Visual representation learning

Bill Freeman, Antonio Torralba, Phillip Isola
6.819 / 6.869
http://www.deeplearningbook.org/

By Ian Goodfellow, Yoshua Bengio and Aaron Courville

November 2016

Today: parts of chapters 14 and 15 (but this lecture is mostly a departure from the book)
Observed image

Drawn from memory

[Bartlett, 1932]

[Intraub & Richardson, 1989]
Observed image

Drawn from memory

[Bartlett, 1932]
[Intraub & Richardson, 1989]
"I stand at the window and see a house, trees, sky. Theoretically I might say there were 327 brightnesses and nuances of colour. Do I have "327"? No. I have sky, house, and trees."
— Max Wertheimer, 1923
Representation learning

Image

X

“Coral”

“Fish”

Compact mental representation
Convolution is pointwise multiplication in the frequency domain.
Representation learning

Good representations are:

1. Compact
2. Explanatory
3. Disentangled
4. Interpretable

[See “Representation Learning”, Bengio 2013, for more commentary]
Classification units

PIT/AIT

V4/PIT

V2/V4

V1/V2

[Serre, 2014]
Classical object recognition

Feature extractors

Edges
Texture
Colors

Classifier

Segments
Parts

"clown fish"
Deep learning

Learned

“clown fish”
What do deep nets internally learn?

"image features" (a vector representation of the image)
Visualizing and Understanding CNNs
[Zeiler and Fergus, 2014]

Gabor-like filters learned by **layer 1**

Image patches that activate each of the **layer 1** filters most strongly
Image patches that activate each of the \textbf{layer 2} neurons most strongly
Image patches that activate each of the layer 3 neurons most strongly
Image patches that activate each of the layer 4 neurons most strongly
Image patches that activate each of the \textbf{layer 5} neurons most strongly
CNNs learned the classical visual recognition pipeline!

Edges
Texture
Colors

Segments
Parts

“clown fish”

![Clown fish image]

![Edges, Texture, Colors]

![Segments, Parts]

![“clown fish” categories]
Object Detectors Emergence in Deep Scene CNNs

[Zhou et al., ICLR 2015]

- For each unit (neuron) in network, find which images it is most selective for (cause it to have highest activation)

- Find which pixels in these images are responsible by occluding regions and seeing which pixels, when occluded, cause activation to change the most

AlexNet
Object Detectors Emergence in Deep Scene CNNs

[Zhou et al., ICLR 2015]

pool 1
Object Detectors Emergence in Deep Scene CNNs

[Zhou et al., ICLR 2015]
Object Detectors Emergence in Deep Scene CNNs

[Zhou et al., ICLR 2015]
Object Detectors Emergence in Deep Scene CNNs

[Zhou et al., ICLR 2015]
Object Detectors Emergence in Deep Scene CNNs

[Zhou et al., ICLR 2015]
Represent image as a vector of neural activations
(perhaps representing a vector of detected texture patterns or object parts)
Disentangling

Layer 1 representation

Layer 6 representation

[DeCAF, Donahue, Jia, et al. 2013]

[Visualization technique : t-sne, van der Maaten & Hinton, 2008]
Investigating a representation via similarity analysis

How similar are these two images?

How about these two?

[Kriegeskorte et al. 2008]
Investigating a representation via similarity analysis

Representational Dissimilarity Matrix

\[
\| h_i - h_j \|
\]

Neural activation vector

[Kriegeskorte et al. 2008]
Investigating a representation via similarity analysis

IT Neuronal Units

Deep net (in particular, HMO)

[Yamins et al., PNAS 2014]
Investigating a representation via similarity analysis

Deep nets and the primate brain both learn similar metric spaces.

Deep nets organize visual information similarly to how our brains do!

[Yamins et al., PNAS 2014]
What do deep nets internally learn?

A CNN is a multiscale, hierarchical representation of data.

Image

Representations!

“Fish”
Transfer learning

“Generally speaking, a good representation is one that makes a subsequent learning task easier.” — Deep Learning textbook
Often, what we will be “tested” on is to learn to do a new thing.
Finetuning starts with the representation learned on a previous task, and adapts it to perform well on a new task.
Finetuning in practice

• Pretrain a network on task A (often object recognition), resulting in parameters $W$ and $b$

• Initialize a second network with some/all of $W$ and $b$

• Train the second network on task B, resulting in parameters $W'$ and $b'$
Finetuning in practice

Pretraining
Object recognition

Finetuning
Place recognition

dolphin
cat
grizzly bear
angel fish
chameleon
clown fish
iguana
elephant
The “learned representation” is just the weights and biases, so that’s what we transfer.
If we keep on finetuning for every new datapoint or task that comes our way, we get **online learning**. Humans seem to do this, we never stop learning.
Supervised object recognition

image X → Learner → "Fish" → label Y
Supervised object recognition

image X \rightarrow \text{Learner} \rightarrow \text{"Fish"}

label Y
Supervised object recognition

\[ \text{image } X \rightarrow \text{Learner} \rightarrow \text{"Fish"} \]

label Y
Supervised object recognition

image X \rightarrow \text{Learner} \rightarrow \text{“Fish”} \rightarrow \text{label Y}

...
Supervised computer vision

Hand-curated training data
+ Informative
- Expensive
- Limited to teacher’s knowledge

Vision in nature

Raw unlabeled training data
+ Cheap
- Noisy
- Harder to interpret
Learning

Data → Learner → Model

Inference

Input → Model → Output
Learning from examples

(aka **supervised learning**)

