Object Prior Embedded Network for Query-Agnostic Image Retrieval

Yikang Li Jen-hao Hsiao Chiuman Ho
OPPO US Research Center
{yikang.li1, mark, chiuman}@oppo.com

Abstract

The Text-to-Image retrieval task plays an important role in bridging the gap between vision and language modalities. This task is challenging and far from being solved, because of the large visual-semantic discrepancy between language and vision. Recent studies on vision-language contrastive learning have shown that it can effectively learn good representations from massive image-text pairs. However, most existing methods simply concatenate image and text features as input and resort to the deep network to learn the visual-semantic relationship between image and text in a brute force manner. The insufficient alignments information pose a challenging weakly-supervised learning task, and results in only limited accuracy in previous methods. Motivated by the observation that the salient objects in an image can be accurately detected and are often mentioned in the paired text, in this paper, we propose a novel cross-attention transformer that uses objects detected in image as anchor points and prior to significantly ease the learning of image-text alignments, and thus boost the text-to-image search accuracy. In addition, unlike the query-dependent architectures adopted by most previous methods, our proposed method is query-agnostic and is thus significantly faster in the inference process. The extensive experiments on Flickr30K and MSCOCO captions datasets demonstrate that our proposed method can outperform the SOTA method while preserving the inference efficiency.

1. Introduction

The Text-to-Image retrieval task is one important research area aiming at bridging the gap between vision and language. Methods have been proposed recently for obtaining a better understanding of the cross-modality alignment and higher accuracy in the cross-modal retrieval task [2, 3, 5, 8, 9, 12, 16, 21, 24]. Pre-training with tremendous image-sentence pair datasets [4, 10, 11, 13, 15, 17, 23] and cross-modal alignment architecture using benchmark datasets [5,9,21] are the two primary streams of methodologies for creating the semantic space for vision and language modality.

The pre-training strategies that use a large number of image-sentence pairs [11, 15, 17] is capable of obtaining better understanding of the semantic meaning of the visual modality. However, directly applying the pre-trained model directly into the Text-to-Image retrieval task is hard because either the high cost of fine-tuning of the large pre-trained model with hundreds of millions of parameters or the unsatisfactory performance of zero-shot classification with pre-trained model on the benchmark datasets is usually not a common solution for this task.

The cross-modal alignment architecture [5,9,21] is often lightweight and more efficient compared to the pre-trained model. Moreover, the cross-modal alignment architecture commonly utilizes the region features to improve the capability of catching the local information so that the model may focus on more details that are significant in matching the image with distinct words. Nonetheless, the commonly used feature extraction networks like ResNet [7], VGG [19], Faster-RCNN [18], and Bottom-up Attention Model [1] are not able to utilize any pre-defined semantic space to obtain a better performance.

Therefore, in this paper, we propose an algorithm to combine the pre-trained semantic image embeddings with regional semantic features via a cross-attention transformer to enhance the performance of the Text-to-Image retrieval task. The well pre-defined semantic space is utilized as the initialization for our method and the model is easier to be trained compared to the pre-trained model. The cross-attention transformer takes the semantic image embedding as the query and the semantic regional features as the key and value to maintain and enhance the global information with local information. Furthermore, the proposed approach is built on a query-agnostic retrieval architecture, which is more efficient than query-dependent cross-modal alignment methods. The flowchart of the proposed model is illustrated in Fig. 1

The main contribution of our paper can be concluded as:

- We propose an algorithm to collaborate the semantic image and region embeddings for Text-to-Image retrieval task.

