CS 688 Project

Hateful Meme Identification using Multi-modal approach

Venkata Sandeep Yerra
Boston University

Agenda

- Introduction
- Hateful Meme Identification
- Dataset Particulars
- Unimodal Approaches
- Multi-modal : Visual Bert
- Multi-modal: MMBT + CLIP (OpenAI)
- Feature Augmentation
  - Facebook Deep Face algorithm
  - Google Entity Detection Model
- Comparison Across models
- Future Work
- References

Introduction

- Why meme ? Is it important for reasons other than fun ?
  - The average millennial looks at 20-30 memes every day
  - Over 60% of the people say they would more likely to buy from a company that uses memes in their marketing
  - The click-through-rate(CTR) of a meme campaign is 14% higher than email marketing
  - CTRs in an average marketing campaign are approx. 6% whereas in meme marketing is roughly 19%
  - In 2020, Global meme industry was valued at $ 2.3 billion and is projected to grow to $ 6.1 billion by 2025
  - Source: https://www.amraandelma.com/meme-statistics/

Hateful Meme Identification

- Task of detecting multimodal hate is both extremely important and particularly difficult
- Relying on just text or just images to determine whether a meme is hateful is insufficient
- Using certain types of images, text, or combinations, a meme can become a multimodal type of hate speech
- FacebookAI has released a dataset to help build systems that better understand multimodal hate speech

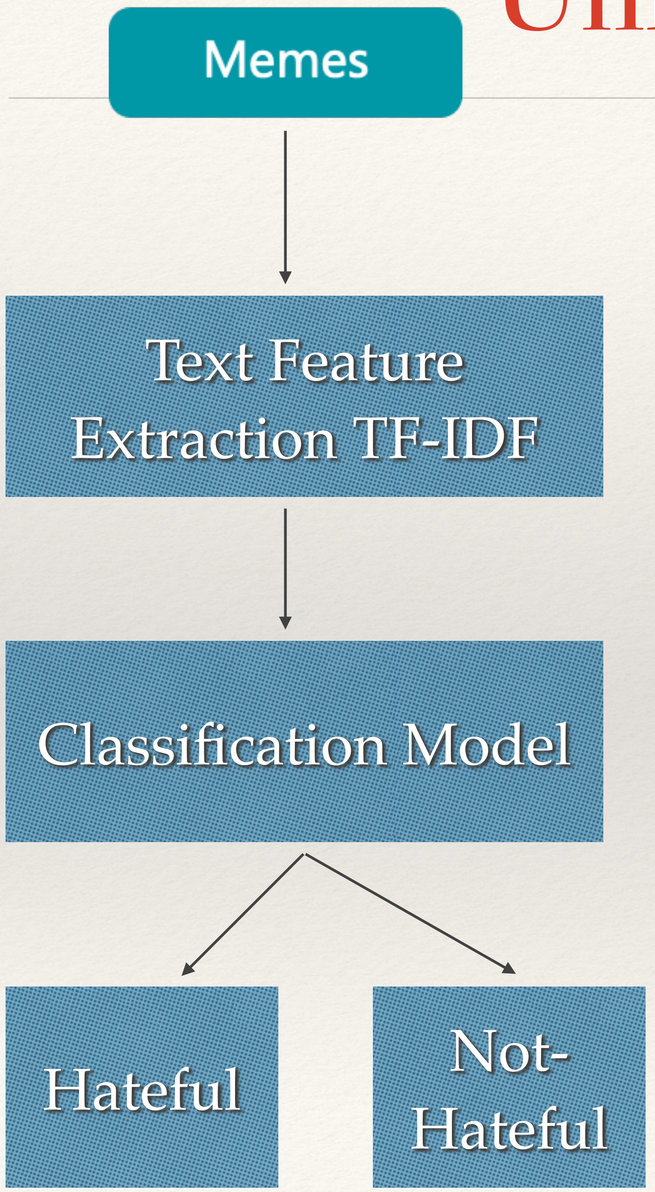
Text and image by themselves are not hateful .
But in combination they can become a form of hate speech

