Integrated Multi-stage Contextual Attention Network for Text-Image Matching

Yi Shao[1], Yawen Chen[1], Yang Zhang[1], Tianlin Zhang[1], Xuan Zhang[2], Ye Jiang[3], Jing Li[1] and Jiande Sun[1,*]

[1]*Shandong Normal University, China*
[2]*Shandong Police College, China*
[3]*Qingdao University of Science and Technology, China*

Abstract

For the Mediaeval competition's NewsImages track, we introduce the Integrated Multi-stage Contextual Attention Network (IMCAN) for effective text-image matching. Our method harnesses the representational power of BERT and Vision Transformer to extract textual and visual features, which are then synergistically enhanced by a series of Transformer encoders integrated with a contextual multi-modal attention mechanism. This architecture significantly improves the alignment and fusion of modality-specific features, crucial for the cross-modal retrieval task. The end-to-end training strategy optimizes the model for precise feature matching, ensuring the robustness of our approach. In our experimental evaluation, we primarily utilize MRR and Mean Recall@K as metrics to measure performance. Comparison with the CLIP model results suggests that there is still much room for model improvement in the text-image matching task, but creative models still show desirability in this Text-Image matching.

1. Introduction

News articles frequently employ a combination of text and image to disseminate information, with textual narratives incorporating visual representations to grab attention and help readers understand the content more intuitively. The process of image-text matching is a measure of the visual and textual similarities that exist between images and accompanying text, which is especially critical for cross-modal retrieval tasks. Although this area has seen considerable progress in recent years, image-text matching continues to pose a significant challenge due to intricate matching patterns and pronounced semantic divergences between the two media. The NewsImages task within MediaEval 2023 has delved into this dynamic.

In this paper, we propose a novel integrated multi-stage contextual attention network (IMCAN) for text-image task, which stands out as a comprehensive solution for the intricate demands of text-image matching, showcasing a deep synergy between both modalities. The multi-stage feature extraction and fusion approach, coupled with the use of cosine similarity to measure text-image distance, effectively selects relevant images with a high degree of similarity to the text, and the model's performance on the three datasets highlights its capabilities in text-image matching.

MediaEval'23: Multimedia Evaluation Workshop, February 1–2, 2024, Amsterdam, The Netherlands and Online
*Corresponding author.
2021020981@stu.sdnu.edu.cn (Y. Shao); chenyawen111@163.com (Y. Chen); 2021317099@stu.sdnu.edu.cn (Y. Zhang); sdnu_tianlinzhang@163.com (T. Zhang); zx@sdpc.edu.cn (X. Zhang); ye.jiang@qust.edu.cn (Y. Jiang); lijingjdsun@hotmail.com (J. Li); jiandesun@hotmail.com (J. Sun)

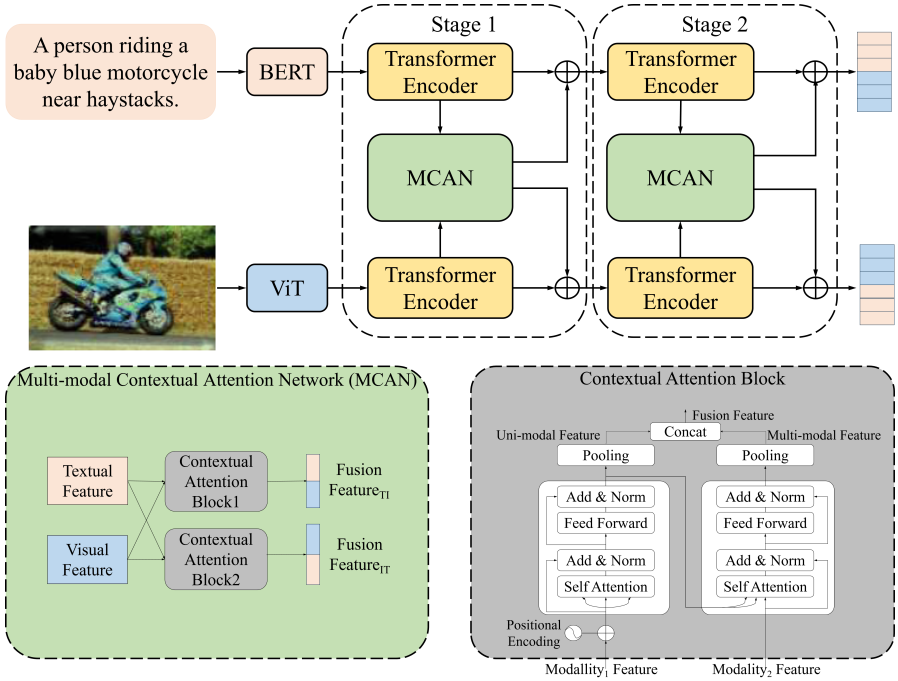

Figure 1: Integrated Multi-stage Contextual Attention Network (IMCAN).

