Argumentative Stance Classification on ImageArg Datasets

Jiaxuan Chen jchenia@connect.ust.hk 21018177
Yingan Chen ychenly@connect.ust.hk 20972900
Hong Chang Su hsuaj@connect.ust.hk 20972649
Yi Lin Ye ylye@connect.ust.hk 20922644

Abstract

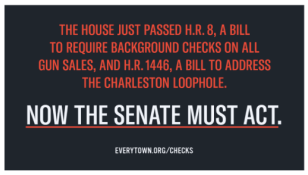
In this paper, we address the ImageArg shared task's Subtask-A: Argumentative Stance (AS) Classification, focusing on classifying multimodal tweets about gun control and abortion. Our methodologies encompass decision-level and feature-level multimodal fusion. The results reveal that our feature-level fusion method, with F1 scores of 0.85 for abortion and 0.8 for gun control, outperforms the decision-level approach, demonstrating its effectiveness in enhancing classification accuracy for this specific task.

1 Introduction

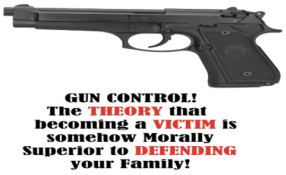
In the dynamic field of natural language processing (NLP), the integration of multimodal elements, particularly visual data alongside text, marks a significant evolution. The ImageArg dataset, introduced by Liu et al. (Liu et al., 2022b), exemplifies this shift, emphasizing the role of images in augmenting text-based persuasive communication. This novel approach addresses the gap in traditional text-centric methodologies, offering a more holistic view of persuasion in social media.

Our study centers on Subtask-A of the ImageArg shared task, part of the 10th Workshop on Argument Mining. This task focuses on two controversial topics, gun control and abortion, and challenges us to classify the argumentative stance of tweets that combine text and image. The primary goal is to determine whether each tweet supports or opposes the given topic, highlighting the synergy between textual and visual information in conveying persuasive messages. Figure 1 presents samples from Subtask A.

Our methodologies for this multimodal classification task involve proposing both decision-level and feature-level multimodal fusion approaches for classification. Our methods surpassed the baseline performances across two evaluation metrics, with the feature-level fusion approach achieving higher F1 scores of 0.85 and 0.8 for abortion and gun control, respectively, compared to 0.78 and 0.73 for the decision-level method.



Support: ANOTHER BIG WIN FOR GUN SAFETY! Just after passing a bill to require background checks on all gun sales #HR8 the U.S House just passed #HR1446, a bill that would address the deadly Charleston loophole. This bill now heads to the Senate.



Oppose: #GunControl The THEORY that becoming a VICTIM is somehow morally superior to DEFENDING yourself or family! We support #SelfDefence #Legalgunownership #SafeCitizen

Figure 1: Examples of Subtask-A: Argument Stance (AS) Classification: the tweet on the left supporting gun control by referencing a house bill for mandatory background checks, and on the right, the tweet opposing gun control in favor of self-defense

2 Related Work

2.1 Multimodal Learning

The field of artificial intelligence has seen significant advancements in multimodal learning, with researchers increasingly focusing on how models process and interpret combinations of text and visual data (Worsley, 2023). Researchers have created techniques to develop strong representative for each modality (Tsai et al., 2019) to achieve better result. Also, fusion techniques was implemented to increase model effectiveness (Liu et al., 2022a).

The interest of multimodel learning is driven by the need to mirror real-world applications where multiple input signals coexist and interact. Particularly, there's been growing interest in joint vision-language models, like OpenAI's CLIP (Radford et al., 2021). These models excel in complex tasks such as image captioning, text-driven image creation and editing, and answering questions based on images. Meanwhile, multimodel stance detec-

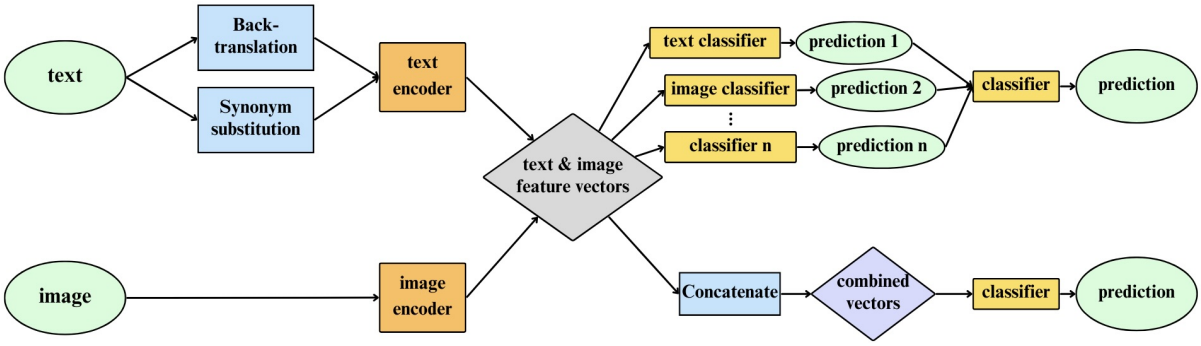


Figure 2: A sketch of our model, with two multimodel fusion method. Top right: multimodel fusion at decision level; Bottom right: multimodel fusion at feature level

