

# Uncertainty-aware Self-training for Text Classification with Few Labels

## 具有不确定性的自我训练，用于很少标签的文本分类

Subhabrata Mukherjee  
苏哈布拉塔慕克吉

Ahmed Hassan Awadallah  
艾哈迈德□哈桑□阿瓦达拉

Microsoft Research AI  
Redmond, WA  
submukhe@microsoft.com  
微软研究 AI Redmond, WA  
简体中文 English

Microsoft Research AI  
Redmond, WA  
hassanam@microsoft.com  
微软研究 A I Redmond , WA  
简体中文 English

## Abstract

Recent success of large-scale pre-trained language models crucially hinge on fine-tuning them on large amounts of labeled data for the downstream task, that are typically expensive to acquire. In this work, we study self-training as one of the earliest semi-supervised learning approaches to reduce the annotation bottleneck by making use of large-scale unlabeled data for the target task. Standard self-training mechanism randomly samples instances from the unlabeled pool to pseudo-label and augment labeled data. In this work, we propose an approach to improve self-training by incorporating uncertainty estimates of the underlying neural network leveraging recent advances in Bayesian deep learning. Specifically, we propose (i) acquisition functions to select instances from the unlabeled pool leveraging Monte Carlo (MC) Dropout, and (ii) learning mechanism leveraging model confidence for self-training. As an application, we focus on text classification on five benchmark datasets. We show our methods leveraging only 20-30 labeled samples per class for each task for training and for validation can perform within 3% of fully supervised pre-trained language models fine-tuned on thousands of labeled instances with an aggregate accuracy of 91% and improving by upto 12% over baselines.

大规模预训练语言模型最近的成功关键在于对下游任务的大量标记数据进行微调，而获取这些数据通常成本很高。在这项工作中，我们研究了作为最早的半监督式学习方法之一的自我训练，以通过利用大规模未标记数据来减少标注瓶颈。标准的自训练机制将实例从未标记池随机抽样到伪标签并增加标记数据。在这项工作中，我们提出了一种利用贝叶斯深度学习的最新进展，结合基础神经网络不确定性估计值来改进自我训练的方法。具体而言，我们建议 (i) 采集函数，利用 Monte Carlo (MC) Dropout 从未标记池中选择实例，(ii) 利用模型信心进行自我训练的学习机制。作为一个应用程序，我们专注于五个基准数据集的文本分类。我们展示了我们的方法，每门课程仅利用 20 - 30 个标记样本进行训练和验证，在经过数千个标记实例微调的完全监督式预训练语言模型中，仅有 3% 的模型可以执行，总准确度达到 91%，并且比基线提高高达 12%。

## 1 Introduction

## 1 引言

Motivation. Deep neural networks are the state-of-the-art for various natural language processing applications. However, one of the biggest challenges facing them is the lack of labeled data to train these complex networks. Not only is acquiring large amounts of labeled data for every task expensive and time consuming, but also it is not feasible to perform large-scale human labeling, in many cases, due to data access and privacy constraints. Recent advances in pre-training help close this gap. In this, deep and large neural networks like BERT [Devlin et al., 2019], GPT-2 [Radford et al., 2019] and RoBERTa [Liu et al., 2019] are trained on millions of documents in a self-supervised fashion to obtain general purpose language representations. A significant challenge now is to fine-tune these models on downstream tasks that still rely on thousands of labeled instances for their superior performance. Semi-supervised learning (SSL) [Chapelle et al., 2010] is one of the promising paradigms to address this shortcoming by making effective use of large amounts of unlabeled data in addition to some labeled data for task-specific fine-tuning. A recent work [Xie et al., 2019] leveraging SSL for consistency learning has shown state-of-the-art performance for text classification with limited labels leveraging auxiliary resources like backtranslation and forms a strong baseline for our work.

动机。深度神经网络是各种自然语言处理应用的最先进技术。然而，他们面临的重大挑战之一是缺乏标记数据来训练这些复杂的网络。不仅为每个任务获取大量标记数据既昂贵又耗时，而且在许多情况下，由于数据访问和隐私限制，执行大规模人工标记也不可行。最近在预培训方面取得的进展有助于缩小这一差距。在此过程中，深度和大型神经网络如 BERT [Devlin et al., 2019]、GPT-2 [Radford et al., 2019] 和 RoBERTa [Liu et al., 2019] 以自我监督的方式对数百万文档进行训练，以获得通用语言表示。目前的一项重大挑战是针对下游任务微调这些模型，这些任务仍然依赖数千个标记实例来实现卓越性能。半监督式学习 (SSL) [Chapelle et al., 2010] 是解决这一缺陷的有前途的范式之一，它通过有效利用大量未标记数据以及一些标记数据进行特定任务的微调。最近一项利用 SSL 进行一致性学习的工作 [Xie et al., 2019] 显示了利用反向翻译等辅助资源的有限标签进行文本分类的最先进性能，并为我们的工作形成了坚实的基准。

Self-training (ST, [III, 1965]) as one of the earliest SSL approaches has recently been shown to obtain state-of-the-art performance for tasks like neural machine translation [He et al., 2019] performing at par with supervised systems without using any auxiliary resources. For self-training, a base model (teacher) is trained on some amount of labeled data and used to pseudo-annotate (task-specific) unlabeled data. The original labeled data is augmented with the pseudo-labeled data and used to train a student model. The student-teacher training is repeated until convergence. Traditionally, 自我训练 (ST, [III, 1965]) 是最早的 SSL 方法之一, 最近已被证明可以在不使用任何辅助资源的情况下, 为神经机器翻译等任务获得最先进的性能 [He et al., 2019]。对于自我训练, 基础模型 (教师) 根据一定量的标记数据进行训练, 并用于伪注释 (特定于任务) 未标记的数据。原始标记数据使用伪标记数据进行扩充, 并用于训练学生模型。学生 - 教师培训重复进行, 直到融合。传统上,

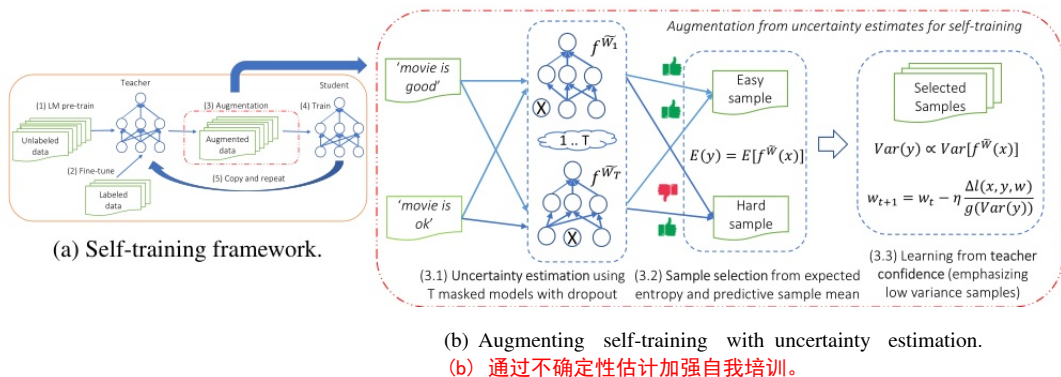


Figure 1: Uncertainty-aware self-training framework.

图 1 : 不确定性感知自训练框架。

self-training mechanisms do not consider the teacher uncertainty or perform any sample selection during the pseudo-labeling process. This may result in gradual drifts from self-training on noisy pseudo-labeled instances [Zhang et al., 2017]. Sample selection leveraging teacher confidence has been studied in curriculum learning [Bengio et al., 2009] and self-paced learning [Kumar et al., 2010] frameworks. These works leverage easiness of the samples to inform a learning schedule like training on easy concepts first followed by complex ones. Since it is hard to assess the “easiness” of a sample, especially in deep neural network based architectures, these works rely only on the teacher model loss while ignoring its uncertainties that can be leveraged for sample selection.

自我训练机制不考虑教师的不确定性，也不在伪标签过程中进行任何样本选择。这可能会导致在杂的伪标签实例中逐渐偏离自我训练 [Zhang et al., 2017]。在课程学习 [Bengio 等人, 2009 年] 和自定进度学习 [Kumar 等人, 2010 年] 框架中研究了利用教师信心的样本选择。这些工作利用简单的样本来告知学习计划，如简单概念的培训，首先是复杂概念的培训。由于很难评估样本的“易用性”，特别是在基于深度神经网络的架构中，这些工作仅依赖于教师模型损失，而忽略了可用于样本选择的不确定性。

Intuitively, if the teacher model already predicts some samples with high confidence, then there is little to gain with self-training if we focus only on these samples. On the other hand, hard examples for which the teacher model has less confidence are hard to rely on for self-training as these could be noisy or too difficult to learn from. In this scenario, the model could benefit from judiciously selecting examples for which the teacher model is uncertain about. However, it is non-trivial to generate uncertainty estimates for non-probabilistic models like deep neural networks. To this end, we leverage recent advances in Bayesian deep learning [Gal and Ghahramani, 2016] to obtain uncertainty estimates of the teacher for pseudo-labeling and improving the self-training process.

直观地说，如果教师模型已经以高可信度预测了一些样本，那么如果我们只关注这些样本，那么自我训练就没有什么好处。另一方面，教师模式缺乏信心的硬例子很难依赖自我训练，因为这些例子可能会很吵或太难学习。在这种情况下，模型可以从明智地选择教师模型不确定的例子中受益。然而，为深度神经网络等非概率模型生成不确定性估计值并不重要。为此，我们利用贝叶斯深度学习的最新进展 [Gal 和 Ghahramani, 2016 年]，以获得教师伪标签的不确定性估计，并改进自我培训过程。

Our task and framework overview. We focus on leveraging pre-trained language models for classification with few labeled samples (e.g.,  $K = \{20, 30\}$ ) per class for training and validation, and large amounts of task-specific unlabeled data. Figure 1(a) shows an overview of the self-training framework for NLU tasks, where augmented data is obtained from hard pseudo-labels from the teacher (e.g., BERT [Devlin et al., 2019]) without accounting for its uncertainty. We extend this framework with three core components in Figure 1(b), namely: (i) Masked model dropout for uncertainty estimation: We adopt MC dropout [Gal and Ghahramani, 2016] as a technique to obtain uncertainty estimates from the pre-trained language model. In this, we apply stochastic dropouts after different hidden layers in the neural network model and approximate the model output as a random sample from the posterior distribution. This allows us to compute the model uncertainty in terms of the stochastic mean and variance of the samples with a few stochastic forward passes through the network. (ii) Sample selection. Given the above uncertainty estimates for a sample, we employ entropy-based measures to select samples that the teacher is most or least confused about to infuse for self-training corresponding to easy- and hard-entropy-aware example mining. (iii) Confident learning. In this, we train the student model to explicitly account for the teacher confidence by emphasizing on the low variance examples. Finally, all of the above are jointly used for end-to-end learning. We adopt BERT as our encoder and show that its performance can be significantly improved by 12% for low-resource settings without using additional resources. Furthermore, we marginally outperform recent models [Xie et al., 2019] that make use of auxiliary resources like back-translation. In summary, our work makes the following contributions. (i) Develops an uncertainty-aware self-training framework for text classification with few labels. (ii) Compares the effectiveness of various sample selection schemes leveraging teacher uncertainty for self-training. (iii) Demonstrates its effectiveness for text classification with few labeled samples on five benchmark datasets.

