DUTIR938 at SemEval-2024 Task 4: Semi-Supervised Learning and Model Ensemble for Persuasion Techniques Detection in Memes

Erchen Yu[1], Junlong Wang[2], Xuening Qiao[1], Jiewei Qi[1], Zhaoqing Li[1], Hongfei Lin[1], Linlin Zong[2], Bo Xu[1*]

[1] School of Computer Science and Technology, Dalian University of Technology, China

[2] School of Software, Dalian University of Technology, China

{yuerchen0809, jlwang, qiao, 1329027682, lizhaoqing}@mail.dlut.edu.cn

{hflin, llzong, xubo}@dlut.edu.cn

Abstract

The development of social platforms has facilitated the proliferation of disinformation, with memes becoming one of the most popular types of propaganda for disseminating disinformation on the internet. Effectively detecting the persuasion techniques hidden within memes is helpful in understanding user-generated content and further promoting the detection of disinformation on the internet. This paper demonstrates the approach proposed by Team DUTIR938 in Subtask 2b of SemEval-2024 Task 4. We propose a dual-channel model based on semi-supervised learning and model ensemble. We utilize CLIP to extract image features, and employ various pretrained language models under task-adaptive pretraining for text feature extraction. To enhance the detection and generalization capabilities of the model, we implement sample data augmentation using semi-supervised pseudo-labeling methods, introduce adversarial training strategies, and design a two-stage global model ensemble strategy. Our proposed method surpasses the provided baseline method, with Macro/Micro F1 values of 0.80910/0.83667 in the English leaderboard. Our submission ranks 3rd/19 in terms of Macro F1 and 1st/19 in terms of Micro F1.

1 Introduction

Social networks play a significant role in our society. The development of social platforms has facilitated the dissemination of information, but it has also fueled the proliferation of disinformation (Da San Martino et al., 2020; Dimitrov et al., 2021). The dissemination mechanism of disinformation involves the use of propaganda techniques. "Propaganda" is defined as a dissemination pattern referring to stakeholders influencing public opinion to support specific agendas and ideas by adopting persuasion techniques, such as disseminating one-sided, biased, or even fake news. Research on detecting propaganda techniques contributes to combating network disinformation (Da San Martino et al., 2021).

Among all types of content on social networks, memes play a significant role. Memes typically exist in the form of images, possibly with overlapping text, and convey information in the form of jokes, irony, etc. In the current era of social media, they spread rapidly and can influence many people without awareness. Memes are one of the most popular content types in online disinformation propaganda activities and serve as a powerful medium for promoting ideological and cognitive persuasion techniques (Moody-Ramirez and Church, 2019). Therefore, research on automatically detecting persuasion techniques hidden in memes is of significant importance, which contributes to understanding user-generated content and further aids in detecting network disinformation.

Subtask 2b of Semeval-2024 Task 4 aims to promote research on computational methods to detect persuasion techniques in memes (Dimitrov et al., 2024), which is modeled as a binary classification problem. Due to the complexity and subjectivity inherent in persuasive language, single multimodal model may struggle to capture all relevant features effectively. To address this issue, we propose a dual-channel model based on semi-supervised learning and model ensemble in this paper. Within the image channel, we use CLIP (Radford et al., 2021) to extract image features from memes. Concurrently, in the text channel, we employ diverse pretrained language models and conduct task-adaptive pretraining utilizing certain corpora provided by the task. We execute feature extraction from the pretrained language model, employing a variety of methodologies to capture sentence features, followed by a concatenation and fusion process of the extracted features, subsequently fed into the classification layer. We implement a semi-supervised pseudo-labeling ap-

*Corresponding Author

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 642–648

proach, wherein pseudo-labels are assigned to the test set, thereby augmenting the training data to achieve data augmentation. Furthermore, to bolster each model's robustness, We employ Fast Gradient Method(FGM) as our adversarial training strategy, introducing perturbations to the embedding layer of the model. Lastly, We design a two-stage soft-voting ensemble strategy to amalgamate the predictions of multiple models, augmenting the model's generalization capacity and performance.

We applied our proposed method to the English dataset. Our proposed method ranked 3rd/19 in terms of Macro F1 and 1st/19 in terms of Micro F1 in the English leaderboard for Subtask 2b.

