Where have we been? Where are we going?

Li Fei-Fei & Jia Deng
The Beginning: CVPR 2009

The Impact of IMAGENET
ImageNet 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

VizSense
Simplifying the Visual Web

clarifai
Toward Data-Driven Medicine

Lunit

MetaMind

DNNresearch

VUNO
A Revolution in Deep Learning

Why Deep Learning is Suddenly Changing Your Life

By Roger Parloff, Sept, 2016
“The \textbf{IMAGENET} of $x$”

\textbf{SpaceNet} \\
DigitalGlobe, CosmiQ Works, NVIDIA

\textbf{MusicNet} \\
J. Thickstun et al, 2017

\textbf{Medical ImageNet} \\
Stanford Radiology, 2017

\textbf{ShapeNet} \\
A.Chang et al, 2015

\textbf{EventNet} \\
G. Ye et al, 2015

\textbf{ActivityNet} \\
F. Heilbron et al, 2015
An Explosion of Datasets

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 IMAGENET
Hardly the First Image Dataset

Segmentation (2001)

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

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

ESP (2006)
Ahn et al. 2006

LabelMe (2005)
Russell et al. 2005

3D Textures (2005)
S. Lazebnik, C. Schmid, J. Ponce

CuRRRET Textures (1999)
K. Dana B. Van Ginneken S. Nayar J. Koenderink

UIUC Cars (2004)
S. Agarwal, A. Awan, D. Roth

CalTech 101/256 (2005)
Fei-Fei et al. 2004

KTH human action (2004)
I. Leptev & B. Caputo

Middlebury Stereo (2002)
D. Scharstein R. Szeliski

3D Textures (2005)
S. Lazebnik, C. Schmid, J. Ponce

CuRRRET Textures (1999)
K. Dana B. Van Ginneken S. Nayar J. Koenderink

CMU/VASC Faces (1998)
H. Rowley, S. Baluja, T. Kanade

COIL Objects (1996)
S. Nene, S. Nayar, H. Murase

KTH human action (2004)
I. Leptev & B. Caputo

Sign Language (2008)
P. Buehler, M. Everingham, A. Zisserman

Middlebury Stereo (2002)
D. Scharstein R. Szeliski

CalTech 101/256 (2005)
Fei-Fei et al. 2004

Griffin et al. 2007

CAVIAR Tracking (2005)
R. Fisher, J. Santos-Victor J. Crowley

Segmentation (2001)

Middlebury Stereo (2002)
D. Scharstein R. Szeliski

CalTech 101/256 (2005)
Fei-Fei et al. 2004

Griffin et al. 2007

KTH human action (2004)
I. Leptev & B. Caputo

Segmentation (2001)

Middlebury Stereo (2002)
D. Scharstein R. Szeliski

CalTech 101/256 (2005)
Fei-Fei et al. 2004

Griffin et al. 2007

KTH human action (2004)
I. Leptev & B. Caputo

Segmentation (2001)

Middlebury Stereo (2002)
D. Scharstein R. Szeliski

CalTech 101/256 (2005)
Fei-Fei et al. 2004

Griffin et al. 2007

KTH human action (2004)
I. Leptev & B. Caputo

Segmentation (2001)

Middlebury Stereo (2002)
D. Scharstein R. Szeliski

CalTech 101/256 (2005)
Fei-Fei et al. 2004

Griffin et al. 2007

KTH human action (2004)
I. Leptev & B. Caputo

Segmentation (2001)

Middlebury Stereo (2002)
D. Scharstein R. Szeliski

CalTech 101/256 (2005)
Fei-Fei et al. 2004

Griffin et al. 2007

KTH human action (2004)
I. Leptev & B. Caputo

Segmentation (2001)

Middlebury Stereo (2002)
D. Scharstein R. Szeliski

CalTech 101/256 (2005)
Fei-Fei et al. 2004

Griffin et al. 2007

KTH human action (2004)
I. Leptev & B. Caputo

Segmentation (2001)
A Profound Machine Learning Problem Within Visual Learning
Machine Learning 101: Complexity, Generalization, Overfitting

![Graph showing the relationship between error and capacity, with zones for underfitting and overfitting, and the optimal capacity where generalization error is minimized.](image-url)
One-Shot Learning

Fei-Fei et al, 2003, 2004
One-shot learning algorithm: Bayesian Variational Inference

Model:
Parameter update

Data: model fitting

prior statistics of $p(\theta)$

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

Fei-Fei et al, 2003, 2004
How Children Learn to See
Underfitting Zone

Overfitting Zone

Training Error

Optimal Capacity

Generalization Error

Capacity

Generalization Gap
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

Global Data Traffic (PB/month)

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.
IMAGENET Comrades

Prof. Kai Li
Princeton

Jia Deng
1st Ph.D. student
Princeton
Step 1: Ontological structure based on WordNet
Step 2: Populate categories with thousands of images from the Internet
Step 3: Clean results by hand
Three Attempts at Launching IMAGENET
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 \sim 10^{2-3})

\[ 40,000 \times 10,000 \times 3 / 2 = 6000,000,000,000 \text{ sec} \]

\[ \approx 19 \text{ years} \]
2nd Attempt:
Human-in-the-Loop Solutions

Towards scalable dataset construction:
An active learning approach

Brendan Collins, Jia Deng, Kui Cai
{macollin, dengjia, li, feifei}@princeton.edu
Department of Computer Science, Princeton University

Abstract. As computer vision research continues to advance, the need for
and greater variation within object categories becomes more evident. Creating
more exhaustive datasets is laborious and monotonous, especially when
in which many images have been automatically labeled or when the relevant
category (typically by automatic internet searches) cannot be turned
out to be noise. We present a dataset construction framework which
employs active, online learning to select images for minimal user input.
This approach achieves scalability by employing the principle of
a well-structured 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.
2\textsuperscript{nd} 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

