wentaorub at Memotion 3: Ensemble learning for Multi-modal MEME classification

Wentao Yu[1], Dorothea Kolossa[2]

[1]Institute of Communication Acoustics, Ruhr University Bochum, Germany
[2]Electronic Systems of Medical Engineering, TU Berlin, Germany

Abstract

Memes, as a new means of creative expression on social networks, provide an appealing multi-modal form of communication. However, some memes are being used to express hatred, which can take a toll on people's mental health and on societal cohesion. This year's Memotion 3.0 challenge provides an English and a mixed Hindi-English meme dataset for three classification tasks: Task A is sentiment analysis to classify a given meme as positive, negative, or neutral. In Task B, emotion classification, a meme should be identified as humorous, sarcastic, offensive, or motivational. Finally, Task C asks to predict the intensity of the emotion classes in Task B. Both text and image data play a role in the identification and classification of hateful memes. While such multi-modality can be helpful in many contexts, here, it also increases the challenge of the classification tasks due to the nature of memes, which often achieve their humorous effects through juxtaposition and irony. To address this difficulty, we adopt a multi-headed self-attention mechanism to integrate the text and image information in a learned, task-adapted manner. The gradient blending algorithm prevents overfitting issues in the multi-modal model. Our uni-modal models, which feed into the attention mechanism, are based on the CLIP model due to its outstanding performance on zero-shot classification tasks. Ultimately, with an ensemble strategy of our two best-performing models, our submission only reaches a 0.3289 weighted F1 score on sub-task A, but it ranks 1st on the two final Tasks B and C, with respective scores of 0.7977 and 0.5982. [1]

Keywords

Ensemble, CLIP, OSCAR, multimodal, memes classification

1. Introduction

It is well-known that multi-modal machine learning can vastly outperform uni-modal learning, at least when the system is set up appropriately. For example, in audio-visual speech recognition, visual information can complement speech signals to significantly improve recognition rates [1, 2, 3]. However, memes often express opinions in an implied manner. The text and image may even have opposite meanings in isolation and can be combined ironically. This characteristic of memes leads to a new type of challenge in automatic classification tasks. In order to study this problem, the Memotion 3.0 challenge provides a Hinglish meme dataset for three meme classification tasks [4, 5, 6, 7, 8, 9, 10].

[1]Our code will be made available at: https://github.com/wentaoxandry/Memotion3.0_challenge.git

De-Factify 2: 2nd Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAAI 2023. 2023 Washington, DC, USA

wentao.yu@rub.de (W. Yu); dorothea.kolossa@tu-berlin.de (D. Kolossa)
https://cognitive-signal-processing.de/index.php/team/ (W. Yu); https://www.tu.berlin/en/mtec (D. Kolossa)

In this work, we consider transfer learning to customize two multi-modal models based on the Transformer model: the CLIP model [11] and the OSCAR model [12]. The text and image encoders from the CLIP model are optimized as two uni-modal (text and image) models. Ultimately, the ensemble strategy is applied for better performance.

The paper is organized as follows: Section 2 introduces the related solutions to the task. Our system framework is described in Section 3, followed by the experimental setup in Section 4. Finally, our results are shown and conclusions are drawn in Sections 5 and 6.

2. Related Work

The transformer model [13] is widely used in natural language processing tasks due to its outstanding performance. In recent years, a number of works have expanded the capability of the transformer model towards multi-modal tasks. For example, the OSCAR model [12] adopts the Faster R-CNN [14] to extract visual embeddings of the detected object regions. In addition, the Faster R-CNN model outputs the detected object tags, which are considered as additional anchor points to improve the learning performance of alignments. Subsequently, an attention mechanism walks through the combined text-image sequence embeddings.

