# Ad在保留通用知识的同时适应语言模型 e

**Zixuan Ke**[1], **Yijia Shao**[2], **HaoweiLin**[2],**Hu Xu**[3],**LeiShu**[1*] **和 Bing Liu**[11]伊
利诺伊大学芝加哥分校计算机科学系[2]北京大学王选计算机研究所[3]
Meta AI[1]{zke4,liub}@uic.edu[2]{shaoyj, linhaowei}@pku.edu.cn[3]
huxu@fb.com

arXiv:2301.08986v1 [cs.CL] 21 Jan 2023

## 摘要

领域自适应预训练（简称领域自适应训练），也称为后训练，旨在利用特定领域的未标注语料库训练预训练通用语言模型（语言模型），使其适应该领域，从而提升该领域下游任务的表现。然而，现有的领域自适应训练方法在某种意义上是盲目的，因为它们没有明确识别语言模型中哪些知识应予保留，哪些应由领域语料库改变。本文表明，现有方法并非最优，并提出一种新方法，通过以下方式对语言模型中的知识进行更有针对性的适应：(1) 根据注意力头的重要性进行软掩码，以尽可能保留语言模型中的通用知识；(2) 对比通用表征与完整表征（包含通用知识和领域知识），学习融合通用知识和领域特定知识的综合表征。实验结果将证明所提方法的有效性。[1]

## 1 引言

像 BERT（Devlin 等，2019）、RoBERTa（Liu 等，2019）和 GPT-3（Brown 等，2020）这类经过预训练的通用语言模型，已成为几乎所有自然语言处理应用的标准组件。研究人员还发现，使用特定领域的未标注语料库进行领域自适应预训练（简称领域自适应训练），以适配语言模型，可以进一步提升其在该领域下游任务中的表现（Gururangan 等，2020；Xu 等，2019a,b；Sun 等，2019；Alsentzer 等，2019）。请注意，领域自适应预训练也称为后训练（Xu 等，2019a）。

现有的领域自适应训练方法只是将相同的预训练目标（即掩码语言模型 （ MLM ）损失）用于基于领域语料库进一步训练语言模型。这些方法并非最优，因为它们没有明确区分领域语料库应当使语言模型保留哪些内容、更新哪些内容。

本文指出，优秀的领域自适应训练方法需要满足两点。一方面，应尽可能保留语言模型中学到的通用语言知识，因为目标域数据通常规模不足以充分学习通用知识。例如，某些词语及其上下文可能很少出现在特定领域中。仅凭领域数据，无法准确学习与它们相关的知识。当这些词语和上下文出现在下游任务中时，系统就会遇到困难。因此，我们需要依赖语言模型中关于它们的知识。现有的领域自适应训练在更新语言模型时缺乏充分指导，因此这类有用的通用知识可能会受到损害。另一方面，由于多义性（同一个词在不同领域中有不同含义），且不同领域也有各自特有的用词方式和上下文，语言模型应针对目标领域进行专门化或适配。优秀的领域自适应训练应在这两方面之间取得平衡，使语言模型适应目标领域，同时尽量避免损害其中有价值的通用知识。

本文提出了一种新颖的技术，以实现更具知情适应的训练，从而（1）尽可能保留语言模型中的通用知识，并（2）根据需要更新语言模型，使其纳入目标领域的特定领域知识。现有领域自适应训练研究的重点一直放在（2）上。正如我们上文所论证的，（1）同样很重要，因为如果只关注（2），可能会破坏一些有用的通用知识，并导致下游任务的结果不尽如人意。为实现（1），系统应根据每个注意力头对通用知识的重要性，限制其梯度更新

*现任职于 Google Researchleishu@google.com
[1] https://github.com/UIC-Liu-Lab/DGA

头[2] 根据其对通用知识的重要性进行约束，从而尽可能保留语言模型中的通用知识。实现（1）后，（2）便能够改变需要更新的那部分通用知识，使语言模型适应目标领域。[3]

在本文中，我们提出了一种名为 DGA（***D**A-training- **G**eneral knowledge preserva-tionand LM **A**daptation*）的新型模型，以实现这一目的。该方法的核心思想是在使语言模型适应特定领域的同时，保留其中的通用语言知识。然而，如何实现这一点并不明显，即如何找出对通用知识至关重要的参数，以及如何保护这些参数。本文提出了一种新颖的基于代理的方法来实现这些目标。其工作方式如下：DGA 首先通过新提出的代理*KL*散度损失(第3.1节）估计语言模型中每个注意力头的重要性。这一重要性分数反映了各注意力头对通用知识的重要程度。基于这些重要性分数，DGA 执行两项关键功能：第一项功能利用这些分数对梯度更新进行软掩码（而非二值掩码或完全阻断），以避免语言模型中重要的通用知识遭到不必要的破坏。这与剪枝不重要的注意力头(Michel 等人，2019年）有关。不过，正如我们将在第2节中所示，剪枝并不能直接应用于领域自适应训练。所提出的软掩码仅约束训练中的反向梯度流。无论在训练还是推理时，都无需对前向传播进行软掩码。这一点很重要，因为利用完整网络中的知识，有助于最大程度地融合预训练通用知识与目标领域的领域特定知识。第二项功能则对比语言模型中通用知识的表示与完整知识（包括通用知识和领域特定知识）的表示，以学习融合表示（第3.2节）。[4]

总而言之，本文作出了两项关键贡献。

**(1).** 本文提出了知情适应的理念，旨在将目标领域的专门知识融入语言模型，同时尽可能减少对原始语言模型中有用通用知识的损害。

**（2）**。本文提出了一种名为 DGA 的新模型，它具备两项新颖功能，可实现更好的领域自适应训练。DGA 会估算注意力头重要性，以保护语言模型中的重要通用知识；并通过对比通用知识与完整知识，将目标领域中的专门知识整合到语言模型中。

据我们所知，文献中此前尚未报道过这些内容。

我们在 6 个不同领域和 10 个基线上进行了大量实验，以证明所提出的 DGA 的有效性。

## 2 相关工作

**领域自适应预训练（领域自适应训练）**。研究人员已将领域自适应训练应用于许多领域，例如评论（Xu 等，2019a，b）、生物医学文本（Lee 等，2020年）、新闻和论文（Gururangan 等，2020年）以及社交媒体（Chakrabarty 等，2019年）。然而，它们都使用相同的掩码语言模型（MLM）损失。我们认为这种做法并非最优，而且尽可能保留语言模型中的通用知识，并将其与目标领域知识相结合也很重要。

将网络剪枝视为重要性计算。众所周知，神经网络中的许多参数都是冗余的，可以进行剪枝（Li 等，2021；Lai 等，2021）。这一点在预训练Transformer中也得到了证实（Chen 等，2020a；Lin 等，2020；Gao 等，2021b；Michel 等，2019；Voita 等，2019）。一种常见的剪枝方法是丢弃绝对值较小的参数（Han 等，2015；Guo 等，2016）。其他方法则在更高层级上对网络进行剪枝。在基于Transformer的模型中，这些方法包括剪除注意力头（Michel 等，2019；Voita 等，2019；McCarley 等，2019），以及剪除标准Transformer层中的子层（Fan 等，2020；Sajjad 等，2020）。然而，上述方法并不适用于我们，因为我们需要使用无标注领域数据计算语言模型的头部重要性，而上述方法都是针对有监督下游任务的。为此，我们提出使用代理KL散度损失。需要注意的是，也可以对Transformer中的其他子层进行剪枝。然而，如第4.3节所示，估计其他层的重要性

[2]我们将在第4 节看到，无需约束其他层中的神经元。[3]这与持续学习（CL）（Chen和Liu，2018年）截然不同，因为持续学习需要保留已有知识，以应对灾难性遗忘（McCloskey和Cohen，1989年）。领域自适应训练可以且应该改变／调整原始语言模型中的通用知识，使其适应目标领域。[4]对比通用知识与仅有的领域特定知识会导致较差的结果（见第4.3节），因为这会使两类知识彼此分离，而非相互融合。

