Objects that sound

unsupervised localization of sources of sounds in images trained from videos

Radek Bartyzal

Let's talk ML in Prague

Date TBA

Achievements

Achievements:

- networks that can embed audio and visual inputs into a common space that is suitable for cross-modal retrieval
- network that can localize the object that sounds in an image, given the audio signal
- training from unlabelled video using only audio-visual correspondence (AVC) as the objective function.

Cross-modal retrieval

Use audio to search in an image.

Self-supervision

Labels are constructed directly from data.

Audio-visual correspondence (AVC)

Input: Pair of a video frame and 1 second of audio represented as a log-spectrogram.

Task: Are they in correspondence or not?

Labels: Obtained directly from video for both positives (matching) and negatives (mismatched) pairs.

Learnt visual and audio representations are:

- discriminative = distinguish matched and mismatched pairs
- semantically meaningful = network has to find semantical match between audio and an image (visual network has only 1 image as input)

Dataset

- publicly available AudioSet dataset
- 10 second clips from YouTube
- authors filtered it for musical instruments, singing and tools, yielding 110 audio classes
- train set = 263k clips
- validation set = 30k clips
- test set = 4.3k clips
- labels are only used for quantitative evaluation purposes of cross-modal retrieval

Previous work - L^3

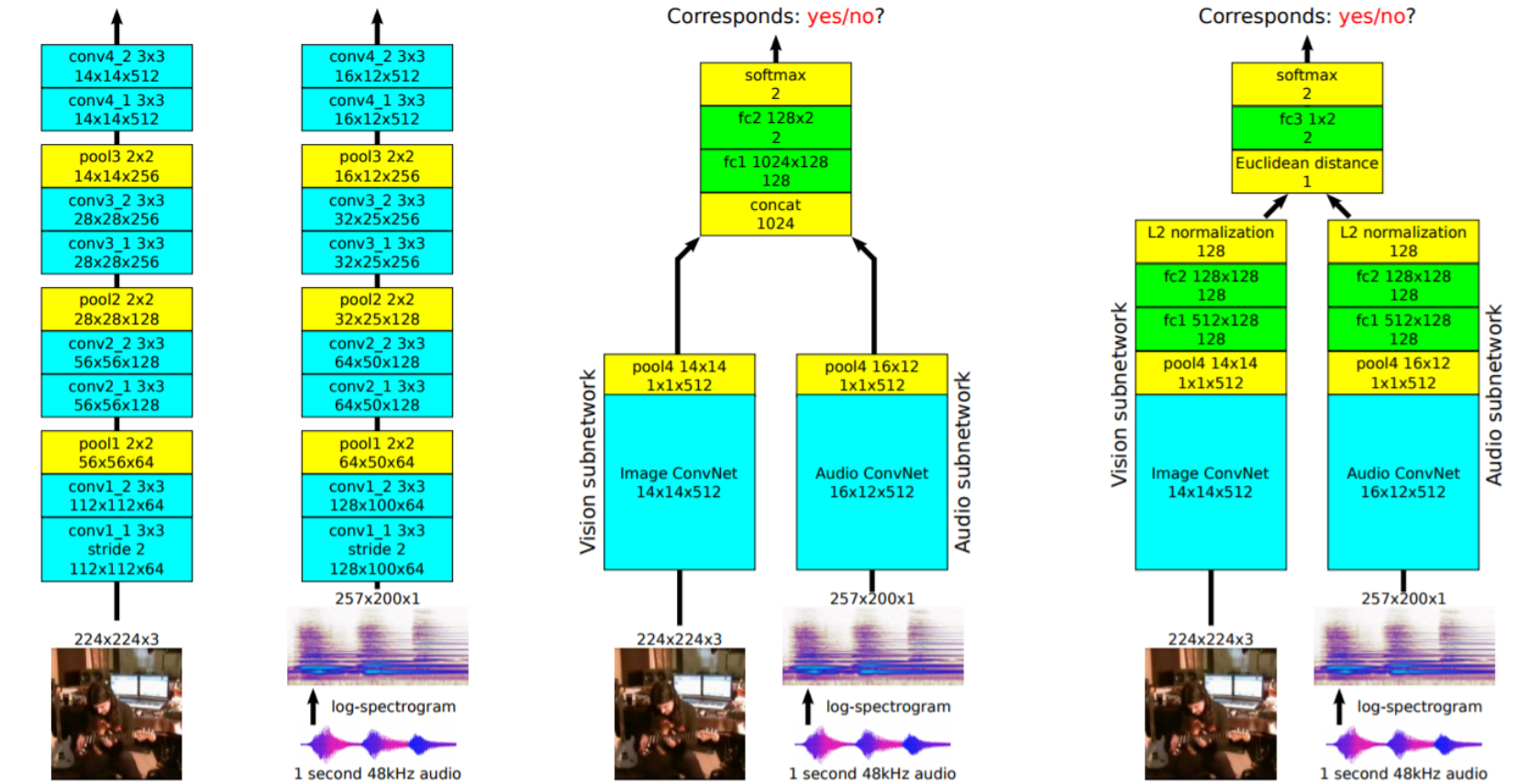
L^3 = Look, Listen and Learn [2]

1. Get embeddings by separate specialized sub-networks.
2. Concatenated embeddings go to fully connected layers that calculate the correspondence score.

Limitations:

- visual and audio representations are not aligned so they cannot be used for crossmodal retrieval - fixed by AVE-Net
- cannot localize the sound source in an image - solved by AVOL-Net

Previous work - L^3



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

Figure: L^3 and AVE-Net descriptions.

