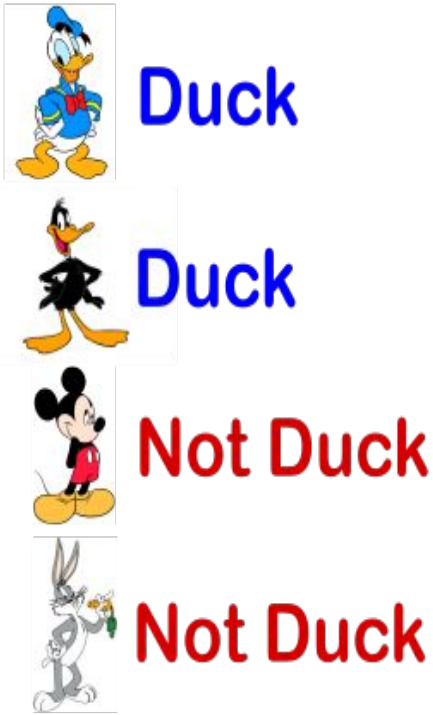
Look, Listen and Learn

Relja Arandjelović, Andrew Zisserman

February 20th, 2018

Greg Rehm, Shahbaz Rezaei

Supervised Learning



Supervised Learning

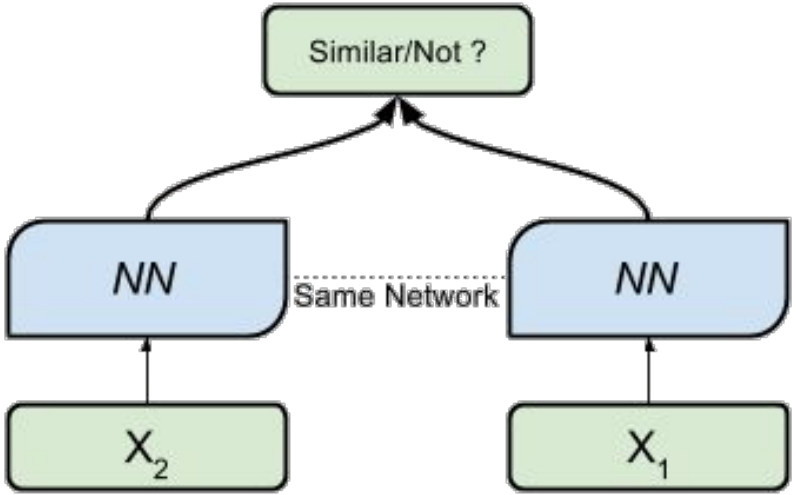
Predictive Model


Vegetation
Street Light
Building
Pole
Street Name/Address – Front
Street Name/Address – Front
Street Light
Street Light
Traffic Sign – Front
Street Light
Traffic Light – Upright (General Traffic)
Traffic Light – Side
Sky
Street Light
Vegetation
Sidewalk
Person
Trash Can
Curb Cut
Lane Marking – Crosswalk
Lane Marking – Driveable
Road
Lane Marking – Driveable
Lane Marking – Others (Arrows/Bicycles Etc.)
Lane Marking – Others (Arrows/Bicycles Etc.)
Lane Marking – Crosswalk
Wire Group
Sky
Pole
Sky
Sky
Traffic Light – Side
Traffic Sign – Back
Street Name/Address – Front
Traffic Light – Upright (General Traffic)
Building
Traffic Sign – Front
Vehicle Group
Person Group
Person
Sidewalk
Road
Lane Marking – Crosswalk
Bicycle
Lane Marking – Crosswalk
Lane Marking

Self-Supervised Learning

- Uses unlabeled data for training
- Do not need annotators to perform extensive ground truth labeling
- Saves time and $$
- How do you perform it in practice for CNN’s though?

Siamese Networks

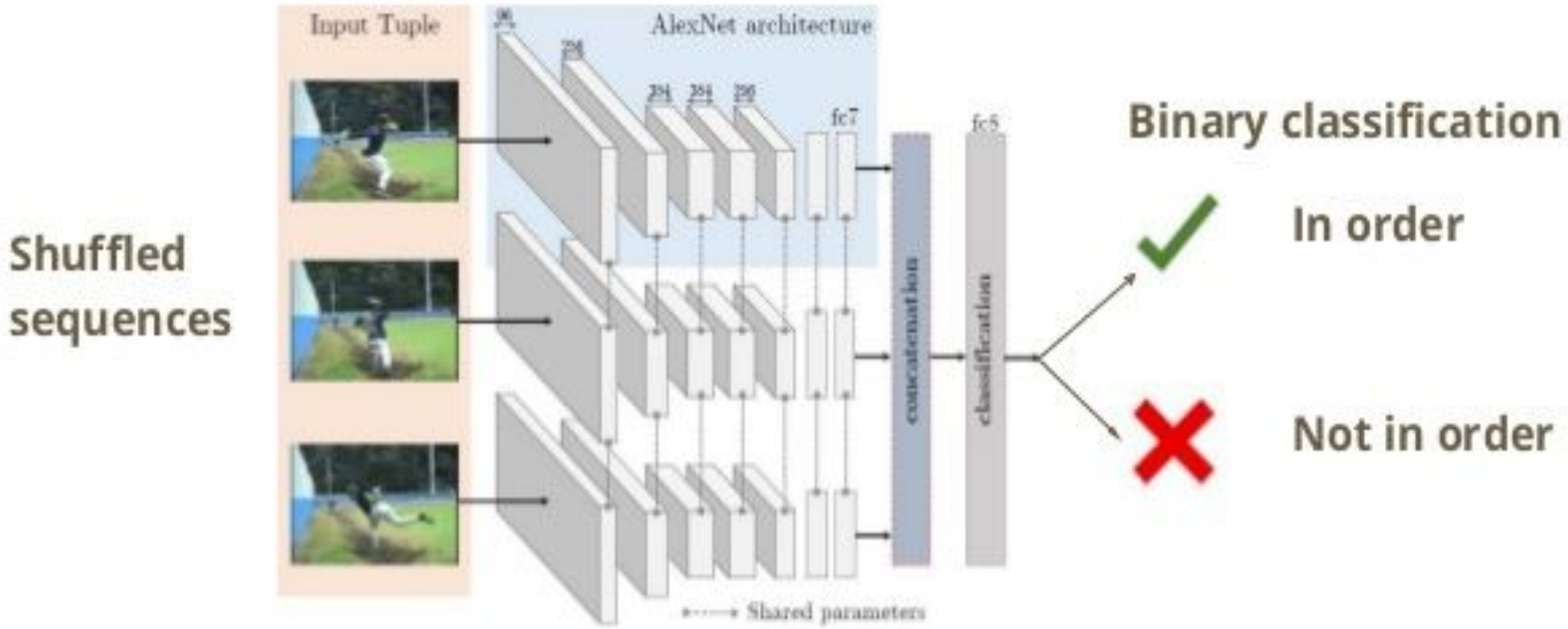


The **Distance Function** decides if the output vectors are close enough to be similar

The **Neural Network** transforms the input into a properties vector

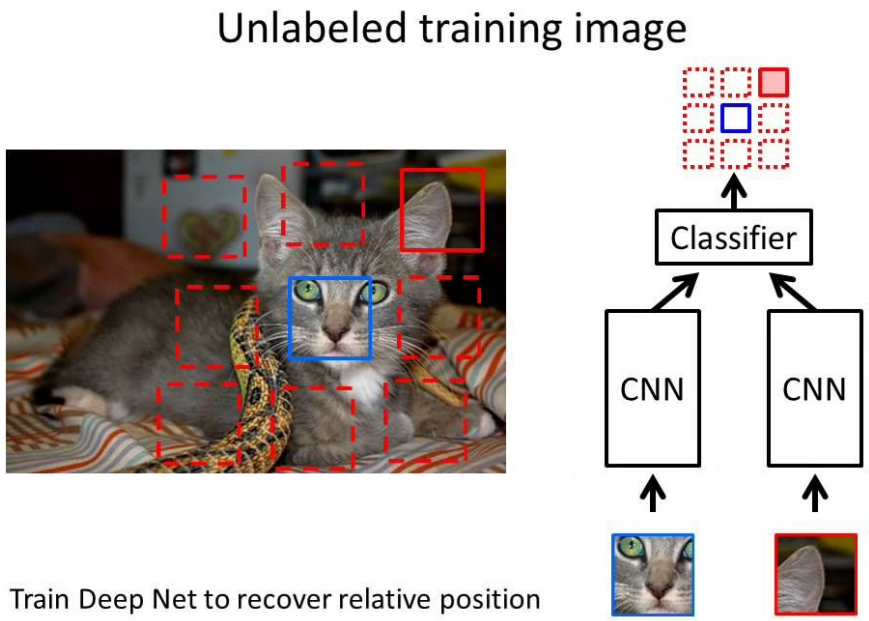
Input Data (image, text, features…)

Can Learn Temporal Ordering



Misra, “Shuffle and Learn: Unsupervised Learning Using Temporal Order Verification.” 2016. Vol. 9905

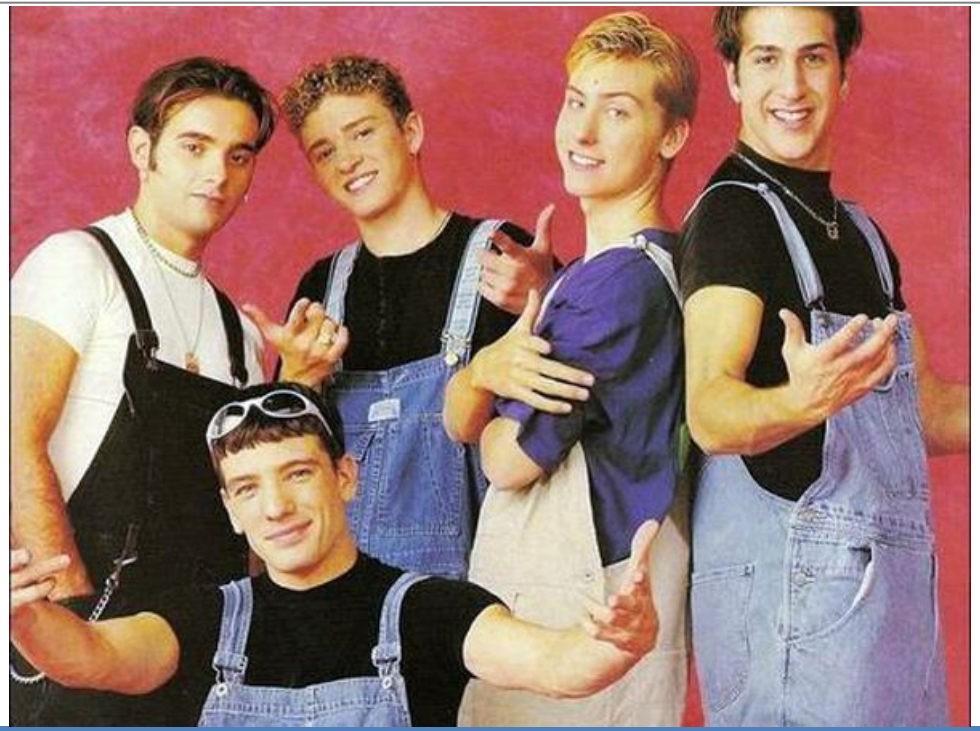
Image Location Prediction



Doersch “Unsupervised Visual Representation Learning by Context Prediction.” (2015):

Can a Self-Supervised Machine Learn what a Boy Band is?

