文章编号：1003-0077（2017）00-0000-00

用于文本分类的均值原型网络

线岩团[1,2] 相艳[1,2] 余正涛[1,2] 文永华[1,2] 王红斌[1,2] 张亚飞[1,2]
（1. 昆明理工大学 信息工程与自动化学院，云南 昆明 650500；
2. 昆明理工大学 云南省人工智能重点实验室，云南 昆明 650500）

摘要：文本分类是自然语言处理的基本任务之一。该文在原型网络基础上，提出了按时序移动平均集成历史原型向量的均值原型网络，并将均值原型网络与循环神经网络相结合提出了一种新的文本分类模型。该模型利用单层循环神经网络学习文本的向量表示，通过均值原型网络学习文本类别的向量表示，并利用文本向量与原型向量的距离训练模型并预测文本类别。与已有的神经网络文本分类方法相比，模型在训练和预测过程中有效利用了样本间的特征相似关系，并具有网络深度浅，参数少的特点。提出的方法在多个公开的文本分类数据集上取得了最先进的分类准率。

关键词：文本分类；均值原型网络；自集成学习

中图分类号：TP391 **文献标识码：**A

Mean Prototypical Networks for Text Classification

XIAN Yantuan[1,2], XIANG Yan[1,2], YU Zhengtao[1,2],WEN Yonghua[1,2], WANG Hongbin[1,2], ZHANG Yafie[1,2]
(1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan, 650500, China;
2. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming, Yunnan, 650500, China)

Abstract: Text classification is a fundamental problems of natural language processing. Based on the prototypical networks, this paper proposes a mean prototype network by ensemble different time steps prototype vectors through moving average, and combines the mean prototype network with a simple RNN neural network to propose a novel text classification model. The model uses a single-layer RNN neural network to learn the vector representation of text, and learns categories vector representation by the mean prototype networks. The model uses the distance between the text vector and the prototype vector to train the model and predict the text category. Compared with the existing neural network text classification method, the model has the characteristics of shallow depth and fewer parameters, and make use of similarity between samples in training and prediction process. The proposed method achieves state-of-the-art results on five benchmark datasets for text classification.

Key words: text classification; mean prototype networks；self-ensemble

0 引言

文本分类是自然语言处理的基本任务，在文本主题分类[1]、情感分析[2]和垃圾邮件检测[3]等领域具有广泛应用。传统机器学习文本分类方法在抽取词袋（bag-of-words，BoW）特征的基础上，通常采用朴素贝叶斯[4]或支持向量机（Support Vector Machine，SVM）分类器[5]进行分类，这类方法存在的主要不足是难以利用文本的词序信息。

近年来，神经网络模型在文本分类任务上取得了显著效果。与传统机器学习方法需要人工设计

收稿收稿日期：2019-06-15；**定稿日期**：2019-08-08
基金项目：国家重点研发计划（2018YFC0830105, 2018YFC0830100）；国家自然科学基金项目(61732005, 61672271, 61562052, 61762056)

特征抽取函数不同，神经网络模型可以从文本的词序列中自动学习分类特征。按网络类型不同，可大致分为卷积神经网络和循环神经网络文本分类模型。这两类模型在词嵌入的基础上，分别采用卷积神经网络和循环神经网络自动学习词、短语（片断）和序列的向量表示，并将该特征向量作为分类的特征。2014 年 Kim 提出了卷积神经网络句子分类模型[6]，取得了较好的分类效果。此后提出的 one-hot CNN 模型[7]、字符级深度卷积模型[8,9],都进一步提升了卷积神经网络文本分类模型的性能。以长短时记忆网络（Long Short Term Memory, LSTMs）为代表的循环神经网络文本分类模型，通常采用预训练词向量[22]初始化模型的嵌入层，取得了很好的效果[11,12]。Xiao 等人提出的融合卷积和循环神经网络的字符级文本分类模型，结合了卷积和循环神经网络的优点，取得了很好的分类性能[13]。Miyato 等人通过引入对抗和虚拟对抗训练方法提升了循环神经网络分类模型的性能[14,15]。与利用卷积和循环单元从词向量序列中学习序列表征不同，FastText[16]和片断嵌入模型[17]直接通过嵌入层学习词和短语的向量表示，并采用平均或求和的方式获得整个序列的向量表征，取得了先进的分类效果。

