Transfer Learning of pre-trained CNN representations

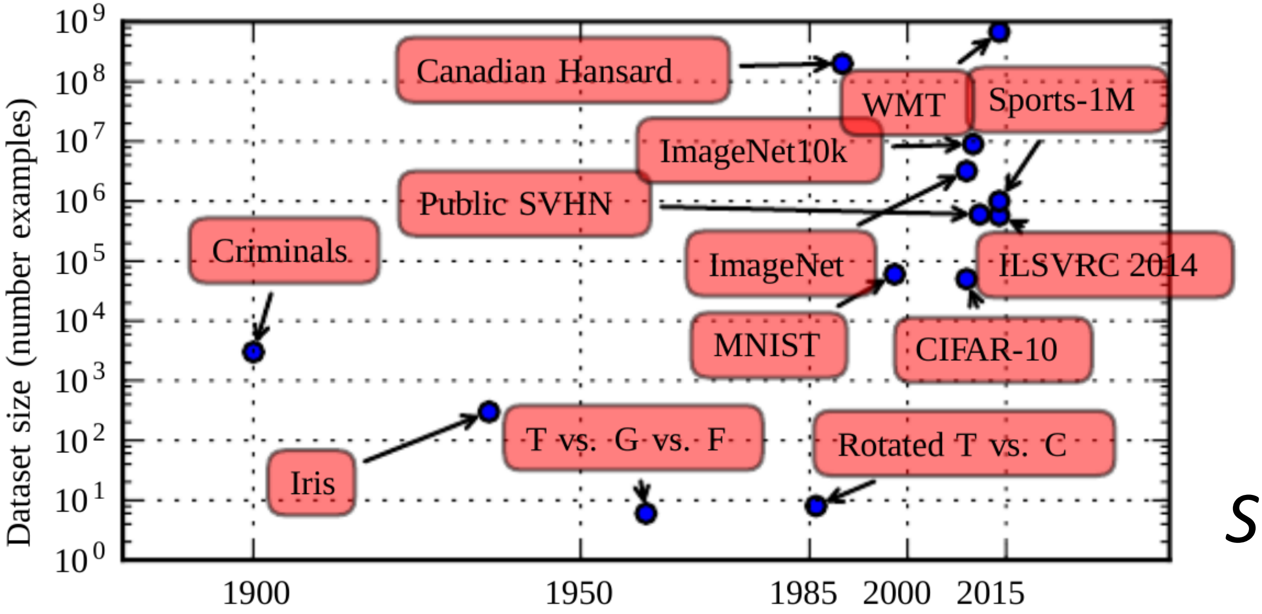
Michał Sokólski

Agenda

1. The problem of large datasets in Deep Learning
2. Transfer Learning & Representation Learning
3. Experiment Overview
4. Initial results

Deep learning works great but...

- *As of 2016, a rough rule of thumb is that a supervised deep learning algorithm will generally achieve acceptable performance with around 5,000 labeled examples per category and will match or exceed human performance when trained with a dataset containing at least 10 million labeled examples* [1]



Source: [1]

Can we get reasonable results using deep learning when the dataset is small?

Sometimes yes.

- Transfer Learning
- Representation Learning

Main idea

- Train model on a large source (upstream) dataset
- Obtain a good representation
- Fine-tune this representation on a smaller destination (downstream) dataset