Can a Self-Supervised Machine Learn what a Boy Band is?



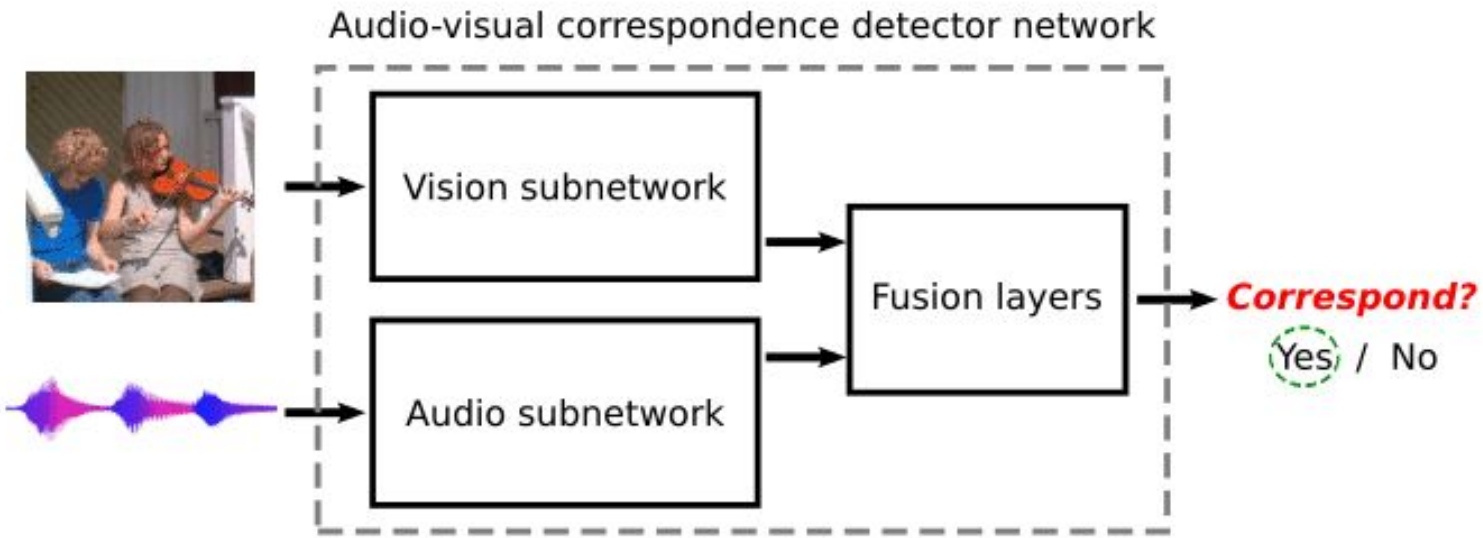
Machine Might Learn This:

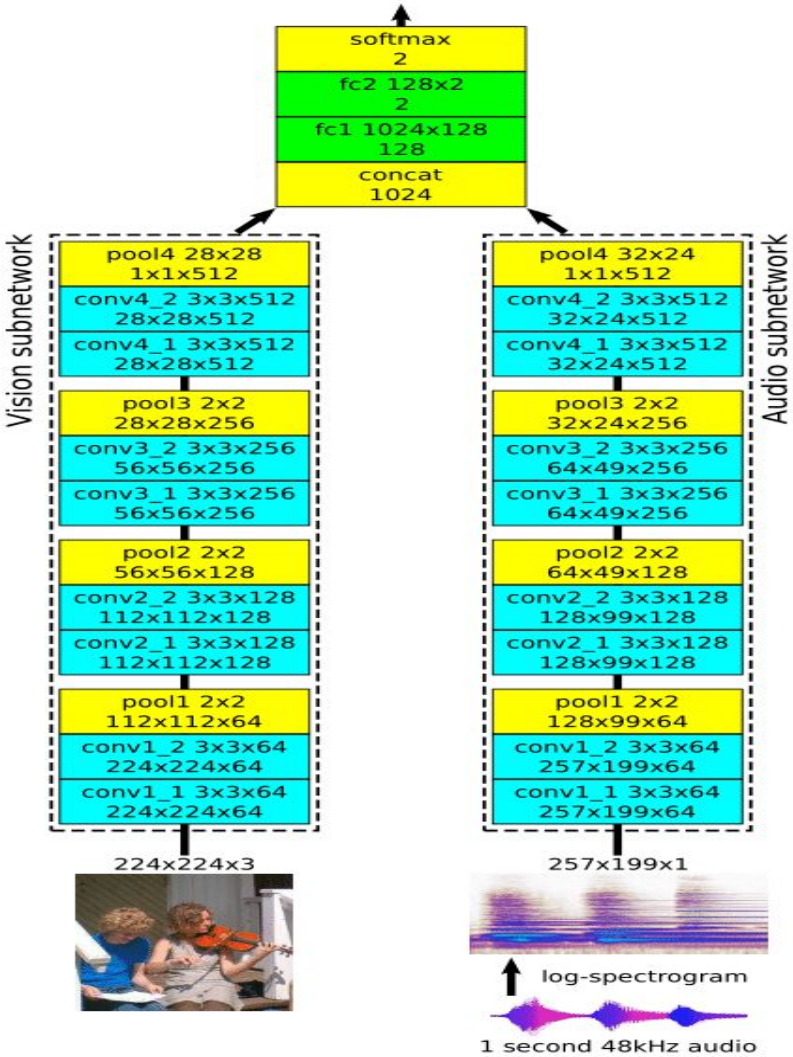
- >4 young males in row
- Silly hair
- Dumb look on face

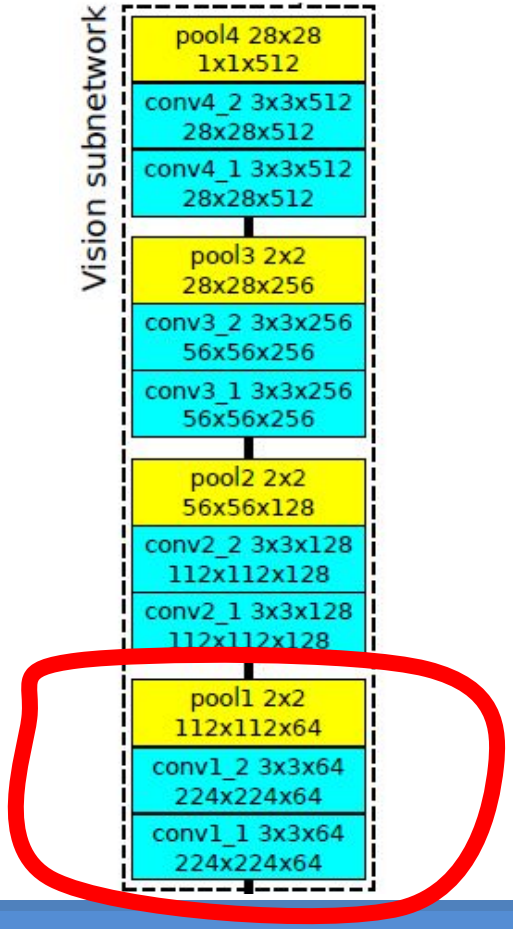
False Positives Abound

Solution: Add Audio

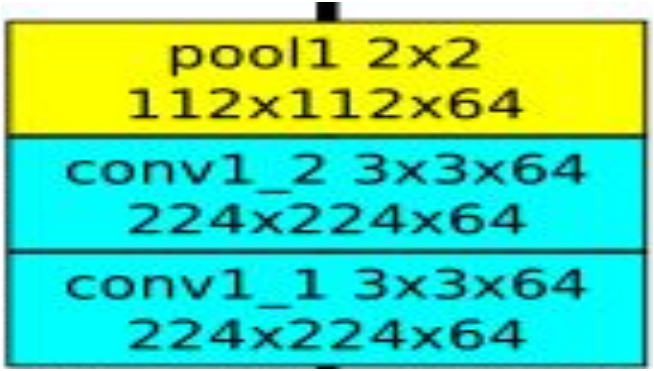
The Main Idea

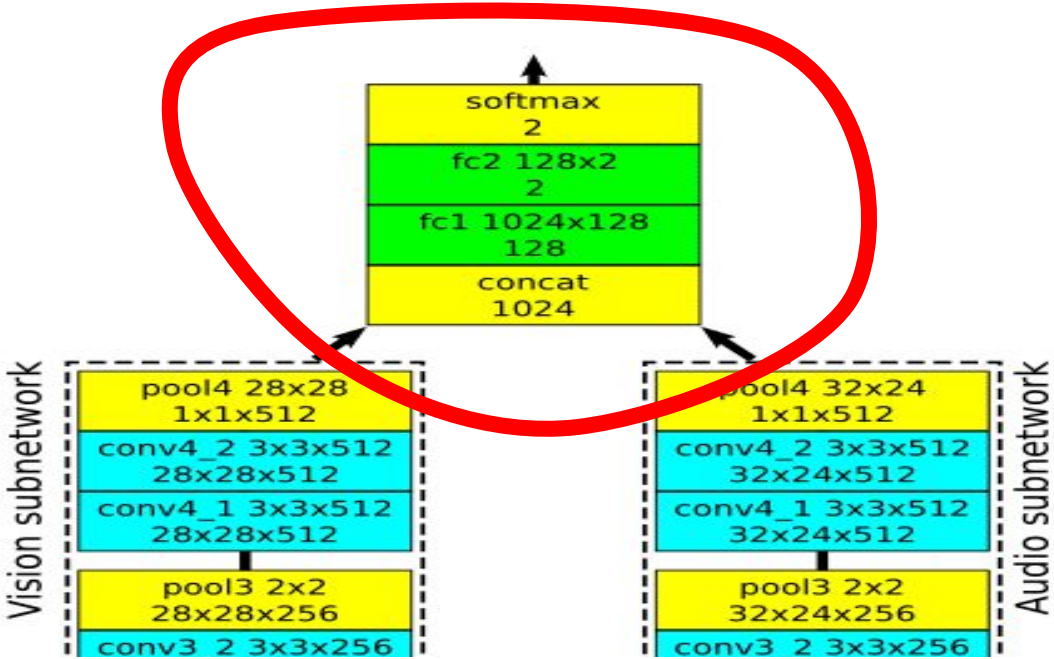



softmax
2
fc2 128x2
2
fc1 1024x128
128
concat
1024
Vision subnetwork
pool4 28x28
1x1x512
conv4_2 3x3x512
28x28x512
conv4_1 3x3x512
28x28x512
pool3 2x2
28x28x256
conv3_2 3x3x256
56x56x256
conv3_1 3x3x256
56x56x256
pool2 2x2
56x56x128
conv2_2 3x3x128
112x112x128
conv2_1 3x3x128
112x112x128
pool1 2x2
112x112x64
conv1_2 3x3x64
224x224x64
conv1_1 3x3x64
224x224x64
224x224x3
Audio subnetwork
pool4 32x24
1x1x512
conv4_2 3x3x512
32x24x512
conv4_1 3x3x512
32x24x512
pool3 2x2
32x24x256
conv3_2 3x3x256
64x49x256
conv3_1 3x3x256
64x49x256
pool2 2x2
64x49x128
conv2_2 3x3x128
128x99x128
conv2_1 3x3x128
128x99x128
pool1 2x2
128x99x64
conv1_2 3x3x64
257x199x64
conv1_1 3x3x64
257x199x64
257x199x1
log-spectrogram
1 second 48kHz audio

