4-2022

On explaining multimodal hateful meme detection models

Ming Shan HEE

Roy Ka-Wei LEE

Wen Haw CHONG
Singapore Management University, whchong.2013@phdis.smu.edu.sg

On Explaining Multimodal Hateful Meme Detection Models

Ming Shan Hee
Singapore University of Technology and Design
Singapore, Singapore
mingshan_hee@mymail.sutd.edu.sg

Roy Ka-Wei Lee
Singapore University of Technology and Design
Singapore, Singapore
roy_lee@sutd.edu.sg

Wen-Haw Chong
Singapore Management University
Singapore, Singapore
whchong.2013@phdis.smu.edu.sg

ABSTRACT

Hateful meme detection is a new multimodal task that has gained significant traction in academic and industry research communities. Recently, researchers have applied pre-trained visual-linguistic models to perform the multimodal classification task, and some of these solutions have yielded promising results. However, what these visual-linguistic models learn for the hateful meme classification task remains unclear. For instance, it is unclear if these models are able to capture the derogatory or slurs references in multimodality (i.e., image and text) of the hateful memes. To fill this research gap, this paper propose three research questions to improve our understanding of these visual-linguistic models performing the hateful meme classification task. We found that the image modality contributes more to the hateful meme classification task, and the visual-linguistic models are able to perform visual-text slurs grounding to a certain extent. Our error analysis also shows that the visual-linguistic models have acquired biases, which resulted in false-positive predictions.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Computer vision representations**.

KEYWORDS

hate speech, hateful memes, multimodal, explainable machine learning

ACM Reference Format:
Ming Shan Hee, Roy Ka-Wei Lee, and Wen-Haw Chong. 2022. On Explaining Multimodal Hateful Meme Detection Models. In *Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3485447.3512260

Disclaimer: *This paper contains violent and discriminatory content that may be disturbing to some readers.*

1 INTRODUCTION

Motivation. Internet memes, which are often presented as images with accompanying text, are increasingly abused to spread hatred under the guise of humor [8, 13, 24]. To fight against the proliferation of hateful memes, Facebook has recently released a large hateful meme dataset and crowdsourced hateful meme classification solutions [13]. The research community has responded enthusiastically as many promising hateful meme classification methods have been proposed [16, 20, 31, 35, 36]. Among the proposed solutions, a popular line of approaches is to apply pre-trained visual-linguistic models such as VisualBERT [17] and VilBERT [21] to perform the hateful meme classification tasks. These methods have yielded promising results. However, what these visual-linguistic models learn for the hateful meme classification task remains unclear.

Research Objectives. Understanding visual-linguistic models is an emerging research area that has garnered much attention from the multimodal research community [3, 7, 18, 22]. Inspired by works that explored the internal behaviors of pre-trained language models [4], Li et al. [18] conducted a quantitative study on whether visual-linguistic models acquire semantic grounding ability during pre-training without explicit supervision. Frank et al. [7] proposed a diagnostic framework to assess the extent to which the visual-linguistic models integrate cross-model information. This paper aims to contribute to the existing literature on visual-linguistic model understanding by applying some of these techniques to investigate how visual-linguistic models understand hateful memes. To the best of our knowledge, this is the first paper that attempts to understand what visual-linguistic models actually learn when training for the hateful meme classification task.

Contributions. Our paper proposed three research questions, which improve our understanding of the internal behaviors of visual-linguistic models trained to perform the hateful meme classification task. Through extensive quantitative and qualitative analyses, we show that (i) the visual-linguistic models have accorded higher importance to the visual modality when performing hateful meme classification; (ii) the visual-linguistic models are able to learn the visual-text slurs grounding; (iii) and the models have acquired biases that adversely affected their hateful meme classification performance.

2 RESEARCH QUESTIONS

This paper aims to improve our understanding of visual-linguistic models applied to perform the hateful meme classification task. Working towards this goal, we formulate three research questions to guide our exploration[1].

RQ1: Modality Attribution. A key characteristic of memes is their multimodality nature, where their underlying message is often communicated via a combination of text and visual information. The multimodality characteristic also motivated the application

WWW '22, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9096-5/22/04…$15.00
https://doi.org/10.1145/3485447.3512260

[1]Code implementation: https://gitlab.com/bottle_shop/safe/ExplainHatefulMeme.

