The Korean Journal of Applied Statistics
2023, Vol. 36, No. 2, 101–114

DOI: https://doi.org/10.5351/KJAS.2023.36.2.101
Print ISSN1225-066X / Online ISSN 2383-5818

Comparative study of data augmentation methods for fake audio detection

KwanYeol Park[a], Il-Youp Kwak[1,a]

[a]Department of Applied Statistics, Chung-Ang University

Abstract

The data augmentation technique is effectively used to solve the problem of overfitting the model by allowing the training dataset to be viewed from various perspectives. In addition to image augmentation techniques such as rotation, cropping, horizontal flip, and vertical flip, occlusion-based data augmentation methods such as Cutmix and Cutout have been proposed. For models based on speech data, it is possible to use an occlusion-based data-based augmentation technique after converting a 1D speech signal into a 2D spectrogram. In particular, SpecAugment is an occlusion-based augmentation technique for speech spectrograms. In this study, we intend to compare and study data augmentation techniques that can be used in the problem of false-voice detection. Using data from the ASVspoof2017 and ASVspoof2019 competitions held to detect fake audio, a dataset applied with Cutout, Cutmix, and SpecAugment, an occlusion-based data augmentation method, was trained through an LCNN model. All three augmentation techniques, Cutout, Cutmix, and SpecAugment, generally improved the performance of the model. In ASVspoof2017, Cutmix, in ASVspoof2019 LA, Mixup, and in ASVspoof2019 PA, SpecAugment showed the best performance. In addition, increasing the number of masks for SpecAugment helps to improve performance. In conclusion, it is understood that the appropriate augmentation technique differs depending on the situation and data.

Keywords: data augmentation, occlusion, deep learning

1. 서론

딥러닝은 컴퓨터 비전 분야에서 이미지 분류, 객체 탐지, 스타일 전송 등 다양한 분야에서 성능을 입증해 왔을 뿐만 아니라 (Wei 등, 2020) 음성 인식이나 분류, 합성 등과 같은 음성 데이터를 다루는 task에서도 성능이 탁월함을 입증해 왔다. 하지만, 많은 매개 변수를 가진 심층 신경망은 대량의 훈련 데이터에 크게 의존한다는 단점이 있다 (Krizhevsky 등, 2012). 그런데, 많은 경우에서 우리는 훈련에 필요한 충분한 양의 데이터를 확보할 수 없다. 이 경우, 과적합 문제가 발생하거나 일반화가 제대로 되지 않는다. 그래서 일반적으로 사용하는 방법이 데이터 증강(data augmentation)이나 일반화(regularization)이다. 일반화와 데이터 증강 외에도 모델 구조 최적화, 전송 학습, 원샷 및 제로샷 학습 등 딥러닝 모델의 성능을 향상시키는 방법들이 있지만 과적합의 본질은 심층 네트워크 모델과 훈련 데이터 사이의 불일치이기 때문에 데이터의 관점에서 바라보는 데이터 증강은 좀 더 효율적이라고 할 수 있으며 (Wei 등, 2020) 새롭게 생성된 훈련 데이터들이 네트워크에 다양성을 제공함으로써 실제 데이터에서 변동이 있을 경우에도 네트워크를 좀 더 견고하게 만들어 줄 수도 있다 (Madhu와 Kumaraswamy, 2019). 그렇기에 오늘날 다양한 데이터 증강 방법이 사용되고 있다.

This research was supported by the National Research Foundation of Korea (NRF) grant funded by Ministry of Science and ICT (RS-2023-00208284). This paper was prepared by extracting part of KwanYeol Park's master's thesis.

[1] Corresponding author: Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06911, Korea. E-mail:ikwak2@cau.ac.kr

Published 30 April 2023 / journal homepage: http://kjas.or.kr

이번 연구에서는 음성 데이터 증강에 occlusion 기반 음성 데이터 증강 방법과 컴퓨터 비전분야에서 사용되는 증강 방법들이 음성위조 탐지문제에서 얼마나 효과적이지 검증하고 비교하고자 한다. 음성 그 자체에 변화를 주는 고전적인 음성 데이터 증강 방법은 noise, shifting, reverse, shuffling, pitch shifting, slicing, speed 및 tempo change 등이 있다 (Choi와 Kwak, 2021). 이러한 방법들은 쉽고 시간이나 자원을 많이 소모하지 않으면서 괜찮은 성능을 보여주었기에 최근까지 널리 사용되었다. 이러한 고전적이고 적인 방법들과 다르게 컴퓨터 비전 연구를 오디오 영역의 문제에 도입하는 것은 생소할 수 있다. 오디오는 1D 신호를 처리하는 반면, 컴퓨터 비전 분야에서는 보통 하나 이상의 2D 이미지를 다룬다. 그렇지만 오디오 분야에서 소리나 음성을 분석할 때 스펙트로그램과 같은 2차원 시간-주파수 표현을 사용하는 경우가 다수 존재한다 (Sukthankar 등, 2006). 음성 데이터를 적절한 방법을 사용하여 전처리를 하면, 음성 데이터 차원을 1D 파형에서 2D 스펙트로그램으로 변경하여 Mixup 등의 컴퓨터 비전 작업을 위해 제안된 데이터 증강 방법 중 일부를 사용할 수 있고, 이들은 오디오 분야에서 활발히 사용되고 있다 (Nam 등, 2022). 그러나 회전, 뒤집기, 크기 조정등 많은 이미지 데이터 증강 방법이 스펙트로그램에 적용되면 사용한 방법과 관련 없는 변환을 초래할 수 있기 때문에 음성 영역에서 모델을 효과적으로 훈련시키기 위해서는 음향 및 신호 처리 영역에 맞는 데이터 증강 방법을 사용하거나 (Nam 등, 2022), 적절한 이미지 증강 기법을 사용할 필요가 있다.

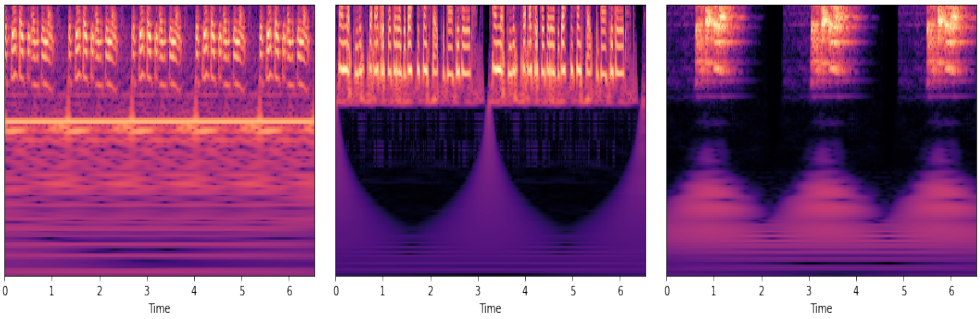
이번 연구에서는 음성위조 탐지문제에 있어서 효율적으로 적용될 수 있는 데이터 증강 기법에 대해 실험 연구를 해보고자 한다. 기본 모형으로는 ASVspoof 2017, 2019 음성위조 탐지 대회에서 각각 1등 2등을 차지한 LCNN 모형을 사용하였고 (Lavrentyeva 등, 2017; Lavrentyeva 등, 2019) 이미지 기반 증강기법으로는 VRM (vicinal risk minimization)기반 데이터 증강 방법인 Mixup (Zhang 등, 2017), Cutout (DeVries와 Taylor, 2017), Cutmix (Yun 등, 2019)등이 occlusion 기반 데이터 증강 방법들이다. 음성 데이터 증강을 위해서도 SpecAugment (Park 등, 2019), Specmix (Kim 등, 2021), FilterAugment (Nam 등, 2022)와 같은 occlusion 기반 증강 방법들이 고안되었으며, 이 중 SpecAugment를 사용해 증강기법들간 성능 비교를 진행해 보았다.

