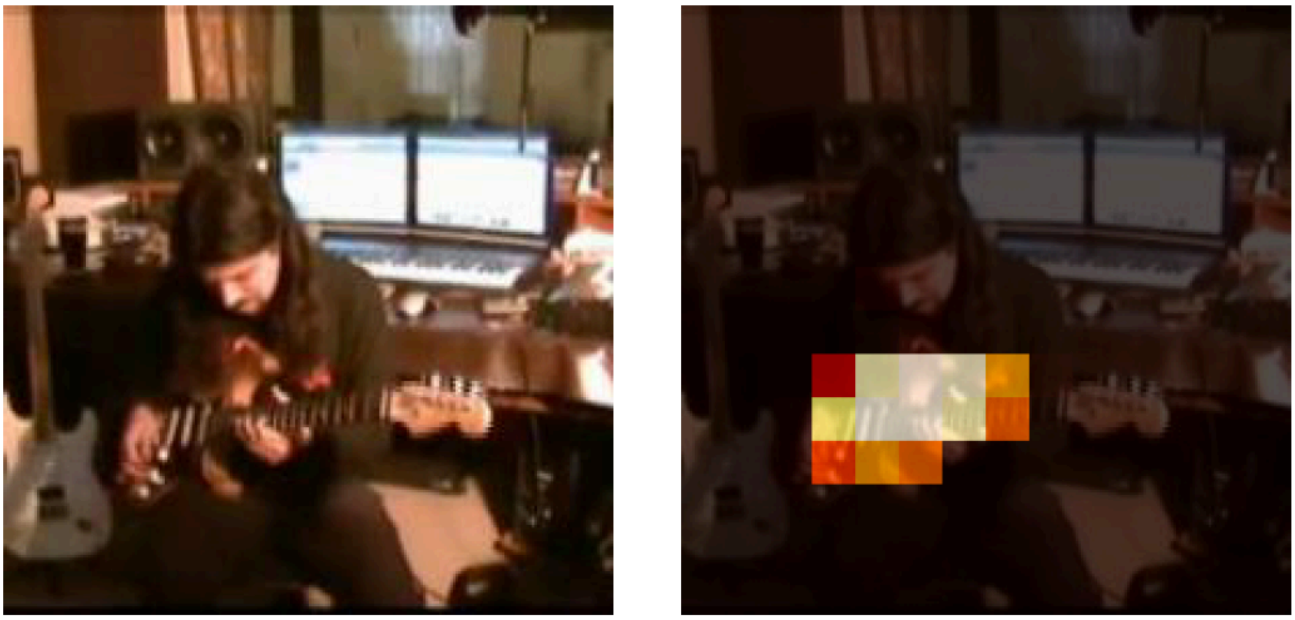
Images That Sound

- https://arxiv.org/pdf/1712.06651.pdf (https://arxiv.org/pdf/1712.06651.pdf)

요약

- Image 와 Sound 간의 상호 대응(Audio-Visual Correspondence) 학습
- Image/Sound Retrival given Sound/Image
- Object localization given Sound



(a) Input image with sound (b) Where is the sound?

Figure 1: **Where is the sound?** Our method learns, without a single labelled example, to, given an input image and sound clip, localize the object that makes the sound.

