Deep Learning for Visual Recognition: Tips for Implementation

Alvaro Soto

Computer Science Department (DCC), PUC

Small Letter in Deep Learning

Small Letter in Deep Learning

- Need special hardware to run: GPUs or TPUs

Small Letter in Deep Learning

- Need special hardware to run: GPUs or TPUs
- Need huge sources of data: Labeled data

GPUs

NVIDIA GeForce 1080 Ti

- 11 GB

NVIDIA GeForce 1080 Ti

- 11 GB
- Arquitectura Pascal

NVIDIA GeForce 1080 Ti

- 11 GB
- Arquitectura Pascal
- 3584 Cuda cores

NVIDIA GeForce 1080 Ti

- 11 GB
- Arquitectura Pascal
- 3584 Cuda cores
- 215 Watts

NVIDIA GeForce 1080 Ti

- 11 GB
- Arquitectura Pascal
- 3584 Cuda cores
- 215 Watts
- Price $650 (USA)

NVIDIA GeForce RTX 2080 Ti

- 11 GB

NVIDIA GeForce RTX 2080 Ti

- 11 GB
- Arquitectura Turing

NVIDIA GeForce RTX 2080 Ti

- 11 GB
- Arquitectura Turing
- 4352 Cuda cores

NVIDIA GeForce RTX 2080 Ti

- 11 GB
- Arquitectura Turing
- 4352 Cuda cores
- 250 Watts

NVIDIA GeForce RTX 2080 Ti

- 11 GB
- Arquitectura Turing
- 4352 Cuda cores
- 250 Watts
- Price $1000 (USA)

NVIDIA TITAN RTX

- 24 GB

NVIDIA TITAN RTX

- 24 GB
- Arquitectura Turing

NVIDIA TITAN RTX

- 24 GB
- Arquitectura Turing
- 4608 Cuda cores

NVIDIA TITAN RTX

- 24 GB
- Arquitectura Turing
- 4608 Cuda cores
- 280 Watts

NVIDIA TITAN RTX

- 24 GB
- Arquitectura Turing
- 4608 Cuda cores
- 280 Watts
- Price $2500 (USA)

Specifications	**TITAN RTX**	TITAN X Pascal	RTX 2080 Ti	RTX 2080	RTX 2070	GTX 1080 Ti
Chip	**TU102**	GP102	TU102	TU104	TU106	GP102
FinFET process	**12 nm**	16 nm	12 nm	12 nm	12 nm	16 nm
CUDA cores	**4,608**	3,584	4,352	2,944	2,304	3,584
Texture Units	**288**	224	272	184	144	224
Tensor Cores	**576**	-	544	368	288	-
RT Cores	**72**	-	68	46	36	-
GPU base clock	**1,350 MHz**	1,417 MHz	1,350 MHz	1,515 MHz	1,410 MHz	1,481 MHz
GPU oost clock	**1,770 MHz**	1,531 MHz	1,545 MHz	1,710 MHz	1,620 MHz	1,582 MHz
Memory Bus	**384-bit**	384-bit	352-bit	256-bit	256-bit	352-bit
Memory Bandwidth	**672 GB/s**	480.4 GB/s	616 GB/s	448 GB/s	448 GB/s	484.4 GB/s
Video memory	**24 GB GDDR6**	12 GB GDDR5X	11 GB GDDR6	8 GB GDDR6	8 GB GDDR6	11 GB GDDR5X
Power consumption	**280 W**	250 W	250 W	215 W	175 W	215 W

GPUs and Memory

At training time, DL techniques have high memory requirements

GPUs and Memory

At training time, DL techniques have high memory requirements

- Parameters: weights and biases

GPUs and Memory

At training time, DL techniques have high memory requirements

- Parameters: weights and biases
- Optimizer: gradients and momentums

GPUs and Memory

At training time, DL techniques have high memory requirements

- Parameters: weights and biases
- Optimizer: gradients and momentums
- Neurons (activations and error for backward pass)

GPUs and Memory

At training time, DL techniques have high memory requirements

- Parameters: weights and biases
- Optimizer: gradients and momentums
- Neurons (activations and error for backward pass)
- Workspace (implementation overhead, input images)

GPUs and Memory

At training time, DL techniques have high memory requirements

- Parameters: weights and biases
- Optimizer: gradients and momentums
- Neurons (activations and error for backward pass)
- Workspace (implementation overhead, input images)

GPUs and Memory

At training time, DL techniques have high memory requirements

- Parameters: weights and biases
- Optimizer: gradients and momentums
- Neurons (activations and error for backward pass)
- Workspace (implementation overhead, input images)

Ex. consider a typical scenario, a network has:

- 100M parameters (weights + biases)