经过分析，我们发现已有神经网络文本分类模型的主要贡献集中在利用了不同网络结构更好地学习文本的特征表示。而具体分类过程都是在特征向量的基础上采用 Softmax 计算文本的类别概率。以上模型的学习过程假设样本间相互独立，特征学习过程中，算法依靠独立的分类标签优化模型参数，未充分利用样本和特征间的关系，不利于模型学习相同类别的共有特征，区分不同类别的差异特征。在分类过程仅依靠文本自身的特征向量计算类别概率，而未考虑文本与其它样本的相似性，也可能有损模型的泛化能力。

从直观上来说，具有相同类别的文本通常在主题、情感和语义上具有共性的特征，而经过神经网络编码得到的特征向量也应具有某种相似性。为了在模型学习过程强化同类样本具有相性的表征，并在学习和预测过程中融入样本的相似性关系，本文借鉴原型网络（Prototypical Networks）[18]思想提出了用于文本分类的均值原型网络模型。

均值原型网络文本分类模型通过神经网络自动学习文本向量表示，将所有文本映到相同向量空间，并使得相同类别的文本具有相似的表示。对于每个类别，模型采用文本特征向量的平均作为该类别的原型向量。预测时文本特征向量与原型向量距离越近则属于该类别的概率越高，反之属于该类别的概率就越低。与原型网络的小样本学习任务不同，文本分类预测时没有参照样本，所以本文提出的均值原型网络将每个类别的原型向量作为模型的参数，类别的原型向量由时序移动平均集成不同时间步的原型向量获得。在此基础上，本文结合循环神经网络构建了新颖的文本分类模型。

本文主要贡献体现在以下三点：

（1）提出了均值原型网络模型，通过时序移动平均集成不同时间步的原型向量，获得能更好表征各类别样本的均值原型向量。

（2）将提出的均值原型网络与循环神经网络结合提出了一种新的文本分类模型。该模型有效利用样本间的关系学习文本的向量表示，并基于文本向量与原型向量的距离实现分类，具有更好的可解释性。相比已有神经网络文本分类模型，本文本提出的方法模型结构简单，参数和超参少。

（3）文本提出方法在多个公开的文本分类数据集上取得了最先进的结果。

1 相关工作

神经网络文本分类模型自动学习特征来分类文本。已有模型通过嵌入、卷积、循环或两者相结合的方式，从文本的词或字符序列中学习用于分类的向量表示[6,10,13,17,18]。与已有的方法类似，文本采用循环神经网络作为编码器学习文本的向量表示。但与通常采用的多层双向 LSTMs 或 GRU 结合池化层学习向量表示不同。本文仅采用单层 GRU（Gated Recurrent Unit）层和 Max-pooling 层获得序列的文本的向量表示，网络结构更加简单。

原型网络在小样本学习（Few-shot Learning）图像分类任务中取得了很好的效果[18]。原型网络模型通过神经网络学习图像的向量表示。在训练与预测时模型通过输入的支撑集（Support Set）向量的平均计算类别的原型向量，然后根据查询集（Query Set）样本特征向量与原型向量距离预测样本分类。与原型网络的小样本学习任务不同，文本分类预测时不输入支撑集，无法根据输入的样本计算原型向量。本文提出的均值原型网络可以作是已有原型网络模型的变种。本文提出的均值原型网络将每个类别的原型向量作为模型的参数，在预测时无需输入支撑集。

本文采用时序移动平均方式计算类别的均值原型向量，其过程可看作是多个历史模型的集成。这一思想与多种自集成模型类似，均值教师模型

（Mean Teachers）通过集成不同时间步学生模型参数获得教师模型，并利用教师-学生网络实现半监督学习[24]。时序集成模型将预测的代理标签视为模型过去预测值的移动平均值[25]。