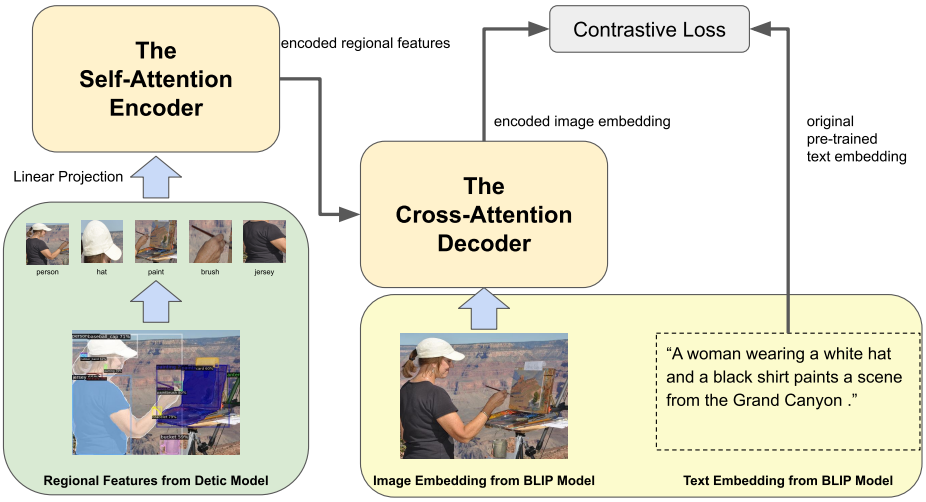

Figure 1. The Flowchart of our proposed algorithm: the regional features are extracted by **Detic** [25] pre-trained model and the sequence of the regional features are encoded with the Self-Attention based encoder. Then the encoded regional features are fed into the Cross-Attention based decoder as key and value while the extracted image embedding from the pre-trained image encoder [11] model are taken as the query to obtain a new encoded image embedding. The model is trained with contrastive loss which is depended on the original text embedding from the pre-trained text encoder [11] model and the encoded image embedding out of the cross-attention decoder.

- To our knowledge, we are the first to utilize a cross-attention transformer to encode the image embeddings as queries and region embeddings as keys and values.
- Our proposed algorithm can outperform the SOTA cross-modal alignment model in Flickr30K and MSCOCO captions dataset.

2. Related Work

In recent years, more and more researchers proposed the pre-training strategy in the multi-modal learning area [4, 10, 11, 13, 15, 17, 23, 25, 26]. The pre-trained model is able to be applied to many downstream tasks like Text-to-Image Retrieval, Visual Question Answering, Image Captioning, and etc. However, the pre-training strategy requires huge amount of paired data [11, 15, 17] which is costly and is more suitable as an initialization rather than fine-tuning in a relatively small benchmark dataset, which may result in the overfitting issue. Furthermore, the alignment of objects and word tokens is usually not known throughout the training process due to the expensive and time-consuming acquisition of massive amounts of paired image-sentence data. As a result, the lack of regional information leads to a degrade of the semantically image understanding in detailed information. Some experiments conducted in this paper show the limitation of directly applying the pre-trained model on the benchmark datasets with linear probe. Details can be found in Section 4

Cross-modal alignment architecture aims at achieving better performance on the benchmark datasets with more sophisticated architecture designing [5, 9, 16, 21, 24]. Compared to the pre-training strategy, the cross-modal alignment commonly utilizes the regional information as the model inputs and thus improving the performance of the Text-to-Image retrieval task. However, the commonly used feature extraction networks like ResNet [7], VGG [19], Faster-RCNN [18], and Bottom-up model [1] do not take into account any pre-defined semantic space so the training process is constrained by the training dataset and the domains between the training and testing set are suffering the shifting problem which will impact the performance.

Our proposed method is designed to overcome the disadvantage of both strategies and we mainly compared our proposed method with the cross-modal alignment architecture methods since we do not fine-tune the feature extraction backbone network during the training process. Detailed information will be discussed in the following section.

3. Methodology

In this paper, we propose a cross-attention transformer to collaborate the pre-trained semantic image embeddings and the semantic regional features. As we mentioned in Section 1 and 2, there are several popular and powerful vision-language pre-training models [11, 15, 17] which are trained with billions of image-sentence pairs. Those well pre-defined semantic spaces are good initialization for our algorithm and we utilize the [11] as the image and text encoder to obtain the features respectively. The image and text features can be represented as:

$$I = \mathcal{F}_{img}(image) \tag{1}$$

$$T = \mathcal{F}_{txt}(text) \tag{2}$$

where the $\mathcal{F}_{img}$ and the $\mathcal{F}_{txt}$ are the pre-trained image and text encoder. The I and T are the image and text features in the pre-defined semantic space respectively.

