RNN Image Captioning

Oscar Brooks

November 6, 2020

Introduction:

In this worksheet we will be using PyTorch to perform image captioning using Recurrent Neural Networks(RNN's). We begin by creating a tokenized vocabulary from captioned images given in the Flickr8k dataset build and compare an RNN and LSTM (long short-term memory) neural network for generating captions.

Question 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)

"Natural Language Processing or NLP is a field of Artificial Intelligence that gives the machines the ability to read, understand and derive meaning from human languages" [1].

As a part of NLP, tokenization and lemmatization are examples of tools which can be used to process language into a more machine friendly form. Different languages contain various rates of inflection within their vocabulary; that is to say that among the total words in the vocabulary, many words may be comprised of the same base word/meaning that is to say they they are derivationally related.

For instance in English. Both "dogs" and "dog" would have the same lemma "dog".

Lemmatization is thus the process of determining the lemma of a word which aims to reduce the inflection of. Below is an example of both the lemmatized and regular caption tokens given in the Flickr8k file.

Image	Regular token	Lemmatized token
	1. A guy snowboarding down a hill . 2. A skier in a ski suit is skiing downhill , near a fence . 3. A skier races down the mountain . 4. A slalom skier moving quickly downhill . 5. Competing skier going down the course leaning to make a sharp turn .	1. A guy snowboard down a hill . 2. A skier in a ski suit be ski downhill , near a fence . 3. A skier race down a mountain . 4. A slalom skier move quick downhill . 5. Compete skier go down a course lean to make a sharp turn .

In order to caption images using our RNN, a vocabulary was required. We use the captions provided in the Flickr8k.token.txt file to produce 3389 unique words which appear more than three times within the file. As we would expect, when this function was run on the Flickr8k.lemma.token.txt file we found this vocabulary to be far smaller with only 2618 unique words fitting this criteria.

As the lemmatized vocabulary is smaller, more concise and clean this could allow for easier computation when being used for training our model. This would also possibly allow for greater bleu scores to be produced as with less words the model is more likely to put suitable words closer together. Although much as this may be easier for the computer to be ”accurate” in its description it deviates from attempting to produce captions similar to (or perhaps better than) that of a real person, a milestone/goal rooted in much of modern AI.

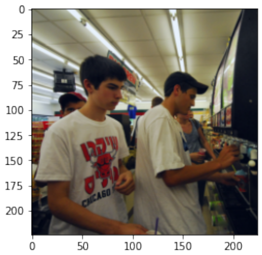
When the lemmatized caption is generated it will be limited to using only lemmas to describe the image. Not only would this make less sense to a person it would be far less descriptive. Another drawback to using a lemmatized vocabulary is that it is very labour intensive. This is because unlike stemming (chopping suffixes off words), lemmatization is very complex and has to usually be completed by hand to ensure that the meanings of words are not skewed or lost. [2]

Question 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)

To produce the generated captions we needed to have our RNN working. During the general cleaning and tokenization of the vocabulary it seemed there were five additional references without a corresponding image in the Flickr8k.token.txt referencing a ”2258277193_586949ec62.jpg” which was not in the dataset provided. Because of this these references were removed from the dataframe before training the model.

Below are examples of images and their corresponding references which our RNN used for training. Below each image and its five references the ”Generated Captions” for each epoch have been displayed with their 4-gram BLEU (Bilingual Evaluation Understudy) scores. When calculating the BLEU score, smoothing method 4 has been used from the nltk package with hopes to combat the inflation in precision for small sentences. [3]

Image 1	Referenced captions
 0 25 50 75 100 125 150 175 200 0 50 100 150 200	1. A group of people are getting fountain drinks at a convenience store . 2. People get their slushies . 3. Several adults are filling their cups and a drink machine . 4. Two boys in front of a soda machine . 5. Two guys getting a drink at a store counter .