并不能提升性能。

**对比学习**。对比学习（Chen 等人，2020b；He 等人，2020年）可以通过最大化正样本对的相似度并最小化负样本对的相似度，学习到良好的表征：

$$\mathcal{L}_{\text{contrast}} = -\log \frac{e^{(\text{sim}(q_i, q_i^+)/\tau)}}{\sum_{j=1}^{N} e^{(\text{sim}(q_i, q_j^+)/\tau)}}, \quad (1)$$

其中，$N$ 是批次大小，$\tau$ 是温度参数，$\text{sim}(\cdot)$ 是相似度度量，且 $q_i$ 和$q_i^+$是正样本对的表征，$x_i$ 和 $x_i^+$（通常，$x_i^+$是 $x_i$的增强样本，例如通过裁剪、删除或同义词替换生成（Gao 等人，2021a））。在无监督对比损失中，负样本是批次中的其他样本，它们出现在分母中。

我们主要使用对比损失，对比原始语言模型中的重要通用知识表征与完整知识（包括通用知识和领域特定知识）的表征，从而实现通用知识与领域特定知识的良好融合。

## 3 DGA 系统概述

如前所述，DGA 不仅计算掩码语言模型损失，还执行另外两项功能：（1）根据注意力头的重要性对其进行软掩码，从而保留语言模型中的重要通用知识。这有助于避免领域自适应训练过程中语言模型中的通用知识遭到潜在破坏（第 3.1节）。不过，挑战在于如何识别语言模型中的通用知识，以及如何保护这些知识。我们将提出一种相应的方法。（2）鼓励模型学习目标领域与语言模型中的通用知识的融合表征（第 3.2节）。如何做到这一点也并非显而易见。我们提出了一种基于对比学习的方法来实现这一点。图 1 展示了 DGA 的概览。

### 3.1 通过软掩码注意力头保留通用知识

**多头注意力**。多头注意力可以说是 Transformer 模型中最重要的组成部分（Vaswani 等人，2017）。我们省略其他部分的细节，读者可参阅原始论文。形式上，设 $\boldsymbol{x} = x^{(1)}, \ldots, x^{(T)}$ 为一个由 $T$ 个实向量组成的序列，其中 $x^{(t)} \in \mathbb{R}^d$，并设 $q \in \mathbb{R}^d$ 为一个查询向量。注意力机制

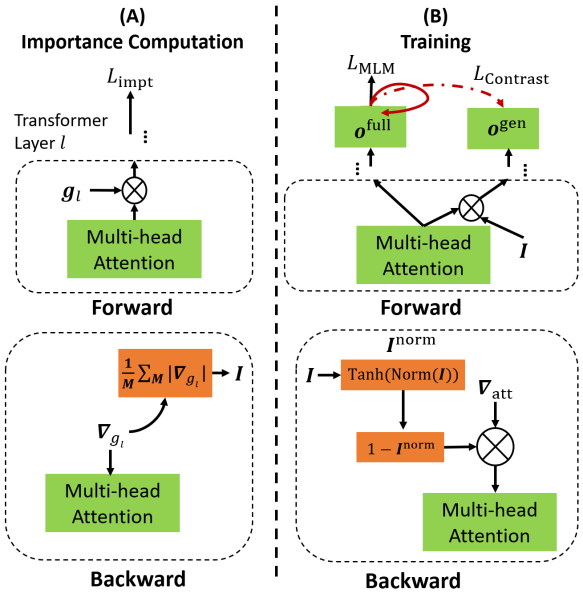


图 1：DGA 示意图。**(A)** 展示了重要性计算过程。具体做法是添加一个门控向量 $\boldsymbol{g}_l$，与多头注意力相乘（式5），并对其训练梯度取平均（式6）。**(B)**展示了 DGA 训练过程。在反向传播中，根据注意力头的重要性对其进行软掩码 $\boldsymbol{I}$（式9和10），以尽可能保留语言模型中的通用知识。在前向传播中，除用于对比损失中的特征学习外，移除添加的门控向量。对比损失通过对比应用了重要性（$\boldsymbol{o}^{\text{gen}}$ 见式12）的通用知识与未应用重要性（$\boldsymbol{o}^{\text{full}}$ 见式14）的完整知识来计算。DGA 的最终目标由掩码语言模型损失和对比损失构成。请注意，我们省略了 Transformer 其他部分的细节，仅关注多头注意力机制。

制定义为

$$\text{att}(\boldsymbol{x}, q) = W_o \sum_{t=1}^{T} \alpha^{(t)}(q) W_v x^{(t)}, \quad (2)$$

其中

$$\alpha^{(t)}(q) = \text{softmax}(\frac{q^T W_q^T W_k x^{(t)}}{\sqrt{d}}). \quad (3)$$

投影矩阵 $W_o, W_v, W_q, W_k \in \mathbb{R}^{d \times d}$是可学习参数。自注意力中的查询向量来自与 $\boldsymbol{x}$相同的序列。一个 Trans- former 包含 $L$ 个相同的层。对于第 $l$ 层，$H_l$个不同的注意力头并行应用，以便让 Transformer 在更多数据上进行训练。简单来说，多头注意力（mhatt）就是在单个 Transformer 架构中同时应用多个注意力头。随后，它们

并行应用，以获得多头注意力。[5]

$$\text{mhatt}_l(\boldsymbol{x}, q) = \sum_{h=1}^{H_l} \text{att}_{lh}(\boldsymbol{x}, q), \tag{4}$$

其中，$h$ 表示 $h^{th}$ 注意力头。请注意，每一层的输入 $\boldsymbol{x}$ 各不相同，因为给定层的输入是上一层的输出。为简化表示，我们对所有层都使用输入 $\boldsymbol{x}$ 。

**头部重要性**。研究人员发现，并非所有注意力头都很重要（Michel 等人，2019年）。我们在注意力头求和过程中引入一个门控向量，$\boldsymbol{g}_l$，其中每个单元都是一个门控变量，$g_{lh}$，用于检测注意力头的重要性。由此得到的重要性分数用于在领域自适应训练中对注意力头进行软掩码处理。

$$\text{gmhatt}_l(\boldsymbol{x}, q) = \sum_{h=1}^{H_l} g_{lh}, \otimes \text{att}_{lh}(\boldsymbol{x}, q) \tag{5}$$

其中，$\otimes$ 表示逐元素相乘。Michel 等人提出了一种基于梯度的头部重要性检测方法（Michel 等人，2019年）。给定一个数据集$D = \{(\boldsymbol{y}_m, \boldsymbol{x}_m)\}_{m=1}^{M}$，包含 $M$ 个样本（$\boldsymbol{y}_m$ 是 $\boldsymbol{x}_m$ 的标签），正如Michel 等人（2019年）在监督学习中所做的那样，使用基于梯度的代理分数来估算一个注意力头的重要性

$$I_{lh} = \frac{1}{M} \sum_{m=1}^{M} |\nabla_{g_{lh}}|, \tag{6}$$

其中，$\nabla_{g_{lh}}$ 是门控变量 $g_{lh}$的梯度，

$$\nabla_{g_{lh}} = \frac{\partial \mathcal{L}_{\text{impt}}(\boldsymbol{y}_m, \boldsymbol{x}_m)}{\partial g_{lh}}, \tag{7}$$

其中，$\mathcal{L}_{\text{impt}}$ 是任务特定／领域特定损失函数。梯度可用作重要性得分，因为如果 $I_{lh}$ 具有较高的值，改变 $g_{lh}$ 可能会对模型产生较大影响。

虽然公式 6 提供了一种计算注意力头重要性的方法，相对于给定的损失$\mathcal{L}_{impt}$，但我们无法直接应用它：如果使用手头的领域数据，并将掩码语言模型损失作为 $\mathcal{L}_{\text{impt}}$，$\nabla_{g_{lh}}$ 则只表示领域特定知识的重要性分数。然而，我们的目标是估计语言模型中通用知识的注意力头重要性，这需要使用训练语言模型时所用的数据来计算 $\mathcal{L}_{\text{impt}}$。

实际上，语言模型的用户无法获取此类数据。此外，公式6 需要标签，但我们在领域自适应训练中使用的领域语料库没有标签。为了解决这些问题，我们提出计算代理*KL*散度损失，用于 $\mathcal{L}_{\text{impt}}$。

