A Fine-tuning Settings

Image-Text Retrieval We adopt the widely used Karpathy split [8] on the COCO caption dataset [11] to conduct our experiments. Specifically, the dataset consists of $113,287$ images for training, $5,000$ images for validation, and $5,000$ images for testing. Each image is associated with 5 human-generated captions. For the Oscar$_\mathrm{B}$ model, we fine-tune with a batch size of 256 for 40 epochs. The initial learning rate is set to $2e^{-5}$ and linearly decreases. For the Oscar$_\mathrm{L}$ model, we fine-tune with a batch size of 128 for 40 epochs. The initial learning rate is set to $1e^{-5}$ and linearly decreases. We use the validation set for parameter tuning. We compare with several existing methods, including DVSA [8], VSE++ [5], DPC [19], CAMP [18], SCAN [9], SCG [15], PFAN [17], Unicoder-VL [10], 12-in-1 [12], UNITER [4].

Image Captioning Though the training objective (*i.e.,* seq2seq) for image captioning is different from that used in pre-training (*i.e.,* bidirectional attention-based mask token loss), we directly fine-tune Oscar for image captioning on COCO without additional pre-training on Conceptual Captions [14]. This is to validate the generalization ability of the Oscar models for generation tasks. We use the same Karpathy split [8]. During training, we randomly select 15% of caption tokens with a maximum of 3 tokens per caption to be masked out. For the Oscar$_\mathrm{B}$ model, we fine-tune with cross-entropy loss for 40 epochs with a batch size of 256 and an initial learning rate of $3e^{-5}$ and then with CIDEr optimization [13] for 5 epochs with a batch size of 64 and initial learning rate of $1e^{-6}$. For the Oscar$_\mathrm{L}$ model, we fine-tune for 30 epochs with a batch size of 128 and an initial learning rate of $1e^{-5}$ and then with CIDEr optimization for another 3 epochs with a batch size of 48 and learning rate of $\{1e^{-6}, 5e^{-7}\}$. We compare with several existing methods, including BUTD [2], VLP [20], AoANet [6].

NoCaps Since NoCaps images are collected from Open Images. We train an object detector using the Open Images training set and applied it to generate the tags. We conduct experiments from BERT model directly without pre-training as required by the task guidelines. For the Oscar$_\mathrm{B}$ model, we train 40 epoch with a batch size of 256 and learning rate $3e^{-5}$; further we perform CIDEr optimization with learning rate $1e^{-6}$ and batch size 64 for 5 epochs. During inference, we use constrained beam search for decoding. We compare Oscar with UpDown [1] on this task.

VQA For VQA training, we random sample a set of 2k images from the MS COCO validation set as our validation set, the rest of images in the training and validation are used in the VQA finetuning. For the Oscar$_\mathrm{B}$ model, we fine-tune for 25 epochs with a learning rate of $5e^{-5}$ and a batch size of 128. For the Oscar$_\mathrm{L}$ model, we fine-tune for 25 epochs with with a learning rate of $3e^{-5}$ and a batch size of 96.

GQA The fine-tuning procedure of GQA is similar to that of VQA. For the Oscar$_\mathrm{B}$ model, we fine-tune for 5 epochs with a learning rate of $5e^{-5}$ and a batch size of 128. We compare with four existing methods, including LXMERT [16], MMN [3], 12-in-1 [12], NSM [7].

NLVR2 For the Oscar$_{\text{B}}$ model, we fine-tune for 20 epochs with learning rate $\{2e^{-5}, 3e^{-5}, 5e^{-5}\}$ and a batch size of 72. For the Oscar$_{\text{L}}$ model, we fine-tune for 20 epochs with learning rate of $\{2e^{-5}, 3e^{-5}\}$ and a batch size of 48.

B Pre-training Corpus

Table 1 shows the statistics of image and text of the corpus.

Table 1: Statistics of the pre-training corpus.

Source	COCO (train)	CC (all)	SBU (all)	Flicker30k (train)	VQA (train)	GQA (bal-train)	VG-QA (train)	Total
Image/Text	112k/560k	3.0M/3.0M	840k/840k	29k/145k	83k/444k	79k/1026k	48k/484k	4.1M/6.5M

C More Results

The enlarged *t*-SNE visualization results of Oscar and baseline (no tags) are shown in Fig. 1 and Fig. 2, respectively.

Acknowledgement

We thank Yonatan Bisk, Hannaneh Hajishirzi, Xiaodong Liu, Sachin Mehta, Hamid Palangi and Arun Sacheti, Rowan Zellers for valuable discussions and comments, and the Microsoft Research Technical Support team for managing the GPU clusters.

