It’s all about the Data!

Alexei A. Efros
UC Berkeley
“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.

Image Classification: 80%+ correct!
Planetary-scale geolocation

Automatic Image Captioning

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

“something is rotten in the state of Denmark”
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
```
<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>2.5%</td>
<td>12.0%</td>
<td>15.0%</td>
<td>23.0%</td>
<td>47.0%</td>
</tr>
<tr>
<td>Im2GPS (new) [20]</td>
<td></td>
<td>21.9%</td>
<td>32.1%</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>18.1%</td>
<td>30.0%</td>
<td>45.6%</td>
<td>65.8%</td>
</tr>
<tr>
<td>PlaNet (91M)</td>
<td><strong>8.4%</strong></td>
<td><strong>24.5%</strong></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
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
Automatic Colorization

Grayscale input High resolution

Colorization of input using average

A. Torralba, R. Fergus, W.T.Freeman. 2008
2 Million Flickr Images
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]
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-person, plants, etc.) Only a few slides were selected which fell into each category, and they were visually distinct.”

Slide by Aude Oliva

“Gist” Only
Sparse Details
Highly Detailed
Massive Memory I: Methods

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

Slide by Aude Oliva
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
<clap!>
<clap!>
10 Minutes Later...
<clap!>
<clap!>
30 Minutes Later...
1 Hour Later...
<clap!>
2 Hours Later...
4 Hours Later...
<clap!>
5:30 Hours Later...
Which one did you see?

(go ahead and shout out your answer)
Visual Cognition

Expert Predictions

- Recognition: 92%
- Memory Results: Replication of Standing (1973)

- Novel: 92%
- Exemplar
- State
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 $X$  

“Shetland Sheepdog”  

label $Y$
classifiers love to cheat

image X

Convolutional Neural Network

label Y

“Shetland Sheepdog”
example: action recognition in video

– input video:

– output: class label

• Picking up cup
• Slicing bread
• **Opening fridge**
• etc

easy 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 $z$)

X

Image
Auto-encoders

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

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

[Hinton & Salakhutdinov, Science 2009]
Data prediction

Some data $X_1$  \rightarrow  \text{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**
  - Input: Data $x_0$
  - Network
  - Output: Data $x_1$

- **Transformation prediction**
  - Input: Data $x_0$, Data $x_1$
  - Network
  - Output: Transformation $T$

- **Meta-supervision**
  - Input: Data $x_0$
  - Network
  - Output: Constraints on Data $x_1$

- Adversarial, cycle, environment, etc
Brains behind the research

Carl Doersch → DeepMind
Phillip Isola → MIT
Richard Zhang → Adobe
Jun-Yan Zhu → CMU
Andrew Owens → U Michigan
Deepak Pathak → CMU
Yu Sun
Allan Jarbi
Taesung Park
Tinghui Zhou
Shiry Ginosar
Deepak Pathak → CMU
Self-Supervision as Data Prediction

Data prediction

\[ \text{Data } x_0 \rightarrow \text{Network} \rightarrow \text{Data } x_1 \]
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

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

Color information: ab channels

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

[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

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

[Zhang, Isola, Efros, ECCV 2016]
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

Split-brain Auto-encoder

-Zhang et al, CVPR 2017

Pathak et al. CVPR 2016
Self-supervision as transformation predication

Data $x_0$ → Network → $T$

Data $x_1$
Context Prediction for Images

Deorsch, Gupta, Efros ICCV 2015
Semantics from a non-semantic task
Relative Position Task

- Randomly Sample Patch
- Sample Second Patch
- 8 possible locations

CNN -> Classifier
CNN CNN

Classifier

Patch Embedding

Input Nearest Neighbors

Correspondence across instances!
Audio-Visual Scene Analysis with Self-Supervised Multisensory Features

ECCV 2018

Andrew Owens        Alexei A. Efros

UC Berkeley
Multisensory representation

3D Convolution

1D Convolution
Idea #1: random pairs
Idea #2: time-shifted pairs
Idea #2: time-shifted pairs

Motion

Loudness

Time →
Self-supervised Training

aligned vs. not-aligned

3D Convolution
3D Convolution
3D Convolution

1D Convolution
1D Convolution
1D Convolution

3D Convolution
Visualizing the location of sound sources

- 3D class activation map
- 3D Convolution
- 3D Convolution
- 3D 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

- ...
CycleGAN, or “there and back aGAN”

[Zhu*, Park*, Isola, Efros. ICCV 2017]
Cycle-Consistency Loss

\[ \|F(G(x)) - x\|_1 \]
Cycle-Consistency Loss

\[||F(G(x)) - x||_1\]

\[||G(F(y)) - y||_1\]
Video
Domain Translation: CG to Real

Grand Theft Auto
CycleGAN: Real to CG
Failure case
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

passive self-supervision

• Good for discovering / calibrating sensing modalities

• But harder to push beyond low-level:
  – Easy things will eventually be learned
  – Hard things are too rare / complex to learn
  – Need a curriculum / teacher / adversary
active self-supervision

• World as adversary
• Competition between an active agent and the complexity of the world
• Will hopefully allow agent to learn increasingly complex things
• Emergent behaviors?
Reinforcement Learning in practice

Typically, RL requires very dense rewards.

- [ Mnih *et al.*, Nature 2015 ]
- [ Jaderberg *et al.*, ICLR 2017; Mirowski *et al.*, ICLR 2017 ]
Rewards are sparse in real world ...
Intrinsic Motivation / Curiosity

[The Scientist in a Crib, Gopnik, Meltzoff & Kuhl, 1999] [Classic definitions and new directions, Ryan et al., 2000] [Six Lessons from Babies, Smith & Gasser, 2005] [Curiosity and Motivation, Silvia, 2012]
Weaning Time (days)

Intelligence Measure

[S. T. Piantadosi, and C. Kidd. PNAS. 2016]
Curiosity-driven Exploration by Self-Supervised Prediction

Deepak Pathak, Pulkit Agrawal, Alexei Efros, Trevor Darrell

ICML 2017
Observation

Agent

Action

Pathak, Agrawal, Efros, Darrell. ICML'17
Agent

Observation

Action

“Down” has no effect
Pathak, Agrawal, Efros, Darrell. ICML'17
Curiosity $\equiv$ Prediction Error

Pathak, Agrawal, Efros, Darrell. ICML'17
Train a model

Predict consequences of the action

agent learns **continually** on its own...

Bad prediction $\rightarrow$ High Curiosity

Pathak, Agrawal, Efros, Darrell. ICML'17
No external reward, only curiosity

At the start of training

After curiosity-driven training
Do these skills generalize?

Trained on Level-1

Testing on Level-2
No external reward, only curiosity

Environment: Atari Games
Curiosity on both sides... makes a rally
Summary

- Magic is in the data
- Machine Learning methods love to cheat
- It’s worth thinking about tasks that discourage cheating, e.g. self-supervision
- And sometimes, not having a task is not a bad idea
Thank You

All models/code on GitHub