TeamOtter at SemEval-2022 Task 5: Detecting Misogynistic Content in Multimodal Memes

Paridhi Maheshwari [*]
Stanford University
paridhi@stanford.edu

Sharmila Reddy Nangi [*]
Stanford University
srnangi@stanford.edu

Abstract

We describe our system for the SemEval 2022 task on detecting misogynous content in memes. This is a pressing problem and we explore various methods ranging from traditional machine learning to deep learning models such as multimodal transformers. We propose a multimodal BERT architecture that uses information from both image and text. We further incorporate common world knowledge from pretrained CLIP and Urban dictionary. We also provide qualitative analysis to support out model. Our best performing model achieves an F1 score of 0.679 on Task A (Rank 5) and 0.680 on Task B (Rank 13) of the hidden test set. Our code is available at https://github.com/paridhimaheshwari2708/MAMI.

1 Introduction

In this era of the internet, memes have become a new form of communication, which predominantly contain an image and a small caption. While their general purpose is to invoke humour or irony, they are also increasingly being used as a source of harmful, offensive and misogynistic content. Detecting such content in an automated manner is an important problem to avoid the spread of hate.

Memes pose a unique multimodal challenge as their underlying implication is not just a simple combination of the image and text, but a subtle inference that comes naturally to humans. Another complexity is that memes are highly contextual and the component image and text pieces might be completely uncorrelated. Understanding this fusion of modalities is a challenging task for machines. Our aim is to automatically identify misogynistic multimodal memes using machine learning.

2 Related Work

The task of identifying misogyny in memes is a relatively new area and is closely related to hate detection. While there has been a lot of work on identifying hateful content in unimodal data (Gandhi et al., 2019; Fortuna and Nunes, 2018), there is little work on multimodal hate detection. Recently, Facebook Hateful Memes Challenge (Kiela et al., 2020) explored fusion of text and vision models along with advanced architectures like cross-modal BERT (Lu et al., 2019). A major problem with these large pretrained models is the domain gap between memes and training data. Some works try to solve this with better pretraining (Zhu, 2020) and disentangling hate from meme representations (Lee et al., 2021). In this work, we build on these technologies for our specific use-case of misogyny detection and incorporate common world knowledge from Urban Dictionary (Wilson et al., 2020) and CLIP (Radford et al., 2021) to address the domain gap.

3 Method

3.1 Baselines

The task of detecting misogynistic content in memes can be posed as a classification task based on visual and textual features. We start with simple baselines, namely SVM, Naive Bayes and Logistic Regression, and also experiment with *unimodal* feature space, i.e, training classifiers with *text only* and *image only* features. For text only models, we incorporate the TF-IDF technique with bag-of-words concepts to compute features. To capture visual cues from images, we leverage pretrained VGG-16 (Simonyan and Zisserman, 2014) for feature extraction. Since memes are a complex combination of text and image, we require cues from both modalities and we therefore, move towards *multimodal* methods for classification.

3.2 Deep Learning Architectures

We leverage various Deep Learning (DL) techniques for this task. We first start with unimodal techniques, namely LSTM and CNN architectures.

[*] These authors contributed equally

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 642 - 647
July 14-15, 2022 ©2022 Association for Computational Linguistics

Model	Binary Classification				Multi-class Multi-label Classification			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Using Only Text Features								
Logistic Regression	**0.802**	**0.802**	**0.801**	**0.801**	**0.844**	**0.800**	0.455	0.524
SVM	0.794	0.807	0.790	0.790	0.838	0.673	0.556	**0.596**
Naive Bayes	0.794	0.795	0.795	0.794	0.827	0.707	0.363	0.371
Using Only Image Features								
Logistic Regression	0.634	0.634	0.634	0.634	0.762	0.462	0.399	0.425
SVM	0.631	0.631	0.631	0.631	0.767	0.468	0.385	0.417
Naive Bayes	0.633	0.632	0.632	0.632	0.677	0.441	0.558	0.480
Using Both Image and Text Features								
Logistic Regression	0.760	0.760	0.759	0.634	0.808	0.583	0.516	0.545
SVM	0.786	0.786	0.786	0.786	0.818	0.593	**0.577**	0.584

Table 1: Evaluation of various baselines.

