Alpha at SemEval-2021 Task 6: Transformer Based Propaganda Classification

Zhida Feng[1,2] *, Jiji Tang[1] *, Jiaxiang Liu[1], Weichong Yin[1], Shikun Feng[1], Yu Sun[1], Li Chen[2]
[1]Baidu Inc., Beijing, China
[2]Wuhan University of Science and Technology, China
{fengzhida, tangjiji, liujiaxiang, yinweichong, fengshikun01, sunyu02}@baidu.com
chenli@wust.edu.cn

Abstract

This paper describes our system participated in Task 6 of SemEval-2021: this task focuses on multimodal propaganda technique classification and it aims to classify given image and text into 22 classes. In this paper, we propose to use transformer-based (Vaswani et al., 2017) architecture to fuse the clues from both image and text. We explore two branches of techniques including fine-tuning the text pre-trained transformer with extended visual features and fine-tuning the multimodal pre-trained transformers. For the visual features, we experiment with both grid features extracted from ResNet(He et al., 2016) network and salient region features from a pre-trained object detector. Among the pre-trained multimodal transformers, we choose ERNIE-ViL (Yu et al., 2020), a two-steam cross-attended transformers model pre-trained on large-scale image-caption aligned data. Fine-tuning ERNIE-ViL for our task produces a better performance due to general joint multimodal representation for text and image learned by ERNIE-ViL. Besides, as the distribution of the classification labels is extremely unbalanced, we also make a further attempt on the loss function and the experiment results show that focal loss would perform better than cross-entropy loss. Lastly, we ranked first place at sub-task C in the final competition.

1 Introduction

Propaganda is usually adopted to influence the audience by selectively displaying the facts to encourage specific synthesis or perception, or using the loaded language to produce emotion rather than emotion itself. It was often associated with materials prepared by governments in the past century. In the internet era, activist groups, companies, religious organizations, the media, and individuals also produce propaganda, and sometimes it can reach very large audiences (Da San Martino et al., 2020). With the recent research interest in detecting "fake news", the detection of persuasion techniques in the texts and images has emerged as an active research area. Most previous work like (Patil et al., 2020) and (Chauhan and Diddee, 2020) have performed the analysis at the language content level only. However, in our daily life, memes consist of images superimposed with texts. The aim of the image in a meme is either to reinforce a technique in the text or to convey one or more persuasion techniques.

SemEval-2021 Task6-c offers a different perspective, multimodal multi-label classification (Dimitrov et al., 2021), identify which of the 22 techniques are used both in the textual and visual content of memes. Since memes are combinations of texts and images, for this propaganda classification task, we proposed to use transformer-based architecture to fuse the clues from both linguistic and visual modalities. Two branches of fine-tuning techniques are explored in this paper. First, a text pre-trained transformer is applied with extended visual features. Specifically, we initialize the transformer with pre-trained text transformers and fine-tune the model with extended visual features including grid features(e.g., ResNet(He et al., 2016)) and region features(e.g., BUTD (Anderson et al., 2018)) from an image feature extraction network and an object detector respectively. Second, pre-trained multimodal transformers from ERNIE-ViL(Yu et al., 2020) are used due to its better multimodal joint representations characterizing cross-modal alignments of detailed semantics.

Our contributions are three-folds:

- We propose to use transformer architecture for fusing the visual and linguistic clues to tackle the propaganda classification task.

*indicates equal contribution.

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 99–104
Bangkok, Thailand (online), August 5–6, 2021. ©2021 Association for Computational Linguistics

- We find that the multimodal pre-trained transformers work better than using text pre-trained transformers with visual features. And the experiment results have shown that fine-tuning the ERNIE-ViL model could achieve state-of-the-art performance for this task.
- Our ensemble result of several models obtains the best score and ranks first in Semeval-2021 Task 6-c multimodal classification task.

