Figure 1: Hateful meme example from [5].

estimate the level of alignment between the two modalities by a simple element-wise multiplication of the pre-trained representations at the token level. Then, the alignment-aware features are processed by a Transformer encoder that estimates intra- and inter-modality associations with the multihead softmax attention mechanism, to produce a rich feature vector. Finally, we also consider external knowledge as input to our model as well as a caption supervision relevant to the background image as a regularizing factor.

2 Related work

Although the topic of meme classification has emerged recently, there are already several papers addressing it and consecutively pushing further the state-of-the-art on pertinent standard benchmarks such as Memotion7k [2], Facebook Hateful Memes [3] and MultiOFF [9].

Exploratory data analysis methods, such as t-SNE [10], have been used for cluster detection [11], and Canonical Correlation Analysis (CCA) [12] and variants (e.g., Kernel CCA [13], Deep CCA [14]) for the study of modality interrelationship [11]. Many modality fusion approaches have been utilized such as early fusion (embedding concatenation [15, 16, 17, 18] or sum/average [19], Gated Multimodal Unit [20], multihead softmax attention [21], Low-rank factorization bilinear pooling [22] and outer product [23, 24]), late fusion [20, 25], hybrid fusion [11, 26] and even inter-task fusion in cases of multi-task learning [21]. Also, a typical approach is to consider pre-trained modality representations using word embeddings (e.g., GloVe, FastText, Word2Vec) [27, 22] and large-scale pre-trained neural networks for text [21, 18], image [18] and multimodal processing [3, 28, 29, 30].

Most of the previous works leverage ensemble approaches to achieve improved performance. The main model ensemble techniques considered are soft voting [29, 25], majority voting [26] as well as other linear and non-linear [29] prediction models optimally combining the ensemble models' predictions. The ensemble members typically are (i) large-scale VL pre-trained models [28], (ii) neural networks trained on k-folds of the training set [31], (iii) from scratch trained neural networks initialized with different seeds [29, 32] or of different hyperparameter settings [30] and, (iv) produced by evolutionary algorithms [31].

Finally, external knowledge has been leveraged by studies to make the models aware of information not provided explicitly by the datasets. The most frequent choices of external knowledge are textual and visual sentiment predictions [20], protected attribute extraction [28, 33], entity linking in the text [23] and the image [28, 33], and fine-grained object detection predictions [30, 31, 33].

3 MemeFier

Here we present MemeFier, a method to address the image meme classification task. Figure 2 illustrates the proposed architecture.