GPUs and Memory

At training time, DL techniques have high memory requirements

- Parameters: weights and biases
- Optimizer: gradients and momentums
- Neurons (activations and error for backward pass)
- Workspace (implementation overhead, input images)

Ex. consider a typical scenario, a network has:

- 100M parameters (weights + biases)
- 10M activations (neurons)

GPUs and Memory

At training time, DL techniques have high memory requirements

- Parameters: weights and biases
- Optimizer: gradients and momentums
- Neurons (activations and error for backward pass)
- Workspace (implementation overhead, input images)

Ex. consider a typical scenario, a network has:

- 100M parameters (weights + biases)
- 10M activations (neurons)
- Batch size=128

GPUs and Memory

At training time, DL techniques have high memory requirements

- Parameters: weights and biases
- Optimizer: gradients and momentums
- Neurons (activations and error for backward pass)
- Workspace (implementation overhead, input images)

Ex. consider a typical scenario, a network has:

- 100M parameters (weights + biases)
- 10M activations (neurons)
- Batch size=128
- Input image=1M

GPUs and Memory

At training time, DL techniques have high memory requirements

- Parameters: weights and biases
- Optimizer: gradients and momentums
- Neurons (activations and error for backward pass)
- Workspace (implementation overhead, input images)

Ex. consider a typical scenario, a network has:

- 100M parameters (weights + biases)
- 10M activations (neurons)
- Batch size=128
- Input image=1M

GPUs and Memory

At training time, DL techniques have high memory requirements

- Parameters: weights and biases
- Optimizer: gradients and momentums
- Neurons (activations and error for backward pass)
- Workspace (implementation overhead, input images)

Ex. consider a typical scenario, a network has:

- 100M parameters (weights + biases)
- 10M activations (neurons)
- Batch size=128
- Input image=1M

Memory requirement:

- At training time (bytes):
  100M*4 + 100M*4*2 + 128*10M*4*2 + 128*1M $\approx$ 11.5Gb

GPUs and Memory

At training time, DL techniques have high memory requirements

- Parameters: weights and biases
- Optimizer: gradients and momentums
- Neurons (activations and error for backward pass)
- Workspace (implementation overhead, input images)

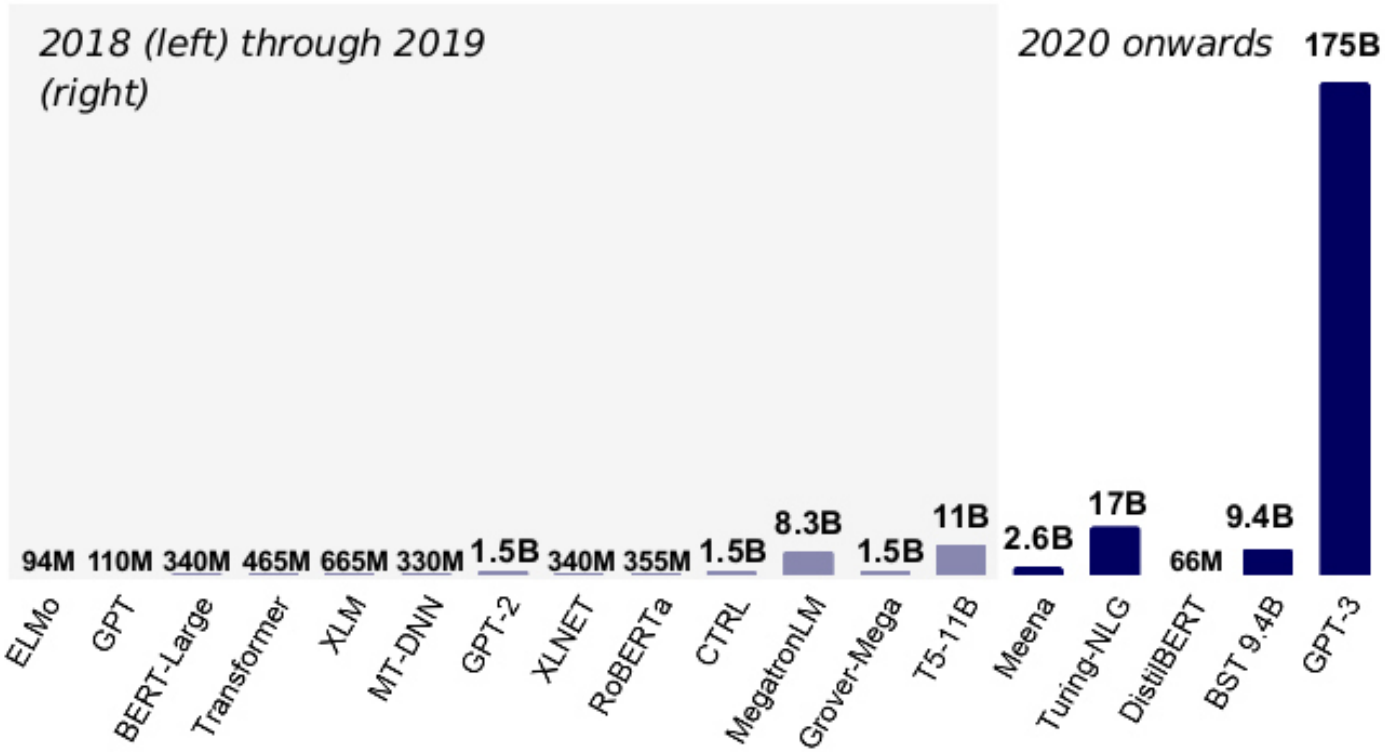
Ex. consider a typical scenario, a network has:

