@DocXavi

Master in Computer Vision Barcelona

[http://pagines.uab.cat/mcv/]

Xavier Giró-i-Nieto

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Signal Theory and Communications
Image Processing Group

Module 6

Deep Learning from Videos

13th March 2018

Deep Learning online courses by UPC:

- [MSc course](#) (2017)
- [BSc course](#) (2018)

Next edition Autumn 2018

- [1st edition](#) (2016)
- [2nd edition](#) (2017)
- [3rd edition](#) (2018)

Summer School (late June 2018)

- [1st edition](#) (2017)
- [2nd edition](#) (2018)

Next edition Winter/Spring 2019

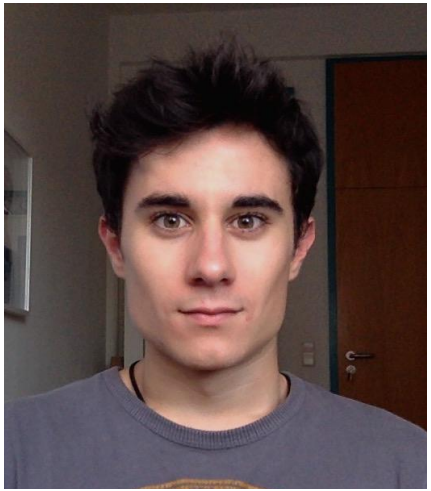
Acknowledegments (Unsupervised)

Insight

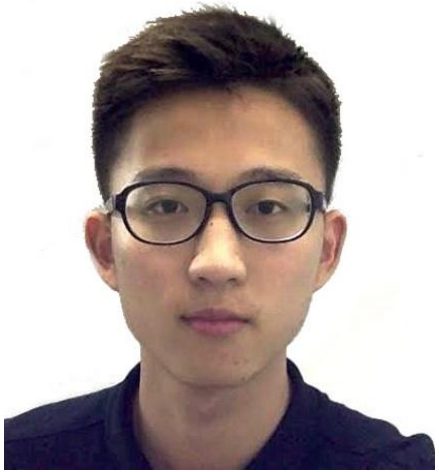
DCU

Kevin McGuinness, “Unsupervised Learning” Deep Learning for Computer Vision.
[Slides 2016] [Slides 2017]

Acknowledgements (feature learning)



Víctor Campos



Junting Pan

Xunyu Lin

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

UNIVERSITY OF MICHIGAN

Carnegie Mellon University

Densely linked slides

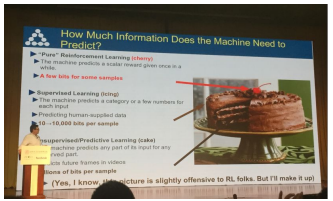
Outline

1. **Unsupervised Learning**
2. Predictive Learning
3. Self-supervised Learning
4. Cross-modal Learning

Unsupervised Learning

- "Pure" Reinforcement Learning (cherry)
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- Supervised Learning (icing)
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- Unsupervised/Predictive Learning (cake)
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

- (Yes, I know, this picture is slightly offensive to RL folks. But I'll make it up)



Slide credit:
Yann LeCun

Yann LeCun

Monday at 10:15 · Edited ·

Statement from a Slashdot post about the AlphaGo victory: "We know now that we don't need any big new breakthroughs to get to true AI"

That is completely, utterly, ridiculously wrong.

As I've said in previous statements: most of human and animal learning is unsupervised learning. If intelligence was a cake, unsupervised learning would be the cake, supervised learning would be the icing on the cake, and reinforcement learning would be the cherry on the cake. We know how to make the icing and the cherry, but we don't know how to make the cake.

We need to solve the unsupervised learning problem before we can even think of getting to true AI. And that's just an obstacle we know about. What about all the ones we don't know about?

#deeplearning #AI #AlphaGo

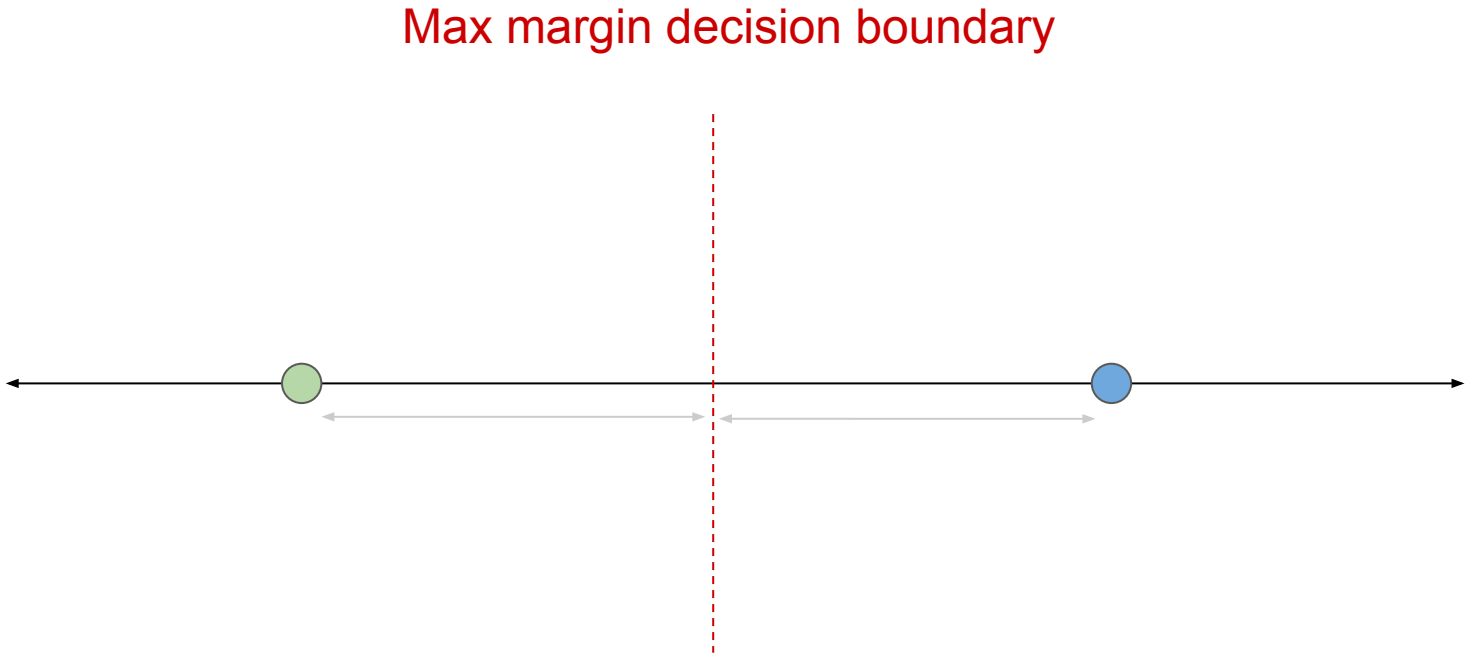
Greff, Klaus, Antti Rasmus, Mathias Berglund, Tele Hao, Harri Valpola, and Juergen Schmidhuber. "Tagger: Deep unsupervised perceptual grouping." NIPS 2016 [video] [code]

Unsupervised Learning

Why Unsupervised Learning?

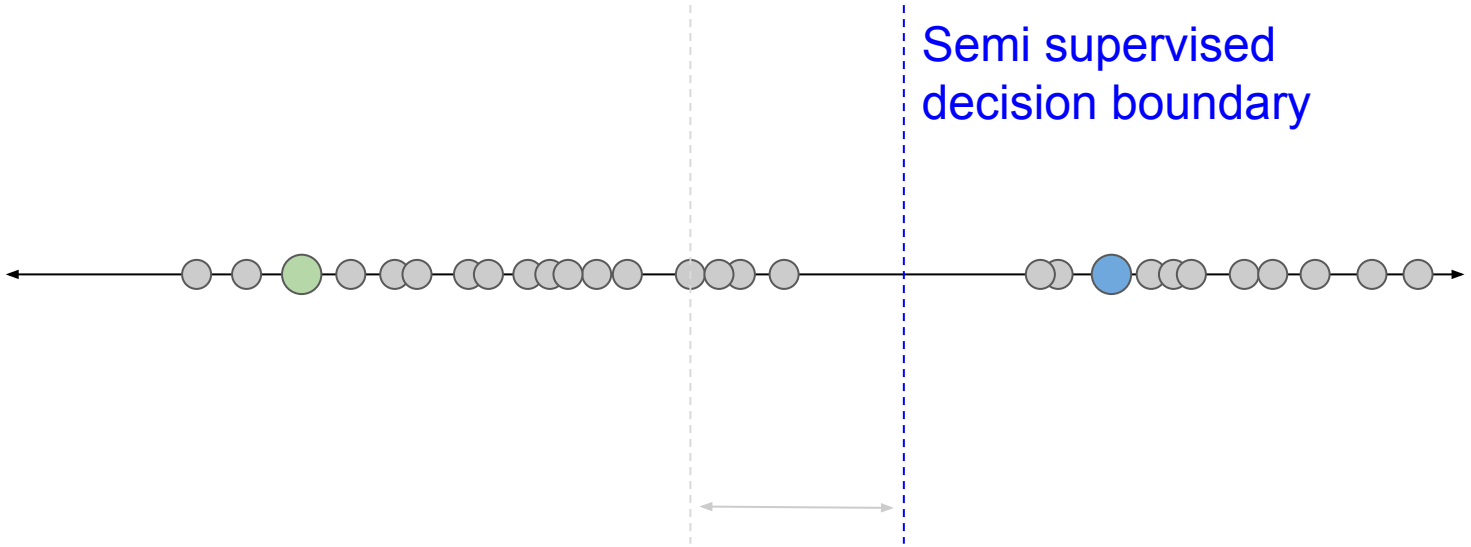
- It is the nature of how intelligent beings percept the world.
- It can save us tons of efforts to build a human-alike intelligent agent compared to a totally supervised fashion.
- Vast amounts of unlabelled data.

How data distribution P(x) influences decisions (1D)



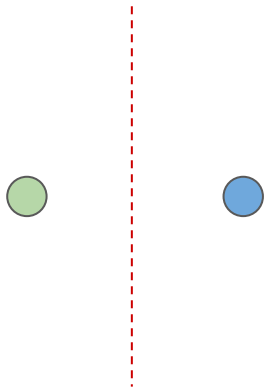
Slide credit: Kevin McGuinness (DLCV UPC 2017)

How data distribution P(x) influences decisions (1D)



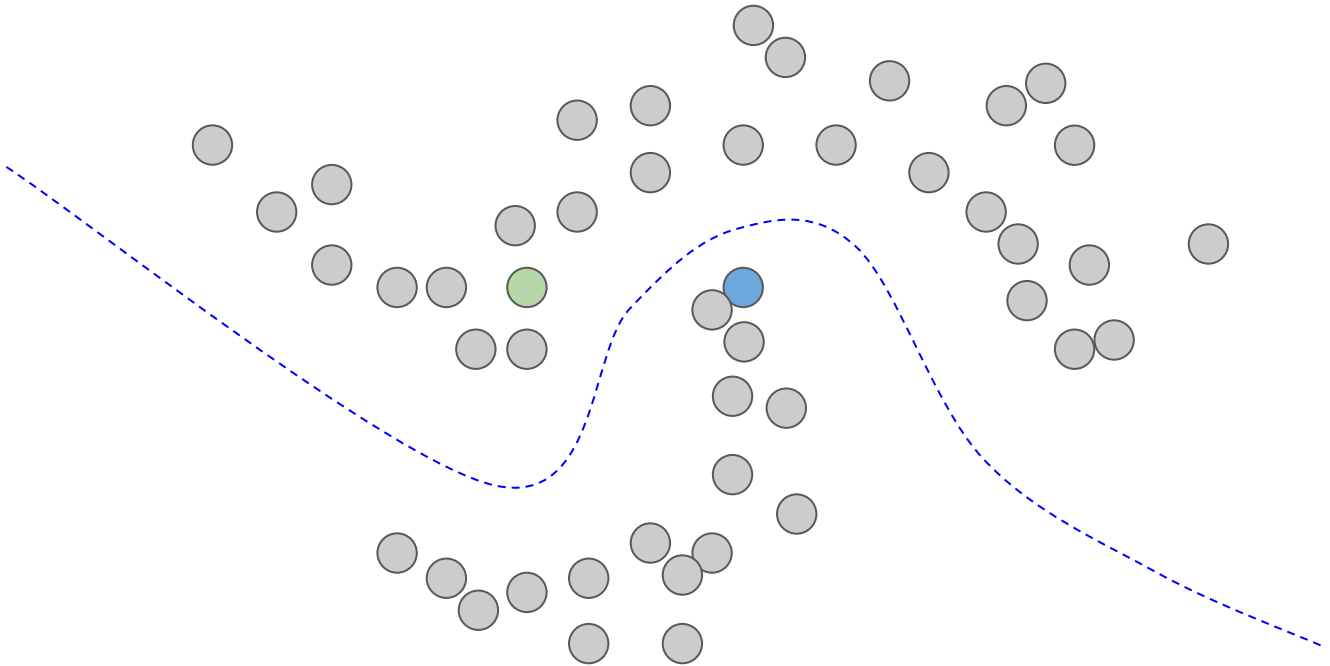
Slide credit: Kevin McGuinness (DLCV UPC 2017)

How data distribution P(x) influences decisions (2D)



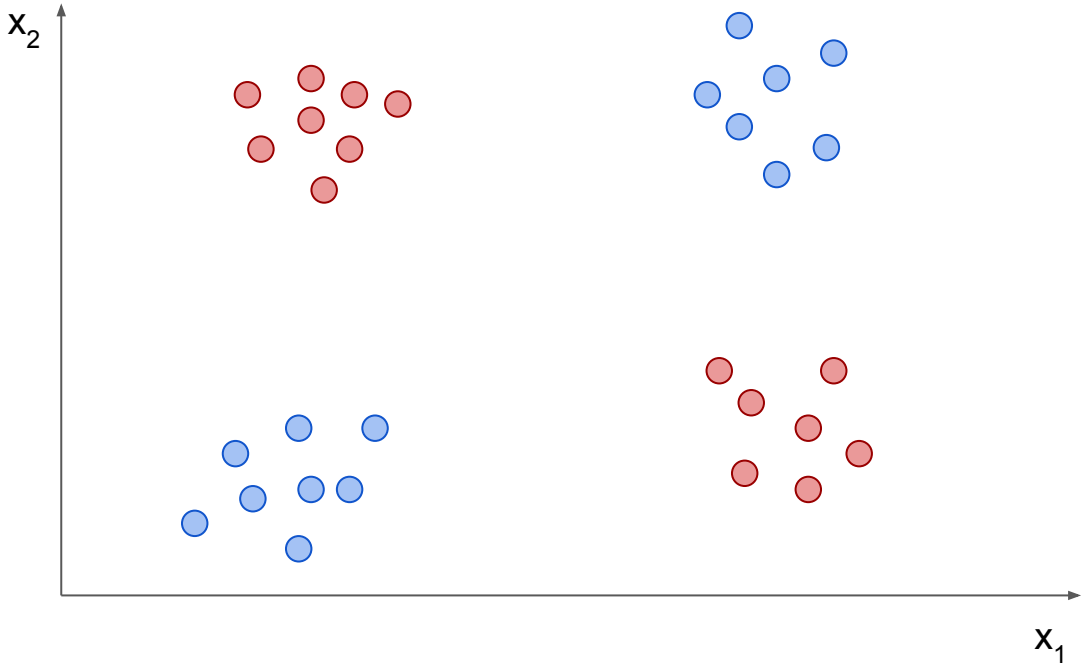
Slide credit: Kevin McGuinness (DLCV UPC 2017)

How data distribution P(x) influences decisions (2D)



Slide credit: Kevin McGuinness (DLCV UPC 2017)

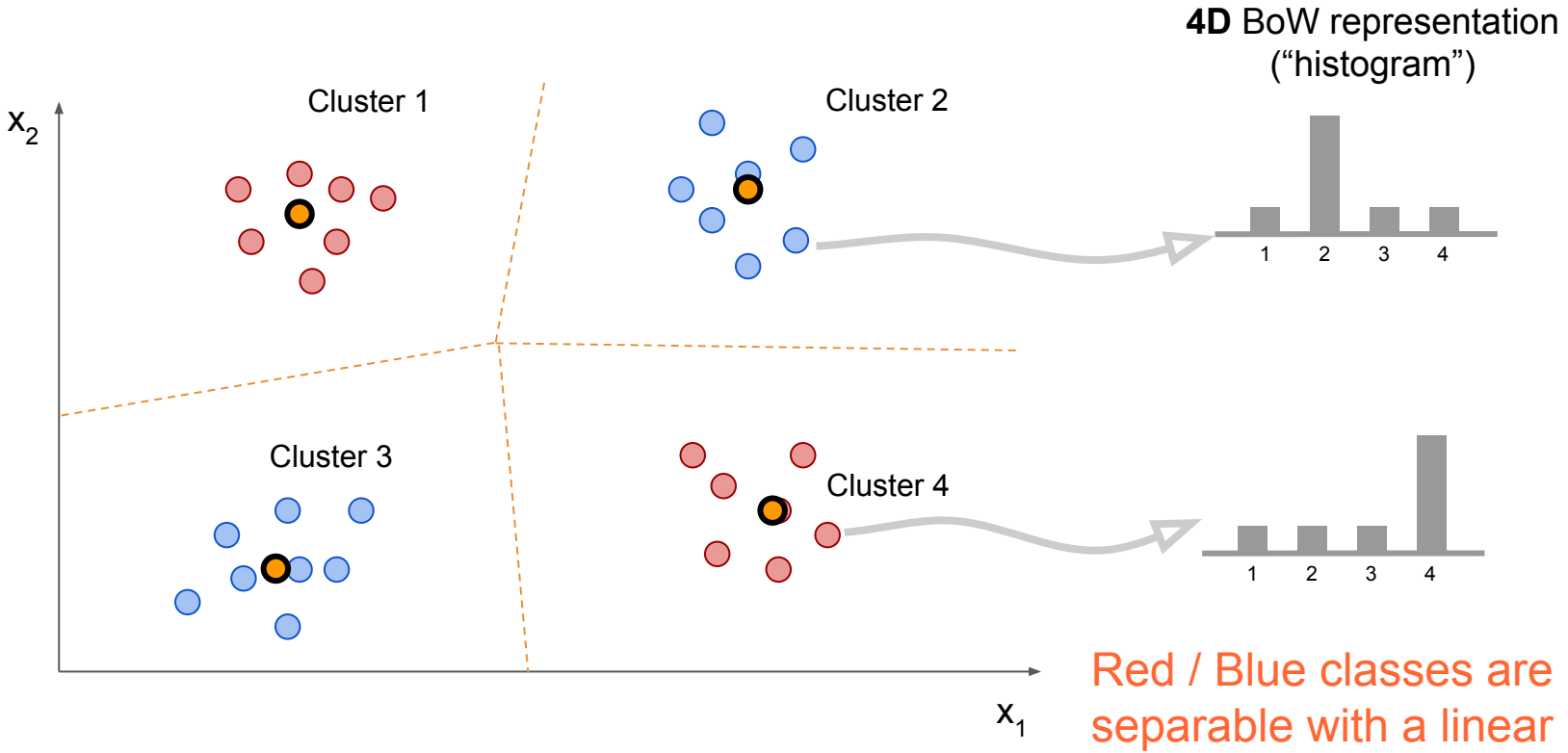
How clustering is valuable for linear classifiers



Red / Blue classes are nt linearly separable in this 2D space :(

Slide credit: Kevin McGuinness (DLCV UPC 2017)

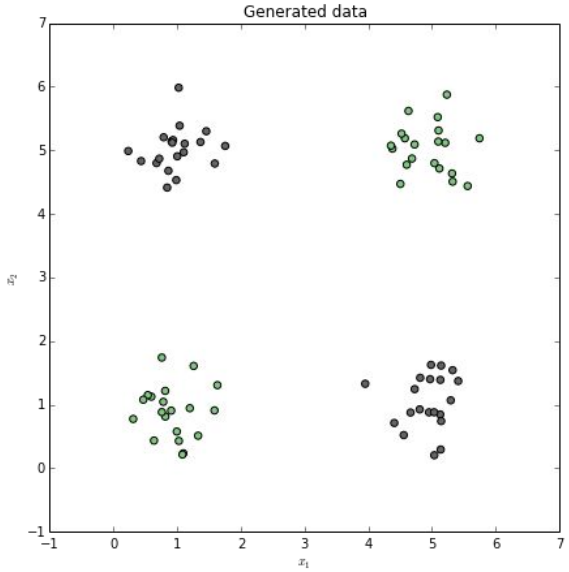
How clustering is valuable for linear classifiers

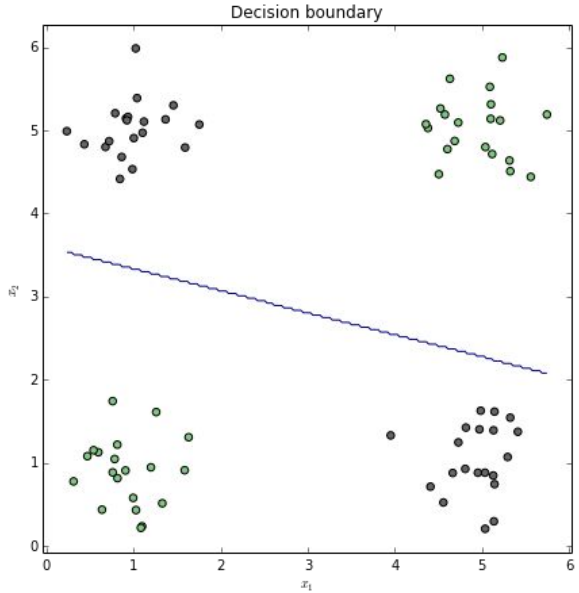


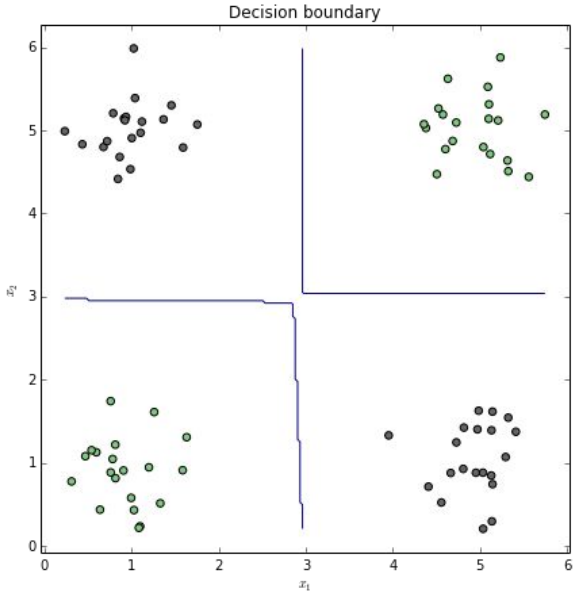
Red / Blue classes are now separable with a linear classifiers in a 4D space :)

Slide credit: Kevin McGuinness (DLCV UPC 2017)

How clustering is valueble for linear classifiers







Example from [Kevin McGuinness Juyter notebook](#).

