Meta Distant Transfer Learning for Pre-trained Language Models

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

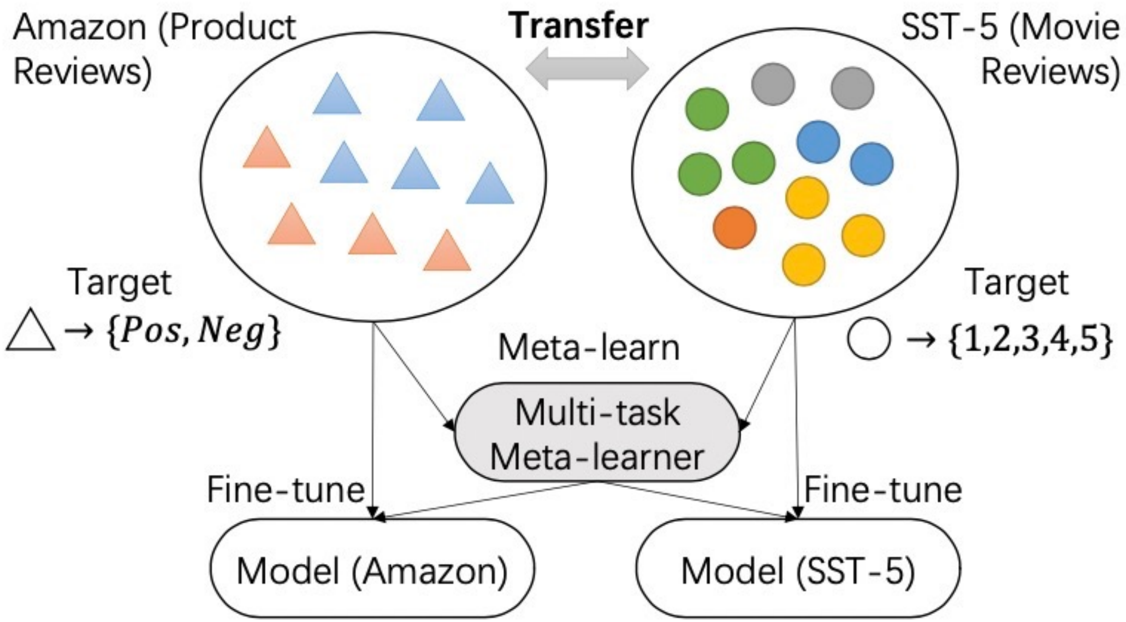
[1] Alibaba Group [2] Zhejiang Lab [3] Zhejiang University

Introduction (1)

✓Transfer learning for Pre-trained Language Models (PLMs)

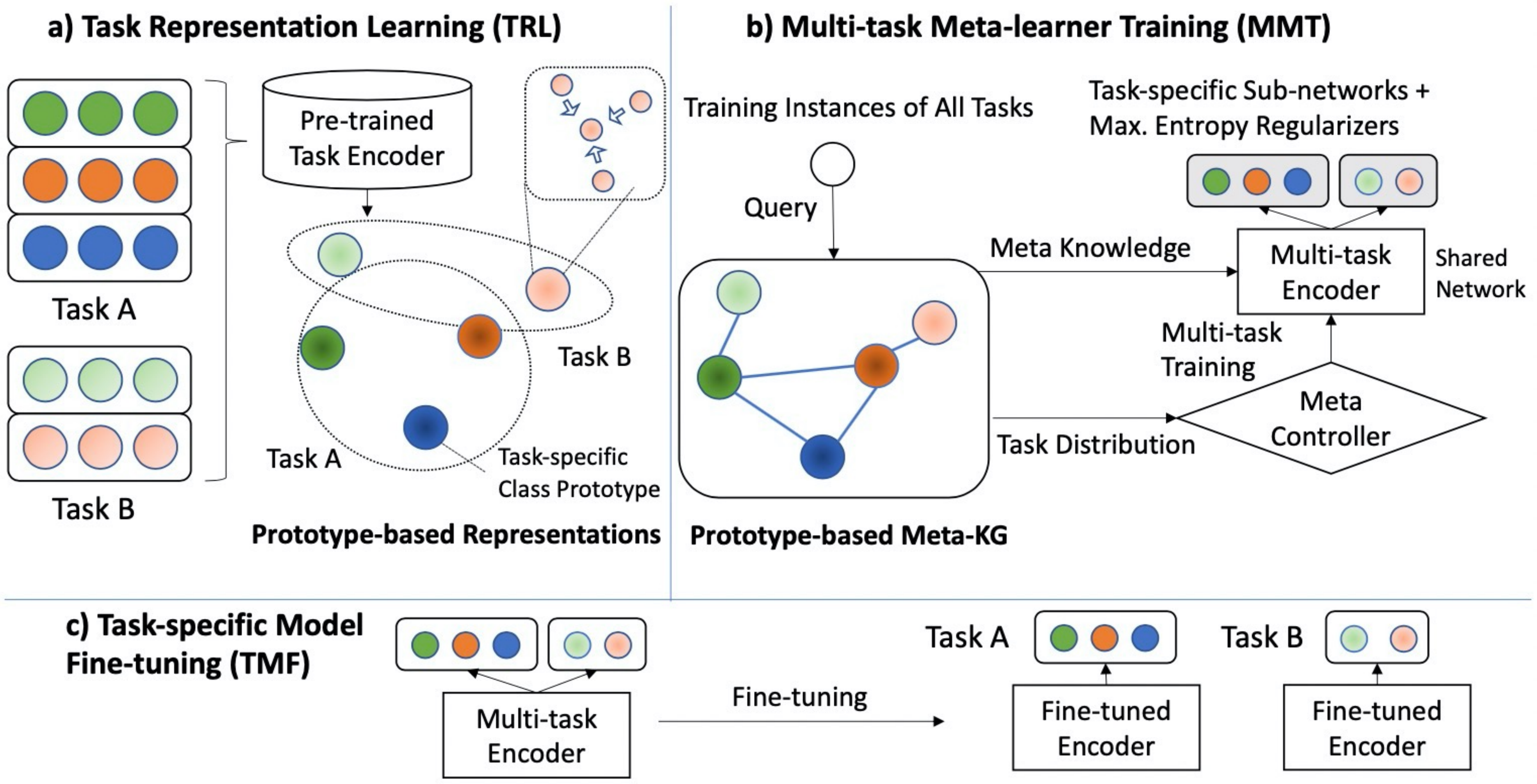
- Fine-tuning by multi-task learning: learning from source-domain datasets may force PLMs to memorize non-transferable knowledge of source domains, leading to the negative transfer effect

