Computer Vision Center

Imaging Knowledge

Centre de Visió per Computador

Self-supervised Learning and Multimodal embeddings

Marçal Rossinyol

4th Summer School on Machine Learning
IIIT Hyderabad
July 11th 2019

Only Center in Europe fully devoted to Computer Vision

24 Years

+130 Staff

€2,7 M€ Income

Facts & Figures 2017

Research

21 Competitive projects obtained:

- **5** R+D projects of national calls (255 K€)
- **1** R+D project of european calls (18 K€)
- **3** R+D of other calls (233 K€)
- **12** HHRR grants (FI, DI, FPI, FPU i Cofund)

CVC is partner of the **Marie Curie Cofund P-Sphere** project within the UAB and has hired 6 post-doctoral researchers in 2017

Tech Transfer

36 new contracts signed, with a total budget of 752.000 €

50 active projects with a total budget of 1.098.000 €

3 license contracts granted for a total amount of 73.000 €

With companies such as: **Qidenus**, **Audi**, **Ficosa**, **Toyota Research Institute**, **Mediapro**, **Casa Tarradellas**, **Aimsun**, **Gas Natural**, **Intel**, **Sensofar**, **CaixaBank**, **Ciments Molins**, etc.

Facts & Figures 2017

Scientific Production

40 JCR indexed Journal articles

54 papers in International Conferences

7 book chapters

4 books

PhD Thesis

6 defended Thesis, within the Informatics PhD programme of the UAB. One of them co-directed with the Caen-Basse University

59 ongoing PhD Thesis. **2** in co-supervision with the Université de Monastir and the Chinese NPU. **12 industrial** PhDs

Media

29 articles in national & international press

4 clips on national and regional TV broadcasters

5 radio interviews

Human Resources

49 Post-doc and senior researchers

59 PhD students broadcasters

HR EXCELLENCE IN RESEARCH

We have been granted with the **HR Excellence in Research** distinction in the year 2015

Tech Transfer

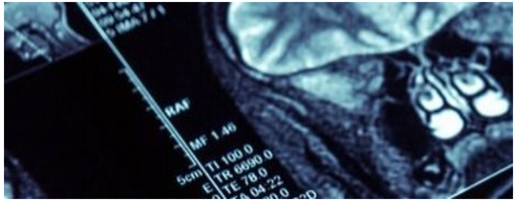
The main asset and guarantee of our work is the confidence placed by our partners for over 21 years, experiencing at first hand our expertise and professionalism

More than **350 projects** and feasibility studies
11 Spin-offs launched
More than **150 companies** among our customers

Spin-offs

Year	Spin-offs
1998	VISIÓ I ROBÒTICA APLICADA (VyRA) Computer Vision Solution
2001	VISUAL CENTURY Video Indexing
2002	ICAR VISION SYSTEMS Systems for personal document
2003	INSPECTA Cork quality Control
2005	DAVANTIS Smart surveillance
2012	CLOUD SIZING SERVICE Sizing clothing
2012	VISUAL TAGGING SERVICES Mobile apps
2014	CROWDMOBILE, SL Crowd Sourcing Solutions (Knowxel)
2015	CARE RESPITE Indoor Intelligent Visual System for Dependence
2016	ORAIN TECHNOLOGIES Intelligent Vending Machines
2019	ALL_READ Scene Text Recognition

Research



Health and well-being

Computer assisted diagnosis, intervention and planning;
Computational models of human vision;
Well-being and ambient assisted living.

Mobility and transport

Advanced driving systems and autonomous driving;
Virtual worlds for ADAS;
Unmanned Aerial Vehicles.

Culture & Experience-based technologies

Cultural heritage (AR/VR)
Reading Systems – Document analysis
Surveillance

Industry 4.0

Quality control
AR/VR technologies for industry 4.0
Robotic Vision

Intelligent reading systems

Self-Supervised Learning

Machines that learn

If we are ever to make a machine that will speak, understand or translate human languages, solve mathematical problems with imagination, practice a profession or direct an organization, either we must reduce these activities to a science so exact that we can tell a machine precisely how to go about doing them or we must develop a machine that can do things without being told precisely how.

R.M. Friedberg. "A Learning Machine". IBM Journal, Jan 1958

Machines that learn

If we are ever to make a machine that will speak, understand or translate human languages, solve mathematical problems with imagination, practice a profession or direct an organization, either we must reduce these activities to a science so exact that we can tell a machine precisely how to go about doing them or we must develop a machine that can do things without being told precisely how.

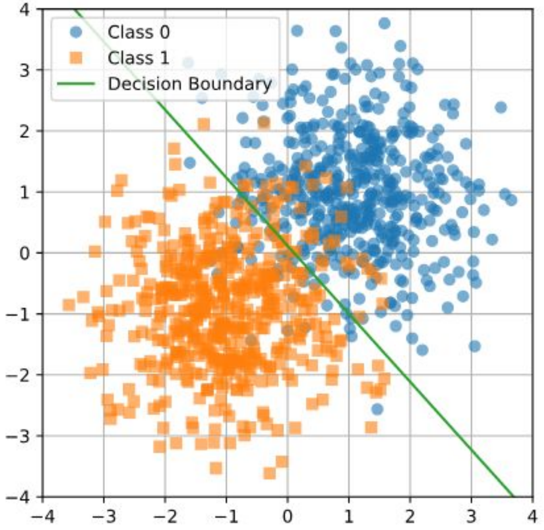
R.M. Friedberg. "A Learning Machine". IBM Journal, Jan 1958

Learning Procedures

- Supervised Learning

$y = f(x)$

Predict label y corresponding to observation x



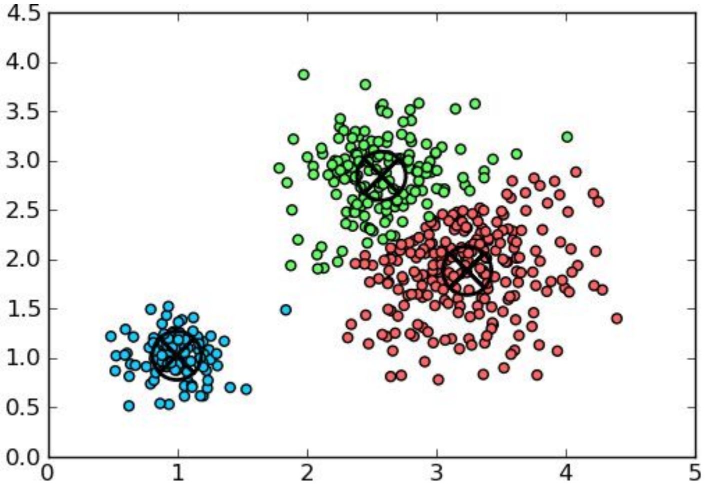
Learning Procedures

- Supervised Learning

$y = f(x)$ **Predict label y corresponding to observation x**

- Unsupervised Learning

$f(x)$ **Estimate the distribution of x**



Learning Procedures

- Supervised Learning

$y = f(x)$ **Predict label y corresponding to observation x**

- Unsupervised Learning

$f(x)$ **Estimate the distribution of x**

- Reinforcement Learning

$y = f(x)$ z **Predict action y based on observation x, to maximize a future reward z**

Self-supervised Learning

Strong Supervision (e.g. ImageNet)

- Features from networks trained on ImageNet can be used for other visual tasks, e.g.
  - detection, segmentation, action recognition, fine grained visual classification
- To some extent, any visual task can be solved now by:
  - Construct a large-scale dataset labelled for that task
  - Specify a training loss and neural network architecture
  - Train the network and deploy
- Self-supervision as an alternative to strong supervision for training

Source: A. Zisserman. Self-supervised learning

Self-supervised Learning

Why Self-supervision?

