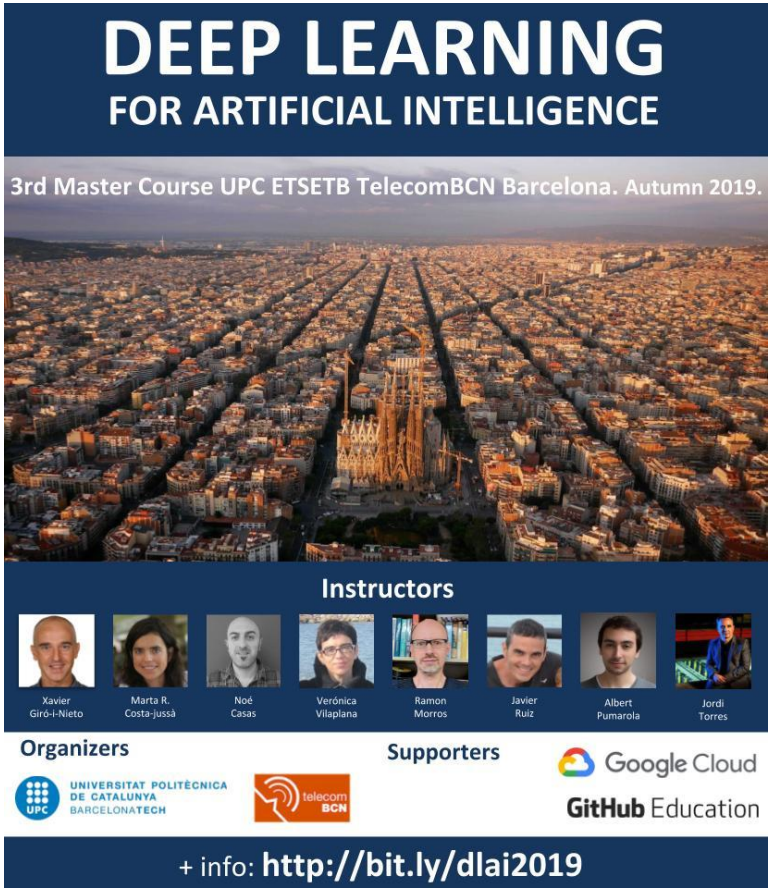


[course site]

#DLBCN

Day 10 Lecture 1

Self-supervised Learning

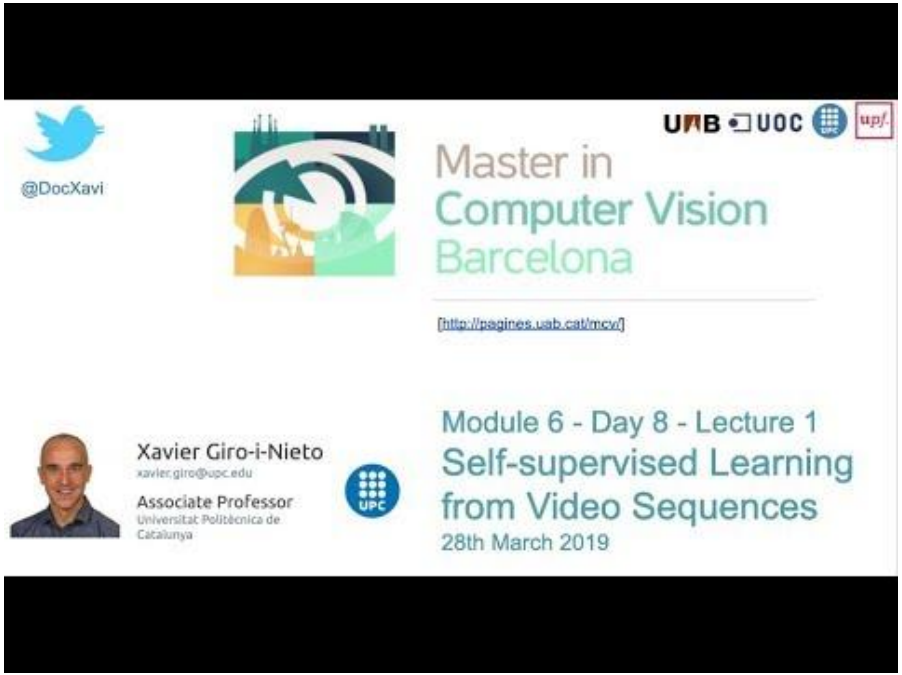
Xavier Giro-i-Nieto
xavier.giro@upc.edu
@DocXavi

Associate Professor
Universitat Politècnica de Catalunya
Technical University of Catalonia

telecom BCN

UPC

Video lecture



Lectures on applied deep learning for video in [Master in Computer Vision Barcelona 2019]


The Revolution will not be Supervised
BAIR
Alexei A. Efros
UC Berkeley

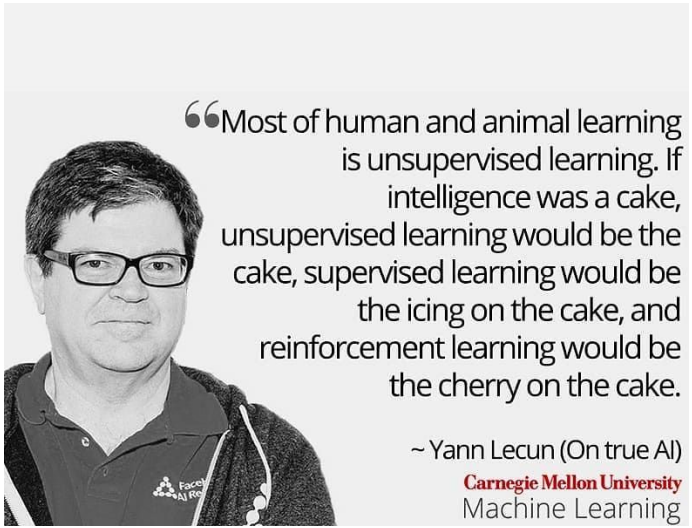
Outline

1. **Unsupervised Learning**
2. Self-supervised Learning
    a. Autoencoder
    b. Temporal regularisations
    c. Temporal verifications
    d. Predictive Learning
    e. Miscellaneous: optical flow, color & multiview

Types of machine learning

Yann Lecun's Black Forest cake

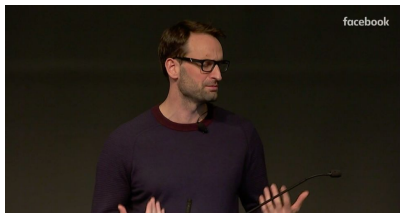
- **"Pure" Reinforcement Learning (cherry)**
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- **Supervised Learning (icing)**
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits 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)**

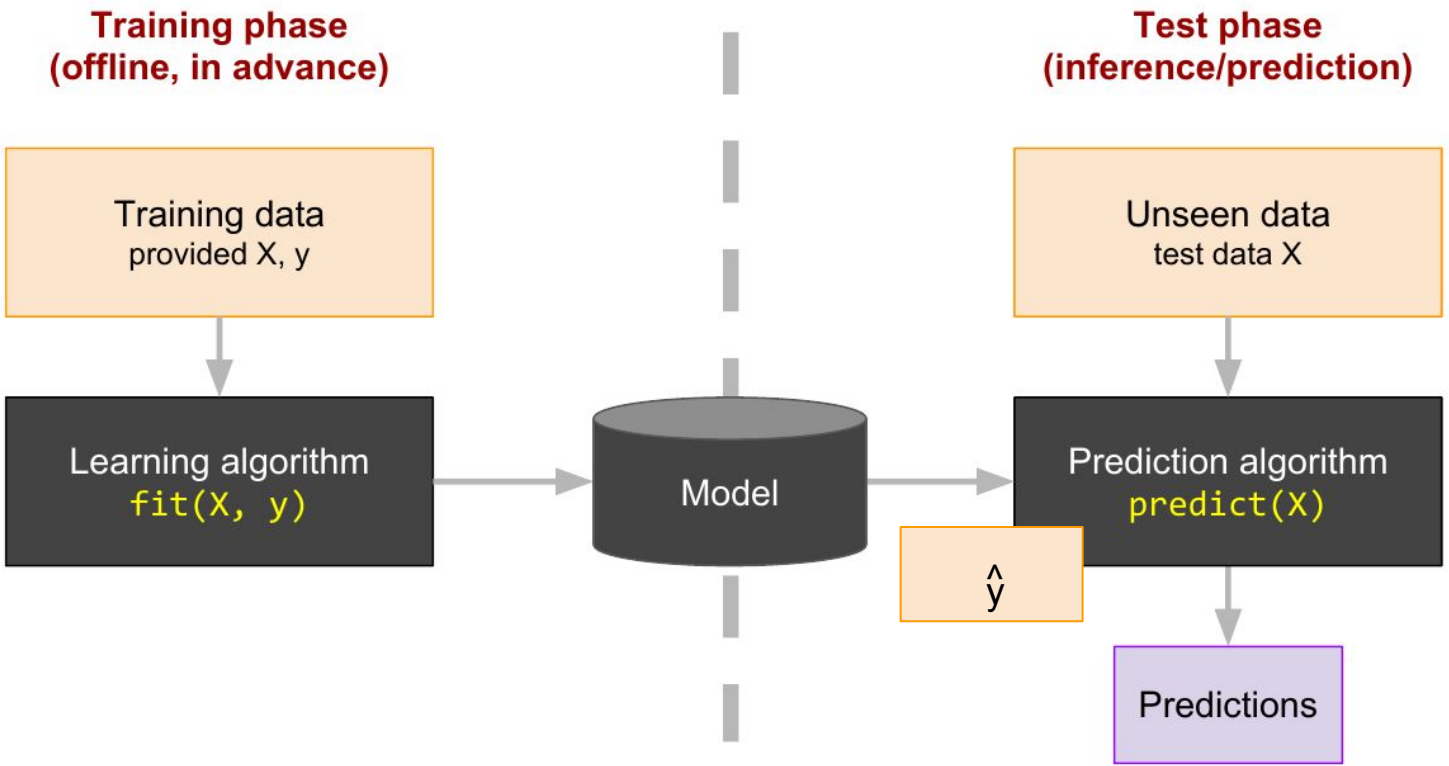
A broader picture of types of learning…

	…with a teacher	…without a teacher
Active agent…	Reinforcement learning (with extrinsic reward)	Intrinsic motivation / Exploration.
Passive agent…	Supervised learning	Unsupervised learning



Slide inspired by Alex Graves (Deepmind) at "Unsupervised Learning Tutorial" @ NeurIPS 2018.

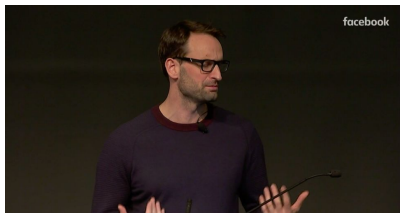
Supervised learning



Slide credit: Kevin McGuinness

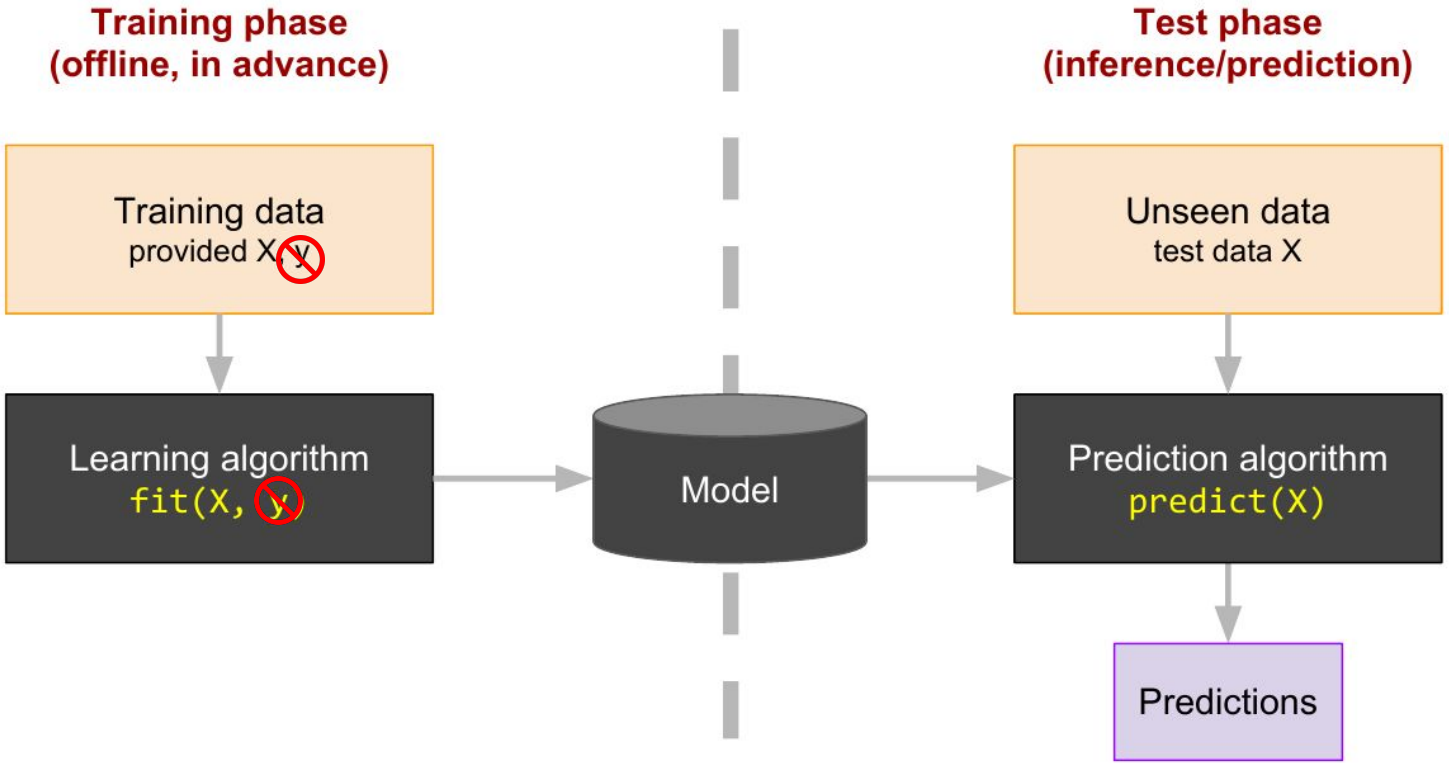
A broader picture of types of learning…

	…with a teacher	…without a teacher
Active agent…	Reinforcement learning (with extrinsic reward)	Intrinsic motivation / Exploration.
Passive agent…	Supervised learning	Unsupervised learning



Slide inspired by Alex Graves (Deepmind) at "Unsupervised Learning Tutorial" @ NeurIPS 2018.

Unsupervised learning



Slide credit: Kevin McGuinness

Unsupervised learning

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

Video lectures on Unsupervised Learning

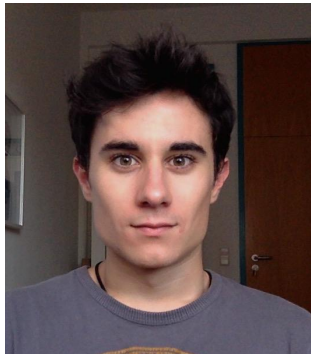
Kevin McGuinness, [UPC DLCV 2016]

Xavier Giró, [UPC DLAI 2017]

Outline

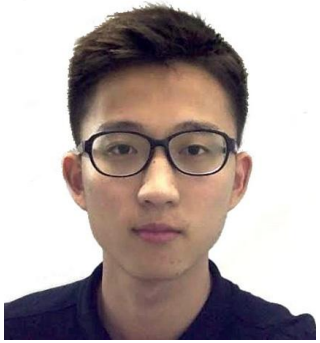
1. Unsupervised Learning
2. **Self-supervised Learning**
    a. Autoencoder
    b. Temporal regularisations
    c. Temporal verifications
    d. Predictive Learning
    e. Miscellaneous: optical flow, color & multiview

Acknowledgements



Víctor Campos

Barcelona Supercomputing Center
Centro Nacional de Supercomputación



Junting Pan

Columbia University
in the City of New York

Xunyu Lin

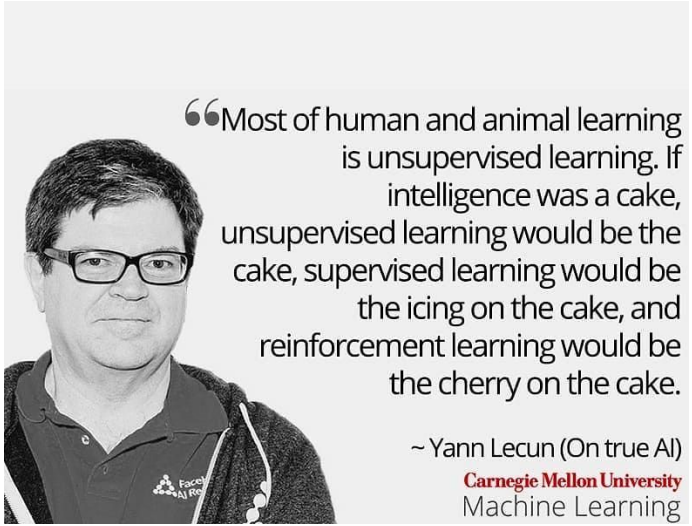
Carlos Arenas

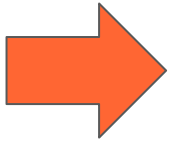
Sebastian Palacio

UNIVERSITAT POLITÈCNICA DE CATALUNYA

German Research Center for Artificial Intelligence

Self-supervised learning





Yann LeCun
@ylecun

I Now call it "self-supervised learning", because "unsupervised" is both a loaded and confusing term.

