Self-Supervised Visual Learning
a.k.a. “It’s all about the Data”!
“AI” is (almost) here!
Deep Learning solves everything!

• “No need to code anymore”

• For any given problem, just:
  1. label some training data
  2. define an objective function
  3. train neural network

  4. sell your startup for millions!
ImageNet Challenge (1000 object classes), Fei-Fei et al.
Planetary-scale geolocation

Automatic Image Captioning

- A group of people posing for a picture on a ski lift
- A pot of broccoli on the stove

Microsoft Research (Fang et al, 2015) + many other groups
BUT…

something seems too good to be true
Image Captioning

“a car parked on the side of the road”
“a car parked on the side of the road”
“a car parked on the side of the road”
Image Classification

• Performance on ImageNet: ~80%
• Performance in the real world: ~30%

“T-Shirt” class in ImageNet
T-Shirts in the real world
Geolocation

im2gps, 2008

• Nearest Neighbors
• 6 million images

PlaNet, 2016

• Deep Net
• 91 million images
## Algorithm vs. Data

<table>
<thead>
<tr>
<th>Method</th>
<th>Street 1 km</th>
<th>City 25 km</th>
<th>Region 200 km</th>
<th>Country 750 km</th>
<th>Continent 2500 km</th>
</tr>
</thead>
<tbody>
<tr>
<td>Im2GPS (orig) [19]</td>
<td>12.0%</td>
<td>15.0%</td>
<td>23.0%</td>
<td>47.0%</td>
<td></td>
</tr>
<tr>
<td>Im2GPS (new) [20]</td>
<td>2.5%</td>
<td><strong>21.9%</strong></td>
<td><strong>32.1%</strong></td>
<td>35.4%</td>
<td>51.9%</td>
</tr>
<tr>
<td>PlaNet (900k)</td>
<td>0.4%</td>
<td>3.8%</td>
<td>7.6%</td>
<td>21.6%</td>
<td>43.5%</td>
</tr>
<tr>
<td>PlaNet (6.2M)</td>
<td>6.3%</td>
<td><strong>18.1%</strong></td>
<td><strong>30.0%</strong></td>
<td>45.6%</td>
<td>65.8%</td>
</tr>
<tr>
<td>PlaNet (91M)</td>
<td><strong>8.4%</strong></td>
<td>24.5%</td>
<td><strong>37.6%</strong></td>
<td><strong>53.6%</strong></td>
<td><strong>71.3%</strong></td>
</tr>
</tbody>
</table>
Data gets little respect…

Data

Features

Learning Algorithm
Face Detection: Early Success Story

- Rowley, Baluja, and Kanade, 1998
  - features: **pixels**, classifier: **neural network**
- Schniderman & Kanade, 1999
  - features: **pairs of wavelet coeff.**, classifier: **naïve Bayes**
- Viola & Jones, 2001
  - features: **haar**, classifier: **boosted cascade**
Our Scientific Narcissism

All things being equal, we prefer to credit our own cleverness.
“Unreasonable Effectiveness of Data” [Halevy, Norvig, Pereira 2009]

• Parts of our world can be explained by elegant mathematics:
  – physics, chemistry, astronomy, etc.

• But some cannot:
  – psychology, biology, economics, AI, etc.

• That’s exactly where we need Magic of Data
Some things are just messy

Navier-Stokes Equation

\[
\frac{\partial \mathbf{u}}{\partial t} = - (\mathbf{u} \cdot \nabla) \mathbf{u} + v \nabla^2 \mathbf{u} - \frac{1}{d} \nabla p + \mathbf{f}
\]

+ weather
+ location
+ …
Brain-dead lookup (aka Nearest Neighbor) often works surprisingly well
Lots of Tiny Images

Lots

Of

Images

A. Torralba, R. Fergus, W.T.Freeman. PAMI 2008
Lots Of Images

A. Torralba, R. Fergus, W.T.Freeman. PAMI 2008
Lots Of Images
a) Input image

b) Neighbors
c) Ground truth
d) Wordnet voted branches
Automatic Colorization

Grayscale input High resolution

Colorization of input using average

A. Torralba, R. Fergus, W.T.Freeman. 2008
Why does it work?
Nearest neighbors from a collection of 20 thousand images
Nearest neighbors from a collection of 2 million images
im2GPS
(using 6 million GPS-tagged Flickr images)

Query Photograph

Im2gps [Hays & Efros, CVPR’08]
6 Million Flickr Images
im2GPS
(using 6 million GPS-tagged Flickr images)

Query Photograph

Visually Similar Scenes

Im2gps [Hays & Efros, CVPR’08]
The Good News

Really stupid algorithms + Lots of Data
= “Unreasonable Effectiveness”

[Halevy, Norvig, Pereira 2009]
But surely the brain can’t remember this much!?
What’s the Capacity of Visual Long Term Memory?

