Image Captioning

Indian Institute of Technology, Kanpur

Course Project – CS698O

Aaditya Aanand – 14002
Pulkit Sharma – 14504

Abstract

The aim of this project is to generate descriptive captions for images. Deep learning has been used via a Convolutional Neural Network coupled with an LSTM based architecture. An image is passed as an input to the CNN, which yields certain annotation vectors. Based on a human vision inspired notion of attention, a context vector is obtained as a function of these annotation vectors, which is then passed as an input to the LSTM. The LSTM network training is separately performed with randomly initialised word and phrase embedding respectively. The output, thereby, obtained is passed to a softmax layer to generate captions. Our results have been evaluated using a metric named METEOR, and word embedding based implementation turns out to be better than its phrase embedding based counterpart.

Dataset

For training and validation purposes, we used Flickr8k dataset which contains 8000 images obtained from Flickr website. Corresponding to each image, five descriptive captions are available for training. Thus, in total, we trained our system on forty thousand captions.

Similarity Score

We used a metric named METEOR for evaluating the performance of our model. METEOR scores a caption generated by our algorithm by aligning it with a reference caption. F – Score multiplied with penalty is termed as METEOR. This alignment is achieved with three considerations: exact, stem and paraphrase matches.

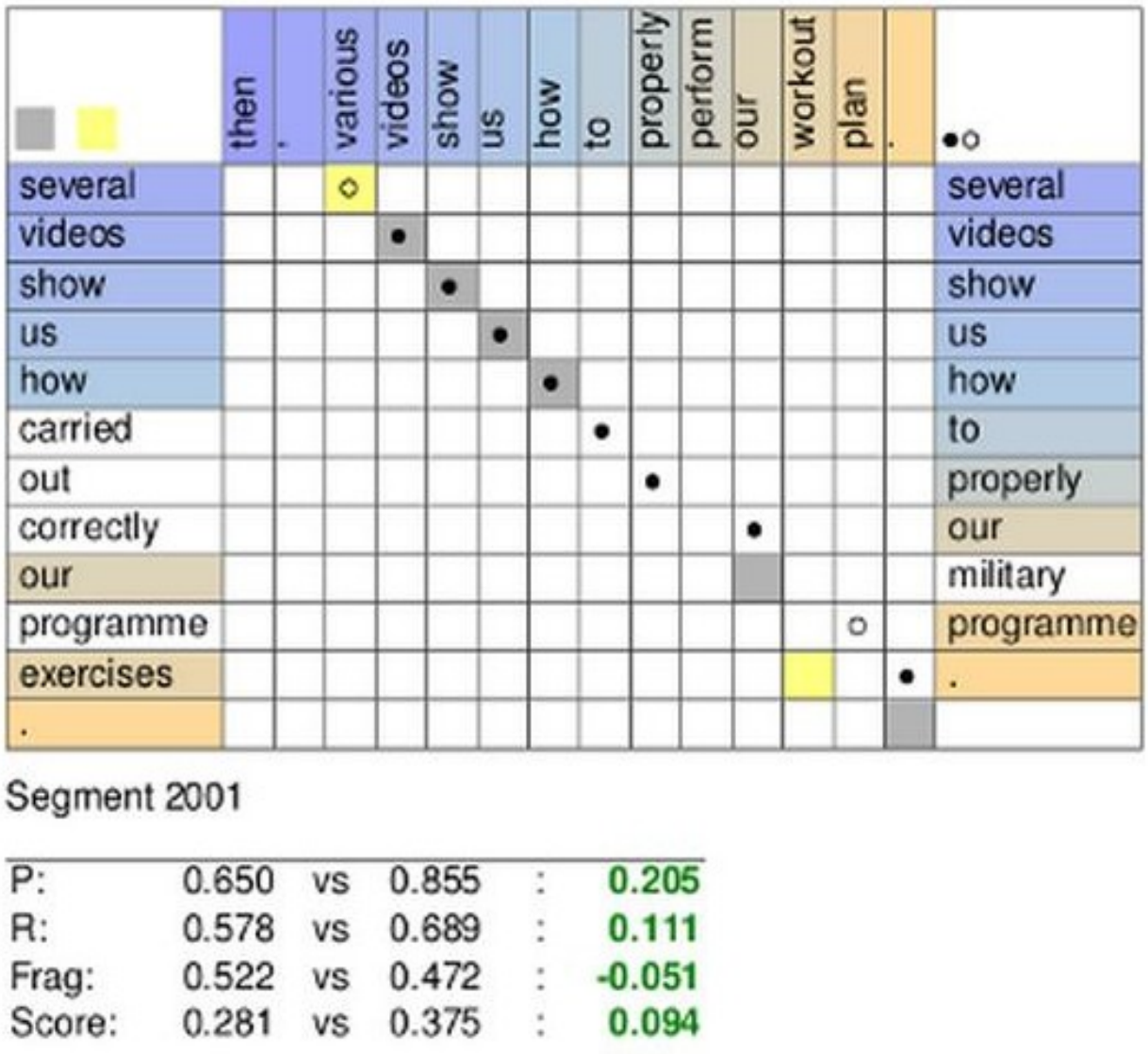

Fig: Example of METEOR applied for sentence scoring Source: Xu et al., “Show, Attend and tell”

Methodology

We have approached the problem of caption generation in two ways: using word embedding and phrase embedding. The implementation using word embedding has been directly derived from the code released by Ryan Kiros. On the other hand, phrase embedding has been implemented by us.

Incorporating Phrase Embedding

Since, the implementation using word embedding was directly available, the first task in front of us was to incorporate phrase embedding in our algorithm. For this purpose, we have used SENNA software to obtain phrases from the captions available to us, as a part of training data. For generating our captions, we are only using Noun, Verb and Prepositional Phrases, and neglecting others.