2 Related work

2.1 Text Transformers

Transformer network (Vaswani et al., 2017) is first introduced in neural machine translation in which encoder and decoder are composed of multi-layer transformers. After then, pre-trained language models, such as BERT (Devlin et al., 2018) and GPT(Radford et al., 2018), adopting transformer encoder as the backbone network, have significantly improved the performance on many NLP tasks. One of the main keys to their success is the usage of transformer to capture the contextual information for each token in the text via self-attention. Later text pre-training works, such as ERNIE2.0 (Sun et al., 2020), RoBERTa (Liu et al., 2019) and XLNET (Yang et al., 2019) are all shared the same multi-layer transformer encoder and mainly put their effort on modification of pre-training task.

2.2 Visual Feature Extraction

Visual feature extractors are mainly composed of plenty of convolutional neural networks (**CNN**) since CNN has a strong ability to extract complex features that express the image with much more details and learn the task-specific features much more efficiently. Existing works can be divided into the following two types which are based on two different image inputs: image grids and object regions. Some of those methods, such as VGG (Simonyan and Zisserman, 2014), ResNet (He et al., 2016) operate attention on CNN features corresponding to a uniform grid of equally-sized image regions. While the other works like Faster R-CNN (Ren et al., 2015) operate a two-stage framework, which firstly identifies the image regions containing the specific objects, and then encodes them with multi-layer CNNs.

2.3 Multimodal Transformers

Inspired by text pre-training models (Devlin et al., 2018), many cross-modal pre-training models for vision-language have been proposed. To integrate visual features and text features, recent multimodal pre-training works are mainly based on two variables of transformers. Some of them, like UNITER (Chen et al., 2019) and VILLA (Gan et al., 2020) use a uniform cross-modal transformer modelling both image and text representations. As fine-tuning on multimodal classification tasks, such as the Visual-question-answering (**VQA**) (Antol et al., 2015) task (a multi-label classification task), unified transformers take textual and visual features as the model input, treat the final hidden state of $h_{[CLS]}$ as the vision-language feature. While the others like Vilbert (Lu et al., 2019), LXMERT (Tan and Bansal, 2019), ERNIE-ViL (Yu et al., 2020) are based on two-stream cross-modal transformers, which bring more specific representations for image and text. These two transformers are applied to images and texts to model visual and textual features independently and then fused by a third transformer in a later stage. The fusion of the final hidden state of $h_{[CLS]}$ and $h_{[IMG]}$ are used to do the classification.

3 Approach

We propose to use a transformer encoder to fuse the clues from both linguistic and visual modalities and our approach is summarized in two branches, the first one is fine-tuning a text pre-trained transformer with extended visual features, and the other one is fine-tuning a multimodal pre-trained model. For the first one, we try two different sets of visual features, grid features based on equally-split patches of the image and salient region features based on an object detector. For the second one, a SoTA multimodal model, ERNIE-ViL (Yu et al., 2020) is applied with a multi-label classification loss. A unified framework for the two branches is shown in Figure 1. We will introduce more details in this section.

3.1 Text Pre-trained Transformer with Visual Features

Our model consists of three parts: a) input feature extractor, b) feature fusion encoder, c) classification encoder.

For the first part, the text is tokenized into subwords to lookup the embedding while the image is processed by a feature extractor, such as a grid feature processor or a salient region feature processor to convert into vision embeddings. The input em-