2 Related Work

Transformer (Vaswani et al., 2017) is a deep learning model based on the self-attention mechanism, which is known for its ability to effectively capture long-range dependencies in sequential data. Based on Transformer, pretrained language models such as BERT (Kenton and Toutanova, 2019) have been proposed, which capture the contextual information of each token in the text through self-attention. Subsequent pretrained language models have mainly been modified on pretraining tasks, such as RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2020), and so on. These text models improve upon BERT by enhancing semantic feature representation in text feature extraction.

Existing image feature extraction methods mostly rely on multiple convolutional neural networks (Li et al., 2021), such as VGG (Simonyan and Zisserman, 2015) and ResNet (He et al., 2016). Vision Transformer (ViT) (Dosovitskiy et al., 2020) is an image processing model based on the traditional Transformer, dividing the input image into image patches and using multi-head self-attention mechanisms to capture global relationships between images. ViT represents a significant advancement of Transformer in the field of computer vision, bringing new ideas and methods to image processing tasks. CLIP (Radford et al., 2021) is a multimodal model which is trained through contrastive learning on 400 million pairs of images and text. CLIP achieves high-performance cross-modal semantic feature extraction for images and text.

3 System Overview

In this section, we will introduce the overall structure of our proposed system. Our system is divided into two stages. (1) Training of our Text-Image Multi-modal Classification Model. During the training process, we introduce several training strategies, including task-adaptive pretraining, pseudo-labeling and adversarial training. (2) Model Ensemble. We use k-fold cross-validation for model training and design a two-stage soft-voting strategy to globally integrate the models obtained in the first stage.

3.1 Model Architecture

The architecture of our model is shown in Figure 1. Our proposed text-image multi-modal classification model can be divided into two modules: feature extraction module and cross-modal fusion module. In the feature extraction module, we employ a parallel architecture to perform feature extraction separately for image and text channels using pretrained models.

For image inputs, we utilize the pretrained CLIP model. Before inputting images into the CLIP image encoder, they are resized and normalized, resulting in one-dimensional features of dimensionality 512. For text inputs, we experiment with various pretrained language models and their variants, including BERT, RoBERTa, DeBERTa, XLM (Conneau et al., 2020), DistilBERT (Sanh et al., 2019), etc. Based on their performance on the validation and dev sets, we ultimately select two general domain models and two models pretrained using political domain related corpora as text encoder models: CLIP text encoder, Deberta-v3-large (He et al., 2022), politicalBiasBERT (Baly et al., 2020), and xlm-twitter-politics-sentiment (Antypas et al., 2023). We select the latter two models because the vast majority of persuasion techniques are reflected in political domain.

The output of the CLIP text encoder is a one-dimensional vector of 512 dimensions, while for three BERT-like models above, the encoder's output is a two-dimensional vector that needs to be converted into a one-dimensional vector through pooling operation. Common pooling operations include cls, pooler, last layer average and first-last layer average.The optimal pooling method is selected for each BERT-like model as the model's pooling strategy. After the pooling layer, the text features are ultimately obtained as one-dimensional features of 512 dimensions.

The cross-modal fusion module aims to integrate features from two modalities. We concatenate the output image features with the text features and em-