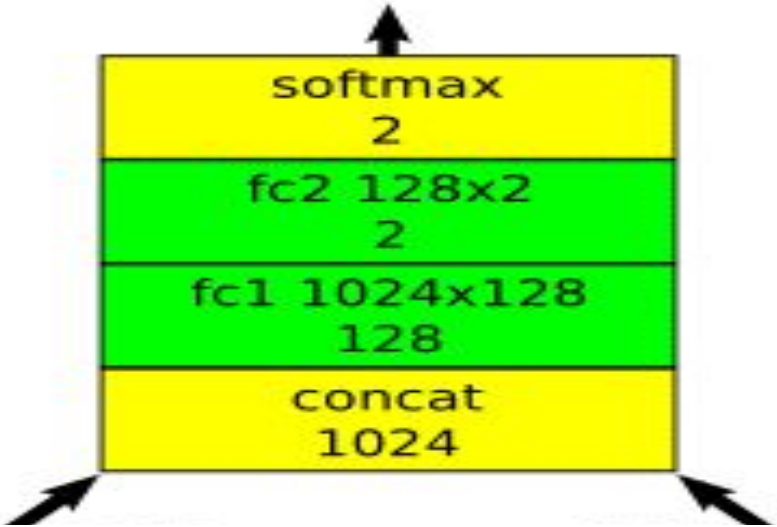

Vision subnetwork
pool4 28x28
1x1x512
conv4_2 3x3x512
28x28x512
conv4_1 3x3x512
28x28x512
pool3 2x2
28x28x256
conv3_2 3x3x256
56x56x256
conv3_1 3x3x256
56x56x256
pool2 2x2
56x56x128
conv2_2 3x3x128
112x112x128
conv2_1 3x3x128
112x112x128
pool1 2x2
112x112x64
conv1_2 3x3x64
224x224x64
conv1_1 3x3x64
224x224x64

Convolution Block (From VGG)




softmax
2
fc2 128x2
2
fc1 1024x128
128
concat
1024
Vision subnetwork
pool4 28x28
1x1x512
conv4_2 3x3x512
28x28x512
conv4_1 3x3x512
28x28x512
pool3 2x2
28x28x256
Audio subnetwork
pool4 32x24
1x1x512
conv4_2 3x3x512
32x24x512
conv4_1 3x3x512
32x24x512
pool3 2x2
32x24x256

Fusion + Classification Layers



Inputs

1 second 48kHz audio

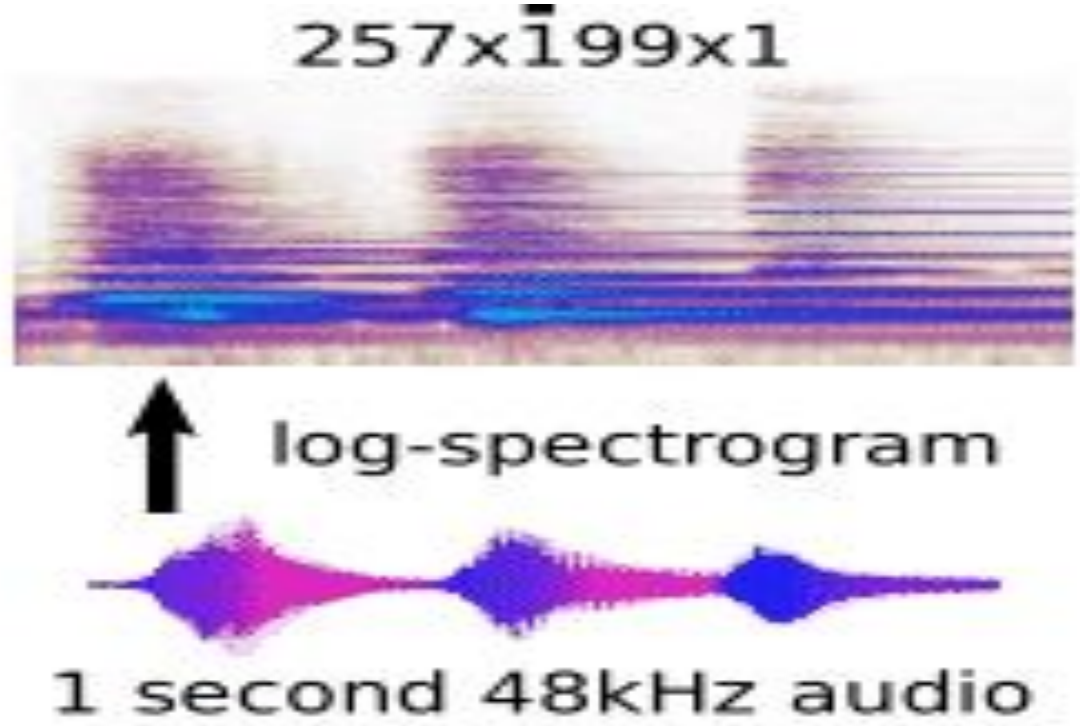
Want to train with overlapping audio and images

And train with non-overlapping audio and images

Image input

224x224x3

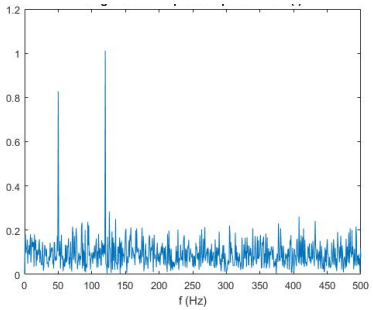
Audio input



How to compute a spectrogram

.01 sec. window

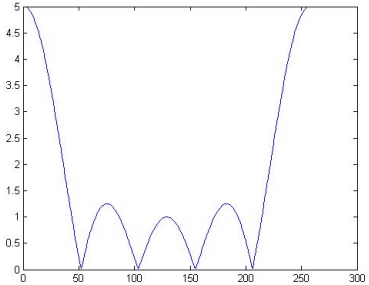
FFT

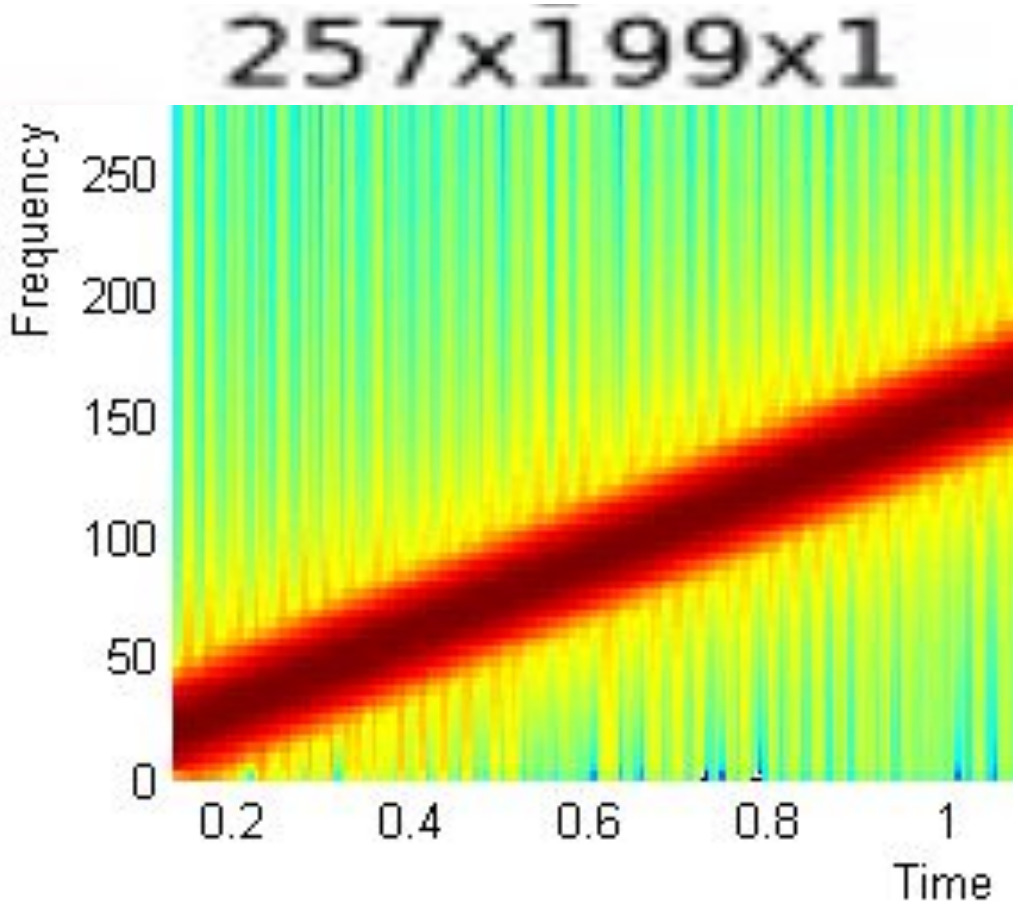


How to compute a spectrogram

Half window overlap

FFT




257x199x1
Frequency
250
200
150
100
50
0
0.2
0.4
0.6
0.8
1
Time

Training Details

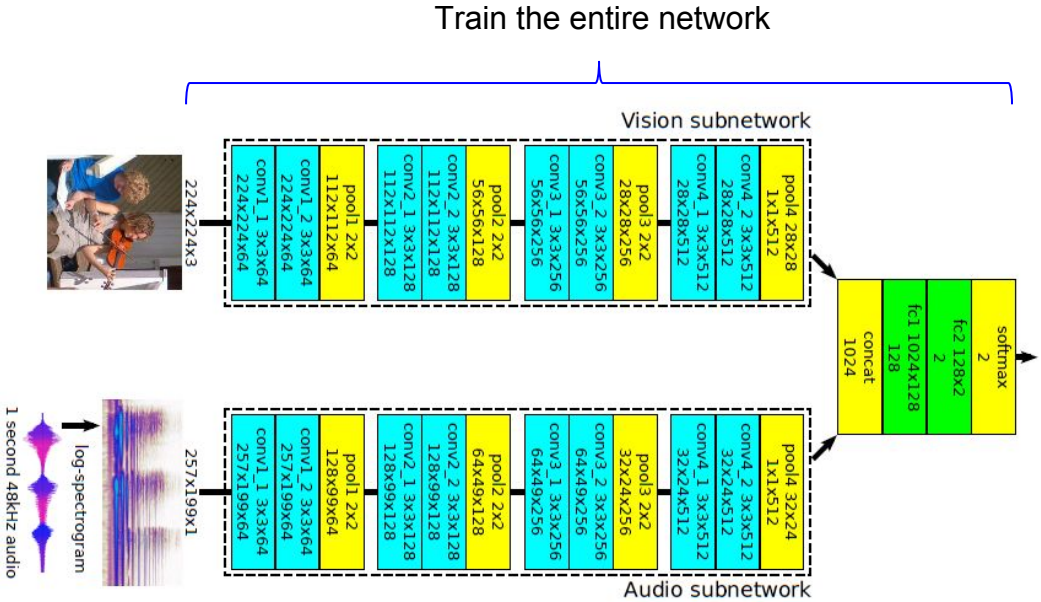
- Uses Adam Optimizer
- Trained with 16 GPUs
- Effective batch size is 256 elements
- Augmentation for images with random cropping/flipping/brightness/saturation changes
- Augmentation for audio with random volume changes of 10%
- Sees **60 million** frame-audio pairs

