Image Caption Generation

Name: Shilpa Gopalakrishna

QUESTION I

1.1 Text preparation

utils.py.

1.2 Extracting image features

extract_features.py file.

1.3 Training DecoderRNN

decoder.py file.

QUESTION II

2.1 Generating predictions on test data

2.1.1 Present three sample test images showing different objects, along with your model's generated captions and the 5 reference captions.

Image	Reference captions	Model generated caption
390987167_2d5905b459 	• a woman wearing a black hat and white sunglasses walks up a snowy and steep path • a woman wearing short sleeves and sunglasses is backpacking up a snowy mountain cliff • a woman in a tank and sunglasses climbs a sunny but snow covered slope • woman climbing a snow covered mountain looking at people behind her • a hiker moves up a snowy mountain	• two people stand on a mountain overlooking a mountain

391723162_3bdeb7ea33 	• a small black and brown dog is standing in the snow • a large black dog is digging in the deep snow • a black dog digging through the snow • a black dog digs in the snow • black dog digging in white snow	• a dog is jumping in the air over a snowy hill
950411653_20d0335946 	• a boy jumps into the pool • an adult watches a child somersault into the pool while another child looks on • a woman throwing a little boy into the pool • woman in blue is tossing a young child into a pool • woman throws little boy into pool as another boy watches	• a boy in a swimming suit is jumping off a diving board into a pool

2.2 Caption evaluation via text similarity

(1) BLEU for evaluation

2.2.1 Report the trained model's performance on the test set using the BLEU method and discuss.

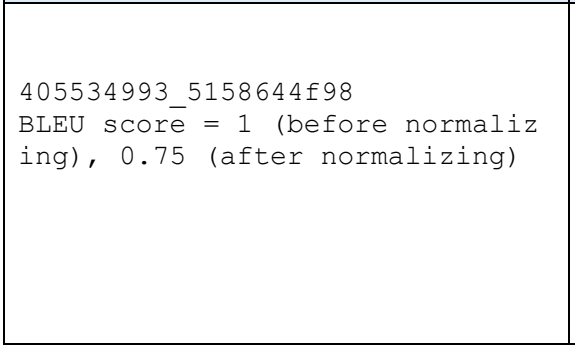
- **Observation of BLEU Score on our test set:**
  - All the 5-reference caption is passed to BLEU function along with the predicted caption.
  - 4-gram method ie Weight of 0.25,0.25,0.25,0.25 was used for BLEU scoring. Each 1-gram, 2-gram, 3-gram & 4-gram of reference caption is compared with the predicted caption. Cumulative scores calculate individual n-gram scores from 1 to 4 and weighs them by calculating the weighted geometric mean.

  - **Model's Performance:**

- Without using smoothing function, the average BLEU score across the 7K records in test set was too less because most of the predicted captions were shorter than the reference captions and the word-to-word match was less.
- On applying the method4 smoothing, the average BLEU score went up to **42%**, this is because smoothing function adds a small empirically determined count if the n-gram match is 0. method4 smoothing was used in the coursework which takes into consideration of the length of the predicted sentence while smoothing.

- **About BLEU score approach in general:**
  - The prediction on test data is evaluated using BLEU method.
  - Sentence_bleu method is used to evaluate sentence to sentence comparison.
  - Code Used:
    - sentence_bleu(reference,candidate,weights=(0.25, 0.25, 0.25, 0.25), smoothing_function=cc.method4)
  - **Weights** used for sentence_bleu is 0.25, 0.25, 0.25, 0.25. This is called cumulative 4-gram Bleu score. Cumulative scores calculate individual n-gram scores from 1 to 4 and weighs them by calculating the weighted geometric mean.
  - **Smoothing function** – BLEU score performs poorly on sentence scoring because it computes a geometric mean of n-gram precisions, if a higher order n-gram precision of a sentence is 0, then the BLEU score of the entire sentence is 0, irrespective of 1-grams or 2-grams are matched.
  - Method4 smoothing has been used here. In method4, if a matched n-gram count is 0, then we use a small value as count instead of 0. The count is determined empirically and in method4 the count is penalized if the sentence is too short.

2.2.2 Present one sample test image with a high BLEU score and one sample with a low score, along with your model's generated captions and the 5 reference captions.