49k Workers from 167 Countries
2007-2010
The Result: IMAGENET Goes Live in 2009
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

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

Carnivore
- Canine
- Dog
- Working Dog
- Husky

Free of Charge
To ensure immediate application and a sense of community
An Emphasis on Community and Achievement

IMAGENET

Large Scale Visual Recognition Challenge (ILSVRC 2010-2017)
Our Inspiration: PASCAL VOC

2005-2012
Our Inspiration: PASCAL VOC

Mark Everingham
1973-2012

Mark Everingham Prize @ ECCV 2016

IMAGENET

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

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of Entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>35</td>
</tr>
<tr>
<td>2011</td>
<td>29</td>
</tr>
<tr>
<td>2012</td>
<td>81</td>
</tr>
<tr>
<td>2013</td>
<td>123</td>
</tr>
<tr>
<td>2014</td>
<td>157</td>
</tr>
<tr>
<td>2015</td>
<td>172</td>
</tr>
</tbody>
</table>

*Number of Entries*
Participation and Performance

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of Entries</th>
<th>Classification Errors (top-5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>35</td>
<td>0.28</td>
</tr>
<tr>
<td>2011</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>123</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>157</td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>172</td>
<td>0.03</td>
</tr>
</tbody>
</table>

- Number of Entries
- Classification Errors (top-5)
Participation and Performance

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of Entries</th>
<th>Classification Errors (top-5)</th>
<th>Average Precision For Object Detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>35</td>
<td>-</td>
<td>0.28</td>
</tr>
<tr>
<td>2011</td>
<td>29</td>
<td>0.03</td>
<td>-</td>
</tr>
<tr>
<td>2012</td>
<td>29</td>
<td>0.23</td>
<td>-</td>
</tr>
<tr>
<td>2013</td>
<td>123</td>
<td>0.23</td>
<td>-</td>
</tr>
<tr>
<td>2014</td>
<td>157</td>
<td>0.66</td>
<td>-</td>
</tr>
<tr>
<td>2015</td>
<td>172</td>
<td>0.03</td>
<td>-</td>
</tr>
<tr>
<td>2016</td>
<td>172</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
What we did to make IMAGENET better
Lack of Details
Lack of Details...ILSVRC Detection Challenge

Statistics

<table>
<thead>
<tr>
<th></th>
<th>PASCAL VOC 2012</th>
<th>ILSVRC 2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object classes</td>
<td>20</td>
<td>200</td>
</tr>
<tr>
<td>Images</td>
<td>5.7K</td>
<td>395K</td>
</tr>
<tr>
<td>Objects</td>
<td>13.6K</td>
<td>345K</td>
</tr>
</tbody>
</table>

- Training Images increased by 70x from 5.7K to 395K.
- Objects increased by 25x from 13.6K to 345K.
- Object classes increased by 10x from 20 to 200.
Evaluation of ILSVRC Detection

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

<table>
<thead>
<tr>
<th>Table</th>
<th>Chair</th>
<th>Horse</th>
<th>Dog</th>
<th>Cat</th>
<th>Bird</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

# images: 400K
# classes: 200
# annotations = 80M!
Evaluation of ILSVRC Detection

Hierarchical annotation

Object Presence

Image

Is there an animal?

Is there a mammal?

Is there a cat?

J. Deng, O. Russakovsky, J. Krause, M. Bernstein, A. Berg, & L. Fei-Fei. CHI, 2014
What does classifying 10K+ classes tell us?
Fine-Grained Recognition

“Cardigan Welsh Corgi”

“Pembroke Welsh Corgi”
Fine-Grained Recognition

[Gebru, 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!

Krizhevsky, Sutskever & Hinton, NIPS 2012

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 ILSVRC-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%.
...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]
A Deep Learning Revolution

Neural Nets

IMAGENET

GPUs

A Deep Learning Revolution
Ontological Structure Structure
Not Used as Much
Wombat

Deng, Krause, Berg & Fei-Fei, CVPR 2012
Maximize Specificity ($f$) Subject to Accuracy ($f$) $\geq 1 - \varepsilon$
Optimizing with a Knowledge Ontology Results in Big Gains in Information at Arbitrary Accuracy

Deng, Krause, Berg & Fei-Fei, CVPR 2012
Relatively Few Works Have Used Ontology

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

ECCV 2012 Best paper Award
Most works still use 1M images to do pre-training.

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
How Humans Compare

How Humans Compare

<table>
<thead>
<tr>
<th>Human</th>
<th>GoogLeNet</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>5.1%</strong> Top-5 error rate</td>
<td><strong>6.8%</strong> Top-5 error rate</td>
</tr>
</tbody>
</table>

**Susceptible to:**
- Fine-grained recognition
- Class unawareness
- Insufficient training data

**Susceptible to:**
- Small, thin objects
- Image filters
- Abstract representations
- Miscellaneous sources

What Lies Ahead
Moving from object recognition...
...to human-level understanding.

Stepping on a scale adds weight and ups the reading.
Inverse Graphics

Image credit: https://www.youtube.com/watch?v=ip-KIzQmcBo (Oliver Villar)
lady
“A lady in pink dress is skiing.”
“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

[Johnson et al., CVPR 2015]
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

Krishna et al. IJCV 2016
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

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

Krishna et al. IJCV 2016
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

2010-2017

10× reduction of image classification error

3× improvement of detection precision
What Happens Now?

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?

ImageNet + Kaggle

ImageNet Object Localization Challenge
ImageNet Object Detection Challenge
ImageNet Object Detection from Video Challenge
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