2. 관련 연구

Occlusion은 데이터 증강 분야는 물론 다양한 분야에서 쓰이고, 컴퓨터 비전 분야와 오디오 분야에서도 쓰이는 방식이다. Occlusion 기반 방법은 컴퓨터 비전 분야에서는 매우 다양한 방식으로 쓰였지만, 음성 데이터 분야에서는 비교적 느린 발전 속도를 보여주었다. 하지만, 음성에 occlusion을 적용하였을 때에도 좋은 성능을 보여주는 방식이 소개됨으로써 이 분야에도 적극적인 연구가 이루어지고 있다.

컴퓨터 비전 분야에서 널리 쓰이는 occlusion 방법은 문자 그대로 이미지의 일부를 가리는 방식부터 이미지의 일부 지역 부분만 가지고도 물체나 얼굴 인식등에 사용된다. 그리고, 폐색에 대한 견고성은 이미지 인식 시스템의 중요한 속성이다. 즉, 강력한 이미지 분류기는 이미지에서 관심 객체의 일부만 보이는 경우에도 문제를 해결할 수 있어야 한다 (Fong과 Vedaldi, 2019). CNN 기반 네트워크는 이미지 분류에서 훌륭한 성능을 보여주지만 충분하지 못한 양의 데이터로 네트워크를 훈련시키면 발생할 수 있는 과적합 문제를 해결하기 위한 가장 일반적인 방법으로 데이터 증강 방식이 사용된다. 이에 인위적으로 occlude 시켜 증강시킨 데이터를 사용하여 네트워크의 성능을 높이는 것이 유용한지 연구가 이루어졌고, Hsu 등 (2021)의 연구나 Zhong 등 (2020)등이 제시한 random erasing 방법, Singh 등 (2018)이 제안한 hide-and-seek 방법, Haut 등 (2019)이 oclusion 기반 데이터 증강 방법이 실제로 객체 인식이나 분류등에 낮은 비용으로도 의미있는 성능 증가를 보여주었으며, CNN 기반 모형들과 상성이 좋았음을 보여주었다.

음성 분야에서는 장애물에 막히거나 회절된 음성, 잡음이 많은 음성을 분류하는데 적용되었고, 컴퓨터 비전 분야처럼 일부를 가리거나 음성의 일부만 가지고 분류하는 task를 처리하는데 occlusion 기반 방법이 도움이 됨을 Ke 등 (2005)이 보여주었다. 이에 occlusion 관련 증강 방법 중 앞서 언급했던 SpecAugment가 이번 연구에 가장 적합하고 직관적인 방법이라고 여겼기에 사용하기로 하였다.



(a) ASV2017 (b) ASV2019 LA (c) ASV2019 PA

Figure 1: *Spectrogram with CQT applied.*

3. 연구 방법

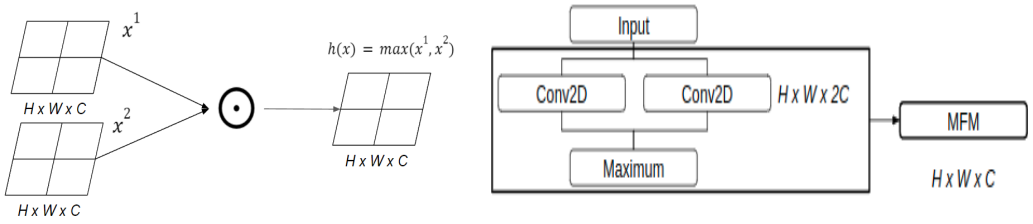
3.1. 데이터셋

ASV (automatic speaker verification) 기술은 스마트폰, 차량, TV, 세탁기, 냉장고, 콜센터 등등 다양한 분야에서 널리 쓰이고 있다. 하지만 spoofing의 위협은 합성 음성, 음성 변조 등의 형태로 보안을 위협하고 있기 때문에 spoofing에 대한 대책이 필요하다. ASVspoof challenge는 spoofing 및 대응 성능의 취약성에 대한 독립적인 평가를 지원하도록 설계되었다. 이 대회는 사전 지식의 부적절한 사용을 가능한 한 방지하면서 다양하고 예측 불가능한 spoofing들을 탐지할 수 있는 대응책을 개발하는 것을 목표로 한다 (Wu 등, 2015). 2015년 첫 challenge에서는 특정 세션 범위 내에서의 spoofing 탐지에 중점을 두었으며, 2017년 challenge에서는 리플레이 공격을 탐지하는 것에 중점을 두었다. 음성 분야에서 리플레이 공격이란, ASV 시스템의 마이크에 재생되거나 표현되는 액세스 시도에 대한 기록으로 구성된다 (Delgado등, 2017). 2019년 challenge는 이전 2015, 2017 challenge를 발전 및 확장시키는 방식으로 이루어졌다. Logical access 시나리오(LA)에서는 text to speech(TTS, 텍스트 음성 변환) 및 voice conversion(VC, 음성 변환) 기술을 사용하여 인간 음성 신호와 기계 생성 음성 신호를 분류하는것을 목표로 하였으며, physical access 시나리오(PA)에서는 인간 음성 신호와 캡처 및 재생된 음성 신호(2017년도와 같은 리플레이 공격 탐지)의 분류를 목표로 하였다.

3.2. 데이터 전처리

음성 데이터를 모형에 넣어 훈련시키기 전에, 먼저 네트워크가 받아들일 수 있는 형태로 데이터를 전처리해야할 필요성이 있다. 음성 피쳐 추출 방법에는 여러가지 방법이 있는데, 본 실험에서 사용한 특징 추출 방법은 constant Q transform (CQT) 방법이다.

