文章编号：1003-0077(2020)09-0078-11

基于改进生成对抗网络的谣言检测方法

李 奥[1,2]，但志平[1,2]，董方敏[1,2]，刘龙文[1,2]，冯 阳[1]

(1. 三峡大学 计算机与信息学院，湖北 宜昌 443002；
2. 三峡大学 水电工程智能视觉监测湖北省重点实验室，湖北 宜昌 443002)

摘 要：传统谣言检测算法存在提取文本语义、关键特征等效果不理想的问题，而一般序列模型在文本检测中无法解决特定语义下的特征提取，导致模型泛化能力差。为解决上述问题，该文提出一种改进的生成对抗网络模型(TGBiA)用于谣言检测，该模型采用对抗训练方式，符合谣言在传播过程中人为增删、夸大和歪曲信息的特点，通过对抗网络生成器和判别器的相互促进作用，强化谣言指示性特征的学习，不断提高模型的学习能力。训练过程中的生成器通过 Transformer 结构代替单一的 RNN 网络，实现语义的提取和特征的学习，同时，在训练过程中的判别器采用基于双向长短期记忆单元的深度网络分类模型，并引入注意力机制来提升对较长时间序列谣言的判断能力。在公开的微博和 Twitter 数据集上的实验结果表明，该文提出的方法比其他现有方法检测效果更好，鲁棒性更强。

关键词：谣言检测；生成对抗网络；注意力机制

中图分类号：TP391　　**文献标识码**：A

An Improved Generative Adversarial Network for Rumor Detection

LI Ao[1,2], DAN Zhiping[1,2], DONG Fangmin[1,2], LIU Longwen[1,2], FENG Yang[1]

(1. School of Computer and Information Technology, China Three Gorges University, Yichang, Hubei 443002, China;
2. Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering, China Three Gorges University, Yichang, Hubei 443002, China)

Abstract: Existing rumor detection algorithms, including general sequential models, are defected in capturing text semantics and key features detection, resulting in poor generalization capability. To address this issue, this paper proposes an improved generative adversarial network model named TGBiA for rumor detection. TGBiA adopts adversarial training method, to capture the development of augmentation, detraction, exaggeration and distortion during its spread. Generator model extracts sequence semantics and features via Transformer instead of RNN. And the discriminator is a classification model based on BiLSTM, with the attention mechanism introduced. Through the mutual promotion of the generator and discriminator, it enables the learning of the indicative features of rumors increasingly. Experimental results on the Weibo and Twitter datasets show that the proposed method is not only outperforms other existing detecting methods but is also more robust.

Keywords: rumor detection; generative adversarial network; attention mechanism

0 引言

谣言会引起公众恐慌和社会动荡，有效检测谣言和扼制谣言的传播有助于社会的安定和健康发展，具有较大的现实意义。谣言具有时序性和特征多样性，传播过程中信息会被不断加工，具有很强的迷惑性。现阶段有许多学者从传统机器学习算法如支持向量机、决策森林等分类器出发构建分类模型，然而，这类算法需预先定义和抽取文本相关特征，对

收稿日期：2019-12-04　定稿日期：2020-02-26

基金项目：国家自然科学基金—新疆联合基金项目(U1703261)；国家自然科学基金(61871258)；国家重点研发计划(2016YFB0800403)；湖北省自然科学基金(2018CFC852)

特定类型的数据集依赖性较强，因而导致分类模型泛化能力差。同时，人为构建特征工程主观性强，工作量大。

为了解决以上问题，国内外学者研究如何利用神经网络中的循环神经网络单元（RNN）和卷积神经网络单元（CNN）来提取谣言特征，避免人为构建特征工程，搭建具有语义理解能力的分类模型。CNN 具有不错的特征提取能力，而 RNN 能很好地解决文本序列模型的时序性，引入 CNN 和 RNN 结构，可以提高谣言检测的性能，如：结合注意力机制的深度循环神经网络[1-3]、多尺度卷积核的 C-GRU 模型[4]等。但是这类模型未结合谣言在网络中传播的特点，在提取语义特征时无法让模型辨析关键特征，并且在现实生活中谣言会随着时序而改变部分边缘信息，增加语义混淆度，从而降低模型对关键特征的提取能力，导致检测效果不理想。

针对现有工作中的不足，Ma 等人提出基于 RNN 的生成对抗网络（GAN）谣言检测模型[5]，通过生成器模拟谣言在时序上的无关特征变化，可以有效训练模型的关键特征提取能力。该模型利用 GRU 对文本序列进行特征学习，采用对抗式训练来提高判别器的分类能力，在 Twitter 数据集上取得了较好的预测效果。但是，该模型没有考虑较长文本序列会导致 RNN 的特征矩阵信息丢失问题，单向的 RNN 只能从前一时刻确定当前时序的语义信息，而谣言文本中词与词之间语义相互依赖现象会导致模型的特征提取能力差，并且 RNN 依赖于时序间的次序进行训练，无法做到并行计算，导致模型训练速度慢。

基于目前谣言检测存在的问题，本文结合谣言传播特点，利用 Transformer 结构[6]可并行化训练特点、注意力机制在长序列信息处理中的优势和双向的长短期记忆网络学习时序特征能力较强的特点，提出了基于 Transformer 结构和注意力机制的双向长短期记忆网络（BiLSTM）的生成对抗网络谣言检测模型 TGBiA。首先，采用 Transformer 结构中的自注意力机制来获取谣言文本的关键信息，并提供给模型做并行化计算；其次，采用双向的长短期记忆网络来设计判别器，有效解决长序列文本的信息丢失问题，并获得上下文语义依赖，充分学习时序语义特征；最后，引入层间注意力机制，对分类能力较差的语义特征进行信息增强。通过在 Twitter 谣言数据集和中文微博谣言数据集上的实验，表明本文提出的模型具有较强的泛化能力，各项指标与现有的方法相比都具有很强的竞争力，预测精度更高。