**代理KL散度损失**。我们需要为 $\mathcal{L}_{\text{impt}}$ 找到一个代理，使其梯度（$\nabla_{g_{lh}}$）可用于计算头部重要性，而无需使用语言模型原始的预训练数据。我们提议使用模型的鲁棒性 作为代理，即尝试检测对语言模型鲁棒性而言重要的头部。其梯度，$\nabla_{g_{lh}}$，随后可表示鲁棒性，进而表示其对语言模型的重要性。我们的理由如下：如果一个 $I_{lh}$（$|\nabla_{g_{lh}}|$的平均值，见公式6）具有较高的值，这表明它对语言模型的鲁棒性很重要，因为改变它可能会导致语言模型发生巨大变化。因此，它是对语言模型而言重要的头部。相反，如果 $I_{lh}$ 的值较小，则它对语言模型而言不太重要，或并不重要。

为了计算语言模型的鲁棒性，我们从目标领域数据中选取一个子集（一个超参数）$\{\boldsymbol{x}_m^{\text{sub}}\}$（领域自适应训练中没有标签），并将输入 $\boldsymbol{x}_m^{\text{sub}}$两次送入语言模型，计算由此得到的两个表征之间的 KL 散度，

$$\mathcal{L}_{\text{impt}} = \text{KL}(f_1(\boldsymbol{x}_m^{\text{sub}}), f_2(\boldsymbol{x}_m^{\text{sub}})), \tag{8}$$

其中 $f_1$ 和 $f_2$ 是使用不同 Dropout 掩码的语言模型。注意，我们无需添加任何额外的 Dropout 来实现 $f$，因为在 Transformer 中，作为输入使用的是独立采样的*Dropout* 掩码。训练 Transformer 时，全连接层和注意力概率上都会使用 Dropout 掩码。因此，只需将相同输入两次馈入 Transformer，就能得到使用不同 Dropout 掩码的两个表征。由于 Dropout 类似于添加噪声，两个表征之间的差异可以视为 Transformer 模型的鲁棒性。图1 （A）展示了如何利用门控向量的梯度计算每个注意力头的重要性 $\boldsymbol{g}_l$。

**领域自适应训练中的注意力头软掩码**。回顾一下，我们希望在领域自适应训练期间，使用头部重要性 $I_{lh}$ 来保留语言模型中的通用知识。给定注意力头 $\text{att}(\boldsymbol{x}, q)$ 和领域自适应训练损失 $\mathcal{L}_{\text{DA-train}}$ （通常为掩码语言模型损失；我们还在第3.2节提出了一个额外的损失），我们可以对其对应的梯度进行“软掩码”

[5]我们遵循（Michel 等人，2019年）中的符号表示法，其中公式 4 中的符号表示法等价于（Vaswani 等人，2017）中的“拼接”公式。

($\nabla_{\text{att}_{lh}}$ [6])，使用注意力头重要性值 $I_{lh}$，

$$\nabla'_{\text{att}_{lh}} = (1 - I_{lh}^{\text{norm}}) \otimes \nabla_{\text{att}_{lh}}, \quad (9)$$

$I_{lh}^{\text{norm}}$通过归一化得出 $I_{lh}$

$$I_{lh}^{\text{norm}} = |\text{Tanh}(\text{Normalize}(I_{lh}))|. \quad (10)$$

归一化 ma kes the $I_{lh}$ have 的均值 o f 0 and 标准差为 1。绝对值为双曲正切确保 $I_{lh}$ 取值范围为 $[0, 1]$。公式 9 意味着通过将对应注意力头的重要性从梯度中逐元素相减，来约束该注意力头的梯度$_{lh}(\boldsymbol{x},\ q)$ 。这称为“软掩码”，因为 $I_{lh}$ 是一个取值范围为 $[0, 1]$ 的实数（而不是二进制值 $\{0, 1\}$），这使模型能够灵活地调整注意力头。这很有用，因为尽管某些注意力头对语言模型很重要，但它们可能与目标领域中的知识相冲突，因此需要进行调整。还要注意，此处的软掩码仅影响反向传播，不用于前向传播（因此前向传播可以使用完整网络，并促进预训练的一般知识和领域特定知识最大限度地融合），但使用对比学习进行特征学习时除外（见下文）。图 1 （B）显示，训练期间注意力头会进行软掩码。

### 3.2 通用知识与完整知识的对比

现在，我们介绍如何通过对比通用知识和完整知识（包括通用知识和领域特定知识），将语言模型中的通用知识与目标领域中的领域特定知识结合起来。我们首先介绍如何从语言模型中获取输入$\boldsymbol{x}$对应的这类知识，然后讨论如何对它们进行对比。

**通过提取语言模型在前向传播中，将注意力头与其重要性分数结合所得的表示 $\boldsymbol{x}$，可以获得输入序列的通用知识**（见式$I_{lh}^{\text{norm}}$10）。直观而言，由于重要性分数体现了每个注意力头对通用知识的重要程度，因此得到的表示反映了 $\boldsymbol{x}$所使用的主要通用知识。形式上，我们将 $I_{lh}^{\text{norm}}$（软掩码）作为门控变量 $g_{lh}$ 代入式5，

$$\text{gmhatt}_l^{\text{gen}}(\boldsymbol{x}, q) = \sum_{h=1}^{H_l} I_{lh}^{\text{norm}} \otimes \text{att}_{lh}(\boldsymbol{x}, q). \quad (11)$$

[6] $\nabla_{\text{att}_{lh}}$ 表示注意力头 att 的梯度$_{lh}(\boldsymbol{x}, q)$，与 $\nabla_{g_{lh}}$ 在公式6 中的梯度不同，后者是门控变量的梯度 $g_{lh}$

给定用于通用知识的注意力头后，我们可以将其代入整个 Transformer，从而得到最终的通用知识（取输入序列中每个词元输出的平均值）。

$$\boldsymbol{o}^{\text{gen}} = \text{Transformer}(\text{gmhatt}^{\text{gen}}(\boldsymbol{x}, q)). \quad (12)$$

（另请参见 $\boldsymbol{o}^{\text{gen}}$ 图1 （B））。

**获取$\boldsymbol{x}$** 中的完整知识（包括通用知识和领域特定知识）也类似。唯一的区别是，在前向传播中，我们提取 $\boldsymbol{x}$的表示时，不会在注意力头上应用重要性软掩码，

$$\text{gmhatt}_l^{\text{full}}(\boldsymbol{x}, q) = \sum_{h=1}^{H_l} \text{att}_{lh}(\boldsymbol{x}, q). \quad (13)$$

同样，我们可以将其接入 Transformer，

$$\boldsymbol{o}^{\text{full}} = \text{Transformer}(\text{gmhatt}^{\text{full}}(\boldsymbol{x}, q)). \quad (14)$$

（参见 $\boldsymbol{o}^{\text{full}}$ 图 1 （B））。请注意，可以使用（$1 - I_{lh}^{\text{norm}}$）来衡量领域特定知识的重要性，并将其与通用知识进行对比。不过，这样会导致结果较差（见表3），详情见脚注 4。

**对比通用知识与完整知识**。众所周知，在正例和负例的帮助下，对比学习有助于学习良好的各向同性表征，适用于下游任务。对于同一输入 $\boldsymbol{x}$，我们将通用表征（$\boldsymbol{o}^{\text{gen}}$）和完整表征（$\boldsymbol{o}^{\text{full}}$）作为正例和负例进行对比，使它们彼此不同，从而促进模型在 $\boldsymbol{o}^{\text{full}}$ 中学习领域特定知识；这些知识尚未包含在通用知识中，同时又与输入的通用知识（$\boldsymbol{o}^{\text{gen}}$）相关并与之整合。