2. Related Work

Text-image matching has attracted extensive research in the multimedia research community, and the emergence of deep learning techniques has significantly improved performance in this area. Tom Sühr et al. [1] devised a method to embed text and image inputs into a unified embedding space, enabling matching of text-image pairs through distance or similarity metrics. Yuta Fukatsu et al. [2] leveraged the ADAPT model for improved cross-modal retrieval in image-text matching, utilizing Swin Transformer and DistilBERT for extracting image and text features, respectively.

Nikolaos Sarafianos et al. [3] introduced TIMAM, a text-image modal adversarial matching method, which exploits both adversarial and cross-modal matching objectives to learn modal invariant feature representations and demonstrates that BERT, a publicly available language model for extracting word embeddings, can be successfully applied to the field of text-to-image matching. Notably, the emergence of the CLIP [4] model has inspired several enhancements in image-text matching models, harnessing its powerful capabilities for improved performance [5].

3. Approach

In the NewsImages track of the Mediaeval 2023 competition, we introduce an IMCAN model for cross-modal retrieval tasks, which integrates multi-stage transformer encoders and a Multi-modal Contextual Attention Network (MCAN) for refining and fusing the textual and visual features extracted via BERT [6] and ViT [7]. The overall architecture is illustrated in 1.

Specifically, we construct a multi-stage Transformer encoder designed to further deepen and refine the textual and visual features extracted via BERT and ViT. Each stage of the encoder concentrates on extracting higher-level semantic representations, thereby enhancing the model's capability to comprehend complex semantic information. This multi-stage feature extraction

Table 1
Official evaluation results of CLIP on three datasets

Dataset	CLIP version	MRR@100	Recall@5	Recall@10	Recall@50	Recall@100
GDELT-1	Standard	0.52019	0.68400	0.77067	0.90800	0.94800
GDELT-2	Standard	0.46503	0.59933	0.69200	0.86133	0.91600
RT	Standard	0.18607	0.25333	0.32000	0.52133	0.60500
	Multilingual	0.12076	0.16833	0.22800	0.41567	0.50400

Table 2
Official evaluation results of IMCAN trained with different loss functions

Dataset	Loss Function	MRR@100	Recall@5	Recall@10	Recall@50	Recall@100
GDELT-1	Cosine Similarity Loss	**0.00567**	**0.00733**	**0.01267**	**0.03867**	**0.07400**
	Triplet Loss	0.00530	0.00533	0.00800	0.03733	0.07333
GDELT-2	Cosine Similarity Loss	**0.00435**	**0.00467**	**0.01000**	**0.04133**	0.07667
	Triplet Loss	0.00372	**0.00467**	0.00800	0.03600	**0.07933**

approach aids our model in capturing richer and more nuanced data features. To facilitate effective fusion of textual and image features, we introduce a Multi-modal Contextual Attention Network (MCAN). This network comprises two contextual attention blocks, which process the text and image information from different Transformer encoding stages. This work to accurately align and integrate features from both modalities, thus enhancing the model's ability to capture cross-modal correlations. MCAN not only fuses multi-modal information but also ensures the richness and consistency of the final feature representation, providing robust support for retrieval tasks.

4. Results and Analysis

4.1. Comparative Experiment

Table 1 shows the performance of CLIP fine-tuned on three datasets. CLIP shows excellent retrieval performance on GDELT-1 and slightly degrades on GDELT-2, but still maintains high retrieval recall on a larger subset. Standard CLIP performs poor inference directly on the German RT data set because it is trained on English data. We respectively used standard CLIP to reason on the English machine-translated text of the RT data set, and used multilingual CLIP to reason on the German text of the RT data set. It can be seen that it is a good strategy to machine-translate the German text into English and then use standard CLIP to fine-tune it.