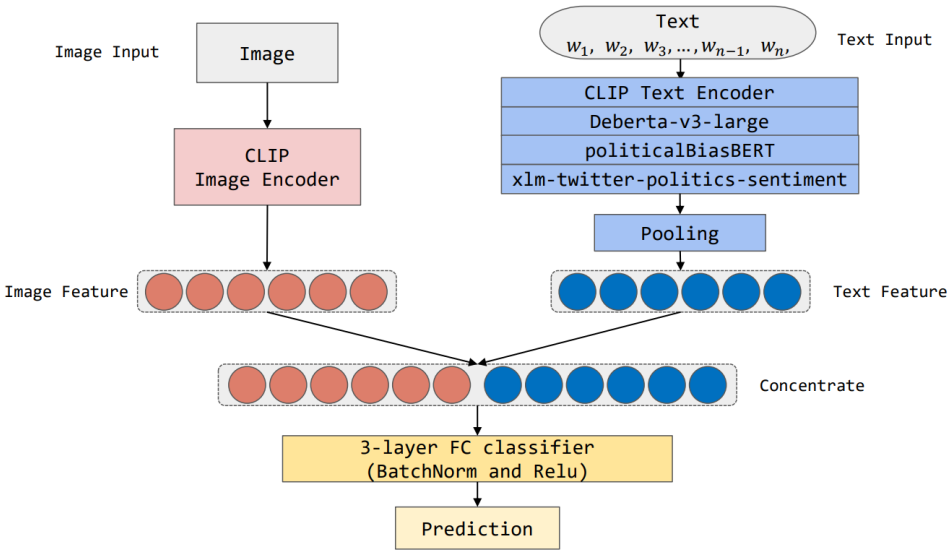

Figure 1: The architecture of our Text-Image Multi-modal Classification Model.

ploy a three-layer fully connected network as the classifier to map the final representation obtained from the fusion layer to a scalar, which is then bounded between 0 and 1 through a sigmoid function. The formula of cross-modal fusion module is defined as:

$$\hat{y}_c = \text{sigmoid}\left(W\left[h^I, h^T\right] + b\right)$$

Where h^I is image feature and h^T is text feature. Finally, the model would be fit with the binary cross-entropy loss function.

3.2 Training strategies

To further improve the performance of our model, we adopt three training strategies:

Task-adaptive Pretraining(TAPT): Certain research has proved that further pretraining models on the unlabeled task data itself or task related data, called TAPT, can improve the performance of model in downstream tasks (Gururangan et al., 2020). In this task, we adopt TAPT for three BERT-like text models. The purpose of this task is to detect persuasion techniques in memes. Therefore, we perform TAPT on text models using the provided dataset. Specifically, we collect and utilize all texts from task1, task2a, and task2b for pretraining our models with masked language model(MLM). By conducting MLM pretraining, these models can not only better fit the distribution in the task, but also learn rich knowledge and semantic information, thereby performing better on the task.

Pseudo-labeling: Pseudo-labeling is a semi-supervised method aimed at predicting labels for unlabeled data using a model trained on labeled data and adding the labeled data to the training set to achieve data augmentation. Considering the diversity of samples in the test sets, we adopt a semi-supervised pseudo-labeling method for data augmentation. We train four models on the original training set and then ensemble the four trained models using soft-voting for inference on unlabeled samples, i.e., test set samples. Subsequently, we consider samples with output probabilities greater than or equal to 0.8. Finally, we obtain 393 pseudo-labeled test set samples, which are re-incorporated into the training set for training.

Adversarial Training: Adversarial training is a common method to improve the robustness of neural networks. By adding perturbations in the embedding layer, we can obtain more stable embedding representations and more universal models, improving the performance of models on unseen data. In this task, we introduce FGM (Miyato et al., 2016) to enhance the model's robustness. The adversarial perturbation δ on s is defined as:

$$\delta = \epsilon \cdot g/\|g\|_2 \quad where \quad g = \nabla_s L(s, y)$$

where ϵ is a constant that controls the degree of perturbation suppression. The idea of FGM is to increase the perturbation direction along the gradient, where increasing along the gradient means maximizing the loss.

3.3 Ensemble Learning

To integrate the learning abilities of each model and improve the generalization ability of the final sys-

Dataset	Negative	Positive
Train	400	800
Validation	50	100
Dev	100	200
Test(Pseudo-labeled)	101	292
All	651	1392

Table 1: The label distribution of the dataset in task2b. Negative means non-propagandistic and positive means propagandistic.

Setting	Value
Epochs	20
Max Sequence Length	128
Batch Size	16
Optimizer	Adam
Learning Rate	5e-5
Dropout	0.5
Weight Decay	0.001

Table 2: Hyper-parameter settings of the experiment.

tem, we design a two-stage soft-voting strategy to ensemble the four models saved from k-fold cross-validation. In this task, due to the small size of the dataset, we use k-fold cross-validation to train models, ensuring that all data participate in training and validation to effectively avoid over-fitting. In the first stage, we average output probability values generated by the four models in each fold of the k-fold cross-validation, resulting in the probability values for each fold. In the second stage, we aggregate the probability values from each fold by averaging, yielding the final output. We determine the optimal binary classification threshold by testing on the validation set for various thresholds.

