We’re looking for interns!

sbell@grokstyle.com
THE PROBLEM: “WHAT IS IT?”

Can we answer this automatically?

Michelle Melwani_Mehra wrote:

What is the countertop? Thank you!

kammererk wrote:

Can someone tell me the manufacturer/style of the range hood. Thanks.

sjheff1 wrote:

What is the color of the subway tile you used in this kitchen? Thank You.

Can we answer this automatically?
THE PROBLEM: "WHERE IS IT USED?"

Tundra 71" Dining Table W/ Extension, Chocolate /CHOCOLATE
4 Questions | 808 Saves

$1,560
Only 2 Left!
Quantity: 1

Product Description:
For the happy dining room, the Tundra extensible table is the perfect match. Featuring the possibility of exchanging the colors of its center piece, it's visual appeal is unmatched. Whatever the size of your diner party, this table can adapt. 4

leskap5 wrote:
great table can you reccomend chairs?
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?”

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

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

Category: Fire pit

Sold by: John T. Unger, LLC
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

Name: Hemel Ring
Category: Hanging light
Sold by: Holly Hunt

“What is it?”

Retrieval

Project

Embedding
SEARCH USING THE EMBEDDING

“Where is it used?”

Project

Retrieve

Embedding
APPROACH 1: DIRECTLY SOLVING FOR EMBEDDING

Learn the positions of all images directly from triplet judgements

Need triplet judgements to project new items

Generalized Nonmetric Multidimensional Scaling
[Agarwal et al 2007]
**APPROACH 2: METRIC LEARNING**

Learn a distance function over features
[Chechik 2010, Garces 2014, O’Donovan 2014, …]

Image $\rightarrow$ Extract Features $\rightarrow$ Project to Embedding $\rightarrow$ Embedding position $x$

(Designed, Fixed) (Learned)

A Similarity Measure for Illustration Style [Garces 2014]
APPROACH 3: DEEP LEARNING (END-TO-END)

Learn end-to-end: a function from image to embedding
DEEP LEARNING HAS TAKEN OVER

Image Classification [Szegedy 2014]  
Video Classification [Karpathy 2014]

Object Detection [Girshick 2015]  
Image Captioning [Vinyals 2015]
DEEP LEARNING HAS TAKEN OVER

Semantic Segmentation [Long 2014]

Color Constancy [Barron 2015]

Surface Normals [Wang 2014]

People Counting [Stewart 2015]
DEEP LEARNING HAS TAKEN OVER

Playing video games [Minh 2015]

Generating house numbers [Gregor 2015]
OVERVIEW

(1) Training the CNN

(2) Gathering training data

(3) Search Results
CONVOLUTIONAL NEURAL NETWORKS (CNN)

“LeNet”
Backpropagation applied to handwritten zip code recognition
[Lecun 1989]

“AlexNet”
ImageNet Classification with Deep Convolutional Neural Networks
[Krizhevsky 2012]

“GoogLeNet”
Going deeper with convolutions
[Szegedy 2014]
CONVOLUTIONAL NEURAL NETWORKS (CNN)

Parameters $\theta$

Embedding position $x \times (256D)$

"GoogLeNet" [Szegedy 2014]
SIAMESE NETWORK AND CONTRASTIVE LOSS

Siamese Architecture
[Chopra 2005, Hadsell 2006]
CONTRASTIVE LOSS: POSITIVE EXAMPLE

In context

Iconic (same)

Loss $L_p$

Embedding

$$L_p(x_q, x_p) = \|x_q - x_p\|_2^2$$
CONTRASTIVE LOSS: NEGATIVE EXAMPLE

In context

Iconic (different)

Loss $L_n$

Margin $m$

Embedding

Parameters $\theta$

$\text{CNN}$

$\text{CNN}$

$\text{CNN}$

$\text{CNN}$

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

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

- Penality for similar images that are far away
- Penality for dissimilar images that are nearby

\[ L_p(x_q, x_p) = \|x_q - x_p\|^2 \]

Margin

\[ L_n(x_q, x_n) = \max(0, m^2 - \|x_q - x_n\|^2) \]

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

[Chopra 2005, Hadsell 2006]
Use object categories to train CNN: helps regularize the CNN representation

Why? We have 3M iconic images but only 100k products in context

Categories give an additional training signal to the remaining 2.9M iconic images

(839 categories, e.g. “Accent chair”, “Pendant Light”, …)
TRAINING: FULL SYSTEM

- Parameters $\theta$
- CNN
- L2
- Contrastive Loss
- Softmax Loss
- Loss

- Category $C_q$
- Embedding $x_q$
- Category $C_n$
- Embedding $x_n$
TRAINING: FULL SYSTEM

Abstract away the complexity:
Write as a large graph (DAG) and compute the gradient with the back-propagation algorithm [Rumelhart et al. 1986]
TRAINING PIPELINE

Image pairs

Image database

Stochastic Gradient Descent

θ

CNN Parameters

Embedding

CNN Parameters

θ
OVERVIEW

(1) Training the CNN

(2) Gathering training data

(3) Search Results
Some scenes (~2%) have product tags.
(a) Full scene
(b) Iconic product images

Problem: what is the spatial extent of each tag?
SPATIAL EXTENT: CROWDSOURCING ON MECHANICAL TURK

