비디오 내 음원 위치 추정 모델의 성능 향상을 위한 클래스 인지 대조 학습 기법 제안

(Class-Aware Contrastive Learning for Improving Performance of Sound Source Localization Model in Videos)

선 주 형‡ (Juhyeong Seon)　김 재 윤‡ (Jaeyoon Kim)　김 주 영§ (Kim Joo Young)

이 영 주§ (Lee Young Joo)　한 혜 경§ (Han Hyekyung)　윤 성 의¶ (Sung-Eui Yoon)

요 약

비디오 상의 음원 위치를 추정하는 신경망 모델 학습은 이미지 및 오디오 멀티 모달 연구의 중요 분야 중 하나이다. 최근 연구들은 대조 학습법 (contrastive learning)을 사용하여 음원 위치 추정 모델을 지도하는 방법을 제안하는데, 이 방법은 서로 다른 비디오는 서로 다른 클래스의 객체를 표현한다고 가정한다. 그러나, 일반적인 학습 데이터셋에는 동일한 객체를 담은 비디오가 존재하기 마련이다. 기존의 학습 과정에는 이러한 비디오들이 학습 배치 내에 함께 존재하여 모델에게 잘못된 지도를 제공할 수 있다. 이러한 문제를 바로잡고자, 본 논문에서는 음원 위치 추정 모델이 비디오 상의 객체 클래스를 미리 예측하여 데이터를 재배치하는 정확한 대조 학습법을 제안한다. 제안하는 방법은 추가적인 레이블 없이도 기존의 음원 위치 추정 모델의 성능을 개선하였다. 음원 위치 추정 연구 분야의 성능 검증 실험을 통해 이를 뒷받침한다.

키워드: 심층 학습, 멀티 모달 학습, 음원 위치 추정, 대조 학습

Abstract

Training neural network models to localize the source of sound in a video is one of the important areas in image and audio multi-modal research. Recent research proposes a method for training the sound source localization model using contrastive learning, which assumes that different videos always represent different objects. However, there will likely be the same object videos in the common training dataset. These videos mislead the model in the previous training stage, existing together in the same batch. To address this issue, this paper proposes a refined contrastive learning approach that accurately regroups the data by predicting object classes shown in the videos. Commendably, this simple approach enhanced the performance of the existing sound source localization model without explicit additional labels. This performance was supported by validation experiments in sound source localization research.

Keywords: Deep Learning, Multi-modal Learning, Sound Source Localization, Contrastive Learning

---

† 이 연구는 LIG NEX1 산학협력과제 지원으로 연구되었음. This research is performed based on the cooperation with KAIST-LIG Nex1 Cooperation.

‡ 비 회 원 : 한국과학기술원 전산학부

{munuwazzi, sqbrq}@kaist.ac.kr

§ 비 회 원 : LIG넥스원 연구원

{jooyoung.kim, youngjoo.lee, hyekyung.han}@lignex1.com

¶ 종신회원 : 한국과학기술원 전산학부 교수

sungeui@kaist.edu

1. 서 론

이미지와 오디오를 융합적으로 이해하는 멀티 모달 학습법은 다방면으로 연구되어왔으며, 수많은 사용자를 보유하는 소셜 네트워크 서비스 (social network service)와 주문형 비디오 (video on demand) 플랫폼에 대한 관심이 증가함에 따라 본 분야의 연구가 더욱 주목받고 있다. 특히, 이미지 상에서 소리의 발원지를 찾는 음원 위치 추정 (sound source localization) 문제는 비디오를 이해하는 데에 중요한 과제로 대두되어왔다. 이러한 연구는 감시 시스템의 탐지 대상 검출 정확도 향상 기술, 사용자 상호작용 기반 비디오 서비스의 원천 기술 등으로 활용될 수 있다.

