# 一种基于预训练模型掩码 Aspect 术语的数据增强方法

石晓瑞 [1]
SHI Xiaorui

**摘 要** 数据增强是解决低资源场景下数据稀缺问题的有效方案。然而，当应用于诸如方面术语提取（ATE）之类的词级别任务时，数据增强方法通常会遭受词标签不对齐的问题，从而导致效果不理想。对此提出了掩码方面语言建模（MALM）作为 ATE 的新型数据增强框架。为了缓解标记、标签错位问题，将 ATE 标签显式注入到句子上下文中，由此经过微调的 MALM 能够显式地调整标签信息来预测掩码的方面标记。因此，MALM 可帮助生成具有新方面的高质量增强数据，提供丰富的层面方面知识。此外，提出了一个两阶段的训练策略来整合这些合成数据。通过实验，证明了 MALM 在两个 ATE 数据集上的有效性，相比基线方法，所提出的 MALM 有显著的性能改进。

**关键词** 数据增强；Aspect 术语提取；预训练模型；掩码方面语言建模；MALM 方法

doi：10.3969/j.issn.1672-9528.2024.02.022

## 0 引言

Aspect 术语提取（ATE）是一项基本的自然语言处理任务，旨在检测句子中提到的方面术语。作为基于 Aspect 术语情感分析的子任务，它是 Aspect 术语级别情感分析的关键构建块。然而，该任务的标记数据量非常有限。因为为每种语言 / 领域手动注释足够的标记数据是昂贵的，所以资源低的 ATE[1-2] 在过去几年中越来越受到研究界的关注。作为低资源场景中数据稀缺的有效解决方案，数据增强通过应用标签保留转换来扩大训练集。

自然语言处理的典型数据增强方法包括：（1）修改词层面 [3]；（2）回译 [4]。尽管它们在句子级别任务上有效，但在应用于 ATE 等词级别任务时，会遇到词标签错位问题。更具体地说，单词级别的修改可能会用与原始标签不匹配的替代方案替换一个方面。反向翻译在很大程度上创建了保留原始内容的增强文本。然而，它依赖于外部单词对齐工具，用于将标签从原始输入传播到增强文本，这已被证明是容易出错的。

为了在词级别任务上增强应用数据，DAI 等人 [5] 提出用同一类的现有方面随机替换句子内的实体。尽管他们的方法避免了词标签错位问题，但 Aspect 术语的多样性并没有增加。此外，被替换的实体可能并不适合原始上下文。LI 等人 [6] 通过仅使上下文多样化来避免词标签错位问题。SONG 等 人 [7] 使用 MASS 替换上下文（具有“O”标签）词，并使 Aspect 术语完全不变。然而，根据 LIN 等人 [8] 对 ATE 的评估结果来看，通过上下文的增强对基于预训练语言模型 LM 的 ATE 模型只有微弱的性能提升。

图 1 为不同掩码策略效果对比。“Add Aspect”表示用额外真实样本中的 Aspect 术语替换当前句内的 Aspect 术语，“Add Context”则表示替换当前句子内容，保持当前 Aspect 术语不变。

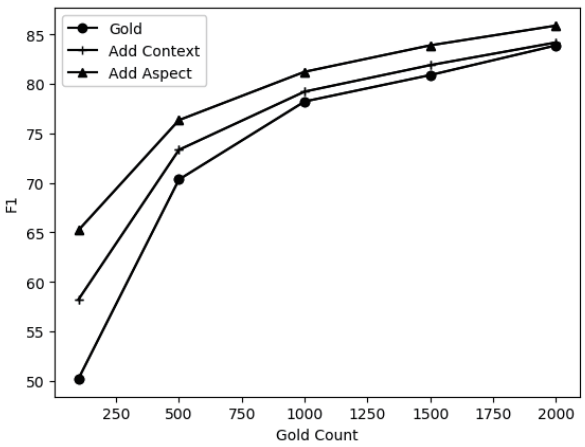


图 1 不同掩码策略效果对比

本文在低资源 ATE 上的初步结果（见图 1）表明，在训练数据中多样化 Aspect 术语比引入更多上下文模式更有效。受上述观察的启发，本文提出了掩码方面语言建模（MALM）作为低资源 ATE 的数据增强框架，它可生成具有不同方面术语的增强数据，同时避免了词标记标签错位问题。本文提出的 MALM 建立在预训练语言模型的掩码语言模型（MLM）

