HATEFUL MEMES

FACEBOOK AI CHALLENGE

Shreya Bhootda, Casey Copeland, Ankita Kundra, Yanqi Liang

Advanced Machine Learning - Ghosh
Fall 2021 Term Project

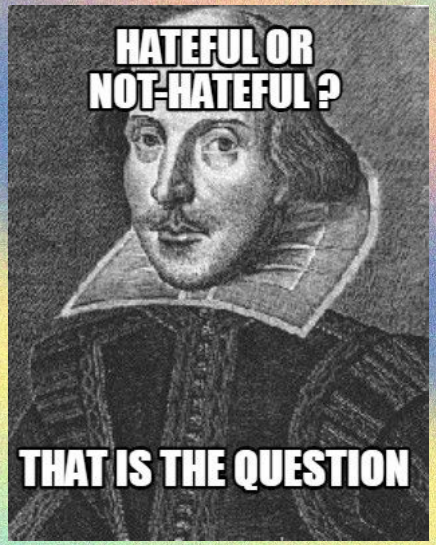


TABLE OF CONTENTS

1

PROJECT OVERVIEW & TASKS

HATE SPEECH ON SOCIAL MEDIA

Hate Speech

- “Any communication that disparages a target group of people based on some characteristics such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics” (Nockleby, 2000)

AI/Machine Learning on Social Media

- Scale of content
- challenges : context, world knowledge, subtle nuances, unwanted bias

Outcome of this challenge

- Multimodal methods that can be applied to a broad set of problems

HATE SPEECH ON SOCIAL MEDIA

MEMES

Multimodal

- Images overlaid with text
- Sophisticated text and image model fusion is necessary for proper classification

Benign Confounders

ACCURACY ACHIEVED BY EXISTING PRE-TRAINED MODELS ON HATEFUL MEMES

TYPE	MODEL	TEST Acc.	AUROC
	Human	84.70	82.65
Unimodal	Image-grid	52.00	52.63
	Image-region	52.13	55.92
	Text BERT	59.20	65.08
Multimodal *(Unimodal pretraining)*	Late fusion	59.66	64.75
	Concat BERT	59.13	65.79
	MMBT-grid	60.06	67.92
	MMBT-region	60.23	70.73
	ViLBERT	62.30	70.45
	Visual BERT	63.20	71.33
Multimodal *(Multimodal pretraining)*	ViLBERT CC	61.10	70.03
	Visual BERT COCO	64.73	71.41

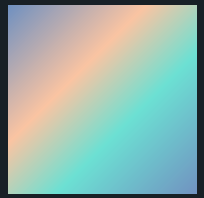
THE COMPETITION

GOAL

Create a large scale classification algorithm that identifies multimodal hate speech in internet memes

PROVIDED BY FACEBOOK AI

Datasets - Train, Dev seen & unseen, Test seen & unseen
Baseline & Starter Kits



TASKS

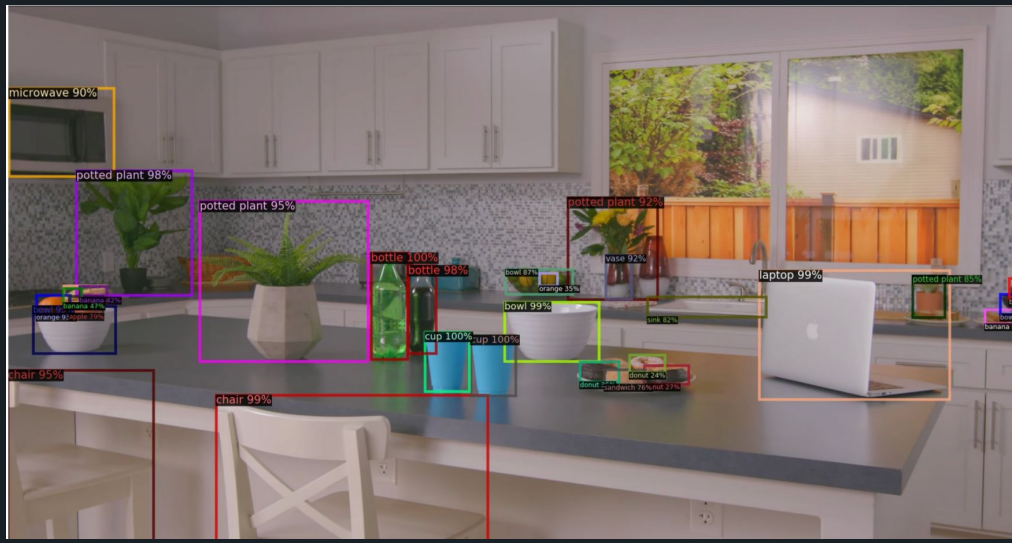
Data Preprocessing, Feature Engineering
Model Selection
AUROC & Accuracy Metrics