Table 1: Distribution of Facebook hateful meme dataset

Train		Test	
Hate	Non-hate	Hate	Non-hate
5,493	3,007	246	254

of visual-linguistic models to perform hateful meme classification. However, it is unclear if the text and visual information contributed equally towards the multimodal classification task. Existing studies have attempted to improve the explainability of deep learning models by attributing the prediction of a deep network to its input features [1, 11, 15, 30]. For instance, Sundararajan et al. [30] proposed an attribution method called *Integrated Gradients* to score the contribution of input features on deep models' prediction. The researchers applied their model on several images and text deep learning models to demonstrate its ability to explain the prediction results of these models. Similar studies were also conducted by investigating the attribution of text and visual features in multimodal tasks such as Visual-question-answering (VQA) [9, 10, 23]. We aim to apply the attribution methods to understand how the different modalities input features contribute to the hateful meme classification task.

RQ2: Visual-Text Slurs Grounding. Hate speech detection tasks have been notoriously challenging due to the ambiguity and variability in natural languages. Adding to the complexity, hate speech also includes unique derogatory terms that are commonly used as insinuations or allegations about members of a specific group. For example, the word *"Dishwasher"* has been associated with females due to the traditional gender ideology where married women would become housewives and do house chores. In doing so, it undermines women's gender equality rights and objectifies them as mere tools. The use of derogatory terms amplifies further in hateful memes. With the additional visual information, the subtle allegations in the derogatory terms are communicated through contextual cues in both modalities. Existing works have investigated the semantic grounding capabilities of visual-linguistic models in VQA and image captioning tasks [7, 18]. We aim to extend these studies to understand the visual-linguistic models' ability to perform visual-text grounding for derogatory terms and slurs used in the hateful memes.

RQ3: Bias and Error Analysis. Data and model biases in text-based hate speech detection tasks have been are widely researched [2, 5, 12, 32, 34]. For instance, Kennedy et al. [12] conducted a study to analyze group identifier biases in hate speech detection models. The researchers found that existing text-based hate speech classifiers are over-sensitivity to group identifiers like "*Muslim*", "*gay*", and "*black*". We aim to extend these studies to conduct a preliminary analysis on the biases in hateful meme classification models. Specifically, we will examine the group identifier biases in both text and visual modalities of the wrongly classified memes.

3 EXPERIMENTS

3.1 Experiment Settings

3.1.1 Dataset. The Facebook hateful meme dataset [13], which was constructed and released by Facebook as part of a challenge to crowd-source multimodal hateful meme classification solutions, is a popular dataset used in many research studies. Therefore, we utilize this dataset in our experiments. The dataset contains $10K$ memes with binary labels (i.e., hateful or non-hateful). As we do not have labels of the memes in the test split, we utilize the *dev-seen* split as the *test* set. Table 1 outlines the distributions of the dataset.

Table 2: Average gradients from text and visual inputs in various models

Model	Text Input		Visual Input	
	Avg	Std. Dev	Avg	Std. Dev
VilBERT	3.183	0.900	4.096	0.774
VisualBERT	3.006	0.804	7.705	2.180
VilBERT CC	3.130	0.882	4.145	0.730
VisualBERT COCO	3.112	0.843	6.444	0.976

3.1.2 Models. VilBERT [21] and VisualBERT [17] are amongst the state-of-the-art visual-linguistic models often used for various multimodal tasks. The two models were also applied as baselines to evaluate the released Facebook hateful meme dataset [13]. Both models use pre-trained text from BERT [6] and image features from $fc6$ layer of Faster-RCNN [25] with ResNeXt-152 as its backbone [33]. These models can also be trained on multimodal objectives as an intermediate step before fine-tuning them for the multimodal hateful meme classification task. For our experiments, we train the **VilBERT** and **VisualBERT** on the hateful meme classification task. We also included two multimodally pre-trained versions of these models and fine-tuned them for the hateful meme classification task. Specifically, we include VilBERT on Conceptual Captions [27] (**VilBERT CC**), and VisualBERT on Microsoft's Common Objects in Context [19] (**VisualBERT COCO**). We trained the models using the Facebook MMF framework [29] and adopt the hyperparameters specified in [13], as the researchers have already performed grid search on numerous hyperparameters.

