计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(1):322−326 [doi: 10.15888/j.cnki.csa.008246]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于条件生成式对抗网络的情感语音生成模型①

崔新明, 贾　宁, 周洁美慧

(大连东软信息学院 计算机与软件学院, 大连 116023)
通信作者: 贾　宁, E-mail: jianing@neusoft.edu.cn

摘　要: 提出了一种基于条件生成对抗网络的情感语音生成技术, 在引入情感条件的基础上, 通过学习语音库中的情感信息, 能够自主生成全新的富有指定情感的语音. 生成式对抗网络是由一个判别网络和一个生成器组成. 使用TensorFlow 作为学习框架,利用条件 GAN 模型对大量情感语音进行训练, 利用语音生成网络 G 和生成网络 D 构成动态“博弈过程”, 更好地学习观测语音情感数据的条件分布. 其生成样本接近原始学习内容的自然语音信号, 具有多样性, 而且能够逼近符合真实情感的语音数据. 所提出的解决方案在交互式情绪二进制动作捕捉 IEMOCAP 语料库和自建情感语料库上进行评估, 并且与现有情感语音生成算法相比显示出提供更准确的结果.

关键词: 条件生成式对抗网络; 条件 GAN 模型; 情感判别; 语音生成模型; TensorFlow 框架

引用格式: 崔新明,贾宁,周洁美慧.基于条件生成式对抗网络的情感语音生成模型.计算机系统应用,2022,31(1):322–326. http://www.c-s-a.org.cn/1003-3254/8246.html

Speech Generation Model Based on Conditional Generative Adversarial Network

CUI Xin-Ming, JIA Ning, ZHOU Jie-Mei-Hui

(School of Computer and Software, Dalian Neusoft Institute of Information, Dalian 116023, China)

Abstract: An affective speech generation technology based on a conditional generative adversarial network (GAN) is proposed in this study. After the introduction of affective conditions and the learning of affective information from the phonetic database, a brand new affective speech with specified emotions can be generated independently. GAN is composed of a discrimination network and a generator. With TensorFlow as the learning framework, the conditional GAN model is employed to train plenty of affective speech, and the speech generation network G and generation network D are used to form a dynamic “game process” for better learning and observation of the conditional distribution of speech emotion data. The generated sample is close to the natural speech signal of the original learning content, which has diversity and can approximate the speech data consistent with the real emotion. The proposed solution is evaluated on the interactive emotional dyadic motion capture (IEMOCAP) corpus and the self-built emotional corpus. It generates more accurate results than the existing affective speech generation algorithms.

Key words: conditional generative adversarial network (GAN); conditional GAN model; emotion discrimination; speech generation model; TensorFlow framework

在人与人或人与计算机的交互中, 话语的表达方式传递着重要的副语言信息, 特别是蕴含着与潜在情感有关的信息. 因此, 需要现代言语分析系统能够分析这种与情感相关的非语言维度, 以及话语本身的信息, 以适应更好的人机交互操作. 近年来, 自动识别口语情感内容和情感语音对话等技术引起了越来越多研究人员的关注. 语音情感识别和对话均是典型的有监督的音频任务, 它们均涉及低级音频特征映射到具有不同

① 基金项目: 辽宁省教育厅校际合作项目 (86896244); 大连市科技计划 (2019RQ120)
收稿时间: 2021-03-09; 修改时间: 2021-04-07, 2021-04-20; 采用时间: 2021-04-26; csa 在线出版时间: 2021-12-17

情感的高级类标签或情感维度的映射. 因此, 带精确标注的数据集在构建和评估语音情感识别系统中是非常重要的. 然而, 在现实生活中, 真实表达情感的开源语料库少之又少, 而且大多数语料库使用的是非汉语语言.

基于此, 本文提出了一个基于条件生成对抗网络 (conditional generative adversarial networks, 条件 GAN) 的语音生成技术, 用于合成海量的汉语情感语音, 并实现情感的精确表达, 在实验过程中, 通过情感识别模型验证了生成情感语音的有效性.