Epoch	Generated Caption	BLEU Score
1	a man in a blue shirt and a woman in a red shirt and a woman in a white	0.2438
2	a man and a woman are standing on a sidewalk	0.4297
3	a man is sitting on a bench	0.2365
4	a man in a black shirt and a woman in a black shirt and a black jacket is standing	0.2365
5	a man in a black shirt and a woman in a black dress	0.3060

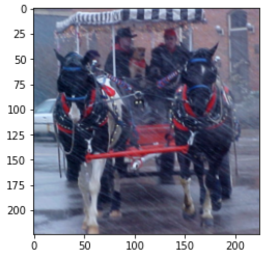
Table 1: RNN Image 1

By comparison of the captions generated we can see that in both cases for images and decoders, that the final of training epoch doesn't necessarily produce the best result in terms of BLEU score. For the image; our largest BLEU score was achieved in the fourth LSTM epoch (‘a man in a blue shirt and a hat is standing in front of a crowd of people’, 0.52) whereas for our second image the RNN produced the highest score (‘a group of people are riding horses on a track’ , 0.55).

Here it is worth noting that both models show greater signs of improvement throughout the training when dealing with image two rather than image one. With image one the network seems to produce sentences about the men

Epoch	Generated Caption	BLEU Score
1	a man in a black shirt and a black shirt and a black and white shirt and a black	0.2203
2	a man in a white shirt and a black shirt and a white shirt and a man in a	0.2382
3	a man in a blue shirt is holding a <unk>	0.4663
4	a man in a blue shirt and a hat is standing in front of a crowd of people	0.5249
5	a man in a black shirt and a woman in a white shirt and a man in a black	0.2415

Table 2: LSTM Image 1

Image 2	Referenced captions
 0 25 50 75 100 125 150 175 200 0 50 100 150 200	1. A cart containig two men being pulled by horses in the rain . 2. A horse driven carriage running through a rainstorm . 3. Black and white horses carry a cart with people through the rain . 4. People enjoy a horse draw open carriage in the rain . 5. The horses pull the carriage , holding people and a dog , through the rain .

Epoch	Generated Caption	BLEU Score
1	a group of people are playing in a grassy field	0.4098
2	a man in a blue shirt and a woman are playing in the snow	0.3950
3	a man is standing on a bench in front of a large crowd	0.2677
4	a group of people are riding horses on a track	0.5523
5	a man is riding a horse on a dirt path	0.3955

Table 3: RNN Image 2

Epoch	Generated Caption	BLEU Score
1	a man is doing a trick on a rock	0.1568
2	a man and a woman are standing on a bench	0.3146
3	a man is riding a horse on a dirt track	0.4152
4	a man is riding a horse with a man in a black jacket	0.5332
5	a man is riding a horse on a track	0.3541

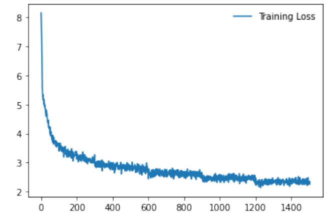
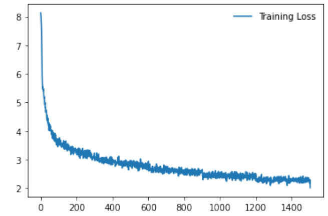
Table 4: LSTM Image 2

in the photo but cannot generate much context with its language however with image two the network starts with noticing the people but later begins to recognise the horses too (across both models). For these examples both models seem to produce sentences of similar lengths.

Question 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)

Illustrated above we showed two images and their resulting RNN and LSTM 4-gram BLEU scores. Now we will compare how the network performs on the test set as a whole, for this assignment we did not check for over training using a validation set and instead have used the whole dataset for either training or testing. This is because the network is being run for five epochs only and will not be subject to much over-training if any at all with the given learning rate (0.001).

RNN Loss	LSTM Loss
	

As for the training loss we can see for both models this decreases as expected from the random weights at the beginning of training to the trained weights at the end. For both models these graphs look quite similar except the LSTM loss curve ends the fifth epoch on a slightly lower value than the RNN at 2.2456 as opposed to 2.3489.