2

MODELING PROCESS

FEATURE EXTRACTION
Detectron2

- Pre-trained Models
  - Faster R-CNN
  - RetinaNet
  - RetinaNet + Faster R-CNN
- VisualGenome Dataset



FEATURE EXTRACTION
Google Vision Web Entity Detection

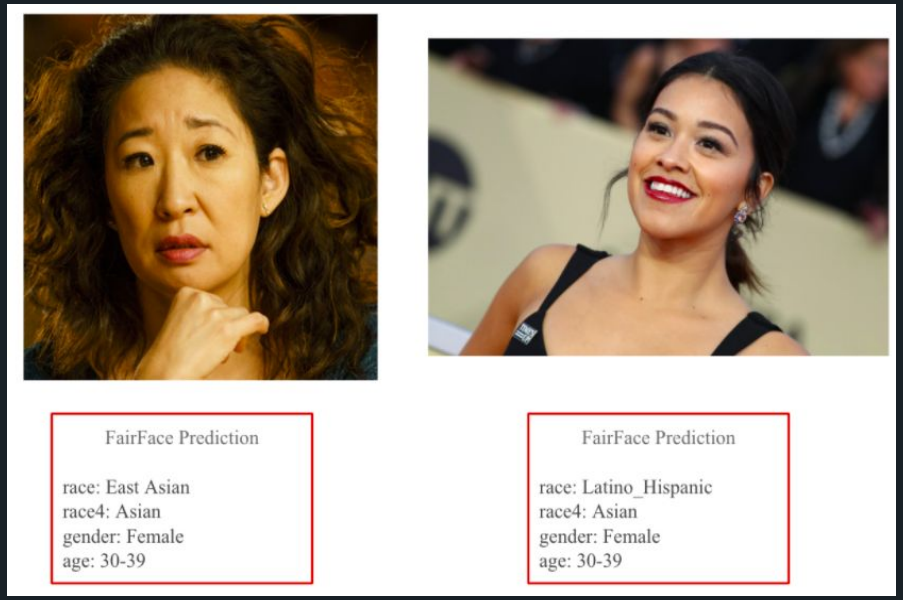
Patch 1 Entity Detection

Naturalization
Immigration
Adjustment of status
Multiple citizenship

Patch 2 Entity Detection

mexico border
Border
Border barrier
United States

FEATURE EXTRACTION
Fairface Classifier



Protected Category	%
Race or Ethnicity	47.1
Religion	39.3
Sexual Orientation	4.9
Gender	14.8
Nationality	9.8
Immigration Status	6.1

MULTIMODAL MODELS

ERNIE-ViL

Incorporates structured knowledge obtained from scene graphs to learn joint representations of vision-language.

