控制与决策
Control and Decision

一种去注意力机制的动态多层语义感知机

刘孝炎, 唐焕玲, 王育林, 窦全胜, 鲁明羽

引用本文:
刘孝炎,唐焕玲,王育林,窦全胜,鲁明羽. 一种去注意力机制的动态多层语义感知机[J]. 控制与决策, 2024, 39(2): 588−594.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.0496

一种去注意力机制的动态多层语义感知机

刘孝炎[1]，唐焕玲[1,2,3†]，王育林[1]，窦全胜[1,2,3]，鲁明羽[4]

(1. 山东工商学院 计算机科学与技术学院，山东 烟台 264005；2. 山东省高等学校协同创新中心: 未来智能计算，山东 烟台 264005；3. 山东工商学院 山东省高校智能信息处理重点实验室，山东 烟台 264005；4. 大连海事大学 信息科学技术学院，辽宁 大连 116026)

摘 要: Transformer在大规模数据集上取得了优异效果,但由于使用多头注意力使得模型过于复杂,且在小规模数据集上效果不理想. 对于多头注意力替换的研究在图像处理领域已取得一些成果,但在自然语言处理领域还少有研究. 为此,首先提出一种去注意力的多层语义感知机(multi-layer semantics perceptron, MSP)方法,其核心创新是使用token序列转换函数替换编码器中的多头注意力,降低模型复杂度,获得更好的语义表达;然后,提出一种动态深度控制框架(dynamic depth control framework, DDCF),优化模型深度,降低模型复杂度;最后,在MSP方法和DDCF的基础上,提出动态多层语义感知机(dynamic multi-layer semantics perceptron, DMSP)模型,在多种文本数据集上的对比实验结果表明,DMSP既能提升模型分类精度,又能有效降低模型复杂度,与Transformer比较,在模型深度相同的情况下,DMSP模型分类精度大幅提升,同时模型的参数量大幅降低.

关键词: 特征表示；语义感知机；动态深度控制；Transformer；文本分类

中图分类号: TP181 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.0496 **开放科学(资源服务)标识码(OSID):**

引用格式: 刘孝炎,唐焕玲,王育林,等. 一种去注意力机制的动态多层语义感知机[J]. 控制与决策, 2024, 39(2): 588-594.

A dynamic multi-layer semantics perceptron without attention mechanism

LIU Xiao-yan[1], *TANG Huan-ling*[1,2,3†], *WANG Yu-lin*[1], *DOU Quan-sheng*[1,2,3], *LU Ming-yu*[4]

(1. School of Computer Science and Technology，Shandong Technology and Business University，Yantai 264005，China；2. Co-Innovation Center of Shandong Colleges and Universities: Future Intelligent Computing，Yantai 264005，China；3. Key Laboratory of Intelligent Information Processing in Universities of Shandong，Shandong Technology and Business University，Yantai 264005，China；4. Information Science and Technology College，Dalian Maritime University，Dalian 116026, China)

Abstract: Transformer has achieved excellent results on large-scale data sets, but it is too complex due to utilizing Multi Head Attention (MHA), and its performance is poor on small-scale data sets. The study on the replacement of MHA is little in the field of natural language processing, although it has made great achievements in the field of image processing. Therefore, a method called multi-layer semantics perceptron (MSP) is proposed. Its major innovation is that instead of MHA, a simple token sequence transformation function is used, thus achieving a better semantic feature representation with lower complexity. Additionally, a dynamic depth control framework (DDCF) is proposed, which is able to optimize the depth of neural networks automatically, as a result the complexity of the model is reduced markedly. Finally, based on the MSP and the DDCF, the dynamic multi-layer semantics perceptron model (DMSP) is proposed. Compared with the Transformer model with same depth, the experimental results on multi-data sets show that the DMSP model achieves better performance significantly, meanwhile, its parameters declines sharply.

Keywords: feature representation；semantics perceptron；dynamic depth control；Transformer；text categorization

0 引 言

使用多头注意力(multi headed attention, MHA)的模型, 如Transformer[1]、BERT (bidirectional encoder representation from Transformers)[2-4]等,广泛应用于文本分类[5-6]、机器翻译[7-9]、命名实体识别[10-12]、图像分类[13-14]等各种领域. 然而,使用MHA的

收稿日期: 2022-03-28；**录用日期:** 2022-11-02.

基金项目: 国家自然科学基金项目(61976124, 61976125, 62176140).

责任编委: 胡清华.

†通讯作者. E-mail: thL01@163.com.

*本文附带电子附录文件,可登录本刊官网该文“资源附件”区自行下载阅览.

Transformer模型缺点同样明显: 1) 依赖大量样本, 训练样本不足时, 特征学习效果较差; 2) 模型复杂, 参数量庞大. 许多研究表明了MHA的可替代性[15-18], 但是存在去除MHA后导致模型性能下降等问题. 目前, 关于MHA的替代研究是一个尚未解决的难点问题.

