Annals of R.S.C.B., ISSN:1583-6258, Vol.24, Issue 1, 2020, Pages. 1400-1414
Received 15 April 2020; Accepted 23 June 2020.

Anger and Stress Recognition from Spectrogram Images Using Pre-TrainedConvolutional Neural Network

[1]Shalini Kapoor, [2]Tarun Kumar

Research Scholar, AKTU, Lucknow, shalini_kapoor311@rediffmail.com, ORCID-ID: 0000-0001-8719-997X

Professor(CSE) Radha Govind Group of Institutions, Meerut, taruncdac@gmail.com ORCID ID: 0000-0001-5033-5544

Abstract

The convolutional neural Network (CNN) has recently shown exceptionally good performance in image classification with a little amount of pre-processing compared to other classification algorithms. Building CNN from a scratch is not only an expensive and complex task but also requires expertise. Pre-trained CNN reduces efforts of creating models from scratch as they are earlier trained on the ImageNet dataset and could work well with the little training set. In the proposed work CNN is used for anger and stress classification through analyzing discriminatory patterns in spectrogram images. In the proposed work performance of pre-trained CNN ResNet50, SqueezeNet, WideResNet, and InceptionV3 are compared in identifying emotions (anger, stress, neutral). Pre- trained Network ResNet50 has shown the highest performance with an accuracy of 99.7%, followed by WideResNet, SqueezeNet, and InceptionV3 with accuracy 85.7%, 57.8%, and 38.5% respectively.

K**eywords**: Deep Learning; Convolution Neural Network; Transfer Learning; Pre-trained Neutral Networks;Spectrogram

1. INTRODUCTION

Speech, as we know is an effortless physiological process that we humans, use to convey the desired message to other creatures. Over the period speech process has not only enabled us to communicate with humans alone but also with pets and other creatures. [1]The speech production process, although seems to be effortless to use, involves five stages i) Conceptual, ii) Syntactic stage iii) Lexical stage iv) Phonological Stage and v) Phonetic stage, which is controlled by the left cerebral hemisphere of the brain and which further controls the muscle articulation and finally helps an individual to deliver the message incorrect levels tone, pitch, volume, timings, and facial gestures. [2]These ingredients finally transmit the desired message to the receiver. The term articulation refers to how you pronounce individual words and how you create speech sounds, varying your pitch, volume, and timing. All the articulations are commanded by the brain are responding to his or her emotional makeup and various patterns such as anger, happiness, stress, shock, and surprise, etc. becomes observable in human speech and conveys our psychological and emotional state of mind.[5]Past studies related to speech emotion recognition have analyzed a variety of features for emotion classification but still, there is no consensus about the features that could best classify the emotional state. In most studies, researchers have proposed their features set for emotion classification. Features engineering is a complex task that requires lots of time and effort. The ability of traditional machine-learning approaches to analyze natural data in its raw form was limited. For decades, designing a feature extractor that turned raw data (such as the pixel values of an image) into an appropriate internal representation or feature required careful engineering and extensive subject

experience.[4] Recently deep learning approaches are replacing conventional methods in several tasks. Deep-learning methods are representation-learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level. Very complex functions can be learned by combining enough of these modifications. Higher layers of representation accentuate characteristics of the input that are significant for discriminating while suppressing irrelevant variations in classification tasks. When an image is represented as an array of pixel values, the learned features in the first layer of representation often represent the presence or absence of edges at specific orientations and places in the image. The second layer recognizes motifs by spotting specific edge arrangements, even if the edge placements are somewhat different. The third layer may group motifs into bigger groups that correspond to components of recognizable things, with following

levels detecting items as a combination of these elements.[6-9] Deep learning is distinguished by the fact that these layers of features are learned from data using a general-purpose learning technique, rather than being built by human engineers. Deep learning is making significant progress in solving issues that have long eluded artificial intelligence's best efforts. The fact that these layers of features are acquired from data using a general-purpose learning technique rather than being designed by human engineers distinguishes deep learning. Deep learning is making tremendous progress in addressing problems that have escaped artificial intelligence's best efforts for a long time. [10] Deep learning has beaten existing machine-learning algorithms in several tasks, as well as breaking records in image recognition and speech recognition. Recently several studies have shown the possibility of using speech spectrogram for emotion classification. The dynamic pattern associated with the different emotional states could be observed well in spectrogram images. In the proposed work deep learning architecture CNN is used for anger and stress recognition. Lots of experimentation using a different number of layers with different hyperparameter settings is required to design optimized CNN. Domain-specific parameter settings are required for enhancing the accuracy and performance of CNN. There is an infinite number of ways to create a deep neural Network. As a result, it is preferable to select something that has previously performed well on a similar challenge, a process known as transfer learning. Pre-trained CNN architecture such as Networks ResNet50, SqueezeNet, WideResNet, and InceptionV3 are earlier trained on the ImageNet dataset which is composed of millions of images and has shown high accuracy in image classification. In the proposed work performance of pre -trained CNN Networks ResNet50, SqueezeNet, WideResNet, and InceptionV3 is evaluated for anger and stress recognition using spectrogram images as input. Pre-trained Networks ResNet50 has shown higher accuracy in anger and stress recognition in comparison to other state-of-the-art techniques used for emotion classification through speech.

The rest of this paper is structured as follows in Section2 contains a related work and preliminaries, in Section3 proposed methodology, Section4 contains experimental setup, in Section5 contains performance evaluation and results, and Section 6 contains conclusion and future work.

Annals of R.S.C.B., ISSN:1583-6258, Vol.24, Issue 1, 2020, Pages. 1400-1414
Received 15 April 2020; Accepted 23 June 2020.

