小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems

基于双向注意力和类生成器的小样本文本分类

王 婷[1]，朱小飞[1]，唐 顾[1]

[1]（重庆理工大学 计算机科学与工程学院，重庆 400054）

E-mail：tingwang2022@stu.cqut.edu.cn

摘 要：在小样本文本分类领域中，查询集和支持集的特征提取是影响分类结果的关键之一，但以往的研究大多忽略了两者之间存在匹配信息且在各自的信息提取中忽略了特征间的重要性程度不同，因此提出了一种新的小样本分类模型。模型结合 GRU 的全局信息提取能力和注意力机制的局部细节学习能力对文本特征进行建模，同时采用双向注意力机制来获取支持样本与查询样本间的交互信息，并创新性的提出“类生成器”用以区分同类样本间的不同重要性同时生成更具判别性的类别表示。此外，为了获得更为清晰的分类界限，还设计了一个原型感知的正则化项来优化原型学习。模型在 2 个小样本分类数据集上进行了实验，均取得了比目前最优基线模型更好的分类效果。

关键词：小样本学习；度量网络；双向注意力；文本分类

中图分类号：TP391 **文献标识码：**A

Few-shot Text Classification Based on Bidirectional Attention and Class Generator

WANG Ting[1]，ZHU Xiao-fei[1]，TANG Gu[1]

[1]（College of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China）

Abstract: In the field of few-shot text classification, the feature extraction of query set and support set is one of the keys to affect the classification results, but most of the previous studies ignored the matching information between the two and ignored the different importance of the features in their respective information extraction, so a new few-shot classification model is proposed. The model combines the global information extraction ability of the GRU and the local detail learning ability of the attention mechanism to model the text features. At the same time, the bidirectional attention mechanism is used to obtain the interactive information between the support samples and the query samples, and innovatively proposes "class generator". It is used to distinguish the different importance among similar instances while generating more discriminative class representations. In addition, to obtain clearer classification boundaries, an prototype-aware regularization term is designed to optimize prototype learning. The model was tested on two few-shot classification datasets, and both achieved better classification results than the current optimal baseline model.

Key words: few-shot learning；metric network；bidirectional attention mechanism；text classification

1 引 言

文本是大数据时代分布最广、体量最大、最易获取的信息载体，如何从大规模的文本数据中抽取出有价值的知识是当前亟待解决的难题。文本分类（Text Classification）是自然语言处理（Natural Language Processing，NLP）领域一个经典的任务，过去，研究人员采用人工手动对文本提取特征进行分类，但是伴随着移动互联网络的发展，文本数据呈爆炸式增长，利用人工手动对文本数据进行标注分类的方式因其耗时长且易受到标注人的主观认知影响而被舍弃，转而利用机器实现对文本数据的自动标注。传统的机器学习方法[1]主要通过人工提取特征构成特征向量，再采用支持向量机[2]、朴素贝叶斯[3]、决策森林[4]等算法从大量训练数据中学习分类器，利用分类器对待标注的文本数据进行分类，但此方法依赖于人为设计的规则和功能，同时该方法忽略了文本数据中的上下文信息，使得建模文本的语义信息变得困难。随着大数据时代的到来，基于深度学习算法的文本分类模型取得了巨大的进展，文本分类任务的准确率不断提升。与传统的方法相比，深度学习算法能够建模文本语义表示，解决计算和数据的局限性，显著提高文本分类的准确率。具有代表性的模型结构有三种，一是基于卷积神经网络[5]（Convolutional Neural Networks，CNN），二是基于循环神

收稿日期：2022-05-05 收修改稿日期：2022-07-04 基金项目：国家自然科学基金项目(62141201)资助；重庆市技术创新与应用发展专项项目(cstc2020jscx-dxwtBX0014)资助；重庆市教委语言文字科研资助项目(yyk20103)资助. 作者简介：**王 婷**，女，1997 年生，硕士研究生，CCF 学生会员，研究方向为文本分类；**朱小飞**(通讯作者)，男，1979 年生，博士，教授，CCF 高级会员，研究方向为自然语言处理和信息检索；**唐 顾**，男，1997 年生，硕士研究生，CCF 学生会员，研究方向为推荐系统.

经网络[6]（Recurrent Neural Network，RNN），三是图神经网络[7]（Graph Neural Network，GNN)。

虽然上述方法取得了重大进展，但是它的成功主要依赖于现有的大量有标签数据，然而在现实生活中，大量的有标签数据是不便获取的，所以这极大地限制了文本分类技术的发展与应用，因此，本文开始探索如何在已有少量标注样本的情况下进行文本分类。

