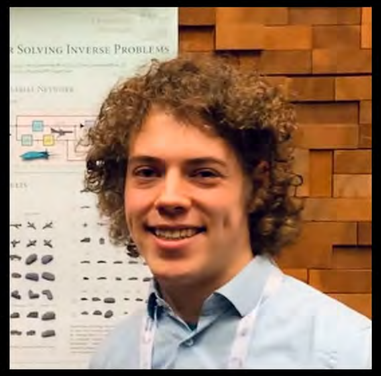
# Differentiable Top-$k$ Classification Learning


Felix Petersen

Hilde Kuehne

Christian Borgelt

Oliver Deussen


ICML
International Conference
On Machine Learning

difftopk

# Differentiable Top-$k$ Classification Learning

# Differentiable Top-$k$ Classification Learning

difftopk

# Differentiable Top-$k$ Classification Learning

difftopk

CNN

$$\begin{bmatrix} 1 \\ 8 \\ -1 \\ 9 \\ 5 \end{bmatrix}$$

# Differentiable Top-$k$ Classification Learning

difftopk

CNN

$$\begin{bmatrix} 1 \\ 8 \\ -1 \\ 9 \\ 5 \end{bmatrix}$$

Diff. Rank

# Differentiable Top-$k$ Classification Learning

difftopk

CNN

$$\begin{bmatrix} 1 \\ 8 \\ -1 \\ 9 \\ 5 \end{bmatrix}$$

Diff. Rank

$$\mathbf{P} \quad \begin{bmatrix} & \\ & \\ & \\ & \\ & \end{bmatrix}$$

# Differentiable Top-$k$ Classification Learning

difftopk

CNN

$$\begin{bmatrix} 1 \\ 8 \\ -1 \\ 9 \\ 5 \end{bmatrix}$$

Diff. Rank

$$\begin{matrix} \text{rank 1} \\ \text{rank 2} \\ \text{rank 3} \\ \text{rank 4} \\ \text{rank n} \end{matrix}$$

$$\mathbf{P} \begin{bmatrix} \\ \\ \\ \\ \\ \end{bmatrix}$$

# Differentiable Top-$k$ Classification Learning

difftopk

CNN

$$\begin{bmatrix} 1 \\ 8 \\ -1 \\ 9 \\ 5 \end{bmatrix}$$

Diff. Rank

$$\begin{bmatrix} \text{rank 1} \\ \text{rank 2} \\ \text{rank 3} \\ \text{rank 4} \\ \text{rank n} \end{bmatrix}$$

$$\mathbf{P} \begin{bmatrix} \text{airplane} \\ \textbf{panda} \\ \text{goldfish} \\ \text{mammal} \\ \text{husky} \end{bmatrix}$$

# Differentiable Top-$k$ Classification Learning

difftopk

CNN

$$\begin{bmatrix} 1 \\ 8 \\ -1 \\ 9 \\ 5 \end{bmatrix}$$

Diff. Rank

$$\mathbf{P} = \begin{array}{c} \\ \text{rank 1} \\ \text{rank 2} \\ \text{rank 3} \\ \text{rank 4} \\ \text{rank n} \end{array} \begin{bmatrix} \text{airplane} & \textbf{panda} & \text{goldfish} & \text{mammal} & \text{husky} \\ .0 & .3 & .0 & .7 & .0 \\ .0 & .6 & .0 & .3 & .1 \\ .1 & .1 & .0 & .0 & .8 \\ .7 & .0 & .2 & .0 & .1 \\ .2 & .0 & .8 & .0 & .0 \end{bmatrix}$$

# Differentiable Top-$k$ Classification Learning

difftopk

CNN

$$\begin{bmatrix} 1 \\ 8 \\ -1 \\ 9 \\ 5 \end{bmatrix}$$

Diff. Rank

$$\mathbf{P} = \begin{array}{l} \\ \text{rank 1} \\ \text{rank 2} \\ \text{rank 3} \\ \text{rank 4} \\ \text{rank n} \end{array} \begin{bmatrix} \text{airplane} & \textbf{panda} & \text{goldfish} & \text{mammal} & \text{husky} \\ .0 & .3 & .0 & .7 & .0 \\ .0 & .6 & .0 & .3 & .1 \\ .1 & .1 & .0 & .0 & .8 \\ .7 & .0 & .2 & .0 & .1 \\ .2 & .0 & .8 & .0 & .0 \end{bmatrix}$$

