LexLIP: Lexicon-Bottlenecked Language-Image Pre-Training for Large-Scale Image-Text Sparse Retrieval: Supplementary Material

Ziyang Luo[1]*, Pu Zhao[2], Can Xu[2], Xiubo Geng[2], Tao Shen[2], Chongyang Tao[2],
Jing Ma[1], Qingwei Lin[2], Daxin Jiang[2]†
[1] Hong Kong Baptist University, Hong Kong SAR, China
[2] Microsoft Corporation
cszyluo@comp.hkbu.edu.hk, majing@hkbu.edu.hk
{pu.zhao,caxu,xigeng,shentao,chongyang.tao,qlin,djiang}@microsoft.com

A. Implementation Details

Table 1 includes the hyperparameters of our models during Lexicon-Bottlenecked Pre-Training and Momentum Lexicon Contrastive Pre-Training. All training experiments are conducted on 8 V100 GPUs.

Dense. This baseline model utilizes dense CLS representations to represent images/texts, and undergoes a pre-training process similar to our LexLIP. Notably, the key distinction between the two models lies in the first bottlenecked pre-training phase, whereby the lexicon-bottlenecked module is omitted. The dense CLS representations are direct input into the masking-style text decoders.

CLIP. It is one of the most well-known SOTA dual-stream retrievers [3], which has achieved notable success. However, we recognize that the pre-training process for this model is resource-intensive, requiring the use of over 400M image-text pairs and 256-512 V100 GPUs, rendering it infeasible for our current study. To address this challenge, we have chosen to re-implement this model utilizing the same-scale pre-training data as our LexLIP. Additionally, to circumvent the need for a significant number of GPUs, we have adopted the momentum contrastive pre-training [1]. These enable us to leverage similar pre-training procedures with relatively fewer resources.

A.1. Small-Scale Retrieval

In this part, our LexLIP and Dense models are pre-trained with 4.3M and 14.3M image-text pairs. Our re-implemented CLIP is pre-trained with 4.3M pairs. Table 2 includes the hyperparameters of our models during fine-tuning (MSCOCO and Flickr30k). After fine-tuning, the checkpoints which have the best performance on the development set are evaluated on the test set.

*Work done during the internship at Microsoft.
†Corresponding author

A.2. Large-Scale Retrieval

In this part, our LexLIP, Dense model, and re-implemented CLIP are pre-trained with 4.2M image-text pairs. Notably, the training set of Flickr30k was excluded, resulting in 0.1M fewer pairs. We evaluate the zero-shot retrieval performance of these models on our Large-Scale Flickr30k test set without any additional fine-tuning. For our LexLIP model, we employ the approach outlined in Section 3.4 to convert all samples into high-dimensional sparse representations, subsequently transforming them into lexicons and frequencies (weights). We conduct the retrieval utilizing the term-based retrieval system Anserini [5]. Similarly, for the BM25 [4] baseline, retrieval is also conducted using Anserini. For the dense retrieval, we first convert all samples into dense vectors and subsequently conduct retrieval using the dense-vector retrieval system Faiss [2].

B. More Lexicon-Weighting Examples

In Figure 1, we introduce more Lexicon-Weighting examples of images and their corresponding captions. We can find that the major features of the images and texts are successfully captured by the lexicons.

References

[1] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9726–9735. Computer Vision Foundation / IEEE, 2020. 1

[2] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Trans. Big Data*, 7(3):535–547, 2021. 1