3.2 Modality Attribution

Inspired by gradient-based researches that use gradients as feature importance [26, 28], we use the gradients to represent the contribution of each modality towards making the model's decision. We reasoned that gradients signify the weightage each input feature has towards making the model's prediction. Hence, we equate the summation of gradients for each input type as the contribution of each modality. Specifically, we first obtain the gradients for the inputs in each modality through backpropagation. Subsequently, we normalize the gradients by their magnitude. Through normalization, we place the gradients for the text and visual modality onto the same unit space and enable us to compare their relative contribution. We attribute the summation of these normalized gradients as the contribution of each modality. Finally, we compute the average and standard deviation across the 500 samples in the test data. We noted empty tokens are present in the text inputs but not in the visual inputs, as the text inputs have various lengths. As these empty tokens do not contain any meaning, we argue that their gradients should not attribute to the contribution of each modality and do not consider these gradients in our analysis.

Table 2 shows the average and standard deviation for the text and visual modality gradients for various hateful meme classification models. We observe that the visual modality consistently has a slightly higher average gradient than the text modality, suggesting that the visual modality contributes more towards the model's

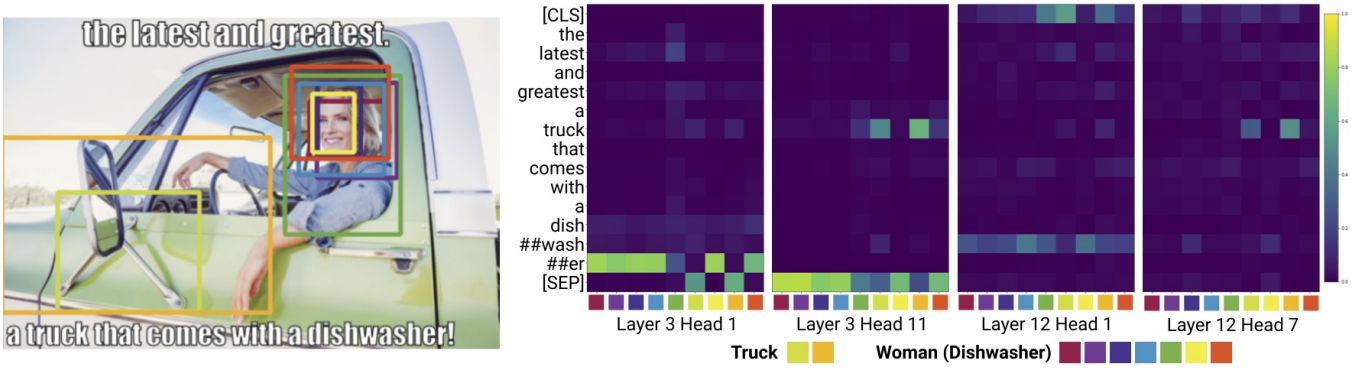

Figure 1: Attention heads in VisualBERT where the text-visual alignments attend to the word "dishwasher" and "truck".

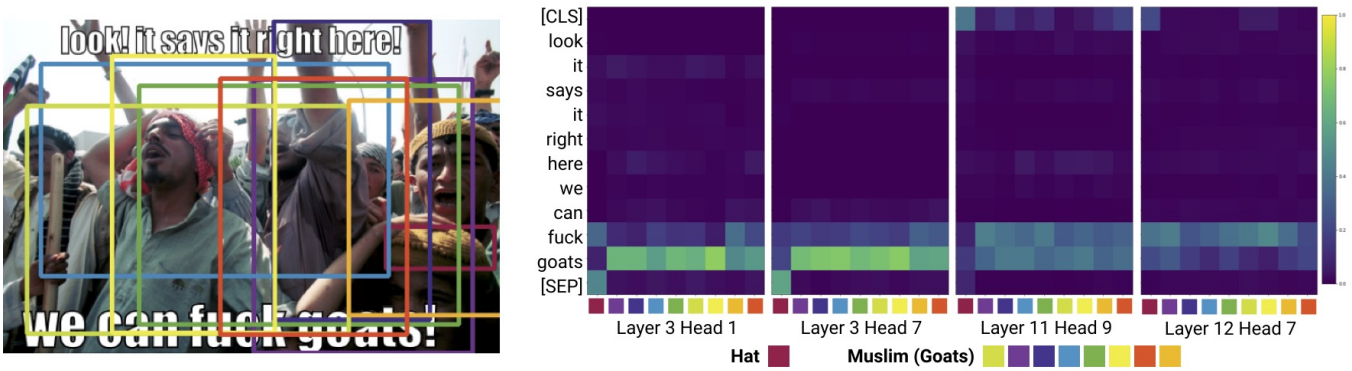

Figure 2: Attention heads in VisualBERT where the text-visual alignments attend to the word "goat" and "f*ck"

prediction. We also noted that the standard deviation for both modalities is relatively low across most models, suggesting that there is little variation in the contribution of each modality across all samples.