Recently, contrastive learning has drawn much attention due to its outstanding performance on zero-shot prediction [15, 16]. In this work, we consider the CLIP model [11] to extract contrastive embeddings, since memes contain various image types, which causes difficulties in classification. The remarkable zero-shot prediction accuracy of the CLIP model can help us to alleviate this problem. The CLIP model uses a pre-trained BERT model to extract text context classification features and a Vision transformer [17] for obtaining image classification features. Contrastive learning is adopted to learn multi-modal embeddings without manual labels by teaching the CLIP model about the similarity of different data points. Assuming a mini training batch with meme OCR texts $\mathbf{T} = \{\mathbf{t}_1, \mathbf{t}_2, \cdots \mathbf{t}_n, \}$ and images $\mathbf{I} = \{\mathbf{i}_1, \mathbf{i}_2, \cdots \mathbf{i}_n, \}$, where n is the batch size, the CLIP model learns to match the OCR text and image as follows: the extracted text classification features $\mathbf{F}_{cls,t} \in \mathbb{R}^{n \times d_t}$ and image classification features $\mathbf{F}_{cls,i} \in \mathbb{R}^{n \times d_i}$ are computed by

$$\begin{aligned} \mathbf{F}_{cls,t} &= \text{encoder}_t(\mathbf{T}), \\ \mathbf{F}_{cls,i} &= \text{encoder}_i(\mathbf{I}), \end{aligned} \tag{1}$$

where d_t and d_i are the attention dimension of text and image encoder, respectively. The classification features are mapped to the same dimension d_e and normalized with an l2 regularization. The contrastive logits $\boldsymbol{x}$ are derived as their scaled, pairwise cosine similarities:

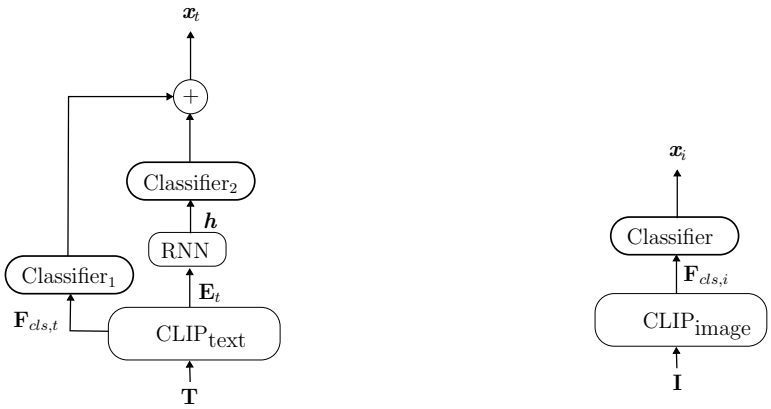
$$\boldsymbol{x} = (\|\mathbf{W}_t \cdot \mathbf{F}_{cls,t}\|_2 \cdot \|\mathbf{W}_i \cdot \mathbf{F}_{cls,i}\|_2^T) \times e^k, \tag{2}$$

where k is a learnable parameter. As in [11], the labels are the one hot encoded labels of the set $\boldsymbol{y} = [1, 2, \cdots n]$. The loss function is defined as:

$$l = 0.5 \cdot \underset{axis=0}{\text{CE}}(\widehat{\boldsymbol{y}}_i, \boldsymbol{y}) + 0.5 \cdot \underset{axis=1}{\text{CE}}(\widehat{\boldsymbol{y}}_t, \boldsymbol{y}), \tag{3}$$

where CE is the cross-entropy, $\widehat{\boldsymbol{y}}_i = \underset{axis=0}{\text{softmax}}(\boldsymbol{x})$ and $\widehat{\boldsymbol{y}}_t = \underset{axis=1}{\text{softmax}}(\boldsymbol{x})$. Finally, the learned text and image classification features $\mathbf{F}_{cls,t}, \mathbf{F}_{cls,i}$ are used in our proposed CLIP-based text, image, and multi-modal models.

3. System Overview



(a) Text model (b) Image model

Figure 1: Uni-modal CLIP models.

Figure 1 illustrates the proposed ***uni-modal*** text and image models. In Figure 1a, the $\text{CLIP}_{\text{text}}$ model outputs the classification embedding $\mathbf{F}_{cls,t}$ for the sequence classification task, and the embedding sequence $\mathbf{E}_t = \left[\mathbf{F}_{cls,t}, \boldsymbol{e}_{t,1}, \boldsymbol{e}_{t,2}, \cdots \boldsymbol{e}_{t,s}\right]$, where $\boldsymbol{e}_{t,j}$, $j \in \{1, \ldots, s\}$ is the embedding vector of the j-th token. All classifiers in Figure 1 have the same structure: a fully connected layer with a dropout layer followed by an output layer. The text model first uses the classification embedding $\mathbf{F}_{cls,t}$ to predict the classification logits. Then 2 BiLSTM layers, whose cell dimension of each layer is $d = 512$, extract the context information $\boldsymbol{h} \in \mathrm{R}^{s \times 2d}$. Another classifier computes the sequence logits from $\boldsymbol{h}$. Finally, the text model logits $\boldsymbol{x}_t$ are computed by summing the classification and sequence logits. The image model logits $\boldsymbol{x}_i$ are computed straightforwardly from the classification embedding $\mathbf{F}_{cls,i}$ by the classifier. The image sequence embeddings

