CORNELL TECH

Deep Learning Clinic

(DLC)

Lecture 6 - Data in DL

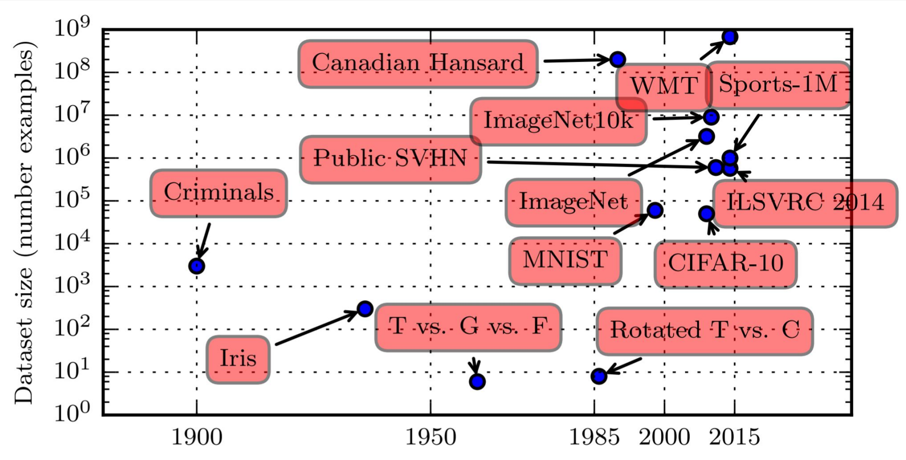
Jin Sun

11/2/2018

Today

- **Overview**
- Existing Dataset
- Build A Dataset
    - Data Collection
    - Annotation
    - Verification
    - Tools
- Amazon MTurk Tutorial

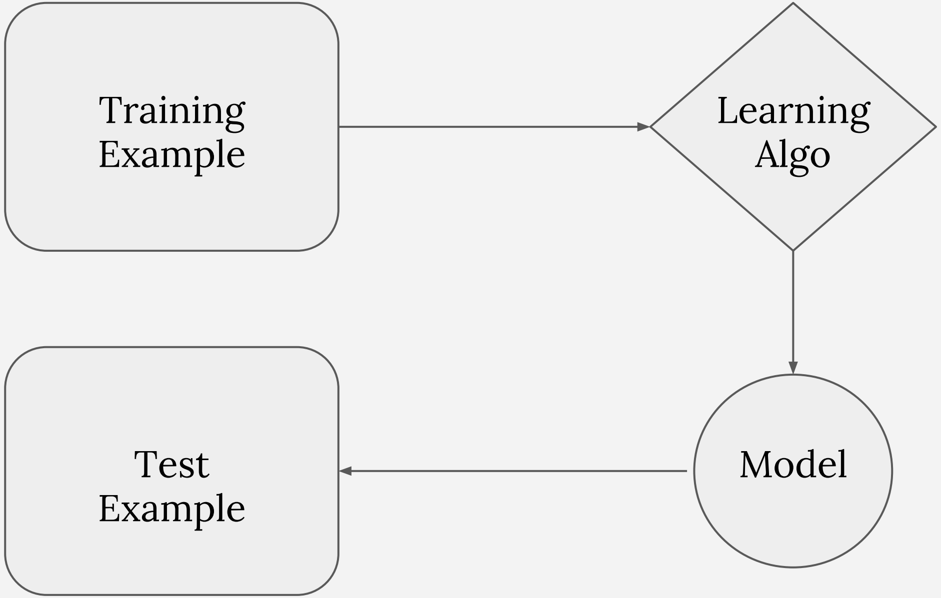
Main Reasons Behind Deep Learning's Success