On the coming pages we have two tables to compare the training of each model. The first column is a histogram of the lengths of sentences generated, the second column gives the average BLEU scores for the whole and loss for each epoch and the final column investigates the 4-gram BLEU score for long, short and medium sentences. Long sentences are defined as exceeding a length of 12 and short sentences are defined to have less than a length of 8.

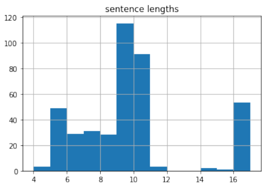
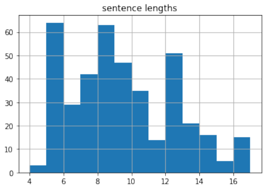
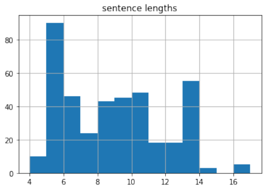
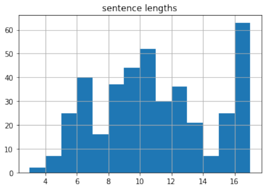
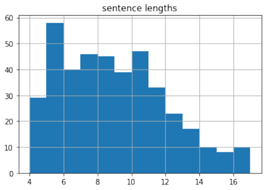
Before looking at the individual tables we can notice that BLEU scores decrease in every case for higher ”gram” methods, where the size of the ”n” used in the n-gram method determines how far a candidate word may be in position to a reference word. For the results given we use a cumulative n-gram score to stay consistent with the cumulative 4-gram BLEU score default. A cumulative n-gram score uses an equal weighting of its own along with each antecedent individual-gram score.

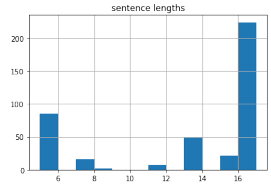
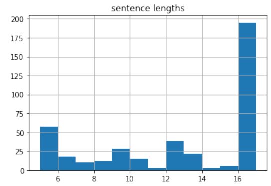
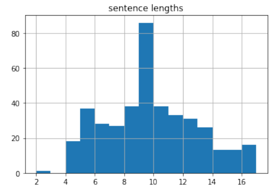
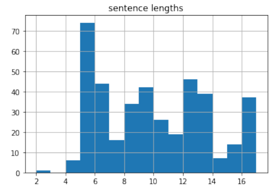
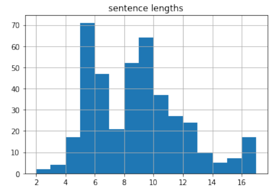
Starting with the RNN we can see that before training most generated sentences are of medium or small length and the vast majority of long sentences are of length 16 with no sentences constructed of length 12 or 13. As training progresses we can see that sentences of all lengths (within the range of those produced) were generated, to form a more normal distribution of sentence length.

When looking at the LSTM we can see near the start of training the vast majority of sentences are of length 16 or 5 without many medium sentences being constructed. Observing the n-gram BLEU scores it is intuitive to think of the 1-Gram scores as only discriminating the candidate against neighbouring words.

Both networks end up the final epoch with a similar average sentence length and distribution of sentence lengths. Surprisingly the LSTM created sentences with the smallest average length and the shortest sentences (with only two words) wheras the RNN produced a shortest sentence of 4 words in its final epoch. This is suprising as I expected the LSTM to create longer sentences as its feedback loop is designed to fix/improve the vanishing gradient problem. The highest scoring 1-Gram score came from RNN in its third epoch however the best 2, 3, 4 and 5-Gram scores were all produced by the LSTM in the final epoch.