UNITER - ITM

UNiversal Image-TExt Representation, uses conditional masking on pre-training tasks

EXTENDED VL-BERT

Extends Transformer model to take both visual and linguistic embedded features as input

OSCAR

Object-Semantics Aligned Pre-Training uses image, text, and object tags for pre training of vision-language tasks

ERNIE-ViL

- Pre-training
  - Scene Graph Prediction Task
- Pre-train dataset
  - Conceptual Captions
  - SBU Captions
- Object/Attribute/Relationship
  - Semantic connection

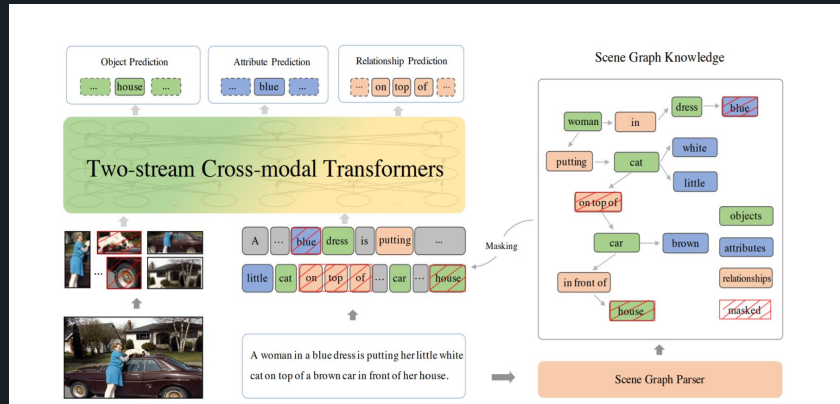
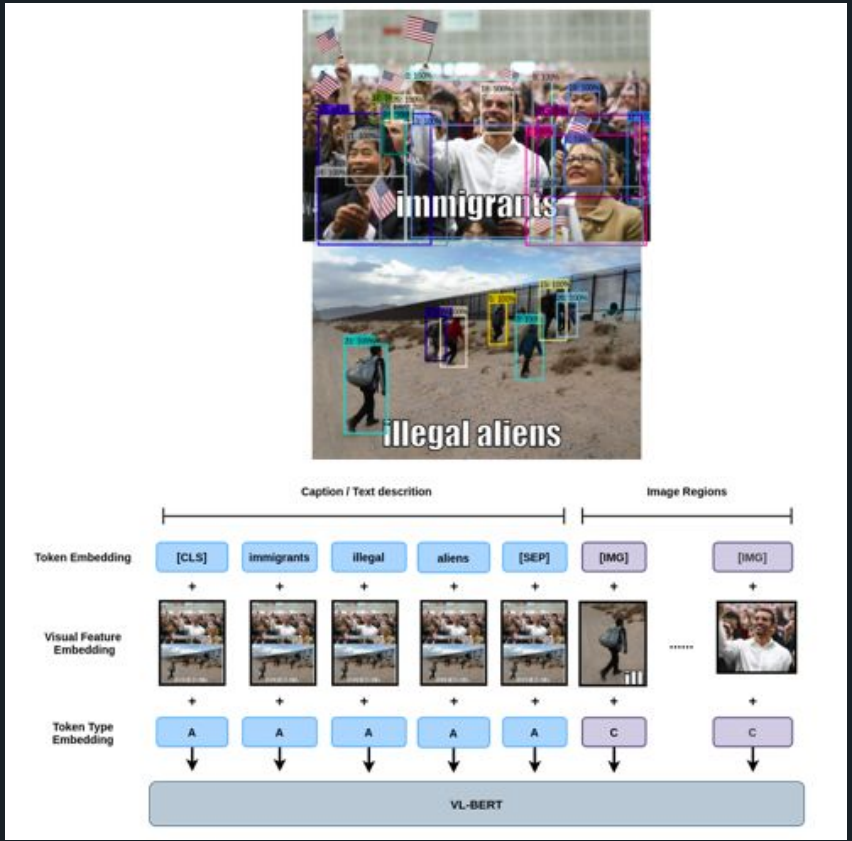
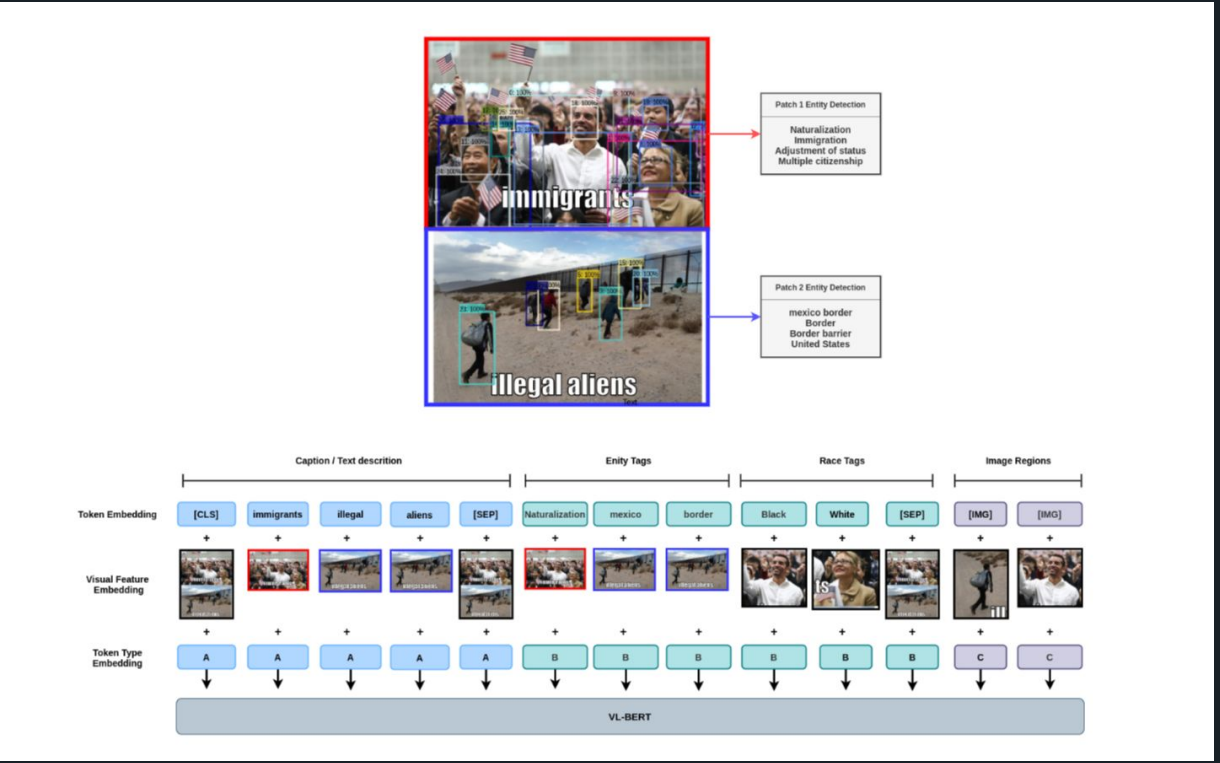