For text, we use the GloVe (Pennington et al., 2014) embeddings to initialize individual words and pass this sequence through an LSTM layer. Finally, this embedding is fed to FC layers that outputs a score for each class. For image, we extract the feature representations from a pretrained VGG-16 (Simonyan and Zisserman, 2014) model and pass through a classifier head which is composed of FC layers. All models are trained end-to-end using binary cross entropy loss for every class independently. Note that we do not pose this as softmax classification as each meme can belong to multiple classes simultaneously, i.e., multi-label classification. To handle class imbalance in the dataset, we give more importance to the positive examples. Specifically, we weigh the positive component of the binary cross entropy loss with the ratio of negative to positive occurrences per class.

Since our data is inherently multimodal, we propose advanced DL methods that incorporate both textual and visual features. This is important because memes are complex entities and the fusion of both modalities is necessary to understand the full meaning of the meme (which might not be apparent from a single modality alone). We experiment with the following mulitmodal networks:

1. **CNN + LSTM**: This architecture does a simple late-fusion of the two unimodal designs. We concatenate image and text features and pass through a FC classifier for prediction.
2. **VQA**: There has been significant work in multimodal learning on Visual Question Answering, which requires subtle reasoning around both modalities to answer complex queries. Given similar reasoning in memes, we experiment with the VQA model (Antol et al., 2015). Both image and text (question) features are transformed to a common space and fused via element-wise product, which is then passed to a FC layer to get class scores (answers).
3. **MUTAN**: This model (Ben-Younes et al., 2017) tries to effectively mix and merge information from the two modalities. It uses a multimodal Tucker decomposition to efficiently parametrize bilinear interactions between visual and textual representations. It demonstrates improved performance on the visual question-answering task by learning interpretable embedding spaces.

Recently, Bidirectional Encoder Representations from Transformers (BERT) models (Devlin et al., 2019) trained on large-corpus have proven to provide state-of-the-art results for diverse NLP applications. Given an input sentence, a pretrained BERT model gives a hidden representation for each token in the sentence along a pooled output for the entire sentence. These representations are rich in contextual knowledge and we explore different ways to use this information as follows:

4. **Concat BERT**: The pooled output for text is concatenated with the image feature, and passed through a FC classifier.
5. **Average BERT**: Similar to the previous setting, but the average of the final hidden state is taken as the text feature.
6. **Gated BERT**: The final hidden state is averaged to get text feature. We combine the text and image feature using a Multimodal Gated Layer (Ovalle et al., 2017), which weights relevance of each modality and combines them to output prediction classes.

3.3 Common World Knowledge

Language in memes is informal and often contains slang words. We propose to use the Urban dictionary which is a crowd-sourced repository of common slangs along with their definitions. Particularly, we initialize our constituent words with embeddings pretrained on the Urban Dictionary (Wilson et al., 2020) instead of GloVe vectors. These features perform well on extrinsic tasks such as sentiment analysis and sarcasm detection where some knowledge of colloquial language is required.

Popular vision algorithms (such as VGG-16) are trained on object detection tasks and they require explicit supervision from labels. This limits their usability. More recently, pretraining on image-text matching (Radford et al., 2021) has gained traction by outperforming other methods. Since the images are crawled from the internet, we believe that the distribution captured by CLIP (Radford et al., 2021) are more relevant and representative of the online media today, and hence, more suitable for our task.