Brown (1991)이 제시한 CQT 방법은 같은 음성 피쳐 추출 방법인 STFT에 비해 음성 신호를 처리하는데 더 유리하다고 알려져 있으며 spoofing 공격을 감지하는 데 널리 사용된다 (Lavrentyeva 등, 2017). 또한 ASVspoof 2017 challenge의 주최측에서 제안한 베이스라인 시스템이 CQT 기반이었던것도 CQT를 사용하는데 영향을 주었다. 그리고, challenge에 참여한 Witkowski 등 (2017), Nagarsheth 등 (2017), Lavrentyeva 등 (2017)은 CQT를 기반으로 전처리하여 높은 순위에 들었는데, 이것도 실험에 영향을 주었다. Constant-Q cepstral coefficients (CQCC) 피쳐는 음성 위조 탐지 분야에서 가장 효과적인 피쳐들 중 하나이다. 이 피쳐의 추출에는 신호에서 장거리 정보를 캡처하는 CQT이 포함된다. 또, CQT피쳐에서 옥타브 파워 스펙트럼을 균일하게 리샘플링하여 CQCC 피쳐를 얻을 수 있는 선형 파워 스펙트럼을 갖는다 (Yang 등, 2018). 다만 앞서 언급했던 Lavrentyeva 등 (2017)이 2019년 challenge에서 발표했던 논문에 따르면, ASVspoof 2019의 두 시나리오에서 CQCC로 피



(a) MFM 2/1 (b) MFM block

Figure 2: *MFM function and MFM block definition.*

쳐 추출을 할 경우, 오히려 성능이 떨어지는 현상이 발생했다고 하였고 (Lavrentyeva 등, 2019), Cheng (2019) 등의 연구에 따르면 CQT 베이스의 모형이 CQCC 모형의 성능보다 리플레이 공격을 탐지하는데 더 효과적인 성능을 보였다고 하였기에, CQT로 전처리를 하였다.

Figure 1은 각각 ASVspoof2017, ASVspoof2019 PA, ASVspoof2019 LA의 샘플 데이터에 CQT를 적용해 전처리를 한 예시이다.

3.3. 모형

Convolutional neural networks (CNN) 기반 아키텍처를 가진 모형들은 특히 대규모 데이터 세트를 가지고 있을 때 이미지 분류에 매우 효과적이라고 알려져 있으며 spoofing 탐지 분야에서는 주로 얼굴 spoofing 탐지에서 사용되었다. 그런데 Zhang 등 (2017)이 음성 spoofing 탐지에 유용하다는 것을 증명해 냈으며, Hershey 등 (2017), Dua 등 (2021), Abdel 등 (2014)도 CNN 기반 모델을 음성 분류에 사용했을 때, 좋은 성능을 낸다는 것을 보여주었다.

Wu 등 (2018)이 제안한 light CNN (LCNN)은 대규모 노이즈 레이블이 있는 데이터에서 심층 얼굴 표현 (representation)을 학습하기 위해 제안되었다. 원래 CNN 모형은 이미지 분류를 위해 설계된 모형이기 때문에 오디오 분류를 위해서는 구조의 수정이 불가피한데, Lavrentyeva 등 (2017)등이 제시한 개량된 LCNN 모형은 2017년과 2019년 challenge에서 매우 높은 성능을 보여줌으로서 LCNN 모형이 음성 데이터를 학습시키에도 적합한 모형임을 입증하였다. LCNN 모형은 압축된 표현을 얻기 위해 ReLU의 대안으로써 Goodfellow 등 (2013)이 제안한 max-out 활성화 함수의 확장인 MFM (max-feature map) 활성화 함수를 사용하여 특징 (feature) 필터를 선택하고, MFM을 CNN의 각 컨볼루션 레이어에 도입함으로써 CNN 모형을 단순화시켰다. MFM은 낮은 자원으로도 계산을 할 수 있어 부담을 덜 준다는 장점이 있다. 아래의 식이 MFM 함수의 식이다.

$$y_{ij} = \max\left(x_{ij}^{k}, x_{ij}^{k+N/2}\right), \tag{3.1}$$

$$\forall i = \overline{1, H}, \quad j = \overline{1, W}, \quad k = \overline{1, N/2}. \tag{3.2}$$

x는 크기 $H \times W \times N$의 입력 텐서이고, y는 크기 $H \times W \times N/2$의 출력 텐서이다. 여기서 i, j는 주파수 및 시간 영역을 나타내고 k는 채널 지수이다. 많은 피쳐 맵을 사용하여 임의의 convex 활성화 함수를 선형으로 근사하는 최대값 활성화와 달리 MFM은 경쟁 관계를 통해 이를 수행한다. MFM은 노이즈 신호와 정보 신호를 분리할 수 있을 뿐만 아니라 두 개의 피쳐 사이에서 피쳐를 선택하는 역할도 수행한다.

Figure 2(a)가 MFM의 layer로, MFM 연산의 결과값은 해당 위치의 신경 노드 사이의 최대값이다. CNN 네트워크에서 MFM 은 생체 인식에서 국소 기능 선택과 유사한 역할을 하고, 최대값 연산은 서로 다른 동일한 차원의 필터에서 학습한 각 위치에서 최적의 기능을 선택하여 정보 손실과 계산 부담을 줄인다 (Wu 등, 2020; Shim 등, 2022). CNN에서 MFM을 사용할 때 드러나는 또 하나의 특징은, MFM 계층의 gradiant가 낮은 동안

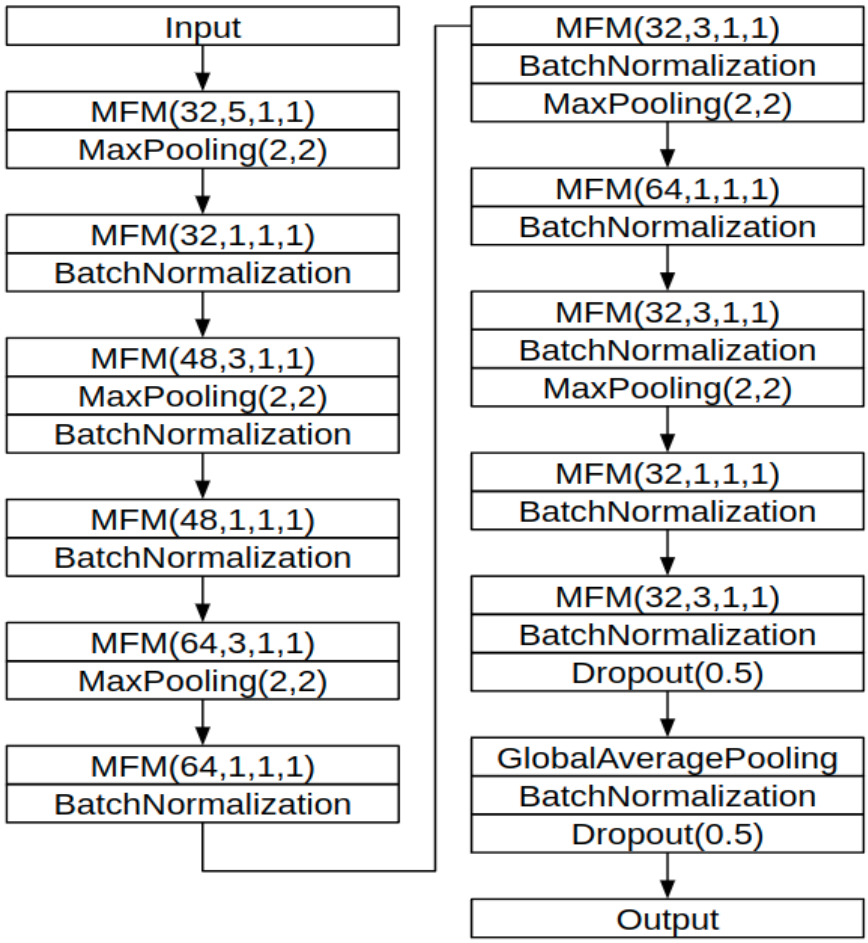


Figure 3: *LCNN model architecture.*

압축된 표현을 얻을 수 있다는 것이다. 그리고, 훈련 단계에서 back propagation을 수행할 때, 확률적 기울기 강하(SGD)는 반응 변수의 뉴런에만 영향을 미친다 (Wu 등, 2020). Figure 2(b)는 본 실험에서 사용한 MFM의 블록의 구조이다.

