

COMP5623 Coursework on Image Caption Generation

Name	Kunhao Liang
Student ID & username	201469850

QUESTION I [40 marks]

1.1 Text preparation [15 marks]

Please submit your *utils.py*.

1.2 Extracting image features [10 marks]

Please submit your *extract_features.py* file.


1.3 Training DecoderRNN [15 marks]



Please submit your *decoder.py* file.

QUESTION II [60 marks]

2.1 Generating predictions on test data [10 marks]

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
	<ol style="list-style-type: none">1. one dog is standing whilst two other dogs are running in the snow2. three dogs are playing around in the snow3. three dogs chasing each other in the snow4. three dogs play in snow5. two dogs play together in the snow	a dog is running on the beach

	<ol style="list-style-type: none"> 1. four dogs play in the snow with the city skyline behind them. 2. the dogs are playing in the snow. 3. three dogs are in the snow , and one is wearing a winter coat. 4. three dogs playing in the snow while one dog wears a jacket. 5. three dogs playing in the snow , with a city in the background. 	<p>a group of people are playing in the snow</p>
	<ol style="list-style-type: none"> 1. a white dog is running along a path outside. 2. a white dog travels along a narrow path in a park setting. 3. a yellow dog running along a forest path. 4. the two-tone dog is running down the trail. 5. white dog traveling alone down a paved path through some woods. 	<p>a white dog is running through the snow</p>



2.2 Caption evaluation via text similarity [30 marks]

(1) BLEU for evaluation

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

The overall average BLEU score is 0.56. This indicates that the trained model could extract some information from reference captions, but fails to generate accurate caption that can describe the image properly.

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
	<ol style="list-style-type: none"> 1. a black dog splashes through the water. 2. a brown and tan dog is running through shallow water. 3. a dog is running in the ocean beside the beach. 4. a dog running through water. 5. dog splashes running across water. 	a black dog is running through the water
One sample with low BLEU score		
	<ol style="list-style-type: none"> 1. two guys are playing horse shoe together. 2. two guys playing horseshoe. 3. two men playing horseshoes. 4. two men wearing jeans and sunglasses are playing horseshoes. 5. two people play horseshoes. 	a boy runs through the grass

(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.

The overall average cosine similarity score is 0.43, which shows that the model could generate fairly accurate captions, but it might fail to capture some keywords included in reference captions.

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
	<ol style="list-style-type: none">1. a bearded man in white clothes is sitting on a long bench2. a man dressed in white sitting on a bench3. a man in a white outfit on a bench4. a man wearing white sits on a wooden bench against a white wall5. an middle eastern man in a white robe is sitting on a wooden bench with his shoes off	a man in a white shirt and tie is sitting on a bench
One sample with low cosine similarity score		
	<ol style="list-style-type: none">1. a man and two boys standing in spraying water2. a shirtless male looks to his right while water flows over him3. two boys in swimsuits standing under running water4. two boys playing in water5. young boys enjoying a spray of water	two boys are playing in a fountain

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	Cosine Similarity Score(rescaled)
Overall Average Score	0.56	0.51
First five scores	1.0, 0.79, 0.61, 0.59, 0.48	0.71, 0.42, 0.14, 0.64, 0.34

BLEU is computed using a couple of ngrams, which can efficiently give us a fairly good result which is close to human evaluation. But the drawback of BLEU is that it calculates the score irrespective of the meaning and the structure of sentence, as well as the synonyms and other expressions that have the same meaning. Cosine similarity has the advantage of being simple, particularly for sparse vectors. However, it does not take into account the size of vectors, which means the differences in values is not fully considered.

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

Cosine similarity score: 0.22; **BLEU score:** 0.72

Predicted caption: two dogs play in the water.

Reference caption:

1. a brown dog leaps into the water
2. a dog leaping off a boat
3. a dog wearing a collar jumping from a platform
4. a grey and brown dog jumps off a dock into a lake
5. a light brown dog with his tail in the air jumps of a pontoon toward the water



For this sample, BLEU gives a high score of 0.72, while its cosine score is only 0.22. Despite the fact that the predicted caption does not describe the picture properly, it receives a high BLEU score because it has many overlaps of n-grams. In contrast, in this case, cosine similarity provides us with a more precise evaluation of the generated caption.

Cosine similarity score: 0.865; **BLEU score:** 0.862

Predicted caption: a man in a blue shirt climbs a rock wall.

Reference captions:

1. a man in a pink shirt climbs a rock face
2. a man is rock climbing high in the air
3. a person in a red shirt climbing up a rock face covered in assist handles
4. a rock climber in a red shirt
5. a rock climber practices on a rock climbing wall



As it can be seen, the predicted caption is very close to what is shown in the image. Hence, it is reasonable that we receive a high cosine score because the word vectors projected in a multi-dimensional space would have a small angle. We also obtain a high BLEU score since it has many overlaps of n-grams.

To sum up, we should pay attention to both evaluation metrics when evaluating the model performance to get an accurate result.

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

No response needed here.