计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(8):269−277 [doi: 10.15888/j.cnki.csa.009174]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

自然语言场景下增量知识构造与遮蔽回放策略①

周　航, 黄震华

(华南师范大学 计算机学院, 广州 510631)
通信作者: 周　航, E-mail: 2020023062@m.scnu.edu.cn

摘　要: 在增量学习中, 随着增量任务的数量增多, 模型在新增任务上训练后, 由于数据分步偏移等一系列问题, 模型对旧任务上所学到的知识发生灾难性遗忘, 致使模型在旧任务上性能下降. 对此, 本文提出了基于知识解耦的类增量学习方法, 分层次的学习不同任务共有知识与特有知识, 并对这两种知识进行动态的结合, 应用于下游的分类任务中. 并在回放学习中运用自然语言模型的遮蔽策略, 促进模型快速回忆起先前任务的知识. 在自然语言处理数据集 AGNews、Yelp、Amazon、DBPedia 和 Yahoo 的类增量实验中, 本文所提出的方法能有效降低模型的遗忘, 提高在各个任务上的准确率等一系列指标.

关键词: 增量学习; 特征学习; 自然语言处理

引用格式: 周航,黄震华.自然语言场景下增量知识构造与遮蔽回放策略.计算机系统应用,2023,32(8):269–277. http://www.c-s-a.org.cn/1003-3254/9174.html

Incremental Knowledge Construction and Mask Replay Strategy in NLP Scenario

ZHOU Hang, HUANG Zhen-Hua

(School of Computer Science, South China Normal University, Guangzhou 510631, China)

Abstract: In increment learning, as the number of tasks increases, the knowledge learned by the model on the old task is catastrophically forgotten after the model is trained on the new task due to a series of problems such as step-by-step data migration, resulting in the degradation of the model performance on the old task. Given this problem, a class-incremental learning method based on knowledge decoupling is proposed in this study. This method can learn the common and unique knowledge of different tasks hierarchically, combine the two kinds of knowledge dynamically, and apply them to the downstream classification tasks. Besides, the mask strategy of the natural language model is used in replay learning, which prompts the model to quickly recall the knowledge of the previous tasks. In class-incremental experiments on NLP datasets—AGNews, Yelp, Amazon, DBPedia and Yahoo, the proposed method can effectively reduce the forgetting of the model and improve the accuracy and other indicators on various tasks.

Key words: increment learning; representation learning; natural language processing (NLP)

随着数据量的骤增, 时刻都涌现着新的任务数据, 而在旧的任务数据上训练所得到的深度学习模型, 其性能随着不断新增的数据持续下降, 急需在新的数据上重新学习; 与此同时, 由于数据的隐私或存储问题, 以往的数据可能无法再次获得, 由此诞生了增量学习场景. 增量学习最先兴起于图像分类场景, 是对人工神经网络模拟人类在不同任务间对已学习的任务“记忆”“推导”能力的一种模仿. 关于增量学习动机, 即避免模型的灾难性遗忘问题的研究最早在文献 [1] 提出, 指出模型在新的任务数据上训练学习, 会影响先前任务的

① 基金项目: 国家自然科学基金 (62172166)
收稿时间: 2023-01-12; 修改时间: 2023-02-09; 采用时间: 2023-02-23; csa 在线出版时间: 2023-06-09
CNKI 网络首发时间: 2023-06-09

性能. 基于此, 学者提出了灾难性遗忘这个概念. 尔后, Li 等人[2] 提出了用蒸馏学习的方法[3] 减缓灾难性遗忘问题. 随着增量学习问题研究的深入, 深度学习模型灾难性遗忘的原因被归因于几方面: 首先, 是学习新任务时的权重漂移现象. 这是在模型规模固定的情况下, 由于更新参数所用的数据发生了改变, 使得相关的网络权重以及相关激活变化, 进而导致网络输出发生较大的改变, 从而影响先前任务的性能; 其次, 是任务间混淆[4] 问题. 这种问题是由于在类增量学习的场景中, 目标是将类与所有任务区分开来. 本文将现有的解决方式分为两大类别: 第 1 类方法为静态的模型结构; 第 2 类为模型结构动态分配的方法.

在静态结构增量场景中, 模型的神经元资源静态固定, 模型的大小或资源是在初次训练前就完成分配的. 这一类方法通过各种手段, 使得模型表征重叠部分相对固定或较小变化. 由于模型参数是以一定程度更新, 而非完全更新或者不加限制的更新, 所以模型理论上是能够既具有先前任务所学习到的知识, 又使得模型能够在新的任务上拟合的. 主要是通过一些正则化方法, 使得模型的参数受其约束, 理论上能够让模型在学习心得数据时巩固先前的知识, 这一系列方法大致可以进一步分为基于数据的正则化方法以及基于先验知识的正则化方法. Li 等人[2] 最先将蒸馏学习方法用以解决增量学习问题. Huang 等人[5] 通过模式迁移学习的方法, 将表征解耦方法运用在任务增量方法上, 解耦后的类别相关信息输入到任务判别器作任务分类, 以辅佐最终重构的表征作增量分类任务. Ke 等人[6] 针对增量任务中不同任务间的相似性程度不同, 提出了针对任务相似性的算法. Zhou 等人[7] 利用了这种新、旧中类别语义关系之间的关联性, 提出了协同运输的增量学习方法. Lee 等人[8] 提出了一组注意力独立机制, 彼此独立的注意力模块相互竞争的学习用于解耦提取到的高维特征, 从而学习任务之间通用的独立机制, 从而避免模型的遗忘问题.