2. Related Work

Recently, there is increasing demand for emotionally aware systems to better understand human behavior. [3] Building a successful, robust and accurate speech emotion recognition using the traditional approach needs to address three important issues a) choice of features b) choice of classifier, and c) choice of emotional speech dataset. Each aspect has a huge impact on the performance and accuracy of the speech emotion recognition system. The steps to build traditional emotion recognition system are shown in Fig.1.

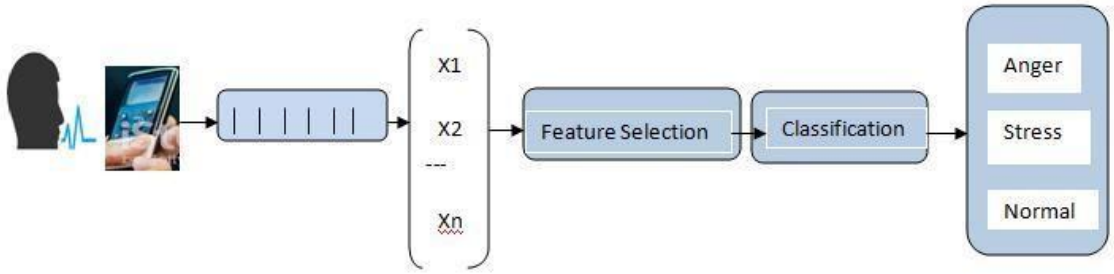


Fig.1. -Steps involved in traditional machine learning

Researchers have used both local features and global features for SER. A large set of acoustic features correlates with emotions. The acoustic features used in speech emotion recognition can be categorized as

prosodic, spectral, and voice quality. One of the major challenges in speech emotion recognition is finding the "best" set of features representative of a particular emotion. Different researchers have used different subsets and types of features for speech emotion classification. Finding a suitable set of features requires time and effort. Previous studies based on speech emotion have directly analyzed raw speech to extract

discriminatory features for speech emotion classification. Image classification approaches, on the other hand, maybe utilized to extract features for sound classification by replacing audio traces with visual representation. Spectrograms and Mel-frequency plots are two of the most prominent visual representations of audio traces. A spectrogram is a two-dimensional representation of sound, the first two dimensions are time and frequency, and a third dimension (pixel intensity) indicates the signal amplitude in a particular frequency. Deep learning approaches are gradually replacing the traditional method of sound classification. Attention is currently focused on the visual representation of sound for pattern recognition. The traditional method uses handcrafted features for pattern recognition with the aim two minimize intra-class variability and maximize interclass variability. Identifying optimized set features using the traditional method is a tough task. The deep learning algorithm has the advantage of automatic learning of the best features from data during training. Deep learning algorithms are more compatible with visual representations and more informative in comparison to hand-engineered features. The advent of CNN with their specialty in image understanding the visual representation of sound is preferable in comparison to traditional features for pattern recognition. Developing efficient CNN architecture from scratch is a tedious task requiring lots of experimentation by using the different configurations in terms of number and types of layer. CNN

Annals of R.S.C.B., ISSN:1583-6258, Vol.24, Issue 1, 2020, Pages. 1400-1414
Received 15 April 2020; Accepted 23 June 2020.

can be fine-tuned by applying different hyperparameter settings which is a tedious and time-consuming task. The solution to the discussed

Issue is transfer learning and a pre-trained Network. Earlier proposed approaches, where transfer learning is used for speech emotion recognition, are [11] proposed a sparse autoencoder for feature transfer learning in speech emotion recognition. In the proposed small dataset in the target domain learns a common emotion-specific mapping rule. This criterion is then applied to emotion-specific data in a different domain to produce newly reconstructed data. Six standard databanks were used to analyse the experimental outcomes. [12] proposed transfer learning for speech emotion identification. The findings of the experiments show that transfer learning outperforms traditional speech emotion classification algorithms. [13] investigated the performance of AlexNet-SVM and FTAlexNet for multiclass speech emotion recognition systems. When evaluated on a popular Berlin Emotional Speech (EMO-DB) database, both algorithms produced state-of-the-art results.[14]This study proposes a strategy for adapting ASR
models in the emotion recognition domain using transfer learning. On the Tedlium2 dataset. [15]This research provides a Time-Delay Neural Network (TDNN) architecture-based transfer learning approach for speech emotion identification. We show that without pretraining on ASR, transfer learning models beat the other approaches significantly. The experiment was successful.[16.] proposed deep learning and transfer learning for speech and emotion recognition. [17] proposed an audiovisual approach using transfer learning for emotion recognition in wild and attained a final accuracy of 57.2% on the test set, which is higher than the baseline of 40.47%. [18]proposed DBNs for transfer learning and demonstrated how a considerable gain in accuracy compared to baseline can be accomplished utilizing the transfer learning technique for cross-corpus emotion recognition by using the experimental data from diverse scenarios.[19] proposed shared-hidden-layer autoencoders for transfer learning for speech emotion recognition.[20] The current research focuses on common yet effective feature transfer learning approaches based on autoencoders, including denoising autoencoders (DAE).[21] Investigated transfer learning for SER and concluded that transfer learning is affected by choice of pretraining task and acoustic settings.