我们的任务和框架概述。我们专注于利用预训练的语言模型进行分类，每个类只有少量标记样本（例如  $K = \{20, 30\}$ ）用于训练和验证，以及大量特定于任务的未标记数据。图 1 (a) 概述了 NLU 任务的自我训练框架，其中从教师的硬伪标签（例如，BERT [Devlin et al., 2019]）中获取增强数据，而不考虑其不确定性。我们通过图 1 (b) 中的三个核心部分扩展了该框架，即：(i) 用于不确定性估计的蒙蔽模型丢弃：我们采用 MC 丢弃 [Gal and Ghahramani, 2016] 作为从预训练语言模型获得不确定性估计的技术。在此，我们在神经网络模型中不同隐藏层之后应用随机丢弃，并将模型输出近似为后验分布中的随机样本。这使我们能够根据随机均值和样本方差计算模型不确定性，并通过网络进行一些随机前向传递。(ii) 抽样选择。鉴于上述样本的不确定性估计值，我们使用基于的测量来选择教师最困惑或最不困惑的样本，以进行与容易和难感知的示例挖掘相对应的自我训练。(iii) 自信学习。在此，我们训练学生模型，通过强调低方差示例来明确解释教师的信心。最后，上述所有内容共同用于端到端学习。我们采用 BERT 作为我们的编码器，并表明在低资源设置下，无需使用额外资源，其性能可显著提高 12%。此外，我们的表现略高于最近利用反向翻译等辅助资源的模型 [Xie et al., 2019]。总之，我们的工作作出了以下贡献。(i) 开发一个认识到不确定性的自我培训框架，用于标签少的文本分类。(ii) 比较利用教师不确定性进行自我培训的各种抽样选择办法的有效性。(iii) 展示其文本分类的有效性，在五个基准数据集上很少有标记样本。

## 2 Background

### 2 背景

Consider  $D_l = \{x_i, y_i\}$  to be a set of  $n$  labeled instances with  $y_i$  being the class label for  $x_i$ . Each  $x_i$  is a sequence of  $m$  tokens:  $x_i = \{x_{i1}, x_{i2} \dots x_{im}\}$ . Also, consider  $D_u = \{x_j\}$  to be a set of  $N$  unlabeled instances. We view  $D_l = \{x_i, y_i\}$  as a group of  $n$  labeled instances, where  $y_i$  is the class label for  $x_i$ . Each  $x_i$  is a sequence of  $m$  tokens:  $x_i = \{x_{i1}, x_{i2} \dots x_{im}\}$ . Also, consider  $D_u = \{x_j\}$  to be a set of  $N$  unlabeled instances.

unlabeled instances, where  $n \ll N$ . For most tasks, we have access to a small amount of labeled data along with a large amount of unlabeled ones.

未标记的实例，其中  $n \ll N$ 。对于大多数任务，我们可以访问少量标记数据以及大量未标记数据。

Self-training starts with a base teacher model trained on the labeled set  $D_l$ . The teacher model is applied to a subset  $S_u \subset D_u$  of the unlabeled data  $D_u$  to obtain pseudo-labeled instances. The augmented data  $D_l \cup S_u$  is used to train a student model. The teacher-student training schedules are repeated till some convergence criterion is satisfied. The unlabeled subset  $S$  is usually selected based on confidence scores of the teacher model. In Section 3.1, we study different techniques to generate this subset leveraging uncertainty of the teacher model. Self-training process can be formulated as: 自我训练从在标记集  $D_l$  上训练的基础教师模型开始。教师模型应用于未标记数据  $D_u$  的子集  $S_u \subset D_u$  以获取伪标记实例。增强数据  $D_l \cup S_u$  用于训练学生模型。师生培训计划重复进行，直至达到某种融合标准。未标记的子集  $S$  通常基于教师模型的置信度得分来选择。在第 3.1 节中，我们研究了不同的技术来利用教师模型的不确定性生成这个子集。自我训练过程可以表述为：

$$\min_W \mathbb{E}_{x_l, y_l \in D_l} [-\log p(y_l | x_l; W)] + \lambda \mathbb{E}_{x_u \in S_u, S_u \subset D_u} \mathbb{E}_{y \sim p(y | x_u; W^*)} [-\log p(y | x_u; W)] \quad (1)$$

where  $p(y | x; W)$  is the conditional distribution under model parameters  $W$ .  $W^*$  is given by the model parameters from the last iteration and fixed in the current iteration. The above optimization function has been used recently in variants of self-training for neural sequence generation [He et al., 2019] and data augmentation Xie et al. [2019].

其中  $p(y | x; W)$  是模型参数  $W$  下的条件分布。 $W^*$  由上次迭代的模型参数给出，并在当前迭代中固定。上述优化函数最近被用于神经序列生成 [He et al., 2019] 和数据增广的自训练变体 Xie et al. [2019]。

Bayesian neural network (BNN) [Gal and Ghahramani, 2015] assumes a prior distribution over its weights, thereby, replacing a deterministic model's weight parameters by a distribution over these parameters. For inference, instead of directly optimizing for the weights, BNN averages over all the possible weights, also referred to as marginalization.

Bayesian neural network (BNN) [Gal and Ghahramani, 2015] 假设了其权重的先验分布，从而通过这些参数的分布取代确定性模型的权重参数。对于推理，BNN 不是直接优化权重，而是对所有可能的权重进行平均，也称为边缘化。

Consider  $f^W(x) \in \mathbb{R}^h$  to be the  $h$ -dimensional output of such a neural network where the model likelihood is given by  $p(y | f^W(x))$ . For classification, we can further apply a softmax likelihood to the output to obtain:  $P(y = c | x, W) = \text{softmax}(f^W(x))$ .

考虑  $f^W(x) \in \mathbb{R}^h$  是这种神经网络的  $h$ -维输出，其中模型似然由  $p(y | f^W(x))$  给出。对于分类，我们可以进一步对输出应用 softmax 似然来获得： $P(y = c | x, W) = \text{softmax}(f^W(x))$ 。

Bayesian inference aims to find the posterior distribution over the model parameters  $p(W | X, Y)$ . Given an instance  $x$ , the probability distribution over the classes is given by marginalization over the posterior distribution as:  $p(y = c | x) = \int_W p(y = c | f^W(x)) p(W | X, Y) dW$ .

贝叶斯推理旨在找到模型参数  $p(W | X, Y)$  上的后验分布。给定一个实例  $x$ ，类上的概率分布通过后验分布上的边缘化给出： $p(y = c | x) = \int_W p(y = c | f^W(x)) p(W | X, Y) dW$ 。

This requires averaging over all possible model weights, which is intractable in practise. Therefore, several approximation methods have been developed based on variational inference methods and stochastic regularization techniques using dropouts. Here, the objective is to find a surrogate distribution  $q_\theta(w)$  in a tractable family of distributions that can replace the true model posterior that is hard to compute. The ideal surrogate is identified by minimizing the Kullback-Leibler (KL) divergence between the candidate and the true posterior.

这需要对所有可能的模型权重求取平均值，这在实践中是难以解决的。因此，基于变分推理方法和使用 dropouts 的随机正则化技术，开发了几种近似方法。这里的目标是在可处理的分布家族中找到替代分布  $q_\theta(w)$ ，该分布可以替代难以计算的后验真实模型。理想替代者通过最小化候选者和真实后者之间的 Kullback - Leibler (KL) 差异来确定。

Consider  $q_\theta(W)$  to be the Dropout distribution [Srivastava et al., 2014] which allows us to sample  $T$  masked model weights  $\{\tilde{W}_t\}_{t=1}^T \sim q_\theta(W)$ . For classification tasks, the approximate posterior can be now obtained by Monte-Carlo integration as:

将  $q_\theta(W)$  视为 Dropout 分布 [Srivastava et al., 2014]，它允许我们采样  $T$  掩模模型权重  $\{\tilde{W}_t\}_{t=1}^T \sim q_\theta(W)$ 。对于分类任务，现在可以通过蒙特卡洛积分获得近似后验值：

$$\begin{aligned} p(y = c | x) &\approx p(y = c | f^W(x)) q_\theta(W) dW \\ &\approx \frac{1}{T} \sum_{t=1}^T p(y = c | f^{\tilde{W}_t}(x)) = \frac{1}{T} \sum_{t=1}^T \text{softmax}(f^{\tilde{W}_t}(x)) \end{aligned} \quad (2)$$

### 3 Uncertainty-aware Self-training

### 3 不确定性自我训练

Given a pre-trained language model as the teacher, we first fine-tune it on the small amount of labeled data. To this end, we use a small batch size to gradually expose the teacher model to the few available labels. Given our low-resource setting, we do not compute uncertainty estimates over the small labeled set. Instead, given the teacher model, we compute uncertainty estimates over each instance from the large unlabeled set as follows. Considering dropouts enabled before every hidden layer in the teacher model, we perform several stochastic forward passes through the network for every unlabeled sample. For computational efficiency, we perform these stochastic passes and hence the self-training over sampled mini-batches.

给定一个预先训练的语言模型作为教师，我们首先对少量标记数据进行微调。为此，我们使用小批量大小来逐步将教师模型暴露在几个可用的标签上。鉴于我们的低资源设置，我们不计算小标记集的不确定性估计值。相反，考虑到教师模型，我们从大型未标记集中计算每个实例的不确定性估计值，如下所示。考虑到教师模型中每个隐藏层之前都启用了 dropout，我们为每个未标记的样本执行多个随机向前传递。为了提高计算效率，我们执行这些随机传递，从而对采样小批量进行自我训练。

For each unlabeled instance  $x_u$ , given  $T$  stochastic forward passes through the network with dropout, each pass  $t \in T$  with corresponding model parameters  $W_t \sim q_\theta(W)$ , generates a pseudo-label given by  $p(y_t^*) = \text{softmax}(f^{W_t}(x_u))$ .

对于每个未标记的实例  $x_u$ ，给定  $T$  随机向前通过带 dropout 的网络，每个传递  $t \in T$  与相应的模型参数  $W_t \sim q_\theta(W)$ ，生成由  $p(y_t^*) = \text{softmax}(f^{W_t}(x_u))$  给出的伪标签。

There are several choices to integrate this pseudo-label for self-training, including considering  $E(y) = \frac{1}{T} \sum_{t=1}^T \text{softmax}(f^{W_t}(x))$  for the soft pseudo-labels as well as discretizing them for hard集成此伪标签进行自我训练有多种选择，包括考虑  $E(y) = \frac{1}{T} \sum_{t=1}^T \text{softmax}(f^{W_t}(x))$  用于软伪标签，以及将其分散用于硬



labels and aggregating predictions from the T passes as:

来自 T 传递的标签和聚合预测如下：

$$y_u = \underset{c}{\operatorname{argmax}} \sum_{t=1}^T \mathbb{I}[\underset{c'}{\operatorname{argmax}} (p(y_t^* = c')) = c] \quad (3)$$

where  $\mathbb{I}(\cdot)$  is an indicator function. Empirically, the hard pseudo-labels work better in our framework with standard log loss. The pseudo-labeled data is used to augment and re-train the model with the steps repeated till convergence. At each self-training iteration, the model parameters  $W^*$  from the previous iteration is used to compute the predictive mean  $E(y)$  of the samples before re-training the model end-to-end on the augmented (pseudo-labeled) data to learn the new parameters  $W$ .

