Exploring a Pre-trained Model for Re-Matching News Texts and Images

Mingliang Liang[1], Martha Larson[1]
[1]Radboud University, Netherlands
mingliang.liang@ru.nl,martha.larson@ru.nl

ABSTRACT

We investigate the use of a pre-trained model to address the task of re-matching images and texts collected from online news sites. Our aim is to explore the potential of pre-training in learning the connection between the visual and textual modalities. Online news is challenging because it covers a large number of semantic concepts and also the correlation between the modalities can be weak. The results show that the proposed method has good performance in text-image retrieval. The performance was 46.58% (R@100).

1 INTRODUCTION

Pre-trained models are used for various vision and language tasks [5, 7, 10, 15]. They have proven effective in downstream tasks, such as image-text retrieval and Visual Question Answering (VQA). Pre-trained models have several advantages. First, they can increase performance on small datasets, where we need to improve the generalization ability of the model. An important example of a domain in which data is limited is news. Topics in the news develop rapidly and the amount of up-to-date data at any given moment is naturally limited. Second, the basic elements of images and text are not always associated in the same way across data sets. Investigating pre-trained models can also help us to better understand the connection between image and text modalities. Third, a pre-trained model can be quickly validated on a new dataset without the need to spend a lot of time and computing power on retraining.

In this paper, we use a pre-trained model to address image-text re-matching, a subtask of NewsImages at MediaEval 2021 [6]. We evaluate our approach on a test set containing 1915 articles that was collected from online news sites. Each article consists of an image and an associated text (title and snippet), cf. Figure 1. In the test data, the text has been disassociated from the images, and the task is to re-match them.

The task is challenging because there is not a 1-to-1 relationship between the concepts depicted in the images and those described in the text. As illustrated by Figure 1, only a few concepts may be common to both image and text. Also, in the collection a very broad range of topics and concepts are present. Our hope is that pre-training introduces prior knowledge that allows more effective learning of the relationship between text and images.

2 RELATED WORK

Cross-modal retrieval generally leverages a common space. Individual modalities express similar semantics differently, and the common space homogenizes the semantic representation. VSE++, DeViSE and CLIP learn a robust shared space, under which the learned-feature representation can be measured between modalities and preserves the correlations in paired samples as well [3, 4, 13].

ViT shows that by mapping sequences of image patches to embeddings, which replaces the word embeddings as input, the transformer can perform very well in visual tasks. [2]. Inspired by ViT, Vilt uses patch projection embedding to encode images [2]. Then, in vision and language interaction tasks, transformers are used instead of dot products for interaction between features [7]. This interaction method can not only increase the model's reasoning speed, but also can capture detailed relationships between vision and language. As mentioned above, we used Vilt to address the subtask of Image-Text Re-Matching at MediaEval 2021.

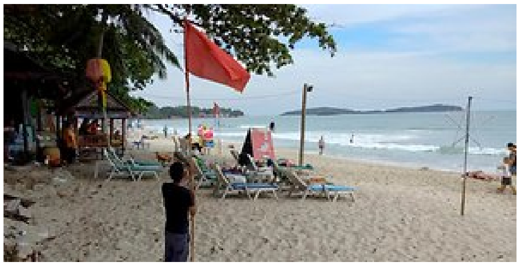

Figure 1: Example news item in the collection. Text associated with the image: "In the face of what is possibly the worst tropical storm in decades, tens of thousands of people in southern Thailand have left their homes and sought protection. The residents on the coast of the province of Nakorn Si Thammarat, in the storm 'Pabuk' on Thursday." This example illustrates a typical case in which few concepts are shared between text and image.

3 APPROACH

3.1 Dataset and data pre-processing

The dataset comprises 7530 training samples and 1915 test samples, released for the MediaEval 2021 Image-Text Re-Matching subtask [5]. We need to crawl the images by ourselves using the URLs provided by the task organizers and drop the *404 Not Found* URLs from the training and test set. For compatibility with the pre-trained model, we translate the text into English using Google Translate.

3.2 Model

In this section, we describe the architecture of the model that we used to address the Image-Text Re-matching subtask. The model

MediaEval'21, December 13-15 2021, Online

The refugee home has been on Langenbergstrasse in Blumenberg since the end of 2014. Originally - and announced by the city at the time - the residential containers were to remain for two years.

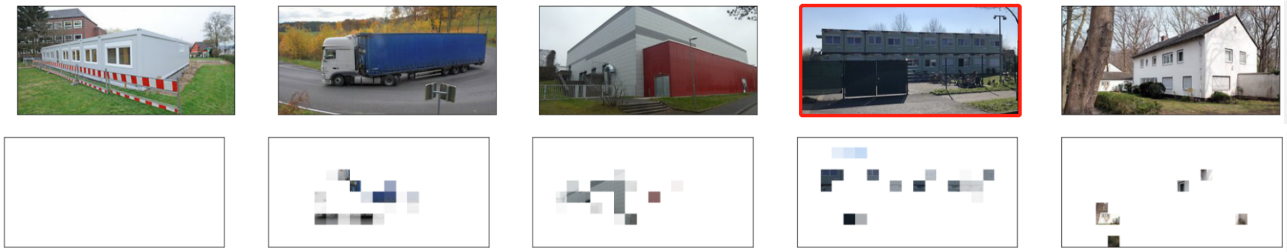

Figure 2: Top5 results returned by text-image retrieval. The word marked in red is the word we selected in the sentence. The picture in the red box is the ground truth. The generated heatmap is in the second row.

consists of a text encoder, an image encoder and a transformer for modality interaction.

