Meta Distant Transfer Learning for Pre-trained Language Models

Chengyu Wang[1,2], Haojie Pan[2], Minghui Qiu[2], Fei Yang[1], Jun Huang[2], Yin Zhang[3]

[1] Zhejiang Lab, [2] Alibaba Group, [3] College of Computer Science and Technology, Zhejiang University

Key Contributions

- The *Meta Distant Transfer Learning* (Meta-DTL) framework is proposed to digest the cross-task, transferable knowledge and alleviates *negative transfer* for pre-trained language models.
- Meta-DTL employs a *task representation learning* procedure to obtain a collection of *prototype vectors* for each task. The *meta-learner* is trained by multi-task learning with rich *meta-knowledge* injected based on prototype vectors to capture the cross-task transferable knowledge.
- In the experiments, we apply Meta-DTL to BERT and ALBERT for three sets of NLP tasks. Experiments show that Meta-DTL consistently outperforms strong baselines.

Introduction

Background. Pre-trained Language Models (PLMs) achieve the state-of-the-art results for a majority of text classification tasks. However, the performance of PLMs on a downstream task may be limited by the availability of the training set. A large number of transfer learning algorithms address tasks across similar sub-domains. For PLMs, these models can be fine-tuned over both source-domain and target-domain datasets by various multi-task training strategies. When there exist large domain gaps and class label differences, these transfer learning solutions are likely to fail. A natural question arises: *how can we transfer knowledge across distant domains with different classification targets for PLM-based text classification*?

Our Work. The *Meta Distant Transfer Learning* (Meta-DTL) framework is proposed. Specially, Meta-DTL employs a *task representation learning* procedure to obtain a collection of *prototype vectors* for each task. To understand how to transfer across these tasks and classes, we construct a *Meta Knowledge Graph* (Meta-KG) to characterize the implicit relations among tasks and classes, based on the representations of multiple tasks. The *meta-learner* in Meta-DTL can be initialized by any PLMs and trained by multi-task learning with rich *meta-knowledge* injected from Meta-KG. Additionally, we design the *Weighted Maximum Entropy Regularizers* to make the model more *task-agnostic* and *unbiased*. Finally, the meta-learner can be fine-tuned to fit each task using its own training set. In this way, the model is able to digest the cross-task, transferable knowledge and alleviates *negative transfer*.