Assumptions for unsupervised learning

To model P(X) given data, it is necessary to make some assumptions

“You can’t do inference without making assumptions”

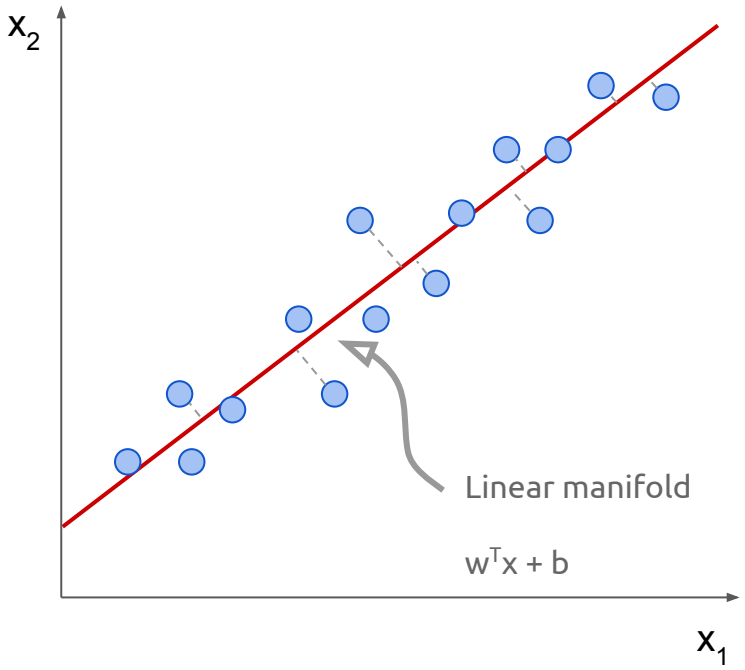
-- David MacKay, Information Theory, Inference, and Learning Algorithms

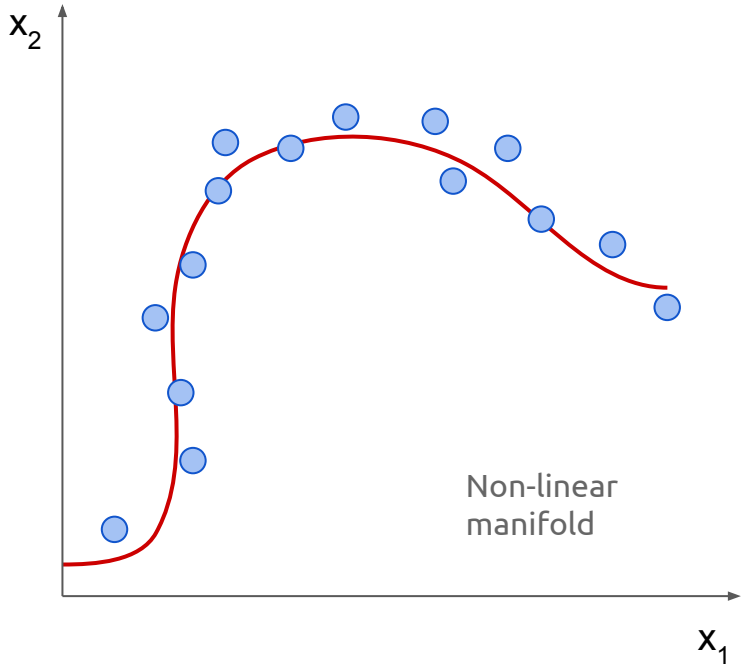
Typical assumptions:

- Smoothness assumption
  - Points which are close to each other are more likely to share a label.
- Cluster assumption
  - The data form discrete clusters; points in the same cluster are likely to share a label
- **Manifold assumption**
  - The data lie approximately on a manifold of much lower dimension than the input space.

Slide credit: Kevin McGuinness (DLCV UPC 2017)

The manifold hypothesis





Slide credit: Kevin McGuinness (DLCV UPC 2017)

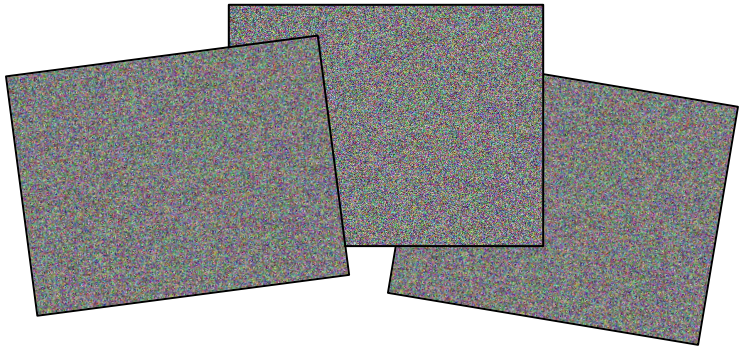
The manifold hypothesis

The data distribution lies close to a low-dimensional manifold

Example: **consider image data**

- Very high dimensional (1,000,000D)
- A randomly generated image will almost certainly not look like any real world scene
  - The space of images that occur in nature is almost completely empty
- Hypothesis: real world images lie on a smooth, low-dimensional manifold
  - Manifold distance is a good measure of similarity

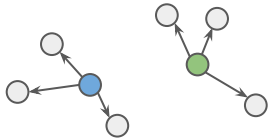
Similar for audio and text



Slide credit: Kevin McGuinness (DLCV UPC 2017)

Assumptions for unsupervised learning

Smoothness assumption



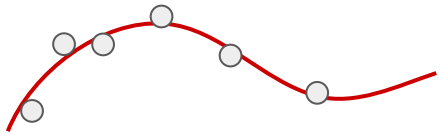
- Label propagation
  - Recursively propagate labels to nearby points
  - Problem: in high-D, your nearest neighbour may be very far away!

Cluster assumption

- Bag of words models
  - K-means, etc.
  - Represent points by cluster centers
  - Soft assignment
  - VLAD
- Gaussian mixture models
  - Fisher vectors

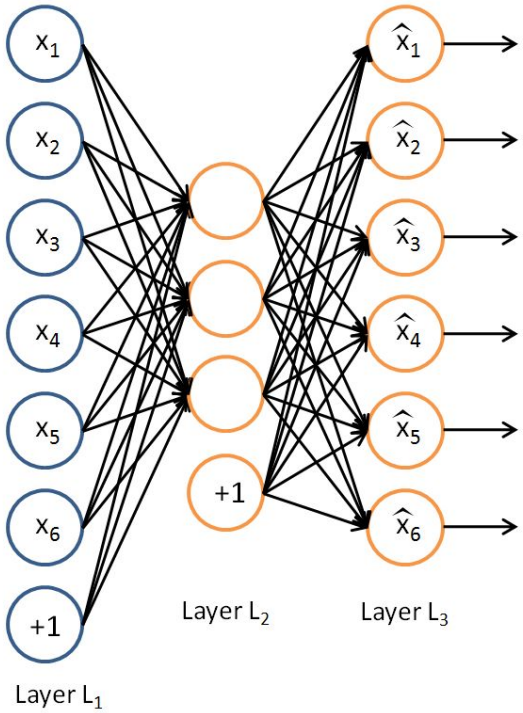
Manifold assumption

- Linear manifolds
  - PCA
  - Linear autoencoders
  - Random projections
  - ICA
- Non-linear manifolds:
  - Non-linear autoencoders
  - Deep autoencoders
  - Restricted Boltzmann machines
  - Deep belief nets



Slide credit: Kevin McGuinness (DLCV UPC 2017)

Autoencoder (AE)

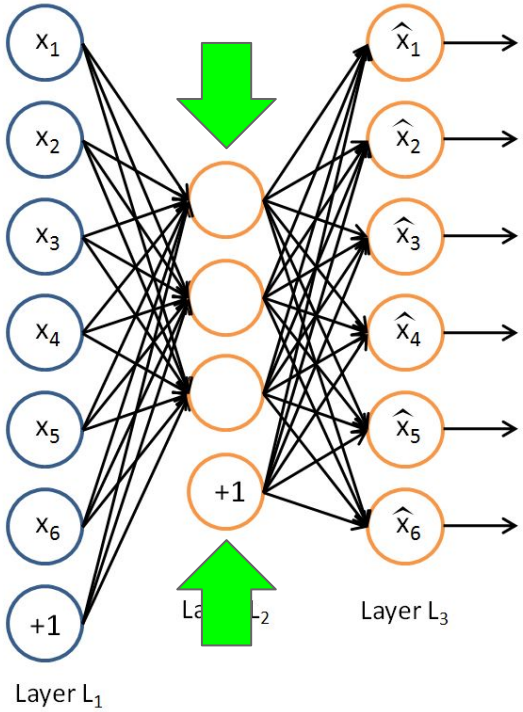


Autoencoders:

- Predict at the output the same input data.
- Do not need labels:

"Deep Learning Tutorial", Dept. Computer Science, Stanford

Autoencoder (AE)



Application #1

Dimensionality reduction:

- Use hidden layer as a feature extractor of any desired size.

"Deep Learning Tutorial", Dept. Computer Science, Stanford

Autoencoder (AE)

Application #2

WHY?

Pretraining:

1. Initialize a NN solving an autoencoding problem.

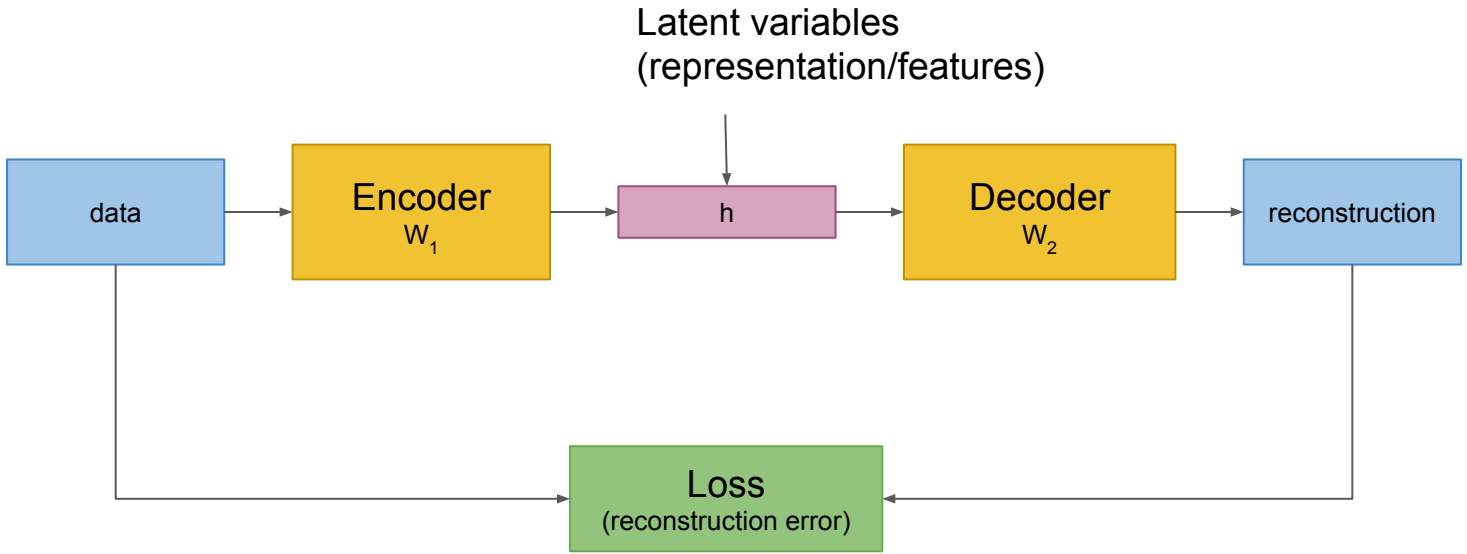


Figure: Kevin McGuinness (DLCV UPC 2017)

Autoencoder (AE)

Application #2

WHY?

Pretraining:

1. Initialize a NN solving an autoencoding problem.
2. Train for final task with “few” labels.

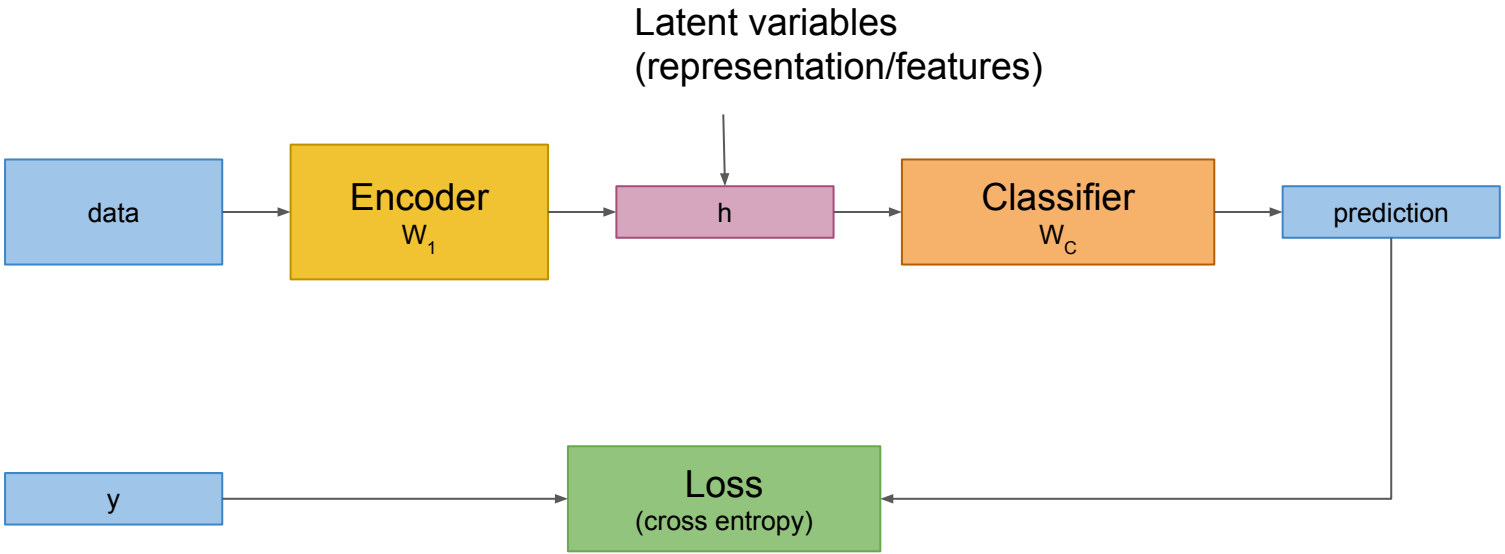


Figure: Kevin McGuinness (DLCV UPC 2017)

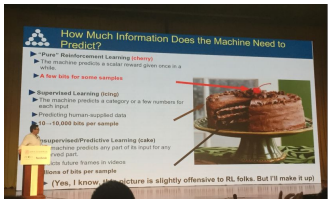
Outline

1. Unsupervised Learning
2. **Predictive Learning**
3. Self-supervised Learning
4. Cross-modal Learning

Unsupervised Feature Learning

- "Pure" Reinforcement Learning (cherry)
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- Supervised Learning (icing)
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- Unsupervised/Predictive Learning (cake)
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

- (Yes, I know, this picture is slightly offensive to RL folks. But I'll make it up)

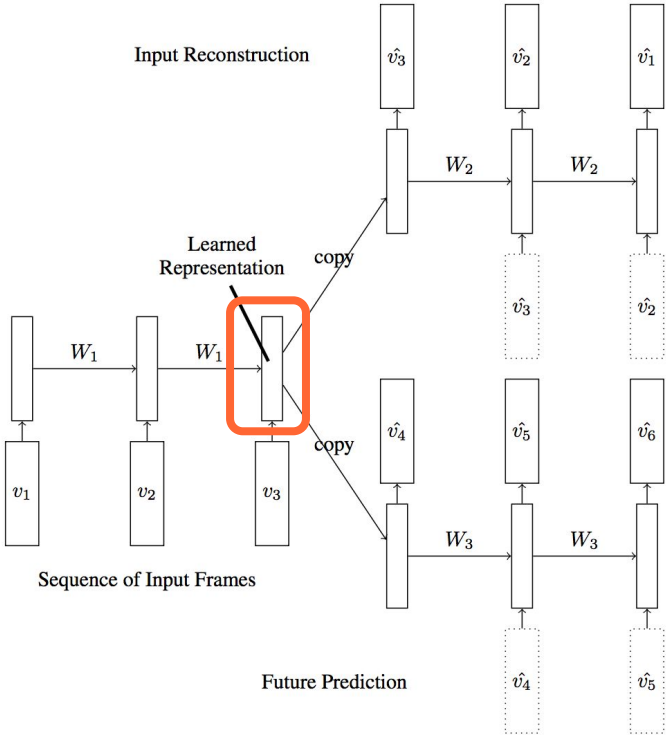


Slide credit:
Yann LeCun

Slide credit: Junting Pan

Frame Reconstruction & Prediction

Unsupervised feature learning (no labels) for...

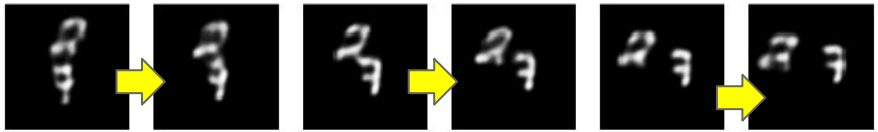


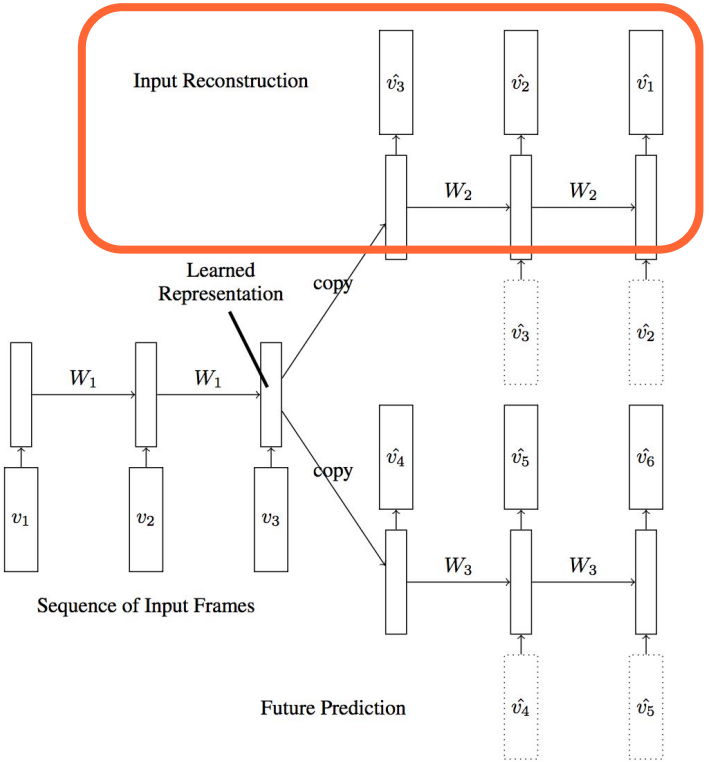
Srivastava, Nitish, Elman Mansimov, and Ruslan Salakhutdinov. "Unsupervised Learning of Video Representations using LSTMs." In ICML 2015. [Github]

Frame Reconstruction & Prediction

Unsupervised feature learning (no labels) for...

...frame prediction.



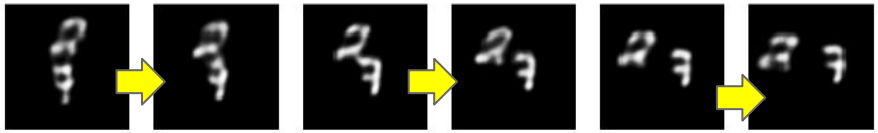


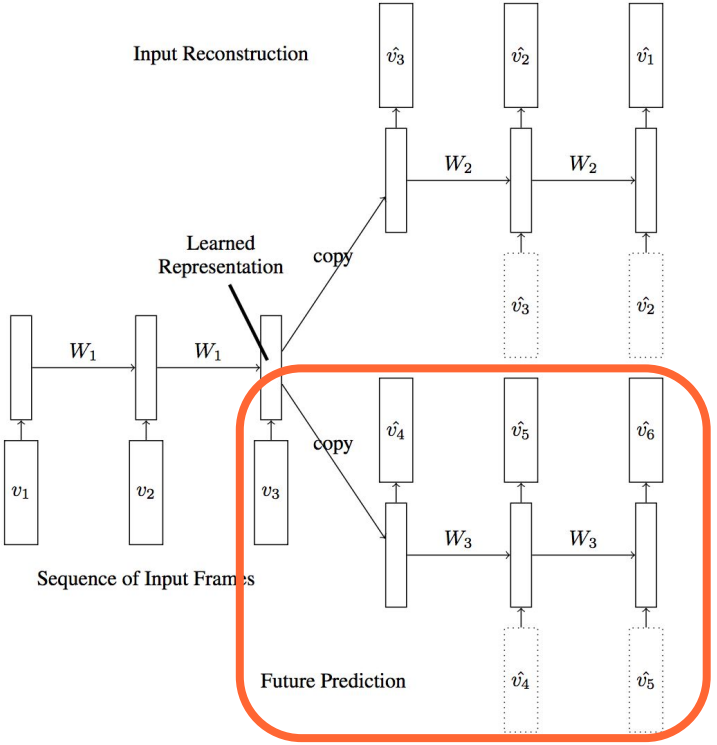
Srivastava, Nitish, Elman Mansimov, and Ruslan Salakhutdinov. "Unsupervised Learning of Video Representations using LSTMs." In ICML 2015. [Github]

Frame Reconstruction & Prediction

Unsupervised feature learning (no labels) for...