- Expense of producing a new dataset for each new task
- Some areas are supervision-starved, e.g. medical data, where it is hard to obtain annotation
- Untapped/availability of vast numbers of unlabelled images/videos
  - Facebook: one billion images uploaded per day
  - 300 hours of video are uploaded to YouTube every minute
- How infants may learn …

Source: A. Zisserman. Self-supervised learning

Self-supervised Learning

What is Self-supervision?

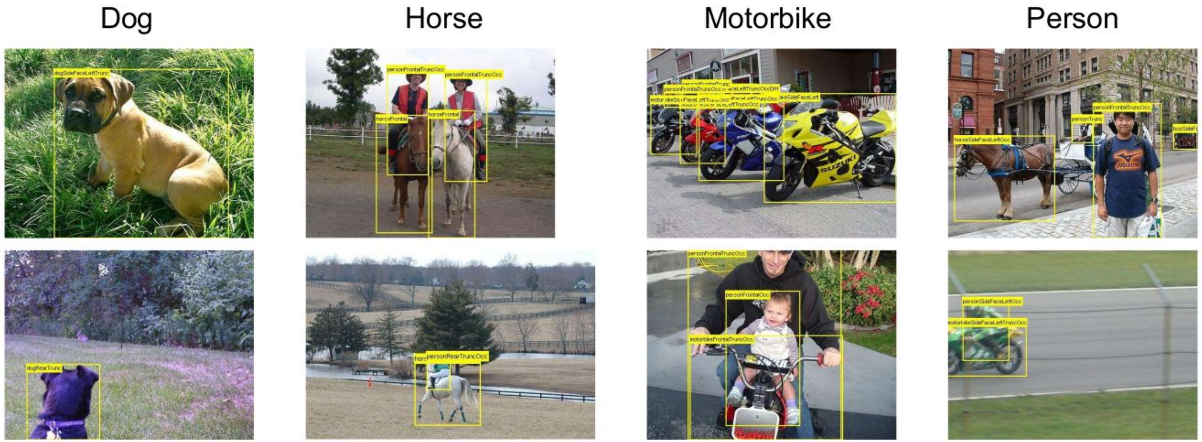
- A form of unsupervised learning where the data provides the supervision
- In general, withhold some part of the data, and task the network with predicting it
- The task defines a proxy loss, and the network is forced to learn what we really care about, e.g. a semantic representation, in order to solve it

Source: A. Zisserman. Self-supervised learning

ExI: Object Detection

Image we want to train an object detection network

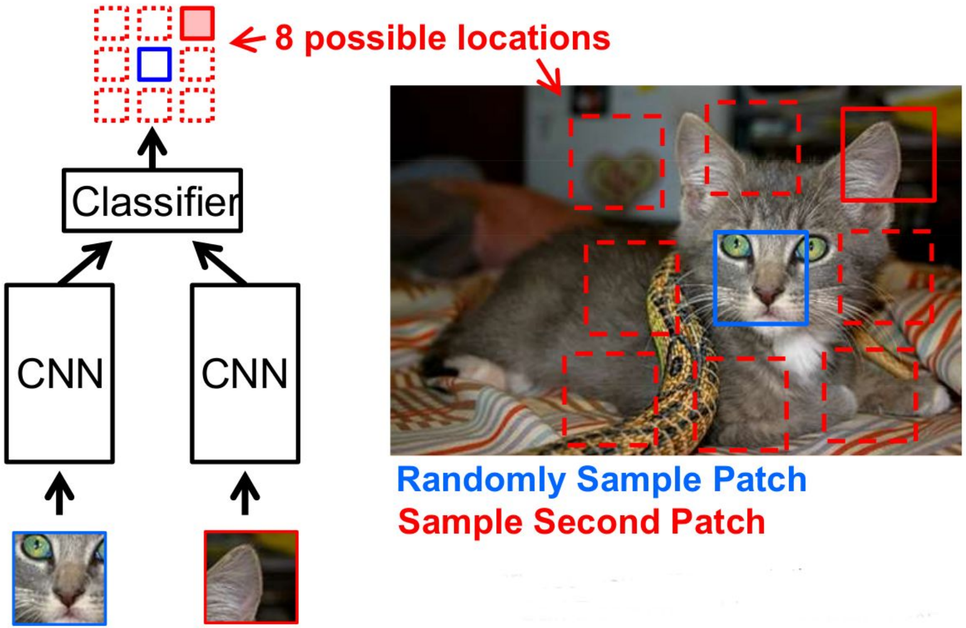
- PASCAL VOC Detection
  - 20 classes (car, bycicle, etc.)
  - Predict bounding boxes and object classes
- Usually pre-trained on ImageNet to get good visual features



Doersch et al. "Unsupervised visual representation learning by context prediction", ICCV15

ExI: Object Detection

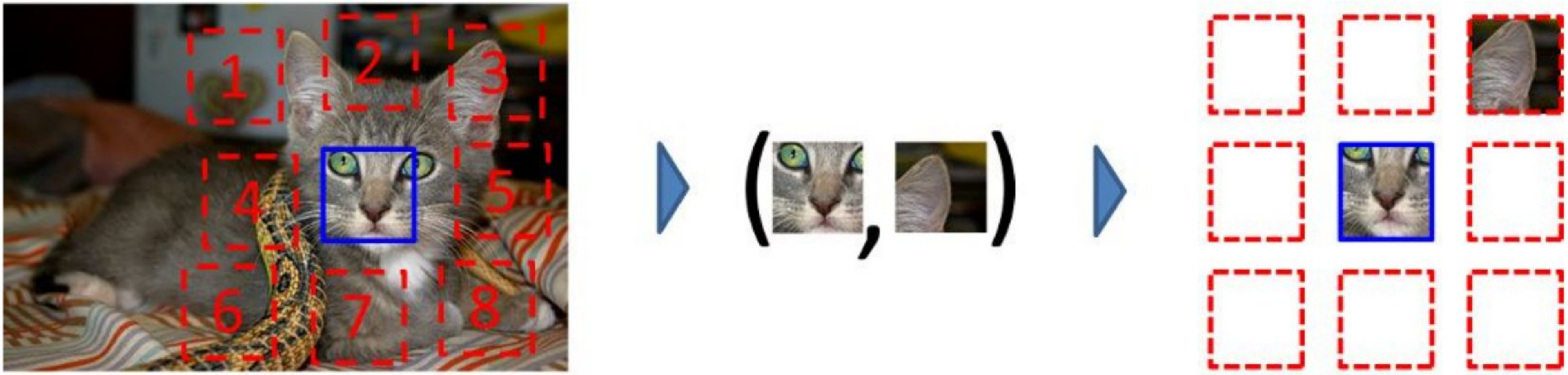
Train network to predict relative position of two regions in the same image



Doersch et al. "Unsupervised visual representation learning by context prediction", ICCV15

