

Ranking and Tuning Pre-trained Models: A New Paradigm of Exploiting Model Hubs

Kaichao You^{1 *}

Yong Liu^{1 *}

Jianmin Wang¹

Michael I. Jordan²

Mingsheng Long^{1 †}

YOUKAICHAO@GMAIL.COM

LIUYONG1095556447@GMAIL.COM

JIMWANG@TSINGHUA.EDU.CN

JORDAN@CS.BERKELEY.EDU

MINGSHENG@TSINGHUA.EDU.CN

¹ School of Software, BNRist, Tsinghua University, Beijing 100084, China.

² Division of Computer Science and Department of Statistics, UC Berkeley, CA 94720-1776, USA

Abstract

Pre-trained model hubs with many pre-trained models (PTMs) have been a cornerstone in deep learning. Although built at a high cost, they are in fact *under-exploited*: practitioners usually pick one PTM from the provided model hub by popularity, and then fine-tune the PTM to solve the target task. This naïve but common practice poses two obstacles to sufficiently exploiting pre-trained model hubs: (1) the PTM selection procedure has no optimality guarantee; (2) only one PTM is used while the rest PTMs are overlooked. Ideally, to maximally exploit pre-trained model hubs, trying all combinations of PTMs and extensively fine-tuning each combination of PTMs are required, which incurs exponential combinations and unaffordable computational budget. In this paper, we propose a new paradigm of exploiting model hubs by ranking and tuning pre-trained models: (1) Our conference work (You et al., 2021) proposed LogME to estimate the maximum value of label evidence given features extracted by pre-trained models, which can rank all the PTMs in a model hub for various types of PTMs and tasks *before fine-tuning*. (2) the best ranked PTM can be fine-tuned and deployed if we have no preference for the model’s architecture, or the target PTM can be tuned by top-K ranked PTMs via the proposed B-Tuning algorithm. The ranking part is based on the conference paper, and we complete its theoretical analysis (convergence proof of the heuristic evidence maximization procedure, and the influence of feature dimension) in this paper. The tuning part introduces a novel Bayesian Tuning (B-Tuning) method for multiple PTMs tuning, which surpasses dedicated methods designed for homogeneous PTMs tuning and sets up new state of the art for heterogeneous PTMs tuning. We believe the new paradigm of exploiting PTM hubs can interest a large audience of the community.

Keywords: Pre-trained Model Hub, Model Ranking, Model Tuning, Transfer Learning

1. Introduction

Deep neural networks (He et al., 2015, 2016; Devlin et al., 2019) trained by large-scale data (Deng et al., 2009; Russakovsky et al., 2015; Merity et al., 2017) and specialized computational devices (Jouppi et al., 2017) has gone beyond human performance on many recognition tasks in both computer vision and natural language processing. Lots of researchers (Donahue et al., 2014; Girshick et al., 2014; Devlin et al., 2019) found that deep neural networks trained

*. The first two authors contribute equally to the paper.

†. Mingsheng Long is the corresponding author.

on a large-scale pre-training task (Yang et al., 2019; Clark et al., 2020; Brown et al., 2020) can produce generic representations (Donahue et al., 2014) that benefit downstream tasks such as object detection (Girshick et al., 2014) and language understanding (Wang et al., 2019). These trained neural networks are known as pre-trained models (PTMs). Readers can refer to dedicated surveys (Han et al., 2021; Qiu et al., 2020; Bommasani et al., 2021) for a holistic history of pre-trained models. The amazing power (Brown et al., 2020) of PTMs, together with the transfer learning paradigm of “pre-training \rightarrow fine-tuning” to leverage PTMs, has transformed both vision (Kornblith et al., 2019) and language (Devlin et al., 2019) communities, and continues to spread its influence to communities like geometric learning (Hu et al., 2020).

The cost of training PTMs varies from hundreds of GPU *hours* (He et al., 2016) to hundreds of GPU *days* (Devlin et al., 2019), which is too expensive for individual researchers and academic labs to afford. Most pre-trained models are provided by technical tycoons, including PyTorch Hub ¹, TensorFlow Hub ² and HuggingFace Transformer Models ³. A collection of pre-trained models is called a “pre-trained model hub” (PTM Hub), which has become an indispensable ingredient in practitioners’ daily development. Take the HuggingFace Transformer library (Wolf et al., 2020) for example, the most popular BERT model (Devlin et al., 2019) is downloaded over 80 million times every month!

Although technical tycoons spend enormous costs on providing a rich PTM Hub for the public, it is somewhat embarrassing that practitioners usually *pick just one PTM by popularity*, meaning that the whole PTM Hub is insufficiently exploited. Figure 1 analyzes the monthly downloads of popular PTMs in the HuggingFace Transformer library. Except for several popular models, most PTMs in the hub are seldomly exploited. The situation in PyTorch Hub and TensorFlow Hub is quite the same: several popular PTMs are frequently downloaded while others receive little attention.

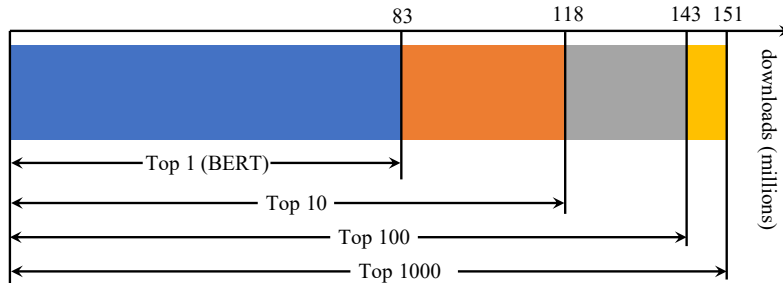


Figure 1: Monthly download statistics (in millions) for top popular models in the HuggingFace Transformer library. The most popular PTM (BERT) takes up more than a half downloads, while other PTMs are much less exploited.

The above naïve practice is far from optimal in two aspects: (1) The optimal PTM selection should be *task-specific*, depending on the compatibility between the pre-trained model and the target task. Not a single PTM can be the best in any downstream tasks (You

1. <https://pytorch.org/hub/>
 2. <https://www.tensorflow.org/hub>
 3. <https://huggingface.co/models>

et al., 2021). (2) Only one PTM can be exploited and the rest PTMs are left on the shelf. Correspondingly, there are two reasons why practitioners resort to the sub-optimal naïve practice: (1) maximally exploiting a PTM hub requires trying all combinations of PTMs and extensively fine-tuning each PTM combination, which is too computationally intensive to afford; (2) even if the humongous computational cost can be paid, it is unclear how to exploit multiple PTMs for transfer learning. Shu et al. (2021) studied the problem in a limited case but a general solution is still lacking, which is discussed in detail in Section 2.4.

To improve the utilization efficiency of a PTM hub, we propose a new paradigm: ranking and tuning pre-trained models. Figure 2 provides an overview of the procedure. It consists of two parts: first, PTMs are *ranked* by a transferability metric; second, top-ranked PTMs are *tuned* according to applications’ requirements. Our preliminary work (You et al., 2021) proposed LogME to estimate the compatibility between PTMs and downstream datasets, and demonstrated its effectiveness on a variety of PTMs and tasks, well supporting the ranking of PTMs. With a provided transferability rank, the best ranked PTM can be fine-tuned if no preference for network architecture is present, or the PTM with desired architecture can be tuned by top-K ranked PTMs via a novel B-Tuning algorithm to sufficiently exploit the model hub.

Compared with the naïve practice, our proposed new paradigm features two significant advantages: first, it provides a *task-adaptive ranking* of all PTMs in a PTM hub on which optimal selection of PTMs can be based; second, it opens the possibility to *exploit multiple PTMs for tuning*, breaking the stereotype that fine-tuning must be tied up with a single PTM. The new paradigm can be interesting to a large audience in the community, as pre-trained models are increasingly important in deep learning.

Besides a new paradigm of exploiting PTM hubs, this paper brings novel theoretical analysis and a new algorithm for multiple PTMs tuning. From the theoretical perspective, we derive the sufficient condition for the evidence maximization algorithm (MacKay, 1992) to converge, and reveal the influence of dimensionality on LogME. The evidence maximization algorithm (MacKay, 1992) has been regarded as a pure heuristic for decades, and to the best of our knowledge, we are the first to successfully derive its convergence condition. From the perspective of algorithm design, we devise a multiple PTMs tuning method named B-Tuning using Bayesian learning, which surpasses the dedicated method (Shu et al., 2021) for homogeneous PTMs and also works for the challenging heterogeneous PTMs setting.

The contributions of this paper are summarized as follows:

1. We propose a new paradigm of exploiting PTM hubs, namely ranking and tuning pre-trained models. It has significant advantages compared with the common practice of naïvely fine-tuning a popular pre-trained model.
2. Concerning ranking PTMs, we propose LogME for transferability assessment, and develop a fast algorithm to accelerate the computation. LogME is easy to interpret and is extremely efficient: it brings at most $3700\times$ speedup in wall-clock time and requires just 1% memory footprint. Theoretical analysis of algorithmic convergence and influence of dimensionality further confirm the rationality of LogME, and lay a theoretical foundation for a decades-long heuristic algorithm in evidence maximization.
3. For tuning PTMs, two possible scenarios are studied. In the academic scenario without specific PTM architecture requirement, the ranking can help select the best ranked pre-trained model for subsequent fine-tuning; in the industrial scenario where specific

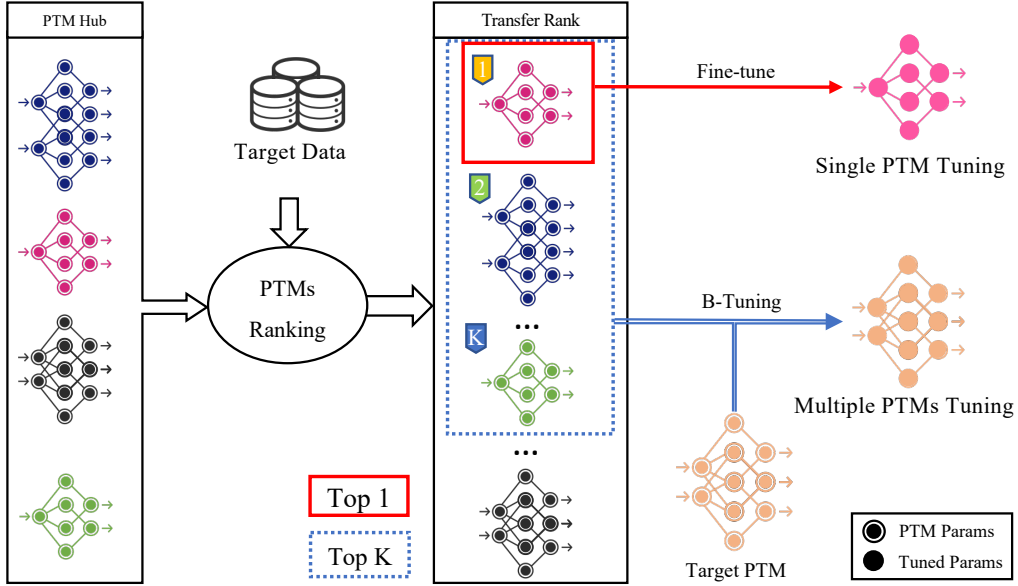


Figure 2: The proposed paradigm of ranking and tuning pre-trained models. PTMs are ranked by their transferability to the target data, then either the best PTM is fine-tuned or the target PTM is tuned by top-K PTMs via the proposed B-Tuning.

PTM architecture is required to meet the computation and energy cost budget, we propose B-Tuning to fine-tune the given pre-trained model with top-K ranked PTMs, even though these PTMs are heterogeneous.

Compared with the conference paper (You et al., 2021) that only proposes LogME for transferability estimation, this paper extends LogME to a paradigm of ranking and tuning pre-trained models. Additional theoretical analyses are available in the ranking part, and a new algorithm is presented in the tuning part. Moreover, LogME is tested against additional tasks like named entity recognition (Sang and De Meulder, 2003) in Section 7.2.5 and prompt learning (Liu et al., 2021) in Section 7.5.

The rest of the paper is organized as follows: Section 2 summarizes related work, Section 3 describes the problem setup, Section 4 introduces LogME (PTM ranking), Section 5 presents LogME’s theoretical results, Section 6 focuses on PTM tuning and the proposed B-Tuning algorithm, Section 7 holds all the experiments, finally Section 8 concludes the paper.

2. Related work

2.1 Transfer learning

Transfer learning (Thrun and Pratt, 1998) consists of transductive transfer, inductive transfer, and task transfer. In transductive transfer, a commonly known paradigm is domain adaptation (Quionero-Candela et al., 2009; Long et al., 2015; Ganin and Lempitsky, 2015), which is concerned with reducing domain shifts by transferring samples, hidden features and so forth. Inductive transfer, namely fine-tuning in deep learning (Erhan et al., 2010; Yosinski

et al., 2014), utilizes prior knowledge (pre-trained models) to improve the performance of target tasks. In task transfer learning (Zamir et al., 2018), researchers investigate how to transfer among tasks rather than pre-trained models. It aims to discover the relevance among tasks (Ben-David and Schuller, 2003) and to exploit the relationship for improvement on the target task. In deep learning, transfer learning usually means inductive transfer with a pre-trained model, the topic we focus on in this paper.

Many previous works (Yosinski et al., 2014; Kornblith et al., 2019; Neyshabur et al., 2020) have revealed the benefit of initializing a deep neural network with a pre-trained model. Apart from the vanilla method (*i.e.*, pure initialization), researchers have proposed more sophisticated fine-tuning techniques like regularization (Li et al., 2018; Chen et al., 2019), additional supervision (You et al., 2020), and carefully designed architecture (Kou et al., 2020). They can further improve transfer learning performance, but generally *these fine-tuning methods do not change the ranking of pre-trained models on downstream tasks*. That is, pre-trained model A would be better than pre-trained model B when those advanced techniques are turned on, if A is better than B after vanilla fine-tuning. For example, on three datasets and four sampling rates from Table 2 in You et al. (2020), better fine-tuning performance primarily indicates better Co-Tuning (their proposed method) performance, implying that the transferability of a pre-trained model is *task-specific* but not *method-specific*. Therefore, our experiments stick to the vanilla fine-tuning during PTM ranking.

2.2 PTMs and PTM hubs

Pre-trained models (PTMs) are generalizable deep networks trained on large-scale data and can be transferred to a series of downstream tasks. They have become a cornerstone in deep learning and sometimes are called foundation models (Bommasani et al., 2021). Typical kinds of PTMs are summarized in the following.

Supervised PTMs. In the ImageNet classification challenge, He et al. (2015) developed the first deep neural network that outperforms human beings. By supervised pre-training on the ImageNet dataset, deep models are marching towards higher accuracy, fewer parameters and lower computation. InceptionNet (Szegedy et al., 2015) makes use of parallel convolutional filters to extract different levels of features. ResNet (He et al., 2016) introduces skip connections to ease the vanishing gradient problem so that much deeper networks can be trained. DenseNet (Huang et al., 2017) is equipped with carefully designed densely connected blocks, with the advantages of better feature reuse and higher parameter efficiency (Huang et al., 2019). MobileNet (Sandler et al., 2018) is a low-parameter, mobile-friendly network structure which is further optimized with the help of network architecture search to become MNASNet (Tan et al., 2019).

Unsupervised PTMs. Although supervised pre-training is the most common practice, the labeling cost of large-scale data is too expensive to afford. As a large amount of unlabeled data on the Internet are available but under-exploited, recently many researchers have sought to apply self-supervised learning (Jing and Tian, 2020) on unlabeled data (Mahajan et al., 2018) with contrastive loss (Gutmann and Hyvärinen, 2010). And then, a family of unsupervised deep models have emerged in recent years. He et al. (2020) proposed Momentum Contrast with a creative queue structure to fully exploit the implicit representations of unlabeled data. Chen et al. (2020a) greatly improved the performance by exploring data

augmentation, multi-layer projection head and empirical designs. Designing better contrastive pre-training strategies is still under active research (Tian et al., 2020).

Language PTMs. Natural language processing, the crown jewel of artificial intelligence, has been revolutionized by language PTMs in recent years. Unsupervised pre-trained models have been well established by training masked language models (Devlin et al., 2019) or autoregressive language models (Yang et al., 2019) on large unlabeled corpora (Merity et al., 2017). Liu et al. (2019) explored many practical details on how to improve the training of language models. Sanh et al. (2019) proposed distillation to get smaller and faster models. These pre-trained language models become an indispensable component in winning submissions on common benchmarks like GLUE (Wang et al., 2018) and SQuAD (Rajpurkar et al., 2016), and have a profound influence in the industry.

PTMs are grouped together to be hosted in PTM Hubs like TorchVision⁴ and Hugging-Face⁵. Technical giants spend humongous resources on training so many PTMs, but sadly speaking, PTM Hubs are rather under-exploited, as quantitatively measured in Figure 1 and detailedly described in the introduction section. The goal of this paper is to develop a new paradigm of exploiting PTM hubs, so that pre-trained models can be sufficiently utilized.

2.3 Assessing the transferability of pre-trained models

Assessing the transferability of PTMs has great significance to guide deep learning practice. It can be used to rank available PTMs and act as a criterion for pre-trained model selection. Yosinski et al. (2014) studied the performance of transferring different layers of a pre-trained model and Kornblith et al. (2019) studied a wide variety of ImageNet PTMs with modern network architectures. These papers aim for a deep understanding (Neyshabur et al., 2020) of transfer learning by expensive and exhaustive fine-tuning with humongous computation cost (Section 7.4) which is hard for practitioners to afford. In most scenarios, practitioners care most about PTMs’ relative ranking on target tasks to guide PTM selection, requiring a practical assessment method that is *efficient*, *accurate* and *general*: a transferability assessment method should be efficient enough compared with brute-force fine-tuning (Zamir et al., 2018), should be accurate enough so that the potentially best models can be identified, and should be general enough to tackle a wide variety of common scenarios.

LEEP (Nguyen et al., 2020) and NCE (Tran et al., 2019) are the first two methods for assessing the transferability of pre-trained models. Nguyen et al. (2020) constructed an empirical predictor by the joint distribution over pre-trained labels and the target labels, and calculated the log expectation of the empirical predictor (LEEP) as the transferability measure. Negative Conditional Entropy (NCE) proposed by Tran et al. (2019) is based on an information-theoretic quantity (Cover, 1999), which reveals the transferability and hardness between different tasks.

However, the above methods leave plenty of room for further improvement. They can only handle classification tasks with supervised pre-trained models. Increasingly popular contrastive pre-trained models and language models are out of prior methods’ scope. In this paper, we propose LogME to greatly extend the applicability of transferability assessment. LogME is fast to compute, less prone to over-fitting and broadly applicable to various

4. <https://pytorch.org/vision/stable/models.html>

5. <https://huggingface.co/models>

Table 1: Applicability of prior methods and LogME proposed in this paper.