Meta-DTL: The Proposed Framework

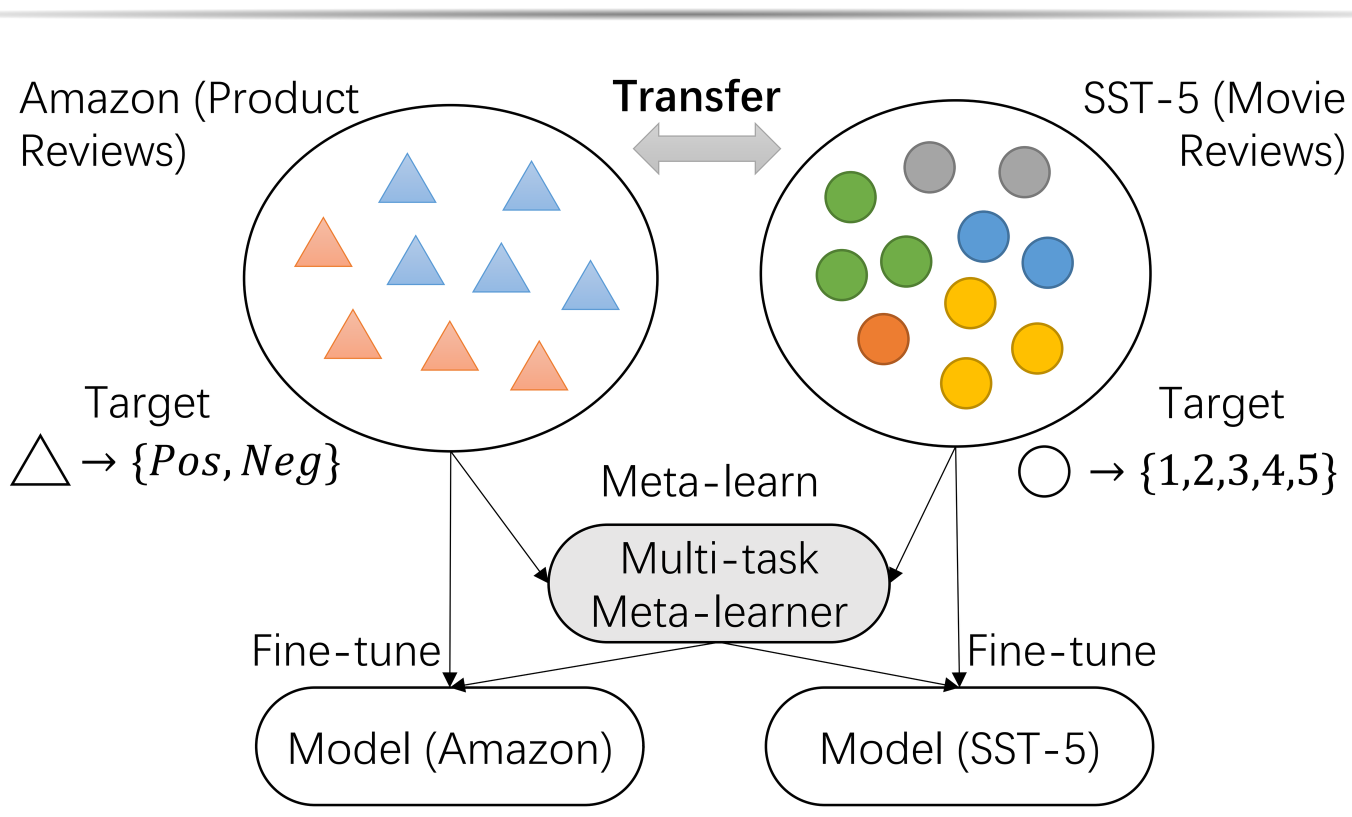

Figure 1:A simple example of *Meta Distant Transfer Learning* for review analysis.

Meta-DTL consists of three modules: i) *Task Representation Learning* (TRL), ii) *Multi-task Meta-learner Training* (MMT), and iii) *Task-specific Model Fine-tuning* (TMF).

