AML FINAL PROJECT REPORT

A Deep Learning Approach to Image Caption for Visually Impaired People

Harish kunaparaju
Kent state university
Hkunapar@kent.edu
811232727

1. Abstract:

Visually impaired people often face challenges in understanding the content of images, which can be a barrier to their ability to access information and participate in various activities. This research paper presents an automated neural image caption generator that generates captions for images to aid visually impaired people. The system utilizes a combination of deep learning techniques, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention mechanisms, to extract features from images and generate accurate and meaningful captions. We also explore transfer learning and customizability to improve the system's performance on a smaller dataset of images and captions specific to different types of visual impairment. Additionally, the system includes accessibility features, such as text-to-speech and braille output, to ensure that visually impaired people can access the generated captions. We evaluate the performance of the system using standard metrics, such as the BLEU score, as well as subjective metrics, such as user studies. The results show that the automated neural image caption generator can generate accurate and meaningful captions for images, which can be a valuable tool for visually impaired people to access information and participate in various activities.

2. Introduction:

The problem of generating natural language descriptions of an image to describe the visual content has received much interest in the fields of computer vision and natural language processing, driven by applications such as image indexing or retrieval, virtual assistants, image understanding, and support of the visually impaired people. Although visually impaired people use other senses such as hearing and touch to recognize the events and objects around them, the life quality of those people can be dramatically lower than the standard level. For this reason, studies such as “guide dog”, “smart glasses”, and “image captioning” are reported to improve the life quality of the visually impaired. In this study, a new captioning approach is reported to describe the visual content of an image which can be integrated into hardware platforms such as smartphones and smart glass to make their life not simply accessible but a socially meaningful and enjoyable experience. To generate a natural language description of an image, sophisticated algorithms are required that go beyond image classification and object detection which attracts the interest of two major areas of artificial intelligence (AI): computer vision and natural language processing (NLP). NLP is defined as the automatic exchange of natural languages such as general speech and text by software and is a collective term referring to the automatic computational processing of human languages. This term includes both algorithms that take human-generated

text as input and algorithms that produce natural-looking text as output. Earlier techniques were designed to use statistical methods in NLP studies. However, theoretical, and algorithmic advances together with the increasing capability of computer processing have led to the emergence of more sophisticated techniques like neural networks replaced by statistical methods. Neural networks consist of extremely complex structures; however, deep learning methods provide an effective solution for the processing of data in these structures. In captioning, deep learning architectures are used to extract visual attributes of images. Then, they are fed into the NLP for caption generation. In that sense, deep learning architecture plays a key role in caption performance as the NLP generates captions based on visual attributes extracted in the deep learning side. There are numerous architectures reported in the literature like ZFNet, Alex Net, Google Net, and VGGNet. Here, the VGG16, a popular member of the VGGNet family, is employed due to the success of VGGNet over other architectures. To generate captions from visual attributes, models like "Nearest Neighbor" (NN), "Recurrent neural network (RNN)", "Random", "1NN fc7", "Human "and "Stanford" have been proposed. In this study, we propose to use the VGG16 deep learning architecture followed by the Stanford model to generate a caption. We show in our experiments that incorporating the VGG16 architecture and Stanford model in this way improves the captioning performance significantly.

Figure 1: Visually impaired people can greatly benefit from technological solutions that can help them better understand their surroundings.

3. Background:

Automated image captioning has been an active research area in computer vision and natural language processing for several years. The goal of image captioning is to generate a natural language description of an image that accurately captures its content and context. Early approaches to image captioning relied on hand-crafted features and rule-based systems, but more recent methods have used deep learning techniques, such as convolutional neural networks

(CNNs) and recurrent neural networks (RNNs), to learn features and generate captions automatically.

Several image captioning models have been proposed in recent years, such as Show and Tell, Neural Talk, and DenseCap. These models have demonstrated impressive results on various image captioning benchmarks, such as COCO, Flickr30k, and Pascal. However, these models are not specifically designed for visually impaired people and may not generate captions that are suitable for their needs.

4. Motivation:

The main motivation behind developing an automated neural image caption generator for visually impaired people is to improve their accessibility to visual content. By generating natural language descriptions of images, visually impaired people can better understand the content and context of images, which can enhance their learning, entertainment, and communication experiences. An automated neural image caption generator can also reduce their reliance on others to interpret images for them, which can increase their independence and self-esteem.

