

# CSC2523: Visual Recognition with Text Introduction

Sanja Fidler

January 14, 2015



UNIVERSITY OF  
**TORONTO**

- **Instructor:**



Sanja Fidler (`fidler@cs.toronto.edu`)

- **Office:** 283B in Pratt
- **Office hours:** Send email for appointment

# Course Information

- **Class time:** Wednesdays at 11am-1pm

- **Location:** HA 410 (Haultain Bldg)

- **Class Website:**

<http://www.cs.utoronto.ca/~fidler/CSC2523.html>

- The class will use Piazza for **announcements** and **discussions**:

<https://piazza.com/utoronto.ca/winter2015/csc2523/home>

- Your grade will **not depend on your participation on Piazza**

- No textbook, so class attendance is encouraged

# Course Prerequisites

## Good to know:

- Computer Vision
- Machine Learning
- Basic NLP

Without this you'll need some catching up to do

- Today we'll do a quick high-level overview of visual recognition topics and techniques with links to papers
- Next lecture will be given by Mohit Bansal covering the basics of NLP and discussing current topics and trends that may be useful for joint image and text modeling

# Requirements and Grading

- This course is a seminar course. We'll be reading papers on diverse topics in the domain of images and text. Thus, how much you learn greatly depends on how prepared everyone comes to class.
- Each student expected to write short reviews of two papers we'll be reading each week, present two papers, and do a project
- **Grading**
  - Participation (attendance, participation in discussions, reviews): 25%
  - Presentation (presentation of papers in class): 35%
  - Project (proposal, final report): 40%
- **Project:**
  - Topics will be posted sometime this week (you can also come up with your own topic)
  - Need to hand in a **report** and do an oral **presentation**
  - Can work **individually** or in **pairs**

# Term Work Dates

<b>Term Work</b>	<b>Due Date</b>
Reviews	one day before class (Tuesdays)
Project Proposal	Feb 15
Project Report	March 28
Project Presentation	Last day of class

- All dates are for 2015. ;)

**Deadline** Reviews / project should be submitted **by 11.59pm on the date they are due**. Anything from 1 minute late to 24 hours will count as **one late day**.

**Lateness** Each student will be given a total of **3 free late days**. This means that you can hand in three of the reviews one day late, or one review three days late. It is up to the you to make a good planning of your work. **After you have used the 3 day budget, the late reviews will not be accepted.**

**Discount** You have a budget of 2 missing reviews without penalty

# Let's begin!

- Introduction to Visual Recognition with Text
  - Motivation
  - Diverse set of topics
- Visual Recognition
  - High-level overview of topics/problems

# Motivation

- Computer Vision is mainly about images, NLP about text
- But these two modalities do not appear in isolation

# Motivation

- Computer Vision is mainly about images, NLP about text
- But these two modalities do not appear in isolation

EDITION: INTERNATIONAL | U.S. | MEXICO | ARABIC

TV: CNN | CNN en Español

Set edition preference

Sign up | Log In

SEARCH

POWERED BY Google

Home | Video | World | U.S. | Africa | Asia | Europe | Latin America | Middle East | Money | World Sport | Entertainment | Tech | Travel | iReport

RADIOMIR 1940 ITALIAN DESIGN SWISS TECHNOLOGY PANERAI LABORATORIO DI IDEE.

December 1, 2014 — Updated 19:20 GMT (03:20 HKT)

**Editor's choice**

**The world's biggest ritual slaughter**

**The week in 36 photos**

**What's Prince Harry's secret?**

**Airstrikes mark shift to ISIS 'capital'**

The U.S.-led coalition fighting ISIS in Syria has stepped up its attacks on the militant Islamist group's de facto capital, the London-based Syrian Observatory for Human Rights says. [FULL STORY](#) | [THE ISIS TERROR THREAT](#)

**THE LATEST**

- Iraq's army weakened from within by 50,000 'ghost' soldiers
- Women recorded fighting sexual harassment
- Watch as alihkhole swallows car
- Boys make horrifying find on beach
- Pilot killed when U.S. jet crashes in Jordan
- World Cup bids: 'Picasso painting offered as kickback'

**DISCOVER MORE**

**PANERAI**

LABORATORIO DI IDEE.

**Welcome to the Muddle East**

**Analysis**

The U.S. will have its hands full in the Mideast for years to come, Aaron David Miller says.

- ISIS to begin minting its own currency
- Has ISIS peaked?

# Motivation

- Images do not appear in isolation



# Motivation

- Images do not appear in isolation



# Why study images and text?

# Why study images and text?

- Goals of AI include development of household robots, visual solutions for the blind, assistive driving
- An autonomous system needs to sense the 3D world and parse it semantically
- And it needs to communicate with the user



# Description Generation of Images/Videos

# Description Generation

- Goal: generate a naturally looking lingual description of a given image/video
- One of the most active research subareas involving images and text

# Description Generation

- Goal: generate a naturally looking lingual description of a given image/video
- One of the most active research subareas involving images and text
- Types of approaches:
  - Generating descriptions via hand-coded templates
  - Learning the grammar/templates (very few approaches, descriptions look less natural)
  - Borrowing a description from the most visually similar image in large dataset
  - Learning generation models via joint embeddings

# Description Generation

- Goal: generate a naturally looking lingual description of a given image/video
- One of the most active research subareas involving images and text
- Types of approaches:
  - Generating descriptions via hand-coded templates
  - Learning the grammar/templates (very few approaches, descriptions look less natural)
  - Borrowing a description from the most visually similar image in large dataset
  - Learning generation models via joint embeddings

# Papers – Image to Text

## Collective Generation of Natural Image Descriptions

Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg, Tamara L. Berg, Yejin Choi  
ACL, 2012

## TREETALK: Composition and Compression of Trees for Image Descriptions

Polina Kuznetsova, Vicente Ordonez, Tamara L. Berg, Yejin Choi  
TACL, 2014

## Im2Text: Describing Images Using 1 Million Captioned Photographs

Vicente Ordonez, Girish Kulkarni, Tamara L. Berg  
NIPS, 2011

## Every Picture Tells a Story: Generating Sentences for Images

A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, D. A. Forsyth  
ECCV, 2010

## I2T: Image Parsing to Text Description

B.Z. Yao, X. Yang, L. Liang, M. W. Lee, S.-C. Zhu  
Proc of IEEE, 2010

## How many words is a picture worth? Automatic caption generation for news images

Y. Feng, M. Lapata  
ACL, 2010

## Corpus-Guided Sentence Generation of Natural Images

Y. Yang, C. L. Teo, H. Daume III, Y. Aloimonos  
EMNLP, 2011

## Multimodal Neural Language Models

Ryan Kiros, Ruslan Salakhutdinov, Richard Zemel  
ICML, 2014

Project page: <http://www.cs.toronto.edu/~rkiros/multimodal.html>

# Papers – Video to Text

## Translating Video Content to Natural Language Descriptions

M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, B. Schiele  
ICCV, 2013

## Video In Sentences Out

Andrei Barbu, Alexander Bridge, Zachary Burchill, Dan Coroiu, Sven Dickinson, Sanja Fidler, Aaron Michaux, Sam Mussman, Siddharth Narayanaswamy, Dhaval Salvi, Lara Schmidt, Jiangnan Shangguan, Jeffrey Mark Siskind, Jarrell Waggoner, Song Wang, Jinlian Wei, Yifan Yin, Zhiqi Zhang  
UAI, 2012

Project page: <https://engineering.purdue.edu/~qobi/mindseye/>

## A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching

P. Das, C. Xu, R. F. Doell, and J. J. Corso  
CVPR, 2013

## Understanding Videos, Constructing Plots: Learning a Visually Grounded Storyline Model from Annotated Videos

A. Gupta, P. Srinivasan, J. Shi, L. S. Davis  
CVPR 2009

## Generating Natural-Language Video Descriptions Using Text-Mined Knowledge

N Krishnamoorthy, G Malkarnenkar, RJ Mooney, K. Saenko, S. Guadarrama  
AAAI, 2013

## YouTube2Text: Recognizing and Describing Arbitrary Activities Using Semantic Hierarchies and Zero-shot Recognition

S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, K. Saenko  
ICCV, 2013

# Papers – Last two months

## Learning a Recurrent Visual Representation for Image Caption Generation

Xinlei Chen, C. Lawrence Zitnick  
(arXiv:1411.5654), Nov, 2014

## From Captions to Visual Concepts and Back

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh Srivastava, Li Deng, Piotr Dollr, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, Geoffrey Zweig  
(arXiv:1411.4952), Nov 2014

## Long-term Recurrent Convolutional Networks for Visual Recognition and Description

Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, Trevor Darrell  
(arXiv:1411.4389), Nov 2014

## Explain Images with Multimodal Recurrent Neural Networks

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Alan L. Yuille  
(arXiv:1410.1090), Oct 2014

## Show and Tell: A Neural Image Caption Generator

Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan  
(arXiv:1411.4555), Nov 2014

## Deep Visual-Semantic Alignments for Generating Image Descriptions

Andrej Karpathy, Li Fei-Fei  
(arXiv:1412.2306), Dec 2014

Project page: <http://cs.stanford.edu/people/karpathy/deepimagesent/>

## Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models

Ryan Kiros, Ruslan Salakhutdinov, Richard S. Zemel  
(arXiv:1411.2539), Nov 2014

# Image Description

A small plane parked in a field with trees in the background.