Representation Learning

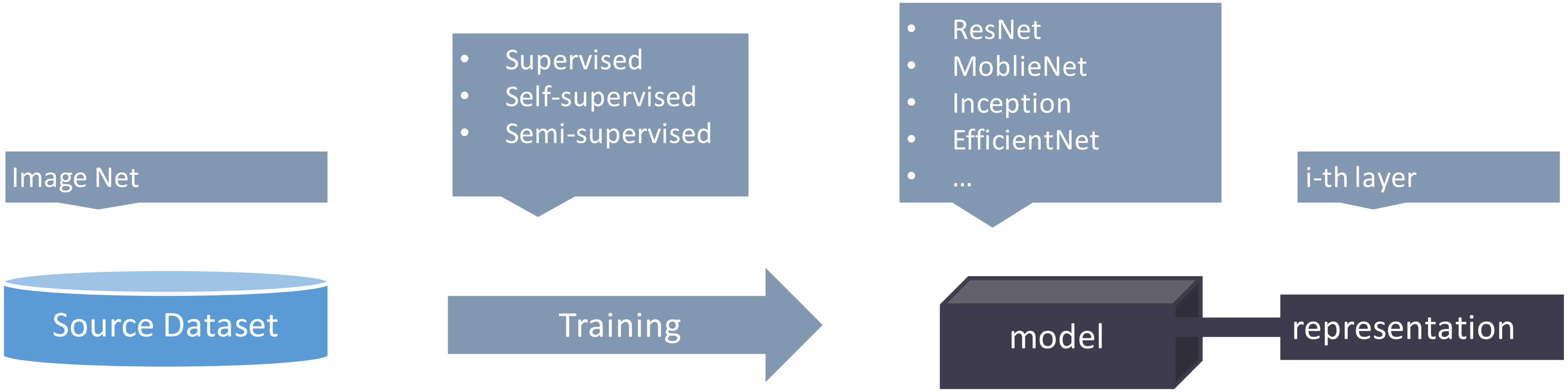


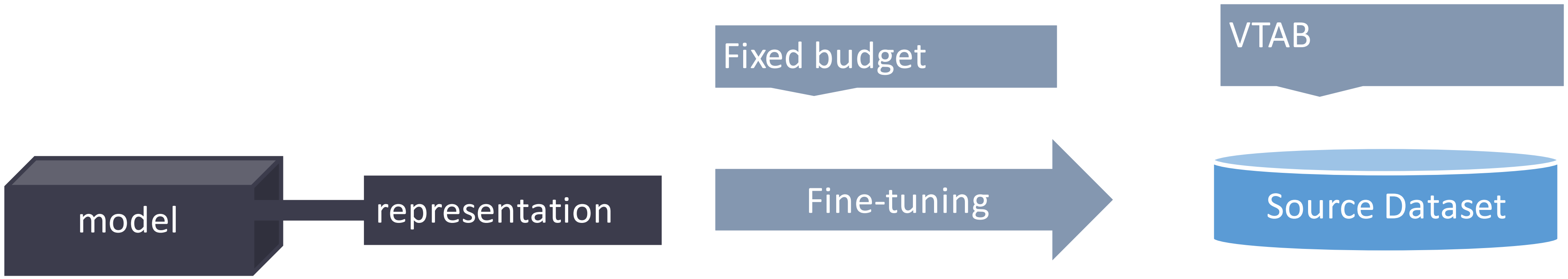
Image Net

- Supervised
- Self-supervised
- Semi-supervised

- ResNet
- MoblieNet
- Inception
- EfficientNet
- …

i-th layer

Transfer Learning



Experiment outline

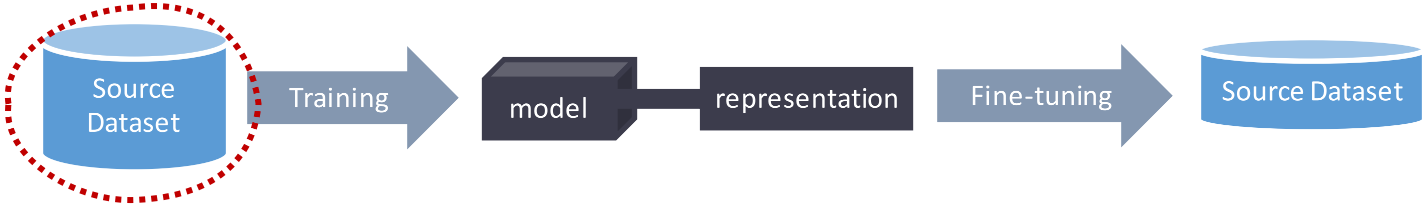
- Goal:

Discover well-generalizable, open source, pre-trained CNN architectures

- Method:

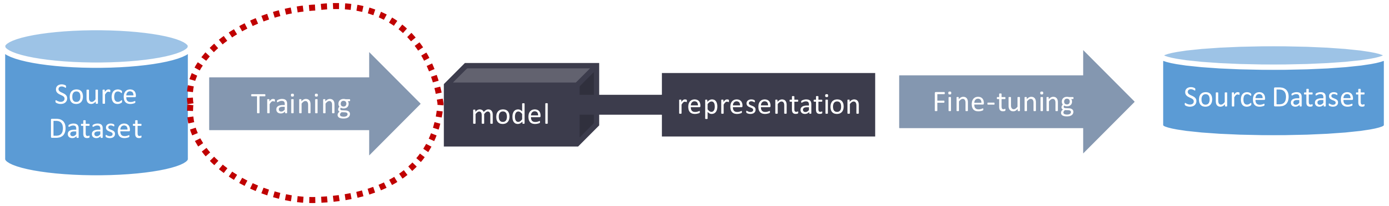
Benchmark as many models on diverse vision tasks, to see how well they generalize.

Source dataset



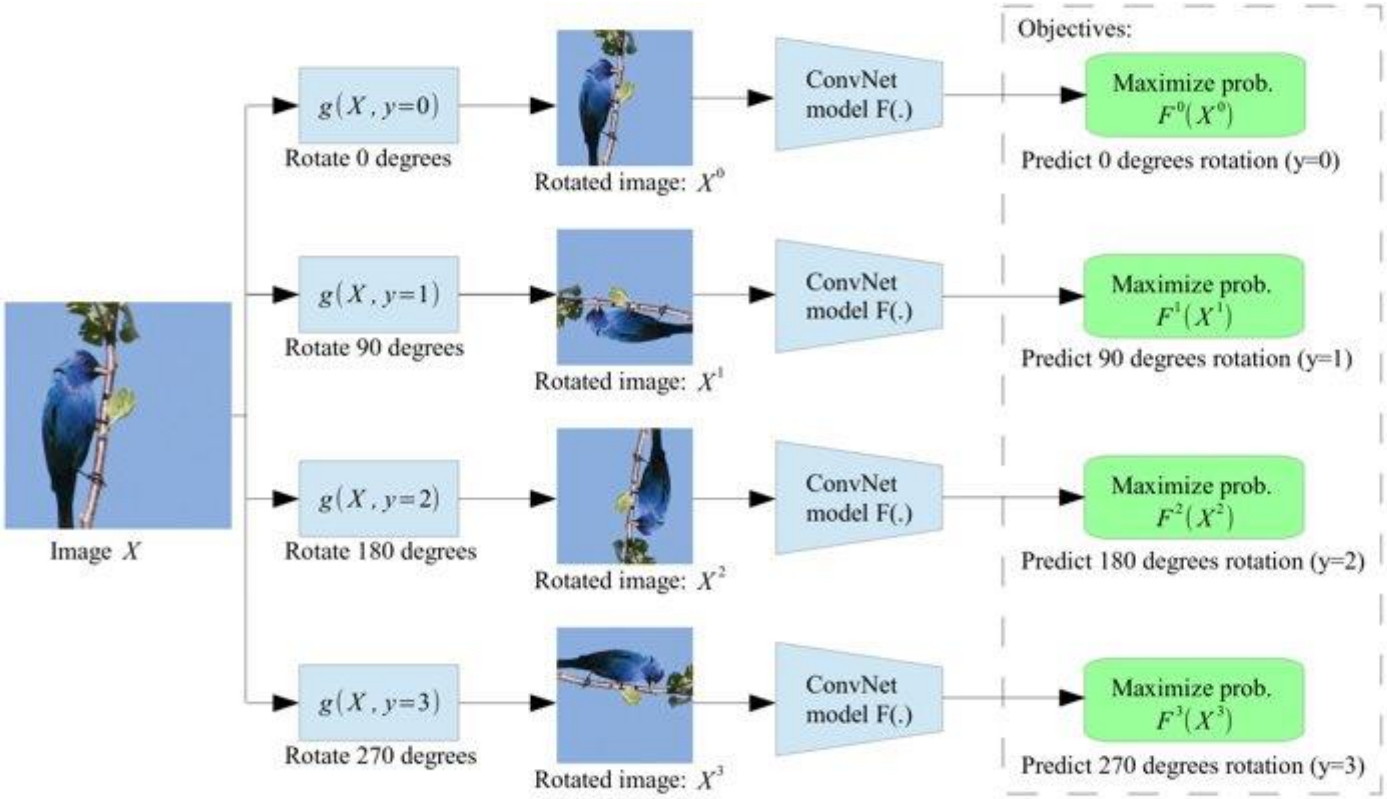
- Different flavors of ImageNet – depending on actual models

