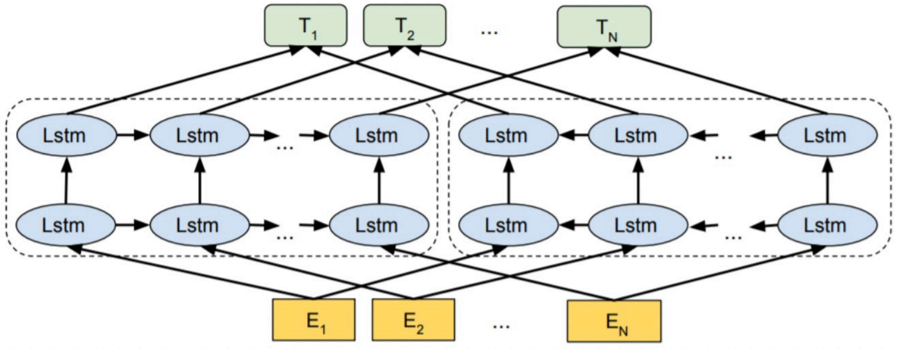
Meta Fine-Tuning Neural Language Models for Multi-Domain Text Mining

Chengyu Wang[1], Minghui Qiu[1], Jun Huang[1], Xiaofeng He[2]

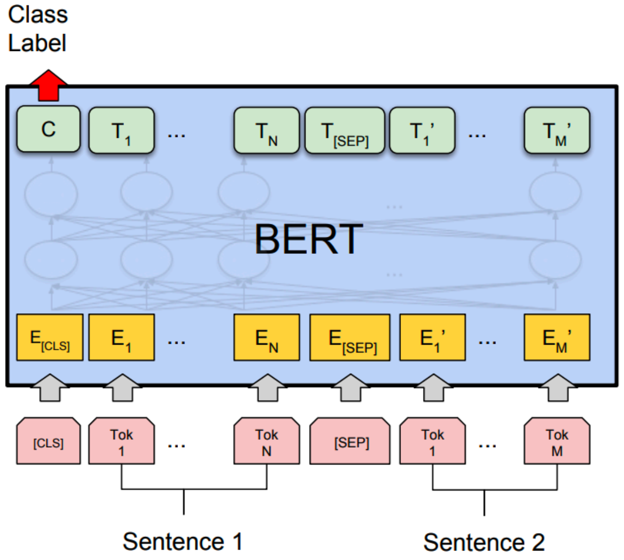
[1] Alibaba Group [2] East China Normal University

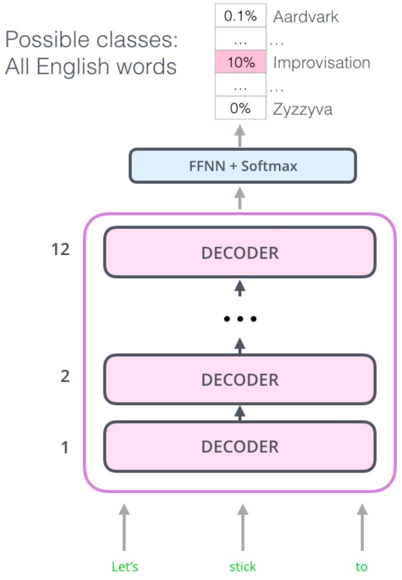
Introduction

✓Pre-trained language models have achieved significant success in NLP.