...frame prediction.





Srivastava, Nitish, Elman Mansimov, and Ruslan Salakhutdinov. "Unsupervised Learning of Video Representations using LSTMs." In ICML 2015. [Github]

Frame Reconstruction & Prediction

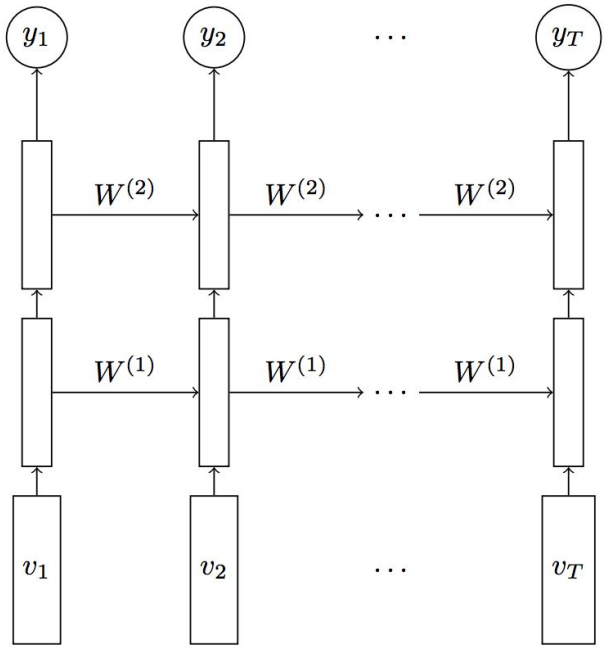


Figure 6. LSTM Classifier.

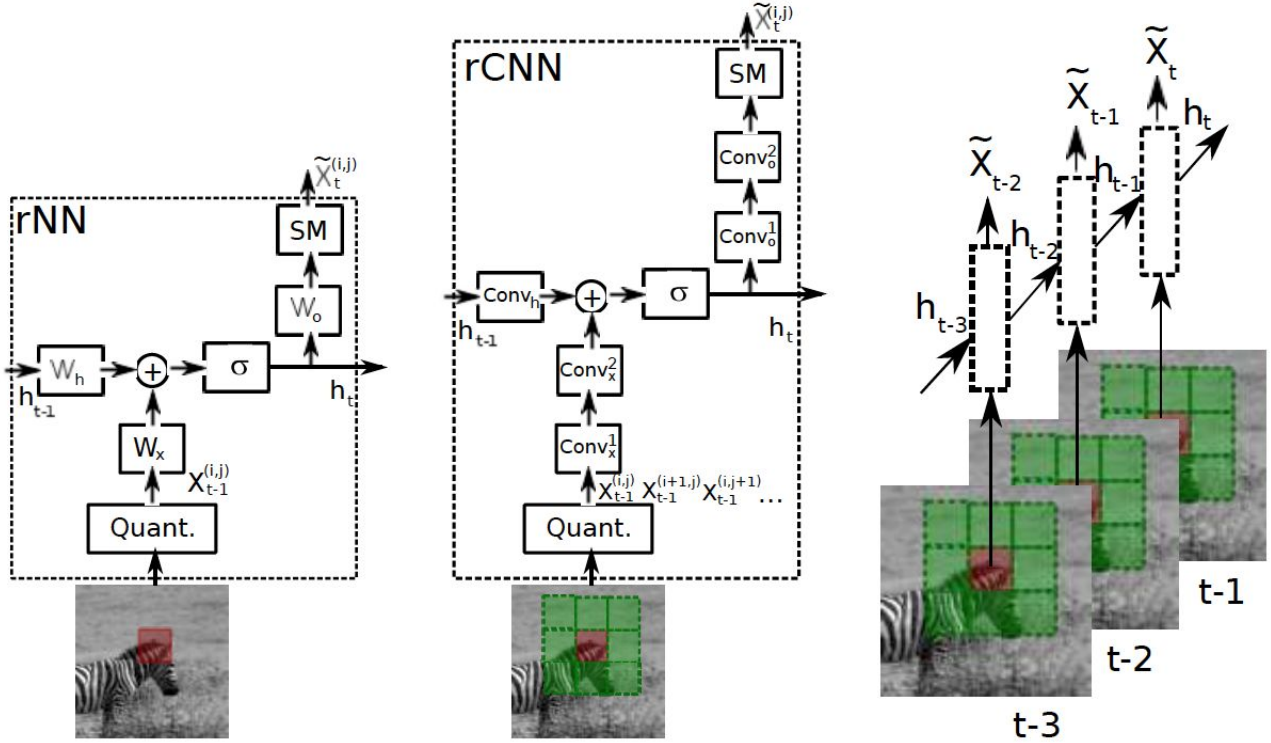
Unsupervised learned features (lots of data) are fine-tuned for activity recognition (little data).

Model	UCF-101 RGB	UCF-101 1- frame flow	HMDB-51 RGB
Single Frame	72.2	72.2	40.1
LSTM classifier	74.5	74.3	42.8
Composite LSTM Model + Finetuning	**75.8**	**74.9**	**44.1**

Table 1. Summary of Results on Action Recognition.

Srivastava, Nitish, Elman Mansimov, and Ruslan Salakhutdinov. "Unsupervised Learning of Video Representations using LSTMs." In ICML 2015. [Github]

Frame Prediction

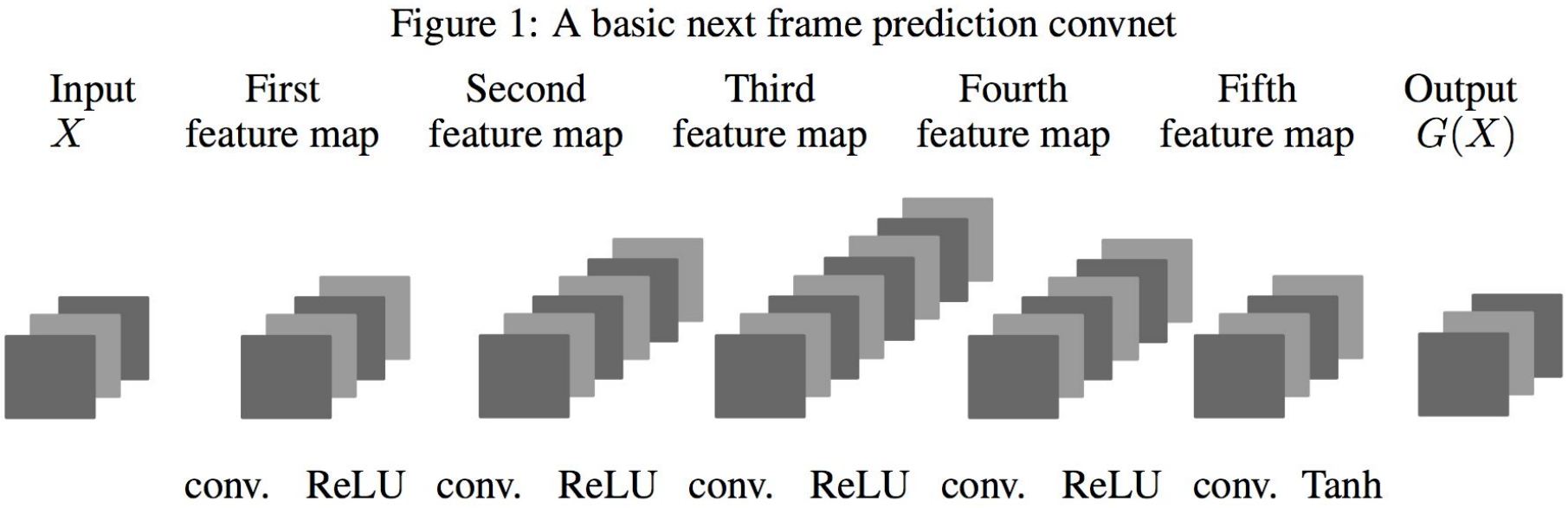


Ranzato, MarcAurelio, Arthur Szlam, Joan Bruna, Michael Mathieu, Ronan Collobert, and Sumit Chopra. "Video (language) modeling: a baseline for generative models of natural videos." arXiv preprint arXiv:1412.6604 (2014).

Frame Prediction

Video frame prediction with a ConvNet.

Figure 1: A basic next frame prediction convnet

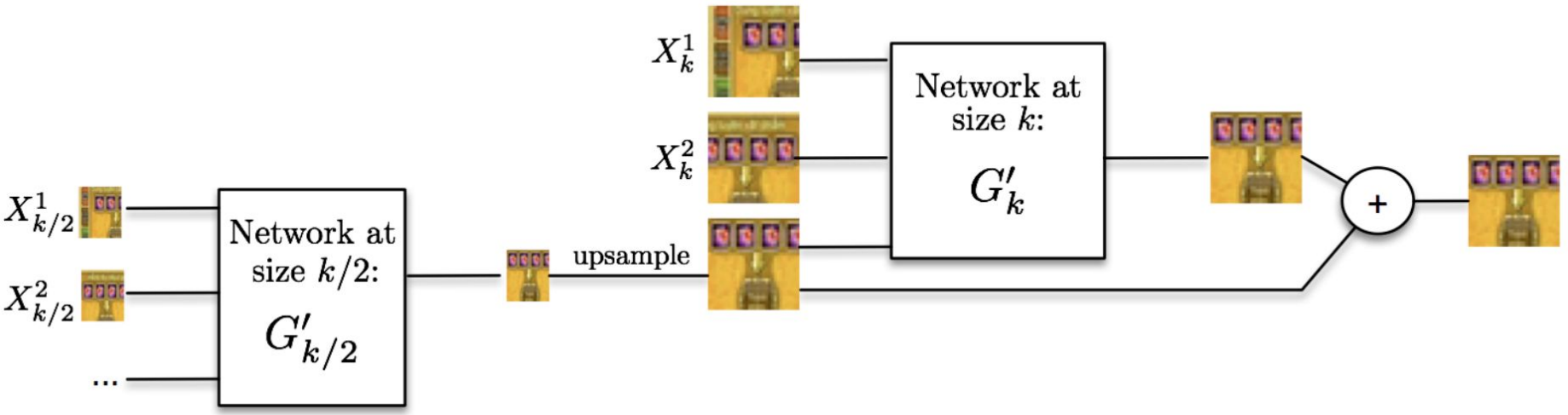


Mathieu, Michael, Camille Couprie, and Yann LeCun. "Deep multi-scale video prediction beyond mean square error." ICLR 2016 [project] [code]

Frame Prediction

The blurry predictions from MSE are improved with **multi-scale architecture**, adversarial learning and an image gradient difference loss function.

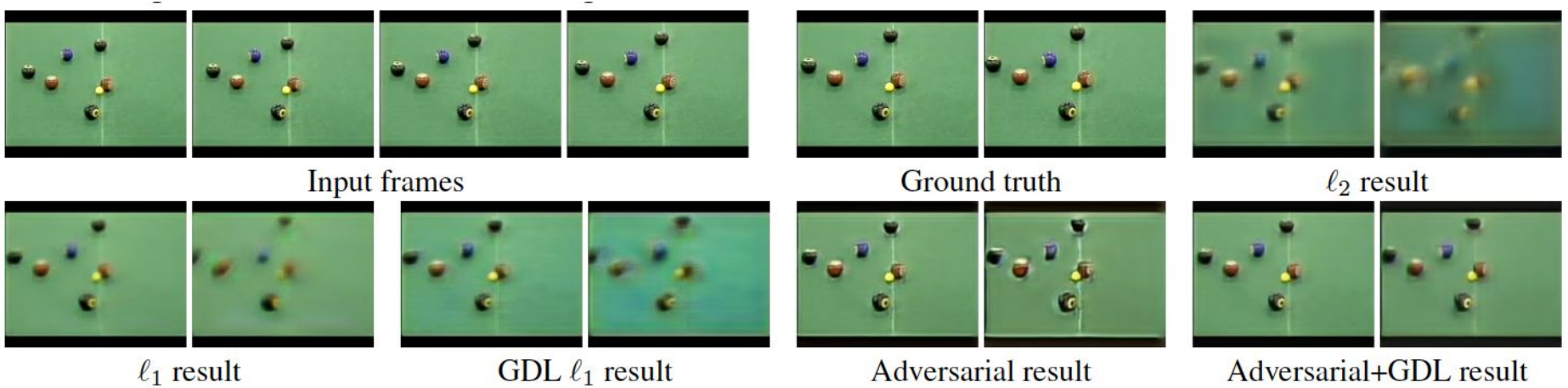
Figure 2: Multi-scale architecture



Mathieu, Michael, Camille Couprie, and Yann LeCun. "Deep multi-scale video prediction beyond mean square error." ICLR 2016 [project] [code]

Frame Prediction

The blurry predictions from MSE (l1) are improved with multi-scale architecture, **adversarial training** and an **image gradient difference loss (GDL)** function.

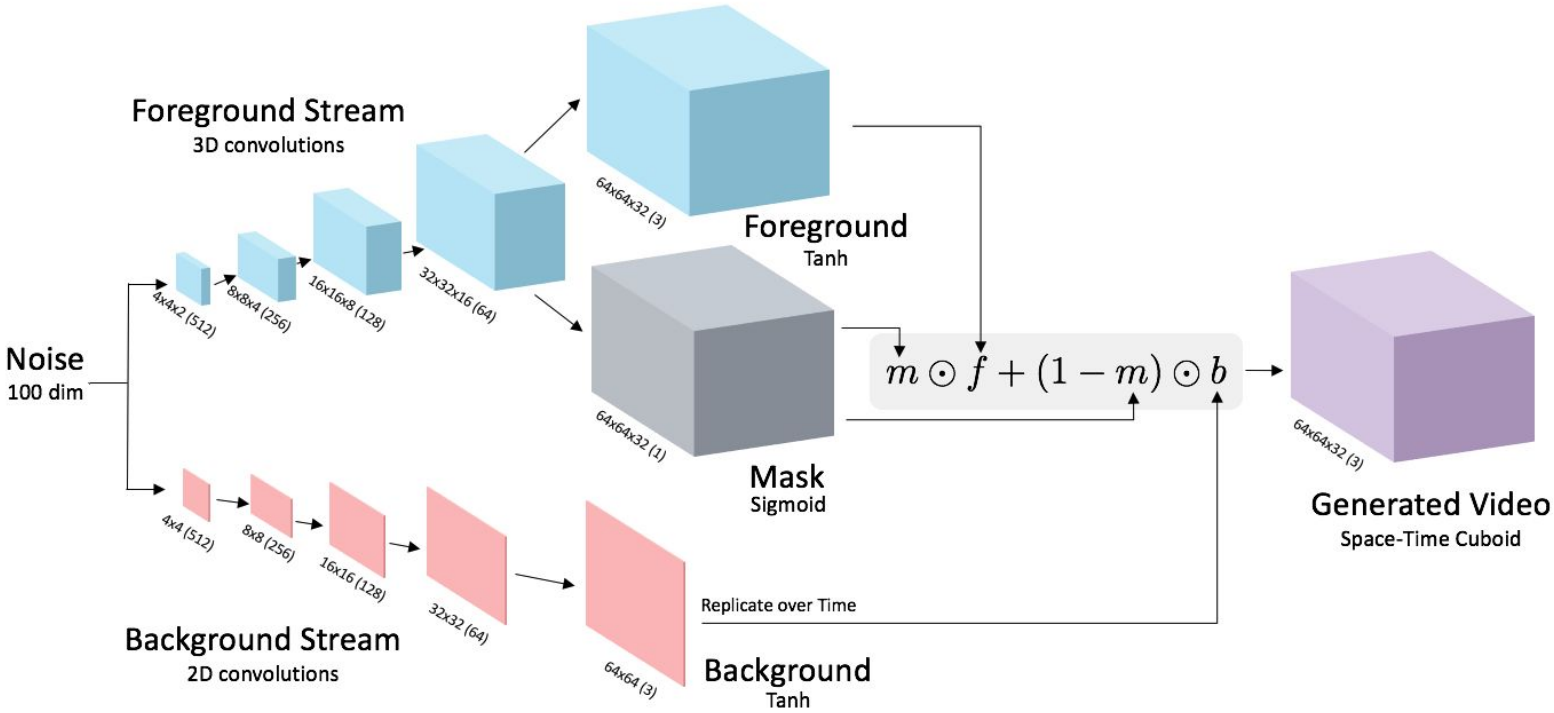


Mathieu, Michael, Camille Couprie, and Yann LeCun. "Deep multi-scale video prediction beyond mean square error." ICLR 2016 [project] [code]

Frame Prediction

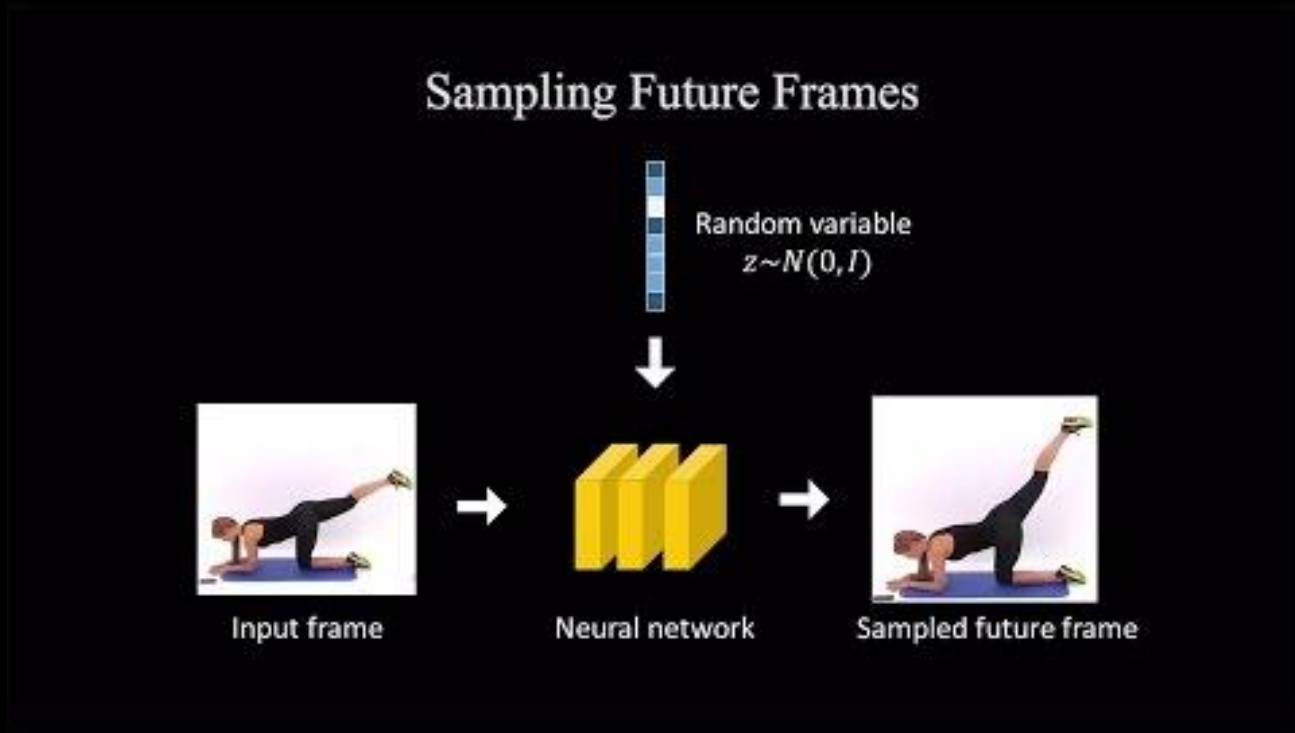
Mathieu, Michael, Camille Couprie, and Yann LeCun. "Deep multi-scale video prediction beyond mean square error." ICLR 2016 [project] [code]

Frame Prediction



Vondrick, Carl, Hamed Pirsiavash, and Antonio Torralba. "Generating videos with scene dynamics." NIPS 2016.

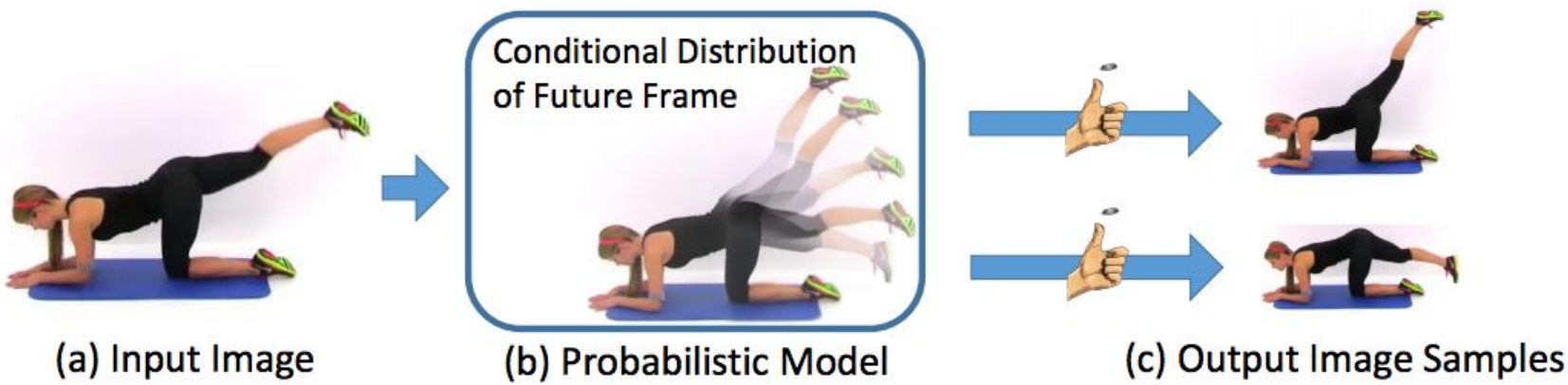
Vondrick, Carl, Hamed Pirsiavash, and Antonio Torralba. "Generating videos with scene dynamics." NIPS 2016.



