COTS: Collaborative Two-Stream Vision-Language Pre-Training Model for Cross-Modal Retrieval
– Supplementary Material –

Haoyu Lu[1,2] Nanyi Fei[1] Yuqi Huo[1] Yizhao Gao[1] Zhiwu Lu[1,2,*] Ji-Rong Wen[1,2]
[1]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
[2]Beijing Key Laboratory of Big Data Management and Analysis Methods
{lhy1998, feinanyi, bnhony, gaoyizhao, luzhiwu, jrwen}@ruc.edu.cn

1. Discussion on Task-Level Interaction

The main goal of cross-modal learning is to find a joint space where images and texts are aligned. Our task-level interaction is also towards this goal. Note that the contrastive loss can be seen as a classification loss: $\mathcal{L}_{\text{I2T}} = -\mathbb{E}_{(v_i,l_i)\sim\mathcal{D}} \log p(v_i|\{l_i\} \cup Q^l)$, where the image v_i is classified to the pseudo class denoted by the paired text l_i among candidates $\{l_i\} \cup Q^l$. As samples in two queues Q^v&Q^l are one-to-one paired, the image and text from each pair form a pseudo class. Thus I2T and T2I tasks can be viewed as classification over the same candidate classes, and aligning the distributions of the two directions is intrinsically sound. This is also indirectly supported by our observation that even when the task-level KL loss is not explicitly applied, its value still decreases during training. In this work, we are thus inspired to use this task-level KL loss to further enhance higher-level interaction (w.r.t. instance-level one). Importantly, the ablation study results in Table 3 do verify the effectiveness of such task-level interaction.

2. Architecture Details

Text and Image Encoders. In our COTS, we adopt the BERT-base [5] model as our text encoder, which contains a total of 12 Transformer layers with 768 hidden units and 12 heads. Meanwhile, we deploy ViT-B/16 [8] as our image encoder. The dimensions of the output vectors of the image and text tokens are both $N_{seq} \times 768$, where N_{seq} is the sequence length. For each image, the final output vector of the [CLS] token is used as the image embedding. And each text embedding is obtained by averaging output vectors of all the text tokens. We then apply a single fully-connected layer for each modality to project the image/text embeddings to a joint cross-modal space. The final dimensions of the image and text embeddings are 256.

Model	R@1	R@5	R@10	MR↓
VSE [12]	5.0	16.4	24.6	1500.0
VSE++ [12]	5.7	17.1	24.8	47.0
W2VV [6]	6.1	18.7	27.5	45.0
GPO [3]	8.7	25.3	35.9	-
HGR [4]	9.2	26.2	36.5	24.0
COOKIE [14]	9.8	28.3	39.6	-
CE [11]	10.0	29.0	41.2	16.0
MMT [9]	10.7	31.1	43.4	15.0
Dual Encoding [7]	11.6	30.3	41.3	17.0
COTS (5.3M)	17.4	38.8	49.7	11.0
COTS (15.3M)	**19.2**	**41.6**	**52.8**	**9.0**

Table 1. Comparison to the state-of-the-arts for text-to-video retrieval on MSR-VTT [15] under the full split setting. Notations: ↓ denotes that lower results are better.

Image Tokenizer. For each raw image, we first apply an average pooling layer to resize it from 384×384 to 192×192. Further, we utilize the pre-trained discrete variational auto-encoder (dVAE) [13] as the image tokenizer to obtain a sequence of 24×24 discrete image tokens. In this work, for performing our cross-modal masked vision modeling (CMVM) in our COTS, we apply a fully-connect layer as the CMVM Head to predict the masked tokens.

3. More Text-to-Video Retrieval Results

In this section, we provide more results for text-to-video retrieval on MSR-VTT [15] under the full split setting.

Implementation Details. We adopt the Adam [10] optimizer with a weight decay of 0.02 for text-to-video retrieval. We select hyper-parameters heuristically due to computational constraint: the batch size is 48, the momentum hyper-parameter $m = 0.99$, temperature $\tau = 0.05$, and the queue size N_Q is 1,444 for finetuning on MSR-VTT. We set the initial learning rate to 5e-5 for the first epoch, and decay

*The corresponding author.