ELMo





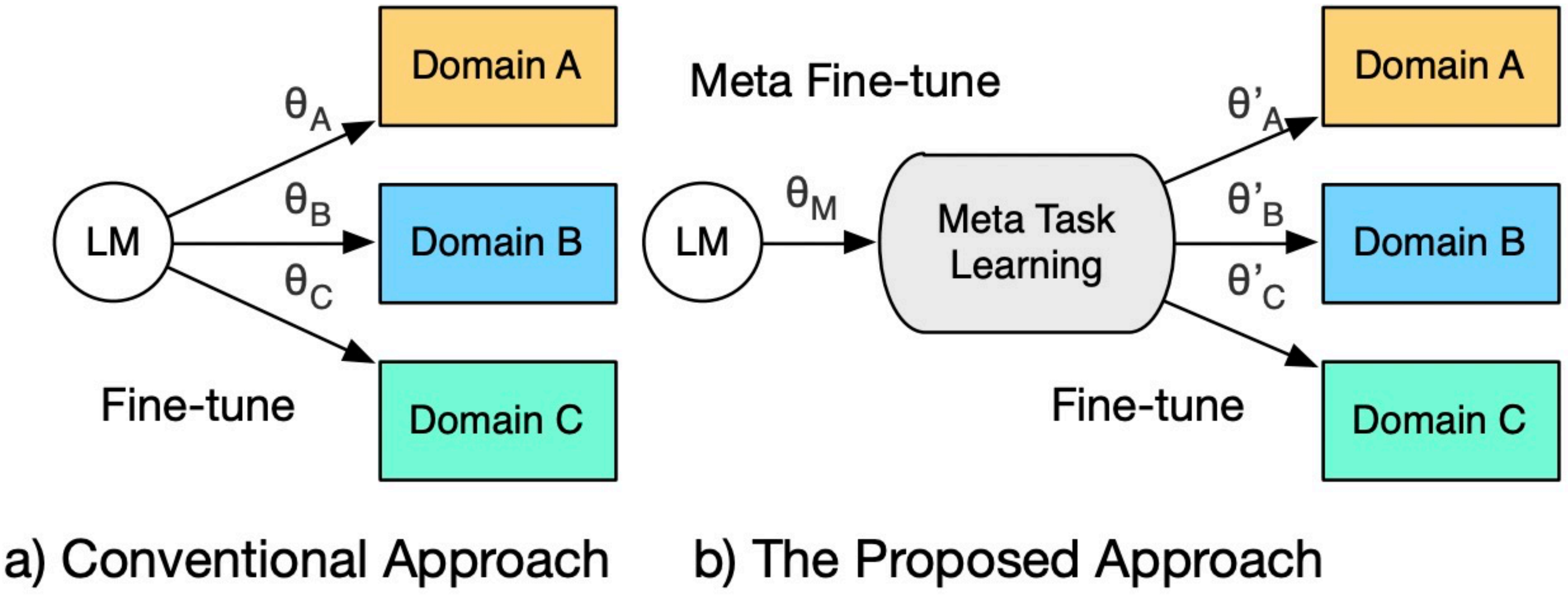
GPT-2

✓A learning gap exists between pre-training and fine-tuning.

✓For a group of similar tasks, parameters of all task-specific models are initialized from the same pre-trained language model.

