IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 8, August 2023

DOI: 10.17148/IARJSET.2023.10836

IMPROVING SPEECH EMOTION RECOGNITION WITH ADVERSARIAL DATA AUGMENTATION NETWORK

Vismaya.S[1], Prof. Sandeep. NK[2]

PG Scholar, Dept. of MCA, Vidya Vikas Institute of Engineering and Technology, Mysuru, Karnataka, India[1]

Assistant professor, Dept. of MCA, Vidya Vikas Institute of Engineering and Technology, Mysuru, Karnataka, India[2]

Abstract: When working with limited training data, training a deep neural network without causing overfitting can be a challenge. To address this issue, a new data augmentation network called the Adversarial Data Augmentation Network (ADAN) has been proposed in this article. The ADAN is based on Generative Adversarial Networks (GANs) and consists of a GAN, an autoencoder, and an auxiliary classifier. These networks are trained adversarially to synthesize class-dependent feature vectors in both the latent space and the original feature space, which can then be used to augment the real training data for training classifiers. Instead of using the conventional cross-entropy loss for adversarial training, the Wasserstein divergence is used to produce high-quality synthetic samples.The proposed networks were applied to speech emotion recognition using EmoDB and IEMOCAP as the evaluation datasets. By making the synthetic latent vectors and the real latent vectors share a common representation, the gradient vanishing problem can be largely alleviated. Results show that the augmented data generated by the proposed networks are rich in emotional information

I. INTRODUCTION

When there is a shortage of training data, training a deep neural network can be challenging due to the likelihood of overfitting. To overcome this challenge, a new data augmentation network called the adversarial data augmentation network (ADAN) has been proposed in this article. The ADAN consists of a generative adversarial network (GAN), an autoencoder, and an auxiliary classifier. These networks are trained adversarially to synthesize class-dependent feature vectors in both the latent space and the original feature space, which can be added to the real training data for training classifiers. Instead of using the conventional cross-entropy loss for adversarial training, the Wasserstein divergence is used to attempt to produce high-quality synthetic samples. The proposed networks were applied to speech emotion recognition using EmoDB and IEMOCAP as evaluation datasets. It was found that by forcing the synthetic latent vectors and the real latent vectors to share a common representation, the gradient vanishing problem can be largely alleviated. Results show that the augmented data generated by the proposed networks are rich in emotion information. Therefore, the resulting emotion classifiers are competitive with state-of-the-art speech emotion recognition systems.

Problem Statement

Speech emotion recognition is a critical component of human-computer interaction and has applications in fields like healthcare, customer service, and entertainment. However, existing models often face challenges in accurately recognizing emotions from speech due to variations in vocal expressions and background noise. Therefore, the scope of this project is to enhance speech emotion recognition by leveraging adversarial data techniques to improve model accuracy and robustness.

II. RELATED WORKS

1. Literature Survey on Recent Advancements in Speech Emotion Recognition

Abstract: This literature survey provides an in-depth examination of the latest developments in the field of Speech Emotion Recognition (SER). It covers a wide range of techniques, from traditional acoustic feature-based models to deep learning approaches. The survey explores the challenges associated with emotion recognition from speech, such as data variability, cultural influences, and cross-lingual issues. By synthesizing research findings, this survey offers insights into the state-of-the-art SER methodologies and their applications across diverse domains.

2. A Comprehensive Review of Datasets for Speech Emotion Recognition

Abstract: This literature survey focuses on the crucial role of datasets in training and evaluating Speech Emotion Recognition (SER) systems. It provides a comprehensive overview of publicly available emotion speech databases,

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 8, August 2023

DOI: 10.17148/IARJSET.2023.10836

analyzing their size, diversity, and relevance to real-world scenarios. The survey assesses the challenges associated with dataset bias, data collection methodologies, and the need for cross-cultural and multilingual datasets. By summarizing dataset-related research, this survey aids in understanding the foundations of SER and its potential for practical applications.