In self-supervised learning, the system learns to predict part of its input from other parts of it input. In...
facebook.com/722677142/post...

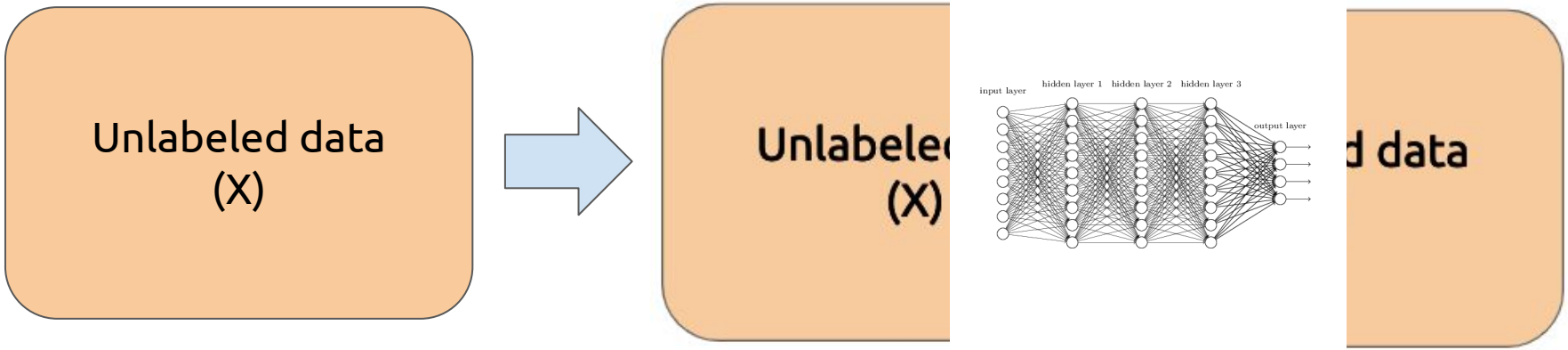
4:40 PM · Apr 30, 2019 · Facebook

(...)
“Doing this properly and reliably is **the greatest challenge in ML and AI** of the next few years in my opinion.”

Self-supervised learning

Self-supervised learning is a form of unsupervised learning where the data provides the supervision.

- A **surrogate task** must be invented by withholding a part of the unlabeled data and training the NN to predict it.



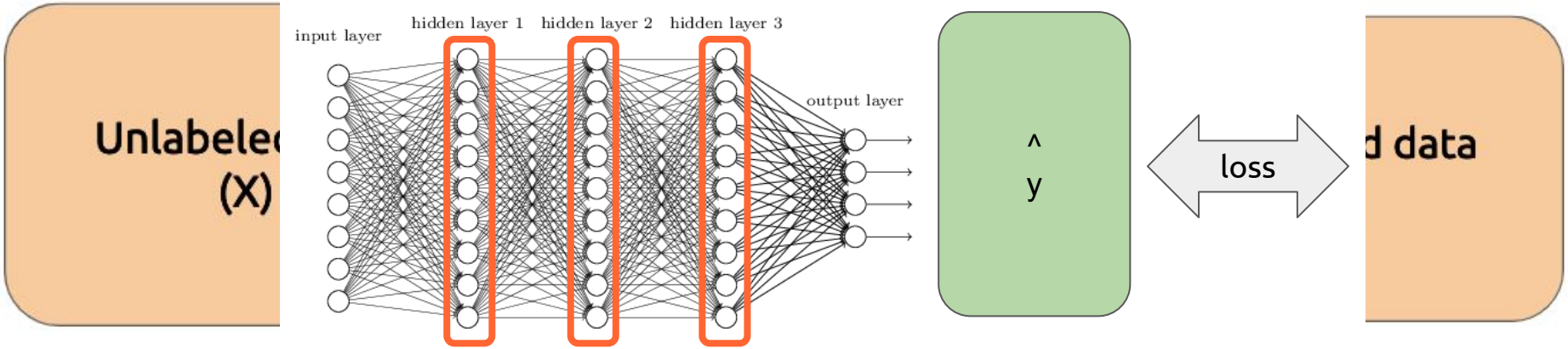
Reference: Andrew Zisserman (PAISS 2018)

Self-supervised learning

Self-supervised learning is a form of unsupervised learning where the data provides the supervision.

- By defining a proxy loss, the NN learns representations, which should be valuable for the actually target task.

Representations learned without labels

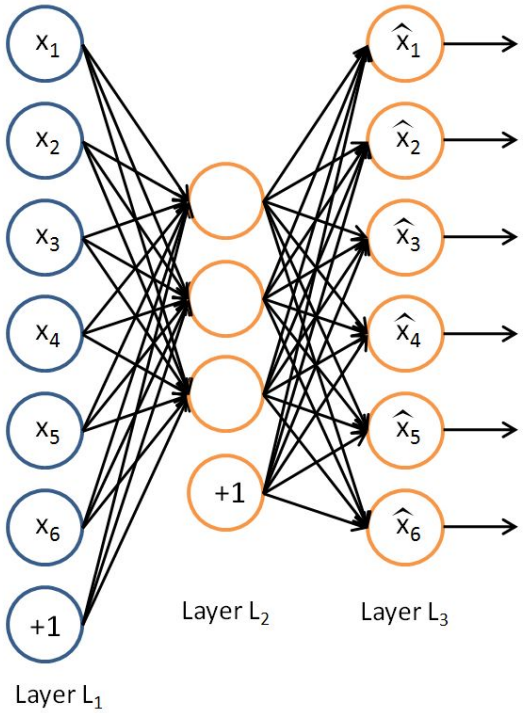


Reference: Andrew Zisserman (PAISS 2018)

Outline

1. Unsupervised Learning
2. Self-supervised Learning
    a. **Autoencoder**
    b. Temporal regularisations
    c. Verifications
    d. Predictive Learning
    e. Miscellaneous: optical flow, color & multiview

Autoencoder (AE)



Autoencoders:

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

Fig: "Deep Learning Tutorial" Stanford

Autoencoder (AE)

What is the use of an autoencoder ?

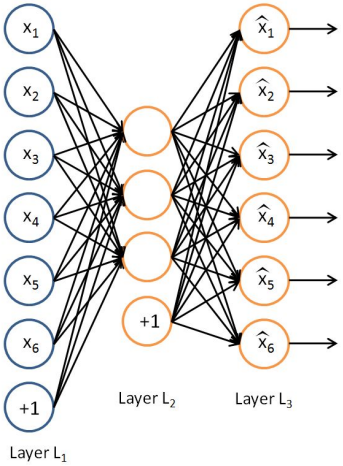
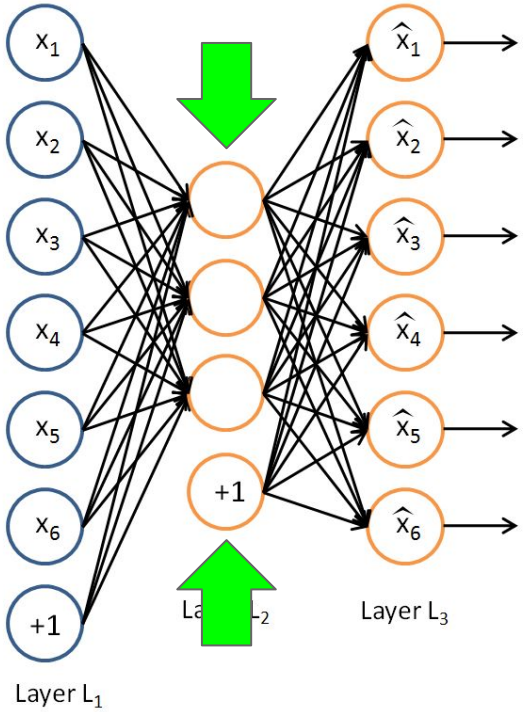


Fig: "Deep Learning Tutorial" Stanford

Autoencoder (AE)



Dimensionality reduction:

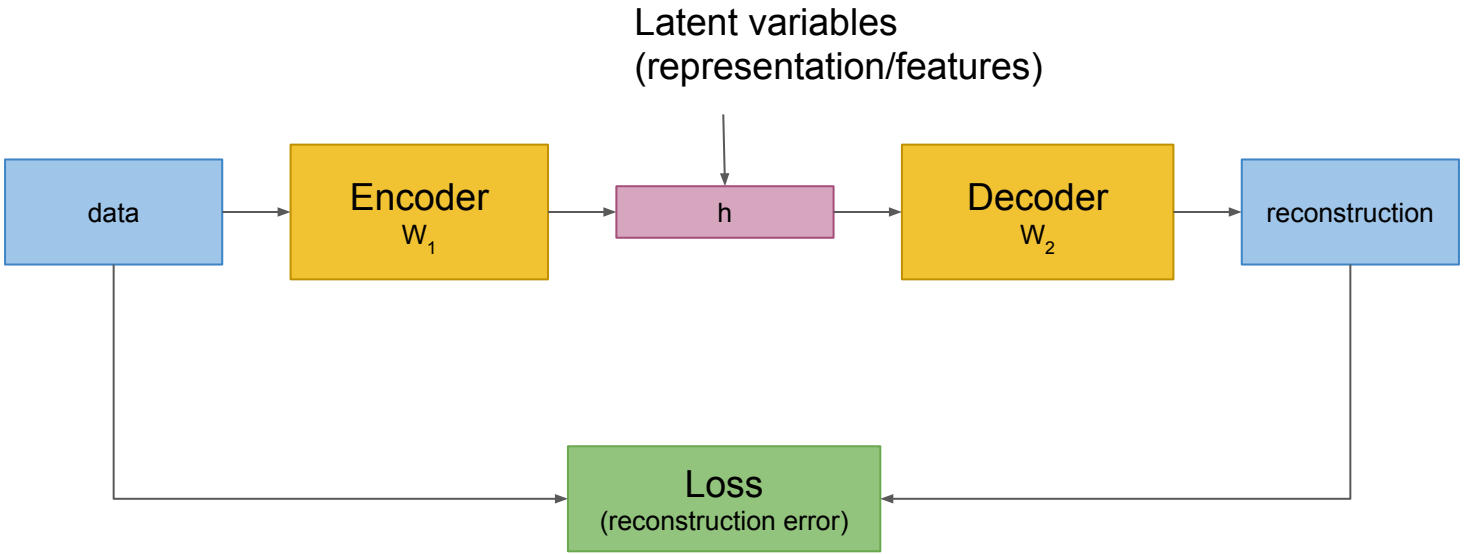
Use the hidden layer as a feature extractor of any desired size.

Fig: "Deep Learning Tutorial" Stanford

Autoencoder (AE)

WHY?

1. Initialize a NN by solving an autoencoding problem.

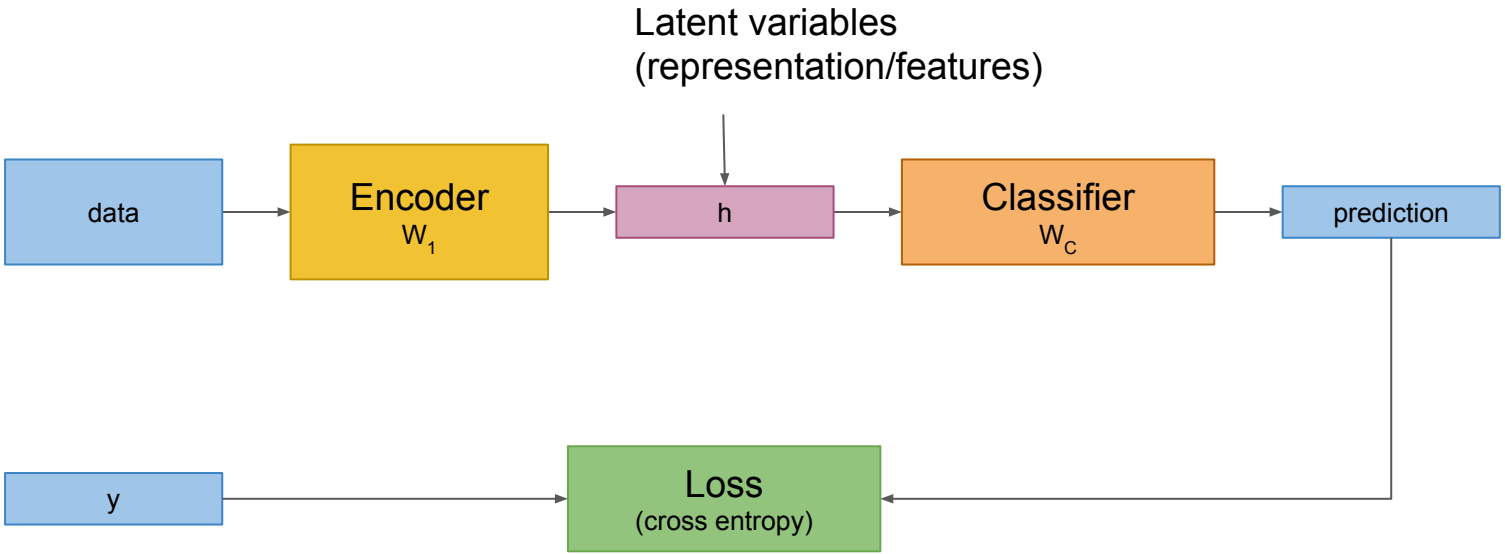


Slide: Kevin McGuinness (DLCV UPC 2017)

Autoencoder (AE)

WHY?

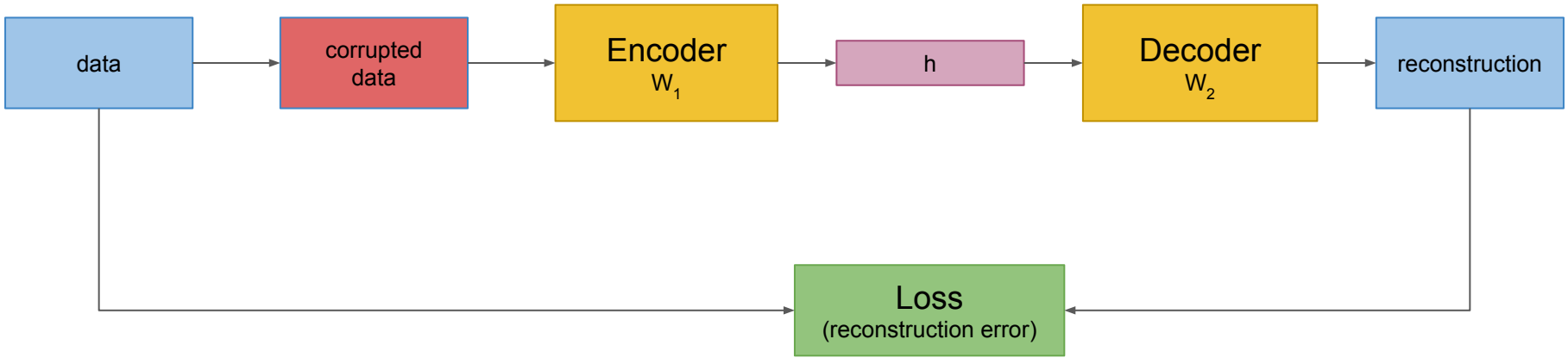
1. Initialize a NN solving an autoencoding problem.
2. Train for final task with “few” labels.



Slide: Kevin McGuinness (DLCV UPC 2017)

Denoising Autoencoder (DAE)

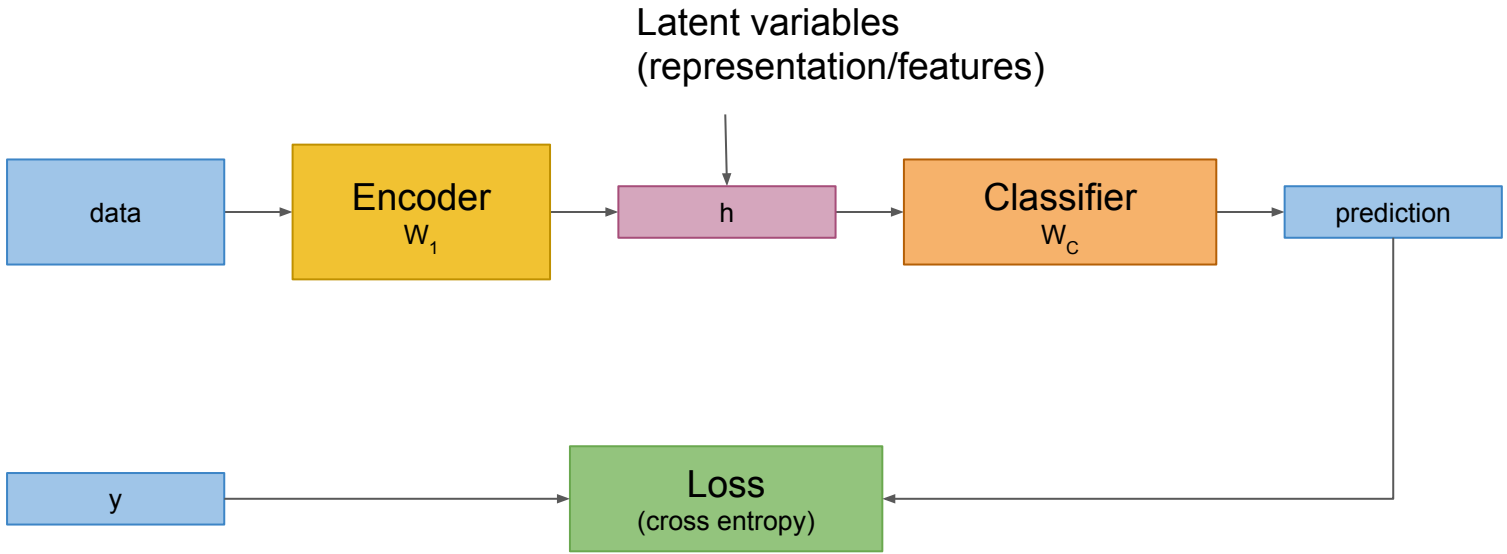
1. Corrupt the input signal, but predict the clean version at its output.
2. Train for final task with “few” labels.