Xue, Tianfan, Jiajun Wu, Katherine Bouman, and Bill Freeman. "Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks." NIPS 2016 [video]

Frame Prediction

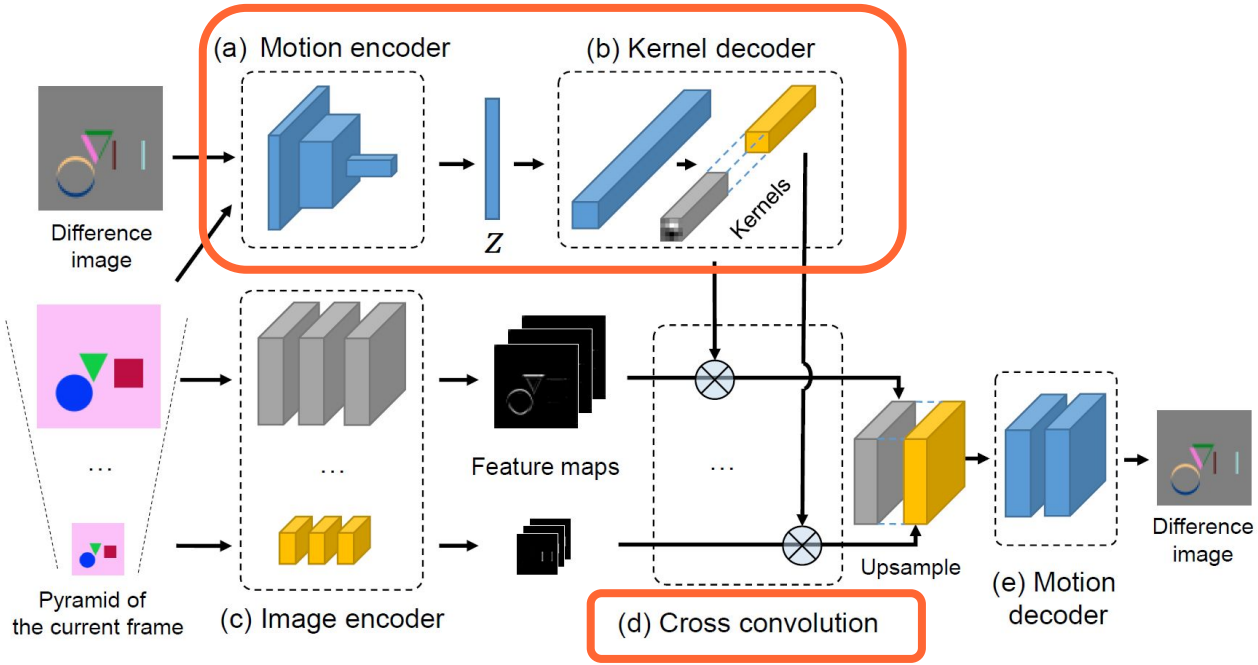
Given an input image, probabilistic generation of future frames with a **Variational AutoEncoder (VAE)**.



Xue, Tianfan, Jiajun Wu, Katherine Bouman, and Bill Freeman. "Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks." NIPS 2016 [video]

Frame Prediction

Encodes image as feature maps, and motion as and **cross-convolutional kernels**.

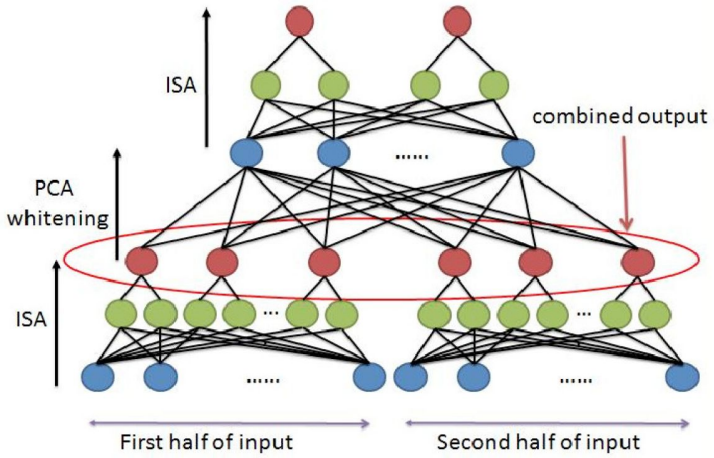


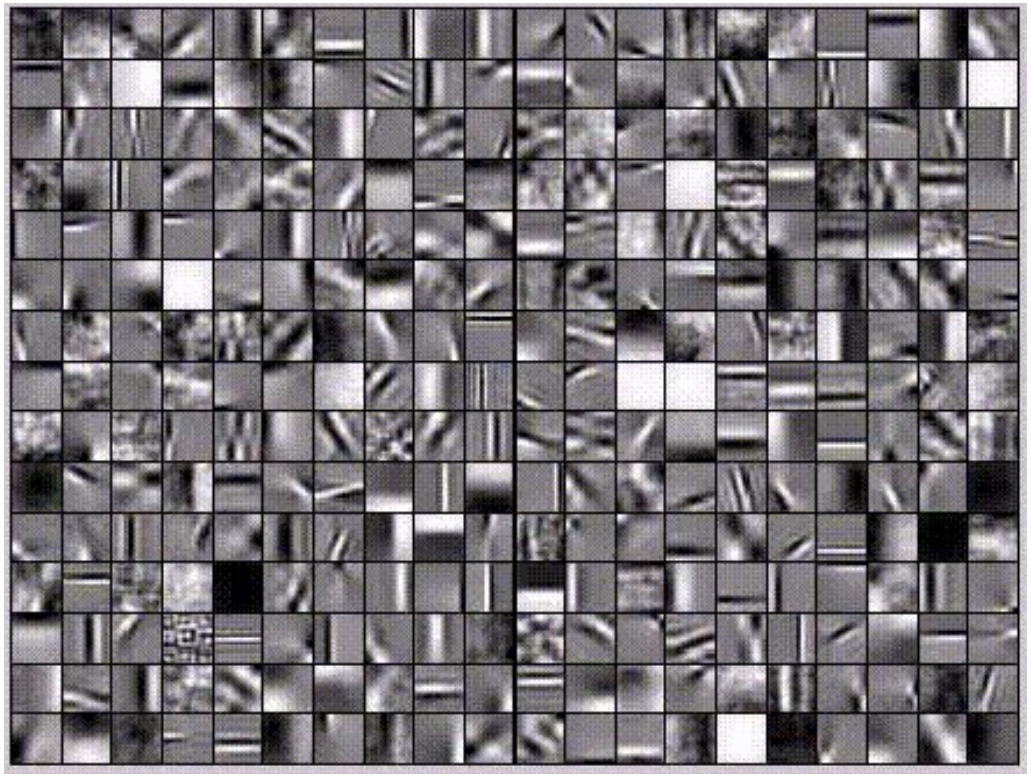
Xue, Tianfan, Jiajun Wu, Katherine Bouman, and Bill Freeman. "Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks." NIPS 2016 [video]

Outline

1. Unsupervised Learning
2. Predictive Learning
3. **Self-supervised Learning**
4. Cross-modal Learning

First steps in video feature learning

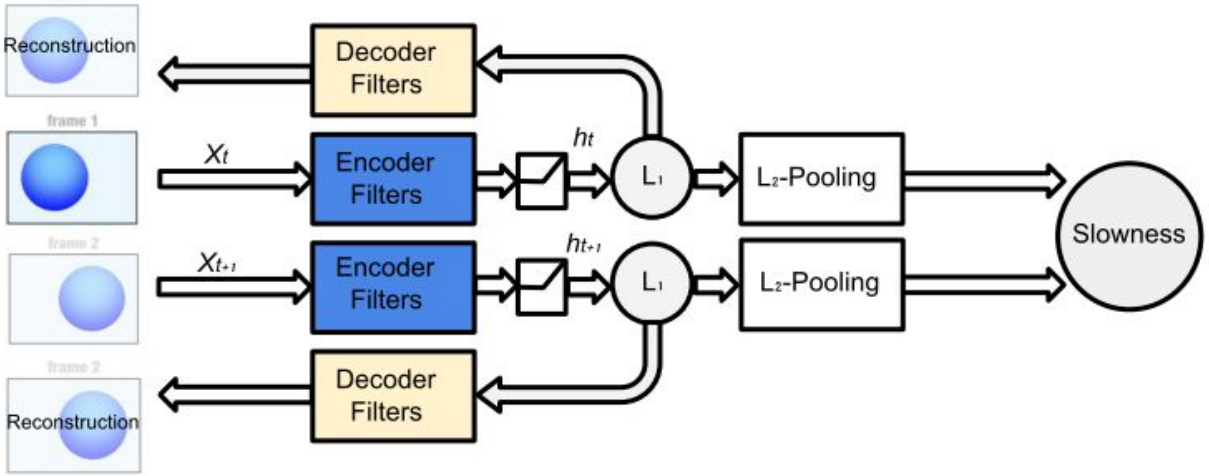




Le, Quoc V., Will Y. Zou, Serena Y. Yeung, and Andrew Y. Ng. "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis." CVPR 2011

Temporal Weak Labels

Assumption: adjacent video frames contain semantically similar information.
Autoencoder trained with regularizations by slowliness and sparisty.



Goroshin, Ross, Joan Bruna, Jonathan Tompson, David Eigen, and Yann LeCun. "Unsupervised learning of spatiotemporally coherent metrics." ICCV 2015.

Temporal Weak Labels

Slow feature analysis

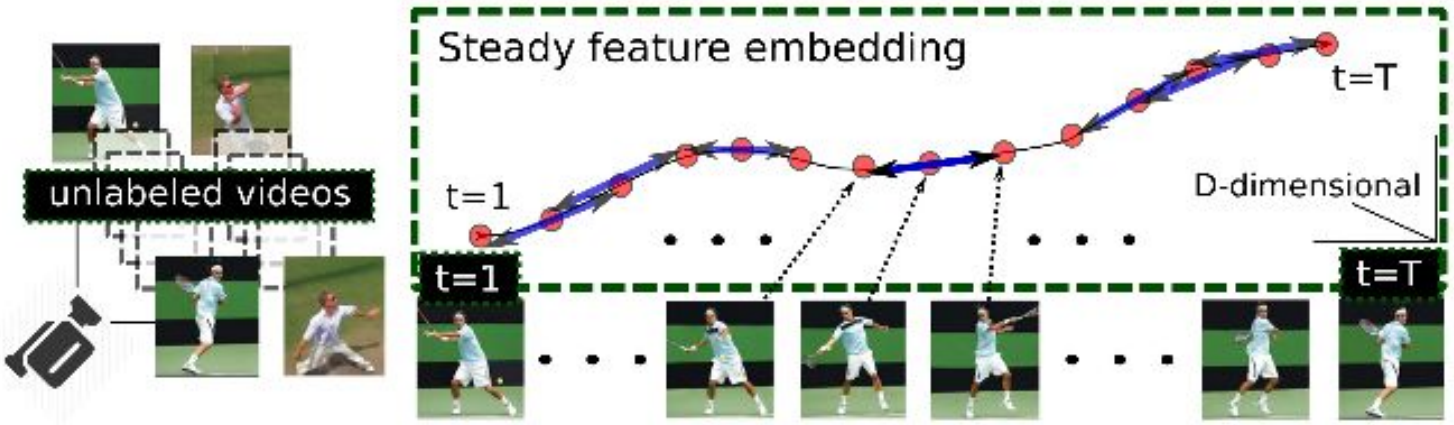
- Temporal coherence assumption: features should change slowly over time in video

Steady feature analysis

- Second order changes also small: changes in the past should resemble changes in the future

Train on triplets of frames from video

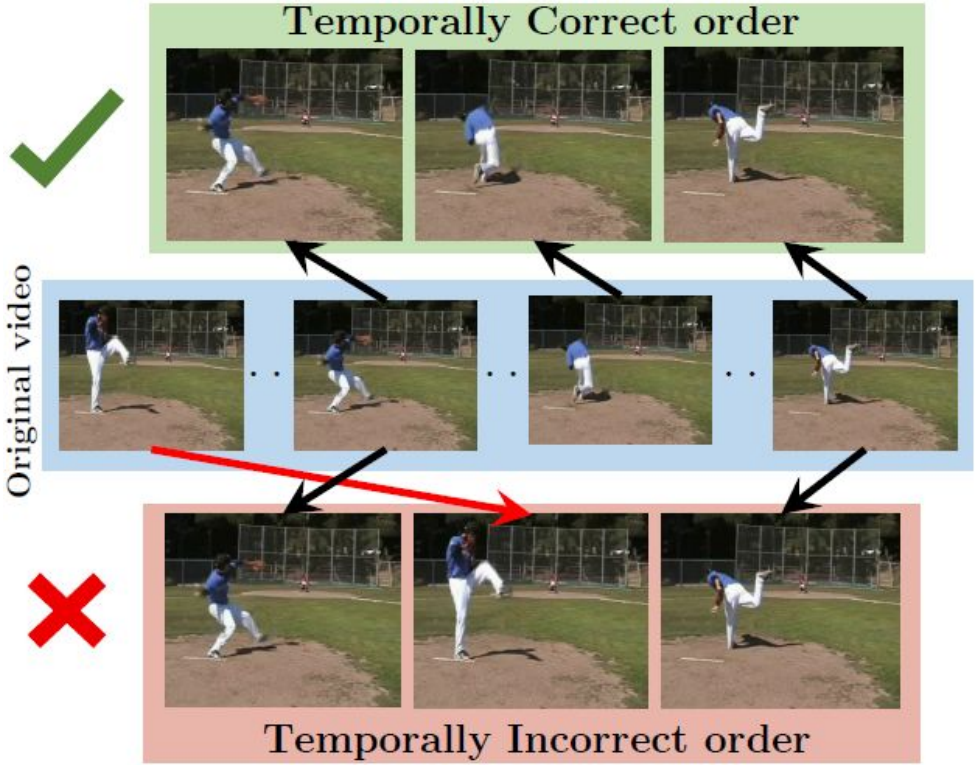
Loss encourages nearby frames to have slow and steady features, and far frames to have different features



Jayaraman, Dinesh, and Kristen Grauman. "Slow and steady feature analysis: higher order temporal coherence in video." CVPR 2016. [video]

Temporal Weak Labels

Temporal order of frames is exploited as the supervisory signal for learning.

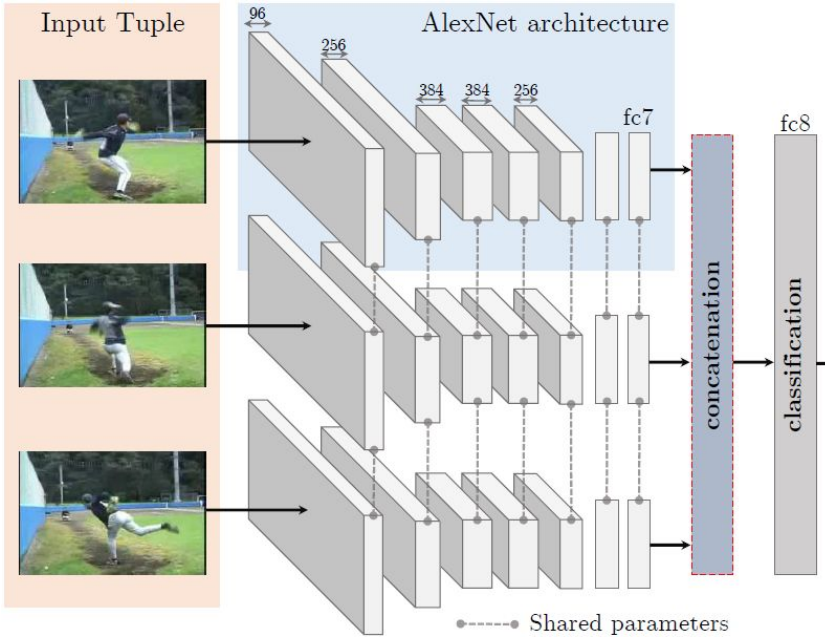


(Slides by Xunyu Lin): Misra, Ishan, C. Lawrence Zitnick, and Martial Hebert. "Shuffle and learn: unsupervised learning using temporal order verification." ECCV 2016. [code]

Temporal Weak Labels

Take temporal order as the supervisory signals for learning

Shuffled sequences



Binary classification

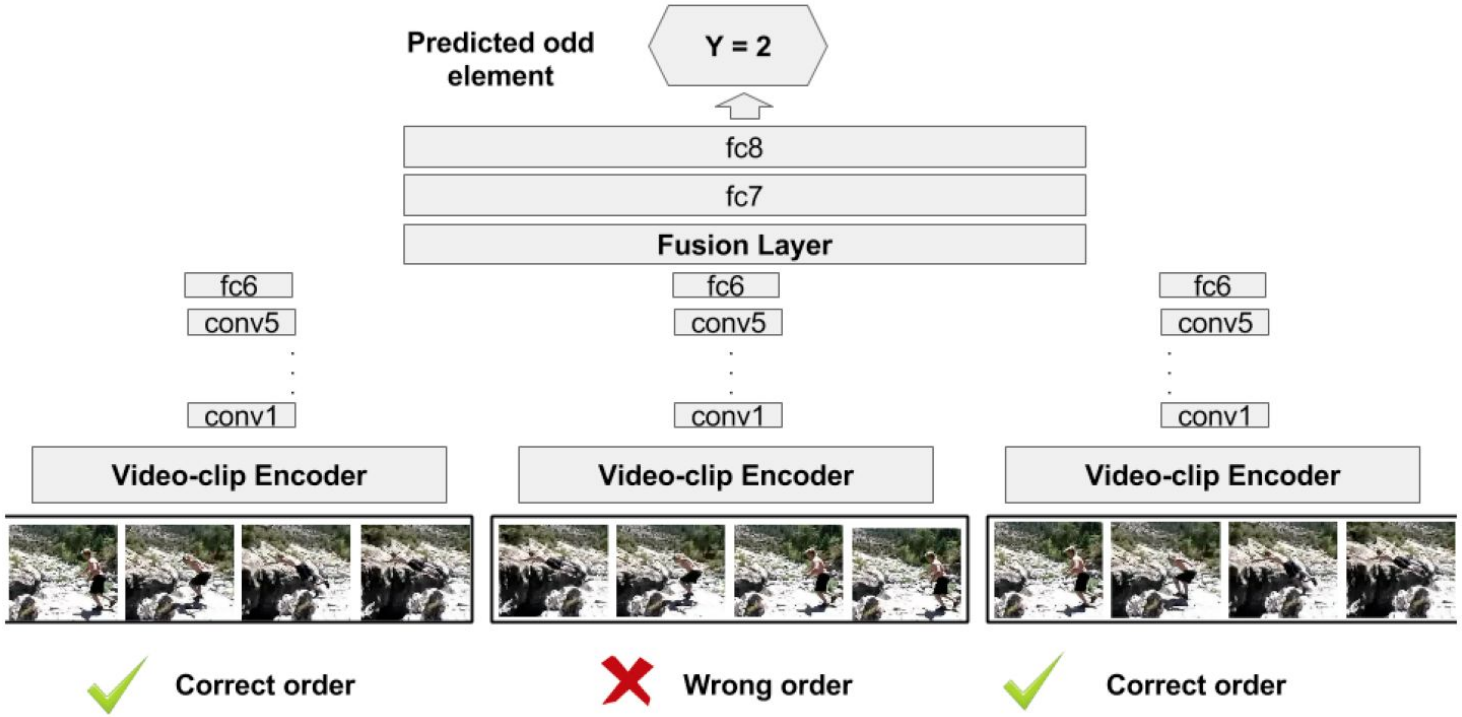
In order

Not in order

(Slides by Xunyu Lin): Misra, Ishan, C. Lawrence Zitnick, and Martial Hebert. "Shuffle and learn: unsupervised learning using temporal order verification." ECCV 2016. [code]

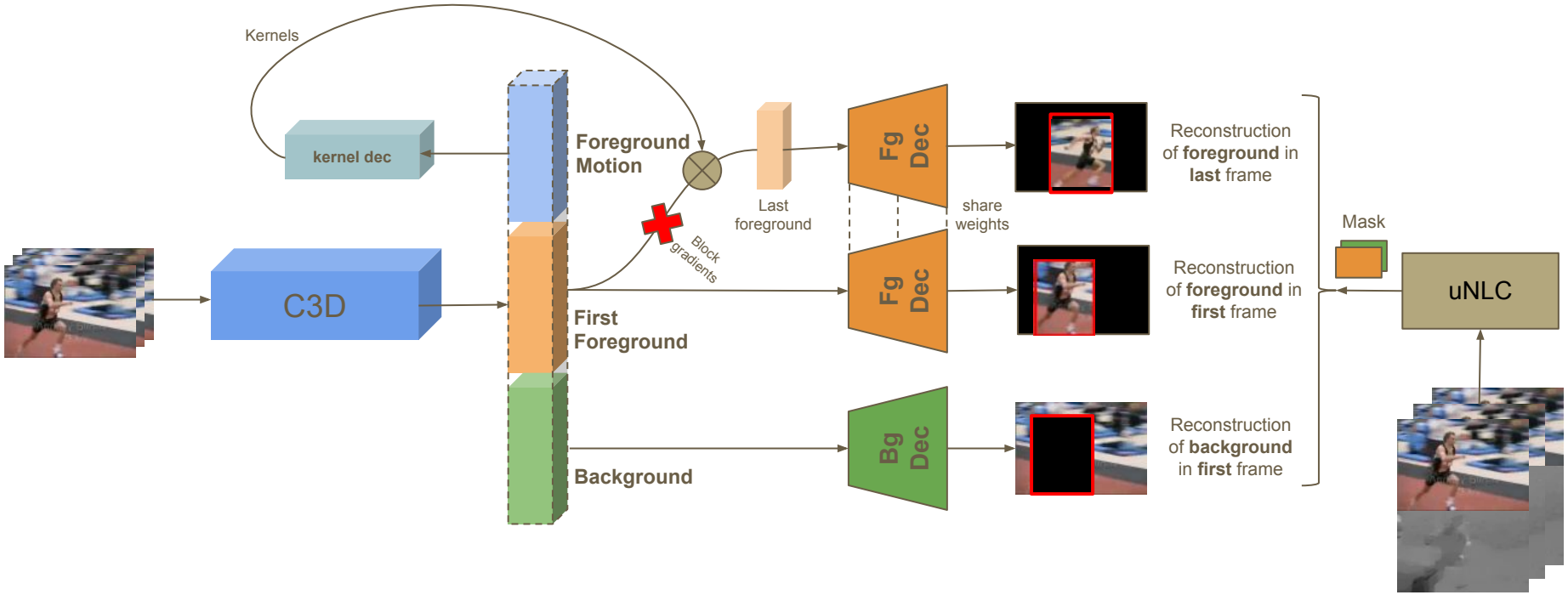
Temporal Weak Labels

Train a network to detect which of the video sequences contains frames wrong order.



