EasyChair Preprint

№ 1110

Novel Deep Learning Architectures: Classification Accuracy Improvement

Rama Murthy Garimella and Inthiyaz Basha Kattubadi

June 8, 2019

Novel Deep Learning Architectures: Classification Accuracy Improvement

Inthiyaz Basha Kattubadi

IIIT-RKValley, RGUKT-AP, AB2-T20, BH1-F25, Idupulapaya, Vempalli, Kadapa, Andhra Pradesh 516330
r141176@rguktrkv.ac.in

Dr. Rama Murthy Garimella

Professor, Department of Computer Science, Mahindra Ecole Centrale, 1A, Survey No:62, Bahadurpally, Hyderabad, Telangana 500043
rama.murthy@mechyd.ac.in

Abstract

In future, emotion classification, object classification etc, by machines will play an important role. In this research paper, we proposed a series connection of Convolutional Neural Network (CNN) and Auto-Encoder (AE) for classification problems. We proposed a total of three architectures. We applied these architectures for the emotion classification. Among the three architectures, two architectures are trained with JAFFE (Japanese Female Facial Expressions), remaining one architecture trained with Berlin Database of Emotional Speech. We attained better classification accuracy than earlier efforts. We expect that such architectures will provide better classification accuracy in other applications also.

Keywords: Convolutional Neural Networks (CNN's), Auto-Encoders (AE's), Classification, Emotion, Series.

1. Introduction

With excellent performance of deep learning systems [particularly Convolution Neural Networks (CNN's)] in various applications, researchers are exploring different architectures. It was realized that increasing the number of hidden layers with a large amount of data leads to better than human accuracy in many applications. In many applications, data augmentation is employed when the training data is limited.

In deep learning research efforts, auto-encoders (traditional as well as convolutional) were employed to provide the dimensionality reduction. Researchers attempted to provide the reason for the high classification accuracy of CNN's. There is a common agreement that convolutional and pooling layers (trained by suitable masks/kernels/filters) are effectively extracting features that are fed to the fully connected layers.

In ensemble classifiers, multiple classifiers are utilized to provide better accuracy. To the best of our knowledge, we are the first group to utilize an interconnection (Series/Parallel) of convolutional/pooling layers, auto-encoders for improving the classification accuracy. The effort was mainly experimental and particularly applied to the problem of emotion classification. In this research paper, we are successful in training such novel deep learning architectures for emotion classification.

This research paper is organized as follows. In Section-II related research work is presented. In Section-III, our proposed architectures are presented. In Section-IV how to build and train these architectures is presented. In Section-V experimental results are discussed. In Section-VI innovative ideas are briefly summarized. The research paper concludes in Section-VII.

2. Related Work

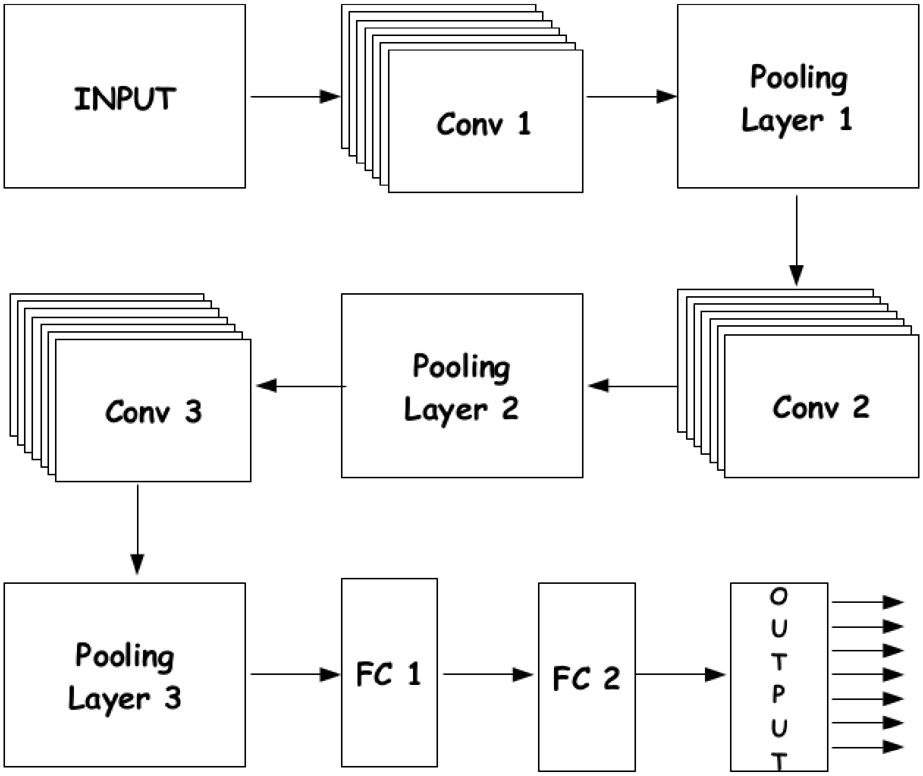
In Machine Learning research, researchers experimented with the idea of interconnecting classifiers for improving classification accuracy. To the best of our knowledge series connection of feature extractor output of convolutional and pooling layers (after training with same data), trained auto-encoder (with trained auto-encoder being the first stage) were utilized for emotion classification in[1]. They found that the classification accuracy was very poor. They concluded that the Series connection of feature extractors fed to fully connected layers in general leads to poor classification accuracy. After contemplating the reason for poor performance, we exchanged the order of auto-encoder and convolution/pooling layers. As expected such a novel deep learning architecture provided good classification accuracy. We are also aware that the utilization of the ensemble of classifiers was recognized as a promising idea.

3. Proposed Architectures

In the following discussion, we provide details of novel deep learning architectures which provide better classification accuracy. These architectures are proposed based on experiments we conducted using CNN's, Auto-Encoders.

The Convolutional Neural Networks and the Auto-Encoders which are used in our architectures are illustrated with the following diagrams.