1 相关工作

语料库规模不足或高度倾斜的数据是语音生成和语音识别过程中的一个常见问题. 在数据采集和标注过程中, 中性语音样本的使用频率远高于含有情感的语音样本, 或者存在大量情感表达有歧义的语音, 这些问题均导致数据集出现高度不平衡的现象. 解决数据不平衡的一种常见方法是使用数据增强技术[1]. Wong 等人[2] 利用过采样和变换生成数据空间合成样本, 有利于数据空间的扩展以进行数字分类. Schluter 等人[3] 评估了七种不同的数据增强技术, 用于检测语音谱图中的情绪表达, 发现音高的偏移量和随机频率的滤波是最有效的情感表达. 此外, 研究表明音调增强有利于环境声音分类[4] 和音乐流派分类[5] 等任务. 对于语音生成等任务, Aldeneh 等人[6] 对原始信号应用了语速的微调, 证明了情感表达的有效性.

除了传统的数据增强方法, 生成对抗网络 (generative adversarial networks, GAN) 作为一种强大的生成模型, 其应用范围越来越广. GAN 通过同时训练两个相互竞争的网络 (一个生成器和一个鉴别器) 来近似数据分布. GAN 的计算流程与结构如图 1 所示. 许多研究集中在提高生成样本的质量和稳定 GAN 训练等领域中[7].

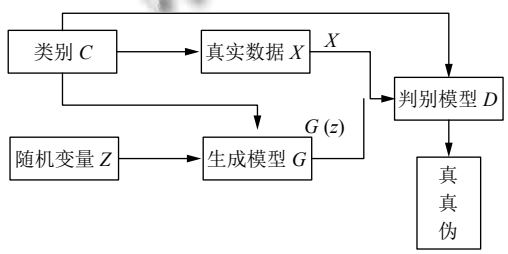

图 1 GAN 的计算流程与结构

生成性对抗网络已经成功地应用于各种计算机视觉任务以及与语音相关的应用, 如语音增强[8] 和语音转换[9]. Sahu 等人部署对抗式自动编码器在压缩特征空间中表示语音, 同时保持情感类之间的区分性[10]. Chang 和 Scherer 利用一个深层卷积 GAN 以半监督的方式学习情感言语的区别表示[11]. Han 等人提出了一个由两个网络组成的条件对抗训练框架. 一个学习预测情绪的维度表示, 而另一个旨在区分预测和数据集的真实标签[12], 其识别精度可以和传统的合成网络相当.

最近, 基于 GAN 的强大合成能力, 许多研究人员开发了 GAN 的变换版本, Antoniou 等人[13] 训练了一个生成类内样本的 GAN. Zhu 等人[14] 设计了 CycleGAN 架构适用于面部表情的情感分类. 对于语音领域, Sahu 等人[15] 综合特征向量用于提高分类器在情感任务中的性能. Mariani 等人[16] 提出了一种条件 GAN 结构来解决数据不平衡问题. 基于上述现有的研究基础, 本文选择使用条件对抗生成网络实现对于情感语音的合成, 其中涉及权重学习和条件 GAN 的微调过程, 此外在实验中设计了一个判别模型用于验证合成情感的有效性.

2 基于条件 GAN 的语音生成模型设计

作为人机交互系统的重要功能之一, 本文在生成对应文本信息的基础上, 针对个体用户的语音模型, 以说话者的低级描述符特征标签为条件, 设计条件对抗生成网络模型生成语音.

在生成过程中, 生成模型的设计起到决定性的作用, 传统的生成模型需要预先获得一个标准模型才可进行数据的生成, 该模型的获取较为困难, 且容易出现误差, 可以采用具有学习和模仿输入数据分布的 GAN 模型实现目标.

GAN 由一个鉴别器和一个发生器组成, 它们协同工作, 以学习目标的基础数据分布, 这种无需预先建模的方法, 对于较大的数据是不可控的. 为了解决 GAN 太过自由这个问题, 可以为 GAN 加一些约束, 即一种带条件约束的 GAN, 在生成模型 (G) 和判别模型 (D) 的建模中均引入条件变量, 指导数据生成过程.

本文利用条件 GAN 模型生成语音情感识别的生成特征向量. 它学习以标签或其所属的情感类为条件 y 的高维特征向量的分布. 在生成网络的基础上, 输入增加了一个情感约束条件 $p(z)$, 输出是一个综合打分或者输出两个分数, 分别表示真实与条件 GAN 的相符程度. 具体网络结构如图 2 所示.

