[Supplementary Material] Multi-Modal Representation Learning with Text-Driven Soft Masks

Jaeyoo Park[1] Bohyung Han[1,2]
Computer Vision Laboratory, ECE[1] & IPAI[2], Seoul National University
{bellos1203, bhhan}@snu.ac.kr

1. Datasets for Pretraining

We use the large-scale image-text corpora, which include MS-COCO [10], Visual Genome [8], SBU Captions [11], and Conceptual Captions [13] for pretraining for fair comparisons with the previous works [2, 5, 9, 17]. Table 8 presents the statistics of the pretraining datasets such as the number of images and captions. Note that the size of the CC3M dataset is slightly different from other works [9, 17] since CC3M is webly-crowded and some of the download links are expired. As a result, the pretraining corpora consist of 4M images and 5.1M image-text pairs.

Table 8. Statistics of pretraining datasets

	MS-COCO	VG	SBU	CC3M
# Images	113K	100K	858K	2.89M
# Captions	567K	769K	858K	2.89M

2. Implementation Details

We employ ViT-B/16 [4] with 12 layers as our vision encoder and initialized it using the weights pretrained on ImageNet-1K [15]. The text encoder and multi-modal encoder are initialized by the first 6 layers and the last 6 layers of $\text{BERT}_{\text{base}}$ [3], respectively. We optimize our model using batch size of 1024 on 8 NVIDIA RTX A6000 GPUs for 30 epochs. We adopt AdamW optimizer with weight decay 0.02. The learning rate is initialized to 0.00002 and is warmed up to 2×10^{-4} for 1,000 iterations. After warm up, the learning rate is decayed to 2×10^{-5} following the cosine scheduling.

In order to generate soft masks, we compensate the total weights of masked regions to meet one-half of the total number of patches. By adopting this method, we can guarantee that the embedding of the image after being softly masked will not diverge too significantly from the original embedding, nor will it be excessively similar to it.

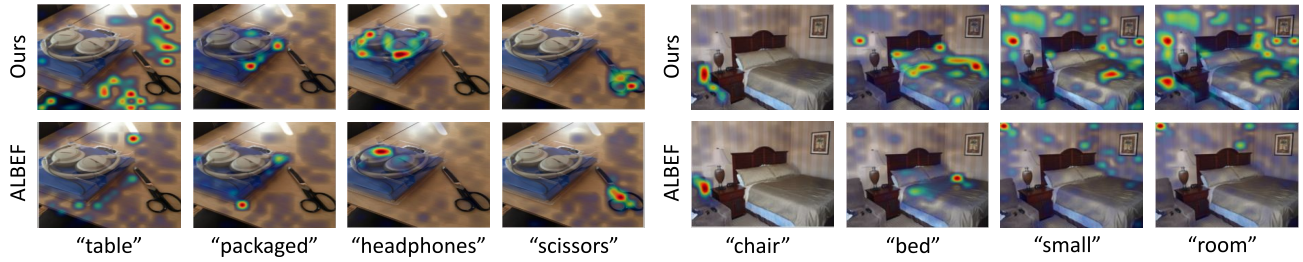
3. Details on Downstream Tasks

Below we provide implementation details of the vision-language downstream tasks to evaluate our pretraining approach, including Image-Text Retrieval (ITR), Visual Entailment (VE), Visual Question Answering (VQA), and Natural Language Visual Reasoning (NLVR). For all downstream tasks, we use AdamW optimizer, RandAugment, cosine learning rate scheduling, and a weight decay, whose hyperparameters are the same as the pretraining procedure.

Image-Text Retrieval (ITR) For the ITR task, we conduct our experiments on the Karpathy split [7] of the Flickr30K [12] and MS-COCO [10] datasets, which consist of 29k/1k/1k and 113k/5k/5k images for train/validation/test, respectively. We fine-tune the model using the ITM and ITC losses, using a batch size of 256, with a learning rate of 0.00001, for 10 epoch for Flickr30K and 5 epochs for MS-COCO. For inference, we first obtain top-k candidates based on similarity scores from the uni-modal encoders, as shown in (1) of the main paper, and then compute their ITM scores given by (4) of the main paper, to rank the candidates, where k is set to 128 for Flickr30K and 256 for MS-COCO.

Visual Entailment (VE) We use the SNLI-VE [16] dataset, which consists of 30k/1k/1k images for train/validation/test, respectively. The dataset is built upon the Flickr30K and Stanford Natural Language Inference (SNLI) [1] datasets. We optimize the pretrained model for 5 epochs using a batch size of 256 with a learning rate of 0.00002.

Natural Language Visual Reasoning (NLVR) We evaluate our model on the NLVR^2 [14] dataset, which consists of 86k/7k/7k examples for train/dev/test, respectively. Since the task requires a pair of images as input, we modify our model by duplicating the transformer block, as mentioned in Section 5.2. We first train the pretrained model using a 4M pretraining corpus for one more epoch to adjust



(a) On the table is packaged headphones and scissors.

(b) A chair and a bed in a small room.

Figure 4. Additional word-conditional Grad-CAM visualization of our method and ALBEF [9] after pre-training.

Table 9. Image-text retrieval results on MS-COCO with Grad-CAM (ours) and normalized cross-attention map. The bold-faced numbers indicate the best performance.

Method	TR@1	IR@1
Cross-Attention	75.68	59.48
Grad-CAM (Ours)	**76.62**	**60.15**

the modified model, using a batch size of 256 with a learning rate of 0.00002. We fine-tuned the adjusted model for 10 epochs, using a batch size of 128 with a learning rate of 0.00002.

Visual Question Answering (VQA) We conduct experiments on VQA2.0 dataset [6], which is based on the images collected from the MS-COCO dataset. The dataset consists of 83k/41k/81k images for train/validation/test, respectively. Following the previous works [9, 17], we utilize both the training and validation sets for training, and also include additional question-answer pairs from the Visual Genome dataset. Since each question in the VQA2.0 dataset is associated with 10 answers, we also weigh the loss for each answer based on its frequency among all answers, following [9]. We fine-tune the pretrained model for 8 epochs, using a batch size of 256, with a learning rate of 0.00002.

4. Design Choice for SoftMask

We conducted an additional experiment to compare Grad-CAM and normalized cross-attention maps to generate soft masks. Table 9 presents that utilizing Grad-CAM outperforms using normalized cross-attention maps. It shows that Grad-CAM provides more suitable guidance than cross-attention maps to generate the soft mask.

5. Additional Qualitative Results

In Figure 4, we show more visualizations of word-conditional Grad-CAM of our method and ALBEF [9] after pre-training. In general, our model provides Grad-CAM with more accurate and diverse attributes of the concepts than ALBEF, which is also presented in Figure 3 in the main paper. However, our model may learn bias towards the object and scene if the most discriminative part is masked with high weights, as we discussed in Section 5.7. For instance, in Figure 4 (b) our model fires on the wall for "bed", since bed usually comes with wall.

References

[1] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In *EMNLP*, 2015. 1

[2] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. In *ECCV*, 2020. 1

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 1

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[5] Jiali Duan, Liqun Chen, Son Tran, Jinyu Yang, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Multi-modal alignment using representation codebook. In *CVPR*, 2022. 1

[6] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 2

[7] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 1

[8] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017. 1

[9] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. 1, 2

[10] 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 *ECCV*, 2014. 1

[11] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. 1

[12] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 1

[13] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 1

[14] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *ACL*, 2019. 1

[15] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 1

[16] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019. 1

[17] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *CVPR*, 2022. 1, 2