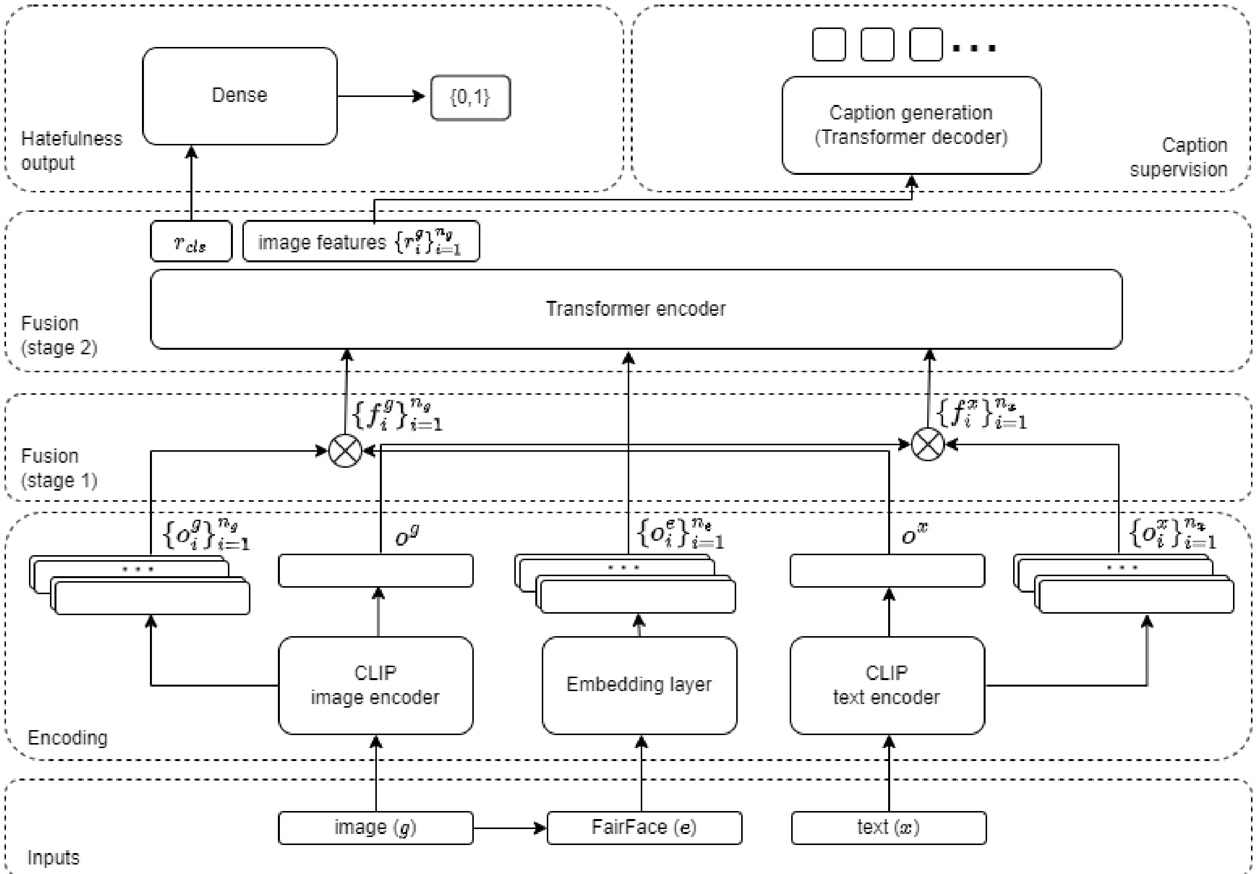

Figure 2: MemeFier's architecture. $\otimes$ denotes broadcastable element-wise multiplication.

3.1 Modality encoding

In [3], it is claimed that feature extractors trained with a multimodal objective tend to outperform combined unimodally pre-trained ones, when the task of interest is multimodal in nature. Here, for modality encoding we consider CLIP[3] [34]. After processing the image $g \in \mathbb{R}^{h \times w \times 3}$ and the text $x \in \mathbb{N}^L$, where h and w are the width and height of the image while L is the text's number of words, with CLIP's image and text encoders, we get the image embedding $o^g \in \mathbb{R}^d$ and its patch embeddings $\{o_i^g\}_{i=1}^{n_g} \in \mathbb{R}^{n_g \times d}$ as well as the text embedding $o^x \in \mathbb{R}^d$ and its token embeddings $\{o_i^x\}_{i=1}^{n_x} \in \mathbb{R}^{n_x \times d}$, where n_g and n_x are the number of patches and the number of tokens, respectively.

3.2 External knowledge retrieval and encoding

Incorporating external knowledge, is inspired by the fact that hate is usually targeted to certain population groups (e.g., muslims, women). Otherwise, the trained models decide purely based on image-text learned correlations. Other researchers have realized this issue as well, for instance in [3] the authors mention:

"*[...] Attacking groups perpetrating hate (e.g. terrorist groups) is also not considered hate. This means that hate speech detection also involves possibly subtle world knowledge.*"

We incorporate external knowledge to our model with the following procedure. For each image we get FairFace [35] predictions regarding gender, race and age of all depicted persons (if any) and denote it by $e \in \mathbb{N}^{3 \times n_p}$, where n_p is the number of persons and $n_e = 3 \times n_p$. We then encode this information through a typical embedding layer that is trained along with the rest of the model and produce the corresponding embeddings $\{o_i^e\}_{i=1}^{n_e} \in \mathbb{R}^{n_e \times d}$.

[3]The ViT-L/14 version provided by `https://github.com/openai/CLIP`.

