#DLUPC

Day 4 Lecture 5

Audio and Vision

Eva Mohedano
eva.mohedano@insight-centre.org

Postdoctoral Researcher
Insight Centre for Data Analytics
Dublin City University

DCU

Insight
Centre for Data Analytics

http://bit.ly/dlcv2018

Contents

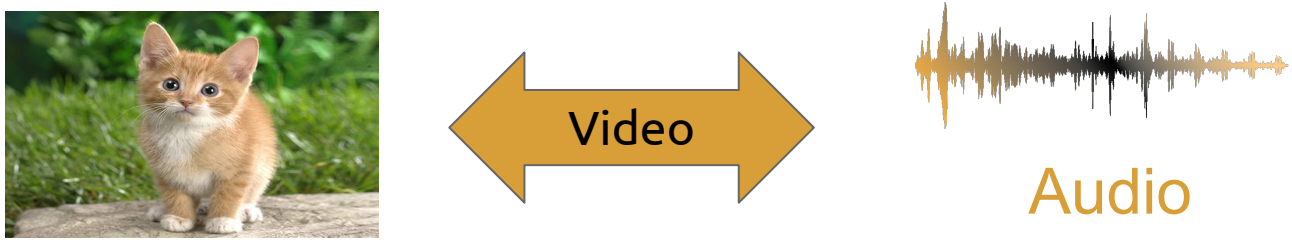
- Feature learning
- Cross-modal retrieval
- Sound source localization
- Sonorization

Contents

- **Feature learning**
- Cross-modal retrieval
- Sound source localization
- Sonorization

Self-supervised learning

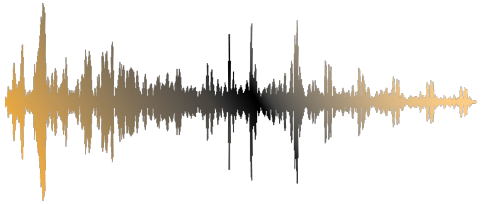
Self-supervised methods **learn features** by training a model to solve a task derived from the **input data itself**, without human labeling



Visual Feature Learning

Vision

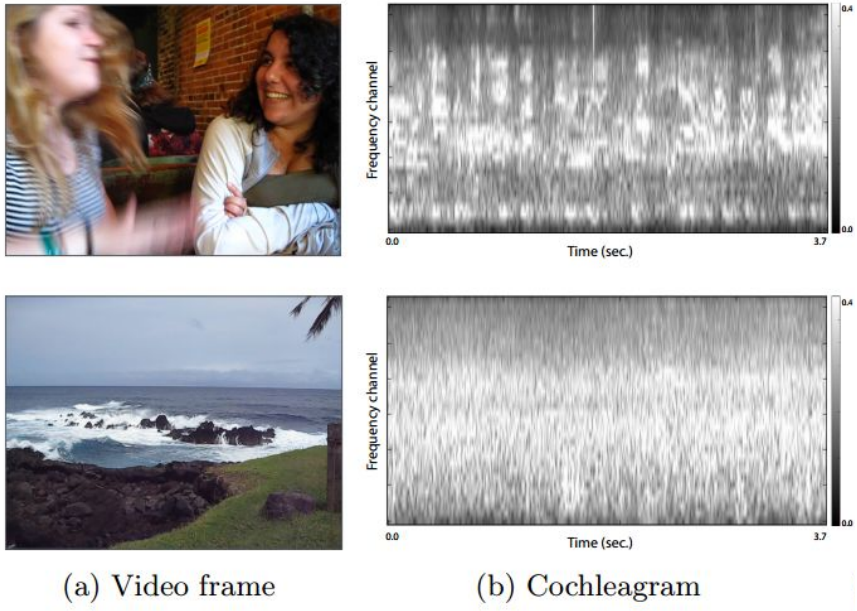
Video



Audio

Visual Feature Learning

Based on the assumption that ambient sound in video is related to the visual semantics.

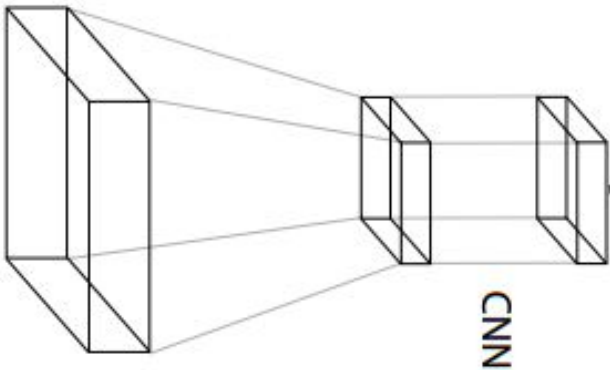


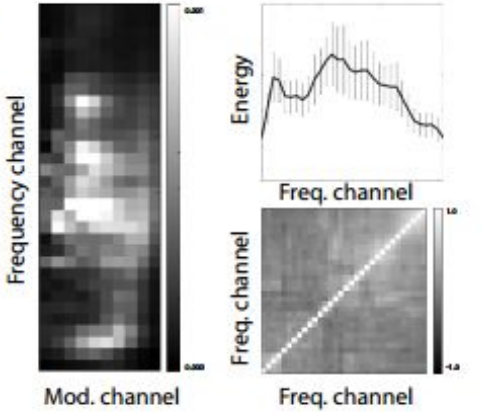
(a) Video frame (b) Cochleagram

Owens, Andrew, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. "Ambient sound provides supervision for visual learning." ECCV 2016

Visual Feature Learning

Use videos to train a CNN that predicts the audio statistics of a frame.





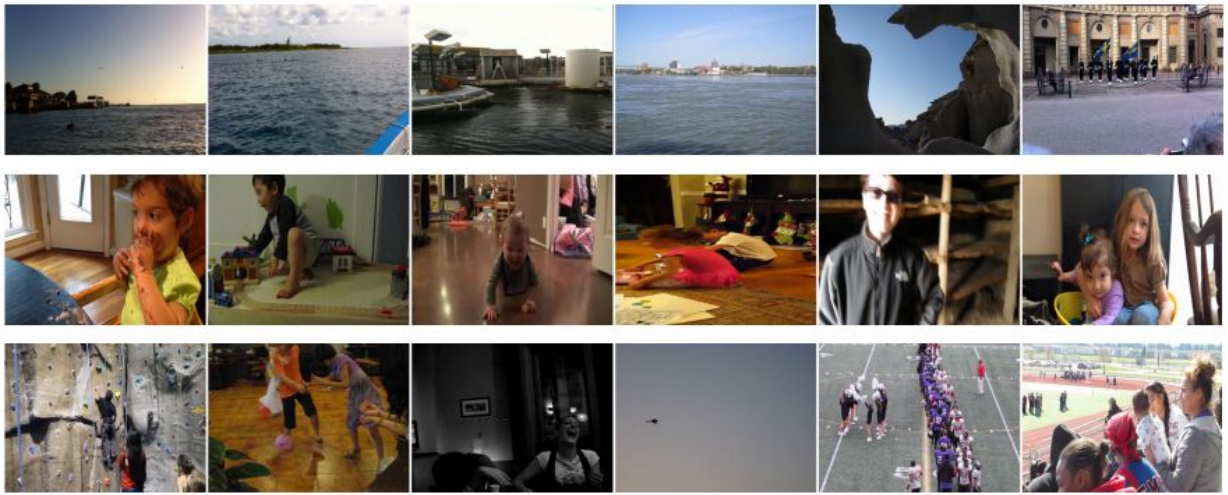
Self-supervised classification problem

Owens, Andrew, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. "Ambient sound provides supervision for visual learning." ECCV 2016

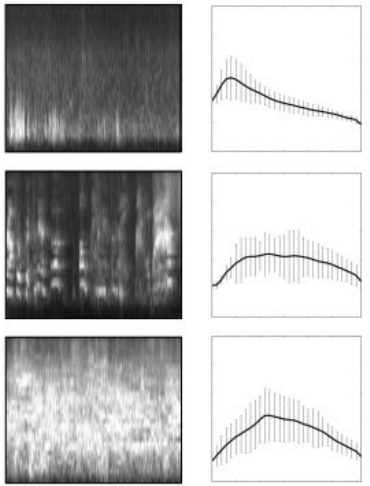
Visual Feature Learning

Task: Use the predicted audio stats to clusters images. Audio clusters built with K-means over training set

Cluster assignments at test time (one row=one cluster)



Average stats



Owens, Andrew, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. "Ambient sound provides supervision for visual learning." ECCV 2016

Visual Feature Learning

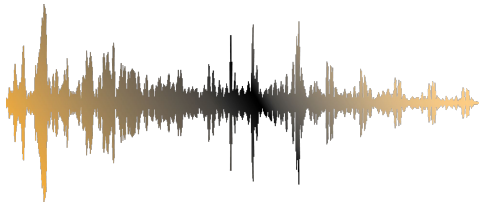
Although the CNN was not trained with class labels, local units with semantic meaning emerge.

