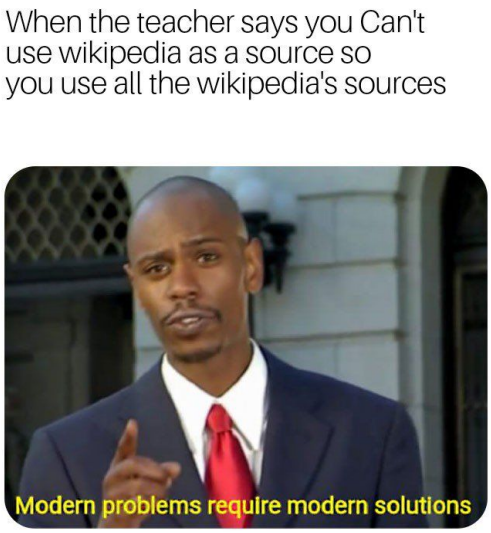
Literature Survey: Hateful Memes Classification

Team Number: 6

Team Members

Vaibhav Garg - 20171005
Akshay Goindani - 20171108
Preet Thakkar - 20171068
Sagar Joshi - 2020701007

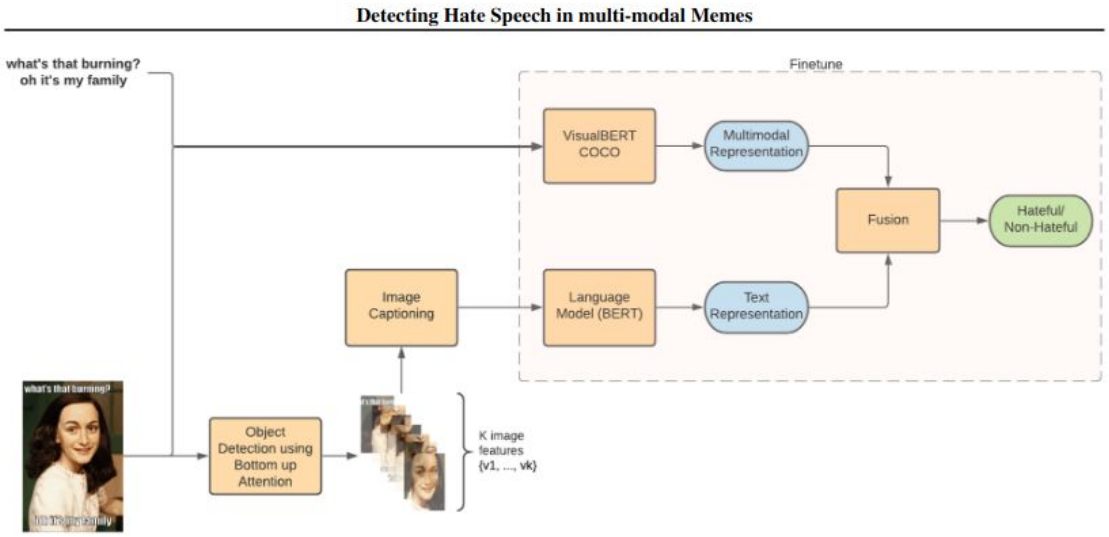
Understanding the Problem Statement



A meme, according to Wikipedia, is an idea, behavior, or style that becomes a fad and spreads by means of imitation from person to person within a culture and often carries symbolic meaning representing a particular phenomenon or theme. In recent times, memes are mostly created for humor, and some of them are capable of being hateful under the combination of pictures and text. Hence, hateful memes detection becomes a crucial task in this era of social media. This problem is different and harder than conventional multimodal tasks because most of the time meme content can only be understood by comprehension of the visual and text content together. Considering only images or only text is insufficient. Technically, it is a binary classification problem that tells whether a meme is hateful or not.

Relevant Research Work

1. Detecting Hate Speech in multimodal Memes

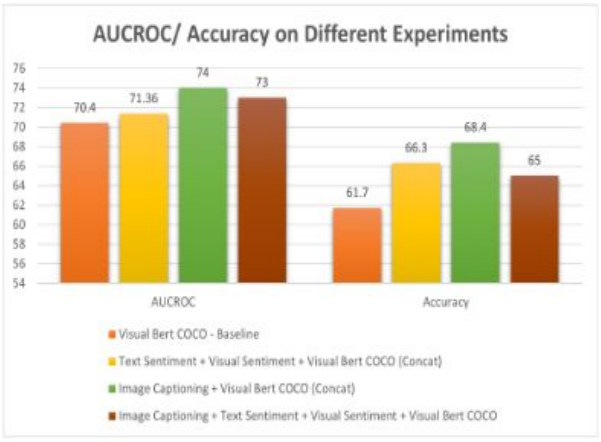


i. In this paper, it was observed that the majority of the data points in the Facebook Meme Dataset which are originally hateful are turned into benign just by describing the image of the meme. This is done because in such a scenario, unimodal models that focus only on the textual or visual modality will fail and only the multimodal models will be able to learn true reasoning.

ii. Many multimodal models give more preference to hate speech (language modality). Therefore image captioning and object detection are used to focus more on the visual modality. Using image captioning and object detection, the actual caption of the meme is extracted and the features from the actual caption are combined with multimodal representation to make the final prediction.

iii. In addition to the above method, unimodal sentiment features (for both image and text) are also used along with the multimodal features from pre-trained networks. This is done to capture the context and relationship between the two modalities.

iv. **Results**



2. Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches: Prize-winning solution to Hateful Memes Challenge

i. In this paper, visualBERT is used to get multimodal representation. The approach is divided into 4 phases - dataset expansion, image encoding, training, and ensemble learning. In the dataset expansion phase, some memes from the Memotion Dataset are selected. The selection is based on the similarity of the memes with the ones in the Hateful Meme Dataset.

