Data is all you need —— 文本分类任务的数据增强

摘要

当前的自然语言处理任务中，采用大规模预训练模型进行微调已经成为了十分普遍的范式。这样的模型通常需要在海量的无标注文本中进行预训练，例如掩码语言建模(Masked Language Modeling)。大量的研究证明，预训练模型可以很好的将学到的知识迁移到下游任务。但下游任务的训练方式往往不同于预训练，这使得我们猜测模型通过下游任务的监督信号并未完全学习到任务文本中的知识。本项工作中我们采用领域自适应预训练(Domain Adaptive Pre-Training, DAPT)和生成伪标签(Pseudo Labeling, PL)两种方式进行数据增强。实验结果表面，我们的方法在多种模型结构，多种预训练方式上均提高了模型效果，说明这两种方法促使模型学到了下游任务中监督信号无法提供的知识。

1. 背景

大规模预训练模型可以从无监督的文本当中学得知识，并以参数的形式保存在模型当中，这个过程被称为预训练。其中掩码语言建模是最为常见的训练方式：通过将文本中的某个词遮盖，由模型预测被遮盖的词的方式，模型可以学得文本中包含的语言知识。这种训练方式也被称为自监督学习。尽管预训练模型在下游任务上表现出了优越的效果，但由于过大的模型参数，较少的下游任务数据等问题，模型易产生过拟合等现象。为了提高模型的鲁棒性、泛化性和模型的效果，通常采取的方法有：从训练方式上，可以采用对抗训练的方式提高鲁棒性。从模型本身，可以采用对参数的范数进行限制。从数据上，可以采用数据增强的方式。

本文主要讨论数据增强的方法。在视觉领域，对数据进行增强通常可以采用扭曲图像，旋转或者对称等方式。类似的，自然语言处理中，典型的数据增强方法是对词进行替换，删除，或者转换语序等。但由于语言的一些性质，还可以采用一些特殊的增强方式，如回翻译(Back Translation)。但随着大规模预训练模型的流行以及一些强模型(如 DeBERTa)的出现，这些方法被实践证明并不能总是有效。简单的数据增强方式可能并不能对模型提供额外的知识，一些方法甚至可能提供错误的监督信号，从而造成模型效果的下降。在本文中我们采用领域自适应的训练方式，使模型在与下游任务相同领域的文本中继续预训练，以及生成伪标签的方式，通过已训练好的模型生成标签，再将带标签的数据重新并入训练集中。两种方法使得不同预训练方式，不同结构的模型效果均有提升。