tion is being increasingly used for social applications such as disaster response with tweet (Ofli et al., 2020) and fake news detection (Sengan et al., 2023). Recently, The ImageArg dataset by Liu et al. (Liu et al., 2022b) was introduced, allowing multimodal learning in argument mining. It combines textual and visual data from Twitter, providing a rich corpus for analyzing argumentative stances in social media.

2.2 Data Augmentation for Text

The main target of data augmentation is to enrich the training set since the limited size of dataset size may influence the generalization capability of the models.

Zhang et al.(Zhang et al., 2016) take a random word from the sentence and replace it with its synonym using a Thesaurus. Jiao et al. (Jiao et al., 2020) use the nearest neighbor words in the GloVec embedding space as the replacement for some word in the sentence. Garg and Ramakrishnan(Garg and Ramakrishnan, 2020) use masked language model to generate Adversarial Examples. Xie et al.(Xie et al., 2020) translate the original text to another Language e.g. French, and then translate it back into an English sentence, which is known as back-translation. Xie et al.(Xie et al., 2017) use blank noising

3 Methods

We applied two multimodel fusion methods for the task. An illustration is shown in Figure 2. Also, we applied two data augmentation methods to increase the size of the dataset since the size of the original training set is limited.

Multimodal Fusion at Decision Level: In our first approach, we employed Multimodal Fusion at the decision level. The fundamental strategy involved training multiple models on the original training set to obtain the corresponding predictions. Subsequently, these predictions were merged to form a new training set for an extra classifier. This approach enables the generation of a final prediction by leveraging the combined insights from the ensemble of models.

Multimodal Fusion at Feature Level: Our second method involved Multimodal Fusion at the feature level. Initially, we used separate text and image encoders to extract feature vectors for the textual and visual components. After the extraction process, we applied pooling techniques to these vectors, followed by concatenation to obtain combined feature vectors. These combined vectors were then fed into a new classifier for training. The trained model, as a result, is equipped to predict whether a given tweet, composed of both text and image, supports or opposes the specified topic.

Data Augmentation: We found the size of ImageArg training set is limited. There are only 888 for abortion and 918 entries for gun control. So we apply two data augmentation methods to increase the sample size. For synonym substitution, we use Python package nlpaug to randomly replace one word in the sentence with its synonym. For back translation, we use Python package googletrans to translate the text into Chinese and then back to English. The generated text is paired with its original image.

4 Experiments

4.1 Baselines and metrics

Baselines: For pure images classification, we use VGG16(Simonyan and Zisserman, 2015). For pure text classification, we use BERT(Devlin et al., 2019) and Naïve Bayes based on FastText(Joulin

Model	Abortion		Gun Control	
	F1-score	Accuracy	F1-score	Accuracy
VGG16	0.71	0.69	0.61	0.63
BERT	0.71	**0.73**	**0.73**	**0.73**
Naïve Bayes	**0.74**	0.71	0.68	0.69
Decision-level Fusion Method (VGG16+BERT+Naïve Bayes)	0.78	0.78	0.73	0.73
Feature-level Fusion Method (VGG16+BERT)	**0.85**	**0.84**	**0.80**	**0.81**
Feature-level Fusion Method (VGG16+BERT, Synonym Substitution)	0.83	0.82	0.79	0.80
Feature-level Fusion Method (VGG16+BERT, Back Translation)	0.80	0.78	0.78	0.78

Table 1: Performance of baselines and fusion methods on the test set on both topics.

et al., 2016). For the decision level fusion method, we use VGG16 as text classifier while Naive Bayes and BERT as text classifiers. For the feature level fusion, we use BERT as text encoder and VGG16 as image encoder.

Metrics: To compare the performance of baselines and fusion methods, we use F1-score and accuracy as the evaluation metrics.