Modality	Pre-train	Target	Method		
			LEEP	NCE	LogME
vision	classification	classification	✓	✓	✓
	classification	regression	✗	✗	✓
	contrastive	classification	✗	✗	✓
	contrastive	regression	✗	✗	✓
language	language modeling	classification	✗	✗	✓

pre-trained models, downstream tasks, and data modalities. Its performance is widely validated in later experiments. Table 1 shows the applicability of several transferability assessment methods. Prior to this paper, for most (4 out of 5) transfer learning settings, task adaptive transferability assessment does not have a decent solution. In addition, LogME’s statistical rigor makes it extensible to multiple PTMs tuning (Section 6.2), which completes the new paradigm of ranking and tuning pre-trained models in this paper.

2.4 Multiple PTMs tuning

Transfer learning in deep learning commonly refers to reusing pre-trained models. The straightforward approach to reuse PTMs is to copy pre-trained parameters for initialization (also known as fine-tuning), which is simple enough to receive tremendous popularity. We call this approach “*single PTM tuning*” because it can only exploit a specific pre-trained model during fine-tuning.

It is widely acknowledged that the success of transfer learning is attributed to the knowledge in the pre-trained model. Considering that there are so many PTMs in a PTM hub, it is rather appealing to transfer multiple PTMs simultaneously, a problem we call “*multiple PTMs tuning*”. With an appropriate method, it is natural to expect that multiple PTMs tuning can outperform single PTM tuning.

Unfortunately, multiple PTMs tuning is under-explored due to its technical difficulty. If multiple PTMs are homogeneous, *i.e.*, they share the same network architecture, the problem would be easier. Researchers in this area focus on how to align and merge multiple PTMs. Singh and Jaggi (2020) define transportation cost between neural representations and minimize the induced Wasserstein distance to align neurons from each PTM. Shu et al. (2021) developed a channel-wise alignment method dedicated to convolutional neural networks. They also proposed a learnable gating function to merge multiple PTMs. Prior to this paper, Shu et al. (2021) hold the state-of-the-art result in homogeneous PTMs tuning.

Heterogeneous PTMs tuning is much more difficult than homogeneous PTMs tuning. It is still an unanswered question because the heterogeneity among PTMs makes it hard to tackle. In practice, however, this challenging problem is more important as pre-trained models in PTM hubs have different architectures and heterogeneity is common.

This paper tries to exploit PTM hubs as maximally as possible. Following the “many could be better than all” (Zhou et al., 2002) principle in ensemble learning, PTMs are first ranked by LogME, then top-K related PTMs from the whole PTM hub are selected, casting

the problem into a multiple PTMs tuning problem. A Bayesian tuning method (B-Tuning, in Section 6.2) is further proposed to solve the multiple PTMs tuning problem. Excitingly, our method is capable of both homogeneous PTMs and heterogeneous PTMs. It fills the blank in heterogeneous PTMs tuning, and surpasses the state-of-the-art method (Shu et al., 2021) dedicated to homogeneous PTMs tuning.

3. Overview of the new paradigm

Suppose a PTM hub has M pre-trained models $\{\phi_k\}_{k=1}^M$, and the transfer learning task is given by its labeled dataset $\mathcal{D} = \{(x_i, Y_i)\}_{i=1}^n$ with n labeled data points. This paper focuses on classification and regression tasks, so the label $Y_i \in \mathbb{R}^C$ is C dimensional. To improve the transfer learning performance on the given task, we aim to develop a solution that can exploit the whole PTM hub.

Before this paper, a naïve but common practice is to select a popular pre-trained model ϕ and to fine-tune it on the target task. To sufficiently exploit the PTM hub, we propose a new paradigm in this paper: **ranking and tuning pre-trained models**. Accordingly, the rest of the method description is divided into two parts: Section 4 focuses on ranking and introduces a transferability metric named LogME; Section 6 focuses on tuning and introduces a novel B-Tuning method for multiple PTMs tuning. Section 5 lies in between to hold theoretical analysis about LogME.

Note that this paper contains heavy math and many notations. For the convenience of readers, all the notations and their meanings are listed in Table 8 for possible reference.

4. Ranking pre-trained models

It requires a transferability metric to rank all the pre-trained models in a PTM hub. But before diving into the transferability metric, the metric itself requires a measurement for its fidelity to the ground-truth transferability performance, which is elaborated in the following.

4.1 How to measure the performance of a transferability metric?

The transfer learning task should have an evaluation metric (accuracy, MAP, MSE *etc.*) to measure the ground-truth transfer performance T_k of fine-tuning ϕ_k with proper hyper-parameter setting. A practical assessment method should produce a score S_k for each pre-trained model ϕ_k (ideally without fine-tuning ϕ_k on \mathcal{D}), and the scores $\{S_k\}_{k=1}^M$ should well correlate with $\{T_k\}_{k=1}^M$ so that top-performing pre-trained models can be selected by simply evaluating the scores.

A perfect pre-trained model assessing method would output $\{S_k\}_{k=1}^M$ with precisely the same order as $\{T_k\}_{k=1}^M$. To measure the deviation from the perfect method, we can use simple metrics like **top-1 accuracy or top-K accuracy** (whether top-K in $\{S_k\}_{k=1}^M$ are also top-K in $\{T_k\}_{k=1}^M$). But top-1 accuracy is too conservative and top-K accuracy is not comparable across different values of M . Therefore, we turn to **rank correlation** (Fagin et al., 2003) to directly measure the correlation between $\{S_k\}_{k=1}^M$ and $\{T_k\}_{k=1}^M$. The prior work (Nguyen et al., 2020) adopted Pearson’s linear correlation coefficient, but neither Pearson’s linear correlation nor its variant (Spearman’s rank correlation) has a simple interpretation (see the interpretation of τ below).

Since the purpose of a transferability metric is to choose a good pre-trained model, we hope T_i is better than T_j if S_i is better than S_j , which can be well captured by Kendall's τ coefficient (Kendall, 1938) as described in the following.

To simplify the discussion, assume larger values of transfer performance T and score S are preferred (*e.g.* accuracy). If this is not the case (*e.g.* transfer performance is measured by mean square error), the negation of T can be considered. For a pair of measures (T_i, S_i) and (T_j, S_j) , the pair is **concordant** if $T_i < T_j \wedge S_i < S_j$ or $T_i > T_j \wedge S_i > S_j$ (concisely speaking, $\text{sgn}(T_i - T_j)\text{sgn}(S_i - S_j) = 1$). The Kendall's τ coefficient is defined by the following equation, which enumerates all $\binom{M}{2}$ pairs and counts the number of concordant pairs minus the number of discordant pairs.

$$\tau = \frac{2}{M(M-1)} \sum_{1 \leq i < j \leq M} \text{sgn}(T_i - T_j)\text{sgn}(S_i - S_j)$$

How to interpret τ (Fagin et al., 2003). The range of τ is $[-1, 1]$. $\tau = 1$ means T and S are perfectly correlated ($S_i > S_j \iff T_i > T_j$), and $\tau = -1$ means T and S are reversely correlated ($S_i > S_j \iff T_i < T_j$). If T and S have correlation of τ , the probability of $T_i > T_j$ is $\frac{\tau+1}{2}$ when $S_i > S_j$.

Pay attention to top-performing models. Since a major application of transferability metric is to select top-performing pre-trained models, discordant / concordant pairs should be weighted more if T_i, T_j, S_i, S_j are larger. This can be taken care of by τ_w (Vigna, 2015), a weighted variant of Kendall's τ . The details of calculating τ_w can be found in the implementation⁶ from SciPy (Virtanen et al., 2020).

In short, we measure the correlation between $\{S_k\}_{k=1}^M$ and $\{T_k\}_{k=1}^M$ by τ_w (Vigna, 2015). Larger τ_w indicates better correlation and better assessment.

4.2 The LogME approach

Since a transferability metric is designed to measure the transferability of pre-trained models, it should produce a score S_k for each pre-trained model ϕ_k independent of the rest of pre-trained models. We thus drop the subscript k in this section.

An important goal of designing transferability metric is to quickly assess a large number of PTMs. We thus treat minimizing assessment time as a priority. First, to avoid expensive gradient optimization, we regard pre-trained model ϕ as a fixed feature extractor. Then features $\{f_i = \phi(x_i)\}_{i=1}^n$ and labels $\{Y_i\}_{i=1}^n$ are all we can get to assess pre-trained models. Note that Nguyen et al. (2020) used a pre-trained classification head h besides the pre-trained representation model ϕ , limiting their method to supervised pre-trained models. In contrast, we only use the pre-trained representation model ϕ so that the proposed method can be applied to any pre-trained model (whether supervised pre-trained or unsupervised pre-trained).

Without gradient optimization, the problem is cast into estimating the compatibility of features $\{f_i = \phi(x_i)\}_{i=1}^n$ and labels $\{Y_i\}_{i=1}^n$, which is discussed in the rest of this section.

4.2.1 EVIDENCE CALCULATION

We first consider a simple case with D -dimensional features $f_i \in \mathbb{R}^D$ and scalar labels $y_i \in \mathbb{R}$. Note that the actual label Y_i can be non-scalar, and how to extend from the scalar label y_i

6. <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.weightedtau.html>

to the vector label Y_i is explained in Section 4.2.2. The feature matrix $F \in \mathbb{R}^{n \times D}$ contains all the features and $y \in \mathbb{R}^n$ denotes all the labels.

A direct measurement of the compatibility between features F and labels y is the probability density $p(y|F)$, which is intractable without a parametrized model. Since the rule-of-thumb transfer learning practice is to add a fully connected layer on top of the pre-trained model, we use a linear model upon features parameterized by w .

A naïve approach to deal with the linear model is to find the best w^* by logistic or linear regression and to assess pre-trained models by likelihood $p(y|F, w^*)$. However, it is well-known that *likelihood is prone to over-fitting* (Bishop, 2006), which is experimentally observed in Section 7.5. A better approach is to use the evidence (marginalized likelihood) $p(y|F) = \int p(w)p(y|F, w)dw$, which integrates over all possible choices of w and is better than simply using one optimal value w^* . This evidence-based approach is an elegant model selection approach and has a rigorous theoretical foundation (Knuth et al., 2015). For $p(w)$ and $p(y|F, w)$, we use the commonly adopted graphical model (Figure 3) specified by two positive hyper-parameters α and β : the prior distribution of the weight is an isotropic multivariate Gaussian $w \sim \mathcal{N}(0, \alpha^{-1}I)$, and the distribution of each observation is a one-dimensional normal distribution $p(y_i|f_i, w, \beta) \sim \mathcal{N}(y_i|w^T f_i, \beta^{-1})$.

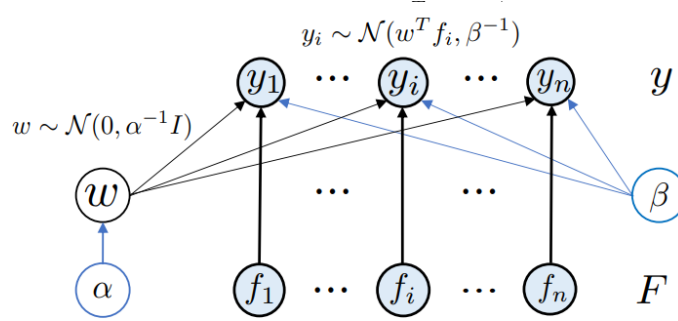


Figure 3: The directed graphical model for calculating evidence.

According to the causal structure in Figure 3 and the basic principles in graphical models (Koller and Friedman, 2009), the evidence can be calculated analytically as follows:

$$p(y|F, \alpha, \beta) = \int p(w|\alpha) \prod_{i=1}^n p(y_i|f_i, w, \beta) dw = \left(\frac{\beta}{2\pi}\right)^{\frac{n}{2}} \left(\frac{\alpha}{2\pi}\right)^{\frac{D}{2}} \int e^{-\frac{\alpha}{2} w^T w - \frac{\beta}{2} \|Fw - y\|^2} dw \quad (1)$$

As $\int e^{-\frac{1}{2}(w^T A w + b^T w + c)} dw = \sqrt{\frac{(2\pi)^D}{|A|}} e^{-\frac{1}{2}c + \frac{1}{8}b^T A^{-1}b}$ when A is positive definite, Equation 1 can be simplified. By taking the logarithm to make the equation simple, Equation 2 shows the logarithm of the evidence as a function of α, β , where $A = \alpha I + \beta F^T F$, $m = \beta A^{-1} F^T y$.

$$\mathcal{L}(\alpha, \beta) = \log p(y|F, \alpha, \beta) = \frac{n}{2} \log \beta + \frac{D}{2} \log \alpha - \frac{n}{2} \log 2\pi - \frac{\beta}{2} \|Fm - y\|_2^2 - \frac{\alpha}{2} m^T m - \frac{1}{2} \log |A| \quad (2)$$

4.2.2 EVIDENCE MAXIMIZATION AND LOGME

A remaining issue of Equation 2 is how to determine α, β . Gull (1989) suggested that we should choose α, β to maximize the evidence, i.e. use $(\alpha^*, \beta^*) = \arg \max_{\alpha, \beta} \mathcal{L}(\alpha, \beta)$. Because m and A are coupled, maximizing $\mathcal{L}(\alpha, \beta)$ is generally a difficult problem. Fortunately, MacKay (1992) proposed a heuristic algorithm to solve the maximization problem: (1)

set initial value of α, β ; (2) evaluate A, m, γ with given α, β : $A = \alpha I + \beta F^T F, m = \beta A^{-1} F^T y, \gamma = \sum_{i=1}^D \frac{\beta \sigma_i^2}{\alpha + \beta \sigma_i^2}$, where σ_i are singular values of F ; (3) maximize α, β by solving $\frac{\partial \mathcal{L}}{\partial \alpha} = 0, \frac{\partial \mathcal{L}}{\partial \beta} = 0$ with m, γ fixed, which yields $\alpha \leftarrow \frac{\gamma}{m^T m}, \beta \leftarrow \frac{n - \gamma}{\|Fm - y\|_2^2}$. The algorithm is called **MacKay’s algorithm** (Algorithm 2), and has been heuristic since it was proposed by MacKay (1992). Section 5.1 gives the first theoretical analysis for the convergence guarantee. Interestingly, the fixed point iteration used for the convergence analysis brings a new and faster algorithm for evidence maximization (Algorithm 3). Please refer to Section 5.1 for details.

After the convergence of evidence maximization, **the logarithm maximum evidence $\mathcal{L}(\alpha^*, \beta^*)$ is used to evaluate the compatibility between features and labels.** Because $\mathcal{L}(\alpha^*, \beta^*)$ scales linearly with n , we normalize it as $\frac{\mathcal{L}(\alpha^*, \beta^*)}{n}$ and term it LogME (logarithm of maximum evidence). Discussion on the influence of dimensionality D is presented in Section 5.2. **LogME can be intuitively interpreted as the average maximum logarithm evidence of labels given the pre-trained features.**

Extending LogME to complex cases. The LogME approach described above starts from a single-target regression. If the target problem is a multivariate-regression task, *i.e.* $Y \in \mathbb{R}^{n \times C}$, we can calculate LogME for each dimension c ($1 \leq c \leq C$) and average them over the C dimension. If the target problem is a classification task with C classes, Equation 1 cannot be calculated analytically (Daunizeau, 2017) with a categorical prior distribution, **but we can convert the labels to one-hot labels and treat the problem as multivariate regression.** Therefore, LogME can be used on both classification and regression tasks. The overall algorithm of LogME is described in Algorithm 1.

4.2.3 COMPUTATIONAL SPEEDUP

Although the Bayesian approach of maximum evidence has many nice properties (Knuth et al., 2015), it inherits the common drawback of Bayesian methods with high computational complexity. The naïve implementation of Algorithm 1 has a complexity of $\mathcal{O}(CD^3 + nCD^2)$. For typical usage with $D \approx 10^3, n \approx 10^4, C \approx 10^3$, the computational cost is 10^{13} , making the wall-clock time comparable to fine-tuning a pre-trained model ϕ .

Our conference paper (You et al., 2021) accelerates the computation by avoiding matrix inversion and matrix-matrix multiplication, as shown in Line 8 of Algorithm 2. In this paper, we present a convergence analysis of the MacKay’s algorithm by fixed point iteration, and find that it can be designed as a faster algorithm for evidence maximization. The algorithm is presented in Algorithm 3 and its rationale is explained in Section 5.1.