在训练样本不足时, 提高Transformer模型特征提取能力的主要方法是数据增强(data augmentation). 其中, 生成式对抗网络(generative adversarial network, GAN)[19-20]是数据增强的代表性方法, 但将GAN模型嵌入到特征学习中, 会导致Transformer模型更复杂. Word/Sentence Mixup[21]是数据增强的另一种代表性方法, 生成的伪标签可能不属于任何已知类. EDA (easy data augmentation)[22]方法, 如随机同义词替换等, 需要计算每个词的相似度, 导致算法时间复杂度提高. 然而, 上述方法或导致模型更加复杂, 或出现未知类别的标签, 增加了Transformer特征学习的困难性.

针对Transformer模型复杂、参数量庞大的问题, ResMLP[15]和MLP-Mixer[16]使用两种多层感知机(MLP)替代Transformer的MHA, 二者在图像分类上有较好的表现. 但用于文本分类时, 特别是训练样本不足时, 效果却很不理想. 虽然能够降低模型的参数量, 但是特征学习效果变差, 这是目前备受关注的尚未解决的新的难点问题.

鉴于此, 本文首先提出一种去注意力的多层语义感知机(multi-layer semantics perceptron, MSP), 使用token序列转换函数替换编码器中的多头注意力, 缓解过拟合. 然后, 提出一种动态深度控制框架(dynamic depth control framework, DDCF), 动态优化神经模型深度, 降低模型复杂度. 在MSP方法和DDCF框架基础上, 提出动态多层语义感知机(dynamic multi-layer semantics perceptron, DMSP), 通过多种文本数据集上的对比实验验证DMSP模型的精度和有效性.

1　多层语义感知机MSP模型

MSP模型的核心创新是利用token序列转换方法$\mathcal{F}_t$替换多头注意力机制, 提出一种新的编码器结构, 并称之为语义感知机(semantics perceptron, SP). 多个语义感知机SP叠加构成了去注意力机制的多层语义感知机MSP, 其结构如图1所示.

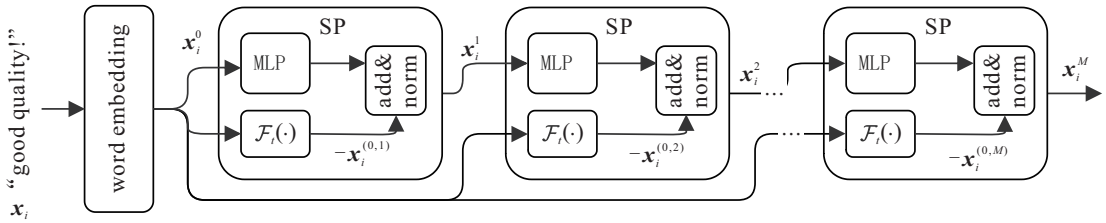

图1　多层语义感知机模型

令$\boldsymbol{x}_i$表示第i个样本, 经过词嵌入表示后得到$\boldsymbol{x}_i^0 \in \boldsymbol{R}^{m\times d}$作为第1层语义感知机的初始输入特征, m是样本$\boldsymbol{x}_i$包含的单词数, d是单词的嵌入维度. 将初始特征$\boldsymbol{x}_i^0$和前一层语义感知机输出$\boldsymbol{x}_i^{j-1}$输入第j层SP中, 学习消除因词序忽略引起的词序语义歧义, 其输出为$\boldsymbol{x}_i^j$.

1.1　token序列转换$\mathcal{F}_t$

在Transformer中, MHA的运算量大. 本文提出token序列转换$\mathcal{F}_t$方法, 该方法对token序列重排, 其时间复杂度为$\mathcal{O}(m)$, 空间复杂度为$\mathcal{O}(d)$. 给定初始token序列$\boldsymbol{x}_i^0 = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m)$, 在每层SP中, 利用$\mathcal{F}_t$对$\boldsymbol{x}_i^0$进行token序列转换, 即

$$\boldsymbol{x}_i^{(0,j)} = \mathcal{F}_t(\boldsymbol{x}_i^0) = (\boldsymbol{x}_{j1}, \boldsymbol{x}_{j2}, \ldots, \boldsymbol{x}_{jm}) \in \boldsymbol{A}_m^m. \quad (1)$$

其中: $\boldsymbol{A}_m^m$是长度为m的token序列$(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m)$的全排列集合; $\boldsymbol{x}_i^{(0,j)} = (\boldsymbol{x}_{j1}, \boldsymbol{x}_{j2}, \cdots, \boldsymbol{x}_{jm})$为$\boldsymbol{x}_i^0$的第$j$次随机排序, 即第$j$次token序列转换结果.

在训练样本不足时, token序列转换$\mathcal{F}_t$可以丰富样本的多种词序表征, 但这种词序可能是有歧义的, 将在后续的多层语义感知机中学习消除.

1.2　多层语义感知机MSP

如图1所示, MSP由语义感知机SP堆叠而成, 其中每一层语义感知机由token序列转换$\mathcal{F}_t$、特征学习MLP及特征融合层(add&norm)组成.

