arXiv:2508.07299v1 [cs.LG] 10 Aug 2025

When Is Prior Knowledge Helpful? Exploring the Evaluation and Selection of Unsupervised Pretext Tasks from a Neuro-Symbolic Perspective

Lin-Han Jia[12], Si-Yu Han[13], Wen-Chao Hu[13], Jie-Jing Shao[12], Wen-Da Wei[13], Zhi Zhou[12], Lan-Zhe Guo[14*], Yu-Feng Li[12*]

[1]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[2]School of Computer Science, Nanjing University, Nanjing, China
[3]School of Artificial Intelligence, Nanjing University, Nanjing, China
[4]School of Intelligence Science and Technology, Nanjing University, Nanjing, China
{jialh,hansy,huwc,shaojj,weiwd,zhouz,guolz,liyf}@lamda.nju.edu.cn
*Corresponding Authors

Abstract

Neuro-symbolic (Nesy) learning improves the target task performance of models by enabling them to satisfy knowledge, while semi/self-supervised learning (SSL) improves the target task performance by designing unsupervised pretext tasks for unlabeled data to make models satisfy corresponding assumptions. We extend the Nesy theory based on reliable knowledge to the scenario of unreliable knowledge (i.e., assumptions), thereby unifying the theoretical frameworks of SSL and Nesy. Through rigorous theoretical analysis, we demonstrate that, in theory, the impact of pretext tasks on target performance hinges on three factors: knowledge learnability with respect to the model, knowledge reliability with respect to the data, and knowledge completeness with respect to the target. We further propose schemes to operationalize these theoretical metrics, and thereby develop a method that can predict the effectiveness of pretext tasks in advance. This will change the current status quo in practical applications, where the selections of unsupervised tasks are heuristic-based rather than theory-based, and it is difficult to evaluate the rationality of unsupervised pretext task selection before testing the model on the target task. In experiments, we verify a high correlation between the predicted performance—estimated using minimal data—and the actual performance achieved after large-scale semi-supervised or self-supervised learning, thus confirming the validity of the theory and the effectiveness of the evaluation method.

Introduction

In the training of machine learning models, a large amount of labeled data is usually required to enable the models to achieve good performance on target tasks. However, in practical applications, data labeling is often expensive and difficult to obtain. Therefore, the utilization of unlabeled data has become the key to further improving performance under limited conditions.

Semi-supervised learning and self-supervised learning (SSL) are two types of methods currently studied to use unlabeled data to assist models in achieving target tasks(Van Engelen and Hoos 2020; Chen et al. 2020). They are usually studied independently but are theoretically the same in essence. The core lies in the fact that both need to set pretext tasks for unlabeled data based on prior assumptions. The main difference is reflected in the different training sequences of pretext tasks and target tasks. The former trains the target task using labeled data while training the pretext task using unlabeled data, while the latter first pre-trains the pretext task using unlabeled data and then fine-tunes the target task using labeled data. The key to the success of SSL lies in whether the selection of unsupervised pretext tasks is reasonable. However, there is currently a lack of theoretically based quantitative indicators to determine the reasonableness of pretext tasks. In practical applications, the selection of pretext tasks is usually heuristic, and the verification of the reasonableness of pretext tasks is usually lagging and difficult. On the one hand, it is necessary to retain verification data from the already scarce labeled data, which seriously affects the effectiveness of verification. On the other hand, it is necessary to train the model on the pretext task using large-scale unlabeled data, which results in high costs. This paper holds that in SSL, how to reliably evaluate the reasonableness of pretext tasks before model training, so as to efficiently select the most applicable pretext task, is an urgent problem to be solved.

Neuro-symbolic (Nesy) learning, like SSL, uses unlabeled data to improve model performance(Manhaeve et al. 2018). Unlike them, which rely on subjective assumptions, Nesy relies on more reliable knowledge. However, Nesy typically assumes that knowledge holds perfectly, which is overly idealistic. Notably, under the latest Bayesian scientific view, knowledge is generally considered probabilistic. Nesy also use a large amount of unlabeled data to gradually train and reduce the inconsistency between the model and knowledge, thereby integrating knowledge into the existing model. This is the same as SSL, which reduce the inconsistency between the model and assumptions by training on pretext tasks. We compared these three learning forms: self-supervised learning, semi-supervised learning, and neuro-symbolic learning. Based on their similar characteristics of needing to combine knowledge that holds with a certain probability to set pretext tasks for unlabeled data, we established a unified theory framework based on knowledge learnability, reliability and completeness. This framework provides a unified description for most scenarios in SSL and Nesy.

Our theoretical framework derives the following conclusions: The benefits of using unlabeled data based on probabilistic knowledge in model training are jointly determined by three factors: knowledge learnability with respect to the model, knowledge reliability with respect to the data, and knowledge completeness with respect to the target. We conduct a thorough theoretical analysis of these three factors and establish their relationships with the generalization error bound of the target task. This sheds light on the mystery of why unsupervised tasks can significantly influence downstream performance, and reveals the intricate connection between prior knowledge and the three fundamental components of traditional machine learning—model, data, and task—highlighting the critical role of prior knowledge as an emerging fundamental component in the next generation of machine learning algorithms.

In summary, from a theoretical perspective, obtaining the three indicators—learnability, reliability, and completeness—of the prior knowledge is sufficient to predict the model's generalization performance after SSL training on a given pretext task. However, in practical applications, empirically estimating these theoretical indicators remains a major challenge. In the current SSL literature, there is a lack of systematic metrics to quantify the relationship between pretext tasks and target tasks. As a result, the selection of pretext tasks often relies on heuristics, and there is no scientific method to determine whether training on a pretext task will benefit the target task or to identify the most suitable pretext task without extensive testing.