본 실험에서는 앞서 언급된 개량된 LCNN 모형을 2019년 challenge에 같은 연구자가 다시 개량한 모형에서 좀 더 나은 일반화 성능을 위해 몇가지를 추가하여 사용하였다. Figure 3 가 본 실험에서 사용한 모델의 구조이다. Lavrentyeva등은 2017년에도 LCNN 모형으로 데이터를 학습시켜 실험을 실시하였는데, 2019년 challenge에서는 softmax 함수와 batch normalization을 추가하여 실험을 진행하였다. 본 실험에서는 이를 참조하여 Figure 3에 제시된 모델을 사용하여 학습을 진행하였다. 과적합 방지를 위해 dropout을 0.5로 설정하고 maxpooling으로 차원을 축소하였으며, 때때로 maxpooling 레이어 뒤에 훈련 과정 중 안정성 및 수렴 속도를 높이기 위해 Ioffe와 Szegedy (2015)이 제시한 batch normalization을 사용하였다 (Lavrentyeva 등, 2019). 인간의 음성 신호와 기계 음성 신호를 분류하는 것은 물론 재생된 음성 신호도 분류해야 하였기 때문에 활성화 함수로 softmax를 사용하였고, 옵티마이저는 ADAM을 사용하였다.

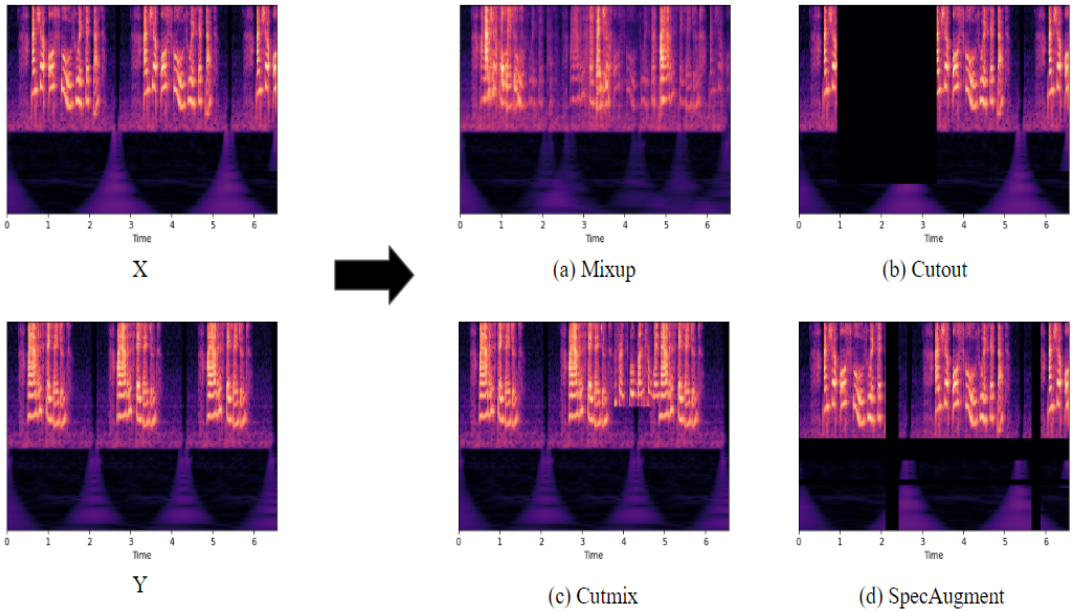


Figure 4: *Data augmentation strategy.*

3.4. 증강 전략

심층 인공 신경망은 효과적으로 학습하기 위해 대량의 훈련 데이터셋을 필요로 하는데, 그러한 훈련 데이터 수집은 종종 비용이 많이 들고 수고스럽다. 데이터 증강은 레이블을 보존하면서 훈련 데이터셋를 인위적으로 부풀려 이 문제를 해결한다. 최근 CNN의 작업 성능을 향상시키기 위해 일반 데이터 증강이 광범위하게 사용되고 있는데 (Krizhevsky 등, 2012), 앞서 소개했던 고전적인 증강 방법은 음향 모델을 훈련할 때 약간의 비효율성을 초래할 수 있으며, 사전 지식이 필요하다. 또한, 이러한 방법은 보다 자연스러운 사운드를 만들어 내기 위해 많은 계산이 필요하고, 그럼에도 불구하고 더 잘 훈련시킨다는 보장도 없다. 따라서 우리는 음향 모델을 훈련하는 데 간단하고 직관적이면서도 효과적인 데이터 증강 방법이 필요하다 (Nam 등, 2022). 본 논문에서는 CNN 모델을 훈련시키는데 효과적이라고 알려진 몇 가지 방법을 소개하고, 이를 사용해볼 것이다.

3.4.1. Mixup

Zhang 등 (2017)이 제시한 Mixup 기법은 기존의 empirical risk minimization (ERM)의 모순을 해결하기 위한 대안인 vicinal risk minimization (VRM)의 한 종류이다 (Chapelle 등, 2000). Mixup은 linear interpolration을 사용하여 두 랜덤 샘플 (x_i, x_j), (y_i, y_j)을 베타 분포에서 추출된 $\lambda \in [0, 1]$를 혼합 비율로 하여 재조합하는 방법이다. Mixup을 통해 모델 복잡성을 제어하거나 훈련 데이터를 증가시켜 일반화 격차를 줄임으로써 억제할 수 있다 (Wei 등, 2020). Mixup은 아래의 식으로 정의될 수 있다.

$$\tilde{x} = \lambda x_i + (1 - \lambda)\, x_j, \tag{3.3}$$

$$\tilde{y} = \lambda y_i + (1 - \lambda)\, y_j. \tag{3.4}$$

여기서 $\tilde{x}$ 와 $\tilde{y}$ 는 mixup에 의해 생성된 데이터이다. Mixup으로 생성된 $\tilde{x}$의 예시는 Figure 4(a)에서 확인할

수 있다. Figure 4 의 λx_i 와 $(1-\lambda)x_j$ 가 혼합되어 Figure 4(a)의 $\tilde{x}$ 가 생성된다.

Mixup이 주로 쓰이는 곳은 컴퓨터 비전 분야이나, 원 논문의 저자는 물론 다양한 연구자들이 audio data에 Mixup을 적용하였다. 결과적으로 audio data에 적용해 본 다른 연구들에서 상당히 뛰어난 성능을 보여주어 많이 활용되고 있다. 또, ERM 방식을 사용하면 각 클래스 중간에 위치하는 데이터에 대해 한쪽 클래스라고 과도하게 확신하는 경향이 있는데, 라벨이 없거나 부정확한 데이터의 분류에도 robust하다고 설명된다. 이번 실험에서 사용하는 데이터도 이에 해당하기 때문에 사용하는데 적합하다고 할 수 있기에 실험에 사용해 보기로 하였다.

