结合 SS-GAN 和 BERT 的文本分类模型

宛艳萍,闫思聪,于海阔,许敏聪
(河北工业大学 人工智能与数据科学学院,天津 300401)

摘 要:BERT 是近年来提出的一种大型的预训练语言模型,在文本分类任务中表现优异,但原始 BERT 模型需要大量标注数据来进行微调训练,且参数规模大、时间复杂度高。在许多真实场景中,大量的标注数据是不易获取的,而且模型参数规模过大不利于在真实场景的实际应用。为了解决这一问题,提出了一种基于半监督生成对抗网络的 BERT 改进模型 GT-BERT。采用知识蒸馏的压缩方法将 BERT 模型进行压缩;引入半监督生成对抗网络的框架对 BERT 模型进行微调并选择最优生成器与判别器配置。在半监督生成对抗网络的框架下增加无标签数据集对模型进行微调,弥补了标注数据较少的缺点。在多个数据集上的实验结果表明,改进模型 GT-BERT 在文本分类任务中性能优异,可以有效利用原始模型不能使用的无标签数据,大大降低了模型对标注数据的需求,并且具有较低的模型参数规模与时间复杂度。

关键词:文本分类;半监督;BERT;生成对抗网络;模型压缩
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1673-629X(2023)02-0187-08
doi:10.3969/j.issn.1673-629X.2023.02.028

A Text Classification Model Based on Semi-supervised Generative Adversarial Networks and BERT

WAN Yan-ping, YAN Si-cong, YU Hai-kuo, XU Min-cong
(School of Artificial Intelligence and Data Science, Hebei University of Technology, Tianjin 300401, China)

Abstract: BERT is a large-scale pre-training language model proposed in recent years, which performs well in text classification tasks. However, the original BERT model requires a large amount of labeled data for fine-tuning training with large parameter scale and high time complexity. In many real scenarios, a large amount of labeled data is not easy to obtain, and the large scale of model parameters is not conducive to practical application in real scenarios. To solve this problem, an improved BERT model GT-BERT based on semi-supervised generative adversarial network is proposed. The BERT model is compressed by the compression method of knowledge distillation, and the framework of semi-supervised generative adversarial network is introduced to fine-tune the BERT model and select the optimal generator and discriminator configurations. In the framework of semi-supervised generative adversarial network, an unlabeled data set is added to fine-tune the model, which makes up for the shortcomings of less labeled data. The experimental results on multiple datasets show that the improved model GT-BERT has excellent performance in text classification tasks, can effectively use unlabeled data that the original model cannot use, greatly reduces the model's demand for labeled data, and has low model parameter size and time complexity.

Key words: text classification; semi-supervised; BERT; generative adversarial network; model compression

0 引 言

文本分类是自然语言处理领域(Natural Language Processing, NLP)中的重要分支之一,其目的是将文本按照一定的分类标准,依次归属到一个或多个类别中。目前文本分类在垃圾邮件检测、情感分析及其他领域有着广泛的应用。目前文本分类方法可分为三类:基于规则的文本分类方法、基于机器学习的文本分类方法和基于深度学习的文本分类方法[1]。

近年来,国内外关于文本分类的研究逐渐由传统机器学习向深度学习转变,许多经典深度学习模型如递归神经网络(Recurrent Neural Network, RNN)[2]、卷积神经网络(Convolutional Neural Networks, CNN)[3]和长短期记忆模型(Long Short-Term Memory, LSTM)[4]被应用到文本分类中。杨兴锐等人[5]提出

收稿日期:2022-02-22 **修回日期**:2022-06-23
基金项目:河北省高等学校科学技术研究重点项目(ZD2014051)
作者简介:宛艳萍(1968-),女(回),硕士生导师,副教授,研究方向为自然语言处理、智能计算;通讯作者:闫思聪(1995-),女,硕士生,研究方向为自然语言处理。

了基于自注意力机制和残差网络(ResNet)的 BiLSTM_CNN 复合模型。通过自注意力赋予卷积运算后信息的权重,进一步提高模型的分类性能。李启行等人[6]基于注意力机制的双通道 DAC-RNN 文本分类模型,利用 Bi-LSTM 通道提取文本中的上下文关联信息,并结合 CNN 通道提取文本中连续词间的局部特征。

随着基于 Transformer[7]的 BERT 模型[8]的出现,一系列大型预训练语言模型如 ELMo[9]、GPT[10]、XLNET[11]和 RoBERTa[12]等显著提升了自然语言处理任务的性能。越来越多的研究者开始采用预训练模型进行文本分类,但这些模型训练需要大量标注数据并且参数规模庞大,这导致了预训练时间过长且计算成本高昂,因此一直为人所诟病。