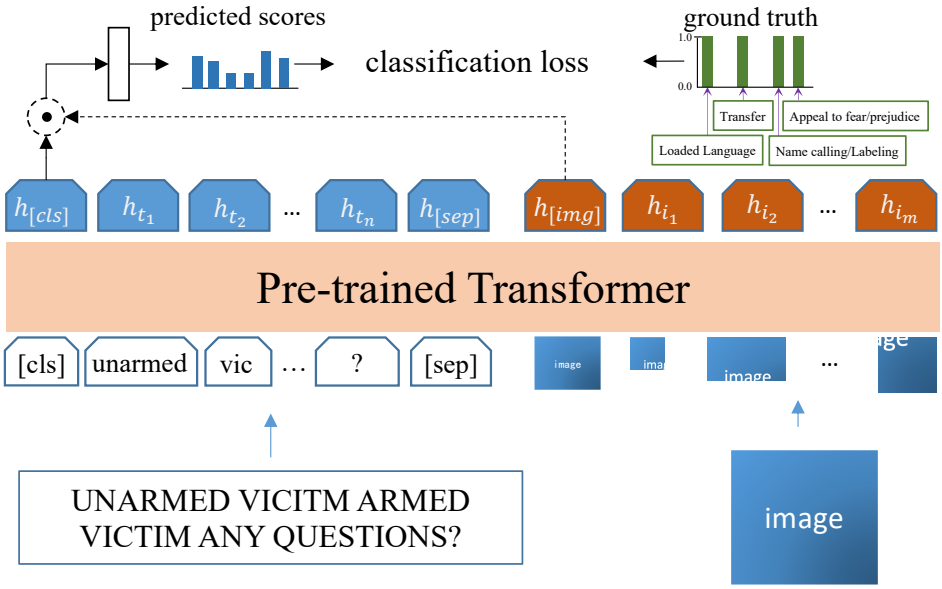

Figure 1: A unified framework used for the multimodal classification task.

beddings are combinations of image embeddings and text embeddings and represented as

$$h_{[CLS]}, h_{t_1}, \cdots, h_{t_n}, h_{[SEP]}, h_{i_1}, \cdots, h_{i_m}, h_{[SEP]}$$

where the $h_{[CLS]}$, $h_{[SEP]}$ are the vector representations of special tokens $[CLS]$ and $[SEP]$ respectively. The $[CLS]$ token is inserted in the beginning of the sequence, which act as an indicator of the whole text, specifically, it is used to perform complete text classification. The $[SEP]$ is a token to separate a sequence from the subsequent one and indicate the end of a text. $h_{t_1}, \cdots, h_{t_n}$ are the text embeddings, and $h_{i_1}, \cdots, h_{i_m}$ are the vision embeddings. For the vision embeddings part, grid features and salient region features are used.

Grid Features Convolutional neural networks have potent capabilities in image feature extraction. The feature map obtained after the image goes through multiple stacked convolution layers contains high-level semantic information. Given an image, we can use a pre-trained CNN encoder, such as ResNet, to transform it to a high-dimensional feature map and flatten each pixel on this feature map to form the final image representation.

Salient Region Features Object detection models are widely used to extract salient image regions from the visual scene. Given an image, we use a pre-trained object detector to detect the image regions. The pooling features before the multi-class classification layer are utilized as the region features. The location information for each region is encoded via a 5-dimension vector representing the fraction of image area covered and the normalized coordinates of the region and then is projected and summed with the region features.

For the second part, the transformer encoder fuses the input text and image embedding, and finally a cross-modal representation of size D is achieved for this sequence.

The last part of our model is the classification encoder and loss function. After obtaining the encoding representation of the image and the text from the transformer encoder, we send the representation of $[CLS]$ through the classification head, which is consisted of a fully connected layer and a $Sigmoid$ activation for predicting the score of each category and loss with the ground truth.

3.2 Multimodal Pre-trained Transformer

Different from a single-modal pre-trained text transformer described above, a multimodal pre-trained transformer for vision-language can learn more efficient presentations. In this part, a SoTA model, ERNIE-ViL, is applied.

For the generation of input embedding of text and image, it is mostly the same as the procedure described in the previous section. Differences are two-folds. First, for the vision feature, a faster R-CNN encoder(Anderson et al., 2018) is used to detect the salient regions while the position infor-

mation is taken into consideration. Second, The text and the visual input embedding is represented as

$$h_{[CLS]}, h_{t_1}, \cdots, h_{[SEP]}, h_{[IMG]}, h_{i_1}, \cdots, h_{[i_m]}$$

where there is a new token $h_{[IMG]}$ represents the feature for the entire image.

