

IM GENET

Where have we been? Where are we going?

LI FEI-FEI & JIA DENG



The Beginning: CVPR 2009



ImageNet: A Large-Scale Hierarchical Image Database

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei
Dept. of Computer Science, Princeton University, USA
{jia Deng, wdong, socher, lija, li, feifei}@cs.princeton.edu

Abstract

The explosion of image data on the Internet has the potential to foster more sophisticated and robust models and algorithms to index, retrieve, organize and interact with images and multimedia data. But exactly how such data can be harvested and organized remains a critical problem. We introduce here a new database called "ImageNet", a large-scale ontology of images built upon the backbone of the WordNet structure. ImageNet aims to populate the majority of the 80,000 synsets of WordNet with an average of 500-1000 clean and full-resolution images. This will result in tens of millions of annotated images organized by the semantic hierarchy of WordNet. This paper offers a detailed analysis of ImageNet in its current state: 12 subnets with 5247 synsets and 3.2 million images in total. We show that ImageNet is much larger in scale and diversity and much more accurate than the current image datasets. Constructing such a large-scale database is a challenging task. We describe the data collection scheme with Amazon Mechanical Turk. Lastly, we illustrate the usefulness of ImageNet through three simple applications in object recognition, image classification and automatic object clustering. We hope that the scale, accuracy, diversity and hierarchical structure of ImageNet can offer unprecedented opportunities to researchers in the computer vision community and beyond.

1. Introduction

The digital era has brought with it an enormous explosion of data. The latest estimations put a number of more than 3 billion photos on Flickr, a similar number of video clips on YouTube and an even larger number for images in the Google Image Search database. More sophisticated and robust models and algorithms can be proposed by exploiting these images, resulting in better applications: face maps to index, retrieve, organize and interact with these data. But exactly how such data can be utilized and organized is a problem yet to be solved. In this paper, we introduce a new image database called "ImageNet", a large-scale ontology of images. We believe that a large-scale ontology of images is a critical resource for developing advanced, large-scale

content-based image search and image understanding algorithms, as well as for providing critical training and benchmarking data for such algorithms.

ImageNet uses the hierarchical structure of WordNet [1]. Each meaningful concept in WordNet, possibly described by multiple words or word phrases, is called a "synonym set" or "synset". There are around 80,000 sense synsets in WordNet. In ImageNet, we aim to provide on average 500-1000 images to illustrate each synset. Images of each concept are quality-controlled and human-annotated as described in Sec. 3.2. ImageNet, therefore, will offer tens of millions of clearly sorted images. In this paper, we report the current version of ImageNet, consisting of 12 "subnets": mammal, bird, fish, reptile, amphibian, vehicle, furniture, musical instrument, geological formation, and flower, fruit. These subnets contain 5247 synsets and 3.2 million images. Fig. 1 shows a snapshot of two branches of the mammal and vehicle subnets. The database is publicly available at <http://www.image-net.org>.

The rest of the paper is organized as follows: We first show that ImageNet is a large-scale, accurate and diverse image database (Section 2). In Section 4, we present a few simple application examples by exploiting the current ImageNet, mostly the mammal and vehicle subnets. Our goal is to show that ImageNet can serve as a useful resource for visual recognition applications such as object recognition, image classification and object localization. In addition, the construction of such a large-scale and high-quality database can no longer rely on traditional data collection methods. Sec. 3 describes how ImageNet is constructed by leveraging Amazon Mechanical Turk.

2. Properties of ImageNet

ImageNet is built upon the hierarchical structure provided by WordNet. In its completion, ImageNet aims to contain in the order of 50 million clearly labeled full-resolution images (500-1000 per synset). At the time this paper is written, ImageNet consists of 12 subnets. Most analysis will be based on the mammal and vehicle subnets.

Scale ImageNet aims to provide the most comprehensive and diverse coverage of the image world. The current 12 subnets consist of a total of 3.2 million clearly annotated

1
of images, but only a subset of 60k images are publicly available. Objects in images should have variable appearances, positions,

page datasets, such as Flickr's Photo 2006 dataset based on the Word 2006. All annotations are from 1/20/06. All images are the work of the authors and the Mechanical Turk workers by Turk and Amazon. © 2009 IEEE. All Rights Reserved. This is the final version.

The Impact of IM GENET

IMGENET on Google Scholar

4,386
Citations

[Imagenet: A large-scale hierarchical image database](#)

[J Deng, W Dong, R Socher, LJ Li, K Li...](#) - Computer Vision and ..., 2009 - [ieeexplore.ieee.org](#)

Abstract: The explosion of image data on the Internet has the potential to foster more sophisticated and robust models and algorithms to index, retrieve, organize and interact with images and multimedia data. But exactly how such data can be harnessed and organized

[Cited by 4386](#) [Related articles](#) [All 30 versions](#) [Cite](#) [Save](#)

2,847
Citations

[Imagenet large scale visual recognition challenge](#)

[O Russakovsky, J Deng, H Su, J Krause...](#) - International Journal of ..., 2015 - Springer

Abstract The **ImageNet** Large Scale Visual Recognition Challenge is a benchmark in object category classification and detection on hundreds of object categories and millions of images. The challenge has been run annually from 2010 to present, attracting participation

[Cited by 2847](#) [Related articles](#) [All 17 versions](#) [Cite](#) [Save](#)

...and many more.

From IMAGENET Challenge Contestants to Startups



A Revolution in Deep Learning

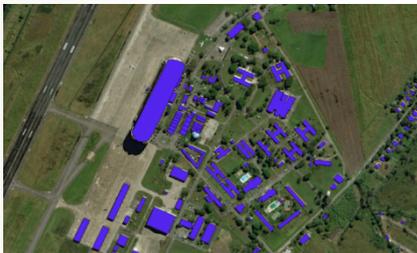


FORTUNE

*Why Deep Learning is Suddenly
Changing Your Life*

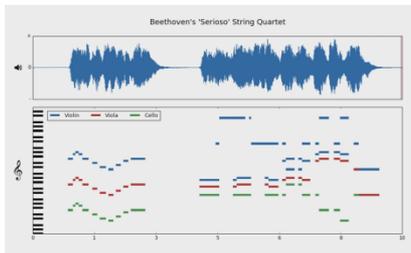
By Roger Parloff, Sept, 2016

“The IMGENET of x ”



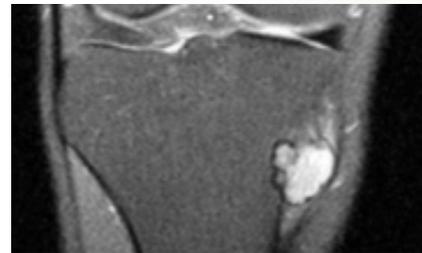
SpaceNet

DigitalGlobe, CosmiQ Works, NVIDIA



MusicNet

J. Thickstun et al, 2017



Medical ImageNet

Stanford Radiology, 2017



ShapeNet

A.Chang et al, 2015



EventNet

G. Ye et al, 2015



ActivityNet

F. Heilbron et al, 2015

An Explosion of Datasets

kaggle™

1627

Hosted Datasets

276

Commercial
Competitions

1919

Student
Competitions

1MM

Data Scientists

4MM

ML Models
Submitted

“Datasets—not algorithms—might be the key limiting factor to development of human-level artificial intelligence.”

ALEXANDER WISSNER-GROSS

Edge.org, 2016

The Untold History of

IMGENET

Hardly the First Image Dataset



Segmentation (2001)

D. Martin, C. Fowlkes, D. Tal, J. Malik.



CMU/VASC Faces (1998)

H. Rowley, S. Baluja, T. Kanade



FERET Faces (1998)

P. Phillips, H. Wechsler, J. Huang, P. Raus



COIL Objects (1996)

S. Nene, S. Nayar, H. Murase



MNIST digits (1998-10)

Y LeCun & C. Cortes



KTH human action (2004)

I. Leptev & B. Caputo



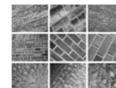
Sign Language (2008)

P. Buehler, M. Everingham, A. Zisserman



UIUC Cars (2004)

S. Agarwal, A. Awan, D. Roth



3D Textures (2005)

S. Lazebnik, C. Schmid, J. Ponce



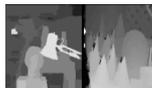
CuRRET Textures (1999)

K. Dana B. Van Ginneken S. Nayar
J. Koenderink



CAVIAR Tracking (2005)

R. Fisher, J. Santos-Victor J. Crowley



Middlebury Stereo (2002)

D. Scharstein R. Szeliski



CalTech 101/256 (2005)

Fei-Fei et al, 2004
Griffin et al, 2007



LabelMe (2005)

Russell et al, 2005



ESP (2006)

Ahn et al, 2006



MSRC (2006)

Shotton et al. 2006



PASCAL (2007)

Everingham et al, 2009



**Lotus Hill
(2007)**

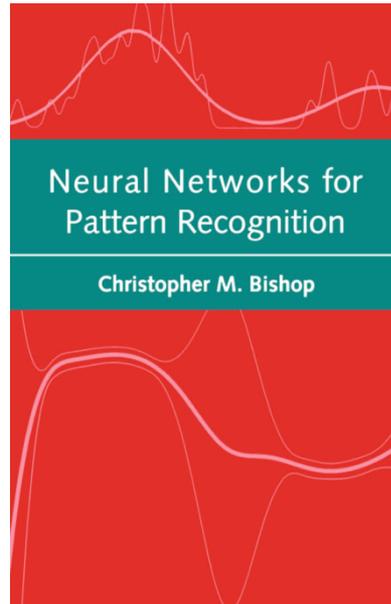
Yao et al, 2007



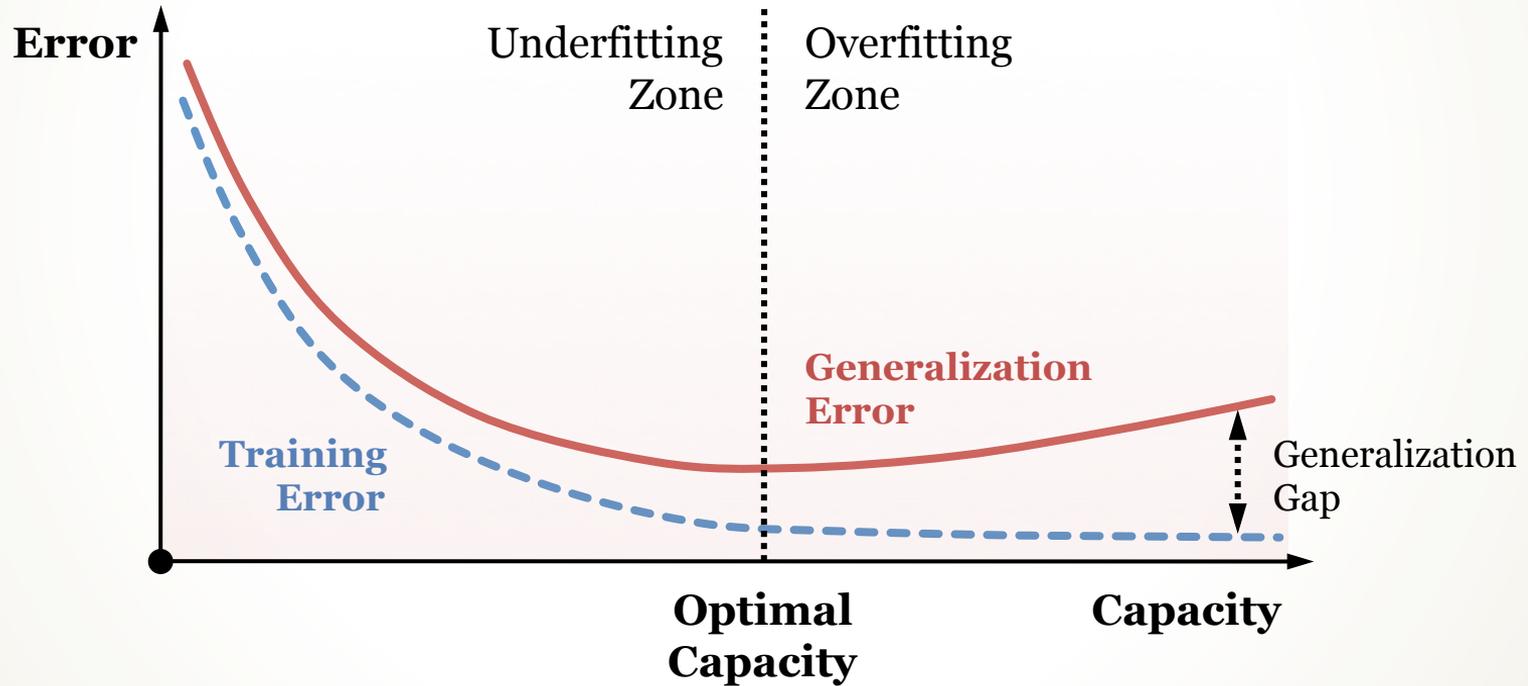
TinyImage (2008)