- 100M parameters (weights + biases)
- 10M activations (neurons)
- Batch size=128
- Input image=1M

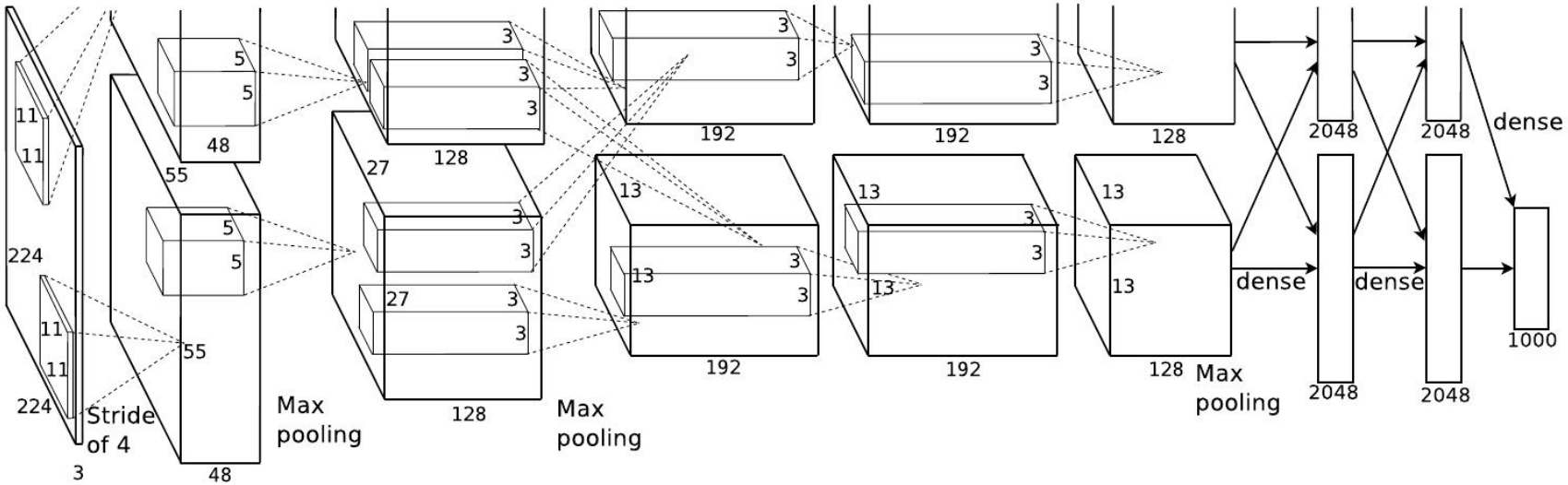
Memory requirement:

- At training time (bytes):
  100M*4 + 100M*4*2 + 128*10M*4*2 + 128*1M $\approx$ 11.5Gb
- At test time (bytes):
  100M*4 + 1M $\approx$ 101 Mb

Deep Learning Madness!