$$\mathbf{E}_i = \left[\mathbf{F}_{cls,i}, \boldsymbol{e}_{i,1}, \boldsymbol{e}_{i,2}, \cdots \boldsymbol{e}_{i,p}\right], \boldsymbol{e}_{i,j}, j \in \{1, \ldots, p\} \tag{4}$$

are utilized in the multi-modal model, where p is the number of patches of the image.

Figure 2 depicts the proposed ***multi-modal model***. The sequence embeddings from the text and image model in Figure 1 are concatenated along the sequence dimension as multi-modal embeddings

$$\mathbf{E} = [\mathbf{E}_t; \mathbf{E}_i], \text{ where } \mathbf{E} \in \mathbb{R}^{(s+q) \times d_e}. \tag{5}$$

The complete embedding sequence $\mathbf{E}$ is fed into six multi-head attention (MHA) blocks. We removed the MHA block's residual connection and dropout layer based on our experimental results. Finally, the classifier, which has the same structure as the text and image classifiers, uses the multi-modal classification embedding $\mathbf{F}_{cls} \in \mathbb{R}^{2d_e}$ to obtain the multi-modal logits $\boldsymbol{x}_{multi}$, where

$$\mathbf{F}_{cls} = [\widetilde{\mathbf{E}}_t[0, :], \widetilde{\mathbf{E}}_i[0, :]]. \tag{6}$$

$\widetilde{\mathbf{E}}_t[0, :]$ and $\widetilde{\mathbf{E}}_i[0, :]$ are the first classification embeddings after 6 MHA blocks.

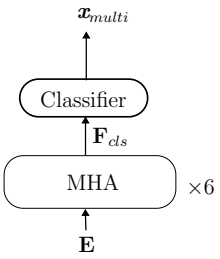


Figure 2: Multi-modal model.

4. Experimental Setup

Table 1 gives an overview of all datasets used to optimize our models. The `ImgFlip575K` [1] dataset was collected from Imgflip, captioned by `Scrapy`. The `MMHS150k` dataset [18] was obtained from Twitter for hate speech detection tasks. Only meme images and OCR texts are available in the `ImgFlip575K` and `MMHS150k` datasets. The `Facebookhm` dataset [19] was created in 2020 for the Facebook hateful memes challenge. The `MAMI` dataset [20] was created for training models to identify and classify misogynous memes. Both the `Facebookhm` and `MAMI` datasets are labeled for binary classification tasks, namely, for identifying hateful or misogynous memes, respectively. Therefore, we combine the `Facebookhm` training and development sets with the `MAMI` training set into the new `pretrainuni` training set. The `Facebookhm` and `MAMI` test sets are correspondingly grouped into the `pretrainuni` test set. The training and development sets of all datasets are combined to form our `pretrainCLIP` training set, while—except for the `Memotion3` test set—all test sets are combined into the `pretrainCLIP` test set.

Table 1
Dataset distribution

Dataset	training set	test set	development set
`ImgFlip575K`	573,948	2000	-
`MMHS150k`	59,252	-	-
`Facebookhm`	8500	3000	640
`MAMI`	10000	1000	
`Memotion3`	7000	1500	1500
`pretrainCLIP`	652,340	6,000	-
`pretrainuni`	19,140	4000	-

The Memotion 3.0 challenge has three sub-tasks. Task A is to classify a meme as positive, negative, or neutral. In Task B, a given meme should be identified as humorous, sarcastic, offensive, or motivational. It is a multi-label classification task, so that a meme can have more than one category. Finally, Task C asks to predict the intensity of the emotion classes in Task B. In our work, we only optimized the models for Task A and Task C. Task B results are obtained

[1] https://github.com/schesa/ImgFlip575K_Dataset

from Task C. For example, we consider a meme as *Humorous* if it is classified as **F**, **VF**, or **H** (detailed in Table 2) and vice versa. Since the Memotion3 dataset contains English as well as mixed Hindi-English memes, we perform back-translation to Hindi and then to English with the Python `translators` package.

