1e-4
- Convergence criterion: Early stopping based on validation loss.
- Various performance parameters: P, R, F-score, Accuracy

Performance Parameters

True Negatives : 43 True Positives : 39 False Positives : 7 False Negatives : 11