我们按如下方式构造对比样本：对于输入 $\boldsymbol{x}_m$，会生成三个对比样本。锚点 $\boldsymbol{o}_m$ 和正样本 $\boldsymbol{o}_m^+$ 均为公式 14中的完整知识，它们基于 Transformer 中独立采样的两个 Dropout 掩码获得（请回忆，这可以通过将 $\boldsymbol{x}_m$ 输入两次来实现（见第 3.1节））。我们将 $\boldsymbol{o}_m^+$和 $\boldsymbol{o}_m$ 视为正样本，因为已有研究表明，Dropout 噪声是有助于提升句子嵌入训练中对齐效果的正样本（Gao 等人，2021a）。负样本 $\boldsymbol{o}_m^-$是通过公式 12 从语言模型中获得的 $\boldsymbol{x}_m$ 的通用知识。给定 $\boldsymbol{o}_m$、$\boldsymbol{o}_m^+$和 $\boldsymbol{o}_m^-$，我们的对比损失为（sim$(\cdot)$ 表示余弦相似度），

$$\mathcal{L}_{\text{contrast}} = -\log \frac{e^{\text{sim}(\boldsymbol{o}_m, \boldsymbol{o}_m^+)/\tau}}{\sum_{j=1}^{N} (e^{\text{sim}(\boldsymbol{o}_m, \boldsymbol{o}_j^+)/\tau} + e^{\text{sim}(\boldsymbol{o}_m, \boldsymbol{o}_j^-)/\tau})}. \quad (15)$$

与公式 1相比，分母中增加了第二项，即将通用知识表示作为额外的负样本。图 1 (B) 中，红色箭头从 $\boldsymbol{o}^{\text{full}}$指向其自身，表示正样本来自将输入输入两次。指向 $\boldsymbol{o}^{\text{gen}}$ 的红色虚线箭头表示对比专门知识与通用知识的负样本。

### 3.3 DGA 目标

DGA 是一种流水线模型：首先，使用一部分领域数据来估算注意力头重要性（$I_{lh}$ 见第 3.1节）。其次，给定注意力头重要性后，我们通过结合传统的掩码语言模型（MLM）损失（包括所提出的用于保留通用知识的软掩码）和所提出的对比损失，计算最终的领域自适应损失：

$$\mathcal{L}_{\text{DA-train}} = \mathcal{L}_{\text{MLM}} + \lambda_1 \mathcal{L}_{\text{contrast}}, \qquad (16)$$

其中， $\lambda_1$ 是用于调整新增项影响的超参数。

## 4 实验

我们遵循（Gururangan 等人，2020年）中的实验设置。RoBERTa（Liu等人，2019年）[7] 用作语言模型。在每项实验中，我们首先对语言模型进行领域自适应训练，然后在下游任务上对其进行微调。最终评估基于下游任务的结果。

### 4.1 数据集与基线

数据集：表 1 展示了用于领域自适应训练的无标签领域数据集及其对应的下游任务分类数据集的统计信息。我们使用了 6 个无标签领域数据集：[8] 3 其中 3 个是评论数据集：Yelp 餐厅（Xu 等，2019a）、Amazon 手机（Ni 等，2019）、Amazon 相机（Ni 等，2019）；另外 3 个是学术论文数据集：ACL 论文（Lo 等，2020）、人工智能论文（Lo 等，2020）和 PubMed 论文[9]。每个无标签领域数据集都有一个对应的下游任务分类数据集[10]：餐厅[11]（Xu 等，2019a）、手机（Ding 等，2008年；Hu 和 Liu，2004）、相机（Ding 等，2008年；Hu 和 Liu，2004）[12], ACL（Jurgens 等人（2018年）中的 ACL-ARC）、人工智能（Luan 等人（2018年）中的 SCIERC）和 PubMed（Kringelum 等人（2016年）中的 CHEMPORT）。

**基线。** 我们考虑了 10 个基线。

(1). **非领域适应训练（RoBERTa）**（Liu 等人，2019年）使用原始 RoBERTa 进行最终任务微调，不进行任何领域自适应训练。

(2).**使用掩码语言模型损失（MLM）进行领域自适应训练**是现有的领域自适应训练方法。据我们所知，现有的领域自适应训练系统均基于掩码语言模型损失。

(3). **使用适配器调优进行领域自适应训练（掩码语言模型（适配器））**在 Transformer 的各层之间添加适配器层以进行领域自适应训练。适配器（Houlsby 等人，2019年）由两个全连接层和一个跳跃连接组成。领域自适应训练期间，Transformer 的参数保持固定，仅训练适配器。瓶颈（适配器）大小设为 64（Houlsby 等人，2019年）。为进行公平比较，在最终任务微调期间，RoBERTa 和适配器均可训练。

(4). **使用提示调优进行领域自适应训练（掩码语言模型（提示））**(Lester 等人，2021）会在原始序列末尾添加一串提示词元。在领域自适应训练中，RoBERTa（语言模型）保持固定，仅训练提示词元。在最终任务微调中，语言模型和训练好的提示都可以训练。我们初始化 100 个词元，并按照 Lester 等人(2021) 中的设置，将领域自适应训练时提示词元的学习率设为 0.3。

(5). **知识蒸馏（掩码语言模型+知识蒸馏）**(Hinton 等人，2015）通过最小化语言模型中的通用知识与领域自适应训练中的专门知识之间的表征偏差来进行蒸馏。我们计算两个模型（语言模型和经过领域自适应训练的模型）中每个词的表征（掩码语言模型预测头之前的输出）之间的 KL 散度，并将其作为蒸馏损失。

（6）。**通过注意力进行的适配蒸馏（掩码语言模型+AdaptedDeiT）**源自 DeiT（Tou-

[7] https://huggingface.co/roberta-base[8]我们对 *PubMed* 进行了下采样，因为其原始数据量极大。总体而言，与以往工作相比（Gururangan 等人，2020年）（其每个领域使用了超过 11GB 的数据），我们的数据集小得多。实验结果表明，较小的数据集已足够，更多数据并无助益，而且所需的计算资源也少得多。[9] https://pubmed.ncbi.nlm.nih.gov/

[10]请注意，我们的结果与 Gururangan 等人，2020年）的表 5 所示结果不同，因为原始测试集规模很小，导致我们观察到的方差非常大。因此，我们重新划分了训练集和测试集（扩大测试集并略微缩小训练集）[11]为与现有研究保持一致（Tang 等人，2016），不使用存在极性冲突的样例（即针对某个方面词同时表达正面和负面情感）。[12]请注意，Ding 等人(2008年）和Hu 和 Liu (2004）分别包含 9 个和 5 个领域。我们从中提取与“手机”和“相机”相关的领域。

| 无标签领域数据集 | | | 终端任务分类数据集 | | | | |
|---|---|---|---|---|---|---|---|
| 来源 | 数据集 | Size | 数据集 | Task | 训练数量 | 测试数量 | 类别数量 |
| 评论 | Yelp 餐厅 | 758MB | 餐厅 | 方面情感分类（ASC） | 3,452 | 1,120 | 3 |
| | Amazon 手机 | 724MB | 手机 | 方面情感分类（ASC） | 239 | 553 | 2 |
| | Amazon 相机 | 319MB | 相机 | 方面情感分类（ASC） | 230 | 626 | 2 |
| 学术论文 | ACL 论文 | 867MB | ACL | 引文意图分类 | 1,520 | 421 | 6 |
| | 人工智能论文 | 507MB | AI | 关系分类 | 2,260 | 2,388 | 7 |
| | PubMed 论文 | 989MB | PubMed | 化学物质-蛋白质相互作用预测 | 2,667 | 7,398 | 13 |

表 1：领域后训练数据集和最终任务监督分类数据集的统计信息（各任务的更多详细信息见附录A）。

vron 等人，2021），这是一种用于视觉 Transformer（ViT）的蒸馏方法（Dosovitskiy 等人，2020年）。我们通过将语言模型表示蒸馏[13] 到新增的蒸馏令牌中，将 DeiT 适配为基于文本的无监督模型，并将 ViT 更改为 RoBERTa。

（7, 8）。**使用序列级对比学习进行领域自适应训练（MLM+SimCSE 和 MLM+InfoWord）**。SimCSE 是一种用于句子嵌入的对比学习方法（Gao 等人，2021a）。我们使用其无监督版本，其中正样本来自同一输入，但使用不同的 Dropout 掩码；负样本则是同一批次中的其他实例。InfoWord（Kong 等人，2020年）是另一种对比学习方法，用于对比跨度级局部表示与序列级全局表示。