Table 2 compares the complexity of calculating LogME with different implementations of evidence maximization. The naïve implementation is biquadratic, You et al. (2021) makes it cubic, and this paper further reduces the number of cubic terms. The optimized algorithms make a time-consuming Bayesian approach fast enough, reducing the wall-clock time by the order of 10^2 (see Section 7.4 for empirical measurement of the speedup). Note that three implementation methods are functionally equivalent and only differ in the computational complexity. Therefore *the fixed point iteration proposed in this paper is always preferred in our implementation.*

Algorithm 1 LogME

-
- 1: **Input:** Pre-trained model ϕ and target dataset $\mathcal{D} = \{(x_i, Y_i)\}_{i=1}^n$
 - 2: **Output:** Logarithm of Maximum Evidence (LogME)
 - 3: Extract features using pre-trained model ϕ : $F \in \mathbb{R}^{n \times D}$, $f_i = \phi(x_i)$, $Y \in \mathbb{R}^{n \times C}$
 - 4: Compute SVD of F : $F = U\Sigma V^T$. Then $F^T F = V \text{diag}\{\sigma^2\} V^T$
 - 5: **for** dimension $c = 1$ to C **do**
 - 6: Let $y = Y^{(c)} \in \mathbb{R}^n$,
 - 7: Calculate the LogME value \mathcal{L}_c by evidence maximization (Algorithm 2 or Algorithm 3).
 - 8: **end for**
 - 9: Return LogME $\frac{1}{C} \sum_{c=1}^C \mathcal{L}_c$
-

Algorithm 2 Evidence Maximization by MacKay's Algorithm

-
- 1: **Input:** Extracted features $F \in \mathbb{R}^{n \times D}$ and corresponding labels $y \in \mathbb{R}^n$
 - 2: **Output:** Logarithm of Maximum Evidence (LogME)
 - 3: **Note:** F has been pre-decomposed into $F = U\Sigma V^T$
 - 4: Initialize $\alpha = 1, \beta = 1$
 - 5: **while** α, β not converge **do**
 - 6: Compute $\gamma = \sum_{i=1}^D \frac{\beta \sigma_i^2}{\alpha + \beta \sigma_i^2}, \Lambda = \text{diag}\{(\alpha + \beta \sigma^2)\}$
 - 7: **Naïve:** $A = \alpha I + \beta F^T F, m = \beta A^{-1} F^T y$
 - 8: **Optimized** by You et al. (2021): $m = \beta(V(\Lambda^{-1}(V^T(F^T y))))$
 - 9: Update $\alpha \leftarrow \frac{\gamma}{m^T m}, \beta \leftarrow \frac{n-\gamma}{\|Fm-y\|_2^2}$
 - 10: **end while**
 - 11: Compute and return $\mathcal{L} = \frac{1}{n} \mathcal{L}(\alpha, \beta)$ using Equation 2
-

Algorithm 3 Evidence Maximization by Optimized Fixed Point Iteration

-
- 1: **Input:** Extracted features $F \in \mathbb{R}^{n \times D}$ and corresponding labels $y \in \mathbb{R}^n$
 - 2: **Output:** Logarithm of Maximum Evidence (LogME)
 - 3: **Require:** Truncated SVD of F : $F = U_r \Sigma_r V_r^T$, with $U_r \in \mathbb{R}^{n \times r}, \Sigma_r \in \mathbb{R}^{r \times r}, V_r \in \mathbb{R}^{D \times r}$.
 - 4: **Compute** the first r entries of $z = U_r^T y$
 - 5: **Compute** the sum of remaining entries $\Delta = \sum_{i=r+1}^n z_i^2 = \sum_{i=1}^n y_i^2 - \sum_{i=1}^r z_i^2$
 - 6: Initialize $\alpha = 1, \beta = 1, t = \frac{\alpha}{\beta} = 1$
 - 7: **while** t not converge **do**
 - 8: Compute $m^T m = \sum_{i=1}^r \frac{\sigma_i^2 z_i^2}{(t + \sigma_i^2)^2}, \gamma = \sum_{i=1}^r \frac{\sigma_i^2}{t + \sigma_i^2}, \|Fm - y\|_2^2 = \sum_{i=1}^r \frac{z_i^2}{(1 + \sigma_i^2/t)^2} + \Delta$
 - 9: Update $\alpha \leftarrow \frac{\gamma}{m^T m}, \beta \leftarrow \frac{n-\gamma}{\|Fm-y\|_2^2}, t = \frac{\alpha}{\beta}$
 - 10: **end while**
 - 11: Compute $m = V_r \Sigma'_r z$, where $\Sigma'_{ii} = \frac{\sigma_i}{t + \sigma_i^2} (1 \leq i \leq r)$.
 - 12: Compute and return $\mathcal{L} = \frac{1}{n} \mathcal{L}(\alpha, \beta)$ using Equation 2
-

Table 2: Complexity of Algorithm 1 with three implementations of evidence maximization. n, C are the number of samples and the number of classes in classification (or the number of target variables in regression) in downstream tasks, and D is the dimension of features produced by a pre-trained model, respectively.

Evidence maximization method	Complexity per while-loop	Overall complexity
naïve implementation	$\mathcal{O}(D^3 + nD^2)$	$\mathcal{O}(nCD^2 + CD^3)$
optimized by You et al. (2021)	$\mathcal{O}(D^2 + nD)$	$\mathcal{O}(nD^2 + nCD + CD^2 + D^3)$
fixed point iteration (this paper)	$\mathcal{O}(n)$	$\mathcal{O}(nD^2 + nCD)$

5. Theoretical analysis of LogME

In this section, we analyze two theoretical aspects of the proposed LogME, which further explains the rationality behind the LogME algorithm and why LogME works.

5.1 Convergence analysis of evidence maximization

Historical remarks: The evidence maximization procedure in Section 4.2.2 was firstly proposed in MacKay (1992) as a heuristic method to maximize the evidence of given data, following the spirit of empirical Bayesian learning (Bishop, 1995). Its theoretical analysis is lacking and remains heuristic in modern machine learning textbooks like Murphy (2012). In 2016, researchers (Li et al., 2016) revealed that if the predictive uncertainty β is known, the maximization over model uncertainty α can be viewed as a special instantiation of the EM algorithm (Dempster et al., 1977). However, pre-determining β is unrealistic, and people prefer to simultaneously maximize α, β , which is hard to analyze. To the best of our knowledge, we are the first to give a theoretical analysis of the MacKay’s algorithm when α, β are simultaneously optimized.

For the convenience of readers, we collect necessary notations here: n is the number of data examples; D is the size of feature dimensionality; $F \in \mathbb{R}^{n \times D}$ is the feature matrix, with $r = \text{rank}(F)$ being its rank; $y \in \mathbb{R}^n$ is the label vector of data examples. We can immediately get: $r \leq \min\{n, D\}$.

The key in our analysis is to take full advantage of the singular value decomposition of the feature matrix $F = U\Sigma V^T$, where $U \in \mathbb{R}^{n \times n}$, $V \in \mathbb{R}^{D \times D}$, and $\Sigma \in \mathbb{R}^{n \times D}$. Note that Σ only has r non-zero entries: $\Sigma_{ii} = \sigma_i > 0$ ($1 \leq i \leq r$) where σ_i^2 is the i -th largest eigenvalue of $F^T F$ and $\sigma_i = 0$ ($r + 1 \leq i \leq \max(n, D)$). To simplify the computation, let $z = U^T y$ be the transformed y under orthogonal basis U , *i.e.* $y = Uz$.

The MacKay’s algorithm (Algorithm 2) consists of a while-loop. The key to analyze the whole algorithm is to analyze each iteration of the while-loop, which is listed in Algorithm 4 for readers’ convenience. During each iteration, new values α', β' are computed based on initial values α, β , which can be regarded as evaluating a vector-valued function $(\alpha', \beta') = g(\alpha, \beta)$.

The MacKay’s algorithm converges if and only if $\alpha' = \alpha, \beta' = \beta$ after some iterations. With F, y as constants, this corresponds to finding the fixed point of the vector-valued function g , *i.e.*, find α, β such that $(\alpha, \beta) = g(\alpha, \beta)$.

Algorithm 4 One iteration of evidence maximization in Algorithm 2.

- 1: Input: α, β ; Output: α', β' for the next iteration.
 - 2: Compute $A = \alpha I + \beta F^T F, m = \beta A^{-1} F^T y, \gamma = \sum_{i=1}^D \frac{\beta \sigma_i^2}{\alpha + \beta \sigma_i^2}$
 - 3: Return $\alpha' = \frac{\gamma}{m^T m}, \beta' = \frac{n - \gamma}{\|Fm - y\|_2^2}$
-

In general, fixed points of vector-valued functions are difficult to analyze and visualize. Fortunately, we find that the vector-valued function $(\alpha', \beta') = g(\alpha, \beta)$ is homogeneous: $g(k\alpha, k\beta) = kg(\alpha, \beta), \forall k > 0$. Let $t = \alpha/\beta$, and $t' = \alpha'/\beta'$, the vector-valued function $(\alpha', \beta') = g(\alpha, \beta)$ induces a scalar function $t' = f(t)$, whose explicit form can be derived in Theorem 1. Evaluating $g(\alpha, \beta)$ is equivalent to calculating $f(\frac{\alpha}{\beta})$, which is easier to analyze.

Theorem 1 Algorithm 4 induces a scalar function (Equation 3) with $t = \frac{\alpha}{\beta}$ and $t' = \frac{\alpha'}{\beta'}$.

$$t' = f(t) = \left(\frac{n}{n - \sum_{i=1}^D \frac{\sigma_i^2}{t + \sigma_i^2}} - 1 \right) t^2 \frac{\sum_{i=1}^n \frac{z_i^2}{(t + \sigma_i^2)^2}}{\sum_{i=1}^n \frac{\sigma_i^2 z_i^2}{(t + \sigma_i^2)^2}} \quad (3)$$

The proof is in Appendix B. By analyzing the fixed point of $f(t)$, we can understand the convergence property of the MacKay’s algorithm. Although $f(t)$ seems very complicated and a complete understanding of its behavior is difficult to derive, surprisingly the existence of fixed point can be guaranteed with an interpretable condition, which is presented in the following Theorem 2.

Theorem 2 If $r < n$ and $\sum_{1 \leq i, j \leq n} (z_i^2 - z_j^2)(\sigma_i^2 - \sigma_j^2) > 0$, then $f(t)$ has a fixed point and thus the MacKay’s algorithm will converge.

The proof is in Appendix C. Theorem 2 requires two conditions to guarantee the fixed point: $r < n$ and $\sum_{1 \leq i, j \leq n} (z_i^2 - z_j^2)(\sigma_i^2 - \sigma_j^2) > 0$. The first condition is easy to interpret and can be easily satisfied. Typically we have $n > D$, therefore $n > D \geq r$ naturally holds. The condition $\sum_{1 \leq i, j \leq n} (z_i^2 - z_j^2)(\sigma_i^2 - \sigma_j^2) > 0$ is new in this paper. Note that $z = U^T y$ and $z_i = U_i^T y$, where U_i (the i -th column of U) is the left-singular vector of the singular value σ_i , which means that z_i is the projection of label vector y in the direction of the left-singular vector for the singular value σ_i . Intuitively speaking, $\sum_{1 \leq i, j \leq n} (z_i^2 - z_j^2)(\sigma_i^2 - \sigma_j^2) > 0$ requires z_i^2 share roughly the same descending order as σ_i^2 . For larger σ_i^2 (*i.e.* smaller i), it means the projection of y in the corresponding left-singular vector should be larger, which can be interpreted as a rigorous way to say that *labels y are meaningful with respect to the features F* . We would like to emphasize that the requirement on the order of z_i^2 is *soft*: strict ordering $z_i^2 \geq z_j^2 \iff i \leq j \iff \sigma_i^2 \geq \sigma_j^2$ certainly assures the convergence condition $\sum_{1 \leq i, j \leq n} (z_i^2 - z_j^2)(\sigma_i^2 - \sigma_j^2) > 0$, but as long as most z_i^2 follow the ordering, the condition can be satisfied. We find that all experiments in this paper admit the convergence condition, *i.e.*, the evidence maximization algorithm are guaranteed to converge if the data is meaningful. For example, Figure 4 plots $f(t)$ on the CIFAR10 dataset, which clearly shows that $\lim_{t \rightarrow \infty} \frac{f(t)}{t} < 1$, so the convergence condition $\sum_{1 \leq i, j \leq n} (z_i^2 - z_j^2)(\sigma_i^2 - \sigma_j^2) > 0$ holds.

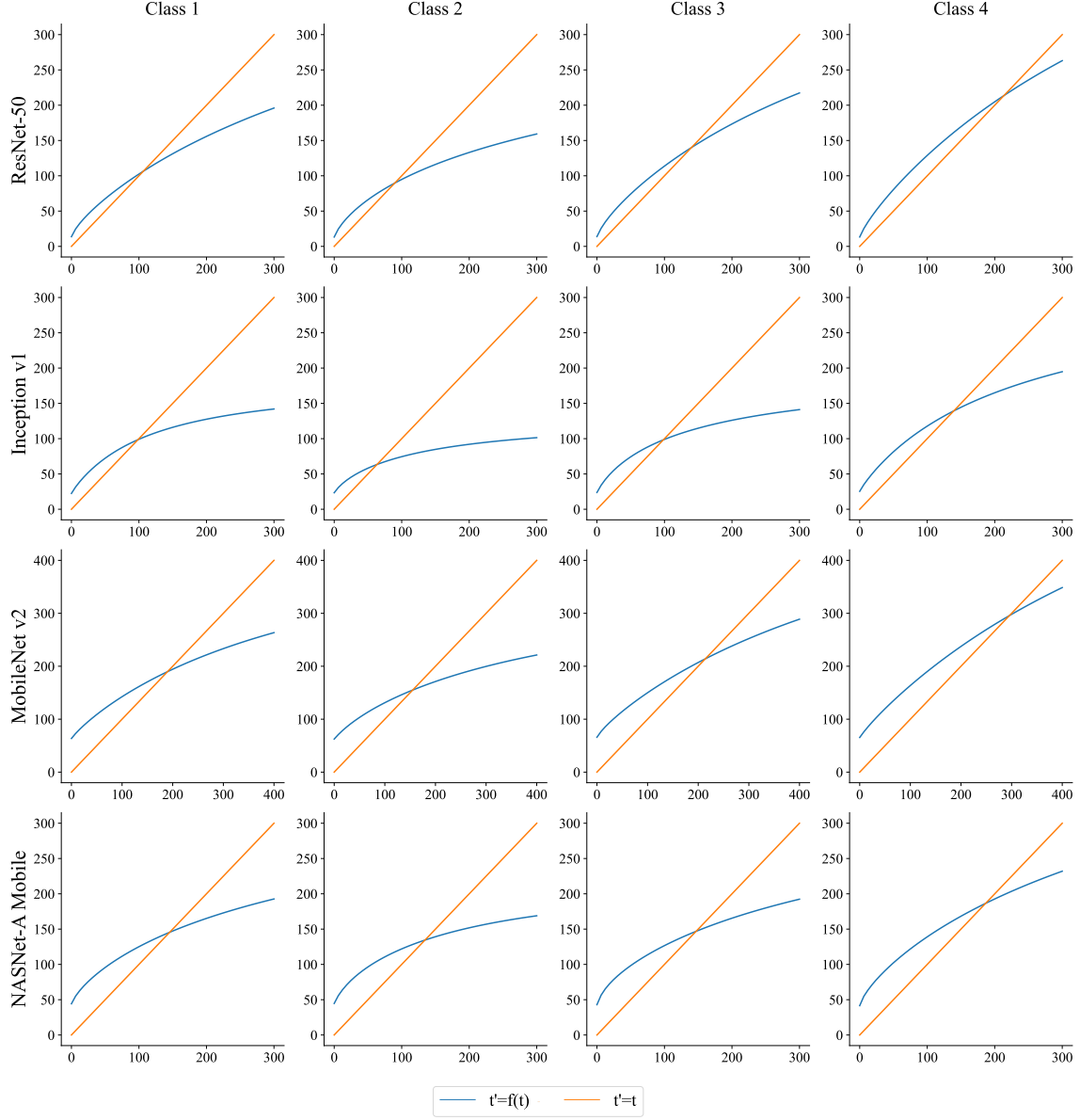


Figure 4: Fixed points of $f(t)$ in Equation 3 for the first 4 classes in CIFAR10 with 4 pre-trained models. Full figure for all the 10 classes in CIFAR10 with 5 pre-trained models can be found in Figure 13, which is omitted here to prettify the layout. We plot $t' = f(t)$ (in blue) and $t' = t$ (in orange), whose intersections are fixed points $f(t) = t$. The existence of fixed points guarantees the convergence of the MacKay’s algorithm for evidence maximization.

Make the fixed point iteration faster. Note that the fixed point iteration Equation 3 requires to explicitly compute $z = U^T y$ with $\mathcal{O}(n^2)$ storage and computation, which would be undesired if n is very large. To make it a practical algorithm, we take advantage of the fact that $\sigma_i = 0$ for $i > r$, and optimize the fixed point update as follows:

$$\begin{aligned} t' = f(t) &= \left(\frac{n}{n - \sum_{i=1}^D \frac{\sigma_i^2}{t + \sigma_i^2}} - 1 \right) t^2 \frac{\sum_{i=1}^n \frac{z_i^2}{(t + \sigma_i^2)^2}}{\sum_{i=1}^n \frac{\sigma_i^2 z_i^2}{(t + \sigma_i^2)^2}} \\ &= \left(\frac{n}{n - \sum_{i=1}^r \frac{\sigma_i^2}{t + \sigma_i^2}} - 1 \right) t^2 \frac{\sum_{i=1}^r \frac{z_i^2}{(t + \sigma_i^2)^2} + \frac{1}{t^2} \sum_{i=r+1}^n z_i^2}{\sum_{i=1}^r \frac{\sigma_i^2 z_i^2}{(t + \sigma_i^2)^2}} \\ &= \left(\frac{n}{n - \sum_{i=1}^r \frac{\sigma_i^2}{t + \sigma_i^2}} - 1 \right) t^2 \frac{\sum_{i=1}^r \frac{z_i^2}{(t + \sigma_i^2)^2} + \frac{1}{t^2} (\sum_{i=1}^n y_i^2 - \sum_{i=1}^r z_i^2)}{\sum_{i=1}^r \frac{\sigma_i^2 z_i^2}{(t + \sigma_i^2)^2}} \end{aligned}$$

Therefore, we can derive a faster algorithm (Equation 4) for the fixed point iteration, which only requires the first r entries of z without computing the full U matrix or the full z vector. This is the exact algorithm we implement in Algorithm 3.

$$t' = f(t) = \left(\frac{n}{n - \sum_{i=1}^r \frac{\sigma_i^2}{t + \sigma_i^2}} - 1 \right) t^2 \frac{\sum_{i=1}^r \frac{z_i^2}{(t + \sigma_i^2)^2} + \frac{1}{t^2} (\sum_{i=1}^n y_i^2 - \sum_{i=1}^r z_i^2)}{\sum_{i=1}^r \frac{\sigma_i^2 z_i^2}{(t + \sigma_i^2)^2}} \quad (4)$$

5.2 Influence of dimensionality

In Section 4.2.2, we normalize the LogME value by the number of examples because Equation 2 clearly scales with n . The influence of feature dimension D is, however, unclear. In this section, we prove that *LogME value will remain unchanged if the feature dimension goes up without introducing more information*. We study two cases: feature duplicate and feature padding, which shows the existence of infinitely many features with arbitrary feature dimension that share the same LogME value, therefore canceling the necessity of dimensionality normalization.

Theorem 3 (feature duplicate) LogME value will remain the same if the feature consists of arbitrary replicas of the original feature. Formally speaking, if the LogME value for $F \in \mathbb{R}^{n \times D}$ and $y \in \mathbb{R}^n$ is \mathcal{L} , then the LogME value for $\tilde{F} = [F, \dots, F] \in \mathbb{R}^{n \times qD}$ and $y \in \mathbb{R}^n$ is also \mathcal{L} . ($q \in \mathbb{N}$ is a natural number to represent the number of replicas.)

Theorem 4 (feature padding) LogME value will remain the same if the feature is padded with arbitrary number of zeros. Formally speaking, if the LogME value for $F \in \mathbb{R}^{n \times D}$ and $y \in \mathbb{R}^n$ is \mathcal{L} , then the LogME value for $\tilde{F} = [F, \mathbf{0}] \in \mathbb{R}^{n \times (D+d)}$ and $y \in \mathbb{R}^n$ is also \mathcal{L} . $d \in \mathbb{N}$ is a natural number and $\mathbf{0} \in \mathbb{R}^{n \times d}$ is a matrix with all zero entries.

The proofs of Theorem 3 and Theorem 4 are in Appendix D and Appendix E, respectively. The core idea is to find the SVD of \tilde{F} .

