国际人工智能会议
AAAI 2021 论文北京预讲会

面向数据稀缺领域BERT知识蒸馏的自动增强方法

冯玲云[1] 邱明辉[2] 李雅亮[2] 郑海涛[1] 沈颖[3*]

[1] 清华大学 [2] 阿里巴巴达摩院 [3] 中山大学

预训练模型

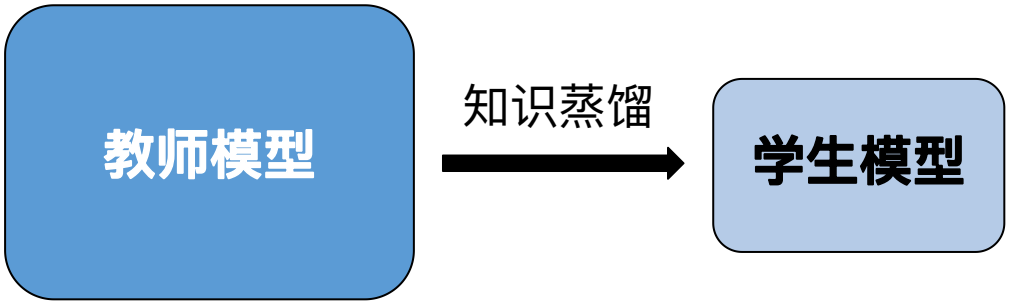
- 预训练模型在自然语言处理任务当中展示了非凡的性能。
- 这些模型中大量的参数导致较高的存储和计算成本，使得在资源受限的应用场景中性能受到制约。

预训练模型

- 预训练模型在自然语言处理任务当中展示了非凡的性能。
- 这些模型中大量的参数导致较高的存储和计算成本，使得在资源受限的应用场景中性能受到制约。

知识蒸馏

- 知识蒸馏的基本思想是在保留教师模型知识的同时，将大的BERT模型压缩为小的学生模型。
- 可有效降低存储和计算成本，并加快推理时间。



知识蒸馏

Pros

- 知识蒸馏的基本思想是在保留教师模型知识的同时，将大BERT模型压缩为小的学生模型。
- 可有效降低存储和计算成本，并加快推理时间。

Cons

- 对于缺少训练数据的目标域，教师模型往往难以将有用的知识传递给学生模型，导致学生模型的性能下降。

数据增强

数据增强（Data augmentation）是处理数据稀缺问题的常用策略。它通过已标记的训练数据来生成新数据，从而扩充目标域的数据。

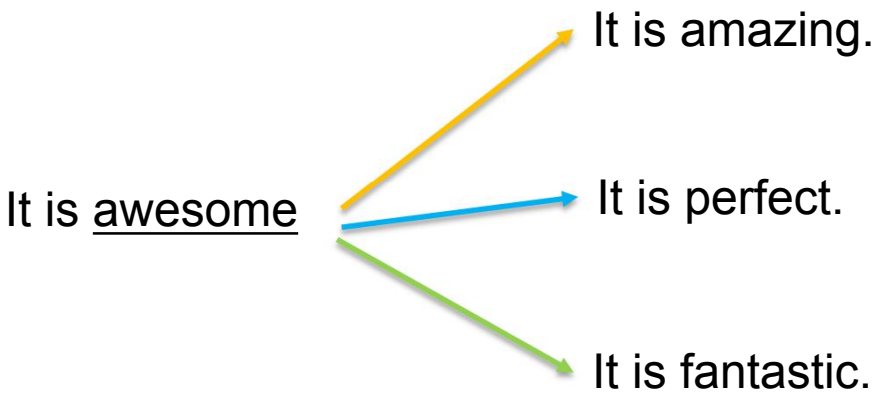
现有的基于数据增强的蒸馏方法大多是手动设计或预定义的：

- **Heuristics based methods** 基于启发式的方法
  Synonym replacement 同义词替换, random insertion随机插入, random swap随机交换, random deletion随机删除
- **Generation based methods** 基于生成的方法
  - Variational Auto-Encoder (VAE)
  - Round-trip translation
  - Paraphrasing
  - Data noising
  - Contextual augmentation

数据增强

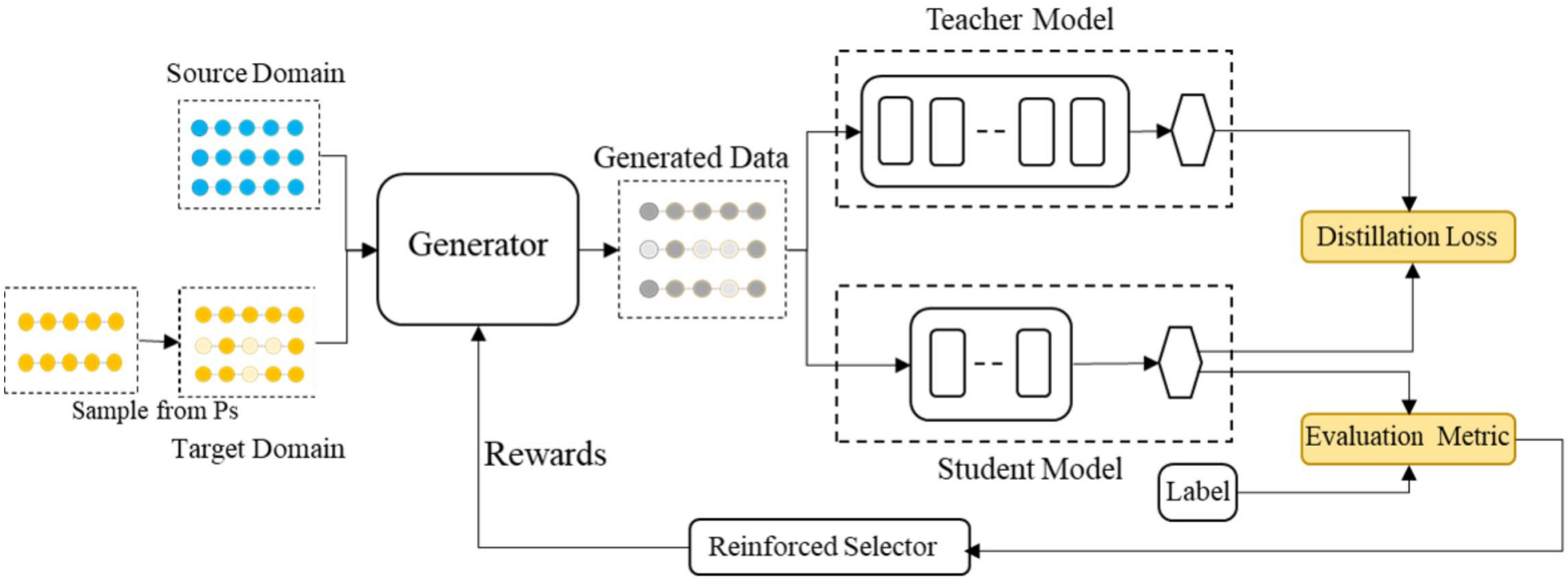
数据增强（Data augmentation）是处理数据稀缺问题的常用策略。
它通过已标记的训练数据来生成新数据，从而扩充目标域的数据。