For the feature fusion part, ERNIE-ViL utilized a two steam cross-modal transformer to fuse the multimodal information. For more details, you may refer to (Yu et al., 2020).

3.3 Criterion

In this task, there are 22 classes and the distribution of positive and negative samples is extremely unbalanced. To solve this problem, we use the focal loss to improve the imbalance of positive and negative samples. For *i-th* class

$$L_{class_i} = \begin{cases} \alpha(1-p)^{\gamma} log(p) & \text{if y=1} \\ (1-\alpha)p^{\gamma} log(1-p) & \text{otherwise} \end{cases}$$

where y is the ground truth; p is model prediction, which is the confidence score of category i; α and γ are hyper-parameters, α is used to control the loss weight of positive and negative samples, and γ is used to scale the loss of difficult and easy samples.

4 Experiment

4.1 Implementation Details

In this task, we choose DeBERTa-large+ResNet50, DeBERTa-large+BUTD and ERNIE-VIL as the final models. We performed all our experiments on a Nvidia Tesla V100 GPU with 32 GB of memory. The models are trained for 20 epochs and we pick the model which has the best performance on validation set.

For the DeBERTa transformer, the Adam optimizer with a learning rate of 3e-5 is used. Also, we have applied the linear warm strategy for the learning rate. We set $\alpha = 0.9$ and $\gamma = 2.0$ for the focal loss. To ensure robustness under a small dataset, we set the threshold to 0.5 instead of performing a threshold search strategy on the validation set. For the pre-trained object detector, we choose Faster R-CNN (Anderson et al., 2018) and name the region features as BUTD in the experimental results.

For the ERNIE-ViL transformers, we use the same input prepossessing methods as (Yu et al., 2020) and choose the large scale model[1] pre-trained on all the four datasets. We finetune on our multimodal classification dataset with a batch of 4 and a learning rate of 3e-5 for 20 epochs.

	Positive(%)	Negative(%)
train	1745(11.55%)	13369(88.45%)
dev	183(13.20%)	1203(86.80%)
test	523(13.49%)	3877(86.51%)

Table 1: Statistics of the positive and negative distribution of the dataset.

Loss Function	Precision	Recall	F1
cross-entropy	76.12	55.74	64.35
focal loss	71.18	66.12	68.56

Table 2: Results of different loss functions.

4.2 Experimental Analysis

4.2.1 DeBERTa with Visual Features

Unbalanced Distribution There are 687/63/200 examples includes 22 categories in the train/validation/test datasets respectively. As shown in Table 1, the distribution of the classes is extremely unbalanced. If the cross-entropy loss is adopted directly during model training(the visual features are from ResNet50), the model output may have a greater chance of predicting the majority class(negative class in this task), which results in a lower recall. To solve this problem, the focal loss is applied. From Table 2, it can be seen that the result with focal loss performs much better than with cross-entropy loss respective to the F1 score.

Visual Features We evaluate the improvement brought by extended visual features and explore different types of visual feature extractors, e.g., from pre-trained image classification networks or pre-trained object detectors. The results are illustrated in Table 3. Firstly, it can be seen that the final score is significantly improved with mixing image features compared with using only text features (Row "w/o vision feature"), which indicates that the visual information is significantly beneficial for recognizing cross-modal propaganda techniques. Then, for features extracted from ResNet, we find that the depth of the network affects the results, especially on the validation dataset, with the best result from ResNet50. The reason may be

[1]the pre-trained model is downloaded from https://github.com/PaddlePaddle/ERNIE/tree/repro/ernie-vil

	dev-F1	test-F1
w/o vision feature	65.73	55.10
ResNet18	65.92	55.59
ResNet50	68.56	55.96
ResNet152	65.91	55.63
BUTD	66.29	56.21

Table 3: The results of using features extracted different networks.

region numbers	Dev F1	Test F1
5	64.91	54.00
10	66.67	54.60
36	67.40	57.14
100	67.45	56.07

Table 4: Results comparisons with different object region number inputs.

that the shallower network has insufficient feature extraction capabilities, and the deeper network is very difficult to train. Finally, the region features from the pre-trained object detector(Row "BUTD") work best with an improvement of 0.25 on the test dataset compared to ResNet50 features.