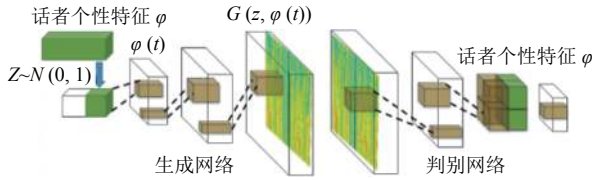

图 2 条件 GAN 模型网络结构

条件 GAN 的目标函数是带有条件概率的二人极小极大值博弈. 如式 (1) 所示.

$$\min_G \max_D V(D,G) = E_{x\sim pdata(x)}[\log D(x|y)] + E_{z\sim p_z(z)}[\log(1-D(G(z|y)))] \quad (1)$$

同一般形式的 GAN 类似, 条件 GAN 也是先训练判别网络, 再训练生成网络, 然后循环此过程, 即两个网络交替的完成训练. 训练判别网络的样本与之前的样本稍有不同, 此时需要这 3 种样本, 分别是:

(1) 与条件相符的真实语音, 期望输出为 1;

(2) 与条件不符的真实语音, 期望输出为 0;

(3) 生成网络生成的输出, 期望输出为 0.

如图 3 所示, 本文中条件 GAN 中生成数据生成以标签为条件. 给定一组数据点及其对应的标签, 真实数据的类信息均为 one-hot 编码. 条件 GAN 学习条件分布, 将含有噪声的语音以及情感条件通过生成器, 生成增强后的语音, 增强后的语音和相似情感的语音通过判别器, 判别器的功能是找出语音的正确类别, 并实现对语音的标记, 生成器的目标则是让判别器实现对于生成语音的合理判别, 在实现过程中, 双方不断的微调各自的结果, 最终达到一种平衡状态, 即生成器生成的语音可以真实的模拟情感, 而判别器可以从语音中识别出的目标情感. 基于此, 可以获得与相似情感语音的情感描述相同的生成结果.

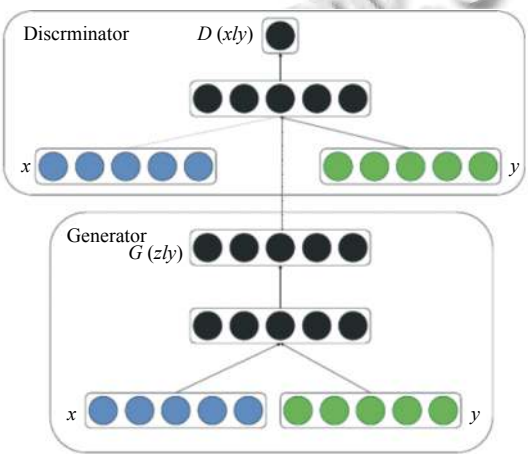

图 3 条件 GAN 生成模型

在条件 GAN 模型初始化时, 需要将学习到的权值传输到 GAN 模块, 编码器权值分别传输到判别器和解码器权值传输到生成器. 对于类条件, 需要计算与每个类的图像相对应的自动编码器的学习潜在向量的均值和协方差矩阵. 此时可用多元正态分布对每一类进行建模, 然后, 从一个特定类的分布中随机抽取一个潜在向量, 并将其作为输入提供给生成器, 因此条件 GAN 需具有明确的类别信息.

另一个关键技术是条件 GAN 的微调过程, 即条件 GAN 是使用少数和大多数类别的训练数据进行微调的. 通过这种方式, 学习类之间共享的特性. 这些特征有助于为少数类创造定性的条件. 在微调期间, 生成器从类条件潜在向量生成器中提取的潜在特征作为输入. 后者以均匀分布的类标签作为输入. 然后, 将一批真实语音和生成的语音转发给判别器. 两个竞争网络中的每一个的目标是使用交叉熵优化其损失函数, 对判别器进行优化后, 使真实语音与正确的类标签相匹配, 生成的语音与假标签相匹配. GAN 微调后, 使用生成器分别为每个类别生成人工情感语音, 以接近大多数类别的情感真实表达.

3 实验及相关结果

本文采用 IEMOCAP 和自建情感语料库分别进行实验分析.