Training methods

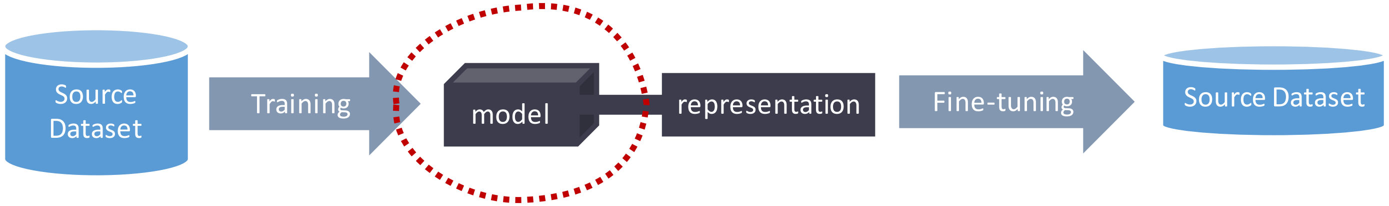


- Supervised learning
- Self-supervised learning
- Semi-supervised learning

Self-supervised learning



Source: https://arxiv.org/abs/1803.07728



Models

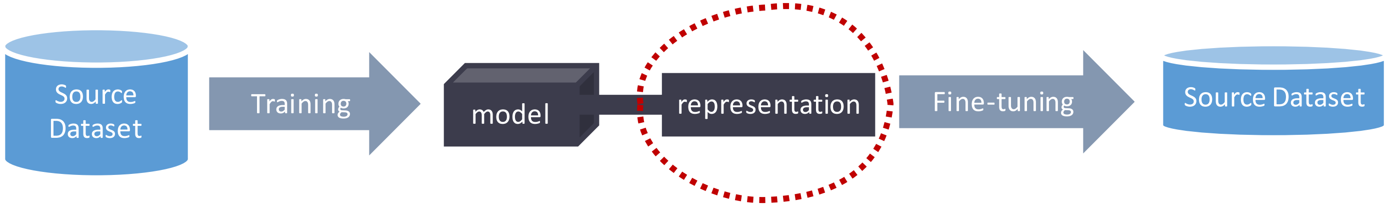
- pytorch.org/docs/stable/torchvision/models.html

(~20 models, supervised)

- github.com/open-mmlab/OpenSelfSup

(~10 models, self-supervised)

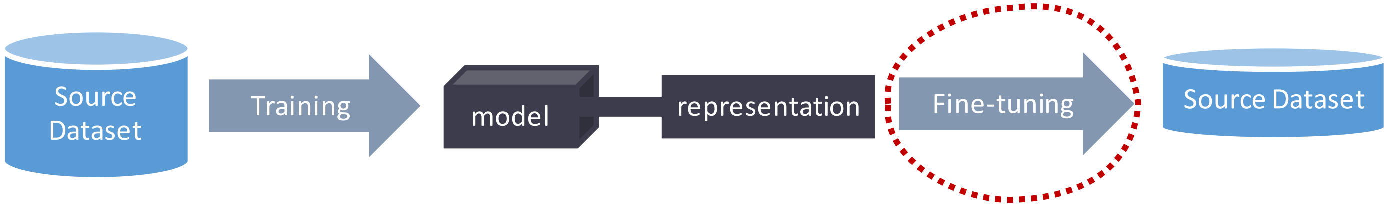
- github.com/rwightman/pytorch-image-models
- (~280 models, supervised, semi-supervised)



Representation

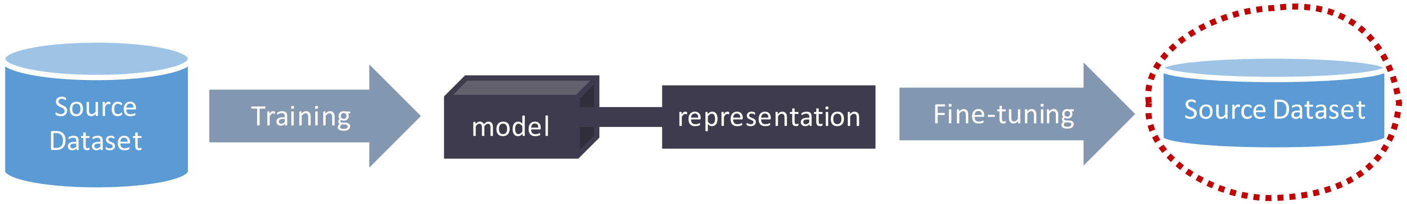
- CNN networks can be divided into backbone and classification head
- For my purposes, I choose second-to last layer as representation
- For example in ResNet family it's 2048-dimensional vector.