1. 北京安融汇达科技有限公司 北京 100872

之上，并且它对掩码的训练句子进行了进一步微调，只有 Aspect 术语标记被随机屏蔽，以促进面向 Aspect 术语的标记替换。使用微调的 MLM 简单地屏蔽和替换 Aspect 术语标记是不够的，因为预测的 Aspect 术语可能与原始标签不一致。

以图 2 所示的句子为例，在掩盖 Aspect 术语“pad penang”后，微调后的 MLM 可以将其预测为“提供食物”。这种预测符合上下文，但与原始标签不一致。为了避免错位，本文的 MALM 还引入了标记序列线性化策略，该策略分别在每个方面标记之前和之后插入一个标记词，并将插入的标记词视为掩码语言建模期间的正常上下文标记。因此，掩码标记的预测不仅可以根据上下文，还可以根据 Aspect 术语的标签来引入标签信息。同时，设计了一个两阶段的训练策略来整合标签增强的 ATE 数据，即首先在合成数据上训练 MALM，然后逐渐进入原始真实数据。通过这些方式，本文的 MALM 不仅可以利用来自预训练的丰富知识来增加 Aspect 术语的多样性，同时大大减少了标记词的标签错位问题。

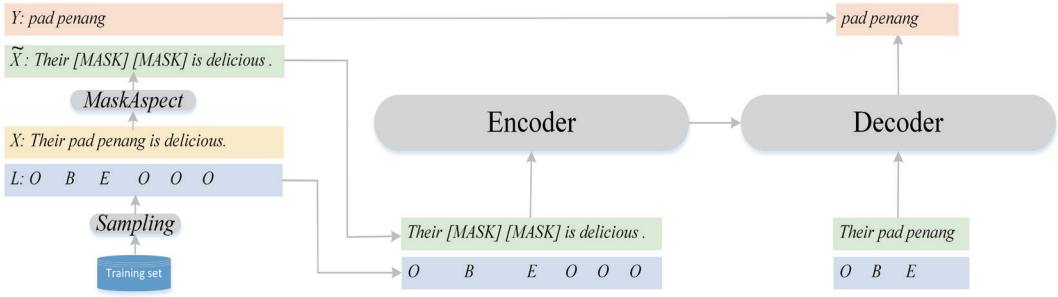


图 2 本文的数据增强方法框架

本文的主要贡献总结如下。

（1）提出了一个新的框架，该框架同时利用句子上下文和新的 Aspect 术语来增强数据，在多个英文 ATE 数据集上评估时，本文模型始终都取得了显著的改进。

（2）提出的标记序列线性化策略，能够有效缓解增强过程中标记词标签错位的问题。

（3）设计了两阶段的训练策略，逐步从合成的 ATE 数据过渡到原始的真实数据，以达到更好的性能。

本文引言介绍数据增强的相关方法和研究现状。第 1 节介绍本文构建的基于掩码 Aspect 术语的语言建模数据增强方法。第 2 节通过对比实验验证了所提模型的有效性。第 3 节总结全文。

## 1 基于预训练模型掩码 Aspect 术语的数据增强方法

图 2 展示了本文提出的数据增强框架的工作流程。首先对模型框架进行简单的介绍（第 1.1 节）；然后，本文在线性化序列上微调提出的 MALM（第 1.2 节）；最后通过掩码方面预测生成不同方面来创建增强数据（第 1.3 节）。该数据首先用于训练 ATE 模型，然后将原始真实数据用于最终训练 ATE 模型，以更好地提高性能（第 1.4 节）。

### 1.1 MALM 方法

本文模型基于预先训练的 transformerr 架构，如 BART 和 T5。编码器由 L 个变换器编码器块组成，包括多头自注意机制、前馈神经网络和层归一化。首先，解码器采用与编码器类似的结构，每个解码器层包括交叉注意子层，交叉关注子层旨在接收来自编码器的输出隐藏状态。然后，以自回归的方式生成输出目标序列。最后，在每个解码时间步长产生预测词，并通过交叉熵损失进行监督训练。

