

# COMP5623 Coursework on Image Caption Generation

Name	PRAVEEN GOPAL REDDY
------	---------------------

## 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.
---

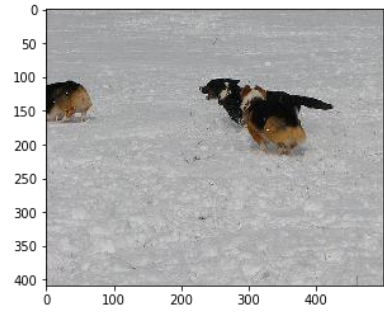
### 1.3 Training DecoderRNN [15 marks]



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
	<ol style="list-style-type: none"><li>1. One dog is standing whilst two other dogs are running in the snow.</li><li>2. Three dogs are playing around in the snow.</li><li>3. Three dogs chasing each other in the snow.</li><li>4. Three dogs play in snow.</li><li>5. Two dogs play together in the snow.</li></ol>	three dogs are running through the snow

	<ol style="list-style-type: none"> <li>1.A man in a pink shirt climbs a rock face.</li> <li>2.A man is rock climbing high in the air.</li> <li>3.A person in a red shirt climbing up a rock face covered in assist handles.</li> <li>4.A rock climber in a red shirt</li> <li>5.A rock climber practices on a rock-climbing wall.</li> </ol>	<p>a man is climbing a rock wall</p>
	<ol style="list-style-type: none"> <li>1.A black dog is swimming while carrying a tennis.</li> <li>2.a black dog swimming in the water with a tennis ball in his mouth</li> <li>3.a black dog swims through the water with a tennis ball in its mouth.</li> <li>4.A dog swims in water with a blue and green tennis ball in its mouth.</li> <li>5.A dog with a ball in its mouth swims in the water.</li> </ol>	<p>a black dog is running through the water</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.

n-gram	bleu score = 0	0<bleu score <=0.50	0.50< bleu score<=1.0	Total test dataset
Bleu 1-gram	5	3020	1990	5015
Bleu 2-gram	5	4315	695	5015
Bleu 3-gram	5	4685	325	5015
Bleu 4-gram	5	4875	135	5015

The above table shows frequency of bleu scores between 0 to 1 on test data.

When we ran script on test data using bleu method, we got following averages for different n-grams using cumulative method:

The overall average for bleu-1-gram is **0.48**.

The overall average for bleu-2-gram is **0.35**.

The overall average for bleu-3-gram is **0.30**.

The overall average for bleu-4-gram is **0.26**.

Clearly, we can see bleu-1-gram has highest average because it checks only one gram match.



. The least average we got for 4-gram bleu. Because it must match more n-grams which is in this case is 4-gram. The cumulative and individual 1-gram BLEU use the same weights, e.g. (1, 0, 0, 0). The 2-gram weights assign a 50% to each of 1-gram and 2-gram and the 3-gram weights are 33% for each of the 1, 2 and 3-gram scores. The weights for the BLEU-4 are 0.25 for each of the 1-gram, 2-gram, 3-gram and 4-gram scores.

In all n-grams the number of test data with score zero is 5. That means, there is no unigram, bigram, trigram, quadram matching between predicted caption and reference caption in case of 1-gram, 2-gram, 3-gram, and 4-gram evaluation, respectively.

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.

The below samples we get result from 4-gram BLEU score.

One sample with high BLEU score		
Image	Reference captions	Model generated caption
	1.A brown dog is running through a brown field.	a brown dog is running through the grass

 <p>Image id = 408233586_f2c1be3ce1 4-gram BLEU score = 0.86</p>	<p>2.A brown dog is running through the field. 3.A brown dog with a collar runs in the dead grass with his tongue hanging out to the side. 4.a brown dog with his tongue wagging as he runs through a field. 5.A dog running in the grass.</p>	
One sample with low BLEU score		
 <p>Image id = 537230454_1f09199476 4-gram BLEU score = 0</p>	<p>1.A boy in black waves his arms while other people are behind him on a field. 2.A boy is making the victory sign with both hands. 3.A boy is standing on a grassy field with his arms raised while others are standing behind him. 4.A young boy poses in a grassy field as onlookers stand in the background. 5.People are standing in a field.</p>	<p>two boys playing soccer</p>

## (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.

Cosine score<0	Cosine score = 0	0<Cosine score <=0.50	0.50<Cosine score<=1.0	Total test dataset
108	25	4006	876	5015

The above table shows frequency of cosine scores between -1 and 1. Overall average cosine score for all test data is **0.34**. The number of cosine scores greater than 0.50 is 876. So, 17% of test data reference captions word vectors has greater than 50% alike words with prediction caption word vectors. If the score is 0.50, we can say that the angle between reference caption and prediction caption is  $\cos(60) = 1/2 = 0.5$ , which indicates 50% similarity between vectors. The number of cosine scores between 0 and 0.50 is 4006, which is 79% of test data captions.