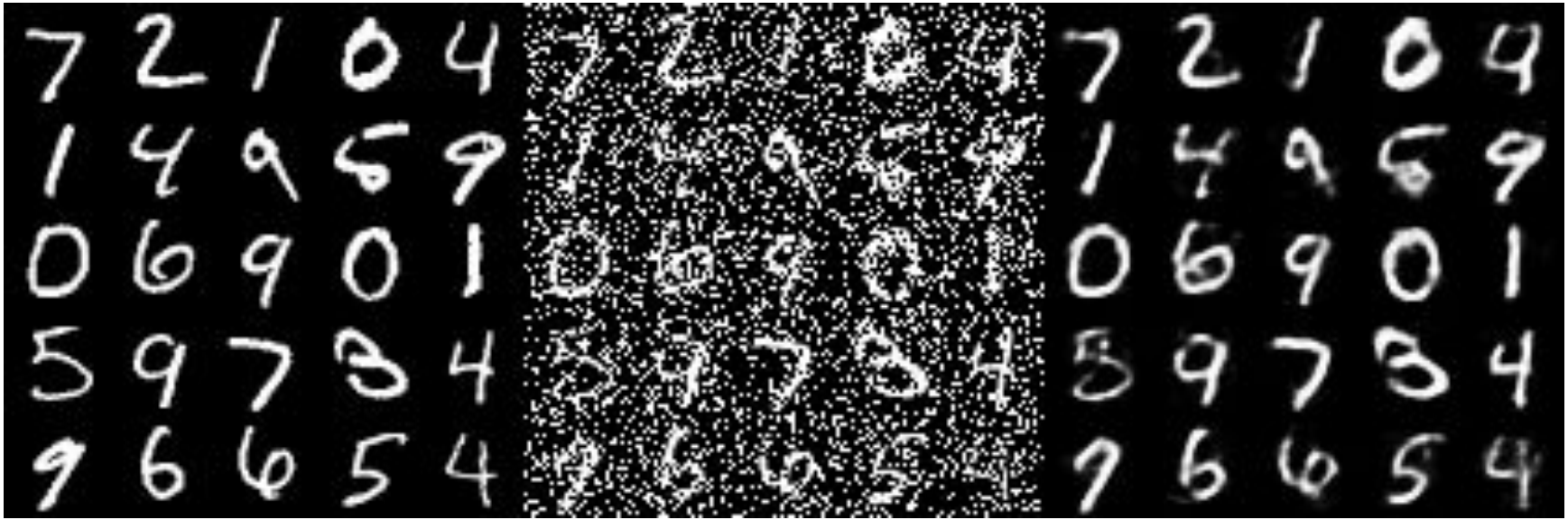
Vincent, Pascal, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. "Extracting and composing robust features with denoising autoencoders." ICML 2008.

Denoising Autoencoder (DAE)

1. Corrupt the input signal, but predict the clean version at its output.
2. Train for final task with “few” labels.



Denoising Autoencoder (DAE)



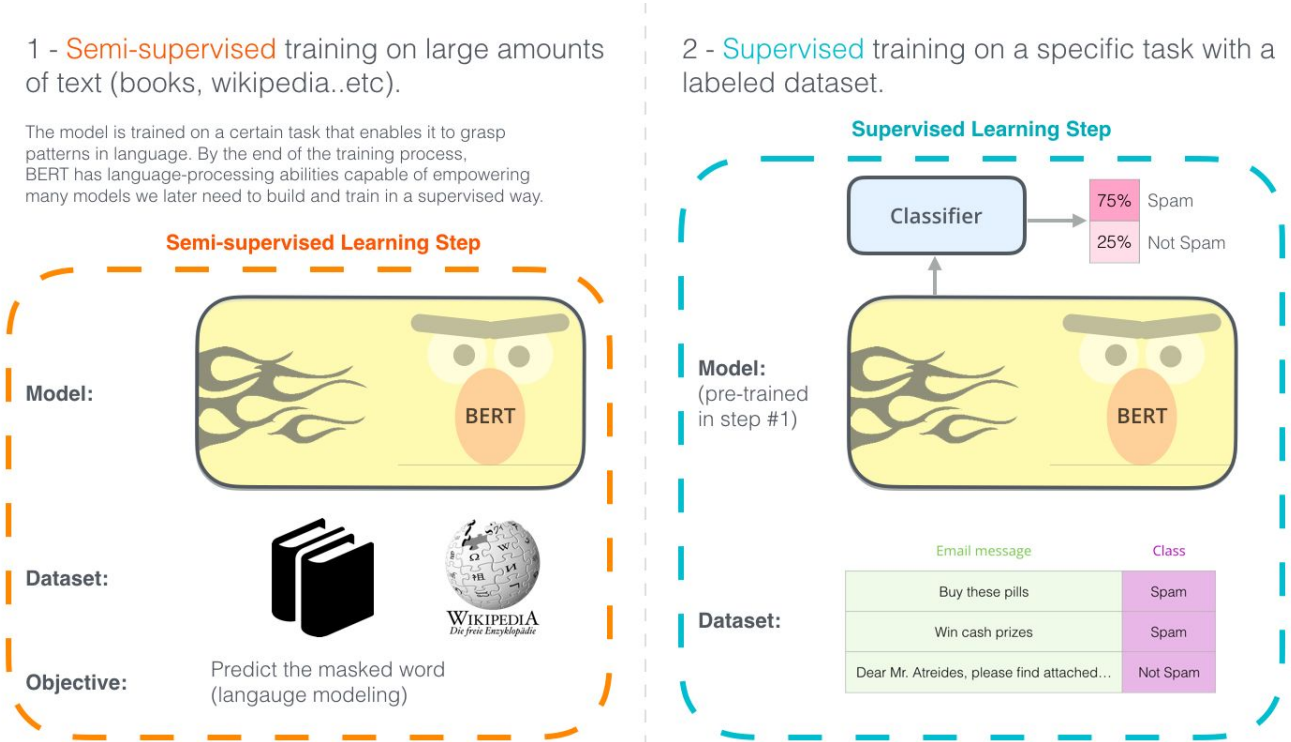
Source: [“Tutorial: Your First Model (DAE)”](#) by Opendeep.org

Eigen, David, Dilip Krishnan, and Rob Fergus. "Restoring an image taken through a window covered with dirt or rain." ICCV 2013.

Masked Autoencoder (MAE)

Show it a sequence of words on input, mask out 15% of the words, and ask the system to predict the missing words (or a distribution of words)

Figure: Jay Alammar



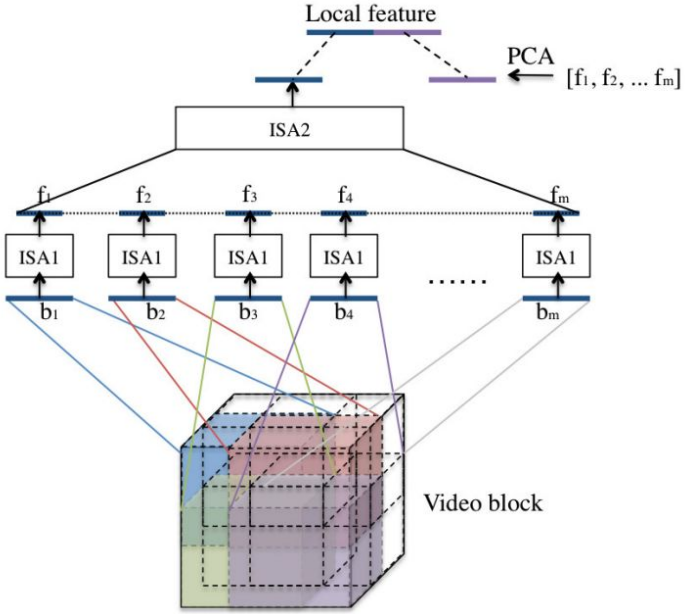
#BERT Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." NAACL 2019 **(best paper award)**.

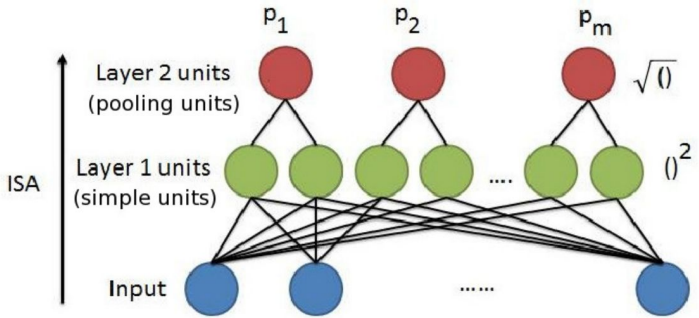
Outline

1. Unsupervised Learning
2. Self-supervised Learning
    a. Autoencoder
    b. Temporal regularisations
    c. Verifications
    d. Predictive Learning
    e. Miscellaneous: optical flow, color & multiview

Temporal regularization: ISA

Video features are learned with convolutions and poolings combined with a Independent Subspace Analysis (ISA).



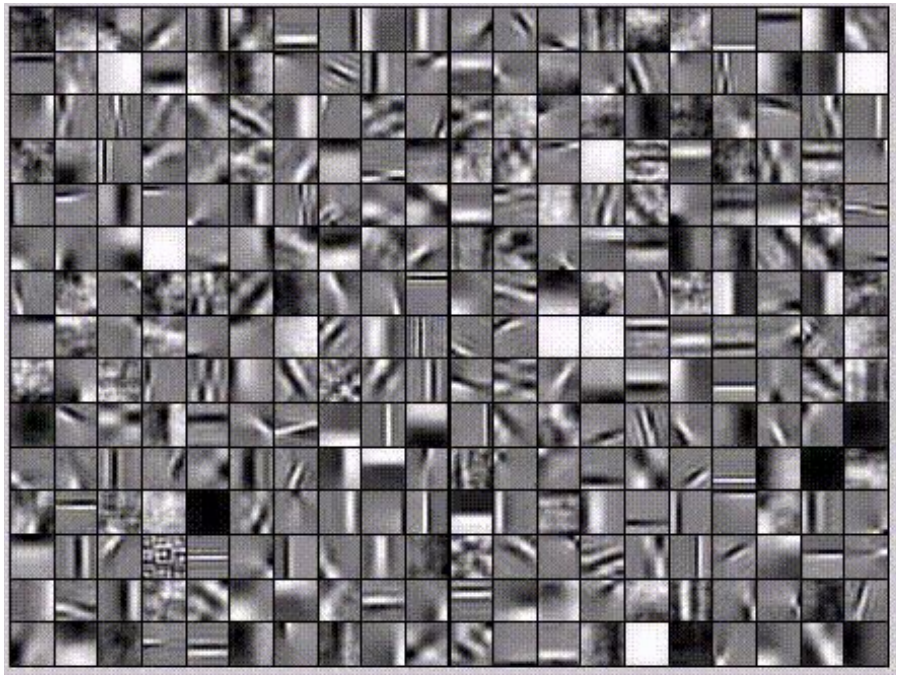


	KTH	Hollywood2	UCF	YouTube
Best published results	92.1%	50.9%	85.6%	71.2%
Our results	93.9%	53.3%	86.5%	75.8%

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 regularization: ISA

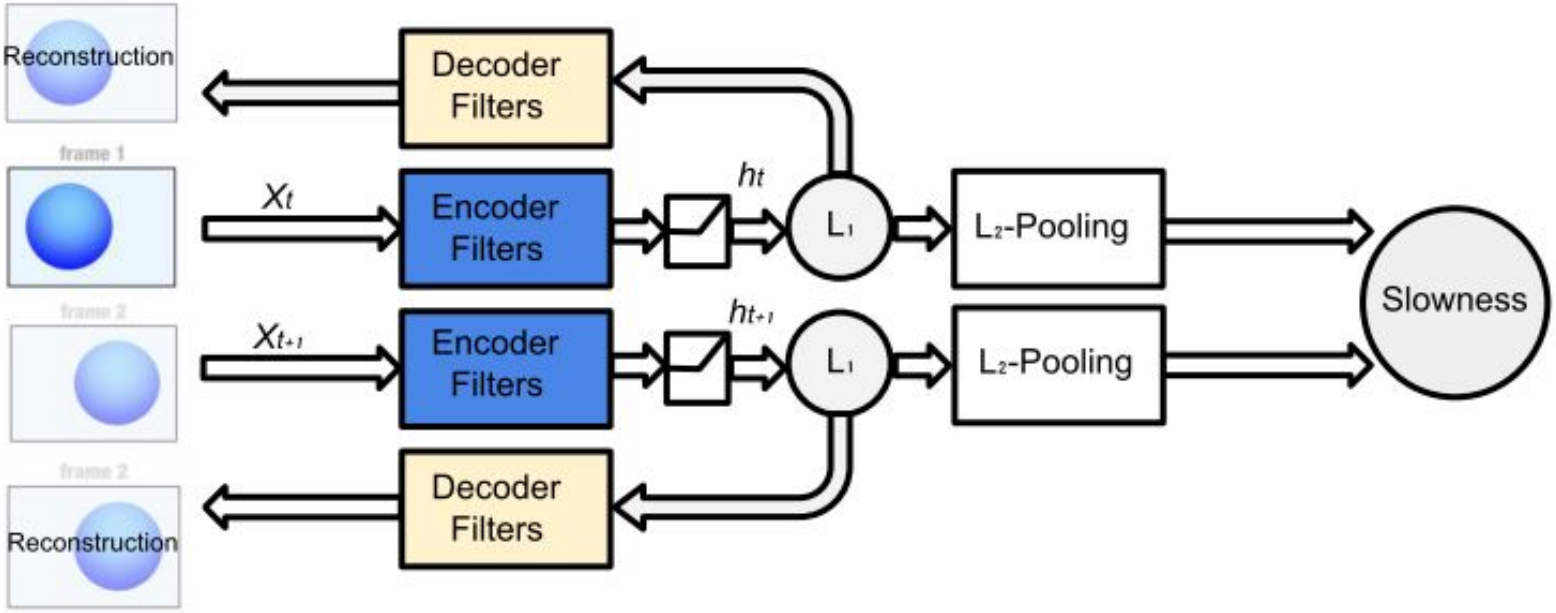
Feature visualizations.