Data

Hardware

Data in Ml/Dl Pipeline



$$e \doteq \mathop{\mathbb{E}}_{(x,y)\ \sim D}[L(y, f(x))]$$

$$\doteq \frac{1}{N}\sum_{n=1}^{N} L(y_n, f(x_n))$$

Importance of Dataset

Benchmark

IMAGENET **Large Scale Visual Recognition Challenge 2017 (ILSVRC2017)**

DET LOC VID **Team information**

Legend:
Yellow background = winner in this task according to this metric; authors are willing to reveal the method
White background = authors are willing to reveal the method
Grey background = authors chose not to reveal the method
Italics = authors requested entry not participate in competition

Object detection (DET)[top]

Task 1a: Object detection with provided training data

Ordered by number of categories won

Team name	Entry description	Number of object categories won	mean AP
BDAT	submission4	85	0.731392
BDAT	submission3	65	0.732227
BDAT	submission2	30	0.723712
DeepView(ETRI)	Ensemble_A	10	0.593084
NUS-Qihoo_DPNs (DET)	Ensemble of DPN models	9	0.656932
KAISTNIA_ETRI	Ensemble Model5	1	0.61022
KAISTNIA_ETRI	Ensemble Model4	0	0.609402
KAISTNIA_ETRI	Ensemble Model2	0	0.608299
KAISTNIA_ETRI	Ensemble Model1	0	0.608278
KAISTNIA_ETRI	Ensemble Model3	0	0.60631
DeepView(ETRI)	Single model A using ResNet for detection	0	0.587519

Russian–English

#	score	range	system
1	0.583	1	AFRL-PE
2	0.299	2	ONLINE-B
3	0.190	3-5	ONLINE-A
	0.178	3-5	PROMT-HYBRID
	0.123	4-7	PROMT-RULE
	0.104	5-8	UEDIN-PHRASE
	0.069	5-8	Y-SDA
	0.066	5-8	ONLINE-G
4	-0.017	9	AFRL
5	-0.159	10	UEDIN-SYNTAX
6	-0.306	11	KAZNU
7	-0.487	12	RBMT1
8	-0.642	13	RBMT4

English–Russian

#	score	range	system
1	0.575	1-2	PROMT-RULE
	0.547	1-2	ONLINE-B
2	0.426	3	PROMT-HYBRID
3	0.305	4-5	UEDIN-UNCNSTR
	0.231	4-5	ONLINE-G
4	0.089	6-7	ONLINE-A
	0.031	6-7	UEDIN-PHRASE
5	-0.920	8	RBMT4
6	-1.284	9	RBMT1

French–English

#	score	range	system
1	0.608	1	UEDIN-PHRASE
2	0.479	2-4	KIT
	0.475	2-4	ONLINE-B
	0.428	2-4	STANFORD
3	0.331	5	ONLINE-A
4	-0.389	6	RBMT1
5	-0.648	7	RBMT4
6	-1.284	8	ONLINE-C

English–French

#	score	range	system
1	0.327	1	ONLINE-B
2	0.232	2-4	UEDIN-PHRASE
	0.194	2-5	KIT
	0.185	2-5	MATRAN
	0.142	4-6	MATRAN-RULES
	0.120	4-6	ONLINE-A
3	0.003	7-9	UU-DOCENT
	-0.019	7-10	PROMT-HYBRID
	-0.033	7-10	UA
	-0.069	8-10	PROMT-RULE
4	-0.215	11	RBMT1
5	-0.328	12	RBMT4
6	-0.540	13	ONLINE-C

