ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 37,No. 2,Mar. 2022,pp. 396－404
DOI:10. 16337/j. 1004-9037. 2022. 02. 012

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

融合LSTM-GRU网络的语音逻辑访问攻击检测

杨海涛[1]，王华朋[1]，牛瑾琳[1]，楚宪腾[1]，林暖辉[2]

(1. 中国刑事警察学院公安信息技术与情报学院，沈阳 110854；2. 广州市刑事科学技术研究所，广州 510030)

摘　要：为进一步提高语音欺骗检测的准确率，提出一种融合LSTM-GRU网络的语音逻辑访问攻击(语音转换、语音合成)检测方法。融合LSTM-GRU网络是由长短期记忆网络(Long short-term memory，LSTM)层、门控循环神经单元(Gated recurrent unit，GRU)层、丢弃层、批归一化层和全连接层串联结合的一种混合网络，其中LSTM层可以解决语音序列中的长时依赖问题，GRU层则可降低模型参数量。实验在ASVspoof2019 LA数据集上进行，提取20维的梅尔倒谱系数特征用于模型训练，在测试阶段使用训练好的LSTM-GRU模型对测试集中的语音进行欺骗检测。与GRU网络及LSTM网络的比较结果表明：LSTM-GRU网络在3种网络模型中正确识别率最高，等错误率(Equal error rate，EER)比ASVspoof2019挑战赛所提供基线系统低27.07%，对逻辑访问攻击语音检测的平均准确率达到98.04%，并且融合LSTM-GRU网络具备训练时间短、防止过拟合及稳定性高等优点。结果证明本文方法可有效应用于语音逻辑访问攻击检测任务中。

关键词：逻辑访问攻击；梅尔倒谱系数；等错误率；LSTM-GRU网络

中图分类号：TP391.4；TN912.3　　**文献标志码**：A

Logical Access Attack Audio Detection Based on LSTM-GRU

YANG Haitao[1]，WANG Huapeng[1]，NIU Jinlin[1]，CHU Xianteng[1]，LIN Nuanhui[2]

(1. Video and Audio Material Examination Department，Criminal Investigation Police University of China，Shenyang 110854，China；2. Criminal Science and Technology Institute of Guangzhou，Guangzhou 510030，China)

Abstract：In order to improve the accuracy of speech spoofing detection，a speech spoofing detection method based on LSTM-GRU network is proposed. LSTM-GRU network is a hybrid network combining long short-term memory (LSTM) layer，gated recurrent unit (GRU) layer，dropout layer，batch normalization layer and dense layer in series. LSTM layer can solve the problem of longtime dependence in speech sequence，while GRU layer can reduce the number of model parameters. The experiment is conducted on the ASVspoof2019 LA dataset，and the 20-dimensional Mel-frequency cepstral coefficient features are extracted for model training. In the test stage，the trained LSTM-GRU model is used for deception detection of the speech in the test set. By comparing with separate GRU and LSTM networks，the results show that：LSTM-GRU network achieves the highest correct recognition rate among the three network models；the equal error rate is 27.07% lower than the baseline system provided by the ASVspoof2019 challenge；the average accuracy of speech detection for logical access attack is 98.04%；

基金项目：国家重点研发计划(2017YFC0821000)；广州市科技计划(2019030004)；辽宁网络安全执法协同创新中心(WXZX-201807003)；司法部司法鉴定重点实验室(司法鉴定科学研究院)开放基金；中国刑事警察学院研究生创新能力提升项目。
收稿日期：2021-06-22；**修订日期**：2021-11-05

LSTM-GRU network has the advantages of short training time, over-fitting prevention and high stability. It is proved that the proposed method can be effectively applied to speech logical access attack detection task.