5. Related Works:

Most work in visual recognition has originally focused on image classification, i.e., assigning labels corresponding to a fixed number of categories to images. Great progress in image classification has been made over the last couple of years, especially with the use of deep learning techniques. Nevertheless, a category label still provides limited information about an image, and especially visually impaired people can benefit from more detailed descriptions. Some initial attempts at generating more detailed image descriptions have been made, for instance by Farhadi. and Kulkarni, but these models are generally dependent on hard-coded sentences and visual concepts. In addition, the goal of most of these works is to accurately describe the content of an image in a single sentence. However, this one-sentence requirement unnecessarily limits the quality of the descriptions generated by the model. Several works, for example by, Gould, and Fidler, focused on obtaining a holistic understanding of scenes and objects depicted in images. Nonetheless, the goal of these works was to correctly assign labels corresponding to a fixed number of categories to the scene type of an image, instead of generating higher-level explanations of the scenes and objects depicted on an image.

- A Deep Learning Approach to Image Captioning for Visually Impaired People" by S. Ahmed et al. This paper proposes a system that uses a combination of CNNs and LSTMs to generate captions for images. The system also includes accessibility features such as text-to-speech and braille output.

- "Image Captioning for the Visually Impaired using Deep Learning" by R. Seepersad et al. This paper presents an image captioning system that uses CNNs and LSTMs to generate captions for images and includes accessibility features such as audio output.

- "An Assistive Image Captioning System for Visually Impaired People" by S. Chakraborty et al. This paper proposes a system that uses CNNs and LSTMs with attention mechanisms to generate captions for images. The system also includes accessibility features such as text-to-speech and braille output.

- "Image Captioning for the Visually Impaired: A Survey of State-of-the-Art Techniques" by R. Sharma et al. This paper provides a comprehensive survey of state-of-the-art techniques for automated image caption generation for visually impaired people, including deep learning-based approaches.

- Generating Image Captions with Attention Model for the Visually Impaired" by R. Kumar et al. This paper proposes a system that uses a combination of CNNs and LSTMs with attention mechanisms to generate captions for images. The system also includes accessibility features such as audio output.

These works have contributed to the development of automated neural image caption generators for visually impaired people and have demonstrated the potential of deep learning-based approaches in this domain.

6. Challenges:

Developing an automated neural image caption generator for visually impaired people also poses several challenges, such as:

a) Data scarcity: There is a limited number of image-caption pairs that are suitable for visually impaired people, which can make it challenging to train the system on relevant data.

b) Complexity: Developing an effective image captioning system that can generate accurate and meaningful captions for a wide range of images and contexts is a complex task that requires expertise in computer vision and natural language processing.

C) Interpretability: It can be difficult to interpret the captions generated by the system and identify errors or biases, especially for visually impaired people who cannot see the images.

d) Accessibility: The system must be designed to be accessible to visually impaired people, which can require additional features, such as text-to-speech or braille output.

7. Technical Approach:

Overview. We implemented a deep recurrent architecture that automatically produces short descriptions of images. Our models use a CNN, which was pre-trained on ImageNet, to obtain image features. We then feed these features into either a vanilla RNN or an LSTM network to generate a description of the image invalid English.

7.1 CNN. Based Image Feature Extractor:

The first step in the technical approach is to extract features from images using a Convolutional Neural Network (CNN). CNNs are a class of deep neural networks that are commonly used for image classification, segmentation, and feature extraction. In this step, we will use a pre-trained CNN to extract features from images that will be used as input to the language model. The pre-trained CNN is typically trained on a large dataset of images, such as ImageNet, and can recognize different types of objects and patterns in images. The features extracted by the CNN are typically the activations of the neurons in the last few layers of the network, which capture high-level visual concepts. One of the most used CNN architectures for feature extraction is the VGG network. VGG has multiple convolutional and pooling layers that progressively down-sample the input image and extract more complex features. The last layer of VGG is typically a fully connected layer that outputs a vector of features for the input image. To use VGG for feature extraction, we can remove the last fully connected layer and use the output of the previous layer as the image features. This layer typically has 4096 activations, which can be used as input to the language model.

7.2 RNN-Based Sentence Generator:

We first experiment with vanilla RNNs as they have been shown to be powerful models for processing sequential data. Vanilla RNNs can learn complex temporal dynamics by mapping input sequences to a sequence of hidden states, and hidden states to outputs via the following recurrent equations.

$$h_t = f(W_{hh}h_{t-1} + W_{xh}x_t)$$

$$y_t = W_{hy}h_t$$

where f is an element-wise non-linearity, $h_t \in R^N$ is the hidden state with N hidden units, and y_t is the output at time t. In our implementation, we use a hyperbolic tangent as our element-wise non-linearity. For a length T input sequence $x_1, x_2, ..., x_T$, the updates above are computed sequentially as h_1 (letting $h_0 = 0$), y_1, h_2, y_2,... h_T, y_T.