大量标注数据一般需要人工标注,获取难度高且费用贵。为解决这个问题,Weston[13]和 Yang[14]等人提出可以利用半监督的方法,在标注数据量较少的情况下,同时利用未标注数据来提高模型的泛化能力。半监督生成对抗网络(Semi-Supervised Generative Adversarial Networks,SS-GANs)[15]利用了半监督的方法对生成对抗网络(Generative Adversarial Network,GAN)[16]进行扩展,在判别器中为每一个样本增加一个新类别并且判断该样本是否为自动生成的数据。在自然语言处理领域中,Croce 等人[17]首次将 SS-GANs 应用于 NLP 任务中,使用 SS-GANs 的方法扩展基于内核的深度体系结构[18]。Croce 等人[19]也从 SS-GANs 的角度对 BERT 模型进行了简单微调,将 BERT 模型提取特征向量,并利用半监督的方法来降低模型训练所需要的标注数据量。针对 BERT 模型参数过大这一问题,目前已经提出许多神经网络压缩技术,一般可以分为权重量化[20]、权重剪枝[21]和知识蒸馏[22]。Jiao 等人[23]提出了更加复杂的特性模型蒸馏损失函数来提高性能。

针对目前 BERT 模型的缺点,利用半监督的方法在生成对抗性环境下,利用少量标注数据和大量无标注数据对 BERT 模型进行训练,提出了一种基于 SS-GAN 的 BERT 改进模型 GT-BERT。该模型利用知识蒸馏的思路将 BERT 模型进行压缩,将文本输入至 SS-GAN 框架下的压缩模型进行训练。从而降低模型对标注数据的依赖以及模型参数规模与时间复杂度。

该文的主要贡献有:第一,提出了一种改进的 BERT 模型,在 SS-GAN 的框架下丰富了 BERT 的微调过程,改进了模型的整体架构,使模型能够有效利用大量无标注数据;第二,改进了 SS-GAN 中生成器与判别器的选择方式,选择不同的生成器与判别器的最优配置来提高整个训练过程的效果,从而提升模型的整体性能;第三,使用模型压缩方法 Bert-of-theseus 对模型进行压缩,有效降低了模型的参数规模与时间复杂度。

1 相关工作

1.1 GAN 网络

GAN 网络通过生成器与判别模型器两个模块间的相互博弈,产生良好的输出。模型架构如图 1 所示。

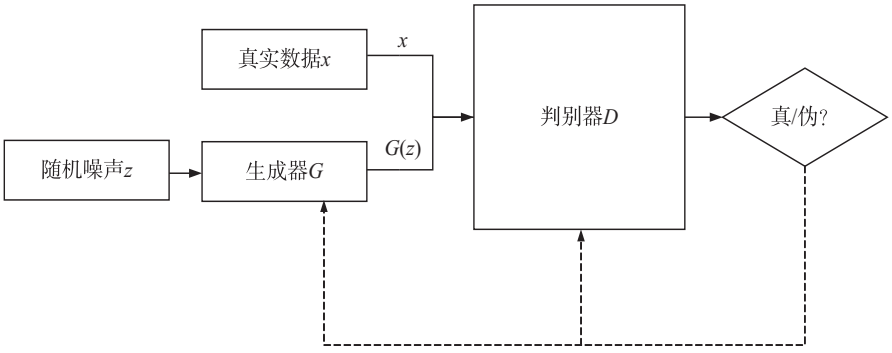

图 1 GAN 网络模型框架结构

图 1 中判别器 D 是一个传统的神经网络分类器,其输入的数据一部分来自真实样本数据集,另一部分来自生成器 G 生成的假样本数据。判别器的训练目标为对于真实样本,判别器输出概率值接近 1,而对于生成器生成的假样本,输出概率值接近于 0。生成器与判别器恰恰相反,其通过训练最大程度地生成接近真实样本的假样本来欺骗判别器。在 GAN 网络中,同时使用两个优化器来最小化判别器与生成器的损失,从而得到良好输出。

1.2 Bert-of-theseus

Bert-of-theseus[24]是基于模块的可替换性,逐步使用小模块替代 BERT 中的大模块来进行模型压缩。在此方法中,被替换模块和替换模块分别被称为前驱模块(predecessor)和后继模块(successor)。

theseus 压缩方法的主要流程如图 2 所示,以将 6 层的 BERT 模型压缩至 3 层模型为例:首先,为两个前驱模块指定一个后继模块,在训练阶段,逐步以一定概率将后继模块替换为所对应的前驱模块。将替换后的

后继模块继续放入模型中与其他剩余的前驱模块共同训练,然后依次将前驱模块替换为后继模块,直至将所有前驱模块替换完毕。当整个训练收敛后,将所有后继模块串联起来作为新模型。假设前驱模型 P 和后继模型 S 都含有 n 个模块,即 $P=\{\mathrm{prd}_1,\mathrm{prd}_2,\cdots,\mathrm{prd}_n\}$,$S=\{\mathrm{scc}_1,\mathrm{scc}_2,\cdots,\mathrm{scc}_n\}$,其中,scc_i 用来替换 prd_i。假设第 i 个模块输入为 y_i ,前驱模型的前向过程为:

$$y_{i+1}=\mathrm{prd}_i(y_i) \tag{1}$$

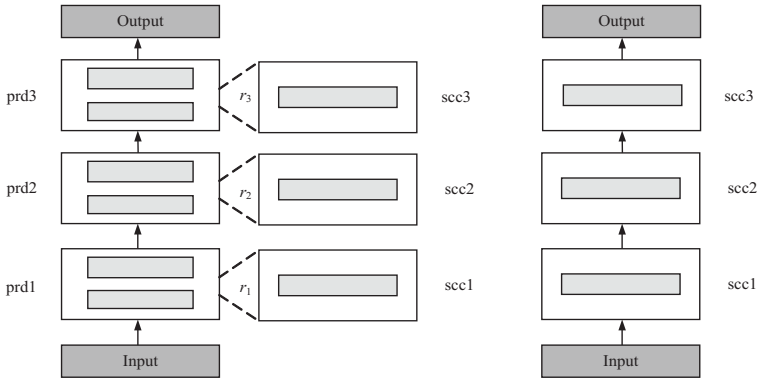

图 2　Bert-of-theseus 压缩流程

压缩时,对于第 $i+1$ 个模块,r_{i+1} 是一个独立服从伯努利分布的变量,即以概率 p 取值为 1,以概率 $1-p$ 取值为 0:

$$r_{i+1}\sim\mathrm{Bernoulli}(p) \tag{2}$$

则第 r_{i+1} 个模块输出为:

$$y_{i+1}=r_{i+1}\odot\mathrm{scc}_i(y_i)+(1-r_{i+1})\odot\mathrm{prd}_i(y_i) \tag{3}$$

其中,$\odot$ 表示逐元素相乘,$r_{i+1}\in\{0,1\}$ 。通过这种方式,前驱模块和后继模块同时训练,训练的损失函数即特定任务所设定的损失函数。

在反向传播时,所有前驱模块的权重参数会被冻结,只有后继模块参数会被更新。前驱模块参数权重冻结后,其嵌入层和输出层在训练阶段会被直接当作后继模块。

当训练收敛时,所有后继模块被串起来组成后继模型 S :

$$S=\{\mathrm{scc}_1,\mathrm{scc}_2,\cdots,\mathrm{scc}_n\} \tag{4}$$

$$y_{i+1}=\mathrm{scc}_i(y_i) \tag{5}$$

由于 scc_i 比 prd_i 规模小,因此整个前驱模型被压缩为更小的后继模型,之后后继模型将会再次使用损失函数进行微调优化,最后使用微调后的后继模型进行后续的文本分类任务。

2　文中模型框架

2.1　SS-GAN

SS-GAN 是对 GAN 网络的一种改进与扩展,利用 GAN 网络进行半监督学习,从而使大量无标注数据能够作为训练数据参与模型训练,降低模型对标注数据的依赖。判别器 D 对输入的每一个样本进行分类,同时区分该样本是否为生成器 G 生成的假样本。SS-GAN 将 GAN 的判别器 D 改为 $k+1$ 类,其中第 $k+1$ 类为异常类:当判别器 D 将输入数据鉴别为真实样本时,将数据分至 $(1,2,\cdots,k)$ 中其中一类;当将输入数据鉴别为生成器 G 生成的假样本时,将数据分至第 $k+1$ 类中。为了训练半监督的 k 类分类器,将生成器 D 的目标扩展如下:

$p_m(\hat{z}=z|\alpha,z=k+1)$ 表示模型 m 中样本数据 α 和假样本数据相关联的概率,$p_m(\hat{z}=z|\alpha,z\in(1,\cdots,k))$ 表示 α 属于目标类中真实数据的概率。D 的损失函数表示为:

$$L_D=L_{D_{\mathrm{sup}}}+L_{D_{\mathrm{unsup}}} \tag{6}$$

其中:

$$L_{D_{\mathrm{sup}}}=-E_{\beta,z\sim p_r}\log[p_m(\hat{z}=z|\alpha,z\in(1,2,\cdots,k))] \tag{7}$$

$$L_{D_{\mathrm{unsup}}}=-E_{\beta,z\sim p_r}\log[1-p_m(\hat{z}=z|\alpha,z=k+1)]-E_{\beta,z\sim p_g}\log[p_m(\hat{z}=z|\alpha,z=k+1)] \tag{8}$$

其中,p_r 和 p_g 分别表示真实数据和生成数据。$L_{D_{\mathrm{sup}}}$ 表示判别器 D 的有监督损失,即在判别过程中判别器 D 将真实样本在 $(1,2,\cdots,k)$ 类中分类错误的损失。$L_{D_{\mathrm{unsup}}}$ 表示 D 的无监督损失,即判别器 D 将无标注的真实样本分至 $k+1$ 类中,或将生成器生成的假样本错误判断为真实样本的损失。利用判别器中间层的输出,使得生成数据与真实数据的特征相匹配,直观上来讲判别器的中间层为特征提取器,用来区别真实数据与生成数据的特征。因此 G 应该尽可能生成近似真实的数据样本,$f(\alpha)$ 表示 D 中间层的输出,G 的特征匹配损失函数定义为:

$$L_{G_{\mathrm{featurematching}}}=\|E_{\alpha\sim p_r}f(\alpha)-E_{\alpha\sim G}f(\alpha)\|_2^2 \tag{9}$$