최근 음원 위치 추정 연구는 레이블이 없는 비디오 데이터를 활용하여 신경망 모델을 학습하는데 주로 데이터 임베딩 분야에서 우수한 성능을 보이는 대조 학습법 (contrastive learning)이 활용된다. 대조 학습은 동일한 객체 **(positive)**가 표현된 이미지와 오디오의 인코딩 유사도를 높이고 상이한 객체 **(negative)**의 유사도는 낮추는 학습을 목표로 하며 선행 연구들은 이를 바탕으로 더욱 발전된 대조 학습법을 제안한다[1-4]. 이러한 방법들은 비디오 레이블을 활용하지 않기에 서로 다른 비디오는 서로 다른 객체 클래스를 표현한다는 가정에 기초한다[5]. 그러나 실제 데이터셋에는 동일한 객체 클래스를 표현하는 상이한 비디오가 존재하며 이러한 경우 모델은 잘못된 지도를 받는데 이를 거짓 음성 (false negative) 문제라 한다 (그림 1). 거짓 음성 문제는 대조 학습을 불안정하게 만들고 정확한 인코딩 생성을 방해하므로[6, 7] 신경망 모델의 음원 위치 추정 성능을 저해한다.

이러한 문제를 해결하고자, 본 논문은 비디오 상의 객체 클래스를 예측하여 활용하는 학습법을 제안한다. 제안하는 모델은 같은 객체 클래스를 표현하는 비디오를 음성**(negative)**이 아닌 양성**(positive)**으로 올바르게 배치하여 혼란을 야기하는 지도를 정정하는 동시에 같은 클래스에 속하는 다변적인 데이터를 학습할 수 있도록 한다. 이 방법을 통해 추가적인 레이블 없이도 대조 학습 기반의 기존 위치 추정 모델 성능을 개선할 수 있다. 본 연구 분야의 벤치마크 성능 측정 실험을 통해 논문이 제안하는 방법이 음원 위치 추정 정확도 및 검출 성공률을 개선함을 보인다.

2. 연구 배경 지식

2.1 관련 연구

비디오 상에서 음원의 위치를 추정하는 신경망 모델 학습법은 다양한 측면으로 연구되어왔다. 음원의 위치를 지도하기 위한 픽셀 마스크는 고비용 구축 작업이 요구되므로 본 분야에서는 이러한 형태의 레이블 없이 음원의 위치를 추정하는 학습법이 활발히 연구되고 있다. 논문[2]는 이미지

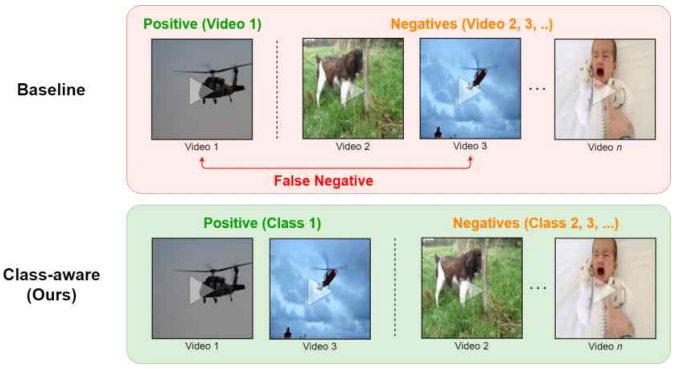

그림 1. 음원 위치 추정 모델 별 대조 학습을 위한 양성 (positive) 및 음성 (negative) 데이터 묶음 방법 비교도
Fig. 1 Comparison of positive and negative data grouping methods for contrastive learning of sound source localization models

상에서 음원과 배경을 분리하는 학습법을 제안하여 대조 학습 (contrastive learning) 기반 음원 위치 추정 연구 분야의 초석을 마련하였다. 논문[1]은 이미지와 소리 임베딩의 유사도 계산과 음원 위치 추정 태스크를 구분하여 학습하는 모델을 제안하였으며 높은 추정 정확도 (CIoU, AUC)를 달성하였다. 이들은 모두 음원 위치 추출 문제에서 대조 학습이 레이블이 없이도 뛰어난 성능을 발휘함을 증명해왔다. 하지만 위 방법들은 비디오 단위로 객체를 구별하기 때문에 다른 비디오에 동일한 객체 클래스가 존재하는 거짓 음성의 경우를 올바르게 학습에 활용하지 못한다. 논문[3]은 이러한 거짓 음성 문제를 겪지 않기 위해 음성 (negative) 데이터 없이 음원 위치 추정을 학습하는 방법을 개발하였다. 하지만 대조 학습에서 양질의 음성 데이터는 정확한 임베딩을 생성하는 데에 중요한 역할을 하기 때문에[6, 7], 논문[3]은 거짓 음성 문제를 회피하기 위해 음성 데이터 활용을 포기한다는 한계점이 존재한다. 논문[5]는 비디오 상의 객체 구별 문제 (instance discrimination)에서 거짓 음성 문제를 해결하여 비디오 임베딩 성능을 개선하였다. 본 논문에서도 비디오 상의 음원 위치 추정 문제에서 거짓 음성 문제를 해결하여 위치 추정 성능을 개선하는 것을 목표로 한다. 거짓 음성 문제를 개선하는 클래스 인지 학습법을 제안하고 이 학습법이 음원 위치 추정 성능에 미치는 영향을 분석하는 실험을 진행한다.