Theorem 3 and Theorem 4 imply that duplicating features or padding features with zeros will not influence LogME value. LogME is capable of aggregating redundant information or filtering out useless information in features, explaining its excellent empirical performance in You et al. (2021).

6. Tuning of pre-trained models

The paradigm we propose in Section 3 consists of two parts: ranking and tuning pre-trained models. So far we have introduced all the technical details in ranking pre-trained models, including a transferability metric named LogME in Section 4.2 and its theoretical analysis in Section 5. In this section we will focus on tuning of pre-trained models.

An additional problem that needs to be resolved before tuning pre-trained models is to select the PTMs to tune. For M pre-trained models, the possible number of subsets is 2^M , which is impractical to fully explore. To overcome the exponential complexity, the PTM ranking can come to rescue. With the PTM ranking, PTM selection can be greedy, namely *top-ranked PTMs are selected first*.

We identify two possible scenarios of tuning pre-trained models: single best PTM tuning and multiple PTMs tuning. (1) *Single best PTM tuning* fits perfectly if there is not any constraint on the network architecture, parameter count, or FLOPs of computation, which is common in academic research. Intuitively, the rest PTMs are inferior to the best ranked PTM so they would not help much. There are many dedicated methods (Chen et al., 2019; Kou et al., 2020; You et al., 2020) to fine-tune a single PTM, which is not the focus of this paper. (2) When we deploy neural networks in industrial applications, typically there are strict constraints on the available memory/computation budget. Therefore the pre-trained model ϕ_t satisfying these constraints is probably not the best ranked. In the past, practitioners could only fine-tune ϕ_t , and the knowledge of the PTM hub $\{\phi_i\}_{i=1}^M$ cannot be exploited. In this paper, we show that it is possible to transfer knowledge from several PTMs to the target pre-trained model ϕ_t during fine-tuning, which we call “multiple PTMs tuning”. This way, even if the target pre-trained model is specified but optimal, we can still improve it by top-performing PTMs in the hub. The later section focuses on the less studied scenario (multiple PTMs tuning).

In multiple PTMs tuning, the target pre-trained model ϕ_t is given, and the top- K ranked PTMs $\{\phi_k\}_{k=1}^K$ are the source of knowledge transfer. Typically $K < M$, *i.e.* not all PTMs are used, because (1) ensemble learning (Zhou et al., 2002) reveals that a subset of appropriate teachers can be better than all the teachers; (2) some PTMs may not be suitable for the target task and would hinder the transfer learning procedure. As for how to choose the hyper-parameter K ($1 \leq K \leq M$), we give empirical guidelines in Section 7.3.

Now suppose we have selected K pre-trained models $\{\phi_k\}_{k=1}^K$, with each pre-trained model ϕ_k transforming input x into a D_k dimensional feature. In general, pre-trained models $\{\phi_k\}_{k=1}^K$ have various network architectures, and dimensionality of features $\{D_k\}_{k=1}^K$ can vary, too. Let ϕ_t be the target architecture, which transforms the input into D_t dimensional feature. Formally, the multiple PTMs tuning problem is to fine-tune a pre-trained model ϕ_t by leveraging selected pre-trained models $\{\phi_k\}_{k=1}^K$, as shown in Figure 2.

To fine-tune a model ϕ_t in a target task, a new output head would be attached after ϕ_t , where target-specific loss is calculated. The target-specific head and loss are necessary

for every possible solution to multiple PTMs tuning, which is represented by a loss term L_{task} . We will not elaborate on L_{task} as it depends on the task. Next we review existing approaches to the problem in Section 6.1 and introduce our method in Section 6.2, with a focus on how to take advantage of selected pre-trained models $\{\phi_k\}_{k=1}^K$.

6.1 Existing approaches to the multiple PTMs tuning problem

A baseline approach is to just fine-tune ϕ_t without considering selected PTMs $\{\phi_k\}_{k=1}^K$. This can serve as a baseline to measure the improvement brought by multiple PTMs tuning.

A knowledge distillation approach to multiple PTMs tuning is knowledge distillation (Hinton et al., 2015) in the feature space via mean-square error. Since the feature dimensions between ϕ_t and $\{\phi_k\}_{k=1}^K$ may be different, a linear transformation module is necessary. The knowledge distillation (KD) method takes advantage of selected pre-trained models by adding a regularization term $L_{KD} = \frac{1}{n} \sum_{i=1}^n \frac{1}{K} \sum_{k=1}^K \|\phi_k(x_i) - W_k \phi_t(x_i)\|_2$, where W_k is learnable parameter to transform D_t dimensional feature $\phi_t(x_i)$ into D_k dimensional vector compatible with $\phi_k(x_i)$. Even if $D_k = D_t$, the semantic of each dimension in ϕ_t and ϕ_k may vary, making it necessary to introduce the transformation parameter W_k . The final loss would be $L_{task} + \lambda L_{KD}$, with λ a hyper-parameter. The KD method is another simple but general baseline in multiple PTM tuning. It can be applied to various PTMs but the performance improvement is limited.

Zoo-tuning for homogeneous PTMs tuning. In the special case when ϕ_t and $\{\phi_k\}_{k=1}^K$ all share the same network architecture, Zoo-tuning proposed by Shu et al. (2021) adaptively aggregates parameters of $\{\phi_k\}_{k=1}^K$ into ϕ_t in a layer-wise fashion, thereby modifying L_{task} to leverage pre-trained models. Zoo-tuning is the current state-of-the-art method for homogeneous PTMs tuning, but it fails to deal with the practical scenario when architectures of ϕ_t and $\{\phi_k\}_{k=1}^K$ are different.

6.2 A Bayesian approach to multiple PTMs tuning (B-Tuning)

We draw lessons from the shortcomings of the knowledge distillation approach and the Zoo-tuning approach mentioned above. Zoo-tuning operates at the level of parameters (layers), thereby limiting itself in the homogeneous case. Knowledge distillation operates at the level of output features, which works for heterogeneous PTMs but aligning features across PTMs is not easy. Taking the advantages from both sides, our approach should operate at the level of features to hide the heterogeneity among PTMs, and should go beyond features to avoid explicitly align features from various pre-trained models. Inspired by the ranking metric (LogME), we propose a B-Tuning approach with the help of posterior predictive distribution in Bayesian regression.

Posterior predictive distribution predicts the label y' of incoming feature f conditioned on all the available training features F and labels y , i.e. $p(y'|f, F, y)$ rather than $p(y'|f)$. In essence, the LogME algorithm (Algorithm 1) estimates α^*, β^* to maximize the model evidence. With pre-computed α^*, β^*, m (byproducts of the LogME algorithm), for a new incoming feature f , the posterior predictive distribution is $p(y'|f, F, y) = \int_w p(y'|w, f) p(w|F, y) dw$, where $p(y'|w, f) \sim \mathcal{N}(w^T f, \beta^{*-1})$ by definition, $p(w|F, y) = \frac{p(w)p(F, y|w)}{\int_{w'} p(w')p(F, y|w') dw'}$ by the Bayesian rule. With further computation we can derive that $p(w|F, y) \sim \mathcal{N}(\beta^* A^{-1} F^T y, A^{-1})$.

Plugging in the distributions of $p(y'|w, f)$ and $p(w|F, y)$, Rasmussen (2003) shows that $p(y'|f, F, y) \sim \mathcal{N}(f^T \beta^* A^{-1} F^T y, f^T A^{-1} f + \beta^{*-1})$, or $p(y'|f, F, y) \sim \mathcal{N}(f^T m, f^T A^{-1} f + \beta^{*-1})$ with $m = \beta^* A^{-1} F^T y$ to make the expression simpler. In short, for extracted features $F \in \mathbb{R}^{n \times D}$ and labels $y \in \mathbb{R}^n$, the LogME algorithm gives α^*, β^*, m , and the posterior predictive distribution is $p(y'|f, F, y) \sim \mathcal{N}(f^T m, f^T A^{-1} f + \beta^{*-1})$.

Posterior predictive distribution depends on the training data F, y and incoming feature f . Let F_k represents features extracted by the pre-trained model ϕ_k , $f_k = \phi_k(x)$ represents the output feature of the current data point extracted by the pre-trained model ϕ_k , then each pre-trained model can produce a posterior predictive distributions $p(y'_k|f_k, F_{\phi_k}, y) \sim \mathcal{N}(f_k^T m_k, f_k^T A_k^{-1} f_k + \beta_k^{*-1})$. Ensembling these distributions, we can get $\bar{y}' = \frac{1}{K} \sum_{k=1}^K y'_k$. Although $\{y'_k\}_{k=1}^K$ admit a simple Gaussian distribution, they are dependent because features F_k come from the same dataset. Therefore it is difficult to derive the exact distribution of \bar{y}' . Fortunately, according to the linearity property of expectation, $\mathbb{E}\bar{y}' = \frac{1}{K} \sum_{k=1}^K \mathbb{E}y'_k = \frac{1}{K} \sum_{k=1}^K f_k^T m_k$, the expectation of \bar{y}' is known.

Note that LogME can be calculated for any pre-trained model, so we can also derive the posterior predictive distribution for ϕ_t : $p(y'_t|f_t, F_t, y) \sim \mathcal{N}(f_t^T m_t, f_t^T A_t^{-1} f_t + \beta_t^{*-1})$. To leverage selected pre-trained models $\{\phi_k\}_{k=1}^K$, we can align the expectation of posterior predictive distributions from ϕ_t and average posterior predictive distributions of $\{\phi_k\}_{k=1}^K$ into a regularization term $L_{Bayesian} = \frac{1}{n} \sum_{i=1}^n \|\mathbb{E}\bar{y}' - \mathbb{E}y'_t\|_2$, where the expectation is taken over the predictive distributions. Extending to multiple classes, the final expression is

$$L_{Bayesian} = \frac{1}{n} \sum_{i=1}^n \frac{1}{C} \sum_{c=1}^C \left(\frac{1}{K} \sum_{k=1}^K f_{k,c}^T m_{k,c} - f_{t,c}^T m_{t,c} \right)^2, \quad (5)$$

where $m_{k,c}, m_{t,c}$ are calculated by the LogME algorithm and keep fixed during training. The final loss would be $L_{task} + \lambda L_{Bayesian}$, with λ a hyper-parameter to trade-off two loss terms. Because the method depends on the Bayesian approach of calculating posterior predictive distribution, we call it Bayesian Tuning, or **B-Tuning**. Figure 5 describes the method in detail. Note that only ϕ_t is updated during B-Tuning while the rest PTMs $\{\phi_k\}_{k=1}^K$ are fixed.

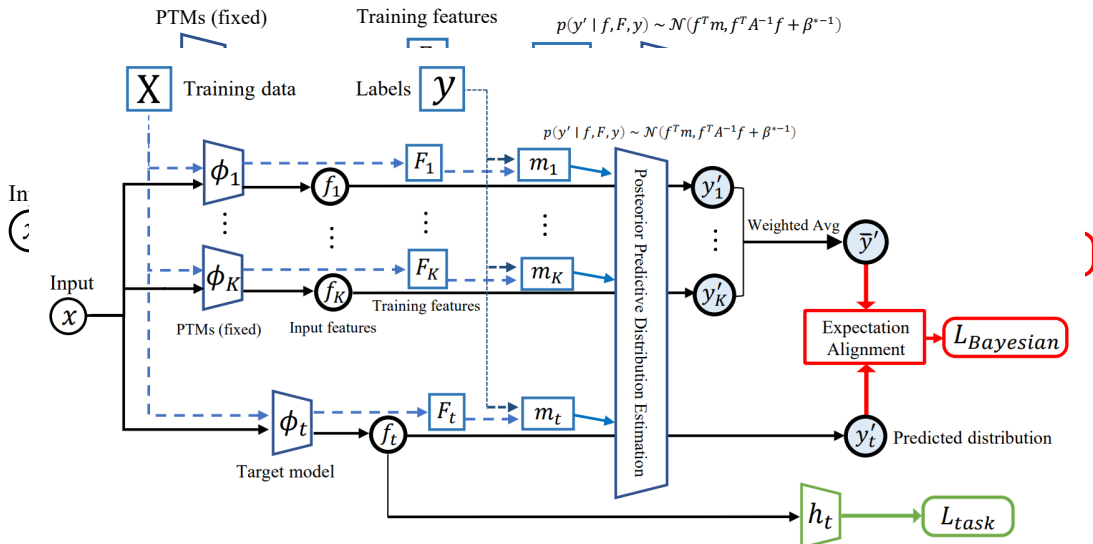


Figure 5: Illustration of B-Tuning. Dashed lines are pre-calculated before tuning.

B-Tuning has several significant advantages over previous methods. First, it hides the heterogeneity among PTMs by operating at the level of features, characterizing itself as a *general* solution to multiple PTMs tuning in both homogeneous and heterogeneous cases. Second, despite its heavy mathematical derivation, B-Tuning has a simple interpretation: it aligns features adaptively with m serving as an attention-like mechanism, canceling the necessity of learning to transform features into a common space like the knowledge distillation approach. The advantages of B-Tuning make it theoretically suitable for multiple PTMs tuning, which is empirically demonstrated in Section 7.3.

7. Experiments

We first present the behaviors of LogME on toy problems in Section 7.1. Experiments on ranking PTMs and tuning PTMs are located in Section 7.2 and Section 7.3 respectively, demonstrating the power of the proposed new paradigm. Section 7.4 quantitatively measures the efficiency of LogME and Section 7.5 compares LogME against a common approach of re-training head over a fixed feature extractor, providing a comprehensive understanding of LogME. Original data for some figures are available in the appendix. Code for LogME is available at <https://github.com/thuml/LogME> and full code will be available later.

7.1 Illustration with toy data

To give readers an intuitive sense of how LogME works, we generate features with increasing noise to mimic the features extracted by pre-trained models with decreasing transferability and to check if LogME can measure the quality of features.

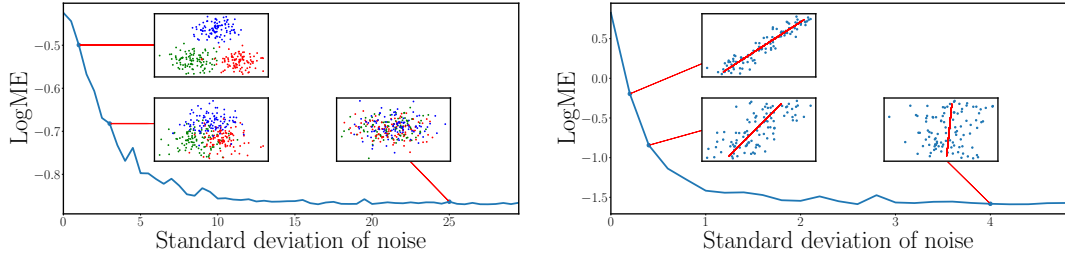


Figure 6: Toy data shows that LogME value goes down with decreasing feature quality.

For classification (Figure 6 left), three clusters in 2-D plane are generated, with colors representing the categories. Initially, the features are well separable so LogME has a large value. Then we add Gaussian noise with increasing variance to the data and the clustering structure in feature space disappears, leading to smaller LogME values as expected.

For regression (Figure 6 right), x is uniformly distributed ($x \sim \mathcal{U}[0, 1]$) and the output $y = 2x + \epsilon$ with observation error $\epsilon \sim \mathcal{N}(0, 0.1^2)$. By adding noise to the feature $x' = x + \mathcal{N}(0, t^2)$, the quality of feature x' becomes worse and it is harder to predict y from x' . With larger t (the standard deviation of noise), LogME becomes smaller as expected.

These toy experiments on synthesized data show that LogME is a good measure of the feature quality, and therefore can provide a good ranking for PTMs in a pre-trained model hub. The solid theoretical foundation behind LogME also brings an effective B-Tuning

method for multiple PTMs tuning, making LogME a core technique in the whole paradigm of ranking and tuning pre-trained models.

7.2 Ranking pre-trained models

Now we focus on the first part of the proposed paradigm: ranking pre-trained models. The goal is to rank pre-trained models so that potentially best PTMs can be selected for the subsequent tuning process. This section attaches great importance to the diversity of pre-trained model and the downstream task. Section 7.2.1 and Section 7.2.2 transfer supervised pre-trained models to classification and regression tasks, respectively; Section 7.2.3 explores unsupervised pre-trained models on both classification and regression; Section 7.2.4 and Section 7.2.5 study pre-trained language models on language understanding tasks and a sequential tagging task, respectively. These extensive experiments demonstrate the generality and effectiveness of the proposed LogME in ranking pre-trained models.

7.2.1 RANKING SUPERVISED PRE-TRAINED MODELS IN CLASSIFICATION TASKS

We use 12 ImageNet pre-trained models available from PyTorch: Inception V1 (Szegedy et al., 2015), Inception V3 (Szegedy et al., 2016), ResNet 34 (He et al., 2016), ResNet 50 (He et al., 2016), ResNet 101 (He et al., 2016), ResNet 152 (He et al., 2016), Wide ResNet 50 (Zagoruyko and Komodakis, 2017), DenseNet 121 (Huang et al., 2017), DenseNet 169 (Huang et al., 2017), DenseNet 201 (Huang et al., 2017), MobileNet V2 (Sandler et al., 2018), and NASNet-A Mobile (Tan et al., 2019). These pre-trained models cover most of the supervised pre-trained models in transfer learning that practitioners frequently use.

For downstream classification tasks, we take 9 commonly used datasets: Aircraft (Maji et al., 2013), Birdsnap (Berg et al., 2014), Caltech (Fei-Fei et al., 2004), Cars (Krause et al., 2013), CIFAR10 (Krizhevsky and Hinton, 2009), CIFAR100 (Krizhevsky and Hinton, 2009), DTD (Cimpoi et al., 2014), Pets (Parkhi et al., 2012), and SUN (Xiao et al., 2010). The description of each dataset and data statistics are listed in Appendix F.

For all the datasets we use, we respect the official train / val / test splits if they exist, otherwise we use 60% data for training, 20% data for validation (hyper-parameter setting to measure the ground-truth transfer learning performance) and 20% data for testing.

To compute the value of transfer performance $\{T_m\}_{m=1}^M$ ($M = 12$), we carefully fine-tune pre-trained models with grid-search of hyper-parameters. As pointed out by Li et al. (2020), learning rates and weight decays are the two most important hyper-parameters. Hence we grid search learning rates and weight decays (7 learning rates from 10^{-1} to 10^{-4} , 7 weight decays from 10^{-6} to 10^{-3} , all logarithmically spaced) to select the best hyper-parameter on the validation set and compute the accuracy on the test set. *It is noteworthy that LogME requires neither fine-tuning nor grid search.* Here we fine-tune pre-trained models to measure the ranking performance of LogME itself, but practitioners can straightforwardly use LogME to evaluate pre-trained models without fine-tuning.