$$P_K = \begin{bmatrix} .5 & .5 & .0 & .0 & .0 \end{bmatrix}$$

# Differentiable Top-$k$ Classification Learning

difftopk

CNN

$$\begin{bmatrix} 1 \\ 8 \\ -1 \\ 9 \\ 5 \end{bmatrix}$$

Diff. Rank

$$\mathbf{P} = \begin{array}{c} \\ \text{rank 1} \\ \text{rank 2} \\ \text{rank 3} \\ \text{rank 4} \\ \text{rank n} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{airplane} & \textbf{panda} & \text{goldfish} & \text{mammal} & \text{husky} \end{array} \\ \begin{bmatrix} .0 & .3 & .0 & .7 & .0 \\ .0 & .6 & .0 & .3 & .1 \\ .1 & .1 & .0 & .0 & .8 \\ .7 & .0 & .2 & .0 & .1 \\ .2 & .0 & .8 & .0 & .0 \end{bmatrix} \end{array}$$

$$P_K = \begin{bmatrix} .5 & .5 & .0 & .0 & .0 \end{bmatrix}$$

$$\mathbb{E}_{k \sim P_K} \left[ \sum_{m=1}^{k} \mathbf{P}_{m,y} \right]$$

# Differentiable Top-$k$ Classification Learning

difftopk

CNN

$$\begin{bmatrix} 1 \\ 8 \\ -1 \\ 9 \\ 5 \end{bmatrix}$$

Diff. Rank

$$\mathbf{P} = \begin{array}{c} \\ \text{rank 1} \\ \text{rank 2} \\ \text{rank 3} \\ \text{rank 4} \\ \text{rank n} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{airplane} & \textbf{panda} & \text{goldfish} & \text{mammal} & \text{husky} \end{array} \\ \begin{bmatrix} .0 & .3 & .0 & .7 & .0 \\ .0 & .6 & .0 & .3 & .1 \\ .1 & .1 & .0 & .0 & .8 \\ .7 & .0 & .2 & .0 & .1 \\ .2 & .0 & .8 & .0 & .0 \end{bmatrix} \end{array} \begin{array}{c} \\ \times 1 \\ \times .5 \\ \times 0 \\ \times 0 \\ \times 0 \end{array}$$

$$P_K = \begin{bmatrix} .5 & .5 & .0 & .0 & .0 \end{bmatrix}$$

$$\mathbb{E}_{k \sim P_K} \left[ \sum_{m=1}^{k} \mathbf{P}_{m,y} \right]$$

# Differentiable Top-$k$ Classification Learning

difftopk

CNN

$$\begin{bmatrix} 1 \\ 8 \\ -1 \\ 9 \\ 5 \end{bmatrix}$$

Diff. Rank

$$\mathbf{P} = \begin{array}{c} \\ \text{rank 1} \\ \text{rank 2} \\ \text{rank 3} \\ \text{rank 4} \\ \text{rank n} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{airplane} & \textbf{panda} & \text{goldfish} & \text{mammal} & \text{husky} \end{array} \\ \begin{bmatrix} .0 & .3 & .0 & .7 & .0 \\ .0 & .6 & .0 & .3 & .1 \\ .1 & .1 & .0 & .0 & .8 \\ .7 & .0 & .2 & .0 & .1 \\ .2 & .0 & .8 & .0 & .0 \end{bmatrix} \end{array} \begin{array}{c} \\ \times 1 \\ \times .5 \\ \times 0 \\ \times 0 \\ \times 0 \end{array}$$