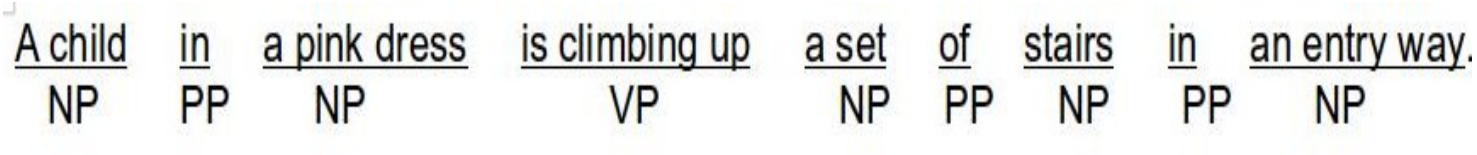

Fig: Illustration of Chunking implemented using Senna Software

SENNA software performs chunking of a reference caption and yields the requisite phrases as output. The embedding of these phrases are obtained by taking the sum of the embedding of words within the phrase. It can be noted from the figure, that the chunking output did not consider the adjectival phrase 'a pink'. It was however, simply embedded in a Noun phrase instead.

Training Procedure

- For annotation vectors, we used a pre trained model of CNN namely Oxford VGGnet trained on Imagenet dataset.

- We are using an LSTM network with 1 hidden layer having 1000 cells. We used mini batches containing captions of same length while training. RMSprop with an adaptive learning rate has been used for updating weights.

- Due to lack of time, we stopped the training after 232 epochs. To compare the results, we trained the model with words as input and halted it after 221 epochs to get a reference for comparing the results.

- Due to large vocabulary size of phrases, we computed the most frequently occurring ones and replaced the rest of the word with a special word (UKN). This reduced our vocabulary to 10000 phrases. The reasoning behind deleting less frequent phrases, is that those phrases are often the modified forms of other phrases. So we made an assumption that the architecture would learn to replace these UKN words with reasonable phrases.

  For ex: Frequent Phrase – A shark
  Less Frequent Phrase – A white shark with big teeth

- We used pre-trained word2vec model trained by google news corpus having 3b words. The vector embedding of phrase is derived by summing up the vector embedding for each word in the phrase.

Beam search for choosing captions

For choosing captions in addition to the conventional way, we implemented Beam search algorithm at the output of LSTM network. Each candidate caption has a certain probability of occurrence. Instead of choosing just the one with the highest probability, Beam search keeps a list of K best candidate choices at the output.

