

1. A common practice in natural language processing is to lemmatize words before creating tokens. While we did not do this for our vocabulary, take a look at Flickr8k.lemma.token.txt and briefly explain what the advantages/disadvantages might be of using lemmatized vs regular tokens. (max. 5 marks)

Each line in the input is divided into multiple individual substrings known as regular tokens. These tokens can contain identifiers, keywords, operators, delimiters, and literals separated by white spaces, whereas lemmatized tokens are root forms of the words with considering the part of speech and context of the word in the sentence.

One of the major advantages of lemmatization is indexing for **easy information retrieval**. Thus it creates an index for words with the same root, hence increasing the speed of the search engine. For example, the indexes for 'better' and 'good' are the same. Whereas **for regular tokens**, the search engine needs to go through all the corpus to find a match for user requirements. Lemmatizer also **considers the context of a word in sentence** thus converting it appropriately. For example, in the sentence, 'A young boy carry a green bucket on a sandy beach with a cloudy sky overhead .', the word 'carry' won't be converted to its root 'car' as done in stemming.

The lemmatizer uses WordNet to generate a correct base word for input. Thus it **creates an additional computational overhead** to create its own knowledge base. Also, it requires to **state the part of speech of a word**, else it considers it as a noun. As shown in the above example, 'better' is converted to 'good' only if it is described as an adjective. Wordnet may **sometimes don't have root for few words** such as 'iPhone' and 'iPhones' thus lemmatizing it to original form. Thus if we use lemmatized token in our vocabulary, the generated captions may represent words in its root form and won't make much sense even though it would be fast enough as compared to regular tokens.

2. Present the sample images and generated caption for each epoch of training for both the RNN and LSTM version of the decoder, including the BLEU scores. (max. 30 marks)

Reference Images are shown in the below table.

**Please note:** Same two images have been used to generate captions using LSTM and RNN network

Image 1	Image 2
	

BELOW SCREENSHOT CONTAINS **REFERENCE CAPTION, GENERATED CAPTION AND BLEU SCORE** FOR BOTH IMAGES:

**BEFORE TRAINING:**

### RNN

```
Reference Captions before training for Image 1
['a', 'group', 'of', 'people', 'walking', 'on', 'a', 'rocky', 'ridge']
['a', 'man', 'and', 'two', 'boys', 'are', 'climbing', 'on', 'a', 'rock', 'in', 'the', 'park']
['two', 'boys', 'and', 'a', 'man', 'holding', 'a', 'white', 'hat', 'climb', 'on', 'a', 'rock', 'with', 'a', 'lawn', 'and', 'shade', 'trees', 'visible', 'in', 'the', 'background']
['two', 'children', 'and', 'a', 'man', 'on', 'a', 'boulder']
['three', 'people', 'climb', 'on', 'a', 'large', 'rock']
Reference Caption before training for Image 2
['two', 'dogs', 'are', 'running', 'through', 'a', 'green', 'yard']
['the', 'two', 'dogs', 'are', 'running', 'through', 'the', 'grass']
['one', 'brown', 'and', 'white', 'dog', 'chasing', 'a', 'black', 'and', 'white', 'dog', 'through', 'the', 'grass']
['two', 'running', 'through', 'the', 'grass']
['two', 'dogs', 'are', 'running', 'through', 'the', 'grass', 'near', 'a', 'house', 'and', 'trees']
Generated caption for Image 1 before training ['rows', 'shades', 'whose', 'pumpkins', 'helmeted', 'bib', 'vegetation', 'does', 'chains', 'parachute', 'biking', 'clothing', 'dune', 's']
Generated caption for Image 2 before training ['monument', 'biting', 'sooner', 'jumps', 'giant', 'them', 'soldier', 'porch', 'bunch', 'winner', 'rainy', 'canal', 'expression', 'paras']
BLEU score for Image 1 before traing 0
BLEU score for Image 2 before traing 0
```