3.4 Joint Learning

In the previous sections, we were considering the two tasks independently and training separate models. Given the synergy between the two tasks, we propose a joint learning framework where we use weight sharing between networks to exploit the commonalities and learn improved features. We propose two approaches to achieve this:

1. **Multi-Task Learning**: We start with a multimodal deep network as a shared embedding layer for both modalities, and followed by two different classifier heads, one for each task.
2. **Hierarchical Learning**: We utilize the inherent hierarchy between the two tasks where the second classifier kicks in only when the probability for "misogynous" class from first classifier is greater than 0.5. The model architecture is same as the multi-task setup, but now the second classifier head for finer categorization is only trained on misogynous items.

4 Experiments and Results

4.1 Task and Dataset

We work on the "*Multimedia Automatic Misogyny Identification*" task (Fersini et al., 2022) at SemEval 2022. The problem comprises of two sub-tasks: (i) *Binary Classification* to categorize a given meme as misogynous or not; (ii) *Multi-class Multi-label Classification* to further classify misogynous memes into fine-grained, overlapping categories (shaming, stereotype, objectification, violence). Our dataset consists of 10,000 memes and we partition them into 70% / 20% / 10% for train, validation and test respectively. We only report metrics on this data split as we do not have the ground truth labels for the competition's hidden test set. We measure the performance using these metrics: average accuracy per class, and weighted-average precision, recall and F1 scores where the weights are determined by the support of that class.

4.2 Textual and Visual Cues

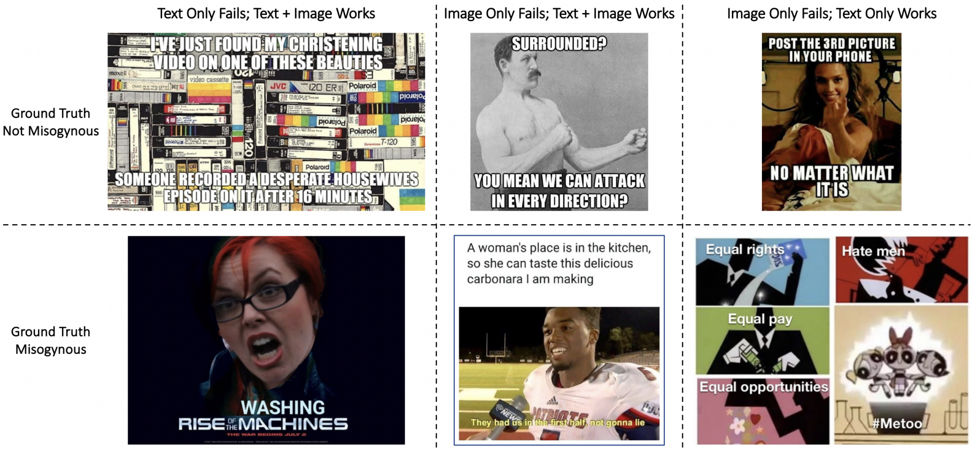
Prior work on detecting sexism in memes (Fersini et al., 2019) use specially curated textual and visual cues. We curated the profanity scores for text using a pretrained model on toxic comment classification (Pearson coeff. -0.05), sentiment polarity from Textblob (Pearson coeff. -0.012), and percentage of skin in images (Pearson coeff. 0.125). Thus, many intuitive cues showed no correlation with misogyny, exemplifying the difficulty of our task.

4.3 Baselines

Table 1 presents the baseline results. We extend these linear models to the multi-label setting as a *one-vs-all* task, where separate classifier are trained for each class. We observe the following: (i) Textual models perform better than image only models. (ii) Performance of text + image models is similar to text only methods, implying that TF-IDF vectors are a strong indicator for meme classification.