[Source: L. Zitnick, NIPS'14 Workshop on Learning Semantics]

# Image Description

A man with a colorful umbrella walking down a street.



[Source: L. Zitnick, NIPS'14 Workshop on Learning Semantics]

# Image Description

A train traveling down train tracks next to a train station.



[Source: L. Zitnick, NIPS'14 Workshop on Learning Semantics]

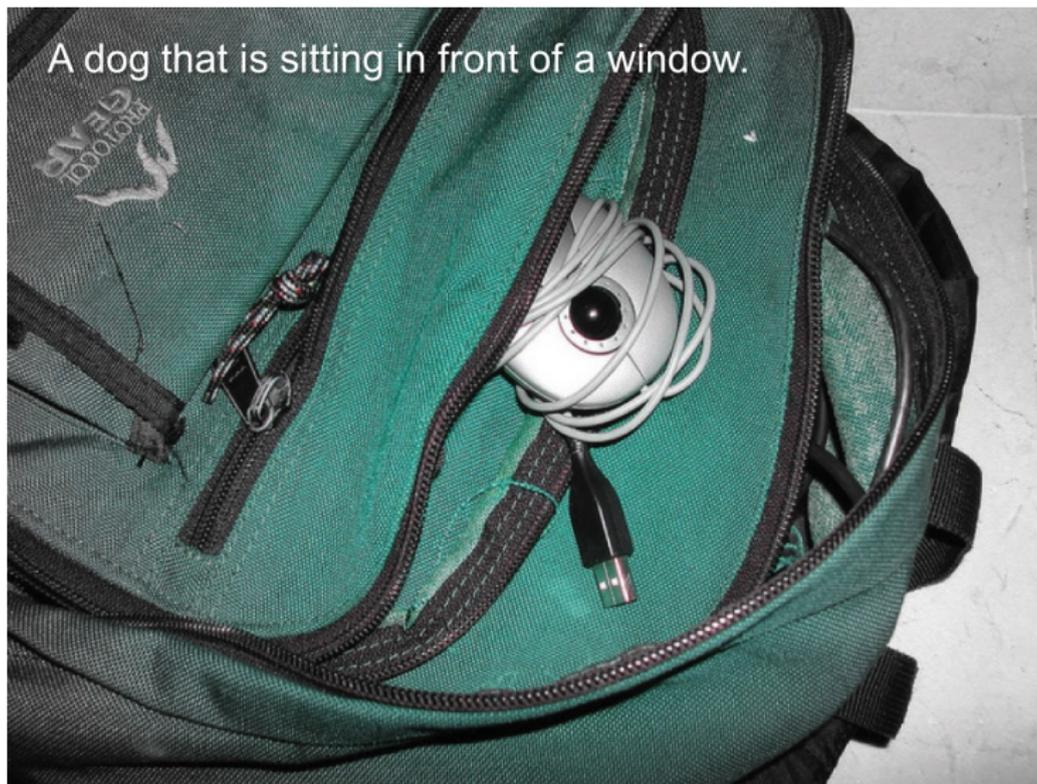
# Image Description

A herd of giraffes walk down the street in the middle of some trees.



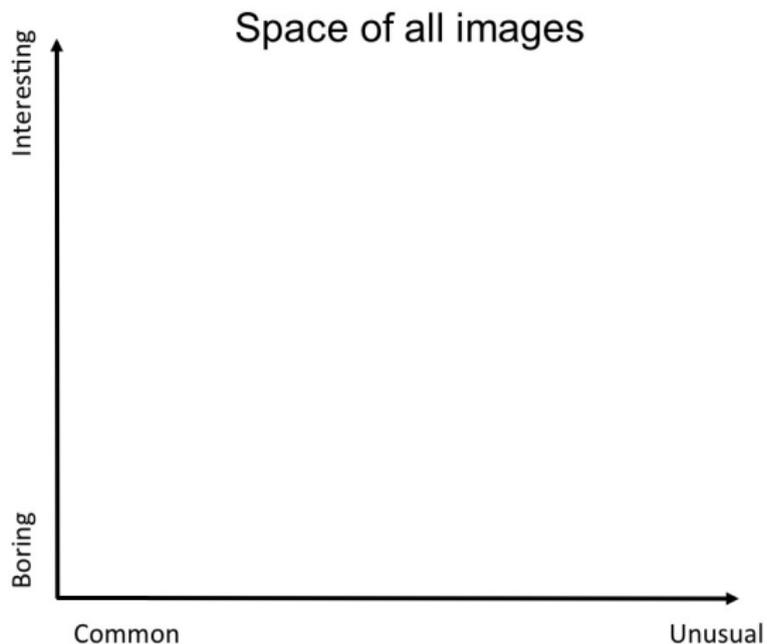
[Source: L. Zitnick, NIPS'14 Workshop on Learning Semantics]

# Image Description



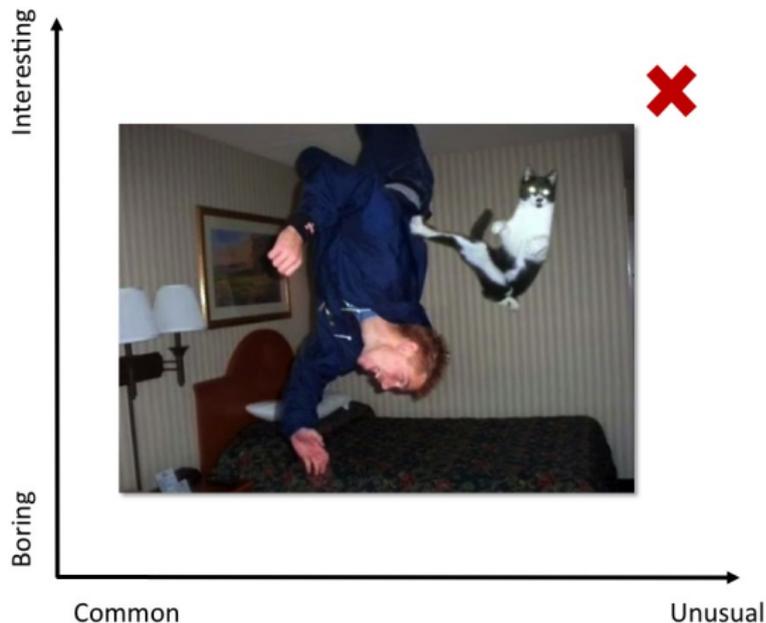
[Source: L. Zitnick, NIPS'14 Workshop on Learning Semantics]

# Image Description



[Source: L. Zitnick, NIPS'14 Workshop on Learning Semantics]

# Image Description



[Source: L. Zitnick, NIPS'14 Workshop on Learning Semantics]

# Image Description



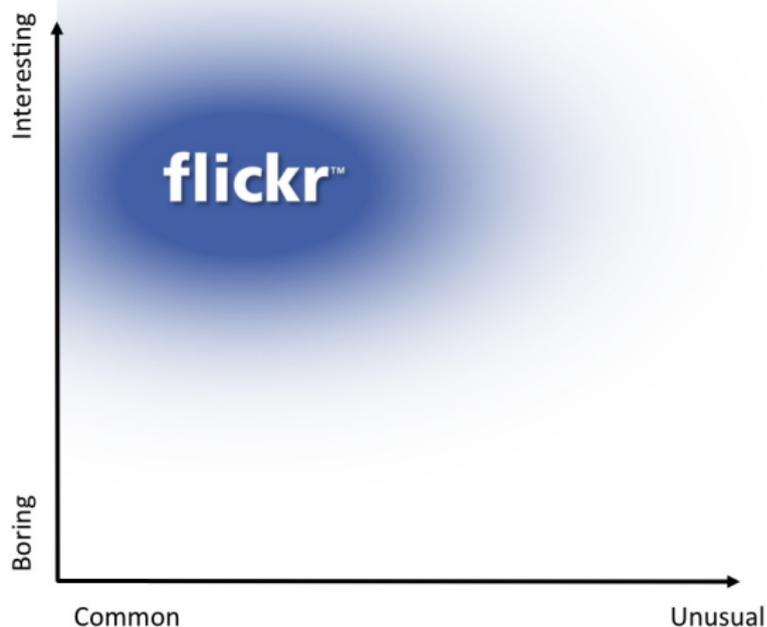
[Source: L. Zitnick, NIPS'14 Workshop on Learning Semantics]