One sample with high BLEU score		
Image	Reference captions	Model generated caption
 405534993_5158644f98 BLEU score = 1 (before normalizing), 0.75 (after normalizing)	• a rock climber • there are two people rock climbing one is on the ground while the other climbs • two men climb a large rock • two people rock climbing	• a man is rock climbing

	• two rock climbers scaling a sheer cliff	
One sample with low BLEU score		
407678652_1f475acd65 BLEU Score=0.14 	• a group of hikers climb rocky terrain • five backpackers hiking along a rocky slope • five hikers walk down the hillside • five people with backpacks hiking down a rocky hill • the hikers descend on the mountain	• two children and a man in a red coat and a blue and white shirt and black pants is

(2) Cosine similarity for evaluation

2.2.3 Report the trained model's performance on the test set using the cosine similarity method, and discuss.

- **Observation of Cosine Similarity Score on our test set:**
  - Here, in the coursework, Sklearn's "sklearn.metrics.pairwise.cosine_similarity" is used to evaluation
  - Cosine similarity is applied to compare the reference caption with predicted caption. Unlike BLEU (takes the tokenized sentence as input), here the input to cosine is a reference caption's numpy array & predicted caption's numpy array.
  - To compute the numpy array for reference & predicted caption, below was coded
    - Retrieve the id for each word from vocab.
    - Pass the idx to embedding and retrieve the vector representation of the word from embedding for each work in the reference caption. (Gives us a tensor of (num_of_words_in_caption, embed_size_256)

- Compute the average of the word vector which gives 1,256 as the response. Convert this to numpy array
- Now, the computed average vectors of reference and predicted is passed to cosine_similarity that outputs similarity value for each reference caption & predicted caption. As there are 5 reference captions for each image, we finally average the 5 cosine_similarity values to get the final similarity score for a single prediction

➢ **Model's Performance:**

- As the cosine value is derived from the embedding vector, the cosine score performs better than BLEU score. Cosine takes into consideration the context of the words in embedding vector space. So, the words close to each other in vector space gets high similarity.
- In our case, cosine similarity did a better comparison on manually verifying the results for few records.

- **About Cosine Similarity score approach in general:**
  - ➢ Unlike BLEU score which used word-overlap based metrics, Cosine can be used with word embedding which considers the meaning of each word as defined by a word embedding that assigns a vector to each word. The sentence-level embeddings between the candidate and target response are compared using a measure such as cosine distance.
  - ➢ In the coursework, the embedding average metric calculates sentence-level embeddings, a method for computing the meanings of phrases by averaging the vector representations of their constituent words.
  - ➢ The cosine similarity between ground truth and predicted captions is computed using their respective sentence level embeddings.

2.2.4 Present one sample test image with a high cosine similarity score and one sample with a low score, along with your model's generated captions and the 5 reference captions.

One sample with high cosine similarity score		
Image	Reference captions	Model generated caption
`799199774_142b1c3bb2` Cosine similarity: `0.80`	• a boy plays in a pool with an inflatable toy • a boy swimming in a pool • a child on a pink raft in a pool • a small boy swims with a pink floatation device in a swimming pool	• a boy in a swimming pool

	• boy playing on a pink raft in a pool	
One sample with low cosine similarity score		
536828916_b763b82949 Cosine similarity: 0.08 	• a curly redheaded girl with a large headband in her hair • a girl with a long curly red hair and green eyes • a redhaired woman looks off camera • a red haired woman looks past the cameraman • the redheaded woman is staring off into the distance	• two women in white shirts and black sunglasses

2.3 Comparing text similarity methods [15 marks]

2.3.1 Compare the model's BLEU and cosine similarity scores on the test set and identify some weaknesses and strengths of each method.

BLEU score:

Strengths:

- BLEU scoring evaluation is said to report a high correlation with human judgments of quality as it does sequential evaluation.
- BLEU scoring is independent of language.
- Can be used as a metric for quick evaluation

Weakness:

- BLEU scores tell us how a system performs on the specific set of reference sentences and the translations selected for the test. As there can be alternative ways to imply the same sentence, it can score a correct predictions/translation poorly. Hence, the scores do not always reflect the actual performance of a system if the content differs from the specific test set.