Convolutional Neural Network - 1

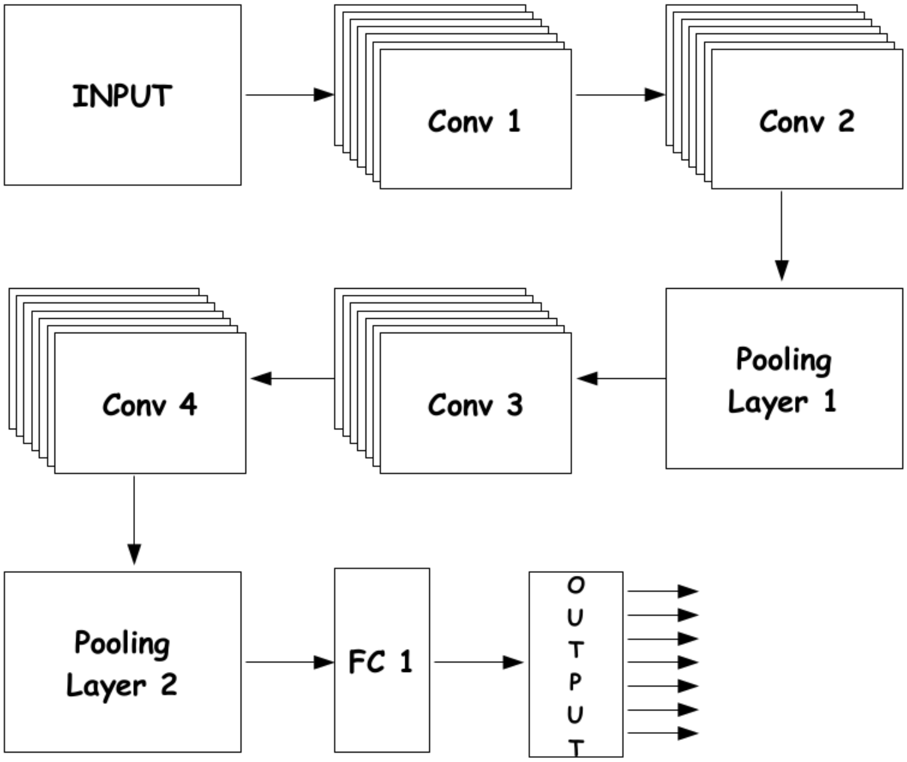


The Convolutional Neural Network is having 3 Convolution layers, 3 Pooling layers and 3 Dense layers including Output layer.

Details of the Architecture :-

INPUT: Here we give our data-set.
CONV 1: We have used kernel size as 3 X 3, Depth (Mask) 16, padding same and Activation ReLU.
CONV 2: We have used kernel size as 3 X 3, Depth (Mask) 8, padding same and Activation ReLU.
CONV 3: We have used kernel size as 3 X 3, Depth (Mask) 8, padding same and Activation ReLU.
Pooling Layer 1,2 and 3: We have used Max Pooling in this architecture with kernel size as 2 X 2 and Padding Same.
FC1: Contains 256 neurons and ReLU as activation function.
FC2: Contains 126 neurons, ReLU as activation function and Dropout as 0.6.
OUTPUT: Output layer contains N neurons representing each class and softmax as the activation function (depending on the number of classes the number of neurons (N) in the output layer changes).

Convolutional Neural Network - 2



The Convolutional Neural Network which is having 4 Convolution layers, 2 Pooling layers and 2 Dense layers including Output layer.

Details of the Architecture :-

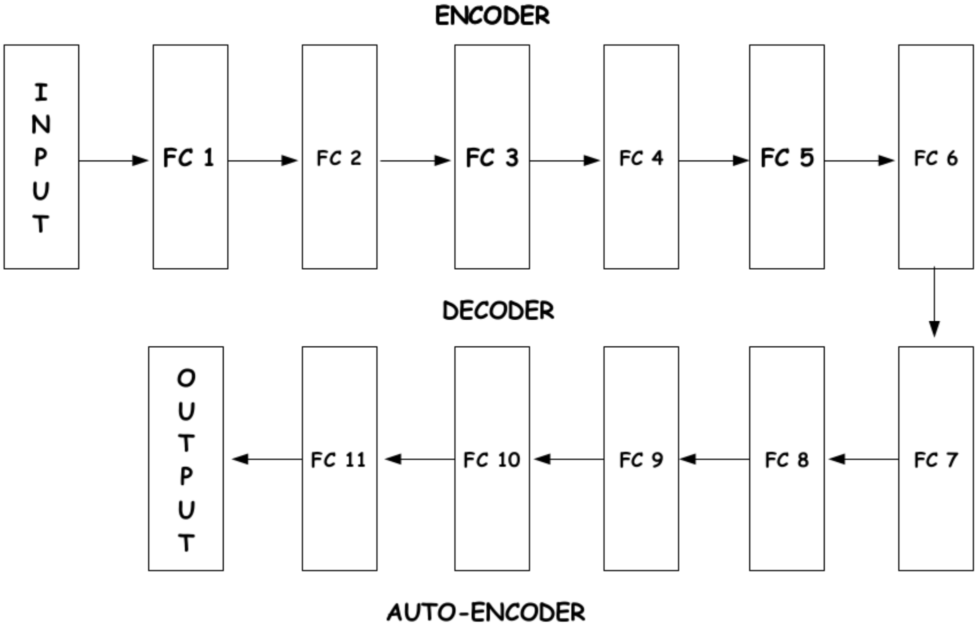
INPUT: Here we give our data-set.
CONV 1: Here we have used kernel size as 13 X 13, Depth 8, padding same and the Activation function as ReLU.
CONV 2: Here we have used kernel size as 13 X 13, Depth 8, padding same and the Activation function as ReLU.
CONV 3: Here we have used kernel size as 13 X 13, Depth 8, padding same and the Activation function as ReLU.
CONV 4: Here we have used kernel size as 3 X 3, Depth 8, padding same and the Activation function as ReLU.