Hindi–English

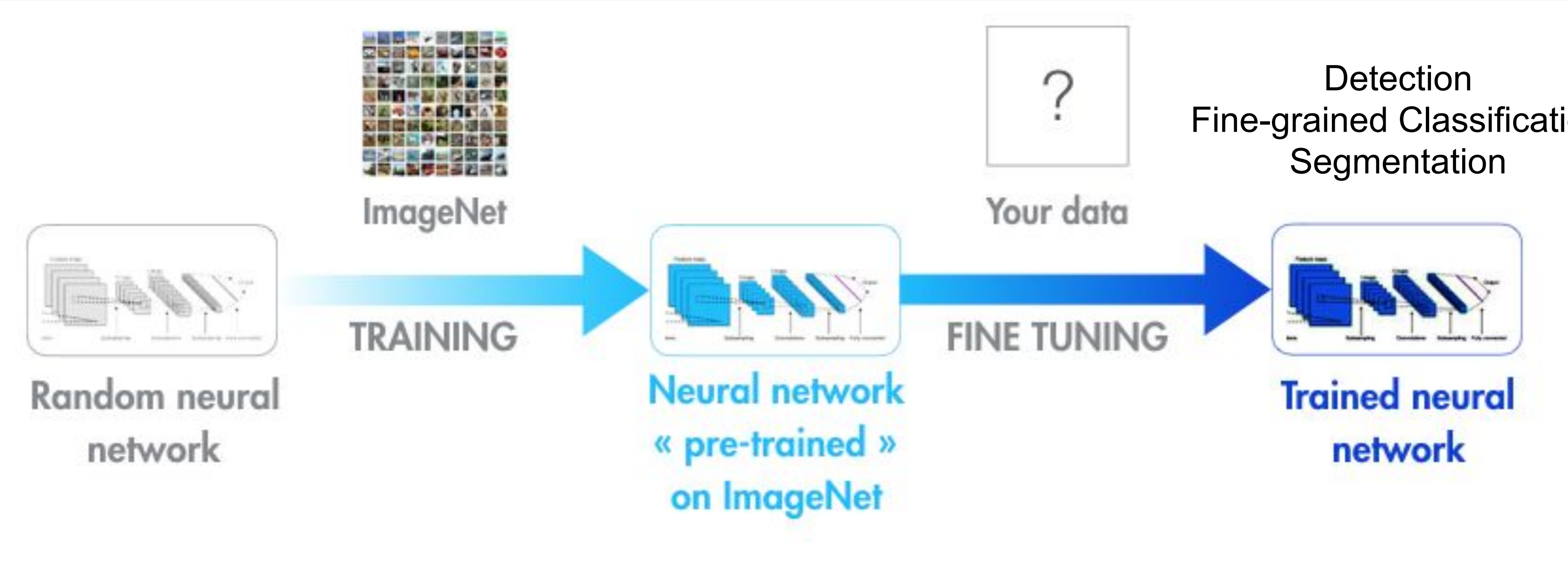
#	score	range	system
1	1.326	1	ONLINE-B
2	0.559	2-3	ONLINE-A
	0.476	2-4	UEDIN-SYNTAX
	0.434	3-4	CMU
3	0.323	5	UEDIN-PHRASE
4	-0.198	6-7	AFRL
	-0.280	6-7	IIT-BOMBAY
5	-0.549	8	DCU-LINGO24
6	-2.092	9	IIIT-HYDERABAD

English–Hindi

#	score	range	system
1	1.008	1	ONLINE-B
2	0.915	2	ONLINE-A
3	0.214	3	UEDIN-UNCNSTR
4	0.120	4-5	UEDIN-PHRASE
	0.054	4-5	CU-MOSES
5	-0.111	6-7	IIT-BOMBAY
	-0.142	6-7	IPN-UPV-CNTXT
6	-0.233	8-9	DCU-LINGO24
	-0.261	8-9	IPN-UPV-NODEV
7	-0.449	10-11	MANAWI-H1
	-0.494	10-11	MANAWI
8	-0.622	12	MANAWI-RMOOV

Importance of Dataset

General Purpose Prior



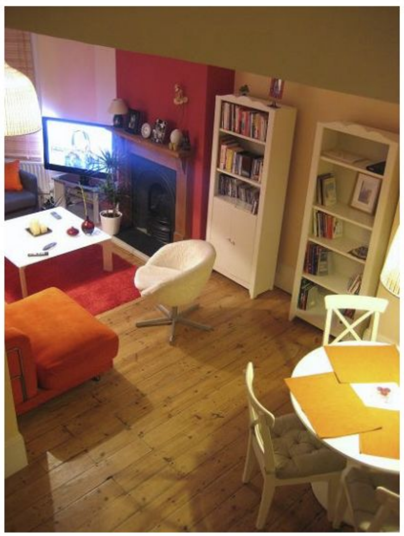
https://medium.com/owkin/transfer-learning-and-the-rise-of-collaborative-artificial-intelligence-41f9e2950657

Importance of Dataset

New Algorithms and Research Problems

Question : How many shelves?

