Video Description and Collision Detection for Visually Impaired

Vinay Kumar Saini (Mentor)[1], Hitesh Kwatra[2], Himanshu Narang[3]

Department of Information Technology, Maharaja Agrasen Institute of Technology

Abstract: ***Using Deep learning techniques, find a new approach that analyses a video and then present it in understandable language using NLP techniques. For most people, watching a brief video and describing what happened is an easy task. For machines, extracting the meaning from video pixels and generating natural-sounding language is a very complex problem. Solutions have been proposed for narrow domains with a small set of known actions and objects. We plan to extract features for each frame, mean pool the features across the entire video and input this at every time step to the LSTM network. The LSTM outputs one word at each time step, based on the video features until it picks the end-of-sentence tag and extends them to generate sentences describing events in videos. They then use a sequence model, specifically a Recurrent Neural Network (RNN), to "decode" the vector into a sentence. In this work, we plan to show that interpreting a visual vector into a set of English words will work same for videos as well as static images. We did this in all the experiments, and it did help quite a lot in terms of generalization. Another set of weights that could be sensibly initialized are We, the word embeddings. We tried initializing them from a large news corpus, but no significant gains were observed, and we decided to just leave them uninitialized for simplicity. Lastly, we did some model level overfitting-avoiding techniques. We tried dropout and ensembling models, as well as exploring the size (i.e., capacity) of the model by trading off number of hidden units versus depth. We also propose collision detection system so that along with getting what is happening around the person, it also gets a collision warning if the distance between the object and the person become smaller than a certain threshold.***

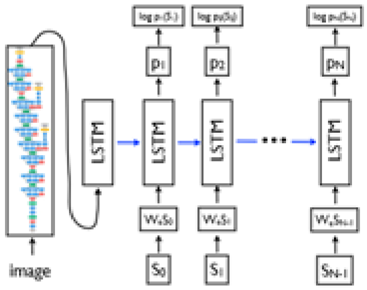
1. Introduction

Need of the Study Because this problem is a million-dollar problem at present and is being researched in many of the premier institutions of the world. Solving the visual symbol grounding problem has long been a goal of artificial intelligence. The field appears to be advancing closer to this goal with recent breakthroughs in deep learning for natural language grounding in static images. In this journal, we propose to translate videos directly to sentences using a unified deep neural network with both convolutional and recurrent structure. Described video datasets are scarce, and most existing methods have been applied to toy domains with a small vocabulary of possible words. For most people, watching a brief video and describing what happened (in words) is an easy task. For machines, extracting the meaning from video pixels and generating natural sounding language is a very complex problem. Solutions have been proposed for narrow domains with a small set of known actions and objects. Shooting, posting, and serving video are relatively easy. Even editing video is easier than it used to be. But captioning has never been easy and has not gotten any easier with the advent of multimedia. Objective: Automatically generating captions to an image shows the understanding of the image by computers, which is a fundamental task of intelligence. For a caption model it not only need to find which objects are contained in the image and also need to be able to expressing their relationships in a natural language such as English. Recently work also achieve the presence of attention, which can store and report the information and relationship between some most salient features and clusters in the image. In our project, we do image-to-sentence generation. This application bridges vision and natural language. If we can do well in this task, we can the utilize natural language processing technologies understand the world in images. In addition, we introduced attention mechanism, which is able to recognize what a word refers to in the image, and thus summarize the relationship between objects in the image. This will be a powerful tool to utilize the massive unformatted image data, which dominate the whole data in the world.

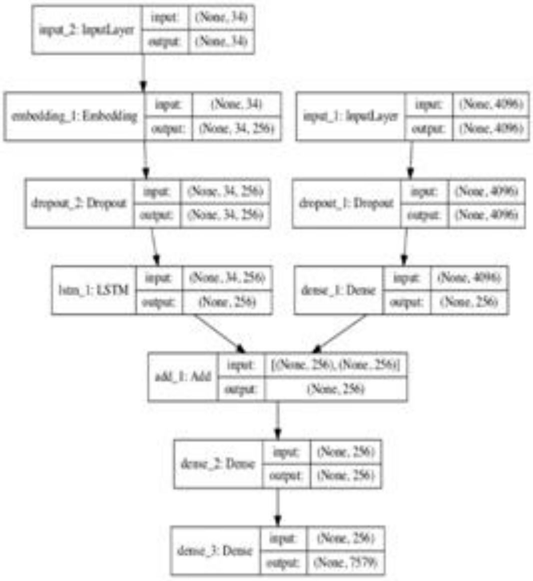
2. Scope of work

A quick glance at an image is sufficient for a human to point out and describe an immense amount of details about the visual scene. However, this remarkable ability has proven to be an elusive task for our visual recognition models.The majority of previous work in visual recognition has focused on labelling images with a fixed set of visual categories and great progress has been achieved in these endeavours. However, while closed vocabularies of visual concepts constitute a convenient modelling assumption, they are vastly restrictive when compared to the enormous amount of rich descriptions that a human can compose. Some pioneering approaches that address the challenge of generating image descriptions have been developed. However, these models often rely on hard- coded visual concepts and sentence templates, which imposes limits on their variety. Moreover, the focus of these works has been on reducing complex visual scenes into a single sentence, which we consider to be an unnecessary restriction.