### LSTM

```
Reference Captions before training for Image 1
['three', 'people', 'climb', 'on', 'a', 'large', 'rock']
['a', 'group', 'of', 'people', 'walking', 'on', 'a', 'rocky', 'ridge']
['two', 'boys', 'and', 'a', 'man', 'holding', 'a', 'white', 'hat', 'climb', 'on', 'a', 'rock', 'with', 'a', 'lawn', 'and', 'shade', 'trees', 'visible', 'in', 'the', 'background']
['a', 'man', 'and', 'two', 'boys', 'are', 'climbing', 'on', 'a', 'rock', 'in', 'the', 'park']
['two', 'children', 'and', 'a', 'man', 'on', 'a', 'boulder']
Reference Caption before training for Image 2
['two', 'dogs', 'are', 'running', 'through', 'a', 'green', 'yard']
['one', 'brown', 'and', 'white', 'dog', 'chasing', 'a', 'black', 'and', 'white', 'dog', 'through', 'the', 'grass']
['two', 'dogs', 'are', 'running', 'through', 'the', 'grass', 'near', 'a', 'house', 'and', 'trees']
['two', 'running', 'through', 'the', 'grass']
['the', 'two', 'dogs', 'are', 'running', 'through', 'the', 'grass']
Generated caption for Image 1 before training ['three', 'children', 'are', 'sitting', 'on', 'a', 'rock', 'overlooking', 'a', 'bridge']
Generated caption for Image 2 before training ['a', 'dog', 'is', 'running', 'through', 'a', 'field', 'of', 'grass']
BLEU score for Image 1 before traing 0.37342112655242105
BLEU score for Image 2 before traing 0.37531192687516973
```

**EPOCH 0:**

### RNN:

```
Epoch: 0 Loss 3.4575994101869703
Caption after epoch 0
Generated caption for Image 1 ['a', 'man', 'in', 'a', 'red', 'shirt', 'and', 'a', 'white', 'shirt', 'and', 'a', 'white', 'dog']
Generated caption for Image 2 ['a', 'man', 'and', 'a', 'dog', 'are', 'playing', 'in', 'a', 'field']
BLEU score for Image 1 0.6206161803077822
BLEU score for Image 2 0.7952707287670506
```

## LSTM:

```
Epoch: 0 Loss 2.1772099008591863
Caption after epoch 0
Generated caption for Image 1 ['a', 'group', 'of', 'people', 'are', 'climbing', 'a', 'rock', 'wall']
Generated caption for Image 2 ['a', 'dog', 'is', 'running', 'on', 'a', 'leash']
BLEU score for Image 1 0.4032989116748133
BLEU score for Image 2 0.7013967267997694
```

## EPOCH 1:

### RNN:

```
Epoch: 1 Loss 2.707636668436551
Caption after epoch 1
Generated caption for Image 1 ['three', 'people', 'are', 'sitting', 'on', 'a', 'rock', 'overlooking', 'a']
Generated caption for Image 2 ['a', 'dog', 'is', 'standing', 'in', 'the', 'snow']
BLEU score for Image 1 0.4518010018049224
BLEU score for Image 2 0.7013967267997694
```

## LSTM:

```
Epoch: 1 Loss 2.060229009568097
Caption after epoch 1
Generated caption for Image 1 ['a', 'group', 'of', 'people', 'are', 'sitting', 'on', 'a', 'rock', 'ledge', 'overlooking', 'a', 'tree']
Generated caption for Image 2 ['three', 'dogs', 'are', 'running', 'on', 'a', 'grassy', 'field']
BLEU score for Image 1 0.2978201796359045
BLEU score for Image 2 0.392814650900513
```

## EPOCH 2:

### RNN:

```
Epoch: 2 Loss 2.4846688552552285
Caption after epoch 2
Generated caption for Image 1 ['two', 'people', 'are', 'standing', 'on', 'a', 'rock', 'overlooking', 'a', 'lake']
Generated caption for Image 2 ['a', 'dog', 'is', 'running', 'through', 'the', 'snow']
BLEU score for Image 1 0.37342112655242105
BLEU score for Image 2 0.40495158902656925
```

## LSTM:

```
Epoch: 2 Loss 1.950141115442067
Caption after epoch 2
Generated caption for Image 1 ['a', 'group', 'of', 'children', 'climbing', 'on', 'a', 'rock', 'wall']
Generated caption for Image 2 ['three', 'dogs', 'are', 'playing', 'with', 'a', 'blue', 'ball', 'in', 'a', 'grassy', 'field']
BLEU score for Image 1 0.4463236137853328
BLEU score for Image 2 0.3882726775222324
```

## EPOCH 3:

### RNN:

```
Epoch: 3 Loss 2.3378434664387244
Caption after epoch 3
Generated caption for Image 1 ['a', 'group', 'of', 'people', 'are', 'standing', 'on', 'a', 'beach']
Generated caption for Image 2 ['a', 'black', 'and', 'white', 'dog', 'with', 'a', 'red', 'collar', 'in', 'a', 'field']
BLEU score for Image 1 0.36889397323344053
BLEU score for Image 2 0.31702331385234306
```

