Codewithzichao@DravidianLangTech-EACL2021: Exploring Multimodal Transformers for Meme Classification in Tamil Language

Zichao Li
School of Software and Microelectronics, Peking University, China
lizichao@pku.edu.cn

Abstract

This paper describes our submission to shared task on Meme Classification for Tamil Language. To address this task, we explore a multimodal transformer for meme classification in Tamil language. According to the characteristics of the image and text, we use different pretrained models to encode the image and text so as to get better representations of the image and text respectively. Besides, we design a multimodal attention layer to make the text and corresponding image interact fully with each other based on cross attention. Our model achieved 0.55 weighted average F1 score and ranked first in this task.

1 Introduction

In recent years, with the prosperity of social media platforms, memes have gradually become a part of online communication. Therefore, it is essential to detect whether memes are offensive to individuals or organizations to ensure the diversity and sustainability of content on the Internet. It it a challenging task to classify whether memes are troll or not. In addition, there has been a lot of work currently focused on English (Truong and Lauw, 2019; Xu et al., 2019; Cai et al., 2019), but little work has been done for Tamil language.

Shared task on Meme Classification for Tamil Language fills this gap. The goal of this shared task is to detect whether memes which are collected from social media platforms are troll or not. Each meme has been annotated with troll or not troll class. Furthermore, a transcription of captions in Latin script for both Tamil is embedded in each image. This is a multimodal classification task that given the image and text pair, systems have to classify this pair into troll or not troll class.

In this paper, we explore a multimodal transformer for meme classification on Tamil language. According to the characteristics of the image and text, we use different pre-trained models to encode the image and text so as to get better representations of the image and text respectively. Besides, due to the particularity of social media text, in many cases we can only understand the meaning of text through the corresponding image, so it is essential to make the text and corresponding image interact fully with each other. To tackle this issue, we design a multimodal attention layer based on cross attention. Our model took first place in this task.

Class	Train	Test
troll	1282	395
not troll	1018	272

Table 1: Statistics of the train and test datasets.

2 Data

The data we used is provided by the organizers of shared task on Meme Classification in Tamil Language (Suryawanshi et al., 2020; Suryawanshi and Chakravarthi, 2021). There are 2300 samples in the training data. The specific statistics of the data are shown in Table 1.

2.1 Text Preprocessing

There are two methods used to preprocess social media text as follow:

- **Noise removal:** Emojis and extra blanks in the training data are removed in advance. Experimental results show that removing these noise can improve the performance of our model. The maximum sequence size is 256.
- **Tokenization:** Texts are tokenized using the sentencepiece toolkit[1] and converted to the corresponding IDs through the vocabulary of XLM-RoBERTa (Conneau et al., 2020).

[1] https://github.com/google/sentencepiece

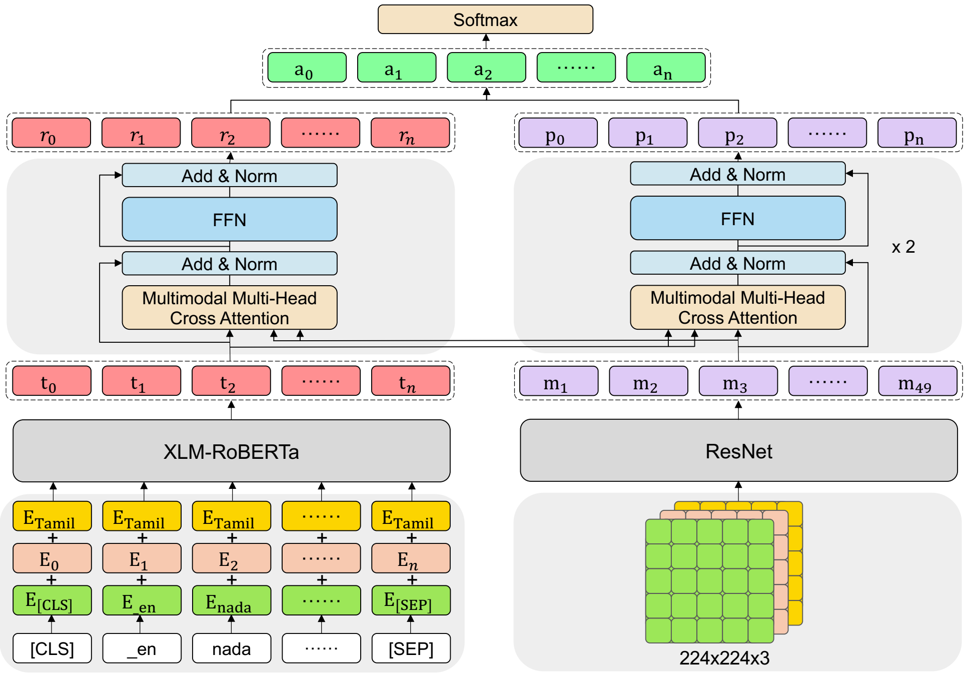

Figure 1: Our model architecture.