Owens, Andrew, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. "Ambient sound provides supervision for visual learning." ECCV 2016

Audio Feature Learning

Vision

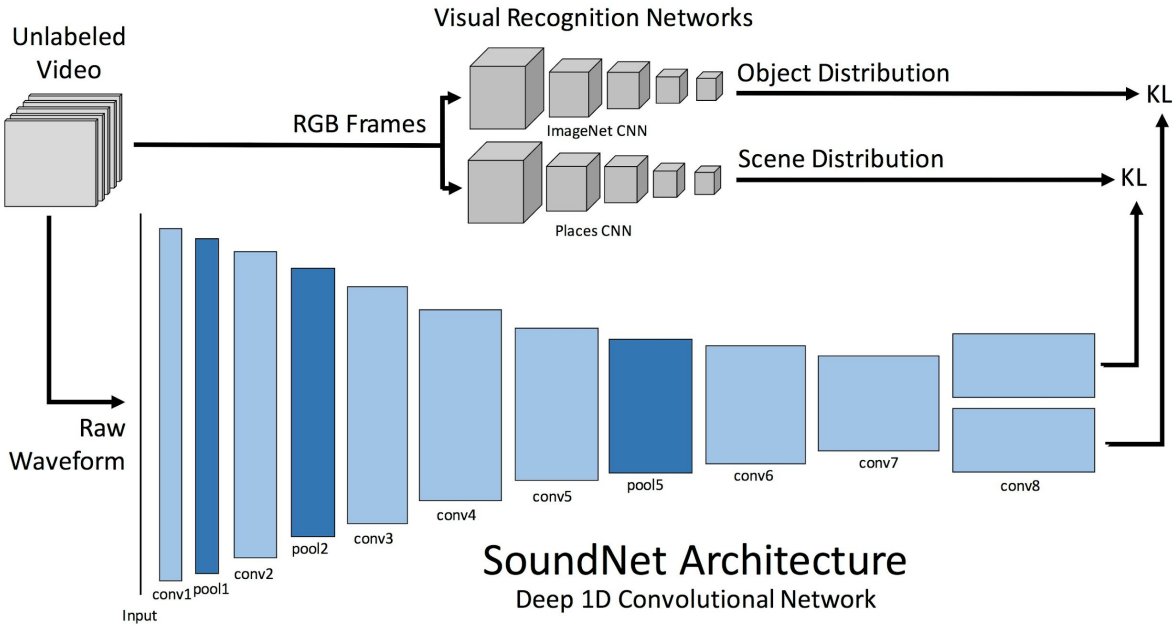
Video



Audio

Audio Feature Learning

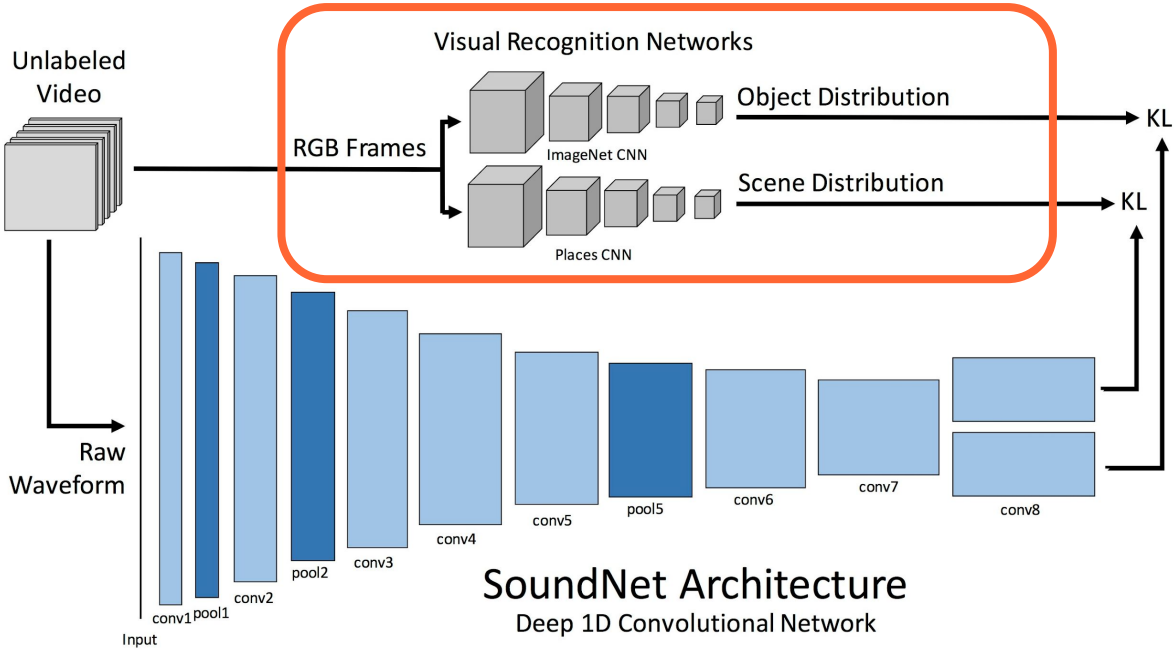
Pretrained visual ConvNets supervise the training of a model for sound representation



Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." NIPS 2016

Audio Feature Learning

Videos for training are unlabeled. Relies on Convnets trained on labeled images.



Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." NIPS 2016

Audio Feature Learning

Recognizing Objects and Scenes from Sound

Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." NIPS 2016

Audio Feature Learning

Hidden layers of Soundnet are used to train a standard SVM classifier that outperforms state of the art.

Method	Accuracy
RG [29]	69%
LTT [21]	72%
RNH [30]	77%
Ensemble [34]	78%
SoundNet	**88%**

Table 3: **Acoustic Scene Classification on DCASE:** We evaluate classification accuracy on the DCASE dataset. By leveraging large amounts of unlabeled video, SoundNet generally outperforms hand-crafted features by 10%.

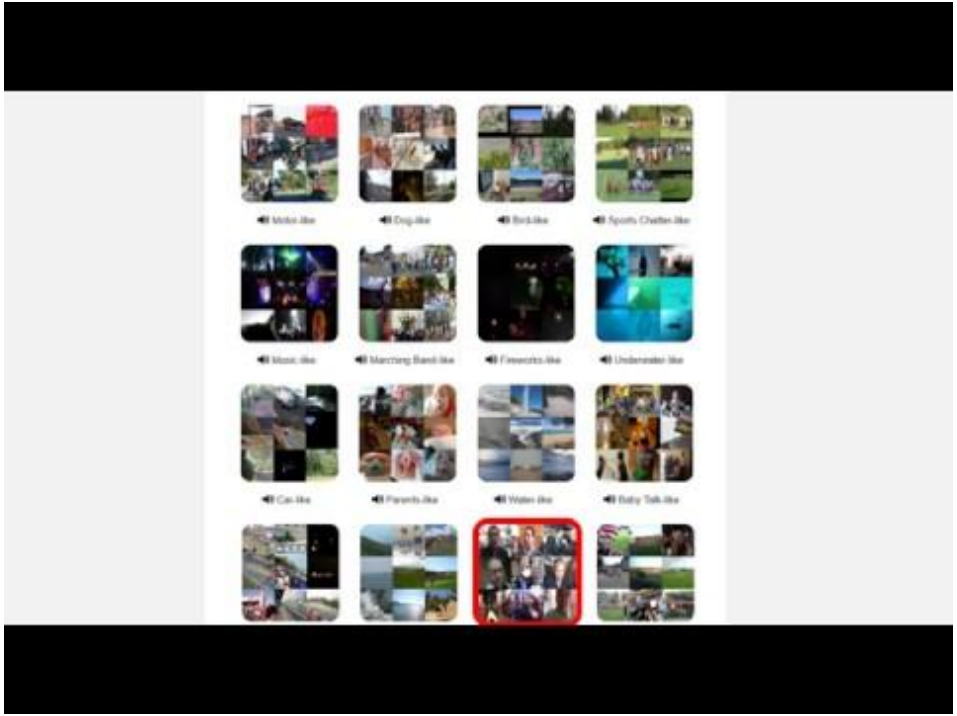
Method	Accuracy on ESC-50	Accuracy on ESC-10
SVM-MFCC [28]	39.6%	67.5%
Convolutional Autoencoder	39.9%	74.3%
Random Forest [28]	44.3%	72.7%
Piczak ConvNet [27]	64.5%	81.0%
SoundNet	**74.2%**	**92.2%**
Human Performance [28]	81.3%	95.7%

Table 4: **Acoustic Scene Classification on ESC-50 and ESC-10:** We evaluate classification accuracy on the ESC datasets. Results suggest that deep convolutional sound networks trained with visual supervision on unlabeled data outperforms baselines.

Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." NIPS 2016

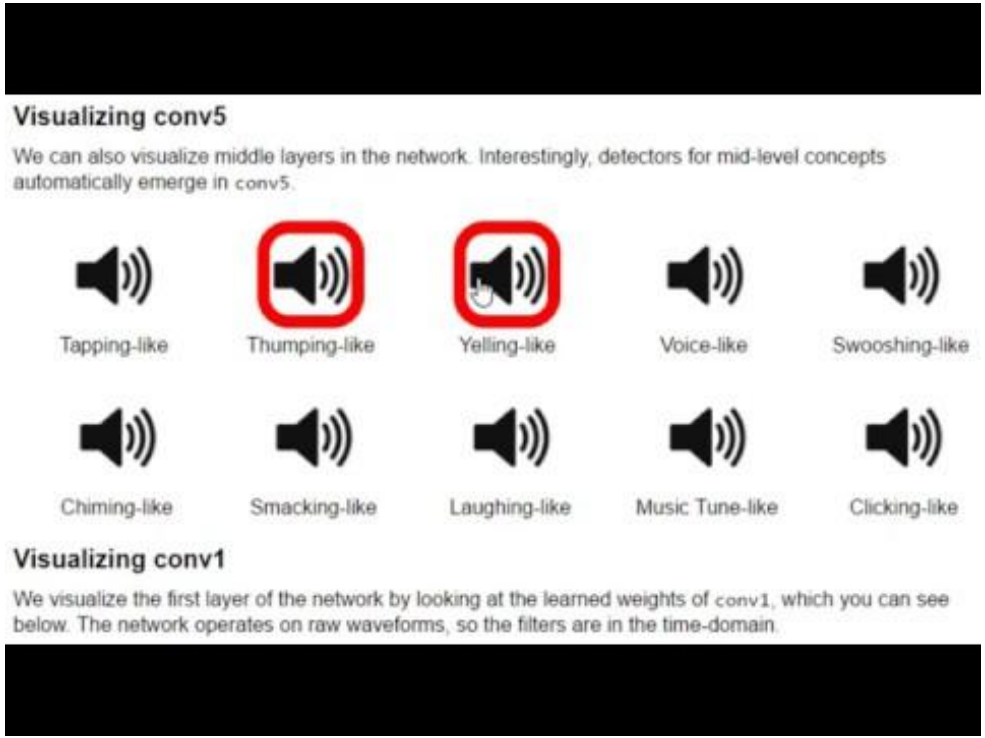
Audio Feature Learning

Hearing sounds that most activate a neuron in the sound network (conv7)



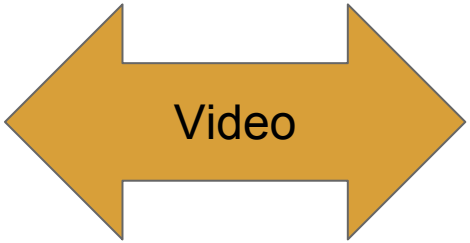
Audio Feature Learning

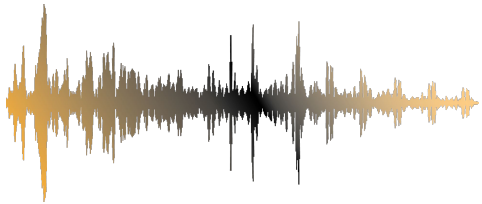
Hearing sounds that most activate a neuron in the sound network (conv5)



Audio & Visual feature learning

Vision

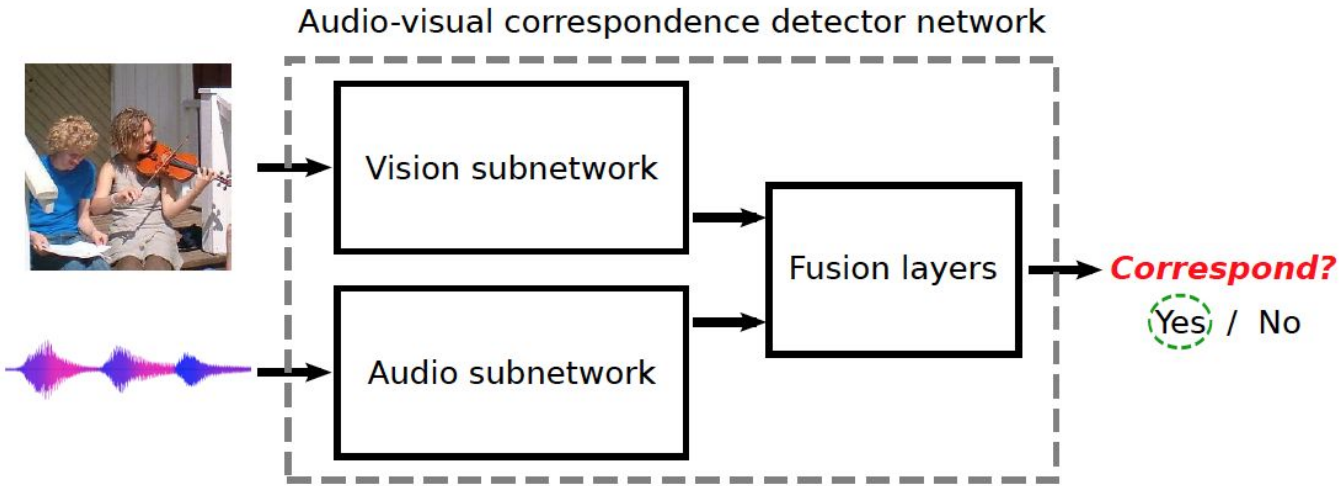


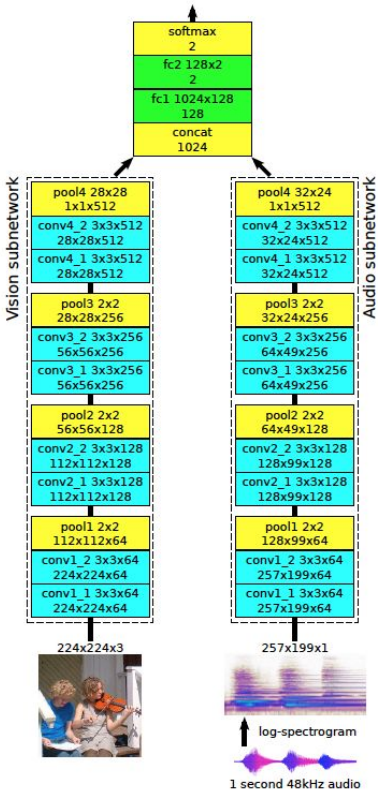


Audio

Audio & Visual feature learning

Audio and visual features learned by assessing **correspondence**.



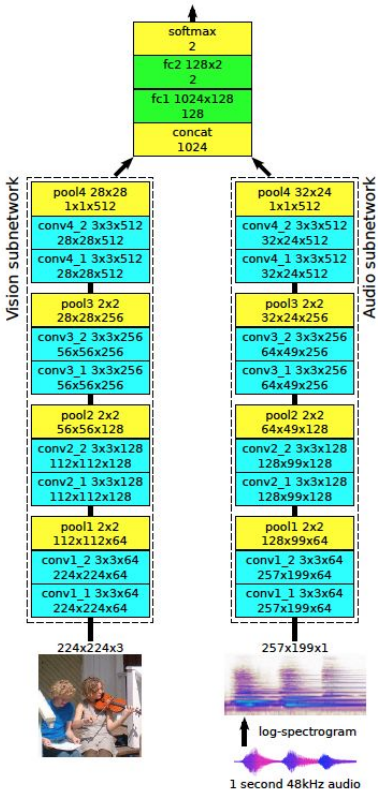


Arandjelović, Relja, and Zisserman, Andrew. "Look, Listen and Learn." *ICCV 2017.*

Audio & Visual feature learning

Audio and visual features learned by assessing **correspondence**.

- Large collection of **unlabelled videos**
- Flickr-SoundNet → Subset of 500k videos
- Kinetics-Sound → Subset of 19k videos from the Kinetics dataset
- 10s length videos
- 1 second of audio as input.
- A visual frame happening within that second is a match



Arandjelović, Relja, and Andrew Zisserman. "Look, Listen and Learn." *ICCV 2017.*

Audio & Visual feature learning

Most activated unit in *pool4* layer of the **visual network**

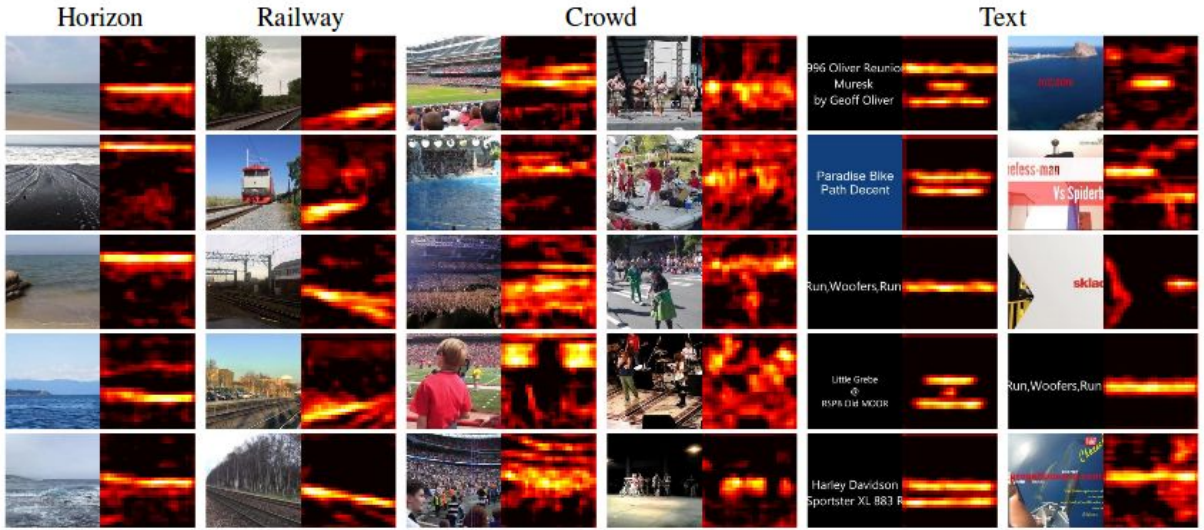
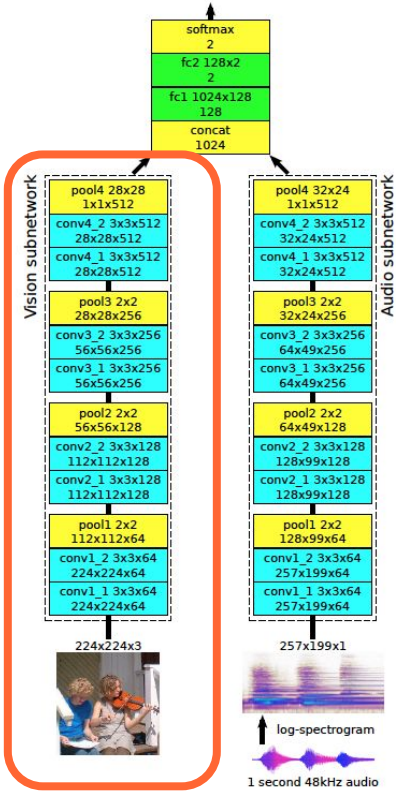


Figure 5. **Learnt visual concepts and semantic heatmaps (Flickr-SoundNet).** Each mini-column shows five images that most activate a particular unit of the 512 in `pool4` of the vision subnetwork, and the corresponding heatmap (for more details see Figures 3 and 4). Column titles are a subjective names of concepts the units respond to.

