#4365

BayesPrompt: Prompting Large-Scale Pre-Trained Language Models on Few-shot Inference via Debiased Domain Abstraction

Jiangmeng Li*, Fei Song*, Yifan Jin, Wenwen Qiang,
Changwen Zheng, Fuchun Sun, Hui Xiong

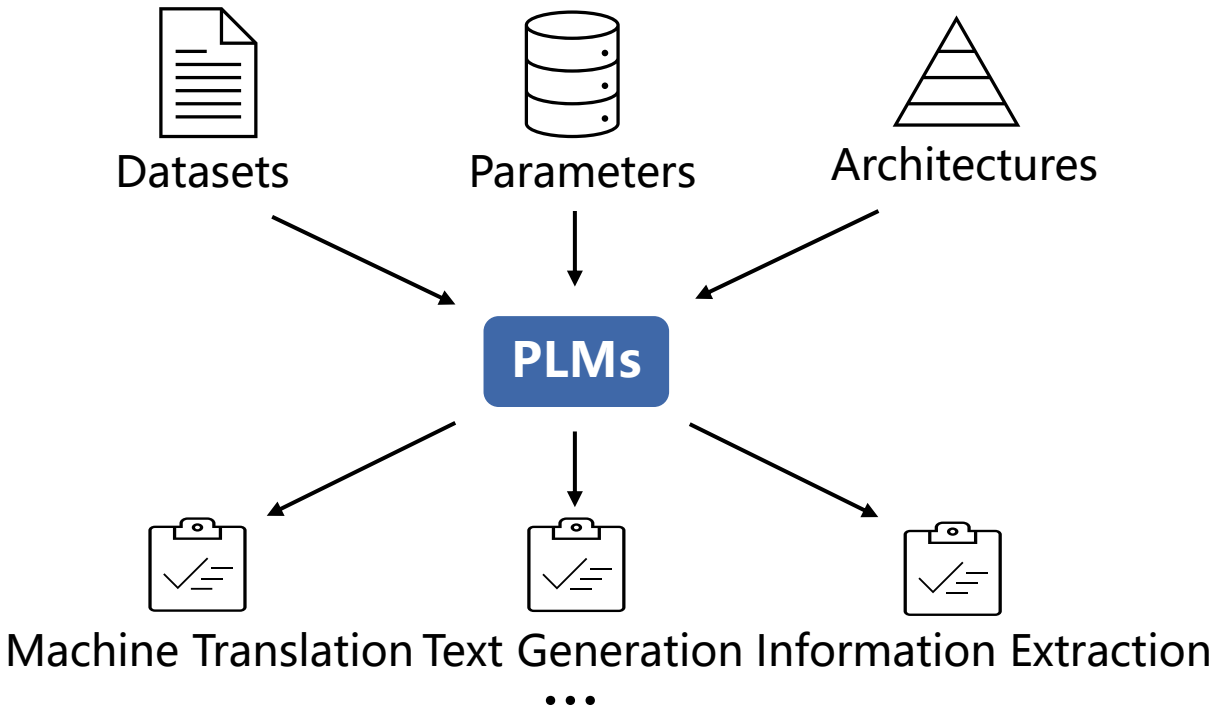
中国科学院大学
University of Chinese Academy of Sciences