4.2 Data preprocessing

For images, we remove those corrupted samples, resize the input to 224x224 pixels, convert the image to RGB format and substract the mean RGB value of the training set from each pixel of the image. For text, we correct the spelling and remove HTML tags, URLs, emails and garbled characters in the text. In the decision-level fusion method, the missing values are filled by 0.5 when combining the predictions by VGG16, BERT and Naïve Bayes.

4.3 Implementation Details

For each model containing BERT, tweet text are tokenized with a maximum length of 60. The experimental setup vary slightly depending on the models. SGD with a learning rate of 1e-5 is used as the optimizer in the VGG16. Adam is used in the decision-level fusion method. AdamW with a learning rate of 5e-4, 1e-4 and 2e-6 are used in BERT and feature-level fusion method. We also apply two data augmentation methods to the feature-level fusion method.

4.4 Experimental results

Table 1 shows the results of our experiments on the test set on both topics. On the whole, the prediction of abortion are more precise than the prediction of gun control.

For the baselines, predicting with pure text will perform better on these two topics than predicting with pure images, which indicates that text contains more related information. At the same time, as a ubiquitous baseline in NLP, BERT does better than the traditional model, Naïve Bayes, especially on the topic of gun control.

Both of our methods outperform the baselines on two evaluation metrics. The decision-level fusion method has an improvement of 0.04 in F1-score and 0.05 in accuracy on abortion, but gets nowhere on gun control. The feature-level fusion method has the best performance, and improves 0.11 in F1-score and accuracy on abortion and 0.07 in F1-score and 0.08 in accuracy on gun control. However, after applying synonym substitution and back translation to the feature-level fusion method, they perform worse than using the original dataset. We think this is due to the imbalance of the training dataset, which usually makes the model predictions tend to the majority label and the model performance will be unstable.

5 Conclusion

We implemented two multimodal fusion methods for multimodal argumentative stance classification on the ImageArg dataset. Both methods outperformed the baselines. We found the feature level fusion method is more suitable for the task, which has a more f1-score. We also test two data augmentation methods: synonym substitution and back translation. Although both data augmentation methods don't work well in our model, they may still be

good methods if we could address data imbalance of the training set.

Limitaions

The labels are unbalanced in the given training set, especially in the abortion dataset, but we didn't deal with the imbalance of the training set when conducting data augmentation.

Our policy for synonym substitution is randomly choosing one word to replace, which could not handle the context information and may suffer from the problem of polysemy. We may use some masked language model to get the right synonym set.

Also, we only conduct data augmentation for text in our experiment. We should also try some data augmentation methods for images to increase image diversity.

We could use more state-of-the-art pre-trained models instead of BERT and VGGNet16, like DeBERTa(He et al., 2021) for text and ViT(Dosovitskiy et al., 2021) for images.

In our feature level fusion method, we only use simple concatenation for feature fusion. We think there are other more advanced feature fusion methods such as multimodal attention.

Since original CNN models focus more on the objects but not the characters in the images, we could take the characters in those images into consideration.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Siddhant Garg and Goutham Ramakrishnan. 2020. Bae: Bert-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling bert for natural language understanding.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification.

Hong Liu, Menglei Jiao, Yuan Yuan, Hanqiang Ouyang, Jianfang Liu, Yuan Li, Chunjie Wang, Ning Lang, Yueliang Qian, Liang Jiang, Huishu Yuan, and Xiangdong Wang. 2022a. Benign and malignant diagnosis of spinal tumors based on deep learning and weighted fusion framework on mri. *Insights into Imaging*, 13.

Zhexiong Liu, Meiqi Guo, Yue Dai, and Diane Litman. 2022b. ImageArg: A multi-modal tweet dataset for image persuasiveness mining. In *Proceedings of the 9th Workshop on Argument Mining*, pages 1–18, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Ferda Ofli, Firoj Alam, and Muhammad Imran. 2020. Analysis of social media data using multimodal deep learning for disaster response.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Sudhakar Sengan, Subramaniyaswamy Vairavasundaram, Logesh Ravi, Ahmad Qasim Mohammad Al-Hamad, Hamzah Ali Alkhazaleh, and Meshal Alharbi. 2023. Fake news detection using stance extracted multimodal fusion-based hybrid neural network. *IEEE Transactions on Computational Social Systems*, pages 1–12.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Learning factorized multimodal representations.

Marcelo Worsley. 2023. *Artificial Intelligence Innovations for Multimodal Learning, Interfaces, and Analytics*, pages 19–35. Springer International Publishing, Cham.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. 2020. Unsupervised data augmentation for consistency training.

Ziang Xie, Sida I. Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, and Andrew Y. Ng. 2017. Data noising as smoothing in neural network language models.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2016. Character-level convolutional networks for text classification.