LT3 at SemEval-2021 Task 6: Using Multi-Modal Compact Bilinear Pooling to Combine Visual and Textual Understanding in Memes

Pranaydeep Singh and Els Lefever
LT3, Language and Translation Technology Team
Department of Translation, Interpreting and Communication – Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
pranaydeep.singh@ugent.be, els.lefever@ugent.be

Abstract

Internet memes have become ubiquitous in social media networks today. Due to their popularity, they are also a widely used mode of expression to spread disinformation online. As memes consist of a mixture of text and image, they require a multi-modal approach for automatic analysis. In this paper, we describe our contribution to the SemEval-2021 Detection of Persuasian Techniques in Texts and Images Task. We propose a Multi-Modal learning system, which incorporates "memebeddings", viz. joint text and vision features by combining them with compact bilinear pooling, to automatically identify rhetorical and psychological disinformation techniques. The experimental results show that the proposed system constantly outperforms the competition's baseline, and achieves the 2nd best Macro F1-score and 14th best Micro F1-score out of all participants.

1 Introduction

Propaganda is a mode of communication by which the interested party pursues the aim of influencing public opinion in favour of a specific agenda or ideas. This is achieved by disseminating one-sided, biased or even fake news. With the advent of social media networks, propagandist text can reach an enormous audience. Given the overload of online text produced on a daily basis, it is not feasible to monitor this manually and researchers have started to investigate automatic methods to detect propaganda in text.

In the SemEval-2020 Task 11 on Detection of Propaganda Techniques in News Articles (Da San Martino et al., 2020), participants were asked to identify 14 different propagandist techniques in news articles. The task attracted a large interest, with 44 teams participating in the task. The best systems all used pre-trained transformers and ensemble techniques. We further refer to Da San Martino et al. (2020b) for a comprehensive review of computational propaganda detection techniques.

More recently, internet memes have emerged as a very popular mode of expression on social networks. While memes initially seemed to be used by users of specific online communities, they have gained popularity very rapidly and are today used by a very large and varied user base. Memes can be used for very different purposes: they can be used as a form of visual rhetoric (Huntington, 2013), for online bullying or trolling (Leaver, 2013), but they can also function as a kind of persuasive device, while the intended message is wrapped in humour (Shifman, 2013). As a result, they form an interesting object of study for automatically detecting propaganda techniques.

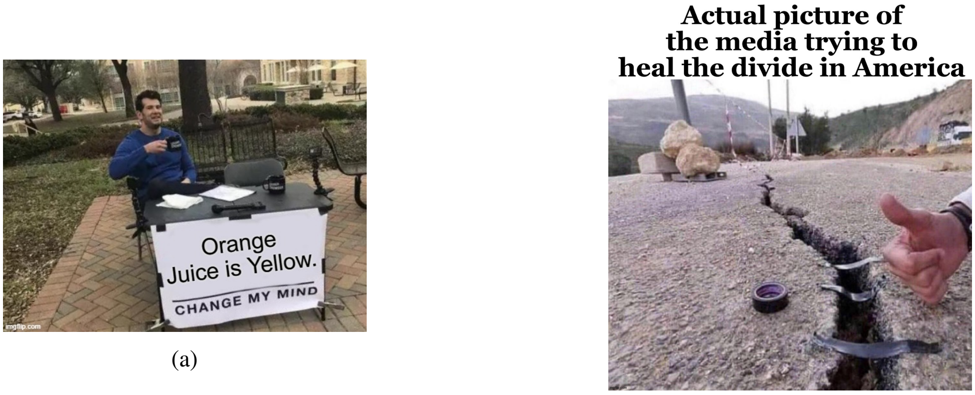
The goal of the SemEval-2021 shared task on the detection of persuasion techniques (Dimitrov et al., 2021) is to build models for identifying rhetorical and psychological techniques that are used to influence social media users in online disinformation campaigns. This paper reports on our participation in Subtask 3, which is a multi-modal task conceived as a multi-label classification problem: given a meme, the system has to identify which of the 22 techniques are used both in the textual and visual content of the meme. To solve subtask 3, we propose a multi-modal multi-task learning system, which incorporates "memebeddings", viz. joint text and vision features combined by means of compact bilinear pooling, to automatically identify rhetorical and psychological disinformation techniques in memes.