Original Image | 8



Visual Q&A

Complementary Image | 11

Importance of Dataset

New Algorithms and Research Problems

Recommendation Systems

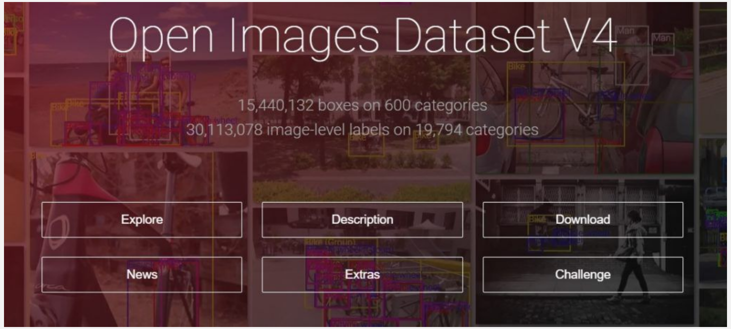
Music, books, videos
Online shopping
Financial
Online Dating
...

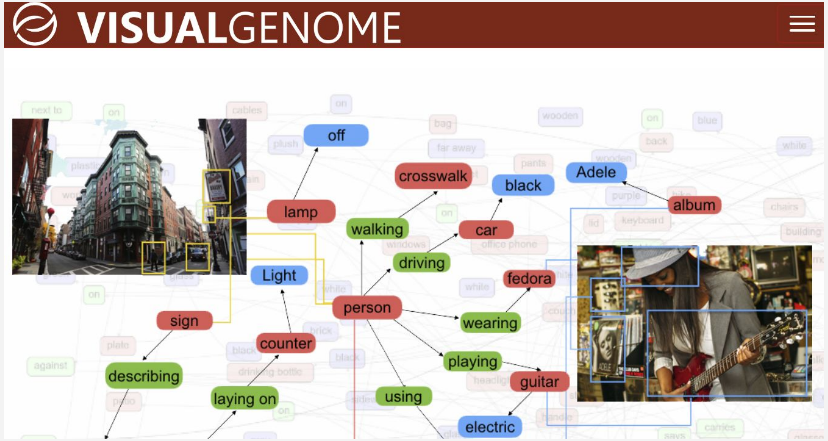
Today

- Overview
- **Existing Dataset**
- Build A Dataset
  - Data Collection
  - Annotation
  - Verification
  - Tools
- Amazon MTurk Tutorial

Existing Dataset - Vision

COCO
Common Objects in Context





Existing Dataset - Natural Language

IMDB Reviews, Sentiment140 (Sentiment Analysis)

1 Billion Word Language Model Benchmark (Language Modeling)

WordNet (Database for English ‘synsets’)

Google Books Ngrams

Existing Dataset - Others

HealthData.gov (Health Care)

OASIS brain images

Data.gov (agriculture, climate, ecosystems, public safety...)

Kaggle Dataset

Today

- Overview
- Existing Dataset
- **Build A Dataset**
  - Data Collection
  - Annotation
  - Verification
  - Tools
- Amazon MTurk Tutorial

Why Build Your Own Dataset

Variation

Existing datasets do not contain enough variety.

E.g., non-traditional lighting and poses.

Annotation

Existing datasets do not provide the required information.

E.g., no lighting condition information in ImageNet.

Case Study:
The Construction of COCO Dataset

COCO Dataset Statistics

What is COCO?

