소리 데이터 분류에 대한 데이터 증대 방법 연구

*장일식 **박구만

*서울과학기술대학교 나노IT디자인융합대학원 정보통신미디어공학전공

**서울과학기술대학교 전자미디어IT공학과

*foreverme@naver.com

A study on data augmentation methods for sound data classification

*Chang, Il-Sik **Park, Goo-man

*Dept. of Information Technology and Media Engineering

**Dept. of Electronics and IT Media Engineering

Seoul National University of Science and Technology

요약

소리 데이터 분류는 단순 소리를 통한 분류, 감정 인식등 다양한 연구가 진행중이다. 심층 신경망에서 데이터의 부족과 과적합 문제를 개선하는 방법으로 데이터 증강은 중요하다. 본 논문에서는 3가지의 소리데이터(UrbanSound8K, RAVDESS, IRMAS)를 사용하였으며, 소리데이터는 멜 스펙트로그램을 통한 변환과정을 거쳐 네트워크 망에 입력된다. 입력된 신호는 다양한 네크워크 신경망(Bidirection LSTM, Bidirection LSTM Attention, Multi-Head Attention, CNN)을 통해 학습되어지며, 각각의 네트워크 신경망에서 데이터 증강 전후의 분류 정확도를 확인 하였다. 다양한 데이터셋과 다양한 네트워크 망에서의 데이터 증강 방법의 결과 비교를 통한 통찰을 얻을수 있을 것이다.

1. 서론

심층 신경망 학습에서 데이터는 학습하는데 중요한 역할을 한다. 훈련 데이터가 충분하지 않으면 과적합등의 문제가 발생하여 좋은 예측이 어렵다. 이러한 문제는 보안하기 위해 모델의 복잡도를 줄이거나 가중치 규제, Dropout, 데이터 증강등의 다양한 방법들을 사용한다. 본 논문에서는 데이터 증강의 방법을 사용하여 3가지의 데이터셋(UrbanSound8K[1], RAVDESS[2], IRMAS[3])에 적용한다. 또한 총 4가지의 네트워크망(Bidirection LSTM, Bidirection LSTM Attention, Multi-Head Attention, CNN)을 사용하여 소리데이터를 학습니다. 소리데이터 분류를 통한 방법으로 다양한 CNN + RNN, TCN, CNN, Transformer 등 다양한 방법들이 존재한다. 하지만 본 논문에서는 다양한 소리 데이터 셋에서 데이터 증강 방법에 대한 네트워크 망에서의 효과를 확인 하기 위함으로 비교적 단순한 네트워크 망으로 구성하였다. 소리 데이터는 시계열 데이터로서 기본적인 접근으로 데이터 증강은 시간 도메인, 주파수 도메인, 시간 주파수 도메인으로 나눌수 있다.[4] 소리데이터를 스펙토그램으로 변환 후 스펙토그램에서 데이터 증강을 하는 방법[5][6]으로 시간축 마스킹, 주파수 마스킹, 미니 배치에서 다른 데이터 셋과 섞는 방법, 일부 구간 대체하는 방법등이 있다. 본 논문에는 다양한 소리 데이터 셋을 이용하여 서로다른 망에서 동일한 데이터 증강 방법을 적용하였을 경우 성능의 비교를 진행하였다. 논문의 구성은 소리 데이터 셋, 데이터 증강 방법, 신경망 구조 및 학습방법, 실험 결과, 결론 및 향후 연구 방향의 순으로 기술한다.

2. 소리 데이터 셋

본 논문에서 사용되는 3가지의 데이터 셋을 사용한다. 첫 번째 UrbanSound8K는 도시 영상에서 8732개의 사운드 데이터를 총 10개의 클래스를 가지는 데이터 셋이다. 두 번째 IRMAS는 악기 소리 데이터셋으로 총 6705개의 3초간격 오디오 파일로 구성되어 있으며 총 11개의 클래스를 가진다. 마지막 RAVDESS은 감성 음성 오디오데이터셋으로 1440개의 파일오 구성되며, 8개의 클래스를 가진다. UrbanSound8K는 3초 이상의 데이터에서 3초만 사용하였다. 표1은 소리데이터 셋 구성을 나타낸다.

