Master in Computer Vision Barcelona

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

Xavier Giro-i-Nieto

@DocXavi
xavier.giro@upc.edu

Associate Professor
Universitat Politècnica de Catalunya

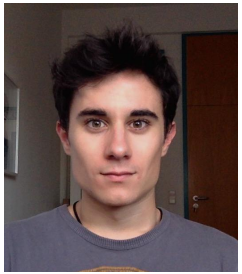
Module 6 - Day 9 - Lecture 4

Self-supervised Learning

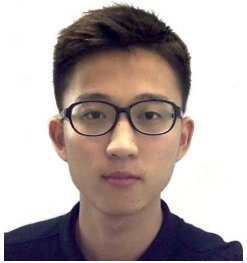
5th April 2022

Acknowledgements

Oscar Mañas



Víctor Campos



Junting Pan

Xunyu Lin

Carlos Arenas

Sebastian Palacio

Video lecture

Yann LeCun
@ylecun

Self-supervised learning, obviously.

Tradueix el tuit

> **Muzaffer Kal WFH attending COVID-19** @DSPonFPGA · 19 de març
> "supervised machine learning doesn't live up to the hype" @ylecun where do we go from here?
>
> medium.com/starsky-roboti...

7:10 a. m. · 20 de març de 2020 · Twitter for Android

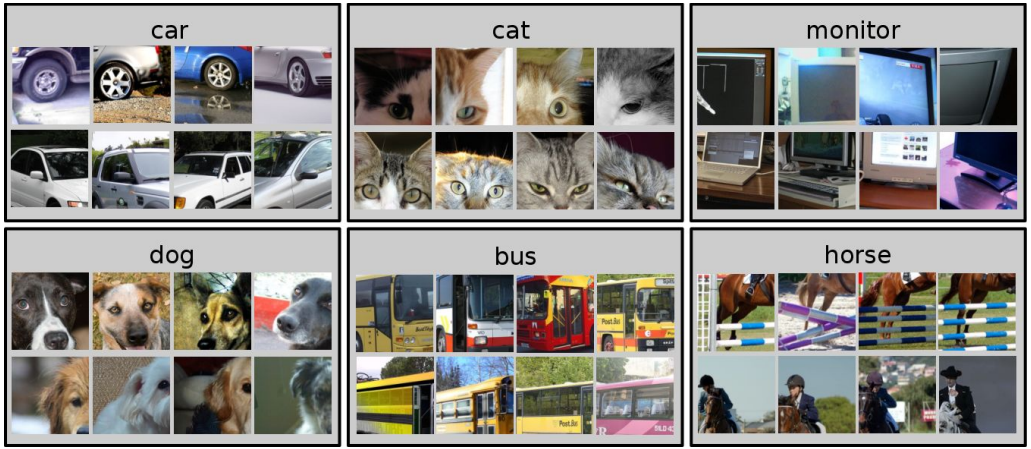
Outline

1. **Transfer Learning**
2. Representation Learning
3. Unsupervised Learning
4. Self-supervised Learning
5. Predictive methods
6. Contrastive methods

Study case: PASCAL VOC 2007 dataset

- Standard classification benchmark, 20 classes, ~10K images, 50% train, 50% test
- Deep networks can have many parameters (e.g. 60M in Alexnet)
- Direct training (from scratch) using only 5K training images can be problematic. Model overfits.

How can we use deep networks in this setting ?



Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. IJCV 2015.

Recap: Transfer Learning

In many cases, annotated data is not enough for training deep neural networks. The model will just overfit to the training data and do not generalize well.

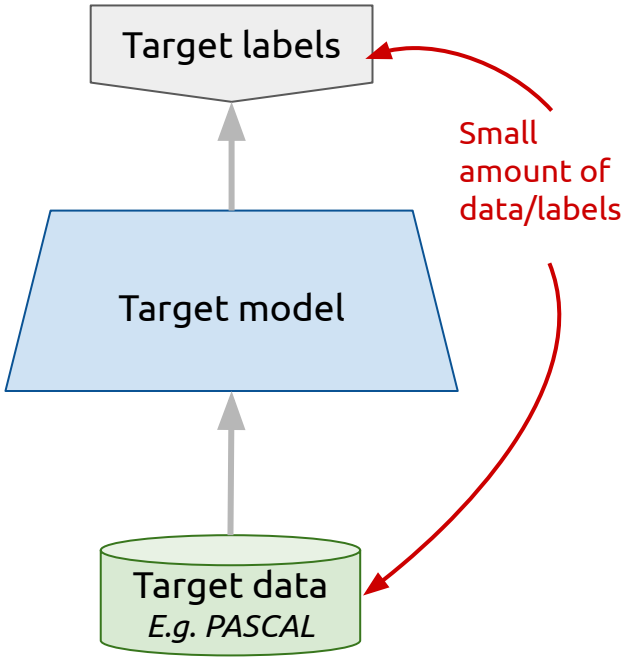
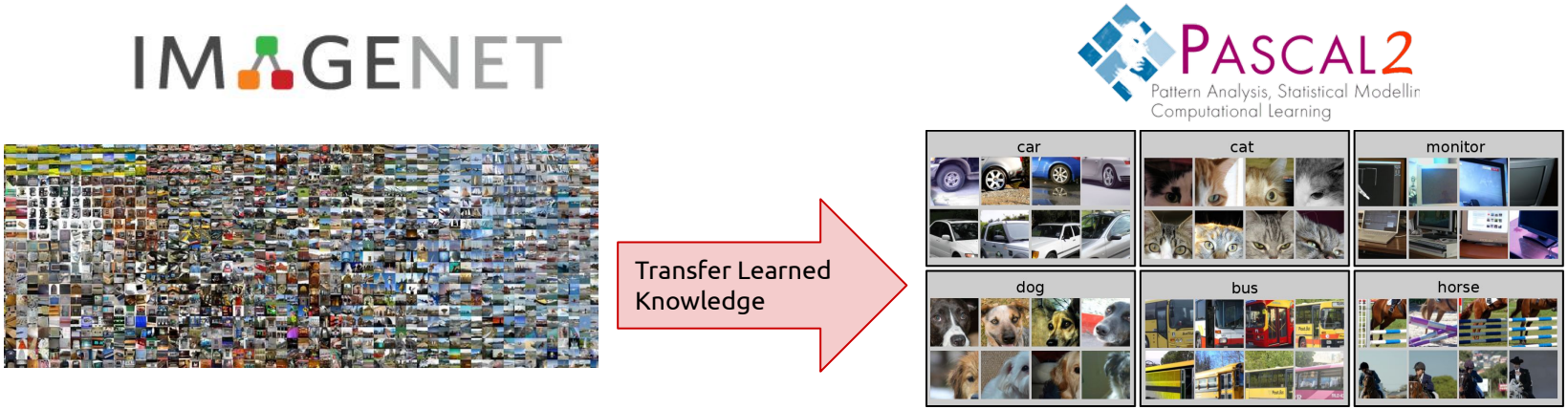


Figure: Kevin McGuinness

Transfer learning: idea

Instead of training a deep network from scratch for your task:

- Take a network trained on a different domain same/different **source task**
- Adapt it for your domain and your **target task**



Source domain / task

Target domain / task

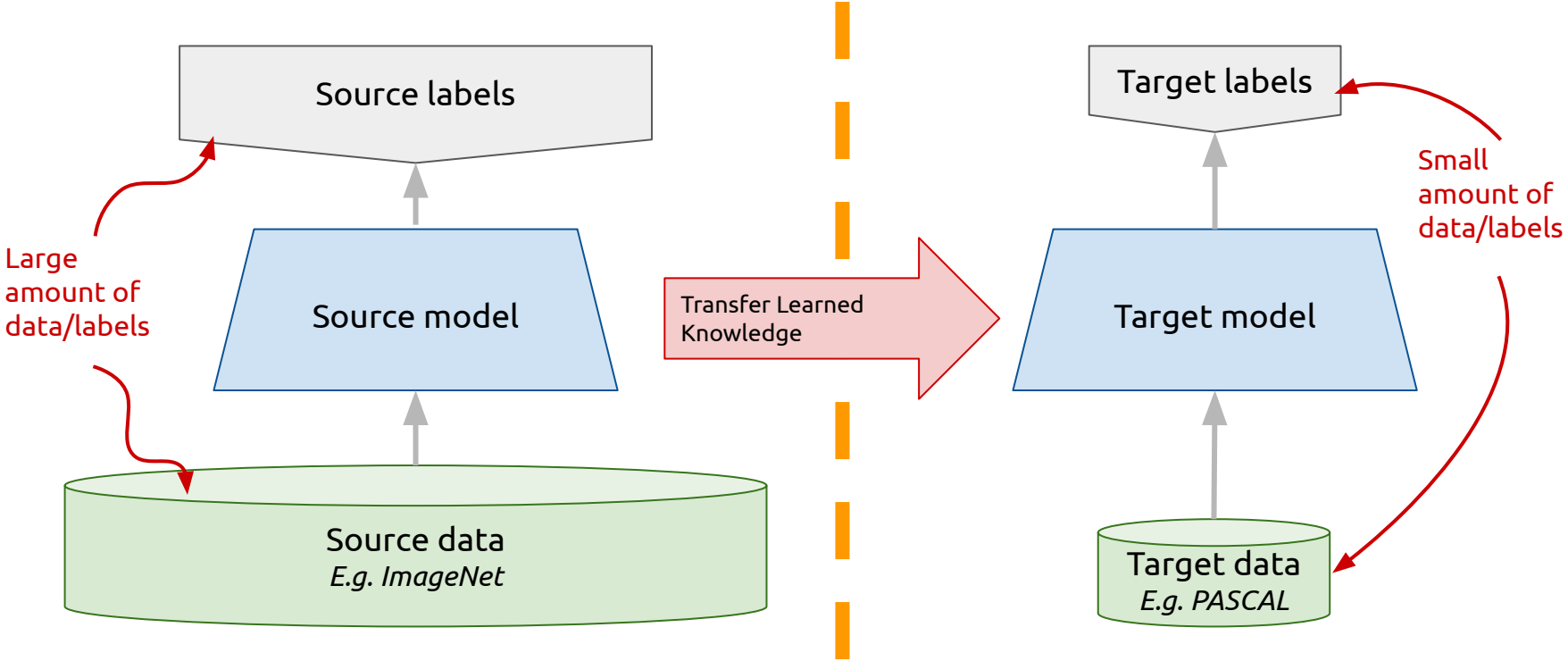
Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang et al. "Imagenet large scale visual recognition challenge." IJCV2015. [web]

Slide credit: Kevin McGuinness

Transfer Learning

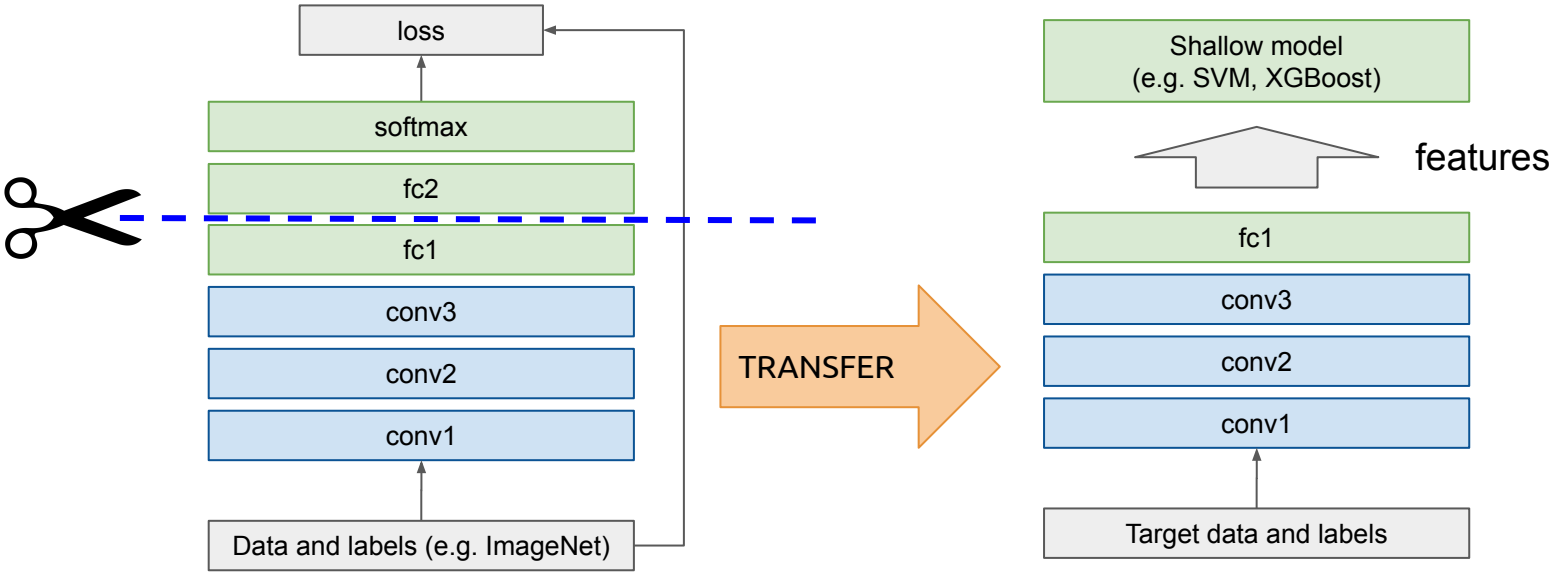
Instead of training a deep network from scratch for your task:

- Take a network trained on a different **source domain and/or task**
- Adapt it for your **target domain and/or task**



“Off-the-shelf”

Idea: use outputs of one or more layers of a network trained on a different task as generic feature detectors. Train a new shallow model on these features.

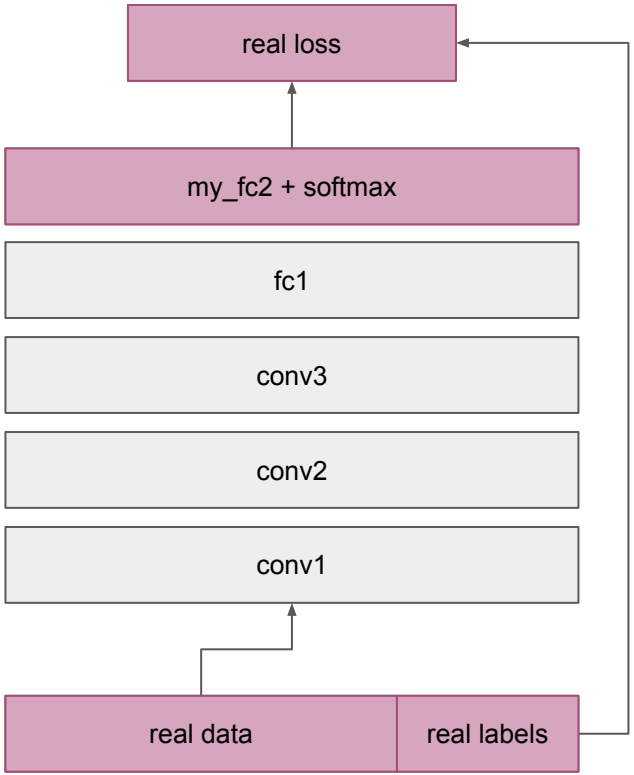


Sharif Razavian, A., Azizpour, H., Sullivan, J., & Carlsson, S. CNN features off-the-shelf: an astounding baseline for recognition. CVPR Workshops 2014.

Linear probing or Fine-tuning

Cut off top layer(s) of network and replace with supervised objective for target domain

- **Linear probing:** Train my_fc2 + softmax only.
- **Fine-tuning:** Backprop gradients to modify parameters of the pre-trained model.



#Decaf Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. . Decaf: A deep convolutional activation feature for generic visual recognition. ICML 2014.
Oquab, M., Bottou, L., Laptev, I., & Sivic, J. Learning and transferring mid-level image representations using convolutional neural networks. CVPR 2014.

The Gelato Bet (Sept 23, 2014)

*"If, by the first day of autumn (Sept 23) of 2015, a method will exist that can match or beat the performance of R-CNN on Pascal VOC detection, without the use of any extra, human annotations (e.g. ImageNet) as pre-training, **Mr. Malik** promises to buy **Mr. Efros** one (1) gelato (2 scoops: one chocolate, one vanilla)."*

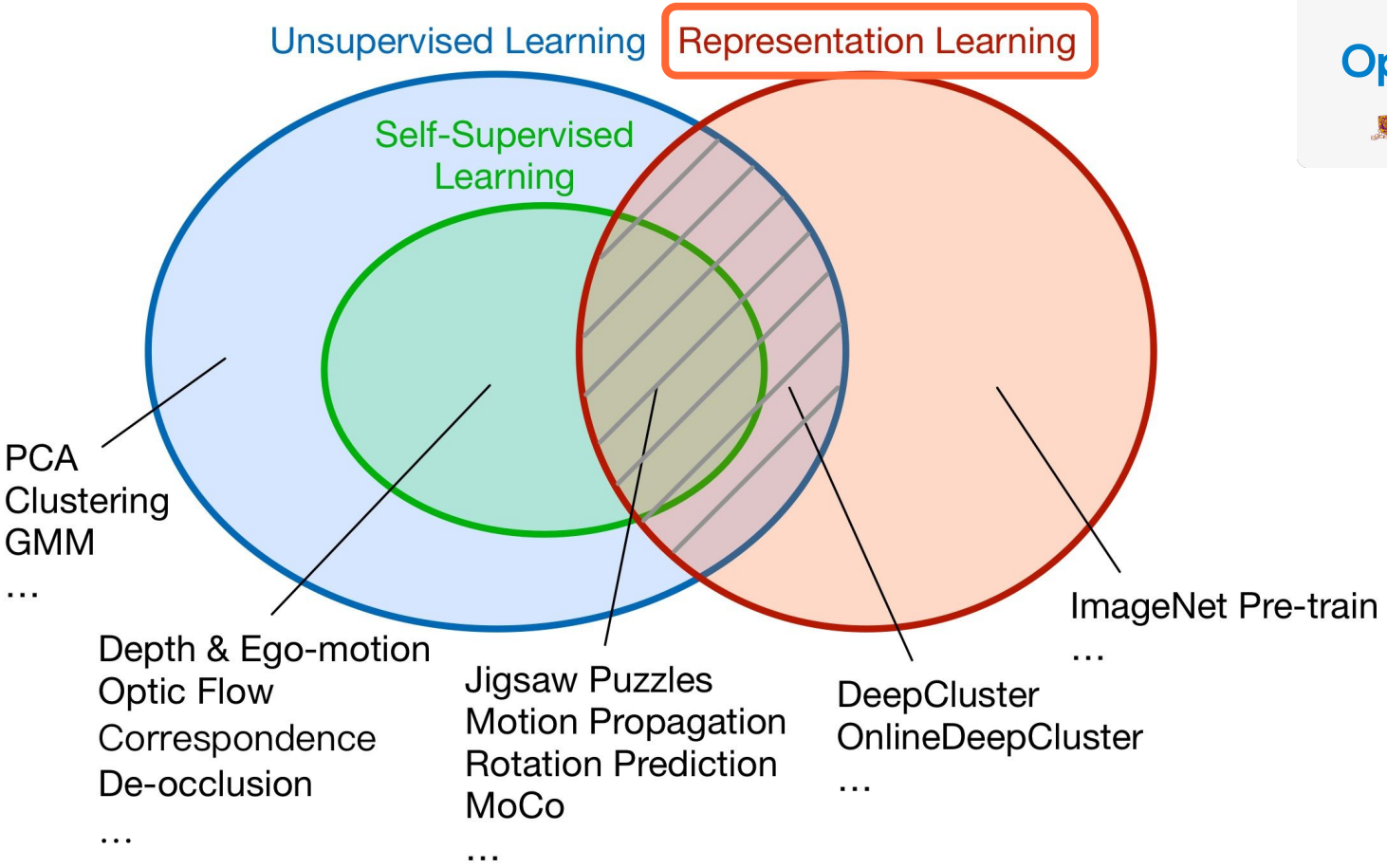
Source: Ankesh Anand, "Contrastive Self-Supervised Learning" (2020)"

Prof. Jitendra Malik

Prof. Alexei A. Efros

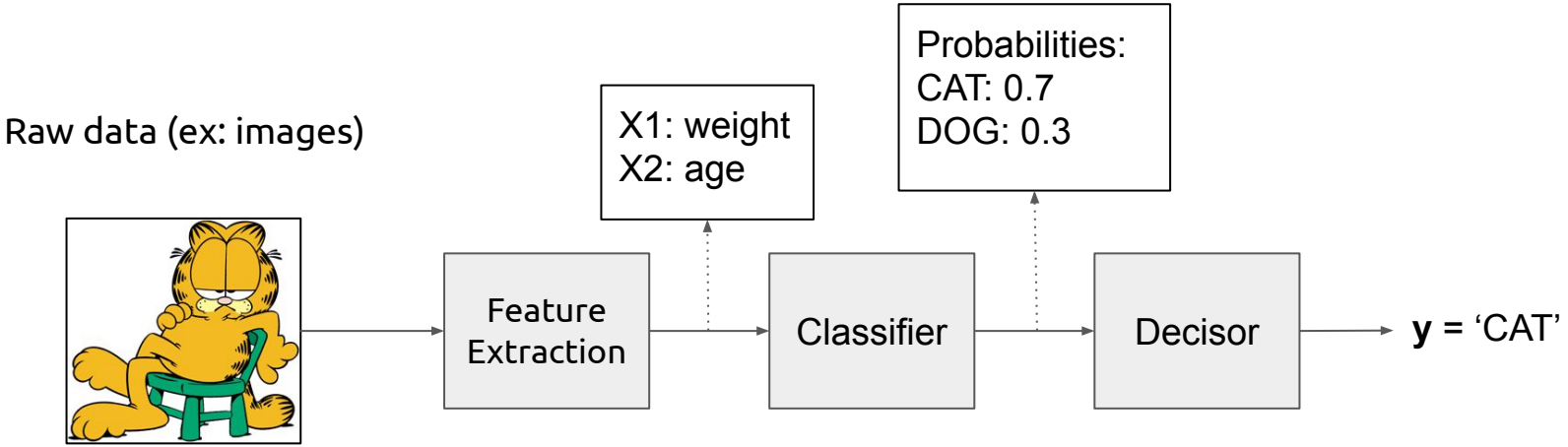
Outline

1. Transfer Learning
2. **Representation Learning**
3. Unsupervised Learning
4. Self-supervised Learning
5. Predictive methods
6. Contrastive methods



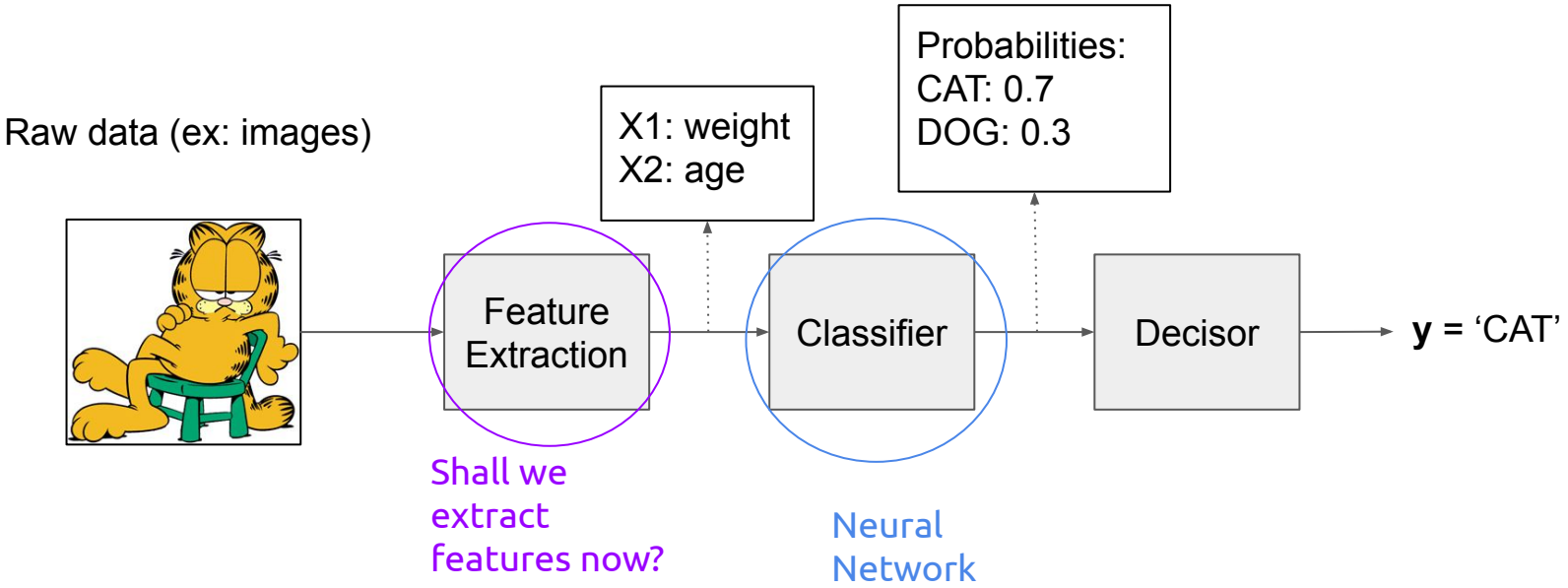
Source: OpenSelfSup by MMLAB (CUHK & Nanyang)

Representation + Learning pipeline



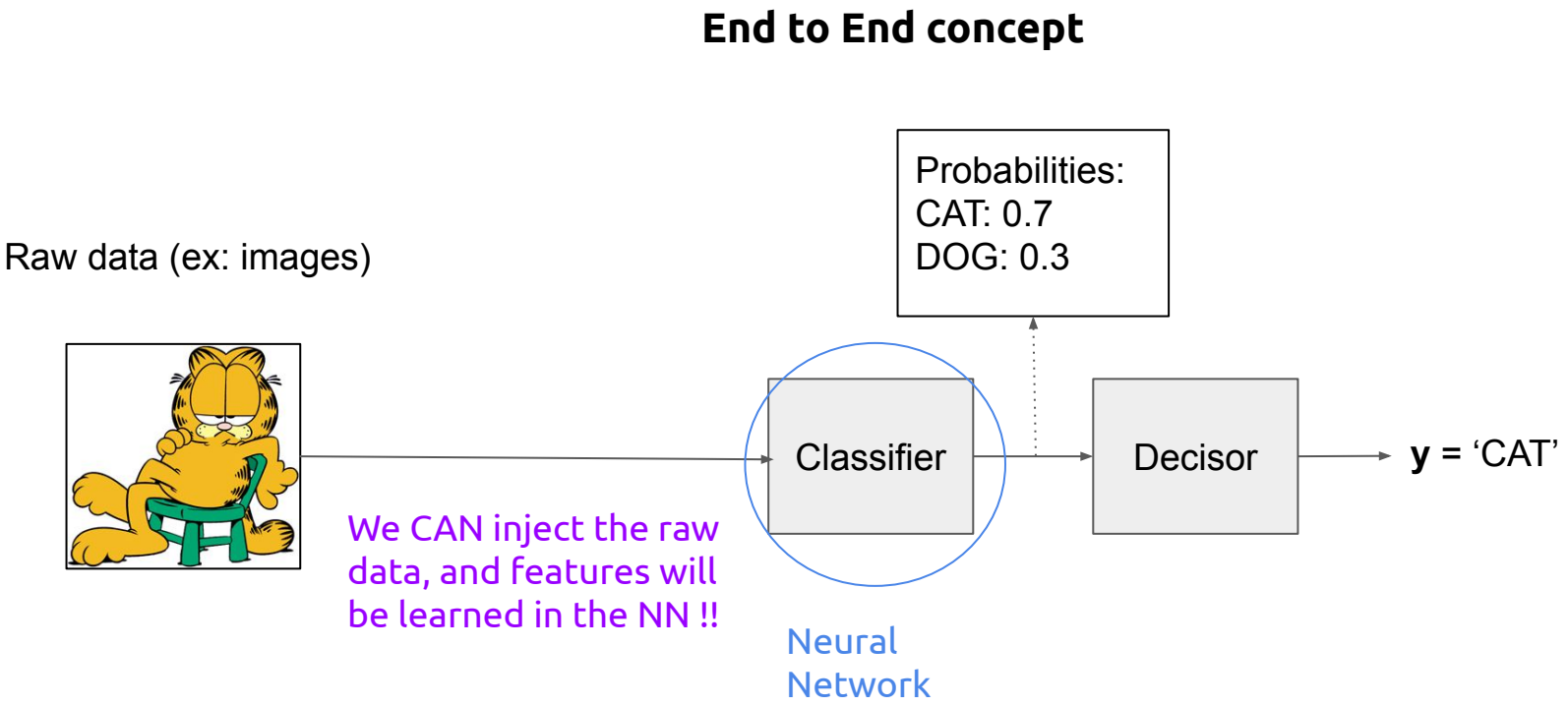
Slide concept: Santiago Pascual (UPC 2019). Garfield is a character created by Jim Davis.