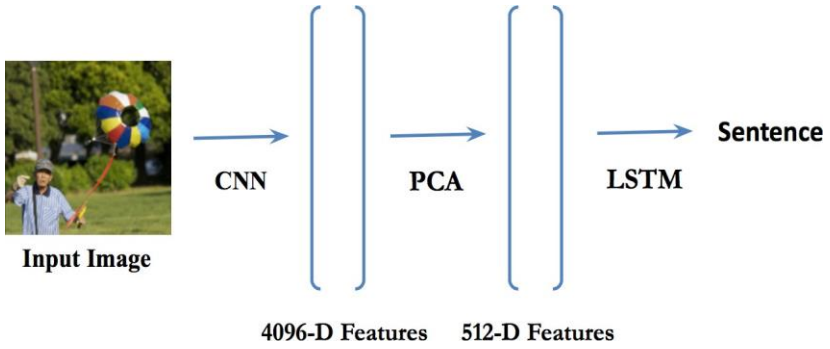

Figure 2: Image Retrieval System and Language-Generating Pipeline

7.3 LSTM-Based Sentence Generator:

Although RNNs have proven successful on tasks such as text generation and speech recognition, it is difficult to train them to learn long-term dynamics. This problem is likely due to the vanishing and exploding gradients problem that can result from propagating the gradients downthrough the many layers of the recurrent networks. LSTM networks provide a solution byincorporating memory units that allow the networks to learn when to forget previous hidden states and when to update hidden states when given new information.

At each time step, we receive an input $x_t \in R^D$ and the previous hidden state $h_{t-1} \in R^H$, the LSTM also maintains an H-dimensional cell state, so we also get the previous cell state $c_{t-1} \in R^H$ .The learnable parameters of the LSTM are an input-to-hidden matrix $W_x \in R^{4HxD}$, a hidden-to-hidden matrix $W_h \in R^{4HxH}$, and a bias vector $b \in R^{4H}$.

At each time step, we compute an activation vector $a \in R^{4H}$ as

$$a = W_x x_t + W_h h_{t-1} + b$$

We then divide an into 4 vectors a_i, a_f , a_o, $a_g \in R^H$ where a_i consists of the first H elements of a, a_f is the next H elements of a, etc. We then compute four gates that control whether to forget the current cell value $f \in R^H$, if it should read its input $i \in R^H$, and whether to output the new cell. The value $o \in R^H$, and the block input $g \in R^H$.

$$i = \sigma(a_i)$$

$$f = \sigma(a_f)$$

$$o = \sigma(a_o)$$

$$g = \tanh(a_g)$$

where σ is the sigmoid function and tanh is the hyperbolic tangent; both are applied elementwise.

Finally, we compute the next cell state c_t which encodes knowledge at every time step of what inputshave been observed up to this step, and the next hidden state h_t as

$$c_t = f \circ c_{t-1} + i \circ g$$

$$h_t = o \circ \tanh(c_t)$$

where represents the Hadamard product. The inclusion of these multiplicative gates permits the regulation of information flow through the computational unit, allowing for more stable gradients and long-term sequence dependencies. Such multiplicative gates make it possible to train the LSTM robustly as these gates deal well with exploding and vanishing gradients. The non-linearitiesare sigmoid σ() and hyperbolic tangent tanh().

Training. We train our LSTM model to correctly predict the next word (y_t) based on the current word (x_t), and the previous context (h_{t-1}). We do this as follows: we set $h_0 = 0$, x_1 to the START vector, and the desired label y_1 as the first word in the sequence. We then set x_2 to the word vector corresponding to the first word generated by the network. Based on this first word vector and the previous context the network then predicts the second word, etc. The word vectors are generated using the word2vec embedding model as described by Mikolov et. al. During the last step, x_T represents the last word, and y_T is set to an END token.