在图1中: $\boldsymbol{x}_i^1, \boldsymbol{x}_i^2, \ldots, \boldsymbol{x}_i^M$分别为样本$\boldsymbol{x}_i$经第1～$M$个SP学习后的特征表示, 其中第$j$层SP输出$\boldsymbol{x}_i^j$为

$$\boldsymbol{x}_i^j = \mathrm{norm}(-\mathcal{F}_t(\boldsymbol{x}_i^0) + \boldsymbol{x}_i^{j-1} + \mathcal{F}_m(\boldsymbol{x}_i^{j-1}, \boldsymbol{\theta}^j)). \quad (2)$$

其中: norm$(\cdot)$表示归一化, $\mathcal{F}_t(\boldsymbol{x}_i^0)$为序列转换结果, $\boldsymbol{x}_i^0$为初始token序列特征表示, $\boldsymbol{x}_i^{j-1}$为第$j-1$层语义感知机SP输出, $\mathcal{F}_m(\boldsymbol{x}_i^{j-1}, \boldsymbol{\theta}^j)$为特征$\boldsymbol{x}_i^{j-1}$经第$j$层SP中的MLP学习后的特征表示, $\boldsymbol{\theta}^j$为第j个MLP

的参数.

包含$\mathcal{F}_t$函数的SP能够感知token位置信息,并生成不同词序的token序列,但生成的token序列的对应语义可能存在歧义,故在式(2)中,在$\mathcal{F}_t(\boldsymbol{x}_i^0)$前加负号学习消减. 然后与上一层SP的输出$\boldsymbol{x}_i^{j-1}$、第$j$层的MLP输出$\mathcal{F}_m(\boldsymbol{x}_i^{j-1},\boldsymbol{\theta}^j)$相加、归一化,输出$\boldsymbol{x}_i^j$为第$j$层SP的学习的特征表示.

2 动态深度控制框架DDCF

DDCF框架结构如图2所示,由词嵌入层(word embedding)、加权特征学习器(weighted feature extractor, WFE)和动态深度控制器(dynamic depth controller, DDC)组成. WFE由特征学习器f^j与基分类器h_ω^j构成.

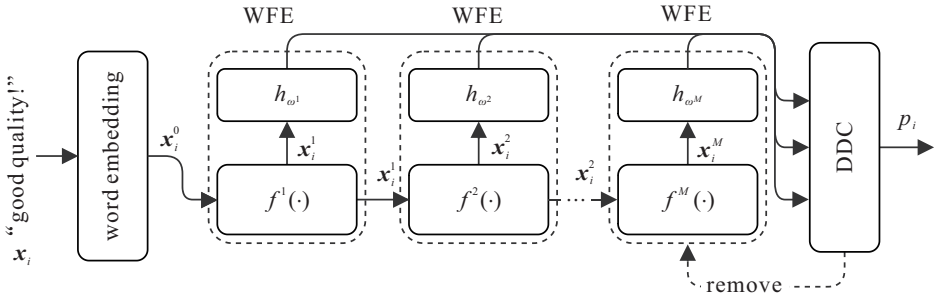

图2 动态深度控制DDCF框架

样本$\boldsymbol{x}_i$经嵌入层得到初始特征$\boldsymbol{x}_i^0$作为第1个WFE的输入,第j个WFE中f^j的输出作为第$j+1$个WFE的输入. 每个WFE使用f^j提取特征后,基分类器h_ω^j对样本$\boldsymbol{x}_i$进行预测,并将预测结果输入DDC. DDC对预测结果加权平均,并动态调整模型深度.

2.1 特征学习函数f^j

令$\boldsymbol{x}_i^j$记作第j个f^j的输出,即

$$\boldsymbol{x}_i^j = f^j(\boldsymbol{x}_i^{j-1}). \tag{3}$$

其中:$\boldsymbol{x}_i^{j-1}$为第$j-1$层f^{j-1}的特征学习结果,f^j为第j个加权特征学习器WFE中的特征学习函数.

2.2 基分类器h_ω^j

如图2的DDCF框架所示,每层WFE中,除了f^j,增加了基分类器h_ω^j,其分类正确率反映了每层特征学习的效果,并作为深度控制的决策依据.

对于样本$\boldsymbol{x}_i$,在第j层经过f^j学习后表示为$\boldsymbol{x}_i^j$,则第j个基分类器h_ω^j对$\boldsymbol{x}_i^j$的分类预测结果记为p_i^j,即

$$\boldsymbol{p}_i^j = h_{\omega^j}(\boldsymbol{x}_i^j). \tag{4}$$

将所有样本的预测结果$\boldsymbol{p}_i^j$输入给DDC,用于深度控制.