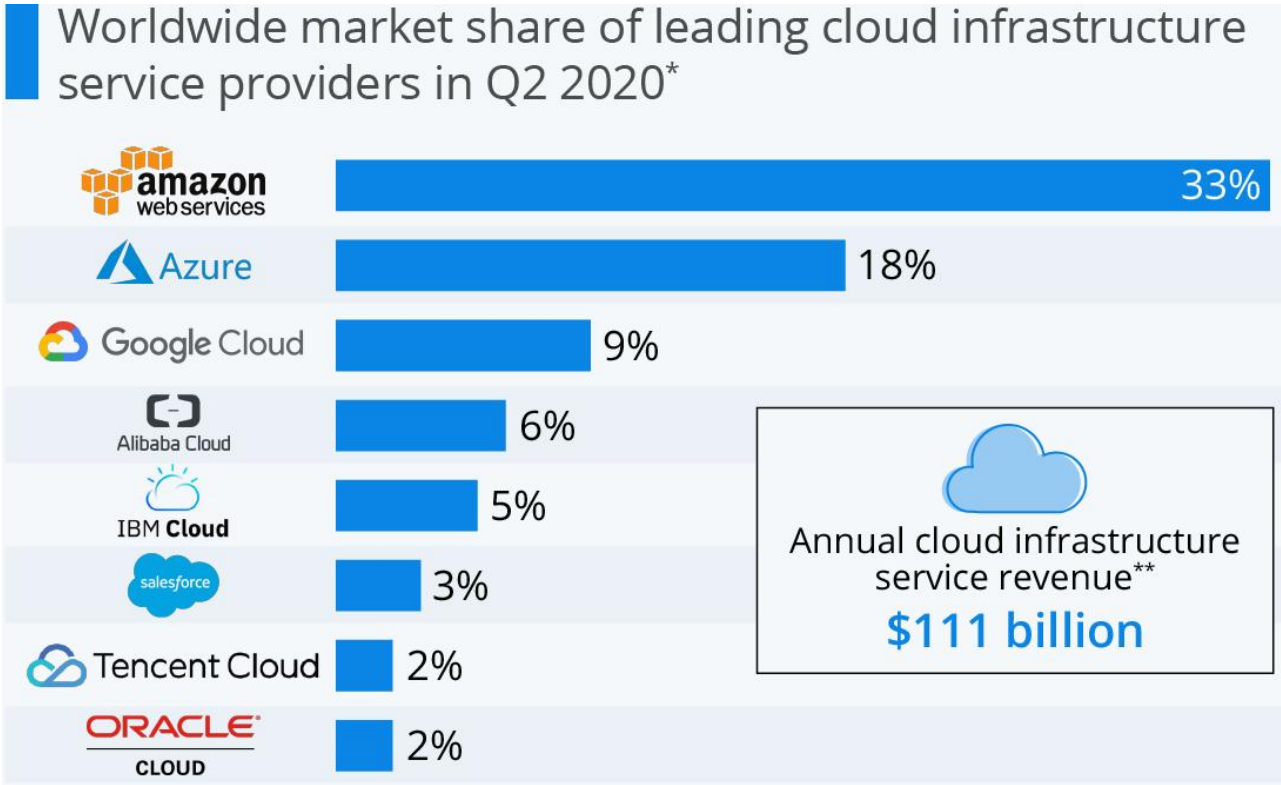


Parallel Training

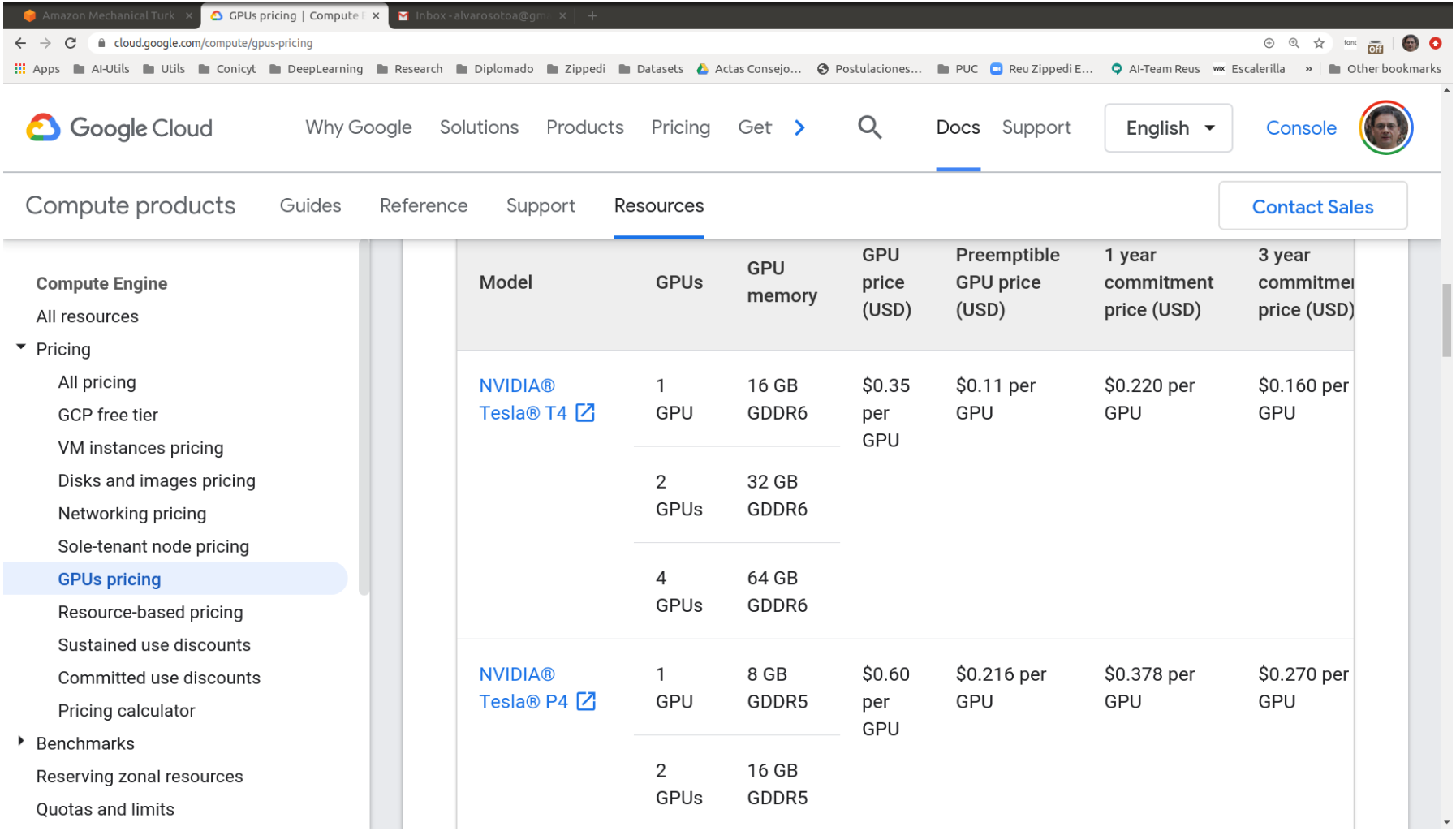


GPUs: Cloud

- Amazon
- Azure
- Google
- Alibaba
- ...