结构扩展的增量学习方法, 通过在新的任务上动态分配并适应新分配的神经元资源, 从而改变新数据在神经元上的表征分步, 使得相较于前任务时的网络有了不同的结构属性, 常见的方法为直接增加神经元个数或者直接以并联或串联增加网络层的方式进行重新训练. Shen 等人[9] 设计了一个面向自然语言处理槽填充的基于扩张结构的增量学习方法. Monaikul 等人[10] 提出了 AddNER 框架以及 ExtendNER 框架, 以解决增量场景在命名实体识别中, 新任务的数据无需重新对旧任务标签进行标记的问题. Yan 等人[11] 设计了一个两阶段动态扩张表征的增量模型框架. Singh 等人[12] 利用一部分的网络参数学习修正参数信息, 使得主网络能够利用这些信息来修正网络参数, 最终能够在新任务数据上拟合.

除了上述常规意义上的对网络模型结构的某些部分 (一般是表征部分) 进行扩张, 本文将基于记忆网络的回放方法[2] 也归纳入模型扩张部分. Castro 等人[13] 对基于回放的增量学习方法进行细化, 完善了旧任务训练案例构建、更新的策略, 提出了一整完整的端到端的基于回放的增量学习框架. 针对分类器的偏置项参数倾向于新任务的问题, Wu 等人[14] 提出利用少量新、旧样本对分类器再次进行纠正训练, 避免因为回放数据过少的类不平衡问题. Lopez-Paz 等人[15] 提出梯度片段记忆算法, 通过约束先前任务损失不增加, 避免模型对先前任务的遗忘. de Masson 等人[16] 提出了稀疏性经验回放策略, 并创新性的将回放数据局部适应步骤增加在模型推断阶段.

但是, 上述方法存在以下几点问题. 首先, 针对任务增量的增量学习场景在测试时经常会由于缺少了任务标签导致性能下降, 且并不具有普遍性, 而没有任务标签的类增量场景更为普遍; 其次, 当模型在第 t 个任务时, 相较于第前 t–1 个任务, 当前任务的数据是足以让模型在当前任务中充分得到训练的, 这一点从深度学习本身以及我们做的消融实验本身都是足以说明的. 而前 t–1 个任务的样本的回放训练才是影响模型性能的关键, 有效利用存储的回放样本使模型快速恢复在先前任务上的性能是回放任务的重点.

对此, 本文提出了基于表征解耦的增量知识构造与遮蔽回放策略方法 (incremental knowledge construction and mask replay, IKCMR). 相较于利用类标签的解耦方法, 本文认为通过类标签解耦的知识并不能很好的指导最终分类, 因为模型所学到的知识其实是针对任务的而不是针对类别的, 而本文的最终目的是利用模型所学习到的知识去进行分类. 受启发于文献 [4] 所述的互补学习系统理论, 模型在有效提取感知时间的统计结构、泛化知识的同时, 还保留了特定任务的记忆或经验. 根据这一理念, 模型解耦后的公共特征与类别特征在学习和记忆方面的互补作用, 公共特征解

耦器通过新的任务数据, 补充学习不同数据中语言、语义信息的共有特征, 类别特征解耦器专注于学习类别相关知识. 而后, 通过特征相似性方法, 本文将学习到的泛化知识与特定知识动态的结合, 以此作为最终分类器的输入. 本文的方法动态地协调了模型可塑性与稳定性平衡的问题. 同时, 为了进一步避免模型遗忘问题, 在回放训练中本文利用缓冲区存储的数据进行学习, 使得模型的类别相关的解耦器能够通过这些少量的回放数据快速回忆起先前任务的知识.

归纳起来, 本文的主要贡献如下.

(1) 针对自然语言处理中更普遍的类别增量问题, 本文基于此提出了类增量特征解耦方法, 在避免了任务标签的同时, 以类的粒度对特征进行解耦, 并依照互补学习理论动态地进行特征重构.

(2) 本文探索了在样例回放学习过程中, 普通的回放方法对比遮蔽回放学习方法对模型快速回忆先前任务知识的效果. 本文的解耦模型配合遮蔽回放学习方法, 能够在保持当前任务良好的情况下, 减少先前任务上知识遗忘造成的性能上的损失.

(3) 本文通过自然语言处理的分类任务上的 5 个公认的数据集验证了 IKCMR 模型的有效性. 实验结果说明, IKCMR 模型比现有的 SOTA 模型在类增量问题上的性能要更加出色. 除此之外, 也通过实验验证了融合了遮蔽回放的 IKCMR 能进一步提高模型的表现.

本文第 1 节将对本文的工作进行具体的描述. 第 2 节会呈现本文的实验结果以及结论部分. 最后, 第 3 节是本文的总结以及未来工作部分.