2.3 动态深度控制器DDC的深度优化策略

DDC对每层的基分类器进行监视,度量当前深度下模型的特征学习能力,动态调整每层基分类器h_ω^j的权重β^j,从而控制模型的最佳深度. 令$\xi\in(0,1)$为基分类器的正确率阈值,则基分类器的惩罚系数μ^j和权重β^j计算分别为

$$\mu^j=\begin{cases}1,\ \alpha^j>\xi;\\ -\infty,\ \alpha^j\leqslant\xi,\ g\geqslant s+1;\\ -1,\ \alpha^j\leqslant\xi,\ g<s+1.\end{cases} \tag{5}$$

$$\beta^j=\frac{\exp(\mu^j\alpha^j)}{\sum\limits_{i=0}^{\gamma}\exp(\alpha^i)}. \tag{6}$$

其中:α^j为第j层h_ω^j的正确率,γ为当前动态学习到的最佳模型深度,g为当前深度下分类正确率大于ξ的基分类器个数,s为分类正确率小于ξ的基分类器个数. DDC深度优化策略按以下3种情况处理:

1) $\alpha^j>\xi$时,$\mu^j=1$,不删除WFE.

2) $\alpha^j\leqslant\xi$且$g\geqslant s+1$时,令所有$\alpha^j\leqslant\xi$的基分类器的$\mu^j=-\infty$,由式(6)可知其权重β^j趋于0,DDC将删除第γ个WFE.

3) 当$\alpha^j\leqslant\xi$且$g<s+1$时,所有$\alpha^j\leqslant\xi$的基分类器$\mu^j=-1$,由式(6)可知其β^j会变小,DDC保留了本层WFE,但加大了对该层基分类器的惩罚.

综上,根据每层特征学习的效果和分类性能,DDC动态调整基分类器的惩罚系数μ^j和权重β^j,实现模型深度的动态优化. DDC对样本$\boldsymbol{x}_i$的分类预测结果进行集成,这与文献[23]思想类似,其结果作为$\boldsymbol{x}_i$的预测结果$\boldsymbol{p}_i$,计算如下:

$$\boldsymbol{p}_i=\sum_{j=1}^{\gamma}\beta^j\boldsymbol{p}_i^j. \tag{7}$$

其中:γ为当前动态学习到的最佳模型深度,β^j为第j个基分类器的权重,$\boldsymbol{p}_i^j$为第j个分类器对样本$\boldsymbol{x}_i$的预

测.

DDCF与其他模型的不同在于通过DDC自主优化模型深度,不需人为设定. 如ResNet[24]等模型,其精度随深度加深提高有限,但复杂度显著提高,甚至如VDCNN (very deep convolutional networks)[25]等模型,当深度大于某个值后,精度反而下降. 而DDC监测特征学习器提取能力下降时,会自主丢弃相应的WFE,从而降低复杂度,并提升分类性能.

3　动态多层语义感知机DMSP

本文在MSP的基础上,采用DDCF动态控制模型深度,提出动态多层语义感知机模型(dynamic multi-layer semantics perceptron, DMSP),其结构如图3所示,描述如算法1所示. 一方面DMSP利用MSP去除MHA,提高特征学习质量,降低时间和空间复杂度;另一方面,使用DDCF自主优化模型深度,再次降低模型复杂度,提升模型分类性能.

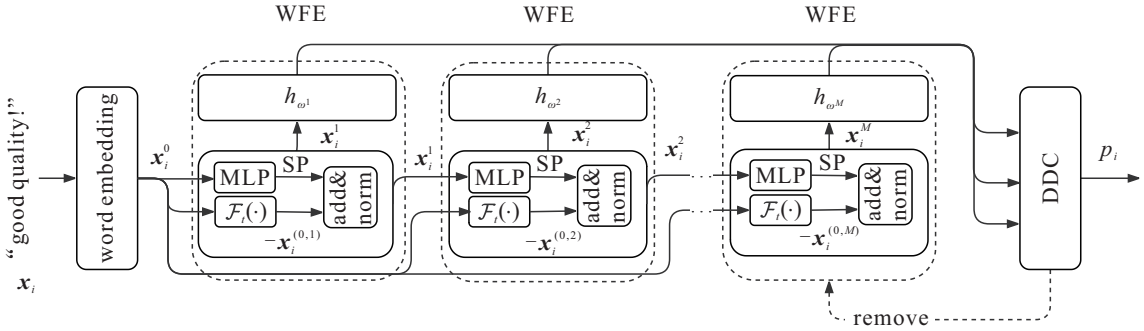

图3　DMSP模型

算法1　动态多层语义感知机DMSP算法

1) input: training set $D=\{(\boldsymbol{x}_i,y_i)\}_{i=0}^{n}$, $\boldsymbol{x}_i\in\boldsymbol{R}^{m\times d}$, $y_i=\{1,2,\ldots,C\}$, threshold ξ, the initial depth of the networks M, iterations R

2) initialization: $\gamma=M,\xi,R,\boldsymbol{\theta},\boldsymbol{\omega}$

3) for $i=1$ to n do

4)　embed sample $\boldsymbol{x}_i$ as $\boldsymbol{x}_i^0$

5) for $r=1$ to R do

6)　$g,s=0,0$

7)　for $j=1$ to γ do

8)　　for $i=1$ to n do

9)　　compute linear sequence transform $\mathcal{F}_t$ acc. to eq 1

10)　　train semantics perceptron $f^j(\boldsymbol{x}_i^{j-1},\boldsymbol{x}_i^0)$ acc. to eq 8