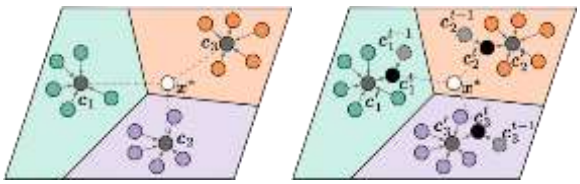
2 本文提出的方法

本节介绍均值原型网络原理和计算过程，以及结合单层循环神经网络设计的文本分类模型。

2.1 均值原型网络

原型网络利用支撑集 S 为每个类别计算原型向量 $\mathbf{c}_k$，然后根据查询集样本 x^* 与原型向量的距离来对样本进行分类，在小样本图像分类任务中取得了很好的效果。原型网络原理参见图 1a 的示意图。

原型网络小样本分类任务中，训练与预测过程都依赖支撑集 S 来获得类别的原型向量。而在文本分类任务的预测阶段不存在支撑集，模型仅根据输入的文本预测其类别。为了解决这一问题，本文将类别的原型向量设计为模型参数，在预测时不再需要支撑集来计算原型向量。预测时直接计算文本特征向量与保存的原型向量的距离，并据此获得文本的类别。



a. 原型网络 b. 均值原型网络

图 1. 原型网络与均值原型网络示意

为了获得合理的类别原型向量，本文采用时序移动平均方式集成历史时间步的原型向量获得时间 t 的均值原型向量 $\mathbf{c}_k^t \in \mathbb{R}^m$，

$$\mathbf{c}_k^t = a\mathbf{c}_k^{t-1} + (1-a)\mathbf{c}_k^{'} \tag{1}$$

其中，m 为原型向量的维度，$\mathbf{c}_k^{t-1}$ 为训练时间 $t-1$ 的均值原型向量，$\mathbf{c}_k^{'}$ 为当前时间 t 对应的支撑集计算得到的原型向量，a 是均值原型向量的衰减系数。

与原型网络一样，原型向量 $\mathbf{c}_k^{'}$ 根据支撑集计算得到。设支撑集为 $S=\{(\mathbf{x}_1,y_1),\cdots,(\mathbf{x}_N,y_N)\}$，其中样本 $\mathbf{x}_i$ 对应的分类标记作 $y_i \in \{1,\cdots,K\}$，K 为类别数量。计算类别 k 原型向量 $\mathbf{c}_k^{'}$ 时，样本由编码函数 f_f 计算得到特征向量 $f_f(\mathbf{x}_i) \in \mathbb{R}^m$，$f$ 是编码函数的参数。本文采用 GRU 神经网络作为文本的编码函数，将在下一节具体介绍。原型向量 $\mathbf{c}_k^{'}$ 由对应类别支撑集 S_k 特征向量的平均得到，

$$\mathbf{c}_k^{'} = \frac{\sum_i f_\phi(\mathbf{x}_i) z_{i,k}}{\sum_i z_{i,k}},\ z_{i,k} = \mathbf{1}[y_i = k] \tag{2}$$

其中，$z_{i,k}$ 是类别 k 的示性函数，用于选择属于类别 k 的样本。

均值原型可看作是样本类别的预测器，样本 $\mathbf{x}^*$ 属于类别 k 的概率可由特征向量 $f_f(\mathbf{x}^*)$ 到均值原型向量 $\mathbf{c}_k^t$ 的欧式距离确定，

$$p(y=k \mid \mathbf{x}^*) = \frac{\exp(-\left\| f_f(\mathbf{x}^*) - \mathbf{c}_k^t \right\|_2^2)}{\sum_{k'} \exp(-\left\| f_f(\mathbf{x}^*) - \mathbf{c}_{k'}^t \right\|_2^2)} \tag{3}$$

其中，$k' \in \{1,\cdots,K\}$ 是类别标签。

模型针对一个训练批次的训练损失 $J(f)$ 定义为各类别查询样本到对应类别概率的负对数的均值：

$$\begin{aligned} J(f) = & -\frac{1}{K}\sum_i \log p(y_i^* \mid x_i^*, \{\mathbf{c}_k^t\}) \\ & + \frac{1}{K^2 - K}\sum_{k \neq k'} \exp(-\left\| \mathbf{c}_k^t - \mathbf{c}_{k'}^t \right\|_2^2) \end{aligned} \tag{4}$$

公式（4）中的第二项是正则项，用于惩罚相近的原型向量，使原型向量相互远离，提高模型的泛化能力。模型通过梯度下降算法最小化损失函数 $J(f)$ 实现。训练过程中 $\mathbf{c}_k^t$ 按公式（1）更新，梯度下降算法不更新 $\mathbf{c}_k^t$。

