Empowering Visually Impaired People using Deep Learning

Savan Visalpara[1], Kashyap Raval[2], Prof Ajaykumar T. Shah [3]
Department of Computer Engineering
[1,2] Alpha College Of Engineering And Technology
[3] HOD, Alpha College Of Engineering And Technology

Abstract- *Deep learning has shown tremendous success in vision, language, and speech related tasks. These advancements also open a room for many applications that can make a significant impact on our lives. In this paper, we show our work on a device that leverages advancements in deep learning to help visually impaired people understand their surroundings. We show that techniques such as image captioning and visual question answering can be combined together to build a device that can act as a virtual assistant to visually impaired people. We call this device "ThirdEye". We also show how it can be challenging to productize such device due to heavy computational requirements. Finally, we conclude this paper by sharing our thoughts on possible future improvements.*

Keywords- smart device for visually impaired people, applied deep learning, image captioning, visual question answering.

I. INTRODUCTION

Machine learning methods are proved to be state-of-the-art in many areas such as computer vision and natural language processing [5, 9, 10]. Since our work is mainly about vision, we will narrow down our focus to computer vision only. Currently, deep learning based methods are state-of-the-art in almost every computer vision problem such as object detection, object segmentation, pose estimation, etc. It has also made it possible to solve tasks which were conceived to be impossible before the advent of deep learning (i.e image captioning [3]). Previous works in smart glasses were limited to object detection, face recognition and similar tasks. Most of these devices used traditional computer vision techniques.

Here, we introduce a new concept for a device that uses image captioning and visual question answering to help visually impaired understand their surroundings better than before. These two features can be used together to build an ideal virtual assistant for visually impaired people. We can also add other features (i.e optical character recognition) into this device. We believe that such a device can be very helpful to visually impaired people, given that we have solved challenges mentioned in this paper.

II. RELATED WORK

There have been a lot of effort to make smart glasses using various technologies. Most of these used traditional computer vision techniques to implement object detection and segmentation related tasks. Some previous work used deep learning methods for tasks such as face recognition. But, there is no previous work in applying deep learning techniques (i.e image captioning) to smart glasses.

III. THIRD EYE

In order to load and process modules for various features such as image captioning, we need to have a small processing unit. If we are interested in developing a smart glasses with features mentioned earlier, then a custom processing chip will be required. It is also possible to use a standard tiny computer (i.e raspberry pi) for this glasses, but we believe it would be comfortless. It is also possible to design a cap that has a processing unit on top and a camera on the front side. For this experiment, we used raspberry pi 3 with a pi camera on the front to capture an image. Flowchart of an ideal interaction with a device is shown in Figure-1.

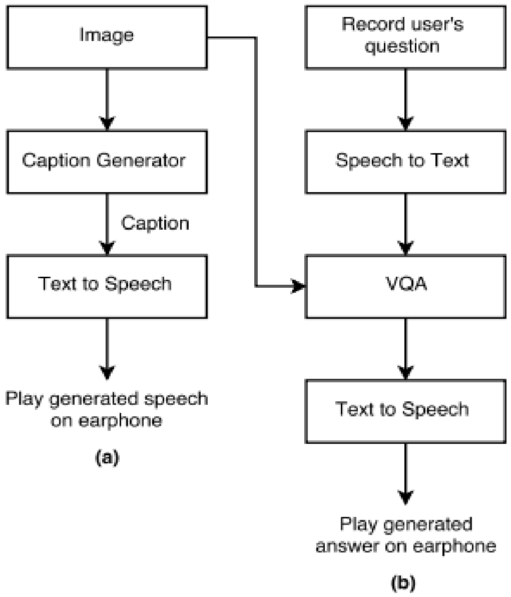

Figure 1. (a) Caption generation process (b) Visual question answering process

A user can invoke image captioning module by pressing a button on a device. First of all, this will capture an image using a pi camera and this image will be sent to image captioning module to generate a caption. If a user finds generated caption vague or if he/she has any questions regarding this image he can invoke a VQA module by pressing a button. Now, a user needs to speak up his question near to the microphone. As soon as the user stops speaking, the recorded speech will be sent to a speech-to-text module and then this text will be sent to a VQA module which takes this text as well as the captured image as an input. VQA module will process this input and an answer will be generated. This text answer will be sent to a text-to-speech module and speech will be played on a connected headphone. If a user has more questions, he/she needs to press the button again and the same process will take place. An important thing to note here is that an image captured during image captioning task needs to be in memory so that it can be used by a VQA module. Once a user asks for another caption by pressing the button, the previous image will be flushed. Although image captioning and visual question answering are different methods, they are synchronized in this case and works as a single unit.

In the following sections, we briefly explain image captioning and visual question answering.

3.1 IMAGE CAPTIONING

Image captioning is a task of generating a caption of a given image [3]. Currently, deep learning based methods are state-of-the-art for this task. An architecture used to train such a model is usually composed of CNN [5] and RNN (i.e LSTM) [11], where CNN acts as a visual feature extractor and RNN acts a caption generator.

3.2 VISUAL QUESTION ANSWERING

Visual question answering is a task of generating a relevant answer of a question given an image and a question as an input [4]. A simple VQA system is shown in Figure 3.