ExI: Object Detection

Train network to predict relative position of two regions in the same image



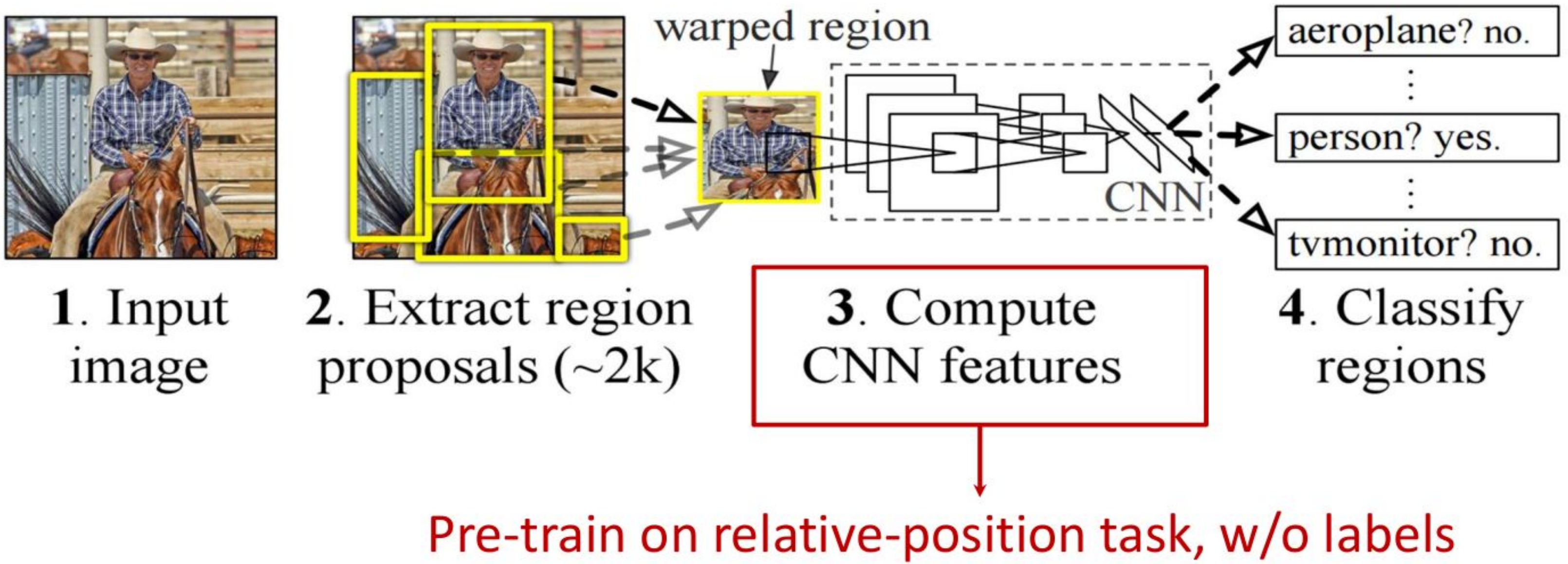
Doersch et al. "Unsupervised visual representation learning by context prediction", ICCV15

ExI: Object Detection

Pre-train CNN using self-supervision (no labels)

Train CNN for detection in R-CNN object category detection pipeline

R-CNN



Doersch et al. "Unsupervised visual representation learning by context prediction", ICCV15

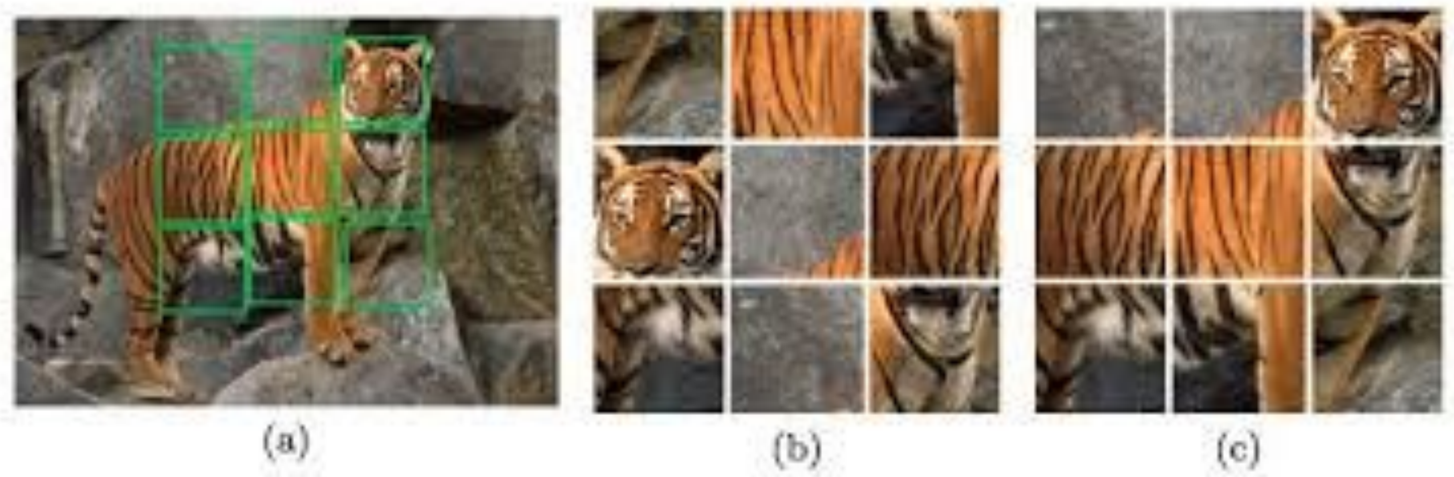
ExI: Object Detection

Pre-train CNN using self-supervision (no labels)

Train CNN for detection in R-CNN object category detection pipeline

	Average Precision
ImageNet labels	56.8%
Self-supervised relative positioning	51.1%
No pretraining	45.6%

ExI: Object Detection



Doersch et al. "Unsupervised visual representation learning by context prediction", ICCV15

ExII: Action Recognition

Imagine we want to train a model for action recognition from video clips

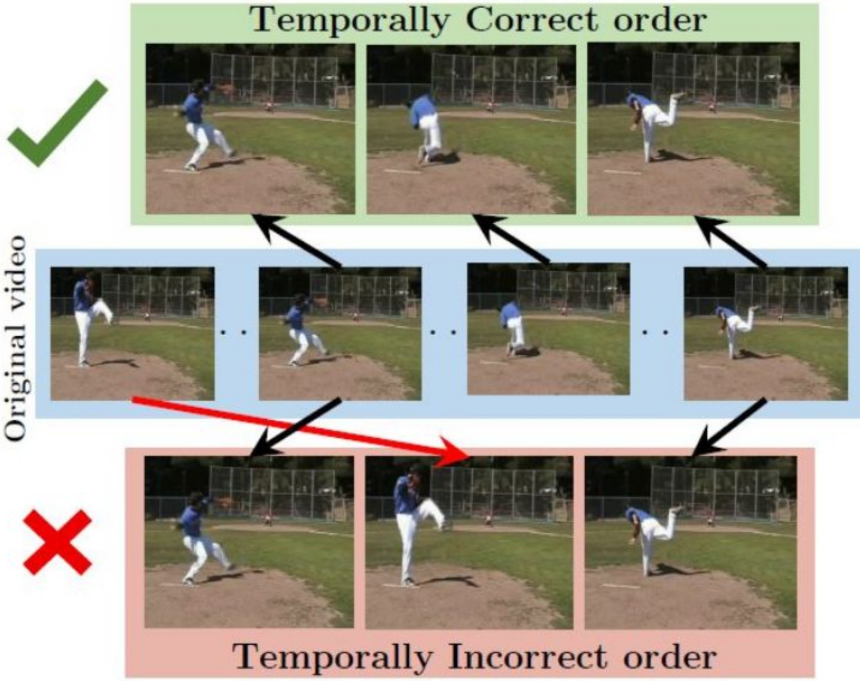
- UCF101 dataset

- HMDB51 dataset

Misra et al. "Shuffle and Learn: Unsupervised Learning using Temporal Order Verification", 2016

ExII: Action Recognition

Self-supervised learning by Temporal Order Verification



Misra et al. "Shuffle and Learn: Unsupervised Learning using Temporal Order Verification", 2016

