Deep Representation Learning for Improving Speech Emotion Recognition

Siddique Latif[1,2]

[1]University of Southern Queensland, Australia
[2]Distributed Sensing Systems Group, Data61, CSIRO Australia
siddique.latif@usq.edu.au

1. Motivation

Speech emotion recognition (SER) is an active area of research with potential applications in healthcare [1, 2], customer centres [3], and designing naturalistic spoken dialog systems [4]. Despite significant progress in machine learning (ML), the performance of state-of-the-art SER systems is quite low. Data scarcity is one of the major challenges in this field [5]. Available SER corpora are relatively small in size compared to datasets in computer vision and other speech-related applications such as speaker identification and speech recognition [4]. This also causes poor generalisation in SER systems against cross-corpus and cross-language settings which limits their performance and deployments in real-life [6].

To solve these issues, we focus on the utilisation of deep representation techniques for SER systems. Particularly, we aim to develop deep learning (DL) models that learn better emotional representation from fewer labelled data samples and can be trained in semi-supervised or weakly supervised settings to enables the effective utilisation of abundantly available unlabelled data to improve performance as well as generalisation of the system. Besides, the generative modelling ability of Generative adversarial networks (GANs) [7] is utilised to generate synthetic data to augment SER systems.

2. Contributions

This section briefly discusses the contributions made towards achieving the improved performance of SER systems.

2.1. Utilising Deep Architectures

The performance of SER systems heavily depends on the quality of input features. A good set of features often provides better performance. Therefore, human knowledge-based feature engineering, which focuses on crafting speech-related features has led to lots of research studies in SER. Recently, the trend of using deep representation learning models for automatic feature learning is growing in the speech community. These models can automatically learn better representations for different speech-related tasks and minimise the dependency on hand-engineered features. In particular, a combination of convolution neural networks (CNNs) and long short term memory (LSTM) networks have gained great traction in learning representations from raw speech. However, the performance of raw speech based SER models is always less compared to the systems trained with hand-engineered features. In our paper [8], we showed the opportunities to improve the performance of raw speech based SER model by exploiting the properties of CNN in contextual modelling. We propose the use of parallel convolutional layers to harness multiple temporal resolutions in the feature extraction block that is jointly optimised with the LSTM based classification block for the emotion recognition task. Our results in Table 1 suggest that the proposed construct can reach the performance of CNN trained with hand-crafted features on both IEMOCAP [9] and MSP-IMPROV [10] datasets. However, we found that the raw speech based SER system needs more data compared to the models trained on other speech representations. Therefore, it is important to utilise data augmentation or unlabelled data in order to improve generalisation.

Table 1: *UAR (%) comparison of proposed approach on raw speech with CNN trained on hand-engineered feature, source [8].*

Models	UAR (/%)	
	IEMOCAP	MSP-IMPROV
CNN+MFBs [11]	61.8±3.0	52.6 ± 3.8
Proposed+raw speech	60.23±3.2	52.43 ±4.1

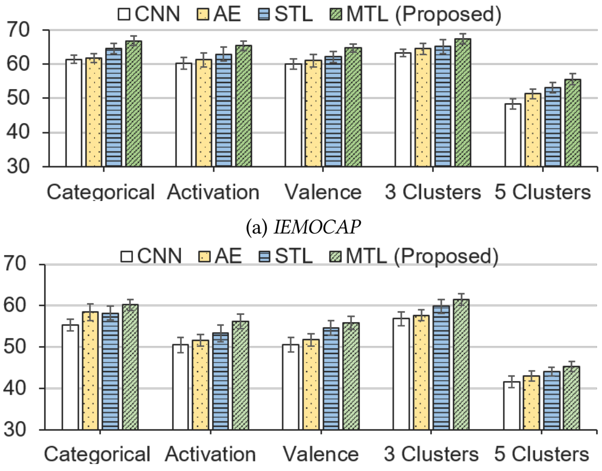
Speech emotion recognition (SER) systems can achieve improved accuracy when the training and test data are identically distributed, but this assumption is frequently violated in practice and the performance of SER systems plummet against unforeseen data shifts [12]. This issue can be solved by learning more complex and generalised representations with very deep architecture. For this, we proposed a deep architecture built on DenseNet [13] and highway networks [14] for robust SER. Our proposed model is a hybrid architecture, where we use a DenseNet for temporal feature extraction, LSTM for context aggregation and fully connected layers in highway configuration for discriminative feature learning. We comprehensively evaluate the architecture in [15] against (1) noise, (2) adversarial attacks and (3) cross-corpus settings. Our evaluations on the widely used IEMOCAP and MSP-IMPROV datasets show promising results when compared with existing studies and state-of-the-art models. Table 2 shows the benchmark results; more detailed analysis can be viewed in our paper [15].

