Image Captioning With Visual Attention

Sambhav Agarwal
Department of Computer Engineering
Maharashtra Institute of Technology,Pune
sambhav03@gmail.com

Himani Mali
Department of Computer Engineering
Maharashtra Institute of Technology,Pune
himani.mali1996@gmail.com

Prithviraj Khelkar
Department of Computer Engineering
Maharashtra Institute of Technology,Pune
pkhelkar26@gmail.com

Soham Mahabaleshwarkar
Department of Computer Engineering
Maharashtra Institute of Technology,Pune
soham.mkar@gmail.com

ABSTRACT
Generating a natural language explanation of an image automatically has attracted the interests of many researchers in recent times because of its significance in practical applications.
Image captioning combines two most important fields of Artificial Intelligence, first is Computer Vision and the second one is natural language processing. Existing approaches start from the gist of an image and converting it into words and because of not focusing on the salient features of the image, sometimes these models are not able to depict the exact relationship between the objects. To overcome these drawbacks, we propose an attention based model that can attend to the objects that are in the foreground of an image.

Keywords: Image Captioning, Visual attention, Neural Network, Natural language processing

1. INTRODUCTION

Automatically creating captions of a picture is a task close to the core of scene understanding- one of the essential objectives of computer vision. Not exclusively should caption generation models be ground-breaking enough to understand the vision difficulties of figuring out which objects are in a picture, they should also be able to analyse the relationship among objects in a scene and put it into meaningful sentence. For this reason, Captioning for quite some time has been seen as a troublesome issue. It is a significant test for Artificial Intelligence algorithms, as it adds up to mimicking the surprising human capacity to pack immense measures of striking visual data into engaging language.
In spite of the difficulties of this task, researchers have been actively trying to find out the solution to image captioning problem. Supported by advances in training neural systems (Krizhevsky et al., 2012) and classifying extensive datasets (Russakovsky et al., 2014), ongoing work has altogether improved the nature of image captioning by utilizing an amalgamation of convolutional neural networks (convnets) to acquire vectorial portrayal of pictures and recurrent neural networks to decipher those portrayals into natural language sentences.

A standout amongst the most inquisitive aspects of the human visual framework is the presence of attention (Rensink, 2000; Corbetta and Shulman, 2002). Instead of compressing a whole picture into a static portrayal, attention takes into consideration the salient features in a picture to dynamically come to the forefront as required. This is particularly significant when there is a too much of messiness in a

picture.Utilizing representations, (for example, those from the top layer of a convnet) that distill data in picture down to the most highlighted items, is one of the successful solution that has been embraced in past work.Unfortunately, this has one potential disadvantage of losing data which could be valuable for more extravagant, progressively graphic captioning. Utilizing more low-level portrayal can help save this data. However working with these features requires an incredible system that can control the model to data significant to the job that needs to be done.

As of late, a few strategies have been proposed for Image Captioning. A large number of these techniques depend on recurrent neural networks and roused by the successful utilization of sequence to sequence training with neural networks for machine translation (Cho et al., 2014; Bahdanau et al., 2014; Sutskever et al., 2014). One noteworthy reason image captioning is appropriate to the encoder-decoder framework (Cho et al., 2014) of machine translation is because it is closely resembling "translating" a picture to a sentence.

There has been a long queue of past work consolidating attention into neural networks for vision associated tasks. Some of them share a similar context as our work incorporate Larochelle and Hinton (2010); Denil et al. (2012); Tang et al. (2014). Specifically, our work straightforwardly broadens the work of Bahdanau et al. (2014); Mnih et al. (2014); Ba et al. (2014).

2. PROPOSED FRAMEWORK

Existing Image Captioning systems sometimes fail to recognise the salient features or actions in an image and hence the captions so produced do not describe the image well.Hence, a need to incorporate the ability to identify the most salient features comes up.

Our system overcomes the same drawback by using the concept of visual attention.

Step 1: Preprocessing using the MS-COCO Dataset

MS-COCO dataset is being used for training and testing purposes.

The images were first pre-processed and then features were extracted from each image. Next InceptionV3 is used (pretrained on Imagenet) to classify each image. We will extract features from the last convolutional layer. Before that is completed, the images must be converted into the format InceptionV3 expects. We will preprocess each image with InceptionV3 and cache the output to disk.

The MS-COCO captions are then tokenized giving us a vocabulary of all the unique words in the data. Word-index mapping is created and vocabulary size is limited.

Step 2: Feature Extraction

The features are extracted from the lower convolutional layer of InceptionV3 where the vector shape is further modified. The system is able to extract features from an image source, for this InceptionV3 is used which has 21 million parameters. We extract the features stored in the files and pass those features through the encoder.