ii. In the Image Encoding phase, 2048 dimensional region-based image features are extracted using ResNeXT-152 based Mask-RCNN model. These features are then projected to the textual embedding space.

iii. A pre-trained visualBert model is used to get the multimodal representations. The visualBert model is fine-tuned during the training phase.

iv. For the alignment between the textual modality and visual modality, the self-attention from the transformer model is used. Image regions and the language are combined and used as an input to the transformer.

v. Classification is done using softmax over the output of a linear transformation (Wx + b) on the output of the transformer.

vi. For ensembling, different models are taken and a majority voting strategy is used to predict the final class label.

vii. **Results**

Type	Model	Validation Acc.	Validation AUROC	Test Acc.	Test AUROC
	Human	-	-	84.70	82.65
Baselines	ViLBERT	62.20	71.13	62.30	70.45
	VisualBERT	62.10	70.60	63.20	71.33
	ViLBERT CC	61.40	70.07	61.10	70.03
	VisualBERT COCO	65.06	73.97	64.73	71.41
Ours	Ensemble	-	-	**76.50**	**81.08**
	Best ensemble model	**70.93**	**75.21**	-	-

3. Classification of Multimodal Hate Speech - The Winning Solution of Hateful Memes Challenge

i. In this paper, an ad-hoc classification algorithm is proposed. Some rules are extracted from the training set and are used to generate pseudo labels and adjust classification probability.

ii. After analyzing the training set, the samples are clustered into 4 categories:
   i. 3-tuple: It consists of 3 memes, the first meme has the same image as the second meme, it also has the same text with the third meme, but the second meme and the third meme does not seem to be related.
   ii. 2-tuple: It consists of 2 memes, they have the same image or the same text, and neither of them belongs to "3-tuple".
   iii. Unimodal hate: It is an image or text that appears more than once in the training set, and the labels of all memes that contain this image or text are hateful.
   iv. Others: It consists of all memes that are not present in either of the above three categories.

iii. According to statistical analysis, the labels of “3-tuple” are 1, 0, and 0, where 1 represents hateful and 0 represents non-hateful, and most of the labels of “2-tuple” are 1 and 0. From the analysis above, the following rules are extracted from the training set:
   i. For samples in "3-tuple", adjust the hateful probabilities to (1,0,0).
   ii. For samples in "2-tuple", adjust the hateful probabilities to (1,0), where the sample has a larger hateful probability adjusted to 1.

   They implement rule 1 to get the label of "3-tuple" in the test set, then merge these samples into the training set, and implement rule 2 to adjust the hateful probability of "2-tuple".

iv. VisualBert with Masked Region Modeling(MRM) is used as a base model. For this, four pre-trained models provided by Multimodal Modular Framework(MMF) were used. Model stacking with K-fold classification is implemented to combine the results of all four models.

v. **Results:** This approach was the winner of Facebook’s Hateful Memes Challenge. An AUROC score of 0.923 and accuracy of 86.8% was achieved. These results are indeed remarkable as the human accuracy for this task is 84.7%.

4. Detecting Hateful Memes Using a Multimodal DeepEnsemble

i. The first main idea in this paper is to leverage the benefits provided by the Transformer based pre-trained models which have been trained on very large datasets from a wide range of domains. These models have been extremely powerful and flexible on NLP tasks and can do a good job at fine-tuning and benefiting the downstream tasks. Also, it isn’t a good idea to train a complex network from scratch on such a small dataset.

ii. The second important idea proposed in this paper is to generate captions from the meme images, these captions are generated using the Show and Tell Model which has been trained on a different corpus. This way for each meme, we have two texts, one is the original meme text and the other is this new caption, now cross attention can be used between these image-text pairs for richer semantics.

iii. **Results**:

Type	Model	Validation Acc.	Validation AUROC	Test Acc.	Test AUROC
	Human	–	–	84.70	82.65
Unimodal	Image-Grid	52.73	58.79	52.00	52.63
	Image-Region	52.66	57.98	52.13	55.92
	Text BERT	58.26	64.65	59.20	65.08
Multimodal (Unimodal Pretraining)	Late Fusion	61.53	65.97	59.66	64.75
	Concat BERT	58.60	65.25	59.13	65.79
	MMBT-Grid	58.20	68.57	60.06	67.92
	MMBT-Region	58.73	71.03	60.23	70.73
	ViLBERT	62.20	71.13	62.30	70.45
	Visual BERT	62.10	70.60	63.20	71.33
Multimodal (Multimodal Pretraining)	ViLBERT CC	61.40	70.07	61.10	70.03
	Visual BERT COCO	65.06	73.97	64.73	71.41
(Phase 1)	UNITER	–	–	**68.70**	**74.14**
(Phase 1)	$\text{UNITER}_{\text{PA}}$	–	–	**68.30**	**75.29**
(Phase 1)	$\text{UNITER}_{\text{PA}}$ Ensemble	–	–	**66.60**	**76.81**
(Phase 2)	VL-BERT + $\text{UNITER}_{\text{PA}}$	**74.53**	**75.94**	**73.90**	**79.21**
(Phase 2)	$\text{UNITER}_{\text{PA}}$ Ensemble	**72.50**	**79.39**	**74.30**	**79.43**