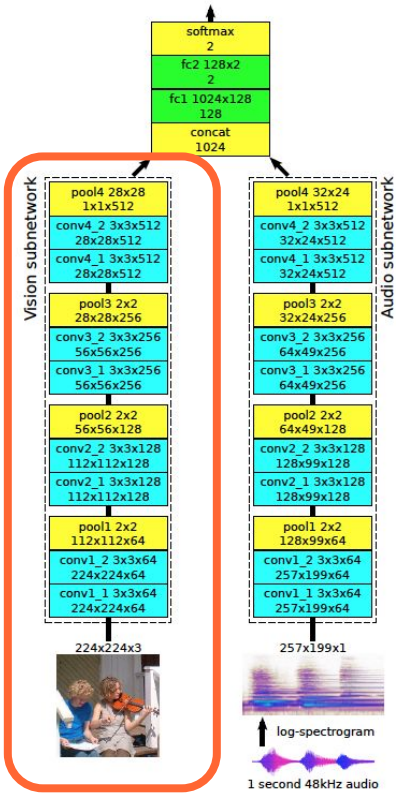


Arandjelović, Relja, and Andrew Zisserman. "Look, Listen and Learn." *ICCV 2017.*

Audio & Visual feature learning

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%

Table 3. **Visual classification on ImageNet.** Following [36], our features are evaluated by training a linear classifier on the ImageNet training set and measuring the classification accuracy on the validation set. For more details and discussions see Section 3.4. All performance numbers apart from ours are provided by authors of [36], showing only the best performance for each method over all parameter choices (*e.g.* Donahue *et al.* [7] achieve 27.1% instead of 31.0% when taking features from `pool5` instead of `conv3`).



Arandjelović, Relja, and Andrew Zisserman. "Look, Listen and Learn." *ICCV 2017.*

Audio & Visual feature learning

Most activated unit in *pool4* layer of the **audio network**

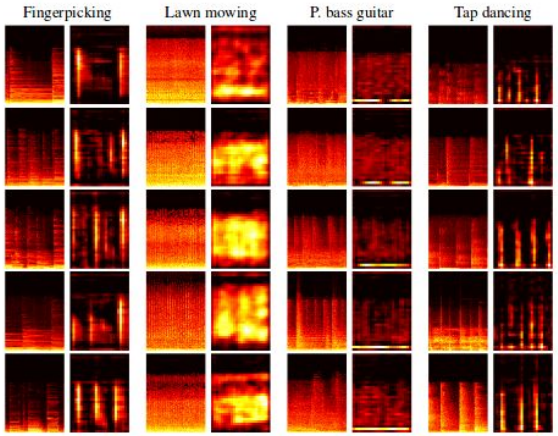
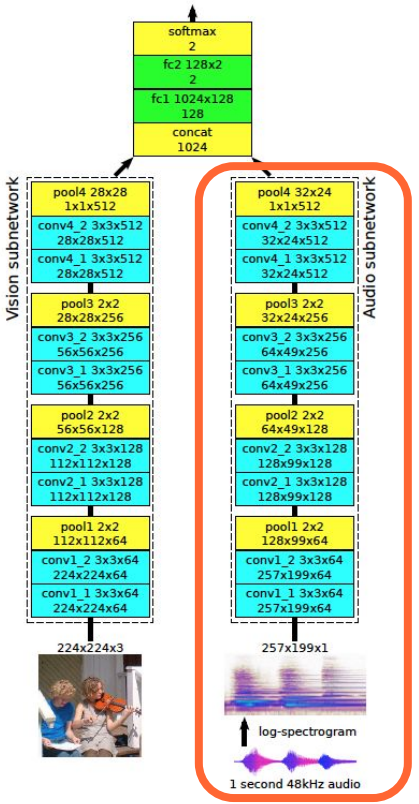


Figure 9. **Audio semantic heatmaps (Kinetics-Sounds).** Each pair of columns shows a single action class (top, "P." stands for "playing"), five log-spectrograms (left) and spectrogram semantic heatmaps (right) for the class. Horizontal and vertical axes correspond to the time and frequency dimensions, respectively. A semantic heatmap is obtained as a slice of activations of the unit from `conv4_2` of the audio subnetwork which shows preference for the considered class.



Arandjelović, Relja, and Andrew Zisserman. "Look, Listen and Learn." *ICCV 2017.*

Audio & Visual feature learning

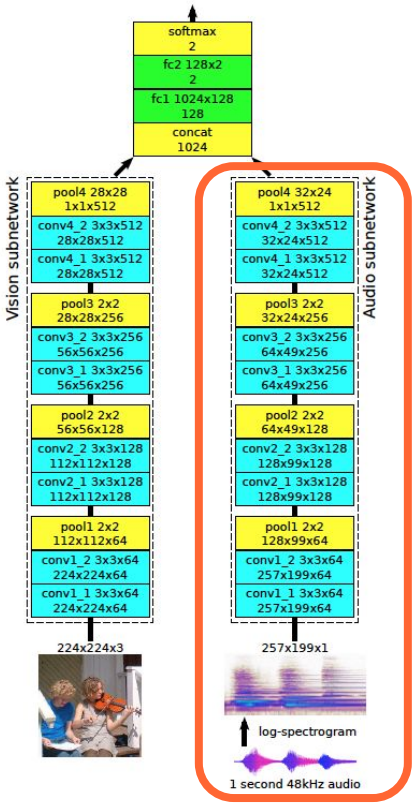
(a) ESC-50

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%**

Table 2. **Sound classification.** "Ours random" is an additional baseline which shows the performance of our network without L^3-training. Our L^3-training sets the new state-of-the-art by a large margin on both benchmarks.



Arandjelović, Relja, and Andrew Zisserman. "Look, Listen and Learn." *ICCV 2017.*

Contents

- Feature learning
- **Cross-modal retrieval**
- Sound source localization
- Sonorization

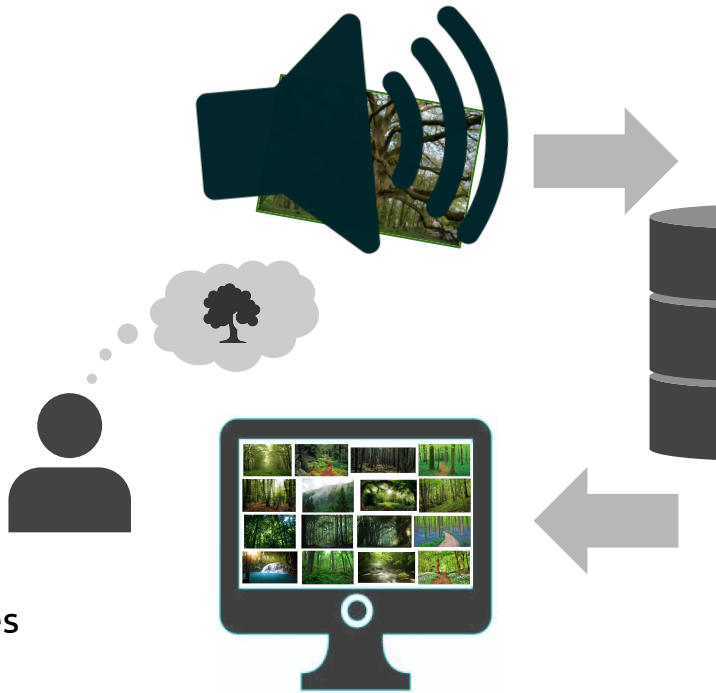
The problem: query by example

Given:

- An example query image that illustrates the user's information need
- A very large dataset of images

Task:

- Rank all images in the dataset according to how likely they are to fulfil the user's information need

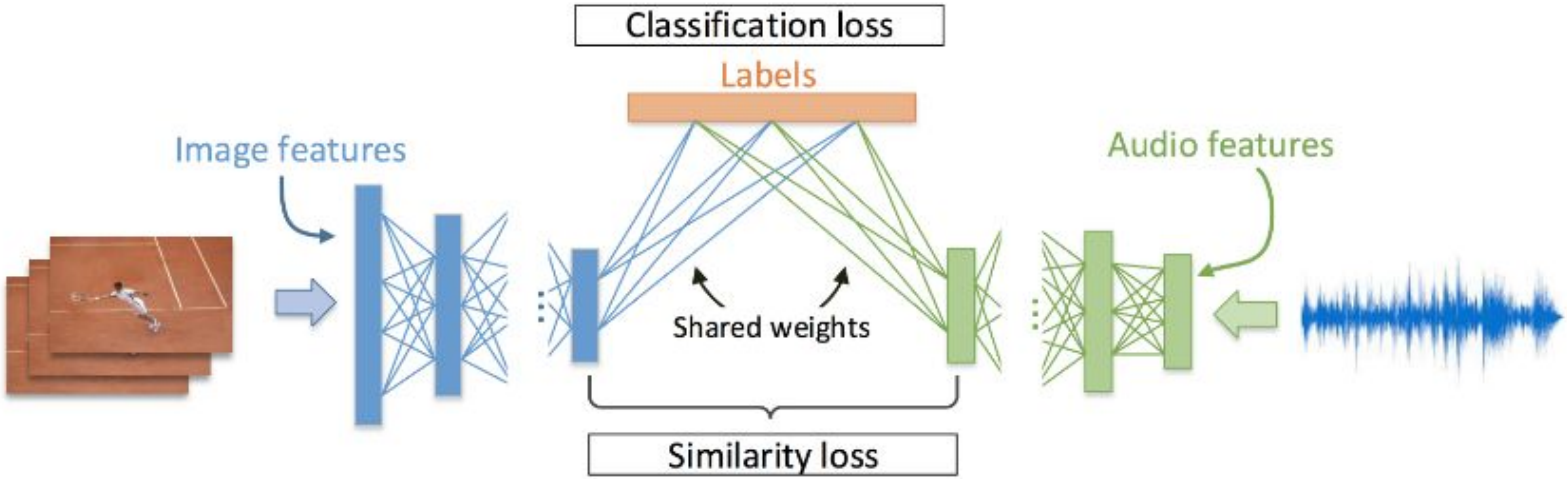


How to make audio and visual features comparable?

