Image-Text Re-Matching with Zero-shot and Finetuning of CLIP

Yuta Fukatsu[1,*], Masaki Aono[1]

[1] *Toyohashi University of Technology, Japan*

Abstract

Images play an important role in the perception of online news. We aim to gain more insight into the interplay of images and texts in different news domains. In this paper, we describe our method toward Image Text Re-Matching based on the CLIP model with zero-shots and finetuning. Specifically, we introduce WISE-FT, a method for linear interpolation of weights between the zero-shot and finetuning models, to improve recall for re-matching. The WISE-FT has been reported to be effective to boost accuracy in CLIP on classification experiments. We obtained certain MRR and Recall with these methods.

1 INTRODUCTION

Online news articles in recent years have been a mixture of text and image-based articles. Images are often added to text articles to attract attention and help readers understand the article intuitively. Typically, studies of multimedia and recommendation systems assume a simple relationship between images and text. For example, in the study of image captions [1], it is assumed that the caption is a textual representation of the landscape of the image. However, news-specific studies point to a more complex relationship [2]. The NewsImages task [3] in MediaEval 2022 investigates this relationship to understand its implications for journalism and news personalization.

In MediaEval 2021, Thien-Tri et al. [4] used CLIP (Contrastive Language-Image Pre-Training) [5]. Thien-Tri et al. did not train on the dataset, but we train on the dataset published by the organizer, resulting in improvement in Recall and MRR was observed. However, news data has a different relationship to image and text dataset such as MSCOCO [6]. Therefore, CLIP's inherent ability may be lost due to shifts in the data distribution caused by finetuning. As a solution we apply WISE-FT [7], which has been reported to be robust to shifts in data distribution in class classification, in our retrieval task.

The paper is organized as follows: in Sec. 2, related studies; in Sec. 3, our approach; in Sec. 4, experimental results and their analysis are presented, and trends are discussed through visualization. Finally, Sec. 5 discusses the conclusions and challenges.

2 RELATED WORK

2.1 CLIP

CLIP (Contrastive Language-Image Pre-Training) [5] a neural network trained by a large dataset of image-text pairs. CLIP can predict the most relevant text given an image without directly optimizing for the dataset in a particular task. This process of making predictions for a different task other than the pre-training task without optimization is called zero-shot. CLIP is very powerful in this zero-shot, comparable to the zero-shot performance of the original ResNet50 in ImageNet. The feature representations obtained by CLIP through pre-training can be used to vectorize data in the retrieval task.

2.2 WISE-FT on classification

CLIP shows consistent accuracy in zero-shot inference across a variety of datasets. Finetuning can also improve accuracy on specific datasets. However, shifts in the data distribution in finetuning can reduce robustness. Wortsman et al [7] introduced WISE-FT, an ensemble of weights for zero-shot and finetuning models, for this problem and found that it improves accuracy in classification problems. The ensemble of weights in WISE-FT is achieved by linear interpolation of weights. For a hyperparameter α, the model weights are determined by the following equation (1), where W_x refers to the model weights.

$$W_{WISE-FT} = (1-\alpha) \times W_{zero-shot} + \alpha \times W_{finetuning} \quad (1)$$

MediaEval'22: Multimedia Evaluation Workshop, January 12–13, 2023, Bergen, Norway and Online
**Corresponding author.*
m.a.larson@tudelft.nl (M. Larson); gareth.jones@computing.dcu.ie (G. Jones); bionescu@imag.pub.ro (B. Ionescu)

3 APPROACH

3.1 Training CLIP

Three datasets are provided by the organizer in this task [3]. Thus, we train CLIP separately on each dataset. Finetuning of CLIP was done using the experimental method of Sedigheh et al [8]. The model of CLIP consists of a Vision Encoder and a Text Encoder. The Vision Encoder can be implemented by CNN based models such as ResNet or by Transformer-based models such as Vision Transformer. We used Vision Transformer Based model with a patch size of 32 as the Vision Encoder. The model structure of CLIP and the hyperparameters in the training are the same for each dataset. In addition, only the Online News portal dataset is in German. Since CLIP is pre-trained on the English dataset, the dataset must be translated in order to benefit from it. We used the method of Jörg et al [9] for German-English translation.

3.2 Applying WISE-FT

News dataset has two characteristics. First, they may differ in content from domain to domain. Second, they contain depending on when the news is released. For the first characteristic, finetuning adapted to the data in that domain is expected to improve Recall. However, for the second characteristic, there is a possibility that finetuning may degrade performance due to differences in data distribution caused by the timing of news releases. Therefore, we use WISE-FT, which has been reported to improve accuracy in classification problems, for the retrieval problem. WISE-FT is a linear interpolation of weights as shown in Equation 1. Since training is performed here for each dataset, the WISE-FT is also performed for each dataset. The hyperparameter α of the linear interpolation is also different for each dataset.