此外,G 的损失也应考虑生成器 D 正确判别出 G 所生成的假样本所产生的损失,因此:

$$L_{G_{\mathrm{unsup}}}=-E_{\alpha\sim G}\log[1-p_m(\hat{z}=z|\alpha,z=k+1)] \tag{10}$$

因此，G 的损失函数为：

$$L_G = L_{G_{\text{featurematching}}} + L_{G_{\text{unsup}}} \tag{11}$$

2.2 GT-BERT

为了使 BERT 模型能够得到广泛的应用，在保证模型分类准确率不降低的情况下，减少模型参数规模并降低时间复杂度，提出一种基于半监督生成对抗网络与 BERT 的文本分类模型 GT-BERT。模型的整体框架如图 3 所示。

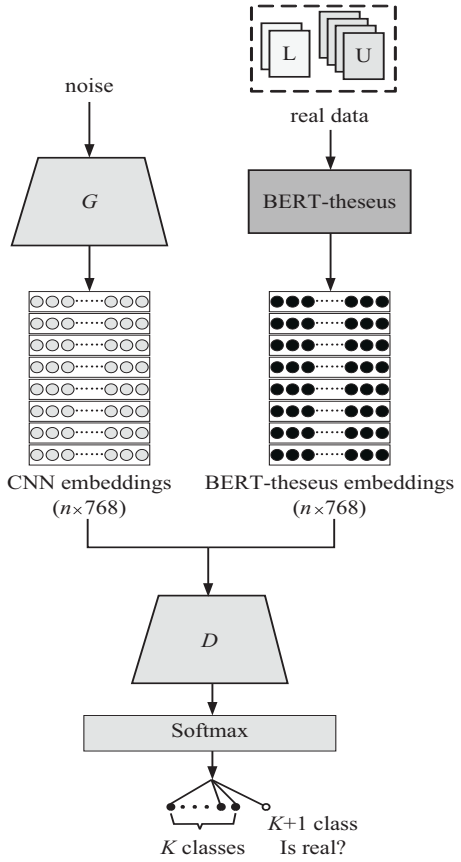

图 3 GT-BERT 模型架构

首先，对 BERT 进行压缩，通过实验验证选择使用 Bert-of-theseus[24] 方法进行压缩得到 BERT-theseus 模型。损失函数设定为文本分类常用的交叉熵损失：

$$L = -\ \text{sum}_{j\in|x|}\sum_{c\in C}[1[z_j = c]\cdot \log p(z_j = c\,|x_j)] \tag{12}$$

其中，x_j 为训练集的第 j 个样本，z_j 是 x_j 的标签，C 和 c 表示标签集合和一个类标签。

接着，在压缩之后，从 SS-GANs 角度扩展 BERT-theseus 模型进行微调。在预训练过的 BERT-theseus 模型中添加两个组件：(1)添加特定任务层；(2)添加 SS-GANs 层来实现半监督学习。本研究假定 K 类句子分类任务，给定输入句子 $s = (t_1, t_2, \cdots, t_n)$，其中开头的 t_1 为分类特殊标记“[CLS]”，结尾的 t_n 为句子分隔特殊标记“[SEP]”，其余部分对输入句子进行切分后标记序列输入 BERT 模型后得到编码向量序列为 $H_{\text{BERT}} = (h_{\text{CLS}}, h_{t_1}, \cdots, h_{t_n}, h_{\text{SEP}})$。

将生成器 G 生成的假样本向量与真实无标注数据输入 BERT-theseus 中所提取的特征向量，分别输入至判别器 D 中，利用对抗训练来不断强化判别器 D。与此同时，利用少量标注数据对判别器 D 进行分类训练，从而进一步提高模型整体质量。

其中，生成器 G 输出服从正态分布的“噪声” h_{fake}，采用 CNN 网络，将输出空间映射到样本空间，记作 $h_{\text{fake}} \in R_d$。判别器 D 也为 CNN 网络，它在输入中接收向量 $h^* \in R_d$，其中 h^* 可以为真实标注或者未标注样本 h_{real}，也可以为生成器生成的假样本数据 h_{fake}。在前向传播阶段，当样本为真实样本时，即 $h^* = h_{\text{real}}$，判别器 D 会将样本分类在 K 类之中。当样本为假样本时，即 $h^* = h_{\text{fake}}$，判别器 D 会把样本相对应的分类于 $K+1$ 类别中。在此阶段生成器 G 和判别器 D 的损失分别被记作 L_G 和 L_D，训练过程中 G 和 D 通过相互博弈而优化损失。

在反向传播中，未标注样本只增加 $L_{D_{\text{unsup}}}$。标注的真实样本只会影响 $L_{D_{\text{sup}}}$，在最后 L_G 和 L_D 都会受到 G 的影响，即当 D 找不出生成样本时，将会受到惩罚，反之亦然。在更新 D 时，改变 BERT-theseus 的权重来进行微调。训练完成后，生成器 G 会被舍弃，同时保留完整的 BERT-theseus 模型与判别器 D 进行分类任务的预测。