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 regularization: Slowliness

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 regularization: Slowliness

Slow feature analysis

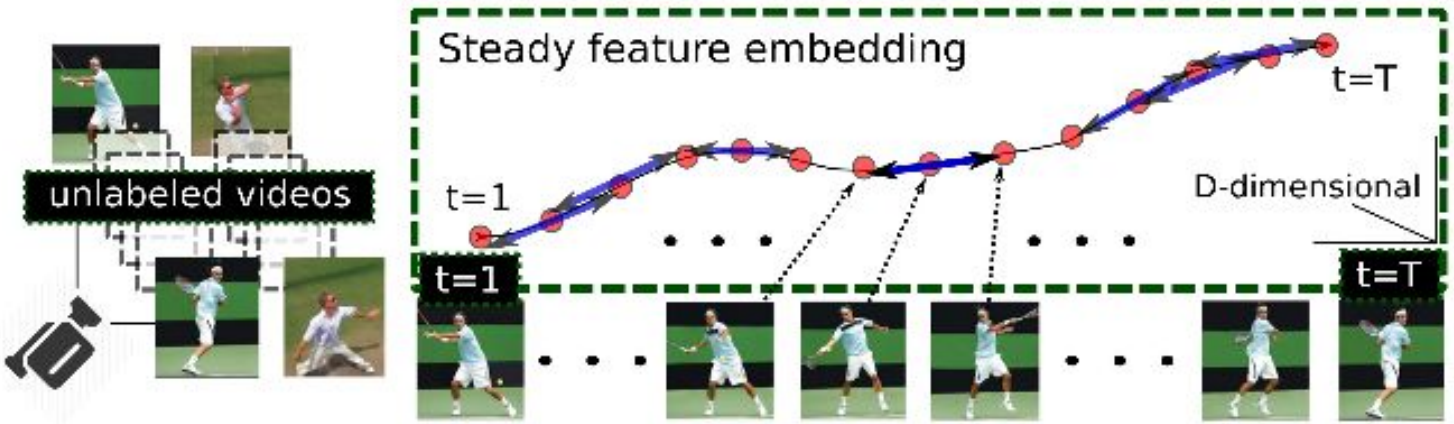
- Temporal coherence assumption: features should change slowly over time in video

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



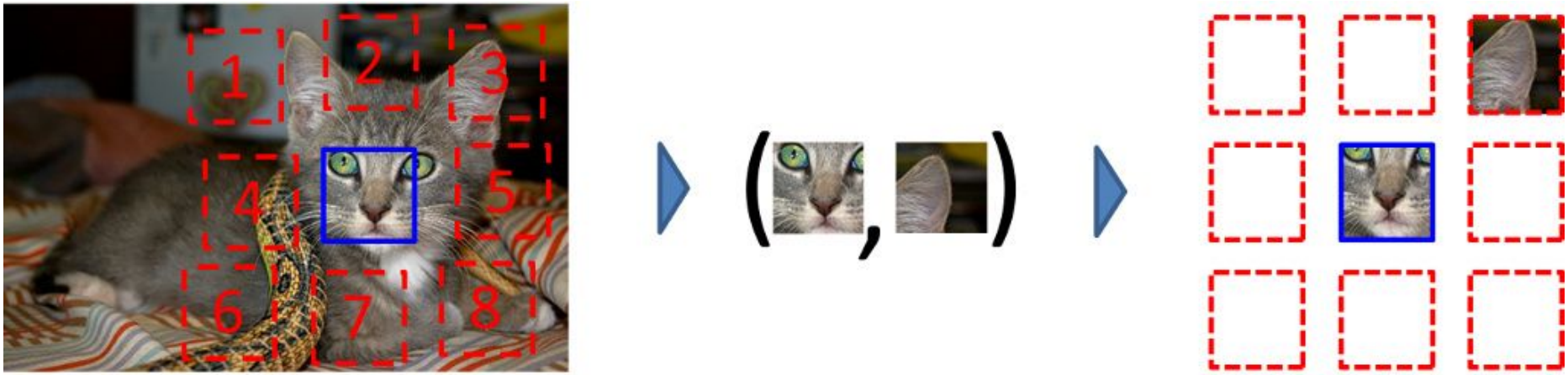
Jayaraman, Dinesh, and Kristen Grauman. "Slow and steady feature analysis: higher order temporal coherence in video." CVPR 2016. [video]

Outline

1. Unsupervised Learning
2. Self-supervised Learning
    a. Autoencoder
    b. Temporal regularisations
    c. Verification
    d. Predictive Learning
    e. Miscellaneous

Spatial verification

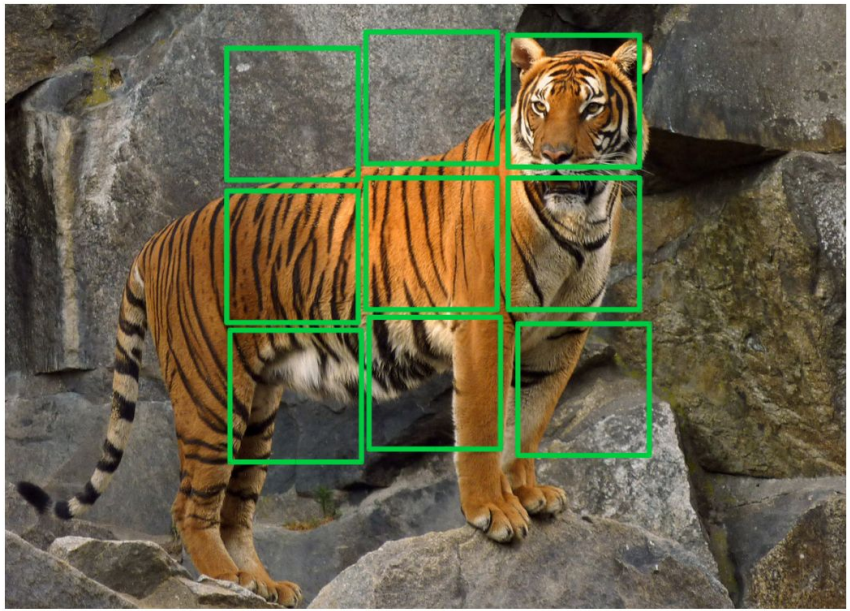
Predict the relative position between two image patches.

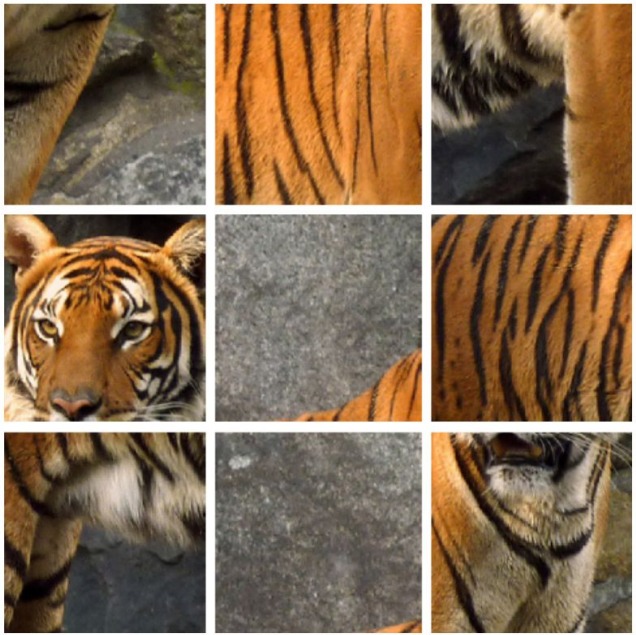


#RelativePosition Doersch, Carl, Abhinav Gupta, and Alexei A. Efros. "Unsupervised visual representation learning by context prediction." ICCV 2015.

Spatial verification

Train a neural network to solve a jigsaw puzzle.

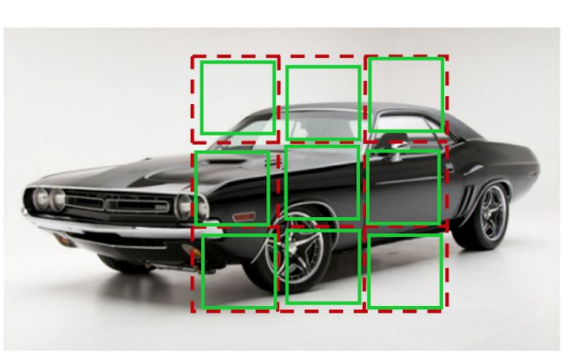




#Jigsaw Noroozi, M., & Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. ECCV 2016.

Spatial verification

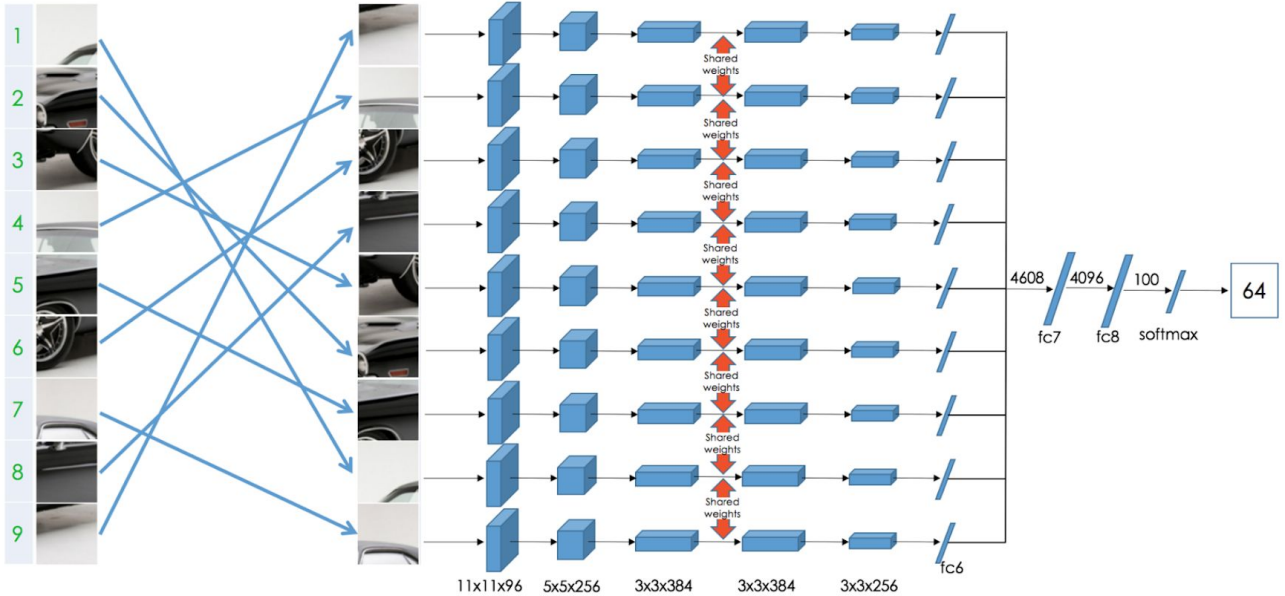
Train a neural network to solve a jigsaw puzzle.



Permutation Set

index	permutation
64	9,4,6,8,3,2,5,1,7

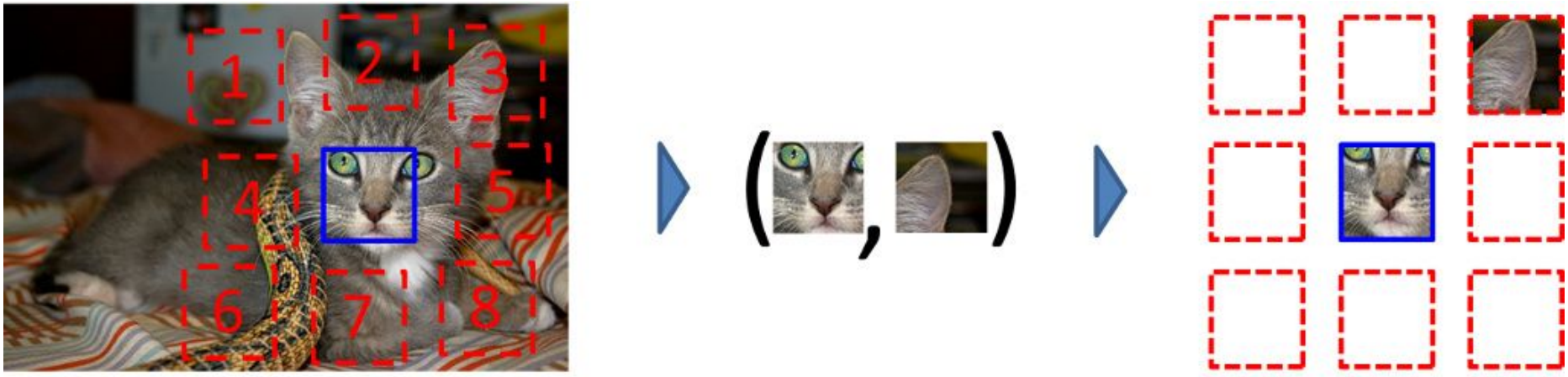
Reorder patches according to the selected permutation



#Jigsaw Noroozi, M., & Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. ECCV 2016.

Spatial verification

What video-specific surrogate tasks could you think about ?

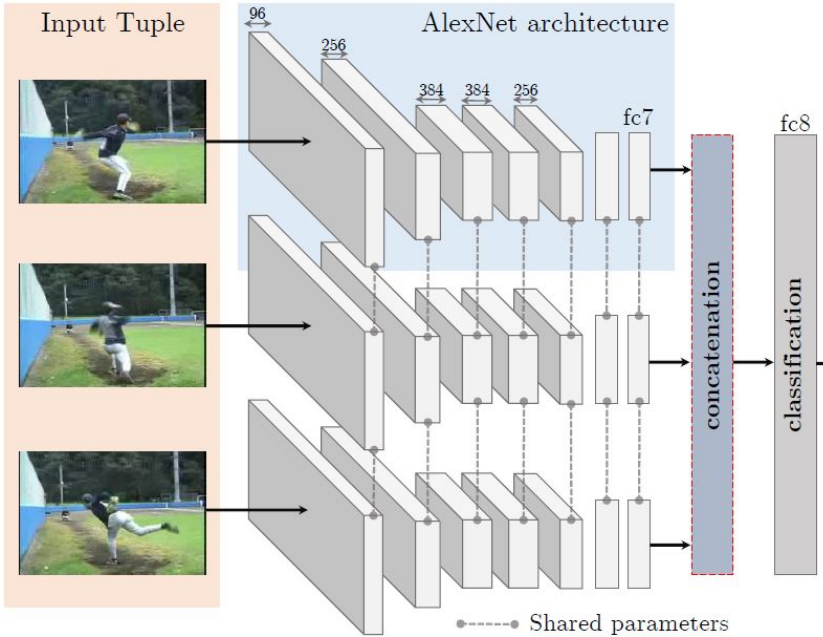


#RelativePosition Doersch, Carl, Abhinav Gupta, and Alexei A. Efros. "Unsupervised visual representation learning by context prediction." ICCV 2015.

Temporal coherence

Take temporal order as the supervisory signals for learning

Shuffled sequences



Binary classification

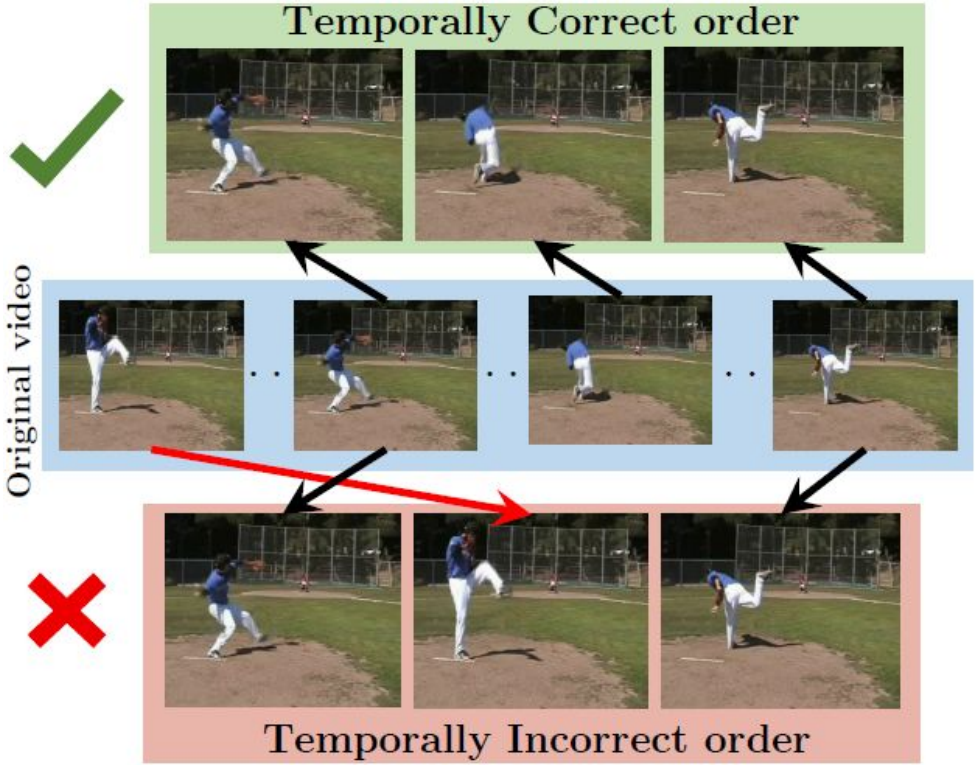
In order

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]

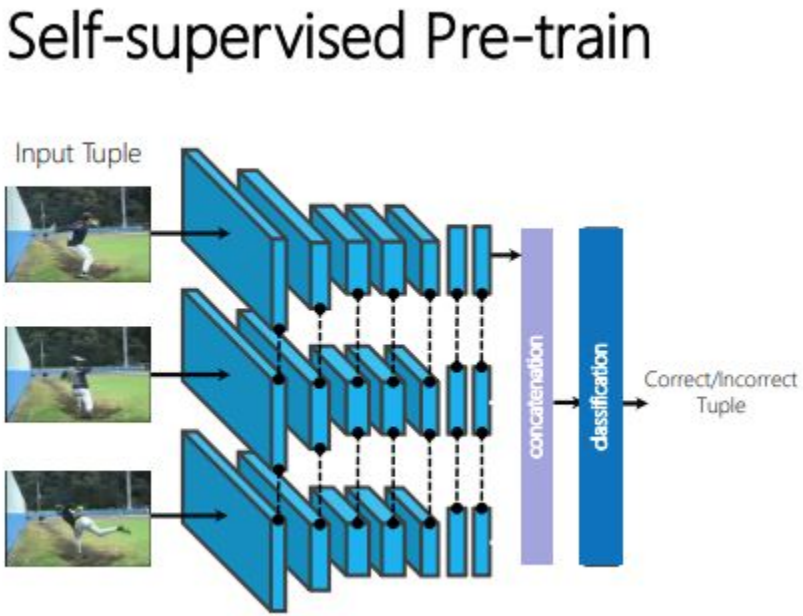
Temporal coherence

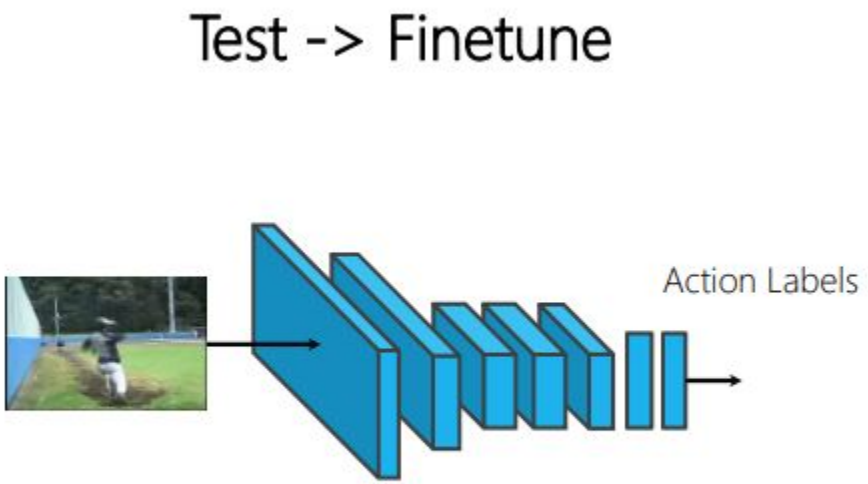
Temporal order of frames is exploited as the supervisory signal for learning.