与原型网络不同，本文方法将均值原型向量 $\mathbf{c}_k$ 作为模型的参数，预测时直接计算测试样本 $\mathbf{x}^*$ 与原型向量 $\mathbf{c}_k$ 的距离，并将距离最近的类别确定预测类别 $\hat{y}$，

$$\hat{y} = \arg\max_k p(k \mid x^*, \{\mathbf{c}_k\}) \tag{5}$$

2.2 GRU 序列编码器

均值原型网通过可学习的编码函数 f_f 将样本映射为与均值原型向量维度一致的特征向量。针对文本分类任务，本文采用一个简单的单层 GRU 循环神经网络作为编码器学习文本的向量表示。编码器网络包括：嵌入层、单向 GRU 层、max-pooling 层和全联接层。

嵌入层将输入的词序列 $\mathbf{x}=(w_1,w_2,\cdots,w_T)$ 映

射为低维稠密的词向量序列 $(\mathbf{v}_1,\mathbf{v}_2,\cdots,\mathbf{v}_T)$，$\mathbf{v}_j \in \mathbb{R}^d$。嵌入层参数为 $W_E \in \mathbb{R}^{|V| d}$，其中 $|V|$ 为词汇表的大小。

本文通过一层单向 GRU 层对词向量序列进一步编码，获得词在序列中的隐含表示 $\mathbf{h}_j \in \mathbb{R}^n$：

$$\mathbf{h}_j = \mathrm{GRU}(\mathbf{h}_{j-1}, \mathbf{v}_j) \tag{6}$$

Max-pooling 层按向量维度抽取序列时间步的最大值获取输入序列最有用的分类特征表示 $\mathbf{h} \in \mathbb{R}^n$，

$$h^l = \max h_j^l,\ j \in [1,T], \forall l \in \{1,\cdots,n\} \tag{7}$$

全联接层将序列的特征向量 $\mathbf{h}$ 映射为最与类别原型向量维度一致的向量表示 $\mathbf{d} \in \mathbb{R}^m$，

$$\mathbf{d} = \tanh(\mathbf{W}_d \mathbf{h} + \mathbf{b}_d) \tag{8}$$

其中，$\mathbf{W}_d$ 和 $\mathbf{b}_d$ 为分别为权重矩阵和偏置，tanh 激活函数将输出值限定为 $(-1,1)$。

表 1 数据集统计

任务	数据集	类别数	平均长度	训练样本数	测试样本数	词汇表大小
情感分析	Yelp Review Polarity	2	156	560,000	38,000	80424
	Yelp Review Full	5	158	650,000	50,000	86482
	Amazon Review Polarity	2	91	3,000,000	650,000	244266
	Amazon Review Full	5	93	3,600,000	400,000	222745
新闻分类	AG's News	4	44	120,000	7,600	31802
	Sogou News	5	579	450,000	60,000	45460
问句分类	Yahoo! Answers	10	112	1,400,000	60,000	216201
本体抽取	DBPedia	14	55	560,000	70,000	138941

本文提出模型的参数为 $\Phi = \{\mathbf{c}, \mathbf{W}_E, \mathbf{W}_{\mathrm{GRU}}, \mathbf{W}_d, \mathbf{b}_d\}$，其中 $\mathbf{c} = \{\mathbf{c}_1,\cdots,\mathbf{c}_K\}$ 为均值原型向量，与深层卷积网络模型[9]参数更少；相比通过嵌入层获得短语表征的二元 FastText 模型[16]和片断嵌入模型[17]，本文模型的参数也更少。

3 实验

本节介绍实验采用的数据集、模型实现细节，介绍提出方法与基准模型的实验结果对比。

3.1 数据集

本文采用文献[19]的数据集来验证提出方法的有效性。数据集分为情感分析、新闻分类、问句分类和本体抽取 4 类任务。各数据集的训练样本数、测试样本数、类别数以及文本的平均词长度统计参见表 1。为了很更好地和基准方法作对比，本文采用了与基准方法一致的准确率作为评价指标。

