

COMP5623 Coursework on Image Caption Generation

Name	Benjamin Jeffrey
Student ID & username	201479413, mm20bwj

QUESTION I [40 marks]

1.1 Text preparation [15 marks]

Please submit your <i>utils.py</i> .

1.2 Extracting image features [10 marks]

Please submit your <i>extract_features.py</i> file.


1.3 Training DecoderRNN [15 marks]



Please submit your <i>decoder.py</i> 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
	<ul style="list-style-type: none">• a black dog on a white carpet next to an orange• a black puppy is playing with an orange on a carpeted floor• a black puppy standing inside beside an orange• an orange on the floor next to a dog• the black dog is standing by the wall next to the orange	a black dog jumps over a hurdle

	<ul style="list-style-type: none"> • two children laying on the floor with a group of wooden blocks above their heads • two children re laying on a rug with some wooden bricks laid out in a square between them • two kids lie on a rug near wooden blocks and smile • two little girls lie on the carpet next to an o made of wooden blocks • two young girls lay on the carpet next to wooden blocks 	<p>two children are sitting on a bench</p>
	<ul style="list-style-type: none"> • a man fishes in the ocean • a man fly fishes in the ocean • a man fly fishing • a man is fishing in kneehigh waves of water • the person is fishing with waves splashing around him 	<p>a man is wakeboarding in a rowboat</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 used nltk package computes the cumulative 4-gram BLEU score by default, which resulted in a low score of 0.065 for our model evaluated over the entire test set. This is due to the equal weighting being given to 1-, 2-, 3- and 4-gram matches using the default settings. Getting a lot of 4-gram matches is quite a high hurdle and maybe is not a realistic expectation to have for our captions, especially since the reference caption can vary quite a bit themselves. We played around with the weighting being given to the 1-, 2-, 3- and 4-gram matches and got a score of 0.52 for single words (1-gram) and 0.32 for single words and word pairs (cumulative 2-gram). We felt most comfortable with using the cumulative 2-gram score, as this does take context into account more than just using single words, but also does not set the bar too high for the level of similarity we expect from the generated captions.

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
	<ul style="list-style-type: none"> • a black dog splashes through the water • a brown and tan dog is running through shallow water • a dog is running in the ocean beside the beach • a dog running through water • dog splashes running across water 	a black dog is running through the water
One sample with low BLEU score		
	<ul style="list-style-type: none"> • a dog carries a leash in its mouth • a fluffy dog carries a black leash in its mouth • a large furry brown dog is walking with a leash in his mouth • a yellow dog carries a black leash • dog walking with his lease in his mouth 	two dogs play together

(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.

Using the word embeddings stored in the decoder to calculate averaged vectors for the captions and then evaluating the similarity of these vectors based on cosine similarity, our model achieved a score of 0.54 when evaluated over the entire test set. This seems like a reasonable score given the task and the generated captions. There aren't any parameters to tweak using this approach.

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
	<ul style="list-style-type: none"> • 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 • boy playing on a pink raft in a pool 	a boy in a swimming pool swims underwater
One sample with low cosine similarity score		
	<ul style="list-style-type: none"> • a toddler starting at a clown walking down a snowy sidewalk • several people including a child and a clown are walking towards a snowy sidewalk • three adults and a toddler stand on a snowy path • three adults and one child are walking along a paved path that is snow covered • two adults a child and a clown walking down a sidewalk 	two boys in blue shirts are running on the beach



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.

Directly comparing the two scores to each other in absolute terms does not make much sense, since the method by which the scores are derived are so different. So, the difference between our BLEU score (0.32) and Cosine Similarity score does not necessarily tell us much. The BLEU score calculation is fast and does not take much in terms of resources, which would be a serious advantage if we used it during training for instance. However, it does not really consider meaning at all and just looks for words appearing in a similar fashion to the reference ones. A human could probably think up a caption quite easily for most cases that is entirely accurate at describing the image but produces a BLEU score of 0. Word embeddings on the other hand are known for

capturing meaning well, which indicates the Cosine Similarity approach might perform better in this aspect. Building the captions from the embeddings and the subsequent calculations are more resource intensive however and we are putting a lot of trust into captions retaining enough information from all the word vectors averaged together. In conclusion both approaches have their advantages and drawbacks, and it never hurts to use multiple evaluation metrics to verify model performance in order to get a more complete picture.

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

One sample with similar scores (BLEU: 0, Cosine Similarity: 0)		
Image	Reference captions	Model generated caption
	<ul style="list-style-type: none"> • a toddler starting at a clown walking down a snowy sidewalk • several people including a child and a clown are walking towards a snowy sidewalk • three adults and a toddler stand on a snowy path • three adults and one child are walking along a paved path that is snow covered • two adults a child and a clown walking down a sidewalk 	two boys in blue shirts are running on the beach
One sample with different scores (BLEU: 0, Cosine Similarity: 0.7)		
	<ul style="list-style-type: none"> • a climber traverses a rope ladder between cliffs • a long woman wearing a white helmet crossing a chain linked bridge • a man is crossing a mountain pass on a metal bridge • a woman is walking across a step ladder bridge between two mountains 	a person in a red jacket climbs a snowy mountain

	<ul style="list-style-type: none"> • a woman traverses a wire bridge across a rocky canyon 	
--	---	--

With the first example both metrics resulted in a score of 0 for the generated caption. This happens quite a lot for BLEU given how the algorithm works, but for Cosine Similarity this is a result of the min-max scaling we applied to bring the score into the same range. Both approaches rightfully detect that the generated caption is terrible with basically no part of it describing what is visible in the picture. In the second example the scores vary quite drastically. The BLEU algorithm computed a score of 0 for the generated caption, which is not very sensible, given the image does show a person in the mountains. On the other hand, the Cosine Similarity score of 0.7 seems far too high as there is not actually any snow in the picture and the woman is not wearing red. Both approaches seem to struggle with this example therefore and neither is very accurate.

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

No response needed here.