# idT5：多语言 T5 Transformer 的印度尼西亚语版本

Mukhlish Fuadi*Institut Teknologi Sepuluh Nopember* 电气工程系印度尼西亚泗水
*Universitas Islam Negeri Maulana Malik Ibrahim*印度尼西亚玛琅
muchad@uin-malang.ac.id

Adhi Dharma Wibawa*Institut Teknologi Sepuluh Nopember* 电气工程系医疗技术系印度尼西亚泗水
adhiosa@te.its.ac.id

Surya Sumpeno*Institut Teknologi Sepuluh Nopember* 电气工程系 计算机工程系印度尼西亚泗水 surya@te.its.ac.id

摘要——**印度尼西亚语有近2亿人使用，是世界上使用人数第十多的语言，但在自然语言处理（NLP）研究中却未得到充分关注。语言资源匮乏阻碍了此前针对印度尼西亚语的研究。Transformer 是一种新型架构，正迅速成为自然语言处理领域的主流，超越了卷积神经网络和循环神经网络等替代方案。T5（文本到文本迁移Transformer）是一种 Transformer 模型，可将所有基于文本的语言问题转换为英语的文本到文本格式。其多语言变体为 mT5（多语言 T5），已在跨语言的许多自然语言处理任务中展现出良好前景。然而，这种多语言模型的规模较大，不利于其在生产应用中的使用，因为此类应用有时只需要支持一种语言。本研究将 mT5 模型调整为仅支持一种语言——印度尼西亚语，从而得到一个专用于印度尼西亚语且体积更小的预训练 T5 模型。为了比较性能，我们采用相同的机制和数据集，对该模型和 mT5 模型进行了微调，以用于情感分析（SA）、问题生成（QG）和问答（QA）任务。基于我们模型的微调模型在情感分析任务上达到了 77.18% 的准确率，比基于 mT5 的模型高 8%，并且在问题生成和问答任务上取得了几乎相同的分数。结果证实，可以构建体积更小的预训练模型，在将模型大小最多缩减 58% 的同时，保持相当的表现。此外，生成的模型所需内存更少、加载速度更快，推理时间也更短。**

关键词——模型压缩、***Transformer***、预训练模型、问题生成、问答、情感分析

## I. 引言

尽管印度尼西亚语是全球使用人数排名第十的语言[1], ，但它在自然语言处理（NLP）研究中的代表性仍然不足。其中的障碍之一是基于印度尼西亚语的资源有限，包括面向印度尼西亚语的预训练语言模型数量有限。

自然语言处理结合语言学、计算和统计学研究，创造出能够像人类一样理解并回应语言的机器 [2]。在自然语言处理中，序列模型被广泛用于改善序列数据（如书面语和口语）的处理，因为它能够处理语言中的序列特性，而这些特性取决于词语之间的关系。循环神经网络是其中一种。

循环神经网络被广泛使用，因为它们采用长短期记忆网络 [3] 和门控循环单元[4] 架构。这两种单元都通过提供更直接的梯度反向传播途径来防止梯度消失问题。这种机制有助于处理较长句子时的计算。网络的高度灵活性使其能够解决各种问题 [5]。然而，这些模型并不完美，因为其固有的重复结构使其难以在多个进程中并行化，而且由于梯度缺失，在长从句中也会出现问题[6]。此外，顺序处理也使训练变得十分繁琐且耗时。

Transformer 是一种能够克服这些问题的新架构 [7]。Transformer 不会出现梯度消失，并且能够处理数据序列任务，同时高效处理远程依赖、促进任务并行化，并加快更大规模网络的训练。凭借将未标注数据中的知识迁移到下游任务的卓越能力，预训练的基于Transformer的语言模型已成为现代自然语言处理系统的重要组成部分。因此，Transformer 正迅速成为自然语言处理的主导架构[8]。

T5（文本到文本迁移Transformer）是一种Transformer变体，可将所有基于文本的语言问题转换为文本到文本格式[9]。这种方法切实可行，因为它无需更改模型架构，就能将知识从高资源任务迁移到低资源任务[10]。T5 模型最初针对英语进行了预训练，之后扩展到多语言设置；mT5 在多个跨语言理解基准上取得了先进的表现[11]。mT5 覆盖 101 种语言，包括印度尼西亚语，并为多个预训练模型提供了若干检查点。其中之一是 mT5-base，拥有 5.8 亿个参数，因此模型规模相对较大。

