Assignment-3 Discussion (Image Captioning)

Singamsetty Sandeep, 213050064
Divyank Pratap Tiwari, 213050029
Sakharam Sahadeo Gawade, 213050027

06 May 2022

Problem Statement (1/2)

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).

Problem Statement (2/2)

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)

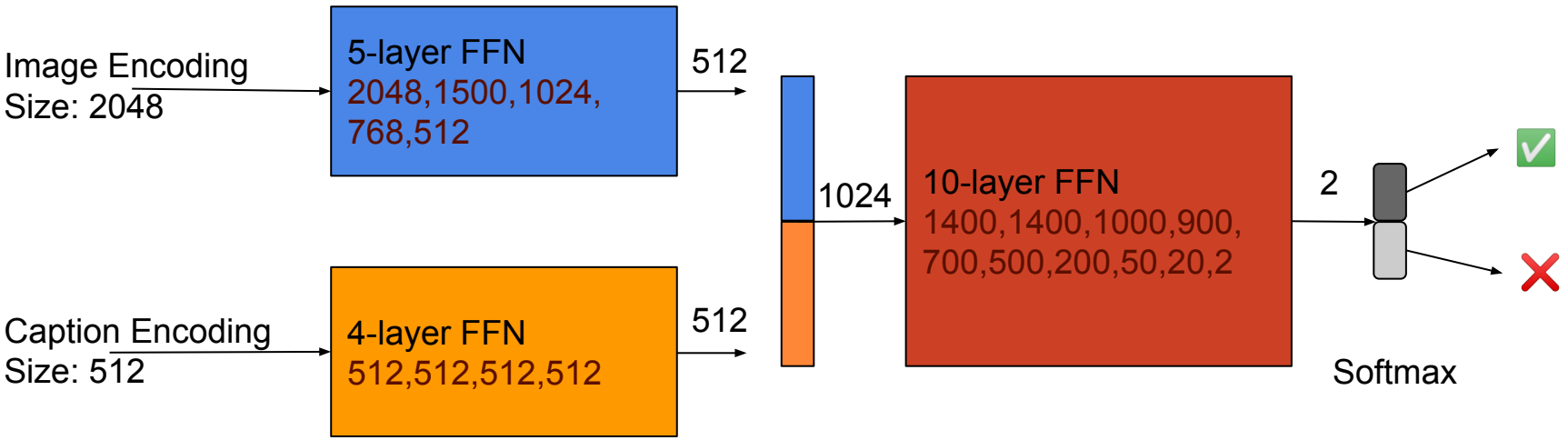
Creation of Negative Samples:

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

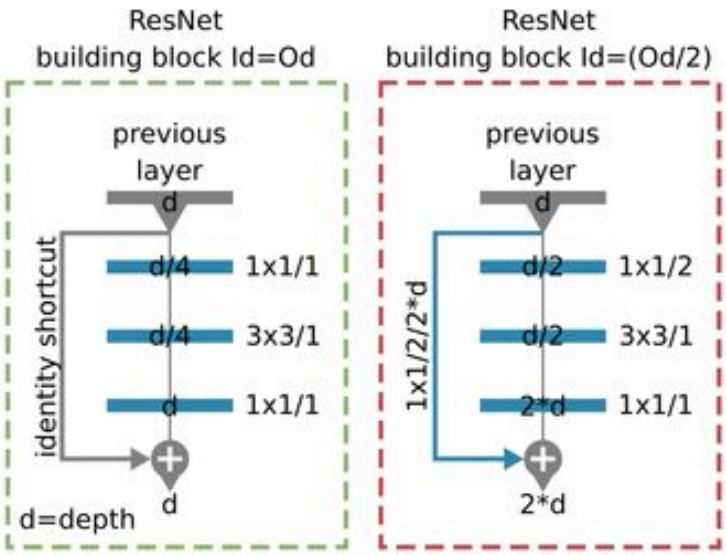
Details of the FFNN N/W

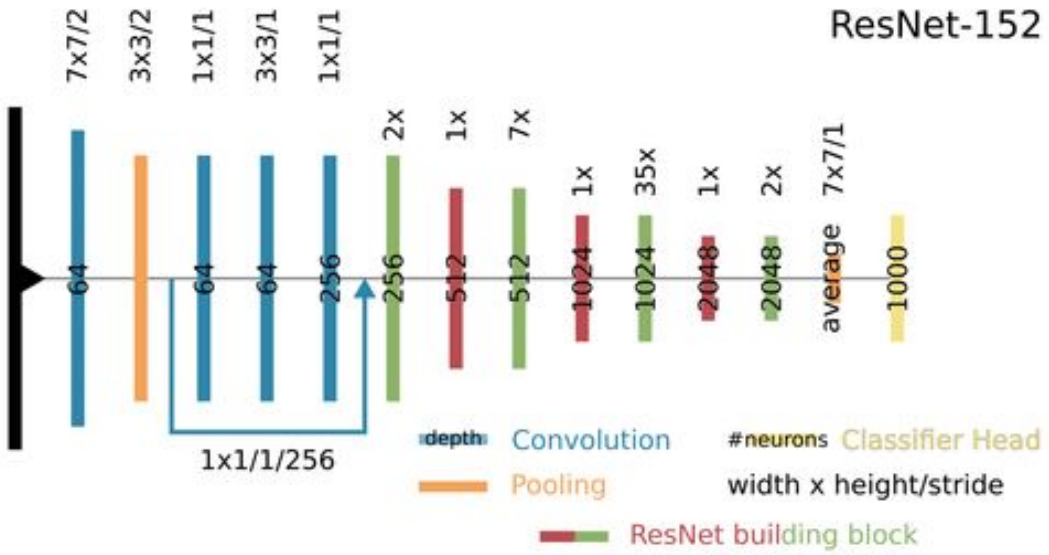
- Layers: Two FFN of 5 and 4 layers followed by 10 layer FFN
- Different Hyper parameters: Learning rate: 10^{-4}, Adam : 10^{-3}
- Model diagram



Details of the CNN N/W

ResNet152





ResNet -152 Architecture

Hoeser, Thorsten & Kuenzer, Claudia. (2020). Object Detection and Image Segmentation with Deep Learning on Earth Observation Data: A Review-Part I: Evolution and Recent Trends. Remote Sensing. 12. 10.3390/rs12101667.

