

DEEP LEARNING FOR ARTIFICIAL INTELLIGENCE

Master Course UPC ETSETB TelecomBCN Barcelona. Autumn 2017.



Instructors



Xavier
Giró-i-Nieto



Marta R.
Costa-jussà



Jordi
Torres



Elisa
Sayrol



Santiago
Pascual
Verónica
Vilaplana
Ramon
Morros
Javier
Ruiz

Organizers



UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH



telecom
BCN



Barcelona
Supercomputing
Center
Centre Nacional de Supercomputació

+ info: <http://dlai.deeplearning.barcelona>

[\[course site\]](#)



#DLUPC

Day 9 Lecture 1

Unsupervised Learning



Xavier Giro-i-Nieto
xavier.giro@upc.edu

Associate Professor
Universitat Politècnica de Catalunya
Technical University of Catalonia



Acknowledgments



The slide title is "The manifold hypothesis". It states: "The data distribution lie close to a low-dimensional manifold". Below this, under "Example: consider image data", there is a bulleted list:

- 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
 - Manhattan distance is a great measure of similarity

Below the list is the text "Similar for audio and text". To the right of the text are two images: one showing four dark grey rectangular blocks and another showing three small thumbnail images of a landscape, a person, and a motorcycle.

At the bottom of the slide, there is a video frame showing a person standing and gesturing in front of a whiteboard.

At the bottom right of the slide area, there is a logo for UPC (Universitat Politècnica de Catalunya) and the text "UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH Departament de Teoria del Sinyal i Comunicacions".

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

Outline

1. Motivation
2. Unsupervised Learning
3. Predictive Learning
4. Self-supervised Learning

Outline

1. Motivation
2. Unsupervised Learning
3. Predictive Learning
4. Self-supervised Learning

Motivation

Alpha Go Takes the Match, 3-0 (i-programmer.info)



Posted by timothy on Saturday March 12, 2016 @10:24AM from the rhyming-singlet dept.



117

[mikejuk](#) writes:

Google's [AlphaGo has won the Deep Mind Challenge](#), by winning the third match in a row of five against the 18-time world champion Lee Se-dol. AlphaGo is now the number three Go player in the world and this is an event that will be remembered for a long time. Most AI experts thought that it would take decades to achieve but now we know that we have been on the right track since the 1980s or earlier. AlphaGo makes use of nothing dramatically new — it learned to play Go using a deep neural network and reinforcement learning, both developments on classical AI techniques. We know now that we don't need any big new breakthroughs to get to true AI. The results of the final two games are going to be interesting but as far as AI is concerned the match really is all over.

Motivation



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?



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

Motivation

Yann Lecun's Black Forest cake



■ "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)

Motivation

We can categorize three types of learning procedures:

1. Supervised Learning:

$$\mathbf{y} = f(\mathbf{x})$$

Predict label y corresponding to observation x

2. Unsupervised Learning:

$$f(\mathbf{x})$$

Estimate the distribution of observation x

3. Reinforcement Learning (RL):

$$\mathbf{y} = f(\mathbf{x})$$

$$\mathbf{z}$$

Predict action y based on observation x , to maximize a future reward z



Motivation

We can categorize three types of learning procedures:

1. Supervised Learning:

$$\mathbf{y} = f(\mathbf{x})$$

2. Unsupervised Learning:

$$f(\mathbf{x})$$

3. Reinforcement Learning (RL):

$$\mathbf{y} = f(\mathbf{x})$$

$$\mathbf{z}$$



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.

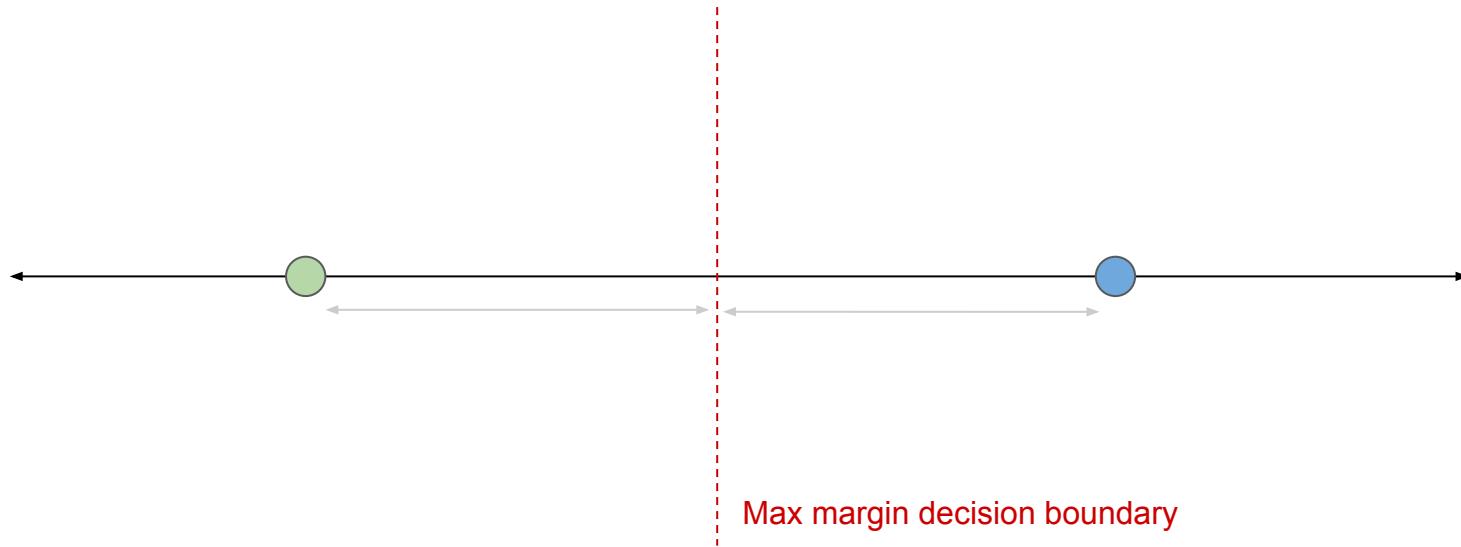
WHY?



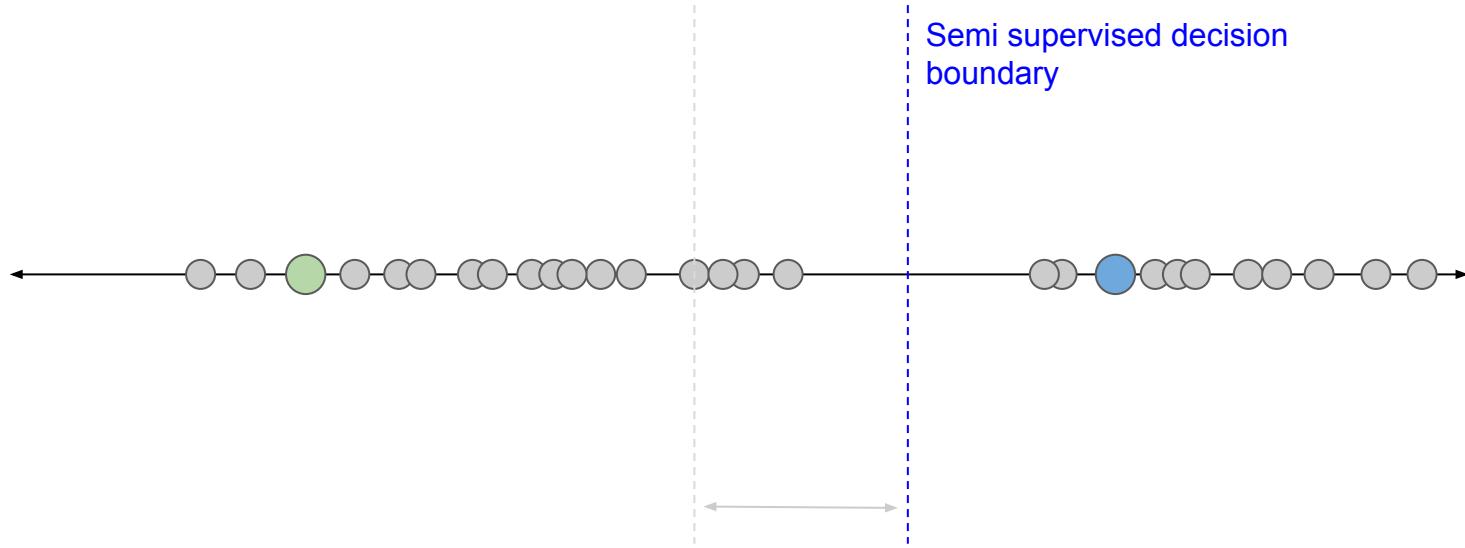
Outline

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

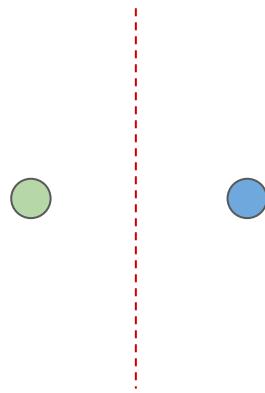
How data distribution $P(x)$ influences decisions (1D)



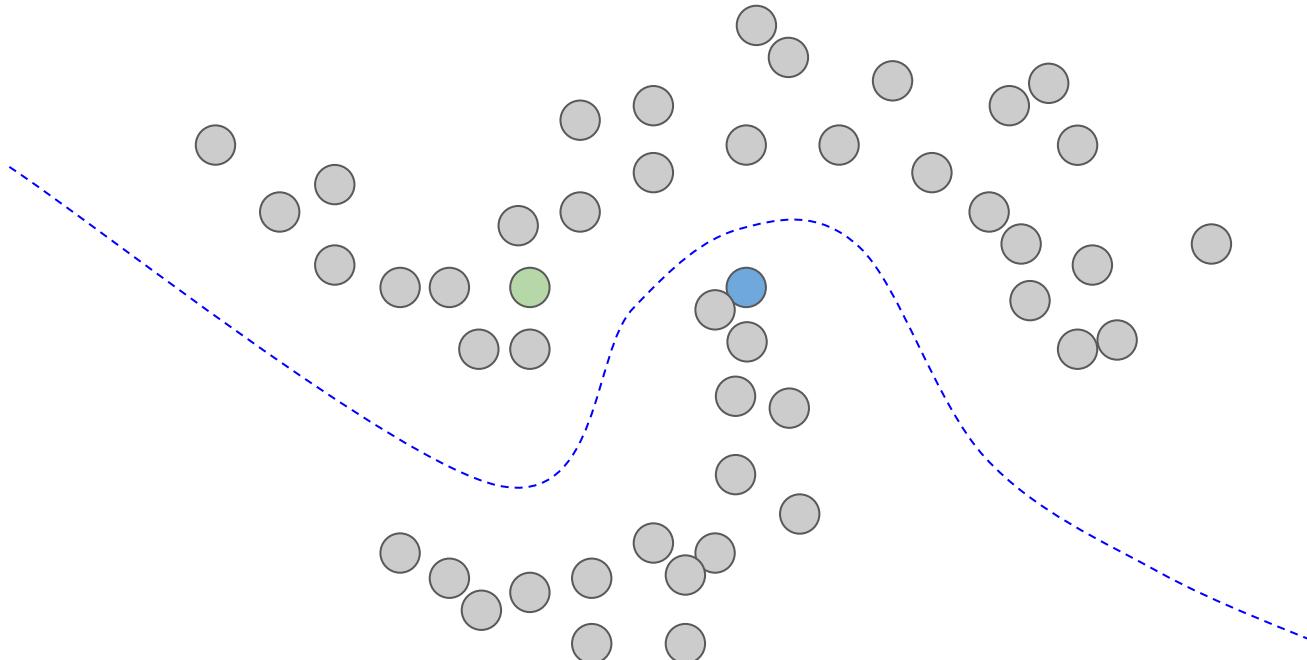
How data distribution $P(x)$ influences decisions (1D)



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



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



How $P(x)$ is valuable for naive Bayesian classifier

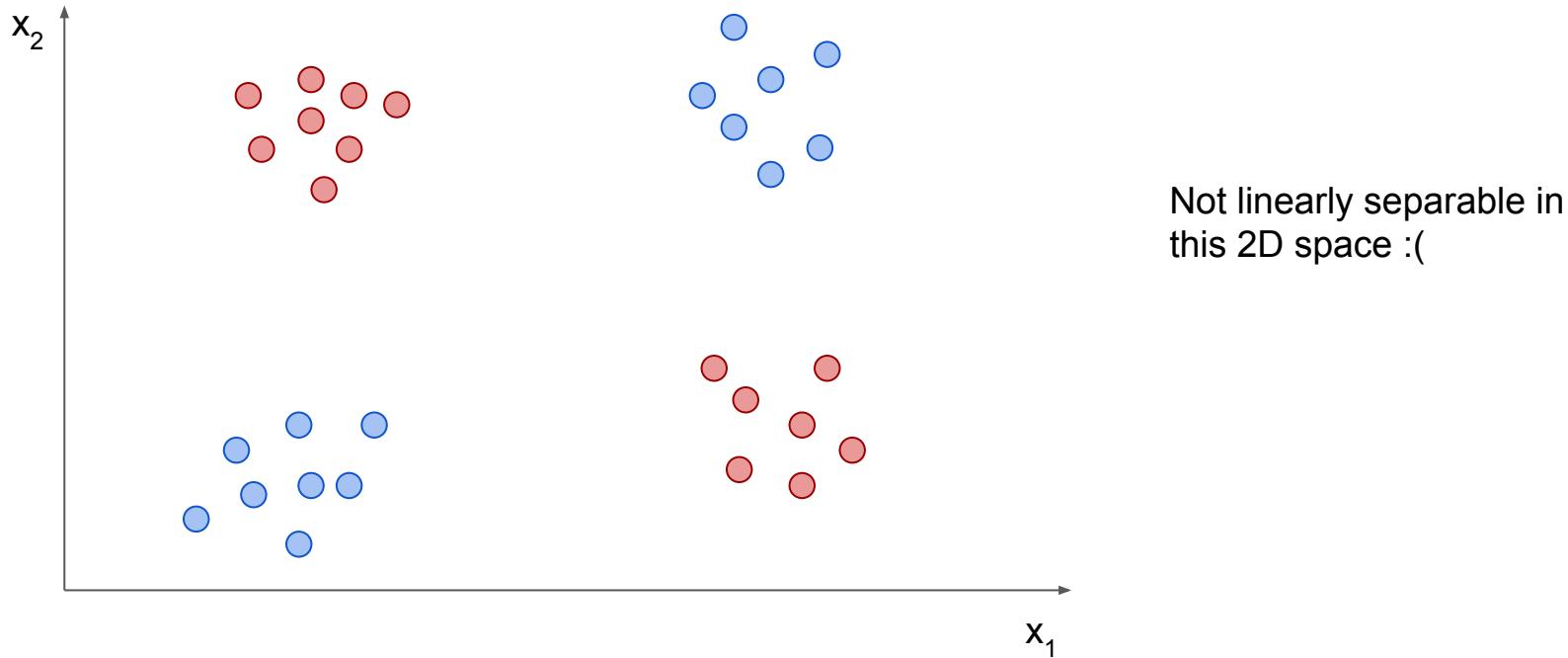
Bayes rule

$$P(Y = y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

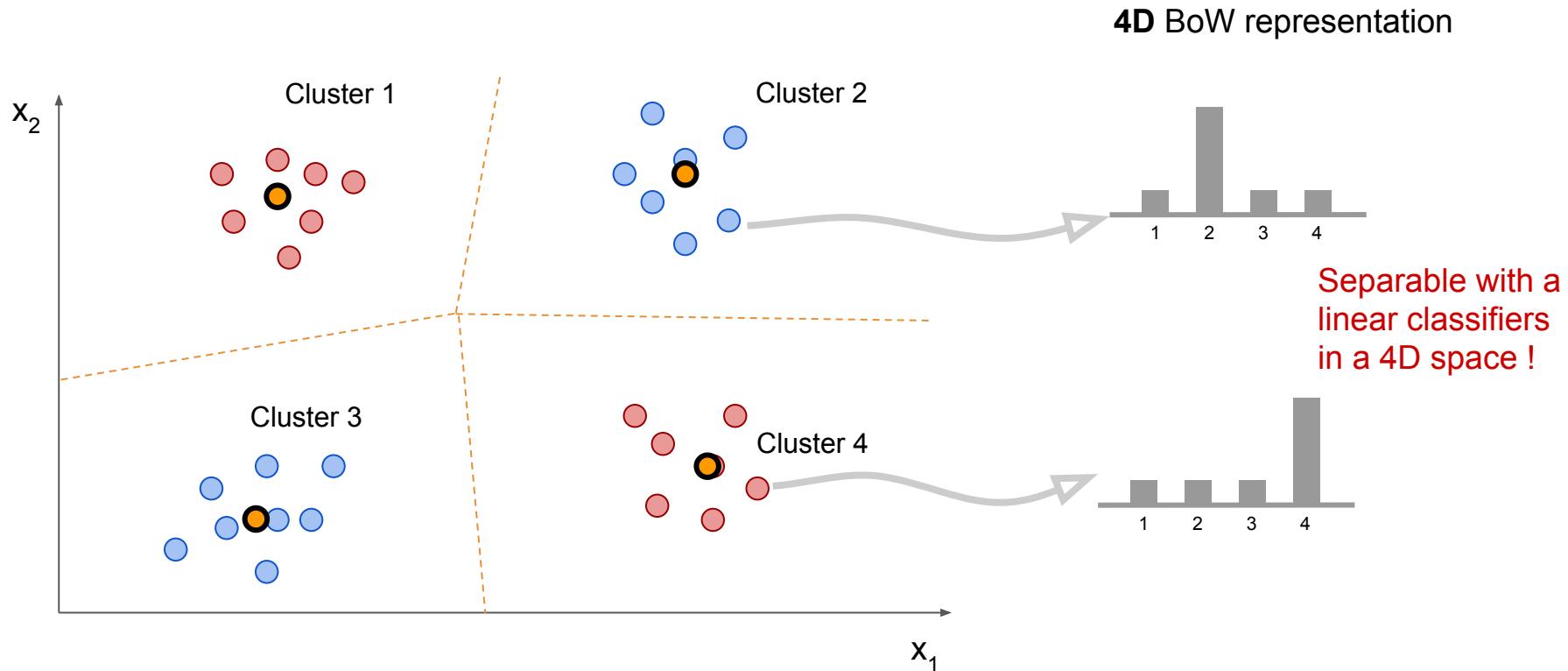
X: Data
Y: Labels

- $P(Y|X)$ depends on $P(X|Y)$ and $P(X)$
- **Knowledge of distribution $P(X)$ can help to predict $P(Y|X)$**
- Good model of $P(X)$ must have Y as an implicit latent variable

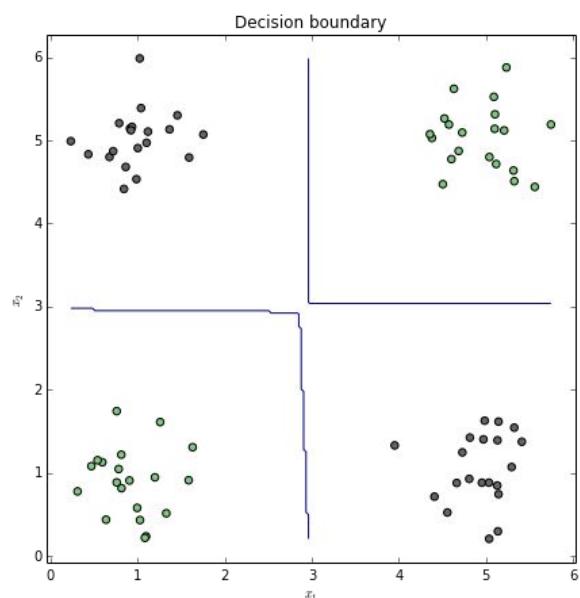
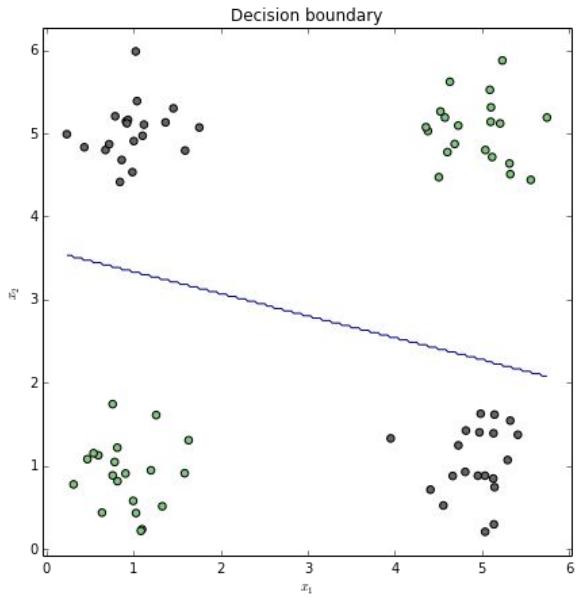
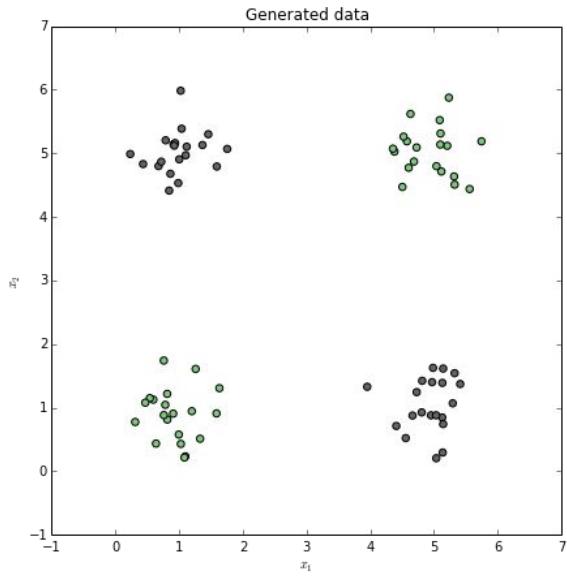
How clustering is valuable for linear classifiers



How clustering is valuable for linear classifiers



How clustering is valuable for linear classifiers



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

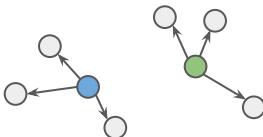
Typical assumptions:

- 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.

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!

