LocTex: Learning Data-Efficient Visual Representations from Localized Textual Supervision

Zhijian Liu
MIT

Simon Stent, Jie Li, John Gideon
Toyota Research Institute

Song Han
MIT

https://loctex.mit.edu/

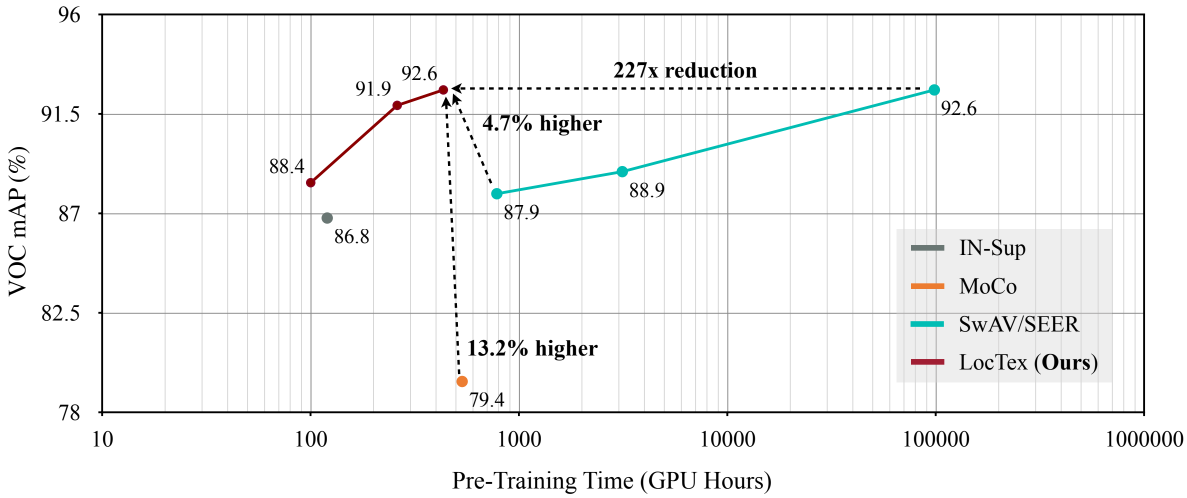
A.1. Annotation Cost

We provide quantitative comparisons between various forms of annotations in Table A1. Here, all annotation costs are estimated on the 118K training images of COCO. Compared with classification and segmentation annotations, localized narratives are cheaper (lower cost) and offer richer information (higher accuracy). It is worth noticing that the annotation cost of localized narratives is dominated by manual transcription. Thus, its cost can be further reduced by **3.6**× with an accurate automatic speech recognition system. Annotating over larger sets of classes can be even more challenging since memorizing and learning to distinguish over a large class hierarchy (*e.g.*, 1000 classes for ImageNet) is very costly.

	Annotations	Cost (hours)	mAP
Multi-label classification	Multi-class labels	11.1K [5, 2]	86.2
Instance segmentation	Segmentation masks	30.0K [5, 2]	82.3
LocTex (Ours)	Localized narratives	**4.7K** [6]	**88.4**

Table A1: Comparison with different forms of annotations.

A.2. Training Efficiency



Most pre-training efforts focus on improving performance and data efficiency. However, some of them suffer from extremely long training time. We conduct a thorough analysis on the training efficiency across the following pre-training methods:

- **LocTex**. We include three variants of our LocTex. One of them is trained only with COCO images for 600 epochs. The remaining two are trained on both COCO and Open Images data, one for 300 epochs and the other one for 500 epochs.
- **SwAV/SEER** [1, 3]. We include three variants of SwAV as well. Two of them are trained with ImageNet data for 200 and 800 epochs, respectively. The other one is trained with 1 billion uncurated Instagram images (for one epoch).