Fortunately, we have identified concrete methods for empirically evaluating the three theoretical indicators, enabling a rough estimation of the model's expected performance on a given pretext task prior to training, at a very low cost. It enables us to predict a model's potential performance on the target task before training or testing begins. This enables cost-effective pretext task selection, leading to improved performance in semi-supervised co-training or self-supervised pre-training(Devlin et al. 2019).

The learnability of the knowledge describes the extent to which the model satisfies the prior knowledge. Since the objective of unsupervised training is inherently to enforce compliance with the prior knowledge, the degree of non-compliance corresponds directly to the model's error on the unsupervised pretext task. This error can be estimated directly from the training loss on the pretext task, reflecting the model's capacity to fit the prior knowledge. Therefore, its empirical evaluation can be performed simply by executing the pretext task on a small set of unlabeled data and computing the loss. The reliability of the knowledge refers to the extent to which the knowledge holds over the data, i.e., whether the assumed knowledge is valid for the current dataset. In semi-supervised or self-supervised learning, knowledge typically stems from assumptions and may only hold for a subset of the data. Evaluating this reliability ideally requires an oracle model that can provide ground-truth labels. When such an oracle is not available, we approximate it using a supervised model trained on a small set of labeled data, which enables an empirical estimate of the knowledge's reliability. The completeness of the knowledge captures how sufficient the knowledge is for solving the target task. It measures whether satisfying the knowledge allows the model to achieve the target objective. This is quantified by how well the model can perform the target task after satisfying the knowledge. As with reliability, evaluating completeness relies on an approximate oracle model. Prior to this evaluation, the outputs of the unsupervised model must be aligned with the target task.

Finally, by taking the product of these three indicators, we can obtain a rough estimation of how beneficial a given unsupervised task is for the downstream target task.

To verify whether the model's target performance, as predicted by the empirical estimates of knowledge learnability, reliability, and completeness, can reliably anticipate its actual performance after large-scale semi-supervised or self-supervised training, we constructed a benchmark consisting of over a hundred pretext tasks. Across these various tasks, we compared the theoretically predicted performance—based on small amounts of data—with the true performance obtained after full-scale semi-supervised or self-supervised training using large datasets. Experimental results show that, even when using only 5 labeled samples and 50 unlabeled samples per task for rapid estimation, the predicted performance is highly correlated with the actual test performance obtained through full semi-supervised or self-supervised training on large-scale data. This demonstrates that the suitability of unsupervised tasks can be effectively assessed at a low cost.

Related Work

Neuro-Symbolic Learning

Research on Nesy can generally be classified into three types (Yu et al. 2023): Reasoning to learning, which focuses on constructing network architectures (Mao et al. 2019) or loss functions based on rules, emphasizing the mapping from X to S; Learning to reasoning, where learning methods are employed to identify reasoning paths (Mao et al. 2019) or accelerate the search process (Wang, Ye, and Gupta 2018), focusing on the reasoning from S to Y; Learning-reasoning: which emphasizes the interaction between the two components and where S is unknown and needs to be obtained based on both X and Y, ensuring consistency between the learning and reasoning processes (Xu et al. 2018; Stammer et al. 2023; Petersen et al. 2021).

Additionally, Nesy faces the issue of shortcuts (Yang et al. 2024; He et al. 2024; Jia et al.). The shortcut problem refers to the phenomenon where a model satisfies symbolic constraints or knowledge rules without truly learning the underlying semantics or reasoning patterns. Instead, the model finds superficial patterns or heuristics in the data that coincidentally conform to the given knowledge, allowing it to "game the system" and pass constraint checks without genuine understanding. Symbolic knowledge in NeSy systems is often incomplete or underspecified, allowing for multiple hypothesis solutions that satisfy the given constraints but differ in their generalization behavior. The shortcut problem is often rooted in incomplete knowledge: when the symbolic constraints are not sufficient to fully determine the correct

behavior, the model fills in the gaps with heuristics learned from data—some of which may be incorrect or unstable. The theoretical work of Nesy has been exploring the gap between satisfying knowledge and satisfying the final target under different circumstances. Many studies have highlighted that knowledge bases may have multiple satisfying solutions, but fail to accurately pinpoint the target one. However, few works have provided viable solutions to address this problem.

Semi-Supervised Learning

With the development of deep learning, deep SSL algorithms have received widespread attention in recent years. As research into semi-supervised learning deepens, researchers have increasingly uncovered a critical flaw: its lack of robustness. This weakness now severely limits its large-scale deployment and further development in real-world settings. Examining the issue from first principles, the root of the robustness problem lies in the inherent drawbacks of how unlabeled data are exploited. Semi-supervised learning hinges on several assumptions that are intended to extract latent structural information from unlabeled data and thereby improve model performance. Yet in the complex and variable conditions of real-world data, these assumptions often fail to hold. There is an urgent need to study robust SSL methods that do not suffer severe performance degradation when unlabeled data is corrupted. Oliver et al.(Oliver et al. 2018) demonstrate that using unlabeled data from unseen classes can actually hurt performance compared to not using any unlabeled data at all. Many subsequent works(Yu et al. 2020; Guo et al. 2020; Saito, Kim, and Saenko 2021; Cascante-Bonilla et al. 2021; Guo and Li 2022; He et al. 2022; Huang, Yang, and Gong 2022) follow this issue and propose many effective strategies.

Although extensive research has been conducted on the robustness and safety of semi-supervised learning, it still fails to prevent cases in real-world applications where incorporating unlabeled data yields no performance improvement—or even leads to significant degradation(Wang et al. 2022; Huang et al. 2023; Jia et al. 2024). This is primarily because prior work has not systematically or scientifically analyzed the impact of unsupervised tasks on the target task, which is precisely the problem our work aims to address.