3.3 Splitting the Dataset and Submitted Runs

In this task, three datasets are provided by the organizer for this task: Online News portals, Twitter, and RSS news feed. We sort each dataset in chronological order, using 80% from the top as training data and the remaining 20% as validation data. The predictions for the test data were conducted by extracting features from all the test data, followed by using the features to compute cosine similarity to obtain the top 100 candidates.

In Run1, we use zero-shot prediction by CLIP pretrained on WebImageText. In Run2, we finetuned CLIP on each dataset and predict. In Run3, we apply WISE-FT to the CLIP trained on each dataset. The parameter alpha is the value with the highest Recall@1 in each validation data. Specifically, 0.5 is used for Online news portals, 0.5 for Twitter, and 0.4 for RSS news feeds.

4 RESULTS AND ANALYSIS

4.1 Submission Result

The results of the submitted runs are summarized in Table 1 for Online News portals, Table2 for Twitter, and Table3 for RSS news feed. The left column shows the name of the Runs. The evaluation metrics shown are MRR@100, Recall@5, Recall@10, Recall@50, and Recall@100. Each column in Table 1 shows the results of zero-shot, finetuning, and WISE-FT of CLIP.

Table 1: **Submission result for Online News portals in the three types of CLIP**

	MRR@100	Recall@5	Recall@10	Recall@50	Recall@100
zero-shot	0.195	0.261	0.347	0.562	0.647
fine-tuning	0.231	**0.329**	0.416	**0.655**	**0.745**
WISE-FT	**0.240**	**0.329**	**0.423**	0.649	0.735

Table 2: **Submission result for Twitter in the three types of CLIP**

	MRR@100	Recall@5	Recall@10	Recall@50	Recall@100
zero-shot	0.446	0.547	0.641	0.802	0.868
fine-tuning	0.472	**0.595**	0.679	**0.857**	**0.915**
WISE-FT	**0.476**	**0.595**	**0.681**	0.848	0.910

Table 3: **Submission result for RSS news feed in the three types of CLIP**

	MRR@100	Recall@5	Recall@10	Recall@50	Recall@100
zero-shot	0.403	0.509	0.603	0.777	0.833
fine-tuning	0.408	0.531	0.621	0.772	0.839
WISE-FT	**0.455**	**0.569**	**0.666**	**0.809**	**0.859**

4.2 Analysis of Each Dataset

4.2.1 Ranking Changes by WISE-FT

The analysis in this chapter uses our own split data for validation. Table 4 shows the average improvement and decrease in ranking from finetuning to WISE-FT. The tabulations here assume of an improvement from zero-shot to finetuning. It should be noted here that the size of each validation dataset is different for each dataset. Recall values improved the most for RSS news feeds, but the average increase in ranking was the lowest.

Table 4: **Average improvement and decrease in ranking when found in the top 100 on validation set**

	improvement	decrease
Online News portals	4.905	6.091
Twitter	3.893	5.158
RSS news feed	1.667	6.778

4.2.2 Visualization of Online News portals

The left side of Figure 1 shows a case that was worsened by finetuning but improved by WISE-FT, and the right side shows a case that was improved by finetuning but worsened by WISE-FT. Here, the original text is too long to display, so only the first sentence is displayed. In cases that were exacerbated by WISE-FT, news about the person was found. In the cases that improved with WISE-FT, news was seen where the text described the scene of the image. It is possible that WISE-FT brought general information into focus.

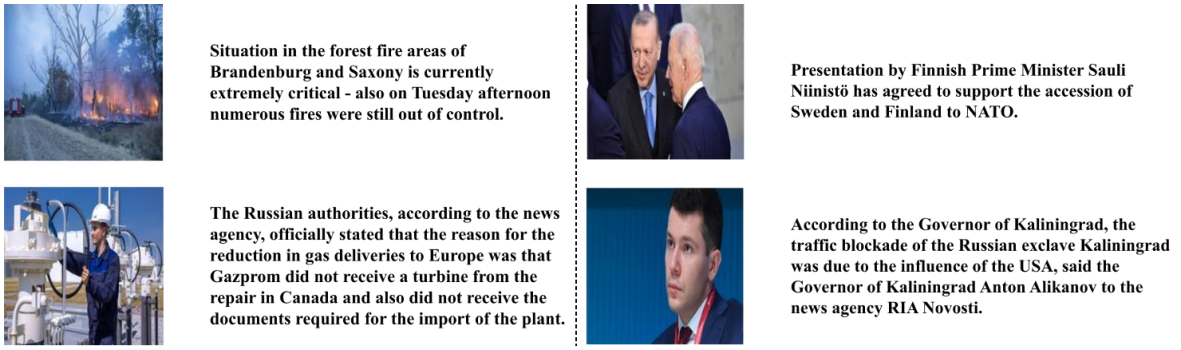

Figure 1: **Image-text pairs for cases that worsened by finetuning but improved by WISE-FT (left) and improved by finetuning but worsened by WISE-FT (right) on Online News portals.**

4.2.2 Visualization of Twitter

The left side of Figure 2 shows a case that was worsened by finetuning but improved by WISE-FT, and the right side shows a case that was improved by finetuning but worsened by WISE-FT. In both cases in the Twitter dataset, words tied to images were found in the text. For example, 'robot', 'missile', and 'van'. However, since the dataset also includes images that are thumbnails of videos, it is possible that the images are not correctly connected to the text.