Table 2 shows the performance of the proposed model IMCAN on two English datasets. Limited by time and computing resources, we only submitted the training results on the English data set to the official in time. The main function of the proposed model is to put text features and image features into the same feature space, so we use cosine similarity loss and triplet loss for training respectively. Specifically, for training with triplet loss, we augment each text and image data separately using the nlpaug package and affine transformation, generating 5 augmented samples as positive sample set. We randomly select 100 samples from all non-matching samples and choose the 5 samples with the lowest loss as negative sample set to construct the triplet. The triplet results shown in Table 2 are the best results obtained by taking the boundary value every 0.1 from 0.1 to 0.9. Table 2 shows that the MRR of the model trained with cosine similarity

Table 3
Performance comparison of ablation models trained with cosine similarity loss on the GDELT-1 dataset

Number of Stages	MRR@100	Recall@5	Recall@10	Recall@50	Recall@100
0	0.0086	0.0059	0.0118	0.0588	0.1175
1	0.0149	**0.0129**	0.0224	0.0964	0.1753
2	**0.0151**	**0.0129**	**0.0282**	**0.1118**	**0.2035**
3	0.0132	**0.0129**	0.0165	0.0671	0.1529
4	0.0097	0.0059	0.0106	0.0682	0.1376

loss is higher than that of triplet loss. When the test subset capacity is small, the recall rate of the model trained by cosine similarity is significantly higher than that of the triplet loss. When the subset reaches 100, the performance of the models trained with the two losses is close. This shows that it is difficult for the triplet loss to make an optimal match among similar samples in this task, but it can still find relatively similar options in a large number of samples.

4.2. Ablation Experiment

According to Table 1 and Table 2, even though the proposed model has obvious commonalities with CLIP in structure, its performance on the English data set lags significantly behind CLIP. In addition to differences in computing resources such as the amount of training data and batch size, we also considered the rationality of the structure of the model itself. Specifically, we organized IMCAN with different numbers of stages as ablation experiments to study whether the current number of stages is reasonable.

Table 3 shows the performance of different numbers of stages. Here, 0 stage means directly computing the similarity between text features extracted by BERT and image features extracted by ViT. These ablation models with different numbers of stages were trained on GDELT-1 dataset with a training set:test set ratio of 8:2 for 100 epochs using cosine similarity loss function and have converged. Table 3 demonstrates that when the number of stages increases from 0 to 2, the MRR and recall rate of all subset sizes show a steady improvement. However, when the number of stages increases from 2 to 4, the MRR and recall rate of all subset sizes show a steady decline. When the number of stages reaches 3 or 4, we found that although the training loss steadily decreased, the MRR and Recall@K of the test set showed slow and severely fluctuating improvements, with this phenomenon more pronounced in the 4-stage model than in the 3-stage model. This indicates that when the number of stages reaches 3 and 4, the model may have memorized too much noise due to the excessive number of parameters, leading to overfitting. Therefore, we did not attempt to use models with 5 or more stages, and considered the 2-stage model as optimal.

ACKNOWLEDGMENTS

Thanks to the organizers of the MediaEval2023, especially to those organizers for NewsImages. This work was supported in part by the Joint Project for Innovation and Development of Shandong Natural Science Foundation (Grant No. ZR2022LZH012), and in part by the Joint Project for Smart Computing of Shandong Natural Science Foundation (Grant No. ZR2020LZH015).

References

[1] T. Sühr, A. Madhavanr, N. J. Avanaki, R. Berk, A. Lommatzsch, Image-Text Rematching for News Items using Optimized Embeddings and CNNs in MediaEval NewsImages 2021, in: Proceedings of the MediaEval workshop, CEUR-WS.org, 2021. URL: https://ceur-ws.org/Vol-3181/paper11.pdf.

[2] Y. Fukatsu, M. Aono, Image-Text Re-Matching Using Swin Transformer and DistilBERT, in: Proceedings of the MediaEval workshop, CEUR-WS.org, 2021. URL: https://ceur-ws.org/Vol-3181/paper26.pdf.

[3] N. Sarafianos, X. Xu, I. A. Kakadiaris, Adversarial Representation Learning for Text-to-Image Matching, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 5814–5824.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning Transferable Visual Models from Natural Language Supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[5] Y. Zhang, Y. Shao, X. Zhang, W. Wan, J. Li, J. Sun, CLIP Pre-trained Models for Cross-modal Retrieval in NewsImages 2022, in: Proceedings of the MediaEval workshop, CEUR-WS.org, 2022. URL: https://2022.multimediaeval.com/paper3975.pdf.

[6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv:1810.04805 (2018).

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, arXiv preprint arXiv:2010.11929 (2020).