주요 키워드

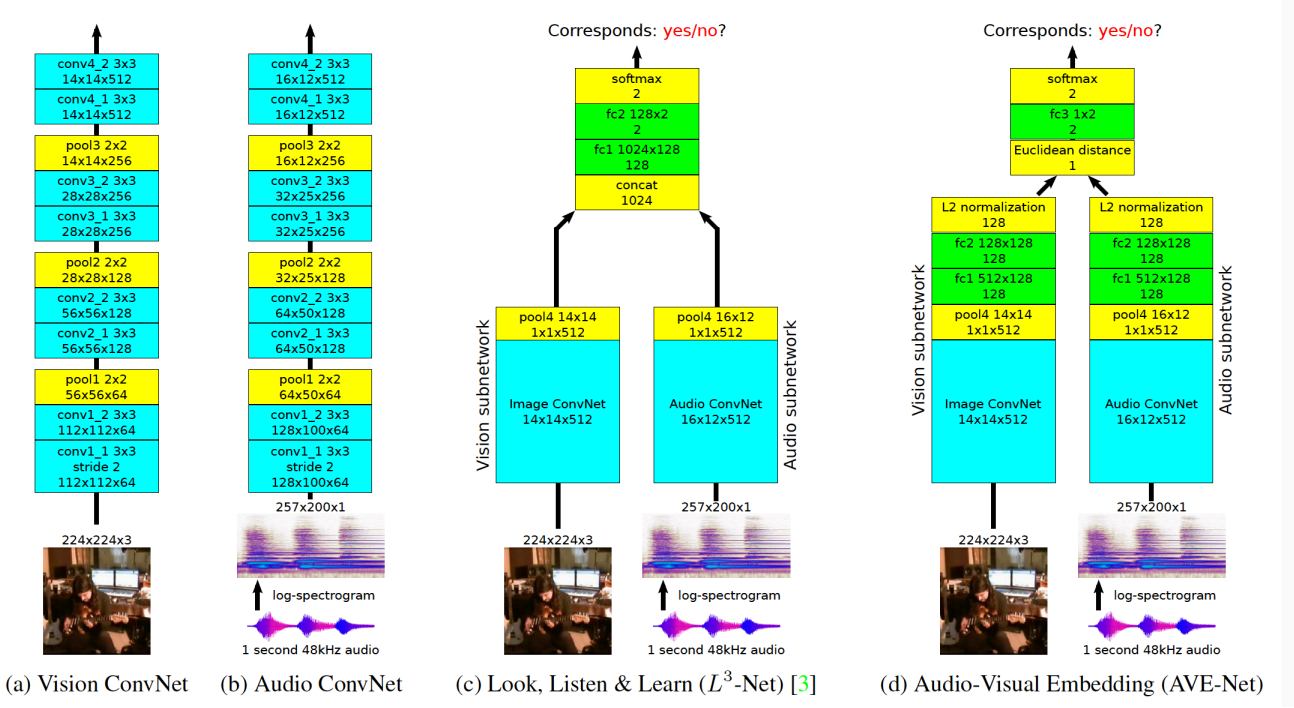
- Cross-Modal Learning
  - image/audio, image/text
- Cross-Modal Self-Supervision
  - youtue video => image/audio correspondence
- Cross-Modal Retrieval

DataSet

- Youtube video clip
  - 주로 악기가 나오는 동영상 위주
  - 부정확한 점이 많다
    - 동영상 설명이 악기 이름 여러개, 어떤 프레임에 어떤 악기 등장인지 명확치 않음
    - 악기가 아니라 앨범 커버나 가수 얼굴 등이 나오는 영상도 있음
  - 클립별로 video label은 있지만 학습에는 쓰지 않고 나중에 성능 측정시에만 사용
    - self-supervision
- 입력 feature
  - 1초 영상/음성 freame
    - movement hint는 없다.

Image/Video 일치 판단 네트워크

- Look,Listen,Learn L3 네트워크랑 유사
- L3랑 다르게 Image embeddeing, Sound embedding Representation Learning에 집중
- correspondece하면 두 embeddeing 사이의 거리가 가까워지게 학습



(a) Vision ConvNet (b) Audio ConvNet (c) Look, Listen & Learn (L^3-Net) [3] (d) Audio-Visual Embedding (AVE-Net)

Image/Video Retrieval

- L3에서는 불가능
- 여기서는 correspondence 를 고려한 embedding 학습을 잘 시킴
- 주어진 image/sound에 대해서 연관된 sound/image N 개 retrieval

