Image captioning model for mobile app

Ankush Govind Chavan[1], Kuldeepsingh Rajpurohit[1], Abhishek Kumar Singh[1], Rishabh Kumar[1], and Mrs. Mansi Bhonsle[2]

[1]UG Students, [2]Assistant Professor, Computer Engineering department,

G.H Raisoni College of Engineering and Management, Pune

Abstract

The latest developments in Deep Learning based Machine Translation and Computer Vision based Object Detection have led to high accuracy Image Captioning models. Although these models are very accurate, they tend to rely on the use of expensive computational power making it difficult to use these models in real-time applications like processing the video streams in real time and extracting the information. In this paper, we carefully follow some of the heuristic strategies and core ideas of Image Captioning and its common methods and present our simple sequence to a sequence based implementation with a remarkable transformation and efficiency such as using beam search instead of greedy search that allows us to implement these on low-end hardware. The proposed system compares the results calculated using a variety of metrics with high-quality models and analyses the reasons behind the model trained on the MS-COCO dataset that are lacking due to trade-off between computation speed and quality. In this proposed system, Restful API endpoint will be created to be used on any device with an internet connection such as a mobile phone, IoT devices, clock, etc, this endpoint used to sent an image to the model running on remote server which in response will generate and sent an caption describing the objects and their relationship with each other in image in a natural language.

Keywords: Neural Networks · Assistive Vision · Caption Generator · Deep Learning · Restful API · Optimization

I. Introduction

Automatically defining image content and their relationships or actions is an important issue for artificial intelligence that connects computer vision and natural language processing. But this can have a profound effect on helping blind people to better understand their surroundings. These pictures can be used to produce captions that can be read aloud to the visually impaired so that they can better understand what is happening around them. This proposed system provides an API endpoint which uses generative model based on a deep recurrent architecture that incorporates the latest advances in computer vision and machine translation and that can be used to create natural sentences describing a previously captured or camera-captured image. The model is trained to increase the chances of interpreting the sentence using the Maximum Likelihood Estimation (MLE) given the training image. What is most impressive about this is that it is a single end model that can be described as predicting captions, given a picture, instead of requiring sophisticated data preparation or a pipeline of specifically designed models.

Not only must the model be able to solve the computer vision challenges of identifying objects in the image, but it must also be smart enough to capture and express object relationships in the natural language. For this reason, image caption generation is considered a serious problem for long. Its purpose is to mimic a person's ability to understand and process large amounts of visual information in descriptive language, making it an attractive problem in the field of AI.

II. System Architecture

In this proposed system, we are creating a RESTful API with a single endpoint that will be used to provide an image to the Image Captioning model running on the server. For creating an API, we will use AWS API Gateway Service and AWS Lambda. AWS Lambda will have a function to send the image received from the API request to the Image Captioning model on the AWS Sagemaker.

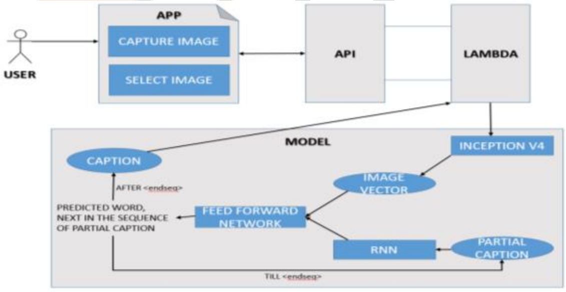

Fig. 1 System Architecture

The starting point of this system will be an application that can run on any platform like a mobile phone, smartwatch, or any IoT devices. This application will send an image through the API request to the AWS Lambda function. The picture that is sent by the application will be a pre-captured image or an image captured by the camera device. The AWS Lambda function will be responsible for transferring this image for further processing to the Image Captioning model where the image is processed, and an appropriate caption describing that image will be generated. This caption will be in a text format. The API response will send this caption back to the application, and this caption will be spoken out loud by the device. Also, the caption will be displayed on the device screen, if that device has a screen.

III. Literature Review

S.No	Paper title	Journal/ Conference Name	Author Name	Problem Discussed	Algorithm/ Technology used	Conclusion
1	Show, Attend and Tell	ICML'16	Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho	Caption generation with visual attention	Attention Based Algorithm	Highly accurate, Slow, Computationally Expensive
2	Show and Tell	IEEE's CVPR'15	Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan	Caption generation using CNN and RNN	CNN and LSTM Based Algorithm	Less accurate, Computationally efficient
3	Deep Visual-Semantic Alignments for Generating Image Descriptions	IEEE's CVPR'15	Andrej Karpathy, Li Fei-Fei	Generating natural language descriptions of images and their regions.	CNN and bidirectional RNN Based Algorithm	Accurate and Fast

III.1. Show and Tell

In this paper, Oriol Vinyals and his team proposes a neural and probabilistic framework to generate descriptions from images. Recent advances in statistical machine translation have shown that, given a powerful sequence model, it is possible to achieve state-of-the-art results by directly maximizing the probability of the correct translation given an input sentence in an "end-to-end" fashion – both for training and inference. These models make use of a recurrent neural network which encodes the variable length input into a fixed dimensional vector, and uses this representation to "decode" it to the desired output sentence. Thus, it is natural to use the same approach where, given an image (instead of an input sentence in the source language), one applies the same principle of "translating" it into its description. Having trained a generative model, an obvious question is whether the model generates novel captions, and whether the generated captions are both d verse and high quality. The result of the human evaluations of the descriptions provided by NIC, as well as a reference system and ground truth on various datasets. We can see that NIC is better than the reference system, but clearly worse than the ground truth, as expected. This shows that BLEU is not a perfect metric, as it does not capture well the difference between NIC and human descriptions assessed by raters. Word embedding can be jointly trained with the rest of the model. It is remarkable to see how the learned representations have captured some semantic from the statistics of the language.