Torralba et al. 2008

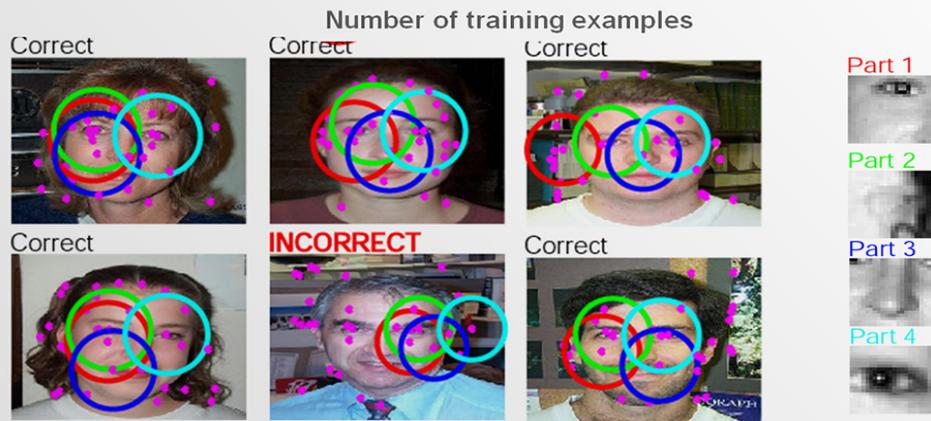
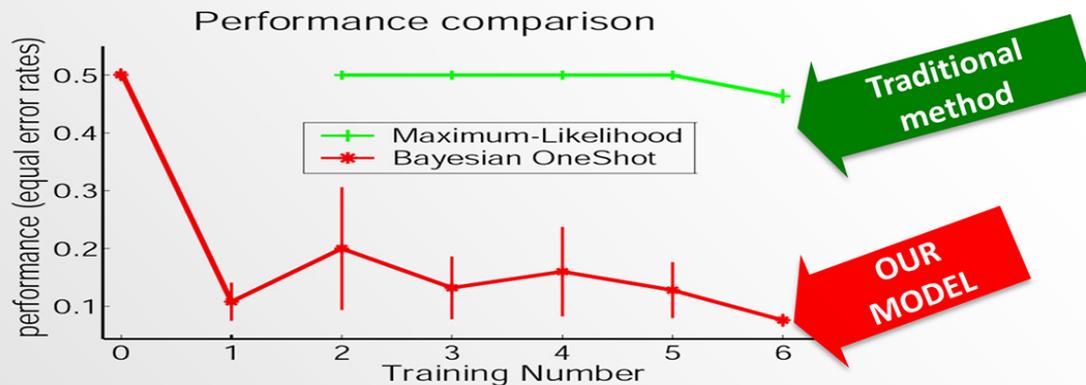
A Profound Machine Learning Problem Within Visual Learning



Machine Learning 101: Complexity, Generalization, Overfitting



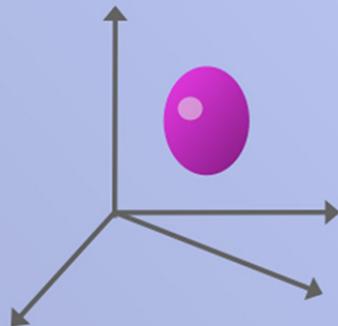
One-Shot Learning



One-shot learning algorithm: Bayesian Variational Inference

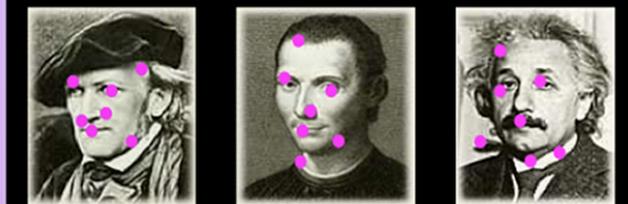
Model
initialization

Model:
Parameter update



new estimate
of $p(\theta|\text{train})$

Data: model fitting

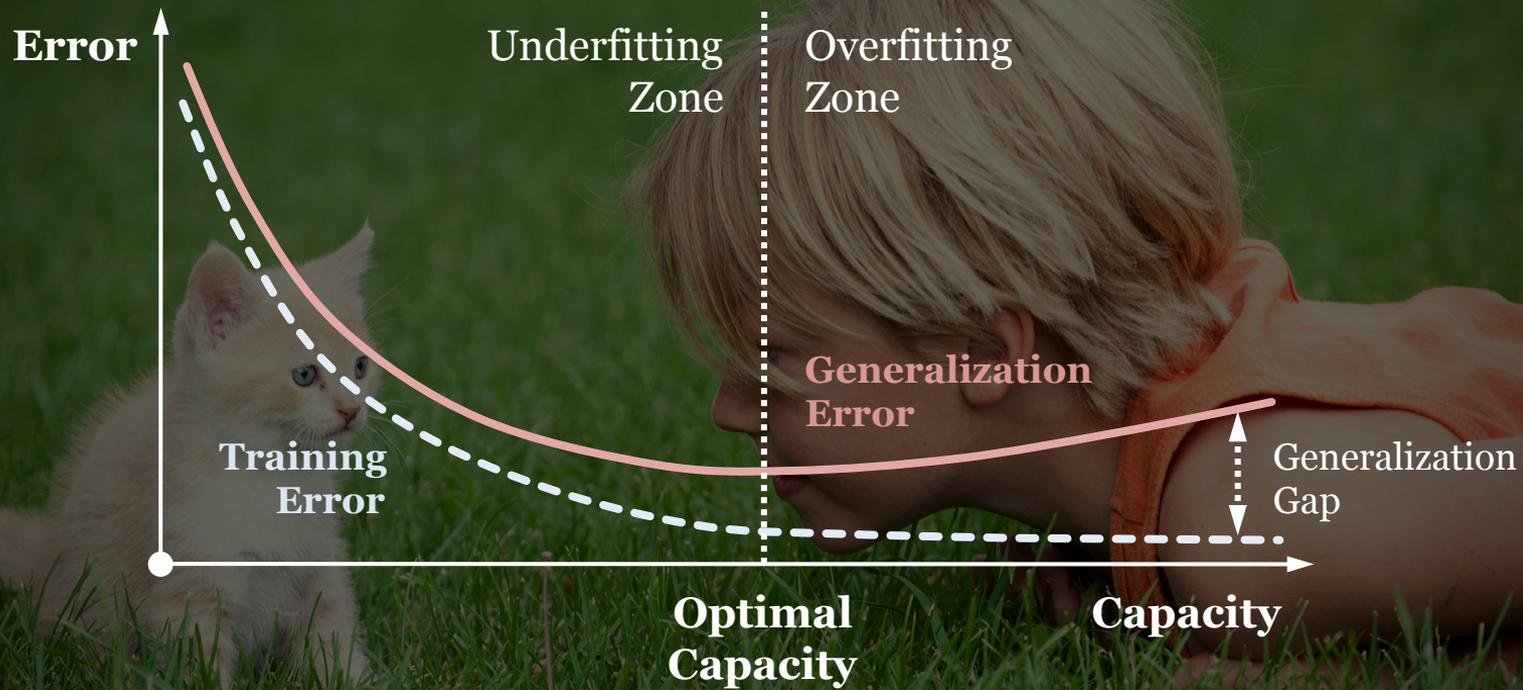


prior statistics of $p(\theta)$



How Children Learn to See





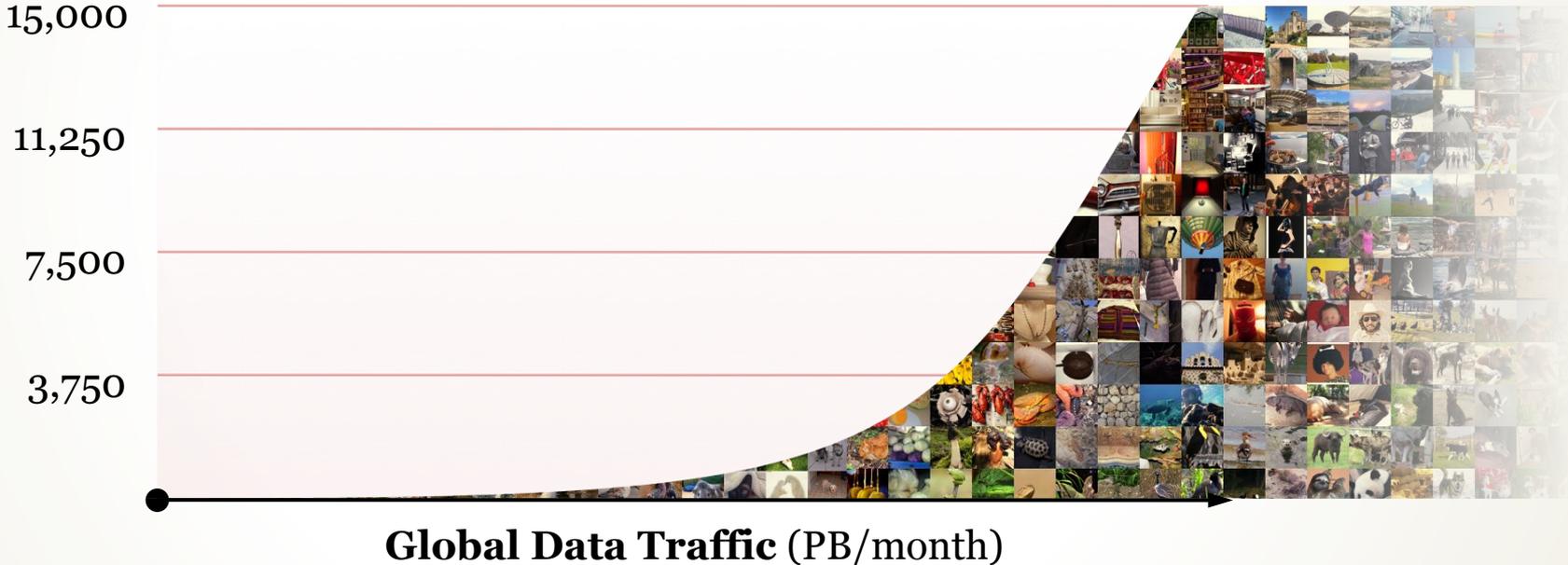
A new way of thinking...

To shift the focus of Machine Learning for visual recognition

from
modeling...

...to data.
Lots of data.

Internet Data Growth 1990-2010



Source: Cisco

What is WordNet?



Original paper by
**[George Miller, et
al 1990]** cited over
5,000 times

Organizes over
150,000 words into
117,000 categories
called *synsets*.

Establishes
ontological and
lexical relationships
in NLP and related
tasks.



Christiane Fellbaum

Senior Research Scholar

Computer Science Department, Princeton

President, Global WordNet Consortium

Individually Illustrated WordNet Nodes



jacket: a short coat



German shepherd: breed of large shepherd dogs used in police work and as a guide for the blind.



microwave: kitchen appliance that cooks food by passing an electromagnetic wave through it.



mountain: a land mass that projects well above its surroundings; higher than a hill.

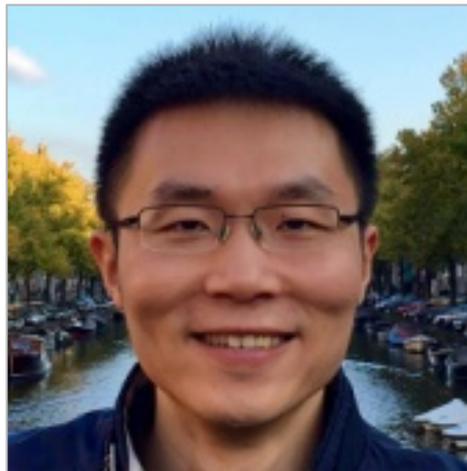
A massive ontology of images to transform computer vision

The background of the right side of the slide is a large, dense mosaic of thousands of small, square images. The images are arranged in a grid and cover the entire right half of the slide. The mosaic is composed of a wide variety of subjects, including landscapes, animals, objects, and abstract patterns, creating a complex and colorful visual texture.