Preliminaries

Transfer Learning

Building a deep learning model from scratch is a complex task as it is difficult. There lots of things to keep in mind to create an effective deep learning model. Decisions while designing a deep learning model ranges from the number of layers, types of layer, their configuration, tuning hyperparameters. Training deep neural Networks requires a huge amount of training data and processing power. Because of the absence of computational infrastructure, it was impossible to create a model that performed exceptionally well on a certain job. Furthermore, even though many people tried, it was hard to gather enough useful data to train a model. Furthermore, despite the efforts of many, gathering enough meaningful data to train a model proved difficult. A short-cut way to achieve this is to re-use the model weights from pre-trained models to train the proposed model that was earlier trained on benchmark datasets, such as the ImageNet for image recognition tasks. Transfer learning is the process of applying models that have been trained on one problem as a

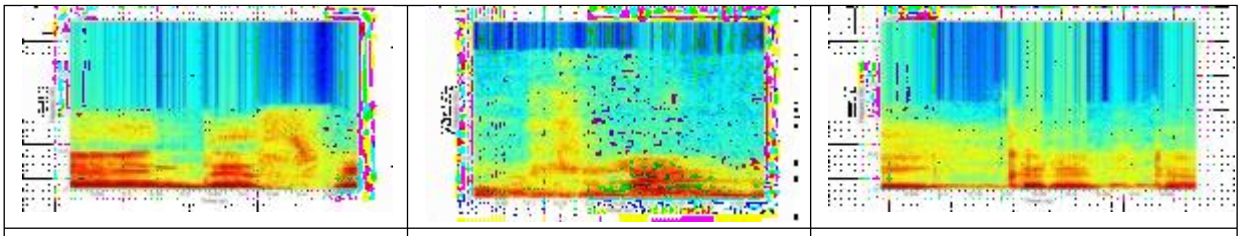
Annals of R.S.C.B., ISSN:1583-6258, Vol.24, Issue 1, 2020, Pages. 1400-1414
Received 15 April 2020; Accepted 23 June 2020.

starting point for a different problem. Transfer learning is adaptable, allowing pre-trained models to be used directly as feature extraction pre-processing or incorporated into completely new models. Transfer learning provides an edge in compensating for the lack of data, reduces model development time, and computing resources. Transfer

earlier used for several tasks such as text clustering, text classification, sentiment classification, and learning collaborative filtering, etc.

Spectrogram

The spectrogram is the image of sound representing changing frequency content of the signal concerning time. The frequency of the signal is represented on the y-axis, and time on the x-axis in the 2D spectrogram. The lowest frequencies are represented at the bottom of the spectrogram while the higher frequencies are represented at the top. The colors are used to represent the amount of energy in the signal. Higher energy regions in spectrogram caused by events such as vocal fold closures, formants, and harmonics are shown using darker color; lighter color such as white is used to represent the region of little energy such as silence. Spectrograms are of two types' narrowband and wideband spectrogram. Narrowband spectrogram is used to show the characteristics of source like the vibration of vocal folds while wideband spectrogram is used to investigate the characteristics of vocal tract such as vocal tract resonance (formants).To generate a spectrogram of sound signal; the signal is first divided into smaller overlapping segments called frames followed by short-Fourier transform.Fig.1 show spectrogram for emotion anger stress and neutral.



Anger	Stress	Normal

Fig.2 Spectrogram of corpus carrying emotion anger, stress, and neutral speech

Hyperparameter

Hyperparameters express "higher-level" structural settings for algorithms. They are decided before fitting the model because they can't be learned from the data. Tuning models, we specifically mean tuning hyperparameters. There is an infinite number of the hyper parameter that governs the learning process few are discussed below.

- The number of the epoch: specify the number of iterations on the dataset its value can be fractional.

- Batch size: is several training examples processed before the model updating. The higher the batch size requires more memory space.

Annals of R.S.C.B., ISSN:1583-6258, Vol.24, Issue 1, 2020, Pages. 1400-1414
Received 15 April 2020; Accepted 23 June 2020.

- Rho: (Applicable only if adaptive_rate is enabled) Specify the adaptive learning rate time decay factor.

- Loss: Evaluate the suitability of a specific algorithm to a given dataset. It helps to reduce the error in the prediction.

- Optimizer: the aim of using an optimizer is to minimize the loss function. This is the most common setting for classification problems.

- Learning Rate: it is an important hyperparameter setting that governs the learning process. A too-small learning rate takes time to converge, setting it too high will lead to an increase in loss. Currently, adaptive learning rates are used at different rates for different model parameters and help the algorithm to converge smoothly.

3. Proposed Methodology

The limitations of building a robust SER system are the lack of huge data set for training a model and limited processing power. An efficient way to deal with such situations is the application of pre-trained Networks and transfer learning. Several pre-trained models exist that were designed for the task of image classification. To handle the SER problem needs to be re-defined as an image classification problem as the majority of current pre-trained Networks were built for image classification. In the proposed performance of four pre-trained Networks (ResNet50, WideResNet, SqueezeNet, and InceptionV3) is investigated for recognizing three emotions anger, stress, and neutral. Steps involved in speech emotion recognition using pre-trained are discussed below

Speech segmentation, transformation into spectrogram images, and dataset uploading

One of the important steps in recognize speech is extracting part of speech that needs to be processed in the proposed work voiced the part of speech is processed for emotion identification. The voiced part of speech is extracted from a digitized speech signal using the voiced activity detection module. The voiced data was further bifurcated into small chunks of size 30ms each. The segmentation speech signal was done in MATLAB. Each chunk belonging to emotional class anger, stress, and neutral was further transformed into color spectrogram images of size 900*1200*3 using Fast Fourier transform. Spectrograms belonging to three emotional classes were stored separately into three folders named anger, stress, and neutral. All three folders were transferred into a folder named emotion. CSV file was created with the path of spectrogram images and emotional labels. CSV file is stored in an emotion folder named emotion. The emotion folder is converted into a zip file and uploaded into deep cognition AI platform. The diagrammatic representation of proposed methodology is shown in Fig.3.

Annals of R.S.C.B., ISSN:1583-6258, Vol.24, Issue 1, 2020, Pages. 1400-1414
Received 15 April 2020; Accepted 23 June 2020.