其中  $\mathbb{I}(\cdot)$  是指标函数。从经验上看，硬伪标签在我们的框架中使用标准日志丢失效果更好。伪标记数据用于扩充和重新训练模型，重复步骤直至收敛。在每次自训练迭代中，上一次迭代中的模型参数  $W^*$  用于计算样本的预测均值  $E(y)$ ，然后在增强（伪标记）数据上端到端重新训练模型以学习新参数  $W$ 。

In order to incorporate the above uncertainty measures in the self-training framework, we modify the loss component over unlabeled data in the original self-training learning process (Equation 1) as:

为了将上述不确定性度量纳入自训练框架，我们将原始自训练学习过程中未标记数据的损失部分修改为：

$$\min_{W, \theta} \mathbb{E}_{x_u \in S_u, S_u \subset D_u} \mathbb{E}_{\tilde{W} \sim q_{\theta}(W^*)} \mathbb{E}_{y \sim p(y|f^{\tilde{W}}(x_u))} [-\log p(y|f^W(x_u))] \quad (4)$$

where  $W^*$  denotes the model parameters from the previous iteration of the self-training process.

其中  $W^*$  表示自训练过程上一次迭代中的模型参数。

### 3.1 Sample Selection

#### 3.1 样本选择

Prior works have leveraged various measures to sample instances based on predictive entropy [Shannon, 2001], variation ratios [Freeman, 1965], standard deviation and more recently based on model uncertainty like Bayesian Active Learning by Disagreement (BALD) [Houlsby et al., 2011]. Consider  $D'_u = \{x_u, y_u\}$  to be the pseudo-labeled dataset obtained by applying the teacher model to the unlabeled data. The objective of the BALD measure is to select samples that maximize the information gain about the model parameters, or in other words, maximizing the information gain between predictions and the model posterior given by:  $\mathbb{B}(y_u, W|x_u, D'_u) = \mathbb{H}[y_u|x_u, D'_u] - \mathbb{E}_{p(W|D'_u)}[\mathbb{H}[y_u|x_u, W]]$ , where  $\mathbb{H}[y_u|x_u, W]$  denotes the entropy of  $y_u$  given  $x_u$  under model parameters  $W$ . Gal et al. [2017] show that the above measure can be approximated with the Dropout distribution  $q_{\theta}(W)$  such that:

以前的工作利用各种测量方法对基于预测 [Shannon, 2001]、变异比 [Freeman, 1965]、标准偏差和最近基于模型不确定性的实例进行采样，例如 Bayesian Active Learning by Disagreement (BALD) [Houlsby et al., 2011]。将  $D'_u = \{x_u, y_u\}$  视为通过将教师模型应用于未标记数据而获得的伪标记数据集。BALD 测度的目标是选择样本，使模型参数的信息增益最大化，换句话说，使预测和模型后验之间的信息增益最大化： $\mathbb{B}(y_u, W|x_u, D'_u) = \mathbb{H}[y_u|x_u, D'_u] - \mathbb{E}_{p(W|D'_u)}[\mathbb{H}[y_u|x_u, W]]$ ，其中  $\mathbb{H}[y_u|x_u, W]$  表示模型参数  $W$  下给定的  $x_u$  的  $y_u$ 。Gal 等人。[2017] 显示上述测度可以近似于 Dropout 分布  $q_{\theta}(W)$ ，这样：

$$\hat{\mathbb{B}}(y_u, W|x_u, D'_u) = -\sum_c \left( \frac{1}{T} \sum_t \hat{p}_c^t \right) \log \left( \frac{1}{T} \sum_t \hat{p}_c^t \right) + \frac{1}{T} \sum_{t,c} \hat{p}_c^t \log(\hat{p}_c^t) \quad (5)$$

where,  $\hat{p}_c^t = p(y_u = c|f^{W_t}(x_u) = \operatorname{softmax}(f^{W_t}(x_u)))$ .

The above measure depicts the decrease in the expected posterior entropy in the output  $y$  space. This results in a tractable estimation of the BALD acquisition function with  $\mathbb{B}(y_u, W|\cdot) \xrightarrow{T \rightarrow \infty} \hat{\mathbb{B}}(y_u, W|\cdot)$ 。上述测度描述了输出  $y$  空间中预期后的减少。这将导致使用  $\hat{\mathbb{B}}(y_u, W|\cdot)$  对 BALD 采集函数进行跟踪估计

$\mathbb{B}(y_u, W|\cdot)$ 。A high value of  $\hat{\mathbb{B}}(y_u, W|x_u, D'_u)$  indicates that the teacher model is highly confused about the expected label of the instance  $x_u$ . We use this measure to rank all the unlabeled instances based on uncertainty for further selection for self-training.

$\mathbb{B}(y_u, W|\cdot)$ 。  $\hat{\mathbb{B}}(y_u, W|x_u, D'_u)$  的高值表示教师模型高度混淆了实例  $x_u$  的预期标签。我们使用此措施根据自我训练的进一步选择的不确定性对所有未标记的实例进行排名。

Class-dependent selection. We can further modify this measure to take into account the expected class label of the instance. This helps in sampling equivalent number of instances per class, and avoids the setting where a particular class is typically hard, and the model mostly samples instances from that class. Given the pseudo-labeled set  $S_u$ , we can construct the set  $\{x_u \in S_{u,c} : y_u = c\}$  for every class  $c$ . Now, we use the BALD measure to select instances from each class-specific set instead of a global selection.

类相关选择。我们可以进一步修改此度量，以考虑实例的预期类标签。这有助于采样每个类的等效实例数，并避免特定类通常很难的设置，而模型主要是从该类中采样实例。给定伪标记集合  $S_u$ ，我们可以为每个类  $c$  构造集合  $\{x_u \in S_{u,c} : y_u = c\}$ 。现在，我们使用 BALD 测度从每个类特定集中选择实例，而不是全局选择。

Selection with exploration. Given the above measure, there are choices to select the pseudo-labeled examples for self-training, including mining easy ones (as in curriculum learning and self-paced learning) and hard ones. To this end, we can select the top-scoring instances for which the model is least or most uncertain about ranked by  $1 - \hat{\mathbb{B}}(y_u, W|x_u, D'_u)$  and  $\hat{\mathbb{B}}(y_u, W|x_u, D'_u)$  respectively. In the former case, if the model is always certain about some examples, then these might be too easy to contribute any additional information. In the latter case, emphasizing only on the hard examples may result in drift due to noisy pseudo-labels. Therefore, we want to select examples with some exploration to balance these schemes with sampling using the uncertainty masses. To this end, given a budget of  $B$  examples to select, we sample instances  $x_u \in S_{u,c}$  without replacement with probability: 选择与探索。考虑到上述措施，可以选择自我训练的伪标记示例，包括挖掘简单示例（如课程学习和自定进度学习）和硬示例。为此，我们可以选择模型对  $1 - \hat{\mathbb{B}}(y_u, W|x_u, D'_u)$  和  $\hat{\mathbb{B}}(y_u, W|x_u, D'_u)$  排名最少或最不确定的得分最高实例。在前一种情况下，如果模型总是确定一些例子，那么这些例子可能太容易贡献任何额外的信息。在后一种情况下，只强调硬例子可能会导致杂的伪标签的漂移。因此，我们希望通过一些探索来选择一些例子，以平衡这些方案与使用不确定性质量的抽样。为此，给定要选择的  $B$  示例的预算，我们采样实例  $x_u \in S_{u,c}$  而不使用概率替换：

$$p_{u,c}^{easy} = \frac{1 - \hat{\mathbb{B}}(y_u, W|x_u, D'_u)}{\sum_{x_u \in S_{u,c}} 1 - \hat{\mathbb{B}}(y_u, W|x_u, D'_u)} \quad (6) \quad p_{u,c}^{hard} = \frac{\hat{\mathbb{B}}(y_u, W|x_u, D'_u)}{\sum_{x_u \in S_{u,c}} \hat{\mathbb{B}}(y_u, W|x_u, D'_u)} \quad (7)$$

Our framework can use either of the above two strategies for selecting pseudo-labeled samples from the unlabeled pool for self-training; where these strategies bias the sampling process towards picking easier samples (less uncertainty) or harder ones (more uncertainty) for re-training.

我们的框架可以使用上述两种策略之一从未标记池中选择伪标记样本进行自我训练；这些策略将采样过程偏向于选择更容易的样本（不确定性较小）或更难的样本（不确定性更大）进行再训练。

### 3.2 Confident Learning

#### 3.2 自信学习

The above sampling strategies select informative samples for self-training conditioned on the posterior entropy in the label space. However, they use only the predictive mean, while ignoring the uncertainty of the model in terms of the predictive variance. Note that many of these strategies implicitly minimize the model variance (e.g., by focusing more on difficult examples for hard example mining). The prediction uncertainty of the teacher model is given by the variance of the marginal distribution, where the overall variance can be computed as:

上述采样策略根据标签空间中的后选择信息丰富的样本进行自我训练。但是，它们仅使用预测均值，而忽略模型在预测方差方面的不确定性。请注意，其中许多策略隐式地最小化了模型方差（例如，通过更多地关注硬示例挖掘的困难示例）。教师模型的预测不确定性由边缘分布的方差给出，其中总方差可以计算为：

$$Var(y) = Var[\mathbb{E}(y|W, x)] + \mathbb{E}[Var(y|W, x)] \quad (8)$$

$$= Var(\text{softmax}(f^W(x))) + \sigma^2 \quad (9)$$

$$\approx \left( \frac{1}{T} \sum_{t=1}^T y_t^*(x)^T y_t^*(x) - E(y)^T E(y) \right) + \sigma^2 \quad (10)$$

where,  $y_t^*(x) = \text{softmax}(f^{W_t}(x))$  and the predictive mean computed as:  $E(y) = \frac{1}{T} \sum_{t=1}^T y_t^*(x)$ .

We observe the total variance can be decomposed as a linear combination of the model uncertainty from parameters  $W$  and the second component results from noise in the data generation process.

我们观察到总方差可以分解为参数  $W$  中模型不确定性的线性组合，而第二个成分则是数据生成过程中噪声的结果。

In this phase, we want to train the student model to explicitly account for the teacher uncertainty for the pseudo-labels in terms of their predictive variance. This allows the student model to selectively focus more on the pseudo-labeled samples that the teacher is more confident on (corresponding to low variance samples) compared to the less certain ones (corresponding to high variance ones). Accordingly, we update the loss function over the unlabeled data in the self-training mechanism given by Equation 4 to update the student model parameters as:

在此阶段，我们希望训练学生模型，以便根据伪标签的预测方差明确解释教师的不确定性。这使得学生模型能够选择性地更多地关注教师更有信心的伪标记样本（对应于低方差样本），而不是较不确定的样本（对应于高方差样本）。相应地，我们更新方程 4 给出的自训练机制中未标记数据的损失函数，将学生模型参数更新为：

$$\min_{W, \theta} \mathbb{E}_{x_u \in S_u, S_u \subset D_u} \mathbb{E}_{\tilde{W} \sim q_{\theta}(W^*)} \mathbb{E}_{y \sim p(y|f^{\tilde{W}}(x_u))} [\log p(y|f^W(x_u)) \cdot Var(y)] \quad (11)$$

In the above equation, the per-sample loss for an instance  $x_u$  is a combination of the log loss  $-\log p(y)$  and (inverse of) its predictive variance given by  $\log \frac{1}{Var(y)}$  with log transformation for scaling. This penalizes the student model more on mis-classifying instances that the teacher is more certain on (i.e. low variance samples), and vice-versa.