3.3 Fusion

We consider a dual-stage modality fusion approach. During stage 1, we produce token-level modality representations that are aware of the level of alignment with the other modality. This kind of fusion has been proposed first in [24] but for image- (sentence-) level representations. More precisely, we compute $f_i^g = o_i^g \otimes o^x$ and $f_i^x = o_i^x \otimes o^g$, where $\otimes$ denotes element-wise multiplication, to get the fused image and text features, respectively. During stage 2, a Transformer encoder $T(\cdot)$ processes f_i^g,f_i^x and o_i^e along with a learnable classification token CLS and produces the corresponding feature representations:

$$\left[r_{cls}, \{r_i^g\}_{i=1}^{n_g}, \{r_i^x\}_{i=1}^{n_x}, \{r_i^e\}_{i=1}^{n_e}\right] = T\left(\left[CLS, \{f_i^g\}_{i=1}^{n_g}, \{f_i^x\}_{i=1}^{n_x}, \{o_i^e\}_{i=1}^{n_e}\right]\right) \tag{1}$$

3.4 Classification

As classification head for the hatefulness output we consider a typical fully-connected and sigmoid-activated layer of one unit $D(r_{cls})$.

3.5 Caption supervision

The vision encoder is likely to learn reduced image features that are advantageous only for the hatefulness detection, ignoring part of the background's semantics and consequently may overfit. To mitigate this potential deviation, we consider an additional supervision signal by reconstructing a description of the background image through a standard image captioning decoder. To this end, we first crop the visual part of the memes extracted with Visual Part Utilization (VPU) [1] and then consider the Once-For-All (OFA) [36] generated caption as target. Finally, a Transformer decoder is used to produce the caption based on the fused image features $\{r_i^g\}_{i=1}^{n_g}$.

4 Experimental setup

4.1 Datasets

Facebook Hateful Memes [3]: It contains 10K memes labeled as hateful or not and is split in 8.5K training data, 0.5K validation data (dev) and 1K test data. The test split labels are not released, thus we evaluate all models on the dev set.

Memotion7k [2]: It contains 9,871 memes, 1K trial data, 6,992 training data and 1,879 test data. It is human annotated for: (a) sentiment prediction, (b) overall emotion prediction and (c) estimation of the corresponding intensities. We randomly sample 10% from the training set for validation and report results on the test data.

MultiOFF [9]: It contains 743 memes manually annotated as offensive or non-offensive. The training set has 445 images and the validation and test sets have 149 each.

4.2 Baselines

We consider three baselines, an image-only, a text-only and a multimodal. The image-only is a ResNet18 [37] pretrained on ImageNet [38], the text-only is a trained-from-scratch LSTM [39], and the multimodal is a combination through early fusion of the above two models. In addition, we report the top performing competitive methods' scores per benchmark for comparison purposes. Unfortunately, there are no other papers reporting results on all the benchmarks considered by this work, thus the comparison is performed against different methods across the datasets. Brief descriptions about the competitive methods can be found in Section 2.

4.3 Hyperparameter tuning

For the baselines we conduct experiments for hyperparameter tuning accounting for different (1) initial learning rates (1e-2, 1e-3, 1e-4, 1e-5), (2) ResNet18 visual feature extractor being pre-trained on ImageNet or not (True, False), (3) number of hidden dimensions for the fully connected as well as for the LSTM (64, 128, 256), (4) number of LSTM layers (1, 3). We report the maximum performance.

For the proposed method, MemeFier, we consider the following hyperparameter grid. (1) Initial learning rates (1e-4, 1e-5), (2) number of epochs (16, 32), (3) contribution of the caption supervision α (cf. Section 4.4) to the final loss function (0.2, 0.8), (4) model dimension (512, 1024), (5) Transformer encoder settings: (5a) 4 heads, 512 feedforward dimension, 1 layer, (5b) 16 heads, 2048 feedforward dimension, 3 layers, (6) Transformer decoder settings: (6a) 64

method	accuracy	AUC
image	0.530	57.3
text	0.544	62.2
multimodal	0.554	61.3
[5]	0.659	74.14
[29]	-	81.56
[31]	-	77.39
[33]	**0.758**	**82.8**
[40]	-	73.93
[41]	-	78.57
[24]	-	82.62
MemeFier	0.736	80.1