4.4 Deep Learning Architectures

The results are tabulated in Table 2 and we make the following observations: (i) Using both image and text significantly improves performance over the unimodal variants. We further provide qualitative comparison of unimodal and multimodal methods in Figure 1, which also illustrates the complexity of the task and the subtle relations between the two modalities. (ii) Similar to the baselines, text only methods give better results than image only methods. (iii) BERT-based models show significant improvement in performance for multi-label classification task. (iv) For the multi-class, multi-label classification problem, there is a skewed distribution of positive and negative examples within a class. Hence, performance varies across different classes, as shown in Figure 2. Here, training the


Text Only Fails; Text + Image Works
Image Only Fails; Text + Image Works
Image Only Fails; Text Only Works
Ground Truth Not Misogynous
Ground Truth Misogynous
I'VE JUST FOUND MY CHRISTENING VIDEO ON ONE OF THESE BEAUTIES
SOMEONE RECORDED A DESPERATE HOUSEWIVES EPISODE ON IT AFTER 16 MINUTES
SURROUNDED?
YOU MEAN WE CAN ATTACK IN EVERY DIRECTION?
POST THE 3RD PICTURE IN YOUR PHONE
NO MATTER WHAT IT IS
WASHING
RISE OF THE MACHINES
They had us in the first half, not gonna lie
Equal rights
Hate men
Equal pay
Equal opportunities
#Metoo

A woman's place is in the kitchen, so she can taste this delicious carbonara I am making

Figure 1: Qualitative evaluation of Text Only, Image Only and Text + Image models.

Model	**Binary Classification**				**Multi-class Multi-label Classification**			
	Accuracy	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**
Using Only Text Features								
LSTM	0.772	0.773	0.770	0.770	0.654	0.504	0.712	0.558
LSTM (CWK)	0.778	0.779	0.776	0.777	0.644	0.453	0.571	0.478
Using Only Image Features								
CNN	0.717	0.727	0.712	0.711	0.673	0.497	0.629	0.524
CNN (CWK)	0.839	0.838	0.838	0.838	0.769	0.580	0.790	0.654
Using Both Image and Text Features								
CNN + LSTM	0.799	0.798	0.799	0.798	0.700	0.522	0.761	0.598
VQA	0.785	0.789	0.787	0.784	0.669	0.503	0.821	0.599
MUTAN	0.821	0.821	0.820	0.820	0.639	0.483	**0.855**	0.594
Concat BERT	0.715	0.719	0.712	0.711	0.780	0.531	0.477	0.501
Average BERT	0.742	0.746	0.740	0.740	0.796	0.563	0.541	0.540
Gated BERT	0.728	0.728	0.727	0.727	0.800	0.587	0.523	0.553
CNN + LSTM (CWK)	0.836	0.837	0.834	0.835	0.770	0.585	0.798	0.658
VQA (CWK)	0.828	0.829	0.826	0.827	0.771	0.592	0.800	0.662
MUTAN (CWK)	0.828	0.827	0.827	0.827	0.781	0.607	0.791	0.670
Concat BERT (CWK)	**0.840**	**0.841**	**0.841**	**0.840**	0.798	0.628	0.723	0.658
Average BERT (CWK)	0.837	0.837	0.836	0.836	0.838	**0.655**	0.700	0.676
Gated BERT (CWK)	0.839	0.839	0.839	0.838	**0.845**	**0.665**	0.710	**0.684**

Table 2: Evaluation of various deep learning architectures. Here, CWK refers to common world knowledge sources, namely Urban Dictionary (Wilson et al., 2020) and CLIP (Radford et al., 2021), for text and image respectively.