1 自然语言场景下的类增量问题方法

首先, 本文对增量学习进行定义. 增量学习的任务通常来说是将模型在一系列非独立同分布的任务上进行训练. 将这一系列的非独立同分布的任务定义为 $\mathcal{T}_{n_t}=\{T_1,\cdots,T_{n_t}\}$, 其中, 第 n 个 task 的 $\mathcal{T}_t\in\left\{\left(x_i^t,y_i^t\right)\right\}_{i=1}^{n_t}$ 包含了输入语言序列 $x_i^t\in X$ 以及其对应的分类标签 $y_i^t\in\mathcal{Y}$. 而增量学习的目标为训练一个模型 θ, 使其能对给定的任意已训练的任务的测试语言序列 $x_j\in\mathcal{T}_t$, 都能预测其对应的标签 $\hat{y}_j=f_{\theta(x_j)}\in\mathcal{Y}$. 模型在训练第 t 个任务的时候, 先前的任务 $\mathcal{T}_{t-1}$ 都不可见, 只有 T_t 和少量存储在 memory buffer 中的样例 $\mathcal{M}_{t-1}$ 可见.

本文的总体框架如图 1 所示, 模型的主体由 3 个部分构成. 首先是低层特征提取部分, 该部分将输入的自然语言序列转化为特征向量; 模型的第 2 个部分是高层特征的解耦以及重构部分, 最后一部分则为模型的分类器部分. 近年, 各种大规模预训练模型在各项自然语言处理任务中效果突出, 例如 BERT[17] 和 GPT-2[18] 等大规模的编码器-解码器[19] 结构的模型. 为了更加高效的搭建模型以及资源限制问题, 本文将预训练语言模型 BERT 作为表征嵌入部分. 在此基础上, 模型的第 2 部分从高层特征表示出发. 首先, 将句子的特征输入公共编码器和类别编码器中, 得到编码后的关于当前任务的类别特征以及公共特征, 例如更好的文本表征、句子间的关系等; 接着, 通过对公共表征与类别表征的表征相似度分析, 并以此为依据将两种特征进行融合, 得到模型结构的相似性, 这种相似性是区别于模型表征的相似性的, 模型表征的相似度其实是较为底层的, 生成其最终的语义特征.

1.1 公共特征网络

公共特征解耦器如何学习到自然语言的共性, 即类别无关的特征是本文要解决的一大难点, 而大规模语言模型的训练给了本文以启发. 大规模预训练模型通过例如下一句预测、遮蔽语言模型等任务, 使模型通过大量不同的语料学习到其共性特征. 类似的, 从理论上来说, 公共特征网络接触过的任务越多, 其能学习到的自然语言特征的共性也就越多. 模型在学习任务 $T_i\in\mathcal{T}_{n_t}=\{T_0,\cdots,T_{n_t}\}$时, 令公共特征网络采用遮蔽语言的任务进行语言特征共性的学习. 具体的, 本文与 Devlin 等人[17] 的实验设置相同, 对样本 $x_i=\{\text{token}_1,\cdots,\text{token}_{\text{mask}},\cdots,\text{token}_n\}\in T_i$, n 为句子长度, 将其中所有 token 以 15% 的概率进行遮蔽处理. 其中, 遮蔽处理的 token 之中, 用 [MASK] 替代的占 80%, 随机用其他 token 替代的占 10%, 还有 10% 的不做改变. 本文将公共特征处理器定义为 $Net_G(\cdot):R^{n\times 768}\to R^{n\times D}$, 其中 D 是特征的维度. 遮蔽后的样本 $\widehat{x_i}$, 经过预训练语言模型处理后, 输入到公共特征网络, 再通过一个激活层, 得到公共特征:

$$\begin{cases}E_G=\left\{e_G^1,\cdots,e_G^n\right\},\ e_G^j\in R^D\\ e_G^{ij}=\tanh\left(Net_G\left(\text{token}_j\right)\right)\end{cases}\tag{1}$$

公共特征网络需要学习和理解样本 x 的内容, 然后通过上下文含义对遮蔽部位进行预测. 通过对遮蔽词向量的预测, 可以监督公共特征网络对不同领域自然语言知识的学习情况, 判断公共特征解耦网络对自然

语言共性特征空间的构建情况. 具体的, 公共特征解耦网络预测样本 $\hat{x}$ 的遮蔽表征 e_G^{mask}, 通过一个辅助解码网络 $Net_D(\cdot)$ 映射到词典中, 检验公共特征网络是否理解了输入文本的含义, 从而使公共特征网络学习到的基于任务 t_i 的语言本身的知识.

$$\widehat{\text{token}}_{\text{mask}} = Net_D\left(e_G^{i,\text{mask}}\right) \tag{2}$$

其中, 对于公共特征网络的训练, 本文采用交叉熵作为预测的监督损失:

$$\mathcal{L}_G = \mathbb{E}_{(x,y)\in T_i}(\widehat{\text{token}}_{i,\text{mask}}, \text{token}_{i,\text{mask}}) \tag{3}$$

此外, 为了进一步缓解模型遗忘问题, 本文对模型参数加上了额外的限制, 即对训练数据的解耦表征采取回归损失:

$$\mathcal{L}_{\text{reg}}^G = \ell_{(x,y)\in T_i} \| Net_G^{i-1}(x) - Net_G(x) \|_2 \tag{4}$$

其中, $Net_G^{i-1}(\cdot)$ 为在任务 T_{i-1} 学习后的解耦网络.