Meta-DTL: The Proposed Framework

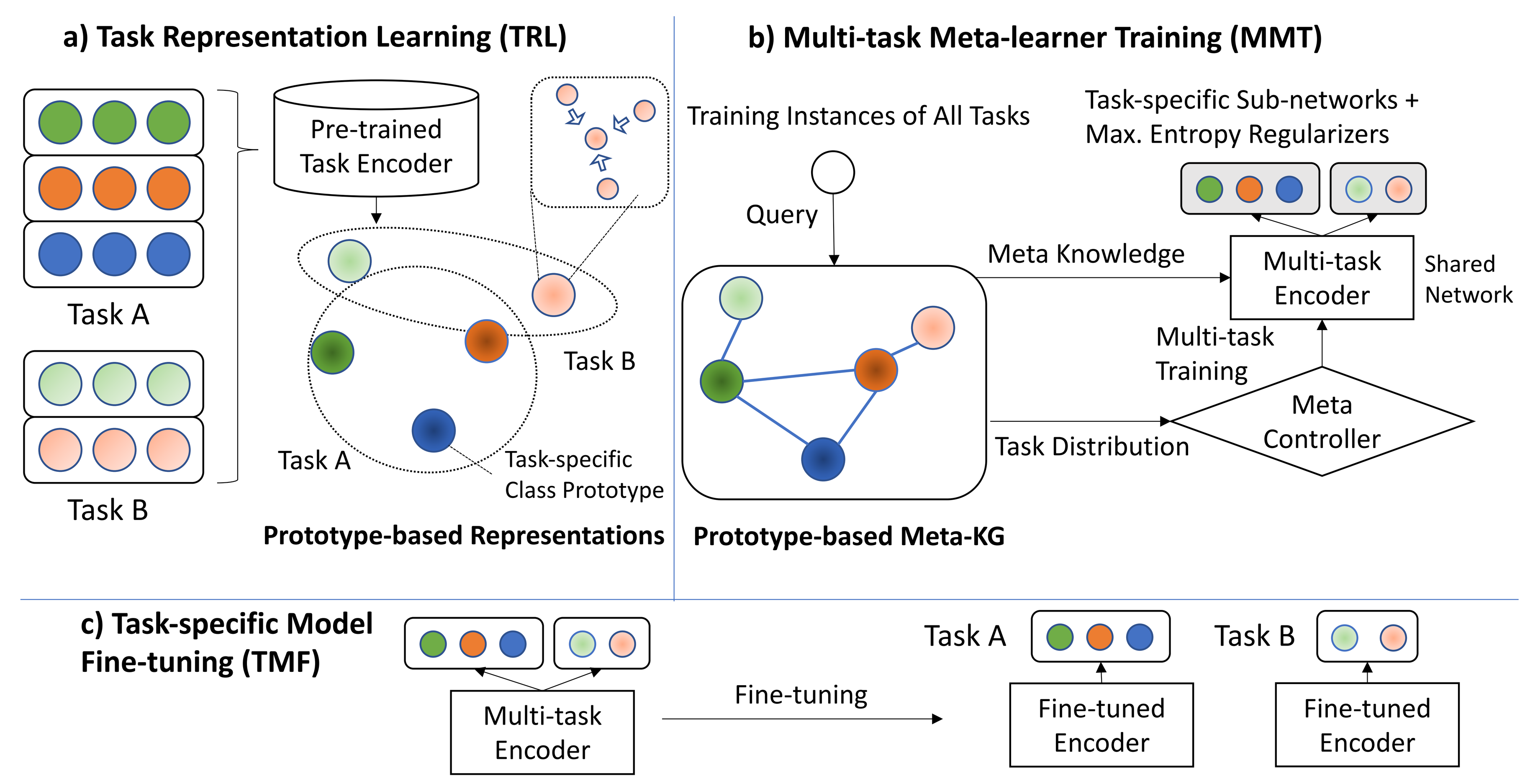

Figure 2:The high-level architecture of Meta-DTL.

Task Representation Learning: Specially, for each task $\mathcal{T}_i$, TRL employs a *pre-trained task encoder* to do a one-pass scan over the training set $\mathcal{D}_i$. It represents each task $\mathcal{T}_i$ as a collection of *prototypical vectors*, denoted as $\mathcal{P}_i = \{\vec{p}_{i,j}\}$ where $\vec{p}_{i,j}$ is the j-th prototypical embedding vector of $\mathcal{T}_i$, corresponding to the j-th class in $\mathcal{D}_i$.

Multi-task Meta-learner Training: We obtain a meta-learner $\mathcal{M}$ that only digests *transferable knowledge* across all the K tasks. We first construct a prototype-based *Meta Knowledge Graph* (Meta-KG, denoted as G) from $\mathcal{P}_i, \cdots, \mathcal{P}_K$, implicitly describing the relations among tasks and classes. For each training instance of all tasks $x_{i,j}$, we query $x_{i,j}$ in G to generate the meta-knowledge score $m_{i,j}$, which represents the degree of the *knowledge transferability* of the input $x_{i,j}$. Additionally, the *Weighted Maximum Entropy Regularizers* (WMERs) are proposed and integrated into the model to make the meta-learner $\mathcal{M}$ more *task-agnostic* and *unbiased*.

Task-specific Model Fine-tuning: In TMF, we fine-tune the meta-learner $\mathcal{M}$ to generate the K classifiers for the K tasks, based on their own training sets $\mathcal{D}_1, \cdots, \mathcal{D}_K$.

Experiments

Key Results. In the experiments, we apply the Meta-DTL framework to BERT and ALBERT for three sets of NLP tasks (seven public datasets in total): i) coarse and fine-grained review analysis across domains; ii) natural language inference (across sentence relation prediction and scientific question answering); and iii) lexical semantics (across hypernymy detection and lexical relation classification). Experiments show that Meta-DTL consistently outperforms strong baselines, regardless of the types of underlying PLMs and downstream NLP tasks.

PLM	Method	Review Analysis Tasks				NLI Tasks			Lexical Semantic Tasks		
		SST-5	Amazon	IMDb	Avg.	MNLI	SciTail	Avg.	Shwartz	BLESS	Avg.
Bert	Single-task	53.4	89.3	95.2	79.3	83.0	92.4	87.7	92.6	93.2	92.9
	Multi-task	53.2	89.8	95.6	79.5	83.8	92.0	87.9	92.8	93.0	92.9
	Task Comb.	53.2	89.5	94.1	78.9	83.7	92.2	87.9	91.3	91.7	91.5
	Meta-FT*	53.6	91.0	95.8	80.1	83.9	93.4	88.6	92.8	93.5	93.1
	Meta-DTL	54.6††	91.8††	98.2††	81.5	84.2†	93.6††	88.9	93.2††	94.8††	94.0
Albert	Single-task	51.0	87.6	93.6	77.4	80.7	88.2	84.4	92.0	90.7	91.3
	Multi-task	50.3	88.1	94.2	77.5	81.0	88.3	84.6	92.4	91.0	91.7
	Task Comb.	49.8	88.0	93.6	77.1	80.8	85.2	83.0	91.4	90.6	91.0
	Meta-FT*	50.8	88.4	95.0	78.0	81.2	88.7	84.9	92.4	91.9	92.1
	Meta-DTL	51.2††	88.8††	97.6††	79.2	82.4††	89.2††	85.8	92.8†	93.4††	93.1

Table 1:General performance of Meta-DTL and all the baselines over all the datasets in terms of accuracy. The p-values of the paired t-tests for each dataset are marked as follows: $^{\dagger\dagger}: p < 0.05$ and $^{\dagger}: 0.05 < p < 0.1$.