# Image Description



[Source: L. Zitnick, NIPS'14 Workshop on Learning Semantics]

# Image Description



[Source: L. Zitnick, NIPS'14 Workshop on Learning Semantics]

# Image Description

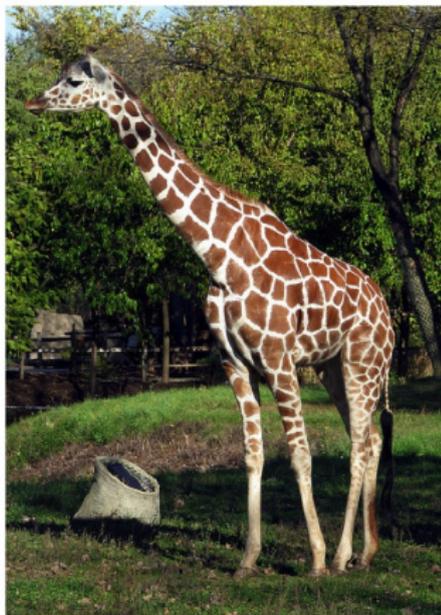


**Figure:** Vemodalen: The Fear That Everything Has Already Been Done

[Source: L. Zitnick, NIPS'14 Workshop on Learning Semantics]

# Image Description

A giraffe standing in the grass next to a tree.



[Source: L. Zitnick, NIPS'14 Workshop on Learning Semantics]

# Image Description

A giraffe standing in the grass next to a tree.

“giraffe”

[Source: L. Zitnick, NIPS'14 Workshop on Learning Semantics]

# Image Description

A giraffe standing in the grass next to a tree.

“giraffe”



[Source: L. Zitnick, NIPS'14 Workshop on Learning Semantics]

# Image Description

A giraffe standing in the grass next to a tree.

“giraffe”



[Source: L. Zitnick, NIPS'14 Workshop on Learning Semantics]

# Image Description

A giraffe standing in the grass next to a tree.

“giraffe”

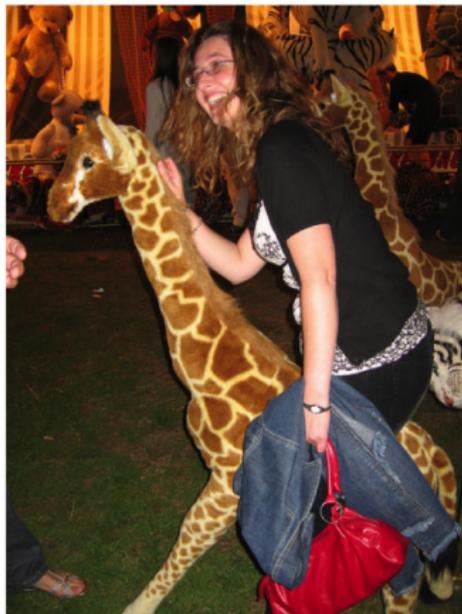


[Source: L. Zitnick, NIPS'14 Workshop on Learning Semantics]

# Image Description

A giraffe standing in the grass next to a tree.

“giraffe”



[Source: L. Zitnick, NIPS'14 Workshop on Learning Semantics]

# Image Description

A giraffe standing in the grass next to a tree.

“giraffe”



[Source: L. Zitnick, NIPS'14 Workshop on Learning Semantics]

# Image Description

A giraffe standing in the grass next to a tree.

“giraffe”

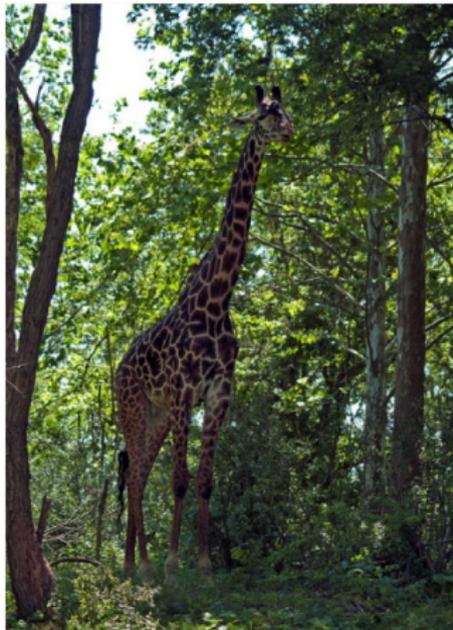


[Source: L. Zitnick, NIPS'14 Workshop on Learning Semantics]

# Image Description

A giraffe standing in the grass next to a tree.

“giraffe”

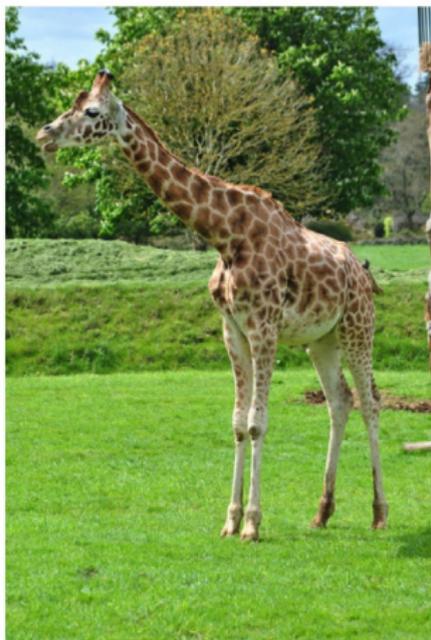


[Source: L. Zitnick, NIPS'14 Workshop on Learning Semantics]

# Image Description

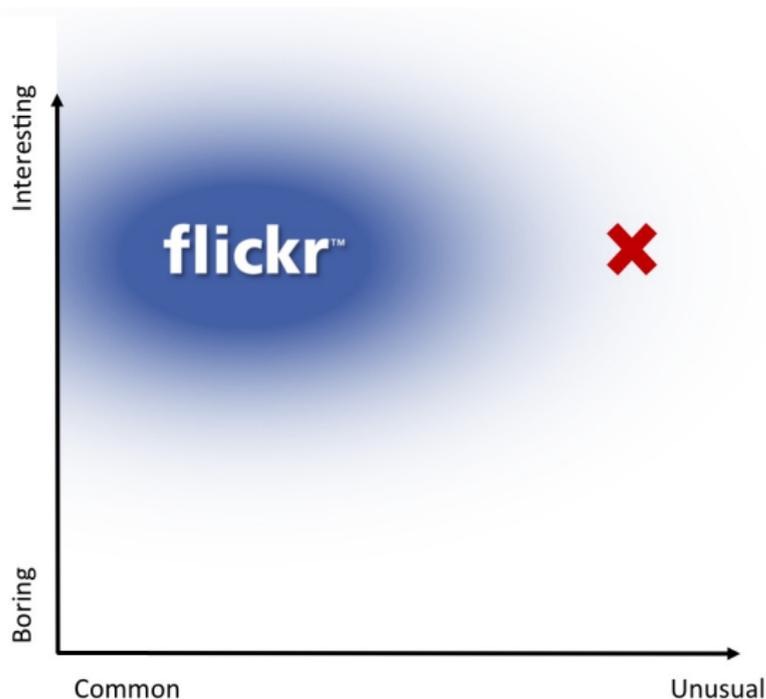
A giraffe standing in the grass next to a tree.

“giraffe”



[Source: L. Zitnick, NIPS'14 Workshop on Learning Semantics]

# Image Description – Main Challenges



[Source: L. Zitnick, NIPS'14 Workshop on Learning Semantics]

# Image Description – Main Challenges

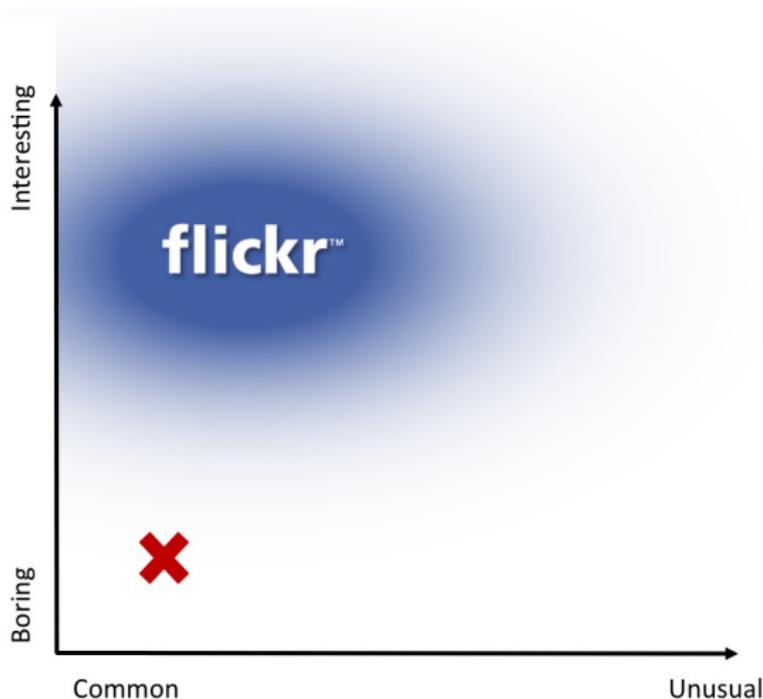
A crazy zebra climbing a giraffe to get a better view.

