ConvNets as a Versatile Tool

CS194: Intro to Comp. Vision and Comp. Photo
Alexei Efros, UC Berkeley, Fall 2021
Project 2 winner!
Classification Loss (e.g. ImageNet)

\[
\begin{align*}
\text{P(image is class \#1)} & \\
\text{P(image is class \#2)} & \\
\text{P(image is class \#F)}
\end{align*}
\]
Regression Loss (e.g. Project 5)

\[
L_2(\hat{Y}, Y) = \frac{1}{2} \sum_{h,w} \| Y_{h,w} - \hat{Y}_{h,w} \|^2
\]

H W C

\[ \rightarrow \]

CNN

\[ \rightarrow \]

1 1 F

Keypoint #1 X
Keypoint #1 Y
Keypoint #2 X
Convert HxW image into a F-dimensional vector

Which pixels in this image are a cat?  
How far is each pixel away from the camera?  
Which pixels of this image are fake?
“Semantic Segmentation”

Each pixel has label, inc. **background**, and **unknown**.
Usually visualized by colors.

Note: don’t distinguish between object **instances**

Encoder-Decoder

Key idea: First **downsample** towards middle of network. Then **upsample** from middle.

**How do we downsample?**
Convolutions, pooling
Missing Details

While the output is HxW, just upsampling often produces results without details/not aligned with the image.

Why?

Information about details lost when downsampling!

Result from Long et al. *Fully Convolutional Networks For Semantic Segmentation*. CVPR 2014
Missing Details

Where is the useful information about the high-frequency details of the image?

Result from Long et al. *Fully Convolutional Networks For Semantic Segmentation*. CVPR 2014
How do you send details forward in the network?
You copy the activations forward.
Subsequent layers at the same resolution figure out how to fuse things.

Result from Long et al. *Fully Convolutional Networks For Semantic Segmentation*. CVPR 2014
Extremely popular architecture, was originally used for biomedical image segmentation.
e.g. Depth Prediction

Instead: give label of depthmap, train network to do regression (e.g., $\|z_i - \hat{z}_i\|$ where $z_i$ is the ground-truth and $\hat{z}_i$ the prediction of the network at pixel $i$).

Input HxWx3
RGB Image

Output HxWx1
Depth Image

True HxWx1
Depth Image
Surface Normals

\[ \mathbf{n} = [n_x, n_y, n_z], \| \mathbf{n} \| = 1 \]

Legend

Image credit: NYU Dataset, Silberman et al. ECCV 2012
Surface Normals

Instead: train normal network to minimize $\|n_i - \widehat{n}_i\|$ where $n_i$ is ground-truth and $\widehat{n}_i$ prediction at pixel $i$.

Input: HxWx3 RGB Image

Output: HxWx3 Normals

Result credit: X. Wang, D. Fouhey, A. Gupta, Designing Deep Networks for Surface Normal Estimation. CVPR 2014
Generic Image-to-Image Translation

- Labels to Street Scene
- Labels to Facade
- BW to Color
- Aerial to Map
- Day to Night
- Edges to Photo
Grayscale image: $L$ channel

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

\[ \mathcal{F} \]

Color information: $ab$ channels

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

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

Grayscale image: $L$ channel

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

Concatenate $(L, ab)$ channels

$(X, \hat{Y})$

Simple L2 regression doesn’t work 😞

L_2(\hat{Y}, Y) = \frac{1}{2} \sum_{h,w} \| Y_{h,w} - \hat{Y}_{h,w} \|_2^2

Slide by Richard Zhang
Slide by Richard Zhang

\[ L_2(\hat{Y}, Y) = \frac{1}{2} \sum_{h,w} \| Y_{h,w} - \hat{Y}_{h,w} \|_2^2 \]
Better Loss Function

\[ \theta^* = \arg\min_{\theta} \ell(\mathcal{F}_\theta(X), Y) \]

- Regression with L2 loss inadequate
  \[ L_2(\hat{Y}, Y) = \frac{1}{2} \sum_{h,w} \| Y_{h,w} - \hat{Y}_{h,w} \|_2^2 \]

- Use per-pixel multinomial classification
  \[ L(\hat{Z}, Z) = -\frac{1}{HW} \sum_{h,w} \sum_q Z_{h,w,q} \log(\hat{Z}_{h,w,q}) \]
Designing loss functions

Input  | Zhang et al. 2016  | Ground truth

Color distribution cross-entropy loss with colorfulness enhancing term.