多语言Transformer是一种能够处理多种语言的自然语言处理模型。使用多语言Transformer的主要优势之一，是它能够通过单一模型处理多种语言；与使用多个单语言模型相比，这种方式可能更加高效且更具成本效益。近期研究结果表明，多语模型的表现可能优于单语模型，尤其是在低资源语言方面 [12]。然而，多语模型通常比单语言模型使用更多数据进行训练。其训练和使用的计算成本也可能更高。

一些现实世界的应用只需处理一种特定语言；因此，我们提出通过减少模型所处理的语言数量来缩减词汇表规模。在本文中，我们提取 mT5，以获得一个仅针对一种语言——印度尼西亚语——的更小型 Transformer 模型。本研究的贡献是一个用于印度尼西亚语的预训练 Transformer 模型。之后，可以针对印度尼西亚语中的各种自然语言处理任务对该模型进行微调。该方法也可以应用于其他语言。

下一节将介绍 Transformer mT5 模型，以及如何基于多语言模型生成针对印度尼西亚语的专用模型。随后，我们通过将所生成模型的性能与 mT5 模型本身进行比较，测试该模型。我们比较了两个模型在问题生成、问答和情感分析任务上的微调结果。此外，还比较了加载时间、推理时间和内存用量。最后，我们在本文结尾分析并总结实验结果。

## II. 相关工作

此前已有多项研究致力于压缩 Transformer 模型的规模。知识蒸馏 [13, 14] 是一种方法，旨在构建架构比原始模型更简单、同时能够模仿其行为的模型。具体做法是将大型教师网络学到的知识迁移给较小的学生模型。该方法已用于减少 BERT 模型的层数 [15]，并且能够在不降低性能的情况下有效压缩模型。模型蒸馏并不局限于简化架构，还 [16] 通过同时训练教师模型和学生模型来减小词汇表和嵌入维度。

[17] 提出的层剪枝策略无需重新训练模型，可直接应用于此前训练好的模型。该剪枝策略在 BERT、XLNet 和 RoBERTa 等广泛使用的架构上，将模型大小减少了 40%，推理时间缩短了 50%，同时保留了高达 98.2% 的原始准确率。该方法考虑了所有较低层，为实现师生架构提供了更好的方式。它不像 [15] 那样，每隔一层抽取一层来初始化学生模型。

[18] 还采用知识蒸馏，利用教师模型最后一个 Transformer 层中的自注意力分布和值之间的关系来指导学生模型训练，使其具备有效性和灵活性。在研究[19], 中，这种方法通过利用多头自注意力关系训练学生模型得到了推广，从而获得更细致的自注意力知识，并取消了对学生模型注意力中心数量的限制。本研究中从 BERT、RoBERTa 和 XLM-R 蒸馏得到的单语和多语言模型取得了具有竞争力的表现。

[20] 通过降低权重表示所用的精度值，减少模型的内存占用，从而将模型权重量化的方法应用于 BERT 模型的 Transformer 架构。不过，该方法在特定硬件设备上使用时才有效。

通过分析和比较，[21] 得出结论：量化和裁剪等传统模型压缩方法有益于 BERT。针对 BERT 的技术，例如知识蒸馏变体以及减少架构超参数数量的方法，也能取得有竞争力的结果。

对于多语言 Transfomer 模型，[22] 缩减了 mBERT 的词汇表，从而减少了主要位于嵌入层中的参数数量。该方法在不降低平均准确率的情况下，最多减少了 45% 的总参数量。该方法的另一个优点是，无需像 [23] 那样从头开始训练模型。[24] 也使用了类似的思路，从 mT5 中提取俄语，但目前尚无同行评审出版物。压缩过程只保留常用词元，并裁剪其余词元。本文采用相同的方法，构建一个规模更小、更适用于印度尼西亚语的模型。

## III. 方法论

多语言 T5（mT5）是 T5 的多语言变体，使用包含 101 种语言（包括印度尼西亚语）的 mC4（多语言巨型清洁爬取语料库）数据集进行了预训练 [11]。该模型包含多种语言，因此规模相当大。根据 Google AI 团队提供并发布于 Hugging Face 仓库网站上的 mT5 模型，mT5-base 版本的大小为 2.33 GB。最小的模型 mT5-small 大小为 1.2 GB。