2.2 기존 음원 위치 추정 학습법

본 논문이 제안하는 클래스 인지 대조 학습법 소개에 앞서, 먼저 기존의 음원 위치 추정 모델 학습법을 소개한다 (그림 2). 음원 위치를 추정하기 위한 비디오는 특정 시점의 이미지 한 프레임과 3초 길이의 오디오로 나뉜다. 그 후 이미지와 오디오는 각각의 인코더를 거쳐 같은 차원을 공유하는 임베딩으로 추출된다. 이때 이미지 인코더는 ImageNet[8]

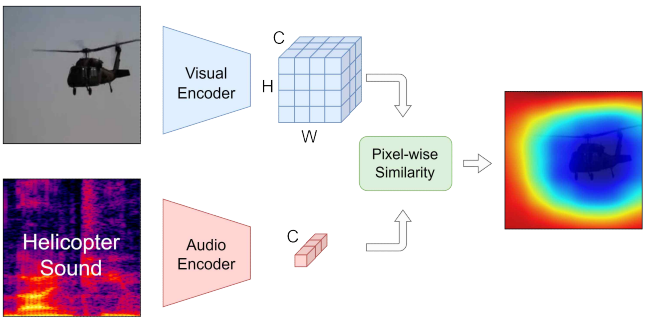

그림 2. 음원 위치 추정 모델의 기본 구조
Fig. 2 The baseline architecture of the sound source localization model

데이터셋으로 사전 학습이 되어있다. 오디오의 인코딩은 오디오 마다 1개가 생성되며 이미지 인코딩은 이미지의 공간 정보를 반영하기 위해 픽셀 수만큼 존재한다. 따라서 두 도메인의 인코딩 간 코사인 유사도를 계산하면 픽셀 수만큼의 유사도 지도가 생성되며 이는 음원 위치 추정 지도로 사용된다. 유사도 지도 상에서 음원 위치에 해당하는 이미지 인코딩이 오디오 인코딩과 높은 유사도를 가지도록 인코더를 학습시킨다. 이때 클래스 레이블이 주어지지 않는 기존 연구 환경에서는 같은 비디오 만이 같은 객체를 표현한다고 가정하기 때문에, 같은 비디오에서 추출된 이미지와 오디오 인코딩은 항상 유사도가 높고 (*Pos*) 다른 비디오 간 인코딩은 유사도가 낮도록 (*Neg*) 학습한다 (1). 이를 바탕으로 하는 대조 학습의 손실 함수 (2)는 다음과 같다:

$$Pos_i = S_{ii}, Neg_i = \frac{1}{N-1}\sum_{i \neq j} S_{ij}, \quad (1)$$

$$L_{contrastive} = -\frac{1}{N}\sum_{i=1}^{N}[\log\frac{\exp(Pos_i)}{\exp(Pos_i)+\exp(Neg_i)}. \quad (2)$$

S_{ij}는 비디오 i의 이미지 인코딩과 비디오 j의 오디오 인코딩 간 코사인 유사도의 픽셀 차원 평균을 나타낸다. N은 학습 배치 내 비디오 수를 의미한다. 이때 비디오 i의 이미지와 비디오 j의 오디오에 같은 객체가 있는 경우도 음성 (*Neg*)으로 잘못 취급하는데, 이는 모델이 객체의 의미를 반영하는 인코딩을 구축하는 데에 혼란을 일으켜 학습을 불안정하게 만든다[6, 7]. 따라서, 다음 장에는 이러한 거짓 음성을 양성으로 올바르게 취급하는 클래스 인지 기반 대조 학습법을 제안한다.