Self-Supervised Learning

Self-supervised learning has emerged as a powerful paradigm for learning useful representations from unlabeled data by constructing pretext tasks that generate supervision signals from the data itself. It has achieved remarkable success in both computer vision and natural language processing.Early SSL approaches designed handcrafted pretext tasks such as predicting image rotations(Komodakis and Gidaris 2018), solving jigsaw puzzles(Noroozi and Favaro 2016), or colorizing grayscale images(Zhang, Isola, and Efros 2016). These tasks encouraged the model to learn spatial and semantic structure without human annotation. More recent advances have shifted toward contrastive learning frameworks, such as SimCLR(Chen et al. 2020), MoCo(He et al. 2020), and InfoNCE-based methods, which aim to maximize agreement between different augmentations of the same image while pushing apart representations of different images. These methods demonstrated that large-scale SSL can approach or even surpass supervised learning on several downstream benchmarks. Beyond contrastive learning, non-contrastive methods like BYOL(Grill et al. 2020), SimSiam(Chen and He 2021), and Barlow Twins(Zbontar et al. 2021) avoid negative samples altogether and still achieve competitive results, raising new questions about the underlying principles of representation learning.

Recent studies have identified a key challenge known as pretext task misalignment—where the objectives or representations learned during the pretext task do not align well with those required for the target task. In such cases, models trained with misaligned pretext tasks may even underperform compared to training from scratch or using only labeled data. This has motivated recent efforts to evaluate the transferability of SSL representations (Kornblith et al. 2019) and to design adaptive or task-aware pretext tasks. In the theoretical research of semi-supervised/self-supervised learning, there are also related works discussing the causal relationship between pretext tasks and target tasks(Lee et al. 2021; Huang et al.; Teng, Huang, and He 2022). However, despite these advances, predicting the effectiveness of a pretext task on a downstream target task remains an open challenge, particularly in semi-supervised or transfer learning settings.

Theory of Prior Knowledge Induction via Unsupervised Pretext Tasks

Problem Formulation

In classical machine learning tasks, algorithms learn a model $f \in \mathcal{F}$ from a hypothesis space $\mathcal{F}$, using labeled data $D_L \sim \mathcal{D}_{x,y}$ drawn from the distribution $\mathcal{D}_{x,y}$ to achieve the target task $Task_{Target}$. However, due to the limited availability of labeled data, this classical paradigm struggles to scale, leading to increasing interest in research areas such as semi-supervised learning and self-supervised learning, where pseudo-tasks $Task_{Pretext}$ are designed for unlabeled data $D_U \sim \mathcal{D}_x$ to assist in training or pretraining. Nonetheless, theoretical studies on how training with $Task_{Pretext}$ affects model performance on $Task_{Target}$ are relatively scarce, resulting in a heavy reliance on heuristic experience for the selection of $Task_{Pretext}$, and requiring a complete training and evaluation process to assess its effectiveness.

From the Bayesian perspective, machine learning is essentially the process of learning $p(y|x)$, whereas using unlabeled data only provides information about $p(x)$, which by itself does not contribute to learning $p(y|x)$. The reason why $Task_{Pretext}$ can be helpful for $Task_{Target}$ lies in its implicit introduction of prior knowledge $\mathcal{K}$ during training. We denote by $((x, \mathcal{K}) \models \text{True})$ the compatibility between x and the knowledge $\mathcal{K}$. Training on $Task_{Pretext}$ can enforce the model to satisfy $\mathcal{K}$, thereby forming a bridge to affect the target task via the relation:

$$\begin{aligned} p(y|x) =& p[y|(y, \mathcal{K}) \models \text{True}] \cdot p[(y, \mathcal{K}) \models \text{True}|(x, \mathcal{K}) \\ & \models \text{True}] \cdot p[(x, \mathcal{K}) \models \text{True}|x] \end{aligned} \tag{1}$$

As a result, the model fits all of $\mathcal{K}$ during training of the protext task. Here, $p((x, \mathcal{K}) \models \text{True}|x)$ reflects the learnability of the knowledge by the model—that is, to what extent the model satisfies the knowledge; $p((y, \mathcal{K}) \models \text{True}|(x, \mathcal{K}) \models \text{True})$ denotes the reliability of the knowledge with respect to the data—that is, to what extent the data conforms to the knowledge; $p(y|(y, \mathcal{K}) \models \text{True})$ represents the completeness of the knowledge with respect to the target—that is, the extent to which the target depends on the knowledge. When the knowledge is fully reliable, $p((y, \mathcal{K}) \models \text{True}|(x, \mathcal{K}) \models \text{True}) = 1$, which means that for a data point (x, y), if x satisfies the knowledge, then y also satisfies the knowledge—there exists no data that violates the knowledge. In Neuro-Symbolic learning, knowledge is considered as reliable. However, in semi-supervised and self-supervised learning, the knowledge is not always reliable—it is only assumed to hold with certain probability, denoted as $\mathcal{K}_A$.

Given a pretext $Task_{Pretext}$ and its corresponding assumed knowledge $\mathcal{K}_A$, if we can accurately estimate its learnability, reliability, and completeness, we can make informed predictions about the performance on the target task.