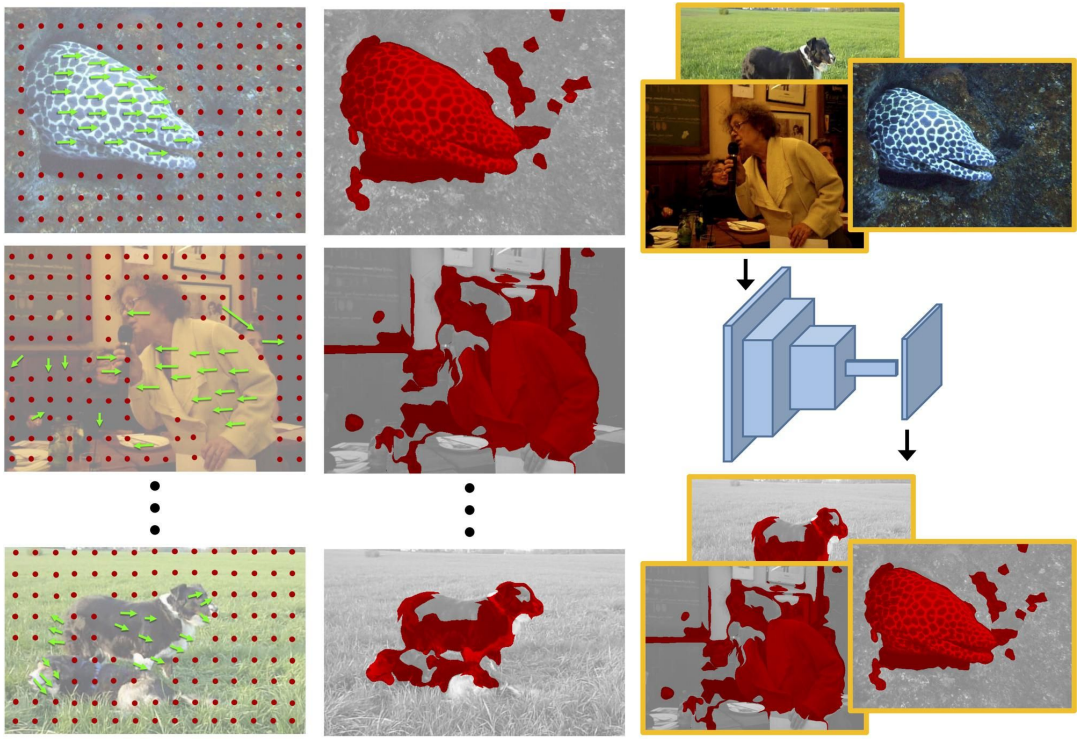
Fernando, Basura, Hakan Bilen, Efstratios Gavves, and Stephen Gould. "Self-supervised video representation learning with odd-one-out networks." CVPR 2017

Spatio-Temporal Weak Labels



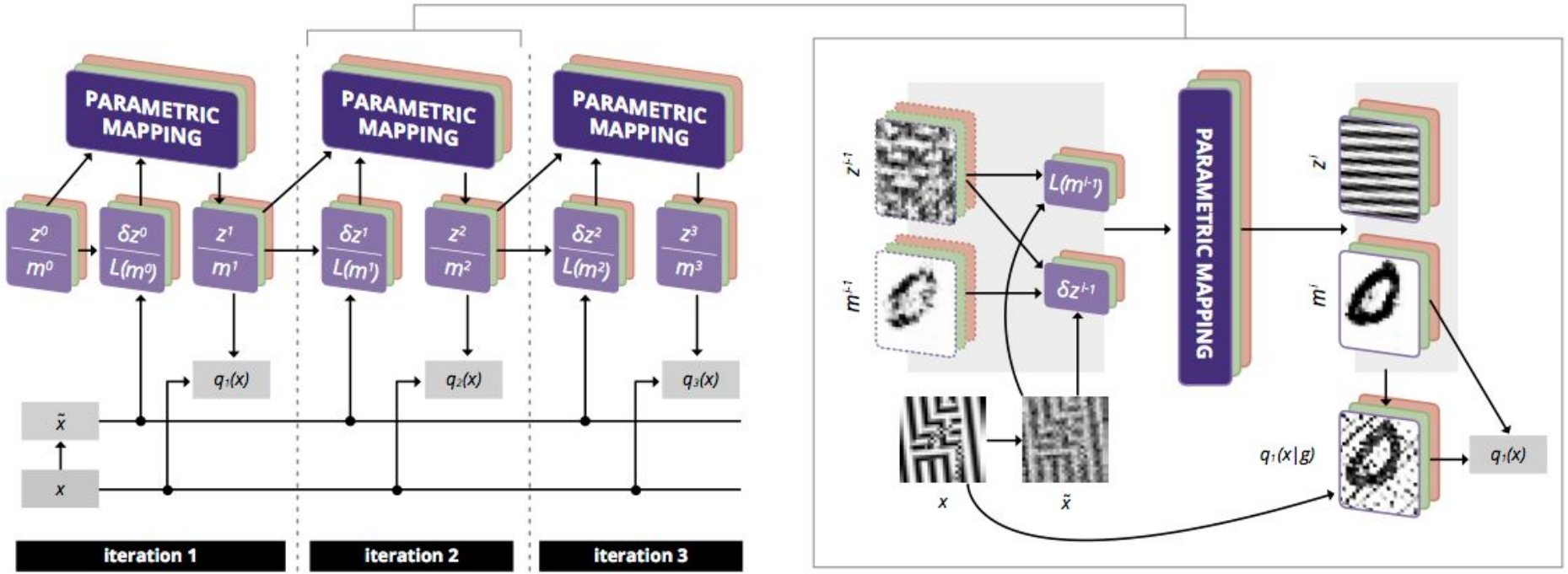
X. Lin, Campos, V., Giró-i-Nieto, X., Torres, J., and Canton-Ferrer, C., "**Disentangling Motion, Foreground and Background Features in Videos**", in *CVPR 2017 Workshop Brave New Motion Representations*

Spatio-Temporal Weak Labels



Pathak, Deepak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. "Learning features by watching objects move." CVPR 2017

Spatio-Temporal Weak Labels



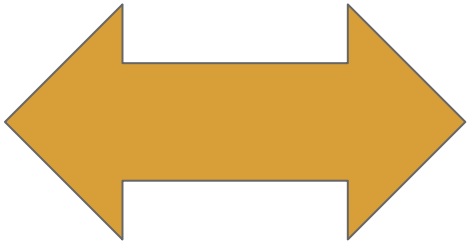
Greff, Klaus, Antti Rasmus, Mathias Berglund, Tele Hao, Harri Valpola, and Juergen Schmidhuber. "Tagger: Deep unsupervised perceptual grouping." NIPS 2016 [video] [code]

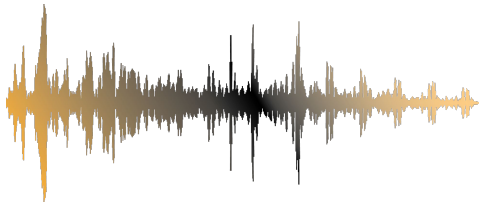
Outline

1. Unsupervised Learning
2. Predictive Learning
3. Self-supervised Learning
4. **Cross-modal Learning**
    - Feature Learning
    - Cross-modal Retrieval
    - Cross-modal Translation

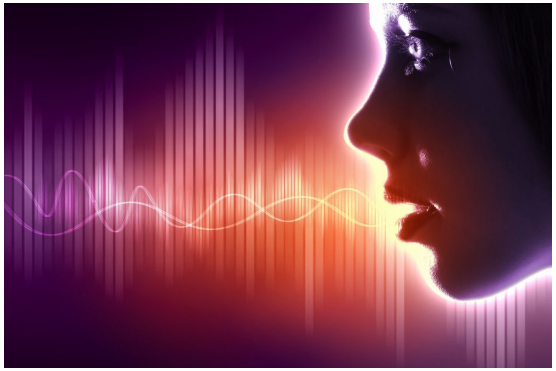
Cross-modal Learning

Vision



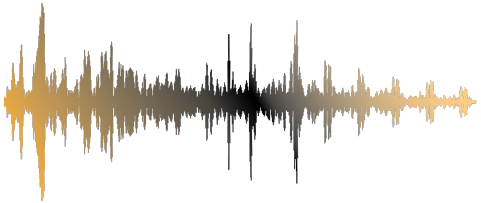


Audio



Speech

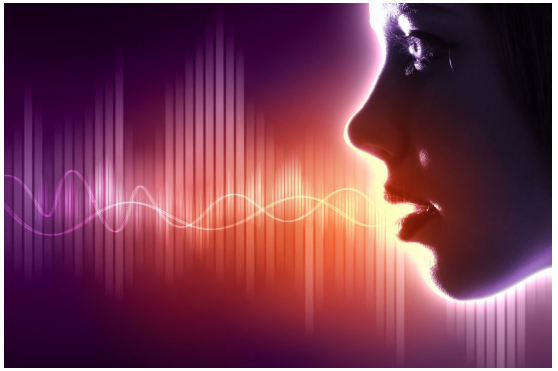
Cross-modal Learning



Audio

Vision

Video



Speech

Synchronization among modalities captured by **video** is exploited in a self-supervised manner.

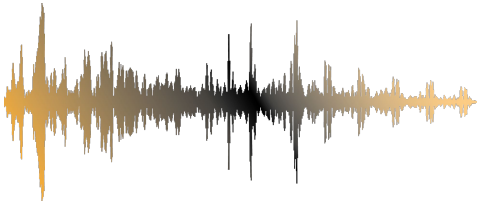
Outline

1. Unsupervised Learning
2. Predictive Learning
3. Self-supervised Learning
4. Cross-modal Learning
    - **Feature Learning**
    - Cross-modal Retrieval
    - Cross-modal Translation

Visual Feature Learning

Vision

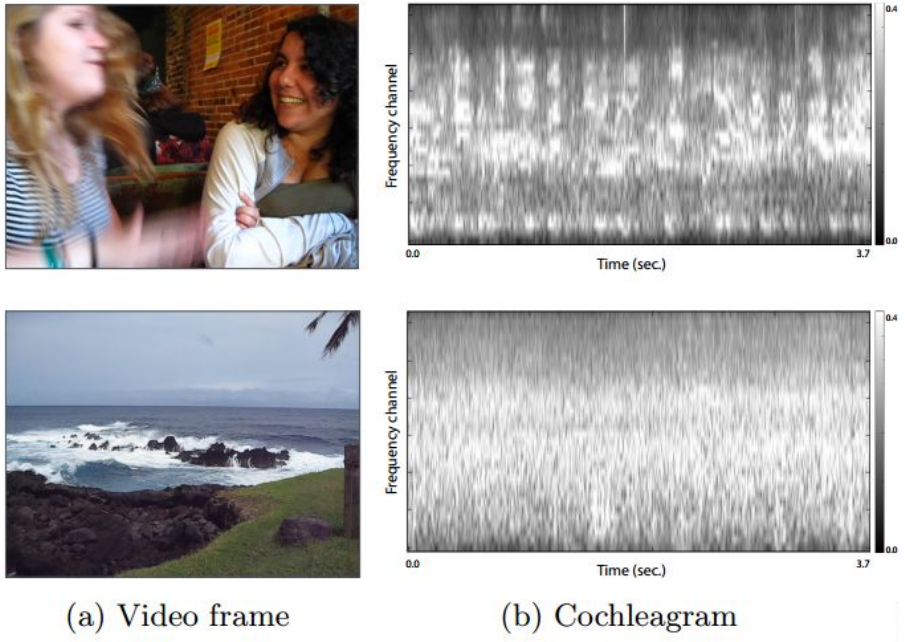
Video



Audio

Visual Feature Learning

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

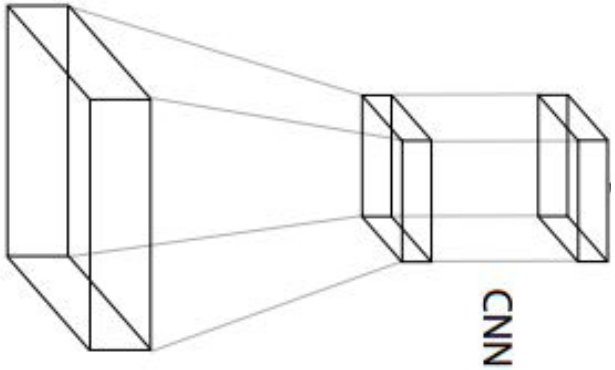


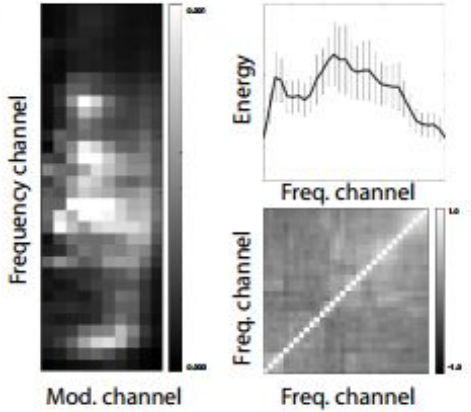
(a) Video frame (b) Cochleagram

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.



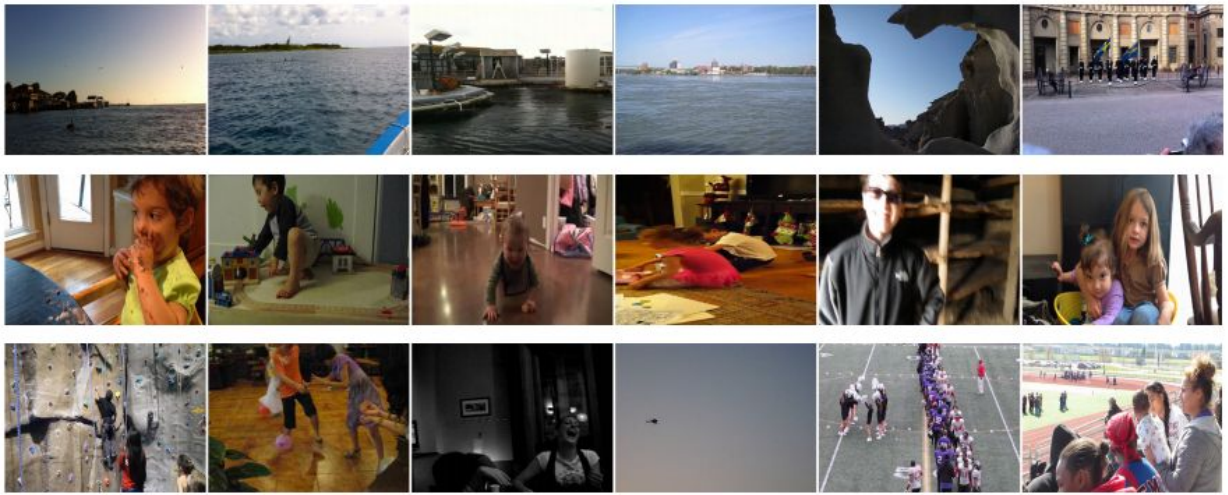


Owens, Andrew, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. "Ambient sound provides supervision for visual learning." ECCV 2016

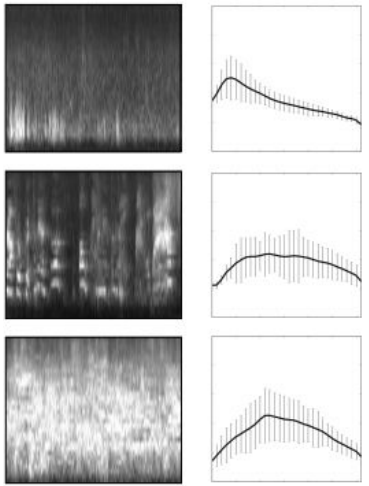
Visual Feature Learning

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)



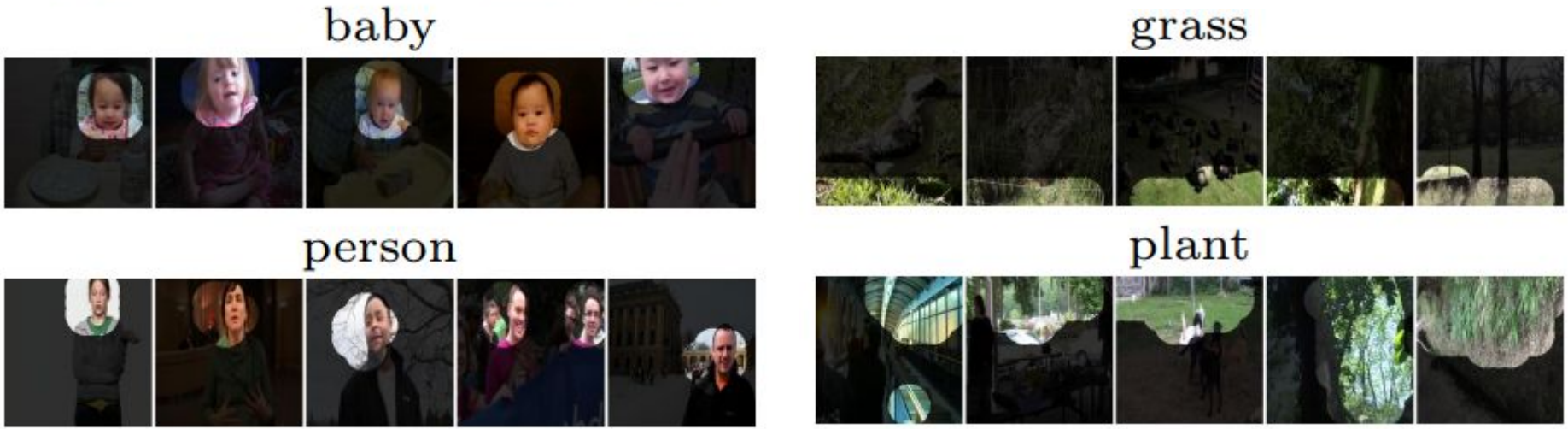
Average stats



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.

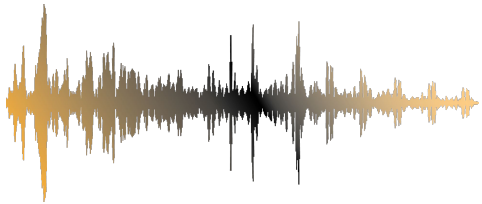


Owens, Andrew, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. "Ambient sound provides supervision for visual learning." ECCV 2016

Audio Feature Learning

Vision

Video



Audio

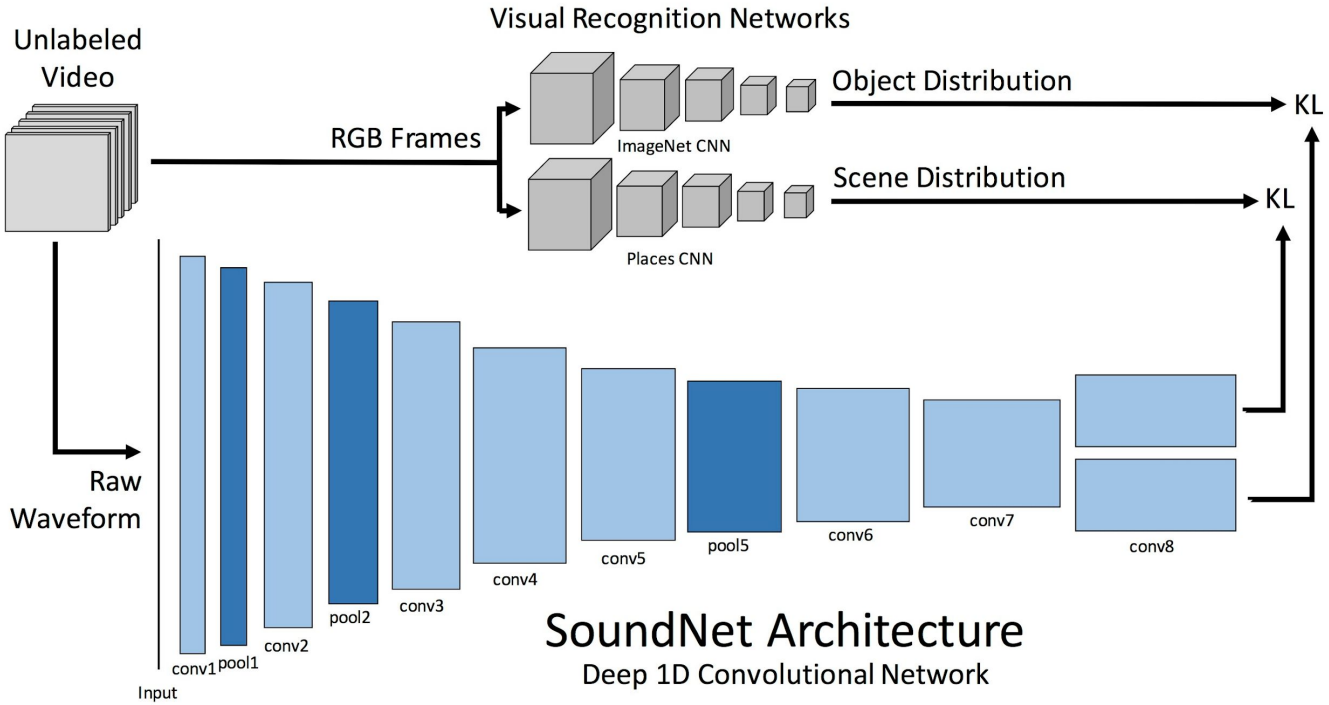
Predicted Objects and Scenes from Sound Only

(Videos are blurred so you can try to recognize yourself!)