References

1. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: nocaps: novel object captioning at scale. In: ICCV (2019)
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR (2018)
3. Chen, W., Gan, Z., Li, L., Cheng, Y., Wang, W., Liu, J.: Meta module network for compositional visual reasoning. arXiv preprint arXiv:1910.03230 (2019)
4. Chen, Y.C., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Learning universal image-text representations. arXiv preprint arXiv:1909.11740 (2019)
5. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improved visual-semantic embeddings. arXiv preprint arXiv:1707.05612 **2**(7), 8 (2017)

6. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on attention for image captioning. In: ICCV (2019)
7. Hudson, D., Manning, C.D.: Learning by abstraction: The neural state machine. In: NeurIPS (2019)
8. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR (2015)
9. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: ECCV (2018)
10. Li, G., Duan, N., Fang, Y., Jiang, D., Zhou, M.: Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. arXiv preprint arXiv:1908.06066 (2019)
11. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV (2014)
12. Lu, J., Goswami, V., Rohrbach, M., Parikh, D., Lee, S.: 12-in-1: Multi-Task vision and language representation learning. arXiv preprint arXiv:1912.02315 (2019)
13. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: CVPR (2017)
14. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Annual Meeting of the Association for Computational Linguistics (2018)
15. Shi, B., Ji, L., Lu, P., Niu, Z., Duan, N.: Knowledge aware semantic concept expansion for image-text matching. In: IJCAI (2019)
16. Tan, H., Bansal, M.: LXMERT: Learning cross-modality encoder representations from transformers. EMNLP (2019)
17. Wang, Y., Yang, H., Qian, X., Ma, L., Lu, J., Li, B., Fan, X.: Position focused attention network for image-text matching. arXiv preprint arXiv:1907.09748 (2019)
18. Wang, Z., Liu, X., Li, H., Sheng, L., Yan, J., Wang, X., Shao, J.: CAMP: Cross-Modal adaptive message passing for text-image retrieval. In: ICCV (2019)
19. Zheng, Z., Zheng, L., Garrett, M., Yang, Y., Shen, Y.D.: Dual-path convolutional image-text embedding with instance loss. arXiv preprint arXiv:1711.05535 (2017)
20. Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J.J., Gao, J.: Unified vision-language pre-training for image captioning and VQA. AAAI (2020)