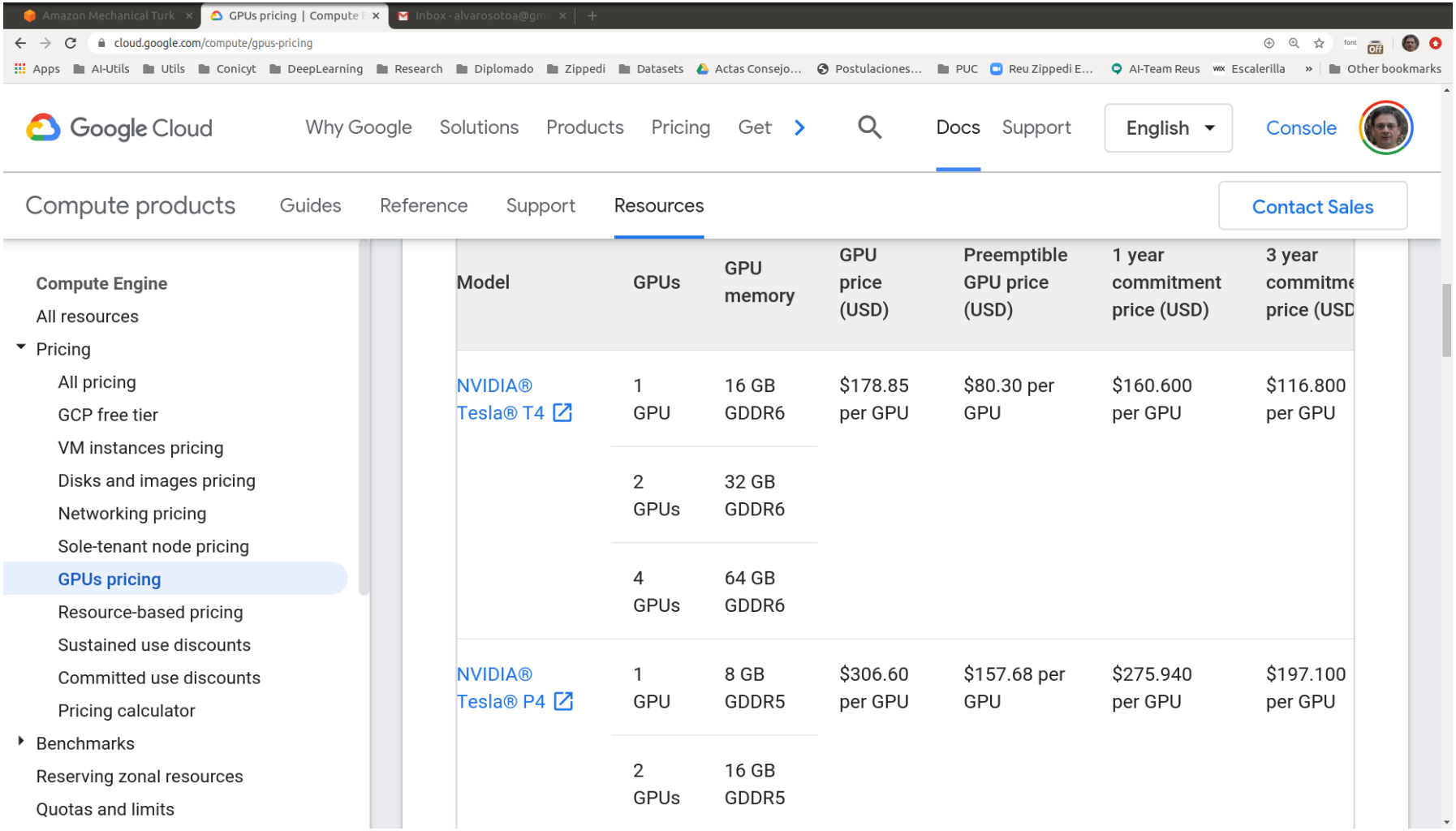


Google Cloud Hourly



Model	GPUs	GPU memory	GPU price (USD)	Preemptible GPU price (USD)	1 year commitment price (USD)	3 year commitmer price (USD)
NVIDIA® Tesla® T4	1 GPU	16 GB GDDR6	$0.35 per GPU	$0.11 per GPU	$0.220 per GPU	$0.160 per GPU
	2 GPUs	32 GB GDDR6				
	4 GPUs	64 GB GDDR6				
NVIDIA® Tesla® P4	1 GPU	8 GB GDDR5	$0.60 per GPU	$0.216 per GPU	$0.378 per GPU	$0.270 per GPU
	2 GPUs	16 GB GDDR5				

Google Cloud Monthly



Model	GPUs	GPU memory	GPU price (USD)	Preemptible GPU price (USD)	1 year commitment price (USD)	3 year commitme price (USD
NVIDIA® Tesla® T4	1 GPU	16 GB GDDR6	$178.85 per GPU	$80.30 per GPU	$160.600 per GPU	$116.800 per GPU
	2 GPUs	32 GB GDDR6				
	4 GPUs	64 GB GDDR6				