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya

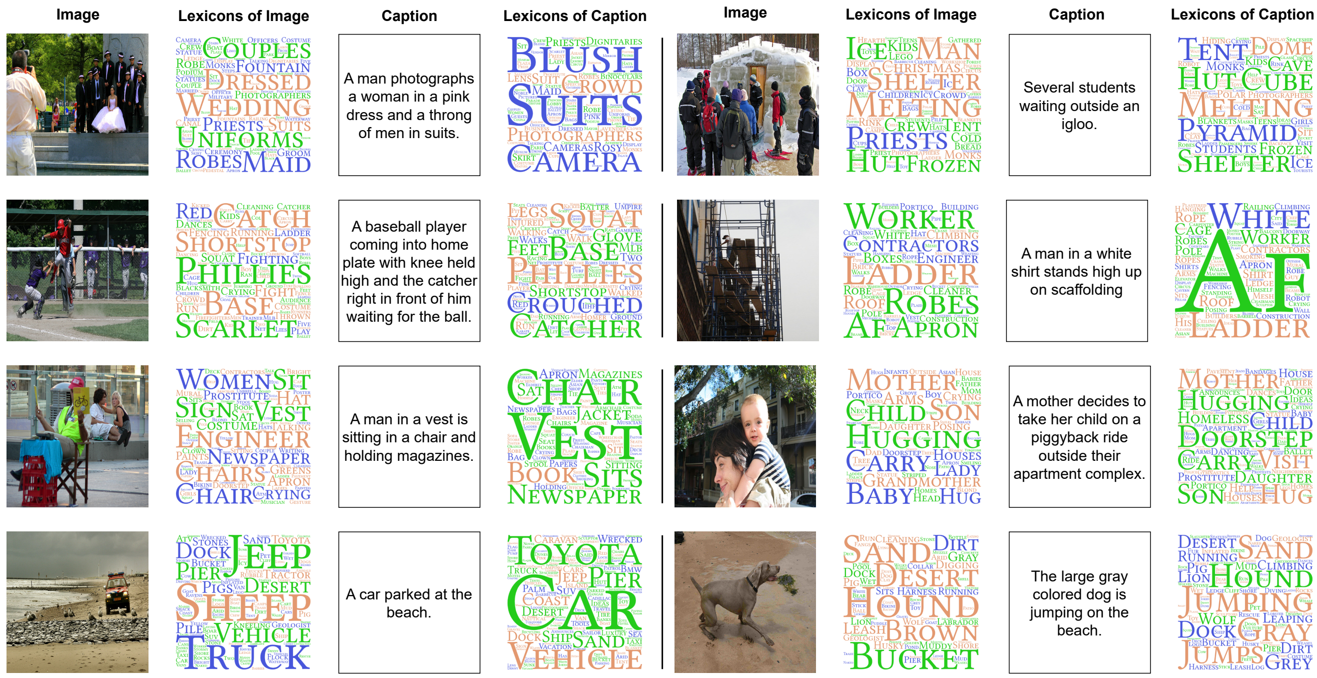

Figure 1: Visualizing the lexicons cloud of more images and their corresponding captions in the Flickr30k test set.

Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 1

[4] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. In Donna K. Harman, editor, *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST), 1994. 1

[5] Peilin Yang, Hui Fang, and Jimmy Lin. Anserini: Enabling the use of lucene for information retrieval research. In Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White, editors, *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 1253–1256. ACM, 2017. 1

Hyperparameters	LexLIP	Dense	CLIP
Lexicon-Bottlenecked Pre-Training			
Epochs	20	20	-
Batch Size	800	800	-
τ	0.05	0.05	-
λ	0.002	0.002	-
LR	5e-5	5e-5	-
LR Decay	Linear	Linear	-
Warmup Steps	10%	10%	-
Max Text Length	40	40	-
Weight Decay	0.01	0.01	-
Dropout Rate	0.1	0.1	-
Gradient Clip	1.0	1.0	-
Encoder Mask Rate	30%	30%	-
Decoder Mask Rate	50%	50%	-
Image Size	224	224	-
Patch Size	16	16	-
Momentum Lexicon Contrastive Pre-Training			
Epochs	10	10	30
Batch Size	2880	2880	2880
Queue Size	11520	11520	11520
τ	0.05	0.05	0.05
λ	0.002	0.002	0.002
m	0.99	0.99	0.99
LR	5e-5	5e-5	5e-5
LR Decay	Linear	Linear	Linear
Warmup Steps	10%	10%	10%
Max Text Length	40	40	40
Weight Decay	0.01	0.01	0.01
Dropout Rate	0.1	0.1	0.1
Gradient Clip	1.0	1.0	1.0
Image Size	224	224	224
Patch Size	16	16	16

Table 1: The hyperparameters of our models during pre-training.

Hyperparameters	Our Models
Fine-Tuning	
Epochs	10
Batch Size	1536
Queue Size	11520
τ	0.05
λ	0.002
m	0.99
LR	5e-5
LR Decay	Linear
Warmup Steps	10%
Max Text Length	40
Weight Decay	0.01
Dropout Rate	0.1
Gradient Clip	1.0
Image Size	384
Patch Size	16

Table 2: The hyperparameters of our models during fine-tuning.