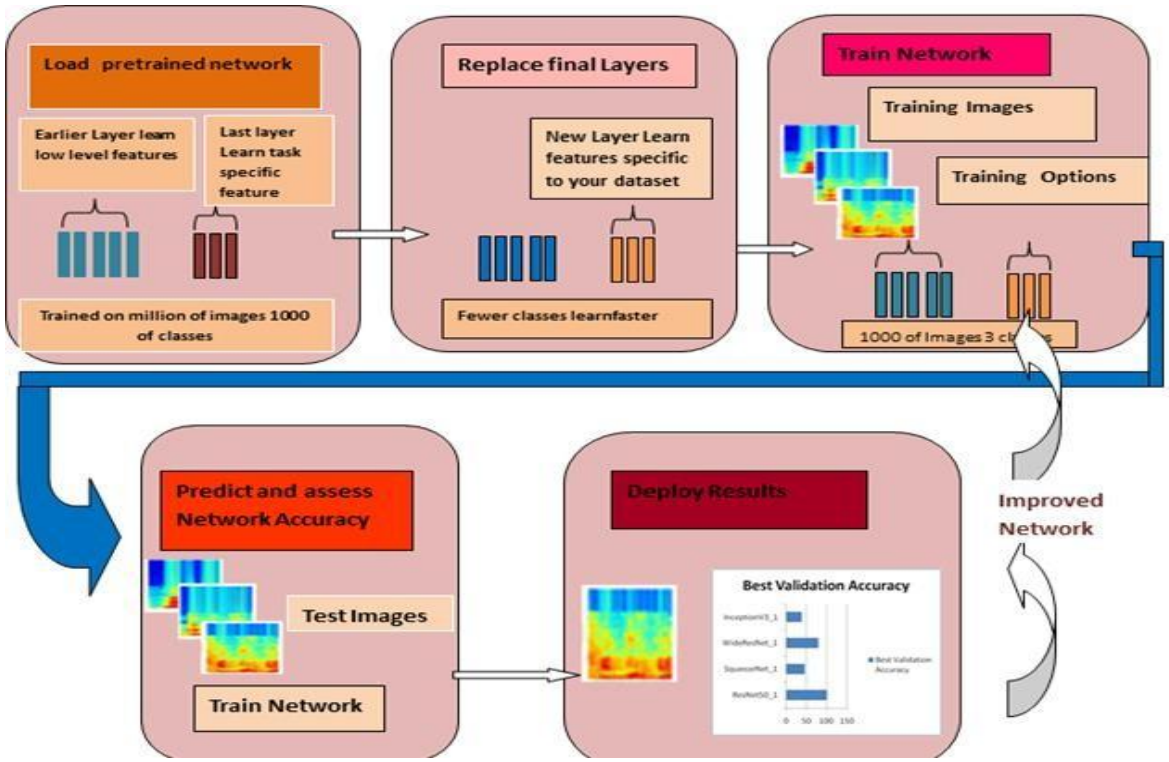


Fig.3: Diagrammatic representation of proposed methodology used for anger and stress recognition using transfer learning

Deep Learning Studio is a deep learning platform for developing and implementing artificial intelligence (AI). You can easily create deep learning models with the simple drag-and-drop interface. Model development is simplified and accelerated by using pre-trained models and built- in helpful features. Using the visual interface, you may import model code and edit the model. As you iterate and tweak hyper-parameters to increase performance, the platform automatically stores each model version. Deep Learning Studio is available in both cloud and desktop versions. With a cloud account, you have an option to rent on-demand high-powered GPUs to train the model. Deep cognition ai provides advanced pre-trained models like ResNet, MobileNet, or build your custom models.

Pre-trained Network topology

The topology of pre-trained models InceptionV3/ ResNet50 /SqueezeNet /WiderResNet is shown in Fig.4, Fig.5, Fig.6, and Fig.7. The Network is composed of the Input layer, pre-trained network, global max-pooling layer and, two dense layers. The first dense layer has an output dimension of 512 and the second dense layer has output dimension three equal to the number of emotional categories. The output of the second dense layer is fed to the softmax layer for classification. Topology of pre-trained models ResNet50, SqueezeNet, WideResNet, and InceptionV3 is shown in Fig.4, Fig.5, Fig.6 and Fig.7 respectively.

Annals of R.S.C.B., ISSN:1583-6258, Vol.24, Issue 1, 2020, Pages. 1400-1414
Received 15 April 2020; Accepted 23 June 2020.