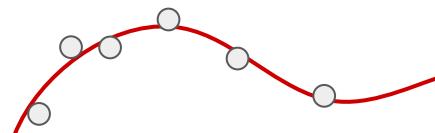


Cluster assumption

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

Manifold assumption

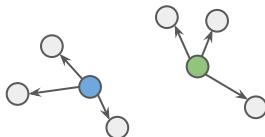
- Linear manifolds
 - PCA
 - Linear autoencoders
 - Random projections
 - ICA
- Non-linear manifolds:
 - Non-linear autoencoders
 - Deep autoencoders
 - Restricted Boltzmann machines
 - Deep belief nets



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!

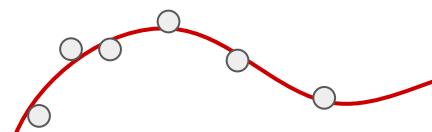


Cluster assumption

- 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



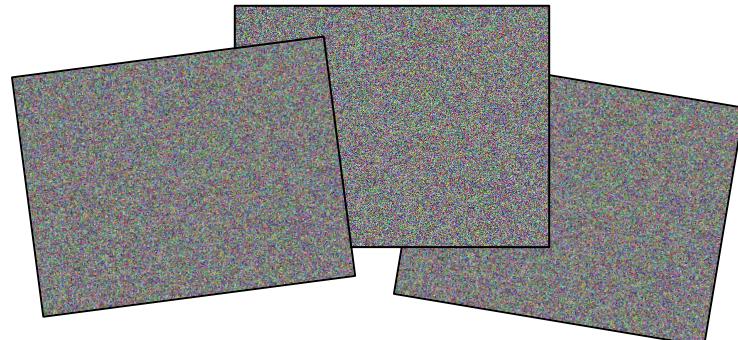
The manifold hypothesis

The data distribution lie close to a low-dimensional manifold

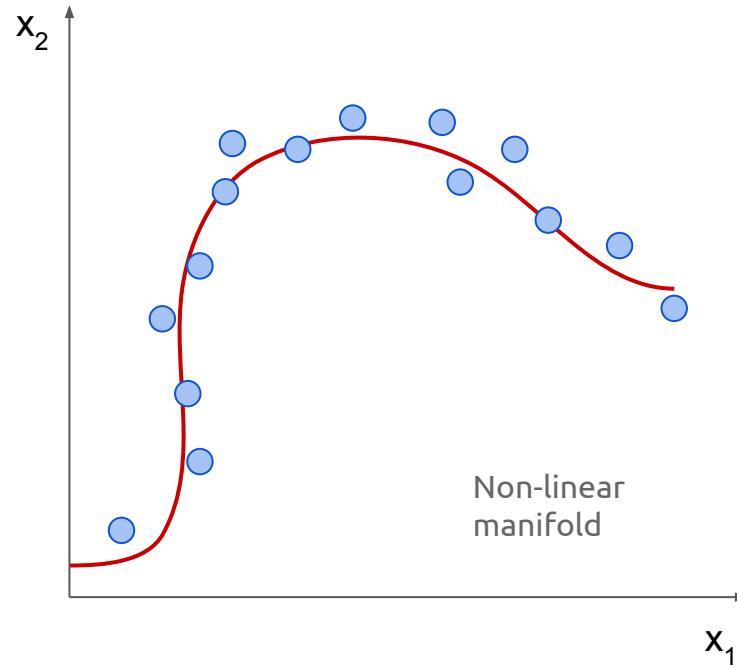
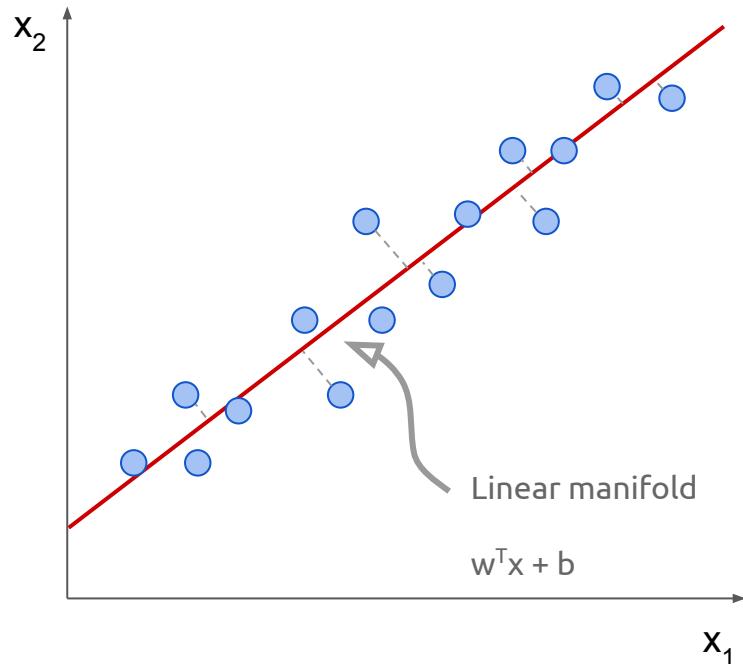
Example: **consider image data**

- 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



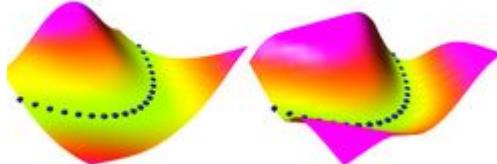
The manifold hypothesis



The manifold hypothesis: Energy-based models

Often intractable to explicitly model probability density

Energy-based model: high energy for data far from manifold, low energy for data near manifold of observed data



Fitting energy-based models

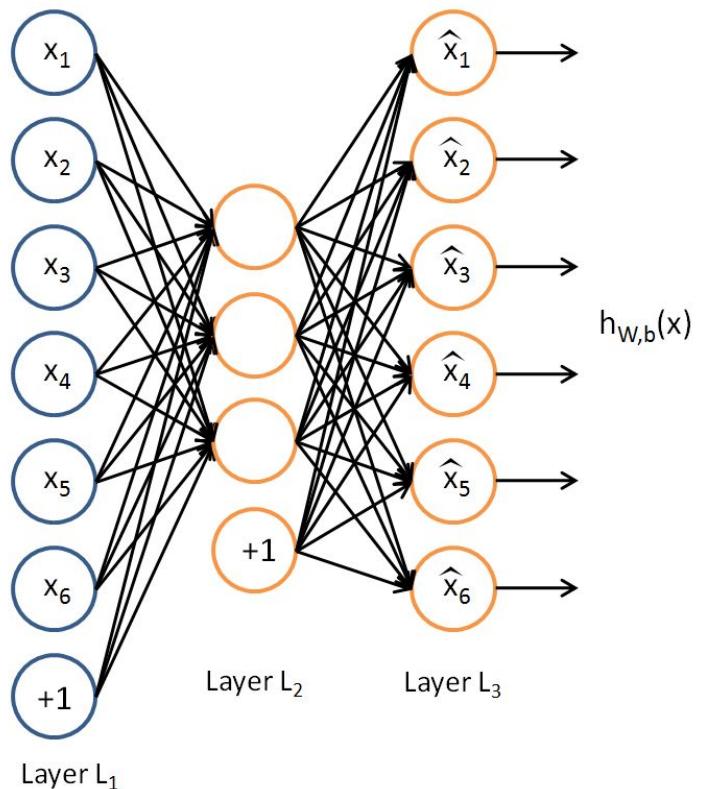
- Push down on area near observations.
- Push up everywhere else.

Examples

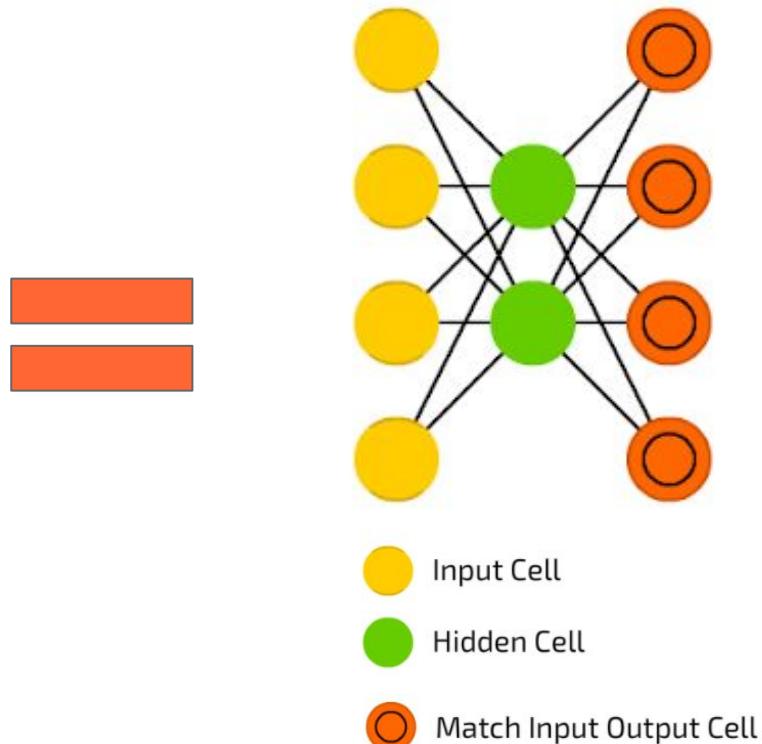
Encoder-decoder models: measure energy with reconstruction error

- **K-Means:** push down near prototypes. Push up based on distance from prototypes.
- **PCA:** push down near line of maximum variation. Push up based on distance to line.
- **Autoencoders:** non-linear manifolds...

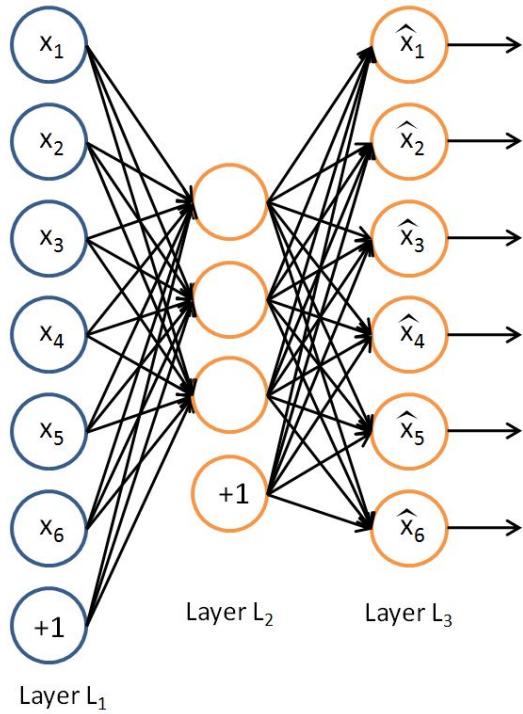
Autoencoder (AE)



Auto Encoder (AE)



Autoencoder (AE)

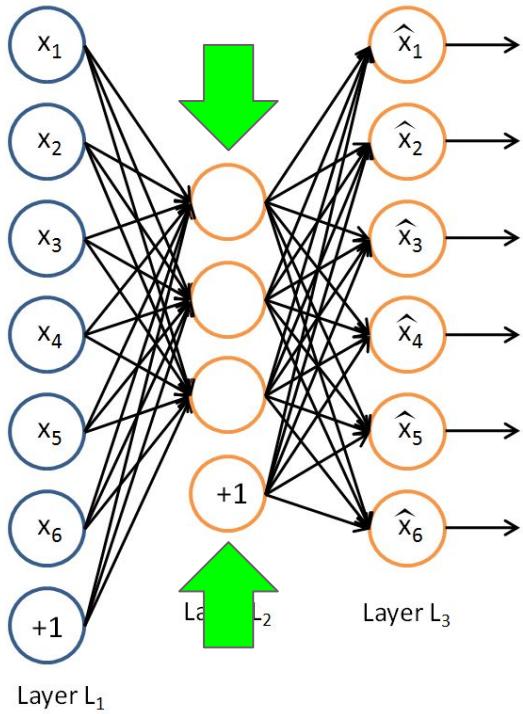


Autoencoders:

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

Autoencoder (AE)

WHY?



Dimensionality reduction:

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

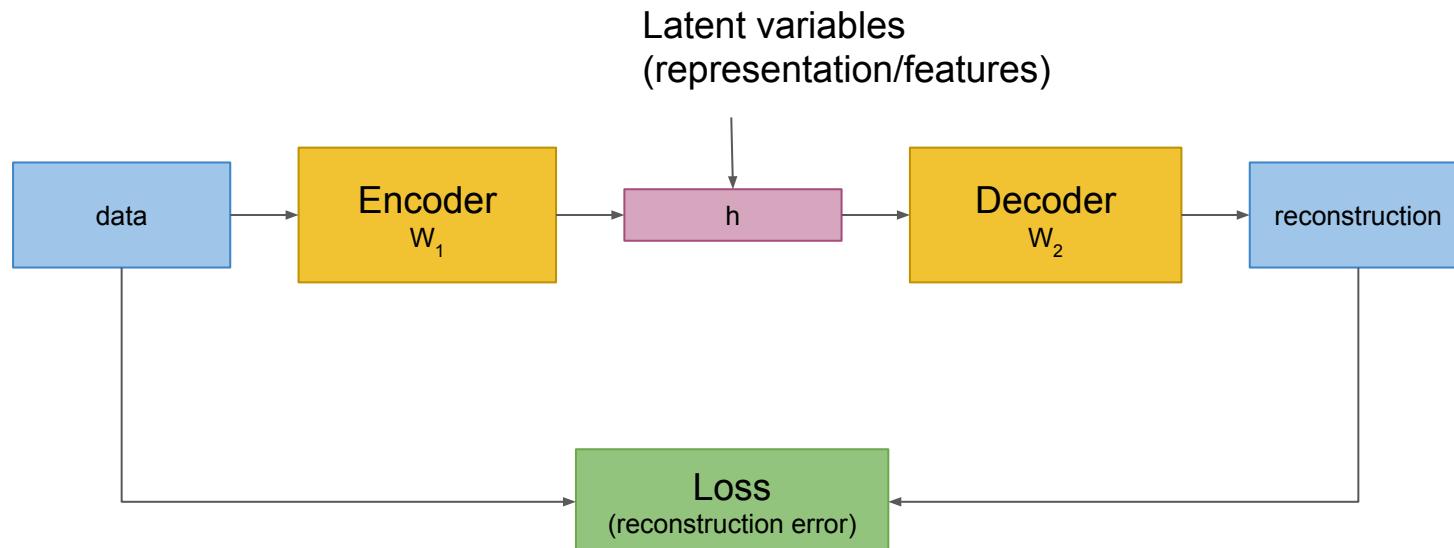
Autoencoder (AE)

WHY?



Pretraining:

1. Initialize a NN solving an autoencoding problem.



Autoencoder (AE)

WHY?



Pretraining:

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

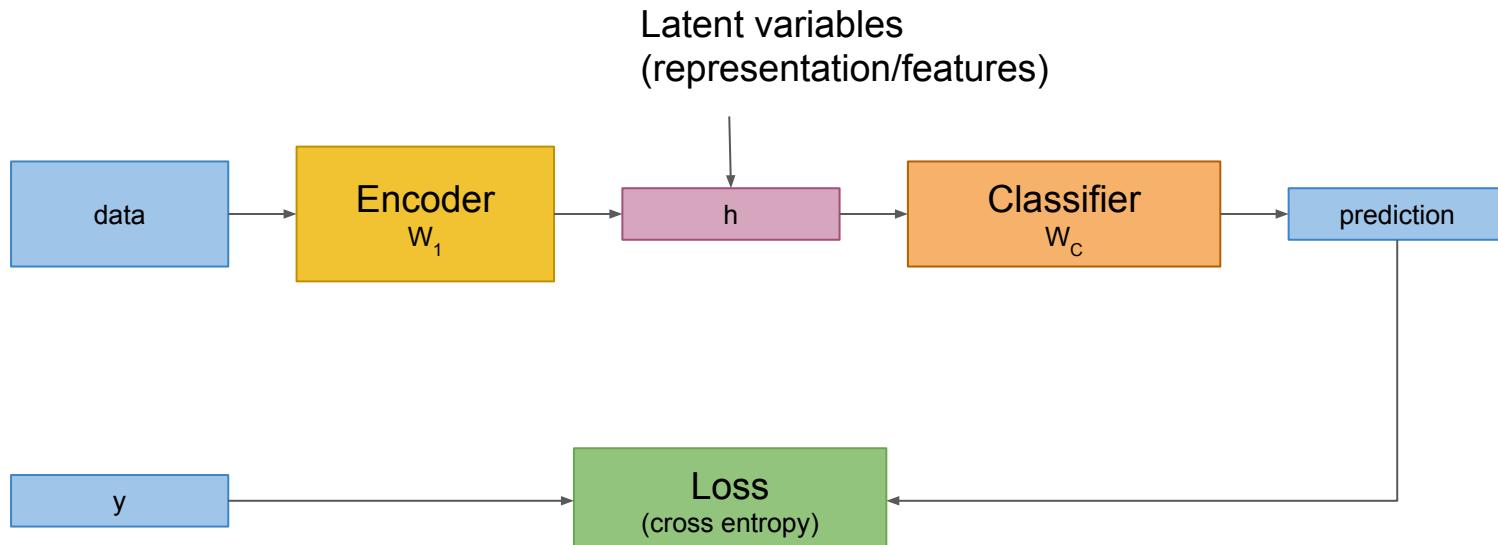
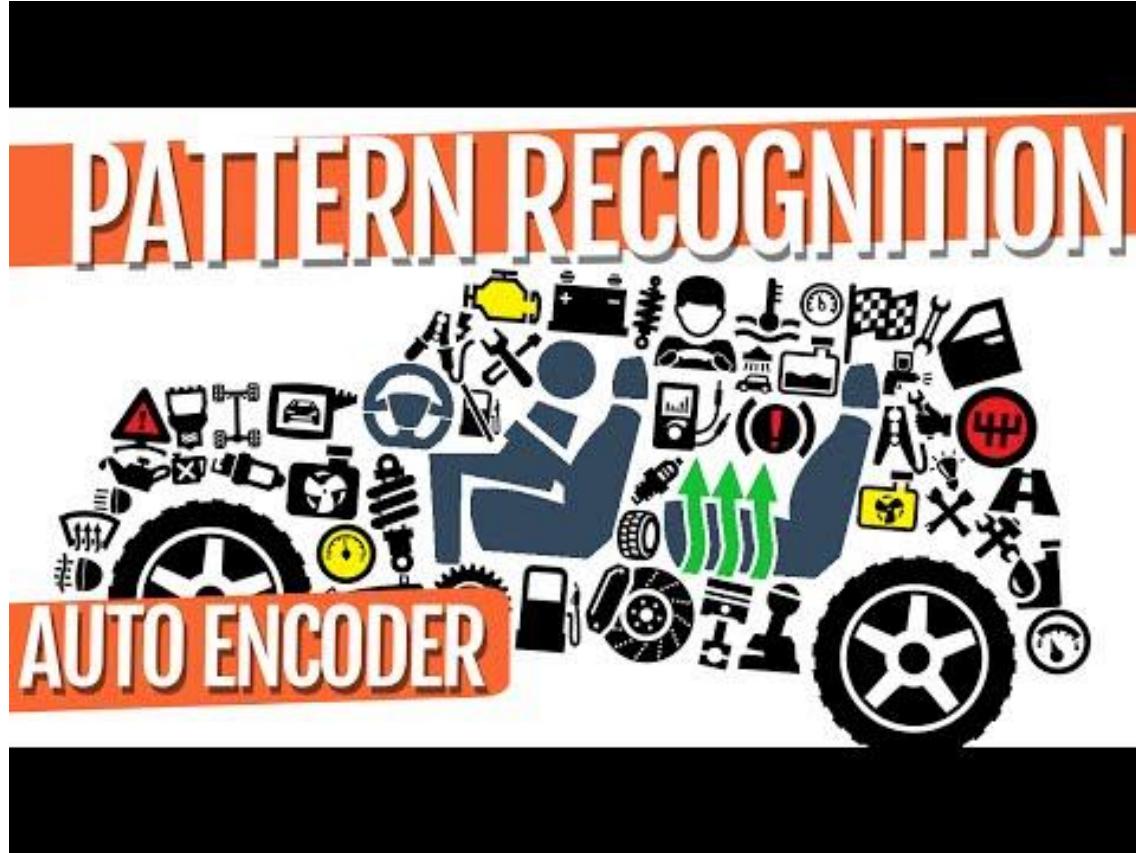


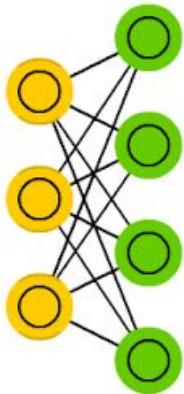
Figure: [Kevin McGuinness \(DLCV UPC 2017\)](#)

Autoencoder (AE)



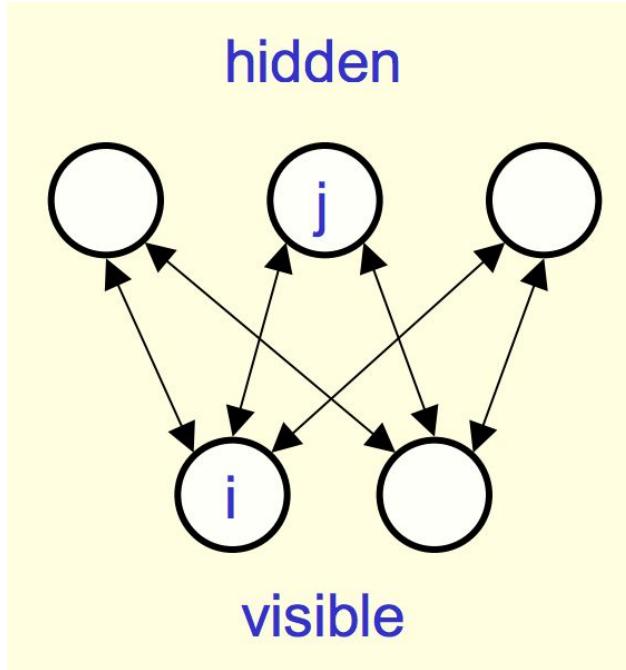
Restricted Boltzmann Machine (RBM)

Restricted BM (RBM)



- Backfed Input Cell
- Probabilistic Hidden Cell
- Hidden Cell
- Match Input Output Cell

Restricted Boltzmann Machine (RBM)

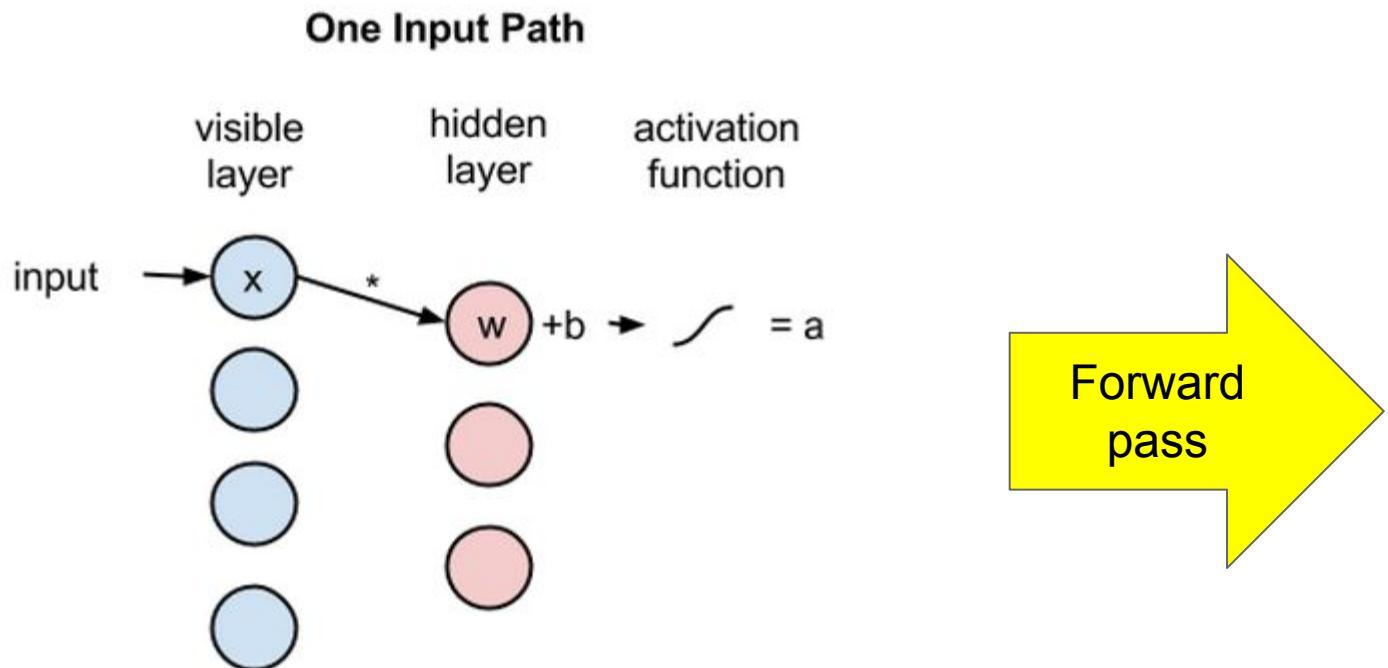


- Shallow two-layer net.
- Restricted = No two nodes in a layer share a connection
- Bipartite graph.
- Bidirectional graph
 - Shared weights.
 - Different biases.

