Spectrogram을 이용한 CNN 기반 음성 감정인식

Speech Emotion Recognition based on CNN using Spectrogram

저자
(Authors)

박소은, 김대희, 권순일, 박능수
Soeun Park, Daehee Kim, Soonil Kwon, Neungsoo Park

출처
(Source)

정보 및 제어 논문집 , 2018.10, 240-241 (2 pages)
INFORMATION AND CONTROL SYMPOSIUM , 2018.10, 240-241 (2 pages)

발행처
(Publisher)

대한전기학회
The Korean Institute of Electrical Engineers

URL

http://www.dbpia.co.kr/Article/NODE07562655

APA Style

박소은, 김대희, 권순일, 박능수 (2018). Spectrogram을 이용한 CNN 기반 음성 감정인식. 정보 및 제어 논문집, 240-241.

Spectrogram을 이용한 CNN 기반 음성 감정인식

박소은[1], 김대희[1], 권순일[2], 박능수[1]
건국대학교 컴퓨터공학과[1]
세종대학교 소프트웨어학과[2]

Speech Emotion Recognition based on CNN using Spectrogram

Soeun Park[1], Daehee Kim[1], Soonil Kwon[2], Neungsoo Park[1]
Dept of Computer Science and Engineering, Konkuk University[1]
Dept of Software, Sejong University[2]

Abstract - 최근 사람과 기계 간의 상호작용을 통하여 개개인에 특화된 서비스를 제공하는 기술에 관한 연구가 활발히 진행되고 있다. 특히 Apple의 Siri나 Amazon의 Alexa와 같이 음성만을 이용한 인공지능 서비스에 관한 관심이 늘어나고 있다. 이에 따라 본 논문에서는 음성을 통해 사람의 감정을 인식하기 위해 CNN(Convolutional Neural Network)을 기반으로 한 음성 감정인식 모델을 제안한다. 음성 파일로부터 spectrogram을 추출하여 학습을 진행하고 학습된 결과를 이용하여 감정을 분류하였다. 총 5가지 감정(화남, 기쁨, 슬픔, 분노, 중립)을 분류한 결과 약 64%의 정확도를 보임을 확인하였다.

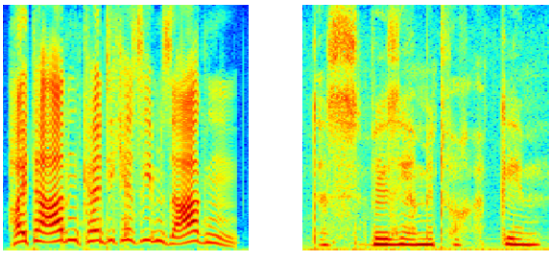
1. 서 론

사람과 기계 간의 상호작용에 관한 연구가 활발히 진행됨에 따라 사람과의 상호작용을 통해 서비스를 제공하는 시스템에 대한 요구도 증가하고 있다. 그중에서도 사람의 감정을 인식하고 개개인에 특화된 서비스를 제공하는 기술에 관한 관심이 높아지고 있다. 사람의 감정을 인식하는 기술에 관한 연구는 많이 진행되고 있으나 대부분 영상과 음성을 결합하여 사용하거나 영상만을 사용한 경우가 많다.[1, 2] 이에 따라 데이터 크기가 큰 영상과는 다르게 휴대전화와 같은 휴대용 개인 기기에서도 사용 가능한 음성만을 사용한 인공지능 서비스가 증대하고 있다.[3, 4] 이를 위해 본 논문에서는 사용자의 음성을 인식해 감정을 분류하기 위해 Spectrogram을 사용한 CNN 기반의 음성감정인식 모델을 제안한다. 실험은 Berlin EMO DB[5]를 사용하여 진행하였으며 총 5가지 감정(화남, 기쁨, 슬픔, 분노, 중립)을 분류한 결과 최대 64%의 정확도를 보였다. 본 논문의 구성은 다음과 같다. 2장에서는 제안한 CNN 기반 음성 감정인식 모델에 관해 설명하고 3장에서는 음성 감정인식 모델을 기반으로 한 실험 결과를 보인다. 마지막으로 4장에서는 결론을 나타낸다.

2. 본 론

2.1 Spectrogram

Spectrogram은 시간이 변화하면서 소리나 다른 신호의 강도나 세기가 변하는 것을 각각 다른 주파수에 따라 시각적으로 표현한 것이다. Spectrogram의 가로축은 시간을 나타내며 세로축은 주파수를 나타내며 소리나 신호의 강도나 세기가 변화함에 따라 spectrogram에서 표현되는 색도 달라진다. 특정 감정을 발화함에 따라 <그림 1>과 같이 다르게 표현되는 spectrogram을 이용하여 음성 데이터에서의 감정을 인식하고자 한다.



