AAEC: An Adversarial Autoencoder-based Classifier for Audio Emotion Recognition

Changzeng Fu*
[1]Osaka University, Japan
[2]Advanced Telecommunications Research Institute International, Japan

Jiaqi Shi
[1]Osaka University, Japan
[2]Advanced Telecommunications Research Institute International, Japan

Chaoran Liu
[1]Advanced Telecommunications Research Institute International, Japan

Carlos Toshinori Ishi
[1]Advanced Telecommunications Research Institute International, Japan
[3]Interactive Robot Research Team, Robotics Project, RIKEN, Japan

Hiroshi Ishiguro
[1]Osaka University, Japan
[2]Advanced Telecommunications Research Institute International, Japan

ABSTRACT

In recent years, automatic emotion recognition has attracted the attention of researchers because of its great effects and wide implementations in supporting humans' activities. Given that the data about emotions is difficult to collect and organize into a large database like the dataset of text or images, the true distribution would be difficult to be completely covered by the training set, which affects the model's robustness and generalization in subsequent applications. In this paper, we proposed a model, Adversarial Autoencoder-based Classifier (AAEC), that can not only augment the data within real data distribution but also reasonably extend the boundary of the current data distribution to a possible space. Such an extended space would be better to fit the distribution of training and testing sets. In addition to comparing with baseline models, we modified our proposed model into different configurations and conducted a comprehensive self-comparison with audio modality. The results of our experiment show that our proposed model outperforms the baselines.

KEYWORDS

audio modality, neural networks, adversarial auto-encoder, emotion recognition

ACM Reference Format:
Changzeng Fu, Jiaqi Shi, Chaoran Liu, Carlos Toshinori Ishi, and Hiroshi Ishiguro. 2020. AAEC: An Adversarial Autoencoder-based Classifier for Audio Emotion Recognition. In *1st International Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop (MuSe'20), October 16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3423327.3423669

*email: changzeng.fu@irl.sys.es.osaka-u.ac.jp; shi.jiaqi@irl.sys.es.osaka-u.ac.jp

MuSe'20, October 16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-8157-4/20/10…$15.00
https://doi.org/10.1145/3423327.3423669

1 INTRODUCTION

In human social interaction, emotion has a great influence on purpose expression, intention understanding, and decision-making. In recent years, automatic emotion recognition has attracted the attention of researchers due to its important role and wide application in supporting humans' activities [3, 15, 20] by tracking user emotions in interactions, generating appropriate responses to users' needs and behaviors, improving the user experience and trustworthiness [17, 20, 25]. With the rapid development and wide use of artificial intelligence technology, a variety of neural network models have been proposed for emotion recognition, especially for speech emotion recognition [4, 8, 10, 24]. Deep learning is used to enable the network to extract emotional features from the data for emotion classification, by using manually extracted features or in an end-to-end manner. Benefiting from the improvement of computing resources and the innovation and optimization of network structure, their performance in emotion recognition tasks has been increasing.

However, due to the limitation of data, sometimes the distribution of training samples cannot cover the test samples, and it is also difficult to estimate the real distribution of the whole data (see Figure 1). This problem will easily result in that the model performing well on the training set, but the prediction accuracy for unknown samples would not be satisfactory, and the ability of robustness and generalization would be poor. Although some methods, e. g. early stopping, dropout, and regularization, have been adopted to alleviate the problem of overfitting, they are still confined by the samples of the training set.

Another solution to overfitting is data augmentation. Data augmentation utilizes the existing training samples to generate new additional data, which can effectively alleviate the overfitting of the model. It can also force the network to learn more robust features and bring stronger generalization ability to the model. The method is widely applied in image-related tasks, where new images are generated by flipping, rotating, scaling, clipping, shifting, etc. This

practice not only plays a key role in increasing training samples on small datasets but also has positive effects on improving the performance of models trained on large datasets [1]. Similarly, for audio data, some data augmentation methods to maintain the validity of the labels are also frequently used, e.g. time stretching, pitch shifting, and noise injection, which do not change the semantics of the labels, so that the network is invariant to these transformations and can be better generalized to unknown data [18].