The limits of vision and language models...



[Source: L. Zitnick, NIPS'14 Workshop on Learning Semantics]

# Image Description – Main Challenges



[Source: L. Zitnick, NIPS'14 Workshop on Learning Semantics]

# Image Description – Main Challenges



[Source: L. Zitnick, NIPS'14 Workshop on Learning Semantics]

# Image Description – Main Challenges



[Source: L. Zitnick, NIPS'14 Workshop on Learning Semantics]

# Comparison: Datasets, Metrics

## Datasets:

- Microsoft Coco (<http://mscoco.org/>) has 5 sentences per image
- UIUC [dataset](#) has 3 sentences per image
- [Abstract Images](#)
- ImageFlickr 8K [dataset](#)
- Flickr30K [dataset](#)
- YouTube [dataset](#) has descriptions of videos

## How to evaluate?

- Metrics: BLEU, ROUGE, METEOR, however these standard measures don't match human judgements well
- This paper provides in-depth analysis and proposes new perceptual metrics:

[Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics](#)

M. Hodosh, P. Young, J. Hockenmaier  
Journal of Artificial Intelligence Research, 2013

- Microsoft is developing the first benchmark for image description generation for dataset CoCo (roughly 120K images), planned release in Feb'15

# Image Generation from Text

WordsEye: An Automatic Text-to-Scene Conversion System

Bob Coyne, Richard Sproat

SIGGRAPH, 2001

*Goal:* Generate a 3D scene given a textual description. What's the motivation?

## **Input text:**

John uses the crossbow. He rides the horse by the store. The store is under the large willow. The small allosaurus is in front of the horse. The dinosaur faces John. A gigantic teacup is in front of the store. The dinosaur is in front of the horse. The gigantic mushroom is in the teacup. The castle is to the right of the store.

# Image Generation from Text

WordsEye: An Automatic Text-to-Scene Conversion System

Bob Coyne, Richard Sproat

SIGGRAPH, 2001

*Goal:* Generate a 3D scene given a textual description. What's the motivation?

## Input text:

John uses the crossbow. He rides the horse by the store. The store is under the large willow. The small allosaurus is in front of the horse. The dinosaur faces John. A gigantic teacup is in front of the store. The dinosaur is in front of the horse. The gigantic mushroom is in the teacup. The castle is to the right of the store.

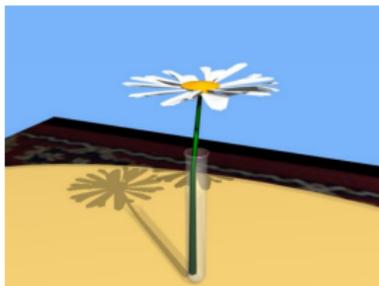


# Image Generation from Text

WordsEye: An Automatic Text-to-Scene Conversion System

Bob Coyne, Richard Sproat

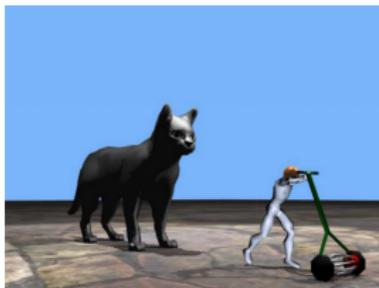
SIGGRAPH, 2001



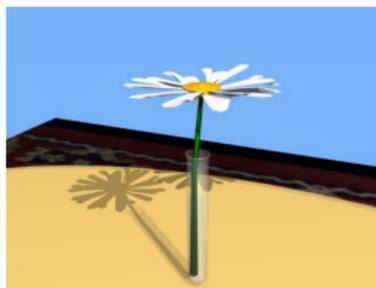
The daisy is in the test tube.



The blue daisy is not in the army boot.



The lawn mower is 5 feet tall. John pushes the lawn mower.  
The cat is 5 feet behind John. The cat is 10 feet tall.



The bird is in the bird cage. The bird cage is on the chair.

# Image Generation from Text

## Learning the Visual Interpretation of Sentences

C. L. Zitnick, D. Parikh, L. Vanderwende

ICCV, 2013

Input Description	Tuples	GT	Full-CRF	BoW	Noun-CRF	Random
Jenny is catching the ball. Mike is kicking the ball. The table is next to the tree.	<pre>&lt;&lt;Jenny&gt;, &lt;catch&gt;, &lt;ball&gt;&gt; &lt;&lt;Mike&gt;, &lt;kick&gt;, &lt;ball&gt;&gt; &lt;&lt;table&gt;, &lt;be&gt;, &lt;&gt;&gt;</pre>					
Mike is sitting next to Jenny. The cat is sitting next to the tree. Jenny is throwing the ball.	<pre>&lt;&lt;Mike&gt;, &lt;sit next to&gt;, &lt;Jenny&gt;&gt; &lt;&lt;cat&gt;, &lt;sit next to&gt;, &lt;tree&gt;&gt; &lt;&lt;Jenny&gt;, &lt;throw&gt;, &lt;ball&gt;&gt;</pre>					
Mike is scared of lightning. It is a stormy day. Jenny is standing on the slide.	<pre>&lt;&lt;Mike&gt;, &lt;be scared&gt;, &lt;&gt;&gt; &lt;&lt;day&gt;, &lt;be, stormy&gt;, &lt;&gt;&gt; &lt;&lt;Jenny&gt;, &lt;stand on&gt;, &lt;slide&gt;&gt;</pre>					

**Figure:** The model takes a sentence and generates a scene by sampling from a CRF.

# Detecting Text in the Wild

# Motivation

- In the last month I did some traveling. First I was getting lost in Hong Kong.
- That involved remembering street names from a map and matching them to the road signs.



# Motivation

- Then there was shopping...



# Motivation

- And sometimes I wished I could read the food ingredients...



# Motivation

- Then there was Christmas back home, which required finding gifts with my name on it. ;)



# Motivation

- After that a few days on the beach. Even beach had signs.



# Motivation

- At work, wouldn't it be cool to have automatic grading?