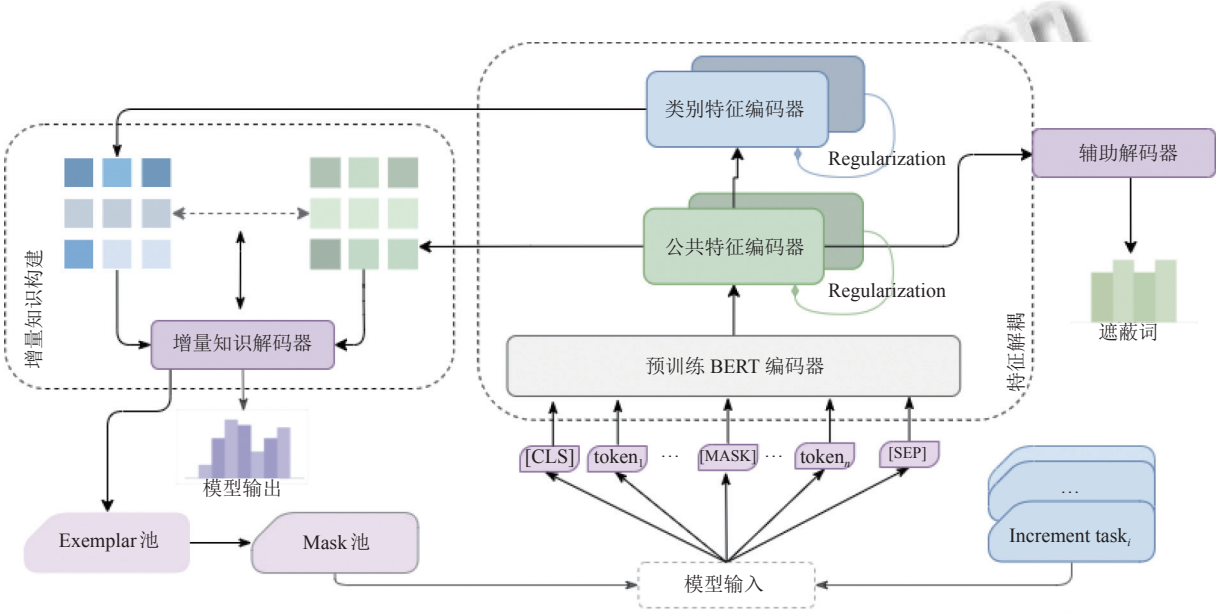

图 1 增量知识构造与遮蔽回放策略方法

1.2 类别特征网络

传统的深度学习模型中, 只要低层特征训练足够充分, 通过训练数据即可直接训练出一个性能优异的网络. 类似的, 通过大规模预料训练的底层特征处理器, 再加上足够的训练数据, 即可在训练任务 t_i 时, 使得类别特征网络记为 $Net_S(\cdot)$ 直接地关注当前的分类任务涉及的类别本身, 而无需再加上类别标签或者其他额外的训练技巧与方式. 具体的, 在训练任务 t_i 时, 在获得输入样本 x_i 的底层表征向量后, 将其输入到类别特征处理器 $Net_S(\cdot)$, 得到类别相关特征 e_S^i:

$$e_S^{ij} = \tanh\left(Net_S\left(\text{token}_j\right)\right) \tag{5}$$

然后, 通过第 3.4 节中的特征融合方法得到融合后的特征表示 E_{mix}, 并将其用作训练分类器 $Cls(\cdot)$ 的特征:

$$E_{\text{mix}}^i = Mix(E_s^i, E_G^i) \tag{6}$$

$$\widehat{y} = Cls\left(E_{\text{mix}}^i\right)|_n^{i=0} \tag{7}$$

最后, 再通过分类器到的标签进行分类损失学习, 同时对类别特征网络和分类器进行训练更新:

$$\mathcal{L}_c = \ell_{(x,y)\in T_i}(\widehat{y}, y) \tag{8}$$

此外, 对于类别特征网络, 本文对网络同样采用了一个回归损失以缓解灾难性以往问题:

$$\mathcal{L}_{\text{reg}}^s = \ell_{(x,y)\in T_i} \left\| Net_S^{i-1}(x) - Net_S(x) \right\|_2 \tag{9}$$

其中, $Net_S^{i-1}(\cdot)$ 为在任务 T_{i-1} 学习后的解耦网络.

1.3 增量知识构造

通过公共特征、类别特征解耦网络得到的特征 $E_G \in R^{n\times d}$ 和 $E_S \in R^{n\times d}$, 本文设计通过衡量特征解耦网络得到的特征, 并以其作为类别特征网络蕴含公共特征信息的置信度参数, 将两个特征进行融合. 具体的, 本文采用 CKA (centered kernel alignment) 方法[20], 衡量来自两个不同网络结构对同一数据源数据的特征相

似度:

$$HSIC(K,L)=\frac{1}{(n-1)^2}\mathrm{tr}(K(E_G,E_S)L(E_G,E_S)) \quad (10)$$

$$Sim(E_G,E_S)=CKA(E_G,E_S)=\frac{HSIC(K,L)}{\sqrt{HSIC(K,K)HSIC(L,L)}} \quad (11)$$

其中, $K_{ij}\left(E_i,E_j\right),L_{ij}(E_i,E_j)$是两个核函数, $HSIC$ 是 Hilbert-Schmidt 独立性准则. 然后, 通过得到的特征相似度为置信度参数, 动态地结合类别特征和公共特征:

$$E=Sim(E_G,E_S)\cdot E_S\oplus(1-Sim(E_G,E_S))\cdot E_G \quad (12)$$

1.4 遮蔽回放策略

在第t $(t>1)$个增量任务的学习过程中, 本文在一定的训练间隔间采取回放的策略, 并辅佐以额外的语言学习任务, 促进模型通过少量回放案例能快速地回忆先前的知识. 回放的样本是通过 K-means 的方法筛取样本$x_i\in\mathcal{T}_{n_t}$, 在筛选完样本后, 本文根据样本在当前模型中的表征, 随机选取词语作为遮蔽对象. 具体的, 在任务t_i的训练回合结束后, 通过 K-means 方法选取一部分样本作为样例 (exemplar), 然后通过 token 级别的随机遮蔽操作对词w_j进行处理:

$$\begin{cases}\mathrm{token}_i=[\mathrm{mask}], & p(\mathrm{token}_j)\geqslant k\\ \mathrm{token}_i=\mathrm{token}_i, & p(\mathrm{token}_j)<k\end{cases} \quad (13)$$

