Identifying Hateful Memes with Multimodal Classification

Alex Fan[1] and Yixuan (Sherry) Wu[1]
[1]Department of Statistics, Research Mentor: Drew Kaul

CS231n: Computer Vision
Poster Presentation, Spring 2022

Background & Data

Problem:

Given a meme, can our model correctly classify it as either hateful or non-hateful based on the combined input of the text and image?

Motivation:

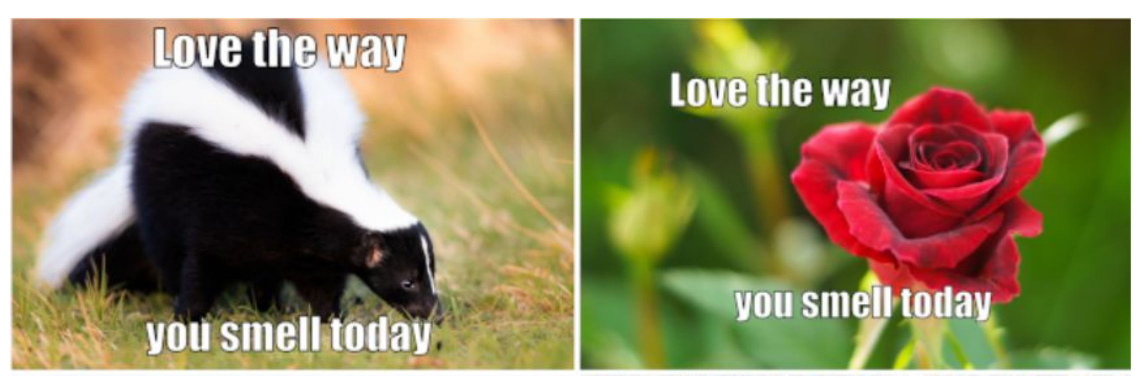
- This task involves *multimodal classification*, which requires the model to effectively combine representations from drastically different modalities.
- Three main methods that have been studied:
  1. Unimodal models
  2. Multimodal models with unimodal pre-trainings
  3. Multimodal models with multimodal pre-trainings

Goal:

Implement and improve on the vanilla VisualBERT model.

Data:

- From Meta's Hateful Meme challenge in 2020
- 11,040: 8500, 540, 2,000 in train, dev, and test
- Each set contains: 10% unimodal hate, and 40% multimodal hate; 20% benign text confounder, 20% benign image confounder, and 10% non-hateful - Challenging Set!



Model Architecture and Training

Baseline

- Resnet152 Pretrain for Images
- Sentence-BERT Pretrain for text
- Concatentation Fusion
- Hidden Size: 1200

VisualBERT

- Detectron2 R-CNN for Images
  - Uses Resnet101 Backbone Model
- BERT Tokenizer for Text
- VisualBERT Pretrain
  - NLVR2-COCO Pretrain Weights
- Hidden Size: 1000