5. Hate Speech in Pixels: Detection of Offensive Memes towards Automatic Moderation

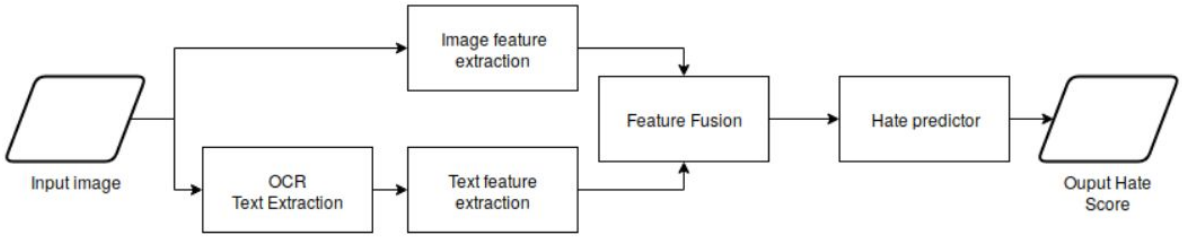


Figure 1: Block diagram of the system

i. This paper introduces a small dataset of 5020 memes, however, the labeling hasn't been done in a very precise and controlled fashion.

ii. Their model is a rather simple one, it processes images using a pre-trained VGG net to get a feature representation and the text is extracted using OCR which is then processed using BERT to get a feature embedding for the sentence. The visual and textual embeddings are concatenated to get a combined representation which is finally fed through MLP to output a hate score.

iii. **Results**:

Table 1: Accuracy results for the three configurations

Model	Max. Accuracy	Smth. Max. Accuracy
Multimodal	0.833	0.823
Image	0.830	0.804
Text	0.761	0.750

6. Hateful Memes Detection via Complementary Visual and Linguistic Networks

i. This paper proposes a solution where it tries to enhance the multi-modal features by introducing a complementary visual and linguistic (CVL) network. The reason behind introducing this network is that image representations with just image features lack the contextual information which can be obtained through the linguistic part, similarly, linguistic features lack the object-level information. Therefore, combining both the representations using some network will enhance the multi-modal features. Also, a lot of extra features are extracted from both image and text to capture contextual level and object-level information.

ii. Image representations are made using 3 different components - object-level representation (objects corresponding to a region of interests (ROIs)), position embedding of the ROIs, and the whole contextual information. For object-level information, the Faster-RCNN model is used as a feature extractor, which is used to obtain ROIs representation. For contextual information, ResNet-152 is used as a feature extractor. The final image representation is the pointwise addition of all

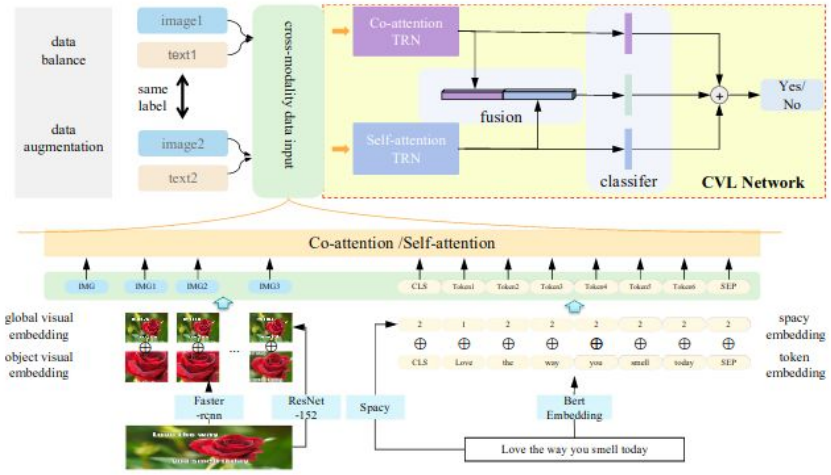


Figure 1: The overall framework of CVL network for hateful memes detection.

three representations.

iii. Text representations are also composed of 2 components. The first component is the word embedding extracted from BERT. The second component is the noun phrase embedding extracted using the spacy toolkit. From the given text, noun

phrases are extracted, then stop words are removed from these phrases and the remaining words in the phrases are known as Spacy Keywords. Using these spacy keywords from the text, the embeddings are extracted.

iv. Finally, the image and text representations are combined using the CVL network. In this work, they have used ViLBERT and VisualBERT with co-attention and self-attention, respectively. The image and text representations are inputs to these 2 models. Before making the final classification, the output of these models is concatenated.

v. **Results:**

Types	Models	Results (%) Acc.	AUROC
	Human	84.70	82.65
Uni-modal	Image-Grid	52.00	52.63
	Image-Region	52.13	55.92
	Text BERT	59.20	65.08
Multi-modal (Uni-modal Pre-training)	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
Multi-modal (Multi-modal Pre-training)	ViLBERT CC	61.10	70.03
	Visual BERT COCO	64.73	71.41
Multi-modal (Uni-modal Pre-training)	ViLBERT with Contextual and Sembedding	65.30	**75.22**
	VisualBERT with Contextual and Sembedding	64.30	72.85
	CVL	**66.20**	75.02

7. A Multimodal Framework for the Detection of Hateful Memes

i. This paper proposes a multimodal framework for hateful meme detection, which improves the performance of current multimodal approaches. They also show the effectiveness of upsampling benign confounders in the dataset to encourage multimodality. Also, ensembling is used to improve the robustness of the model.

ii. Most existing multimodal approaches are either late fusion or early fusion. In late fusion models, the image and text modalities are processed independently and then both the representations are combined using some network or simple concatenation before the final classification layer. In early fusion models, complex models are used to process both image and text jointly. Early fusion models perform better than the late fusion models. Multiple early fusion models such as LXMERT, UNITER and Oscar are present, but from their experiments they have shown that UNITER outperforms both LXMERT and Oscar. Moreover, the base model of UNITER (less number of parameters) performs better than the large counterpart because of better generalization and less overfitting.

iii. They extract the image features using the Faster-RCNN model and for text features they use pre-trained BERT. Using the image and text features, these models are fine-tuned on the hateful meme dataset. The UNITER-base model

was also fine-tuned using the supervised binary classification on the hateful meme dataset.