Representation + Learning pipeline



Slide concept: Santiago Pascual (UPC 2019). Garfield is a character created by Jim Davis.

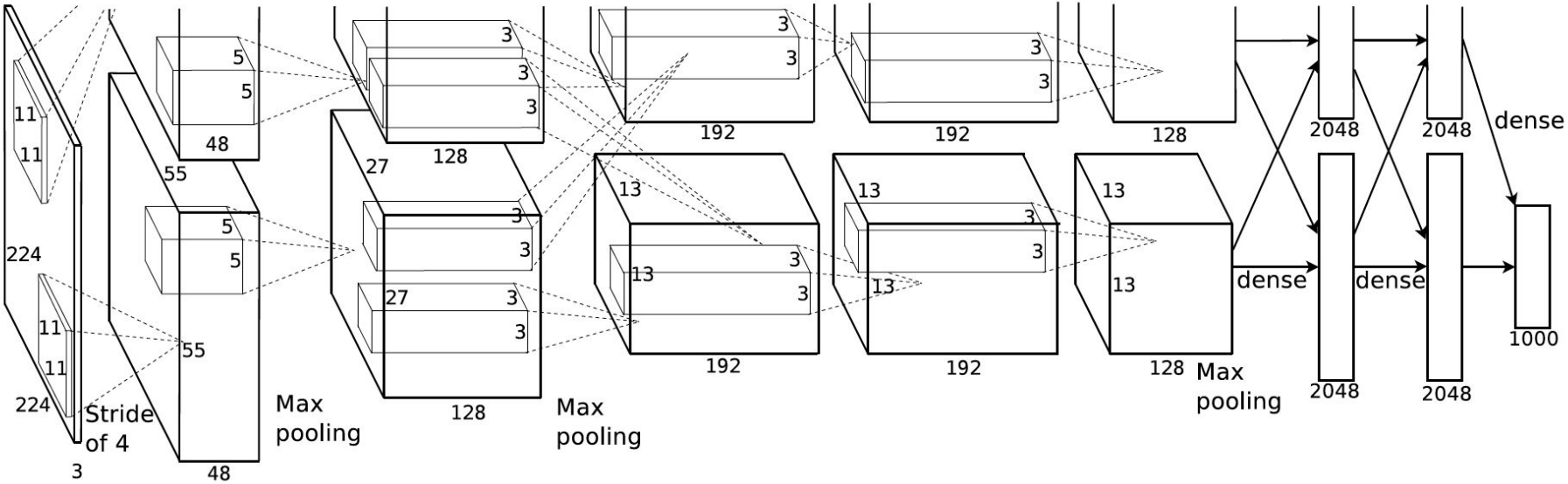
End-to-end Representation Learning



Slide concept: Santiago Pascual (UPC 2019). Garfield is a character created by Jim Davis.

the BIG BANG

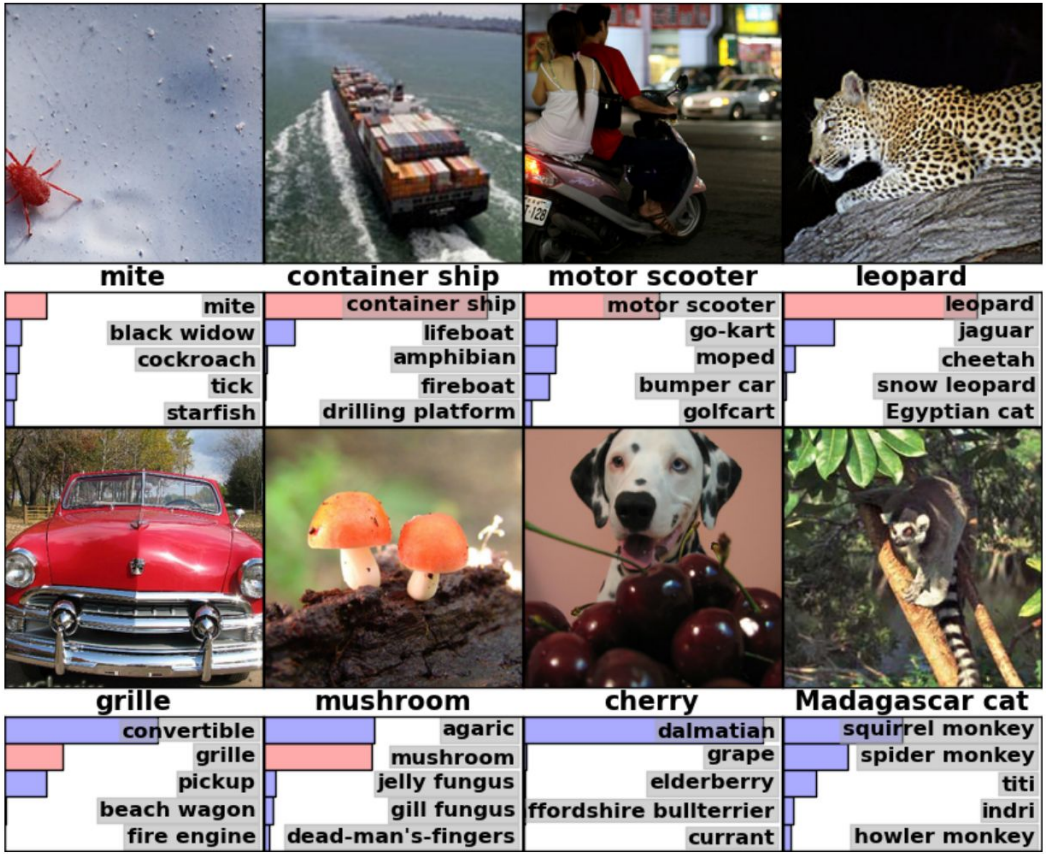
76,278 citations
(February 2021)



Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." NIPS **2012**

IMAGENET

- 1,000 object classes (categories).
- Labeled Images:
  - 1.2 M train
  - 100k test.

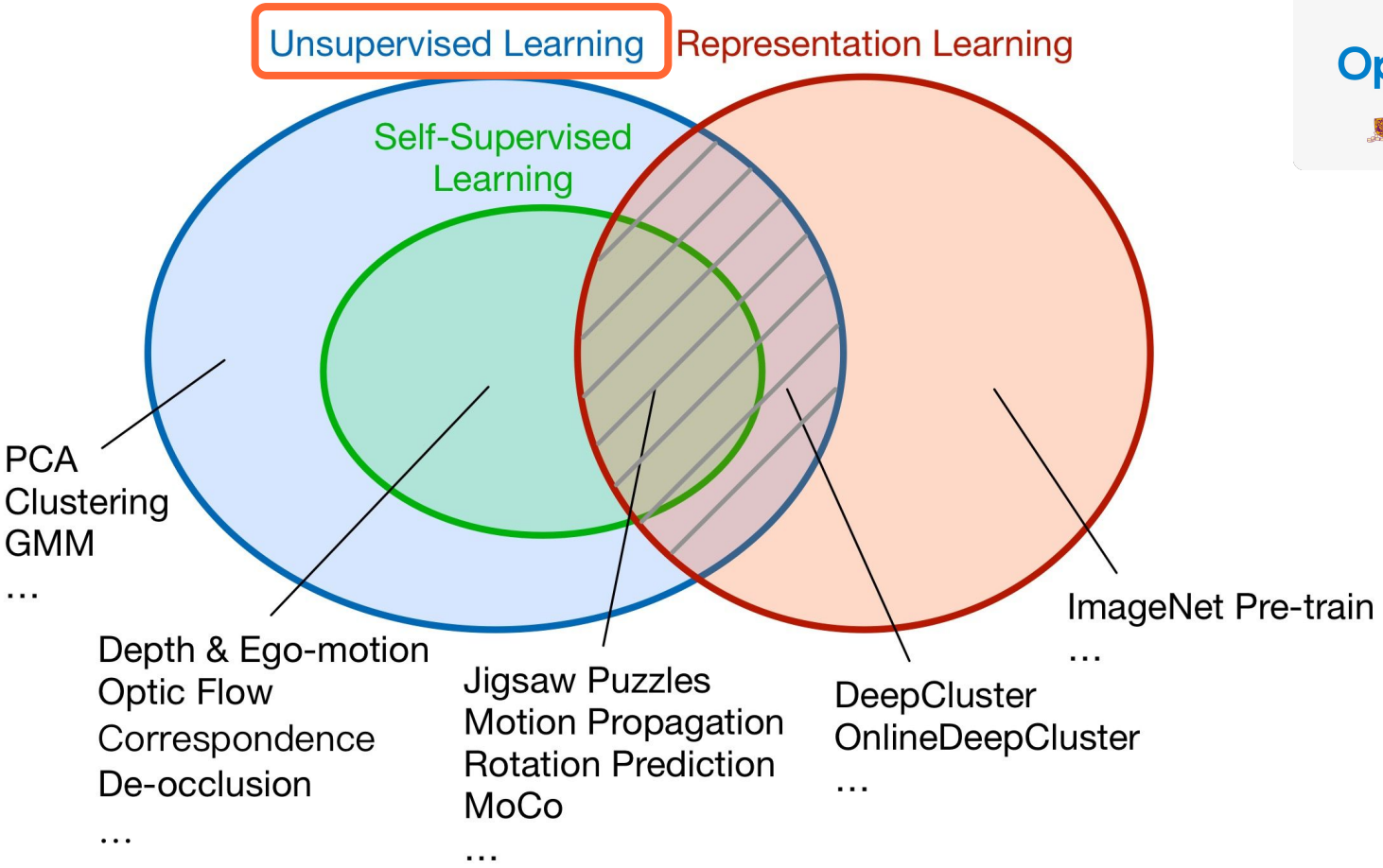


Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang et al. "Imagenet large scale visual recognition challenge." International Journal of Computer Vision 115, no. 3 (2015): 211-252. [web]

Slides from Yann LeCun (ICLR 2020) [tweet by @lefterav]

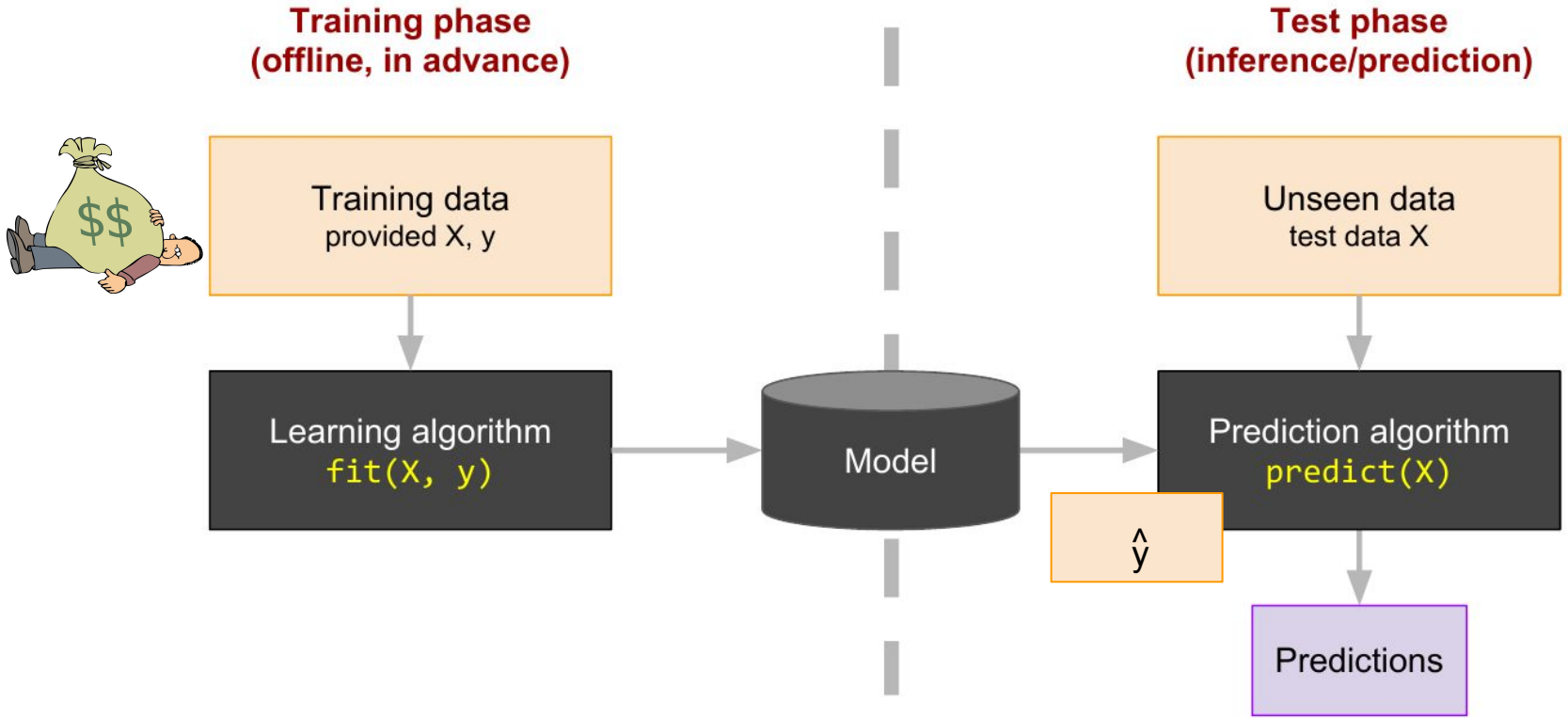
Outline

1. Transfer Learning
2. Representation Learning
3. **Unsupervised Learning**
4. Self-supervised Learning
5. Predictive methods
6. Contrastive methods



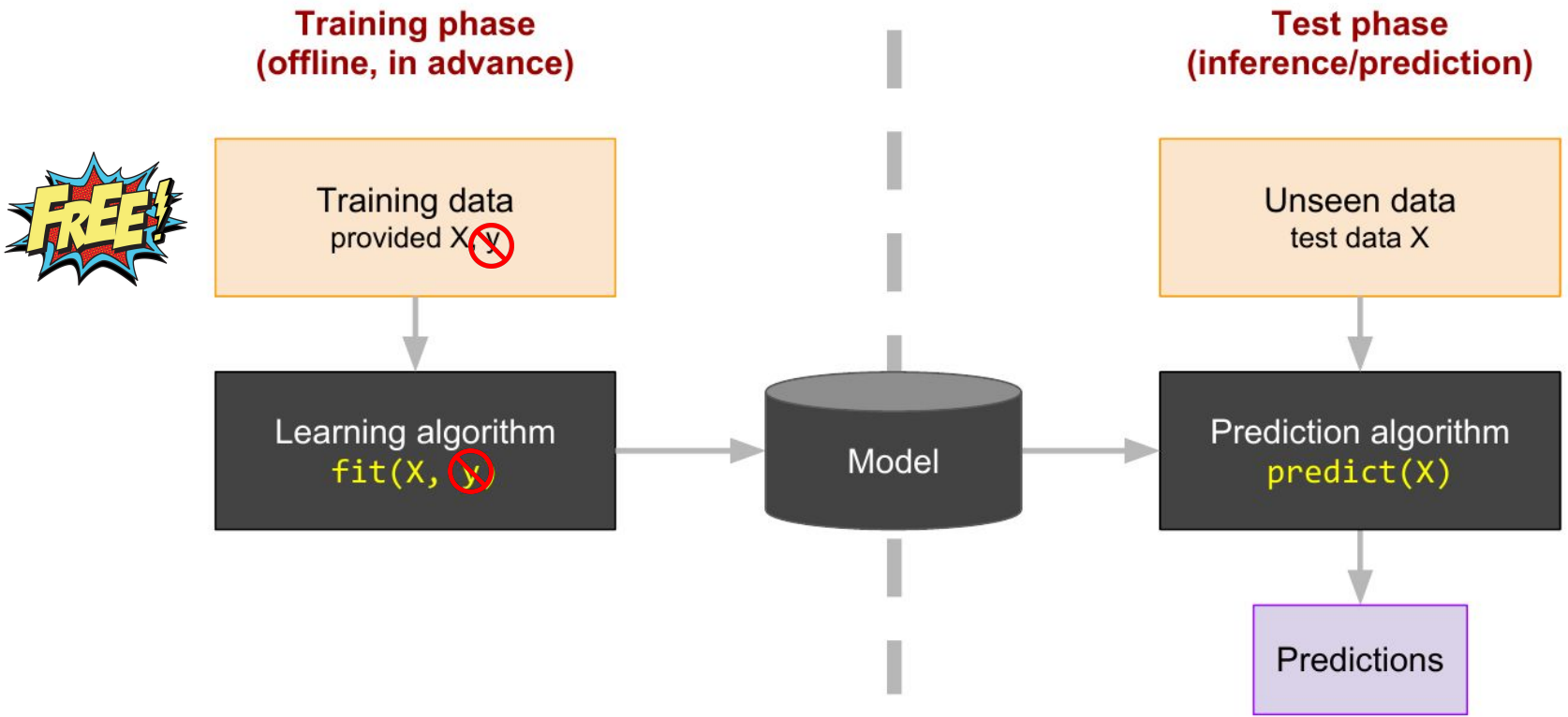
Source: OpenSelfSup by MMLAB (CUHK & Nanyang)

Supervised learning



Slide credit: Kevin McGuinness

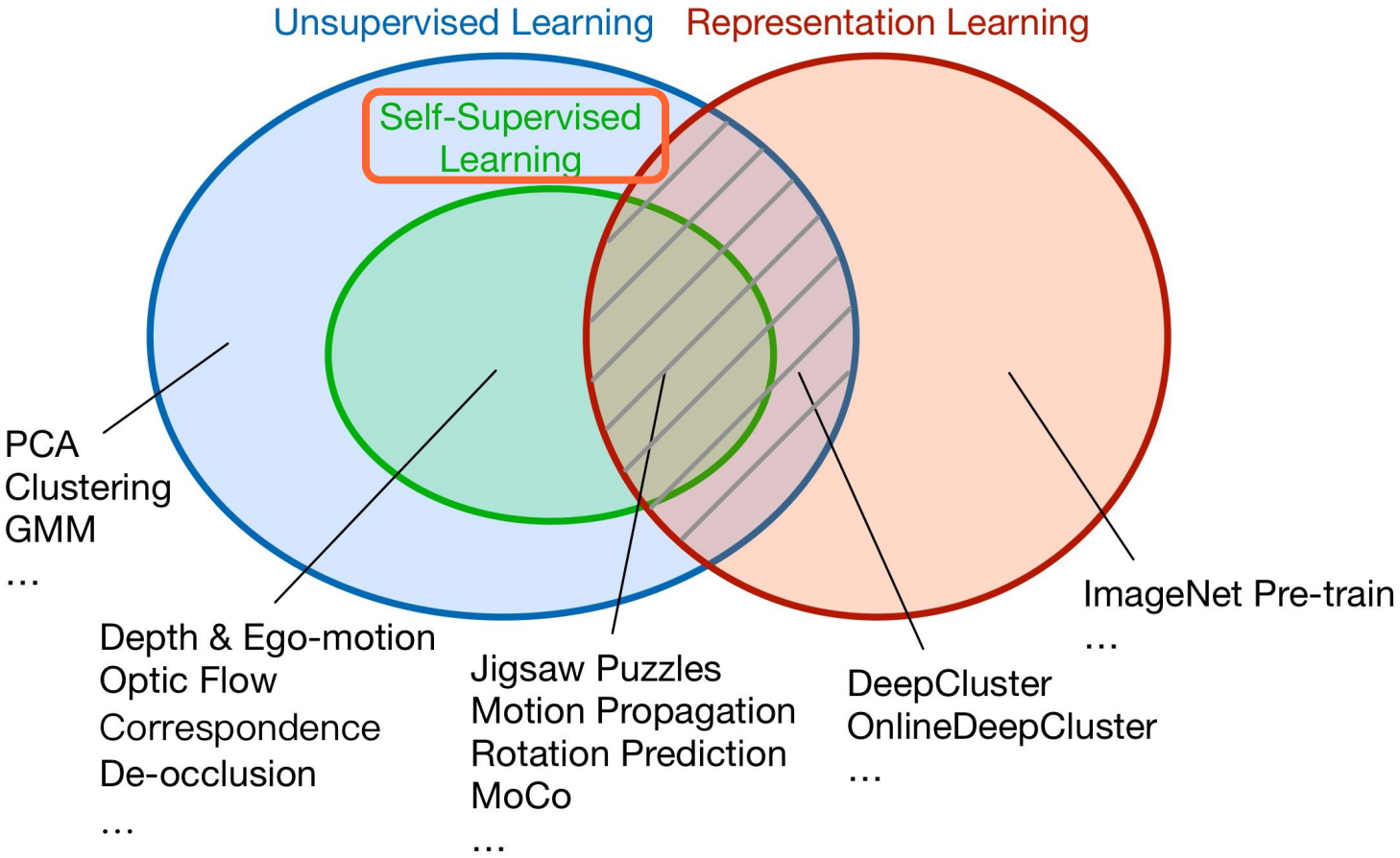
Unsupervised learning



Slide credit: Kevin McGuinness

Outline

1. Representation Learning
2. Unsupervised Learning
3. **Self-supervised Learning**
4. Predictive methods
5. Contrastive methods



Source: OpenSelfSup by MMLAB (CUHK & Nanyang)

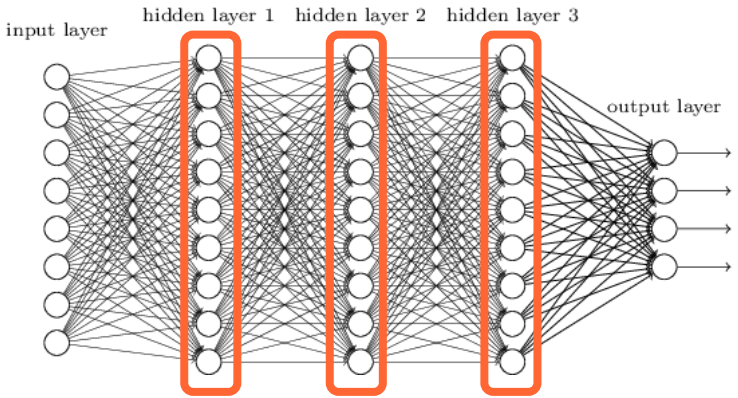
Self-supervised learning

Self-supervised learning is a form of unsupervised learning where raw data (no annotations) is enough to compute the gradients.

- A **pretext (or surrogate) learning task** must be designed.
- The NN **learns representations**, which should be transferable to the actual **downstream task of interest**.

Unlabeled data
(X)

Representations learned without labels

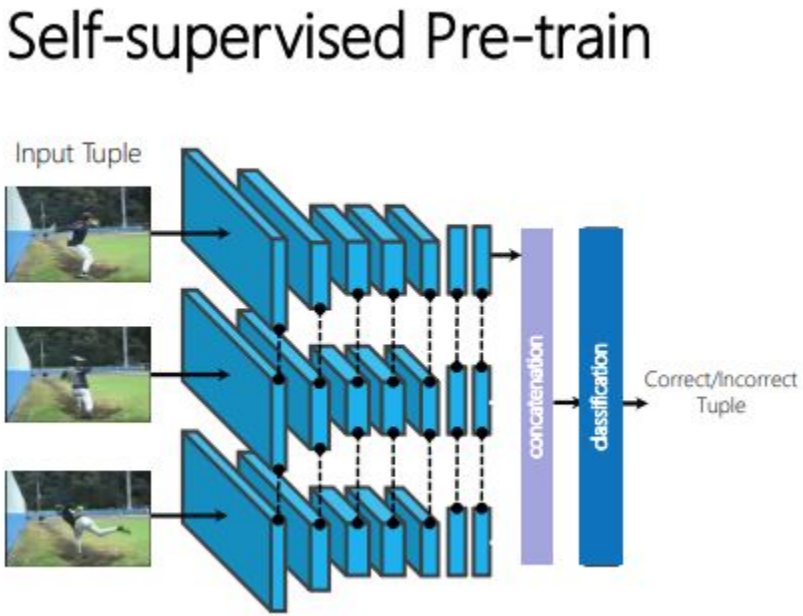


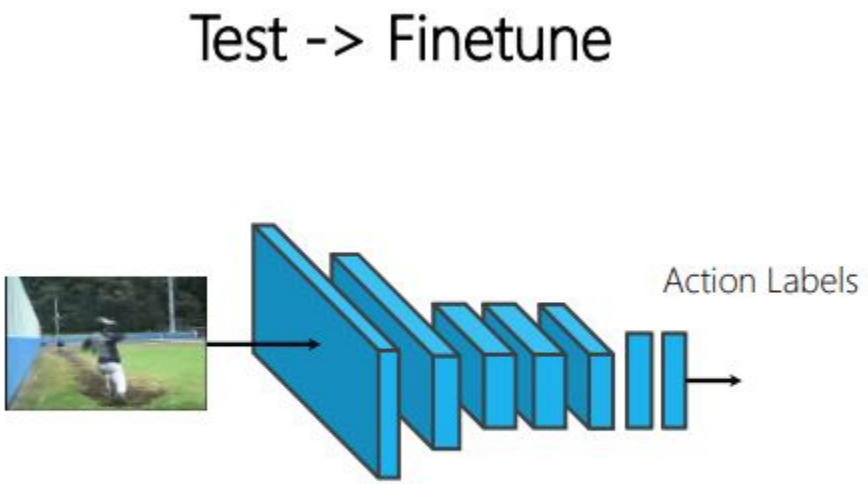
Reference: Andrew Zisserman (PAISS 2018)

Self-supervised Learning

- Scalability.
- Avoid ambiguities in annotation criteria.

Self-supervised + Transfer Learning





#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):

Outline

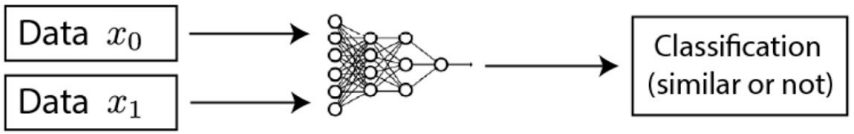
1. Transfer Learning
2. Representation Learning
3. Unsupervised Learning
4. Self-supervised Learning
5. **Predictive methods**
6. Contrastive methods

Predictive vs Contrastive Methods

Generative / Predictive

Loss measured in the output space

Contrastive



Loss measured in the representation space

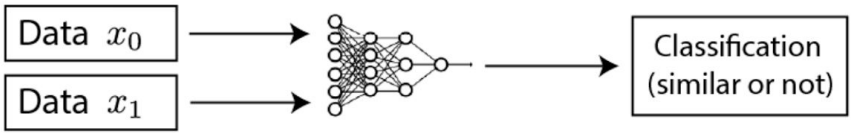
Source: Ankesh Anand, "Contrastive Self-Supervised Learning" (2020)"

Predictive vs Contrastive Methods

Generative / Predictive

Loss measured in the output space

Contrastive



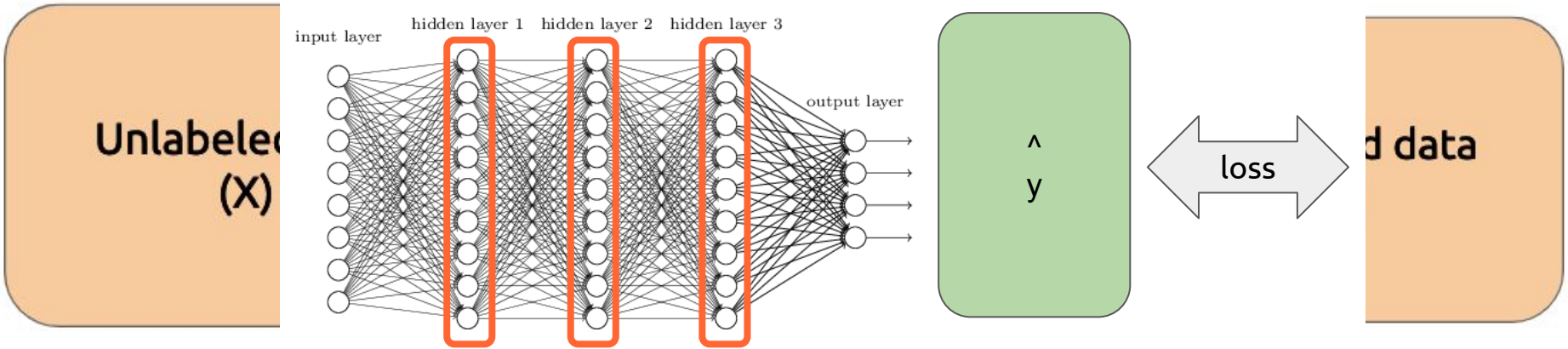
Loss measured in the representation space

Source: Ankesh Anand, "Contrastive Self-Supervised Learning" (2020)"

Predictive vs Contrastive Methods

Predictive Methods:
A **pretext (or surrogate) task** is defined by withholding a part of the unlabeled data and training the NN to predict it.