Caption grading notoriously difficult and thus the quality of BLEU scoring is sometimes quite difficult to interpret and can lead to some questionable conclusions. Earlier in Question 2 for example, where the RNN had its BLEU scores calculated for Image 2. First epoch produced ”a group of people are playing on a grassy field” and was scored 0.4098; Fifth epoch, ”a man is riding on a horse on a dirt path” was scored 0.3955. Although this is not a large reduction in BLEU score I believe the fifth epoch caption to much better describe the photo than the first.

RNN sentence lengths	BLEU scores & loss on the whole set	Specific sentence lengths
	- 1-Gram BLEU score: 0.8686 - 2-Gram BLEU score: 0.7368 - 3-Gram BLEU score: 0.6061 - 4-Gram BLEU score: 0.4987 - 5-Gram BLEU score: 0.4170 - Final epoch loss: 3.3056	- Short sentence scores: 0.5561 - Medium sentence scores : 0.5030 - Long sentence scores: 0.3655 - Average sentence length: 9.36
	- 1-Gram BLEU score: 0.8664 - 2-Gram BLEU score: 0.7471 - 3-Gram BLEU score: 0.6262 - 4-Gram BLEU score: 0.5265 - 5-Gram BLEU score: 0.4468 - Final epoch loss: 2.9715	- Short sentence scores: 0.5842 - Medium sentence scores : 0.5051 - Long sentence scores: 0.4655 - Average sentence length: 9.05
	- 1-Gram BLEU score: 0.8761 - 2-Gram BLEU score: 0.7616 - 3-Gram BLEU score: 0.6403 - 4-Gram BLEU score: 0.5413 - 5-Gram BLEU score: 0.4638 - Final epoch loss: 3.1017	- Short sentence scores: 0.5591 - Medium sentence scores : 0.5439 - Long sentence scores: 0.4863 - Average sentence length: 8.43
	- 1-Gram BLEU score: 0.8554 - 2-Gram BLEU score: 0.7262 - 3-Gram BLEU score: 0.6075 - 4-Gram BLEU score: 0.5128 - 5-Gram BLEU score: 0.4398 - Final epoch loss: 2.5118	- Short sentence scores: 0.5739 - Medium sentence scores : 0.5383 - Long sentence scores: 0.4208 - Average sentence length: 10.61
	- 1-Gram BLEU score: 0.8590 - 2-Gram BLEU score: 0.7490 - 3-Gram BLEU score: 0.6364 - 4-Gram BLEU score: 0.5433 - 5-Gram BLEU score: 0.4690 - Final epoch loss: 2.3489	- Short sentence scores: 0.5427 - Medium sentence scores : 0.5504 - Long sentence scores: 0.5161 - Average sentence length: 8.48

LSTM sentence lengths	BLEU scores & loss on the whole set	Specific sentence lengths
	• 1-Gram BLEU score: 0.7997 • 2-Gram BLEU score: 0.6260 • 3-Gram BLEU score: 0.4893 • 4-Gram BLEU score: 0.3870 • 5-Gram BLEU score: 0.3151 • Final epoch loss: 2.1850	• Short sentence scores: 0.5280 • Medium sentence scores : 0.4829 • Long sentence scores: 0.3358 • Average sentence length: 13.34
	• 1-Gram BLEU score: 0.8267 • 2-Gram BLEU score: 0.6682 • 3-Gram BLEU score: 0.5416 • 4-Gram BLEU score: 0.4456 • 5-Gram BLEU score: 0.3760 • Final epoch loss: 2.3794	• Short sentence scores: 0.6080 • Medium sentence scores : 0.5173 • Long sentence scores: 0.3532 • Average sentence length: 12.71
	• 1-Gram BLEU score: 0.8755 • 2-Gram BLEU score: 0.7633 • 3-Gram BLEU score: 0.6529 • 4-Gram BLEU score: 0.5630 • 5-Gram BLEU score: 0.4899 • Final epoch loss: 2.6957	• Short sentence scores: 0.5751 • Medium sentence scores : 0.5705 • Long sentence scores: 0.5181 • Average sentence length: 9.34
	• 1-Gram BLEU score: 0.8713 • 2-Gram BLEU score: 0.7582 • 3-Gram BLEU score: 0.6491 • 4-Gram BLEU score: 0.5589 • 5-Gram BLEU score: 0.4861 • Final epoch loss: 2.8011	• Short sentence scores: 0.6067 • Medium sentence scores : 0.5685 • Long sentence scores: 0.4728 • Average sentence length: 9.58
	• 1-Gram BLEU score: 0.8672 • 2-Gram BLEU score: 0.7659 • 3-Gram BLEU score: 0.6634 • 4-Gram BLEU score: 0.5756 • 5-Gram BLEU score: 0.5041 • Final epoch loss: 2.2456	• Short sentence scores: 0.5643 • Medium sentence scores : 0.6044 • Long sentence scores: 0.4713 • Average sentence length: 8.40