近几年来，小样本文本分类问题受到专家学者们的关注，逐渐成为业界重要的研究方向。所谓小样本文本分类方法，其根本目标是希望机器能够像人类一样仅通过学习少量样本特征就能够实现准确的文本分类。现有的小样本分类方法主要分为以下五种：元学习、数据增强、图神经网络、提示学习和度量网络。基于元学习的小样本分类方法旨在学习一个通用的模型，使得这个模型在面对新旧任务时都可以在很少的梯度下降后达到较优解，其中最具代表性的是 Finn[8]等人在 2017 年提出的一种通用的元学习框架(Model-Agnostic Meta-Learning，MAML)，虽然该类方法较为简单，但在实际应用中目标数据与源数据之间存在差异可能会导致过拟合现象。基于数据增强的小样本分类方法是指借助已有的少量有标签样本，生成更多的增强数据用于训练，缓解样本不足的情况，帮助模型更好的进行训练，常用的生成式方法有生成对抗网络[9]和自训练[10]，但由于其生成了新的数据所以可能会引入噪声反而降低模型的分类准确率。基于图神经网络的小样本分类方法旨在借助图神经网络的消息传递思想将有标签样本的标签信息传递至无标签的样本上但其存在模型复杂度较高的问题。基于提示学习的小样本分类方法旨在通过构造提示模板和标签映射向输入增加“提示信息”，从而提升小样本分类的准确率，但其依赖于人工设计的模板与标签词，换言之，选择不同的模板与标签词都会对实验结果造成影响且该方法更适用于文本输入较短，包含类别数较少的英文数据集[11]。基于度量学习的方法直观易懂，其核心思想是在同一个嵌入空间中，通过给定的距离度量函数测量支持集与待分类的测试样本间的距离，以此来进行分类，距离相近则说明样本同属于一个类别，间隔较远则说明不属于同一类。常用的距离函数[12]包括欧几里得度量、皮尔逊相关系数、余弦相似度等。其中基于度量网络的小样本分类方法是本研究的基础与重点，具体研究内容将在第二章阐明。经典的度量网络结构有四种，分别是：双生神经网络[13]（Siamese neural network），匹配网络[14]（Matching network），原型网络[15]（Prototypical network）以及关系网络[16]（Relation network）。双生神经网络由两个相同结构、共享权值的神经网络连接而成。当训练样本与测试样本组成一对作为输入，分别通过两个神经网络后会输出其高纬特征向量表示，通过比较两个表征之间的距离来衡量两者的相似度，两个样本同属一类则标注为 1，否则标注为 0。双生神经网络并不是对输入进行分类而是进行区分，通过计算损失函数，最小化同类样本损失实现分类。随后，匹配网络被提出，Orid Vinyals 等人[14]设计了一个通用的 end-to-end 的网络框架，结合 LSTM 和注意力机制来捕获样本的表征，再使用余弦相似函数度量查询样本与支持样本之间的相似性，实现小样本分类目标。在训练阶段，模型要求支持集和查询集的数据分布必须相同，在训练的时候让匹配网络只学习每一个类别的少量样本，保证和测试过程的一致性。当标签分布存在较大误差时，该方法的分类效果会大打折扣。2017 年，另一种适用于小样本文本分类的网络架构-原型网络[15]被提出，该网络能够应用于不同的小样本数据集，是一种简单、高效的小样本的学习方式。原型网络的目标是学习到一个向量空间来实现文本分类任务，它的主要思想是先将所有的样本通过映射至低维的向量空间中，再对同属一个类别的多个样本求均值作为类别原型表示，针对每个待分类的查询样本，采用欧氏距离计算类别原型与查询向量之间的距离来确定分类结果。与以往固定的度量方法不同，FloodSung 等人[16]进一步研究了一种可迁移的深度度量网络-关系网络，整个网络由两部分组成，第一部分是特征提取模块，用于提取样本的特征信息，第二部分自适应度量模块，通过输出查询集特征信息与各个支持集特征信息之间的相似性得分，从而判断是否同属于一个类别。

目前通过各种途径，已有一些方法实现了在小样本场景下完成文本分类任务。尽管这些方法取得了一定的成效但是仍然面临着以下挑战：首先，有效标注样本数量少，文本语义稀疏，上下文信息未被充分挖掘，特异性表征提取不到位；其次，以往的研究大多忽略了支持样本与查询样本之间存在匹配信息且在各自的信息提取中忽略了特征间的重要性程度不同，最后，原始的原型网络模型难以生成更具区分性的类别原型表示，为解决上述问题，本文提出了一种新的小样本分类方法。

针对传统的网络结构无法捕获文本深层的语义信息且无法有效提取样本的重要特征，设计了一个嵌入注意力机制的双向循环神经网络（Bi-AGRU）作为特征提取器，同时考虑到支持集与查询集之间存在交互信息且在各自的信息提取中忽略了特征间的重要性程度不同，因此提出了融合支持样本和查询样本的双向注意力网络，除此之外，由于小样本学习场景中缺乏标注样本，所以如果同一类中支持样本之间的距离较远，则难以捕捉它们的共同特征并生成具有代表性的类别原型，如果不同类的支持实例在特征空间中彼此接近，则生成的原型是无法区分的，因此构造了一个结合双向 LSTM 和注意力机制的类生成器，通过非线性映射使原型向量的生成不易受到支持集中噪声的影响，并设计了原型感知的正则化项对模型进行优化。

综上所述，本文的主要创新点与贡献点如下：

（1）提出了一种新的小样本文本分类模型（Few-shot classification model based on bidirectional attention and class

generator,简称 BACG-FC 模型），结合注意力机制的局部细节学习能力和门控循环单元的序列建模能力对文本进行特征提取，使得模型可以全面建模文本的深层语义信息；

（2）构建了双向注意力网络，通过从 query2support 和 support2query 两个方向上计算注意力来获取支持样本与查询样本间的交互信息；

（3）提出了一个由双向长短期记忆网络和注意力机制构成的“类生成器”，用以更好地划分类别界限，生成更具区分性的类别表示；

（4）将本文提出的模型分别应用于 ARSC 和 FewRel 数据集，均取得了比目前最优基线模型更好的分类效果。

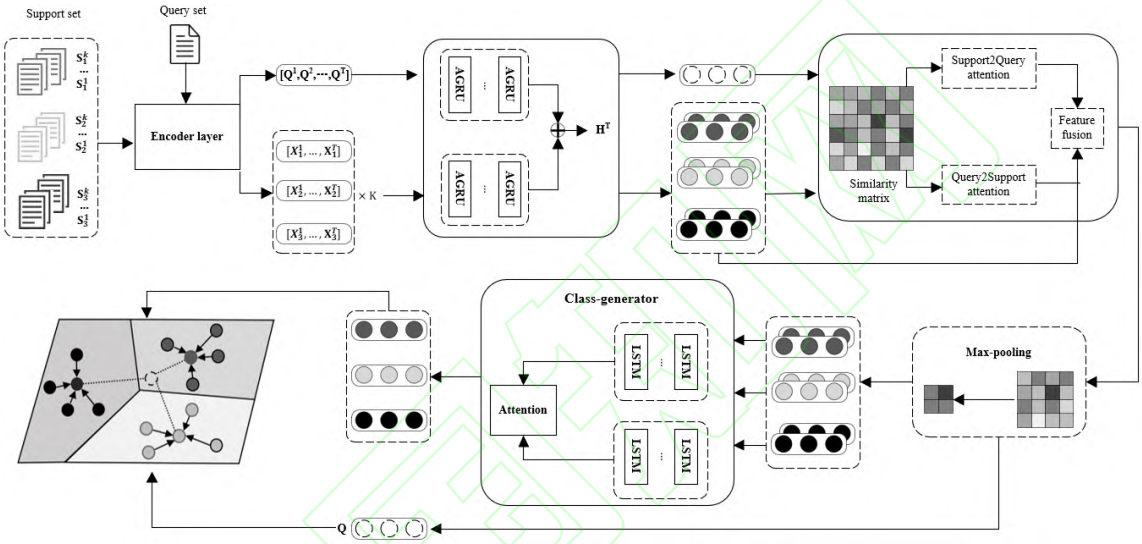

图 1 基于双向注意力和类生成器的小样本分类模型

Fig.1 Few-shot classification model based on bidirectional attention and class generator

