@DocXavi

Master in Computer Vision Barcelona

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

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

Associate Professor
Universitat Politècnica de Catalunya

Module 6 - Day 8 - Lecture 1

Self-supervised Learning from Video Sequences

28th March 2019

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)

Slide credit:
Yann LeCun

Supervised learning

Slide credit: Kevin McGuinness

Unsupervised learning

Slide credit: Kevin McGuinness

Types of machine learning

We can categorize three types of learning procedures:

1. Supervised Learning:

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

Predict label y corresponding to observation x

2. Unsupervised Learning:

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

Estimate the distribution of observation x

3. Reinforcement Learning (RL):

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

$$\mathbf{z}$$

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

Types of machine learning

We can categorize three types of learning procedures:

1. Supervised Learning:

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

2. Unsupervised Learning:

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

3. Reinforcement Learning (RL):

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

$$\mathbf{z}$$

Why Unsupervised Learning ?

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

Assumptions for Unsupervised Learning

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

“You can’t do inference without making assumptions”

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

Slide: Kevin McGuinness (DLCV UPC 2017)

Assumptions for Unsupervised Learning

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

“You can’t do inference without making assumptions”

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

Typical assumptions:

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

Slide: Kevin McGuinness (DLCV UPC 2017)

The manifold hypothesis

Slide: Kevin McGuinness (DLCV UPC 2017)

The manifold hypothesis

The data distribution lie close to a low-dimensional manifold

Example: **consider image data**

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

Similar for audio and text

Slide: Kevin McGuinness (DLCV UPC 2017)

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

University of Michigan

Universitat Politècnica de Catalunya

German Research Center for Artificial Intelligence

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. Temporal 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?

Pretraining:

1. Initialize a NN by solving an autoencoding problem.

Slide: Kevin McGuinness (DLCV UPC 2017)

Autoencoder (AE)

WHY?

Pretraining:

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

Slide: Kevin McGuinness (DLCV UPC 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

Temporal regularization: ISA

Features are learned with a Independent Subspace Analysis (ISA). Uses convolution and pooling operations.

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

Features are learned unsupervisedly by considering 3D (space+time) video blocks.

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. Temporal verification
    d. Predictive Learning
    e. Miscellaneous: optical flow, color & multiview

Related work on still images

A surrogate task is defined by exploiting the spatial context.

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

Related work on still images

What video-specific surrogate tasks could you think about ?

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

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

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

(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

Dataset	Initialization	Mean Classification Accuracy
UCF101	Random	38.6
	Shuffle & Learn	50.2
	ImageNet pre-trained	**67.1**

(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 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 coherence

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.

Temporal coherence

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
    a. Autoencoder
    b. Temporal regularisations
    c. Temporal verification
    d. Frame Prediction
    e. Miscellaneous: optical flow, color & multiview

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**

Slide credit:
Yann LeCun

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

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) 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 where green frames indicate conditioned input and red frames indicate generations.
Generations from the MCNet of Villegas et al. (2017), are shown for comparison.

#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

A CNN video architecture learns disentangles features for appearance & motion when trained for frame prediction.

#DisNet Carlos Arenas, Victor Campos, Sebastian Palacio, Xavier Giro-i-Nieto, [“Video Understanding through the Disentanglement of Appearance and Motion”](#) MSc thesis, ETSETB TelecomBCN 2018.

Frame Prediction + Disentangled features

A synthetic dataset of moving MNIST digits was built to have access to virtually an infinite amount of data.

- **Train:**
  - 5000 clips: (0) Horizontal
  - 5000 clips: (3) Vertical
- **Validation:**
  - 500 clips: (3) Horizontal
  - 500 clips: (0) Vertical
- **Bounding angle:** 180º
- **Speed:** 8 pixels/frame
- **Size:** original scale 1:1

#DisNet Carlos Arenas, Victor Campos, Sebastian Palacio, Xavier Giro-i-Nieto, "Video Understanding through the Disentanglement of Appearance and Motion" MSc thesis, ETSETB TelecomBCN 2018.

Outline

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

Noisy labels from motion (optical flow)

Noisy labels can be built with optical flow computed with a handcrafted tool.
NN somehow regularizes the noise present in the annotations

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

Noisy labels from color

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]

Noisy labels from color

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]

Noisy labels from color

Learned embeddings can be clustered to track objects.

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

Temporal + Multiview Weak Labels

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.

Temporal + Multiview Weak Labels

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.

Outline

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

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

Deep Learning courses @ UPC TelecomBCN:

- MSc course [2017] [2018]
- BSc course [2018] [2019]

Next edition: Autumn 2019

- 1st edition (2016)
- 2nd edition (2017)
- 3rd edition (2018)
- 4th edition (2019)

Registration open for 2019

- 1st edition (2017)
- 2nd edition (2018)
- 3rd edition - NLP (2019)

Registration open for 2019

Deep Learning for Professionals @ UPC School

Next edition starts November 2019. Sign up here.