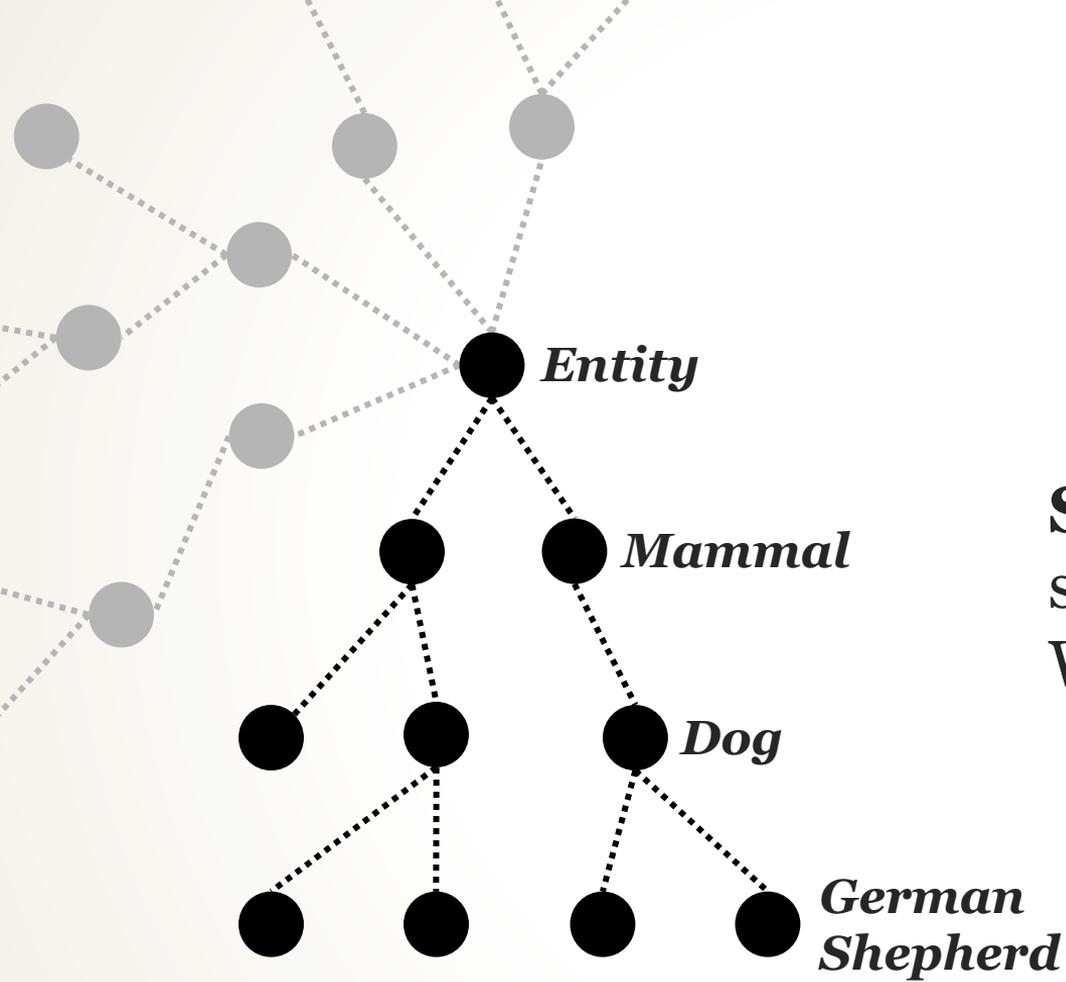
IM GENET Comrades



Prof. Kai Li
Princeton



Jia Deng
1st Ph.D. student
Princeton



Step 1: Ontological structure based on WordNet

Dog



German Shepherd

Step 2: Populate categories with thousands of images from the Internet

Dog

German Shepherd



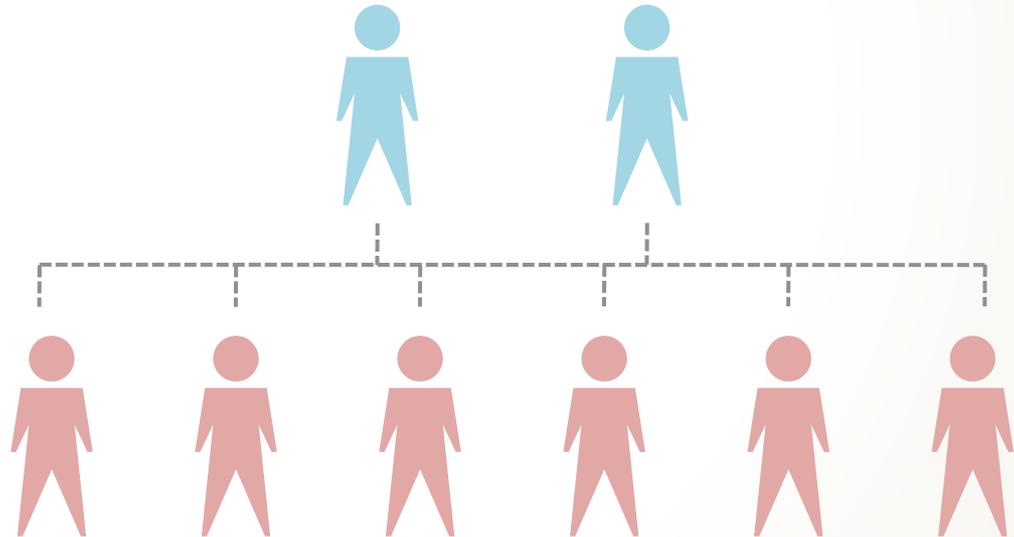
Step 3: Clean results by hand

Three Attempts at Launching IMGENET

1st Attempt: The Psychophysics Experiment

**ImageNet PhD
Students**

**Miserable
Undergrads**



1st Attempt: The Psychophysics Experiment

- # of synsets: **40,000** (subject to: imageability analysis)
- # of candidate images to label per synset: **10,000**
- # of people needed to verify: **2-5**
- Speed of human labeling: **2 images/sec** (one fixation: ~200msec)
- **Massive parallelism (N ~ 10²⁻³)**

$$40,000 \times 10,000 \times 3 / 2 = 6000,000,000 \text{ sec} \quad \frac{\approx 19 \text{ years}}{N}$$

2nd Attempt: Human-in-the-Loop Solutions

Towards scalable dataset construction: An active learning approach

Brendan Collins, Jia Deng, Kai
{bmcollin, dengjia, li, feifei}

Department of Computer Science, Princeton

Abstract. As computer vision research continues to advance, more and greater variation within object categories and more exhaustive datasets are necessary. Having such datasets is laborious and monotonous in which many images have been automatically collected from the internet (typically by automatic internet search engines) and are often irrelevant. We present a novel approach which employs active, online learning to collect relevant images from noise. We present a dataset which employs active, online learning to collect relevant images from noise. The principle advantage of this approach is its scalability. We demonstrate that our approach is superior to the state-of-the-art, with scalable performance.

1 Introduction

Though it is difficult to foresee the future of computer vision, it is clear that its trajectory will include examining a great variety of images (such as objects or scenes), that the complexity

OPTIMOL: automatic Online Picture collection via Incremental Model Learning

Li-Jia Li¹, Gang Wang¹ and Li Fei-Fei²

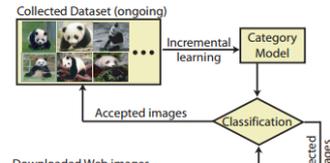
¹ Dept. of Electrical and Computer Engineering, University of Illinois Urbana-Champaign, USA

² Dept. of Computer Science, Princeton University, USA

ljali3@uiuc.edu, gwang6@uiuc.edu, feifeili@cs.princeton.edu

Abstract

A well-built dataset is a necessary starting point for advanced computer vision research. It plays a crucial role in evaluation and provides a continuous challenge to state-of-the-art algorithms. Dataset collection is, however, a tedious and time-consuming task. This paper presents a novel automatic dataset collecting and model learning approach



2nd Attempt: Human-in-the-Loop Solutions



Machine-generated
datasets can only match
the best algorithms of
the time.



Human-generated
datasets transcend
algorithmic limitations,
leading to better
machine perception.

3rd Attempt: A Godsend Emerges

**ImageNet PhD
Students**



**Crowdsourced
Labor**

amazon **mechanical turk**TM
Artificial Artificial Intelligence

49k Workers *from* 167 Countries
2007-2010

The Result: IMAGENET Goes Live in 2009

IMAGENET SEARCH

14,187,122 images, 21841 synsets indexed

Home About Explore Download

Not logged in. [Login](#) | [Signup](#)

Yellow sand verbena, *Abronia latifolia*

Plant having hemispherical heads of yellow trumpet-shaped flowers, found in coastal dunes from California to British Columbia

200 pictures 15.34% Popularity Percentile Wordnet IDs

Numbers in brackets: (the number of synsets in the subtree).

- ImageNet 2011 Fall Release (32326)
 - plant, flora, plant life (4486)
 - phytoplankton (2)
 - microflora (0)
 - crop (9)
 - endemic (0)
 - holophyte (0)
 - non-flowering plant (0)
 - plantlet (0)
 - wildling (141)
 - wildflower, wild flower (140)
 - sagebrush buttercup, Ra
 - pasqueflower, pasque fic
 - meadow rue (0)
 - sand verbena (6)
 - snowball, sweet sand
 - sweet sand verbena,
 - yellow sand verbena,
 - beach pancake, Abro
 - beach sand verbena,
 - desert sand verbena,
 - trailing four o'clock, trailir
 - red maids, redmaids, Ca
 - siskiyou lewisia, Lewisia
 - bitterroot, Lewisia rediviv
 - pussy-paw, pussy-paws,
 - flame flower, flame-flowe
 - woolly daisy, dwarf daisy
 - heartleaf arnica, Arnica c
 - Arnica montana (0)

Reemap visualization Images of the Synset Downloads



Images of children synsets are not included. All images shown are thumbnails. Images may be subject to copyright.

Prev 1 2 3 4 5 6 Next

© 2010 Stanford Vision Lab, Stanford University, Princeton University, support@image-net.com, Copyright infringement

IM  GENET

What We Did Right

While Others Targeted Detail...



LabelMe

Per-Object Regions and Labels
Russell et al, 2005



Lotus Hill

Hand-Traced Parse Trees
Yao et al, 2007

...We Targeted Scale

SUN, 131K

[Xiao et al. '10]

LabelMe, 37K

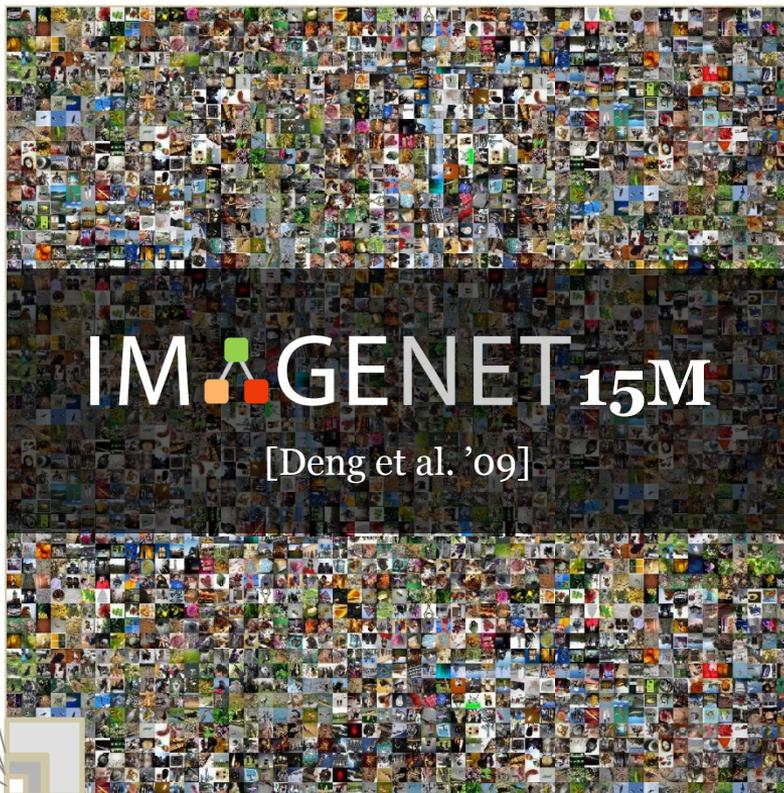
[Russell et al. '07]

PASCAL VOC, 30K

[Everingham et al. '06-'12]

Caltech101, 9K

[Fei-Fei, Fergus, Perona, '03]



Additional IMAGENET Goals



- Carnivore
- Canine
 - Dog
 - Working Dog
 - Husky



High Resolution

To better replicate human visual acuity

High-Quality Annotation

To create a benchmarking dataset and advance the state of machine perception, not merely reflect it

