Large Scale Fine-Grained Categorization and Domain-Specific Transfer Learning

Yin Cui[1,2], Yang Song[3], Chen Sun[3], Andrew Howard[3], Serge Belongie[1,2]

[1] Department of Computer Science, Cornell University [2] Cornell Tech [3] Google AI

Introduction

Fine-Grained Visual Categorization (FGVC)

- On large-scale dataset: little prior work.
- On small-scale dataset: fine-tuning a network from ImageNet pre-training.

Contributions

- A simple training scheme for large-scale FGVC.
  - Best performance on iNaturalist 2017.
- A measure to quantify domain similarity.
- We demonstrate higher domain similarity leads to better transfer learning performance.
  - Better than ImageNet pre-training.
  - SOTA on 7 popular small-scale FGVC datasets.

Large-Scale FGVC - Image Resolution

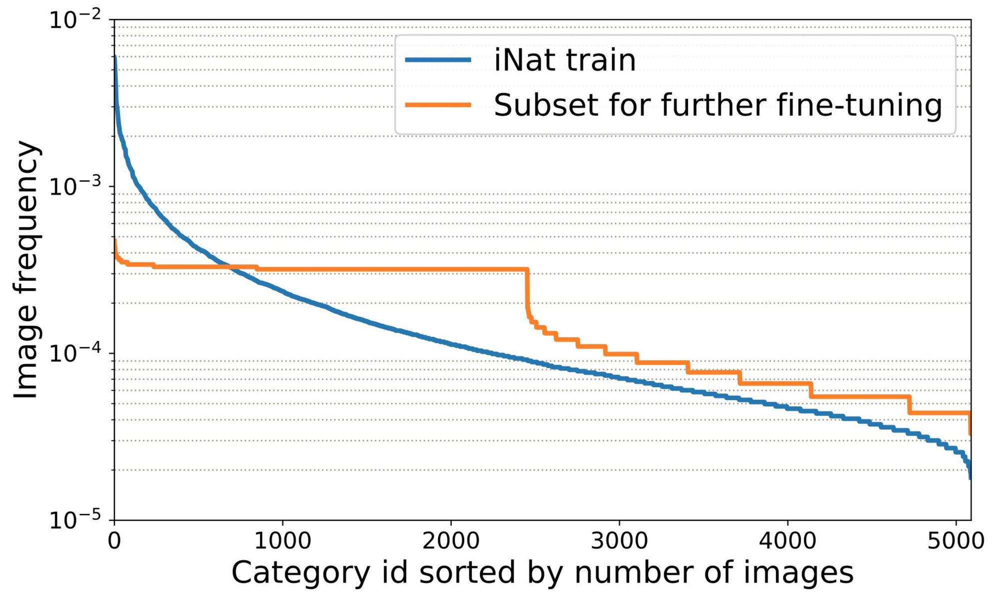
Input Res.	Networks
224 × 224	AlexNet [33], VGGNet [48], ResNet [20]
299 × 299	Inception [51, 52, 50]
320 × 320	ResNetv2 [21], ResNeXt [61], SENet [23]
331 × 331	NASNet [72]

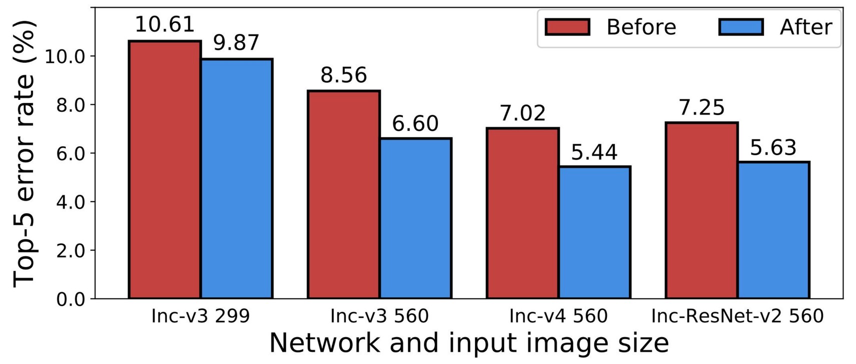
- Why not higher? Heavily tuned for ImageNet:
  - Most ImageNet images are 500 x 375.
  - MAX center crop size = 375 x 0.875 = 328.
- Higher resolution → Richer information and details that are especially important for FGVC.
- We show higher input resolution (e.g., 448, 560) leads to significant improvement on iNaturalist.