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

Audio Feature Learning: SoundNet

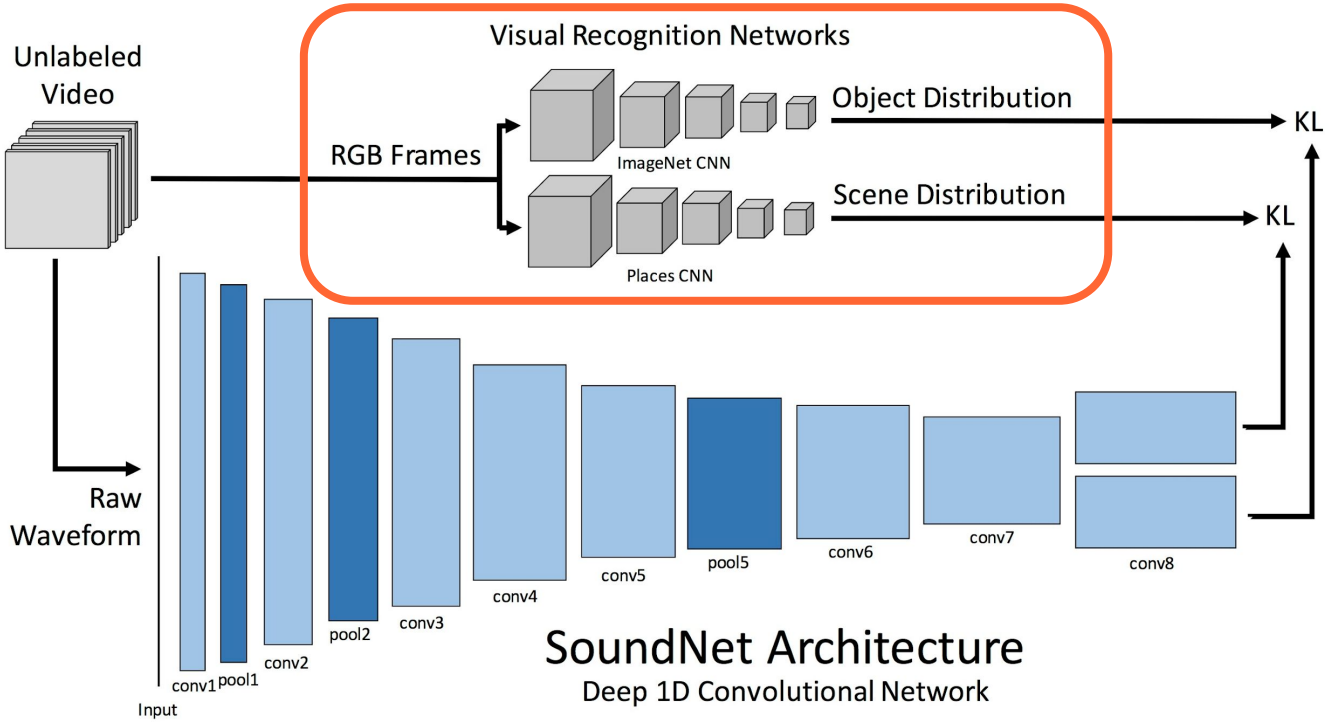
Pretrained visual ConvNets supervise the training of a model for sound representation



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

Audio Feature Learning: SoundNet

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: SoundNet

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%.

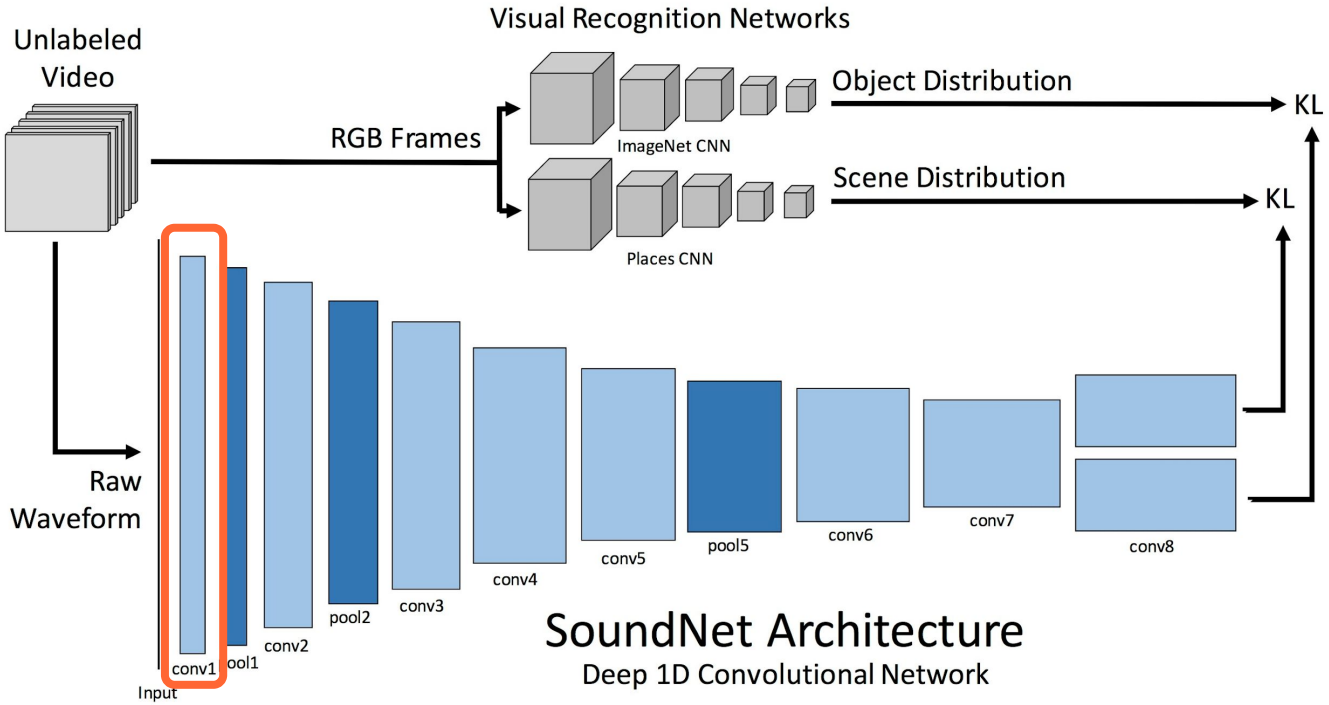
Method	Accuracy on ESC-50	Accuracy on ESC-10
SVM-MFCC [28]	39.6%	67.5%
Convolutional Autoencoder	39.9%	74.3%
Random Forest [28]	44.3%	72.7%
Piczak ConvNet [27]	64.5%	81.0%
SoundNet	**74.2%**	**92.2%**
Human Performance [28]	81.3%	95.7%

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: SoundNet

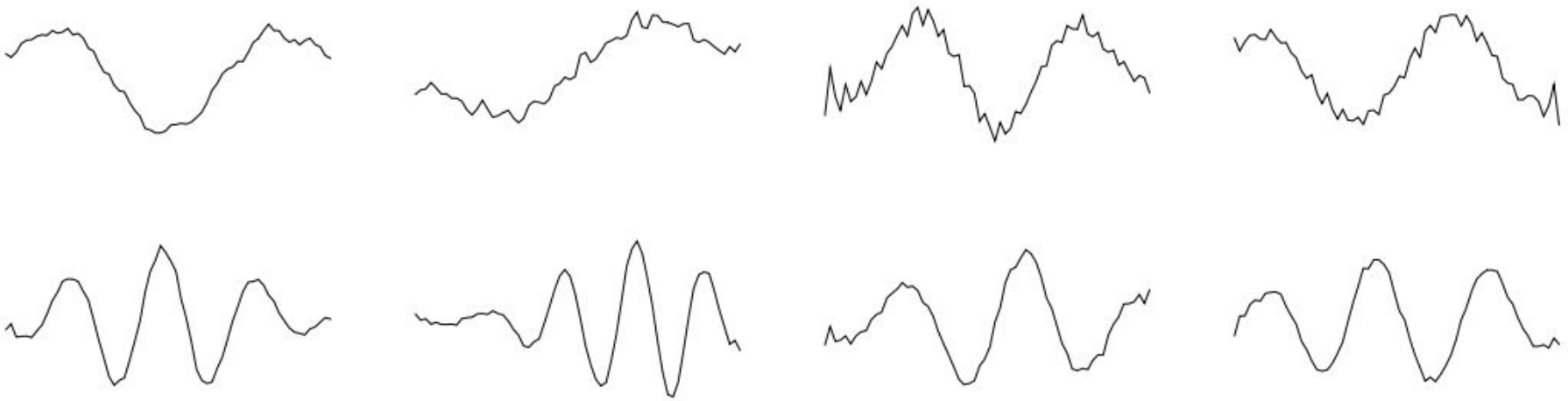
Visualization of the 1D filters over raw audio in conv1.



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

Audio Feature Learning: SoundNet

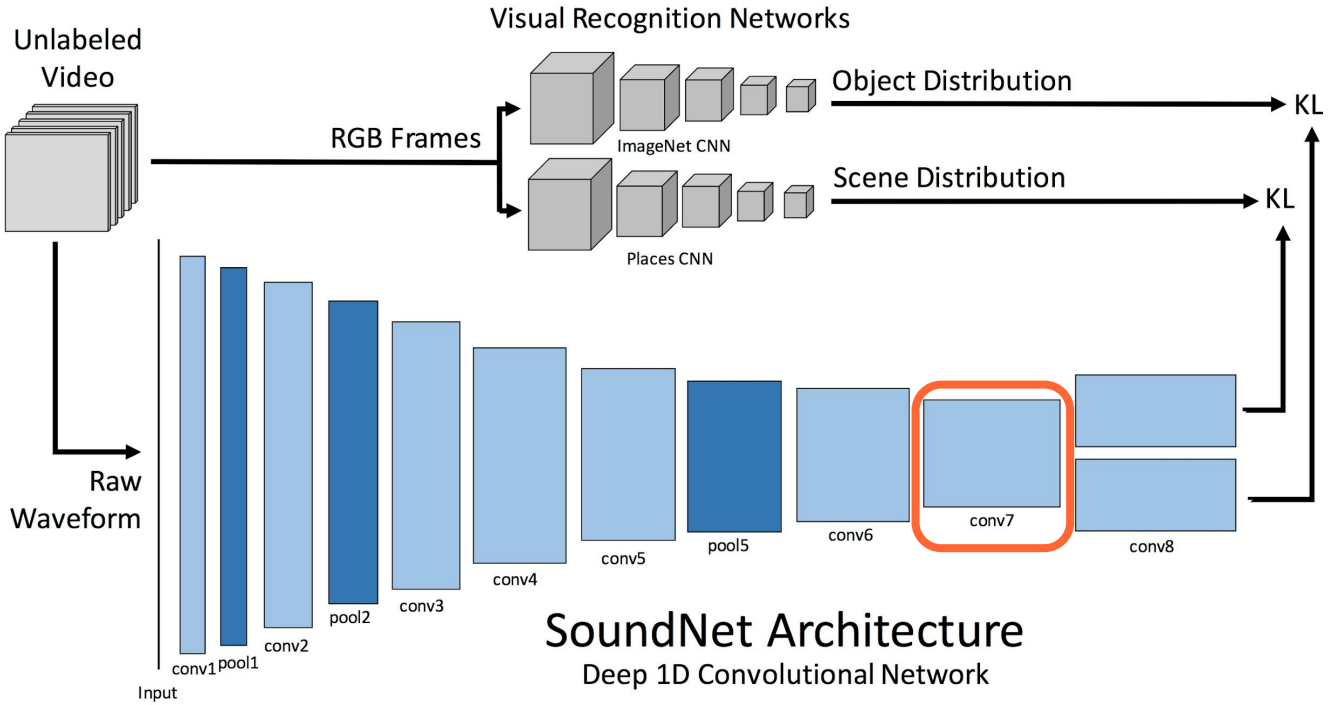
Visualization of the 1D filters over raw audio in conv1.



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

Audio Feature Learning: SoundNet

Visualize samples that mostly activate a neuron in a late layer (conv7)



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

Audio Feature Learning: SoundNet

Visualization of the video frames associated to the sounds that activate some of the last hidden units (conv7):

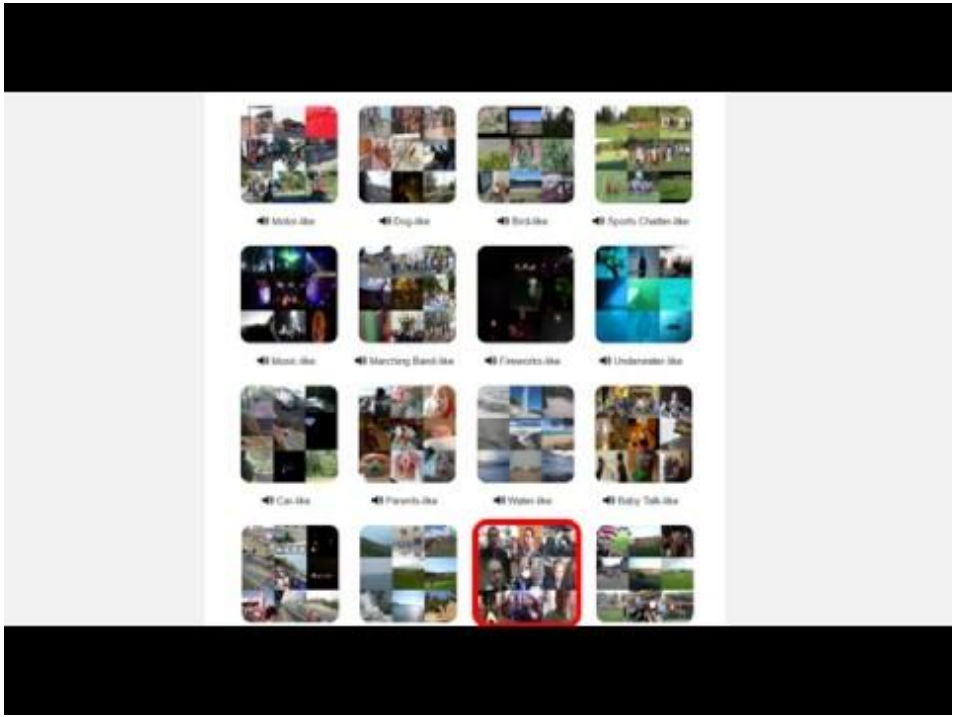
Baby Talk

Bubbles

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

Audio Feature Learning: SoundNet

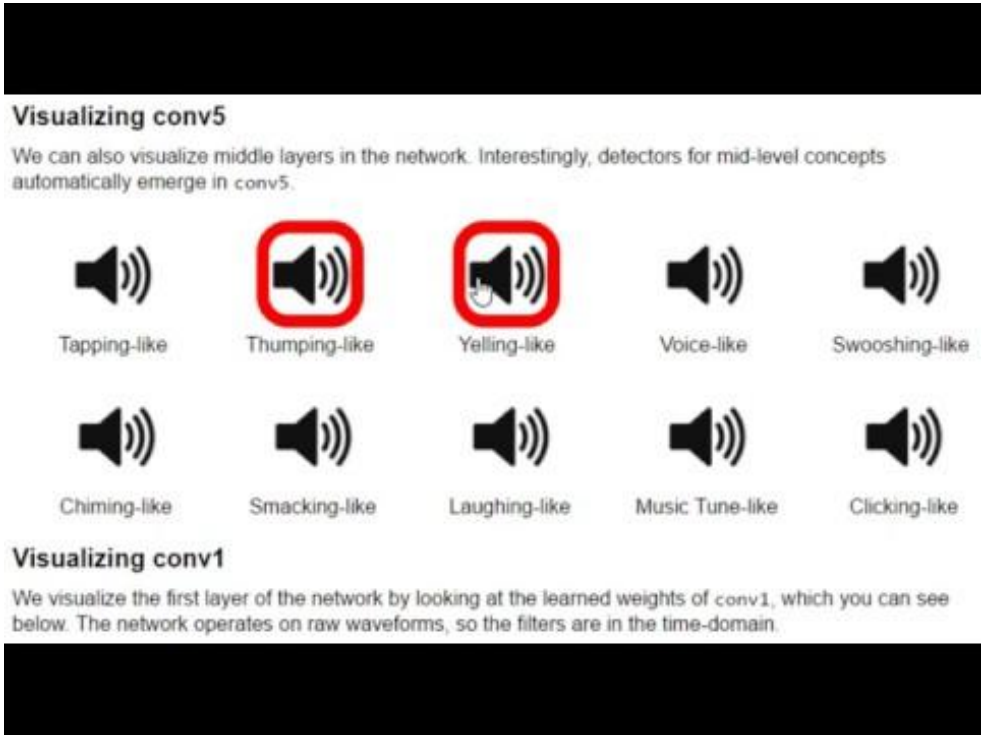
Hearing sounds that most activate a neuron in the sound network (conv7)