- Users are shown the iconic image and tag in context
- Users draw a bounding circle to set the zoom level (1 click)
- Users draw a bounding box (4 clicks)
- Can label “mismatch”
SPATIAL EXTENT: QUALITY CONTROL

Duplication:

• Multiple workers draw boxes
• Until any 2 agree with IOU > 0.7

Sentinels:

• Secret test items (1 out of 7)
• Block low accuracy workers

IOU(A, B) = \frac{\text{Intersection}(A,B)}{\text{Union}(A, B)}
CROWDSOURCING PIPELINE

(1) Browse houzz.com, remove near-duplicates:

3M products images
180k product tags

(2) Crowdsourcing on Mechanical Turk:

100k boxes with high agreement

(3) Sample image pairs for training:
For each box:

- Add 1 positive pair: *(in-context box, tagged iconic image)*
- Add 20 negative pairs: *(in-context box, random other iconic image)*

80% from same category, 20% from different category
CROPPING IMAGES FOR CNN INPUT

Original image

Crop with padding (repeat with different amounts)

Warp to square (256x256)

Random 224x224 sub-crops
OVERVIEW

(1) Training the CNN

(2) Gathering training data

(3) Search Results
Train the CNN for 200,000 mini-batches…

(go for a very long walk and come back)

Runtime: milliseconds
EMBEDDING VISUALIZED IN 2D: CHAIRS

- Eclectic
- Office Chairs
- Sofas
- Leather
- Thin legs
RESULTS

(1) “What is it?”

(2) “Where is it used?”

(3) Cross-category search

(4) Quantitative Evaluation
RESULTS: "WHAT IS IT?"
RESULTS: “WHAT IS IT?”

In context

Iconic

Top 4 results:
RESULTS: “WHAT IS IT?”
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

To the left is a picture labeled ‘In context’, which shows a lighting fixture in an actual setting. To the right, the ‘Iconic’ category is displayed with a picture of a lighting fixture, followed by the ‘Top 4 results’ section:

1. **Tiara Oval Suspension by Harco Loor**
   - Tags: Dining Room, Contemporary, View jpg, Wall Sconces, Unknown style
   - View detail
   - Price: 1,792.880

2. **'Apollon' Black Shaded 6-light Crystal Chandelier**
   - Tags: Chandeliers, Contemporary, View jpg
   - View detail
   - Price: 1,895.330

3. **Authentic Deer and Elk Antler Banquet Table Chandelier**
   - Tags: Chandeliers, Unknown style, View jpg
   - View detail
   - Price: 1,911.520

4. **Vermeer Hexagonal Pendant**
   - Tags: Pendant Lighting, Modern, View jpg, View detail
   - Price: 2,035.000
**COMPARISON:** TRAINED ONLY ON IMAGENET

**In context**

**Iconic**

**Top 4 results:**

1. **Tiara Oval Suspension by Harco Loor**
   - **Artwork**
   - **Contemporary**
   - **View jpg**
   - **2 boxes**

2. **MERGING TOGETHER**
   - **Artwork**
   - **Contemporary**
   - **View jpg**
   - **View detail**

3. **“Elephant View” Artwork**
   - **Artwork**
   - **Contemporary**
   - **View jpg**
   - **View detail**

4. **Terzani Argent N925 Chandelier**
   - **Chandeliers**
   - **Contemporary**
   - **View jpg**
   - **View detail**

5. **Mensa Hanging Light Fixture**
   - **Chandeliers**
   - **Modern**
   - **View jpg**
   - **View detail**
RESULTS: FAILURE CASE

In context
RESULTS: FAILURE CASE

In context

Iconic

Top 4 results:
RESULTS: “WHERE IS IT USED?”

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

“LEM Piston Stool | Design Within Reach”
SEARCHING ACROSS CATEGORIES

<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>
How do we evaluate the results?

In the list of 3M products, what is the rank of the iconic image?
How do we evaluate the results?

Even when the top product is correct, the rank may be large.

Rank: 440 / 3,387,555 (top 0.01%)
QUANTITATIVE EVALUATION

Variations of training with a Siamese architecture

Baselines [Razavian 2014]

Baselines [Azizpour 2014]

Random guessing
ABLATION STUDY

L2 Normalization

Training with categories

Siamese network

Variations of training with a Siamese architecture

Baselines [Razavian 2014]

Baselines [Azizpour 2014]

Random guessing

Mean recall @ k

Top k (log scale)
LIMITATIONS AND FUTURE WORK

• Better metrics to evaluate performance (e.g. measure precision)
• Efficiently collecting more training data (we are limited by what is tagged)
• Reduce crowdsourcing costs (less duplication)
• Explicitly model and train for style similarity/compatibility
CONTRIBUTIONS

Visual search for products: “what is it?” and “where is it used?”

• Crowdsourced pipeline to collect image pairs across domains
• Use deep learning to learn a high quality embedding
• Applications in interior design and product search
THANK YOU

Acknowledgements:

- NSERC PGS-D
- National Science Foundation
- NVIDIA
- Intel Science and Technology Center for Visual Computing
- Amazon AWS for Education
- Houzz users
We’re looking for interns!

sbell@grokstyle.com
USING MORE DIMENSIONS PERFORMS THE SAME

4096 dimensions
256 dimensions
(16x faster for almost the same performance)
ARCHITECTURE VARIATIONS

(Our system)

[Chrop 2005, Hadsell 2006, ...]

[Weston 2008]

[Razavian 2014, ...]