Text encoder: The text is processed in the same way as BERT [1]. A modal-type embedding is introduced to distinguish different modalities. The text endcoder consists of three parts: word embedding, token position embedding, and modality-type embedding. First, it converts the input tokens to embeddings with a word embedding matrix. Then, the word embedding is summed with the position embedding and modality-type embedding to create the input of the encoder [10, 16].

Image encoder: Inspired by ViT, the model use a 32 × 32 patch projection to embed the image. The method of patch projection is slicing the image into patches, flattening the patches and mapping them into D dimensions with a trainable linear projection. In other words, patch projections, rather than regions or grid features with high weights, are used to create the image to image embedding [2, 7]. Similar to the text encoder, the image embedding, position embedding, and modality-type embedding are summed.

Modality interaction schema: After we have the image and text embeddings, a transformer is used for both intra-modal and inter-modal interaction. It outputs a contextualized feature sequence for prediction. The image is replaced by a different image with probability of 0.5. Then, we compute the binary loss function.

Loss function: We keep the loss function identical to BERT and it contains two parts: the first is image-text matching (ITM) and the second is masked language modeling (MLM). Words in the sentence are randomly masked with the probability of 0.15 before being input into the model.

4 RESULTS AND ANALYSIS

The task asks participants to predict a ranked list of images corresponding to each text and report the *Recall@K* result. Our results are shown in Table 1. The model uses four datasets for pre-training: Microsoft COCO (MSCOCO), Visual Genome (VG), SBU Captions (SBU), and Google Conceptual Captions (GCC) [8, 9, 11, 14]. In the experiment, we kept the default parameters of the model and fine-tuned it on single GPU from Google Colab with a batch size of 32. We load the pre-trained model and fine-tune it on the task dataset. We also train the model without loading the pre-trained model. The result without the pre-trained model is much worse.

Table 1: Submission result. Comparing pre-trained and not pre-trained model, pre-trained model has better performance.

Method	R@10	R@50	R@100	MRR@100
ViLT (pre-trained)	0.1347	0.3342	0.4658	0.0596
ViLT (no pre-trained)	0.0397	0.1264	0.2183	0.0174

The result demonstrates that the model is learning useful features from the pre-trained datasets. In turn, this means that the characteristics of the pre-training dataset provide an adequate match with the news domain. Our experiment confirms the benefits of pre-training for the NewsImages rematching task.

To better understand the performance of cross-modal alignment, we inspected the top 5 results of our text-image retrieval for a selection of test texts. We also generated a keyword heatmap for each image in the results. Figure 2 shows a random test text from this analysis. The fourth result (with the border) is correct according to the ground truth. However, we can see that many of the other four images that are returned match the text with respect to the word "containers", which is a key concept in the text. This example demonstrates that the pre-training, the model makes our retrieval results closer to the real description scene of the text, even though they may not be the correct match.

5 CONCLUSION AND FUTURE WORK

In this work, we explored the performance of the pre-trained model in the text-image re-matching subtask of NewsImages at MediaEval 2021. Compared to the model without pre-training, the pre-trained model achieved better performance, as expected with a relatively small dataset. However, there is still a big gap compared to the performance improvement delivered by pretraining when text-image retrieval is carried out on other datasets, such as MSCOCO and Flickr30K [9, 12]. In future work, we will try to collect more data in the news domain to help the model improve its performance. We would like to develop a new pre-trained model with more complex modality interaction.

REFERENCES

[1] 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)*.

[2] 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. In *International Conference on Learning Representations*.

[3] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*.

[4] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomás Mikolov. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*.

[5] Benjamin Kille, Frank Hopfgartner, Torben Brodt, and Tobias Heintz. 2013. The Plista dataset. In *Proceedings of the 2013 international news recommender systems workshop and challenge*.

[6] Benjamin Kille, Andreas Lommatzsch, Özlem Özgöbek, Mehdi Elahi, and Duc-Tien Dang-Nguyen. 2021. News Images in MediaEval 2021. In *Proc. of the MediaEval 2021 Workshop, Online, 13-15 December 2021*.

[7] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Vol. 139.

[8] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and others. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* 123, 1 (2017).

[9] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[10] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*.

[11] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*.

[12] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*.

[13] 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. In *Proceedings of the 38th International Conference on Machine Learning*.

[14] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

[15] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*.