General Feed-Forward Block

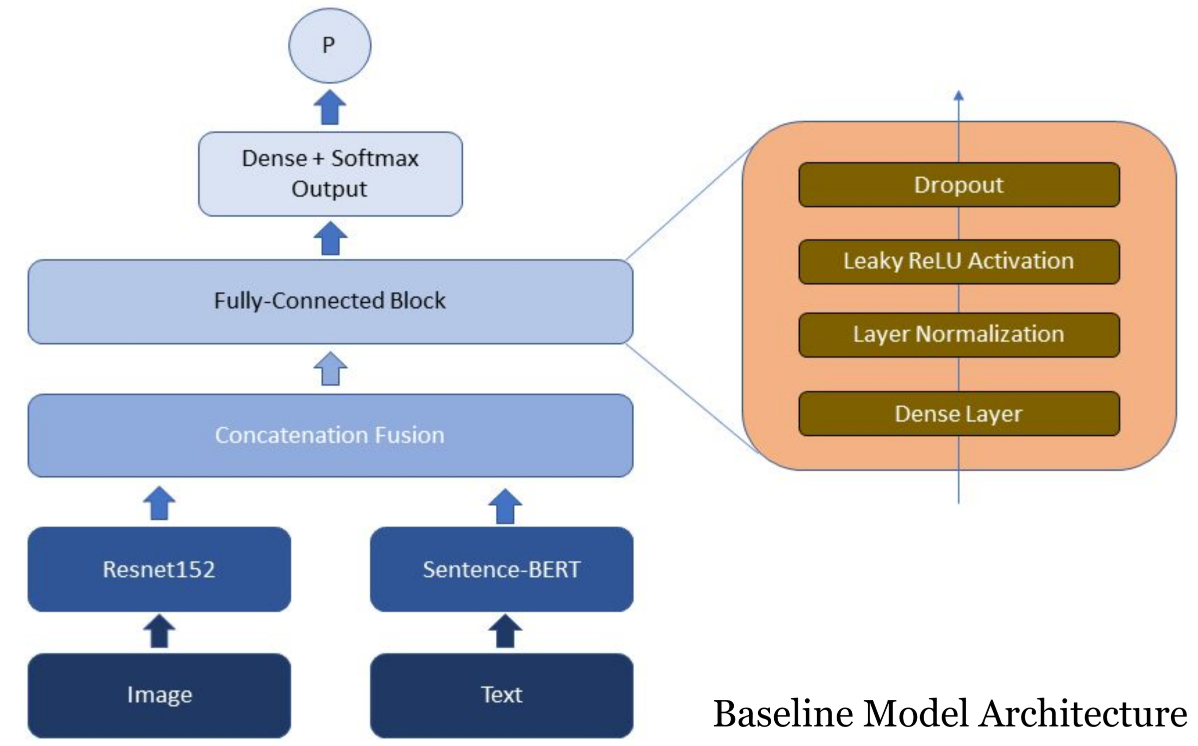
- Linear Layer (Kaiming)
- LayerNorm
- Leaky ReLU
- Dropout (tuned)

Feature Extraction