Datasets

- Flickr-SoundNet
  - Unlabelled dataset of Flickr video
  - Contains over 2 million videos
  - A random subset of 500K was used
- Kinetics-Sounds
  - Manually labelled dataset containing 34 classes
    - Playing instruments (guitar, violin, etc.)
    - Using tools (lawn mowing, shoveling snow, etc.)
    - Performing actions (tap dancing, bowling)
  - Clean, but contains noises:
    - Load background music
    - Human voice
    - Videos with soundtrack

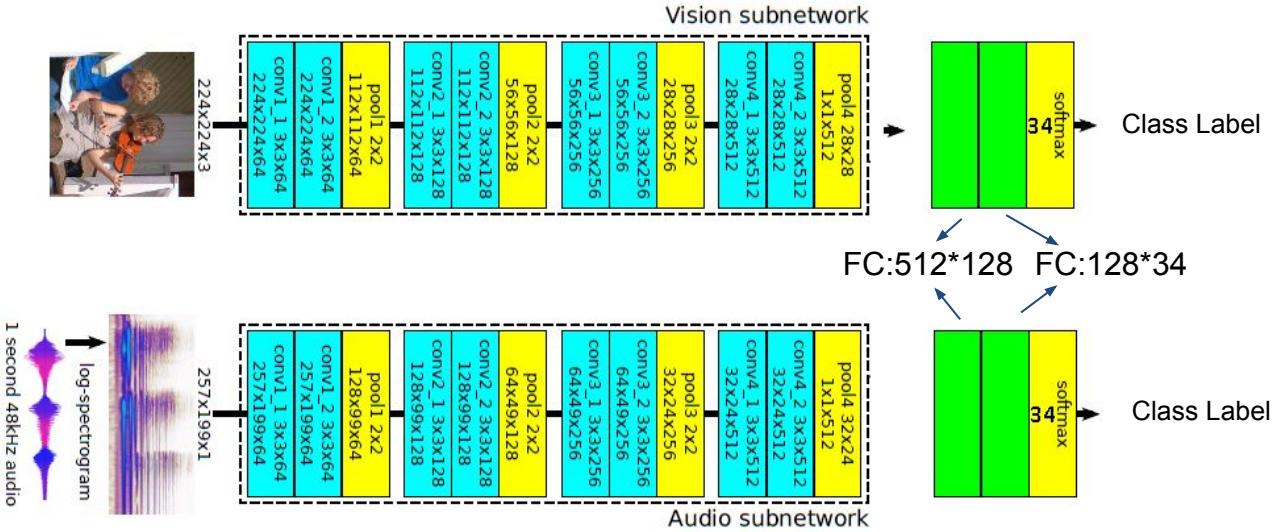
3.2 Audio Visual Correspondence

L3-Net: Trian the whole network in unsupervised manner.



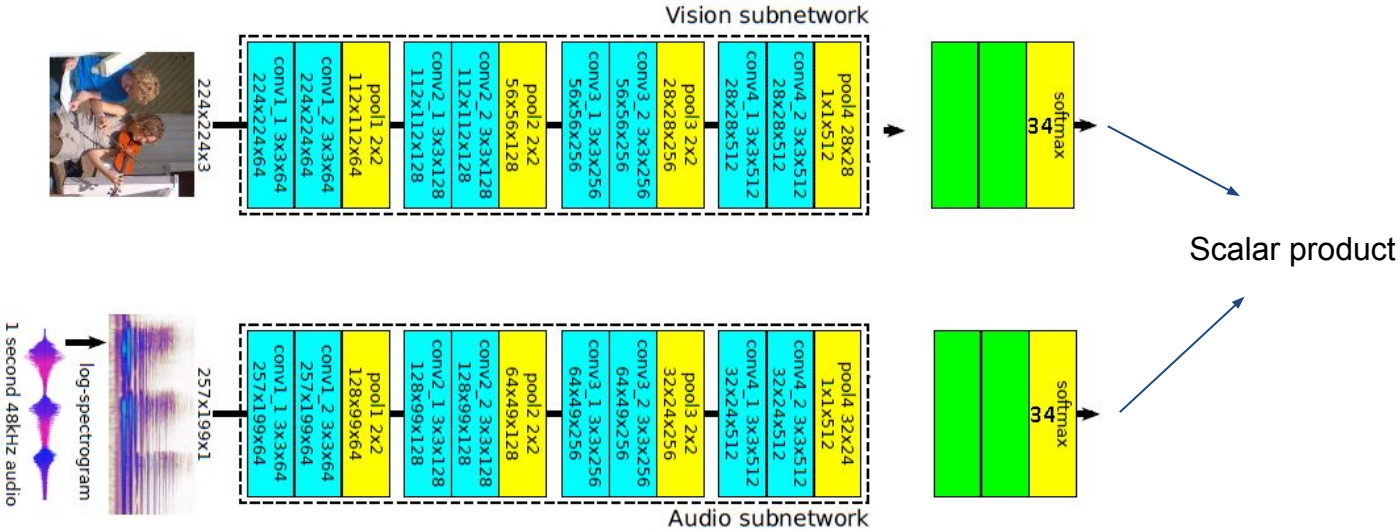
3.2 Audio-Visual Correspondence (AVC)

Supervised Direct (**Training**): Training two networks independently using Kinetics-Sounds (34 classes).



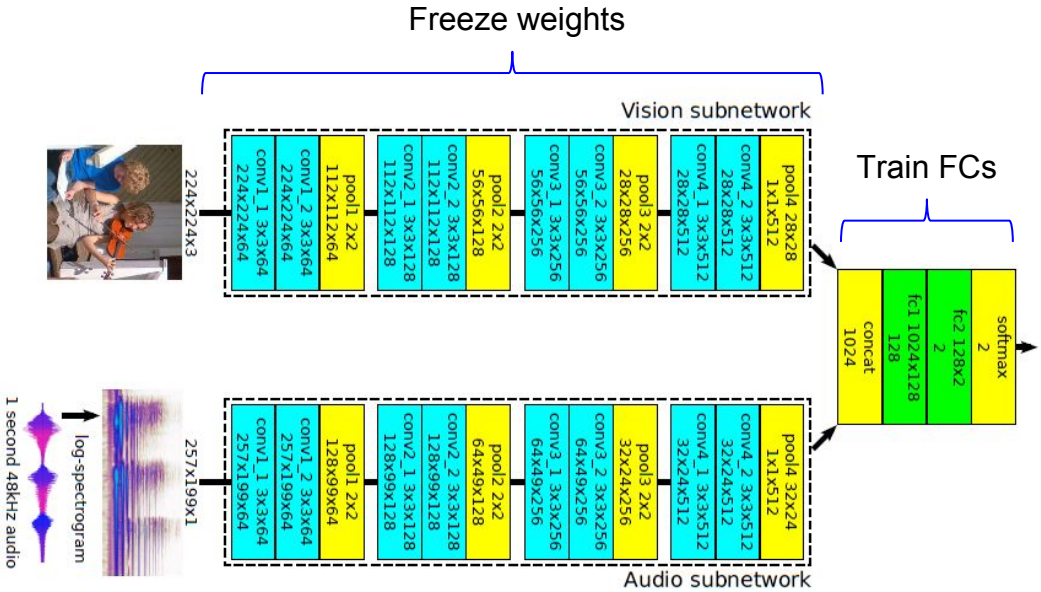
3.2 Audio Visual Correspondence

Supervised Direct (**Testing**): Scalar product of two networks' softmax output.



3.2 Audio Visual Correspondence

Supervised Pretraining (**Training**): Combine two trained networks, freeze the trained networks' weights, then train the new fully connected layers.

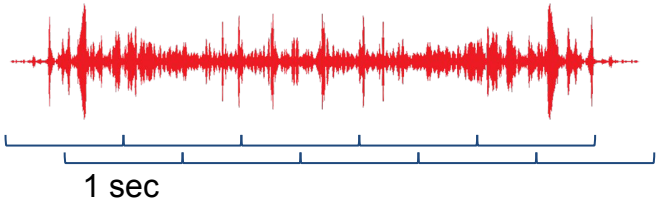


3.2 Audio Visual Correspondence Results

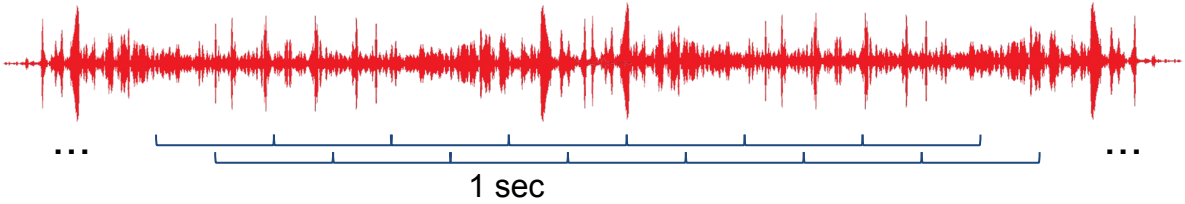
Method	Flickr-SoundNet	Kinetics-Sounds
Supervised direct	–	65%
Supervised pretraining	–	74%
L^3-Net	78%	74%

3.3 Audio Features

Environmental Sound Classification (ECS-50): 2000 audio clips, 5 seconds each, equally balanced between 50 classes. They extracted 10 Equally spaced 1 sec subclips.

