Analyzing hateful memes - Report

Saideekshith Vaddineni(2022101110)

1 Tasks Attempted

a) Object Detection:

- Goal: Detect objects in memes using computer vision.
- Tasks:
  - Apply object detection algorithms to identify elements in memes.
  - Catalog detected objects, analyze their frequency and distribution.
  - (Bonus) Develop a simple classification system using object catalog and meme toxicity labels.

b) Caption Impact Assessment:

- Goal: Analyze how captions affect object detection accuracy.
- Tasks:
  - Determine impact of text overlays on object detection.
  - Explore methods to mitigate caption impact.

c) Classification System Development:

- Goal: Develop a system to classify memes.
- Suggestion: Classify memes vs. non-memes using provided dataset.

d) BONUS TASK: Predict meme toxicity based on image text.

2 Tasks Description

a) Object Detection:

- Model Used: Used Ultralytics YOLOv8 model for object detection in the image. This model classifies objects detected in an object into 80 classes.
- Methodology:

  - For this task, I calculated the frequency of objects in all images both hateful and non-hateful memes. If an object is present in a hateful meme, its frequency increases by 1, and if its present in a non-hateful meme, its frequency decreases by 1.
  - I did the same thing with pairs of objects by checking if they were present in the hateful or non-hateful memes and calculate their frequency accordingly.
  - Then to calculate toxicity score of an image, I detected the objects in the image and did the following steps:
    - Made 2 lists, list1 and list2.
    - If an object is present in the image append its frequency to list1 and if a pair of objects is present in the image append its frequency to list2.
    - The after the entire process, we calculate toxicity score of the image in the by first calculating sum of elements in list1 and list2 separately and then final score by taking weighted average of these sums(Giving weight 0.6666 to list2 sum and weight 0.33333 to list1 sum)

.

- Observations: The mean and median of images in the test set which are non-toxic are 5613.52 and 90 respectively and for the images which are toxic, the mean and median are 12152.06 and 164 respectively. This indicates that the toxic memes have higher toxicity score in general than non-toxic memes showing that our methodology is accurate.

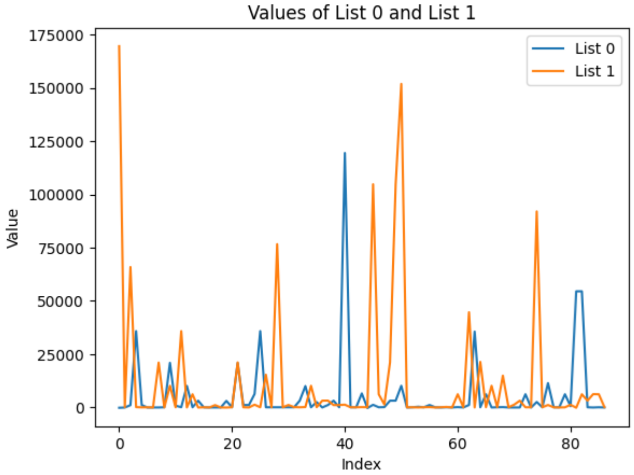

Figure 1: Toxicity Score Graph

- Prediction Model(Bonus): Based on the frequency of objects in images in the train set, I made a csv file with 81 columns(80 columns for different classes of objects detected and 1 column for whether its hateful or not) and each of the 80 columns has frequency of that object for that image. Then I used an XGBClassifier model and trained it based on the data in the csv file. Then based on the test set, I made a csv with 80 columns for it and each column contains frequency of that object in that image. I tested my model on this test csv and got an accuracy of 0.537 .

b) Caption Impact Assessment:

- For this task, I ran the aboject detection algorithm over 10 memes and their original templates and analyzed the ratio of objects detected in the meme template but not detected in the meme to the total number of objects in the meme template. Then the average ratio of this is 0.254.

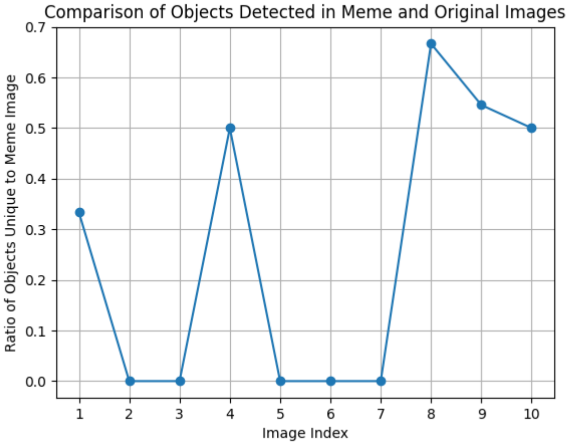

Figure 2: Object Detection Graph

- To minimize the impact of captions, I used image inpainting. I found a model called Kandinsky and even created a mask image(white captions with black background) to inpaint the meme. But due to heavy memory usage, I couldn't use this code.

c) Classification System Development:

- Model Used: VG19 model
- Methdology: Collected around 800 images from the internet used trained the VG19 model using these images and same number of

memes. I used Adam optimizer and binary cross entropy loss function. The trained model is stored in a file. I used 10 epochs for the training.

- Results: My test set has about 13 images and my model works for 12 of them. The one it doesn't work for is an image but it predicts it as a meme(Erony.jpeg).

d) BONUS TASK: Predict meme toxicity based on image text

- Model Used: BertForSequenceClassification model
- Methodology: Preprocessed all the text by replacing slang words and contracted words. Then trained the BertForSequenceClassification model on the training set and saved the model. For the training, batch size was 16, epochs was 3, learning rate was 0.00002, optimizer was Adam optimizer and loss function was binary cross entropy.
- Results: After testing the model over the test set(385 images) I got an accuracy score of 0.545. Text only analysis produces more accuracy than image only analysis on this test data but there isn't much difference.

e) Combining image prediction and text prediction(Multimodal Analysis)

Combined both image and text prediction to get the final prediction by taking average of their prediction probability and comparing it with 0.5 . If greater than or equal to 0.5, then its a hateful meme otherwise its a non-hateful meme. I got 0.52 accuracy with this multimodal analysis.