CS230

Image Contextualization for the Visually Impaired

Aditya Dusi, Asish Koruprolu, and Hitha Revalla

Department of Electrical Engineering, Stanford University
(adusi,asishk,hitha)@stanford.edu

Abstract

Humans are gifted with a complex visual system that is fast, accurate and performs highly convoluted tasks with only a little conscious thought. Mimicking this behavior, to convey real-time scene information to the visually impaired is a novel task. In this paper, we present an end-to-end system that performs object detection and classification, caption generation and speech synthesis to generate natural sentences describing an image. This model combines a convolutional and a recurrent neural network and pipelines these results to a text-to-speech generator to produce image descriptions for a given image. Architecture search is performed on the given networks and the performance of various architectures are compared using BLEU, SPICE, CIDEr, ROGUE_L and METEOR scores.

1 Introduction

Vision is the most important sensory stimulus. It helps us perceive the world and make sense of it. Vision impairment or vision loss is the lack of this perception. Having to deal with vision loss is challenging and it highly undermines a person's understanding of the surroundings. Technological advances now have the power to impact such lives for the better. Leveraging the omnipresence of computer assisted devices, building a descriptive image captioning module on such devices, can make the world more palpable to the visually impaired. Indeed, the description must capture not only the objects contained in an image, but also must express how various objects relate to each other. Their key attributes and the activities they are involved in must also be apprehended. Moreover, the above semantic knowledge has to be expressed in a natural language like English, which means that a language model is needed to buttress the visual understanding.

The model takes images from the *COCO* dataset as inputs for training. These images have bounding boxes around objects, their category IDs and 5 descriptive captions. The model learns on these images to output the best descriptive sentence of the image.

2 Related work

Object detection and image captioning are revolutionizing a lot of areas of research. The network presented in this paper is based on the work done by Karpathy and Fei Fei [1]. The authors use a combination of a convolutional neural network (CNN) and a recurrent neural network (RNN); the CNN performs object-detection and classification while the RNN converts the output vector from the CNN to a sentence. This specific paper explains about the intricacies of aligning a word/phrase to various objects in the input image. The key takeaway is the idea of multi-modal embedding, i.e, a mixture of media like image and the word embeddings. This work outperformed the state of the art (at that time) on two fronts, that is overcoming the hard-coded sentence structures (essentially templates) and secondly the generation of captions rich in semantic information.

CS230: Deep Learning, Fall 2018, Stanford University, CA.

Another notable work in this field is by Redmon et al., (2016) [2]. Here, the authors propose a new algorithm to perform real time object detection and classification for images and videos. The novelty of their algorithm in encapsulated in the title- *You Only Look Once*. Conventional algorithms at the time made multiple passes though the image, trying to identify different objects in different passes. *YOLO* just makes a single pass through the image and identifies various objections along with some confidence attached to those predictions.

A seminal paper in image captioning uses an inception network to perform the task of image captioning [1]. Inception nets perform well at image captioning as they use convolutions of different sizes; filters of different sizes work differently for objects of different sizes. These nets compute the convolution with filters of various sizes and then perform a depth concatenation. Inception nets, however, are computationally very expensive and hence the authors propose an alternative architecture- use simple convolution layers in the starting layers of the neural network and then switch to utilizing inception blocks after a certain depth in the network.

3 Dataset and Features

There are a couple of data sets tailored for the task of image captioning- *MS COCO* [3], *Pascal VOC* [4] and *ImageNet* [5].

Pascal VOC is a standardized dataset that comprises of objects from over 20 different classes. While this data set could potentially be a good candidate to train on, 20 classes are too less and in a regular day-to-day scenario, one encounters objects from many more classes.

ImageNet is one of the largest visual database designed for use in visual object recognition software research that contains over 14 million images with subjects spanning over $20,000$ classes. Training over such a large and diverse dataset not only requires larger compute resources but also is unnecessary for classifying day-to-day objects.

The MS COCO 2014 dataset has $80,000$ images binned in 80 classes. These images reflect everyday scenes and provide contextual information. COCO dataset provides bounding boxes around objects along with 5 sentences describing each of the images. The exact data split and usage is described in section 5.

