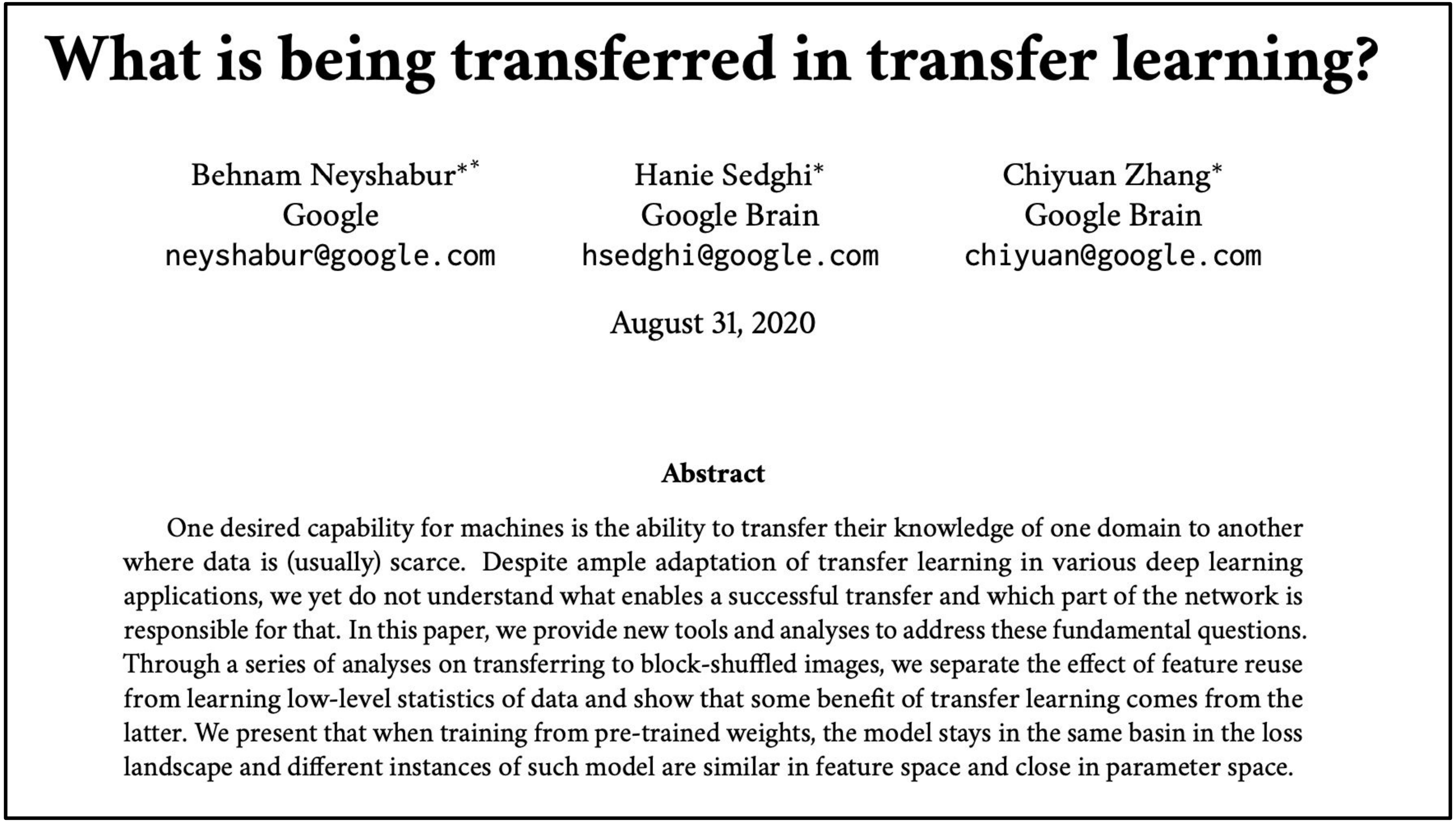
Paper Review:

What is being transferred in transfer learning?

Seyed Iman Mirzadeh

seyediman.mirzadeh@wsu.edu

Disclaimer: The master theme for this presentation is borrowed and edited from the DMG theme.



What is being transferred in transfer learning?

