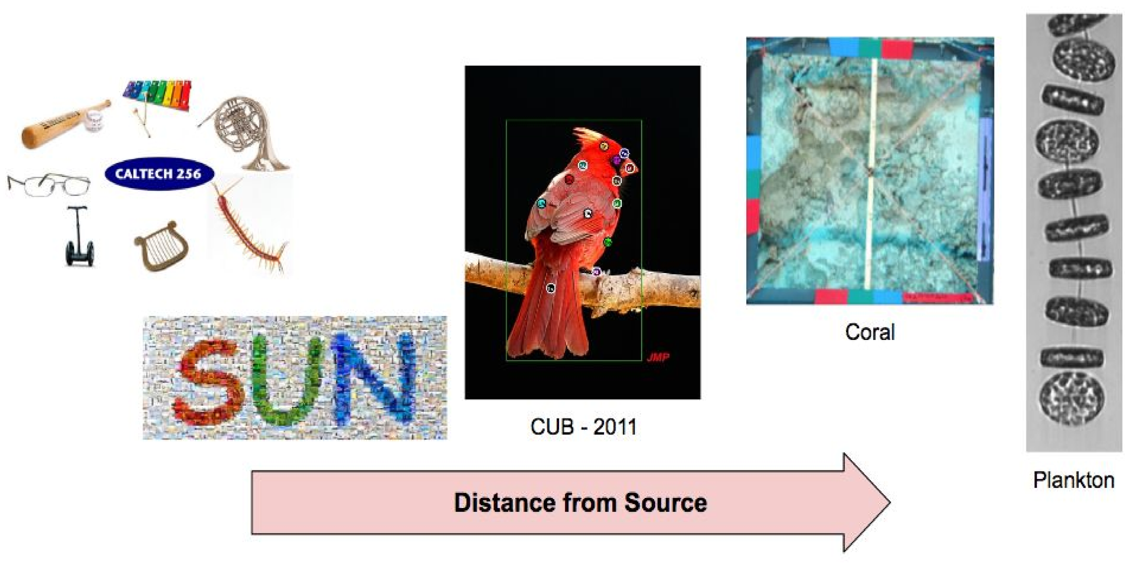
Best Practices for Fine-Tuning Visual Classifiers to New Domains

Brian Chu*, Vashisht Madhavan*, Oscar Beijbom, Judy Hoffman, Trevor Darrell

Department of Computer Science
University of California, Berkeley

Introduction

- Fine-tuning deep networks pre-trained on ImageNet is one of the most popular supervised transfer methods. However, there is no analysis as to how the fine-tuning procedure changes with the target dataset.
- We analyze fine-tuning along two axes
  - **Dataset distance** - how far the target data is from ImageNet
  - **Training data** - how many labeled examples we have for fine-tuning



Datasets

- The datasets listed below are in order from closest to furthest from ImageNet according to our distance measures.
- We consider a range of images per-class to understand the effects of low, medium, and high amounts of labeled target examples on fine-tuning

Dataset	# Categories	Classification Task	# Images per Class Val	Test	Train
Caltech256	256	Object	2	25	1, 10, 25, 53
SUN397	397	Scene	2	25	1, 10, 50, 70
MITIndoor	67	Scene	2	*	1, 10, 25, 75
CUB-200	200	Object (fine-grained)	2	*	5, 20, 35
Coral	9	Coral	50	300	10, 50, 200, 450
Plankton	103	Plankton	50	85	1, 10, 300, 550
Yosinski	500	Object	20	*	1, 10, 25, 53, 120

Table 1. Properties of the Datasets used and Their Training Splits. These datasets vary in terms of their object diversity and amount of training data available

Distance from Source

- These common domain shift metrics were computed using the mean fc7 responses of these datasets using a pretrained AlexNet
- All metrics show the same ordering of datasets, even though pairwise distances differ
- **CNN** - a three layer convolutional network that classifies whether image is from source or target.

Dataset	Cosine distance	MMD	Linear SVM	CNN
Yosinski	0.003	2.3	57.3%	51.0%
Caltech-256	0.071	10.6	71.4%	69.0%
SUN397	0.194	17.8	81.5%	76.4%
MIT-Indoor	0.307	23.9	90.0%	84.5%
CUB-200	0.358	37.2	92.9%	86.5%
Coral	0.455	38.7	97.3%	99.4%
Plankton	0.534	39.1	97.2%	99.7%

Table 2. Here we compare various commonly used metrics for measuring domain shift. What is interesting is that these metrics preserve the same ordering across datasets.