在上述等式中，实例  $x_u$  的每个样本损失是对数损失  $-\log p(y)$  和具有用于缩放的对数变换的  $\log \frac{1}{Var(y)}$  给出的预测方差的组合。这更多地惩罚学生模型，因为教师更确定的错误分类实例（即低方差样本），反之亦然。

## 4 Experiments

### 4 实验

Encoder. Pre-trained language models like BERT [Devlin et al., 2019], GPT-2 [Radford et al., 2019] and RoBERTa [Liu et al., 2019] have shown state-of-the-art performance for various natural language processing tasks. In this work we adopt one of these namely, BERT as our base encoder or teacher model to start with. We initialize the teacher model with the publicly available pre-trained checkpoint from Wikipedia. To adapt the teacher language model for every downstream task, we further continue pre-training on task-specific unlabeled data  $D_u$  using the original language modeling objective. The teacher is finally fine-tuned on task-specific labeled data  $D_l$  to give us the base model for self-training.

编码器。BERT [Devlin et al., 2019]、GPT-2 [Radford et al., 2019] 和 RoBERTa [Liu et al., 2019] 等预训练语言模型在各种自然语言处理任务中表现出最先进的性能。在这项工作中，我们采用其中之一，即 BERT 作为我们的基础编码器或教师模型。我们使用来自维基百科的公开预训练检查点初始化教师模型。为了使教师语言模型适应每个下游任务，我们进一步使用原始语言建模目标继续对特定于任务的未标记数据  $D_u$  进行预训练。教师最终对特定于任务的标记数据  $D_l$  进行了微调，以便为我们提供自我训练的基础模型。

Datasets. We perform large-scale experiments with data from five domains for different tasks as summarized in Table 1. SST-2 [Socher et al., 2013], IMDB [Maas et al., 2011] and Elec [McAuley and Leskovec, 2013] are used for sentiment classification for movie reviews and Amazon electronics product reviews respectively. The other two datasets Dbpedia [Zhang et al., 2015] and Ag News [Zhang et al., 2015] are used for topic classification of Wikipedia sample K labeled instances from Train data, and 数据集。我们针对不同任务使用来自五个领域的数据进行大规模实验，如表 1 所示。SST-2 [Socher et al., 2013]、IMDB [Maas et al., 2011] 和 Elec [McAuley 和 Leskovec, 2013] 分别用于电影评论和亚马逊电子产品评论的情感分类。另外两个数据集 Dbpedia [Zhang et al., 2015] 和 Ag News [Zhang et al., 2015] 用于对来自 Train 数据的维基百科样本 K 标记实例进行主题分类。

Table 1: Dataset summary (W: avg. words / doc).

表 1：数据集汇总（W：平均值）中文（简体）

Dataset	Class	Train	Test	Unlabeled	#W
IMDB	2	25K	25K	50K	235
DBpedia	14	560K	70K	-	51
AG News	4	120K	7.6K	-	40
Elec	2	25K	25K	200K	108

and news articles respectively. For every dataset, we add remaining to the Unlabeled data in Table 1.

Evaluation setting. For self-training, we fine-tune the base model (teacher) on K labeled instances for each task to start with. Specifically, we consider  $K = 30$  instances for each class for training and 评价设定。对于自我训练，我们对每个任务开始时的 K 标记实例的基础模型（教师）进行微调。具体来说，我们考虑每个类的  $K = 30$  实例用于训练和

Table 2: Accuracy comparison of different models for text classification on five benchmark datasets. All models use the same BERT-Base encoder and pre-training mechanism. All models (except ‘all train’) are trained with 30 labeled samples for each class and overall accuracy aggregated over five different runs with different random seeds. The accuracy number for each task is followed by the standard deviation in parentheses and percentage improvement ( ) over the base model.

表 2：不同模型在五个基准数据集中进行文本分类的准确性比较。所有模型都使用相同的 BERT - Base 编码器和预训练机制。所有模型（除了 "所有火车"）都为每个类使用 30 个标记样本进行训练，并在不同随机种子的 5 次不同试验中汇总总体精度。每个任务的精确度数后面跟着括号中的标准差和基本模型的百分比改进（%）。

Dataset	All train	30 labels per class for training and for validation			
	BERT	BERT (base)	UDA	Classic ST	UST (our method)
SST	92.12	69.79 (6.45)	83.58 (2.64) (↑ 19.8)	84.81 (1.99) (↑ 21.5)	<b>88.19</b> (1.01) (↑ 26.4)
IMDB	91.70	73.03 (6.94)	<b>89.30</b> (2.05) (↑ 22.3)	78.97 (8.52) (↑ 8.1)	89.21 (0.83) (↑ 22.2)
Elec	93.46	82.92 (3.34)	89.64 (2.13) (↑ 8.1)	89.92 (0.36) (↑ 8.4)	<b>91.27</b> (0.31) (↑ 10.1)
AG News	92.12	80.74 (3.65)	85.92 (0.71) (↑ 6.4)	84.62 (4.81) (↑ 4.8)	<b>87.74</b> (0.54) (↑ 8.7)
DbPedia	99.26	97.77 (0.40)	96.88 (0.58) (↓ 0.9)	98.39 (0.64) (↑ 0.6)	<b>98.57</b> (0.18) (↑ 0.8)
Average	93.73	80.85 (4.16)	89.06 (1.62) (↑ 10.2)	87.34 (3.26) (↑ 8.0)	<b>91.00</b> (0.57) (↑ 12.6)

similar for validation, that are randomly sampled from the corresponding Train data in Table 1. We also show results of the final model on varying  $K \in \{20, 30, 50, 100, 500, 1000\}$ . We repeat each experiment five times with different random seeds and data splits, use the validation split to select the best model, and report the mean accuracy on the blind Test data. We implement our framework in Tensorflow and use four Tesla V100 gpus for experimentation. We use Adam [Kingma and Ba, 2015] as the optimizer with early stopping and use the best model found so far from the validation loss for all the models. Hyper-parameter configurations with detailed model settings presented in Appendix. We report results from our UST framework with easy sample selection strategy employing Equation 6, unless otherwise mentioned.

类似于验证，从表 1 中的相应 Train 数据随机抽样。我们还展示了最终模型在变化的  $K \in \{20, 30, 50, 100, 500, 1000\}$  上的结果。我们使用不同的随机种子和数据拆分重复每个实验五次，使用验证拆分选择最佳模型，并在盲测试数据上报告平均准确度。我们在 Tensorflow 中实现我们的框架，并使用四个 Tesla V100 gpus 进行实验。我们使用 Adam [Kingma and Ba, 2015] 作为优化器，并使用迄今为止所有模型验证损失中发现的最佳模型。附录中提供了详细模型设置的超参数配置。除非另有说明，否则我们会报告 UST 框架的结果，并采用方程 6 的简单样本选择策略。

Baselines . Our first baseline is BERT-Base with 110 MM parameters fine-tuned on K labeled samples  $D_l$  for downstream tasks with a small batch-size of 4 samples, and remaining hyper-parameters retained from its original implementation. Our second baseline, is a recent work UDA [Xie et al., 2019] leveraging backtranslation <sup>1</sup> for data augmentation for text classification. UDA follows similar principles as Virtual Adversarial Training (VAT) [Miyato et al., 2017] and consistency training [Laine and Aila, 2017, Sajjadi et al., 2016] such that the model prediction for the original instance is similar to that for the augmented instance with a small perturbation. In contrast to prior works for image augmentation (e.g., flipping and cropping) UDA leverages backtranslation for text augmentation. In contrast to other baselines, this requires auxiliary resources in terms of a trained NMT system to generate the backtranslation. Our third baseline is the standard self-training mechanism without any uncertainty. In this, we train the teacher model on  $D_l$  to generate pseudo-labels on  $D_u$ , train the student model on pseduo-labeled and augmented data, and repeat the teacher-student training till convergence. Finally, we also compare against prior SSL works – employing semi-supervised sequence learning [Dai and Le, 2015], adversarial training [Goodfellow et al., 2015, Miyato et al., 2017], variational pre-training [Gururangan et al., 2019], reinforcement learning [Li and Ye, 2018], temporal ensembling and mean teacher models [Laine and Aila, 2017, Tarvainen and Valpola, 2017, Sajjadi et al., 2016], layer partitioning [Li and Sethy, 2019] and delta training [Jo and Cinarel, 2019] – on these benchmark datasets on the same Test data and report numbers from corresponding works. 基线。我们的第一个基线是 BERT - Base，其 110 MM 参数在带有 K 标记的样本  $D_l$  上进行了微调，用于具有 4 个小批量样本的下游任务，并保留了原始实现中的剩余超参数。我们的第二个基线是 UDA [Xie et al., 2019] 最近的一项工作，利用反向翻译 1 进行文本分类的数据增强。UDA 遵循虚拟对抗训练 (VAT) [Miyato et al., 2017] 和一致性训练 [Laine and Aila, 2017, Sajjadi et al., 2016] 的类似原则，使得原始实例的模型预测与具有小扰动的增强实例的模型预测相似。与以前的图像增广（例如翻转和裁剪）相比，UDA 利用反向翻译进行文本增广。与其他基线相比，这需要辅助资源，即经过训练的 NMT 系统来生成反向翻译。我们的第三个基准是标准的自我培训机制，没有任何不确定性。在此，我们在  $D_l$  上训练教师模型以在  $D_u$  上生成伪标签，在 ps eduo 标记和增强数据上训练学生模型，并重复师生训练直到收。最后，我们还比较了之前的 SSL 工作 —— 采用半监督序列学习 [Dai and Le, 2015]、对抗式训练 [Goodfellow et al., 2015, Miyato et al., 2017]、变分预训练 [Gururangan et al., 2019]、强化学习 [Li and Ye, 2018]、时间汇编和平均教师模型 [Laine and Aila, 2017, Tarvainen and Valpola, 2017, Sajjadi et al., 2016]、层分割 [Li and Sethy, 2019] 和 delta 训练 [Jo and Cinarel, 2019] —— 基于相同测试数据和相应工作报告编号的这些基准数据集。

Overall comparison. Table 2 shows a comparison between the different methods. We observe that the base teacher model trained with only 30 labeled samples for each class for each task has a reasonable good performance with an aggregate accuracy of 80.85%. This largely stems from using BERT as the encoder starting from a pre-trained checkpoint instead of a randomly initialized encoder, thereby, demonstrating the effectiveness of pre-trained language models as natural few-shot learners. We observe the classic self-training approach leveraging unlabeled data to improve over the base model by 8%. The UDA model leverages auxiliary resources in the form of backtranslation from an NMT system for augmentation to improve by over 10%. Finally, our uncertainty-aware self-training mechanism obtains the best performance by improving more than 12% over the base model without any additional resources. Our method reduces the overall model variance in terms 总体比较。表 2 显示了不同方法之间的比较。我们观察到，基础教师模型仅针对每个任务的每节课进行了 30 个标记样本的训练，具有合理的良好性能，聚合精度为 80.85%。这主要源于使用 BERT 作为编码器，从预训练的检查点开始，而不是随机初始化的编码器，从而证明了预训练语言模型作为自然的少数镜头学习者的有效性。我们观察到经典的自我训练方法，利用未标记的数据，比基础模型提高 8 %。UDA 模型利用来自 NMT 系统的反向翻译形式的辅助资源进行增强，以提高 10% 以上。最后，我们的不确定性感知型自我训练机制比基础模型提高 12% 以上，无需任何额外资源即可获得最佳性能。我们的方法减少了整体模型方差