Table 2
Memotion3 dataset class distribution. +: Postive; **!**: Neurtal; -: Negative; **NF**: Not funny; **F**: Funny; **VF**: Very funny; **H**: Hilarious; **NS**: Not Sarcastic; **LS**: Little Sarcastic; **VS**: Very Sarcastic; **ES**: Extremely Sarcastic; **NO**: Not Offense; **S**: Slight; **VO**: Very Offensive; **HO**: Hateful Offensive; **NM**: Not Motivational; **M**: Motivational.

	Task A			Task C													
set	+	!	-	Humorous				Sarcastic				Offensive				Motivational	
				NF	**F**	**VF**	**H**	**NS**	**LS**	**VS**	**ES**	**NO**	**S**	**VO**	**HO**	**NM**	**M**
train	2275	2970	1755	1010	3393	2038	559	1476	1953	3021	550	4264	1935	610	191	6170	830
val	341	579	580	99	973	375	53	123	977	376	24	641	804	44	11	1457	43

All models are trained using the PyTorch library [21]. The AdamW optimizer [22] is used for backpropagation. The CLIP model (detailed in Section 2) is first optimized using the `pretrainCLIP` dataset, where the contrastive loss is the objective function. The pre-trained CLIP model thus learns to match the meme image and text and is denoted as CLIP_{pre}. For uni-modal training, the `pretrainuni` dataset is used to fine-tune the the CLIP component models $\text{CLIP}_{\text{pre,t}}$ and $\text{CLIP}_{\text{pre,i}}$ within the two overall model structures as shown in Figure 1. The model parameters of the CLIP component models `CLIP-text` and `CLIP-image` are initialized by those of the respective CLIP_{pre} model and optimized on the Memotion3 dataset. The multi-modal model `CLIP-multi` parameters are then initialized by the uni-modal models. We train one model for Task A with an output dimension of 3 and four models for the four aspects of Task C with output dimensions 4, 4, 4, and 2. The dropout rate in all classifiers is 0.1. The attention dimensions of the text encoder and image encoder are 512 and 768, respectively. The dataset for Task A is balanced. Therefore, we simply use the cross-entropy (CE) as the loss function for Task A.In contrast, the dataset for Task C is quite imbalanced. The focal loss (F) function [23] is therefore selected as the loss function for training the respective classifiers.

In this work, we adopt Gradient-Blending [24] (GB) to reduce the effect of overfitting. The multi-modal model (Figure 2) is based on the text and image model (Figure 1). Therefore, the text and image model logits $\boldsymbol{x}_t$ and $\boldsymbol{x}_i$ are also available in the multi-modal model. Taking the gradient of the blended loss

$$GB = \sum_i w_i \text{CE}_i, \tag{7}$$

where $i \in$[text, image, multi-modal], produces the blended gradient. It should be emphasized that the multi-modal predictions are only obtained from the multi-modal logits $\boldsymbol{x}_{multi}$. Finally, Table 3 gives an overview of the use of the loss functions in training all models.

We use the Python RAY [2] package to find the best-performing hyperparameters. The training process is carried out on NVIDIA's Volta-based DGX-1 multi-GPU system, using 3 TeslaV100 GPUs with 32 GB memory each.

[2]https://github.com/ray-project/ray

Table 3
Loss functions overview in different models.

