Ambient Sound Provides Supervision for Visual Learning

Andrew Owens[1], Jiajun Wu[1], Josh H. McDermott[1], William T. Freeman[1,2], and Antonio Torralba[1]

[1]MIT & [2]Google Research

ECCV 2016

Presented by An T. Nguyen

Introduction

Problem

- Learn Image Representation without labels ...
- ... that useful for a real task (e.g. Object Recognition).

Introduction

Problem

- Learn Image Representation without labels ...
- ... that useful for a real task (e.g. Object Recognition).

Idea

- Set up a pretext task.
- To solve pretext task, model must learn good representation.

Introduction

Problem

- Learn Image Representation without labels ...
- ... that useful for a real task (e.g. Object Recognition).

Idea

- Set up a pretext task.
- To solve pretext task, model must learn good representation.

Learn to predict a "natural signal"...

- ...that available for 'free'.

Introduction

Problem

- Learn Image Representation without labels ...
- ... that useful for a real task (e.g. Object Recognition).

Idea

- Set up a pretext task.
- To solve pretext task, model must learn good representation.

Learn to predict a “natural signal”...

- ...that available for ‘free’.
- This paper: Sound.

Introduction

Problem

- Learn Image Representation without labels ...
- ... that useful for a real task (e.g. Object Recognition).

Idea

- Set up a pretext task.
- To solve pretext task, model must learn good representation.

Learn to predict a "natural signal"...

- ...that available for 'free'.
- This paper: Sound.
- Others: Camera motion.
  (Agrawal et. al., Jayaraman & Grauman, 2015)

Data

Yahoo Flickr Creative Commons 100 Million Dataset.
(Thomee et. al. 2015)

Data

Yahoo Flickr Creative Commons 100 Million Dataset.
(Thomee et. al. 2015)

- 360,000 video subset.
- Sample one image per 10sec.
- Extract 3.75 sec of sound around.
- 1.8 mil. train examples.

Examples 1

(flickr.com/photos/41894173046@N01/4530333858)

Sound

Video

Examples 2

(flickr.com/photos/42035325@N00/8029349128)

Sound

Video

Examples 3

(flickr.com/photos/zen/2479982751)

Sound

Video

Challenges

- Sound is sometimes indicative of image.
- But sometimes not.

Challenges

- Sound is sometimes indicative of image.
- But sometimes not.

Sound producing objects

- outside image.
- not always produce sound.

Challenges

- Sound is sometimes indicative of image.
- But sometimes not.

Sound producing objects

- outside image.
- not always produce sound.

Video

- is edited.
- has noisy, background sound.

Challenges

- Sound is sometimes indicative of image.
- But sometimes not.

Sound producing objects

- outside image.
- not always produce sound.

Video

- is edited.
- has noisy, background sound.

Question: What representation can we learn?

Represent sound

Pre-process

Represent sound

Pre-process

- Filter waveform ... (mimic human ear).

Represent sound

Pre-process

- Filter waveform ... (mimic human ear).
- Compute statistics (e.g. mean of each freq. channel).

Represent sound

Pre-process

- Filter waveform ... (mimic human ear).
- Compute statistics (e.g. mean of each freq. channel).
- $\rightarrow$ sound texture: 502-dim vector.

Represent sound

Pre-process

- Filter waveform ... (mimic human ear).
- Compute statistics (e.g. mean of each freq. channel).
- $\rightarrow$ sound texture: 502-dim vector.

Two labeling models

1. Cluster sound texture (k-mean).

Represent sound

Pre-process

- Filter waveform ... (mimic human ear).
- Compute statistics (e.g. mean of each freq. channel).
- $\rightarrow$ sound texture: 502-dim vector.

Two labeling models

1. Cluster sound texture (k-mean).
2. PCA, 30 projections, threshold $\rightarrow$ binary codes.

Represent sound

Pre-process

- Filter waveform ... (mimic human ear).
- Compute statistics (e.g. mean of each freq. channel).
- $\rightarrow$ sound texture: 502-dim vector.

