Capstone - Eye for the Blind

UoA-MSDS C5 - Sep 2022 to Sep 2024
Author: Vu Truong

Agenda

Problem overview

Solution

Objective

Results

Process

Shortcoming and Challenges

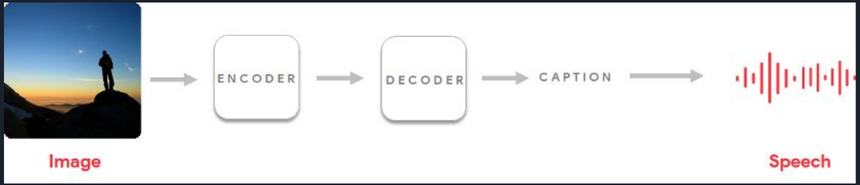
Appendix

Problem overview

Today, in the world of social media, millions of images are uploaded daily. Some of them are about your friends and family, while some of them are about nature and its beauty. Imagine a condition where you are not able to see and enjoy these images — a problem that blind people face on a daily basis. According to the World Health Organization (WHO), it has been reported that there are around 285 million visually impaired people worldwide and out of these 285 million, 39 million are totally blind.

In an initiative to help such people experience the beauty of the images, Facebook had earlier launched a unique feature that can help blind people operate their app on their respective mobile phones and the feature would explain (by speaking out) to the blind person the contents of an image that their friends have posted on Facebook.

Solution



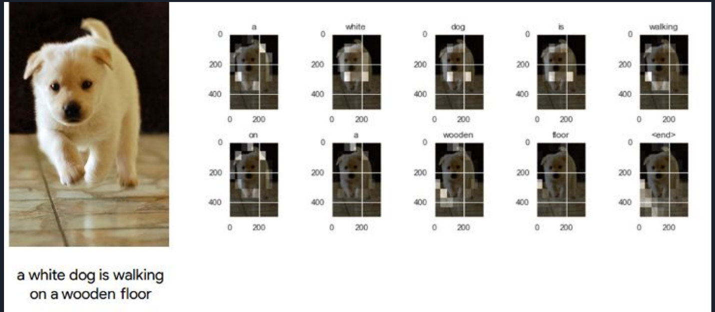
01 Make a machine learning model similar to what Facebook made, specifically making the blind person know the contents of an image in front of them

02 The model is a CNN-RNN based model. The model will convert the contents of an image and will give the output in the form of audio.

03 Inner implementation, we are converting the image to text description first and then using a simple text to speech API, the extracted text description/caption will be converted to audio.