Figure 2: Illustration of Scene Graph Prediction tasks for ERNIE-ViL. Given detected regions of the image and token sequence of the text, ERNIE-ViL uses a two-stream cross-modal Transformers network to model the joint vision-language representations. Based on the scene graph parsed from the text using Scene Graph Parser, we construct Object Prediction, Attribute Prediction and Relationship Prediction tasks to learn cross-modal detailed semantics alignments.

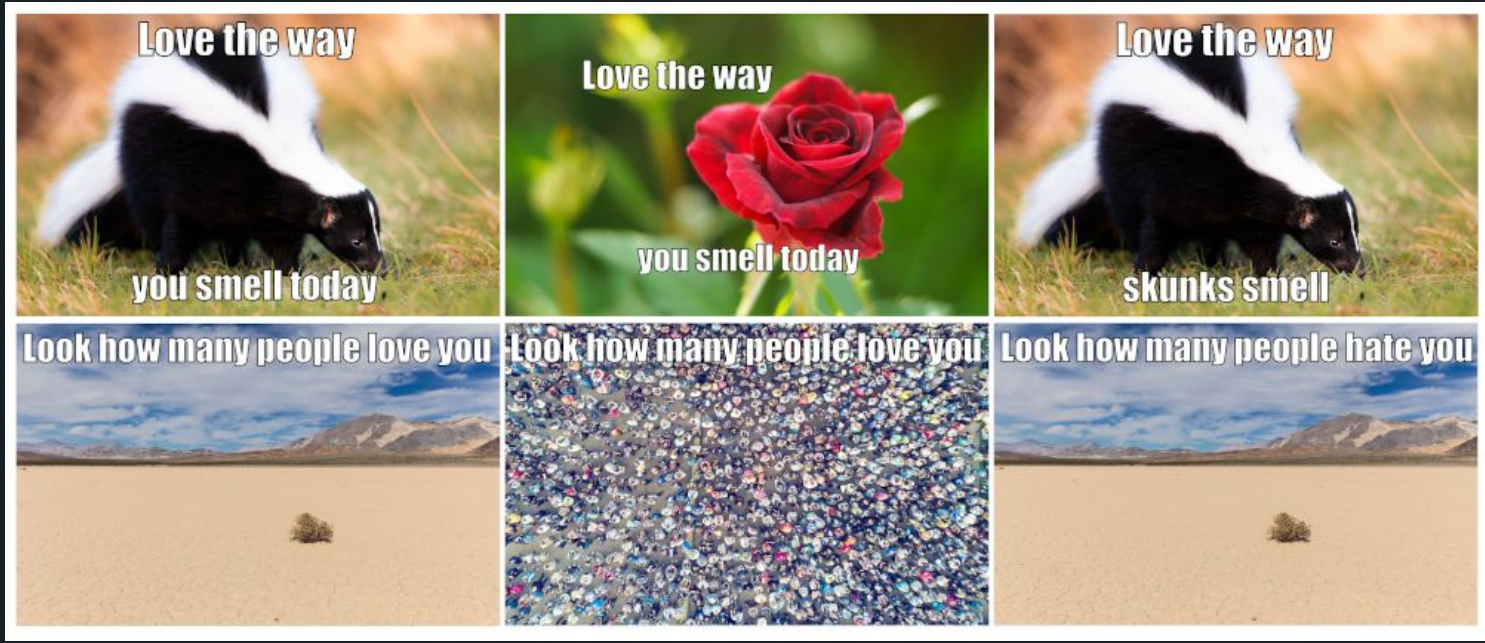
VL-BERT


immigrants
illegal aliens
Caption / Text descrition
Image Regions
Token Embedding
[CLS]
immigrants
illegal
aliens
[SEP]
[IMG]
[IMG]
Visual Feature Embedding
Token Type Embedding
A
A
A
A
A
C
C
VL-BERT

Extended VL-BERT



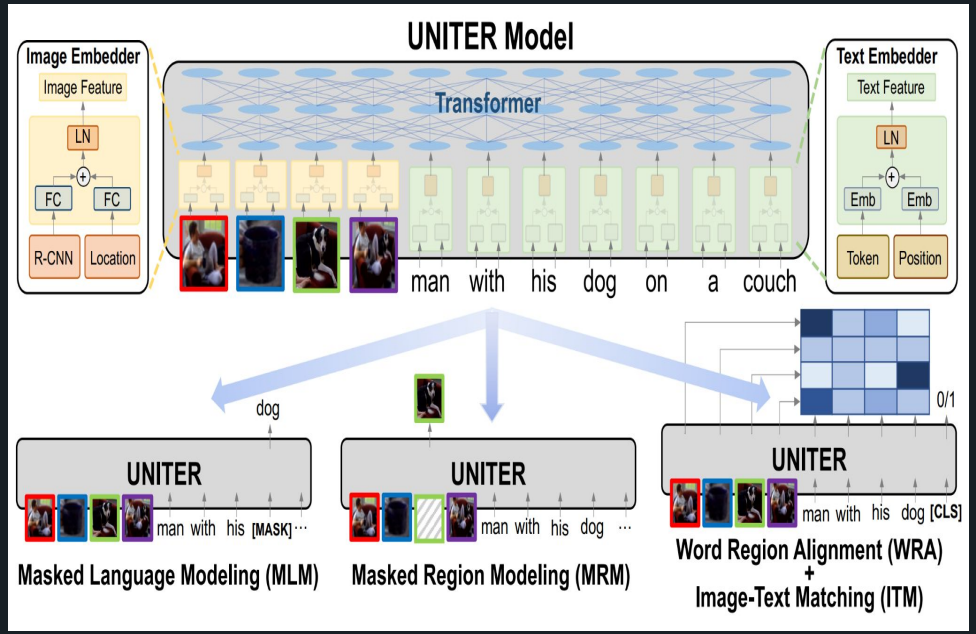
UNITER - ITM
Text Confounders?!