其中, $p(\mathrm{token}_j)$为取得对样本$examplar_i$的 token w_j进行遮蔽操作的概率, 若概率大于k, 则将原样例样本$examplar_i$中该词进行遮蔽处理, 随机替换成 mask 标签, 并将遮蔽后的样本进行临时保存, 以在之后的回放中进行预测学习. 与传统的随机选取的样本回放的任务不同, 基于预测遮蔽语言任务的遮蔽回放学习方法使模型通过预测遮蔽对象, 从而在少量样本的回放学习过程中, 快速“回忆”起先模型在先前任务上学习到的该领域的知识. 模型对遮蔽词的预测通过交叉熵损失训练监督:

$$\mathcal{L}_m=\mathbb{E}_{(x,y)\in\mathcal{T}_{n_t}}(\widehat{\mathrm{token}}_{i,\mathrm{mask}},\mathrm{token}_{i,\mathrm{mask}}) \quad (14)$$

模型总体训练优化目标是上述所有的损失的总和:

$$\mathcal{L}=\alpha_c\mathcal{L}_c+\alpha_G\mathcal{L}_G+\alpha_{\mathrm{reg}}\mathcal{L}_{\mathrm{reg}}^S+\alpha_{\mathrm{reg}}\mathcal{L}_{\mathrm{reg}}^G+\alpha_m\mathcal{L}_m \quad (15)$$

算法 1 展示了本文关于增量知识构造和基于遮蔽回放策略的类增量学习方法的完整过程.

算法 1. 增量知识构造和基于遮蔽回放策略算法

输入: $(x_j,y_j)\in T_i$.
输出: 增量训练后的 Net_G、Net_S 和$Cls(\cdot)$模型.

Begin
1. 分别用随机种子$seed_G$、$seed_S$、$seed_{Cls}$和$seed_D$初始化Net_G、Net_S、$Cls(\cdot)$、Net_D;
2. **For** 增量任务 $T_i\in\mathcal{T}_{n_t}$ **do**
3. **For** 增量任务 T_i 的 epoch **do**
4. 从增量任务 t_i 中选取一批数据(x_j,y_j);
5. 随机对 x_j 中的词进行遮蔽操作, 得到 $\hat{x}_j$;
6 初始化 $step$=0;
7. **If** $step$!=0 and $step\%replayfrequency$==0
8. 从记忆池筛选一批 exemplars $(x_j,y_j)\in\mathcal{T}_{i-1}$
9. 将选取的 exemplars 与输入数据拼接;
10. **Else**
11. 通过式 (1) 得到公共特征 E_G;
12. 通过式 (5) 得到任务特征 E_S;
13. 通过式 (10)–式 (12), 根据公共和任务特征E_G、E_S得到增量特征 E;
14. $Cls(\cdot)$通过式 (7) 预测 $\hat{y}$;
15. Net_D 通过式 (2) 预测 $\widehat{\mathrm{token}}_{\mathrm{mask}}$;
16. 通过式 (15) 计算损失 $\mathcal{L}$;
17. 根据损失$\mathcal{L}$与学习率ρ更新模型m_θ的参数;
18. $step$+=0;
19. **End For**
20. 选择 1% 已训练过的数据作为 exemplars;
21. 通过式 (13) 得到遮蔽后的样本;
22. **End For**
23. Return $m_{\theta'}$.
End

2 实验分析

本节中进行了完整的实验和分析. 首先, 讨论了本文的实施细节和训练细节. 接下来, 本文的实验结果与 SOTA 方法进行了比较. 最后, 本文进行消融研究并对结果进行分析.

2.1 模型实现

在类增量学习的实验设置中, 由于实际场景中任务的规模和数量是未知的. 基于所有任务的验证集数据的最优超参数方法, 例如网格搜索, 是过于乐观的. 在一定程度上, 模型的规模与模型的性能是成正比的. 所以通过上述的方法调整得到的模型并不具有代表性和说服力. 结合以上原因, 本文采用预训练的 BERT (https://huggingface.co/bert-base-uncased) 为底层的特征提取器, 并采用传统的线性层作为知识解耦网络, 这样更具有说服力和通用性. 分类器是由线性层与 Softmax 激活函数组成.

由于计算资源的限制, 本文遵循 Huang 等人[5] 的数据集设置. 具体来说, 本文对每个类别随机抽取 2 000

个训练实例. 本文对表 1 所示的任务序列进行了实验. 前 3 个是长度为 3 的任务序列, 遵循 Huang 等人[5] 的实验设置; 其他是长度为 5 的任务序列, 遵循 de Masson 等人[16] 的实验设置. 本文的实验环境为 11 GB 内存的 NVIDIA 3080Ti 上进行. 对于以前的任务的回放训练频率设置为每 10 个训练 step 一次. 增量训练和回放训练中, batchsize 和最大序列长度分别为 8 和 256. 超参数α_G、α_C 和 α_P都设置为 1. 回放训练中, 超参数α_{reg}设置为 2.0, 当前任务训练中, 则设置为 0.25. 本文采用 AdamW[21] 作为模型的优化器. 学习率和权重衰减分别被设置为 3E–5 和 0.01. 所有实验结果是 3 轮实验的平均.