清華大學
Tsinghua University

香港科大（广州）
HKUST（GZ）

Introduction

Background



Development Bottleneck

few-shot scenarios

Intrinsic Reason

over-multitudinous conceptual knowledge

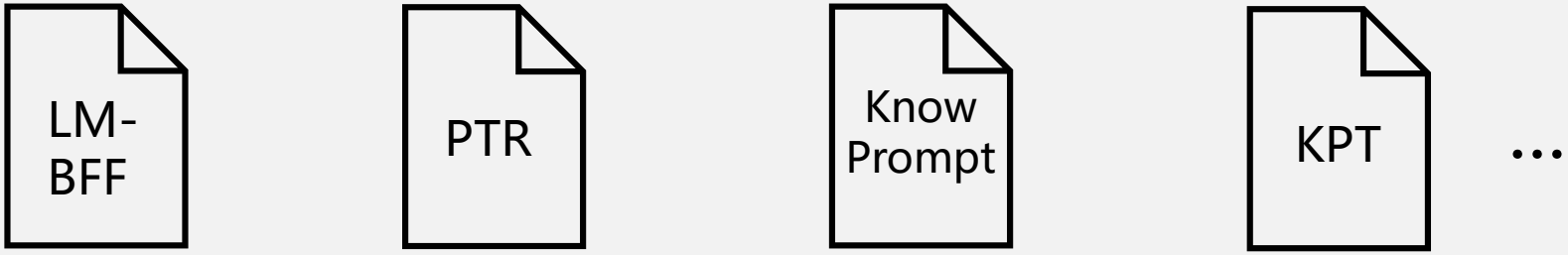
- **Over-multitudinous conceptual knowledge:** The knowledge contained by PLMs exhibits inherent polysemy.
- Domain-irrelevant knowledge may interfere with the inference on downstream tasks, especially for few-shot datasets.

Introduction

Related Work

GPT-3 (Brown et al., 2020)

Prompt-tuning



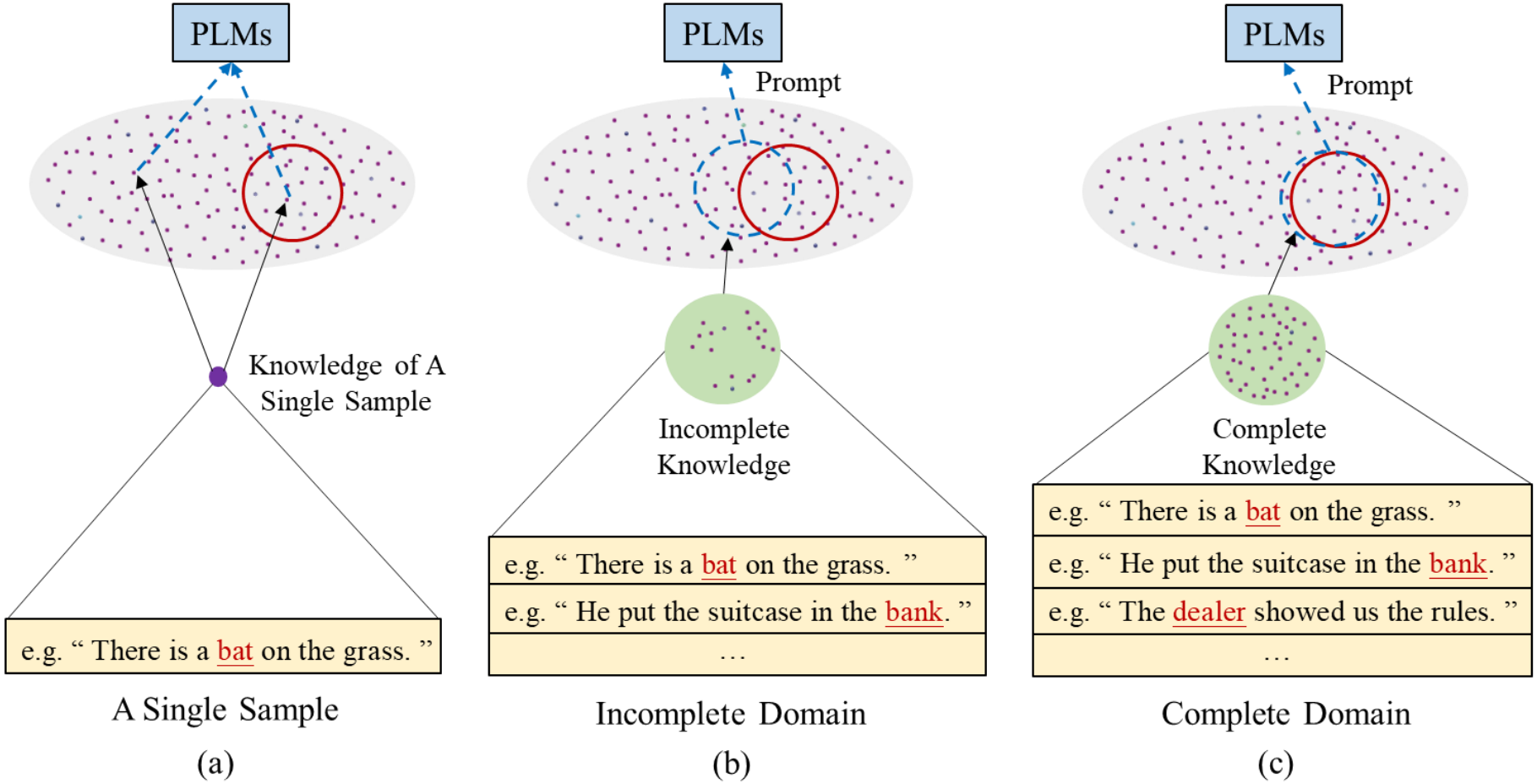
(Gao et al., 2021) (Han et al., 2022) (Chen et al., 2022b) (Hu et al., 2022)

