# Meta-KD: A Meta Knowledge Distillation Framework for Language Model Compression across Domains

Haojie Pan*, **Chengyu Wang***, Minghui Qiu, Yichang Zhang, Yaliang Li, Jun Huang

Alibaba Group

# Introduction (1)

✓Knowledge distillation for pre-trained language models (PLMs)

- Distilling the knowledge from a large teacher model to a small student model
- Difficult to capture knowledge from other domains

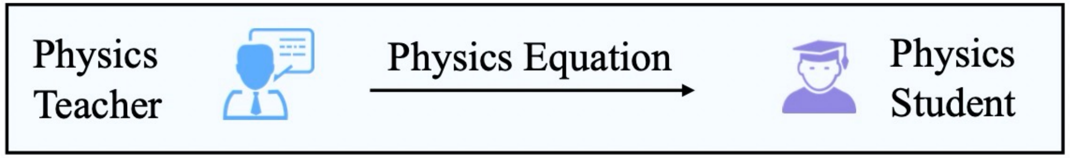
✓Cross-domain knowledge distillation

- Teachers of other domains may pass non-transferable knowledge to the student model, hence harming the performance

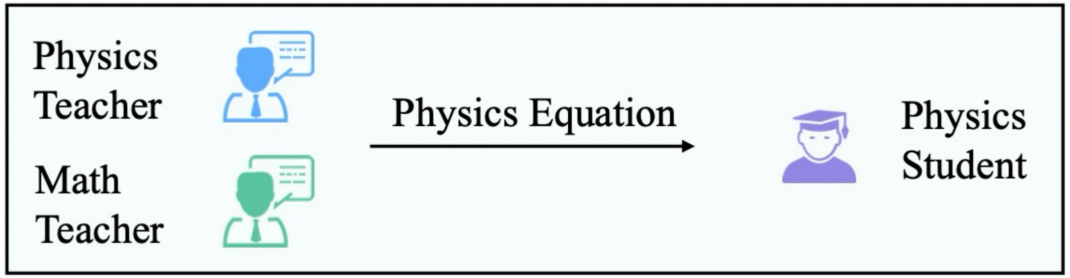
# Introduction (2)

✓Our idea: Meta Knowledge Distillation (Meta-KD)

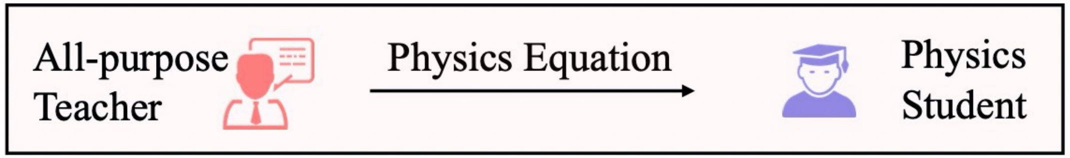
- Meta-teacher learning: learning a meta-teacher model that captures transferable knowledge across domains
- Meta-distillation: learning a student model over a domain-specific dataset with the selective guidance from the meta-teacher




(a) Learning from an in-domain teacher.




(b) Learning from multiple teachers of varied domains.




(c) Learning from meta-teacher with multi-domain knowledge.

**Motivation example**

# Meta-teacher Learning

✓Learning instance-level transferable knowledge

- Compute prototype scores to select transferable instances across domains

$$t_k^{(i)} = \alpha \cos(p_k^{(m)}, h(X_k^{(i)})) + \zeta \sum_{k'=1}^{K(k' \neq k)} \cos(p_{k'}^{(m)}, h(X_k^{(i)}))$$