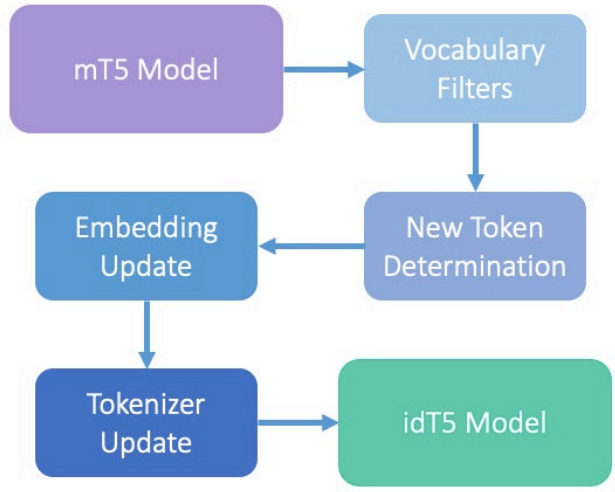


图 1. 提取流程。

为了获得印度尼西亚语版本的 mT5，我们从 mT5-base 中筛选出印度尼西亚语词汇表，然后更新嵌入层和分词器，以构建合适的模型。各阶段的详细流程如图 1 所示。

### A. 选择词汇表

我们使用原始分词器处理印度尼西亚语语料库。由于印度尼西亚语文本中经常使用英语，我们在模型中保留少量英语词元。为了计算不同词元的频率，我们从莱比锡语料库 [25] 中获取了印度尼西亚语和英语句子语料。

**词汇表筛选器**。根据印度尼西亚语语料库中的词元数量，我们发现模型词汇表中只有29%被使用，英语词汇表中的使用率为27%。有趣的是，

词汇表之间的重叠率超过76%。这可能是因为在印度尼西亚语文本中，有时会出现英语单词或其他使用拉丁字母的词语。此外，排名前20K的词元占印度尼西亚语语料库的98%以上。对于英语，统计结果也类似。

**新词元的确定**。根据词汇表筛选器，本研究将词元数量确定为 30K，其中包括原始分词器中排名靠前的 1K 个词元、英语词元中的 10K 个（已覆盖超过 95% 的英语语料库），以及 T5 使用的 100 个特殊词元。至于印度尼西亚语词元，其余词元尚未纳入这三组词元，因此词元总数为 30K，即多语言版本中 250K 个词元的 12%。至于其他被移除的词元，根据这 30K 个词元，纳入的印度尼西亚语词元共有 24,471 个。这是因为存在重叠词元。

### *B.* 更新模型

**嵌入更新**。为更新神经网络，我们将其输入和输出参数替换为预先定义的参数，从而将模型大小缩减了 58%，由 2330 MB 降至 977 MB。新模型有 2.44 亿个参数，相当于多语言模型的 42%。

**分词器更新**。由于 T5 使用 Sentencepiece 分词器，因此我们使用 Protobuf 表示来更新分词器。

生成的模型已上传至 Transformers Hub，方便自然语言处理社区使用，尤其是与印度尼西亚语相关的社区[1]。该模型是一个预训练模型，与 mT5 模型类似。因此，要完成自然语言处理任务，需要对其进行微调，因为 mT5 仅通过无监督任务进行训练，以预测缺失的词语。

预训练模型是指已在大型数据集上完成训练的模型，无需进一步训练即可用于各种自然语言任务[26]。预训练模型很有用，因为从头开始训练模型可能耗时较长，而且会增加计算成本。微调则是一种自然语言处理技术，用于使预训练模型适应特定任务[27]。在微调过程中，会使用新的数据集进一步训练预训练模型，以提升其在特定任务上的表现。与从头开始训练模型相比，这种方式通常效率更高。

## IV. 结果与讨论

我们得到的模型（idT5）是对 mT5 的改编，因此它和 mT5 一样，也是一个预训练模型，必须经过微调才能完成自然语言处理任务。idT5 模型是 mT5 的较小版本，因此其性能可能不如原版 mT5。本节将通过简单的比较，了解 idT5 在针对自然语言处理任务进行微调后与 mT5 相比的表现。我们针对两种模型微调了多个自然语言处理任务：问题生成（QG）、问答（QA）和情感分析。为了公平、准确地比较性能，我们在每个模型中使用完全相同的机制、数据集和配置。所有任务的微调都只进行一个训练轮次，因为这样做只是为了比较两种模型在相同条件下的性能，而不是为了获得最佳性能。所有微调和比较均使用 Hugging Face Transformer Trainer 类 [8] 与 PyTorch [28] 在 Google Colab 上进行。