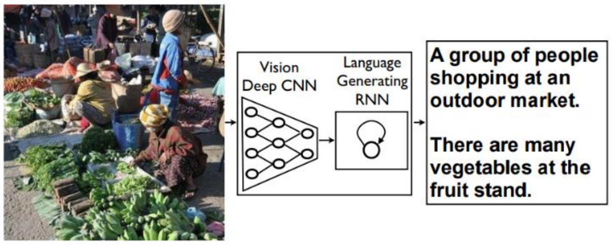
Vision
Deep CNN
Language
Generating
RNN
A group of people shopping at an outdoor market.
There are many vegetables at the fruit stand.

Figure 2. A Neural Image Caption Generator (NIC) model [3].

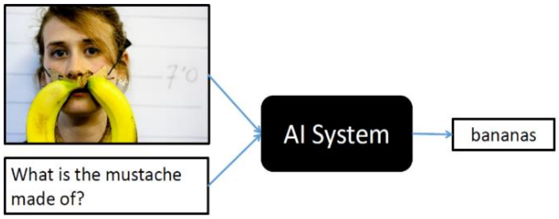

Figure 3. A simple VQA system [4].

IV. CHALLENGES

This type of system comes with a big issue of inference speed. Deep learning models used in such tasks usually have millions of parameters and that requires a lot of computation to generate an answer. There are two ways to improve inference speed. One is to have a better hardware and the second one is to optimize deep learning models as much as possible. There are several chips which are specially designed for deep learning. For example, Nvidia and Intel have released their chips specially designed to accelerate the speed of deep learning based programs. Since we used TensorFlow to develop and train our models, we had to find ways to optimize this code. We used TensorFlow because we found that it is the best match for our project and it is also widely used in production. One of the ways to optimize TensorFlow graph is to freeze it and the other one is to perform some operations that can improve the inference speed. Some of these operations are stripping out parts of the graph that are never reached and folding batch normalization operations into the pre-calculated weights. Fortunately, TensorFlow provides a script that implements all such operations that can improve the inference speed.

Another challenge is of accuracy. We need to make sure that deep learning models used in such devices are accurate enough to generate relevant captions. Since users might take actions based on content generated by these models, it is very important that these models are robust to unseen data as well as adversarial inputs [12].

V. FUTURE WORK

Below we discuss some of the features that we believe can make this device more helpful.

Optical Character Recognition- OCR is the task of extracting text from an image. Having this feature in this device will allow visually impaired people to read from any kind of source.

Danger Alert- Currently, we are working to understand whether this feature is feasible or not. The idea is to find situations that can hurt users using various computer vision and image processing techniques. This can be more effective if these techniques are applied continuously (i.e live video streaming) and that leads us to our next feature.

Video Captioning- Capturing an image every time you want to generate a caption might not be comfortable for every user. For this reason, we believe video captioning [7, 8] will be of great help to visually impaired people.

One of the things that we are constantly working on is the accuracy of machine learning models. We need to make sure existing machine learning models are highly accurate and robust against unseen data.

VI. CONCLUSION

In this paper, we introduced a new idea for a device that leverages deep learning methods such as image captioning and visual question answering to help visually impaired people understand their surroundings. This device can be a glasses or any custom device that can be placed on a cap. Machine learning models used in such tasks usually have millions of parameters and it poses a computational challenge for us. It is very important to build accurate and robust models since users are likely to take actions based on answers generated by these models. Recent work in adversarial examples also makes it inevitable to build models that are robust to adversarial examples. We believe we can add more such features that can help visually impaired people live like a normal people.

REFERENCES

[1] Stuart Elder and Alex Vakaloudis. A technical evaluation of devices for smart glasses applications. In Internet Technologies and Applications, 2015.

[2] Nanoka Sumi and Vasily Moshnyaga. A novel face recognition for smart glasses. In IEEE Region 10 Symposium, 2016.

[3] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan. Show and Tell: A Neural Image Caption Generator. In CVPR, 2015.

[4] Aishwarya Agrawal, Jaisen Lu, Stanislaw Antol, Margaret, Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh. Visual Question Answering. In ICCV, 2015.

[5] Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton. ImageNet classification with deep convolutional neural networks. In Neural Information Processing Systems, 2012.

[6] Bo Dai and Dahua Lin. Contrastive learning for image captioning. In Neural Information Processing Systems, 2017.

[7] Jiajun Sun, Jing Wang, Ting-chun Yeh. Video Understanding: From Video Classification to Captioning.

[8] Da Zhang, Hamid Maei, Xin Wang, Yuan-Fang Wang. Deep Reinforcement Learning for Visual Object Tracking in Videos. Arxiv pre-print, Jan 2017.

[9] Alex Graves, Abdel-rahman Mohamed, Geoffrey E. Hinton. Speech Recognition with Deep Recurrent Neural Networks. In ICASSP, 2013.

[10] Yonghui Wu, et al. Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.

[11] Sepp Hochreiter and Jurgen Schmidhuber. Long Short-Term Memory. In Neural Computation, 1997.

[12] Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Cho-Jui Hsieh. Show-and-Fool: Crafting Adversarial Examples for Neural Image Captioning. Arxiv pre-print, Dec 2017.