Experiments

- We randomly initialize(RI) the top **1, 3, and 5** layers of a pre-trained AlexNet
- The layers weights that are copied are then fine-tuned towards the target dataset

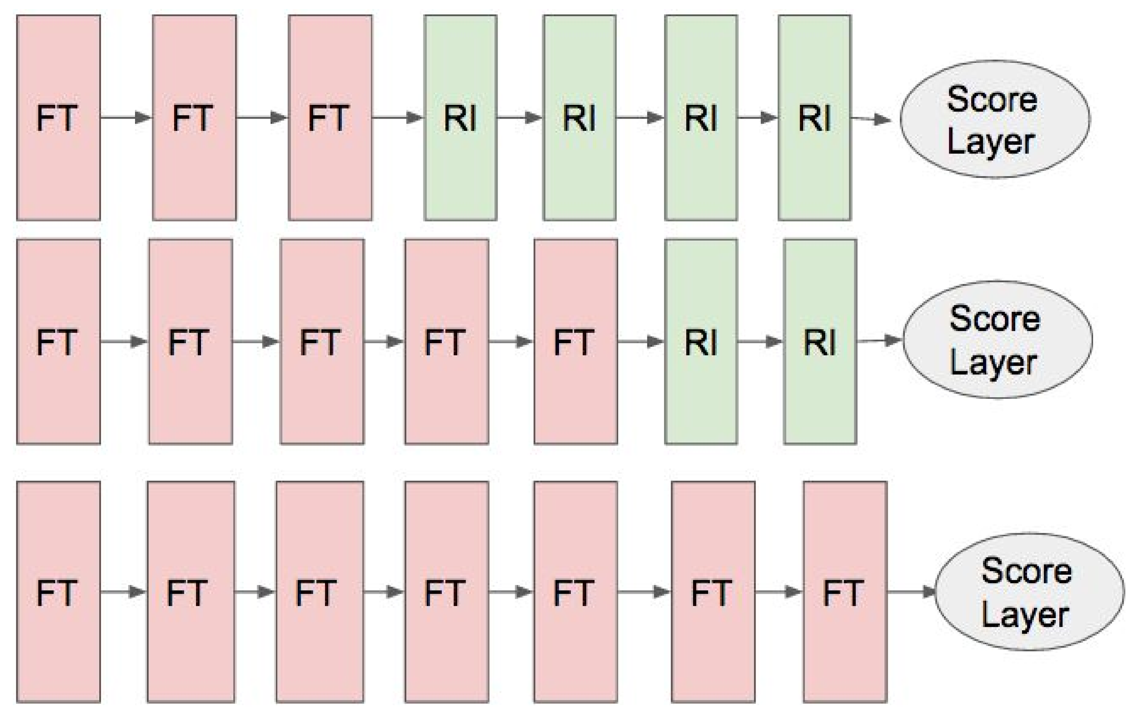


Figure 1. For notation, FT(a-b) denotes that layers a-b are copied and fine-tuned. R(a-b) denotes that layers a-b are randomly initialized and learned. Here we have FT(1-3)R(4-8) (top) , FT(1-5) R(4-8)(middle), FT(1-3) R(8) (bottom)

- We also copy and freeze layer weights in the same network configurations we used for fine-tuning
- This helped us understand how well AlexNet features trained on ImageNet directly transfer AND how these frozen layers interact with randomly initialized ones

