DNN, CNN, RNN, CRNN 각 모델과 오디오 특징들에 따른 정확도 및 속도 비교

개요:

- DNN, CNN, RNN, CRNN 각 모델들 간의 성능 비교
- Audio Data 로부터 추출할 수 있는 다양한 특징들에 대해서 음성인식 성능 비교

1. 사용 데이터 셋: Speech Commands v0.01
2. 사용 특징:

Feature	설명
Spectral Centroid	소리의 밝기 성분을 나타내고 주로 음악 분류 등에 사용된다.
Spectral Contrast	주파수 성분을 옥타브를 이용해 구분한다. 이 논문의 저자는 클래식 음악 분류를 위해 Mel_Spectrogram 의 대체 수단으로서 어느 정도 효율이 있는가를 서술하였는데, Mel_Spectrogram 을 뒤집어 놓은 모습과 유사하다.
Mel Spectrogram	인간의 청각 특성을 반영하여 저주파를 더 상세하게 구분하는 특징
MFCC	Mel Spectrogram 을 DCT 를 취하여 압축한 형태의 결과물
STFT	Short Time Fourier Transform, 짧은 시간의 FT 결과를 시간에 따라 나열하여 시간별 주파수 특징 정보를 나타냄

3. Model

i. DNN

Layer	Input Layer	16
	Drop Out	25%
	Hidden Layer 1	64
	Drop Out	25%
	Hidden Layer 2	64
	Drop Out	25%
	Output Layer	30 (# of classes)
Optimizer	Adam	Learning Rate 0.0001

FEATURE	모델 속도	ACCURACY	F1 SCORE
MEL SPECTROGRAM	42us/sample	0.8003	0.8008
STFT	102us/sample	0.5789	0.5770
SPECTRAL CONTRAST	39us/sample	0.3627	0.3574
SPECTRAL CENTROID	66us/sample	0.0363	0.0025
MFCC	482us/sample	0.6666	0.6714

ii. CNN

Layer	1D Convolutional Layer	64filters, kernel size : 7
	1D Convolutional Layer	128filters, kernel size : 7
	Max Pooling	pool size : 2
	1D Convolutional Layer	256filters, kernel size : 3
	1D Convolutional Layer	512filters, kernel size : 3
	Max Pooling	pool size : 2
	Drop Out	50%
	Fully Connected Layer	1024
	Drop Out	50%
	Fully Connected Layer	1024
	Drop Out	50%
	Output Layer	30 (# of classes)
Optimizer	Adam	Learning Rate 0.0001

FEATURE	모델 속도	ACCURACY	F1 SCORE
MEL SPECTROGRAM	101us/sample	0.8876	0.8880
STFT	230us/sample	0.8931	0.8935
SPECTRAL CONTRAST	24us/sample	0.2149	0.2056
SPECTRAL CENTROID	62us/sample	0.5042	0.5096
MFCC	31us/sample	0.8310	0.8315

iii. RNN

Layer	LSTM	256
	Drop Out	50%
	Time Distributed Dense Layer	128
	Time Distributed Dense Layer	64
	Time Distributed Dense Layer	32
	Time Distributed Dense Layer	16
	Drop Out	50%
	Fully Connected Layer	256
	Drop Out	50%
	Fully Connected Layer	256
	Output Layer	30 (# of classes)
Optimizer	Adam	Learning Rate 0.0001

FEATURE	모델 속도	ACCURACY	F1 SCORE
MEL SPECTROGRAM	79us/sample	0.8495	0.8500
STFT	195us/sample	0.8967	0.8499
SPECTRAL CONTRAST	7us/sample	0.2636	0.2410
SPECTRAL CENTROID	62us/sample	0.4749	0.4800
MFCC	23us/sample	0.7700	0.7908

iv. CRNN

Layer	1D Convolutional Layer	64filters, kernel size : 7
	1D Convolutional Layer	128filters, kernel size : 7
	Max Pooling	pool size : 2
	1D Convolutional Layer	256filters, kernel size : 3
	1D Convolutional Layer	512filters, kernel size : 3
	Max Pooling	pool size : 2
	LSTM	256
	Drop Out	50%
	Time Distributed Dense Layer	128
	Time Distributed Dense Layer	64
	Time Distributed Dense Layer	32
	Time Distributed Dense Layer	16
	Output Layer	30 (# of classes)
Optimizer	Adam	Learning Rate 0.0001

FEATURE	모델 속도	ACCURACY	F1 SCORE
MEL SPECTROGRAM	115us/sample	0.8404	0.8404
STFT	255us/sample	0.8477	0.8482
SPECTRAL CONTRAST	24us/sample	0.2044	0.1816
SPECTRAL CENTROID	91us/sample	0.3807	0.3842
MFCC	34us/sample	0.8025	0.8037

4. 결론

주로 음악 분류에 쓰이는 Spectral Contrast, Spectral Centroid 의 경우 넓은 주파수 대역의 특징을 구분하는 것에는 의미가 있을 수 있지만, 음성 데이터와 같은 좁은 주파수 영역에 대한 특징 추출 성능이 매우 떨어지는 것을 알 수 있습니다.

Mel Spectrogram 이 MFCC 보다 전반적인 정확도는 높지만 모델 속도 면에서 MFCC 가 압도적이기 때문에 음성 데이터를 처리하는데 있어서 MFCC 가 어느 정도 중요한 것인지 확인할 수 있습니다.

DNN 을 제외한 CNN, RNN, CRNN 에 대해서 RNN 의 속도가 조금 빠른 편인 것을 제외하고 큰 차이가 보이지는 않습니다. 그러나 PC 성능의 한계로 RNN 과 CRNN 학습 시 weight 업데이트가 되지 않거나, GPU 가 멈춰버리는 등의 물리적인 에러가 지속적으로 발생하여 RNN 과 CRNN 의 결과에 대한 신뢰도가 낮기 때문에 확답할 수는 없습니다.