2 BACG-FC 模型

本节详细介绍了所提模型的具体实现过程，包括问题的定义、整体模型架构以及各模块的细节说明。首先模型的输入是多个支持样本和查询样本，通过词嵌入模块获得样本的固定词嵌入表示，再在双向门控循环单元的更新门中引入注意力分数，替换原始的更新门，得到 Bi-AGRU，借助其捕获对应的文本级特征表示，之后使用双向注意力网络融合支持集与查询集的匹配信息，得到支持感知向量表示和查询感知向量表示，将原始特征向量与支持感知向量或查询感知向量进行拼接融合，得到最终的支持样本特征表示和查询样本特征表示，然后采用带有注意力机制的双向 LSTM 作为类生成器，生成更具代表性的原型表示，最后度量查询样本与类别原型之间的相似性实现小样本文本分类。整体架构如图 1 所示。

2.1 问题定义

本文将数据集 D 分为训练 D_{train} 集和测试集 D_{test}，其中训练集和测试集所包含的样本类别各不相同，两者都有各自的标签集合 Y_{train} 和 Y_{test}。针对小样本分类问题，其旨在训练出一个可以从 D_{train} 中学习先验知识的分类器，学习过程主要分为两个阶段：元训练和元测试。元训练阶段需从同一任务分布中划分出多个元子任务 T_i，再从训练集 D_{train} 中随机抽取包含 N 个类别，每个类别 K 个样本，一共 $N \times K$ 个样本的子集 S 进行训练，然后将来自 T_i 的 N 个类别上的剩余样本作为测试集 Q 进行测试，为避免引起混淆，将元子任务中的训练集定义为“支持集”（support set），测试集定义为“查询集”（query set）。

2.2 特征提取模块

特征提取模块包括单词嵌入模块和上下文编码模块。假设支持集和查询集中每个文本都包含 T 个单词，在单词嵌入模块中，本文使用 GloVe[17]预训练词向量获取每个单词的固定嵌入表示，表达式如下：

$$x^i = f(w^i) \tag{1}$$

其中，f表示映射函数，w^i表示文本中的第i个单词，$x^i \in R^d$为经过映射后的第i个单词的向量表示。第k个支持样本表示为$X_k = [x^1, ..., x^T]_k \in R^{T\times d}$，查询样本表示为$Q = [q^1, ..., q^T] \in R^{T\times d}$。

在上下文编码模块中，本文应用了一个嵌入注意力机制的双向 GRU 模型细化单词的表示，具体过程如下，首先是将第k个文本序列作为输入，然后使用注意力机制得到第t个时间步的输入向量x^t的注意力分数α_k^t，最后将注意力得分与 GRU 中的更新门结合构建 AGRU(如图 2 所示)，实现公式如下：

$$U_k = WX_k + b \tag{2}$$

$$\alpha_K = softmax(U_k) \tag{3}$$

其中，$\alpha_K = [\alpha_k^1, ..., \alpha_k^T]$，$k$表示第$k$个样本，$T$表示样本中单词个数。

$$z_t = \sigma(W^z[h_{t-1}, x_t]) \tag{4}$$

$$z_t^{'} = \alpha_k^t * z_t \tag{5}$$

$$r_t = \sigma(W^r[h_{t-1}, x_t]) \tag{6}$$

$$h_{t-1}^{'} = h_{t-1} * r_t \tag{7}$$

$$\tilde{h_t} = \tanh\left(W^{\tilde{h}}[h_{t-1}^{'}, x_t]\right) \tag{8}$$

$$h_t = \left(1 - z_t^{'}\right) * h_{t-1} + z_t * \tilde{h_t} \tag{9}$$

其中，W、W^z、W^r、$W^{\tilde{h}}$为可训练的权重参数，b是注意力机制的偏置项，α_i为注意力得分，σ为$sigmoid$函数，通过这个函数可以将数据转换为$0\sim1$范围内的数值，z_t是控制更新的门，范围是$0\sim1$，掌控着多少前一时刻的状态是当前需要记忆存储的，$z_t^{'}$为结合了注意力分数的更新门，r_t是控制重置的门，范围是$0\sim1$，控制着多少新的状态只是旧状态的拷贝，h_{t-1}、$h_{t-1}^{'}$、h_t是隐层状态表示。

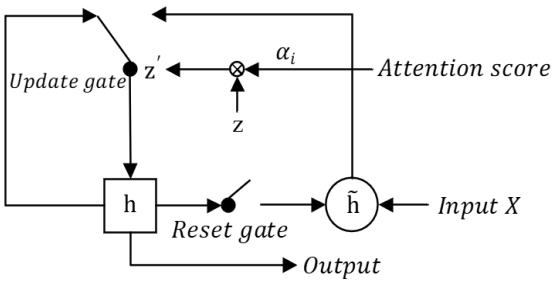

图 2 AGRU 结构图

Fig.2 Structure of AGRU

为了能够提取更为全面的特征信息，本文将 AGRU 双向化。对于每个输入x_i，可以得到它的正向隐层状态表示$\overrightarrow{h_i}$和反向隐层状态表示$\overleftarrow{h_i}$，其计算公式如下：