The observation that the visual modality contributes more in the hateful meme classification results concurs with Frank et al. [7]'s findings where proposed vision-and-language models tend to attend to visual information more than text information. A possible reason for this observation could be the ratio of text features to image features. We note that internet memes generally have relatively short text. For instance, the samples in the validation dataset have an average number of 14 words and a maximum number of 54 words. Conversely, the image features always comprise the top 100 image regions extracted from existing object detection techniques. Hence, the number of text features is always lower than the number of image features across the samples, and the models may have leverage on the modality with more features.

3.3 Visual-Text Slurs Grounding

The visual-linguistic models offer a simplistic solution to learn interaction across modalities. It uses a stack of Transformer layers that implicitly aligns the text input and visual input with self-attention. Recent works have also established that input alignments within VisualBERT's attention weights often capture intricate associations in the Transformers' architecture (e.g., visual-text entity grounding, visual-text syntactic grounding, etc.) [18]. Extending from these studies, we should observe visual-text alignments for slurs within the attention weights after fine-tuning the visual-linguistic models on the hateful meme classification task.

As our investigation aims to understand the visual-text grounding in the hateful meme classification tasks, we visualize the visual-text alignments within the models' attention weights for evidence of slurs grounding. Specifically, we display the bounding region on the image for each visual feature. To obtain the bounding region for each visual feature, we used the predicted coordinates from the final layer of the Faster-RCNN as they originate from the model's intermediate layer fc6. However, visualizing 100 bounding regions would clutter the image and would be impractical to make any valuable observation. Therefore, we choose to visualize the top 9 features ranked by their contribution to the models' decisions.

There are many slurs embedded in the hateful memes. We select and examine two common slurs, "*dishwasher*" and "goat-f*cker", that target female and Muslim communities, respectively. Specifically, we search the hateful memes that contain the words "*dishwasher*", "*goat*", and "goat-f*cker". From the retrieved memes, we examine the visual bounding regions that have attention weight aligned to the selected keywords. We then show the 4 heads of the attention layers that best demonstrate visual-text alignments for slurs and entities. The visual-text alignment will reveal the models' multimodal understanding of the slurs embedded in the memes.

The word "*dishwasher*" is a sexist slur commonly found in hate speech targeted at the female gender, used to stigmatize women as housewives. In Figure 1, the meme describes a scenario where one undermines women's gender equality rights and addressed

[CLS] I ordered a dish ##wash ##er and they sent me a sandwich maker [SEP]

(a)

[CLS] goat meat is very popular in Islamic countries [SEP]

(b)

Figure 3: Two examples of non-hateful memes wrongly classified by VisualBERT

the woman as "*dishwasher*". We observed that the model implicitly forms alignments between the subword "##er" for *dishwasher* and the seven image segments containing the woman in lower layers. The model displays residual visual-text alignments for the slur using a different subword "##wash" by the end of the computation (i.e., layer 12), demonstrating the strong presence of visual-text alignment for image of woman and the slur "*dishwasher*".

The phrases "*goat-humper*" and "*goat-f*cker*" is one of the many offensive slurs commonly targeted at the Muslim religion, accusing Muslim men of having sexual relationships with goats (i.e., an act of bestiality). Figure 2 shows a meme that suggests the Muslim men have sexual relationships with goats by associating the text "*we can f*ck goats*" with Muslim men in the image. While the phrase "*goat-humper*" and "*goat f*cker*" does not appear in this meme, we could infer the underlying allegations and identify the relevant keywords. We observed that the word "*f*ck*" and the word "*goat*" assigns significant attention weights to the bounding regions containing Muslim men in the early layers. The model displays residual visual-text alignment for the word "*f*ck*" and the word "*goat*" in the later layers (i.e., layer 11 and 12), albeit more attention weights are assigned to the word "*f*ck*". Combining the observations, the model demonstrates presence of visual-text alignment for the relevant keywords that suggest the slurs "*goat-humper*" and "*goat-f*cker*".

We also observed that the early layers demonstrate a cleaner alignment to the slur terms. For example, in Figure 1, the unrelated visual segments to the subword "##er" and the word "truck" are aligned to the separator token. A similar observation can be made for the unrelated visual segments to the word "goat" in Figure 2. Based on recent researches, these alignments to the separator tokens can be seen as no-operation (no-op) as they do not substantially impact the model's output [4, 14].