Table 2: *UAR (%) of different models, source [15].*

Model	IEMOCAP	MSP-IMPROV
CNN	61.5 ± 2.3	52.6 ± 2.5
CNN-LSTM	62.1 ±1.8	53.1 ± 2.3
DenseNet	63.2 ± 1.7	54.5 ± 1.9
DenseNet-LSTM	63.5 ± 1.5	55.6 ± 1.6
Proposed	**64.1 ± 1.3**	**56.2 ± 1.5**
CNN-LSTM [16]	62.0	–
CNN [17]	61.4	55.3

2.2. Utilising Additional Data

Semi-supervised training of deep models enables the utilisation of unlabelled data and leads performance improvement of the system. Use of additional unlabelled data also improves the generalisation which helps classifiers to show robustness against unseen data shifts in real-time applications. Here, we proposed a multi-task semi-supervised model [17] that can effectively exploit the abundantly available unlabelled speech data in order to improve the performance of SER. The proposed model was adversarially trained to learn generalised representations for two auxiliary tasks along with emotion classification as the primary task. We propose to use speaker and gender recognition as auxiliary tasks to operate the model on any larger speech corpus which has speaker and gender labels. We demonstrated that the SER performance can be significantly enhanced by simultaneously training emotion classification task with additional auxiliary tasks having an availability of a large amount of data. The proposed model is rigorously evaluated for both categorical and dimensional emotion classification tasks. Experimental results in Figure 1 demonstrate that the proposed model achieves state-of-the-art performance on two publicly available datasets. For more results, our paper [17] can be consulted.



(a) *IEMOCAP*

(b) *MSP-IMPROV*

Figure 1: *Benchmarking results of the proposed multi-task model (MTL) against a single task implementation of the same model (STL), single task implementation by CNN, and a single-task semi-supervised implementation of an autoencoder (AE) using leave-one-speaker-out scheme; source [17].*

Generative adversarial networks (GANs) [7] have gained a lot of attention in the ML community due to their ability to learn and mimic data distributions. They have shown great performance in image generation [18], translation [19], and enhancement [20, 21]; and also in speech generation [22], and conversion [23]. However, the unavailability of larger labelled datasets causes convergence problems in vanilla GANs while generating the synthetic feature vector to augment SER systems [24]. To solve this issue, we propose to use a data augmentation technique combined with a GAN to improve the generation of synthetic samples [25]. Particularly, we propose to use data augmentation technique called “mixup” [26] to train a GAN for synthetic emotional feature generation and also for learning emotional representation in compressed size. To the best of our knowledge, this was the first work to investigate mixup to augment GANs.

The proposed framework can effectively utilise mixup while training a GAN, which augments the representation learning as well as synthetic feature vector generation of GAN. We present a detailed analysis by evaluating the SER performance on (i) a compressed representation, (ii) synthetic samples, and (iii) by using generated samples to augment the training data. Results for within-corpus and cross-corpus setting using two emotional datasets show that the proposed framework performs better compared to recent studies. We trained deep neural network (DNN) classifiers for emotion classification using: (i) only real features, (ii) only synthetic features, and (iii) both real and synthetic features. Here, real features show the representations extracted with openSMILE [27] and augmented by the mixup scheme. Results are reported in Table 3 and 4, which show that synthetic samples by the proposed model can help improve SER performance. More results can be reviewed in [25].

Table 3: *Results for cross-validation evaluation on IEMOCAP, source [25].*

Studies	Real	Syn.	Real+Syn.
Sahu et al. [24]	59.42	34.09	60.29
Bao et al. [28]	59.48 ± 0.71	46.59 ± 0.75	60.37 ±0.70
Ours	**60.51±0.57**	45.75 ± 0.81	**61.05 ±0.68**

Table 4: *Results for cross-corpus evaluation, source [25].*

Studies	Real	Syn.	Real+Syn.
Sahu et al. [24]	45.14	33.96	45.40
Bao et al. [28]	45.58 ± 0.40	41.58 ± 1.29	46.52±0.43
Ours	**46.0±0.57**	**42.15 ± 1.12**	**46.60 ±0.45**

3. Discussion and Future Works

The mentioned results of our previous studies on speech emotion recognition show the potentials of using deep representation learning techniques. We found in [8] that deep models can capture emotional attributes in an end to end fashion trained directly on raw speech. To achieve robustness in SER, very deep architectures can be utilised, as validated in [15], which can learn more complex and robust representation. To increase generalisation abilities in SER, additional data can be utilised. Multi-task learning provides a good way to utilise additional unlabelled data for auxiliary tasks, which lead to performance improvement (see our paper [17]). Synthetic features/data from generative models (e.g., GANs) can be utilised to augment SER systems for performance improvements, as shown in [25].