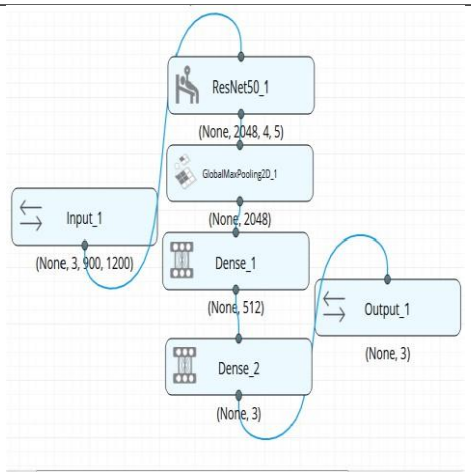


Fig.4:ResNet50

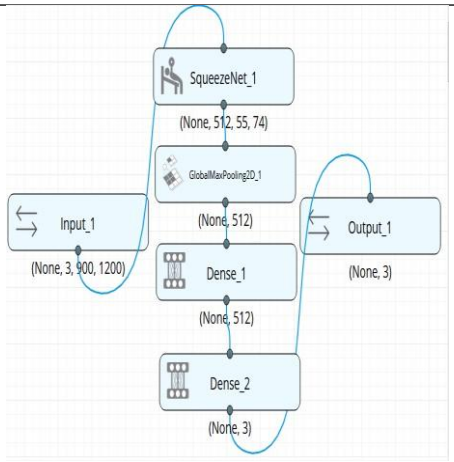


Fig.: SqueezeNet

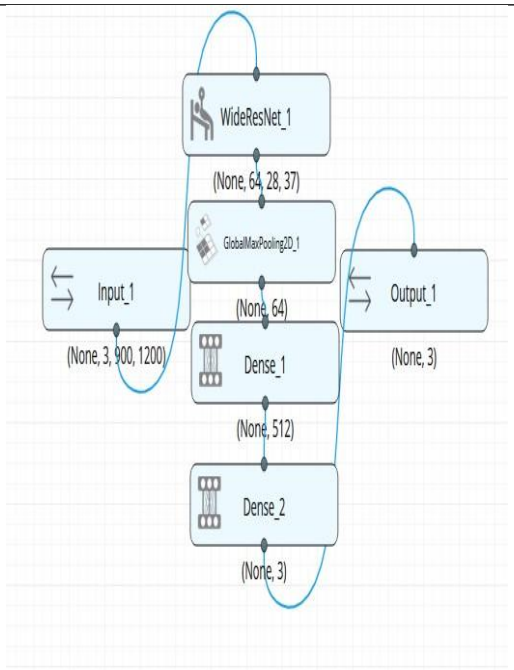


Fig. 6: WideResNet

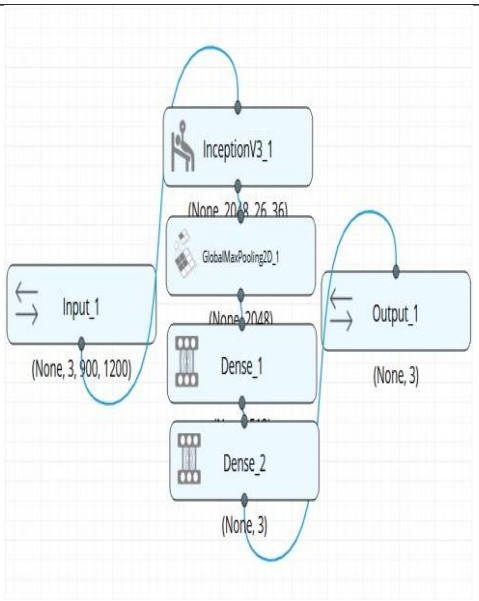


Fig.7:InceptionV3

Model training and testing

Dataset is first split into training, validation, and test set. Model is trained, validated, and tested on the dataset proposed in Section 4.1.

4. Experimental Setup

Dataset

The dataset used to experiment is Interactive Emotional Dyadic Motion Capture (IEMOCAP). The database is acted is a recently collected acted, multimodal, and multispeaker database from USC's SAIL lab. It has about 12 hours of audiovisual content, including video, speech, facial motion

Annals of R.S.C.B., ISSN:1583-6258, Vol.24, Issue 1, 2020, Pages. 1400-1414
Received 15 April 2020; Accepted 23 June 2020.

capture, and text transcriptions. It consists of dyadic sessions in which actors execute improvisations or scripted scenarios that have been carefully chosen to elicit a response from the audience. Multiple annotators have annotated the IEMOCAP database with categorical categories like anger, happiness, sadness, neutrality, and dimensional labels like valence, activation, and dominance. The dataset is available at https://sail.usc.edu/iemocap/

Experiments

Four experiments were performed to evaluate the performance of ResNet50, WideResNet, InceptionV3**,** and SqueezeNet in recognizing emotions anger, stress, and neutral. Each pre-trained model is fed with spectrogram images belonging to emotions anger, stress, and neutral. Each pre-trained model is trained on 1800 spectrograms, tested, and validated on 600 spectrograms each. The platform used to experiment is deep cognition ai.

Hyperparameters setting before implementing pre-trained networks are shown in Table.1.The number of epochs used to train each pre-trained model is set to 10, batch size taken is 32, loss function used is categorical cross-entropy, optimizer chosen is adelta due to its adaptive learning rate, the learning rate is set to .0001, the value of epsilon is set to 1e-08, decay is set to 0 and rho is 0.95. GPU used to implement pre-trained Networks is GPU-K80-12GB.

Table.1. Hyper parameter settings

Hyperparameters	**Values**
Number of Epochs	10
Batch Size	32
Loss Function	Categorical_crossentropy
Optimizer	Adadelta
Learning Rate	.0001
Epsilon	1e-08
Decay	0
Rho	0.95

5. Performance Evaluation and Results

Annals of R.S.C.B., ISSN:1583-6258, Vol.24, Issue 1, 2020, Pages. 1400-1414
Received 15 April 2020; Accepted 23 June 2020.

Performance evaluation measure

Classification accuracy and loss are the two important measures to evaluate any model architecture. Expectation from any good model architecture is higher classification accuracy and low training and validation loss. During the training process, the validation set is used after each epoch to calculate the validation loss and accuracy Performance of each pre-trained model is evaluated based on accuracy and loss.

Accuracy of model is the ratio of correct prediction upon the total number of predictions. A faulty guess results in a loss. To put it another way, the loss is a statistic that indicates how inaccurate the model's forecast was for a single case. The loss is 0 if the model's forecast is perfect; otherwise, the loss is bigger. The purpose of training a model is to discover a set of weights and biases that have a low loss across all cases on average.