Figure: Geoffrey Hinton (2013)

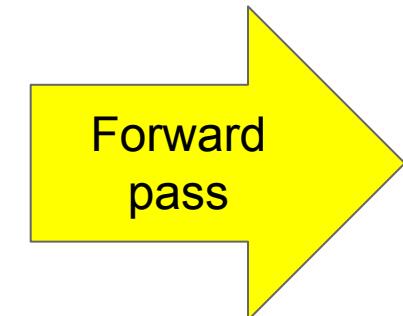
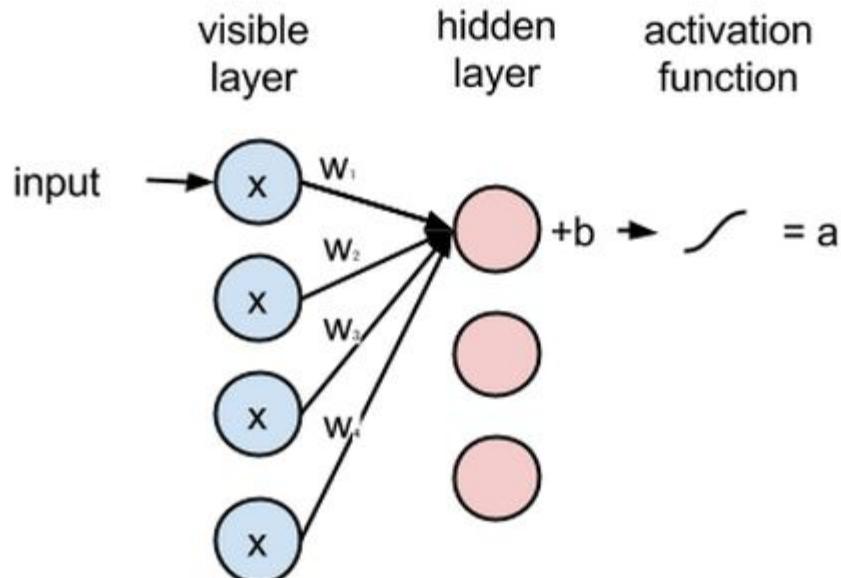
Salakhutdinov, Ruslan, Andriy Mnih, and Geoffrey Hinton. "[Restricted Boltzmann machines for collaborative filtering.](#)" Proceedings of the 24th international conference on Machine learning. ACM, 2007.

Restricted Boltzmann Machine (RBM)

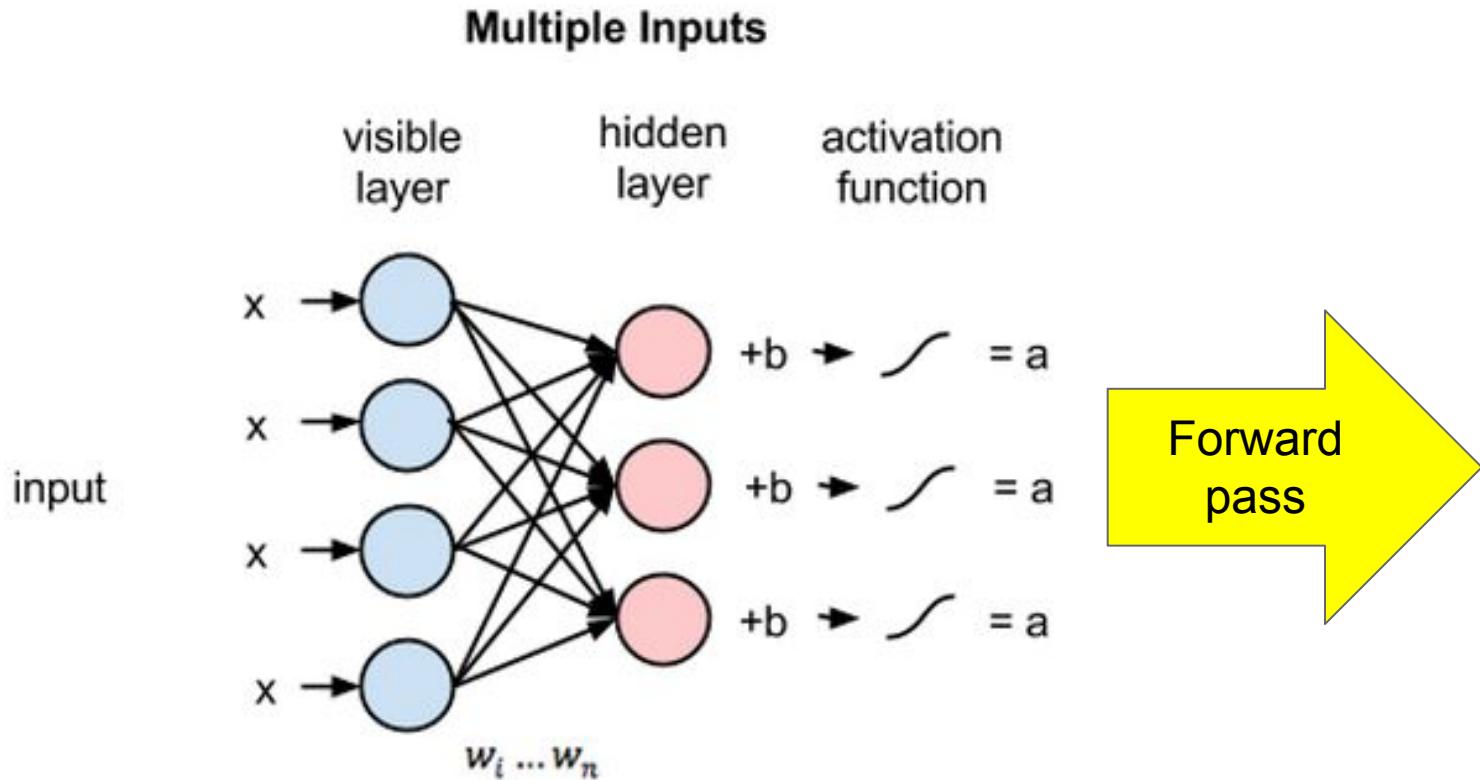


Restricted Boltzmann Machine (RBM)

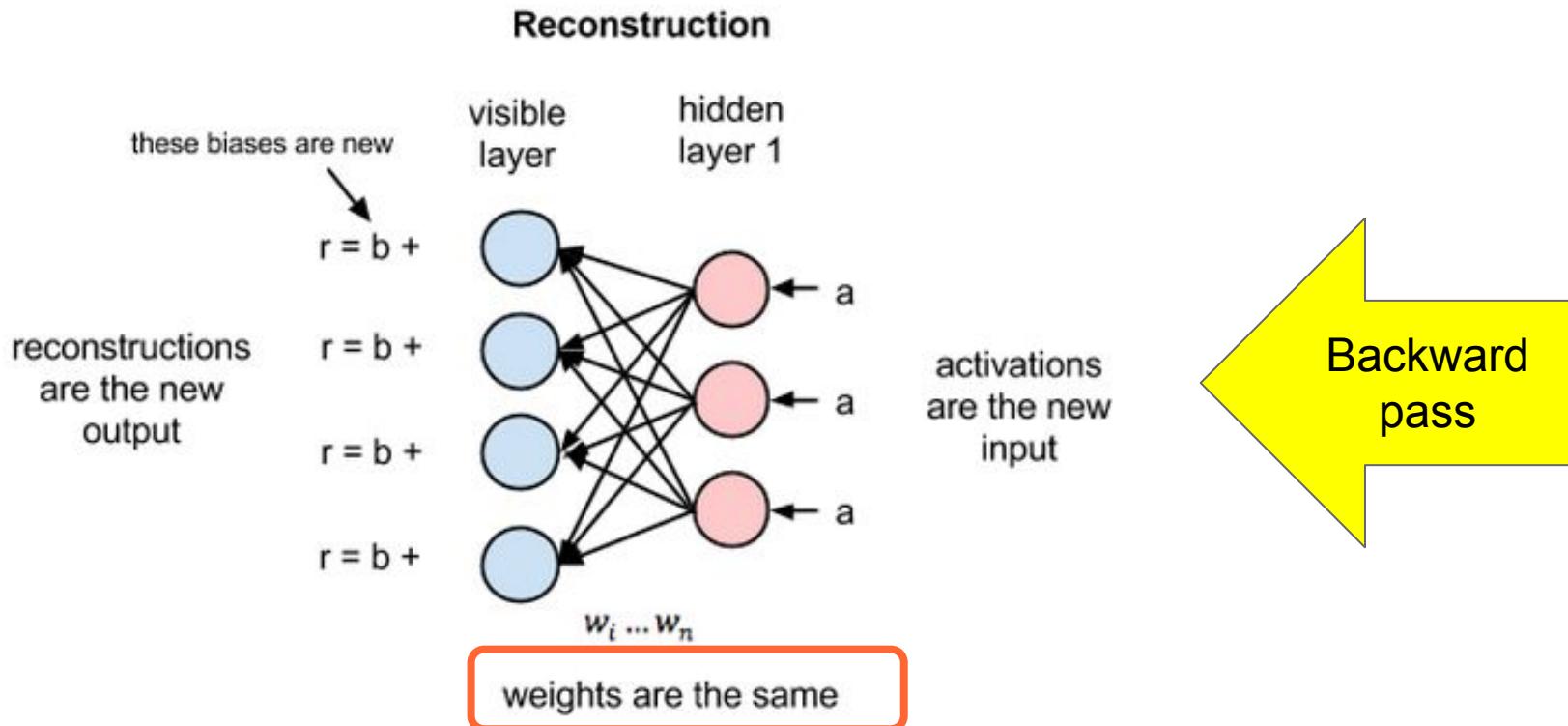
Weighted Inputs Combine @Hidden Node



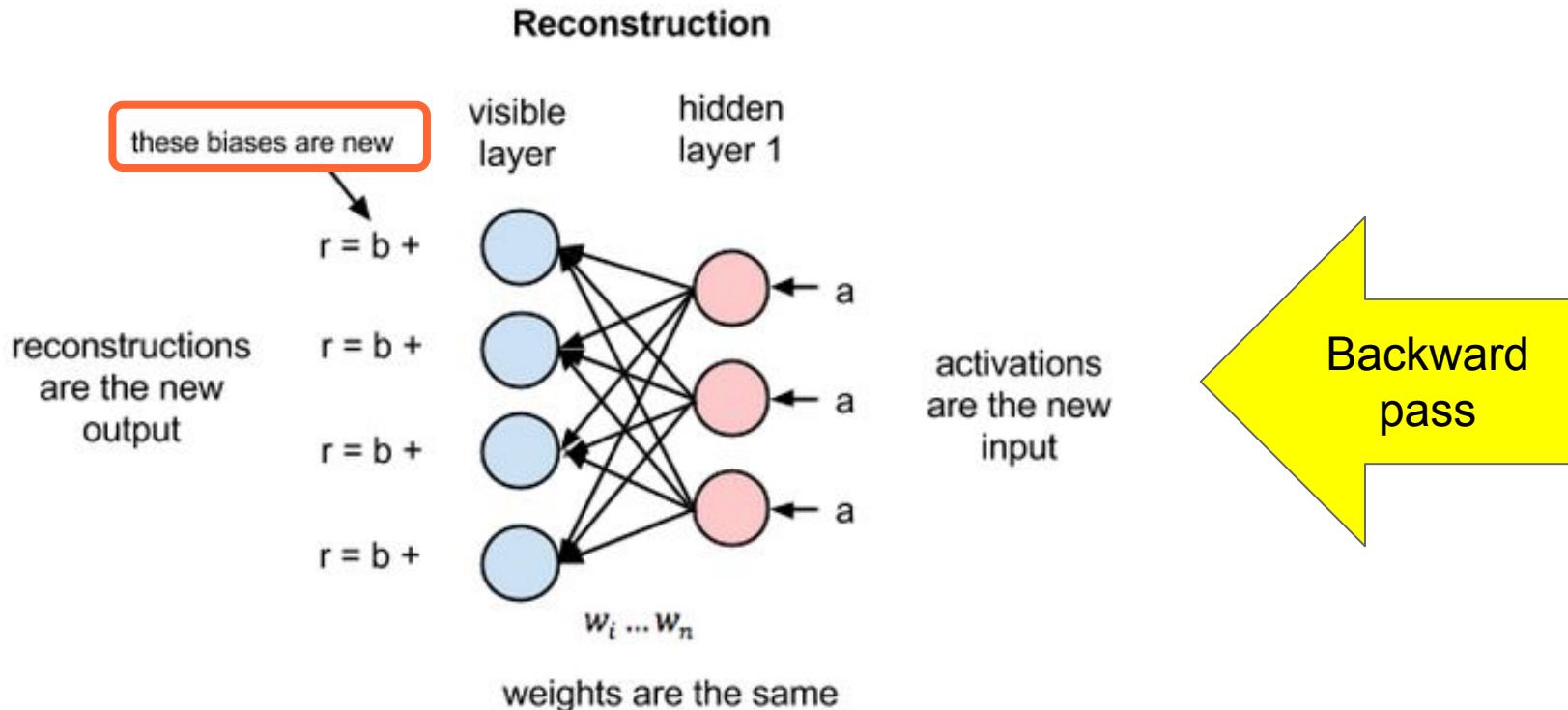
Restricted Boltzmann Machine (RBM)



Restricted Boltzmann Machine (RBM)



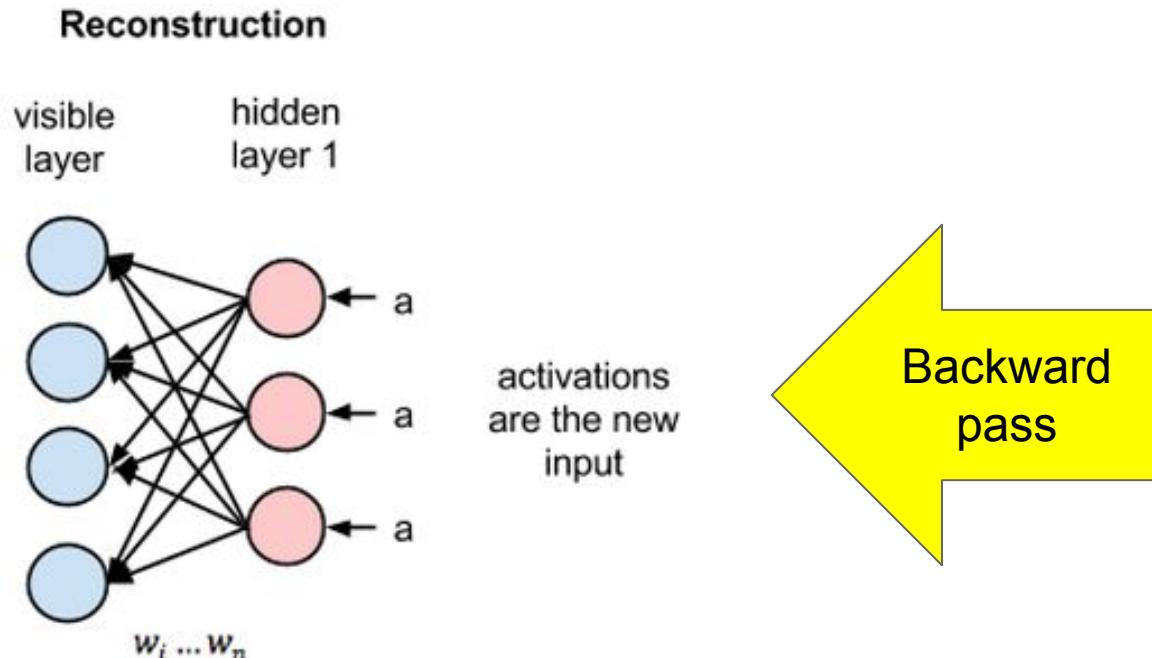
Restricted Boltzmann Machine (RBM)



Restricted Boltzmann Machine (RBM)

The reconstructed values at the visible layer are compared with the actual ones with the [KL Divergence](#).

$$D_{\text{KL}}(P\|Q) = - \sum_i P(i) \log \frac{Q(i)}{P(i)},$$



Restricted Boltzmann Machine (RBM)

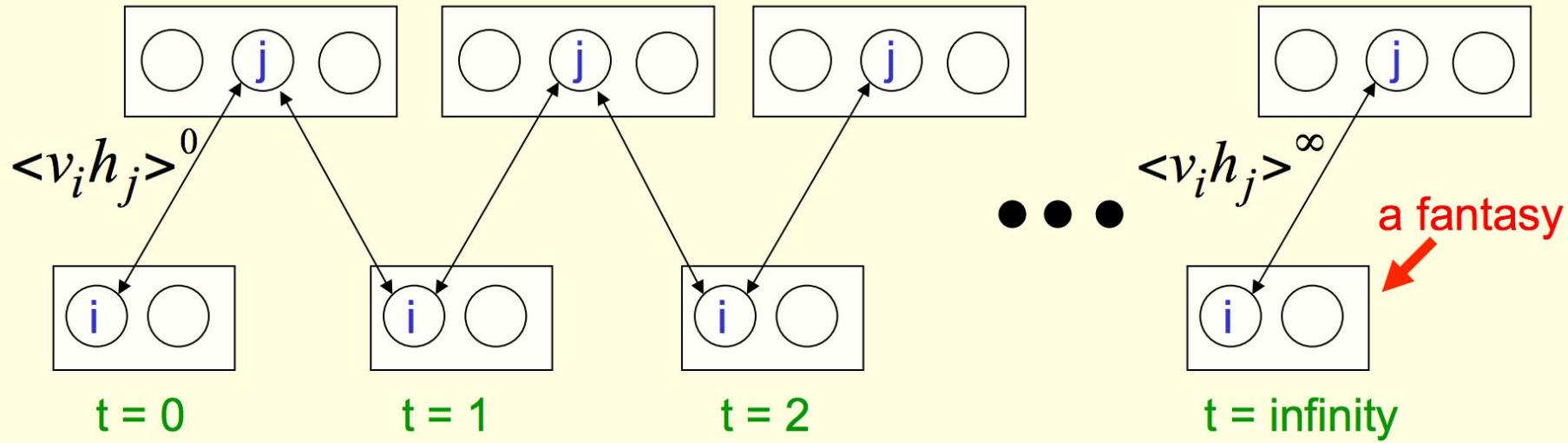
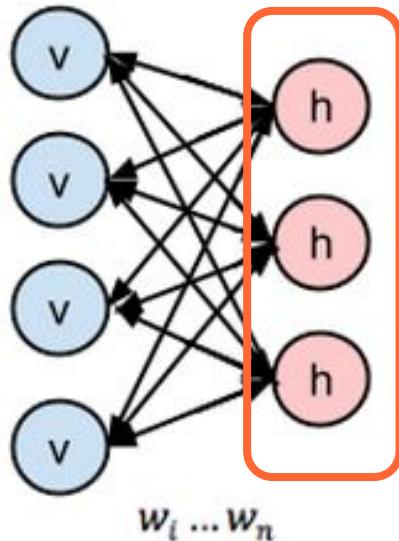


Figure: Geoffrey Hinton (2013)

Salakhutdinov, Ruslan, Andriy Mnih, and Geoffrey Hinton. "[Restricted Boltzmann machines for collaborative filtering.](#)" Proceedings of the 24th international conference on Machine learning. ACM, 2007.

