Seeing Out of the Box:
End-to-End Pre-training for Vision-Language Representation Learning
Supplementary Material

Zhicheng Huang[1,2]*, Zhaoyang Zeng[3]*, Yupan Huang[3]*, Bei Liu[4], Dongmei Fu[1,2], Jianlong Fu[4]
[1]School of Automation and Electrical Engineering, University of Science and Technology Beijing
[2]Beijing Engineering Research Center of Industrial Spectrum Imaging
[3]Sun Yat-sen University, [4]Microsoft Research Asia

1. Dataset Statistics

Here we first summarize the detailed train/test image and text numbers of our pre-training and downstream datasets in Table 9. Then we provide a detailed comparisons of pre-training dataset usage of recent VLPT works in Table 10.

We follow UNITER [1] to classify pre-training datasets into two classes of "in-domain" and "out-of-domain". MSCOCO Captions (MSCOCO)[7] and **V**isual **G**enome Dense Captions (VG) [4] are typical in-domain datasets for many VL downstream tasks (e.g., image-text retrieval). In contrast, Conceptual Captions [10] and SBU Captions [9] are out-of-domain datasets which are noisier than in-domain datasets. We show the dataset usage of recent VLPT works in Table 10. For example, VisualBERT [6], LXMERT [12] and UNITER [1] pre-train with in-domain datasets. Among them, UNITER [1] additionally use out-of-domain data for model training. The ablation study of UNITER [1] shows that the additional usage of out-of-domain further improves performance.

In our work, we focus on in-domain datasets as they are commonly used in many VL tasks (e.g., image-text retrieval) and adopted by many VLPT works (e.g., VisualBERT [6], LXMERT [12] and UNITER [1]). When comparing with UNITER, we fairly compare with its in-domain pre-training results if they are provided. Otherwise, our "in-domain" dataset setting is inferior to the "in-domain+out-of-domain" pre-training setting of UNITER, and our results are not directly comparable.

We plan to include out-of-domain data in our pre-training data as a future work.

Table 9: Statistics of different datasets. Notation "*" denotes Karpathy split [3].

Dataset	Split		#Image (K)	#Text (K)
VG	train		105.9	472.7
COCO	train		82.8	414.1
	val	restval*	30.5	152.6
		val*	5.0	25.0
		test*	5.0	25.0
VQA2.0	train		82.8	443.8
	val		40.5	214.4
	test-dev		81.4	447.8
	test-std			
	test-challenge			
NLVR^2	train		103.2	86.4
	dev		8.2	7.0
	test-P		8.1	7.0
Flickr30K	train*		29.0	145.0
	val*		1.0	5.0
	test*		1.0	5.0
SNLI-VE	train		29.8	529.5
	val		1.0	17.9
	test		1.0	17.9

2. Implementation Details

We adopt two strategies to **speed up the training procedure**. First, we adopt mixed-precision training to reduce memory cost and speed up training procedure. Second, we re-organize the input data in one mini-batch Within a mini-batch, we only forward an image once to the visual backbone if it has multiple corresponding texts, while concatenating it with each text into cross-modal transformers. For example, an image will be paired with four texts in each batch during pre-training, including two positive pairs and two negative pairs. We only apply MLM and MVM on the positive image-text pairs.

[*] Equal Contribution. This work was performed when Zhicheng Huang, Zhaoyang Zeng and Yupan Huang were visiting Microsoft Research Asia as research interns.

Table 10: Statistics on the datasets used in recent vision-and-language pre-training works.

	In-domain		Out-of-domain	
Dataset	Visual Genome [4]	MSCOCO [7]	Conceptual Captions [10]	SBU [9]
Caption/Image Num	5,060K/101K	533K/106K	3,000K/3,000K	990K/990K
Unified VLP [13]			✓	
ViLBERT [8]			✓	
VLBERT [11]			✓	
Unicoder-VL [5]			✓	✓
VisualBERT [6]		✓		
LXMERT [12]	✓	✓		
UNITER [1]	✓	✓	✓	✓
Ours	✓	✓		

3. Visualization of Visual Dictionary

To show the semantic of visual dictionary (VD) items, we visualize the image patches that are grouped in each indices. We have shown two examples in the paper, and in the supplementary material, we randomly select ten more indices from the VD. From the visualization shown in Figure 4, we can find that each item in VD has meaningful and consistent semantics. In other words, our model is able to learn unified representations to represent different semantics of the image even though we do not have object bounding box annotations for supervision.

