AutoDL: Transfer Learning

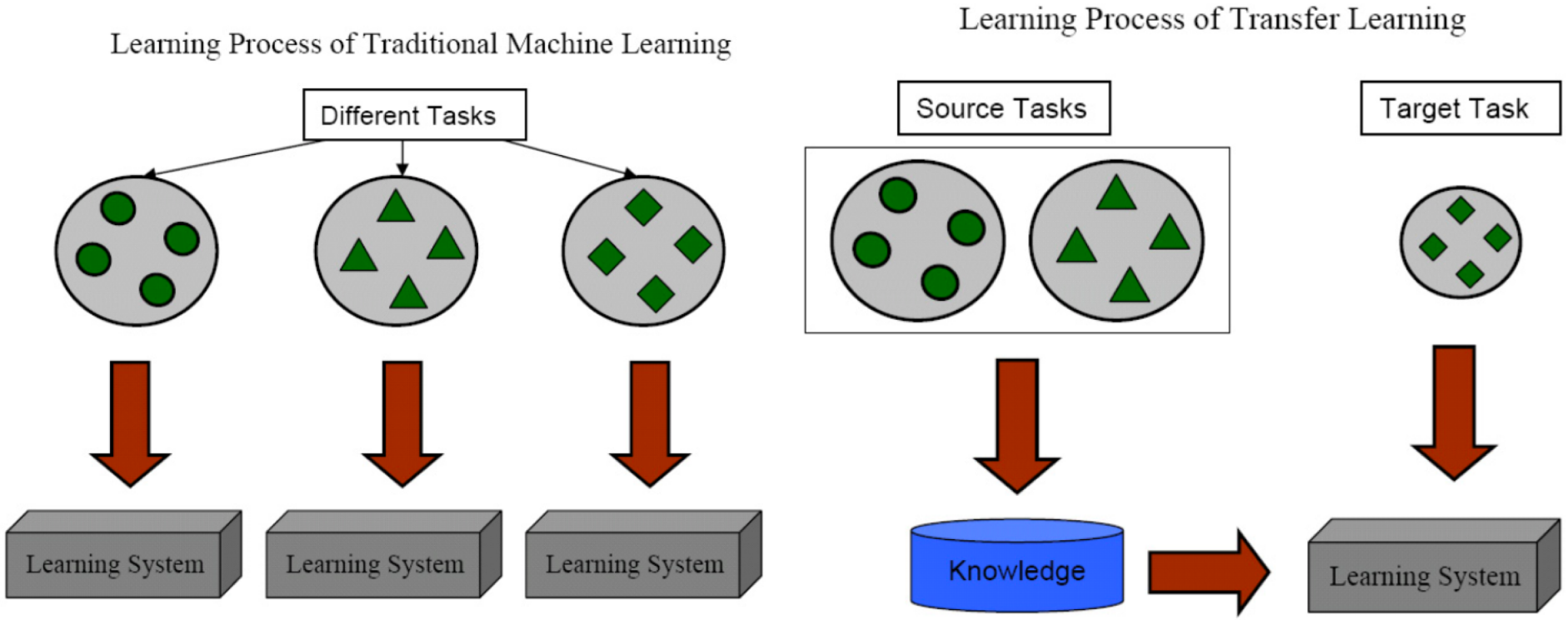
Xingjian Li
Big Data Lab, Baidu Research

Overview

- Problem Definition
- Empirical Studies
- Theoretical Understandings
- Novel Algorithms
- Applications

Background about Transfer Learning

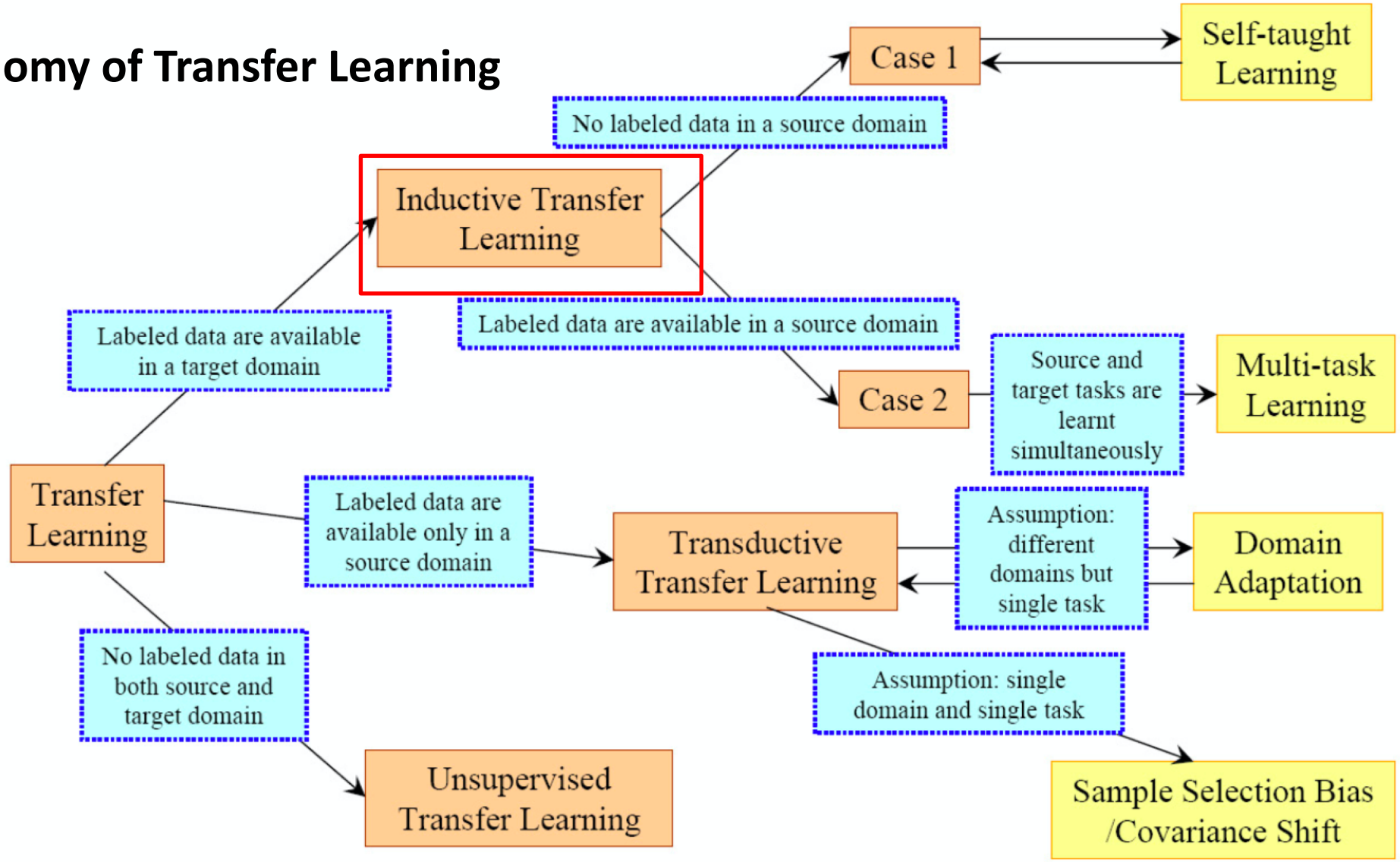
- Problem Background
  - Improve model performance when training dataset is small
  - Use source task to help target task



Sinno Jialin Pan, Qiang Yang, et al. A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10):1345– 1359, 2010.

Background about Transfer Learning

Taxonomy of Transfer Learning



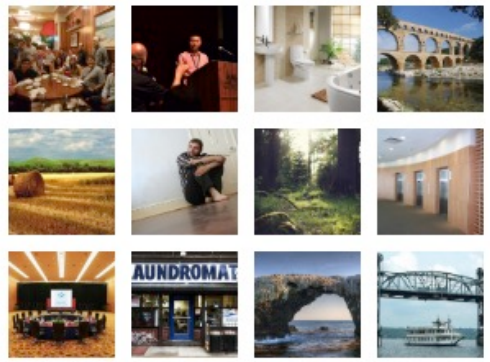
Sinno Jialin Pan, Qiang Yang, et al. A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10):1345– 1359, 2010.

Background about Transfer Learning

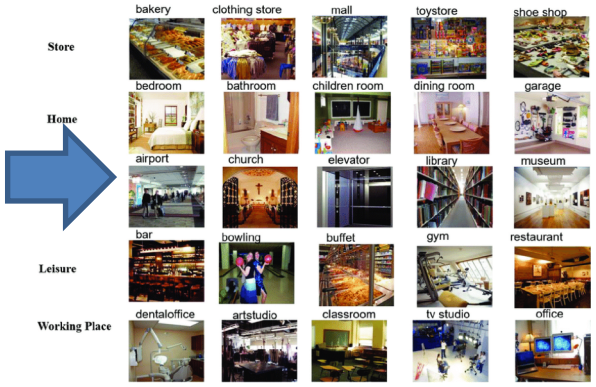
Common Practice of Inductive Transfer Learning