Audio-Visual Embedding Network (AVE-Net)

Changes compared to L^3:

1. fully connected layers are moved to the sub-networks
2. L2 normalization on top of them
3. 128-D embedding out of each sub-network
4. euclidean distance of the normalized embeddings
5. 1 FC layer followed by softmax

The tiny FC layer scales and shifts the distance to calibrate it for the subsequent softmax. The bias of the FC essentially learns the threshold on the distance above which the two features are deemed not to correspond.

Audio-Visual Embedding Network (AVE-Net)

- The single value that summarizes whether the image and the audio correspond, forces the two embeddings to be aligned.
- Use of the distance during training makes the features "aware" of the distance metric, therefore making them amenable to retrieval.
- Entire network trained from scratch.
- Parameter-free euclidean distance.

Results

normalized discounted cumulative gain (nDCG)

Measures quality of the ranked list of the top k retrieved items.

Each item in the test set is used as a query and the average nDCG@30 is reported.

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**

Figure: "Our AVENet beats all baselines convincingly." [1] Ehm...

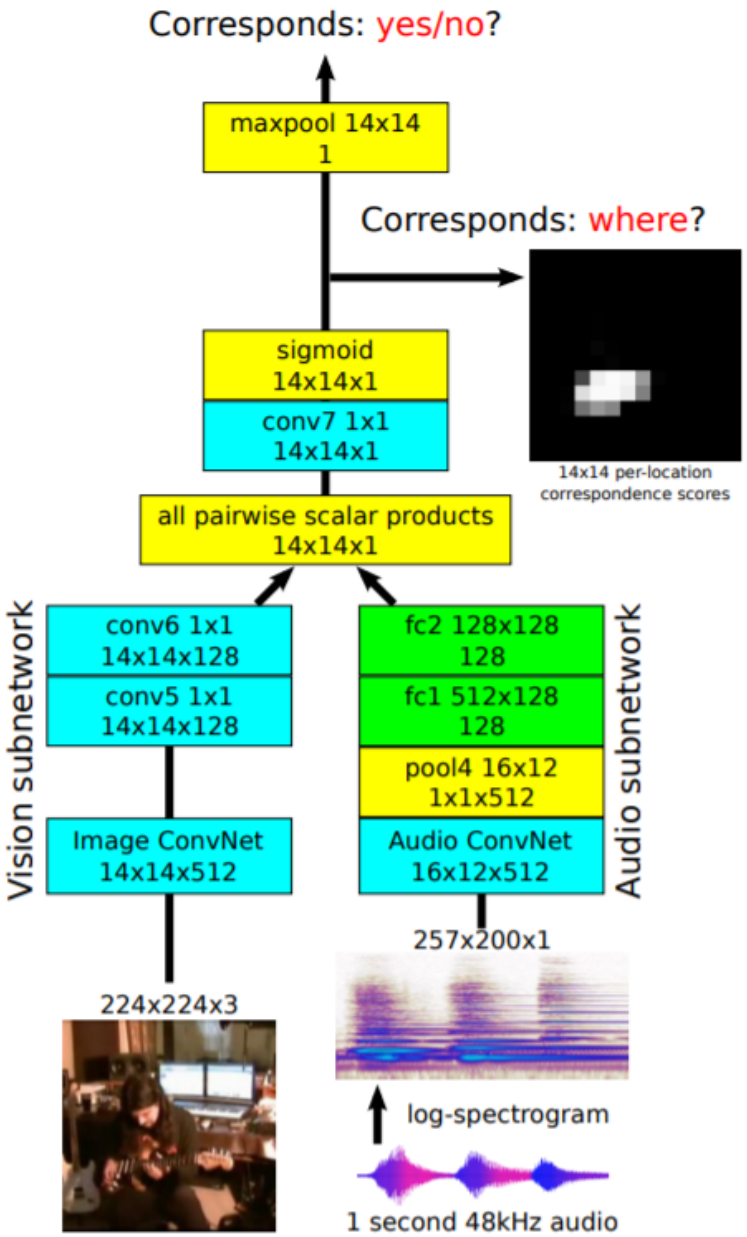
Extending the AVE-Net

Extensions:

- use 25 frames instead of 1
- add 10s of optical flow to the 1 frame

Both had better performance on AVC task: 85% vs AVE-Net 82% vs L^3 82% but it does not translate into better crossmodal retrieval scores.

Audio-Visual Object Localization (AVOL-Net)



Audio-Visual Object Localization (AVOL-Net)

1. 14x14 grid of local 128-D visual embeddings
2. 14x14 distances to the single 128-D audio embedding
3. calibration FC layer turned into a 1x1 Fully Convolutional layer
4. convolutional softmax producing image-audio correspondence scores
5. max correspondence score is used for the training by AVC task

Results

Figure: First row - correctly localized both flute and piano sound.

Sampling of negative pairs

- Positive pair is a frame and one second of audio with the frame in the middle.
- Videos have 25fps.
- Positive audio samples start at multiples of 0.04s.
- Getting a negative audio pair randomly allows the network to cheat. It recognizes that negative samples do not start at multiples of 0.04.
- Cheating allowed 87.6% vs correct 81.9% at AVC task.
- Solution = sample negative audio samples also at the multiples of 0.04.

Sources

1. Arandjelović, Relja, and Andrew Zisserman. "Objects that Sound." arXiv preprint arXiv:1712.06651 (2017). Accessible from: https://arxiv.org/abs/1712.06651

2. Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017. Accessible from: https://arxiv.org/abs/1705.08168