4. Discussion

For image-text retrieval task, the traditional approaches [2] first project an image and a text to a common representation space and then correlate their representations by late fusion. For example, the widely-used late fusion method is calculating cosine similarity based on a dot-product operation, which is simple and fast. In contrast, Transformer-based approaches early fuse the image and text by a multi-layer Transformer to get an united representation. The unified representation captures the deep relation between an image and a text with self-attention mechanism, thus is able to achieve a better result than the late fusion representation. However, the early fusion Transformer-based approaches cannot produce separate representation for images and texts, thus suffers from slow speed due to exhaustive computation of each possible image-text combination. Our model as well as other vision-language pre-training models are based on Transformers, and the inference speed has become a bottleneck for applying these models to real-world search engines. For future works, we are curious about how we could speedup the Transformer-based approaches in image-text retrieval task.

References

[1] 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 *ECCV*, pages 104–120, 2020. 1, 2

[2] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2017. 2

[3] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015. 1

[4] 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, 2

[5] Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, pages 11336–11344, 2019. 2

[6] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A simple and performant baseline for vision and language. *arXiv:1908.03557*, 2019. 1, 2

[7] 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*, pages 740–755, 2014. 1, 2

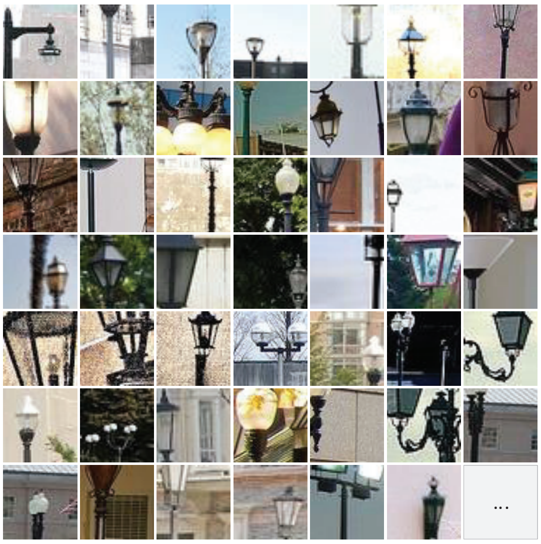
[8] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, pages 13–23, 2019. 2

[9] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. In *NeurIPS*, pages 1143–1151, 2011. 1, 2

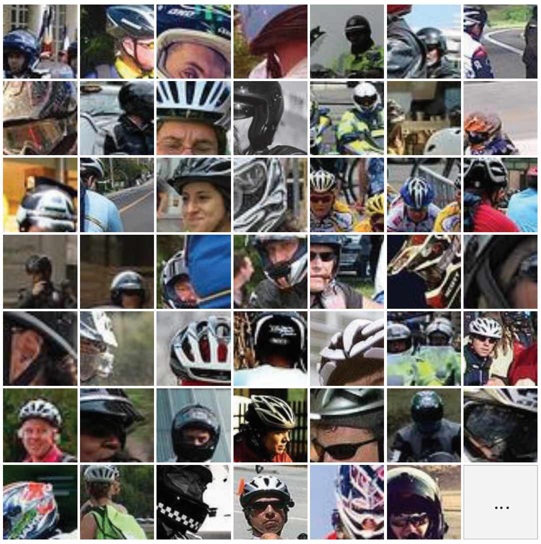
[10] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565, 2018. 1, 2

[11] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-training of generic visual-linguistic representations. In *ICLR*, 2019. 2