Key words: logical access attack; Mel-frequency cepstral coefficients; equal error rate; LSTM-GRU network

引　言

科学技术的发展给人们带来便利的同时也产生了新的问题,例如语音作为生物识别技术的重要环节在日常生活中常常被人恶意利用,以进行诈骗、造谣和煽动公众情绪等。语音欺骗方法很早就产生,其类型主要包括:语音模仿、语音回放、语音合成和语音转换[1]。近年来人们开始重视语音欺骗检测。自动说话人识别欺骗攻击与防御对策挑战赛(Automatic speaker verification spoofing and countermeasures challenge,ASVspoof)于2015年第一次举办,主要关注于逻辑访问(Logical access,LA),包括语音合成(Text to speech, TTS)和语音转换(Voice conversion, VC)检测[2]。随后的ASVspoof2017注重于物理访问(Physical access,PA)区分真实音频和回放音频[3]。ASVspoof2019则涵盖了LA和PA[4]。在这几个挑战赛中度量标准都是等错误率(Equal error rate, EER),包括语音合成、语音转化的逻辑访问攻击语音因其逼真性而被广泛应用[5],这也给不法分子提供了便利条件。传统机器学习的语音欺骗检测主要使用高斯混合模型和i-vector,前者具有训练速度快、准确度高的优点,但由于语料不够,抗信道干扰差;后者则对全局差异进行建模,除信道的干扰,放宽了对训练语料的限制[1,6-7]。随着深度学习的快速发展,深度神经网络(Deep neural network,DNN)被应用于语音欺骗检测。Villalba等使用DNN对提取的率波库(Filter bank, FBank)及相对相移(Relative phase shift, RPS)特征进行检测,在10种欺骗语音检测结果中有9种EER低于0.05%,取得了非常好的效果[8]。卷积神经网络(Convolutional neural networks, CNN)在图像领域的成功应用为语音处理提供了更多思路。Lavrentyeva等使用CNN的变种LCNN进行语音回放检测,并在ASVspoof2017挑战赛中取得语音回放检测第一名的成绩[9],证明了CNN在语音欺骗检测中的能力。处理语音时序数据能力较强的是循环神经网络(Recurrent neural network,RNN),RNN通过循环单元和门限结构使其具有记忆性。Gomez-Alanis等使用CNN-RNN的混合模型对噪声鲁棒性语音进行欺骗检测,取得了较好的效果[10]。该团队在后来的研究中使用GRU-RNN的混合模型对回放语音、转换语音及合成语音进行欺骗检测,其结果都比ASVspoof2019提供的基线系统更优[11]。但是RNN在处理长时依赖问题时易出现梯度消失和梯度爆炸的现象[12]。Hochreiter等提出的长短期记忆网络则是为了解决这一问题[13]。在ASVspoof2017挑战赛中,Li团队使用了基于注意力机制的LSTM结构取得较好的结果[14]。Cho等于2014年提出的门控循环神经单元是长短期记忆网络(Long short-term memory, LSTM)的变种中改动较大的一种[15]。Chen等使用门控循环神经单元(Gated recurrent unit, GRU)在ASVspoof2017数据集上进行试验,EER为9.81%,表现突出[16]。文献[17-18]则对LSTM和GRU网络模型进行了比较,发现两者的能力相当,但相比于LSTM网络,GRU的张量操作更少,训练速度更快,泛化能力更强。

在语音逻辑访问攻击检测的任务中,单一的神经网络结构在进行逻辑访问攻击检测时存在着一定的局限性,因此混合网络模型成为研究热点。在处理语音序列中,LSTM网络和GRU网络能够更好地处理语音序列中的长时依赖问题,进而提高网络的性能。由于两种网络结构相似,在融合时能够正确获取语音信息。为进一步提高语音欺骗检测的准确率,本文将LSTM网络及GRU网络进行融合,提出一种融合LSTM-GRU网络模型进行语音欺骗检测研究。

1 门控循环神经网络

门控循环神经网络是在传统DNN的基础上加入了门控机制用来控制神经网络中信息的传递，可以解决长时依赖关系问题，避免了梯度消失和梯度爆炸。

1.1 长短期记忆网络

LSTM网络结构由一系列的记忆单元组成，记忆单元通常包含一个自连接记忆单元来存储网络的时间状态。LSTM拥有3个门（输入门、输出门和遗忘门）来保护和控制单元状态，也就是控制信息的流动，其中：输入门决定记忆单元内保存什么新信息；输出门决定要输出的单元状态信息；遗忘门决定要忘记什么内容。图1所示为LSTM记忆单元结构。在时间步长t处，LSTM可表示为