**What we know...**

Standing (1973)
10,000 images
83% Recognition

... people can remember thousands of images

**What we don’t know...**

... what people are remembering for each item?

According to Standing

“Basically, my recollection is that we just separated the pictures into distinct thematic categories: e.g. cars, animals, single-person, 2-people, plants, etc.) Only a few slides were selected which fell into each category, and they were visually distinct.”

Slide by Aude Oliva
Showed 14 observers 2500 categorically unique objects

1 at a time, 3 seconds each

800 ms blank between items

Study session lasted about 5.5 hours

Repeat Detection task to maintain focus

Followed by 300 2-alternative forced choice tests
Massive Memory Experiment I

A stream of objects will be presented on the screen for ~ 3 second each.

Your primary task:

Remember them ALL!

*afterwards you will be tested with...*

- Completely different objects...
- Different exemplars of the same kind of object...
- Different states of the same object...
Massive Memory Experiment I

Your other task: Detect exact repeats anywhere in the stream
Ready?

(Seriously, get ready to clap. The images go by fast...)
<clap!>
10 Minutes Later...
<clap!>
<clap!>
30 Minutes Later...
1 Hour Later...
<clap!>
2 Hours Later...
<clap!>
4 Hours Later...
<clap!>
5:30 Hours Later...
Which one did you see?

(go ahead and shout out your answer)
Recognition Memory Results

Replication of Standing (1973)

92%
Recognition Memory Results

Brady, et al. (2008), PNAS
So, do humans have “photographic memory”? 
Clap your hands when you see an image repeat

Ready?
<clap!>
<clap!>
<clap!>
<clap!>
<clap!>
<clap!>
Oliva, unpublished
This is where ConvNets come in

• Huge capacity
  – effectively unlimited?

• Learned feature space
  – focuses on what’s “important”
With huge capacity, huge temptations…

• ConvNets can memorize anything
  – e.g. “Rethinking generalization” [Zhang et al, 2017]

• ConvNets love to cheat…

• Semantic supervision might be to blame
classifiers love to cheat

Convolutional Neural Network

image $X$  \leftrightarrow  \text{“Shetland Sheepdog”}  \leftrightarrow  \text{label $Y$}
classifiers love to cheat

image $X$  

Convolutional Neural Network

“Shetland Sheepdog”

label $Y$

Gatys et al, 2017
example: action recognition in video

– input video:

– output: class label
  • Picking up cup
  • Slicing bread
  • Opening fridge
  • etc

example by David Fouhey
semantic supervision == memorization

We are raising a generation of algorithms who can only “cram for the test” (set)
Why do we have vision?

• “To see what is where by looking”
  – Aristotle, Marr, etc
• “To make babies who make babies, etc”
  – Darwin, Dawkins, etc.
Why do we have vision?

- “To see what is where by looking”
  - Aristotle, Marr, etc.
- “To predict the world”
  - Moshe Bar, Jan Koenderink, etc.
- “To make babies who make babies, etc”
  - Darwin, Dawkins, etc.
The world as supervision

Try to predict some aspect of the world that we interact with / have effect on:

– What’s gonna happen next?
– What’s to my left?
– What can I touch?
– What will make a sound?
– Etc.
Self-Supervision

Drawing Hands, M.C. Escher, 1948
Auto-encoders

Compressed image code (vector \( z \))
Auto-encoders

Compressed image code (vector $\mathbf{z}$)

$\mathbf{X}$

Image

$\hat{\mathbf{X}}$

Reconstructed image

[e.g., Hinton & Salakhutdinov, Science 2006]
Data compression

[Hinton & Salakhutdinov, Science 2009]
Data prediction

Some data $X_1$ → Other data $\hat{X}_2$
Self-Supervision in Multisensory Learning

Supervised
- implausible label

Unsupervised
- limited power

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

Virginia De Sa, “Learning classification from unlabelled data”, NIPS 1994
(Partial) Taxonomy of Self-Supervision