Pooling Layer 1 and 2: Here we have used Max Pooling in this architecture with kernel size as 2 X 1 and Padding Same.

FC1: Contains 64 neurons, ReLU as activation function and Dropout as 0.2.

OUTPUT: Output layer contains N neurons representing each class and Softmax as the Activation function (depending on the number of classes the number of neurons (N) in the output layer changes).

Traditional Auto-Encoder



The Traditional Auto-Encoder which is having 13 Dense layers including input, output layer.

Details of the Architecture :-

INPUT: Here we give our data-set.

FC1: Contains 512 neurons, ReLU as activation function and Dropout as 0.1.

FC2: Contains 256 neurons, ReLU as activation function and Dropout as 0.2.

FC3: Contains 128 neurons, ReLU as activation function and Dropout as 0.3.

FC4: Contains 64 neurons, ReLU as activation function and Dropout as 0.4.

FC5: Contains 32 neurons, ReLU as activation function and Dropout as 0.5.

FC6: Contains 16 neurons, ReLU as activation function.

FC7: Contains 32 neurons, ReLU as activation function and Dropout as 0.4.

FC8: Contains 64 neurons, ReLU as activation function and Dropout as 0.3.

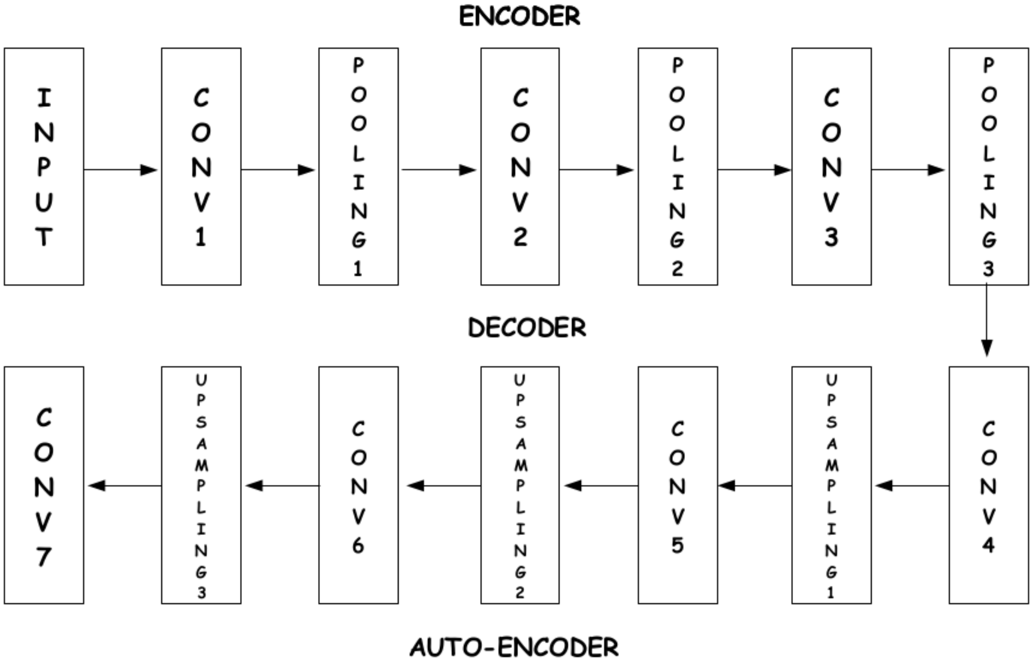
FC9: Contains 128 neurons, ReLU as activation function and Dropout as 0.2.

FC10: Contains 256 neurons, ReLU as activation function and Dropout as 0.1.

FC11: Contains 512 neurons, ReLU as activation function.

Output: Contains N neurons, sigmoid as activation function (N is equal to input dimension's).

Convolutional Auto-Encoder



The Convolutional Auto-Encoder which is having 7 Convolution layers, 3 Pooling layers and 3 Sampling layers.

Details of the Architecture :-

INPUT: Here we give our data-set.

CONV 1:- We have used kernel size as 3 X 3, Depth (Mask) 16, padding same and Activation ReLU.

CONV 2:- We have used kernel size as 3 X 3, Depth (Mask) 8, padding same and Activation ReLU.

CONV 3:- We have used kernel size as 3 X 3, Depth (Mask) 8, padding same and Activation ReLU.

POOLING 1,2 and 3:- We have used Max Pooling in this architecture with kernel size as 2 X 2 and Padding Same.

CONV 4:- We have used kernel size as 3 X 3, Depth (Mask) 8, padding same and Activation ReLU.

CONV 5:- We have used kernel size as 3 X 3, Depth (Mask) 8, padding same and Activation ReLU.

CONV 6:- We have used kernel size as 3 X 3, Depth (Mask) 16, padding same and Activation ReLU.