$$\overrightarrow{h_i} = \overrightarrow{AGRU}(x_i) \tag{10}$$

$$\overleftarrow{h_i} = \overleftarrow{AGRU}(x_i) \tag{11}$$

为了从两个方向捕获信息，选择在两侧连接单词w_i的隐藏表示：$H_i = \left[\overrightarrow{h_i}; \overleftarrow{h_i}\right]$，得到双向的语义信息。为了方便表述，这里使用$S^k \in R^{T_k\times 2d_h}$表示支持集中第$k$个样本的输出矩阵，表达式见式(12)，使用$\breve{Q} \in R^{T_q\times 2d_h}$来表示查询集的输出矩阵，表达式见式（13），$d_h$是 AGRU 的隐层大小，$T_k$表示第$k$个支持样本包含的单词数，$T_q$表示查询样本包含的单词数。

$$S^k = BiAGRU(X_k) \tag{12}$$

$$\breve{Q} = BiAGRU(Q) \tag{13}$$

2.3 双向注意力网络

双向注意力层考虑查询样本与每个支持样本间的匹配信息，以交互的方式对它们进行编码，从 support set 到 query set 和从 query set 到 support set 两个方向上计算注意力。双向注意力计算的前提是得到一个共享的词级相似性矩阵$M \in R^{T_q\times T_k}$，这个相似性矩阵的含义是计算查询样本与第k个支持样本之间的逐词相似度，该矩阵计算公式如下：

$$M_{ij}^k = \breve{Q}_i \cdot S_j^{kT} \tag{14}$$

其中，Q_i表示$\breve{Q}$的第i行，S_j^k表示S^k的第j行。

支持集到查询集的注意力强调的是查询样本中哪些词与支持样本中的单词更相关，通过公式（15）的计算，可以获得包含所有查询样本信息的第k个支持样本的查询感知表示$\tilde{S}^k$，T_k表示第k个支持样本中包含的单词数。

$$\tilde{S}^k = \frac{\exp\left(M_{ij}^k\right)}{\sum_{j=1}^{T_k}\exp\left(M_{ij}^k\right)}S \tag{15}$$

查询集到支持集的注意力计算方式与上述类似，通过计

算支持样本中哪些词与查询样本中的单词更加相关得到查询样本的支持感知表示 $\tilde{Q}$，具体计算过程见公式（16），T_q表示查询样本中包含的单词数。

$$\tilde{Q}=\frac{\exp(M_{ij}^{k})}{\sum_{i=1}^{T_q}\exp(M_{ij}^{k})}\tilde{Q} \tag{16}$$

然后融合每个支持样本和查询样本的原始特征表示和感知向量表示。针对支持样本，其融合表示为式（17），其中，$g(\cdot)$表示 $ReLU$。

$$\bar{S}^{k}=g([S^{k};\tilde{S}^{k};S^{k}\odot\tilde{S}^{k}]W) \tag{17}$$

针对查询样本，其融合表示为式（18），其中，W是一个可训练的参数矩阵，$[;]$为拼接操作，⊙表示哈达玛积，$g(\cdot)$表示 $ReLU$。

$$\bar{Q}=g([Q;\tilde{Q};Q\odot\tilde{Q}]W) \tag{18}$$

2.4 度量网络模块

度量网络模块是基于原型网络进行改进优化的。传统的原型网络先通过对每个类的支持集样本求取均值得到类别原型表示，再通过计算查询样本与每个类原型之间的距离来实现分类的。但是考虑到传统方法生成的类别原型向量易受到支持集中个别噪声数据的影响而丢失准确性，且每个支持样本对于类别原型的贡献程度是不同的，因此在计算类别原型时采用融入注意力机制的双向 LSTM 作为类原型生成器以获得更具代表性的原型表示（如图 3 所示），再通过度量类别原型与查询样本之间的距离，实现文本分类。

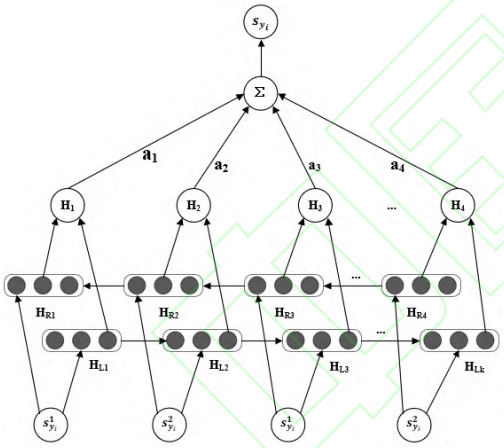

图 3　类生成器结构图

Fig.3　Structure of class generator

首先，通过最大池化将支持样本和查询样本分别汇总为单个特征向量 $\bar{s}^{k}$和 $\bar{q}$，其次使用类原型生成器为每个类生成类别原型表示 $\{\bar{S}_1,\bar{S}_2,\ldots,\bar{S}_N\}$，最后根据查询特征向量和类别原型的相似性预测查询样本属于哪个类，具体计算过程如下：

$$\bar{s}^{k}=maxpooling(\bar{S}^{k}) \tag{19}$$

$$\bar{q}=maxpooling(\bar{Q}) \tag{20}$$

$$\bar{S}_i=AttBiLSTM(\bar{s}_i^{1},\bar{s}_i^{2},\ldots,\bar{s}_i^{k}) \tag{21}$$

$$h(\bar{S}_i,\bar{q})=v^{T}(ReLU(W[\bar{S}_i;\bar{q}])) \tag{22}$$

$$p(y_i|\{\bar{S}_j\}_{j=1}^{N},\bar{q})=\frac{\exp(h(\bar{S}_i,\bar{q}))}{\sum_{j=1}^{N}\exp(h(\bar{S}_j,\bar{q}))} \tag{23}$$

其中，N表示支持集的类别数，K表示支持集的样本数，$maxpooling(.)$表示最大池化，v、W为可学习的超参数，$[;]$为拼接操作。

在训练过程中，采用交叉熵损失[18] $loss_{T_i}$来优化模型，Q 表示每个训练轮次中采样的查询集，$|D_Q|$表示查询样本的数量。除此之外，本文还设计了一个原型感知的正则化项 $loss_{proto}$来进行优化，使得类内距离更为接近，类间距离更为疏远。具体公式如下：

$$loss_{T_i}=-\frac{1}{|D_Q|}\sum_{\bar{q}\in Q}\log(p(y_i|\{\bar{S}_j\}_{j=1}^{N},\bar{q})) \tag{24}$$

$$loss_{proto}=\frac{1}{NK}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{K}\max(0,d(\bar{S}_i,\bar{s}_i^{k})+\gamma-d(\bar{S}_i,\bar{s}_j)) \tag{25}$$