4 Experimental setup

4.1 Data description and Evaluation

The dataset is provided by Subtask 2b of Semeval-2024 Task 4. In the original dataset partition, there are 1200 samples in the training set, 150 samples in the validation set, 300 samples in the dev set, and 600 samples in the test set. After pseudo-labeling, 101 samples in the test set are labeled as non-propagandistic and 292 samples are labeled as propagandistic. We aggregate the training set, validation set, dev set, and the test set augmented with pseudo-labels into a new dataset for k-fold cross-validation. The label distribution of this dataset is shown in Table 1.

The official evaluation metrics for this task are Macro F1 and Micro F1, with a focus primarily on the performance of Macro F1 in our experiments. Both Macro F1 and Micro F1 performances are presented in the final test set ranking.

4.2 Implementation

During validation, we conduct 8-fold cross-validation on the dataset and consistently use the average measure from the first fold as the new validation set to evaluate the performance of our model. We preserve the model parameters to achieve optimal performance. During the testing phase, we train each model separately and predict the test set through the two-stage soft-voting strategy we purposed, resulting in the final prediction by averaging the probabilities from all trained models in 8 folds. After comparing the overall performance under different thresholds, we set the final threshold as 0.5, which yields the optimal average performance on the validation set under 8-fold cross-validation.

We implement our model using the transformer package[1]. We select the following four models as text encoder models: CLIP text encoder, Deberta-v3-large, xlm-twitter-politics-sentiment and politicalBiasBERT. And we select CLIP ViT-L/14@336px as the image encoder model. We sequentially combine each text encoder with the image encoder, and four final models are sequentially as follows: CLIP_{img}+$\text{CLIP}_{\text{text}}$, CLIP_{img}+$\text{DeBERTa}_{\text{text}}$, CLIP_{img}+XLM_{text} and CLIP_{img}+$\text{BERT}_{\text{text}}$. As for the pooling method, we test various options and selected the optimal one for each model based on performance on the validation set. Specifically, CLIP_{img}+$\text{DeBERTa}_{\text{text}}$ and CLIP_{img}+$\text{BERT}_{\text{text}}$ performed optimally with first-last layer average pooling, while CLIP_{img}+XLM_{text} performs optimally with pooler. We employ early stopping to retain the model parameters that exhibited the best performance on the validation set. Details of the hyper-parameter settings are provided in Table 2. We used a weighted binary cross-entropy loss using the class distribution. By default, We set ϵ to 1.0 in FGM. All experiments are conducted on an RTX 4090 with 24GB of memory.

5 Results

The overview statistics of four different models on the new validation set are shown in Table 3. Among all base models, CLIP_{img}+$\text{CLIP}_{\text{text}}$

[1]https://huggingface.co/

Setting	$CLIP_{img}$+$CLIP_{text}$	$CLIP_{img}$+$DeBERTa_{text}$	$CLIP_{img}$+XLM_{text}	$CLIP_{img}$+$BERT_{text}$
Base	0.8524	0.8375	0.8505	0.8385
+FGM	**0.8534**	0.8555	0.8435	0.8455
+TAPT	/	**0.8574**	0.8465	0.8505
+FGM+TAPT	/	0.8515	**0.8515**	**0.8586**

Table 3: Macro F1 for four models on the new validation set. "Base" indicates no training strategy added, "+FGM" indicates the addition of FGM, "+TAPT" indicates the addition of TAPT.

performs best, and $CLIP_{img}$+XLM_{text} likewise exhibits impressive performance. Additionally, $CLIP_{img}$+$BERT_{text}$+FGM+TAPT attains the highest Macro F1 score of 0.8586. The politicalBiasBERT model with FGM and TAPT strategies showcased its robust reasoning capability in detecting persuasion techniques in memes.