## LSTM:

```
Epoch: 3 Loss 1.8488462795054794
Caption after epoch 3
Generated caption for Image 1 ['three', 'boys', 'are', 'climbing', 'a', 'rock', 'wall']
Generated caption for Image 2 ['a', 'dog', 'is', 'running', 'on', 'the', 'grass']
BLEU score for Image 1 0.5410822690539396
BLEU score for Image 2 0.5091996654452335
```

## EPOCH 4:

### RNN:

```
Epoch: 4 Loss 2.2235807271494816
Caption after epoch 4
Generated caption for Image 1 ['a', 'man', 'and', 'woman', 'sit', 'on', 'a', 'bench']
Generated caption for Image 2 ['two', 'dogs', 'play', 'with', 'a', 'ball', 'in', 'the', 'grass']
BLEU score for Image 1 0.45966135761245924
BLEU score for Image 2 0.6104735835807844
```

### LSTM:

```
Epoch: 4 Loss 1.75479755922742
Caption after epoch 4
Generated caption for Image 1 ['three', 'young', 'men', 'are', 'standing', 'on', 'a', 'rock', 'in', 'front', 'of', 'a', 'large', 'rock', 'formation']
Generated caption for Image 2 ['a', 'dog', 'is', 'running', 'in', 'the', 'grass']
BLEU score for Image 1 0.2533654946448647
BLEU score for Image 2 0.5091996654452335
```

### 3. Compare training using an RNN vs LSTM for the decoder network (loss, BLEU scores over test set, quality of generated captions, performance on long captions vs. short captions, etc.). (max. 30 marks)

#### Performance on long captions vs. short:

The performance of LSTM is better compared to RNN when it comes to long captions. When it comes to the architecture of each network, RNN lacks memory using which LSTM can look back many time steps thus learning Long Term Dependencies. As seen in the above example, the last epoch for First Image, generated caption by RNN is far less accurate compared to LSTM. Here, since the distance between the two contexts is more, the vanishing gradient causes the network to learn less.

Whereas when it comes to short captions, RNN learns more information between the two contexts since there isn't much gap between them. As seen in the last epoch of Second Image, RNN predicts more accurately compared to LSTM. Thus LSTM is better when captions are large.

#### Quality of generated captions:

The quality of captions for both the network is commendable. Both the captions are good grammatically but LSTM has a slight edge over RNN since it learns and stores output from previous layers. BLEU-1 scores of the same image for both the network are excellent but there is a considerable difference between the BLEU-4 score for RNN and LSTM. LSTM can produce much better quality when it comes to sentence formation as seen in the below example:

RNN	LSTM
BLEU Score 1 0.6333861926251716	BLEU Score 1 0.5625
BLEU Score 2 0.5046941987886602	BLEU Score 2 0.33541019662496846
BLEU Score 3 0.44531055832596395	BLEU Score 3 0.48627319717182316
BLEU Score 4 0.38662527162788296	BLEU Score 4 0.5791460926441345
<b>Reference Caption</b>	<b>Reference Caption</b>

[[ 'the', 'small', 'boy', 'is', 'walking', 'on', 'a', 'street', 'lined', 'with', 'people'], [ 'a', 'little', 'boy', 'in', 'a', 'blue', 'shirt', 'looking', 'at', 'the', 'camera', 'concrete', 'in', 'the', 'background'], [ 'a', 'small', 'boy', 'looking', 'at', 'the', 'camera'], [ 'a', 'little', 'boy', 'walks', 'around', 'with', 'a', 'white', 'and', 'black', 'shorts', 'outside'], [ 'a', 'boy', 'in', 'a', 'blue', 'shirt']]]	[[ 'a', 'little', 'boy', 'walks', 'around', 'with', 'a', 'white', 'and', 'black', 'shorts', 'outside'], [ 'a', 'small', 'boy', 'looking', 'at', 'the', 'camera'], [ 'a', 'boy', 'in', 'a', 'blue', 'shirt'], [ 'a', 'little', 'boy', 'in', 'a', 'blue', 'shirt', 'looking', 'at', 'the', 'camera', 'concrete', 'in', 'the', 'background'], [ 'the', 'small', 'boy', 'is', 'walking', 'on', 'a', 'street', 'lined', 'with', 'people']]]
<b>Generated Caption</b> [ 'a', 'little', 'boy', 'in', 'a', 'red', 'shirt', 'is', 'holding', 'a']	<b>Generated Caption</b> [ 'a', 'little', 'girl', 'in', 'a', 'pink', 'shirt', 'and', 'blue', 'jeans', 'is', 'sitting', 'on', 'a', 'wooden', 'bench']

### Objects in Images

Another factor affecting the performance of both networks in the number of objects in the picture. Thus the model finds it difficult to generate a sequence of words that relate to each other. In the case of LSTM, it generates a few words which are not present in the image. However, RNN seems to produce words that are appropriate to the image while missing some of the.