Representations learned without labels

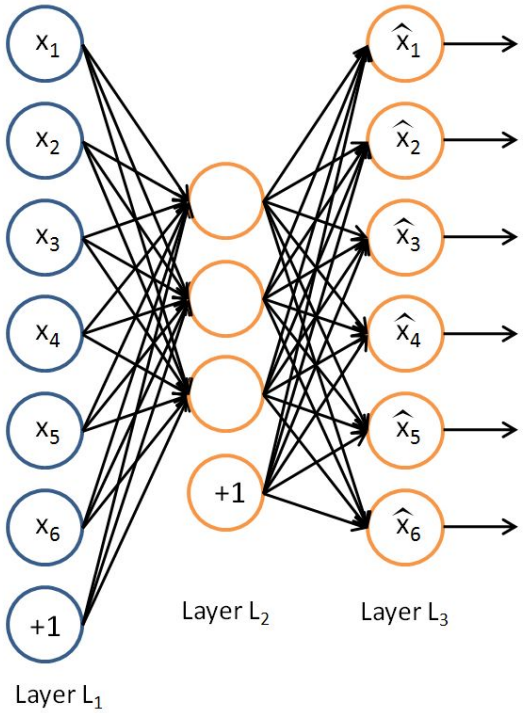


Reference: Andrew Zisserman (PAISS 2018)

Outline

1. Unsupervised Learning
2. Self-supervised Learning
3. Predictive Methods
    - **(Autoencoder)**

Autoencoder (AE)



Autoencoders:

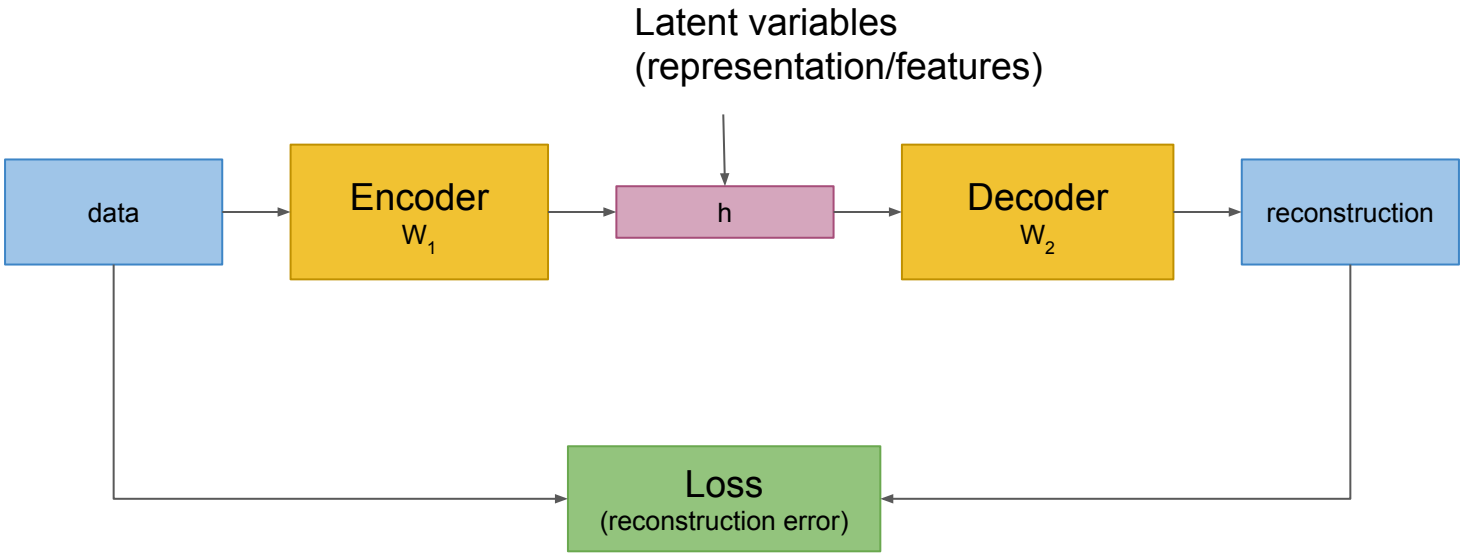
- Predict at the output the same input data.
- Do not need labels.

Fig: "Deep Learning Tutorial" Stanford

Autoencoder (AE)

WHY?

1. Initialize a NN by solving an autoencoding problem.

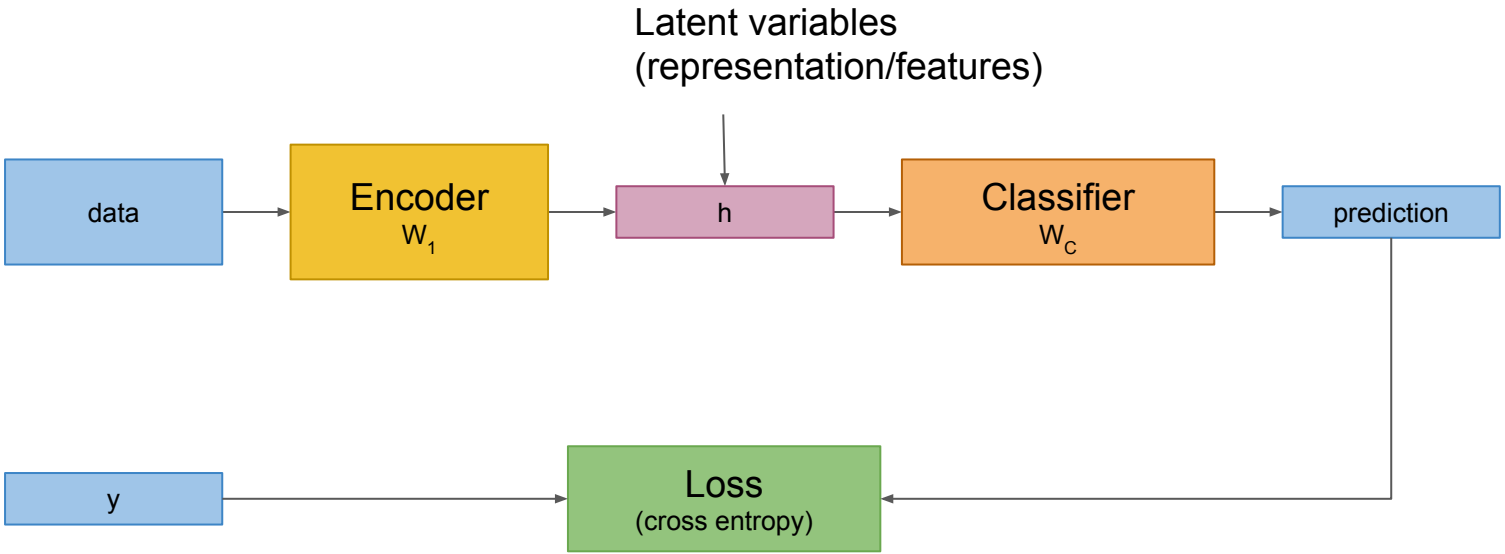


Slide: Kevin McGuinness (DLCV UPC 2017)

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)

Autoencoder (AE)

What other application does the AE target address ?

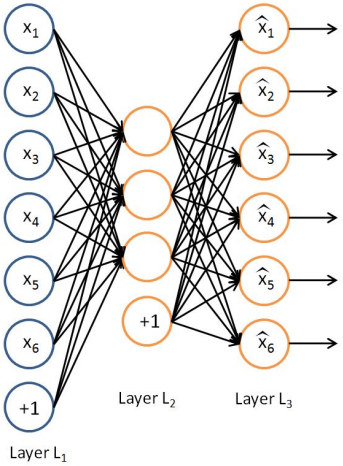
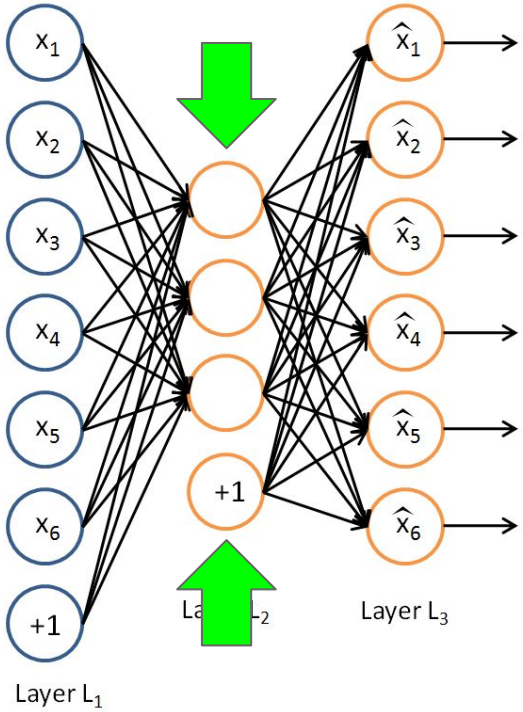


Fig: "Deep Learning Tutorial" Stanford

Autoencoder (AE)



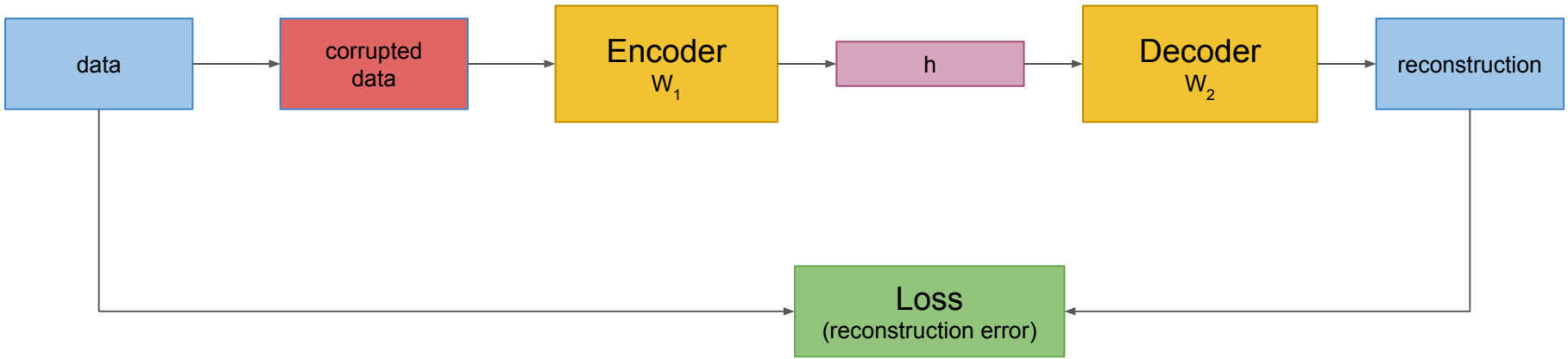
Dimensionality reduction:

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

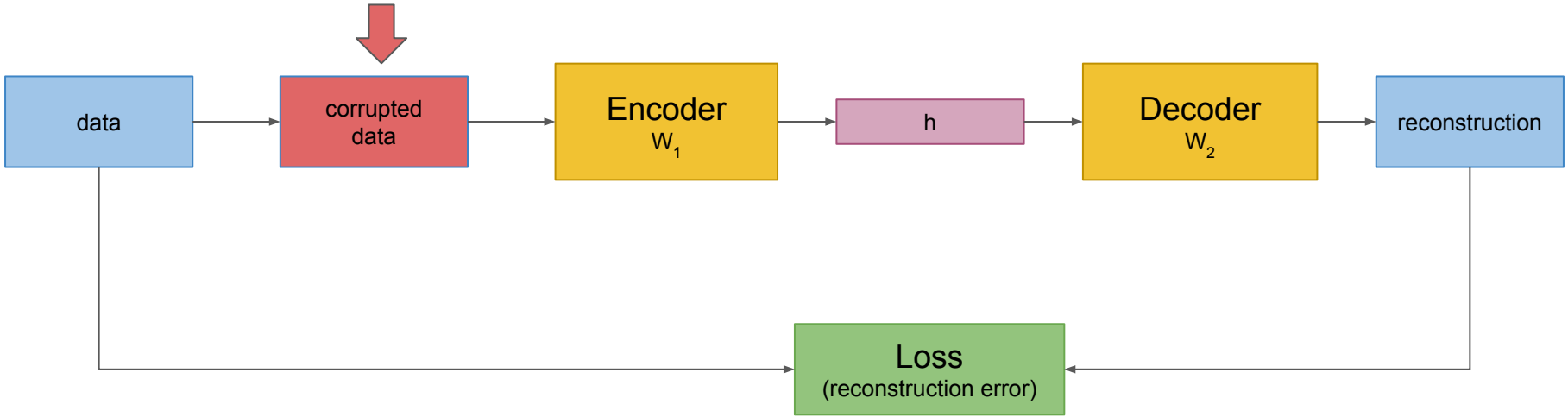
Fig: "Deep Learning Tutorial" Stanford

Denoising Autoencoder (DAE)

Corrupt the input signal, but predict the clean version at its output.

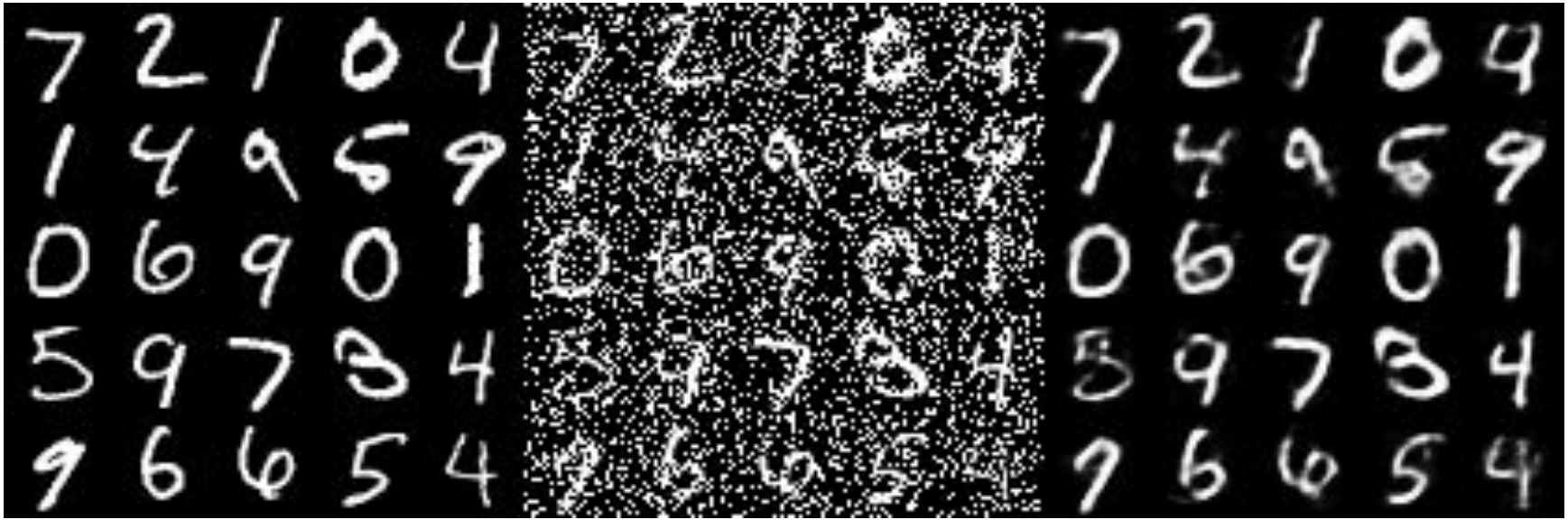


Vincent, Pascal, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. "Extracting and composing robust features with denoising autoencoders." ICML 2008.



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

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.

Outline

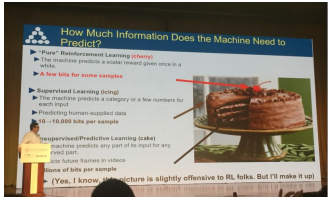
1. Unsupervised Learning
2. Self-supervised Learning
3. Predictive Methods
    - **Future Prediction**

Future Prediction

How Much Information is the Machine Given during Learning?

Y. LeCun

- **"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**
- **Self-Supervised Learning (cake génoise)**
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

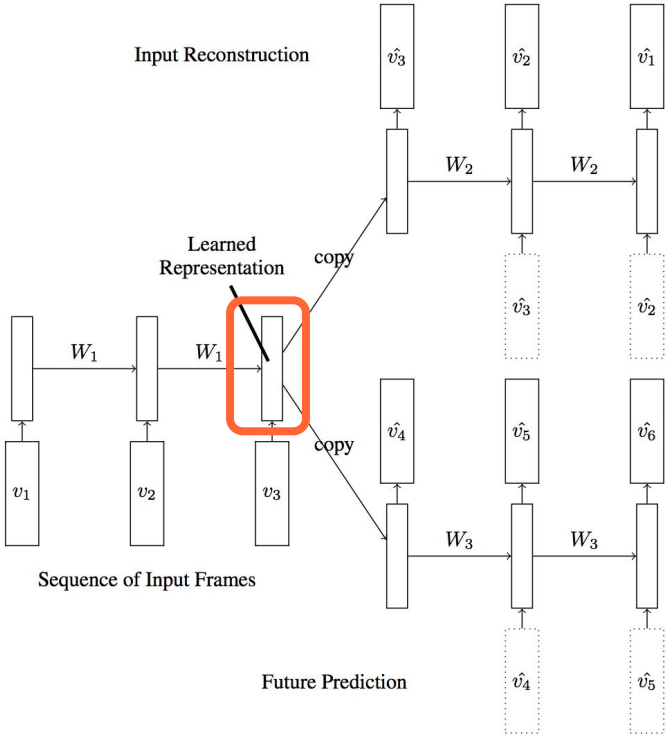


Slide credit:
Yann LeCun

[Full set of slides at @ylecun]

Future Prediction: Video Frames

Learning video representations (features) by...

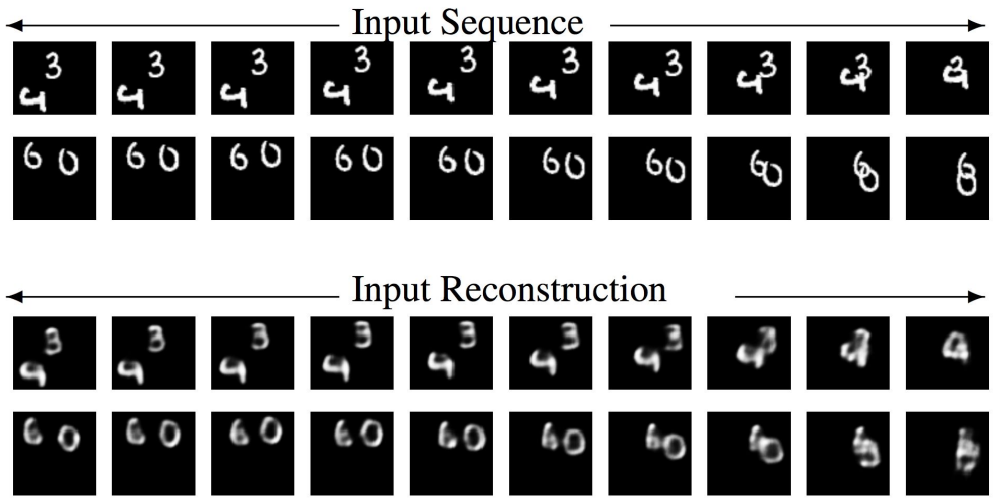


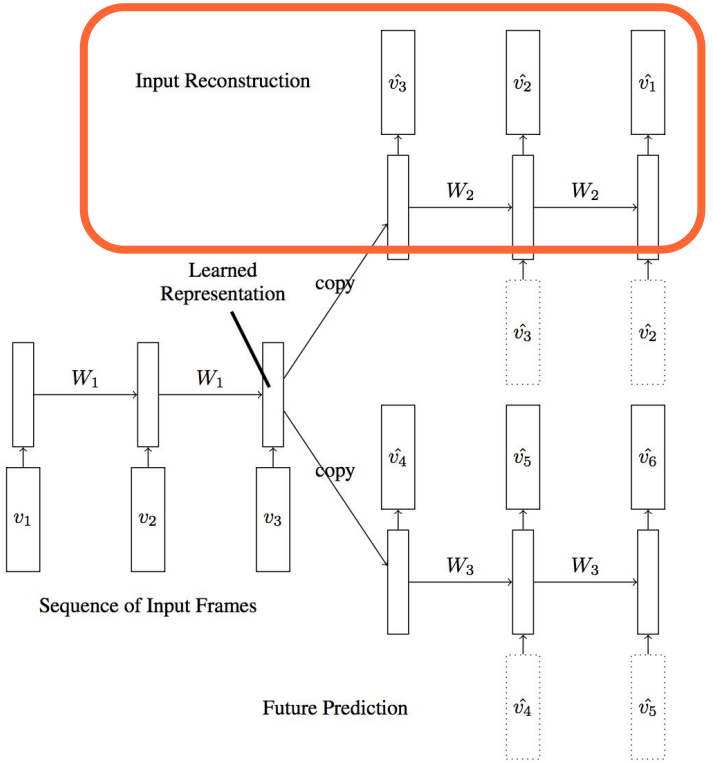
Srivastava, Nitish, Elman Mansimov, and Ruslan Salakhutdinov. "Unsupervised Learning of Video Representations using LSTMs." In ICML 2015. [Github]

Future Prediction: Video Frames

Learning video representations (features) by...

(1) frame reconstruction (AE):



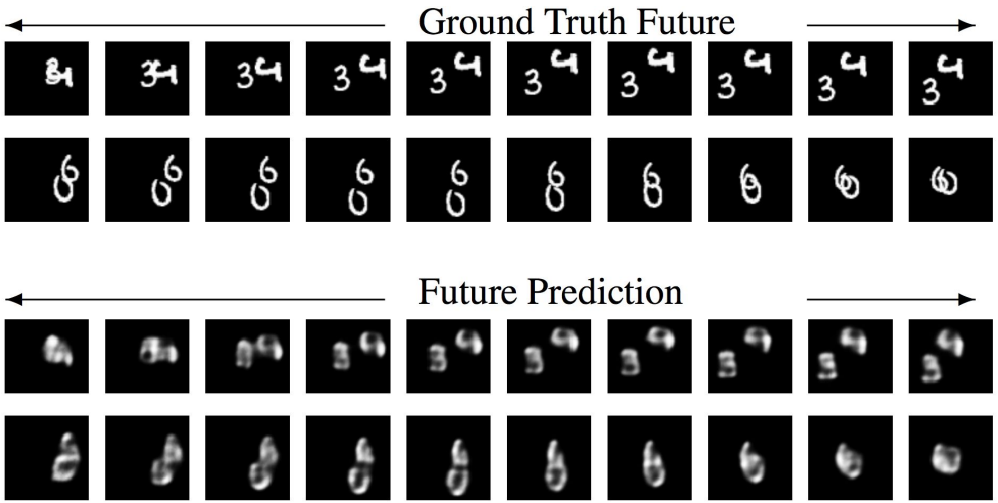


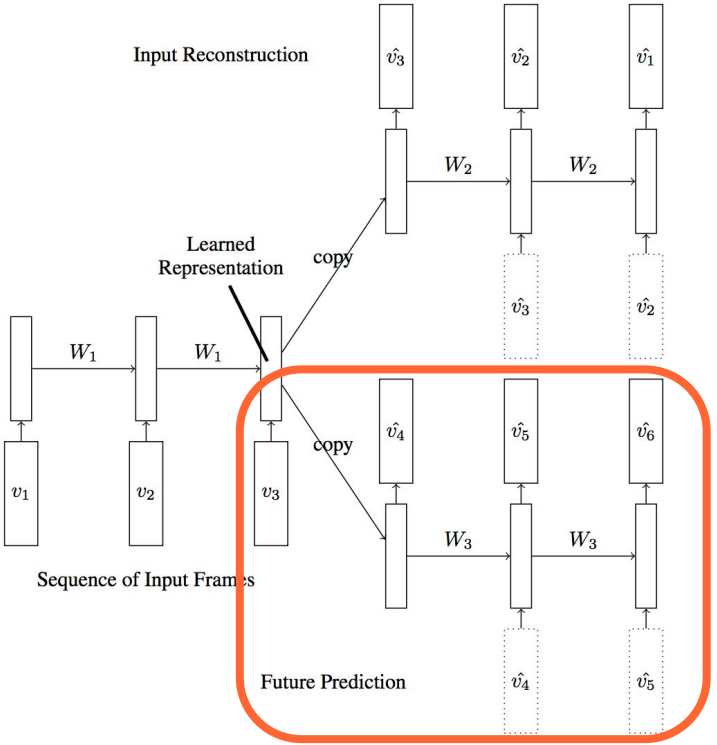
Srivastava, Nitish, Elman Mansimov, and Ruslan Salakhutdinov. "Unsupervised Learning of Video Representations using LSTMs." ICML 2015. [Github]

Future Prediction: Video Frames

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]

Future Prediction: Video Frames

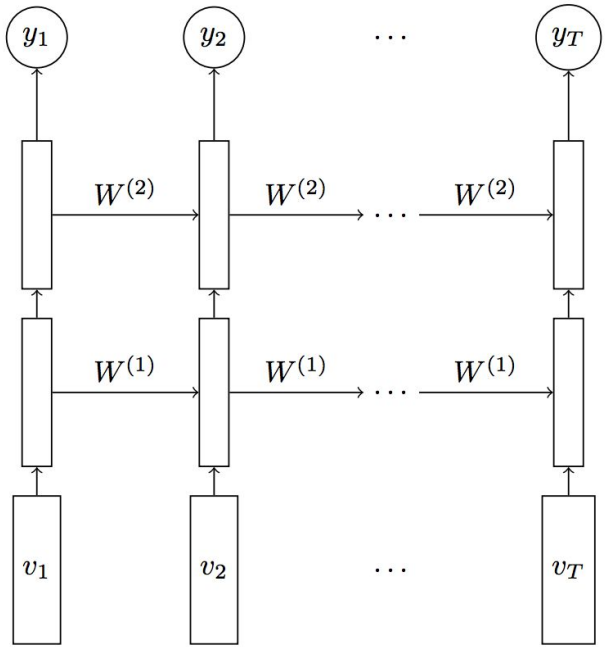


Figure 6. LSTM Classifier.