Restricted Boltzmann Machine (RBM)

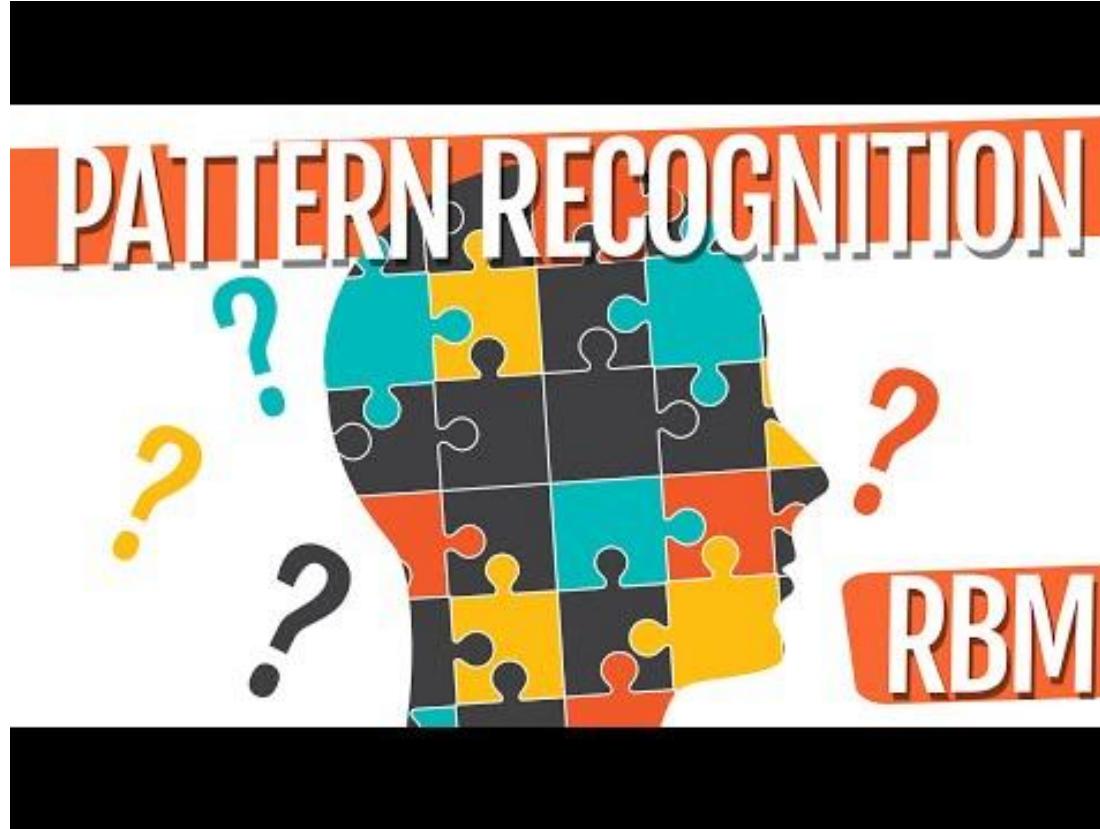
WHY?



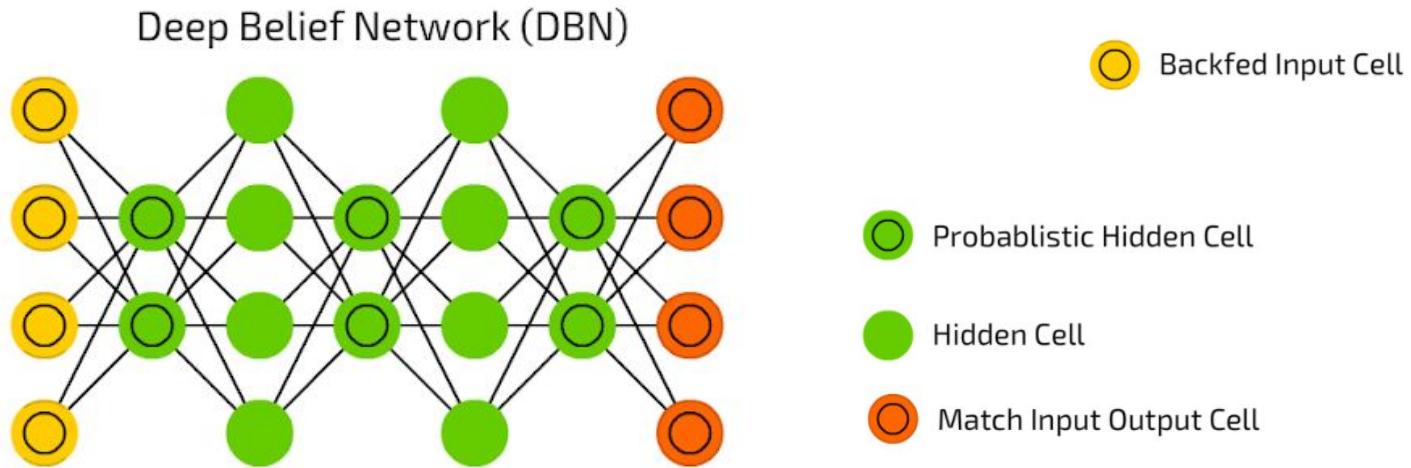
RBMs are a specific type of autoencoder.



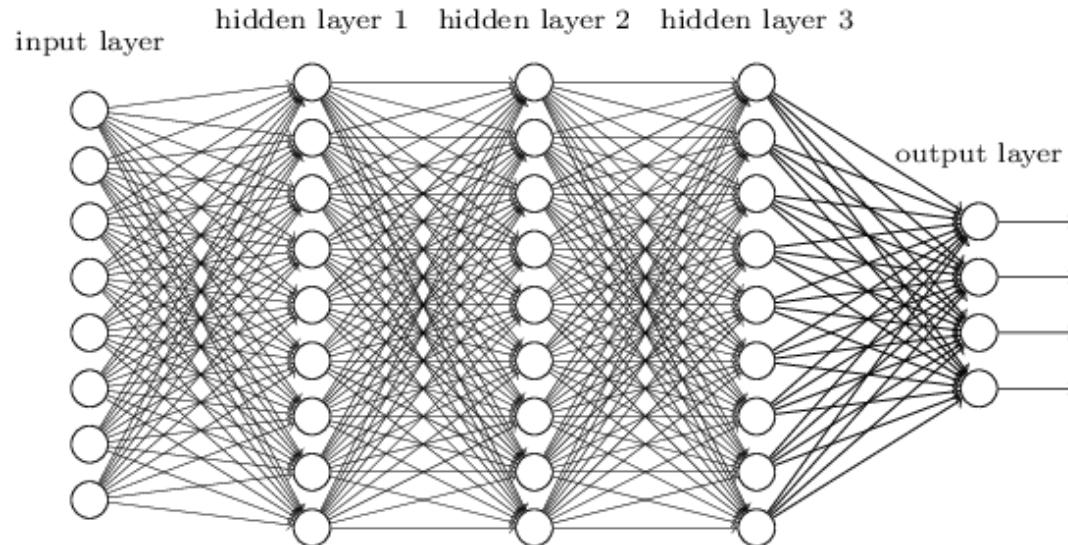
Restricted Boltzmann Machine (RBM)



Deep Belief Networks (DBN)



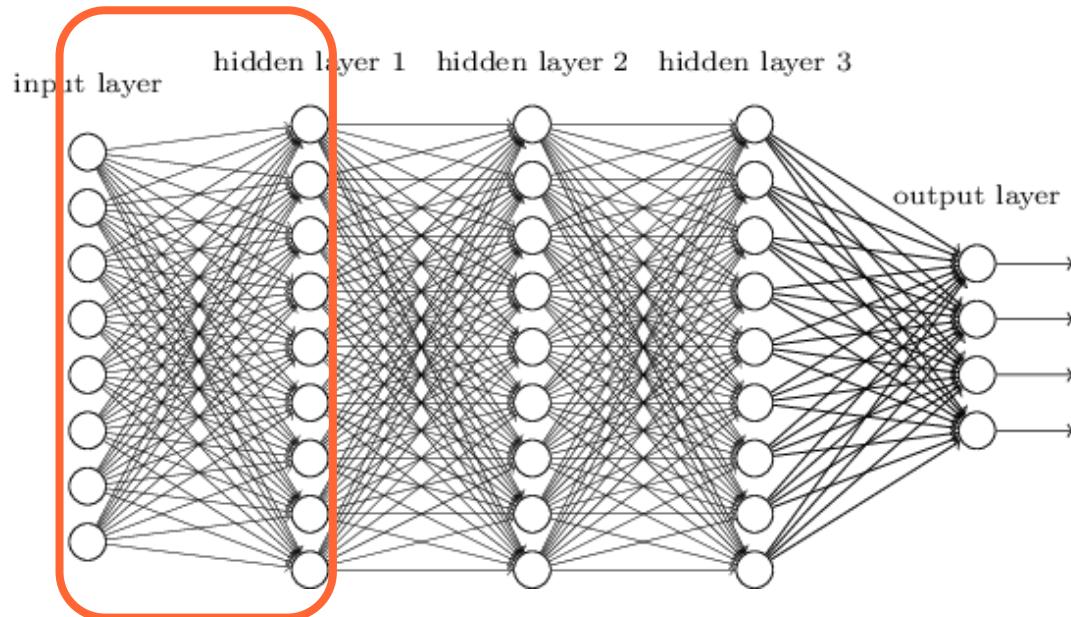
Deep Belief Networks (DBN)



- Architecture like an MLP.
- Training as a stack of RBMs.

Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "[A fast learning algorithm for deep belief nets.](#)" Neural computation 18, no. 7 (2006): 1527-1554.

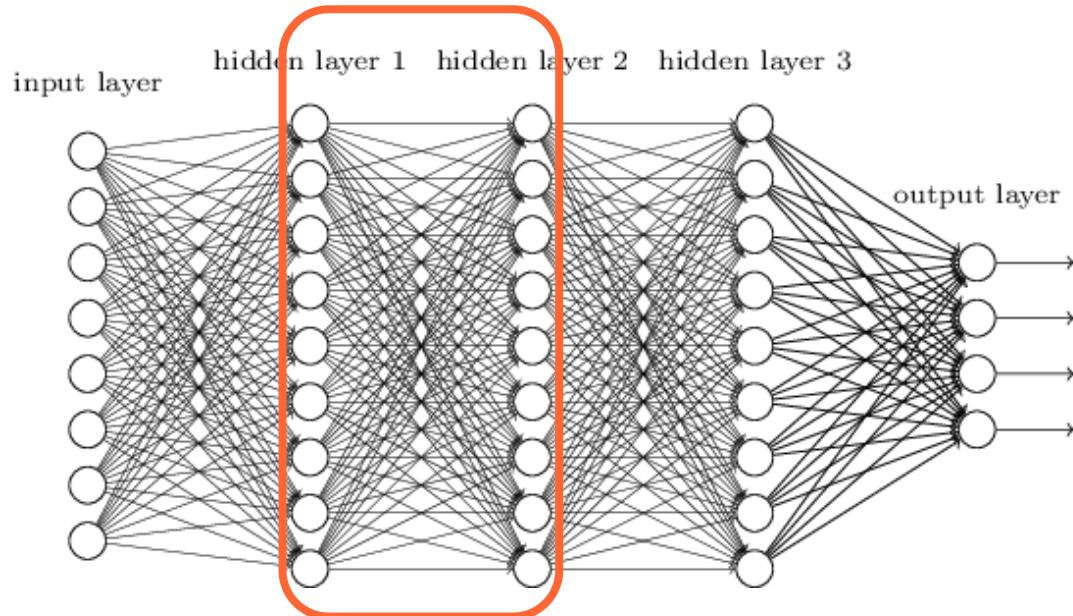
Deep Belief Networks (DBN)



- Architecture like an MLP.
- Training as a stack of RBMs.

Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "[A fast learning algorithm for deep belief nets.](#)" Neural computation 18, no. 7 (2006): 1527-1554.

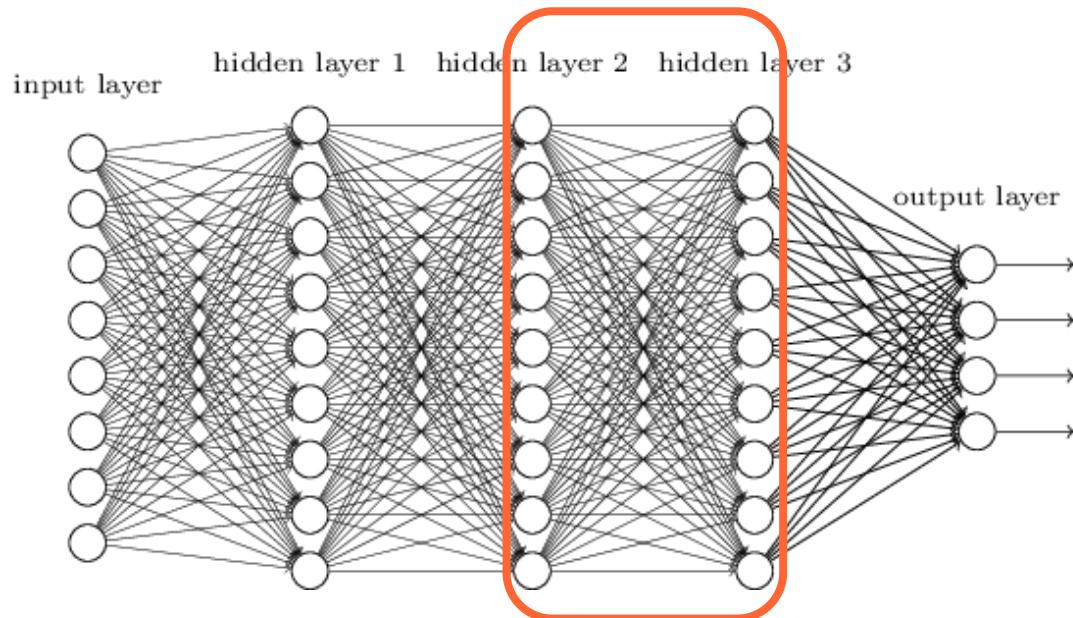
Deep Belief Networks (DBN)



- Architecture like an MLP.
- Training as a stack of RBMs.

Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "[A fast learning algorithm for deep belief nets.](#)" Neural computation 18, no. 7 (2006): 1527-1554.

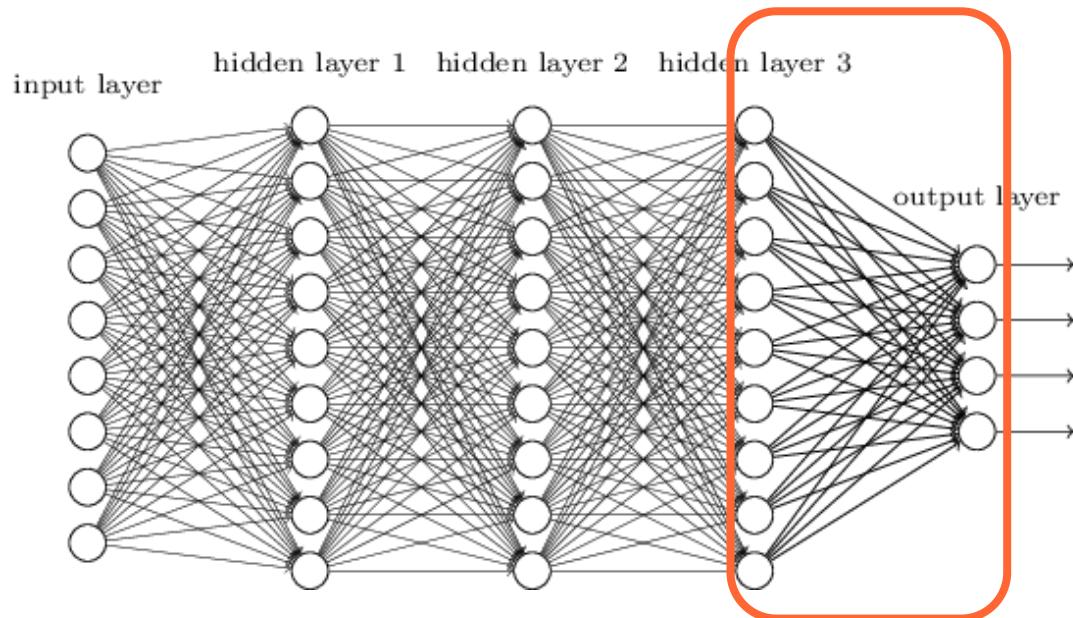
Deep Belief Networks (DBN)



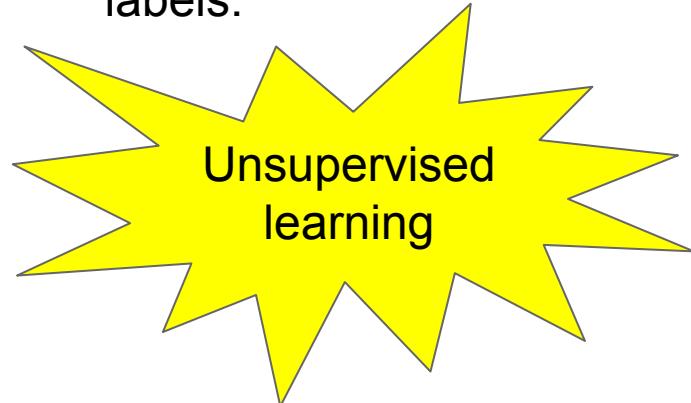
- Architecture like an MLP.
- Training as a stack of RBMs.

Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "[A fast learning algorithm for deep belief nets.](#)" Neural computation 18, no. 7 (2006): 1527-1554.

Deep Belief Networks (DBN)

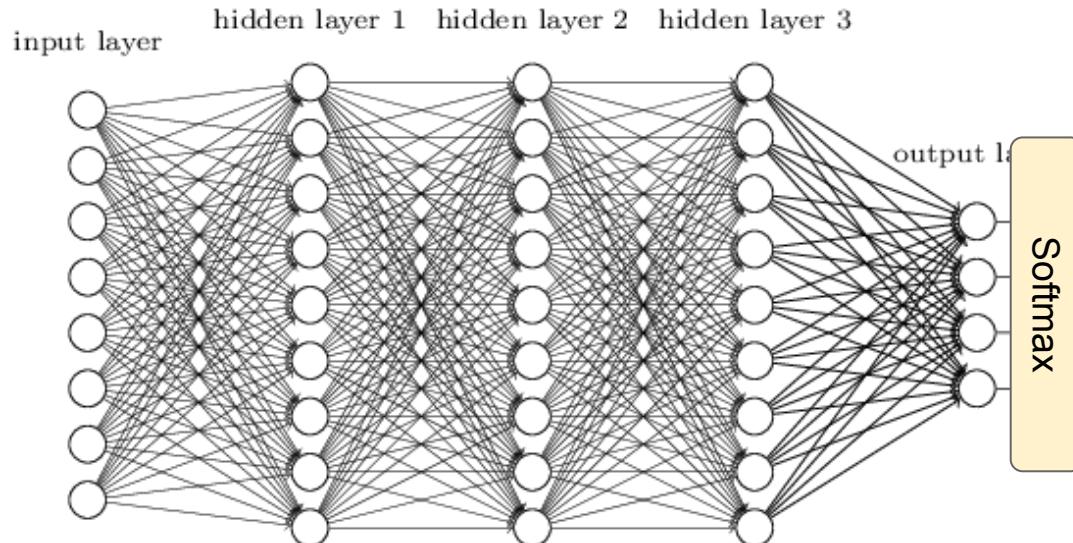


- Architecture like an MLP.
- Training as a stack of RBMs...
- ...so they do not need labels:

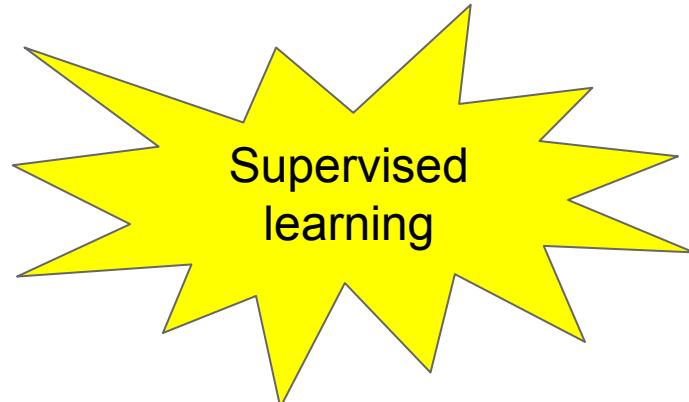


Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "[A fast learning algorithm for deep belief nets.](#)" Neural computation 18, no. 7 (2006): 1527-1554.

Deep Belief Networks (DBN)

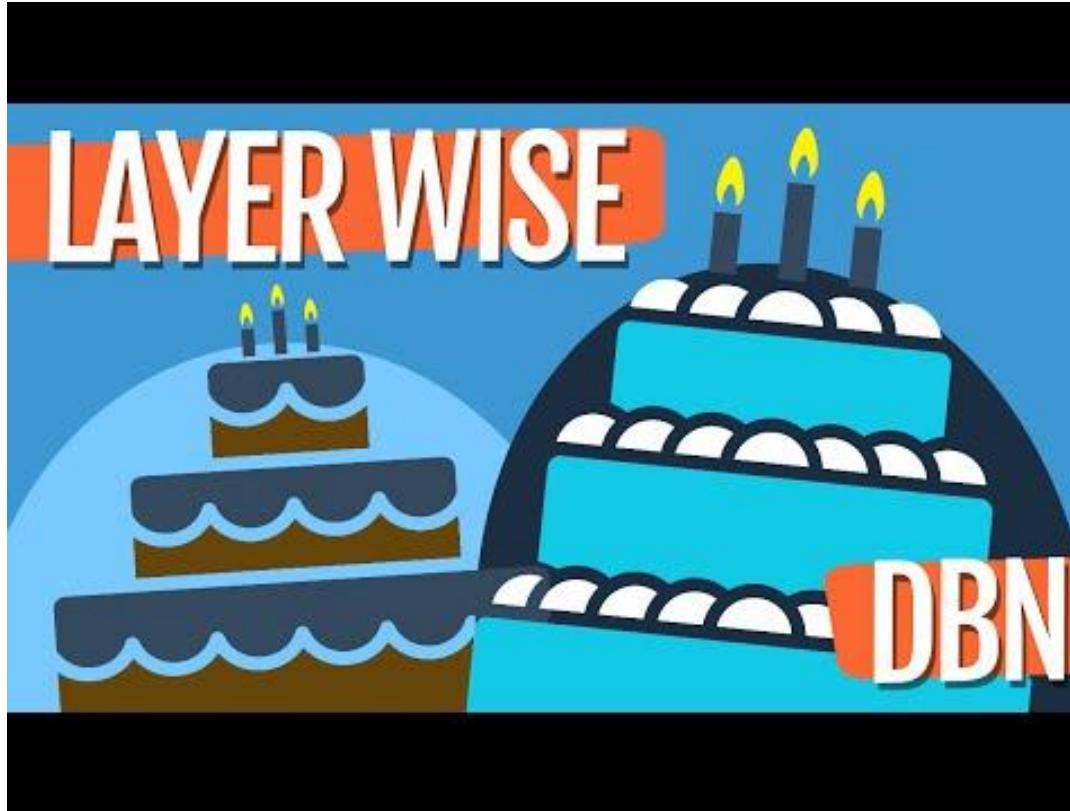


After the DBN is trained, it can be fine-tuned with a reduced amount of labels to solve a supervised task with superior performance.



Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "[A fast learning algorithm for deep belief nets.](#)" Neural computation 18, no. 7 (2006): 1527-1554.

Deep Belief Networks (DBN)



Deep Belief Networks (DBN)



Geoffrey Hinton, "[Introduction to Deep Learning & Deep Belief Nets](#)" (2012)
Geoffrey Hinton, "[Tutorial on Deep Belief Networks](#)". NIPS 2007.

Deep Belief Networks (DBN)

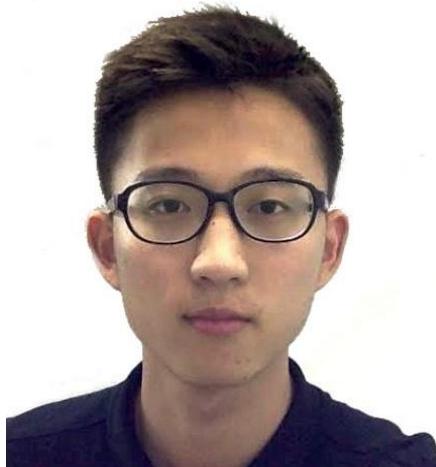
Geoffrey Hinton



Outline

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

Acknowledgments



Junting Pan



UNIVERSITAT POLITÈCNICA
DE CATALUNYA



Barcelona
Supercomputing
Center

Centro Nacional de Supercomputación



Xunyu Lin

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK



Unsupervised Feature Learning

■ "Pure" Reinforcement Learning (cherry)

- ▶ The machine predicts a scalar reward given once in a while.

▶ **A few bits for some samples**

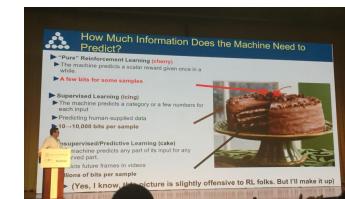
■ Supervised Learning (icing)

- ▶ 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

Unsupervised Feature Learning

■ "Pure" Reinforcement Learning (cherry)

- ▶ The machine predicts a scalar reward given once in a while.

- ▶ **A few bits for some samples**

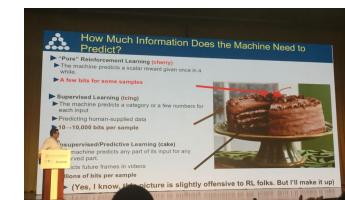
■ Supervised Learning (icing)

- ▶ 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)



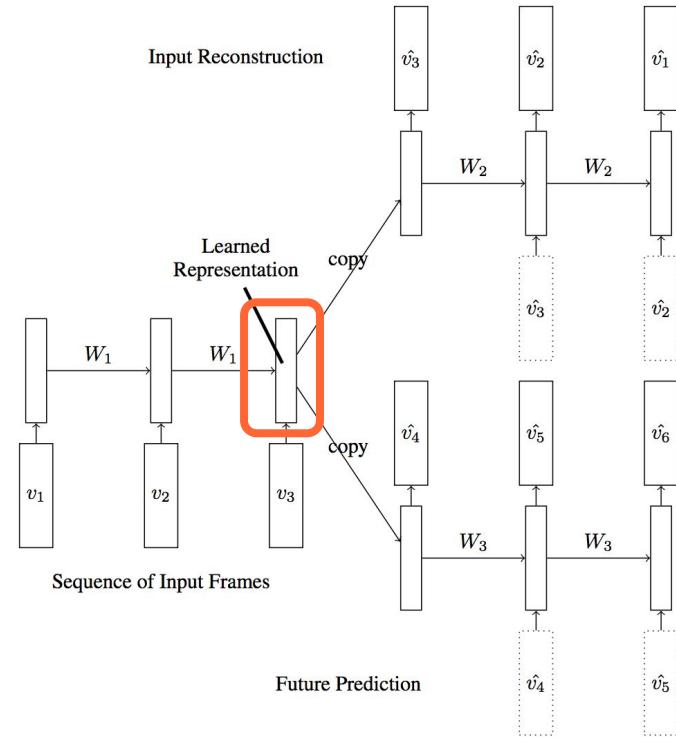
Slide credit:
Yann LeCun



Slide credit: Junting Pan

Frame Reconstruction & Prediction

Unsupervised feature learning (no labels) for...

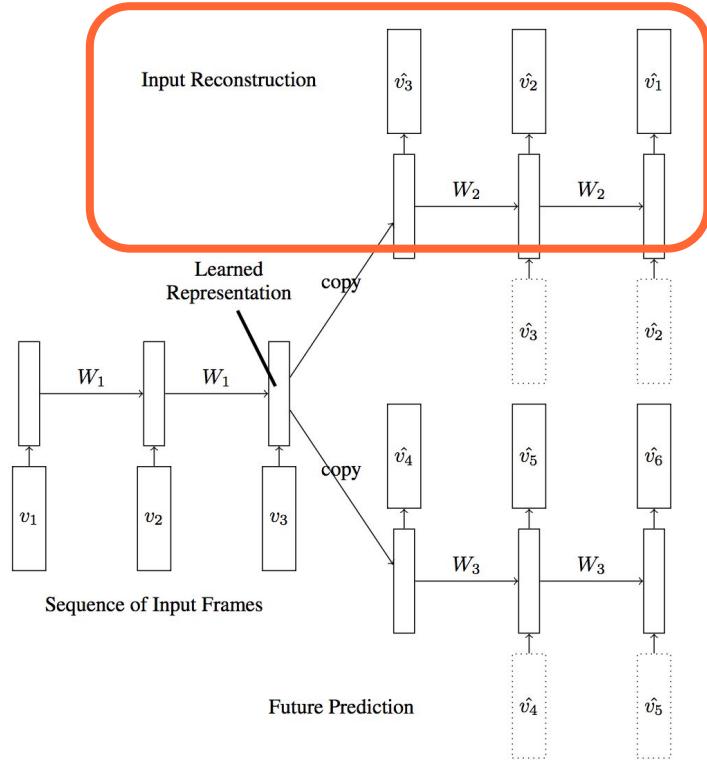
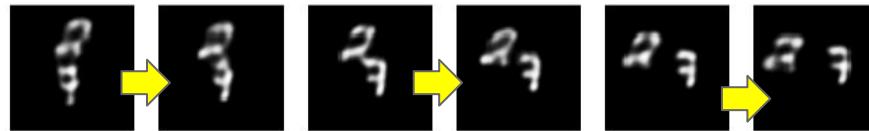


Srivastava, Nitish, Elman Mansimov, and Ruslan Salakhutdinov. "[Unsupervised Learning of Video Representations using LSTMs.](#)" In ICML 2015. [\[Github\]](#)

Frame Reconstruction & Prediction

Unsupervised feature learning (no labels) for...

...frame prediction.

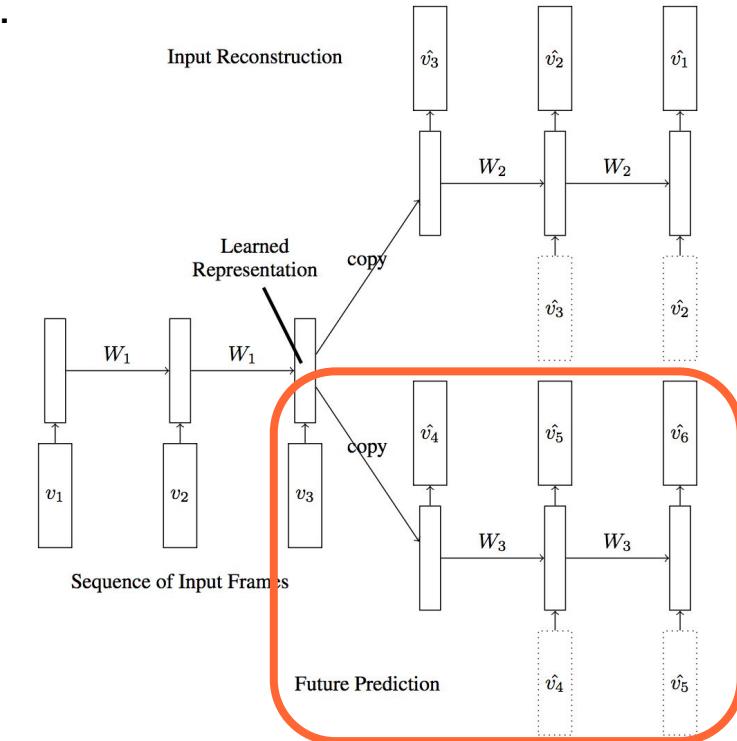
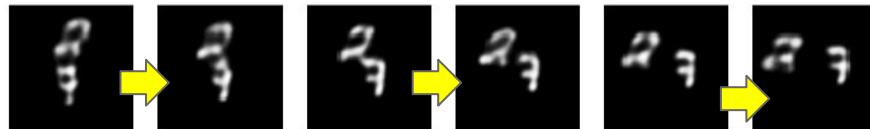


Srivastava, Nitish, Elman Mansimov, and Ruslan Salakhutdinov. "[Unsupervised Learning of Video Representations using LSTMs.](#)" In ICML 2015. [\[Github\]](#)

Frame Reconstruction & Prediction

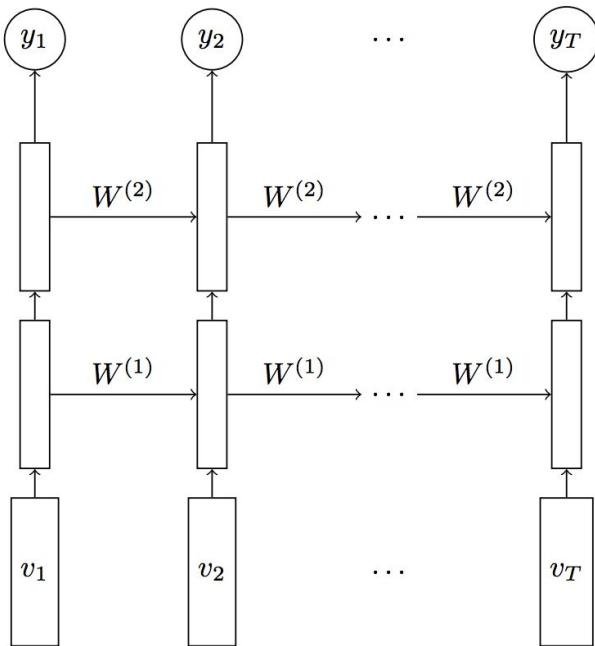
Unsupervised feature learning (no labels) for...

...frame prediction.



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

Frame Reconstruction & Prediction



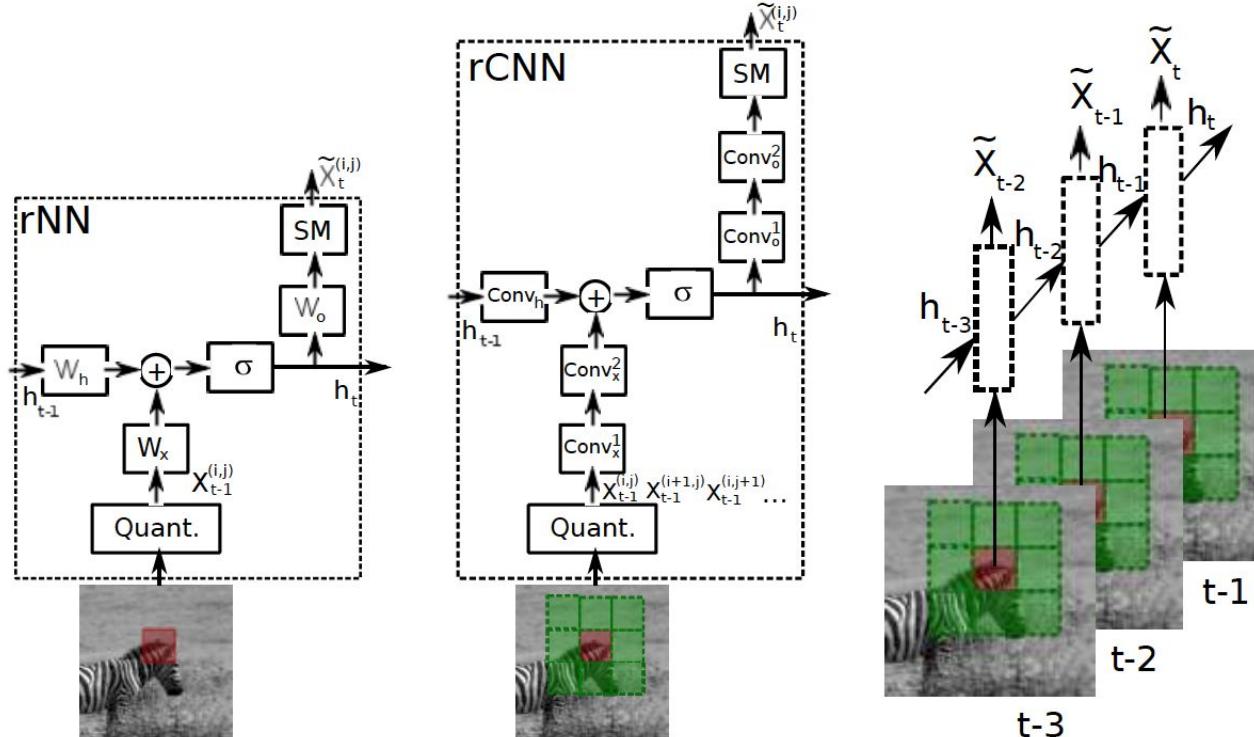
Unsupervised learned features (lots of data) are fine-tuned for activity recognition (little data).

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

Table 1. Summary of Results on Action Recognition.

Figure 6. LSTM Classifier.

Frame Prediction

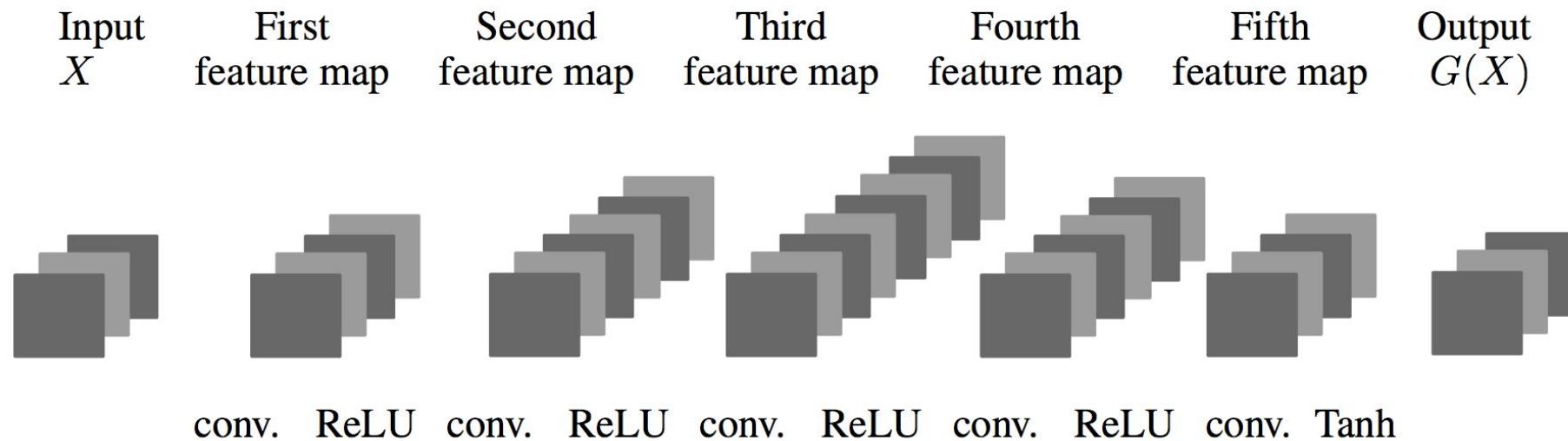


Ranzato, MarcAurelio, Arthur Szlam, Joan Bruna, Michael Mathieu, Ronan Collobert, and Sumit Chopra. ["Video \(language\) modeling: a baseline for generative models of natural videos."](#) arXiv preprint arXiv:1412.6604 (2014).

Frame Prediction

Video frame prediction with a ConvNet.

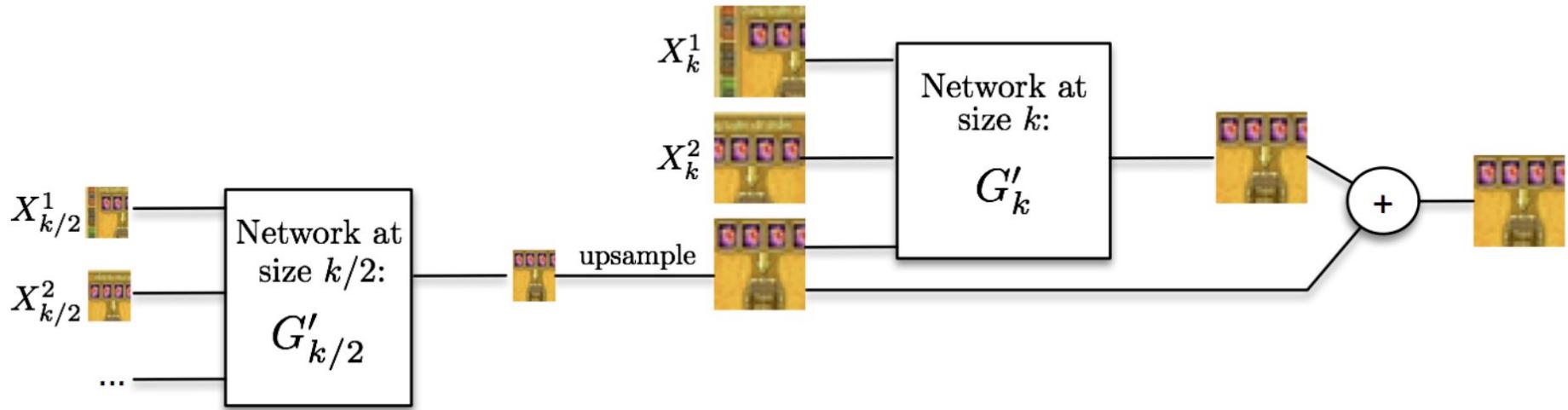
Figure 1: A basic next frame prediction convnet



Frame Prediction

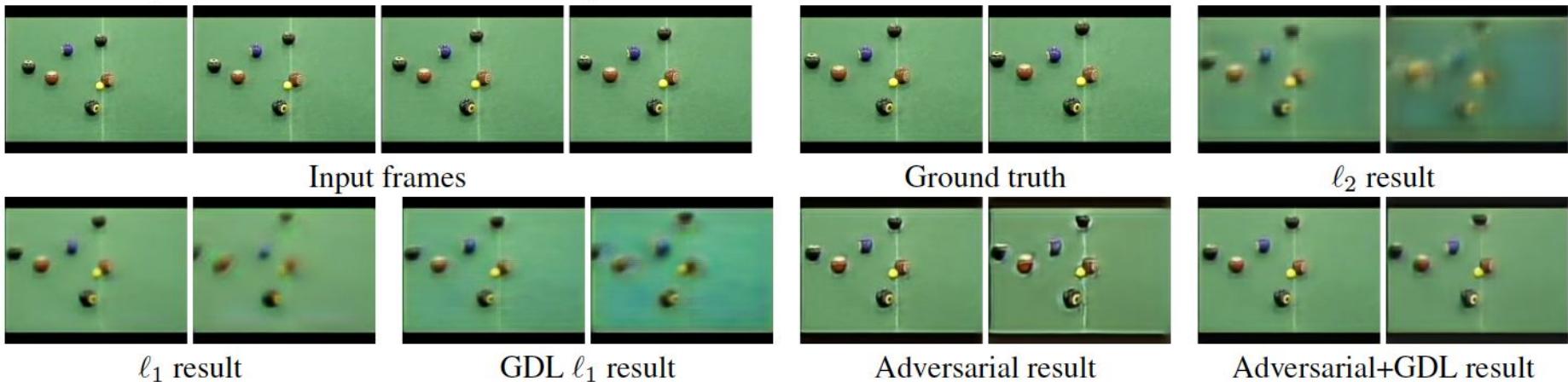
The blurry predictions from MSE are improved with multi-scale architecture, adversarial learning and an image gradient difference loss function.

Figure 2: Multi-scale architecture



Frame Prediction

The blurry predictions from MSE (ℓ_1) are improved with multi-scale architecture, adversarial training and an image gradient difference loss (GDL) function.

