A Sight for Sore Eyes: Automated Image Captioning for the Visually Impaired

Victoria Beckeman
University of Michigan
1221 Beal Ave, Ann Arbor, MI 48109
vgbeck@umich.edu

Elisa Simon
University of Michigan
1221 Beal Ave, Ann Arbor, MI 48109
elisasmn@umich.edu

Eray Sabuncu
University of Michigan
1221 Beal Ave, Ann Arbor, MI 48109
esabuncu@umich.edu

Erika Webb
University of Michigan
1221 Beal Ave, Ann Arbor, MI 48109
eriwebb@umich.edu

Introduction

For our project, we chose to create a model that creates automated image captioning. The main motivation for our project is two-fold: 1. Improve accessibility of the products in our society today with better assistive technologies, and 2. Increase the searchability of items by expanding the tags on documents to also include the item's visual components.

Accessibility

For accessibility, image captioning enhances accessibility for individuals with visual impairments by providing textual descriptions of visual content that can then be translated to audio by text-to-speech technologies. This is incredibly important for improving the accessibility of websites, textbooks, and social media, which depend heavily on visuals.

Content Searchability and Retrieval

Currently, most information is retrievable by a text-dependent search function, but a lot of items (such as photos) don't include text and are therefore harder to locate. Automated image captioning can create tags for visuals, which helps with search engine optimization and personal item retrieval as these tags can enable efficient searching and categorization of images based on their content, making it easier to organize and retrieve relevant visual information.

Our Project's Contributions

Our specific contributions to our project are to recreate the "Show And Tell" paper (described in more detail in the next section) to the best of our ability. We attempted to experiment with our model's effectiveness by incorporating various modifications to the current such as changing the activation function and dropout rate. With our final design, we were able to determine that the ReLU activation function with a higher dropout rate has the highest effectiveness based on the variations we tested. This model we recreated can define captions for visuals relatively accurately, however, it can be improved with more advanced techniques to implement it into current technologies. Our captions were far from perfect and included junk data we were forced to truncate.

Related Work

With so many use cases, the popularity of image captioning has grown exponentially within computer vision. The "Show and Tell" paper significantly advanced the field, and since its publication in 2015, numerous works have further contributed to the realm of image captioning.

The model used in "Show and Tell" is an end-to-end neural network that takes an image as input and generates a natural language description. The architecture combines

convolutional neural networks (CNNs) for image feature extraction and long short-term memory (LSTM) networks for sequence generation. The datasets used for training and testing are Pascal VOC, Flickr 8k, Flickr 30k, MSCOCO, and SBU. In the encoding phase, a pre-trained CNN captures relevant input image features and turns the image into a fixed-size feature vector. In the decoding phase, the captured image features are given to a recurrent neural network (RNN) with LSTM networks. Then a sequence of words that forms the image description is generated by the RNN, while the LSTM, conditioned on the image features, learns to predict subsequent words in the description based on the context of prior words. For training, maximum likelihood estimation (MLE) and cross-entropy loss are utilized to maximize the likelihood of generating a true caption for the input image. During inference, beam search explores different word sequences to improve the contextual relevance of the image description. Lastly, qualitative and quantitative (ranking metrics, BLEU model) evaluation methods are used to measure the accuracy of the generated captions (Vinyals et al).

The encoder-decoder framework, specifically the usage of CNN as encoder and LSTM as decoder (as demonstrated in "Show and Tell") has experienced tremendous expansion and advancement over the last decade. This framework excels at creating contextually relevant natural language descriptions for images. The use of LSTM allows for an understanding of sequential dependencies and the generation of captions that are semantically and syntactically coherent. The framework is also flexible and generalizes well to a variety of images. However, it is not without limitations, particularly in terms of time and space complexity efficiency. Additionally, its performance is significantly reduced when dealing with complex images due to a lack of semantic understanding of the information and objects depicted in the images (Ahmad et al).

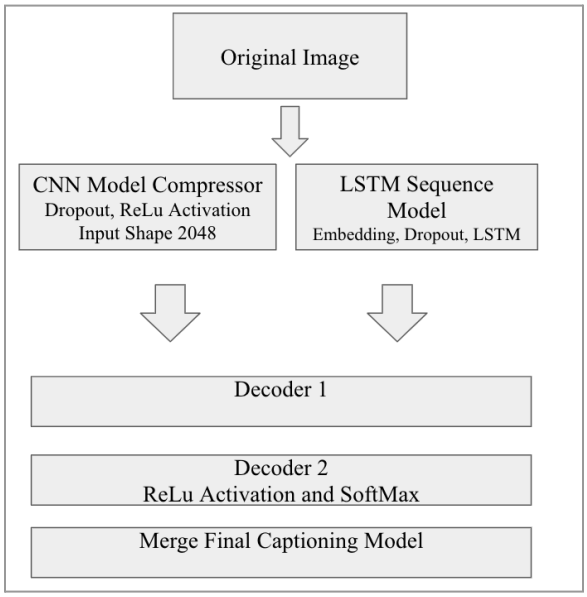
Our project drew inspiration from the framework in "Show and Tell" and referenced other websites and previous GitHubs to create an image captioning model similar to the one described in "Show and Tell".

