Supplementary Material for "Billion-Scale Pretraining with Vision Transformers for Multi-Task Visual Representations"

Josh Beal Hao-Yu Wu Dong Huk Park Andrew Zhai Dmitry Kislyuk
Pinterest
{jbeal, rexwu, dhukpark, andrew, dkislyuk}@pinterest.com

1. Implementation Details

Pretraining is implemented using PyTorch [3] on 8 p3dn.24xlarge Amazon EC2 instances with a total of 64 Tesla V100 GPUs, while the fine-tuning uses PyTorch on a single p3dn.24xlarge Amazon EC2 instance with 8 Tesla V100 GPUs. We use DistributedDataParallel for multi-GPU training. We use automatic mixed precision for all of our experiments, and channels-last memory format for the ResNeXt experiments, in order to improve the training throughput. All of our model training runs and performance benchmarks use PyTorch 1.7.1, CUDA 11.0, and cuDNN 8.

Vision Transformer pretraining uses a warmup phase of 10k steps, total batch size of 8192, base learning rate (LR) of 8e-4, and linear decay LR schedule of 2 epochs in length, such that around 2.6B images are processed during the main phase of training. We train using the AdamW [2] optimizer with a weight decay value of 0.05. Vision Transformer Unified Visual Embedding fine-tuning uses a warmup phase of 5k steps, base LR of 0.24, and cosine decay LR schedule of 20 epochs in length. We fine-tune using the SGD optimizer with a base LR of 0.24 and weight decay of 1e-4 for the non-sparse parameters. Vision Transformer ImageNet fine-tuning uses a warmup phase of 5k steps, base LR of 0.03, cosine decay LR schedule of 50k steps, SGD optimizer, and zero weight decay.

ResNeXt-101 pretraining uses a warmup phase of 15k steps, total batch size of 12288, base learning rate of 0.03, and step LR schedule of 20 steps and $\gamma = 0.5$. We train using the LARS [6] optimizer with a weight decay value of 1e-4. The hyperparameters of the ResNeXt-101 Unified Visual Embedding fine-tuning are largely the same as [4], except the base learning rate is 0.03.

For pretraining, we use the Inception [5] random crop strategy, whereas for Unified Visual Embedding fine-tuning we apply horizontal mirroring, random crops, and color jitter to the resized images. For ImageNet fine-tuning we directly apply the data augmentation strategy that is specified in the original work on Vision Transformers [1].

For ablations on the sample count of the pretraining dataset, we linearly interpolate the training schedule length between the minimum and maximum value, i.e., 100 epochs on the 13M dataset and 2 epochs on the 1.3B dataset.

References

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

[2] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018.

[3] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

[4] Raymond Shiau, Hao-Yu Wu, Eric Kim, Yue Li Du, Anqi Guo, Zhiyuan Zhang, Eileen Li, Kunlong Gu, Charles Rosenberg, and Andrew Zhai. Shop the look: Building a large scale visual shopping system at pinterest. In *KDD*, 2020.

[5] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[6] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.