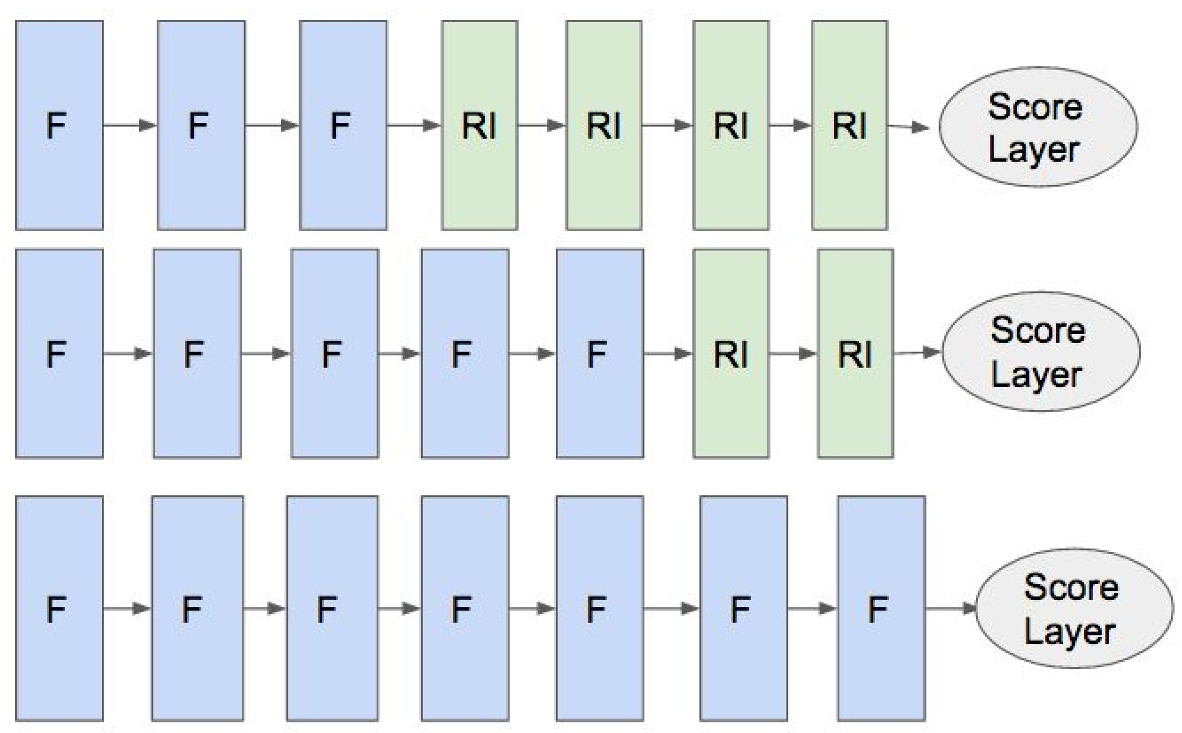


Figure 2. In the same way as before, F(a-b) denotes that layers a-b are copied and fine-tuned. R(a-b) denotes that layers a-b are randomly initialized and learned. Here we have F(1-3)R(4-8) (top) , F(1-5) R(4-8)(middle), F(1-3) R(8) (bottom)

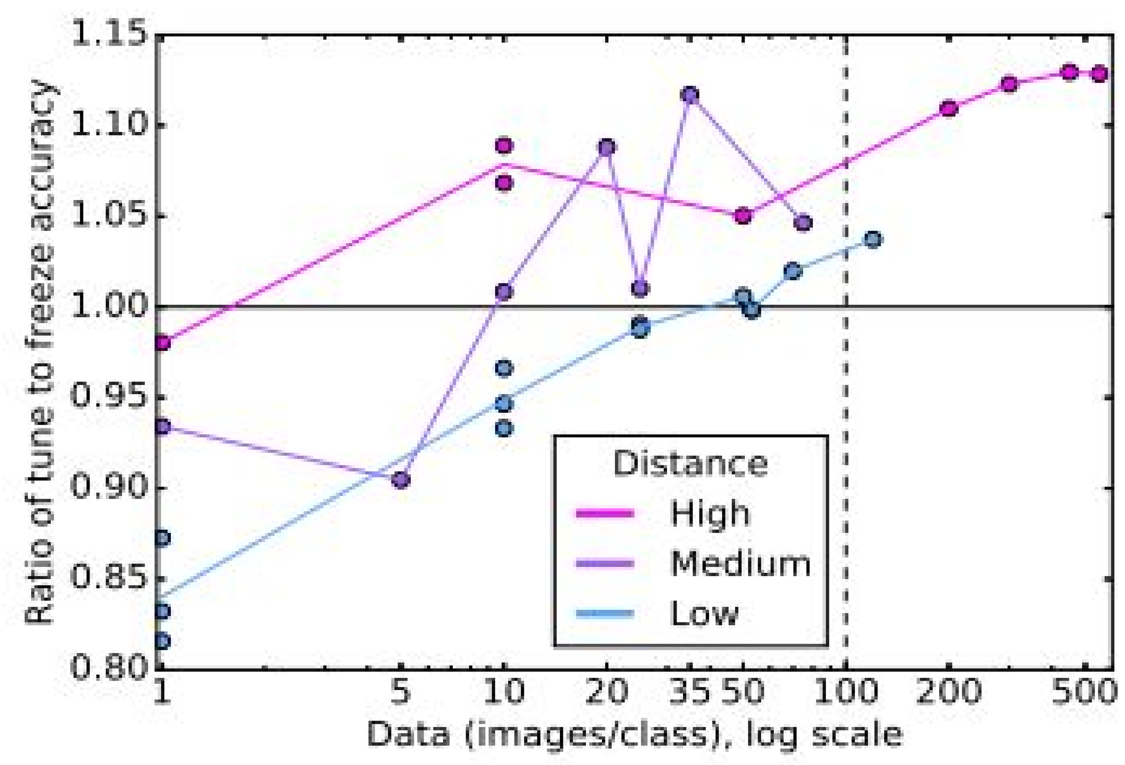
Conclusions

- **Copy all layers except the classification layer.** This is often standard practice, though we are the first to provide comprehensive evidence across a variety of datasets and many different operating points of the amount of labeled data available in the target dataset.
- **Fine-tune the copied layers**. We find that even with very few examples, fine-tuning is possible and beneficial. T**he exception being if the dataset distance is small and there is only a small amount of training data**. In this case, freeze the copied layers.