(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 coherence

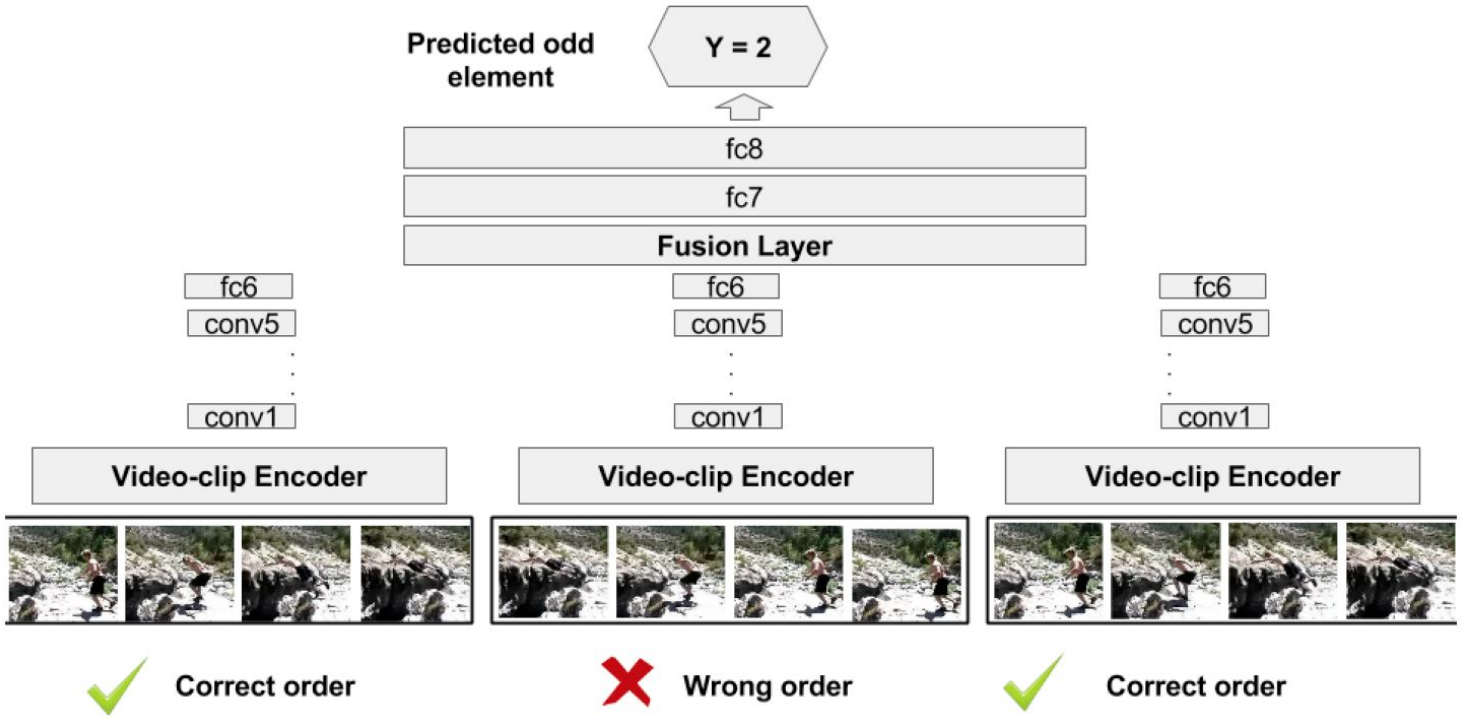




#Shuffle&Learn Misra, Ishan, C. Lawrence Zitnick, and Martial Hebert. "Shuffle and learn: unsupervised learning using temporal order verification." ECCV 2016. [code] (Slides by Xunyu Lin):

Temporal verification

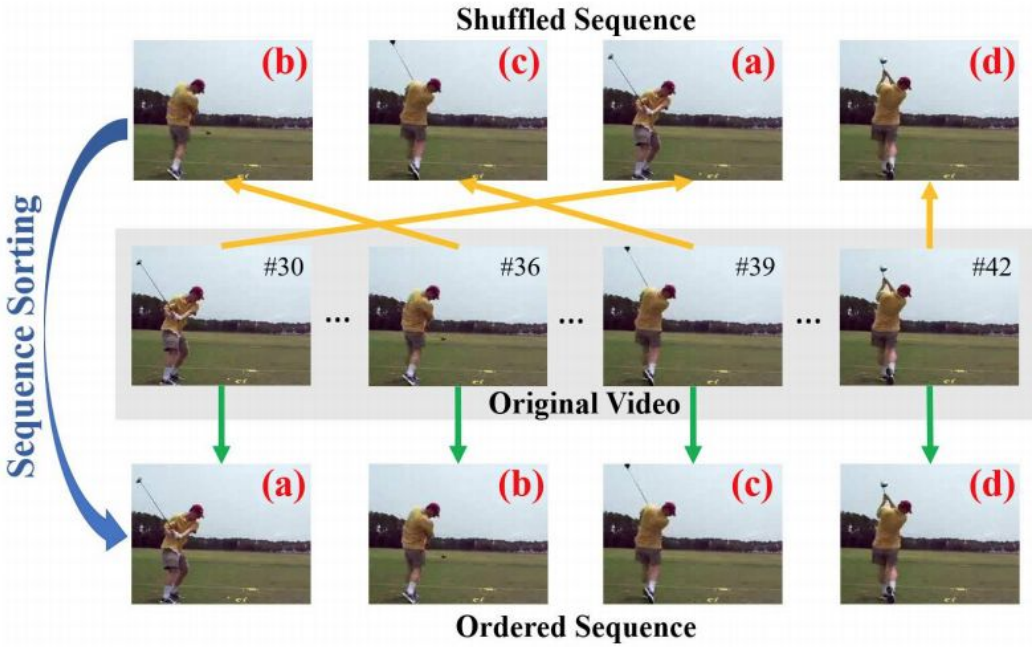
Train a network to detect which of the video sequences contains frames in the wrong order.



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

Temporal sorting

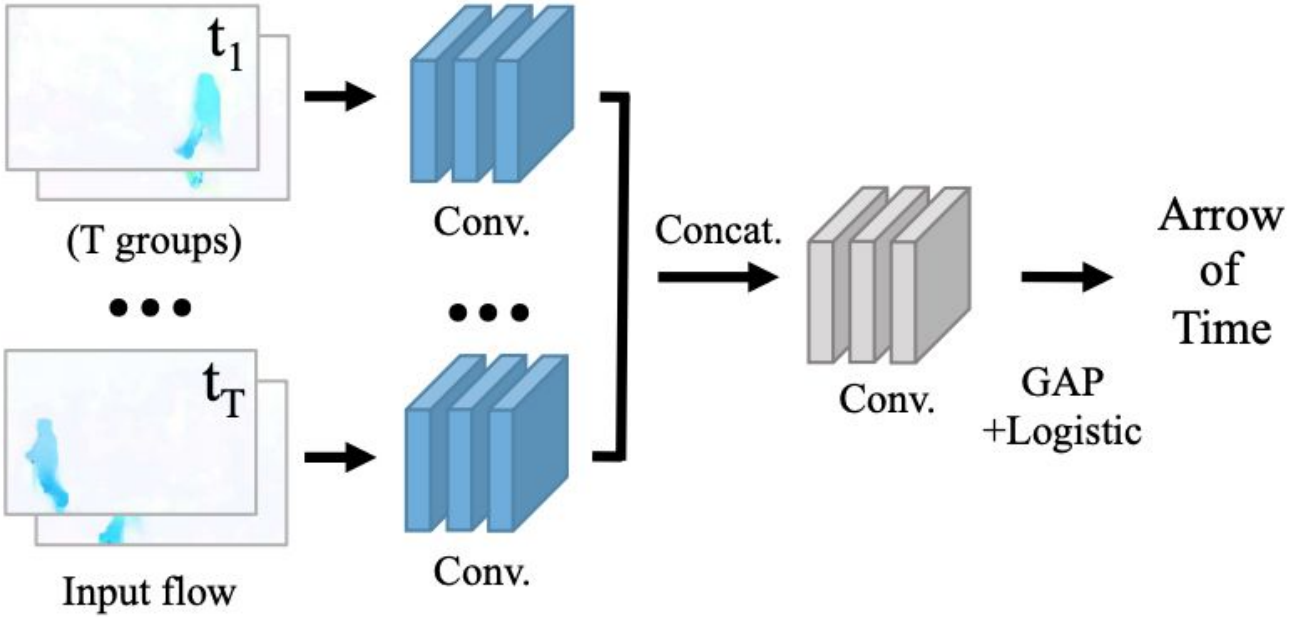
Sort the sequence of frames.



Lee, Hsin-Ying, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. "Unsupervised representation learning by sorting sequences." ICCV 2017.

Arrow of time

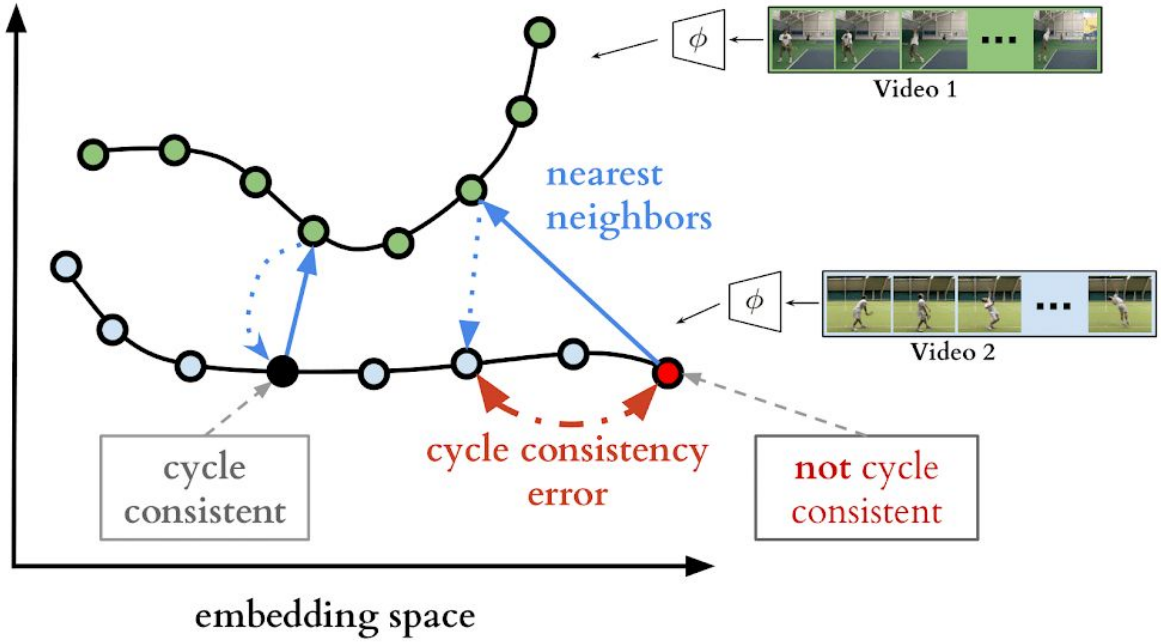
Predict whether the video moves forward or backward.



#T-CAM Wei, Donglai, Joseph J. Lim, Andrew Zisserman, and William T. Freeman. "Learning and using the arrow of time." CVPR 2018.

Temporal coherence

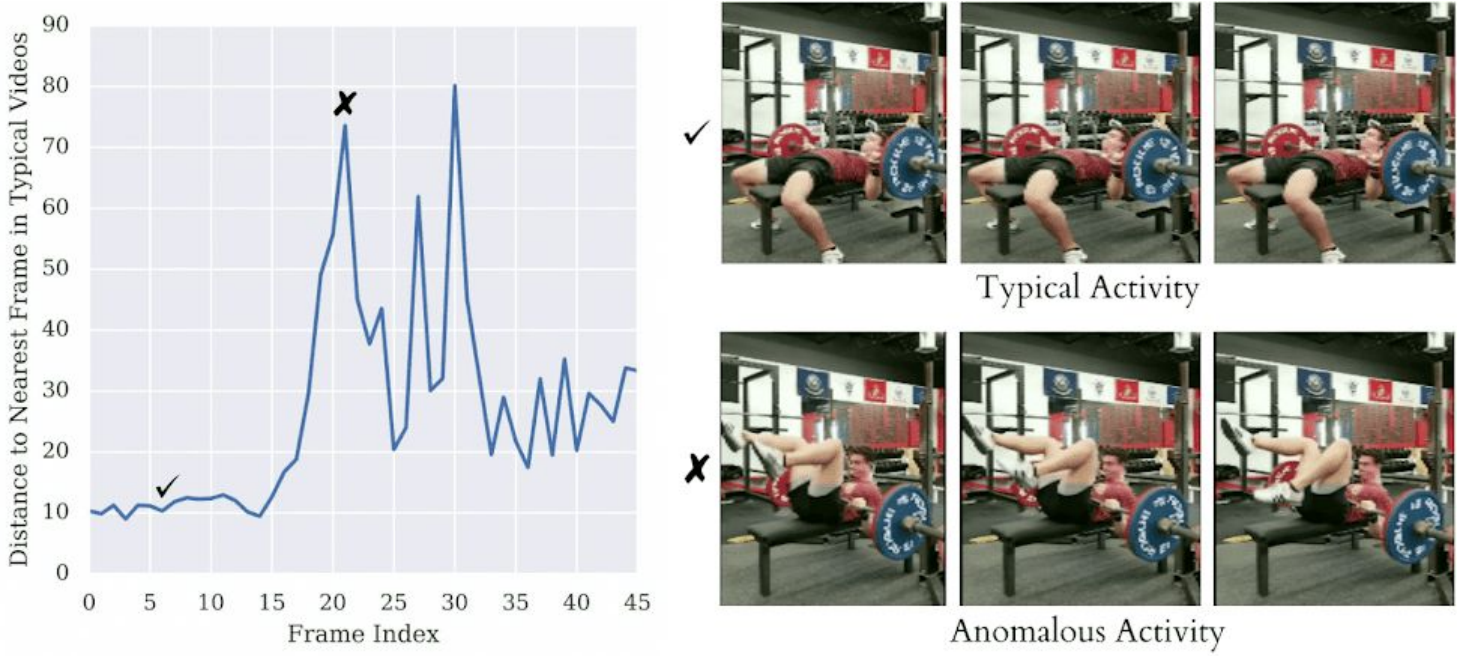
Training: Learn video frame embeddings that must satisfy a temporal alignment between videos depicting the same activity.



#TCC Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, Andrew Zisserman, "Temporal Cycle-Consistency Learning". CVPR 2019

Temporal coherence

Application: Anomaly detection.



