

Attention Beam: An Image Captioning Approach (Student Abstract)

Anubhav Shrimal, Tanmoy Chakraborty

Department of Computer Science & Engineering
IIT-Delhi, India
{anubhav18033, tanmoy}@iiitd.ac.in

Abstract

The aim of image captioning is to generate textual description of a given image. Though seemingly an easy task for humans, it is challenging for machines as it requires the ability to comprehend the image (computer vision) and consequently generate a human-like description for the image (natural language understanding). In recent times, encoder-decoder based architectures have achieved state-of-the-art results for image captioning. Here, we present a heuristic of beam search on top of the encoder-decoder based architecture that gives better quality captions on three benchmark datasets: Flickr8k, Flickr30k and MS COCO.

Introduction

Image captioning is an active research area as it provides a gateway for scene understanding where the task is not just object recognition but also to capture the relations between the objects present in the image. Convolutional Neural Networks (CNNs) are known to perform well for feature extraction in images. Long Short Term Memory Networks (LSTMs) have shown great potential in natural language modeling and text generation tasks. The idea to combine the two into an encoder-decoder architecture for image generation was first proposed by (Vinyals et al. 2014; Karpathy and Fei-Fei 2015) in which the pre-trained CNN was used to extract the latent features of an image and represent it in a reduced form which are then fed to a modified RNN coupled with the word embedding inputs and history of the RNN to generate sequence of words, i.e., caption for the image. The extension of this work (Xu et al. 2015) introduced a visual attention network along with the encoder-decoder framework. The intuition was that while captioning an image, rather than looking at the complete image at once, one can look over different regions at each time step to caption it. The objective of attention network was to provide an attention map for the image pixels at each time step of caption generation which allowed the model to look into specific regions of the image while captioning.

We further extend the architecture mentioned above by using beam search (Zhou et al. 2018) at the time of caption generation. It helps in finding the most optimal caption that

can be generated by the model instead of greedily choosing the word with best score at each decoding step. Though beam search has been previously used for image captioning (Ma et al. 2019), we show that using this simple heuristic search along with better training schemes such as teacher forcing gives better scores for different evaluation metrics such as BLEU-1,2,3,4, METEOR, CIDEr and ROUGE-L.

Our code and dataset available at <https://bit.ly/2kUU4g8> and a demo video is available at <https://youtu.be/bO4bvjYyvQE>. A graphical user interface is also created to consume the trained model (see supplementary¹).

Proposed Approach

We propose an encoder-decoder attention based architecture. The encoder is a ResNet-101 model pre-trained on ImageNet dataset. We remove the final classification layer of the model to use it as a feature extractor. The decoder is an LSTM model which takes the feature vector extracted by the encoder as input along with the attention map given by the visual attention model. The attention model gives a weight between 0 and 1 to each pixel in the image. The weighted image along with the word embedding is fed to the LSTM model at each time step which then gives a hidden state and a predicted word for current time step. It is then used by attention and LSTM network for the next decoding step (see supplementary architecture diagram). We use soft attention where the weights of the pixels add up to 1. If there are P pixels in our encoded image, then at each time step t , $\sum_p \alpha_{p,t} = 1$, where $\alpha_{p,t}$ denotes the probability or importance of pixel p at time step t . The other attention mechanism is to use hard attention in which we choose to just sample some pixels from a distribution defined by α . However, it is non-deterministic and non-stochastic. It gives only marginal improvements as compared to soft attention. The following optimizations and heuristics are applied in the proposed model:

- Doubly Stochastic Regularization loss function is used for the attention network. The motivation is to encourage the weights at a single pixel p to sum to 1 across all time steps T so that the model attends to every pixel over the course of generating the entire sequence: $\sum_t \alpha_{p,t} \approx 1$.