Results

After training the model for both the word and phrase embedding cases, we compared the captions generated by both with reference captions that we had in our dataset. We used METEOR evaluation metric to score the quality of captions generated. Table illustrates the METEOR score we obtained with our implementations.

Input	Vocabulary	METEOR
Phrases	10000	0.0595
Phrases	36220	0.041
Words	9630	0.067
Words using Beam Search	9630	0.089

Table: METEOR score obtained on evaluation using word and phrase embedding

For phrase embedding, we used two different vocabulary sizes: 10,000 and 36,220. Here, captions generated using the smaller vocabulary size were found to be better, having a METEOR score of 0.0595 in comparison with a score of 0.041. Following figure shows the value of the error function on the test dataset, plotted against the number of epochs for word, phrases with 10000 vocabulary and phrases with 36,220 vocabularies as inputs respectively. For the implementation using word embedding, the error starts from 64 and begins to fall until it reaches 45. For the implementation based on phrase embedding, the error function does not fall by any reasonable extent when the vocabulary used has a size of 36,220. It stays almost equal to 48. On the other hand, with a phrase vocabulary of 10,000, the error falls from 41 to approximately 37.

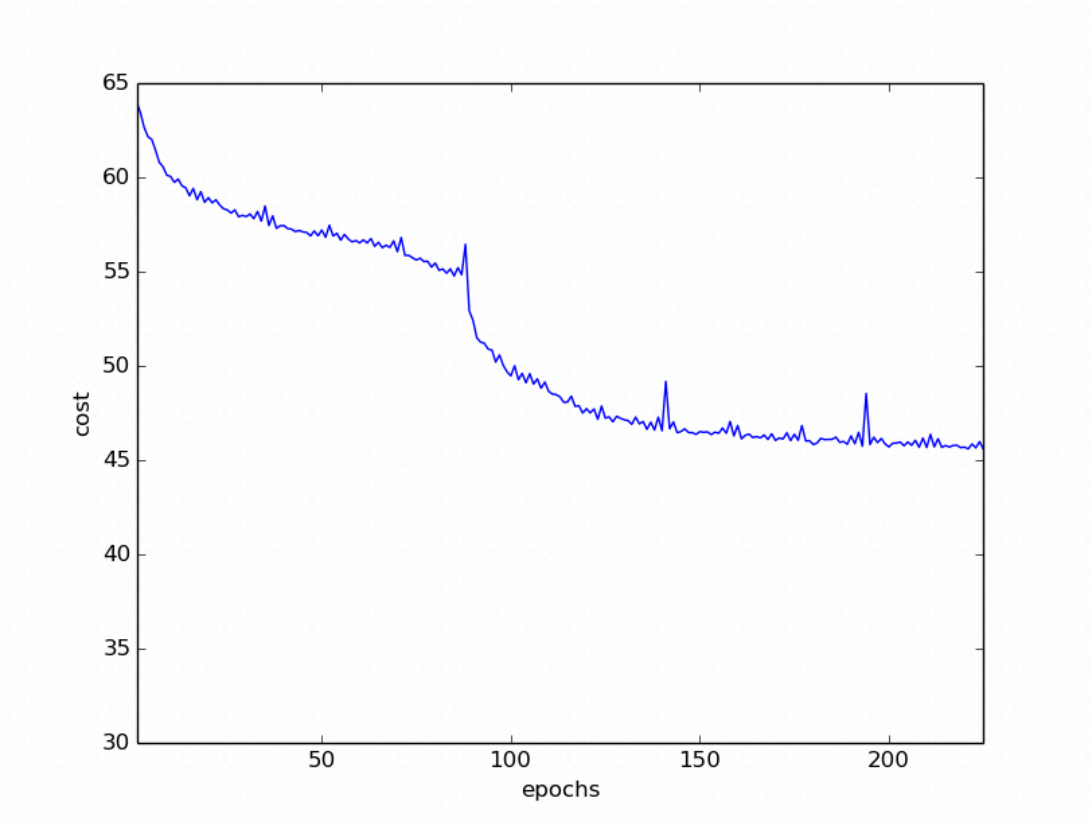

Fig: Error change with epochs for word inputs

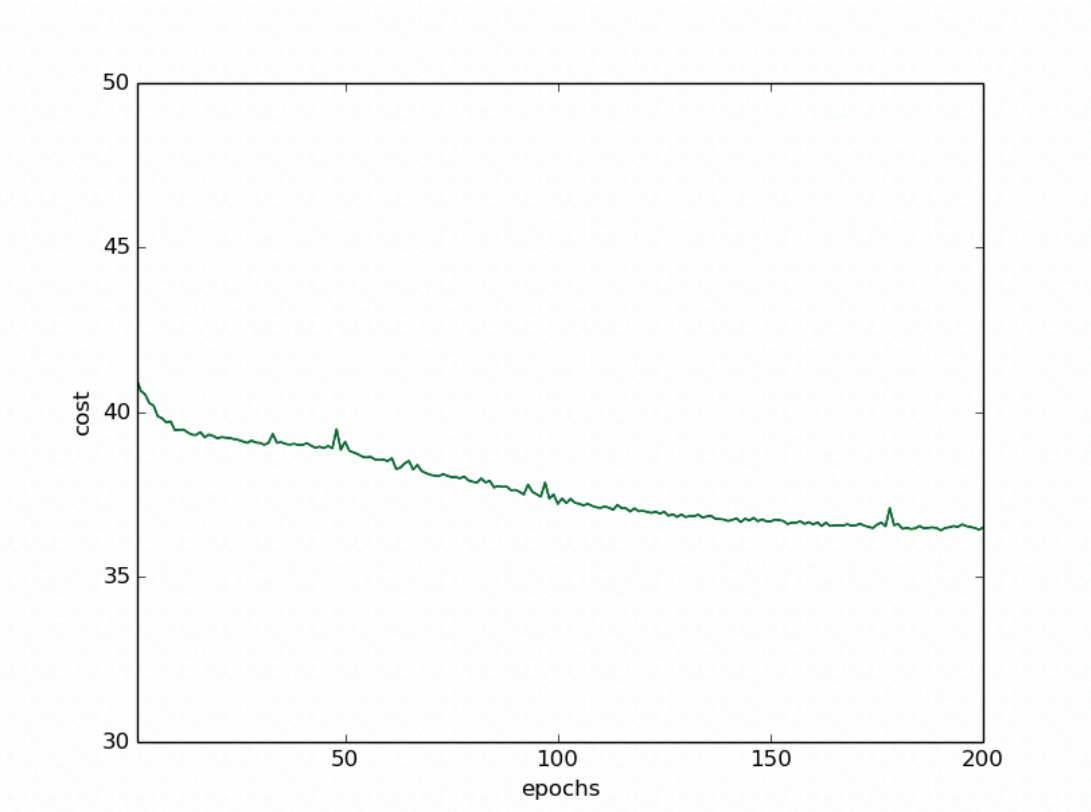

Fig: Error change with epochs for phrase inputs when vocabulary size is 10,000

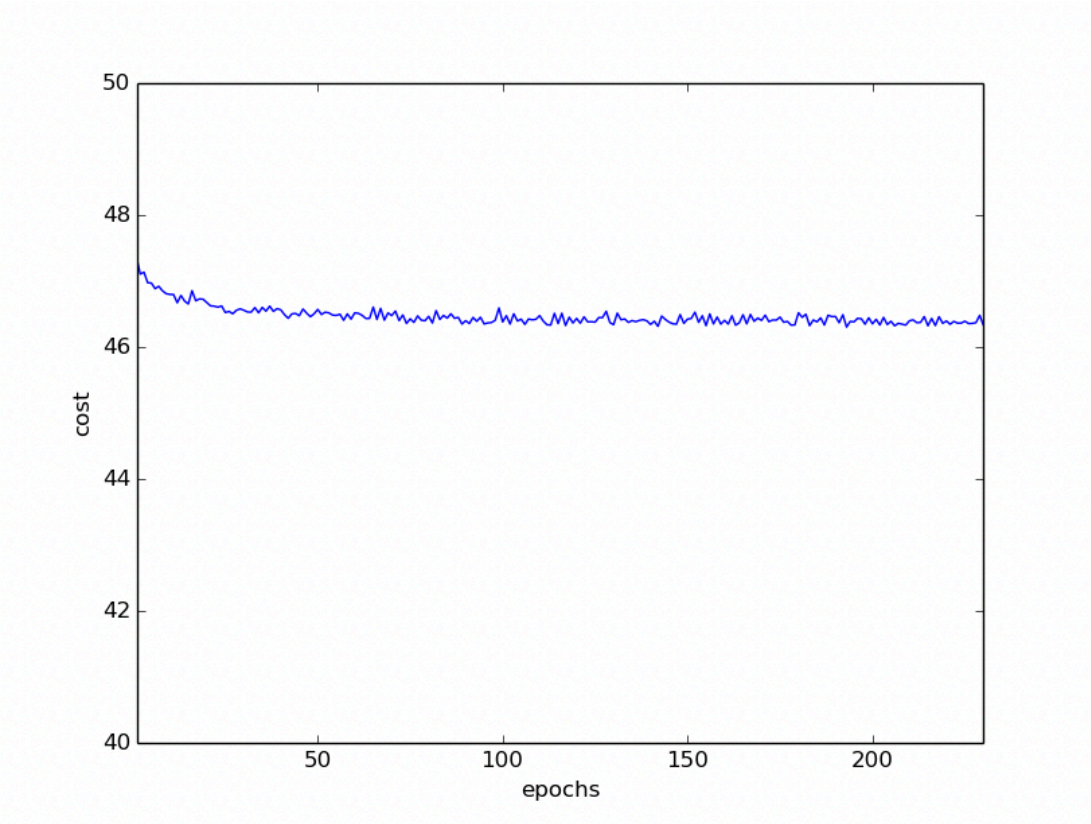

Fig: Cost change with epochs for phrase inputs when vocabulary size is 36,220

Inferences

From the results that we obtained, the following inferences can be made:

- Xu obtained a METEOR score of 0.203 on Flickr8k dataset, whereas we obtained 0.089. This reduced METEOR score can be explained by the difference in the number of epochs that we used. Our model was trained on just 195 epochs against 5000 epochs.

- The two tests on different phrase vocabulary sizes, show a surprising difference. This difference can most likely be attributed to the problem of over-fitting. Given that we had a training set of 8000 images with 5 captions per image, the total number of captions being used for training is 40000. Using only 40000 captions for training a model with a vocabulary size of 36,220 (caption to vocabulary ratio being 1.1) is very likely to lead to over-fitting, and thus poor quality of results. On the contrary, when a vocabulary size of 10,000 is used (caption-vocabulary ratio being 4), over fitting is not that likely to occur.