4.2.2 ERNIE-ViL

We compare the performance between ERNIE-ViL with different object region inputs, which are number dynamic ranges between 0 and 36 with a fixed confidence threshold of 0.2 and constantly fixed 5, 10, or 100 boxes. The results are illustrated in Table 4.

Results show that a larger box number can always achieve better performance within a certain range. Utilizing 0-36 boxes leads to huge performance improvement with a 3.14 and 2.54 on Test-F1 compared with using constant 5 boxes and constant 10 boxes respectively. It can be concluded that more object regions in a certain range can provide more useful information. However, the performance with 100 boxes is worse than that with 0-36 boxes. The reason may lie in that there are not enough objects in the task sample. The extracted low-confidence object regions may mislead the multimodal model, therefore fuse useless or harmful visual features with text features. As a result of that, brings a performance decrease on the final score.

Models	Dev-F1	Test-F1
DeBERTa + ResNet50	68.56	55.96
DeBERTa + BUTD	66.29	56.21
ERNIE-VIL	67.40	57.14
Ensemble	69.12	58.11

Table 5: Final ensemble result.

4.3 Ensemble Results

The performance comparison between our two branches of approach is shown in Table 5. It can be concluded that fine-tuning the multimodal pre-trained transformer (Row "ERNIE-ViL") works better than fine-tuning text pre-trained transformers with visual features (Row "DeBERTa + BUTD"). Overall, fine-tuning ERNIE-ViL has achieved state-of-the-art performance for this multimodal classification task.

Since the training dataset is small, we train multiple models under various model structures and different parameter configurations to take full advantage of the training dataset and increase the diversity of models. We choose three models of all model structures and all parameter configuration that performs best on the validation set and then ensemble them together. After performing ensemble strategy on those three models, both validation and test scores increases. As a result of that, we achieved a 58.11 score at F1 in the test set and ranked first place in the task competition.

5 Conclusion

We explore two branches to fine-tune pre-trained transformers to jointly modelling texts and images for the propaganda classification task. The first branch, fine-tuning pre-trained text transformer with visual feature, obtain significant performance improvement compared to text classification which validate the importance of visual clues for this task. Visual features from object detector yield slightly better results than grid features from ResNet. Importantly, fine-tuning pre-trained multimodal transformers obtain the best single model performance. And this improvement further validates the claim made by previous work that vision-language pre-training learned general joint representation needed for multimodal tasks. Besides, since the distribution of the classification labels is extremely unbalanced, we also make a further attempt on the loss function. Training models with focal loss can lead to a huge performance improvements than training with cross entropy loss.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Aniruddha Chauhan and Harshita Diddee. 2020. Psuedoprop at semeval-2020 task 11: Propaganda span detection using bert-crf and ensemble sentence level classifier. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1779–1785.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Learning universal image-text representations.

Giovanni Da San Martino, Alberto Barrón-Cedeno, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. Semeval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414.

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*.

Dimiter Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Task 6 at semeval-2021: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation*, SemEval '21, Bangkok, Thailand.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. *arXiv preprint arXiv:2006.06195*.

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*, pages 770–778.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.

Rajaswa Patil, Somesh Singh, and Swati Agarwal. 2020. Bpgc at semeval-2020 task 11: Propaganda detection in news articles with multi-granularity knowledge sharing and linguistic features based ensemble learning. *arXiv preprint arXiv:2006.00593*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975.

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, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

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.