Textual Concept Expansion for Text-Image Matching within Online News Content

Mingliang Liang[1], Martha Larson[2]

[1]*Radboud University, Netherlands*

Abstract
We investigate a **Textual Concept Expansion (TCE)** approach to address the NewsImages task in MediaEval'22. Specifically, we use a pre-trained multi-label classifier to predict concepts beyond the words in the captions to enrich the captions. We explore **TCE** because it leverages commonsense knowledge, which can improve the performance in news dataset. The results show that the proposed method achieve a strong performance in text-image retrieval in NewsImages task.

1. Introduction

The goal of the NewsImages task is to learn the relationship between images and articles. Task participants design and implement systems that return images that are related to a query article [1]. The task is challenging because there is complex relationship between images and articles. Specifically, due to the nature of news, not everything depicted in the image is described in the article. As a result, information related to the image is missing from the article. The loose connection of images and articles in news dataset prompted us to explore external knowledge to enrich the articles by textual concept expansion (TCE) [2], which can provide possible co-occurrence concepts related to the images.

In this paper, we propose an new approach to address of NewsImages task [1], which extends the concepts for the articles by a multi-label classifier that pre-trained on MS COCO dataset [3]. We combine these concepts text from the articles as the input of text encoder. Then we fine-tuned our model on a pre-trained model which was pre-trained on a 4M dataset [4].

In tasks of vision-and-language (VL), such as image-text retrieval and visual question answering(VQA), co-attention (cross-modal attention) and merged-attention transformer have been shown to have strong performance in learning the relationship between image and text [5, 6, 4, 7, 8]. The co-attention transformer layer proposed by ViLBERT [7] allows the model to have a deep interaction between different modalities. VisualBERT [9] combines image regions and language with a transformer to align image and text, which is called merged-attention. In this paper, we will apply merged-attention to the News Images dataset.

In additional, vision-and-language pre-training (VLP) has become a popular approach to tackle image-text retrieval task [10, 5, 6, 4, 7, 8]. Learning pre-trained representations from large numbers of image-text pairs can lead to better baseline performance of the model in vision-and-language tasks. Also, the pre-trained model has demonstrated substantial improvement in performance on the NewsImages task dataset [11, 12] in MediaEval 2021. For this reason, we also take advantage of pre-trained model to fine-tune our model to obtain a strong performance on the NewsImages task MediaEval 2022.

---

*Corresponding author.

✉ m.liang@cs.ru.nl (M. Liang); m.larson@cs.ru.nl (M. Larson)

CEUR Workshop Proceedings (CEUR-WS.org)

2. Approach

In this section, we begin with a brief introduction to the NewsImages dataset and data pre-processing. Then we present the technical details of the pre-trained model and our textual concept expansion (TCE) approach.

2.1. Dataset and data pre-processing

The dataset for the NewsImages task includes about 9300 training samples and 4500 test samples released for the MediaEval 2022 [13], the dataset crawled from three different news sources website: Online News portals (rt), Twitter (tw), and RSS news feed (rss). The articles of rt news feed come from German news publisher. Therefore, we translated the German text into English via Google Translate in order to keep consistent with the rest of the text.

2.2. Textual Concept Expansion

We tackled the NewsImages task with a pre-trained merged-attention model with Triple Contrastive Learning (TCL) as a baseline model [4]. Then, we expand the model with textual concept expansion.

Pre-trained Model with Triple Contrastive Learning (TCL): TCL applies triple contrastive learning both cross-modally and intra-modally, which can maintain the similarity of image-text pairs and similar samples from same modality [4]. TCL has a vision encoder, a text encoder, and a fusion encoder. For the both vision encoder and text encoder, TCL utilize two separate data augmentation operators to generate the inputs of a encoder and a momentum encoder [14]. The outputs of both the vision and text encoders are fed into the fusion encoder, which predicts whether the image-text pairs match. TCL is pre-trained on 4.0M images and 5.1M image-text pairs, which consists of four datasets, inculde MS COCO [3], Visual Genome (VG) [15], Conceptual Captions(CC) [16], and SBU Captions [17]. Additionally, we also tested the zero-shot performance of TCL that was not fine-tuned in the target dataset.