3. Training Details



We have taken Flickr8k dataset of captioned images for the training of our video description. Each image in the dataset has four captions to it. The task of telling what is happening in an image is harder than object classification and data driven approaches have only recently become dominant thanks to datasets as large as ImageNet. Not to overfit the data for the description of each image we have used VGG16 pretrained for extracting the features of images.The text of the captions is pre-processed by removing all the stop-words and punctuations.



The idea behind the project is to make an Image caption generating model with a great accuracy. To make an Image caption generating model we would need train a Neural Network implementing state of the art concept for image processing as well as handling continuous data. While training the model we would first require to extract the features of the training images with the help of CNN (Convolutional Neural Network), once the features are extracted, the word embedding along with the features will be passed through RNN (Recurrent Neural Networks) layers. After training the model its time to generate the caption given an image. For the getting the best caption out of various captions Beam search would be used and the captions getting best BLEU (Bilingual Evaluation Understudy) score would be the final caption output. The caption we get as output, given an image is then delivered to the user through headphone like device in the form of voice message.Once model is trained frames from the live video are extracted skipping a few for each interval of time.

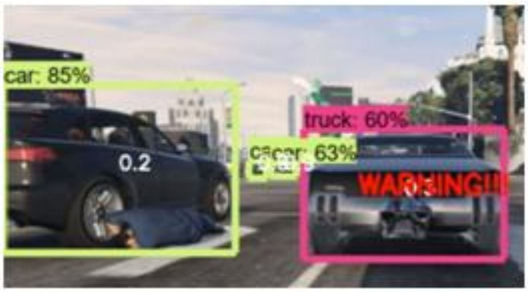
For collision detection we have used Tensorflow object detection API, we predicted the bounding box of each object present in the image. Once we get the bounding we calculate the distance between the user and the object detected by relative increase/decrease in size.

Image Captioning Results

startseq man in black shirt and jeans is standing by an adult in black and white shirt endseqstartseq man in black shirt and jeans is standing by an adult in black and white shirt endseq

startseq dog is running through the snow endseq

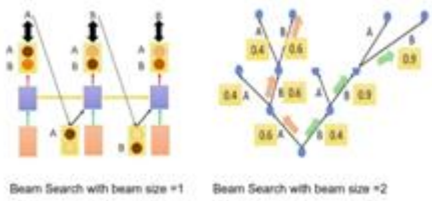
Collision Detection Results



4. Conclusion

We have presented an end-to-end neural network system that can automatically view an image and generate a reasonable description in plain English. NIC is based on a convolution neural network that encodes an image into a compact representation, followed by a recurrent neural network that generates a corresponding sentence. The model is trained to maximize the likelihood of the sentence given the image. Experiments on several datasets show the robustness of NIC in terms of qualitative results (the generated sentences are very reasonable) and quantitative evaluations, using either ranking metrics or BLEU, a metric used in machine translation to evaluate the quality of generated sentences. It is clear from these experiments that, as the size of the available datasets for image description increases, so will the performance of approaches like NIC. Furthermore, it will be interesting to see how one can use unsupervised data, both from images alone and text alone, to improve image description approaches.

Right now we have used greedy method for generating descriptions to improve result, we propose to use beam search for generating the sentences by following more accurate methodology of predicting the next word in a given sentence.



We have also presented a collision detection system based on object detection in images and calculating their distances with the concept on relative size in the image.

5. Limitations

- The dataset chosen for the training purposes has huge effect on the description of the test images. For example the dataset we used, Flickr8k had a lot of dog pictures in it so our final model was biased towards dogs.
- The camera used should meet minimum quality requirements else if the images generated are grainy the product would give unexpected and worst results.
- The collision warning widely depends upon the resolution and capturing angle of the camera.
- The correctness of objects detected is dependent upon the neural architecture we choose. For research purposes we have used various pretrained models available at Tensorflow model Zoo. Out of all neural nets ResNet gave the most accurate results but MobileNet happened to be the fastest, giving fair amount of true positives.

6. Future Scope

We believe the system we have proposed is going to make difference in the lives of visually impaired people by making them connect more to their surrounding by giving them descriptions of what is happening around them. This will help bridging the gap between the normal and the visually impaired people. They would no longer be guessing about the collisions but our system would tell them, if there's any obstacles on their way.

References

[1] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using amazon's mechanical turk. In NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 139– 147, 2010.

[2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, 2014.

[3] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229, 2013.

[4] R. Socher, A. Karpathy, Q. V. Le, C. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describ- ing images with sentences. In ACL, 2014.

[5] Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In NIPS, 2014.

[6] R. Vedantam, C. L. Zitnick, and D. Parikh. CIDEr: Consensusbasedimage description evaluation. In arXiv:1411.5726, 2015.

[7] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu. I2t: Image parsing to text description. Proceedings of the IEEE, 98(8), 2010.

[8] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From im- age descriptions to visual denotations: New similarity met- rics for semantic inference over event descriptions. In ACL, 2014. [34]W. Zaremba, I. Sutskever, and O. Vinyals. Recurrent neural network regularization. In arXiv:1409.2329, 2014.