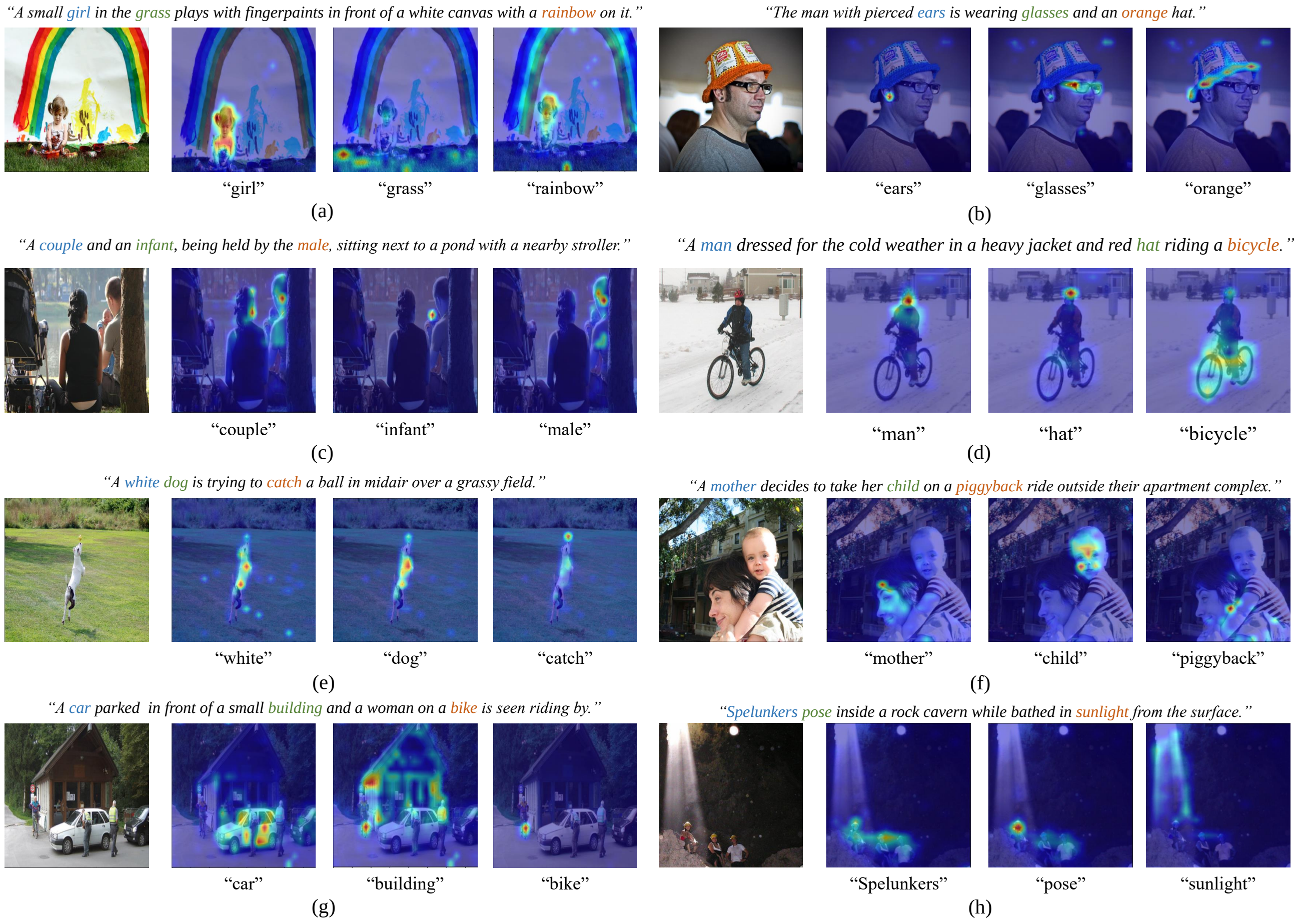

Figure 1. Visualizations of attention maps of our COTS using GAE [2] on images responding to individual words.

the learning rate linearly in the rest epochs. For each video, we extract the feature embeddings of 16 frames and take the average embedding as the video representation (we only employ half of the frames used in Frozen in Time [1]).

Full Split Results. Table 1 presents the comparative results for text-to-video retrieval on MSR-VTT under the full split setting (as in COOKIE [14]). It can be observed that: (1) Our COTS (5.3M) significantly outperforms all competitors by large margins on all evaluation metrics, which clearly validates the general applicability and the transfer ability of our COTS. (2) Compared with the latest model Dual Encoding [7], our COTS achieves higher results by 5.8% (17.4% vs. 11.6%) on R@1 and 8.4% (49.7% vs. 41.3%) on R@10. This also demonstrates the effectiveness of COTS. (3) When leveraging a larger pre-training dataset, our COTS (15.3M) further improves the performance.