Cross-modal retrieval

YouTube 8M

Video-level features



Surís, Didac, Amanda Duarte, Amaia Salvador, Jordi Torres, and Xavier Giró-i-Nieto. "Cross-modal Embeddings for Video and Audio Retrieval." arXiv preprint arXiv:1801.02200 (2018).

Cross-modal retrieval

$$L = L_{cos} + \lambda L_{class}$$

Similarity loss

$$similarity = \cos(x, z) = \frac{\sum_{k=1} x_k z_k}{\sqrt{\sum_k^N x_k^2}\sqrt{\sum_i^N z_k^2}}$$

$$L_{cos}((\Phi^a, \Phi^i), y) = \\ = \begin{cases} 1 - \cos(\Phi^a, \Phi^i), & \text{if} \quad y = 1 \\ \max(0, \cos(\Phi^a, \Phi^i) - \alpha), & \text{if} \quad y = -1 \end{cases}$$

Classification Regularization

$$L_{class}(p^i, p^a, c^i, c^a) = -\sum_k (p_k^i \log(c_k^i) + (p_k^a \log(c_k^a))$$

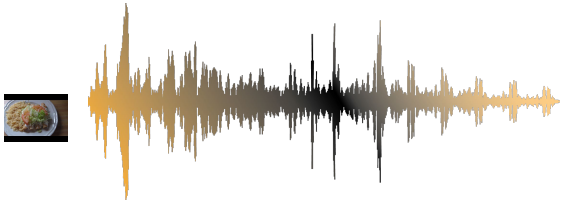
Surís, Didac, Amanda Duarte, Amaia Salvador, Jordi Torres, and Xavier Giró-i-Nieto. "Cross-modal Embeddings for Video and Audio Retrieval." arXiv preprint arXiv:1801.02200 (2018).

Cross-modal retrieval

Visual feature

Best match

Audio feature

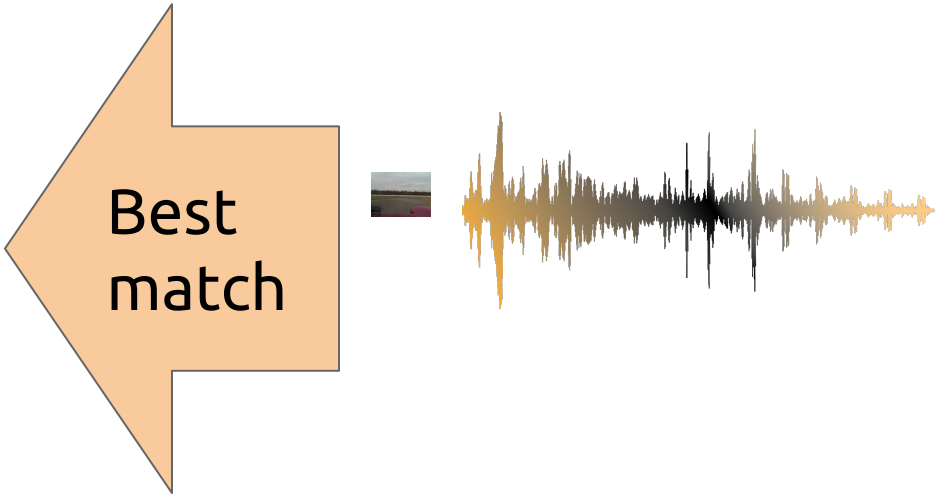


Surís, Didac, Amanda Duarte, Amaia Salvador, Jordi Torres, and Xavier Giró-i-Nieto. "Cross-modal Embeddings for Video and Audio Retrieval." arXiv preprint arXiv:1801.02200 (2018).

Cross-modal retrieval

Visual feature

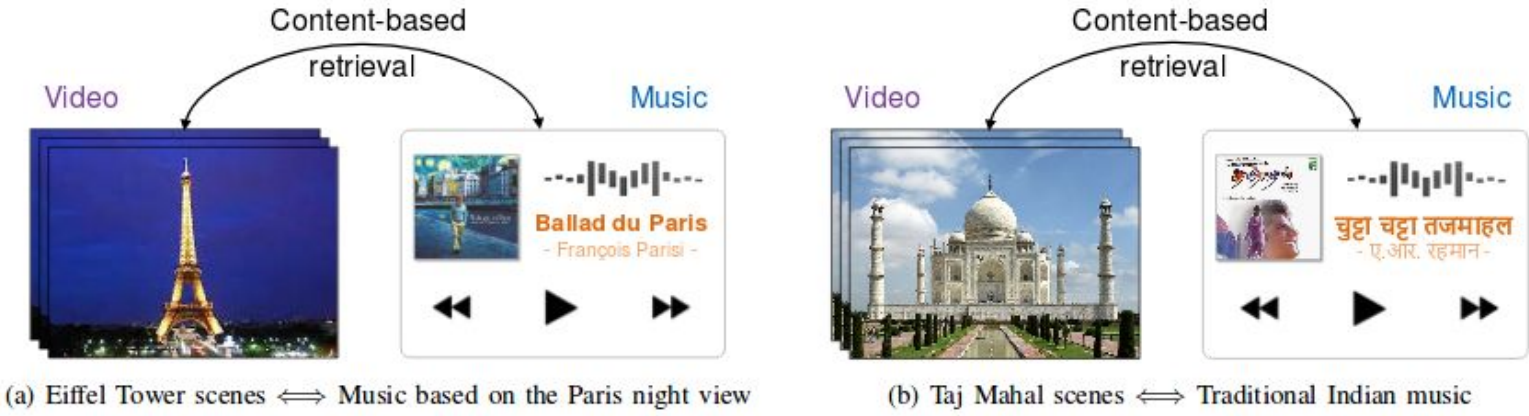
Audio feature



Surís, Didac, Amanda Duarte, Amaia Salvador, Jordi Torres, and Xavier Giró-i-Nieto. "Cross-modal Embeddings for Video and Audio Retrieval." arXiv preprint arXiv:1801.02200 (2018).

Cross-modal retrieval

Video-level features



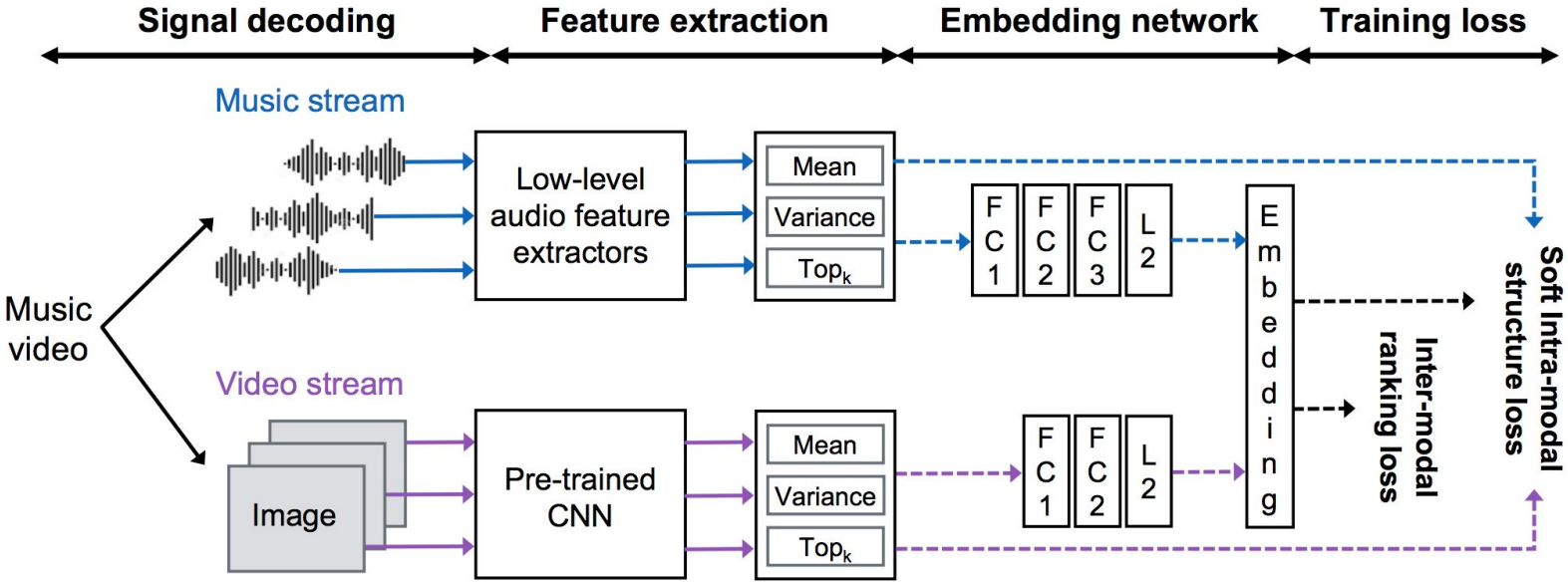
(a) Eiffel Tower scenes ⟺ Music based on the Paris night view