		50% Training Data						100% Training Data					
	# Pretrain Images	AP^{bbox}	AP^{bbox}_{50}	AP^{bbox}_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{bbox}	AP^{bbox}_{50}	AP^{bbox}_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}
Random Init	–	28.4	46.1	29.9	25.7	43.2	26.8	36.1	55.0	38.9	31.8	52.0	33.9
IN-Sup (10%)	128K	31.2	50.0	32.7	27.9	46.8	29.2	37.7	56.9	40.6	33.0	53.4	35.3
VirTex [2]	118K	35.5	54.8	37.9	31.1	51.6	32.7	39.8	59.4	42.6	34.6	56.1	36.7
LocTex (Ours)	118K	**36.1**	**55.7**	**38.6**	**31.6**	**52.4**	**33.1**	**40.6**	**60.6**	**44.1**	**35.2**	**57.0**	**37.4**
IN-Sup (50%)	640K	34.9	54.4	37.0	30.8	51.0	32.5	39.7	59.4	43.3	34.6	56.2	36.8
VirTex [2]	118K(×5)	36.6	55.9	39.3	31.9	52.6	33.6	40.8	60.5	44.2	35.2	57.0	37.6
LocTex (Ours)	809K	**37.5**	**57.2**	**40.4**	**32.7**	**54.0**	**34.7**	**41.4**	**61.3**	**44.9**	**35.8**	**57.7**	**38.4**
IN-Sup (100%)	1.28M	36.3	56.1	38.8	31.9	52.7	33.7	40.2	60.0	43.5	35.0	56.4	37.4

Table A2: Additional results of instance segmentation on COCO under 50% and 100% data settings.

- **MoCo** [4]. We include the baseline MoCo self-supervised pre-training on ImageNet for 200 epochs.
- **IN-Sup**. We follow the standard ImageNet supervised pre-training (as in `torchvision`) for 90 epochs.

The training time is measured on 8 NVIDIA V100 GPUs for all pre-training methods except SwAV/SEER. For SwAV/SEER, we directly adopt the statistics from their official GitHub repository* and paper [3]: the two trained with ImageNet data use 32 NVIDIA V100 GPUs, and the scaled-up one uses 512 NVIDIA V100 GPUs.

Results. Our LocTex pre-training is more efficient than ImageNet supervised pre-training while achieving more than 1% higher accuracy. Compared with the self-supervised pre-training, the improvement is more significant: *i.e.*, we achieve the same linear classification accuracy (92.6) as the scaled-up SwAV with **227**× less training time. In terms of data efficiency, the scaled-up SwAV requires 1 billion unlabeled images from Instagram while our LocTex makes use of 809K images with low-cost localized textual annotations. This suggests that supervised pre-training can be much more computationally efficient, and its annotation cost is also affordable if the form of annotation is chosen carefully (which is discussed in the main paper).

A.3. Additional Results on COCO Instance Segmentation

In Table A2, we provide additional results of instance segmentation on COCO under 50% and 100% data settings. The experimental setup is exactly the same as in the main paper where we scale the training schedule linearly with the dataset size.

Results. The overall trend is the same as the one under 10% and 20% settings (which is presented in the main paper). With the same amount of labelled images, our LocTex always achieves the highest performance compared with ImageNet supervised pre-training and VirTex pre-training methods. It achieves more than 1% higher (box or mask) AP than the full ImageNet supervised pre-training baseline while using only half of the annotated images. Under the 100% data setting, our LocTex is able to push the instance segmentation performance from 40.2% to 41.4% in box AP and from 35.0% to 35.8% in mask AP.

A.4. Additional Visualizations of Learned Image-Caption Attention Maps

In Figure A1, we provide additional visualizations of learned image-caption attention maps on COCO. Note that these visualizations are picked randomly from COCO `val2017`. The only constraint we apply is to ensure that there are at least six entities in the image for visualization purposes. We observe that the learned attention map is able to localize the instances fairly accurately, even for some small instances (*e.g.*, cap in the second example), which is especially useful for the downstream object detection and instance segmentation tasks.

References

[1] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *NeurIPS*, 2020.