A Long-Standing Challenge

The limited and discrete semantic information contained in the training samples from downstream domains can barely support the conventional trainable prompts to acquire sufficient supervision, such that the guidance of the generated prompts is trivial to PLMs. Especially, such a challenge further exacerbates the performance of PLMs in few-shot scenarios.

Introduction

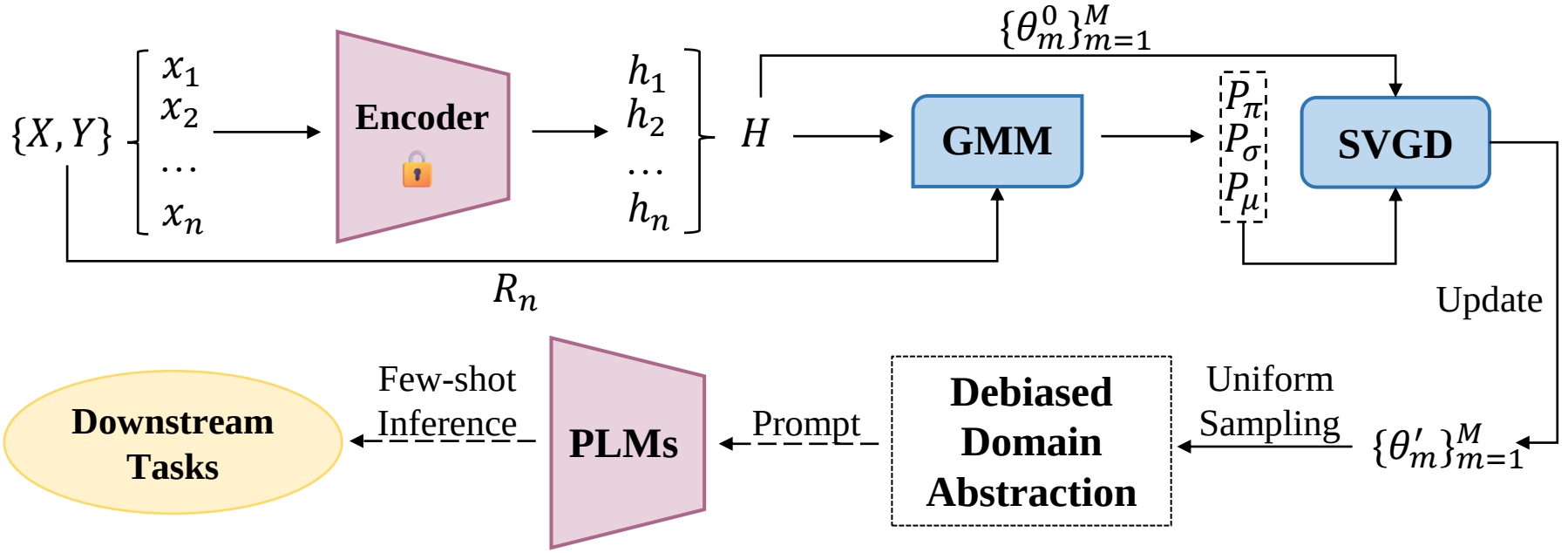
Distribution Perspective



Motivation: We intuitively explore to approximate the complete training domains on downstream tasks in a debiased manner, and then abstract such