Free of Charge

To ensure immediate application and a sense of community

An Emphasis on Community and Achievement

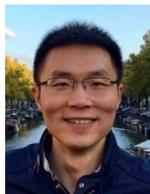
IM  GENET

**Large Scale Visual Recognition Challenge
(ILSVRC 2010-2017)**

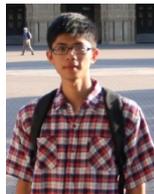
ILSVRC Contributors



Alex Berg
UNC Chapel Hill



Jia Deng
Univ. of Michigan



Zhiheng Huang
Stanford



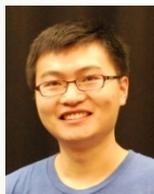
Aditya Khosla
Stanford



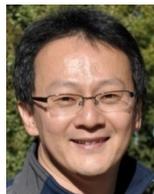
Jonathan Krause
Stanford



Fei-Fei Li
Stanford



Wei Liu
UNC Chapel Hill



Sean Ma
Stanford



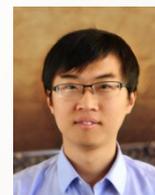
Eunbyung Park
UNC Chapel Hill



Olga Russakovsky
Stanford



Sanjeev Satheesh
Stanford



Hao Su
Stanford

Our Inspiration: PASCAL VOC



PASCAL2

Pattern Analysis, Statistical Modelling and
Computational Learning

2005-2012

Our Inspiration: PASCAL VOC

**Mark
Everingham**
1973-2012

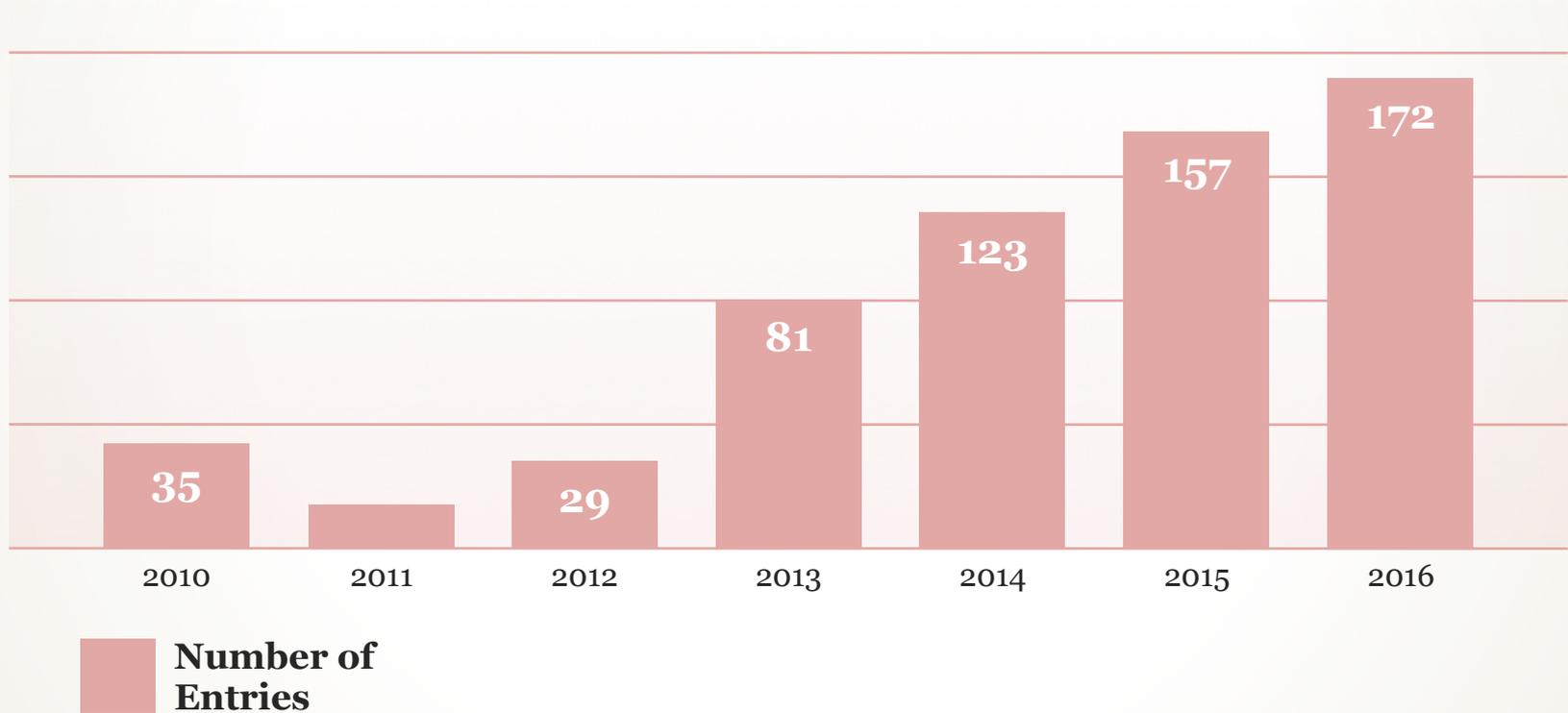


**Mark Everingham
Prize @ ECCV 2016**

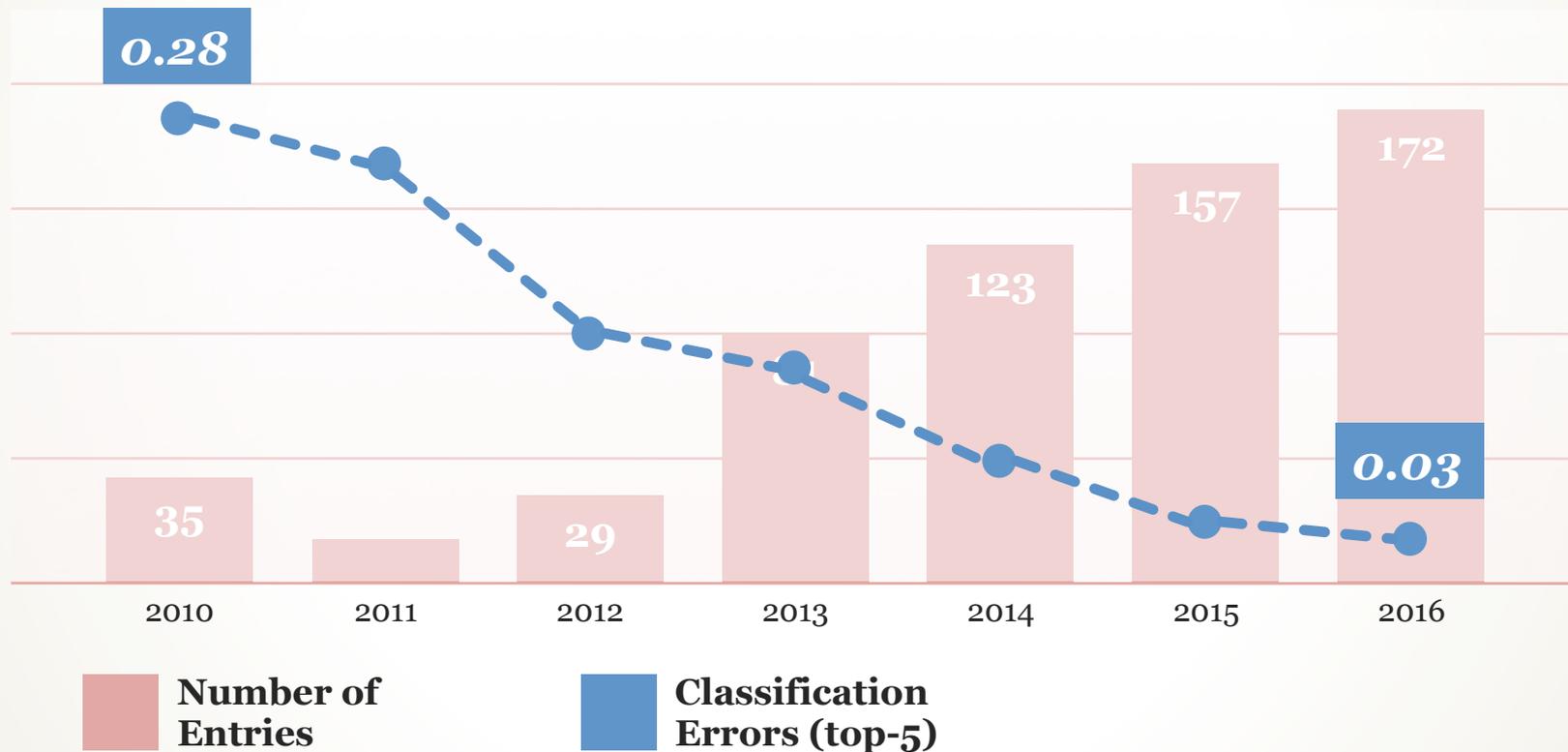
IM  **GENET**

Alex Berg, Jia Deng, Fei-Fei Li, Wei Liu,
Olga Russakovsky

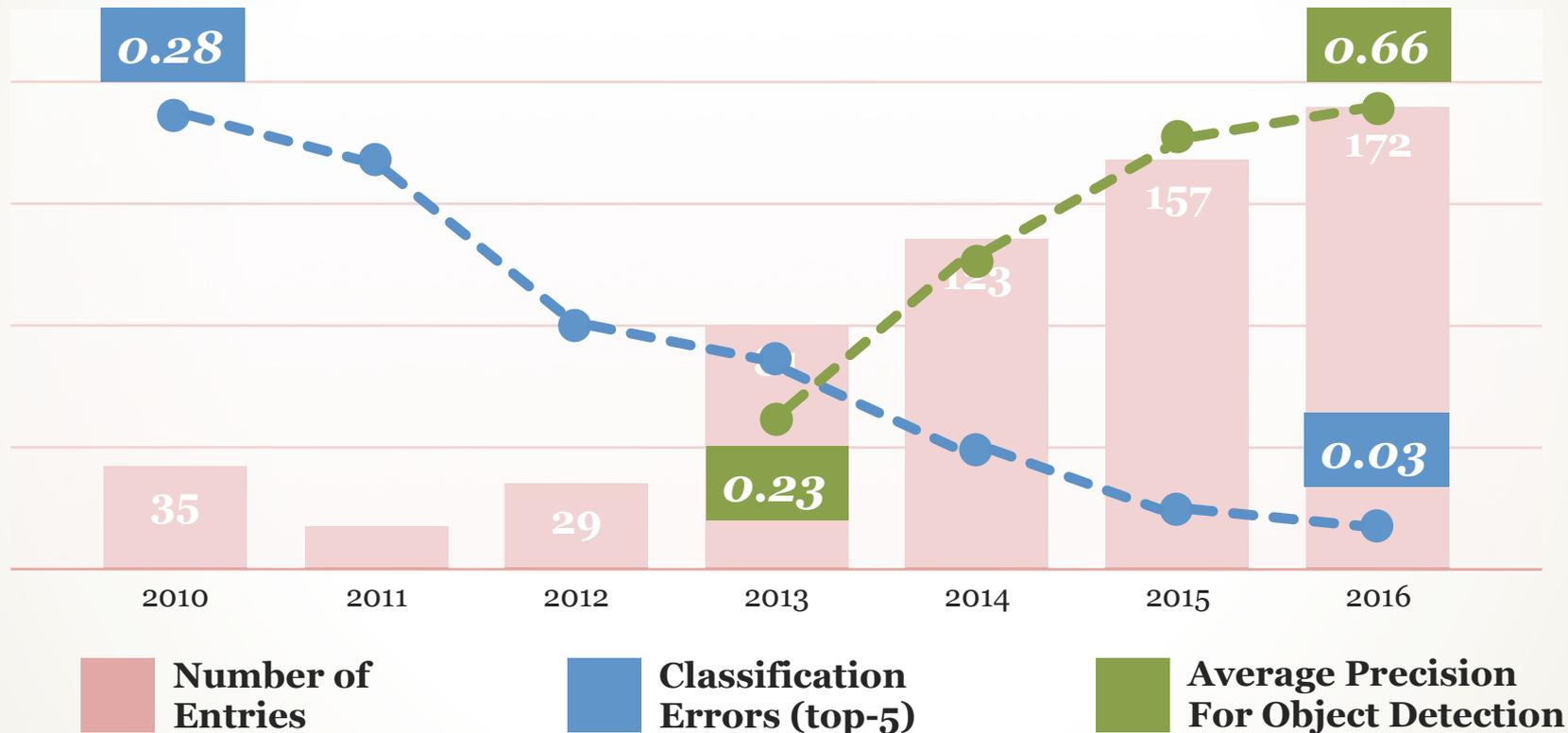
Participation and Performance



Participation and Performance

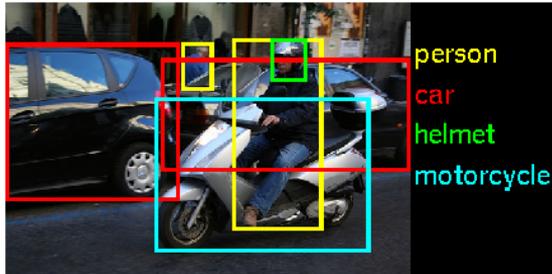
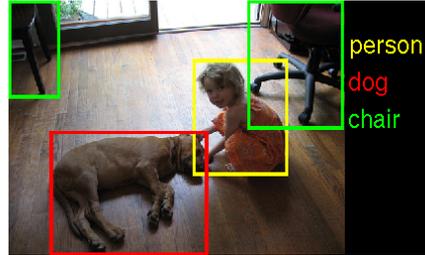
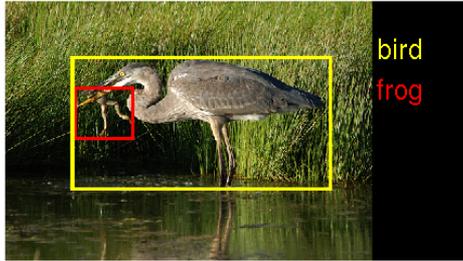


Participation and Performance



What we did to make
IM  GENET better

Lack of Details...ILSVRC Detection Challenge



| Statistics | PASCAL VOC 2012 | | ILSVRC 2013 |
|----------------|--------------------|-----|----------------|
| Object classes | 20 | 10x | 200 |
| Images | 5.7K | 70x | 395K |
| Objects | 13.6K | 25x | 345K |

Evaluation of ILSVRC Detection

Need to annotate the presence of all classes
(to penalize false detections)



| Table | Chair | Horse | Dog | Cat | Bird |
|-------|-------|-------|-----|-----|------|
| + | + | - | - | - | - |
| + | - | - | - | + | - |
| + | + | - | - | - | - |