Two labeling models

1. Cluster sound texture (k-mean).
2. PCA, 30 projections, threshold $\rightarrow$ binary codes.

Given an image

1. Predict sound cluster.
2. Predict 30 binary codes (multi-label classification).

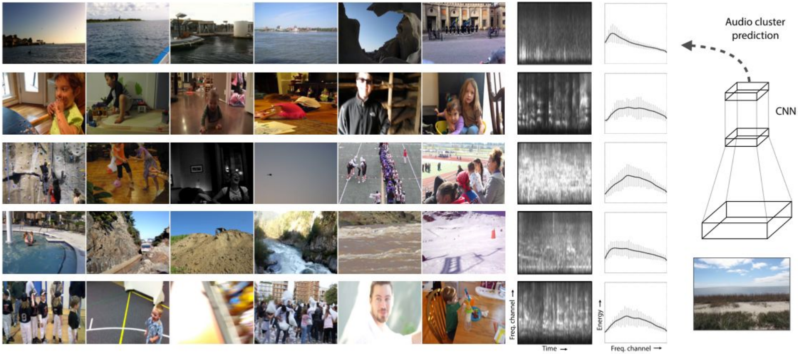
Training

Training

Convolutional Neural Network

- Similar to (Krizhevsky et. al. 2012).
- Implemented in Caffe.

Training



(a) Images grouped by audio cluster (b) Clustered audio stats. (c) CNN model

Visualizing neurons (in upper layers)

Visualizing neurons (in upper layers)

Method: for each neuron

Visualizing neurons (in upper layers)

Method: for each neuron

1. Find images with large activation.

Visualizing neurons (in upper layers)

Method: for each neuron

1. Find images with large activation.
2. Find locations with large contribution to activation.

Visualizing neurons (in upper layers)

Method: for each neuron

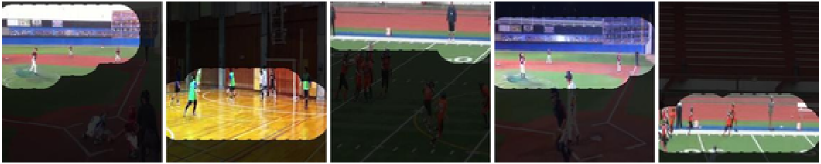
1. Find images with large activation.
2. Find locations with large contribution to activation.
3. Highlight these regions.

Visualizing neurons (in upper layers)

Method: for each neuron

1. Find images with large activation.
2. Find locations with large contribution to activation.
3. Highlight these regions.
4. Show to human on AMT.