本文的主要贡献如下：

（1）提出将 Transformer 结构和生成对抗网络相结合的自然语言分类模型；

（2）利用注意力机制增强文本语义特征表达，相比其他类似模型，该模型具有可并行化特点；

（3）提出的方法在公开的 Twitter 数据集和微博数据集上都取得了很好的预测准确度。

1　相关工作

近些年，为了抑制谣言的网络传播，一些学者开始研究谣言的相关特征，手工构建有效的特征工程。如 Qazvinian 等人[7]在 2011 年对 Twitter 中产生的谣言进行检测，提取了推文内容特征、用户行为特征和微博标签特征，实验结果表明这些特征组合能有效地检测出谣言，尤其是使用推文内容特征效果最好；Takahashi 等人[8]在 2012 年对 Twitter 上产生的谣言进行分析，发现谣言推文的转发率要远高于正常推文的转发率，而且谣言短文本和正常推文的关键词分布也有所不同。刘知远等人[9]在 2015 年通过统计微博谣言转发、评论、举报时间、举报者和发布者等信息来对数据进行聚类，对谣言文本进行语义分析，并对谣言传播的时序进行统计，确定其关键性特征，构建自动辟谣框架。毛二松等人[10]在 2016 年总结 Qazvinian 选取的关键特征，将推文情感倾向性特征、传播树结构特征、用户特征作为谣言检测的重要特征，构建了基于决策树和 SVM 的集成分类器。王志宏等人[11]在 2019 年构建基于内容、用户、结构以及事件流行度、模糊度、流传度的特征工程，采用以 RBF 为核函数的 SVM 模型，在微博数据集上取得了不错的效果。

关键特征的提取一直是谣言检测领域研究的重点。Ma 等人[5]在 2016 年提出利用递归神经网络来提取特征信息，让网络去学习序列的相关特征，这比人为分析确定特征要更加客观、可靠，在 Twitter 数据集上取得了不错的效果。而单一的 RNN 可能忽略了谣言的传播特征，考虑到谣言随时间的动态变化情况，Ma 等人在 2018 年提出基于树结构的递归神经网络[12]，分析了推文内容特征和评论特征，在 Twitter 数据集上取得了 0.737 的准确度，但是该方法在构造树上依赖大量评论，导致节点在递归神经网络训练中出现信息丢失现象，不利于递归网络的训练。

由于卷积在图像特征提取方面效果不错，刘政

等人[13]尝试利用CNN来实现谣言检测，采用句向量结合多层卷积实现了谣言检测框架。最近李等人[4]结合CNN和RNN的优势，提出C-GRU谣言检测模型，用CNN提取句子的特征，利用GRU学习句子的语义信息。但是该方法没有考虑到谣言随时间会发生细微变化的特点，也无法学习序列的关键特征。Ma等人[14]基于RNN构造的生成对抗神经网络模型，考虑到信息随时间的一些变化情况，在生成器和判别器上都使用GRU，在Twitter上取得了0.86的精度，但是该模型的生成器结构与本文有较大区别，单向的GRU在长序列文本上会出现信息丢失现象，没有考虑到序列内部词汇之间的特征关系，这使得生成器丧失隐藏关键特征的能力，无法训练分辨能力强的判别器。

以上这些工作均有效提高了谣言检测的效果，但各个方法没有完全考虑到如何让网络本身模拟谣言在现实生活中传播的变化情况，去学习序列的关键特征。

2 本文模型

本文提出的生成对抗模型旨在提高模型对关键特征的学习能力。对抗网络中的生成器基于现有的谣言样本，通过对这些现有样本进行语义学习，构建关键特征，生成富含表现特征的文本样例，促使判别器提高自己的语义特征识别能力，并且通过模拟谣言在传播过程中的信息流失和混淆，来提高判别器识别谣言的能力。整体网络模型架构如图1所示，网络整体分为数据预处理部分、生成器网络、判别器网络以及输出层部分。

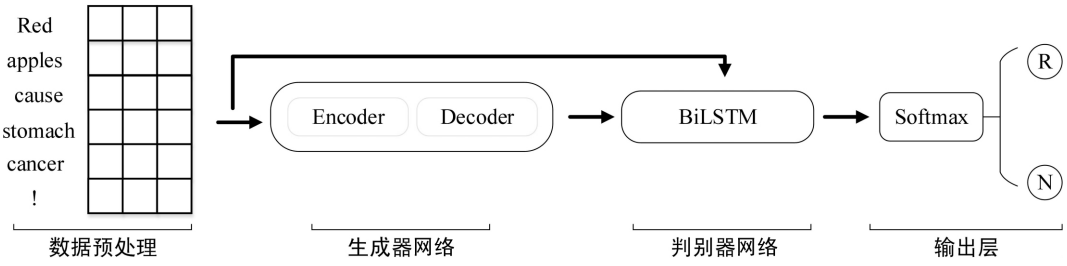

图1 整体网络模型架构

2.1 数据预处理

为了将谣言文本的位置信息嵌入到词向量中，对于数据集中的谣言短文本序列，通过对数据特殊符号清洗、分词、词频过滤、向量化后，数据被分成多个由词向量组成的序列 $\boldsymbol{S}_i=\{\boldsymbol{w}_1,\cdots,\boldsymbol{w}_n\}$，其中，$\boldsymbol{w}_i$ 代表序列中的语义词向量，i 代表每个词向量的具体位置。这 n 个词组代表了这一条序列的所有信息。

将单词在序列中的位置信息标记为pos，利用三角函数对位置信息进行编码，如式(1)所示。

$$\begin{cases} x_{(\mathrm{pos},2i+1)}=\cos\left(\mathrm{pos}*\dfrac{1}{10\,000^{\frac{2i}{d_{\mathrm{model}}}}}\right) \\ x_{(\mathrm{pos},2i)}=\sin\left(\mathrm{pos}*\dfrac{1}{10\,000^{\frac{2i}{d_{\mathrm{model}}}}}\right) \end{cases} \tag{1}$$

其中，d_{model} 为序列中词的维度，i 代表词向量的位置，$x_{(\mathrm{pos},i)}$ 表示处于pos位置上的第 i 个单词的位置编码，在通过计算得到一组位置信息向量 $\mathrm{wp}_i=\{x_1,\cdots,x_n\}$，结合上一步得到的单词语义向量 $\boldsymbol{S}_i$，最终输入到模型中的是单词向量 $\boldsymbol{M}_i=S_i+\mathrm{wp}_i$。下一步构建生成器，采用Transformer的encoder-decoder结构，将序列关键特征进行表现或隐藏，以提高分类器的谣言语义鉴别能力。