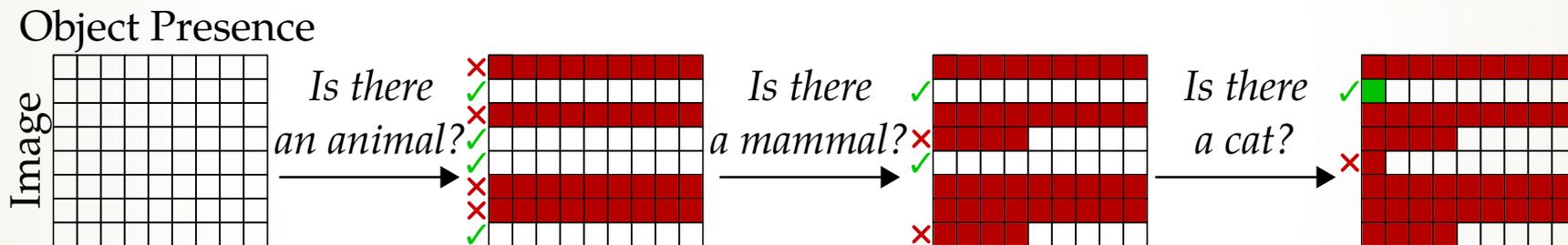
images: 400K

classes: 200

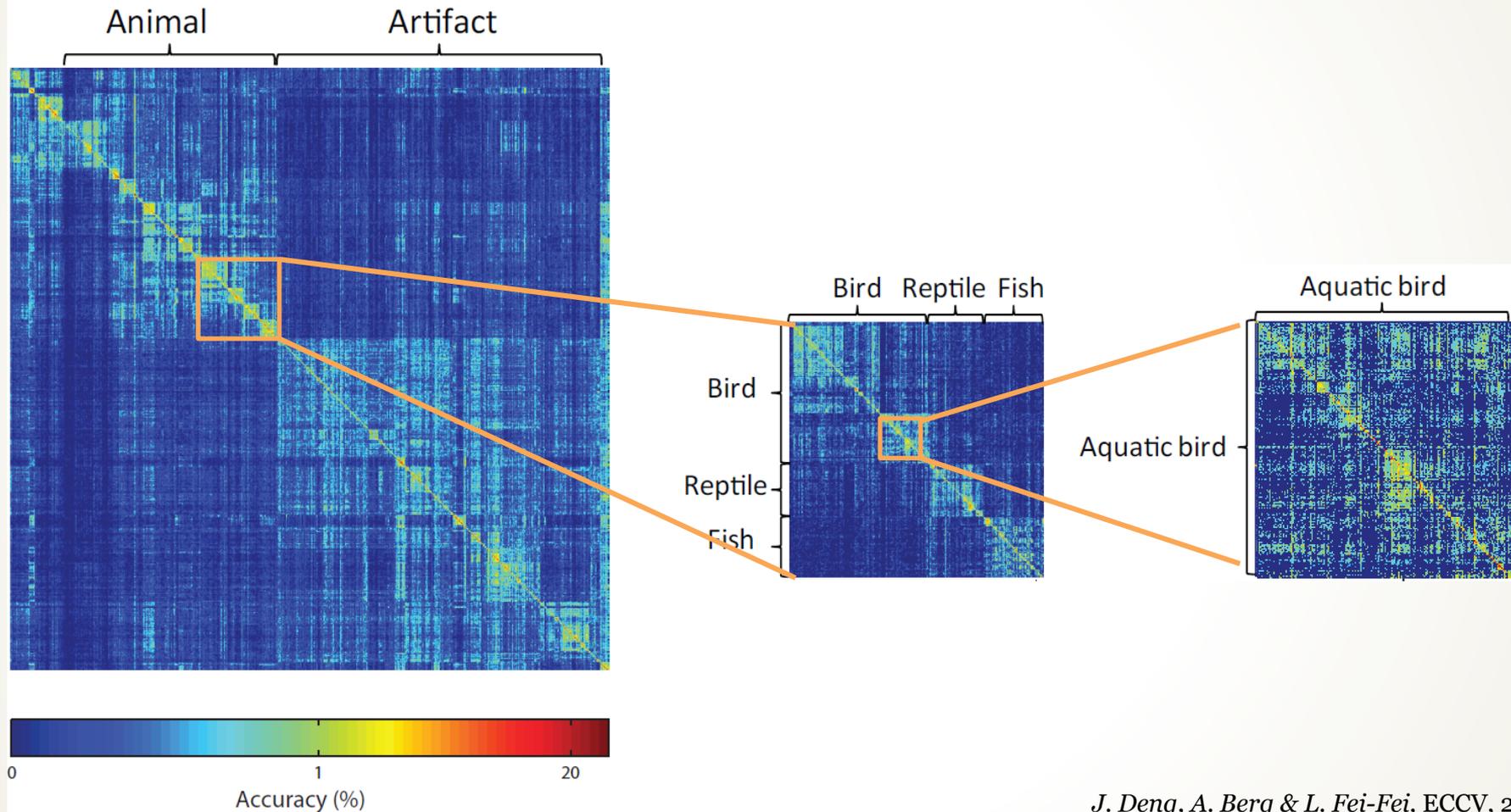
annotations = 80M!

Evaluation of ILSVRC Detection

Hierarchical annotation



What does classifying 10K+ classes tell us?



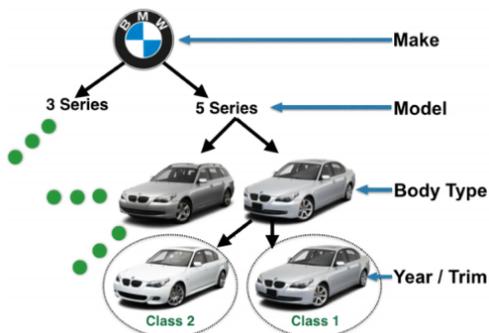
Fine-Grained Recognition



Fine-Grained Recognition

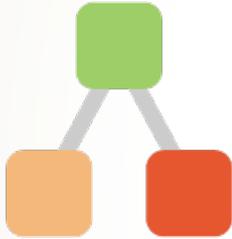
IMAGENET  cars

[Geburu, Krause, Deng, Fei-Fei, CHI 2017]



2567 classes
700k images

Expected Outcomes



ImageNet becomes a benchmark



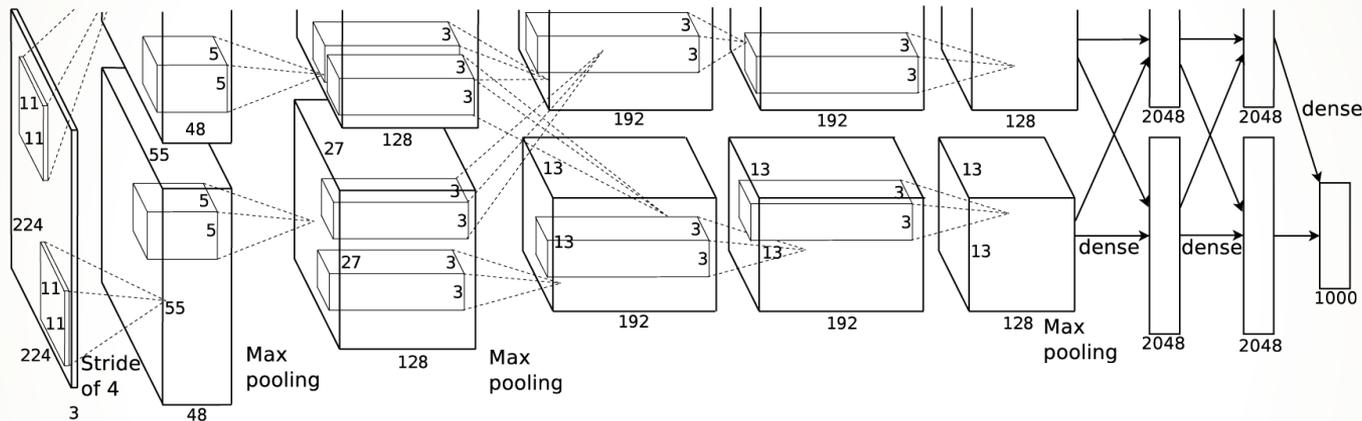
Breakthroughs in object recognition



Machine learning advances and changes dramatically

Unexpected Outcomes

Neural Nets are Cool Again!



13,259
Citations

[Imagenet classification with deep convolutional neural networks](#)

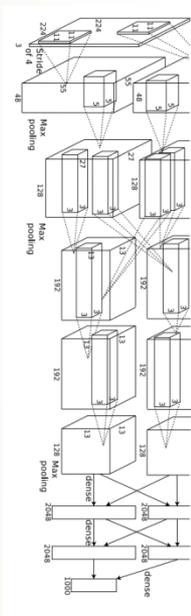
A Krizhevsky, I Sutskever, GE Hinton - Advances in neural ..., 2012 - papers.nips.cc

Abstract We trained a large, deep convolutional neural network to classify the 1.3 million high-resolution images in the LSVRC-2010 **ImageNet** training set into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 39.7% and 18.9%

[Cited by 13259](#) [Related articles](#) [All 95 versions](#) [Cite](#) [Save](#)

...And Cooler and Cooler ☺

“AlexNet”



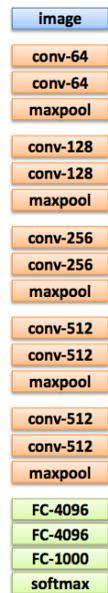
[Krizhevsky et al. NIPS 2012]

“GoogLeNet”



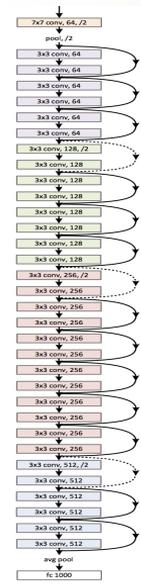
[Szegedy et al. CVPR 2015]

“VGG Net”

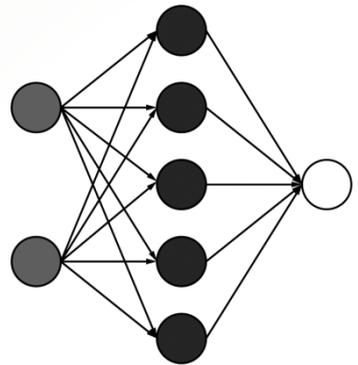


[Simonyan & Zisserman,
ICLR 2015]

“ResNet”



[He et al. CVPR 2016]