	Task A	Task C			
		Humorous	Sarcastic	Offensive	Motivational
Text	CE	F	F	F	F
Image	CE	F	F	F	F
Multi-modal	GB	GB	GB	GB	GB

Table 4
Weighted F1 score on validation set

Models	Overall	Humorous	Motivation	Offensive	Sarcastic
CLIP-text0	**0.3724**	0.5237	0.9568	0.4645	0.4750
CLIP-text1	0.3617	0.5244	**0.9572**	0.4520	0.4422
CLIP-image0	0.3469	0.5261	**0.9572**	0.4811	**0.5302**
CLIP-image1	0.3492	0.5221	**0.9572**	0.4611	0.5255
CLIP-multi0	0.3666	0.5225	**0.9572**	**0.4966**	0.4803
CLIP-multi1	0.3691	0.5198	**0.9572**	0.4858	0.4786
OSCAR0	0.3684	**0.5302**	0.9532	0.4553	0.4011
OSCAR1	0.3380	0.5181	0.9532	0.4525	0.3843
CLIP-multi0 without GB	0.3764	0.5186	0.9419	0.4590	0.4512
Ensemble-1	0.3725	0.5345	**0.9572**	0.4974	0.6170
Ensemble-2	**0.4453**	**0.5362**	**0.9572**	**0.5052**	**0.6259**

5. Results

This work considers the CLIP-text, CLIP-image (in Figure 1), CLIP-multi (in Figure 2), and OSCAR models. For better performance, majority voting is adopted to ensemble different models' decisions. Ensemble-1 fuses the prediction decisions of the candidate models CLIP-text0, CLIP-image0, CLIP-multi0, and OSCAR0, while Ensemble-2 also takes CLIP-text1, CLIP-image1, CLIP-multi1, and OSCAR1 into consideration. We iterate over all possible model combinations and adopt majority voting on the validation set to find the best performance model combinations. Then, these combinations are used to fuse the test set predictions.

Table 4 lists the weighted F1 score on the validation set. For Task A ("Overall" column in Table 4), the CLIP-text model performs better than the CLIP-image model. The score of the CLIP-multi setup lies between those of the former two models. Ensemble-2 improves the weighted F1 score to 0.4453. The model for motivation classification has scores above 0.9, because the binary classification dataset is imbalanced. Comparing the best-performing text and image models (CLIP-text0 and CLIP-image0), the image model shows a slightly better performance in Task C. The CLIP-multi0 model without GB training performs far worse than its gradient-blending counterpart. Overall, Ensemble-2 shows the best performance in Task A and Task C. Ultimately, the strategy of ensembling the top two models yields a 0.3289 (5th) weighted F1 score on Task A, 0.7977 (1st rank) on Task B and 0.5982 (also 1st rank) on Task C.

6. Conclusion

This work proposes a multi-modal CLIP-based meme classification system, which owes its capabilities on this rather small dataset to the outstanding zero-shot performance of the CLIP model. The text model combines the CLIP model text encoder with 2 BiLSTM layers; the image model is fine-tuned on the Memotion 3.0 dataset. The proposed multi-modal model integrates the text and image embeddings from the text and image encoders in 6 multi-head self-attention blocks. Gradient blending prevents the fusion model from overfitting. The OSCAR model is used both as a baseline model and as a participant model in our ensemble strategy, which further serves to improve the system performance. Our ensembe of the top two models yields a clearly better accuracy than one single model, winning Task B and C in the Memotion 3.0 challenge. The experimental results of the challenge do indicate, however, that sentiment analysis in memes is difficult for machine learning. The next goal of our work is therefore to develop mechanisms for understanding multi-modal, contrasting information, e.g. conveying irony, to improve sentiment classification performance for memes and social media posts.

Acknowledgments

The work was supported by the PhD School "SecHuman - Security for Humans in Cyberspace" by the federal state of NRW, and partially funded by the Deutsche Forschungsgemeinschaft (DFG – German Research Foundation) [Project-ID 429873205] and by the German Federal Ministry of Education and Research ["noFake", Grant No: 16KIS1519]. The authors are responsible for the content of this publication.

References

[1] W. Yu, S. Zeiler, D. Kolossa, Multimodal integration for large-vocabulary audio-visual speech recognition, in: Proc. 28th European Signal Processing Conf. (EUSIPCO), IEEE, 2021, pp. 341–345.

[2] W. Yu, S. Zeiler, D. Kolossa, Fusing information streams in end-to-end audio-visual speech recognition, in: Proc. ICASSP, IEEE, 2021, pp. 3430–3434.

[3] W. Yu, B. Boenninghoff, J. Roehrig, D. Kolossa, Rubcsg at SemEval-2022 Task 5: Ensemble learning for identifying misogynous MEMEs, arXiv preprint arXiv:2204.03953 (2022).