现有的基于数据增强的蒸馏方法大多是手动设计或预定义的：



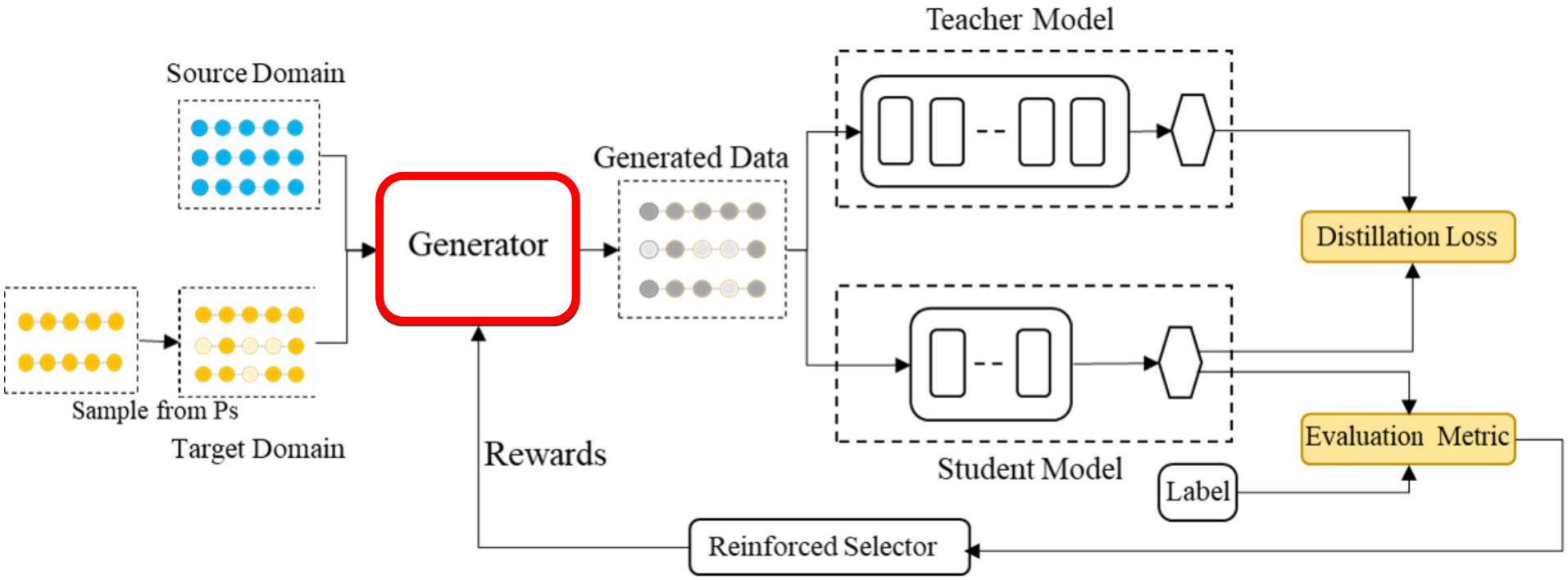
为BERT知识蒸馏设计有效的数据增强方法的研究还需进一步加深

模型概览



本研究提出了一种方法（**L2A**），以实现数据稀疏域BERT知识蒸馏的数据增强。

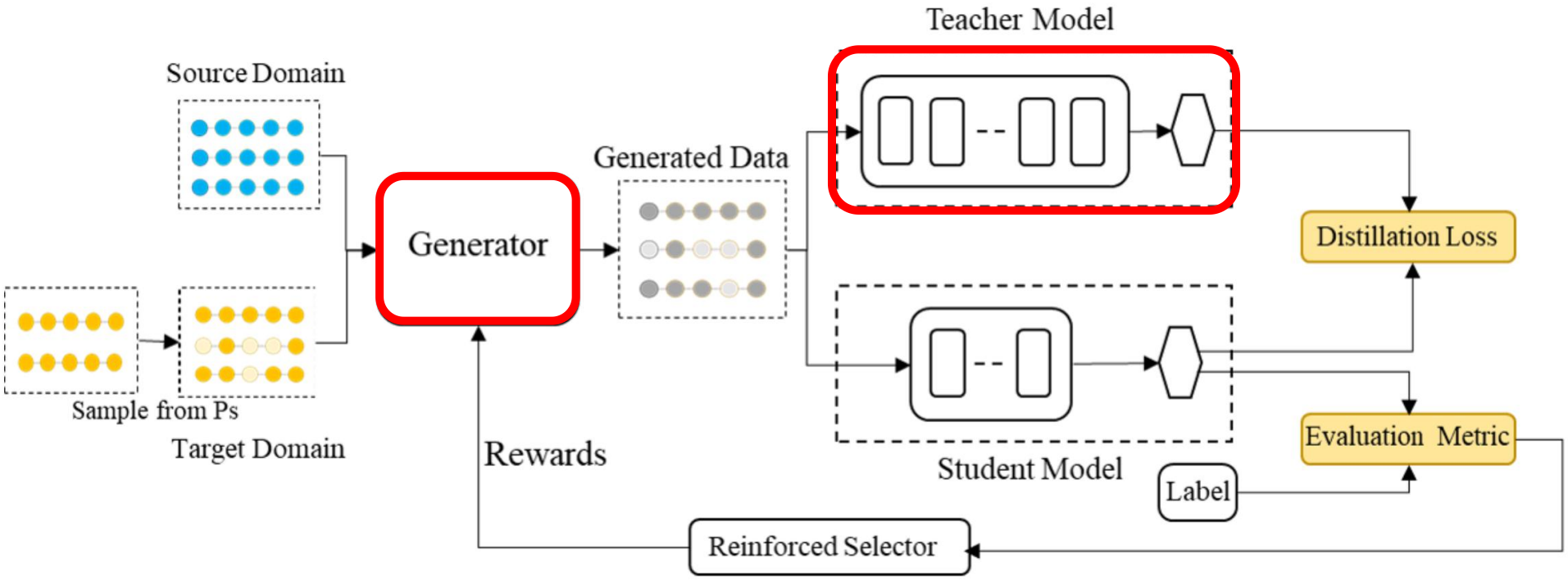
模型概览



本研究提出了一种方法（L2A），以实现数据稀疏域BERT知识蒸馏的数据增强。

- 数据生成器从源域和目标域生成数据，以供教师模块指导学生模块。

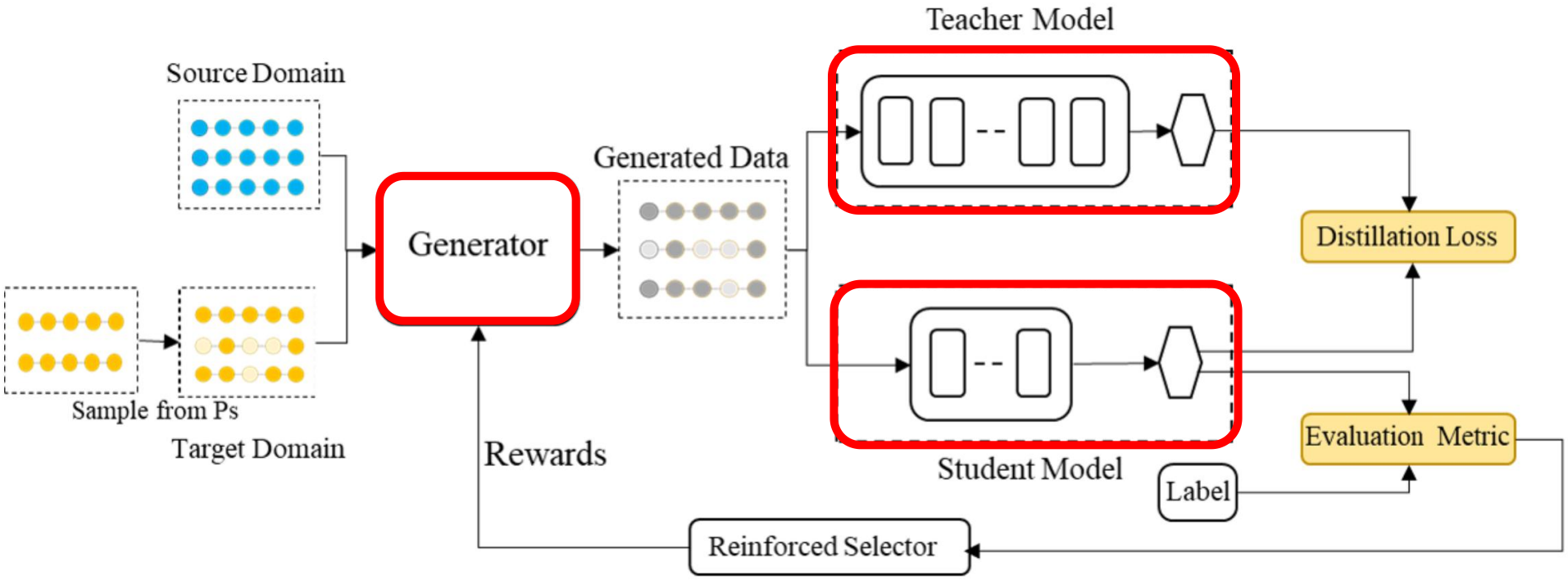
模型概览



本研究提出了一种方法（L2A），以实现数据稀疏域BERT知识蒸馏的数据增强。

- 数据生成器从源域和目标域生成数据，以供教师模块指导学生模块。

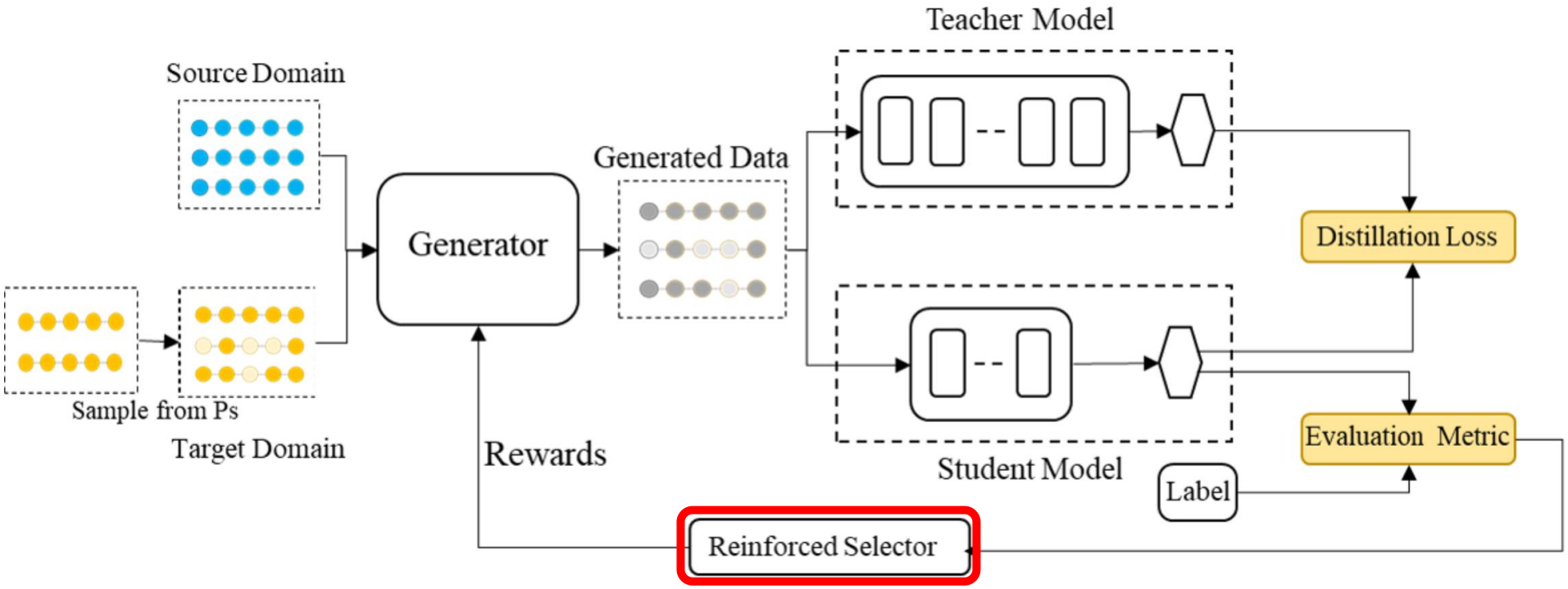
模型概览



本研究提出了一种方法（L2A），以实现数据稀疏域BERT知识蒸馏的数据增强。

- 数据生成器从源域和目标域生成数据，以供教师模块指导学生模块。

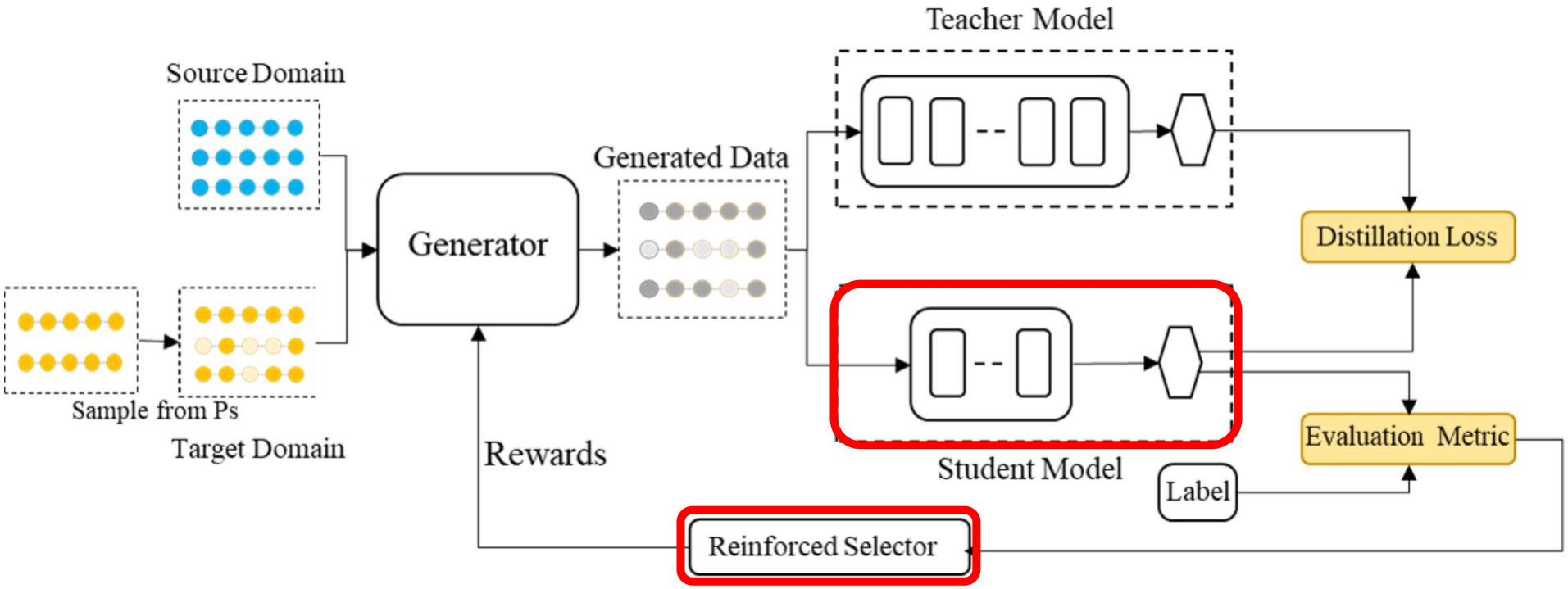
模型概览



本研究提出了一种方法（L2A），以实现数据稀疏域BERT知识蒸馏的数据增强。

- 数据生成器从源域和目标域生成数据，以供教师模块指导学生模块。
- 强化选择器根据学生模块的性能改进数据生成器的增强策略。

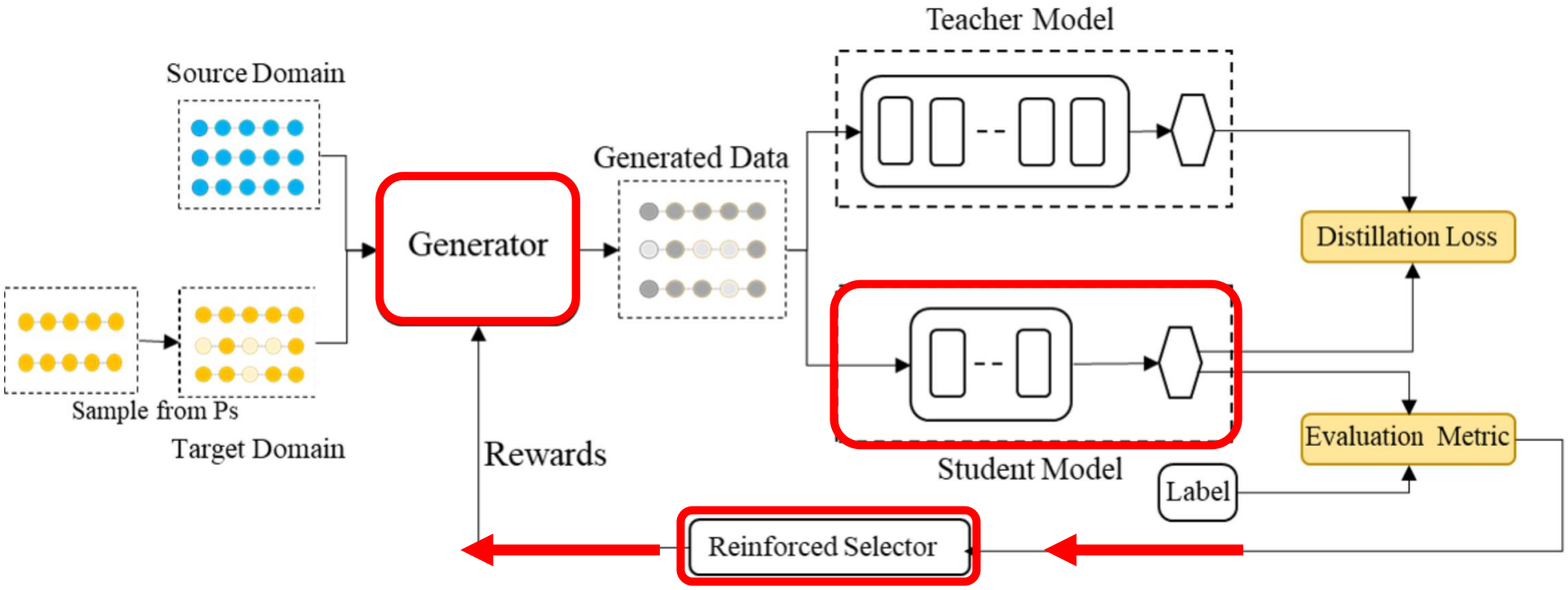
模型概览



本研究提出了一种方法（L2A），以实现数据稀疏域BERT知识蒸馏的数据增强。

- 数据生成器从源域和目标域生成数据，以供教师模块指导学生模块。
- 强化选择器根据学生模块的性能改进数据生成器的增强策略。

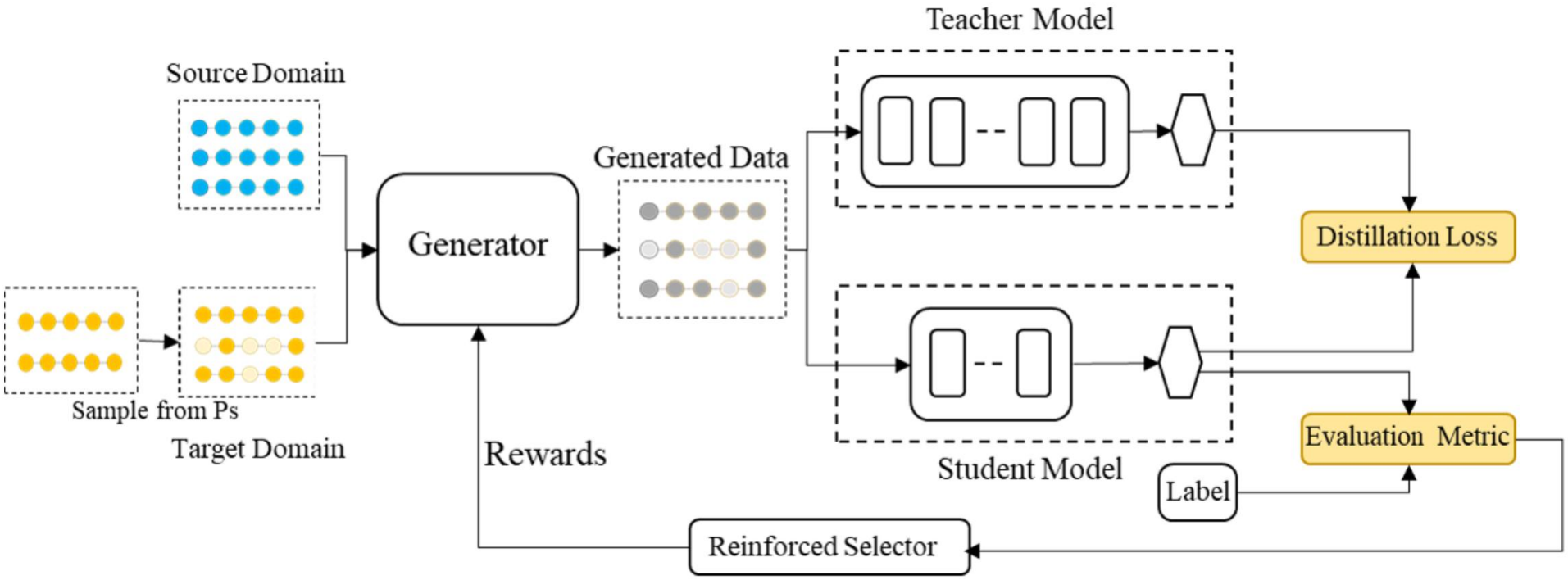