표 1. 소리데이터 셋 구성

	UrbanSound8K		RAVDESS		IRMAS	
	Train	Test	Train	Test	Train	Test
1	897	100	88	7	350	38
2	201	17	128	11	458	47
3	881	99	163	12	413	38
4	627	70	158	10	575	62
5	745	85	120	6	692	68
6	883	90	167	12	620	62
7	32	2	105	10	656	65
8	760	78	173	13	566	60
9	824	77			521	56
10	900	100			530	50
11					701	77

각 데이터 셋의 클래스 이름은 UrbanSound8K : air_conditioner, car_horn, children_playing, dog_bark, drilling, engine_idling, gun_shot, jackhammer, siren, street_music. RAVDESS : neutral, calm, happy, sad, angry, fearful, disgust, surprised. IRMAS : cello, clarinet, flute, acoustic guitar, electric guitar, organ, piano, saxophone , trumpet, violin, human singing voice 순으로 정의 된다. 클래스간 데이터가 불균형 하기 때문에 클래스 불균형 문제를 해결하는 방법으로 오버 샘플링 기법을 사용하였다.

3. 데이터 증강 방법

데이터 증강 방식으로 소리 데이터를 증강하는 방식과 멜스펙트로그램으로 변환 후 특징 벡터를 이용하여 증강하는 방식으로 나눌수 있다. 본 논문에서는 두가지 방법을 모두 사용하여 데이터 증강을 하였다. 소리데이터에 대한 증강 방법으로는 소리 데이터 일정 구간을 자르는 방법, 소리 데이터의 위치를 지정하여 해당 위치에서 랜덤하게 크기를 정하고 각 점에 대하여 슬플라인 보간을 한 후 소리데이터와 곱하는 크기에 대한 보간, 소리 정보에 평균이 0이고 표준편차가 1인 가우시안 정규분포의 값에 일정크기의 값을 더해서 입력데이터에 더하는 즉 노이즈를 추가하는 방법등을 사용한다. 스펙트로그램으로 변경된 특징 벡터에 대한 데이터 증강 방식으로는 SpecAugment[5] 방법에서 사용된 시간 마스킹, 주파수 마스킹 방법을 사용하였으며, 마스킹 영역은 평균값으로 적용하였다. 또한 SpecAugment++[6] 방법에서 사용된 혼합 마스킹, 커팅 마스킹 방법을 사용하였다.

4. 신경망 구조 및 학습 방법

시계열 처리에 대한 다양한 방법의 딥러닝 네트워크가 존재한다. 본 논문에서는 소리데이터를 직접 사용하지 않고 멜스펙토그램으로 변경 후 변경된 특징 데이터를 네트워크 망의 입력으로 한다. 본 논문에서 사용한 네트워크 망은 4가지로 Bidirection LSTM, Bidirection LSTM Attention, Multi-Head Attention, CNN 이다. 앞의 3가지 방법은 시계열 데이터를 처리하는 네트워크 망이고 마지막 CNN은 2차원의 특징 데이터를 가지고 학습을 하는 구조이다. Bidirection LSTM, Bidirection LSTM Attention는 입력크기는 멜스펙토그램의 주파수 크기로 사용한다. 히든벡터의 크기는 64로 설정하였다. 2개의 양방향 LSTM 레이어를 중첩하여 사용하고, 시계열 데이터에 효과적인 레이어 정규화를 거쳐 선형결합을 거쳐 최종 클래스로 분류하는 구조이다. Multi-Head Attention은 트랜스포머 구조의 인코더만 사용하는 구조로 피쳐 입력의 크기는 멜스펙토그램의 주파수 크기로 설정하고, 멀티 헤드의 수는 4개로 설정하였다. 피드포워드 네트워크의 차원은 512로 설정하였다. 마지막 CNN 네트워크 망은 커널 크기가 3,, 스트라이드 1, 출력 체널수가 32인 콘볼루션, 배치 정규화, 활성화 함수로 Relu, 커널 크기 2, 스트라이드 2인 최대 풀링의 구조를 2개를 사용하였다. 2번째 구조의 출력 체널수는 64로 설정하였다. 콘볼루션 출력은 입력크기에 상관없이 출력을 고정하도록 글로벌 평균 풀링을 거친 후 선형결합을 거쳐 최종 클래스로 분류한다.

