

用于小样本跨语言文本分类的元对比学习框架

郭建铭, 赵或然, 刘功申

(上海交通大学网络空间安全学院, 上海 200240)

摘要: 众多的安全风险问题均为文本分类问题, 国际场景下的舆情分析等风控问题涉及多种语言, 是一大难点。先前的研究表明, 通过跨语言语义知识迁移可以显著提高小样本文本分类任务的性能。然而, 跨语言文本分类的发展仍面临着一系列挑战。获得语义无关的文本表征是一项困难的任务。不同语言之间的语法结构和句法规则引起文本表征的差异, 因此提取通用的语义信息较为困难。此外, 跨语言文本分类的标签数据十分稀缺。在很多现实场景中, 只能获得少量的标记数据, 这严重降低了许多方法的性能。因此需要有效的方式能够在小样本情况下准确地迁移知识, 提高分类模型的泛化能力。为应对这些挑战, 提出了集成对比学习和元学习的框架, 该框架集成了对比学习和元学习的优势, 利用对比学习来提取与语言无关的通用语义信息, 同时利用元学习快速泛化的优势来改善小样本场景中的知识迁移。此外, 提出了基于任务的数据增强方法, 以进一步提高所提框架在小样本跨语言文本分类中的性能。通过在两个广泛使用的多语言文本分类数据集上进行大量实验, 证实了所提方法能够有效提升文本分类的准确性, 可有效应用于风控安全领域。

关键词: 跨语言文本分类; 元学习; 对比学习; 小样本

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2024043

Contrastive meta-learning framework for few-shot cross-lingual text classification

GUO Jianming, ZHAO Yuran, LIU Gongshen

School of Cyber Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Abstract: Many security risk control issues, such as public opinion analysis in international scenarios, have been identified as text classification problems, which are challenging due to the involvement of multiple languages. Previous studies have demonstrated that the performance of few-shot text classification tasks can be enhanced through cross-lingual semantic knowledge transfer. However, the advancement of cross-lingual text classification is faced with several challenges. Firstly, it has been found difficult to obtain language-agnostic representations that perform

收稿日期: 2023-08-09; 修回日期: 2023-12-25

通信作者: 刘功申, lgshen@sjtu.edu.cn

基金项目: 国家自然科学基金项目 (U21B2020); 上海市科技计划项目 (22511104400)

Foundation Items: The National Natural Science Foundation of China (U21B2020), Shanghai Science and Technology Plan (22511104400)

引用格式: 郭建铭, 赵或然, 刘功申. 用于小样本跨语言文本分类的元对比学习框架[J]. 网络与信息安全学报, 2024, 10(3): 107-116.

Citation Format: GUO J M, ZHAO Y R, LIU G S. Contrastive meta-learning framework for few-shot cross-lingual text classification[J]. Chinese Journal of Network and Information Security, 2024, 10(3): 107-116.

well in cross-lingual transfer. Moreover, the differences in grammatical structure and syntactic rules between different languages cause variations in text representation, making it difficult to extract general semantic information. Additionally, the scarcity of labeled data has been identified as a severe constraint on the performance of most existing methods. In many real-world scenarios, only a small amount of labeled data is available, which has been found to severely degrade the performance of many methods. Therefore, effective methods are needed to accurately transfer knowledge in few-shot situations and improve the generalization ability of classification models. To tackle these challenges, a novel framework was proposed that integrates contrastive learning and meta-learning. Within the framework, contrastive learning was utilized to extract general language-agnostic semantic information, while the rapid generalization advantages of meta-learning were leveraged to improve knowledge transfer in few-shot settings. Furthermore, a task-based data augmentation method was proposed to further improve the performance of the framework in few-shot cross-lingual classification. Extensive experiments conducted on two widely used multilingual text classification datasets show that the proposed method outperforms several strong baselines. This indicates that the method can be effectively applied in the field of risk control and security.

Keywords: cross-lingual text classification, meta-learning, contrastive learning, few-shot

0 引言

文本分类在安全风控中有着众多应用,如风险评估、舆情分析、恶意行为识别、禁限售商品分类等。而随着国际化的发展,国际场景下多语言的安全风控也受到了非常多的关注。跨语言文本分类(cross-lingual text classification, CLTC)是自然语言处理领域十分热门的研究方向^[1-3],在安全风控这一领域应用广泛。现有的大多数文本分类方法依赖大量目标语言的标记数据^[4-6],这限制了它们在小样本文本分类场景下的性能。然而语言的分布是长尾的,大量的标注数据资源集中在中文、英语等高资源语种,因此尝试通过跨语言文本分类来解决小语种上的分类问题。由于语言通过不同的表达形式共享相同的语义,跨语言文本分类旨在将其他高资源语言中与任务相关的语义知识转移至目标语言,即使目标语言中只有少量标记实例,也能提高分类准确度。

早期的研究主要集中在减小源语言语义空间和目标语言空间之间的差距^[7-8]。随着多语言预训练语言模型(multilingual pre-trained language models, mPLMs)的发展,预训练后微调已成为文本分类的新范式^[9-10]。引入判别器到 mPLMs 中用于生成语言无关的文本表征被证明是可行的^[11]。为了解决标记实例稀缺的问题, X-MAML 被应用于跨语言领域^[12]。此外,元学习成为跨语言小样本学习的研究重点,一部分对该框架的修改证明是有效的^[2]。

