Visual Data on the Internet

With slides from James Hays, Antonio Torralba, Frederic Heger and Deepak Pathak

Cassandra Jones:
https://youtu.be/5H7WrlBrDRg?t=161

CS194: Image Manipulation & Computational Photography
Alexei Efros and Jun-Yan Zhu, UC Berkeley, Fall 2017
300 million images uploaded daily

300 hours uploaded every minute

13 billion images uploaded daily

1.5 million images uploaded daily

90% of net traffic will be visual!
Big issues

• What is out there on the Internet? How do we get it? What can we do with it?
• How do we compute distances between images?
Subject-specific Data

Photos of Coliseum

Portraits of Bill Clinton
Much of Captured World is “generic”
Generic Data

street scenes

Food plates

generic
data
pedestrians
The Internet as a Data Source

- Social Networking Sites (e.g. Facebook)
- Image Search Engines (e.g. Google, Bing)
- Photo Sharing Sites (e.g. Instagram, Flickr, Picasa, Panoramio, photo.net)
- Computer Vision Databases (e.g. CalTech 256, PASCAL VOC, LabelMe, Tiny Images, image-net.org, Places365, LSUN, etc.)
Is Generic Data useful?

A motivating example…
[Hays and Efros. Scene Completion Using Millions of Photographs. SIGGRAPH 2007 and CACM October 2008.]
Efros and Leung result
Scene Matching for Image Completion
The Algorithm
Scene Descriptor
Scene Descriptor

Scene Gist Descriptor (Oliva and Torralba 2001)
Scene Descriptor

Scene Gist Descriptor
(Oliva and Torralba 2001)
2 Million Flickr Images
... 200 total
Context Matching
Graph cut + Poisson blending
Result Ranking

We assign each of the 200 results a score which is the sum of:

The scene matching distance

The context matching distance (color + texture)

The graph cut cost
... 200 scene matches
Why does it work?
Nearest neighbors from a collection of 20 thousand images
Nearest neighbors from a collection of 2 million images
“Unreasonable Effectiveness of Data”  
[Halevy, Norvig, Pereira 2009]

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

• But much cannot
  – psychology, economics, genetics, etc.

• Enter The Data!
  – Great advances in several fields:
    • e.g. speech recognition, machine translation
    • Case study: Google
• A.I. for the postmodern world:
  – all questions have already been answered...many times, in many ways
  – Google is dumb, the “intelligence” is in the data
How about **visual** data?

- **Text is simple:**
  - clean, segmented, compact, 1D

- **Visual data is much harder:**
  - Noisy, unsegmented, high entropy, 2D/3D

Quick Overview

- Comparing Images
- Uses of Visual Data
- Deep Learning with Big Visual Data
- The Dangers of Data
Distance Metrics

\[ \text{Euclidian distance of 5 units} \]

\[ \text{Grayvalue distance of 50 values} \]

\[ \text{?} \]
SSD says these are not similar
Gist of a scene

• Need a full image descriptor, to capture the context

• But still want it to be not too high-dimentional (else nothing will look similar)
Make them tiny!
Tiny Images

- 80 million tiny images: a large dataset for non-parametric object and scene recognition
Tiny Images pack a punch!

4x4  8x8  16x16  32x32  64x64

Bedroom  Beach  Bedroom  Bedroom  Bedroom
c) Segmentation of 32x32 images
Image Segmentation (by humans)
Human Scene Recognition

Correct recognition rate vs Image resolution for color and grayscale images.

- Color image
- Grayscale

a) Scene recognition
Tiny Images Project Page

http://groups.csail.mit.edu/vision/TinyImages/
Powers of 10

Number of images on my hard drive: \(10^4\)

Number of images seen during my first 10 years: \(10^8\)
(3 images/second \(\times\) 60 \(\times\) 60 \(\times\) 16 \(\times\) 365 \(\times\) 10 = 630720000)

Number of images seen by all humanity: \(10^{20}\)
106,456,367,669 humans\(^1\) \(\times\) 60 years \(\times\) 3 images/second \(\times\) 60 \(\times\) 60 \(\times\) 16 \(\times\) 365 =

Number of photons in the universe: \(10^{88}\)

Number of all 32x32 images: \(10^{7373}\)
256 \(32^{32^3}\times\) \(\sim\) \(10^{7373}\)
Scenes are unique
But not all scenes are so original
But not all scenes are so original
Lots Of Images

7,900
Lots Of Images
Lots Of Images
Automatic Colorization Result

Grayscale input High resolution

Colorization of input using average

A. Torralba, R. Fergus, W.T.Freeman. 2008
Automatic Orientation

• Many images have ambiguous orientation
• Look at top 25% by confidence:
• Examples of high and low confidence images:
Automatic Orientation Examples

A. Torralba, R. Fergus, W.T. Freeman. 2008
Tiny Images Discussion

• Why SSD?
• Can we build a better image descriptor?
Image Representations: Histograms

global histogram
- Represent distribution of features
  - Color, texture, depth, …

Images from Dave Kauchak
Image Representations: Histograms

Joint histogram
- Requires lots of data
- Loss of resolution to avoid empty bins

Marginal histogram
- Requires independent features
- More data/bin than joint histogram
Adaptive binning

- Better data/bin distribution, fewer empty bins
- Can adapt available resolution to relative feature importance
Red Car Retrievals (Color histograms)

\[ \chi^2(h_i, h_j) = \frac{1}{2} \sum_{m=1}^{K} \frac{(h_i(m) - h_j(m))^2}{h_i(m) + h_j(m)} \]

Histogram matching distance
Image Representations: Clustering

Clusters / Signatures / “Visual Words”, Textons

- “super-adaptive” binning
- Does not require discretization along any fixed axis
Capturing the “essence” of texture

