Image-Text Re-Matching Using Swin Transformer and DistilBERT

Yuta Fukatsu
Department of Computer Science & Engineering
Toyohashi University of Technology
Toyohashi, Aichi, Japan
fukatsu.yuta.ye@tut.jp

Masaki Aono
Department of Computer Science & Engineering
Toyohashi University of Technology
Toyohashi, Aichi, Japan
aono@tut.jp

ABSTRACT

In recent years, the news media has become multimodal. The relationship between text and images in news is complex and needs to be understood. In this paper, we work on Image-Text Re-Matching to understand the relationship between images and text, and apply and improve the image retrieval method, ADAPT. Improvements are made by reconsidering the feature extraction methods in image retrieval. We employ Swin Transformer for image feature extraction and DistilBERT for text feature extraction. According to the report from organizers, our runs resulted in MRR@100 score of 0.0789 and Recall@100 score of 0.5781 for test set.

1 INTRODUCTUIN

Online news articles in recent years have mixed components, consisting of texts and images. It is often the case that images are added to text articles to attract attention and to help readers understand the articles intuitively. Usually, in research on multimedia and recommendation systems, a simple relationship between images and text is assumed. As an example, in the study of image captioning [1], the caption is assumed to be a literal representation of the image landscape. However, news-specific studies have pointed out a more complex relationship [2]. The NewsImages task of MediaEval 2021, investigates this relationship to understand its impact on journalism and news personalization. Our team (KDEval 2021) participated in subtask 1, Image-Text Re-Matching. In this task, links between a series of articles and images have been removed.

In MediaEval 2020 [3], metric learning was introduced. We thus adopt a metric learning based method is inspired by ADAPT [4]. Re-Matching is performed by a text-based image retrieval method. We reconsider and experiment with image feature extraction and text feature extraction in ADAPT [4] for NewsImages. After reconsidering the feature extraction method, we confirmed that the best results are obtained by using Swin Transformer for image feature extraction and DistilBERT for text feature extraction.

2 RELATED WORK

2.1 ADAPT

ADAPT, which one of the image-to-text (text-to-image) alignment model are used for cross-modal retrieval. ADAPT takes a text (image) as input and then searches for the closest image (text) and outputs it. In ADAPT, the features for the input modality are used to recalculate the features for the other modality.

2.2 DistilBERT

DistilBERT [5] is a distillation of the BERT (Bidirectional Encoder Representations from Transformers) model, which is a natural language model that can understand context backwards and forwards and has been pre-trained on a large scale. However, BERT has the disadvantage that the model is too large for its performance, so DistilBERT achieves lightweight and speedup by distilling the model.

2.3 Swin Transformer

Swin Transformer [7] is a type of Vision Transformer, an image recognition model that introduces the concept of Transformer, which has been successful in natural language. Vision Transformer can benefit from the Transformer by dividing images into patches and treating them like words in NLP. Swin Transformer is a model that solves the shortcoming of Vision Transformer, that is, the fixed size patches are insufficient for recognizing objects of various sizes.

3 APPROACH

As shown in next sections, we reconsider the feature extraction methods used in ADAPT for the specific case of news articles and explain our method.

3.1 Reconsidering of Text Feature Extraction

GloVe embedding and bi-directional GRU are used in ADAPT to extract text features considering contextual information. However, even with context-aware methods

MediaEval'21, 13-15 December 2021, Online

using bi-directional GRU, there is a limitation on maintaining context information with distant words. Especially in news articles, it is highly likely that the text tends to be long. Thus, we have newly adopted DistilBERT as our text feature extraction method, which can handle longer texts and can obtain better features due to its rich pre-training. DistilBERT is also lighter than plain BERT, which would be more practical for applications to real-time search and recommendation.

3.2 Reconsidering of Image Feature Extraction

The image feature extraction in ADAPT is based on a Faster R-CNN pre-trained on the Visual Genome dataset [6]. This method uses 36 objects as features with high confidence in the image. However, the images given in news articles are often abstract or imaginative of the article content. Therefore, we cannot obtain useful features in such cases, or we need to extract 36 objects using extremely low confidence thresholds. Thus, we decided to reconsider how to acquire useful features while retaining the advantages of ADAPT, which is more efficient than attention-based methods by using spatial-level features. To deal this problem, we adopted the Swin Transformer. By using Swin Transformer, it is possible to obtain spatial-level and meaningful features.