<그림 1> 각 감정에 따른 spectrogram

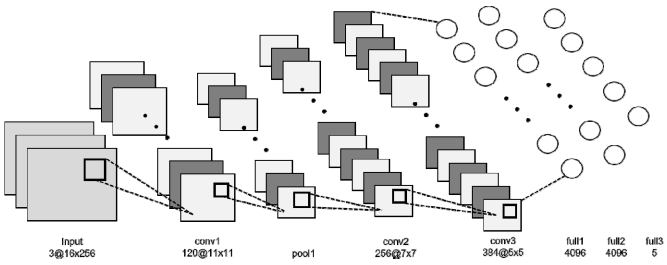
각각의 음성 데이터로부터 spectrogram을 추출 후 시간 축을 일정 간격으로 나누어 16x256형태를 입력 값으로 사용하였다. 음성 파일마다 시간이 다르므로 각 파일을 나누었을 때 나오는 개수도 다르다.

2.2 CNN을 이용한 음성 감정인식 기계학습 알고리즘

Spectrogram은 이미지 형태로 나타나기 때문에 이미지 인식에 특화된 CNN(Convolutional Neutral Network)를 사용하여 학습을 진행하였다.[6] 학습 시 데이터 개수의 불균형에 따른 치우침을 막고자 loss 값을 계산 시 weight function을 적용하여 이를 해결하였다. 또한, 추론을 위해서는 파일별로 작게 나누어진 이미지 각각이 분류되는 정보를 이용하여 하나의 파일이 분류되는 감정을 결정한다.

2.2.1 CNN 모델

CNN 모델은 현재 이미지 분류에서 특화된 모델로 이미지를 입력 데이터로 사용하여 high-level 특징값을 추출하여 학습을 진행한다. 본 논문에서 음성 감정인식을 위해 사용한 CNN 모델은 아래 <그림 2>와 같다.



<그림 2> CNN 모델

CNN 모델은 convolutional layer 3층, max pooling layer 1층 그리고 fully connected layer 3층으로 이루어져 있다. 첫 번째 convolutional layer는 11x11 커널 120개로 이루어져 있으며 활성 함수는 rectified linear units(ReLU)를 사용하였다. 그다음에는 3x1 크기에 stride 크기가 2인 max pooling layer가 나온다. 두 번째 convolutional layer는 7x7 커널 256개로 이루어져 있다. 세 번째 convolutional layer는 5x5 커널 384개로 구성되어 있다. 마지막 convolutional layer 뒤에는 fully connected layer(FC)가 이어지는데 각각 4096, 4096, 5의 뉴런을 가지고 있다. FC 다음에는 softmax를 사용하여 각 감정이 나올 확률이 몇 %인지를 나타내게 된다.

2.2.2 Weighted 학습

Spectrogram을 16x256 형태로 나누어 입력 값으로 사용할 경우 각 음성 파일의 길이가 다르므로 나뉘는 파일의 개수도 다르게 된다. 이 경우 대부분 음성 파일의 길이가 긴 슬픔 감정의 파일의 개수는 많이 나뉘지만 음성 파일의 길이가 짧은 화남 감정의 파일 개수는 적게 나뉘게 된다. 이 상태에서 학습을 진행하게 되면 슬픔의 감정에 치우쳐 학습되는 문제가 발생한다. 이를 해결하기 위해 학습 과정에 loss 값을 계산하는 단계에서 각 감정의 개수에 따라 weight 값을 다르게 주었다. 각 감정별 weight 값은 아래 식 (1)과 같이 계산하였다. $total_{input}$은 전체 데이터의 개수이며 $total_{class}$는 분류할 감정 종류의 개수이다. $dass_{input}[i]$는 각 감정에 있는 데이터의 개수를 나타낸다.

$$weight[i] = \frac{total_{input}}{total_{class}} / dass_{input}[i] \quad (1)$$

전체 데이터의 개수를 감정 종류의 개수로 나누면 데이터가 균일하게 분포되었을 시에 나올 감정별 데이터 개수가 나온다. 이 값을 현재 각 감정에 있는 데이터의 개수로 나누면 감정별 weight값을 계산할 수 있다.

2.2.3 Majority voting 알고리즘

하나의 spectrogram은 여러 개로 나누어져 학습 및 추론이 이루어진다. 나누어진 각 이미지는 가장 높은 확률을 보인 감정으로 분류된다. 나누어진 이미지가 분류된 감정을 기반으로 하나의 파일에 대한 추론을 진행하게 된다. 하나의 음성 파일로부터 나누어진 이미지 개수의 30% 이상이 같은 감정으로 분류되었다면 음성 파일은 해당 감정으로 분류되게 된다. 어떠한 감정도 30%가 넘지 않을 때에는 중립의 감정으로 분류하였다. 또한, 중립 감

정의 경우에는 나누어진 파일 개수의 50% 이상이 중립으로 분류되면서 다른 감정으로 30% 이상이 분류되지 않았을 때만 중립으로 분류한다.