其中，γ是一个超参数，$\bar{s}_i^{k}$表示属于第 i 类的第 k 个样本表示，$\bar{s}_j$表示不属于第 i 类的任一样本表示，$\bar{S}_i$表示第 i 类的类别原型。最终，损失函数 L 定义如式（26）：β是一个超参数，用于控制两部分损失的相对重要性。

$$L=loss_{T_i}+\beta loss_{proto} \tag{26}$$

3　实 验

本节介绍了研究过程中所使用的 2 个小样本数据集、实验环境的配置信息、超参的设置详情及评价指标，并对 2 个数据集上的实验结果进行了分析。

3.1 实验数据集

为了验证所提方法的有效性与适用性，本文在 2 个公开数据集上进行了对比实验，两个实验数据集的统计信息如表 1 所示，Dataset 表示数据集，Num.train 表示训练样本数，Num.test 表示测试样本数，Vocab size 表示词汇数量，Avg.len 表示文本的平均长度。

亚马逊评论情感分类数据集[19](Amazon Review Sentiment Classification，ARSC)由 Yu 等人提出，该数据集由 23 种亚马逊商品的评论数据组成，针对每一种商品，构建了三个具有不同评分阈值的二分类任务，评分阈值分别设置为 5 星、4 星和 2 星。基于此，共构建了 69 个分类任务，为了进行评估，本文从 4 个领域（书籍、DVD、电子产品、厨房）中选择 12 个任务作为元测试集，其余 57 个任务作为元训练集[20]。对于目标任务，创建了 2-way 5-shot 学习问题。

实体关系抽取数据集[21](Few-shot Relation classification，FewRel)覆盖了 100 种关系，每种关系 700 个注释实例，本次实验使用公开发布的 80 种关系，48 种关系作为训练集，12 种关系作为验证集，剩下的 20 种关系进行预测。

3.2 实验环境搭建

本研究的实验配置为：Ubuntu 20.04.3 操作系统，AMD Ryzen 5 PRO 3500U w/ Radeon Vega Mobile Gfx 2.10 GHz 的计算机，NVIDIA RTX 1080Ti 的 GPU，Python 3.7.5 的开发环境以及 PyTorch 1.3.1 的学习框架。

3.3 实验设置

本文在 ARSC 数据集上进行了 2-way 5-shot 的实验，词编码阶段采用 300 维的 GloVe 词向量进行初始化，最大句长设置为 128，在 FewRel 数据集上构造 5-way 5-shot 任务，采用 50 维的 GloVe 词向量进行初始化，最大句长设置为 40。在特征提取模块，设置 GRU 的隐层状态大小为 128。为了避免训练过度还设置了早停，模型的所有参数均采用随机梯度下降策略[22]（SGD）进行优化。初始学习率设为 0.1，学习率衰减步长为 3,000，衰减率为 0.1，为了防止小样本常出现的过拟合现象，本研究设置了 dropout 参数为 0.2。模型一共训练 30,000 轮，每 1,000 轮进行一次测试，每次测试阶段包含 1,000 轮，定义每轮实验结果作为单轮准确率，每次测试阶段的平均准确率作为该模型的阶段准确率，取最好的阶段准确率作为模型的结果。另外，对于超参数 γ和 β的取值在 3.6.3 章节进行了实验，最终选定 γ=1，β=1。

表 1 数据集统计

Table 1 Dataset statistics

Dataset	Num.train	Num.test	Vocab.size	Avg.len
ARSC	119,745	18,627	206,913	98.62
FewRel	42,000	14,000	124,577	24.99

3.4 评价指标

本文所采用的评价指标为正确率(ACC)，正确率表示在所有样本中预测正确的样本数量占总样本数量的比例。A 表示分类器预测标签为正，实际标签也为正的样本数，B 表示分类器预测标签为负而实际标签也为负的样本数，C 表示分类器预测标签为正而实际标签为负的样本数，D 表示分类器预测标签为负而实际标签为正的样本数。

$$ACC = \frac{A+B}{A+B+C+D}$$

3.5 对比模型

在本次研究中，选取了多种小样本分类模型作为基线模型，分别在 ARSC 和 FewRel 数据集上进行了实验，下面对基线模型进行简要介绍。

Proto Net:[15]原型网络，通过学习一个嵌入函数将所有样本映射到统一的向量空间中，并根据支持样本的句子嵌入均值来生成类别原型，最后通过设定好的度量函数来判断查询样本的类别。

Relation Net:[15]关系网络，采用神经网络进行距离度量，使得模型进行端到端的训练并通过汇总支持集中的样本向量来计算类别向量。

ROBUSTTC-FSL:[19]根据各子任务间的差异实施聚类操作，不同的子任务类别自动生成对应的度量方式。

DC-GNN:[20]一种双通道图神经网络模型，借助图的标签传播机制，通过共享两通道的信息传播矩阵解决了元学习框架下的监督信息稀疏化，缓解了图神经网络中过度平滑问

题。

MAML:[23]采用元学习方法解决小样本问题的经典模型之一，是一种与模型无关的算法，可以兼容各种模型并且适用于各种任务。该算法最大限度地提高了新任务损失函数的敏感性,因此当参数发生微小变化时便可以大大改善任务的损失，实现快速的收敛。

GNN:[24]一种采用图神经网络解决小样本学习的算法，适用于非结构化数据,其核心思想是借助图结构将有标注样本的标注信息传递至待标注样本中，实现最终的分类。

SNAIL:[25]一种简单且通用的元学习器架构，利用时序卷积神经网络和软注意力学习支持样本的标签信息,借助学习到的信息对序列的最后一个样本进行预测,该方法可以快速的学习和吸收以往的经验，显著的提升性能。