Table 1: Model performance on Facebook Hateful Memes dataset (dev seen) in terms of accuracy and AUC scores.

method	a	b	c
image	0.333	0.502	0.315
text	0.350	0.481	0.279
multimodal	0.346	0.493	0.310
[27]	0.355	-	-
[42]	0.345	0.518	0.317
[17]	0.352	0.515	0.323
[22]	0.368	-	-
[43]	0.353	-	-
[21]	0.366	0.469	-
[23]	0.370	-	-
MemeFier	**0.396**	**0.519**	**0.343**

Table 2: Model performance on Memotion7k dataset in terms of F1 score.

method	accuracy	F1
image	0.638	0.619
text	0.571	0.508
multimodal	0.671	0.626
[9]	-	0.48
[33]	-	0.646
[23]	-	**0.671**
MemeFier	**0.685**	0.625

Table 3: Model performance on MultiOFF dataset in terms of accuracy and AUC scores.

input dimension, 4 heads, 64 feedforward dimension, 1 layer, (6b) 256 input dimension, 16 heads, 256 feedforward dimension, 3 layers.

4.4 Implementation details

For the baselines, training images are resized to 256, randomly cropped at 224, randomly horizontally flipped and standardized using the ImageNet statistics. Validation images are only resized to 224 and standardized. For text, we consider lowercasing and removing punctuation, numbers and double spaces. The vocabulary size is determined by the number of words having at least 5 occurrences in the corpus and the maximum sequence length by the 90% quantile of the distribution of lengths. Binary cross entropy is employed for the binary classification tasks and the multi-label classification tasks, while categorical cross entropy is employed for the multi-class classification tasks. We use the Adam optimizer, training for 10 epochs, with batch size 128 and the learning rate is reduced by a factor of 10 at epoch 5.

For MemeFier, we consider CLIP's image preprocessing pipeline and the same text preprocessing pipeline used for the baselines. For the text vocabulary we consider the same approach as for the baselines while for the captions vocabulary we consider all words and the actual maximum sequence length. The model comprises almost 29M parameters that we optimize using the Adam optimizer, binary cross entropy loss for the hatefulness output and categorical cross entropy for the caption supervision. Batch size is set to 32 and the learning rate is reduced by a factor of 10 after half epochs. The losses are combined as below:

$$\mathcal{L} = \mathcal{L}_{hate} + \alpha \cdot \mathcal{L}_{caption}$$

α denotes the caption supervision contribution, $\mathcal{L}_{hate}$ denotes binary cross-entropy, and $\mathcal{L}_{caption}$ denotes categorical cross-entropy.

4.5 Evaluation protocol

We report F1 score and accuracy for Memotion7k and MultiOFF, and AUC for Facebook Hateful Memes.

5 Results

Table 1 presents MemeFier performance compared to state-of-the-art on the dev seen split of the Facebook Hateful Memes dataset. In terms of accuracy the multimodal baseline outperforms the two unimodal baselines while in terms of AUC the text based approach performs better than both the image and the multimodal one. State-of-the-art methods exhibit varying performance between 74.1-82.8 in terms of AUC, while MemeFier performs comparably resulting in 80.1 AUC.

Table 2 illustrates the performance of MemeFier and competitive methods on the Mememotion7k dataset in terms of macro F1 score. Accuracy is not reported by other papers on this dataset so we do not provide a dedicated table here; however, some marginal gains were observed with MemeFier. Regarding the baselines, we observe that the combination of input modalities does not lead to best results in most of the cases. The latter entails the dominance of one of the two modalities in terms of exploitable information for solving the task of interest and it has already been demonstrated in previous papers as well [44]. However, MemeFier effectively exploits both modalities and outperforms all baselines and state-of-the-art methods in all tasks (a, b and c) of this dataset.

MemeFier	**80.1**
- External knowledge	78.7
- Caption supervision	79.1
- Fusion stage 1	72.8
- Fusion stage 2	67.4

Table 4: Ablation analysis of MemeFier on Facebook Hateful Memes dataset. The performance (AUC) of the proposed method after removing (denoted by hyphen) components one by one.