Recently, Generative Adversarial Networks (GANs) have been widely used to generate extra training samples in speech-related tasks[2, 9, 16]. In these works, the generative model learns a mapping function from input distribution to generation distribution, and the discriminative model tries to judge whether the generated samples come from the real distribution or false distribution. In the training set, a generative distribution is optimized through the game between the generative model and the discriminative model, and the new data is generated from the distribution close to the original samples of the training set.

In this work, we focus on taking the audio feature as input and output the prediction class of arousal and valence. The contributions of this paper are as following:

- We propose a model based on adversarial autoencoder (AAE) [11] and an emotional classifier with an adversarial training strategy. Through the adversarial training process, the augmented samples are generated based on the training data, then fed into the emotional classifier with the original data, which aims to improve the robustness and generalization of the classifier and mitigate the problem of overfitting to a certain extent. The performance of the model exceeds the baseline models.
- We investigate different configurations of the proposed model to find out how it is possible to optimize the generalization. Each configuration has been analyzed to find a method that can not only do the data augmentation but also extend the data space within reason, so that the possible distribution of generated data can not only include the distribution of training samples but also include part of the distribution space of test samples.

The rest of the paper is arranged as follows: In Section 2, we briefly discuss the related work of data augmentation using GAN, and put forward the hypothesis that the distribution of their generated samples is different from ours; Section 3 describes the network structures of AAEC and each module; the experiment and results are described in Section 4; we discuss the results of the experiment in Section 5; finally, we conclude with a brief summary and provide some directions for future research.

2 RELATED WORKS AND HYPOTHESIS

2.1 GAN based audio data augmentation

Data augmentation method based on GAN is used in various audio related tasks, where the practice plays an important role in increasing sample size and improving network performance.

Madhu et al. [9] applied the Generative Adversarial Networks based on WaveGAN [5] for data augmentation in the task of environmental sound classification, and proved that compared with time stretching, pitch shifting, additive background noise and dynamic range compression, the data augmentation method based on GAN has a more active and effective role in improving the classification accuracy.

Chatziagapi et al. [2] modified and improved a conditional GAN architecture (BAGAN [12]) to generate synthetic spectrograms for minority classes in speech emotion recognition tasks and improve the quality of the generated spectrograms. Through experiments, they concluded that the proposed method can solve the problem of data imbalance in speech emotion recognition more effectively than standard augmentation technology.

In the work of Qian et al. [16], the augmentated data from basic GAN and conditional GAN is generated frame by frame based on spectrogram feature level for robust speech recognition. The experimental results showed that the data augmentation method can improve the system performance under all tested noise conditions, and greatly improved the robustness of the system, which is also better than the traditional manual data augmentation strategy.

However, because these models simply fit the distribution of the original data and therefore only generate new samples close to the original sample distribution of the training set, the generalization of the models is still limited.

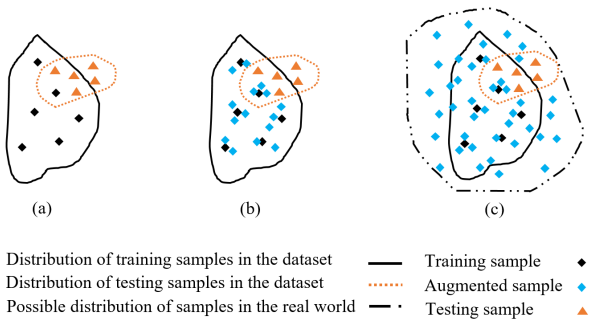

Figure 1: The hypothetical distribution diagram. (a) is the hypothetical data distribution of the dataset; (b) is the hypothetical data distribution after augmentation with existing methods, by which increasing the variety of training samples; (c) is the ideal hypothetical data distribution that extends the boundary of the distribution in the dataset to a possible real distribution.