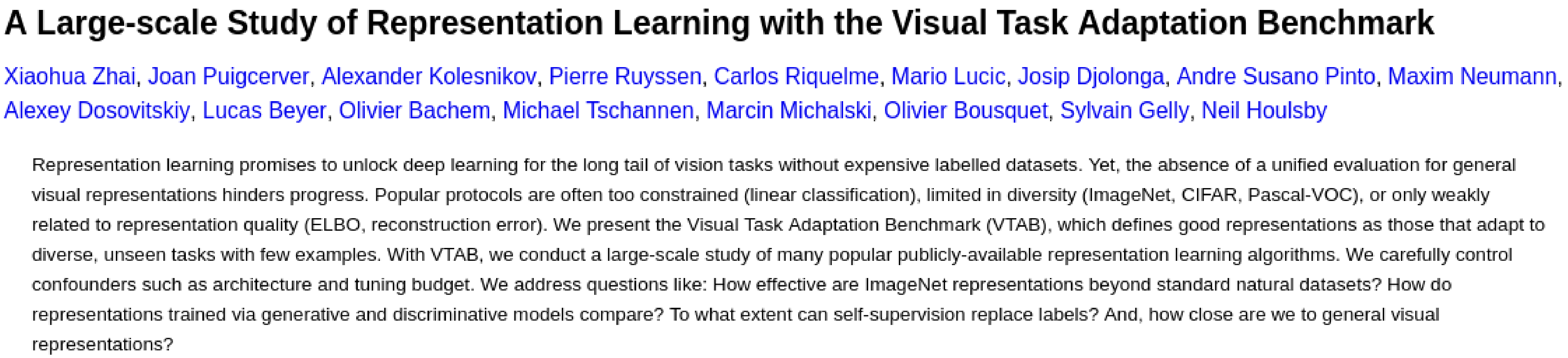
Fine-tuning



- Very small hyperparameter grid search
- Two learning rates, two training durations

VTAB [2]





A Large-scale Study of Representation Learning with the Visual Task Adaptation Benchmark

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, Neil Houlsby

Representation learning promises to unlock deep learning for the long tail of vision tasks without expensive labelled datasets. Yet, the absence of a unified evaluation for general visual representations hinders progress. Popular protocols are often too constrained (linear classification), limited in diversity (ImageNet, CIFAR, Pascal-VOC), or only weakly related to representation quality (ELBO, reconstruction error). We present the Visual Task Adaptation Benchmark (VTAB), which defines good representations as those that adapt to diverse, unseen tasks with few examples. With VTAB, we conduct a large-scale study of many popular publicly-available representation learning algorithms. We carefully control confounders such as architecture and tuning budget. We address questions like: How effective are ImageNet representations beyond standard natural datasets? How do representations trained via generative and discriminative models compare? To what extent can self-supervision replace labels? And, how close are we to general visual representations?

Visual Task Adaptation Benchmark

- Good representations = able to adapt to diverse and small datasets
- 19 diverse vision tasks with a small (1K) training and validation sets
- Comparison of 18 different methods of representation learning (mostly trained on ResNet50):