3. 클래스 인지 음원 위치 추정 학습법

논문[6]은 대조 학습 기반의 이미지 임베딩 문제에서 클래스 레이블을 활용하여 거짓 음성을 줄이는 것이 보다 정확히 객체의 의미를 반영하는 이미지 임베딩에 기여함을 보였다. 하지만 이를 음원 위치 추정 문제에 적용하기 위해서는 비디오 단위의 객체 클래스 레이블이 필요하다는 한계점이 있다. 이에 본 논문은 비디오 클래스 레이블을 활용하지 않는 새로운 방법을 제안한다. 먼저 ImageNet[8]으로 사전 학습된 이미지 인코더를 활용하여 훈련 데이터셋의 비디오를 이미지 클래스로 분류한다. 이렇게 분류된 클래스를 활용하여 서로 다른 비디오라도 같은 객체 클래스로 분류된 경우 두 비디오의 유사도를 높일 수 있는 학습법을 제안한다:

$$Pos_i = \frac{1}{N_{class(i)}}\sum_{j \in class(i)} S_{ij},$$
$$Neg_i = \frac{1}{N - N_{classs(i)}}\sum_{j \notin class(i)} S_{ij}. \quad (3)$$

*Class(i)*는 비디오 i에 표현된 객체 클래스를, $N_{Class(i)}$는 배치 내 *class(i)*에 속한 비디오 수를 나타낸다. 제안하는 방법은 기존에 음성 (*Neg*)으로 잘못 분류되었던 데이터를 추출하여 양성 (*Pos*) 학습에 올바르게 제공하므로 모델에 혼란을 야기하던 지도를 정정하는 동시에 클래스 별로 다양한 데이터를 제공한다. 이는 모델의 일반화 성능을 향상하여 테스트 데이터셋에서 보다 정확한 음원 위치 추정을 달성할 수 있다.

다음 장에서는 클래스 인지 기반의 대조 학습법이 기존의 음원 위치 추정 성능에 미치는 영향을 분석하기 위해 음원 위치 추정 벤치마크에서의 성능 검증 프로토콜로 모델 간의 성능을 비교 분석한다.

4. 음원 위치 추정 성능 실험

4.1. 성능 검증 환경: 모델, 데이터셋 및 성능 지표

본 연구의 성능 우수성 비교 검증을 위해 논문[1, 2]와 동일한 평가 프로토콜을 준수한다. 논문[1]의 모델 (이하 SLAVC)를 베이스라인으로 선정하고 본 논문에서 제안한 학습법을 적용한 새로운 모델을 학습한다. 정확한 성능 비교를 위해 베이스라인 모델은 공개된 코드를 바탕으로 본 논문과 동일한 환경에서 학습한 결과를 보고한다. 이미지 및 오디오 인코더는 ResNet-18[9] 구조이고 이미지 인코더는 ImageNet에 사전 학습된 모델로 초기화한다. 모델 학습에는 14만 장의 비디오가 포함된 VGG-SS[10] 데이터셋을 활용하며 추가적인 비디오 레이블을 학습에 활용하지 않는다. 모델이 예측하는 클래스 수는 ImageNet과 동일한 1000개로 설정한다. 모델 테스트에는 벤치마크 데이터셋인 Flickr SoundNet[11] (이하 Flickr) 및 VGG-SS와 논문[1]에서 제안한 각각의 Extended 버전을 활용한다. Extended 버전은 서로 다른 비디오의 이미지와 오디오를 합성한 비디오를 추가한 테스트 셋으로, 모델이 오디오와 상관없이 이미지 상의 객체 위치를 추정하는지 판단하기 위해 고안되었다. 음원 위치 추정의 성능 측정 기준으로는 Consensus Intersection-over-Union (CIoU), Average Precision (AP), Area Under Curve (AUC), F1-Score를 활용한다.

표 1. VGG-SS 및 Flickr-SoundNet 벤치마크 상에서 음원 위치 추정 성능 결과

Table 1 Comparison of sound source localization results on Flickr-SoundNet and VGG-SS benchmark

Method	Flickr-SoundNet				VGG-SS			
	CIoU*↑	AUC*↑	AP↑	F1↑	CIoU*↑	AUC*↑	AP↑	F1↑
SLAVC [1]	0.832	0.6448	0.8629	90.8	0.3889	**0.3985**	0.4264	56
SLAVC + Class Aware (ours)	**0.848**	**0.6454**	**0.8873**	**91.8**	**0.3924**	0.3976	**0.4332**	**56.4**

표 2. 확장된 VGG-SS 및 Flickr-SoundNet 벤치마크[1] 상에서 음원 위치 추정 성능 결과

Table 2 Comparison of sound source localization results on Extended Flickr-SoundNet and Extended VGG-SS benchmark [1]