3. Machine Learning and Deep Learning Approaches in Speech Emotion Recognition: A Survey

Abstract: This literature survey explores the application of machine learning and deep learning techniques in Speech Emotion Recognition (SER). It reviews research on feature extraction, classification algorithms, and deep neural network architectures tailored for emotion recognition from speech signals. The survey discusses the advantages and limitations of each approach, emphasizing their impact on SER accuracy and robustness. By analyzing existing studies, this survey provides a comprehensive overview of the role of machine learning and deep learning in advancing SER technology.

4. Cross-Cultural Speech Emotion Recognition: Challenges and Solutions

Abstract: This literature survey addresses the complexities of Cross-Cultural Speech Emotion Recognition (SER) and the unique challenges it poses. It reviews research on cultural influences, linguistic variations, and emotion expression differences in speech across diverse populations. The survey also explores cross-cultural dataset creation, model adaptation techniques, and the development of culturally sensitive SER systems. By summarizing cross-cultural SER studies, this survey offers insights into the need for context-aware emotion recognition in a globalized world.

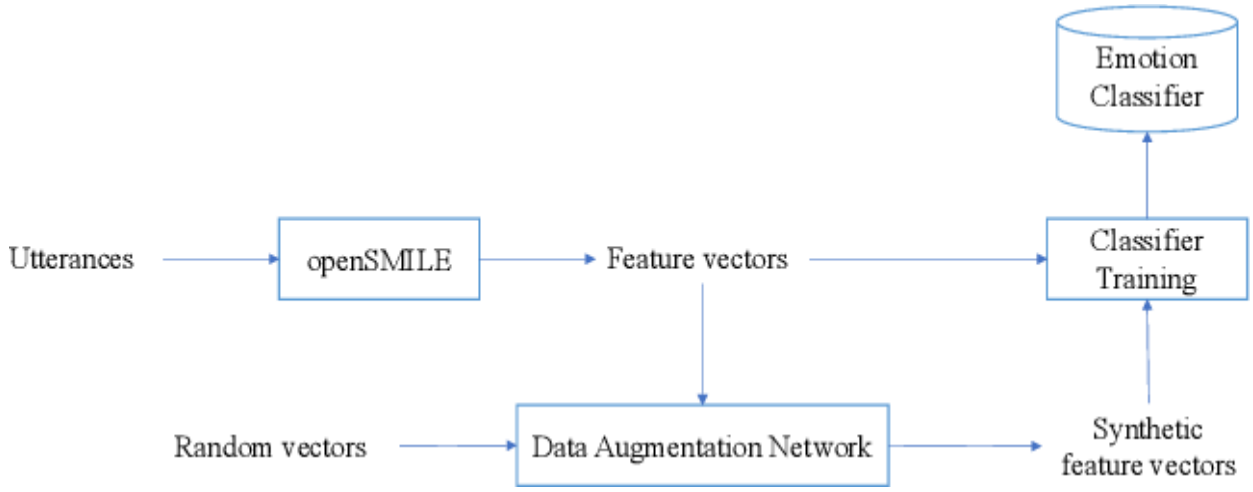
5. Real-Time Speech Emotion Recognition: Techniques and Applications

Abstract: This literature survey focuses on the advancements in real-time Speech Emotion Recognition (SER) systems and their practical applications. It reviews real-time feature extraction methods, efficient algorithms, and hardware acceleration techniques for on-the-fly emotion analysis. The survey discusses SER's role in human-computer interaction, virtual assistants, and healthcare applications. By synthesizing real-time SER research, this survey provides a comprehensive understanding of the potential for immediate emotion recognition in various domains.

III. SYSTEM ANALYSIS

The figure shows that while the model can capture the entire distribution of the dataset, it struggles to match the distribution of the individual classes. The data generated from the AAE follows a predefined mixture of Gaussian distributions and not the actual data distribution.In contrast, the proposed ADAN model not only forms clusters in the latent space but also generates synthetic samples that follow the real data distribution. Due to space limitations, we present only one t-SNE plot for WADAN, which is similar to that of ADAN.It appears that the discriminator has been optimized and converged too early, while the generator is still struggling to optimize. This highlights the challenge of training a standard GAN, where the learning between the generator and discriminator must be carefully balanced.Since the generated samples are significantly different from the real samples, especially when high-dimensional synthetic data points are produced from a low-dimensional random distribution, the discriminator can easily distinguish them.