<sup>1</sup>A sentence is translated to a foreign language followed by backtranslation to the source language. Due to noise injected by Neural Machine Translation systems, backtranslation is often a paraphrase of the original. 1 句子被翻译成外语，然后反向翻译到源语言。由于神经机器翻译系统注入的噪音，反向翻译通常是原始版本的释义。



Table 3: Ablation analysis of our framework with different sample selection strategies and components including class-dependent sample selection with exploration (Class) and confident learning (Conf) for uncertainty-aware self-training with 30 labeled examples per class for training and for validation. 表 3：对我们的框架进行消融分析，包括不同的样本选择策略和组件，包括基于类别的样本选择与探索（类）和自信学习（Conf），用于不确定性感知自我训练，每个类有 30 个标记示例用于训练和验证。

	SST	IMDB	Elec	AG News	Dbpedia	Average
BERT Base	69.79	73.03	82.92	80.74	97.77	80.85
Classic ST (Uniform)	84.81	78.97	89.92	84.62	98.39	87.34
UST (Easy)	88.19	89.21	<b>91.27</b>	<b>87.74</b>	<b>98.57</b>	<b>91.00</b>
- removing <i>Class</i>	87.33	87.22	89.18	86.88	98.27	89.78
- removing <i>Conf</i>	86.73	<b>90.00</b>	90.40	84.17	98.49	89.96
UST (Hard)	88.02	88.49	90.00	85.02	98.56	90.02
- removing <i>Class</i>	80.45	89.28	90.07	83.07	98.46	88.27
- removing <i>Conf</i>	<b>88.48</b>	87.93	88.74	84.45	98.26	89.57

of both implicit reduction by sample selection and explicit reduction by accounting for the sample variance for confident learning. This is demonstrated in a consistent performance of the model across different runs with an aggregated (least) standard deviation of 0.57 across different runs of the model for different tasks with different random seeds. UDA with its consistency learning closely follows suit with an aggregated standard deviation of 1.62 across different runs for different tasks. Classic self-training without any such mechanism shows high variance in performance across runs with different seeds. In Table 4, we show the results from other works on these datasets as reported in [Li and Ye, 2018, Jo and Cinarel, 2019, Li and Sethy, 2019, Gururangan et al., 2019]<sup>2</sup>. Our UST framework outperforms them while using much less training labels per class (shown by K).

通过样本选择进行隐式归约和通过考虑样本方差进行显式归约，实现自信学习。这体现在模型在不同试验中的一致性性能上，对于具有不同随机种子的不同任务，模型在不同试验中的聚合（最小）标准差为 0.57。UDA 的一致性学习与不同任务的不同运行中 1.62 的聚合标准差密切相关。没有任何此类机制的经典自训练显示不同种子试验的性能差异很大。在表 4 中，我们展示了有关这些数据集的其他工作结果，如 [Li and Ye, 2018, Jo and Cinarel, 2019, Li and Sethy, 2019, Gururangan et al., 2019]<sup>2</sup>。我们的 UST 框架性能优于它们，但每个类使用的训练标签要少得多（K 表示）。

Ablation analysis. We compare the impact of different components of our model for self-training with 30 labeled examples per class for each task for training and for validation with results in Table 3. Sampling strategies. The backbone of the sample selection method in our self-training framework is given by the BALD measure [Houlsby et al., 2011] that has been shown to outperform other active sampling strategies leveraging measures like entropy and variation ratios in Gal et al. [2017] for image classification. We use this measure in our framework to sample examples based on whether the model is confused about the example or not by leveraging sampling strategies in Equations 7 or 6 and optimized by self-training with Equation 11 – denoted by UST (Hard) and UST (Easy) respectively in Table 3. In contrast to works in active learning that find hard examples to be more informative than easy ones for manual labeling, in the self-training framework we observe the reverse with hard examples often contributing noisy pseudo-labels. We compare this with uniform sampling in the classic ST framework, and observe that sample selection bias (easy or hard) benefits self-training.

消融分析。我们将模型中不同组件对自我训练的影响与每个训练任务和验证任务的 30 个标记示例与表 3 中的结果进行比较。抽样策略。在我们的自我训练框架中，样本选择方法的支柱来自 BALD 测量 [Houlsby 等人，2011 年]，该测量已被证明优于其他主动采样策略，利用 Gal 等人的和变异比等测量。[2017] 用于图像分类。我们在我们的框架中使用此度量来采样示例，基于模型是否对示例感到困惑，利用公式 7 或 6 中的采样策略，并通过使用公式 11 进行自我训练进行优化 – 分别在表 3 中表示为 UST（硬）和 UST（易）。与主动学习中发现硬例子比简单的手工标签更有信息的工作相反，在自我训练框架中，我们观察到硬例子往往造成杂的伪标签。我们将其与经典 ST 框架中的均匀采样进行比较，并观察到样本选择偏差（容易或困难）有利于自我训练。

Class-dependent selection with exploration. In this, we remove the class-dependent selection and exploration with global selection of samples based on their easiness or hardness for the corresponding UST sampling strategy. Class-dependent selection ameliorates model bias towards picking samples from a specific class that might be too easy or hard to learn from with balanced selection of samples across all the classes, and improves our model on aggregate.

基于类别的选择与探索。在此过程中，我们根据样品的易操作性或硬度为相应的 UST 采样策略消除了依类别选择和探索的可能性。基于类别的选择通过在所有类别中均衡选择样本，改善了从特定类别中挑选样本的模型偏差，这可能太容易或难以学习，并改进了我们的聚合模型。

Confident learning. In this, we remove confident learning from the UST framework. Therefore, we optimize the unlabeled data loss for self-training using Equation 4 instead of Equation 11 that is used in all other UST strategies. This component helps the student to focus more on examples the teacher is confident about corresponding to low-variance ones, and improves the model on aggregate.

自信学习。在这一点上，我们从 UST 框架中删除了自信的学习。因此，我们使用公式 4 而不是所有其他 UST 策略中使用的公式 11 来优化未标记的数据丢失。该组件帮助学生更专注于教师有信心对应于低方差示例的示例，并改进了聚合模型。

Overall, we observe that each of the above uncertainty-based sample selection and learning strategies outperform the classic self-training mechanism selecting samples uniform at random.

总的来说，我们观察到，上述基于不确定性的样本选择和学习策略都优于经典的随机选择样本均匀的自训练机制。

Impact of K labeled examples. From Figure 2a, we observe the self-training accuracy to gradually improve with increase in the number of labeled examples per class to train the base teacher model leading to better initialization of the self-training process. With only 20 labeled examples for each task for training and for validation, we observe the aggregate performance across five tasks to be 89.27% with further improvements with more labeled data coming from IMDB and AG news datasets.

K 标记例子的影响。从图 2a 中，我们观察到，随着每个班级标记示例数量的增加，自我训练的准确性逐渐提高，以训练基础教师模型，从而更好地初始化自我训练过程。由于每个任务仅有 20 个用于训练和验证的标记示例，我们观察到五个任务的汇总性能为 89.27%，并进一步改进了来自 IMDB 和 AG 新闻数据集的更多标记数据。

Impact self-training iterations. Figure 2b shows increase in self-training accuracy of UST over iterations for a single run. In general, we observe the self-training performance to improve rapidly initially, and gradually converge in 15-20 iterations. We also observe some models to drift a bit while continuing the self-training process and similar for consistency learning in UDA beyond a certain point. This necessitates the use of the validation set for early termination based on validation loss.

影响自我训练迭代。图 2b 显示了单次试验中 UST 自训练精度的提高。一般情况下，我们观察到自我训练性能最初会迅速提高，并在 15 - 20 次迭代中逐渐收敛。我们还观察到一些模型在继续自我训练过程时有点偏离，类似于 UDA 中的一致性学习超出了某一点。这就需要使用基于验证损失的提前终止验证集。

<sup>2</sup>Note that these models use different encoders and pre-training mechanisms.

2 请注意，这些模型使用不同的编码器和预训练机制。

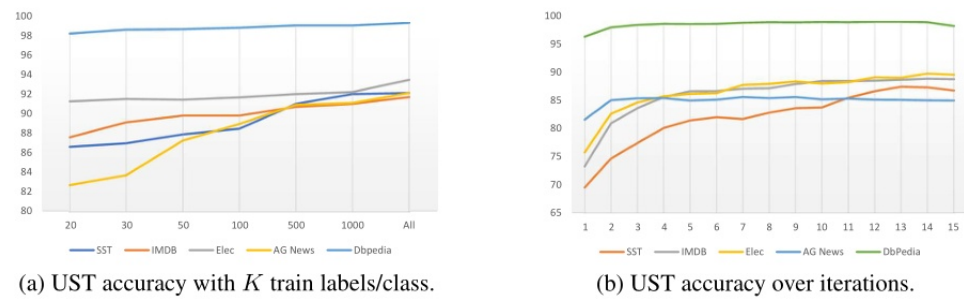


Figure 2: Improvement in UST accuracy with more training labels and epochs.  
图 2：使用更多训练标签和迭代周期提高 UST 精度。

Table 4: SSL methods with K training labels per class (Adv: Adversarial, Parti: Partitioning, Temp: Temporal) <sup>2</sup>. UST performs better than competing methods while using less training labels.

表 4：每个类带 K + 训练标签的 SSL 方法（Adv：对抗性，Parti：分区，Temp：时间性）<sup>2</sup>。UST 在使用较少的训练标签时比竞争方法表现得更好。

Datasets	Model	K Labels	Acc.
IMDB	<b>UST (ours)</b>	<b>30</b>	<b>89.2</b>
	Variational Pre-training	200	82.2
	Reinforcement + Adv. Training	100	82.1
	SeqSSL + Self-training	100	79.7
	SeqSSL	100	77.4
	Layer Parti. + Temp. Ensembling	100	75.9
	SeqSSL + Adv. Training	100	75.7
	Delta-training	212	75.0
	Layer Parti. + II Model	100	69.3
AG News	<b>UST (ours)</b>	<b>30</b>	<b>87.7</b>
	Variational Pre-training	200	83.9
	Reinforcement + Adv. Training	100	81.7
	SeqSSL + Self-training	100	78.5
	SeqSSL	100	76.2
	SeqSSL + Adv. Training	100	73.0
DBpedia	<b>UST (ours)</b>	<b>30</b>	<b>98.6</b>
	Reinforcement + Adv. Training	100	98.5
	SeqSSL + Self-training	100	98.1
	SeqSSL + Adv. Training	100	96.7
	SeqSSL	100	96.1

## 5 Related Work

### 5 相关工作

Semi-supervised learning has been widely used in many different flavors including consistency training [Bachman et al., 2014, Rasmus et al., 2015, Laine and Aila, 2017, Tarvainen and Valpola, 2017], latent variable models [Kingma et al., 2014] for sentence compression [Miao and Blunsom, 2016] and code generation [Yin et al., 2018]. More recently, consistency-based model like UDA Xie et al. [2019] has shown promising results for few-shot learning for classification leveraging auxiliary resources like paraphrasing and back-translation (BT) [Sennrich et al., 2016].