尽管这些方法取得了成功,但它们仍然面临一些挑战。首先,促进语义知识的跨语言迁移需要获取与语言无关的文本表征,这是困难的任务,因为多语言预训练模型往往忽略多语言对齐知识,从而限制了对不同语言的同一类别文本内部关系的探索^[10]。由于受限于通用预训练任务,它们的表示可能包含除语义之外的其他维度的信息,如句法信息、语法信息等^[13-15]。其次,现有方法仍然较依赖标记数据,它们以足够的辅助语言标记语料为前提。然而,在大部分分类任务中获取大量的多语言标记数据是不容易的。

因此,本文提出了元对比学习框架,并从两个方面尝试解决这些问题。首先将对比学习与元学习相结合,以更好地提取与语言无关的语义表征。其次,针对提出的框架设计了跨语言数据增强方法来减少对标记数据的依赖,从而使该框架更具适应性和多样性。

根据之前的研究,对比学习和元学习之间存在着密切的关系^[16-17]。MAML(model-agnostic meta-learning)^[18]类似于在分类任务中使用对比学习目标函数的元学习器,在这种设定下,查询集特征被拉向相同类别的支持集特征,而远离不同类别的支持集特征。因此,通过将对比学习集成到元学习中,期望两种方法能够相辅相成并增强知识迁移的能力,以解耦文本表征中的高维度信息,并提取与任务相关的语义信息,同时利用元学习来实现泛化和知识传递的能力。在元训练阶段通过对比学习获得的任务相关的语义表示,

将其作为元知识传递到元测试阶段。此外，受到对比学习中数据增强的关键性^[19-21]以及元学习任务增强的重要性的启示^[22-23]，本文为跨语言场景设计了基于任务的数据增强方法。在这个方法中，首先构建了通用词典和领域相关的词典，然后选择表示语义的词，并用其同义词替换它们，从而生成具有等效语义的额外文本。这些同义词参考之前构建的词典。通过对比学习和元学习的应用，本文提出的框架在具有有限标记数据的辅助语言中展现出了有效的自监督学习能力。

本文在MLDoc和Amazon Reviews数据集上进行了大量实验，这些数据集在跨语言文本分类领域的评估中被广泛使用。本文方法明显优于几种同类方法。此外，还设计了几个对比实验和可视化实验，以揭示本文方法改进的潜在原因。

1 研究现状

1.1 跨语言文本分类

跨语言文本分类领域受到多语言预训练模型发展的影响。其中多语言BERT (multilingual BERT, mBERT)^[24]使用包含104种语言的语料库进行训练，在零样本和小样本跨语言迁移任务中表现出强大的性能。为了解决mBERT训练过程中不同语言之间信息不兼容的问题，Facebook先后提出了3种跨语言预训练模型。XLM (cross-lingual language model)^[9]将之前的掩码语言模型任务扩展为翻译语言模型任务，该任务以一对平行语料库为输入，使模型能够同时利用两种语言的信息来预测掩码词。在XLM的基础上，XLM-R (XLM-RoBERTa)^[25]增加了模型的深度和宽度，并使用包含100种语言的语料库进行训练。mBART (multilingual BART)^[26]采用两种噪声添加方法，使用由25种语言组成的单语言数据集进行序列到序列的预训练。在CLTC中涉及文本理解，需要使用编码器。因此，本文选择XLM-R作为基础学习器，因为它在各种下游任务上表现出卓越的性能。

多语言预训练模型是利用大量无标数据进行无监督训练得到的，在完成文本分类等任务时需要一部分标记数据进行微调，在小样本甚至零样本场景下效果较差。许多方法已被用来进一步增强mPLMs的知识传递能力，从而提高它们在

CLTC上的性能。其中，在CLTC中，常见的方法是知识蒸馏^[27-29]。这些方法通常涉及构建教师—学生网络，并通过蒸馏传递知识。教师模型可以是高性能的单语言模型或多语言模型。然而，这种知识蒸馏方法依赖于教师模型，导致计算成本增加和性能波动。此外，一些方法利用图神经网络 (graph neural network, GNN) 学习语言之间的关系，并转移语义知识^[30-31]。然而，这类方法往往依赖于特定的图结构数据，并需要大量标记数据，使其难以有效处理新任务。

此外，元学习在文本分类任务中发挥着重要作用。Heijden等^[2]证明了一组元学习方法，如ProtoNet (prototypical networks)^[32]、MAML^[18]、Reptile^[33]和ProtoMAML^[34]在CLTC上表现良好，并提出了名为Foprotomamln的新型方法。因为元学习具有优越的泛化性能，并且可以快速适应新领域而不引入新的结构，所以本文选择遵循元学习方法，并提出改进方法。