NVIDIA® Tesla® P4	1 GPU	8 GB GDDR5	$306.60 per GPU	$157.68 per GPU	$275.940 per GPU	$197.100 per GPU
	2 GPUs	16 GB GDDR5				

Google Cloud

Google expands its cloud with new regions in Chile, Germany and Saudi Arabia

Frederic Lardinois @fredericl / 2:11 PM GMT-3 • December 21, 2020

Comment

Microsoft Azure Cloud

INICIO TRENDTIC LIVE REVISTAS TRANSFORMACIÓN DIGITAL INDUSTRIAS & MERCADOS E-SALUD OPINIÓN

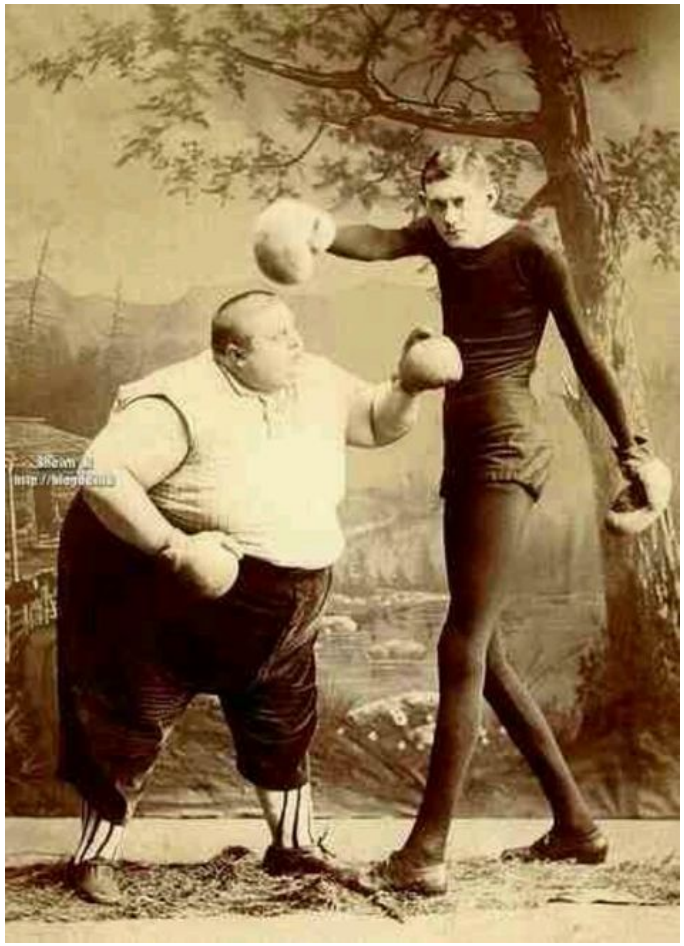
NOTICIAS Y MÁS

MICROSOFT ANUNCIA "TRANSFORMA CHILE" PARA ACELERAR EL CRECIMIENTO Y LA TRANSFORMACIÓN DE LOS NEGOCIOS

TRENDTIC | 10 diciembre, 2020 at 06:09 0

- ***Incluyendo una nueva región de data center, el compromiso de capacitar a más de 180.000 personas y un consejo asesor.***