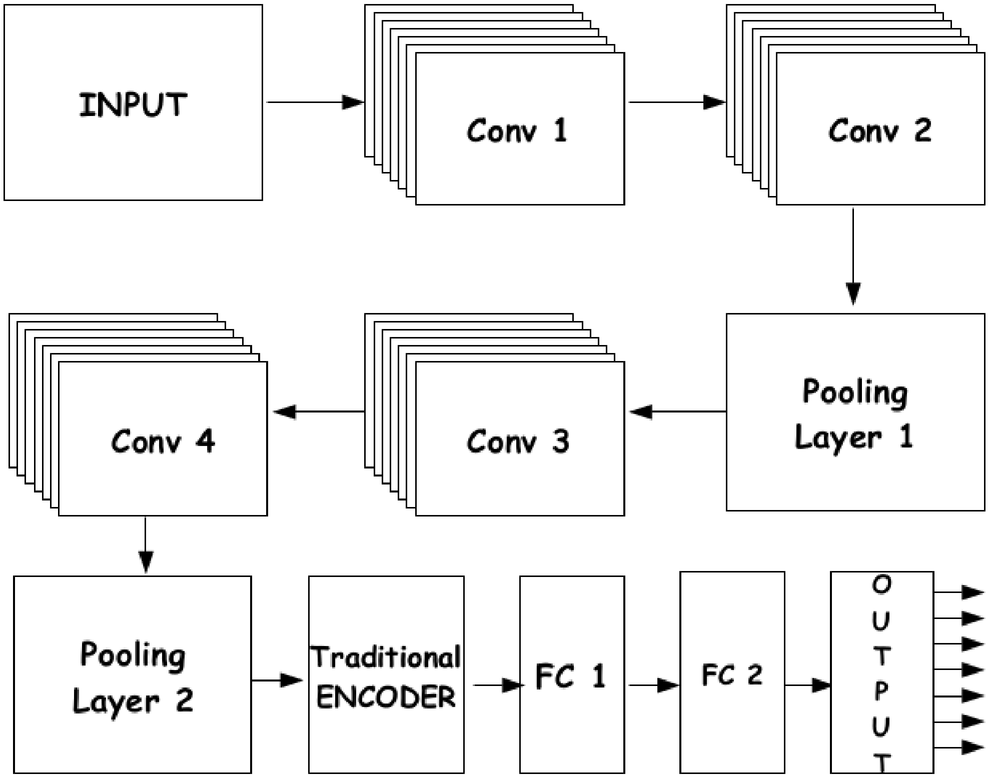
CONV 7:- We have used kernel size as 3 X 3, Depth(Mask) 1, padding same and Activation sigmoid.

UP SAMPLING 1,2 and 3:- Up Sampling with kernel size as 2 X 2 and Padding Same.

3.1. Input Image based Emotion Classification

Based on our intuition, we proposed the following architectures involving series connection of Convolutional Neural Network and Traditional/Convolutional Auto-Encoder. We discovered that such architectures provide better accuracy in image based emotion classification application.

3.1.1. Architecture 1: Series Combination of Convolutional Neural Network and Traditional Auto-Encoder



Series Combination of Convolutional Neural Network and Traditional Auto-Encoder

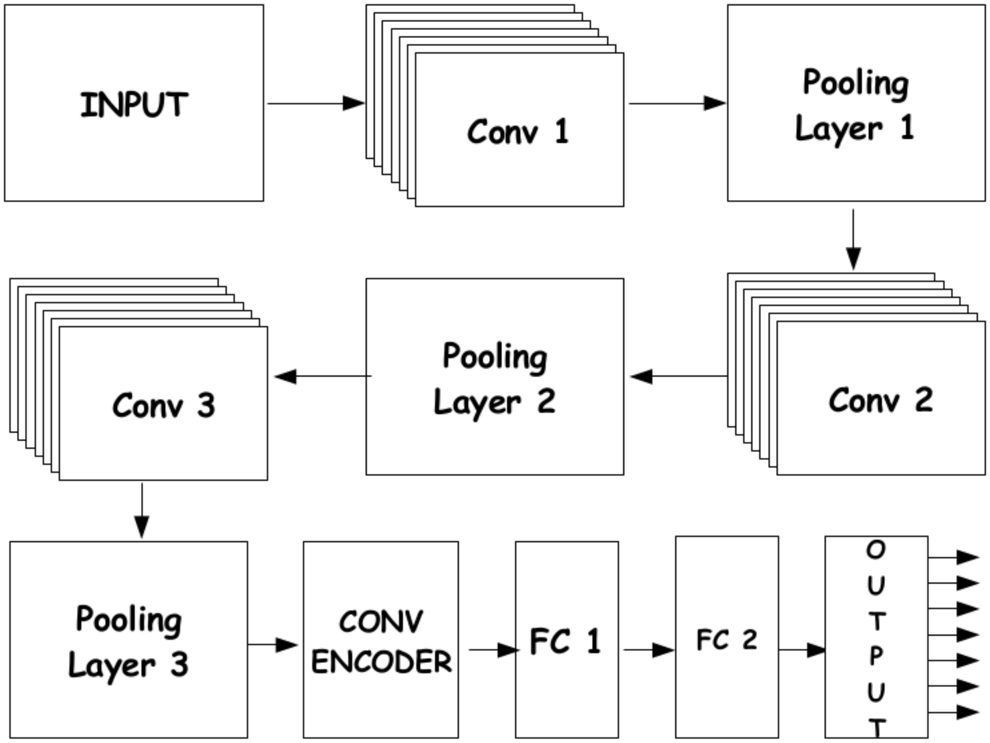
Details of Series combination of CNN and Traditional Auto-Encoder :-

FC1: Contains 256 neurons and ReLU as activation function.

FC2: Contains 126 neurons, ReLU as activation function and Dropout as 0.6.

OUTPUT: Output layer contains 8 neurons representing each class and softmax as the activation function.