Dataset Details

- How was dataset created by Facebook?
  - Licensed images from Getty images
  - Third-party annotators created new memes similar to ones shared on social media sites
  - Examples cover wide range of protected categories
    - Religion, gender etc
  - Distribution in dataset reflects real-world distribution in original examples in social media

- Particulars
- Total Data: 10,000 images in jsonline format
  - Dataset division
    - Train.jsonl (8500)
    - Dev_seen.jsonl (1040)
- Features in dataset
  - Id: Indicating id of the data
  - Img path
  - Text
  - Label : 0 for non-hateful and 1 for hateful
- Confounders added

Unimodal Approaches (1/2)



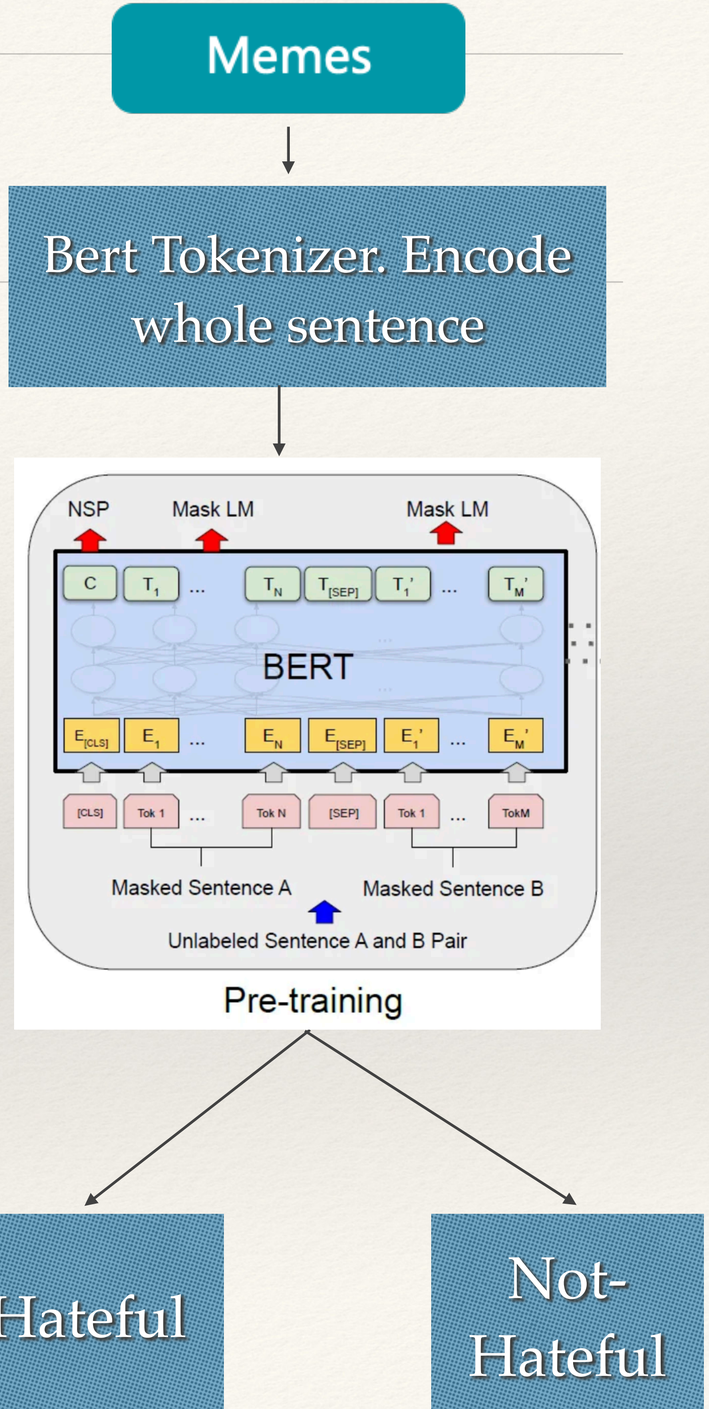
Using TF-IDF, Extract features measuring Importance of each word

Features	Model	Validation Accuracy	Validation ROC-AUC Score
TF-IDF	Logistic Reg	0.5270	0.5018
TF-IDF	Naive Bayes	0.5413	0.5301
TF-IDF	Random Forest	0.5682	0.5040
TF-IDF	SVM	0.5673	0.5024
TF-IDF	KNN	0.575	0.5169