IV. PROPOSED METHOD



IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 8, August 2023

DOI: 10.17148/IARJSET.2023.10836

The proposed networks have been used for speech emotion recognition with EmoDB and IEMOCAPA datasets as evaluation datasets. It was discovered that by making the synthetic and real latent vectors share a common representation, the problem of gradient vanishing can be greatly reduced. Additionally, the results indicate that the augmented data generated by the proposed networks contain rich emotional information. As a result, the emotion classifiers produced by the proposed networks are competitive with the state-of-the-art speech emotion recognition systems. Our training strategy involves dynamically changing the number of training epochs between the generator and the discriminator instead of fixing it. Rather than optimizing the number of training epochs, our proposed network aims to improve learning stability by helping the generator learn the target distribution. Specifically, the generator in our proposed network learns the distribution of a latent representation produced by a simultaneously trained encoder rather than learning a predefined distribution.

V. CONCLUSION

In conclusion, the research on "Improving Speech Emotion Recognition With Adversarial Data Augmentation Network" emphasizes the potential of using adversarial data augmentation techniques to enhance speech emotion recognition systems. By generating synthetic speech samples that closely resemble real data, the adversarial network addresses the challenges posed by limited labeled datasets, variability in speech, and the difficulty in capturing subtle emotional cues.

FUTURE ENEHANCEMENT

One potential area for future developments is the creation of techniques to facilitate transfer learning across diverse languages and cultures. By utilizing the knowledge and insights gained from one language or culture, it may be feasible to enhance the accuracy of emotion recognition in other languages and cultures, thus expanding the scope and practicality of the system.

REFERENCES

[1]. G. E. Hinton et al., "Deep neural networks for acoustic modeling in speech recognition", IEEE Signal Process. Mag., vol. 29, no. 6, pp. 82-97, Oct. 2012.
[2]. K.He,X.Zhang,S.RenandJ.Sun,"Deepresidual learning forimage recognition", Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 770-778, Jun. 2016.
[3]. B.Alipanahi,A.Delong,M.T.WeirauchandB.J.Frey,"Predictingthesequence specificities of DNA-and RNA-binding proteins by deep learning", Nat. Biotechnol., vol. 33, pp.831-838,Jul. 2015.
[4]. C.Bussoetal.,"IEMOCAP: Interactive emotional dyadic motion capture database",Lang. Resour. Eval., vol. 42, no. 4, pp. 335-359, Dec. 2008.
[5]. E. Cambria, "Affective computing and sentiment analysis", IEEE Intell. Syst., vol. 31, no. 2, pp. 102-107, Mar. 2016.
[6]. I.Chaturvedi, R. Satapathy, S. Cavallari and E. Cambria, "Fuzzy commonsense reasoning for multimodal sentiment analysis", Pattern Recognit. Lett., vol. 125, pp. 264-270, Jul. 2019. Show in Context CrossRef GoogleScholar
[7]. B. Schuller, G. Rigoll and M. Lang, "Hidden Markov model-based speech emotion recognition",Proc.IEEEInt.Conf.Acoust.SpeechSignalProcess.(ICASSP),pp.1,Apr.2003.
[8]. I.Luengo,E.NavasandI.Hernaez,"Feature analysis and evaluation for automatice motion identification in speech",IEEETrans.Multimedia,vol.12,no.6,pp.490-501,Oct.2010.Show in Context View Article Google Scholar
[9]. K.Han,D.YuandI.Tashev,"Speech emotion recognition using deep neural network and extreme learning machine", Proc. Interspeech, pp. 223-227, Sep. 2014. Show in Context Google Scholar
[10]. S.Ghosh,E.Laksana,L.-P.MorencyandS.Scherer,"Representationlearningforspeech emotion recognition", Proc. Interspeech, pp. 3603-3607, Sep. 2016