Method

Our model was trained on the Flickr8k dataset. The specific training subset contained 6000 images. All of the images and captions had to be processed through a variety of helper functions and the descriptions had to be cleaned (remove punctuation, convert to lowercase, remove hanging " 's " etc). The encoder was built using a pre-trained Xception model to extract features and then shaped with different parameter choices in our define_model() function to encode and decode. This function begins by setting up two input layers, one for the image features and another for the caption word sequence. Dropout is applied to the image features for regularization and a fully connected layer reduces the dimensionality to 256 nodes with ReLu activation. This is processed with an embedding layer, another dropout, and an LSTM layer with 256 units. The outputs of the image and sequence models are merged followed by another dense layer with ReLu activation. The final output layer predicts the next word in the sequence using softmax activation. This is compiled with categorical cross-entropy loss and the Adam optimizer. Our baseline is built with these mentioned loss activations and a dropout rate of 0.5. We chose to experiment with these hyperparameters to see the effects on the output and loss rates. Some of the shortcomings of our model include overfitting. With more time and research we could attempt early stopping or freezing layers to prevent the data from being overfit. Another shortcoming was the fact that we trained on a smaller data set than the paper

that we were modeling, this is due to personal computing power limitations and extended runtime. We wanted to be able to test many ideas and our set-up was not cohesive with a larger extended dataset. In addition, if the "correct" captions that we calculated our BLEU scores with had a wider range we might see a difference in results.

Figure 1. CNN Model



Experiments

We experimented with our model on images handpicked from the remaining 2000 pictures from the Flickr8k dataset it was not trained on. We ran an ablation study on our model three different times with a different combination of activation functions and dropout rates. One variation used the Rectified Linear Unit (ReLU) activation functions in the input layers with a dropout rate of 0.5 and the other two variations used the Exponential Linear Unit (ELU) activation function with a dropout rate of 0.2 and then again with 0.5. Most sources highlighted that the ReLU activation function would be more effective than the ELU. The ReLU function is more computationally efficient, can enable more effective learning of the image features in convolutional layers with better gradient propagation during backpropagation, and its sparse properties can be beneficial for representing and learning more distinctive features in the images and better focus on the most relevant parts of the image. However, the ELU's inclusion of negative values can avoid dead neurons during training, facilitate training optimization because it is comparably smoother than ReLU, and could be more advantageous for capturing various features and nuances in images with its flexibility. We also decided on dropout rates of 0.2 and 0.5 to experiment with finding a balance between preventing overfitting by introducing randomness during training while still allowing the model to retain more information from the data.

Table 1. Results

			
ReLU (0.5) *(loss = 3.0676)*	Dog in grass with stick in mouth and toy *bleu: 0.490*	Man riding surfboard *bleu: 0.474*	Standing on snowy mountain with sun *bleu: 0.595*
ELU (0.5) *(loss = 2.6727)*	Brown dog running through the grass carrying a stick in mouth *bleu: 0.574*	Red & white surfboard in water while man in red and yellow surfboard in the background by wave *bleu: 0.00*	Climber without up snapping at the peak of snowcapped peak looking at the peak *bleu: 0.209*
ELU (0.2) *(loss = 2.1827)*	Yellow dogs are running through field of yellow flowers *bleu: 0.478*	Beagle catching his legs into the water *bleu: 0.116*	Skier covered mountain looking left *bleu: 0.297*

We maintain 10 epochs for each training session, the same softmax activation function for the output layer to normalize all our output, and the same loss function of categorical cross entropy to properly encapsulate the wide variety of labeling within our dataset of images. The model's effectiveness is evaluated with a BLEU (Bilingual Evaluation Understudy) score which is a metric used to evaluate the quality of machine-generated text by comparing N-tuples of words. Similar to the "Show and Tell" paper, we calculated a BLEU score for each generated caption compared to a user-generated caption. The score is between 0 and 1 with the caption being more accurate the closer it is to 1.
Table 1 highlights a few examples of images and their generated captions with corresponding BLEU scores. The model with a ReLU function and dropout rate of 0.5 outperformed the other two models with an ELU activation function by consistently generating captions with higher BLEU scores. Even though it had the highest loss score, we infer that the ReLU function did better capture the more distinct features of the images and reduced overfitting with the higher dropout rate. This allowed it to better perform in images it was not trained on. Our results are not directly comparable to the "Show and Tell" paper because the authors used the BLEU-4 metric which is a more advanced variation of our simple BLEU metric.

Conclusions

Our project addresses the dual motivation of improving accessibility for the visually impaired and enhancing visual content searchability. We contribute to a more inclusive digital landscape by developing an automated image captioning model. Inspired by the "Show and Tell" paper, our recreated model, combining convolutional neural networks (CNNs) and long short-term memory networks (LSTMs), proves effective in generating accurate captions for visuals. Our experiments, exploring various parameters and model modifications, demonstrate the model's robustness and potential for practical implementation.

Despite our challenges of overfitting, limited training data, and junk words in the outputted caption, our baseline model exhibits promising performance. We believe our current model needs some future refinement before being deployed into existing technologies. Some areas of further research would include increasing our dataset's size and image variety to improve our model's performance as well as to create early stopping mechanisms to reduce overfitting.

References

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," CV Foundation, https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Vinyals_Show_and_Tell_2015_CVPR_paper.pdf (accessed Dec. 11, 2023).

[2] S. Gupta, "Step by Step Guide to Build Image Caption Generator using Deep Learning," Analytics Vidhya, https://www.analyticsvidhya.com/blog/2021/12/step-by-step-guide-to-build-image-caption-generator-using-deep-learning/#h-testing-the-image-caption-generator-model (accessed Dec. 10, 2023).

[3] Yumi, "Develop an image captioning deep learning model using Flickr 8K data," Yumi's Blog, https://fairyonice.github.io/Develop_an_image_captioning_deep_learning_model_using_Flickr_8K_data.html (accessed Dec. 10, 2023).

[4] R. Ahmad, M. Azhar, and H. Sattar, "An Image captioning algorithm based on the Hybrid Deep Learning Technique (CNN+GRU)," Arxiv, https://arxiv.org/pdf/2301.02440.pdf. (accessed Dec. 13, 2023).