2.2 Hypothesis

We hypothesize that the distribution of training samples and testing samples in the dataset are intersecting but might not totally overlap with each other due to the limitation of collected data (as shown in Figure 1(a)). For the existing methods, the distribution of the generated feature samples is almost the same as that of the training samples, and therefore it is difficult to generate extra samples in the different distribution space of testing samples with training samples (see Figure 1(b)). On the other hand, if the generator and classifier are trained in each iteration, some approximate samples with distance from the real distribution will be generated when the mapping ability of the generator has not yet converged, and

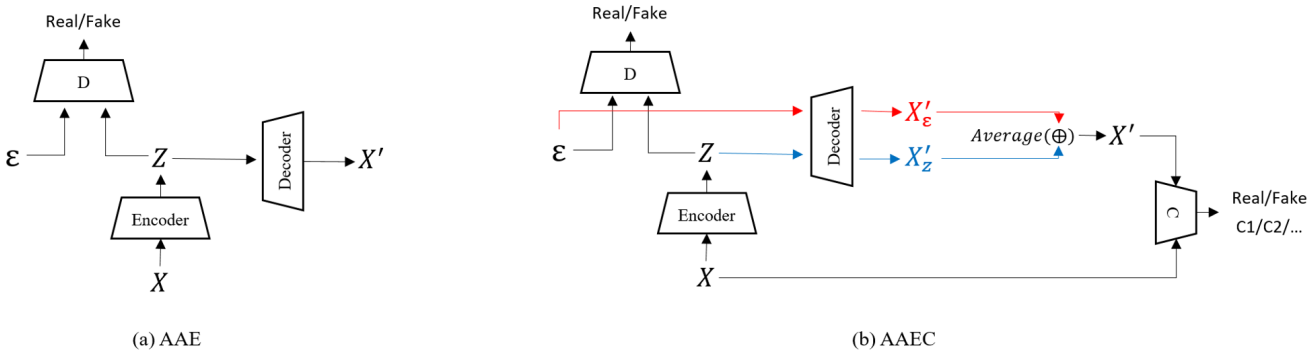

Figure 2: The architecture of the original adversarial autoencoder (AAE) and the proposed adversarial autoencoder-based classifier (AAEC). D and C are annotations for discriminator and classifier respectively. $\mathcal{E}$ indicates the Gaussian noise and Z stands for the extracted hidden feature of X. X' is the reconstructed sample of X yielded by the decoder.

these samples may appear in the area that the original training data cannot cover. In this way, it may be possible to generate data in a larger distribution space while retaining the main features of the training set samples to a certain extent, which can expand the boundary of the generated data distribution and improve the probability of the generated data appearing in the testing data distribution. Ideally, the augmented data will extend the boundary of the data distribution to include both the training data distribution and the test data distribution, which makes the distribution closer to the distribution in the real world (see Figure 1(c)).

3 PROPOSED METHOD

In this section, we demonstrate the structure of our proposed adversarial autoencoder-based classifier (AAEC) and the neural networks' architecture for each module.

3.1 Adversarial Autoencoder-based Classifier

Figure 2 (a) shows the structure of the original AAE [11], which is an autoencoder regularized by matching the aggregated posterior (Z) to an arbitrary prior ($\mathcal{E}$). The *Encoder* encodes X into a representation Z, the *Decoder* works as a generator that generates the reconstructed sample X', the discriminator D distinguishes whether $\mathcal{E}$ or Z comes from Gaussian Distribution, $\mathcal{E}$ for real while Z for fake. The adversarial network that guides Z to match $\mathcal{E}$, meanwhile, attempts to minimize the reconstruction error between X' and X. Speaking of our proposed AAEC, we made some modifications based on AAE (see Figure 2 (b)) inspired by Conditional GAN (CGAN) [13] and Auxiliary Classifiers GAN (AC-GAN) [14]. Firstly, we added a classifier to classify the label of the original sample X and the reconstructed sample X', as well as to distinguish X' (fake) from X (real). Secondly, in order to improve the diversity of data, mending the possible data distribution that can be presented by testing samples but be lack of in training samples, while retaining the main features to a certain extent, we separately fed $\mathcal{E}$ and Z to the *Decoder* and averaged the reconstructed $X'_{\mathcal{E}}$ with X'_Z (see Equation 1).

$$\begin{aligned} Z &= Encoder(X) \\ X'_{\mathcal{E}} &= Decoder(\mathcal{E}) \\ X'_Z &= Decoder(Z) \\ X' &= Average(X'_{\mathcal{E}}, X'_Z) \end{aligned} \tag{1}$$