表 1 增量实验的增量任务序列

序列编号	任务序列
1	AGNews → Yelp → Yahoo
2	Yelp → Yahoo → AGNews
3	Yahoo → AGNews → Yelp
4	AGNews → Yelp → Amazon → Yahoo → DBpedia
5	Yelp → Yahoo → Amazon → DBpedia → AGNews
6	DBpedia → Yahoo → AGNews → Amazon → Yelp
7	Yelp → AGNews → DBpedia → Amazon → Yahoo

2.2 实验分析

Replay[5,22]: 在 Finetune 方法的基础上, 该方法在对新任务进行增量学习时, 对以前的任务数据存储并对模型回放训练.

MBPA[5,23]: 该方法用情节性记忆模块来增强 BERT. 其利用 K-近邻来选择在测试时用于局部适应的例子.

LAMOL[5,24]: 该方法提出同时学习任务并产生训练样本. 模型学习生成用于样本数据作为回放的伪样本. 本文和 Sun 等人 [24] 一样用 Q&A 格式的数据来喂养 LAMOL.

IDBR[5]: 该方法用两个网络模块增强了 BERT, 并用作持续学习文本分类问题. IDBR 也利用了数据重放的方法.

L2P[25]: 该方法使模型在不同的任务中动态地通过提示进行学习. 该方法通过对提示词的优化, 使得模型再数据回放时提升对模型预测的准确性.

可以观察到, 直接对不同任务的序列进行微调会遭受大量的遗忘, 而在增量学习步骤中简单地存储和重放百分之一的示例, 有助于防止灾难性的遗忘. 另一种传统的增量学习方法, 正则化, 也在一定程度上缓解了灾难性遗忘. 但是他们的整体性能下降了大约 10%. 本文也将我们的方法与 SOTA 方法进行了比较. 本文为它们提供了一些额外的设置, 如局部适应策略[23] 等. 为 IDBR 和 LAMOL 提供了额外的任务标识符, 使其预测任务更加容易, IFCPR 仍然有一定的优势高过其他方法. 可以观察到, 即使没有任务标识符或测试时的额外局部适应方法, 本文的方法也比所有基线方法高出 1 个百分点左右. 此外, 本文对类别特征网络以及增量知识构造环节的回归损失也进行了实验, 从试验结果说明, 对上述两个网络结构参数进行一定的约束也能有效地避免模型在增量任务序列中对先前任务的遗忘.

本文在所有的增量任务序列中对方法进行测试与评估, 表 2 和表 3 分别展示了模型在长度为 3 与 5 的类增量任务中的性能. 本文方法在准确率、遗忘率[26] 方面一直优于所有比较的方法. 平均遗忘率 (average forgetting rate, *AFR*) 的定义如式 (16)、式 (17) 所示:

$$AFR = \frac{1}{i-1}\sum_{j=1}^{i-1} F_{\{i,j\}} \tag{16}$$

$$F_{i.j} = Acc_{\text{best}} - Acc_{i,j}, \forall j < i \tag{17}$$

其中, $F_{i,j}$表示在模型在结束任务i的训练后, 对之前学习的任务j遗忘的程度. Acc_{best}表示模型在学习任务i之前, 其在增量任务j上取得的最佳测试精度. $Acc_{i,j}$是学习任务i后在任务j上的测试精度. 从实验结果可以看出:

1) 模型直接地在增量任务序列上进行微调会遭受较为严重的遗忘从而导致性能的下降, 而在增量任务的训练过程中穿插少量地样本回放, 即便是百分之一的示例也有助于防止模型灾难性的遗忘. 另一种传统的增量学习方法, 即正则化方法, 也能够在一定程度上缓解了灾难性遗忘. 但是他们的总体表现下降了 10% 左右.

2) 本文将 IKCMR 的方法与 SOTA 方法进行了比较, 如表 4 所示. 即便为 SOTA 提供了一些额外的设置, 如为 IDBR方法和 LAMOL 方法提供了额外的任务标识符、为 MBPA 方法提供测试时的局部适应等使其预测任务更加容易, IKCMR 仍然以明显的优势胜过它们. 可以观察到, 即使没有任务标识符或测试的额外便利, 本文的方法也比所有基线方法高出 1 个百分点.

3) 除了准确性之外, IKCMR 在新任务的遗忘方面也有更好的表现. 计算了 IKCMR 在新的增量步骤 (任务) 上训练后的遗忘率, 如表 5 所示. IKCMR 的遗忘率在一些任务序列中对第一个所学任务的遗忘甚至接近零. 即便增量任务序列的影响客观存在, IKCMR 在防

止遗忘的策略中也起到了很大作用. 不难发现, 即使某些任务的遗忘率大大增加 (约 2%), 模型的遗忘率仍能在一定范围内保持稳定. 这充分证明了 IKCMR 能够在其他增量任务和数据回放期间进行学习和复习.