3.2 基准模型

本文对比了多种典型的有监督文本分类模型，包括 Zhang 等人提出的 n 元 TFIDF 模型(ngrams TFIDF)[19]，字符级的卷积神经网络模型(char-CNN)[13]，字符级的卷积循环神经网络模型(char-CRNN)[8]，深层卷积模型(VDCNN)[9]，判别 LSTM 模型(D-LSTM)[20]；二元 FastText 模型(bigram FastText)[16]。片断嵌入模型（W.C.region.emb）[17]。表 2 中基准模型的实验结果来源于对应文献。

3.3 实现细节

本文采用与文献[6]一样的预处理方法，在过滤特殊字符后将所有单词转为小写。文档频次为 1 的词被视作未登录词。本文随机从训练数据中抽取 10%的样本作为验证集，按验证集损失最小原则选择最优模型。

模型采用 PyTorch 框架[21]实现，并在 Nvidia Tesla P100 GPU 上完成模型的训练和测试。除非特别说明，所有实验数据集采用以下相同的超参数设置，嵌入层维度为 100，GRU 层隐状态维度为 384，原型向量维度为 64，嵌入层和 GRU 层的 dropout 分别设为 0.5 和 0.1。

本文对比了随机初始化或预训练词向量对模型的影响。在使用预训练词向量时，除了 Sogou News 和 Amazon 数据集外，本文采用 Wikipedia 2014 和 Gigaword 5 数据预训练的 glove 词向量[1]初始化模型的嵌入层参数。针对 Sogou News 和 Amazon 数据集，本文利用其训练集通过 Glove 算法[22]训练得到词向量，作为嵌入层词向量的初始值。

[1] glove.6B.zip, https://nlp.stanford.edu/projects/glove

模型采用小批次方法训练，每批数据按类别随机采样64个样本,其中32个样本作为支撑集，余下的32个样本作为查询集。在训练过程中，除了Sogou News数据集最大序列长度限制为2000外，其余数据集使用整个序列进行训练和预测，不做截断处理。本文采用Adam梯度下降算法[23]训练模型，初始学习速率设为10^{-3}，$b_1=0.9$，$b_2=0.99$，$e_{adam}=10^{-8}$。均值原型向量的衰减系数设为$a=0.99$。论文提出方法的实验代码可通过互联网下载[2]。

3.4 实验结果

表2展示模型在各个数据集上的测试样本的分类准确率，其中Mean-Proto为本文方法，+GloVe表示用预训练词向量初始化嵌入层参数。实验取三次运行结果的平均作为最终准确率。从实验结果可以看出，预训练词向量可以在一定程度提升模型性能，其原因可能是预训练词向量可以为模型提供更好的初始值。通过观察实验过程发现，加入预训练词向量后，模型可以更快的收敛，收敛过程也更加平稳。

与基准系统相比，本文提出的方法在6个数据集上都获得了更优的分类准确率。模型性能在多个数据集上准确率提升显著，特别是在Yelp Review Full和Yahoo! Answers数据集上获得了1.2%的性能提升。在Amazon Review Full和Sogou数据集上，本文提出方法的准确率也分别接近于性能最好的两个基准系统（VDCNN和W.C.region.emb）。

[2] https://github.com/xianyt/mean_prototype

表 2 实验结果

模型	Yelp P.	Yelp F.	Amz. P.	Amz. F.	AG	Sogou	Yah. A.	DBP
ngrams TFIDF	95.4	54.8	91.5	52.4	92.4	97.2	68.5	98.7
char-CRNN	94.5	61.8	94.1	59.2	91.4	95.2	71.7	98.6
bigram-FastText	95.7	63.9	94.6	60.2	92.5	96.8	72.3	98.6
VDCNN	95.7	64.7	95.7	63.0	91.3	96.8	73.4	98.7
D-LSTM	92.6	59.6	-	-	92.1	94.9	73.7	98.7
W.C.region.emb	96.4	64.9	95.1	60.9	92.8	97.6	73.7	98.9
Mean-Proto	96.6	65.5	94.9	60.4	93.0	97.0	74.3	98.8
Mean-Proto+GloVe	**96.8**	**66.1**	**95.8**	62.5	**93.2**	97.5	**74.9**	**99.0**