iv. There are two types of benign confounders - image confounders and text confounders. In image confounders, 2 samples have the same images, and the text is modified in such a way that the label of the 2 samples differ. Similarly, text confounders are defined. From their observations, performance on text confounders is poor, therefore their major focus is on upsampling text confounders. The upsampling takes place during fine-tuning the UNITER-base model. Also, they have modified the cross-entropy loss because the hateful meme dataset is imbalanced (36% hateful memes and 64% non-hateful memes), they have used re-weighting strategy where higher weight is given to the loss of hateful classes and lower weight is given to the loss of non-hateful classes. The sum of these 2 weights is 1.

$$\mathcal{L}_{HW}(\boldsymbol{\theta}) = \sum_{i=1}^{N} -\left[\alpha_{pos} \cdot y_i \log f_{\boldsymbol{\theta}}(x_i) + \alpha_{neg} \cdot (1 - y_i) \log\left(1 - f_{\boldsymbol{\theta}}(x_i)\right)\right]$$

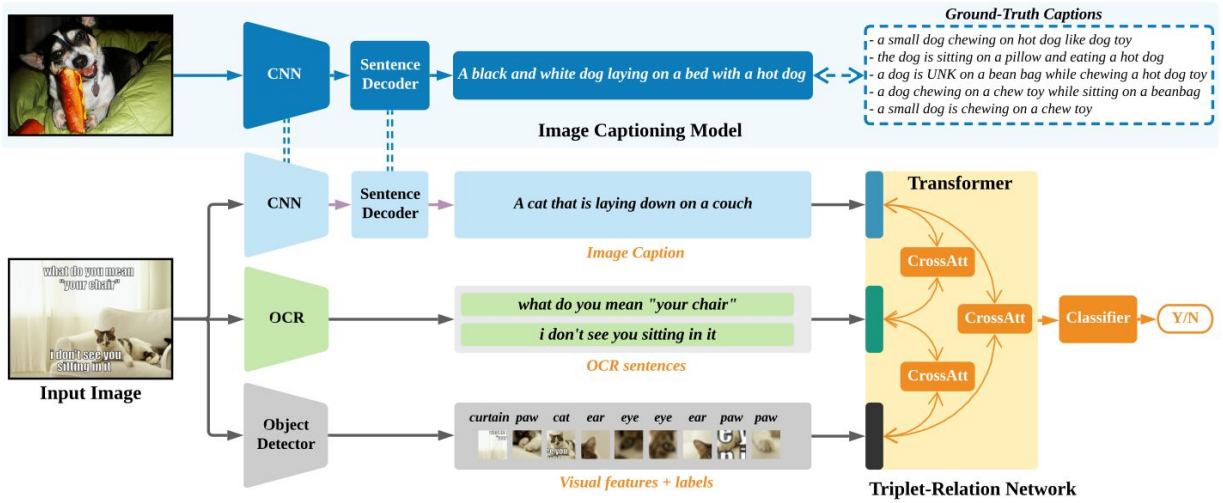
v. Since, the multimodal datasets have very few samples therefore they have used an ensemble of UNITER-base models for better generalization. The weight of these models is defined using evolutionary algorithms.

vi. Along with the loss defined above, they have added another loss term. The second loss term is related to the margin ranking loss. For each sample (x) , its text confounder (x1) is used, the training is performed on the pair (x, x1). The objective of the margin ranking loss is to predict a higher probability score for the meme that is labeled as hateful. At testing time, inference is done without pairing.

vii. **Results:**

Model	AUROC		
	Development	Phase 1	Phase 2
ViLBERT CC	70.07	70.03	–
VisualBERT COCO	73.97	71.41	–
UNITER	78.04	74.73	–
LXMERT	72.33	–	–
Oscar	72.00	–	–
$\text{UNITER}_{\text{CV10}}$	79.81	–	–
$\text{UNITER}_{\text{CV10 + CFU}}$	79.64	–	–
$\text{UNITER}_{\text{CV10 + CFU + HW}}$	80.01	78.60	–
$\text{UNITER}_{\text{CV15 + CFU + HW}}$	80.65	79.06	–
$\text{UNITER}_{\text{CV30 + CFU + HW}}$	81.36	78.98	–
$\text{UNITER}_{\text{CV15 + CFU + HW + MRL}}$	80.44	78.14	–
$\text{UNITER}_{\text{CV15 + CFU + HW + YOLO}}$	80.67	78.21	–
$\text{UNITER}_{\text{ENSEMBLE 1}}$	81.71	79.13	80.33
$\text{UNITER}_{\text{ENSEMBLE 2}}$	81.76	79.10	80.40
$\text{UNITER}_{\text{FINAL}}$	77.39	79.07	**80.53**

8. Multimodal Learning for Hateful Memes Detection



i. In this paper, the proposed model consists of an image captioner, an object detector, a triplet-relation network, and a classifier. They consider three different knowledge extracted from each meme: image caption, meme text, and visual features (using object detection).

ii. The proposed triplet-relation network models the triplet-relationships among caption, objects, and OCR sentences, adopting the cross-attention model to learn the more discriminative features from cross-modal embeddings.

iii. Sentence embeddings are generated by first extracting the image features with an image encoder and then visual features are decoded into a sentence using a sentence decoder. Finally, image caption embeddings and OCR text embeddings are concatenated. Image embeddings are generated using a pre-trained Faster R-CNN.

