1. Write a 2-page (excluding pictures) summary of what transfer learning (10 points)

Strong performance using deep learning usually requires a large amount of task specific data and compute. These per-task requirements can make new tasks prohibitively expensive. Transfer learning offers a solution: task-specific data and compute are replaced with a pre-training phase. A network is trained once on a large, generic dataset, and its weights are then used to initialize subsequent tasks which can be solved with fewer data points, and less compute.

Transfer learning for visual task, where the input is an image. The classifier could be an image e.g. An image of a cat or that of a lung (like what we did in explainability assignment) which is done with CNN's that take in the images by many layers of convolution especially residual networks. This works fine when dealing with large data sets. A problem arises when you have a small dataset with few labeled samples where the model could learn from and this isn't enough to learn from big models that would perform well, making you settle for a less performing model.

The solution is transfer learning. In transfer learning, you take a large dataset and you train your CNN on it. You then take the CNN that you gain from the large dataset as a starting point and perform what is called Fine Tuning (FT) on the small dataset, training for a few steps on the small dataset and adapt it to the small dataset that you plan to train it on.

This works with the assumption that the images in the large dataset are overlapping with those in the small dataset, that they share some similarity and share features. If that's the case, when you FT(fine tune) on the small dataset, you can reuse the features, with some readjustment and learn how to map the features to the output which is now different to the original test. You wont have to rediscover the features. That what's makes transfer learning useful.

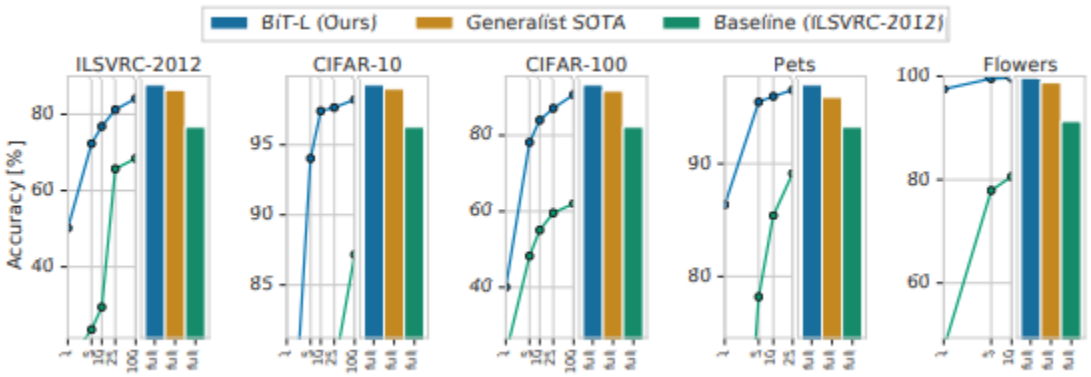
The first phase is called pre-training (PT) and final phase is called fine tuning (FT).

The goal is the following: pre-train on a large supervised source dataset and fine-tune the weights on the target task. (BiT-L 300M images – JFT dataset, BiT-M 14M images -IN21K dataset and BiT – S - 1.3M images ImageNet dataset)

Take pre-trained model (BiT-L) and Fine Tune (FT) it to these datasets (ILSVRC-2012, CIFAR-10, CIFAR-100, Pets and Flowers). Taking the full dataset outperforms State of the Art dataset (SOTA). It outperforms some of the specialist models. Not all of them. Taking 5,10,25,100 labels per test still achieves good results, however not as well as using full pretrained dataset.

In the big dataset, they remove all images that appear in the downstream task. These images are scraped from on the internet, so remove images of exact duplicates. These datasets might be apart of one another, but the results are show

improvement as seen in the diagram below.



X-axis: Number of Samples per Data Class

Fig. 1: Transfer performance of our pre-trained model, BiT-L, the previous stateof-the-art (SOTA), and a ResNet-50 baseline pre-trained on ILSVRC-2012 to downstream tasks. Here we consider only methods that are pre-trained independently of the final task (generalist representations), like BiT. The bars show the accuracy when fine-tuning on the full downstream dataset. The curve on the left-hand side of each plot shows that BiT-L performs well even when transferred using only few images (1 to 100) per class.