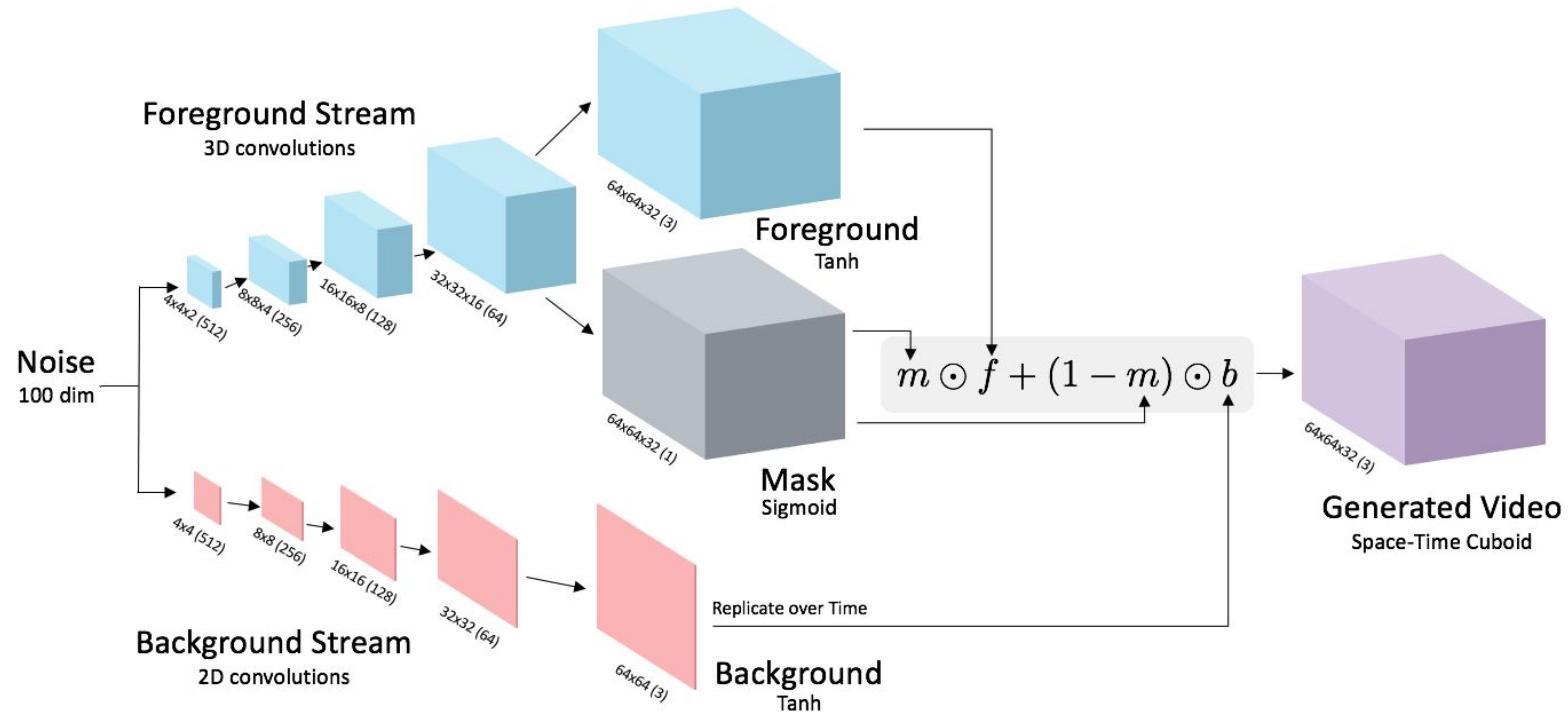


Frame Prediction

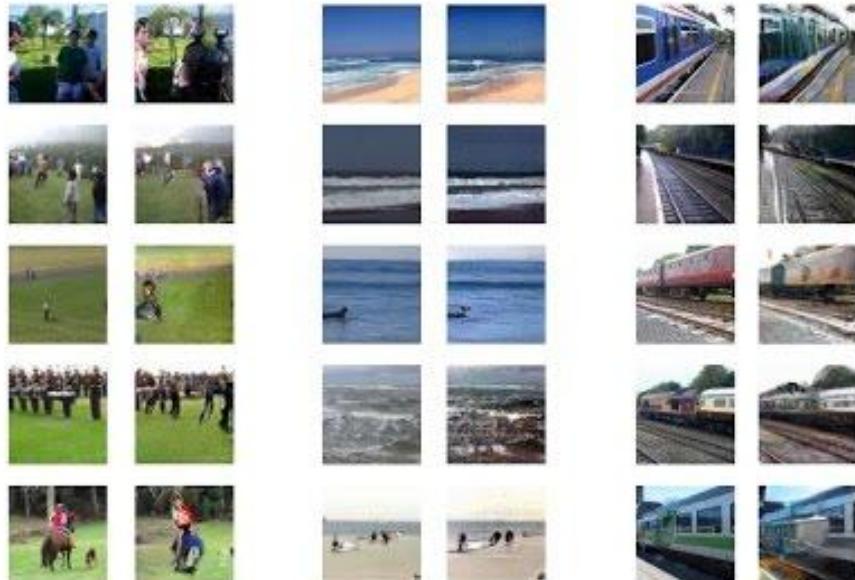


Mathieu, Michael, Camille Couprie, and Yann LeCun. "[Deep multi-scale video prediction beyond mean square error.](#)"
ICLR 2016 [\[project\]](#) [\[code\]](#)

Frame Prediction

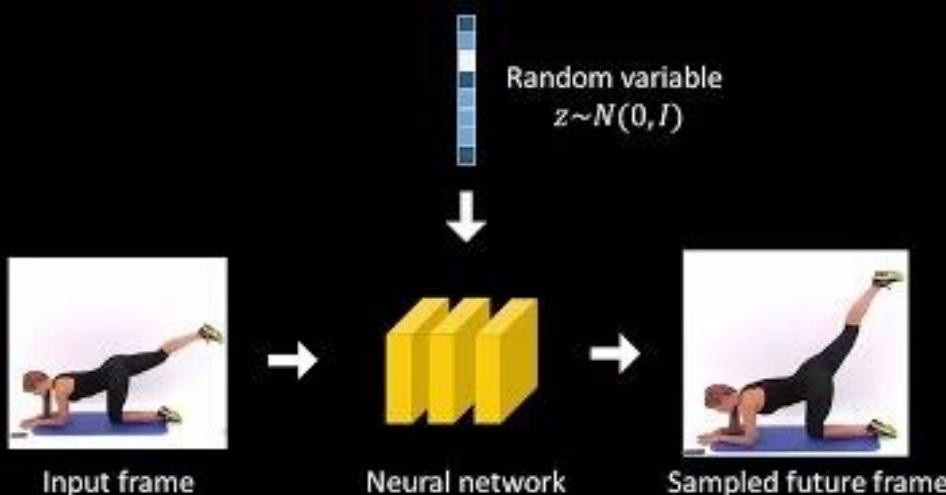


Frame Prediction



Vondrick, Carl, Hamed Pirsiavash, and Antonio Torralba. "[Generating videos with scene dynamics.](#)" NIPS 2016. 99

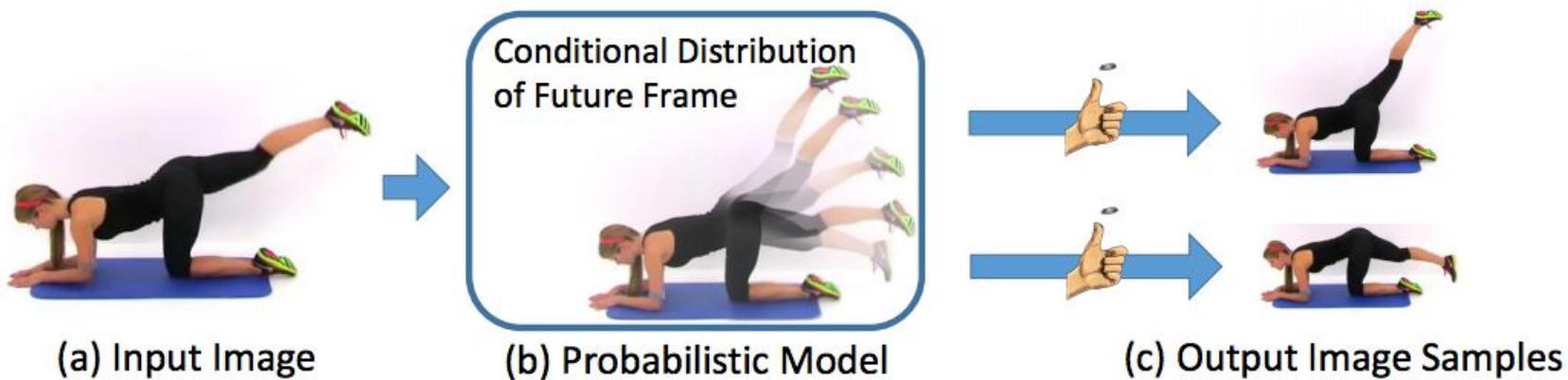
Sampling Future Frames



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

Frame Prediction

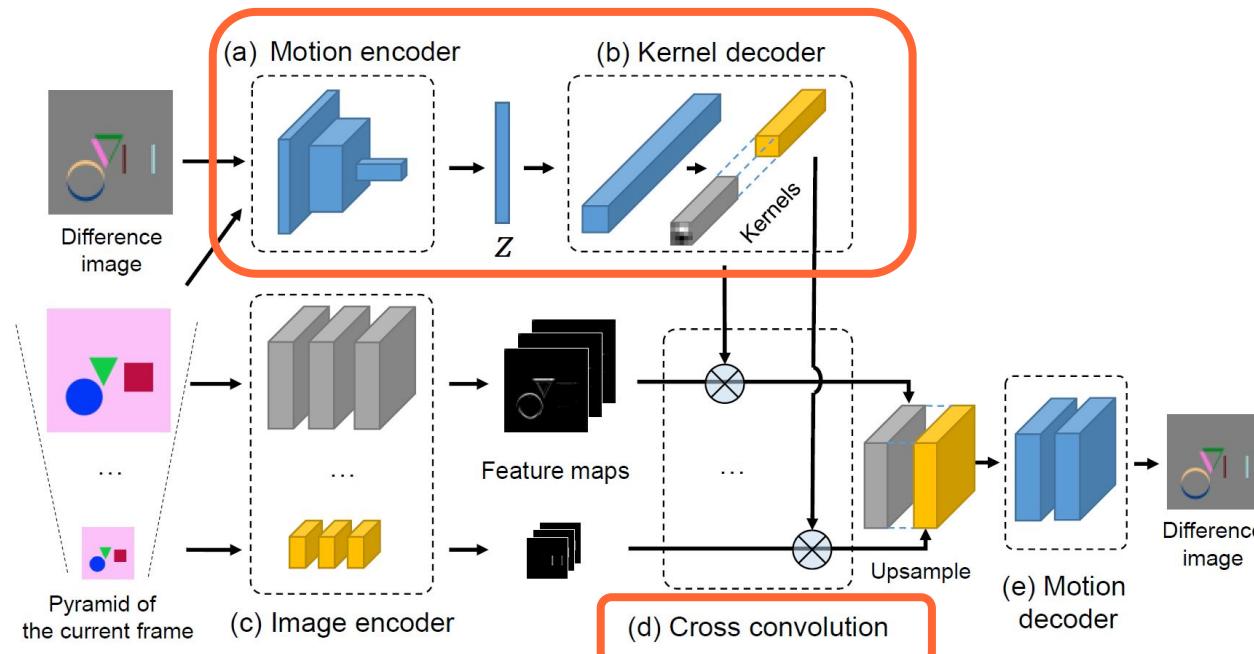
Given an input image, probabilistic generation of future frames with a Variational AutoEncoder (VAE).



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.

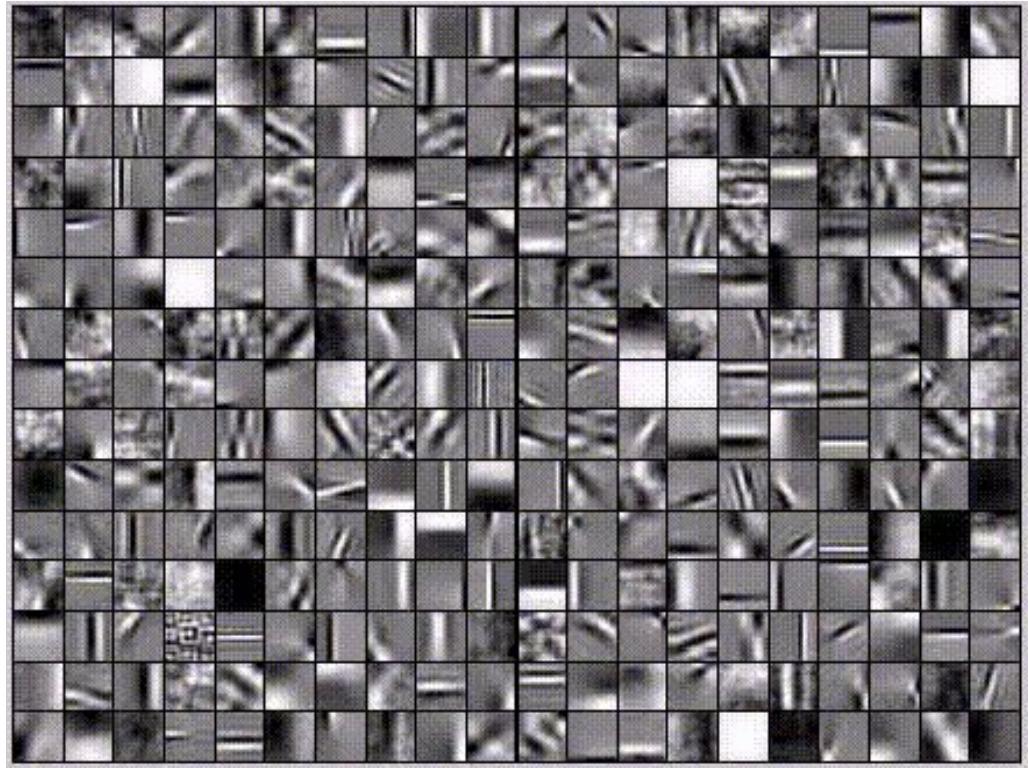
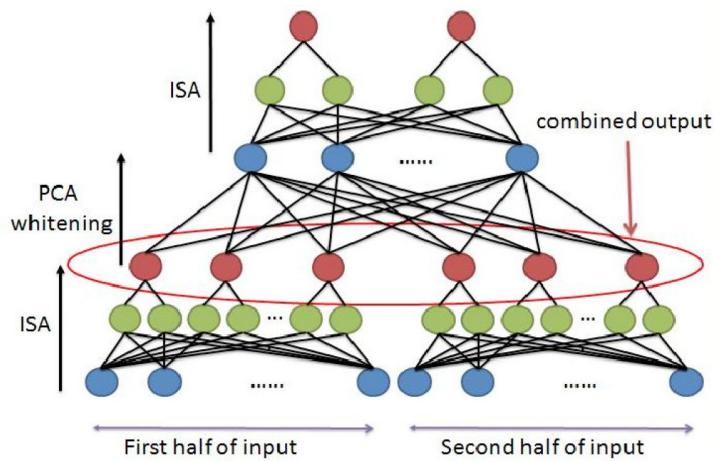


Xue, Tianfan, Jiajun Wu, Katherine Bouman, and Bill Freeman. "[Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks.](#)" NIPS 2016 [\[video\]](#)

Outline

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

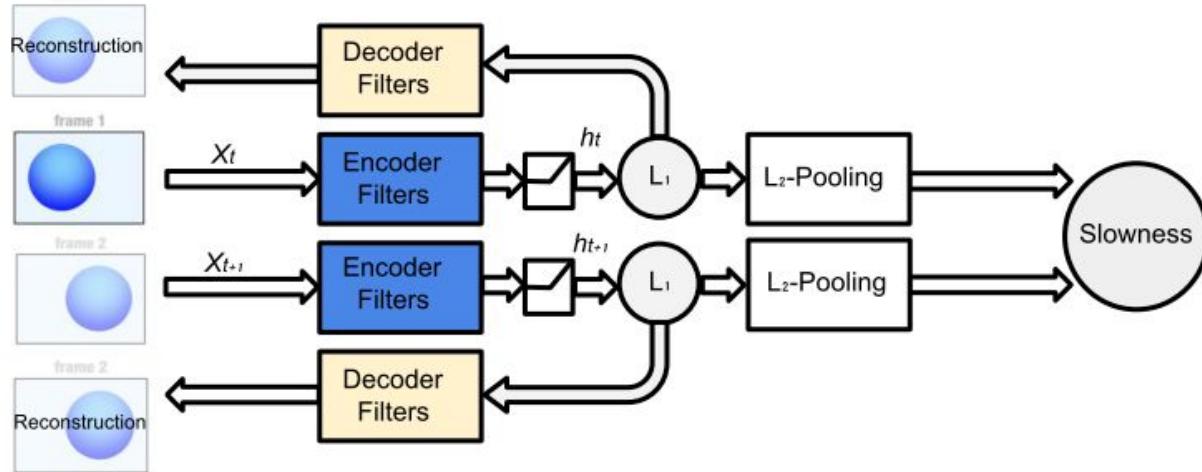
First steps in video feature learning



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 slowness and sparsity.



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

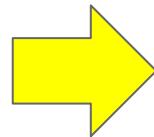
Temporal Weak Labels

Slow feature analysis

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

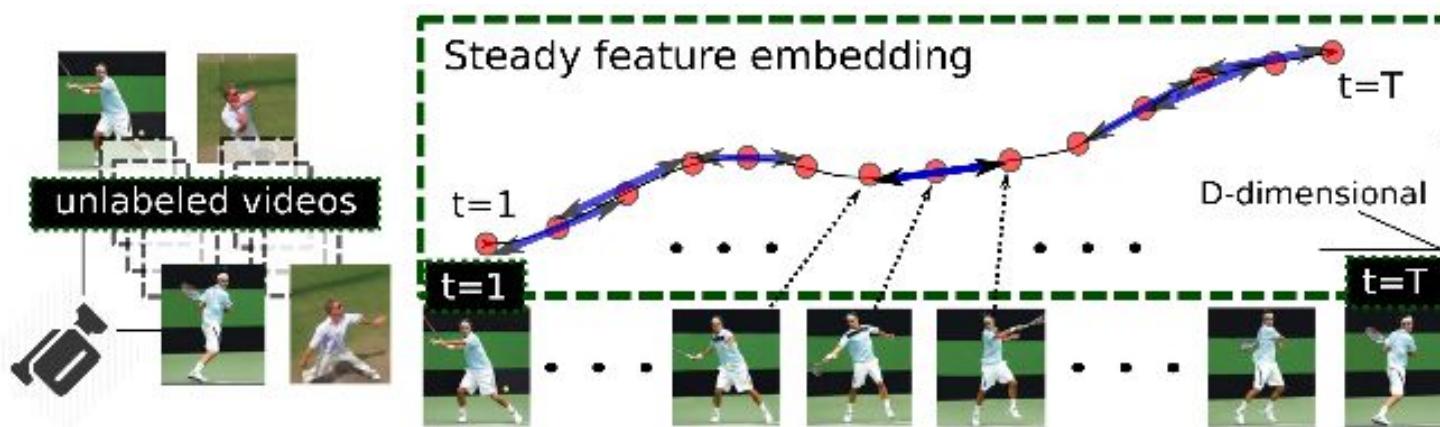
Steady feature analysis

- 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

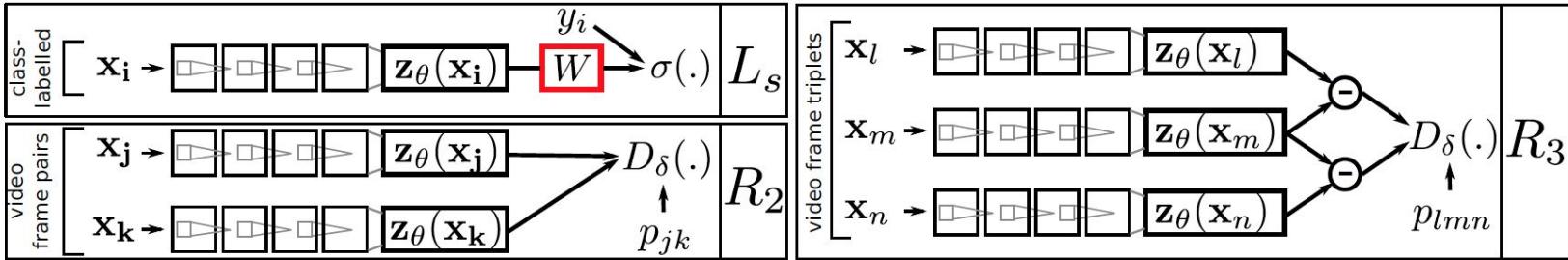


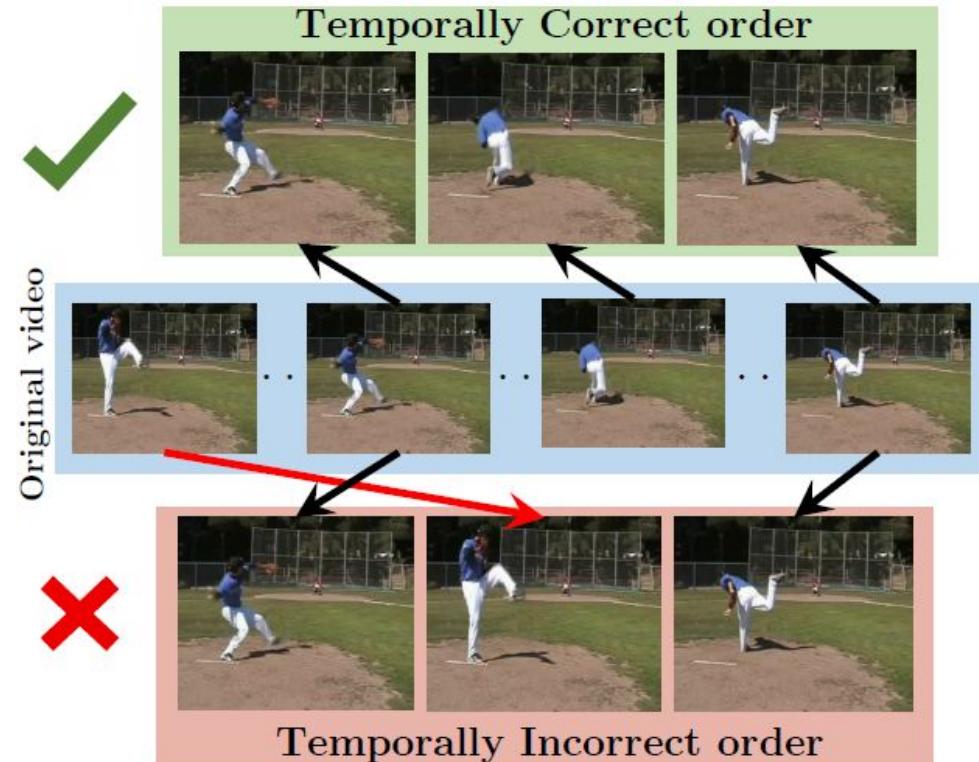
Figure 2: “Siamese” network configuration (shared weights for the \mathbf{z}_θ layer stacks) with portions corresponding to the 3 terms L_s , R_2 and R_3 in our objective. R_2 and R_3 compose the unsupervised loss L_u in Eq (1). L_s is the supervised loss for recognition in static images.