Equation 2 demonstrates the objective function of our proposed method, where D stands for the discriminator, $Decoder(\mathcal{E}, Z)$ is the notation that combines $Decoder(\mathcal{E})$ and $Decoder(Z)$. Different from the conventional GAN than only contains 2 players G and D, we add a classifier C, which is composed with two parts (C_c, C_d), to not only do the classification (annotated as C_c), but also to distinguish X' from real X (annotated as C_d). Given an input X, D aims to estimate the representation extracted by *Encoder* and $\mathcal{E}$, maximize the possibility of recognizing fake and real; while C is also doing the maximization, it attempts to distinguish X' reconstructed by *Decoder* as fake and assigns a class to both X and X'. We defined that the class of the reconstructed X' is the same as the original X. *Encoder*, *Decoder* is trained to maximally confuse D and C_d into believing that representations and samples it generates come from the real data space.

$$\begin{aligned} \min_{Encoder,Decoder,C_c} \max_{D,C_d} &= \mathbb{E}_{\mathcal{E}}[\log D(\mathcal{E})] + \mathbb{E}_Z[\log(1 - D(Z))] \\ &+ \mathbb{E}_X[\log C_c(X)] + \mathbb{E}_x[\log C_d(X)] \\ &+ \mathbb{E}_{\mathcal{E}Z}[\log C_c(Decoder(\mathcal{E}, Z))] \\ &+ \mathbb{E}_{\mathcal{E}Z}[\log(1 - C_d(Decoder(\mathcal{E}, Z)))] \end{aligned} \tag{2}$$

3.2 Modules Architecture

Figure 3 and 4 show the neural networks we designed for each module. Since the long-short-term memory network has been confirmed to be useful in audio processing [6, 21, 22], we hereby employed one of its variants, gated recurrent units (GRU), to build the encoder and classifier, the number of cells was set to 64. The input of encoder was the power spectrogram extracted from audio, we fed them into a bi-directional GRU for processing the time-series

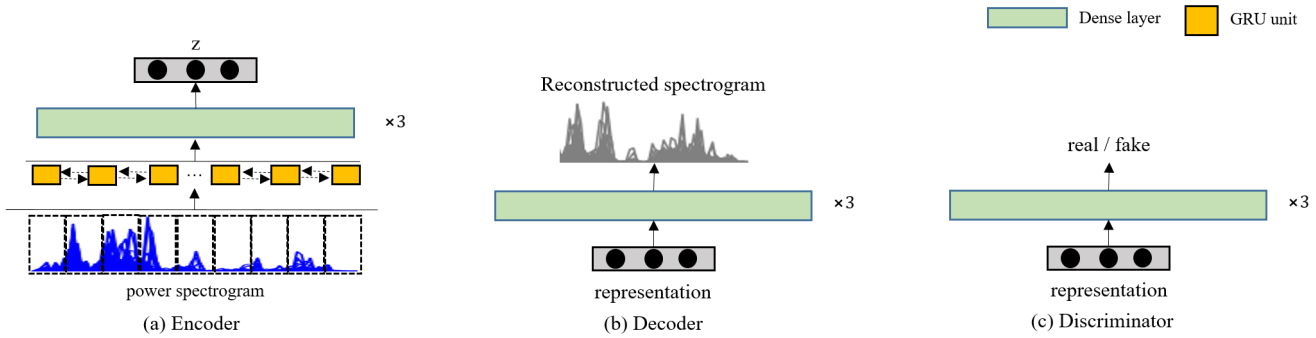

Figure 3: The neural network architectures for the proposed (a) Encoder, (b) Decoder and (c) Discriminator.

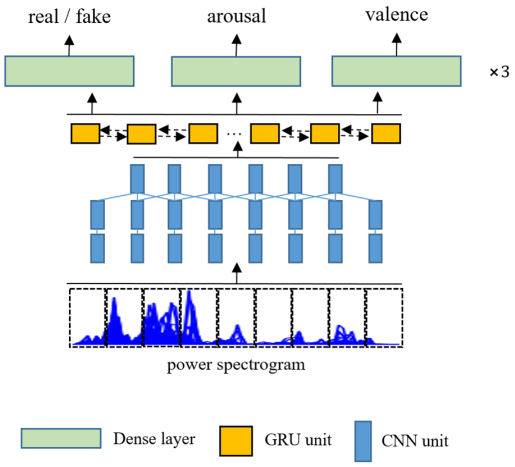

Figure 4: The neural network architecture for the proposed classifier.

information. After that, three dense layers were adopted to generate the hidden representation Z, the numbers of units for each dense layer were 512, 256, 128. The decoder and discriminator were simply constructed with three dense layers. The numbers of units for each dense layer in the decoder were 64, 64, 38400, while the numbers of units for each dense layer in the discriminator were 512, 256, 1.