$$P_K = \begin{bmatrix} .5 & .5 & .0 & .0 & .0 \end{bmatrix}$$

$$p = \begin{bmatrix} .0 & .6 & .0 & .85 & .05 \end{bmatrix}$$

$$\mathbb{E}_{k \sim P_K} \left[ \sum_{m=1}^{k} \mathbf{P}_{m,y} \right]$$

# Differentiable Top-$k$ Classification Learning

difftopk

CNN

$$\begin{bmatrix} 1 \\ 8 \\ -1 \\ 9 \\ 5 \end{bmatrix}$$

Diff. Rank

$$\mathbf{P} = \begin{array}{c} \\ \text{rank 1} \\ \text{rank 2} \\ \text{rank 3} \\ \text{rank 4} \\ \text{rank n} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{airplane} & \textbf{panda} & \text{goldfish} & \text{mammal} & \text{husky} \end{array} \\ \begin{bmatrix} .0 & .3 & .0 & .7 & .0 \\ .0 & .6 & .0 & .3 & .1 \\ .1 & .1 & .0 & .0 & .8 \\ .7 & .0 & .2 & .0 & .1 \\ .2 & .0 & .8 & .0 & .0 \end{bmatrix} \end{array} \begin{array}{c} \\ \times 1 \\ \times .5 \\ \times 0 \\ \times 0 \\ \times 0 \end{array}$$

$$P_K = \begin{bmatrix} .5 & .5 & .0 & .0 & .0 \end{bmatrix}$$

$$p = \begin{bmatrix} .0 & .6 & .0 & .85 & .05 \end{bmatrix}$$

$$\arg\max_{\Theta} \quad \mathbb{E}_{X,y\sim\mathcal{D}}\left[\log\left(\mathbb{E}_{k\sim P_K}\left[\sum_{m=1}^{k}\mathbf{P}_{m,y}\right]\right)\right]$$

# Differentiable Top-$k$ Classification Learning

difftopk

CNN

$$\begin{bmatrix} 1 \\ 8 \\ -1 \\ 9 \\ 5 \end{bmatrix}$$

Diff. Rank

$$\mathbf{P} = \begin{array}{c} \\ \text{rank 1} \\ \text{rank 2} \\ \text{rank 3} \\ \text{rank 4} \\ \text{rank n} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{airplane} & \textbf{panda} & \text{goldfish} & \text{mammal} & \text{husky} \end{array} \\ \begin{bmatrix} .0 & .3 & .0 & .7 & .0 \\ .0 & .6 & .0 & .3 & .1 \\ .1 & .1 & .0 & .0 & .8 \\ .7 & .0 & .2 & .0 & .1 \\ .2 & .0 & .8 & .0 & .0 \end{bmatrix} \end{array} \begin{array}{c} \\ \times 1 \\ \times .5 \\ \times 0 \\ \times 0 \\ \times 0 \end{array}$$

$$P_K = \begin{bmatrix} .5 & .5 & .0 & .0 & .0 \end{bmatrix}$$

$$p = \begin{bmatrix} .0 & .6 & .0 & .85 & .05 \end{bmatrix}$$

$$\arg\max_{\Theta} \quad \mathbb{E}_{X,y\sim\mathcal{D}}\left[\log\left(\mathbb{E}_{k\sim P_K}\left[\sum_{m=1}^{k}\mathbf{P}_{m,y}\right]\right)\right] \qquad \mathcal{L}(X,y) = -\log\left(\sum_{k=1}^{n} P_K(k)\left(\sum_{m=1}^{k}\mathbf{P}_{m,y}(f_\Theta(X))\right)\right)$$

# Experiments: Fine-Tuning on ImageNet

# Experiments: Fine-Tuning on ImageNet

| ImageNet-1K | $P_K$ | Top-1 | Top-5 |
|---|---|---|---|
| Softmax | $[1, 0, 0, 0, 0]$ | 86.06 | 97.795 |
| Smooth top-5 loss | $[0, 0, 0, 0, 1]$ | 85.15 | 97.540 |
| Top-5 NeuralSort | $[0, 0, 0, 0, 1]$ | 33.37 | 94.748 |
| Top-5 SoftSort | $[0, 0, 0, 0, 1]$ | 18.23 | 94.965 |
| Top-5 SinkhornSort | $[0, 0, 0, 0, 1]$ | 85.65 | **97.991** |
| Top-5 DiffSortNets | $[0, 0, 0, 0, 1]$ | 69.05 | 97.389 |