ImageNet

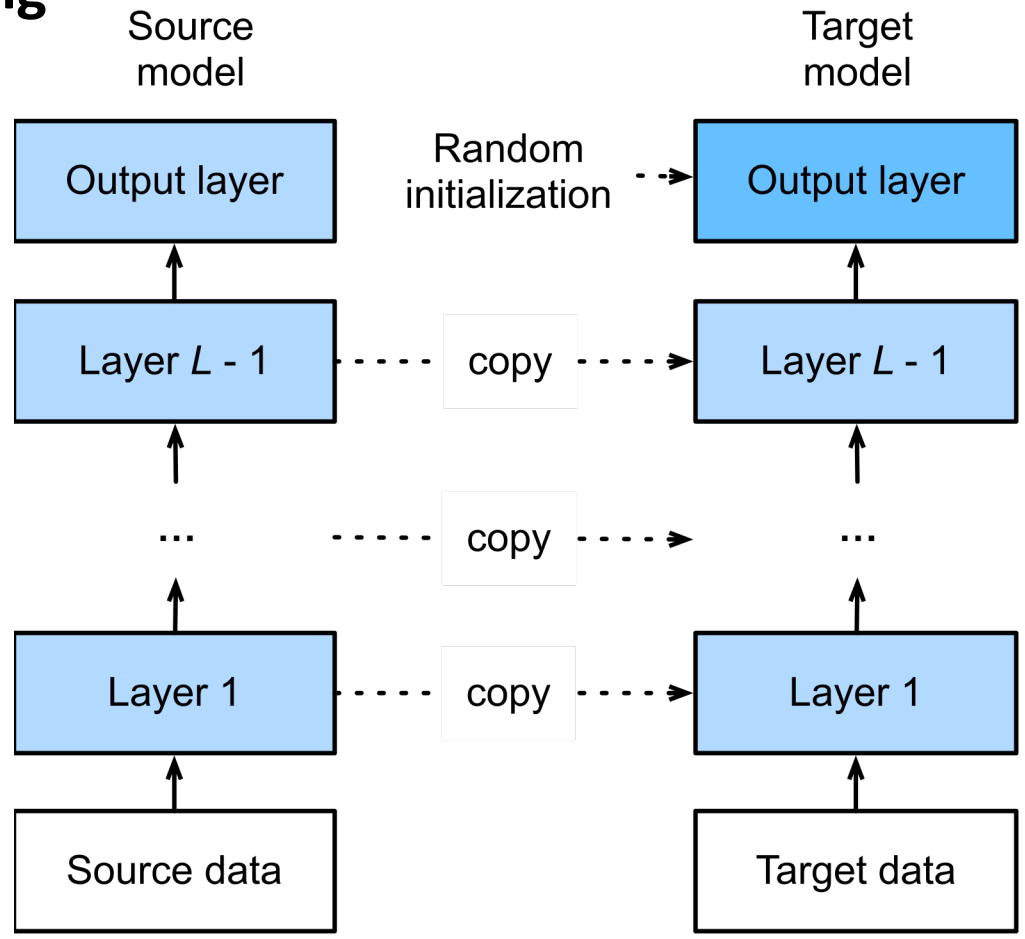
Flower 102



MIT Places

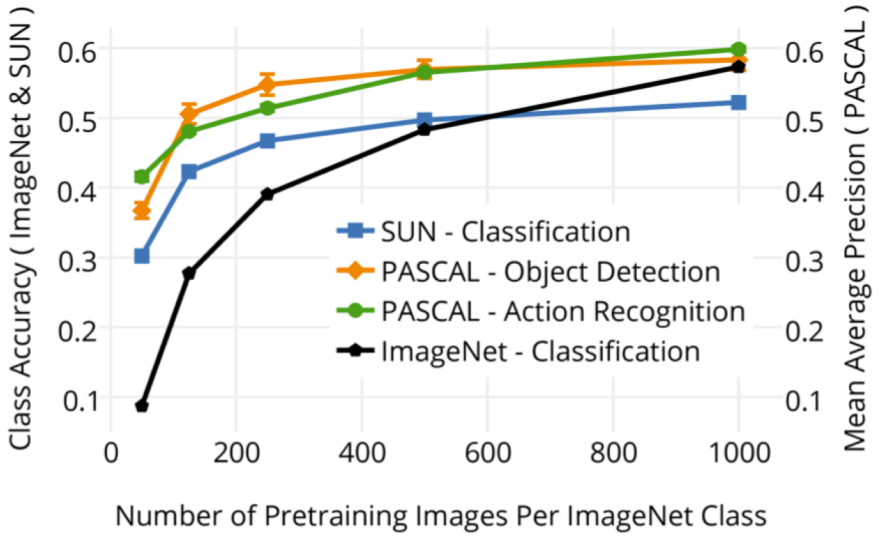


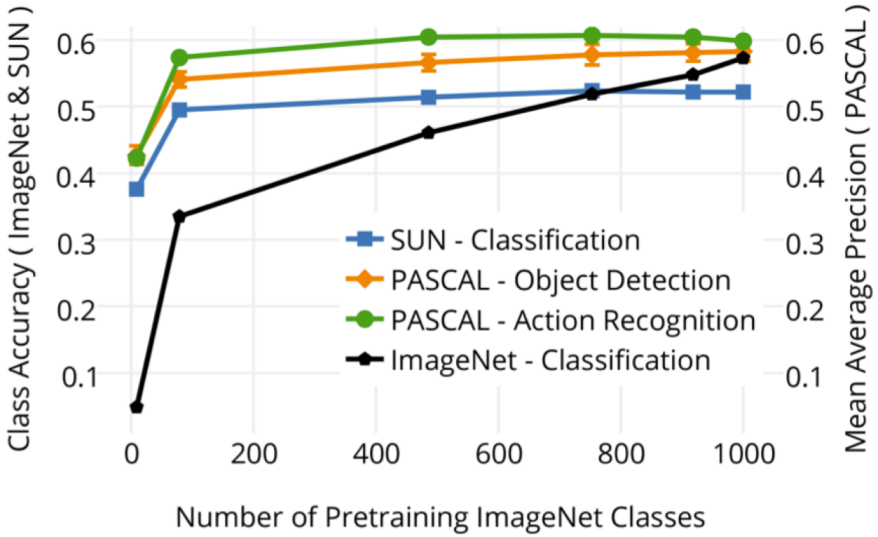
MIT Indoor 67

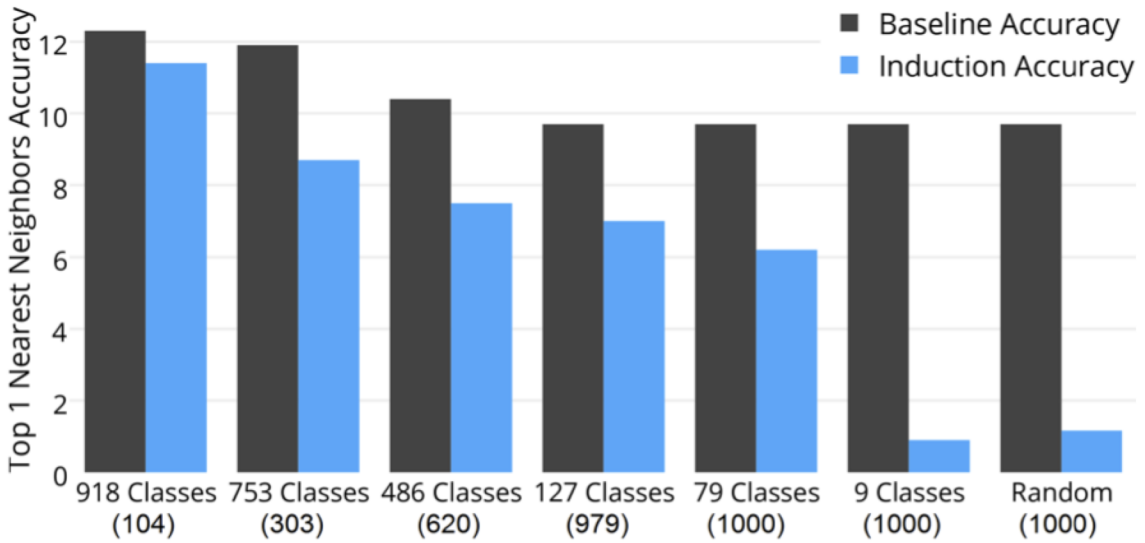


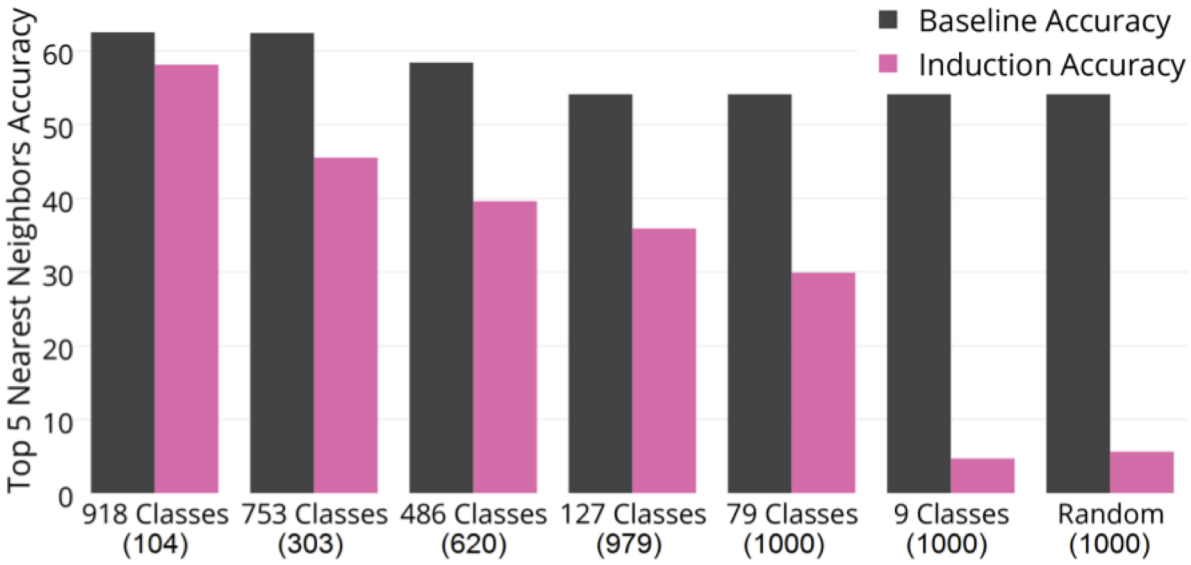
https://d2l.ai/chapter_computer-vision/fine-tuning.html

Empirical Studies - Dataset



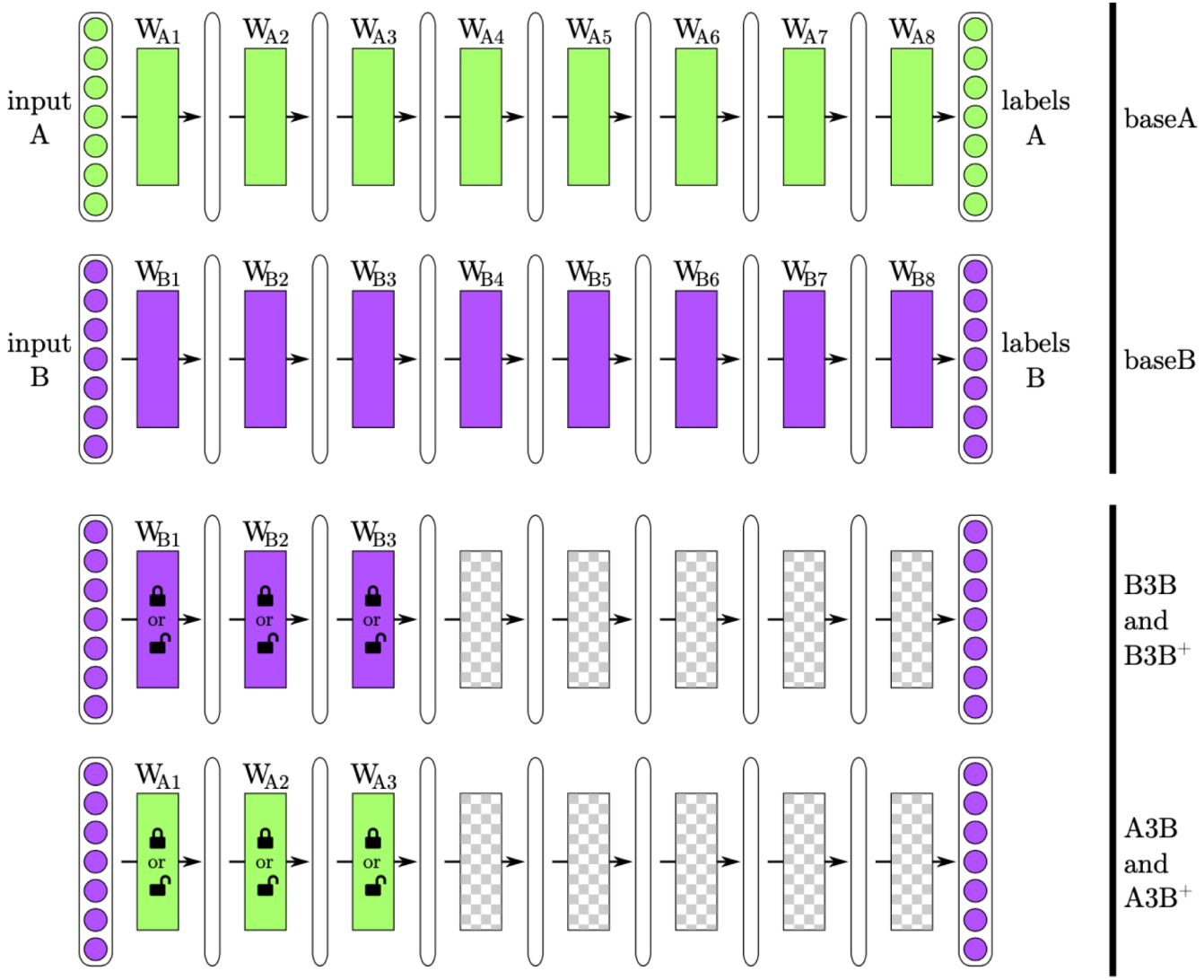


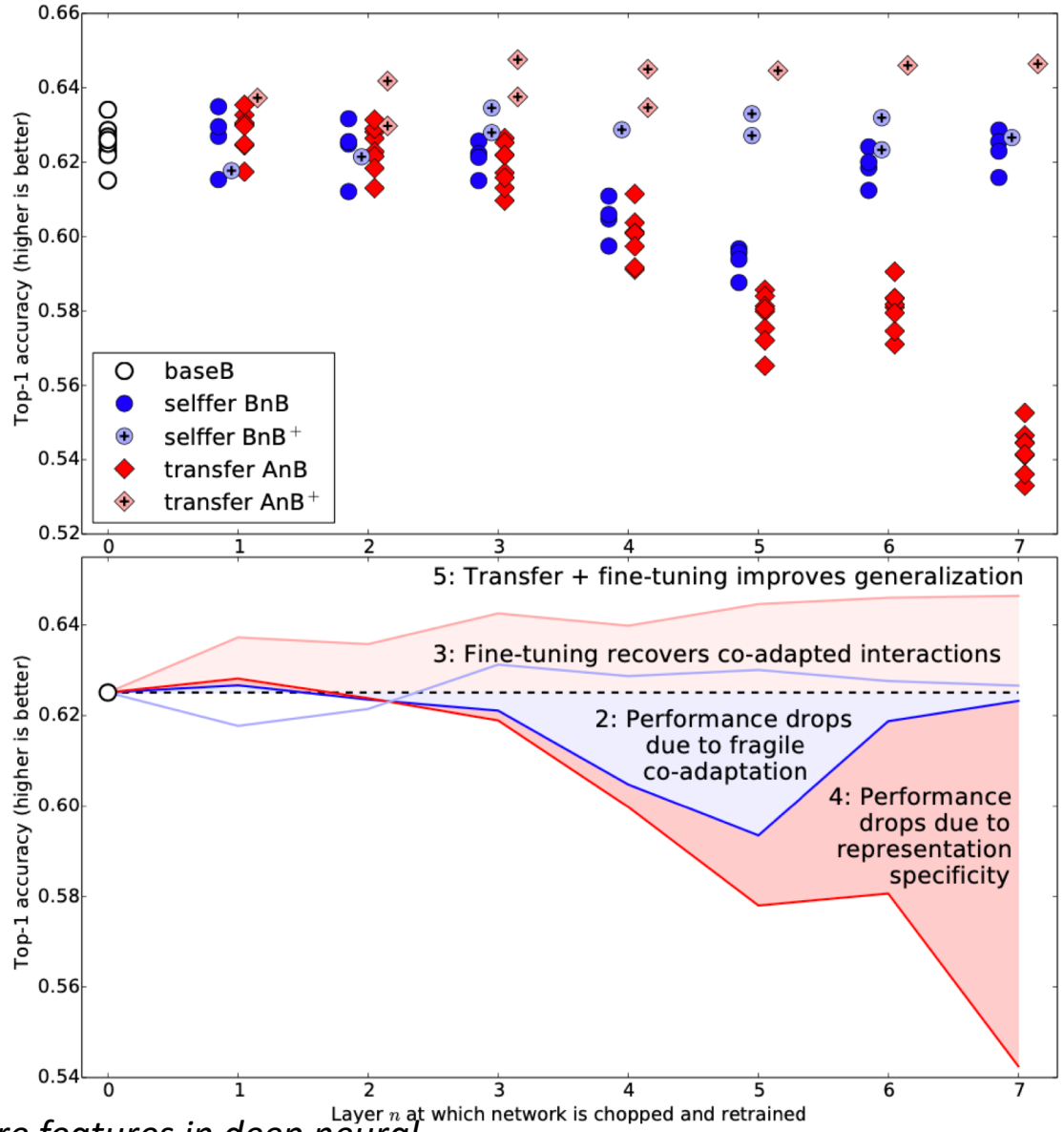




Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? arXiv preprint arXiv:1608.08614, 2016.