半监督式学习已广泛应用于许多不同的口味，包括一致性训练 [Bachman 等人，2014 年，Rasmus 等人，2015 年，Laine 和 Aila，2017 年，Tarvainen 和 Valpola，2017 年]，用于句子压缩的潜在变量模型 [Kingma 等人，2014 年] [Miao 和 Blunsom，2016 年] 和代码生成 [Yin 等人，2018 年]。最近，基于一致性的模型，如 UDA Xie 等。[2019] 在利用转述和反向翻译（BT）等辅助资源进行分类的少量学习方面取得了可喜的成果 [Sennrich et al.，2016]。

Sample selection. One of the earlier works in neural networks leveraging easiness of the samples for learning is given by curriculum learning [Bengio et al., 2009]. This is based on the idea of learning easier aspects of the task first followed by the more complex ones. However, the main challenge is the identification of easy and hard samples in absence of external knowledge. Prior work leveraging self-paced learning [Kumar et al., 2010] and more recently self-paced co-training [Ma et al., 2017] leverage teacher confidence (or lower model loss) to select easy samples during training. In a similar flavor, some recent works have also focused on sample selection for self-training leveraging meta-learning [Li et al., 2019] and active learning [Panagioti Mastoropoulou, 2019, Chang et al., 2017] based on teacher confidence. However, all of these techniques rely on only the teacher confidence while ignoring the uncertainty associated with its predictions. There are also works on anti-curriculum learning (or hard example mining) [Shrivastava et al., 2016] that leverage hardness of the samples.

样品选择 早期的神经网络利用样本的易用性进行学习的工作之一是课程学习 [Bengio et al.，2009]。这是基于学习任务中更简单的方面，然后是更复杂的方面的想法。然而，主要挑战是在缺乏外部知识的情况下识别容易和坚硬的样品。以前的工作利用自定进度学习 [Kumar et al.，2010] 和最近的自定进度联合培训 [Ma et al.，2017] 利用教师信心（或降低模型损失）在培训期间选择简单样本。在类似的风格中，最近的一些作品也侧重于利用金属学习 [Li 等人，2019 年] 和基于教师信心的主动学习 [Panagioti Mastoropoulou，2019 年，Chang 等人，2017 年] 进行自我训练的样本选择。然而，所有这些技术都只依赖于教师的信心，而忽略了与预测相关的不确定性。还有一些关于反课程学习（或硬例子挖掘）的工作（Shrivastava 等人，2016 年），以利用样品的硬度。

Uncertainty in neural networks. A principled mechanism to generate uncertainty estimates is provided by Bayesian frameworks. A Bayesian neural network Gal and Ghahramani [2016] replaces神经网络的不确定性。贝叶斯框架提供了生成不确定性估计的原则机制。贝叶斯神经网络 Gal 和 Ghahramani [2016 年] 取代



a deterministic model’s weight parameters with distributions over model parameters. Parameter optimization is replaced by marginalisation over all possible weights. It is difficult to perform inference over BNN’s as the marginal distribution cannot be computed analytically, and we have to resort to approximations such as variational inference to optimize for variational lower bound [Graves, 2011, Blundell et al., 2015, Hernández-Lobato et al., 2016, Gal and Ghahramani, 2015].

具有模型参数分布的确定性模型的权重参数。参数优化被所有可能权重的边际化所取代。由于无法分析计算边缘分布，因此很难对 BNN 进行推理，我们必须求助于变分推理等近似值来优化变分下界 [Graves, 2011, Blundell et al., 2015, Hernández-Lobato et al., 2016, Gal and Ghahramani, 2015]。

## 6 Conclusions

### 6 结论

In this work we developed an uncertainty-aware framework to improve self-training mechanism by exploiting uncertainty estimates of the underlying neural network. We particularly focused on better sample selection from the unlabeled pool based on posterior entropy and confident learning to emphasize on low variance samples for self-training. As application, we focused on task-specific fine-tuning of pre-trained language models with few labels for text classification on five benchmark datasets. With only 20-30 labeled examples and large amounts of unlabeled data, our models perform close to fully supervised ones fine-tuned on thousands of labeled examples. While pre-trained language models are natural few-shot learners, we show their performance can be improved by upto 8% by classic self-training and by upto 12% on incorporating uncertainty estimates in the framework.

在这项工作中，我们开发了一个不确定性感知框架，通过利用底层神经网络的不确定性估计来改进自我训练机制。我们特别注重基于后和自信学习从未标记池中更好地选择样本，以强调用于自我训练的低方差样本。作为应用，我们专注于针对特定任务对预训练语言模型进行微调，在五个基准数据集上只有很少的文本分类标签。仅有 20 - 30 个标记示例和大量未标记数据，我们的模型在数千个标记示例上进行微调，接近完全监督的模型。虽然预训练的语言模型是自然的少数镜头学习者，但我们表明，通过传统的自我训练，他们的性能可以提高高达 8%，而在框架中纳入不确定性估计则可以提高 12%。

## 7 Appendix

### 7 附录

#### 7.1 Pseudo-code

##### 7.1 伪代码

**Algorithm 1:** Uncertainty-aware self-training pseudo-code.

---

Continue pre-training teacher language model on task-specific unlabeled data  $D_u$  ;  
 Fine-tune model  $f^W$  with parameters  $W$  on task-specific small labeled data  $D_l$  ;  
**while** *not converged* **do**  
     Randomly sample  $S_u$  unlabeled examples from  $D_u$  ;  
     **for**  $x \in S_u$  **do**  
         **for**  $t \leftarrow 1$  **to**  $T$  **do**  
              $W_t \sim Dropout(W)$  ;  
              $y_t^* = softmax(f^{W_t}(x))$ ;  
         **end**  
         Compute predictive sample mean  $E(y)$  and predictive sample variance  $Var(y)$  with Equation 11 ;  
         Compute BALD acquisition function with Equation 6 ;  
     **end**  
     Sample  $R$  instances from  $S_u$  employing sample selection with Equations 7 or 8 ;  
     Pseudo-label  $R$  sampled instances with model  $f^W$  ;  
     Re-train model on  $R$  pseudo-labeled instances with Equation 12 and update parameters  $W$  ;  
**end**

Teacher-student training: In our experiments, we employ a single model for self-training. Essentially, we copy teacher model parameters to use as the student model and continue self-training. Although, some works initialize the student model from scratch.

师生培训：在我们的实验中，我们采用单一模型进行自我培训。本质上，我们复制教师模型参数用作学生模型并继续自我训练。不过，有些作品从头开始初始化学生模型。

Sample size. Ideally, we need to perform  $T$  stochastic forward passes for each sample in the large unlabeled pool. However, this is too slow. For computational efficiency, at each self-training iteration, we select  $S_u$  samples randomly from the unlabeled set, and then select  $R \in S_u$  samples from therein based on uncertainty using several stochastic forward passes.

样本大小。理想情况下，我们需要为大型未标记池中的每个样本执行  $T$  随机前向传递。然而，这太慢了。为了提高计算效率，在每次自训练迭代中，我们从未标记集中随机选择  $S_u$  样本，然后使用几个随机前向通道根据不确定性从其中选择  $R \in S_u$  样本。

#### 7.2 Hyper-parameters

##### 7.2 超参数

We do not perform any hyper-parameter tuning for different datasets and use the same set of hyper-parameters as shown in Table 5.

我们不对不同的数据集执行任何超参数调整，而是使用相同的超参数集，如表 5 所示。

Also, we retain parameters from original BERT implementation from <https://github.com/google-research/bert>.

此外，我们从 <https://github.com/google-research/bert> 保留原始 BERT 实现的参数。

Table 5: Hyper-parameters  
表 5 : 超参数

Dataset	Sequence Length
SST-2	32
AG News	80
DBpedia	90
Elec	128
IMDB	256

Sample size for selecting $S_u$ samples from unlabeled pool for forward passes in each self-training iteration	16384
Sample size for selecting $R$ samples from $S_u$ for each self-training iteration	4096
Batch size for fine-tuning base model on small labeled data	4
Batch size for self-training on $R$ selected samples	32
T	30
Softmax dropout	0.5
BERT attention dropout	0.3
BERT hidden dropout	0.3
BERT output hidden size $h$	768
Epochs for fine-tuning base model on labeled data	50
Epochs for self-training model on unlabeled data	25
Iterations for self-training	25

Table 6: UDA batch size.  
表 6 : UDA 批量大小。

Dataset	Batch size
SST-2	32
AG News	32
DBpedia	32
Elec	16
IMDB	8

UDA configuration. Similar to all other models, we add validation data to UDA to select the best model parameters based on validation loss. We retain all UDA hyper-parameters from <https://github.com/google-research/uda>. We use the same sequence length for every task as in our models and select the batch size as in Table 6.

UDA 配置。与所有其他模型类似，我们向 UDA 添加验证数据，以根据验证损失选择最佳模型参数。我们保留来自 <https://github.com/google-research/uda> 的所有 UDA 超参数。我们对每个任务使用与模型相同的序列长度，并选择表 6 中的批量大小。

Note that our UDA results are worse than that reported in the original implementation due to different sequence length and batch sizes for hardware constraints. We select the maximum batch size permissible by the V100 gpu memory constraints given the sequence length.

请注意，由于硬件约束的序列长度和批量大小不同，我们的 UDA 结果比原始实现中报告的结果差。我们选择给定序列长度的 V100 GPU 内存约束所允许的最大批处理大小。

References

参考资料

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 4171–4186, 2019. URL <https://aclweb.org/anthology/papers/N19/N19-1423/>.

Jacob Devlin, Ming-Wei Chang, Kenton Lee 和 Kristina Toutanova。BERT：用于语言理解的深度双向变压器的预训练。计算语言学协会北美分会 2019 年会议纪要：人类语言技术，NAACL-HLT 2019，明尼阿波利斯，明尼苏达州，美国，2019 年 6 月 2-7 日，第 1 卷（长篇和短篇论文），第 4171-4186 页，2019 年。网址：<https://aclweb.org/anthology/papers/N19/N19-1423/>。

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei 和 Ilya Sutskever。语言模型是不受监督的多任务学习者。2019 年。

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. CoRR, abs/1907.11692, 2019. URL <http://arxiv.org/abs/1907.11692>.

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning. 2010.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training, 2019.

H. J. Scudder III. Probability of error of some adaptive pattern-recognition machines. IEEE Trans. Inf. Theory, 11(3):363–371, 1965. doi: 10.1109/TIT.1965.1053799. URL <https://doi.org/10.1109/TIT.1965.1053799>.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc’Aurelio Ranzato. Revisiting self-training for neural sequence generation, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, 2017. URL <https://openreview.net/forum?id=Sy8gdB9xx>.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In Andrea Pohoreckyj Danyluk, Léon Bottou, and Michael L. Littman, editors, Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009, volume 382 of ACM International Conference Proceeding Series, pages 41–48. ACM, 2009. doi: 10.1145/1553374.1553380. URL <https://doi.org/10.1145/1553374.1553380>.

M. P. Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, Advances in Neural Information Processing Systems 23, pages 1189–1197. Curran Associates, Inc., 2010. URL <http://papers.nips.cc/paper/3923-self-paced-learning-for-latent-variable-models.pdf>.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016, volume 48 of JMLR Workshop and Conference Proceedings, pages 1050–1059. JMLR.org, 2016. URL <http://proceedings.mlr.press/v48/gal16.html>.

Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. CoRR, abs/1506.02158, 2015. URL <http://arxiv.org/abs/1506.02158>.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res., 15(1): 1929–1958, 2014. URL <http://dl.acm.org/citation.cfm?id=2670313>.