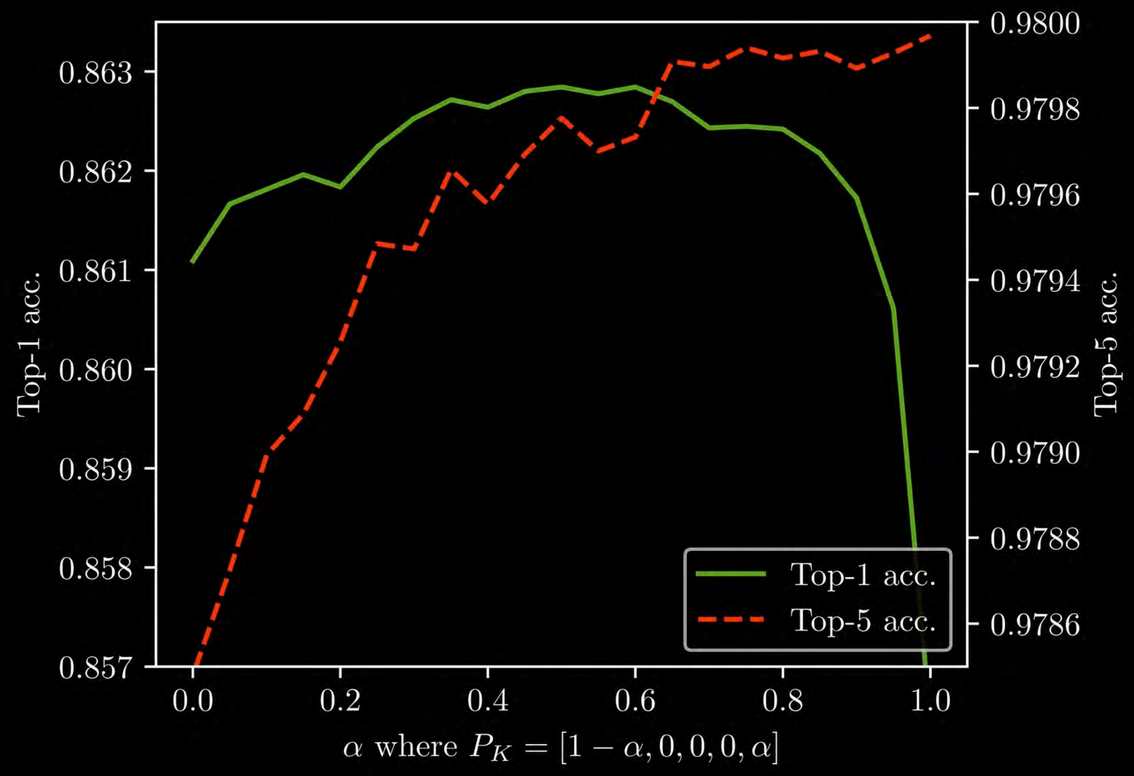
# Experiments: Fine-Tuning on ImageNet

| ImageNet-1K | $P_K$ | Top-1 | Top-5 |
|---|---|---|---|
| Softmax | $[1, 0, 0, 0, 0]$ | 86.06 | 97.795 |
| Smooth top-5 loss | $[0, 0, 0, 0, 1]$ | 85.15 | 97.540 |
| Top-5 NeuralSort | $[0, 0, 0, 0, 1]$ | 33.37 | 94.748 |
| Top-5 SoftSort | $[0, 0, 0, 0, 1]$ | 18.23 | 94.965 |
| Top-5 SinkhornSort | $[0, 0, 0, 0, 1]$ | 85.65 | **97.991** |
| Top-5 DiffSortNets | $[0, 0, 0, 0, 1]$ | 69.05 | 97.389 |
| Top-$k$ NeuralSort | $[.5, 0, 0, 0, .5]$ | 86.30 | 97.896 |
| Top-$k$ SoftSort | $[.5, 0, 0, 0, .5]$ | 86.26 | 97.963 |
| Top-$k$ SinkhornSort | $[.5, 0, 0, 0, .5]$ | **86.29** | **97.971** |
| Top-$k$ DiffSortNets | $[.5, 0, 0, 0, .5]$ | 86.24 | 97.937 |

