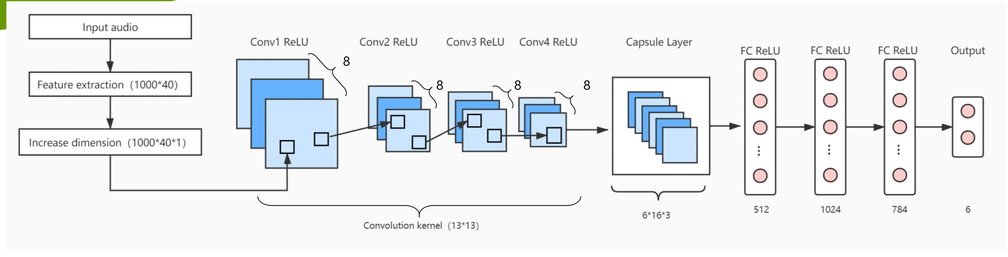
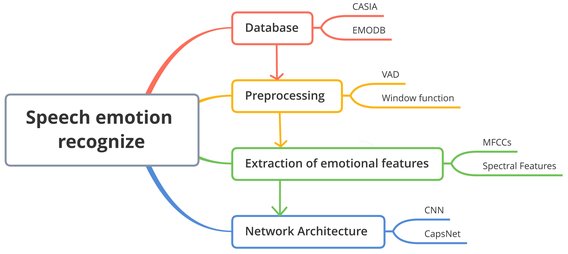
The Application of Capsule Neural Network Based CNN for Speech emotion recognition

Xin-Cheng Wen, Kun-Hong Liu, Wei-Ming Zhang and Kai Jiang
School of Informatics, Xiamen University, Xiamen, China



1 Background

Speech emotion recognition (SER) is an important and challenging task. It requires that a machine learning model process a person's speech signals and to judge his emotional state accurately. Due to the high dimensionality of the audio data, the extracted features are always noisy. Besides, the abstraction of audio features makes it impossible to fully use the inherent relationship among audio features.



2 Methods

This paper proposes a model that combines a convolutional neural network (CNN) and a capsule network (CapsNet), named as CapCNN. The advantage of CapCNN lies in that it provides a solution for time sensitivity, and gives the overall characteristics.

3 Results

The proposed method can extract the time series and space information effectively, achieving a classification accuracy of 81.90% and 82.90% on the CASIA and EMODB. CapCNN can well handle the high-order data with a large scale of features and retain the most valuable information. And it also performs well on the task sensitive to time series.

	Angry	Fear	Happy	Neutral	Sad	Surprise	Average
LSTM+MFCC	66.67	58.33	82.86	71.05	53.33	68.57	65.83
CNN+MFCC	51.28	27.08	37.14	52.63	31.11	48.57	40.42
CNN+Spectrogram	79.49	35.42	25.71	52.63	44.44	54.29	48.33
CNN+Spectral Features	74.36	70.83	68.57	73.68	60	82.86	71.25
CapCNN+MFCC	74.36	62.5	74.29	92.1	55.56	80	72.08
CapCNN+Spectrogram	79.49	64.59	80	97.36	57.78	80	75.42
CapCNN+Spectral Features	89.74	77.08	85.71	92.1	66.67	85.71	82.08

	Angry	Fear	Happy	Neutral	Sad	Average
LSTM+MFCC	71.88	66.67	70	64.29	72.73	69.51
CNN+MFCC	81.25	26.67	60	50	36.36	57.32
CNN+Spectrogram	53.13	13.33	50	64.29	27.27	43.9
CNN+Spectral Features	96.88	0	10	92.86	63.64	63.41
CapCNN+MFCC	88.89	71.43	50	66.67	50	70.83
CapCNN+Spectrogram	87.5	66.67	30	92.86	81.82	76.83
CapCNN+Spectral Features	87.75	86.67	40	85.71	100	82.93

40 50 60 70 80 90 100

4 Conclusion

This paper proposed a new framework for SER based on CNN and CapsNet. As is found that the local feature information in audio is of great significance to the SER task, and different features of audio would contribute to the recognition task differentially, such as frequency, tone, and so on. The proposed CapCNN combined CNN and CapsNet to overcome the shortcomings of CNN, so as to obtain better results in the field of SER. In experiments, CapCNN is compared with some widely used deep learning modes, and experimental results confirm that our algorihtms can achieve good results on both data sets due to the full utilization of the local feature information.

This study shows that the application of CapsNet to the SER task is promising. So in the future, we would try some new CapsNet structure/algorithm in this field.