5. 실험 결과

데이터증강에 따른 실험을 위해 두가지 방법을 사용하였다. 첫 번째 방법은 소리데이터에 대한 증강을 하기 위해선 소리데이터를 멜스펙토그램으로 변경 후 변경된 특징벡터를 네트워크 망의 입력으로 사용하였다. 하지만 이렇게 할 경우 실제 많은 학습시간이 소요된다. 본 논문의 경우 3가지 데이터 셋과 4가지 네트워크 망이 있기 때문에 총 12번의 학습이 필요하다. 그래서 두 번째 방법으로 소리 데이터를 멜스펙토그램으로 변경한 후 특징 벡터을 미리 학습 데이터로 저장하였다. 소리데이터 증강 및 소리데이터 증강과 특징 벡터 증강을 통한 결과의 경우 첫 번째 방법을 사용하였고, 두 번째 방법은 데이터 증강을 사용하지 않은 모델을 통해 멜스펙토그램으로 변환 할 경우 주파수의 크기를 테스트하여 적절한 값을 추출한 경우와 특징 벡터만 증강하는 방식에서 사용하였다. 네크워크 망의 입력은 초기 z-점수 정규화를 한다. 표 2는 데이터 증강을 적용하지 않은 상태에서 멜스펙토그램의 주파수 크기에 대한 실험 결과를 나타낸다. 다양한 네크워크 망중 Multi-Head Attention의 결과를 보여준다. 데이터 증강 실험에는 주파수의 크기를 64로 선정하였다. WA는 Weight Accuracy를 나타내고, UA는 Unweight Accuracy, TA는 Test Accuracy를 나타낸다. 표 3은 소리데이터 증강에 대한 실험 결과를 나타낸다.

표 2 멜스펙토그램의 주파수 크기에 대한 실험 결과

Dataset	Network	Freq Num	WA	UA	TA
Urban Sound8K	mt_att	256	0.72	0.68	0.92
		128	0.69	0.64	0.78
		64	0.7	0.64	0.79
		32	0.72	0.68	0.83
		16	0.65	0.63	0.71
		8	0.52	0.46	0.59
		4	0.32	0.23	0.41
RAVDESS	mt_att	256	0.58	0.57	0.68
		128	0.57	0.56	0.62
		64	0.56	0.56	0.59
		32	0.48	0.46	0.46
		16	0.37	0.35	0.35
		8	0.29	0.25	0.32
		4	0.23	0.2	0.19
IRMAS	mt_att	256	0.58	0.59	0.83
		128	0.54	0.55	0.8
		64	0.52	0.52	0.62
		32	0.51	0.52	0.57
		16	0.45	0.45	0.48
		8	0.4	0.41	0.42
		4	0.32	0.32	0.33

표 3 소리 데이터 증강에 대한 실험 결과

Dataset	Network	WA	UA	TA
Urban Sound8K	brnn	0.65	0.59	0.82
	brnn+att	0.68	0.62	0.8
	mt_att	0.69	0.64	0.84
	cnn	0.68	0.65	0.67
RAVDESS	brnn	0.34	0.31	0.32
	brnn+att	0.39	0.37	0.42
	mt_att	0.55	0.56	0.59
	cnn	0.39	0.38	0.35
IRMAS	brnn	0.42	0.43	0.5
	brnn+att	0.44	0.45	0.5
	mt_att	0.48	0.48	0.54
	cnn	0.34	0.34	0.37

표 4 특징 벡터 증강 적용 전후의 실험 결과

Dataset	Network	Aug	WA	UA	TA
Urban Sound8K	brnn	aug1	0.62	0.58	0.74
		aug2	0.65	0.59	0.72
		aug3	0.64	0.58	0.7
	brnn+att	aug1	0.67	0.61	0.73
		aug2	0.66	0.6	0.73
		aug3	0.68	0.62	0.67
	mt_att	aug1	0.69	0.64	0.78
		aug2	0.7	0.64	0.75
		aug3	0.69	0.63	0.75
	cnn	aug1	0.69	0.65	0.68
		aug2	0.62	0.57	0.6
		aug3	0.63	0.56	0.59
RAVDESS	brnn	aug1	0.43	0.43	0.38
		aug2	0.42	0.41	0.35
		aug3	0.39	0.37	0.34
	brnn+att	aug1	0.36	0.33	0.38
		aug2	0.36	0.33	0.36
		aug3	0.4	0.38	0.4
	mt_att	aug1	0.56	0.56	0.54
		aug2	0.43	0.42	0.3
		aug3	0.43	0.42	0.42
	cnn	aug1	0.41	0.38	0.38
		aug2	0.36	0.33	0.31
		aug3	0.38	0.36	0.32
IRMAS	brnn	aug1	0.5	0.5	0.69
		aug2	0.47	0.47	0.49
		aug3	0.48	0.48	0.51
	brnn+att	aug1	0.52	0.53	0.67
		aug2	0.49	0.5	0.51
		aug3	0.47	0.48	0.49
	mt_att	aug1	0.52	0.52	0.58
		aug2	0.44	0.44	0.42
		aug3	0.43	0.43	0.41
	cnn	aug1	0.37	0.38	0.42
		aug2	0.33	0.33	0.36
		aug3	0.33	0.34	0.35