Behnam Neyshabur[**]
Google
neyshabur@google.com

Hanie Sedghi[*]
Google Brain
hsedghi@google.com

Chiyuan Zhang[*]
Google Brain
chiyuan@google.com

August 31, 2020

Abstract

One desired capability for machines is the ability to transfer their knowledge of one domain to another where data is (usually) scarce. Despite ample adaptation of transfer learning in various deep learning applications, we yet do not understand what enables a successful transfer and which part of the network is responsible for that. In this paper, we provide new tools and analyses to address these fundamental questions. Through a series of analyses on transferring to block-shuffled images, we separate the effect of feature reuse from learning low-level statistics of data and show that some benefit of transfer learning comes from the latter. We present that when training from pre-trained weights, the model stays in the same basin in the loss landscape and different instances of such model are similar in feature space and close in parameter space.

https://arxiv.org/pdf/2008.11687.pdf

Punchlines

- **Motivation**: Despite ample adaptation of transfer learning in various deep learning applications, we yet do not understand what enables a successful transfer and which part of the network is responsible for that.

- **Goal**: provide new tools and analyses to address these fundamental questions

- **Result**: when training from pre-trained weights, the model stays in the same basin in the loss landscape and different instances of such model are similar in feature space and close in parameter space.

Setup

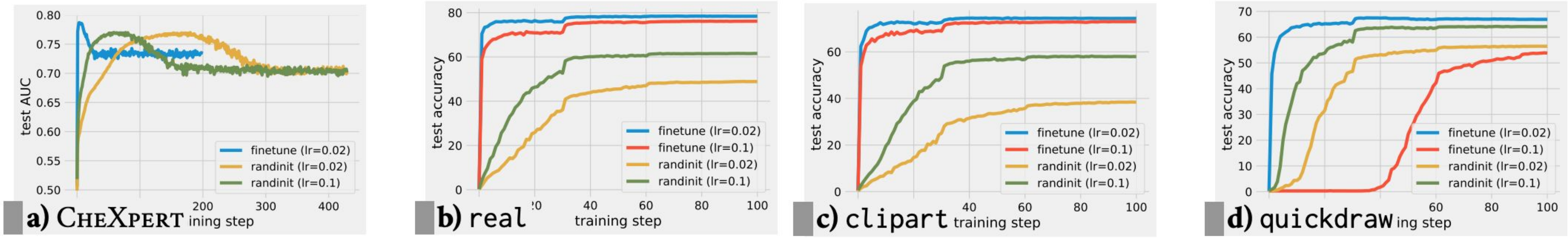


Figure 2: Learning curves comparing random initialization (RI-T) and finetuning from ImageNet pre-trained weights(P-T). For CheXpert, finetune with base learning rate 0.1 is not shown as it failed to converge.

Experiment 1: Role of feature reuse

- The benefits of transfer learning are generally believed to come from reusing the pre-trained feature hierarchy.

- However, this intuition cannot explain why in many successful applications of transfer learning, the target domain could be visually very dissimilar to the source domain (e.g., imagenet to chest rays)

- Question: How to test if feature reuse is important ?
    - Take a guess!