Each model was hyper-parameter optimized on parameter grid to get the best possible fit. The final model was retrained using the best parameters and above scores are on validation set

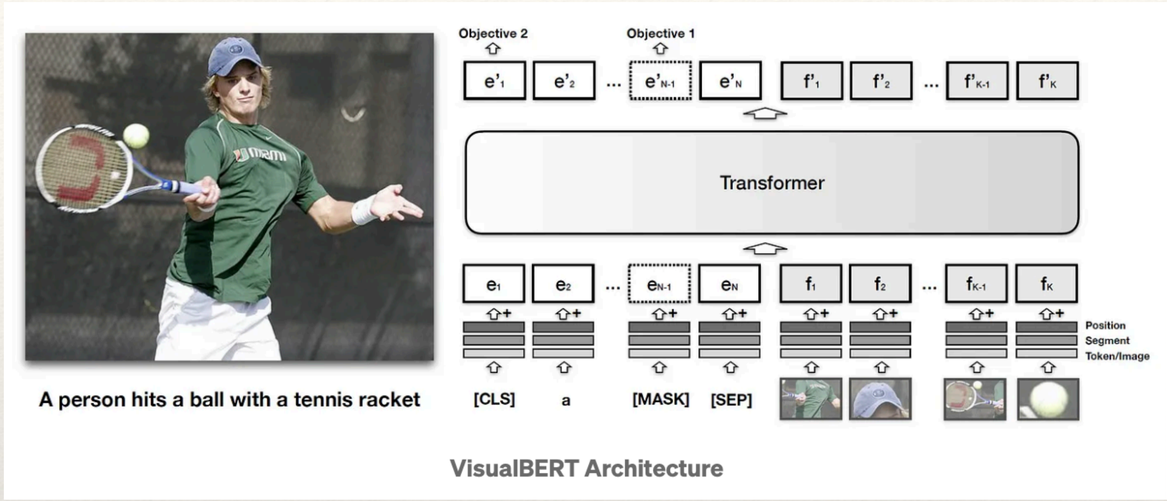
Unimodal Appoach : BERT (2/2)

- Bert Based model
  - Why is this better?
    - Trained on 2500M words in Wikipedia and 800M from books
    - Two modeling methods
      - Masked Language Model (MLM)
      - Next Sentence Prediction
  - Tokenize and encode whole sentences
  - About model implementation
    - Used bert-base-uncased from Transformer Library
      - Optimizer: Adam, Learning Rate: 1e-5, Epochs=30, batch_size=32
  - Model Results:
    - Accuracy: 0.558
    - AUC: 0.55



Multi-modal Approaches (1/2)

- Visual BERT
  - Multi-modal vision and language model
  - Applications
    - Visual-Question Answering
    - Multiple-choice
    - Visual Reasoning
    - Region-to-phrase correspondence
  - VisualBERT is trained using COCO, which consists of images paired with captions
  - Available in Transformers Library: uclanlp/visualbert-nlvr2-coco-pre : NLVR is a dataset for joint reasoning about natural language and images, with a focus on semantic diversity, compositionality, and visual reasoning challenges. The task is to **determine whether a natural language caption is true about a pair of images**