Visualizing neurons



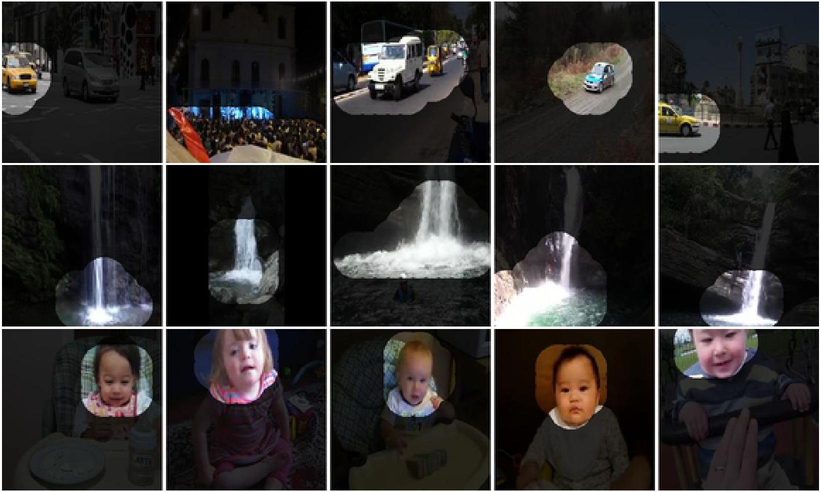
Visualizing neurons

Visualizing neurons

Visualizing neurons

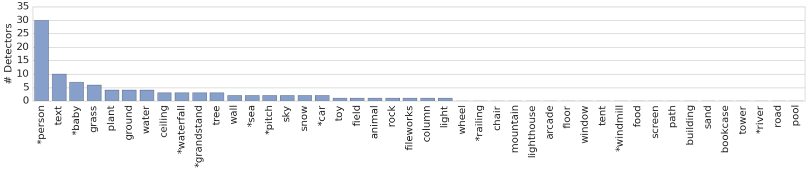
Visualizing neurons

Visualizing neurons



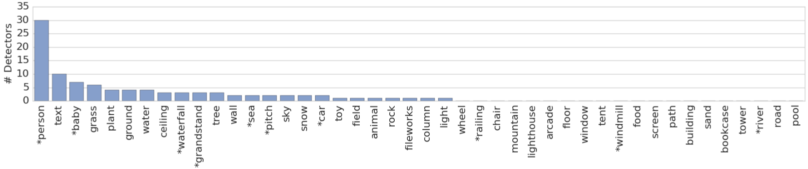
Detectors Histogram

Sound

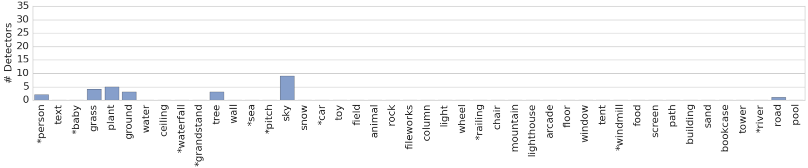


Detectors Histogram

Sound

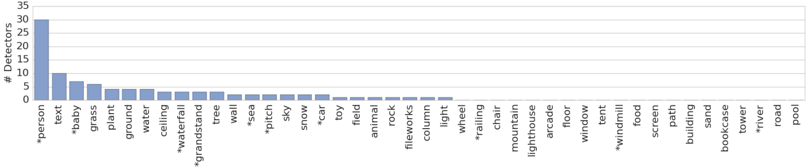


Ego Motion

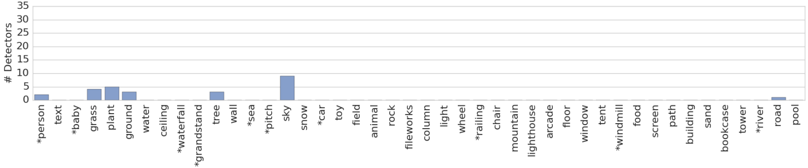


Detectors Histogram

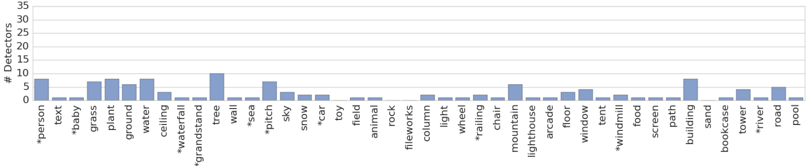
Sound



Ego Motion



Labeled Scenes (supervised)



Observations

- Each method learn some kinds of representations...
- ...depend on the pretext task.

Observations

- Each method learn some kinds of representations...
- ...depend on the pretext task.

Representation learned from sound

- Objects with distinctive sound.
- Complementary to other methods.

Object/Scene Recognition (1-vs-rest SVM)

1. Agrawal et.al. 2015
4. Doersch et.al 2015
20. Krähenbühl et.al. 2016
35. Wang & Gupta 2015

Object/Scene Recognition (1-vs-rest SVM)