IEMOCAP 数据集包含两个演员之间的对话记录, 数据的总量是来自 10 个发言者的 12 小时音频和视频信息, 其中注释了 11 类情感标签 (愤怒, 幸福, 悲伤, 中立, 惊讶, 恐惧, 沮丧和兴奋) 和尺寸标签 (激活和效价的值) 从 1 到 5). 本文事先完成了数据过滤步骤: 保留的样本中, 其中至少两个注释者就情感标签达成一致, 丢弃话语注释不同的情绪的样本, 并仅使用注释为中性, 愤怒, 高兴和悲伤的样本, 产生 4 490 个样本 (愤怒的 1 103, 高兴的 595, 中性的 1 708 和悲伤的 1 084), 其中使用 4 个会话进行训练, 余下的 1 个会话用于验证和测试.

自建情感语料库是由本校 60 余名学生及教师录制的语音样本集合, 数据总量为包含 40 条文字样本的 21 000 条音频数据, 每个音频时长约 2–4 s, 均采用中性, 愤怒, 高兴和悲伤 4 种情感进行录制. 每类的样品数量较为均衡. 与 IEMOCAP 数据集的处理方法类似, 只保留至少两个注释者的情感标签一致的数据, 同时

使用 5 折交叉验证的方式进行测试.

对于 IEMOCAP 的评估实验, 我们使用 5 折交叉验证, 即使用 4 个会话进行训练, 使用 1 个会话进行测试. 此设置是相关 SER 库中 IEMOCAP 的常见做法. 我们使用 80%–20% 的切分策略, 分别进行训练和验证.

此处选择深度学习方法建立情感特征验证模型, 使用 TensorFlow 作为开发框架. 此处采用了双向 3 层的长短期记忆网络 (long-short term memory, LSTM) 模型, 双向是指存在两个传递相反信息的循环层, 第 1 层按时间顺序传递信息, 第 2 层按时间逆序传递信息. 它意味着过去和未来的信息均可以成功捕获, 这是由于情感表达的时序因素可以由当前时刻的前后若干帧的信息共同决定. 因此按照上述思路设计了 3 层双向 LSTM 模型, 利用条件 GAN 模型获得的新语音进行模型的训练和参数学习. 此模型的结构如图 4 所示.

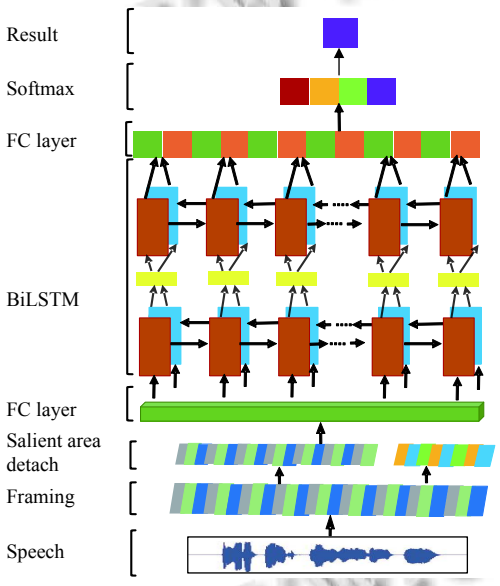

图 4 语音情感验证模型

设计相关的实验验证其有效性, 本节将对比情感语音的识别效果. 采用加权精度 (weighted accuracy, WA) 和未加权精度 (unweighted accuracy, UA) 作为指标, 采用图 4 中描述的情感分类模型进行测试. IEMOCAP 语料库和自建语料库的测试结果见表 1 和表 2.

表 1 IEMOCAP 测试结果 (%)

模型	WA	UA
基线: 原始语音	73.4	74.2
GAN模型生成语音	72.1	72.5
条件GAN模型生成语音	73.2	73.8

表 2 自建语料库测试结果 (%)

模型	WA	UA
基线: 原始语音	70.3	70.8
GAN模型生成语音	69.6	70.8
条件GAN模型生成语音	70.6	70.2

从 IEMOCAP 语料库和自建语料库的测试结果可以看出, 使用条件 GAN 模型生成语音的情感识别效果高于使用 GAN 模型生成语音的结果, 而且, 它们的测试精度与原始语音的精度相当, 由此可以得出结论, 当前的条件 GAN 模型可以生成与原始语音信号表达情感相似的信号.

进一步对比验证生成语音与原始语音的相似度及可用性. 分别使用原始语音、生成语音和上述语音集合作为训练数据, 采用生成语音、原始语音和混合语音作为测试数据, 搭建如图 4 所示的情感分类模型. IEMOCAP 语料库和自建语料库的测试结果见表 3 和表 4.