本文关注的任务是文本分类任务。不同于传统的文本分类将整个文本归于某个类，我们的任务是对文本中的每个词进行分类。从任务的复杂度角度来讲，我们的任务更加困难和复杂。

2. 方法

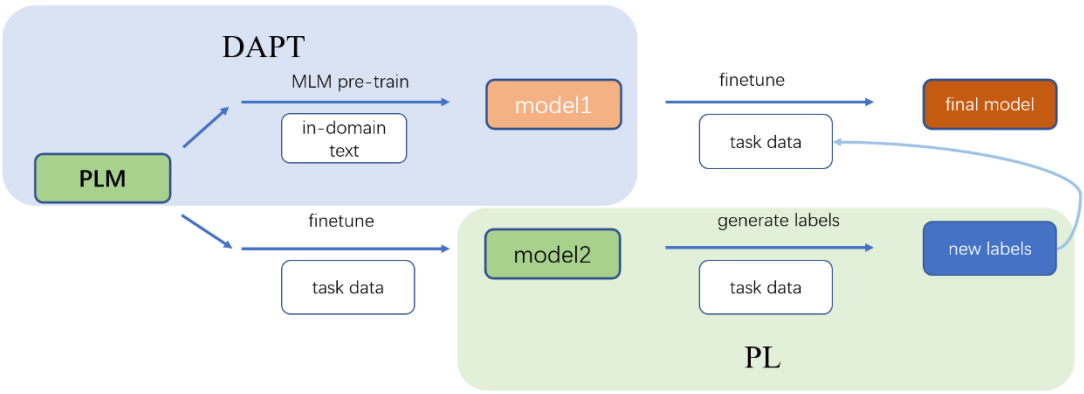

图 1：我们的方法如图示。蓝色区域为 DAPT 步骤，绿色区域为 PL 步骤。

领域自适应预训练(DAPT)是指将预训练模型在与下游任务文本相似的文本上继续进行预训练，通过持续的预训练将与下游任务相关的额外知识注入模型。与 DAPT 类似的还有任务自适应预训练(Task Adaptive Pre-Training, TAPT)，其与 DAPT 主要的不同点在于：TAPT 预训练的文本来自下游任务的训练集，而 DAPT 则是来自同一领域的文本。DAPT 的过程如图 1 所示。预训练模型首先在相同领域文本继续预训练，其训练方式可以采用掩码语言建模，或者其他训练方式，训得的模型再在下游任务上进行微调。

生成伪标签(PL)是指先将预训练模型在下游任务上进行微调，再将微调过的模型在无标签的数据上进行预测，将模型预测所得的结果作为标签，从而得到新的训练数据。PL 的过程如图 1 所示。

我们的方法将 DAPT 和 PL 相结合，首先将预训练模型在相似的文本上继续预训练，对 DAPT 所得的模型在下游任务上进行微调，并利用微调所得的模型生成伪标签并得到新的训练数据。将新的训练数据与原训练集合并得到新的训练集，再利用新的训练集微调 DAPT 所得的模型。整个过程如图 1 所示。

3. 实验

3.1 NBME – Score Clinical Patient Notes 数据集

我们实验所用的数据集来自 Kaggle 平台的比赛 NBME – Score Clinical Patient Notes 的数据集。该数据集由美国医师资格考试委员会(National Board of Medical Examiners, NBME)提供。该任务要求给定一段诊断笔记和一个特征文本，识别诊断笔记中与特征文本对应的关键词。我们将其转化为对每个词的分类任务：如果该词与特征文本对应则为正例，否则为负例。除了标注的数据外，该数据集还提供了 42146 个未标注的段落，我们的 DAPT 在这些未标注的段落上进行训练。

3.2 模型选择

我们选择 DeBERTa 作为基线模型。DeBERTa 模型由微软在 2021 年提出，其采

	dev			test		
	vanilla	DAPT	PL	vanilla	DAPT	PL
RoBERTa	82.07	85.16	84.15	81.73	84.26	83.71
BART	80.09	81.68	82.46	79.67	81.28	81.95
Longformer	80.88	83.02	83.42	80.64	82.73	82.74
Big Bird RoBERTa	77.67	82.49	80.06	76.91	82.58	79.62
DeBERTa	86.17	87.03	86.99	85.32	86.43	86.14
DeBERTa-v3	85.70	86.24	86.68	84.96	85.87	86.09

表 1：对不同模型使用 DAPT 和 PL 后验证集和测试集上的模型效果，vanilla 表示不使用 DAPT 或 PL

用解绑的注意力机制，并在 SuperGLUE 基准上获得了超越人类水平的表现。最新的 DeBERTa-v3 版本采用了类似 ELECTRA 的训练方式并进一步提升了模型性能。作为对比，我们也在采用掩码语言建模训练方式的模型，不同结构的模型上应用了 DAPT 和 PL，实验结果显示这些模型的性能均有提升。

3.3 评估指标

我们采用与比赛相同的评估指标，即 F1 指标。该指标同时考虑了召回率与正确率。由于我们的分类是在词级别进行的，因此 F1 指标的计算也是在词级别进行。

3.4 参数设置

对于 DAPT，我们在未标注的 40k 个段落上训练 10 轮，学习率为 2e-5 并线性降低到 0，批大小为 32，每一轮保存一个模型并评估困惑度(Perplexity, PPL)，最终选择 PPL 最低的模型作为微调的模型。对于微调过程，包括 PL 和未使用 PL 的微调，选择 20%的训练数据作为验证集并训练 5 轮，学习率为 2e-5 并余弦降低到 1e-6，最终选择在验证集上表现最好的模型作为最终模型。由于数据的测试集并未公开，我们将最终模型在 Kaggle 平台上进行测试并获得模型在测试集上的表现。