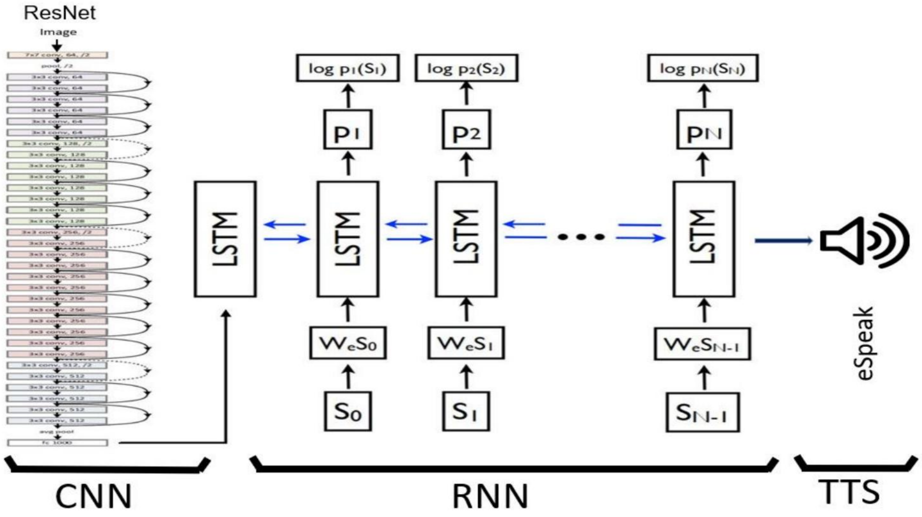

Figure 1: Neural Network Architecture

4 Architecture

The architectures used in the CNN and RNN part of the end-to-end model are elucidated in Figure 1. The CNN is used for extracting features from the images and the RNN generates captions from those extracted features.

4.1 Convolutional Neural Network (CNN) Architecture Search

The model initially used the *VGG*-16 network architecture [6]. *VGG*-16, the simplest of the VGG models was chosen as the performance of both the networks is almost similar. The architecture is simple, using only 3 x 3 convolutions throughout. Although the network is relatively shallow, the use of 2 Fully-Connected layers at the end significantly increase the number of parameters to around 138 million. This makes the network difficult to train and various techniques have to be applied such as breaking the network into many smaller chunk and training each of them separately.

This is enough impetus to switch to newer architectures like *Residual Networks* (ResNets). In the *ResNet*-101 (has 101 layers with learnable parameters) model, there is no restriction on the filter size for the convolutions and these networks are easier to train as they have a skip connection between every three layers that assuages the issue of vanishing activations/ gradients, unlike in *VGG*-16. The number of parameters is also significantly less (around 12.4 million) as they don't use any Fully-Connected layers. This is the architecture that was subsequently utilized for the CNN portion of this work.

4.2 Recurrent Neural Network (RNN) Architecture Search

The model first utilized a uni-directional long short-term memory (LSTM) based RNN as described by the Neural Image Caption generator (Show and Tell) [7]. This is a generative model based on deep recurrent architecture that generates natural sentences. Following this, a reinforcement learning based model was adopted to generate captions [8]. This is a self critical sequence training (SCST) model which uses the reward of the current model's inference algorithm to baseline the reinforce algorithm. Finally, the top-down attention LSTM in pair with a language LSTM model based on mimicking the human visual system, proved to give the best scores as shown in the results section [9]. This top-down mechanism RNN generates captions based on the CNN's prediction of attention distribution on image regions.

4.3 Text to Speech

For the specific application of generating auditory descriptions, captions generated by the RNN are fed to a speech synthesizer. There are a plethora of open source packages that do this- eSpeak [18], GNU talk [19], ASRA toolbox [10] and CSTR Merlin [11]. Each of these are based on different speech synthesis techniques- formant based, concatenative and rule based (Neural Network). eSpeak was the best candidate for this task as it is simple to interpret and doesn't need a lot of pre-processing or training compared to the others.