Table 3 illustrates the performance of MemeFier and competitive methods on MultiOFF. Image surpasses text, while the combination of the two modalities leads to the best outcome both in terms of accuracy and F1 score. The proposed method's performance is better than the baselines in terms of accuracy while almost identical in terms of F1 score. Also, it is comparable but lower than the state of the art.

Finally, Table 4 provides an ablation analysis. We ablate the external knowledge input, the caption supervision, the first and the second fusion stage, respectively. We observe that all MemeFier components are necessary to provide the best performance as removing any of them results in reduced accuracy.

6 Conclusions

In this work we propose MemeFier, a deep learning-based architecture to address the task of Internet image meme fine-grained classification. Our method utilizes a dual-stage modality fusion module to first estimate modality alignment at token level via element-wise multiplication and then search for input dependencies at various scales through a typical Transformer encoder. In addition, external knowledge pertinent to protected attributes of the depicted persons is incorporated for joint processing and caption supervision is imposed as a regularization factor. Our pipeline competes and in some cases surpasses state-of-the-art methodologies on three widely-adopted meme classification benchmarks.

Acknowledgments

This work is partially funded by the Horizon 2020 European project MediaVerse under grant agreement no. 957252.

References

[1] Christos Koutlis, Manos Schinas, and Symeon Papadopoulos. Memetector: Enforcing deep focus for meme detection. *arXiv preprint arXiv:2205.13268*, 2022.

[2] Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas Pykl, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. Semeval-2020 task 8: Memotion analysis-the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, 2020.

[3] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Casey A Fitzpatrick, Peter Bull, Greg Lipstein, Tony Nelli, Ron Zhu, et al. The hateful memes challenge: Competition report. In *NeurIPS 2020 Competition and Demonstration Track*, pages 344–360. PMLR, 2021.

[4] Lambert Mathias, Shaoliang Nie, Aida Mostafazadeh Davani, Douwe Kiela, Vinodkumar Prabhakaran, Bertie Vidgen, and Zeerak Waseem. Findings of the WOAH 5 shared task on fine grained hateful memes detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 201–206, Online, August 2021. Association for Computational Linguistics.

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

[6] Shivam Sharma, Firoj Alam, Md Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, Tanmoy Chakraborty, et al. Detecting and understanding harmful memes: A survey. *arXiv preprint arXiv:2205.04274*, 2022.

[7] Anusha Chhabra and Dinesh Kumar Vishwakarma. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, pages 1–28, 2023.

[8] Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444, 2023.

[9] Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 32–41, 2020.

[10] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[11] Lisa Bonheme and Marek Grzes. SESAM at SemEval-2020 task 8: Investigating the relationship between image and text in sentiment analysis of memes. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 804–816, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[12] Xinghao Yang, Weifeng Liu, Wei Liu, and Dacheng Tao. A survey on canonical correlation analysis. *IEEE Transactions on Knowledge and Data Engineering*, 33(6):2349–2368, 2021.

[13] Pei Ling Lai and Colin Fyfe. Kernel and nonlinear canonical correlation analysis. *International journal of neural systems*, 10(05):365–377, 2000.

[14] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1247–1255, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[15] Xiaoyu Guo, Jing Ma, and Arkaitz Zubiaga. NUAA-QMUL at SemEval-2020 task 8: Utilizing BERT and DenseNet for Internet meme emotion analysis. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 901–907, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[16] Sunil Gundapu and Radhika Mamidi. Gundapusunil at SemEval-2020 task 8: Multimodal memotion analysis. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1112–1119, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[17] Yingmei Guo, Jinfa Huang, Yanlong Dong, and Mingxing Xu. Guoym at SemEval-2020 task 8: Ensemble-based classification of visuo-lingual metaphor in memes. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1120–1125, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[18] Nayan Varma Alluri and Neeli Dheeraj Krishna. Multi modal analysis of memes for sentiment extraction. In *2021 Sixth International Conference on Image Information Processing (ICIIP)*, volume 6, pages 213–217, 2021.