### Loss, BLEU Score

	RNN		LSTM	
Training Loss	Epoch	Loss	Epoch	Loss
	1	3.45	1	2.17
	2	2.70	2	2.06
	3	2.48	3	1.95
	4	2.33	4	1.84
	5	2.22	5	1.75
Bleu Score	Epoch	Avg. Bleu Score Over Test Set	Epoch	Avg. Bleu Score Over Test Set
	1	0.76	1	0.72
	2	0.73	2	0.77
	3	0.74	3	0.76
	4	0.7426	4	0.76
	5	0.7528	5	0.729

**4. Among the text annotations files downloaded with the Flickr8k dataset are two files we did not use: ExpertAnnotations.txt and CrowdFlowerAnnotations.txt. Read the readme.txt to understand their contents, then consider and discuss how these might be incorporated into evaluating your models. (max. 10 marks)**

Currently, BLEU is used to automate the process of caption evaluation however it can be sometimes misleading, A reason for these might be the quality of data captions present in the training set. Below two files can be used to increase the quality and quantity of the training dataset considering the factors mentioned below.







### ExpertAnnotations.txt:

As described in read me file, the ExpertAnnotations file contains 5 columns where the first column contains Image IDs of the picture given to the expert while the second column consists of random captions IDs, using which we can derive its correct Image IDs. Thus Caption id 1287475186\_2dee85f1a5.jpg#2 is basically the second caption for Image with ID 1287475186\_2dee85f1a5. The last three columns contain values ranging from 1-4 annotated by experts where 1 depicts incorrect information and 4 describes the correct description.

These annotations can be used to evaluate our model. If an expert gives values less than 3 for a caption, it means it either describes only a part of an image or it doesn't describe the image at all. These captions would be equivalent to generated captions before any training, which would definitely reduce the BLEU score. Hence we won't consider these captions to evaluate our model. On the other hand, if a caption is having a minimum of 3 points from at least 2 experts we can consider it to evaluate our model.

Also, if the caption ID and Image ID are the same, the values tend toward the higher side. However, sometimes if an Image ID and Caption ID are different but the caption describes the image appropriately to some extent as shown in Fig.1, then the values by some experts would still be good. Thus these captions can be used as a reference caption against the generated caption to calculate BLEU score for the image ID in column 1 also it would increase the knowledge base for reference captions.

Image ID and Image	Original Image for caption, Caption Id and Caption	Expert 1	Expert 2	Expert 3
1119015538_e8e796281e.jpg 	416106657_cab2a107a5.jpg2  A white dog runs in the grass	4	4	4
115684808_cb01227802.jpg 	3185409663_95f6b958d8.jpg#2  A little girl runs on the wet sand near the ocean	3	3	4





**Fig.1.**



### CrowdFlowerAnnotations.txt.

The CrowdFlowerAnnotations file contains 5 columns where the first and second columns representing Image IDs and random Caption IDs with Image ID of correct picture respectively. As seen, these values in these two columns can be different at times. The third column portrays the percentage of Yesses out of the total judgments. The fourth and fifth columns show a total number of Yeses and Noes respectively.

From the given file we can see the judges have been asked to provide their input in binary format either 'Yes' or 'No' unlike in **ExpertAnnotations.txt** where they could provide in range 1-4. This format would result in inconsistency since it would depend on an individual's perception. Even if the majority of authorities are in favor of the caption, it still won't be dependable as shown in the 1st row of Fig.2, where the image caption doesn't depict exactly the given image but still the percentage of 'yes' is more than 50%. Similarly, as shown in the second row of Fig.2, even though the Image ID and Caption IDs are the same, there is a difference in opinion among the judges.

However, there can be seen a few captions where the correct description is provided for unknown images as shown in the third row in Fig.2. Hence we can use these captions where the percentage of 'Yeses' is 100% thus ensuring data integrity and increasing BLEU score. These captions can be used to increase the dataset for the given Image ID.

Image ID	Original Image for Caption, Caption Id, Caption	Percentage of Yes
2283966256_70317e1759.jpg 	3642220260_3aa8a52670.jpg#2  A black dog walks on the beach near the rocks	0.66666666 666667
1096395242_fc69f0ae5a.jpg 	1096395242_fc69f0ae5a.jpg#2 	0.66666666 666667

	A young boy with his foot outstretched aims a toy at the camera in front of a fireplace.	
1119015538_e8e796281e.jpg 	494792770_2c5f767ac0.jpg#2  a dog runs through the long grass.	100

**Fig.2.**