### *A.* 问题生成

问题生成是一项自然语言处理任务，模型经过训练后，能够根据给定的输入文本自动生成问题 [29]。这项技术可用于多种应用场景，例如提升阅读理解能力或构建问答系统。问题生成是自然语言处理中最具挑战性的任务之一 [30-32]。其中的难点之一是，系统必须能够准确识别文本，从而生成结构良好、切题且可回答的问题 [33]。

在问题生成和问答任务中，为了生成基于 idT5 的模型和基于 mT5 的模型，我们采用了与 [34] 相同的方法。数据集方面，我们使用了 TyDi QA 黄金篇章任务数据集，这是一个用于阅读理解的数据集 [35]。该数据集包含 11 种语言，但我们只使用印度尼西亚语数据集。训练集包含 5702 个印度尼西亚语语境及数量相同的问题。开发集则包含 565 个语境及数量相同的问题。我们使用 TyDi QA 训练集和开发集对问题生成和问答任务进行微调，超参数为：学习率 1e-4，训练轮次 1。

表 I. 根据 TyDiQA 开发集生成的问题

| **1** | |
|---|---|
| **上下文 & 答案** | 仙人掌是对以下成员的称呼<br>开花植物科 仙人掌科。<br><br>（仙人掌是对开花植物的称呼<br>科 仙人掌科。） |
| **目标问题** | 仙人掌植物的学名是什么?<br>（仙人掌植物的学名是什么？） |
| ***mT5*** | 仙人掌是什么意思?<br>（仙人掌是什么意思？） |
| ***idT5*** | 仙人掌是什么?<br>（什么是 *Cactus*？） |
| **2** | |
| **上下文与答案** | 这种病毒名为 人类免疫缺陷<br>病毒（简称 HIV），它是一种<br>会削弱人体免疫力的病毒。人<br>感染这种病毒的人会更容易患上<br>机会性感染，也更容易患上肿瘤。<br>尽管现有的治疗方法可以<br>减缓病毒的进展速度，但这种疾病<br>仍无法彻底治愈。<br><br>（这种病毒本身称为人类免疫缺陷病毒<br>（简称 *HIV*），这是一种会削弱<br>人体免疫力的病毒。接触过这种病毒的人容易受到<br>机会性感染的侵袭<br>或容易患上肿瘤。尽管现有的治疗方法可以<br>减缓病毒的发展速度，但这种疾病<br>尚未得到彻底治愈。） |
| **目标问题** | HIV 的全称是什么?<br>（*HIV* 代表什么？） |
| ***mT5*** | HIV 是什么意思?<br>（*HIV* 是什么意思？） |
| ***idT5*** | 什么病毒会削弱人体的<br>免疫力?<br>（什么病毒会削弱人体的免疫力?<br>人体？） |

[1] https://huggingface.co/muchad/idt5-base

为了比较这两个微调模型，我们使用 TyDi QA 开发集对它们进行了评估。为了量化这两个模型的性能，我们使用了 [36] 提供的评估包，该评估包会计算双语评估替补（BLEU）[37] 和面向摘要评估的召回率导向替补（ROUGE）[38]。BLEU 通过统计机器生成的问题中有多少词语出现在人工参考问题中来衡量精确率。ROUGE 衡量召回率，即人工参考问题中有多少词语出现在机器生成的问题中。生成问题示例见表 I。

表 II.　　问题生成性能比较

| 模型 | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE-L |
|---|---|---|---|---|---|
| mT5 | **34.39** | 26.61 | 21.32 | 17.25 | **36.63** |
| idT5 | 34.35 | **27.08** | **21.79** | **17.71** | 36.40 |

问题生成任务的性能比较结果如表 II 所示。使用 BLEU 和 ROUGE-L 指标比较得分时，两个微调模型（基于 mT5 和基于 idT5）的得分几乎完全相同，差异极小。基于 mT5 的模型在 BLEU-1 和 ROUGE-L 上表现更好，而基于 idT5 的模型在 BLEU-2、BLEU-3 和 BLEU-4 上表现更好。总体而言，这些结果表明，尽管基于 idT5 的模型规模更小，但其在问题生成任务上的性能与基于 mT5 的模型相当。