domains to generate discriminative prompts, thereby providing the de-ambiguous guidance for PLMs.

Method

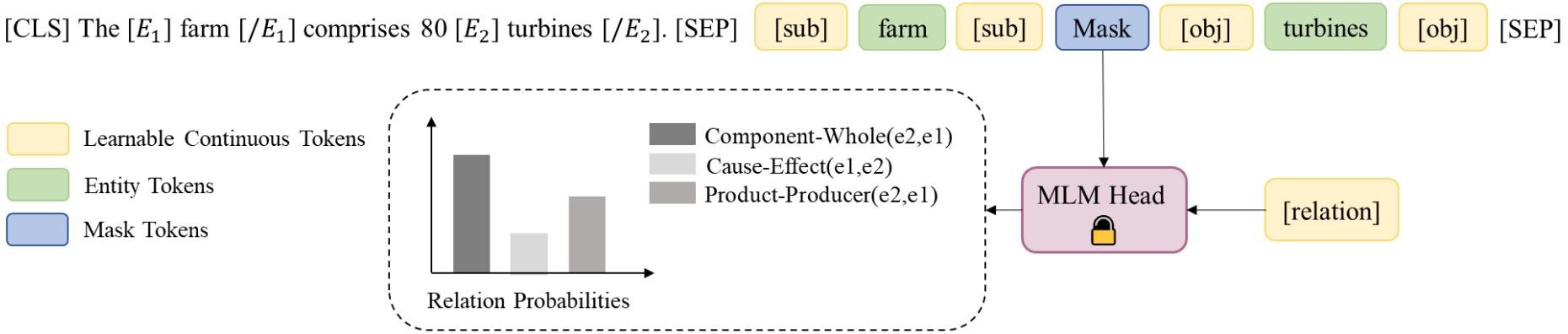
Overview

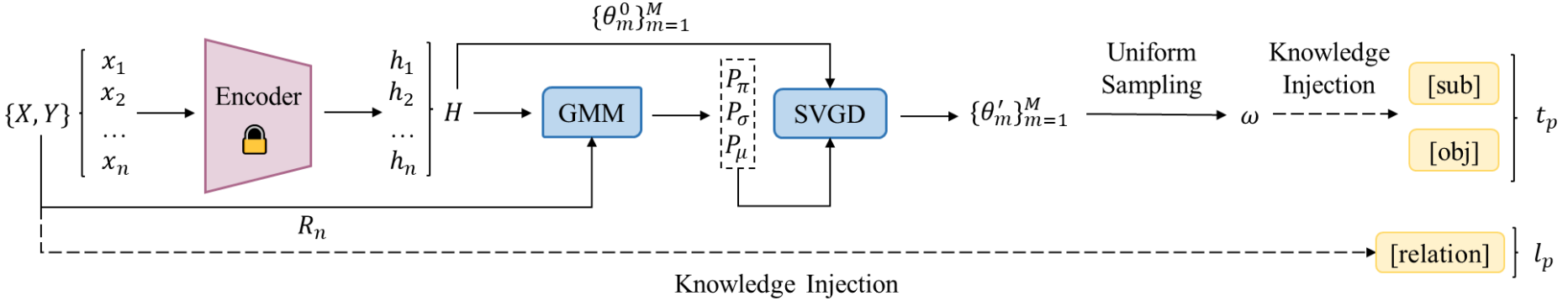


- **GMM:** The Gaussian Mixture Model (GMM) is constructed to fit the sample distribution.
- **SVGD:** The distribution approximation is achieved by using Stein Variational Gradient Descent (SVGD), which is a general-purpose Bayesian inference algorithm.
- **Debiased Domain Abstraction:** Abstracting a feature that can well describe the actual downstream domain based on the debiased factual distribution.