2 System Architecture

2.1 Task Overview

The SemEval 2021 Task 6 for detection of persuasion techniques in texts and images revolved

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 1051–1055
Bangkok, Thailand (online), August 5–6, 2021. ©2021 Association for Computational Linguistics



(a)

(b)

Figure 1: Examples of memes where the analysis of only the text (*a*) or both the text and image (*b*) are required to automatically detect the correct persuasion technique.

around identifying 22 rhetorical and psychological disinformation techniques in internet memes. These techniques cover a wide array of phenomenons like *Causal Oversimplification*, *Exaggeration/Minimisation*, *Name Calling/labeling* or *Presenting irrelevant data (red herring)*. For a full list of all categories, we refer to the task description paper (Dimitrov et al., 2021).

While some techniques in certain contexts may be accurately found just by processing the textual modality, it is very difficult to consistently identify all of the techniques without complete visual and textual context. Figure 1 shows examples of the two different cases, where in the first image, it is fairly obvious that the text contains all necessary information to predict the propaganda techniques accurately, whereas in the second meme, the textual modality is not sufficient to provide all information required to correctly predict the label. To tackle the task at hand, our approach incorporates information from both domain-related text and visual pre-training, and finally combines the two modalities using Multi-modal Compact Bilinear (MCB) Pooling (Fukui et al., 2016).

2.2 Proposed Models

Our multi-modal system is composed of three submodules:

1. **The visual pre-processor**: the visual module uses a ResNet-51 architecture (He et al., 2016) which is pre-trained to identify subreddits (E.g. /r/*motivation*, /r/*pets* /r/*politics*) from around 6200 Reddit memes.

2. **The text pre-processor**: the text module uses a pre-trained BERT-large-uncased model (Devlin et al., 2019), fine-tuned on the PTC Corpus (Martino et al., 2020a) from the SemEval 2020 Shared Task.

3. **Integration Network**: the two sets of embeddings from the first two modules are combined with MCB pooling.

2.2.1 Visual Embeddings

We decided to use the Resnet-51 architecture for the Visual pre-processor. This model was trained to predict one of the 18 sub-reddits the memes were scraped from. We hypothesized that the learned embeddings are able to distinguish certain elements of the meme, since the model is forced to encode the sub-reddit it comes from, and the sub-reddits represent the genre (E.g. /r/politics, /r/sports) or the emotion associated with the meme (E.g. /r/motivation, /r/dankmemes).

2.2.2 Textual Embeddings

For the text pre-processor we used a pre-trained bert-large-uncased model from the HuggingFace transformers package[1]. We fine-tuned the model with additional linear layers for the multi-label task of predicting propaganda techniques in the PTC Corpus. BERT based fine-tuned models are often used for a lot of text classification tasks and obtain state-of-the-art performances in a large number of NLP tasks like GLUE (Wang et al., 2018) and SQuAD (Rajpurkar et al., 2016).

2.2.3 Combined Embeddings

We train the final model with the combined embeddings from the visual and text pre-processor,

[1] https://huggingface.co/transformers/

for the final task of multi-label prediction of the 22 propaganda techniques. While the visual pre-processor and textual pre-processor become excellent feature generators individually, combining the embeddings from two modalities with different dimensions (768d for text and 1024d for images) becomes very complicated.

While a dot product will be simple and efficient to compute, it will only encode a linear mapping of features, i.e first order interactions where every visual feature only interacts with just one textual feature and not multiple features. In addition, a dot product cannot be computed with two vectors having different dimensions. A cross product on the other hand, computes the relation of every feature from one modality to every feature from the second modality. While this is closer to the representation we need, the cross product of two vectors, with 768 and 1024 dimensions respectively, will be 786,432 dimensional. To use this large a vector, the classification model would need billions of parameters, making it almost impossible to train such a model in practice.