The classifier is mainly composed of convolutional layer and long-short-term memory network since the combination of them has been widely implemented in speech emotion recognition studies [7, 23]. In our proposed classifier, we first fed the power spectrogram to a three-layer convolutional network, the numbers of filters for each layer were 512, 256, 128 with kernel size as 3, 1, 3 respectively and strides were all equal to 1. Then, the extracted features were passed to a bi-directional GRU with the cells number as 64. Finally, there were three independent networks composed of three dense layers for the sake of distinguishing real or fake, the classification of the arousal dimension, and the classification of the valence dimension. The numbers of units for each dense layer of the distinguishing network were 512, 32, 1, while the numbers of units for each dense layer in the classification networks were 512, 32, 3.

4 EXPERIMENT

4.1 Dataset and Preprocessing

We trained and evaluated on Multimodal Emotion-Target Engagement Sub-challenge (MuSe-Topic) [19]. There are two tasks in this dataset, predicting 10 classes of domain-specific and 3 classes of valence and arousal from a car introduction video. In this study, we only focus on the emotion prediction task based on audio modality. Following the train–test splitting that has done with the original dataset, we processed 4207 training samples and 1335 testing samples. The evaluation scoring standard for the trained model combines (0.34·) unweighted average recall (UAR) and (0.66·) F1 independent for each prediction, which is identical with the baseline.

As for the audio preprocessing, we extracted the power spectrogram from each video segment as features for later training use. We firstly set the maximal length of each utterance to 15s. The longer utterances were cut at 15s and the shorter ones were padded with zeros. The sampling rate was set to 16000Hz. For each frame, a Fast Fourier transform (FFT) of length 1024 with a hop length of 512 was calculated.

4.2 Settings and Baselines

As Algorithm 1 illustrates the training procedures, we iteratively trained each module in AAEC. More specifically, we first trained the *Encoder* and *Discriminator* for mapping Z to $\mathcal{E}$ with K epochs. Then, we trained the *Decoder* individually to generate new samples. After K epochs, the $Classifier$ started to be trained on original and reconstructed samples. The number of total training iterations was set to 500 and K_1, K_2, K_3 were set to 10, 20, 20 respectively in our experiment.

The other configurations for self-comparing of AAEC are shown in Table 1, note that the notation G stands for the proposed encoder-decoder, while G_{ori} is a aberration for the original encoder-decoder presented in Figure 2 (a) . To verify whether C could completely replace D, the discriminator was removed in AAE_C. To confirm whether C could take advantage of adversarial training for improving performance, we designed AAE+C that C was taken out from the iteration and trained afterward. The AAEC′ was designed to

Table 1: Results of F1, Unweighted Average Recall (UAR), and Combined ($0.66 \cdot F1 + 0.34 \cdot UAR$) for the 3-class valence and arousal predictions and the Combined (mean) of valence and arousal.

Model	Player	Modality	c-Arousal			c-Valence			Combined
			F1	UAR	Combined	F1	UAR	Combined	
- ours									
AAEC	D, C, G	audio	**39.80**	40.06	39.89	40.72	**40.89**	40.78	**40.33**
AAE_C	C, G	audio	38.88	38.86	38.87	39.25	39.23	39.24	39.05
AAE+C	D, G	audio	38.15	38.41	38.24	37.69	39.01	38.14	38.19
AAEC′	D, C, G_{ori}	audio	38.48	38.56	38.51	37.77	37.78	38.11	38.31
Classifier	-	audio	34.66	34.81	34.71	35.92	36.16	36.00	35.36
- Stappen et al. [19]									
SVM	-	audio	34.29	34.07	34.21	**42.30**	40.18	**41.83**	38.02
LSTM+Self-ATT	-	audio	34.60	35.00	34.74	37.54	36.78	37.28	36.01
MMT	-	audio, visual, text	39.02	**40.52**	**40.12**	37.30	37.87	37.50	38.81