4. More Attention Visualization Results

More visualizations of attention maps obtained by our COTS are shown in Figure 1. It can be observed that our COTS has the ability to well locate different objects (e.g., "girl" in Figure 1(a), "sunlight" in Figure 1(h)) and even capture fine-grained information (e.g., "ears" in Figure 1(b), "hat" in Figure 1(d), and "bike" in Figure 1(g)). Interestingly, our COTS can also capture color concepts (e.g., "orange" in Figure 1(b), "white" in Figure 1(e)) and actions (e.g., "catch" in Figure 1(e), "pose" in Figure 1(h)). Moreover, as shown in Figure 1(c) and (f), our COTS can correctly determine human information (i.e., gender and age). Overall, these visualization results demonstrate that our two-stream based COTS is able to identify multiple objects (and even fine-grained information) without introducing any cross-modal module like single-stream models.

5. Visualization of Momentum Similarity Scores in Adaptive Momentum Filter

As we have mentioned in Section 3.3, we propose an adaptive momentum filter (AMF) module to filter noisy image-text pairs based on their momentum similarity scores. We visualize the momentum similarity scores of several image-text pairs sampled from CC12M in Figure 2. It can be seen that for each image-text pair with a high

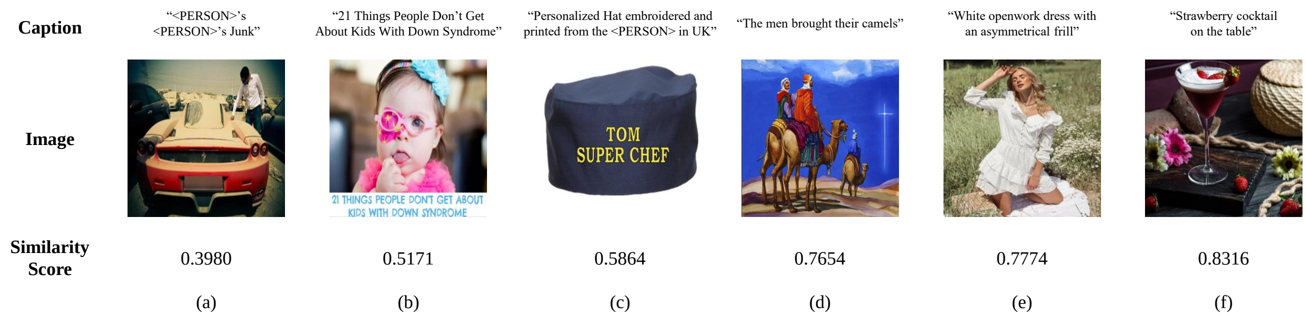

Figure 2. Examples of momentum similarity scores of several image-text pairs sampled from CC12M.

similarity score, there is a strong semantic correlation between its image and text (as shown in Figure 2(d)–(e)). On the contrary, the low similarity score typically indicates that the paired image and text have a weak semantic correlation or even no semantic correlation (as shown in Figure 2(a)–(c)). Specifically, in Figure 2(a), there is a man touching his red car in the image, while the corresponding caption is "<PERSON>'s <PERSON>'s Junk". Since the text is totally meaningless, such image-text pair could have negative effects on vision-language pre-training and thus needs to be filtered/removed. Overall, the similarity scores calculated by the AMF module are well in line with human judgements, indicating the effectiveness of our AMF.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (61976220 and 61832017), Beijing Outstanding Young Scientist Program (BJJWZYJH012019100020098), and Large-Scale Pre-Training Program 468 of BAAI.

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, pages 1728–1738, 2021. 2

[2] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *ICCV*, pages 397–406, 2021. 2

[3] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *CVPR*, pages 15789–15798, 2021. 1

[4] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *CVPR*, pages 10635–10644, 2020. 1

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2018. 1

[6] Jianfeng Dong, Xirong Li, and Cees G. M. Snoek. Predicting visual features from text for image and video caption retrieval. *IEEE Trans. Multim.*, pages 3377–3388, 2018. 1

[7] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. Dual encoding for video retrieval by text. *TPAMI*, 2021. 1, 2

[8] 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. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[9] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, pages 214–229, 2020. 1

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

[11] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *BMVC*, page 279, 2019. 1

[12] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K. Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *ICMR*, pages 19–27, 2018. 1

[13] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021. 1

[14] Keyu Wen, Jin Xia, Yuanyuan Huang, Linyang Li, Jiayan Xu, and Jie Shao. COOKIE: Contrastive cross-modal knowledge sharing pre-training for vision-language representation. In *ICCV*, pages 2208–2217, 2021. 1, 2

[15] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016. 1