  - generative,

  - from-scratch,

  - self-supervised,

  - semi-supervised,

  - supervised

VTAB evaluation protocol

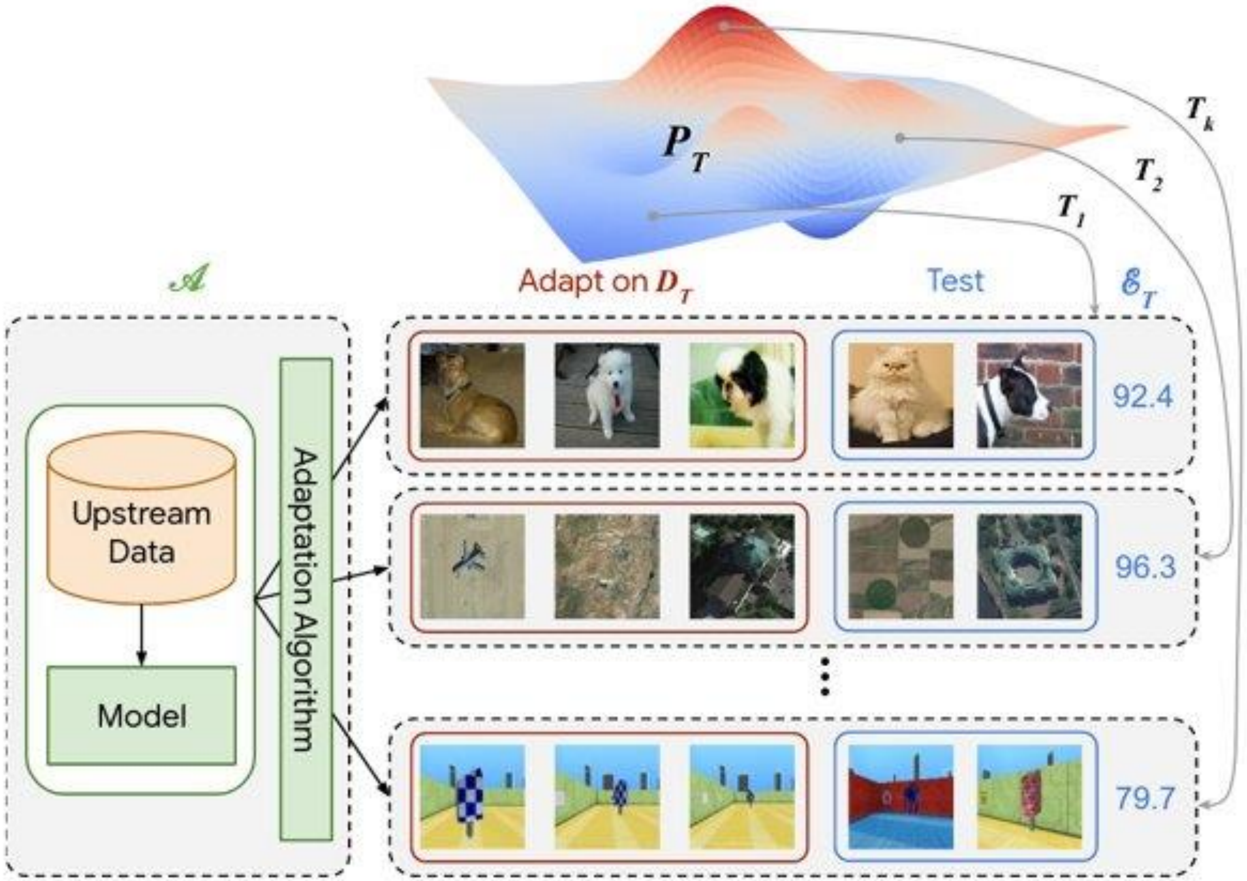
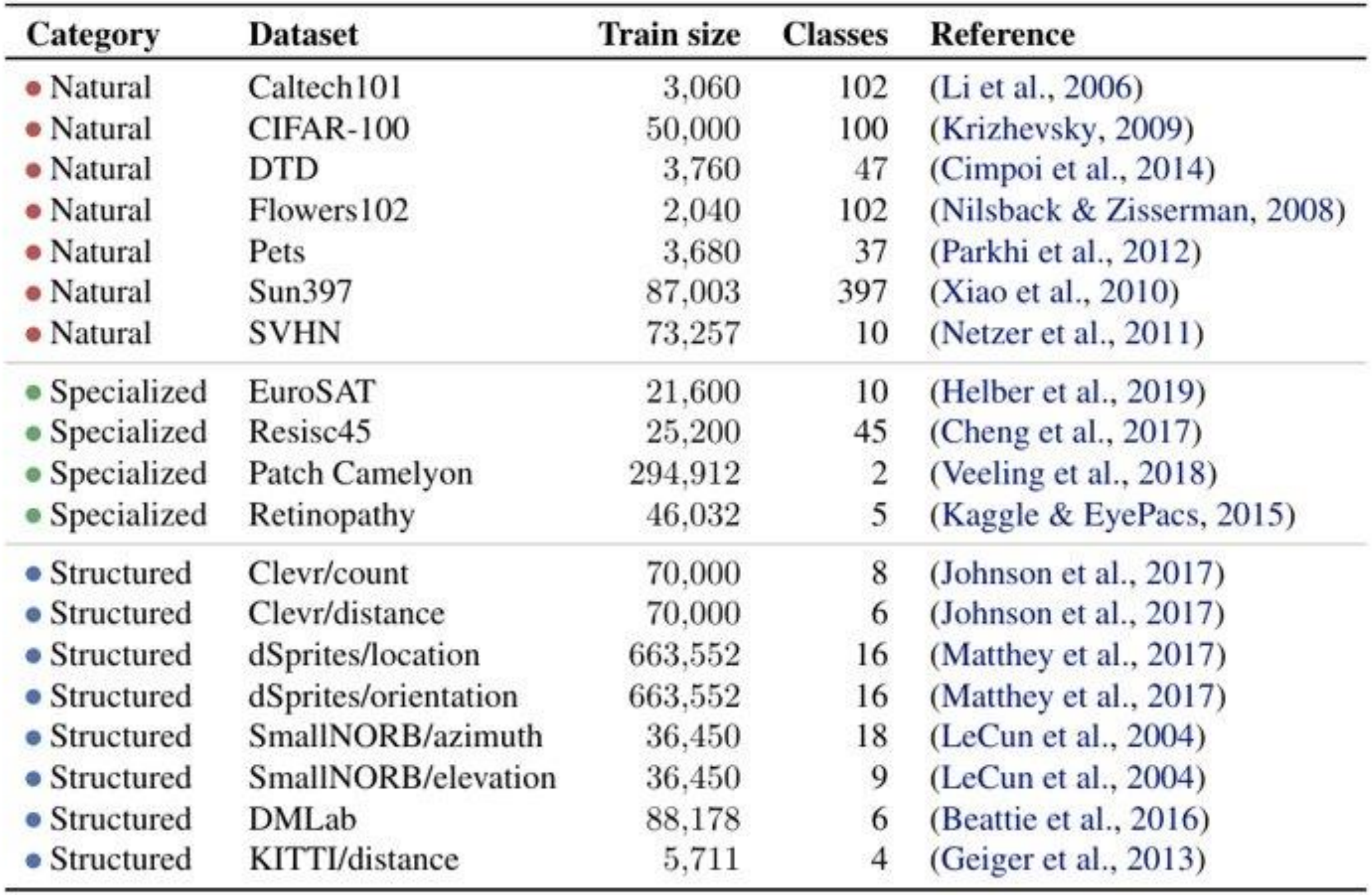


Figure 1. Overview of the VTAB evaluation protocol.