# Experiments: Fine-Tuning on ImageNet

difftopk

| ImageNet-1K | $P_K$ | Top-1 | Top-5 |
|---|---|---|---|
| Softmax | $[1, 0, 0, 0, 0]$ | 86.06 | 97.795 |
| Smooth top-5 loss | $[0, 0, 0, 0, 1]$ | 85.15 | 97.540 |
| Top-5 NeuralSort | $[0, 0, 0, 0, 1]$ | 33.37 | 94.748 |
| Top-5 SoftSort | $[0, 0, 0, 0, 1]$ | 18.23 | 94.965 |
| Top-5 SinkhornSort | $[0, 0, 0, 0, 1]$ | 85.65 | **97.991** |
| Top-5 DiffSortNets | $[0, 0, 0, 0, 1]$ | 69.05 | 97.389 |
| Top-$k$ NeuralSort | $[.5, 0, 0, 0, .5]$ | 86.30 | 97.896 |
| Top-$k$ SoftSort | $[.5, 0, 0, 0, .5]$ | 86.26 | 97.963 |
| Top-$k$ SinkhornSort | $[.5, 0, 0, 0, .5]$ | **86.29** | **97.971** |
| Top-$k$ DiffSortNets | $[.5, 0, 0, 0, .5]$ | 86.24 | 97.937 |