3.1. Future Works

Future works include exploring self-supervised training of deep representation learning models. Most importantly, we aim to use other modalities such as text and visual data in a self-supervised way. To improve SER performance against cross-corpus and cross-language emotion recognition further efforts are needed. Therefore, in our ongoing work, we are developing a DL model that can learn corpus and language invariant representation for effective SER.

4. Acknowledgements

I would like to thank and acknowledge my PhD mentors Dr Rajib Rana, Dr Sara Khalifa, Prof. Raja Jurdak, Prof. Julien Epps, and Prof. Björn Schuller for their valuable discussions and motivations. Also, I would like to acknowledge the PhD funding provided by the University of Southern Queensland (USQ), Australia and Data61, CSIRO Australia.

5. References

[1] R. Rana, S. Latif, R. Gururajan, A. Gray, G. Mackenzie, G. Humphris, and J. Dunn, "Automated screening for distress: A perspective for the future," *European Journal of Cancer Care*, p. e13033, 2019.

[2] S. Latif, J. Qadir, A. Qayyum, M. Usama, and S. Younis, "Speech technology for healthcare: Opportunities, challenges, and state of the art," *IEEE Reviews in Biomedical Engineering*, pp. 1–1, 2020.

[3] F. Burkhardt, J. Ajmera, R. Englert, J. Stegmann, and W. Burleson, "Detecting anger in automated voice portal dialogs," in *Ninth International Conference on Spoken Language Processing*, 2006.

[4] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. W. Schuller, "Deep representation learning in speech processing: Challenges, recent advances, and future trends," *arXiv preprint arXiv:2001.00378*, 2020.

[5] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. MüLler, and S. Narayanan, "Paralinguistics in speech and language—state-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, pp. 4–39, 2013.

[6] S. Latif, J. Qadir, and M. Bilal, "Unsupervised adversarial domain adaptation for cross-lingual speech emotion recognition," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 732–737.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[8] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct modelling of speech emotion from raw speech," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH 2019)*. International Speech Communication Association (ISCA), 2019, pp. 3920–3924.

[9] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[10] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2017.

[11] Z. Aldeneh and E. M. Provost, "Using regional saliency for speech emotion recognition," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 2741–2745.

[12] S. Latif, A. Qayyum, M. Usman, and J. Qadir, "Cross lingual speech emotion recognition: Urdu vs. western languages," in *2018 International Conference on Frontiers of Information Technology (FIT)*. IEEE, 2018, pp. 88–93.

[13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[14] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *Deep Learning workshop, International Conference on Machine Learning (ICML)*, 2015.

[15] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Deep architecture enhancing robustness to noise, adversarial attacks, and cross-corpus setting for speech emotion recognition," *Interspeech*, 2020.

[16] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms." in *INTERSPEECH*, 2017, pp. 1089–1093.

[17] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps, and B. W. Schuller, "Multi-task semi-supervised adversarial autoencoding for speech emotion recognition," *IEEE Transactions on Affective Computing*, 2020.

[18] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807.

[19] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2849–2857.

[20] M. Usman, S. Latif, M. Asim, B.-D. Lee, and J. Qadir, "Retrospective motion correction in multishot mri using generative adversarial network," *Scientific Reports*, vol. 10, no. 1, pp. 1–11, 2020.

[21] S. Latif, M. Asim, M. Usman, J. Qadir, and R. Rana, "Automating motion correction in multishot mri using generative adversarial networks," *Medical Imaging meets NIPS (MED-NIPS)*, 2018.

[22] P. Chandna, M. Blaauw, J. Bonada, and E. Gómez, "Wgansing: A multi-voice singing voice synthesizer based on the wasserstein-gan," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.

[23] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," in *Proc. Interspeech 2017*, 2017, pp. 3364–3368.

[24] S. Sahu, R. Gupta, and C. Espy-Wilson, "On enhancing speech emotion recognition using generative adversarial networks," *Proc. Interspeech 2018*, pp. 3693–3697, 2018.

[25] S. Latif, M. Asim, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Augmenting generative adversarial networks for speech emotion recognition," *Interspeech*, 2020.

[26] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[27] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[28] F. Bao, M. Neumann, and N. T. Vu, "Cyclegan-based emotion style transfer as data augmentation for speech emotion recognition," *Manuscript submitted for publication*, pp. 35–37, 2019.