[Zhang, Isola, Efros, ECCV 2016]
Designing loss functions

Image colorization

Cross entropy loss, with colorfulness term

Super-resolution

“semantic feature loss” (VGG feature covariance matching objective)
Universal loss?
Generative Adversarial Network (GANs)

Real photos vs Generated images

[Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville, Bengio 2014]
Generator

[Goodfellow et al., 2014]
\( \mathbf{G} \) tries to synthesize fake images that fool \( \mathbf{D} \)

\( \mathbf{D} \) tries to identify the fakes

[Goodfellow et al., 2014]
\[
\arg \max_D \mathbb{E}_{x,y} \left[ \log D(G(x)) + \log(1 - D(y)) \right]
\]

[Goodfellow et al., 2014]
$\mathbf{G}$ tries to synthesize fake images that fool $\mathbf{D}$:

$$\arg\min_{\mathbf{G}} \mathbb{E}_{\mathbf{x},\mathbf{y}} \left[ \log D(G(\mathbf{x})) + \log(1 - D(\mathbf{y})) \right]$$

[Goodfellow et al., 2014]
\[ \text{arg min}_G \text{max}_D \mathbb{E}_{x,y} \left[ \log D(G(x)) + \log(1 - D(y)) \right] \]

\[ \text{G tries to synthesize fake images that fool the best D:} \]

[Goodfellow et al., 2014]
G’s perspective: D is a loss function.

Rather than being hand-designed, it is learned.

[Goodfellow et al., 2014]
[Isola et al., 2017]
arg min_{G} \max_{D} \mathbb{E}_{x,y} \left[ \log D(G(x)) + \log(1 - D(y)) \right]

[Goodfellow et al., 2014]
\[
\arg\min_G \max_D \mathbb{E}_{x,y} \left[ \log D(G(x)) + \log(1 - D(y)) \right]
\]

[Goodfellow et al., 2014]
\[
\arg \min_G \max_D \mathbb{E}_{x,y} \left[ \log D(G(x)) + \log(1 - D(y)) \right]
\]

[Goodfellow et al., 2014]
[Isola et al., 2017]
\[
\text{arg min}_G \max_D \mathbb{E}_{x,y} \left[ \log D(x, G(x)) + \log(1 - D(x, y)) \right]
\]

[Goodfellow et al., 2014]
[Isola et al., 2017]
arg min_G max_D \mathbb{E}_{x,y} [ \log D(x, G(x)) + \log(1 - D(x, y)) ]

[Goodfellow et al., 2014]
[Isola et al., 2017]
\[
\arg\min_G \max_D \mathbb{E}_{x,y} \left[ \log D(x, G(x)) + \log(1 - D(x, y)) \right]
\]

[Goodfellow et al., 2014]
[Isola et al., 2017]
$$\arg\min_G \max_D \mathbb{E}_{x,y} \left[ \log D(x, G(x)) + \log(1 - D(x, y)) \right]$$

[Goodfellow et al., 2014]
[Isola et al., 2017]
BW → Color

Data from [Russakovsky et al. 2015]
Data from [maps.google.com]
Labels → Facades

Data from [Tylecek, 2013]
Labels $\rightarrow$ Facades

Data from [Tylecek, 2013]
Day → Night

Data from [Laffont et al., 2014]
Thermal → RGB

Input  Ground-truth  Output
Edges $\rightarrow$ Images

Edges from [Xie & Tu, 2015]
Sketches $\rightarrow$ Images

Trained on Edges $\rightarrow$ Images

Data from [Eitz, Hays, Alexa, 2012]
Image-to-image translation in PyTorch (e.g., horse2zebra, edges2cats, and more)