Experiment 1: Role of feature reuse

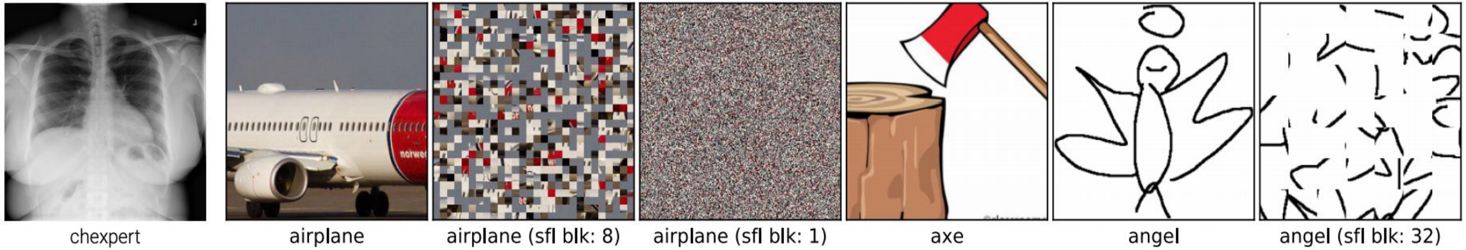
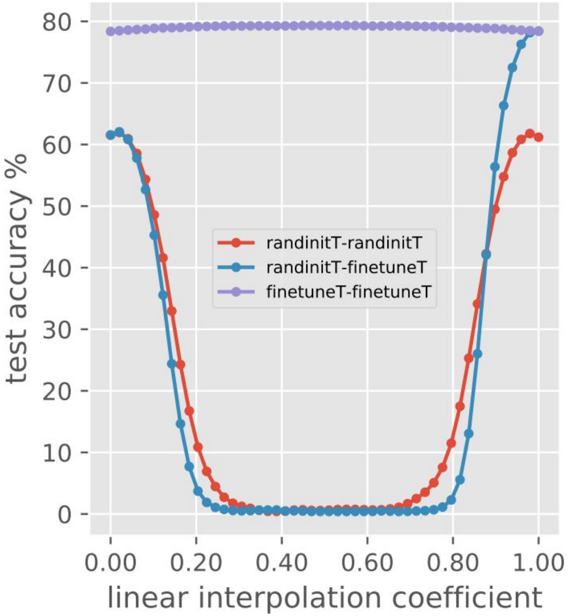
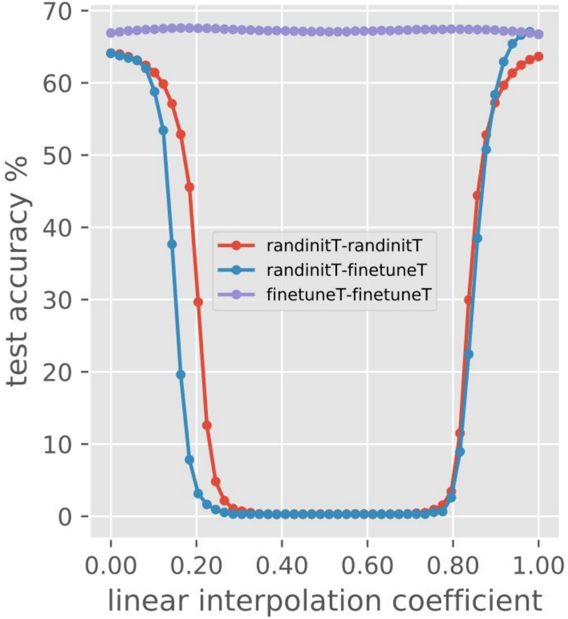


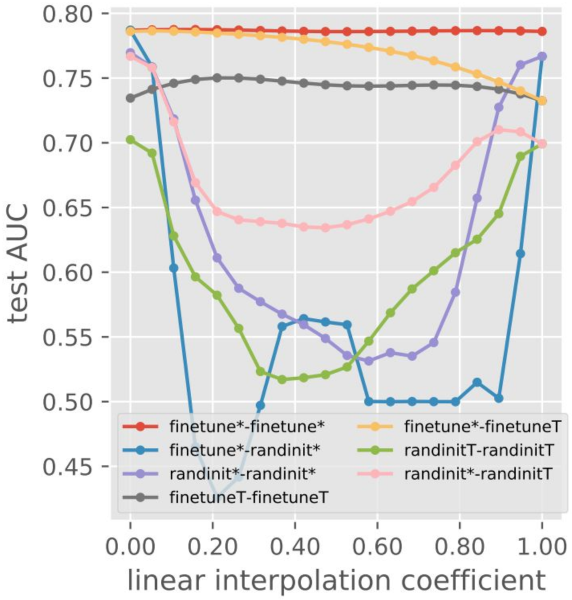
Figure 1: Sample images of dataset used for transfer learning downstream tasks. Left most: an example from CheXpert. The next three: an example from the DomainNet `real` dataset, the same image with random shuffling of 8 × 8 blocks and 1 × 1 blocks, respectively. The last three: examples from DomainNet `clipart` and `quickdraw`, and a 32 × 32 block-shuffled version of the `quickdraw` example.

Result: Although the benefit of transfer learning diminishes, still, it's helpful!