III.2. Show, Attend and Tell

In this paper, Kelvin Xu and his team introduces an attention based model that automatically learns to describe the content of images. One of the most curious facets of the human visual system is the presence of attention. Rather than compress an entire image into a static representation, attention allows for salient features to dynamically come to the forefront as needed. Unfortunately, this has one potential drawback of losing information which could be useful for richer, more descriptive captions. This paper attempts to incorporate two forms of attention variants: a "hard" stochastic attention mechanism trainable by maximizing an approximate variational lower bound or equivalently by REINFORCE and a "soft" deterministic attention mechanism trainable by standard back-propagation methods. A few challenges exist for comparison, which we explain here. The first is a difference in choice of convolutional feature extractor. For identical decoder architectures, using more recent architectures such as GoogLeNet or Oxford VGG can give a boost in performance over using the AlexNet. The second challenge is a single model versus ensemble comparison. While other methods have reported performance boosts by using ensembling, in our results we report a single model performance. In conclusion we saw, an attention based approach that gives the state in the art performance on three benchmarks datasets using the BLUE and METEOR metric. We also saw how the learned attention can be exploited to give more interpretability into the models generation process, and demonstrate that the learned alignments correspond very well to human intuition.

III.3. Deep Visual Sematic Alignment for Generating Image Description

In this paper, Andrej Karpathy and Li Fei-Fei presents a model that generates natural language description of images and their regions. This approach leverages datasets of images and their sentence descriptions to learn about the inter-modal correspondences between language and visual data. This alignment model is based on a novel combination of Convolutional Neural Networks over image regions, bidirectional Recurrent Neural Networks over sentences, and a structured objective that aligns the two modalities through a multimodal embedding. Then describing a Multimodal Recurrent Neural Network architecture that uses the inferred alignments to learn to generate novel descriptions of image regions. Comparing this BRNN model to other captioning models like Google NIC gives a very interesting result, though the Google NIC results surpasses the accuracy of BRNN by a margin but considering the fact that BRNN prioritizes simplicity and speed at a slight cost in performance. And comparing BRNN model with any other retrieval baseline models like nearest neighbour, LRCN etc. BRNN makes better accuracy.

IV. Conclusion

By studying and understanding the researches made in the image recognition, we have used the optimized approach for implementing our proposed system. These papers are served as a important source of relevant information for our proposed system.

V. References

[1] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan (2015) Show and tell: A neural image caption generator. CVPR 1, 2

[2] K. Xu (2016) Show, attend and tell: Neural image caption generation with visual attention. inProc. Int. Conf. Mach. Learn.

[3] Andrej Karpathy, Li Fei Fei (2015) Deep Visual-Semantic Alignments for Generating Image Descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence (April 2017), vol 39, issue 4:664–676.

[4] Ayush Yadav, Rutgers The State University of New Jersey (2017) Camera2Caption: A real-time image caption generator.

[5] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan (2016) Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge PAMI

[6] Wei, Y.; Xia, W.; Huang, J.; Ni, B.; Dong, J.; Zhao, Y.; Yan, S. CNN: Single-label to Multi-label. arXiv 2014, arXiv:1406.5726.

[7] Lin, T.; Maire, M.; Belongie, S.J.; Bourdev, L.D.; Girshick, R.B.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. arXiv 2014, arXiv:1405.0312.

[8] Hodosh, M.; Young, P.; Hockenmaier, J. Framing Image Description As a Ranking Task: Data, Models and Evaluation Metrics. J. Abbr. 2013, 47, 853–899.

[9] Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Trans. Assoc. Comput. Linguist. 2014, 2, 67–78.

[10] Vedantam, R.; Zitnick, C.L.; Parikh, D. CIDEr: Consensus-based Image Description Evaluation. arXiv 2014, arXiv:1411.5726.

[11] Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. SPICE: Semantic Propositional Image Caption Evaluation. In ECCV; Springer: Cham, Switzerland, 2016; pp. 382–398.

[12] 13 Denkowski, M.; Lavie, A. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In Proceedings of the Ninth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Baltimore, MD, USA, 26–27 June 2014; pp. 376–380.

[13] Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the ACL-04 Workshop Text Summarization Branches Out, Barcelona, Spain, 25–26 July 2004; pp. 74–81.

[14] Papineni, K.; Roukos, S.; Ward, T. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318. 20. Farhadi, A.; Hejrati, M.; Sadeghi, M.A.