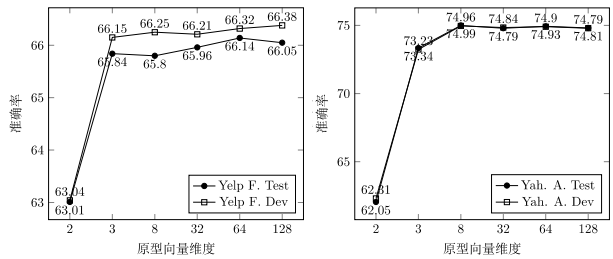
4 实验分析

本节在默认超参设置下，分析了不同原型向量维度、GRU 隐状态维度、原型向量衰减系数、词嵌入维度,以及损失正则项对模型性能的影响。实验取三次运行的平均值作为最终结果。

4.1 原型维度的影响

在本文提出的方法中，均值原型向量是样本分类的重要依据。本文以 Yelp Review Full 和 Yahoo! Answers 数据集为例，对比了不同原型向量维度对模型性能的影响。

从图 2 的结果可以出，在原型向量维度小于等于 3 时，模型对两个数据集的性能具有明显差异,模型对 Yahoo! Answers 数据集的性能很差；而在原型向量维度大于 3 后，模型在上验证集和测试集上的性能平稳。出现这种现象的主要原因可能是两个数据集的类别数差异较大，而过小的原型向量维度限制了其表征能力，在类别数较多时难以区分多个类别。而原型向量维度大于 3 后，其表征能力迅速增强，所以对模型性能影响也迅速减弱。



a. Yelp F. b. Yah. A.

图 2.原型向量维度的影响

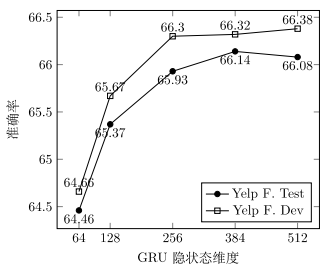


图 3.GRU 层隐状态维度的影响

4.2 GRU 层隐状态维度的影响

在固定其它超参数不变的前提下，本文对比了 GRU 层隐状态维度对模型性能的影响。从图 3 的实验结果可看出随着维度增大，模型的性能稳步提升，但维度大于 384 后模型对测试集的性能略有下降。但总体来说模型对 GRU 层隐状态维度对模型性能的影响不明显，在维度大于 256 后，模型对 Yelp Review Full 测试集的准确率差异在 0.2%以内。

4.4 原型衰减系数的影响

本文对比了不同原型衰减率对模型的影响，图 4 的实验结果表明，衰减率模型性能的影响不明显，模型对 Yelp Review Full 验证集和测试集的准确率基本稳定在 66.3%和 66.0%。

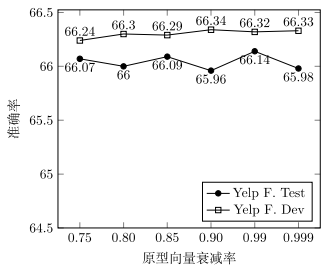


图 4.原型衰减系数的影响

4.3 词嵌入维度的影响

在 GRU 层和原型向量维为默认值的条件下，本文对比了不同词嵌入维度下，模型对 Yelp Review Full 验证集和测试集的准确率。从图 5 的实验结果可以出,随着词嵌入层维度从 50 增大到 300，模型对验证集的性能略有下降，在 100 维时测试集获得最优性能。

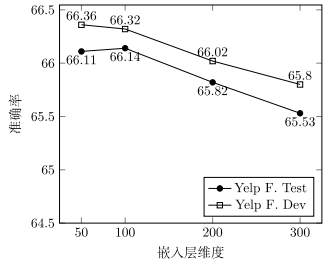
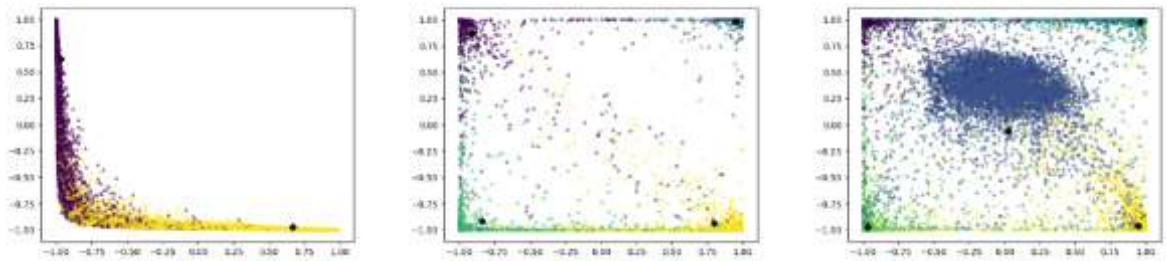