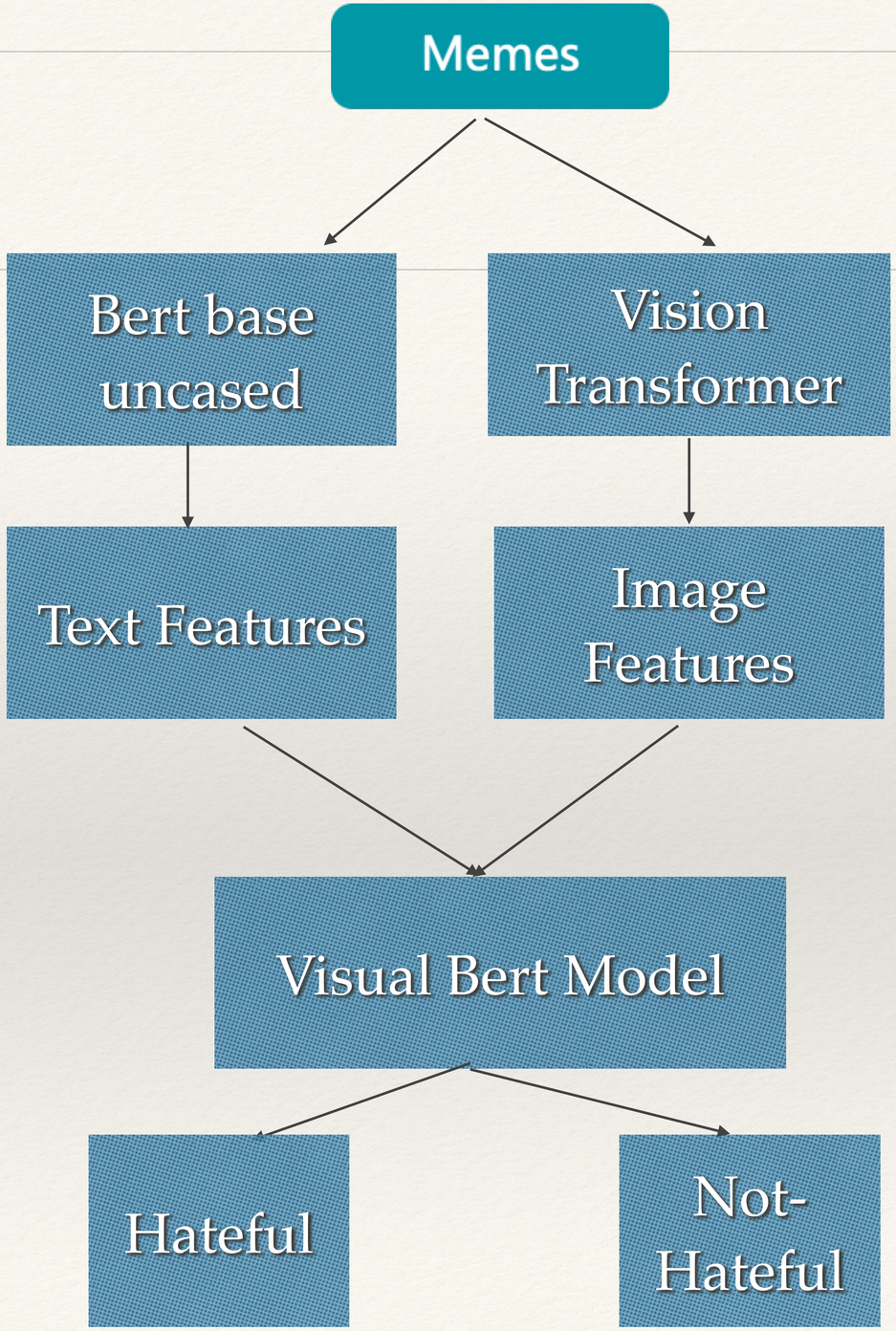
VisualBERT Architecture

- Pre-Training
- **Masked Language modeling with image**: Elements of text input are masked but Image vectors are not masked
- **Sentence-Image prediction**: For COCO, where there are multiple captions corresponding to one image, we provide text segment consisting of two captions
- One of the caption is describing the image, while the other has a 50% chance to be another corresponding caption and a 50% chance to be a randomly drawn caption.
- The model is trained to distinguish these two situations