Step 3: CNN Encoder

The vector is then passed through the CNN encoder which consists of a single fully connected layer. The RNN attends over the image to predict the next word. A joint image-sentence embedding is used where

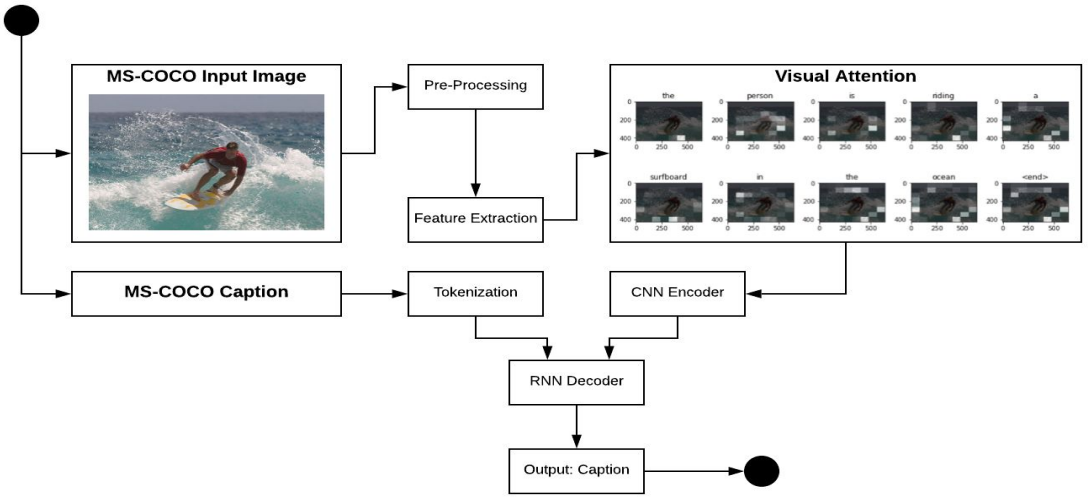


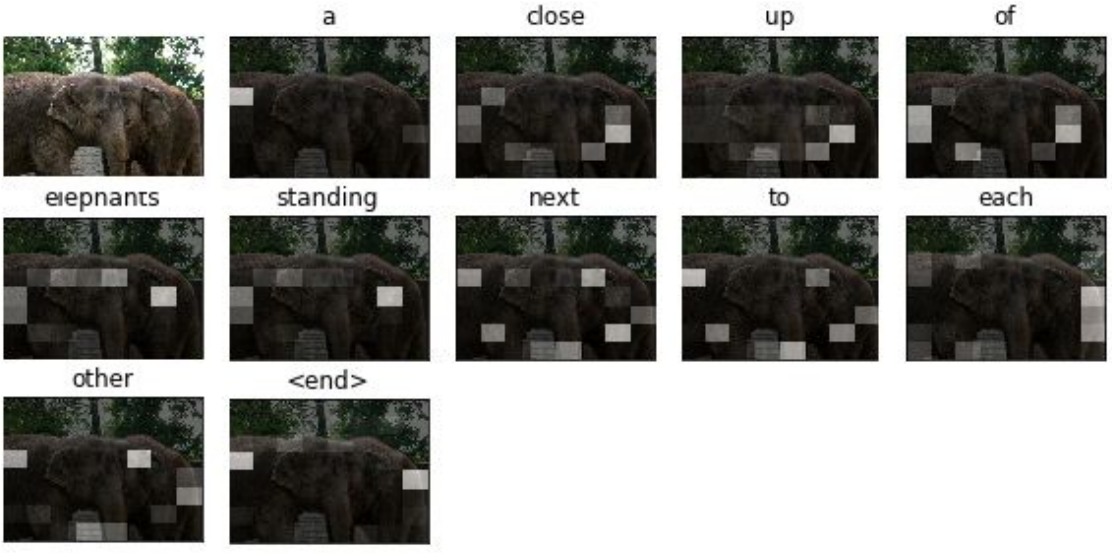
Fig 2.1 System Architecture

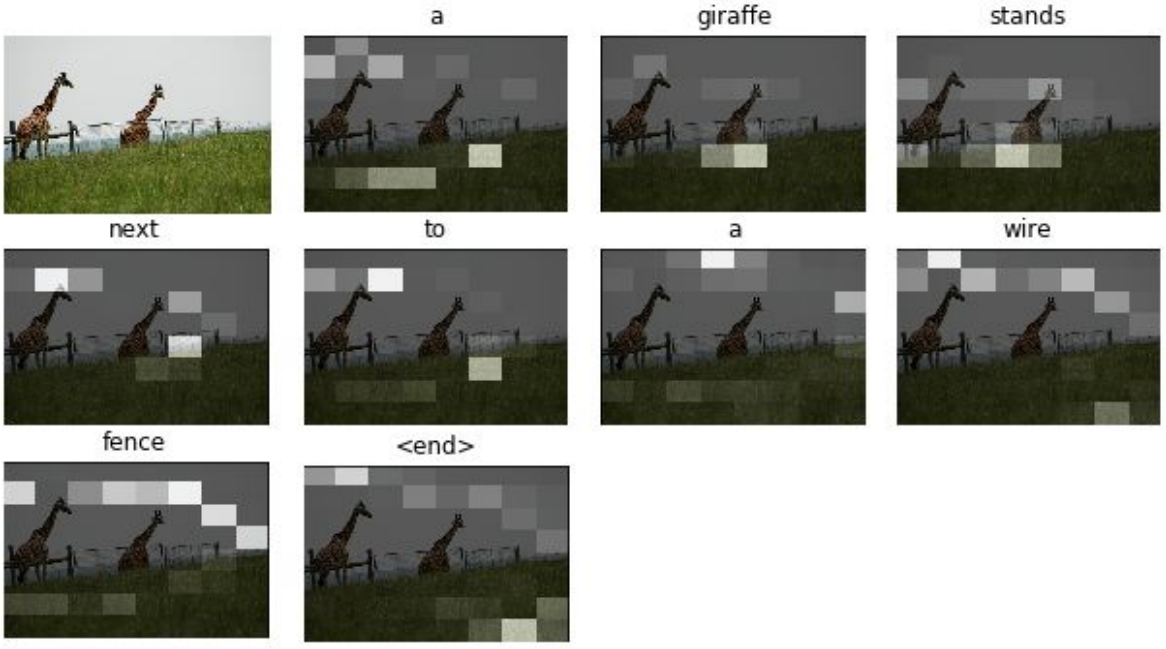
sentences are encoded using long short-term memory RNN. The encoder output, hidden state and the decoder input is then passed onto the decoder.

Step 4: RNN Decoder

The decoder returns the predictions and the decoder hidden state. The decoder hidden state is then passed back into the model and the predictions are used to calculate the loss. Use teacher forcing to decide the next input to the decoder. Teacher forcing is the technique where the target word is passed as the next input to the decoder. The final step is to calculate the gradients and apply it to the optimizer and back propogate.

3. RESULTS





4.CONCLUSION

We have developed a neural network architecture for image captioning. We have used technologies such as Convolutional Neural Networks, Recurrent Neural Networks with attention model to increase the accuracy and viability of this project. Our Model distinguishes itself from the years of previous research in the field of image captioning by singling out the salient features of the image, and hence only generating the most relevant captions that better describe the said image.

5. REFERENCES

[1] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. ”Show, attend and tell: Neural image caption generation with visual attention.”In ICML, 2015.

[2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. ”Show and tell: A neural image caption generator.” In CVPR, 2015.

[3] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick. ”Microsoft coco captions: Data collection and evaluation server.” In arXiv:1504.00325, 2015.

[4] Q. You, H. Jin, Z.Wang, C. Fang, and J. Luo. ”Image captioning with semantic attention.” In CVPR, 2016.