3.1.2. Architecture 2: Series Combination of Convolutional Neural Network and Convolutional Auto-Encoder



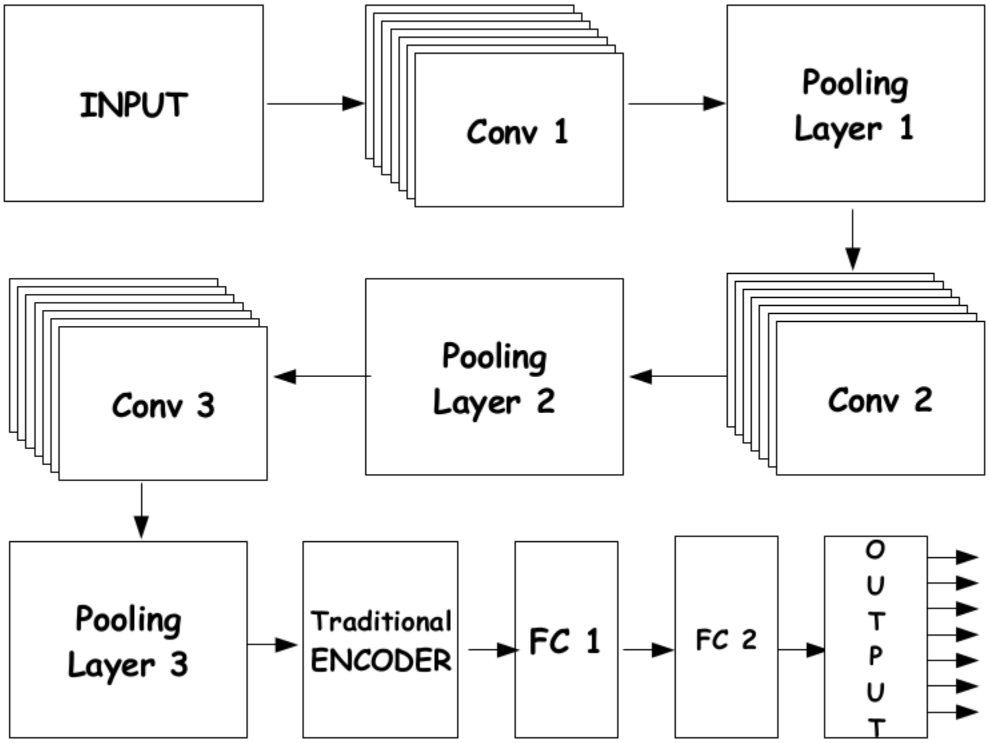
Series Combination of Convolutional Neural Network and Convolutional Auto-Encoder

Details of Series combination of CNN and Convolutional Auto-Encoder :-
FC1: Contains 256 neurons and ReLU as activation function.
FC2: Contains 126 neurons, ReLU as activation function and Dropout as 0.6.
OUTPUT: Output layer contains 8 neurons representing each class and softmax as the activation function.

3.2. Speech Signal based Emotion Classification

As in case '3.1' above, we realized that the following architecture involving series connection of Convolutional Neural Network and traditional Auto-Encoder provides better accuracy in speech signal based emotion classification.

3.2.1. Series Combination of Convolutional Neural Network and Traditional Auto-Encoder



Series Combination of Convolutional Neural Network and Traditional Auto-Encoder

Details of Series combination of CNN and Traditional Auto-encoder :-

FC1: Contains 256 neurons and ReLU as activation function.

FC2: Contains 126 neurons, ReLU as activation function and Dropout as 0.6.

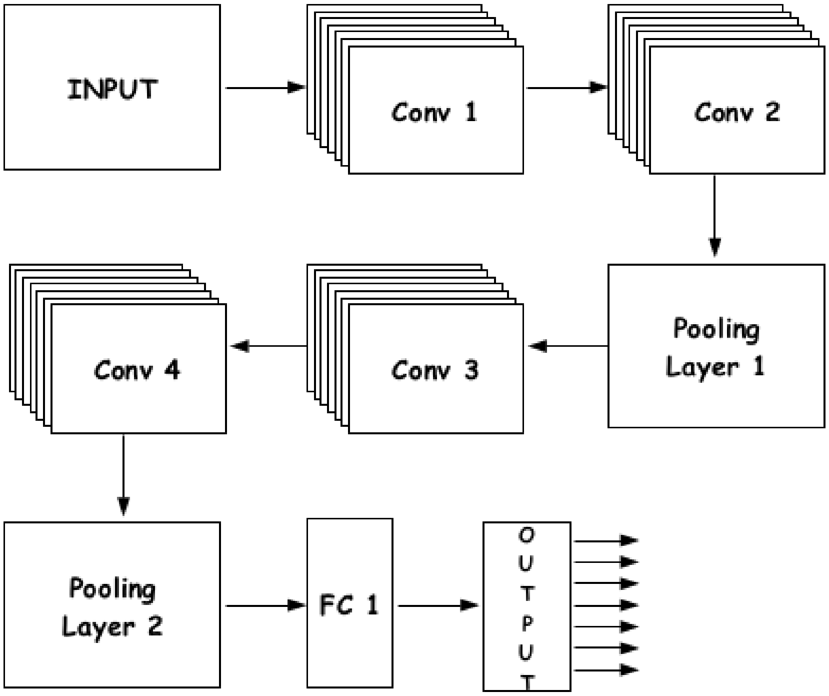
OUTPUT: Output layer contains 8 neurons representing each class and softmax as the activation function.

4. How to build and train these architectures

Here we are going to explain how to build these architectures and also, how to train them. Let me explain for one architecture. The procedure is the same for the remaining architectures.

Input Image based Emotion Classification
Architecture 1: Series Combination of Convolution Neural Network and Traditional Auto-Encoder:

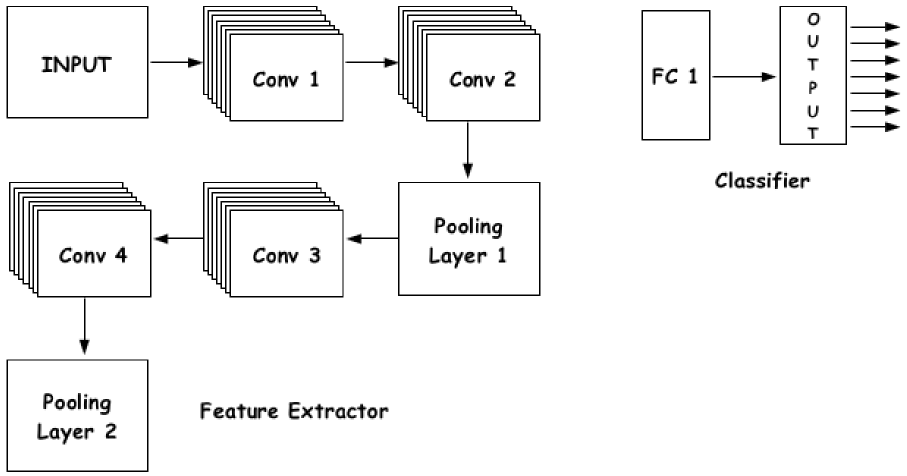
1. Construct the Convolution Neural Network architecture. The architecture is illustrated with the following diagram. Details are given in Proposed Architectures, Convolution Neural Network - 2.