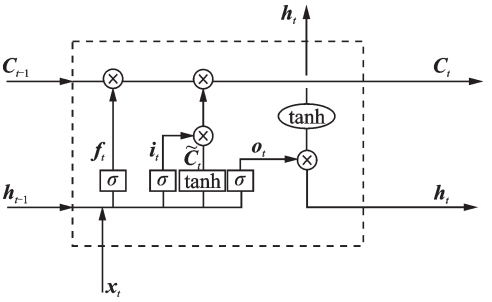

图1 LSTM记忆单元

Fig.1 LSTM memory cell

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_f\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_f) \tag{1}$$

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_i\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_c) \tag{2}$$

$$\tilde{C}_t=\tanh(W_C.[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_c) \tag{3}$$

$$\boldsymbol{C}_t=\boldsymbol{f}_t*\boldsymbol{C}_{t-1}+\boldsymbol{i}_t*\tilde{\boldsymbol{C}}_t \tag{4}$$

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_o\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_o) \tag{5}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t*\tanh(\boldsymbol{C}_t) \tag{6}$$

式中：激活函数使用的是Sigmoid函数（σ）和双曲正切函数（tanh）；$\boldsymbol{i}_t$、$\boldsymbol{o}_t$、$\boldsymbol{f}_t$、$\boldsymbol{C}_t$、$\tilde{\boldsymbol{C}}_t$分别表示为输入门、输出门、遗忘门、记忆单元内容和新记忆单元内容；$\boldsymbol{W}$表示权重矩阵；$\boldsymbol{b}$表示偏置向量，比如$\boldsymbol{b}_i$表示输入门的偏置向量；$\boldsymbol{h}_t$为时间t时的隐层向量。

1.2 门控循环神经单元

GRU与LSTM的结构相似但是结构更简单，张量操作更少。它引入了重置门和更新门的概念，从而修改了循环神经网络中隐藏状态的计算方式，图2所示为GRU的记忆单元结构。GRU通过直接在当前网络的状态$\boldsymbol{h}_t$和上一时刻网络的状态$\boldsymbol{h}_{t-1}$之间添加一个线性的依赖关系，来解决梯度消失和梯度爆炸的问题，表达式为

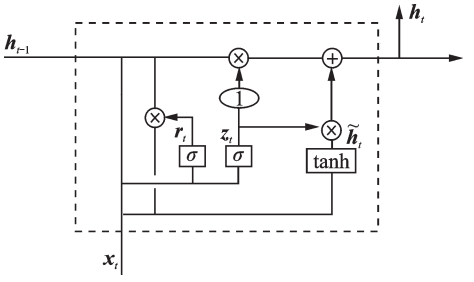

图2 GRU记忆单元

Fig.2 GRU memory cell

$$\boldsymbol{r}_t=\sigma\left(\boldsymbol{W}_{xr}\boldsymbol{x}_t+\boldsymbol{W}_{hr}\boldsymbol{h}_{t-1}+\boldsymbol{b}_r\right) \tag{7}$$

$$\boldsymbol{z}_t=\sigma\left(\boldsymbol{W}_{xh}\boldsymbol{x}_t+\boldsymbol{W}_{hz}\boldsymbol{h}_{t-1}+\boldsymbol{b}_z\right) \tag{8}$$

$$\tilde{\boldsymbol{h}}_t=\tanh\left(\boldsymbol{W}_{xh}\boldsymbol{x}_t+\boldsymbol{W}_{hh}\left(\boldsymbol{r}_t\odot\boldsymbol{h}_{t-1}\right)+\boldsymbol{b}_h\right) \tag{9}$$

$$\boldsymbol{h}_t=\boldsymbol{z}_t\odot\boldsymbol{h}_{t-1}+\left(1-\boldsymbol{z}_t\right)\odot\tilde{\boldsymbol{h}}_t \tag{10}$$