Empirical Studies - Architecture





Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In Advances in neural information processing systems, pages 3320–3328, 2014.

Empirical Studies - Training

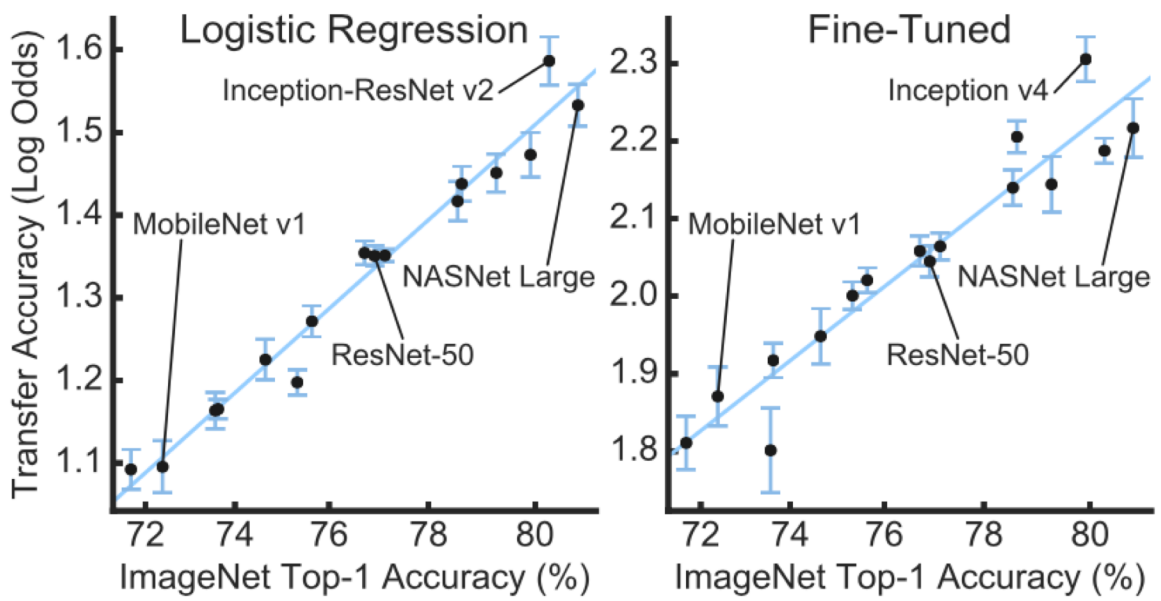


Figure 1. Transfer learning performance is highly correlated with ImageNet top-1 accuracy for fixed ImageNet features (left) and fine-tuning from ImageNet initialization (right). The 16 points in each plot represent transfer accuracy for 16 distinct CNN architectures, averaged across 12 datasets after logit transformation (see Section 3). Error bars measure variation in transfer accuracy across datasets. These plots are replicated in Figure 2 (right).

Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2661–2671, 2019.

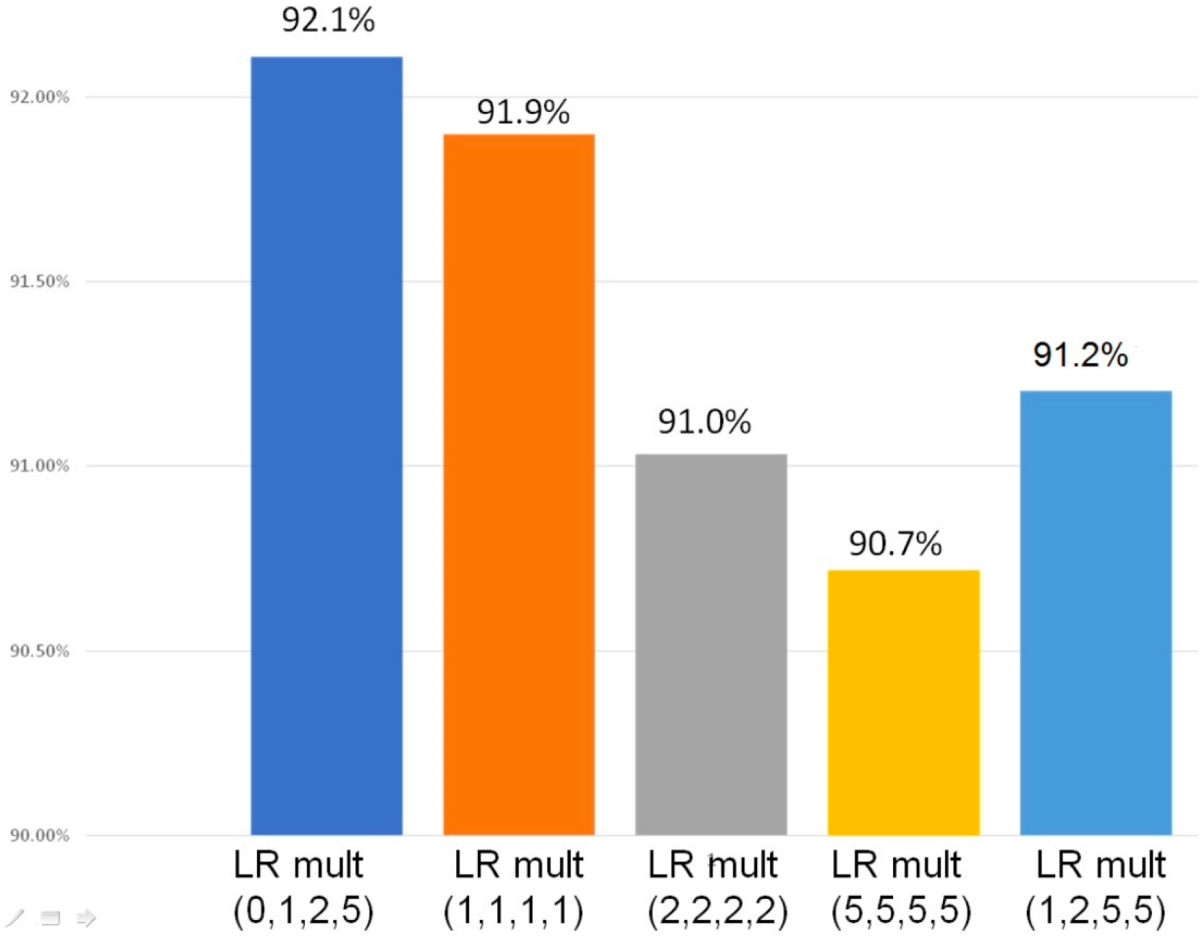
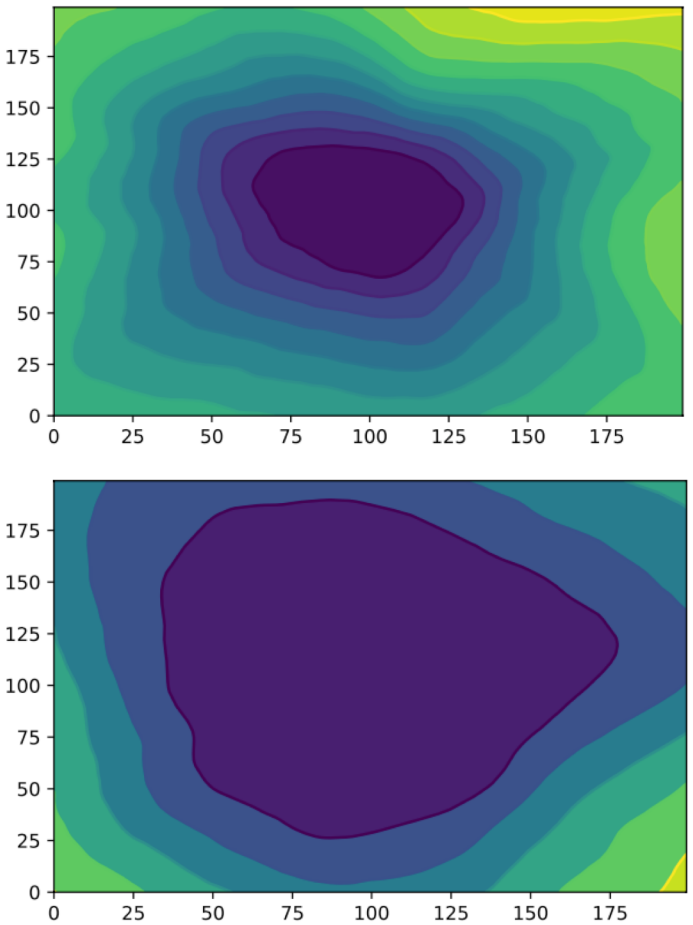


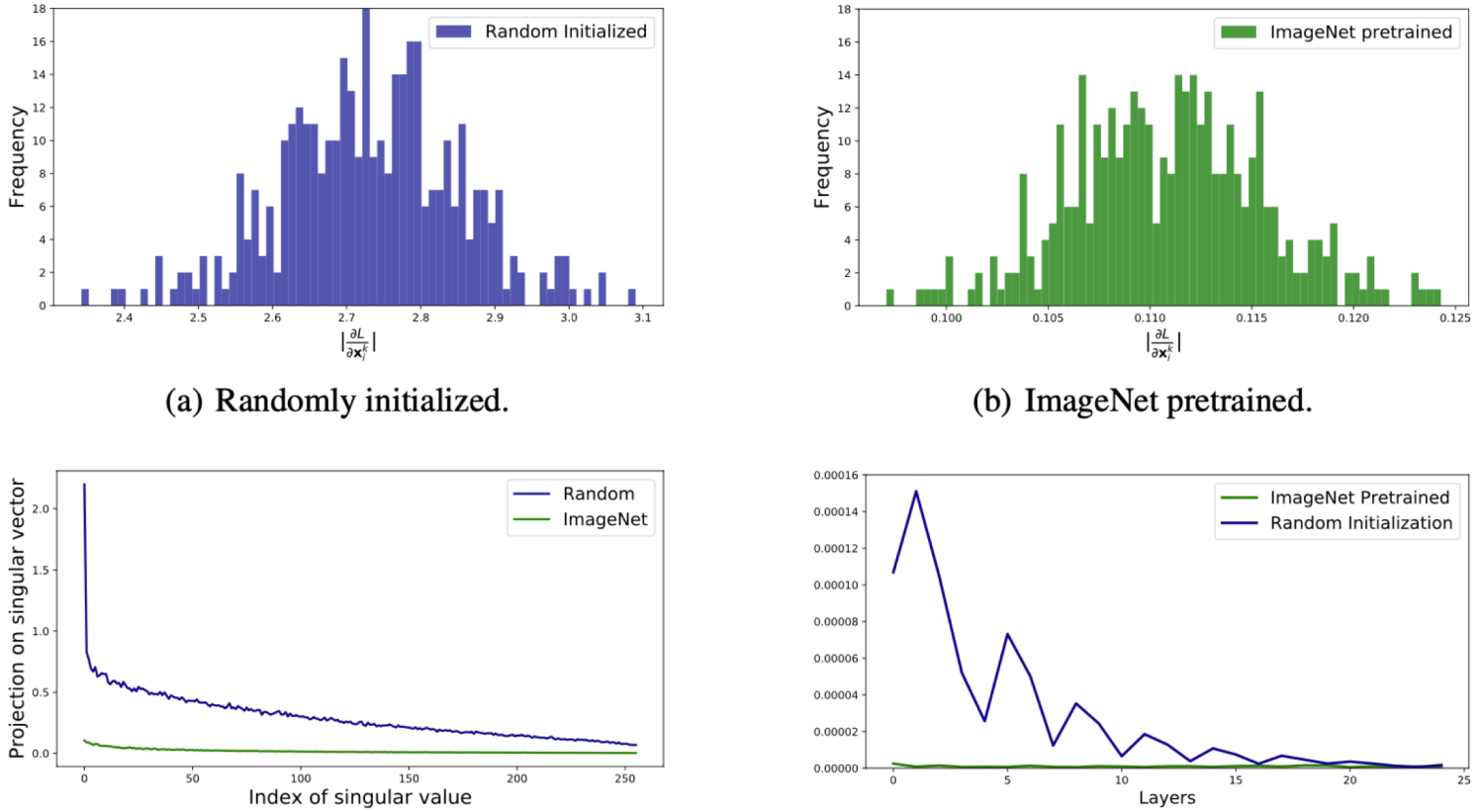
Fig. 10: Top-1 accuracy with varying inner LR mult and fixed outer LR mult at 20

Parijat Dube, Bishwaranjan Bhattacharjee, Elisabeth Petit-Bois, and Matthew Hill. Improving transferability of deep neural networks. arXiv preprint arXiv:1807.11459, 2018

Theoretical Studies - Understanding



Upper: Randomly initialized; Lower: Transfer Learning



(a) Randomly initialized. (b) ImageNet pretrained.