1.2 跨语言数据增强

越来越多的研究表明，在文本理解任务中，数据增强起到非常重要的作用^[35-38]。同时，元学习被发现受益于任务增强^[22-23]。考虑跨语言设置中的CLTC问题，代码混合 (code-switching) 成为广泛用于数据增强的方法。ALM (alternating language modeling)^[38]利用来自不同语言的代码混合句子，在预训练过程中捕捉词汇和短语的丰富跨语言上下文，而不是简单地进行连接。Manifold Mixup^[39]利用语义插值作为额外的训练信号，得到具有平滑决策边界的神经网络，适用于多个表示层级。文献[40]提出了数据增强框架，生成多语言代码混合数据以对mBERT进行微调。Krishnan等^[41]通过随机翻译来增强单语言的代码混合，以增强多语言模型在下游任务微调中的泛化能力。Feng等^[42]认为代码混合的标记可能导致新句子的语法不一致。因此，他们不仅进行标记替换，还考虑上下文和切换标记之间的相似性。

关于元学习，文献[22]认为迁移学习引入了过拟合的潜在来源，并提出了信息理论框架来进行元增强，还证明了元增强对元规则化技术的好处。Yao等^[23]提出了两种任务增强方法，即

MetaMix 和 Channel Shuffle, 并证明了它们增强元学习泛化的能力。为了进一步提高小样本情况下的性能, 本文不局限于代码混合, 而是将其与元学习结合起来, 提出了新的基于任务的数据增强方法。

2 方案设计

本节将主要介绍元对比学习框架以及相对应的跨语言数据增强方法。

2.1 跨语言文本分类任务

小样本跨语言文本分类的目标是从源语言和辅助语言中学到的知识传递给在训练期间未见的小样本目标语言。给定一个用于多语言文本分类的数据集 $D = \{D_c\}$, $c \in D$, 其中 D_c 表示属于类别 c 的样本包含 n 种语言, $D_c = \{D_c^l\}$, $l \in L = \{l_1, l_2, \dots, l_n\}$ 。

其中, 源语言是指用于微调模型的源语言, 辅助语言指元训练过程中使用的语种。这两者都使模型学习语义知识并将其传递给目标语言。它们在实验过程中根据不同的使用方法进行区分。目标语言则用于评估模型。

遵循元学习的范式, 本文从任务分布 $p(\tau)$ 中随机抽样任务 (或称为 episode), 该任务与数据集相关, 以构建元学习数据集 $D_{\text{meta}} = \{T_1, T_2, \dots, T_n\}$ 。

每个任务 T_i 包含一个支持集 S_i 和一个查询集 Q_i 。支持集用于在内部优化中进行快速适应, 而查询集用于评估性能并计算相对于初始模型参数的损失。本文采用 n -way、 k -shot 的设置来描述训练批次, 即包含 n 个不同类别的例子, 每个类别有 k 个实例。在本文的架构中, 支持集和查询集都采用这种设置。

与之前的工作保持一致^[2,43], 本文采用前文

提到的 CLTC 工作中的 XLM-R 作为基础学习器, 并用它来表示文本。为了方便后续使用, 本文将 XLM-R 的文本模块表示为 $f(\cdot)$, 并将其所有参数表示为 θ 。

2.2 元对比学习框架

本文提出的元对比学习框架将对比学习与元学习相结合, 以促进从辅助语言向目标语言的无关语义知识的传递。元对比学习框架的整体流程如图 1 所示。元学习是一种学习范式, 专注于使学习器快速适应新任务。本文采用元学习方法是因为它在小样本学习和知识传递方面的有效性。此外, 本框架利用对比学习在元训练阶段解耦任务相关的语义表示。在元测试阶段, 本文对目标语言进行单步微调, 并通过计算与已有类别表征的质心 (prototype) 来确定类别。

2.2.1 元训练

在元训练阶段, 本文框架采用 FoMAML (first-order MAML) 方法, 模型参数 θ 的更新步骤如式(1)所示。

$$\theta = \theta - \beta \frac{1}{|B|} \sum_{(S_i, Q_i) \in B} \left(\nabla_{\theta_i^{(r)}} L_i(f_{\theta_i^{(r)}}) \right) \quad (1)$$

其中, $L_i(f_{\theta_i^{(r)}})$ 表示在支持集 S_i 上更新基础学习器 r 步后在查询集 Q_i 上的损失, B 是根据任务分布 $p(\tau)$ 采样的一批任务样本, β 是学习率。

本文的目标是提取包含丰富语义信息的语言无关的文本表征。关键是解耦高维的语义特征和其他语言相关的信息, 如语法和句法信息等。为了实现这个目标, 本文基于有监督对比学习提出了将同标签文本作为正例对的采样方法。具体地, 在包含 $n \times k$ 个实例的支持集 S 中, 本文的损失函数可以表示为:

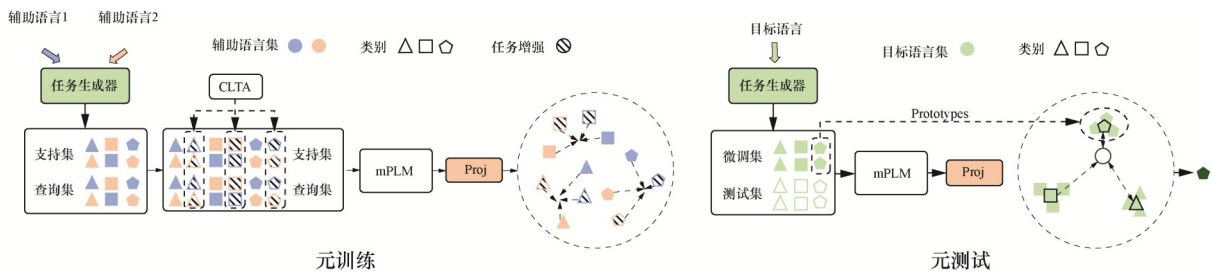


图 1 元对比学习框架的整体流程

Figure 1 The overall workflow of the contrastive meta-learning framework

$$L_{\text{sup}} = \sum_{i \in I} - \log \left(\frac{1}{|P_i|} \sum_{p \in P(i)} \frac{\exp(h_i h_p \tau)}{\sum_{a \in I} \exp(h_i \cdot h_a / \tau)} \right) \quad (2)$$

其中, $P_i = \{p \in I \mid \{i\} : y_p = y_i\}$ 是与 i 标签相同的所有正样本在支持集中的索引。编码后的文本表征为 $h_i = \text{Normalize}(f(x_i)) \in R^{768}$, $\tau \in R^+$ 是一个标量温度系数, $\cdot$ 表示内积, 用作两个文本表示之间的相似度度量。在式(2)中, 本文将在特征空间中具有相似语义的文本拉近, 而不考虑它们的语言。相应地, 本文计算查询集中 $n \times k$ 个文本实例的损失, 如式(2)所示, 这将指导外层循环的梯度更新。为了更好地与对比学习合作, 本框架在 PLM 之后添加了一个投影层, 表示映射到应用对比损失的空间。在这个阶段, 该元对比学习框架将帮助语义与其他信息分离开, 产生更多与任务相关的元知识。

2.2.2 元测试

在元测试阶段, 本文将在元训练过程中获得的元知识转移到目标语言分类任务中, 在小样本情况下进行评估。在目标语言上利用式(1)进行参数更新。由于融入了对比学习, 模型无法直接输出每个类别的概率, 因此无法直接评估分类标签。因此, 本文采用 ProtoNet^[32]中使用的方法来计算每个类别的质心。对于类别 c , 质心定义为:

$$\mu_c = \frac{1}{|S_c|} \sum_{(x,y) \in S_c} f_\theta(x) \quad (3)$$

其中, S_c 是支持集中标签为 c 的实例子集。

本文通过式(4)计算测试集中每个实例的每个类别的概率。

$$p(y = c|x) = \frac{\exp(f_\theta(x)\mu_c)}{\sum_{c' \in C} \exp(f_\theta(x)\mu_{c'})} \quad (4)$$

其中, c 表示所有类别。通过这种方式, 可以像往常一样, 通过将标签分配给具有最高概率的类别来预测 x 的类别。

2.3 跨语言任务增强

以往的研究表明, 数据增强可以提高分类性能并减少对标记数据的依赖。此外, 任务增强在元学习中被证明是重要的。考虑元学习将学习单元从样本转换为任务, 而 CLTC 任务涉及跨语言情景, 本文提出了跨语言任务增强 (cross-

lingual task augmentation, CLTA) 方法。

任务增强方法的基础是观察到给定句子的语义在不同语言中保持一致。因此, 用同义词替换句子中的特定单词不会改变其基本含义。这类方法相比反向翻译类的方法可以将一句话用多语言的形式展现, 进而可以在同一个语义空间中学习不同语言的上下文依赖关系。然而现有方法, 如代码混合 (code-switch), 通常使用随机单词采样并采用共同的词汇表。随机采样引入了增强过程的不稳定性, 可能会错过更具关键语义的单词。同时, 使用共同的词汇表可能产生多义性问题, 可能会替换掉与领域无关的单词。

为了解决上述问题, 本文提出了利用词频—逆文档频率 (term-frequency-inverse document frequency, TF-IDF) 来辅助单词选择和创建任务特定词典。该方法涉及构建一个外部词典集, 其中包含通用和特定领域的单词翻译。对于通用词典, 本文利用 Facebook MUSE 等开源词典。这些词典为不同领域中常用的单词提供了宝贵的参考。为了捕捉特定领域的单词, 本文根据其在给定领域中的频率和稀有性来评估单词的重要性, 并生成特定领域的多语言词典。该词典专门针对目标领域生成, 允许更准确和相关的单词替换。

考虑 MLDoc 语料库中 TF-IDF 的分布, 本文选择具有最高 25% TF-IDF 分数的单词来构建与领域相关的词典。在确定了具有语义代表性的单词后, 继续使用翻译 API 构建相应的特定领域词典, 其被设计成允许映射涵盖辅助语言和目标语言的多语言词典。不同语言中的多个可选单词有助于增加任务增强的丰富性, 并让模型学习更多信息。本文认为领域相关词典中的单词具有内在的语义代表性。因此, 不再采用随机抽样, 而是有意地从这些词典中选择单词。