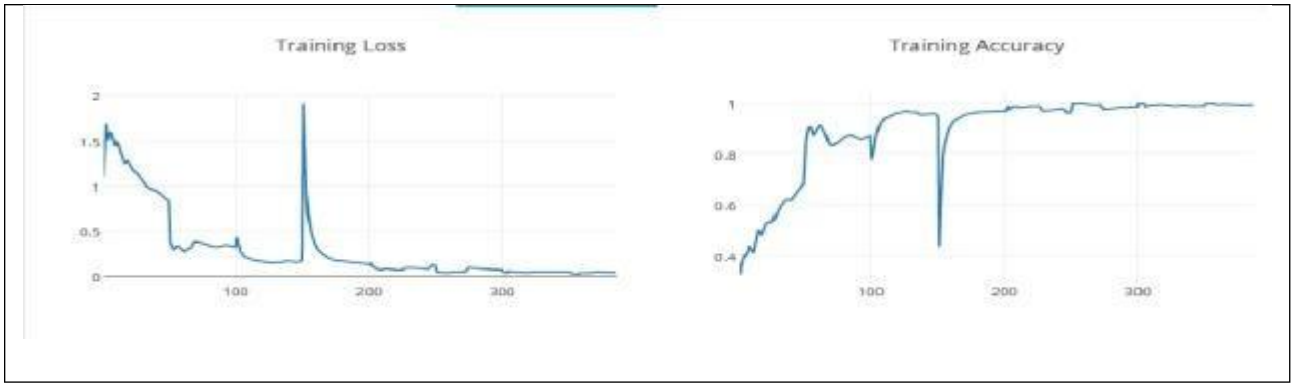
Results

In the proposed work performance of all four pre-trained models is evaluated based on training & validation accuracy and training and validation loss.

ResNet50 showed the highest training accuracy and validation accuracy of 99.7% and 99.8% and minimum training and validation loss 0.0236 and 0.23. The second performance is shown by WideResNet with the training & validation accuracy of 85.7% and 78.9% and training and validation loss of 6.12 and 6.23. The third position is achieved by SqueezeNet with training and validation accuracy of 57.8% & 46.7% with training and validation loss of 7.2 and 9.6 and at the fourth position is InceptionV3 with training & validation accuracy of 38.5% and 38.5% and training and validation loss of 9.8 and 9.8.

Plots showing training & validation accuracy, and training & validation loss obtained using ResNet50, SqueezeNet, WideResNet, and InceptionV3 are shown in Fig.8, Fig.9, Fig.10, and Fig.11.

The x-axis of each plot shows the number of epochs and the y-axis is used for either training accuracy or validation accuracy or training loss, or validation loss.



Annals of R.S.C.B., ISSN:1583-6258, Vol.24, Issue 1, 2020, Pages. 1400-1414
Received 15 April 2020; Accepted 23 June 2020.

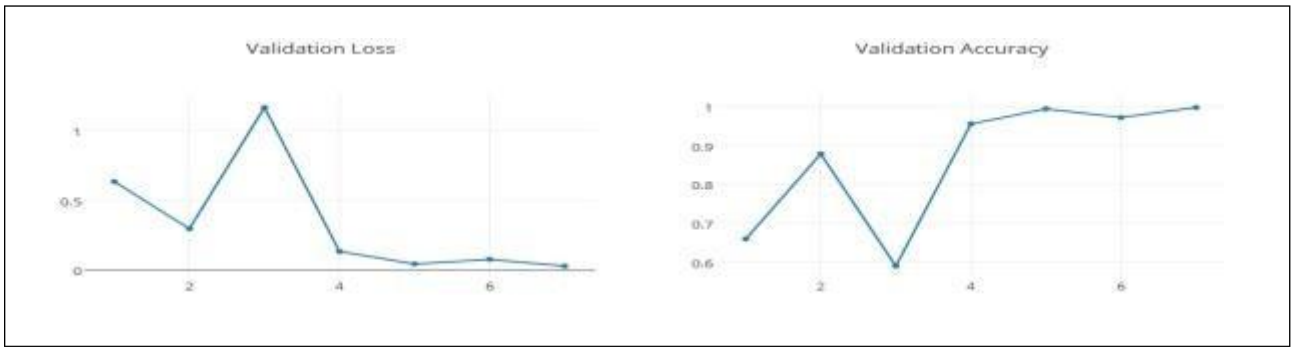
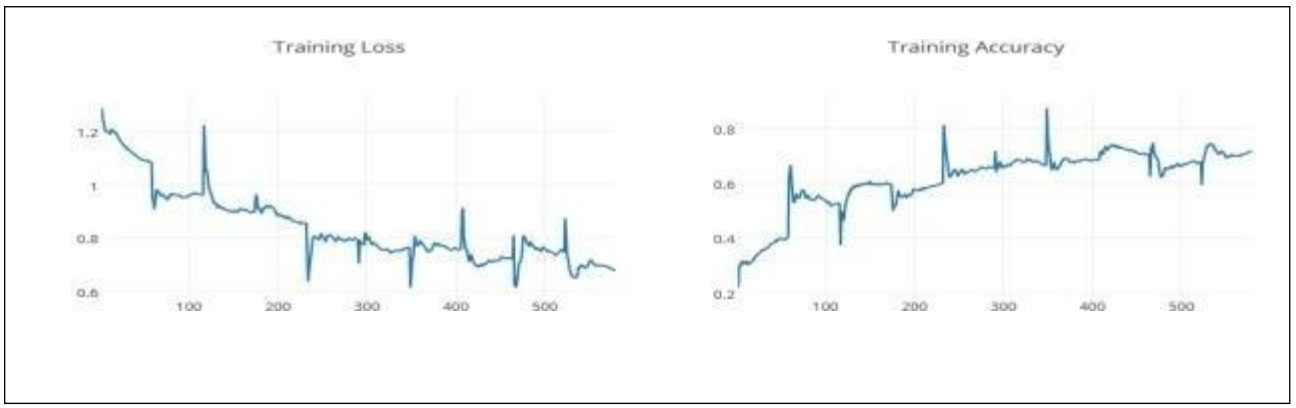


Fig.8: Plots showing training accuracy, training loss, validation accuracy and validation loss for recognizing emotions anger, stress, and neutral using RestNet50

Fig.9: Plots showing training accuracy, training loss, validation accuracy and validation loss for recognizing emotions anger, stress, and neutral using SqueezeNet



Annals of R.S.C.B., ISSN:1583-6258, Vol.24, Issue 1, 2020, Pages. 1400-1414
Received 15 April 2020; Accepted 23 June 2020.