(c) Projection of weight on components of gradient. (d) Scale of gradient in different layers.

Figure 7: The stabilization of gradient by pretrained weights. (a) (b) Distribution of the magnitude of gradient in the 25th layer of ResNet-50. (c) Magnitude of the projection of weight matrices on the singular vectors of gradient. (d) Scaling of the gradient in different layers through back-propagation.

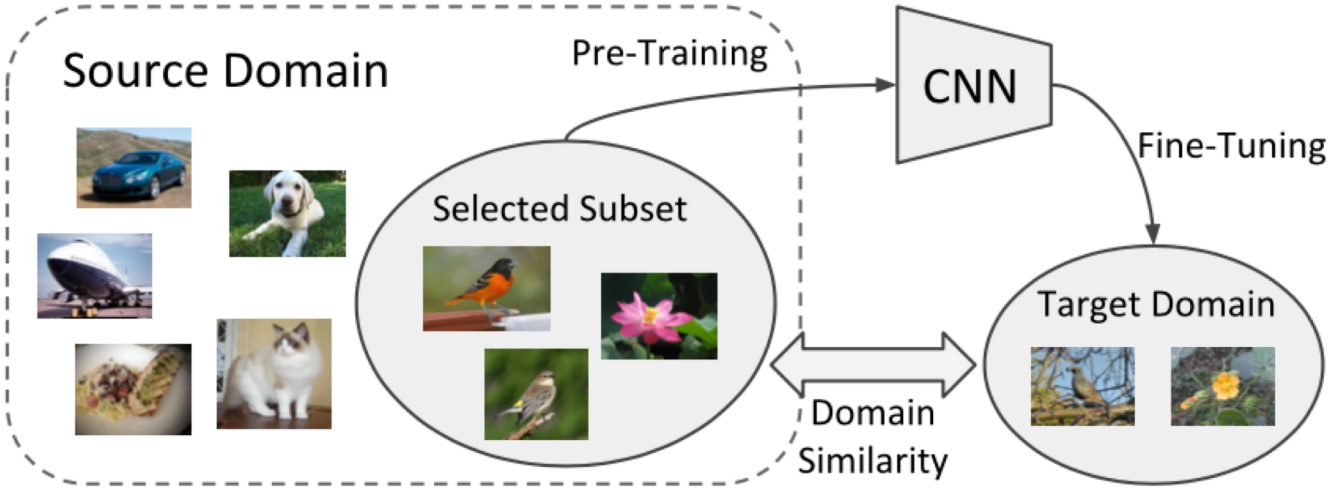
Hong Liu, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Towards understanding the transferability of deep representations. arXiv preprint arXiv:1909.12031, 2019.

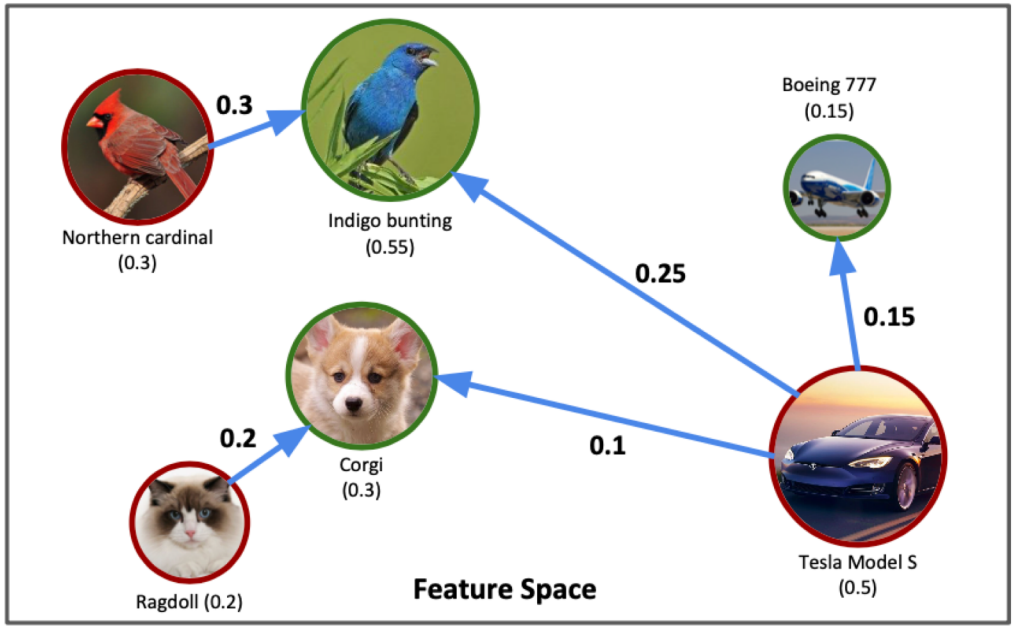
Transfer Learning Algorithms

- Data Transfer
- Weight Transfer
- Feature Transfer
- Meta Transfer

Data Transfer

How to exploit bigger potential of the source dataset ?





	CUB200	Stanford Dogs	Flowers-102	Stanford Cars	Aircraft	Food101	NABirds
ImageNet	82.84	84.19	96.26	91.31	85.49	88.65	82.01
iNat	89.26	78.46	97.64	88.31	82.61	88.80	87.91
ImageNet + iNat	85.84	82.36	97.07	91.38	85.21	88.45	83.98
Subset A (832-class)	86.37	84.69	97.65	91.42	86.28	88.78	84.79
Subset B (585-class)	88.76	85.23	97.37	90.58	86.13	88.37	87.89

Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4109–4118, 2018.

Data Transfer

How to exploit bigger potential of the source dataset ?

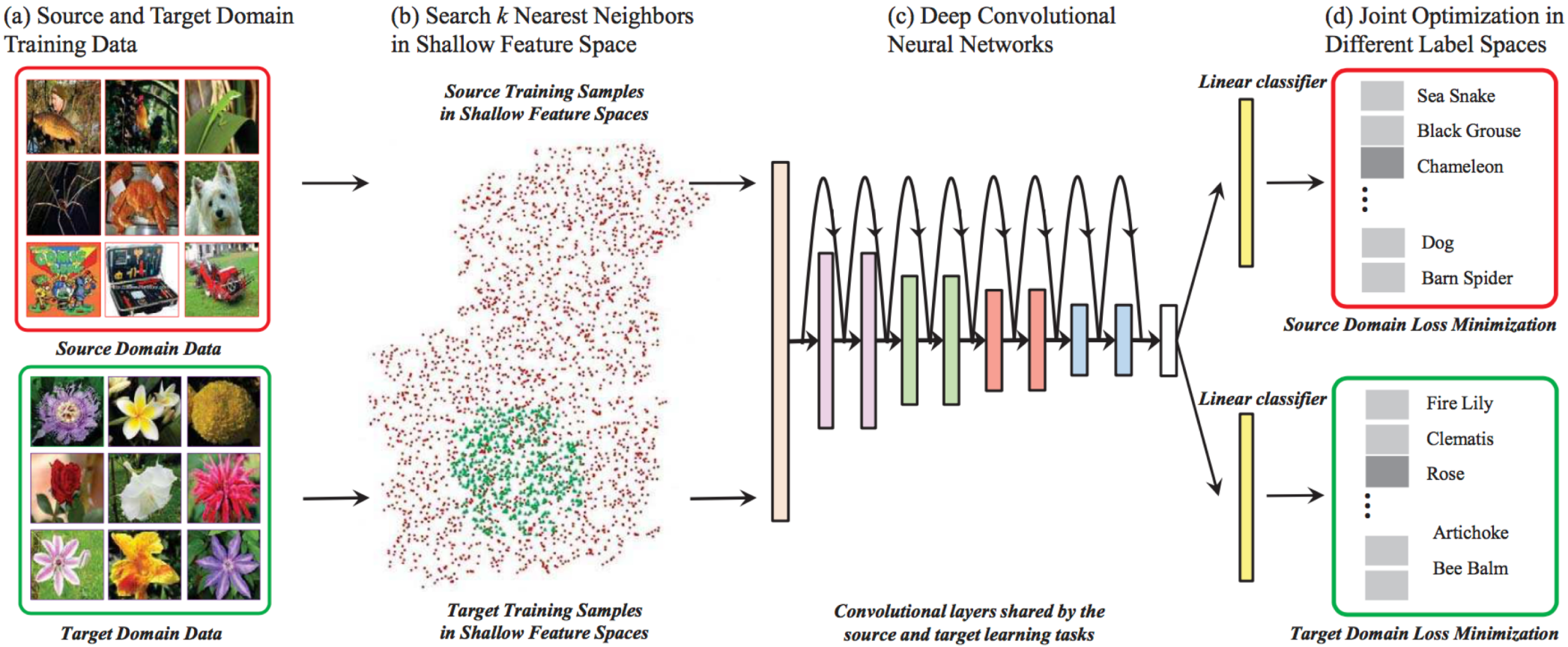


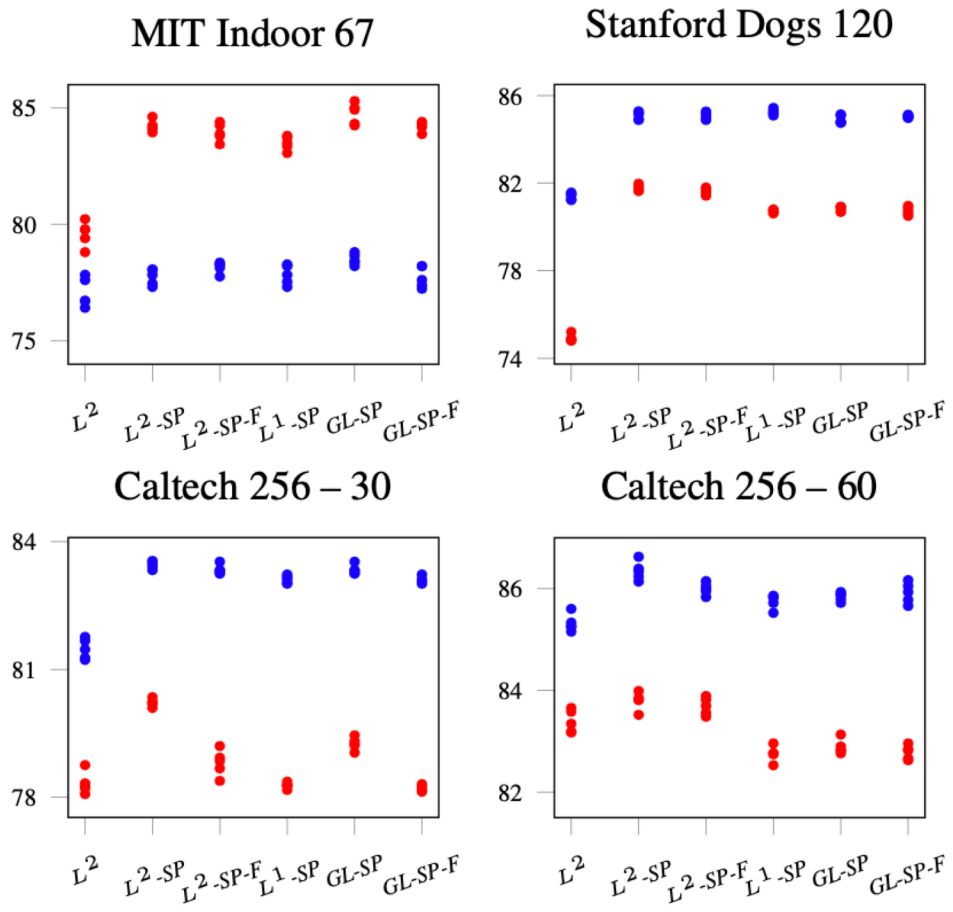
Figure 1. Pipeline of the proposed selective joint fine-tuning. From left to right: (a) Datasets in the source domain and the target domain. (b) Select nearest neighbors of each target domain training sample in the source domain via a low-level feature space. (c) Deep convolutional neural network initialized with weights pre-trained on ImageNet or Places. (d) Jointly optimize the source and target cost functions in their own label spaces.