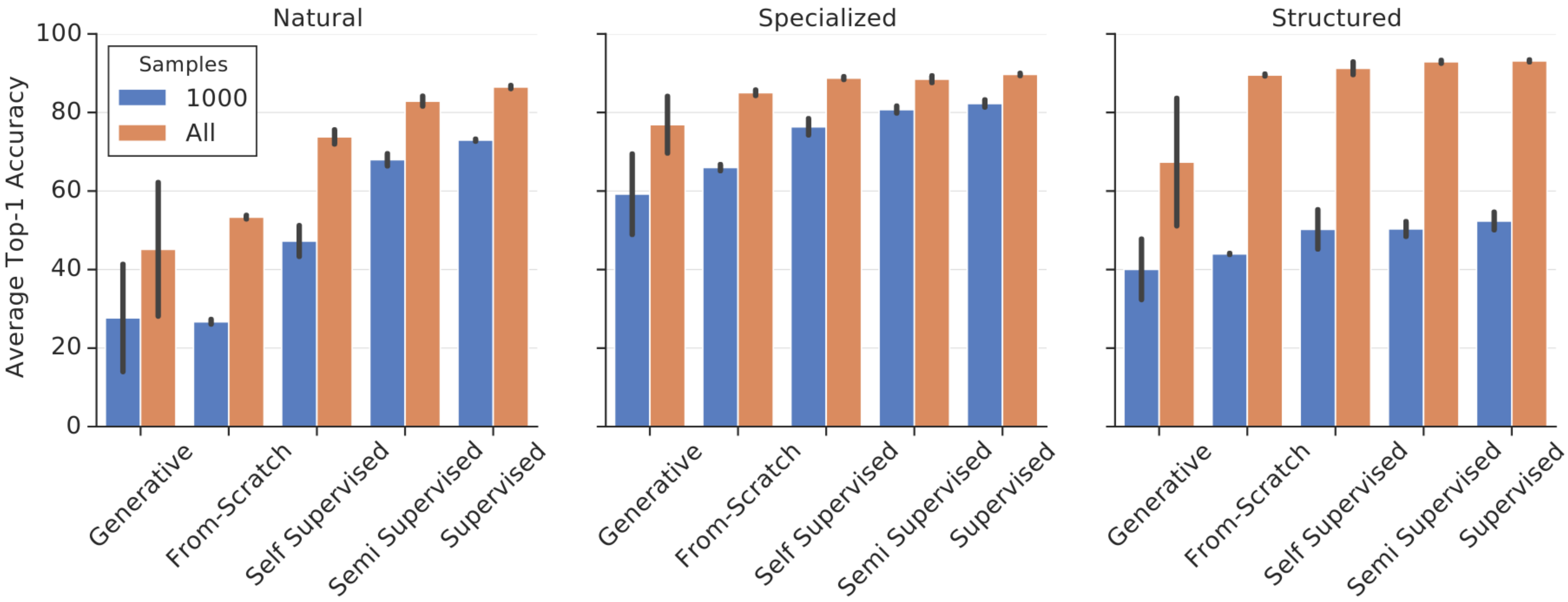
VTAB datasets



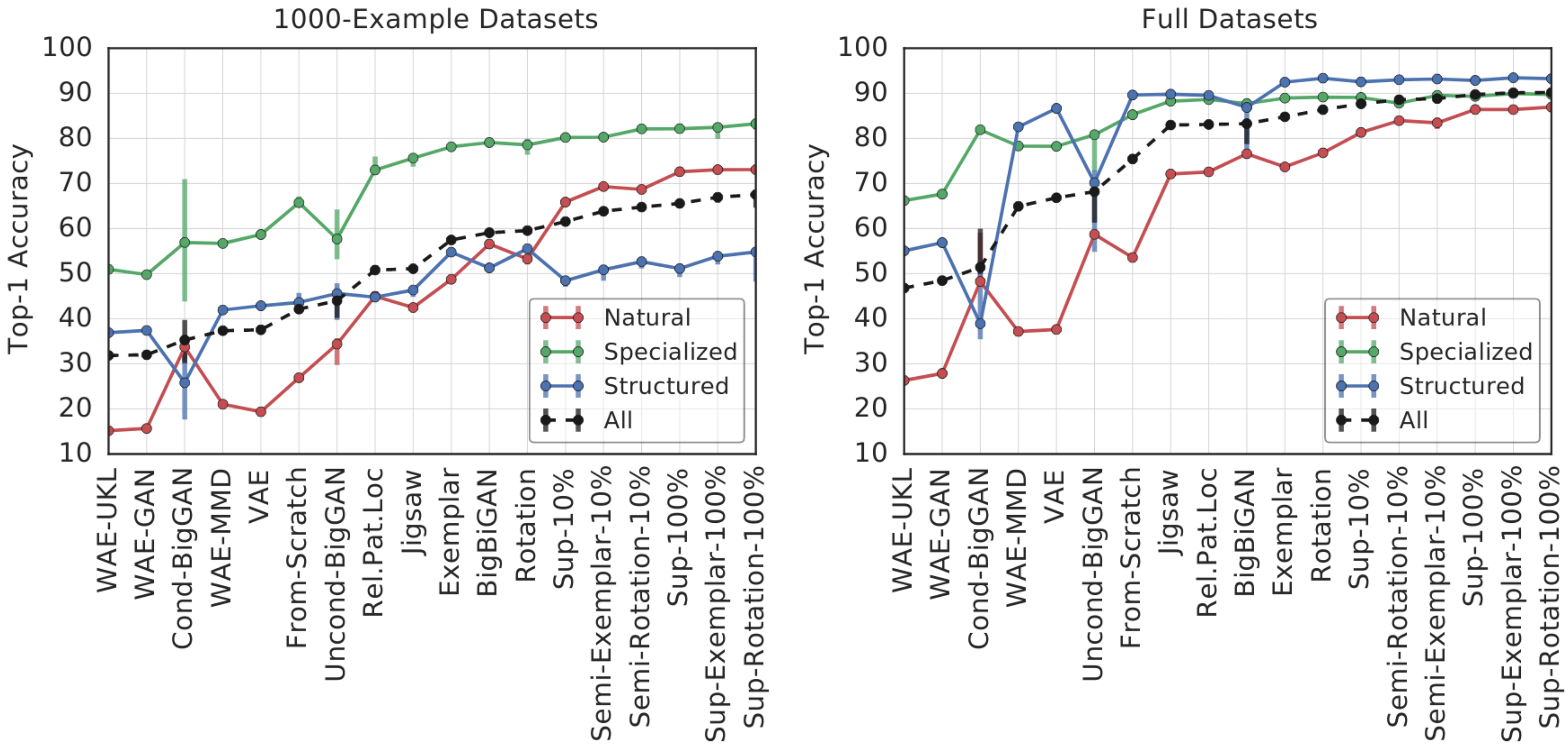
Category	Dataset	Train size	Classes	Reference
● Natural	Caltech101	3,060	102	(Li et al., 2006)
● Natural	CIFAR-100	50,000	100	(Krizhevsky, 2009)
● Natural	DTD	3,760	47	(Cimpoi et al., 2014)
● Natural	Flowers102	2,040	102	(Nilsback & Zisserman, 2008)
● Natural	Pets	3,680	37	(Parkhi et al., 2012)
● Natural	Sun397	87,003	397	(Xiao et al., 2010)
● Natural	SVHN	73,257	10	(Netzer et al., 2011)
● Specialized	EuroSAT	21,600	10	(Helber et al., 2019)
● Specialized	Resisc45	25,200	45	(Cheng et al., 2017)
● Specialized	Patch Camelyon	294,912	2	(Veeling et al., 2018)
● Specialized	Retinopathy	46,032	5	(Kaggle & EyePacs, 2015)
● Structured	Clevr/count	70,000	8	(Johnson et al., 2017)
● Structured	Clevr/distance	70,000	6	(Johnson et al., 2017)
● Structured	dSprites/location	663,552	16	(Matthey et al., 2017)
● Structured	dSprites/orientation	663,552	16	(Matthey et al., 2017)
● Structured	SmallNORB/azimuth	36,450	18	(LeCun et al., 2004)
● Structured	SmallNORB/elevation	36,450	9	(LeCun et al., 2004)
● Structured	DMLab	88,178	6	(Beattie et al., 2016)
● Structured	KITTI/distance	5,711	4	(Geiger et al., 2013)