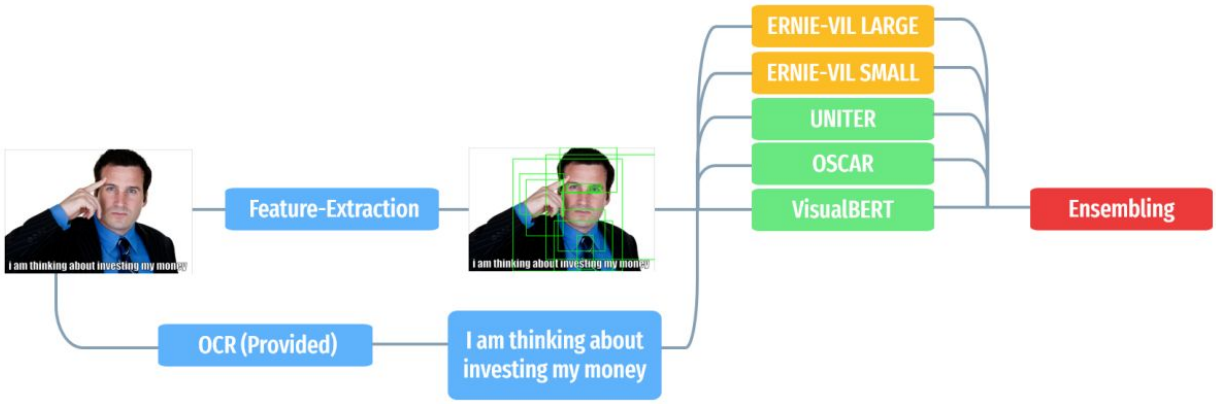
iv. The triplet-relation network is essentially a transformer network that is used to model the cross-modality relationships between image features and two textual features.

v. A joint representation of the textual and visual content obtained from the transformer model is fed to a fully connected layer followed by a softmax layer to get the prediction probability. A binary cross-entropy (BCE) loss function is used as the final loss function for meme detection.

vi. **Results:**

Inputs	Model	AUROC (Test)
	Human [36]	82.65
Image	Image-Grid [36]	52.63
	Image-Region [36]	55.92
Text	Text BERT [36]	65.08
Image + Text	Late Fusion	64.75
	Concat BERT [36]	65.79
	MMBT-Grid [36]	67.92
	MMBT-Region [36]	70.73
	ViLBERT [36]	70.45
	Visual BERT [36]	71.33
	ViLBERT CC [36]	70.03
	Visual BERT COCO [36]	71.41
Image + Text + Caption	Ours (V+L)	**73.42**
	Ours (V&L)	**74.80**

9. Vilio: State-of-the-art Visio-Linguistic Models Applied To Hateful Memes

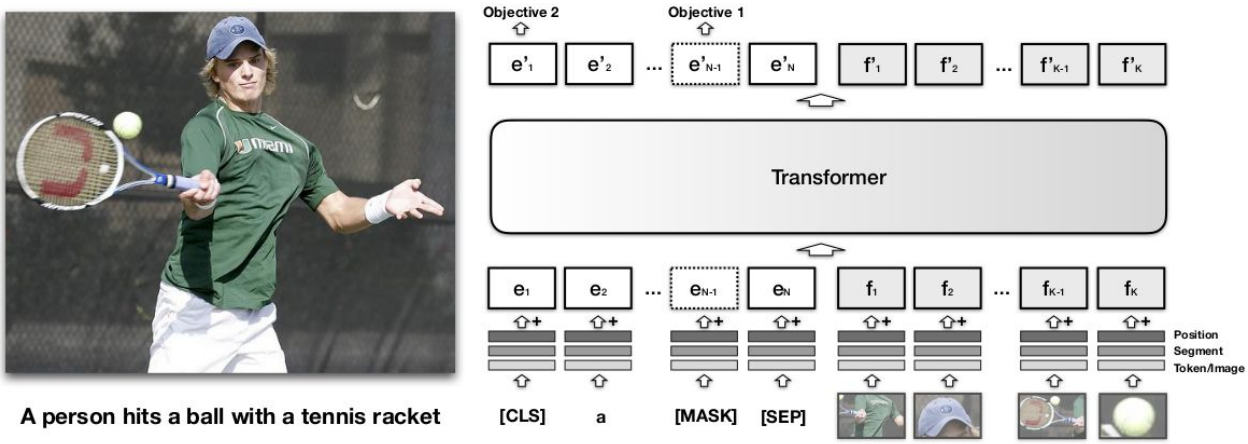


i. In this paper, the proposed model uses an ensemble technique.
ii. In the first step, image features are extracted using the detectron2 framework.
iii. Next, meme text together with the features are fed into the ensemble models.
iv. The ensemble contains five models: ERNIE-ViL (small, large), OSCAR, UNITER, and VisualBERT.
v. Vilio model is able to achieve an AUROC score of 0.8252 on the test set.
vi. **Results:**

Source	Model	**Validation** AUROC	**Test** AUROC
Hateful Memes Baseline	Human	-	82.65
	ViLBERT	71.13	70.45
	VisualBERT	70.60	71.33
	ViLBERT CC	70.07	70.03
	VisualBERT COCO	73.97	71.41
Vilio	VisualBERT	75.49	75.75
	OSCAR	77.16	77.30
	UNITER	77.75	78.65
	ERNIE-ViL Base	78.18	77.02
	ERNIE-ViL Large	78.76	80.59
	Ensemble	**81.56**	**82.52**

10. VisualBERT: A Simple and Performant Baseline for Vision and Language

i. A transformer-based framework for dealing with text and image data is proposed, which proves to be better or on par with other state of the art methods, even though being significantly simpler.
ii. The transformer model performs an early fusion of the image and text embeddings. An ablation study shows the choice of early fusion to be important for multiple interactions between vision and language.



iii. The training of the VisualBERT model was performed on a BERT initialization with the following steps:

i. Task-agnostic pre-training - It consisted of two training objectives on the COCO dataset:
   a) Masked language modeling on the text component.
   b) Sentence-image prediction, where given an image and a true caption, the model determines whether another given caption is relevant or not.

ii. Task-specific pre-training - Masked language modeling is performed using image objective of the target domain.

iii. Fine-tuning - The transformer performance is maximised on task-specific training objective.