Training data

\[
\{x_1, y_1\}, \{x_2, y_2\}, \{x_3, y_3\}, \ldots
\]

\[
f^* = \arg \min_{f \in \mathcal{F}} \sum_{i=1}^{N} \mathcal{L}(f(x_i), y_i)
\]
Learning without examples

(includes unsupervised learning and reinforcement learning)

\[
\begin{align*}
\text{Data} & \\
\{x_1\} & \rightarrow \\
\{x_2\} & \rightarrow \\
\{x_3\} & \rightarrow \\
\ldots & \\
\end{align*}
\]
Unsupervised Representation Learning

Data

\[
\begin{align*}
\{x_1\} & \quad \rightarrow \quad \text{Learner} \\
\{x_2\} & \\
\{x_3\} & \\
\ldots & \\
\end{align*}
\]

\rightarrow \quad \text{Representations}
Unsupervised Representation Learning

X

Image

“Coral”

“Fish”

Compact mental representation
Unsupervised Representation Learning

$X$

Image

compressed image code (vector $z$)
Unsupervised Representation Learning

Image $X$ → compressed image code (vector $z$) → Reconstructed image $\hat{X}$

"Autoencoder"

[e.g., Hinton & Salakhutdinov, Science 2006]
Autoencoder

\[ \hat{X} = \mathcal{F}(X) \]

Image

\[ \text{arg min}_{\mathcal{F}} \mathbb{E}_X[||\mathcal{F}(X) - X||] \]

Reconstructed image

[e.g., Hinton & Salakhutdinov, Science 2006]
\[ \hat{X} = \mathcal{F}(X) \]

[e.g., Hinton & Salakhutdinov, Science 2006]
Autoencoder

Data \( \{x_i\}_{i=1}^N \) → \( f \)

**Learner**

Objective
\[ \mathcal{L}(f(x), x) = \|f(x) - x\|^2 \]

Hypothesis space
CNN

Optimizer
SGD
Data compression

[Data] \( \xrightarrow{\text{compression}} \) [Compressed Data]

[Hinton & Salakhutdinov, Science 2009]
Label prediction

e.g., image classification
Data prediction

$X_1$

Some data

$\hat{X}_2$

Other data
Grayscale image: $L$ channel

$X \in \mathbb{R}^{H \times W \times 1}$

Color information: $ab$ channels

$\hat{Y} \in \mathbb{R}^{H \times W \times 2}$

[Zhang, Isola, Efros, ECCV 2016]
Deep Net “Electrophysiology”

[Zeiler & Fergus, ECCV 2014]
[Zhou et al., ICLR 2015]
Stimuli that drive selected neurons (conv5 layer)
Is the code informative about object class \( y \)?

Logistic regression:
\[
y = \sigma(Wz + b)
\]
Self-supervised learning

Common trick:

- Convert “unsupervised” problem into “supervised” empirical risk minimization

- Do so by cooking up “labels” (prediction targets) from the raw data itself

Escher, 1948
Multisensory self-supervision

Supervised
- implausible label

"cow"
Target

Input

Self-Supervised
- derives label from a co-occurring input to another modality

Input 1

Input 2

moo


[see also “Six lessons from babies”, Smith and Gasser 2005]
The allegory of the cave
“Multiview” self-supervised learning

Data \( \{x_i\}_{i=1}^N \) \( \rightarrow \) Learner

Objective
\[ \arg\min_f \sum_i D(f(g(x_i)) - h(x_i)) \]

Distance function

\( g \) and \( h \) are two “views” of the data \( x \), e.g., two different image channels

\( \rightarrow f \)
Goal: Set up a pre-training scheme to induce a “useful” representation 

[Slide credit: Richard Zhang]
Unsupervised visual representation learning by context prediction

[Carl Doersch, Abhinav Gupta, Alexei A. Efros, ICCV 2015]
Context as Supervision

[Collobert & Weston 2008; Mikolov et al. 2013]

house, where the professor lived without his wife and child; or so he said jokingly sometimes: “Here’s where I live. My house.” His daughter often added, without resentment, for the visitor’s information, “It started out to be for me, but it’s really his.” And she might reach in to bring forth an inch-high table lamp with fluted shade, or a blue dish the size of her little fingernail, marked “Kitty” and half full of eternal dust, but she was sure to replace these, after they had been admired, pretty near exactly where they had been. The little house was very orderly, and just big enough for all it contained, though to some tastes the bric-à-brac in the parlor might seem excessive. The daughter’s preference was for the store-bought gimmicks and appliances, the toasters and sweepers of Lilliput, but she knew that most adult vi...
Context Prediction for Images

A

B

[Slide credit: Carl Doersch]
Semantics from a non-semantic task

[Slide credit: Carl Doersch]
Relative Position Task

8 possible locations

Randomly Sample Patch
Sample Second Patch

[Slide credit: Carl Doersch]
Patch Embedding (representation)

Input | Nearest Neighbors

Note: connects *across* instances!

[Slide credit: Carl Doersch]
Prediction hypothesis

1. To survive, biological agents are constantly trying to anticipate, to predict sensations.

2. This trains up representations useful for prediction — surfaces, objects, events!

Henri Cartier-Bresson
Summary

1. Deep nets learn *representations*, just like our brains do.

2. This is useful because representations transfer — they act as prior knowledge that enables quick learning on new tasks.

3. Representations can also be learned without labels, which is great since labels are expensive and limiting.

4. Without labels there are many ways to learn representations. We saw:
   1. representations as compressed codes
   2. representations that are shared across sensory modalities
   3. representations that are predictive of the future
Yann LeCun’s cake:

1. Cake is unsupervised representation learning
2. Frosting is supervised transfer learning
3. Cherry on top is reinforcement learning (model-based RL)