2.2 基于Transformer结构的生成器

生成器由两层Transformer结构组成，主要结构为encoder-decoder，encoder具有自动编码器的功能，它创建原始数据的隐藏或压缩表示，将谣言文本序列编码为向量矩阵，输出的向量将原始数据压缩为低维数据，而decoder则基于压缩的向量矩阵来重建输入数据，并在原有信息基础上进行解码序列。但这样重建序列可能会导致特征的信息与原始序列不一样，因此需要注意力机制来引导网络去关注谣言序列的关键特征，学习关键特征和边缘特征，进而解码得到基于关键特征的谣言序列，凸显边缘特征或加入无关谣言的噪声来模拟谣言随时间的变化特性，进而增强判别器的谣言识别能力。

2.2.1 注意力机制

为了将序列词组的不同特征信息表示出来，更有效地提取谣言序列关键特征，实现并行化训练，在

序列词组经过 embedding 时，会初始化三个状态矩阵 $\boldsymbol{W}_q$、$\boldsymbol{W}_k$、$\boldsymbol{W}_v$，它们主要解决传统 RNN 提取特征时，对时序过于依赖的问题。$\boldsymbol{W}_q$、$\boldsymbol{W}_k$、$\boldsymbol{W}_v$ 矩阵的并行计算可以同时得出所有时序的输出权重和特征矩阵表示，这三个矩阵依次与词向量 $\boldsymbol{M}_i$ 进行矩阵运算，如式(2)～式(4)所示。

$$\boldsymbol{x}_q = \boldsymbol{M}_i \cdot \boldsymbol{W}_q \tag{2}$$

$$\boldsymbol{x}_k = \boldsymbol{M}_i \cdot \boldsymbol{W}_k \tag{3}$$

$$\boldsymbol{x}_v = \boldsymbol{M}_i \cdot \boldsymbol{W}_v \tag{4}$$

通过上面的公式用三个向量来表示 $\boldsymbol{M}_i$，即 $\boldsymbol{M}_i = \{\boldsymbol{x}_q, \boldsymbol{x}_k, \boldsymbol{x}_v\}$。进入 Self-Attention 层，依次将每一个词向量的 q 和 k 分向量进行矩阵运算得到 score，如式(5)所示。

$$\text{score}_i = \boldsymbol{x}_q \cdot \boldsymbol{x}_k \tag{5}$$

进行归一化后输入到 softmax 函数里得到当前词向量对谣言序列输出向量 $\boldsymbol{c}_i$ 转换的权重，式(6)为 attention 的计算公式，其中 n 为词的长度，取值范围为 $n \in [1, 512]$。

$$\begin{cases} \boldsymbol{a}\ (\text{score}_i, x_v)_i = \text{softmax}\left(\dfrac{\text{score}_i}{\sqrt{n}}\right) * \boldsymbol{x}_v \\ \boldsymbol{c}_i = \sum\limits_{i=1}^{n} \boldsymbol{a}_i \end{cases} \tag{6}$$

经过 Self-Attention 层之后，进入全连接层进行特征提取。

2.2.2　全连接层以及层间注意力机制

为了提升谣言的关键特征和边缘特征对分类结果的影响力，防止特征之间信息融合导致特征丢失，全连接层采用独立的前向网络单元分别对每一个自注意力层的输出进行前向运算，如图 2 所示。

图 2　全连接层前向运算过程

图 2 中，Linear 表示线性层，$\boldsymbol{c}_i$ 表示自注意力层输出的第 i 个谣言序列词向量。通过独立的线性层能很好地保存词的特征，decoder 层将谣言信息进行特征增强，通过层间的注意力机制来使信息增强偏向谣言的边缘特征，将谣言文本序列向对立类别转化，使序列特征具有很强的迷惑性。层间注意力的输入一部分来自 decoder 自注意力层的输出 $\boldsymbol{D}_i = \{d_1, \cdots, d_n\}$，另一部分则来源于 encoder 结构输出的谣言语义压缩特征向量 $\boldsymbol{E}_i = \{e_1, \cdots, e_n\}$，层间注意力计算，如式(7)所示。

$$\begin{cases} a(d_q, e_k, e_v) = \text{softmax}\left(\dfrac{d_q \cdot e_k}{\sqrt{n}}\right) * e_v \\ \text{Out}'_i = \sum\limits_{i=1}^{n} a_i \end{cases} \tag{7}$$

其中，Out′ 表示层间注意力输出的词向量。通过 decoder 结构，原始序列的信息在很大程度上得到了保留，并且通过解码进行特征重编码，使新得到的序列 $\bar{\boldsymbol{S}}_i$ 能在原来基础上产生新的语义信息。对于谣言 X_r 和非谣言 X_n 的文本，该编解码器类似构造了如下非线性函数，使得原始特征映射到新的特征空间中，如式(8)所示。

$$f(x) = \begin{cases} x_r \rightarrow x_{\text{rn}} & \text{if } \text{S}_i \text{ is rumors} \\ x_n \rightarrow x_{\text{nr}} & \text{if } \text{S}_i \text{ is non-rumors} \end{cases} \tag{8}$$

通过网络的训练，需要让这两层 Transformer 结构不断学习序列的特征，并且让谣言的文本序列尽可能保存其关键特征，凸显其边缘特征或加入新的噪声，使其特征分布尽可能像非谣言的序列特征分布，同理，也希望非谣言的文本序列能在原有的语义基础上，尽可能伪装成谣言的语义，隐藏序列关键特征，使得生成器具有迷惑判别器的能力。图 3 为生成器整体结构图，从图 3 中可以看到，encoder 和 decoder 结构上最大的不同在于，decoder 多了一层 encoder-decoder attention，这一层主要是为了在解码时，能考虑上一层的输入和来自 encoder 的输入，最后通过线性层接到 decoder 得到序列，为进行特征隐藏后的谣言序列。

2.3　判别器模型

判别器由双向递归神经网络和 Self-Attention 构成。由于谣言数据集序列长短不一，普通的 RNN 网络能够存取的上下文信息范围有限，在处理较长序列时，隐藏层输入对于网络输出的影响会随着网络环路的不断递归而衰退，而 LSTM 结构[15-16]正好解决这一问题。此外，单向的循环网络在时序上往往忽略未来的词信息，在提取特征信息时如能查阅序列的上下文来标注信息，这对于分类任务来说，将提高模型对特征的获取能力[17]。

