DD-TIG at Constraint@ACL2022: Multimodal Understanding and Reasoning for Role Labeling of Entities in Hateful Memes

Ziming Zhou[2], **Han Zhao**[1], **Jingjing Dong**[2], **Jun Gao**[1], **Xiaolong Liu**[1]
[1]DD-TIG
[2]Peking University
{zhaohan,gaojun_i,xlongliu}@didiglobal.com
{zhouziming,djj}@stu.pku.edu.cn

Abstract

The memes serve as an important tool in online communication, whereas some hateful memes endanger cyberspace by attacking certain people or subjects. Recent studies address hateful memes detection while further understanding of relationships of entities in memes remains unexplored. This paper presents our work at the Constraint@ACL2022 Shared Task: Hero, Villain and Victim: Dissecting harmful memes for semantic role labelling of entities. In particular, we propose our approach utilizing transformer-based multimodal models through a visual commonsense reasoning (VCR) method with data augmentation, continual pretraining, loss re-weighting, and ensemble learning. We describe the models used, the ways of preprocessing and experiments implementation. As a result, our best model achieves the Macro F1-score of 54.707 on the test set of this shared task[1].

1 Introduction

Memes are getting popular as a communication tool on social media platforms for expressions of opinions and emotions, conveying a subtle message through multimodal information from both images and texts. However, memes are increasingly abused to spread hate instigate social unrest and therefore seem to be a new form of expression of hate speech on online platforms (Bhattacharya, 2019).

Automatic hateful memes detection is difficult since it primarily requires context and external knowledge to understand online speech, which sometimes can be very short and contains nuanced meaning (Pramanick et al., 2021). A new type of challenging task has been introduced by The Hateful Memes Challenge (Kiela et al., 2020) proposed by Facebook AI to leverage machine learning models to address hateful memes detection problems, which can only be solved by joint reasoning and understanding of visual and textual information (Zhu, 2020).

In previous studies, researchers focus on binary classification problems, labelling a meme as hateful or non-hateful based on image and text features (Afridi et al., 2020). Moreover, the relationships of entities in memes remain unexplored, and the task of role labelling of entities in hateful memes can be more sophisticated.

The Constraint@ACL2022 Shared Task: Hero, Villain and Victim: Dissecting harmful memes for semantic role labelling of entities offers us a perspective on this issue (Sharma et al., 2022). This task aims to promote the detection and classification of glorified, vilified or victimized entities within a meme. The shared dataset concerns memes from US Politics domains and Covid-19. Covid-19-related online hostile content especially demands to be detected as early as possible after their appearance on social media.

In this paper, we present our work on this task. Specifically, mainstream multimodal models of transformer-based architecture are applied through a visual commonsense reasoning (VCR) method, with the leverage of continual pretraining to fit models with our dataset. Then, data augmentation and loss re-weighting are implemented to improve the performance of models. The predictions from variant models are combined in a machine learning method to produce final results.

2 Related Work

Hateful memes understanding and reasoning is a vision and language task. Current state-of-the-art Vision-Language machine learning models are based on the transformer architecture (Vaswani et al., 2017). Multimodal models learn the joint visual and textual representations through self-supervised learning that utilize large-scale unlabelled data to conduct auxiliary tasks (Chen et al., 2022), including masked language modelling based

[1]https://github.com/zjl123001/DD-TIG-Constraint

Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations, pages 12 - 18
May 27, 2022 ©2022 Association for Computational Linguistics

on randomly-masked sub-words, masked region prediction and image-text matching. Among these models, there are two prevalent approaches: single-stream and dual-stream (Du et al., 2022).

In single-stream architecture, the representations of two modalities are learned by a single transformer encoder. Particularly, the text embeddings $L = \{w_1, w_2, w_3, \cdots, w_l\}$ and image features $V = \{o_1, o_2, o_3, \cdots, o_k\}$ are concatenated together as $X = \{L \parallel V\}$, added some special embeddings to indicate position and modalities, and fed into a transformer-based encoder.