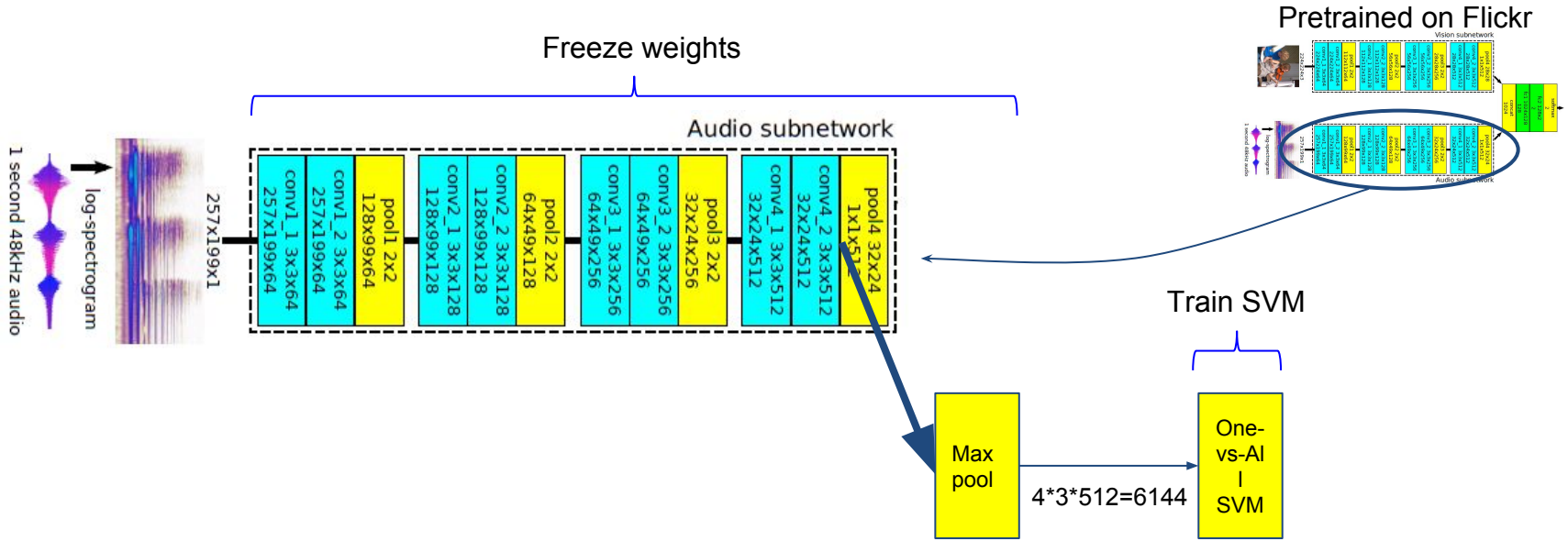


Detection and Classification of Acoustic Scenes and Events(DCASE): 110 audio clips, 30 seconds each, with 10 classes. 60 subclips are extracted



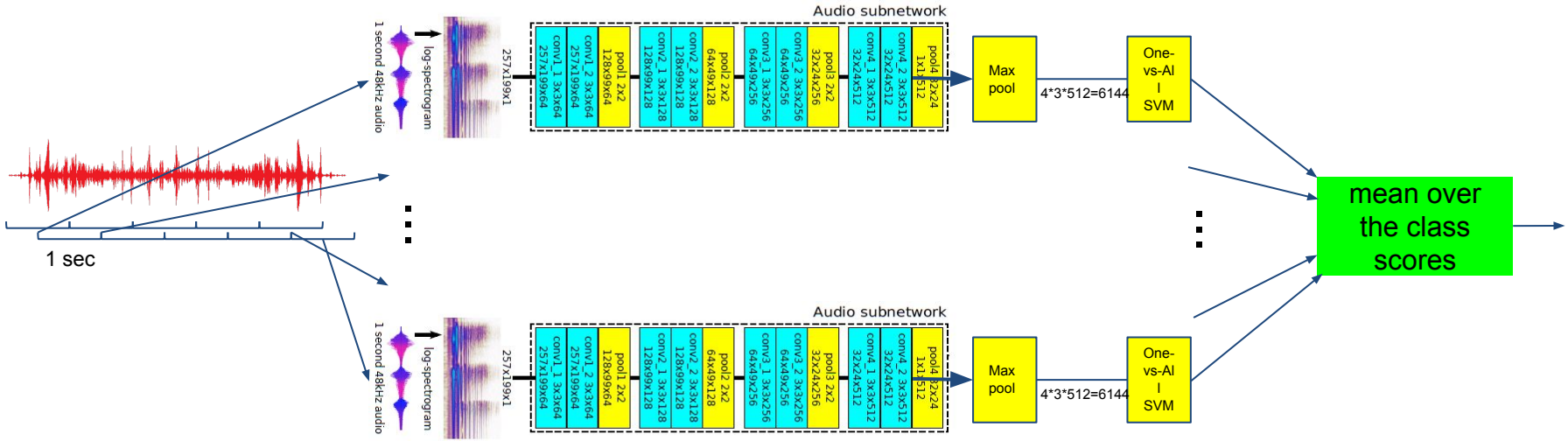
3.3 Audio Features

L3-Net (Training): Audio subnetwork of L3-Net which was pretrained on Flickr-SoundNet is used.



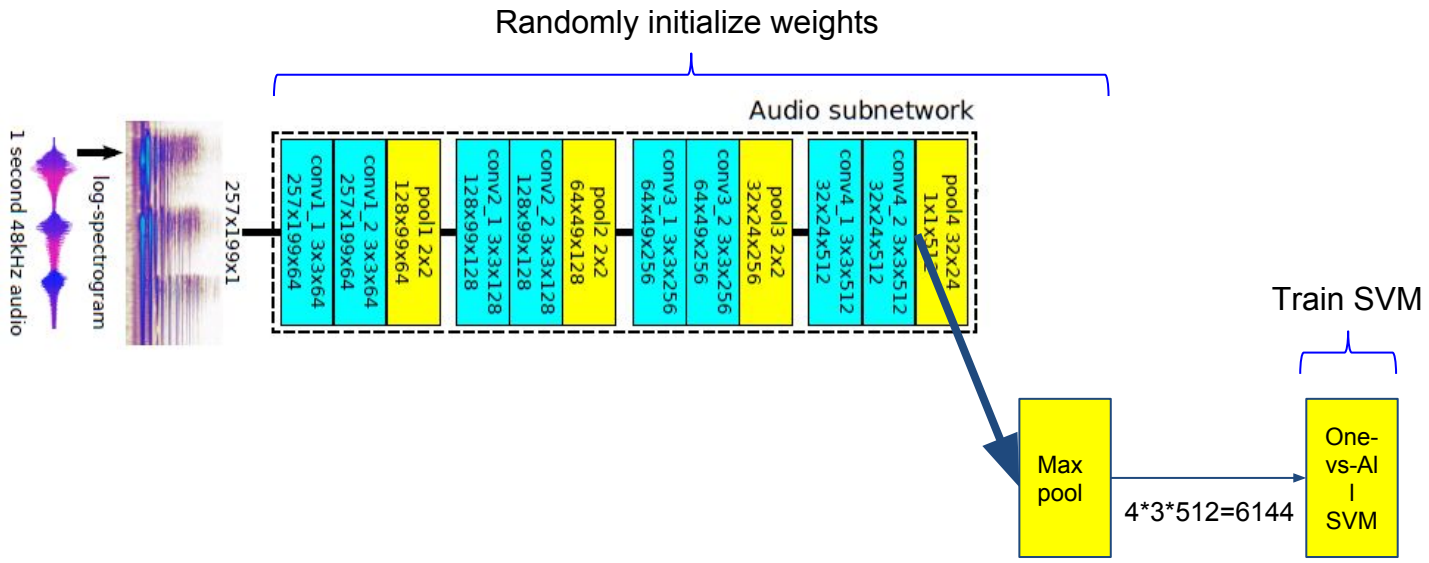
3.3 Audio Features

L3-Net (Testing): Audio subnetwork of L3-Net which was pretrained on Flickr-SoundNet is used.



3.3 Audio Features

L3-Net (random): Audio subnetwork of L3-Net which was initialized randomly is used.



3.3 Audio Features

(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%*

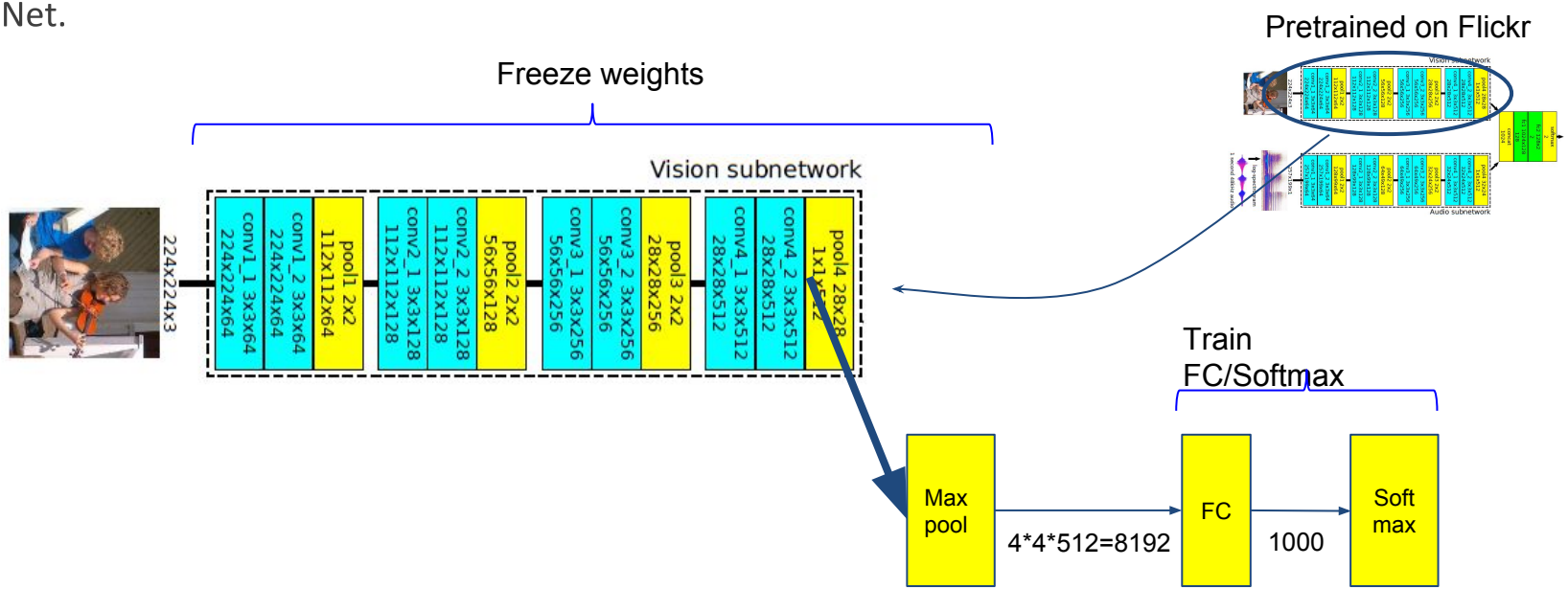
(b) DCASE

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

[Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." ICCV 2017]