图 5.词嵌入维度的影响

总体来说，在嵌入层维度大小对模型性能的影响不明显，模型对验证集和测试集的性能相对平稳，最高与最低准确率的差异小于 0.7%。嵌入层在 50-300 维条件下，本文提出的方法在 Yelp Review Full 数据集上的准确率都超过了基准模型。

4.5 正则项的影响

实验结果表明损失函数中的正则项有助于提高模型的泛化能力。去掉正则项后，模型在 Yelp Review Full 数据集的实验结果上有 0.2%-0.3% 的性能损失。



(a) Yelp P., Acc:96.3% (b) AG, Acc: 92.6% (c)Sogou News 96.4%

图 6 分类结果可视化

4.5 可视化

本节通过文本特征向量和原型向量的可视化展示模型的编码和分类效果，验证模型对文本的分类性能。为了方便可视化展示，本文在保持其它超参不变的前提下，将原型向量的维度指定为 2 维,并用彩色散点图展示样本特征向量和原型向量特征在特征空间中的分布情况。

图 6 分别展示了 Yelp Review Polarity,AG's News 和 Sogou News 数据集测试样本特征向量与类别均值原型向量在二维空间的分布情况，模型对这两个数据集的分类准确率分别为 96.3%、92.6%和 96.4%。从图 5 可以看出各类别的测试样本体现出的明显的分类效果，不同类别的样本集中分布在对应的原型向量周围。各类别的均值原型向量分相互远离,分散于二维空间的不同位置，加强了样本的分类效果,提高了模型的泛化能力。图 6 中数据集的可视化结果体现出了直观清晰的分类效果，其均值原型向量分别位于二维空间的四个角及中心，而对应类别的样本集中分布于均值原型向量周围。

本文通过 max-pooling 层的输出来展现词在文本特征向量中的重要程度，一个词被 max-pooling 层选中的次数越多，则表明其对分类的贡献越大。表 3 展示了 AG's News 数据集测试样列的 Max-pooling 可视化结果。从图 7 中可以看出,本文提出的模型通过单层 GRU 和 Max-pooling 层可以较好地关注与分类相关的词。例如,在“Sports”样例中模型关注了“olympic”、“gold”“medal”等词，在“Sci/Tech”样例中则关注了“phone”、“driver”、“intel”和“technology”等词，表明模型关注词与文本的分类具有较好的一致性。

图 7 AG 数据集 Max-pooling 可视化

类别	样列
World	ivory coast 's army bombs rebel towns (reuters) reuters government warplanes and helicopter gunships pounded rebel held towns in northern ivory coast for a second day on friday , fueling fears of a slide into all out war in the world 's top cocoa grower .
Sports	today in athens <unk> <unk> van <unk> of the netherlands wipes a tear after winning the gold medal in the women <num> s road cycling individual time trial at the vouliagmeni olympic centre in athens on wednesday .
Business	stocks seen flat as earnings pour in us stock futures pointed to a flat market open thursday as a rush of quarterly earnings reports painted a mixed picture for corporate profits amid lingering worries over the high price of oil .
Sci/Tech	will your next cell phone have a hard drive ? hitachi global storage technologies and intel are pushing the development of an interface technology that they hope will smooth the adoption of compact hard drives into mobile phones , pdas , and digital music players , the companies say .

以上的可视化展示从一个侧面表明本文采用的基于距离的分类方法，结合简单循环神经网络编码,能有效的获得文本分类特征,具有一定的可解释性。

5 结论

本文提出了按时序移动平均集成历史原型向量的均值原型网络，并结合简单的循环网络提出了一种新的文本分类模型。模型在多个公开文本分类数据集上取得了最优的分类结果。相比基准方法，本文提出的模型在训练和预测过程中有效利用的样本间的特征相似关系，使模型在结构更简单、参数更少的前提下获得更好的性能。

参考文献

[1] Lewis D, Yang Y, Rose T G, et al. RCV1: A New Benchmark Collection for Text Categorization Research[J]. Journal of Machine Learning Research, 2004: 361-397.