式中：$\boldsymbol{r}_t$、$\boldsymbol{z}_t$、$\boldsymbol{x}_t$分别表示重置门、更新门和输入向量；$\odot$表示Hadamard Product，也就是操作矩阵中对应的元素相乘；其他变量含义与LSTM网络相同。

2 融合LSTM-GRU网络的检测系统

2.1 LSTM-GRU网络结构

LSTM通过自身的3个门控装置来控制数据信息在网络间的流通并以此解决长时依赖问题,但是由于LSTM网络设置的参数过多,每1个细胞里面都有4个全连接层,在实际应用过程中,如果时间跨度较大而LSTM网络层次又深则会容易出现过拟合现象,并且对计算机的运算能力要求也较大。GRU为LSTM的简化,它引入了更新门和重置门来处理数据信息,相比于LSTM设置的参数更少,减少过拟合风险,但是在处理大数据集的情况下表现不如LSTM。在此本文将两种网络结构进行串联处理,提出一种融合LSTM-GRU的网络结构。

LSTM-GRU网络是由单层LSTM网络及单层GRU网络串联形成的一种混合网络结构,如图3所示。数据输入LSTM层后依次通过输入门、输出门和遗忘门,使用sigmoid函数和tanh函数进行信息的更迭处理后进入GRU层;GRU层中的更新门和重置门对信息进行矩阵相乘处理,输入到Dropout层,丢弃一些神经节点防止过拟合;随后进行归一化处理,再输入到全连接层;最后通过使用softmax函数的分类层进行真假语音分类。

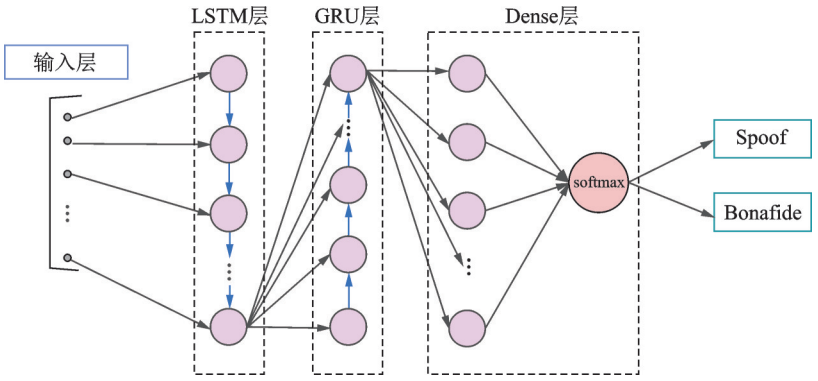

图3 LSTM-GRU网络结构

Fig.3 LSTM-GRU network structure

2.2 检测系统的评价指标

评价语音欺骗检测性能的常用指标是EER。EER是错误拒绝率(False rejection rate, FRR)和错误接受率(False acceptance rate, FAR)相等时的数值。EER是衡量生物识别系统性能的重要指标,能够同时反映出系统的安全性和准确性[19]。FRR、FAR、EER的计算表示为

$$\mathrm{FRR}(\theta)=\frac{\mathrm{num}[s]<\theta}{N_{\mathrm{bonafide}}} \tag{11}$$

$$\mathrm{FAR}(\theta)=\frac{\mathrm{num}[s]>\theta}{N_{\mathrm{spoofed}}} \tag{12}$$

$$\mathrm{EER}=\mathrm{FRR}\left(\theta_{\mathrm{EER}}\right)=\mathrm{FAR}\left(\theta_{\mathrm{EER}}\right) \tag{13}$$

式中:N_{bonafide}、N_{spoofed}分别表示真语音的总数及假语音的总数;$\mathrm{num}[s]<\theta$表示攻击样本中得分小于θ的数量;$\mathrm{num}[s]>\theta$表示攻击样本中得分大于θ的数量。当EER数值越小,反映其系统性能越好。