Data prediction

Transformation prediction

Meta-supervision

Adversarial, cycle, environment, etc.
Self-Supervision as Data Prediction

Data prediction

Data $x_0$ → Network → Data $x_1$
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]
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]
Ansel Adams, Yosemite Valley Bridge
Our result
Migrant Mother
Dorothea Lange
1936
Grayscale image: L channel

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

Semantics? Higher-level abstraction?

\[ \mathcal{F} \]

Information: ab channels

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

\[ L \rightarrow \mathcal{F} \rightarrow ab \]

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

[Zeiler & Fergus, ECCV 2014]
[Zhou et al., ICLR 2015]
Hidden Unit (conv5) Activations

sky

trees

water
Hidden Unit (conv5) Activations

faces

dog

faces

flowers
Self-supervision as data prediction

Zhang et al., ECCV 2016

Pathak et al., CVPR 2016

Zhang et al., CVPR 2017

Split-brain Auto-encoder
Self-supervision as transformation predication

Data $x_0$ \rightarrow Network \rightarrow T

Data $x_1$ \rightarrow Network \rightarrow T
Context Prediction for Images

Deorsch, Gupta, Efros ICCV 2015
Semantics from a non-semantic task
CNN Classifier

Patch Embedding

Input Nearest Neighbors

Correspondence *across* instances!
Image Splice Detection via Learned Self-Consistency

Minyoung Huh*1,2  Andrew Liu*1  Andrew Owens1  Alexei A. Efros1
UC Berkeley1  Carnegie Mellon University2
Learning same-image consistency

Same Imaging Pipeline?

Consistent

Inconsistent
Might be too slow…

Negative examples are too easy!
CameraMake: NIKON CORPORATION
CameraModel: NIKON D5300
ColorSpace: sRGB
ExifImageLength: 3947
ExifImageWidth: 5921
Flash: No
FocalLength: 31.0mm
WhiteBalance: Auto
CompressedBitsPerPixel: 2

CameraMake: EASTMAN KODAK COMPANY
CameraModel: KODAK EASYSHARE CX7300
ColorSpace: sRGB
ExifImageLength: 1544
ExifImageWidth: 2080
Flash: No (Auto)
FocalLength: 5.9mm
WhiteBalance: Auto
CompressedBitsPerPixel: 181/100
Learning EXIF self-consistency

CameraMake: Apple
CameraModel: iPhone 4s
ColorSpace: sRGB
ExifImageLength: 2448
ExifImageWidth: 3264
Flash: Flash did not fire
FocalLength: 107/2
WhiteBalance: Auto
ExposureTime: 1/2208

CameraMake: NIKON CORPORATION
CameraModel: NIKON D3200
ColorSpace: Uncalibrated
ExifImageLength: 2472
ExifImageWidth: 3091
Flash: Flash did not fire
FocalLength: 90
WhiteBalance: Auto
ExposureTime: 1/100

Same WhiteBalance?
Same vs. different EXIF attributes

…
Manipulated Image

Affinity Matrix

Normalized Cuts

Normalized Cuts

Manipulated Region
Manipulated Region

Our detection

Ground Truth
Our detection

Ground Truth

Manipulated Region
Input Image
Manipulated Region

Our detection

Ground Truth

Hays, Efros SIGGRAPH 2007
Authentic Images

Image Consistency Map
Failures

Image

Ground Truth

Consistency Map
Audio-Visual Scene Analysis with Self-Supervised Multisensory Features
ECCV 2018

Andrew Owens        Alexei A. Efros
UC Berkeley
Same audio, different video!
Multisensory representation
Idea #1: random pairs
Idea #2: time-shifted pairs
Idea #2: time-shifted pairs

Graph showing time vs. motion and loudness
Self-supervised Training

aligned vs. not-aligned

3D Convolution
3D Convolution
3D Convolution

3D Convolution

1D Convolution
1D Convolution
1D Convolution
Visualizing the location of sound sources

3D class activation map

3D Convolution

3D Convolution

3D Convolution

3D Convolution

1D Convolution

1D Convolution

1D Convolution

1D Convolution

1D Convolution
On/off-screen source separation
Meta-supervision

\[ F(x) = y \]
- direct supervision

\[ F(x) \in Y \]
- GANs

\[ G(F(x)) = x \]
- cycle-consistency

\[ \cdots \]
Dense Semantic Correspondence
Traditional Pairwise Methods

- SIFT flow: Liu et al., ECCV 2008
- Generalized PatchMatch: Barnes et al., ECCV 2010
- Deformable Spatial Pyramid: Kim et al., CVPR 2013
Collection Correspondence