#TCC Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, Andrew Zisserman, "Temporal Cycle-Consistency Learning". CVPR 2019

Temporal coherence

Application: Fine-grained retrieval

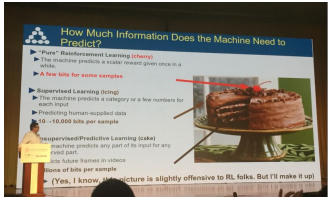
#TCC Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, Andrew Zisserman, "Temporal Cycle-Consistency Learning". CVPR 2019

Outline

1. Unsupervised Learning
2. Self-supervised Learning
    a. Autoencoder
    b. Temporal regularisations
    c. Temporal verification
    d. Predictive Learning
    e. Miscellaneous

Predictive 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

Next word prediction (language model)

Self-supervised learning

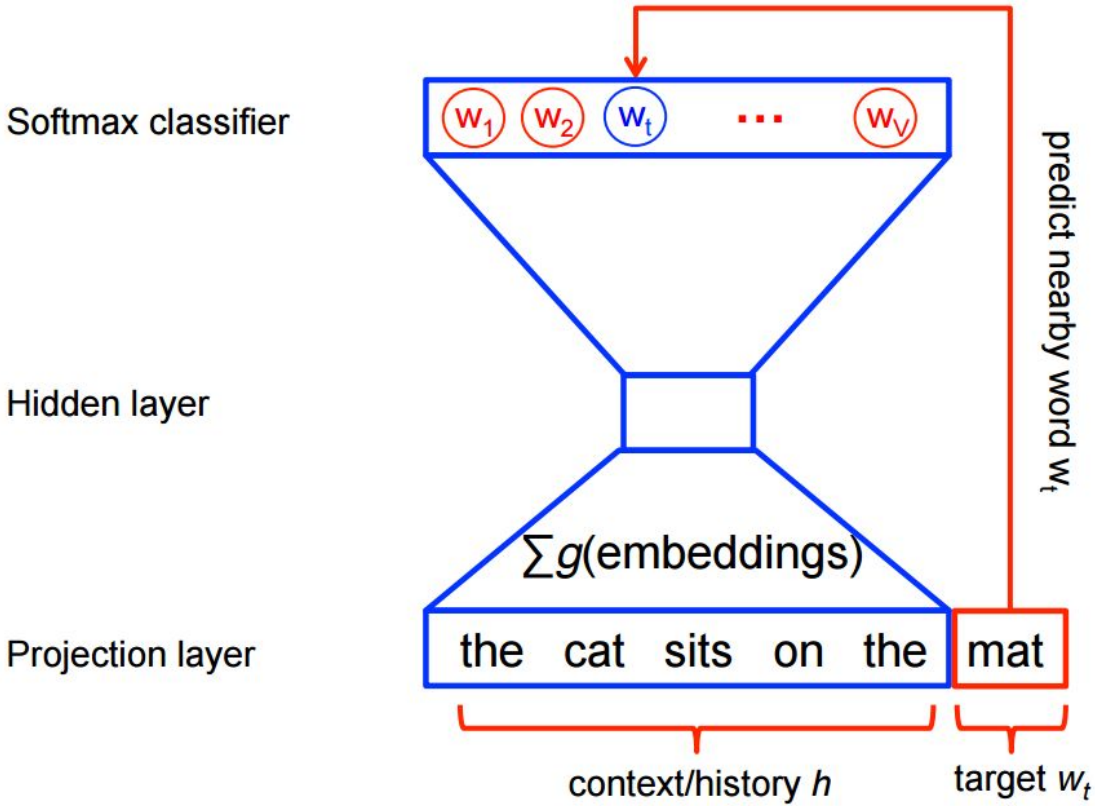
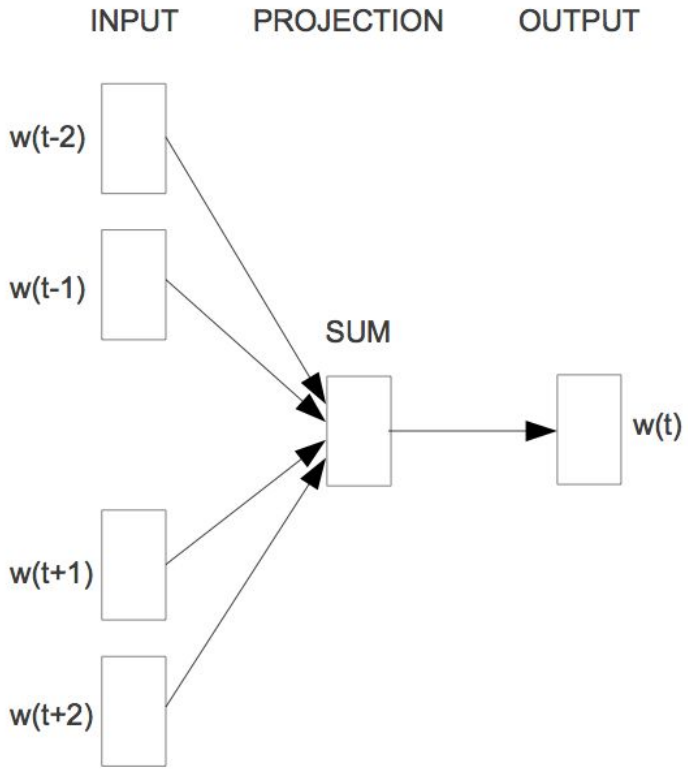


Figure:
TensorFlow tutorial

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. "A neural probabilistic language model." Journal of machine learning research 3, no. Feb (2003): 1137-1155.

Missing word prediction (language model)

Self-supervised learning



`the cat climbed a tree`

Given context:

`a, cat, the, tree`

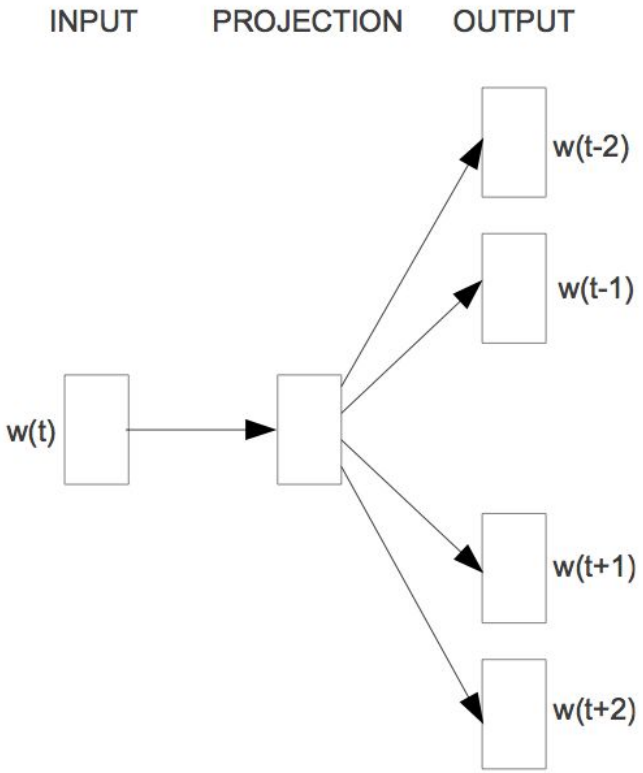
Estimate prob. of

`climbed`

#word2vec #continuousbow Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality." NIPS 2013

Context Words Prediction (language model)

Self-supervised learning



`the cat climbed a tree`

Given word:

`climbed`

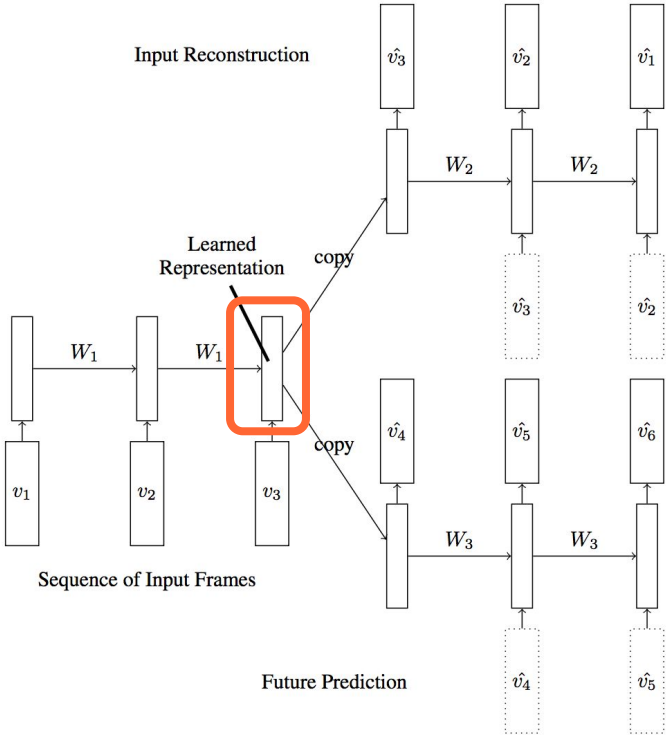
Estimate prob. of context words:

`a, cat, the, tree`

#word2vec #skipgram Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality." NIPS 2013

Frame Prediction

Learning video representations (features) by...

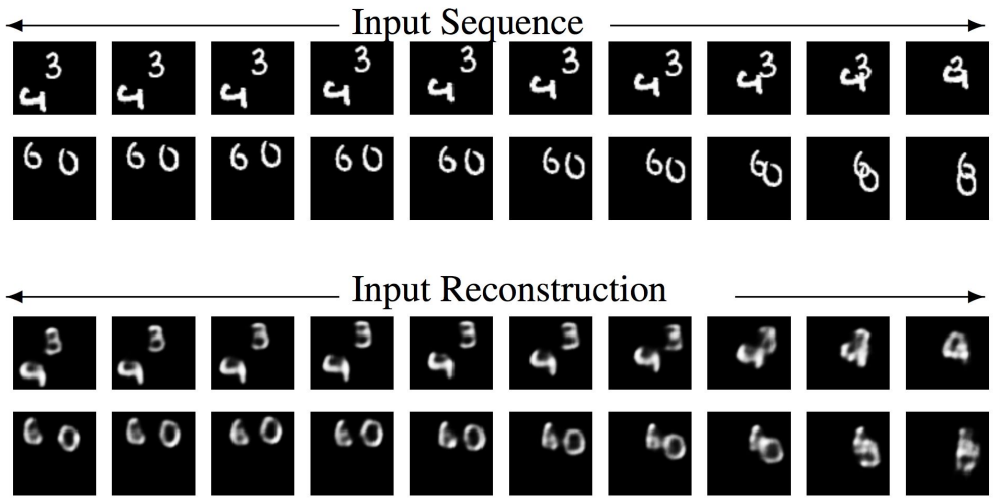


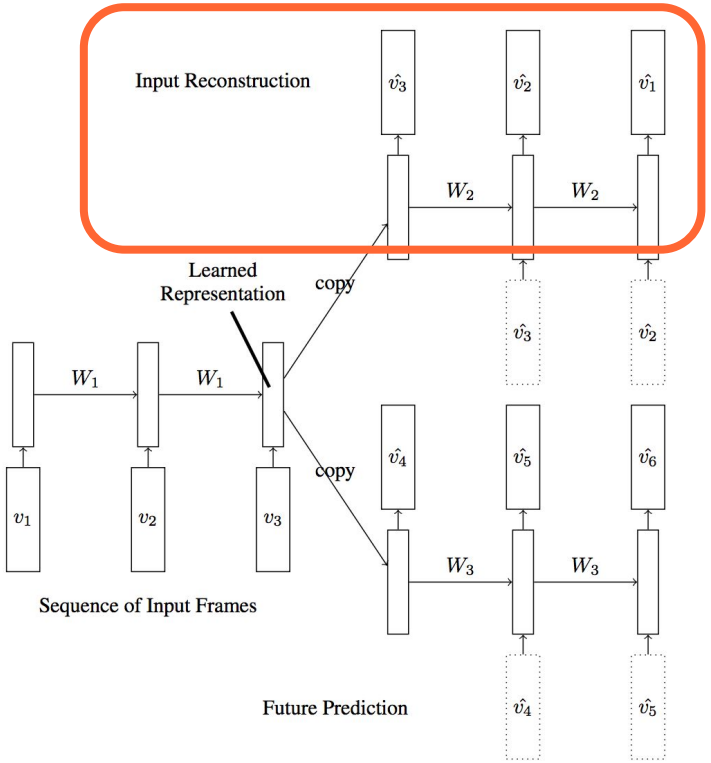
Srivastava, Nitish, Elman Mansimov, and Ruslan Salakhutdinov. "Unsupervised Learning of Video Representations using LSTMs." In ICML 2015. [Github]

Frame Prediction

Learning video representations (features) by...

(1) frame reconstruction (AE):



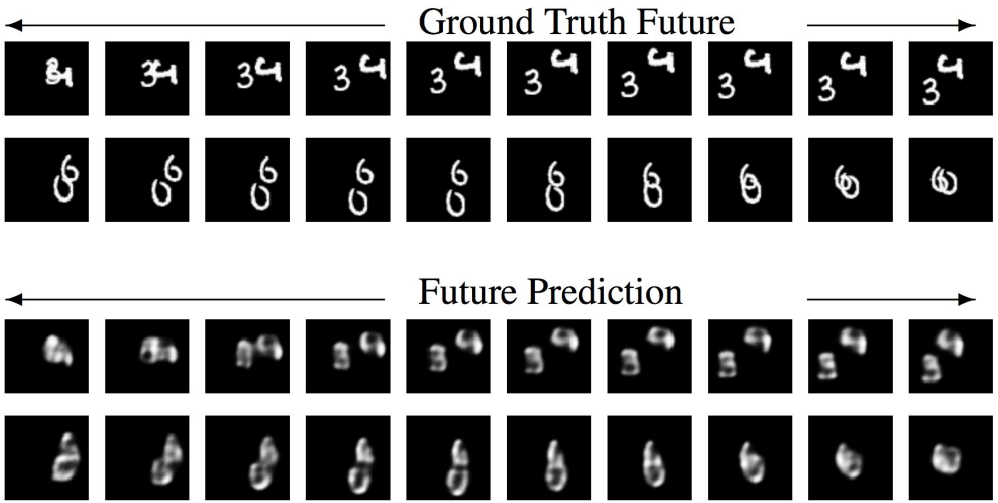


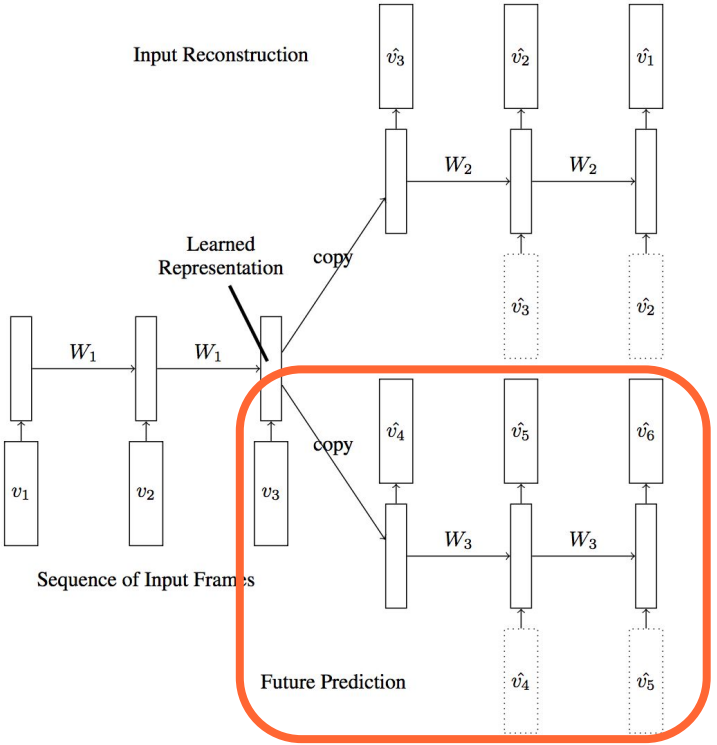
Srivastava, Nitish, Elman Mansimov, and Ruslan Salakhutdinov. "Unsupervised Learning of Video Representations using LSTMs." ICML 2015. [Github]

Frame Prediction

Learning video representations (features) by...

(2) frame prediction





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

Frame Prediction

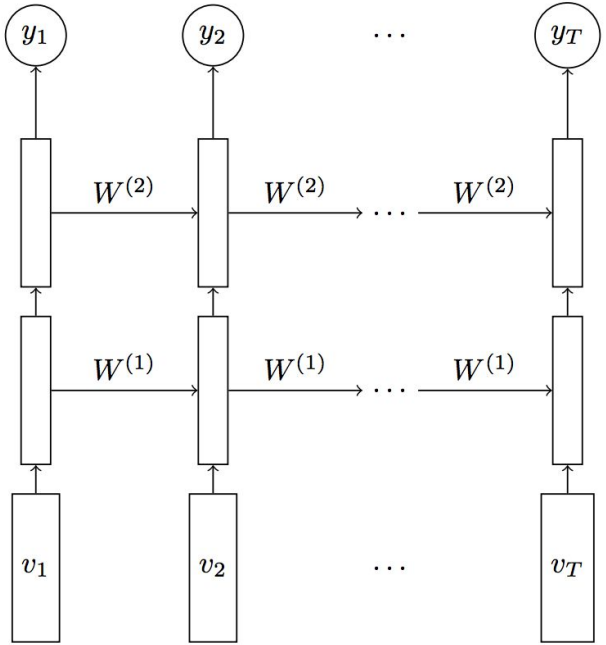


Figure 6. LSTM Classifier.

Unsupervised learned features (lots of data) are fine-tuned for activity recognition (small 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**

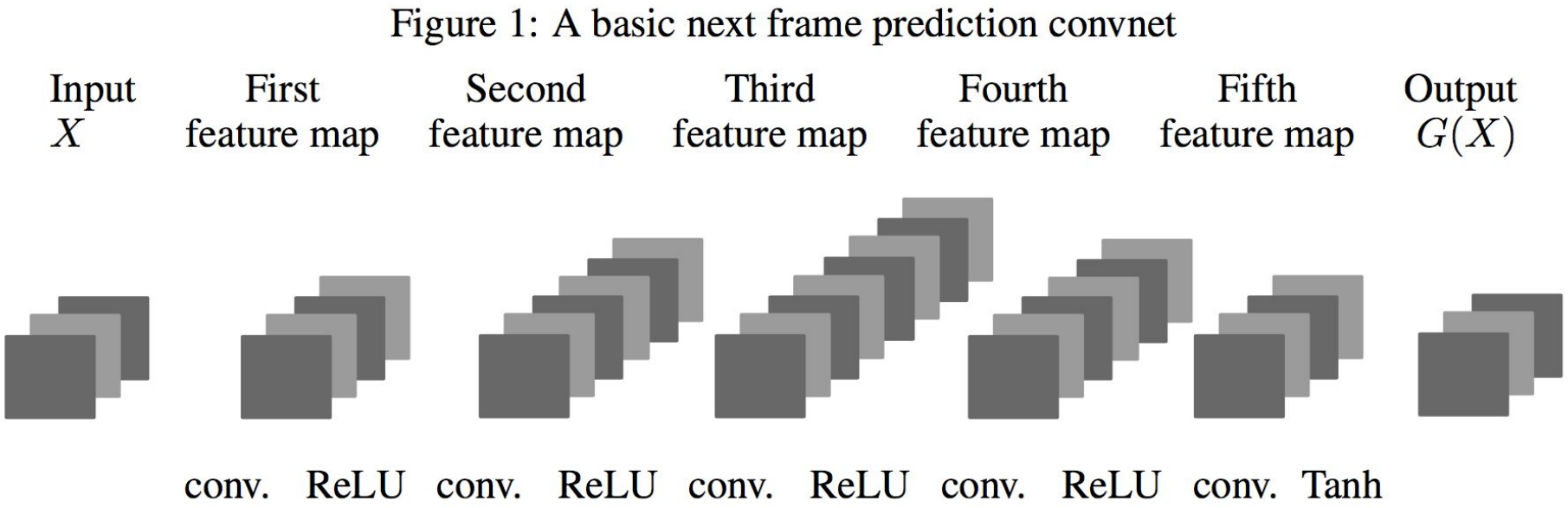
Table 1. Summary of Results on Action Recognition.