3.4.2. Cutout

DeVries와 Taylor (2017)가 제시한 무작위 영역을 지우기 위한 regional dropout 방법 중 하나인 Cutout 기법은 이미지의 일부분을 잘라내어 0으로 치환하는 전략으로, occlusion 계열 전략이다. 다만, 일반적으로 occlusion은 대상 문자를 덮는 스크래치, 점 또는 낙서의 형태를 취하는 반면, cutout은 전체 영역을 완전히 방해하기 위해 제로 마스크를 사용한다. 또, cutout은 객체 폐색 문제를 해결하는 방법중 하나로, occlude된 샘플를 고려하는 새로운 이미지를 생성함으로써 실제 상황에서 occlusion이 발생하는 경우에 대해 모델을 더 잘 대비시켜주며, 결정을 내릴 때 이미지 컨텍스트를 더 많이 고려하게 된다. 결론적으로 cutout을 사용하면, CNN 기반 네트워크의 견고성을 늘려주고, 전반적인 성능을 향상시킬 수 있다. 또, 구현하기 매우 쉽고, 다른 데이터 증강 전략이나 정규화와 함께 사용하면 모델 성능을 더욱 향상시킬 수 있다. 이에 우리는 네트워크의 성능을 향상시키는데 도움이 된다고 판단하였기 때문에 사용하기로 하였다.

λ는 베타 분포에서 추출되며 잘라내는 부분 영역과 전체 영역의 비율을 조절하는 파라미터이다. 잘라낼 박스의 가로, 세로 길이인 r_w와 r_h를 계산하는 데 사용된다. Bounding box의 중심 위치인 r_x와 r_y는 균일 분포에서 추출된다(W와 H는 각각 가로와 세로차원의 크기를 나타낸다).

$$r_x \sim \mathrm{Unif}(0, W), \quad r_w = W\sqrt{1-\lambda}, \tag{3.5}$$

$$r_y \sim \mathrm{Unif}(0, H), \quad r_h = H\sqrt{1-\lambda}, \tag{3.6}$$

bounding box는 r_x와 r_y를 중심으로 가로와 세로의 길이가 r_w와 r_h 가 되도록 하되, 그림 영역을 벗어나는 부분은 제외한다. M은 bounding box 부분이 0, 나머지는 1으로 정의된 masking matrix 라고 할 때, Cutout 기법은 아래와 같이 수행된다.

$$\tilde{x} = M \odot x_i, \tag{3.7}$$

여기서 x_i 와 M은 같은 차원의 행렬이며, $\odot$ 연산은 element-wise 곱을 나타낸다. 즉, 주어진 x_i에 해당 연산을 취한 $\tilde{x}$는 Figure 4(b)와 같은 모습을 나타내게 된다.

3.4.3. Cutmix

Yun 등 (2019)이 제시한 Cutmix 기법은 Mixup 기법과 Cutout 기법을 조합한 방식이라고 이해하면 쉽다. Mixup은 분류 성능은 확실히 향상되지만 혼합 샘플은 다소 부자연스럽게 보이는 경향이 있고, cutout은 잘리는 영역만큼 정보 손실이 일어난다. 또, 두 전략 모두 상대적으로 덜 중요한 영역에 집중하는 경향을 보이는 것이 단점으로 꼽힌다. Cutmix는 이미지 영역을 다른 train 이미지의 패치로 교체하여 문제를 이 문제들을 다소 해결한다. Cutmix는 두 개의 서로 다른 train image의 일부를 마스킹하고, 그 파트를 다른 train image patch로 채워 넣고 라벨을 섞어서 새로운 이미지를 만드는 방식인데, 추가된 patch의 부분적인 모습만 보고 물체를 추정할 수 있기 때문에 localization 성능이 향상 되는 결과도 보인다. Cutout과 같은 방식으로 bounding box를

Table 1: Results for Mixup hyperparameter tuning

	λ	ASVspoof2017	ASVspoof2019 LA	ASVspoof2019 PA
Baseline		0.396(0.024)	0.032(0.001)	0.006(0.001)
	0.1	0.352(0.031)	0.027(0.002)**	0.006(<0.000)
	0.3	0.340(0.036)	0.027(0.003)	0.005(0.001)*
	0.5	0.340(0.017)	0.025(0.004)	0.005(0.001)**
	0.7	0.334(0.013)**	0.023(0.001)**	0.005(0.001)*

Table 2: Results for Cutout hyperparameter tuning

	λ	ASVspoof2017	ASVspoof2019 LA	ASVspoof2019 PA
Baseline		0.396(0.024)	0.032(0.001)	0.006(0.001)
	0.1	0.343(0.028)**	0.024(0.003)	0.005(<0.000)*
	0.3	0.342(0.030)	0.028(0.003)	0.005(<0.000)
	0.5	0.415(0.028)	0.024(0.003)**	0.005(<0.000)**
	0.7	0.335(0.028)	0.027(0.003)*	0.005(<0.000)*

결정하되, 샘플과 라벨은 아래와 같이 계산된다.

$$\tilde{x} = M \odot x_i + (1 - M) \odot x_j \,, \tag{3.8}$$

$$\tilde{y} = p y_i + (1 - p)\, y_j \,. \tag{3.9}$$

생성되는 데이터 $\tilde{x}$ 는 Figure 4(c)와 같이 x_i에서 잘려진 bounding box 부분이 0 이 아닌 x_j 부분의 정보가 채워지게 된다. 따라서 생성되는 라벨 $\tilde{y}$ 역시 bounding box의 면적에 비례해 정의되어야 하며, $p = 1 - (r_w r_h)/WH$ 의 식을 계산하고, target label을 조정하는데 사용한다.

3.4.4. SpecAugment

SpecAugment는 Park 등 (2019)이 제시한 기법으로, log mel spectrogram을 입력으로 받아 time wrapping, frequency masking, time masking을 사용하여 증강시킨다. SpecAugment는 파형에 데이터 증강을 적용하는 대신 log mel spectrogram에 직접 적용할 수 있는 time wrapping, frequency masking, time masking을 제안하였다. 그리고 이 전략을 사용할 경우 입력된 특성 공간에 직접 적용되어 이해 및 사용이 용이하다는 장점도 있다. Audio masking은 일반적으로 희미하지만 가청 소리가 더 큰 다른 가청 소리, 즉 마스커(masker)가 있을 때 들리지 않게 되는 효과로 정의된다. 하지만, Park 등 (2019)의 논문에서는 컴퓨터 비전 분야의 occlusion과 거의 동일한 의미로 사용되었다.

Time warping은 tensorflow의 sparse image warp fucntion으로 사용된다. 먼저 τ time step이 있는 log mel spectrogram이 주어지면 이를 시간 축이 수평이고 주파수 축이 수직인 이미지로 본다. 그리고, spectrogram의 정중앙을 통과하는 수평선에 사용자가 지정한 $(w, \theta - W)$ 사이의 임의의 점을 찍고, 해당 점을 기준으로 왼쪽, 혹은 오른쪽으로 균일한 분포에서 선택된 거리 w에 의해 오디오 정보를 이동시키는 것이 바로 time warping 이다 (Park 등, 2019).