# Experiments: Fine-Tuning on ImageNet

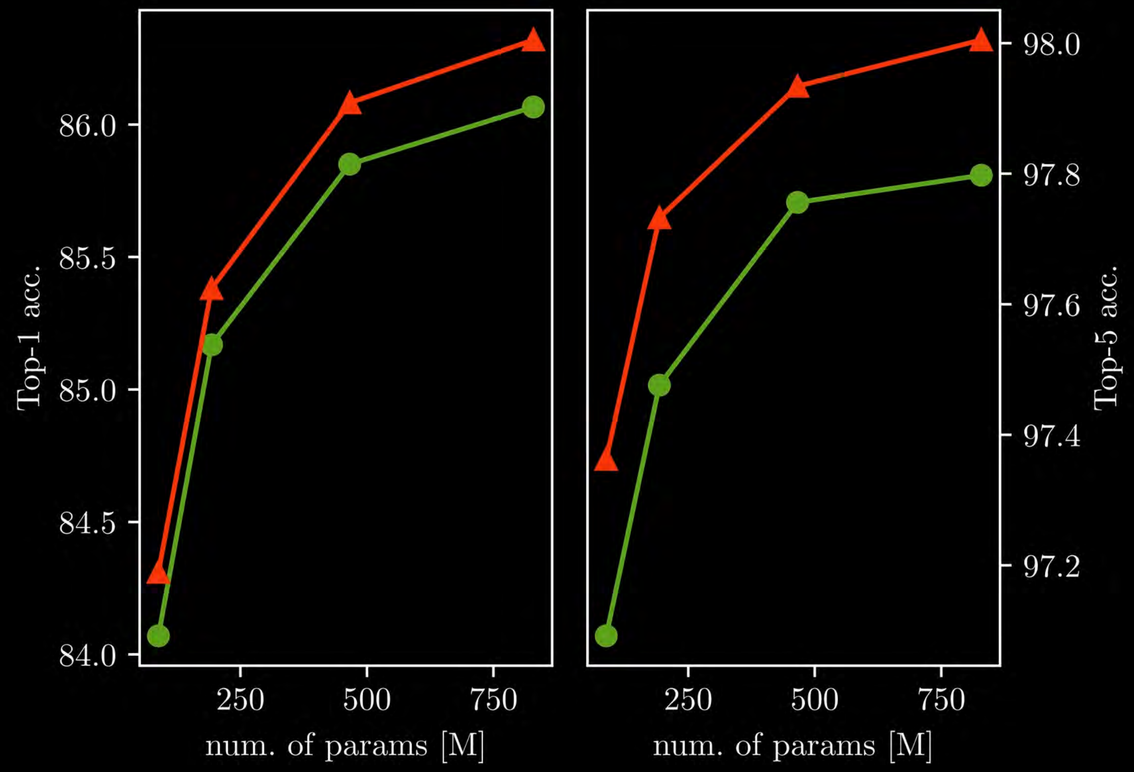

| ImageNet-1K | Public | Top-1 | Top-5 |
|---|---|---|---|
| ResNet50 | ✓ | 79.26 | 94.75 |
| ResNet152 | ✓ | 80.62 | 95.51 |
| ResNeXt-101 32x48d WSL | ✓ | 85.43 | 97.57 |
| ViT-L/16 | ✓ | 87.76 | −− |
| Noisy Student EfficientNet-L2 | ✓ | **88.35** | **98.65** |
| BiT-L | ✗ | 87.54 | 98.46 |
| CLIP (w/ Noisy Student EffNet-L2) | ✗ | ≈ 88.4 | −− |
| ViT-H/14 | ✗ | 88.55 | −− |
| ALIGN (EffNet-L2) | ✗ | 88.64 | **98.67** |
| Meta Pseudo Labels (EffNet-L2) | ✗ | 90.20 | ≈ 98.8 |
| ViT-G/14 | ✗ | 90.45 | −− |
| CoAtNet-7 | ✗ | **90.88** | −− |

# Experiments: Fine-Tuning on ImageNet


| ImageNet-1K | Public | Top-1 | Top-5 |
|---|---|---|---|
| ResNet50 | ✓ | 79.26 | 94.75 |
| ResNet152 | ✓ | 80.62 | 95.51 |
| ResNeXt-101 32x48d WSL | ✓ | 85.43 | 97.57 |
| ViT-L/16 | ✓ | 87.76 | −− |
| Noisy Student EfficientNet-L2 | ✓ | **88.35** | **98.65** |
| BiT-L | ✗ | 87.54 | 98.46 |
| CLIP (w/ Noisy Student EffNet-L2) | ✗ | ≈ 88.4 | −− |
| ViT-H/14 | ✗ | 88.55 | −− |
| ALIGN (EffNet-L2) | ✗ | 88.64 | **98.67** |
| Meta Pseudo Labels (EffNet-L2) | ✗ | 90.20 | ≈ 98.8 |
| ViT-G/14 | ✗ | 90.45 | −− |
| CoAtNet-7 | ✗ | **90.88** | −− |
| ResNeXt-101 32x48d WSL | | 86.06 | 97.80 |
| Top-$k$ SinkhornSort | | **86.22** | 97.99 |
| Top-$k$ DiffSortNets | | 86.21 | **98.00** |