There are many implementations in single-stream models, such as VisualBERT (Li et al., 2019), UNITER (Chen et al., 2020), OSCAR (Li et al., 2020).

In dual-stream models, the image and text features are first sent to two independent encoders. Then two features are separately fed into cross-modal transformer layers, where the query vectors are from one modality while the key and value vectors are from another. They are responsible for exchanging the information and aligning the semantics between the two modalities L and V. The formula of cross-modal transformer layers is represented as follows.

$$L_i^m = CrossAtt_{L-V}(L_i^{m-1}, \{V_1^{m-1}, \cdots, V_k^{m-1}\}) \quad (1)$$

$$V_i^m = CrossAtt_{V-L}(V_i^{m-1}, \{L_1^{m-1}, \cdots, L_l^{m-1}\}) \quad (2)$$

where m is the m^{th} cross-attention layer, k is the number of visual tokens, and l is the length of text tokens.

Following each cross-attention layer, there is also a layer computing the self-attention of each modality independently. Features are combined at the end of the model.

Several dual-stream models have been proposed in former studies, such as LXMERT (Tan and Bansal, 2019), ERNIE-Vil (Yu et al., 2020), DeVLBERT (Zhang et al., 2020), VilBERT (Lu et al., 2019),

3 Task Definition

Given the image and transcribed text of a meme, the role of a certain entity in this meme will be determined as hero, villain, victim or other, which can be interpreted as a multi-class classification task.

- **Input:** a meme image V, text transcriptions L, a entity E
- **Output:** $y \in \{hero, villain, victim, other\}$

The official evaluation measure for the shared task is the macro-F1 score for the multi-class classification.

4 Data Composition

The dataset provided in this task is a combination of memes from Covid-19 and US Politics domain. Every sample in the train and validation set contains an image, a transcription of texts and a list of entities with annotated labels. The shared task organizers provide the definitions for each class [2]:

- **Hero:** the entity is presented in a positive light, glorified for its actions.
- **Villain:** the entity is portrayed negatively, e.g., in an association with adverse traits like wickedness, cruelty, hypocrisy, etc.
- **Victim:** the entity is portrayed as suffering the negative impact of someone else's actions or conveyed implicitly within the meme.
- **Other:** the entity is not a hero, a villain, or a victim.

We present the distribution of entities' roles in Table 1.

	Covid-19			
	Hero	**Villain**	**Victim**	**Other**
Train	190	662	360	6022
Val	20	81	48	674
Test	21	124	58	1087
	US Politics			
	Hero	**Villain**	**Victim**	**Other**
Train	285	1765	550	7680
Val	34	224	73	915
Test	31	226	56	830

Table 1: Numbers of sample for each role label in Covid-19 and US Politics domain

There is a considerable imbalance in the distribution of entities' roles where the "other" class accounts for more than 80 percent of the whole

[2] https://codalab.lisn.upsaclay.fr/competitions/906

dataset. Meanwhile, the distribution of entities' frequency also shows a disparity. We present some most frequent entities with their roles distribution in Figure 1.

	Hero	Villiain	Victim	Other
donald trump	47	560	68	708
coronavirus	3	68	12	661
joe biden	22	183	17	587
barack obama	39	90	28	488
mask	-	-	-	326
work from home	-	-	-	272
2020	-	35	-	167
democratic party	-	161	24	115

Figure 1: Roles distribution of most frequent entities

5 System Descriptions

5.1 Preparation

For visual feature preprocessing, we use the pretrained Mask-RCNN model provided in the detectron2 framework[3] to obtain the object detection based region feature embedding $V = [o_1, o_2, \cdots, o_k]$ of images. Detectron2 is proposed by Facebook AI with state-of-the-art detection and segmentation algorithms. Specifically, 50 boxes of 2048 dimensions region-based image features are extracted for every meme. For the text transcriptions, we make the content lower-case and remove punctuation and stopwords with NLTK library (Loper and Bird, 2002).