	Inc-v3 299	Inc-v3 448	Inc-v3 560
Top-1 (%)	29.93	26.51	25.37
Top-5 (%)	10.61	9.02	8.56

Large-Scale FGVC - Long-Tailed Distribution

- Real-world fine-grained datasets are long-tailed:
  - Few classes have most data, whereas most classes have few data.
- How to deal with the long-tail? Two-stage training:
  1. Train on the original dataset for feature learning.
  2. Fine-tune on a balanced subset for transferring from head classes to tail classes.



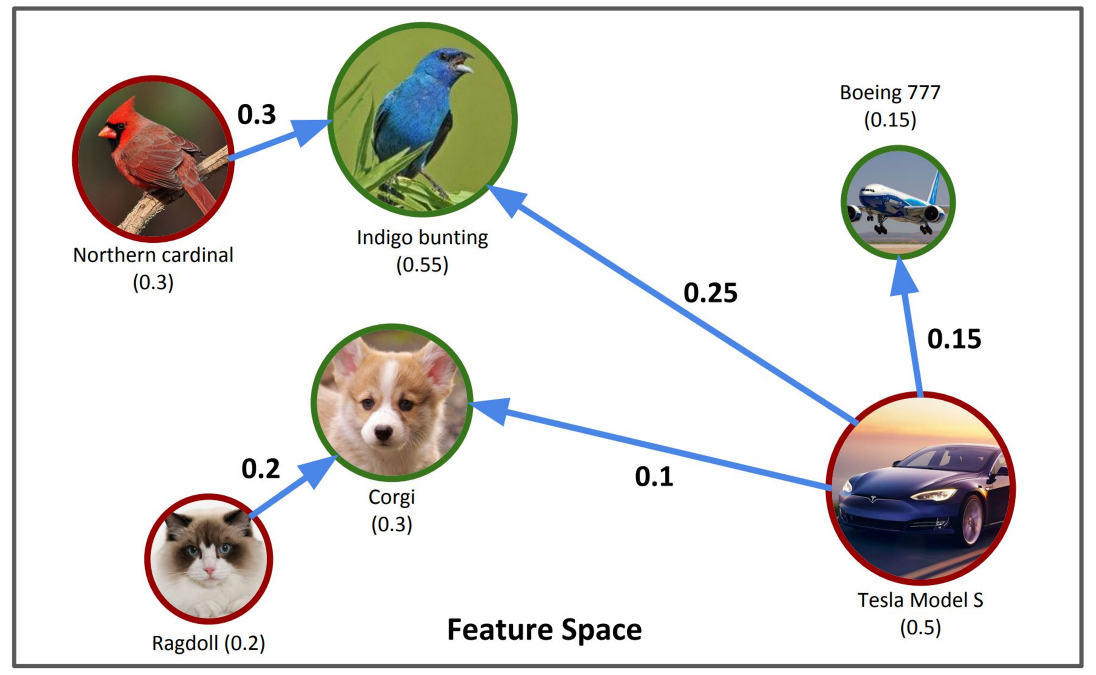


	Before FT		After FT	
	Top-1	Top-5	Top-1	Top-5
Head: $\geq$ 100 imgs	19.28	5.79	17.33	4.87
Tail: $<$ 100 imgs	29.89	9.12	24.15	6.41

Domain Similarity in Transfer Learning

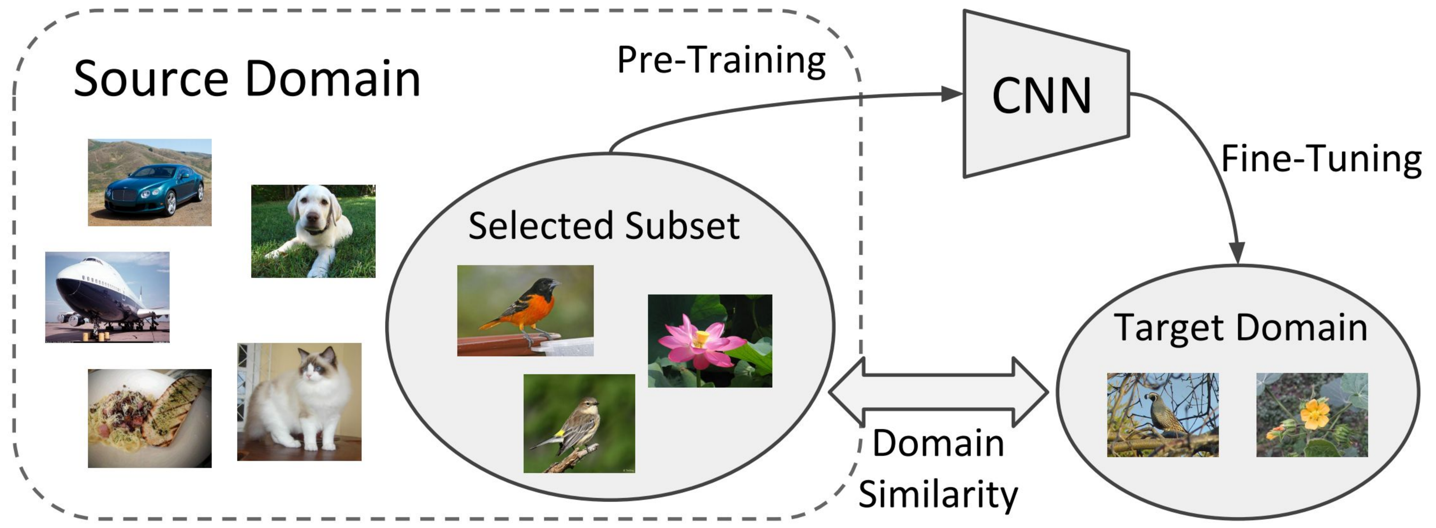
- Transfer learning as transporting a set of images from source domain to target domain.
- Define domain similarity by Earth Mover's Distance (**EMD**), based on distance of image feature.