Method

Implementation





The Loss Function: $J = -\frac{1}{|X|}\sum_{x \in X} ylogP([MASK] = M(y)|T(x))$

Experiments

Few-Shot Setting

F1 scores (%) of prompt-tuning models with different settings

		Few-Shot Setting							
Datasets	Split	FINE-TUNING	GDPNET	PTR	KnowPrompt	RetrievalRE	**BayesPrompt**	Δ(B-K)	Δ(B-R)
SemEval	K=1	18.5(±1.4)	10.3(±2.5)	14.7(±1.1)	28.6(±6.2)	33.3(±1.6)	**35.1**(±2.9)	**+4.3**	**+1.23**
	K=5	41.5(±2.3)	42.7(±2.0)	53.9(±1.9)	66.1(±8.6)	69.7(±1.7)	**71.6**(±3.3)		
	K=16	66.1(±0.4)	67.5(±0.8)	80.6(±1.2)	80.9(±1.6)	**81.8**(±1.0)	**81.8**(±1.2)		
TACRED	K=1	7.6(±3.0)	4.2(±3.8)	8.6(±2.5)	17.6(±1.8)	19.5(±1.5)	**22.5**(±2.5)	**+3**	**+1.27**
	K=5	16.6(±2.1)	15.5(±2.3)	24.9(±3.1)	28.8(±2.0)	30.7(±1.7)	**31.4**(±0.6)		
	K=16	26.8(±1.8)	28(±1.8)	30.7(±2.0)	34.7(±1.8)	36.1(±1.2)	**36.2**(±0.8)		
TACREV	K=1	7.2(±1.4)	5.1(±2.4)	9.4(±0.7)	17.8(±2.2)	18.7(±1.8)	**21.9**(±2.0)	**+2.43**	**+1.37**
	K=5	16.3(±2.1)	17.8(±2.4)	26.9(±1.5)	30.4(±0.5)	30.6(±0.2)	**31.2**(±0.8)		
	K=16	25.8(±1.2)	26.4(±1.2)	31.4(±0.3)	33.2(±1.4)	35.3(±0.3)	**35.6**(±0.7)		

- Δ**(B-K)** denotes the comparison between BayesPrompt and KnowPrompt, and Δ**(B-R)** denotes the comparison between BayesPrompt and RetrievalRE.
- On average, BayesPrompt beats KnowPrompt by **3.24%** among benchmark datasets. For RetrievalRE, BayesPrompt achieves an average improvement of **1.29%** among benchmark datasets.