3.4 Bias and Error Analysis

To analyze the bias of the hateful meme classification models, we conduct an error analysis on the wrongly predicted non-hateful memes by the models. Similar to the existing works that studied biases in text-based hate speech detection models, we are especially interested in the false positives because it may reveal the features that the models are overly sensitive to when performing hateful meme prediction. Specifically, we conduct this analysis by inspecting the bounding regions for the critical visual features and contributions of individual word tokens for text features. The visualization of bounding regions for the visual features follows the same strategy specified in 3.3. Whereas for the text features, we used Integrated Gradients [30] to visualize the contribution of each word towards making the model's prediction. Specifically, words are highlighted in green and red to indicate their attribution to hateful and non-hateful prediction, and the color intensity represents the level of attribution.

Figure 3 shows two examples of non-hateful memes wrongly classified by VisualBERT. In Figure 3(a), we observed that the subword tokens for "*dishwasher*" and word token "*sandwich*" have a high contribution towards making the model predict the meme as hateful. Upon inspecting the text-visual alignment, these text tokens also assign significant attention weights to the bounding region containing the woman. From the observations, we can infer that the model has likely learned a bias where the presence of keywords such as "*dishwasher*" and images of women would render the model to predict the meme as hateful. The bias also exposes a deeper issue with the visual-linguistic hateful meme classification model; even though the models are able to learn the visual-text slurs grounding, over-sensitivity to such grounding may also introduce bias and results in false positive.

The model also exhibits bias for the group identifier term "Islamic" in Figure 3(b). Inspecting the text-visual alignments, we observed that the word "Islamic" does not assign much attention weights to any bounding regions. We postulate that the bias for this group identifier term happens in the unimodal text space rather than the multimodal space.

4 CONCLUSION

We have presented an analysis of applying visual-linguistic models on the hateful meme classification task. Our analysis showed that the image modality contributes more to the hateful meme classification task, and the visual-linguistic models can perform visual-text slurs grounding. Nevertheless, the visual-linguistic models have also acquired biases, which resulted in false-positive predictions. For our future work, we will extend our analysis to multiple hateful meme datasets, and benchmark more models. We would also explore debiasing techniques to reduce the multimodal biases in the visual-linguistic models and improve their hateful meme classification performance.

REFERENCES

[1] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2018. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30-May 3, 2018, Conference Track Proceedings*. OpenReview. net.

[2] Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *The World Wide Web Conference*. 49–59.

[3] Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *European Conference on Computer Vision*. Springer, 565–580.

[4] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What Does BERT Look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. 276–286.

[5] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*. 25–35.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[7] Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 9847–9857.

[8] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1470–1478.

[9] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6904–6913.

[10] Yash Goyal, Akrit Mohapatra, Devi Parikh, and Dhruv Batra. 2016. Towards transparent ai systems: Interpreting visual question answering models. *arXiv preprint arXiv:1608.08974* (2016).

[11] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.

[12] Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing Hate Speech Classifiers with Post-hoc Explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5435–5442.

[13] 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 (2020).

[14] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. *arXiv preprint arXiv:2004.10102* (2020).

[15] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896* (2020).

[16] 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*. 5138–5147.

[17] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).

[18] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020. What Does BERT with Vision Look At?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5265–5275.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[20] Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. 2020. A multimodal framework for the detection of hateful memes. *arXiv preprint arXiv:2012.12871* (2020).

[21] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*.

[22] Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. 2021. Seeing past words: Testing the cross-modal capabilities of pretrained v&l models on counting tasks. In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*. 32–44.

[23] Badri Patro, Shivansh Patel, and Vinay Namboodiri. 2020. Robust explanations for visual question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1577–1586.

[24] Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Detecting Harmful Memes and Their Targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2783–2796.

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015), 91–99.

[26] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[27] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2556–2565.

[28] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[29] Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2020. MMF: A multimodal framework for vision and language research. https://github.com/facebookresearch/mmf.

[30] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*. PMLR, 3319–3328.

[31] Riza Velioglu and Jewgeni Rose. 2020. Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches: Prize-winning solution to Hateful Memes Challenge. *arXiv preprint arXiv:2012.12975* (2020).

[32] Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting Racial Bias in Hate Speech Detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*. 7–14.

[33] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1492–1500.

[34] Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah A Smith. 2021. Challenges in Automated Debiasing for Toxic Language Detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 3143–3155.

[35] Yi Zhou, Zhenhao Chen, and Huiyuan Yang. 2021. Multimodal Learning For Hateful Memes Detection. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 1–6.

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