Method	Extended Flickr-SoundNet				Extended VGG-SS			
	CIoU*↑	AUC*↑	AP↑	F1↑	CIoU*↑	AUC*↑	AP↑	F1↑
SLAVC [1]	0.400	0.6328	0.6721	64.4	**0.1871**	**0.3927**	0.2576	33.6
SLAVC + Class Aware (ours)	**0.418**	**0.6404**	**0.6871**	**67.1**	0.1870	0.3917	**0.2596**	**33.7**

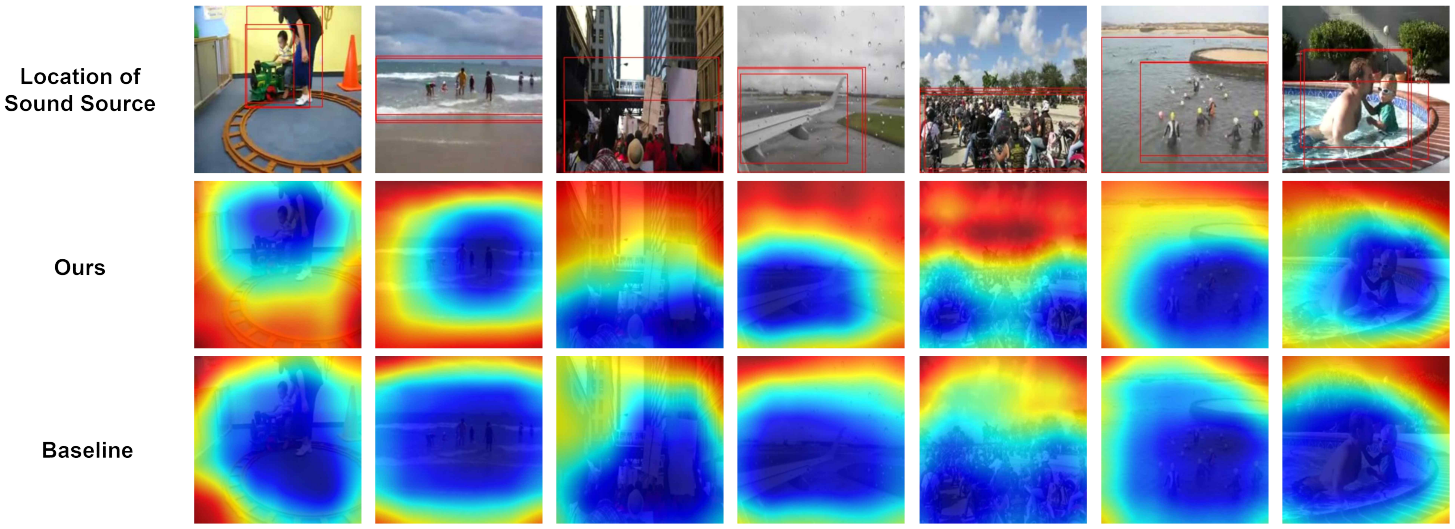

그림 3. 모델 별 Flickr-SoundNet 테스트 데이터의 음원 위치 추정 결과 예시

Fig. 3 Examples of estimated sound source location results on the Flickr-SoundNet test dataset by models

CIoU는 모델이 추정한 음원 위치와 테스트 데이터에 bounding box로 표시되는 실제 음원 위치 영역 간의 교집합을 합집합으로 나누어 계산하므로 위치 추정 정확도와 직결된 중요한 성능 지표이다. 또한, 음원 위치 추정 정확도인 CIoU 값을 바탕으로 AP, F1-Score (precision-recall curve), AUC (ROC curve)를 계산함으로써 모델이 적절히 음원 검출에 성공하였는가를 파악할 수 있다. 해당 연구 분야에서는 CIoU와 AUC 지표가 주로 사용된다[2-4, 11, 12]. 검증치와 관련된 자세한 설명은 논문[1]에서 확인할 수 있다.