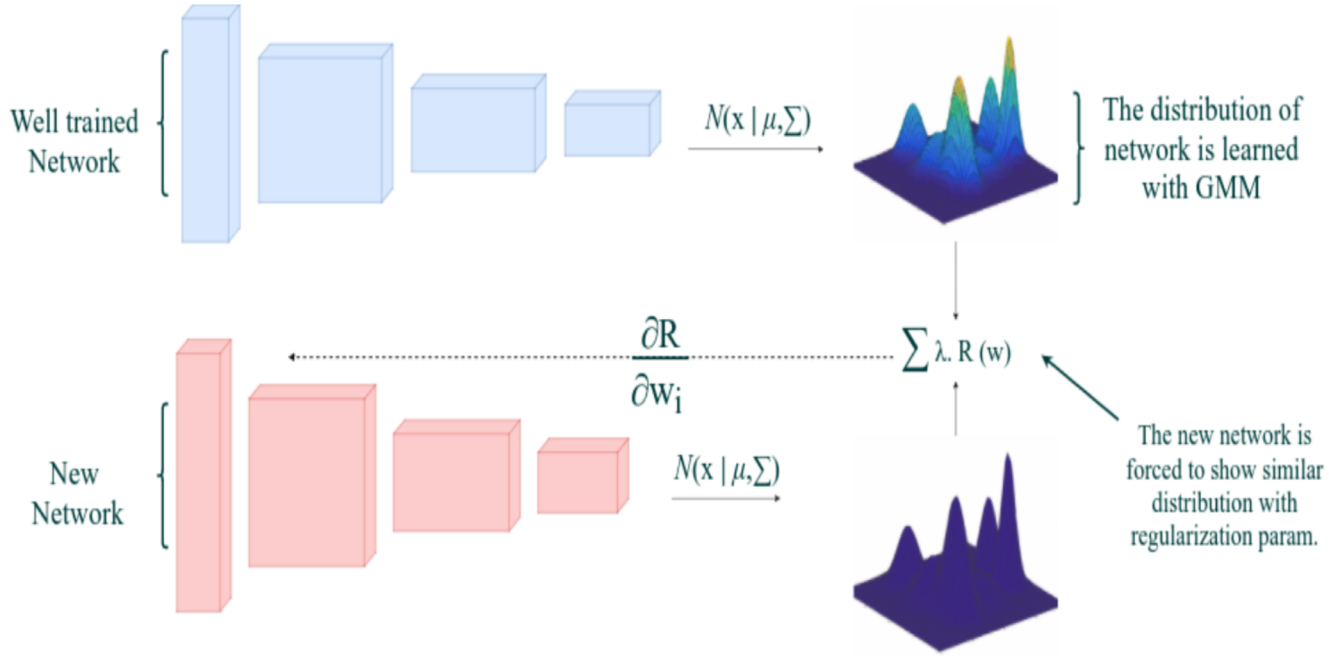
Ge W, Yu Y. Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning

Weight Transfer

How to extract dark knowledge in pre-trained weights

L2-SP $\Omega(w) = \frac{\alpha}{2}\left\|w_S - w_S^0\right\|_2^2 + \frac{\beta}{2}\left\|w_{\bar{S}}\right\|_2^2$

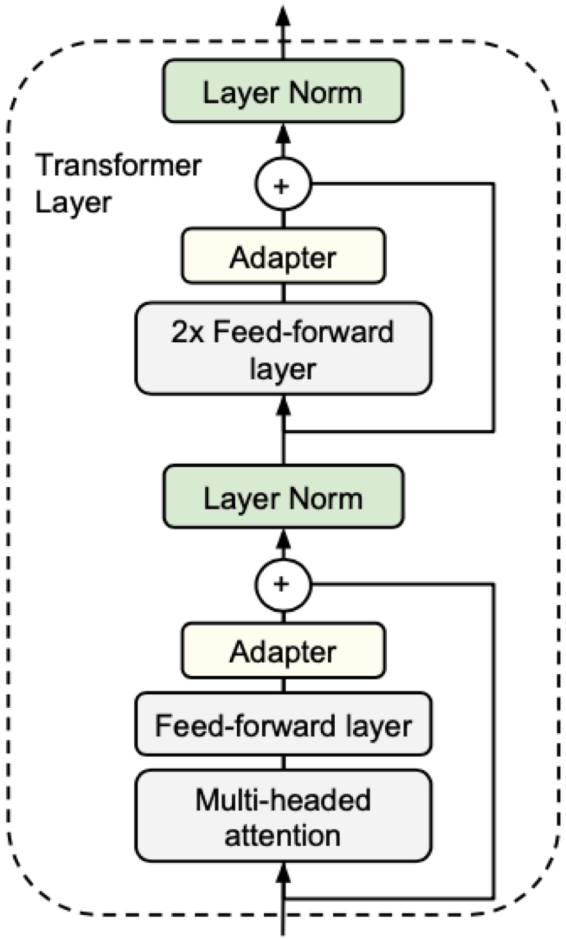


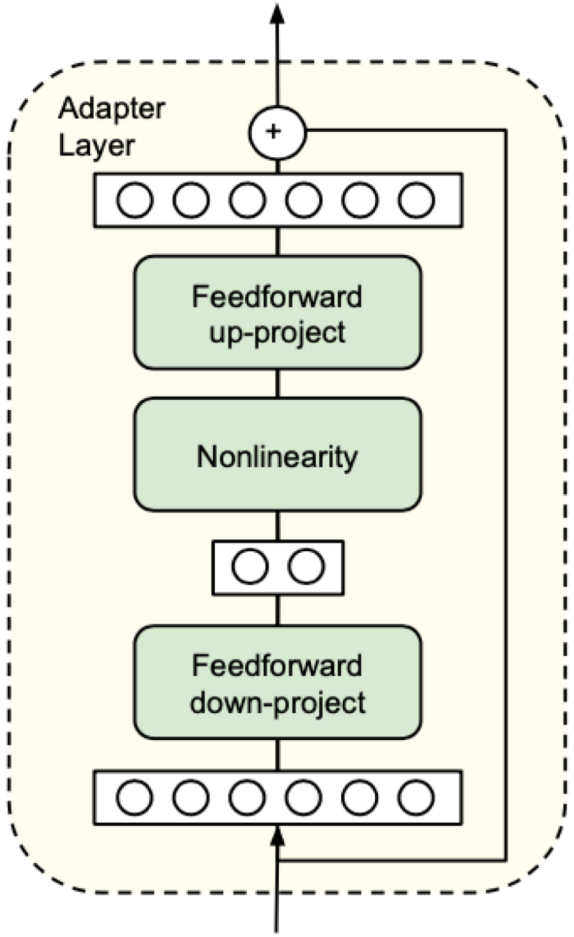


Li X, Grandvalet Y, Davoine F. Explicit inductive bias for transfer learning with convolutional networks

Aygun M, Aytar Y, Kemal Ekenel H. Exploiting convolution filter patterns for transfer learning

Weight Transfer





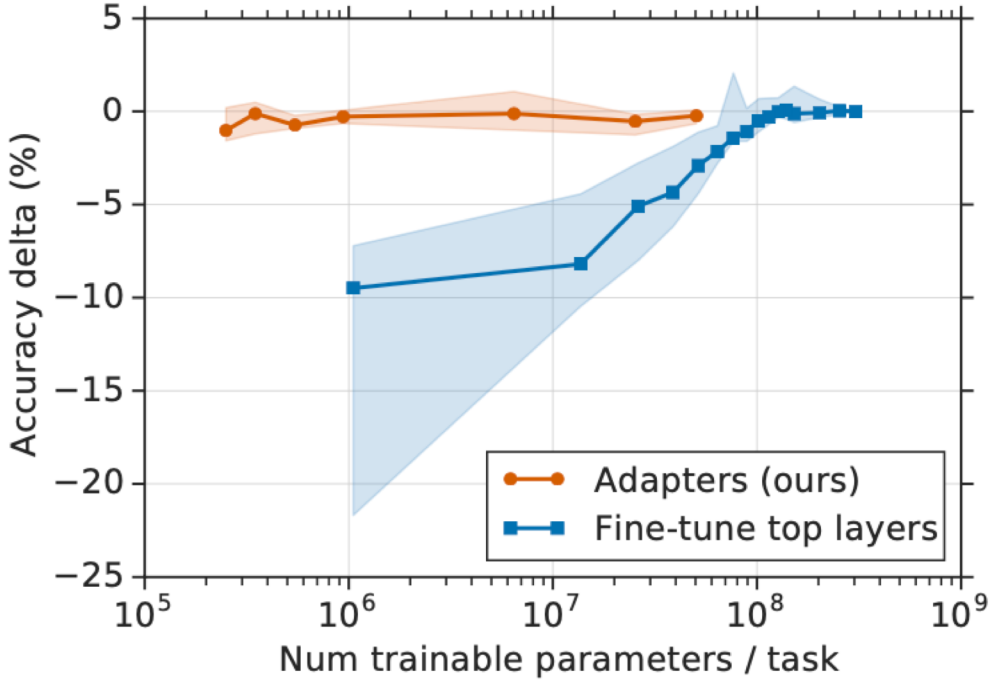


Figure 1. Trade-off between accuracy and number of trained task-specific parameters, for adapter tuning and fine-tuning. The *y*-axis is normalized by the performance of full fine-tuning, details in Section 3. The curves show the 20th, 50th, and 80th performance percentiles across nine tasks from the GLUE benchmark. Adapter-based tuning attains a similar performance to full fine-tuning with two orders of magnitude fewer trained parameters.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Syl- vain Gelly. Parameter-efficient transfer learning for nlp. arXiv preprint arXiv:1902.00751, 2019.

Feature Transfer

Knowledge Distillation on Features

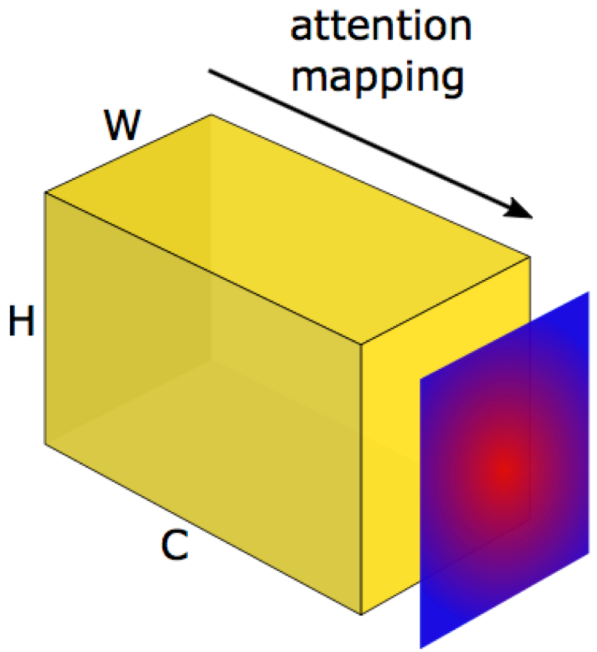
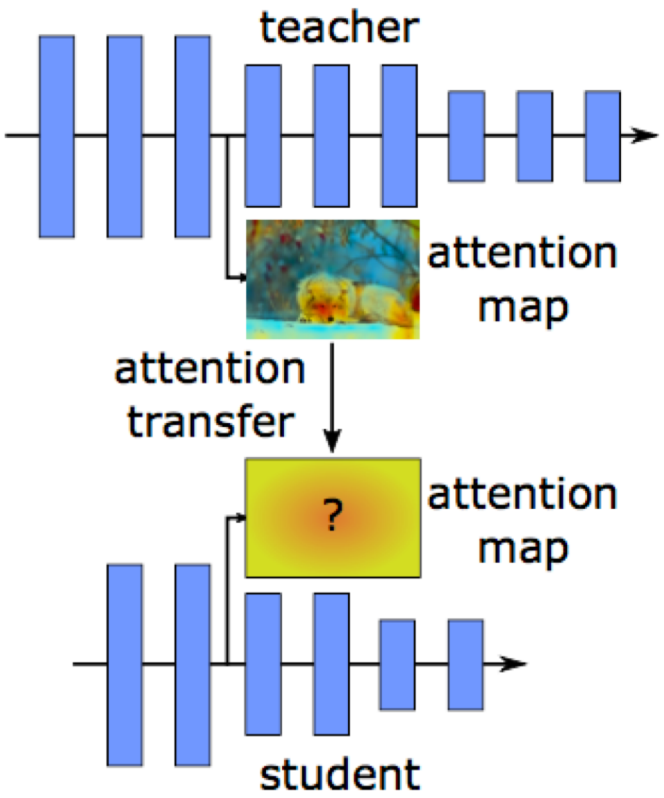