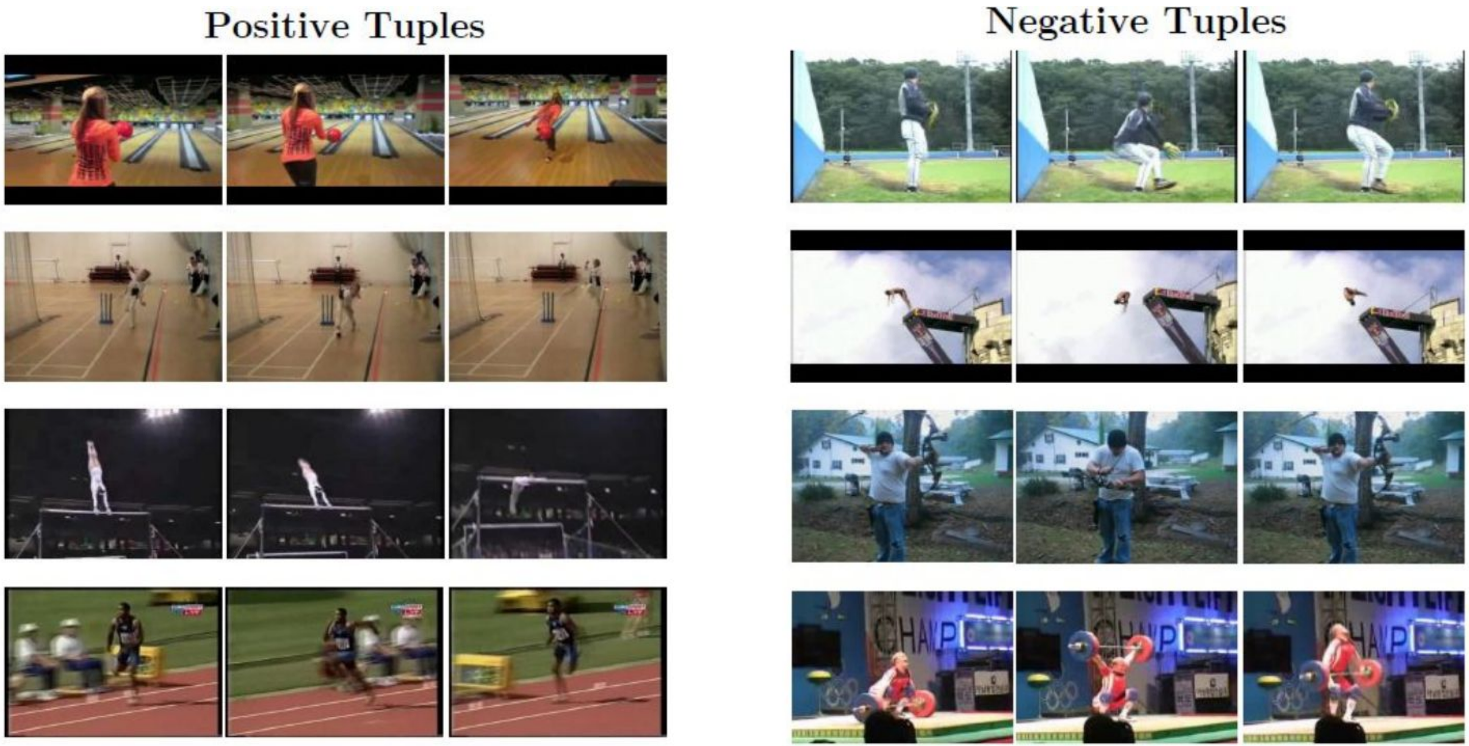
ExII: Action Recognition

Self-supervised learning by Temporal Order Verification

Misra et al. "Shuffle and Learn: Unsupervised Learning using Temporal Order Verification", 2016

ExII: Action Recognition

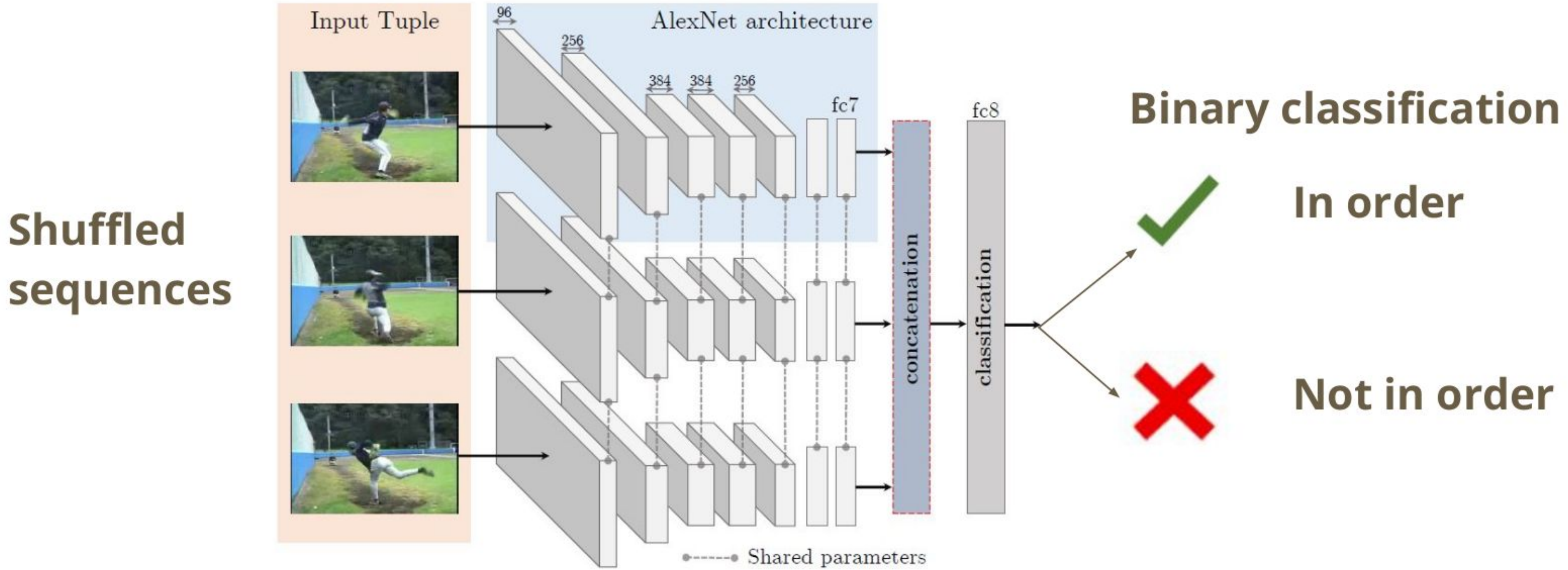
Self-supervised learning by Temporal Order Verification



Misra et al. "Shuffle and Learn: Unsupervised Learning using Temporal Order Verification", 2016

ExII: Action Recognition

Take temporal order as the supervisory signal for learning



Misra et al. "Shuffle and Learn: Unsupervised Learning using Temporal Order Verification", 2016

ExII: Action Recognition

- Comparison to random initialization & transfer learning

Dataset	Initialization	Mean Accuracy
UCF101	Random	38.6
	(Ours) Tuple verification	**50.2**
HMDB51	Random	13.3
	UCF Supervised	15.2
	(Ours) Tuple verification	**18.1**

+ 11.6 %

+ 4.8 %

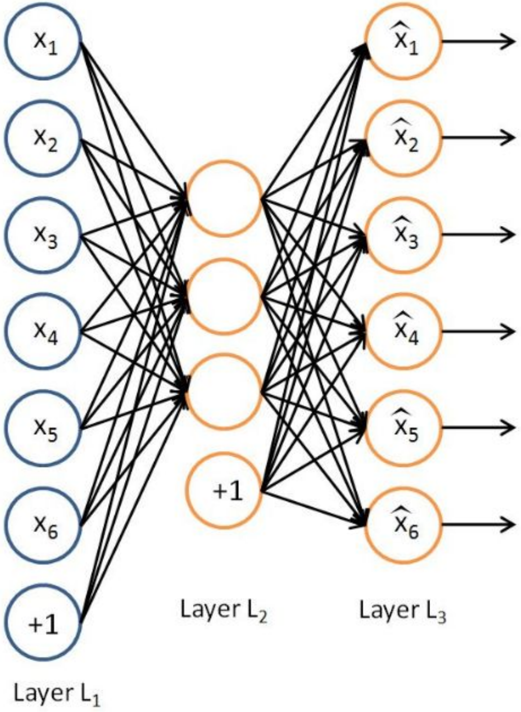
- Pre-trained on ImageNet and finetuned on UCF-101 gives an accuracy of 67.1%.
- Pre-trained on ImageNet and finetuned on HMDB-51 gives an accuracy of 28.5%.