TPN:[26]利用转导的思想，为整个语料库构建一个无向权重图，通过标签传播的方式得到预测结果。

Meta Network:[27]借助高阶元学习器来监督训练过程，利用损失梯度生成快权重，有助于模型快速适应新的任务。

Induction Network:[28]它通过将元学习框架与动态路由算法相结合来学习广义的类别表示，整体是 end-to-end 的元训练，具有良好的可扩展性。

MEDA:[29]通过在元学习中引入数据增强方法，生成置信度高的增强样本以增加新类别的样本数量,提高模型在小样本情况下的泛化能力。MEDA-PN 表示采用原型网络进行度量。

3.6 实验结果分析

3.6.1 对比实验

表 2 和表 3 是不同模型在实体关系抽取数据集（FewRel）和亚马逊评论情感分类数据集（ARSC）上的评测结果（粗体部分表示在 2 个数据集上的最优结果）。

实验结果显示，在 FewRel 数据集上，DC-GNN 的准确率明显高于 GNN 和 Proto Net，其原因在于它融合了支持本的全局特征与查询样本的标签信息，而在 ARSC 数据集上，Induction Network 相较于度量网络的两大经典模型-Proto Net 和 Relation Net，分类准确率分别提升了 17.46%和 2.56%，其原因有两个，一是提供了一个可学习的非线性分类器，在分类的能力上要优于传统的线性分类器，二是融合了动态路由算法，针对类别原型表示进行了改进，将每个类别中的样本表示凝练成了更具代表性的类别表征，获得了更优的分类性能，这验证了好的类别表征能够提升模型的性能。与 Induction Network 相比， MEDA-PN 的分类准确率达 85.68%，其主要原因在于该模型提出了一个球生成器，用以生成更多的样本进行训练，从而改善了模型的性能。而本章所提的方法在 5-way 5-shot 和 2-way 5-shot 的设定下均优于所有的基准模型,这是因为我们的方法在原型网络的身上充分吸取了教训,在建模时重视查询集和支持集的交互信息并针对不同的特征分配了不同的注意力权值,且不再采用简单的均值法来获取类别表征，有效提高了模型的分类性能。同时本文为了探究不同预训练词向量对模型性能的影响，在两个少样本数据集上进行了对比实验（实验结果见表 2 和表 3 粗体部分）。从结果中我们不难看出，当使用了 Bert[30]作为预训练词嵌入表示时，两个数据集上的准确率均有所提升，这是因为相较于 GloVe，Bert 提供了更高质量的词嵌入来表示上下文的语义信息。

表 2 不同模型在 FewRel 数据集上的准确率对比

Table 2 Comparison of the accuracy of different models on the FewRel dataset

Methods	Acc(%)
MAML(2017)	65.73
TPN(2018)	67.50
Meta Network(2017)	71.02
SNAIL(2018)	78.07
GNN(2018)	79.47
Proto Net(2017)	85.10
DC-GNN(2021)	85.86
Ours(GloVe)	**86.25**
Ours(Bert)	**87.59**

表 3 不同模型在 ARSC 数据集上的准确率对比

Table 3 Comparison of the accuracy of different models on the ARSC dataset

Methods	Acc(%)
Proto Net(2017)	68.17
MAML(2017)	78.33
GNN(2018)	82.61
SNAIL(2018)	82.57
Relation Net(2018)	83.07
ROBUSTTC-FSL(2018)	83.12
Induction Network(2019)	85.63
MEDA-PN(2021)	85.68
Ours(GloVe)	**86.54**
Ours(Bert)	**86.90**

3.6.2 消融实验

我们通过去除模型的特定部分来进行消融实验以验证其影响。

首先，探究了特征提取模块的效果。在 ARSC 数据集上，通过观察 LSTM、GRU、双向 LSTM、双向 GRU 和双向注意力 GRU 这五种不同结构充当上下文特征提取模块时模型准确率的变化，验证双向注意力 GRU 的有效性。对比结果如图 4 所示，相较于单向结构，采用双向结构的准确率增长了约 1.8 个百分点，产生该结果的原因在于双向结构更能捕获文本的全局信息，在进行分类时不仅考虑到了文本上文的信息，同时还结合了文本末端的信息，有助于实现正确的分类。例如在 ARSC 数据中，因单向结构仅关注上文信息，因此容易将亚马逊购物评论“我很喜欢这条裙子，但是它缩水太严重了，我没法穿出去。”（译文）判定为积极态度，然而结合上下文看来该评论应判定为消极态度。所以，相较于单向提取特征，双向结构在处理文本分类任务中更具优势。除此以外，还观察到在处理小样本文本数据时，无论是单向或双向，GRU 结构的准确率均高于 LSTM 结构，这说明了 GRU 结构更擅长于提取文本信息，同时不难注意到嵌入了注意力机制的双向 GRU 效果要比双向 GRU 更好，这是因为注意力机制有助于捕获重要的局部信息，再结合双向 GRU 捕获全局信息，有效的提升了模型分类的准确性。

其次，为了证明双向注意力模块的有效性，在 FewRel 数据集上分别就无注意力特征提取模块、自注意力特征提取模块和双向注意力特征提取模块进行了对比实验，结果如表 4 所示。由表 4 可知，当模型去除注意力模块后(w/o att)，准确率大幅下降，因此注意力机制是模型不可或缺的一部分，在特征提取模块发挥重要作用，生成更具区分性的语义表征，从而提升模型的分类性能，此外，还发现融入了自注意力模块(Self-att)的模型准确率为 85.71%，与融入了双向注意力模块(Bi-att)的准确率相比低了约 0.6 个百分点，这是因为双向注意力模块不仅专注于支持集的信息，还考虑到了查询集的信息，通过在两个方向上使用注意力机制，学习支持集与查询集之间的联系。

表 4 不同注意力模块在 FewRel 数据集上的准确率

Table 4 Accuracy of different attention modules on FewRel dataset

Attention module	Acc(%)
w/o att	84.13
Self-att	85.71
Bi-att	86.25

然后，为了探索类别数和样本数对于不同模型性能的影响，我们在 FewRel 数据集上进行了实验。首先验证不同测试类别数对于分类准确率的影响。固定样本数为 5，即 5-shot，设置测试类别的数量范围为 5 到 10，同时选取 Proto Net、GNN、DC-GNN 作为对比模型，实验结果如图 5 所示。