4.2. 성능 실험 결과

표 1은 음원 위치 추정 모델 별 성능 측정 결과를 보여준다. 본 논문에서 제안한 모델은 베이스라인 모델[1]의 일부 성능을 개선하였다. CIoU 값을 Flickr 데이터셋에서 +0.016, VGG-SS 데이터셋에서 +0.0045, F1 score 값은 Flickr에서 +1.0, VGG-SS에서 +0.4, AP 값은 Flickr에서 +0.0244, VGG-SS에서 +0.0068 만큼 개선하였다. 이러한 결과는 제안하는 클래스 인지 대조 학습이 음원 위치 추정 성능 개선에 도움이 됨을 보인다. 다만, AUC 값은 Flickr에서 +0.0006, VGG-SS에서 -0.0009 로 기존과 비슷한 성능을 유지하였다. 높은 AUC 값을 나타내기 위해서는 모델이 음원 영역 상의 멀티 모달 임베딩 유사도를 높이고 그 외의 부분은 낮추어야 한다. 즉, 객체와 배경에 대한 학습이 보다 정확히 이루어져야 하는 어려움 때문에 제안하는 모델은 AUC 성능 개선에 한계를 보였다.

표 2는 모델의 논문[1]에서 제안한 Extended Flickr와 Extended VGG-SS 테스트 벤치마크 성능을 보여준다. 이 데이터셋은 이미지 상의 객체와 매칭되지 않는 소리를 인위적으로 합성하였을 때 음원 위치 추정 모델이 이를

적절히 구별해내는 지를 파악하기 위해 고안되었다. 본 논문이 제안한 방법은 Extended Flickr 벤치마크 상의 모든 성능 지표에서 기존 모델 (SLAVC) 보다 1~4% 만큼 개선된 성능을 보이는데, 이는 모델이 비디오의 클래스를 인지하고 학습하므로 소리와 이미지 상의 객체 클래스가 다른 경우를 구별하는 능력이 향상됨을 시사한다. 하지만, 이미지 상의 객체 위치나 크기가 다양하게 존재하는 Extended VGG-SS 데이터셋[1]에서는 주요 지표인 CIoU와 AUC에서 성능 개선을 보이지 못했다.

그림 3은 Flickr 테스트 데이터셋에 모델이 음원 위치를 추정한 결과 예시를 보여준다. 기존 베이스라인 모델 (baseline)은 이미지 상의 음원과 인접한 배경까지 음원이라고 판단하지만 본 논문이 제안한 모델 (ours)은 음원을 보다 정교히 예측하는 점을 확인할 수 있다. 그림 3의 3열 이미지에서 사람이 들고 있는 피켓을 음원으로 예측하는 오류를 개선하거나 4열의 이미지에서 음원으로 표현된 배경 영역을 확연히 줄인 점에서 이를 확인할 수 있다. 이는 기존에 거짓 음성으로 분류되던 동일 객체 비디오가 양성으로 올바르게 제공되어, 학습 중에 객체의 방향이나 형태가 다양한 데이터를 학습하여 모델이 배경과 음원을 잘 구분함을 추론할 수 있다. 하지만, 두 모델은 여전히 음원의 형태를 정확하게 표현하지 못하고 대략적인 위치만을 표시하는 한계를 보인다. 이는 해상도가 낮은 이미지 임베딩을 보간하여 사용하는 현 음원 위치 추정 모델 방식이 원인이다. 이를 해결하기 위해서는 복호 모델 (decoder) 혹은 해상도가 높은 이미지 임베딩을 활용해야 하는데, 이는 연산량 및 모델 복잡도의 상승을 유발하므로 픽셀 마스크 등의 추가적인 레이블 활용이 불가피하다[13].

4.3 클래스 분류 정확도가 음원 위치 추정 성능에 미치는 영향 분석

본 논문은 비디오 프레임 내에 등장하는 객체의 클래스를 예측하여 비디오를 분류하는 전략을 제안하였다. 제안하는 기법은 클래스 분류의 정확성에 의존하므로 이에 대한 실험을 추가로 수행하였다.

표 3은 비디오 클래스 분류 모델을 변경하여 데이터를 분류한 이후 학습한 음원 위치 추정 모델의 성능 변화를 나타낸다. 클래스 분류 모델의 정확성이 일정 수준 이하로 감소할 경우 음원 위치 예측 모델의 성능이 저하됨을 알 수 있다. Top-1 정확도가 상대적으로 낮은 모델 (AlexNet[14], SqueezeNet[15])을 이용하여 비디오를 분류하였을 때에는 비디오 분류를 진행하지 않는 기본 모델 (baseline) 보다도 더 저조한 성능을 나타냈다. 그러나, 본 논문에서 제안한 ResNet-18 모델을 활용하였을 때에는 기본 모델보다 대부분 우수한 성능 수치를 달성하였다. 이는 본 논문이 제안하는 음원 위치 예측 학습 방법에서 비디오의 정확한 분류가 거짓 음성 문제 해결에 있어 중요하게 작용한다는 점을 시사한다.