Figure 2: **Image-text pairs for cases that worsened by finetuning but improved by WISE-FT (left) and improved by finetuning but worsened by WISE-FT (right) on Twitter.**

4.2.3 Visualization of RSS news feed

The left side of Figure 3 shows a case that was worsened by finetuning but improved by WISE-FT, and the right side shows a case that was improved by finetuning but worsened by WISE-FT. Here, the original text is too long to display, so only the first sentence is displayed. Despite the higher accuracy compared to other datasets, the text is not a direct expression to the image in both cases.

Figure 3: **Image-text pairs for cases that worsened by finetuning but improved by WISE-FT (left) and improved by finetuning but worsened by WISE-FT (right) on RSS news feed.**

5 CONCLUSIONS

We adopted CLIP and implemented finetuning and WISE-FT. As a result, we achieved MRR@100 score of 0.240, Recall@100 score of 0.735 for Online News portals test set, MRR@100 score of 0.476 and Recall@100 score of 0.595 for Twitter test set, and MRR@100 score of 0.455 and Recall@100 score of 0.859 for RSS news feed test set. We confirmed that the zero-shot method can obtain a constant Recall. Furthermore, in all datasets, finetuning increased Recall more than zero-shot. This indicates that CLIP pre-trained on datasets with different domains has some effect on the news dataset, and that learning to adapt to the news dataset is an effective method. On the other hand, the improvement in Recall by WISE-FT was significant only for the RSS news feed dataset. This may be because the hyperparameter α in the linear interpolation of WISE-FT was determined by the validation data, which did not result in optimal model weights for the test data. Alternatively, the finetunig may have over-adapted to the content within a certain time period.

Future work is needed to understand why similar images were attached to each news article. For example, the two images on the left in Figure 3 are images of buildings, but in the news articles they have different information such as place names. For this reason, the application of methods other than deep learning, such as pre-associating images to unique expressions such as place names, may improve Recall.

REFERENCES

[1] MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A Comprehensive Survey of Deep Learning for Image Captioning. ACM Comput. Surv. 51, 6, Article 118 (Feb. 2019). https://doi.org/10.1145/3295748

[2] Nelleke Oostdijk, Hans van Halteren, Erkan Bas, ar, and Martha Larson. The Connection between the Text and Images of News Articles:New Insights for Multimedia Analysis. In Proceedings of The 12th Language Resources and Evaluation Conference. 4343–4351.

[3] Benjamin Kille, Andreas Lommatzsch, Özlem Özgöbek, Mehdi Elahi and Duc-Tien Dang-Nguyen. News Images in MediaEval 2022. Proc. of the MediaEval 2022 Workshop, Bergen, Norway and Online, 12-13 January 2023.

[4] Thien-Tri Cao, Nhat-Khang Ngo, Thanh-Danh Le , Tuan-Luc Huynh, ,Ngoc-Thien Nguyen, Hai-Dang Nguyen , Minh-Triet Tran. 2021. HCMUS at MediaEval 2021: Fine-tuning CLIP for Automatic News-Images Re-Matching. In Proceedings of the MediaEval 2021 Workshop, Online, 13-15 December 2021

[5] 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. CoRR abs/2103.00020 (2021). arXiv:2103.00020 https://arxiv.org/abs/2103.00020

[6] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick. 2014. Microsoft COCO: Common Objects in Context, in: European Conference on Computer Vision, Springer, 2014, pp. 740–755. doi:10.1007/978-3-319-10602-1_48.

[7] Wortsman, Mitchell and Ilharco, Gabriel and Kim, Jong Wook and Li, Mike and Kornblith, Simon and Roelofs, Rebecca and Gontijo-Lopes, Raphael and Hajishirzi, Hannaneh and Farhadi, Ali and Namkoong, Hongseok and Schmidt, Ludwig. 2021. Robust fine-tuning of zero-shot models. arXiv preprint arXiv:2109.01903. https://arxiv.org/abs/2109.01903

[8] Sedigheh Eslami, Gerard de Melo, Christoph Meinel, Does CLIP Benefit Visual Question Answering in the Medical Domain as Much as it Does in the General Domain? CoRR abs/2112.13906 (2021). arXiv:2112.13906, https://arxiv.org/abs/2112.13906

[9] Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – Building open translation services for the World. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.