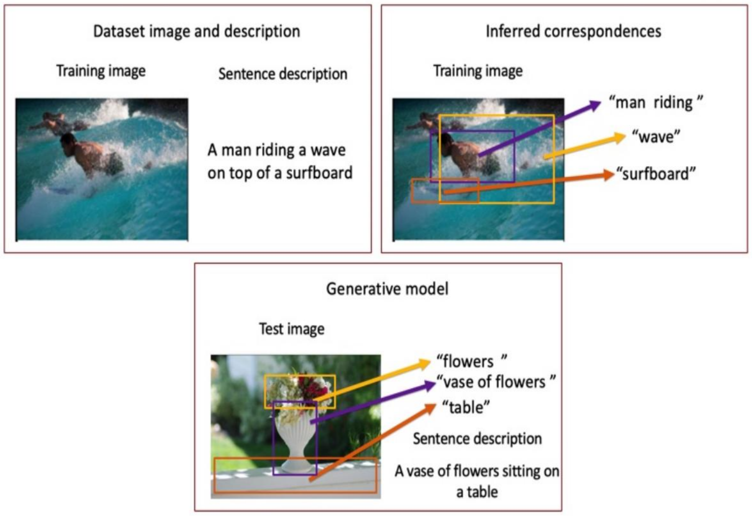

Figure 2: Bounding boxes and Images Captioning preview

5 Methods

The input image is fed to the CNN. The model maps the latent alignment between sentence segments and the region of the image. It infers these correspondences and then learns to generate novel descriptions. These descriptions are generated by sampling from a given distribution at each time step of the RNN and then by performing BEAM search to find out which group of words describes the image most accurately. The training data has $82,783$ examples and the validation set has $40,504$ examples [3]. The Karpathy test split is used. Each epoch comprises of $11,328$ images and the network randomly samples 6000 images and trains on them before validating the performance of the network with 5000 images [17]. Figure 2 shows various objects along with bounding and captions describing them. It also shows how the neural network breaks down a caption to interpret the images.

6 Results

The performance of the models is scored on the relevance of the generated description to the input image. The score should also consider the closeness to a human generated text. Figure 4 shows the captions generated by the neural network. The following metrics were identified to bring out a comparative study of the different models used: BLEU_1, BLEU_2, BLEU_3, BLEU_4, SPICE, CIDEr, ROGUE_L and METEOR [12-16].

Each of the models- Top-Down, Show and Tell and Att2in2 are compared to the current state of the art and their corresponding scores are plotted in Figure 3. The corresponding hyper parameters for the CNN are shown in Table 1. The RNN parameters are finally set to the same preset parameters in the repository as they work best. A minibatch size of 16 is used along with techniques like learning rate decay and gradient clipping in order to avoid over-fitting. Top-Down model is the best performing model when compared to state of the art. The performance gap may be due to the dataset- this work uses *COCO* but ImageNet is bigger and has more classes and hence the network can generalize better; the price is that the network has to be trained for weeks on high-end GPUs. The network also reflects the bias in the dataset- when are always associated with 'mobile phones', umbrella with 'rain' and pillars/ buildings with 'clock towers'. *MS COCO* also lacks object from some categories like tigers, specific vegetables like pumpkins, etc. Rather than creating another dataset, transfer learning was employed by picking up a ResNet trained on ImageNet.