We compare LogME against LEEP (Nguyen et al., 2020) and NCE (Tran et al., 2019). Prior to this paper, LEEP and NCE are the only two methods to rank PTMs without fine-tuning, and they can only be used to rank supervised pre-trained models in classification tasks. We use LEEP, NCE and LogME to compute scores $\{S_m\}_{m=1}^M$ by applying 12 pre-

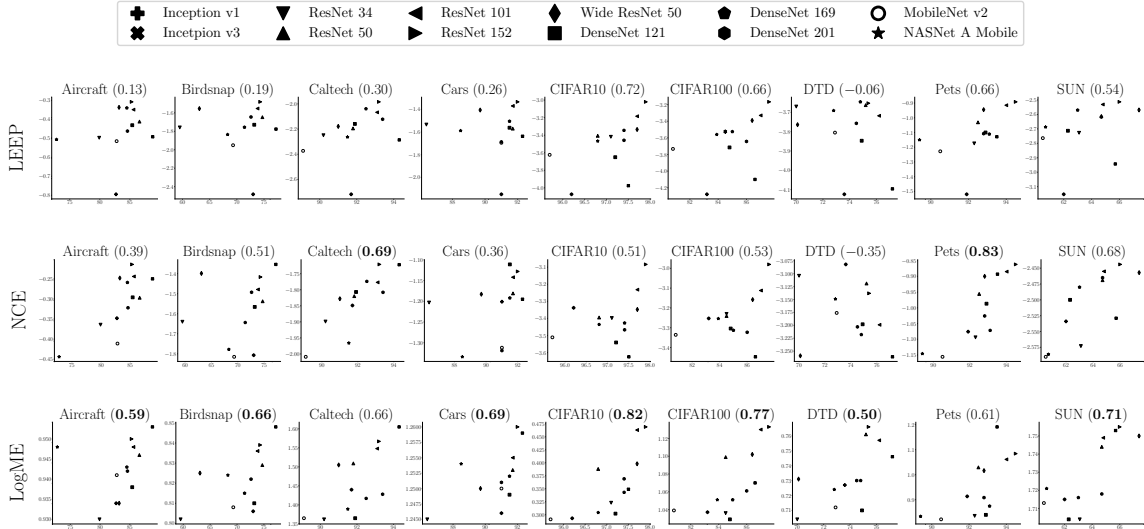


Figure 7: Correlation (τ_w) between fine-tuned accuracy (X-axis) and scores produced by three methods (Y-axis) for ranking PTMs on 9 datasets with 12 pre-trained models. One row for each method, one column for each dataset (with τ_w in the parenthesis near the dataset name), and one marker for each pre-trained model. **The best τ_w in each dataset is marked in bold.**

trained models to the datasets. The correlation τ_w between scores and fine-tuned accuracies are presented in Figure 7.

We can find that LogME has consistently better correlation than LEEP, and outperforms NCE on most datasets (7 datasets out of 9 datasets). Note that LEEP and NCE even show a negative correlation in DTD (Cimpoi et al., 2014), because they rely on the relationship between classes of the pre-trained task and the target task while DTD classes (textures) are very different from ImageNet categories (objects). In contrast, LogME still performs reasonably well for DTD.

According to the definition and interpretation of τ_w in Section 4.1, τ_w rank correlation can be translated into $\frac{\tau_w + 1}{2}$ probability of correct comparison. The smallest τ_w of LogME in Figure 7 is around 0.5, so the probability of a pre-trained model ϕ_A transferring better than ϕ_B is at least 75% if ϕ_A has a larger LogME. For most tasks τ_w of LogME is 0.7 or 0.8, so the probability of correct selection is 85% or 90%, sufficient for practical usage.

7.2.2 RANKING SUPERVISED PRE-TRAINED MODELS IN A REGRESSION TASK

Besides extensive classification tasks considered above, this section shows how LogME can be used to assess pre-trained models for a regression task. The two prior methods (LEEP and NCE) depends on the category relationship between pre-trained categories and downstream categories, therefore they are not applicable to regression tasks.

The regression task we use is the dSprites (Matthey et al., 2017) dataset from the Visual Task Adaptation Benchmark (Zhai et al., 2020) which is commonly used for evaluating the quality of learned representations. The input is an image containing a sprite (heart, square,

and ellipse) with varying scale, orientation, and position. Pre-trained models are transferred to predict four scalars (scale, orientation, and (x, y) positions) together, and mean square error (MSE) on the test data is reported. The supervised pre-trained models are the same as Section 7.2.1 and hyper-parameter setting scheme follows.

Results are plotted in Figure 8. It is clear that LogME and MSE are well correlated and the correlation coefficient $\tau_w = 0.79$ is very large: if a pre-trained model ϕ_A has larger LogME than ϕ_B , with 89.5% probability ϕ_A is better (has smaller MSE) than ϕ_B after actually fine-tuning.

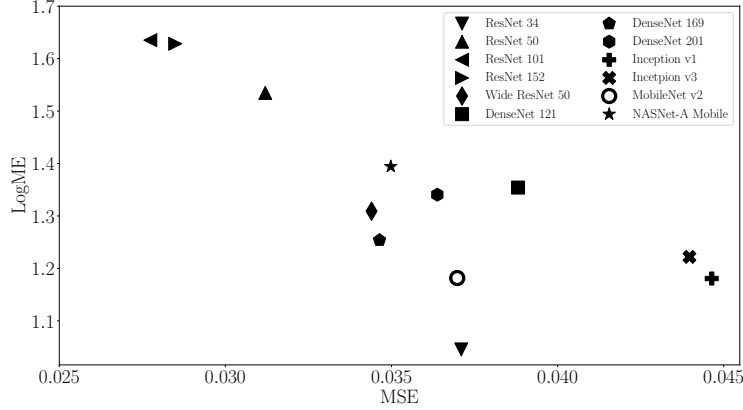


Figure 8: Supervised pre-trained models transferred to dSprites.

7.2.3 RANKING CONTRASTIVE PRE-TRAINED MODELS IN DOWNSTREAM TASKS

The recently emerging unsupervised pre-trained models (He et al., 2020) have attracted much attention due to the potential of utilizing humongous unlabeled data in the Internet. They use the contrastive loss (Gutmann and Hyvärinen, 2010) to inject supervision signal into pre-training with unlabeled data, and they feature a projection head with continuous output. Practitioners may be interested in ranking contrastive pre-trained models, but LEEP and NCE cannot be extended to deal with the projection head of contrastive-based unsupervised pre-trained models, either.

Since LogME only requires features extracted from pre-trained models, it can be applied to contrastive pre-trained models. To demonstrate this, we use four popular models pre-trained with various training schemes: MoCo V1 (He et al., 2020) with momentum contrast, MoCo V2 (Chen et al., 2020b) with an MLP projection head and strong data augmentation, MoCo 800 trained with 800 epochs as suggested by Chen et al. (2020a), and SimCLR (Chen et al., 2020a) with carefully designed implementation.

Aircraft (Maji et al., 2013), the first dataset (alphabetically) in Section 7.2.1 is used as the classification task, and dSprites (Matthey et al., 2017) is used as the regression task. Results are shown in Table 3. SimCLR on dSprites is not reported because it does not converge after several trials. LogME gives the *perfect order* of both transferred accuracy and MSE. Note that the ground-truth order on transfer learning performance in Aircraft (MoCo V1 < MoCo V2 < MoCo 800) is different from the order in dSprites (MoCo V1 < MoCo 800 < MoCo V2), emphasizing that ranking pre-trained models is *task adaptive*.

We also observe that LogME values of unsupervised pre-trained models are similar (the difference is smaller than their supervised counterparts in Section 7.2.1), mainly because unsupervised features are not very discriminative.

Table 3: Use LogME to rank unsupervised pre-trained models.

Pre-trained Network	Aircraft		dSprites	
	Accuracy (%)	LogME	MSE	LogME
MoCo V1	81.68	0.934	0.069	1.52
MoCo V2	84.16	0.941	0.047	1.64
MoCo 800	86.99	0.946	0.050	1.58
SimCLR	88.10	0.950	-	-
$\tau_w: 1.0$		$\tau_w: 1.0$		

7.2.4 RANKING PRE-TRAINED LANGUAGE MODELS IN THE GLUE BENCHMARK

To further demonstrate the generality of LogME, we show how LogME can work for pre-trained language models. Again, existing works (LEEP and NCE) cannot deal with these pre-trained language models.

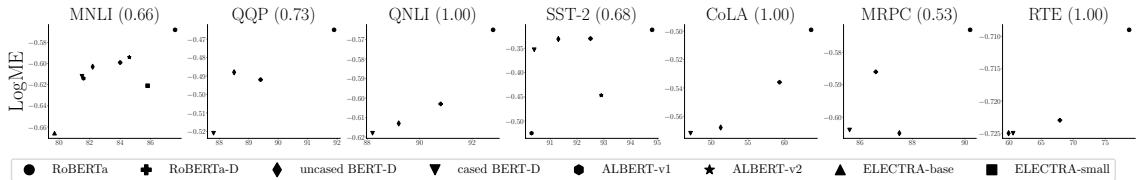


Figure 9: Correlation (τ_w) between fine-tuned accuracy (X-axis) and LogME value (Y-axis) in 7 GLUE tasks with 8 popular pre-trained language models. One sub-figure for each task (with its τ_w in the parenthesis), and one marker for each PTM.

Here we take an alternative approach of evaluating the transfer performance $\{T_m\}_{m=1}^M$. We do not fine-tune pre-trained models ourselves, but directly use accuracies tuned by others, and check if LogME can correlate well with the results. The HuggingFace [Model Hub](#) generously provides lots of pre-trained language models and even provides carefully tuned transfer learning results in some GLUE (Wang et al., 2018) tasks for some models. We take pre-trained models that have GLUE performance tuned by the HuggingFace organization, and select the top 8 downloaded models: RoBERTa (Liu et al., 2019), RoBERTa-D, uncased BERT-D, cased BERT-D, ALBERT-v1 (Lan et al., 2020), ALBERT-v2 (Lan et al., 2020), ELECTRA-base (Clark et al., 2020), and ELECTRA-small (Clark et al., 2020) (“D” means distilled version). The LogME on seven GLUE classification tasks together with fine-tuned accuracy are plotted in Figure 9. Some models only have results for certain tasks and we keep them as they are. Even though these accuracy numbers are tuned by the HuggingFace organization, LogME perfectly estimates the ranking of transfer performance for 3 tasks (with $\tau_w = 1$), showing the surprising effectiveness of LogME in ranking pre-trained models.

7.2.5 RANKING PRE-TRAINED LANGUAGE MODELS IN A SEQUENTIAL TAGGING TASK

So far we only considered simple classification and regression tasks. It would be valuable to extend LogME to deal with structural output tasks such as object detection and semantic segmentation. Next we show how LogME can be used in a special sequential tagging task where both the input and the output is structural. How to deal with a general task with structural output is left as future work.

The specific task we consider in this section is named entity recognition (Sang and De Meulder, 2003). It requires the model to predict the entity label (person, location, organization, etc.) of every token in a sentence, therefore the output is structural. What’s more, pre-trained language models would produce an embedding for each token, so the input is also structural and each output matches an input token. Actually the named entity recognition task is sometimes referred to as “token-level classification”. To apply LogME to the task, we can simply flatten the token dimension so that n represents the number of tokens rather than the number of sentences.

We use the same PTMs as in Section 7.2.4, and the dataset is CoNLL-2003 (Sang and De Meulder, 2003) whose performance is measured by F-1 score. Table 4 holds the results. The rank correlation τ_w is 0.20, smaller than results in previous sections. The small τ_w is caused by an outlier RoBERTa (Liu et al., 2019), which has the largest F-1 score with a relatively small LogME value. Without RoBERTa, the rank correlation τ_w is 0.56, on par with previous sections. It is noteworthy that if we select the best PTM according to the LogME value, ALBERT-v1 will be used and its performance is the best. From this perspective, LogME is decently useful. In general, how to deal with structural tasks better would be an interesting problem for future research.

Table 4: Ranking pre-trained models in named entity recognition (CoNLL-2003 task).

PTM	RoBERTa	RoBERTa-D	uncased BERT-D	cased BERT-D	ALBERT-v1	ALBERT-v2	ELECTRA-base	ELECTRA-small	τ_w
F-1 score (%)	97.4	96.6	96.8	95.5	97.0	97.4	97.2	91.9	
LogME	0.685	0.723	0.783	0.623	0.834	0.809	0.746	0.646	0.20

7.3 Tuning pre-trained models

Now we turn to the second part of the proposed paradigm: tuning pre-trained models. As mentioned earlier in Section 6, most academic researchers care little about the practical constraints of deployed models, therefore they can use the best ranked (according to the LogME value) PTM straightforward. This paper is concerned about the practical usage scenario, where computational constraints require us to use a specific PTM but we still want to leverage the knowledge from other PTMs in the pre-trained model hub.

Experiments in this section are designed to compare three methods of tuning multiple PTMs: the knowledge distillation approach, the Zoo-tuning approach and the proposed B-Tuning. We first conduct experiments with multiple homogeneous PTMs where three methods are applicable, then dive into the practical case of multiple heterogeneous PTMs.

7.3.1 TUNING MULTIPLE HOMOGENEOUS PTMs

To compare the proposed B-Tuning against existing methods (knowledge distillation and Zoo-tuning), we follow the experimental setup of Zoo-tuning (Shu et al., 2021) to use five homogeneous pre-trained models. They are ResNet-50 models trained by different pre-training tasks: (1) Supervised pre-trained on ImageNet (He et al., 2016); (2) Unsupervised pre-trained by MoCo (He et al., 2020); (3) MaskRCNN model (He et al., 2017); (4) DeepLab V3 (Chen et al., 2017); (5) KeyPoint detection model pre-trained on COCO (Lin et al., 2014). These five pre-trained models make up the PTM hub. The dataset we use is Aircraft (Maji et al., 2013), the first dataset (alphabetically) in Section 7.2.1.

Following the experimental setup of Shu et al. (2021), all the five PTMs are used as the teacher models, and three methods on multiple PTM tuning are reported in the first row of Table 5. Zoo-tuning performs better than vanilla knowledge distillation, and the proposed B-Tuning even surpasses Zoo-tuning, setting a new state-of-the-art result for multiple PTMs tuning.

Table 5: Accuracy (%) of multiple PTMs tuning in Aircraft, with different teacher models and tuning methods. As a baseline, single PTM fine-tune achieves 82.99% accuracy.

method teacher models	Knowledge Distillation	Zoo-tuning	B-Tuning
all PTMs from the PTM hub	83.02	83.26	83.44
top-3 PTMs (ranked by LogME)	84.55	-	84.70

Shu et al. (2021) used **all** five PTMs to tune the target pre-trained model, **because they do not know how to select PTMs**. An important contribution of this paper is to rank PTMs by a transferability metric (LogME) so that potentially best models can be selected. Therefore we use LogME to rank the five PTMs, selecting the top-3 PTMs as the teacher models in the subsequent tuning. The results are in the second row of Table 5. **Surprisingly, selecting top-3 PTMs brings a significant performance improvement, demonstrating the effectiveness of our *new paradigm of “ranking and tuning pre-trained models”*.**

To demonstrate the effectiveness of multiple PTMs transfer, we conduct the Fine-tune with Knowledge Distillation experiments. That is, we fine-tune ResNet-50 supervisedly on Aircraft dataset, in addition to vanilla distillation loss in the feature space with 3 teachers picked up from the homogeneous PTMs zoo, which assists the model to learn abundant knowledge from different vision domains while fine-tuned supervisedly. To evaluate the effectiveness of LogME in multiple PTMs selection, we compare the top-3 models selected by LogME as teachers against trying out all teacher models combination from the zoo, the top-1 test accuracies are presented in Figure 10. As the baseline, we also show the result of single ResNet-50 obtained by the Pre-training → Fine-tuning paradigm.

To quantitatively measure the benefit of selecting top-3 PTMs ranked by LogME, we present an in-depth study in Figure 10. There are $10 = \binom{5}{3}$ combinations in selecting 3 PTMs out of 5, and we tried all of the combinations. To avoid hidden confounders, we use the **vanilla Knowledge Distillation** approach to exploit multiple PTMs. We observe two conclusions from Figure 10: (1) **transferring the knowledge from multiple PTMs consistently outperforms fine-tuning a single pre-trained model (82.99%)**, which adheres to our intuition

that utilizing the rich knowledge from various PTMs is better than fine-tuning alone. (2) the best combination achieves 84.64% accuracy, but usually it is too expensive to try all the combinations (10 tuning runs). Instead, we can use LogME to select the top-3 PTMs and then perform tuning just once, which achieves a pretty good accuracy of 84.55%.

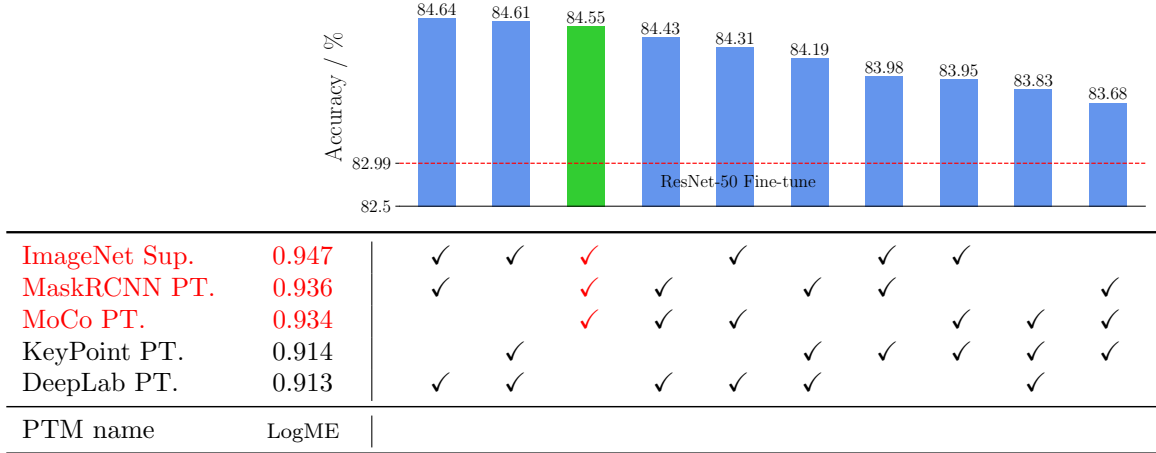


Figure 10: Accuracy of knowledge distillation with 3 PTM teachers. We tried all of the $\binom{5}{3} = 10$ combinations of selecting 3 PTM teachers. Selecting top-3 PTMs according to LogME achieves the third best performance among 10 combinations.