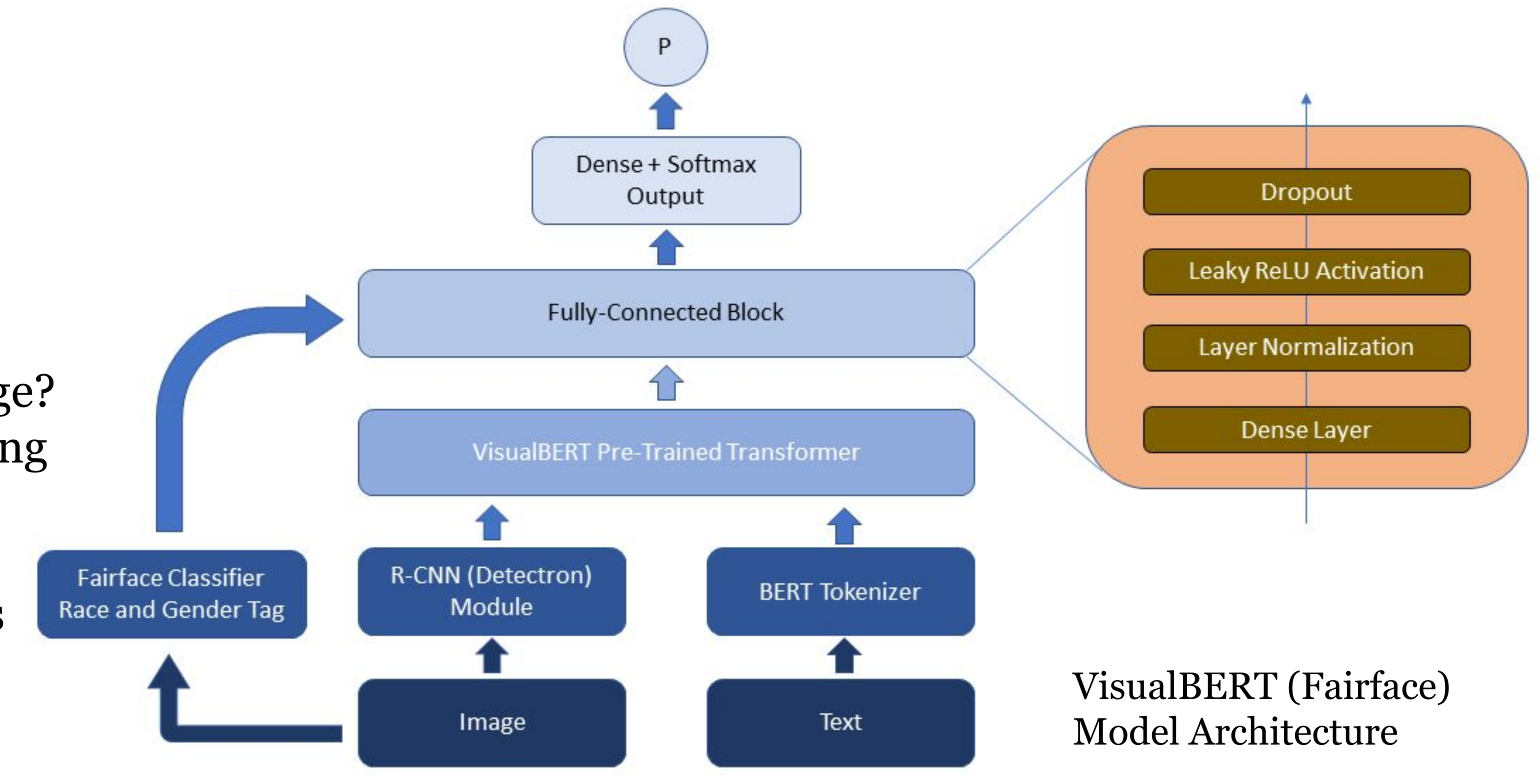
- FairFace
  - Face in Image?
  - If Face, what race/gender/age?
  - Concatenated after embedding extraction