3 实 验

3.1 数据集设置

本节验证了 GT-BERT 在不同数据集下文本分类的性能。分别在 20 News Group (20N)、Stanford Sentiment Treebank (SST-5)、Movie Review sentiment classification data (MR) 和 TREC 四个数据集上进行实验。具体数据集设置如表 1 所示。

表 1 数据集设置

数据集	20N	SST-5	MR	TREC
训练集	11 314	8 500	7 108	5 450
测试集	7 532	2 356	3 554	500
类别数	20	5	2	50

为了更好地研究模型在小样本数据下的分类性能，并且避免抽取小样本而带来原有数据集类别分布失衡的情况发生，该文在保证实验数据集与原有数据集中各个数据类别比例相同的情况下，分别抽取数据集中 2%、10% 和 50% 的三组数据进行训练，来对比不同数据量下各个模型的训练效果。此外，还为 GT-BERT 模型额外提供一组无标签数据（$|U| = 100|L|$，其中 $|U|$ 表示无标签数据量，$|L|$ 为有标签数据量）

参与训练。针对特定的数据集,设定了该数据集常用的评价指标,即 SST-5、TREC 和 MR 数据集的评价指标为准确率(accuracy),20N 数据集的评价指标为 F1 值。

3.2 Baseline 设置

将 GT-BERT 模型与目前较为先进的文本分类模型进行了对比实验,对比模型分别为 BERT、GAN-BERT、CNN、LSTM、RoBERTa、XLNet、ALBERT[25]、BERT-theseus[24]、GAN-Bt、GAN-DB、GAN-MB、GAN-TB。其中 GAN-Bt、GAN-DB、GAN-MB 和 GAN-TB 分别为将 BERT-theseus、DistillBERT[26]、MobileBERT[27]和 TinyBERT[23]压缩模型放入半监督 GAN 网络的环境下进行微调所得到的新模型。

3.3 实验过程

使用 BERT-base 作为训练起点。首先进行压缩,在模型压缩中,采用渐进式替换策略来改进 Bert-of-theseus 方法中固定概率 p 的替换方法,后继模块替换前驱模块概率 p_d 会随着时间增加:

$$p_d = \min(1,\varphi(t)) = \min(1,kt + b) \tag{13}$$

其中,t 为训练步长,$k>0$ 为系数,b 为基本替换概率。在压缩的初始阶段,当 $\varphi(t) < 1$ 时,替换模块的数量是 p_d,n 个后继模块的平均学习率为:

$$\grave{l_\alpha} = \left(\frac{n \cdot p_d}{n}\right) \cdot l_\alpha = (kt + b) \cdot l_\alpha \tag{14}$$

其中,l_α 为初始学习率,$\grave{l_\alpha}$ 为考虑所有后继模块的学习率。

在扩展阶段,G 和 D 都为 CNN,激活函数为 leaky-relu 函数。G 的输出为噪声,是由从服从 $N(0,1)$ 中提取的噪声向量通过 CNN 后变成的 768 维向量组成的。D 也为一个 CNN,隐藏层激活函数与 D 相同,在最后一层利用 softmax 进行分类。在训练过程中 epoch 设置为 4,学习率为 2×10^{-5}(20N 数据集的学习率为 5e-6),batch size 为 32,dropout 设定为 0.1,其他超参数和 Devlin 等人[8]设定的超参数一致。

除去有特殊说明的情况外,最终每一个模型都在各个测试集上执行 5 次独立测试,并计算 5 次测试得到的平均值。

3.4 结果与分析

3.4.1 模型文本分类性能比较

表 2 展示了各个模型在四种数据集上针对文本分类任务的性能对比。如表 2 所示:GT-BERT 模型在 20N 数据集所有数据设置下以及其他数据集绝大部分情况下相比于其他模型都取得了最佳性能。

表 2　模型分类性能对比

模型	20N			SST-5			TREC			MR		
	2%	10%	50%	2%	10%	50%	2%	10%	50%	2%	10%	50%
BERT	32.5	68.1	83.1	25.1	42.6	52.3	10.2	24.5	72.8	53.5	70.8	86.1
CNN	28.6	60.8	76.5	22.4	38.5	48.6	7.6	20.3	65.6	51.8	68.9	83.2
LSTM	25.6	61.5	73.5	20.6	36.8	46.4	8.1	20.7	63.2	52.4	69.5	84.2
Bi-LSTM	26.4	60.6	73.6	21.3	37.6	47.6	8.3	21.4	63.4	52.7	70.2	84.4
RoBERTa	30.2	65.1	82.3	22.5	39.3	49.3	9.6	21.2	66.3	52.5	71.5	87.1
XLNet	35.9	69.5	85.8	26.2	40.2	52.8	11.2	25.4	73.6	55.6	72.7	88.2
ALBERT	26.6	64.8	79.2	24.6	40.9	48.4	8.9	22.6	64.1	50.4	69.5	84.3
GAN-BERT	51.0	67.5	83.7	35.6	48.7	51.2	45.6	71.3	79.6	64.5	73.4	85.2
BERT-theseus	31.5	67.9	82.9	23.1	40.3	51.1	9.3	23.8	70.6	52.8	68.6	84.7
GAN-Bt	42.2	68.6	83.5	34.8	46.5	50.4	43.2	70.6	78.8	65.6	72.5	84.6
GAN-DB	49.6	65.4	80.6	31.8	43.2	48.1	41.3	68.6	76.2	64.1	70.8	83.5
GAN-MB	46.5	66.5	80.1	30.9	45.6	49.8	40.8	69.2	74.5	62.4	70.2	83.6
GAN-TB	42.2	63.2	76.4	30.2	41.2	45.1	41.6	66.5	73.4	63.5	71.5	83.4
GT-BERT	54.2	70.5	86.2	36.2	49.4	52.4	45.9	72.6	78.9	66.3	75.8	85.3