Research Question: *how can we transfer knowledge across distant domains with different classification targets for PLM-based text classification?*



Introduction (2)

✓Our idea: the Meta-DTL framework



Task Representation Learning

✓Learning the prototypical vector for each class in each task

- The input includes both the text and the class label

$$\vec{p}_{i,j} = \frac{1}{|\mathcal{D}_{i,j}|} \sum_{x_{i,j} \in \mathcal{D}_{i,j}} \mathcal{E}(x_{i,j}, c_{i,j})$$

PLM Encoding Function

Multi-task Meta-learner Training

✓Obtaining the meta-knowledge

- Considering both the instance-level and the class-level meta-knowledge

$$\alpha_{i,j} = \max_{\vec{p}_{m,n} \in \tilde{\mathcal{P}}_i} \cos(\mathcal{E}(x_{i,j}, c_{i,j}), \vec{p}_{m,n}) \qquad \beta_{i,j} = \max_{\vec{p}_{m,n} \in \tilde{\mathcal{P}}_i} \cos(\vec{p}_{i,j}, \vec{p}_{m,n})$$

✓Training the meta-learner

- Weighted cross-entropy loss

$$\mathcal{L}_{CE}(x_{i,j}) = -\sum_{c \in \mathcal{C}_i} \mathbf{1}_{(c_{i,j}=c)} m_{i,j} \log \tau_c(x_{i,j})$$

- Weighted Maximum Entropy Regularizer

$$\mathcal{L}_{ME}(x_{i,j}) = -\sum_{c \in \mathcal{C}_i} \frac{m_{i,j}}{|\mathcal{C}_i|} \log \tau_c(x_{i,j})$$

Task-specific Model Fine-tuning

✓Fine-tuning the meta-learner for specific tasks

- The dataset-level loss function

$$\mathcal{L}^*(\mathcal{T}_i) = -\sum_{x_{i,j}\in\mathcal{D}_i}\sum_{c\in\mathcal{C}_i}\mathbf{1}_{(c_{i,j}=c)}\log\tau_c^*(x_{i,j})$$

Experiments (1)

✓Experimental datasets

Name	Task Description	Classification Label Set	#Train	#Dev.	#Test
SST-5	Fine-grained movie review analysis	{1, 2, 3, 4, 5}	8,544	1,101	2,210
Amazon	Coarse-grained product review analysis	{positive, negative}	7,000	500	500
IMDb	Coarse-grained movie review analysis	{positive, negative}	23,785	1,215	25,000
MNLI	NLI across multiple genres	{entailment, neutral, contradiction}	382,702	10,000	9,815
SciTail	Scientific question answering	{entailment, neutral}	23,596	1,304	2,126
Shwartz	Hypernymy detection	{hypernymy, non-hypernymy}	20,335	1,350	6,610
BLESS	Lexical relation classification	{event, meronymy, random, co-hyponymy, attribute, hypernymy}	18,582	1,327	6,637

Experiments (2)

✓Overall experiments

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**

Experiments (3)

✓Ablation Study

Task	w/o.IMK	w/o.WMER	Full
SST-5	**54.0**	53.8	54.6
Amazon	90.6	**90.8**	91.8
IMDb	97.0	**97.6**	98.2
MNLI	84.0	**84.1**	84.2
SciTail	**92.9**	92.7	93.6
Shwartz	91.8	**92.2**	93.2
BLESS	93.5	**93.8**	94.8
Avg.	86.4	**86.6**	87.2

✓Learning with Small Data

- Using a small number of MNLI training samples

PCT	Single	Meta-FT*	Meta-DTL
1%	62.5	64.1	**66.5 (+4.0%)**
2%	67.5	68.2	**69.8 (+2.3%)**
5%	72.8	73.8	**74.2 (+1.4%)**
10%	75.8	76.2	**77.6 (+1.8%)**
20%	80.4	80.8	**81.4 (+1.0%)**

Conclusion

- ✓We present the Meta-DTL framework for few-shot learning across tasks with distant domains and labels.
- ✓Experiments confirm the effectiveness of Meta-DTL over various NLP tasks.
- ✓Future work includes:
  - ✓Using Meta-DTL in other application scenarios and other NLP tasks
  - ✓Exploring how Meta-DTL can be applied to other PLMs apart from BERT-style models

THANKS

--------- Q&A Section --------