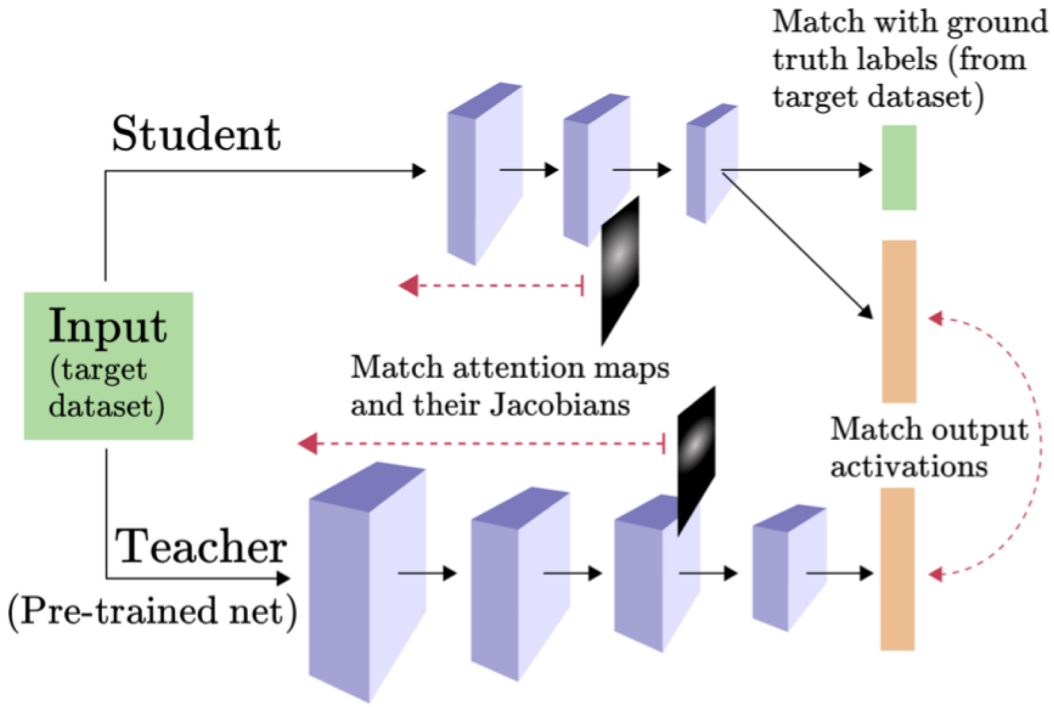
Figure 3: Attention mapping over feature dimension.



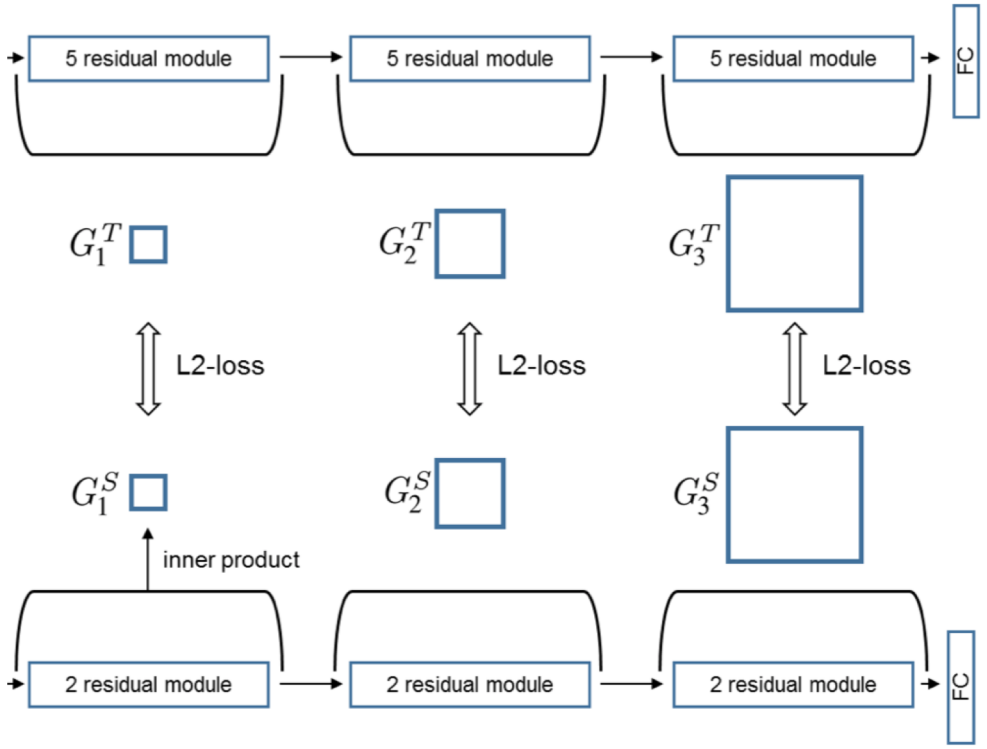
Zagoruyko S, Komodakis N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer

Feature Transfer

Knowledge Distillation on Features



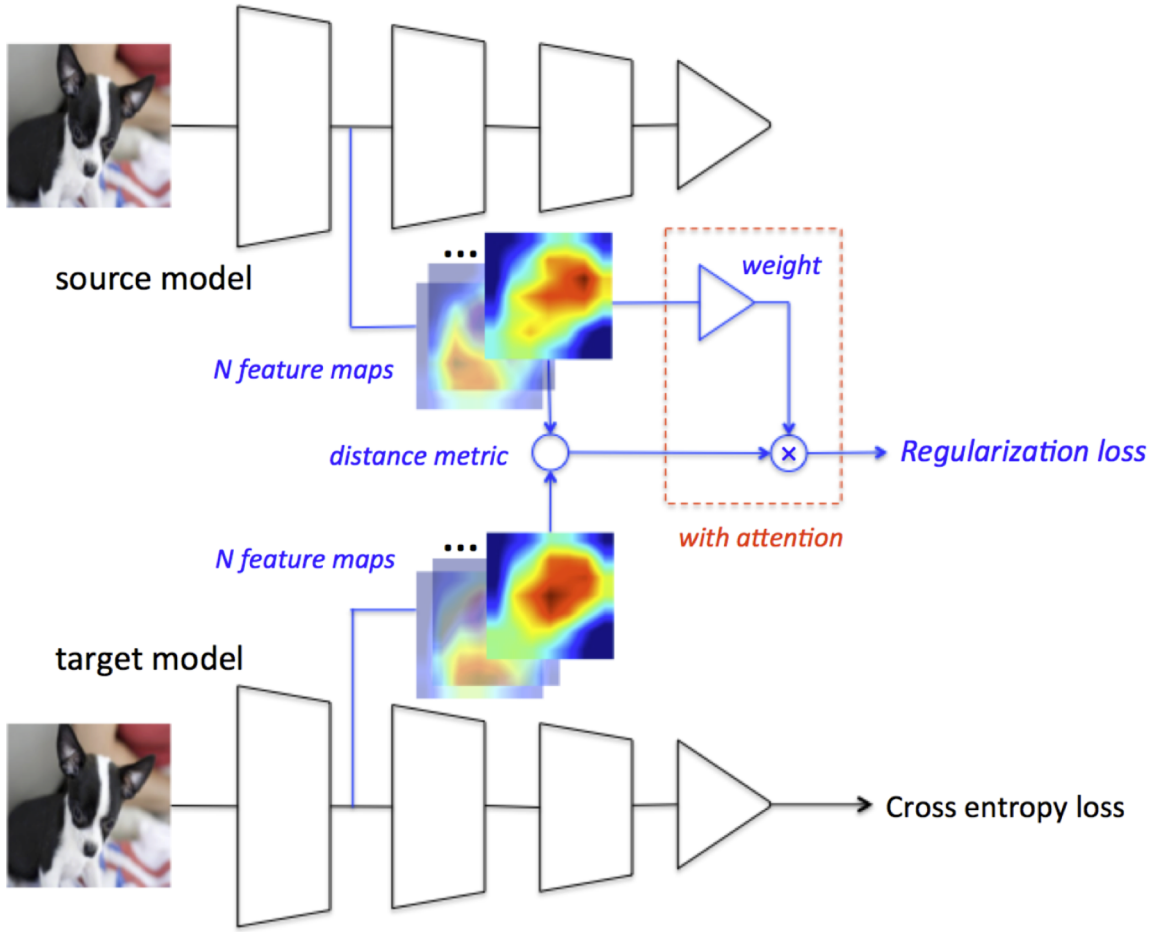
Srinivas S, Fleuret F. Knowledge transfer with jacobian matching

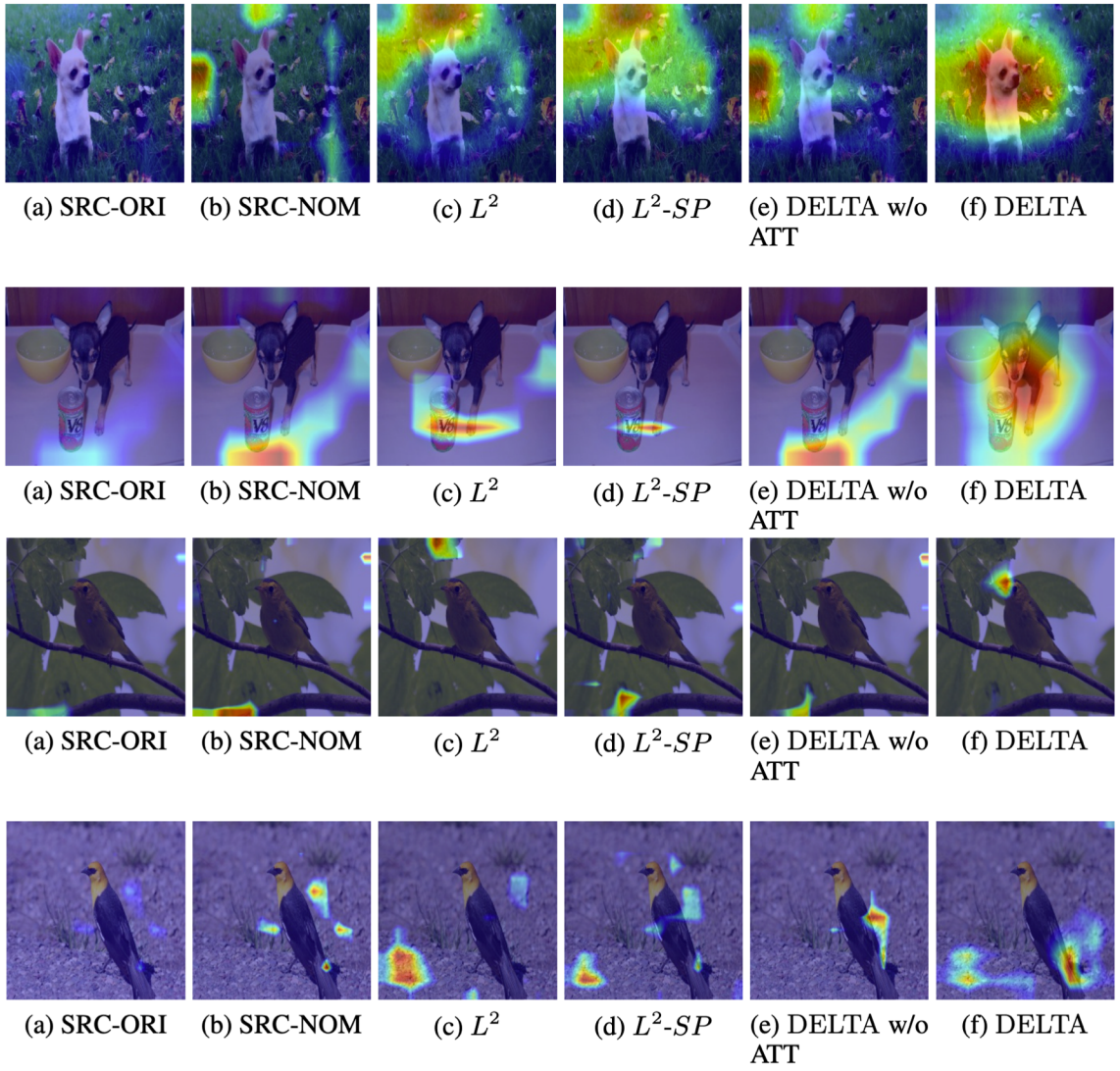


Yim J, Joo D, Bae J, et al. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning

Feature Transfer

Preserving knowledge by middle feature alignment





Xingjian Li, Haoyi Xiong, Hanchao Wang, Yuxuan Rao, Liping Liu, and Jun Huan. DELTA: deep learning transfer using feature map with attention for convolutional networks. International Conference on Learning Representations, 2019.