11)　　classify $\boldsymbol{x}_i$ acc. to eq 4

12)　　if $\alpha^j>\xi$ then $g=g+1$

13)　　else $s=s+1$

14)　　compute μ^j acc. to eq 5, β^j acc. to eq 6

15)　if $g>s+1$ then remove the last WFE by DDC, $\gamma=\gamma-1$

16)　get emsemble classifier acc. to eq 7

17)　optimize $\boldsymbol{\omega}^*,\boldsymbol{\theta}^*$ acc. to eq 9

18) output: the best $\boldsymbol{\omega}^*,\boldsymbol{\theta}^*,\boldsymbol{\beta},\gamma$

每层WFE的特征学习器f^j采用SP方法作为特征学习方法. 给定样本$\boldsymbol{x}_i$,经过第j层f^j学习,其输出$\boldsymbol{x}_i^j$计算式为

$$\boldsymbol{x}_i^j=f^j(\boldsymbol{x}_i^{j-1},\boldsymbol{x}_i^0)=\\ \text{norm}(-\mathcal{F}_t(\boldsymbol{x}_i^0)+\boldsymbol{x}_i^{j-1}+\mathcal{F}_m(\boldsymbol{x}_i^{j-1},\boldsymbol{\theta}^j)). \quad (8)$$

其中: f^j为第j层特征学习函数; $\boldsymbol{x}_i^{j-1}$为第$j-1$层语义感知机SP的输出; $\boldsymbol{x}_i^0$为初始特征; norm$(\cdot)$为归一化; $\mathcal{F}_t(\boldsymbol{x}_i^0)$为序列转换结果; $\mathcal{F}_m(\boldsymbol{x}_i^{j-1},\boldsymbol{\theta}^j)$为经第$j$层SP中的MLP学习后的特征表示, $\boldsymbol{\theta}^j$为第j个MLP的参数.

第j个层基分类器h_ω^j对$\boldsymbol{x}_i$的分类预测结果$\boldsymbol{p}_i^j=h_{\omega^j}(\boldsymbol{x}_i^j)$,其分类预测结果输入至DDC,用来自主控制模型深度. DDC根据设置的正确率阈值$\xi\in(0,1)$,按照式(5)和(6)调整基分类器的惩罚系数μ^j和权重β^j,即分3种情况动态自主控制模型深度,并按照式(7)对分类结果进行集成.

DMSP模型采用交叉熵损失函数,目标函数如下:

$$\boldsymbol{\omega}^*,\boldsymbol{\theta}^*=\\ \min_{\boldsymbol{\omega},\boldsymbol{\theta}}-\frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{C}y_i^{(k)}\log(h_{\omega^\gamma}(f^\gamma(\boldsymbol{x}_i^{\gamma-1},\boldsymbol{\theta}^\gamma))^{(k)}). \quad (9)$$

其中: n为训练样本数; C为类别数; $y_i^{(k)}$为样本$\boldsymbol{x}_i$的真实标签,且为第k类; $\boldsymbol{x}_i^\gamma$为样本$\boldsymbol{x}_i$在深度为γ的DMSP模型下的最终特征表示; γ为DMSP模型动态学习到的最佳模型深度; $h_{\omega^\gamma}(\boldsymbol{x}_i^\gamma)^{(k)}$为模型将样本$\boldsymbol{x}_i$预测为第$k$类的概率; $\boldsymbol{\omega},\boldsymbol{\theta}$分别为基分类器$h_\omega^\gamma$和特征学习器$f^\gamma$的模型参数. 采用Adam算法优化模型参数.

如算法1所示,将初始特征输入到线性序列转换$\mathcal{F}_t$中,按式(1)得到序列转换结果,然后按式(8)训练所有语义感知机,并以式(4)训练所有基分类器. 在每轮迭代后,计算各基分类器h_ω^j正确率α^j,并累计

正确率大于ξ和小于ξ的基分类器个数,分别记为g、s. 然后按式(5)和(6)计算基分类器惩罚系数μ^j和权重β^j,由DDC决策是否移除最后一层WFE. 接着按式(7)对分类预测集成,按式(9)优化参数$\boldsymbol{\omega}^j, \boldsymbol{\theta}^j$. 实验中使用Adam优化算法得到最优模型参数$\boldsymbol{\omega}^*, \boldsymbol{\theta}^*$,最优基分类器权重$\boldsymbol{\beta}$和最优模型深度$\gamma$. 在DDC控制下得到的最优模型的基分类器权重向量$\boldsymbol{\beta} = \{\beta^1, \beta^2, \ldots, \beta^\gamma\}$,其中$\beta^j$为第$j$个基分类器权重.