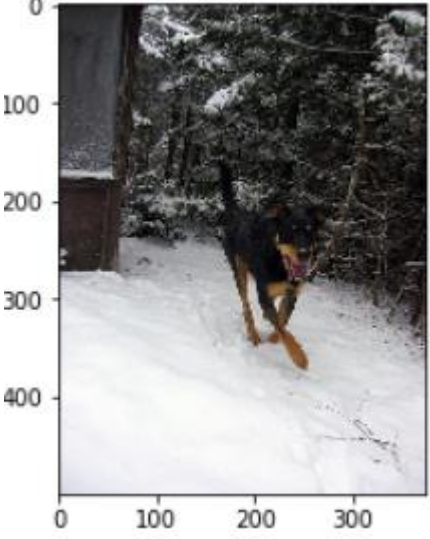

We also see there are 25 dissimilar vectors whose cosine score is 0 which indicates the angle between vectors is perpendicular or orthogonal vectors.

Normally cosine-similarity between frequency vectors cannot be negative as word-counts cannot be negative, but we have use word-embeddings with random vectors so we can have negative values. We see there are 108 negative values. A negative value indicates the angle between vectors is greater than 90 degree and vectors similarity is completely opposite to each other.

The highest score we got using cosine similarity on test data is **0.82** and lowest score is **-0.14**.

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"> <li>1. A black and brown dog walks through the snow near a building.</li> <li>2. A black dog, running in the snow.</li> </ol>	a black dog is running through the snow

 <p>Image id = 445655284_c29e6d7323 Cosine score = 0.82</p>	<ol style="list-style-type: none"> <li>3. A black dog running in the snow by some trees.</li> <li>4. A large black and tan dog is running across the snow in a wooded area.</li> <li>5. The black and brown dog is running through the snow.</li> </ol>	
One sample with low cosine similarity score		
 <p>Image id = 432248727_e7b623adbf Cosine score = -0.14</p>	<ol style="list-style-type: none"> <li>1. A brown dog howls.</li> <li>2. A brown dog in the grass field is looking up at the sky.</li> <li>3. A brown dog is crouching and looking up in a field of grass.</li> <li>4. A large brown dog is crouching in a grassy field.</li> <li>5. Brown dog crouching in grass looking up.</li> </ol>	<p>two dogs run through a field</p>

## 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.

For comparison with cosine similarity, we have used bleu 4-gram score and cosine similarity rescaled between 0 to 1 range. The below table shows, frequency of scores between BLEU and Cosine similarity.

Method	Score = 0	$0 < \text{Score} \leq 0.50$	$0.50 < \text{Score} \leq 1.0$	Total test dataset
4-gram BLEU	5	4875	135	5015
COSINE (rescale)	0	4919	96	5015

The number of values that matched between 4-gram BLEU and cosine similarity scores by considering precision to 1 is **524** and the number of values that not matched is **4491**.

### Some examples that matched:

The imageid is :385186343\_464f5fc186 ,matching score is :0.3 and caption is :the dog is wearing a pink sweater

The imageid is :390992388\_d74daee638 ,matching score is :0.3 and caption is :the woman is hiking up a snowy hill

### Some examples that not matched:

The imageid is :377872472\_35805fc143, score of Cosine and BLEU is :0.4 & 0.3 and caption is :two dogs play together in the snow

The imageid is :377872672\_d499aae449, score of Cosine and BLEU is :0.5 & 0.3 and caption is :three dogs playing in the snow with a city in the background

### disadvantages of BLEU method:

- BLEU does not measure meaning. It looks for exact word to word match. For example Consider reference caption: “the quick brown fox” and predicted caption: “the fast brown fox”. As you see quick and fast word is same meaning, but BLEU method will penalize the weight of mis-matched n-gram.
- BLEU does not consider sentence structure. For example, depending on the reference caption when you compared with “I’m bad at football” and “football bad I’m at” the bleu gives same score for both.

### advantages of BLEU method:

- Off course BLEU is faster than human translation when evaluating quality of text on large data set.



- b) BLEU is language independent and easy to understand and more over it is commonly used algorithm for machine translation systems.


### Disadvantages of cosine similarity method:

- a) Cosine similarity does not consider magnitude of vector. It only checks direction and angle between vectors. Thus, ignoring high frequency words. This way a long document with many words can be similar to a short document with fewer words but similar frequencies.


### advantages of cosine similarity method:

- a) Cosine similarity gives best test similarity irrespective of size of the text like (frequent words appearing many times) and far apart by Euclidean distance. Cosine still can have smaller angle and get better similarity.
- b)

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

One example with similar score		
Image	Reference captions	Model generated caption
 <p>Image id = 405331006_4e94e07698 Matching Score = 0.49</p>	<ol style="list-style-type: none"> <li>1. A man in a blue cowboy hat is riding a white horse.</li> <li>2. A man in blue is riding a horse on a dirt road.</li> <li>3. A man wearing a blue hat and shirt is riding a white horse.</li> <li>4. A person in a blue cowboy hat rides a horse down a dirt trail.</li> <li>5. The person in the blue shirt and blue hat is riding a white horse.</li> </ol>	a man in a blue shirt and blue shorts is riding a bike on a dirt path
One example where score not matching		
	1. A boy is swimming	a young girl is



 <p>Image id = 897621891_efb1e00d1d BLEU 4-gram score = 0.43 Cosine score = 0.41</p>	<p>underwater holding a toy in his hand.</p> <p>2. A little boy swimming underwater with a toy in his hand</p> <p>3. A little boy swims underwater.</p> <p>4. A little boy underwater in a pool , holding a plastic dinosaur.</p> <p>5. Child swimming underwater with a toy in his hand.</p>	<p>swimming under water in a pool</p>
---	---	---------------------------------------

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

*No response needed here.*