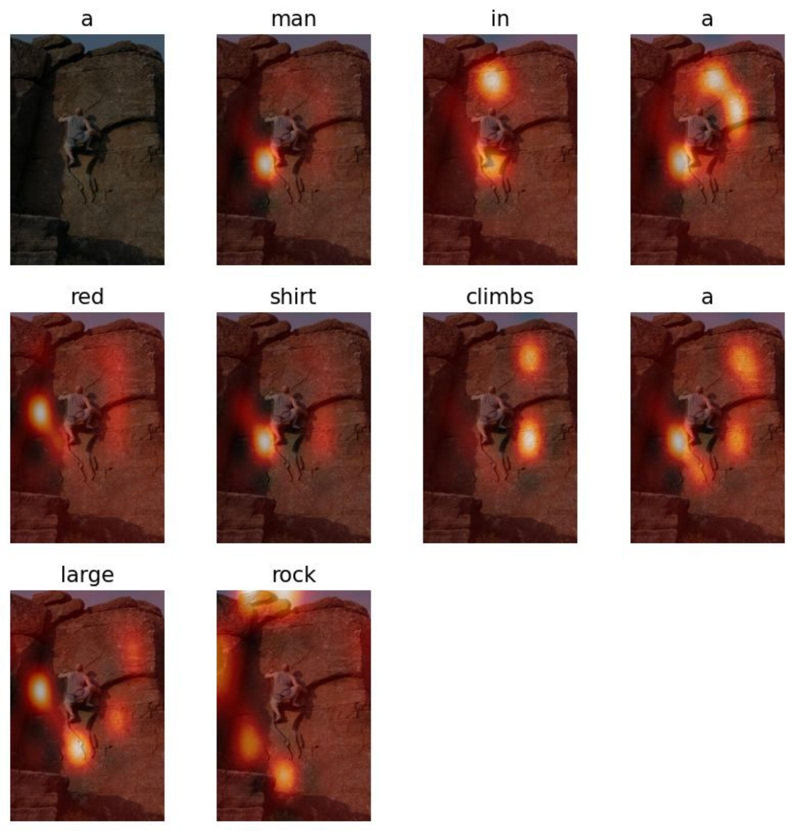
Objective

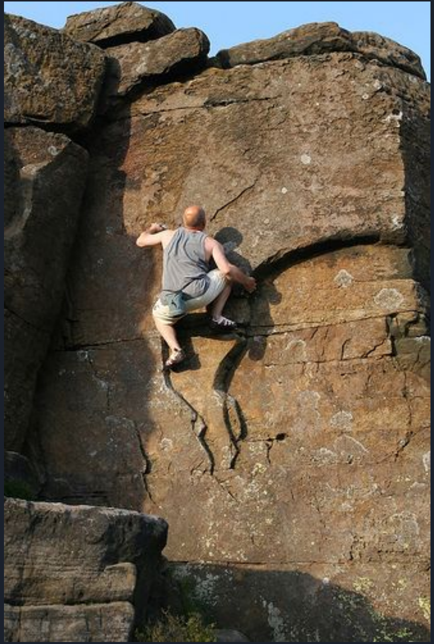
- Understandable prediction by visualizing how the model see the image and predict each word.
- Caption covers the main elements in the image and the context.



