On the Transferability of Representations in Neural Networks Between Datasets and Tasks

Haytham M. Fayek, Lawrence Cavedon, Hong Ren Wu

School of Engineering | School of Science, RMIT University

haytham.fayek@ieee.org

Abstract

Deep networks, composed of multiple layers of hierarchical distributed representations, tend to learn low-level features in initial layers and transition to high-level features towards final layers. Paradigms such as transfer learning, multi-task learning, and continual learning leverage this notion of generic hierarchical distributed representations to share knowledge across datasets and tasks. Herein, we study the layer-wise transferability of representations in deep networks across a few datasets and tasks and note some interesting empirical observations.

1. Introduction

Deep networks, composed of multiple layers of hierarchical distributed representations, tend to learn low-level features in initial layers and transition to high-level features towards final layers.

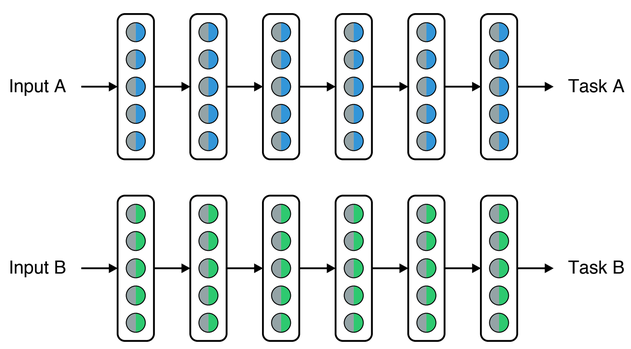
Similar low-level features commonly appear across various datasets and tasks, while high-level features are somewhat more attuned to the dataset or task at hand, which makes low-level features more generic and easier to transfer from one dataset or task to another [Yosinski et al., 2014].

Paradigms such as transfer learning, multi-task learning, and continual learning leverage this notion of generic hierarchical distributed representations to share knowledge across datasets and tasks [Fayek et al., 2016].

2. Gradual Transfer Learning

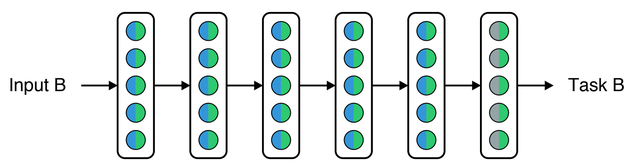
Gradual transfer learning is a methodology for quantifying the layer-wise transferability of representations between two datasets or tasks as follows.

Two primary neural network models of L layers are trained independently, one for each dataset or task.

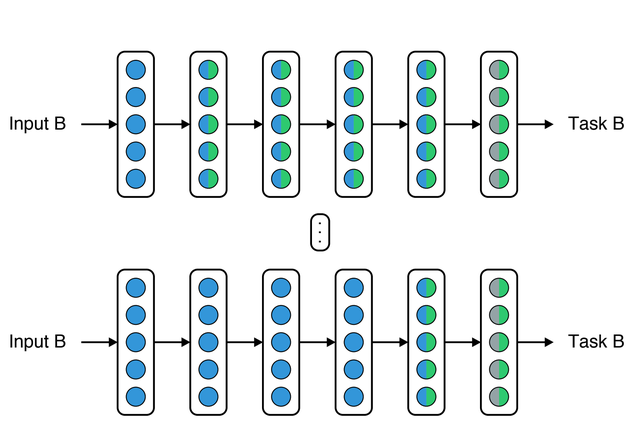


The learned parameters in all layers of the trained model, except the output layer, are copied to a new model for the (other) secondary dataset or task. The first $l_c \in \{0, \ldots, L_H\}$ layers are held constant and the remaining layers are fine-tuned for the secondary dataset or task, where $L_H = L - 1$.

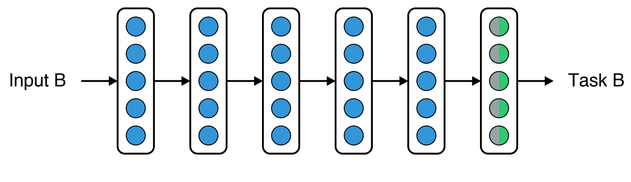
When $l_c = 0$, the primary model is an initialization to the secondary model.



The process is repeated iteratively for $1 \leq l_c < L_H$.



When $l_c = L_H$, the primary model is a feature extractor to the secondary model.



If the constant transferred layers l_c are relevant to the secondary dataset or task, one can expect an insignificant or no drop in performance relative to the primary model trained independently, and vice versa.

3. Experiments

The CIFAR-10, CIFAR-100, and SVHN datasets are used to study how task relatedness can influence the layer-wise transferability.