5.2 Vision and Language Models

Four mainstream multimodal models of VL transformer architectures are applied in this work, namely: VisualBERT, UNITER, OSCAR, and ERNIE-Vil.

VisualBERT (Li et al., 2019), known as the first image-text pre-training model, uses the visual features extracted by Faster R-CNN, concatenates the visual features and textual embeddings, and then feeds the concatenated features to a single transformer initialled by BERT.

UNITER (Chen et al., 2020) learns contextualized joint representation of both visual and textual modalities through local alignment in the reconstruction of masked tokens/regions across modalities, powering heterogeneous downstream V+L tasks with joint multimodal embeddings.

OSCAR (Li et al., 2020), instead of simply using image-text pair, adds object tags detected from the image and represent the image-text pair as a <Word, Tag, Image> triple to help the fusion encoder better align different modalities.

ERNIE-Vil (Yu et al., 2020), as a typical dual-stream model, enhances the model with the application of scene utilizing scene graphs of visual scenes, which can learn the joint representations characterizing the alignments of the detailed semantics across vision and language.

For domain adaptation, we carry out continual pretraining on our dataset to reduce the distribution gap between the pretraining dataset and our memes dataset. Masked Language Modeling (MLM) pretraining task is taken on pretraining VisualBERT-large, UNITER-large, and OSCAR-large model.

5.3 VCR Implementation

Visual Commonsense Reasoning (VCR) focuses on a higher-order cognitive and commonsense understanding of relationships of the visual components in the image (Zellers et al., 2019). Former studies take a question, answer choices and an image into models to predict the right answer as a multi-class classification problem (Su et al., 2019). We modify this method's input and output format to conduct our experiments.

As can be seen in Figure 2, we concatenate the given entity and text tokens as the textual input with a separate token $[SEP]$, while different segment embedding will be added respectively to indicate their states. Then, textual input and visual will be concatenated in the single-stream model like VisualBERT. They would be separately sent into encoders in the dual-stream model like ERNIE-Vil. In the single-stream model, the final output feature of $[CLS]$ element is taken. In the dual-stream model, textual and visual features are fused through sum or multiplication. Then, features are fed to a linear layer with softmax to predict the role of the given entity.

The final objective is to minimize the cross-entropy (CE) loss between the predicted distribution and the targeted role category, which can be formally defined as:

[3] https://github.com/facebookresearch/detectron2

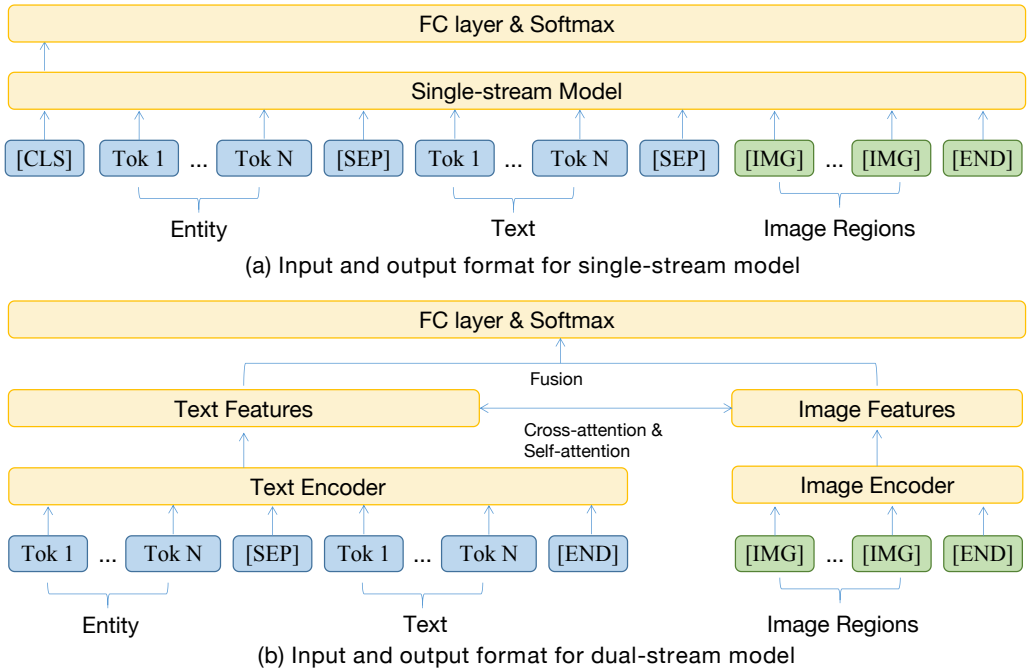

Figure 2: The input and output format of our system

$$p(x) = \frac{exp(g(x)_i)}{\sum_{j=1}^{N} exp(g(x)_j)} \tag{3}$$

$$L = -\sum logp(x) \cdot y \tag{4}$$

where $g(x)$ is the output of the FC layer and N is the number of labels.

5.4 Loss Re-weighting

A loss re-weighting strategy has been applied in our experiment since the "other" class accounts for the overwhelming majority of entries in samples, while hero, villain, and victim roles shall be stressed. Thus, our new loss function is defined as follows:

$$L = -\sum \alpha \cdot \log p(x) \tag{5}$$

$$\alpha = \begin{cases} \alpha_{neg} & y = other \\ \alpha_{pos} & else \end{cases} \tag{6}$$

where α_{neg} and α_{pos} are the weights for the "other" role and "non-other" role respectively as $\alpha_{neg} < \alpha_{pos}$ and $\alpha_{neg} + \alpha_{pos} = 1$.

5.5 Data Augmentation

We adopt the data augmentation with the back-translation strategy. Specifically, the provided text of each meme is paraphrased with Baidu translation API: English-Chinese-English and English-French-English. Diverse sentences are produced for each meme to enrich our dataset.

5.6 Ensemble Learning

We train these four base models with different seeds to produce a total of 16 models. The predicted scores on validation set are generated by all models. Then, a SVM model is trained with the predictions and true labels. In the testing phase, the predictions on the test set are fed into the trained SVM model to make final ensemble predictions.

5.7 Experimental Setting

For continual pretraining on VisualBERT, OSCAR, and UNITER, each word in the text transcriptions is randomly masked at a probability of 15 percent. The final output feature corresponding to the masked word is fed into a classifier over the whole vocabulary, driven by softmax cross-entropy loss.

We finetune all models with a focal loss (Lin et al., 2017) and a batch size of 16. The max sequence length is set at 256. The Adam optimizer is used with a learning rate of 1e-5 and 10 percent linear warm-up steps. VisualBERT, OSCAR, and UNITER are trained for 10 epochs and ERNIE-Vil models are trained for 10000 steps. The weights with the best scores on the validation set are saved and used for inference on the test set.

Source	Model	Macro F1-score
Original model	VisualBERT-large	47.8
	UNITER-large	48.8
	OSCAR-large	48.5
	ERNIE-Vil-large	50.9
Continual pretrained model	VisualBERT-large	48.2
	UNITER-large	49.9
	OSCAR-large	49.2
	Ensemble	**54.7**

Table 2: Results of models in our systems

6 Results and Discussion

In Table 2, we present the results of our experiments in a step by step manner. We started with finetuning base models provided by original authors. Then, VisualBERT-large, UNITER-large, and OSCAR-large models are pretrained on our dataset with MLM task and finetuned on our task. After that, ensemble learning is implemented to combine results of various models. We evaluate our models using official metrics Macro F1-score on test set.