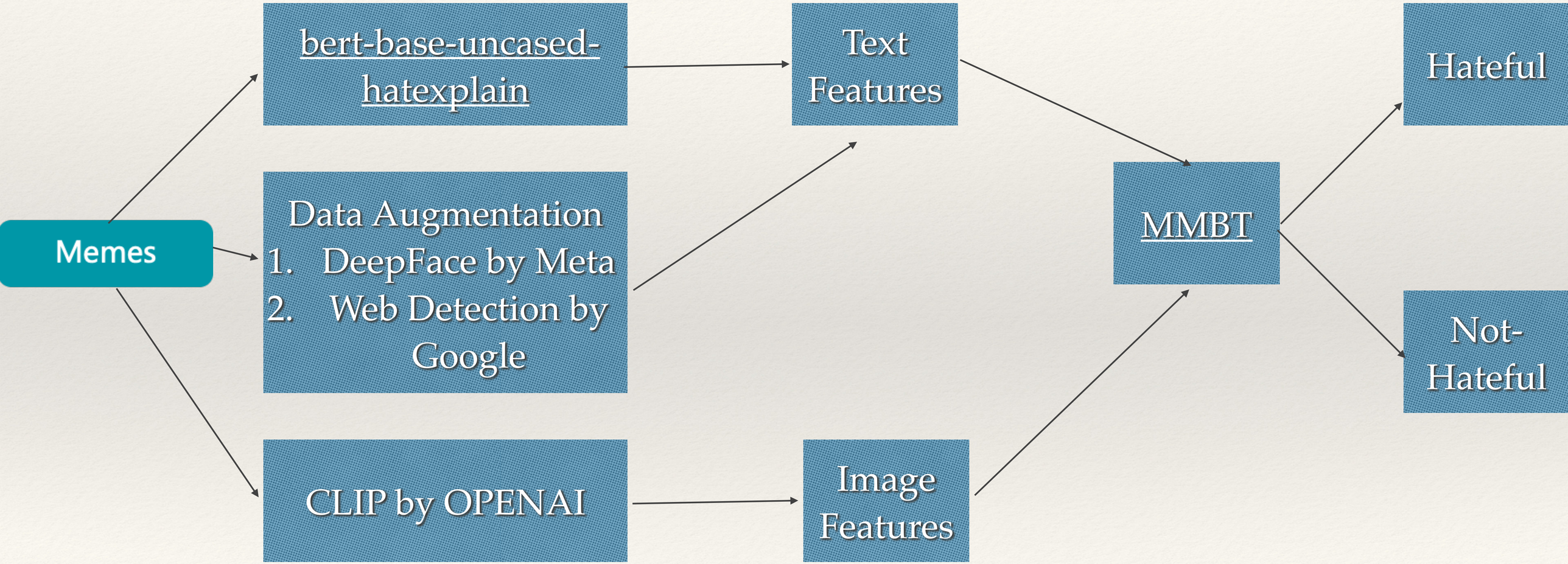
Visual BERT Approach

- Visual Bert model requires 2 inputs
- Text Features: bert-base-uncased
- Visual Features: google/vit-base-patch16-224-in21k: The Vision Transformer (ViT) is a transformer encoder model (BERT-like) pretrained on a large collection of images in a supervised fashion, namely ImageNet-21k, at a resolution of 224x224 pixels

- Model Performance
- Validation Data Accuracy: 0.610
- Validation AUC: 0.613
- Improvement but still not great