Textual Concept Expansion: The information in online news articles the accompanying images is often complementary. To address this challenges, we propose to use Textual Concept Expansion (TCE) [2], which expands a text through more concepts that may co-occur in the same context. In additional, TCE as a method of query expansion has demonstrated effective on image-text matching task [2]. Here, we attempt to explore the effectiveness TCE in text-image matching for online news. For this purpose, we train a pre-trained multi-label classifier on MS COCO dataset, which can predict the co-occurrence concepts to enrich text. Specifically, when we train the multi-label classifier, we first select the top q most frequent concepts of three types (i.e., Object, Motion and Property) from the MS COCO data set as concepts vocabulary $v = \{c_1, c_2, \ldots, c_q\}$. Then, we label the training dataset. In MS COCO dataset, each image has five captions that created by different people to describe the image, contain complementary concepts. As we know, objects that appear in the same scene are closer together in terms of commonsense knowledge than objects that do not appear in the scene. Therefore, we merge the captions of each image in MS COCO, $T_j = \{t_1, t_2, \ldots, t_i\}$, i is the caption numbers of each image, j is the image number. The target labels of classifier is $y_j = \{c_1, c_2, \ldots, c_m\}$, m is the number of co-occurrence concepts for T_j. Finally, we use a pre-trained BERT [18] model and add a multi-label classification layer to train the multi-label classifier by a multi-label classification loss:

Table 1

Evaluation of text-image retrieval approaches on test dataset. We first evaluate the zero-shot performance with pre-trained model, which the model was not fine-tuned, on the NewsImages datasets. We then fine-tune the model on the NewsImages dataset without (Fine-Tune) and with (Fine-Tune + TCE) TCE and the compare the performance.

Dataset	Approach	Image Retrieval				MRR@100
		R@5	R@10	R@50	R@100	
rss	zero-shot	15.33	19.93	37.27	46.53	11.39
	Fine-Tune	18.07	24.40	42.87	55.47	13.40
	Fine-Tune + TCE	20.60	26.93	47.40	58.40	14.73
rt	zero-shot	5.93	08.53	19.00	26.60	4.29
	Fine-Tune	10.33	14.00	28.40	36.67	7.59
	Fine-Tune + TCE	9.67	14.00	30.00	39.13	7.33
tw	zero-shot	21.13	26.93	43.47	51.67	15.48
	Fine-Tune	28.54	35.67	56.20	67.53	21.58
	Fine-Tune + TCE	30.00	37.07	57.67	68.93	22.31

$$Loss = \sum_{c=1}^{C} y^c \log(\sigma(\hat{y}^c)) + (1 - y^c) \log(1 - \sigma(\hat{y}^c)) \tag{1}$$

When we fine-tune the pre-trained TCL model, the pre-trained multi-label classifier be used to predict the concepts of an input caption. Then, we combine each caption and its predicted concepts to use as the input of the text encoder of TCL in order to fine-tune our model. We set the confidence score of the multi-label classifier to 0.1 to select the prediction concepts.

3. Results and Analysis

The task asks participants to predict a ranked list of images corresponding to each text and report the text-image $recall@k$ results. In all of our experiments, we kept the default parameters of the TCL [4] and fine-tuned it on two 3090Ti GPUs with a batch size of 16.

First, we evaluate the TCL that pre-trained on 4M image-pair dataset without fine-tuning on NewsImages dataset. As shown in first experiment Table 1, the zero-shot result of text-image retrieval tasks on NewsImages dataset achieves a strong baseline performance. Then, we fine-tune the model on NewsImages dataset, the MRR@100 increases from 11.39, 4.29, 15.48 to 13.40, 7.59, 21.58 on three test data separately. Next, we go on to evaluate the performance of TCE fine-tune on pre-trained TCL. Comparing the row 2 and row 3 of rss and tw in Table 1, TCE can further improve the performance in rss and tw test data on MRR@100. And rt test data, TCE can improve the performance on R@50 and R@100, from 28.40, 36.67 to 30.00, 39.13. TCE is more improved than rt in the rss and tw datasets. We conjecture that the training dataset for the multi-label classifier is closer to rss and tw than rt. This encourages us to train a more generalization TCE model to expand caption and improve the performance of most vision-language tasks.

The results of our experiments show that textual concept expansion works well on NewsImages datasets, although we pre-trained our multi-label classifier on another dataset that was not very close to the news dataset. The predicted concepts can help the vision and language model to learn more common and general knowledge.

Visualisation Analysis : To better understand the TCE, we give same visualisation examples to show the expand concepts of captions. As shown in the Table 2, The "airplane, passenger

Table 2
The examples of expand the concepts of captions in NewsImages dataset. The first and second columns are image-text pair, and the third column is our expand concepts.