由表 2 可知,在使用较少的标注数据时,BERT 模型与其他单独的预训练模型的效果都不尽如人意,而 GAN-BERT 等其他将预训练模型放入半监督环境下微调所得到的模型此时分类效果整体较为良好,尤其在使用 50% 的 TREC 数据集进行训练,GAN-BERT 模型取得了最优效果。这证明将预训练语言模型放入半监督生成对抗框架下利用额外的无标注数据进行微调,能够有效提升模型在低标注数据量下的模型分类性能。

该文着重对比了 GT-BERT 与 BERT、GAN-

BERT 模型,在 20N 数据集、SST-5 数据集和 TREC 数据集下,当使用 2% 的数据进行训练时,BERT 模型的 F1 值和准确率分别仅为 32.5%、25.1% 和 10.2%。而 GT-BERT 和 GAN-BERT 模型相对应值分别为 54.2%、36.2%、45.9% 和 51.0%、35.2%、45.6%,远高于 BERT 模型。因此,当标注数据使用量低时,BERT 模型的分类效果较差,与其他两个模型的差距较大。随着训练使用的标注数据量的增加,三者分类性能随之提升同时之间的差距逐渐缩小,但绝大多数情况下 GT-BERT 仍取得了最佳效果。尤其在 20N 与 TREC 数据集下,当只采用数据集中 2% 的数据(约 220 条和 109 条数据)训练时,GT-BERT 模型性能指标分别为 54.2% 和 45.9%。相较于 BERT 模型提升了 21.7 百分点和 35.7 百分点。这可能表明,当涉及到大量类别时,即当分类任务越复杂时,半监督学习方法的提升效果越好。

3.4.2 压缩方法对比实验

BERT-theseus 与 BERT 在各个数据集下表现大致相似,但参数量却降低一倍[24]。由表 2 可得:与其他将其他压缩模型放入 SS-GANs 框架下微调后模型相比,GAN-Bt 在各个数据集下都获得了良好结果。与原始未经压缩的 GAN-BERT 相比,GAN-Bt 在 20N 数据集下使用 10% 的数据时,F1 值为 68.6%,在 MR 数据集下使用 2% 的数据时,准确率为 65.6%,都略高于 GAN-BERT 模型,其他情况下两者相似。因此,选用 theseus 方法对 BERT 进行压缩并放入 SS-GANs 的环境下进行微调。

3.4.3 生成器与判别器最佳配置选择实验

为了探究不同生成器和判别器对模型的影响并获取最优配置,针对不同的生成器判别器配置组合在四种数据集下进行实验,结果如表 3 所示。为了不增加模型整体参数规模与运行时间,生成器与判别器选择 MLP、CNN 和 LSTM 这三种传统神经网络。表中 GAN-Bt-MC 表示 GAN-Bt 模型的生成器为 MLP、判别器 CNN,GAN-Bt-MM 表示 GAN-Bt 模型的生成器、判别器均为 MLP,其他模型的表示方式与此同理(GT-BERT 模型中生成器和判别器皆为 CNN。GAN-Bt 模型生成器与判别器为 MLP,为了实验对比方便明显,在此表示为 GAN-Bt-MM)。

表 3 不同生成器与判别器对模型的影响

模型	20N			SST-5			TREC			MR		
	2%	10%	50%	2%	10%	50%	2%	10%	50%	2%	10%	50%
GAN-BERT	51.0	67.5	83.7	35.6	48.7	51.2	45.6	71.3	79.6	64.5	73.4	85.2
GAN-Bt-MM	42.2	68.6	83.5	34.8	46.5	50.4	43.2	70.6	78.8	65.6	72.5	84.6
GAN-Bt-MC	51.8	69.2	85.4	33.5	48.9	50.6	45.1	72.3	78.7	63.6	73.7	85.8
GAN-Bt-ML	48.5	68.3	84.2	34.8	47.3	49.1	42.5	70.8	74.6	65.5	74.4	82.6
GAN-Bt-CM	52.1	67.2	83.5	32.5	48.6	48.5	43.8	72.3	76.4	64.8	73.1	83.5
GAN-Bt-CL	49.6	65.3	81.9	34.7	48.3	48.4	43.0	71.5	75.8	64.5	74.4	83.6
GAN-Bt-LM	53.2	70.8	85.6	35.2	49.2	51.6	44.8	71.8	77.2	63.3	75.3	84.1
GAN-Bt-LC	48.8	68.5	83.2	36.4	48.8	52.2	45.5	73.2	77.0	64.2	73.5	84.3
GAN-Bt-LL	51.6	70.2	84.4	34.3	47.9	50.8	46.2	70.1	76.2	65.4	72.7	83.4
GT-BERT	54.2	70.5	86.2	36.2	49.4	52.4	45.9	72.6	78.9	66.3	75.8	85.3

