Image Caption Generator

Shree Murthy, Rahul Sura, Dylan Inafuku

2023-12-16

Table of Contents

1. Introduction

This report aims to discuss the model building and evaluation of an Image Caption Generator trained on the Flicker 8k Dataset. The goal of this model is to take in an image and generate a caption that describes the image. This model is built using an Encoder/Decoder structure. The motivation for this model is to develop a model that can help create “Alt Text” to help visually impaired people understand images on the internet. This model can also be used to help people who are learning a new language understand the meaning of an image.

1.1. Alt Text

Alt text is a short description of an image that is used by screen readers to describe images to visually impaired people. The goal is to make the internet more accessible and useable for everyone. HubSpot has a great article on how to write alt text for images. It explains that alt text should be short yet descriptive. The goal of alt text would be to help people access wide variety of websites without worrying about not being able to understand the images. Over 2.2 billion people have a vision impairment or blindness. This is a huge number of people who are not able to access the internet in the same way as people who are not visually impaired. Furthermore, users who use a screen reader report they are ~61% less likely to retain all the information presented by an image. With this information in mind, we sought to create a model that could be helpful generating captions for images to help people who are visually impaired understand images on the internet.

2. Analysis

3. Methods

4. Results

5. Reflection

When tackling the problem of image captioning, we first analyzed what we already knew. Throughout CPSC 393, we learned CNNs, LSTMs, and transfer learning models. However, we never really put all those ideas together (i.e Model Fusion) which was a necessary step for this project. The closest thing we learned to multiple models working together were transformers. With that initial knowledge, we sought to blend CNN and LSTMs together to create the image caption model. Also, we could use transfer learning to help with the mundane tasks like feature extractions. Once, we figured out this baseline we built an effective model that could generate captions for the images. Nevertheless, this approach still had some limitations. The dataset we used had over 8000 images; however, most of the pictures were of people or household animals. Furthermore, most of the people were white or had pale complexions. This could've easily poisoned our model training stage because it could assume that when it sees a person they are automatically white, or have a pale complexion. This is not the ideal when trying to create a model that can be used by everyone and for any type of pictures. Also, the pictures were mainly outdoors and people doing activities. This also limits the model's ability to learn because it may not learn what features are important to tell the interior of a house, school, workplace environment, etc. Our world has various activities and tasks that people and animals do. While, this model showcases the capabilities to train on an image dataset to generate somewhat accurate captions, it is not the best model to use for a wide variety of images.

To improve this model, it would be ideal to have two things: a training environment with access to a large number of GPUs. We trained this model on one NVIDA GPU, while it trained quickly we didn't have extra GPUs to handle a model that could be trained on a larger image dataset (maybe like 10,000 images with various features).

We hope this model creates a simple baseline for future researchers to use a simple, yet efficient, model design and tweak the hyperparameters and dataset information to create a more robust and effective model that can be used by all to generate alt text at an effective rate.

Overall, given our problem statement, i.e generate alt text, we created a sufficient model that can hold up to the task. However, we recognize there are limitations to this approach and hope that future researchers, as well as ourselves, can improve upon this model.

6. References

References are placed in order of appearance in the report.

1. HubSpot. HubSpot
2. Sharif, et. al. Understanding Screen-Reader Users' Experiences with Online Data Visualizations