Big Data Analytics: Hate Detection in Memes

Amisha Jindal, Edward Carlson, Pascal Bakker

Introduction

Identification of hate in memes poses a challenge because the text or image in isolation may not be offensive on their own, but the combination of the image and text may make it offensive. Facebook AI has created the Hateful Memes dataset to aid researchers in understanding multimodal hate speech and build better systems.

Figure 1: Mean memes (left), benign image confounders (middle) and benign text confounders (right), ***for illustrative purposes only.***

Goal: To classify offensive and hateful memes. The challenge focuses on fine-tuning existing pre-trained large scale multimodal models.

Dataset

The dataset is 4GB and contains about 10,000 memes.

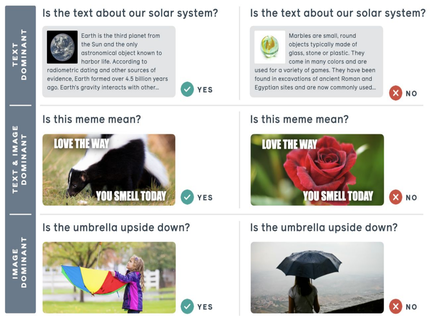
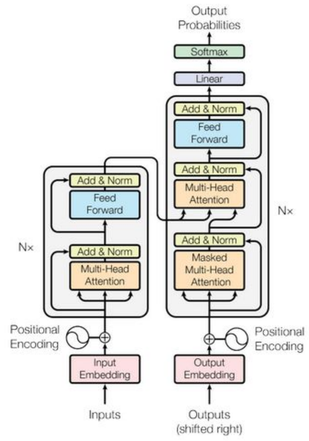


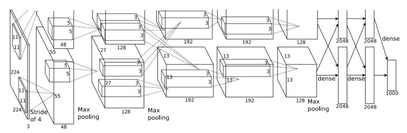
Figure 2: Three kinds of image data in the project

There are three kinds of image data in the scope of this project: text, image, and **text and image dominant**. Our project focuses on text and image dominant memes.

Methodology

A. BERT + AlexNet





B. ResNet + VGG + LSTM

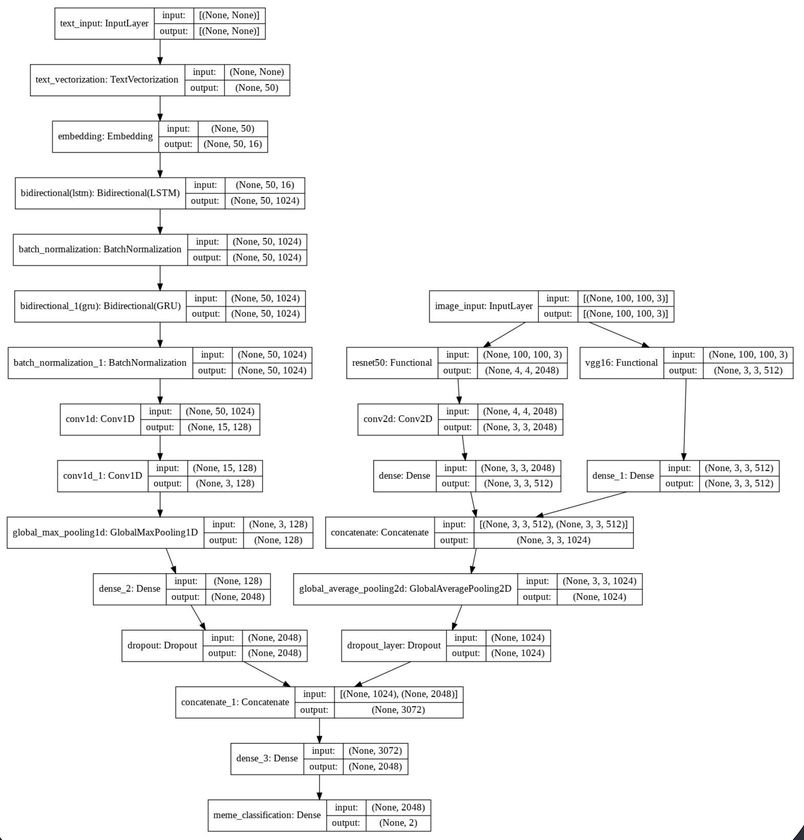


Figure 3: Combined Model Architecture for ResNet, VGG and LSTM

Experiments and Results

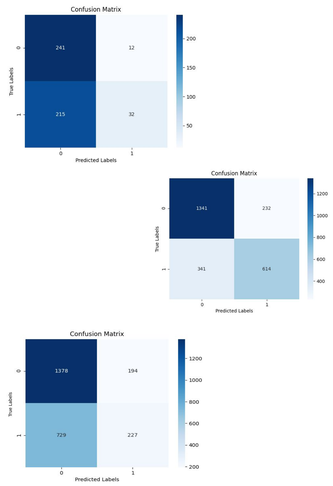


Figure 4: Predicted vs Actual labels for BERT (top), AlexNet (middle) and Alex - BERT Ensemble (bottom)

BERT (Text):

Trained for 15 epochs, batch size 8

Accuracy: 55%

AlexNet (Image):

Trained for 20 epochs, batch size 4

Accuracy: 63%

BERT + AlexNet:

Trained for 10 epochs, batch size 4

Accuracy: 77%

Conclusion: Beats both BERT and AlexNet (individually) by a high margin.

ResNet (Image):

Trained for 30 epochs, batch size 64.

Accuracy: 52.9%

LSTM (Text):

Trained for 30 epochs, batch size 32.

Accuracy: 89%

ResNet + VGG + LSTM:

Trained for 25 epochs, batch size 256.

Accuracy: 64.47%

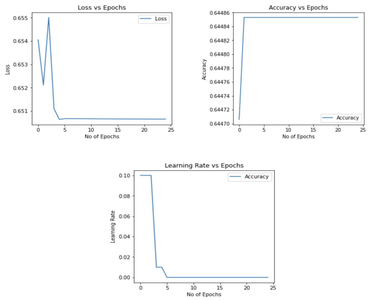


Figure 5: Classification Plots for Combined Model

Conclusion: Better than ResNet but less than LSTM (individually).

Discussion and Future Scope

- Higher accuracy with combined models
- Using user activity as a metric to judge if a meme is hateful
- Utilizing multimodal models
- Recurrent architectures for video memes