- CPSC 8430: Deep Learning –
- Homework 2 –
- Video caption generation –

-By Shounak Kulkarni
(shounak@g.clemson.edu)

- https://github.com/ShounaKulkarni/Deep_Learning_8430_Homework2.git -

❖ Problem Statement – Create a video caption for an input video using sequential-to-sequential model.
Input – video clip.
Output – Meaningful captions which will describe the video clip.

❖ Gathering all the requirements –

- Python, palmetto jupyter notebook instance with 2 GPU's and 8 cores.
- Libraries – torch, scipy, numpy, pandas, pickle.
- Datasets –
  MSVD - 1450 videos for training and 100 videos for testing.
  Dataset given by professor in google drive – "MLDS_hw2_1_data.tar.gz" - https://drive.google.com/file/d/1RevHMfXZ1zYjUm4fPU1CfFKAjyMJdgJ/view
- .json for Training and testing purpose.

❖ Approach towards solving this homework:
Model – seq-to-seq Model.
This model will consume sequence of video frames and give sequence of words for the output.

❖ Data Structures –
The seq2seq model requires a dictionary with specific tokens.

Tokenization:

1) <PAD>: It is used to pad sentences to ensure they are the same length.
2) <BOS>: This Indicates the beginning of a sentence and used to generate the output sentence.
3) <EOS>: This token Indicates the end of a sentence and is used to signal the end of the output sentence.
4) <UNK>: It is used as a placeholder for unknown words in the dictionary. Paraphrase.

5) The "word_dict" is a vocabulary generated from the training label file that indicates the frequency of each word. Words with a frequency of less than four are disregarded.
6) The "w2i" dictionary maps each word in the vocabulary to a unique index.
7) The "i2w" dictionary is the reverse mapping of "w2i" and maps each index in the vocabulary to its corresponding word.

❖ Models and details –

1) The sequential-to-sequential model consists of two layers: the encoder and decoder, which are designed using the Gated Recurrent Units (GRU).
   GRU is preferred over LSTM because it has fewer training parameters, which makes it faster and consumes less memory.
2) This is particularly useful in datasets with shorter sequences of data, where the GRU can achieve similar accuracy as LSTM but with faster processing times. The encoder layer takes in the video as input and encodes it into the necessary format, while the decoder layer is used to segment the captions based on the beginning and ending tokens.
3) The decoder then processes the captions and generates actual words by performing video processing over the words.
4) Attention Layer - The Attention Layer is an important feature of the sequential-to-sequential model that enables the model to selectively focus on specific parts of the input sequence at each decoding time step.
   The Attention Layer is based on the research presented in the paper "Attention-based convolution neural network for semantic relation extraction" authored by Shen and Huang.
   The Attention Layer achieves this functionality by computing a matching function between the hidden state of the decoder and the output of the encoder. This matching function is then transformed by a softmax activation function to produce attention weights that determine which parts of the input sequence are most relevant to the current decoding step. The final hidden state produced by the Attention Layer is then used as input to the next time step of the decoder, improving the accuracy and relevance of the generated captions.

***

The dataset is loaded to “/scratch1/shounak/” instance on palmetto as mentioned in code. The user can change the path in.sh file in hardcoded way. The output will be generated in palmettos jupyter notebook local instance.

***

- The sequential-to-sequential model was trained with the following parameters:

  1) Epochs = 100
  2) Learning Rate = 0.0001
  3) Batch Size = 128
  4) Hidden Layers = 512
  5) Optimizer = Adam Optimizer
  6) Dropout = 0.3
  7) Teacher Learning Ratio = 0.7
  8) Vocabulary Size = n > 4

Training was done for batches with the interval of 50 clips till 950 video clips. Then killed the process.

- Caption dimension: (24235,2)
- Captions max length: 40.
- Average length of captions: 7.723198762545078
- Unique tokens: 6451
- ID of 28th video: _txL575S_OA_13_23.avi
- Shape of features of 28th video: (888,608)

The model achieved a BLEU score of 0.6974300347116 on the test data that was used for evaluation.
Highest bleu scores: [0.6994, 0.4390]
The loss of the model in the last epoch was recorded as 2.27886.

- References/Bibliography –

  - https://gist.github.com/vsubhashini/38d087e140854fee4b14.
  - https://paperswithcode.com/paper/attention-based-convolutional-neural-network-2