표 3. 비디오 클래스 분류 모델 별 음원 위치 추정 모델 성능 비교

Table 3 Comparison of sound source localization performance with different video class classification models

Classification model (Top-1 Accuracy on ImageNet)	Flickr-SoundNet		VGG-SS	
	CIoU↑	AUC↑	CIoU↑	AUC↑
baseline (None)	0.832	0.6448	0.389	**0.3985**
AlexNet (56.5)	0.800	0.6124	0.356	0.3830
SqueezeNet (58.0)	0.776	0.6002	0.348	0.3762
ResNet-18 (69.7) (ours)	**0.848**	**0.6454**	**0.392**	0.3976

4.4 한계점 및 향후 연구

본 논문은 클래스 인지 대조 학습법을 제안하여 음원 위치 추정 모델의 벤치마크 성능을 개선하며, 이를 통해 거짓 음성 문제가 음원 위치 추정의 정확도 향상에 중요한 문제임을 시사하였다. 이에 따라 본 논문에서는 거짓 음성 문제를 해결하기 위한 클래스 인지 대조 학습의 한계점을 분석하고 이를 개선하는 향후 연구를 제안한다.

클래스 인지 대조 학습법에는 여전히 모델에게 잘못된 클래스 정보를 전달하는 두 가지의 경우가 존재한다. 먼저 클래스 분류 모델이 객체의 클래스를 잘못 분류하는 경우이며 4.3장에서 이러한 경우가 위치 추정 성능에 미치는 영향을 분석하였다. 이 문제는 사전 학습 모델로부터 임시 클래스 레이블을 생성하여 활용하는 연구에 공통되는 한계점이며, 학계에서는 이러한 레이블 오류에도 강건한 모델을 학습하는 방법이 다양하게 연구되어왔다[16]. 이러한 연구를 본 논문이 제안하는 클래스 인지 기반 대조 학습에 효과적으로 적용하는 것은 새로운 향후 연구 주제이다.

두 번째 경우는 이미지 상에 여러 객체가 존재하여 음원이 아닌 다른 객체를 분류하는 경우이다. 이는 제안한 모델이 VGG-SS 데이터셋에서의 CIoU 성능 지표를 개선하지 못한 점에서 간접적으로 나타난다. VGG-SS 데이터셋은 Flickr 데이터셋보다 객체의 수와 각각의 위치 및 크기 등이 다양하게 존재하기 때문에[1], 비디오 이미지 상의 객체로만 비디오를 분류하였을 때 모델에게 음원 객체가 아닌 다른 객체의 클래스를 전달할 여지가 크다. 이러한 문제를 해결하기 위해서는 오디오와 이미지를 함께 활용하는 보다 정확한 객체 분류 기법을 도입하는 것이 적절하다. 오디오에 사전 훈련된 분류 모델을 추가로 활용해 오디오에서 검출된 객체와 이미지에서 검출된 객체 간 대응을 판단하거나 두 모델의 예측을 융합하는 기법 (late fusion) 등을 활용할 수 있다[17]. 이러한 연구 주제는 음원 위치 추정 모델의 대조 학습을 더욱 정확하게 지도하여 높은 정확도를 이끌어낼 것으로 기대된다.

5. 결론

본 논문에서는 심층 학습 모델이 비디오 상에서 음원의 위치를 추정할 때 비디오의 클래스를 예측하여 활용하는 발전된 대조 학습법 (contrastive learning)을 제안하였다. 각 비디오가 서로 다른 객체를 담고 있다는 기존 연구의 가정에 한계를 극복하고자, 비디오를 객체 클래스 단위로 분류하여 모델에게 보다 정확한 데이터 지도를 제공하였다. 제안하는 방법은 추가적인 레이블 없이도 기존의 대조 학습 기반 모델의 음원 위치 추정 성능을 개선하였으며, 음원 위치 추정 연구 분야의 벤치마크 성능 검증을 통해 이를 뒷받침하였다. 이는 거짓 음성 (false negative) 데이터가 음원 위치 추정의 정확도 향상에 중요한 문제임을 시사하며, 앞으로는 더욱 정확한 객체 클래스 분류 기법과 효과적인 클래스 활용 기법 등의 연구가 필요할 것이다.