3 实验

3.1 数据集

本文在两个多语言数据集 MLDoc^[44]和 Amazon Review^[45]上评估了本文提出的方法。

MLDoc 提供了路透社语料库 (reuters corpus) 的一个子集, 其中包含 8 种类别分布较为平衡的语言, 包括德语、英语、西班牙语、法语、意大利语、日语、俄语和中文。每篇新闻手

动标注 4 个类别（公司、经济、政府和市场）之一。每种语言的训练集有 10 000 个实例，验证集有 1 000 个实例，测试集有 4 000 个实例。本文将英语设定为源语言，并将西班牙语设定为验证语言。其余的语言分为两组，分别是德语、意大利语、日语、法语、俄语、中文。他们分别用作辅助语言集和目标语言。模型在源语言上进行微调。在辅助语言集上进行元训练后，在验证语言的验证集上表现最佳的模型将被选择应用于元测试阶段。除非另有说明，本文的分析实验都是在 MLDoc 数据集上进行的，其中德语被设定为目标语言。

Amazon Review 数据集包含德语、英语、法语和日语的产品评论，涵盖书籍、DVD 和音乐 3 个领域，每个样本包含 1 条评论以及从 1 到 5 的分数，表示情感分数。遵循之前研究所采用的设置^[2]，本文将得分大于 3 的实例归类为正标签，得分小于 3 的归类为负标签。经过预处理后，该数据集包含每种语言的训练集和测试集各 6 000 个样本。受限于仅有 4 种语言，本文将英语设定为源语言，其余的语言依次设置为辅助语言、验证语言和目标语言。

需要强调的是，目标语言的数据在实验中始终保持小样本，每种语言随机抽取 64 个样本。

3.2 实现细节

本文采用 HuggingFace 提供的预训练模型 XLM-RoBERTa 的基础版本作为基础学习器，并使用 PyTorch 库创建所有模型。所有实验都在单张 NVIDIA RTX 3090 GPU 上进行。

对于元学习方法，本文使用辅助语言训练 20 轮。在每轮训练结束时，对模型在验证语言上进

行小样本（64 个样本）微调，然后在每轮训练结束后，使用 5 个不同的种子在验证集上进行评估。一轮训练包含 200 更新步。每个支持集和查询集都包含 16 个样本，内循环中的更新步数设置为 5。微调过程中采用 Adam 优化器，学习率设置为 3×10^{-5} 。为了避免过拟合，本文采用早停技术，当模型在验证集上连续 5 轮没有性能提升时停止训练。对于除了元学习方法以外的基线模型，本文在相应的数据集上进行 20 轮的微调训练，并在每轮训练结束后进行验证。

3.3 主要结果

两个数据集上每个语种的分类准确率如表 1 所示。本文框架在两个数据集上均取得了最高的平均准确率。

此外，通过任务增强，可以观察到在 MLDoc 数据集上平均准确率增加了 0.5%，在 Amazon Review 数据集上增加了 0.2%。这些改进凸显了所提出的基于任务的数据增强对提高框架性能的重要性。与 FoMAML 比较得知，监督对比学习有效地提取了与任务相关的语义表示，并增强了模型的知识传递能力。此外，非指定任务的性能证明了跨语言知识传递的可行性。本文方法在日语上未能取得最佳性能，这可能归因于辅助语言与目标语言之间的实质性差异。语种簇之间的差异影响了语言无关信息的提取。此外，考虑不同语言之间类别分布的不平衡，领域漂移和异构任务分布可能对结果产生影响^[2]。另外，配备了元学习的方法获得了更好的准确率，表明它们能够利用目标语言中宝贵的跨语言信息，并在小样本情况下快速进行泛化。

表 1 两个数据集上每个语种的分类准确率
Table 1 Classification accuracy for each language in the two datasets

方法	MLDoc							Amazon Review			
	德语	法语	意大利语	日语	俄语	中文	平均值	德语	法语	日语	平均值
Non-episodic	95.1%	91.2%	81.6%	81.3%	71.7%	89.2%	85.0%	90.1%	89.6%	86.3%	88.7%
ProtoNet ^[32]	95.3%	91.5%	81.9%	82.2%	76.0%	88.0%	85.8%	90.5%	91.6%	86.9%	89.7%
Reptile ^[33]	94.8%	91.0%	80.2%	82.0%	72.7%	89.9%	85.1%	90.8%	91.3%	87.8%	90.0%
FoMAML ^[33]	95.5%	91.6%	81.5%	80.2%	77.3%	91.3%	86.2%	91.3%	91.5%	87.7%	90.2%
foProtoMAMLn ^[2]	95.8%	92.5%	82.7%	79.1%	79.4%	91.6%	86.9%	92.0%	91.9%	88.6%	90.8%
元对比学习框架	96.0%	93.1%	83.4%	78.9%	81.0%	91.7%	87.4%	92.8%	92.1%	88.1%	91.0%
w/ CLTA	96.2%	93.1%	84.0%	80.9%	81.8%	91.2%	87.9%	92.9%	92.1%	88.7%	91.2%

4 分析与讨论

4.1 框架优势

不同输入参数和训练方法的分类准确率如表2所示。

表2 不同输入参数和训练方法的分类准确率
Table 2 Classification accuracy with different input parameters and training methods