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

Frame Prediction

Video frame prediction with a ConvNet.

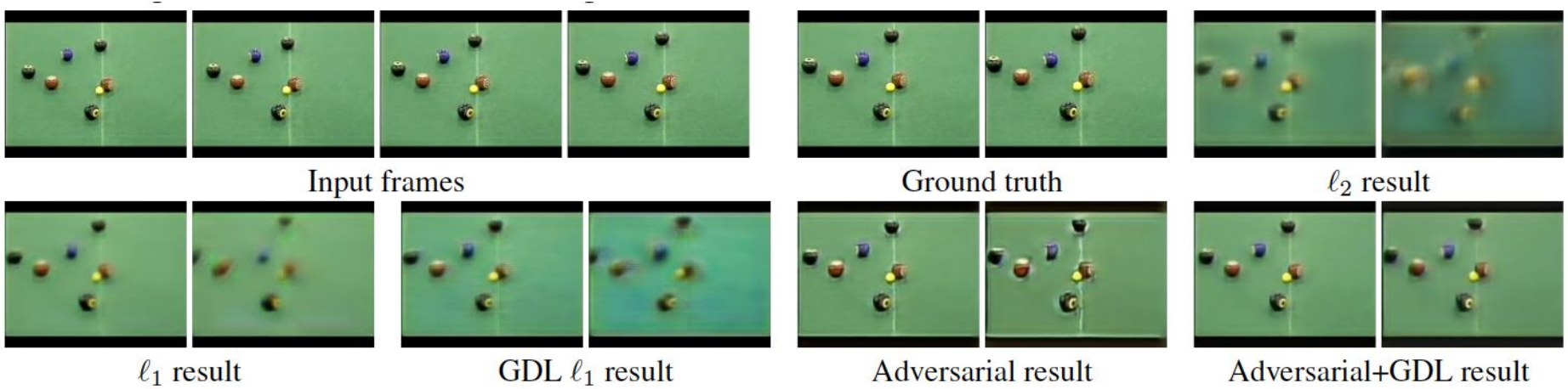
Figure 1: A basic next frame prediction convnet



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 (L1 loss) 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 + Disentangled features

The model learns to disentangle (“separate”) the visual features that correspond to the:

Object Content (class)

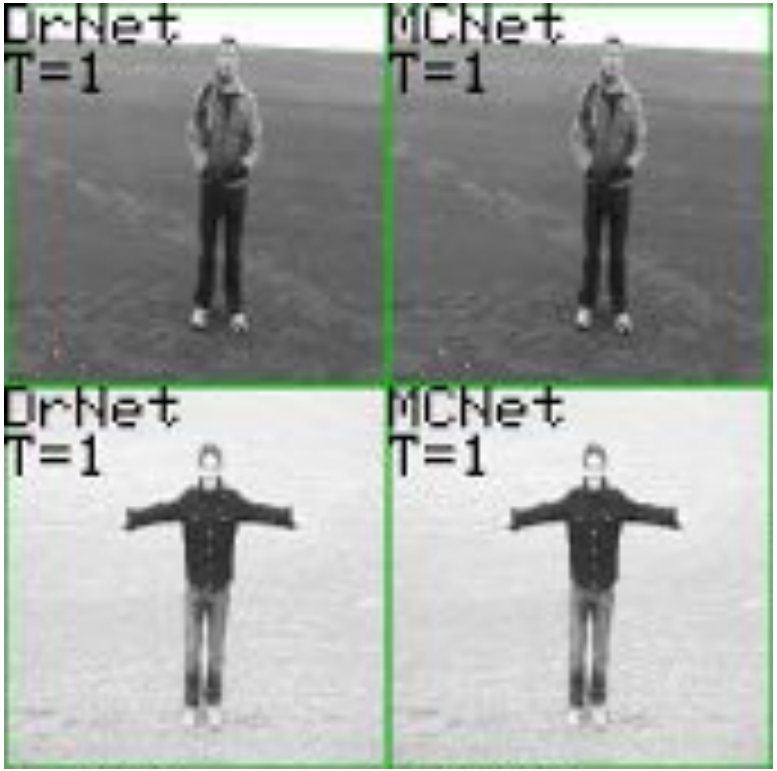
Object Pose (wrt the camera)

#DrNet Denton, Emily L. "Unsupervised learning of disentangled representations from video." NIPS 2017.

Frame Prediction + Disentangled features

100 step video generation on KTH dataset where:

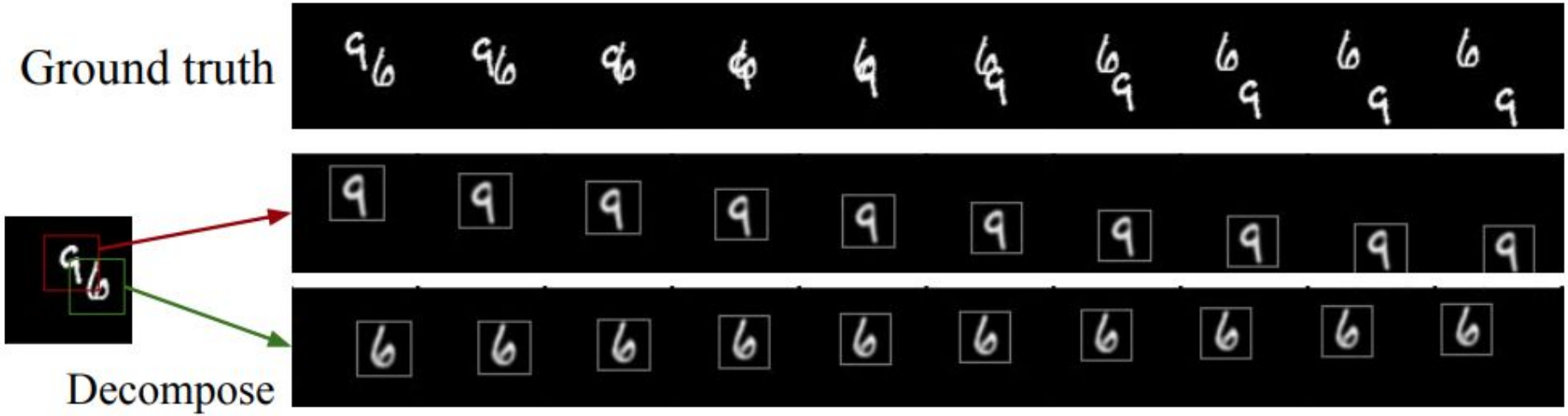
- green frames indicate reconstructions
- red frames indicate frame predictions.



#DrNet Denton, Emily L. "Unsupervised learning of disentangled representations from video." NIPS 2017.
#MCNet R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee. Decomposing motion and content for natural video sequence prediction. In *ICLR*, 2017

Frame Prediction + Disentangled features

Decompositional Disentangled Predictive Auto-Encoder (DDPAE):



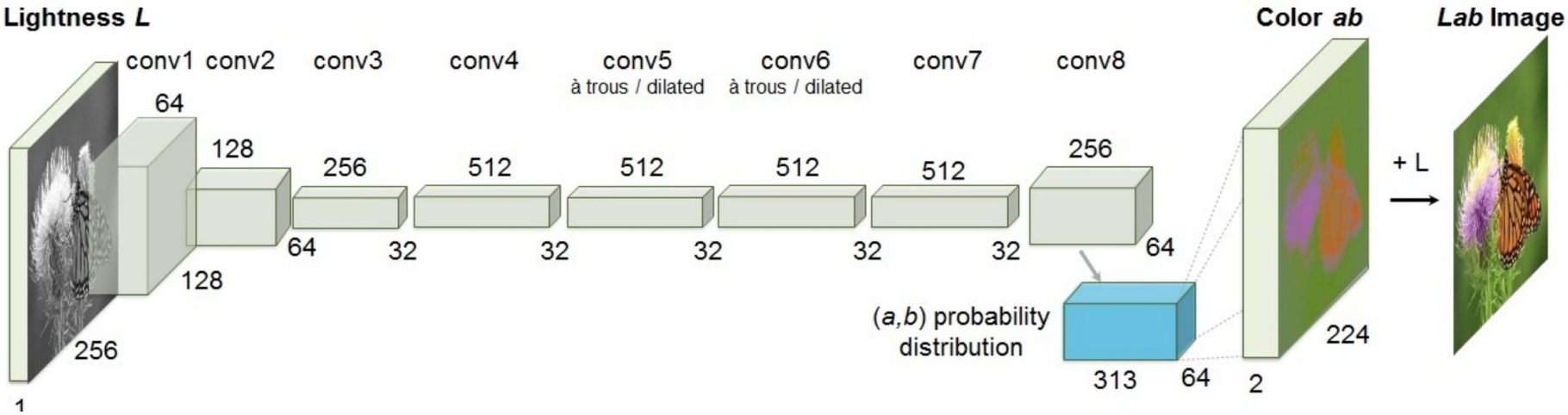
#DDPAE Hsieh, Jun-Ting, Bingbin Liu, De-An Huang, Li F. Fei-Fei, and Juan Carlos Niebles. "Learning to decompose and disentangle representations for video prediction." NeurIPS 2018.

Outline

1. Unsupervised Learning
2. Self-supervised Learning
    a. Autoencoder
    b. Temporal regularisations
    c. Temporal verification
    d. Frame Prediction
    e. Miscellaneous

Colorization (still image)

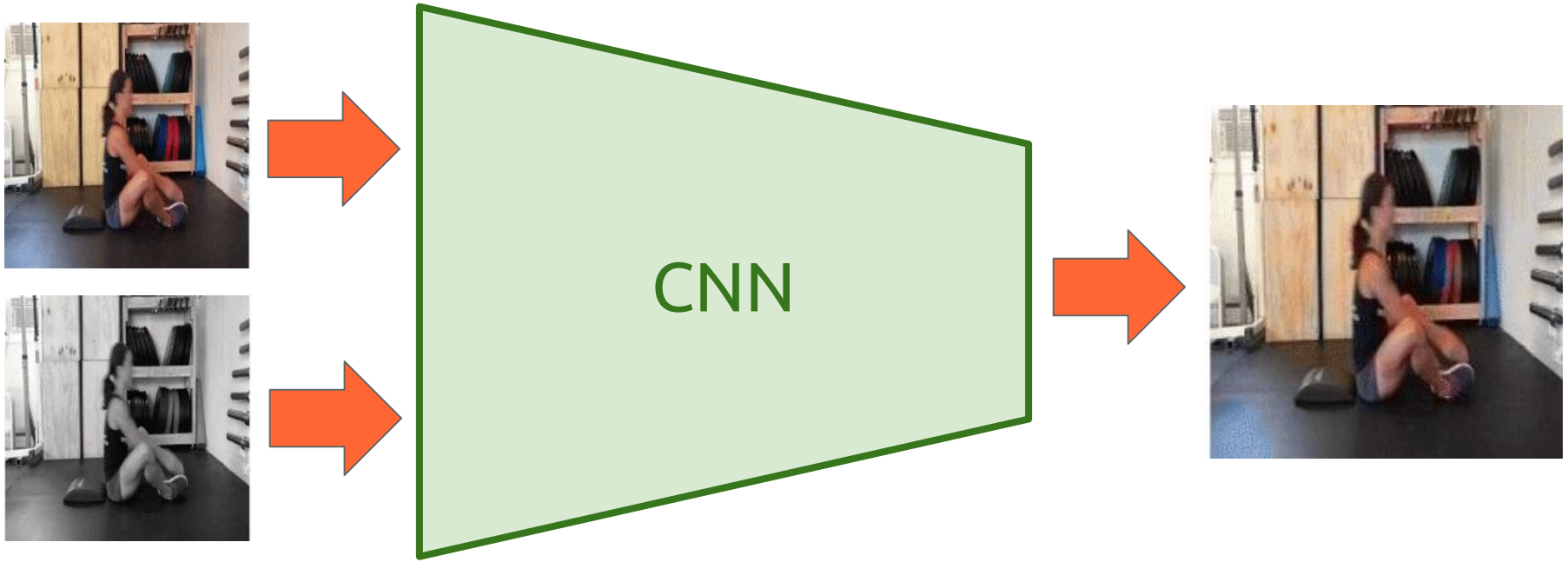
A NN is trained to colorize a gray scale image..



Zhang, R., Isola, P., & Efros, A. A. Colorful image colorization. ECCV 2016.

Colorization (video)

A NN is trained to colorize a video frame, given the color of the first frame of the video sequence.



Vondrick, Carl, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. "Tracking emerges by colorizing videos." ECCV 2018. [blog]

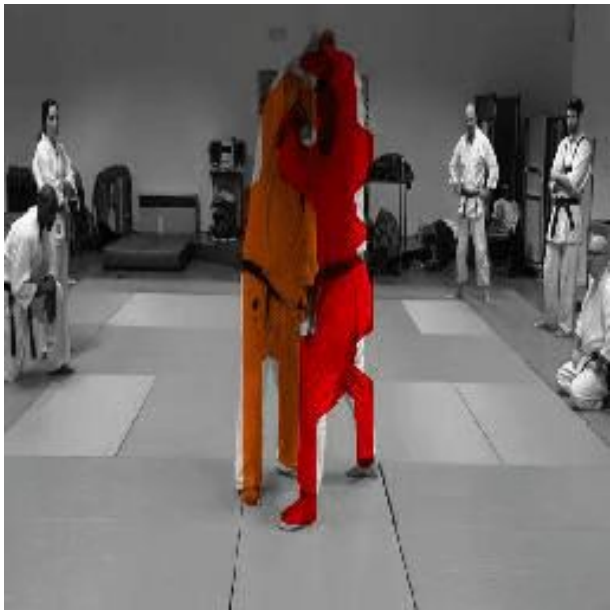
Colorization (video)

Training: A NN is trained to colorize a video frame, given the color of the first frame of the video sequence.

Vondrick, Carl, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. "Tracking emerges by colorizing videos." ECCV 2018. [blog]

Colorization (video)

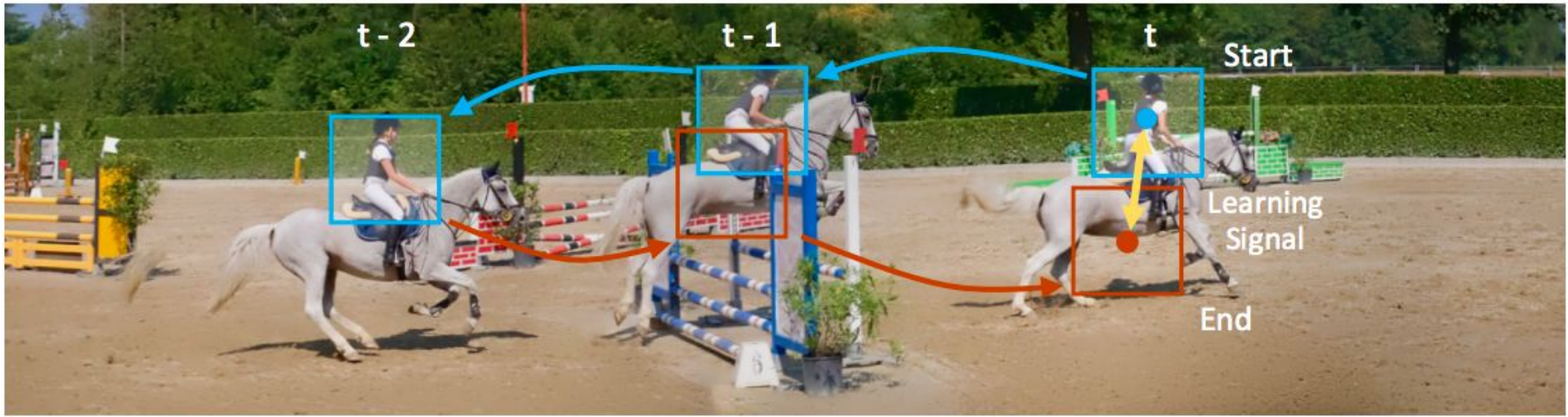
Inference: Video Object tracking with the learned pixel embeddings.



Vondrick, Carl, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. "Tracking emerges by colorizing videos." ECCV 2018. [blog]

Cycle-consistency

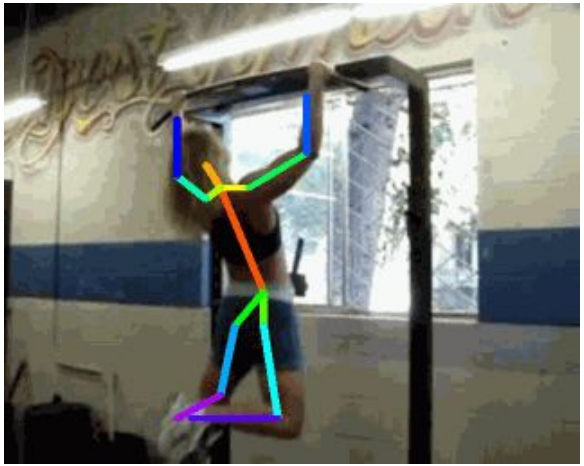
Training: Tracking backward and then forward to learn from the divergence of the predictions.