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

CLIP: Contrastive Language-Image Pre-training
ViT: Vision Transformer

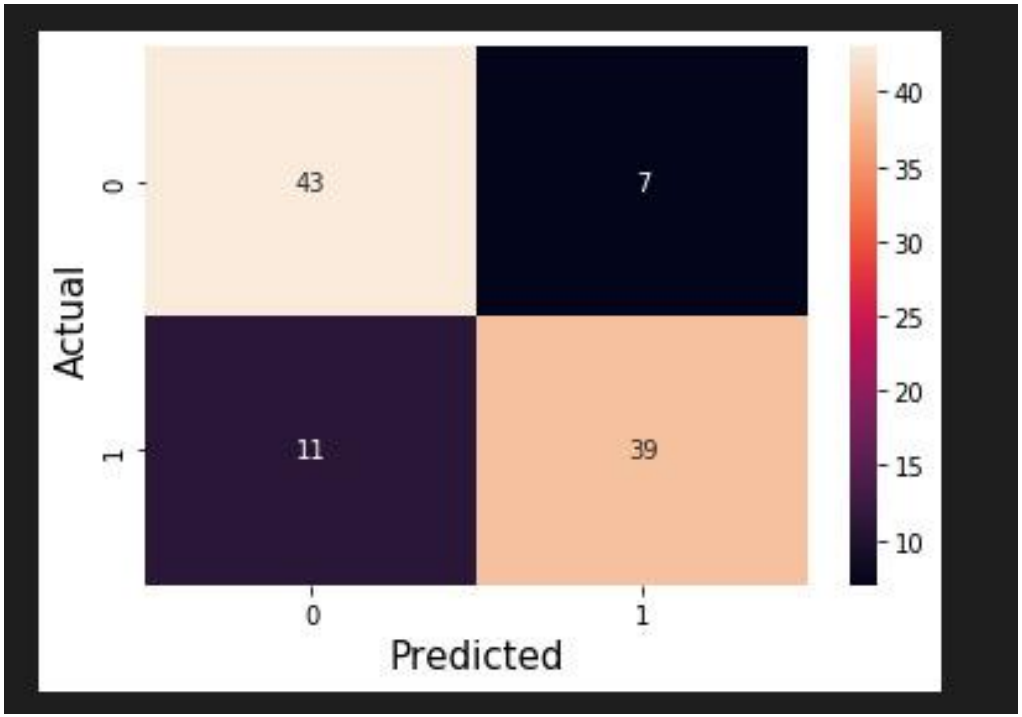
Training details (hyper-parameters)

- 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

	precision	recall	f1-score	support
0.0	0.7963	0.8600	0.8269	50
1.0	0.8478	0.7800	0.8125	50
accuracy			0.8200	100
macro avg	0.8221	0.8200	0.8197	100
weighted avg	0.8221	0.8200	0.8197	100

Performance Parameters



True Negatives : 43

True Positives : 39

False Positives : 7

False Negatives : 11

Error Analysis - False Negatives

A young girl inhales with the intent of blowing out a candle.

(Actual 1 , Predicted 1)

Girl blowing out the candle on an ice-cream

(Actual 1, Predicted 0)

[False Negative]

Error Analysis - False Negatives

a woman is holding a cat in her kitchen

(actual 1, predicted 1)

A girl smiles as she holds a cat and wears a brightly colored skirt.

(actual 1, predicted 0) [False Negative]

Major Takeaway

Just because a model's accuracy is low, it doesn't necessarily mean model is erroneous.

It could be because of incorrect ground truth also.

Error Analysis - False Positives

a young boy barefoot holding an umbrella touching the horn of a cow
(Actual 0 , Predicted 1)
[False Positive]

A man with a red helmet on a small moped on a dirt road.
(Actual 1, Predicted 1)

Similarity b/w above sentences:**0.503**

Error Analysis - False Positives

A young boy stares up at the computer monitor.
(Actual 1 , Predicted 1)

A young girl is preparing to blow out her candle.
(Actual 0, Predicted 1)
[False Positive]

Major Takeaway

Some times model might get confused because of the ambiguity involved in the test instance.

Good Examples

Performance Analysis of the model

Performance Analysis - True Positives

Food cooks in a pot on a stove in a kitchen.

(Actual 1 , Predicted 1)

[True Positive]

Some food sits in a pot in a kitchen.

(Actual 1 , Predicted 1)

[True Positive]

Performance Analysis - True Negatives

A woman in a room with a cat.

(Actual 0 , Predicted 0)

[True Negative]

A group of people sitting at desk using computers.

(Actual 0 , Predicted 0)

[True Negative]

a young boy barefoot holding an umbrella touching the horn of a cow.

(Actual 1 , Predicted 1)

Background Slides

Other hyperparameters of CLIP-VIT B/32

Hyperparameter	Value
Batch size	32768
Vocabulary size	49408
Training epochs	32
Maximum temperature	100.0
Weight decay	0.2
Warm-up iterations	2000
Adam β_1	0.9
Adam β_2	0.999 (ResNet), 0.98 (ViT)
Adam ϵ	10^{-8} (ResNet), 10^{-6} (ViT)