For the pre-trained regional feature extraction, we utilize the **Detic** [25], which is an open-vocabulary object detection model, to collect the semantic regional information. The open-vocabulary object detection task can be regarded as the 'zero-shot learning' in object detection task since the **Detic** is trained with semantic label information in the pre-trained space [17] as the box label classification weights. As a result, the feature extracted from each proposal can be thought of as holding semantic meaning. Furthermore, since the object detection model is able to detect very small objects with a low resolution which are normally trivial for the Text-to-Image retrieval task, we constrain the number of detected regional features by their area proportion to the entire image size. One of the extracted regional features can be represented as following:

$$O_i = \begin{cases} \mathcal{F}_{obj}(image), & \text{if } \frac{Area(obj_i)}{Area(image)} > \alpha \\ \text{skip}, & \text{otherwise} \end{cases} \tag{3}$$

where the $\mathcal{F}_{obj}$ is the regional feature extraction function from the **Detic** which obtain the features for each valid bounding box proposal. The $Area(obj_i)$ and $Area(image)$ is the area of ith object and the entire image respectively, and α is the pre-defined area proportion. From the Eq. 3 we can find that the number of regional features varies depending on the image, thus we pre-define a list with a fixed number to store all the regional features of all the images and pad the list with zeros and mask if the number of regional features is less than the fixed number just as what the language models commonly do. More detailed settings will be disclosed in Section 4. Therefore, the regional features can be presented as following:

$$O = [O_0, O_1, ..., O_i] \tag{4}$$

where i is the fixed length of the list of the regional features.

Then all the features are taken as inputs for the cross-attention transformer architecture. The proposed cross-attention transformer is consists of two structures, the **encoder** and the **decoder**. The encoder takes the regional features as input and the regional features are projected into the same space, which the semantic image embeddings hold, before being fed into the encoder. The projection can be defined as:

$$\hat{O} = \mathcal{F}_{proj}(O) = [\hat{O_0}, \hat{O_1}, ..., \hat{O_i}] \tag{5}$$

where the $\mathcal{F}_{proj}$ is the linear projection from the pre-trained semantic object space to the semantic image space.

Encoder: the encoder is constructed with several attention blocks which is built upon the Multi-Head Self-Attention mechanism (**MHSA**) with layer normalization (**LN**), feed forward network (**FFN**) and residual shortcuts. The query, key, and the value of the **MHSA** are all the $\hat{O}$ and the corresponding masks. Therefore, each attention block can be described as following:

$$q = k = v = \mathbf{LN}(\hat{O}) \tag{6}$$

$$X = \hat{O} + \mathbf{MHSA}(q, k, v, mask) \tag{7}$$

$$\hat{X} = X + \mathbf{FFN}(X) \tag{8}$$

where q, k, v is the query, key, and value for the **MHSA** respectively. The $mask$ is the mask for the sequence of the regional features. The $\hat{X}$ is the final output of each attention block module. For the **Encoder**, we stack several attention blocks together to encoder the sequence of regional features.

Decoder: the decoder has the same architecture as the encoder. The only difference is the input query changes from the regional features to the semantic image embeddings (i.e. I in Eq. 1) and the key and value is the encoded regional features (i.e. $\hat{X}$ from Eq. 8). It can be explained as the image embeddings are weighted by the semantic regional features so that the image embeddings will contain semantic local information. The attention block in the decoder can be presented as following:

$$q = \mathbf{LN}(I), k = v = \mathbf{LN}(\hat{X}) \tag{9}$$

$$Y = I + \mathbf{MHSA}(q, k, v, \mathbf{None}) \tag{10}$$

$$\hat{Y} = Y + \mathbf{FFN}(Y) \tag{11}$$

where **None** in Eq. 10 means the mask is not required for the image query and $\hat{Y}$ is the final output of the decoder.