由表 3 可知,生成器与判别器都为 CNN 时,模型在绝大多数数据集设置下取得了最优性能。此外,当判别器为 CNN 时,各个模型取得了相对良好的结果。如 GAN-Bt-MC 在 MR 数据集下使用 50% 的数据时,准确率达到 85.8%;GAN-Bt-LC 在 SST-5 数据集与 TREC 数据集下,分别使用 2% 和 10% 的数据时,各自性能指标分别达到 36.4% 和 73.2%。当生成器为 LSTM 时,各个模型取得了不错的训练结果,如 GAN-Bt-LM 在使用 10% 的 20N 数据集的情况下,F1 值为 70.8%;GAN-Bt-LL 在 TREC 数据集使用率为 2% 时,准确率为 46.2%。虽然当仅考虑单独的生成器或判别器时,或许 GAN-Bt-LC 配置为最优选择,但整体而言 GT-BERT 性能仍远高于 GAN-Bt-LC。原因可能为与 CNN 相比,LSTM 参数较多、训练难度较大。实验中为了公平起见,该文对于生成器与判别器的更新频率参数设置相同,这有可能导致 GT-BERT 模型最终结果最优。因此,选择生成器与判别器都为 CNN 作为最终模型的配置。

3.4.4 模型参数规模与时间复杂度

为了比较模型的参数规模,就目前几种模型的层数与参数进行了对比,见表 4。

由表 4 可以看到,新模型的参数规模相较于原始

BERT 模型以及未进行压缩的 GAN-BERT 模型有很大降低。BERT 模型与 GAN-BERT 模型为 120 M 参数,而 GT-BERT 模型参数仅有 66 M 大小。虽然 ALBERT、TinyBERT 和 MobileBERT 模型的参数规模较小,但各个数据集中分类性能较差,在标记数据量小时,不能很好地应用于实际生产中。

表 4　模型参数规模对比

模型	层数	参数/M
BERT	12	110
GAN-BERT	12	110
ALBERT	12	12
TinyBERT	4	15
DistillBERT	6	66
MobileBERT	24	25
GT-BERT	6	66

图 4 为在四个数据集下、使用 10% 的数据时,各个模型之间文本预测时长的对比图。由图 4 可知,GT-BERT 的预测耗时在不同数据集下都明显小于其他模型。在 SST-5 数据集下,预测时间仅为 48 s,其他模型中耗时最低的 BERT 为 60 s,耗时最高的 XLNet 为 80 s。

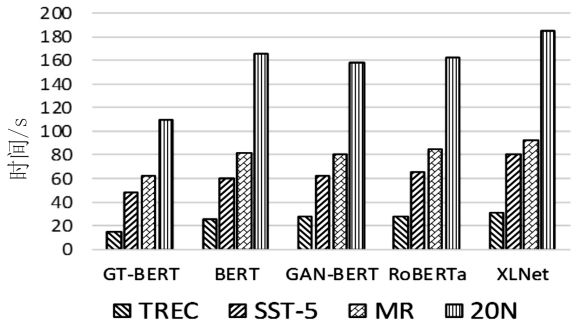

图 4　模型预测时间对比

由上述结果可以看出,在标注数据量较低的小样本数据集中,所提模型可以有效地利用无标注数据,提升模型的分类性能,同时降低模型的参数规模和时间复杂度。

4　结束语

该文提出了一种用于文本分类任务的 GT-BERT 模型。首先,使用 theseus 方法对 BERT 进行压缩,在不降低分类性能的前提下,有效降低了 BERT 的参数规模和时间复杂度。然后,引入 SS-GAN 框架改进模型的训练方式,使 BERT-theseus 模型能有效利用无标注数据,并实验了多组生成器与判别器的组合方式,获取了最优的生成器判别器组合配置,进一步提升了模型的分类性能。实验结果表明,提出的 GT-BERT 模型能够有效利用无标注数据,降低了模型对标注数据的依赖程度,在标注数据量较小的数据集上的文本分类性能显著高于其他模型,并且拥有较低的参数规模与时间复杂度。

参考文献:

[1] MINAEE S, KALCHBRENNER N, CAMBRIA E, et al. Deep learning--based text classification: a comprehensive review[J]. ACM Computing Surveys, 2021, 54(3): 1-40.

[2] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder - decoder for statistical machine translation[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Doha: ACL, 2014: 1724-1734.

[3] SHIN H C, ROTH H R, GAO M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning[J]. IEEE Transactions on Medical Imaging, 2016, 35(5): 1285-1298.

[4] SHI X, CHEN Z, WANG H, et al. Convolutional LSTM network: a machine learning approach for precipitation nowcasting[C]//Advances in neural information processing systems. Montreal: MIT Press, 2015: 802-810.