a white dog is walking on a wooden floor

Results

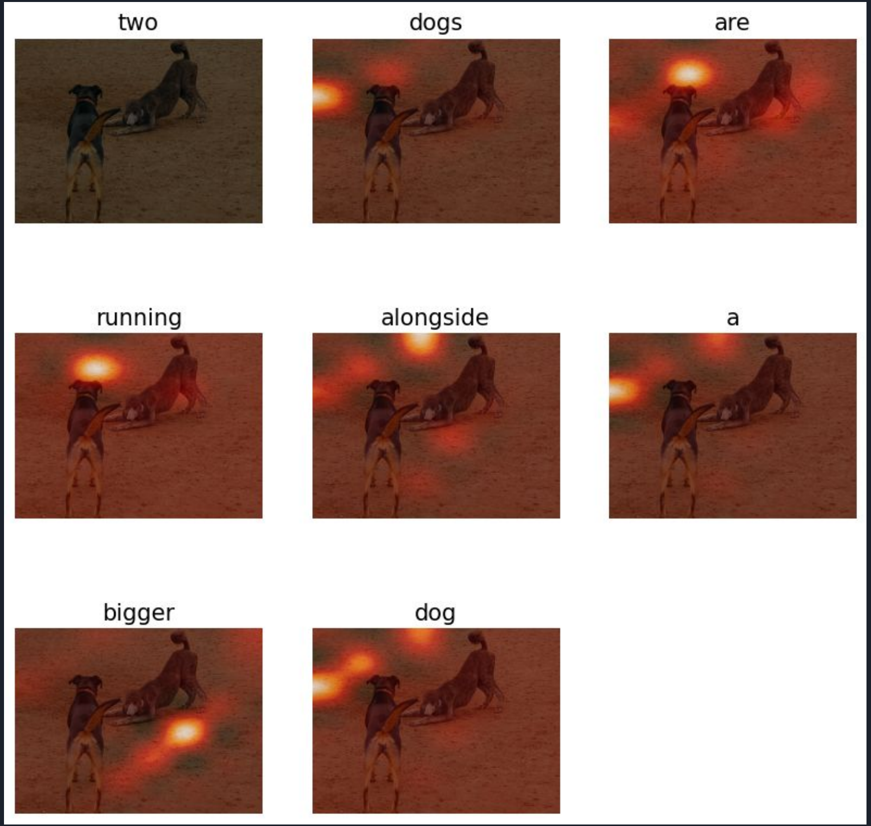




Real Caption: man climbing a rock wall

Prediction Caption: a man in a red shirt climbs a large rock

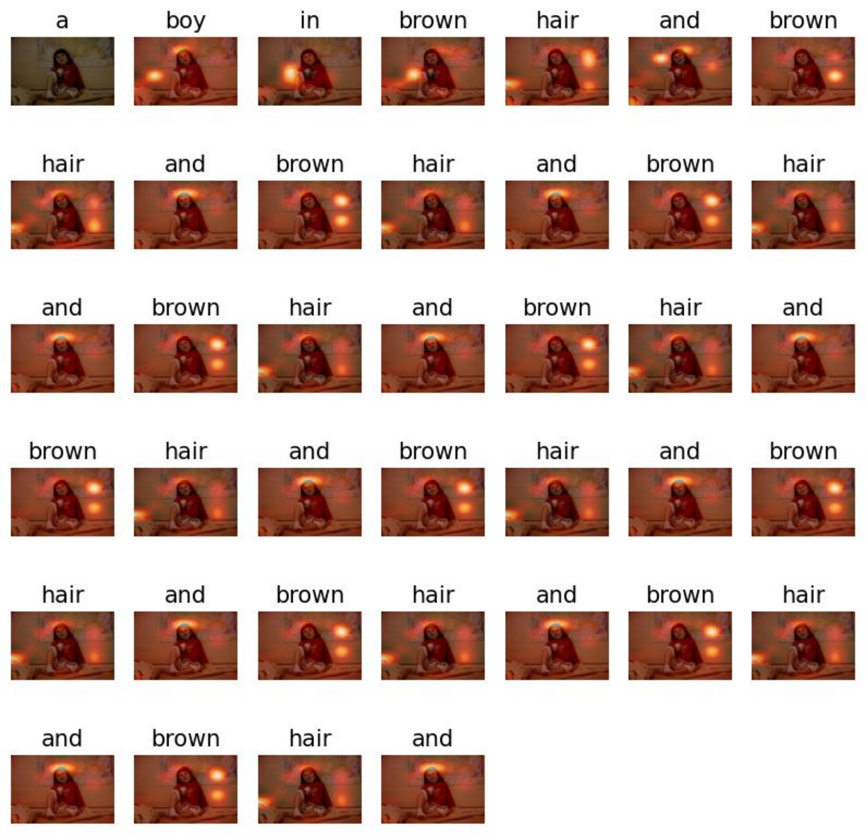
Results

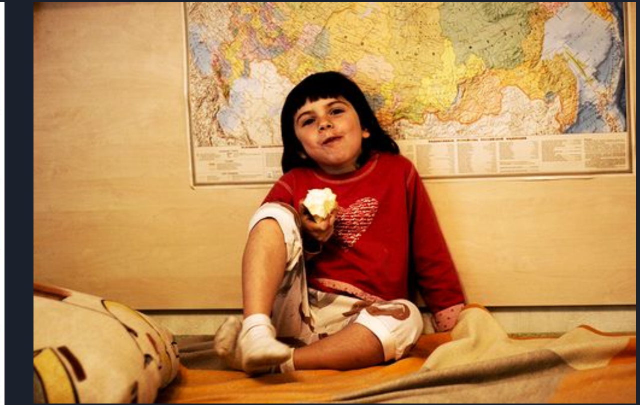


Real Caption: two dogs play with each other

Prediction Caption: two dogs are running alongside a bigger dog

Results





Real Caption: a smiling child sits against a wall on a blanket and eats a snack

Prediction Caption: a boy in brown hair and brown hair and brown hair and brown hair and brown hair and brown hair and brown hair and brown hair and brown hair and brown hair and brown hair and brown hair and