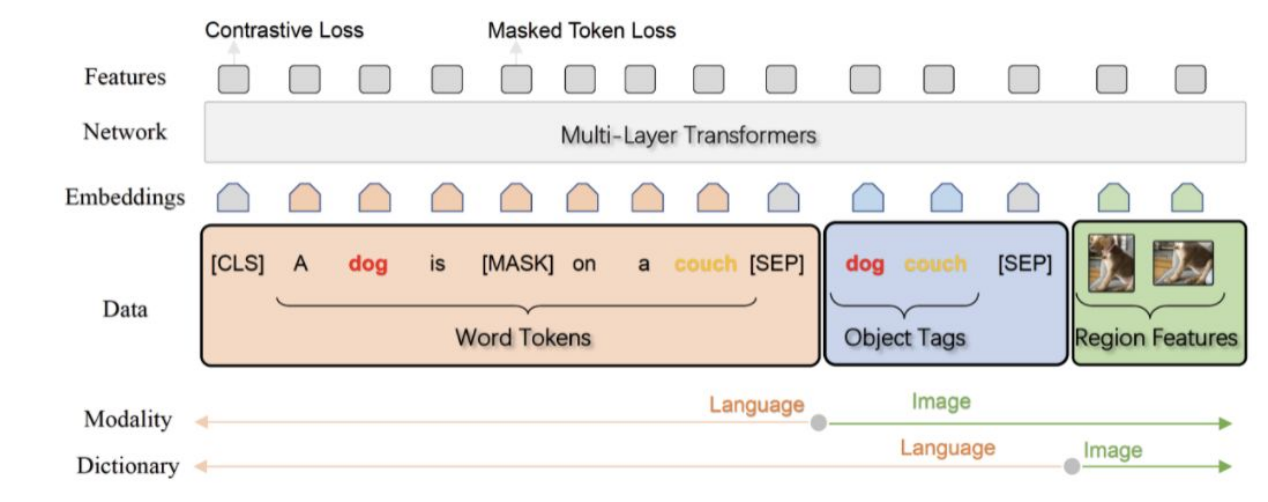
Solution - Image Text Matching

UNITER - ITM

- UNiversal Image-TExt Representation Learning
- Inputs both image and textual tokens
- Image Embedder - R-CNN, FC, Normalization
- Text Embedder - BERT
- ITM



OSCAR



- Find additional meaning through object tags to extract accurate features for V+L tasks
- Pre-trained on 6.5 million text-image pairs
- Classification of memes

3

APPLICATION & ONGOING WORK

FACEBOOK AI

IMPLEMENTATION OF COMPETITION RESULTS

MULTIMODAL PROGRESS

Decrease Model Bias

More research into image-text pre training methods and classification models

FACEBOOK AI

Better enforce community guidelines by accurately classifying multimodal hate

Create a safer platform as Facebook receives ongoing criticism

SOCIAL MEDIA INDUSTRY

With the exponential increase in social media content and users, algorithms must also improve across all platforms to police content and ensure safety

NEXT TO COMPLETE

MODELS

Ongoing modeling process

ACCURACY

Obtain accuracy for individual models and comparison

ADDITIONS

Ensemble methods

BLOG

What we have done and conclusion of our research

THANKS!

Questions?

CREDITS

Chen, Yen-Chun. “UNITER: UNiversal Image-TExt Representation Learning.” *ArXiv*, Cornell University, Sept. 2019.