Experimental results highlight the effectiveness of TAPT and FGM strategies. Introducing FGM to the $CLIP_{img}$+$CLIP_{text}$, $CLIP_{img}$+$DeBERTa_{text}$ and $CLIP_{img}$+$BERT_{text}$ lead to significant performance enhancements. Both the $CLIP_{img}$+$DeBERTa_{text}$ and $CLIP_{img}$+$BERT_{text}$ models demonstrate performance improvements after introducing TAPT, and $CLIP_{img}$+$DeBERTa_{text}$with TAPT achieves the second-best overall performance. When both TAPT and FGM are employed simultaneously, all three models with BERT-like text encoder demonstrate substantial performance improvements compared with base model, while $CLIP_{img}$+XLM_{text} and $CLIP_{img}$+$BERT_{text}$ achieve their respective optimal performance levels, showcasing the effectiveness of adding FGM and TAPT to base model.

Regarding model ensemble, our analysis, illustrated in Table 4, demonstrates a significant enhancement in performance with the adoption of our two-stage global integration approach. Moreover, the scalability and robustness of our integration method are validated by the observed performance scalability relative to the number of integrated models. Results shows that all model ensemble performed best, which demonstrates the effectiveness of our model ensemble strategy.

We employed all-model ensemble as the final model of our system and submitted the result file of the test set predicted by the final model. The official rankings are shown in Table 5. We ranked 3rd in terms of Macro F1 and 1st in terms of Micro F1. Results show that our system demonstrates outstanding performance in detecting persuasion techniques in memes, and the integration of TAPT, FGM and model ensemble techniques further enhances the detection capability of our system.

Method	Macro F1	Micro F1
$CLIP_{img}$+XLM_{text}	0.8515	0.8711
Two-Model Ensemble	0.8515	0.8711
Three-Model Ensemble	0.8596	0.8789
All-Model Ensemble	**0.8654**	**0.8789**

Table 4: Results for model ensemble on the new validation set. Two-model ensemble refers to integrating $CLIP_{img}$+XLM_{text} and $CLIP_{img}$+$BERT_{text}$. Three-model ensemble adds $CLIP_{img}$+$DeBERTa_{text}$. For all-model ensemble, all four models are integrated for ensemble.

Rank	Team	Macro F1	Micro F1
1	LMEME	0.81030	0.82500
2	SuteAlbastre	0.80964	0.83500
3	**DUTIR938**	**0.80910**	**0.83667**
4	BCAmirs	0.80337	0.82500
5	Snarci	0.79860	0.82667

Table 5: Results of top 5 teams for subtask2b English leaderboard on the test set.

6 Conclusion

The paper presents our system designed for Subtask 2b of Semeval-2024 Task 4. We propose a dual-channel model based on semi-supervised learning and model ensemble. Our framework leverages multiple pretrained models served as feature extractors for images and texts. We integrate a semi-supervised pseudo-labeling approach for data augmentation, and introduce TAPT and FGM adversarial training to significantly enhance the model's performance and robustness. Finally, to enhance the generalization capability of our system, we design a two-stage soft-voting model ensemble strategy. Our system achieves excellent performance in detecting persuasion techniques in memes, and we ranked 3rd in terms of Macro F1 and 1st in terms of Micro F1 in the test set for Subtask 2b. Our future research will be directed towards exploring the cross-modal fusion mechanisms within the model.

7 Acknowledgments

We thank our anonymous reviewers for their helpful comments. This work was supported by grant from the Natural Science Foundation of China(N0.62006034), the Ministry of Education Humanities and Social Science Project (No.22YJC740110), the Fundamental Research Funds for the Central Universities (No.DUT23YG136).

References

Dimosthenis Antypas, Alun Preece, and Jose Camacho-Collados. 2023. Negativity spreads faster: A large-scale multilingual twitter analysis on the role of sentiment in political communication. *Online Social Networks and Media*, 33:100242.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard 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.

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.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2021. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4826–4832.

Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov, and Giovanni Da San Martino. 2024. Semeval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, SemEval 2024, Mexico City, Mexico.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Semeval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2022. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. 2021. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*.

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*.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

Mia Moody-Ramirez and Andrew B Church. 2019. Analysis of facebook meme groups used during the 2016 us presidential election. *Social Media+ Society*, 5(1):2056305118808799.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark,

et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

K Simonyan and A Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society.

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.