3.3 Training and Submitted Runs

In subtask 1, Batch1 to Batch3 over three periods are provided by organizers as training data and Batch4 is as test data. Thus, we used Batch1 and Batch2 as training data, Batch3 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 the Run1, we used DistilBERT pre-trained on German for text feature extraction and Faster R-CNN trained on Visual Genome dataset for image feature extraction. In Run2, we changed the image feature extraction to Swin Transformer pre-trained on ImageNet 21K [8]. In Run3 we changed the batch size from 105 to 32.

4 RESULT AND ANALYSIS

The results of the submitted runs are summarized in Table 1. 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. In the table, Recall@k is written as R@k for simplicity.

Table2 shows the comparison of evaluation metrics between the data against Batch3 treated as validation data and the test data. The results demonstrate that there is no significant difference in distribution between the data provided by organizers for training and the test data. This led us to perform several analyses on Batch3.

Figure 1 and 2 show the word frequency when found in the top 5 search results and the word frequency when not found in the top 100. The words displayed here are limited to the nouns (lemma of tag) in the text that were extracted using Tree Tagger [9]. Comparing the two figures, we can see that there is no significant difference in the words that frequently appear in the success and failure cases. Therefore, the performance of the method in this paper for articles with similar content is considered to be low.

Table 1 : Submission result

	MRR@100	R@5	R@10	R@50	R@100
Run 1	0.0466	0.0642	0.1081	0.3159	0.4637
Run 2	0.0738	0.0971	0.1629	0.4318	0.5749
Run 3	**0.0789**	**0.1044**	**0.1687**	**0.4371**	**0.5781**

Table 2 : Comparison between Batch3 as validation data and Batch4 as test data

	MRR@100	R@5	R@10	R@50	R@100
Batch3	0.0695	0.0956	0.1653	0.4014	0.5357
Batch4	0.0789	0.1044	0.1687	0.4371	0.5781

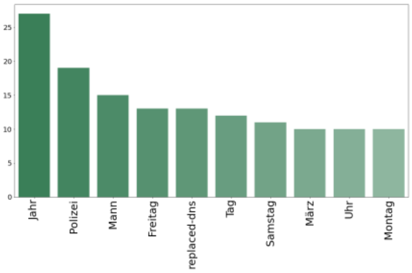

Figure 1 : Top 10 most frequent words found in the top 5

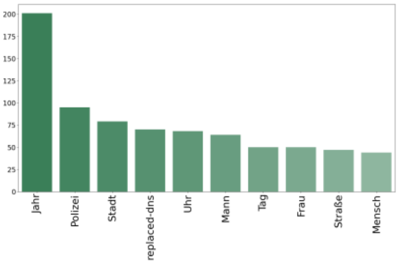

Figure 2 : Top 10 most frequent words not found in the top 100

5 CONCLUSION AND FUTUREWORKS

We changed the image feature extraction in ADAPT to Swin Transformer and the text feature extraction to DistilBERT. With this change, we achieved MRR@100 score of 0.07885 and Recall@100 score of 0.57807. This means that using our retrieval method, we can find relevance with some accuracy of 50% for matching images and text. Looking at the word frequency against successful and unsuccessful search results, the same words are frequently used, and we need to improve our search method for articles with similar contents.

REFERENCES

[1] MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. 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.2020. 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] Quang-Thuc Nguyen, Tuan-Duy Nguyen, Thang-Long Nguyen-Ho, Anh-Kiet Duong, Xuan-Nhat Hoang, Vinh-Thuyen Nguyen-Truong, Hai-Dang Nguyen, Minh-Triet Tran. 2020. HCMUS at MediaEval 2020:Image-Text Fusion for Automatic News-Images Re-Matching. *In Proceedings of the MediaEval 2020 Workshop, Online, 14-15 December 2020.*

[4] Wehrmann, J., Kolling, C. and C Barros, R. 2020. Adaptive Cross-Modal Embeddings for Image-Text Alignment. *Proceedings of the AAAI Conference on Artificial Intelligence*. 34, 07 (Apr. 2020), 12313-12320. DOI:https://doi.org/10.1609/aaai.v34i07.6915.

[5] Sanh, Victor, Lysandre Debut, Julien Chaumond and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv* abs/1910.01108 (2019)

[6] Krishna, R., Zhu, Y., Groth, O. et al. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. In *International Journal of Computer Vision* 123, 32–73 (2017). DOI:https://doi.org/10.1007/s11263-016-0981-7

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10012-10022.

[8] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, Lihi Zelnik-Manor. 2021. ImageNet-21K Pretraining for the Masses. *ArXiv, abs/2104.10972*.

[9] Helmut Schmid. 1995. Improvements In Part-of-Speech Tagging With an Application To German. In *Proceedings of the ACL SIGDAT-Workshop*, pp 47-50