Misra et al. "Shuffle and Learn: Unsupervised Learning using Temporal Order Verification", 2016

ExIII: Digit Recognition

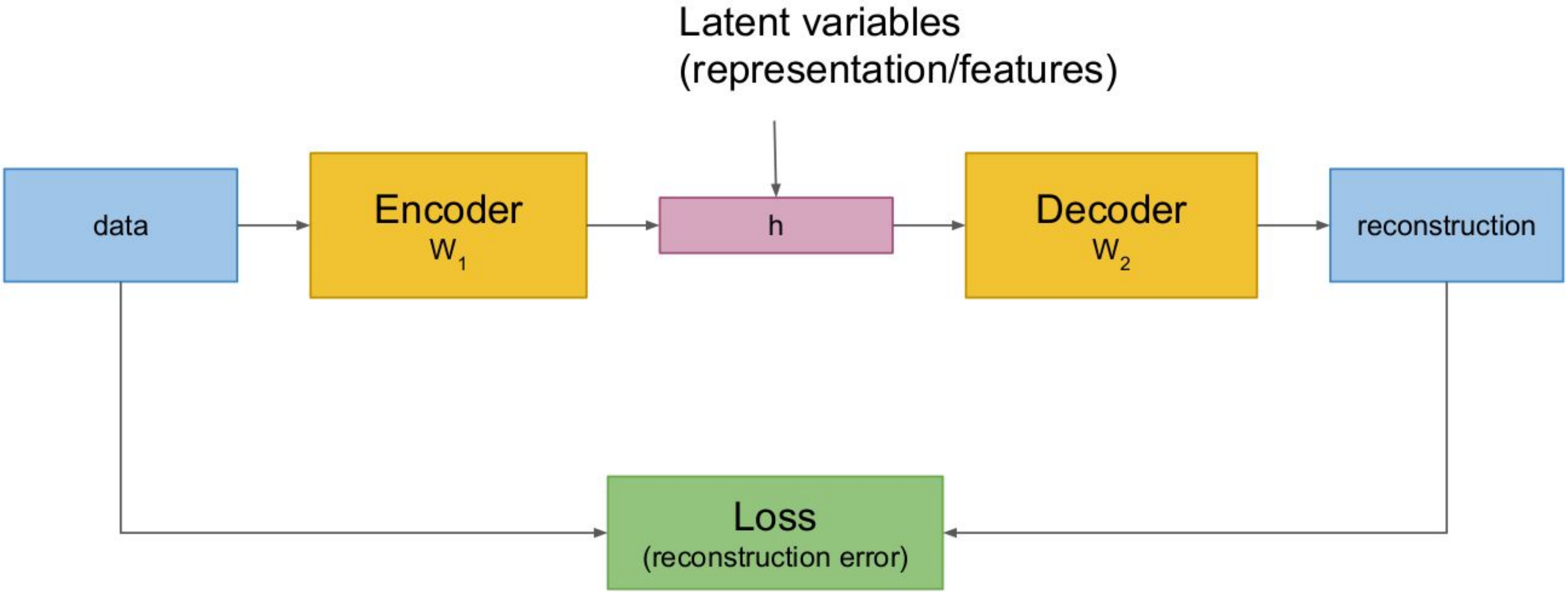
Use of autoencoder intermediate layers as self-supervised feature extraction

- e.g. MNIST numbers



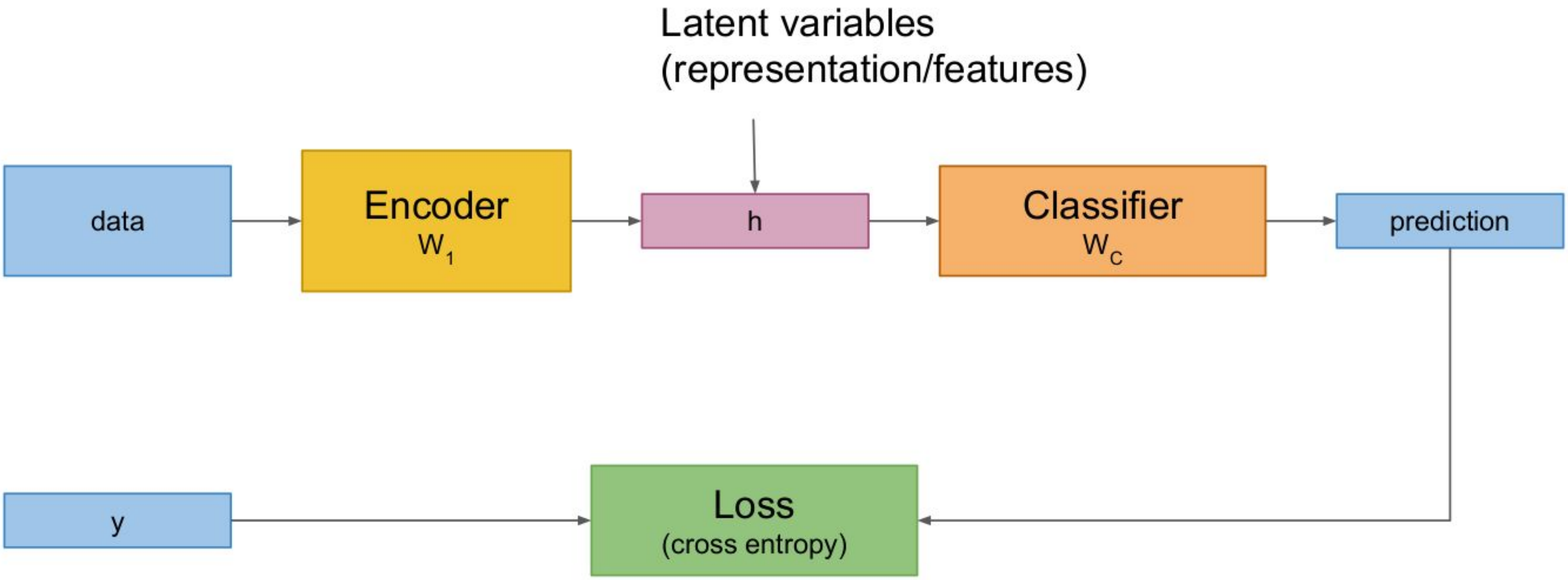
Source: K. McGuiness. "Unsupevised Learning"

ExIII: Digit Recognition



Source: K. McGuiness. "Unsupevised Learning"