表 2 长度为 3 的增量任务序列实验 (%)

方法	准确率			平均准确率
	序列1	序列2	序列3	
Finetune1	25.79	36.56	41.01	34.45
Regularization1	71.5	70.88	72.93	71.77
Replay1	69.32	70.25	71.31	70.29
MBPA2	71.09	71.22	71.20	71.17
LAMOL2	71.24	71.62	71.32	71.39
IDBR1	71.80	72.72	73.08	72.53
L2P2	72.11	73.02	73.20	72.78
IKCMR	**73.44**	**73.13**	**73.22**	**73.26**
Upper-bound1	74.16	74.16	74.16	74.16

表 3 长度为 5 的增量任务序列实验 (%)

方法	准确率				平均准确率
	序列4	序列5	序列6	序列7	
Finetune1	32.37	32.22	26.44	30.12	30.29
Regularization1	72.28	73.03	72.92	72.89	72.78
Replay1	68.25	70.52	70.24	70.33	69.84
MBPA2	72.19	72.55	72.34	72.17	72.31
LAMOL2	73.38	73.45	73.35	73.37	73.39
IDBR1	72.63	73.72	73.23	73.34	73.23
L2P2	73.67	73.61	73.62	73.53	73.60
IKCMR	**73.76**	**73.65**	**73.69**	**74.24**	**73.84**
Upper-bound1	75.09	75.09	75.09	75.09	75.09

表 4 IKCMR 回归损失实验 (%)

设置	序列1	序列2	序列3
IKCMR w/o $\mathcal{L}_{\text{reg}}^{S}$	73.12	72.32	72.47
IKCMR w/o $\mathcal{L}_{\text{reg}}^{G}$	73.11	72.74	72.38
IKCMR w/o $\mathcal{L}_{\text{reg}}$	72.84	72.20	72.29

表 5 IKCMR 对第 1 个任务的平均遗忘率 (%)

增量任务序列	序列1	序列2	序列3	序列4	序列5	序列6	序列7
平均遗忘率	1.33	1.74	2.97	1.97	2.36	0.19	2.35

2.3 消融实验

特征融合策略的影响: 本文首先分析了方法的特征构建和遮蔽重放策略的组成部分, 并展示了它们对最终性能的影响. 所有这些消融研究都是在固定内存设置下进行的. 本文评估了 7 种特征融合策略.

$\text{IKCMR}_{\text{Concat}}$: 该方法是直接地将类别特征和一般特征连接起来, 重建最终的特征. 其他模块保持不变.

$\text{IKCMR}_{\cos}$: 该方法中, 本文利用余弦相似度来衡量类别特征和一般特征的差距, 并将其作为重构特征的权重. 其他模块保持不变.

$\text{IKCMR}_{\text{dot}}$: 该方法中, 本文使用类别特征与一般特征的点积得到的投影作为权重来重建特征. 其他模块保持不变.

$\text{IKCMR}_{\text{bilinear}}$: 该方法中, 本文使用双线性插值的特征作为权重来重建特征.

$\text{IKCMR}_{\text{pearson}}$: 该方法中, 本文使用特征的皮尔逊相关系数作为权重来重建特征, 而其他模块保持不变.

$\text{IKCMR}_{\text{euc}}$: 该方法中, 本文使用特征的欧氏距离作为权重来重建特征, 而其他模块则保持不变.

$\text{IKCMR}_{\text{CKA}}$: 该方法中, 本文使用特征的 CKA 距离作为向量的相关指数, 并将其作为权重来重建特征, 其他模块保持不变.

实验结果如表 6 所示, 特征融合策略使模型对不同的任务不敏感, 导致了不同的性能, 双线性插值的融合策略的融合效果最差. 余弦的融合策略在短序列的任务上取得了较好的平均性能, 我们认为基于角度的特征融合方法在任务序列较短时能取得较好的效果, 但是在高相似度的较长的增量任务中, 角相似度方法会造成严重的 CF 问题. 欧氏距离判别法和皮尔逊系数法的性能较为稳定. 本文采取的中心核对齐方法具有一定的不变性, 是通过计算数据点之间的相似性得到的, 其综合性能表现最好.

表 6 特征融合策略的消融实验 (%)

消融方法	序列1	序列2	序列3	序列4	序列5	序列6	序列7
$\text{IKCMR}_{\text{Concat}}$	71.80	72.72	73.08	72.63	73.70	73.23	73.34
$\text{IKCMR}_{\cos}$	72.45	72.89	73.38	74.01	73.36	73.40	73.43
$\text{IKCMR}_{\text{dot}}$	72.75	72.82	72.96	73.76	73.08	73.10	73.39
$\text{IKCMR}_{\text{bilinear}}$	72.19	72.42	72.28	72.45	73.10	73.18	73.25
$\text{IKCMR}_{\text{pearson}}$	72.44	72.79	72.95	73.91	74.11	73.29	73.79
$\text{IKCMR}_{\text{euc}}$	72.14	72.61	72.86	73.47	74.19	73.81	73.28
$\text{IKCMR}_{\text{CKA}}$	**73.44**	**73.13**	**73.22**	**73.76**	**73.65**	**73.69**	**74.24**

基于角度的特征融合方法在任务序列较短时能取得较好的效果, 但是在高相似度的较长的增量任务中, 角相似度方法会造成严重的 CF 问题. 欧氏距离判别法和皮尔逊系数法的性能较为稳定. 本文采取的中心核对齐方法具有一定的不变性, 是通过计算数据点之间的相似性得到的, 其综合性能表现最好.