### 1.2 微调 MALM

为了最大限度地减少与原始标签不兼容的标记词的生成数量，设计了一种标记序列线性化策略，以在掩码语言建模期间明确考虑标签信息。具体来说，如图 2 所示，在每个 Aspect 术语之前和之后添加标签标记，并将它们视为正常的上下文标记。产生的线性化序列用于进一步微调本文的 MALM，使其预测时可以利用这些插入的标签标记信息。请注意，将标签标记的嵌入初始化为与标签名称语义相关的标记（例如，⟨B⟩ 表示“开始”）。通过这样的方式，线性化序列在语义上更接近自然句子，并且可以降低对线性化序列进行微调的难度。

与 MLM 不同，MALM 在微调期间仅屏蔽了 Aspect 术语标记信息。在每个微调开始时，以掩码率 $\eta$ 随机掩码线性化句子 $X$ 中的 *Aspect* 术语标记。然后，给定掩码后的句子 $\tilde{X}$ 作为输入，通过最大化被掩码的 Aspect 方面标记的概率来训练 MALM 模型，即重构线性化序列 $X$：

$$\max_{\theta} \log p_{\theta}(X|\tilde{X}) \approx \sum_{i=1}^{n} m_i \log p_{\theta}(x_i|\tilde{X}) \tag{1}$$

### 1.3 数据生成

为了生成增强的 ATE 训练数据，应用微调后的 MALM 来替换原始训练样本中的 Aspect 术语。具体来说，给定一个掩码后的序列，MALM 输出词汇表中每个标记词作为掩码 Aspect 术语的概率。然而，由于 MALM 在相同的训练集上进行了微调，直接选择最可能的 Aspect 术语作为替换很可

能会返回原始训练样本中的掩码 Aspect 术语标记，并且可能无法产生新的增强句子。因此，从概率分布的前 $k$ 个最可能的分布中随机抽样替换。首先给定掩码标记的概率分布 $P(x_i|\tilde{X})$，选择 $k$ 个最可能的候选者的集合 $V_ik \subseteq V$。然后，通过来自 $V_ik$ 的随机抽样来获取替换 $x_i$。最后，在获得生成的序列后，移除标签标记并将剩余部分用作增强训练样本。对于原始训练集中的每个句子，重复上述生成过程 $R$ 轮以生成 $R$ 个增强样本。

为了增加数据的多样性，采用了与训练时不同的掩码策略。对于包含 $n$ 个标记的每个 Aspect 术语，从高斯分布 $N(\mu, \sigma^2)$ 中随机采样一个动态掩码率 $\varepsilon$，其中高斯方差 $\sigma^2$ 设置为 $1/n^2$。因此，同一个句子在每一轮 $R$ 扩充中都会有不同的掩码结果，从而产生更多不同的扩充样本数据。

为了从增强数据中去除噪声和信息量较少的样本，生成的增强数据会经过后处理。具体来说，使用已有的真实训练样本训练 ATE 模型，并使用它自动为每个数据增强的句子分配 ATE 标签。仅保留那些预测标签与其原始标签一致的增强句子作为最终的增强样本。

### 1.4 ATE 模型训练

为了有效地结合后处理的增强训练样本，首先在这些增强样本上训练 ATE 模型。然后，在真实的训练集上再次训练 ATE 模型。这样，不仅可以利用这些增强数据，还可以消除这些合成数据带来的负面影响。

## 2 实验分析

### 2.1 实验数据

本文在两个广泛使用的数据集上进行实验，来自 SemEval 2014 Task 4 的笔记本电脑数据集和来自 SemEval 2016 Task 5[9] 的餐厅评论数据集，用于评估。 两个数据集的统计数据如表 1 所示，清楚地表明两个数据集中只有有限数量的样本。

表 1 实验数据集

| 数据集 | 训练集 | | 测试集 | |
|---|---|---|---|---|
| | 句子数目 | 术语数目 | 句子数目 | 术语数目 |
| Laptop | 3045 | 2358 | 800 | 654 |
| Restaurant | 2000 | 1743 | 676 | 622 |