Image	Caption	TCE
	takes plane during ferocious storm and ends up landing in 'wrong country'	sitting white large down blue background air flying day ground sky taking airplane plane passenger airport
	Ireland beat 14-man England 32-15 in Six Nations	man two people group young field other side couple green playing men game together ball ready tennis court four play players
	could fully live in this solar-powered house on wheels	sitting next white top small front has street black sits area building wooden parked back middle lot car parking bike home house

and airport" in the row 1, the "man, people, ball and player" in row 2 and the "building and car" are very useful to recall the right image. As we observe from the Table 2, the concepts are not appear in the captions, but we can train a model to learn the commonsense knowledge to expand the captions.

4. Discussion and Outlook

In this work, we have explored the performance of the textual concept expansion (TCE) on the text-image matching task of NewsImages at MediaEval 2022. Compared with the model without expanded concepts, the captions that are expanded with concepts achieve better performance. In the future work, we would like to train a stronger and better generalizable textual concept expansion model to predict more useful concepts for the caption in the Vision-language tasks.

References

[1] B. Kille, A. Lommatzsch, Özlem Özgöbek, M. Elahi, D.-T. Dang-Nguyen, News Images in MediaEval 2021, in: Proc. of the MediaEval 2021 Workshop, Online, 13-15 December 2021, 2021.

[2] M. Liang, Z. Liu, M. Larson, Textual Concept Expansion with Commonsense Knowledge to Improve Dual-Stream Image-Text Matching, in: International Conference on Multimedia Modeling, 2023.

[3] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: Common Objects in Context, in: Proceedings of the European Conference on Computer Vision, 2014.

[4] J. Yang, J. Duan, S. Tran, Y. Xu, S. Chanda, L. Chen, B. Zeng, T. Chilimbi, J. Huang, Vision-Language Pre-Training with Triple Contrastive Learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[5] J. Li, R. R. Selvaraju, A. D. Gotmare, S. Joty, C. Xiong, S. Hoi, Align before Fuse: Vision and Language Representation Learning with Momentum Distillation, in: Advances in neural information processing systems, 2021.

[6] W. Kim, B. Son, I. Kim, ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision, in: International Conference on Machine Learning, 2021.

[7] J. Lu, D. Batra, D. Parikh, S. Lee, ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks, in: Advances in neural information processing systems, 2019.

[8] Z.-Y. Dou, Y. Xu, Z. Gan, J. Wang, S. Wang, L. Wang, C. Zhu, N. V. Peng, Z. Liu, M. Zeng, An Empirical Study of Training End-to-End Vision-and-Language Transformers, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[9] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, Visualbert: A Simple and Performant Baseline for Vision and Language, arXiv:1908.03557 (2019).

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning Transferable Visual Models from Natural Language Supervision, in: International Conference on Machine Learning, 2021.

[11] M. Liang, M. Larson, Exploring a Pre-trained Model for Re-Matching News Texts and Images, in: Proceedings of the MediaEval Benchmarking Initiative for Multimedia Evaluation 2021, CEUR Workshop Proceedings, 2021.

[12] C. Bartolomeu, R. Nóbrega, D. Semedo, NewsSeek-NOVA at MediaEval 2021: Context-enriched Multimodal Transformers for News Images Re-matching, in: Proceedings of the MediaEval Benchmarking Initiative for Multimedia Evaluation 2021, CEUR Workshop Proceedings, 2021.

[13] A. Lommatzsch, B. Kille, O. Özgöbek, Y. Zhou, J. Tešić, C. Bartolomeu, D. Semedo, L. Pivovarova, M. Liang, M. Larson, NewsImages: Addressing the Depiction Gap with an Online News Dataset for Text-Image Rematching, in: Proceedings of the 13th ACM Multimedia Systems Conference, 2022.

[14] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum Contrast for Unsupervised Visual Representation Learning, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[15] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, M. S. Bernstein, L. Fei-Fei, Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations, International Journal of Computer Vision (2017).

[16] P. Sharma, N. Ding, S. Goodman, R. Soricut, 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), 2018.

[17] V. Ordonez, G. Kulkarni, T. Berg, Im2Text: Describing Images Using 1 Million Captioned Photographs, in: Advances in neural information processing systems, 2011.

[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, 2019.