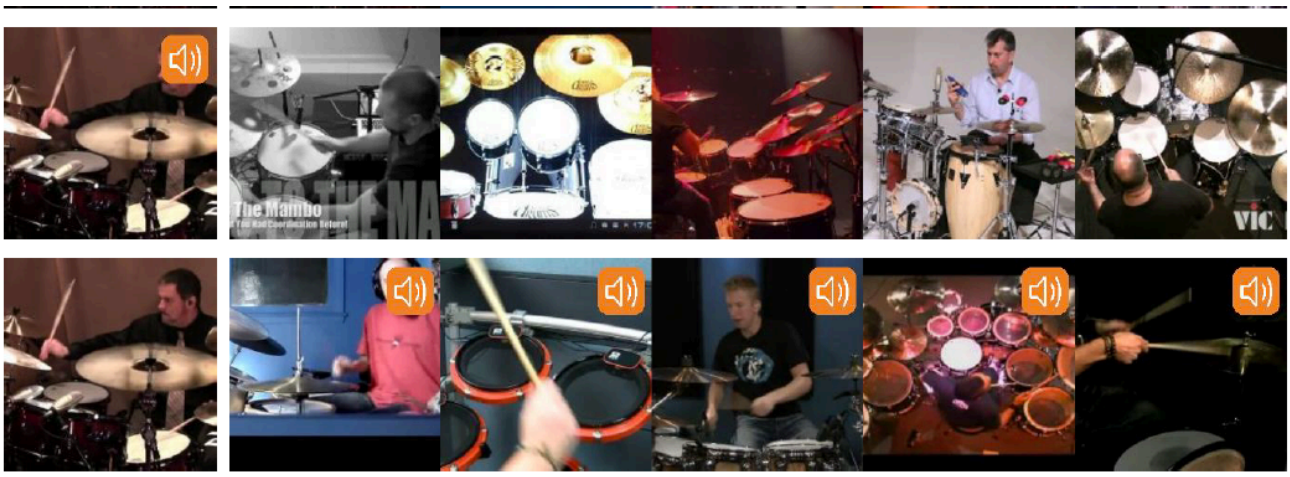


Figure 3: **Cross-modal and intra-modal retrieval.** Each

Method	im-im	im-aud	aud-im	aud-aud
Random chance	.407	.407	.407	.407
L^3-Net [3]	.567	.418	.385	.653
L^3-Net with CCA	.578	.531	.560	.649
VGG16-ImageNet [29]	.600	–	–	–
VGG16-ImageNet + L^3-Audio CCA	.493	.458	.464	.618
AVE-Net	**.604**	**.561**	**.587**	**.665**

- Retrieval 성능 측정
  - the normalized discounted cumulative gain (nDCG).
    - https://en.wikipedia.org/wiki/Discounted_cumulative_gain (https://en.wikipedia.org/wiki/Discounted_cumulative_gain)

$$\mathrm{CG_p} = \sum_{i=1}^{p} rel_i$$

$$\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)} = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i+1)}$$

$$\mathrm{nDCG_p} = \frac{DCG_p}{IDCG_p},$$

where IDCG is ideal discounted cumulative gain,

$$\mathrm{IDCG_p} = \sum_{i=1}^{|REL|} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

Localizing objects that sound

- where is the object that is making the sound?
- use Multiple Instance Learning
  - AVC 를 target signal 로 하면서도
  - 중간 layzer에서 image의 각 부분별로 sound 연관 중요도를 측정
    - CNN의 마지막 레이어, 14*14 가 local region-level image descriptors