- BLEU does not aim to measure overall translation quality but focuses on strings. Hence BLEU scores may be more accurate for corpus level comparisons than at sentence level.
- The BLEU metric gives higher scores to sequential matching words. A string of four words in the prediction matching the reference captions in the exact order will have more of a positive impact on the BLEU score than a string of two matching words. An accurate translation receives a lower score if it uses different, but correct words or matching words in a different word order.
- If the predicted text has the exact words in different order, then the BLEU scores less which is again not a right approach.

Cosine similarity:

Strengths:

- Cosine is distance measure that disregards differences in **magnitude** and focuses on the proportions of features, it is a suitable metric for text comparison.
- Contextual information adds to the final score than word to word match.
- Works well with the hundreds of dimensions in a word embedding vector.

Weakness:

- As we are averaging the vector of words from reference & predicted captions, there is an information loss, and the resulting score may not be accurate, and the score may not reflect the actual distance in the vector space.
- Vectors are usually normalized to use this metric effectively, but word embeddings from various models carry semantic significance in a vector's length as well as the direction and as cosine disregards differences in magnitude, leads to inaccurate scores.

Comparison between 2 scores with respect to our test set:

- Whenever BLEU score resulted in higher accuracy, it means that there is mostly word to word match in the prediction. So, what ever scored high in BLEU score can be considered as a good prediction and matches with the reference captions.
- When BLEU score showed less score, then we need to check if the prediction has synonyms or if the prediction is framed using different words but provides similar meaning to reference caption.
- When Cosine score is high, it means that the context of the prediction matches well with the context of the prediction. Though the exact words or order might not be used in the prediction, the overall meaning of the prediction would reflect the ground truth context.
- When cosine is negative or near to 0, it indicates that the predicted words are no where related to reference caption and prediction is done poorly.
- There were certain totally unrelated and poor prediction that mostly both BLEU and Cosine gave low score.

- Cosine did not score high on an average whereas BLEU gave a high score or a low score on an average.

2.3.2 Show one example where both methods give similar scores, and another example where they do not and discuss.

Similar Score from BLEU & Cosine:

- **Image -** 407678652_1f475acd65
- **BLEU score = 0.14**
- **Cosine score = 0.05**
- **Actual captions:**
  - a group of hikers climb rocky terrain
  - five backpackers hiking along a rocky slope
  - five hikers walk down the hillside
  - five people with backpacks hiking down a rocky hill
  - the hikers descend on the mountain
- **Predicted caption:**
  - two children and a man in a red coat and a blue and white shirt and black pants is

- **Explanation:**
  - As the predicted caption is nowhere related to the reference caption, both the scoring metrics resulted in low score.
  - BLEU metric checks for word to work sequence in 1/2/3/4 grams approach. The prediction has no word correlation with the reference, hence the low BLEU score.
  - Cosine resulted in low score too because there is no word or context in prediction that reflects the context of reference caption. Even if there was one word something like hill/hike then cosine would have scored a bit higher.

Differing Score from BLEU & Cosine:

- **Image -** 693785581_68bec8312a
- **BLEU score = 1.2 (before normalizing), 0.72 (after normalizing)**
- **Cosine score = 0.12**
- **Actual captions**:
  - a boy does a flip while another boy stands on the matt
  - a boy standing on a red mat with a person upside down in the background
  - a child does a flip behind a small boy
  - a young boy is standing near another doing a back flip
  - two kids playing on mats
- **Predicted caption**:
  - two children are playing soccer

- **Explanation:**
  - Cosine scores for the prediction against each reference is as below.
    - `0.05, 0.01,0.02,0.06,`**`0.44`**
    - Average `= 0.12`
  - Cosine score for the last prediction which has kids and play matches and hence the score for last reference is 0.44. The overall average has come down due to the absence of most of the context words in other sentences.
  - BLEU score seems to have found a better match word to word in terms of the last reference caption and hence the score has gone up. Also, to note here that the number of the words in prediction is same as that of 5th reference caption, so this might have been a factor also for the high score. If the sentence was too long or fewer words in prediction, then BLEU would have given less score.

Marks reserved for overall quality of report. [5 marks]

No response needed here.