模型概览



本研究提出了一种方法（L2A），以实现数据稀疏域BERT知识蒸馏的数据增强。

- 数据生成器从源域和目标域生成数据，以供教师模块指导学生模块。
- 强化选择器根据学生模块的性能改进数据生成器的增强策略。

模型概览



本研究提出了一种方法（L2A），以实现数据稀疏域BERT知识蒸馏的数据增强。

- 数据生成器从源域和目标域生成数据，以供教师模块指导学生模块。
- 强化选择器根据学生模块的性能改进数据生成器的增强策略。

这些模块以交互方式工作，共同提高目标领域学生模块的性能。

知识蒸馏

注意力层蒸馏 $\mathcal{L}_{att} = \frac{1}{h}\sum_{i=1}^{h} MSE(A_i^s - A_i^t),$

A_i 表示BERT最后一层的第i 个self-attention head对应的注意力矩阵，h是attention heads的数量

隐含层蒸馏 $\mathcal{L}_{hidden} = MSE(H^s W - H^t),$

H_s和H_t分别表示学生网络和教师网络的最后一层的输出

Dark knowledge蒸馏

$$\mathcal{L}_{dark} = -\sum_i \frac{\exp(g_i^t/T_{KD})}{\sum_j \exp(g_j^t/T_{KD})} \log \frac{\exp(g_i^s/T_{KD})}{\sum_j \exp(g_j^s/T_{KD})},$$

学生网络BERT最后一层预测层的输出logit g_s，与教师网络logit g_t之间的软交叉熵损失 T_{KD}为控制输出分布平滑度的超参数

最终知识蒸馏损失 $\mathcal{L}_{KD} = \mathcal{L}_{att} + \mathcal{L}_{hidden} + \mathcal{L}_{dark}.$

数据增强–数据生成器