AUC(Area under the curve)是机器学习常用的二分类评测手段,指的是ROC(Receiver operating

characteristic)曲线下的面积[20]。ROC曲线通过真正例率与假正例率两项指标,可以用来评估分类模型的性能。AUC的计算公式为

$$\mathrm{AUC}=\frac{\sum_{i\in \mathrm{positiveClass}}\mathrm{rank}_i-\frac{M(1+M)}{2}}{M\times N} \tag{14}$$

式中:rank_i代表第i条样本的序号;M、N分别代表正样本的个数和负样本的个数。ROC曲线下的面积介于0.1和1之间;AUC越接近于1说明模型越好。

2.3 特征提取

本文选取梅尔频率倒谱系数(Mel-frequency cepstral coefficients, MFCC)作为训练神经网络特征。MFCC考虑了人耳对不同频率的感受程度[21],在语音信号处理领域应用广泛,其提取过程如图4所示。

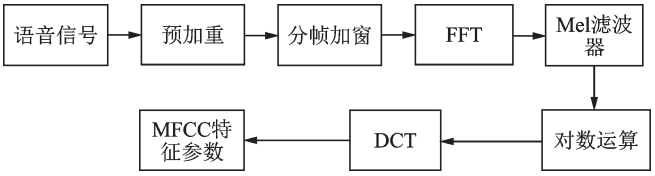

图4 MFCC提取过程

Fig.4 MFCC extraction process

3 逻辑访问攻击检测实验

3.1 实验环境

本文基于Ubuntu18.04.4LTS系统,使用Jupyter Notebook软件运行环境,Tensorflow2.2框架,硬件配置采用Intel Xeon(R) Gold 6132 CPU处理器,NVIDIA Tesla P4显卡。

3.2 数据库

本文针对语音合成及语音转换两种语音逻辑访问攻击的欺骗方法进行检验,采用ASV spoof 2019数据集中的LA数据库。该数据库是基于VCTK数据库进行开发的,划分为3个子集:训练集、开发集和验证集,本文采用训练集进行实验。训练集由20名(8男12女)不同说话人组成,采样率为16 kHz,共计23 580个音频文件。提取MFCC特征后从特征集中随机选取60%(25 345个)特征数据作为本次实验的训练集,20%(8 449个)特征数据作为本次实验的验证集,20%(8 449个)特征数据作为本次实验的测试集。

3.3 实验参数设置

实验中提取MFCC作为训练神经网络的语音特征。在语音提取过程中,MFCC的特征维度设置为20维,选择二维离散余弦变换,每50帧语音为特征长度组成1个序列。

在神经网络模型的选择上采用GRU、LSTM和LSTM-GRU混合模型分别对提取到的MFCC特征进行对比实验。实验控制单一变量,LSTM-GRU设置的网络参数及结构如表1所示。设置的LSTM-GRU网络第1层为LSTM层,具有64个隐藏节点,输入数据的维度为20维;第2层为具有128个隐藏节点的GRU层,用来将信息传递到下一层,激活函数为Relu;第3层使用了Dropout,随机丢弃50%用来防止过拟合;第5层为Batch normalization,减少网络计算量使其学习率更稳定地进行梯度传

播;第5层为全连接层,含有128个隐藏节点;第6层为分类层,激活函数为softmax。网络的迭代周期分别设置为400、1 000,batch-size对应分别设置为128、256,即网络一次训练128或256个数据。学习率的设定使用指数衰减法,初始学习率设置为0.01,衰减系数为0.96,衰减速度为100,优化器使用adam ,通过梯度衰减学习率可以使模型更稳定运行。

表1 LSTM-GRU网络结构参数

Table 1 LSTM-GRU network structure parameters

Layer(type)	Output shape	Parameter
LSTM	(None,None,64)	21 760
GRU	(None,128)	74 112
Dropout	(None,128)	0
Batch normalization	(Batch(None,128))	512
Dense	(None,256)	33 024
Dense	(None,2)	514

3.4 实验结果及分析

GRU、LSTM及LSTM-GRU三种网络模型分别在训练周期为400、1 000下对提取到的MFCC特征训练结果如表2、3所示。结果分析所用评价指标为EER、AUC和准确度。