Model	**Binary Classification**				**Multi-class Multi-label Classification**			
	Accuracy	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**
Multi-Task Learning								
CNN + LSTM (CWK)	0.832	0.833	0.830	0.831	0.750	0.598	0.810	0.662
VQA (CWK)	**0.845**	**0.845**	**0.844**	**0.844**	0.771	0.601	0.790	0.663
MUTAN (CWK)	0.836	0.835	0.835	0.835	0.769	0.585	0.803	0.658
Average BERT (CWK)	0.843	0.843	0.843	0.843	0.757	0.598	**0.816**	0.665
Concat BERT (CWK)	0.838	0.838	0.838	0.838	0.734	0.580	0.804	0.646
Gated BERT (CWK)	0.842	0.842	0.842	0.842	0.725	0.571	0.803	0.639
Hierarchical Learning								
CNN + LSTM (CWK)	0.839	0.838	0.839	0.838	0.773	0.590	0.804	0.663
VQA (CWK)	0.843	0.842	0.842	0.842	0.770	0.598	0.794	0.662
MUTAN (CWK)	0.829	0.828	0.829	0.828	0.772	0.587	0.777	0.653
Average BERT (CWK)	0.842	0.842	0.842	0.842	**0.799**	**0.613**	0.769	**0.671**
Concat BERT (CWK)	0.838	0.838	0.838	0.838	0.796	0.609	0.751	0.661
Gated BERT (CWK)	0.835	0.835	0.835	0.835	0.788	0.599	0.761	0.658

Table 3: Evaluation of various deep learning architectures using joint learning techniques.

models with weighted cross entropy plays an important role in the precision-recall trade off. Note that F1 score is a better measure than accuracy because of the class imbalance.

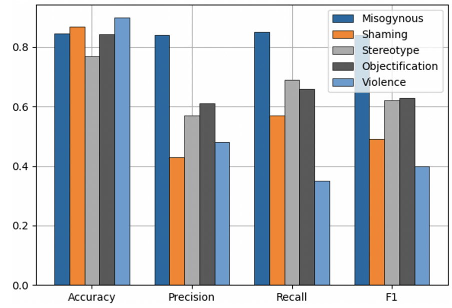
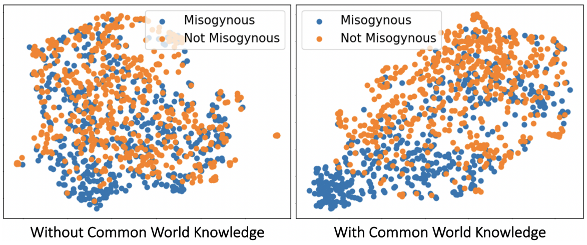


Figure 2: Class-wise performance across metrics.



Without Common World Knowledge

With Common World Knowledge

Figure 3: Multimodal features projected into two dimensional space along with misogynous-or-not labels.

4.5 Common World Knowledge

Incorporating Common World Knowledge (CWK) from Urban Dictionary and CLIP provides a substantial boost in performance across all models, and this is because visiolinguistic models are able to learn more discriminative features. To illustrate this, we consider two set of multimodal features (with and without CWK) and run dimensionality reduction using Uniform Manifold Approximation and Projection (McInnes et al., 2018). We visualize the feature space in lesser variables and plot the misogynous-or-not class in Figure 3. It can be seen that the features without CWK are not able to differentiate between the classes, whereas features with CWK result in better separation, and are therefore, more effective for our task. We provide further qualitative evidence in Figure 4.

4.6 Joint Learning

The results for multi-task and hierarchical learning are presented in Table 3. We observe that there is an improvement in the binary classification task, and we reason that the joint learning paradigm provides significantly new information about subclasses from the multi-class setting to the binary task. However, results for the multi-class setting are comparable to the independent models.