回放策略的影响: 为了探究遮蔽重放策略对模型性能的影响, 本文对传统的回放策略与遮蔽回放策略

进行消融实验. 具体如下: 传统的回放策略、带有遮蔽语言模型任务的回放策略. 关于回放策略的消融研究结果如表 7 所示. 结果显示, 采用遮蔽回放策略的 IKCMR 有助于减轻遗忘的程度, 这也验证了遮蔽回放策略对性能提升的有效性与必要性. 在数据回放的采样效果是有限的前提下, 在数据回放训练期间增加训练任务可以有效地提高样本的利用率. 由此可以证明, 本文的遮蔽回放策略能够使模型在有限的回放样本中恢复更多的记忆.

表 7 回放策略的消融实验 (%)

模型	序列1	序列2	序列3	序列4	序列5	序列6	序列7
IKCMR w/o MLM	72.78	72.56	72.66	72.33	72.74	73.22	73.12
IKCMR	73.44	73.13	73.22	73.76	73.54	73.62	74.24

3 结论与展望

本文提出了一种面向自然语言处理的新型类增量学习方法. IKCMR 将特征以类为粒度进行解耦, 并根据互补学习理论动态地重构特征, 有效地缓解灾难性遗忘问题. 同时, IKCMR 利用遮蔽回放的方法, 快速调用以前任务的知识. 在 5 个数据集上的实验表明, IKCMR 在类增量学习任务上优于现有的最先进方法. 进一步的分析表明, CKA 的融合策略可以在跨特征重建过程中提取并结合更多有用的信息, 而遮蔽重放策略可以大幅提高增量学习的性能. 我们相信, 本文的方法可以扩展到 NLP 的其他增量学习任务, 如关系提取和命名实体识别. 本文计划对这些任务也进行进一步的研究.

参考文献

1 McCloskey M, Cohen NJ. Catastrophic interference in connectionist networks: The sequential learning problem. Psychology of Learning and Motivation, 1989, 24: 109–165.

2 Li ZZ, Hoiem D. Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(12): 2935–2947. [doi: 10.1109/TPAMI.2017.2773081]

3 Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv:1503.02531, 2015.

4 Masana M, Liu XL, Twardowski B, *et al*. Class-incremental learning: Survey and performance evaluation on image classification. arXiv:2010.15277, 2020.

5 Huang YF, Zhang YZ, Chen JA, *et al*. Continual learning for text classification with information disentanglement based regularization. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. ACL, 2021. 2736–2746.

6 Ke ZX, Liu B, Huang XC. Continual learning of a mixed sequence of similar and dissimilar tasks. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 18493–18504.

7 Zhou DW, Ye HJ, Zhan DC. Co-transport for class-incremental learning. Proceedings of the 29th ACM International Conference on Multimedia. Chengdu: ACM, 2021. 1645–1654.

8 Lee E, Huang CH, Lee CY. Few-shot and continual learning with attentive independent mechanisms. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 9435–9444.

9 Shen YL, Zeng XY, Jin HX. A progressive model to enable continual learning for semantic slot filling. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: ACL, 2019. 1279–1284.

10 Monaikul N, Castellucci G, Filice S, *et al*. Continual learning for named entity recognition. Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(15): 13570–13577.

11 Yan SP, Xie JW, He XM. DER: Dynamically expandable representation for class incremental learning. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 3013–3022.

12 Singh P, Mazumder P, Rai P, *et al*. Rectification-based knowledge retention for continual learning. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 15277–15286.

13 Castro FM, Marín-Jiménez MJ, Guil N, *et al*. End-to-end incremental learning. Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 241–257.

14 Wu Y, Chen YP, Wang LJ, *et al*. Large scale incremental learning. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 374–382.

15 Lopez-Paz D, Ranzato MA. Gradient episodic memory for continual learning. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long

Beach: Curran Associates Inc., 2017. 6470–6479.
16 de Masson AC, Ruder S, Kong LP, *et al*. Episodic memory in lifelong language learning. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 1177.
17 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2018. 4171–4186.
18 Radford A, Wu J, Child R, *et al*. Language models are unsupervised multitask learners. OpenAI Blog, 2019, 1(8): 9.
19 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.
20 Kornblith S, Norouzi M, Lee H, *et al*. Similarity of neural network representations revisited. Proceedings of the 2019 International Conference on Machine Learning. 2019. 3519–3529.
21 Zhang GQ, Kenta N, Kleijn WB. Extending AdamW by leveraging its second moment and magnitude. arXiv:2112.06125, 2021.
22 Wang H, Xiong WH, Yu M, *et al*. Sentence embedding alignment for lifelong relation extraction. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019. 796–806.
23 Yogatama D, de Masson d'Autume C, Connor J, *et al*. Learning and evaluating general linguistic intelligence. arXiv:1901.11373, 2019.
24 Sun FK, Ho CH, Lee HY. LAMOL: Language modeling for lifelong language learning. Proceedings of the 8th International Conference on Learning Representations. Addis Ababa: ICLR, 2019.
25 Wang ZF, Zhang ZZ, Lee CY, *et al*. Learning to prompt for continual learning. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 139–149.
26 Mai ZD, Li RW, Jeong J, *et al*. Online continual learning in image classification: An empirical survey. Neurocomputing, 2022, 469: 28–51. [doi: 10.1016/j.neucom.2021.10.021]

(校对责编: 孙君艳)