Image Caption Generator

Nusrat Nawshin
The George Washington University
Spring 2023 - Capstone Project

Introduction

People with weak eyesight or no eyesight find it challenging to stay aware of their surroundings. A system that provides a brief synopsis of the visuals around them would be very helpful to them and remove their accessibility constraints. With a goal of an improved assistive system, my caption generator is going to employ image processing and deep learning techniques to identify objects in the images, and then natural language processing (NLP) will be used for transforming that information into a relevant sentence that describes the entire image and the scenario.

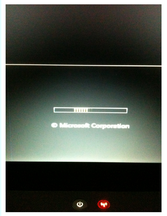
Dataset

VizWiz Dataset: Contains images captured by visually impaired people and each images are paired with 5 captions

Training Set: 23,431 images & 117,155 captions
Validation Set: 3,875 images & 96,876 captions
Test Set: 3,875 images & 96,876 captions

1. A red Hello Kitty button is laying on a table.
2. A red Hello Kitty toy sits on a brown wooden table.
3. An objected depicting a red cat with a white face and purple bow.
4. A red Hello Kitty box on a wood table.
5. A child's toy Hello Kitty is on a table.

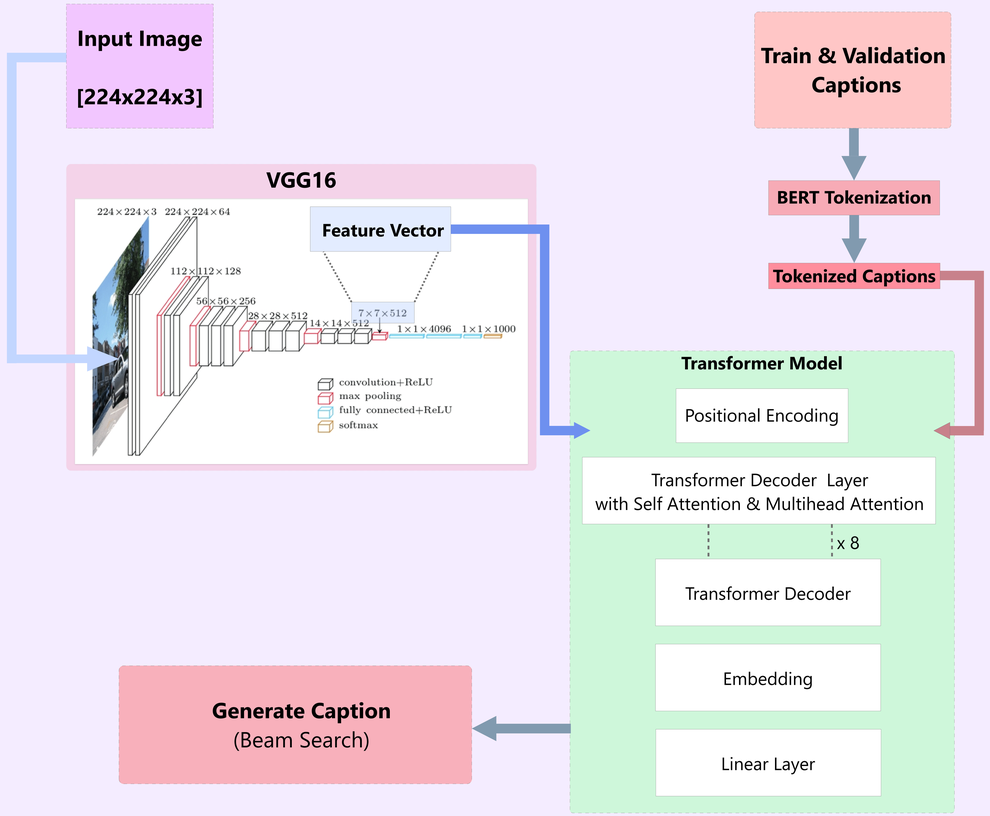
1. A pump bottle of body lotion is on the floor.
2. Bottle of itching medicinal lotion with a white and blue label.
3. A bottle of Sarna anti itch lotion is resting on green carpet.
4. A bottle of lotion has white and black writing.
5. A bottle of lotion is laying on a carpeted floor.



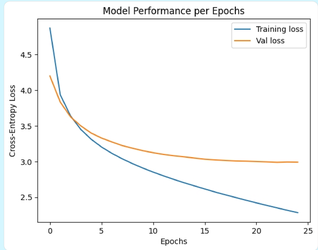
1. A computer display screen branded for Microsoft Corporation.
2. A computer screen is displaying a window with a message.
3. A black computer monitor start up screen showing Microsoft Corporation.
4. A computer display shows Microsoft on it's screen.
5. Laptop screen displaying a black screen with white text and a white loading bar.

1. The side of a brown glass tincture bottles in a man's hand.
2. A brown colored glass bottle featuring the text garden Rue.
3. A person is holding up a bottle of medicine.
4. A bottle of dietary supplement with a brand name ending in "Garden".
5. Small bottle of supplement by the brand garden.

Model



Evaluation Metrices & Results



Validation Set
ROUGE 2 F Measure: **0.27**
ROUGE L F Measure: **0.41**
BLEU: **0.0013**
Meteor: **0.0234**

Test Set
ROUGE 2 F Measure: **0.34**
ROUGE L F Measure: **0.44**
BLEU: **0.0012**
Meteor: **0.0235**

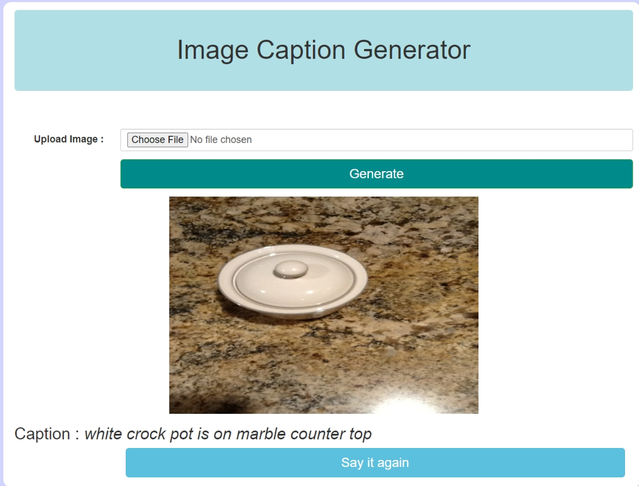
Evaluation Metrices:
Rouge 2: Bigram overlap between words
Rouge L: Longest Matching sequence
BLEU: Counting matching n-grams
Meteor: Unigram matching

Predicted Captions

tube of lotion is on top of table

can of food is on top of table

User Interface



Conclusion

In this work, I developed a multi-model neural network that consequently learns to generate captions form the content of images. At first I have extracted features from the images using VGG16's convolution layers output and the captions were tokenized using BERT tokenizer. After that these two inputs were passed in a Transformer model with 16 multi-head attention and 8 transformer decoder layers. Model was trained till 22nd epochs with Adam optimizer with learning rate 0.00001. Finally through beam search the model generates the captions. A scene graph or object detection mechanism may be added to the model to further improve it, however advanced techniques could not be implemented due to computational limitations.

Author

Nusrat Nawshin, Graduate Student, Columian College of Arts and Sciences, George Washington University (nawshinnusrat@gwu.edu)