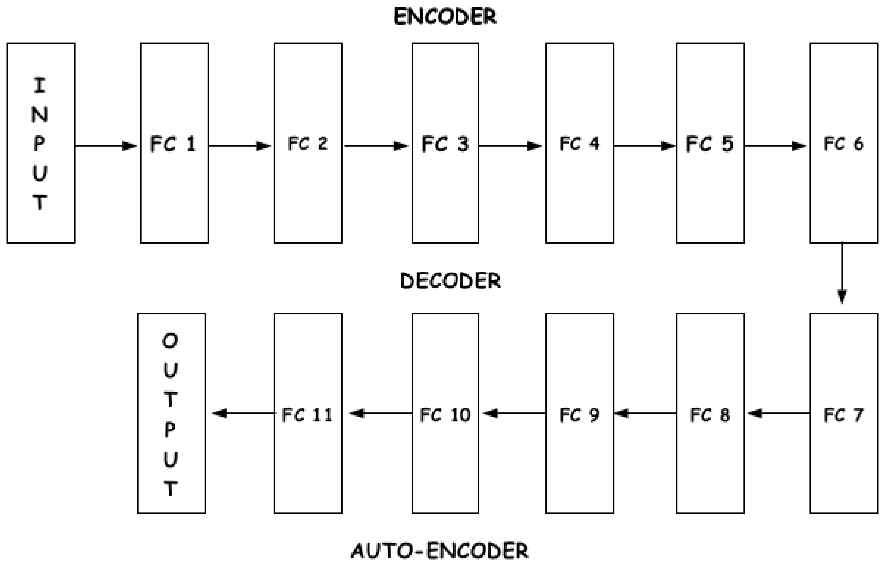
Convolutional Neural Network

2. Train the model.
3. Check if the model is overfitting or not by plotting the graph between training and testing accuracy.

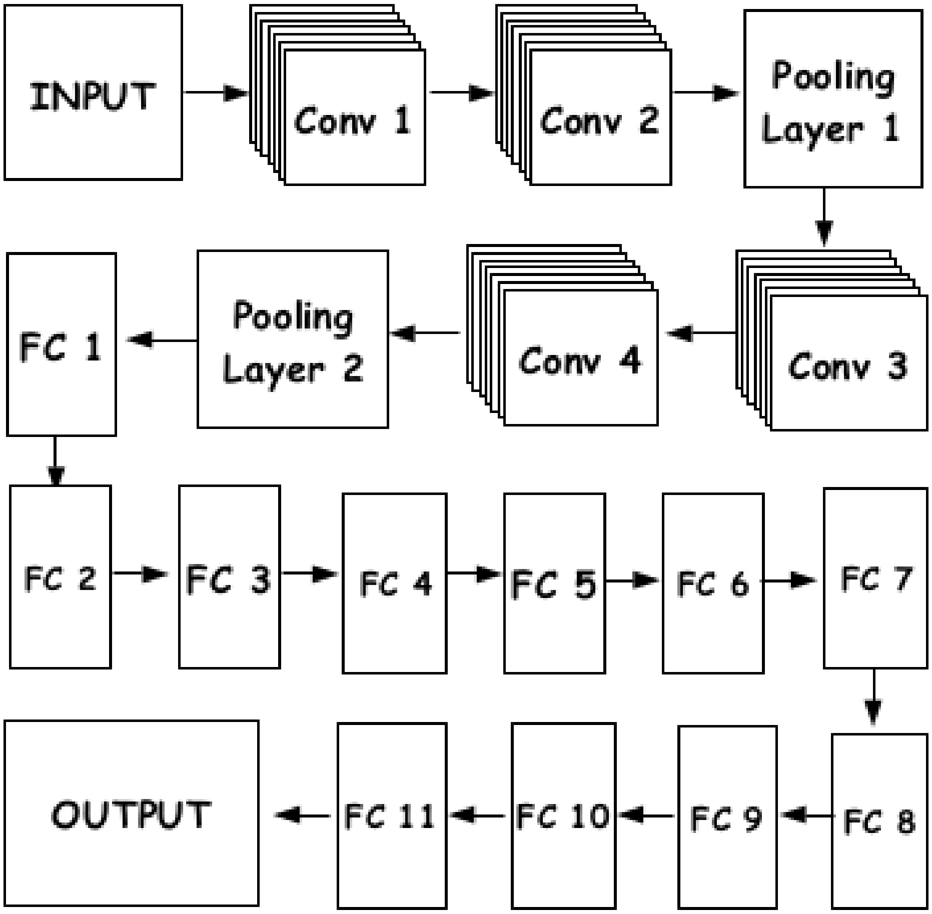
4. Saved the weights of Convolution Neural Network.
5. Divide the Convolutional Neural Network into feature extractor (from Input to Pooling Layer 2) and classifier (from FC1 to Output).



6. Built a traditional Auto-Encoder. The architecture is illustrated with the following diagram. Details are given in Proposed Architectures, Traditional Auto-Encoder.