Additional Details

- All Models trained for 10 epochs
- Training set manually balanced

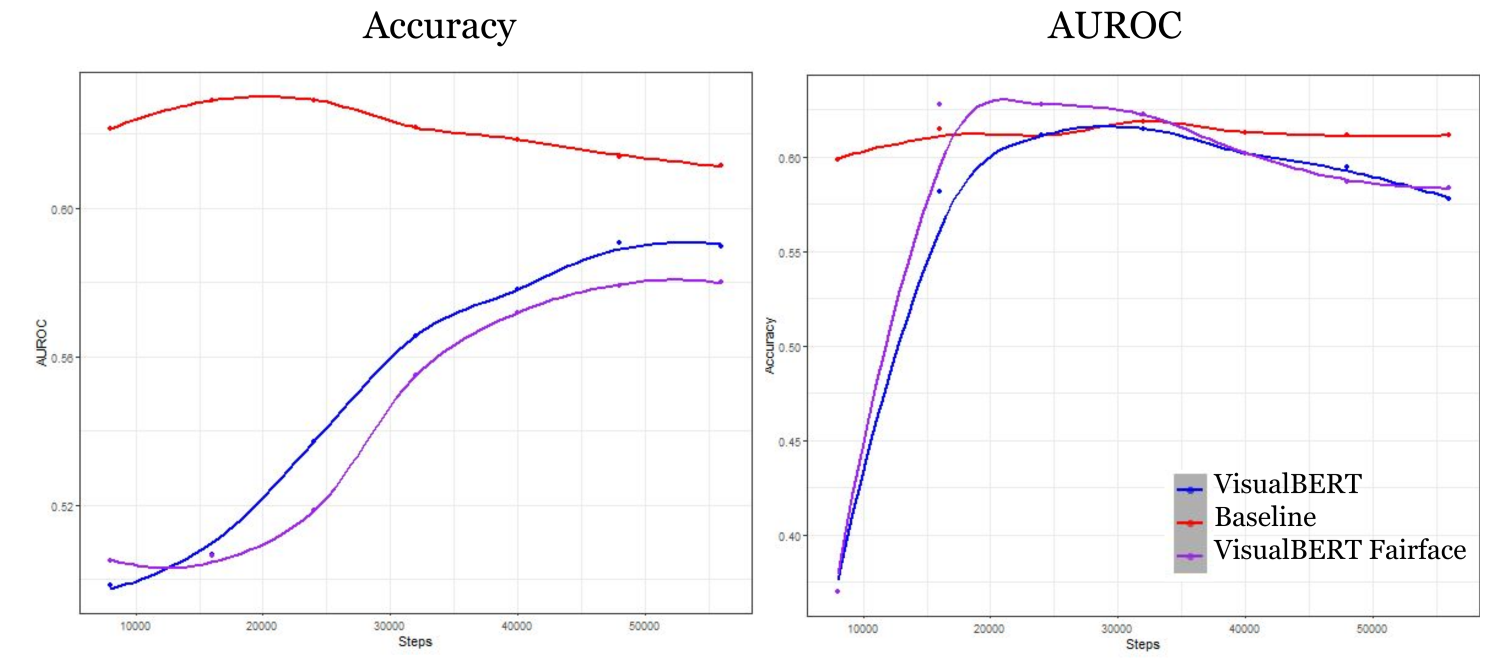


Baseline Model Architecture



VisualBERT (Fairface) Model Architecture

Results



Dev Accuracy and AUROC of three different models vs. Steps

Mixed results from the different models:

- Baseline Achieves 59.8% accuracy on test set (60.5% AUROC)
- VisualBERT accuracy reduces to 59.1%, with corresponding 57.1% AUROC
- VisualBERT with FairFace features has 62.4% accuracy, but 57.4% AUROC (examined in Analysis)

Tuned Hyperparameters via Validation:

- LR = 1e-3 for baseline
- LR = 1e-5 for VisualBERT
- pDrop = 0.2
- Adam Betas = (0.9, 0.999)

Key Takeaways:

- Results suggest large underfit of the data, stemming from small sample size and limited feature space most likely.
- Pre-training alone cannot overcome complex nature of meme understanding for machines

Analyses

Predicted	Non-Hateful	Hateful
Actual		
None-Hateful	1202(60.1%)	48(2.4%)
Hateful	705(34.24%)	45(2.25%)

Confusion matrix of VisualBERT Fairface

Predicted	Non-Hateful	Hateful
Actual		
None-Hateful	870 (43.5%)	380 (19.0%)
Hateful	424 (21.2%)	326 (16.3%)

Confusion matrix of Baseline

- Each model wrongly classifies a meme as non-hateful more often than wrongly classifies a meme as hateful.
- From Model 1 to 3, more likely to classify a hateful meme as non-hateful → which explains why AUC curve is lower for Model 3 (VisualBERT Fairface)

- Images that all models wrongly classified as hateful.
- Offensive languages largely contribute to the classification regardless of image meanings.
- Models are still somewhat incapable of considering textual and image representations in coherence.

- Images that the vanilla VisualBERT classified correctly, but not the VisualBERT Fairface.
- VB Fairface relies too heavily on facial tags, and these tags could be wrong in the first place.
- Models are trained to be biased because of biases in our society.

- Images that VisualBERT Fairface classified correctly but not the vanilla VisualBERT. All are actually non-hateful.
- The bottom left may demonstrate that racial tags may be helpful.

Future Work

Due to time and compute limitations, there is still room for improvement. These include:

1. Train on more data if available
2. Explore more advanced fusion techniques, since we use early fusion with simple concatenation (e.g. CNN or RNN-based fusion)
3. Explore other feature extractions (e.g. Web Entity Detection)
4. Improve the current model architecture to decrease probability of wrongly classifying memes as non-hateful
5. Use the current feature tags but fuse it prior to the VisualBERT pre-training

References

[1] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *arXiv:2005.04790, 2020.*
[2] Luinian Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and KaiWei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv:1908.03557 [cs.CV],* 2019.
[3]Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *ICML*, pages 689–696, 2011.