本文采用 BiLSTM 来实现判别器，具体结构如图 4 所示。

首先，用 $a_0 = (h_0, \text{ce}_0)$ 对正向和反向的 LSTM 层状态进行初始化，其中，h_0 为 LSTM 隐藏状态、ce_0 为 LSTM 元胞状态，$LSTM$ 单元状态 a_0 采用正态分布随机初始化。经过生成器伪装后的谣言序列

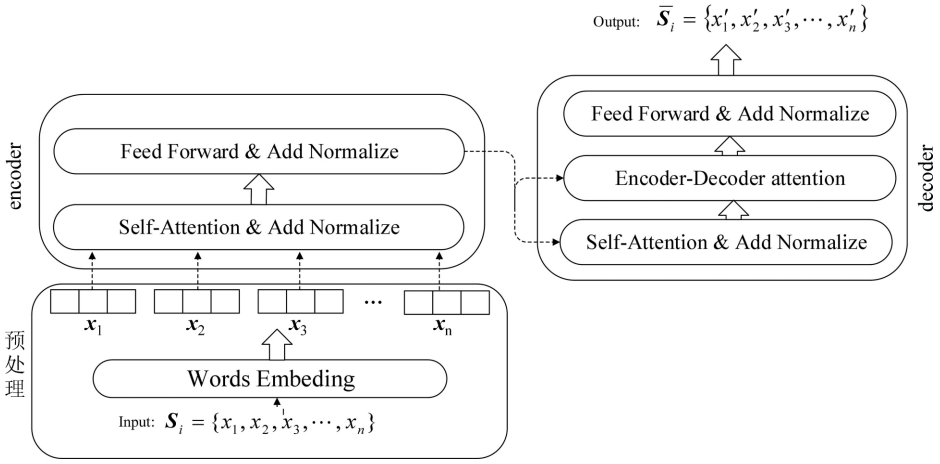

图 3 TGBiA 模型的生成器结构图

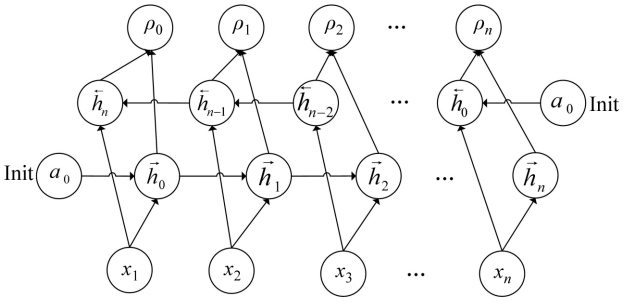

图 4 BiLSTM 的双向网络结构

数据 $X_i = \{x_1, \cdots, x_n\}$ 进入 BiLSTM 的 LSTM 结构单元，LSTM 单元内部的输入门对序列进行信息提取，如式(9)、式(10)所示。

$$i_t = \sigma(\boldsymbol{W}_i \cdot [h_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_i) \tag{9}$$

$$C_t = \tanh(\boldsymbol{W}_c \cdot [h_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_c) \tag{10}$$

式中，h_{t-1} 为前一个时间序列状态，x_t 为当前 t 时刻的输入，也是经 Transformer 处理的谣言序列词向量。C_t 是当前时刻输入词语和上一时刻输出的语义信息，$\boldsymbol{W}_i$，$\boldsymbol{W}_c$ 是输入门对谣言序列的解析矩阵。为了滤掉一些和分类无关的特征信息，保留关键特征，将 C_t，$\boldsymbol{x}_t$ 输入到遗忘门，进行如下特征过滤计算：

$$f_t = \sigma(\boldsymbol{W}_f \cdot [h_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_f) \tag{11}$$

式中，$\boldsymbol{W}_f$，$\boldsymbol{b}_f$ 分别为遗忘门的权重矩阵。经过信息提取和过滤后，计算得到我们新的词信息 $\widetilde{C}_t$，如式(12)所示。

$$\widetilde{C}_t = f_t * C_{t-1} + i_t * C_t \tag{12}$$

得到要输出的特征矩阵，LSTM 控制这个状态信息和输入特征向量对当前时刻输出 o_t 的影响，接下来进入输出门，最终得到此时序谣言词向量 ρ_t，如式(13)～式(15)所示。

$$o_t = \sigma(\boldsymbol{W}_o [h_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_o) \tag{13}$$

$$\overleftrightarrow{\rho}_t = o_t * \tanh(\widetilde{C}_t) \tag{14}$$

$$\boldsymbol{\rho}_t = [\overrightarrow{\rho} : \overleftarrow{\rho}] \tag{15}$$

其中，$\boldsymbol{W}_o$ 是输出门的权重矩阵，$\overrightarrow{\rho}$ 为 LSTM 正向或反向得到的序列向量，$\overrightarrow{\rho}$ 为 LSTM 正向得到的序列向量，$\overleftarrow{\rho}$ 为 LSTM 反向得到的序列向量。BiLSTM 的最后一层得到所有时序的输出 $T_i = \{\boldsymbol{\rho}_i \mid 0 < i < n\}$，考虑到谣言检测理应受到关键特征的影响，而不是一些边缘特征，因此，在 BiLSTM 输出层引入注意力层[18]，如图 5 所示，让网络去学习关键特征的权重，分类得到的 N 代表非谣言，R 代表谣言。

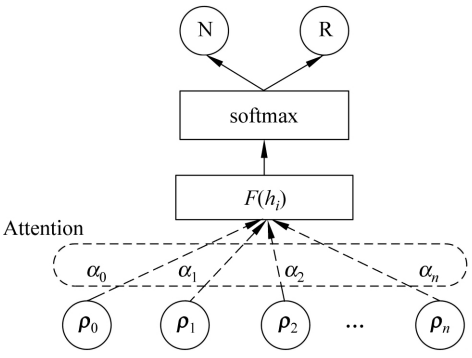

图 5 BiLSTM 的输出层

在输入序列中，构造特征矩阵 $\ddot{s}$，对每一个 LSTM 单元的输出 $\boldsymbol{\rho}_i$ 分别计算出影响因子 e_i，如式(16)、式(17) 所示。