Question 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)

The ExpertAnnotations.txt and CrowdFlowerAnnotations.txt are both records of human judgement on whether a caption fits a given image. The first two columns of both files are comprised an image id the caption id which is being used to describe the image. The next three columns of ExpertAnnotations.txt are expert judged scores ranging from 1 to 4 with 1 being the worst (the caption doesnt fit the image) and 4 being the best (the caption properly describes the image). For the CrowdFlowerAnnotations.txt, the third column is the ”yes-votes”percentage a pairing has been given by the crowd annotators. The numbers of ”yes”:”no” votes given as a ratio in the final two columns with a ”yes” permitted to be scored with ”minor mistakes”, each image-caption pair has at least three votes. [4]

Human judgements such as these could be possibly incorporated into the model training as captions can evidently be used to describe more than one image and thus extra captions being linked/verified to describe an image could contribute to an increasing in the training set.

As we have seen in already in an example at the end of question 3, BLEU scores aren’t infallible and can often fail to correlate with human judgements. This motivated Yin Cui.et.al to devise a new metric co-inspired by the COCO challenge in 2015 (captioning challenge) which could be used in place of BLEU which could score captions with a more realistic/human judgement. It has been proposed that by using a convolutional neural net as a discriminator this could fulfil the role of the loss function to train the RNN/LSTM. In a similar notion to the original Turing Test, the discriminator would be trained to discriminate between the human-certified captions(from ExpertAnnotations.txt or CrowdFlowerAnnotations.txt.) and the RNN-generated captions for a given image. [5] The idea here is that with a well trained loss function, the caption generator would need to improve to fool the learned loss function once again that it was human (or as human as it can be). This cycle would not only produce a good caption generator but also a loss function trained to act like a human critique. Yin Cui.et.al also discuss how their approach can be flexibly robust to ”pathological sentence structure” as they propose identifying pathological transforms to use as negative training examples (encouraging the network to not create them).

References:

[1] Rabby, S., 2019. What Is NLP & What Do NLP Scientists Do?. [online] Towards Data Science. Available at:<https://towardsdatascience.com/whatnlpscientistsdo-905aa987c5c0 > [Accessed 5 May 2020].

[2] Heidenreich, H., 2018. Stemming? Lemmatization? What?. [online] Medium. Available at: <https://towardsdatascience.com lemmatization-what-ba782b7c0bd8 > [Accessed 5 May 2020].

[3] Kite.com. 2020. Code Faster With Line Of Code Completions, Cloudless Processing. [online] Available at: $< https : //kite.com/python/docs/nltk.bleu_score.SmoothingFunction.method4 >$ [Accessed 5 May 2020].

[4] Hodosh, M., 2013. Framing Image Description As A Ranking Task: Data, Models And Evaluation Metrics. [online] Available at: <https://www.aclweb.org/anthology/C18-1179.pdf> [Accessed 5 May 2020].

[5] Cui, Y., 2018. Learning To Evaluate Image Captions. [online] Arxiv.org. Available at: <https://arxiv.org/pdf/1806.06422.pdf [Accessed 5 May 2020].