Frequency masking은 frequency mask parameter F를 정하고 $[0, F]$의 범위 안에서 임의의 값 f를 추출한 후 $[f_0, f_0 + f)$ 만큼 마스킹한다. f_0는 Unif$(0, W - f)$의 분포로부터 임의로 추출된다. 여기서 W 는 시간축 차원의 크기이다. W는 Cutout 설명부분에서 인풋 x_i의 가로축 크기 W 와 의미적으로 동일하다.

Time masking은 time mask parameter T를 정하고 $[0, T]$의 범위 안에서 임의의 값 t를 추출하고 $[t_0, t_0 + t)$ 만큼 마스킹한다. t_0는 Unif$(0, H - t)$의 범위 안에서 임의로 추출된다. H는 Cutout 설명부분에서 인풋 x_i의

Table 3: Results for Cutmix hyperparameter tuning

	λ	ASVspoof2017	ASVspoof2019 LA	ASVspoof2019 PA
Baseline		0.396(0.024)	0.032(0.001)	0.006(0.001)
	0.1	0.230(0.008)**	0.038(0.003)	0.012(0.001)**
	0.3	0.241(0.021)**	0.042(0.004)*	0.012(0.001)**
	0.5	0.240(0.014)**	0.043(0.002)**	0.012(0.001)**
	0.7	0.233(0.013)**	0.045(0.005)**	0.011(0.001)**

Table 4: Results for SpecAugment hyperparameter tuning

	Mask	Frequency, time	ASVspoof2017	ASVspoof2019 LA	ASVspoof2019 PA
Baseline			0.396(0.024)	0.032(0.001)	0.006(0.001)
	2	[10,15]	0.312(0.024)*	0.032(0.002)	0.005(<0.000)*
	2	[15,70]	0.356(0.009)	0.032(0.004)	0.005(<0.000)*
	2	[27,70]	0.312(0.017)**	0.030(0.001)	0.004(<0.000)**
	2	[27,100]	0.332(0.007)**	0.032(0.002)	0.004(<0.000)**
	3	[10,15]	0.305(0.016)*	0.035(0.002)	0.004(<0.000)**
	3	[15,70]	0.347(0.013)	0.034(0.002)	0.004(<0.000)**
	3	[27,70]	0.309(0.035)	0.034(0.003)	0.004(<0.000)**
	3	[27,100]	0.307(0.011)**	0.033(0.002)	0.004(<0.000)**

Table 5: Results of best hyperparameter settings for each data augmentation

	ASVspoof2017	ASVspoof2019 LA	ASVspoof2019 PA
Baseline	0.396	0.032	0.006
Mixup	0.334	0.023	0.005
Cutout	0.335	0.024	0.005
Cutmix	0.230	0.038	0.011
Specaug	0.305	0.030	0.004

세로축 크기 H 와 의미적으로 동일하다.

4. 실험

우리는 앞서 설명했던 세 개의 데이터셋을 각각 데이터 증강 방법을 사용하여 증강시킨 후, Figure 3의 LCNN 모형에 넣고 훈련을 시켰다. 그리고, 성능 비교를 위해 ASVspoof challenge에서 평가 지표로 사용하였던 EER을 사용하였다. EER (equal error rate)은 생체인식 기술에서 성능을 나타내는 지표로, 오인식률 (false acceptance rate; FAR)과 오거부율 (false rejection rate; FRR)이 같아지는 포인트로 정의되며, EER이 낮을수록 정확하다고 간주된다. 일반적으로 spoofing CM의 신뢰성은 EER 메트릭을 사용하여 측정된다 (Kinnunen 등, 2020).

먼저, 각 데이터셋에서 baseline 모형의 성능은 5번 실행해보았을 때, 평균(표준 편차)가 ASVspoof2017은 0.396(0.024), ASVspoof2017 LA는 0.032(0.001), PA는 0.006(0.001)로 측정되었고, baseline 결과와 augmentation 결과 간의 차이가 있는지 t-test에 Bonferroni 교정을 통해 검정해보았다. 통계적으로 유의미한 차이가 있는 결과는 표에 *(p-value <0.1 인 경우) 또는 ** (p-value < 0.05인 경우) 표기하였다.

Mixup, Cutout, Cutmix의 hyperparameter (λ)는 각각 두 spectrogram 이미지가 섞이는 비율, spectrogram 이

미지를 잘라내는 비율, 잘라내고 다음 spectrogram 이미지의 일부를 붙이는 비율로 정의된다. 각 증강 전략에 따라 최적의 hyperparameter (λ)가 다르다는 가정 하에, hyperparameter (λ)를 0.1부터 0.7까지 0.2씩 증가시켜 가며 실험을 진행하였다. SpecAugment의 경우, Park 등 (2019)의 논문에서 진행했던 hyperparameter를 5번씩 실험해보고, 실험의 추이에 따라 연구자가 생각하는 더 적합한 파라미터를 설정하고 실험해 보았다.

4.1. Mixup hyperparameter tuning에 따른 성능 비교

Table 1은 Mixup을 적용할 때 다른 λ값에 따라 성능을 비교해 본 표이다. 모든 경우에서 Mixup을 적용하였을 때 성능이 향상되었고, λ가 0.7 인 경우 세개의 데이터에서 모두 baseline에 비해 유의미한 성능 차이를 보였다. 이미지 데이터에 대해 실험한 Zhang 등 (2017)에서는 0.2 ~ 0.4 사이에서 가장 좋은 성능을 보였지만 음성 데이터를 훈련시킨 모형 쪽에서는 λ가 0.7인 것이 더 좋은 성능을 보였다. 이는 Tomilov 등 (2021)의 모형에서 0.6 ~ 0.8 사이가 좋은 결과를 보였다는 이전의 가설과도 일치한다.

4.2. Cutout hyperparameter tuning에 따른 성능 비교

Table 2는 Cutout에 λ을 다르게 하여 적용해본 결과이다. 각 데이터셋에 Cutout을 사용해 본 결과, ASVspoof2017 λ가 0.1일 때 가장 좋은 성능을 보이고, ASVspoof2019의 두 데이터셋의 경우, λ가 0.5일 때 가장 유의미한 성능 증가를 볼 수 있다. 다만, ASVspoof2017 데이터셋에 적용했을 때, λ가 증가할수록 성능이 나빠지는 것을 관측할 수 있다.

4.3. Cutmix hyperparameter tuning에 따른 성능 비교

Cutmix의 경우, ASVspoof2017에 적용하였을 때 Table 3를 보면 다른 증강 전략에 비해 리플레이 공격을 탐지하는데 평균적으로 좋은 성능을 보여주지만, ASVspoof 2019 LA, PA 데이터셋에 적용할 경우 전략을 적용하지 않은 경우보다도 성능이 낮아지는 것을 관측할 수 있었다. 그리고 이 결과들이 모두 t-test를 적용한 결과 p-value가 유의하게 나옴으로써, 증강 전략을 도입한 결과 통계적으로 차이가 있다고 할 수 있다. 결론적으로, ASVspoof2017 데이터셋에는 Cutmix 전략을 적용하는 것이 가장 효과적인 증강 전략이나, ASVspoof2019의 두 시나리오 데이터셋에 도입하기에는 좋은 전략이 아닌 것으로 보인다.