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

Audio Feature Learning: SoundNet

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

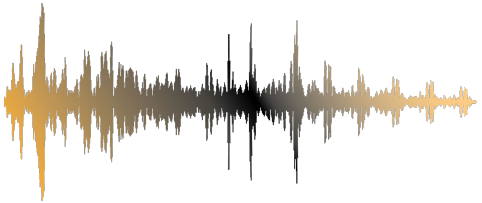


Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." NIPS 2016

Audio & Visual Feature Learning

Vision

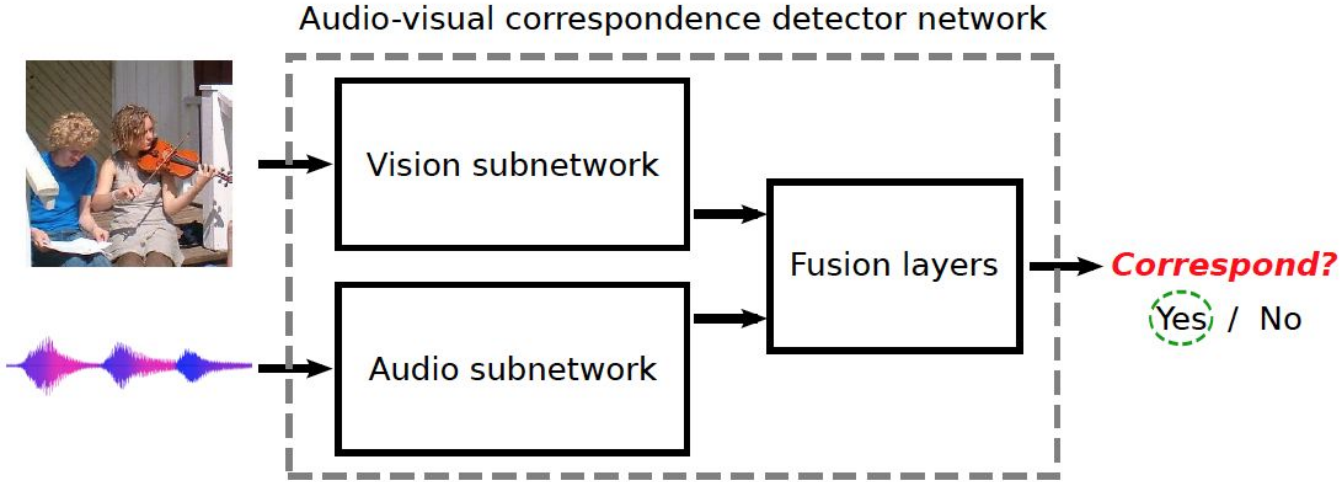
Video

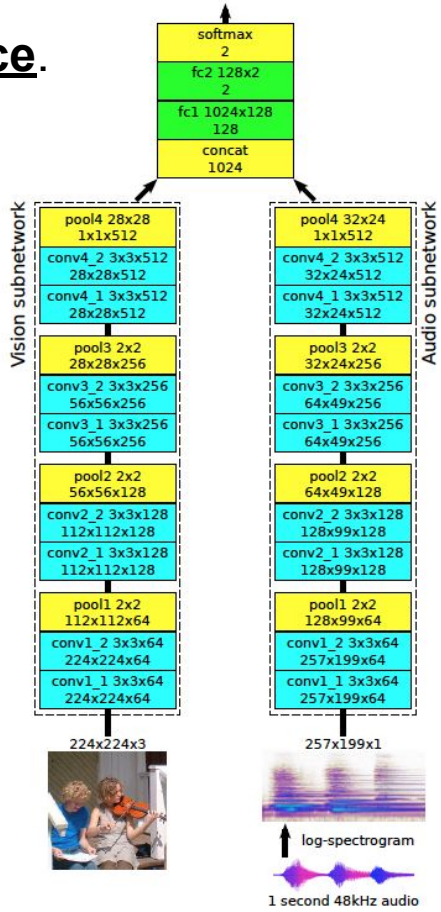


Audio

Audio & Visual Feature Learning

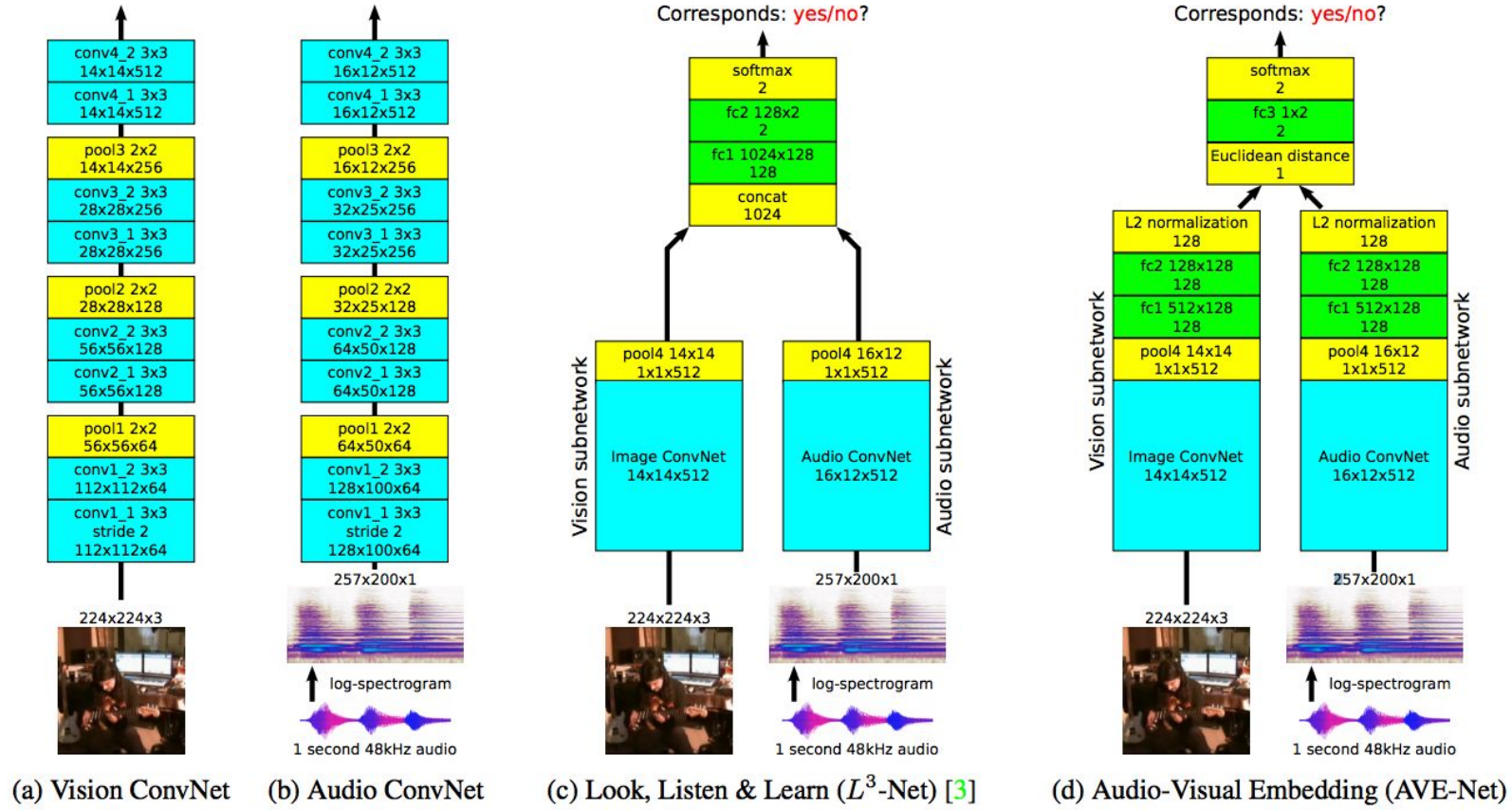
Audio and visual features learned by assessing **correspondence**.





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

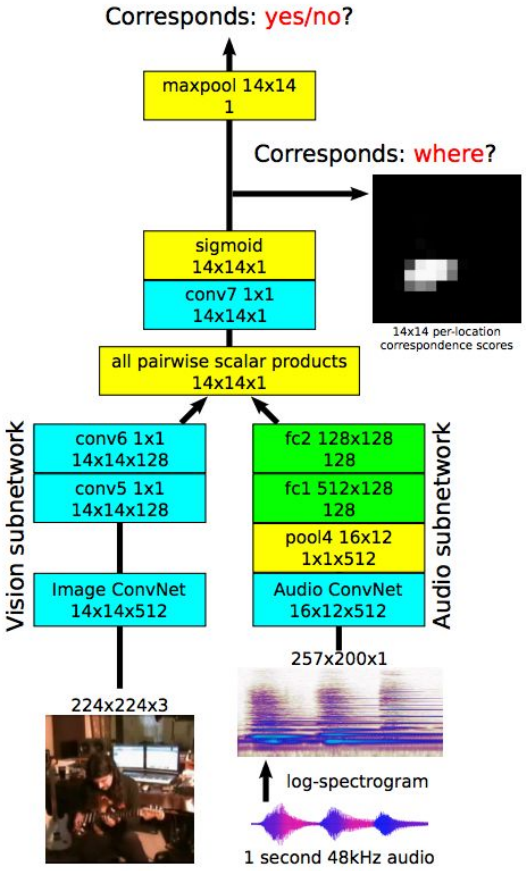
Audio & Visual Feature Learning

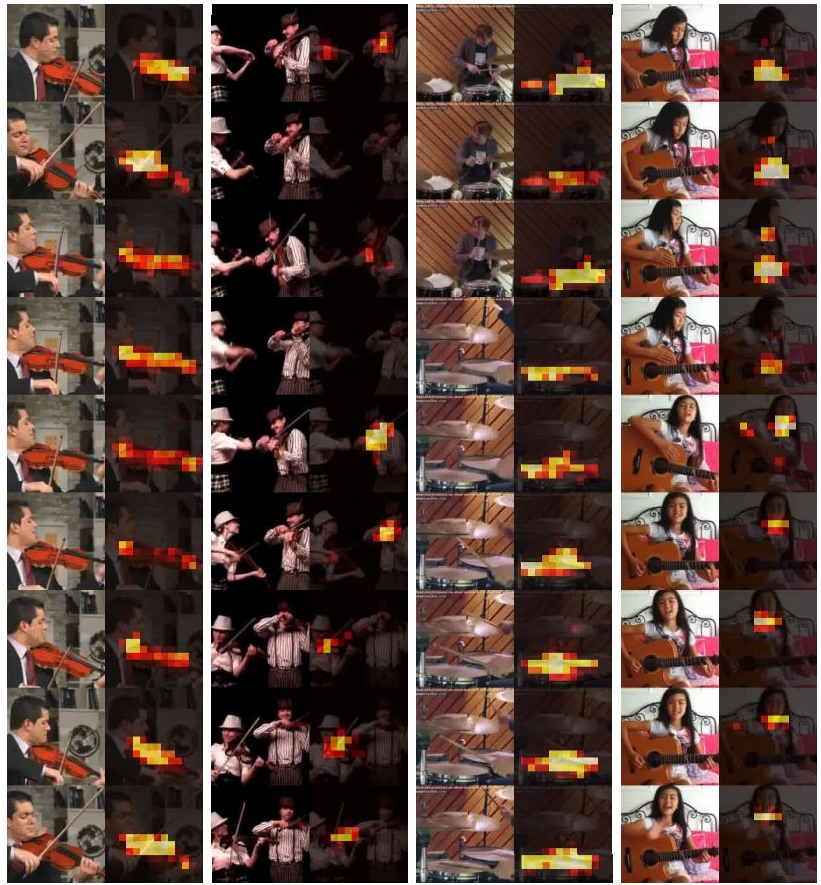


(a) Vision ConvNet (b) Audio ConvNet (c) Look, Listen & Learn (L^3-Net) [3] (d) Audio-Visual Embedding (AVE-Net)

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

Audio & Visual Feature Learning





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

Objects that Sound

Relja Arandjelović[1] , Andrew Zisserman[1,2]
[1]DeepMind [2]University of Oxford

Frames are processed completely independently, motion information is not used, and there is no temporal smoothing

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

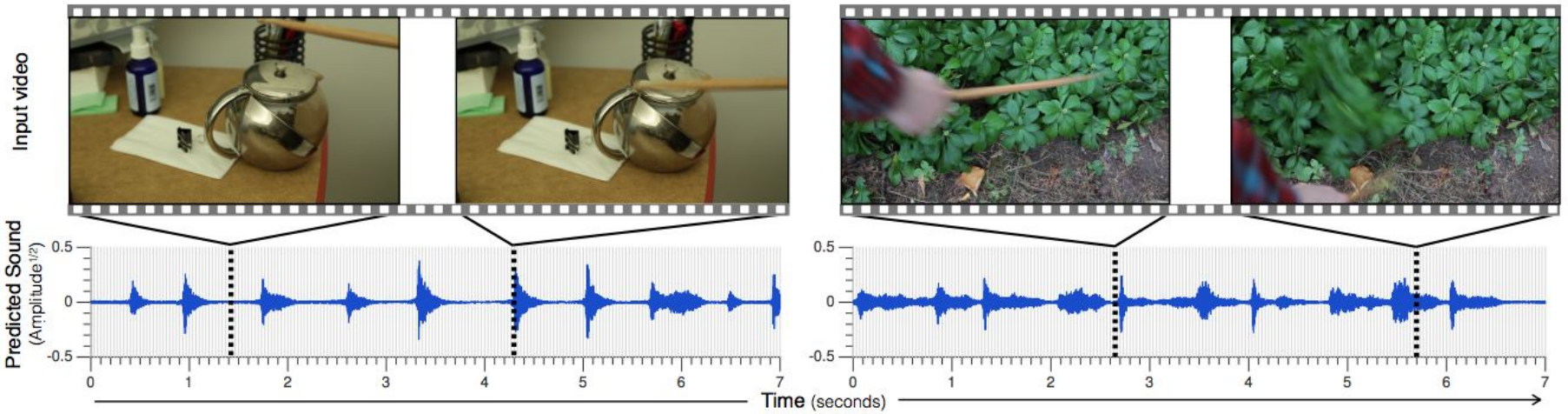
Outline

1. Unsupervised Learning
2. Predictive Learning
3. Self-supervised Learning
4. Cross-modal Learning
    - Feature Learning
    - **Cross-modal Retrieval**
    - Cross-modal Translation

Owens, Andrew, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman. "Visually indicated sounds." CVPR 2016.

Cross-modal Retrieval

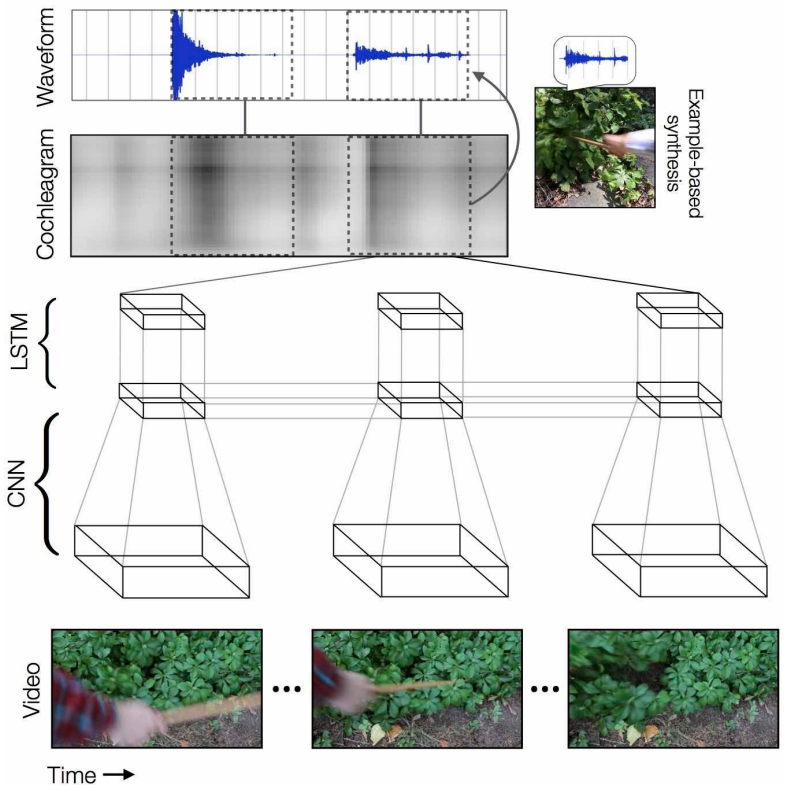
Learn synthesized sounds from videos of people hitting objects with a drumstick.



Owens, Andrew, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman. "Visually indicated sounds." CVPR 2016.

Cross-modal Retrieval

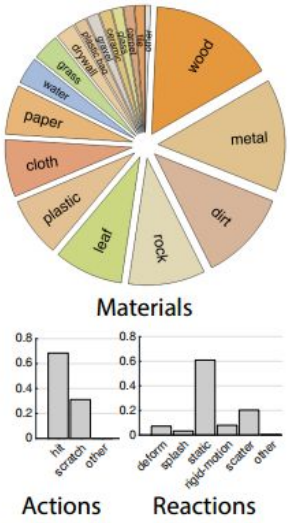
Not end-to-end



Owens, Andrew, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman. "Visually indicated sounds." CVPR 2016.

Cross-modal Retrieval

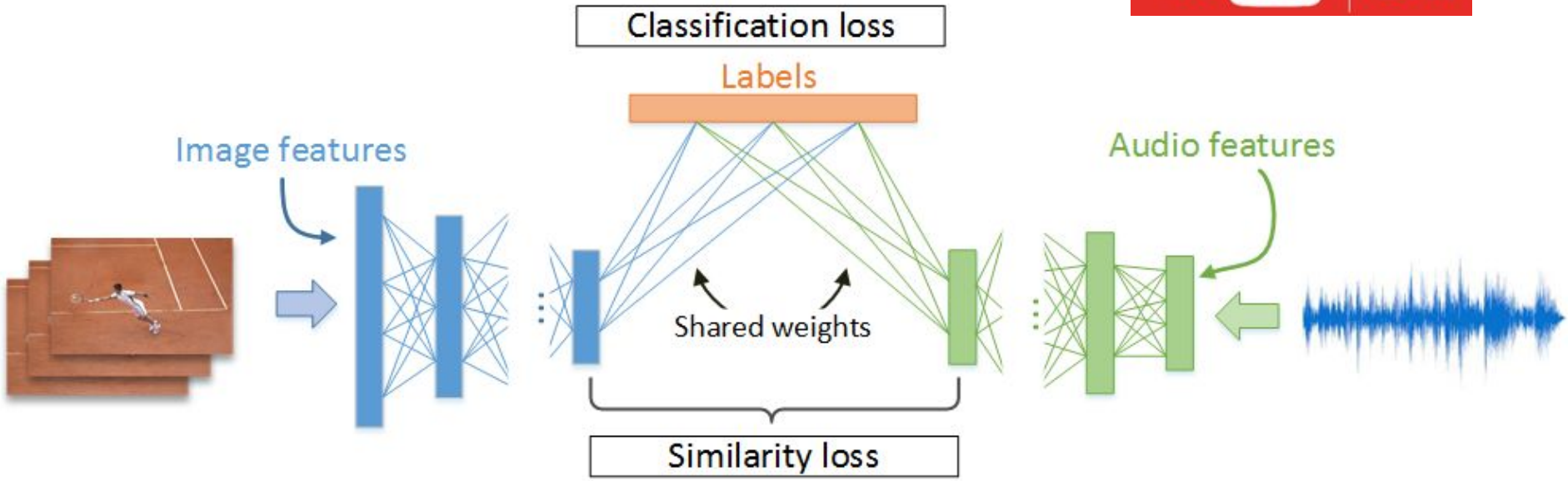
The Greatest Hits Dataset



Owens, Andrew, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman. "Visually indicated sounds." CVPR 2016.

Cross-modal Retrieval

[Paper draft]



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).

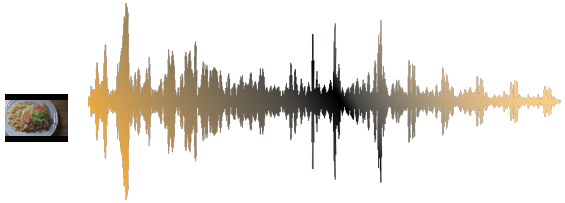
Cross-modal Retrieval

Video sonorization

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).

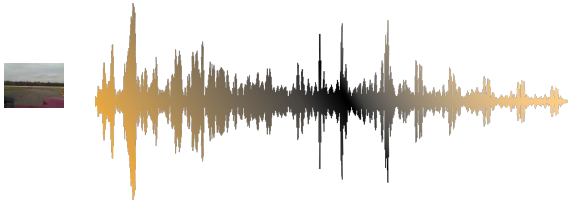
Cross-modal Retrieval

Audio coloring

Visual feature

Audio feature

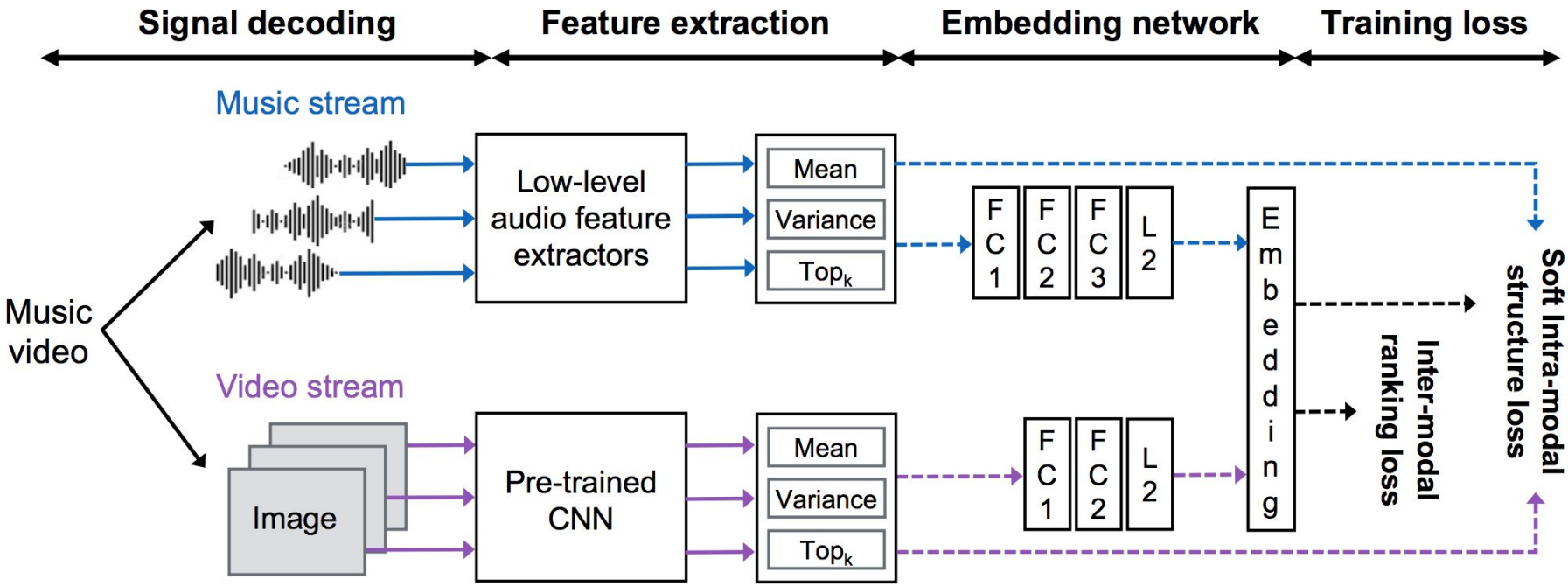
Best match



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).

Hong, Sungeun, Woobin Im, and Hyun S. Yang. "Deep Learning for Content-Based, Cross-Modal Retrieval of Videos and Music." (submitted)

Cross-modal Retrieval



Hong, Sungeun, Woobin Im, and Hyun S. Yang. "Deep Learning for Content-Based, Cross-Modal Retrieval of Videos and Music." 2017 (submitted)

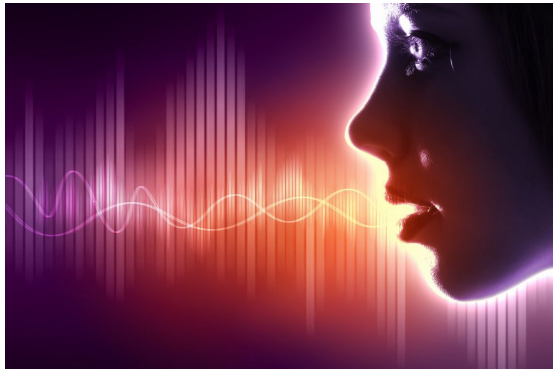
Outline

1. Unsupervised Learning
2. Predictive Learning
3. Self-supervised Learning
4. Cross-modal Learning
    - Feature Learning
    - Cross-modal Retrieval
    - **Cross-modal Translation**

Cross-modal Translation

Vision

Video

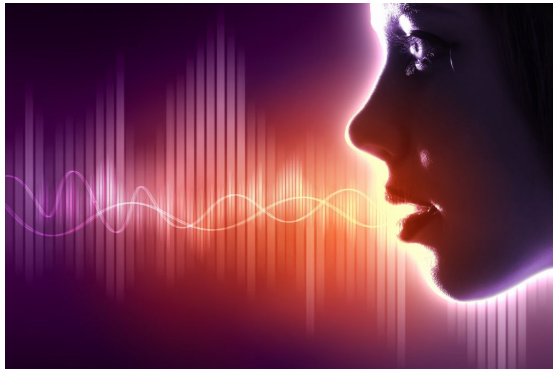


Speech

Cross-modal Translation

Vision

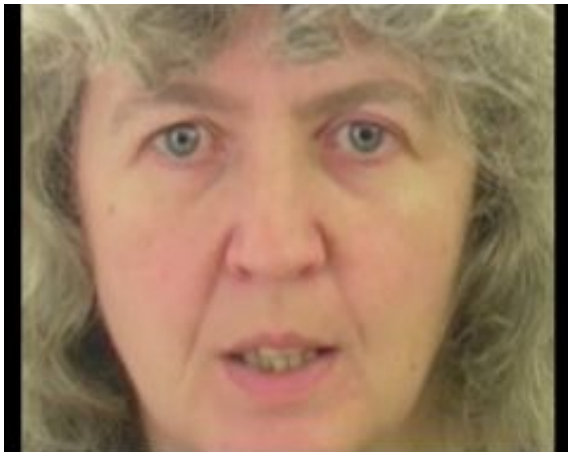
Video



Speech

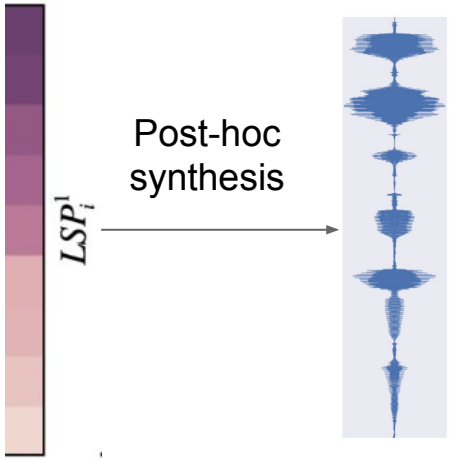
Ephrat, Ariel, Tavi Halperin, and Shmuel Peleg. "Improved speech reconstruction from silent video." In ICCV 2017 Workshop on Computer Vision for Audio-Visual Media. 2017.

