Towards Few-Shot Adaptation of Foundation Models via Multitask Finetuning

Zhuoyan Xu, Zhenmei Shi, Junyi Wei, Fangzhou Mu, Yin Li, Yingyu Liang

Motivation

Few-Shot Learning: Pretraining + Fine Tuning

Problem Setup

Hidden representation data model

- Class \( y \in \mathcal{C} \) over distribution \( p \sim \mathcal{U}(y) \)
- \( \phi \in \Phi \) hypothesis class of representation functions: e.g. ResNet, ViT
- \( g(x) \equiv W(x) \) as prediction logits of \( \mathcal{C} \)

Definition 1 (Diversity and Consistency (Informal))

Consider the latent feature space. Diversity refer to the coverage of the finetuning tasks on the target task. Consistency refer to similarity.

Theorem (Multitask finetuning loss (Informal))

Suppose in pretraining we have empirical pretraining loss \( \mathcal{L}_{pre} \). After sufficient multitask finetuning and get \( \phi' \), the error will be \( \mathcal{L}(\phi') \leq \mathcal{C} \). Theorem will be \( \mathcal{C}(\phi) \leq \mathcal{O}(\alpha) \).

Theoretical Analysis

Definition 1 (Diversity and Consistency (Informal))

Consider the latent feature space. Diversity refer to the coverage of the finetuning tasks on the target task. Consistency refer to similarity.

Theorem (Multitask finetuning loss (Informal))

Suppose in pretraining we have empirical pretraining loss \( \mathcal{L}_{pre} \). After sufficient multitask finetuning and get \( \phi' \), the error will be \( \mathcal{L}(\phi') \leq \mathcal{C} \). Theorem will be \( \mathcal{C}(\phi) \leq \mathcal{O}(\alpha) \).

Practical Solution

Pretrained Method

- MoCo v3, DINO v2, supervised pretraining

Model

- ViT-S, ViT-B, ResNet50

Dataset

- miniImageNet, tieredImageNet, DomainNet, Meta-dataset

Take-Home Message

We provide the theoretical justification and practical solution of multitask finetuning for adapting pretrained foundation models to downstream tasks with limited labels.

Key Intuition

1. Pre-training uses unlabeled and noisy data for general purpose learning, where the model learns general structure rather than task-specific knowledge. Its performance on a specific task may not be perfect.

2. Despite the target data is limited, we have a clear understanding of the target task and its associated data.

○ We actively select extra data from a relevant source that covers target data characteristic features.

○ We then design specialized tasks for multitask finetuning to equip the model to address the specific types of target tasks effectively.

Table 1: Results evaluating our task selection algorithm on Meta-dataset using MoCo v2 backbone. No Cons.: Ignore consistency. No Div.: Ignore diversity. Random: Ignore both consistency and diversity.