Task	Img/frame dims	#Classes	Recog. Task	#Train	#Test	Unsup. Input Type	#Pairs (1:3)	#Triplets (1:1)	Datasets→	NORB	KITTI	HMDB
NORB→NORB	96×96×1	25	object	150	8100	pose-reg. images	50,000	75,000	SFA-1 [30]	0.95	31.04	2.70
KITTI→SUN	32×32×1	397	scene	2382	7940	car-mounted video	100,000	100,000	SFA-2 [14]	0.91	8.39	2.27
HMDB→PASCAL-10	32×32×3	10	action	50	2000	web video	100,000	100,000	SSFA (ours)	0.53	7.79	1.78

Table 1: Left: Statistics for the unsupervised and supervised datasets ($\mathcal{U} \rightarrow \mathcal{S}$) used in the recognition tasks (positive to negative ratios for pairs and triplets indicated in headers). **Right:** Sequence completion normalized correct candidate rank η . Lower is better. (See Sec 4.2.)

Temporal Weak Labels

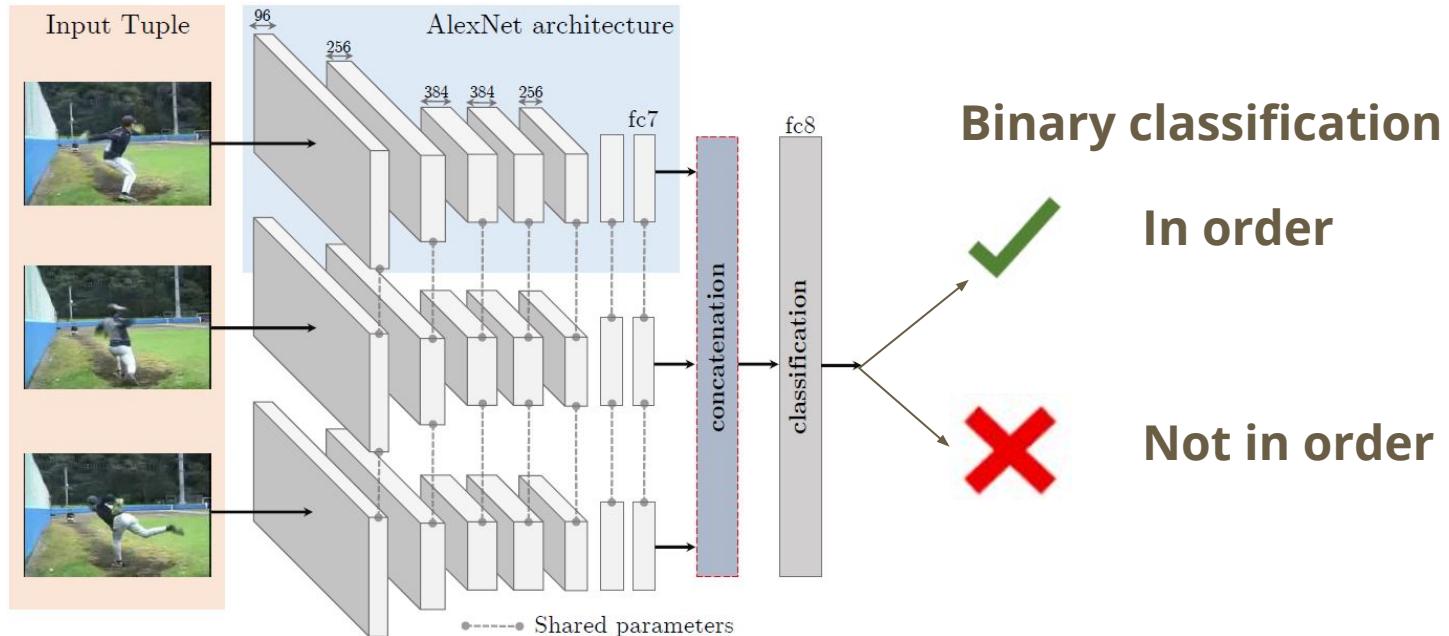
Temporal order of frames is exploited as the supervisory signal for learning.



Temporal Weak Labels

Take temporal order as the supervisory signals for learning

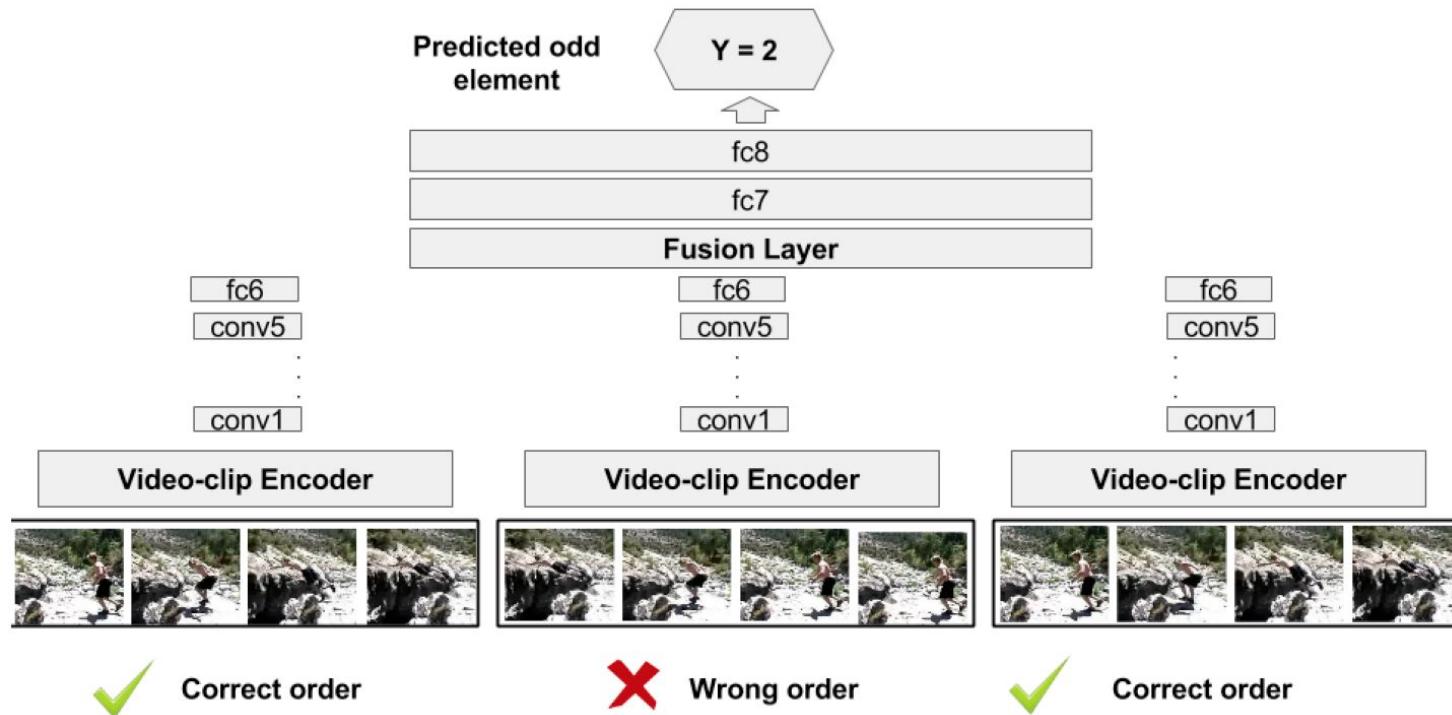
Shuffled
sequences



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

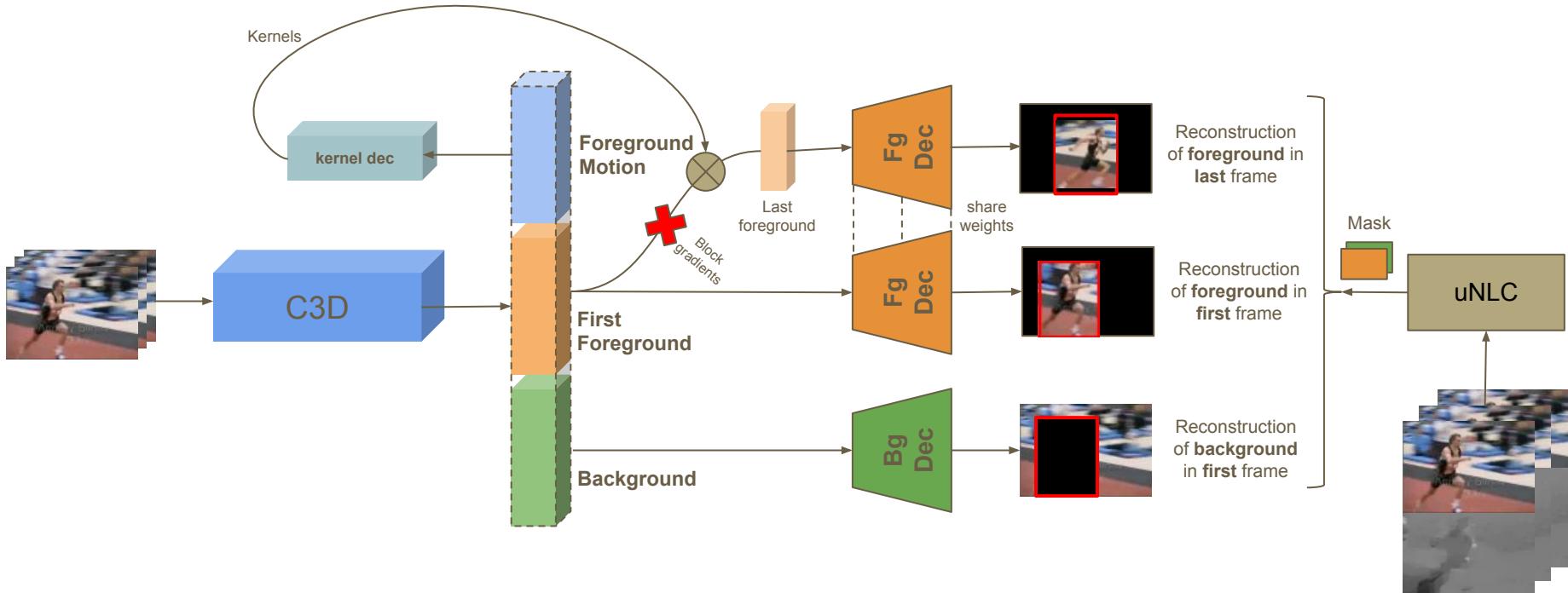
Temporal Weak Labels

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

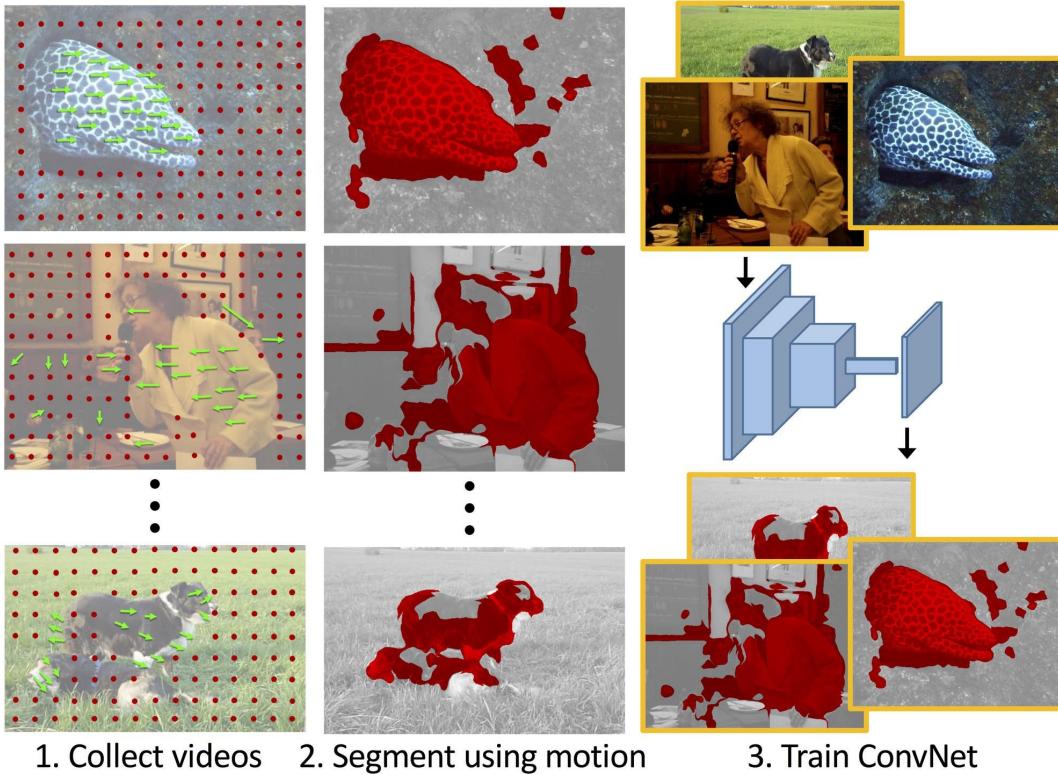


Fernando, Basura, Hakan Bilen, Efstratios Gavves, and Stephen Gould. "[Self-supervised video representation learning with odd-one-out networks.](#)" CVPR 2017

Spatio-Temporal Weak Labels

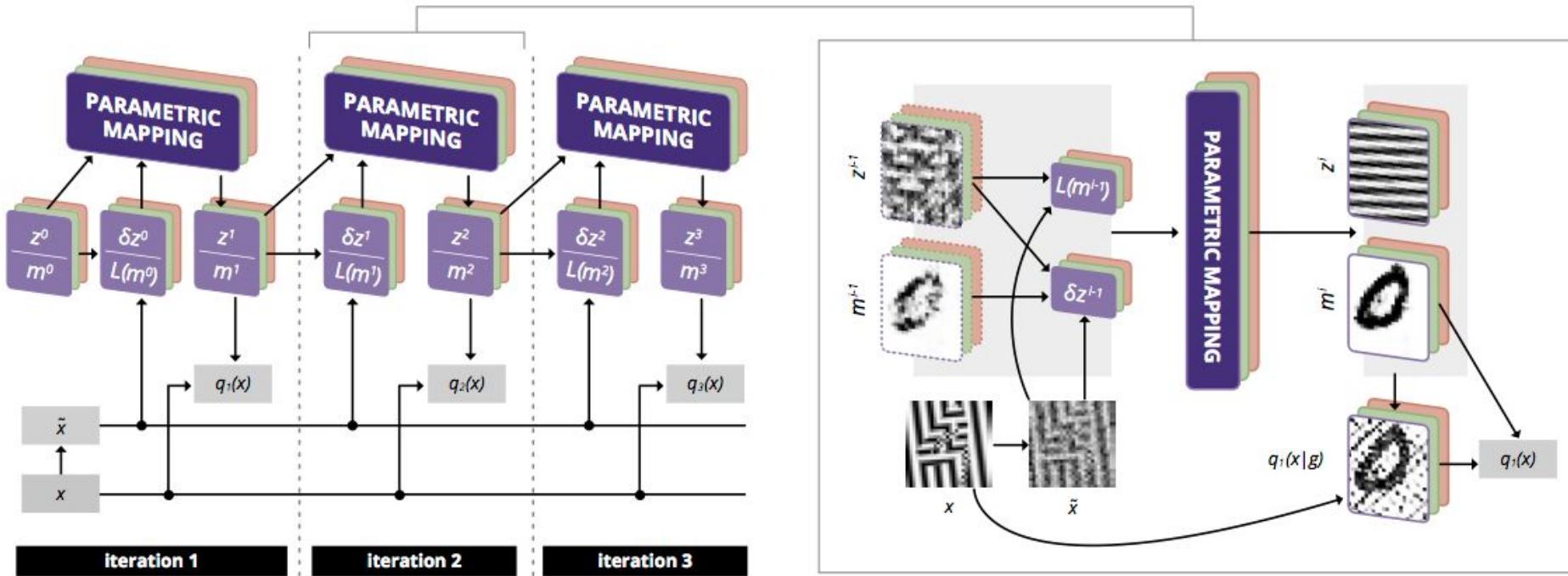


Spatio-Temporal Weak Labels



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 [\[video\]](#) [\[code\]](#)

Predicted Objects and Scenes from Sound Only

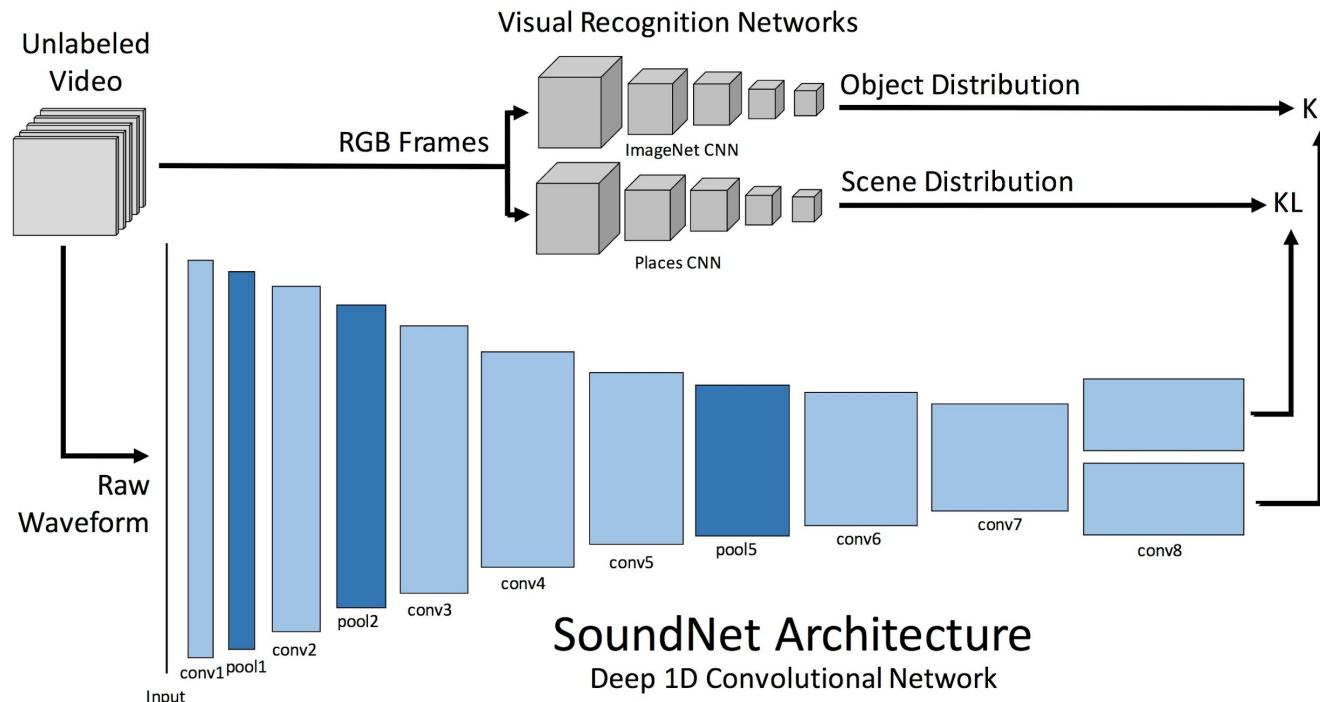


(Videos are blurred so you can try to recognize yourself!)

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

Audio Features from Visual weak labels

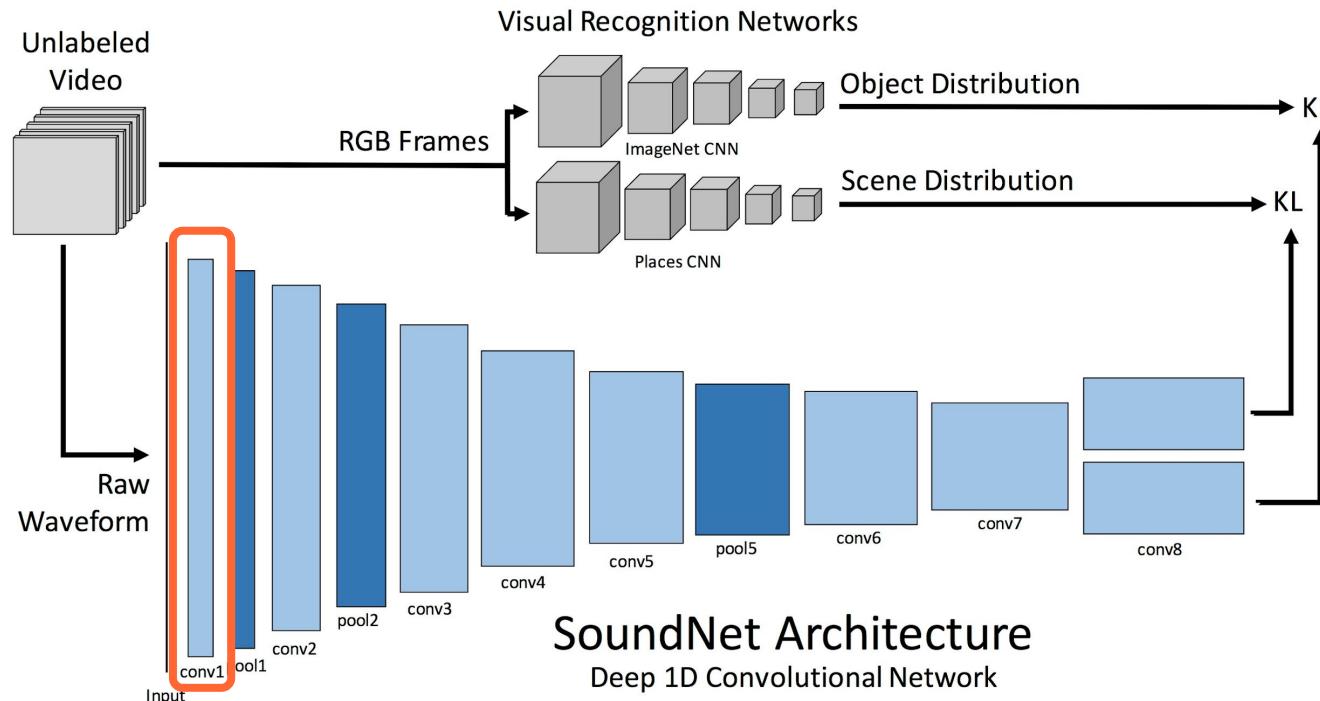
Object & Scenes recognition in videos by analysing the audio track (only).