Theoretical Derivation

Let R_{Target} denote the generalization error of the model on the target task:

$$R_{Target}(f) = P_{x,y\sim\mathcal{D}_{x,y}}(f(x) \neq y) \tag{2}$$

Let $R_{Unlearnable}$ denote the generalization error of the model on the unsupervised pretext task, which characterizes the learnability of the assumed prior knowledge $\mathcal{K}_A$:

$$R_{Unlearnable}(f, \mathcal{K}_A) = P_{x\sim\mathcal{D}_x}(f(x), \mathcal{K}_A \not\models True) \tag{3}$$

Considering the case where the knowledge is both reliable and complete, the pretext task is equivalent to the target task—i.e., satisfying the knowledge is tantamount to accomplishing the target. Therefore, in the ideal case, the model's generalization error on the target task is equivalent to the learnability of the knowledge associated with the pretext task. It thus follows that:

$$R_{Target}(f) \leq R_{Unlearnable}(f, \mathcal{K}_A) \tag{4}$$

Next, we consider the case where the knowledge is not reliable. In this case, the generalization performance on the target task can no longer be determined solely by the model's degree of compliance with the assumed knowledge $\mathcal{K}_A$. This is because data that do not actually satisfy the knowledge may be incorrectly identified as satisfying it under the influence of $\mathcal{K}_A$, leading to an overestimation of the model's expected degree of knowledge satisfaction relative to the true degree. The overestimated portion corresponds to data that the learner fits as conforming to the knowledge, while in fact they contradict it. To reflect the unreliability of the knowledge over the data distribution, we define the following unreliability metric.

$$\begin{aligned} &R_{Unreliable}(f, \mathcal{K}_A) \\ =&P_{x,y\sim\mathcal{D}_{x,y},f(x),\mathcal{K}_A\models True}(y, \mathcal{K}_A \not\models True) \end{aligned} \tag{5}$$

This indicates that when the knowledge is unreliable, the generalization error on the target task depends not only on the portion of the data where the knowledge is unlearnable, but also on the portion where the knowledge is learnable but inconsistent with the actual ground truth:

$$\begin{aligned} &R_{Target}(f) \leq R_{Unlearnable}(f, \mathcal{K}_A) \\ &+ (1 - R_{Unlearnable}(f, \mathcal{K}_A)) \cdot R_{Unreliable}(f, \mathcal{K}_A) \end{aligned} \tag{6}$$

Next, we consider the case where the knowledge is incomplete. In such cases, even if the model fully satisfies all true knowledge, it does not necessarily guarantee good performance on the target task. This indicates that multiple sets of predictions may satisfy the given knowledge, and additional, more complete knowledge is required to distinguish the correct target predictions. This issue is a common form of the shortcut problem in neuro-symbolic learning. We define the incompleteness error as the portion where, despite the knowledge being both learnable and consistent with the data, the model still fails to correctly predict the target.

$$\begin{aligned} &R_{Incomplete}(f, \mathcal{K}_A) \\ =&P_{x,y\sim\mathcal{D}_{x,y},f(x),\mathcal{K}_A\models True,y,\mathcal{K}_A\models True}(f(x) \neq y) \end{aligned} \tag{7}$$

Finally, when the knowledge is potentially unreliable and incomplete, the generalization error on the target task is jointly determined by three factors: (1) the portion of the data where the knowledge is unlearnable, (2) the portion where the knowledge is learnable but inconsistent with the ground truth, and (3) the portion where the knowledge is both learnable and consistent with the data, yet still fails to yield accurate predictions. This leads to the following formulation:

$$\begin{aligned} &R_{Target}(f) \\ \leq&R_{Unlearnable}(f, \mathcal{K}_A) + (1 - R_{Unlearnable}(f, \mathcal{K}_A)) \\ &\cdot R_{Unreliable}(f, \mathcal{K}_A) + (1 - R_{Unlearnable}(f, \mathcal{K}_A)) \\ &\cdot (1 - R_{Unreliable}(f, \mathcal{K}_A)) \cdot R_{Incomplete}(f, \mathcal{K}_A) \\ =&1 - (1 - \hat{R}_{Unlearnable}(f, \mathcal{K}_A)) \cdot (1 - \hat{R}_{Unreliable}(f, \mathcal{K}_A)) \\ &\cdot (1 - \hat{R}_{Incomplete}(f, \mathcal{K}_A)) \end{aligned} \tag{8}$$

Efficient Evaluation of the Impact of Pretext Tasks on Target Task Performance

Based on our theoretical results, we conclude that the generalization error of a model trained with an unsupervised pretext task on a downstream target task is determined by the product of three theoretical factors: the learnability, reliability, and completeness of the corresponding knowledge. However, these theoretical indicators are typically difficult to compute directly in real-world applications. In our experiments, we aim to approximate these three indicators using only a small amount of labeled data and unlabeled data, with minimal computational cost. This enables us to assess, prior to training and without relying on test performance, whether a given unsupervised pretext task is likely to benefit the current learning objective. This section introduces the estimation methods for each of the three indicators.

The Estimation of Learnability

By definition, learnability describes the extent to which the knowledge is learned by the model, and it primarily concerns

whether the knowledge is satisfied by the current model. Since the objective of unsupervised training is precisely to make the model conform to this knowledge, the degree of unsatisfiability essentially represents the model's general error on the pretext task which can be directly estimated by the training error on the task, reflects the model's ability to fit the prior knowledge. Therefore, in this step, we directly train the model on the pretext task using a small sampled unlabeled dataset $\tilde{D}_U \sim D_U$, and obtain a model $f_{Pretext}$ that minimizes the pretext task error:

$$\hat{f}_{Pretext} = \min_f \frac{1}{|\tilde{D}|} \sum_{x \in \tilde{D}} \mathbb{I}(f(x), \mathcal{K}_A \not\models True) \quad (9)$$

Subsequently, the training error of $f_{Pretext}$ on the sampled unlabeled dataset $\tilde{D}_U \sim D_U$ can be used as an estimate of $R_{Unlearnable}$:

$$\hat{R}_{Unlearnable} = \frac{1}{|\tilde{D}_U|} \sum_{x \in \tilde{D}_U} \mathbb{I}(\hat{f}_{Pretext}(x), \mathcal{K}_A \not\models True) \quad (10)$$

For example, in consistency regularization for semi-supervised learning and contrastive learning for self-supervised learning, the pretext task involves applying two transformations to the model and minimizing the difference between their predictions. The corresponding prior knowledge is that the label should remain unchanged after two different transformations of the same data. $\hat{R}_{Unlearnable}$ represents the proportion of unlabeled data for which the model still fails to achieve prediction consistency after consistency-based optimization.