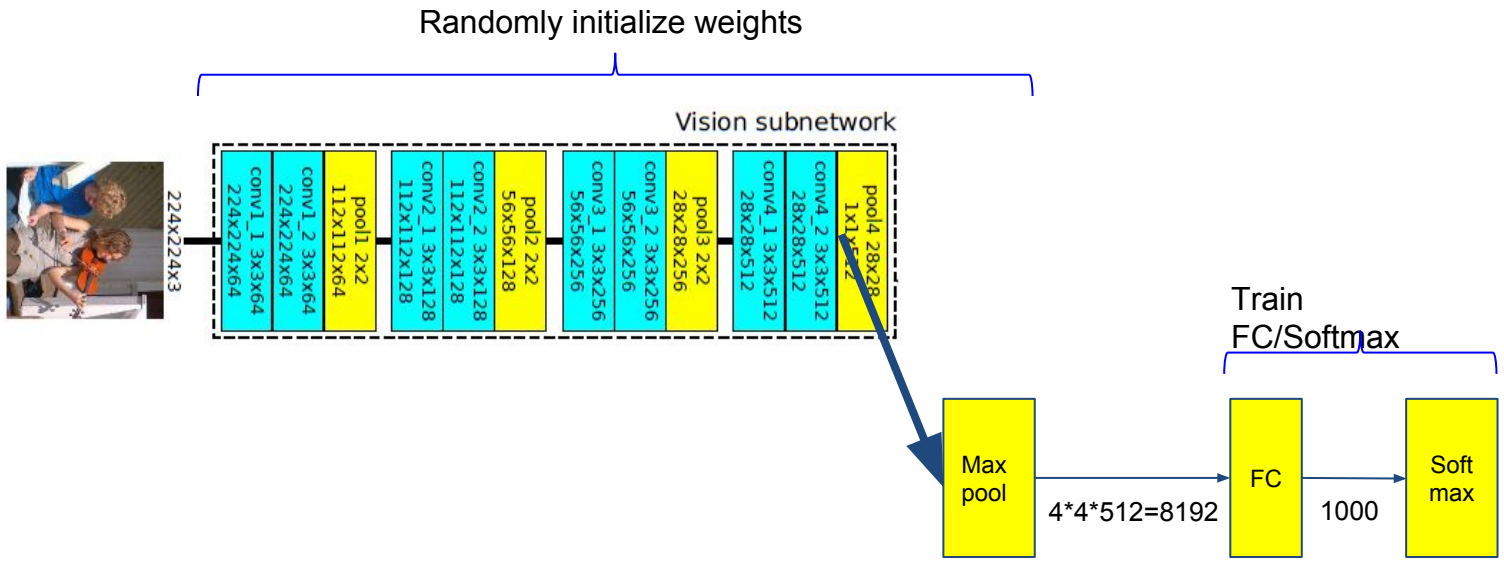
3.4 Visual Features

L3-Net (Training): Visual subnetwork of L3-Net which was pretrained on Flickr-SoundNet is trained of ImageNet.



3.4 Visual Features

L3-Net (random): Visual subnetwork of L3-Net with random weights is trained on ImageNet.

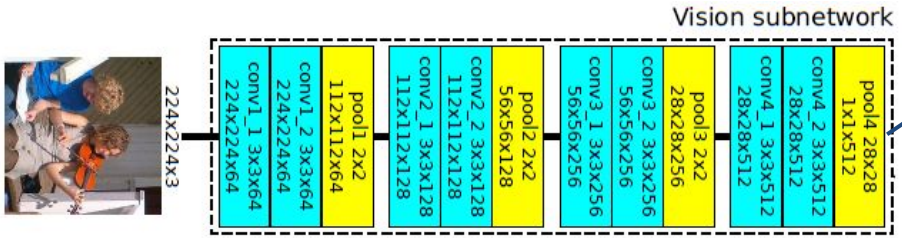


3.4 Visual Features

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%

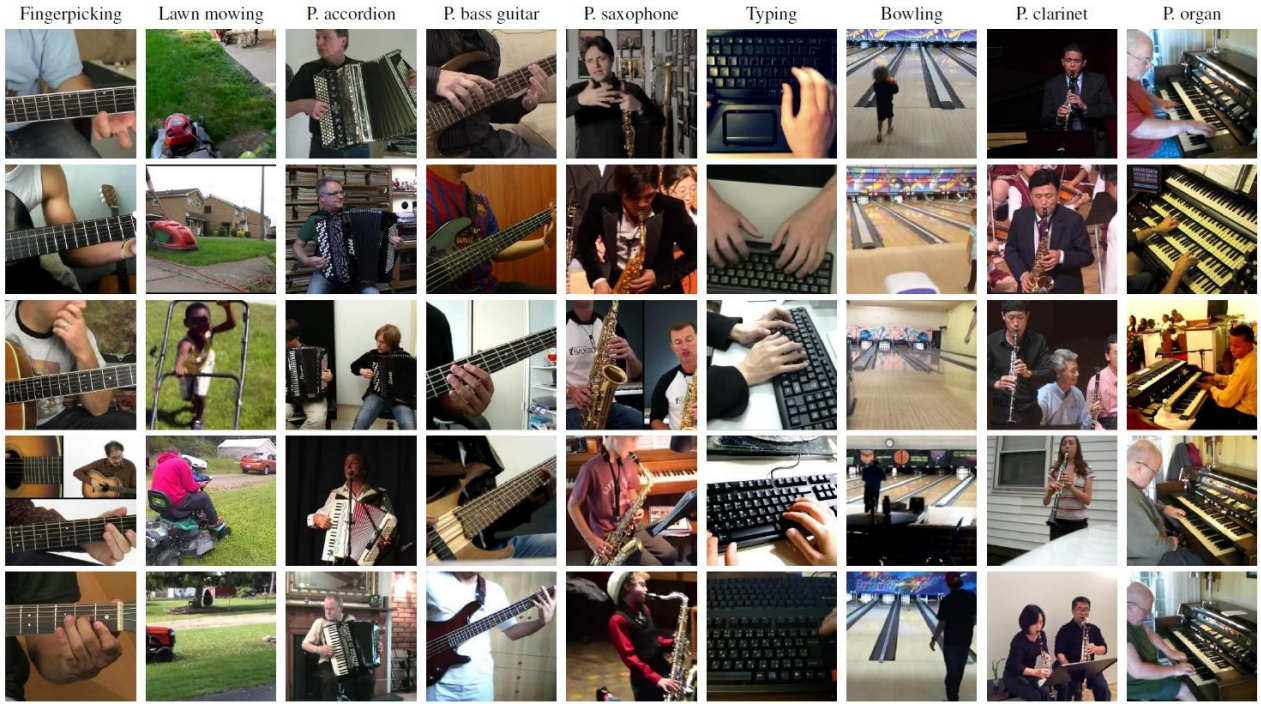
[Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." ICCV 2017]

3.5.1 Qualitative Analysis - Vision Features



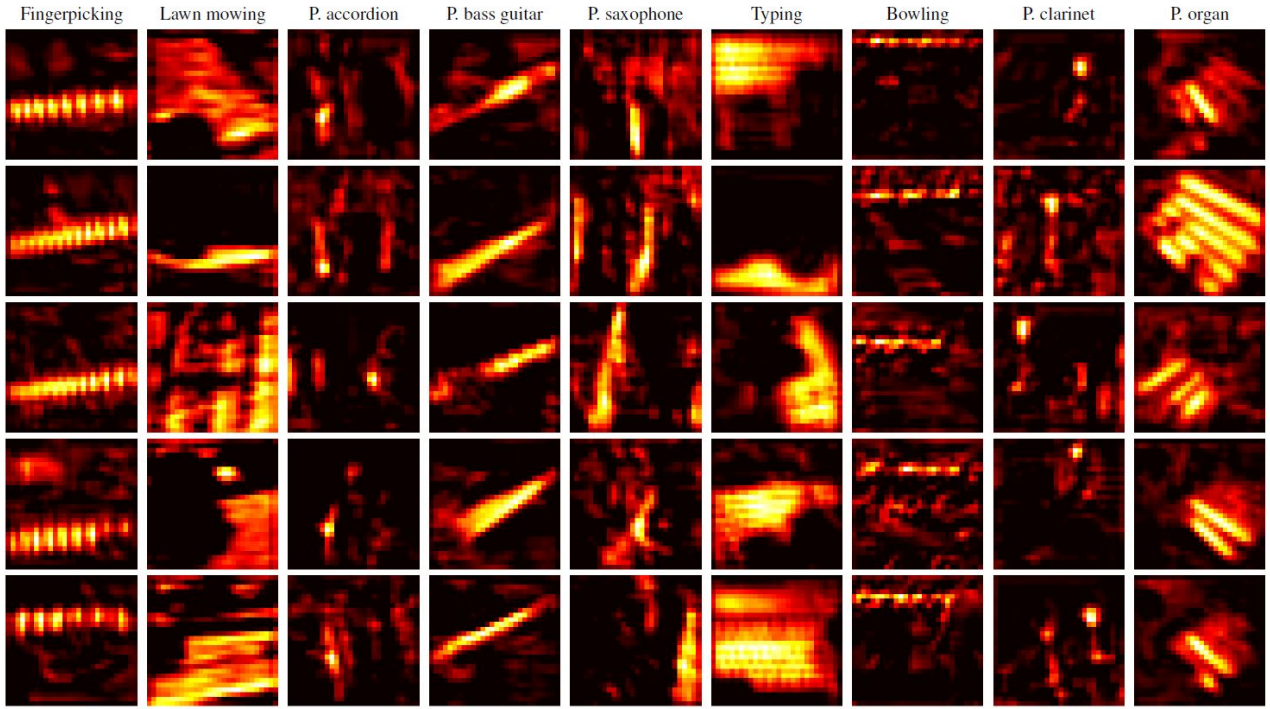
For each particular unit in “Pool4”, test images with highest activation were chosen.

3.5.1 Vision Features on Kinetics-Sounds



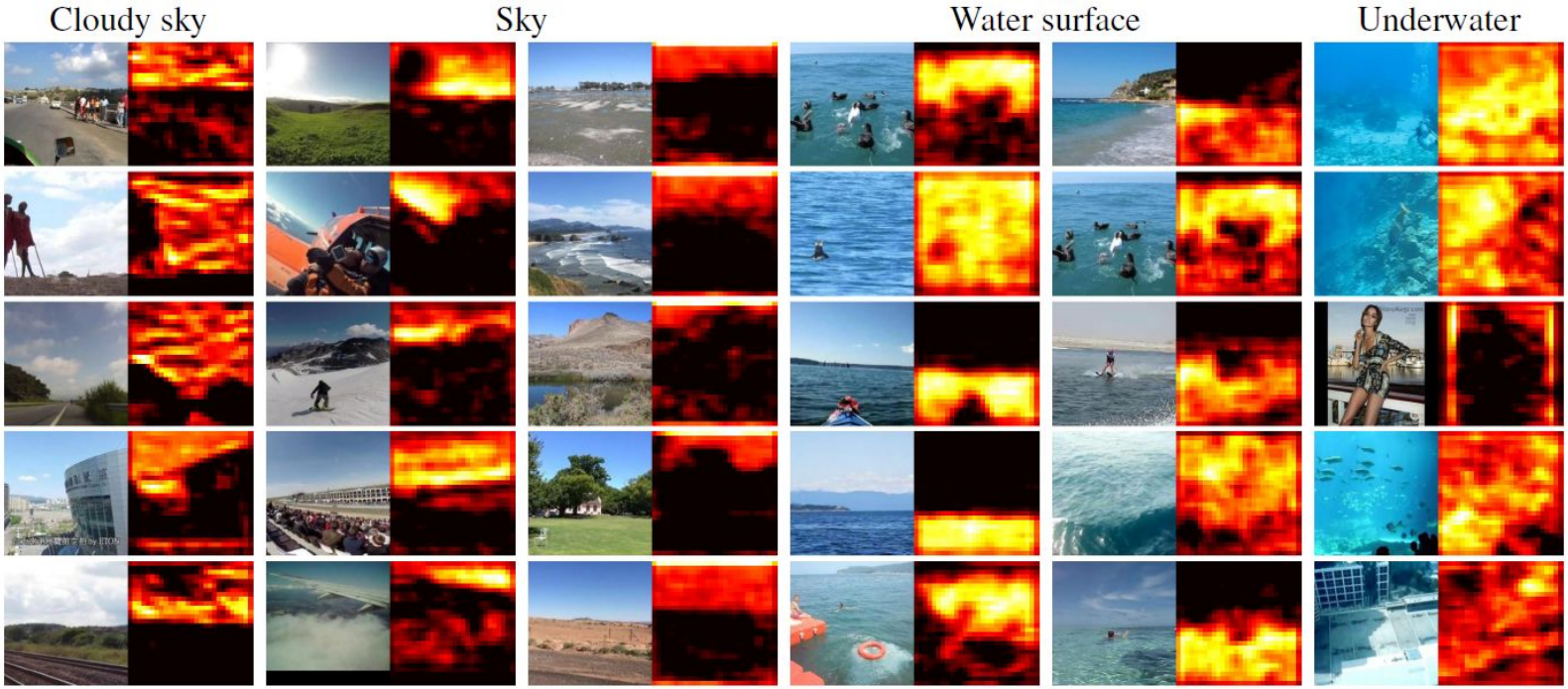
[Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." ICCV 2017]