**Within-domain Class Centroid** | **Out-of-domain Class Centroid**

✓Learning feature-level transferable knowledge

- Add a domain-adversarial loss to make the PLM more domain-invariant

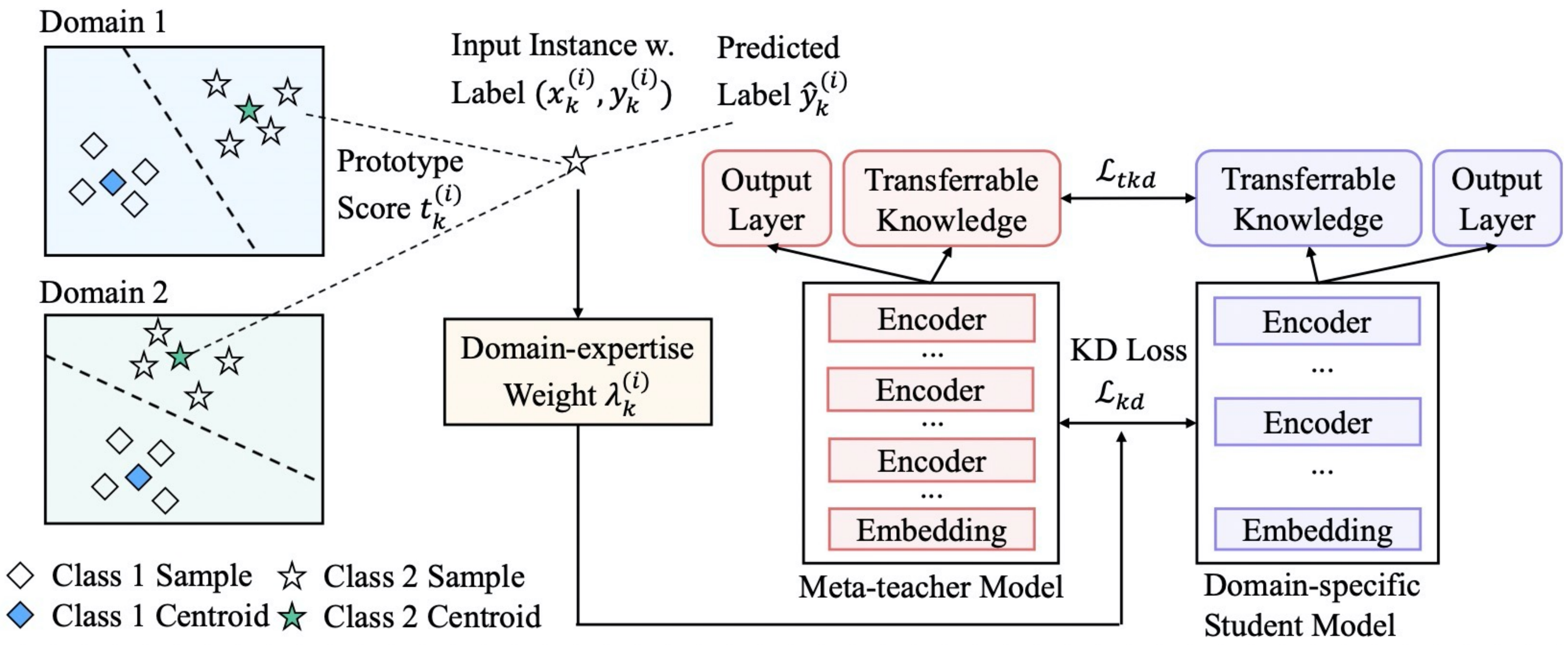
$$\mathcal{L}_{DA}(X_k^{(i)}) = -\sum_{k=1}^{K} \mathbf{1}_{k=z_k^{(i)}} \cdot \log \sigma(h_d(X_k^{(i)}))$$

# Meta-distillation

✓New loss functions and factors for knowledge distillation

- Transferable knowledge distillation loss
- Domain expertise weights $\lambda_k^{(i)} = \dfrac{1 + t_k^{(i)}}{\exp^{(\hat{y}_k^{(i)} - y_k^{(i)})^2} + 1}$

How well the meta-teacher can supervise the student on a specific input



# Experiments (1)

✓Datasets and experimental settings

- Teacher model: BERT-base (L=12, H=768, A=12, #Para.=110M)
- Student model: BERT-small ((L=4, H=312, A=12, #Para.=14.5M)

| Dataset | Domain | #Train | #Dev | #Test |
|---|---|---|---|---|
| MNLI | Fiction | 69,613 | 7,735 | 1,973 |
| | Gov. | 69,615 | 7,735 | 1,945 |
| | Slate | 69,575 | 7,731 | 1,955 |
| | Telephone | 75,013 | 8,335 | 1,966 |
| | Travel | 69,615 | 7,735 | 1,976 |
| Amazon Reviews | Book | 1,631 | 170 | 199 |
| | DVD | 1,621 | 194 | 185 |
| | Elec. | 1,615 | 172 | 213 |
| | Kitchen | 1,613 | 184 | 203 |