	Model	Dev
	VisualBERT	66.7
C1	VisualBERT w/o Grounded Pre-training	63.9
	VisualBERT w/o COCO Pre-training	62.9
C2	VisualBERT w/o Early Fusion	61.4
C3	VisualBERT w/o BERT Initialization	64.7
C4	VisualBERT w/o Objective 2	64.9

Table 5: Performance of the ablation models on NLVR^2. Results confirm that task-agnostic pre-training (C1) and early fusion of vision and language (C2) are essential for VisualBERT.

iv. Experiments on four vision-and-language tasks were performed:

i. VQA - Correctly answer the question, given an image and a question.

Model	Test-Dev	Test-Std
Pythia v0.1 (Jiang et al., 2018)	68.49	-
Pythia v0.3 (Singh et al., 2019)	68.71	-
VisualBERT w/o Early Fusion	68.18	-
VisualBERT w/o COCO Pre-training	70.18	-
VisualBERT	70.80	71.00
Pythia v0.1 + VG + Other Data Augmentation (Jiang et al., 2018)	70.01	70.24
MCAN + VG (Yu et al., 2019b)	70.63	70.90
MCAN + VG + Multiple Detectors (Yu et al., 2019b)	72.55	-
MCAN + VG + Multiple Detectors + BERT (Yu et al., 2019b)	72.80	-
MCAN + VG + Multiple Detectors + BERT + Ensemble (Yu et al., 2019b)	75.00	75.23

Table 1: Model performance on VQA. VisualBERT outperforms Pythia v0.1 and v0.3, which are tested under a comparable setting.

ii. VCR - It is composed of two tasks, question answering and answer justification, with questions focusing on visual common-sense.

Model	Q → A		QA → R		Q → AR	
	Dev	Test	Dev	Test	Dev	Test
R2C (Zellers et al., 2019)	63.8	65.1	67.2	67.3	43.1	44.0
B2T2 (Leaderboard; Unpublished)	-	72.6	-	75.7	-	55.0
VisualBERT w/o Early Fusion	70.1	-	71.9	-	50.6	-
VisualBERT w/o COCO Pre-training	67.9	-	69.5	-	47.9	-
VisualBERT	70.8	71.6	73.2	73.2	52.2	52.4

Table 2: Model performance on VCR. VisualBERT w/o COCO Pre-training outperforms R2C, which enjoys the same resource while VisualBERT further improves the results.

iii. NLVR - Given a pair of images and a caption, determine if the caption is true for the images.

Model	Dev	Test-P	Test-U	Test-U (Cons)
MaxEnt (Suhr et al., 2019)	54.1	54.8	53.5	12.0
VisualBERT w/o Early Fusion	64.6	-	-	-
VisualBERT w/o COCO Pre-training	63.5	-	-	-
VisualBERT	67.4	67.0	67.3	26.9

Table 3: Comparison with the state-of-the-art model on NLVR^2. The two ablation models significantly outperform MaxEnt while the full model widens the gap.

iv. Flickr30K Entities - Given spans from a sentence and an image, select the bounding regions from the image they correspond to.

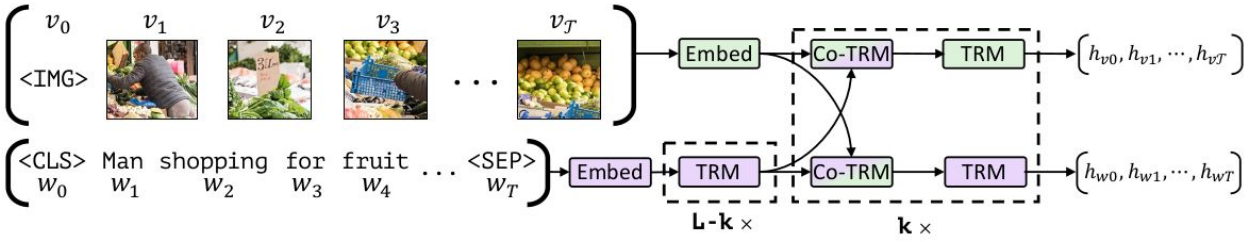
Model	R@1		R@5		R@10		Upper Bound	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test
BAN (Kim et al. 2018)	-	69.69	-	84.22	-	86.35	86.97	87.45
VisualBERT w/o Early Fusion	70.33	-	84.53	-	86.39	-		
VisualBERT w/o COCO Pre-training	68.07	-	83.98	-	86.24	-	86.97	87.45
VisualBERT	70.40	71.33	84.49	84.98	86.31	86.51		

Table 4: Comparison with the state-of-the-art model on the Flickr30K. VisualBERT holds a clear advantage over BAN.

11. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks

i. A transformer-based approach which aims to jointly represent images and text with two parallel BERT-style models is proposed.

ii. The architecture consists of novel co-attentional transformer layers which enable information exchange between the two modalities by passing the keys and values of one modality to the other modality's multi-headed transformer block.



iii. The task-agnostic pre-training is done using Conceptual Captions dataset and consists of two tasks:

i. Masked multi-modal modelling - 15% of word and image region inputs are masked, with text inputs handled the same way as BERT. For masked image regions, the model is trained to output a probability distribution over semantic classes corresponding to the input region.

ii. Multi-modal alignment task - An image-text pair is given, and the model has to make a binary prediction if they are aligned or not.

iv. Four vision-and-language tasks are used to illustrate transfer learning, two of which are VQA and VCR. The other two tasks are as below:

i. Ground referring expressions - Given a natural language referring expression, predict the localized image region.

ii. Caption-based image retrieval - Determine the correct image from a pool for a particular caption input.