- pytorch
- gan
- cycleGAN
- pix2pix
- deep-learning
- computer-vision
- computer-graphics
- image-manipulation
- image-generation
- generative-adversarial-network
- gans

- 223 commits
- 3 branches
- 0 releases
- 26 contributors

Commit Log:
- taesung89 Update README.md
  - Latest commit 69e1730
  - Data:
    - datasets: Multiple changes regarding option management. See below.
    - imgs: add edges2cats demo
    - models: TestModel now supports model_suffix option that can change the name of the model.
Twitter-driven research: #pix2pix

Brannon Dorsey @brannondorsey

Mario Klingemann @quasimondo

Bertrand Gondouin @bgondouin
Scott Eaton (http://www.scott-eaton.com/)
“Do as I Do”
Everybody Dance Now

Caroline Chan, Shiry Ginosar, Tinghui Zhou, Alexei A. Efros
UC Berkeley
Results

Source Subject
*Challenging due to missed detections

https://www.youtube.com/watch?v=PCBTZh41Ris&feature=youtu.be
CEO: our own Dr. Tinghui Zhou
Paired training examples

\{ \}
\{ \}
\{ \}
\{ \}
\{ \}
\{ \}

Unpaired training examples

\[ X \]
\[ \]
\[ \]
\[ \]
\[ \]

\[ Y \]
\[ \]
\[ \]
CycleGAN, or “there and back aGAN”

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

\[ D_Y \]

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

\[ \| F(\mathbf{G}(\mathbf{x})) - \mathbf{x} \|_1 \]

\[ \| \mathbf{G}(F(\mathbf{y})) - \mathbf{y} \|_1 \]
Video
Collection Style Transfer

Photograph
© Alexei Efros

Monet

Van Gogh

Cezanne

Ukiyo-e
CG to Real

Grand Theft Auto
Real to CG
Shallower depth of field
Failure case
Learning an Embedding

![Diagram showing two CNN models connected by a shared representation and a shared weight matrix W.](image-url)
Learning an Embedding

Matching

shared representation

CNN

Shared W

CNN

Image 1

Image 2
Siamese Network w/ Contrastive Loss

Siamese Architecture
[Chopra 2005, Hadsell 2006]
LEARNING VISUAL SIMILARITY FOR PRODUCT DESIGN WITH CONVOLUTIONAL NEURAL NETWORKS

SEAN BELL AND KAVITA BALA
CORNELL UNIVERSITY
THE PROBLEM

(1) “What is this?”

(2) “Where is it used?”

Name: ”Great Bowl O’Fire Sculptural Fire Bowl”

Category: Fire pit

Sold by: John T. Unger, LLC
THE PROBLEM

(1) “What is this?”

(2) “Where is it used?”

Name: ”Great Bowl O’Fire Sculptural Fire Bowl”
Category: Fire pit
Sold by: John T. Unger, LLC

Challenge: determine whether these are the same product (different resolution, viewpoint, color, lighting, occlusions)
TWO KINDS OF IMAGES

Iconic

(From a product website)

In context

(Cropped from a scene photo)
PROJECTING INTO A JOINT EMBEDDING

Iconic

In context

Project

Embedding
SEARCH USING THE EMBEDDING

“What is it?”

Retrieval

Project

Embedding

Name: Hemel Ring
Category: Hanging light
Sold by: Holly Hunt
SEARCH USING THE EMBEDDING

“Where is it used?”

Project

Retrieve

Embedding
CONTRASTIVE LOSS: POSITIVE EXAMPLE

In context

Iconic (same)

Parameters $\theta$

CNN

$\mathbf{X}_p$

$\mathbf{X}_q$

Loss $L_p$

Embedding

$L_p(\mathbf{x}_q, \mathbf{x}_p) = \|\mathbf{x}_q - \mathbf{x}_p\|_2^2$
CONTRASTIVE LOSS: NEGATIVE EXAMPLE

In context

Iconic (different)

