Differentially Private Video Activity Recognition
Supplementary Material

1. The State-of-the-Art in Differential Privacy

Although great strides have been made in developing training methods for differential privacy, most of the prior studies focused on small datasets with shallow neural networks and large-scale differential privacy methods remain under-explored (see Table 1 in the main manuscript). In the few papers that investigate large-scale models and datasets [3, 4], transfer learning [1, 2, 5] is used as the training paradigm. Due to the noise added to the gradient, it is difficult to learn useful feature representations from scratch, especially for large vision models with a large number of parameters [6]. Here, we aim to make differentially-privacy training more practicable at scale using transfer learning by taking the following into consideration. Our primary objective is to devise models that can provide a good privacy-utility trade-off in large-scale and privacy-sensitive vision tasks. Additionally, the training and deployment of our model should be accomplished with reasonable resources, including pre-trained datasets, training time, computational power, and memory requirements.

2. Comparison with Image-Based Approaches

The performance of the image-based method on UCF-101 is shown below. Note that our paper has also included direct comparisons to the Multiscale Vision Transformers (MViT) model [17] under two commonly used training paradigms: training from scratch and full fine-tuning.

Scheme	#Trainable Params.	Pre-train	#Clips	Top-1 Accuracy
From scratch	21.7M	-	1	12.12
Full fine-tune	21.7M	IN-1k	1	61.62
Adapter	929K	IN-1k	8	69.697
Selective fine-tune	58.1K	IN-1k	8	**81.818**

Table 1. **Experiments on UCF-101 using a ViT-S/16-224 model pre-trained on ImageNet.** We adopt $\delta = 10^{-5}$ and $\epsilon = 5$.

3. Hyperparameters

It is well known that the hyperparameters play a critical role in DP training yet they are difficult to tune. In order to facilitate comparison between methods, we pre-define the training epochs for each dataset. And we fix the clipping norm as $C = 1$. We only search for the optimal learning rate on the CIFAR-100 dataset with a fixed $\epsilon = 1$. We perform a grid search over the learning rate between $[10^{-4}, 10^{-2}]$. We directly apply these tuned parameters to other tasks. The complete list of hyperparameters used are detailed in Table 2.

Dataset	Architecture	Batch size	#Epochs	Scheme	Learning rate
CIFAR-10	ViT-S/16-224	1,024	10	Full fine-tune	3e-4
				Adapter	1e-3
CIFAR-100	ViT-S/16-224	1,024	50	From scratch	3e-3
				Full fine-tune	3e-4
				Linear probe	1e-3
				Sparse fine-tune	1e-3
				Adapter	1e-3
CIFAR-100	ConvNeXt-T	1,024	50	From scratch	5e-3
				Full fine-tune	7e-4
				Linear probe	1e-3
				Sparse fine-tune	1e-3
				Adapter	1e-3
CIFAR-100	ResNet-50-GN	1,024	50	From scratch	3e-3
				Full fine-tune	7e-4
				Linear probe	5e-3
				Sparse fine-tune	1e-3
ImageNet	ViT-S/16-224	1,024	90	Full fine-tune	1e-4
				Linear probe	5e-4
				Sparse fine-tune	5e-4
				Adapter	5e-4
ImageNet	ViT-S/16-224	65,536	70	Full fine-tune	3e-4
				Sparse fine-tune	5e-3
				Adapter	5e-3
ImageNet	ViT-B/16-224	65,536	70	Full fine-tune	1e-4
				Sparse fine-tune	5e-3
				Linear probe	5e-3
CheXpert	ViT-B/16-224	256	10	From scratch	1e-4
				Full fine-tune	3e-4
				Sparse fine-tune	1e-3
				Adapter	1e-3
CheXpert	ConvNeXt-T	256	10	From scratch	3e-3
				Full fine-tune	3e-4
				Sparse fine-tune	1e-3
				Adapter	3e-3
UCF-101	MViT-B/16×4	16	10	Full fine-tune	3e-4
				Sparse fine-tune	3e-4
				Adapter	3e-4
HMDB-51	MViT-B/16×4	16	10	Full fine-tune	7e-4
				Sparse fine-tune	4e-3
				Adapter	9e-4

Table 2. **Hyperparameters.** We include the batch size, number of epochs, and learning rate for each setting.

References

[1] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655. PMLR, 2014. 1

[2] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016. 1

[3] Alexey Kurakin, Steve Chien, Shuang Song, Roxana Geambasu, Andreas Terzis, and Abhradeep Thakurta. Toward training at imagenet scale with differential privacy. *arXiv preprint arXiv:2201.12328*, 2022. 1

[4] Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. *arXiv preprint arXiv:2110.05679*, 2021. 1

[5] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009. 1

[6] Florian Tramer and Dan Boneh. Differentially private learning needs better features (or much more data). *arXiv preprint arXiv:2011.11660*, 2020. 1