After this estimation, we proceed to the next metric by selecting the subset of unlabeled data for which the knowledge is satisfied.

$$\tilde{D}_U^{Learnable} = \{x \in \tilde{D}_U \mid \hat{f}_{Pretext}(x), \mathcal{K}_A \models True\} \quad (11)$$

The Estimation of Reliability

According to its definition, Reliability describes the extent to which the knowledge is consistent with the data, with a particular focus on whether the knowledge holds true on the current dataset. In semi-supervised or self-supervised learning, since the knowledge originates from assumed priors, it is inherently valid only for a subset of the data. However, during unsupervised training, it is typically assumed that all data conform to the knowledge assumptions. This mismatch may cause certain samples—whose true labels conflict with the knowledge—to be forcibly fitted by the model as if they satisfy the knowledge, even though they do not. Such samples inevitably fail to contribute positively to model performance. To empirically estimate the degree to which the knowledge is violated, we need to determine whether a conflict exists between the data's true label and the imposed knowledge. This is denoted as:

$$\hat{R}_{Unreliable} = \frac{1}{|\tilde{D}_U^{Learnable}|} \sum_{x \sim \tilde{D}_U^{Learnable}} \mathbb{I}(y_x, \mathcal{K}_A \not\models True) \quad (12)$$

Here, y_x denotes the true label of an unlabeled data sample x. However, since y_x is unknown in practice, this poses a challenge for estimating $\hat{R}_{Unreliable}$. Ideally, we would expect to have access to an oracle model that can directly provide the true label y_x for any unlabeled sample x. Given the impracticality of such an oracle, we approximate it by training a model on a small amount of labeled data corresponding to the target task. This pretext model serves as a practical substitute for the oracle during evaluation, i.e.,

$$\hat{f}_{Target} = \min_f \frac{1}{|D_L|} \sum_{x,y \in D_L} \mathbb{I}(f(x) \neq y) \quad (13)$$

Consequently, this enables a more refined empirical estimation of $\hat{R}_{Unreliable}$, given by:

$$\hat{R}_{Unreliable} = \frac{1}{|\tilde{D}_U^{Learnable}|} \sum_{x \sim \tilde{D}_U^{Learnable}} \mathbb{I}(\hat{f}_{Target}(x), \mathcal{K}_A \not\models True) \quad (14)$$

After completing the estimation, we select the data samples whose predicted true labels are consistent with the knowledge base, and use them for the estimation of the next indicator, i.e.,

$$\tilde{D}_U^{Reliable} = \{x \in \tilde{D}_U^{Learnable} \mid \hat{f}_{Target}(x), \mathcal{K}_A \models True\} \quad (15)$$

The Estimation of Completeness

By definition, Completeness describes the extent to which the knowledge is sufficient for the target task, focusing on whether the knowledge adequately covers the target. In the ideal case, the pretext task is identical to the target task, or the target task is a subset of the pretext task. In such scenarios, successfully solving the pretext task is equivalent to accomplishing the target task, and the prior knowledge associated with the pretext task is considered complete. However, in practice, even if the model fully learns the given knowledge, it may still fall short of achieving the target task perfectly. Completeness quantifies the extent to which the target can be achieved using only reliable knowledge. To empirically estimate completeness, we need to assess whether the model trained on the pretext task truly has the potential to solve the target task. As before, since the true label y_x is unknown, we approximate it using $\hat{f}_{\text{Target}}(x)$, i.e.,

$$\hat{R}_{Incomplete} = \frac{1}{|\tilde{D}_U^{Reliable}|} \sum_{x \sim \tilde{D}_U^{Reliable}} \mathbb{I}(\hat{f}_{Target}(x) \neq \hat{f}_{Pretext}(x)) \quad (16)$$

It is worth noting that $\hat{f}_{\text{Pretext}}$ is the model trained solely on the pretext task, and its prediction targets are not inherently aligned with those of the actual target task. Therefore, an additional target alignment process is required. Specifically, we use $\hat{f}_{\text{Target}}$ as a feature extractor to obtain embeddings on the labeled dataset D_L. For each class label, we compute the

mean of its corresponding embeddings to form prototypes. The distance between a sample's embedding and each class prototype is then used for classification. This process helps to align $\hat{f}_{\text{Pretext}}$ with $\hat{f}_{\text{Target}}$ in terms of the target task.

The model's test performance after full semi-supervised or self-supervised learning can ultimately be predicted based on the three estimated indicators:

$$\hat{R}_{Target} = 1 - (1 - \hat{R}_{Unlearnable}) \cdot (1 - \hat{R}_{Unreliable}) \cdot (1 - \hat{R}_{Incomplete}) \quad (17)$$

Experiments

To verify whether the model's target performance—estimated based on empirical measures of knowledge learnability, reliability, and completeness—can accurately predict the model's actual performance after large-scale semi-supervised or self-supervised training, we constructed a benchmark consisting of over one hundred pretext tasks. Under different pretext tasks, we compared the quickly estimated performance (based on theoretical results with limited data) against the actual performance of models trained with a large amount of data using full semi-supervised or self-supervised procedures. To minimize estimation cost, we sampled only a very small subset from the large pool of unlabeled data for prediction in low-label scenarios. The experimental results align with our expectations, demonstrating that, regardless of whether the setting is semi-supervised learning (where unsupervised pseudo-tasks and supervised target tasks are trained jointly) or self-supervised learning (where pseudo-task pretraining is followed by fine-tuning on the target task), the final model performance is highly correlated with our theoretical estimates based on limited data.