Introduction

✓Our solution: meta fine-tuning

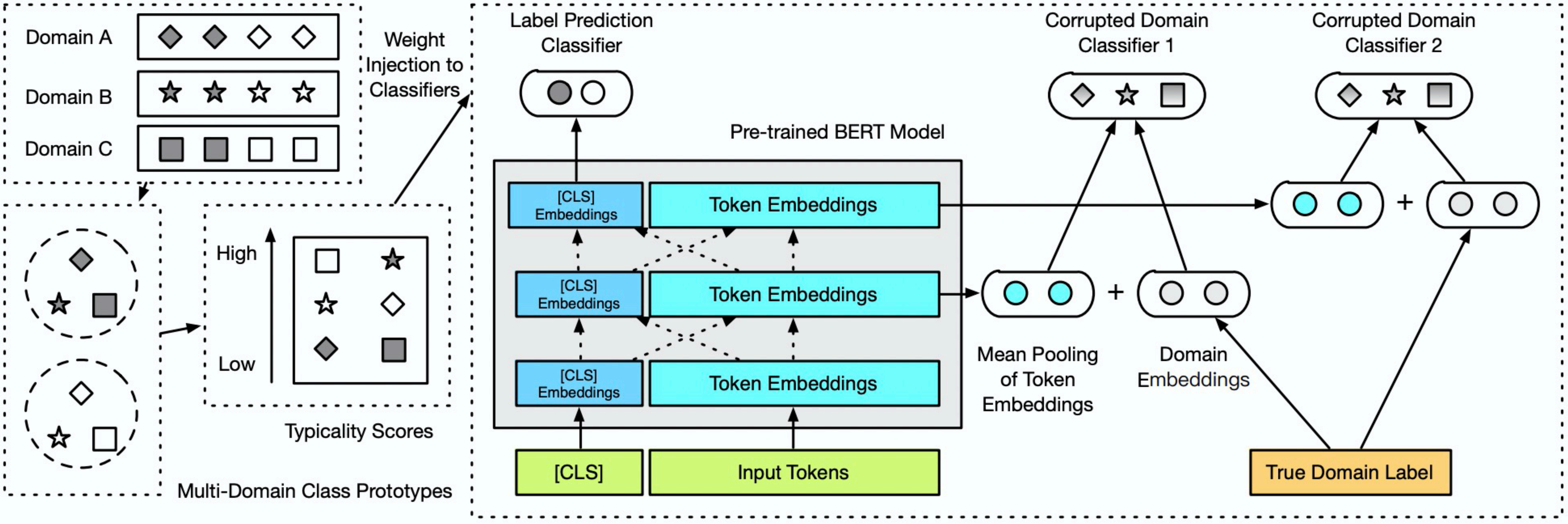


- Target of meta fine-tuning: learning the transferable knowledge across all domains

Key Ideas of Meta Fine-tuning

✓Learning from Typicality

✓Learning Domain-invariant Representations

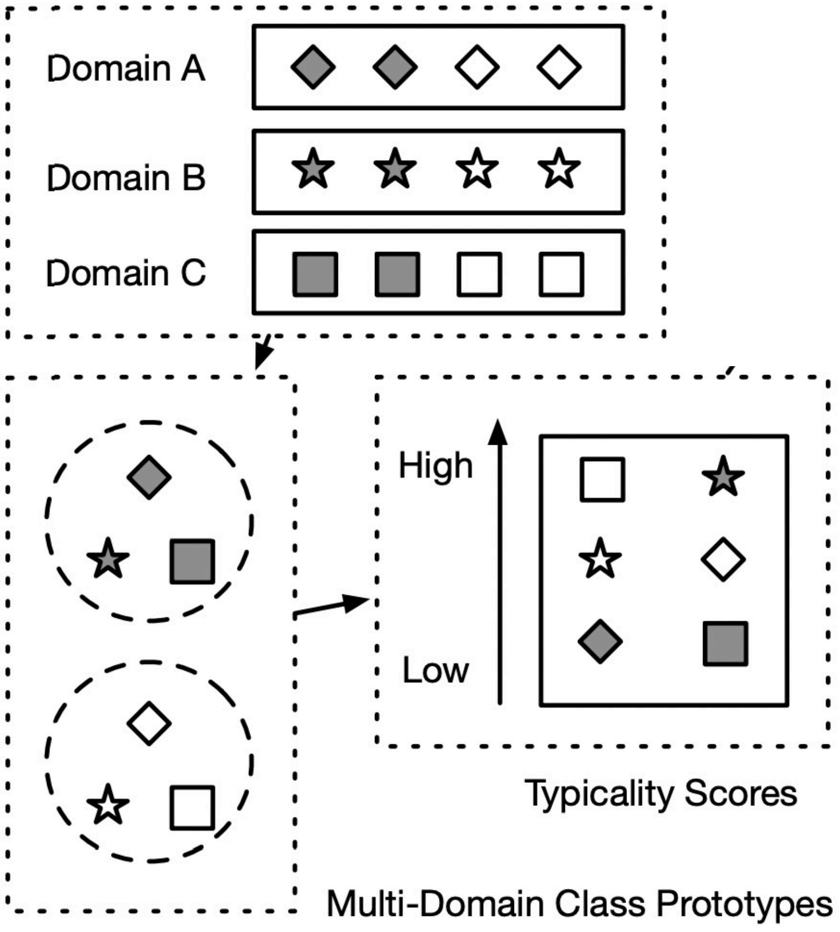


Learning from Typicality

✓An instance is typical if and only if it is:

- Close to its in-domain class centroid in the embedding space
- Not far from its out-of-domain class centroids in the embedding space

$$t_i^k = \alpha \cos(\mathcal{E}(x_i^k), \mathbf{c}_m^k) + \frac{1-\alpha}{K-1} \cdot \sum_{\tilde{k}=1}^{K} \mathbf{1}_{(\tilde{k} \neq k)} \cos(\mathcal{E}(x_i^k), \mathbf{c}_m^{\tilde{k}})$$



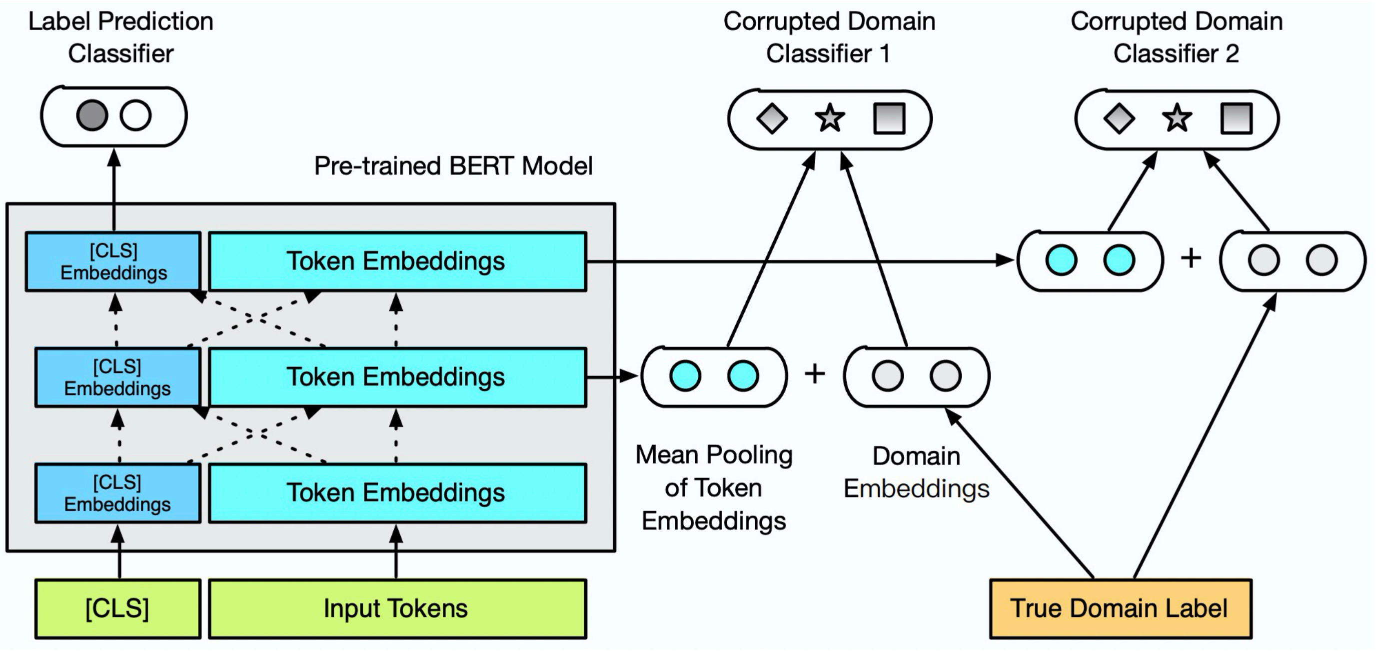
Learning Domain-invariant Representations

✓Corrupted domain classifiers

- Input: true domain embeddings, BERT embeddings
- Output: corrupted domain labels

Goal: forcing BERT to hide domain-specific information

$$\mathcal{L}_{SDC} = -\frac{1}{|L_s| \cdot |\mathcal{D}|} \sum_{(x_i^k, y_i^k) \in \mathcal{D}} \sum_{l \in L_s} \sum_{k=1}^{K} \mathbf{1}_{(k=z_i)} t_i^k \cdot \log \delta_{z_i}(\mathbf{h}_l(x_i^k) + \mathcal{E}_D(k))$$