（9, 10）。**使用令牌感知对比学习进行领域自适应训练（MLM+TaCL 和 MLM+TaCO）**。TaCL（Su 等人，2021）和 TaCO（Fu 等人，2022）是近期提出的两种方法，通过令牌感知对比损失改进 BERT 预训练。[14] 为进行公平比较，我们将骨干模型更换为 RoBERTa。

### 4.2 实现细节

**架构**。我们采用 $\text{RoBERTa}_{\text{BASE}}$ 作为骨干语言模型（12 层，每层包含 12 个注意力头）。我们使用掩码语言模型头进行领域自适应训练。RoBERTa 的下游任务微调遵循标准做法。对于三个方面级情感分类任务（见表 1），我们采用（Xu 等，2019a）中的方面级情感分类形式，其中方面（例如“声音”）与评论句（例如“声音很棒”）通过 </s>连接。

**超参数**。除非另有说明，所有实验均使用相同的超参数。最大输入长度设置为 164，这一长度足以应对所有数据集。Adam 优化器用于领域自适应训练和最终任务微调。最大序列长度设置为 164，这一长度足以应对我们的下游任务，并且只需要适度的计算资源。

**领域自适应预训练（领域自适应训练）**。学习率设为 1e-4，批次大小为 256。我们针对每个领域训练 2.5K 步，大致相当于完整遍历一次领域数据，这遵循（Gu- rurangan 等，2020年；Xu 等，2019a）的做法。用于计算 $\{\boldsymbol{x}_m^{\text{sub}}\}$ 以确定第 $\mathcal{L}_{\text{impt}}$ 节中头部重要性的子集数据3.1设为 164 万个词元，这在我们的实验中已足够。 $\lambda_1$ 在公式 16 中设为 1， $\tau$ 在公式 15 中设为 0.05。

**下游任务微调**。学习率设为 1e-5，批次大小设为 16。我们在下游任务微调数据集上进行训练：餐厅任务训练 5 个 epoch；ACL、人工智能和 PubMed 任务训练 10 个 epoch；手机和相机任务训练 15 个 epoch。我们直接采用最后一个 epoch 的结果，因为根据经验，上述 epoch 数量能够带来稳定且收敛的结果。

### 4.3 评估结果与消融研究

我们在表2中报告了 10 个基线在 6 个数据集上的下游任务结果。

**DGA 的优越性**。我们的 DGA 始终优于所有基线。得益于所提出的更具知情性的适应方法，DGA 的表现优于广泛使用的传统领域自适应训练基线掩码语言模型（MLM）。我们还发现，平均而言，MLM 的表现显著优于 RoBERTa（非领域自适应训练）（见最后一列）。下文将进一步讨论结果中的其他发现。

(1). 在 DGA 中对整个语言模型进行训练有助于取得更好的结果。使用适配器（掩码语言模型（适配器））和提示（掩码语言模型（提示））的效果不一。这是因为适配器和提示的可训练参数不足，而且这些参数是随机初始化的，因此可能难以训练。

（2）DGA 也优于基于蒸馏的系统：MLM+AdaptedDeiT 和 MLM+KD，这些系统试图保留已有知识。这

[13]我们取其词元输出的平均值作为序列的表示。SimCSE 基线也采用相同做法。[14]TaCL 和 TaCO 并非领域自适应训练模型。它们通过使用与训练语言模型相同的数据进行预训练，以提升语言模型。我们将数据替换为目标域数据。

| 领域<br>模型 | 相机 | | 手机 | | 餐厅 | | AI | | ACL | | PubMed | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MF1 | Acc. | MF1 | Acc. | MF1 | Acc. | MF1 | Acc. | MF1 | Acc. | 微平均 F1 | |
| RoBERTa | 78.82 | 87.03 | 83.75 | 86.08 | 79.81 | 87.00 | 60.98 | 71.85 | 66.11 | 71.26 | 72.38 | 73.64 |
| MLM | 84.39 | 89.90 | 82.59 | 85.50 | 80.84 | 87.68 | 68.97 | 75.95 | 68.75 | 73.44 | 72.84 | 76.40 |
| 掩码语言模型（适配器） | 83.62 | 89.23 | 82.71 | 85.35 | 80.19 | 87.14 | 60.55 | 71.38 | 68.87 | 72.92 | 71.68 | 74.60 |
| 掩码语言模型（提示） | 85.52 | 90.38 | 84.17 | 86.53 | 79.00 | 86.45 | 61.47 | 72.36 | 66.66 | 71.35 | 73.09 | 74.98 |
| 掩码语言模型+知识蒸馏 | 82.79 | 89.30 | 80.08 | 83.33 | 80.40 | 87.25 | 67.76 | 75.46 | 68.19 | 72.73 | 72.35 | 75.26 |
| 掩码语言模型+AdaptedDeiT | 86.86 | 91.37 | 83.08 | 85.64 | 79.70 | 86.84 | 69.72 | 76.83 | 69.11 | 73.35 | 72.69 | 76.86 |
| 掩码语言模型+SimCSE | 84.91 | 90.35 | 83.46 | 86.08 | 80.88 | 87.59 | 69.10 | 76.25 | 69.89 | 74.30 | 72.77 | 76.84 |
| 掩码语言模型+TaCL | 81.98 | 88.88 | 81.87 | 84.92 | 81.12 | 87.50 | 64.04 | 73.18 | 63.18 | 70.31 | 69.46 | 73.61 |
| 掩码语言模型+TaCO | 84.50 | 90.22 | 82.63 | 85.32 | 79.27 | 86.68 | 59.73 | 71.22 | 63.66 | 70.36 | 72.38 | 73.69 |
| 掩码语言模型+InfoWord | 87.95 | 91.92 | 84.58 | 86.84 | 81.24 | 87.82 | 68.29 | 75.92 | 68.58 | 73.68 | 73.21 | 77.31 |
| DGA | **88.52** | **92.49** | **85.47** | **87.45** | **81.83** | **88.20** | **71.99** | **78.06** | **71.01** | **74.73** | **73.65** | **78.74** |

表2：我们报告了所有数据集上的宏平均F1（微平均 F1）和准确率结果，但 PubMed 领域中的 CHEMPORT 除外；对于后者，我们遵循 Gururangan 等人(2020年); Dery等人(2021); Beltagy等人(2019年)。结果为 5 个随机种子的平均值（标准差见附录 B）。平均列（平均值）为所有数据集的微平均 F1（对于 PubMed，则为 Micro-F1）的平均值。

并不令人意外，因为领域自适应训练的目标并不仅仅是保留已有知识，还要根据需要对其进行调整或改变，以适应目标领域。DGA 专门通过软掩码和知识对比来实现这一点。

(3). DGA 中的对比学习比其他对比方法（掩码语言模型+SimCSE、掩码语言模型+TaCL、掩码语言模型+TaCO 和掩码语言模型+InfoWord）更有效。这表明，在知识整合中对比通用知识与完整知识十分重要。

**代理KL散度损失的有效性**。我们使用提出的代理*KL*散度损失来计算头部重要性，从而识别语言模型中的通用语言知识，而无需使用该语言模型原始的预训练数据（第3.1节）。

在评估中，我们关注代理方法的效果。由于我们没有用于预训练 RoBERTa 的数据，因此无法直接评估该代理方法的质量。在这里，我们提供了一些间接证据，以展示该代理方法在计算语言模型中各单元对通用知识的重要性方面的有效性。

我们使用来自不同领域的数据进行了一项单独实验，比较应用代理方法后各注意力头的重要性得分向量。对于每个领域 $i$，我们将其重要性向量与其他每个领域的重要性向量进行比较，然后对余弦相似度取平均，得到领域 $i$ 的值。餐厅的结果为 0.92，ACL、人工智能和手机的结果均为 0.91，PubMed 为 0.89，相机为 0.92。我们发现，不同领域得出的重要性值相近，这间接表明我们的代理方法能够识别共同通用知识。