Experiments

Standard Setting

Standard RE performance of F1 scores (%) on benchmarks

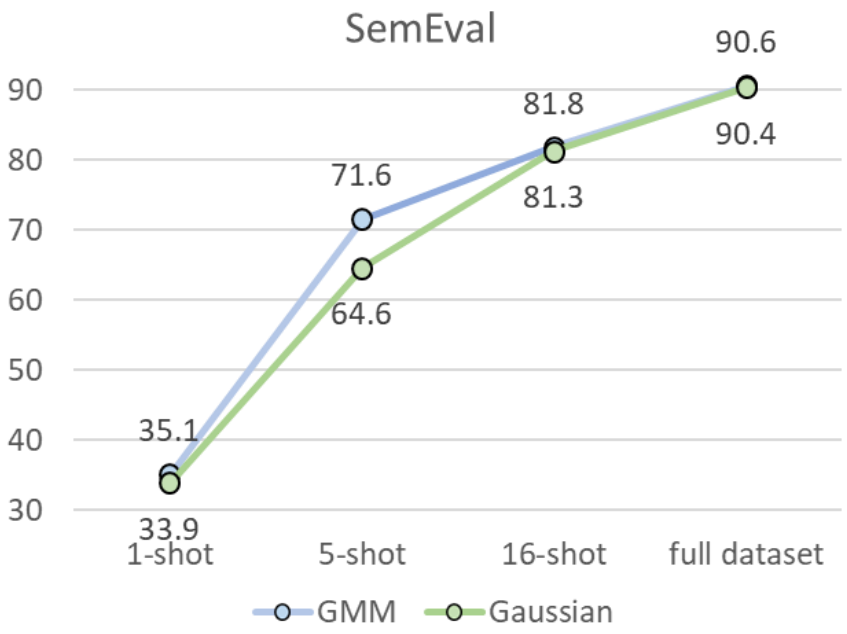
Standard Setting						
Methods	Extra Data	SemEval	TACRED	TACREV	RE-TACRED	Average
Fine-tuning pre-trained models						
FINE-TUNING	w/o	87.6	68.7	76.0	84.9	79.3
SPANBERT	w/	-	70.8	78.0	85.3	78.0
KNOWBERT	w/	89.1	71.5	79.3	89.1	82.3
LUKE	w/	-	72.7	80.6	-	76.7
MTB	w/	89.5	70.1	-	-	79.8
GDPNET	w/o	-	71.5	79.3	-	75.4
Prompt-tuning pre-trained models						
PTR	w/o	89.9	72.4	81.4	90.9	83.7
KnowPrompt	w/o	90.2	72.4	82.4	91.3	84.1
RetrievalRE	w/o	90.4	72.7	82.7	**91.5**	84.3
BayesPrompt	w/o	**90.6**	**72.9**	**83.0**	91.4	**84.5**

Experiments

Ablation Study

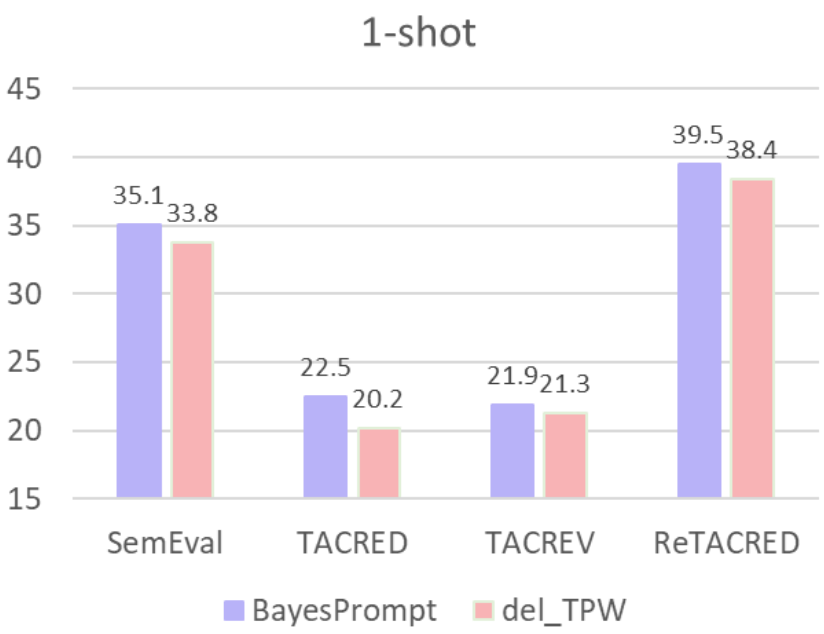
Distribution Assumption

[GMM vs. Gaussian]



Discriminative Prompts

[W/ vs. W/O]



Theoretical Insights

Theoretical Insights With Connection To Domain Adaptation

- The gap between prompting problem and domain adaptation

Domain Adaptation

The "domain adaptation" is learning from a source data distribution a well-performing model on a different (but related) target data distribution

BayesPrompt

BayesPrompt aims to fit the distribution of a few-shot domain, but it is not going to align the distributions of the target few-shot domain and the domain of PLMs.