我们训练数据生成器 $G_\theta(z|x)$，以生成用于学生模块的数据增强样本，从而更好地向老师模块学习。

强化学习损失： $\mathcal{L}_{RL,\theta} = -\mathbb{E}_{z\sim G_\theta(z|x)}[R_\phi(z)] - \alpha\mathbb{H}(P_{G_\theta(z|x)})$,

$R_\phi(z)$ ：Reward函数

$P_{G_\theta(z|x)}$ ：生成器G_θ基于原始样本 x 生成新样本 z 的概率

$\mathbb{H}(P_{G_\theta(z|x)})$ ：正则项，控制生成样本的多样性

RAML损失 (Reward Augmented Maximum Likelihood) (Norouzi et al. 2016):

$$\mathcal{L}_{RAML,\theta} = KL(Q_\phi(z)||P_{G_\theta}(z|x)) + constant$$
$$\propto -\mathbb{E}_{z\sim Q_\phi(z)}[\log P_{G_\theta}(z|x)],$$

$Q_\phi(z)$：指数回报分布, $Q_\phi(z) = \frac{1}{Z}\exp(R_\phi(z)/\alpha)$

当生成器分布$P_{G_\theta}(z|x)$=指数回报分布 $Q_\phi(z)$ 时取到最优值

直接从指数回报分布$Q_\phi(z)$进行采样很困难，因此我们定义了一个静态分布$P_s(z)$，然后设计奖励函数$R_\phi(z)$来训练数据增强模型：

L2A（知识蒸馏中数据增强任务）损失函数：

$$\mathcal{L}_{L2A,\theta} \propto -\mathbb{E}_{z\sim P_s(z)}[R_\phi(z)\log P_{G_\theta}(z|x)]$$
$$= -\mathbb{E}_{z\sim P_s(z)}[(\log P_s(z) + \pi_\varphi(z))\log P_{G_\theta}(z|x)].$$