Fat vs Thin System



Embedded GPUs

Developer Kits: Jetson Family

NVIDIA Jetson Modules

Nano

2-4GB
128-core Maxw.
5W/10W
70 x 45mm
$129

TX1

4GB
256-core Maxw.
10W
70 x 45mm
$100

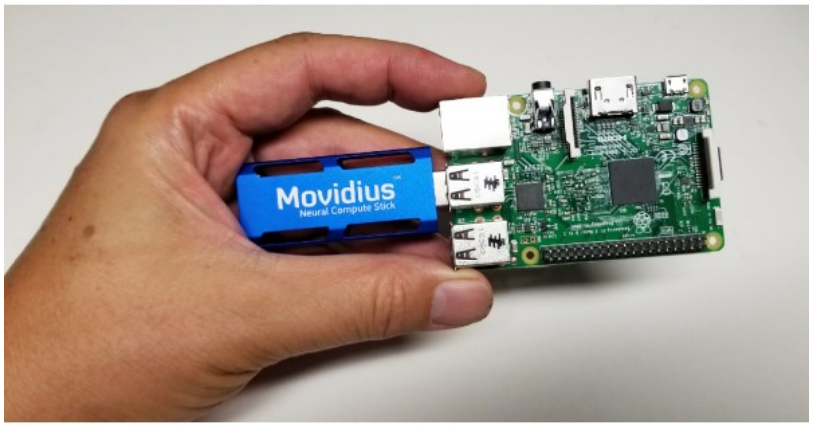
TX2

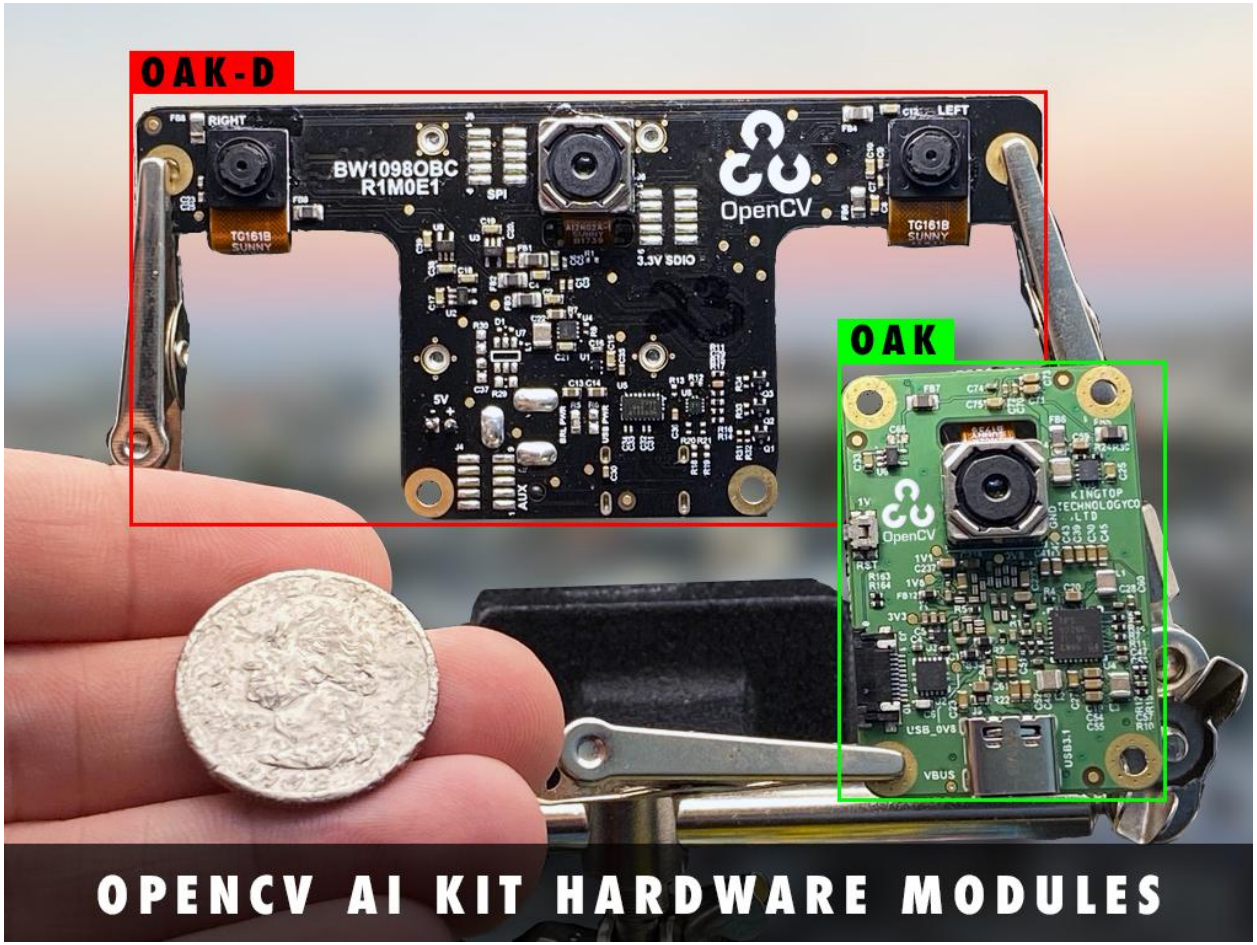
8GB
256-core Pascal
10-20W
87 x 50mm
$400