- Do the theoretical assumptions on a shared label space from domain adaptation hold in prompt- tuning?

In the prompt-tuning scenario, the downstream domain can be treated as the target domain, and the specific subset of the PLM domain can be treated as the source domain, i.e., the domain distribution alignment is performed between the specific subset of the PLM domain and the downstream domain, which have the shared labels. However, the downstream domain can be bounded by the discrete data, but the specific subset of the PLM domain cannot be certainly determined, such that conventional domain adaptation methods cannot be directly leveraged to achieve our objective.

Theoretical Insights

Theoretical Validity of BayesPrompt

Proposition C.1. *Let* $\mathcal{P}(Z)$ *be the set of Borel probability measures. For* $\mathring{\mathcal{P}}^f_{PLM}(Z)$, $\mathcal{P}^f_{DS}(Z) \in \mathcal{P}(Z)$, *there exists a pseudometric, i.e.,* $\mathrm{d}\left(\mathring{\mathcal{P}}^f_{PLM}(Z), \mathcal{P}^f_{DS}(Z)\right)$, *satisfying the negative, symmetric, and triangle inequality conditions. Furthermore,* $\mathrm{d}\left(\mathring{\mathcal{P}}^f_{PLM}(Z), \mathcal{P}^f_{DS}(Z)\right) = 0$ *holds, when* $\mathring{\mathcal{P}}^f_{PLM}(Z) = \mathcal{P}^f_{DS}(Z)$.

Corollary C.2. *Let* $\mathcal{P}^f_L$ *be the distribution of the domain containing the label information, which is stratified from* $\mathcal{P}^f_{DS}$. *For* $\mathring{\mathcal{P}}^f_{PLM}(Z)$, $\mathcal{P}^f_{DS}(Z)$, $\mathcal{P}^f_L \in \mathcal{P}(Z)$, *we have*

$$\mathrm{d}\left(\mathring{\mathcal{P}}^f_{PLM}(Z), \mathcal{P}^f_{DS}(Z)\right) \leq \mathrm{d}\left(\mathring{\mathcal{P}}^f_{PLM}(Z), \mathcal{P}^f_L(Z)\right) + \mathrm{d}\left(\mathcal{P}^f_L(Z), \mathcal{P}^f_{DS}(Z)\right), \tag{12}$$

Theorem C.3. *Suppose* $\mathring{\mathcal{D}}_{PLM}$ *and* $\mathcal{D}_{DS}$ *share a labeling function, i.e.,* $\mathcal{L}: Z \to Y$. *For the predictor functions* $\forall h \in \mathcal{H}$, *we have the following inequality:*

$$\varepsilon^f_{h,\mathcal{L}}\left(\mathcal{P}^f_{DS}(Z)\right) \leq \varepsilon^f_{h,\mathcal{L}}\left(\mathring{\mathcal{P}}^f_{PLM}(Z)\right) + \mathrm{d}\left(\mathring{\mathcal{P}}^f_{PLM}(Z), \mathcal{P}^f_{DS}(Z)\right) + \eta, \tag{13}$$

where $\eta = \varepsilon^f_{h^\star,\mathcal{L}}\left(\mathcal{P}^f_{DS}(Z)\right) + \varepsilon^f_{h^\star,\mathcal{L}}\left(\mathring{\mathcal{P}}^f_{PLM}(Z)\right)$ *and* $h^\star$ *is the ideal joint predictor shared by the two domains after training.*

Compared with benchmark approaches, BayesPrompt derives **the tighter classification error upper bound** on the downstream inference of PLMs.

Thanks !