3. 실 험

본 실험에서 CNN 모델에 사용한 파라미터 값들은 <표 1>과 같다. 배치 사이즈(Batch size)는 한 학습 단계에서 몇 개의 입력 데이터 간격으로 back-propagation을 진행하는지를 결정하는 값이다. 에폭(Epoch) 값은 모든 학습 데이터가 다 사용되는 데 필요한 반복 횟수를 나타낸다. 학습률(Learning rate)은 학습을 진행 시 학습에 사용되는 파라미터 값을 얼마나 빠르게 혹은 느리게 갱신할 것인지에 대한 값이다. 입력 데이터의 크기는 나뉜 파일의 크기에 RGB 세 개의 채널을 곱한 값이다. 마지막으로 입력 데이터는 감정별로 60개씩 총 300개를 사용하였다.

<표 1> CNN 모델에서 사용한 parameter 값

Parameter	Value
Batch size	128
Epoch	30
Learning rate	0.01
Input size	[16, 256, 3]
Number of input	300

실험을 진행한 환경은 아래 <표 2>와 같다.

<표 2> 실험 환경

OS	Ubuntu 14.04-64bit
CPU	Intel Xeon CPU E5-2620
GPU	Nvidia Geforce GTX 1080 Ti
Framework	Tensorflow[7]
	1.2

실험을 위해 Berlin EMO DB를 사용하였다. 전체 DB 중 각 5가지 감정(화남, 기쁨, 슬픔, 분노, 중립)별로 60개의 데이터를 추려내었고 각 음성 파일은 spectrogram 형태로 추출한 뒤 16x256형태로 나누어 입력 값으로 사용하였다. 전체 음성 파일 중 80%는 학습 데이터로 나머지 20%는 테스트 데이터로 사용하였다.

실험 결과 5가지 감정을 분류했을 때 최대 64%의 정확도를 보였다. 화남과 기쁨 감정은 95% 이상 분류되었고 중립 감정도 80% 이상 분류되어 기존 연구보다 높은 정확도를 보였다. 하지만 분노와 기쁨 감정은 화남 감정으로 많이 분류되어 50% 이하의 정확도를 보였다. 실험 결과는 아래 <그림 3>과 같다. A는 화남, H는 기쁨, N은 중립, S는 슬픔 그리고 F는 분노를 각각 나타낸다.

	A	H	N	S	F
A	14	0	0	0	0
H	12	0	0	0	1
N	0	2	10	0	0
S	0	0	1	10	0
F	5	0	1	0	5

<그림 3> 5가지 감정을 분류한 결과를 나타내는 confusion matrix

4. 결 론

본 논문은 CNN을 기반으로 spectrogram을 입력 값으로 사용하여 사용자의 음성을 인식하여 감정을 분류하는 음성 감정인식 기계학습 알고리즘을 제안하였다. 총 3개의 convolutional layer와 1개의 max pooling layer 그리고 3개의 fully connected layer로 이루어진 CNN 모델을 사용하였고, 데이터의 불균형을 해결하기 위해 학습 시 weight 값을 적용하였다. 학습 후 감정을 예측하기 위해서 majority voting 방법을 사용하여 감정을 분류하였다. 실험 결과 총 5가지 감정(화남, 기쁨, 슬픔, 분노, 중립을 분류하여 최대 64%의 정확도를 보였다. 본 논문은 향후 화남의 감정으로 치우치는 기쁨과 분노의 감정을 분류하기 위해 추가적인 특징값을 사용해 RNN(Recurrent Neural Network)로 감정을 분류할 예정이다. 두 감정의 분류 결과를 보고 추가로 CNN과 RNN을 결합하여 전체 감정을 분류하는 모델을 설계할 계획이다.

사 사

이 논문은 2017년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No. IITP-2017-0-00189, 방송통신산업기술개발)

[참 고 문 헌]

[1] Abhinav D., O. V. Ramana M., Roland G., Jyoti J., Tom G., “Video and Image based Emotion Recognition Challenges in the Wild: EmotiW 2015“, International Conference on Multimodal Interaction, pp. 423-426, 2015

[2] H. Go, K. Kwak, D. Lee, M. Chun, “Emotion recognition from the facial image and speech signal,” in SICE 2003 Annual Conference, vol. 3, pp. 2890-2895, 2003

[3] George T., Fabien R., Raymond B., Erik M., Mihalis A. N., Bjorn S., Stefanos Z., “Adieu feautres? End-to-end speech emotion recognition using a deep convolutional recurrent network”, 41st IEEE International Conference on Acoustics, Speech and Signal Processing“, pp. 5200-5204, 2016

[4] Kun H., Dong Y., Ivan T., “Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine”, INTERSPEECH, pp. 223-227, 2014

[5] Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W. F., Weiss B., “A Database of German Emotional Speech”, INTERSPEECH, pp. 1517-1520, 2005

[6] Krizhevsky A., Sutskever I., Hinton G. E., “ImageNet Classification with Deep Convolutional Neutral Networks”, Advances in Neural Information Processing Systems, vol. 25, pp. 1090 - 1098, 2012

[7] Martın A., Ashish A., Paul B., et al., “TensorFlow: Large-scale machine learning on heterogeneous systems”, 2015