Algorithm 1 AAEC training procedure

optimizer = $Adam(lr = .2e-4,\ beta\_1 = 0.5)$
for number of total training iterations **do**
  for K_1 epochs **do**
    train *Encoder* and *Discriminator*
  end for
  for K_2 epochs **do**
    train *Decoder*
  end for
  for K_3 epochs **do**
    train $Classifier$
  end for
end for
extract trained $Classifier$ and fine-tune on original dataset with optimizer = $SGD(lr = .5e-5)$

compare the effect of the proposed encoder-decoder to the original one. Moreover, we also trained a pure classifier with the proposed architecture to verify the effectiveness of AAEC. The baseline models presented in Table 1 are chosen from the reported works done by Stappen et al. [19], which contain single modality emotion recognition and multi-modality emotion recognition. *MMT* is the abbreviation for Multi-model Transformer.

4.3 Results

Table 1 shows the comparison results, where the bold font denotes the best performance. From the statistic results yield by our models, we can see that the adversarial training genre models (AAEC, AAE_C, AAEC′) outperform the pure classifier, also the classifier trained on augmented data (AAE+C) has a better performance by increasing about 2.6% than the classifier alone. Moreover, the comparison between AAEC and AAE+C suggests that the model takes the advantages of adversarial training by increasing the performance about 2.2%. Furthermore, from the results of AAEC and AAE_C we can see that the model containing discriminator improves the combined score by around 0.8%, which might imply that even though the work of D and C overlap to a certain extent, D still can not completely be replaced. Additionally, by comparing the performance of AAEC with AAEC′, the proposed modification (separately feeding $\mathcal{E}$ and Z to the *Decoder* and averaging the reconstructed $X'_{\mathcal{E}}$ with X'_Z for obtaining X) achieves better performance (40.33%), which is around 2% higher than the score of AAEC′ (38.31%).

Compared to the performance of baselines, the proposed model achieves the best combined score, which is about 2.2% higher than audio-modality SVM, about 4.3% higher than audio-modality LSTM with self-attention, about 1.5% higher than MMT with three modalities.

5 DISCUSSION

We randomly generated some samples in each iteration using the generator of AAEC and visualized the samples based with principal component analysis (PCA) to compare the distribution of the augmented samples and the original samples, as shown in Figure 5. The shown samples continuously accumulated following the training process.

From Figure 5 we can see that the original distribution of the training set cannot fully cover the distribution of the test set in the MuSe-Topic dataset, which meets the conjecture of Figure 1. Our method makes up for this problem to a certain extent. Through the process of adversarial training, when the generator's ability has not converged during the early period, it would generate some almost-true-like samples that have a certain offset with the real distribution to expand the distribution of training data and supplement the data that was not accessible by the classifier. This could be the reason why AAEC performs better than the pure classifier and the classifier trained on the augmented data (AAE+C). At the later stage of the training iteration, we can see from Figure 5 (the second row) that most of the samples the model generated are almost within the real distribution, which can be regarded as true-like samples used to augment the training data. This stage also has positive effects on improving the performance of the classifier. Referring to the results achieved by AAEC and AAE_C, AAE_C is not as good as