[5] Ba, Jimmy Lei, Mnih, Volodymyr, and Kavukcuoglu, Koray : Multiple ob- ject recognition with visual attention. arXiv:1412.7755, December 2014.

[6] J. Johnson, A. Karpathy, and L. Fei-Fei. ”Densecap: Fully convolutional lo- calization networks for dense captioning.” In CVPR, 2016.

[7] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney,K. Saenko, and T. Darrell. ”Deep compositional captioning: Describing novel object categories with- out paired training data.” In CVPR,2016.

[8] J. Donahue, L. A. Hendricks, S. Guadarrama,M. Rohrbach, S. Venugopalan,K. Saenko, and T. Darrell.” Long-term recurrent convolutional networks for visual recognition and description.” In CVPR,2016

[9] Marco Pedersoli, Thomas Lucas, Cordelia Schmid, Jakob Verbeek Areas of At- tention

for Image Captioning, 2017 IEEE International Conference on Computer Vision (ICCV).
[10] Joseph Redmon, Ali Farhadi. YOLO9000: Better, Faster, Stronger , IEEE Con- ference on Computer Vision and Pattern Recognition (CVPR),2017.
[11] Andrej Karpathy, Li Fei-Fei Deep Visual-Semantic Alignments for Generat- ing Image Descriptions. , IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2015.
[12] Moses Soh Learning CNN-LSTM Architectures for Image Caption Generation. , IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2016.
[13] Ryan Kiros, Ruslan Salakhutdinov, Richard S. Zemel. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models, NIPS 2014 deep learning workshop.
[14] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh Srivastava, Li Deng, Piotr Dollr, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence
Zitnick, Geoffrey Zweig. From Captions to Visual Concepts and Back, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
[15] Jifeng Dai, Kaiming He, Jian Sun.
BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation, ICCV ’15 Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV).
[16] Christian Szegedy Alexander Toshev Dumitru Erhan.Deep Neural Networks for Object Detection, Published in NIPS 2013.