Experiment 2: Loss landscape of models







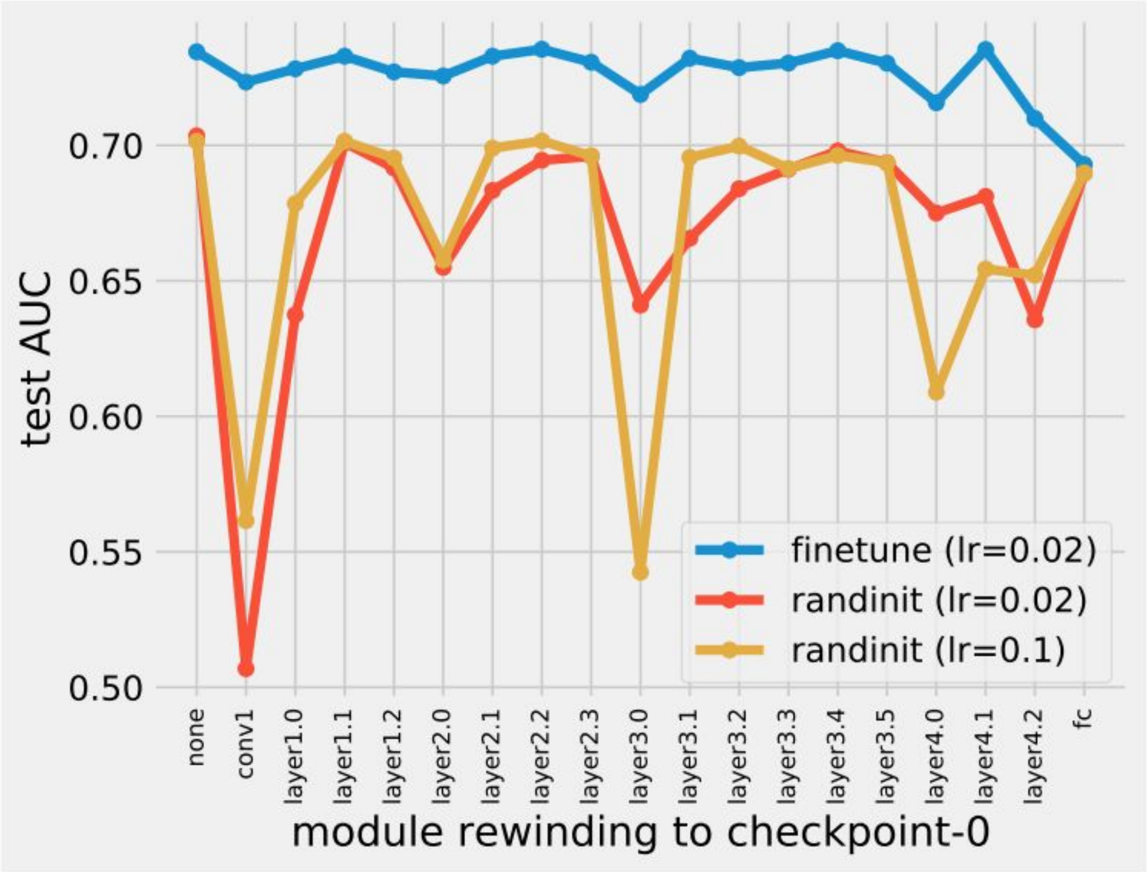
Experiment 3: Module Criticality

- Different layers of the network show different robustness to perturbation of their weight values.

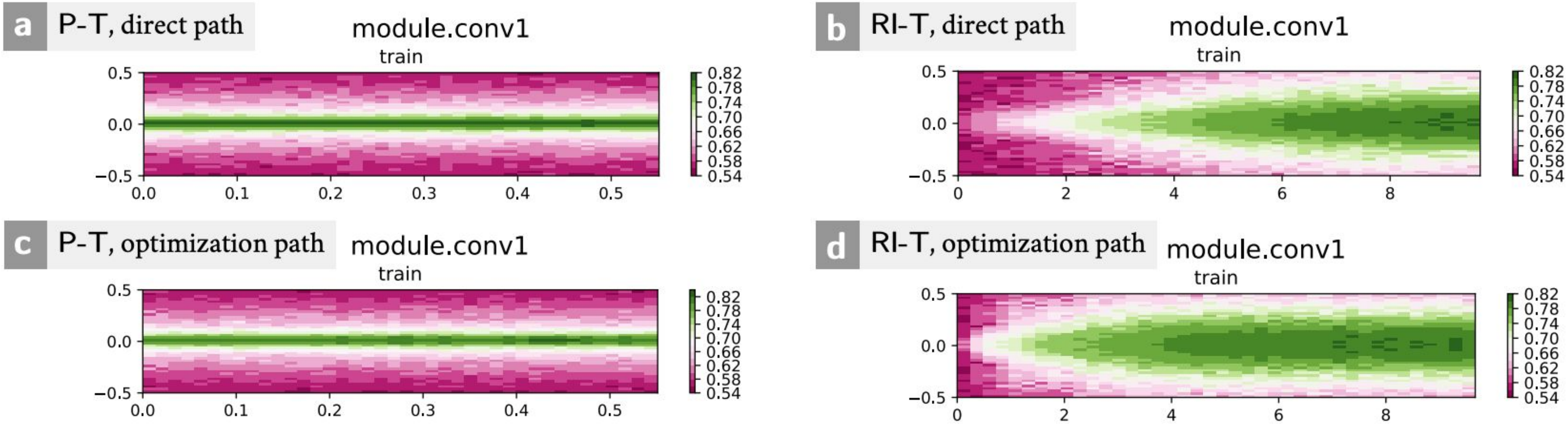
- Experiment: consider a trained network, take one of the modules and rewind its value back to its initial value while keeping the weight value of all other modules fixed at trained values.