我们还计算了代理方法的重要性分数分布。在应用代理方法后，对于 6 个领域中的每一个，大约 20% 的注意力头受到强力保护（$0.8 \leq I_{lh}^{\text{norm}} \leq 1.0$），另有 20% 受到中度保护（$0.6 \leq I_{lh}^{\text{norm}} < 0.8$），这表明它们承载着通用知识。手机、人工智能、相机和餐厅的分布相似，而 ACL 和 PubMed 受到的保护略少。这可以理解，因为 PubMed 和 ACL（医学或自然语言处理出版物）可能不如其他领域常见，语言模型中的通用知识对它们的涵盖也较少。

**消融研究**。为了更好地理解 DGA，我们希望了解：（1）约束其他层中的神经元是否有帮助（所提出的 DGA 仅约束注意力头）；（2）DGA 的收益来自何处。为回答（1），我们对标准 Transformer 中不同层的训练施加约束。在表3 （第 3–5 行）中，“**H**”、“**I**”和“**O**”分别表示标准 Transformer 层中的注意力头、中间层和输出层。“**E**”表示嵌入层。包含“H、I、O、E”组合的方括号表示应用软掩码的位置（DGA 仅在注意力头上应用软掩码）。可以看到，这些方法的结果与 DGA 相近或更差，这表明注意力头更能体现重要知识。为回答（2），我们进行了以下消融实验：（i）**DGA（无对比）**，不使用对比损失，仅根据注意力头重要性对反向传播应用软掩码。（ii）**DGA（随机掩码）**，使用随机生成的注意力头重要性分数，并据此进行软掩

| 领域<br>模型 | 相机 | | 手机 | | 餐厅 | | AI | | ACL | | PubMed | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MF1 | Acc. | MF1 | Acc. | MF1 | Acc. | MF1 | Acc. | MF1 | Acc. | 微平均 F1 | |
| RoBERTa | 78.82 | 87.03 | 83.75 | 86.08 | 79.81 | 87.00 | 60.98 | 71.85 | 66.11 | 71.26 | 72.38 | 73.64 |
| MLM | 84.39 | 89.90 | 82.59 | 85.50 | 80.84 | 87.68 | 68.97 | 75.95 | 68.75 | 73.44 | 72.84 | 76.40 |
| DGA（H、I） | 86.79 | 91.60 | 84.21 | 86.40 | 81.32 | 87.91 | 71.07 | 77.36 | 69.50 | 73.82 | 73.34 | 77.71 |
| DGA（H、I、O） | 88.04 | 92.01 | 85.85 | 87.63 | 81.45 | 87.79 | 71.54 | 77.61 | 70.52 | 74.58 | 73.10 | 78.42 |
| DGA（H、I、O、E） | 87.05 | 91.60 | 83.74 | 86.11 | 80.64 | 87.61 | 72.64 | 78.17 | 71.24 | 74.96 | 73.54 | 78.14 |
| DGA（无对比） | 86.19 | 90.89 | 84.48 | 86.65 | 81.70 | 87.93 | 68.25 | 75.49 | 69.31 | 73.73 | 72.72 | 77.11 |
| DGA（随机掩码） | 82.07 | 89.30 | 83.86 | 86.33 | 80.60 | 87.52 | 69.51 | 76.64 | 69.59 | 73.73 | 72.92 | 76.43 |
| 集成（语言模型+掩码语言模型） | 85.22 | 90.64 | 85.15 | 87.23 | 79.86 | 86.98 | 65.10 | 74.43 | 68.56 | 73.44 | 72.60 | 76.08 |
| DGA（领域特定） | 88.06 | 92.04 | 83.45 | 85.82 | 81.72 | 87.90 | 68.00 | 75.57 | 70.91 | 75.06 | 73.17 | 77.55 |
| DGA | **88.52** | **92.49** | **85.47** | **87.45** | **81.83** | **88.20** | **71.99** | **78.06** | **71.01** | **74.73** | **73.65** | **78.74** |

表3：消融实验结果——5个随机种子的平均值。标准差报告于附录B。

码和对比学习。（iii）**集成（语言模型+掩码语言模型）**在经 MLM 领域适配训练的 RoBERTa（传统领域自适应训练）和原始 RoBERTa（语言模型）上分别进行下游任务微调，拼接二者的输出并取其平均值。（iv）**DGA（领域特定）**指对领域特定知识与通用知识进行对比的变体（见第 3.2节）。[15]

表3 表明，完整的 DGA 始终能取得最佳结果，这说明每个组成部分都有贡献。其他观察结果如下：

(1) DGA 的增益部分来自新颖的软掩码：可以看到，平均而言，DGA（无对比）的表现优于传统的领域自适应训练（掩码语言模型）。此外，我们基于梯度的掩码具有信息量：可以看到，在所有数据集上，DGA（随机掩码）的表现都不如 DGA（无对比）。DGA（无对比）的表现甚至优于集成；集成会在最终任务微调期间，直接结合原始语言模型和传统领域自适应训练模型所提供的信息

(2) 除了软掩码，对比通用知识与完整知识也有帮助。可以看到，DGA 在所有数据集上的表现都优于 DGA（无对比）和 DGA（领域特定）。

## 5 结论

本文认为，有效的领域自适应训练方法应将目标领域知识有效地融入语言模型中的通用知识。现有方法并未明确做到这一点。本文提出了一种新方法 DGA，通过以下方式实现这一目标：(1) 估算语言模型中注意力头的重要性，并利用重要性分数在领域自适应训练中对注意力头进行软掩码，以尽可能保留语言模型中的重要知识；(2) 对比通用知识与完整知识。大量实验结果证明了所提出的 DGA 方法的有效性。

[15]我们没有 DGA（无软掩码）这一变体，因为我们的对比学习依赖于软掩码。如果移除软掩码，对比损失将不再包含额外的负样本，而我们的 DGA 就会变成 MLM+SimCSE。

## 6 局限性

尽管 DGA 效果显著，但仍存在一些局限性。首先，DGA 主要致力于使语言模型适应给定的目标领域，并未考虑对其他领域的泛化。例如，逐步或持续地使语言模型适应越来越多的领域，从而提升其用途，将会很有意思（Ke等人，2020年，2021，2022）。其次，语言模型中对通用知识至关重要的参数，是通过一种基于模型鲁棒性的代理方法计算得出的。尽管这种方法相当有效，但探索其他方法以进一步改进这一点也很有意义。我们将在未来的工作中研究这些问题，因为使语言模型专业化并加以改进是一项重要的任务。

## 致谢

Zixuan Ke 和 Bing Liu 的工作部分得到了三项美国国家科学基金会（NSF）资助：IIS-1910424、IIS-1838770 和 CNS-2225427。

## 参考文献

Emily Alsentzer、John R Murphy、Willie Boag、Wei-Hung Weng、Di Jin、Tristan Naumann 和 Matthew McDermott。2019年。公开可用的临床BERT嵌入。*arXiv* 预印本 *arXiv:1904.03323*。

Iz Beltagy、Kyle Lo 和 Arman Cohan。2019年。SciBERT：一种面向科学文本的预训练语言模型。

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared DKaplan、PrafullaDhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda

Askell 等，2020年。语言模型是少样本学习器。神经信息处理系统进展。

Tuhin Chakrabarty、Christopher Hidey 和 Kathleen McKeown。2019年。恕我直言，微调可提升主张检测效果。*arXiv* 预印本 *arXiv:1905.07000*。

Tianlong Chen、Jonathan Frankle、Shiyu Chang、Sijia Liu、Yang Zhang、Zhangyang Wang 和 Michael Carbin。2020a。预训练 BERT 网络的彩票假设。神经信息处理系统进展，33：15834–15846。

Ting Chen、Simon Kornblith、Mohammad Norouzi 和 Geoffrey Hinton。2020b。用于视觉表征对比学习的简单框架。载于国际机器学习会议，第1597–1607页。PMLR。

Zhiyuan Chen 和 Bing Liu。2018年。终身机器学习。人工智能与机器学习综合讲义，12(3)：1–207。

Lucio M Dery、Paul Michel、Ameet Talwalkar 和 Graham Neubig。2021。我们应该进行预训练吗？论证将面向最终任务的训练作为替代方案。*arXiv* 预印本 *arXiv:2109.07437*。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019。BERT：用于语言理解的深度双向 Transformer 预训练。发表于*NAACL-HLT*。