sampling probability
抽样概率

受强化选择器控制的policy策略，
$\pi_\varphi(z) \in [0,1]$

约束搜索空间Constrained Search Space

减少搜索空间，使训练过程稳定

$$P_s(z|x) = P(d, o, w|x) = P(d|x)P(o|d, x)P(w|o, d, x).$$

$$P(d|x) = \frac{exp\{-d/\alpha\}c(d, m)}{\sum_{e=0}^{m} exp\{-e/\alpha\}c(e, m)},$$

对编辑距离 *d* 进行采样

$$P(o|d, x) = d/m.$$

根据采样的编辑距离
选择位置 *o* 进行替换

$$P(w|o, d, x) = \frac{\exp(P_{\text{BERT}}(w)/T)}{\sum_j \exp(P_{\text{BERT}}(w_j)/T)},$$

在位置 *o* 填充单词 *w*

强化选择器Reinforced Selector

- 为每个增强数据提供评估
- 利用学生模型结果的反馈进行自动更新

State状态：教师模型和学生模型的输出

Action动作：每个输入样本的二元决策（0或1）

Reward奖励：

$$r_t = L(y_i, f^s(x_i)) - L'(y_i, f^s(x_i)),$$

更新后的学生模型评估结果　　先前学生模型的评估结果

对于分类任务，将 L 设置为目标域验证数据的准确性。

对于回归任务，将L设置为预测得分和真实得分之间的相关系数。

- 对于episode中的每个batch，累积reward为：

$$r(\tau) = \sum_{k=0}^{T-t} \gamma^k r_{t+k},$$

在时间T内的动作序列　　discount factor

- 强化选择器根据policy策略π_φ执行操作，最大程度地提升reward

$$\varphi^* = \arg\max_{\varphi} \mathbb{E}_{\tau \sim \pi_\varphi(\tau)} r(\tau).$$

训练过程

Algorithm 1: Learning to Augment for Data-Scarce Domain BERT Compression

Require : training set $\mathcal{D}$, validation data $\mathcal{D}^v$

Initialize the KD module and reinforced selector;
Construct the distribution P_s using Eq. 10;
Sample from P_s and get training data D';
for *each batch* x_b *in* D' **do**
    Obtain teacher model output $f^t(x_b)$, student model output $f^s(x_b)$ and get state s_b;
    Augment data using Eq. 9 and obtain action a_b;
    Update the student model using Eq. 1;
    Obtain the reward r_b using Eq. 13;
    Store (s_b, a_b, r_b) in episode history H;
end
for *each turple* (s_b, a_b, r_b) *in the history H* **do**
    Obtain the accumulated reward using Eq. 14;
    Update policy π_φ using Eq. 15.
end