MCB Pooling combines the computation efficiency of the dot product with the higher order representation of the cross product. It was first implemented for the task of Visual Question Answering (Antol et al., 2015), which also focuses on jointly encoding textual and visual content. It centers around constructing count sketch projections of the vectors by means of the Fast Fourier Transform (FFT) to reduce them to lower-dimensional vectors without loosing a lot of information. Once the vectors are projected to lower-dimensions, computing the cross-product becomes feasible again. Pham et al. (2013) demonstrated, though, that the count sketch of the outer product of two count sketch projections is the same as the convolution of the two count sketch projections, as shown in Equation 1:

$$\Psi(x \otimes q, h, s) = \Psi(x, h, s) * \Psi(q, h, s) \qquad (1)$$

where x and q are the embeddings from the textual and visual modality respectively, $\Psi(x, h, s)$ represents the count sketch projection of a vector x, and $*$ represents the convolution operator. Figure 2 summarizes the proposed system architecture.

2.3 Experimental Setup

The ResNet-51 model for the visual pre-processor was trained with Stochastic Gradient Descent and penalized with the cross-entropy loss. For the initial state we used the model pre-trained on ImageNet, and fine-tuned by replacing the classification layer.

The BERT model for the text embeddings was fine-tuned by freezing the pre-trained model and adding a linear layer as well as a classification layer. The model was trained with the AdamW optimizer, with a rate decay of 0.01, and penalized with cross-entropy as well.

The final MCB model uses embeddings from both models and combines them into a single vector of 8000 dimensions with MCB, then passes them through two linear layers of sizes 2048 and 1024 respectively, followed by a classification layer for the multi-label output for the 22 techniques. This final model is optimized with Adam and penalized with cross-entropy as well. We used the train set released by the task organizers for training, and the development set as a validation set for optimizing hyper-parameters.

3 Results

Table 1 summarizes the key results of our multi-modal approach. While combining the visual and textual embeddings with a simple weighted averaging consistently beats the task baselines by a significant margin, using MCB Pooling results in a considerable performance increase over weighted averaging, both for Macro and Micro F1-scores. In addition, when analyzing samples that are misclassified by the weighted averaging approach, but correctly predicted by MCB Pooling, we noticed around 40 percent of the examples required combining information from both the visual and textual modalities like Figure 1(b).

While the MCB model is able to better pool the understanding of both modalities, it still fails when a complex concept or inference is involved. Figure 3a represents a common meme format frequently found in the memes we were able to obtain from Reddit. Consequently, we expect the model has sufficiently learnt the corresponding visual features to come to a correct prediction. Figure 3b, however, requires some complex visual ideas the model has to infer in combination with the text, which it fails to do frequently. We believe that jointly training visual and textual embeddings (making the learning more coherent and not disjointed between the two modalities), instead of simply attempting to combine independent vi-

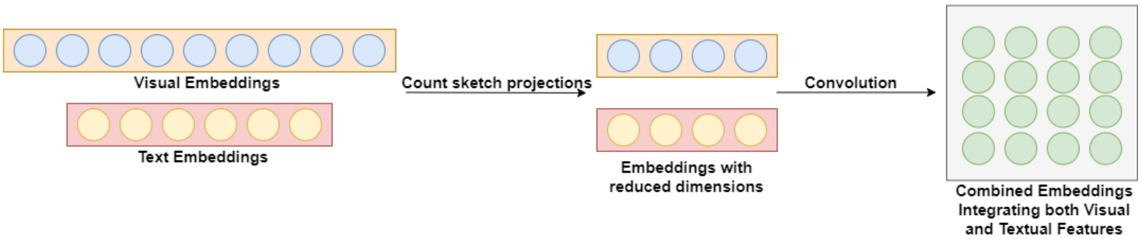
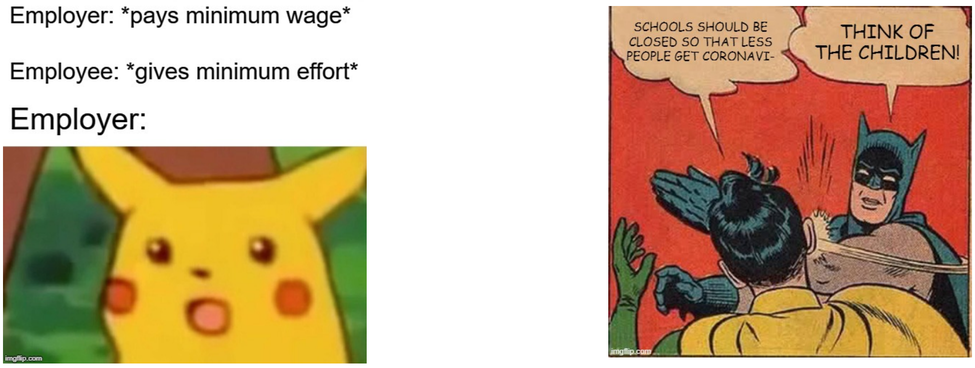