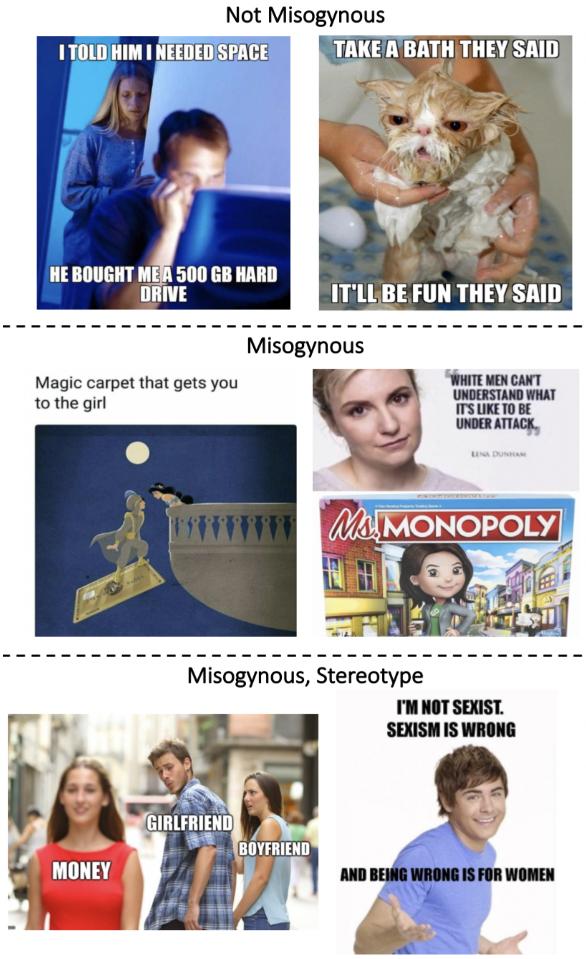

Not Misygynous
I TOLD HIM I NEEDED SPACE
HE BOUGHT ME A 500 GB HARD DRIVE
TAKE A BATH THEY SAID
IT'LL BE FUN THEY SAID
Misogynous
Magic carpet that gets you to the girl
"WHITE MEN CAN'T UNDERSTAND WHAT IT'S LIKE TO BE UNDER ATTACK"
LENA DUNHAM
Ms. MONOPOLY
Misogynous, Stereotype
MONEY
GIRLFRIEND
BOYFRIEND
I'M NOT SEXIST. SEXISM IS WRONG
AND BEING WRONG IS FOR WOMEN

Figure 4: Examples where Gated BERT fails, but Gated BERT with CWK classifies the memes correctly.

5 Conclusion and Future Work

Our work focused on the task of misogyny detection in multimodal memes. We demonstrated that using a combination of visual and textual, i.e, multimodal features outperforms the unimodal counterparts. In addition to simple baselines, we have also experimented with advanced DL architectures inspired from VQA and multimodal transformers. Further, we have shown how incorporating common world knowledge from Urban dictionary and pretrained CLIP can significantly help in identifying misogynistic content, along with qualitative evidence. Finally, the proposed joint learning paradigm can exploit the synergy between the two tasks, instead of training models independently.

Acknowledgements

We thank the anonymous reviewers for providing feedback on our manuscript. This work was done as part of the CS 229: Machine Learning course taught at Stanford University. We gratefully acknowledge the support and feedback from our instructors Andrew Ng, Moses Charikar and Carlos Guestrin, and the entire course staff.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics.

Elisabetta Fersini, Francesca Gasparini, and Silvia Corchs. 2019. Detecting sexist meme on the web: A study on textual and visual cues. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 226–231. IEEE.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 Task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Shreyansh Gandhi, Samrat Kokkula, Abon Chaudhuri, Alessandro Magnani, Theban Stanley, Behzad Ahmadi, Venkatesh Kandaswamy, Omer Ovenc, and Shie Mannor. 2019. Image matters: Detecting offensive and non-compliant content/logo in product images. *arXiv preprint arXiv:1905.02234*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33:2611–2624.

Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. 2021. Disentangling hate in online memes. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5138–5147.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

John Edison Arevalo Ovalle, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. 2017. Gated multimodal units for information fusion. In *ICLR (Workshop)*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Steven Wilson, Walid Magdy, Barbara McGillivray, Kiran Garimella, and Gareth Tyson. 2020. Urban dictionary embeddings for slang nlp applications. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4764–4773.

Ron Zhu. 2020. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *arXiv preprint arXiv:2012.08290*.