Table 1: Transfer task results for our ViLBERT model compared with existing state-of-the-art and sensible architectural ablations. † indicates models without pretraining on Conceptual Captions. For VCR and VQA which have private test sets, we report test results (in parentheses) only for our full model. Our full ViLBERT model outperforms task-specific state-of-the-art models across all tasks.

		VQA [3]	VCR [25]			RefCOCO+ [32]			Image Retrieval [26]			ZS Image Retrieval		
	Method	test-dev (test-std)	Q→A	QA→R	Q→AR	val	testA	testB	R1	R5	R10	R1	R5	R10
SOTA	DFAF [36]	70.22 (70.34)	-	-	-	-	-	-	-	-	-	-	-	-
SOTA	R2C [25]	-	63.8 (65.1)	67.2 (67.3)	43.1 (44.0)	-	-	-	-	-	-	-	-	-
SOTA	MAttNet [33]	-	-	-	-	65.33	71.62	56.02	-	-	-	-	-	-
SOTA	SCAN [35]	-	-	-	-	-	-	-	48.60	77.70	85.20	-	-	-
Ours	Single-Stream†	65.90	68.15	68.89	47.27	65.64	72.02	56.04	-	-	-	-	-	-
Ours	Single-Stream	68.85	71.09	73.93	52.73	69.21	75.32	61.02	-	-	-	-	-	-
Ours	ViLBERT†	68.93	69.26	71.01	49.48	68.61	75.97	58.44	45.50	76.78	85.02	0.00	0.00	0.00
Ours	ViLBERT	**70.55 (70.92)**	**72.42 (73.3)**	**74.47 (74.6)**	**54.04 (54.8)**	**72.34**	**78.52**	**62.61**	**58.20**	**84.90**	**91.52**	**31.86**	**61.12**	**72.80**

Table 2: Ablation study of the depth of our model with respect to the number of Co-TRM→TRM blocks (shown in a dashed box in Fig. 1). We find that different tasks perform better at different network depths – implying they may need more or less context aggregation.

	VQA [3]	VCR [25]			RefCOCO+ [32]			Image Retrieval [26]			ZS Image Retrieval [26]		
Method	test-dev	Q→A	QA→R	Q→AR	val	testA	testB	R1	R5	R10	R1	R5	R10
ViLBERT (2-layer)	69.92	72.44	**74.80**	**54.40**	71.74	**78.61**	62.28	55.68	84.26	90.56	26.14	56.04	68.80
ViLBERT (4-layer)	70.22	**72.45**	74.00	53.82	72.07	78.53	**63.14**	55.38	84.10	90.62	26.28	54.34	66.08
ViLBERT (6-layer)	**70.55**	72.42	74.47	54.04	**72.34**	78.52	62.61	58.20	84.90	**91.52**	31.86	61.12	72.80
ViLBERT (8-layer)	70.47	72.33	74.15	53.79	71.66	78.29	62.43	**58.78**	**85.60**	91.42	**32.80**	**63.38**	**74.62**

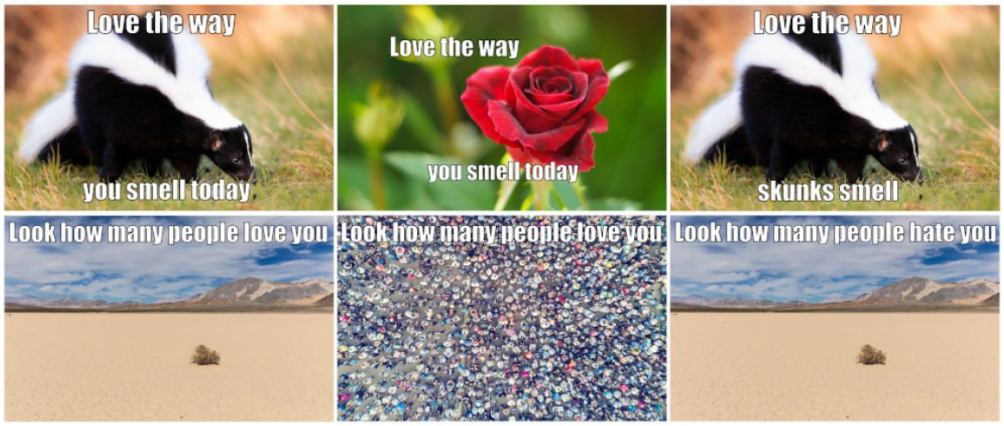
Table 3: Transfer task results for ViLBERT as a function of the percentage of the Conceptual Captions dataset used during pre-training. We see monotonic gains as the pretraining dataset size grows.

	VQA [3]	VCR [25]			RefCOCO+ [32]			Image Retrieval [26]			ZS Image Retrieval [26]		
Method	test-dev	Q→A	QA→R	Q→AR	val	testA	testB	R1	R5	R10	R1	R5	R10
ViLBERT (0 %)	68.93	69.26	71.01	49.48	68.61	75.97	58.44	45.50	76.78	85.02	0.00	0.00	0.00
ViLBERT (25 %)	69.82	71.61	73.00	52.66	69.90	76.83	60.99	53.08	80.80	88.52	20.40	48.54	62.06
ViLBERT (50 %)	70.30	71.88	73.60	53.03	71.16	77.35	61.57	54.84	83.62	90.10	26.76	56.26	68.80
ViLBERT (100 %)	**70.55**	**72.42**	**74.47**	**54.04**	**72.34**	**78.52**	**62.61**	**58.20**	**84.90**	**91.52**	**31.86**	**61.12**	**72.80**

Available Datasets

1. **[The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes]**
   This paper by Facebook AI introduces the Hateful Memes Challenge Dataset, it goes through the detailed process in which these memes have been collected and annotated with a very precise definition of “hate speech”. It consists of a total of 10,000 memes with