참 고 문 헌

[1] S. Mo, and P. Morgado., "A Closer Look at Weakly-Supervised Audio-Visual Source Localization", *Advances in Neural Information Processing Systems*, pp. 37524-37536, 2022.

[2] H. Chen, W. Xie, T. Afouras, A. Nagrani, A. Vedaldi, and A. Zisserman, "Localizing Visual Sounds the Hard Way", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16867-16876, 2021.

[3] Z. Song, Y. Wang, J. Fan, T. Tan, and Z. Zhang, "Self-Supervised Predictive Learning: A Negative-Free Method for Sound Source Localization in Visual Scenes", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3222-3231, 2022.

[4] Y. Lin, H. Tseng, H. Lee, Y. Lin, and M. Yang, "Unsupervised Sound Localization via Iterative Contrastive Learning", *Computer Vision and Image Understanding*, pp. 103602, 2023.

[5] P. Morgado, I. Misra and N. Vasconcelos, "Robust Audio-Visual Instance Discrimination", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12934-12945, 2021.

[6] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised Contrastive Learning", *Advances in Neural Information Processing Systems*, pp. 18661-18673, 2020.

[7] T. Huynh, S. Kornblith, M. R. Walter, M. Maire, and M. Khademi, "Boosting Contrastive Self-Supervised Learning with False Negative Cancellation", *In Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2785-2795, 2022.

[8] J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei, "ImageNet: A large-scale hierarchical image database", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 248-255, 2009.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016.

[10] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721-725, 2020.

[11] A. Senocak, T. H. Oh, J. Kim, M. H. Yang, and I. S. Kweon, "Learning to localize sound source in visual scenes", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4358-4366, 2018.

[12] A. Senocak, H. Ryu, J. Kim, and In So Kweon, "Less Can Be More: Sound Source Localization With a Classification Model", *In Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3308-3317, 2022.

[13] ZHOU, Jinxing, et al., "Audio–visual segmentation", *European Conference on Computer Vision*, pp. 386-403, 2022.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", *Advances in Neural Information Processing Systems*, 2012.

[15] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size", *arXiv preprint arXiv:1602.07360*, 2016.

[16] Y. Lyu, and I. W. Tsang, "Curriculum Loss: Robust Learning and Generalization against Label Corruption", *arXiv preprint arXiv:1905.10045*, 2020.

[17] Pandeya, Y. Raj, and J. Lee, "Deep learning-based late fusion of multimodal information for emotion classification of music video", *Multimedia Tools and Applications*, pp. 2887-2905, 2021.

선 주 형

2022년 한국과학기술원 전산학부(학사). 2022년~현재 한국과학기술원 전산학부 석사과정 재학 중. 관심 분야는 컴퓨터 비전, 멀티 모달 학습, 비디오 인공지능

김 재 윤

2017년 충남대학교 컴퓨터공학과(학사). 2019년 한국과학기술원 전산학부(석사). 2019년~현재 한국과학기술원 전산학부 박사과정 재학 중. 관심 분야는 컴퓨터 비전, 이미지 검색, 기계 학습

김 주 영

2007년 이화여자대학교 컴퓨터공학(학사). 2009년 이화여자대학교 컴퓨터정보통신공학(석사). 현재 LIG넥스원 무인체계연구소 수석연구원. 관심 분야는 객체 인식, 자율제어, 유무인 복합체계

이 영 주

2018년~2022년 언맨드솔루션 주임연구원. 현재 LIG넥스원 무인체계연구소 선임연구원. 관심 분야는 컴퓨터 비전, 센서 퓨전, 자율주행

한 혜 경

2020년 고려대학교 전기전자공학부(학사), 2022년 고려대학교 정보보호(석사), 현재 LIG넥스원 무인체계연구소 연구원. 관심 분야는 신호 처리, 컴퓨터 비전, 멀티미디어 포렌식, 기계 학습

윤 성 의

1999년 서울대학교 Computer Science (학사). 2001년 서울대학교 Computer Science(석사). 2005년 University of North Carolina at Chapel Hill(박사). 2005년 ~ 2007년 Lawrence Livermore National Laboratory, USA. 2007년 ~ 현재 한국과학기술원 전산학부 교수. 관심 분야는 렌더링, 이미지 검색, 컴퓨터 비전, 로보틱스