[5] 杨兴锐,赵寿为,张如学,等. 结合自注意力和残差的 BiLSTM_CNN 文本分类模型[J]. 计算机工程与应用,2022,58(3):172-180.

[6] 李启行,廖　薇,孟静雯. 基于注意力机制的双通道 DAC-RNN 文本分类模型[J/OL]. 计算机工程与应用:1-9[2022-03-17]. http://kns.cnki.net/kcms/detail/11.2127.tp.20210420.1354.070.html.

[7] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is

all you need[C]//Advances in neural information processing systems. Long Beach:MIT Press,2017:5998-6008.
[8] DEVLIN J,CHANG M W,LEE K,et al. Bert:pre-training of deep bidirectional transformers for language understanding [J]. arXiv:1810.04805,2018.
[9] PETERS M E,NEUMANN M,IYYER M,et al. Deep contextualized word representations[J]. arXiv:1802.05365,2018.
[10] RADFORD A,WU J,CHILD R,et al. Language models are unsupervised multitask learners [J]. OpenAI blog,2019,1(8):9.
[11] YANG Z,DAI Z,YANG Y,et al. XLNet:generalized autoregressive pretraining for language understanding [J]. arXiv:1906.08237,2019.
[12] LIU Y,OTT M,GOYAL N,et al. Roberta:a robustly optimized bert pretraining approach[J]. arXiv:1907.11692,2019.
[13] WESTON J,RATLE F,MOBAHI H,et al. Deep learning via semi-supervised embedding[M]//Neural networks:tricks of the trade. Berlin:Springer,2012:639-655.
[14] YANG Z,COHEN W,SALAKHUDINOV R. Revisiting semi-supervised learning with graph embeddings[C]//International conference on machine learning. New York City:PMLR,2016:40-48.
[15] SALIMANS T,GOODFELLOW I,ZAREMBA W,et al. Improved techniques for training gans[J]. Advances in Neural Information Processing Systems,2016,29:2234-2242.
[16] GOODFELLOW I,POUGET-ABADIE J,MIRZA M,et al. Generative adversarial nets[J]. Advances in Neural Information Processing Systems,2014,1050:10.
[17] CROCE D,CASTELLUCCI G,BASILI R. Kernel-based generative adversarial networks for weakly supervised learning [C]//International conference of the italian association for artificial intelligence. [s.l.]:Springer,2019:336-347.
[18] CROCE D,FILICE S,CASTELLUCCI G,et al. Deep learning in semantic kernel spaces[C]//Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1:long papers). Vancouver:ACL,2017:345-354.
[19] CROCE D,CASTELLUCCI G,BASILI R. GAN-BERT:generative adversarial learning for robust text classification with a bunch of labeled examples[C]//Proceedings of the 58th annual meeting of the association for computational linguistics. [s.l.]:ACL,2020:2114-2119.
[20] HAN S,MAO H,DALLY W J. Deep compression:compressing deep neural networks with pruning,trained quantization and Huffman coding[J]. arXiv:1510.00149,2015.
[21] GONG Y,LIU L,YANG M,et al. Compressing deep convolutional networks using vector quantization [J]. arXiv:1412.6115,2014.
[22] HINTON G,VINYALS O,DEAN J. Distilling the knowledge in a neural network[J]. arXiv:1503.02531,2015.
[23] JIAO X,YIN Y,SHANG L,et al. Tinybert:distilling bert for natural language understanding [J]. arXiv:1909.10351,2019.
[24] XU C,ZHOU W,GE T,et al. Bert-of-theseus:compressing bert by progressive module replacing[J]. arXiv:2002.02925,2020.
[25] LAN Z,CHEN M,GOODMAN S,et al. Albert:a lite bert for self-supervised learning of language representations[J]. arXiv:1909.11942,2019.
[26] SANH V,DEBUT L,CHAUMOND J,et al. DistilBERT,a distilled version of BERT:smaller,faster,cheaper and lighter [J]. arXiv:1910.01108,2019.
[27] SUN Z,YU H,SONG X,et al. Mobilebert:a compact task-agnostic bert for resource-limited devices [J]. arXiv:2004.02984,2020.

(上接第 186 页)

sociation for computational linguistics:human language technologies. Seattle:NAACL,2016:1480-1489.
[13] LUO L,YANG Z,YANG P,et al. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition[J]. Bioinformatics,2018,34(8):1381-1388.
[14] YANG J,ZHANG Y,LI L,et al. YEDDA:a lightweight collaborative text span annotation tool [J]. arXiv:1711.03759,2017.
[15] WU G,TANG G,WANG Z,et al. An attention-based BiLSTM-CRF model for Chinese clinic named entity recognition[J]. IEEE Access,2019,7:113942-113949.
[16] WANG Haocheng,WANG Yafeng,HU Yue. Bidirectional Viterbi decoding algorithm for OvTDM[J]. 中国通信:英文版,2020,17(7):183-192.
[17] 潘理虎,赵彭彭,龚大立,等. 煤矿事故案例命名实体识别方法研究[J]. 计算机技术与发展,2022,32(2):154-160.