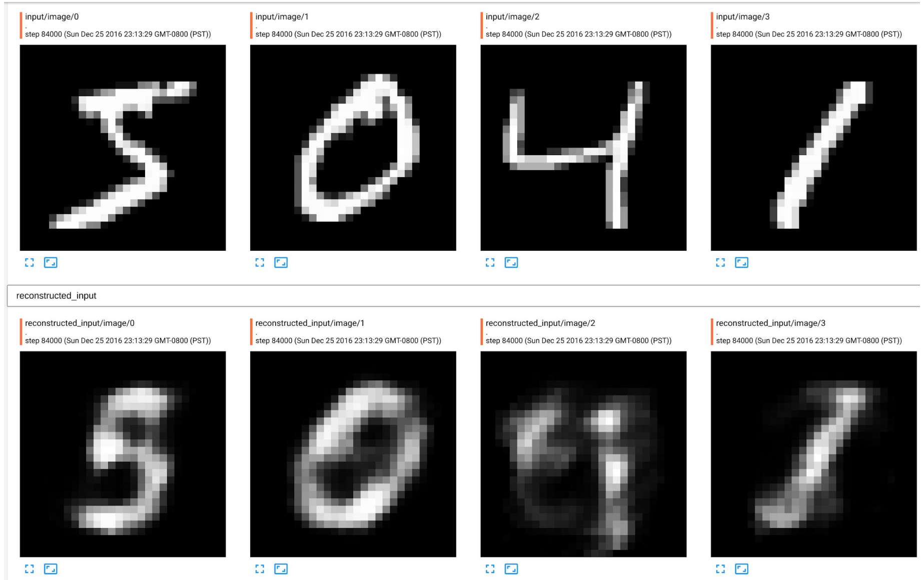
ExIII: Digit Recognition

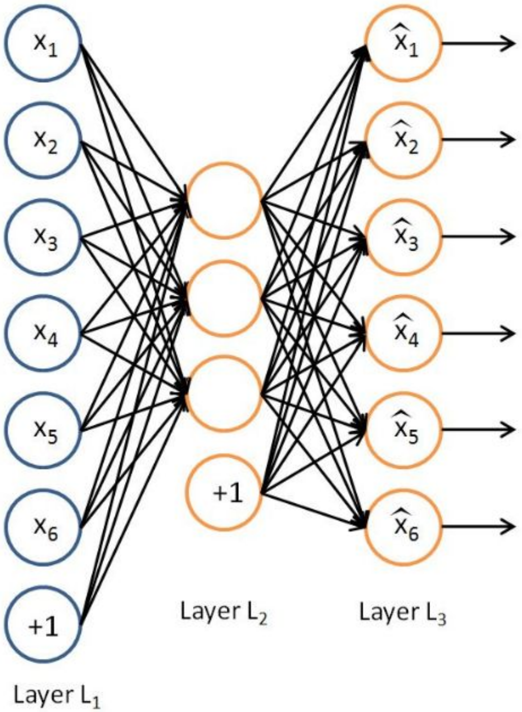


Source: K. McGuiness. "Unsupevised Learning"

ExIII: Digit Recognition

Can easily simulate training data by transforming images: 8.7% error MNIST w/ 100 examples



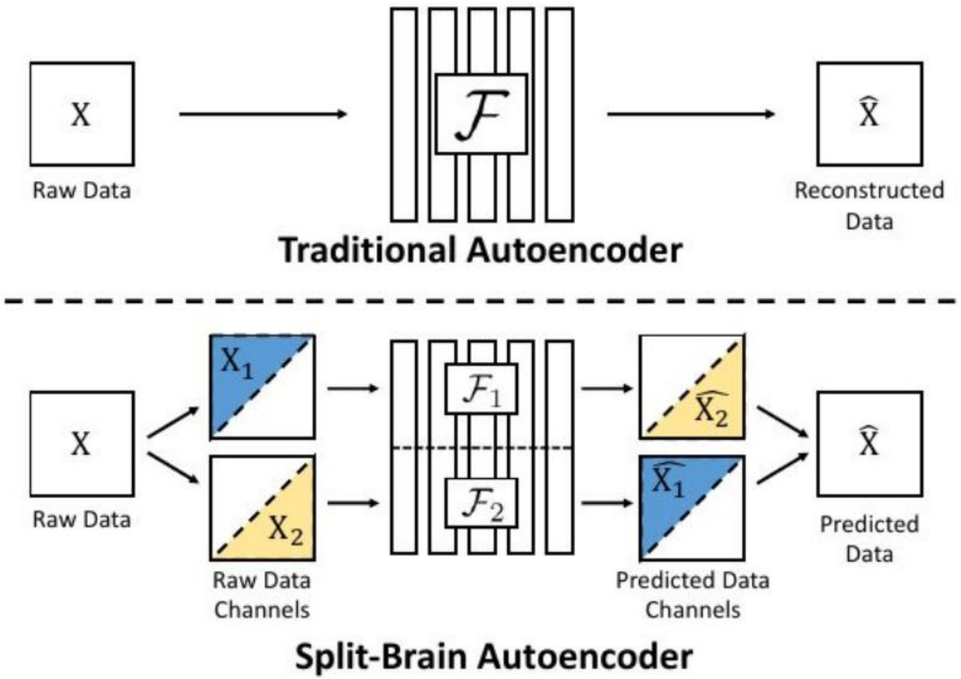


Source: K. McGuiness. "Unsupevised Learning"

Split-brain autoencoders

Simultaneously train two networks to predict one part of the data from the other.

Concat two networks and use features for other tasks.



Zhang et al. "Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction" 2016

Split-brain autoencoders

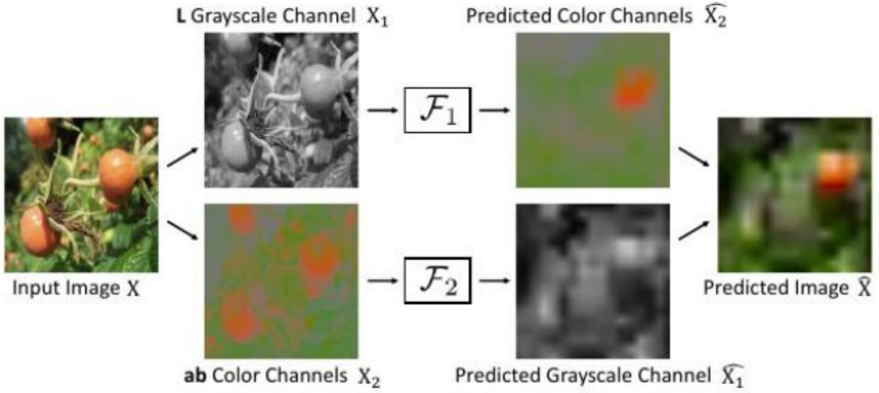
Simultaneously train two networks to predict one part of the data from the other.

Concat two networks and use features for other tasks.

Many possible proxy tasks:

- Predict chrominance from luminance

- Predict depth from RGB.



Zhang et al. "Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction" 2016

Split-brain autoencoders

Beat several state of the art self-supervised approaches on several datasets

Task Generalization on ImageNet Classification [37]					
Method	**conv1**	**conv2**	**conv3**	**conv4**	**conv5**
ImageNet-labels [26]	19.3	36.3	44.2	48.3	50.5
Gaussian	11.6	17.1	16.9	16.3	14.1
Krähenbühl et al. [25]	17.5	23.0	24.5	23.2	20.6
[1]Noroozi & Favaro [31]	19.2	30.1	34.7	33.9	28.3
Doersch et al. [8]	16.2	23.3	30.2	31.7	29.6
Donahue et al. [9]	**17.7**	24.5	31.0	29.9	28.0
Pathak et al. [35]	14.1	20.7	21.0	19.8	15.5
Zhang et al. [49]	13.1	24.8	31.0	32.6	31.8
Lab→Lab	12.9	20.1	18.5	15.1	11.5
Lab(drop50)→Lab	12.1	20.4	19.7	16.1	12.3
L→ab(cl)	12.5	25.4	32.4	33.1	32.0
L→ab(reg)	12.3	23.5	29.6	31.1	30.1
ab→L(cl)	11.6	19.2	22.6	21.7	19.2
ab→L(reg)	11.5	19.4	23.5	23.9	21.7
(L,ab)→(ab,L)	15.1	22.6	24.4	23.2	21.1
(L,ab,Lab)→(ab,L,Lab)	15.4	22.9	24.0	22.0	18.9
Ensembled L→ab	11.7	23.7	30.9	32.2	31.3
Split-Brain Auto (reg,reg)	17.4	27.9	33.6	34.2	32.3
Split-Brain Auto (cl,cl)	**17.7**	**29.3**	**35.4**	**35.2**	**32.8**