2.2 Image Preprocessing

Similar to the image preprocessing method in ImageNet (Deng et al., 2009), each image is cropped and scaled so that the dimension size of each image is 224×224×3.

3 Proposed Model

In this section, we will present our model for meme classification on Tamil language. Our model is mainly divided into three layers: encoding layer, multimodal attention layer and prediction layer. Encoding layer is used to obtain word representations and image representations. Multimodal attention layer is used to make the text and corresponding image interact fully with each other. Prediction layer is used to get the probabilities of all classes. Overall model architecture is shown in Figure 1.

3.1 Encoding Layer

Text Encoding: Compared with RNN such as LSTM (Hochreiter and Schmidhuber, 1997) or GRU (Chung et al., 2014), the pre-trained language model like XLM-RoBERTa can learn better contextual representations of text and also helps in fighting vanishing and exploding gradient descent. Therefore, we use XLM-RoBERTa as the encoder of the social media text. Given a sentence $X = \{x_i\}_{i=0}^{n}$, where n is equal to 256 and x_i is the sum of token embedding, position embedding and language embedding of token at position i and the dimension size of x_i is d, we can obtain contextual representation $T = \{t_i\}_{i=0}^{n}$ for each word:

$$T = XLM{-}RoBERTa(X). \tag{1}$$

Image Encoding: Since ResNet (He et al., 2016) is currently the most widely used image feature extraction network, we use ResNet with 152 layers as the encoder of the image to extract the features for each image $I \in R^{224\times224\times3}$. The output of the last layer which is denoted as $\hat{M} = \{\hat{m}_i\}_{i=1}^{49}$ is used to represent an image:

$$\hat{M} = ResNet(I). \tag{2}$$

Besides, we use a linear transformation to make the dimension size of the image representations consistent with the word representations: $M = W_M^T \hat{M}$, where $W_M \in R^{2048*d}$ and the length of M is 49.

3.2 Multimodal Attention Layer

This layer is the core of our model. Due to the particularity of social media text, although we have obtained representations of the image and text respectively, in many cases we can only understand the meaning of text through the corresponding image, so it is essential to make the text and corresponding image interact fully with each other. To tackle this issue, we designed a multimodal attention layer based on cross attention (Tsai et al., 2019).

Team	Metric	Precision	Recall	F1-score
Codewithzichao	weighted avg	**0.57**	**0.60**	**0.55**
IIITK	weighted avg	0.56	0.59	0.54
NLP@CUET	weighted avg	0.55	0.58	0.52
SSNCSE_NLP	weighted avg	0.58	0.60	0.50
Simon_work	weighted avg	0.53	0.58	0.49

Table 2: Top-5 of the official leader-board in shared task for meme classification in Tamil language. Systems are ordered by weighted average F1 score.

Model	Metric	Precision	Recall	F1-score
Our submission	weighted avg	**0.57**	**0.60**	**0.55**
w/o Multimodal Attention Layer	weighted avg	0.56	0.59	0.54

Table 3: Ablation study of our model in the test dataset. w/o means without.

Multimodal Multi-Head Cross Attention: Similar to multi-head self attention used in Transformer (Vaswani et al., 2017), multimodal multi-head cross attention (**MMHCA**) has three input vectors: $Q = \{w_i\}_{i=1}^{n_Q}$, $K = \{w_i\}_{i=1}^{n_K}$, $V = \{w_i\}_{i=1}^{n_V}$, where w_i refers to a d dimension vector and n_K is equal to n_V. Attention results are defined as:

$$Attn_i(Q_i, K_i) = softmax(\frac{(W_{Q_i}Q_i)(W_{K_i}K_i^T)}{\sqrt{d/m}}) \tag{3}$$

$$V_i^{attn} = Attn_i(Q_i, K_i)(W_{V_i}V_i) \tag{4}$$

$$MMHCA(Q, K, V) = [V_1^{attn}; ...; V_m^{attn}]. \tag{5}$$

Attentive Word Representations: To obtain word representations for each image, we use MMHCA, which take M as queries and T as keys and values. The attentive word representations are defined as :

$$\hat{R} = LN(M + MMHCA(M, T, T)); \tag{6}$$

$$R = LN(\hat{R} + FFN(\hat{R})), \tag{7}$$

where LN is the layer normalization (Ba et al., 2016) and FFN is the feed-forward network. We use MMHCA again to obtain final attentive word representations.