...for real images

We don’t want an actual texture realization, we want a texture invariant

What are the tools for capturing statistical properties of some signal?
Multi-scale filter decomposition

Filter bank

Input image
Filter response histograms
Start with a noise image as output

Main loop:

- Match pixel histogram of output image to input
- Decompose input and output images using multi-scale filter bank (Steerable Pyramid)
- Match subband histograms of input and output pyramids
- Reconstruct input and output images (collapse the pyramids)
Image Descriptors (more later)

- Blur + SSD
- Color / Texture histograms
- Gradients + Histogram (GIST, SIFT, HOG, etc)
- “Bag of Visual Words”
Gist Scene Descriptor
Gist Scene Descriptor

Gist scene descriptor
(Oliva and Torralba 2001)

Hays and Efros, SIGGRAPH 2007
Gist Scene Descriptor

Gist scene descriptor
(Oliva and Torralba 2001)

Hays and Efros, SIGGRAPH 2007
Recap: Using lots of data!

Trick: If you have enough images, the dataset will contain very similar images that you can find with simple matching methods.
im2gps (Hays & Efros, CVPR 2008)

6 million geo-tagged Flickr images
How much can an image tell about its geographic location?
Example Scene Matches
Voting Scheme
im2gps
Data-driven categories
Population density ranking
Figure 4. Global land cover classification map.
Barren or sparsely populated
Urban and built up
Savannah
Water
Using Data for Image Creation...
Semantic Photo Synthesis [EG’06]

Semantic Photo Synthesis
Photo Clip Art [SG’07]

Inserting a single object -- still very hard!

Lalonde et al, SIGGRAPH 2007
Photo Clip Art [SG’07]

Use database to find well-fitting object

Lalonde et al, SIGGRAPH 2007
Camera parameters

Assume

- flat ground plane
- all objects on ground
- camera roll is negligible (consider pitch only)

Camera parameters: height and orientation
Bad sketch! + Huge Dataset → Images, Associated Info

image matching → Contours from Similar Images
ShadowDraw

http://www.youtube.com/watch?v=zh_HUdQwow
Explore Visual Data
AverageExplorer

http://www.youtube.com/watch?v=1QgL_aPPCpM
Internet Data + Deep Learning
Context Encoders: Feature Learning by Inpainting

Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, Alexei Efros, CVPR 2016
Image Painting with Deep Learning
Context Encoders

- Encoder can be substituted with any network architecture like AlexNet etc.
- Decoder is a set of UpConv/deconv/frac-strided-conv layers
Inpainting Generation with L2 Loss

- Model learns rough structure of the scene but output space is an average (blurry) image.
Nearest Neighbors with L2 Loss

- Nearest Neighbors are good in the representation space
- But generated output is blurry => Artifact of L2!!
**Need Mode Picking => GAN**

G tries to synthesize fake images that fool D

D tries to identify the fakes

---

[Goodfellow et al., 2014]
Network Architecture

- Surrounding pixel -> center region
- Trained on 1m ImageNet images.
Decoupled Loss

- Reconstruction $L_2$ loss ensures "correctness"
- Adversarial Loss ensures "realness"
Indirect Conditioning Results
Indirect Conditioning Results
Indirect Conditioning Results
More Results
More Results
Compare w/ Content Aware Fill
(algorithm in Adobe Photoshop)
Open Source

[GitHub](https://github.com/pathak22/context-encoder)
Interactive Completion Application

• Real-time with GPUs.

• http://hi.cs.waseda.ac.jp/~iizuka/projects/completion/en/
The Dangers of Data
Bias

- Internet is a tremendous repository of visual data (Flickr, YouTube, Picasssa, etc)
- But it’s **not** random samples of visual world
- Many sources of bias:
  - Sampling bias
  - Photographer bias
  - Social bias
Flickr Paris
Real Paris
Real Notre Dame
Sampling Bias

- People like to take pictures on vacation
Sampling Bias

People like to take pictures on vacation
Photographer Bias

• People want their pictures to be recognizable and/or interesting

VS.
Photographer Bias

- People follow photographic conventions

VS.
Social Bias

Mildred and Lisa

Source: U.S. Social Security Administration

Gallagher et al CVPR 2008
Social Bias

“100 Special Moments” by Jason Salavon
Social Bias

Gallagher et al, CVPR 2008

Gallagher et al, CVPR 2009
Reducing / Changing Bias

- Autonomous capture methods can reduce / change bias
  - But it won’t go away completely
- Sometimes you can just pick your data to suit your problem, but not always…
Thank You.

• Prof. Efros will come back on Thursday!
  – Project awards.

• Project 5 is due tonight!

• Project 6 will be out tomorrow.
Backup Slides
Results: Decoupled Loss

Input Image  L2 Loss  Adversarial Loss  Joint Loss
ImageNet Results

Has artifacts near the boundary!
Boundary Artifacts

- Discriminator is not conditioned on input!
Indirect Conditioning w/ Context

Input Area = $I$
Target Area = $O$
Indirect Conditioning w/ Context

Input Area = $I$
Target Area = $O$
Predict Area = $O_{extra}$
Overlap = $O_{extra} - O$

Penalize L2 loss heavily in overlapping area
=> so that network memorizes input in the overlap
Penalize L2 loss heavily in overlapping area
=> so that network memorizes input in the overlap

**INDIRECT CONSISTENCY**

- Adversary ensures $O_{extra}$ is coherent (to be realistic)
- L2 Penalty ensures predicted $O - O_{extra}$ is same as overlapping area with input $I$

$\Rightarrow$ Predicted $O$ is coherent with $I$