Table 2. Description of the datasets used for the tasks in VTAB.

VTAB results



VTAB results



VTAB conclusions

- ImageNet labels are indeed effective for natural and specialized tasks
- Generative representations are less promising
- Self-supervised learning appears promising, even outperforming supervision on some tasks that require structured understanding
- Self-supervision can almost (but not quite) replace 90% of ImageNet labels
- Larger models are usually better
- VTAB-1K results approximate well VTAB results

300 models are a lot

There are methods to approximate quality of transfer learning

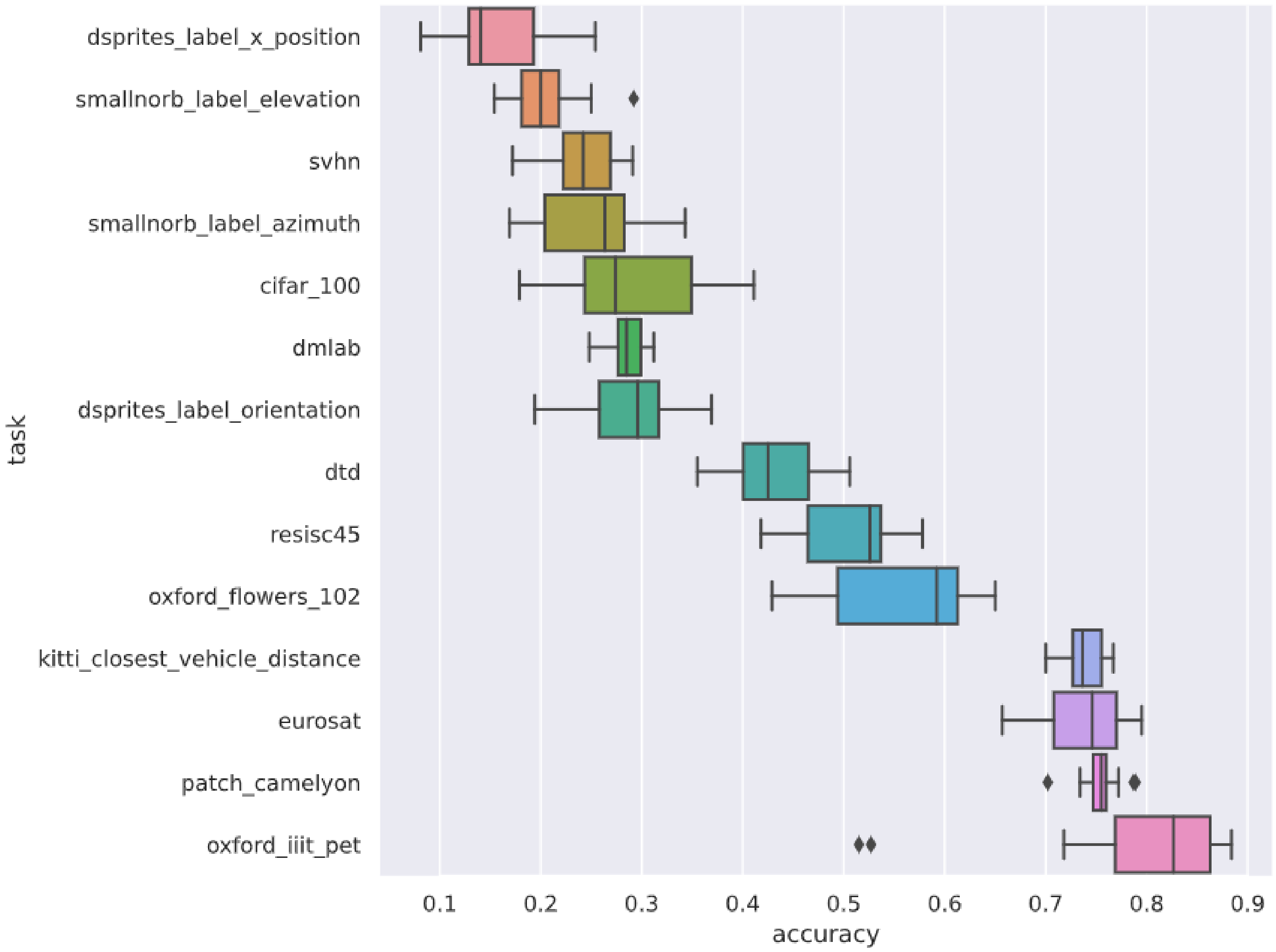
- Linear evaluation
- kNN
- LEEP from LEEP: A New Measure to Evaluate Transferability of Learned Representations
- NLEEP from Ranking Neural Checkpoints

