What is being Transferred in Transfer Learning?

Behnam Neyshabur*, Hanie Sedghi*, Chiyuan Zhang*
Google Brain

Understanding Transfer Learning

- One desired capability of machines is to transfer their knowledge or understanding of a domain it is trained on to another domain where data is (usually) scarce or a fast speed of convergence is needed.
- Plethora of successful applications.
- We yet do not understand:
  - what enables a successful transfer?
  - which parts of the network are responsible for that?

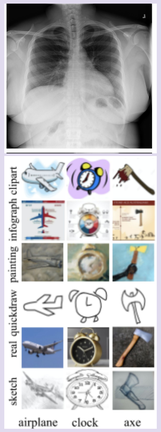
We address these fundamental questions and propose new tools and analysis.

Summary of Results

- Both feature-reuse and low-level statistics of the data are important.
- Finetune models make similar mistakes on target domain, they have similar features and are surprisingly close in the L2 distance in parameter space.
- Finetune models are in the same basins in the loss landscape, while models trained from random initialization are in a different basin.
- Lower layers are in charge of general features and 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.

Problem Setup

- Target domains that are intrinsically different and diverse:
- CheXpert: a medical imaging dataset of chest x-rays considering 5 different diseases.
- DomainNet: designed to probe transfer learning for diverse visual representations. The domains range from real images to sketches, clipart and painting samples. 345 classes
- Two initialization scenarios:
  - Pre-trained on ImageNet (Finetune)
  - Start from random initialization (RandInit)

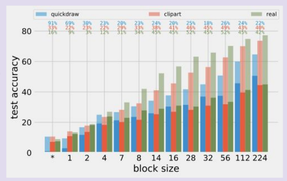


Role of Feature Reuse

- Learning curves: Closer to imagenet, better performance. Finetune converges faster.
- Feature re-use: the network uses the features learned on source domain, to perform tasks on target domain.
- Experiment: We partition the image of the downstream tasks into equal sized blocks and shuffle the blocks randomly. The shuffling disrupts visual features in those images.

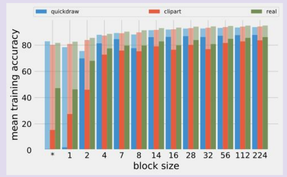
- Feature reuse plays a very important role!

  especially when the downstream task shares similar visual features with the pre-training domain.



- There are other factors at play!

  low-level statistics of the data that are not ruined in the shuffling lead to the significant benefits of transfer learning, especially on optimization speed.

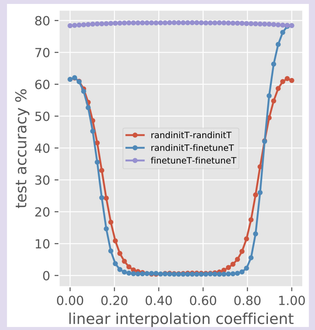


Opening up the model

- Investigating mistakes: two Finetune models have strictly fewer uncommon mistakes.
- Feature similarity: Finetune models are more similar in feature space. and even when Randinit models are showing high accuracy, they are not that similar in the feature space.
- Distance in parameter space: Randinit models are farther from each other compared to two finetune models. This trend is seen in individual modules too. The distance between modules increases as we move towards higher layers in the network.

Performance barriers & basins in the loss landscape

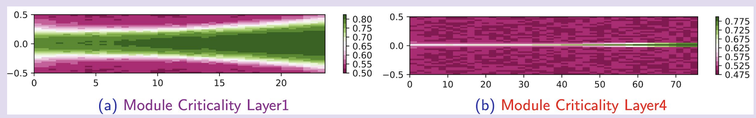
- *Any* two minimizers of a deep network can be connected via a **non-linear** low-loss path.
- We evaluate a series of models along the **linear** interpolation of the two weights.
- Performance barriers are generally expected between two unrelated NN models. But, When the two solutions belong to the same flat basin of the loss landscape, performance barrier is absent.



- **Finetune models reside in the same basin.**
- **RandInits end up in a different basin, even if starting from same random seed.**

Module criticality

- Module criticality as a measure of generalization ability [Chatterji et al 2020]
- As we move from the input towards the output, we see tighter valleys, i.e., modules become more critical
- Lower layers are in charge of more general features while higher layers have features that are more specialized for the target domain.



Code available at: https://github.com/google-research/understanding-transfer-learning