- Module called critical, if the performance of the model drops significantly after rewinding, while for others the performance is not impacted.

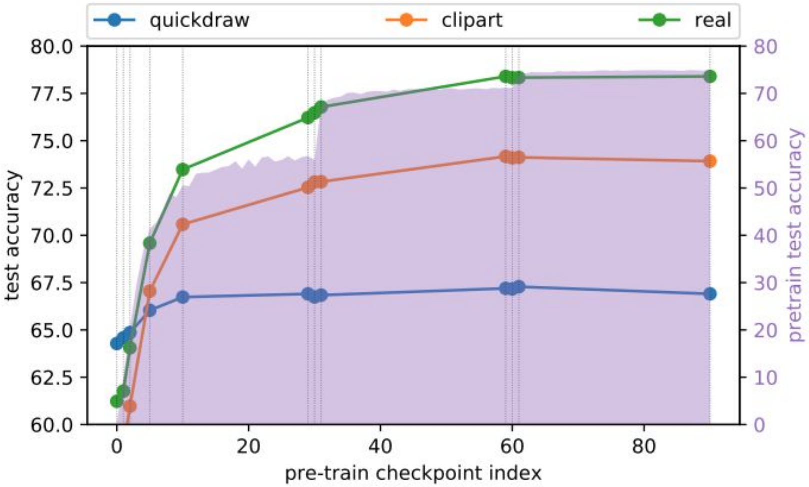
Experiment 3: Module Criticality (2)

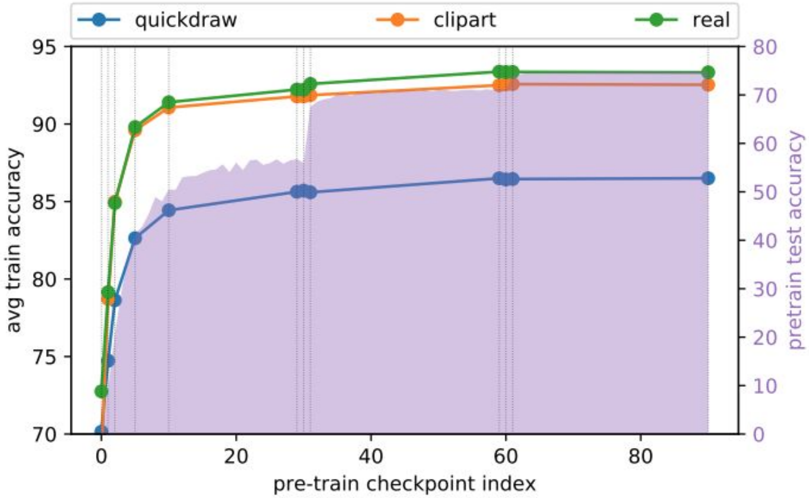


Experiment 3: Module Criticality (3)



Experiment 4: Which pre-trained checkpoint is most useful for transfer learning?





Summary

- For a successful transfer both feature-reuse and low-level statistics of the data are important.

- Models trained from pre-trained weights make similar mistakes on target domain, have similar features and are surprisingly close in l2 distance in the parameter space. They are in the same basins of the loss landscape.

- Models trained from random initialization do not live in the same basin, make different mistakes, have different features and are farther away in l2 distance in the parameter space

Summary (2)

- Modules in the lower layers are in charge of general features and modules in higher layers are more sensitive to perturbation of their parameters.

- One can start from earlier checkpoints of pre-trained model without losing accuracy of the fine-tuned model. The starting point of such phenomena depends on when the pre-train model enters its final basin.

Thank You!

Contact: seyediman.mirzadeh@wsu.edu