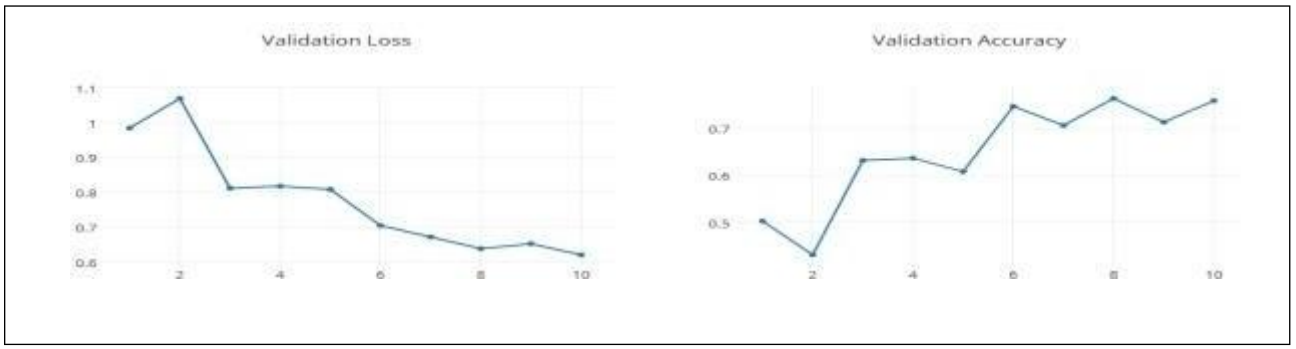


Fig.10: Plots showing training accuracy, training loss, validation accuracy and validation loss for recognizing emotions anger, stress, and neutral using WideResNet

Fig.11: Plots showing training accuracy, training loss, validation accuracy and validation loss for recognizing emotions anger, stress, and neutral using InceptionV3

Table II carries a summary of training accuracy, validation accuracy, training loss, and validation loss obtained by all four models.

TABLE II: Summary of training & validation accuracy and training and validation loss obtained using ResNet50, SqueezeNet, WideResNet, and InceptionV3

Model	Training Accuracy	Validation Accuracy	Training Loss	Validation Loss

Annals of R.S.C.B., ISSN:1583-6258, Vol.24, Issue 1, 2020, Pages. 1400-1414
Received 15 April 2020; Accepted 23 June 2020.

ResNet50_	99.7	99.8	0.0236	0.023
SqueezeNet_	57.8	46.3	7.2	9.6

WideResNet_	85.7	78.9	6.12	6.23
InceptionV3_	38.5	38.5	9.8	9.8

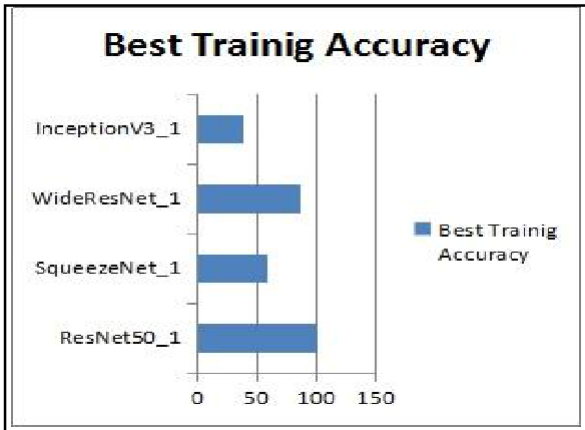


Fig.12 : Comparison of best training accuracy of InceptionV3, WideResNet, SqueezeNet, and ResNet50

Fig-13: Comparison of minimum training loss of InceptionV3, WideResNet, SqueezeNet and ResNet50