kNN Evaluation

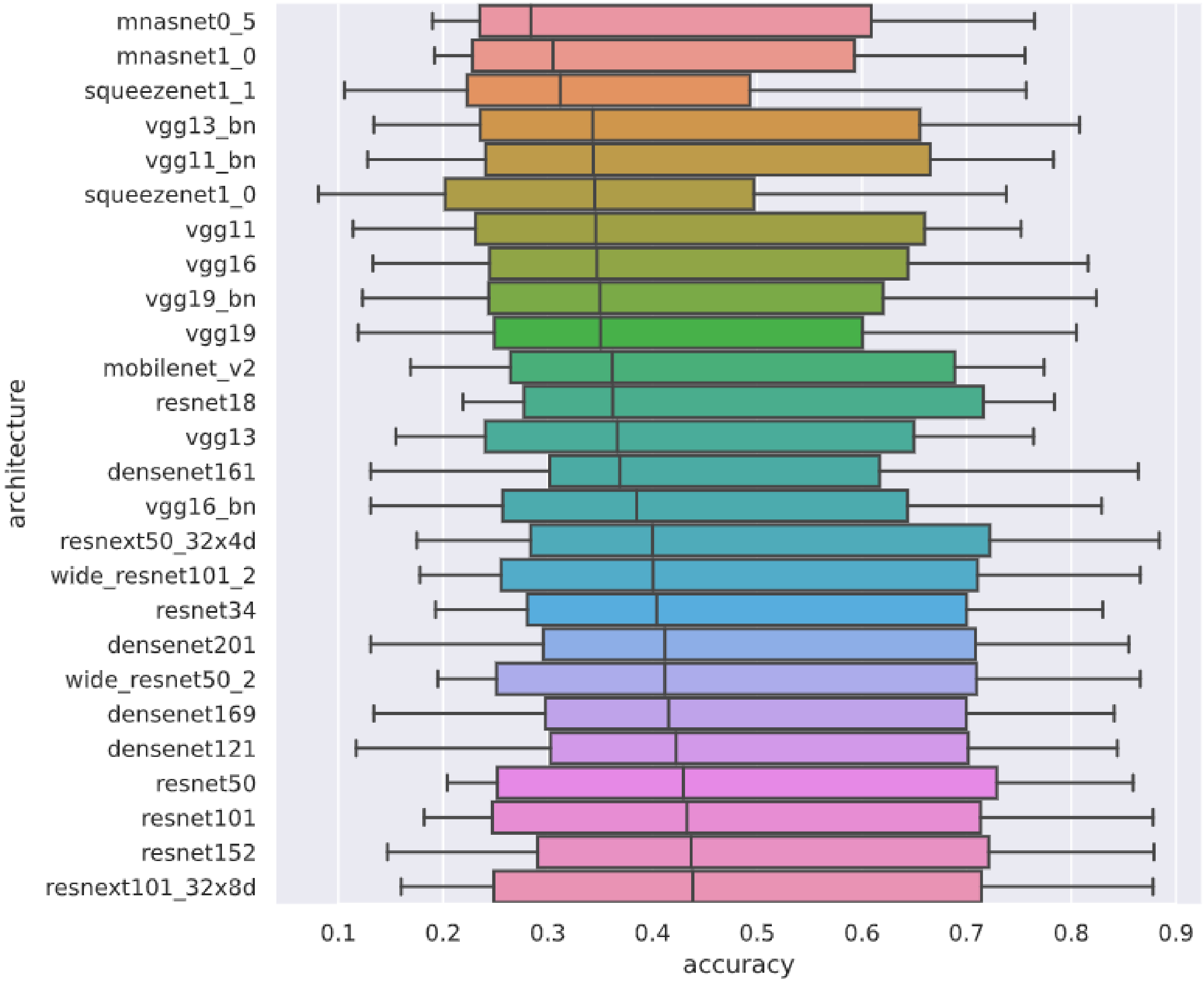
Method presented in Scalable Transfer Learning with Expert Models

1. Discard classifier head to get the representation
2. Run LOOCV using kNN(k=1)
3. Get mean accuracy

Initial results on Torchvision and 16 tasks



Initial results on Torchvision and 16 tasks



Architecture	Top-5
resnet152	9
resnet50	8
resnet101	7
resnext50_32x4d	5
wide_resnet101_24	4
resnext101_32x8d	4
densenet161	4

Future plans

- Complete kNN evaluation for the rest of the networks
- Run finte-tuning on the best ~20 models
- Rank models using Elo-based Predictive Power

Q&A

Bibliography

1. Deep Learning (2016) - Ian Goodfellow and Yoshua Bengio and Aaron Courville
2. A Large-scale Study of Representation Learning with the Visual Task Adaptation Benchmark