3.5.1 Vision Features on Kinetics-Sounds



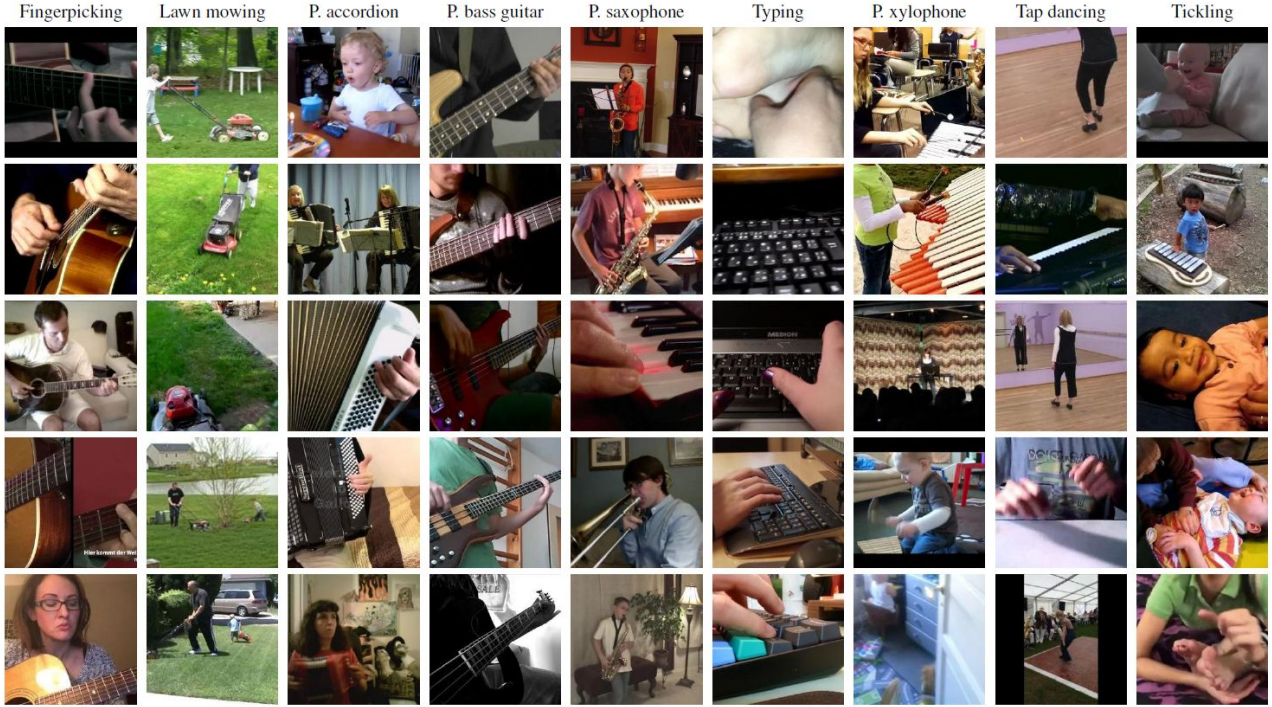
[Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." ICCV 2017]

3.5.1 Vision Features on Flickr-SoundNet



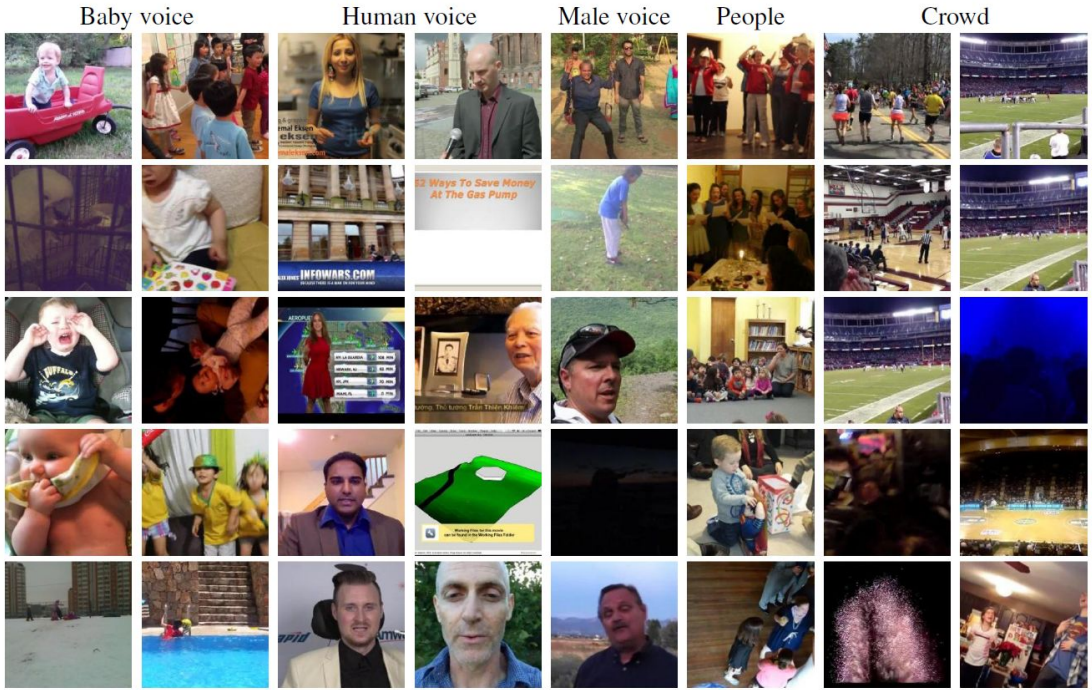
[Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." ICCV 2017]

3.5.1 Audio Features on Kinetics-Sounds



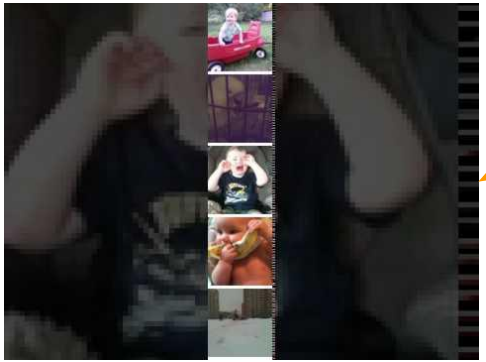
[Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." ICCV 2017]

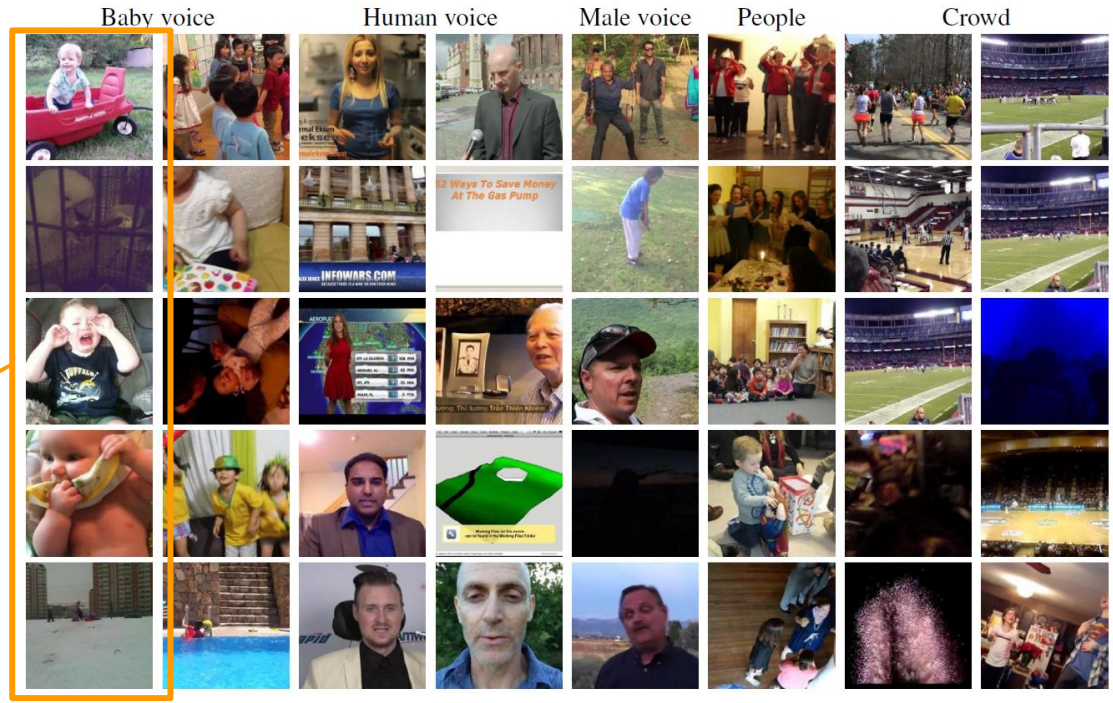
3.5.1 Audio Features on Flickr-SoundNet



[Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." ICCV 2017]