由表2、3可以看出,在准确度上3种模型均有不错的效果,其中LSTM-GRU模型所达到的准确度最高分别为98.12%和97.96%;在AUC指标上LSTM-GRU表现也超过GRU和LSTM。在等错误率表现上LSTM-GRU网络模型最低,表现最优分别为5.9%和7.1%。通过比较这3种模型的各项评判指标可以发现,训练周期为400时,三者都比ASV2019挑战赛所提供的基线系统EER=8.09%要低。其中GRU比基线系统低17.18%;LSTM比基线系统低17.18%;LSTM-GRU比基线系统低27.07%。训练周期为1 000时GRU比基线系统低3.58%;LSTM表现较差;LSTM-GRU比基线系统低12.2%。由此可以得出在GRU、LSTM和LSTM-GRU三种网络中LSTM-GRU网络表现最佳。

表2 训练周期400下3种模型实验结果

Table 2 Experimental results of three models under 400 epochs %

模型	GRU	LSTM	本文模型
准确度	97.83	98.05	98.12
AUC	93.30	93.30	94.10
EER	6.70	6.70	5.90

表3 训练周期1 000下3种模型实验结果

Table 3 Experimental results of three models under 1 000 epochs %

模型	GRU	LSTM	本文模型
准确度	97.38	89.51	97.96
AUC	92.30	50.40	92.90
EER	7.80	49.60	7.10

比较两种周期对3种模型的结果影响可以发现:在周期为400时GRU、LSTM及LSTM-GRU这3种网络模型的结果均比周期为1 000条件下的要好。训练周期为400下的GRU、LSTM-GRU的等错误率分别比训练周期为1 000的等错误率低14.1%、16.9%。可以看出这3种模型在相对较小的训练周期下能够达到更好的训练结果。

在训练周期为400次时GRU和LSTM的表现相近,LSTM网络在准确度上比GRU略高,LSTM-GRU网络较前两者的表现都更加优秀,等错误率比前两者分别低11.94%、11.94%,AUC指标分别比前两者高0.85%和0.85%。在训练周期为1 000时LSTM表现差,准确率低,AUC为50.4%,EER为49.6%并出现过拟合的现象。而GRU及LSTM-GRU均表现稳定且LSTM-GRU性能优于GRU,EER比GRU低8.97%,AUC比GRU高0.65%。在进行周期长、数据多的情况下,LSTM-GRU比GRU、LSTM表现都更好,其稳定性好,准确度高。

图5为2×2的混淆矩阵,能够清晰地显示LSTM-GRU对真假语音的区分准确率。纵坐标表示真

实标签,横坐标表示预测标签。图中数值表示预测值被归为某一类的比例,位于对角线上的数值越大表示有越多的序列被正确归类。图中所示:欺骗语音有99%被正确归类,真实语音有89%被正确归类,有0.01%的欺骗语音和11%的真实语音被错误分类。

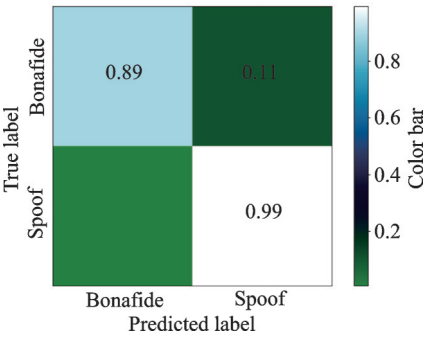

图5 混淆矩阵

Fig.5 Confusion matrix

图6、7为LSTM-GRU网络模型、周期为400训练过程的识别准确度变化曲线及损失大小变化曲线,为每次处理完128个数据的分类准确度及训练损失大小变化,得到交叉熵损失函数值。可以看出,在迭代50个周期后,准确度变化曲线及损失大小变化曲线进入收敛状态,识别准确率训练集稳定在100%附近,测试集准确率稳定在98%附近。交叉熵损失函数值训练集稳定0%附近,测试集稳定在0.075%附近,测试结果准确率为98.12%。说明LSTM-GRU网络对于欺骗语音检测具有良好的潜力,适用于大规模数据库,同时也反映出该网络模型不容易出现梯度爆炸或梯度消失具有稳定性。