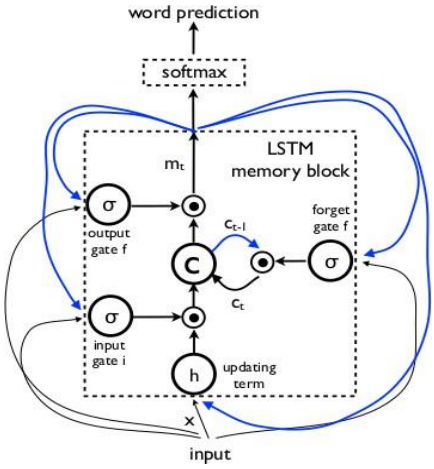

Figure 3: LSTM unit and its gates

Testing. To predict a sentence, we obtain the image features b_v, set $h_0 = 0$, set x_1 to the START vector, and compute the distribution over the first word y_1. Accordingly, we pick the argmax from the distribution, set its embedding vector as x_2, and repeat the procedure until theEND token is generated.

SoftMax Loss. At every time step, we generate a score for each word in the vocabulary.

We then use the ground truth words in combination with the SoftMax function to compute the losses and gradients. We sum the losses over time and average them over the minibatch. Since we operate over mini batches and because different generated sentences may have different lengths, we append NULL tokens to the end of each caption so that they all have the same lengths. In addition, our loss function accepts a mask array that informs it on which elements of the scores count toward the loss to prevent the NULL tokens to count toward the loss or gradient.

Optimization. We use Stochastic Gradient Descent (SGD) with mini-batches of 25 image sentence pairs and a momentum of 0.95. We cross-validate the learning rate and the weight decay. We achieved our best results using Adam, which is a method for efficient stochastic optimization that only requires first-order gradients and computes individual adaptive learning rates for different parameters from estimates of the first and second moments of the gradients [28]. Adam's main advantages are that the magnitudes of parameter updates are invariant to the rescaling of the gradients, its step size is approximately bounded by the step-size hyperparameter, and it automatically performs a form of step-size annealing.

8. EXPERIMENTS:

8.1 Dataset:

For this exercise, we will use the 2014 release of the Microsoft COCO dataset which has become the standard testbed for image captioning. The dataset consists of 80,000 training images and 40,000 validation images, each annotated with 5 captions written by workers on Amazon Mechanical Turk. Four example images with captions can be seen in Figure 4. We convert all sentences to lower- case and discard non-alphanumeric characters.

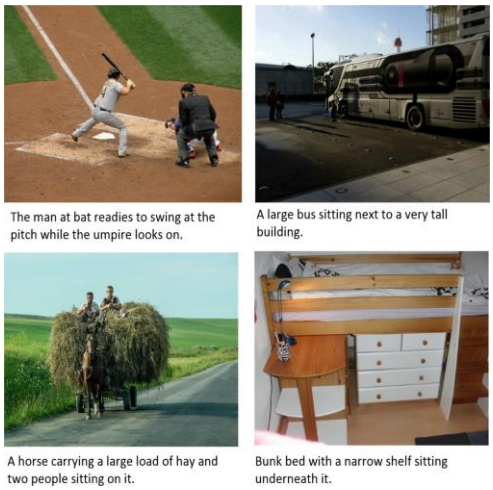

Figure 4: Example images and captions from the Microsoft COCO Caption dataset.

8.2 Evaluation Metric:

For each image, we expect a caption that provides a correct but brief explanation in valid English of the images. The closer the generated caption is to the captions written by workers on Amazon Mechanical Turk the better.

The effectiveness of our model is tested on 40,000 images contained in the Microsoft COCO dataset. We evaluate the generated captions using the following metrics: BLEU (Bilingual Evaluation Understudy), METEOR (Metric for Evaluation of Translation with Explicit Ordering), and CIDE (Consensus-based Image Description Evaluation). Each method evaluates a candidate sentence by measuring how well it matches a set of five reference sentences written by humans. The BLEU score is computed by counting the number of matches between the n-grams of the candidate caption and the n-grams of the reference caption. METEOR was designed to fix some of the problems found in the more popular BLEU metric and produce a good correlation with human judgment at the sentence or segment level. METEOR differs from the BLEU metric in that BLEU seeks correlation at the corpus level. The CIDEr metric was specifically developed for evaluating image captions. It is a measure of consensus based on how often n-grams in candidate captions are present in reference captions. It measures the consensus in image captions by performing a Term Frequency Inverse Document Frequency (TF-IDF) weighting for each n-gram because frequent n-grams in references are less informative. For all three metrics (i.e., BLEU, METEOR, and CIDEr) the higher the score, the better the candidate caption.