## *B.* 问答

问答系统旨在自动回答人类以自然语言提出的问题。这些系统可以利用预结构化数据库或自然语言文档集合来生成答案[39]。问答系统通常用于创建能够理解并响应用户查询的智能助手或聊天机器人 [40]。问答系统通常结合自然语言理解、信息检索和知识表示技术，以提供准确的答案。

表 III.　　问答性能比较

| 模型 | EM | F1 |
|---|---|---|
| mT5 | **60.88** | **74.92** |
| idT5 | 60.53 | 73.64 |

在问答任务中，我们以 TyDi QA 开发集中的答案为参照，使用 F1 分数和精确匹配分数评估基于 idT5 和 mT5 的模型生成的答案。精确匹配（EM）指标衡量预测结果与真实答案完全匹配的百分比。相比之下，F1 分数指标评估预测结果与真实答案之间的平均重合度[41]。基于 mT5 和 idT5 的问答模型的分数对比如表 III 所示。在此任务中，基于 mT5 的模型优于基于 idT5 的模型，但两者在精确匹配分数和 F1 分数上的差异都很小。表 IV 展示了两个模型生成的答案示例。

表 IV.　　TyDiQA 开发集生成的答案

| **1** | |
|---|---|
| **上下文 & 目标答案** | Rumah Baanjung （Ba'anjung）是班查尔族和巴古迈达雅克族传统住宅的统称。[1] 班查尔族通常将其住宅称为 Rumah Banjar 或 Rumah Bahari。<br><br>（*Rumah Baanjung* （*Ba'anjung*）是……的统称 班查尔族和达雅克族的传统房屋 巴古迈部族。*[1]* 班查尔族通常将他们的传统房屋称为班查尔房屋或海上住宅。） |
| **问题** | 班查尔族传统房屋叫什么？（班贾尔传统住宅叫什么名字？） |
| ***mT5*** | Baanjung（*Baanjung*） |
| ***idT5*** | 海上住宅（海上住宅） |
| **2** | |
| **背景与目标答案** | 黑莓首次在印度尼西亚推出的时间是 2004年12月中旬 由运营商 Indosat 和 Starhub 公司。Starhub 公司是 RIM 的体现，RIM 是主要合作伙伴黑莓。随后，黑莓市场上又出现了国内另外两家大型运营商，即 Excelcom 和 Telkomsel。[2]<br><br>*(*黑莓于年 中 *2004年12月* 由 *Indosat* 运营商和 *Starhub* 公司推出。*Starhub* 公司是 *RIM* 的化身 即黑莓的主要合作伙伴。黑莓 市场随后因另外两家大型运营商的加入而活跃起来，它们是该国的 *Excelcom* 和 *Telkomsel*。*[2])* |
| **问题** | 黑莓是什么时候开始进入印度尼西亚的？*(*黑莓什么时候开始进入印度尼西亚？*)* |
| ***mT5*** | 2004年12月 *(2004年12月)* |
| ***idT5*** | 2004年12月中旬 *(2004年12月中旬)* |

## *C.* 情感分析

情感分析是一种文本分类方法，根据情感倾向（如观点或情绪）对文本进行分类 [42]。这可用于各种应用，例如识别客户评论中的情感，或检测社交媒体帖子中的情绪基调。情感分析通常使用在大型数据集上训练的机器学习算法来完成，这些数据集中的文本均标注了相应的情感。随后，这些算法便可根据情感自动对新文本进行分类。

在此任务中，我们使用了 SmSA 数据集[43]，，该数据集收集了来自印度尼西亚各类在线平台的评论和评价。该数据集包含三种可能的情感：正面、负面和中性。不过，在本次评估中，我们仅使用了正面和负面数据。因此，我们得到的训练集包含 9852 个句子（总计 11000 个），验证集包含 1129 个句子（总计 1260 个），测试集包含 412 个句子（总计 500 个）。为了在 idT5 和 mT5 模型上对情感分析任务进行微调，我们采用以下超参数：学习率为 3e-4，训练轮次为 1。

表 V.　　情感分析性能比较

| 模型 | 准确率 | F1 |
|---|---|---|
| mT5 | 68.93 | 67.51 |
| idT5 | **77.18** | **77.01** |

为了比较各模型在情感分析任务上的性能，我们采用了准确率和 F1 宏平均分数。表 V 显示，基于 idT5 的模型在这两项指标上均优于基于 mT5 的模型。基于 idT5 的模型在准确率和 F1 分数上均比基于 mT5 的模型高 8%。这些结果表明，在情感分析任务上，基于 idT5 的模型优于基于 mT5 的模型。