- Source domain (red)
- Target domain (green)
- Size: number of images.
- Blue arrows: optimal flow by solving EMD.



Domain-Specific Transfer Learning

- Source domain: ImageNet + iNaturalist.
- Target domain: 7 fine-grained datasets.
- Select a subset from source domain by similarity.

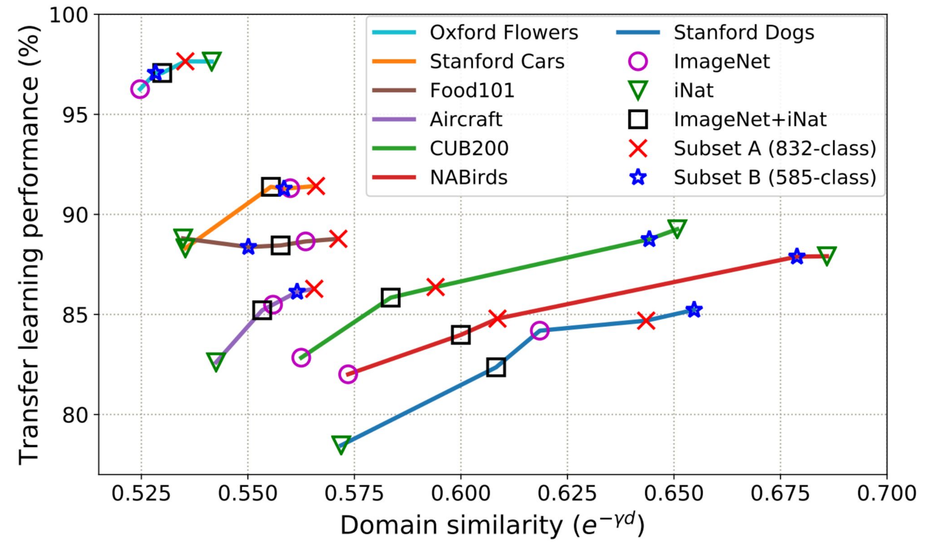


Transfer Learning Performance

	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

- ImageNet and iNaturalist are highly biased!

- Similar source domain leads to better transfer learning.
- A novel direction of studying how to do better pre-training.



- SOTA performance with off-the-self networks!

Method	CUB200	Stanford Dogs	Stanford Cars	Aircrafts	Food101
Subset B (585-class): Inception-v3	**89.6**	86.3	93.1	89.6	90.1
Subset B (585-class): Inception-ResNet-v2 SE	89.3	**88.0**	**93.5**	**90.7**	**90.4**
Krause *et al.* [30]	82.0	-	92.6	-	-
Bilinear-CNN [36]	84.1	-	91.3	84.1	82.4
Compact Bilinear Pooling [17]	84.3	-	91.2	84.1	83.2
Zhang *et al.* [68]	84.5	72.0	-	-	-
Low-rank Bilinear Pooling [29]	84.2	-	90.9	87.3	-
Kernel Pooling [11]	86.2	-	92.4	86.9	85.5
RA-CNN [16]	85.3	**87.3**	92.5	-	-
Improved Bilinear-CNN [35]	85.8	-	92.0	88.5	-
MA-CNN [69]	**86.5**	-	92.8	89.9	-
DLA [65]	85.1	-	**94.1**	**92.6**	**89.7**