Feature Transfer

Discussion

About 90% parameter vectors of DELTA have larger distance from the stating point than L2-SP, and a small number of filters is driven very far away from the initial value (as shown at the left end of the curves in Figure 3).

Using attention, we allow “unactivated” convolution filters to be re-used for better image classification.

Effect of “unactivated channel re-usage”

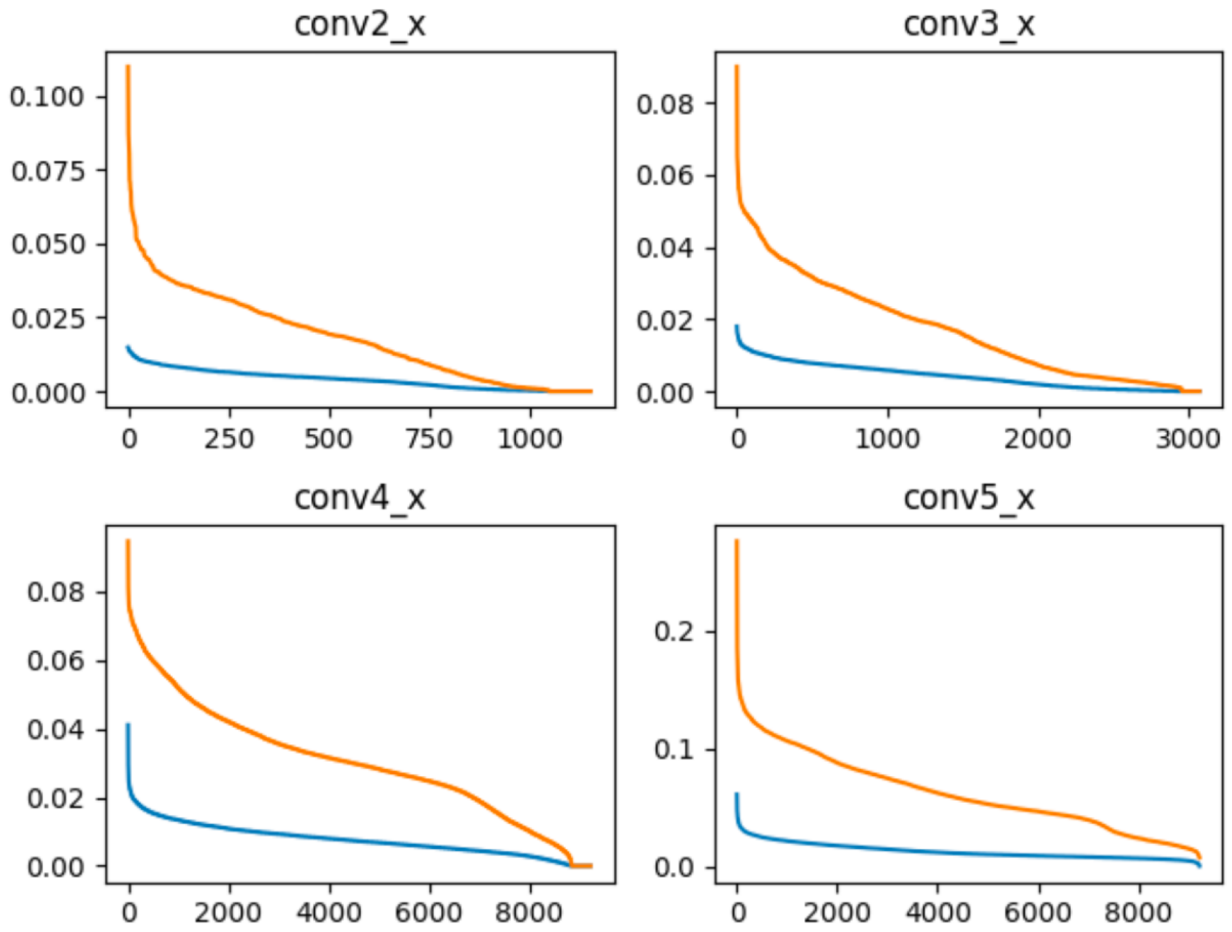
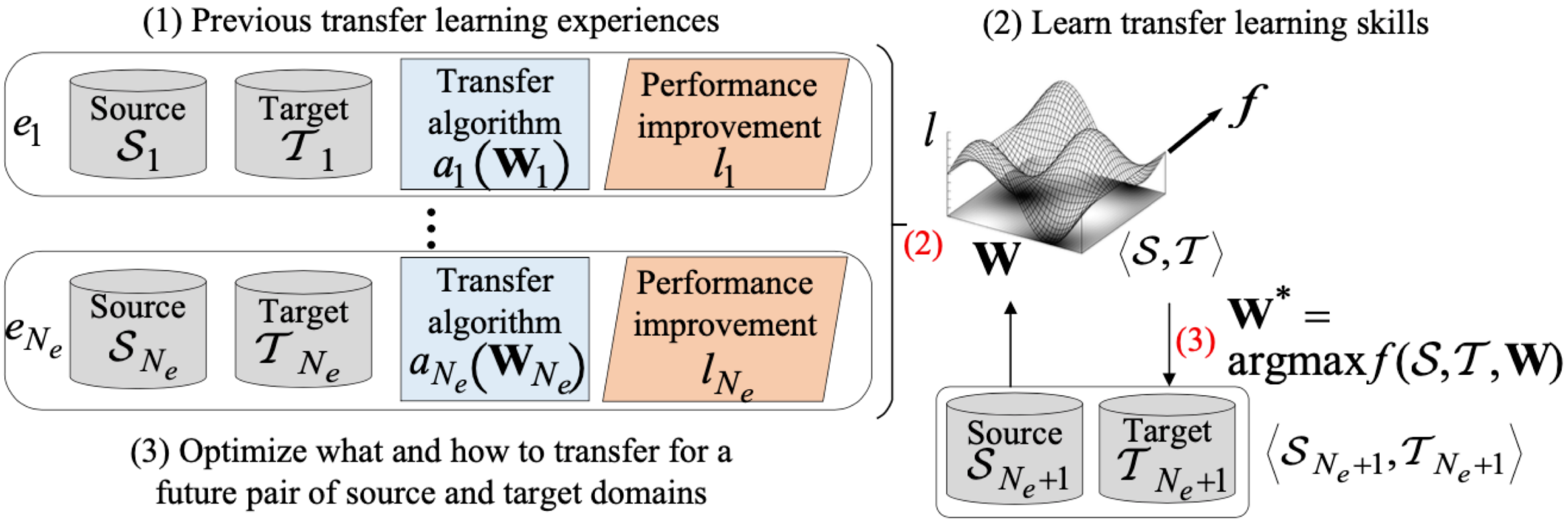


Figure 3: Distribution of the distance of parameters from the starting point. In ResNet-101, conv2_x, conv3_x, conv4_x, conv5_x represent for four main stages each of which has stacked convolution layers. The blue line represents for the result of L^2-SP, and the orange line for DELTA.

Meta Transfer

Learning to choose algorithms based on historical experiences



Ying, W., Zhang, Y., Huang, J., & Yang, Q. Transfer learning via learning to transfer

Meta Transfer

Learning to adapt algorithms based on historical training steps

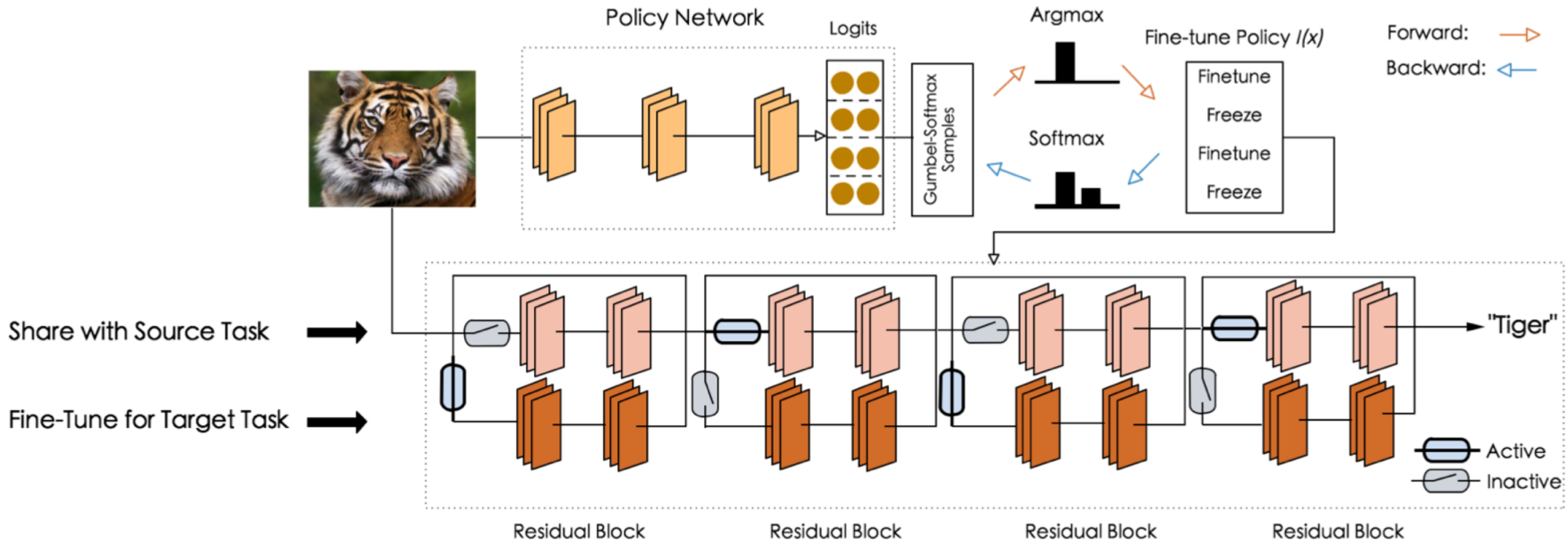


Figure 2. Illustration of our proposed approach. The policy network is trained to output routing decisions (fine-tune or freeze parameters) for each block in a ResNet pre-trained on the source dataset. During learning, the fine-tune vs. freeze decisions are generated based on a Gumbel Softmax distribution, which allows us to optimize the policy network using backpropagation. At test time, given an input image, the computation is routed so that either the fine-tuned path or the frozen path is activated for each residual block.

Yunhui Guo, Honghui Shi, Abhishek Kumar, Kristen Grauman, Tajana Rosing, and Rogerio Feris. Spottune: transfer learning through adaptive fine-tuning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4805–4814, 2019.

AutoDL Transfer Application

- AutoDL Transfer has been used in multiple Baidu products and services
- Over 1 million calls per month

EasyDL
http://ai.baidu.com/ezdl/

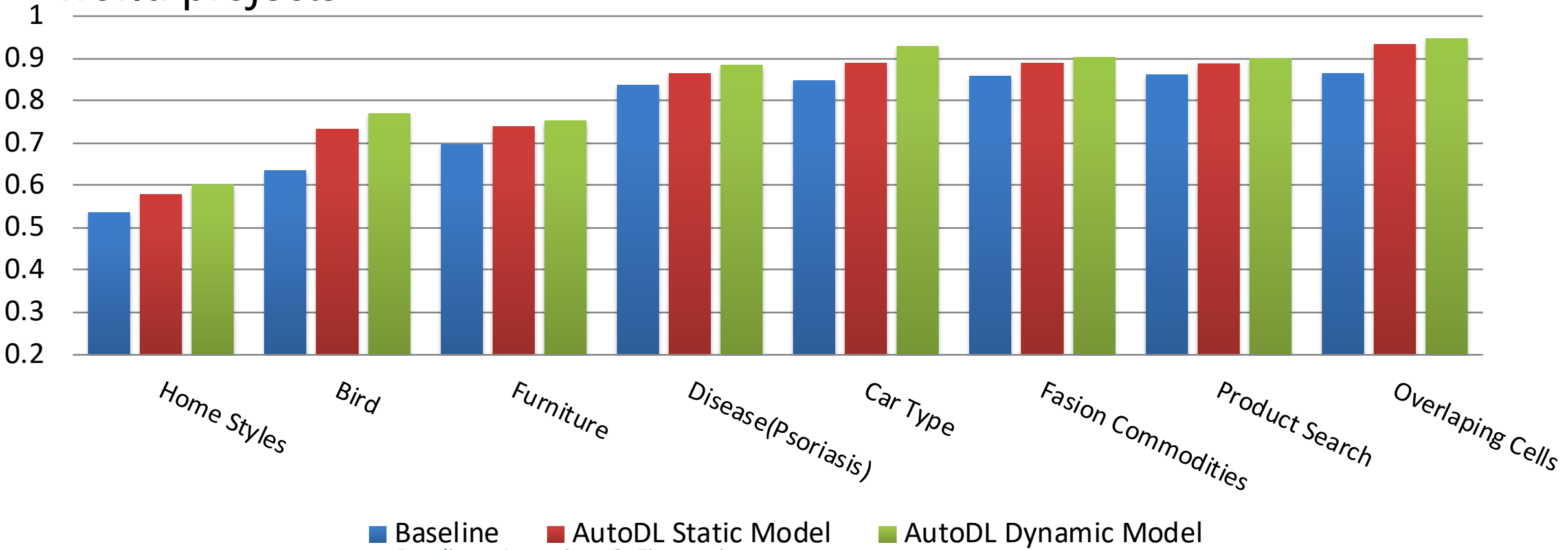
AIStudio
http://aistudio.baidu.com

Jarvis
http://di.baidu.com/product/jarvis?castk=LTE%3D#solution-sec0

Baidu Cloud ML
https://console.bce.baidu.com/bml/#/bml/autoDL

AutoDL Transfer Application

- AutoDL-Transfer achieves significant improvements on numerous real world projects