根据对三项任务性能的比较，结果表明，基于 idT5 的模型在问题生成任务上的表现与基于 mT5 的模型相近，但在情感分析任务上的表现优于基于 mT5 的模型。基于 mT5 的模型在问答任务上的表现略优于基于 idT5 的模型。总体而言，结果表明，体积比 mT5 模型小 58% 的基于 idT5 的模型，在特定自然语言处理任务上的表现可以达到与基于 mT5 的模型相当或更好的水平。

我们希望为某一种语言专门构建模型，因为有时其实现只需专注于该语言的任务。与多语言 Transformer 模型相比，单语言模型体积更小，也更节省内存。本文通过减少多语言 Transformer 模型的参数数量，构建了体积更小、内存需求更低的模型。内存限制至关重要，尤其是在公共云平台上部署 Transformer 时。此外，小模型的加载速度比大模型快，这也有助于提升应用程序的部署速度。我们采用的方法还可以缩短推理时间。

表 VI.　开销比较

| 模型 | 加载（秒） | 序列长度 = 128 | | 序列长度 = 512 | |
|---|---|---|---|---|---|
| | | 内存 (MB) | 推理（秒） | 内存 (MB) | 推理（秒） |
| mT5 | 12.11 | 4322 | 0.239 | 8094 | 1.103 |
| idT5 | **7.82** | **2170** | **0.158** | **3436** | **0.784** |

表 VI 展示了两个模型的加载时间、内存分配和推理时间。加载时间是在 Google Colab（2 个虚拟CPU @ 2.2GHz，13GB RAM）上测量的，每个模型测试 10 次并取平均值。idT5 模型的模型大小较小、参数较少，因此加载时间比 mT5 模型更短。

内存分配和推理时间是在 Google Colab 上使用批大小 = 8 进行测量的，该环境配备 GPU（Tesla T4 – 16GB）。测试采用了序列长度= 128 和序列长度 = 512。在这两种序列长度下，idT5 所需的内存分配都更少，甚至不到 mT5 模型所需内存的一半。对于 mT5 模型，序列长度为 512 时所需的内存比序列长度为 128 时多 46.6%。而对于 idT5，序列长度从 128 增至 512 时，所需内存仅增加 36.8%。

推理时间取 6 次测试的平均值。尽管此方法不会改变模型架构，但与 [22]，中提出的无法缩短推理时间的方法相比，它仍能缩短推理时间。

根据这项测试，特定语言模型 idT5 的加载时间更短，所需的内存分配更少，并且推理时间也比同类多语言模型 mT5 更短。这些结果表明，在某些场景下，特定语言模型可能比多语言模型更具性能优势。这类模型在针对特定语言需求部署模型、在公共云平台上部署模型，或面临内存限制时会很有帮助。

## V. 结论与未来工作

使用多语言 Transformer 的一项优势是，无需为每种语言单独使用特定模型，它就能处理多种语言的文本。这在需要处理多种语言文本的应用中很有用。然而，对于只需处理一种语言的任务，多语言模型的效率较低。模型体积较大，会影响内存分配和处理速度。在公共云平台上部署 Transformer 时，内存限制也是需要考虑的因素。

本文提出了一种简单的方法，可从多语言 Transformer 模型中创建针对特定语言的单语言模型。利用此方法，我们能够生成针对特定语言的模型，其体积小于原始多语言模型。我们使用 mT5-base Transformer 模型展示了该方法，并提取了印度尼西亚语和少量英语，因为英语词汇常用于印度尼西亚语文本。最终模型比原始多语言模型小 58%，展示了此方法在创建更紧凑、针对特定语言的模型方面的潜力。

在本研究中，我们针对三项自然语言处理任务（问题生成、问答和情感分析）对较小模型和 mT5 模型进行了微调，以比较它们的性能。我们发现，较小模型在这些任务上的表现与 mT5 模型相近。此外，尽管模型架构没有改变，较小模型所需的内存更少，加载时间更短，推理时间也更短。这些结果表明，这种方法有望从多语言模型中创建更小、更高效的单语言模型。

作为未来的研究方向，值得在更广泛的模型和任务上评估这种方法，以观察结果是否一致。将这种方法应用于其他语言，也有助于创建适用于各种任务的特定语言模型。该模型有望推动印度尼西亚语 Transformer 在实际应用中的使用，我们计划在未来继续探索其潜力。