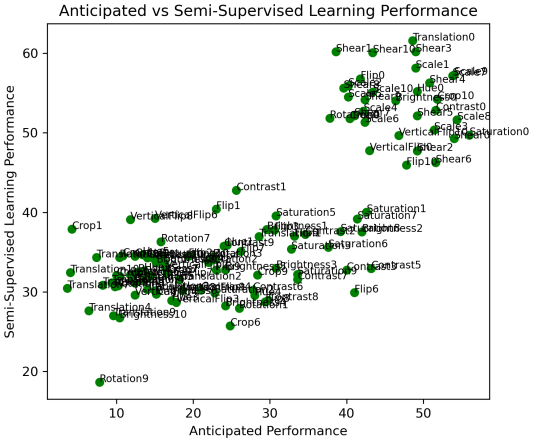

Figure 1: This figure illustrates the the comparison between predicted performance and final actual performance after full semi-supervised learning on CIFAR10.

Experimental Setting

We selected 11 commonly used unsupervised pretext tasks in current self-supervised and semi-supervised learning settings, and applied different augmentation strength parameters, ultimately constructing a benchmark containing 115 unsupervised pretext tasks which are shown in table 1. Under the same dataset, model, and hyperparameters, we used different pretext tasks to perform both performance estimation and actual training. These unsupervised pretext tasks include RandomResizedCrop, RandomRotation, Translate, Shear, Scale, Brightness, Contrast, Saturation, Hue, RandomHorizontalFlip, and RandomVerticalFlip—all of which are widely adopted in consistency-based semi-supervised learning and contrastive self-supervised pretraining. The prior knowledge introduced by these pretext tasks assumes that the label remains consistent after data transformations; specifically, the model is expected to produce identical predictions for two independently augmented views of the same input sample. We conducted experiments on the CIFAR-10 and CIFAR-100 datasets. For each class, we selected only 5 labeled examples—i.e., only 50 labeled samples in total for CIFAR-10 and 500 for CIFAR-100. For performance estimation based on theoretical results, we used 50 unlabeled examples per class—i.e., 500 unlabeled samples for CIFAR-10 and 5,000 for CIFAR-100. In the actual semi-supervised and self-supervised training, all the remaining training samples were used as unlabeled data—i.e., 49,950 samples for CIFAR-10 and 49,500 samples for CIFAR-100. All evaluations were conducted on the full test sets. Ultimately, we compared the estimated performance against the actual performance of self-supervised and semi-supervised learning on the complete test sets.

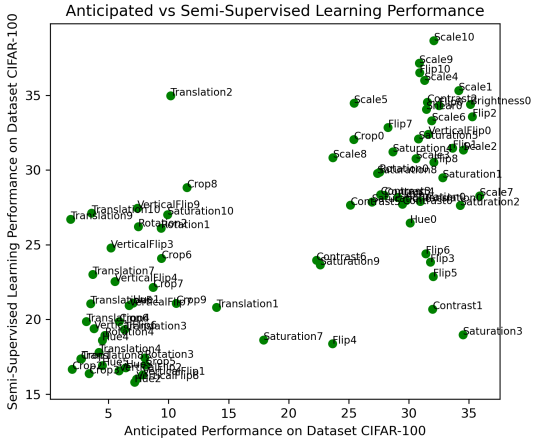

Figure 2: This figure illustrates the the comparison between predicted performance and final actual performance after full semi-supervised learning on CIFAR100.

In all experiments, we used ViT-B with a patch size of 16 as the baseline model. The batch size was set to 64 and the epoch is set to 5, the optimizer was Adam, the learning rate was fixed at 5e-5, and the weight decay was set to 0.01. The loss function for unsupervised pretext tasks was uniformly set to mean squared error (MSE) loss, and the loss function for the supervised target task was cross-entropy.

We use the Pearson correlation coefficient as the metric

Table 1: Candidate Pretext-Tasks for Experiments

Name	Strength	Operation
RandomResizedCrop	$\{s \mid s \in \mathbb{Z},\ 0 \leq s \leq 10\}$	RandomResizedCrop(scale=(s/10, s/10))
RandomRotation	$\{s \mid s \in \mathbb{Z},\ 0 \leq s \leq 10\}$	RandomRotation(degrees=s/10*180)
Translate	$\{s \mid s \in \mathbb{Z},\ 0 \leq s \leq 10\}$	RandomAffine(translate=(s/10, s/10))
Shear	$\{s \mid s \in \mathbb{Z},\ 0 \leq s \leq 10\}$	RandomAffine(shear=(s/10, s/10))
Scale	$\{s \mid s \in \mathbb{Z},\ 1 \leq s \leq 10\}$	RandomAffine(degrees=0, scale=(s/10, s/10))
Brightness	$\{s \mid s \in \mathbb{Z},\ 0 \leq s \leq 10\}$	ColorJitter(brightness=s/10)
Contrast	$\{s \mid s \in \mathbb{Z},\ 0 \leq s \leq 10\}$	ColorJitter(contrast=s/10)
Saturation	$\{s \mid s \in \mathbb{Z},\ 0 \leq s \leq 10\}$	ColorJitter(saturation=s/10)
Hue	$\{s \mid s \in \mathbb{Z},\ 0 \leq s \leq 5\}$	transforms.ColorJitter(hue=s/10)
RandomHorizonFlip	$\{s \mid s \in \mathbb{Z},\ 0 \leq s \leq 10\}$	transforms.RandomHorizontalFlip(p=s/10)
RandomVerticalFlip	$\{s \mid s \in \mathbb{Z},\ 0 \leq s \leq 10\}$	transforms.RandomVerticalFlip(p=s/10)

to evaluate the degree of correlation between predicted performance and true performance. Under the same conditions, this metric can be used to assess the effectiveness of the task selection method.