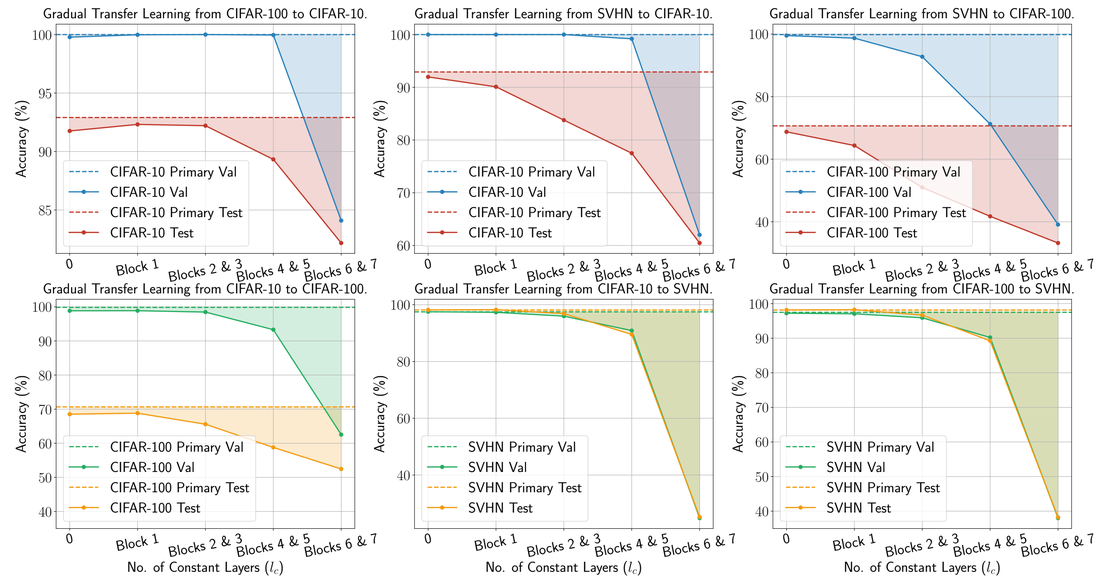


Figure 1: *Classification accuracy of gradual transfer learning between the CIFAR-10, CIFAR-100, and SVHN datasets using DenseNets.*

The ASR (TIMIT) task and the SER (IEMOCAP) task are used to study the influence of the neural network architecture on the layer-wise transferability.

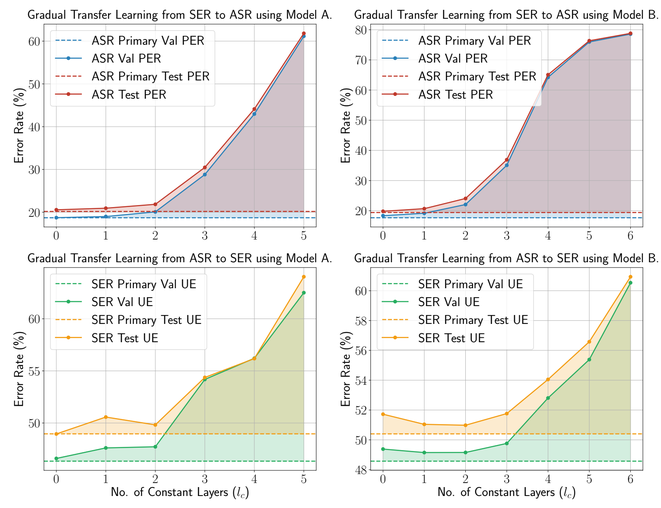


Figure 2: *Phone Error Rate (PER) and Unweighted Error (UE) of gradual transfer learning between the ASR (TIMIT) and SER (IEMOCAP) tasks. The architecture of Model A is similar to AlexNet. The architecture of Model B is similar to VGGNet.*

4. Discussion

The following observations highlight the importance of curriculum methods and structured approaches to designing systems for multiple tasks in paradigms that incorporate learning multiple tasks to maximize the knowledge transfer and minimize the interference between datasets or tasks.

- The layer-wise transferability between two datasets or tasks can be non-symmetric.
- The nature and relationship of the datasets or tasks involved are more influential on the layer-wise transferability of representations compared with other factors such as the architecture of the neural network.
- The layer-wise transferability of representations can be used as a proxy for quantifying task relatedness.

Acknowledgments. H. M. Fayek was funded by the Vice-Chancellor's Ph.D. Scholarship (VCPS) from RMIT University. This research was undertaken with the assistance of resources and services from the National Computational Infrastructure (NCI), which is supported by the Australian Government. The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of one of the Tesla K40 GPUs used for this research.

References

Haytham M. Fayek, Margaret Lech, and Lawrence Cavedon. On the correlation and transferability of features between automatic speech recognition and speech emotion recognition. In *Interspeech*, pages 3618–3622, 2016.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems (NIPS)*, pages 3320–3328. Curran Associates, Inc., 2014.

NeurIPS 2018, Continual Learning Workshop, 32nd Neural Information Processing Systems, 7 December 2018, Montréal, Canada.