4.4. SpecAugment hyperparameter tuning에 따른 성능 비교

Table 4는 SpecAugment의 mask의 개수와 frequecy를 여러가지를 적용해본 결과이다. 원 논문에서의 실험은 mask를 2개로 고정시키고, time과 frequecy mask의 크기를 약간씩 변형시켜서 적용시켜 보았다. 그런데, mask의 크기를 변동시키는 것뿐만 아니라 mask의 개수 그 자체를 변형시켜보는 것도 성능의 영향을 줄 수 있다고 생각했기에 mask의 개수를 새로운 hyperparameter로 정하고 실험을 해 보았다. 결과적으로, mask의 크기를 늘렸을 때, ASVspoof2019 LA 데이터셋을 제외한 ASVspoof2017과 PA 데이터셋은 평균적인 성능이 증가하는 것을 볼 수 있었다. ASV2019 LA 데이터셋에 대해서는 SpecAugment를 적용하면 증강 전략 도입 후 성능에 변화가 있다는 가설을 기각할 수 없다는 결과가 나왔다.

4.5. 실험 데이터별 가장 좋은 성능을 보이는 증강 기법

Table 5는 각 데이터셋과 증강 기법 별로 가장 성능이 좋았던 결과들을 기술하였다. ASVspoof2017, ASVspoof2019 LA, PA 데이터 별로 각각 Cutmix, Mixup, SpecAugment 가 우수한 성능을 보였다. 음성 위조 탐지 문제에 있어서는 데이터 상황별로 잘 동작하는 증강 기법이 다른 것으로 보이고, 대체로 증강 기법의 사용이 성능 향상에 도움을 주는 것으로 나타났다.

ASVspoof2017 데이터셋을 사용하여 리플레이 공격을 탐지하고 분류해내는 task에서 가장 좋은 성능을 보인 증강 전략은 cutmix로 ratio를 0.1로 하였을 때 다른 증강 기법을 사용했을때보다 두드러지게 높은 성능을 보여 주었고, SpecAugument 가 뒤를 이었다. Cutout과 Mixup은 비슷한 정도의 성능 향상을 보였다.

ASVspoof2019 LA 데이터셋을 증강시키는 전략 중, 사람의 음성 신호와 기계 음성 신호를 분류하는 성능이 가장 좋았던 전략은 Mixup으로, ratio를 0.7로 하였을 때 최대 0.00627만큼의 EER이 감소하는 효과를 볼 수 있었다. Cutout도 거의 비슷한 수준의 성능 향상을 보였고, SpecAugment는 성능 변화가 거의 없는 것을, Cutmix에서는 성능이 나빠지는 경우를 관측할 수 있었다.

ASVspoof2019 PA 데이터셋을 증강시키는 전략 중, 사람의 음성 신호와 캡처 및 재생된 음성 신호를 분류하는 성능이 가장 좋았던 전략은 SpecAugment로, mask를 3개로, frequency와 time mask의 두께를 27,100으로 했을 때가 좋았다. Cutmix를 제외한 다른 증강 전략을 도입해 보았을 때에도 EER이 감소하여 성능이 증가함을 보였으나, SepcAugment가 유독 성능 증가폭이 높았다.

5. 결론

본 연구에서는 음성 위조 탐지 문제에 있어서 적용 가능한 다양한 occlusion 기반 데이터 증강 기법들을 실험해 보았다. 세 가지 다른 음성 위조 탐지 데이터에 대해 Mixup, Cutout, Cutmix, SpecAugment 기법을 비교해 보았고, 그 결과 ASVspoof2017에서는 cutmix가, ASVspoof2019 LA 데이터에 있어서는 Mixup과 Cutout이, ASVspoof2019 PA 데이터에 있어서는 SpecAugment, Mixup, Cutout이 우수한 성능을 보였다. 음성 위조 탐지 문제에 있어서 데이터 상황별로 잘 동작하는 증강 기법이 다른 것을 확인할 수 있었다. 또, 음성 분야 딥러닝 문제에 있어서 잘 사용되어 오고 있는 Mixup, SpecAugment 외에도 Cutout 방법도 고려해 볼 만한 방법이라는 것을 확인하였다.

결과적으로, 음성 데이터를 위한 SpecAugment 증강 방법도 높은 성능을 보여주지만, 데이터셋에 따라 컴퓨터 비전 관련 데이터 증강 방법들인 Mixup, Cutout 등도 모델의 성능을 향상시키는 데에 도움이 됨을 알 수 있었다. Occlusion 기반 증강 기법이 기존에 성능이 검증된 컴퓨터 비전 분야뿐만 아니라 음성 분야에도 효과적인 방법이라는 것도 알 수 있었다. 음성 위조 탐지 모형에서의 추천하는 hyperparameter는 Mixup의 경우 0.7, Cutout 의 경우 0.7, Cutmix의 경우 0.5–0.7의 사용을 추천하며, SpecAugment 의 경우는 masking 의 갯수 3, 시간 주파수 방향 masking size를 27, 100 정도로 잡을 때, 전반적인 데이터들에서 좋은 결과를 나타내는 것을 확인하였다.

References

Abdel-Hamid O, Mohamed AR, Jiang H, Deng L, Penn G, and Yu D (2014). Convolutional neural networks for speech recognition, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **22**, 1533–1545.

Brown JC (1991). Calculation of a constant Q spectral transform, *The Journal of the Acoustical Society of America*, **89**, 425–434.

Chapelle O, Weston J, Bottou L, and Vapnik V (2000). Vicinal risk minimization, *Advances in Neural Information Processing Systems*, **13**, Cambridge MA, USA.

Cheng X, Xu M, and Zheng TF (2019). Replay detection using CQT-based modified group delay feature and ResNeWt network in ASVspoof 2019. In *Proceedings of 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Lanzhou, China, 540–545.

Choi HJ and Kwak IY (2021). Data augmentation in voice spoofing problem, *The Korean Journal of Applied Statistics*, **34**, 449–460.

Delgado H, Todisco M, Sahidullah M, Evans N, Kinnunen T, Lee KA, and Yamagishi J (2017). ASVspoof 2017 Version 2.0: Meta-data analysis and baseline enhancement, *Odyssey 2018-The Speaker and Language Recognition Workshop*.

DeVries T and Taylor GW (2017). Improved regularization of convolutional neural networks with Cutout, Available from: arXiv preprint arXiv

Dua M, Jain C, and Kumar S (2021). LSTM and CNN based ensemble approach for spoof detection task in automatic speaker verification systems, *Journal of Ambient Intelligence and Humanized Computing*, **13**, 1985–2000.

Fong R and Vedaldi A (2019). Occlusions for effective data augmentation in image classification. In *Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, Seoul, Korea, 4158–4166.

Goodfellow I, Warde-Farley D, Mirza M, *et al.* (2013). Maxout networks, In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, Atlanta, Georgia, USA, 1319–1327.

Haut JM, Paoletti ME, Plaza J, Plaza A, and Li J (2019). Hyperspectral image classification using random occlusion data augmentation, *IEEE Geoscience and Remote Sensing Letters*, **16**, 1751–1755.