In conclusion, this section demonstrates three facts: (1) multiple PTMs tuning is better than single PTM fine-tuning; (2) selecting top-ranked PTMs according to LogME is better than using all the PTMs, because “many could be better than all” (Zhou et al., 2002); (3) B-Tuning is superior to knowledge distillation and Zoo-tuning.

7.3.2 TUNING MULTIPLE HETEROGENEOUS PTMS

A more general and attractive application of multiple PTMs tuning is transferring knowledge from a large PTMs hub with heterogeneous PTMs. The problem can be decomposed into three parts: (1) rank all the PTMs; (2) select a proper number of teacher PTMs; (3) exploit a tuning method to tune the target PTM by the selected PTMs. The proposed LogME in this paper can help to rank all the PTMs, and the proposed B-Tuning can deal with the multiple PTMs tuning problem. So this section tries to provide some guidelines on how to select a proper number of teacher PTMs.

To avoid confounders, we still study the problem in the Aircraft dataset. The PTM hub consists of the 12 PTMs used in Section 7.2.1. The 12 PTMs are ranked by their LogME values, and the target PTM ϕ_t is the most common ResNet-50. Top- K PTMs are used in B-Tuning to fine-tune the target model, with K varying from 1 to 12. Results are plotted in Figure 11, where the X-axis is the value of K .

We have three observations with respect to Figure 11: (1) tuning with multiple PTMs is consistently better than single PTM fine-tuning (accuracy 82.99%), so it is worthwhile to explore advanced tuning techniques like B-Tuning. (2) tuning with all of the 12 PTMs

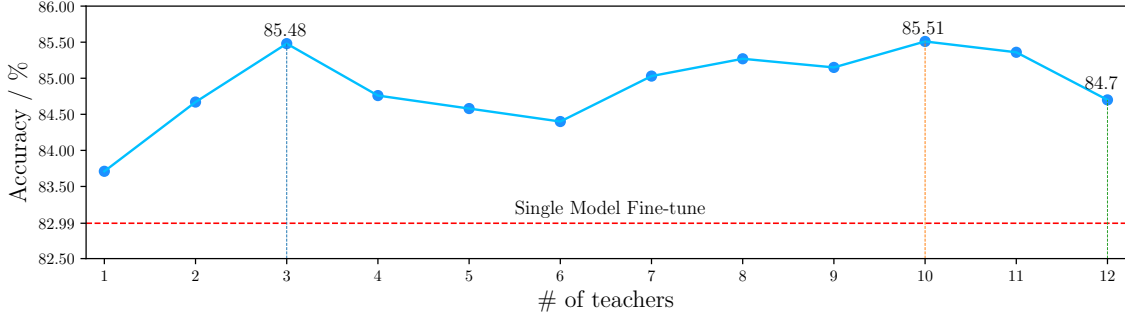


Figure 11: A study on the number of PTMs to use in multiple PTMs tuning.

(accuracy 84.7%) is inferior to tuning with top 10 PTMs (accuracy 85.51%), which again confirms the “many could be better than all” principle (Zhou et al., 2002) and emphasizes the importance of selecting proper PTMs. (3) tuning with top 3 PTMs achieves almost the same accuracy as tuning with top 10 PTMs, but the latter incurs a much larger computational cost, since at least a forward pass of each PTM is required to be exploited.

Given the above evidence, we recommend that $K = 3$ in practice to balance the computational cost and the performance improvement during tuning pre-trained models. The new paradigm thus becomes: (1) rank all PTMs by LogME; (2) select top-3 PTMs; (3) tune the target PTM with the top-3 PTMs using B-Tuning.

7.4 Efficiency of LogME

A theoretically sound algorithm is usually complex and hard to compute, as is the case for LogME without computational optimization. Fortunately, we successfully reduce the computational complexity when analyzing its theoretical convergence property in Section 5.1 by the fixed point iteration. The algorithmic complexity is presented in Table 2, and Table 6 presents the wall-clock time speedup measured in Aircraft with ResNet-50. The naïve implementation is very slow. Our conference paper (You et al., 2021) proposed an optimization scheme for matrix multiplication and matrix inversion, which brings $61.7\times$ speedup. This paper further proposes the fixed point iteration algorithm, which results in a much larger speedup ($131.5\times$). Thanks to the significant optimization, LogME is equipped with both theoretical soundness and fast computation.

Table 6: Quantitative measurement of computational speedup in evidence maximization.

	Wall-clock time (second)	Speedup
evidence maximization (naïve implementation)	802.5 ± 5.6	-
evidence maximization (optimized by You et al. (2021))	13.1 ± 0.7	$61.7\times$
evidence maximization (fixed point iteration, proposed)	6.1 ± 0.7	$131.5\times$

Next we quantitatively measure the wall-clock and memory footprint of LogME in both computer vision and natural language processing. Results are shown in Table 7. ResNet 50

on Aircraft is used for computer vision, and RoBERTa-D on MNLI task is used for NLP. The cost of computing ground-truth transferability T_m (fine-tuning with hyper-parameter search) serves as the upper bound of ranking pre-trained models. We also list the cost of extracting features by pre-trained models as a reference, which is the lower bound of ranking pre-trained models. The cost for the rest models and datasets vary, but the proportion is similar. Note that, because carelessly tuned hyper-parameters cannot tell good models from bad models, it is necessary to attribute the cost of hyper-parameter search to brute-force fine-tuning while LogME does not need hyper-parameter setting.

Table 7: Computational cost and memory footprint of LogME.

	wall-clock time		memory footprint	
Computer Vision	fine-tune (upper bound)	161000s	fine-tune (upper bound)	6.3 GB
	extract feature (lower bound)	37s	extract feature (lower bound)	43 MB
	LogME	43s	LogME	53 MB
	benefit	3700 \uparrow	benefit	120 \uparrow
Natural Language Processing	fine-tune (upper bound)	100200s	fine-tune (upper bound)	88 GB
	extract feature (lower bound)	1130s	extract feature (lower bound)	1.2 GB
	LogME	1136s	LogME	1.2 GB
	benefit	88 \uparrow	benefit	73 \uparrow

It is clear that **brute-force fine-tuning is computationally expensive**, requiring about a day for one dataset with one pre-trained model. Selecting the best pre-trained model out of 12 models would cost 12 days. Extracting features is very cheap and costs much less. Note that the cost of LogME in Table 7 contains the cost of extracting features, and we can see that the additional time-cost of LogME compared to feature extraction is rather small, which means that *LogME’s cost is very close to the lower bound* (the cost of feature extraction). In computer vision, the wall-clock time of LogME is reduced by more than 3700 \times speedup with 120 \times less memory footprint. In the NLP domain, feature extraction is much slower and therefore the wall-clock time speedup is not as striking as computer vision, but the speedup still reaches two orders of magnitude.

In summary, LogME is efficient in terms of both wall-clock time and memory footprint, thanks to the optimized algorithm (fixed point iteration) brought by the theoretical analysis.

7.5 Comparing LogME to re-training head

A naïve way to measure the relationship between features and labels is to train a linear classification / regression head for the downstream task, and to use the head’s performance as an assessment (sometimes it is called “*linear probing*” or “*linear protocol evaluation*”). *This straightforward idea does not work well because it is prone to over-fitting.* To alleviate over-fitting, grid search for its hyper-parameters (such as the strength of L2 regularization) should be tuned extensively on a validation set, making head re-training inefficient and inappropriate if data are limited. In the following section, we summarize why re-training head is inferior to LogME from three perspectives, which partially explains why the important problem of ranking and tuning PTMs was under-explored in the past.

(1) LogME is more efficient than re-training head. In the Caltech dataset, we extract features from 12 pre-trained models, train softmax regressors with tuned hyper-

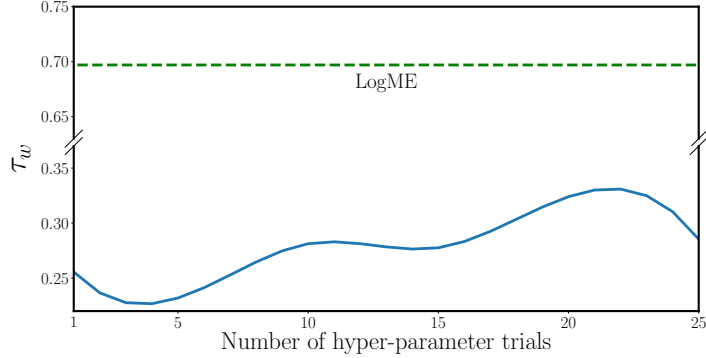


Figure 12: The correlation of re-training head *w.r.t.* the number of hyper-parameter trials. It is clear that re-training head is much worse than LogME.

parameters (the L2 regularization strength), and plot the correlation between the best head accuracy and the transfer performance *w.r.t.* the number of hyper-parameter trials in Figure 12. The correlation of LogME is plotted as a reference. Computing LogME requires $3\times$ less time than re-training a head with one fixed hyper-parameter, and re-training head with **exhaustive hyper-parameter search** is still much inferior to LogME. Therefore re-training head is not appropriate in ranking pre-trained models.

(2) Re-training head does not work well with limited training data. Because re-training head follows a supervised learning paradigm, it suffers in low-shot learning scenarios. For example, prompt learning (Liu et al., 2021) is an active research area in natural language processing, where researchers try to discover the capacity of fixed pre-trained models by only a few training data. Take sentiment classification task SST-2 (Socher et al., 2013) for example, prompt learning (Liu et al., 2021) extracts the sentence embedding E_S for each sentence S , and compare E_S with word embeddings E_P, E_N of a positive anchor word P (*i.e.* “good”, “fantastic”) and a negative anchor word N (*i.e.* “bad”, “awful”). The decision rule is: sentence S contains positive sentiment $\iff E_S^T E_P > E_S^T E_N$. In this case, searching proper anchor words on validation data yields 61.4% accuracy (complete results are available in Table 12). Training a simple classification head on limited training data upon sentence embedding E_S and tuning the weight-decay hyper-parameter on validation set is also a doable idea, but it does not work well empirically because of the limited training data. With 10 sentences for training, re-training head only achieves 51.64% accuracy. Meanwhile, we can apply the LogME approach proposed in this paper (or to be specific, the posterior predictive distribution introduced in Section 6.2), which does not require any hyper-parameter setting. This way, we can combine training data with validation data to compute the predictive weight m for each class, and use m as the embedding of a virtual “anchor word”, which results in 79.24% accuracy, *a huge improvement over re-training head and manually selected anchor words!* The superior performance of LogME is interpretable: we analyzed the predictive weight m for the negative sentiment class, and find that it is closest to the embedding of “dump”, “: (”, “doomed”, “worse”, “worse”. To our surprise,

it can discover that “:(” contains negative sentiment, as “:(” is a cyber word to express unhappy emotion.

(3) Re-training head does not have a clear metric. As a side issue, even if we re-train a head for the downstream task, it is unclear what quantity of the head should be used to measure pre-trained models. Since the performance of downstream tasks is evaluated by accuracy and MSE in transfer learning, it may somewhat cause over-fitting if we use the accuracy and MSE of the re-trained head. Indeed, in Figure 12, when the number of hyper-parameter trials increases, the correlation can even go down, showing the effect of somewhat over-fitting. In contrast, LogME has a unified modeling of label evidence, which has a clear statistical underpinning.

8. Conclusion

Pre-trained models are universally acknowledged as the foundation of modern deep learning, but the research is rather focused on how to exploit a given pre-trained model. Since pre-trained models are gathered into PTM hubs, we study how to exploit the PTM hub as a whole, coming up with a new paradigm of ranking and tuning pre-trained models. The ranking part mainly includes a theoretically sound and computationally efficient transferability metric named LogME. The theoretical investigation of LogME solved the decades-long unanswered question about the convergence of the MacKay’s algorithm. The transferability metric LogME is then further extended to be a multiple PTMs tuning method named B-Tuning, which consists of the tuning part of the paradigm.

The rich experiments confirm the effectiveness of the proposed methods in ranking (LogME *v.s.* brute-force fine-tuning, LEEP and NCE), selection (PTMs with top-3 LogME *v.s.* exponentially many combinations), and tuning (B-Tuning *v.s.* Zoo-tuning and Knowledge Distillation), making the new paradigm of exploiting PTM hubs attractive for our community.

Acknowledgments

We would like to dedicate our thanks to Ximei Wang, Xinyang Chen, Yang Shu at Tsinghua University, Yi Zeng at Peking University, and Yonglong Tian at MIT for helpful discussions. This work was supported in part by the National Natural Science Foundation of China under Grants 62022050 and 62021002, in part by the Beijing Nova Program under Grant Z201100006820041, in part by the MOE Innovation Plan and the BNRist Innovation Fund.

Appendix A. Notation Table

Notations used in this paper are listed in the following table. We tried our best to avoid notation conflicts, but 2 notation conflicts are unavoidable due to the heavy math in this paper: (1) f_i represents features of x_i and $f(t)$ represents the fixed point iteration function. (2) w represents parameters in the linear head and τ_w uses w as its subscript.

Table 8: Notations used in this paper.

notation	dimensionality	meaning
i, j, k	\mathbb{N}	running subscripts
M, n	\mathbb{N}	the number of PTMs and samples
K	\mathbb{N}	the number of selected PTMs for subsequent tuning
C, D	\mathbb{N}	the dimension of label and extracted feature
ϕ	—	a pre-trained model
x_i	—	an input sample
$f_i = \phi(x_i)$	\mathbb{R}^D	extracted feature of an input example
$F = [f_1, \dots, f_n]^T$	$\mathbb{R}^{n \times D}$	stacked features of f_i
Y_i	\mathbb{R}^C	label of x_i
y_i	\mathbb{R}	a component of Y_i
y	\mathbb{R}^n	the label component for all n samples
y'	—	predictive distribution of an input sample
$\bar{y}' = \frac{1}{K} \sum_{k=1}^K y'_k$	—	average of y'
T	\mathbb{R}	ground-truth transfer performance
S	\mathbb{R}	score produced by a transferability metric
τ	\mathbb{R}	Kendall's rank correlation
τ_w	\mathbb{R}	weighted Kendall's rank correlation
w	\mathbb{R}^D	parameter in linear head
α/β	\mathbb{R}	hyper-parameter of the Bayesian linear model
$A = \alpha I_D + \beta F^T F$	$\mathbb{R}^{D \times D}$	a quantity in calculating LogME
$m = \beta A^{-1} F^T y$	\mathbb{R}^D	a quantity in calculating LogME
γ	\mathbb{R}	a quantity in calculating LogME
$\mathcal{L} = \mathcal{L}(\alpha, \beta)$	\mathbb{R}	log evidence given α, β
α^*, β^*	\mathbb{R}	α, β to achieve maximum evidence
U, Σ, V	—	matrices in SVD ($F = U \Sigma V^T$)
σ	\mathbb{R}	diagonal entries in Σ
r	\mathbb{N}	the rank of matrix F
$z = U^T y$	\mathbb{R}^n	transformation of y under U
$t = \frac{\alpha}{\beta}$	\mathbb{R}	a quantity in convergence proof
t', α', β'	\mathbb{R}	the value of t, α, β after an iteration
$f(t)$	—	the fixed point iteration function
$\tilde{\mathcal{L}}, \tilde{F}$	—	\mathcal{L}, F for duplicated or padded features
W	—	transformation matrix in distillation

Appendix B. Proof of Theorem 1

Theorem 1: Algorithm 4 induces a scalar function $t' = f(t) = \left(\frac{n}{n - \sum_{i=1}^D \frac{\sigma_i^2}{t + \sigma_i^2}} - 1 \right) t^2 \frac{\sum_{i=1}^n \frac{z_i^2}{(t + \sigma_i^2)^2}}{\sum_{i=1}^n \frac{\sigma_i^2 z_i^2}{(t + \sigma_i^2)^2}}$.

Proof Let's express all symbols in a unified form with respect to $\alpha, \beta, \Sigma, z, U, V$:

- $A = \alpha I + \beta F^T F = V(\alpha I + \beta \Sigma^T \Sigma) V^T$
- $A^{-1} = V \Sigma_{inv} V^T$ where $(\Sigma_{inv})_{ii} = \frac{1}{\alpha + \beta \sigma_i^2}$ ($1 \leq i \leq D$)
- $m = \beta A^{-1} F^T y = \beta V \Sigma_{inv} \Sigma^T z$
- $m^T m = z^T \Sigma_m z$ with $\Sigma_m = \beta^2 \Sigma \Sigma_{inv}^2 \Sigma^T$ and $(\Sigma_m)_{ii} = \frac{\beta^2 \sigma_i^2}{(\alpha + \beta \sigma_i^2)^2}$, so $m^T m = \sum_{i=1}^n \frac{\beta^2 \sigma_i^2 z_i^2}{(\alpha + \beta \sigma_i^2)^2}$
- $Fm = \beta U \Sigma \Sigma_{inv} \Sigma^T z$, $Fm - y = U \Sigma_{res} z$ with $\Sigma_{res} = \beta \Sigma \Sigma_{inv} \Sigma^T - I$, $(\Sigma_{res})_{ii} = -\frac{\alpha}{\alpha + \beta \sigma_i^2}$
- $\|Fm - y\|_2^2 = (Fm - y)^T (Fm - y) = z^T (\Sigma_{res})^2 z = \sum_{i=1}^n \frac{\alpha^2 z_i^2}{(\alpha + \beta \sigma_i^2)^2}$
- $\gamma = \sum_{i=1}^D \frac{\beta \sigma_i^2}{\alpha + \beta \sigma_i^2} = \sum_{i=1}^D \frac{\sigma_i^2}{t + \sigma_i^2}$

Putting them together, we have

$$t' = \frac{\alpha'}{\beta'} = \frac{\gamma}{n - \gamma} \frac{\|Fm - y\|_2^2}{m^T m} = \left(\frac{n}{n - \sum_{i=1}^D \frac{\sigma_i^2}{t + \sigma_i^2}} - 1 \right) t^2 \frac{\sum_{i=1}^n \frac{z_i^2}{(t + \sigma_i^2)^2}}{\sum_{i=1}^n \frac{\sigma_i^2 z_i^2}{(t + \sigma_i^2)^2}} = f(t)$$

■

Appendix C. Proof of Theorem 2

Theorem 2: If $r < n$ and $\sum_{1 \leq i, j \leq n} (z_i^2 - z_j^2)(\sigma_i^2 - \sigma_j^2) > 0$, then $f(t)$ has a fixed point and thus the MacKay's algorithm will converge.

Proof The theorem can be proved by studying the behavior of $f(t)$ near 0 and ∞ .

We have $\lim_{t \rightarrow 0} f(t) = \frac{r}{n-r} \frac{\sum_{i=r+1}^n z_i^2}{\sum_{i=1}^r z_i^2} > 0$, which is a constant and positive number.