Speech Generation from Video



Frame from a silent video

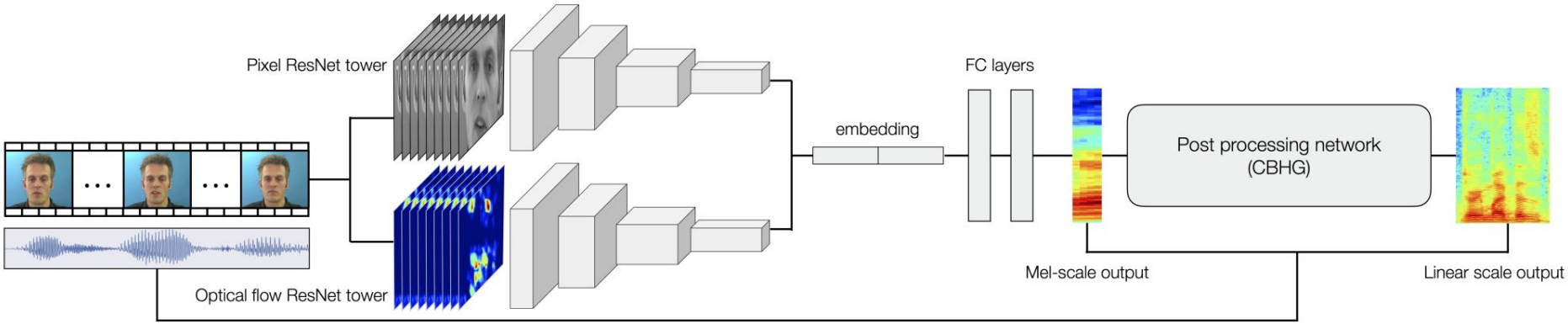
CNN (VGG)



Audio feature

Ephrat et al. Vid2speech: Speech Reconstruction from Silent Video. *ICASSP 2017*

Speech Generation from Video

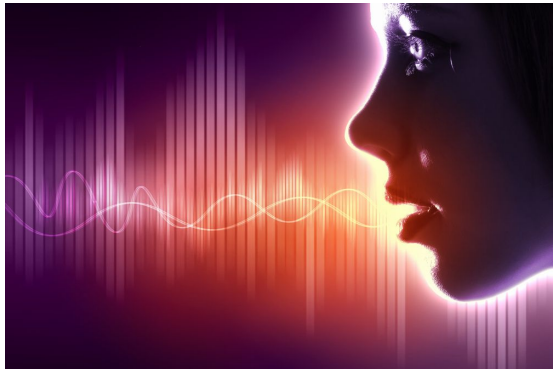


Ephrat, Ariel, Tavi Halperin, and Shmuel Peleg. "Improved speech reconstruction from silent video." In ICCV 2017 Workshop on Computer Vision for Audio-Visual Media. 2017.

Cross-modal Translation

Vision

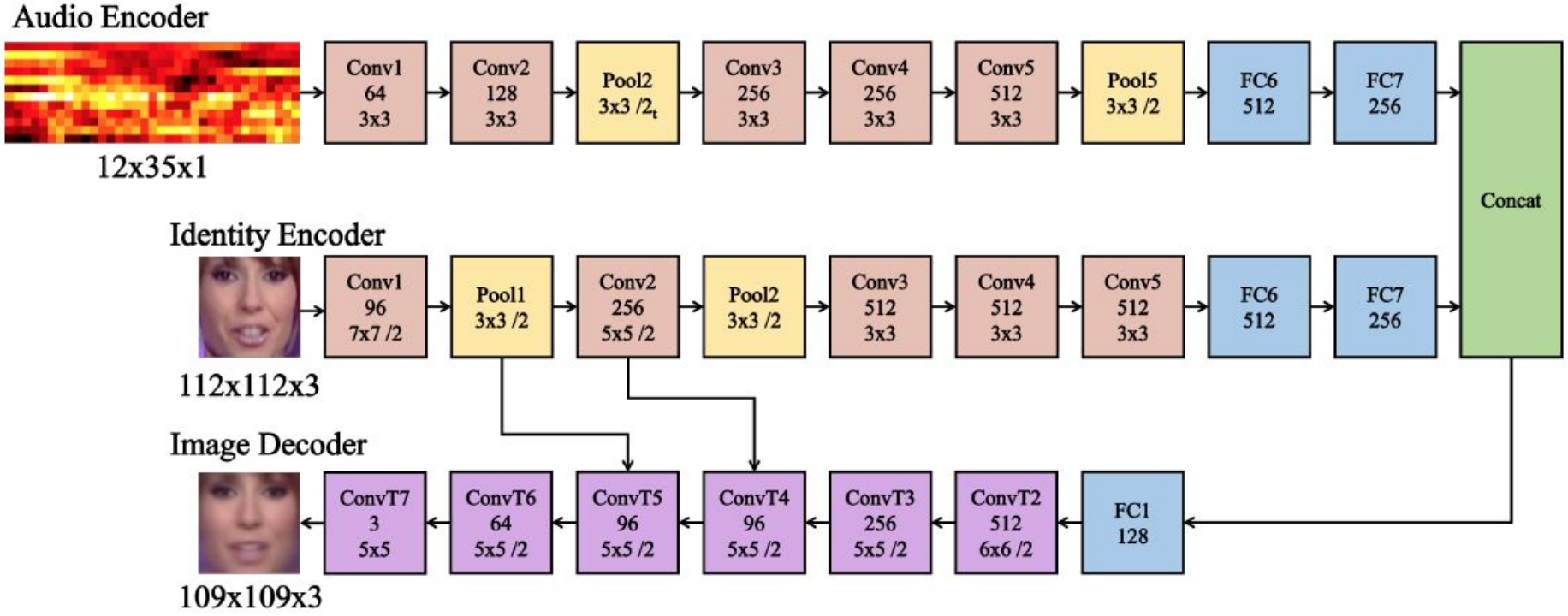
Video



Speech

Chung, Joon Son, Amir Jamaludin, and Andrew Zisserman. "You said that?." BMVC 2017.

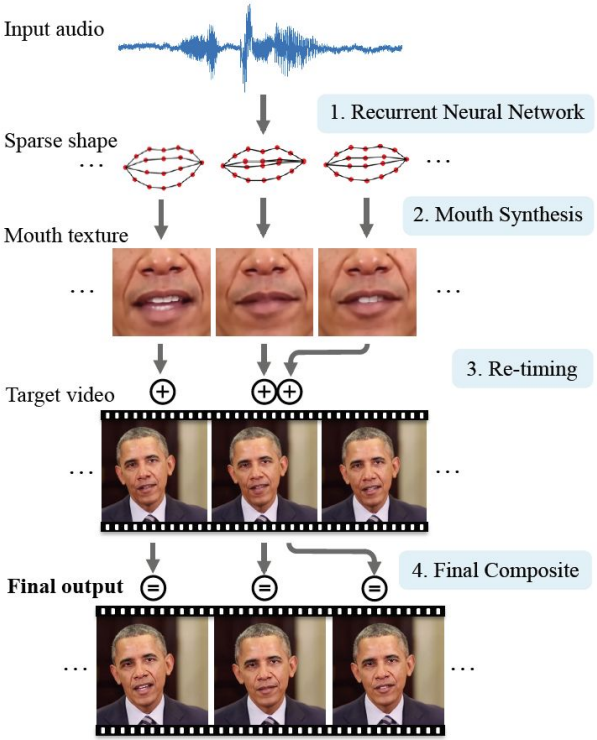
Speech to Video Synthesis (mouth)



Chung, Joon Son, Amir Jamaludin, and Andrew Zisserman. "You said that?." BMVC 2017.

Karras, Tero, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. "Audio-driven facial animation by joint end-to-end learning of pose and emotion." SIGGRAPH 2017

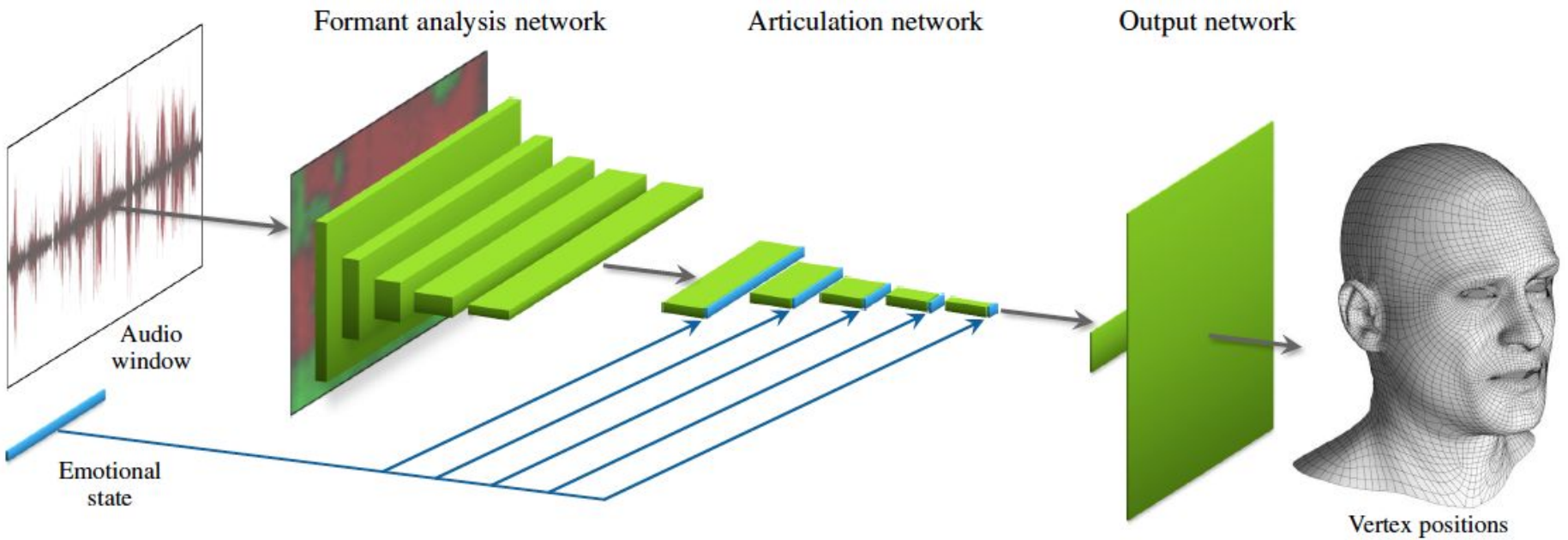
Speech to Video Synthesis (mouth)



Suwajanakorn, Supasorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. "Synthesizing Obama: learning lip sync from audio." SIGGRAPH 2017.

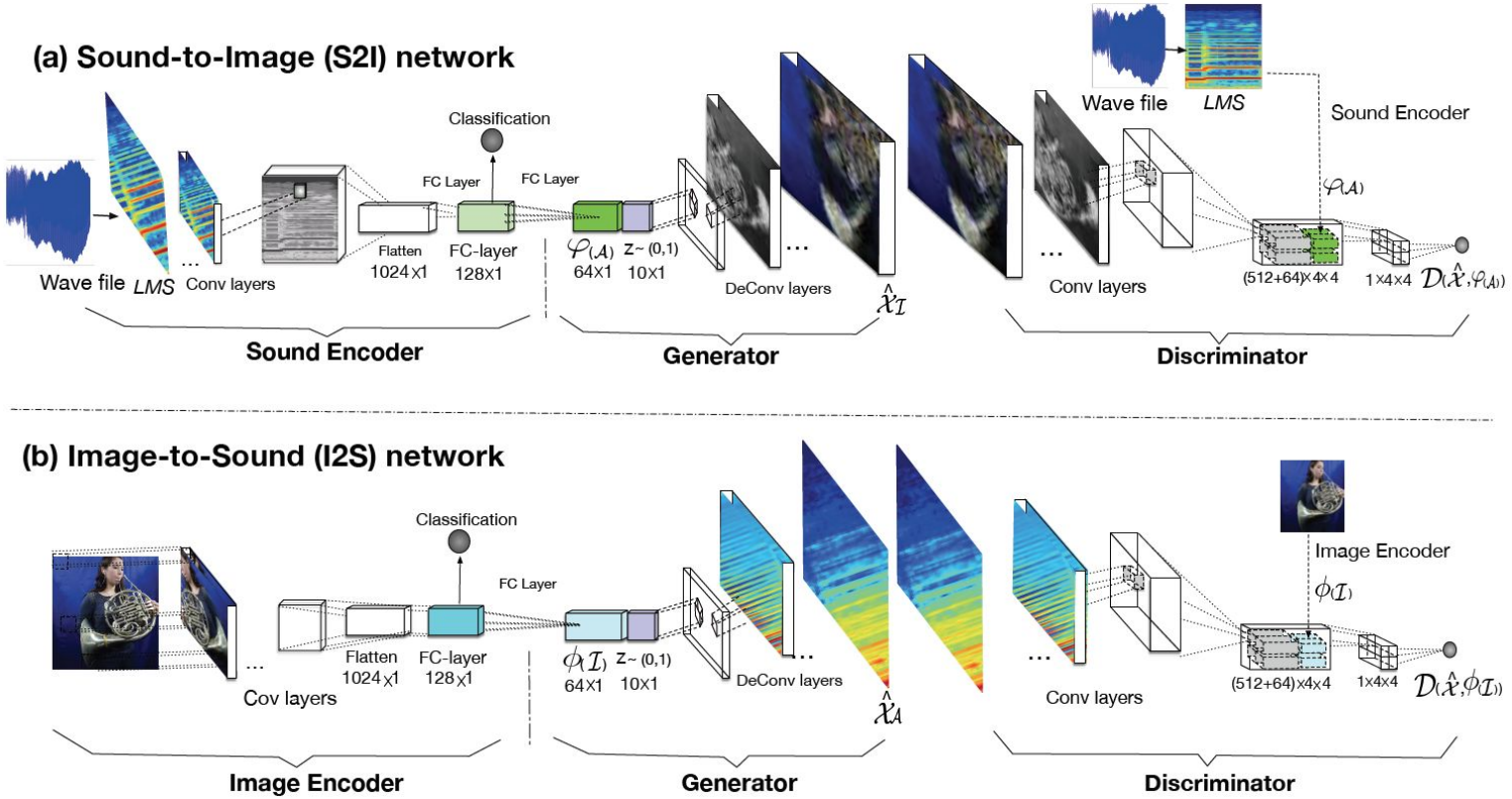
Karras, Tero, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. "Audio-driven facial animation by joint end-to-end learning of pose and emotion." SIGGRAPH 2017

Speech to Video Synthesis (pose & emotion)



Karras, Tero, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. "Audio-driven facial animation by joint end-to-end learning of pose and emotion." SIGGRAPH 2017

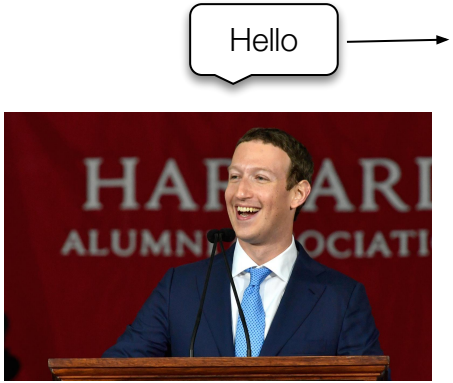
Audio & Visual Generation



L. Chen, S. Srivastava, Z. Duan and C. Xu. Deep Cross-Modal Audio-Visual Generation. ACM Multimedia Thematic Workshops 2017.

Speech2Signs (under work)

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Signal Theory and Communications
Image Processing Group



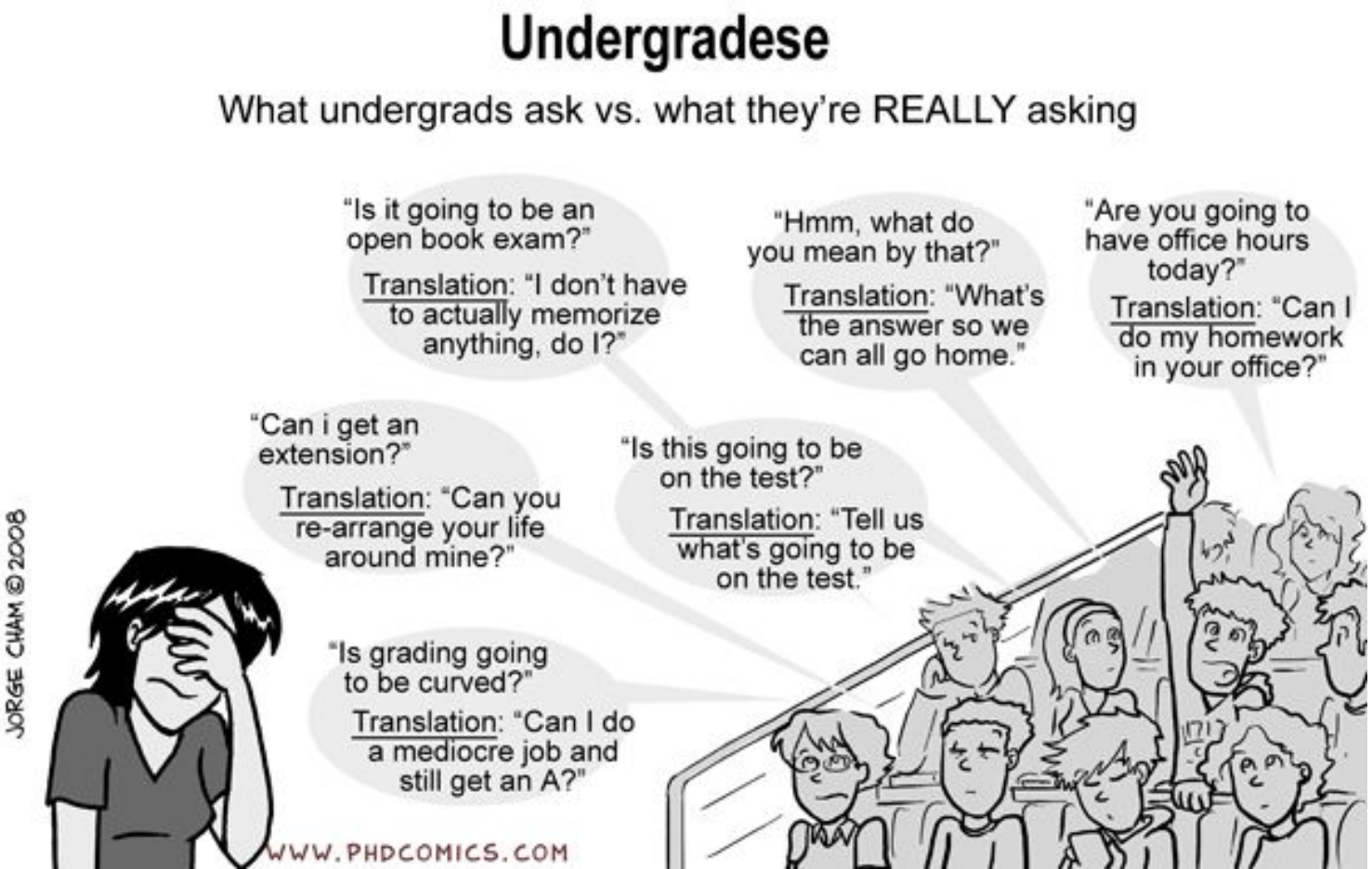
Hello → ASR → "Hello" → NMT → SLPA → Video Generator

Caffe2 facebook research

Outline

1. Unsupervised Learning
2. Predictive Learning
3. Self-supervised Learning
4. Cross-modal Learning

Questions ?


Undergradese
What undergrads ask vs. what they're REALLY asking
"Is it going to be an open book exam?"
Translation: "I don't have to actually memorize anything, do I?"
"Hmm, what do you mean by that?"
Translation: "What's the answer so we can all go home."
"Are you going to have office hours today?"
Translation: "Can I do my homework in your office?"
"Can i get an extension?"
Translation: "Can you re-arrange your life around mine?"
"Is this going to be on the test?"
Translation: "Tell us what's going to be on the test."
"Is grading going to be curved?"
Translation: "Can I do a mediocre job and still get an A?"
JORGE CHAM © 2008
WWW.PHDCOMICS.COM