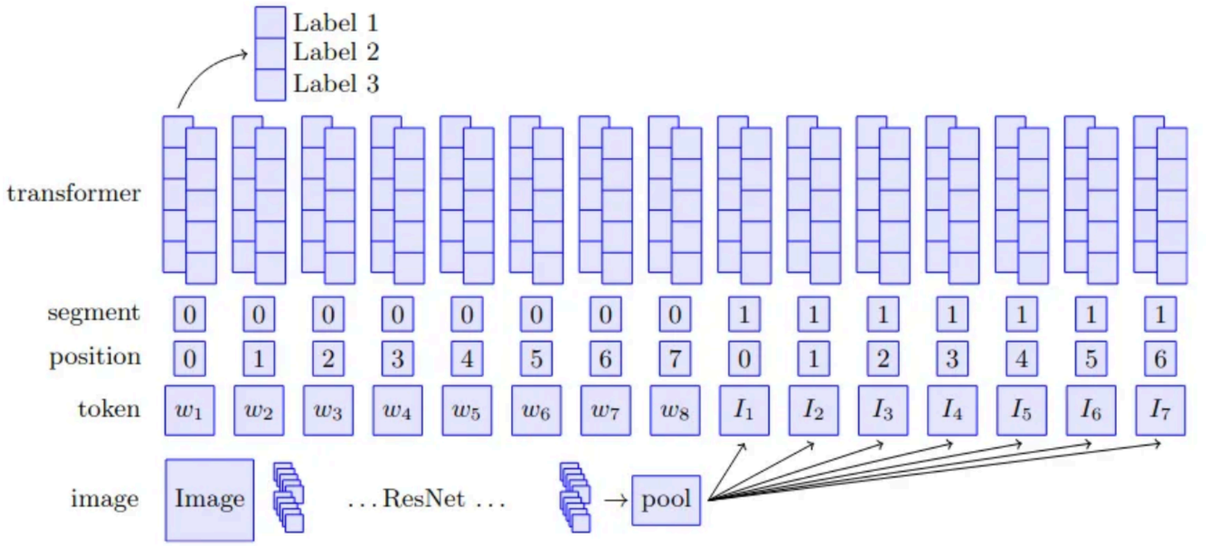


Model Architecture Change



MMBT + CLIP Approach

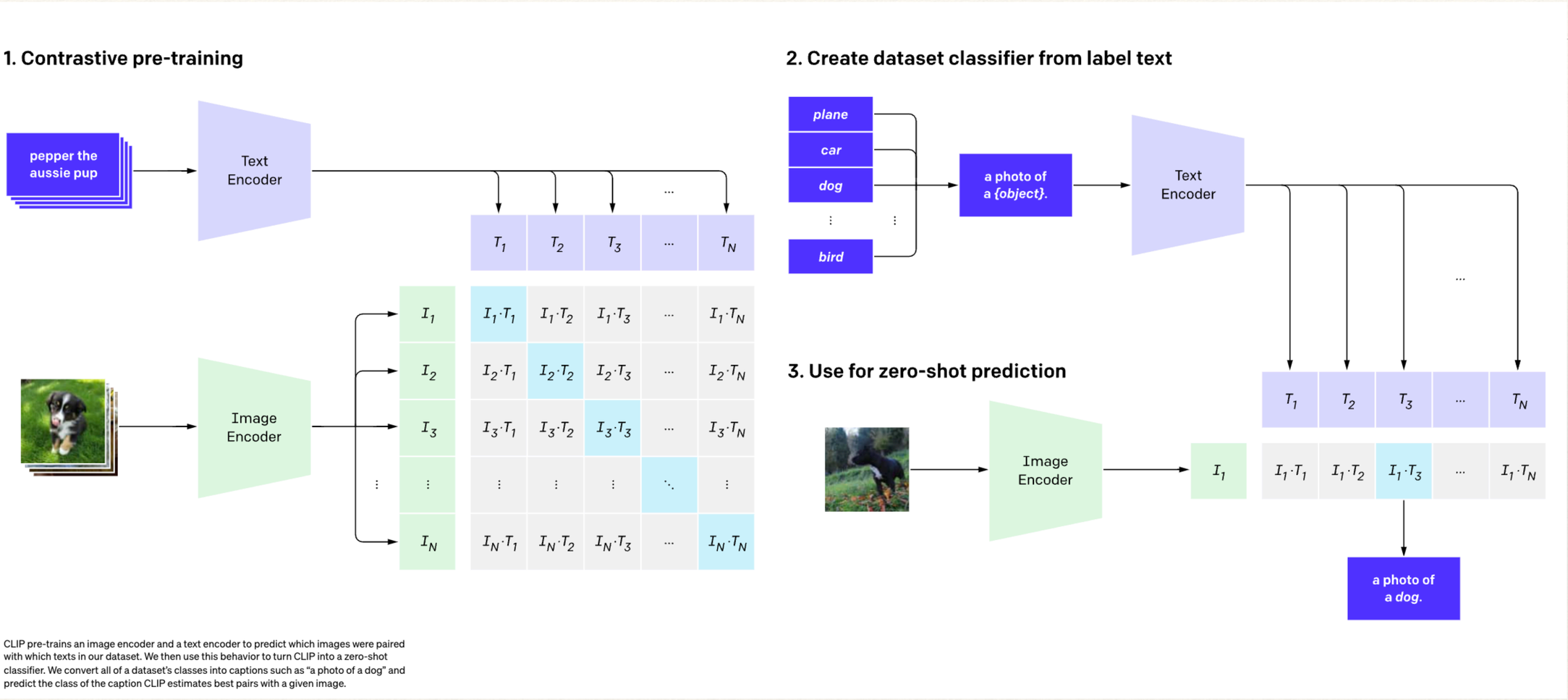
- Multi-modal BiTransformers (MMBT)
- MMBT uses text features from Bert and Resnet for image encoding
- MMBT is flexible for feature embeddings. Can use CLIP for generating image features.
- Better than ViL Bert/ comparable performance for Hard Problems (where vision and text are giving opposite predictions)