### 2.2 基线模型

为了对比本文的数据增强方法，使用原始训练集和增强训练集来训练几个 ATE 模型。 这些模型的详细信息如下。

BiLSTM-CRF 是一种流行的序列标记任务模型。它的结构包括一个 BiLSTM，然后跟一个 CRF 层 [10]。该模型的词嵌入由 GloVe-840B-300d 初始化并在训练期间固定。隐藏大小设置为 300，使用 adam[11]，学习率为 1e-4，L2 权重衰减为 1e-5，来训练模型。

Seq2Seq4ATE[12] 是首次尝试将序列到序列模型应用于 Aspect 术语的提取。编码器和解码器均采用 GRU。编码器将源语句作为输入，解码器生成标签序列作为结果。该方法还配备有门控单元网络和位置感知注意网络。

BERT for token classification[13] 使用线性层的预训练得到。使用公开代码实现该模型，并使用预训练的 BERT-BASE-UNCASE 模型初始化其参数。在以下段落中，将该模型称为 BERT-FTC。

DE-CNN[14] 使用两种类型的词嵌入：通用嵌入和领域特定嵌入。前者采用 GloVe-840B-300d，后者在评论语料库上进行训练，它们被拼接到一起并送进 4 个堆叠卷积层的 CNN 模型。

BERT-PT[15] 利用预训练模型 BERT 的权重进行初始化。为了适应 ABSA 领域知识和任务特定知识，在大规模无监督领域数据集和机器阅读理解数据集上对其进行后训练。

DECNN+SoftProtoE[16] 使用软模板的方法来解决方面词和上下文词之间表现的长尾分布问题，并取得了较好的效果。

BERT-RC[17] 提出采用 MRC 的方式来联合训练 ATE 和情感分析任务，以达到端到端训练的目的，表现出了良好的性能。

PSTD[18] 受课程学习的启发，提出将传统的自我训练提炼为渐进式的自我训练方法。具体地说，基本模型在每次迭代时推断渐进子集上的伪标签，随着迭代的进行，子集中的样本变得越来越难，数量也越来越多。到目前为止，这是 ATE 提取的最新方法。

上述模型都是开源的，本文实验中使用了它们的默认设置。

### 2.3 实验结果与分析

将原始训练集与四个生成的数据集中的每一个数据集结合起来，并获得四个增强训练集，每个训练集的大小都是原始训练集的两倍。对于每个模型，分别在四个增强训练集上对其进行训练，并在测试集上取它们的平均 $F_1$ 分数。通过将此分数与在原始训练集上训练的模型进行比较，可以对比增强数据是否改进了模型。

如表 2 所示，所有利用增强数据的模型效果都得到了改进。即使对于最先进的 BERT-RC 和 PSTD 模型，本文的增强数据也带来了显著的性能改进，这证实了该增强方法可以生成有用的句子，用于训练更强大的方面术语词提取模型。

表 2 不同数据集上的 $F_1$ 分数

| 模型 | Laptop | | Restaurant | |
|---|---|---|---|---|
| | 原样本 | 增强样本 | 原样本 | 增强样本 |
| BiLSTM-CRF | 73.42 | **74.89** | 69.16 | **71.88** |
| Seq2Seq4ATE | 76.68 | **79.56** | 73.71 | **74.79** |
| BERT-FTC | 79.39 | **81.78** | 74.75 | **76.49** |
| DE-CNN | 81.08 | **81.95** | 74.52 | **75.78** |
| BERT-PT | 84.59 | **85.87** | 77.49 | **80.91** |
| DECNN+SoftProtoE | 83.19 | **85.23** | 76.98 | **79.98** |
| BERT-RC | 81.84 | **83.89** | 75.47 | **78.56** |
| PSTD | 86.91 | **88.41** | 82.56 | **84.50** |

2.4 不同的增强方法对比

为了对比不同的增强方法，进行了如下实验，如表 3 所示。DATA_synonym 表示通过用同义词随机替换标记而获得的数据集。DATA_BERT、DATA_GPT-2、DATA_Context 和 DATA_Aspect 分别表示由 BERT、GPT-2、掩码上下文和掩码方面（本文的增强方法）生成的数据集。BERT 被用作实现模型，较低的流利性分数意味着更流利的句子。