输入参数	训练方法	准确率
在源语言上微调	FoMAML	95.50%
在源语言+辅助语言上微调	FoMAML	94.73%
在源语言上对比学习	FoMAML	94.57%
在源语言上微调	元对比学习	95.95%

其中微调是指利用交叉熵损失在目标语言标注数据上进行小样本训练，对比学习是指利用 infoNCE 损失在目标语言标注数据上进行训练。在 FoMAML 之前进行微调和对比学习模型性能会分别降低 0.77% 和 0.93%。这表明仅通过标签进行聚类时会覆盖原始表示中包含的有价值的信息，尤其这不是细粒度的分类任务。两阶段训练方法会带来一些问题，如在辅助语言中过拟合以及对先前学习的知识的灾难性遗忘。然而，本文框架有效地解决了这些问题，并实现了 0.45% 的准确率提升。比较结果表明，这种提升不仅仅是由于引入了对比学习，而是所提出的框架成功地将任务相关的信息与其他噪声因素分离，并探索了不同语言中句子之间的内部关系。

4.2 表征优势

本文引入了对比学习来提取与任务相关的语义表示，并将这些表示作为元知识从元训练阶段传递到元测试阶段。为了展示语义表示的优越

性，本文将不同模型生成的语义空间的维度降低后进行可视化。本文选择原始预训练的 XLM-RoBERTa、FoMAML 和本文提出的框架进行对比。对于每个类别，本文从 MLDoc 测试集中随机采样 200 个例子，包括辅助语言和目标语言。然后，本文对模型生成的表示 $H_{\text{model}}^{d \times N} = [h_1, h_2, \dots, h_N]$ 应用主成分分析（principal component analysis, PCA），将维度从 d_{model} 降低到 50。

随后，本文使用 t 分布随机邻域嵌入算法（t-SNE）在二维平面上绘制表示，可视化结果如图2所示。在原始 XLM-RoBERTa 的空间中，实例的文本表征按照语言聚集，而类别在内部语言区域内分隔。然而，在 FoMAML 模型生成的空间中，情况发生了变化。在元学习方法的帮助下，可以观察到不同类别的明显聚集为簇，尽管它们的语言不同。在本文框架中，与 FoMAML 相比，不同类别之间的边界更为明显，表明了本文方法的有效性。在结合对比学习方法后，本文提出的方法能够更好地识别属于同一类别的实例之间的相似性，而不考虑它们来自哪种语言。

此外，本文在图3中量化聚类性能，以展示本文框架在提取表示方面的优势。实验结果表明，与 FoMAML 相比，本文框架在轮廓系数上实现了更高且更快的增长。这表明该框架能够有效地学习同一类别实例之间的内在联系，从而提取出更优秀的与任务相关的语义表示。

4.3 任务增强优势

为了证明所提出的跨语言任务增强方法的有效性，本文在 3 种语言上评估了该方法，并将其性能与其他两种数据增强方法进行了比较。在 code-switch 方法中^[40]，基于通用词典进行随机词

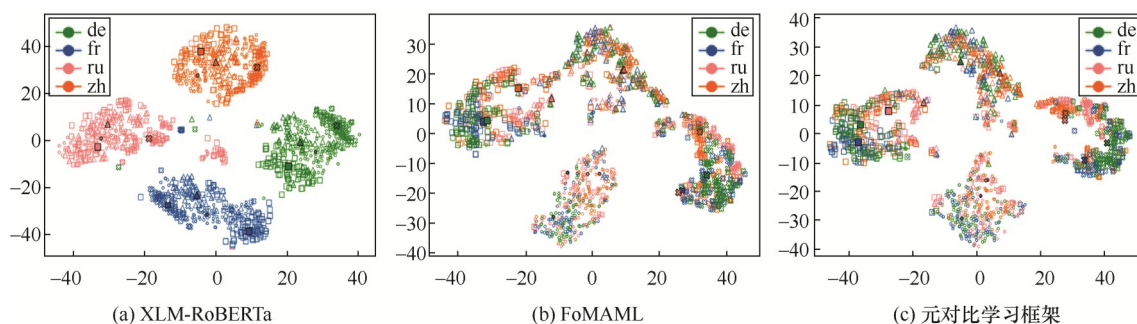


图2 使用不同方法训练模型生成的表征空间的可视化结果

Figure 2 The visualization result of spaces generated by models trained with different methods

汇替换。为了进一步扩展词典, 本文通过反向翻译生成了大量伪标签, 该方法简称为pseudo。如表3所示, pseudo略优于code-switch, 表明更丰富的词典可以改善数据增强的性能。而CLTA在所有3个任务中均取得了最高的准确率, 这证明了选择更具语义表示的词汇进行替换的有效性。数据增强示例如表4所示。