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]

Future Prediction: Words

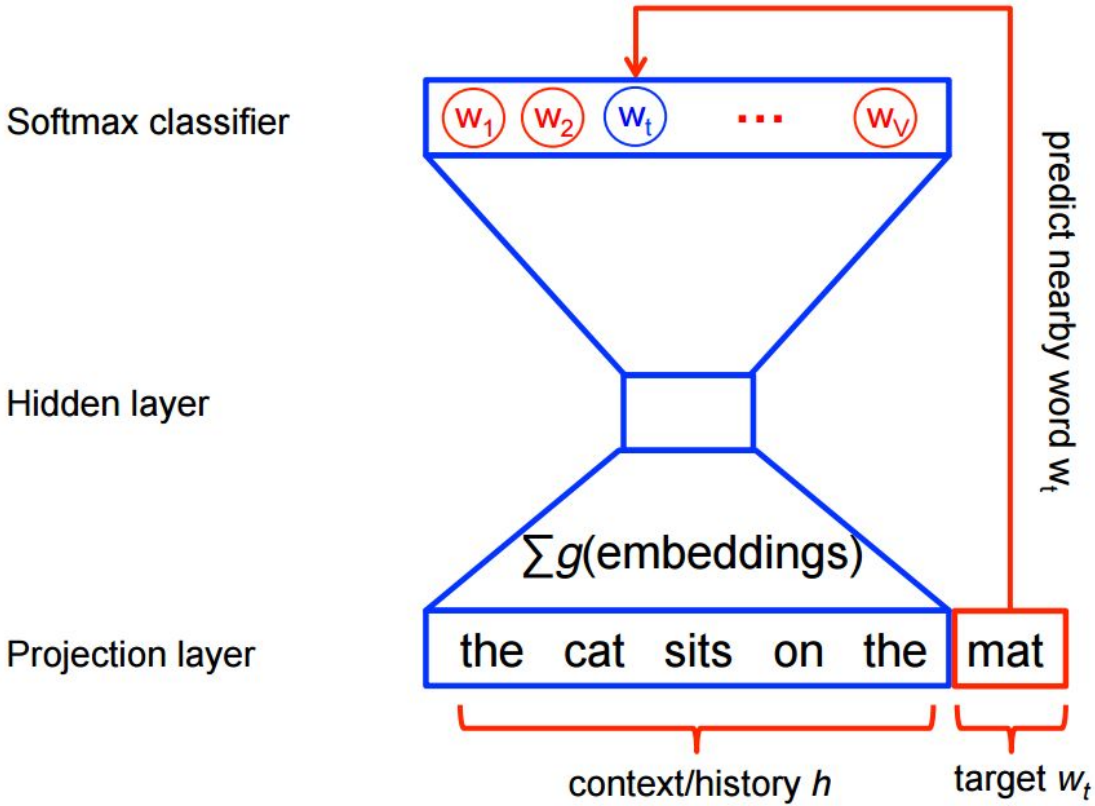


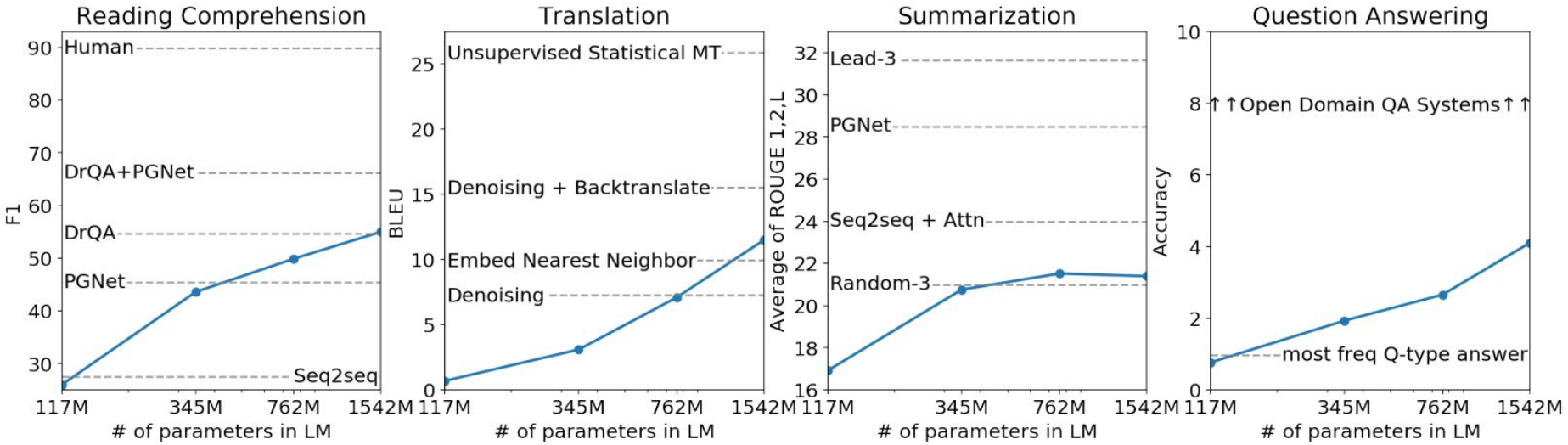
Figure:
TensorFlow tutorial

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.

Future Prediction: Words

"GPT-2 is trained with a simple objective: predict the next word, given all of the previous words within some text."

Zero-shot task performances
(GPT-2 was never trained for these tasks)



#GPT-2 Alec Radford, Jeffrey Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage, Ilya Sutskever, "Better Language Models and Their Implications". OpenAI Blog 2019.

Future Prediction: Words

Training a very large language model with a very large amount of data can result in impressive results:

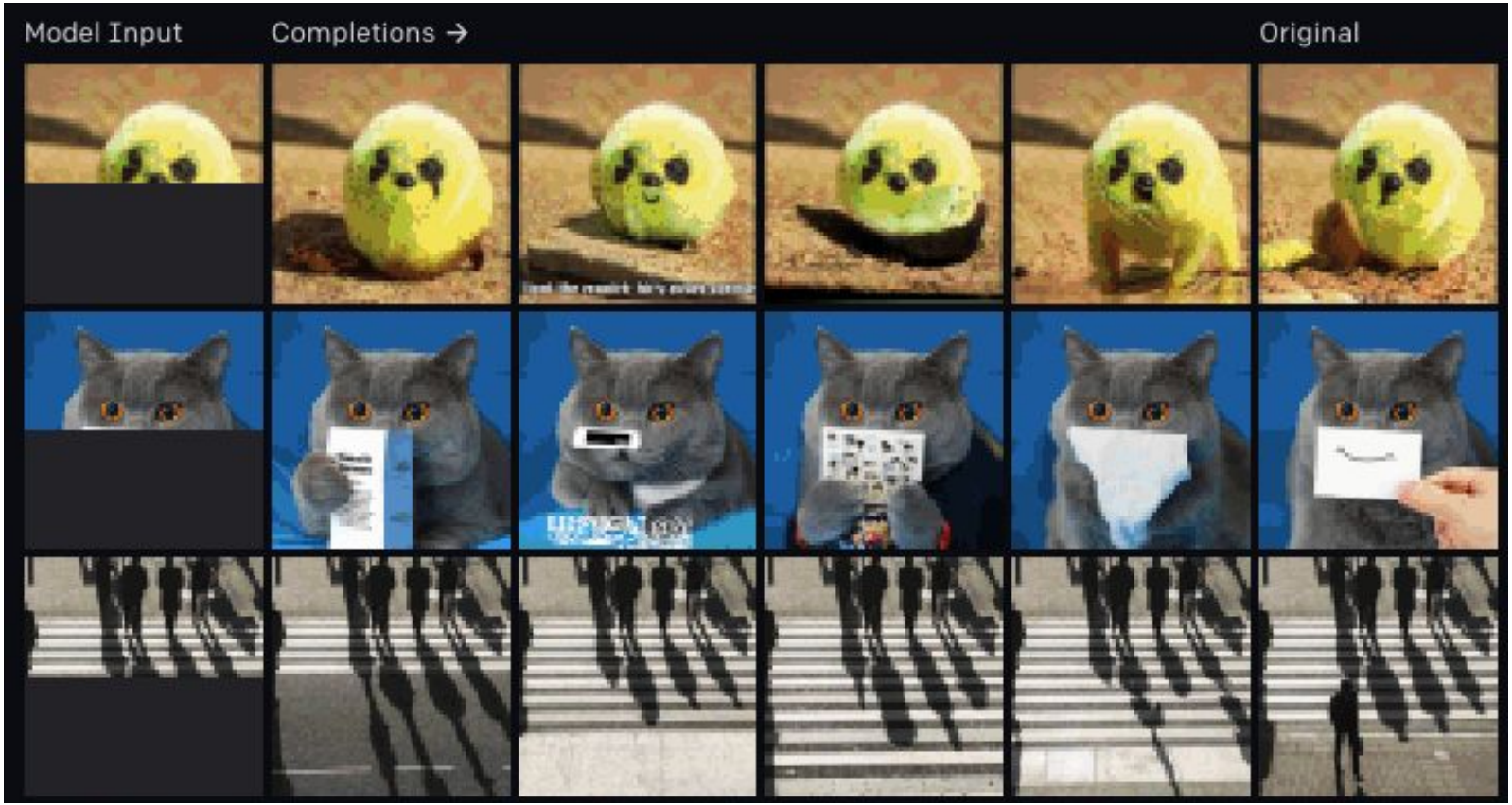
Video Source: @mattshumer_

OpenAI

[OpenAI API]

#GPT-3 Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Agarwal, S. Language models are few-shot learners. NeurIPS 2020 (best paper award).

Future Prediction: Pixels



#iGPT Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., & Sutskever, I. Generative Pretraining from Pixels. ICML 2020.

Future Prediction: Pixels

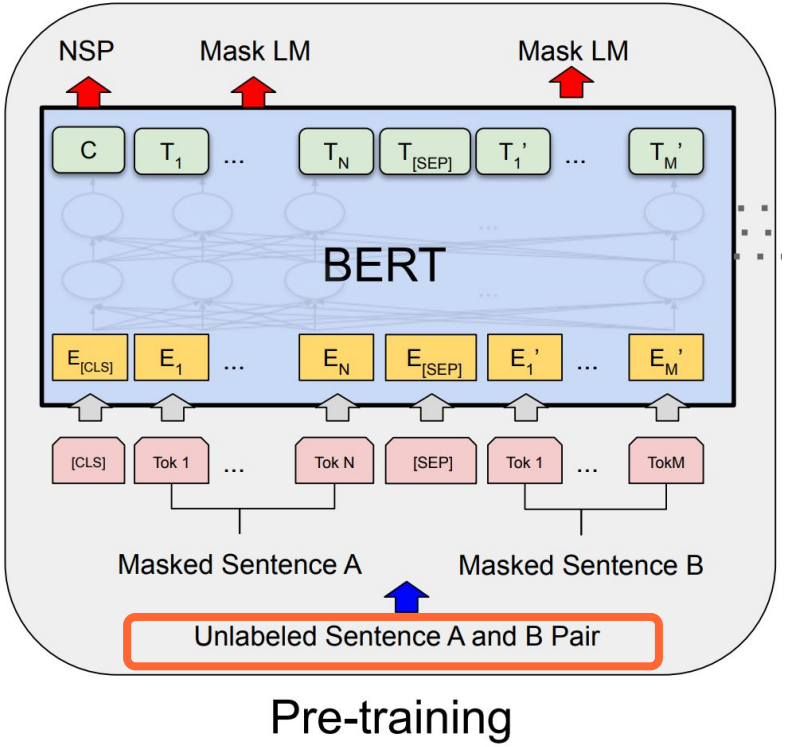
EVALUATION	MODEL	ACCURACY	PRE-TRAINED ON IMAGENET W/O LABELS	PRE-TRAINED ON IMAGENET W/ LABELS
CIFAR-10 Linear Probe	ResNet-152[50]	94.0		✓
	SimCLR[12]	95.3	✓	
	iGPT-L 32x32	**96.3**	✓	
CIFAR-100 Linear Probe	ResNet-152	78.0		✓
	SimCLR	80.2	✓	
	iGPT-L 32x32	**82.8**	✓	

#iGPT Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., & Sutskever, I. Generative Pretraining from Pixels. ICML 2020.

Future Prediction: True / False

The BERT language model is pre-trained for a **binarized next sentence prediction task**:

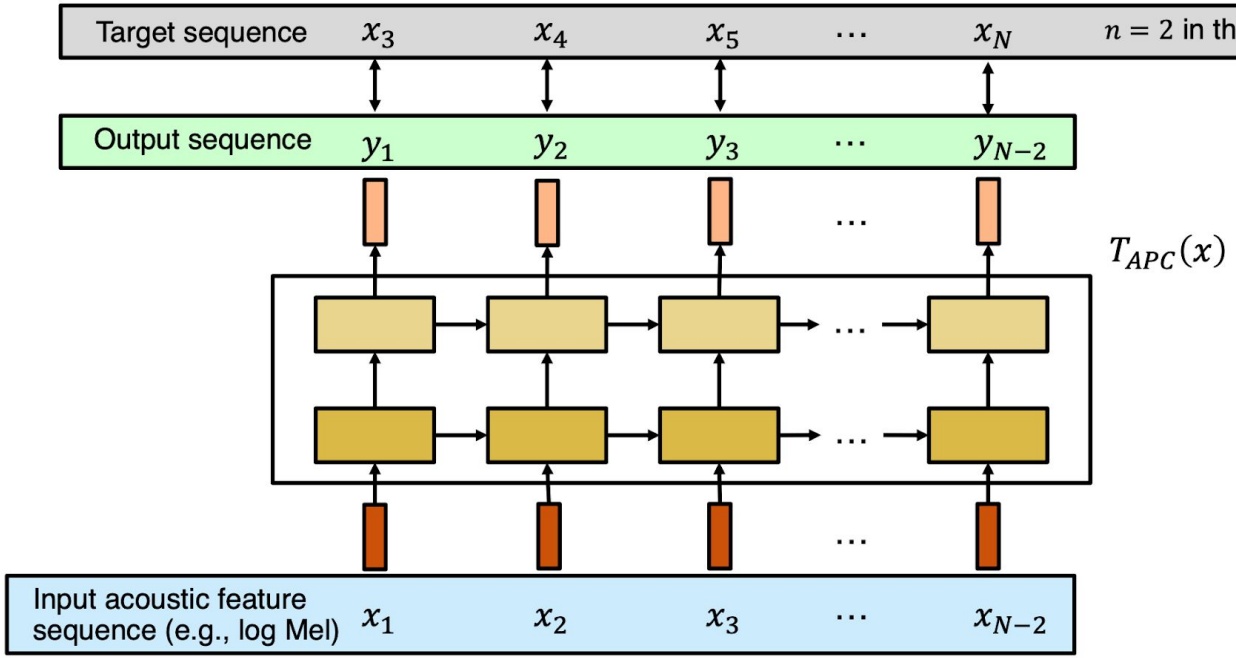
Select some sentences for A and B, where 50% of the data B is the next sentence of A, and the remaining 50% of the data B are randomly selected in the corpus, and learn the correlation.



#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)**.

Future Prediction: Audio Spectrogram

Predict spectrogram *n* samples in the future.



- **Training**

$$\underset{\{RNN,W\}}{\operatorname{argmin}} \sum_{i=1}^{N-n} |x_{i+n} - y_i|,$$

$$y_i = RNN(x_i) \cdot W$$

- **Feature extraction**

Take RNN output of each time step:

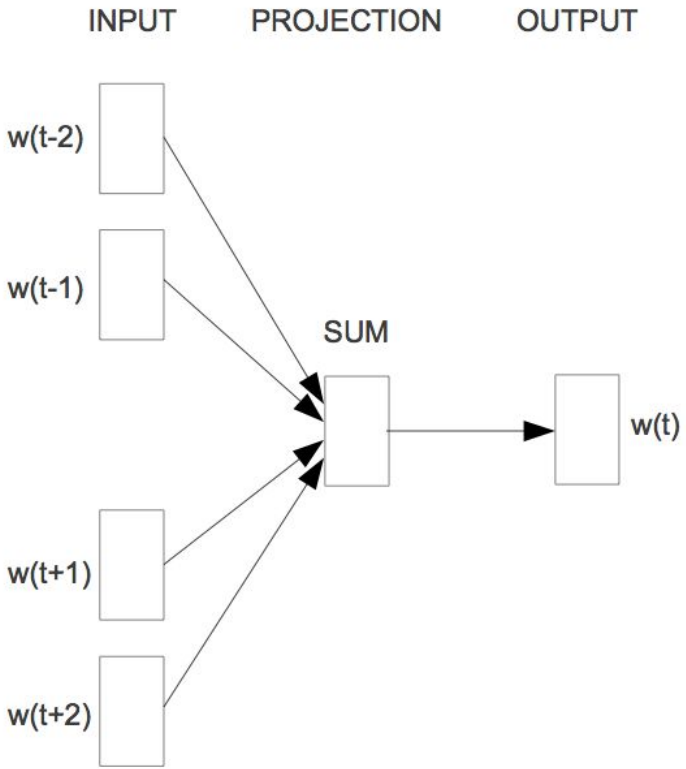
$$T_{APC}(x_i) = RNN(x_i) \; \forall \; i = 1,2, \ldots, N$$

#APC Chung, Y. A., Hsu, W. N., Tang, H., & Glass, J. An Unsupervised Autoregressive Model for Speech Representation Learning. INTERSPEECH 2019.

Outline

1. Unsupervised Learning
2. Self-supervised Learning
3. Predictive Methods
    - **Unmasking**

Unmask: One word given a context



`the cat climbed a tree`

Given context:

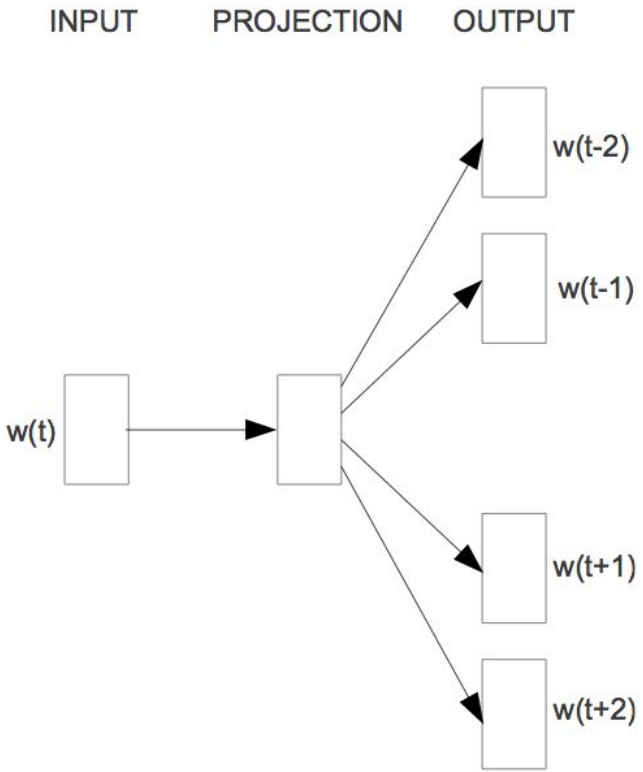
`a, cat, the, tree`

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

Unmask: A context given a word



`the cat climbed a tree`

Given word:

`climbed`

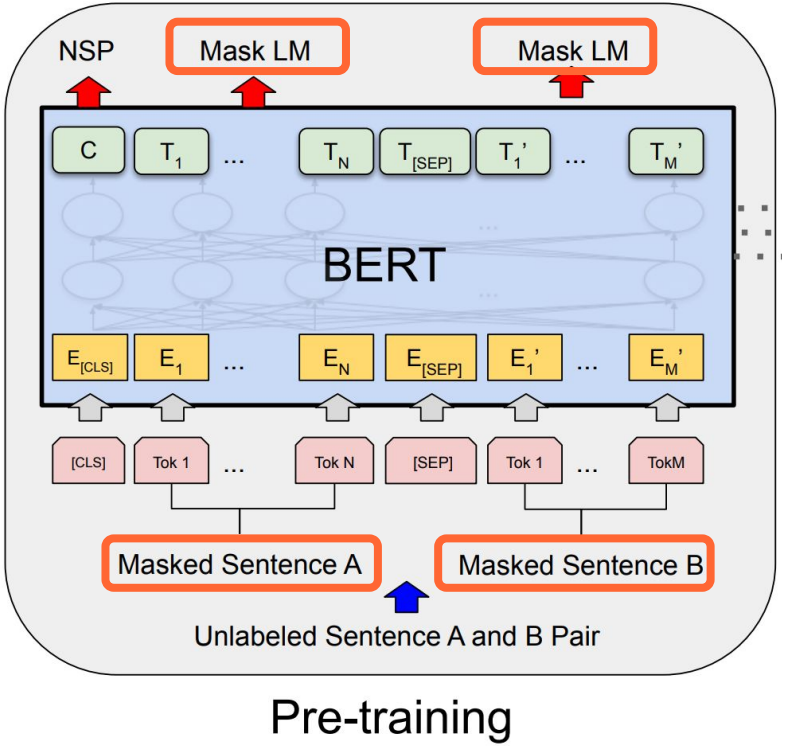
Estimate prob. of context words:

`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

Unmask: Some words given a context

Take 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)

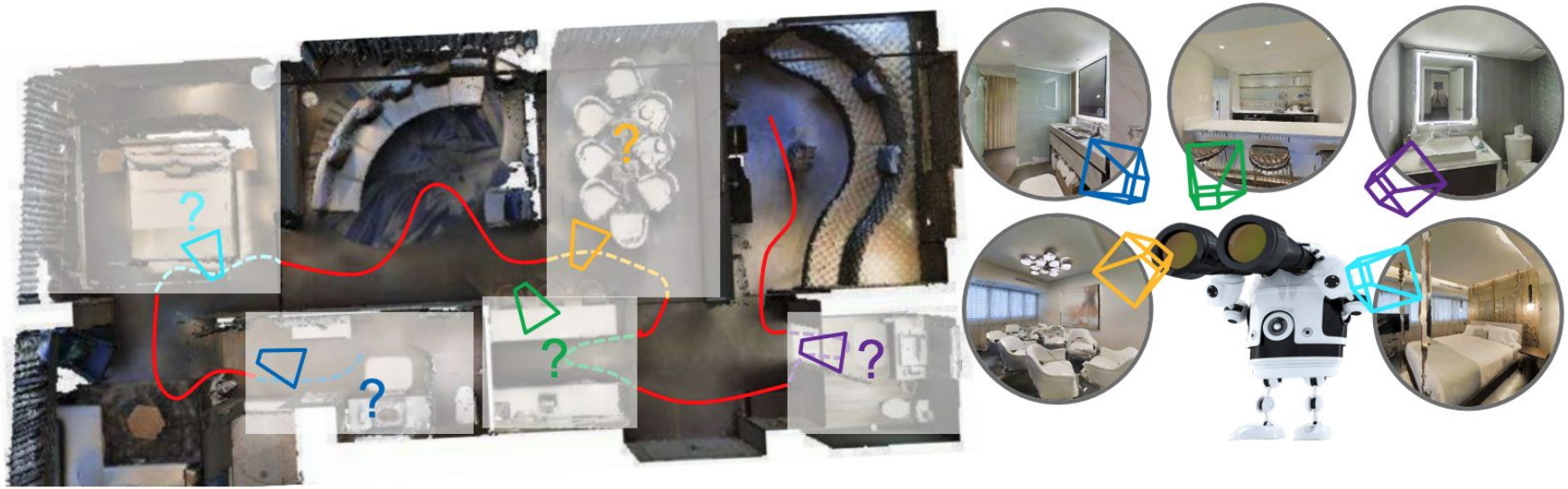


Pre-training

#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)**.

Unmask: Embodied agent trajectories

Mask out portions of an agent's trajectory and predict them from the unmasked portions, conditioned on the agent's camera poses.



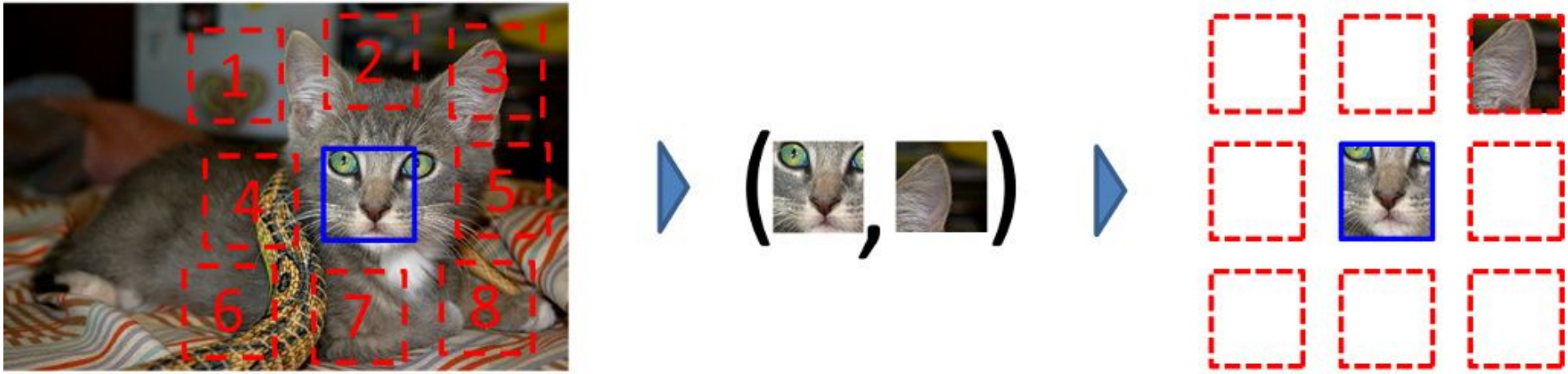
#EPC Ramakrishnan, S. K., Nagarajan, T., Al-Halah, Z., & Grauman, K. (2021). Environment Predictive Coding for Embodied Agents. arXiv preprint arXiv:2102.02337.aa

Outline

1. Transfer Learning
2. Unsupervised Learning
3. Self-supervised Learning
4. Predictive Methods
    - **Spatial relations**

Spatial verification

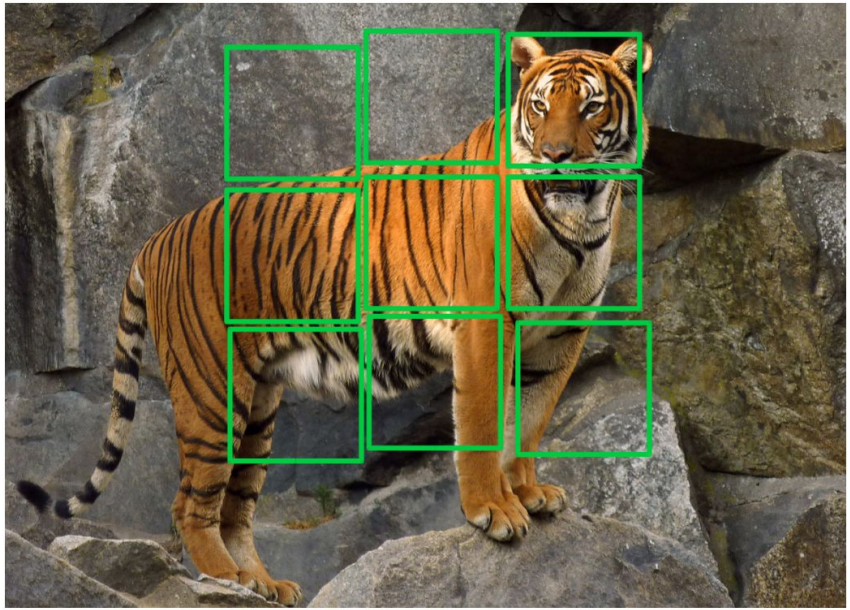
Predict the relative position between two image patches.

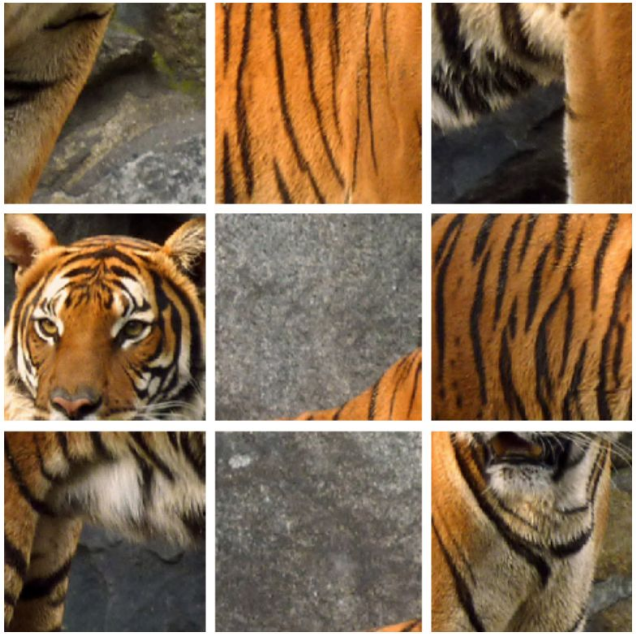


#RelPos Doersch, Carl, Abhinav Gupta, and Alexei A. Efros. "Unsupervised visual representation learning by context prediction." ICCV 2015.

Spatial verification

Train a neural network to solve a jigsaw puzzle.

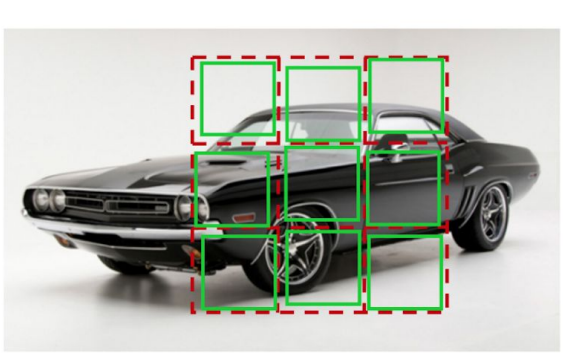




#Jigsaw Noroozi, M., & Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. ECCV 2016.