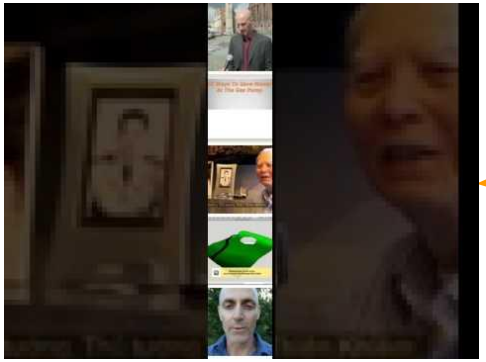
3.5.1 Audio Features on Flickr-SoundNet

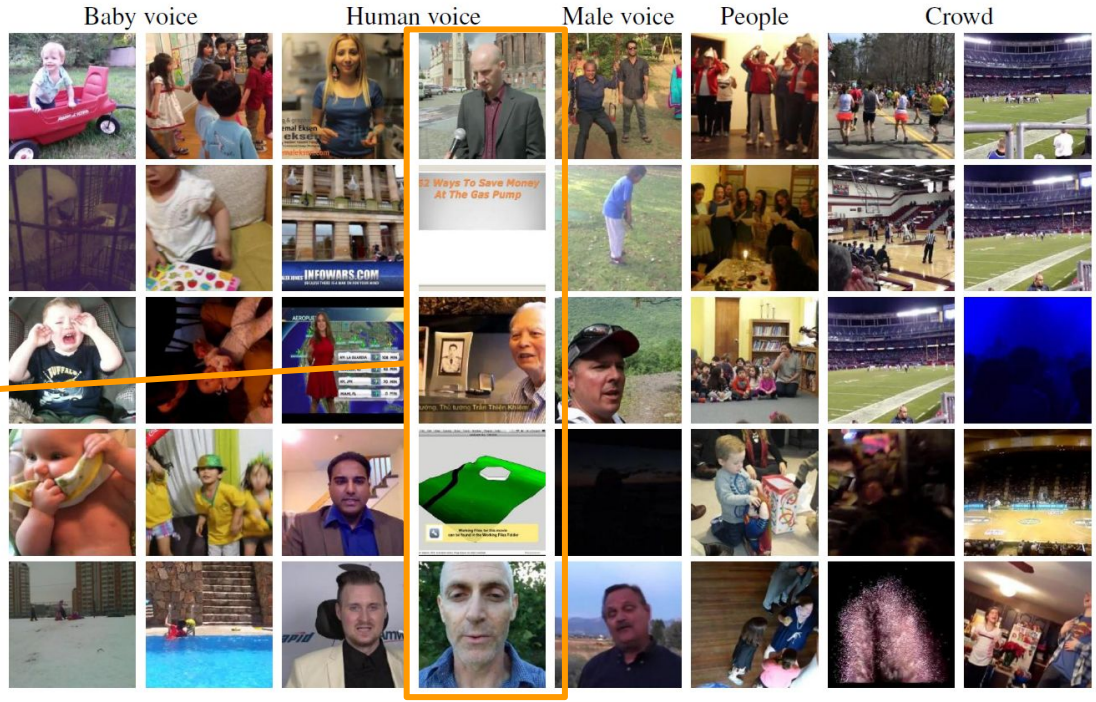




[Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." ICCV 2017]

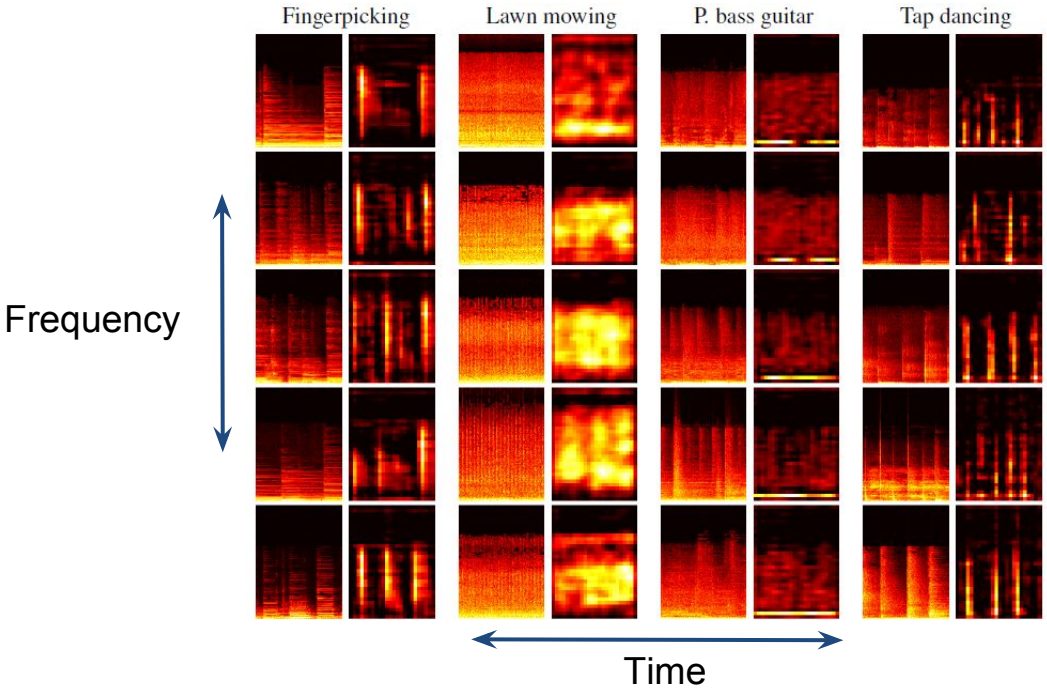
3.5.1 Audio Features on Flickr-SoundNet





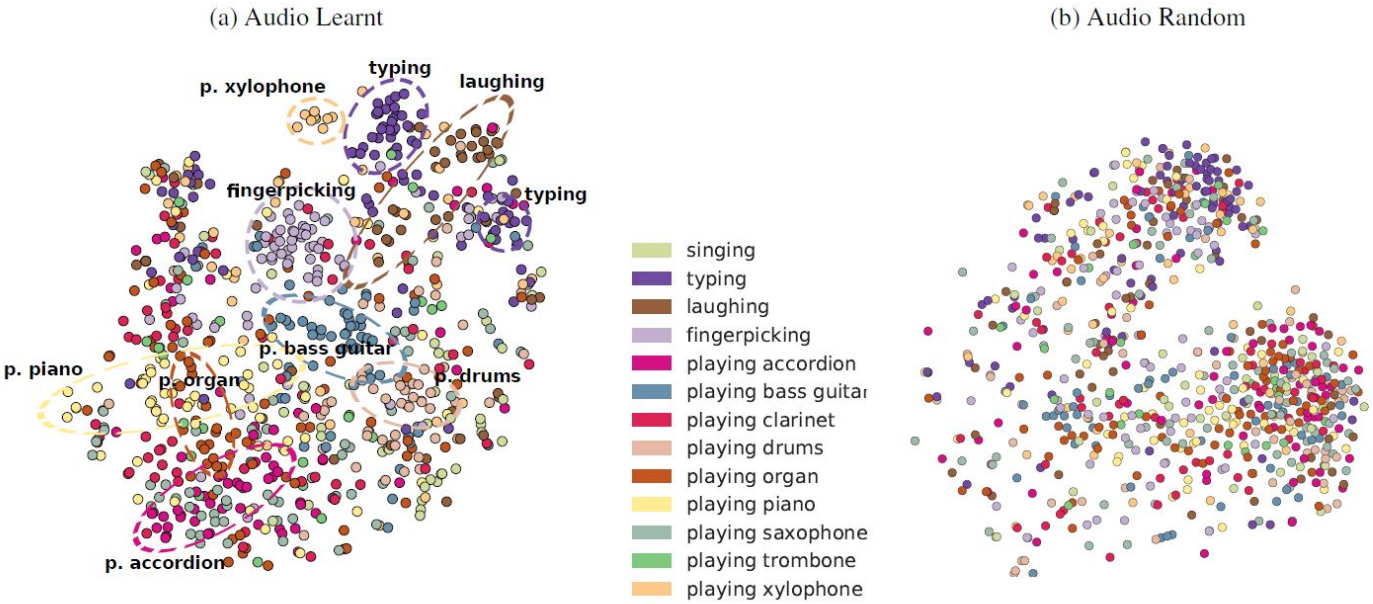
[Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." ICCV 2017]

3.5.1 Audio Features on Kinetics-Sounds



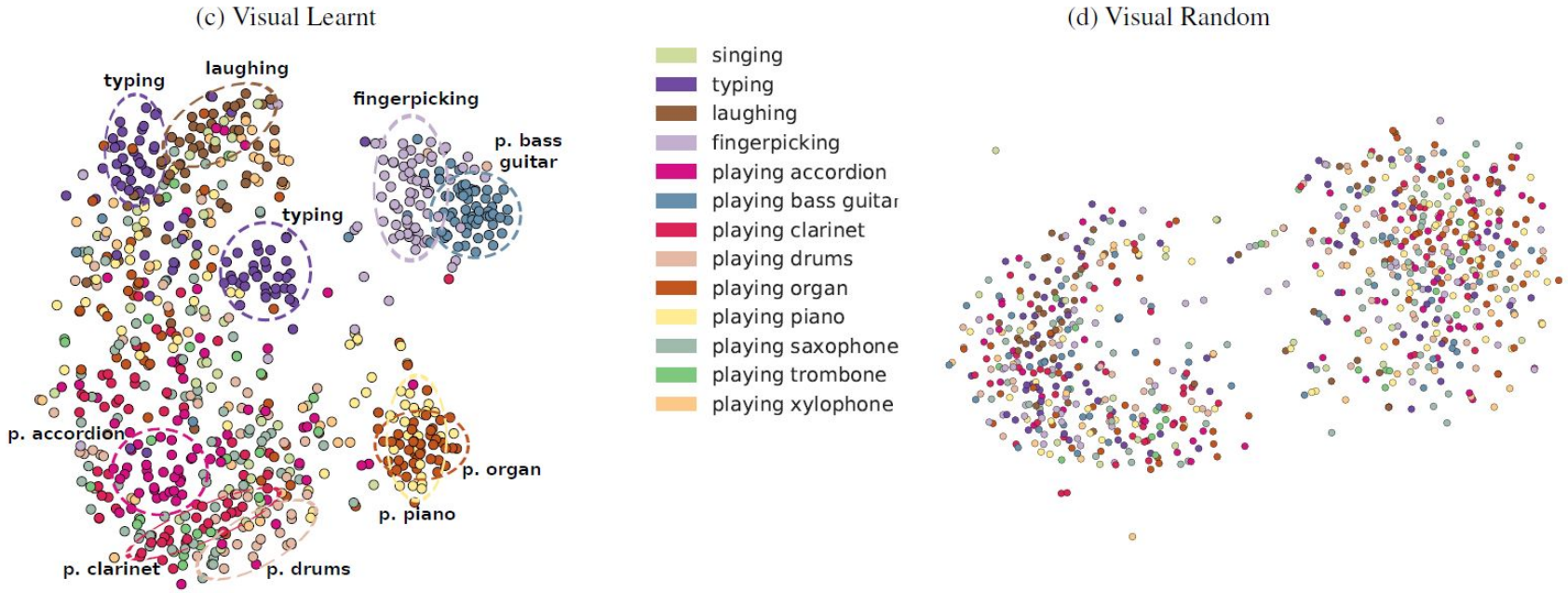
[Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." ICCV 2017]

T-SNE on Audio



[Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." ICCV 2017]

T-SNE on Video



[Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." ICCV 2017]

Strength & Weakness

- *Strengths*
  - Introduces a new method to utilize audio to train video networks, and vice-versa
  - Lots of training data is available for self-supervised models
- *Weakness*
  - Same architecture for both subnetworks
  - Uses VGG architecture (why not at least ResNet? Maybe a little lazy because authors are in VGG group)
  - Not enough explanation on why their model works or fails on some tasks
  - Not exploiting temporal feature of video in visual subnetwork
  - Might not be able to learn finer details of an action (guitar picking v. power chords)
  - You can train anything quickly if you work for Google :(