Improving the transferability of self-supervised learning models

Throughout the last decade, humanity has witnessed unprecedented growth in the capabilities of deep neural networks. As such, they have been widely adopted as the state-of-the-art approach to Artificial Intelligence, and enhance the work, entertainment, and healthcare of billions of people on a daily basis. However, deep neural networks depend on large amounts of labeled data to learn how to perform their tasks. This presents a bottleneck in deploying deep learning models for real-world problems – it is often difficult or costly to label enough data to train a model for a novel task from scratch in a purely supervised fashion. Fortunately, supervised learning is not the only way of training deep learning models.

Self-supervised learning (SSL) is a paradigm of learning representations from unlabeled data, with the intention of *transferring* them to specific downstream tasks defined by labeled data (see Figure 1). In practice, fine-tuning such pretrained models requires much smaller labeled datasets and computational resources compared to training them from scratch. SSL opens up possibilities to pretrain models on vast amounts of unlabeled data, which has led to successes in domains ranging from natural language processing to computer vision.

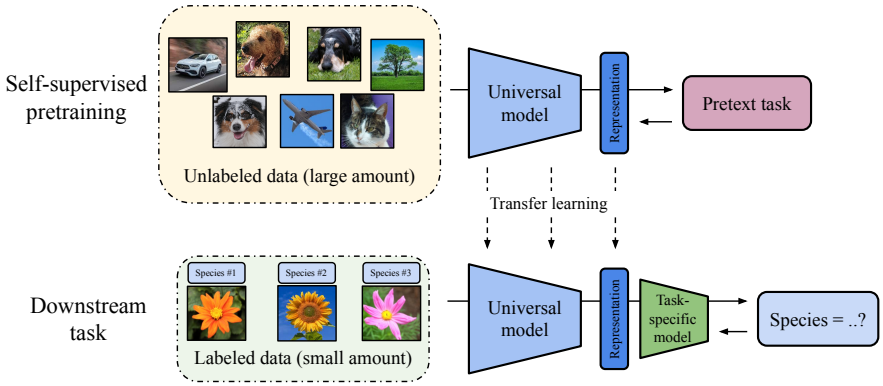

Figure 1: Self-supervised learning trains deep neural networks for building *expressive data representations* by solving *pretext tasks*, which do not require the training data to be labeled. Afterwards, the pretrained model can be re-used as a foundation for models solving different kinds of downstream tasks expressed by labeled data. Since we transfer a pretrained model, we do not need a large amount of labeled data to tune it for specific tasks.

Representations trained through SSL are, however, no exception to common pitfalls which may occur during transfer learning. For instance, the transferred model inherits not only a robust representation but also biases learned during the pretraining task. For example, pretraining with a contrastive objective of invariance to color jittering will result in a model less sensitive to changes in colors, which will poorly generalize to color-dependent tasks, such as flower classification.

The above capabilities and limitations of SSL inspire us to focus this research project on **improving the transferability of self-supervised learning models**. The main direction we would like to pursue in our research is the problem of mitigating the biases induced in the representations by specific pretraining objectives, such as augmentation invariance. The expected results of our research have the potential to affect other machine learning domains, where generalization is an important issue, such as Meta-Learning and Continual Learning.

In our experiments, we will make use of a wide array of contemporary deep learning techniques, such as conditioning neural networks with different data modalities, Hypernetworks, Transformers, and attention mechanisms. We will also explore the concept of network conditioning inspired by the mechanism of training conditional generative models, which has not been previously used within the self-supervised learning paradigm.

We hope that this project will lead to discovering new facts about self-supervised learning and improved ways to train more generalizable foundational models, which can be easily re-used for a plethora of tasks, enabling the application of Deep Learning to more real-world problems.