Jordyn Brooks
CPSC 8430 - Deep Learning
Homework 2
Github link: https://github.com/brooksjk/Deep_Learning - under folder HW2/hw2_1

In this project, a sequence-to-sequence (Seq2Seq) model with attention mechanisms is trained and evaluated for video caption generation. The model's performance is measured using the BLEU score, a widely used metric for evaluating the quality of machine-translated text compared to human-generated reference translations. The results are calculated based on the predictions generated by the trained model and the provided test captions.

Seq2Seq Model:

Model Architecture:

Encoder: Extracts features from the input video sequences.

Decoder: Generates captions based on the features provided by the encoder.

Attention Mechanism: Allows the decoder to focus on specific parts of the input while generating each word in the output sequence.

Data:

“MLDS_hw2_1_data”

Dataset: The model was trained on a dataset of video captions. Each video in the dataset is associated with multiple human-generated captions, which serve as the ground truth during evaluation.

training_data/training_label.json: Used to train the Seq2Seq model.

testing_data/testing_label.json: Contains video features and human-generated captions, used for evaluating the model performance.

Reference Captions: Captions from the test set used to compute BLEU scores.

<u>Methodology</u>:

- Loading the Trained Model:
  - The trained Seq2Seq model was generated by running the training.py file using train.sh then loaded from the saved .h5 file using PyTorch.
- Evaluation Metric:
  - BLEU Score was calculated using the following parameters:
    - n-gram: The BLEU score was computed for unigrams (n=1) in this project.
    - Precision: The ratio of overlapping words between the predicted captions and the reference captions.
    - Brevity Penalty: A penalty applied to short predictions to prevent the model from generating very short sentences that match the reference captions only partially.
- BLEU Score Calculation:
  - For each video:
    - Predicted captions were compared against the reference captions provided in the test set.
    - The BLEU score was computed using a custom BLEU function from bleu_eval.py, which calculates n-gram precision, brevity penalty, and geometric mean of precisions.
    - The original implementation of clip_count contained a potential issue where the max() function could raise a ValueError when the argument passed to it (a sequence of matching n-grams across reference captions) was empty. This situation occurred when no matching n-grams were found between the candidate caption and the reference captions, leading to an empty sequence for max() to evaluate. This was mediated by modifying the code to account for cases where there are no matches for the candidate n-gram in any of the reference captions.
    - Final Evaluation: The average BLEU score across all videos was computed to assess the overall performance of the model.

Results:

Predictions: The predictions generated by the model were saved in the output file test_output.txt in the following format:

Test ID	Caption
12345	a boy is playing football
67890	a dog is running in the park

Average BLEU score: 0.7187544630684313

Discussion:

- Strengths
  - The model was able to generate reasonable captions for most videos in the test set. The use of attention allowed the decoder to focus on relevant parts of the video when generating the caption, leading to higher accuracy in some cases.
  - A BLEU score of ~0.719 signifies that the generated captions from the model were 71.9% accurate when compared to the reference captions.
- Weaknesses
  - There were instances where the generated captions were either too short or incomplete, resulting in lower BLEU scores for those cases. The brevity penalty significantly impacted the BLEU score when the model produced captions shorter than the reference.