Claude E. Shannon. A mathematical theory of communication. ACM SIGMOBILE Mob. Comput. Commun. Rev., 5(1):3–55, 2001. doi: 10.1145/584091.584093. URL <https://doi.org/10.1145/584091.584093>.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer 和 Veselin Stoyanov。Roberta：稳健优化的 BERT 预训练方法。CoRR, abs/1907.11692, 2019。网址：<http://arxiv.org/abs/1907.11692>。Olivier Chapelle, Bernhard Scholkopf 和 Alexander Zien。半监督学习。2010 年。谢其哲, 戴子行, 爱德华霍维, 明唐卢, 和 Quoc V. Le。用于一致性训练的无监督数据增强, 2019 年。H. J. Scudder III。一些自适应模式识别机器的错误概率。IEEE Trans. inf. 理论, 11 (3) : 363 – 371, 1965。doi : 10.1109 / TIT.1965 . 1053799。网址：<https://doi.org/10.1109/TIT.1965.1053799>。何俊贤, 顾嘉涛, 沈佳君, 和 Marc ' Aurelio Ranzato。回顾神经序列生成的自我训练, 2019 年。Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht 和 Oriol Vinyals。理解深度学习需要重新思考一般化。2017 年 4 月 24 日至 26 日, 在法国土伦举行的第五届学习代表国际会议上, 会议记录。OpenReview.net, 2017。网址：<https://openreview.net/forum?id=Sy8gdB9xx>。Yoshua Bengio, Jérôme Louradour, Ronan Collobert 和 Jason Weston。课程学习。Andrea Pohoreckyj Danyluk, Léon Bottou, 和 Michael L. Littman, 编辑, Proceedings of the 26 Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, 2009 年 6 月 14 – 18 日, ACM International Conference Proceeding ACM, 2009。doi : 10.1145 / 1553374.1553380。网址：<https://doi.org/10.1145/1553374.1553380>。M. P. Kumar, Benjamin Packer 和 Daphne Koller。针对潜在变量模型的自定进度学习。在 J. D. Lafferty, C. K. I. Williams, J. Shawe – Taylor, R. S. Zemel 和 A. Culotta, 编辑, 神经信息处理系统的进展 23, 第 1189 – 1197 页。Curran Associates, Inc., 2010 年。网址：<http://papers.nips.cc/paper/3923> Yarin Gal 和 Zoubin Ghahramani。辍学作为贝叶斯近似: 表示深度学习中的模型不确定性。在 Maria – Florina Balcan 和 Kilian Q. Weinberger, 编辑, 第 33 届机器学习国际会议论文集, ICML 2016, 纽约, 美国, 2016 年 6 月 19 – 24 日, JMLR 研讨会和会议论文集第 48 卷, 第 1050 – 1059 页。JMLR.org, 2016 年。网址：<http://procedures.mlr.press/v48/gal16.html>。Yarin Gal 和 Zoubin Ghahramani。贝叶斯卷积神经网络与伯努利近似变分推理。CoRR, abs/1506.02158, 2015。网址 <http://arxiv.org/abs/1506.02158>。Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever 和 Ruslan Salakhutdinov。Dropout：防止神经网络过拟合的简单方法。J. Mach. 学习. Res., 15 ( 1 ) : 1929 – 1958, 2014。网址 <http://dl.acm.org/citation.cfm?id=2670313>。Claude E. Shannon。通信的数学理论。ACM SIGMOBILE 手机 计算. 社区. 修士, 5 ( 1 ) : 3 – 55, 2001。doi : 10.1145 / 584091.584093。网址：<https://doi.org/10.1145/584091.584093>。



Linton G Freeman. Elementary applied statistics . 1965.

Neil Houlsby, Ferenc Huszar, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. CoRR, abs/1112.5745, 2011. URL <http://arxiv.org/abs/1112.5745> .

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In Doina Precup and Yee Whye Teh, editors, Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, volume 70 of Proceedings of Machine Learning Research , pages 1183–1192. PMLR, 2017. URL <http://proceedings.mlr.press/v70/gal17a.html> .

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL, pages 1631–1642. ACL, 2013. URL <https://www.aclweb.org/anthology/D13-1170/> .

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 2011, Portland, Oregon, USA, pages 142–150, 2011. URL <http://www.aclweb.org/anthology/P11-1015> .

Julian J. McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In Seventh ACM Conference on Recommender Systems, RecSys ’13, Hong Kong, China, October 12-16, 2013, pages 165–172, 2013. doi: 10.1145/2507157.2507163. URL <https://doi.org/10.1145/2507157.2507163> .

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada , pages 649–657, 2015. URL <http://papers.nips.cc/paper/5782-character-level-convolutional-networks-for-text-classification> .

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 2015. URL <http://arxiv.org/abs/1412.6980> .

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. Adversarial training methods for semi-supervised text classification. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, 2017. URL [https://openreview.net/forum?id=r1X3g2\\_xl](https://openreview.net/forum?id=r1X3g2_xl) .

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings . OpenReview.net, 2017. URL <https://openreview.net/forum?id=BJ6oOfqge> .

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, pages 1163–1171, 2016.

Andrew M. Dai and Quoc V. Le. Semi-supervised sequence learning. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada , pages 3079–3087, 2015. URL <http://papers.nips.cc/paper/5949-semi-supervised-sequence-learning> .

Neil Houlsby , Ferenc Huszar , Zoubin Ghahramani 和 Máté Lengyel 。 贝叶斯主动学习用于分类和偏好学习。 CoRR , abs / 1112.5745 , 2011 。 网址 <http://arxiv.org/abs/1112.5745> 。 Yarin Gal 、 Riashat Islam 和 Zoubin Ghahramani 。 利用图像数据进行深度贝叶斯主动学习。 在 Doina Precup 和 Yee Whye Teh , 编辑,第 34 届机器学习国际会议论文集, ICML 2017 , 悉尼,新南威尔士州,澳大利亚, 2017 年 8 月 6 - 11 日,机器学习研究论文集第 70 卷,第 1183 - 1192 页。 PMLR , 2017 年。 网址: <http://proceedings.mlr.press/v70/gal17a.html> 。 Richard Socher 、 Alex Perelygin 、 Jean Wu 、 Jason Chuang 、 Christopher D. Manning 、 Andrew Y. Ng 和 Christopher Potts 。 针对情绪树库的语义复合性的递归深度模型。 在 2013 年自然语言处理经验方法会议论文集, EMNLP 2013 年, 2013 年 10 月 18 日至 21 日,西雅图君悦酒店,西雅图,华盛顿,美国, SIGDAT 会议, ACL 的一个特殊兴趣小组,第 1631 - 1642 页。 ACL , 2013 年。 网址: <https://www.aclweb.org/anthology/D13-1170/> 。 Andrew L. Maas 、 Raymond E. Daly 、 Peter T. Pham 、 Dan Huang 、 Andrew Y. Ng 和 Christopher Potts 。 用于情绪分析的学习词向量。 在计算语言学协会第 49 届年会:人类语言技术,会议论文集, 2011 年,美国俄勒冈州波特兰,页 142 - 150 , 2011 年。 网址: <http://www.aclweb.org/anthology/P11-1015> 。 Julian J. McAuley 和 Jure Leskovec 。 隐藏因素和隐藏主题:通过评论文本了解评级维度。 在第七届 ACM 会议上推荐系统, RecSys ' 13 , 香港,中国,2013 年 10 月 12 - 16 日,页 165 - 172 , 2013 。 doi : 10.1145 / 2507157.2507163 。 网址: <https://doi.org/10.1145/2507157.2507163> 。 张翔 , Junbo Jake Zhao , Yann LeCun 。 用于文本分类的字符级卷积网络。 神经信息处理系统的进展 28 : 2015 年神经信息处理系统年度会议, 2015 年 12 月 7 日至 12 日,加拿大魁北克省蒙特利尔,第 649 - 657 页, 2015 年。 网址: <http://papers.nips.cc/paper/5782-character-level-convolutional-networks-for-text-classification> 。 Diederik P. Kingma 和 Jimmy Ba 。 Adam : 随机优化的方法。 在第三届国际会议上学习表示, ICLR 2015 , 圣地亚哥, CA , 美国, 2015 。 网址 <http://arxiv.org/abs/1412.6980> 。 Takeru Miyato , Andrew M. Dai 和 Ian J. Goodfellow 。 半监督文本分类的对抗式训练方法。 2017 年 4 月 24 日至 26 日, 2017 年 4 月 24 日至 26 日在法国土伦举行的第五届国际学习代表会议。 网址 [https://openreview.net/forum?id=r1X3g2\\_xl](https://openreview.net/forum?id=r1X3g2_xl) 。 Samuli Laine 和 Timo Aila 。 半监督学习的时间集合。 2017 年 4 月 24 日至 26 日,法国土伦, 2017 年 ICLR 第 5 届国际学习代表会议,会议记录。 OpenReview.net , 2017 。 网址: <https://openreview.net/forum?id=BJ6oOfqge> 。 Mehdi Sajjadi 、 Mehran Javanmardi 和 Tolga Tasdizen 。 用于深度半监督学习的随机变换和扰动正则化。 在 Daniel D. Lee 、 Masashi Sugiyama 、 Ulrike von Luxburg 、 Isabelle Guyon 和 Roman Garnett , 编辑,《神经信息处理系统的进展 29 : 神经信息处理系统年度会议 2016 》, 2016 年 12 月 5 日至 10 日,西班牙巴塞罗那,第 1163 - 1171 页, 2016 年。 Andrew M. Dai 和 Quoc V. Le 。 半监督序列学习。 在 Corinna Cortes , Neil D. Lawrence , Daniel D. Lee , Masashi Sugiyama 和 Roman Garnett , 编辑,《神经信息处理系统的进展 28 : 2015 年神经信息处理系统年度会议》, 2015 年 12 月 7 - 12 日,加拿大魁北克省蒙特利尔,页 3079 - 3087 , 2015 年。 网址: <http://papers.nips.cc/paper/5949-semi-supervised-sequence-learning> 。

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. URL <http://arxiv.org/abs/1412.6572>.

伊恩·J·古德瑞特, 乔纳森·施伦斯, 和克里斯蒂安·塞格迪。解释和利用对抗性例子。在 Yoshua Bengio 和 Yann LeCun, 编辑, 第三届国际学习代表会议, ICLR 2015, 圣地亚哥, 加利福尼亚州, 美国, 2015 年 5 月 7 日至 9 日, 会议记录, 2015 年。网址 <http://arxiv.org/abs/1412.6572>。

Suchin Gururangan, Tam Dang, Dallas Card, and Noah A. Smith. Variational pretraining for semi-supervised text classification. ACL, 2019.

Yan Li and Jieping Ye. Learning adversarial networks for semi-supervised text classification via policy gradient. In Yike Guo and Faisal Farooq, editors, Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018, pages 1715–1723. ACM, 2018. doi: 10.1145/3219819.3219956. URL <https://doi.org/10.1145/3219819.3219956>.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings. OpenReview.net, 2017. URL <https://openreview.net/forum?id=ry8u21rtl>.

Alexander Hanbo Li and Abhinav Sethy. Semi-supervised learning for text classification by layer partitioning. CoRR, abs/1911.11756, 2019. URL <http://arxiv.org/abs/1911.11756>.

Hwiyeol Jo and Ceyda Cinarel. Delta-training: Simple semi-supervised text classification using pretrained word embeddings. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pages 3456–3461. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1347. URL <https://doi.org/10.18653/v1/D19-1347>.

Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, pages 3365–3373, 2014. URL <http://papers.nips.cc/paper/5487-learning-with-pseudo-ensembles>.

Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, pages 3546–3554, 2015. URL <http://papers.nips.cc/paper/5947-semi-supervised-learning-with-ladder-networks>.

Diederik P. Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, pages 3581–3589, 2014. URL <http://papers.nips.cc/paper/5352-semi-supervised-learning-with-deep-generative-models>.

Yishu Miao and Phil Blunsom. Language as a latent variable: Discrete generative models for sentence compression. In Jian Su, Xavier Carreras, and Kevin Duh, editors, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016, pages 319–328. The Association for Computational Linguistics, 2016. doi: 10.18653/v1/d16-1031. URL <https://doi.org/10.18653/v1/d16-1031>.

Pengcheng Yin, Chunting Zhou, Junxian He, and Graham Neubig. Structvae: Tree-structured latent variable models for semi-supervised semantic parsing. In Iryna Gurevych and Yusuke Miyao, editors, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers, pages 754–765. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1070. URL <https://www.aclweb.org/anthology/P18-1070/>.

Suchin Gururangan、Tam Dang、达拉斯卡和 Noah A. Smith。半监督文本分类的变量预训练。ACL, 2019 年。李燕和叶洁平。通过策略梯度学习用于半监督文本分类的对抗网络。郭毅和费萨尔·法鲁克, 编辑, 第 24 届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集, KDD 2018, 英国伦敦, 2018 年 8 月 19 - 23 日, 第 1715 - 1723 页。ACM, 2018。doi: 10.1145 / 3219819.3219956。网址: <https://doi.org/10.1145/3219819.3219956>。Antti Tarvainen 和 Harri Valpola。平均教师是更好的榜样: 体重平均一致性目标改善半监督的深度学习结果。2017 年 4 月 24 - 26 日, 在法国土伦举行的第五届学习表达国际会议上, 研讨会跟踪论文集。OpenReview.net, 2017 年。网址 <https://openreview.net/forum?id=ry8u21rtl>。Alexander Hanbo Li 和 Abhinav Sethy。通过图层分区进行文本分类的半监督学习。CoRR, abs/1911.11756, 2019。网址 <http://arxiv.org/abs/1911.11756>。Hwiyeol Jo 和 Ceyda Cinarel。Delta 训练: 使用预训练词嵌入的简单半监督文本分类。2019 年 11 月 3 - 7 日, 中国香港, 2019 年自然语言处理经验方法会议暨第九届自然语言处理国际联合会议论文集 (EMNLP - IJCNLP 2019) 编辑昆太郎、江景、吴文强、万晓俊。计算语言学协会, 2019。doi: 10.18653/v1/D19-1347。网址: <https://doi.org/10.18653/v1/D19-1347>。Philip Bachman, Ouais Alsharif 和 Doina Precup。学习伪集成。在 Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence 和 Kilian Q. Weinberger, 编辑, 《神经信息处理系统的进展 27: 2014 年神经信息处理系统年度会议》, 2014 年 12 月 8 日至 13 日, 加拿大魁北克省蒙特利尔, 第 3365 - 3373 页, 2014 网址: <http://papers.nips.cc/paper/5487-learning-with-pseudo-ensembles>。Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola 和 Tapani Raiko。利用阶梯网络进行半监督式学习。在 Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama 和 Roman Garnett, 编辑, 《神经信息处理系统的进步 28: 2015 年神经信息处理系统年度会议》, 2015 年 12 月 7 - 12 日, 加拿大魁北克省蒙特利尔, 第 3546 - 3554 页, 2015 年。网址: <http://papers.nips.cc/paper/5947-semi-supervised-learning-with-ladder-networks>。Diederik P. Kingma, Shakir Mohamed, Danilo Jimenez Rezende 和 Max Welling。具有深度生成模型的半监督学习。在 Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence 和 Kilian Q. Weinberger, 编辑, 《神经信息处理系统的进展 27: 2014 年神经信息处理系统年度会议》, 2014 年 12 月 8 日至 13 日, 加拿大魁北克省蒙特利尔, 第 3581 - 3589 页, 2014 网址: <http://papers.nips.cc/paper/5352-semi-supervised-learning-with-deep-generative-models>。Yishu Miao 和 Phil Blunsom。语言作为潜在变量: 用于句子压缩的离散生成模型。Jian Su, Xavier Carreras, and Kevin Duh, 编辑, 2016 年自然语言处理经验方法会议论文集, EMNLP 2016, 美国德克萨斯州奥斯汀, 2016 年 11 月 1 - 4 日, 第 319 - 328 页。计算语言学协会, 2016 年。doi: 10.18653/v1/d16-1031。网址: <https://doi.org/10.18653/v1/d16-1031>。Pengcheng Yin, Chunting Zhou, Junxian He 和 Graham Neubig。Structvae: 用于半监督语义解析的树结构潜在变量模型。Iryna Gurevych 和 Yusuke Miyao, 编辑, 计算语言学协会第 56 届年会论文集, ACL 2018, 澳大利亚墨尔本, 2018 年 7 月 15 - 20 日, 第 1 卷: 长篇小说, 第 754 - 765 页。计算语言学协会, 2018。doi: 10.18653/v1/P18-1070。网址: <https://www.aclweb.org/anthology/P18-1070/>。

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers. The Association for Computer Linguistics, 2016. doi: 10.18653/v1/p16-1009. URL <https://doi.org/10.18653/v1/p16-1009>.

Rico Sennrich, Barry Haddow 和 Alexandra Birch。使用单语数据改进神经机器翻译模型。在计算语言学协会第 54 届年会的会议记录中, ACL 2016, 2016 年 8 月 7 - 12 日, 德国柏林, 第 1 卷: 长篇论文。计算语言学协会, 2016 年。doi: 10.18653/v1/p16-1009。网址: <https://doi.org/10.18653/v1/p16-1009>。

Fan Ma, Deyu Meng, Qi Xie, Zina Li, and Xuanyi Dong. Self-paced co-training. In Doina Precup and Yee Whye Teh, editors, Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pages 2275–2284, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL <http://proceedings.mlr.press/v70/ma17b.html>.

Xinzhe Li, Qianru Sun, Yaoyao Liu, Qin Zhou, Shibao Zheng, Tat-Seng Chua, and Bernt Schiele. Learning to self-train for semi-supervised few-shot classification. In Advances in Neural Information Processing Systems 32, pages 10276–10286. Curran Associates, Inc., 2019. URL <http://papers.nips.cc/paper/>

9216-learning-to-self-train-for-semi-supervised-few-shot-classification.pdf

Emmeleia Panagiota Mastoropoulou. Enhancing deep active learning using selective self-training for image classification. Master’s thesis, KTH, School of Electrical Engineering and Computer Science (EECS), 2019.

Haw-Shiuan Chang, Erik G. Learned-Miller, and Andrew McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, pages 1002–1012, 2017.

Abhinav Shrivastava, Abhinav Gupta, and Ross B. Girshick. Training region-based object detectors with online hard example mining. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pages 761–769. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.89. URL <https://doi.org/10.1109/CVPR.2016.89>.

Alex Graves. Practical variational inference for neural networks. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain, pages 2348–2356, 2011. URL <http://papers.nips.cc/paper/4329-practical-variational-inference-for-neural-networks>.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In Francis R. Bach and David M. Blei, editors, Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, volume 37 of JMLR Workshop and Conference Proceedings, pages 1613–1622. JMLR.org, 2015. URL <http://proceedings.mlr.press/v37/blundell15.html>.

José Miguel Hernández-Lobato, Yingzhen Li, Mark Rowland, Thang D. Bui, Daniel Hernández-Lobato, and Richard E. Turner. Black-box alpha divergence minimization. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016, volume 48 of JMLR Workshop and Conference Proceedings, pages 1511–1520. JMLR.org, 2016. URL <http://proceedings.mlr.press/v48/hernandez-lobatob16.html>.

范马, 孟德宇, 谢齐, 李泽娜, 董宣义。自定进度联合培训。在 Doina Precup 和 Yee Whye Teh, 编辑, 第 34 届机器学习国际会议论文集, 机器学习研究论文集第 70 卷, 第 2275 - 2284 页, 澳大利亚悉尼国际会议中心, 2017 年 8 月 6 日至 11 日。PMLR。网址: <http://proceedings.mlr.press/v70/ma17b.html>。李新哲、孙千如、刘耀耀、周秦、郑世宝、蔡达生、Bernt Schiele。学习自我训练, 进行半监督的几次射击分类。在神经信息处理系统的进步 32, 页 10276 - 10286。□ 2019 Curran Associates, Inc. 版权所有。网址: <http://papers.nips.cc/paper/9216-learning-to-self-train-for-semi-supervised-few-shot-classification.pdf>。Emmeleia Panagiota Mastoropoulou。使用选择性自我训练进行图像分类, 增强深度主动学习。硕士论文, KTH, 电气工程与计算机科学学院 (EECS), 2019 年。Haw - Shiuan Chang, Erik G. Learned - Miller 和 Andrew McCallum。主动偏置: 通过强调高方差样本训练更准确的神经网络。Isabelle Guyon、Ulrike von Luxburg、Samy Bengio、Hanna M. Wallach、Rob Fergus、S. V. N. Vishwanathan 和 Roman Garnett, 编辑, 《神经信息处理系统的进步 30: 2017 年神经信息处理系统年度会议》, 2017 年 12 月 4 日至 9 日, 美国加利福尼亚州长滩, 第 1002 - 1012 Abhinav Shrivastava, Abhinav Gupta 和 Ross B. Girshick。使用在线硬示例挖掘训练基于区域的物体检测器。2016 年 IEEE 计算机视觉和模式识别会议, CVPR 2016, 美国内华达州拉斯维加斯, 2016 年 6 月 27 - 30 日, 第 761 - 769 页。IEEE 计算机学会, 2016。doi: 10.1109/CVPR.2016.89。网址: <https://doi.org/10.1109/CVPR.2016.89>。Alex Graves。神经网络的实用变分推理。在 John Shawe - Taylor、Richard S. Zemel、Peter L. Bartlett、Fernando C. N. Pereira 和 Kilian Q. Weinberger, 编辑, 《神经信息处理系统的进展 24: 2011 年第 25 届神经信息处理系统年会》。2011 年 12 月 12 日至 14 日在西班牙格拉纳达举行的会议记录, 第 2348 - 2356 页, 2011 年。网址: <http://papers.nips.cc/paper/4329-practical-variational-inference-for-neural-networks>。Charles Blundell, Julien Cornebise, Koray Kavukcuoglu 和 Daan Wierstra。神经网络中的权重不确定性。Francis R. Bach 和 David M. Blei, 编辑, 《第 32 届机器学习国际会议论文集》, ICML 2015, 法国里尔, 2015 年 7 月 6 日至 11 日, JMLR 研讨会和会议论文集第 37 卷, 第 1613 - 1622 页。JMLR.org, 2015 年。网址: <http://proceedings.mlr.press/v37/blundell15.html>。José Miguel Hernández - Lobato、Yingzhen Li、Mark Rowland、Thang D. Bui、Daniel Hernández - Lobato 和 Richard E. Turner。黑盒阿尔法发散最小化。在 Maria - Florina Balcan 和 Kilian Q. Weinberger, 编辑, 第 33 届机器学习国际会议论文集, ICML 2016, 纽约, 美国, 2016 年 6 月 19 - 24 日, JMLR 研讨会和会议论文集第 48 卷, 第 1511 - 1520 页。JMLR.org, 2016 年。网址: <http://proceedings.mlr.press/v48/hernandez-lobatob16.html>。