测试阶段,DMSP模型加载最优模型参数$\boldsymbol{\omega}^*, \boldsymbol{\theta}^*$,对于新样本$\hat{\boldsymbol{x}}_i$,根据式(1)进行词序转换得到$\hat{\boldsymbol{x}}_i^{(0,j)}$,按式(8)由每层语义感知机SP进行学习,输出$\hat{\boldsymbol{x}}_i^j$,并由相应基分类器$h_\omega^j$对$\hat{\boldsymbol{x}}_i^j$进行分类. 最后根据最优基分类器权重$\boldsymbol{\beta}$,按式(7)对各基分类器的分类结果加权平均,得到样本$\hat{\boldsymbol{x}}_i$的类别概率.

假设DMSP模型的深度为γ,分类阈值为ξ,一个WFE所需的参数量记为τ. 在第r轮迭代中,由γ个WFE组成的DMSP模型参数量记为$\psi_r = \gamma\tau$.

在第r轮迭代后,如果存在满足条件$\alpha^j \leqslant \xi$且$g \geqslant s+1$的基分类器,即属于DDC的第2)种情况,则按DDC控制删除最后一层WFE,那么第$r+1$轮DMSP模型的参数量及浮点运算次数均减少τ,即$\psi_{r+1} = \psi_r - \tau$.

如果处于DDC的其他两种情况,则有$\psi_{r+1} = \psi_r$.

DDC控制下的第r轮和第$r+1$轮的参数量和浮点运算次数比较如表1所示.

表 1 DDC控制下模型参数量及浮点运算次数对比

Eopch	DDC三种情况	param.	FLOPs
r	/	$\gamma\tau$	$\gamma\tau$
$r+1$	1) $\alpha^j > \xi$	$\gamma\tau$	$\gamma\tau$
	2) $\alpha^j \leqslant \xi, g \geqslant s+1$	$\psi_r - \tau$	$\psi_r - \tau$
	3) $\alpha^j \leqslant \xi, g < s+1$	$\gamma\tau$	$\gamma\tau$

在第r轮迭代中,DMSP模型的参数量为$\psi_r = \gamma\tau$. 第r轮迭代后,若满足DDC控制条件情况2),删除最后一层WFE,即模型深度$\gamma = \gamma - 1$,则第$r+1$轮时,参数量降为$\psi_{r+1} = \psi_r - \tau$. 对比第$r$轮,第$r+1$轮的参数量及浮点运算次数均减少$\tau$. 若满足DDC控制条件1)和3),深度$\gamma$不变,则参数量和浮点数计算量不变. 这表明DMSP模型可动态控制模型深度,进而降低参数量,加速模型训练.

4 实验结果及分析

选择AGNews (https://huggingface.co/datasets/ag_news), Amazon (https://huggingface.co/datasets/amazon_us_reviews), Sogou (https://huggingface.co/datasets/sogou_news), 20 newsgroups (https://huggingface.co/datasets/newsgroup)四个数据集. 统一对数据集中每个类别的数据打散后,随机抽取85 %作为训练数据,剩余组成测试集,简记为AGN.4、 Ama.4、 So.4和20 ngp.4. 如表2所示,每个类随机取1 000个样本,不足1 000取该类全部样本.

表 2 实验数据集及各类样本数描述

category	dataset			
	Ama.4	AGN.4	So.4	20 ngp.4
0	baby	world	科技	alt.ath.
1	beauty	sports	国际	soc.rel.chr.
2	home	business	国内	comp.gra.
3	sports	sci/Tec	财经	sci.med

为验证所提出方法的有效性,与下列方法进行对比实验,描述如下.

1) Transformer方法: 为文献[1]所提出的模型.

2) MLP方法: 使用MLP学习特征,最后一层接一个softmax分类器,称MLP模型.

3) SMSP方法: 使用MSP方法学习特征,最后一层接一个softmax分类器,没有采用DDFC优化深度,称静态多层语义感知机.

4) DC方法: 使用MLP方法学习特征,采用动态深度控制框架DDCF模型,称动态MLP分类模型.

5) SMSP-E方法: 使用MSP方法进行特征提取,每层接一个softmax分类器,最后集成分类,称静态集成多层语义感知机,未采用动态深度控制框架DDCF.

6) DMSP方法: 使用MSP方法和DDCF框架的动态多层语义感知机.

其中,方法3)～方法6)为本文提出的方法及变形,是为了作针对性的对比实验. 分类评价指标主要采用正确率(accuracy)、宏召回率(macro-recall)和宏平均-F1值(macro-F1),模型复杂度指标采用模型参数量(parameters)及浮点运算次数(FLOPs),下文分别简记为Acc.、Recall、F1、Param.和F..

4.1 动态深度控制框架DDCF的有效性验证

DC模型和MLP模型均使用MLP方法学习特征,但区别在于,DC模型使用了DDCF框架优化模型深度,而MLP模型没有. 实验中,DC模型自主优化深度,将其分类结果与MLP模型的最好分类结果对比,旨在验证DDCF方法的有效性,在4种数据集上的对比实验结果如表3所示.