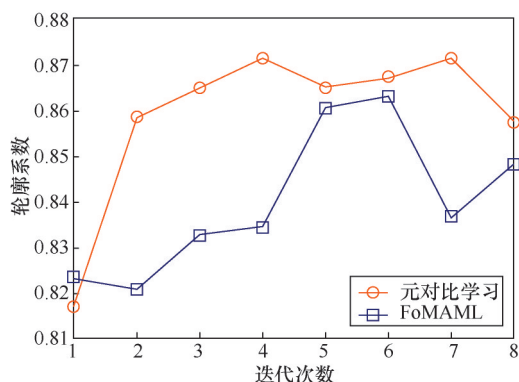


图3 两种方法生成的文本表征的轮廓系数

Figure 3 Silhouette coefficient of representations generated by two methods

表3 不同数据增强方法在MLDoc上的分类准确率

Table 3 Classification accuracy of different data augmentation in MLDoc

数据增强方法	法语	日语	俄语	平均值
code-switch	78.83%	75.72%	75.02%	76.25%
pseudo	80.37%	76.97%	76.03%	77.79%
CLTA	81.88%	77.17%	77.25%	78.76%

表4 数据增强示例

Table 4 Data augmentation examples

源句	Origin & Les cessions de titres de participations des sociétés resteront taxées au taux réduit de 19%, précise lundi le ministère de l'économie.
增强结果	Remplacement & Les cessions de 股票 de participations des 公司 permanecerán taxées au steuersätze réduit de 19%, esprimere mondays le ministry de l'économie

4.4 消融实验

表5显示的消融实验结果表明, 所有组件都对性能的提升有贡献。具体来说, 没有元学习方法的帮助, 准确率下降了2.7%。这凸显了元学习在小样本情况下的关键作用。此外, 将对比学习方法整合到框架中, 准确率提升了1.0%, 这

证明了本文框架的有效性。

表5 消融实验结果

Table 5 The ablation results

方法	德语	意大利	中文	均值
元对比学习+CLTA	96.2%	84.0%	91.2%	90.47%
元对比学习	96.0%	83.4%	91.7%	90.37%
对比学习	94.6%	79.5%	89.2%	87.77%
元学习	95.5%	81.5%	91.3%	89.43%

5 结束语

本文提出了用于小样本跨语言文本分类的训练框架, 该框架将对比学习与FoMAML相结合, 这有助于探索不同语言之间的句子内在关系。此外, 本文还提出了高效的跨语言任务增强方法, 进一步提高了方法的性能。在MLDoc和Amazon Review数据集上进行的大量实验验证了所提方法的有效性, 本文方法可有效应用于安全风险领域。

参考文献:

- [1] PIKULIAK M, ŠIMKO M, BIELIKOVÁ M. Cross-lingual learning for text processing: a survey[J]. Expert Systems with Applications, 2021, 165: 113765.
- [2] HEIJDEN N, YANNAKOUDAKIS H, MISHRA P, et al. Multilingual and cross-lingual document classification: a meta-learning approach[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. 2021: 1966-1976.
- [3] XU Y M, CAO H, DU W Z, et al. A survey of cross-lingual sentiment analysis: methodologies, models and evaluations[J]. Data Science and Engineering, 2022, 7(3): 279-299.
- [4] MINAE S, KALCHBRENNER N, CAMBRIA E, et al. Deep learning: based text classification[J]. ACM Computing Surveys, 2022, 54(3): 1-40.
- [5] TANEJA K, VASHISHTHA J. Comparison of transfer learning and traditional machine learning approach for text classification[C]//Proceedings of 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom). 2022: 195-200.
- [6] LIN Y X, MENG Y X, SUN X F, et al. BertGCN: transductive text classification by combining GNN and BERT[C]//Proceedings of Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021: 1456-1462.
- [7] CHEN X L, SUN Y, ATHIWARATKUN B, et al. Adversarial deep averaging networks for cross-lingual sentiment classification[J]. Transactions of the Association for Computational Linguistics, 2018, 6: 557-570.