Method	VOC Cls. (%mAP)				SUN397 (%acc.)			
	max5	pool5	fc6	fc7	max5	pool5	fc6	fc7
Sound (cluster)	36.7	45.8	44.8	44.3	**17.3**	**22.9**	20.7	14.9
Sound (binary)	**39.4**	**46.7**	**47.1**	**47.4**	17.1	22.5	**21.3**	**21.4**
Sound (spect.)	35.8	44.0	44.4	44.4	14.6	19.5	18.6	17.7
Texton-CNN	28.9	37.5	35.3	32.5	10.7	15.2	11.4	7.6
K-means [20]	27.5	34.8	33.9	32.1	11.6	14.9	12.8	12.4
Tracking [35]	33.5	42.2	42.4	40.2	14.1	18.7	16.2	15.1
Patch pos. [4]	26.8	46.1	-	-	9.8	22.2	-	-
Egomotion [1]	22.7	31.1	-	-	9.1	11.3	-	-
ImageNet [21]	**63.6**	**65.6**	**69.6**	**73.6**	29.8	34.0	37.8	37.8
Places [39]	59.0	63.2	65.3	66.2	**39.4**	**42.1**	**46.1**	**48.8**

(a) Image classification with linear SVM

1. Agrawal et.al. 2015
4. Doersch et.al 2015
20. Krähenbühl et.al. 2016
35. Wang & Gupta 2015

Object/Scene Recognition (1-vs-rest SVM)

Method	VOC Cls. (%mAP)				SUN397 (%acc.)			
	max5	pool5	fc6	fc7	max5	pool5	fc6	fc7
Sound (cluster)	36.7	45.8	44.8	44.3	**17.3**	**22.9**	20.7	14.9
Sound (binary)	**39.4**	**46.7**	**47.1**	**47.4**	17.1	22.5	**21.3**	**21.4**
Sound (spect.)	35.8	44.0	44.4	44.4	14.6	19.5	18.6	17.7
Texton-CNN	28.9	37.5	35.3	32.5	10.7	15.2	11.4	7.6
K-means [20]	27.5	34.8	33.9	32.1	11.6	14.9	12.8	12.4
Tracking [35]	33.5	42.2	42.4	40.2	14.1	18.7	16.2	15.1
Patch pos. [4]	26.8	46.1	-	-	9.8	22.2	-	-
Egomotion [1]	22.7	31.1	-	-	9.1	11.3	-	-
ImageNet [21]	**63.6**	**65.6**	**69.6**	**73.6**	29.8	34.0	37.8	37.8
Places [39]	59.0	63.2	65.3	66.2	**39.4**	**42.1**	**46.1**	**48.8**

(a) Image classification with linear SVM

Comparable Performance to Others

1. Agrawal et.al. 2015
4. Doersch et.al 2015
20. Krähenbühl et.al. 2016
35. Wang & Gupta 2015

Object Detection (Pretrain Fast-RCNN)

1. Agrawal et.al. 2015
4. Doersch et.al 2015
20. Krähenbühl et.al. 2016
35. Wang & Gupta 2015

Object Detection (Pretrain Fast-RCNN)

Method	(%mAP)
Random init. [20]	41.3
Sound (cluster)	44.1
Sound (binary)	43.3
Motion [35,20]	44.0
Egomotion [1,20]	41.8
Patch pos. [4,20]	46.6
Calib. + Patch [4,20]	**51.1**
ImageNet [21]	**57.1**
Places [39]	52.8

(b) Finetuning detection

1. Agrawal et.al. 2015
4. Doersch et.al 2015
20. Krähenbühl et.al. 2016
35. Wang & Gupta 2015

Object Detection (Pretrain Fast-RCNN)

Method	(%mAP)
Random init. [20]	41.3
Sound (cluster)	44.1
Sound (binary)	43.3
Motion [35,20]	44.0
Egomotion [1,20]	41.8
Patch pos. [4,20]	46.6
Calib. + Patch [4,20]	**51.1**
ImageNet [21]	**57.1**
Places [39]	52.8

(b) Finetuning detection

Similar Performance to Motion

1. Agrawal et.al. 2015
4. Doersch et.al 2015
20. Krähenbühl et.al. 2016
35. Wang & Gupta 2015

Discussion

Sound

- is abundant.
- can learn good representations.
- complementary to visual info.

Discussion

Sound

- is abundant.
- can learn good representations.
- complementary to visual info.

Future work

- Other sound representations.
- What object/scene detectable by sound?

Bonus: Visually Indicative Sound

(Owens et. al. 2016, vis.csail.mit.edu)