Hsu CY, Lin LE, and Lin CH (2021). Age and gender recognition with random occluded data augmentation on facial images, *Multimedia Tools and Applications*, **80**, 11631–11653.

Ioffe S and Szegedy C (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift, *International Conference on Machine Learning*, **37**, 448–456.

Yang J, Das RK, and Li H (2018). Extended constant-Q cepstral coefficients for detection of spoofing attacks. In *Proceedings of 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Honolulu, HI, USA, 1024–1029.

Ke Y, Hoiem D, and Sukthankar R (2005). Computer vision for music identification. In *Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, 597–604.

Kim G, Han DK, and Ko H (2021). Specmix: A mixed sample data augmentation method for training with time-frequency domain features, Available from: arXiv preprint arXiv:2108.03020

Kinnunen T, Delgado H, Evans N, *et al.* (2020). Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **28**, 2195–2210.

Krizhevsky A, Sutskever I, and Hinton GE (2012). Imagenet classification with deep convolutional neural networks, *Communications of the ACM*, **60**, 84–90.

Lavrentyeva G, Novoselov S, Malykh E, Kozlov A, Kudashev O, and Shchemelinin V (2017). Audio replay attack detection with deep learning frameworks, In *Interspeech 2017* (pp. 82–86).

Lavrentyeva, G, Novoselov S, Tseren A, Volkova M, Gorlanov A, and Kozlov A (2019). STC antispoofing systems for the ASVspoof2019 challenge, *Interspeech 2019*, 1033–1037.

Madhu A and Kumaraswamy S (2019). Data augmentation using generative adversarial network for environmental sound classification. In *Proceedings of 27th IEEE European Signal Processing Conference (EUSIPCO)*, A Coruna, Spain, 1–5.

Nam H, Kim SH, and Park YH (2022). FilterAugment: An acoustic environmental data augmentation method. In *Proceedings of ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore,4308–4312.

Nagarsheth P, Khoury E, Patil K, and Garland M (2017). Replay attack detection using DNN for channel discrimination, *Interspeech 2017*, 97–101.

Park DS, Chan W, Zhang Y, Chiu C-C, Zoph B, Cubuk ED, and Le QV (2019). SpecAugment: A simple data augmentation method for automatic speech recognition, Available from: arXiv preprint arXiv:1904.08779

Shim HJ, Jung JW, Kim JH, and Yu HJ (2022). Attentive max feature map and joint training for acoustic scene classification. In *Proceedings of ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, 1036–1040.

Singh KK, Yu H, Sarmasi A, Pradeep G, Lee YJ (2018). Hide-and-Seek: A data augmentation technique for weakly-supervised localization and beyond, Available from: arXiv preprint arXiv:1811.02545

Sukthankar R, Ke Y, and Hoiem D (2006). Semantic learning for audio applications: A computer vision approach. In *Proceedings of 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, New York, NY, USA, 112–112.

Tomilov A, Svishchev A, Volkova M, Chirkovskiy A, Kondratev A, and Lavrentyeva G (2021). STC Antispoofing Systems for the ASVspoof2021 Challenge. In *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, (pp. 61–67).

Wei S, Zou S, and Liao F (2020). A comparison on data augmentation methods based on deep learning for audio classification, *Journal of Physics: Conference Series*, **1453**, 012085.

Witkowski M, Kacprzak S, Zelasko P, Kowalczyk K, and Gałka J (2017). Audio replay attack detection using high-frequency features, *Interspeech 2017*, 27–31.

Wu X, He R, Sun Z, and Tan T (2018). A light cnn for deep face representation with noisy labels, *IEEE Transactions on Information Forensics and Security*, **13**, 2884–2896.

Wu Z, Kinnunen T, Evans N, Yamagishi J, Hanilci C, Sahidullah Md, and Sizov A (2015). ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge, *Sixteenth Annual Conference of the International Speech Communication Association*, 2037–2041.

Yun S, Han D, Chun S, Oh SJ, Yoo Y, and Choe J (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, Seoul, Korea, 6023–6032.

Zhang C, Yu C, and Hansen JH (2017). An investigation of deep-learning frameworks for speaker verification antispoofing, *IEEE Journal of Selected Topics in Signal Processing*, **11**, 684–694.

Zhang H, Cisse M, Dauphin YN, and Lopez-Paz D (2017). Mixup: Beyond empirical risk minimization, Available from: arXiv preprint arXiv

Zhong Z, Zheng L, Kang G, Li S, and Yang Y (2020). Random erasing data augmentation, In *Proceedings of the AAAI conference on artificial intelligence*, Hilton New York Midtown, NY, USA, 13001–13008.

Received November 7, 2022; Revised December 2, 2022; Accepted December 13, 2022

음성위조 탐지에 있어서 데이터 증강 기법의 성능에 관한 비교 연구

박관열[a], 곽일엽[1,a]

[a]중앙대학교 응용통계학과

요 약

데이터 증강 기법은 학습용 데이터셋을 다양한 관점에서 볼 수 있게 해주어 모형의 과적합 문제를 해결하는데 효과적으로 사용되고 있다. 이미지 데이터 증강기법으로 회전, 잘라내기, 좌우대칭, 상하대칭등의 증강기법 외에도 occlusion 기반 데이터 증강 방법인 Cutmix, Cutout 등이 제안되었다. 음성 데이터에 기반한 모형들에 있어서도, 1D 음성 신호를 2D 스펙트로그램으로 변환한 후, occlusion 기반 데이터 기반 증강기법의 사용이 가능하다. 특히, SpecAugment는 음성 스펙트로그램을 위해 제안된 occlusion 기반 증강 기법이다. 본 연구에서는 위조 음성 탐지 문제에 있어서 사용될 수 있는 데이터 증강기법에 대해 비교 연구해보고자 한다. Fake audio를 탐지하기 위해 개최된 ASVspoof2017과 ASVspoof2019 데이터를 사용하여 음성을 2D 스펙트로그램으로 변경시켜 occlusion 기반 데이터 증강 방식인 Cutout, Cutmix, SpecAugment를 적용한 데이터셋을 훈련 데이터로 하여 CNN 모형을 경량화시킨 LCNN 모형을 훈련시켰다. Cutout, Cutmix, SpecAugment 세 증강 기법 모두 대체적으로 모형의 성능을 향상시켰으나 방법에 따라 오히려 성능을 저하시키거나 성능에 변화가 없을 수도 있었다. ASVspoof2017 에서는 Cutmix, ASVspoof2019 LA 에서는 Mixup, ASVspoof2019 PA 에서는 SpecAugment 가 가장 좋은 성능을 보였다. 또, SpecAugment는 mask의 개수를 늘리는 것이 성능 향상에 도움이 된다. 결론적으로, 상황과 데이터에 따라 적합한 augmentation 기법이 다른 것으로 파악된다.

주요용어: 데이터 증강 기법, Occlusion, 딥러닝

이 성과는 2023년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. RS-2023-00208284).

이 논문은 박관열의 석사논문의 일부를 발췌하여 작성하였음.

[1]교신저자: (06911) 서울시 동작구 흑석로 84, 중앙대학교 경영경제대학 응용통계학과. E-mail: ikwak2@cau.ac.kr