LR	Optimizer	α	β	grad. clip	LR dec
10^{-4}	*Adam*	0.8	0.999	0.1	10^{-5}

Table 1: Hyperparameters used in the network

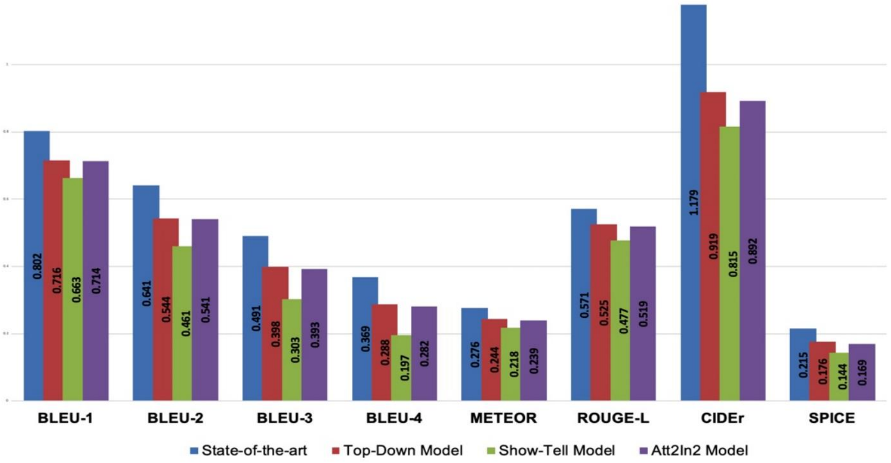

Figure 3: Scores of various models when compared to State of the art

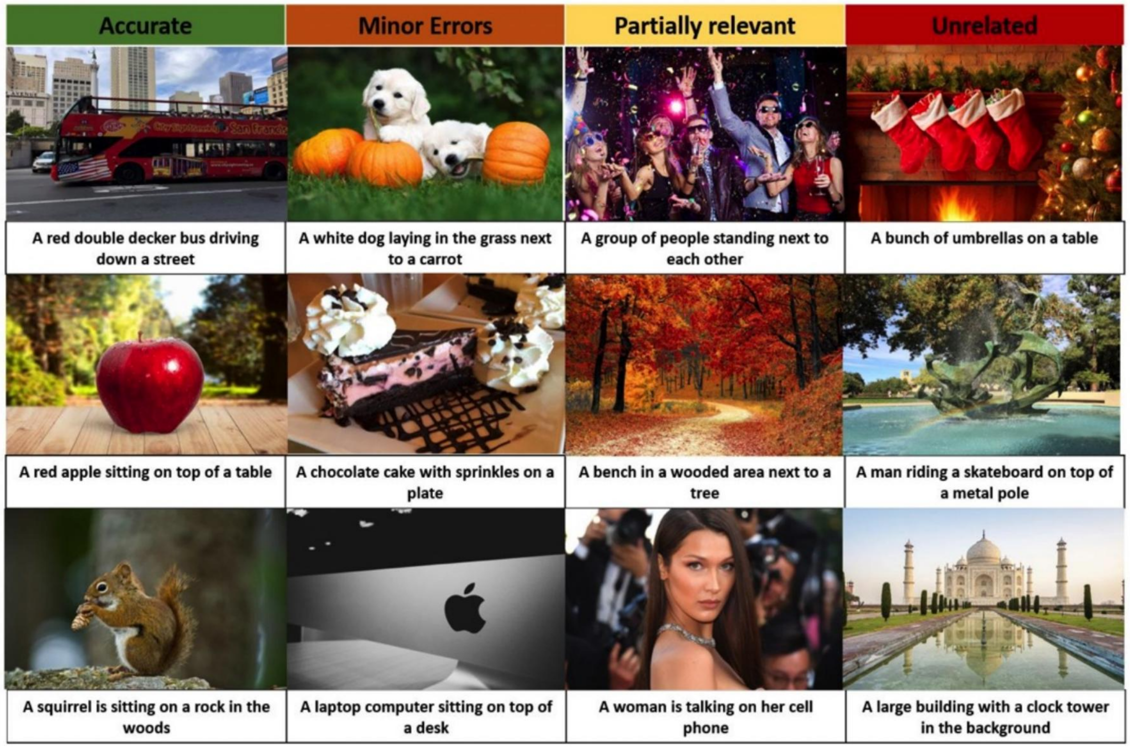
Accurate
Minor Errors
Partially relevant
Unrelated
A red double decker bus driving down a street
A white dog laying in the grass next to a carrot
A group of people standing next to each other
A bunch of umbrellas on a table
A red apple sitting on top of a table
A chocolate cake with sprinkles on a plate
A bench in a wooded area next to a tree
A man riding a skateboard on top of a metal pole
A squirrel is sitting on a rock in the woods
A laptop computer sitting on top of a desk
A woman is talking on her cell phone
A large building with a clock tower in the background

Figure 4: Sample Output captions from the model

7 Conclusions

Various base models were trained and modified to increase the performance metrics. Initially a *VGG-Net* was used for the image captioning task but later a switch was made to ResNets. Models namely, Show and Tell, Top-Down and Att2in2 were used. The Top-Down model outperforms the other architecture models because it enables deeper image understanding using the concept of attention similar to the way the human eye perceives images.