$$e_i = a(\ddot{s}, \boldsymbol{\rho}_i) \tag{16}$$

$$\alpha_i = \frac{\exp(e_i)}{\sum_{i=0}^{n} \exp(e_i)} \tag{17}$$

归一化后得到每一时序的权重，然后按照权重

对每一时序加权求和，如式(18)所示。

$$F(\rho_i)=\sum_{i=0}^{n}\alpha_i\boldsymbol{\rho}_i \tag{18}$$

最后经过 softmax 层计算谣言和非谣言的类别概率。

2.4　GAN 分类模型

在进行模型训练时，生成器尽可能将文本序列解码成对立的类别文本，在上述生成器模型中，定义了映射函数 $f(X)$，表示将谣言伪装为非谣言，或将非谣言伪装为谣言的机制。但是单一的映射可能会导致序列被过度解码，原始语义可能会被改变为目标语义[14]，从而降低判别器的分类能力。为此，需要将生成器的映射函数改为可逆的，即将 Transformer 的输出作为其输入，保证得到原始输入。这里定义映射关系，如式(19)所示。

$$\varphi_s(X)=\begin{cases}f_s(X_r)\rightarrow X_r\\ f_s(X_n)\rightarrow X_n\end{cases} \tag{19}$$

生成器的可逆性保证了解码之后的特征能还原原始特征，并生成混乱的特征序列来达到 Transformer 网络的目的，生成器的损失函数由基于原始序列的预测分类差值 $f_s(X)$、基于生成序列的预测分类差值 $\varphi_s(X)$ 以及它们的欧氏距离组成，如式(20)所示。

$$L_G=D(f_S(X))+D(\varphi_S(X))+\frac{1}{n}\sum_{i=0}^{n}\|f_S(X)-\varphi_S(X)\| \tag{20}$$

其中，D 代表判别器模型。判别器使用 BiLSTM 模型，其损失值由原始分类差值和对生成器生成序列的判别值组成，如式(21)所示。

$$L_D=D(X)+D(f_S(X)) \tag{21}$$

作为分类的生成网络对抗模型，希望生成器能伪装谣言和非谣言序列，使其特征分布尽可能地偏向对立类别，使原始序列和伪装序列的差别尽可能地增大。而对于判别器而言，它需要训练其网络以判别出原始序列的类别，并尽力将伪装后的序列和原始序列归为一类，最大化它们的特征共性，这些特征就是需要判别器去学习的关键特征。对抗网络模型的优化如式(22)、式(23)所示。

$$V(G,D)=\omega L_G+(1-\omega)L_D \tag{22}$$

$$G*=\min_G\max_D V(G,D) \tag{23}$$

其中，ω 定义为权衡对抗网络中的损失系数，默认值为 0.5，V 表示模型关于生成器损失函数 L_G 和判别器损失函数 L_D 的关系，G^* 表示模型优化方式。在网络训练时，先训练 D 网络参数，让生成器的序列类别和期望目标类别差异尽可能大，接着优化 G 网络，使生成序列能伪装成功，让对抗网络中的判别器无法分辨出谣言伪装序列和非谣言的特征差异。

训练按照图 6 算法，网络的参数初始化符合特定分布，数据预处理后进入网络训练，设置小批量更新网络的权重和偏差。对于生成器和判别器，训练过程中要分别进行参数更新，以免影响谣言特征的学习。每轮训练都采用 Adam 算法[19] 对所有 weights 和 biases 更新，通过自适应调整学习率来优化网络收敛的速度。

Algorithm 1: 生成对抗网络训练算法

Data: 谣言数据集:Twitter,Weibo,Pheme
Input: 短文本：S_i，标签：{0,1}，网络初始学习率 ϵ

读取配置文件并初始化网络的 weights 和 biases;
初始化生成器 ΘG 和判别器 ΘD；
R ← 谣言文本；
N ← 非谣言文本；
for *epoch* $=0\rightarrow Max\_Iters$ **do**
　for *each mini-batch* $\{X_r\leftarrow DataSet[R]\}$ *and* $\{X_n\leftarrow DataSet[N]\}$ **do**
　　网络预测文本的类别概率；
　　$\varphi_{X_r}, f_{X_r}\leftarrow \Theta G(X_r)$;
　　$\varphi_{X_n}, f_{X_n}\leftarrow \Theta G(X_n)$;
　　计算生成器和判别器的损失值;
　　$L_G\leftarrow \Theta D(\varphi_X)+\frac{1}{n}\sum_{i=0}^{n}\|\varphi_X-f_X\|$;
　　$L_D\leftarrow \Theta D(X)+\Theta D(f_X)$;
　　反向计算生成器的参数梯度 $\nabla(\Theta G)$;
　　更新生成器的可训练参数:$\Theta G\leftarrow \Theta G-\epsilon\nabla(\Theta G)$;
　　反向计算判别器的参数梯度 $\nabla(\Theta D)$;
　　更新判别器的可训练参数:$\Theta D\leftarrow \Theta D-\epsilon\nabla(\Theta D)$;
　end
end

图 6　TGBiA 模型训练算法

3　实验结果与分析

3.1　实验数据

实验结果表明，模型训练迭代次数达到 150 次以上时，网络的交叉熵[20]损失值基本保持不变。所以最大迭代次数设置为 200，采用 mini-batch 值为 16 来更新网络参数效果较好。实验设置初始学习率为 0.01，采取自适应递减策略更新学习率。中文和英文的词向量维度设置为 50。

实验数据采用公开的 Twitter 数据集和 Weibo 数据集，数据集的详细信息如表 1 所示。

表 1 中 Twitter 数据集来源于 Ma 等人在 2016 年公开的数据集，这些数据大部分取自在线辟谣网站 www.snopes.com，其中部分非谣言事件来源于

Castillo等人[21]和Kwon等人[22]公开的数据集。数据集谣言和非谣言事件数量相对平衡，在992个事件中，谣言有498件，非谣言有494件。微博数据集来源于中国的新浪微博社交平台，新浪社区管理中心提供了大量谣言信息，同时整理Ma等人利用爬虫收集到的非谣言事件。Pheme数据集包含突发新闻期间发布的Twitter谣言和非谣言的集合，数据来源于Zubiaga等人[23]在2016年收集的谣言与非谣言数据集，包含2 246条新闻事件，其中谣言和非谣言事件数量相等，均为1 123条。