Dataset & Task Generalization on Places Classification [50]					
Method	**conv1**	**conv2**	**conv3**	**conv4**	**conv5**
Places-labels [50]	22.1	35.1	40.2	43.3	44.6
ImageNet-labels [26]	22.7	34.8	38.4	39.4	38.7
Gaussian	15.7	20.3	19.8	19.1	17.5
Krähenbühl et al. [25]	21.4	26.2	27.1	26.1	24.0
[1]Noroozi & Favaro [31]	23.0	32.1	35.5	34.8	31.3
Doersch et al. [8]	19.7	26.7	31.9	32.7	30.9
Wang & Gupta [46]	20.1	28.5	29.9	29.7	27.9
Owens et al. [33]	19.9	29.3	32.1	28.8	29.8
Donahue et al. [9]	**22.0**	28.7	31.8	31.3	29.7
Pathak et al. [35]	18.2	23.2	23.4	21.9	18.4
Zhang et al. [49]	16.0	25.7	29.6	30.3	29.7
L→ab(cl)	16.4	27.5	31.4	32.1	30.2
L→ab(reg)	16.2	26.5	30.0	30.5	29.4
ab→L(cl)	15.6	22.5	24.8	25.1	23.0
ab→L(reg)	15.9	22.8	25.6	26.2	24.9
Split-Brain Auto (cl,cl)	21.3	**30.7**	**34.0**	**34.1**	**32.5**

Zhang et al. "Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction" 2016

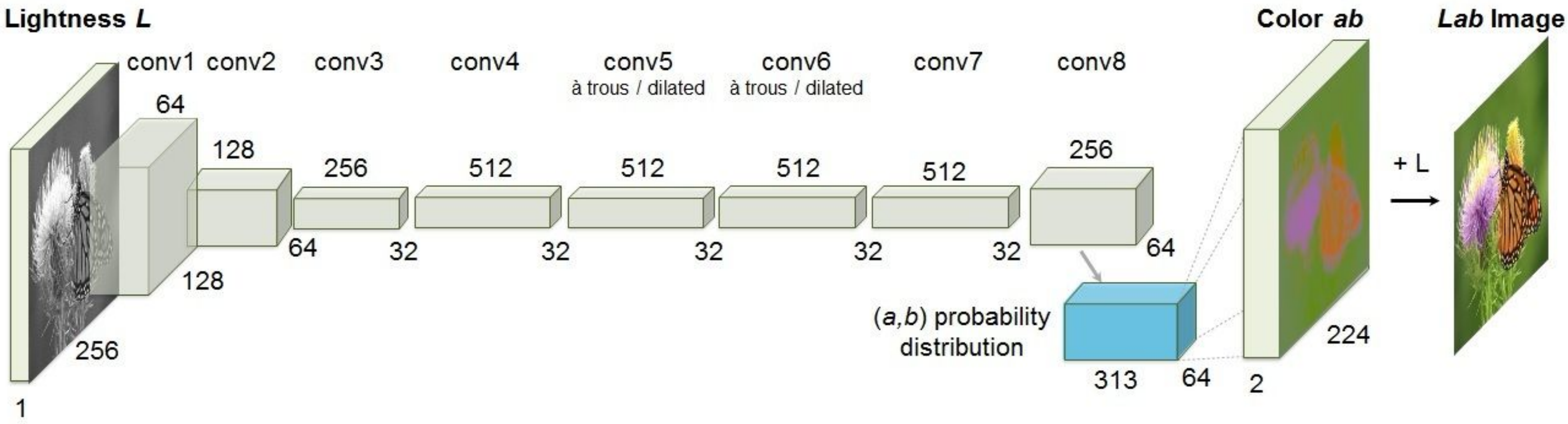
ExIV: Colorization

Take an RGB image, convert it to grayscale and make the network predict its colors

Zhang et al. "Colorful Image Colorization" ECCV 2016

ExIV: Colorization

Take an RGB image, convert it to grayscale and make the network predict its colors



Zhang et al. "Colorful Image Colorization" ECCV 2016

ExIV: Colorization

Let's test it!

Zhang et al. "Colorful Image Colorization" ECCV 2016

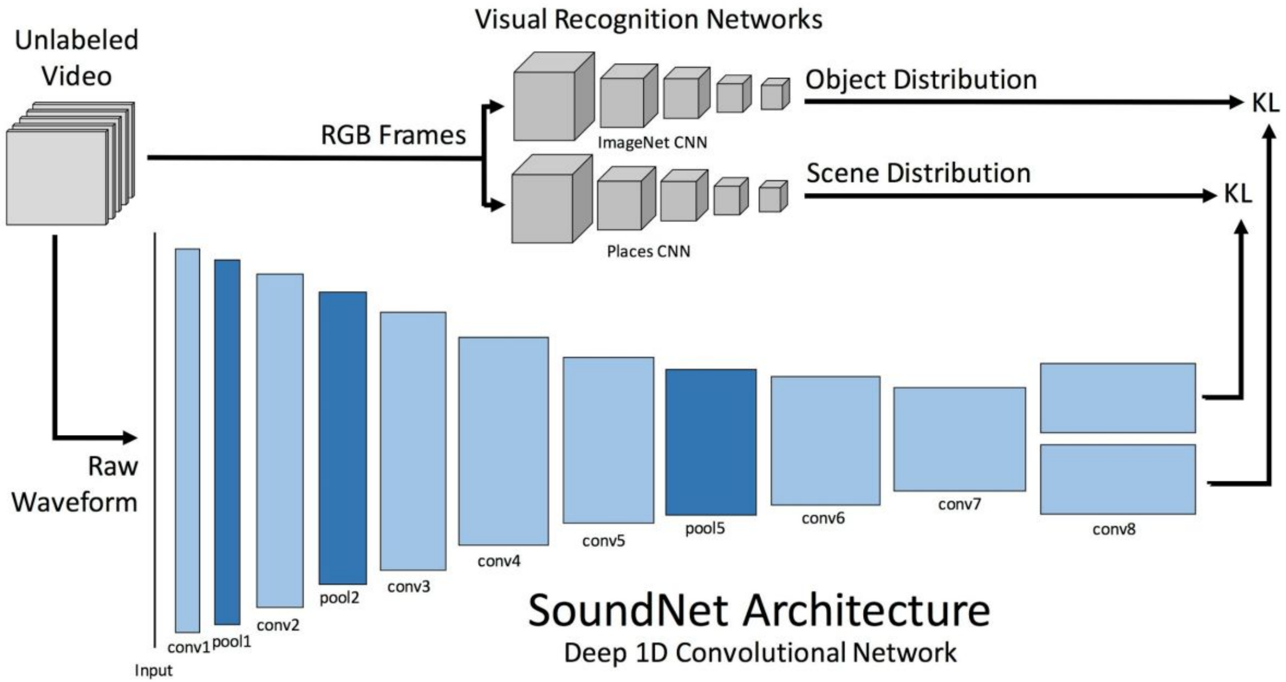
ExV: Predicting from sound

Using sound as a supervisory signal from videos we can

- Infer object and scene classes just hearing the videos
- Object localization from sound
- Make sound (speech) and video synthesis