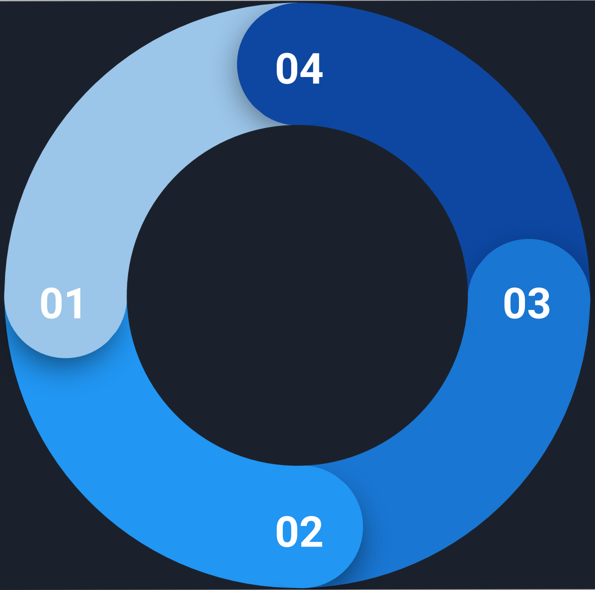
Process

Prepare dataset

From source flickr8k dataset, transform the images using imagenet to extract important features

Prototype

Train base CNN-RNN model with attention mechanism.



Finetune

Implement complete model and finetune on different parameters.

Evaluate

Plotting attention map over the input image and calculating BLEU score for generated caption

Shortcoming and Challenges

- Given the size of dataset (Flickr8k) and the variations of data inside, the model has not reached its potential as testing on unseen images shows signs of overfit.
- The model is rather complex and training time took very long on large number of epochs. Fine tuning even took longer.
- To speedup the training using distributed resources (multiple GPUs or TPUs), we must follow tensorflow best practices or switch to other frameworks like Pytorch. In the time constraint of this capstone and the resources provided by Kaggle, the optimization is not the best possible.
- We can enhance the model further just by stacking more layers and add Dropout + BatchNorm to the layers to achieve better result, but will take more time and require even more powerful training machines.

Appendix

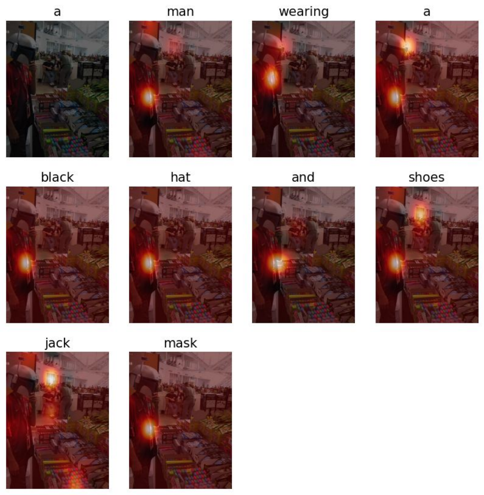
**Please see attached Jupyter notebook for more details on the implementation*

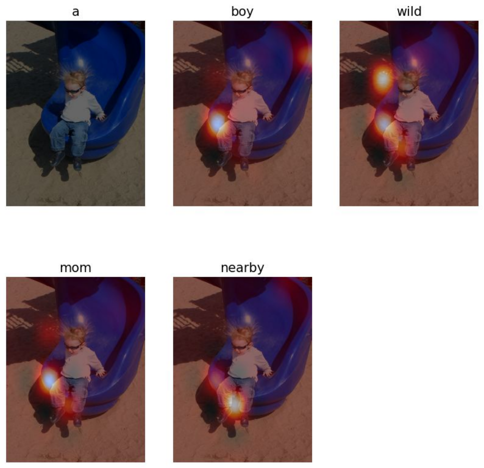
**Or the online backup on Kaggle using the link below*

https://www.kaggle.com/code/innoobwetrust/vu-truong-capstone-2-eye-for-the-blind?kernelSessionId=195337969

Thank you!

a young child wearing sunglasses is at the base of a blue slide





a young man wearing a large space helmet is looking at candy in a market