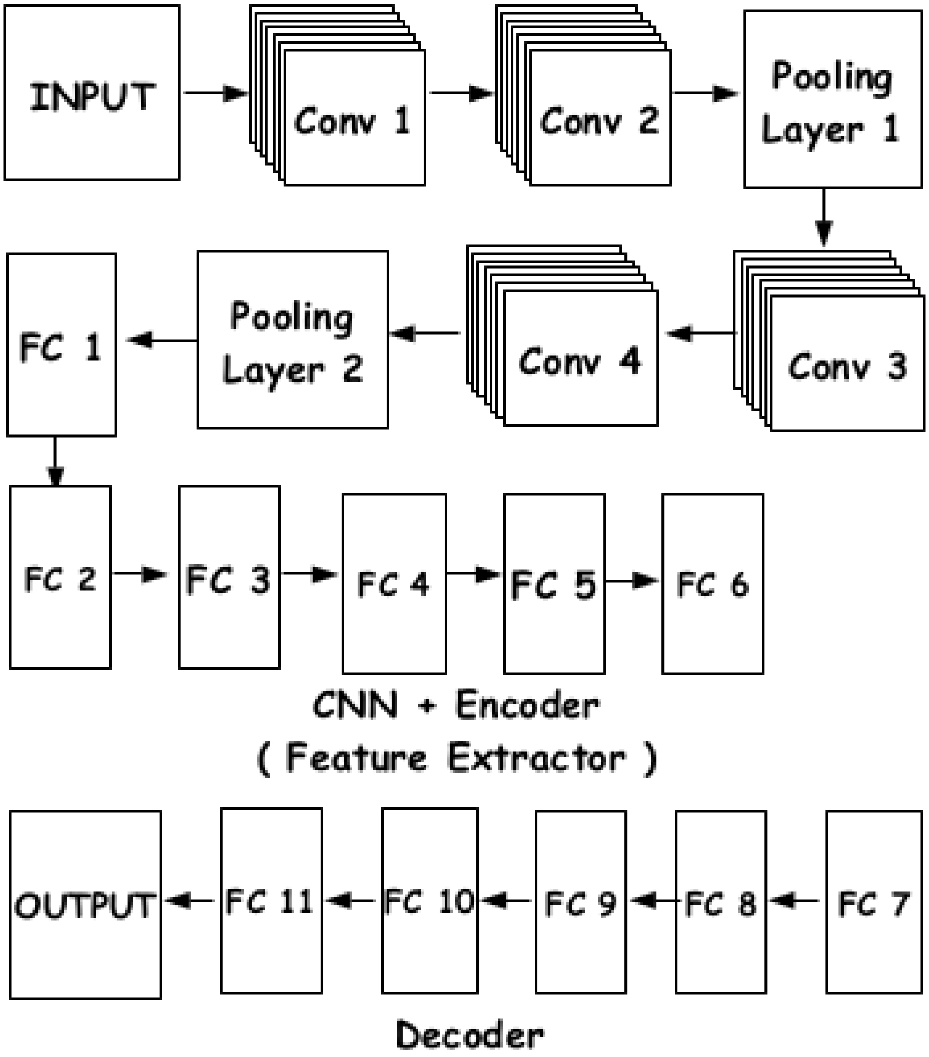
7. Cascade it to the feature extractor (give the output of feature extractor as input to the auto-encoder). This is our updated model.



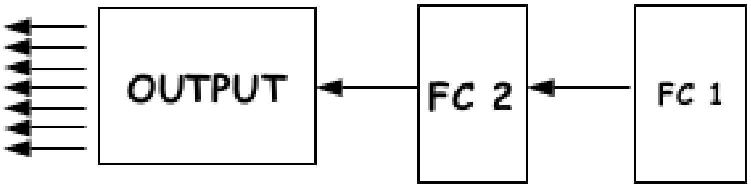
CNN + Auto-Encoder

8. After updating the model, freeze the weights of the feature extractor and kept them as noniterable.
9. Now, again train the whole model.
10. After training the model, apply fine-tuning.
11. Again divide the model into two parts. The first part is feature extractor, Encoder (from the input of CNN to traditional Auto-Encoders FC 6 layer)

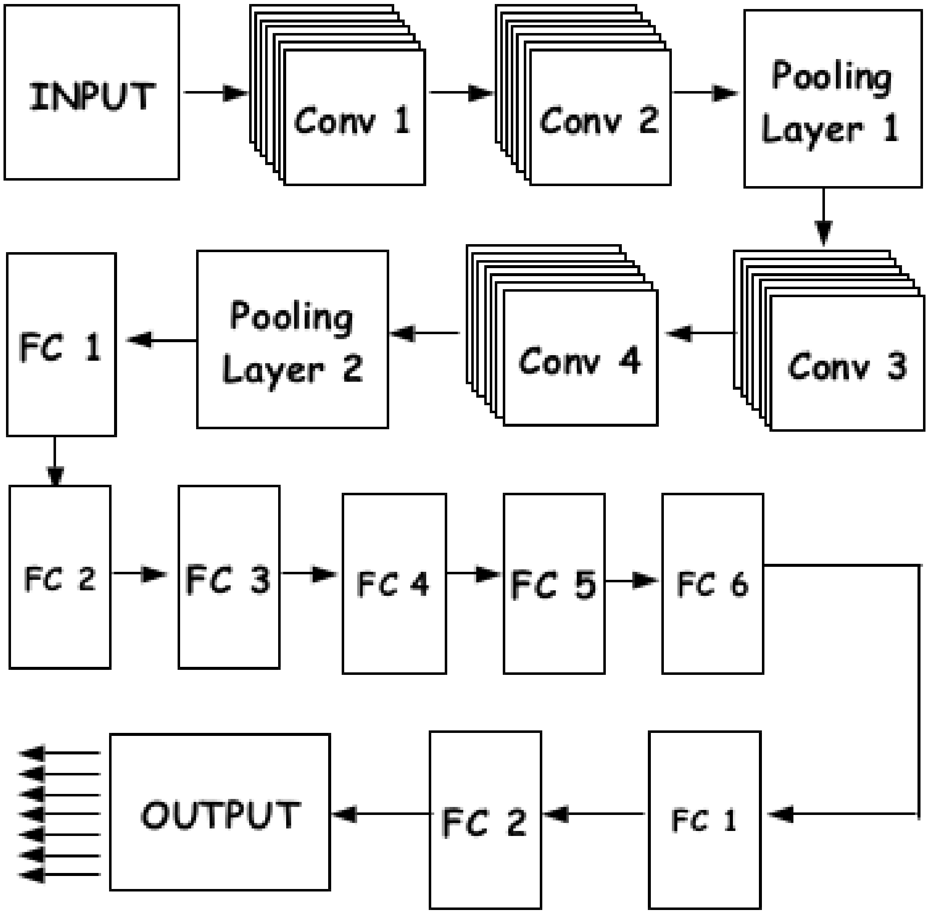
and the second part is the Decoder (traditional Auto-Encoders FC7 to Output layer).