✓Experimental results

- Results on MNLI

| Method | Fiction | Government | Slate | Telephone | Travel | Average |
|---|---|---|---|---|---|---|
| $\text{BERT}_\text{B}$-single | 82.2 | 84.2 | 76.7 | 82.4 | 84.2 | 81.9 |
| $\text{BERT}_\text{B}$-mix | 84.8 | 87.2 | 80.5 | 83.8 | 85.5 | 84.4 |
| $\text{BERT}_\text{B}$-mtl | 83.7 | 87.1 | 80.6 | 83.9 | 85.8 | 84.2 |
| Meta-teacher | 85.1 | 86.5 | 81.0 | 83.9 | 85.5 | 84.4 |
| $\text{BERT}_\text{B}\text{-single} \xrightarrow{\text{TinyBERT-KD}} \text{BERT}_\text{S}$ | 78.8 | 83.2 | 73.6 | 78.8 | 81.9 | 79.3 |
| $\text{BERT}_\text{B}\text{-mix} \xrightarrow{\text{TinyBERT-KD}} \text{BERT}_\text{S}$ | 79.6 | 83.3 | 74.8 | 79.0 | 81.5 | 79.6 |
| $\text{BERT}_\text{B}\text{-mtl} \xrightarrow{\text{TinyBERT-KD}} \text{BERT}_\text{S}$ | 79.7 | 83.1 | 74.2 | 79.3 | 82.0 | 79.7 |
| $\text{Multi-teachers} \xrightarrow{\text{MTN-KD}} \text{BERT}_\text{S}$ | 77.4 | 81.1 | 72.2 | 77.2 | 78.0 | 77.2 |
| $\text{Meta-teacher} \xrightarrow{\text{TinyBERT-KD}} \text{BERT}_\text{S}$ | 80.3 | 83.0 | **75.1** | 80.2 | 81.6 | 80.0 |
| $\text{Meta-teacher} \xrightarrow{\text{Meta-distillation}} \text{BERT}_\text{S}$ | **80.5** | **83.7** | 75.0 | **80.5** | **82.1** | **80.4** |

# Experiments (2)

✓ Results on Amazon (full data)

| Method | Books | DVD | Electronics | Kitchen | Average |
|---|---|---|---|---|---|
| $\text{BERT}_\text{B}$-single | 87.9 | 83.8 | 89.2 | 90.6 | 87.9 |
| $\text{BERT}_\text{B}$-mix | 89.9 | 85.9 | 90.1 | 92.1 | 89.5 |
| $\text{BERT}_\text{B}$-mtl | 90.5 | 86.5 | 91.1 | 91.1 | 89.8 |
| Meta-teacher | 92.5 | 87.0 | 91.1 | 89.2 | 89.9 |
| $\text{BERT}_\text{B}\text{-single} \xrightarrow{\text{TinyBERT-KD}} \text{BERT}_\text{S}$ | 83.4 | 83.2 | 89.2 | 91.1 | 86.7 |
| $\text{BERT}_\text{B}\text{-mix} \xrightarrow{\text{TinyBERT-KD}} \text{BERT}_\text{S}$ | 88.4 | 81.6 | 89.7 | 89.7 | 87.3 |
| $\text{BERT}_\text{B}\text{-mtl} \xrightarrow{\text{TinyBERT-KD}} \text{BERT}_\text{S}$ | 90.5 | 81.6 | 88.7 | 90.1 | 87.7 |
| $\text{Multi-teachers} \xrightarrow{\text{MTN-KD}} \text{BERT}_\text{S}$ | 83.9 | 78.4 | 88.7 | 87.7 | 84.7 |
| $\text{Meta-teacher} \xrightarrow{\text{TinyBERT-KD}} \text{BERT}_\text{S}$ | 89.9 | 84.3 | 87.3 | **91.6** | 88.3 |
| $\text{Meta-teacher} \xrightarrow{\text{Meta-distillation}} \text{BERT}_\text{S}$ | **91.5** | **86.5** | **90.1** | 89.7 | **89.4** |

✓ Results on Amazon (no fiction domain data when training the meta-teacher)

| Method | Accuracy |
|---|---|
| $\text{BERT}_\text{B}$-s (fiction) | 82.2% |
| Meta-teacher (w/o fiction) | 81.6% |
| $\text{BERT}_\text{B}\text{-s (fiction)} \xrightarrow{\text{TinyBERT-KD}} \text{BERT}_\text{S}$ | 78.8% |
| $\text{BERT}_\text{B}\text{-s (govern)} \xrightarrow{\text{TinyBERT-KD}} \text{BERT}_\text{S}$ | 75.3% |
| $\text{BERT}_\text{B}\text{-s (telephone)} \xrightarrow{\text{TinyBERT-KD}} \text{BERT}_\text{S}$ | 75.6% |
| $\text{BERT}_\text{B}\text{-s (slate)} \xrightarrow{\text{TinyBERT-KD}} \text{BERT}_\text{S}$ | 77.1% |
| $\text{BERT}_\text{B}\text{-s (travel)} \xrightarrow{\text{TinyBERT-KD}} \text{BERT}_\text{S}$ | 74.1% |
| $\text{Meta-teacher} \xrightarrow{\text{TinyBERT-KD}} \text{BERT}_\text{S}$ | **78.2%** |

# Conclusion

- ✓We present the Meta-KD framework for knowledge distillation across domains.
- ✓Experiments confirm the effectiveness of Meta-KD over various NLP tasks.
- ✓Future work includes:
  - ✓Using Meta-KD in other application scenarios
  - ✓Applying other meta-learning techniques to knowledge distillation for PLMs

# THANKS

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