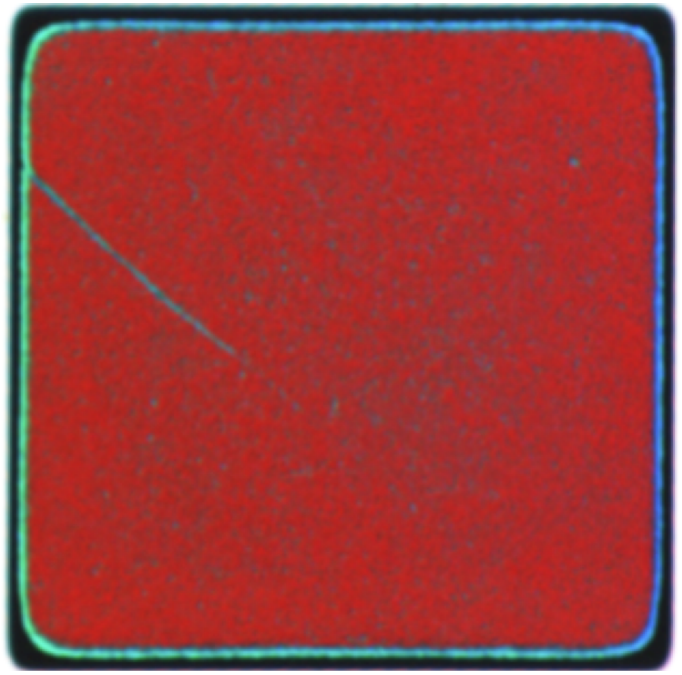
Baseline： Inception v3+Fine-tuning

AutoDL Static Model： AutoDL searches a model architecture from one dataset and then directly use it on another dataset

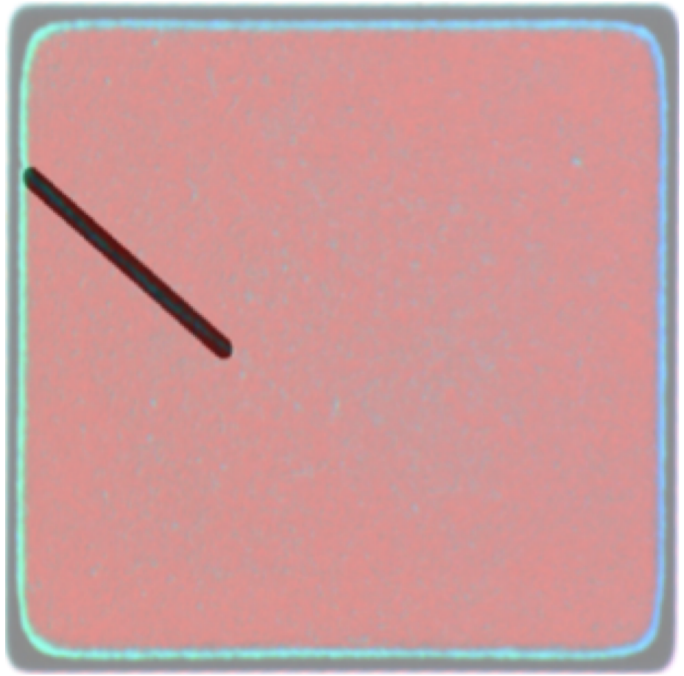
AutoDL Dynamic Model： AutoDL searches the best model on each dataset

AutoDL Application I: Semantic Analysis

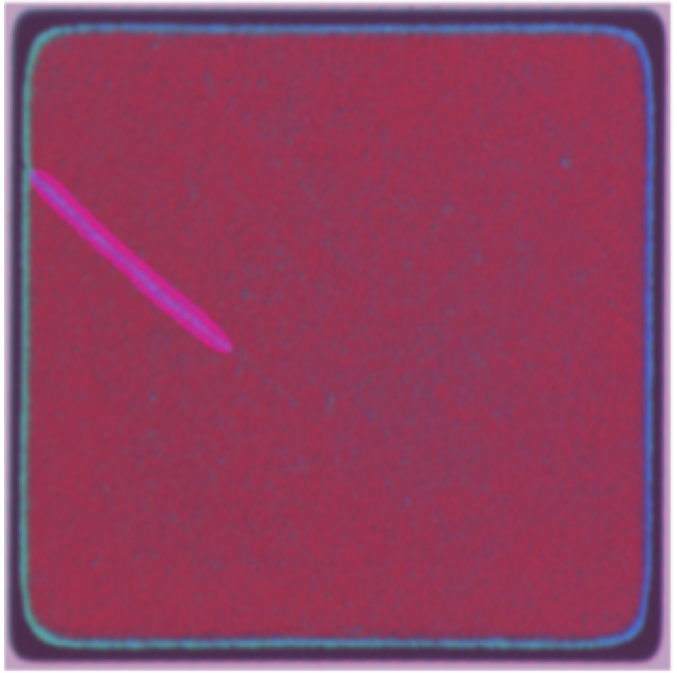
- Industrial production: Products defects labeling (Semantic analysis task)
- AutoDL improves mean IOU from 0.66 to 0.69



Original Sample



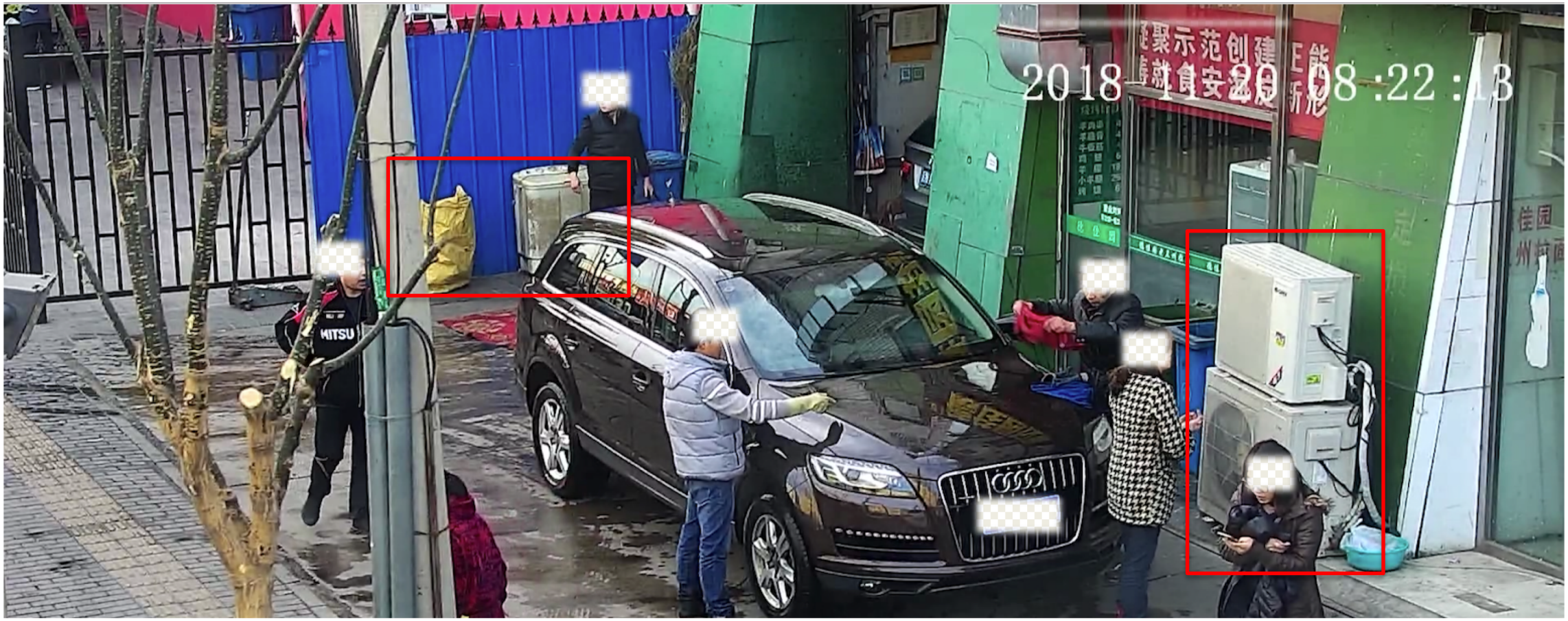
Labeled by Human



Labeled by AutoDL

AutoDL Application II Object Detection

- Smart City: Identify improperly disposed items (Object Detection)
- AutoDL improves model's mAP from 0.41 to 0.48



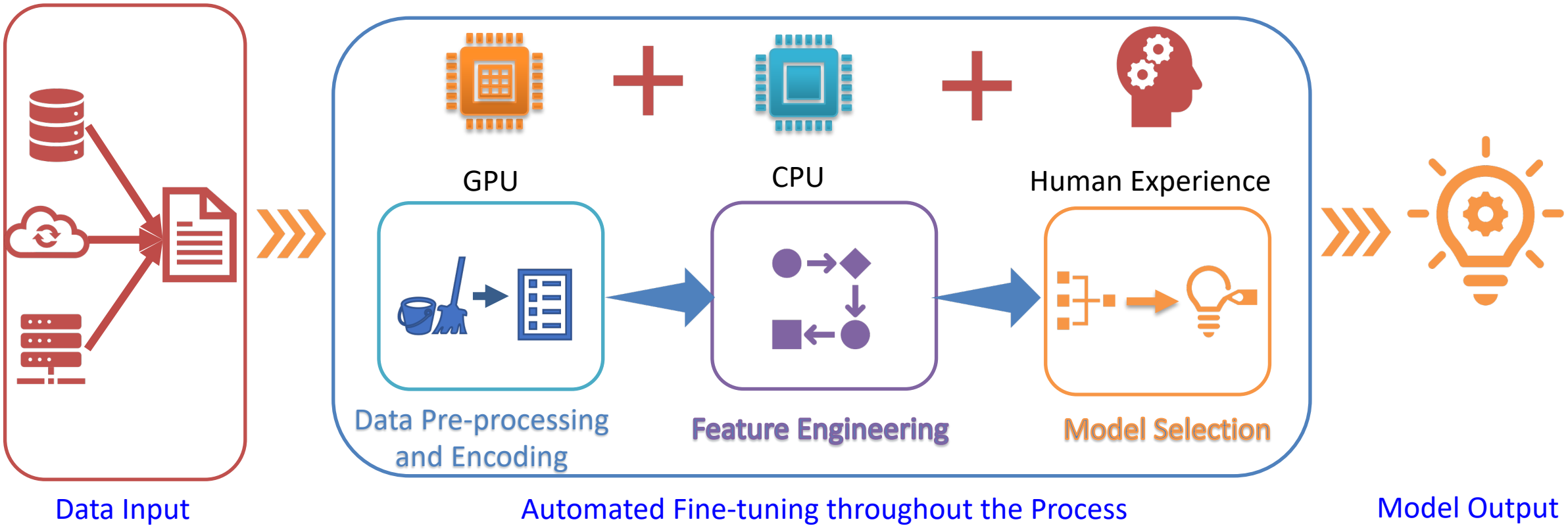
AutoDL Applicaiton III ：Fine-tuning Hyperparameters in NLP Transfer Learning

- AutoDL Finetuner
  - Used for transfer learning tasks, where the fine-tuning of 12 sets of crucial hyperparameters is automated
  - Based on Baidu’s Blackbox optimization algorithm P-SHE2[1], and already made available on Padddle Hub

Sentimental Analysis in Chinese	Precision	Resources Needed
Fine-tuning by Human Experts	95.4%	-
Random Search	94.9%	40 GPU Hours
AutoDL Finetuner	95.2%	40 GPU Hours

1. Gradient-Free Learning with a Continuous-time Approach, under review

AutoDL Application IV : In Structured Data



- **Automated fine-tuning throughout the process**
- **Supports 48 types of common data pre-processing methods**

Case Study: A client used AutoDL on a loan risk control model, whose AUC was improved from 0.6265 to 0.6504

Thank you

- For more information:
  - Dr. Dejing Dou, Head of Big Data Lab, Baidu Research
  - Email: doudejing@baidu.com