The entire model is guided by the same contrastive loss in the [11, 15, 17]. The contrastive loss takes similarity matrix of the image and text embeddings as input and try to optimize the cosine similarity between paired and non-paired data. The contrastive loss can be represented as following:

$$\mathcal{L}_{I-T} = -\frac{1}{M}\sum^{M} log\frac{\exp\left(S(\hat{Y}, T)/\tau\right)}{\sum^{M}\exp S(\hat{Y}, T)} \tag{12}$$

$$\mathcal{L}_{T-I} = -\frac{1}{M}\sum^{M} log\frac{\exp\left(S(T, \hat{Y})/\tau\right)}{\sum^{M}\exp S(T, \hat{Y})} \tag{13}$$

$$\mathcal{L} = \frac{\mathcal{L}_{I-T} + \mathcal{L}_{T-I}}{2} \tag{14}$$

where M is the batch size, the τ is the temperature hyper-parameter and S is the similarity matrix of (image, text) pairs which is computed by text embeddings T and re-weighted image embeddings $\hat{Y}$. $\mathcal{L}_{I-T}$ and $\mathcal{L}_{T-I}$ is the image-to-text and the text-to-image loss respectively.

Table 1. The Performance of the Retrieval Task on **Flickr30K**

Method name	Image-to-Text Retrieval			Text-to-Image Retrieval		
	R@1 ↑	R@5 ↑	R@10 ↑	R@1↑	R@5 ↑	R@10↑
SCAN [9]	67.4	90.3	95.8	48.6	77.7	85.2
SGM [20]	71.8	91.7	95.5	53.5	79.6	86.5
CAAN [24]	70.1	91.6	97.2	52.8	79.0	87.9
DPRNN [3]	70.2	91.6	95.8	55.5	81.3	88.2
MMCA [21]	74.2	92.8	96.4	54.8	81.4	87.8
IMRAM [2]	74.1	93.0	96.6	53.9	79.4	87.2
SHAN [8]	74.6	93.5	96.9	55.3	81.3	88.4
SGRAF [5]	77.8	94.1	97.4	58.5	83.0	88.8
MEMBER [12]	77.5	94.7	97.3	59.5	84.8	91.0
DIME [16]	81.0	95.9	98.4	63.6	88.1	93.0
PreTrain-linear-prob	82.3	96.9	98.8	65.1	89.9	93.4
Ours	**88.9**	**98.2**	**99.2**	**73.6**	**92.9**	**96.2**

Table 2. The Performance of the Retrieval Task on **MSCOCO 5K** (*note: since DPRNN [3] and SHAN [8] do not present the results on the MSCOCO 5K dataset, so we do not cite them in the table)

Method name	Image-to-Text Retrieval			Text-to-Image Retrieval		
	R@1 ↑	R@5 ↑	R@10 ↑	R@1↑	R@5 ↑	R@10↑
SCAN [9]	50.4	82.2	90.0	38.6	69.3	80.4
SGM [20]	50.0	79.3	87.9	35.3	64.9	76.5
CAAN [24]	52.5	83.3	90.9	41.2	70.3	82.9
MMCA [21]	54.0	82.5	90.7	38.7	69.7	80.8
IMRAM [2]	53.7	83.2	91.0	39.7	69.1	79.8
SGRAF [5]	58.8	84.8	92.1	41.6	70.9	81.5
MEMBER [12]	54.5	82.3	90.1	40.9	71.0	81.8
DIME [16]	59.3	85.4	91.9	43.1	73.0	83.1
PreTrain-linear-prob	62.2	86.6	92.7	47.4	75.2	84.4
Ours	**67.8**	**89.0**	**94.2**	**52.4**	**78.5**	**86.7**

Table 3. The performance of inference speed on **MSCOCO 5K**

Method	Time (second)
query-dependent	6349.5
query-agnostic	**167.2**

4. Settings and Experiments

In this paper, we evaluate our algorithm with the Text-to-Image retrieval task on two benchmark datasets: **Flickr30K** [22] and **MSCOCO captions** [14]. There are 123,287 photos in the **MSCOCO** dataset, and each image has five annotated captions. There are 113,287 photos for training, 5000 images for validation, and 5000 images for testing in the dataset. The results are reported by testing on the full 5K images. There are 31,783 photos in the **Flickr30K** collection, each with 5 captions. We use the train-test split as described in [6].

In our paper, we adopt the commonly used Recall at K (R@K), which is defined as the proportion of queries with a ground-truth ranking in the top K. Our evaluation metrics are R@1, R@5, and R@10 demonstrating in the paper.

