Real Time Multimodal Classification for Social Media Notifications

Mertay Dayanc, Computer Science
Mentor: Yezhou Yang, Assistant Professor – Tejas Gokhale, Graduate Assistant/Associate
School of Computing, Informatics, and Decision System Engineering

Introduction

Main goal of this research is to create a multimodal classification model for classification of social media notifications. Although image input gives many information about the context of the notification, textual input can also tell various content. These two inputs together can tell much more about the content of the notification that user received.

Before starting anything else, literature review was made to find out what has been done so far relating to the problem that this research aims to solve. There were two beneficial papers to investigate. Which was proposing early, late and common space fusion which gave the idea on how to approach this problem. First researcher attempted to implement the algorithm that was explained on the paper.

Dataset

Initially researcher used the dataset, which was used on the research that was investigated on literature review. Which consists of 3193 text and labels. Also, there were 1054 image and text input with their labels {creepy, gore, happy, rage} which are collected from Reddit. Besides these, in order to implement image captioning model, researcher have used MS_COCO2014 dataset which was consisting of 12.3GB.

Model

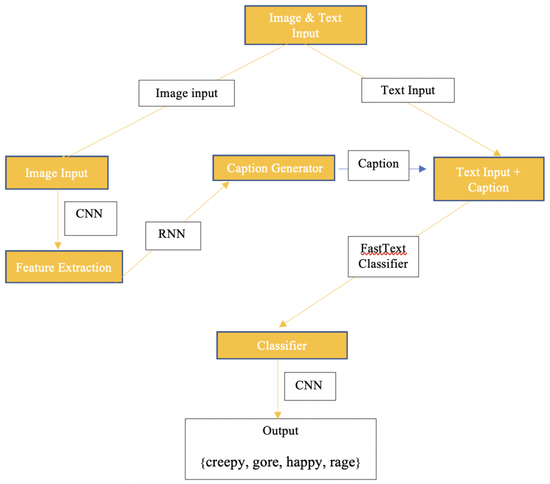

Figure 1. Architecture of the model, generate the image caption if it exists and then append that result into text classifier input if it exists.

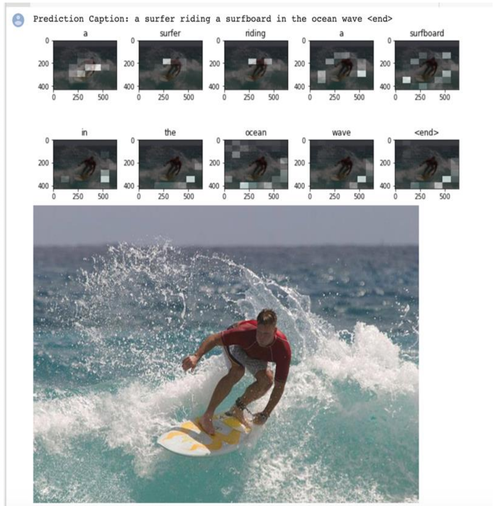

Figure 2. Example caption

After having all the data that was needed researcher focused on the model. First, text classifier is implemented using 2554 texts and labels leveraging the FastText text classifier developed by Facebook. Then researcher used 639 text and labels to validate the text classifier. Due to the limited number of test and validation data this model was only performing at the accuracy **%67.61**. Which wasn't really good accuracy. After implementation of the Text classifier, researcher started working on the Image captioning model. Researcher found a useful article on TensorFlow's website to implement a caption generator. In that large MS_COCO2014 dataset researcher only used 10,000 images for training. However, the results were satisfying.

Results

Model	Accuracy				
	Creepy	Gore	Happy	Rage	Overall
Text-based Classifier	**%71.8**	**%78.87**	%81.86	%65	**%74.28**
Image-based Classifier	%45.11	%38.03	%26.98	%18.57	%32.17
Main Classifier	%62.41	%68.38	**%82.79**	**%66.43**	%70

Figure 3. Results

Discussion

Researcher observed that Text and image together performs well on 2 classes however overall text-based performs better using this model. Possible problem might be due to computational limitations on training the Image Captioning model using 10k images from MS_COCO. Also, those pictures were completely different from the test dataset. Because, dataset was collected from reddit, and they had classes like Creepy and Gore which has disgusting pictures. Whereas, MSCOCO has more daily life like pictures. Which possibly impacted the performance on Creepy and Gore Test. However, Text and image together performs better on test where we had more daily life like pictures {happy, rage}. Also, this model performed poorly on tasks which only had images. Hence, this should be improved maybe by pooling the features of texts and images then training the model with those common features.

Acknowledgments: Thank you to Dr. Yezhou Yang and Tejas Gokhale for their mentorship. This project is funded by FURI department.

FURI

Ira A. Fulton Schools of Engineering
Arizona State University