12. Now, our updated feature extractor is first part.
13. Now, give this updated feature extractor output as input to the three layer Perceptron classifier.



Classifier



Feature Extractor + Classifier

14. Freeze the weights of feature extractor and kept them as noniterable.
15. Again train the whole model.

16. After training the model for third time, again apply fine-tuning.
17. Now construction of our model is completed. Save the weights of our model.

5. Experimental Results

We have trained all the architectures as explained in the Section-IV.

5.1. Input Image based Emotion Classification

5.1.1. Accuracy - Architecture 1

By training this Series Combination of Convolution Neural Network and Traditional Auto-Encoder, we have experienced classification accuracy of 85-88%.

5.1.2. Accuracy - Architecture 2

By training this Series Combination of Convolution Neural Network and Convolution Auto-Encoder, we have experienced classification accuracy of 80-84%.

5.2. Speech Signal based Emotion Classification

5.2.1. Accuracy - Architecture 1

By training this Series Combination of Convolution Neural Network and Traditional Auto-Encoder, we have experienced classification accuracy of 70-73%.

6. Discussion on Innovative Ideas

It is well recognized in the deep learning research community that convolutional/pooling layers (in CNN's), as well as traditional/convolutional auto-encoders essentially, act as feature extractors associated with the masks/kernels/filters. Based on such insight, we proposed a serial/parallel connection of such feature extractors as novel deep learning architectures in[1][2][3]. We realized in this research paper that order of cascading the feature extractors has a significant impact on classification accuracy. The potential reason for such successful deep learning architecture is explored in the following:

1. Dimensionality reduction of features extracted by convolutional/pooling layers by auto-encoder traditional/convolutional.
2. Novel architecture ensures finer discrimination of features enabling increased classification accuracy.
3. It is well known that the training process of Artificial neural networks results in fine tuning the weights for classification.

   In this research paper the convolutional/pooling layer based feature extractor (with weights being freezed after training process) connected in cascade with traditional/convolutional auto-encoder and resulting architecture trained all over to get better classification accuracy. The most probable reason for better accuracy is that the cascading process enables arriving at (class) decision boundaries that are matched to the training data as best as possible. The encoder-decoder pair in the traditional/convolutional encoder ensures arriving at decision boundaries that are well separated based on dimensionality reduction of features.

7. Conclusion

In this research paper we have introduced a total of three architectures, for the emotion classification problems. After training our architectures we conclude that, by connecting two are more feature extractors in series we will obtain better classification accuracy, cascading has the effect of extracting essential features that are fed to the fully connected layers.

8. References

[1] Siva Prasad Raju Bairaju, Soumya Ari, Rama Murthy Garimella, "Emotion Detection using Visual Information with Deep Auto-Encoders". IEEE

Sponsored 5^{th} International Conference for convergence in Technology 2019 (I2CT) [Accepted]

[2] Inthiyaz Basha Kattubadi, Dr. Rama Murthy Garimella,"Emotion Classification: Novel Deep Learning Architectures", 5^{th} International Conference on Advance Computing and Communication Systems2019 (ICACCS)[Accepted]

[3] Siva Prasad Raju Bairaju, Soumya Ari, Rama Murthy Garimella, "Facial Emotion Detection using Deep Auto-Encoders".(IEEE Xplore)Proc.International Conference in Electrical,Electronics&Communication Engineering. (ICRIEECE-2018) [Accepted]

[4] Yann LeCun, Yoshua Bengio, Geoffrey Hinton,(2015/5), "Deep Learning". To download the paper access this link https://doi.org/10.1038/nature14539.

[5] Yann LeCun, Yoshua Bengio,1995/4, "Convolutional networks for images, speech, and time series", Book: "The handbook of brain theory and neural networks", "Convolutional networks for images, speech, and time series".

[6] S. Alizadeh, A. Fazel, "Convolutional Neural Networks for Facial Expression Recognition", CoRR, abs/1704.06756,2017.

[7] Ariel Ruiz-Garcia, Mark Elshaw, Abdulrahman Altahhan, Vasile Palade, "Stacked Deep Convolutional Auto-Encoders for emotion recognition from Facial expressions", 2017 International Joint Conference on Neural Networks(IJCNN).

[8] G A Rajesh Kumar ; Ravi Kant Kumar ; Goutam Sanyal, "Facial emotion analysis using deep convolutional neural network",International Conference on Signal Processing and Communication 2017 (ICSPC-2017) .

[9] M.Matsugu , K.Mori , Y.Mitari , Y.Keneda, "Facial expression recogni-

tion combined with robust face detection in a convolutional neural network", "Proceedings of the International Joint Conference on Neural Networks, 2003".

[10] Dinh Viet Sang , Le Tran Bao Cuong , Do Phan Thuan, "Facial smile detection using convolutional neural net-works ", 9th International Conference on Knowledge and Systems Engineering 2017 (KSE-2017)".

[11] B Subarna , Daleesha M Viswanathan, "Real Time Facial Expression Recognition Based on Deep Convolutional Spatial Neural Networks", International Conference on Emerging Trends and Innovations In Engineering And Techno-logical Research 2018 (ICETIETR-2018).

[12] Andre Teixeira Lopes et al, "A Facial Expression Recognition System Using Convolutional Networks", Vol. 00, pg. 273-280,2015.

[13] Arushi Raghuvanshi, Vivek Choksi, "Facial Expression Recognition with Convolutional Neural Networks", CS23 1n Course Projects, Winter 2016.

[14] Prudhvi Raj Dachapally, "Facial Emotion Detection Using Convolutional Neural Networks and Representational Autoencoder Units", Published 2017 in ArXiv.

[15] https://medium.com/@chrisprinzz/facial-emotion-detection-using-deep-learning-44dbce28349c