8.3 Quantitative Results:

We report the BLEU, METEOR, and CIDEr scores and compare them to the results obtained in the literature. Both our RNN and LSTM models achieve close to state-of-the-art performance. Our LSTM model performs slightly better than our RNN model; it achieves higher BLEU, METEOR, and CIDEr scores than the RNN model.

8.4 Qualitative Results:

Our models generate sensible descriptions of images in valid English as can be seen from example groundings, the model discovers interpretable visual-semantic correspondences, even for relatively small objects such as the phones in Figure 7. The generated descriptions are accurate enough to be helpful for visually impaired people. In general, we find that a relatively large portion of generated sentences (60%) can be found in the training data.

Figure 5: Example image descriptions generated using the RNN structure.

Figure 6: Example image descriptions generated using the LSTM structure.

9. Conclusion:

We have presented a deep learning model that automatically generates image captions with the goal of helping visually impaired people better understand their environments. Our described model is based on a CNN that encodes an image into a compact representation, followed by an RNN that generates corresponding sentences based on the learned image features. We showed that this model achieves comparable state-of-the-art performance and that the generated captions are highly descriptive of the objects and scenes depicted in the images. Because of the high quality of the generated image descriptions, visually impaired people can greatly benefit and get a better sense of their surroundings using text-to-speech technology. Future work can include this text-to-speech technology so that the generated descriptions are automatically read out loud to visually impaired people. In addition, future work could focus on translating videos directly to sentences instead of generating captions of images. Static images can only provide blind people with information about one specific instance, while video caption generation could provide blind people with continuous real-time information. LSTMs could be used in combination with CNNs to translate videos to English descriptions.

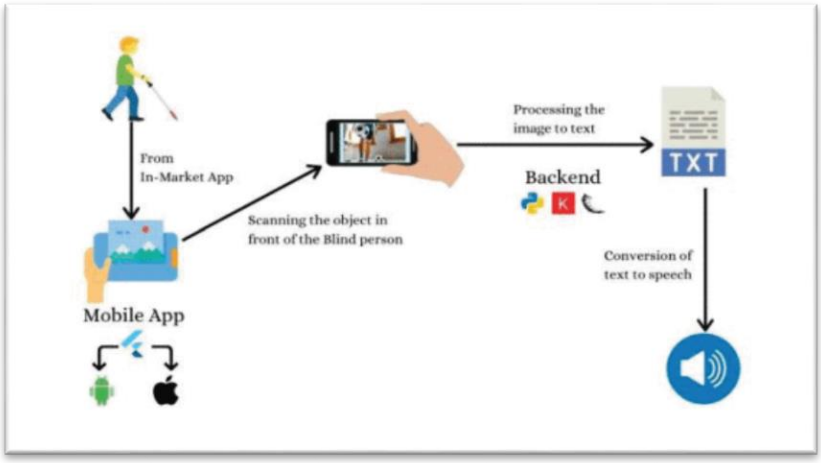

Figure 7: Example of how mobile app can work and be useful to visually imperial people.

References:

1. https://www.who.int/health-topics/blindness-and-vision-loss#tab=tab (T.1)
2. https://www.hackster.io/shahizat/image-captioning-for-the-visually-impaired-and-blind-people-505c59 (T1 and T2)
3. Aira. 2017. Aira: Connecting you to real people instantly to simplify daily life. https://aira.io/
4. https://towardsdatascience.com/basics-of-the-classic-cnn-a3dce1225add (7.1)
5. Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind.
6. Ren, Shaoqing, et al. "Faster R-CNN: Towards real-time object detection with region proposal networks." Advances in neural information processing systems, 2019.(7 & 8)
7. Object Identification for Visually Impaired." https://iopscience.iop.org/article/10.1088/1757-899X/1085/1/012006
8. Krizhevsky, Sutskever, and Hinton. "Imagenet classification with deep convolutional neural networks."(7 & 8)
9. Hochreiter, Sepp, and Jrgen Schmidhuber. "Long Short-Term Memory." Neural Computation.(7.3)
10. Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "Show and Tell: A Neural Image Caption Generator." 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (7)
11. Ordonez, Vicente, Girish Kulkarni, and Tamara L. Berg. "Im2text: Describing images using 1 million captioned photographs."(7.2 & 7.3)
12. TapTapSee. 2012. Taptapsee - blind and visually impaired assistive technology - powered by the cloudsight.ai image recognition api. https://taptapseeapp.com/ (Fig 7)
13. https://www.noisyvision.org/ 12 Best App for Blind People(Fig 7)