표 5 소리 데이터, 특징 벡터 증강 적용 전후의 실험 결과

Dataset	Network	Freq Num	Aug	WA	UA	TA
Urban Sound8K	brnn	64	no aug	0.63	0.57	0.88
			aug	0.64	0.57	0.87
	brnn+att	64	no aug	0.68	0.64	0.74
			aug	0.63	0.57	0.79
	mt_att	64	no aug	0.7	0.64	0.79
			aug	0.7	0.64	0.86
	cnn	64	no aug	0.7	0.67	0.68
			aug	0.67	0.63	0.64
RAVDESS	brnn	64	no aug	0.41	0.4	0.4
			aug	0.41	0.4	0.38
	brnn+att	64	no aug	0.38	0.35	0.35
			aug	0.41	0.37	0.38
	mt_att	64	no aug	0.56	0.56	0.59
			aug	0.54	0.54	0.52
	cnn	64	no aug	0.38	0.38	0.4
			aug	0.38	0.37	0.39
IRMAS	brnn	64	no aug	0.51	0.51	0.64
			aug	0.43	0.44	0.5
	brnn+att	64	no aug	0.51	0.52	0.66
			aug	0.44	0.44	0.5
	mt_att	64	no aug	0.51	0.52	0.57
			aug	0.48	0.47	0.53
	cnn	64	no aug	0.35	0.37	0.4
			aug	0.33	0.34	0.37

멜스펙토그램의 주파수 크기는 64, 소리 데이터 일정 구간을 자르는는 방법의 확률은 0.1, 구간의 최대 길이는 10%, 소리데이터 크기 증강 구간은 4개, 크기는 최대 0.2로 설정하였다. 노이즈 증강 확률은 0.2, 크기값은 0.05로 설정하였다. 표4는 특징 벡터 증강에 대한 실험 결과를 나타낸다. aug1은 시간 마스킹, 주파수 마스킹 방법만 사용, aug2는 혼합 마스킹, 커팅 마스킹 방법사용, aug3은 두가지 모두 사용하였을때의 방법이다. 마스킹 방법의 설정은 마스킹 개수는 2개, 마스킹의 길이는 최대 전체 길이의 5%로 설정 하였다. 표 5는 각 소리 데이터별 네트워크 망에 대한 시간 데이터 증강과 특징 벡터 증강을 함께 했을 경우에 대한 실험 결과를 나타낸다. 특징벡터 증강은 aug1의 방법을 적용하여 실험하였다.

5. 결론 및 향후 연구 방향

본 논문에서는 다양한 딥러닝 방법을 통해 3가지의 소리 데이터 셋을 사용하여 소리 데이터에 대한 증강 및 멜스펙토그램으로 변경된 특징 벡터의 증강 방법을 통한 실험 결과를 통해 데이터의 수와 특징 그리고 네트워크 망에 대한 통찰력을 가질수 있었다. 데이터 증강을 통한 성능을 향상하기 위해선 데이터 셋의 종류와 네트워크 망의 구조에 맞추어 다양한 실험이 필요해 보인다. 향후 연구 방향으로는 다양한 특징 벡터의 사용 및 서로 다른 네트워크의 혼합하여 소리 데이터 분류에 대한 연구를 지속하고, 전이 학습, 자기지도학습을 통한 부족한 학습 데이터에 대한 연구와 클래스간 불균형 문제를 해결하기 위한 연구를 진행할 예정이다.

참고문헌

[1] J. Salamon, C. Jacoby and J. P. Bello, "A Dataset and Taxonomy for Urban Sound Research", 22nd ACM International Conference on Multimedia, Orlando USA, Nov. 2014.

[2] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.

[3] Bosch, J. J., Janer, J., Fuhrmann, F., & Herrera, P. “A Comparison of Sound Segregation Techniques for Predominant Instrument Recognition in Musical Audio Signals”, in Proc. ISMIR (pp. 559-564), 2012

[4]Qingsong Wen, Liang Sun, Fan Yang, Xiaomin Song, Jingkun Gao, Xue Wang, Huan Xu. “Time Series Data Augmentation for Deep Learning: A Survey”, Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21), 2021

[5] Park, Daniel S., et al. "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition." Proc. Interspeech 2019 (2019): 2613-2617.

[6] Helin Wang and Yuexian Zou and Wenwu Wang. “SpecAugment++: A Hidden Space Data Augmentation Method for Acoustic Scene Classification”, arXiv:2103.16858v3 [eess.AS] 15 Jun 2021