Spatial verification

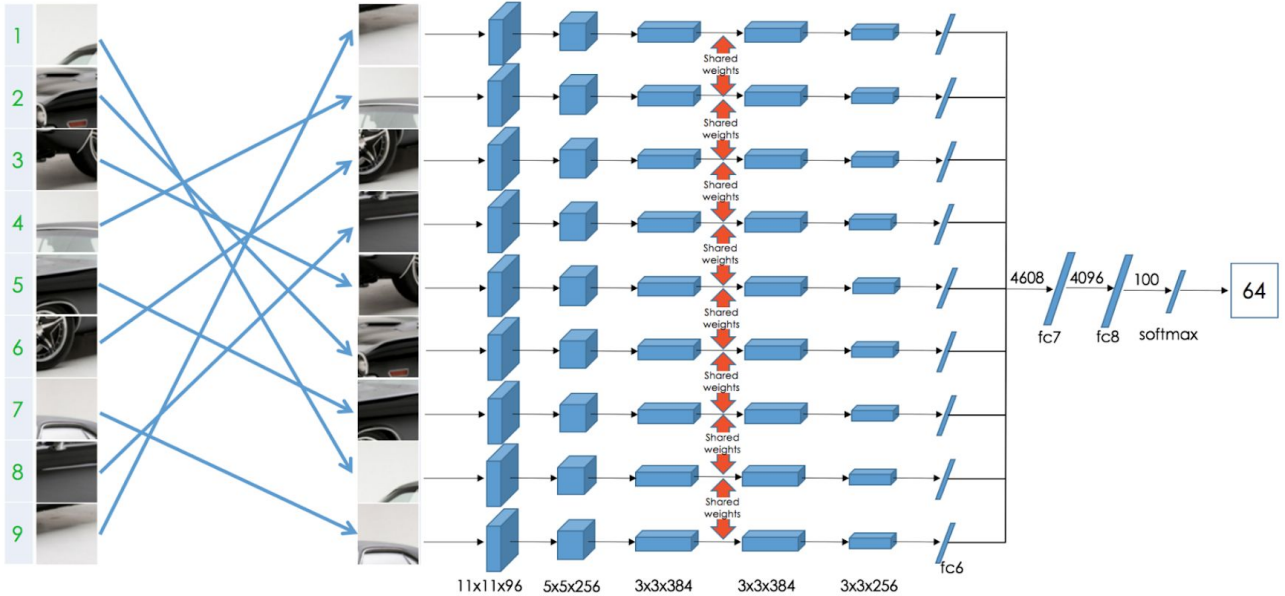
Train a neural network to solve a jigsaw puzzle.



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.

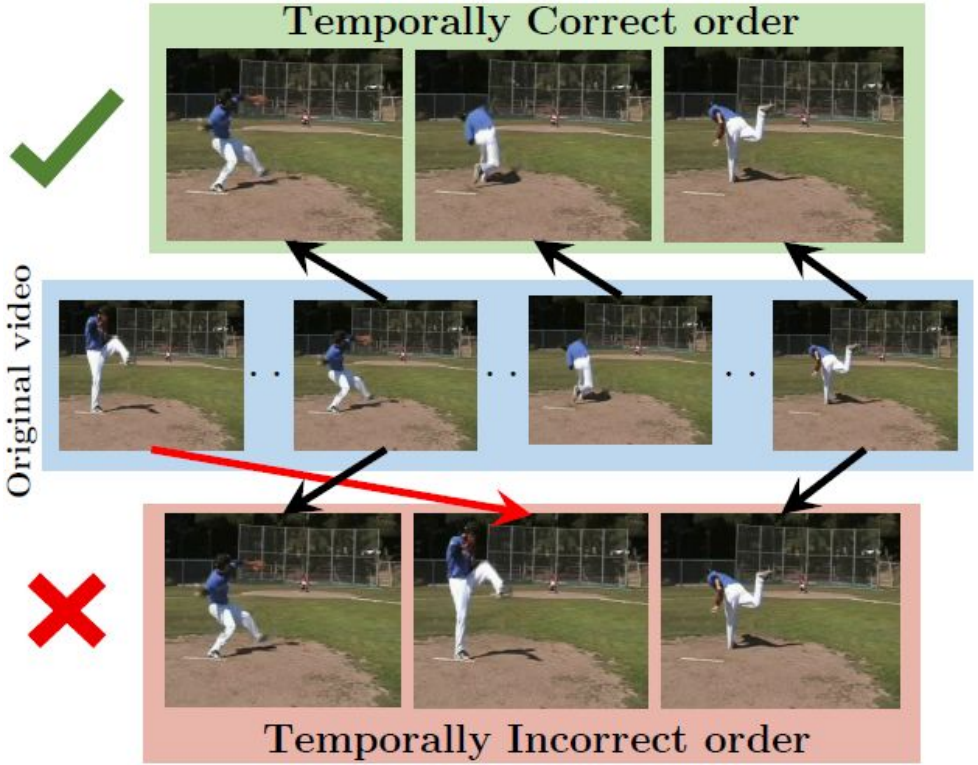
Outline

1. Transfer Learning
2. Unsupervised Learning
3. Self-supervised Learning
4. Predictive Methods
    - **Temporal relations**

Could you think about self-supervised learning tasks exploiting temporal relations ?

Temporal coherence

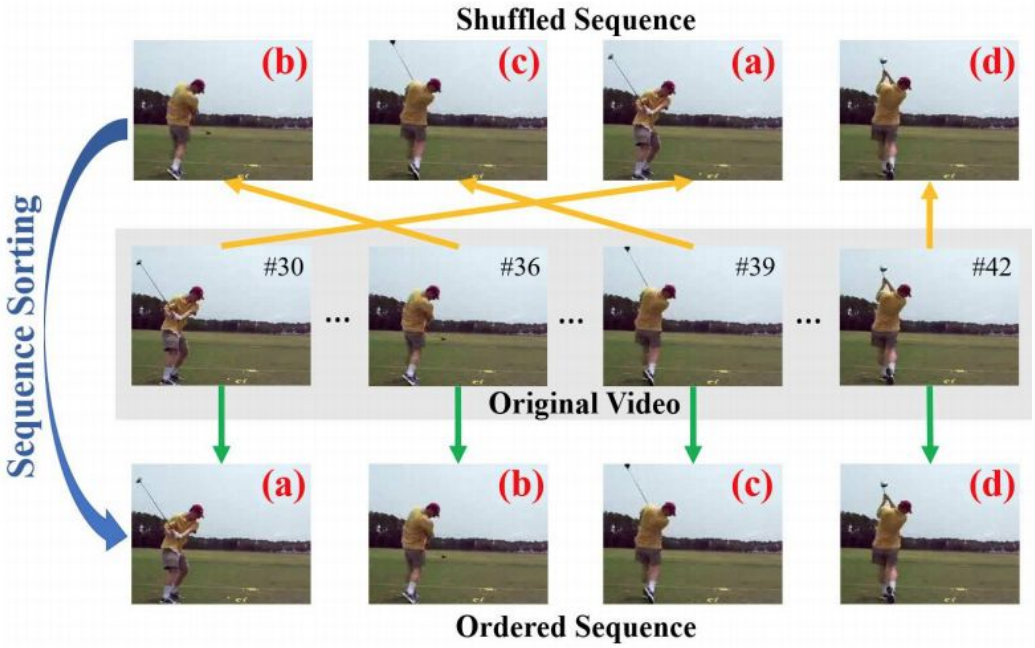
Temporal order of frames is exploited as the supervisory signal for learning.



#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 sorting

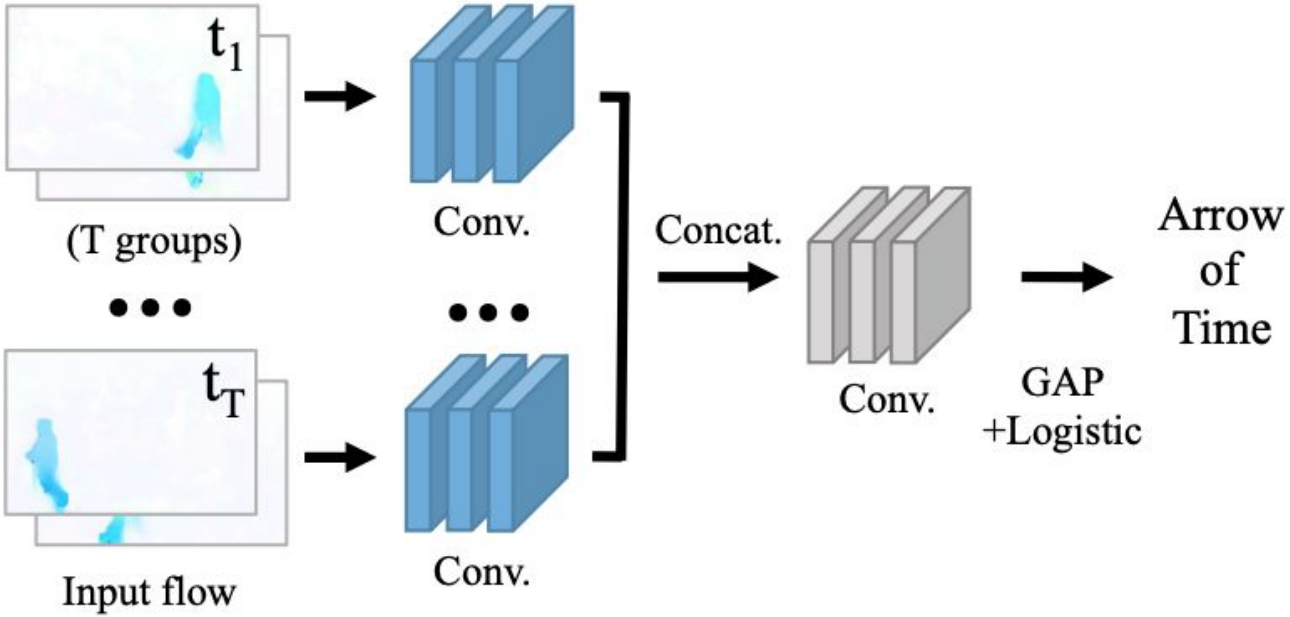
Sort the sequence of frames.



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.



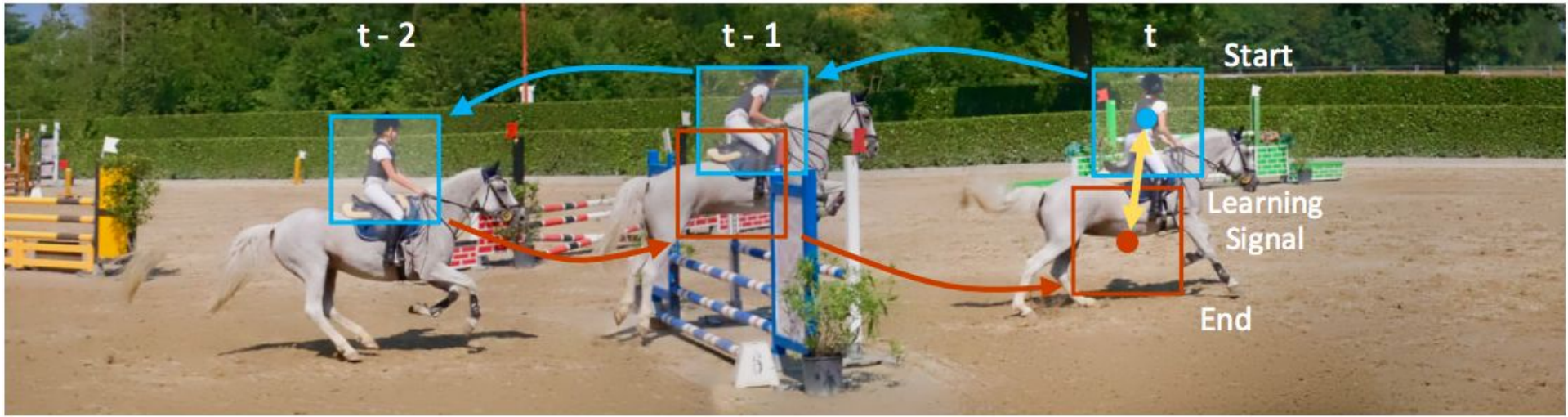
#T-CAM Wei, Donglai, Joseph J. Lim, Andrew Zisserman, and William T. Freeman. "Learning and using the arrow of time." CVPR 2018.

Outline

1. Unsupervised Learning
2. Self-supervised Learning
3. Predictive Methods
    - **Cycle Consistency**

Cycle-consistency

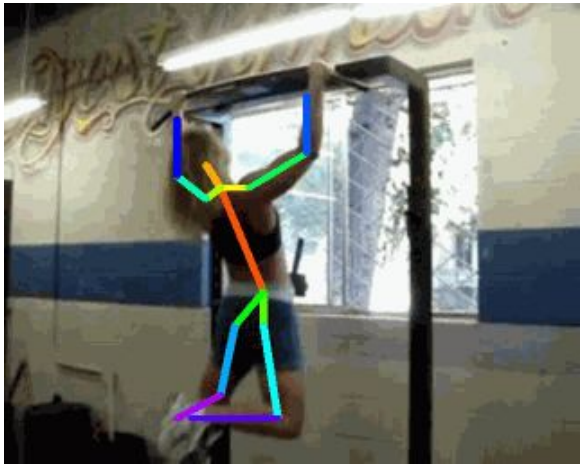
Pre-text task: Tracking backward and then forward with pixel embeddings, to later train a NN to match the pairs of associated pixel embeddings.



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

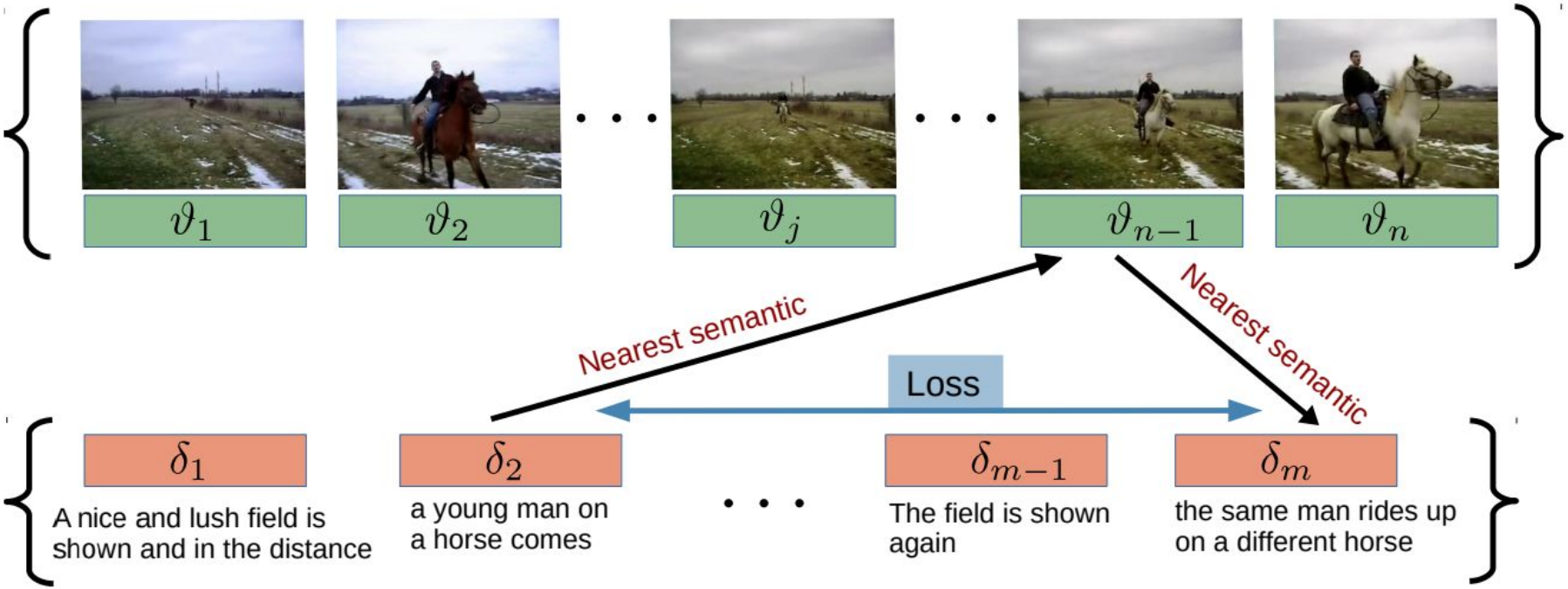
Cycle-consistency

Downstream task: Track pixels across video sequences.



#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



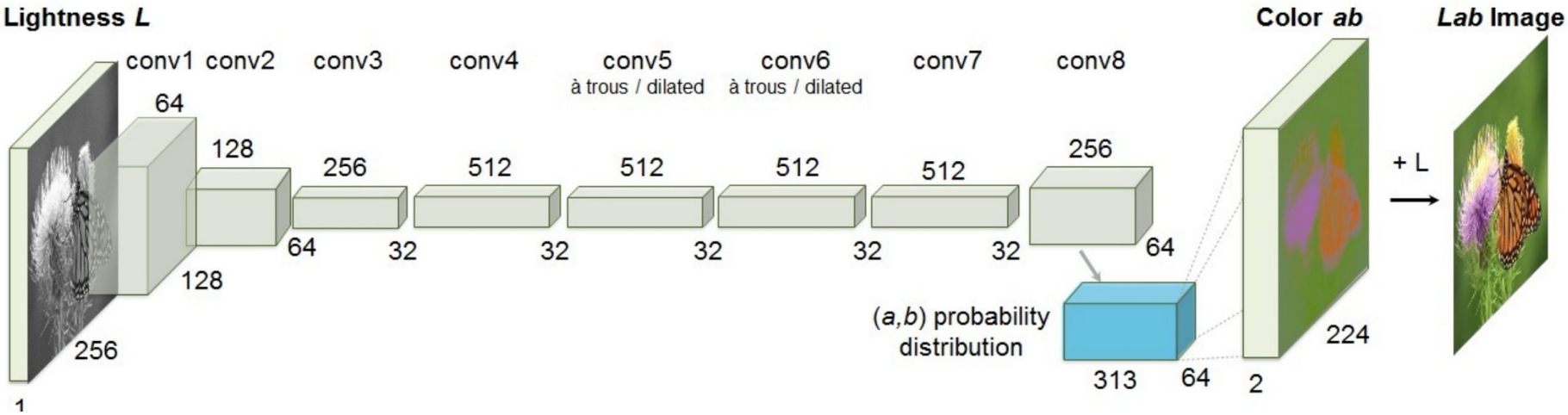
Ging, S., Zolfaghari, M., Pirsiavash, H., & Brox, T. (2020). Coot: Cooperative hierarchical transformer for video-text representation learning. arXiv preprint arXiv:2011.00597.

Outline

1. Unsupervised Learning
2. Self-supervised Learning
3. Predictive Methods
    - **Colorization**

Colorization (still image)

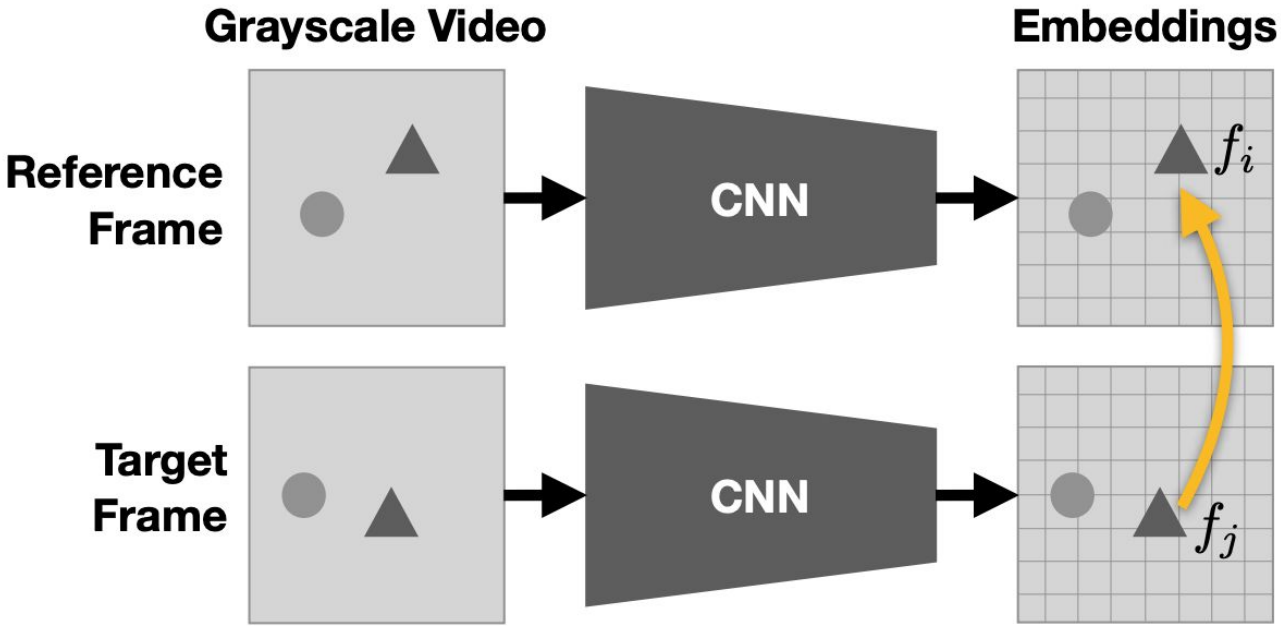
Surrogate task: Colorize a gray scale image.



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

Pixel tracking

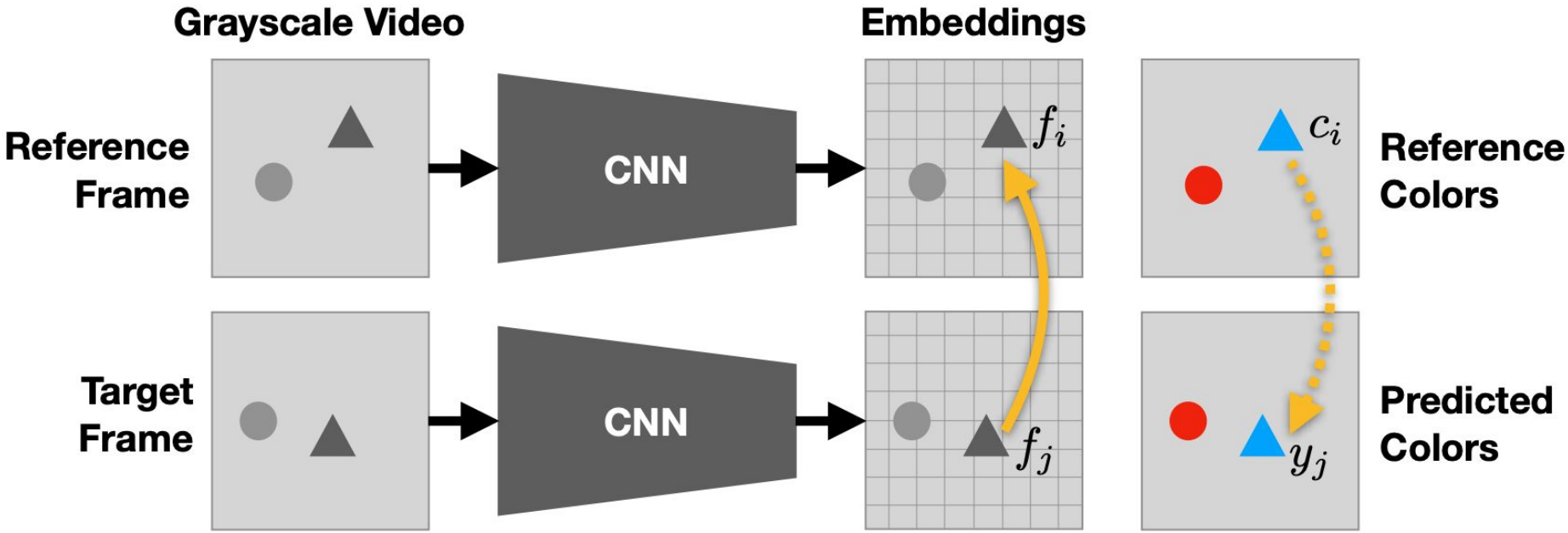
Pre-training: A NN is trained to point the location in a reference frame with the most similar color...



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

Pixel tracking

Pre-training: A NN is trained to point the location in a reference frame with the most similar color… computing a loss with respect to the actual color.



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

Pixel tracking

For each grayscale pixel to colorize, find the color of most similar pixel embedding (representation) from the first frame.

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

Pixel tracking

Application: colorization from a reference frame.

Reference frame

Grayscale video

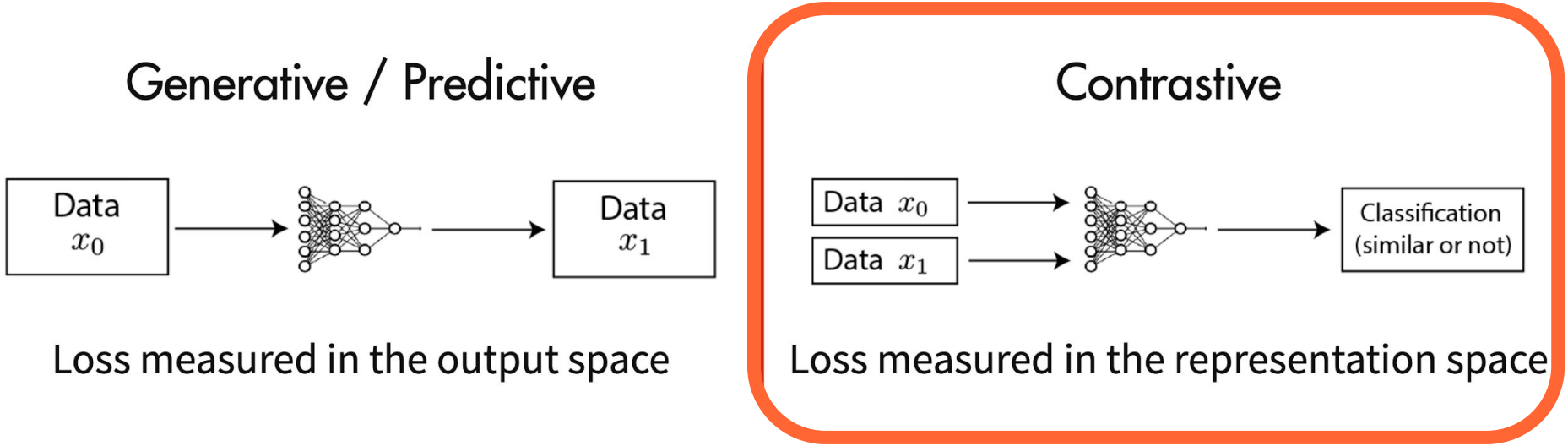
Colorized video

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

Outline

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

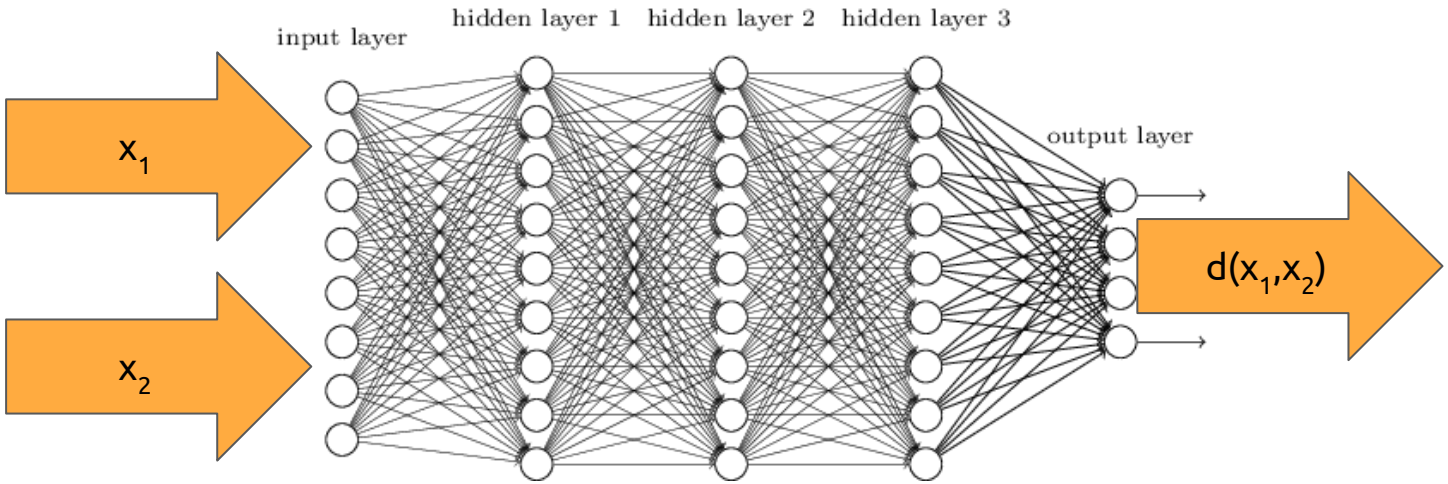
Predictive vs Contrastive Methods



Source: Ankesh Anand, "Contrastive Self-Supervised Learning" (2020)"

Ranking Losses

Ranking losses aim at predicting the relative distance (or similarity) between samples.