(b) Taj Mahal scenes ⟺ Traditional Indian music

Hung-IM Music-Video: Large-scale video-music pair dataset 200k video-music pairs
Based on all videos tagged with “music video” in Y8M

Sungeun Hong, Woobin Im and Hyun Seung Yang: CBVMR: Content-Based Video-Music Retrieval Using Soft Intra-Modal Structure Constraint, ICMR 2018 - Code available

Cross-modal retrieval

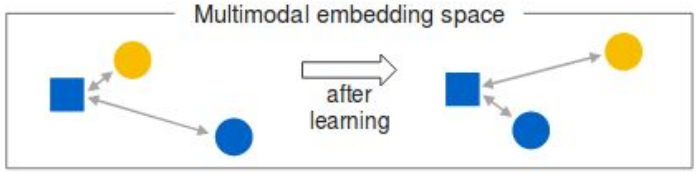


Sungeun Hong, Woobin Im and Hyun Seung Yang: CBVMR: Content-Based Video-Music Retrieval Using Soft Intra-Modal Structure Constraint, ICMR 2018 - Code available

Cross-modal retrieval

Inter-modal ranking constraint

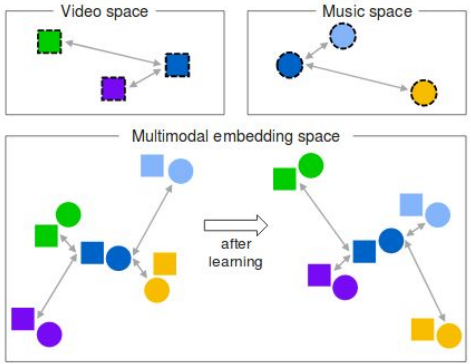
$$d(v_i, m_i) + e < d(v_i, m_j)$$



Soft intra-modal structure constraint

$$d(v_i, v_j) < d(v_i, v_k) \quad \text{if } d(\widetilde{v}_i, \widetilde{v}_j) < d(\widetilde{v}_i, \widetilde{v}_k)$$

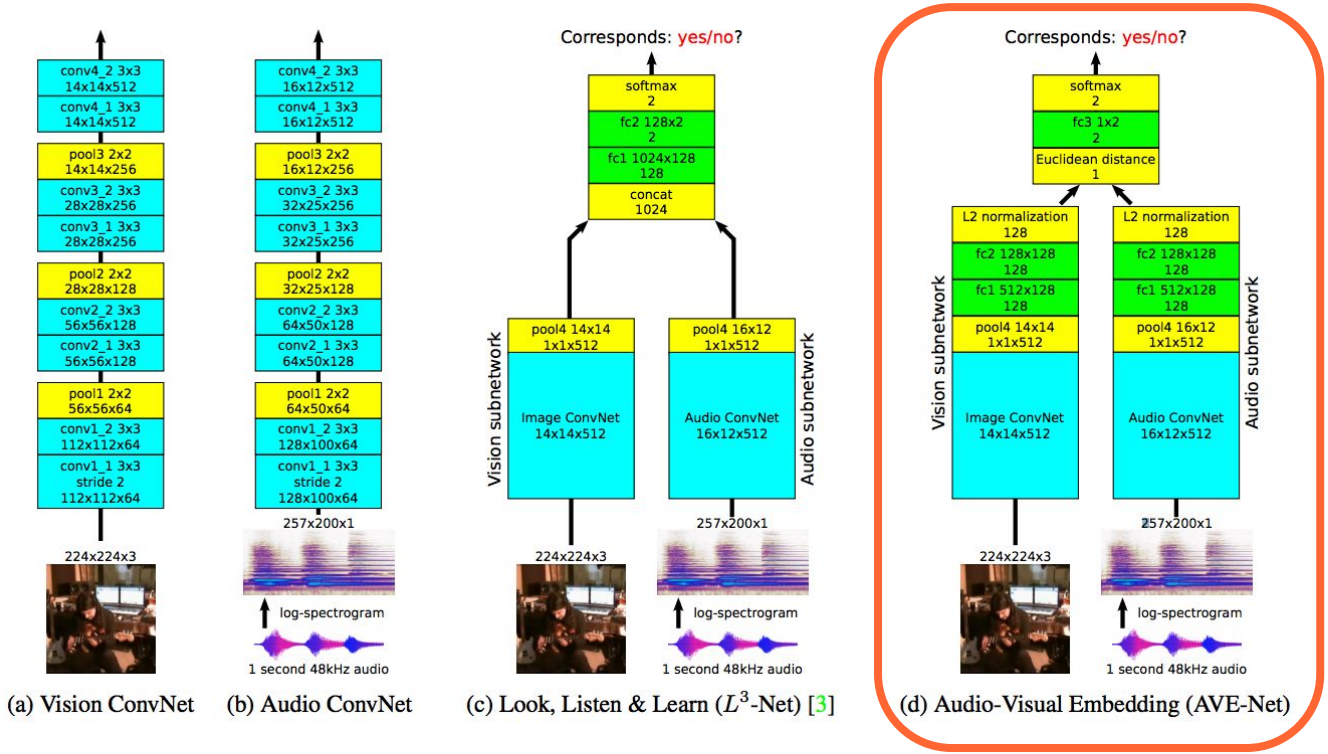
$$d(m_i, m_j) < d(m_i, m_k) \quad \text{if } d(\widetilde{m}_i, \widetilde{m}_j) < d(\widetilde{m}_i, \widetilde{m}_k)$$



Sungeun Hong, Woobin Im and Hyun Seung Yang: CBVMR: Content-Based Video-Music Retrieval Using Soft Intra-Modal Structure Constraint, ICMR 2018 - Code available

Cross-modal retrieval

Unsupervised feature learning



Arandjelović, Relja, and Andrew Zisserman. "Objects that Sound." arXiv preprint arXiv:1712.06651 (2017).

Cross-modal retrieval

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**

Table 1: **Cross-modal and intra-modal retrieval.** Comparison of our method with unsupervised and supervised baselines in terms of the average nDCG@30 on the AudioSet-Instruments test set. The columns headers denote the modalities of the query and the database, respectively, where *im* stands for *image* and *aud* for *audio*. Our AVE-Net beats all baselines convincingly.

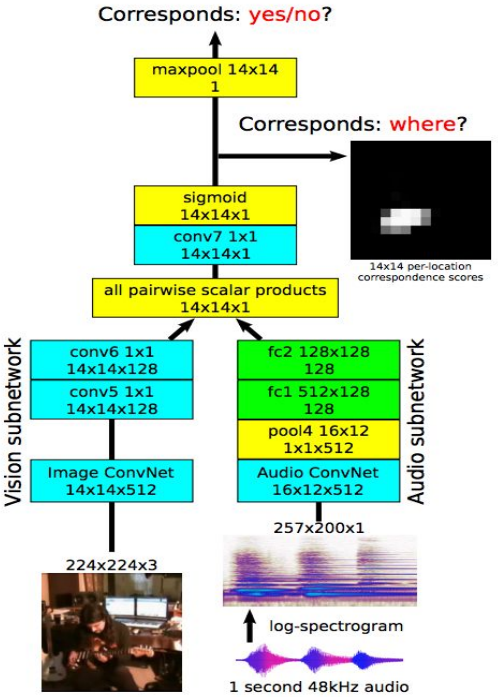
Arandjelović, Relja, and Andrew Zisserman. "Objects that Sound." arXiv preprint arXiv:1712.06651 (2017).