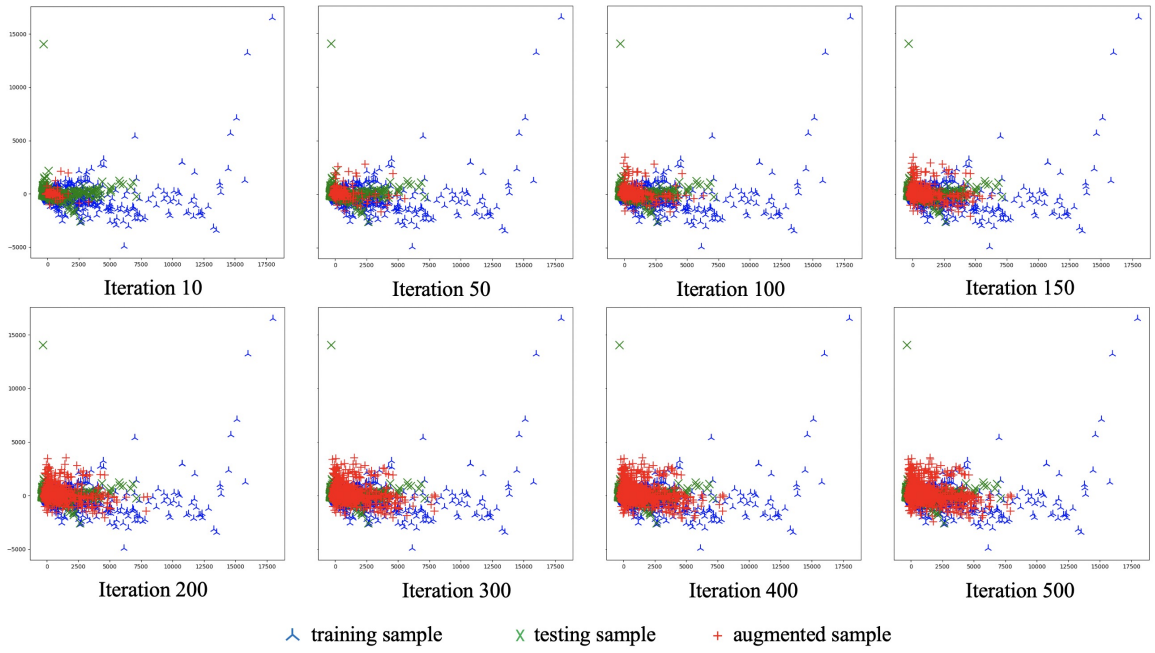

Figure 5: Data visualization. We randomly generated new samples and visualized them in each iteration. The generated samples shown in the figure are continuously accumulated following the training process.

AAEC. We suspect this is because, when C and D are all working for learning the distribution mapping, the true-like sample that the generator in AAEC can generate in the later training stage would be closer to the real distribution than that generated by AAE_C, in which only C was used to learn the distribution mapping. These samples help the classifier to refocus back to the real sample distribution after seeing the possible sample distribution expanded in the early training period. Therefore, it may not be a particularly good choice to blindly expand the possible sample distribution to compensate for the incompleteness of the original training data. The model still needs abilities and methods to refocus back on the most real distribution. In addition, there is an important tip when making the model focus back on the real sample. We may need to let the generator generate some data in the real data distribution but reasonably make them as far as possible from the original real samples, in other words, to fill the area in the real distribution that the original data cannot cover. This guess can be corroborated to a certain extent by the results of AAEC and AAEC′. Without the perturbation of averaging the generated $X'_{\mathcal{E}}$ with X'_{Z}, the new sample that AAEC′ can generate will be very close to the original true sample. It means that when the model refocuses on the true sample distribution, the sample diversity that AAEC′ can provide is lower than that of AAEC.

In the experiment, we also tried to integrate other modalities with audio modality, such as facial expressions, texts. But no matter whether it is a single modality or multi-modality, it could not outperforms the model with only audio modality. We may consider that this may be due to the type of this dataset, which is a video set for introducing cars, the model cannot capture much emotional information from facial expressions and texts. But this problem still worth exploring more deeply.

6 CONCLUSION

In this paper, we propose a model based on adversarial autoencoder and an emotional classifier with an adversarial training strategy. The experiment suggests that our proposed method can not only augment training samples within the existing true distributions but also reasonably extended the data space to a possible distribution that can better fit training and testing sets. The results show that our model outperforms the baseline of single modality and also achieves a better score than the multi-modality baseline model.

7 ACKNOWLEDGMENTS

This work was partly supported by Grant-in-Aid for Scientific Research on Innovative Areas JP20H05576, and JST, ERATO, Grant Number JPMJER1401.