#TimeCycle Xiaolong Wang, Allan Jabri, Alexei A. Efros, "Learning Correspondence from the Cycle-Consistency of Time". CVPR 2019. [code] [Slides by Carles Ventura]

Cycle-consistency

Inference: Video Object tracking with the learned pixel embeddings.

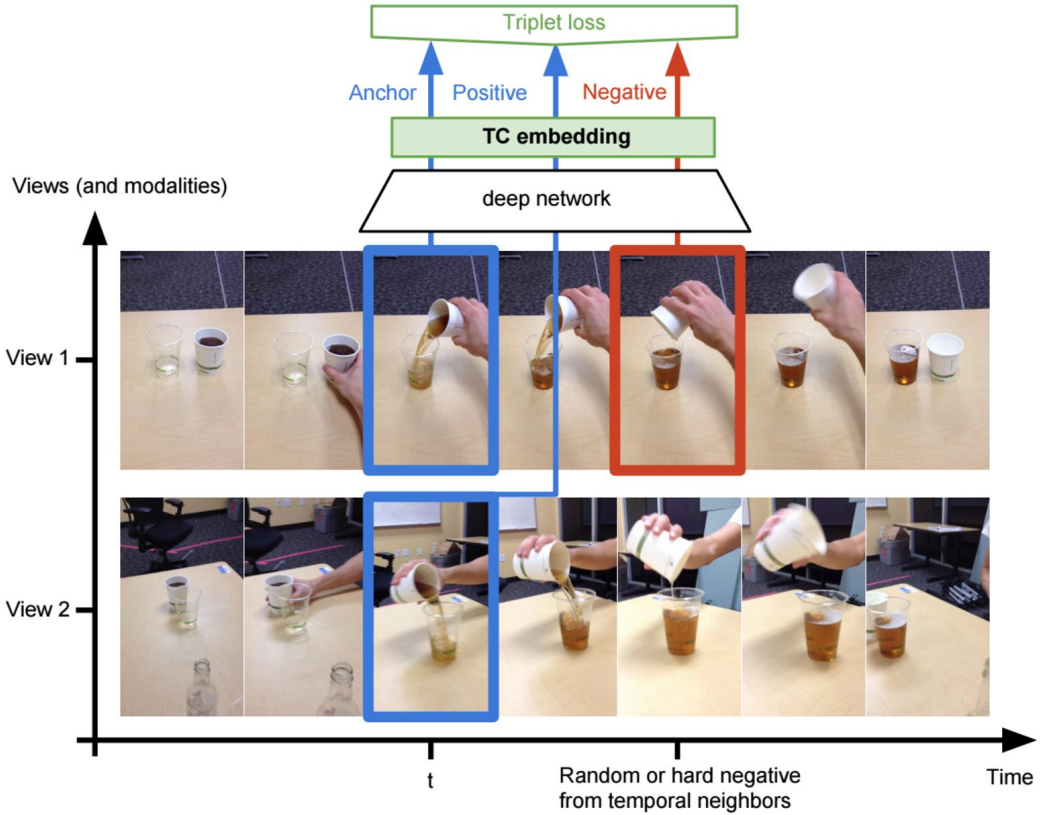


#TimeCycle Xiaolong Wang, Allan Jabri, Alexei A. Efros, "Learning Correspondence from the Cycle-Consistency of Time". CVPR 2019. [code] [Slides by Carles Ventura]

Multiview + Time

Training: Learn features that are:

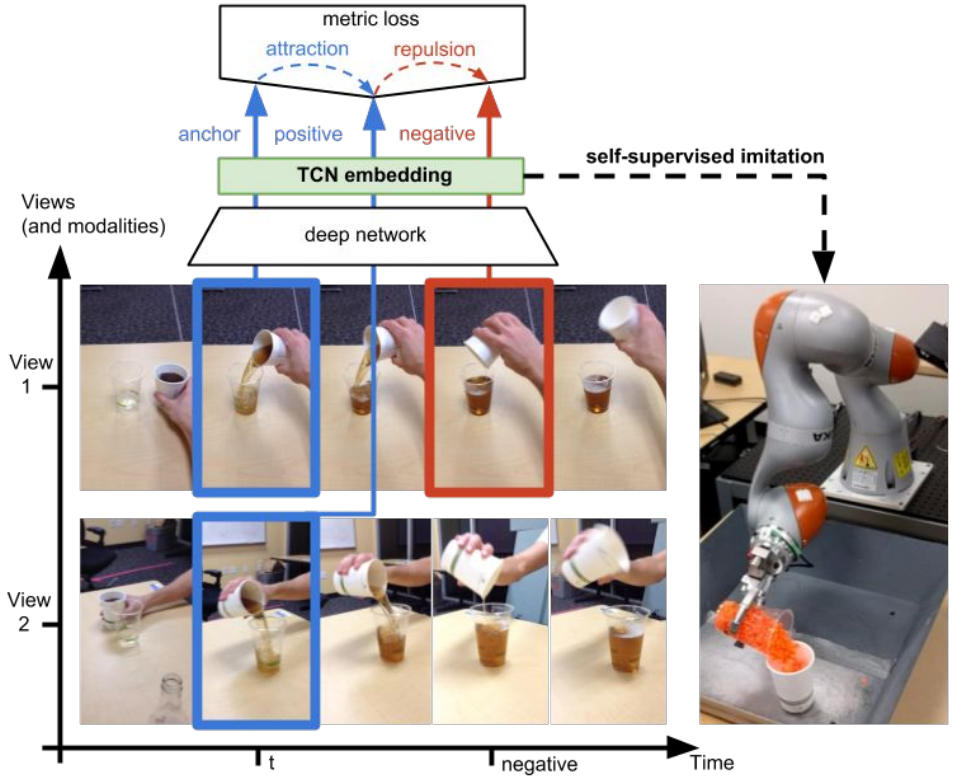
- Invariant to viewpoint.
- Very sensitive to temporal ordering



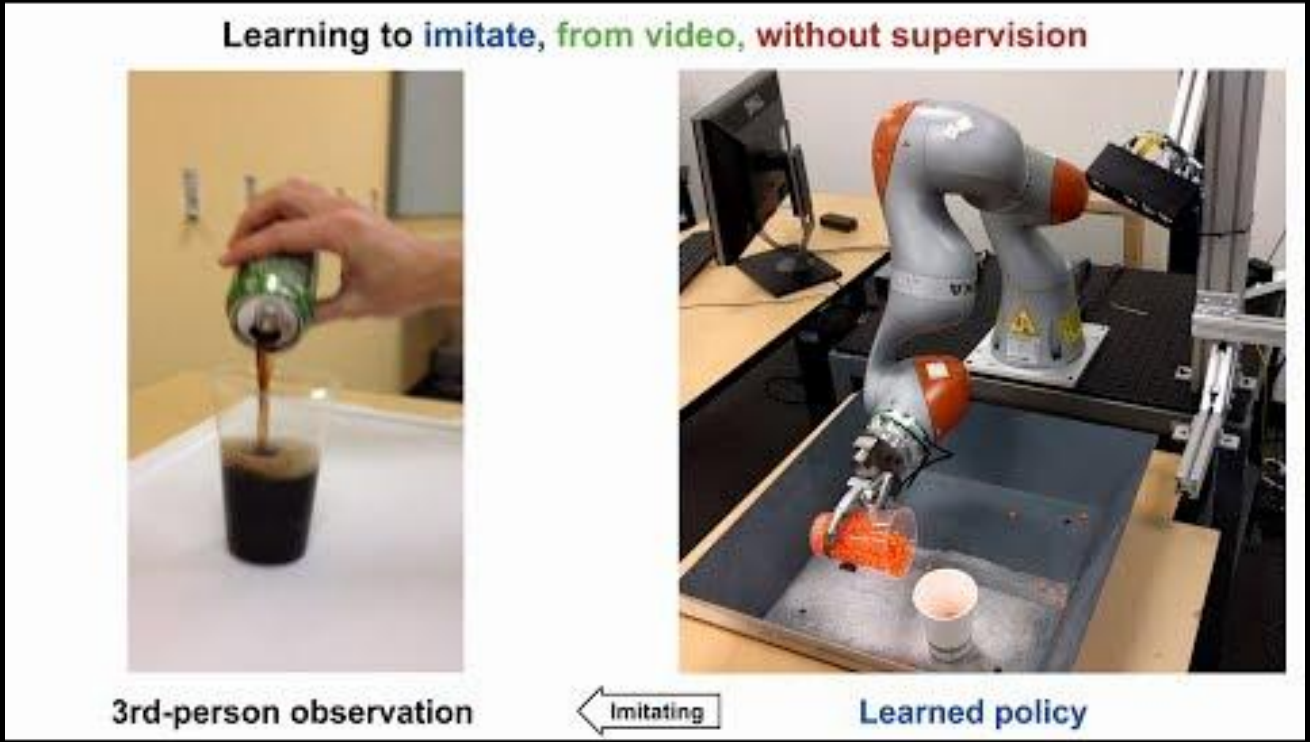
Sermanet, Pierre, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. "Time-contrastive networks: Self-supervised learning from video." ICRA 2018.

Multiview + Time

Inference: Imitation learning for a robotic arm.



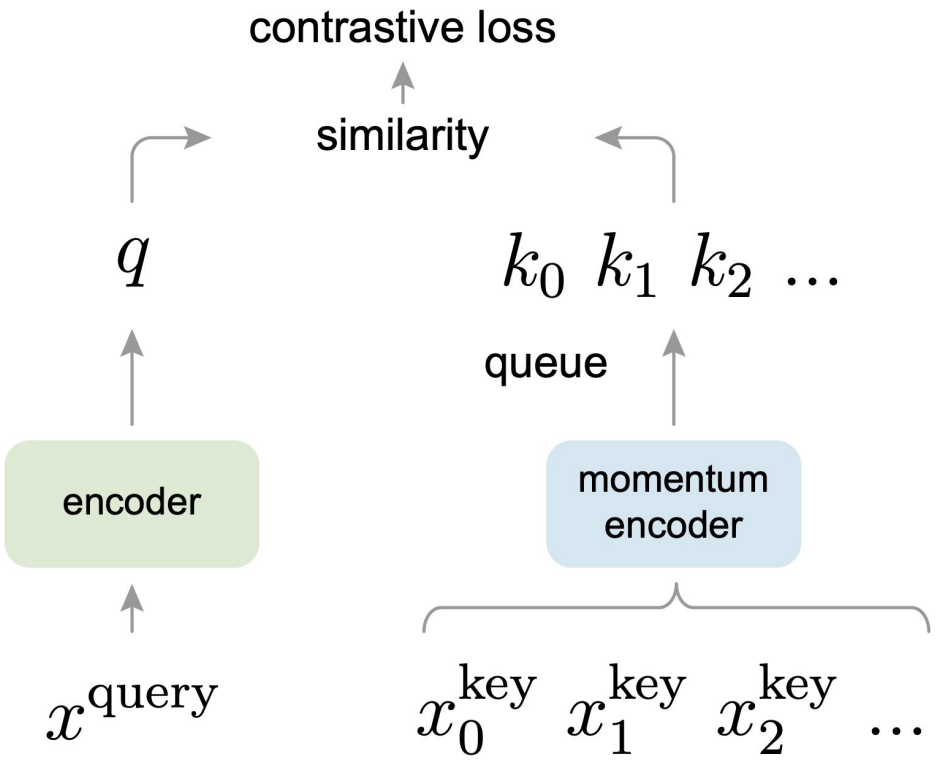
Sermanet, Pierre, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. "Time-contrastive networks: Self-supervised learning from video." ICRA 2018.



Sermanet, Pierre, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. "Time-contrastive networks: Self-supervised learning from video." ICRA 2018.

Momentum Contrast (MoCo)

"This suggests that the gap between unsupervised and supervised representation learning has been largely closed in many vision tasks.!



He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. . Momentum Contrast for Unsupervised Visual Representation Learning. arXiv preprint arXiv:1911.05722. (2019)

Outline

1. Unsupervised Learning
2. Self-supervised Learning
    a. Autoencoder
    b. Temporal regularisations
    c. Temporal verification
    d. Frame Prediction
    e. Miscellaneous

Bonus track

Self-supervised Audiovisual Learning [slides] [video]

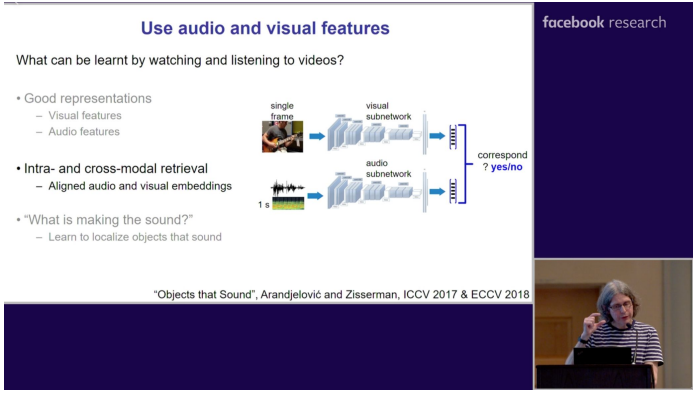
Suggested reading

Alex Graves, Kelly Clancy, “Unsupervised learning: The curious pupil”. Deepmind blog 2019.

Suggested talks

Demis Hassabis (Deepmind)

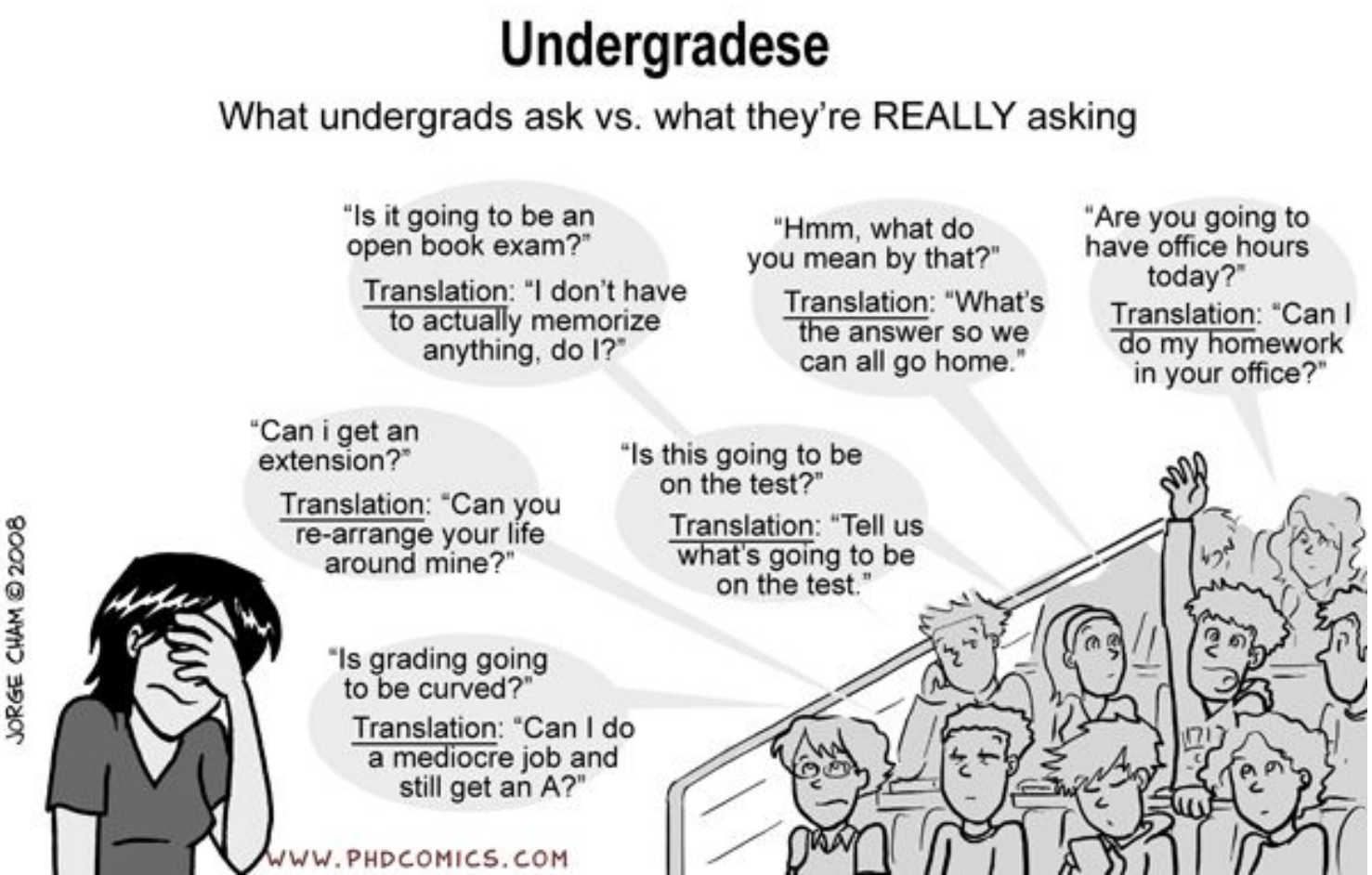
Aloysha Efros (UC Berkeley)



Andrew Zisserman (Oxford/Deepmind)

Yann LeCun (Facebook AI)

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