- Congealing: Learned-Miller, PAMI 2006
- Collection Flow: Kramelmacher-Shlizerman et al., CVPR 2012
- Object discovery and segmentation: Rubinstein et al., CVPR 2013
- Compositional Image Model: Mobahi et al., CVPR 2014
- Object discovery and localization: Cho et al., CVPR 2015
- FlowWeb: T. Zhou et al., CVPR 2015
- Multi-image Matching: X. Zhou et al., ICCV 2015
Labels for CNN Training?

CNN

Infeasible to label in large-scale
Learning Dense Correspondence via 3D-guided Cycle Consistency

Tinghui Zhou\textsuperscript{1}, Philipp Krähenbühl\textsuperscript{1}, Mathieu Aubry\textsuperscript{2}, Qixing Huang\textsuperscript{3}, Alexei A. Efros\textsuperscript{1}

UC Berkeley\textsuperscript{1}, ENPC ParisTech\textsuperscript{2}, TTI-Chicago\textsuperscript{3}

CVPR 2016
Cycle-consistency as Supervision

• Composite flows along a cycle should be zero
Cycle-consistency as Supervision

- Composite flows along a cycle should be zero
- 2-cycle consistency: \( F_{i,j} \circ F_{j,i} = 0 \)
Cycle-consistency as Supervision

• Composite flows along a cycle should be zero
• 2-cycle consistency:  $F_{i,j} \circ F_{j,i} = 0$
• 3-cycle consistency:  $F_{i,k} \circ F_{k,j} \circ F_{j,i} = 0$
Cycle-consistency as Supervision

- Composite flows along a cycle should be zero
- 2-cycle consistency: $F_{i,j} \circ F_{j,i} = 0$
- 3-cycle consistency: $F_{i,k} \circ F_{k,j} \circ F_{j,i} = 0$
Cycle-consistency as Supervision

- Composite flows along a cycle should be zero
- 2-cycle consistency: $F_{i,j} \circ F_{j,i} = 0$
- 3-cycle consistency: $F_{i,k} \circ F_{k,j} \circ F_{j,i} = 0$
Could be consistent but wrong...

Need an anchor edge!
Synthetic Correspondence as the Anchor

3D CAD Model

Viewpoint
Renderer

Correspondence
from renderer
3D-guided Cycle Consistency

\[ \hat{F}_{s_1,s_2} = F_{s_1,r_1} \circ F_{r_1,r_2} \circ F_{r_2,s_2} \]
3D-guided Cycle Consistency

\[ \min \sum_{<s_1,s_2,r_1,r_2>} \mathcal{L} \left( \tilde{F}_{s_1,s_2} - F_{s_1,r_1} \circ F_{r_1,r_2} \circ F_{r_2,s_2} \right) \]
Network Architecture

Source

Weight Sharing

Target

Flow field
OUR RESULT
SIFTFLOW
CycleGAN, or “there and back aGAN”

[X]

[Y]

[G]

[F]

[D_X]

[D_Y]

[Zhu*, Park*, Isola, Efros. ICCV 2017]
Learning Correspondence from the Cycle-consistency of Time

Xiaolong Wang      Allan Jabri      Alexei Efros
CMU      UC Berkeley
Goal: Learn Correspondence without Human Supervision
The visual world exhibits continuity
Learning to Track

$\mathcal{F}$: a deep tracker
Supervision: Cycle-Consistency in Time

Track backwards

Track forwards, back to the future
Supervision: Cycle-Consistency in Time

Backpropagation through time along the cycle
Visualization of Training
Test Time: Nearest Neighbors in Feature Space $\phi$
Test Time: Nearest Neighbors in Feature Space $\phi$
Instance Mask Tracking

DAVIS Dataset

Pose Keypoint Tracking

JHMDB Dataset
Comparison

Our Correspondence

Optical Flow
## Pose Keypoint Tracking

### JHMDB Dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>PCK @.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optical Flow</td>
<td>45%</td>
</tr>
<tr>
<td>Vondrick et al.</td>
<td>45%</td>
</tr>
<tr>
<td>Ours</td>
<td>58%</td>
</tr>
</tbody>
</table>

Texture Tracking

DAVIS Dataset

Semantic Masks Tracking

Video Instance Parsing Dataset