Experimental Analysis

		Images per Class L (1-20)	M (21-99)	H ($\geq$ 100)
Cosine Distance	L (0.0-0.2)	Freeze	Try Freeze or Tune	Tune
	M (0.2-0.4)	Try Freeze or Tune	Tune	Tune
	H (0.4-1.0)	Try Freeze or Tune	Tune	Tune

Table 3 . Our matrix here outlines recommendations based on distance and amount of per class data. Try Tune or Freeze means that there is a tradeoff between training speed and accuracy. Tuning gives slightly higher accuracy but takes longer to train, whereas freezing layers makes for quick training but slightly lower accuracy



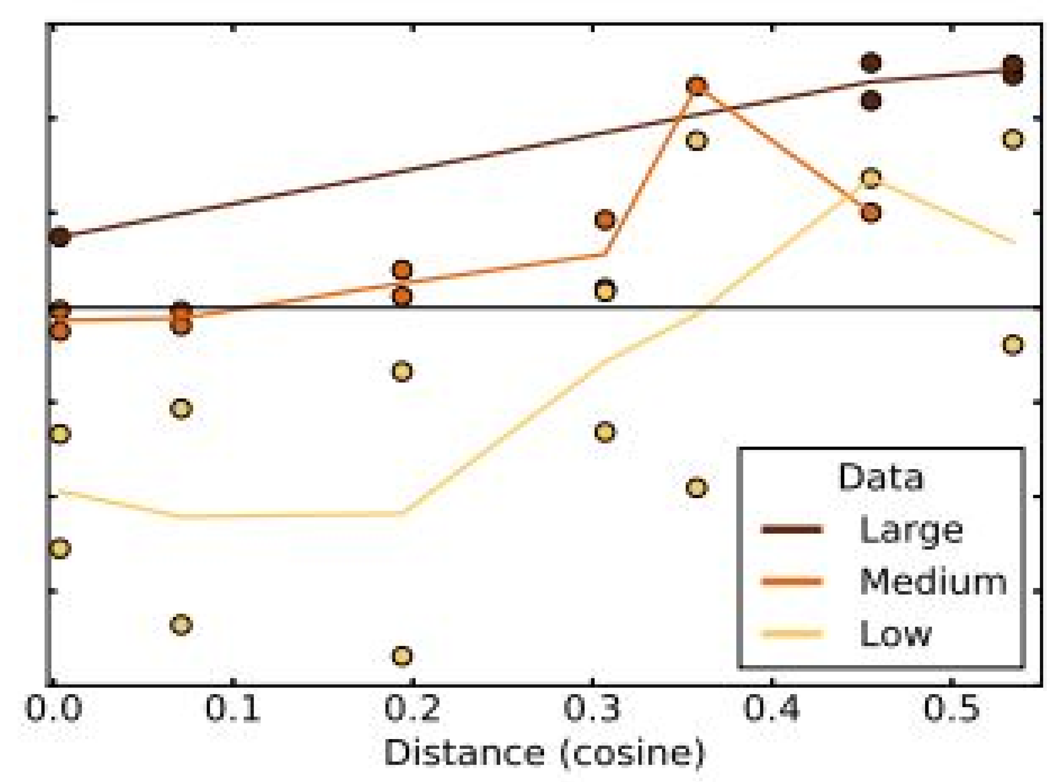


Figure 3. The solid black line denotes when tuning is just as good as freezing in terms of accuracy. The points above the line along both axes (distance and data size) give evidence that fine-tuning performs better than just freezing layers

References

Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Proceedings of the European Conference on Computer Vision (ECCV). (2014)

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: International Conference in Machine Learning (ICML). (2014)

Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Comput. Vis. Image Underst. (April 2007)

Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology (2007)

Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Largescale scene recognition from abbey to zoo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2010)

Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in Neural Information Processing Systems (NIPS). (2014)

Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014)

Agrawal, P., Girshick, R., Malik, J.: Analyzing the performance of multilayer neural networks for object recognition. In: Proceedings of the European Conference on Computer Vision (ECCV). (2014)

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Advances in Neural Information Processing Systems (NIPS). (2014)

Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS). (2012)

Azizpour, H., Razavian, A., Sullivan, J., Maki, A., Carlsson, S.: Factors of transferability for a generic convnet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2015)