Attentive Image Representations: To obtain image representations for each word, we use MMHCA, which take T as queries and M as keys and values. The attentive image representations are defined as :

$$\hat{P} = LN(M + MMHCA(T, M, M)); \tag{8}$$

$$P = LN(\hat{P} + FFN(\hat{P})). \tag{9}$$

Hyper-parameters	Value
Batch size	8
Epoch	50
Learning rate	2e-5
Gradient clipping	0.25
Dropout rate	0.2

Table 4: Hyper-parameters of our model.

After obtaining attentive word representations and attentive image representations, we concatenate them as the output of this layer: $A = [R; P]$.

3.3 Prediction Layer

To classify each image and text pair, we feed A to a average-over-time pooling layer and then use softmax to get the probabilities of all classes:

$$Z = AvgPool(A); \tag{10}$$

$$P(y|X, I) = softmax(WZ + b), \tag{11}$$

where A is the output of multimodal attention layer. We use focal loss (Lin et al., 2017) to train our model:

$$L = -\sum_{\{X,I\} \in S} \alpha(1 - P(y|X, I))^{\gamma} \log P(y|X, I), \tag{12}$$

where S refers to the train dataset.

4 Experiment and Results

4.1 Experimental Settings

We use Pytorch (Paszke et al., 2017) and HuggingFace's transformers (Wolf et al., 2020) to implement our model. We use XLM-RoBERTa and

ResNet-152 as encoders for the text and image, respectively. We use mixed precision training based on Apex library[2]. AdamW (Loshchilov and Hutter, 2019) optimizer is used to optimize our model with a learning rate at 2e-5. We use 5-fold cross validation to obtain better performance. We use adversarial training (i.e. FGM (Goodfellow et al., 2015)) to further improve the robustness and generalization ability of our model. We list all hyper-parameters of our model in Table 4. we conduct the experiments on NVIDIA Tesla T4 GPUs. Our code is available at Github[3].

4.2 Results and Ablations

The top five results in this task have been shown in Table 2. Our model achieved 0.55 weighted average F1 score and ranked first. Besides, our result is 0.01 higher than the second.

In addition, to prove the effectiveness of multimodal attention layer, we conduct the ablation experiment. The ablation result is shown in Table 3. When multimodal attention layer is removed, the final result drops by 0.01, which indicates that multimodal attention layer is considerably useful.

5 Conclusion

In this paper, we present a multimodal transformer for meme classification on Tamil language. Using ResNet and XLM-RoBERTa for the image and text, we obtain better representations of the image and text. Besides, we design a multimodal attention layer to make the text and corresponding image interact fully with each other. Finally, our model took first place in this task which demonstrates the effectiveness of our model. In future research, we will explore ways to better filter irrelevant information in the image and text to obtain better performance.

References

Jimmy Ba, J. Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *ArXiv*, abs/1607.06450.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515, Florence, Italy. Association for Computational Linguistics.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online.

J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

I. Loshchilov and F. Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Adam Paszke, S. Gross, Soumith Chintala, G. Chanan, E. Yang, Zachary Devito, Zeming Lin, Alban Desmaison, L. Antiga, and A. Lerer. 2017. Automatic differentiation in pytorch. In *NIPS-W*.

Shardul Suryawanshi and Bharathi Raja Chakravarthi. 2021. Findings of the shared task on Troll Meme Classification in Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John Philip McCrae, and Paul Buitelaar. 2020. A dataset for troll classification of TamilMemes. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 7–13, Marseille, France. European Language Resources Association (ELRA).

Quoc-Tuan Truong and Hady W. Lauw. 2019. Vistanet: Visual aspect attention network for multimodal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):305–312.

[2] https://github.com/NVIDIA/apex

[3] https://github.com/codewithzichao/Multimodal-Transformers

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Florence, Italy.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online.

Nan Xu, Wenji Mao, and Guandan Chen. 2019. Multi-interactive memory network for aspect based multimodal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):371–378.