## 参考文献

[1] I. Ghosh。《排名：全球使用人数最多的100种语言》。https://www.visualcapitalist.com/100-most-spoken-languages/（访问于2022年11月25日）。[2] N. Ranjan、K. Mundada、K. Phaltane 和 S. Ahmad，《自然语言处理技术综述》，*International Journal of Computer Applications*，第134卷，第8期，第6-9页，2016年。[3] S. Hochreiter 和 J. Schmidhuber，《长短期记忆》，*Neural computation*，第9卷，第8期，第1735-1780页，1997年。[4] K. Cho等，《使用循环神经网络编码器-解码器学习用于统计机器翻译的短语表示》，*arXiv*预印本 *arXiv:1406.1078*，2014年。[5] K. Greff、R. K. Srivastava、J. Koutník、B. R. Steunebrink 和 J. Schmidhuber，《LSTM：搜索空间奥德赛》，*IEEE* 神经网络与学习系统汇刊，第28卷，第10期，第2222-2232页，2016年。

[6] A. Gillioz、J. Casas、E. Mugellini 和 O. Abou Khaled，《NLP任务的基于 Transformer 的模型概述》，载于2020年第15届计算机科学与信息系统会议（FedCSIS），2020年：IEEE，第179-183页。[7] A. Vaswani 等，《Attention Is All You Need》，arXiv /cs.CL/, 20172017年。[在线]。可访问：http://arxiv.org/abs/1706.03762。[8] T. Wolf 等，《Huggingface 的 transformers：最先进的自然语言处理》，arXiv预印本 arXiv:1910.03771，2019年。[9] C. Raffel 等，《通过统一的文本到文本 Transformer 探索迁移学习的极限》，arXiv预印本 arXiv:1910.10683，2019年。[10] A. Elmadany 和 M. Abdul-Mageed，《AraT5：用于阿拉伯语生成的文本到文本 Transformer》，载于第60届计算语言学协会年会论文集（第1卷：长篇论文），2022年，第628-647页。[11] L. Xue 等，《mT5：大规模多语言预训练文本到文本 Transformer》，arXiv预印本 arXiv:2010.11934，2020年。[12] A. Conneau 等，《大规模无监督跨语言表示学习》，arXiv预印本 arXiv:1911.02116，2019年。[13] C. Bucilu ă、R. Caruana 和 A. Niculescu-Mizil，《模型压缩》，载于第12届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集，2006年，第535-541页。[14] G. Hinton、O. Vinyals 和 J. Dean，《蒸馏神经网络中的知识》，arXiv预印本 arXiv:1503.02531，第2卷，第7期，2015年。[15] S. Sun、Y. Cheng、Z. Gan 和 J. Liu，《用于 BERT 模型压缩的渐进式知识蒸馏》，arXiv预印本 arXiv:1908.09355，2019年。[16] S. Zhao、R. Gupta、Y. Song 和 D. Zhou，《利用最优子词和共享投影实现极致语言模型压缩》，2019年。[17] H. Sajjad、F. Dalvi、N. Durrani 和 P. Nakov，《 Poor man's bert：更小、更快的 Transformer 模型》，arXiv 预印本 arXiv:2004.03844，2020年。[18] W. Wang、F. Wei、L. Dong、H. Bao、N. Yang 和 M. Zhou，《MiniLM：用于预训练 Transformer 任务无关压缩的深度自注意力蒸馏》，神经信息处理系统进展，第33卷，第5776-5788页，2020年。[19] W. Wang、H. Bao、S. Huang、L. Dong 和 F. Wei，《MiniLMv2：用于压缩预训练 Transformer 的多头自注意力关系蒸馏》，arXiv预印本 arXiv:2012.15828，2020年。[20] S. Shen 等，《Q-BERT：基于海森矩阵的 BERT 超低精度量化》，载于 AAAI 人工智能会议论文集，2020年，第34卷，第05期，第8815-8821页。[21] P. Ganesh 等，《压缩大规模基于 Transformer 的模型：以 BERT 为例》，计算语言学协会会刊，第 9 卷，第 1061-1080 页，2021 年。[22] A. Abdaoui、C. Pradel 和 G. Sigel，《按需加载：多语言 BERT 的小型版本》，arXiv预印本 arXiv:2010.05609，2020年。[23] S. Mehta、R. Koncel-Kedziorski、M. Rastegari 和 H. Hajishirzi，《Define：用于神经序列建模的深度因子分解输入词元嵌入》，arXiv预印本 arXiv:1911.12385，2019年。[24] D. Dale。《如何将多语言 T5 模型适配为单一语言模型》。https://towardsdatascience.com/how-to-adapt-a-multilingual-t5-model-for-a-single-language-b9f94f3d9c90（访问于2022年6月4日）。[25] D. Goldhahn、T. Eckart 和 U. Quasthoff，《在莱比锡语料库中构建大型单语词典：从100种语言到200种语言》，载于第八届国际语言资源与评估会议（LREC'12）论文集，2012年，第759-765页。[26] J. Devlin、M.-W. Chang、K. Lee 和 K. Toutanova，《BERT：用于语言理解的深度双向 Transformer 预训练》，arXiv预印本 arXiv:1810.04805，2018年。