MMBT architecture, from Supervised Multimodal Bitransformers for Classifying Images and Text paper

	MM-IMDB	-Hard	FOOD-101	-Hard
MMBT	61.6±.2 / 66.8±.1	**65.3±.4 / 68.6±.4**	92.1±.1	**92.4±.5**
MMBT-Large	*63.2±.2 / 68.0±.2*	*68.2±.5 / 70.3±.4*	*93.2±.1*	*93.4±.3*
ViLBert-VQA	60.0±.3 / 66.4±.2	62.7±.6 / 66.2±.4	92.1±.1	92.4±.3
ViLBert-VCR	61.6±.3 / 67.6±.2	63.4±.9 / 66.9±.4	92.1±.1	92.1±.3
ViLBert-Refcoco	61.4±.3 / 67.7±.1	63.4±.5 / 67.1±.4	92.2±.1	92.1±.3
ViLBert-Flickr30k	61.4±.3 / 67.8±.1	63.4±.9 / 67.0±.5	92.2±.1	92.2±.3
ViLBert	**63.0±.2 / 68.6±.1**	**65.4±1. / 68.6±.4**	**92.9±.1**	**92.9±.3**

CLIP



CLIP pre-trains an image encoder and a text encoder to predict which images were paired with which texts in our dataset. We then use this behavior to turn CLIP into a zero-shot classifier. We convert all of a dataset's classes into captions such as "a photo of a dog" and predict the class of the caption CLIP estimates best pairs with a given image.

Across a suite of 27 datasets measuring tasks such as fine-grained object classification, OCR, activity recognition in videos, and geo-localization, we find that CLIP models learn more widely useful image representations. CLIP models are also more compute efficient than the models from 10 prior approaches that we compare with.

Feature Augmentation - Deepface

- Extract features from Image using Facebook Deepface (Open Source)
- DeepFace is a facial recognition system developed by Facebook researchers.
- Uses a nine-layer neural network with over 120 million connection weights to identify human faces in digital images. DeepFace was trained on four million images uploaded by Facebook users
- Regression Model
  - Age
- Classification Model
  - Emotion
  - Gender
  - Race