Xiaowen Ding、Bing Liu 和 Philip S Yu。2008年。一种用于意见挖掘的整体性基于词典的方法。发表于*2008*年网络搜索与数据挖掘国际会议论文集。

Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly 等。2020年。一张图像胜过 16×16 个词：用于大规模图像识别的 Transformer。*arXiv* 预印本 *arXiv:2010.11929*。

Angela Fan、EdouardGrave 和 Armand Joulin。2020年。通过结构化 dropout 按需减少 Transformer 深度。载于国际学习表征会议。

Zhiyi Fu、Wangchunshu Zhou、Jingjing Xu、Hao Zhou 和 Lei Li。2022。超越掩码语言建模的上下文表征学习。载于 *ACL*。

Tianyu Gao、Xingcheng Yao 和 Danqi Chen。2021a。Simcse：简单的句子嵌入对比学习。*arXiv* 预印本 *arXiv:2104.08821*。

Yang Gao、Nicolo Colombo 和 Wei Wang。2021b。通过剪枝进行适配：关于 bert 的案例研究。*arXiv* 预印本 *arXiv:2105.03343*。

Yiwen Guo、Anbang Yao 和 Yurong Chen。2016年。用于高效深度神经网络的动态网络手术。神经信息处理系统进展，29。

Suchin Gururangan、Ana Marasovic、Swabha Swayamdipta、Kyle Lo、Iz Beltagy、Doug Downey 和 Noah A. Smith。2020年。不要停止预训练：将语言模型适配到领域和任务。在*ACL*。Song Han、Jeff Pool、John Tran 和 William Dally。2015年。学习权重和连接以构建高效神经网络。神经信息处理系统进展，28。Kaiming He、Haoqi Fan、Yuxin Wu、Saining Xie 和 Ross Girshick。2020年。用于无监督视觉表征学习的动量对比。在*IEEE/CVF*计算机视觉与模式识别会议论文集，第9729–9738页。

Geoffrey Hinton、Oriol Vinyals、Jeff Dean 等。2015。提炼神经网络中的知识。*arXiv* 预印本 *arXiv:1503.02531*，2(7)。

Neil Houlsby、Andrei Giurgiu、Stanislaw Jastrzebski、Bruna Morrone、Quentin de Laroussilhe、Andrea Gesmundo、Mona Attariyan 和 Sylvain Gelly。2019年。面向自然语言处理的参数高效迁移学习。载于 *ICML*。

Minqing Hu 和 Bing Liu。2004年。客户评论的挖掘与总结。载于 *ACM SIGKDD* 论文集。

David Jurgens、Srijan Kumar、Raine Hoover、Daniel A. McFarland 和 Dan Jurafsky。2018年。通过引文框架衡量科学领域的演变。*TACL*。

Zixuan Ke、Haowei Lin、Yijia Shao、Hu Xu、Lei Shu 和 Bing Liu。2022。用于少样本学习的语言模型持续训练。载于 自然语言处理实证方法会议（*EMNLP*）。

Zixuan Ke、Bing Liu 和 Xingchang Huang。2020年。相似与不相似任务混合序列的持续学习。载于*NeurIPS*。

Zixuan Ke、Bing Liu、Nianzu Ma、Hu Xu 和 Lei Shu。2021年。实现持续学习中的遗忘防止与知识迁移。*NeurIPS*。

Lingpeng Kong、Cyprien de Masson d'Autume、Lei Yu、Wang Ling、Zihang Dai 和 Dani Yogatama。2020年。从互信息最大化的角度看语言表征学习。发表于*ICLR*。

Jens Kringelum、Sonny Kim Kjaerulff、Søren Brunak、Ole Lund、Tudor I Oprea 和 Olivier Taboureau。2016年。Chemprot-3.0：全球化学物质与疾病图谱。*Database*，2016。

Cheng-I Jeff Lai、YangZhang、Alexander H Liu、Shiyu Chang、Yi-Lun Liao、Yung-Sung Chuang、Kaizhi Qian、Sameer Khurana、David Cox 和 Jim Glass。2021。Parp：针对自监督语音识别的剪枝、调整与再次剪枝。*NeurIPS*，34。

Jinhyuk Lee、Wonjin Yoon、Sungdong Kim、Donghyeon Kim、Sunkyu Kim、Chan Ho So 和 Jaewoo Kang。2020。BioBERT：一种用于生物医学文本挖掘的预训练生物医学语言表示模型。《生物信息学》，36(4):1234–1240。Brian Lester、Rami Al-Rfou 和 Noah Constant。2021。规模对于参数高效提示调优的力量。载于*EMNLP*。Jiaoda Li、Ryan Cotterell 和 Mrinmaya Sachan。2021。Transformer 注意力头的可微子集剪枝。《计算语言学协会汇刊》，9:1442–1459。Zi Lin、Jeremiah Zhe Liu、Zi Yang、Nan Hua 和 Dan Roth。2020。通过谱归一化恒等先验剪除 Transformer 模型中的冗余映射。*arXiv* 预印本 *arXiv:2010.01791*。Bing Liu。2015。情感分析：挖掘观点、情感与情绪。剑桥大学出版社。Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019。RoBERTa：一种经过稳健优化的 BERT 预训练方法。*CoRR*。Kyle Lo、Lucy Lu Wang、Mark Neumann、Rodney Kinney 和 Daniel S. Weld。2020。S2ORC：Semantic Scholar 开放研究语料库。载于*ACL*。Yi Luan、Luheng He、Mari Ostendorf 和 Hannaneh Hajishirzi。2018。用于构建科学知识图谱的实体、关系和共指多任务识别。载于*ACL*。JS McCarley、Rishav Chakravarti 和 Avirup Sil。2019。基于 BERT 的问答模型的结构化剪枝。*arXiv* 预印本 *arXiv:1910.06360*。Michael McCloskey 和 Neal J Cohen。1989年。联结主义网络中的灾难性干扰：序列学习问题。载于学习心理学与动机。Paul Michel、Omer Levy 和 Graham Neubig。2019。十六个注意力头真的比一个好吗？神经信息处理系统进展，32。Jianmo Ni、Jiacheng Li 和 Julian J. McAuley。2019。利用远程标注评论和细粒度方面为推荐结果提供依据。载于*EMNLP*，第188–197页。计算语言学协会。Hassan Sajjad、Fahim Dalvi、Nadir Durrani 和 Preslav Nakov。2020。丢弃预训练 Transformer 模型的层所产生的影响。*arXiv* 预印本 *arXiv:2004.03844*。Yixuan Su、Fangyu Liu、Zaiqiao Meng、Lei Shu、Ehsan Shareghi 和 Nigel Collier。2021。TaCL：利用词元感知对比学习改进 BERT 预训练。*arXiv* 预印本 *arXiv:2111.04198*。Chi Sun、Xipeng Qiu、Yige Xu 和 Xuanjing Huang。2019年。如何微调 BERT 进行文本分类？发表于中国计算语言学大会，第194–206页。施普林格。Duyu Tang、Bing Qin 和 Ting Liu。2016。基于深度记忆网络的方面级情感分类。发表于*EMNLP*。Hugo Touvron、Matthieu Cord、Matthijs Douze、Francisco Massa、Alexandre Sablayrolles 和 Hervé Jégou。2021。通过注意力训练数据高效的图像 Transformer 并进行蒸馏。发表于国际机器学习大会，第10347–10357页。PMLR。Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017。注意力就是你所需要的一切。发表于*NeurIPS*。Elena Voita、David Talbot、Fedor Moiseev、Rico Sennrich 和 Ivan Titov。2019。分析多头自注意力：专用注意力头承担了主要工作，其余的可以剪枝。*arXiv* 预印本 *arXiv:1905.09418*。Hu Xu、Bing Liu、Lei Shu 和 Philip S. Yu。2019a。用于评论阅读理解和基于方面的情感分析的 BERT 后训练。发表于*NAACL-HLT*。Hu Xu、Bing Liu、Lei Shu 和 Philip S Yu。2019b。评论对话阅读理解。*arXiv* 预印本 *arXiv:1902.00821*。