Experiments

✓Sentence pair classification (MNLI)

Method	Telephone	Government	Slate	Travel	Fiction	Average
BERT (S)	82.5	84.9	78.2	83.1	82.0	82.1
BERT (Mix)	83.1	85.2	79.3	85.1	82.4	83.0
BERT (MTL)	83.8	86.1	**80.2**	85.2	83.6	83.8
BERT (Adv)	81.9	84.2	79.8	82.0	82.2	82.0
MFT (DC)	**84.2**	**86.3**	**80.2**	**85.8**	**84.0**	**84.1**
MFT (TW)	83.8	**86.5**	**81.3**	83.7	**84.4**	**83.9**
MFT (Full)	**84.6**	**86.3**	**81.5**	**85.4**	**84.6**	**84.5**

✓Sentence classification (Amazon Reviews)

Method	Book	DVD	Elec.	Kit.	Avg.
BERT (S)	90.7	88.2	89.0	85.7	88.4
BERT (Mix)	91.8	**89.4**	87.8	88.4	89.3
BERT (MTL)	**92.2**	89.0	88.3	88.2	89.0
BERT (Adv)	89.3	87.4	86.5	86.7	87.5
MFT (DC)	90.6	**89.4**	**92.5**	**88.7**	**90.3**
MFT (TW)	90.4	88.9	**94.5**	**89.1**	**90.7**
MFT (Full)	91.2	88.8	**94.8**	**89.2**	**91.0**

Experiments

✓Few-shot learning

Genre	With MFT? No	With MFT? Yes	Improvement	With MFT? No	With MFT? Yes	Improvement	With MFT? No	With MFT? Yes	Improvement
Training data	**5%** of the original			**10%** of the original			**20%** of the original		
Telephone	70.5	74.7	**4.2%**↑	74.1	76.4	**2.3%**↑	75.9	79.8	**3.9%**↑
Government	76.5	78.1	**1.6%**↑	78.8	81.0	**2.2%**↑	80.5	82.9	**2.4%**↑
Slate	64.2	69.8	**5.7%**↑	67.6	71.8	**4.2%**↑	71.8	74.1	**2.3%**↑
Travel	71.9	75.4	**3.5%**↑	74.8	78.1	**3.3%**↑	78.3	80.3	**2.0%**↑
Fiction	69.7	73.8	**4.1%**↑	73.3	76.6	**3.3%**↑	76.2	78.4	**2.2%**↑
Average	70.5	74.4	**3.9%**↑	73.7	76.8	**3.1%**↑	76.5	79.1	**2.6%**↑