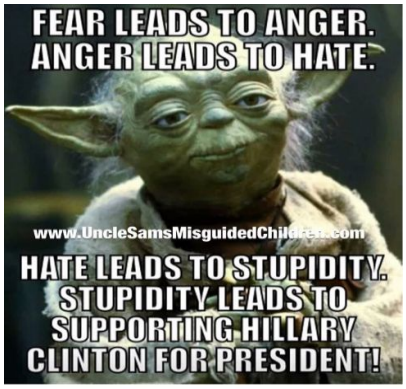
each meme corresponding to an image and associated text. The dataset focuses strongly on the multimodal aspect of the problem and therefore introduces examples that cannot possibly be solved using a unimodal approach as shown below:



Type	Model	Validation Acc.	Validation AUROC	Test Acc.	Test AUROC
	Human	-	-	84.70	82.65
Unimodal	Image-Grid	52.73	58.79	52.00±1.04	52.63±0.20
	Image-Region	52.66	57.98	52.13±0.40	55.92±1.18
	Text BERT	58.26	64.65	59.20±1.00	65.08±0.87
Multimodal (Unimodal Pretraining)	Late Fusion	61.53	65.97	59.66±0.64	64.75±0.96
	Concat BERT	58.60	65.25	59.13±0.78	65.79±1.09
	MMBT-Grid	58.20	68.57	60.06±0.97	67.92±0.87
	MMBT-Region	58.73	71.03	60.23±0.87	70.73±0.66
	ViLBERT	62.20	71.13	62.30±0.46	70.45±1.16
	Visual BERT	62.10	70.60	63.20±1.06	71.33±1.10
Multimodal (Multimodal Pretraining)	ViLBERT CC	61.40	70.07	61.10±1.56	70.03±1.77
	Visual BERT COCO	65.06	73.97	64.73±0.50	71.41±0.46

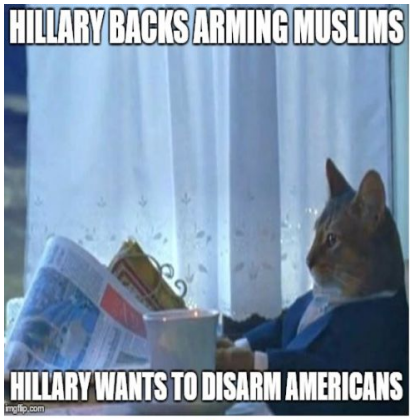
2. **Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text**

   This dataset uses images and pre-processed captions from an existing Kaggle dataset of memes specific to the 2016 US Presidential Elections and annotates them as hateful/non-hateful using 8 annotators having agreement scores in the range of 0.4-0.5. After considering both - image and caption, memes intended to be a personal attack, homophobic/racial/minority abuse are labeled offensive. Some of such examples are as below.



(b) Example of meme intended for personal attack.

(c) Example of meme intended for Homophobic abuse



(b) Example of meme intended for attacking minorities

(a) Example of meme intended for Racial abuse

The dataset contains a total of 743 samples with 445 training samples (187 offensive v/s 258 non-offensive) and 149 samples (59 offensive v/s 90 non-offensive) each in the validation and test sets. The paper runs through several approaches applied for the classification task, separately for text-based classification, image-based classification, and then three multimodal approaches using an early fusion of image and text embeddings. Among all approaches, CNN on the text was seen to have the best recall value of 84% and a decent precision score of 39%. A classifier having just VGG-16 for image data had the lowest recall score of 16%. Multimodal approaches were seen to have a fair precision v/s recall balance as compared to other models, but recall not reaching as high as CNN on the text. The best precision scores, however, were produced by basic Logistic Regression and Naive Bayes models dealing with just text data.

Type	Classifier	P	R	F
Text	LR	0.58	0.40	0.48
	NB	0.52	0.45	0.49
	DNN	0.47	0.54	0.50
	Stacked LSTM	0.39	0.42	0.40
	BiLSTM	0.42	0.23	0.30
	CNN	0.39	0.84	0.54
Image	VGG16	0.41	0.16	0.24
Multi	Stacked LSTM + VGG16	0.40	0.66	0.50
	BiLSTM + VGG16	0.40	0.44	0.41
	CNNText + VGG16	0.38	0.67	0.48

Table 2: Precision, recall and F1-score for the baseline and multimodal classifiers.

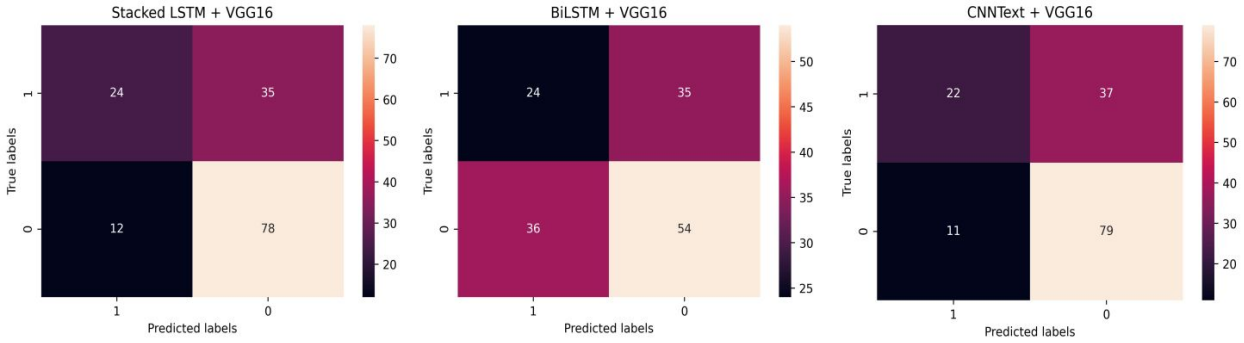


Figure 3: Confusion matrix for Multimodal classifier with Stacked LSTM, BiLSTM and CNN.