When t approaches infinity, we find that $\lim_{t \rightarrow \infty} \frac{f(t)}{t} = \frac{\sum_{i=1}^n \sigma_i^2}{n} \frac{\sum_{i=1}^n z_i^2}{\sum_{i=1}^n \sigma_i^2 z_i^2}$ is constant, which means $f(t)$ behaves linearly when t is large enough.

Noticing a trick used in proving the Chebyshev's Sum Inequality (Hardy et al., 1952), we can get $\sum_{1 \leq i, j \leq n} (z_i^2 - z_j^2)(\sigma_i^2 - \sigma_j^2) = 2n \sum_{i=1}^n \sigma_i^2 z_i^2 - 2(\sum_{i=1}^n \sigma_i^2)(\sum_{i=1}^n z_i^2)$. The condition $\sum_{1 \leq i, j \leq n} (z_i^2 - z_j^2)(\sigma_i^2 - \sigma_j^2) > 0$ thus translates into $\frac{\sum_{i=1}^n \sigma_i^2}{n} \frac{\sum_{i=1}^n z_i^2}{\sum_{i=1}^n \sigma_i^2 z_i^2} < 1$, which means $f(t)$ increases linearly with a slope smaller than 1 (i.e. $\lim_{t \rightarrow \infty} \frac{f(t)}{t} = \frac{\sum_{i=1}^n \sigma_i^2}{n} \frac{\sum_{i=1}^n z_i^2}{\sum_{i=1}^n \sigma_i^2 z_i^2} < 1$).

In summary, When t approaches 0, it is assured that $\lim_{t \rightarrow 0} f(t) > t = 0$; when t is large enough, it is assured that $f(t) < t$. **Putting these two conditions together, they imply the existence of a fixed point $t_0 > 0$ such that $f(t_0) = t_0$.** ■

Appendix D. Proof of Theorem 3

Theorem 3: LogME value will remain the same if the feature consists of arbitrary replicas of the original feature. Formally speaking, if the LogME value for $F \in \mathbb{R}^{n \times D}$ and $y \in \mathbb{R}^n$ is \mathcal{L} , then the LogME value for $\tilde{F} = [F, \dots, F] \in \mathbb{R}^{n \times qD}$ and $y \in \mathbb{R}^n$ is also \mathcal{L} . ($q \in \mathbb{N}$ is a natural number to represent the number of replicas.)

Proof Since LogME is calculated via an iterative algorithm, we prove the theorem by an iterative invariant (a quantitative relation that holds after every while-loop iteration).

Preliminary: SVD of \tilde{F} . We have already known the SVD of F is $F = U\Sigma V^T$, and σ_i is the i -th largest eigenvalue of FF^T . Since $\tilde{F}\tilde{F}^T = qFF^T$, duplicated feature \tilde{F} has singular values $\tilde{\sigma}_i^2 = \begin{cases} q\sigma_i^2 & 1 \leq i \leq D \\ 0 & D+1 \leq i \leq qD \end{cases}$, and its left orthogonal matrix is the

same as F : $\tilde{U} = U$. The right orthogonal matrix of \tilde{F} is somewhat complicated. Let's find an orthogonal matrix $Q_{q \times q}$, whose entries in the first column are $\frac{1}{\sqrt{q}}$. Entries in the other columns do not matter, as long as $Q_{q \times q}$ is a valid orthogonal matrix. For example, we can

use $Q_{2 \times 2} = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$, and $Q_{3 \times 3} = \begin{bmatrix} \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & \frac{2}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \end{bmatrix}$. Then the right orthogonal

matrix of \tilde{F} is $\tilde{V} = Q_{q \times q} \otimes V$, where \otimes is the Kronecker product of two matrices. Using the

block matrix form of Kronecker product, we can write down \tilde{V} as $\tilde{V} = \begin{bmatrix} \frac{1}{\sqrt{p}}V & \dots & \dots \\ \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{p}}V & \dots & \dots \end{bmatrix} \in$

$\mathbb{R}^{qD \times qD}$, with the first D columns of \tilde{V} corresponding to singular values $\sqrt{q}\sigma_i, 1 \leq i \leq D$,

and the other $(q-1) \times D$ columns of \tilde{V} are orthogonal basis with respect to singular values $\sigma_i = 0$. In summary, if the SVD of F is $F = U\Sigma V^T$, then the SVD of $\tilde{F} = [F, \dots, F]$ is

$$\tilde{F} = \tilde{U}\tilde{\Sigma}\tilde{V}^T, \text{ where } \tilde{U} = U, \tilde{\Sigma} = [\sqrt{q}\Sigma, 0, \dots, 0], \tilde{V} = \begin{bmatrix} \frac{1}{\sqrt{q}}V & \dots & \dots \\ \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{q}}V & \dots & \dots \end{bmatrix} = Q_{q \times q} \otimes V.$$

Iterative invariant: if we apply Algorithm 2 to both \tilde{F} and F , with a small change that we initialize $\tilde{\alpha} = q, \tilde{\beta} = 1$, then $\tilde{\alpha} = q\alpha, \tilde{\beta} = \beta$ holds before Line 5. Suppose $\tilde{\alpha} = q\alpha, \tilde{\beta} = \beta$ holds before a while-loop, then we can have:

$$\begin{aligned} \tilde{\gamma} &= \sum_{i=1}^{qD} \frac{\tilde{\beta}\tilde{\sigma}_i^2}{\tilde{\alpha} + \tilde{\beta}\tilde{\sigma}_i^2} = \sum_{i=1}^D \frac{q\beta\sigma_i^2}{q\alpha + q\beta\sigma_i^2} = \sum_{i=1}^D \frac{\beta\sigma_i^2}{\alpha + \beta\sigma_i^2} = \gamma \\ \tilde{\Lambda} &= \text{diag} \left\{ \tilde{\alpha} + \tilde{\beta}\tilde{\sigma}_i^2 \right\}, \tilde{\alpha} + \tilde{\beta}\tilde{\sigma}_i^2 = \begin{cases} q(\alpha + \beta\sigma_i^2) & 1 \leq i \leq D \\ q\alpha & D+1 \leq i \leq qD \end{cases} \end{aligned}$$

$$\begin{aligned}
 \tilde{m} &= \tilde{\beta} \tilde{A}^{-1} \tilde{F}^T y = \beta \tilde{V} \tilde{\Lambda}^{-1} \tilde{V}^T \tilde{V} \tilde{\Sigma}^T \tilde{U}^T y = \beta \tilde{V} \tilde{\Lambda}^{-1} \tilde{\Sigma}^T U^T y \\
 &= \beta \left(\begin{bmatrix} \frac{1}{\sqrt{q}} V & \cdots & \cdots \\ \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{q}} V & \cdots & \cdots \end{bmatrix} \begin{bmatrix} \frac{1}{q} \Lambda^{-1} & \\ & \frac{1}{q\alpha} I_{(q-1) \times D} \end{bmatrix} \begin{bmatrix} \sqrt{q} \Sigma^T \\ 0 \\ \cdots \\ 0 \end{bmatrix} \right) U^T y \\
 &= \begin{bmatrix} \frac{1}{q} V \Lambda^{-1} U^T y \\ \cdots \\ \frac{1}{q} V \Lambda^{-1} U^T y \end{bmatrix} = \begin{bmatrix} \frac{1}{q} m \\ \cdots \\ \frac{1}{q} m \end{bmatrix}
 \end{aligned}$$

$$\text{Therefore } \tilde{m}^T \tilde{m} = \frac{1}{q} m^T m, \tilde{F} \tilde{m} = [F, \dots, F] \begin{bmatrix} \frac{1}{q} m \\ \cdots \\ \frac{1}{q} m \end{bmatrix} = Fm.$$

After the while-loop iteration, $\tilde{\alpha}' = \frac{\tilde{\gamma}}{\tilde{m}^T \tilde{m}} = \frac{\gamma}{\frac{1}{q} m^T m} = q\alpha'$, $\tilde{\beta}' = \frac{n-\tilde{\gamma}}{\|\tilde{F}\tilde{m}-y\|_2^2} = \frac{n-\gamma}{\|Fm-y\|_2^2} = \beta'$, then the iterative invariant $\tilde{\alpha} = q\alpha, \tilde{\beta} = \beta$ still holds. Therefore we know that when the algorithm converges, $\tilde{\alpha}^* = q\alpha^*, \tilde{\beta}^* = \beta^*$. The corresponding maximum evidence is

$$\begin{aligned}
 \tilde{\mathcal{L}} &= \frac{n}{2} \log \tilde{\beta}^* + \frac{qD}{2} \log \tilde{\alpha}^* - \frac{n}{2} \log 2\pi - \frac{\tilde{\beta}^*}{2} \|\tilde{F}\tilde{m} - y\|_2^2 - \frac{\tilde{\alpha}^*}{2} \tilde{m}^T \tilde{m} - \frac{1}{2} \log |\tilde{A}^*| \\
 &= \frac{n}{2} \log \beta^* + \frac{qD}{2} \log(q\alpha^*) - \frac{n}{2} \log 2\pi - \frac{\beta^*}{2} \|Fm - y\|_2^2 - \frac{\alpha^*}{2} m^T m - \frac{1}{2} \log |\tilde{\Lambda}^*| \\
 &= \frac{n}{2} \log \beta^* + \frac{qD}{2} \log(q\alpha^*) - \frac{n}{2} \log 2\pi - \frac{\beta^*}{2} \|Fm - y\|_2^2 - \frac{\alpha^*}{2} m^T m \\
 &\quad - \frac{1}{2} \log |\Lambda^*| - \frac{1}{2} \log \left(q^D (q\alpha^*)^{(q-1)D} \right) \\
 &= \mathcal{L} - \frac{D}{2} \log \alpha^* + \frac{qD}{2} \log(q\alpha^*) - \frac{1}{2} \log \left(q^D (q\alpha^*)^{(q-1)D} \right) \\
 &= \mathcal{L}
 \end{aligned}$$

By the convergence analysis in Section 5.1, initialization of α, β only changes the initial value of t , which does not impact the convergence value of the fixed point iteration. Therefore, we can conclude that duplicating features will not change the value of LogME.

Although the above proof targets at Algorithm 2, it is straightforward to adapt the proof to Algorithm 3. ■

Appendix E. Proof of Theorem 4

Theorem 4: LogME value will remain the same if the feature is padded with arbitrary number of zeros. Formally speaking, if the LogME value for $F \in \mathbb{R}^{n \times D}$ and $y \in \mathbb{R}^n$ is \mathcal{L} , then the LogME value for $\tilde{F} = [F, \mathbf{0}] \in \mathbb{R}^{n \times (D+d)}$ and $y \in \mathbb{R}^n$ is also \mathcal{L} . $d \in \mathbb{N}$ is a natural number and $\mathbf{0} \in \mathbb{R}^{n \times d}$ is a matrix with all zero entries.

Proof The proof follows the same idea as Theorem 3, but the SVD of \tilde{F} is simpler than Theorem 3. If the SVD of F is $F = U\Sigma V^T$, then the SVD of $\tilde{F} = [F, \mathbf{0}]$ is $\tilde{F} = \tilde{U}\tilde{\Sigma}\tilde{V}^T$, where $\tilde{U} = U, \tilde{\Sigma} = [\Sigma, \mathbf{0}], \tilde{V} = \begin{bmatrix} V \\ W \end{bmatrix}$, with $W \in \mathbb{R}^{d \times d}$ an orthogonal matrix that satisfies

$$W^T W = I_d. \text{ Note that } \tilde{\Sigma} = [\Sigma, \mathbf{0}] \text{ translates into } \tilde{\sigma}_i^2 = \begin{cases} \sigma_i^2 & 1 \leq i \leq D \\ 0 & D+1 \leq i \leq D+d \end{cases}.$$

Iterative invariant: if we apply Algorithm 2 to both \tilde{F} and F , with the same initialization $\tilde{\alpha} = 1, \tilde{\beta} = 1$, then $\tilde{\alpha} = \alpha, \tilde{\beta} = \beta$ holds before Line 5. Suppose $\tilde{\alpha} = \alpha, \tilde{\beta} = \beta$ holds before a while-loop, then we can have:

$$\begin{aligned} \tilde{\gamma} &= \sum_{i=1}^{D+d} \frac{\tilde{\beta}\tilde{\sigma}_i^2}{\tilde{\alpha} + \tilde{\beta}\tilde{\sigma}_i^2} = \sum_{i=1}^D \frac{\beta\sigma_i^2}{\alpha + \beta\sigma_i^2} = \gamma \\ \tilde{\Lambda} &= \text{diag} \left\{ \tilde{\alpha} + \tilde{\beta}\tilde{\sigma}_i^2 \right\}, \tilde{\alpha} + \tilde{\beta}\tilde{\sigma}_i^2 = \begin{cases} \alpha + \beta\sigma_i^2 & 1 \leq i \leq D \\ \alpha & D+1 \leq i \leq D+d \end{cases} \end{aligned}$$

$$\begin{aligned} \tilde{m} &= \tilde{\beta}\tilde{A}^{-1}\tilde{F}^T y = \beta \begin{bmatrix} V \\ W \end{bmatrix} \begin{bmatrix} \Lambda^{-1} & \\ & \frac{1}{\alpha} I_d \end{bmatrix} \begin{bmatrix} V^T \\ W^T \end{bmatrix} \begin{bmatrix} F^T \\ \mathbf{0}_{n \times d}^T \end{bmatrix} y = \begin{bmatrix} m \\ \mathbf{0}_{d \times 1} \end{bmatrix} \\ \tilde{m}^T \tilde{m} &= m^T m, \tilde{F} \tilde{m} = [F, \mathbf{0}_{n \times d}] \begin{bmatrix} m \\ \mathbf{0}_{d \times 1} \end{bmatrix} = Fm \end{aligned}$$

After the while-loop iteration, $\tilde{\alpha}' = \frac{\tilde{\gamma}}{\tilde{m}^T \tilde{m}} = \frac{\gamma}{m^T m} = \alpha', \tilde{\beta}' = \frac{n - \tilde{\gamma}}{\|\tilde{F} \tilde{m} - y\|_2^2} = \frac{n - \gamma}{\|Fm - y\|_2^2} = \beta'$, then the iterative invariant $\tilde{\alpha} = \alpha, \tilde{\beta} = \beta$ still holds. Therefore we know that when the algorithm converges, $\tilde{\alpha}^* = \alpha^*, \tilde{\beta}^* = \beta^*$. The corresponding maximum evidence is

$$\begin{aligned} \tilde{\mathcal{L}} &= \frac{n}{2} \log \tilde{\beta}^* + \frac{D+d}{2} \log \tilde{\alpha}^* - \frac{n}{2} \log 2\pi - \frac{\tilde{\beta}^*}{2} \|\tilde{F} \tilde{m} - y\|_2^2 - \frac{\tilde{\alpha}^*}{2} \tilde{m}^T \tilde{m} - \frac{1}{2} \log |\tilde{A}^*| \\ &= \frac{n}{2} \log \beta^* + \frac{D+d}{2} \log \alpha^* - \frac{n}{2} \log 2\pi - \frac{\beta^*}{2} \|Fm - y\|_2^2 - \frac{\alpha^*}{2} m^T m - \frac{1}{2} \log |\tilde{\Lambda}^*| \\ &= \frac{n}{2} \log \beta^* + \frac{D+d}{2} \log \alpha^* - \frac{n}{2} \log 2\pi - \frac{\beta^*}{2} \|Fm - y\|_2^2 - \frac{\alpha^*}{2} m^T m \\ &\quad - \frac{1}{2} \log |\Lambda^*| - \frac{1}{2} \log (\alpha^*)^d \\ &= \mathcal{L} + \frac{d}{2} \log \alpha^* - \frac{d}{2} \log \alpha^* \\ &= \mathcal{L} \end{aligned}$$

■

Appendix F. Detailed descriptions of the datasets

Aircraft: The dataset contains fine-grained classification of 10,000 aircraft pictures which belongs to 100 classes, with 100 images per class.

Birdsnap: The dataset contains 49,829 images of 500 species of North American birds.

Caltech: The dataset contains 9,144 pictures of objects belonging to 101 categories. There are about 40 to 800 images per category. Most categories have about 50 images.

Cars: The dataset contains 16,185 images of 196 classes of cars. The data is split into 8,144 training images and 8,041 testing images.

CIFAR 10: The dataset consists of 60,000 32x32 colorful images in 10 classes, with 6,000 images per class. There are 50,000 training images and 10,000 test images.

CIFAR 100: The dataset is just like the CIFAR 10, except it has 100 classes containing 600 images each.

DTD: The dataset contains a collection of 5,640 textural images in the wild, annotated with a series of human-centric attributes. It has 47 classes and 120 images per class.

Pets: The dataset contains 7,049 images of cat and dog species which belongs to 47 classes, with around 200 images per class.

SUN: The dataset contains 39,700 scenery pictures with 397 classes and 100 samples per class.

Appendix G. Original Results in Figures

Original results in figures are shown in the Table 9, Table 10, and Table 11.