REFERENCES

[1] Antreas Antoniou, Amos Storkey, and Harrison Edwards. 2017. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340* (2017).
[2] Aggelina Chatziagapi, Georgios Paraskevopoulos, Dimitris Sgouropoulos, Georgios Pantazopoulos, Malvina Nikandrou, Theodoros Giannakopoulos, Athanasios Katsamanis, Alexandros Potamianos, and Shrikanth Narayanan. 2019. Data Augmentation Using GANs for Speech Emotion Recognition.. In *INTERSPEECH*. 171–175.
[3] Jaejin Cho, Raghavendra Pappagari, Purva Kulkarni, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Deep neural networks for emotion recognition combining audio and transcripts. *arXiv preprint arXiv:1911.00432* (2019).
[4] Dongyang Dai, Zhiyong Wu, Runnan Li, Xixin Wu, Jia Jia, and Helen Meng. 2019. Learning discriminative features from spectrograms using center loss for

speech emotion recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7405–7409.

[5] Chris Donahue, Julian McAuley, and Miller Puckette. 2018. Synthesizing audio with generative adversarial networks. *arXiv preprint arXiv:1802.04208* 1 (2018).

[6] Changzeng Fu, Thilina Dissanayake, Kazufumi Hosoda, Takuya Maekawa, and Hiroshi Ishiguro. 2020. Similarity of Speech Emotion in Different Languages Revealed by a Neural Network with Attention. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*. IEEE, 381–386.

[7] Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, and Julien Epps. 2019. Direct modelling of speech emotion from raw speech. *arXiv preprint arXiv:1904.03833* (2019).

[8] Runnan Li, Zhiyong Wu, Jia Jia, Sheng Zhao, and Helen Meng. 2019. Dilated residual network with multi-head self-attention for speech emotion recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6675–6679.

[9] Aswathy Madhu and Suresh Kumaraswamy. 2019. Data Augmentation Using Generative Adversarial Network for Environmental Sound Classification. In *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 1–5.

[10] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6818–6825.

[11] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. 2015. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644* (2015).

[12] Giovanni Mariani, Florian Scheidegger, Roxana Istrate, Costas Bekas, and Cristiano Malossi. 2018. Bagan: Data augmentation with balancing gan. *arXiv preprint arXiv:1803.09655* (2018).

[13] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).

[14] Augustus Odena, Christopher Olah, and Jonathon Shlens. 2017. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*. 2642–2651.

[15] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. MELD: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508* (2018).

[16] Yanmin Qian, Hu Hu, and Tian Tan. 2019. Data augmentation using generative adversarial networks for robust speech recognition. *Speech Communication* 114 (2019), 1–9.

[17] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. I know the feeling: Learning to converse with empathy. (2018).

[18] Justin Salamon and Juan Pablo Bello. 2017. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters* 24, 3 (2017), 279–283.

[19] Lukas Stappen, Alice Baird, Georgios Rizos, Panagiotis Tzirakis, Xinchen Du, Felix Hafner, Lea Schumann, Adria Mallol-Ragolta, Björn W Schuller, Iulia Lefter, Erik Cambria, and Ioannis Kompatsiaris. 2020. MuSe 2020 Challenge and Workshop: Multimodal Sentiment Analysis, Emotion-target Engagement and Trustworthiness Detection in Real-life Media. In *1st International Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop, co-located with the 28th ACM International Conference on Multimedia (ACM MM)*. ACM.

[20] Ilaria Torre, Emma Carrigan, Rachel McDonnell, Katarina Domijan, Killian McCabe, and Naomi Harte. 2019. The Effect of Multimodal Emotional Expression and Agent Appearance on Trust in Human-Agent Interaction. In *Motion, Interaction and Games*. 1–6.

[21] Jianyou Wang, Michael Xue, Ryan Culhane, Enmao Diao, Jie Ding, and Vahid Tarokh. 2020. Speech Emotion Recognition with Dual-Sequence LSTM Architecture. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6474–6478.

[22] Yue Xie, Ruiyu Liang, Zhenlin Liang, Chengwei Huang, Cairong Zou, and Björn Schuller. 2019. Speech emotion classification using attention-based LSTM. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27, 11 (2019), 1675–1685.

[23] Jianfeng Zhao, Xia Mao, and Lijiang Chen. 2019. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control* 47 (2019), 312–323.

[24] Ziping Zhao, Zhongtian Bao, Yiqin Zhao, Zixing Zhang, Nicholas Cummins, Zhao Ren, and Björn Schuller. 2019. Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition. *IEEE Access* 7 (2019), 97515–97525.

[25] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Thirty-Second AAAI Conference on Artificial Intelligence*.