[2] Pang B, Lee L. Opinion Mining and Sentiment Analysis[J]. Foundations and Trends in Information Retrieval, 2008, 2(1): 1-135.

[3] Sahami M, Dumais S, Heckerman D, et al. A Bayesian approach to filtering junk e-mail[C]//Learning for Text Categorization: Papers from the 1998 workshop. 1998, 62: 98-105.

[4] McCallum A, Nigam K. A comparison of event models for naive bayes text classification[C]//AAAI-98 workshop on learning for text categorization. 1998, 752(1): 41-48.

[5] Joachims T. Text categorization with support vector machines: Learning with many relevant features[C]//European conference on machine learning. Springer, Berlin, Heidelberg, 1998: 137-142.

[6] Kim Y. Convolutional Neural Networks for Sentence Classification[C]//Empirical methods in natural language processing, 2014: 1746-1751.

[7] Johnson R, Zhang T. Effective Use of Word Order for Text Categorization with Convolutional Neural Networks[C]//North American chapter of the association for computational linguistics, 2015: 103-112.

[8] Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification[C]//Advances in neural information processing systems. 2015: 649-657.

[9] Conneau A, Schwenk H, Barrault L, et al. Very Deep Convolutional Networks for Natural Language Processing[J]. arXiv: Computation and Language, 2016.

[10] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119.

[11] Socher R, Perelygin A, Wu J, et al. Recursive deep models for semantic compositionality over a sentiment treebank[C]//Proceedings of the 2013 conference on empirical methods in natural language processing. 2013: 1631-1642.

[12] Dai A M, Le Q V. Semi-supervised sequence learning[C]//Advances in neural information processing systems. 2015: 3079-3087.

[13] Xiao Y, Cho K. Efficient character-level document classification by combining convolution and recurrent layers[J]. arXiv preprint arXiv:1602.00367, 2016.

[14] Miyato T, Dai A M, Goodfellow I J, et al. Adversarial Training Methods for Semi-Supervised Text Classification[C]//International conference on learning representations, 2017.

[15] Miyato T, Maeda S, Ishii S, et al. Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018: 1-1.

[16] Joulin A, Grave E, Bojanowski P, et al. Bag of Tricks for Efficient Text Classification[C]//Conference of the European chapter of the association for computational linguistics, 2017: 427-431.

[17] Qiao C, Huang B, Niu G, et al. A New Method of Region Embedding for Text Classification [C]//International Conference on Learning Representations. 2018.

[18] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning[C]//Advances in Neural Information Processing Systems. 2017: 4077-4087.

[19] Zhang X, Lecun Y. Text Understanding from Scratch[J]. arXiv: Learning, 2015.

[20] Yogatama D, Dyer C, Ling W, et al. Generative and discriminative text classification with recurrent neural networks[J]. arXiv preprint arXiv:1703.01898, 2017.

[21] Paszke A, Gross S, Chintala S, et al. Automatic differentiation in PyTorch.[C]//NIPS Autodiff Workshop, 2017.

[22] Pennington J, Socher R, Manning C D, et al. Glove: Global Vectors for Word Representation[C]//Empirical methods in natural language processing, 2014: 1532-1543.

[23] Kingma D P, Ba J. Adam: A Method for Stochastic Optimization[C]//International conference on learning representations, 2015.

[24] Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results[C]//Advances in neural information processing systems. 2017: 1195-1204.

[25] Laine S, Aila T. Temporal Ensembling for Semi-Supervised Learning[C]// international conference on learning representations, 2017.

线岩团（1981—），博士研究生，讲师，主要研究领域为自然语方处理、信息抽取、机器翻译。
E-mail：xianyantuan@qq.com

相艳（1979—），博士研究生，讲师，主要研究领域为自然语方处理、信息抽取、机器翻译。
E-mail：50691012@qq.com

余正涛（1970—），通讯作者，博士，教授，主要研究领域为自然语言处理、信息检索、机器翻译、机器学习。
E-mail：ztyu@hotmail.com