- [8] XU R C, YANG Y M. Cross-lingual distillation for text classification[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017: 1415-1425.
- [9] ALEXIS C, GUILLAUME L. Cross-lingual language model pre-training[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019: 7059-7069.
- [10] JIANG X Z, LIANG Y B, CHEN W Z, et al. XLM-K: improving cross-lingual language model pre-training with multilingual knowledge[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(10): 10840-10848.
- [11] HUANG L Z, MA D H, LI S J, et al. Text level graph neural network for text classification[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 3444-3450.
- [12] NOORALAHZADEH F, BEKOULIS G, BJERVA J, et al. Zero-shot cross-lingual transfer with meta learning[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 4547-4562.
- [13] CONNEAU A, KRUSZEWSKI G, LAMPLE G, et al. What you can cram into a single vector: probing sentence embeddings for linguistic properties[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018: 2126-2136.
- [14] JAWAHAR G, SAGOT B, SEDDAH D. What does BERT learn about the structure of language[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 3651-3657.
- [15] KÖHN A. What's in an embedding? Analyzing word embeddings through multilingual evaluation[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 2067-2073.
- [16] CHIA H K, CHUI W C, CHEN P Y. MAML is a noisy contrastive learner in classification[C]//Proceedings of the Tenth International Conference on Learning Representations. 2022: 1-8.
- [17] NI R, SHU M, SOURIS H, et al. The close relationship between contrastive learning and meta-learning[C]//Proceedings of International Conference on Learning Representations. 2022: 9-16.
- [18] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//Proceedings of the 34th International Conference on Machine Learning. 2017: 1126-1135.
- [19] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[C]//Proceedings of the 37th International Conference on Machine Learning. 2020: 1597-1607.
- [20] TIAN Y L, SUN C, POOLE B, et al. What makes for good views for contrastive learning[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020: 6827-6839.
- [21] YOU Y N, CHEN T L, SUI Y D, et al. Graph contrastive learning with augmentations[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020: 5812-5823.
- [22] JANARTHANAN R, ALEX I, ERIC J. Meta-learning requires meta-augmentation[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020: 5705-5715.
- [23] YAO H, HUANG L, ZHANG L, et al. Improving generalization in meta-learning via task augmentation[C]//International Conference on Machine Learning, ICML. 2020: 1-8.
- [24] JACOB D, MING-WEI C, KENTON L, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. 2019: 4171-4186.
- [25] CONNEAU A, KHANDELWAL K, GOYAL N, et al. Unsupervised cross-lingual representation learning at scale[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 8440-8451.
- [26] LIU Y H, GU J T, GOYAL N, et al. Multilingual denoising pre-training for neural machine translation[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 726-742.
- [27] ROUDITCHENKO A, CHUANG Y S, SHVETSOVA N, et al. C2KD: cross-lingual cross-modal knowledge distillation for multilingual text-video retrieval[C]//Proceedings of ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2023: 1-5.
- [28] WANG Z Y, LIU X, YANG P J, et al. Cross-lingual text classification with heterogeneous graph neural network[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021: 612-620.
- [29] ZHANG J Y, LIANG T, WAN M Y, et al. Curriculum knowledge distillation for emoji-supervised cross-lingual sentiment analysis[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022: 864-875.
- [30] GHOSH S, MAJI S, DESARKAR M S. GNoM: graph neural network enhanced language models for disaster related multilingual text classification[C]//Proceedings of the 14th ACM Web Science Conference 2022. 2022: 55-65.
- [31] VAN DER HEIJDEN N, SHUTOVA E, YANNAKOUDAKIS H. K-hop neighborhood regularization for few-shot learning on graphs: a case study of text classification[C]//Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. 2023: 1187-1200.
- [32] JAKE S, KEVIN S, RICHARD Z. Prototypical networks for few-shot learning[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS, 2017: 4080 - 4090.
- [33] ALEX N. On first-order meta-learning algorithms[EB]. arXiv Preprint. arXiv: 1803.02999 (2018).
- [34] ELENIT, ZHU T, VINCENT D, et al. Meta-Dataset: a dataset of datasets for learning to learn from few examples[C]//Proceedings of International Conference on Learning Representations. 2020: 1-8.
- [35] HEIKE A, NGOC T, TANJA S. Combination of recurrent neural networks and factored language models for code-switching lan-

- guage modeling[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. 2013: 206-211.
- [36] CHEN J A, TAM D, RAFFEL C, et al. An empirical survey of data augmentation for limited data learning in NLP[J]. Transactions of the Association for Computational Linguistics, 2023, 11: 191-211.
- [37] LI Y, FUNG P. Improved mixed language speech recognition using asymmetric acoustic model and language model with code-switch inversion constraints[C]//Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. 2013: 7368-7372.
- [38] YANG J, MA S M, ZHANG D D, et al. Alternating language modeling for cross-lingual pre-training[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 9386-9393.
- [39] VERMA V, LAMB A, BECKHAM C, et al. Manifold mixup: better representations by interpolating hidden states[C]//Proceedings of the 36th International Conference on Machine Learning, 2019: 6438-6447.
- [40] QIN L B, NI M H, ZHANG Y, et al. CoSDA-ML: multi-lingual code-switching data augmentation for zero-shot cross-lingual NLP [C]//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization. 2020: 3853-3860.
- [41] KRISHNAN J, ANASTASOPOULOS A, PUROHIT H, et al. Multilingual code-switching for zero-shot cross-lingual intent prediction and slot filling[C]//Proceedings of the 1st Workshop on Multilingual Representation Learning. 2021: 211-223.
- [42] FENG Y K, LI F, KOEHN P. Toward the limitation of code-switching in cross-lingual transfer[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022: 5966-5971.
- [43] HARDALOV M, ARORA A, NAKOV P, et al. Few-shot cross-lingual stance detection with sentiment-based pre-training[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(10): 10729-10737.
- [44] HOLGER S, LI X. A corpus for multilingual document classification in eight languages[C]//Proceedings of the Eleventh International Conference on Language Resources and Evaluation. 2018.
- [45] PRETTENHOFER P, STEIN B. Cross-language text classification using structural correspondence learning[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. 2010: 1118-1127.

[作者简介]



郭建铭 (1999-), 男, 江苏泰州人, 上海交通大学硕士生, 主要研究方向为自然语言处理。



赵彧然 (1998-), 男, 河南安阳人, 上海交通大学硕士生, 主要研究方向为自然语言处理。



刘功申 (1974-), 男, 山东聊城人, 上海交通大学教授, 博士生导师。主要研究方向为人工智能安全、自然语言处理、信息安全。