Experimental Results

Finally, we summarized the empirical indicators of knowledge learnability, reliability, and completeness for all 115 unsupervised pretext tasks under limited data evaluation. Based on these three indicators, we computed the predicted target performance. We also recorded the actual performance obtained through full semi-supervised and self-supervised learning processes evaluated on real test data. To intuitively illustrate the correlation between the predicted and actual performance, we plotted scatter diagrams. Figures 1 and 2 show the comparison between predicted performance and final actual performance after full semi-supervised learning on different pretext tasks for the CIFAR-10 and CIFAR-100 datasets, respectively. Figures 3 and 4 show the comparison between predicted performance and final actual performance after full self-supervised learning for different pretext tasks on the CIFAR-100 dataset.

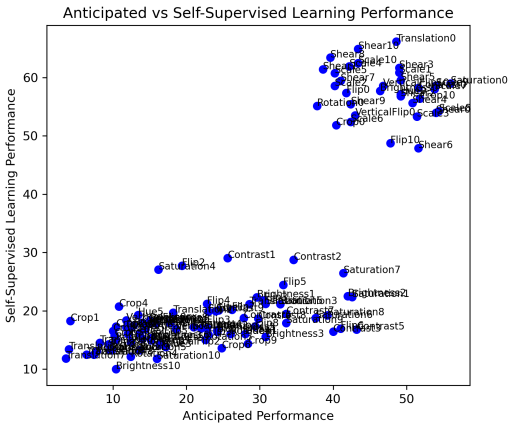

Figure 3: This figure illustrates the the comparison between predicted performance and final actual performance after full self-supervised learning on CIFAR10.

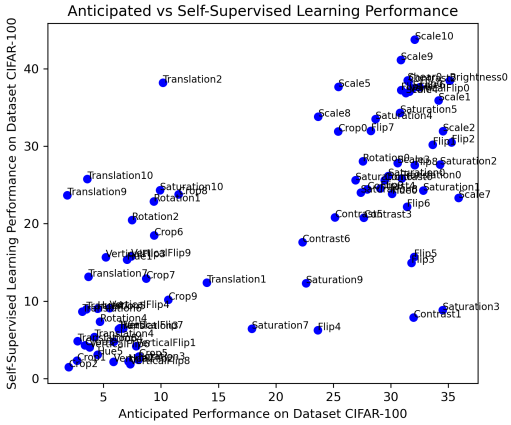

Figure 4: This figure illustrates the the comparison between predicted performance and final actual performance after full self-supervised learning on CIFAR100.

From the experimental results, it is evident that whether using semi-supervised learning and supervised target tasks are executed jointly—or self-supervised learning—the final test performance is highly correlated with the predicted performance derived from the three empirical theoretical indicators, as long as the same pretext task is used. The Pearson correlation coefficient between predicted and actual performance for semi-supervised learning reaches 0.785 and 0.717 on CIFAR-10 and CIFAR-100, respectively; for self-supervised learning, the correlation coefficients are 0.820 and 0.724 on CIFAR-10 and CIFAR-100, respectively.

Conclusion

We conducted an in-depth investigation into the underlying reasons why unlabeled data can enhance the performance of target tasks, specifically how training on unsupervised pretext tasks incorporates prior knowledge into models. Through rigorous theoretical analysis, we demonstrate that, in theory, the impact of unsupervised tasks on final performance hinges on three factors: knowledge learnability with respect to the model, knowledge reliability with respect to the data, and knowledge completeness with respect to the

target. This sheds light on the mystery of why unsupervised tasks can significantly influence downstream performance, and reveals the intricate connection between prior knowledge and the three fundamental components of traditional machine learning—model, data, and task—highlighting the critical role of prior knowledge as an emerging fundamental component in the next generation of machine learning.

More importantly, we propose empirical methods to quantify the theoretical indicators of learnability, reliability, and completeness in applications. These methods enable low-cost, approximate estimation of the three factors, allowing us to predict a model's potential performance on the target task before training or testing begins. This makes it feasible to rapidly assess the applicability of unsupervised pretext tasks in practical scenarios, dramatically reducing trial-and-error risks and costs, and shifting task selection from heuristic-based to theory-based. Our experiments confirm a high correlation between the predicted performance—estimated using minimal data—and the actual performance achieved after large-scale semi-supervised or self-supervised learning.

It is not difficult to foresee that in the future, the scale of models, data and targets will continue to grow, while reliable labels will not increase accordingly. This means that unsupervised training will occupy an increasingly significant role in the overall process. As the starting point of training, the selection of pretext tasks based on prior knowledge will become increasingly important throughout the machine learning pipeline. In the future, we plan to increase our efforts in the design of unsupervised tasks, aiming to enable models to learn more complex prior knowledge from unlabeled data, especially under more complex models, data, and targets.

References

Cascante-Bonilla, P.; Tan, F.; Qi, Y.; and Ordonez, V. 2021. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, 6912–6920.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, 1597–1607.

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15750–15758.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33: 21271–21284.