Table 9: Original results in Figure 4.

	task	ResNet-34	ResNet-50	ResNet-101	ResNet-152	WideResNet-50	DenseNet-121	DenseNet-169	DenseNet-201	Inception v1	Inception v3	MobileNet v2	NASNet-A Mobile	τ_w
Aircraft	Accuracy	79.9	86.6	85.6	85.3	83.2	85.4	84.5	84.6	82.7	88.8	82.8	72.8	-
	LEEP	-0.497	-0.412	-0.349	-0.308	-0.337	-0.431	-0.340	-0.462	-0.795	-0.492	-0.515	-0.506	0.13
	NCE	-0.364	-0.297	-0.244	-0.214	-0.248	-0.296	-0.259	-0.322	-0.348	-0.250	-0.411	-0.444	0.39
	LogME	0.930	0.946	0.948	0.950	0.934	0.938	0.943	0.942	0.934	0.953	0.941	0.948	0.59
		59.5	74.7	73.8	74.3	63.1	73.2	71.4	72.6	73.0	77.2	69.3	68.3	-
Birdsnap	Accuracy	59.5	74.7	73.8	74.3	63.1	73.2	71.4	72.6	73.0	77.2	69.3	68.3	-
	LEEP	-1.758	-1.647	-1.553	-1.481	-1.554	-1.729	-1.756	-1.645	-2.483	-1.776	-1.951	-1.835	0.19
	NCE	-1.640	-1.538	-1.479	-1.417	-1.399	-1.566	-1.644	-1.493	-1.807	-1.354	-1.815	-1.778	0.51
	LogME	0.802	0.829	0.836	0.839	0.825	0.810	0.815	0.822	0.806	0.848	0.808	0.824	0.66
		90.2	91.8	93.1	93.2	91.0	91.9	92.5	93.4	91.7	94.3	89.1	91.5	-
Caltech	Accuracy	90.2	91.8	93.1	93.2	91.0	91.9	92.5	93.4	91.7	94.3	89.1	91.5	-
	LEEP	-2.249	-2.195	-2.067	-1.984	-2.179	-2.159	-2.039	-2.122	-2.718	-2.286	-2.373	-2.263	0.30
	NCE	-1.899	-1.820	-1.777	-1.721	-1.828	-1.807	-1.774	-1.808	-1.849	-1.722	-2.009	-1.966	0.69
	LogME	1.362	1.509	1.548	1.567	1.505	1.365	1.417	1.428	1.440	1.605	1.365	1.389	0.66
		86.4	91.7	91.7	92.0	89.7	91.5	91.5	91.0	91.0	92.3	91.0	88.5	-
Cars	Accuracy	86.4	91.7	91.7	92.0	89.7	91.5	91.5	91.0	91.0	92.3	91.0	88.5	-
	LEEP	-1.534	-1.570	-1.370	-1.334	-1.406	-1.562	-1.505	-1.687	-2.149	-1.637	-1.695	-1.588	0.26
	NCE	-1.203	-1.181	-1.142	-1.128	-1.183	-1.111	-1.192	-1.319	-1.201	-1.195	-1.312	-1.334	0.36
	LogME	1.245	1.253	1.255	1.260	1.250	1.249	1.252	1.251	1.246	1.259	1.250	1.254	0.69
		97.1	96.8	97.7	97.9	97.7	97.2	97.4	97.4	96.2	97.5	95.7	96.8	-
CIFAR10	Accuracy	97.1	96.8	97.7	97.9	97.7	97.2	97.4	97.4	96.2	97.5	95.7	96.8	-
	LEEP	-3.418	-3.407	-3.184	-3.020	-3.335	-3.651	-3.345	-3.458	-4.074	-3.976	-3.624	-3.467	0.72
	NCE	-3.398	-3.395	-3.232	-3.084	-3.348	-3.541	-3.427	-3.467	-3.338	-3.625	-3.511	-3.436	0.51
	LogME	0.323	0.388	0.463	0.469	0.398	0.302	0.343	0.369	0.293	0.349	0.291	0.304	0.82
		84.5	84.5	87.0	87.6	86.4	84.8	85.0	86.0	83.2	86.6	80.8	83.9	-
CIFAR100	Accuracy	84.5	84.5	87.0	87.6	86.4	84.8	85.0	86.0	83.2	86.6	80.8	83.9	-
	LEEP	-3.531	-3.520	-3.330	-3.167	-3.391	-3.715	-3.525	-3.643	-4.279	-4.100	-3.733	-3.560	0.66
	NCE	-3.230	-3.241	-3.112	-2.980	-3.158	-3.304	-3.313	-3.323	-3.253	-3.447	-3.336	-3.254	0.53
	LogME	1.036	1.099	1.130	1.133	1.102	1.029	1.051	1.061	1.037	1.070	1.039	1.051	0.77
		70.0	75.2	76.2	75.4	70.1	74.9	74.8	74.5	73.6	77.2	72.9	72.8	-
DTD	Accuracy	70.0	75.2	76.2	75.4	70.1	74.9	74.8	74.5	73.6	77.2	72.9	72.8	-
	LEEP	-3.670	-3.663	-3.718	-3.653	-3.764	-3.847	-3.646	-3.757	-4.124	-4.096	-3.805	-3.691	-0.06
	NCE	-3.104	-3.119	-3.199	-3.138	-3.259	-3.198	-3.218	-3.203	-3.082	-3.261	-3.176	-3.149	-0.35
	LogME	0.704	0.761	0.757	0.766	0.731	0.710	0.730	0.730	0.727	0.746	0.712	0.724	0.50
		92.3	92.5	94.0	94.5	92.8	92.9	93.1	92.8	91.9	93.5	90.5	89.4	-
Pets	Accuracy	92.3	92.5	94.0	94.5	92.8	92.9	93.1	92.8	91.9	93.5	90.5	89.4	-
	LEEP	-1.174	-1.031	-0.915	-0.892	-0.945	-1.100	-1.111	-1.108	-1.520	-1.129	-1.228	-1.150	0.66
	NCE	-1.094	-0.956	-0.885	-0.862	-0.900	-0.987	-1.072	-1.026	-1.076	-0.893	-1.156	-1.146	0.83
	LogME	0.835	1.029	1.061	1.084	1.016	0.839	0.874	0.908	0.913	1.191	0.821	0.833	0.61
		63.1	64.7	64.8	66.0	67.4	62.3	63.0	64.7	62.0	65.7	60.5	60.7	-
SUN	Accuracy	63.1	64.7	64.8	66.0	67.4	62.3	63.0	64.7	62.0	65.7	60.5	60.7	-
	LEEP	-2.727	-2.611	-2.531	-2.513	-2.569	-2.713	-2.570	-2.618	-3.153	-2.943	-2.764	-2.687	0.54
	NCE	-2.573	-2.469	-2.455	-2.444	-2.457	-2.500	-2.480	-2.465	-2.534	-2.529	-2.590	-2.586	0.68
	LogME	1.704	1.744	1.749	1.755	1.750	1.704	1.716	1.718	1.715	1.753	1.713	1.721	0.71

Table 10: Original results in Figure 5.

task	ResNet-34	ResNet-50	ResNet-101	ResNet-152	WideResNet-50	DenseNet-121	DenseNet-169	DenseNet-201	Inception v1	Inception v3	MobileNet v2	NASNet-A Mobile	τ_w	
dSprites	MSE	0.037	0.031	0.028	0.028	0.034	0.039	0.035	0.036	0.045	0.044	0.037	0.035	-
	LogME	1.05	1.53	1.64	1.63	1.31	1.35	1.25	1.34	1.18	1.22	1.18	1.39	0.79

Table 11: Original results in Figure 6.

task		RoBERTa	RoBERTa-D	uncased BERT-D	cased BERT-D	ALBERT-v1	ALBERT-v2	ELECTRA-base	ELECTRA-small	τ_w
MNLI	Accuracy	87.6	84.0	82.2	81.5	81.6	84.6	79.7	85.8	-
	LogME	-0.568	-0.599	-0.603	-0.612	-0.614	-0.594	-0.666	-0.621	0.66
QQP	Accuracy	91.9	89.4	88.5	87.8	-	-	-	-	-
	LogME	-0.465	-0.492	-0.488	-0.521	-	-	-	-	0.73
QNLI	Accuracy	92.8	90.8	89.2	88.2	-	-	-	-	-
	LogME	-0.565	-0.603	-0.613	-0.618	-	-	-	-	1.00
SST-2	Accuracy	94.8	92.5	91.3	90.4	90.3	92.9	-	-	-
	LogME	-0.312	-0.330	-0.331	-0.353	-0.525	-0.447	-	-	0.68
CoLA	Accuracy	63.6	59.3	51.3	47.2	-	-	-	-	-
	LogME	-0.499	-0.536	-0.568	-0.572	-	-	-	-	1.00
MRPC	Accuracy	90.2	86.6	87.5	85.6	-	-	-	-	-
	LogME	-0.573	-0.586	-0.605	-0.604	-	-	-	-	0.53
RTE	Accuracy	78.7	67.9	59.9	60.6	-	-	-	-	-
	LogME	-0.709	-0.723	-0.725	-0.725	-	-	-	-	1.00

Appendix H. Full Figure in Convergence Analysis

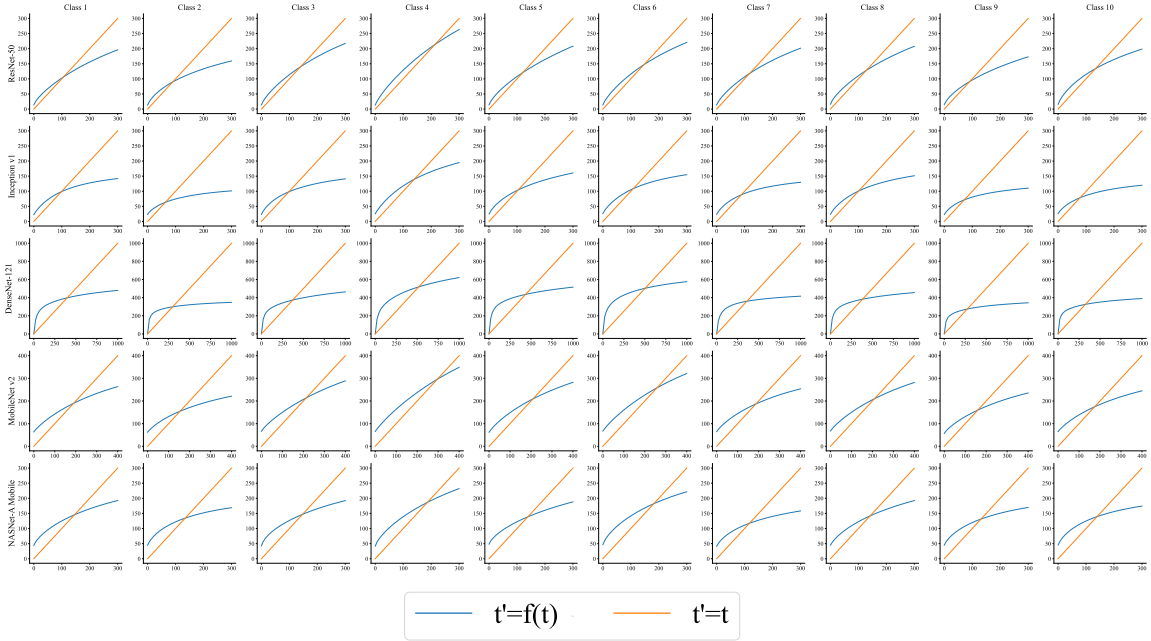


Figure 13: Fixed points of $f(t)$ in Equation 3 for all 10 classes in CIFAR10 with 5 pre-trained models. We plot $t' = f(t)$ (in blue) and $t' = t$ (in orange), whose intersections are fixed points. The existence of fixed points guarantees the convergence of the evidence maximization procedure in LogME.

Appendix I. Complete results in prompt learning

Table 12: Complete results in prompt learning with manually selected anchor words.

accuracy anchor P \ anchor N	negative	bad	ill	evil	poor
positive	49.8	52.8	49.1	49.1	60.9
good	51.0	50.9	49.0	52.4	50.9
fine	51.7	51.0	49.1	54.5	50.9
great	55.4	53.1	49.1	61.4	51.0
nice	51.6	50.6	49.1	51.0	50.8

References

- S. Ben-David and R. Schuller. Exploiting task relatedness for multiple task learning. In *COLT*, 2003.
- T. Berg, J. Liu, S. Woo Lee, M. L. Alexander, D. W. Jacobs, and P. N. Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *CVPR*, 2014.
- C. M. Bishop. *Neural networks for pattern recognition*. 1995.
- C. M. Bishop. *Pattern recognition and machine learning*. 2006.
- R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Kohd, M. Krass, R. Krishna, R. Kudritipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang. On the **Opportunities and Risks of Foundation Models**. *arXiv:2108.07258 [cs]*, 2021.
- T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language Models are Few-Shot Learners. In *NeurIPS*, 2020.
- L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv:1706.05587 [cs]*, 2017.
- T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*, 2020a.
- X. Chen, S. Wang, B. Fu, M. Long, and J. Wang. Catastrophic Forgetting Meets Negative Transfer: Batch Spectral Shrinkage for Safe Transfer Learning. In *NeurIPS*, 2019.
- X. Chen, H. Fan, R. Girshick, and K. He. Improved Baselines with Momentum Contrastive Learning. *arXiv:2003.04297 [cs]*, 2020b.
- M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

- K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *ICLR*, 2020.
- T. M. Cover. *Elements of information theory*. 1999.
- J. Daunizeau. Semi-analytical approximations to statistical moments of sigmoid and softmax mappings of normal variables. *arXiv preprint arXiv:1703.00091*, 2017.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1977.
- J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.
- J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 2019.
- J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014.
- D. Erhan, A. Courville, Y. Bengio, and P. Vincent. Why does unsupervised pre-training help deep learning? In *AISTATS*, 2010.
- R. Fagin, R. Kumar, and D. Sivakumar. Comparing top k lists. In *SODA*, 2003.
- L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR Workshops*, 2004.
- Y. Ganin and V. Lempitsky. Unsupervised Domain Adaptation by Backpropagation. In *ICML*, 2015.
- R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.
- S. F. Gull. Developments in maximum entropy data analysis. In *Maximum entropy and Bayesian methods*. 1989.
- M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010.
- X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, L. Zhang, W. Han, M. Huang, Q. Jin, Y. Lan, Y. Liu, Z. Liu, Z. Lu, X. Qiu, R. Song, J. Tang, J.-R. Wen, J. Yuan, W. X. Zhao, and J. Zhu. Pre-Trained Models: Past, Present and Future. *arXiv:2106.07139 [cs]*, 2021.
- G. H. Hardy, J. E. Littlewood, G. Pólya, and G. Pólya. *Inequalities*. 1952.
- K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015.

- K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *ICCV*, 2017.
- K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.
- G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network, 2015.
- W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande, and J. Leskovec. Strategies for Pre-training Graph Neural Networks. In *ICLR*, 2020.
- G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. In *CVPR*, 2017.
- G. Huang, Z. Liu, G. Pleiss, L. Van Der Maaten, and K. Weinberger. Convolutional networks with dense connectivity. *TPAMI*, 2019.
- L. Jing and Y. Tian. Self-supervised visual feature learning with deep neural networks: A survey. *TPAMI*, 2020.
- N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, and A. Borchers. In-datacenter performance analysis of a tensor processing unit. In *ISCA*, 2017.
- M. G. Kendall. A new measure of rank correlation. *Biometrika*, 1938.
- K. H. Knuth, M. Habeck, N. K. Malakar, A. M. Mubeen, and B. Placek. Bayesian Evidence and Model Selection. *Digital Signal Processing*, 2015.
- D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. 2009.
- S. Kornblith, J. Shlens, and Q. V. Le. Do better imagenet models transfer better? In *CVPR*, 2019.
- Z. Kou, K. You, M. Long, and J. Wang. Stochastic Normalization. In *NeurIPS*, 2020.
- J. Krause, J. Deng, M. Stark, and L. Fei-Fei. Collecting a large-scale dataset of fine-grained cars. 2013.
- A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, 2009.
- Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *ICLR*, 2020.
- C. Li, Y. Mao, R. Zhang, and J. Huai. On hyper-parameter estimation in empirical Bayes: a revisit of the MacKay algorithm. In *UAI*, 2016.
- H. Li, P. Chaudhari, H. Yang, M. Lam, A. Ravichandran, R. Bhotika, and S. Soatto. Rethinking the Hyperparameters for Fine-tuning. In *ICLR*, 2020.

- X. Li, Y. Grandvalet, and F. Davoine. Explicit Inductive Bias for Transfer Learning with Convolutional Networks. In *ICML*, 2018.
- T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.
- P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *arXiv:2107.13586 [cs]*, 2021.
- Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- M. Long, Y. Cao, J. Wang, and M. Jordan. Learning Transferable Features with Deep Adaptation Networks. In *ICML*, 2015.
- D. J. MacKay. Bayesian interpolation. *Neural computation*, 1992.
- D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018.
- S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi. Fine-Grained Visual Classification of Aircraft. *arXiv:1306.5151 [cs]*, 2013.
- L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner. dsprites: Disentanglement testing sprites dataset, 2017.
- S. Merity, C. Xiong, J. Bradbury, and R. Socher. Pointer Sentinel Mixture Models. In *ICLR*, 2017.
- K. P. Murphy. *Machine learning: a probabilistic perspective*. 2012.
- B. Neyshabur, H. Sedghi, and C. Zhang. What is being transferred in transfer learning? In *NeurIPS*, 2020.
- C. Nguyen, T. Hassner, M. Seeger, and C. Archambeau. LEEP: A New Measure to Evaluate Transferability of Learned Representations. In *ICML*, 2020.
- O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012.
- X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 2020.
- J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset shift in machine learning*. 2009.
- P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.

- C. E. Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, 2003.
- O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein. Imagenet large scale visual recognition challenge. *IJCV*, 2015.
- M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *CVPR*, 2018.
- E. T. K. Sang and F. De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *NAACL*, 2003.
- V. Sanh, L. Debut, J. Chaumond, and T. Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.
- Y. Shu, Z. Kou, Z. Cao, J. Wang, and M. Long. Zoo-Tuning: Adaptive Transfer from A Zoo of Models. In *ICML*, 2021.
- S. P. Singh and M. Jaggi. Model fusion via optimal transport. In *NeurIPS*, 2020.
- R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, 2013.
- C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. 2015.
- C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception Architecture for Computer Vision. In *CVPR*, 2016.
- M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le. Mnasnet: Platform-aware neural architecture search for mobile. In *CVPR*, 2019.
- S. Thrun and L. Pratt. Learning to Learn: Introduction and Overview. In *Learning to Learn*. 1998.
- Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola. What Makes for Good Views for Contrastive Learning? In *NeurIPS*, 2020.
- A. T. Tran, C. V. Nguyen, and T. Hassner. Transferability and hardness of supervised classification tasks. In *ICCV*, 2019.
- S. Vigna. A Weighted Correlation Index for Rankings with Ties. In *WWW*, 2015.
- P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, Í. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

- A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *EMNLP*, 2018.
- A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *NeurIPS*, 2019.
- T. Wolf, J. Chaumond, L. Debut, V. Sanh, C. Delangue, A. Moi, P. Cistac, M. Funtowicz, J. Davison, and S. Shleifer. Transformers: State-of-the-art natural language processing. In *EMNLP*, 2020.
- J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.
- Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, 2019.
- J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *NeurIPS*, 2014.
- K. You, Z. Kou, M. Long, and J. Wang. Co-Tuning for Transfer Learning. In *NeurIPS*, 2020.
- K. You, Y. Liu, J. Wang, and M. Long. LogME: Practical Assessment of Pre-trained Models for Transfer Learning. In *ICML*, 2021.
- S. Zagoruyko and N. Komodakis. Wide residual networks. 2017.
- A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese. Taskonomy: Disentangling Task Transfer Learning. In *CVPR*, 2018.
- X. Zhai, J. Puigcerver, A. Kolesnikov, P. Ruysen, C. Riquelme, M. Lucic, J. Djolonga, A. S. Pinto, M. Neumann, A. Dosovitskiy, L. Beyer, O. Bachem, M. Tschannen, M. Michalski, O. Bousquet, S. Gelly, and N. Houlsby. A Large-scale Study of Representation Learning with the Visual Task Adaptation Benchmark. *arXiv:1910.04867 [cs, stat]*, 2020.
- Z.-H. Zhou, J. Wu, and W. Tang. Ensembling neural networks: many could be better than all. *Artificial intelligence*, 2002.