Loss $L_n$

Margin $m$

Embedding

$$L_n(x_q, x_n) = \max (0, m^2 - ||x_q - x_n||_2^2)$$
CONTRASTIVE LOSS: **ALL TOGETHER**

\[ L(\theta) = \sum_{(x_q, x_p)} L_p(x_q, x_p) + \sum_{(x_q, x_n)} L_n(x_q, x_n) \]

- **Penalty for similar images that are far away**
  \[ L_p(x_q, x_p) = \|x_q - x_p\|_2^2 \]

- **Penalty for dissimilar images that are nearby**
  \[ L_n(x_q, x_n) = \max(0, m^2 - \|x_q - x_n\|_2^2) \]

Minimize \( L(\theta) \) with **stochastic gradient descent and momentum**

[Chopra 2005, Hadsell 2006]
TRAINING PIPELINE

Image pairs → Stochastic Gradient Descent

Image database → CNN

\( \theta \) → CNN Parameters → Embedding
RESULTS: “WHAT IS IT?”

In context
RESULTS: “WHAT IS IT?”

In context

Iconic

Top 4 results:
RESULTS: “WHAT IS IT?”

In context
RESULTS: “WHAT IS IT?”

In context

Iconic

Top 4 results:
COMPARISON: TRAINED ONLY ON CATEGORIES

In context

Iconic

Top 4 results:

Tiara Oval Suspension by Harco Loor
- Dining Room
- Contemporary
- 1 View jpg
- Wall Sconces
- Unknown style
- 1 View jpg
- 2 boxes

'Apollon' Black Shaded 6-light Crystal Chandelier
- Chandeliers
- Contemporary
- View jpg
- View detail

Authentic Deer and Elk Antler Banquet Table Chandelier
- Chandeliers
- Unknown style
- View jpg
- View detail

Harco Loor | Tiara Oval HL 15 Suspension Light
- Pendant Lighting
- Modern
- View jpg
- View detail

Vermeer Hexagonal Pendant
- Pendant Lighting
- Modern
- View jpg
- View detail

1792.880
1895.330
1911.520
2035.000
COMPARISON: TRAINED ONLY ON IMAGENET

In context

Iconic

Top 4 results:

1. Tiara Oval Suspension by Harco Loor
   - Artwork
   - Contemporary
   - Dining Room
   - Wall Sconces
   - Unknown style
   - 2 boxes

2. "Elephant View" Artwork
   - Artwork
   - Contemporary
   - Paintings
   - View jpg

3. Terzani Argent N925 Chandelier
   - Chandeliers
   - Contemporary
   - Modern
   - View jpg

4. Mensa Hanging Light Fixture
   - Chandeliers
   - Contemporary
   - View jpg
RESULTS: FAILURE CASE
RESULTS: FAILURE CASE

In context

Iconic

Top 4 results:

Porta Romana Bianca Pendant Light
- Staircase
- Transitional
- 1 View jog
- Pendant Lighting
- Contemporary
- View jog
- 2 boxes

Murray Feiss Khloe Modern Contempo Mini Pendant
- Pendant Lighting
- Transitional
- 2 View jog

Uttermost Metauro Wood Desk Lamp
- Desk Lamps
- Transitional
- View jog

Cycling Blur Wall Mural - 42 Inches W x 40 Inches H
- Prints And Posters
- Contemporary
- View jog

Metauro Wood Desk Lamp
- Desk Lamps
- Contemporary
- View jog
RESULTS: “WHERE IS IT USED?”

“Maskros Pendant Lamp”
RESULTS: “WHERE IS IT USED?”

"LEM Piston Stool | Design Within Reach"
<table>
<thead>
<tr>
<th>Query $I_q$</th>
<th>Dining chairs</th>
<th>Armchairs</th>
<th>Rocking chairs</th>
<th>Bar stools</th>
<th>Table lamps</th>
<th>Outdoor lighting</th>
<th>Bookcases</th>
<th>Coffee tables</th>
<th>Side tables</th>
<th>Floor lamps</th>
<th>Rugs</th>
<th>Wallpaper</th>
</tr>
</thead>
</table>

**Top-1 nearest neighbor from different object categories**