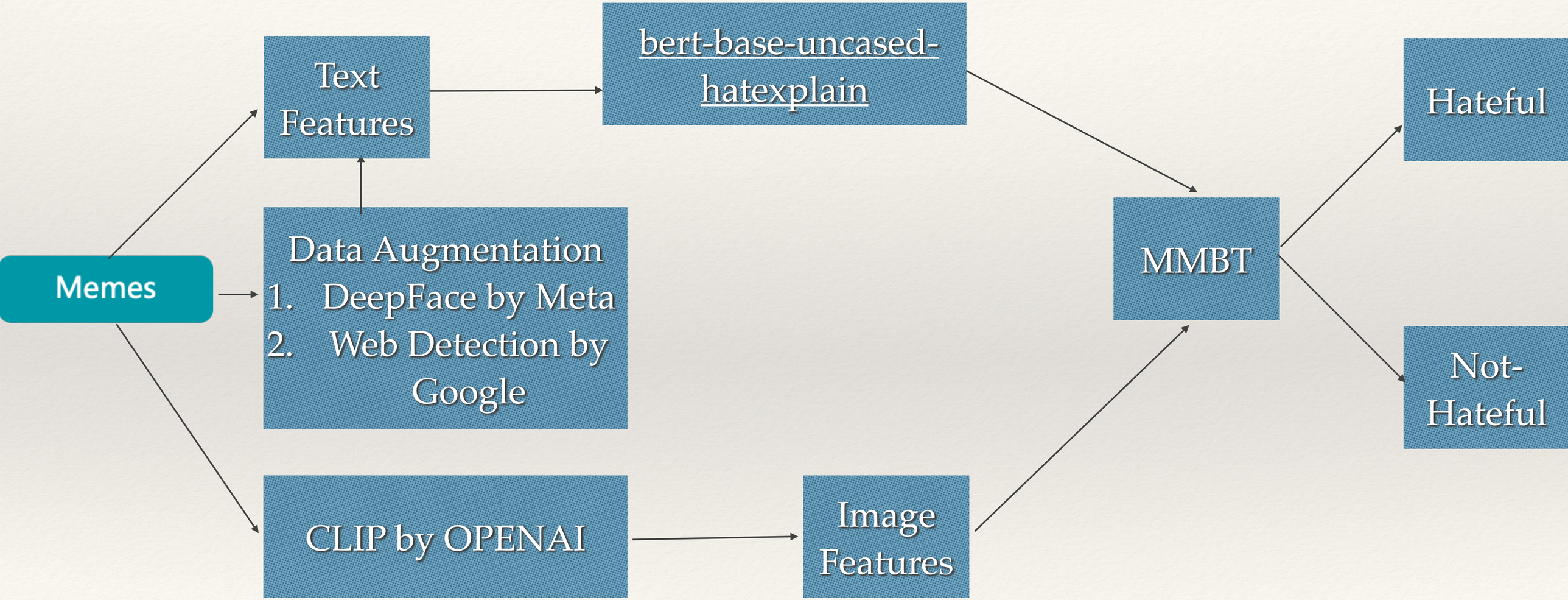

```
{
   "emotion":{
      "angry":7.603101671639384e-14,
      "disgust":2.7474185705216866e-21,
      "fear":1.688688161735822e-14,
      "happy":100.0,
      "sad":4.205067717644173e-10,
      "surprise":7.103817571484745e-13,
      "neutral":4.4851553027136504e-08
   },
   "dominant_emotion":"happy",
   "age":31,
   "gender":"Woman",
   "race":{
      "asian":0.9087088517844677,
      "indian":1.1444833129644394,
      "black":0.09399998234584928,
      "white":66.56872034072876,
      "middle eastern":16.655877232551575,
      "latino hispanic":14.628209173679352
   },
   "dominant_race":"white"
}
```

Feature Augmentation - Web Detection

- Web entities detection with Google Cloud Vision API (free tier)
- Web Detection detects Web references to an image and best guess based on image
- Gives context of image instead of just giving objects identified in image

Best Guess Labels: rio carnival 2019 dancers

Model Change



MMBT+CLIP

- Model Details
- Text Augmented with Features from Deepface and Google Web Detection model
- Parameters
  - Optimizer: Madgrad : https://github.com/facebookresearch/madgrad
  - Learning Rate: 2e-4
  - Batch size: 16
  - Epochs: 5 (kept low as exceeding time on Colab free tier)
- Validation
  - Accuracy: 0.7010
  - AUC: 0.7843

Performance

Features	Model	Accuracy	ROC-AUC Score
Text	Logistic Reg	0.5270	0.5018
Text	Naive Bayes	0.5413	0.5301
Text	Random Forest	0.5682	0.5040
Text	SVM	0.5673	0.5024
Text	KNN	0.575	0.5169
Text	BERT	0.558	0.55
Text+ Image	Visual BERT	0.613	0.61
Text + Image	MMBT + CLIP	0.70	0.78

Future Work

- Other approaches to the problem include
  - Better Feature Processing: In-painting image using OCR methodology to be able to better extract features
  - Dataset Augmentation: Using memotion dataset which has been pre-trained on hateful text yields better results. But sadly the labeling is not accurate hence manual labeling was performed on on features
  - Ensembling methods
    - Ensemble a combination of Vil-BERT, UNITER, Ernie-Vil on augmented features

References

- Visual Bert
- CLIP
- Facebook Meme competition
- Dataset Augmentation Approach
- CLIP + MMBT approach
- Facebook Deep face link
- Google Web detection model

Thank You