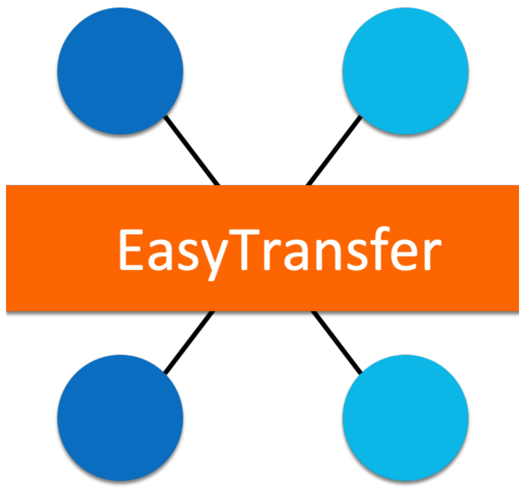
✓Case study

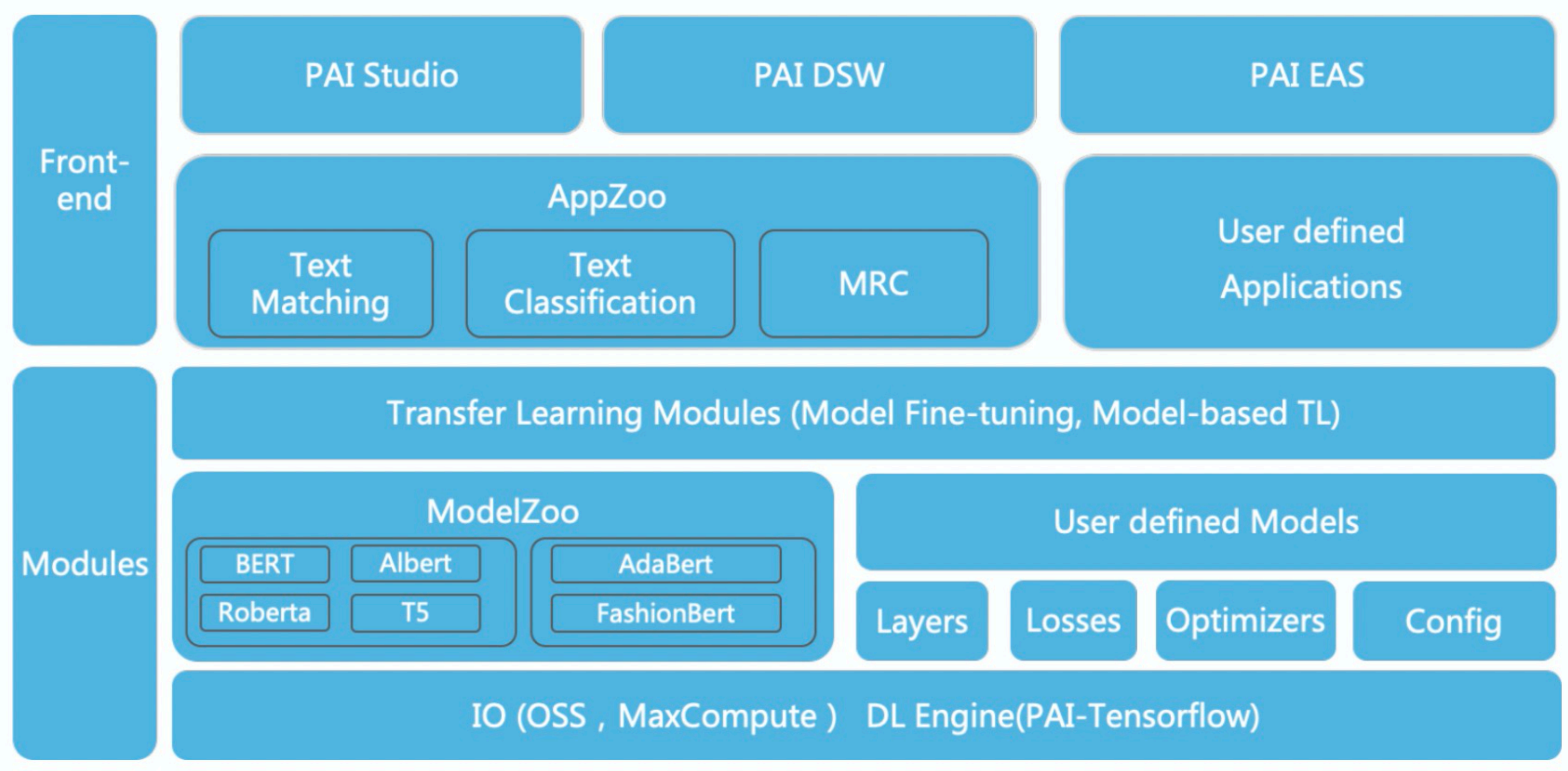
Typicality	Domain	Label	Review Text
Low	Book	NEG	More hate books. How could anyone write anything so wrong.
	Kitchen	NEG	The spoon handle is crooked and there are marks/damage to the wood. Avoid.
	Kitchen	NEG	The glass is quite fragile. I had two breaks.
High	Kitchen	POS	I would recommend them to everyone..and at their price, it's a HUGE DEAL!
	Electronics	NEG	What a waste of money. For $300 you shouldn't HAVE to buy a protection plan for...
	Electronics	NEG	Do not waste your money, this was under recommendations, but I would NOT...

Open Source

✓Meta Fine-tuning is integrated into the EasyTransfer library.

A transfer learning framework for NLP applications





https://github.com/alibaba/EasyTransfer

THANKS

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