ERM, Stability, and Transfer Learning

Alok Shah and Thomas Zhang
Electrical and Systems Engineering
University of Pennsylvania

November 28, 2024

Introduction

The growing popularity of foundation models owes itself to remarkable capabilities of deep neural networks (DNNs) to learn a variety of domain-specific tasks with superhuman performance.

Under the pretrain-finetune regime, where large deep neural networks (DNNs) are trained in whole once on a wide variety of data (pertaining), and then the final layers are retrained on task-specific data.

Literature attributes this success to the backbone learning an underlying representation, producing separable features for the head to understand.

Representation learning aims to understand (insert some stuff)

The goal is transfer learning, that is, learning a representation conducive to retraining on a holdout task resulting in solid performance.

Despite emprical success, there remain many open questios regarding multitask and transfer learning. This work will examine: - How the choice of optimization algorithm influences task transfer

We'll measure our succes using the concept of excess risk ratio from train task to holdout task.

This paper, we aim to bridge this gap by investigating the dynamics of transfer learning and multitask optimization through the lens of excess risk bounds.

Typically, the pretrain-finetune paradigm is implemented using the emprical risk minimization. This seeks to minimize (insert some stuff) Prior works assumes that ERM on both the training and test sets

This paper will show that: - ERM is insuffieict to enforce task transfer: – we construct a counterexample showing both theoretically and empirically that the ERM solution results in unbounded excess test risk, and hence poor trasnferability

By final submission we aim to: - Introduce stability as a key notion to characterize transferabiltiy – will analyze the implicit bias of ERM algorithms, and use this to formulate an adapative sampling algorithm relying principles in Differential Privacy (DP) – apply information-theoretic principles to obtain bounds of performance.

Problem Statement

Consider a scenario with T different tasks where for each task $t \in [T]$ has data drawn from a distribution $x \sim P^{(t)}(x)$ with outputs $y = F_\star^{(t)} \circ g_\star(x)$. We're also given a held-out task T_0 with distribution $P^{(0)}(x)$ and corresponding output $y = F_\star^{(0)} \circ g_\star(x)$. The objective is to fit F and g to learn the tasks under a pretrain-finetune regime.

During pretraining, we perform empirical risk minimization (ERM) on some loss function ℓ over the training data by fitting a body g and a task-specific head $F^{(t)}$ for $t \in [T]$ through stochastic gradient descent. Formally:

$$\hat{g}, \hat{F}^{(t)} = \arg\min_{g,F} \sum_{i=1}^{T} \sum_{i=1}^{N} \ell\left(y, F^{(t)} \circ g(x_i^{(t)}\right) \tag{1}$$

During finetuning, $\hat{g}$ is fixed and a new head $F^{(0)}$ is fine-tuned on the held-out task:

$$\hat{F}^{(0)} = \arg\min_{F} \sum_{i=i}^{N} \ell\left(y, F^{(0)} \circ \hat{g}(x)\right) \tag{2}$$

We seek to analyze the performance of this setup using the concept of excess risk, defined as:

$$\mathrm{ER}^{(t)}(F, g) = \mathbf{E}_t\left[\ell\left(y, F \circ g(x)\right)\right] - \mathbf{E}_t\left[\ell\left(y, F_\star \circ g_\star(x)\right)\right)\right] \tag{3}$$

Prior literature assumes that small excess risk on the training tasks implies small excess on the held out task. Specifically $\forall g \in \mathcal{G}$ and $\nu > 0$

$$\frac{1}{T} \sum_{t=1}^{T} \inf_F \mathrm{ER}^{(t)}(F, g) \geq \nu \inf_F \mathrm{ER}^{(0)}(F, g) \tag{4}$$

where we refer to ν as the transfer coefficient. We also observe that this quite assumption is quite (too) general, as it would make more sense to hold over just the set $\mathcal{G} \subseteq G$, or the reachable g via ERM.

While prior literature demonstrates that $\nu \approx 0$, prior empirical results demonstrate otherwise by way of the strong performance of pretrained DNNs on downstream tasks. This project aims to reconcile these differences whether ν exhibits algorithmic stability with respect to ERM with SGD both theoretically and empirically.

Objectives

The key objectives of this project are:

- To develop a formal understanding of how pretraining on multiple tasks affects the excess risk on a held-out task

- To analyze the transfer coefficient, ν, and investigate its stability across different ERM runs

- To provide theoretical insights that explain the empirical success of neural networks and transformers in pretrain-finetune regimes

Related Work

-

Methodology

To achieve these objectives, the following approach will be employed:

1. Step 1: [Description of the first step]
2. Step 2: [Description of the second step]
3. Step 3: [Description of subsequent steps]

Timeline

The timeline for this project is estimated as follows:

Task	**Duration**
Literature Review	[X Weeks]
Data Collection	[X Weeks]
Implementation	[X Weeks]
Testing	[X Weeks]
Report Writing	[X Weeks]

Expected Outcomes

The expected outcomes of this project are:

- Outcome 1: [Describe the first expected result]
- Outcome 2: [Describe the second expected result]

Conclusion

In conclusion, this project aims to [restate the purpose of the project] by [briefly summarizing the approach]. The proposed methodology is expected to address the stated problem and produce meaningful outcomes.