Neural Nets

IMAGENET



GPUs

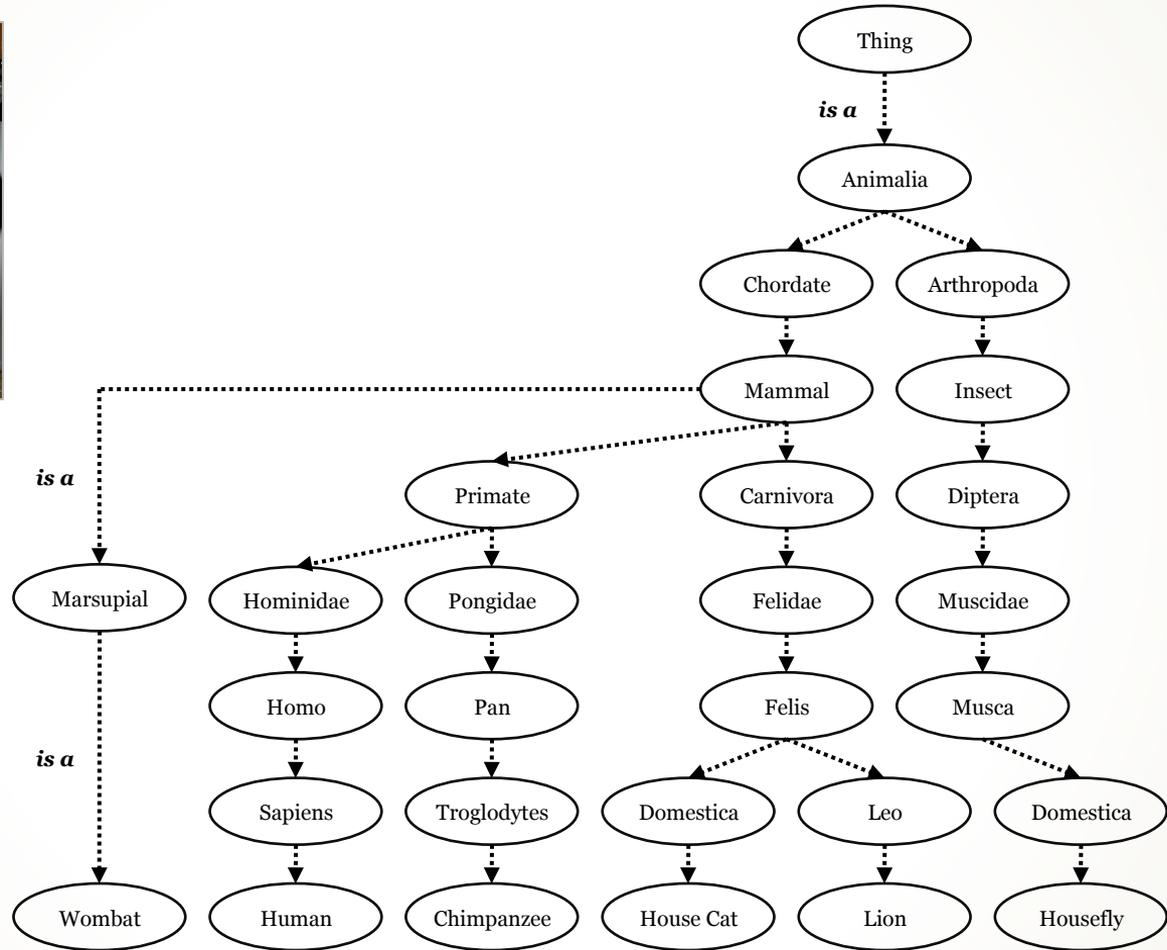
A Deep Learning Revolution

Ontological Structure Structure Not Used as Much



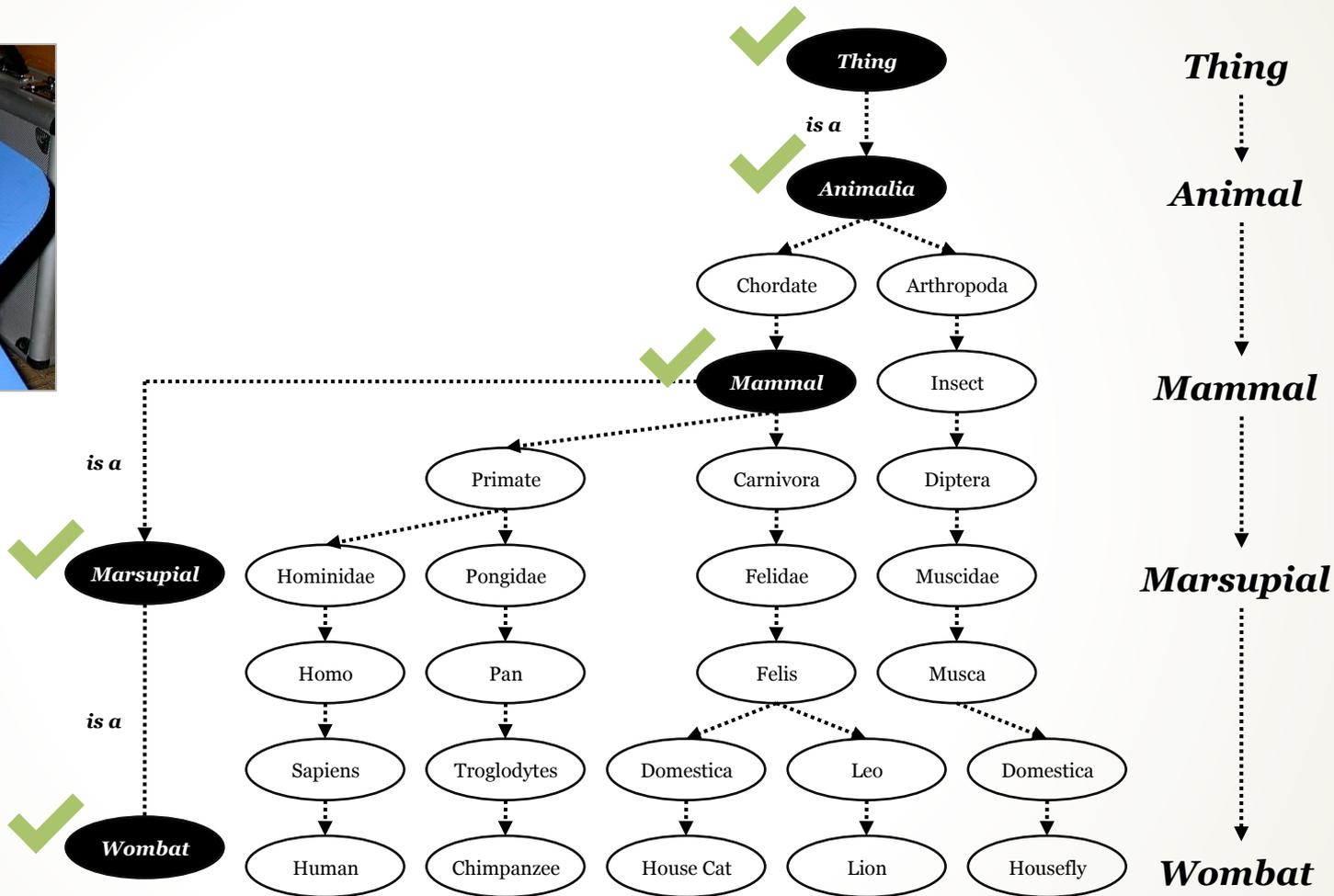


Wombat





Wombat





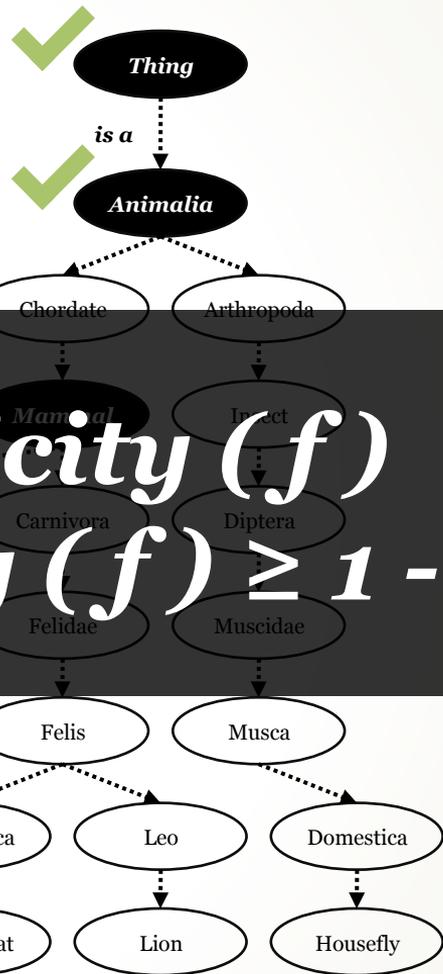
Wombat

is a

Marsupial

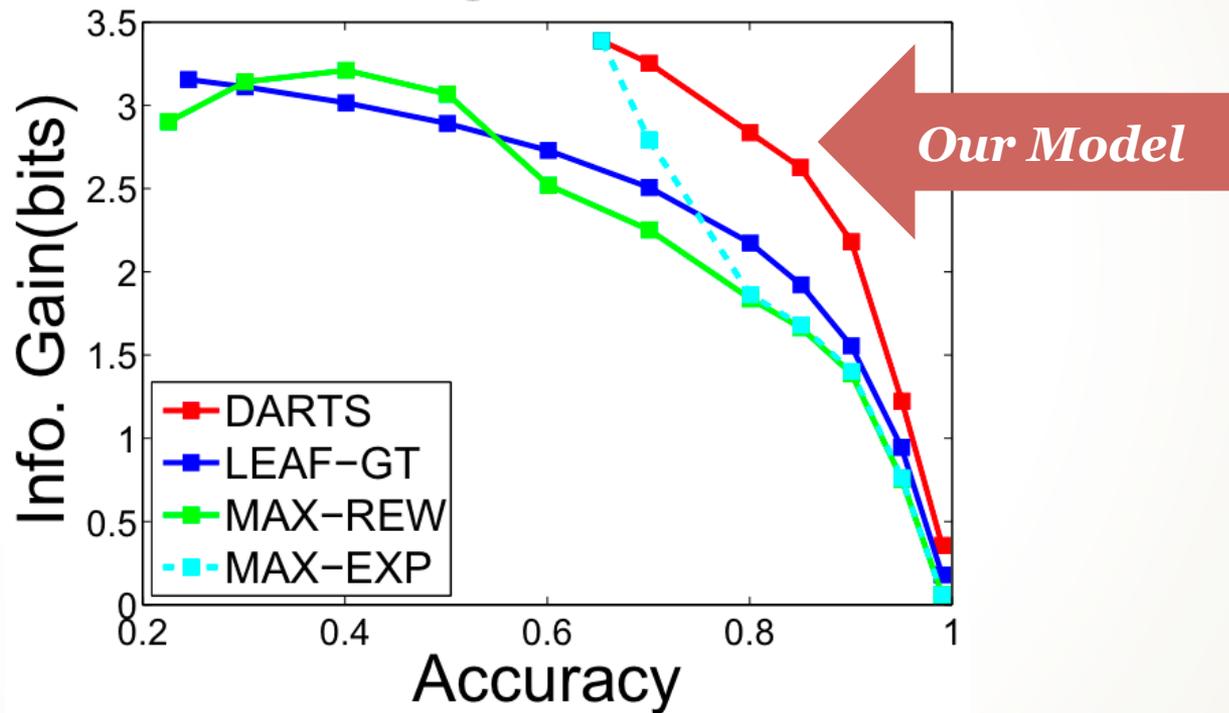
is a

Wombat



Maximize Specificity (f)
Subject to Accuracy (f) $\geq 1 - \epsilon$

Optimizing with a Knowledge Ontology Results in Big Gains in Information at Arbitrary Accuracy



Relatively Few Works Have Used Ontology



Kuettel, Guillaumin, Ferrari.
**Segmentation Propagation in
ImageNet. ECCV 2012**

ECCV 2012
Best paper Award

About 93 results (0.07 sec)

Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition

Search within citing articles

Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition

[S.Guadarrama](#), N.Krishnamoorthy... - Proceedings of the ..., 2013 - cv-foundation.org

Abstract Despite a recent push towards large-scale object recognition, activity recognition remains limited to narrow domains and small vocabularies of actions. In this paper, we tackle the challenge of recognizing and describing activities* in-the-wild". We present a Cited by 129 Related articles All 13 versions Cite Save

Reasoning about object affordances in a knowledge base representation

[Y.Zhu](#), [A.Faloutsos](#), [L.Fei-Fei](#) - European conference on computer vision, 2014 - Springer

Abstract Reasoning about objects and their affordances is a fundamental problem for visual intelligence. Most of the previous work casts this problem as a classification task where separate classifiers are trained to label objects, recognize attributes, or assign affordances. Cited by 78 Related articles All 7 versions Cite Save

TREETALK: Composition and Compression of Trees for Image Descriptions.

[P.Kuznetsova](#), [V.Ordonez](#), [T.L.Berg](#), [Y.Choi](#) - TACL, 2014 - pdfs.semanticscholar.org

Abstract We present a new tree based approach to composing expressive image descriptions that makes use of naturally occurring web images with captions. We investigate two related tasks: image caption generalization and generation, where the former is an Cited by 65 Related articles All 12 versions Cite Save More

From large scale image categorization to entry-level categories

[V.Ordonez](#), [J.Dana](#), [Y.Choi](#), [A.C.Berg](#)... - Proceedings of the IEEE ... , 2013 - cv-foundation.org

Abstract Entry level categories the labels people will use to name an object were originally defined and studied by psychologists in the 1980s. In this paper we study entrylevel categories at a large scale and learn the first models for predicting entry-level categories for Cited by 63 Related articles All 48 versions Cite Save

Integrating Language and Vision to Generate Natural Language Descriptions of Videos in the Wild.

[J.Thomason](#), [S.Venugopalan](#)... - ..., 2014 - ai2-s2-pdfs.s3.amazonaws.com

Abstract This paper integrates techniques in natural language processing and computer vision to improve recognition and description of entities and activities in real-world videos. We propose a strategy for generating textual descriptions of videos by using a factor graph Cited by 59 Related articles All 12 versions Cite Save More

Hierarchical Semantic Labeling for Task-Relevant RGB-D Perception.

[C.Wu](#), [L.Lenz](#), [A.Saxena](#) - Robotics: Science and systems, 2014 - pdfs.semanticscholar.org

Abstract—Semantic labeling of RGB-D scenes is very important in enabling robots to perform mobile manipulation tasks, but different tasks may require entirely different sets of labels. For example, when navigating to an object, we may need only a single label Cited by 47 Related articles All 13 versions Cite Save More



Most works still use 1M images
to do pre-training

1M Images

15M Images Total

“First, we find that the performance on vision tasks still increases linearly with orders of magnitude of training data size.”

C. Sun et al, 2017

Revisiting Unreasonable Effectiveness of Data in Deep Learning Era

Chen Sun¹, Abhinav Shrivastava^{1,2}, Saurabh Singh¹, and Abhinav Gupta^{1,2}

¹Google Research
²Carnegie Mellon University

Abstract

The success of deep learning in vision can be attributed to: (a) models with high capacity; (b) increased computational power; and (c) availability of large-scale labeled data. Since 2012, there have been significant advances in representation capabilities of the models and computational capabilities of GPUs. But the size of the biggest dataset has surprisingly remained constant. What will happen if we increase the dataset size by $10\times$ or $100\times$? This paper takes a step towards clearing the clouds of mystery surrounding the relationship between ‘enormous data’ and deep learning. By exploiting the JFT-300M dataset which has more than 375M noisy labels for 300M images, we investigate how the performance of current vision tasks would change if this data was used for representation learning. Our paper delivers some surprising (and some expected) findings. First, we find that the performance on vision tasks still increases linearly with orders of magnitude of training data size. Second, we show that representation learning (or pre-training) still holds a lot of promise. One can improve performance on any vision tasks by just training a better base model. Finally, as expected, we present new state-of-the-art results for different vision tasks including image classification, object detection, semantic segmentation and human pose estimation. Our sincere hope is that this inspires vision community to not undervalue the data and develop collective efforts in building larger datasets.