8 Future Work

As an extrapolation to this project, the model can be ported on to a mobile platform to generate auditory descriptions for the visually impaired. The use of COCO dataset showed a few errors in generalization. Given more time and compute resources, training could be performed over a larger and a more diverse dataset like ImageNet. Building a new dataset by appending vocal description of objects, an end-to-end system could also be built.

9 Contributions

Aditya Dusi worked on the RNN architechture search and TTS implementation. Asish Koruprolu sifted through previous GitHub implementations to evaluate the best fit for this application and debugging the repositories. Hitha Revalla worked on the implementing hyperparamter tuning experiments and CNN architecture search. All three individuals worked equally and collaboratively on brainstorming for the idea, initial stage planning, poster and the final report.

The code described in this paper is available at https://github.com/AsishKoruprolu/Image2Speech.

References

[1] Andrej Karpathy & Li Fei Fei," *Deep Visual-Semantic Alignments for Generating Image Descriptions*, IEEE Transactions on Pattern Analysis and Machine Intelligence", Vol. 39 issue. 4, (2017).

[2] Joseph Redmon, Santosh Divvala, Ross Girshick & Ali Farhadi," *You Only Look Once: Unified, Real-Time Object Detection*", IEEE Conference on Computer Vision and Pattern Recognition, (2016).

[3] Tsung-Yi Lin *et al.* ," *Microsoft COCO: Common Objects in Context*", European Conference on Computer Vision (2014), pp74-755.

[4] Everingham, M. and Eslami, S. M. A. and Van Gool, L. and Williams, C. K. I. and Winn, J. & Zisserman, A., "*The Pascal Visual Object Classes Challenge: A Retrospective*", International Journal of Computer Vision, Vol. 111, pg. 98-136, (2015).

[5] Deng, J. and Dong, W. and Socher, R. and Li, L.-J. and Li, K. and Fei-Fei, L., "*ImageNet: A Large-Scale Hierarchical Image Database*", IEEE Conference on Computer Vision and Pattern Recognition, (2009).

[6] K. Simonyan & A. Zisserman," *Very Deep Convolutional Networks for Large-Scale Image Recognition*", ImageNet Large Scale Visual Recognition Challenge, (2014).

[7] Oriol Vinyals, Alexander Toshev, Samy Bengio & Dumitru Erhan,"*Show and Tell: A Neural Image Caption Generator*", IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, (2015).

[8] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross and V. Goel, "*Self-Critical Sequence Training for Image Captioning*", IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, (2017).

[9] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould & Lei Zhang,"*Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering*", IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, (2018).

[10] Jyh-Shing Roger Jang," *Audio Signal Processing and Recognition*", Google London, Massachusetts Institute of Technology, Feb. (2017).

[11] Zhizheung Wu, Oliver Watts, Simon King," *Merlin: An Open Source Neural Network Speech Synthesis System* ", In proc. of 9th ICSA Speech Synthesis Workshop (SSW9), Sunnyvale, CA, USA (2016).

[12] Ramakrishna Vedantam, C. Lawrence Zitnick & Devi Parikh," *CIDEr: Consensus-based image description evaluation*", IEEE Conference on Computer Vision and Pattern Recognition, (2015).

[13] Peter Anderson, Basura Fernando, Mark Johnson & Stephen Gould," *SPICE: Semantic Propositional Image Caption Evaluation*", IEEE Conference on Computer Vision and Pattern Recognition, (2016).

[14] Satanjeev Banerjee & Alon Lavie," *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments*", in Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, 43rd Annual Meeting of the Association of Computational Linguistics, Ann Arbor, Michigan, (2005).

[15] Lin, Chin-Yew & Franz Josef Och.," *Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics*", In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, (2004).

[16] Papineni, K., Roukos, S., Ward, T. & Zhu, W. J.," *BLEU: a method for automatic evaluation of machine translation*". 40sth Annual meeting of the Association for Computational Linguistics, pp. 311–318, (2002).

[17] Ruotian Luo," *An Image Captioning codebase in Pytorch*", GitHub repository, (2017).

[18] *eSpeak: Speech Synthesizer*. http://espeak.sourceforge.net

[19] *GNU talk*. http://gnutalk.sourceforge.net