- The results on word embedding indicate that applying Beam search on LSTM output make the captions more informative. However, this claim is still debatable as the number of epochs over which our model was trained was not reasonably large.

Caption Generation Results

Fig: **Actual:** A tan dog jumps into water
Using word as input: A brown dog is walking in the water
Using phrase as input: A dog walking a shallow river looking into in the water grey shorts the horizon with in down water

References

[1] XU, K., BA, J., KIROS, R., CHO, K., COURVILLE, A. C., SALAKHUTDINOV, R., ZEMEL, R. S., AND BENGIO, Y. Show, attend and tell: Neural image caption generation with visual attention. *CoRR abs/1502.03044* (2015).

[2] COLLOBERT, R., WESTON, J., BOTTOU, L., KARLEN, M., KAVUKCUOGLU, K., AND KUKSA, P. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research 12* (2011), 2493–2537.

[3] DENKOWSKI, M., AND LAVIE, A. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation* (2014).

[4] KIROS, R., SALAKHUTDINOV, R., AND ZEMEL, R. Multimodal neural language models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)* (2014), T. Jebara and E. P. Xing, Eds., JMLR Workshop and Conference Proceedings, pp. 595–603.

[5] LEBRET, R., PINHEIRO, P. O., AND COLLOBERT, R. Phrase-based image caption- ing. *arXiv preprint arXiv:1502.03671* (2015).