表1 数据集的信息统计表

信息	数据集		
	Twitter	Pheme	Weibo
用户	491 229	29 387	2 746 818
推文	1 101 985	102 579	3 805 656
事件	992	2 246	4 664
谣言	498	1 123	2 313
非谣言	494	1 123	2 351

微博数据集样本较多，这里仅从中列举出部分谣言和非谣言样本，如表2所示。

表2 微博数据集中的部分样本

微博ID	微博内容	标签	所属集
3557662195382754	中国人转起来！我同意，我转发！一定不去进影院，大家一起为《贞子》票房为零，做努力！中国人拍的《金陵十三钗》在日本小鬼子票房为零。小日本拍的《贞子》3D将于5月12日在中国大陆上映。而5月12日既是南京大屠杀纪念日，又是国难日。勿忘国耻!!作为中国人，敢不敢让贞子3D 5月12日票房为零	谣言	训练集
3489110939758010	今天，小日本通过钓鱼岛国有化，无视中国主权，当初中国人拍的《金陵十三钗》在小鬼子的票房为零。小日本拍的《贞子》3D将于9月12日在中国大陆上 映。而9月12日既是南京大屠杀纪念日，又是国难日。勿忘国耻!!作为中国人，敢不敢让《贞子》3D 9月12日票房为零。朋友们，抵制日货，请大家转发！	谣言	训练集
3487319200055757	国人拍的《金陵十三钗》小鬼子票房为零。小日本拍的《贞子》3D将于9月18日在中国大陆上映。而9月18日既是南京大屠杀纪念日，又是国难日。勿忘国耻！作为中国人，敢不敢让贞子3D9月18日票房为零	谣言	测试集
3922141743754003	【央视快讯：#泰国载中国人巴士翻车# 已至10余人遇难】据泰媒报道，当地时间20日中午11:30左右，一辆载中国游客的巴士发生翻车事故。已造成10余人遇难，多人受伤，当地救援队伍正在实施救援。中国驻清迈领馆工作人员正在赶赴现场了解情况。具体情况，央视记者正继续跟进。	非谣言	训练集
3922137142166302	据泰国头条新闻报道，12月20日一辆载有中国游客的旅游巴士，在清迈山区的公路上，与一辆轿车相撞后翻下山去。目前相关人员正在进行抢救，伤亡不明。	非谣言	测试集

以表2中的前三条样本为例分析来看，其中前两条谣言样本在训练过程中，训练的分类模型提取“《金陵十三钗》在小鬼子票房为零”这样的关键特征，而不同日期的南京大屠杀纪念日等关联信息一直变化着干扰读者，在测试过程中分类模型通过检测到测试样本的关键特征就能准确预测样本的分类，即使测试样本中包含正确日期的南京大屠杀纪念日等关联信息，也不影响其被判定为谣言。这些谣言企图添加一些关联信息来混淆视听，让读者分辨不清关键信息，这也反映出谣言在传播过程中本身的关键信息并没有改变，只是边缘特征信息在变化的特点，如果分类模型无法正确检测出关键特征，那么在预测时，精度就会变低，泛化能力就会变差。

3.2 实验设置

为验证本文方法的有效性，分别在这三个数据集上进行实验，将实验结果与其他现有方法进行比对和分析。实验划分训练集和测试集，测试集和训

练集的比例为 1∶4，每一个数据集中的谣言和非谣言样本比例为 1∶1。划分方式利用程序的随机种子，对样本随机分配。

本文模型实验方法和其他对比实验方法如下：

（1）**DTC**：Castillo 等人[21]提出的基于手动选择的微博的全局统计特征构建决策树分类器。

（2）**SVM-RBF**：Yang 等人[24]基于手动选择特征的微博统计特征，提出基于 RBF 核函数的 SVM 分类器。

（3）**GRU、GAN-GRU**：Ma 等人[14]提出的通过基于 GRU 构建生成对抗网络的谣言检测模型。

（4）**TG-R，TG-BiR，TG-BiA**：本文提出的基于 Transfomer 结构的 GAN 模型，判别器分别采用 LSTM，BiLSTM 和 BiLSTM＋Attention 结构进行实验。