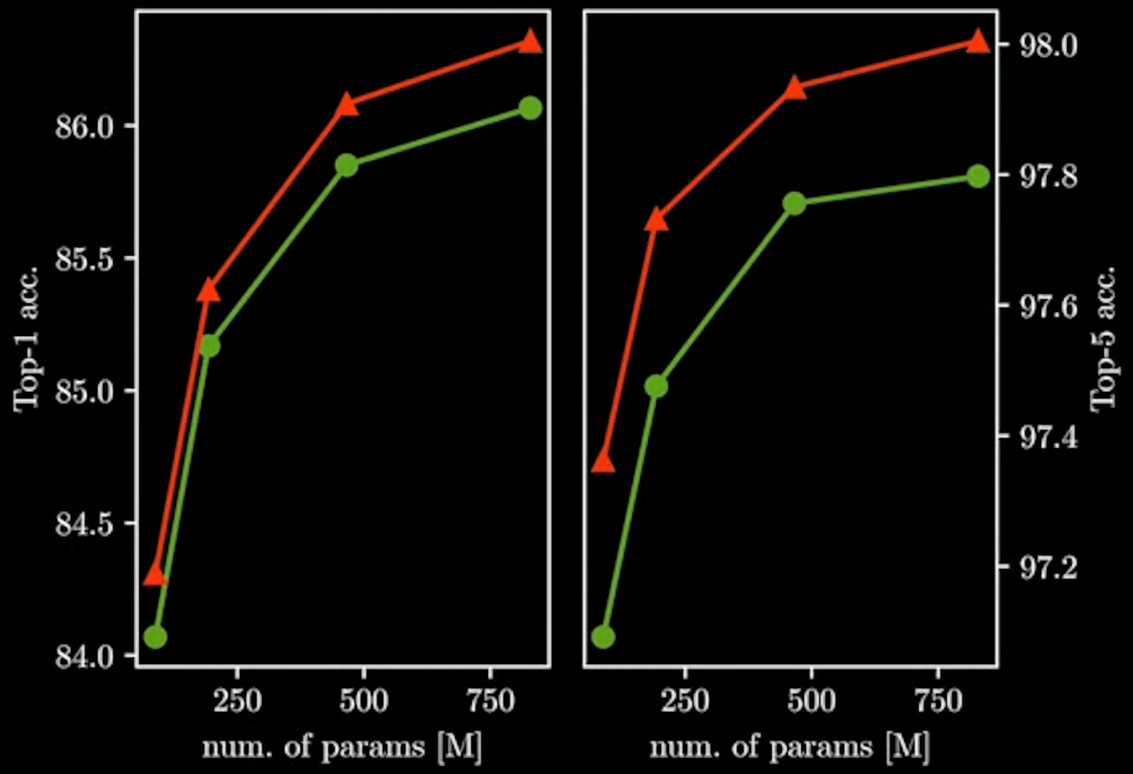
# Experiments: Fine-Tuning on ImageNet

| ImageNet-1K | Public | Top-1 | Top-5 |
|---|---|---|---|
| ResNet50 | ✓ | 79.26 | 94.75 |
| ResNet152 | ✓ | 80.62 | 95.51 |
| ResNeXt-101 32x48d WSL | ✓ | 85.43 | 97.57 |
| ViT-L/16 | ✓ | 87.76 | – – |
| Noisy Student EfficientNet-L2 | ✓ | **88.35** | **98.65** |
| BiT-L | ✗ | 87.54 | 98.46 |
| CLIP (w/ Noisy Student EffNet-L2) | ✗ | ≈ 88.4 | – – |
| ViT-H/14 | ✗ | 88.55 | – – |
| ALIGN (EffNet-L2) | ✗ | 88.64 | **98.67** |
| Meta Pseudo Labels (EffNet-L2) | ✗ | 90.20 | ≈ 98.8 |
| ViT-G/14 | ✗ | 90.45 | – – |
| CoAtNet-7 | ✗ | **90.88** | – – |
| ResNeXt-101 32x48d WSL | | 86.06 | 97.80 |
| Top-$k$ SinkhornSort | | **86.22** | 97.99 |
| Top-$k$ DiffSortNets | | 86.21 | **98.00** |



# Experiments: Fine-Tuning on ImageNet

| ImageNet-1K | Public | Top-1 | Top-5 |
|---|---|---|---|
| ResNet50 | ✓ | 79.26 | 94.75 |
| ResNet152 | ✓ | 80.62 | 95.51 |
| ResNeXt-101 32x48d WSL | ✓ | 85.43 | 97.57 |
| ViT-L/16 | ✓ | 87.76 | – – |
| Noisy Student EfficientNet-L2 | ✓ | **88.35** | **98.65** |
| BiT-L | ✗ | 87.54 | 98.46 |
| CLIP (w/ Noisy Student EffNet-L2) | ✗ | ≈ 88.4 | – – |
| ViT-H/14 | ✗ | 88.55 | – – |
| ALIGN (EffNet-L2) | ✗ | 88.64 | **98.67** |
| Meta Pseudo Labels (EffNet-L2) | ✗ | 90.20 | ≈ 98.8 |
| ViT-G/14 | ✗ | 90.45 | – – |
| CoAtNet-7 | ✗ | **90.88** | – – |
| ResNeXt-101 32x48d WSL | | 86.06 | 97.80 |
| Top-$k$ SinkhornSort | | **86.22** | 97.99 |
| Top-$k$ DiffSortNets | | 86.21 | **98.00** |
| Noisy Student EfficientNet-L2 | | 88.33 | 98.65 |
| Top-$k$ SinkhornSort | | 88.32 | 98.66 |
| Top-$k$ DiffSortNets | | **88.37** | **98.68** |

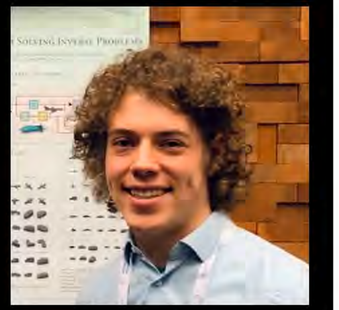


# Thank you!

difftopk

https://github.com/Felix-Petersen/difftopk

**Felix Petersen**
University of Konstanz
@FHKPetersen
Petersen.ai



Universität Konstanz

MIT-IBM Watson AI Lab

PARIS LODRON UNIVERSITÄT SALZBURG

GOETHE UNIVERSITÄT FRANKFURT AM MAIN