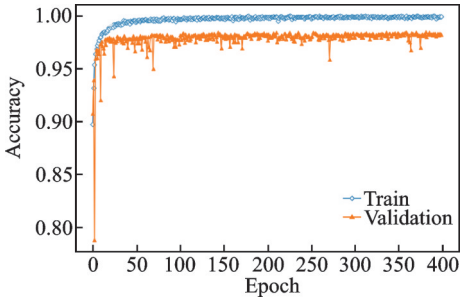

图6 训练过程中准确度变化曲线

Fig.6 Accuracy curves during training

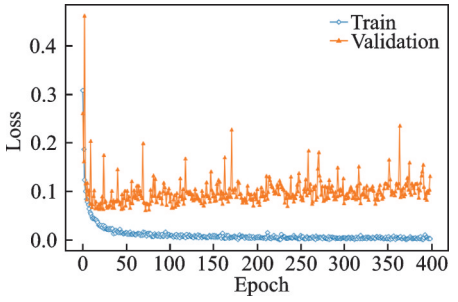

图7 训练过程中损失大小变化曲线

Fig.7 Loss curves during training

在实际应用中模型的运算量及运算速度十分重要。为验证模型的快速准确性,在训练周期为400下将3种模型的参数量、训练速度及测试所用时长进行比较。每次调用程序前在终端使用kill PID命令释放GPU内存保证运行环境一致,同时网络参数设置不变保证变量唯一,实验结果如表4所示。

表4 训练周期400下3种模型运算性能比较

Table 4 Comparison of operation performance of three models under 400 epochs

模型	GRU	LSTM	本文模型
参数量	125 058	154 626	130 306
每个周期训练时长/s	11.07	9.82	10.75
分类时长/s	1.17	1.09	1.04

由表4可看出本文模型的参数量和训练每个周期所费时长均介于GRU和LSTM网络之间,说明本文方法的运算量和损耗时间处于合理范围内,在应用模型进行真假语音分类过程中本文模型耗时最短。综上所述,本文提出的融合LSTM-GRU网络在语音逻辑访问攻击检测任务中能够快速准确地识别伪造语音。

4 结束语

本文提出了一种融合LSTM-GRU网络的语音逻辑访问攻击检测方法。通过比较GRU、LSTM与LSTM-GRU这3种网络模型在ASVspoof2019逻辑访问数据库上的表现可见,基于LSTM-GRU网络

的等错误率在设置的两种实验条件下分别为5.9%、7.1%,准确度分别为98.12%、97.96%,在3种网络模型中表现最好。实验中设置训练周期分别为400和1 000,通过比较3种模型在相对长训练周期下的表现,发现LSTM-GRU抗过拟合性强、准确率高。比较3种网络的运算性能并结合LSTM-GRU模型的训练情况,发现该网络模型不容易出现梯度爆炸或梯度消失,具有良好的稳定性,能够快速准确地对真假语音进行分类,可适用于大规模数据库。LSTM-GRU网络可为语音逻辑访问攻击检测提供新的方法。

参考文献:

[1] 张雄伟,李嘉康,孙蒙,等.语音欺骗检测方法的研究现状及展望[J].数据采集与处理,2020,35(5):807-823.
ZHANG Xiongwei, LI Jiakang, SUN Meng, et al. Speech anti-spoofing:The state of the art and prospect[J]. Journal of Data Acquisition and Processing,2020,35(5):807-823.

[2] WU Z, KINNUNEN T, EVANS N, et al. ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge[J]. IEEE Journal on Selected Topics in Signal Process, 2017, 11(4): 588-604.

[3] KINNUNEN T, SAHIDULLAH M, DELGADO H, et al. The ASVspoof 2017 challenge: Assessing the limits of replayspoofing attack detection[C]//Proceedings of Conference of the International Speech Communication Association (INETSPEECH). Stockholm, Sweden: [s.n.], 2017: 20-24.