COCO is a large-scale object detection, segmentation, and captioning dataset. COCO has several features:

- Object segmentation
- Recognition in context
- Superpixel stuff segmentation
- 330K images (>200K labeled)
- 1.5 million object instances
- 80 object categories
- 91 stuff categories
- 5 captions per image
- 250,000 people with keypoints

Collaborators

Tsung-Yi Lin Google Brain

Genevieve Patterson MSR, Trash TV

Matteo R. Ronchi Caltech

Yin Cui Cornell Tech

Michael Maire TTI-Chicago

Serge Belongie Cornell Tech

Lubomir Bourdev WaveOne, Inc.

Ross Girshick FAIR

James Hays Georgia Tech

Pietro Perona Caltech

Deva Ramanan CMU

Larry Zitnick FAIR

Piotr Dollár FAIR

Sponsors

CVDF

Microsoft

facebook

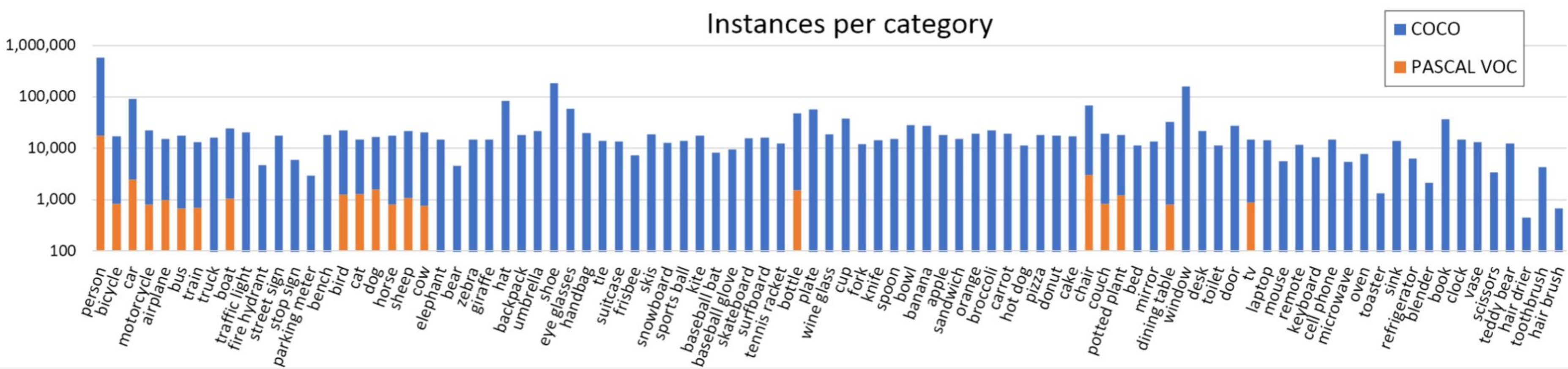
Mighty Ai

Data Collection

Identify Object Categories

PASCAL VOC + frequently used words for objects + survey on 4-8 years old children = 272 candidates

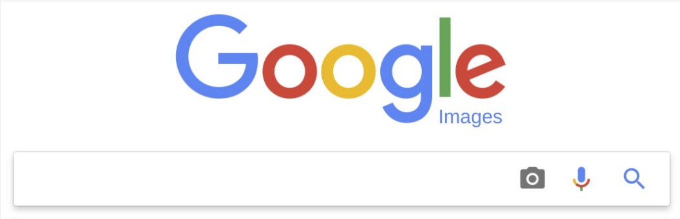
Voting to get final categories: 91.