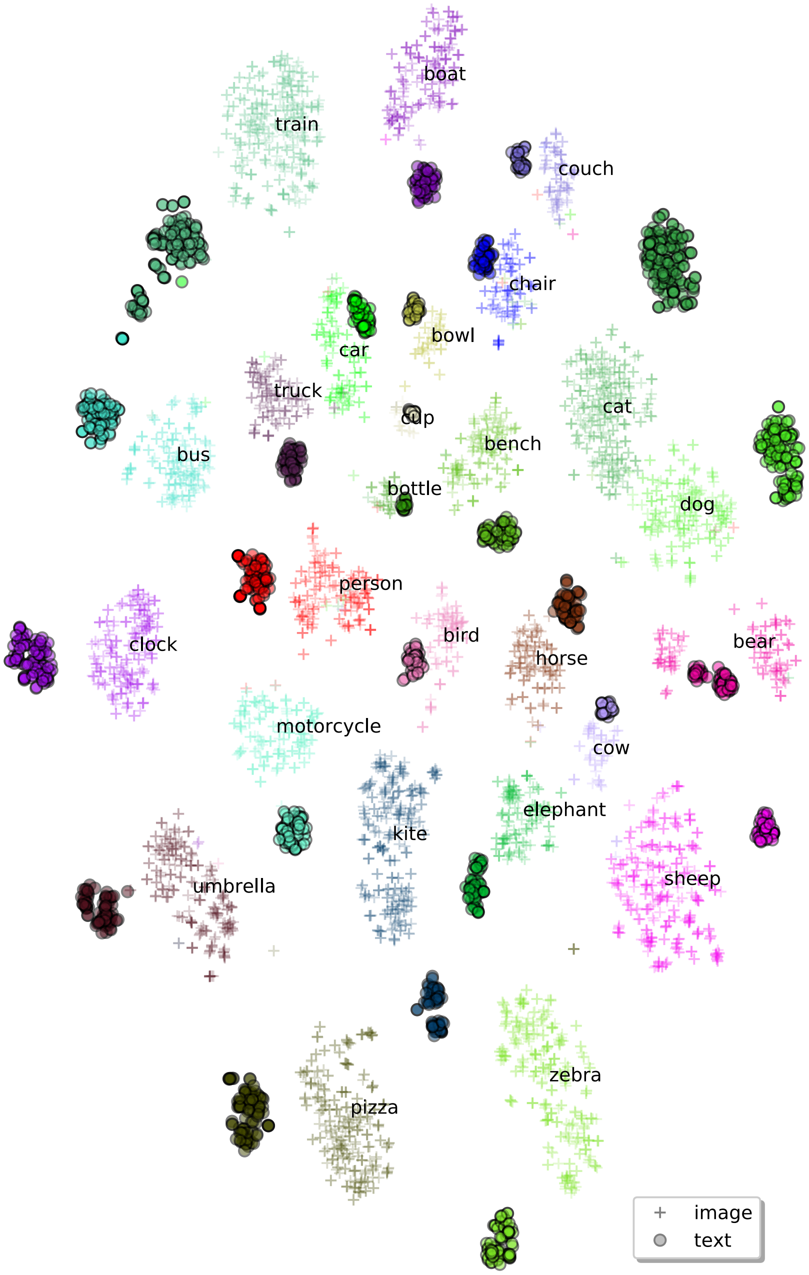

Fig. 1: Feature visualization of Oscar. We observe small distances between text and image features of the same object; some of them are perfectly aligned, as demonstrated by the overlapping regions.

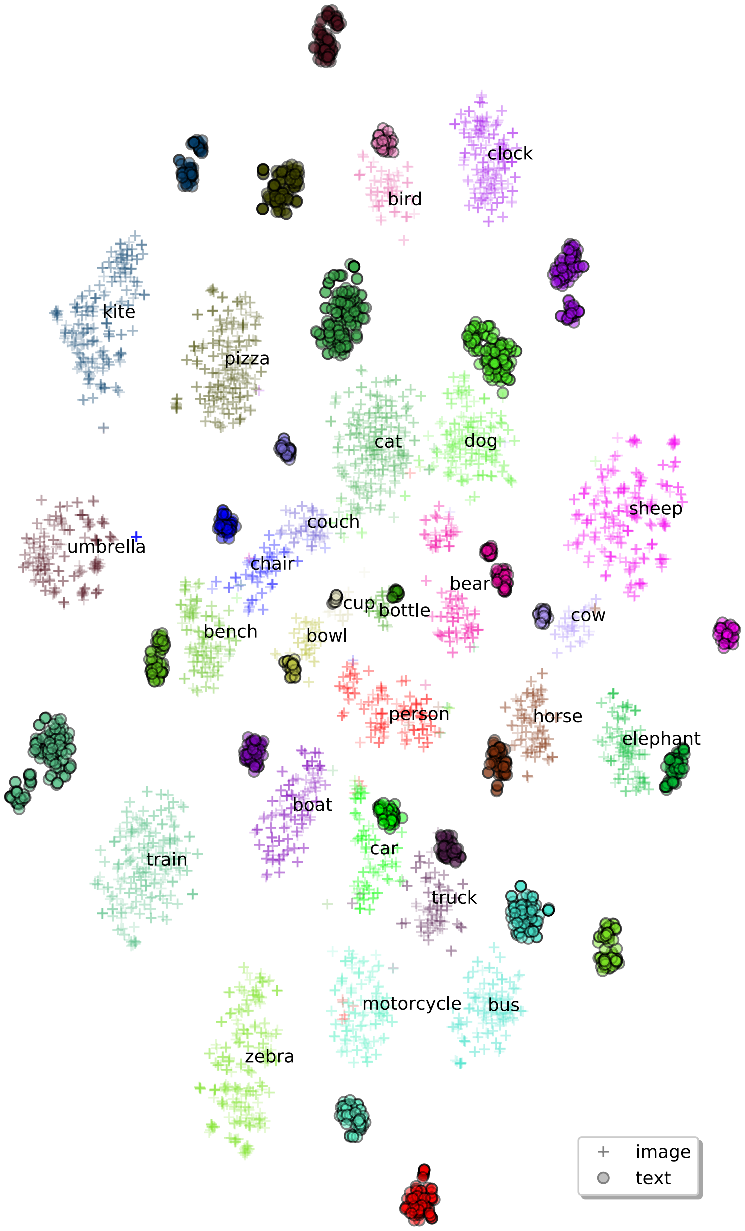

Fig. 2: Feature visualization of baseline (no tags). For several object classes, their text and image features are largely separated (*e.g.,* person, umbrella, zebra). The distance of image features between some objects is too small (*e.g.,* bench, chair, couch).