pretraining

Authors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Synonyms

pre-training

Abstract

Pretraining is the training of a model on a large, generic dataset to produce model parameters that are reused as the starting point for a later learning task. The loss minimized is typically self-supervised, so no hand-annotated labels are needed and the dataset can be scaled up cheaply. For a large language model (LLM), pretraining minimizes the average next-token prediction loss over a large text corpus. Its output is a set of model parameters $\widehat{\mathbf{w}}$, equivalently a learned feature transformation, capturing structure shared across many learning tasks. That is what a foundation model provides, and fine-tuning adapts it to one learning task. Generic features are therefore learned once rather than once per task.

Definition

Before a deep net is asked to detect tumors in a few hundred annotated scans, it is fitted to millions of ordinary photographs that have nothing to do with tumors. That first step is pretraining: the training of a model on a large, generic dataset to produce model parameters that are reused as the starting point for a later learning task.

The loss minimized during pretraining is typically self-supervised (see self-supervised learning): the supervisory signal is constructed from the input data itself, so no hand-annotated labels are needed and the dataset can be scaled up cheaply. For a large language model (LLM), pretraining minimizes the average next-token prediction loss over a large text corpus (Brown et al., 2020). The photographs in the tumor example come from ImageNet, a corpus of more than three million labeled images at its release (Deng et al., 2009).

The output is a set of model parameters $\widehat{\mathbf{w}}$, or equivalently a learned feature transformation $\phi\colon \mathcal{X} \to \mathbb{R}^d$, that captures structure shared across many learning tasks. This is what a foundation model provides: the same ϕ is reused across learning tasks by transfer learning, for instance by appending a small task-specific hypothesis and adapting it through fine-tuning on a task-specific dataset (see Fig. 1). Generic features are therefore learned once, and each later learning task starts from an informative initialization instead of from scratch.

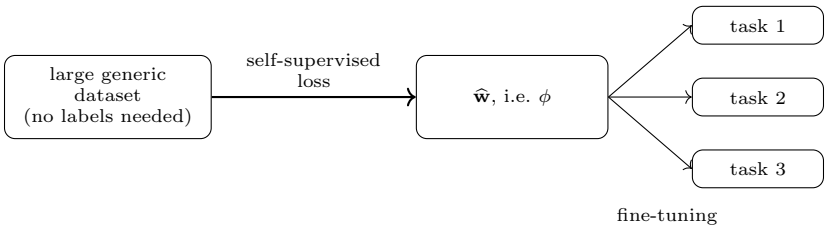

Figure 1: Pretraining fits model parameters $\widehat{\mathbf{w}}$ to a large dataset that needs no labels, because the loss is built from the data itself. The learned feature transformation ϕ is then reused: each later learning task starts from it and is adapted by fine-tuning on its own small dataset, so generic features are learned once rather than once per task

Cross References

fine-tuning, self-supervised learning, foundation model, transfer learning, deep net

References

1. Brown et al. (2020). Language Models are Few-Shot Learners. In: Adv. Neural Inf. Process. Syst., pp. 1877–1901.

2. Deng et al. (2009). ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255.