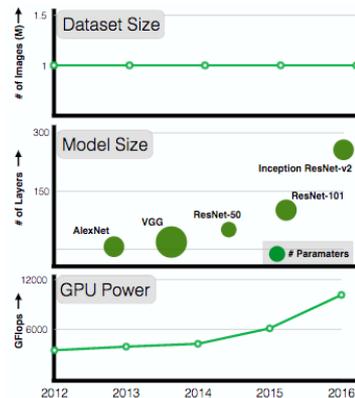


Figure 1. The Curious Case of Vision Datasets: While GPU computation power and model sizes have continued to increase over the last five years, size of the largest training dataset has surprisingly remained constant. Why is that? What would have happened if we have used our resources to increase dataset size as well? This paper provides a sneak-peek into what could be if the dataset sizes are increased dramatically.

ously, while both GPUs and model capacity have continued to grow, datasets to train these models have remained stagnant. Even a 101 layer ResNet with significantly more

How Humans Compare



How Humans Compare

Human

5.1%
Top-5 error rate

Susceptible to:

- Fine-grained recognition
- Class unawareness
- Insufficient training data

GoogLeNet

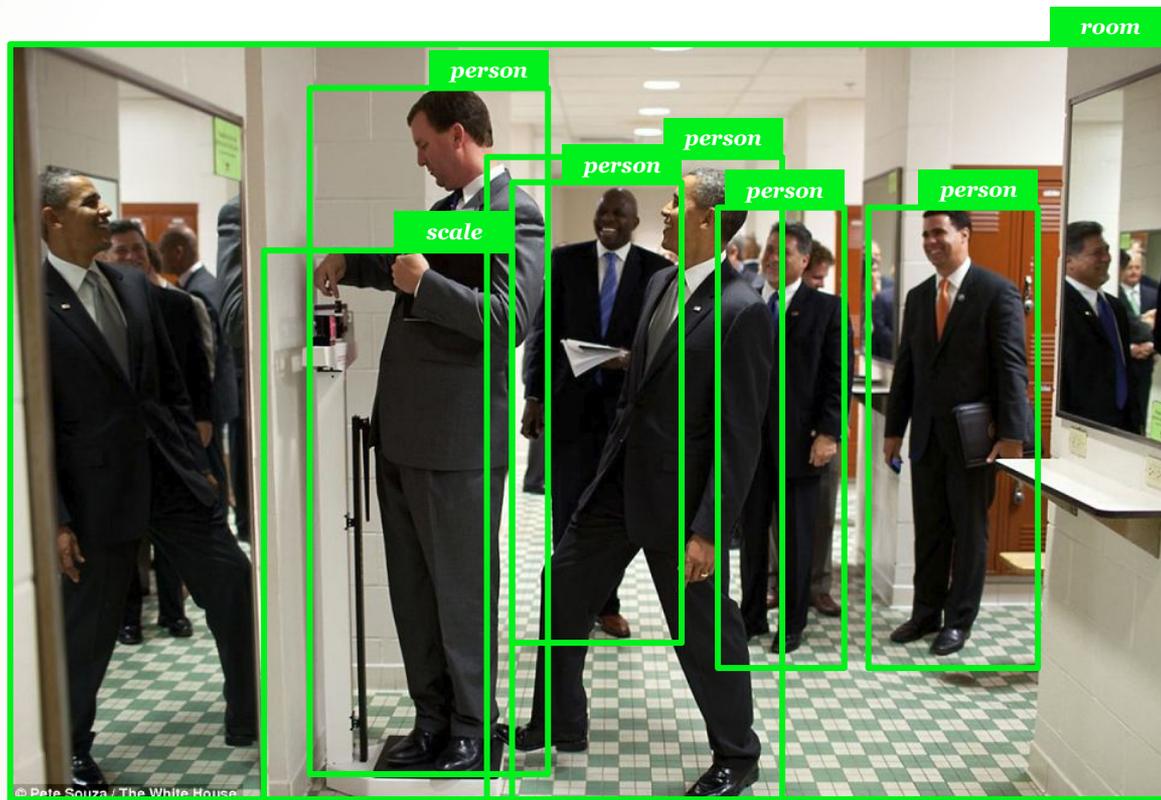
6.8%
Top-5 error rate

Susceptible to:

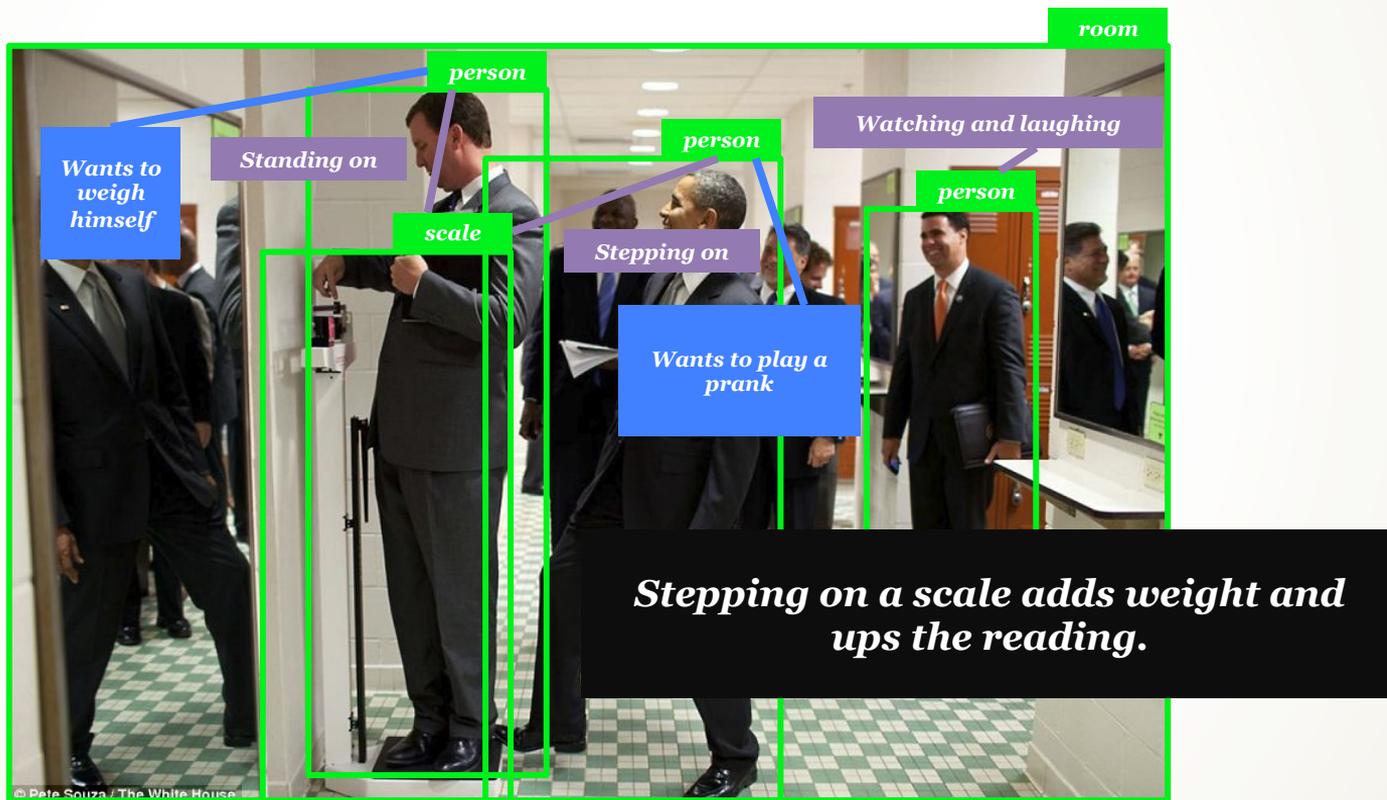
- Small, thin objects
- Image filters
- Abstract representations
- Miscellaneous sources

What Lies Ahead

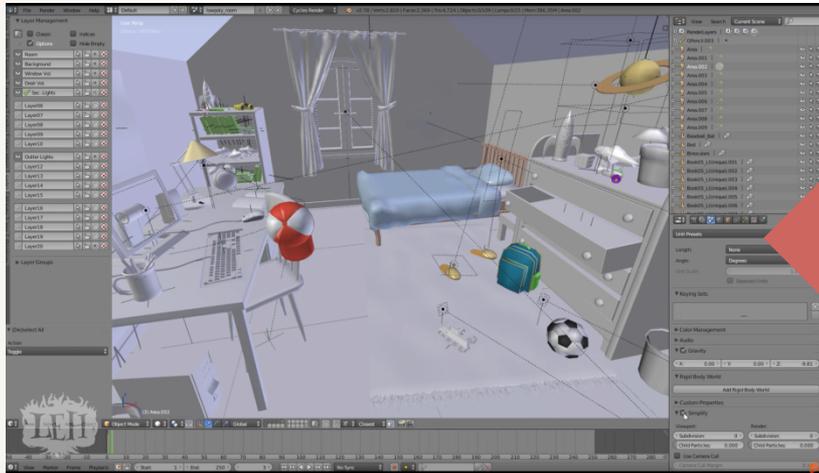
Moving from object recognition...



...to human-level understanding.



Inverse Graphics



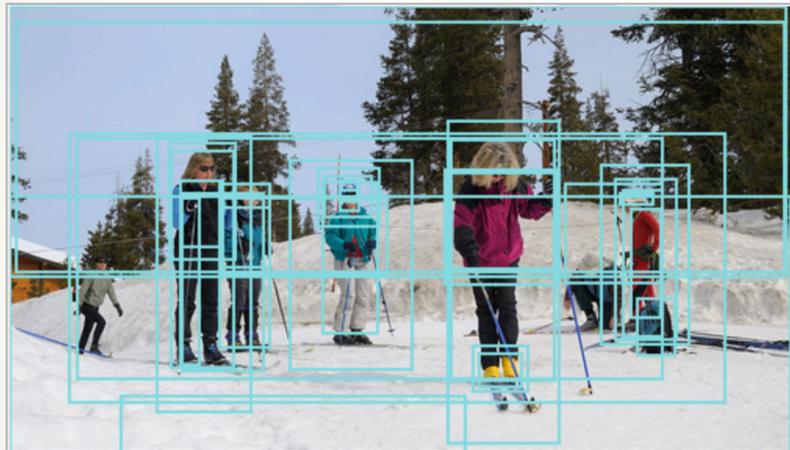
IMAGENET



lady

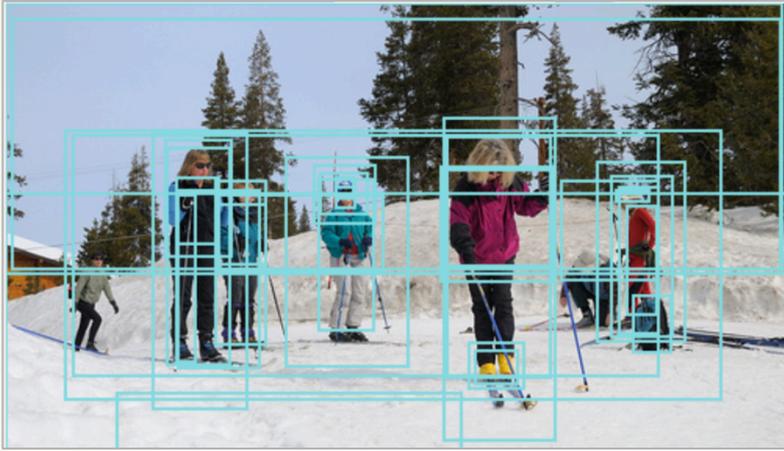


Microsoft COCO
Common Objects in Context



leaves sky snow building
jacket ski bag sunglasses vest
hat tree lady head pole
equipment boots glove coat ...

“A lady in pink dress is skiing.”



leaves sky snow building
jacket ski bag sunglasses vest
hat tree lady head pole
equipment boots glove coat ...