[27] N. Houlsby 等人，《面向自然语言处理的参数高效迁移学习》，载于《国际机器学习会议》，2019：机器学习研究论文集，第2790-2799页。[28] A. Paszke 等人，《PyTorch：一种命令式风格的高性能深度学习库》，《神经信息处理系统进展》，第32卷，2019年。[29] V. Rus、Z. Cai 和 A. Graesser，《问题生成：多年评估活动的实例》，载于 QGSTEC 研讨会论文集，2008年。[30] B. Das、M. Majumder、A. A. Sekh 和 S. Phadikar，《面向主观题考试的自动问题生成与答案评估》，《认知系统研究》，第72卷，第14-22页，2022年3月1日，doi：https://doi.org/10.1016/j.cogsys.2021.11.002。[31] M. Blšták 和 V. Rozinajová，《采用机器学习方法、基于句子结构分析的自动问题生成》，《自然语言工程》，第1-31页，2021年，doi：10.1017/S1351324921000139。[32] M. Last 和 G. Danon，《自动问题生成》，《WIREs 数据挖掘与知识发现》，第10卷，2020年7月27日，doi：10.1002/widm.1382。[33] P. Nema、A. K. Mohankumar、M. M. Khapra、B. V. Srinivasan 和 B. Ravindran，《我们再问一次：用于自动问题生成的精炼网络》，arXiv 预印本 arXiv:1909.05355，2019年。[34] M. Fuadi 和 A. D. Wibawa，《使用文本到文本转换器从印度尼西亚语文本自动生成问题》，载于《2022年国际电气与信息技术会议》（IEIT），2022年9月15日至16日，2022年，第84-89页，doi：10.1109/IEIT56384.2022.9967858。[35] J. H. Clark 等人，《TyDi QA：类型学多样化语言中信息检索式问答的基准测试》，《计算语言学协会会刊》，第8卷，第454-470页，2020年。[36] S. Sharma、L. E. Asri、H. Schulz 和 J. Zumer，《无监督指标在面向任务的对话中评估自然语言生成的相关性》，arXiv 预印本 arXiv:1706.09799，2017年。[37] K. Papineni、S. Roukos、T. Ward 和 W.-J. Zhu，《BLEU：一种用于自动评估机器翻译的方法》，载于《计算语言学协会第四十届年会论文集》，2002年，第311-318页。[38] C.-Y. Lin，《ROUGE：用于自动评估摘要的软件包》，载于《文本摘要的拓展》，2004年，第74-81页。[39] Y. Chali、S. A. Hasan 和 S. R. Joty，《利用句法、浅层语义和扩展字符串子序列核改进用于复杂问答的基于图的随机游走》，《信息处理与管理》，第47卷，第6期，第843-855页，2011年。[40] M. A. C. Soares 和 F. S. Parreiras，《问答技术、范式与系统的文献综述》，《沙特国王大学学报：计算机与信息科学》，第32卷，第6期，第635-646页，2020年。[41] P. Rajpurkar、J. Zhang、K. Lopyrev 和 P. Liang，《SQuAD：100,000+道用于文本机器阅读理解的问题》，arXiv 预印本 arXiv:1606.05250，2016年。[42] M. Devika、C. Sunitha 和 A. Ganesh，《情感分析：不同方法的比较研究》，《计算机科学程序》，第87卷，第44-49页，2016年。[43] B. Wilie 等人，《IndoNLU：用于评估印度尼西亚语自然语言理解的基准测试与资源》，arXiv 预印本 arXiv:2009.05387，2020年。