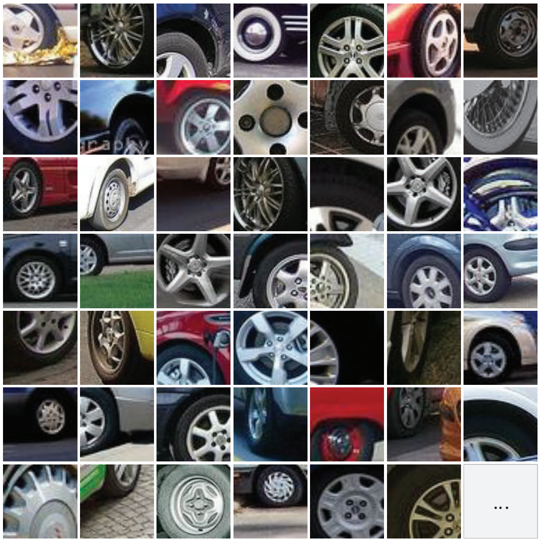
[12] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, pages 5103–5114, 2019. 1, 2



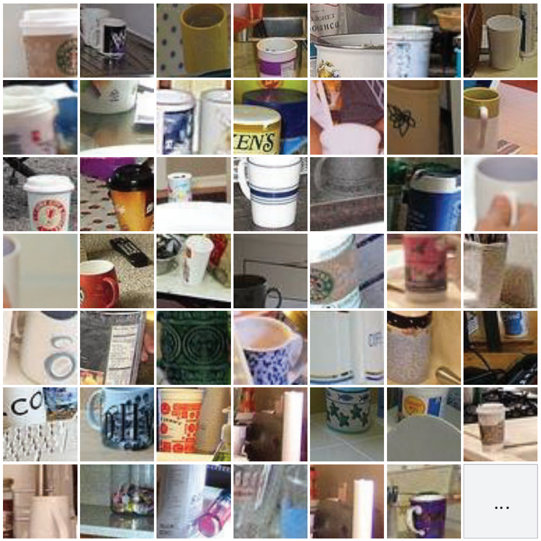
Id=74



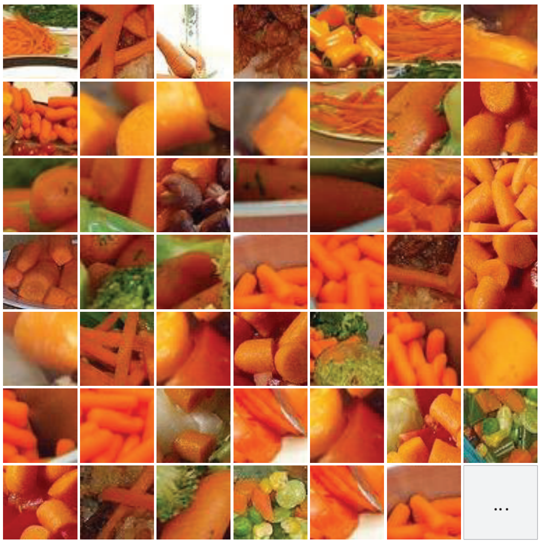
Id=183



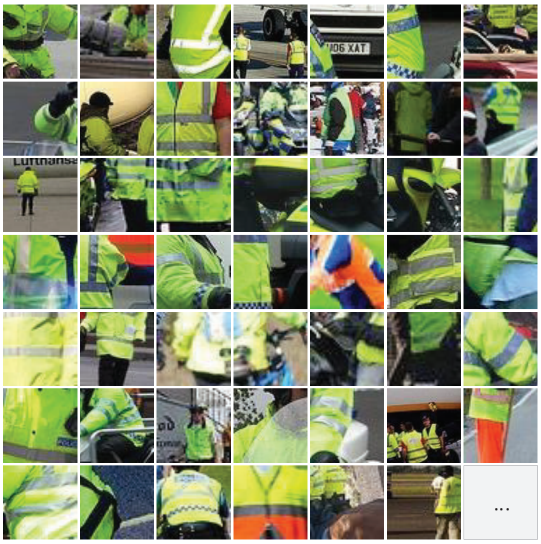
Id=229



Id=731



Id=1162



Id=1237

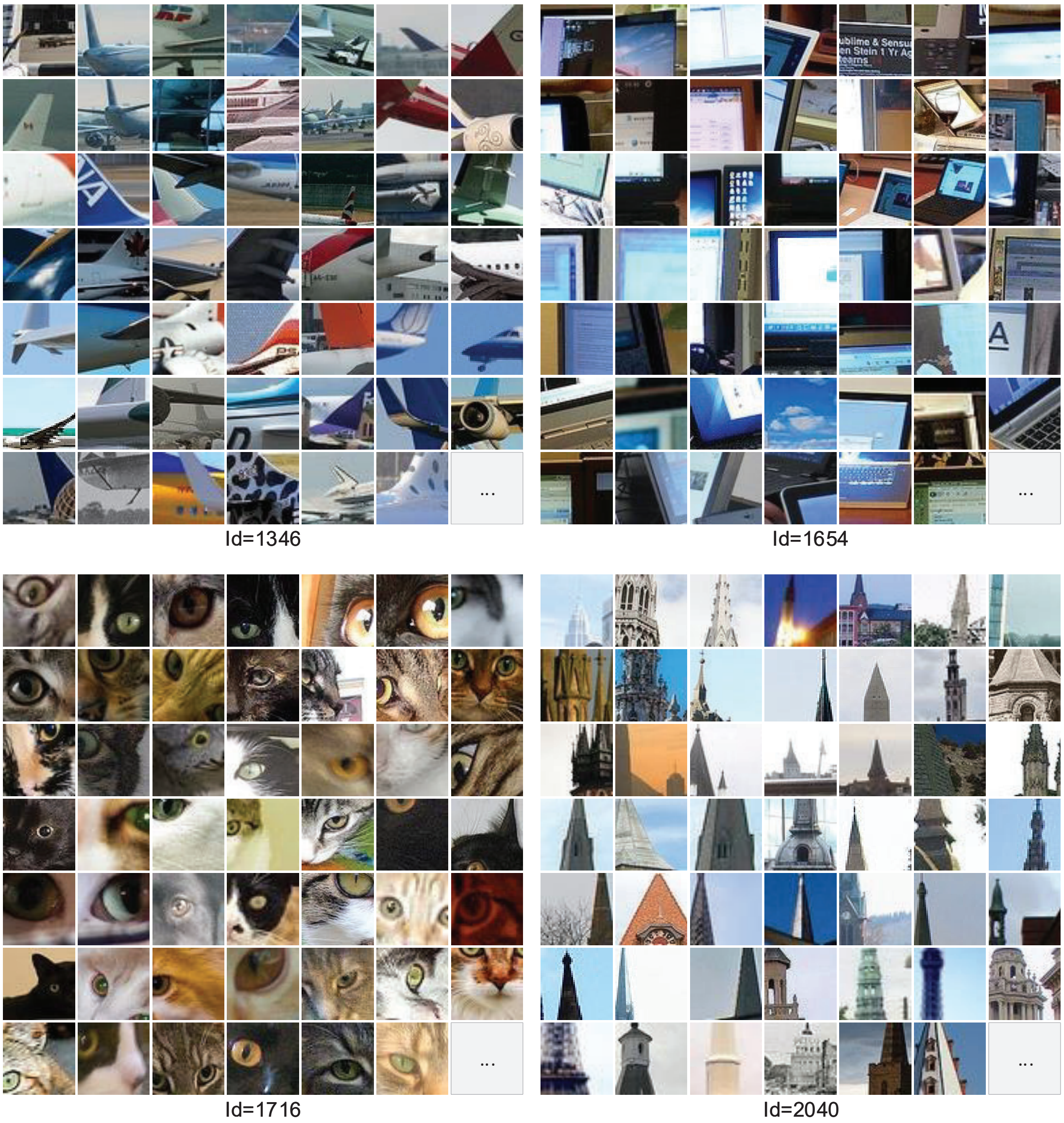


Figure 4: Visualization of visual dictionary (VD) we have learned by SOHO. Apart from the two indices we have shown in the paper, we randomly select another ten indices in the visual dictionary to present in this supplementary material. From the above results we can find that, our visual dictionary is learned to group meaningful and consistent semantics of image patches into different indices. Thus, each index can reflect an abstraction of visual semantics. [Best viewed in color.]

[13] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. In *AAAI*, pages 13041–13049, 2020. 2