Hinge or Margin Losses

Only pairs (x1,x2) closer than a **margin m** contribute to the penalty.

$$l_N(d_D(\mathbf{x}_1, \mathbf{x}_2)) = \max(0, m - d_D(\mathbf{x}_1, \mathbf{x}_2))$$

Hinge loss

What is the use of the Hinge Loss ?

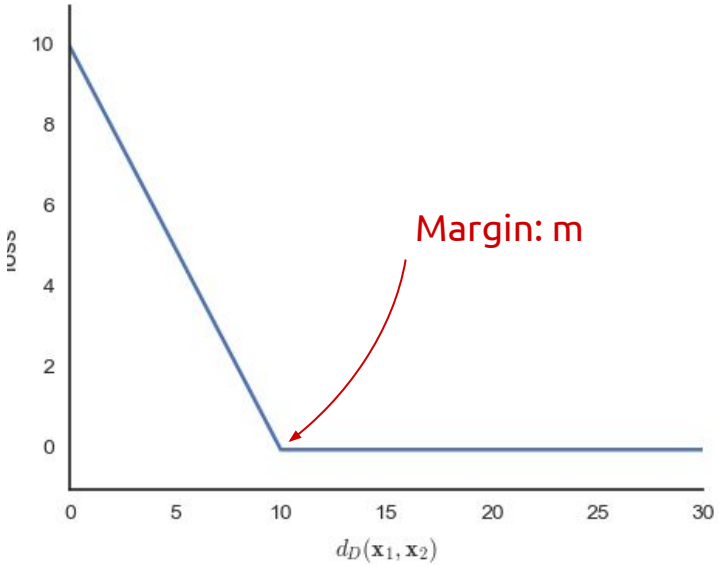
Hinge or Margin Losses

Only pairs (x1,x2) closer than a **margin m** contribute to the penalty.

$$l_N(d_D(\mathbf{x}_1, \mathbf{x}_2)) = \max(0, m - d_D(\mathbf{x}_1, \mathbf{x}_2))$$

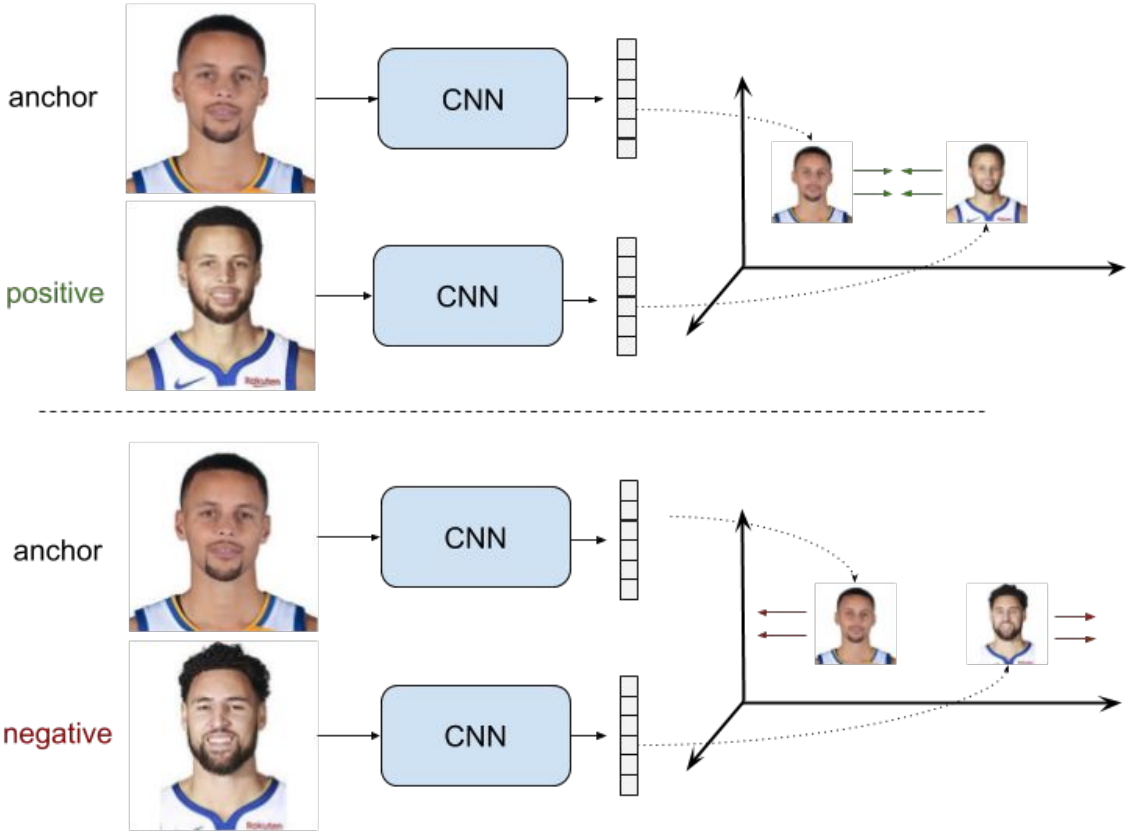
Hinge loss

A hinge



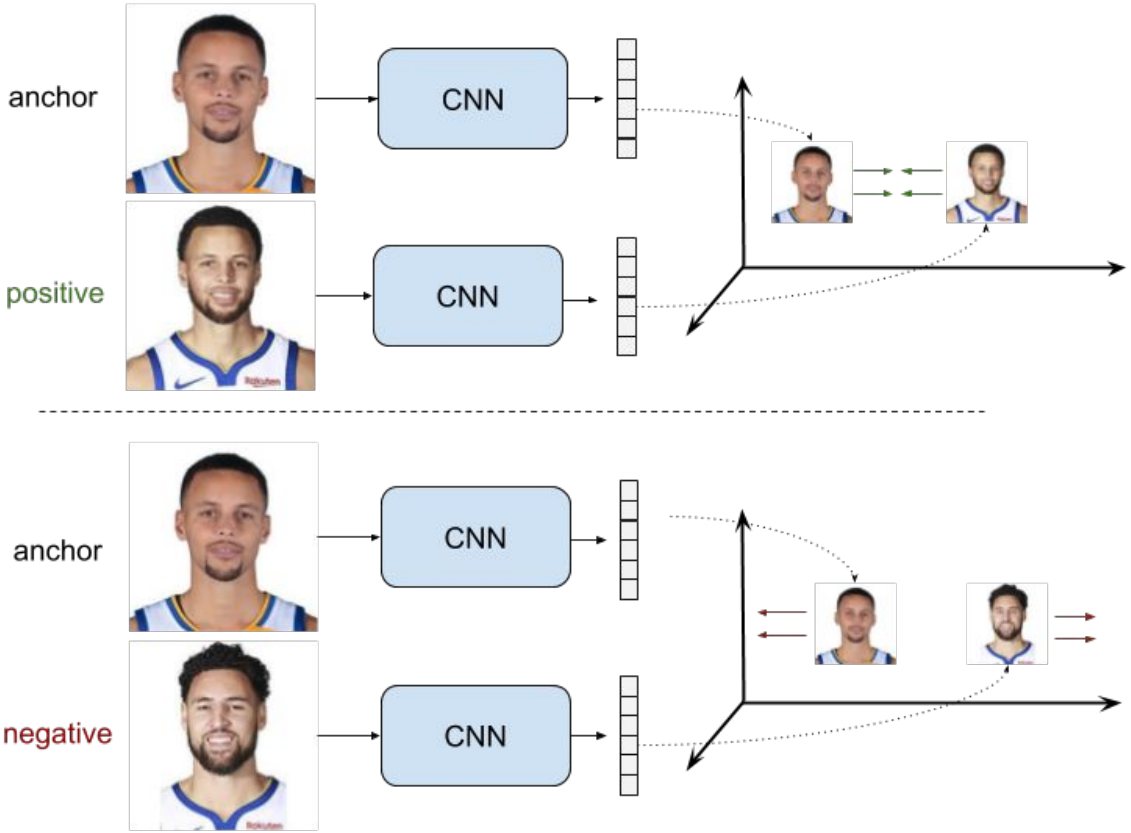
Intuition: there is nothing new to learn from those **easy negative** pairs which are already far enough.

Pairwise Ranking Loss



Source: Raul Gómez, "Understanding Ranking Loss, Contrastive Loss, Margin Loss, Triplet Loss, Hinge Loss and all those confusing names" (2019)

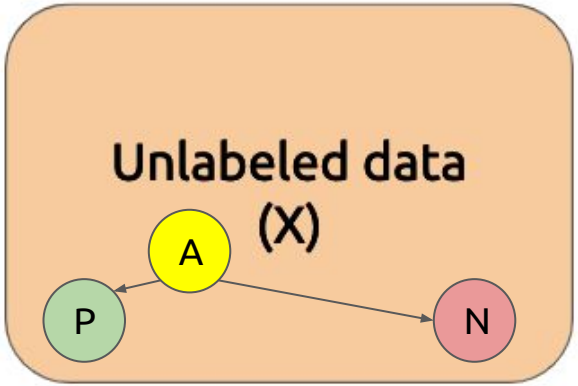
Pairwise Ranking Loss = Contrastive Loss

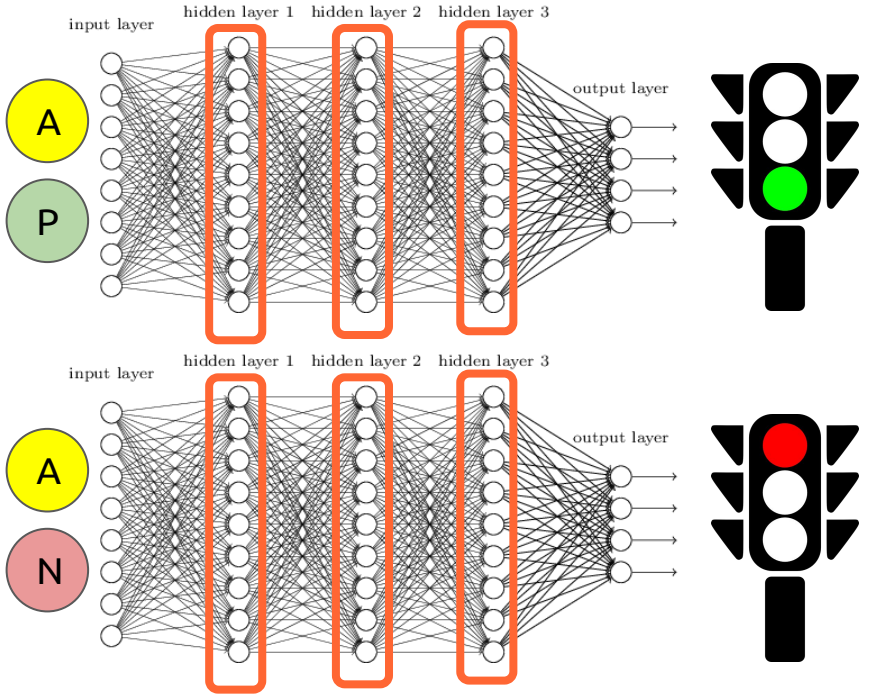


Source: Raul Gómez, "Understanding Ranking Loss, Contrastive Loss, Margin Loss, Triplet Loss, Hinge Loss and all those confusing names" (2019)

Pairwise Ranking Loss = Contrastive Loss

Contrastive Self-Supervised Learning Methods:
The **pretext (or surrogate) task** corresponds to predicting which pairs of samples are similar, focusing the design in the definition of "similar".

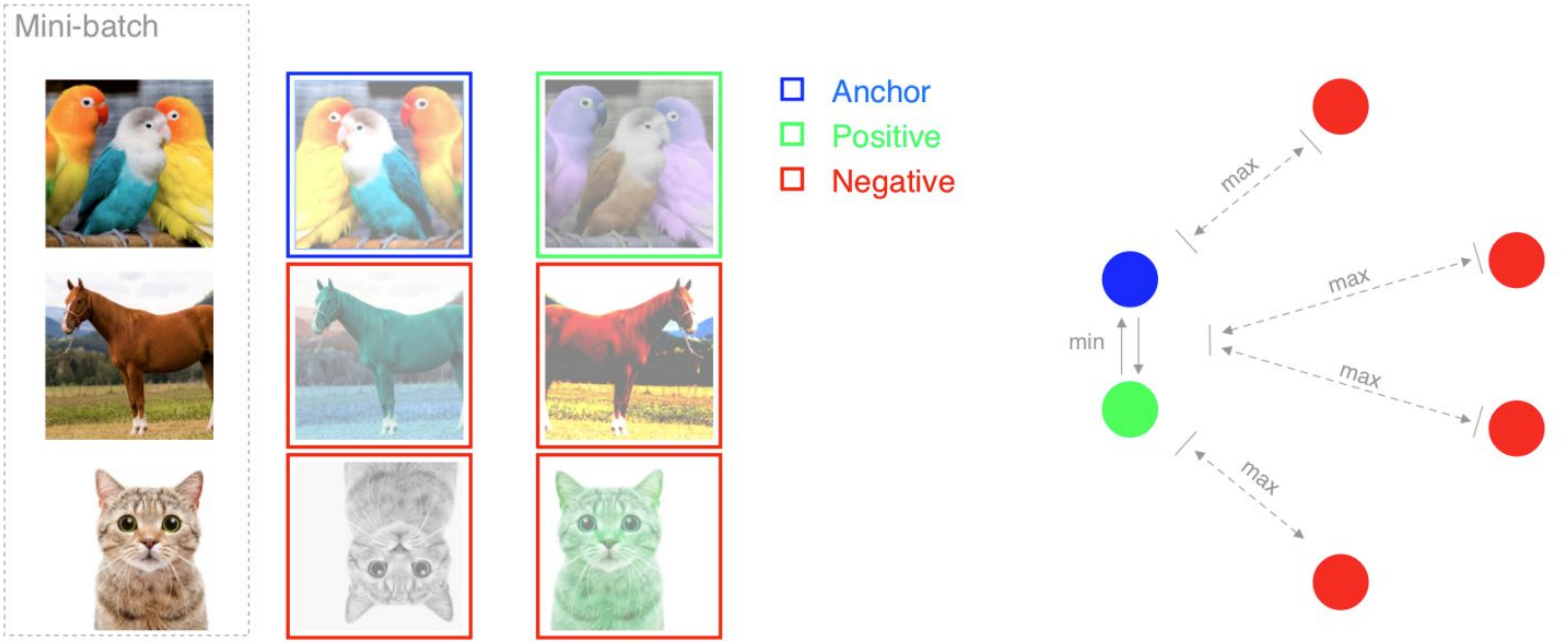




Reference: Andrew Zisserman (PAISS 2018)

Pairwise Ranking Loss = Contrastive Loss

Typically, pairs of positive and negative pairs are augmented in each training mini-batch.



Falcon, W., & Cho, K. (2020). A Framework For Contrastive Self-Supervised Learning And Designing A New Approach. arXiv preprint arXiv:2009.00104.

Contrastive Loss

Learn a NN $G_w(X)$ from pairs (X_1, X_2) which are similar (Y=0) or dissimilar (Y=1).

$$D_W(\vec{X}_1, \vec{X}_2) = \|G_W(\vec{X}_1) - G_W(\vec{X}_2)\|_2$$

Should the distance D_w be high or low when Y=0 ?

When Y=0, G_w will learn to minimize D_w, for example, expressed as:

$$(1 - Y)\frac{1}{2}(D_W)^2$$

Hadsell, Raia, Sumit Chopra, and Yann LeCun. "Dimensionality reduction by learning an invariant mapping." CVPR 2006.

Contrastive Loss

Learn a NN $G_w(X)$ from pairs (X_1, X_2) which are similar (Y=0) or dissimilar (Y=1).

$$D_W(\vec{X}_1, \vec{X}_2) = \|G_W(\vec{X}_1) - G_W(\vec{X}_2)\|_2$$

Should the distance D_w be high or low when Y=1 ?

When Y=1, G_w will learn to maximize D_w, for example, expressed as:

$$(Y)\frac{1}{2}\{max(0, m - D_W)$$

Hadsell, Raia, Sumit Chopra, and Yann LeCun. "Dimensionality reduction by learning an invariant mapping." CVPR 2006.

Contrastive Loss

The contrastive loss to minimize can be expressed as:

$$L(W, Y, \vec{X}_1, \vec{X}_2) = (1 - Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}\{max(0, m - D_W)\}^2$$

Y=0, similar pair (X_1,X_2)

Y=1, dissimilar pair (X_1,X_2)

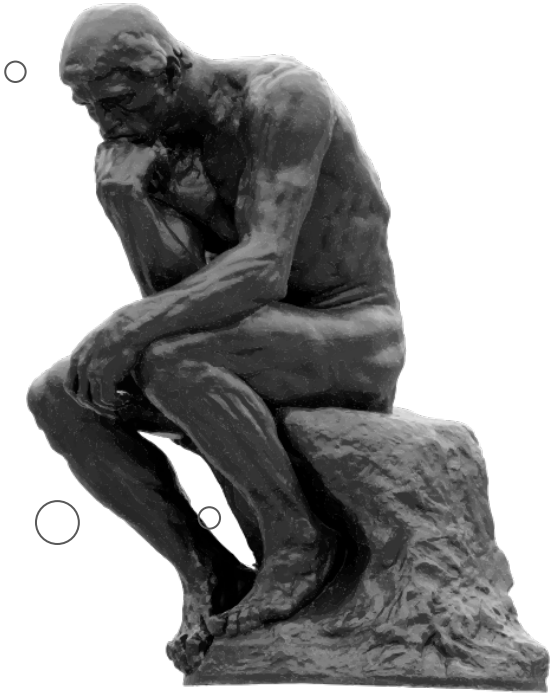
Hadsell, Raia, Sumit Chopra, and Yann LeCun. "Dimensionality reduction by learning an invariant mapping." CVPR 2006.

Contrastive Loss

Learn a NN $G_w(X)$ with pairs (X_1,X_2) which are similar (Y=0) or dissimilar (Y=1).

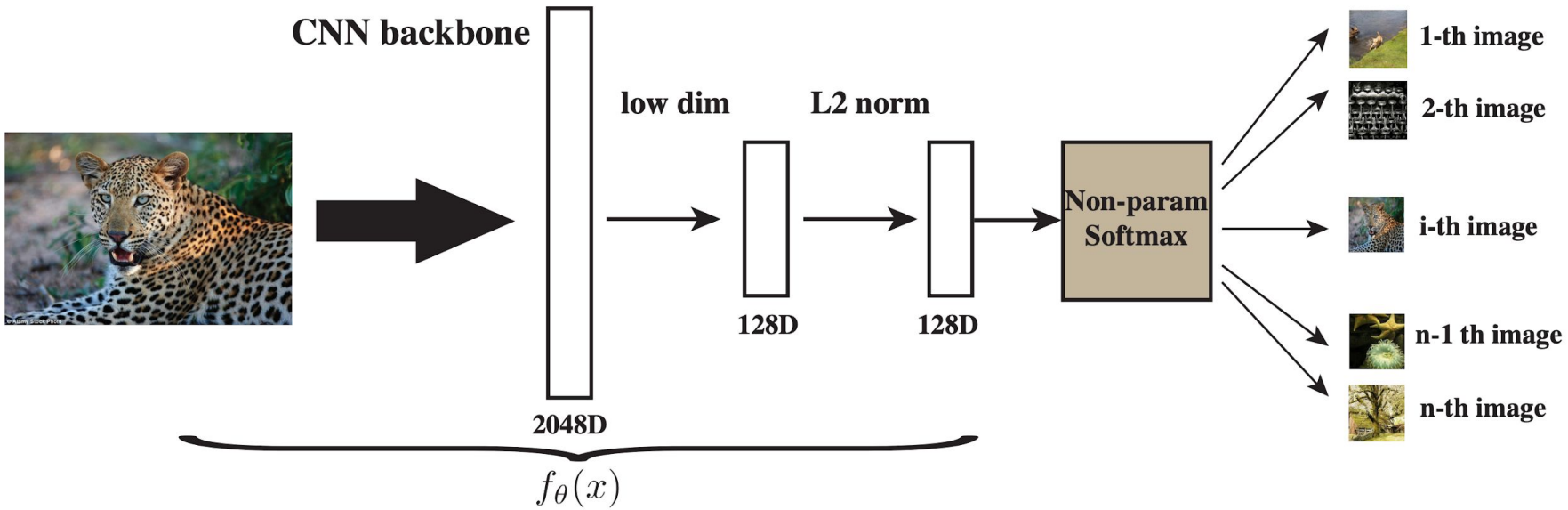
What are pairs (X_1,X_2) ?

How to select similar and dissimilar pairs (X_1,X_2) ?



Momentum Contrast (MoCo)

Pretext task: Treat each image instance as a distinct class of its own and train a classifier to distinguish whether **two crops (views)** belong to the same image.



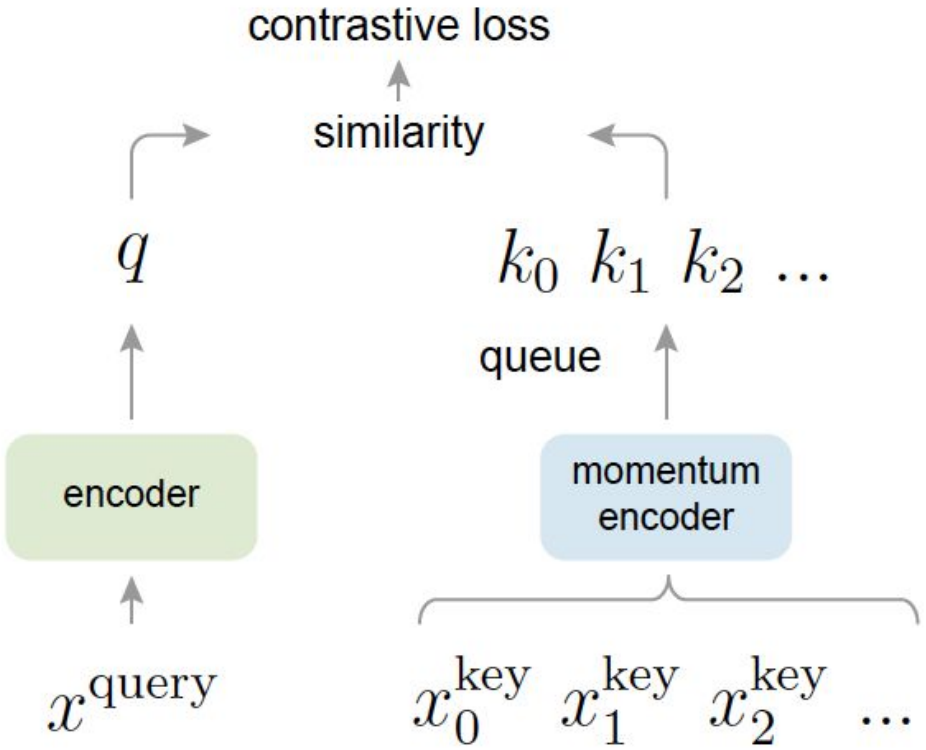
Wu, Z., Xiong, Y., Yu, S. X., & Lin, D. Unsupervised feature learning via non-parametric instance discrimination. CVPR 2018.

Momentum Contrast (MoCo)

Learn a contrastive loss by building training batches between:

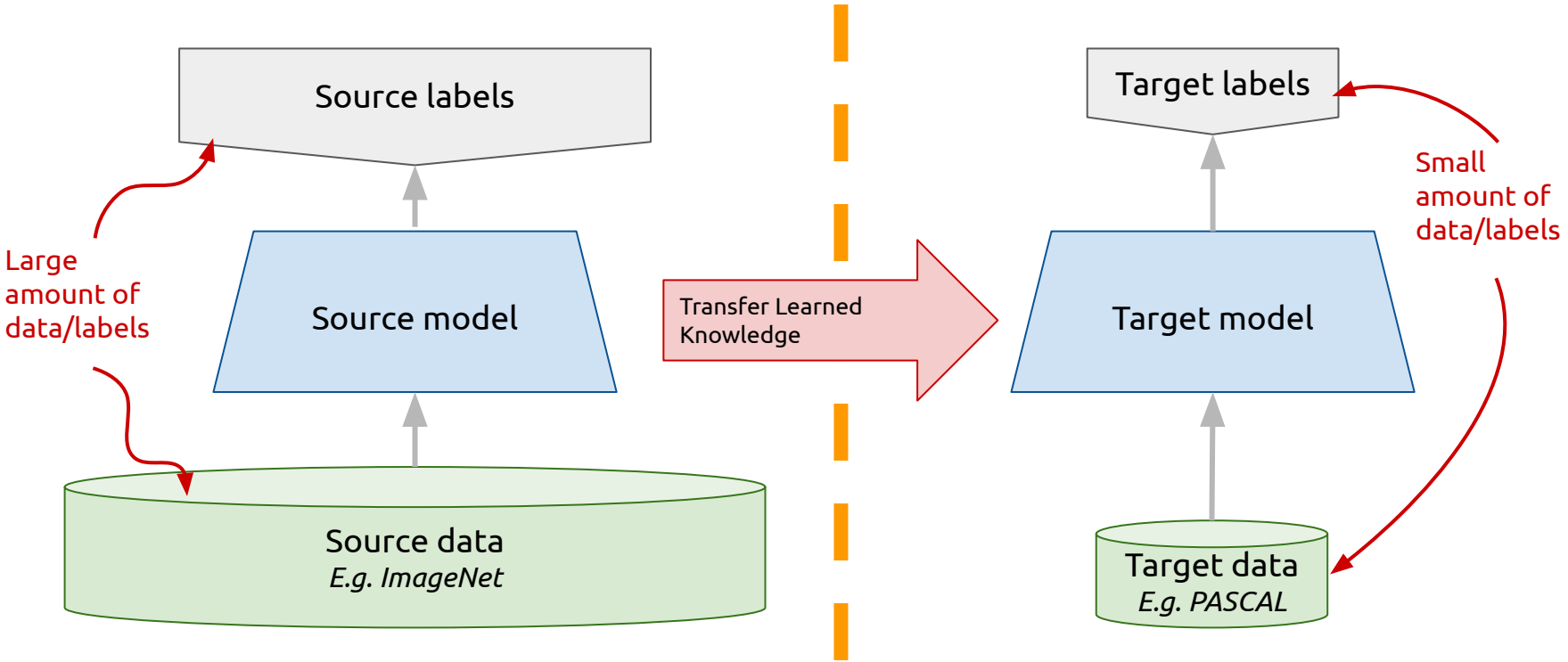
- Encoded features of a query image.
- A running queue of encoded keys.

Losses over representations (instead of pixels) facilitate abstraction.