Data Collection

Collect Images For Each Object Category

Iconic Images

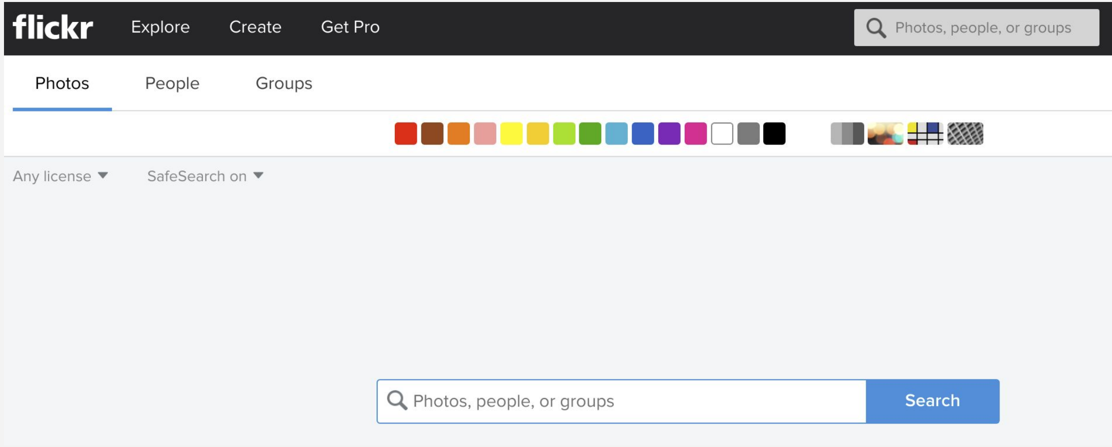


Data Collection

Collect Images For Each Object Category

328,000 images in total.

Non-Iconic Images



Data Annotation

How to label over 2.5 million object instances in 300K+ images?

Crowdsourcing.

Data Annotation

How to label over 2.5 million object instances in 300K+ images?

Crowdsourcing.

(a) Category labeling

8 Workers Per Image

~20k Worker Hours

Data Annotation

How to label over 2.5 million object instances in 300K+ images?

Crowdsourcing.

8 Workers Per Image

~10k Worker Hours

(b) Instance spotting

Data Annotation

How to label over 2.5 million object instances in 300K+ images?

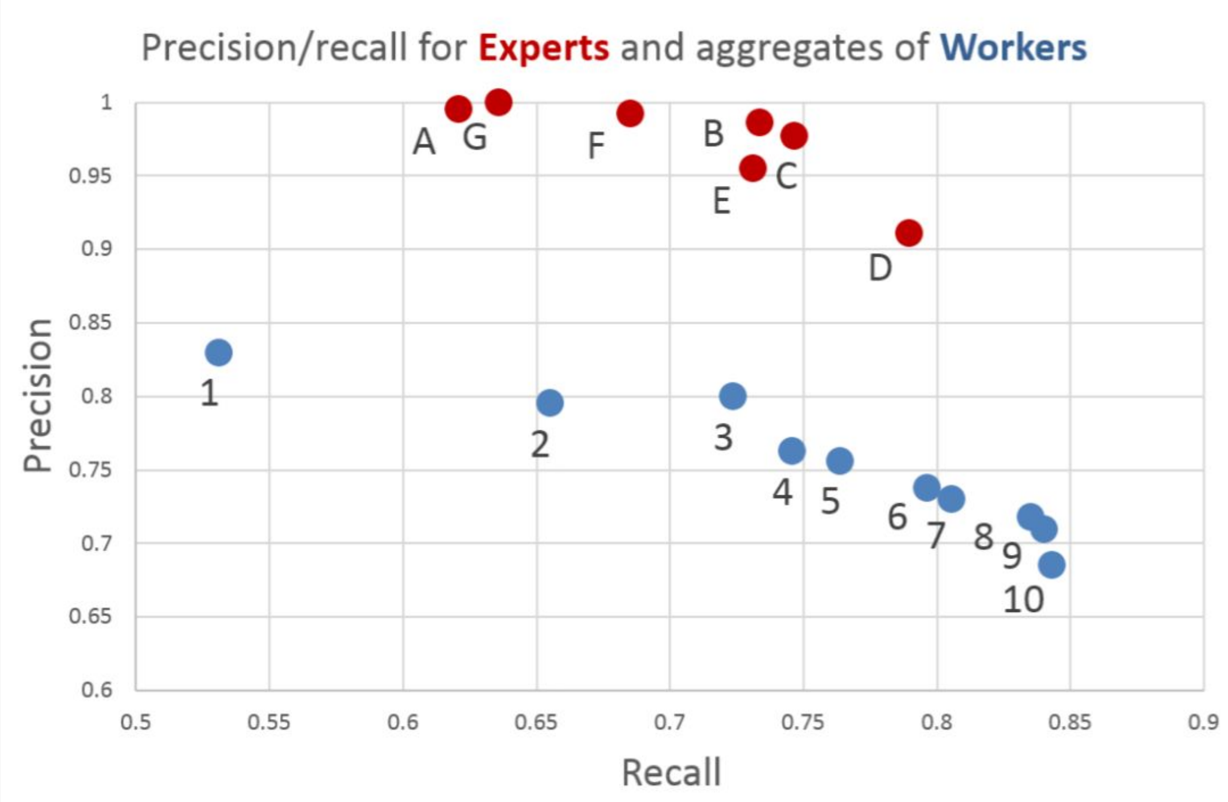
Crowdsourcing.