表 3 不同的增强方法对比结果（$F_1$/BLEU/ 流畅度）

| 数据集 | Laptop | Restaurant |
|---|---|---|
| 原数据 | 81.08/100.00/203 | 74.75/100.00/158 |
| DATA_synonym | 76.93/76.68/558 | 73.74/78.22/438 |
| DATA_BERT | 79.32/69.93/461 | 74.30/68.57/435 |
| DATA_GPT-2 | 66.09/54.86/328 | 63.30/58.60/314 |
| DATA_Context | 80.61/69.70/242 | 75.41/68.42/257 |
| DATA_Aspect | 82.69/50.33/150 | 77.98/50.24/160 |

GPT-2 的 $F_1$ 分数是最差的，因为它的召回分数很低。这符合 GPT-2 的架构和语言建模目标，它没有编码器来编码标签信息。在这种情况下，解码步骤是不可控的，无法生成适合标签序列的句子。相比之下，本文的框架同时包含一个用于编码句子和标签序列的编码器，以及一个根据编码器输出生成句子的解码器。也就是说，该解码器同时利用了上下文信息和方面标签信息，使数据增强是有条件的和可控的。

BERT 在这项任务中的流畅度表现最差。这可以归因于它在生成过程中的独立性假设，这意味着所有被屏蔽的词都是独立重建的，可能导致不连贯的词序列。相比之下，本文的方法以自回归方式生成序列，每个解码步骤都基于其前一步的结果，确保生成的新句子是流畅的。

基于替换的方法没有考虑句子上下文，导致流利度得分很差。此外，在 WordNet 等词汇数据库中，同义词的选择也有限。因此，这种基于替换的方法只能产生有限多样性的句子，较低的 BLEU 分数也证实了这点。

总而言之，本文数据增强模型从其编码器—解码器架构和掩码序列到序列生成机制中受益匪浅，该机制是可控的，以确保用于方面词提取的数据增强是合理有效的。

2.5 参数影响分析

为了确定微调掩码率 $\eta$ 和生成掩码参数 $\mu$ 的最佳设置，对 [0.3, 0.5, 0.7] 范围内的两个超参数进行网格搜索。按照前文方法微调 MELM 并在 Laptop 上生成英语增强数据。增强数据用于训练 ATE 模型，并记录其在英语开发集上的性能。如表 4 所示，当 $\eta = 0.7$ 和 $\mu = 0.5$ 时，实现了最佳开发集 $F_1$，并将其用于本工作的其余部分。

表 4 在 Laptop 数据集上的开发集 $F1$ 分数，使用 BERT-PT 模型调试超参数 $\mu$、$\eta$

| | $\eta$ | | |
|---|---|---|---|
| | 0.3 | 0.6 | 0.7 |
| $\mu$= 0.3 | 84.90 | 84.64 | 85.08 |
| $\mu$=0.5 | 85.16 | 85.06 | **85.87** |
| $\mu$=0.7 | 84.94 | 85.09 | 85.37 |

增强轮数 $R$。合并来自多轮的增强数据会增加实体的多样性，直到它在某个点饱和。继续添加更多的增强数据开始放大增强数据中的噪声并导致性能下降。为了确定最佳增强轮数 $R$，将不同数量的增强数据与英语黄金数据合并以训练 ATE 标注器，$R$ 的范围是 1 ～ 6。如表 5 所示，开发集 $F_1$ 随着增加的增强数据量直到 $R$=3，并开始进一步下降。因此，所有实验选择 $R$=3。

表 5 在 Laptop 数据集上的开发集 $F_1$ 分数，使用 BERT-PT 模型调试增强轮数

| R | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Dev $F_1$ | 85.35 | 85.36 | **85.87** | 85.72 | 85.59 | 85.39 |

2.6 案例分析

表 6 中展示了几个增强示例，以更直观地说明本文的增强方法的效果。观察到掩码片段的内容在增强后可以从它们的原始形式发生较大变化。在某些情况下，情绪极性甚至是相反的。尽管如此，新的上下文仍然适用于方面术语，使它们成为合格的并且多样化的方面术语词提取的新训练样本。原句和增强样本里加粗文本分别表示屏蔽的方面项和生成的片段。