[4] P. Patwa, S. Ramamoorthy, N. Gunti, S. Mishra, S. Suryavardan, A. Reganti, A. Das, T. Chakraborty, A. Sheth, A. Ekbal, C. Ahuja, Findings of Memotion 2: Sentiment and emotion analysis of memes, in: Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, ceur, 2022.

[5] S. Mishra, S. Suryavardan, P. Patwa, M. Chakraborty, A. Rani, A. Reganti, A. Chadha, A. Das, A. Sheth, M. Chinnakotla, A. Ekbal, S. Kumar, Memotion 3: Dataset on sentiment and emotion analysis of codemixed Hinglish memes, in: Proc. Defactify 2: 2nd Workshop on Multimodal Fact-Checking and Hate Speech Detection, CEUR, 2023.

[6] S. Mishra, S. Suryavardan, M. Chakraborty, P. Patwa, A. Rani, A. Chadha, A. Reganti, A. Das, A. Sheth, M. Chinnakotla, A. Ekbal, S. Kumar, Overview of memotion 3: Sentiment

and emotion analysis of codemixed hinglish memes, in: Proc. Defactify 2: 2nd Workshop on Multimodal Fact-Checking and Hate Speech Detection, CEUR, 2023.
[7] S. Suryavardan, S. Mishra, P. Patwa, M. Chakraborty, A. Rani, A. Reganti, A. Chadha, A. Das, A. Sheth, M. Chinnakotla, A. Ekbal, S. Kumar, Factify 2: A multimodal fake news and satire news dataset, in: Proc. Defactify 2: 2nd Workshop on Multimodal Fact-Checking and Hate Speech Detection, CEUR, 2023.
[8] S. Suryavardan, S. Mishra, M. Chakraborty, P. Patwa, A. Rani, A. Chadha, A. Reganti, A. Das, A. Sheth, M. Chinnakotla, A. Ekbal, S. Kumar, Findings of Factify 2: Multimodal fake news detection, in: Proc. Defactify 2: 2nd Workshop on Multimodal Fact-Checking and Hate Speech Detection, CEUR, 2023.
[9] C. Sharma, D. Bhageria, W. Scott, S. Pykl, A. Das, T. Chakraborty, V. Pulabaigari, B. Gamback, SemEval-2020 Task 8: Memotion analysis–the visuo-lingual metaphor!, arXiv preprint arXiv:2008.03781 (2020).
[10] S. Ramamoorthy, N. Gunti, S. Mishra, S. Suryavardan, A. Reganti, P. Patwa, A. DaS, T. Chakraborty, A. Sheth, A. Ekbal, et al., Memotion 2: Dataset on sentiment and emotion analysis of memes, in: Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR, 2022.
[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: Proc. ICML, PMLR, 2021, pp. 8748–8763.
[12] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al., Oscar: Object-semantics aligned pre-training for vision-language tasks, in: Proc. ECCV, Springer, 2020, pp. 121–137.
[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in Neural Information Processing Systems 30 (2017).
[14] R. Girshick, Fast r-cnn, in: Proc. ICCV, 2015, pp. 1440–1448.
[15] Z. Han, Z. Fu, S. Chen, J. Yang, Contrastive embedding for generalized zero-shot learning, in: Proc. CVPR, 2021, pp. 2371–2381.
[16] H. Jiang, R. Wang, S. Shan, X. Chen, Transferable contrastive network for generalized zero-shot learning, in: Proc. CVPR, 2019, pp. 9765–9774.
[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).
[18] R. Gomez, J. Gibert, L. Gomez, D. Karatzas, Exploring hate speech detection in multimodal publications, in: Proc. IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 1470–1478.
[19] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, Advances in Neural Information Processing Systems 33 (2020) 2611–2624.
[20] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, Semeval-2022 task 5: Multimedia automatic misogyny identification, in: Proc. SemEval-2022, 2022, pp. 533–549.
[21] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin,

N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Advances in Neural Information Processing Systems (2019).
[22] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).
[23] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proc. ICCV, 2017, pp. 2980–2988.
[24] W. Wang, D. Tran, M. Feiszli, What makes training multi-modal classification networks hard?, in: Proc. CVPR, 2020, pp. 12695–12705.