3.5 实验结果分析

实验结果如表 1 所示。其中大部分模型采用的是掩码语言建模方式进行预训练的，除此之外，BART 还使用了随机删除词，连续文本掩码，打乱句序等预训练方式，DeBERTa-v3 则使用 ELECTRA 风格的预训练方式。模型训练过程的困惑度变化如图 2 所示。尽管 BART 和 DeBERTa-v3 并未采用掩码语言建模方式进行预训练，DAPT 仍能降低模型的困惑度。大部分模型结构采用了原始 Transformer 结构，但 Longformer 采用了局部和全局注意力机制，Big Bird 采用块稀疏注意力机制，DeBERTa 则采用了解绑的注意力机制。DAPT 和 PL 在所有的模型上均极大提升了效果，说明这两种方法对不同的预训练方式，不同的模型结构均有效。

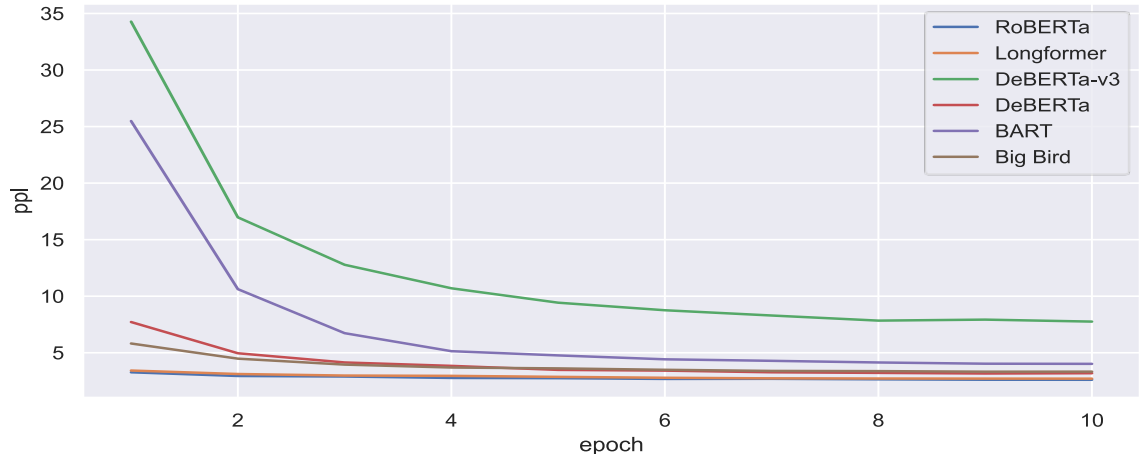

图 2：各模型 DAPT 训练过程中困惑度变化

		dev	test
DeBERTa-v3-large	w/ DAPT & PL	88.91	88.40
	w/ DAPT	88.67	88.08
	w/ PL	88.59	88.05
	-	88.31	87.30

表 2：采用 DeBERTa-v3-large 的消融实验结果，“-”表示未使用数据增强方法

3.6 消融实验

我们使用 DeBERTa-v3-large 作为基线模型，并对两种方法分别进行消融实验，所得结果如表 2 所示，可以发现两种方法效果相差并不大。进一步对对比实验的结果进行分析可发现，不同模型，不同参数量，两种方法的增益大小不同，且相对大小也不同。这两种方法的优劣有待进一步探究。

4. 结论

我们采用领域自适应预训练和生成伪标签两种方式对下游任务进行数据增强，实验表明，对不同的预训练方式，不同结构的模型，我们采用的方法均不同程度地在微调的基础上提高模型表现，说明模型在我们的方法下学到了下游任务监督信号无法提供的知识。

5. 附录

5.1 参考文献

方法主要参考下面两篇文献：

1. Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks
2. Self-training with Noisy Student improves ImageNet classification