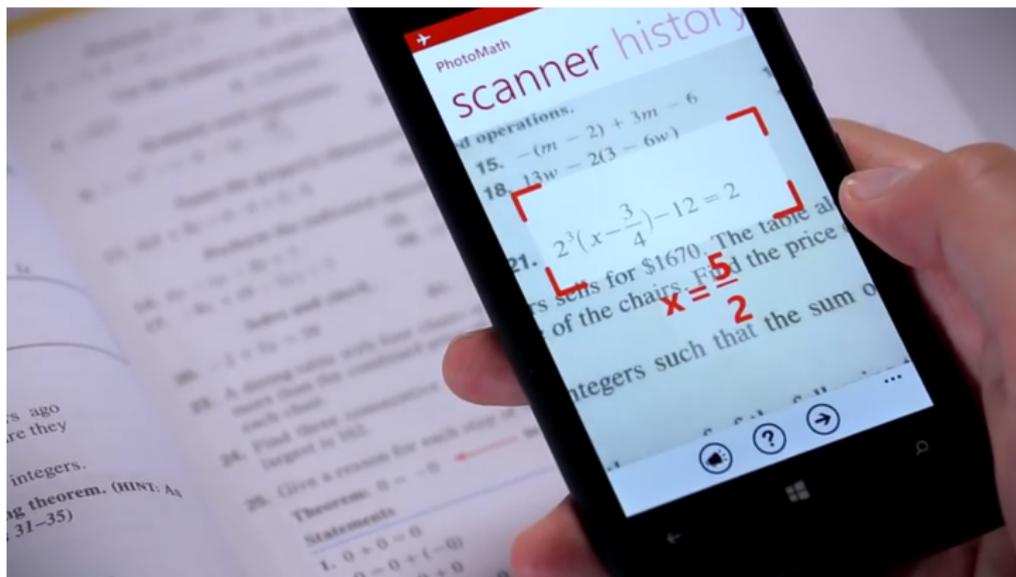


Figure: Photomath: <https://photomath.net/>

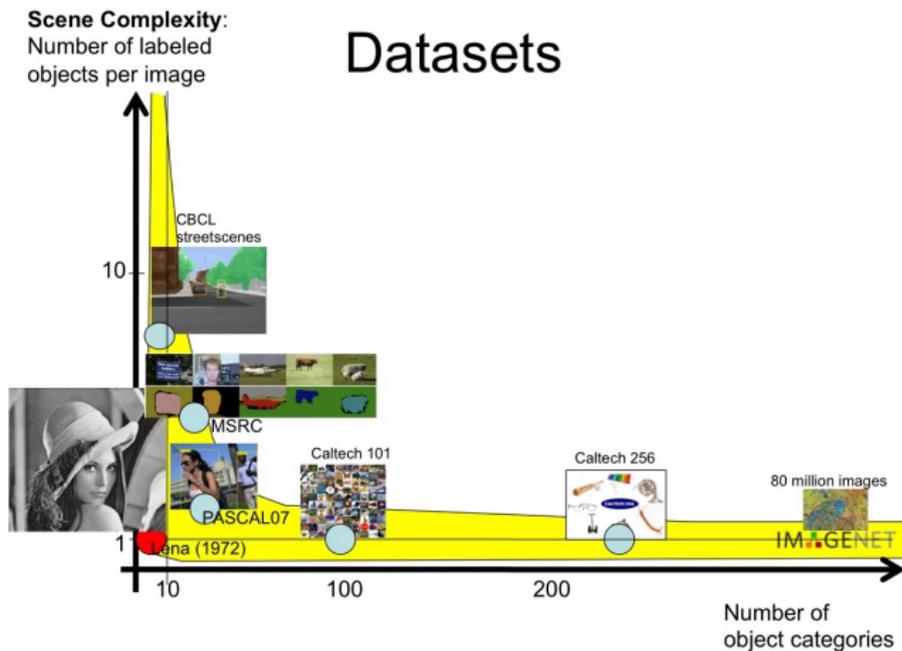
# Spotting Text in Images

- Parsing text in images is useful for several applications
- The problem involves:
  - Localizing regions with text in an image (possibly generating multiple hypotheses and verifying them with a more complex model)
  - Parsing the text into letters/words
- Challenges
  - Text can be in various viewpoints (need to deal with perspective effects)
  - Various fonts
  - Various languages
  - Open vocabulary (depends on the task)
  - For a real application, may need to be close-to real time

# Learning Visual Models via Text

# Motivation

- In the era of “big data”, there is a trend to collect really large datasets



[Source: A. Torralba, “Beware, Humans in the Loop”]

# Motivation

- Someone needs to label this data... Most common annotation source these days is Mechanical Turk (cheap).



Labeling to get a Ph.D.



Labeling for money  
(Sorokin, Forsyth, 2008)



Labeling because it  
gives you added value



Visipedia  
(Belongie, Perona, et al)

Just for labeling



[Source: A. Torralba, "Beware, Humans in the Loop"]

# Motivation

- MT is sometimes amazing

## 1 cent

Task: Label one object in this image



[Source: A. Torralba, "Beware, Humans in the Loop"]

# Motivation

- MT is sometimes amazing

## 1 cent

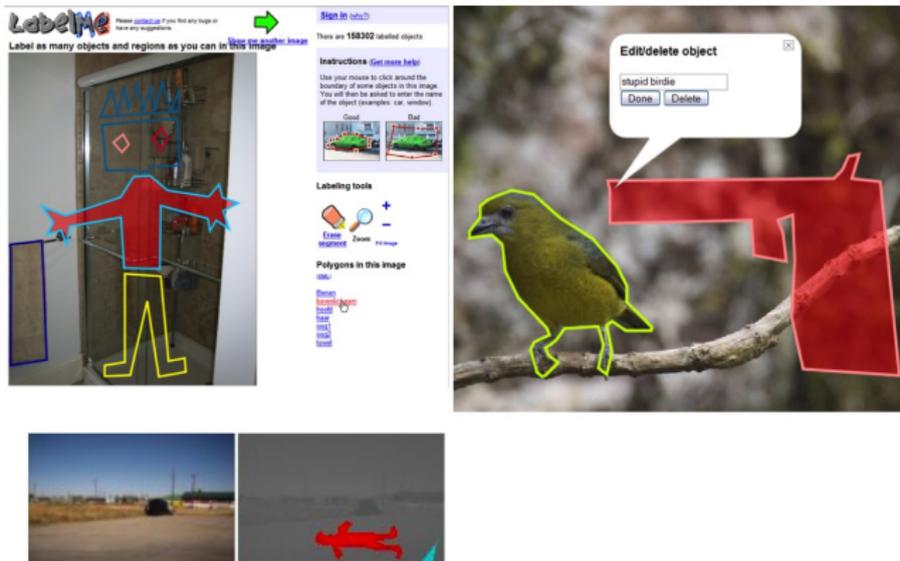
Task: Label one object in this image



[Source: A. Torralba, "Beware, Humans in the Loop"]

# Motivation

- MT is sometimes amazing
- but most of the time unreliable



[Source: A. Torralba, "Beware, Humans in the Loop"]

# Motivation

- MT is sometimes amazing
- but most of the time unreliable

Flip a coin

Requester: ROBERT C MILLER      Reward: \$0.01 per HIT      HITs Available: 3      Duration: 5 minutes

Qualifications Required: None

---

Please flip an actual coin and type either H or T below.

After 50 HITs:



And 50 more:



Experiment by Rob Miller

[Source: A. Torralba, "Beware, Humans in the Loop"]

# Motivation

- MT is sometimes amazing
- but most of the time unreliable

Choose the given item.

Requester: SimpleSphere    Reward: \$0.01 per HIT    HITs Available: 1    Duration: 60 minutes  
Qualifications Required: None

Please click button B:

Results of 100 HITS

A: 2  
B: 96  
C: 2

Experiment by Greg Little

[Source: A. Torralba, "Beware, Humans in the Loop"]

- Hire high quality workers and train them → Expensive (e.g. labeling KITTI was 30K EUR)

- Hire high quality workers and train them → Expensive (e.g. labeling KITTI was 30K EUR)
- Ask your mum to help:

Notes on image annotation

A. Barriuso and A. Torralba

arXiv:1210.3448, 2012

- Hire high quality workers and train them → Expensive (e.g. labeling KITTI was 30K EUR)
- Ask your mum to help:

## Notes on image annotation

A. Barriuso and A. Torralba  
arXiv:1210.3448, 2012

- Stop relying on detailed, laborious annotation → e.g., learn visual models from textual descriptions of images

- More and more datasets have descriptions:
  - Flickr has tags
  - Microsoft Coco (<http://mscoco.org/>) has 5 sentences per image
  - UIUC [dataset](#) augments 1000 PASCAL images with sentences

in a restaurant kitchen, a woman prepare pizzas  
a worker puts toppings on a pizza at a pizza shop  
a woman making pizzas inside of a professional kitchen.  
a woman making a pizza that sits in front of her  
an employee in a red shirt sprinkling cheese on a pizza



a couple of white teddy bears sitting together.  
two stuffed animals with stitched on paces and colored paws.  
two white teddy bears one has pink feet the other blue.  
a pair of white, boy and girl teddy bears  
there are two stuffed animals sitting next to each other



Figure: Examples from Microsoft Coco.