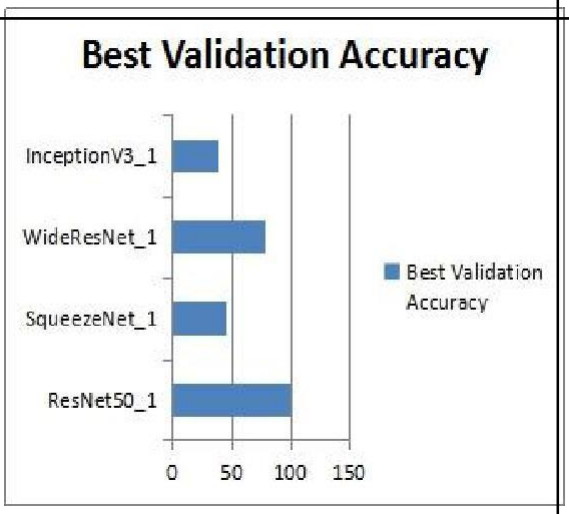


Fig.14: Comparison of best validation accuracy of InceptionV3, WideResNet,

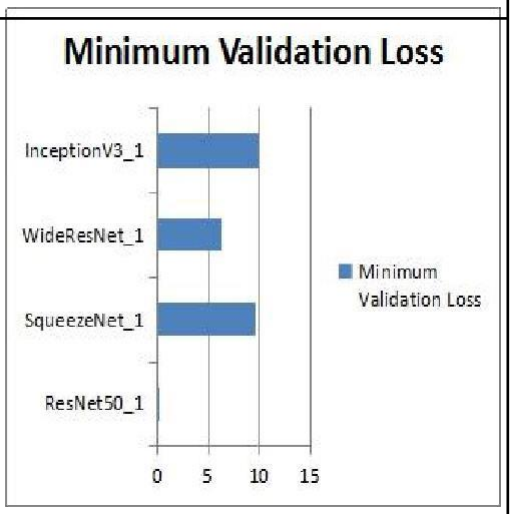


Fig.15: Comparison of minimum validation loss of InceptionV3,

SqueezeNet, and ResNet50	WideResNet, SqueezeNet, and ResNet50

6. Conclusion and future work

In the proposed work we have tried to investigate that whether the representation of sound as spectrogram images is a better approach for speech emotion classification in comparison to traditional features. We also tried analyzing the efficiency and ease of using a pre-trained Network for speech emotion classification. We also compared the performances of the pre-trained Network in classifying emotions anger, stress, and neutral. No doubt pre-trained Networks are easy to implement and show better performance in comparison to custom-designed CNN but the major limitation of implementing pre-trained Networks is the requirement of huge processing power and memory. Due to the scarcity of real-world datasets, the majority of current studies focus on transferring learning across lab-collected datasets. Transfer learning is the process of taking knowledge and insight from one issue domain and applying it to another. Although frequently applied in practice, transfer learning theory is still in its infancy. Previous research has shown that transfer learning approaches may be used to develop reliable emotion detection systems, with better results than in-domain learning (training and testing models on samples from the same dataset). Future work in this direction could be to investigate the proposed technique on the data obtained from real life settings.

References:

[1]. Elandeef, E. A. E., & Hamdan, A. H. E. (2021). Spoken English Production and Speech Reception Processes from Sentence Structure Perspective. International Journal of Linguistics, Literature and Translation, 4(3), 33-41.
[2]. Abbaschian, B. J., Sierra-Sosa, D., & Elmaghraby, A. (2021). Deep learning techniques for speech emotion recognition, from databases to models. Sensors, 21(4), 1249.
[3]. Swain, M.; Routray, A.; Kabisatpathy, P. Databases, features and classifiers for speech emotion recognition: A review. Int. J. Speech Technol. 2018, 21, 93–120, doi.org/10.1007/s10772-018-9491-z. [CrossRef]
[4]. Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. Computational intelligence and neuroscience, 2018.
[5]. Huang, C., Gong, W., Fu, W., & Feng, D. (2014). A research of speech emotion recognition based on deep belief Network and SVM. Mathematical Problems in Engineering, 2014.
[6].Hinton, G. E., Osindero, S., & Teh, Y. W., "A fast learning algorithm for deep belief Nets," Neural computation, 18 (7), 1527 1554 (2006). https://doi.org/10.1162/neco.2006.18.7.1527 Google Scholar
[7] .LeCun, Y., Bengio, Y., & Hinton, G., "Deep learning.," Nature, 521 (7553), 436 444 (2015). https://doi.org/10.1038/nature14539 Google Scholar
[8] Schmidhuber, J., "Deep learning in neural Networks: An overview.," Neural Networks, 61 85 – 117 (2015). https://doi.org/10.1016/j.neuNet.2014.09.003 Google Scholar

[9] Sze, V., Chen, Y. H., Yang, T. Y., & Emer, J. S., "Efficient processing of deep neural Networks: A tutorial and survey.," Proceedings of the IEEE, 105 (12), 2295 –2329 (2017). https://doi.org/10.1109/JPROC.2017.2761740 Google Scholar
[10]. Zacarias-Morales, N., Pancardo, P., Hernández-Nolasco, J. A., & Garcia-Constantino, M. (2021). Attention-Inspired Artificial Neural Networks for Speech Processing: A Systematic Review. Symmetry, 13(2), 214.
[11]. Deng, J., Zhang, Z., Marchi, E., & Schuller, B. (2013, September). Sparse autoencoder-based feature transfer learning for speech emotion recognition. In 2013 humaine association conference on affective computing and intelligent interaction (pp. 511-516). IEEE.
Proposed transfer learning for speech emotion detection.
[12]. Song, P., Jin, Y., Zhao, L., & Xin, M. (2014). Speech emotion recognition using transfer learning. IEICE TRANSACTIONS on Information and Systems, 97(9), 2530-2532.
[13]. Stolar, M. N., Lech, M., Bolia, R. S., & Skinner, M. (2017, December). Real time speech emotion recognition using RGB image classification and transfer learning. In 2017 11th International Conference on Signal Processing and Communication Systems (ICSPCS) (pp. 1-8). IEEE.
[14]. Sahoo, Sourav, Puneet Kumar, Balasubramanian Raman, and Partha Pratim Roy. "A segment level approach to speech emotion recognition using transfer learning." In Asian Conference on Pattern Recognition, pp. 435-448. Springer, Cham, 2019.
[15]. Zhou, S., & Beigi, H. (2020). A transfer learning method for speech emotion recognition from automatic speech recognition. arXiv preprint arXiv:2008.02863.
[16]. Ouyang, X., Kawaai, S., Goh, E. G. H., Shen, S., Ding, W., Ming, H., & Huang, D. Y. (2017, November). Audio-visual emotion recognition using deep transfer learning and multiple temporal models. In Proceedings of the 19th ACM international conference on multimodal interaction (pp. 577-582).
[17]. Deng, J., Zhang, Z., & Schuller, B. (2014, August). Linked source and target domain subspace feature transfer learning--exemplified by speech emotion recognition. In 2014 22nd International Conference on Pattern Recognition (pp. 761-766). IEEE.
[18]. Latif, S., Rana, R., Younis, S., Qadir, J., & Epps, J. (2018). Transfer learning for improving speech emotion classification accuracy. arXiv preprint arXiv:1801.06353.
[19]. Deng, J., Xia, R., Zhang, Z., Liu, Y., & Schuller, B. (2014, May). Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition. In 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 4818-4822). IEEE.
[20]. Deng, J., Frühholz, S., Zhang, Z., & Schuller, B. (2017). Recognizing emotions from whispered speech based on acoustic feature transfer learning. IEEE Access, 5, 5235-5246.
[21]. Triantafyllopoulos, A., & Schuller, B. W. (2021, June). The Role of Task and Acoustic Similarity in Audio Transfer Learning: Insights from the Speech Emotion Recognition Case. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7268-7272). IEEE.on Intelligent Data and Security (IDS) (pp. 96-99). IEEE.