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

Audio Features from Visual weak labels

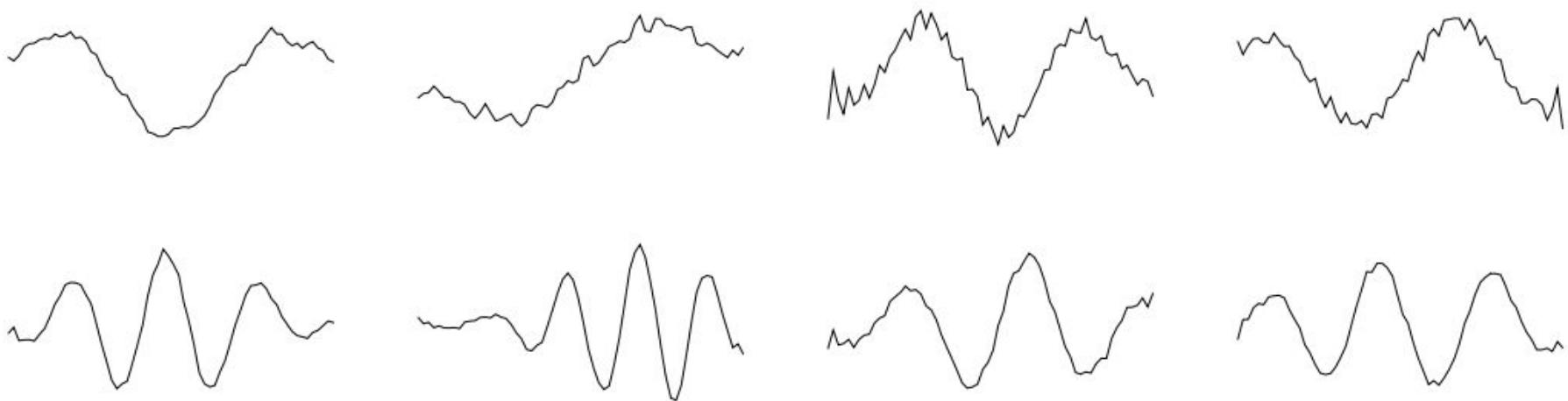
Visualization of the 1D filters over raw audio in conv1.



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

Audio Features from Visual weak labels

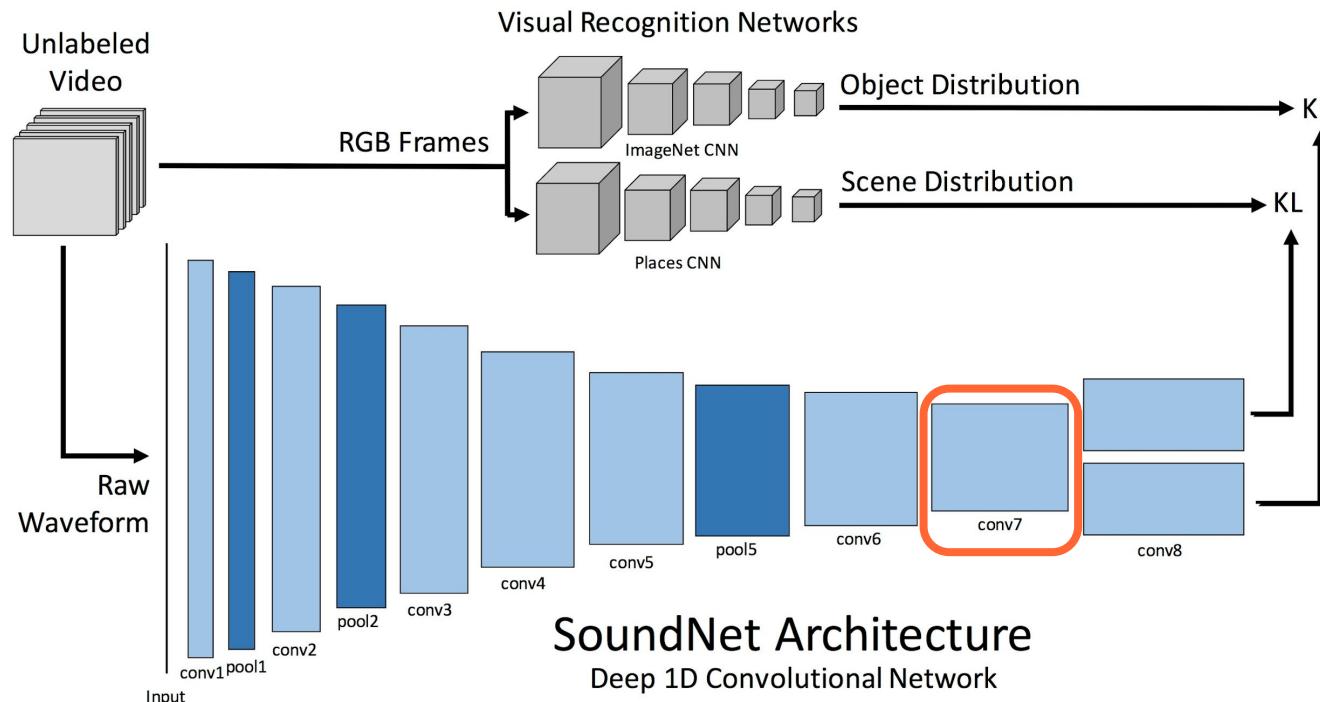
Visualization of the 1D filters over raw audio in conv1.



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

Audio Features from Visual weak labels

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 Features from Visual weak labels

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



Baby Talk

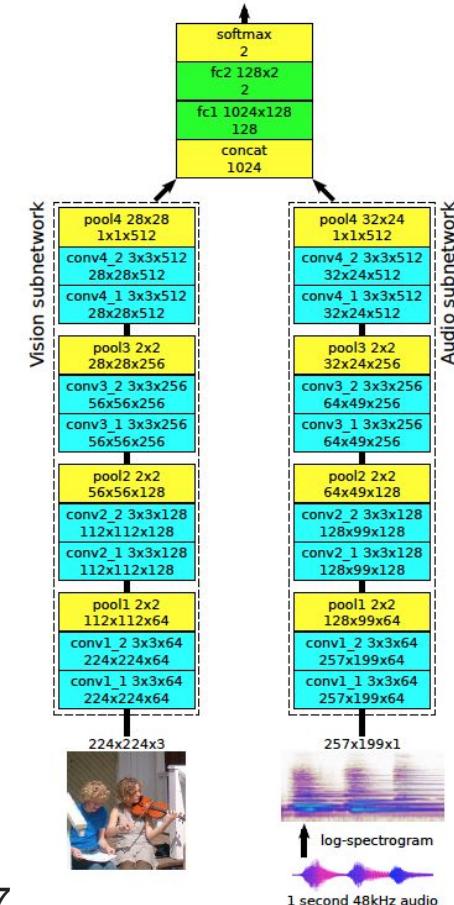
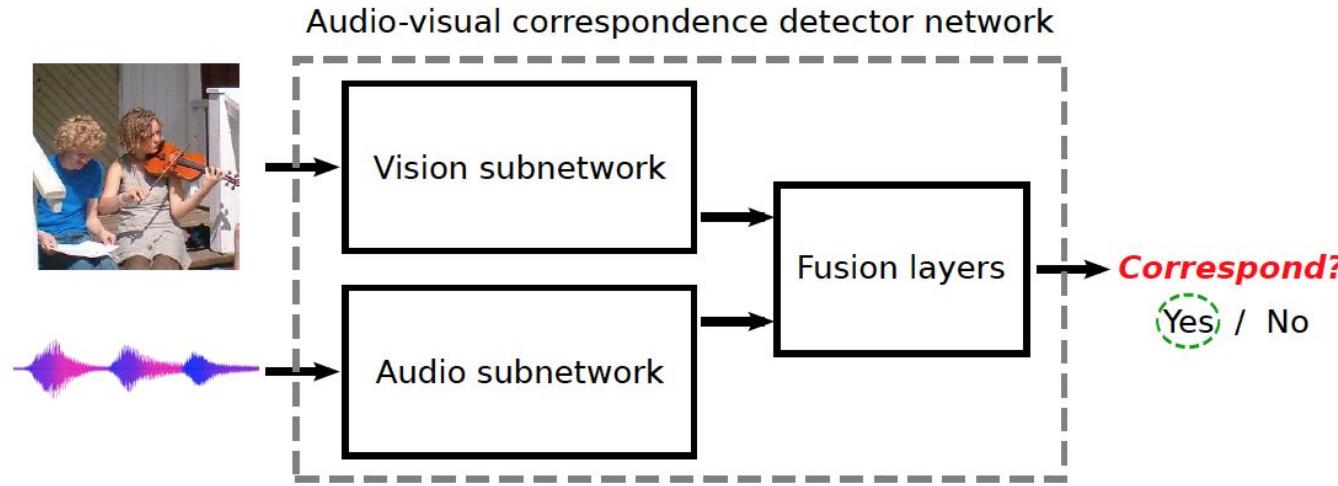


Bubbles

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

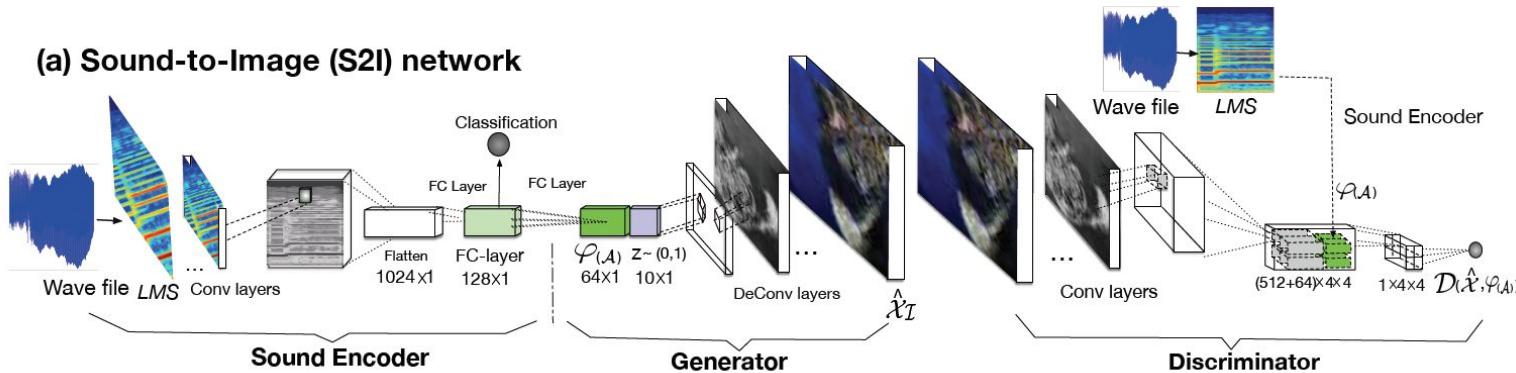
Audio & Visual features from alignment

Audio and visual features learned by assessing alignment.

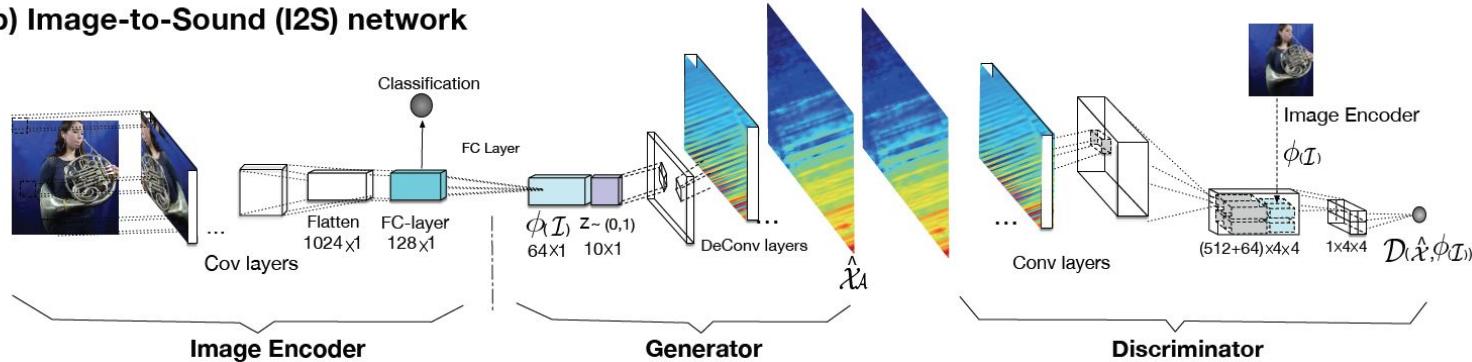


Audio & Visual features from alignment

(a) Sound-to-Image (S2I) network



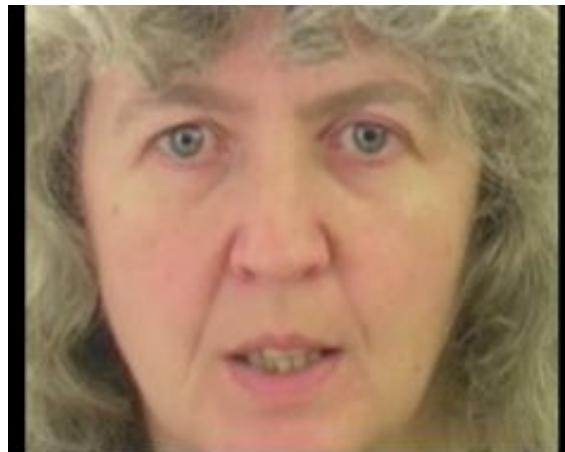
(b) Image-to-Sound (I2S) network



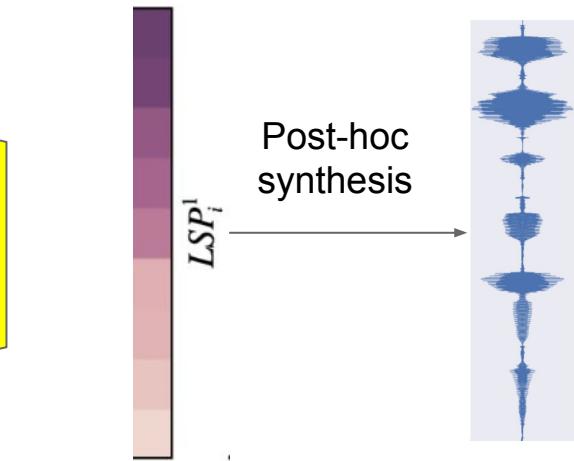
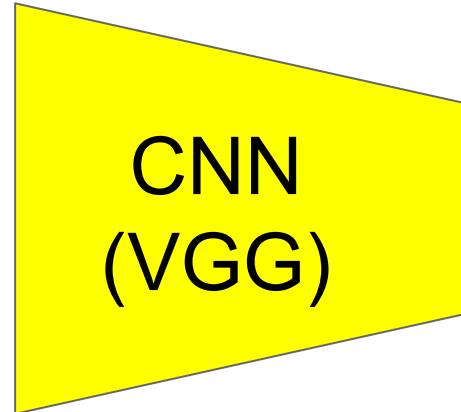


Ephrat, Ariel, and Shmuel Peleg. "Vid2speech: Speech Reconstruction from Silent Video." ICASSP 2017

Video to Speech Representations



Frame from a
silent video



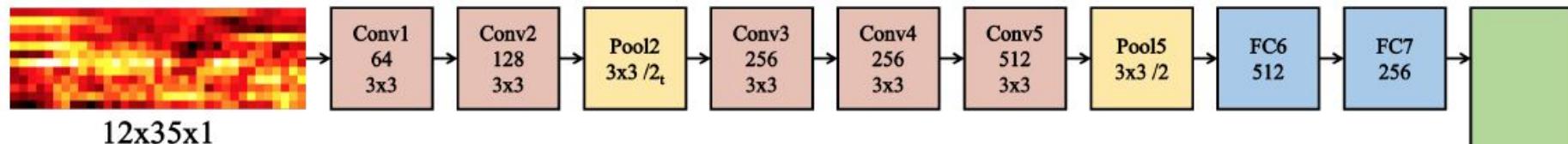
Audio feature



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

Speech to Video Synthesis (mouth)

Audio Encoder



Identity Encoder

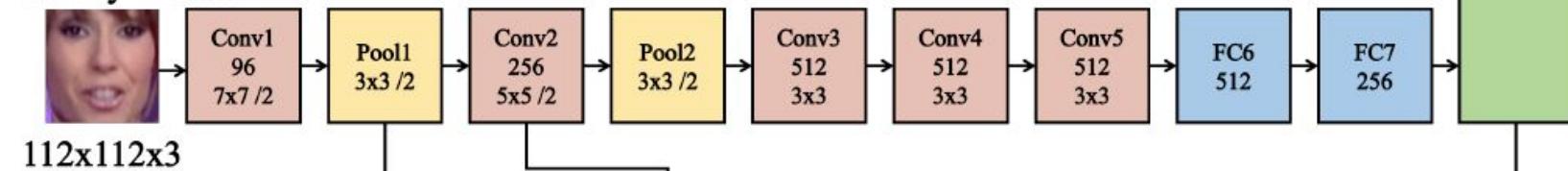
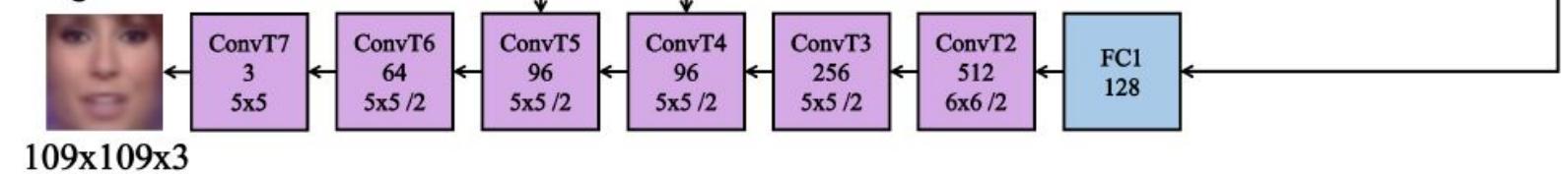
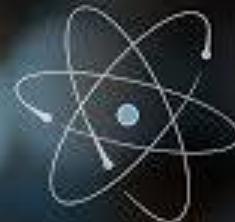


Image Decoder





TWO MINUTE PAPERS

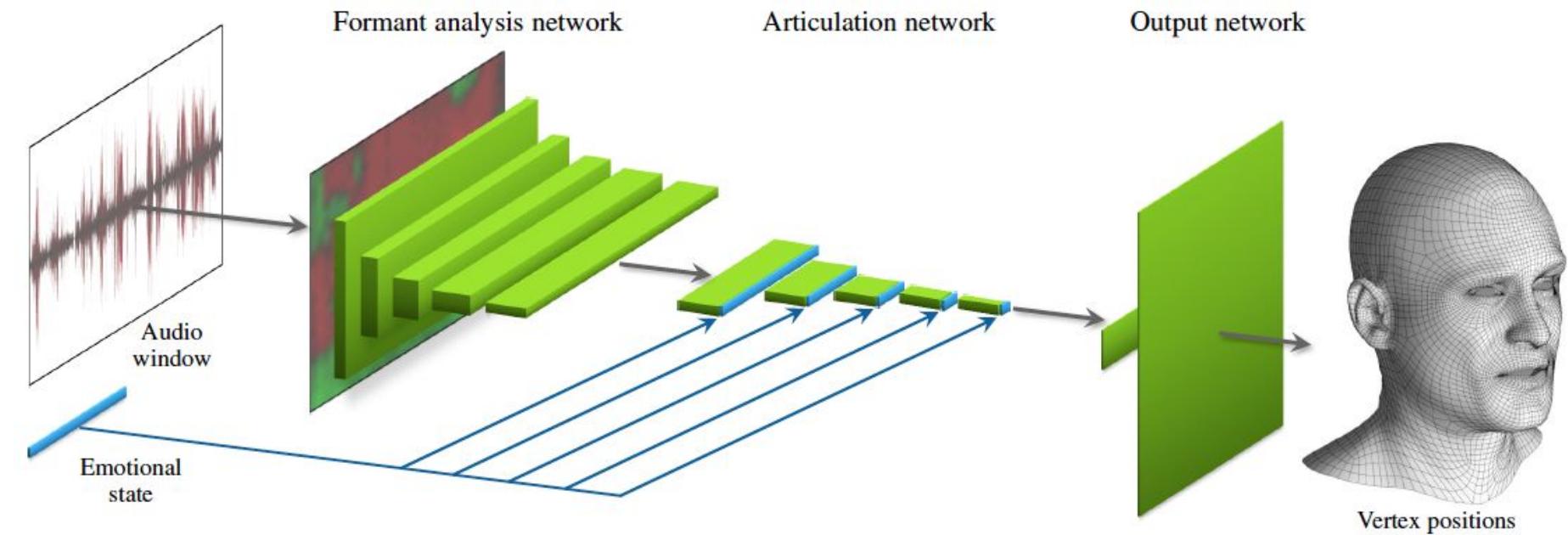
WITH KÁROLY ZSOLNAY-FEHÉR (KZF)

AI CREATES FACIAL ANIMATION FROM AUDIO

Disclaimer: I was not part of this research project; I am merely providing commentary on this work.

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



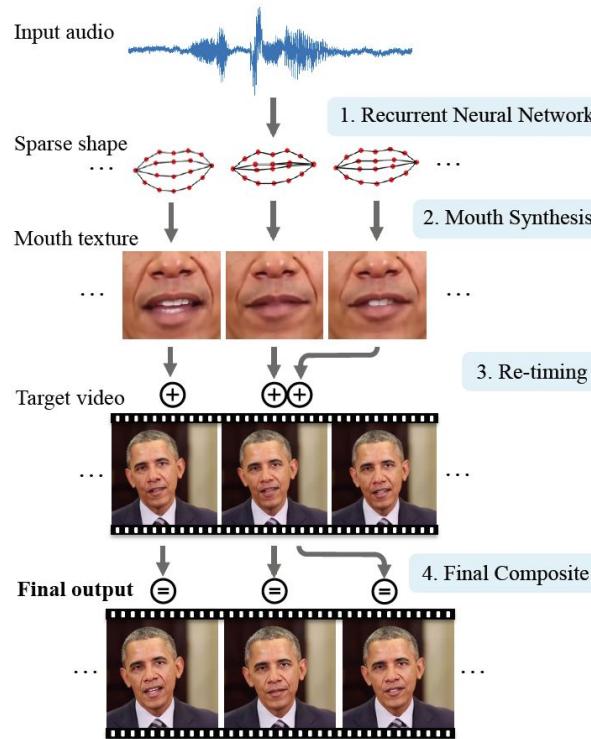
Without Re-timing



With Re-timing
(Our Result)

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.

Outline

1. Motivation
2. Unsupervised Learning
3. Predictive Learning
4. Self-supervised Learning