An expensive task. Only 1 worker per image.

Training stage enforced.

(c) Instance segmentation

Data Verification



Tools

Data Source

Google/Bing Search, Flickr, Instagram, Google Map/Streetview, Satellite

Visual Annotation

VGG Image Annotator, Video Annotation Tool, Build your own (HTML+JS)

Crowdsourcing

Amazon MTurk

Today

- Overview
- Existing Dataset
- Build A Dataset
    - Data Collection
    - Annotation
    - Verification
    - Tools
- **Amazon MTurk Tutorial**

Amazon Mechanical Turk Tutorial

amazon mechanical turk beta

https://www.mturk.com

On Demand

Over 500K workers, 24x7

Speed

Work is done in parallel

Scalable

No minimum project size

Qualification

Set prerequisite to workers

MTurk Concepts

Requesters

Person creates tasks for Workers to work on.

Human Intelligence Tasks (HITs)

HIT is a single, self-contained task.

Assignment

Multiple Workers can be assigned to a single HIT.

A Worker can only accept a HIT once and submit one assignment per HIT.

Workers

Person completes assignments.

Approval and Payment

After assignment submission, if you approve the work, the HIT reward is draw from your MTurk account.

Qualification

Anyone can register as a worker. You can set qualification types such as approval rate to control the quality of submissions.

Common Use Cases

Image/Video Processing

MTurk is well-suited for processing images. While difficult for computers, it is a task that is extremely easy for people to do. In the past, companies have used MTurk to:

Tag objects found in an image to improve your search or advertising targeting

Review a set of images to select the best picture to represent a product

Audit user-uploaded images or videos to moderate content

Classify objects found in satellite imagery

Data Verification and Clean-up

Companies with large online directories or catalogs are using MTurk to identify duplicate entries and verify item details. Examples of this have included:

Removing duplicate content from business listings

Identifying incomplete or duplicate product listings in a catalog

Verifying restaurant details such as phone numbers or hours of operation

Converting unstructured data about locations into well-formed addresses

Common Use Cases

Information Gathering

The diversification and the scale of the MTurk workforce allows you to gather a breadth of information that would be almost impossible to do otherwise such as:

Allowing people to ask questions from a computer or mobile device about any topic and have Workers return the results

Filling out market research or survey data on a variety of topics

Writing content for websites

Finding specific fields or data elements in large legal and government documents

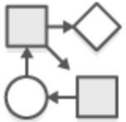
Data Processing

Companies take advantage of the power of the MTurk workforce to understand and intelligently respond to different types of data including:

Audio editing and transcription

Human powered translation services

Rating the accuracy of results for a search engine



Categorizing information to match a given schema or taxonomy

Example: Data Labeling Using MTurk

Tutorial 1
Tutorial 2

1. Setup

Python and Boto3 (AWS SDK).

2. Accounts

AWS and MTurk (Also need to link the two).

Purchasing Prepaid HITs.

3. Creating Tasks

Define a HIT and its reward.

4. Retrieving Results

Verify result, Add a bonus

Summary

- Overview
- Existing Dataset
- Build A Dataset
  - Data Collection
  - Annotation
  - Verification
  - Tools
- Amazon MTurk Tutorial