ERNIE-Vil has been the SoTA model on the multimodal task leaderboard and in this task it also achieves competitive performance at 50.9 on the test set without further continual pretraining, which outperforms all the single-stream models by over 2 in Macro F1-score. We consider that through incorporating structured knowledge obtained from scene graphs during cross-modal pretraining, ERNIE-Vil learns more knowledge which benefits the downstream task.

Meanwhile, VisualBERT-large, UNITER-large, and OSCAR-large models shows improvements in performance through continual pretraining, which can be interpreted as domain adaptation on our dataset.

Ensemble learning remarkably raises our score by 3.5 than the best single model, which achieves the best score for our submission in this task.

6.1 Error Analysis

A classification report is presented in table 3, which allows us to do further assessments on our system.

Our system has a relatively poor performance on the class Hero. On the one hand, we interpret it as a lack of sample of this class in the training set. It is insufficient for our model to learn the features of this class. On the other hand, through observing bad cases, we find some memes need considerable external knowledge about history and politics, which can even be challenging for human beings to comprehend and do classification.

	precision	recall	f1-score	support
Hero	0.31	0.33	0.32	52
Villain	0.55	0.50	0.52	350
Victim	0.44	0.41	0.43	114
Other	0.88	0.89	0.89	1917
Macro-avg	0.54	0.53	0.54	2433

Table 3: An classification report for our final submission

6.2 Future Directions

In our experiment, we use an End2End solution to do roles classification, concatenating the entity with input sequence as a <entity, text, image> triplet. However, we do not directly point out the entity's corresponding region in the image. Some other researchers (Li et al., 2020) have discussed this problem: it is naturally weakly-supervised learning since there are no explicitly labelled alignments between regions or objects in an image and words or phrases in the text. We hypothesize that our model can not align some unusual entities correctly with its image and text. Moreover, comprehending a meme in the political domain heavily relies on knowledge, while the size of the whole dataset is relatively small, so our continual pretraining on a task-specific dataset is far from sufficient. There are two directions for further development of our system on this issue. On the one hand, more in-domain data can be incorporated to enlarge the dataset. On the other hand, knowledge-based models or external knowledge sources can be introduced to help the model understand the background and reason the relations of entities.

7 Conclusion

In this paper, we have exploited a VCR approach to tackle the role labelling of entities in hateful memes, which is a novel task in multimodal understanding and reasoning. Four popular transformer-based multimodal models, VisualBERT, UNITER, OSCAR, and ERNIE-Vil are applied as base models while strategies like loss re-weighting and data augmentation are implemented during the training of models. Then, continual pretraining is taken for domain adaptation and achieves better performance. Ensemble learning of variant models achieves the impressive Macro F1-score of 0.5470 on the final (unseen) test set.

References

Tariq Habib Afridi, Aftab Alam, Muhammad Numan Khan, Jawad Khan, and Young-Koo Lee. 2020. A multimodal memes classification: A survey and open research issues. In *The Proceedings of the Third International Conference on Smart City Applications*, pages 1451–1466. Springer.

Prithvi Bhattacharya. 2019. Social degeneration through social media: A study of the adverse impact of 'memes'. In *2019 Sixth HCT Information Technology Trends (ITT)*, pages 44–46. IEEE.

Feilong Chen, Duzhen Zhang, Minglun Han, Xiuyi Chen, Jing Shi, Shuang Xu, and Bo Xu. 2022. Vlp: A survey on vision-language pre-training. *arXiv preprint arXiv:2202.09061*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. 2022. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*.

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.

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*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

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.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md Akhtar, Preslav Nakov, Tanmoy Chakraborty, et al. 2021. Detecting harmful memes and their targets. *arXiv preprint arXiv:2110.00413*.

Shivam Sharma, Tharun Suresh, Atharva Jitendra, Himanshi Mathur, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022. Findings of the constraint 2022 shared task on detecting the hero, the villain, and the victim in memes. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations - CONSTRAINT 2022, Collocated with ACL 2022*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731.

Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. 2020. Devlbert: Learning deconfounded visio-linguistic representations. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4373–4382.

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