| 领域 | 相机 | | 手机 | | 餐厅 | | AI | | ACL | | PubMed |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 模型 | MF1 | Acc. | MF1 | Acc. | MF1 | Acc. | MF1 | Acc. | MF1 | Acc. | 微平均 F1 |
| RoBERTa | ±0.0403 | ±0.0179 | ±0.0210 | ±0.0154 | ±0.0117 | ±0.0049 | ±0.0646 | ±0.0347 | ±0.0192 | ±0.0096 | ±0.0071 |
| MLM | ±0.0479 | ±0.0298 | ±0.0165 | ±0.0103 | ±0.0096 | ±0.0056 | ±0.0117 | ±0.0086 | ±0.0218 | ±0.0118 | ±0.0035 |
| 掩码语言模型（适配器） | ±0.0165 | ±0.0110 | ±0.0265 | ±0.0181 | ±0.0102 | ±0.0068 | ±0.0551 | ±0.0288 | ±0.0142 | ±0.0099 | ±0.0055 |
| 掩码语言模型（提示） | ±0.0243 | ±0.0138 | ±0.0126 | ±0.0087 | ±0.0060 | ±0.0035 | ±0.0301 | ±0.0124 | ±0.0068 | ±0.0108 | ±0.0028 |
| 掩码语言模型+知识蒸馏 | ±0.0295 | ±0.0158 | ±0.0320 | ±0.0230 | ±0.0099 | ±0.0070 | ±0.0345 | ±0.0224 | ±0.0292 | ±0.0155 | ±0.0093 |
| 掩码语言模型+AdaptedDeiT | ±0.0187 | ±0.0122 | ±0.0160 | ±0.0101 | ±0.0048 | ±0.0022 | ±0.0250 | ±0.0179 | ±0.0065 | ±0.0079 | ±0.0086 |
| 掩码语言模型+SimCSE | ±0.0114 | ±0.0077 | ±0.0098 | ±0.0065 | ±0.0029 | ±0.0016 | ±0.0086 | ±0.0056 | ±0.0054 | ±0.0071 | ±0.0027 |
| 掩码语言模型+TaCL | ±0.0218 | ±0.0103 | ±0.0230 | ±0.0159 | ±0.0105 | ±0.0059 | ±0.0275 | ±0.0156 | ±0.0713 | ±0.0394 | ±0.0118 |
| 掩码语言模型+TaCO | ±0.0456 | ±0.0232 | ±0.0166 | ±0.0134 | ±0.0077 | ±0.0052 | ±0.0675 | ±0.0380 | ±0.0207 | ±0.0128 | ±0.0099 |
| 掩码语言模型+InfoWord | ±0.0267 | ±0.0139 | ±0.0272 | ±0.0191 | ±0.0170 | ±0.0089 | ±0.0344 | ±0.0219 | ±0.0070 | ±0.0079 | ±0.0072 |
| DGA | ±0.0095 | ±0.0047 | ±0.0127 | ±0.0094 | ±0.0052 | ±0.0040 | ±0.0127 | ±0.0081 | ±0.0079 | ±0.0080 | ±0.0034 |

表4：所提出的DGA模型与基线在六项实验中对应指标的标准差。

| 领域 | 相机 | | 手机 | | 餐厅 | | AI | | ACL | | PubMed |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 模型 | MF1 | Acc. | MF1 | Acc. | MF1 | Acc. | MF1 | Acc. | MF1 | Acc. | 微平均 F1 |
| RoBERTa | ±0.0403 | ±0.0179 | ±0.0210 | ±0.0154 | ±0.0117 | ±0.0049 | ±0.0646 | ±0.0347 | ±0.0192 | ±0.0096 | ±0.0071 |
| MLM | ±0.0479 | ±0.0298 | ±0.0165 | ±0.0103 | ±0.0096 | ±0.0056 | ±0.0117 | ±0.0086 | ±0.0218 | ±0.0118 | ±0.0035 |
| DGA（H、I） | ±0.0373 | ±0.0210 | ±0.0032 | ±0.0039 | ±0.0054 | ±0.0045 | ±0.0095 | ±0.0048 | ±0.0094 | ±0.0073 | ±0.0049 |
| DGA（H、I、O） | ±0.0167 | ±0.0092 | ±0.0182 | ±0.0155 | ±0.0055 | ±0.0033 | ±0.0093 | ±0.0075 | ±0.0080 | ±0.0070 | ±0.0056 |
| DGA（H、I、O、E） | ±0.0237 | ±0.0123 | ±0.0270 | ±0.0187 | ±0.0099 | ±0.0050 | ±0.0109 | ±0.0089 | ±0.0067 | ±0.0057 | ±0.0079 |
| DGA（无对比） | ±0.0433 | ±0.0251 | ±0.0135 | ±0.0106 | ±0.0060 | ±0.0040 | ±0.0197 | ±0.0119 | ±0.0132 | ±0.0093 | ±0.0050 |
| DGA（随机掩码） | ±0.0879 | ±0.0413 | ±0.0335 | ±0.0235 | ±0.0096 | ±0.0044 | ±0.0153 | ±0.0090 | ±0.0105 | ±0.0059 | ±0.0052 |
| 集成 | ±0.0332 | ±0.0178 | ±0.0199 | ±0.0139 | ±0.0035 | ±0.0031 | ±0.0236 | ±0.0103 | ±0.0061 | ±0.0028 | ±0.0046 |
| DGA（领域特定） | ±0.0137 | ±0.0070 | ±0.0259 | ±0.0200 | ±0.0031 | ±0.0018 | ±0.0128 | ±0.0071 | ±0.0108 | ±0.0067 | ±0.0043 |
| DGA | ±0.0095 | ±0.0047 | ±0.0127 | ±0.0094 | ±0.0052 | ±0.0040 | ±0.0127 | ±0.0081 | ±0.0079 | ±0.0080 | ±0.0034 |

表 5：所提出的 DGA模型及其在六项实验中的消融实验对应指标的标准差。

## A 数据集详情

正文中的表2给出了各数据集中样例的数量。这里我们补充介绍四种下游任务的更多细节。

(1) **(手机、相机和餐厅)方面情感分类（ASC）**定义如下（Liu，2015）：给定一个方面或产品特征（例如，图像质量在相机评论中），以及领域或产品类别（例如相机）中包含该方面的评论句子，判断该句子表达的关于该方面的情感或极性是积极、消极还是中性（无观点）（对于手机和相机，数据中只有消极和积极极性）。

(2) **(ACL) 引文意图分类**定义如下：给定一个引用句（包含引文的句子），判断该句子表达的引文功能是否属于“背景”、“动机”、“用途”、“扩展”或“比较或对比未来”。

(3) **（人工智能）关系分类**定义如下：给定一个包含一对实体的句内词序列跨度，判断该跨度表达的关系是否属于“特征”“连词”“评估对象”“下位词”“用途”“部分”或“比较”。

(4) **（PubMed）化学物质-蛋白质相互作用分类**定义如下：给定一个包含一对化学物质和蛋白质的跨度，判断该跨度表达的化学物质-蛋白质相互作用是否属于“下调剂”“底物”“间接上调剂”“间接下调剂”“激动剂”“激活剂”“产物”“激动剂-激活剂”“抑制剂”“上调剂”“底物产物”“激动剂-抑制剂”或“拮抗剂”。

## B 标准差

表4 报告了表2 （正文中）中相应结果的标准差，这些结果来自 DGA 及所考察的基线在 5 个不同随机种子下运行的实验。我们可以看到 DGA 的结果稳定。一些基线（例如，人工智能领域的 RoBERTa、相机领域的 MLM，以及 ACL 中的 MLM+TaCL）可能会有相当大的标准差。

表5报告了 DGA 与所考虑基线在 5 个随机种子运行中的对应结果标准差（见主论文中的表3）。可以看出，DGA 的结果很稳定。一些基线（例如 Camera 上的 DGA（随机掩码）和 DGA（无对比））的标准差可能相当大。