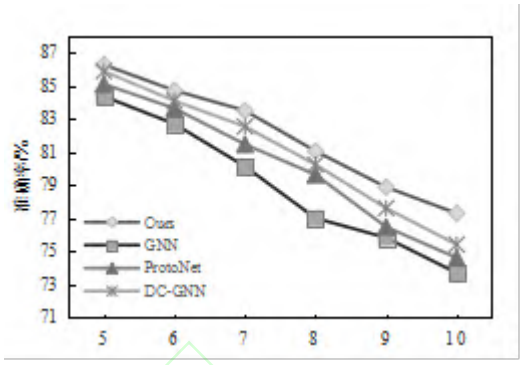

图 5 N-way 5-shot 下各模型准确率对比结果(N 取值范围为 5-10)

Fig.5 Accuracy comparison results of each model under N-way 5-shot (N ranges from 5 to 10)

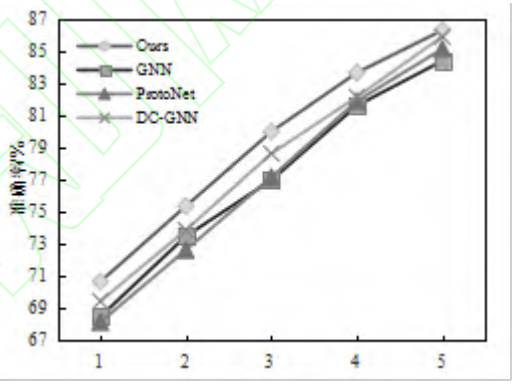

图 6 5-way N-shot 下各模型准确率对比结果(N 取值范围为 1-5)

Fig.6 Accuracy comparison results of each model under 5-way N-shot (N ranges from 1 to 5)

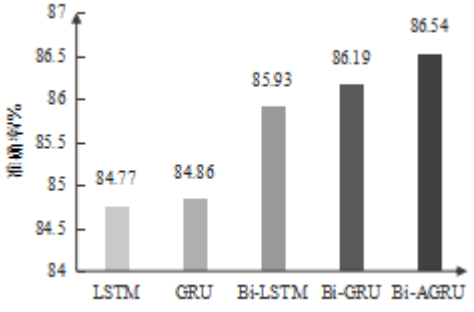

图 4 不同特征提取模块在 ARSC 数据集上的准确率

Fig.4 Accuracy of different feature extraction modules on ARSC dataset

然后验证不同的样本数对于分类准确率的影响。固定测试类别数为 5，即 5-way，设置样本数的范围为从 1 到 5，同样选择 Proto Net、GNN、DC-GNN 作为对比模型，实验结果如图 6 所示。可以看出，本文提出的模型在所有设置下均优于三个对比模型，同时可以注意到随着测试类别数的增多，所有模型的准确率均随之下降，随着支持样本数的增多，所有模型的准确率均随之上升。值得注意的是，本文提出的模型在测试类别数过多或者样本数过少的情况下依然优于所有的对比模型，验证了所提方法的有效性与健壮性。

最后，分析了不同类别原型对于模型准确率的影响，实验结果如表 5 所示。由于小样本任务具有支持样本稀缺的特殊性，所以当样本嘈杂噪声较大时，容易出现个别样本表示远离其他同类样本表示导致生成的类别原型出现巨大误差，降低模型的性能。因此，为了获得更为准确的类别原型，本文提出了一个类原型生成器，使得类别原型的生成更为灵活，同时也能将更多的注意力集中在那些与查询相关的样本上，减少噪声的影响。可以看出，使用类原型生成器替代均值原型和注意力原型能够正确定位与查询样本最为相似的样本，有效提高模型的准确率。

表 5　在 ARSC 数据集上不同类别原型对于模型准确率的影响

Table 5　The influence of different types of prototypes on the model accuracy on the ARSC dataset

Prototypes	Acc(%)
均值原型	85.93
注意力原型	86.27
类原型生成器	86.54

3.6.3　参数敏感性实验

在这一章中探究了超参数 γ 和 β 的取值对于模型分类准确率的影响，首先固定 β 的值为 1，在[0.8、1.0、1.2]中探究 γ 的最佳取值，然后固定 γ 的值为 1，在[0.8、1.0、1.2]中探究 β 对于模型性能的影响，结果如表 6 和表 7 所示，不难看出，当超参数 γ 和 β 均取 1 时，模型的分类效果是最好的。

表 6　超参数 γ 对于模型性能的影响

Table 6　The effect of hyperparameter γ on model performance

γ	0.8	1.0	1.2
ARSC	86.09	86.54	86.20
FewRel	85.92	86.25	86.12

表 7　超参数 β 对于模型性能的影响

Table 7　The effect of hyperparameter β on model performance

β	0.8	1.0	1.2
ARSC	85.96	86.54	86.30
FewRel	85.88	86.25	86.01

4　结束语

基于已有的小样本实现准确的文本分类是 NLP 领域一个正在攻克的难题。本研究针对传统的网络结构无法有效提取样本重要特征设计了一个嵌入注意力机制的双向循环神经网络作为特征提取器，同时对查询集和支持集之间的相互依赖关系进行了探索，提出了融合支持样本和查询样本的双向注意力网络。在此基础上，还对原型网络中原型向量的表示进行了改进，构造了一个类原型生成器，改变了原有的均值计算法与加权求和计算法，使原型向量的生成更为灵活，并设计了原型感知的正则化项对模型进行优化以提升模型分类的准确性。

虽然本研究取得了一些进展，但仍有诸多待改进之处，在后续的研究中有待重视：

（1）在小样本文本分类任务中，本文致力于挖掘文本深层语义信息，但是支持集所含信息不足以支撑模型进行更细粒度的分类研究，未来应该考虑结合文本增强技术和外部知识来增强不同任务间的文本表示，弥补训练数据不足的问题。

（2）分类结果的好坏不仅仅与建模的算法相关，还取决于模型超参数的设置。目前超参数的设置大多是人为通过不断地实验得到，该过程既耗时又耗力且收效甚微。如果能够实现超参数的自动化调优，将会大大减少人工手动调参的时间还有助于模型实现更好的分类效果。

（3）本文设计了一个原型感知的正则化项来进行优化，使得类内距离更为接近，类间距离更为疏远。但同时有监督的对比学习也被提出用于拉近类内距离，拉远类间距离，未来可以考虑使用对比学习来进行类生成器的优化。

References:

[1] Zhao Hai-yan, Cao jie, Chen Qing-kui, et al. Methods for hierarchical multi-label text classification[J]. Journal of Chinese Computer Systems, 2022, 43(4):673-683.