表 6 通过本文的增强方法生成的示例

| |
|---|
| 原句：Also, the **space bar** makes a noisy click every time you use it.<br>增强样本：Also, the **windows 8** makes a noisy click every time you use it. |
| 原句：The **hinge design** forced you to place various connections all around the computer, left right ...<br>增强样本： The **MacBook Pro** forced you to place various connections all around the computer, left right ... |
| 原句：Their **pad penang** is delicious and everything else is fantastic.<br>增强样本： Their **cosi sandwiches** is delicious and everything else is fantastic. |
| 原句： I am learning the finger options for the **mousepad** that allow for quicker browsing of web pages.<br>增强样本： I am learning the finger options for the *mouse* that allow for quicker browsing of web pages. |
| 原句： I charge it at night and skip taking the *wires* with me because of the good **battery life**.<br>增强样本：I charge it at night and skip taking the **cord** with me because of the good **portable battery**. |

## 3 结语

在本文中，提出了一种用于方面术语提取的条件数据增强方法。将其表述为条件生成问题，并提出了一个掩码的序列到序列生成模型来实现它。与现有的增强方法不同，本文的方法可以控制生成合格有效的句子，并允许产生更多多样化的新句子。本文提出的 MALM 作为数据增强方面术语提取的框架，通过标记序列线性化，在预测掩码 Aspect 术语时能够明确地使用标签信息。因此，本文的 MALM 有效地缓解了标记标签错位问题，并通过利用预训练语言模型生成具有新 Aspect 术语的增强样本。两个评论数据集的实验结果证实了该方法在这种条件增强场景中的有效性。同时，利用定性研究来分析这种增强方法的工作原理，并通过测试其他语言模型来解释为什么本文的掩码序列到序列生成框架是更加优越的。此外，所提出的增强方法并不是当前任务所独有的，同样可以应用于其他序列标记任务，例如分块和命名实体识别任务。

**参考文献：**


[1] LIANG Y, MENG F, ZHANG J, et al. An iterative multi-knowledge transfer network for aspect-based sentiment analysis[C]//Findings of the EMNLP 2021.New Orleans: Association for Computational Linguistics,2021: 1768-1780.

[2] 张雪 , 孙宏宇 , 辛东兴 , 等 . 自动术语提取研究综述 [J]. 软件学报 , 2020, 31(7): 2062-2094.

[3] KOBAYASHI S. Contextual augmentation: data augmentation by words with paradigmatic relations[C]//Proceedings of the NAACL. New Orleans: Association for Computational Linguistics,2018: 452-457.

[4] SENNRICH R, HADDOW B, BIRCH A. Improving neural machine translation models with monolingual data[C]//Proceedings of the ACL. New Orleans: Association for Computational Linguistics,2016: 86-96.

[5] DAI X, ADEL H. An analysis of simple data augmentation for named entity recognition[C]//Proceedings of the COLING. Stroudsburg:Association for Computational Linguistics, 2020: 3861-3867.

[6] LI K, CHEN C, QUAN X, et al. Conditional augmentation for aspect term extraction via masked sequence-to-sequence generation[C]//Proceedings of the ACL. Stroudsburg:Association for Computational Linguistics,2020: 7056-7066.

[7] SONG K, TAN X, QIN T, et al. MASS: masked sequence to sequence pre-training for language generation[C]//ICLR.New York:International Conference on Machine Learning,2019: 5926-5936.

[8] LIN H, LU Y, TANG J, et al. A rigorous study on named entity recognition: can fine-tuning pretrained model lead to the promised land?[C]//Proceedings of the EMNLP. Stroudsburg:Association for Computational Linguistics,2020: 7291-7300.

[9] PONTIKI M, GALANIS D, PAPAGEORGIOU H, et al. Semeval-2016 task 5: aspect based sentiment analysis[C]//International workshop on semantic evaluation. Stroudsburg:Association for Computational Linguistics,2016: 19-30.

[10] LAFFERTY J, MCCALLUM A, PEREIRA F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//International Conference on Machine Learning.San Francisco:Morgan Kaufmann Pub, 2001:282-289.

[11] KINGMA D P, BA J. Adam: a method for stochastic optimization[EB/OL].(2014-11-22)[2023-07-30].https://arxiv.org/abs/1412.6980.