数据集

为模拟数据稀缺领域

Natural Language Inference (NLI)

- Source: MultiNLI
- Target: SciTail

每个类随机选择40个实例

Paraphrase Identification (PI)

- Source: Quora question pairs
- Target: CIKM AnalytiCup 2018

目标域相对较小，保持不变

Text classification

- Source: SST-2
- Target: RT

每个类随机选择40个实例

Review helpfulness prediction

- Amazon review dataset
- Source: Electronics domain
- Target: Watches domain

以原始数据的1%作为训练数据

基线模型

BERT 压缩方法

- MINILM (Wang et al. 2020)
- DistIBERT (Sanh et al. 2019)
- BERT–PKD (Sun et al. 2019)
- TinyBERT (Jiao et al. 2019)
- BiLSTMSOFT (Tang et al. 2019)

也进行了数据增强

数据增强方法

- EDA (Wei and Zou. 2019)
- CBERT (Wu et al.2019)

消融实验

- L2A w/o src
- L2A w/o tgt
- L2A w/o Latt
- L2A w/o Lhidden
- L2A w/o Ldark

结果分析

Method	Model Size	NLI		PI		Text Classification		Regression Task	
		ACC	F1	ACC	F1	ACC	F1	P.	S.
Student-FT	14.5M	0.7380	0.6928	0.8844	0.7435	0.7188	0.7309	0.3878	0.3225
$BiLSTM_{SOFT}$	10.1M	0.5890	0.5006	0.8622	0.7009	0.4839	0.6522	-	-
DistilBERT	52.2M	0.6891	0.5648	0.8991	0.7775	0.6776	0.6966	0.4048	0.3343
$BERT_4$-PKD	52.2M	0.5809	0.5819	0.9041	0.7956	0.6173	0.5189	0.4466	0.3778
$BERT_6$-PKD	67M	0.6980	0.6201	0.9060	0.8040	0.6370	0.6311	0.4482	0.3923
MINILM	33M	0.7512	0.6314	0.9024	0.7858	0.7020	0.7022	0.4441	0.4132
TinyBERT	14.5M	0.7319	0.6143	0.8787	0.7274	0.7235	0.7392	0.2653	0.2139
EDA	14.5M	0.7465	0.6375	0.9030	0.7920	0.7254	0.7428	0.4554	0.3887
CBERT		0.7469	0.6820	0.8925	0.7654	0.7366	0.7020	0.4680	0.3891
L2A	14.5M	**0.7827**	**0.7152**	**0.9195**	**0.8275**	**0.7798**	**0.7614**	**0.4852**	**0.4204**

L2A优胜于基础学生模型、BERT知识蒸馏方法和启发式数据增强方法。

Experiments

Method	Model Size	NLI		PI		Text Classification		Regression Task	
		ACC	F1	ACC	F1	ACC	F1	P.	S.
Teacher-FT	109M	0.7639	0.6935	**0.9225**	**0.8359**	0.7573	0.7604	**0.4874**	**0.4264**
Student-FT	14.5M	0.7380	0.6928	0.8844	0.7435	0.7188	0.7309	0.3878	0.3225
$\text{L2A}_{w/o\ tgt}$	14.5M	0.7714	0.7025	0.8549	0.6730	0.7610	0.7442	0.4715	0.4119
$\text{L2A}_{w/o\ src}$		0.7615	0.6955	0.9144	0.8165	0.7526	0.7523	0.4757	0.3992
L2A		**0.7827**	**0.7152**	0.9195	0.8275	**0.7798**	**0.7614**	0.4852	0.4204

Teacher-FT和Student-FT分别为teacher和student的预训练模型加上具体任务的训练数据来训练新的模型，即fine tuning
w/o src 训练和测试都用target domain的数据
w/o tgt训练用source domain数据，测试用target domain数据

- L2A在所有任务中均胜过L2A w/o src和L2A w/o tgt，这表明基于源域或目标域信息的数据增强可帮助提高模型性能。
- 在数据稀缺的领域，适当的数据扩充可以显着提高学生的表现，甚至可以取得与老师相当或更好的成绩。

消融实验–数据多寡的影响

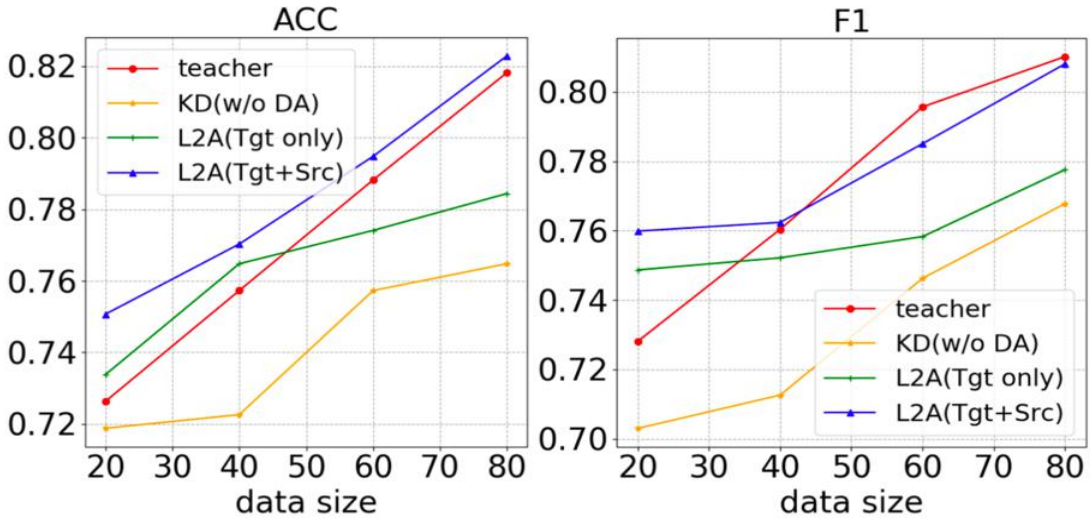


Figure 2: Ablation study on different target domain data sizes.