Contents

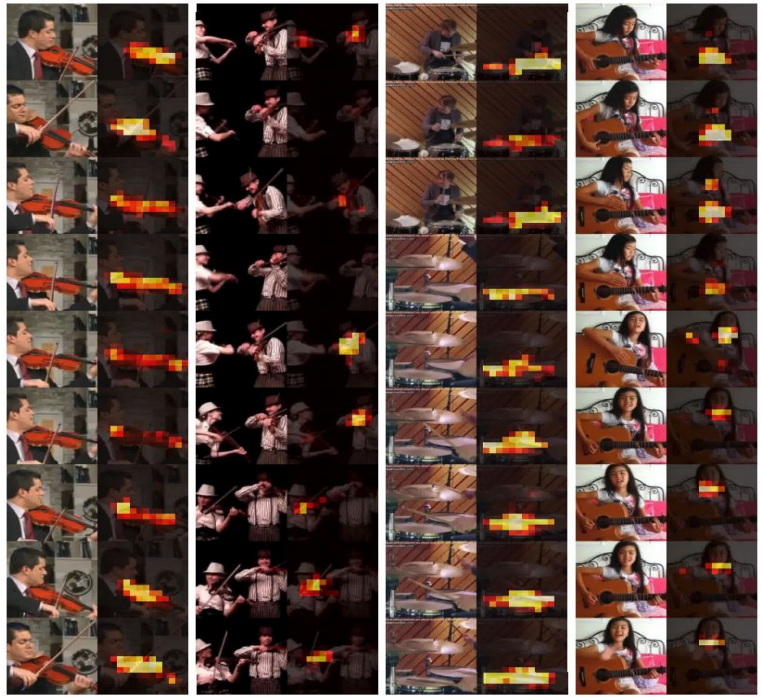
- Feature learning
- Cross-modal retrieval
- **Sound source localization**
- Speech generation

Sound source localization

Unsupervised feature learning



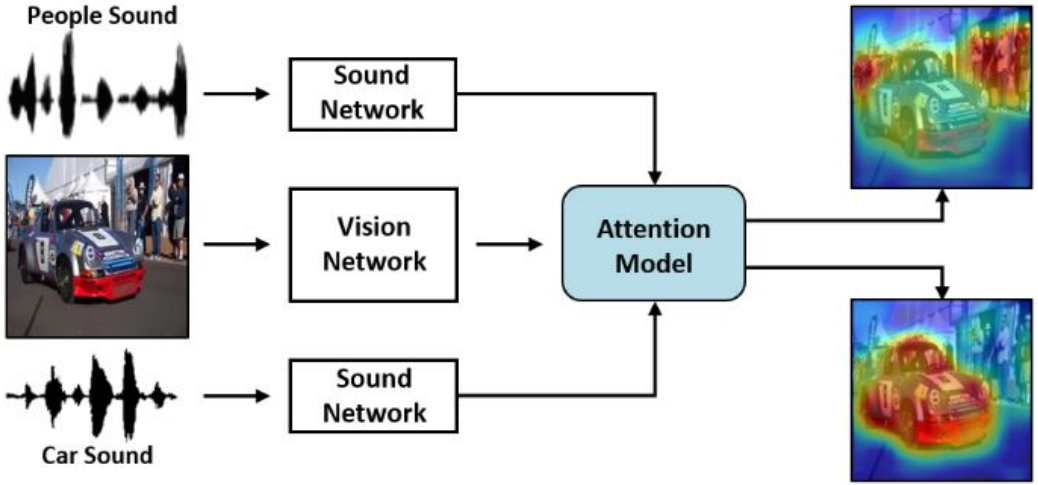
Audio-visual Object localization Network

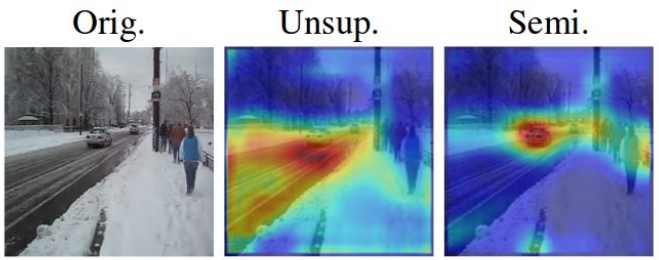


Arandjelović, Relja, and Andrew Zisserman. "Objects that Sound." arXiv preprint arXiv:1712.06651 (2017).

Sound source localization

Semi-supervised feature learning

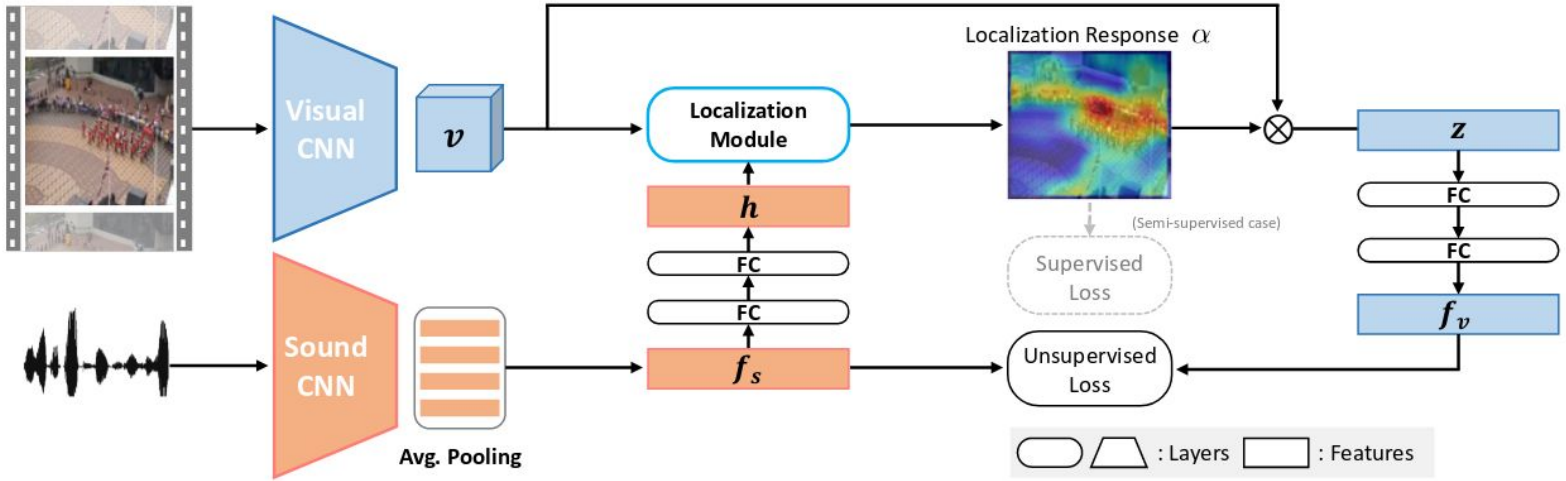




Senocak, A., Oh, T.H., Kim, J., Yang, M.H. and Kweon, I.S. Learning to Localize Sound Source in Visual Scenes. *CVPR 2018*

Sound source localization

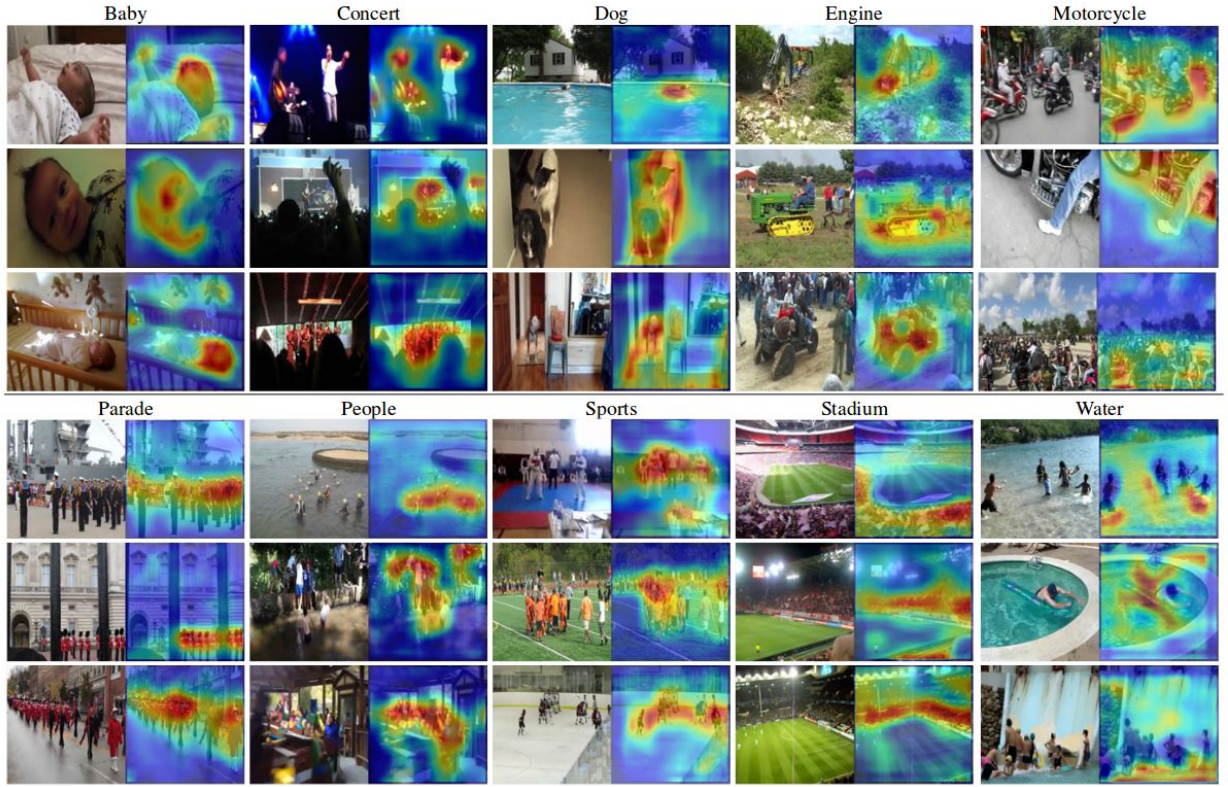
Semi-supervised feature learning



- Flickr-SoundNet subset of 144k frame pairs for training
- New dataset from Flickr with manual localization annotations (2.5k frames)
  - Manually annotate those frame with non-matching audio (edited or not present object)

Senocak, A., Oh, T.H., Kim, J., Yang, M.H. and Kweon, I.S. Learning to Localize Sound Source in Visual Scenes. *arXiv preprint 2018*

Sound source localization



Senocak, A., Oh, T.H., Kim, J., Yang, M.H. and Kweon, I.S. Learning to Localize Sound Source in Visual Scenes. *CVPR 2018*