[12] CHO K, VAN MERRIENBOER B, GÜLÇEHRE Ç, et al. Learning phrase representations using rnn encoder-decoder for statistical machine translation[C]//EMNLP.Stroudsburg:Association for Computational Linguistics, 2014:1724-1734.

# LPGEMM: 低精度通用矩阵乘法计算模拟框架研究

黄浩岚[1,2,3] 罗铁清[1] 文 梅[2,3] 曹亚松[2,3] 时 洋[2,3]
HUANG Haolan LUO Tieqing WEN Mei CAO Yasong SHI Yang

**摘 要** 通用矩阵乘（GEMM）算子是AI模型的核心计算，使用低精度数值格式加速GEMM对加速模型的推理和训练有重要影响。由于并不总是有合适的硬件可供选择，而且人们可能希望实验尚未在硬件中实现的新GEMM计算行为，但很难通过构建硬件的方式去进行不同计算行为的GEMM模拟，如何在算子内部进行细粒度模拟还没有被深入研究。通过提出LPGEMM——一个低精度GEMM计算模拟框架来模拟GEMM的计算过程，重新编写了GEMM算子，实现了可变分组累加长度以及低精度累加器，同时还实现了训练和推理全过程的GEMM相关数据统计，来支持用户探索模型精度的下限。实验结果证实了相较于此前的一些工作，所提出的方法模拟最高可减少56%的平均误差。

**关键词** 深度学习；用户探索模型；通用矩阵乘；低精度

doi：10.3969/j.issn.1672-9528.2024.02.023

## 0 引言

人们对神经网络表现能力的更高追求，推动了学界对大模型的探索，近年来的模型参数量已增长至千亿级。这同时促使了产业界对包括GPU和FPGA设备在内的高性能神经网络加速器，以及对降低模型的数值精度从而加速计算的追求。

1. 湖南中医药大学 湖南长沙 410208
2. 国防科技大学 湖南长沙 410073
3. 先进微处理器芯片与系统重点实验室 湖南长沙 410073

当前一代的深度神经网络（DNN），例如transformer和ResNet等，在很大程度上依赖于通用矩阵乘法（GEMM），GEMM在DNN模型中的计算占比高达95%以上[1]，GEMM的加速计算对整个模型计算的加速具有重要意义，低精度计算作为压缩与加速深度神经网络（DNN）中最成功的技术之一，被广泛应用于GEMM加速。

由于并不总是有合适的硬件可供选择，而且人们可能希望实验尚未在硬件中实现的新GEMM计算行为（例如用分组累加替代逐元素累加），人们很难通过构建硬件的方式去

[13] KENTON J D M W C, TOUTANOVA L K. BERT: Pre-training of deep bidirectional transformerrs for language understanding[C]//Proceedings of NAACL. Stroudsburg:Association for Computational Linguistics, 2019: 4171-4186.

[14] XU H, LIU B, SHU L, et al. Double embeddings and cnn-based sequence labeling for aspect extraction[C]//Proceedings of the ACL. Stroudsburg:Association for Computational Linguistics, 2018: 592-598.

[15] XU H, LIU B, SHU L, et al. BERT post-training for review reading comprehension and aspect-based sentiment analysis[C]//Proceedings of the NAACL. Stroudsburg:Association for Computational Linguistics, 2019: 2324-2335.

[16] CHEN Z, QIAN T. Enhancing aspect term extraction with soft prototypes[C]//Proceedings of the EMNLP.Stroudsburg:Association for Computational Linguistics, 2020: 2107-2117.

[17] MAO Y, SHEN Y, YU C, et al. A joint training dual-mrc framework for aspect based sentiment analysis[C]//Proceedings of the AAAI. Palo Alto:Association for the Advancement of Artificial Intelligence, 2021: 13543-13551.

[18] WANG Q, WEN Z, ZHAO Q, et al. Progressive self-training with discriminator for aspect term extraction[C]//Proceedings of the EMNLP.Stroudsburg:Association for Computational Linguistics, 2021: 257-268.

【作者简介】

石晓瑞（1993—），女，河南周口人，学士，研究方向：人工智能和实体提取。

（收稿日期：2023-11-09）