ExV: Predicting from sound

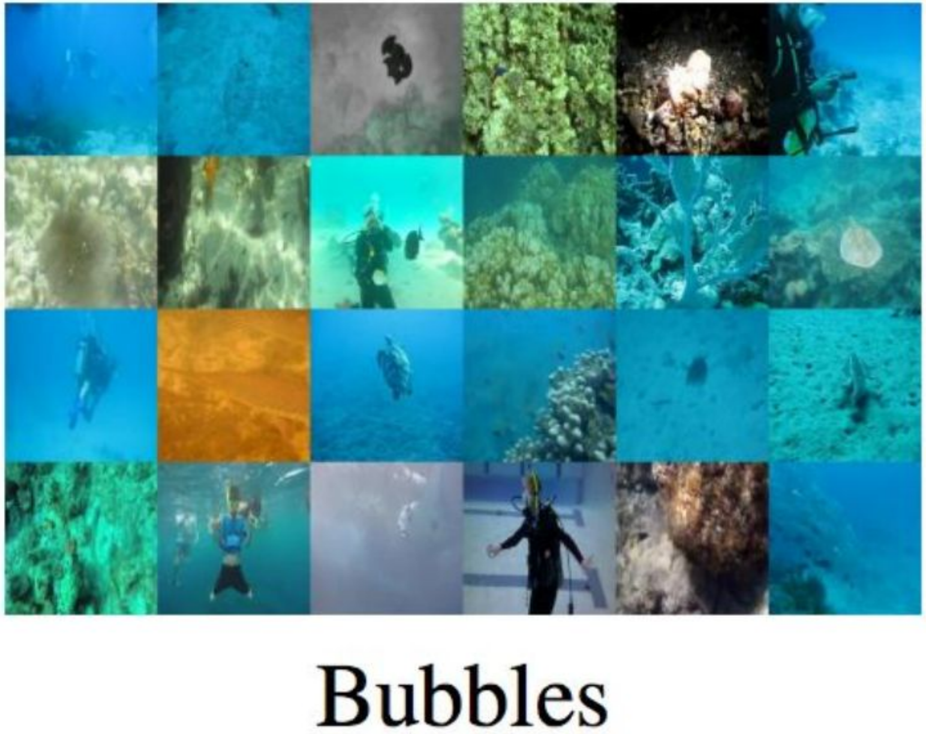
Infer object and scene classes just hearing the videos



Yusuf et al. "Soundnet: Learning sound representations from unlabeled video." NIPS 2016.

ExV: Predicting from sound

Infer object and scene classes just hearing the videos

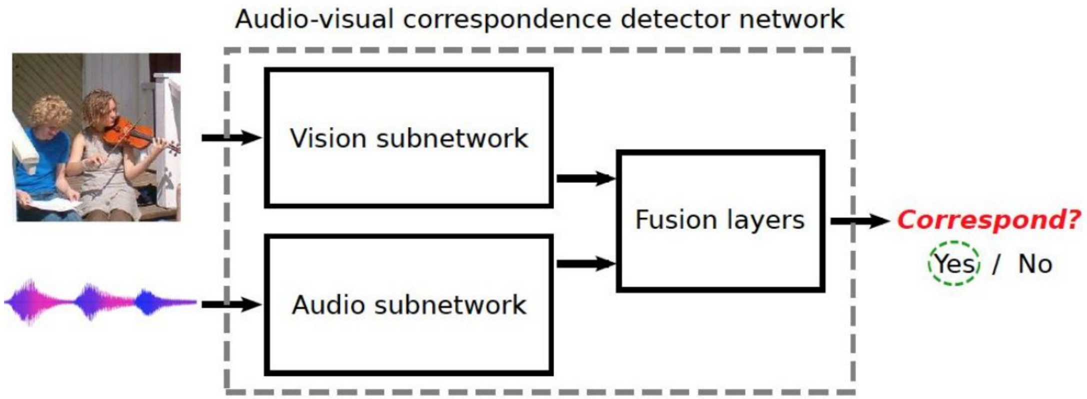


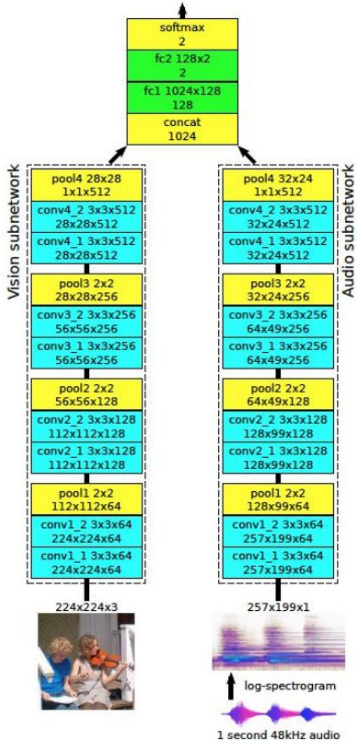
Yusuf et al. "Soundnet: Learning sound representations from unlabeled video." NIPS 2016.

ExV: Predicting from sound

"Object" localization from sound

Audio and visual features learned by assessing alignement.

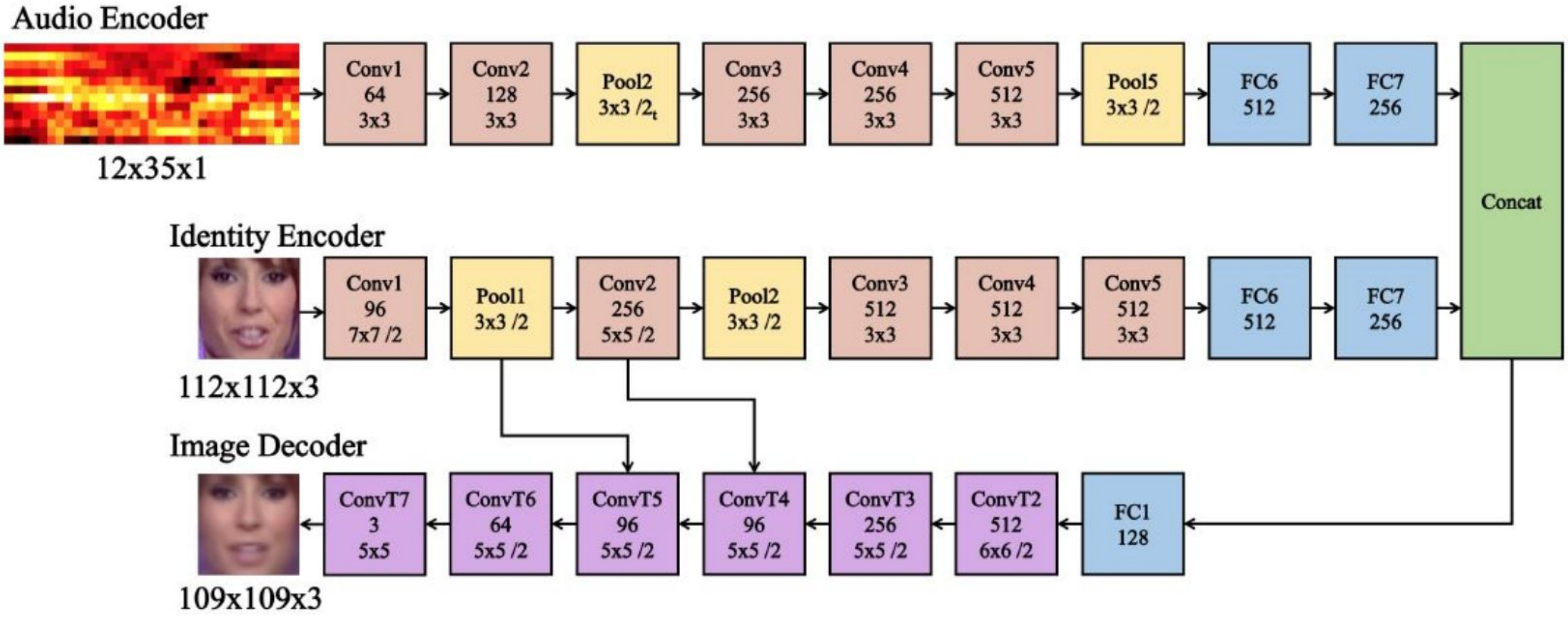




Arandjelovićet al. "Look, Listen and Learn." ICCV 2017
Senocak et al. "Learning to localize sound source in visual scenes" CVPR2018

ExV: Predicting from sound

Make sound (speech) video synthesis



Chung et al. "You said that?." BMVC 2017.