Visual Embeddings
Text Embeddings
Count sketch projections
Embeddings with reduced dimensions
Convolution
Combined Embeddings Integrating both Visual and Textual Features

Figure 2: Conceptual overview of the system architecture

	Macro-F1	Micro-F1
Baseline	0.051	0.070
Feature Combination with Weighted Averaging	0.112	0.248
Multi-modal Compact Bilinear Pooling (our system)	**0.263**	**0.331**

Table 1: Final Micro and Macro F1-scores for SemEval 2021 Task 6 sub-task 3


Employer: *pays minimum wage*
Employee: *gives minimum effort*
Employer:
imgflip.com
SCHOOLS SHOULD BE CLOSED SO THAT LESS PEOPLE GET CORONAVI-
THINK OF THE CHILDREN!
imgflip.com

(a)

(b)

Figure 3: Examples of failed and successful classification by the MCB Pooling model. The model succeeds for standard meme templates and basic visual ideas (example *a*) but fails to understand complex visual ideas that require a joint inference from the text (example *b*)

sual and textual information, would solve this issue. However, joint training can get computationally expensive and would require a much larger dataset.

4 Conclusion and Future Work

This paper presents the multi-modal approach we proposed for automatically detecting persuasion techniques in memes. As memes combine text and images to obtain the desired effect, we built a system where visual and textual embeddings are combined to classify 22 different propaganda techniques. The experimental results show that combining textual and visual embeddings by weighted averaging already beats the baseline. These results, however, are considerably improved by combining both embedding sets by means of MCB Pooling.

In future work, we will investigate how we can incorporate additional semantic information in our model. A first step could consist of integrating more explicit argumentation information into our model. As these propaganda techniques use psychological and rhetorical techniques, we believe it might be interesting to include argumentation structures such as logical fallacies, where the reasoning is flawed, and by consequence the conclusion cannot be drawn from the premise(s) in the text. To this end, we will build on recent work on automatic fallacy detection (Habernal et al., 2017). In addition, we also aim to include automatic emotion detection features, as writers of propagandist text often use emotional language to convince their readers. Finally, we will investigate an approach to jointly train visual and textual embeddings, rather then combining separate embedding sets, as our error analysis showed that efficient analysis of memes often requires a combined approach.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Giovanni Da San Martino, Alberto Barrón-Cedeño, 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, Barcelona (online). International Committee for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 457–468, Austin, Texas. Association for Computational Linguistics.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 7–12, Copenhagen, Denmark. Association for Computational Linguistics.

K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Heidi E Huntington. 2013. Subversive memes: Internet memes as a form of visual rhetoric. *AoIR Selected Papers of Internet Research*, 3.

Tama Leaver. 2013. Fcj-163 olympic trolls: Mainstream memes and digital discord? *The Fibreculture Journal*, (22 2013: Trolls and The Negative Space of the Internet).

G. Da San Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, and P. Nakov. 2020a. Semeval-2020 task 11: Detection of propaganda techniques in news articles.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrn-Cede±o, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020b. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4826–4832. International Joint Conferences on Artificial Intelligence Organization. Survey track.

Ninh Pham and Rasmus Pagh. 2013. Fast and scalable polynomial kernels via explicit feature maps. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 239–247.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Limor Shifman. 2013. Memes in a digital world: Reconciling with a conceptual troublemaker. *Journal of Computer-Mediated Communication*, 18(3):362–377.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.