- 当样本数据量较小时，蒸馏性能的改进愈加明显，甚至优于大型教师模型。
- 表明L2A可有效地帮助BERT进行数据稀缺领域的知识蒸馏。

消融实验–蒸馏函数的影响

Table 3: Ablation study on different distillation objectives.

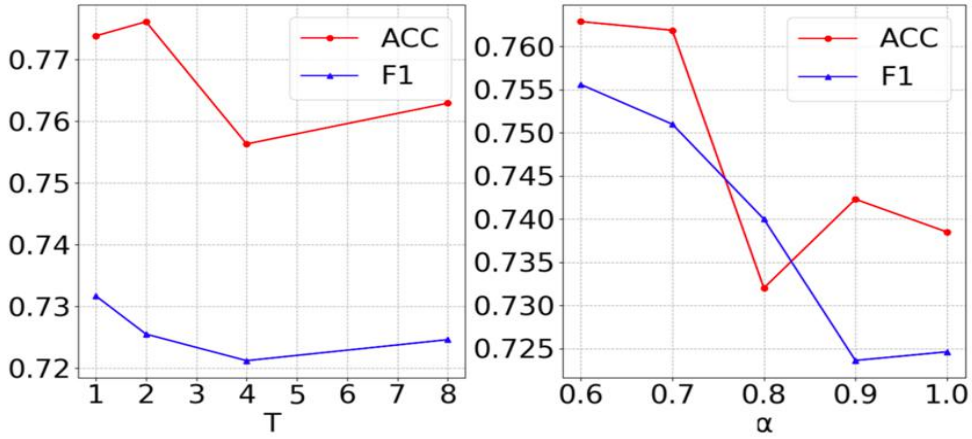
	$\mathcal{L}_{KD}$	w/o $\mathcal{L}_{att}$	w/o $\mathcal{L}_{hidden}$	w/o $\mathcal{L}_{dark}$
ACC	**0.7798**	0.7563	0.7629	0.7647
F1	**0.7614**	0.7506	0.7608	0.7433

- 实验证明，在知识蒸馏中，所使用的的注意力层蒸馏、隐含层蒸馏、Dark knowledge蒸馏都起着一定的作用。
- 该模型通过组合所有蒸馏实现了最佳性能。

消融实验–参数灵敏度分析

$$P_s(z|x) = P(d, o, w|x) = P(d|x)P(o|d, x)P(w|o, d, x).$$

$$P(d|x) = \frac{exp\{-d/\alpha\}c(d, m)}{\sum_{e=0}^{m} exp\{-e/\alpha\}c(e, m)}, \qquad P(w|o, d, x) = \frac{\exp(P_{\text{BERT}}(w)/T)}{\sum_j \exp(P_{\text{BERT}}(w_j)/T)},$$



- T控制着围绕实际数据的搜索空间，最好将T设置为一个较小的值。
- 随着温度 α 变大，增强数据的质量变差。这是因为较高的温度会鼓励生成更多与原始数据分布相距较远的样本，从而导致性能下降。

- 所提出的方法可以利用跨领域信息并根据学生模型的性能自动增强数据。
- 所提出的模型从静态分布中生成样本，极大地减少了搜索空间并使训练过程稳定。
- 在不同任务上进行的实验表明，我们的模型优于包含最新蒸馏方法和数据增强方法在内的基线模型，其性能甚至超过教师模型。

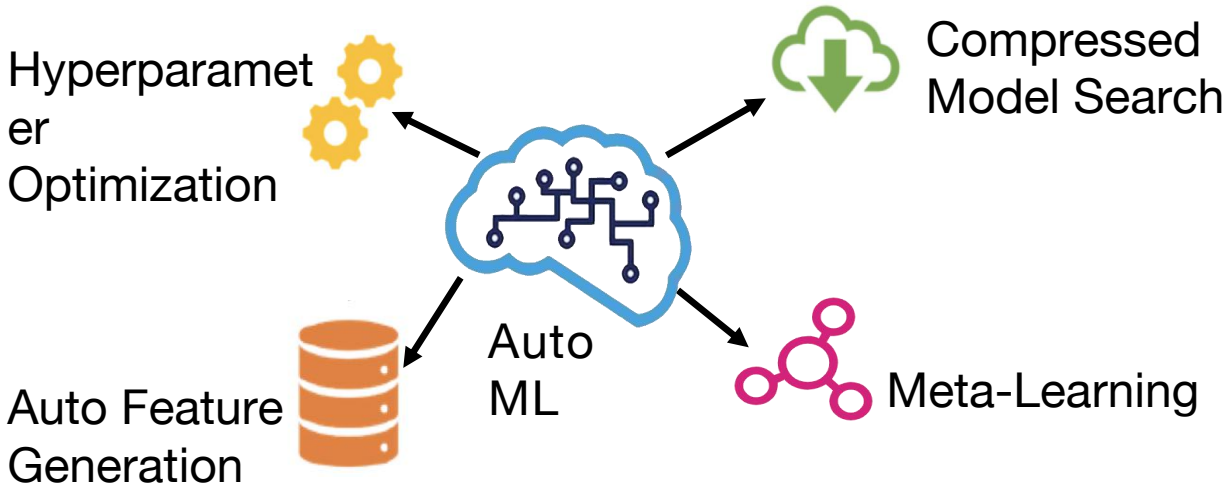
沈 颖

Email:
sheny76@mail.sysu.edu.cn

AutoML: Other Works

Learning to Mutate with Hypergradient Guided Population, NeurIPS 2020.

AdaBERT: Task-Adaptive BERT Compression with D-NAS, IJCAI 2020



Interactive Feature Generation via Learning Adjacency Tensor of Feature Graph.
https://arxiv.org/abs/2007.14573

Automated Relational Meta-learning, ICLR 2020.
https://arxiv.org/abs/2001.00745

国际人工智能会议
AAAI 2021 论文北京预讲会

THANKS

2020.12.19