从表3的数据可以看出,相比于MLP的最好情况,除了在So.4上DC模型的分类结果差一些,在其他3个数据集上,DC模型的分类结果明显优于MLP模型. 例如,在Ama.4上,初始深度为5的DC模型,经

表 3　不同深度下的DC与MLP模型的正确率和参数量比较

dataset	DC				MLP			
	Dep.	ACC./%	Recall/%	F1/%	Dep.	ACC./%	Recall/%	F1/%
Ama.4	5 → 4	**92.05**	**92.85**	**91.93**	4	88.50	88.38	88.44
AGN.4	5 → 4	**80.42**	**80.33**	**80.39**	3	77.25	77.31	77.21
So.4	5 → 3	81.45	81.36	81.39	1	**86.69**	**87.08**	**86.74**
20ngp.4	5 → 3	**82.11**	**81.18**	**82.25**	2	77.98	77.87	77.91

DDCF优化后,其最终深度为4,与MLP模型在该数据集上的最好分类结果对比,DC的accuracy、macro-recall、macro-F1值分别提高3.55%、4.47%、3.49%. 在AGN.4和20 ngp.4数据集上的比较也是类似的.

4.2　模型复杂度与分类性能的综合比较

本节将从模型复杂度和分类性能两方面,对各种模型进行综合评估. 实验中,各模型均取正确率最高的情况,结果如表4所示.

表 4　在Ama.4、AGN.4、So.4和20ngp.4数据集上各模型的复杂度及分类结果比较

dataset	model	complexity			Acc./%
		Dep.	Param.(M)	F.(G)	
Ama.4	Transformer	1 → 1	0.16	1.00	83.20
	MLP	4 → 4	0.27	1.73	88.50
	SMSP	10 → 10	0.38	2.41	**91.62**
	SMSP-E	5 → 5	0.37	2.36	87.91
	DC	5 → 4	0.30	1.89	**92.05**
	DMSP	5 → 4	0.30	1.89	**90.10**
AGN.4	Transformer	1 → 1	0.16	1.00	68.18
	MLP	3 → 3	0.14	0.89	77.25
	SMSP	12 → 12	0.46	2.88	**82.85**
	SMSP-E	3 → 3	0.22	1.42	79.03
	DC	5 → 4	0.30	1.89	**80.42**
	DMSP	5 → 3	0.22	1.42	**79.75**
So.4	Transformer	1 → 1	0.16	1.00	80.98
	MLP	1 → 1	0.15	0.94	86.69
	SMSP	12 → 12	0.46	2.88	**92.51**
	SMSP-E	3 → 3	0.22	1.42	**87.50**
	DC	5 → 3	0.22	1.42	81.45
	DMSP	5 → 5	0.22	1.44	**91.47**
20ngp.4	Transformer	1 → 1	0.16	1.00	73.09
	MLP	2 → 2	0.15	0.94	77.98
	SMSP	8 → 8	0.31	1.94	79.90
	SMSP-E	3 → 3	0.22	1.42	**83.89**
	DC	5 → 3	0.22	1.42	**82.11**
	DMSP	5 → 5	0.22	1.44	**84.86**

Transformer、MLP、SMSP和SMSP-E模型深度固定,而DC和DMSP模型使用动态深度控制器(DDC)优化模型. 单从分类结果方面比较,SMSP模型、DC模型和DMSP模型分类结果明显优于MLP模型和Transformer模型. 单从模型复杂度上比较,MLP、DC、Transformer、DMSP比SMSP、SMSP-E优越. 综合考量分类结果和模型复杂度,则DMSP模型最优. 具体分析如下.

如表4所示,在20 ngp.4上,DMSP的参数量仅比Transformer多0.07 M,但其accuracy比Transformer高11.77%. DMSP与MLP相比,二者参数量持平,但DMSP的分类性能明显大幅超越MLP. 在其他3个数据集上结果类似.

综上所述,多种数据集上的对比实验结果验证了DMSP模型的有效性:一方面,采用去除注意力机制的多层语义感知机MSP方法,丰富了样本的词序语义,提高了模型的泛化能力;另一方面,采用DDCF框架优化了模型深度,降低了模型复杂度,能够在模型复杂度与模型精度间找到一个最佳平衡点.

5　结　论

本文首先提出了一种去注意力的多层语义感知机(MSP)方法,去除了Transformer位置编码器,使用token序列转换函数替换编码器中多头注意力,降低了模型复杂度,并有效提升了模型的语义特征表达. 然后提出了一种动态深度控制DDCF框架,通过自主优化模型深度,降低了模型参数量,同时提高了模型分类性能. 利用MSP特征学习方法和DDCF,进一步提出了DMSP模型,可在模型精度与复杂度之间找到最佳平衡点. 4种数据集上的对比实验结果验证了所提方法的有效性. 综合复杂度和分类性能,在较小规模的文本训练集上,DMSP模型优于Transformer. 从模型在移动端的应用角度,在模型小型化的趋势背景下,DMSP为模型小型化提供了一种可行方法. 下一步将在其他任务上,在多层语义感知机MSP方法和动态深度控制框架DDCF的基础上展开进一步的探索研究.