Settings: we fix the length i of the list of the regional features as 5 and 10 for Flickr30K and MSCOCO respectively. The area ratio α of selecting regional feature is set to 0.1. The temperature in loss function is 1. The number of attention blocks in the encoder and decoder is set to 6 and the dimension of pre-trained image embeddings and regional features are 256 and 512 respectively. The number of attention heads in MHSA is 8 and the batch size is 32. We utilize the AdamW optimizer and the learning rate is set to be 10^{-6}.

Experiments: we compare our algorithm with several cross-modal alignment architectures, SGRAF [20], SCAN [9], CAAN [24], DPRNN [3], MMCA [21], IMRAM [2], DIME [16], SGM [20], MEMBER [12], SHAN [8]. The performances of Text-to-Image retrieval task on two datasets are shown in the Table. 1 and Table. 2.

From both tables, we can observe that our algorithm can outperform the other baseline methods by at least 7.9% in $R@1$ on **Flickr30K** and 8.5% in $R@1$ on **MSCOCO 5K** dataset. Furthermore, we conduct the linear probe experiment with the pre-trained model [11] on both datasets. Surprisingly, only a simple linear probe learning method of the pre-trained model can outperform the SOTA cross-modal alignment methods. Nevertheless, the performance of our proposed model demonstrates that regional or local information is crucial in the Text-to-Image retrieval task as well.

In Table. 3, we conduct a simple but interesting experiment to show the difference in inference speed between query-dependent and query-agnostic models. We build a dummy model with the same architecture except for the final loss function. The contrastive loss is replaced with a Binary Cross-Entropy loss that requires both text and image information, making the model query-dependent rather than query-agnostic. Then we run the evaluation on the **MSCOCO** and the inference speed is shown in the Table. 3. We can see that the query-agnostic model is much faster than the query-dependent model in inference stage.

Ablation: another ablation experiment is conducted by learning with only **Detic** features. The results are not shown in the table since the search ranking acts like a random guess. This inferior performance reflects that the semantic image features is crucial for improvement of those cross-modal alignment methods in Text-to-Image retrieval task.

5. Conclusion

In this paper, we propose an algorithm that combines the pre-trained semantic regional features with semantic image embeddings by utilizing the cross-attention based transformer for the Text-to-Image retrieval task. The experiments on both Flickr30K and MSCOCO show the better capability of the proposed model in the acquisition of image understanding. Furthermore, unlike the common query-dependent cross-modal alignment methods, our model can do the inference process in a query-agnostic fashion which is significantly faster.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 1, 2

[2] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12655–12663, 2020. 1, 4

[3] Tianlang Chen and Jiebo Luo. Expressing objects just like words: Recurrent visual embedding for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10583–10590, 2020. 1, 4

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 1, 2

[5] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. *arXiv preprint arXiv:2101.01368*, 2021. 1, 2, 4

[6] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013. 4

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2

[8] Zhong Ji, Kexin Chen, and Haoran Wang. Step-wise hierarchical alignment network for image-text matching. *arXiv preprint arXiv:2106.06509*, 2021. 1, 4

[9] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018. 1, 2, 4

[10] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021. 1, 2

[11] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 1, 2, 3, 4

[12] Jiangtong Li, Liu Liu, Li Niu, and Liqing Zhang. Memorize, associate and match: Embedding enhancement via fine-grained alignment for image-text retrieval. *IEEE Transactions on Image Processing*, 30:9193–9207, 2021. 1, 4

[13] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 1, 2

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4

[15] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021. 1, 2, 3

[16] Leigang Qu, Meng Liu, Jianlong Wu, Zan Gao, and Liqiang Nie. Dynamic modality interaction modeling for image-text retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1104–1113, 2021. 1, 2, 4

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1, 2

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 2

[20] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1508–1517, 2020. 4

[21] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10941–10950, 2020. 1, 2, 4

[22] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 4

[23] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 1, 2

[24] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z Li. Context-aware attention network for image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3536–3545, 2020. 1, 2, 4

[25] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Phillip Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. *arXiv preprint arXiv:2201.02605*, 2022. 2, 3

[26] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755, 2020. 2