实验采用 5 折交叉验证，设置常用的评价指标为：正确率 Accuracy、准确率 Precision 以及召回率 Recall。实验结果如图 7、图 8 所示。

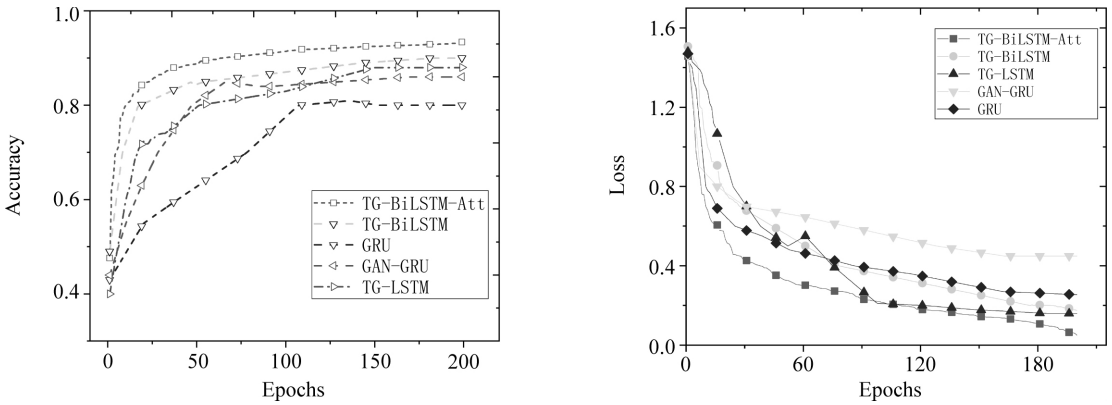

图 7　为模型在新浪微博数据集训练时，accuracy 和 loss 随 epoch 的变化情况

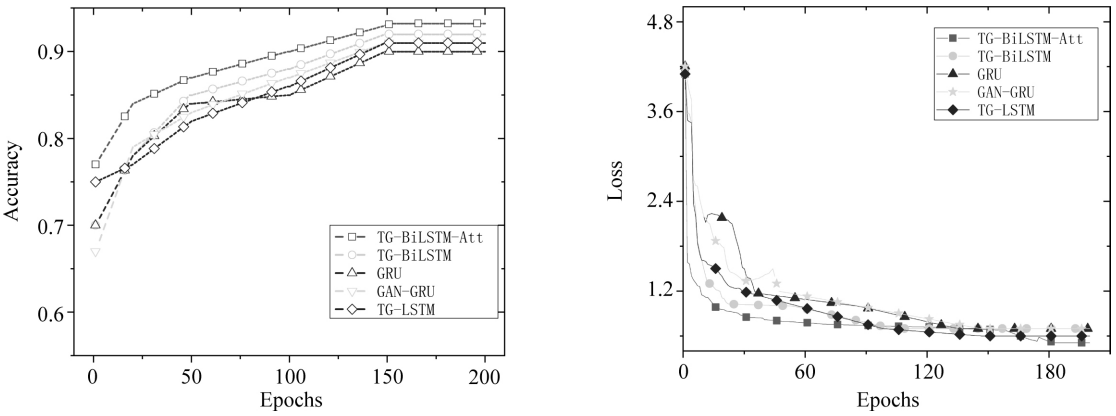

图 8　模型在 Twitter 数据集上训练时，accuracy 和 loss 随 epoch 的变化情况

从图 7、图 8 中可以直观地看出，在三个数据集上，TG-BiLSTM-attention 模型检测谣言的正确率 Accuracy 在进行 200 个时序周期之后达到最高正确率，损失函数值随时序变化明显。

图 7 展示了各个模型在微博数据集上的训练情况，在最初时段，各个模型的正确率都在 0.4～0.5 之间，随着训练的进行，各个模型的学习特征能力也显著上升，其中 TG-BiLSTM-attention、TG-BiLSTM、TG-LSTM、GAN-GRU 增长最快，但是增长幅度最大的是 TG-BiLSTM-attention。同样利用生成对抗网络来训练模型，加入了 attention 机制的 TG-BiLSTM-attention 分类效果明显，这说明 attention 机制对于序列的语义识别和特征提取有帮助。在 TG-LSTM和 GAN-GRU 对比中，起初 TG-LSTM 模型预测精度等指标没有 GAN-GRU 好，这可能是由于 Transformer 结构在最初阶段提取的特征相比 RNN 要丰富一些，导致判别器在最初无法鉴别谣言的关键特征，从而识别正确率低下。而随时序进行运算时，对抗网络中的生成器和判别器的能力不断进化，Transformer 结构很好地增强了判别器的谣言检测能力，使其正确率明显高于 GAN-GRU。从损失过程图中可以看到，损失值下降最快的是

TG-BiLSTM-attention模型，而其他模型则在前期基本同步，到后期下降速度和幅度则明显变小。

图8是各个模型在Twitter数据集上的实验结果，从中可以看出，在谣言预测正确率方面，通过GAN-GRU和TG-LSTM对比可以发现，在生成网络对抗模型中，将RNN换成Transformer结构，可以明显提升模型预测正确率。这说明Transformer结构在序列的关键特征提取上明显强于RNN。

在三个数据集上的谣言检测结果如表3所示。

表3 谣言检测结果(R：谣言；N：非谣言)

(a) 在Twitter数据集上的实验结果

模型	类别	Accuracy	Precision	Recall
DTC	R	0.731	0.724	0.757
	N		0.739	0.704
SVM-RBF	R	0.722	0.856	0.526
	N		0.663	0.914
GRU	R	0.835	0.821	0.858
	N		0.852	0.812
GAN-GRU	R	0.863	0.843	0.892
	N		0.885	0.833
TG-R	R	0.892	0.891	0.899
	N		0.893	0.883
TG-BiR	R	0.910	0.921	0.901
	N		0.900	0.912
TG-BiA	R	0.932	0.946	0.951
	N		0.928	0.899

(b) 在新浪微博数据集上的实验结果

模型	类别	Accuracy	Precision	Recall
DTC	R	0.831	0.847	0.815
	N		0.815	0.847
SVM-RBF	R	0.818	0.822	0.812
	N		0.815	0.824
GRU	R	0.908	0.871	0.958
	N		0.953	0.858
GAN-GRU	R	0.911	0.933	0.920
	N		0.889	0.913
TG-R	R	0.924	0.922	0.899
	N		0.926	0.943
TG-BiR	R	0.939	0.941	0.951
	N		0.890	0.882
TG-BiA	R	0.942	0.956	0.953
	N		0.898	0.899

(c) 在Pheme数据集上的实验结果

模型	类别	Accuracy	Precision	Recall
DTC	R	0.581	0.582	0.573
	N		0.579	0.588
SVM-RBF	R	0.704	0.724	0.675
	N		0.687	0.734
GRU	R	0.742	0.737	0.753
	N		0.754	0.730
GAN-GRU	R	0.781	0.773	0.796
	N		0.791	0.766
TG-R	R	0.824	0.812	0.805
	N		0.827	0.843
TG-BiR	R	0.859	0.742	0.791
	N		0.903	0.889
TG-BiA	R	0.910	0.916	0.889
	N		0.898	0.914

从表3中可以看出，本文的TG-BiA(TG-BiLSTM-attention)模型在Twitter、微博、Pheme数据集上的预测正确率Accuracy依次为0.932、0.942、0.910。从整体上看，深度神经网络算法相比传统分类算法在谣言检测上效果更好，其中DTC表现得最差，在三个数据集上预测正确率Accuracy都没能达到0.850。表3(a)显示，在生成对抗网络的生成器中将RNN替换为Transfomer结构时，即模型GAN-GRU与TG-R相比，模型预测正确率Accuracy上升了0.029，谣言预测准确率Precision-R和非谣言预测准确率Precision-N分别提升了0.048和0.008。而将判别器的单向LSTM改为双向LSTM时，TG-BiR(TG-BiLSTM)模型的预测正确率Accuracy达到了0.910的效果，说明序列模型的语义信息要通过上下文语义来获得，单靠上文推断关键特征是不可靠的。而从表3(a)中数据可以看出，在TG-BiR(TG-BiLSTM)引入attention变为TG-BiA(TG-BiLSTM-attention)之后，谣言的预测精度Precision-R达到最高为0.946，而总体的正确率Accuracy也上升了0.022，说明注意力机制增强了模型对序列的理解能力。从表3(b)中，可以看出TG-R(TG-LSTM)在谣言预测精度Precision-R和非谣言预测精度Precision-N上都达到了0.920以上，相比GAN-GRU模型，效果要好。在总体的预测正确率Accuracy上，TG-BiA(TG-BiLSTM-attention)达到最高为0.942，谣言的预测精度Precision-R达到最高为0.956。在表3(c)中，可以看出，模型TG-BiA

