Learning Transferable Features with Deep Adaptation Networks

Mingsheng Long[12], Yue Cao[1], Jianmin Wang[1], and Michael I. Jordan[2]

International Conference on Machine Learning, 2015

Introduction

- Goal: Enhance the transferability of features from task-specific layers
- Proposed a Deep Adaptation Network DAN architecture
  - General features can generalize well to a novel task; however, for specific features they cannot bridge the domain discrepancy
- Some ways to enhance feature transferability:
  - By mean-embedding matching, feature transferability can be enhanced substantially
  - Utilizing multi-layer representations across domains in a reproducing kernel Hilbert space

Main Breakthrough

- *Generalizes deep CNN to the domain adaptation*
- Deep adaptation of multiple task-specific layers, including output
- Optimal adaptation using multiple kernel two-sample matching

Deep Learning For Domain Adaptation

- None or very weak supervision in the *target* task (new domain)
  - Target classifier cannot be reliably trained due to over-fitting
  - Fine-tuning is impossible as it requires substantial supervision
- Generalize related supervised source task to the target task
  - Deep networks can learn transferable features for adaptation
- Hard to find big source task for learning deep features from scratch
  - Transfer from deep networks pre-trained on unrelated big dataset
  - Transferring features from distant tasks better than random features

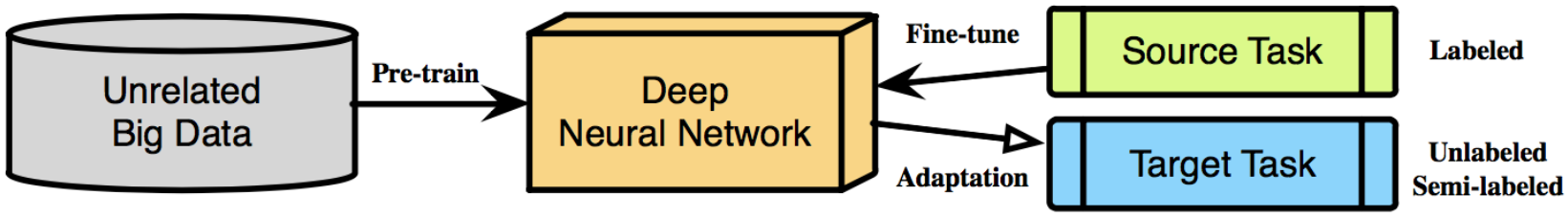

Figure: *Deep Learning for Domain Adaptation Workflow*

How Transferable Are Deep Features?

- Transferability is restricted by (Yosinski et al. 2014; Glorot et al. 2011)
- Specialization of higher layer neurons to original task (new task)
- Disentangling of variations in higher layers enlarges task discrepancy
- Transferability of features decreases while task discrepancy increases

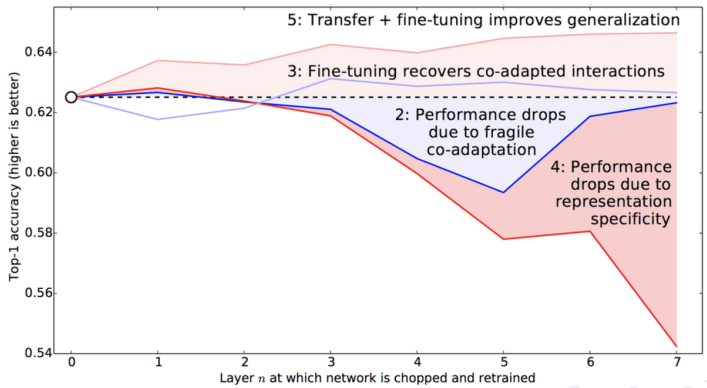

Figure: Transferability of features decreases while task discrepancy increases

Deep Adaptation Network (DAN)

Key Observations (AlexNet) (Krizhevsky et al. 2012)

- Comprised of five convolutional layers $conv1 - conv5$ and three fully connected layers $fc6 - fc8$
- Convolutional layers learn general features: safely transferable
  - Safely freeze $conv1 - conv3$ & fine tuned $conv4 - conv5$
- Fully-connected layers fit task specificicy: NOT safely transferable
  - Deeply adapt $fc6 - fc8$ using statistically optimal two-sample matching

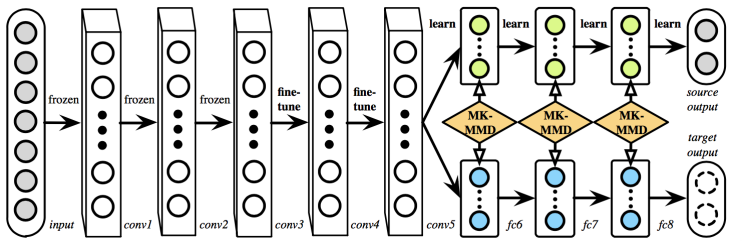

Figure: The DAN architecture for learning transferable features