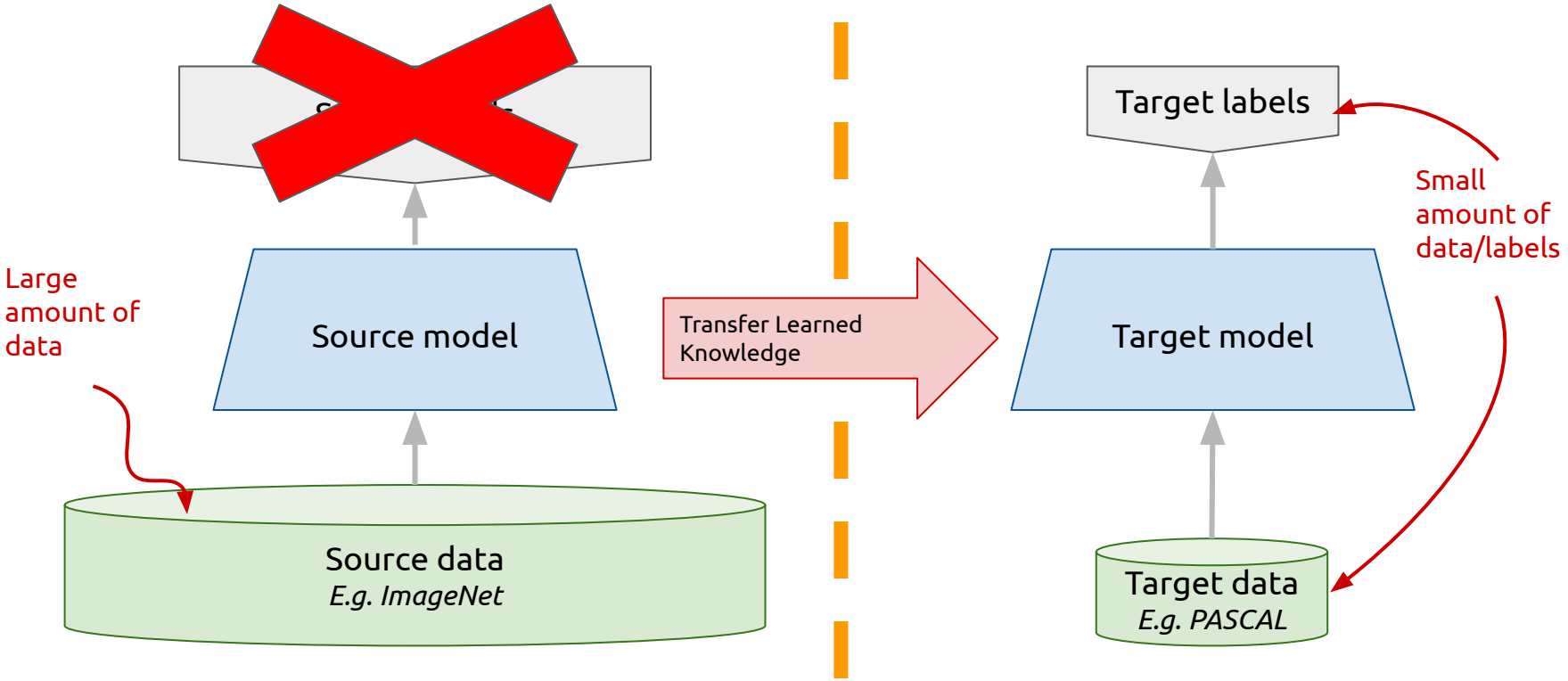
#MoCo He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. . Momentum Contrast for Unsupervised Visual Representation Learning. CVPR 2020. [code]

Transfer Learning from ImageNet labels (super. IM-1M)



Slide credit: Kevin McGuinness

Transfer Learning without ImageNet labels (MoCo IN-1M)



Slide concept: Kevin McGuinness

Fine-tune task (super. IN-1M vs MoCo IN-1M)

Transferring to Pascal VOC object detection with Faster R-CNN model: Unsupervised MoCo training outperformed supervised ImageNet pre-training.

pre-train	AP_{50}	AP	AP_{75}
random init.	60.2	33.8	33.1
super. IN-1M	81.3	53.5	58.8
MoCo IN-1M	81.5 (+0.2)	55.9 **(+2.4)**	62.6 **(+3.8)**

#MoCo He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. . Momentum Contrast for Unsupervised Visual Representation Learning. CVPR 2020. [code]

The Gelato Bet (Sept 23, 2014)

*"If, by the first day of autumn (Sept 23) of 2015, a method will exist that can match or beat the performance of R-CNN on Pascal VOC detection, without the use of any extra, human annotations (e.g. ImageNet) as pre-training, **Mr. Malik** promises to buy **Mr. Efros** one (1) gelato (2 scoops: one chocolate, one vanilla)."*

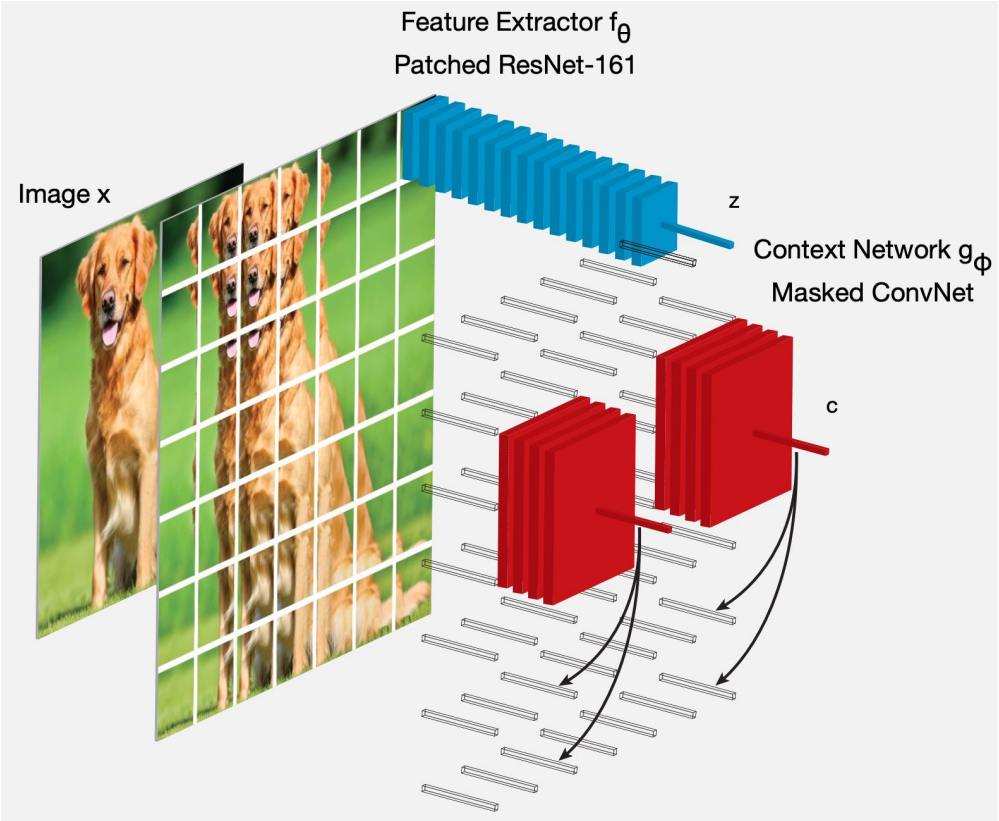
Who won The Gelato Bet ?

Source: Ankesh Anand, "Contrastive Self-Supervised Learning" (2020)"

Prof. Jitendra Malik

Prof. Alexei A. Efros

Spatial Prediction Task



1) Image is divided in patches.
2) Each patch is encoded independently (in blue).
3) Features in the upper half of the images are aggregated with a context network (in red).
4) Rows of context vectors must predict the “blue” vectors below.

#CPCv2 Hénaff, Olivier J., Ali Razavi, Carl Doersch, S. M. Eslami, and Aaron van den Oord. "Data-efficient image recognition with contrastive predictive coding." arXiv preprint arXiv:1905.09272 (2019).

Correlated views by data augmentation

(a) Original

(b) Crop and resize

(c) Crop, resize (and flip)

(d) Color distort. (drop)

(e) Color distort. (jitter)

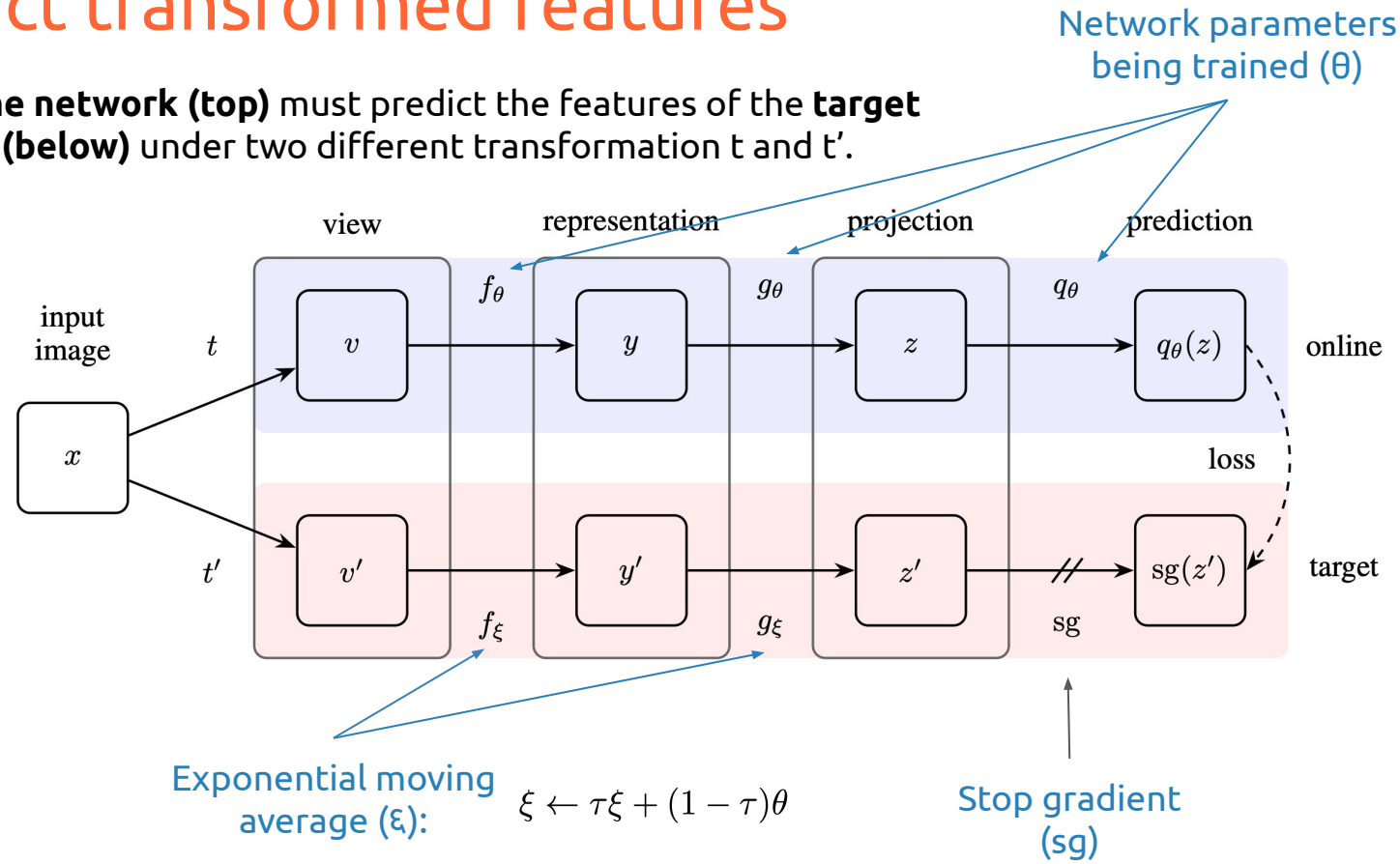
(i) Gaussian blur

Original image cc-by: Von.grzanka

#SimCLR Chen, Ting, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. "A simple framework for contrastive learning of visual representations." ICML 2020. [tweet]

Predict transformed features

The **online network (top)** must predict the features of the **target network (below)** under two different transformation t and t'.



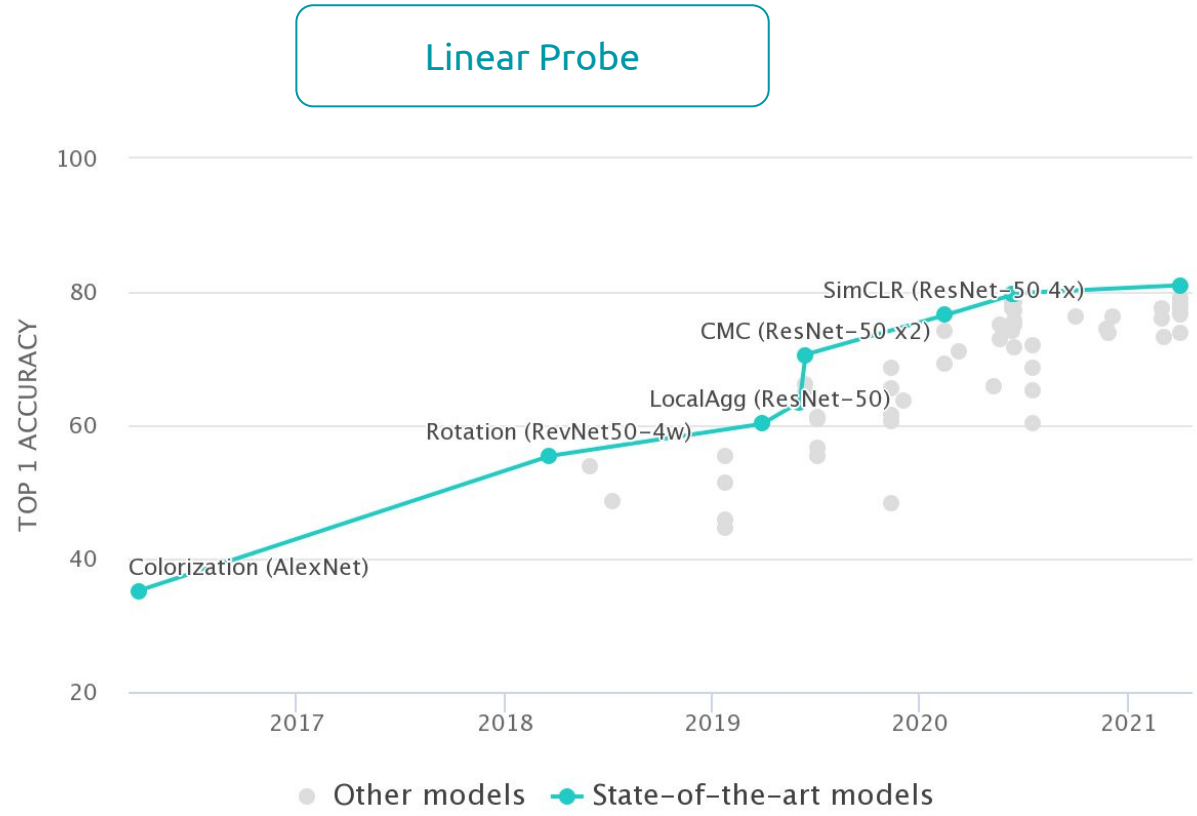
#BYOL Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., ... & Piot, B. (2020). Bootstrap your own latent: A new approach to self-supervised learning. arXiv preprint arXiv:2006.07733. [tweet] [review]

Cluster features

Patches from different views (transformations) of a source image are independently assigned to clusters of images. The cluster assignments must be consistent over time.
Pretext task: predicting the cluster of one version of the image with the other version:

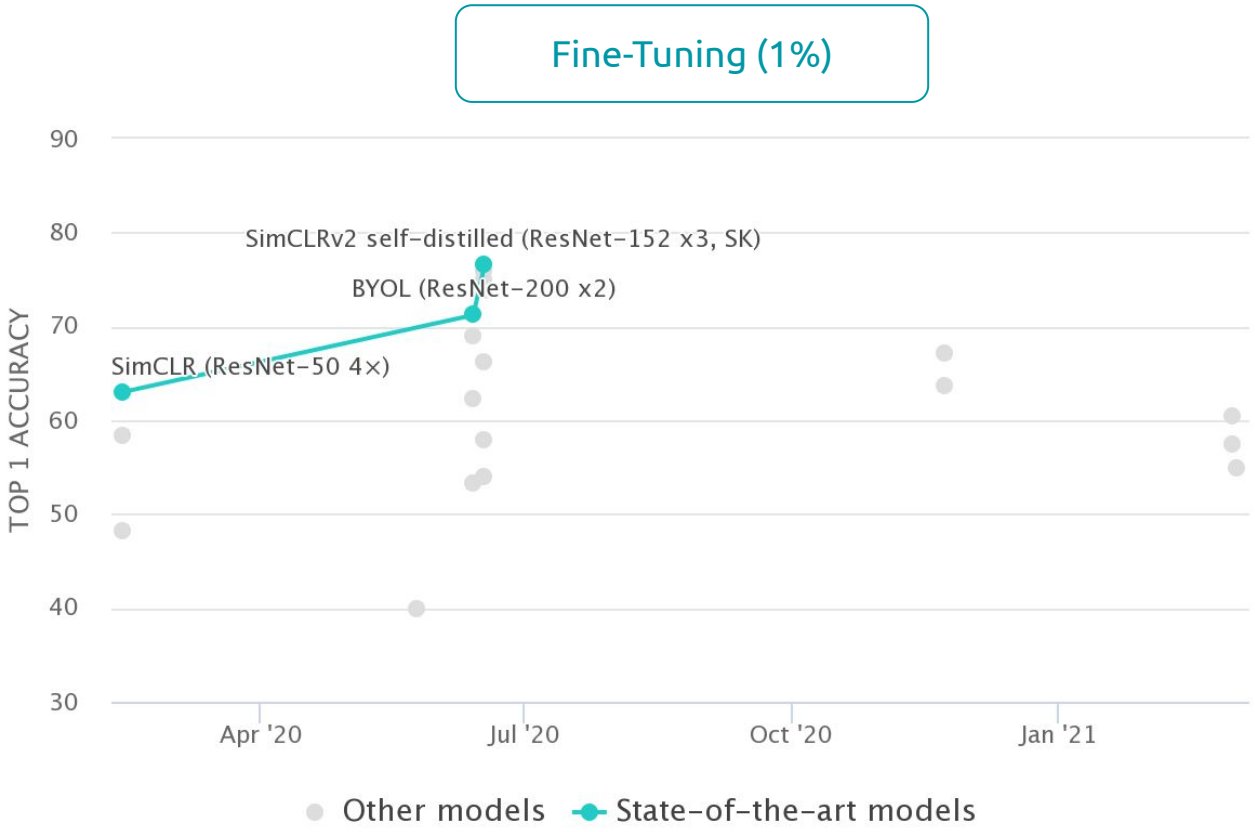
#SWAV Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. arXiv preprint arXiv:2006.09882. [code] [tweet] [video]

State of the Art



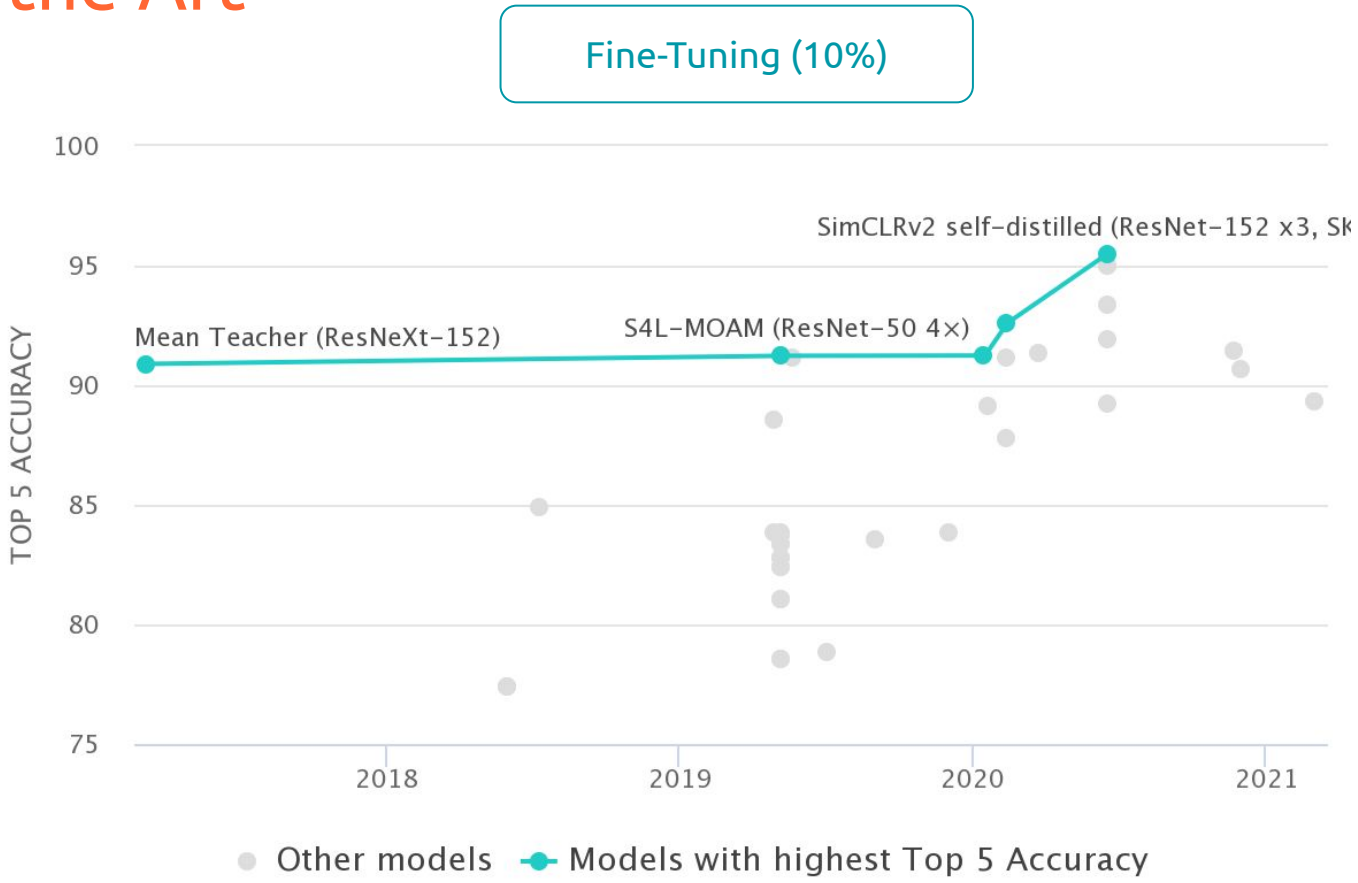
Source: [Papers with Code]

State of the Art



Source: [Papers with Code]

State of the Art



Source: [Papers with Code]

Triplet Ranking Loss (reminder)

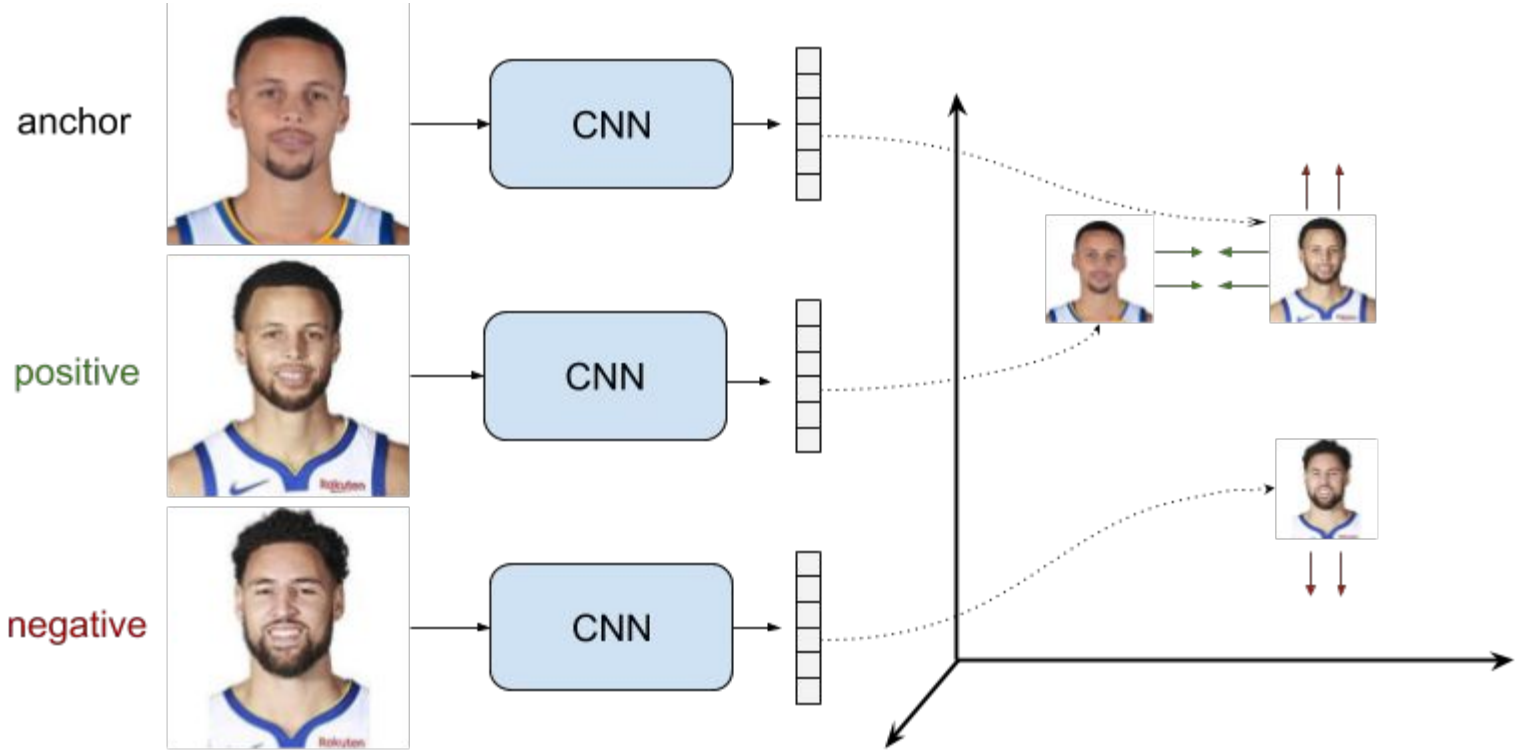


Figure: Raul Gómez, "Understanding Ranking Loss, Contrastive Loss, Margin Loss, Triplet Loss, Hinge Loss and all those confusing names" (2019)

Schultz, M., & Joachims, T. (2004). Learning a distance metric from relative comparisons. NeurIPS 2003.

Triplet Ranking Loss = Contrastive Loss (too)

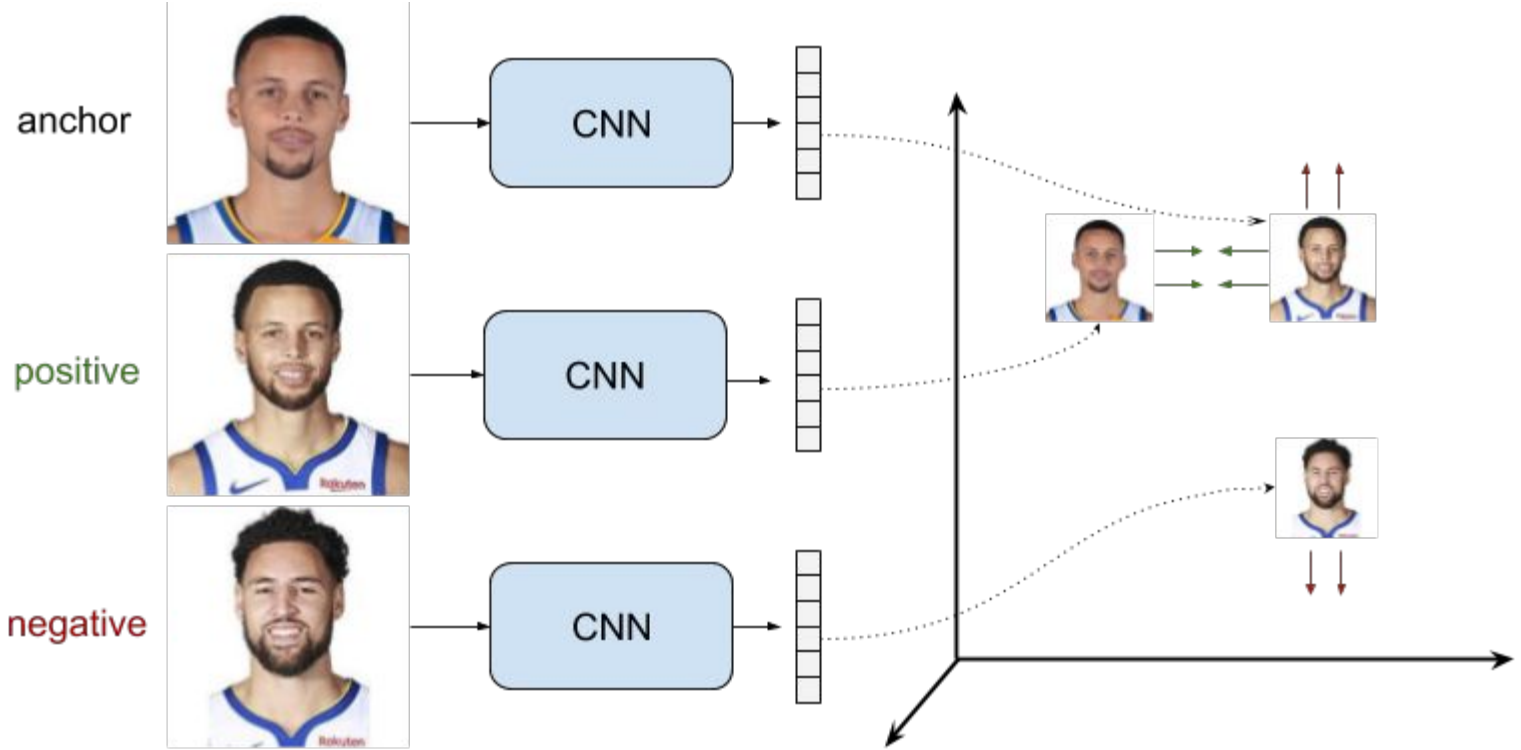


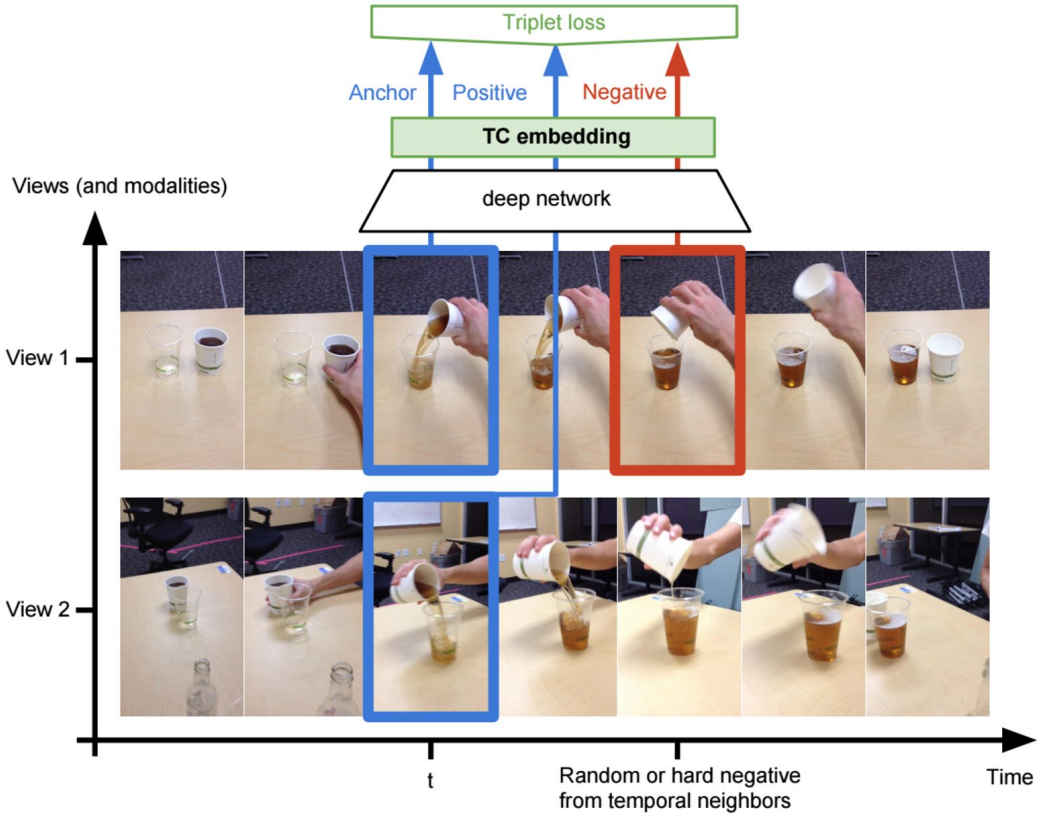
Figure: Raul Gómez, "Understanding Ranking Loss, Contrastive Loss, Margin Loss, Triplet Loss, Hinge Loss and all those confusing names" (2019)

Schultz, M., & Joachims, T. (2004). Learning a distance metric from relative comparisons. NeurIPS 2003.