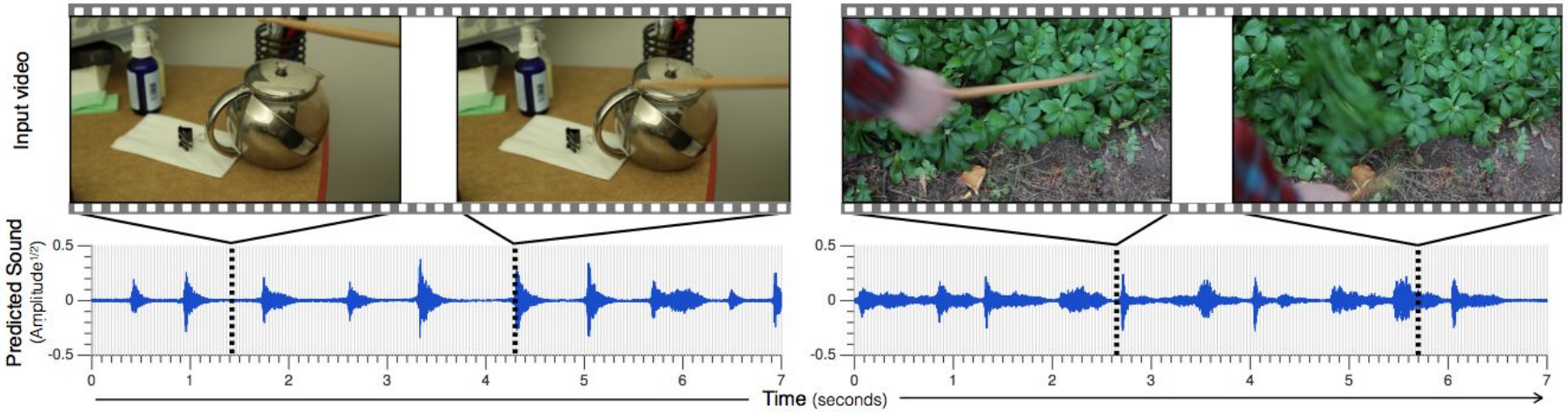
Senocak, A., Oh, T.H., Kim, J., Yang, M.H. and Kweon, I.S. Learning to Localize Sound Source in Visual Scenes. *CVPR 2018*

Contents

- Feature learning
- Cross-modal retrieval
- Sound source localization
- **Sonorization**

Sonorization

Learn synthesized sounds from videos of people hitting objects with a drumstick.

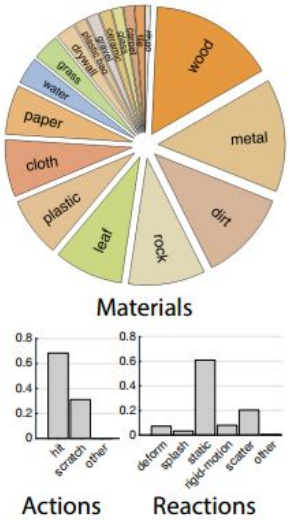


Are features learned when training for sound prediction useful to identify the object's material?

Owens et al. Visually indicated sounds. CVPR 2016

Sonorization

The Greatest Hits Dataset



Owens et al. [Visually indicated sounds.](#) CVPR 2016

Sonorization

The Greatest Hits Dataset

Owens et al. Visually indicated sounds. CVPR 2016

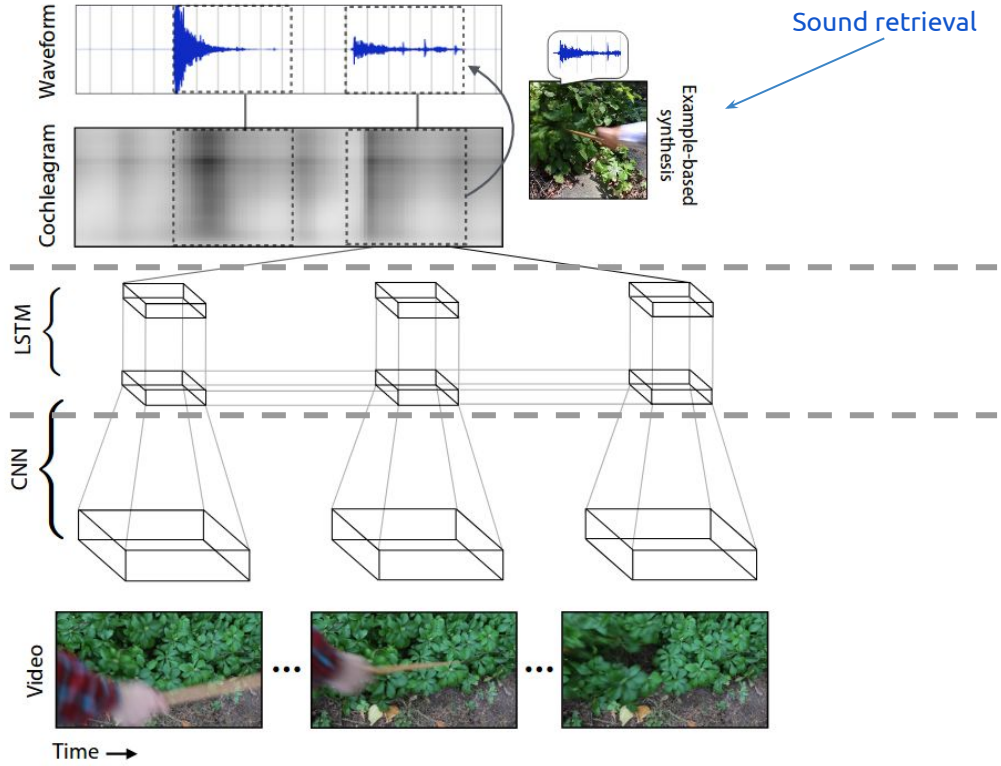
Sonorization

Sound Generation

2-layer LSTM
For sound feature regression

AlexNet FC7
For image feature extraction

Two stream CNN
At each time step, RGB frame (one stream) and
3-spatiotemporal image (current, prev and following)

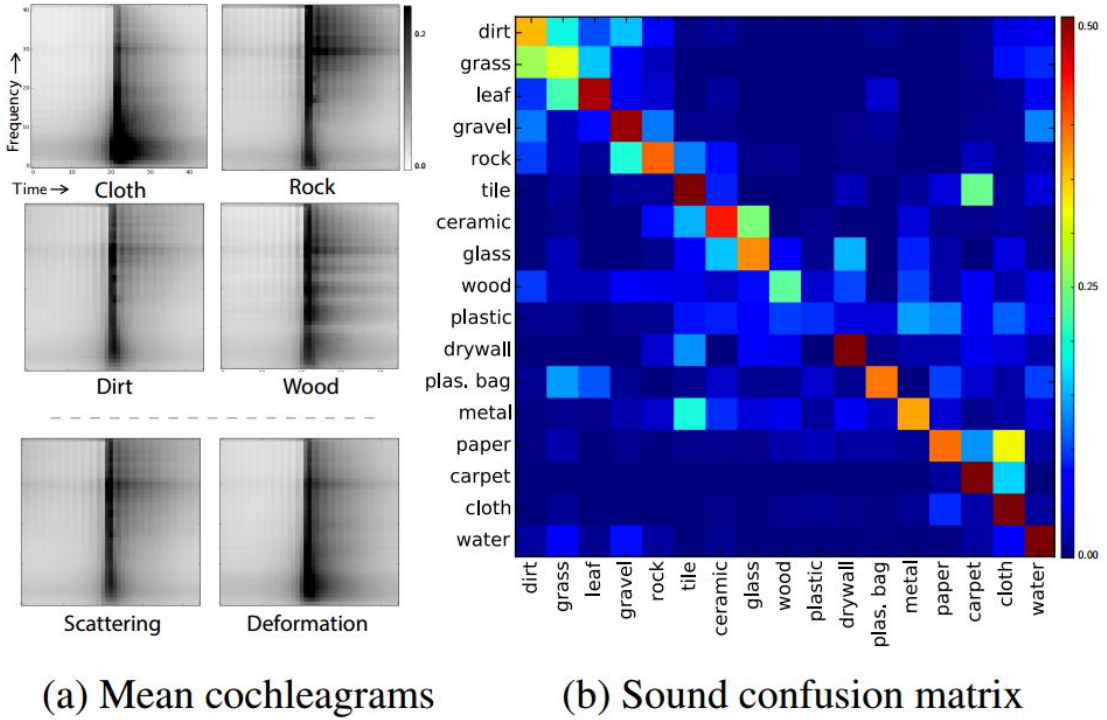


Owens et al. Visually indicated sounds. CVPR 2016

Sonorization

Owens et al. Visually indicated sounds. CVPR 2016

Sonorization



(a) Mean cochleagrams (b) Sound confusion matrix

Owens et al. Visually indicated sounds. CVPR 2016

Visual to Sound

Model trained to **directly predict raw signal**

Inputs RGB + Optical flow

LSTM

SampleRNN for sound generation

Dataset:
Visually Engaged and Grounded AudioSet (VEGAS)

Zhou, Y., Wang, Z., Fang, C., Bui, T. and Berg, T.L., 2017. Visual to Sound: Generating Natural Sound for Videos in the Wild. *arXiv 2017*
Mehri, S., Kumar, K., Gulrajani, I., Kumar, R., Jain, S., Sotelo, J., Courville, A. and Bengio, Y., 2016. SampleRNN: An unconditional end-to-end neural audio generation model. *ICLR 2017*

Contents

- Feature learning
- Cross-modal retrieval
- Sound source localization
- Sonorization

Questions?