[19] Arup Baruah, Kaushik Das, Ferdous Barbhuiya, and Kuntal Dey. IIITG-ADBU at SemEval-2020 task 8: A multimodal approach to detect offensive, sarcastic and humorous memes. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 885–890, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[20] Pradyumna Gupta, Himanshu Gupta, and Aman Sinha. DSC IIT-ISM at SemEval-2020 task 8: Bi-fusion techniques for deep meme emotion analysis. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 876–884, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[21] Yazhou Zhang, Lu Rong, Xiang Li, and Rui Chen. Multi-modal sentiment and emotion joint analysis with a deep attentive multi-task learning model. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty, editors, *Advances in Information Retrieval*, pages 518–532, Cham, 2022. Springer International Publishing.

[22] Gitanjali Kumari, Amitava Das, and Asif Ekbal. Co-attention based multimodal factorized bilinear pooling for Internet memes analysis. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 261–270, National Institute of Technology Silchar, Silchar, India, December 2021. NLP Association of India (NLPAI).

[23] Qi Zhong, Qian Wang, and Ji Liu. Combining knowledge and multi-modal fusion for meme classification. In Björn Þór Jónsson, Cathal Gurrin, Minh-Triet Tran, Duc-Tien Dang-Nguyen, Anita Min-Chun Hu, Binh Huynh Thi Thanh, and Benoit Huet, editors, *MultiMedia Modeling*, pages 599–611, Cham, 2022. Springer International Publishing.

[24] Gokul Karthik Kumar and Karthik Nandakumar. Hate-CLIPper: Multimodal hateful meme classification based on cross-modal interaction of CLIP features. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 171–183, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[25] Ingroj Shrestha and Jonathan Rusert. NLP_UIOWA at SemEval-2020 task 8: You're not the only one cursed with knowledge - multi branch model memotion analysis. In *Proceedings of the Fourteenth Workshop on Semantic*

Evaluation, pages 891–900, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[26] Li Yuan, Jin Wang, and Xuejie Zhang. YNU-HPCC at SemEval-2020 task 8: Using a parallel-channel model for memotion analysis. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 916–921, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[27] Vishal Keswani, Sakshi Singh, Suryansh Agarwal, and Ashutosh Modi. IITK at SemEval-2020 task 8: Unimodal and bimodal sentiment analysis of Internet memes. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1135–1140, Barcelona (online), December 2020. International Committee for Computational Linguistics.

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

[29] Niklas Muennighoff. Vilio: State-of-the-art visio-linguistic models applied to hateful memes. *arXiv preprint arXiv:2012.07788*, 2020.

[30] Riza Velioglu and Jewgeni Rose. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *arXiv preprint arXiv:2012.12975*, 2020.

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

[32] Vlad Sandulescu. Detecting hateful memes using a multimodal deep ensemble. *arXiv preprint arXiv:2012.13235*, 2020.

[33] Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. Disentangling hate in online memes. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 5138–5147, New York, NY, USA, 2021. Association for Computing Machinery.

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

[35] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021.

[36] Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. Once for all: Train one network and specialize it for efficient deployment. In *International Conference on Learning Representations*, 2020.

[37] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.

[38] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009.

[39] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[40] Yi Zhou, Zhenhao Chen, and Huiyuan Yang. Multimodal learning for hateful memes detection. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6, 2021.

[41] Efrat Blaier, Itzik Malkiel, and Lior Wolf. Caption enriched samples for improving hateful memes detection. *arXiv preprint arXiv:2109.10649*, 2021.

[42] George-Alexandru Vlad, George-Eduard Zaharia, Dumitru-Clementin Cercel, Costin Chiru, and Stefan Trausan-Matu. UPB at SemEval-2020 task 8: Joint textual and visual modeling in a multi-task learning architecture for memotion analysis. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1208–1214, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[43] Sofiane Ouaari, Tsegaye Misikir Tashu, and Tomáš Horváth. Multimodal feature extraction for memes sentiment classification. In *2022 IEEE 2nd Conference on Information Technology and Data Science (CITDS)*, pages 285–290. IEEE, 2022.

[44] Kaushik Amar Das, Arup Baruah, Ferdous Ahmed Barbhuiya, and Kuntal Dey. KAFK at SemEval-2020 task 8: Extracting features from pre-trained neural networks to classify Internet memes. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1148–1154, Barcelona (online), December 2020. International Committee for Computational Linguistics.