[2] John C P. Sequential minimal optimization: a fast algorithm for training support vector machines[J]. MSRTR: Microsoft Research, 1998, 3(1): 88-95.

[3] Heckerman D. A tutorial on learning with Bayesian networks[J]. Innovations in Bayesian Networks, 2008, 156(1): 33-82.

[4] Kontschieder P, Fiterau M, Criminisi A, et al. Deep neural decision forests[C]//IEEE International Conference on Computer Vision,2015: 1467-1475.

[5] Wang J, Wang Z, Zhang D, et al. Combining knowledge with deep convolutional neural networks for short text classification[C]//26th International Joint Conference on Artificial Intelligence (IJCAI),2017, 350.

[6] Liu P, Qiu X, Huang X. Recurrent neural network for text classification with multi-task learning[C]//25th International Joint Conference on Artificial Intelligence (IJCAI),2016: 2873-2879.

[7] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[C]//5th International Conference on Learning Representations(ICLR),2017.

[8] Finn C, Abbeel P, Levine S, et al.Model-agnostic meta-learning for fast adaptation of deep networks[C]//International Conference on Machine Learning, 2017: 1126-1135.

[9] Creswell A, White T, Dumoulin V, et al. Generative adversarial networks:an overview[J].IEEE Signal Processing Magazine, 2018, 35(1): 53-65.

[10] Xie Q, Luong M T, Hovy E, et al. Self-training with noisy student improves imagenet classification[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition,2020: 10687-10698.

[11] Gao T, Fisch A, Chen D. Making pre-trained language models better few-shot learners[C]//59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers),2021: 3816-3830.

[12] Zhao K L, Jin X L, Wang Y Z. Survey on few-shot learning[J]. Journal of Software,2021,32(2):349-369.

[13] Koch G, Zemel R, Salakhutdinov R. Siamese neural networks for one-shot image recognition[C]//32nd International Conference on Machine Learning (ICML) Deep Learning Workshop,2015, 2: 0.

[14] Vinyals O, Blundell C, Lillicrap T, et al. Matching networks for one shot learning[C] //30th International Conference on Neural Information Processing Systems. 2016:3637-3645.

[15] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning[C]//31th International Conference on Neural Information Processing Systems,2017:4077-4087.

[16] Sung F, Yang Y, Zhang L, et al. Learning to compare: relation network for few-shot learning[C]//IEEE Conference on Computer Vision and Pattern Recognition,2018: 1199-1208.

[17] Pennington J, Socher R, Manning C D. Glove: global vectors for word representation[C]//Conference on Empirical Methods in Natural Language Processing (EMNLP),2014: 1532-1543.

[18] Qi Song-zhe, Huang Xian-ying, Zhu Xiao-fei. Aspect-based-sentiment analysis model based on weight enhancement[J]. Journal of Chinese Computer Systems, 2022, 43(4): 747-753.

[19] Mo Yu, Xiaoxiao Guo, Jinfeng Yi, et al. Diverse few-shot text classification with multiple metrics.[C]//Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,2018:1206–1215.

[20] Wang Yang-gang, Qiu Xi-peng,Huang Xuan-xuan, et al. Few-shot text classification with dual channel graph neural network[J]. Journal of Chinese Information Processing, 2021, 35(7): 89-97,108.

[21] Han X, Zhu H, Yu P, et al. FewRel: a large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation[C]//Conference on Empirical Methods in Natural Language Processing (EMNLP),2018: 4803-4809.

[22] Lei Y, Tang K. Learning rates for stochastic gradient descent with nonconvex objectives[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(12): 4505-4511.

[23] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks[C]// 34nd International Conference on Machine Learning (ICML),2017: 1126-1135.

[24] Satorras V G, Estrach J B. Few-shot learning with graph neural networks[C]//6th International Conference on Learning Representations(ICLR),2018.

[25] Mishra N, Rohaninejad M, Chen X, et al. A simple neural attentive meta-learner[C]//6th International Conference on Learning Representations(ICLR),2018.

[26] Liu Y, Lee J, Park M, et al. Learning to propagate labels: transductive propagation network for few-shot learning[C]//6th International Conference on Learning Representations(ICLR),2018.

[27] Munkhdalai T, Yu H. Meta networks[C]//34nd International Conference on Machine Learning (ICML),2017: 2554-2563.

[28] Geng R, Li B, Li Y, et al. Induction networks for few-shot text classification[C]//Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),2019: 3904-3913.

[29] Sun P, Ouyang Y, Zhang W, et al. MEDA: meta-learning with data augmentation for few-shot text classification[C]//30th International Joint Conference on Artificial Intelligence (IJCAI),2021: 3929-3935.

[30] Kenton J D M W C, Toutanova L K. Bert: pre-training of deep bidirectional transformers for language understanding[C]//Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT),2019: 4171-4186.

附中文参考文献：

[1] 赵海燕,曹杰,陈庆奎,等. 层次多标签文本分类方法[J]. 小型微型计算机系统, 2022, 43(4):673-683.

[8] 赵凯琳, 靳小龙, 王元卓. 小样本学习研究综述[J]. 软件学报, 2021,32(02):349-369.

[18] 齐嵩喆,黄贤英,朱小飞.基于权重增强的方面级情感分析模型[J].小型微型计算机系统, 2022, 43(4):747-753.

[20] 王阳刚, 邱锡鹏, 黄萱菁,等. 基于双通道图神经网络的小样本文本分类[J]. 中文信息学报, 2021, 35(7): 89-97+108.