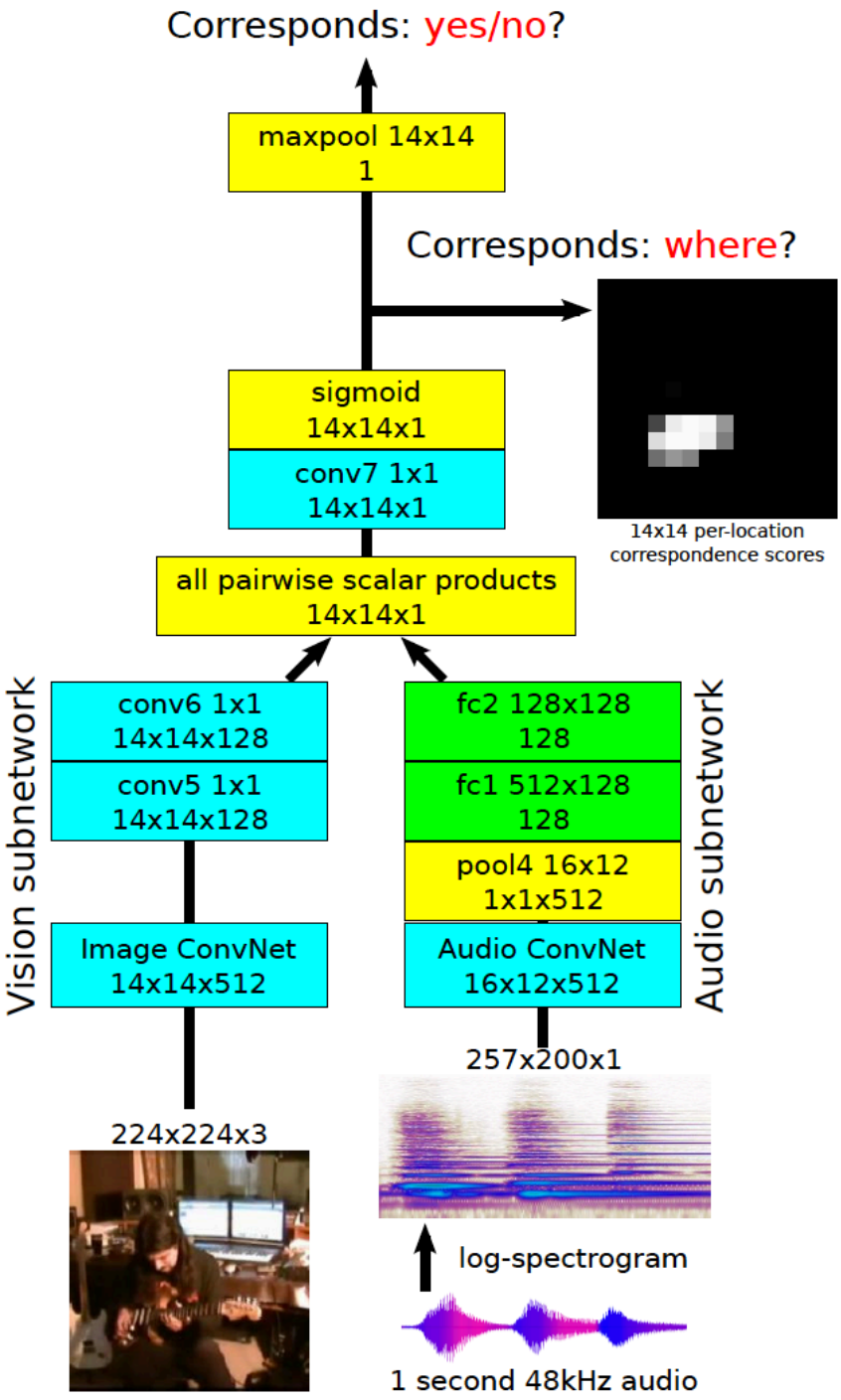
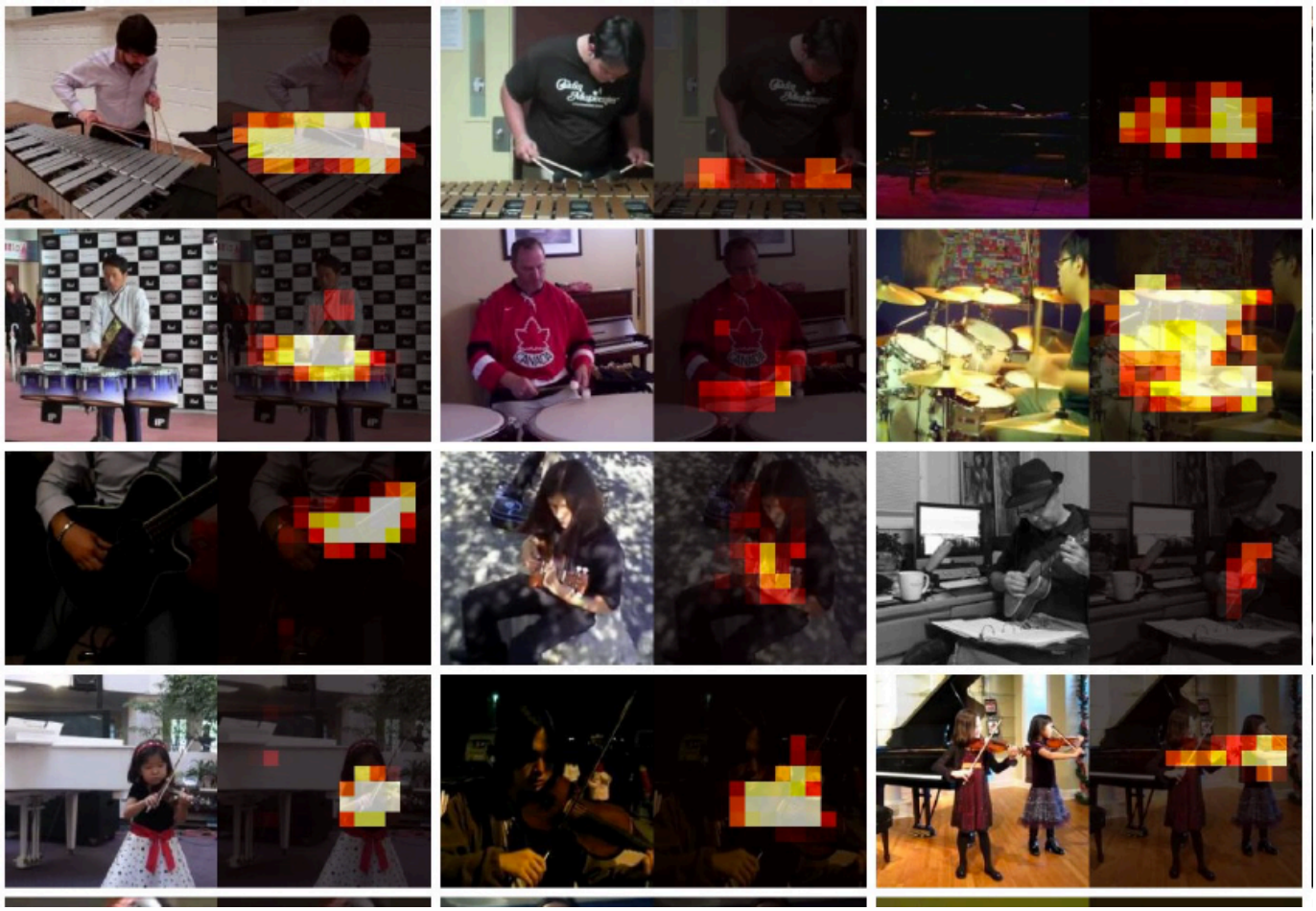
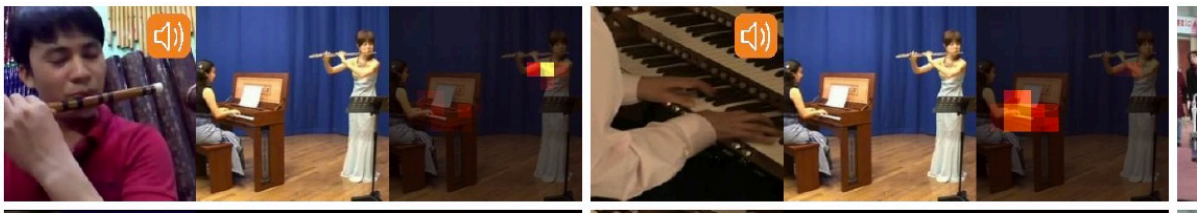


Figure 4: **Audio-Visual Object Localization (AVOL-Net).**



- 아래는 image와 sound 가 각각 다른 video에서 온 경우



- 아래는 multi-frame에서 각각의 sound object localization한 경우