Multi-camera + Temporal ordering

Pre-Training: Learn features with a triplet loss by:

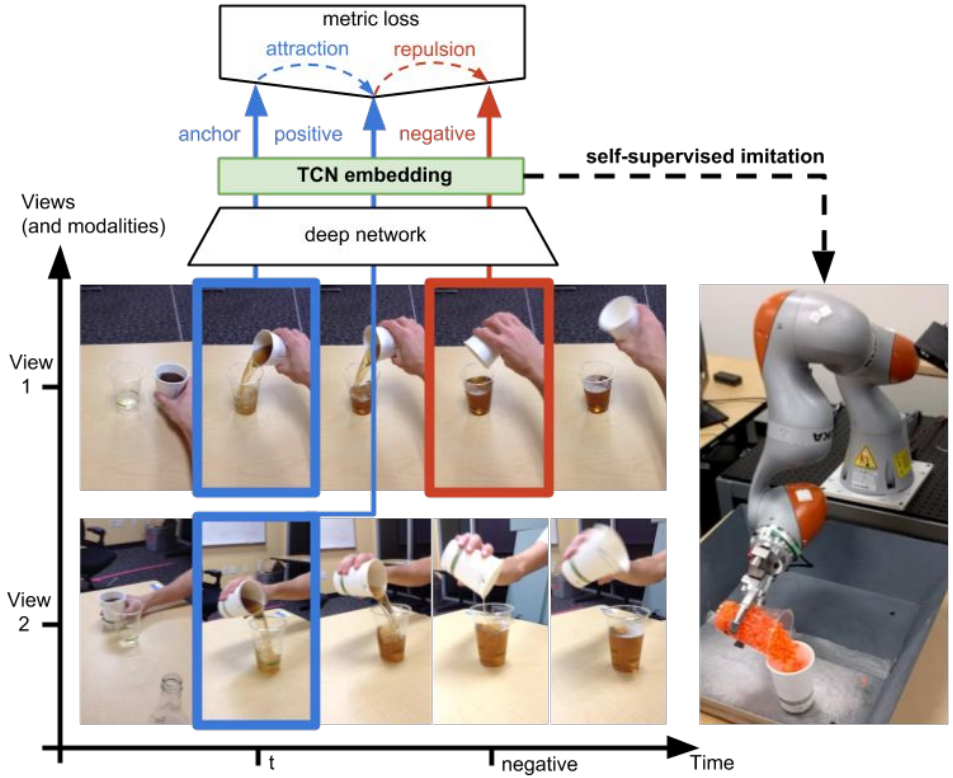
- Positive: Same time-stamp from different cameras, to achieve viewpoint invariance.
- Negative: Other frames from the same camera, to be sensitive to temporal ordering



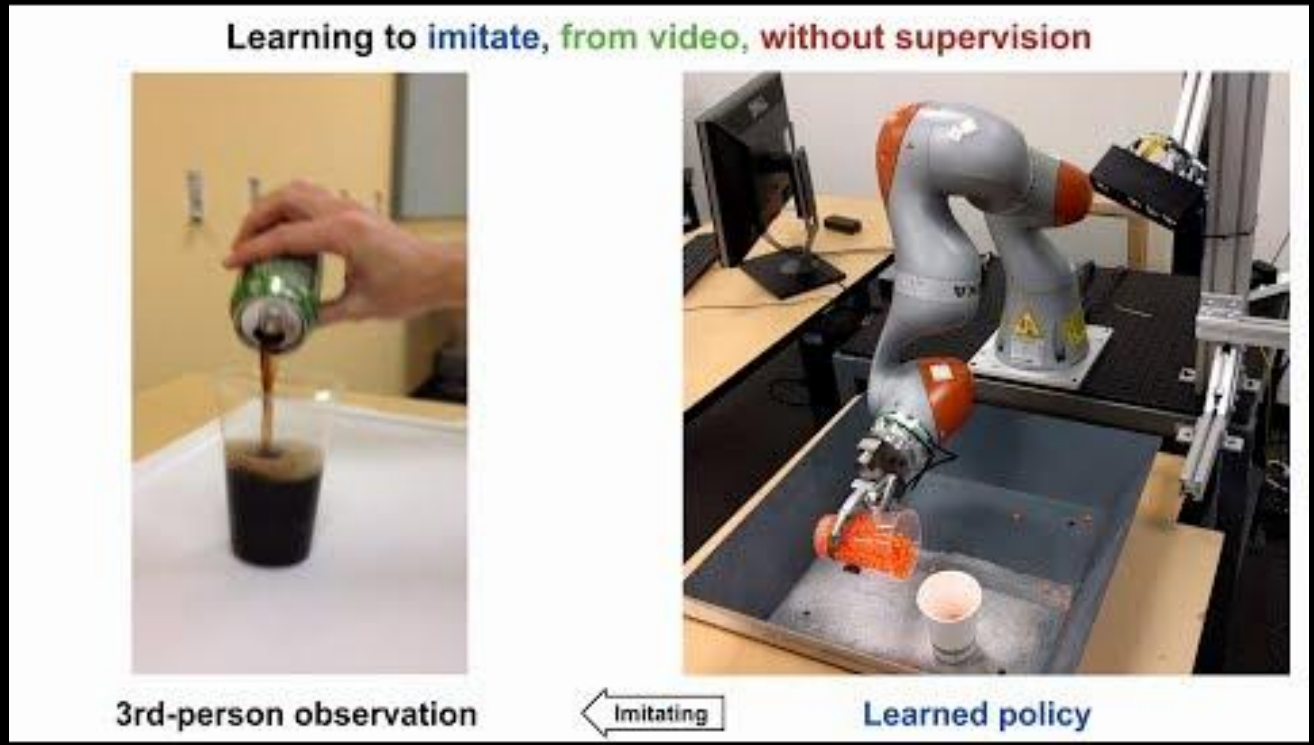
#TCN Sermanet, Pierre, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. "Time-contrastive networks: Self-supervised learning from video." ICRA 2018.

Multi-camera + Temporal ordering

Inference: Imitation learning for a robotic arm.



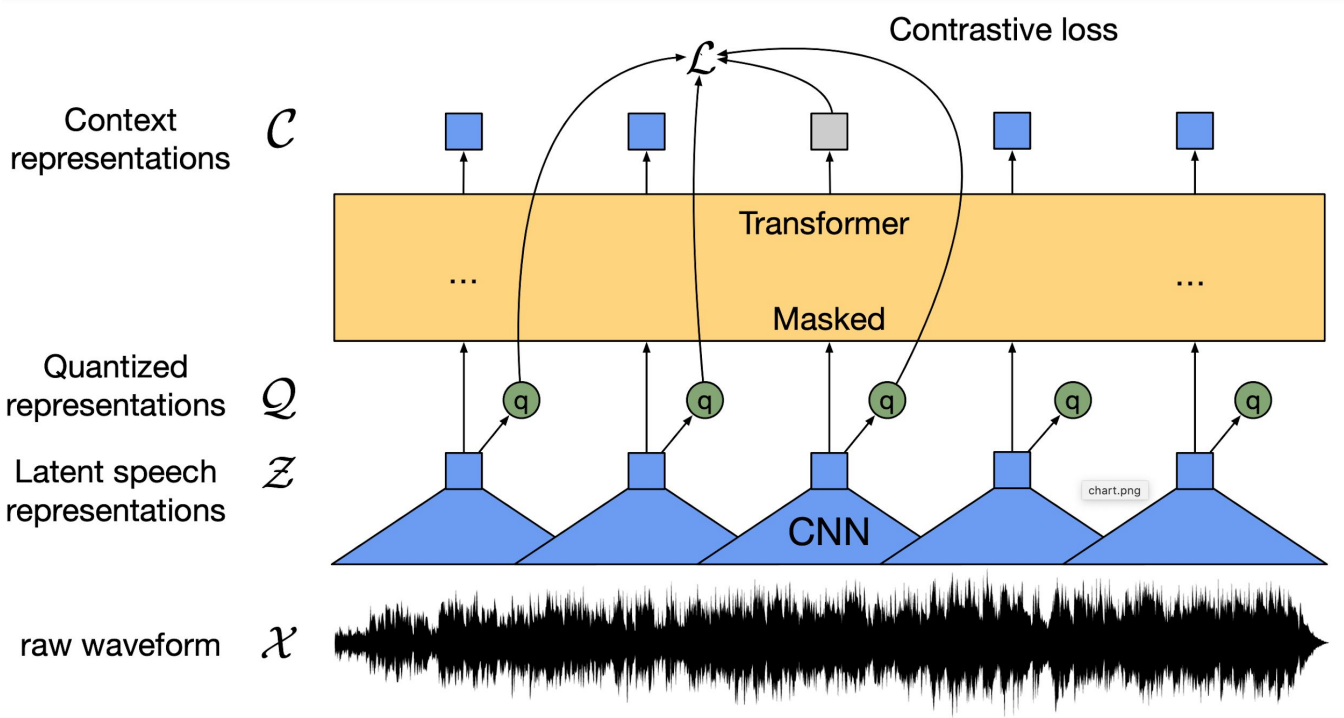
#TCN Sermanet, Pierre, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. "Time-contrastive networks: Self-supervised learning from video." ICRA 2018.



#TCN 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.

Masking + Contrastive loss for speech

Predict whether a sample belong to the future of the same sequence.

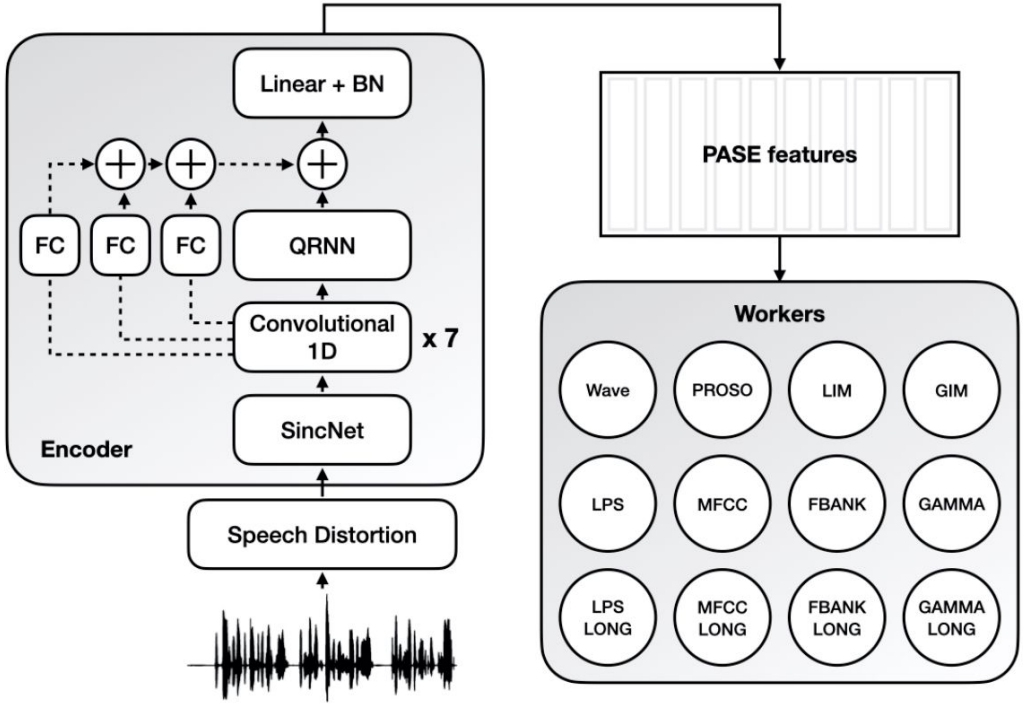


#Wav2Vec Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. arXiv preprint arXiv:2006.11477. [tweet] [code]

Outline

1. Transfer Learning
2. Representation Learning
3. Unsupervised Learning
4. Self-supervised Learning
5. Predictive methods
6. Contrastive methods
7. **Self-supervised Learning @ UPC**

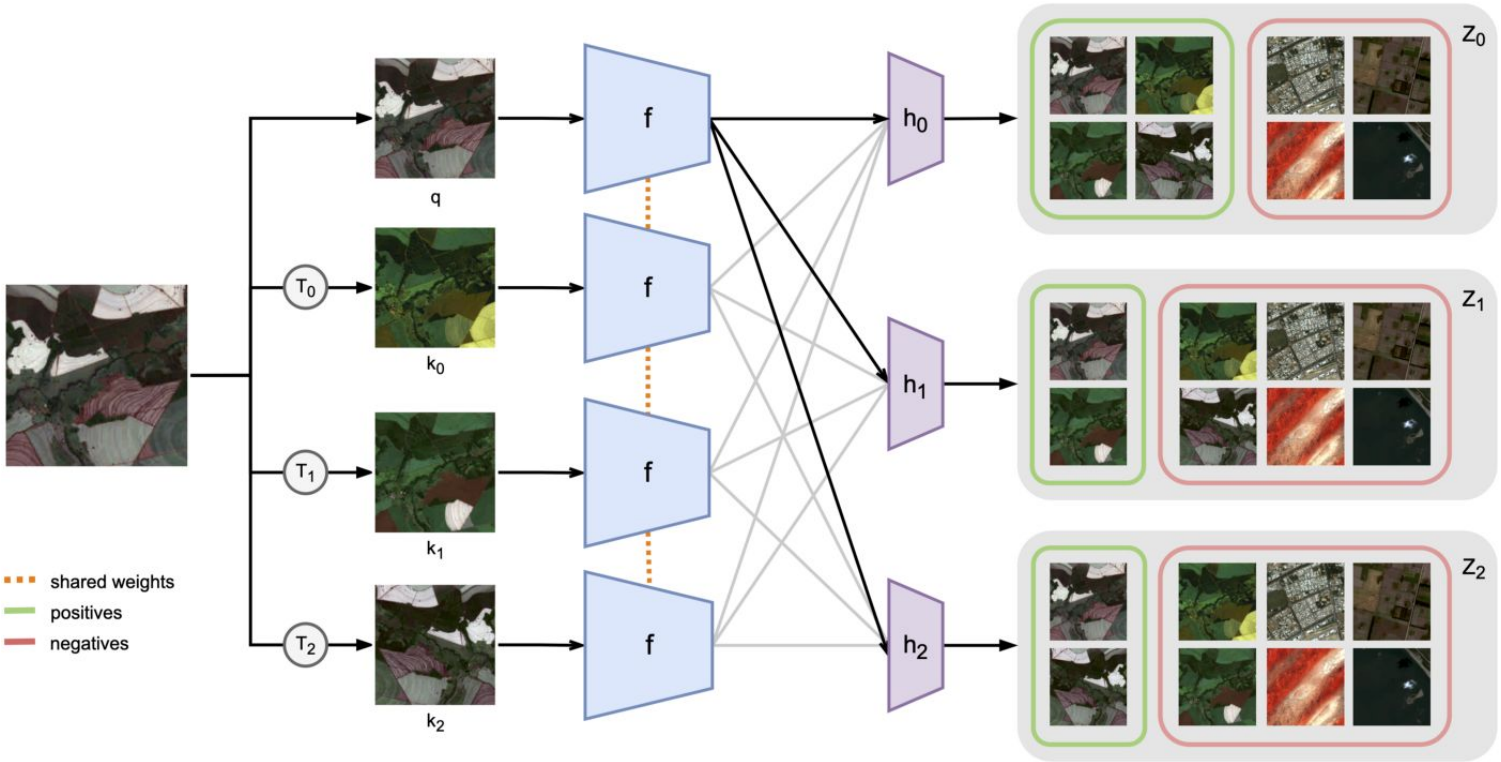
Predictive Method: Handcrafted Workers



#PASE Pascual, S., Ravanelli, M., Serrà, J., Bonafonte, A., & Bengio, Y. (2019). Learning problem-agnostic speech representations from multiple self-supervised tasks. INTERSPEECH 2019. [code]

Figure: Gerard I. Gallego, "End-to-end Speech Translation with Self-supervised Speech Representations" . UPC TelecomBCN 2020.

Contrastive Methods: Seasonal Contrast



#SeCo Mañas, O., Lacoste, A., Giro-i-Nieto, X., Vazquez, D. and Rodriguez, P.. Seasonal Contrast: Unsupervised Pre-Training from Uncurated Remote Sensing Data. ICCV 2021.[PyTorch Lightning]

Bonus Track

Xavier Giró, [“Self-Supervised Audio-Visual Learning”](). UPC TelecomBCN DLAI, 2020.

Homework for December 3

Víctor Campos, **“Towards RL that scales: Autonomous acquisition and transfer of knowledge”**.

UPC TelecomBCN DRL, 2020.

"Inefficient Data Efficient".

CVPR Deepvision 2020.

"The InfoNCE loss in self-supervised learning" [slides] [tweet]

NeurIPS SSL Workshop 2020.

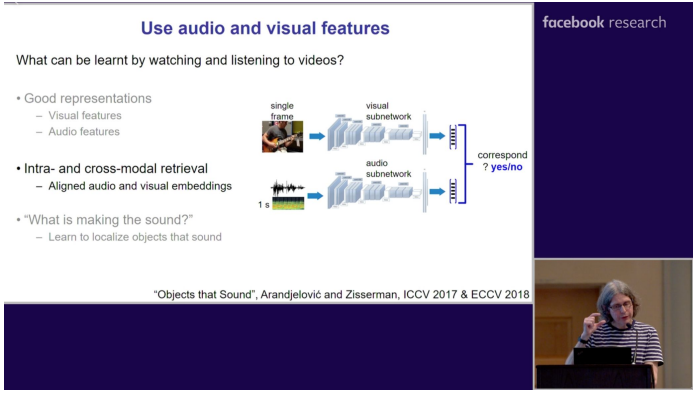
Outline

1. Transfer Learning
2. Representation Learning
3. Unsupervised Learning
4. Self-supervised Learning
5. Predictive methods
6. Contrastive methods
7. Self-supervised Learning @ UPC

Suggested talks

Demis Hassabis (Deepmind)

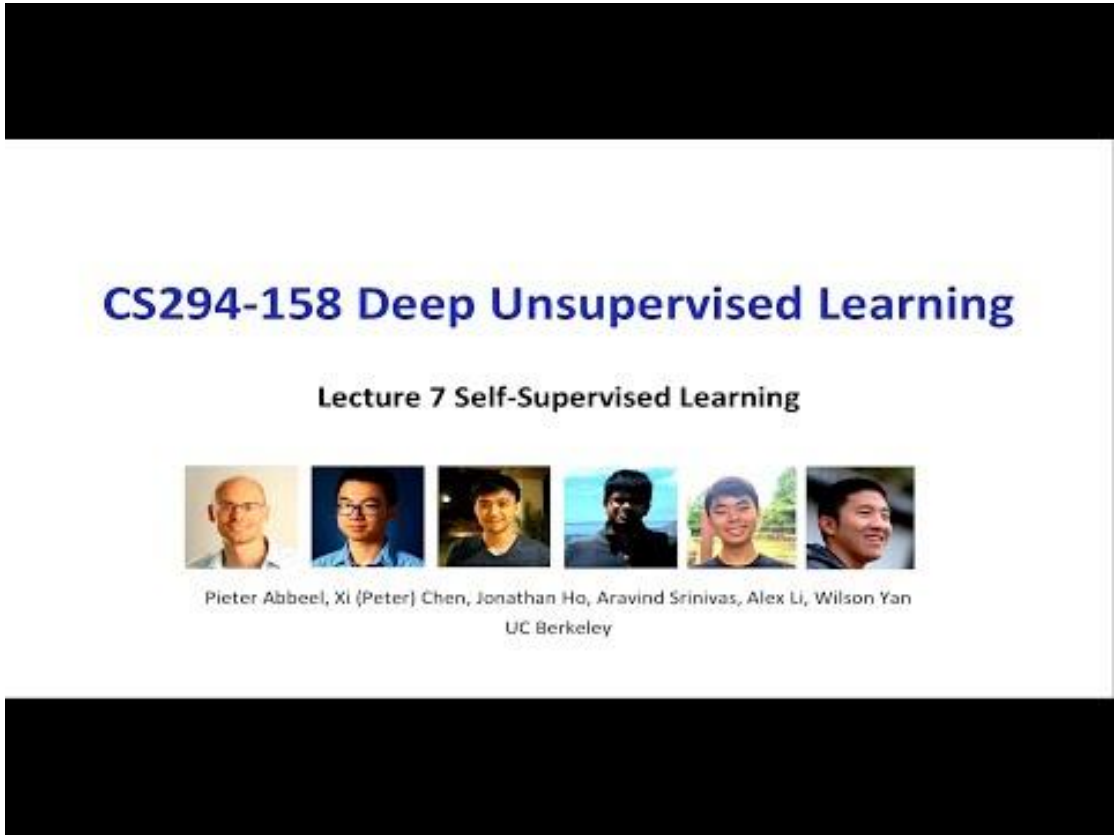
Aloysha Efros (UC Berkeley)



Andrew Zisserman (Oxford/Deepmind)

Yann LeCun (Facebook AI)

Related lecture



Aravind Srinivas, [“Deep Unsupervised Learning”](), UC Berkeley 2020

Suggested reading

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

Related events & publications

Events

- Self-supervised Learning: What's Next. ECCV Workshop 2020.
- Self-supervised Learning - Theory and Practice. NeurIPS Workshop 2020.

Publications

- Le-Khac, Phuc H., Graham Healy, and Alan F. Smeaton. "Contrastive representation learning: A framework and review." IEEE Access (2020).
- Purushwalkam, S., & Gupta, A. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. NeurIPS 2020.
- YHH Tsai, Y Wu, R Salakhutdinov, LP Morency,"Self-Supervised Learning from a Multiview Perspective". ICLR 2021

Transfer Learning: Video-lectures

Kevin McGuinness (UPC DLCV 2016)

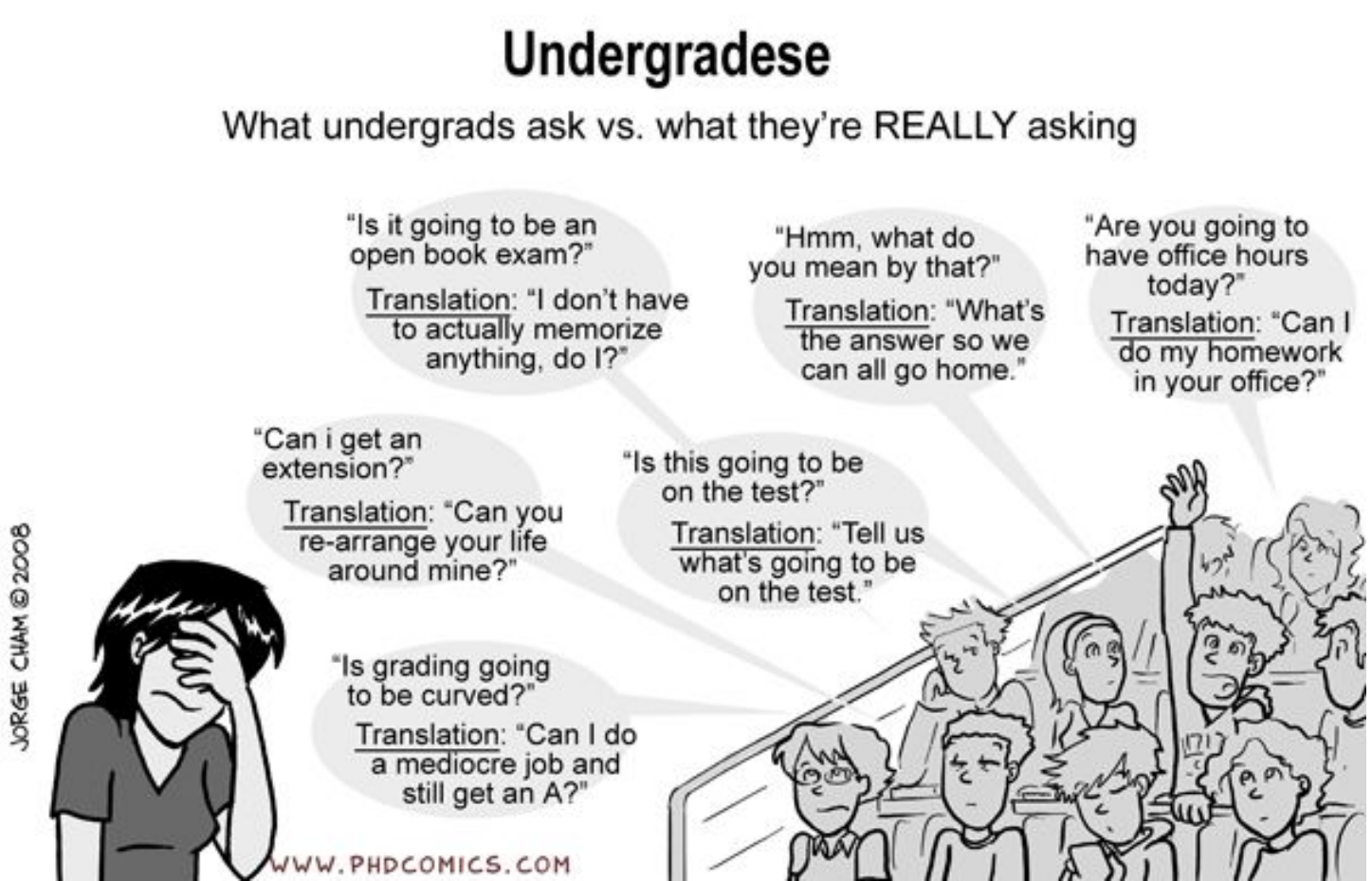
Ramon Morros (UPC DLAI 2018)

Video lectures on Unsupervised Learning

Kevin McGuinness, [UPC DLCV 2016]

Xavier Giró, [UPC DLAI 2017]

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