[2] Karan Desai and Justin Johnson. VirTex: Learning Visual Representations from Textual Annotations. In *CVPR*, 2021.

*https://github.com/facebookresearch/swav

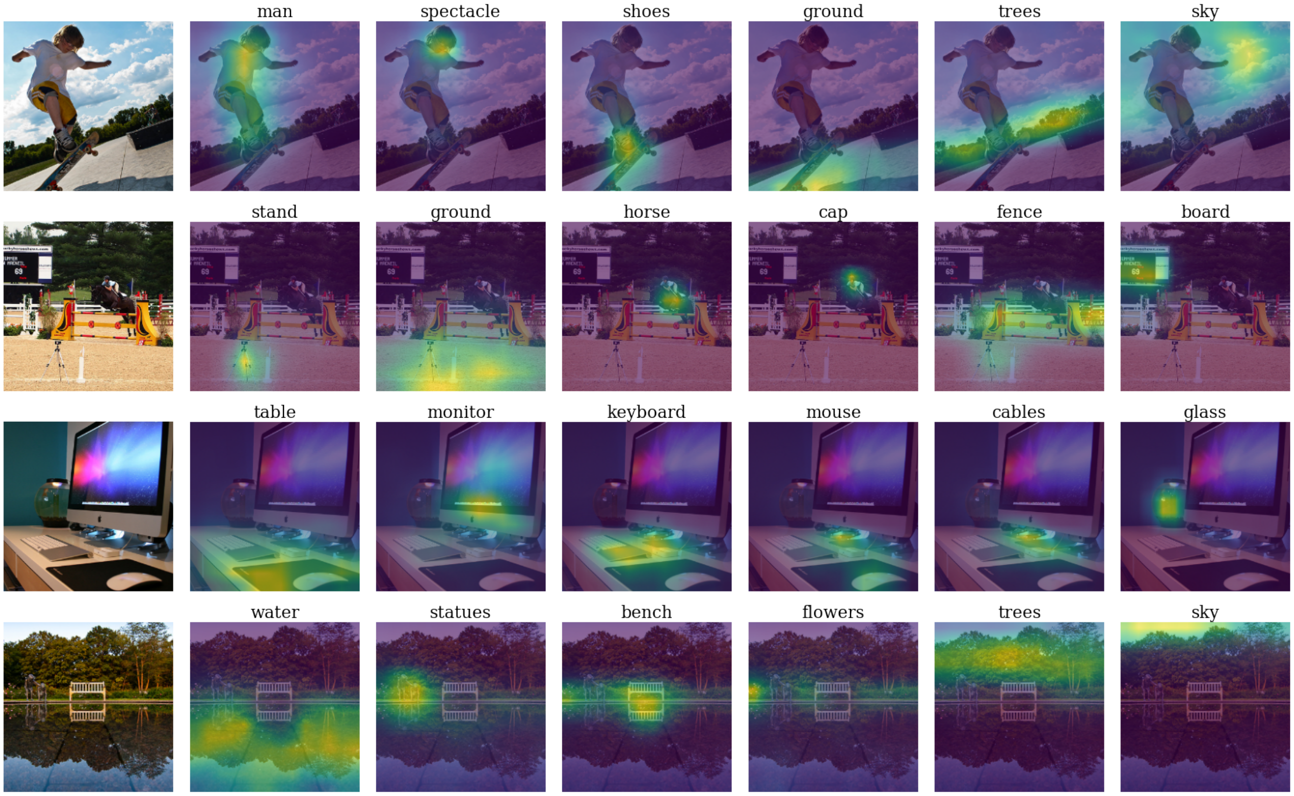


Figure A1: Additional visualization of learned image-caption attention maps (on COCO `val2017`).

[3] Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, and Piotr Bojanowski. Self-supervised Pretraining of Visual Features in the Wild. *arXiv*, 2021.

[4] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *CVPR*, 2020.

[5] 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.

[6] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting Vision and Language with Localized Narratives. In *ECCV*, 2020.