参考文献(References)

[1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. Proceedings of the 31st International Conference on Neural Information Processing Systems. New York, 2017: 6000-6010.

[2] Cui Y M, Che W X, Liu T, et al. Pre-training with whole word masking for Chinese BERT[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514.

[3] Lin J, Nogueira R, Yates A. Pretrained transformers for text ranking: BERT and beyond[J]. Synthesis Lectures on Human Language Technologies, 2021, 14(4): 1-325.

[4] Sun Z Q, Yu H K, Song X D, et al. MobileBERT: A compact task-agnostic BERT for resource-limited devices[C]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, 2020: 2158-2170.

[5] Huang Y R, Chen J J, Zheng S M, et al. Hierarchical multi-attention networks for document classification[J]. International Journal of Machine Learning and Cybernetics, 2021, 12(6): 1639-1647.

[6] Liu Y, Cheng H, Klopfer R, et al. Effective convolutional attention network for multi-label clinical document classification[C]. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, 2021: 5941-5953.

[7] Zhang B, Xiong D Y, Su J S. Neural machine translation with deep attention[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(1): 154-163.

[8] Gao Y Q, Nikolov N I, Hu Y H, et al. Character-level translation with self-attention[C]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, 2020: 1591-1604.

[9] Song L F, Gildea D, Zhang Y, et al. Semantic neural machine translation using AMR[J]. Transactions of the Association for Computational Linguistics, 2019, 7: 19-31.

[10] Zhao Y C, Meng K, Liu G S. A multi-channel graph attention network for Chinese NER[C]. International Conference on Neural Information Processing. Cham, 2021: 203-214.

[11] Yan S, Chai J P, Wu L Y. Bidirectional GRU with multi-head attention for Chinese NER[C]. 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference. Chongqing, 2020: 1160-1164.

[12] Han X Y, Zhang Y, Zhang W K, et al. An attention-based model using character composition of entities in Chinese relation extraction[J]. Information, 2020, 11(2): 79.

[13] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]. 2021 IEEE/CVF International Conference on Computer Vision. Montreal, 2022: 9992-10002.

[14] Liu Z, Hu H, Lin Y T, et al. Swin transformer V2: Scaling up capacity and resolution[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, 2022: 11999-12009.

[15] Touvron H, Bojanowski P, Caron M, et al. ResMLP: Feedforward networks for image classification with data-efficient training[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(4): 5314-5321.

[16] Tolstikhin I, Houlsby N, Kolesnikov A, et al. MLP-mixer: An all-MLP architecture for vision[J/OL]. 2021, arXiv: 2105. 01601.

[17] Lee-Thorp J, Ainslie J, Eckstein I, et al. FNet: Mixing tokens with Fourier transforms[J/OL]. 2021, arXiv: 2105. 03824.

[18] Guo M H, Liu Z N, Mu T J, et al. Beyond self-attention: External attention using two linear layers for visual tasks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(5): 5436-5447.

[19] Guo J X, Lu S D, Cai H, et al. Long text generation via adversarial training with leaked information[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1): 5141−5148.

[20] Wu Q Y, Li L, Yu Z. TextGAIL: Generative adversarial imitation learning for text generation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(16): 14067-14075.

[21] 唐焕玲, 宋双梅, 刘孝炎, 等. 基于u-wordMixup的半监督深度学习模型[J]. 控制与决策, 2023, 38(6): 1646-1652.
(Tang H L, Song S M, Liu X Y, et al. Semi-supervised deep learning model based on u-wordMixup[J]. Control and Decision, 2023, 38(6): 1646-1652.)

[22] Wei J, Zou K. EDA: Easy data augmentation techniques for boosting performance on text classification tasks[J/OL]. 2019, arXiv: 1901. 11196.

[23] 唐焕玲, 鲁明羽, 邬俊. 基于投票信息熵的AdaBoost改进算法[J]. 控制与决策, 2010, 25(4): 487-492.
(Tang H L, Lu M Y, Wu J. Improved AdaBoost algorithm based on vote entropy[J]. Control and Decision, 2010, 25(4): 487-492.)

[24] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 770-778.

[25] Conneau A, Schwenk H, Barrault L, et al. Very deep convolutional networks for text classification[J/OL]. 2016, arXiv: 1606. 01781.

作者简介

刘孝炎 (1997−), 男, 硕士生, 从事机器学习、人工智能、数据挖掘等研究, E-mail: lxy15058247683@aliyun.com;

唐焕玲 (1970−), 女, 教授, 博士, 从事机器学习、人工智能、数据挖掘等研究, E-mail: thL01@163.com;

王育林 (1998−), 男, 硕士生, 从事机器学习、人工智能、数据挖掘等研究, E-mail: ylinwang@yeah.net;

窦全胜 (1971−), 男, 教授, 博士, 从事机器学习、人工智能、演化计算等研究, E-mail: li_dou@163.com;

鲁明羽 (1963−), 男, 教授, 博士生导师, 从事机器学习、人工智能、数据挖掘等研究, E-mail: lumingyu@dlmu.edu.cn.