# Representative Approaches

- Learn object/scene/attributes models given image tags:

[Towards Total Scene Understanding: Classification, Annotation and Segmentation in an Automatic Framework](#)

L.-J. Li, R. Socher, L. Fei-Fei  
CVPR, 2009

[Multimodal Learning with Deep Boltzmann Machines](#) N. Srivastava, R. Salakhutdinov

NIPS, 2012

Project page: <http://www.cs.toronto.edu/~nitish/multimodal/>

[On Learning to Localize Objects with Minimal Supervision](#)

H. O. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, T. Darrell  
ICML, 2014

- Grounding nouns and prepositions:

[Beyond Nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers](#)

A. Gupta, L. S. Davis  
ECCV, 2008

[Matching Words and Pictures](#)

K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, M. I. Jordan  
JMLR, 2003

- Learn people' names in videos (using scripts):

[Joint person naming in videos and coreference resolution in text](#)

V. Ramanathan, A. Joulin, P. Liang, L. Fei-Fei  
ECCV, 2014

# Representative Approaches

- Learn actions and roles using descriptions:

[Video Event Understanding using Natural Language Descriptions](#)

V. Ramanathan, P. Liang, L. Fei-Fei  
ICCV, 2013

[Grounded Language Learning from Video Described with Sentences](#)

H. Yu, J.M. Siskind  
ACL, 2013

[Understanding Videos, Constructing Plots: Learning a Visually Grounded Storyline Model from Annotated Videos](#)

A. Gupta, P. Srinivasan, J. Shi, L. S. Davis  
CVPR 2009

- Learn alignment between words in sentences and image regions:

[Deep Visual-Semantic Alignments for Generating Image Descriptions](#)

A. Karpathy, L. Fei-Fei  
(arXiv:1412.2306), Dec 2014

- Learn concepts from questions and answers:

[A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input](#)

M. Malinowski, M. Fritz  
NIPS, 2014

# Challenges

Challenges when learning from descriptions:

Challenges when learning from descriptions:

- Descriptions sometime talk about entities that are not depicted (e.g., memories, high-level semantics)
- People do not typically describe everything in the image
- Vision is not solved, typically many region proposals per image, most do not contain true objects
- For robotic applications, e.g. vision for the blind of household robots:
  - The system needs to figure out that it doesn't know a new object/concept
  - Learn new concepts continuously and incrementally through dialog

# Solutions

- Stop relying on detailed, laborious annotation → e.g., learn visual models from textual descriptions of images
- Zero-shot learning via text (e.g., encyclopedia entries, attribute tags)

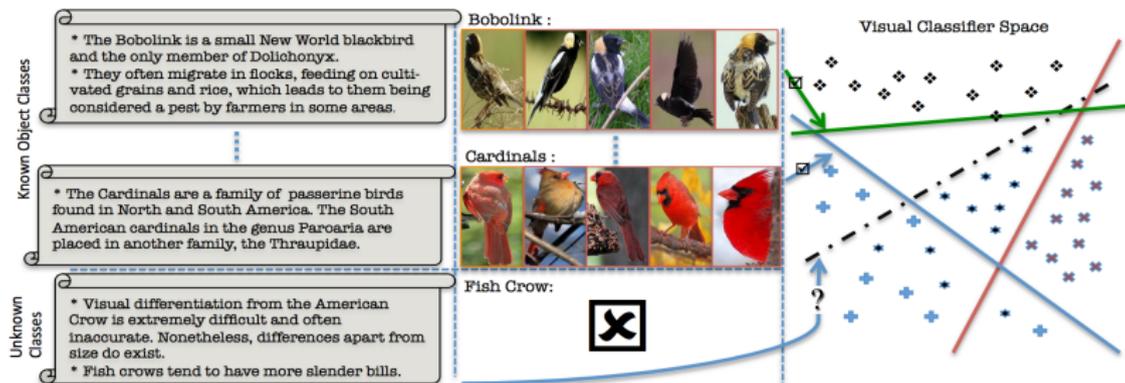


Figure: From Elhoseiny et al., ICCV'13

- Using descriptions:

## Write a Classifier: Zero-Shot Learning Using Purely Textual Descriptions

M. Elhoseiny, B. Saleh, A. Elgammal

ICCV, 2013

## DeViSE: A Deep Visual-Semantic Embedding Model

A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M.'A. Ranzato, T. Mikolov

NIPS, 2013

- Using attributes:

## Attribute-Based Classification for Zero-Shot Visual Object Categorization

C. H. Lampert, H. Nickisch, S. Harmeling

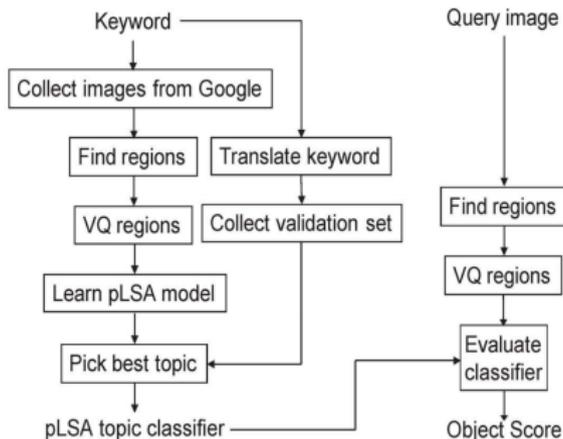
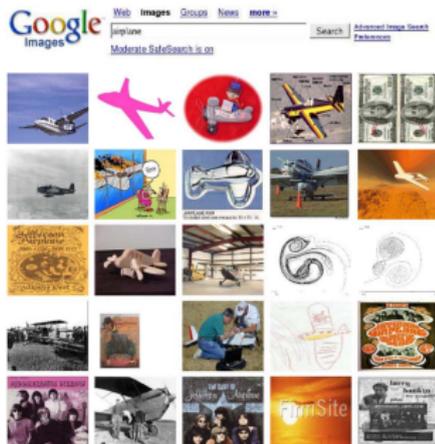
TPAMI, 2014

# Solutions

- Stop relying on detailed, laborious annotation → e.g., learn visual models from textual descriptions of images
- Zero-shot learning via text (e.g., encyclopedia entries, attribute tags)
- Can you learn object models by querying a search engine?

# Solutions

- Stop relying on detailed, laborious annotation → e.g., learn visual models from textual descriptions of images
- Zero-shot learning via text (e.g., encyclopedia entries, attribute tags)
- Can you learn object models by querying a search engine?



## Learning Object Categories From Internet Image Searches

R. Fergus, L. Fei-Fei, P. Perona, A. Zisserman

Proc of IEEE, 2010

# Problems

- Biased images, noisy results
- Some words have multiple senses: problem of *word-sense disambiguation*



Figure: Google search results with query *mouse*

# Representative Approach for WSD

- Wordnet to lookup the number of visual senses
- Use text surrounding the query term in the corresponding webpages to disambiguate

## Unsupervised Learning of Visual Sense Models for Polysemous Words

K. Saenko, T. Darrell  
NIPS, 2008

## Joint Image and Word Sense Discrimination for Image Retrieval

Aurelien Lucchi, Jason Weston  
ECCV, 2012

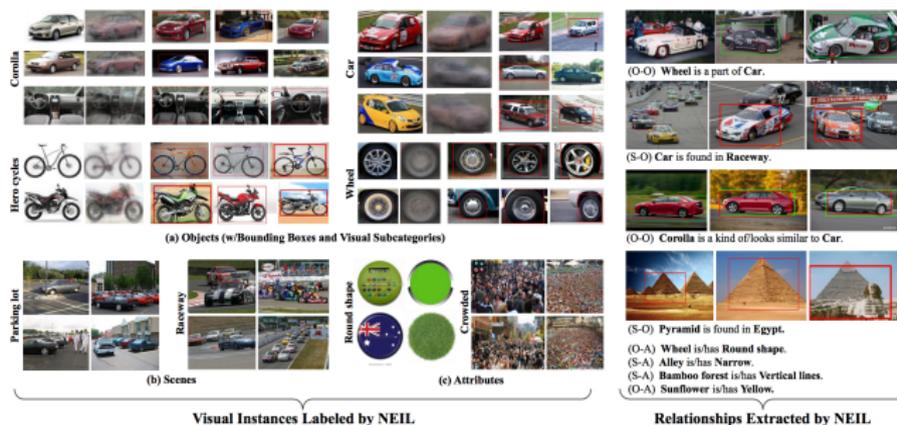
- Inferring senses and visual models for polysemous words for captioned images

## Word Sense Disambiguation with Pictures

K. Barnard, M. Johnson  
AI, 2005

# NEIL: Never Ending Image Learning

- Learns object models by querying a search engine
- Deals with noise in retrieval, polysemy, variations in viewpoints, etc
- Running since July 15, 2014, Analyzed 5 million Images, Labeled 0.5 million images and Learned 3000 Common sense relationships.
- 2,702 Concepts, 1,002,026 Bounding boxes, 8,685 Visual Models, 2,201,468 Images, 517,450 Segmentations, 4,695 Visual Relationships



NEIL: Extracting Visual Knowledge from Web Data

X. Chen, A. Shrivastava, A. Gupta  
ICCV, 2013

Project page: <http://www.neil-kb.com/>

# Learning Motivation

## Inferring the Why in Images

H. Pirsiavash, C. Vondrick, A. Torralba  
TR, 2014



**Human Label:** sitting on bench in a train station because he is waiting

**Top Predictions:**

1. sitting on bench in a park because he is waiting
2. holding a tv in a park because he wants to take
3. holding a seal in a park because he wants to protest
4. holding a guitar in a park because he wants to play



**Human Label:** sitting on chair in a dining room because she wants to eat

**Top Predictions:**

1. sitting near table in dining room because she wants to eat
2. sitting on a sofa in a dining room because she wants to eat
3. holding a cup in a dining room because she wants to eat
4. sitting on a cup in a dining room because she wants to eat

**Figure:** Learning motivation of people by mining the knowledge stored in massive amounts of text. Using language models estimated on billions of webpages, the approach is able to acquire common knowledge about peoples experiences, such as their interactions with objects, their environments, and their motivations.