[4] TODISCO M, WANG X, VESTMAN V, et al. ASVspoof 2019: Future horizons in spoofed and fake audio detection [EB/OL]. (2019-04-09) [2020-03-20]. https://arxiv.org/abs/1904.05441.

[5] 潘孝勤, 芦天亮, 杜彦辉, 等. 基于深度学习的语音合成与转换技术综述[J]. 计算机科学, 2021, 48(8): 200-208.
PAN Xiaoqin, LU Tianliang, DU Yanhui, et al. Overview of speech synthesis and voice conversion technology based on deep learning[J]. Computer Science, 2021, 48(8): 200-208.

[6] REYNOLDS D A, QUATIERI T F, DUNN R B. Speaker verification using adapted Gaussian mixture models[J]. Digital Signal Processing, 2000, 10(1/2/3): 19-41.

[7] CURELARU F. Evaluation of the standard i-vectors based speaker verification systems on limited data[C]//Proceedings of 2018 International Conference on Communications (COMM). [S.l]: IEEE, 2018: 101-106.

[8] VILLALBA J, MIGUEL A, ORTEGA A, et al. Spoofing detection with DNN and one-class SVM for the ASVspoof 2015 challenge[C]//Proceedings of Sixteenth Annual Conference of the International Speech Communication Association. [S.l]: [s.n.], 2015.

[9] LAVRENTYEVA G, NOVOSELOV S, MALYKH E, et al. Audio replay attack detection with deep learning frameworks [C]//Proceedings of Conference of the International Speech Communication Association(INETSPEECH). [S.l]: [s.n.], 2017: 82-86.

[10] GOMEZ-ALANIS A, PEINADO A M, GONZALEZ J A, et al. A deep identity representation for noise robust spoofing detection[C]//Proc Interspeech.[S.l]: [s.n.], 2018: 676-680.

[11] GOMEZ-ALANIS A, PEINADO A M, GONZALEZ J, et al. A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection[C]//Proc Interspeech.[S.l]: [s.n.], 2019: 1068-1072.

[12] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436.

[13] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[14] LI J, ZHANG X, SUN M, et al. Attention-based LSTM algorithm for audio replay detection in noisy environments[J]. Appl Sci, 2019, 9: 1539.

[15] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[EB/OL]. (2014-08-10) [2021-02-22]. https://arxiv.org/abs/1406.1078.

[16] CHEN Z, ZHANG W, XIE Z, et al. Recurrent neural networks for automatic replay spoofing attack detection[C]//

Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l]: IEEE, 2018: 2052-2056.

[17] SHEWALKAR A, NYAVANANDI D, SIMONE A. Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU[J]. Journal of Artificial Intelligence and Soft Computing Research, 2019, 9(4): 235-245.

[18] CHUNG J, GULCEHRE C, CHO K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv 2014[EB/OL]. (2014-08-15) [2021-02-22]. https://arxiv.org/abs/1412.3555.

[19] SINGH N, KHAN R A, RAJSHRE E. Equal error rate and audio digitization and sampling rate for speaker recognition system [J]. Journal of Computational and Theoretical Nanoscience, 2014, 20(5): 1085-1088.

[20] JIN H, LING C X. Using AUC and accuracy in evaluating learning algorithms[J]. IEEE Transactions on Knowledge & Data Engineering, 2005, 17(3): 299-310.

[21] DAVIS S V, MERMELSTEIN P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences[J]. IEEE Trans, Acoust, Speech, Signal Process, 2003, 28(4): 357-366.

作者简介:

杨海涛(1998-),男,硕士研究生,研究方向:深度学习、语音处理,E-mail: 724862715@qq.com。

王华朋(1979-),**通信作者**,男,教授,研究方向:说话人识别、深度学习,E-mail:huapeng.wang@hotmail.com。

牛瑾琳(1997-),女,硕士研究生,研究方向:深度学习、语音处理。

楚宪腾(1999-),男,硕士研究生,研究方向:深度学习、说话人识别。

林暖辉(1975-),男,高级工程师,研究方向:语音处理、声纹鉴定。

(编辑:刘彦东)