“Detectron2: A PyTorch-Based Modular Object Detection Library.” *Facebook AI*, https://ai.facebook.com/blog/-detectron2-a-pytorch-based-modular-object-detection-library-/. Accessed 26 Oct. 2021.

DrivenData. “Competition: Hateful Memes: Phase 1.” *DrivenData*, https://www.drivendata.org/competitions/64/hateful-memes/. Accessed 26 Oct. 2021.

DrivenDataOrg. “GitHub - Drivendataorg/Hateful-Memes.” *GitHub*, https://github.com/drivendataorg/hateful-memes/. Accessed 26 Oct. 2021.

“Hateful Memes Challenge and Dataset.” *Facebook AI*, http://ai.facebook.com/blog/hateful-memes-challenge-and-data-set/. Accessed 29 Nov. 2021.

HimariO. “GitHub - HimariO/HatefulMemesChallenge.” *GitHub*, https://github.com/HimariO/HatefulMemesChallenge. Accessed 26 Oct. 2021.

Kiela, Douw, et al. *The Hateful Memes Challenge: Competition Report*. 2020.

Li, Chunyuan. “Objects Are the Secret Key to Revealing the World between Vision and Language - Microsoft Research.” *Microsoft Research*, https://www.facebook.com/microsoftresearch/, 15 May 2020, https://www.microsoft.com/en-us/research/blog/objects-are-the-secret-key-to-revealing-the-world-between-vision-and-language/.

Li, Xiujun. “Oscar: Object-Semantics Aligned Pre-Training for Vision-Language Tasks.” *ArXiv*, Cornell University, Apr. 2020.

Microsoft. “GitHub - Microsoft/Oscar: Oscar and VinVL.” *GitHub*, https://github.com/microsoft/Oscar. Accessed 29 Nov. 2021.

Muennighoff. “GitHub - Muennighoff/Vilio: 🥶Vilio: State-of-the-Art VL Models in PyTorch & PaddlePaddle.” *GitHub*, https://github.com/Muennighoff/vilio. Accessed 26 Oct. 2021.

Research, Facebook. “GitHub - Facebookresearch/Detectron2: Detectron2 Is a Platform for Object Detection, Segmentation and Other Visual Recognition Tasks.” *GitHub*, https://github.com/facebookresearch/detectron2. Accessed 29 Nov. 2021.

CREDITS

---. “GitHub - Facebookresearch/Detectron2: Detectron2 Is FAIR’s next-Generation Platform for Object Detection, Segmentation and Other Visual Recognition Tasks.” *GitHub*, https://github.com/facebookresearch/detectron2. Accessed 26 Oct. 2021.

---. “Mmf/Projects/Hateful_memes at Main · Facebookresearch/Mmf · GitHub.” *GitHub*, https://github.com/facebookresearch/mmf/tree/main/projects/hateful_memes. Accessed 26 Oct. 2021.

Shulga, Dima. “BERT to the Rescue!. A Step-by-Step Tutorial on Simple Text... | by Dima Shulga | Towards Data Science.” *Medium*, Towards Data Science, 5 June 2019, https://towardsdatascience.com/bert-to-the-rescue-17671379687f.

Su, Weijie. “VL-BERT: Pre-Training of Generic Visual-Linguistic Representations.” *ArXiv*, Cornell University, Aug. 2019.

“VisualBERT — Transformers 4.11.3 Documentation.” *Hugging Face – The AI Community Building the Future.*, https://huggingface.co/transformers/model_doc/visual_bert.html. Accessed 26 Oct. 2021.

Wiggers, Kyle. “AI Still Struggles to Recognize Hateful Memes, but It’s Slowly Improving | VentureBeat.” *VentureBeat*, VentureBeat, 1 Dec. 2020, https://venturebeat.com/2020/12/01/ai-still-struggles-to-recognize-hateful-memes-but-its-slowly-improving/.

Yu, Fei. “ERNIE-ViL: Knowledge Enhanced Vision-Language Representations Through Scene Graph.” *ArXiv*, Cornell University, June 2020.

APPENDIX