(TG-BiLSTM-attention)预测正确率 Accuracy 能达到 0.910，远高于其他模型。

从实验结果中找出原例子样本的预测结果，如表 4 所示，可以看出各个模型的预测结果，其中后面两个模型的检测结果与实际标签一致。这说明模型没有被边缘特征所干扰，能准确地学习到谣言的关键特征，利用关键特征去分类样本。

表 4　谣言样本检测结果

微博 ID	模型				
	GRU	GAN-GRU	TG-R	TG-BiR	TG-BiA
3557662195382754	R	R	R	R	R
3489110939758010	N	R	N	R	R
3487319200055757	R	N	R	R	R
3922141743754003	R	N	N	N	N
3922137142166302	N	N	R	N	N

从以上数据分析来看，与其他现有方法相比，本文模型在谣言检测上效果最好，总样本的正确率和正负样例的预测精度都达到了很高。因此，本文模型在谣言检测上效果要更好。

4　总结

本文提出了基于 Transformer 结构的生成器和 BiLSTM＋Attention 构造的判别器组成的生成对抗网络方法用于谣言检测。在生成对抗网络模型的生成器上，设计了两层的 Transformer 结构，对序列信息进行自编码和解码。在判别器上，通过采用 BiLSTM 结构，双向的深度循环神经网络可以更好地理解序列的上下文相关信息，并且，在 LSTM 层引入注意力机制，来加强关键特征对谣言判别结果的影响。结果也证明，本模型在谣言检测中表现十分出色。

未来，将考虑从以下两个方向来改进模型：

(1) 可以考虑其他优秀的语言模型，如 BERT、XLNet 等，对现有模型做进一步改进；

(2) 现实生活中的谣言不仅有文字，也有图片信息。考虑在模型中加入图片信息，建立多模态模型进一步提高谣言检测性能。

参考文献

[1] 廖祥文，黄知，杨定达，等. 基于分层注意力网络的社交媒体谣言检测[J]. 中国科学：信息科学，2018，48(11)：1558-1574.

[2] Ajao O，Bhowmik D，Zargari S. Fake news identification on Twitter with hybrid CNN and RNN models [C]//Proceedings of the 9th International Conference on Social Media and Society，2018：226-230.

[3] Chen W，Zhang Y，Yeo C K，et al. Unsupervised rumor detection based on users' behaviors using neural networks[J]. Pattern Recognition Letters，2018，105：226-233.

[4] 李力钊，蔡国永，潘角. 基于 C-GRU 的微博谣言事件检测方法[J]. 山东大学学报(工学版)，2019，49(02)：102-106，115.

[5] Ma J，Gao W，Mitra P，et al. Detecting rumors from microblogs with recurrent neural networks[C]//Proceedings of the International Joint Conference on Artificial Intelligence，2016：3818-3824.

[6] Vaswani A，Shazeer N，Parmar N，et al. Attention is all you need[C]//Proceedings of Advances in Neural Information Processing Systems，2017：5998-6008.

[7] Qazvinian V，Rosengren E，Radev D R，et al. Rumor has it：Identifying misinformation in microblogs[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing，2011：1589-1599.

[8] Takahashi T，Igata N. Rumor detection on Twitter [C]//Proceedings of the 6th International Conference on Soft Computing and Intelligent Systems，and The 13th International Symposium on Advanced Intelligence Systems，2012：452-457.

[9] 刘知远，张乐，涂存超，等. 中文社交媒体谣言统计语义分析[J]. 中国科学：信息科学，2015，45(12)：1536-1546.

[10] 毛二松，陈刚，刘欣，等. 基于深层特征和集成分类器的微博谣言检测研究[J]. 计算机应用研究，2016，33(11)：3369-3373.

[11] 王志宏，过弋. 微博谣言事件自动检测研究[J]. 中文信息学报，2019，33(06)：132-140.

[12] Ma J，Gao W，Wong K-F. Rumor detection on Twitter with tree-structured recursive neural networks [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics，2018：1980-1989.

[13] 刘政，卫志华，张韧弦. 基于卷积神经网络的谣言检测[J]. 计算机应用，2017，37(11)：3053-3056，3100.

[14] Ma J，Gao W，Wong K-F. Detect rumors on Twitter by promoting information campaigns with generative adversarial learning[C]//Proceedings of the World Wide Web Conference，2019：3049-3055.

[15] Gers F. Long short-term memory in recurrent neural networks[D].PhD Diss, Universität Hannover, 2001.
[16] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
[17] Schuster M. Bi-directional recurrent neural networks for speech recognition[C]//Proceeding of IEEE Canadian Conference on Electrical and Computer Engineering, 1996: 7-12.
[18] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [J]. arXiv preprint arXiv:1409.0473, 2014.
[19] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv: 1412.6980, 2014.
[20] De Boer P-T, Kroese D P, Mannor S, et al. A tutorial on the cross-entropy method[J]. Annals of Operations Research, 2005, 134(1): 19-67.
[21] Castillo C, Mendoza M, Poblete B. Information credibility on Twitter[C]//Proceedings of the 20th International Conference on World Wide Web, 2011: 675-684.
[22] Kwon S, Cha M, Jung K, et al. Prominent features of rumor propagation in online social media[C]//Proceedings of the IEEE 13th International Conference on Data Mining, 2013: 1103-1108.
[23] Zubiaga A, Liakata M, Procter R. Learning reporting dynamics during breaking news for rumour detection in social media[J]. arXiv preprint arXiv:1610.07363, 2016.
[24] Yang F, Liu Y, Yu X, et al. Automatic detection of rumor on Sina Weibo[C]//Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, 2012: 13.

李奥(1994—),**硕士研究生,主要研究领域为自然语言处理**。
E-mail: liao@ctgu.edu.cn

董方敏(1965—),**博士,教授,主要研究领域为图形图像处理、智能信息处理**。
E-mail:fmdong@ctgu.edu.cn

但志平(1976—),**通信作者,博士,教授,主要研究领域为模式识别、自然语言处理**。
E-mail: zp_dan@ctgu.edu.cn