¹Supplementary is available at <https://arxiv.org/abs/2011.01753>

Dataset	Model	Evaluation metric						
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	ROUGE-L
Flickr8k (Hodosh, Young, and Hockenmaier 2013)	Vinyals et al. ^{†Σ}	63	41	27	—	—	—	—
	Xu et al. (Soft Attention)	67	44.8	29.9	19.5	18.93	—	—
	Xu et al. (Hard Attention)	67	45.7	31.4	21.3	20.3	—	—
	Ours (Beam = 1)	60.8	43	29.4	19.8	20.9	50.7	46.4
	Ours (Beam = 4)	64	45.8	32.2	22.3	21	55.3	47.1
Flickr30k (Young et al. 2014)	Vinyals et al. ^{†Σ}	66.3	42.3	27.7	18.3	—	—	—
	Xu et al. (Soft Attention)	66.7	43.4	28.8	19.1	18.49	—	—
	Xu et al. (Hard Attention)	66.9	43.9	29.6	19.9	18.46	—	—
	Ours (Beam = 1)	65.1	46.4	32.5	22.7	20.3	48	46
	Ours (Beam = 4)	67.4	49.5	36	26	20.1	52	47
MS COCO (Lin et al. 2014)	Vinyals et al. ^{†Σ}	66.6	46.1	32.9	24.6	—	—	—
	Xu et al. (Soft Attention)	70.7	49.2	34.4	24.3	23.9	—	—
	Xu et al. (Hard Attention)	71.8	50.4	35.7	25	23.04	—	—
	Ma et al. (Beam = 3)	70.6	54.0	40.6	30.5	25.3	97.1	52.8
	Ours (Beam = 1)	77.1	61.4	47.1	35.9	27.9	114.8	57.3
Ours (Beam = 4)	77.9	62.8	49.7	39.3	28.7	120.3	58.5	

Table 1: Performance of all the competing methods for image caption generation: – indicates unknown metric; † indicates a different split; Σ indicates an ensemble. *Beam = 1* is same as not using beam search.

- Fine-tune the final layers of ResNet-101 with a smaller learning rate for the purpose of image captioning as it is originally trained for image classification on ImageNet.
- Use Teacher Forcing to train the decoder in which the ground-truth captions are used as input to the decoder at each time step instead of the word predicted in the previous time step. This speeds up the training time by a significant margin.
- Beam search for better captions. A beam width k , (in our case $k = 4$), is chosen. The algorithm selects the word sequence which has the highest cumulative score of all the words in its sequence as the caption (see supplementary).

Results

Data: The experiments are performed using three benchmark datasets – Flickr8k, Flickr30k and MS COCO, which have 8,000, 30,000 and 82,783 images, respectively. Due to the unavailability of standardized splits for Flickr30k and MS COCO, we use the splits provided in (Karpathy and Fei-Fei 2015).

Quantitative Analysis: We use BLEU-1,2,3,4, METEOR, CIDEr and ROUGE-L as our evaluation metric (see supplementary for formulae). The results with various baselines are shown in Table 1. Beam search is also used by (Ma et al. 2019), but our model gives better results due to the other optimizations and heuristics in the training step.

Qualitative Analysis: Figure 1 shows captions generated by different competing methods. We also compare captions generated with and without beam search, low CIDEr score captions, and visualise the attention network weights (see supplementary).

Conclusion

We proposed beam search heuristic for better caption generation for images on three benchmark datasets which shows that it beats the state-of-the-art approach. The heuristic can be applied to any given image captioning model as well as other language modeling tasks.




Image			
Xu et al.	a woman is sitting at a table with a large pizza	a dog is standing on a hardwood floor	a giraffe standing in a forest with trees in the background
Ours (Beam = 4)	a woman standing in front of a table of food	a dog that is laying under a bed	a giraffe standing in the middle of a forest

Figure 1: Captions generated by different competing methods.

References

- Hodosh, M.; Young, P.; and Hockenmaier, J. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR* 47: 853–899.
- Karpathy, A.; and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 3128–3137. doi:10.1109/CVPR.2015.7298932.
- Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*, 740–755. Springer.
- Ma, Z.; Yuan, C.; Cheng, Y.; and Zhu, X. 2019. Image-to-Tree: A Tree-Structured Decoder for Image Captioning. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, 1294–1299. IEEE.
- Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2014. Show and Tell: A Neural Image Caption Generator. *CoRR* abs/1411.4555. URL <http://arxiv.org/abs/1411.4555>.
- Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2048–2057.
- Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL* 2: 67–78.
- Zhou, G.; Luo, P.; Cao, R.; Xiao, Y.; Lin, F.; Chen, B.; and He, Q. 2018. Tree-structured neural machine for linguistics-aware sentence generation. In *AAAI*.