Guo, L.-Z.; and Li, Y.-F. 2022. Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding. In *Proceedings of the 39th International Conference on Machine Learning*, 8082–8094.

Guo, L.-Z.; Zhang, Z.-Y.; Jiang, Y.; Li, Y.-F.; and Zhou, Z.-H. 2020. Safe deep semi-supervised learning for unseen-class unlabeled data. In *Proceedings of the 37th International Conference on Machine Learning*, 3897–3906.

He, H.-Y.; Sun, H.; Xie, Z.; and Li, M. 2024. Ambiguity-Aware Abductive Learning. In *Proceedings of the 41st International Conference on Machine Learning*, 18019–18042.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

He, R.; Han, Z.; Lu, X.; and Yin, Y. 2022. Safe-student for safe deep semi-supervised learning with unseen-class unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14585–14594.

Huang, W.; Yi, M.; Zhao, X.; and Jiang, Z. ???? Towards the Generalization of Contrastive Self-Supervised Learning. In *Proceedings of the 11th International Conference on Learning Representations*.

Huang, Z.; Sidhom, M.-J.; Wessler, B.; and Hughes, M. C. 2023. Fix-A-Step: Semi-supervised Learning From Uncurated Unlabeled Data. In *International Conference on Artificial Intelligence and Statistics*, 8373–8394.

Huang, Z.; Yang, J.; and Gong, C. 2022. They are not completely useless: Towards recycling transferable unlabeled data for class-mismatched semi-supervised learning. *IEEE Transactions on Multimedia*.

Jia, L.-H.; Guo, L.-Z.; Zhou, Z.; and Li, Y.-F. 2024. Realistic evaluation of semi-supervised learning algorithms in open environments. In *Proceedings of the 12th International Conference on Learning Representations*.

Jia, L.-H.; Hu, W.-C.; Shao, J.-J.; Guo, L.-Z.; and Li, Y.-F. ???? Verification Learning: Make Unsupervised Neuro-Symbolic System Feasible. In *Proceedings of the 42nd International Conference on Machine Learning*.

Komodakis, N.; and Gidaris, S. 2018. Unsupervised representation learning by predicting image rotations. In *Proceedings of the 6th International Conference on Learning Representations*.

Kornblith, S.; Norouzi, M.; Lee, H.; and Hinton, G. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, 3519–3529.

Lee, J. D.; Lei, Q.; Saunshi, N.; and Zhuo, J. 2021. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34: 309–323.

Manhaeve, R.; Dumancic, S.; Kimmig, A.; Demeester, T.; and De Raedt, L. 2018. Deepproblog: Neural probabilistic logic programming. In *Advances in Neural Information Processing Systems*, 3753–3763.

Mao, J.; Gan, C.; Kohli, P.; Tenenbaum, J. B.; and Wu, J. 2019. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In

Proceedings of the 7th International Conference on Learning Representations.

Noroozi, M.; and Favaro, P. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, 69–84.

Oliver, A.; Odena, A.; Raffel, C. A.; Cubuk, E. D.; and Goodfellow, I. 2018. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, 3239–3250.

Petersen, F.; Borgelt, C.; Kuehne, H.; and Deussen, O. 2021. Learning with algorithmic supervision via continuous relaxations. *Advances in Neural Information Processing Systems*, 34: 16520–16531.

Saito, K.; Kim, D.; and Saenko, K. 2021. Openmatch: Open-set semi-supervised learning with open-set consistency regularization. In *Advances in Neural Information Processing Systems*, 25956–25967.

Stammer, W.; Friedrich, F.; Steinmann, D.; Brack, M.; Shindo, H.; and Kersting, K. 2023. Learning by self-explaining. *arXiv preprint arXiv:2309.08395*.

Teng, J.; Huang, W.; and He, H. 2022. Can pretext-based self-supervised learning be boosted by downstream data? a theoretical analysis. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 4198–4216.

Van Engelen, J. E.; and Hoos, H. H. 2020. A survey on semi-supervised learning. *Machine Learning*, 109(2): 373–440.

Wang, X.; Ye, Y.; and Gupta, A. 2018. Zero-Shot Recognition via Semantic Embeddings and Knowledge Graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6857–6866.

Wang, Y.; Chen, H.; Fan, Y.; Sun, W.; Tao, R.; Hou, W.; Wang, R.; Yang, L.; Zhou, Z.; Guo, L.-Z.; et al. 2022. Usb: A unified semi-supervised learning benchmark for classification. In *Advances in Neural Information Processing Systems*, 3938–3961.

Xu, J.; Zhang, Z.; Friedman, T.; Liang, Y.; and Broeck, G. 2018. A semantic loss function for deep learning with symbolic knowledge. In *Proceedings of the 35th International Conference on Machine Learning*, 5502–5511.

Yang, X.-W.; Wei, W.-D.; Shao, J.-J.; Li, Y.-F.; and Zhou, Z.-H. 2024. Analysis for Abductive Learning and Neural-Symbolic Reasoning Shortcuts. In *Proceedings of the 41st International Conference on Machine Learning*, 56524–56541.

Yu, D.; Yang, B.; Liu, D.; Wang, H.; and Pan, S. 2023. A survey on neural-symbolic learning systems. *Neural Networks*, 166: 105–126.

Yu, Q.; Ikami, D.; Irie, G.; and Aizawa, K. 2020. Multi-task curriculum framework for open-set semi-supervised learning. In *Proceedings of the 16th European Conference on Computer Vision*, 438–454.

Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; and Deny, S. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning*, 12310–12320.

Zhang, R.; Isola, P.; and Efros, A. A. 2016. Colorful image colorization. In *European Conference on Computer Vision*, 649–666.