Look, Listen and Learn

https://arxiv.org/abs/1705.08168

0. Introduction

- 인간은 시각과 청각을 동시에 활용해 환경을 이해함
- 기존 딥러닝 모델은 시각·청각을 독립적으로 처리하는 경우가 많음
- 본 논문은 영상과 오디오 간의 동기화(self-supervised learning)를 활용한 멀티모달 학습 기법 제안
- 레이블 없는 방대한 영상 데이터를 활용하여 시청각 표현 학습 가능

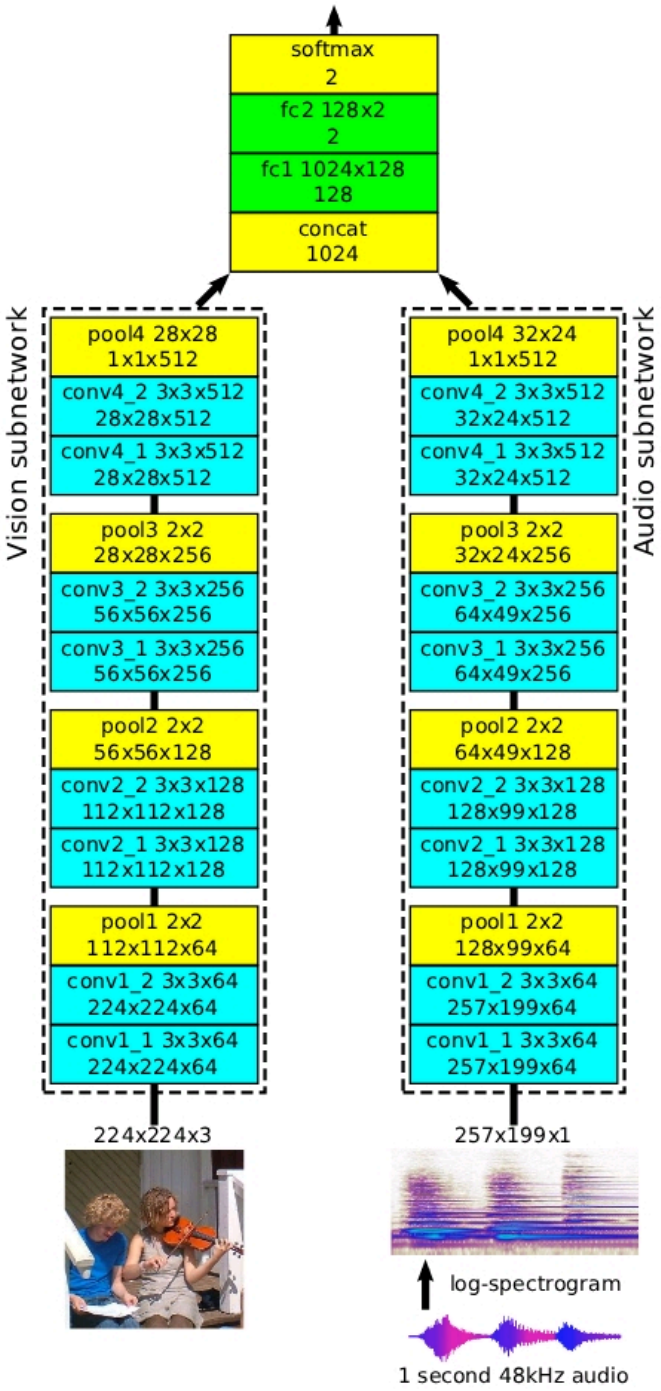
1. Overview

- 제안 모델 :
  - Look, Listen and Learn (L3-Net)
- 학습 방식 :
  - 짧은 비디오 클립 입력 (영상 프레임 + 오디오)
  - 두 모달리티가 동기화된 쌍인지 아닌지 판별하는 pretext task 수행
- 목표 :
  - 멀티모달 임베딩 공간 학습
- 장점 :
  - 별도의 라벨 필요 없이 self-supervised 방식으로 학습 가능

2. Challenges

- 비디오 데이터는 잡음(noise) 많음 (배경음, 화질 저하)
- 오디오와 비디오의 시간적 alignment 문제
- 라벨 없는 대규모 데이터에서 효과적인 학습 방법 설계가 필요

3. Method



- 구조:
  - 두 개의 CNN (영상용 / 오디오용) → 임베딩 추출 후 융합
- 학습 objective :

- Positive pair : 실제 같은 클립의 오디오와 비디오
- Negative pair : 무작위로 매칭된 오디오-비디오
- Binary classification loss (동기화 여부 예측)

- 학습 후 :
  - 임베딩을 다양한 downstream task에 활용

4. Experiments

- 데이터셋 :
  - AudioSet, Flickr-SoundNet 등
- downstream task :
  - 소리 구분 (Sound Classification)
  - 음원 위치 추정 (Sound Source Localization)
  - 크로스모달 검색 (Audio→Video, Video→Audio retrieval)
- 비교 baseline :
  - 랜덤 초기화 CNN, 단일 모달 CNN

5. Results

Method	Flickr-SoundNet	Kinetics-Sounds
Supervised direct	–	65%
Supervised pretraining	–	74%
L^3-Net	78%	74%

(a) ESC-50

Method	Accuracy
SVM-MFCC [26]	39.6%
Autoencoder [2]	39.9%
Random Forest [26]	44.3%
Piczak ConvNet [25]	64.5%
SoundNet [2]	74.2%
Ours random	62.5%
Ours	**79.3%**
Human perf. [26]	*81.3%*

(b) DCASE

Method	Accuracy
RG [27]	69%
LTT [19]	72%
RNH [28]	77%
Ensemble [32]	78%
SoundNet [2]	88%
Ours random	85%
Ours	**93%**

Method	Top 1 accuracy
Random	18.3%
Pathak *et al.* [24]	22.3%
Krähenbühl *et al.* [16]	24.5%
Donahue *et al.* [7]	31.0%
Doersch *et al.* [6]	31.7%
Zhang *et al.* [36] (init: [16])	32.6%
Noroozi and Favaro [21]	34.7%
Ours random	12.9%
Ours	32.3%

- L3-Net은 self-supervised임에도 불구하고 supervised baseline에 근접한 성능 달성
- 특히 음원 위치 추정과 크로스모달 검색에서 큰 성능 향상
- 오디오와 비디오 representation 학습에 모두 강력함

6. Insight

- 멀티모달 self-supervised 학습의 가능성 입증

- 대규모 라벨 없는 데이터 활용의 새로운 방향 제시

Look, Listen and Learn