# Learning Affordances?

- Robotic platforms typically use reinforcement learning on RGB+depth data to learn concepts

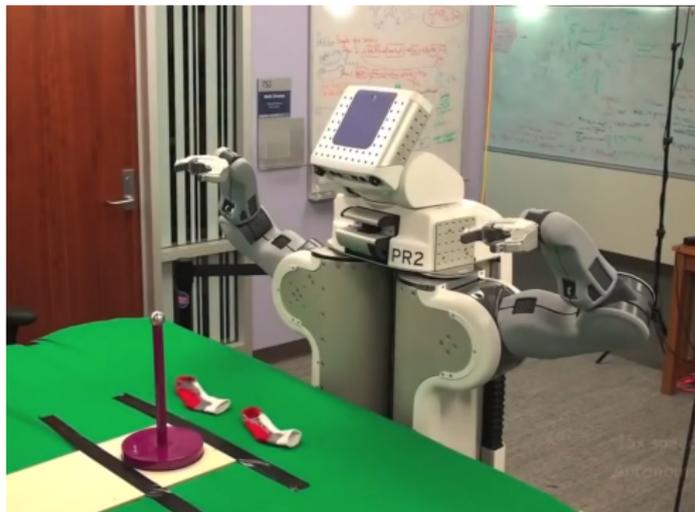


Figure: Sockification by Wang et al., <https://www.youtube.com/watch?v=KKUaVzf30qw>

# Learning Affordances?

- Robotic platforms typically use reinforcement learning on RGB+depth data to learn concepts
- It could be useful to also use spoken/lingual instructions to help learn e.g., affordances

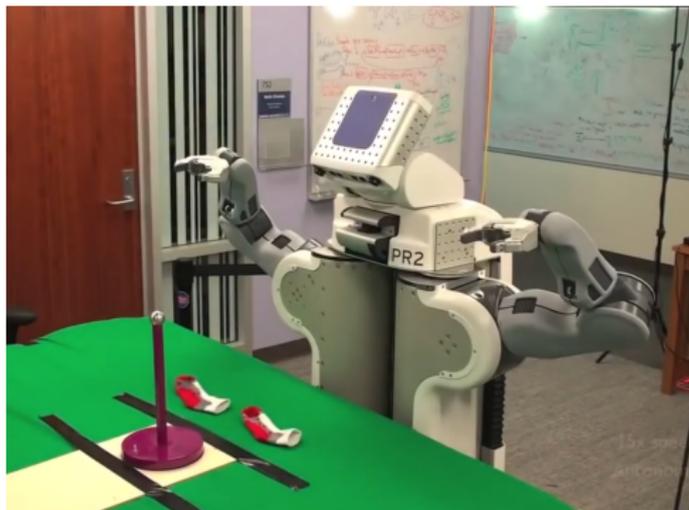
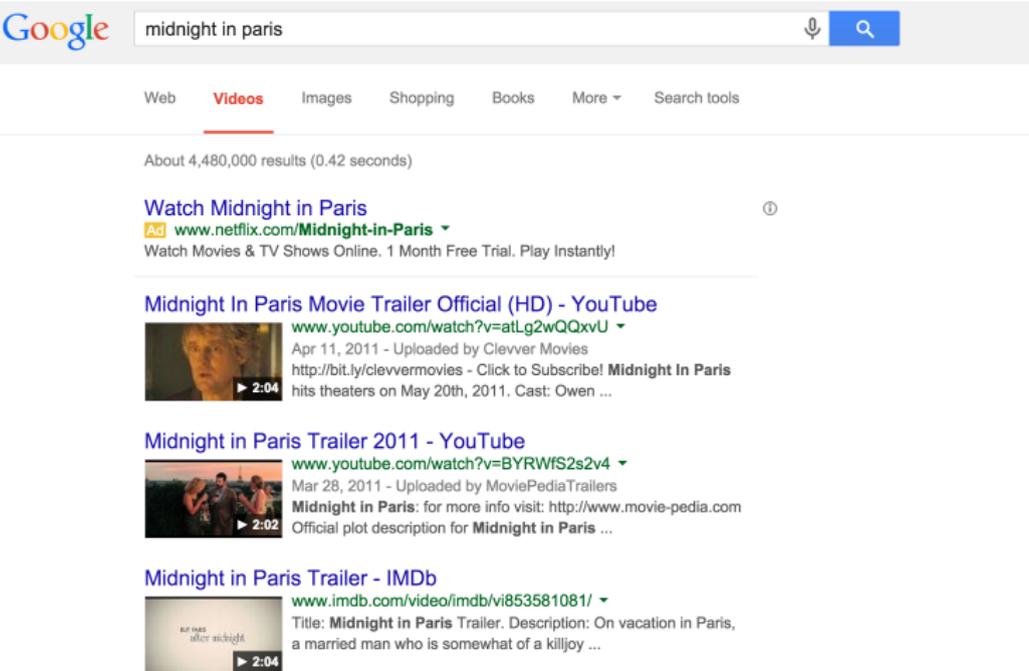


Figure: Sockification by Wang et al., <https://www.youtube.com/watch?v=KKUaVzf30qw>

# Visual Retrieval with Natural Lingual Queries

- Search engines are typically very good with one or two tags



The screenshot shows a Google search interface with the query "midnight in paris" entered in the search bar. Below the search bar, there are navigation tabs for "Web", "Videos", "Images", "Shopping", "Books", "More", and "Search tools". The "Videos" tab is selected and highlighted with a red underline. Below the tabs, it says "About 4,480,000 results (0.42 seconds)".

The search results are as follows:

- Watch Midnight in Paris** (Ad) [www.netflix.com/Midnight-in-Paris](http://www.netflix.com/Midnight-in-Paris) ⌵  
Watch Movies & TV Shows Online. 1 Month Free Trial. Play Instantly!
- Midnight In Paris Movie Trailer Official (HD) - YouTube** Ⓞ  
[www.youtube.com/watch?v=atLg2wQQxvU](http://www.youtube.com/watch?v=atLg2wQQxvU) ⌵  
Apr 11, 2011 - Uploaded by Clevver Movies  
<http://bit.ly/clevvermovies> - Click to Subscribe! **Midnight In Paris** hits theaters on May 20th, 2011. Cast: Owen ...
- Midnight in Paris Trailer 2011 - YouTube**  
[www.youtube.com/watch?v=BYRWfS2s2v4](http://www.youtube.com/watch?v=BYRWfS2s2v4) ⌵  
Mar 28, 2011 - Uploaded by MoviePediaTrailers  
**Midnight in Paris**: for more info visit: <http://www.movie-pedia.com>  
Official plot description for **Midnight in Paris** ...
- Midnight in Paris Trailer - IMDb**  
[www.imdb.com/video/imdb/vi853581081/](http://www.imdb.com/video/imdb/vi853581081/) ⌵  
Title: **Midnight in Paris** Trailer. Description: On vacation in Paris, a married man who is somewhat of a killjoy ...

**Query:** Midnight in Paris

- Still sort of work with slightly longer queries

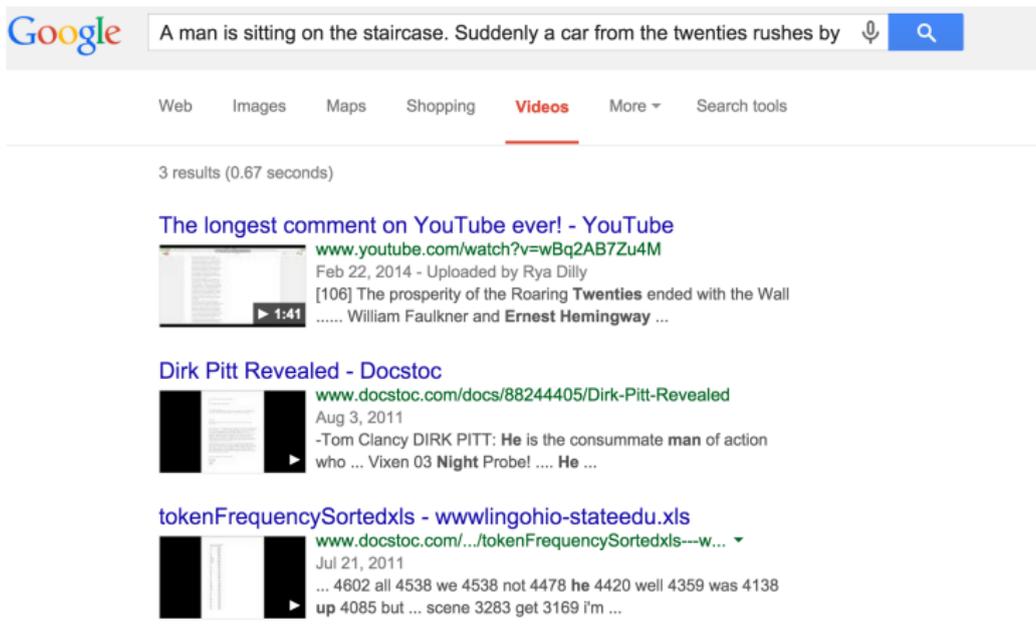
The screenshot shows a Google search interface. The search bar contains the text "Scene in midnight in paris of a man sitting on staircase". Below the search bar, the "Videos" tab is selected. The results show three video entries:

- Minuit a paris - scene de la voiture - YouTube**  
Thumbnail: A man sitting on a staircase.  
URL: [www.youtube.com/watch?v=vbdpxbmf2WA](http://www.youtube.com/watch?v=vbdpxbmf2WA)  
Date: Aug 31, 2012 - Uploaded by Frederic A  
Description: Minuit a paris - **scene** de la voiture ... "Midnight in Paris" (Amazon.com) ... Midnight in the Paris-best **scene** ...
- Woody Allen's "Midnight in Paris" - rehearsal for scene with ...**  
Thumbnail: A scene from the movie "Midnight in Paris".  
URL: [www.youtube.com/watch?v=nvACVii1Q-o](http://www.youtube.com/watch?v=nvACVii1Q-o)  
Date: Aug 5, 2010 - Uploaded by writerinparis99  
Description: In early August 2010, this **scene** for Woody Allen's film "Midnight in Paris" (scheduled release 2011) was ...
- Lady Gaga vomits four times on stage as she continues to ...**  
Thumbnail: Lady Gaga performing on stage.  
URL: [www.dailymail.co.uk/.../Lady-Gaga-vomits-times-st...](http://www.dailymail.co.uk/.../Lady-Gaga-vomits-times-st...)  
Date: Oct 8, 2012  
Description: The pop performer, 26, had just begun to strut down a set of stairs on to the ... in the Obama administration ...

**Query:** Scene in Midnight in Paris of a man sitting on staircase

# Motivation

- Typically completely fail with longer descriptions



The screenshot shows a Google search interface. The search bar contains the text "A man is sitting on the staircase. Suddenly a car from the twenties rushes by" followed by a microphone icon and a search button. Below the search bar are navigation tabs for "Web", "Images", "Maps", "Shopping", "Videos" (which is highlighted with a red underline), "More", and "Search tools". The search results section shows "3 results (0.67 seconds)".

The first result is titled "The longest comment on YouTube ever! - YouTube" with a thumbnail showing a video player. The URL is [www.youtube.com/watch?v=wBq2AB7Zu4M](http://www.youtube.com/watch?v=wBq2AB7Zu4M), dated Feb 22, 2014, uploaded by Rya Dilly. The description includes "[106] The prosperity of the Roaring Twenties ended with the Wall ..... William Faulkner and Ernest Hemingway ...".

The second result is titled "Dirk Pitt Revealed - Docstoc" with a thumbnail showing a document. The URL is [www.docstoc.com/docs/88244405/Dirk-Pitt-Revealed](http://www.docstoc.com/docs/88244405/Dirk-Pitt-Revealed), dated Aug 3, 2011. The description includes "-Tom Clancy DIRK PITT: He is the consummate man of action who ... Vixen 03 Night Probe! .... He ...".

The third result is titled "tokenFrequencySortedxls - wwwlingohio-stateedu.xls" with a thumbnail showing a spreadsheet. The URL is [www.docstoc.com/.../tokenFrequencySortedxls---w...](http://www.docstoc.com/.../tokenFrequencySortedxls---w...), dated Jul 21, 2011. The description includes "... 4602 all 4538 we 4538 not 4478 he 4420 well 4359 was 4138 up 4085 but ... scene 3283 get 3169 f'm ...".

**Query:** A man is sitting on the staircase. Suddenly a car from the twenties rushes by and picks him up. That is the night he meets Ernest Hemingway.

# Representative Approaches

- Little work on this topic
- Infers objects, actions, their attributes, spatial relations from images, and nouns, verbs, adverbs and prepositions from text, and performs matching:

[Visual Semantic Search: Retrieving Videos via Complex Textual Queries](#)

Dahua Lin, Sanja Fidler, Chen Kong, Raquel Urtasun  
CVPR 2014

- Aligns plot synopses from fan sites with movies / TV series, performs retrieval based on alignment:

[Aligning Plot Synopses to Videos for Story-based Retrieval](#)

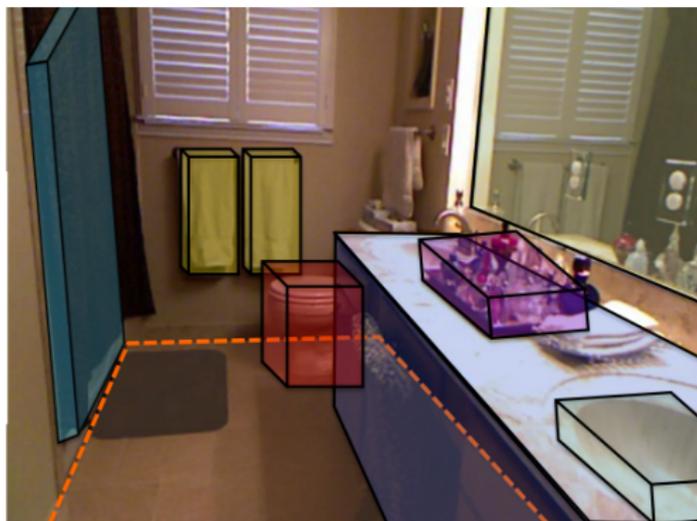
Makarand Tapaswi, Martin Baeuml, Rainer Stiefelhagen  
International Journal of Multimedia Information Retrieval (IJMIR), 2014

# Using Text to Improve Visual Parsing

# Scene Understanding with Natural Text

Understanding what you are told:

- Exploit the information in the provided description
- Determining which visual objects the text is referring to

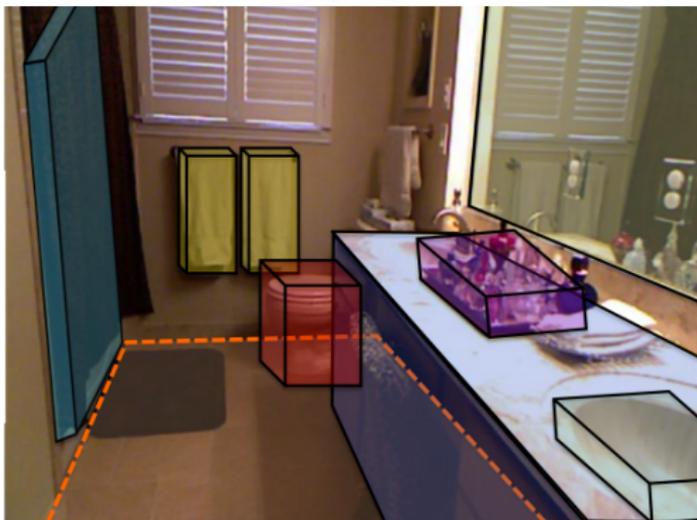


“Next to the toilet in the bathroom are two white towels. Bring them to me!”

# Scene Understanding with Natural Text

Understanding what you are told:

- Exploit the information in the provided description
- Determining which visual objects the text is referring to



“Next to the toilet in the bathroom are two white towels. Bring them to me!”

# Representative Approaches

- Use nouns, attributes and prepositions from text to boost object detectors in an image

## [A Sentence is Worth a Thousand Pixels](#)

S. Fidler, A. Sharma, R. Urtasun  
CVPR, 2013

## [What are you talking about? Text-to-Image Coreference](#)

C. Kong, D. Lin, M. Bansal, R. Urtasun, S. Fidler  
CVPR, 2014

- “Seeing” described actions/objects in videos:

## [Seeing What You're Told: Sentence-Guided Activity Recognition In Video](#)

N. Siddharth, Andrei Barbu, Jeffrey Siskind  
CVPR, 2014

- Improving action recognition using noun-verb statistics:

## [Robots with Language: Multi-Label Visual Recognition Using NLP](#)

Y. Yang, C. L. Teo, C. Fermuller, Y. Aloimonos  
ICRA, 2013

# Topics that Involve Images and Text

- Detecting text in images
- Generating textual descriptions of images/videos
- Visual retrieval based on complex textual queries
- Word-sense disambiguation
- Text to image/video alignment
- Learning visual models via text
- Using text to improve visual parsing
- Questions and answers