表 3 IMOCAP 测试结果 (%)

训练数据	测试数据	WA	UA
原始语音	生成语音	73.6	74.1
生成语音	原始语音	74.3	74.8
原始语音+生成语音	原始语音+生成语音	74.6	75.1

表 4 自建语料库测试结果 (%)

训练数据	测试数据	WA	UA
原始语音	生成语音	71.3	71.8
生成语音	原始语音	71.6	71.2
原始语音+生成语音	原始语音+生成语音	72.8	73.2

从测试结果可以看出生成语音与原始语音相似且可用, 使用生成语音与原始语音的集合可以进一步提升语音情感识别的有效性.

最后, 将原始数据与生成数据结合进一步对情感语音分类, 现有方法测试结果如表 5, 自建语料库模型测试结果如表 6. 由测试结果可以看出自建语料库后对情感语音进行分类, 并对比现有方法, 自建语料库后的模型相比现有方法识别精度有所提升.

表 5 现有方法测试结果

情感	数量	UA (%)
中性	1 528	70.9
愤怒	1 049	70.6
高兴	496	71.0
悲伤	1 254	70.7

表 6 自建语料库测试结果

情感	数量	UA (%)
中性	1 728	73.4
愤怒	1 149	73.1
高兴	596	73.0
悲伤	1 154	73.3

4 结论与展望

本文利用条件 GAN 模型对大量情感语音进行训练, 其生成样本接近原始学习内容的自然语音信号的情感表达, 通过实验表明, 使用 IEMOCAP 语料库和自建语料库获得的新语音具有与原语音相匹配的情感表达.

在未来的研究过程中, 作者将进一步扩充现有的情感语料库, 同时进一步改进条件 GAN 模型, 以提升生成新语音的情感表达效果.

参考文献

1 Cubuk ED, Zoph B, Mane D, *et al*. AutoAugment: Learning augmentation policies from data. arXiv: 1805.09501v1, 2018.

2 Wong SC, Gatt A, Stamatescu V, *et al*. Understanding data augmentation for classification: When to warp? arXiv: 1609.08764, 2016.

3 Schlüter J, Grill T. Exploring data augmentation for improved singing voice detection with neural networks. Proceeding of International Society for Music Information Retrieval Conference. Malaga, 2015. 121–126.

4 Salamon J, Bello JP. Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Processing Letters, 2017, 24(3): 279–283. [doi: 10.1109/LSP.2017.2657381]

5 Aguiar LR, Costa YMG, Silla NC. Exploring data augmentation to improve music genre classification with ConvNets. 2018 International Joint Conference on Neural Networks (IJCNN). Rio de Janeiro: IEEE, 2018. 1–8.

6 Aldeneh A, Provost EM. Using regional saliency for speech emotion recognition. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans: IEEE, 2017. 2741–2745. [doi: 10.1109/ ICASSP.2017.7952655].

7 Zhu JY, Park T, Isola P, *et al*. Unpaired image-to-image translation using cycle-consistent adversarial networks. 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017. 2242–2251.

8 Pascual S, Bonafonte A, Serrà J. SEGAN: Speech enhancement generative adversarial network. arXiv: 1703.09452, 2017

9 Kaneko T, Kameoka H. Parallel-data-free voice conversion using cycle-consistent adversarial networks. arXiv: 1711.11293, 2017.

10 Sahu S, Gupta R, Sivaraman G, *et al*. Adversarial auto-encoders for speech based emotion recognition. arXiv: 1806.02146, 2018.

11 Chang J, Scherer S. Learning representations of emotional speech with deep convolutional generative adversarial networks. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans: IEEE, 2017. 2746–2750.

12 Han J, Zhang ZX, Ren Z, *et al*. Towards conditional adversarial training for predicting emotions from speech. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary: IEEE, 2018. 6822–6826.

13 Antoniou A, Storkey A, Edwards H. Data augmentation generative adversarial networks. arXiv: 1711.04340, 2017.

14 Zhu XY, Liu YF, Qin ZC, *et al*. Data augmentation in emotion classification using generative adversarial networks. arXiv: 1711.00648v5, 2017.

15 Sahu S, Gupta R, Espy-Wilson C. On enhancing speech emotion recognition using generative adversarial networks. arXiv: 1806.06626, 2018.

16 Mariani G, Scheidegger F, Istrate R, *et al*. BAGAN: Data augmentation with balancing GAN. arXiv:1803.09655, 2018.