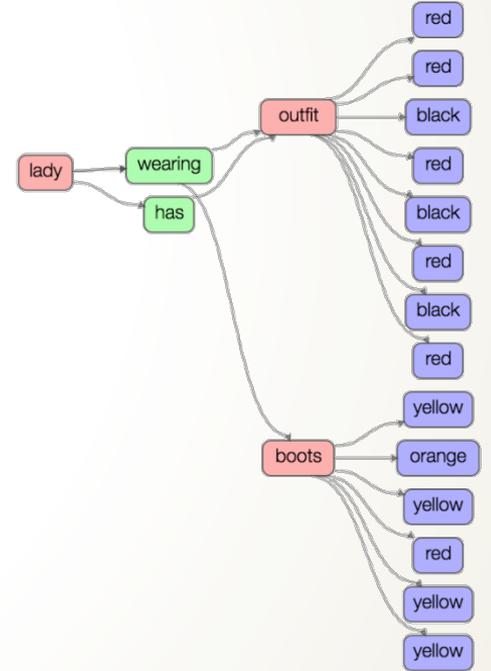
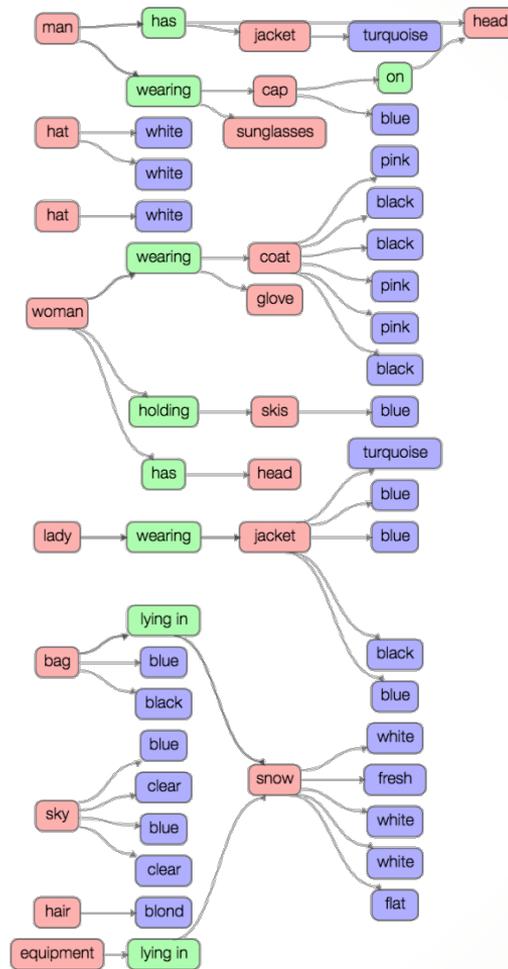
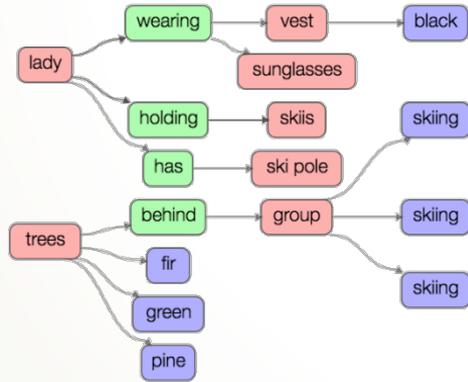
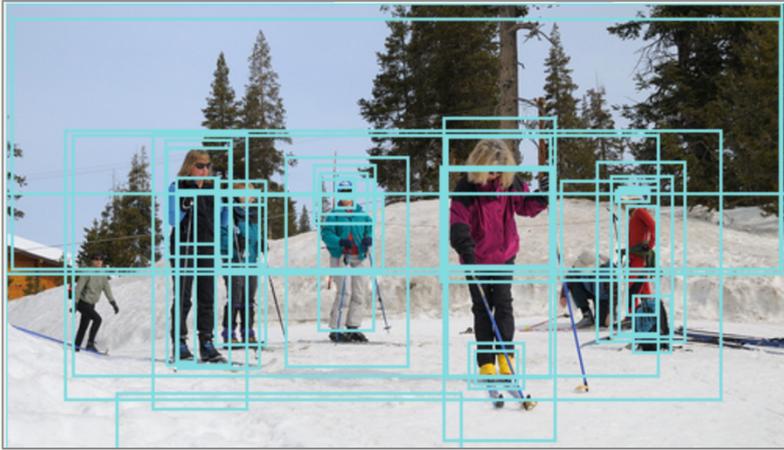
“A lady in pink dress is skiing.”

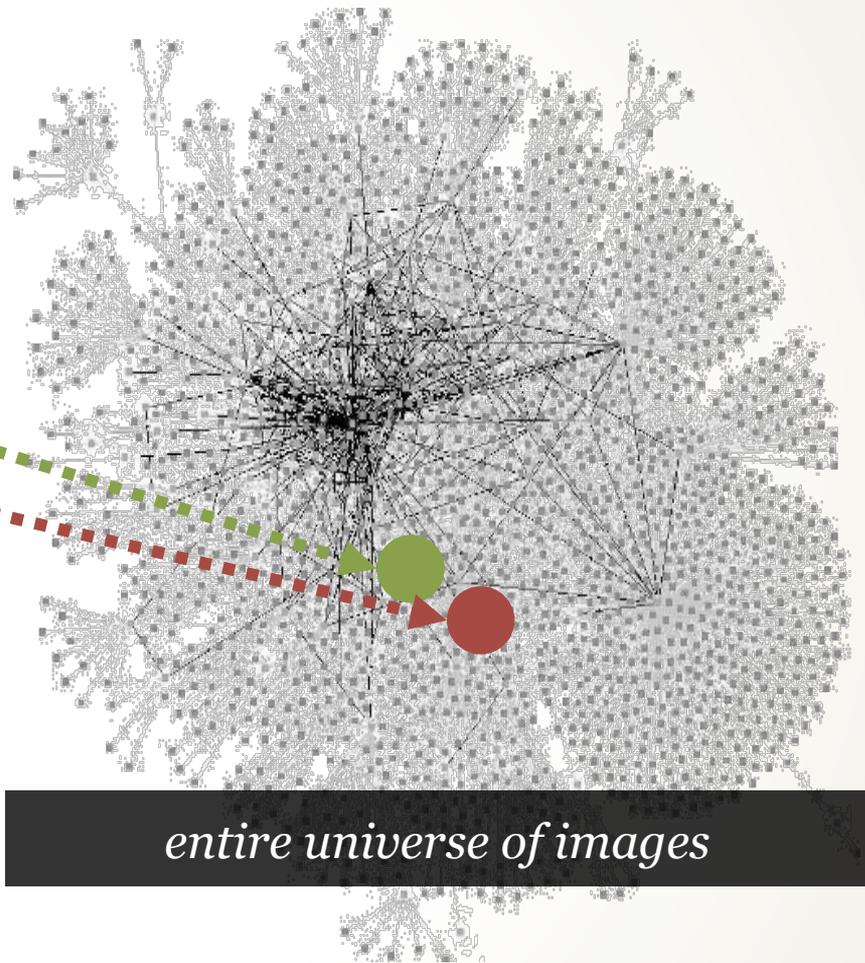
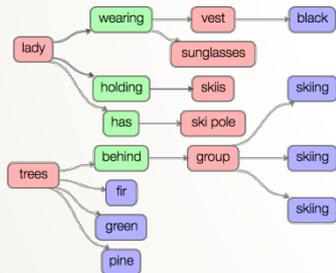
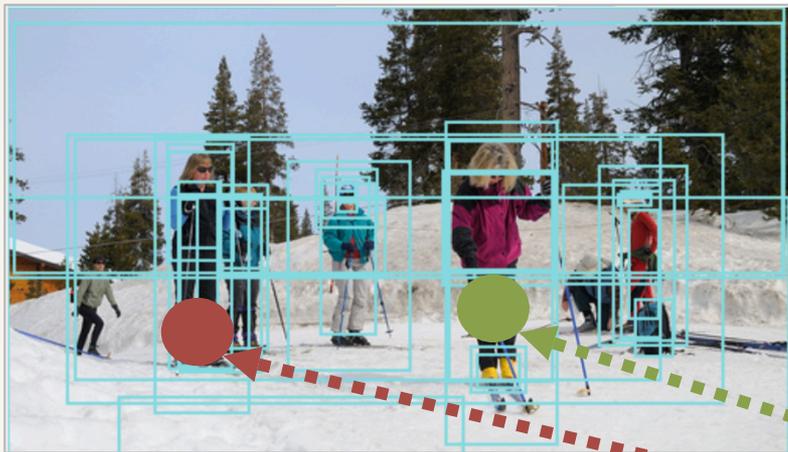
“A man standing.” “A clear blue sky at a ski resort.” “A snowy hill is in front of pine trees.”
“There are several pine trees.” “A group of people getting ready to ski.”

Q: What is the man in the center doing? **A:** *Standing on a ski.*

Q: What is the color of the sky? **A:** *Blue* **Q:** Where are the pine trees? **A:** *Behind the hill.*

<woman wear coat> <trees be green> <trees behind group (of people)>
<man has jacket> <boots be yellow> <lady hold skis> ...





entire universe of images



Visual Genome Dataset

A dataset, a knowledge base, an ongoing effort to connect structural image concepts to language.

Specs

- 108,249 images (COCO images)
- 4.2M image descriptions
- 1.8M Visual QA (7W)
- 1.4M objects, 75.7K obj. classes
- 1.5M relationships, 40.5K rel. classes
- 1.7M attributes, 40.5K attr. classes
- Vision and language correspondences
- Everything mapped to WordNet Synset

Goals

- Beyond nouns
 - Objects, verbs, attributes
- Beyond object classification
 - Relationships and contexts
- Sentences and QAs
- From Perception to Cognition

Visual Genome Dataset

A dataset, a knowledge base, an ongoing effort to connect structural image concepts to language.

DenseCap & Paragraph Generation

Karpathy et al. CVPR'16
Krause et al. CVPR'17

Relationship Prediction

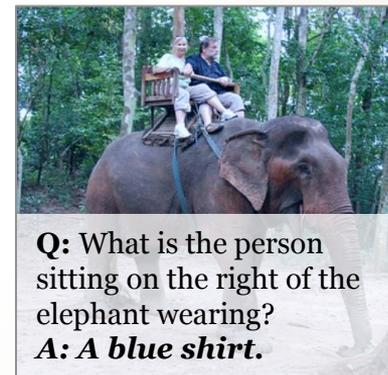
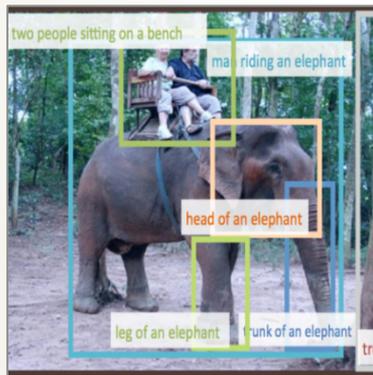
Krishna et al. ECCV'16

Image Retrieval w/ Scene Graphs

Johnson et al. CVPR'15
Xu et al. CVPR'17

Visual Q&A

Zhu et al. CVPR'16



Visual Genome Dataset

A dataset, a knowledge base, an ongoing effort to connect structural image concepts to language.

Workshop on Visual Understanding by Learning from Web Data 2017

*26 July 2017 | Honolulu, Hawaii
in conjunction with CVPR 2017*

<http://www.vision.ee.ethz.ch/webvision/workshop.html>



Q: What is the person sitting on the right of the elephant wearing?
A: A blue shirt.

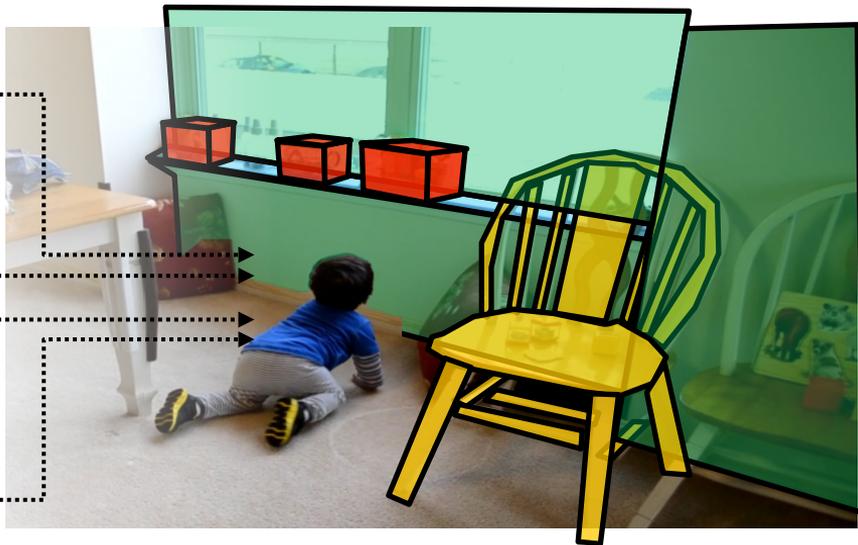
The Future of Vision and Intelligence

Vision

Language

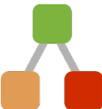
Understanding

Action



Agency:
The integration
of perception,
understanding
and action

Eight Years of Competitions

IMGENET

2010-2017

10×

reduction of image
classification error

3×

improvement of
detection precision

What Happens Now?

IM  GENET + kaggle™

We're passing the baton to **Kaggle**: a community of more than 1M data scientists.

Why?
democratizing data is vital to
democratizing AI.

image-net.org remains live at Stanford.

What Happens Now?

IM  GENET + kaggle™

ImageNet **Object Localization** Challenge

ImageNet **Object Detection** Challenge

ImageNet **Object Detection from Video** Challenge

IMAGENET Contributors/Friends/Advisors

Alex Berg
Michael Bernstein
Edward Chang
Brendan Collins
Jia Deng
Minh Do
Wei Dong
Alexei Efros
Mark Everingham
Christiane Fellbaum
Adam Finkelstein
Thomas Funkhouser
Timnit Gebru

Derek Hoiem
Zhiheng Huang
Andrej Karpathy
Aditya Khosla
Jonathan Krause
Fei-Fei Li
Kai Li
Li-Jia Li
Wei Liu
Sean Ma
Xiaojuan Ma
Jitendra Malik
Dan Osherson

Eunbyung Park
Chuck Rosenberg
Olga Russakovsky
Sanjeev Satheesh
Richard Socher
Hao Su
Zhe Wang
Andrew Zisserman

49k Amazon Mechanical Turk Workers



“This is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning.”

WINSTON CHURCHILL