Xavier

32GB
512-core Volta
10-30W
105 x 105 mm
$900-1000

Edge Devices from Google and Intel





- Start selling 12/2020
- Two versions: OAK-D and OAK-1
- OAK-D incorporates a depth sensor
- Able to run visual DL algorithms in real time

CAMPAIGN

SPECIAL PRICE

OAK-1

4K AI CAMERA MODULE

$99 USD

MSRP $199 SAVE $100!

CAMPAIGN

SPECIAL PRICE

OAK-D

SPATIAL AI CAMERA MODULE

$149 USD

MSRP $299 SAVE $150!

https://www.kickstarter.com/projects/opencv/opencv-ai-kit

Mobile Devices

Several Options:

- Apple Bionic GPU's (Apple)
- Qualcomm Adreno (Google)
- ARM Mali (Samsung)
- ...

So far not real time performance
for visual recognition applications

Algoritmic alternatives: Model Compression

- Model distillation
- Sparsity
- Decrease data precision

Datasets

Imagenet

14,197,122 images, 21841 synsets indexed

Explore Download Challenges Publications Updates About

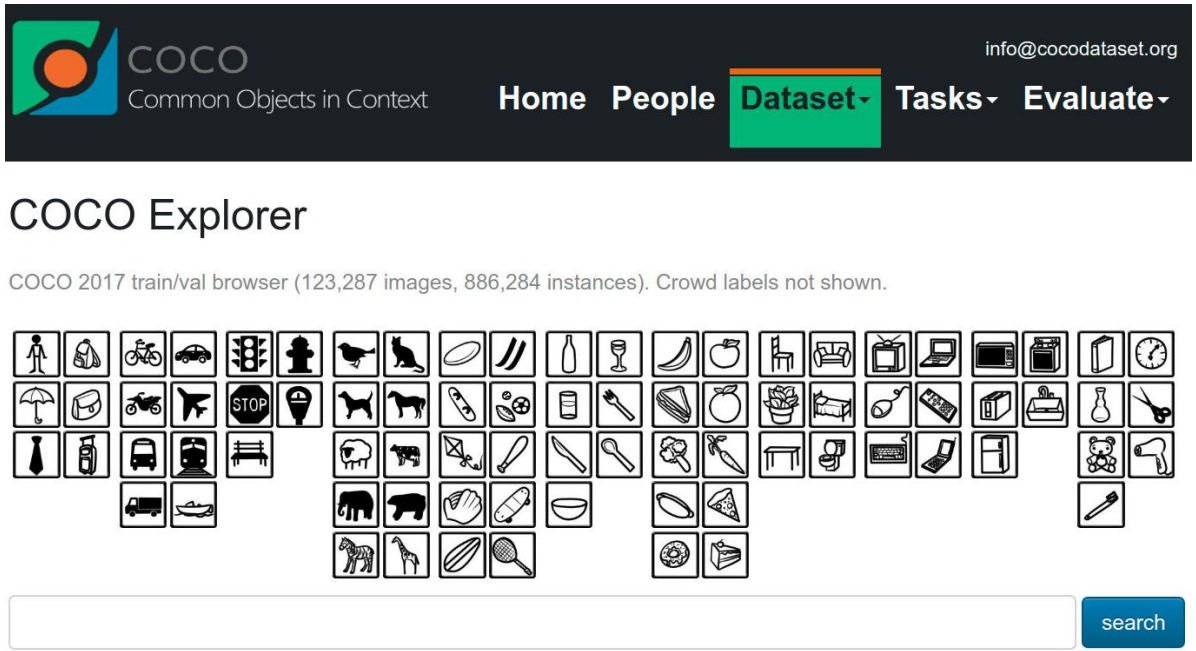
Not logged in. Login | Signup

ImageNet is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. Currently we have an average of over five hundred images per node. We hope ImageNet will become a useful resource for researchers, educators, students and all of you who share our passion for pictures.
Click here to learn more about ImageNet, Click here to join the ImageNet mailing list.

SEARCH

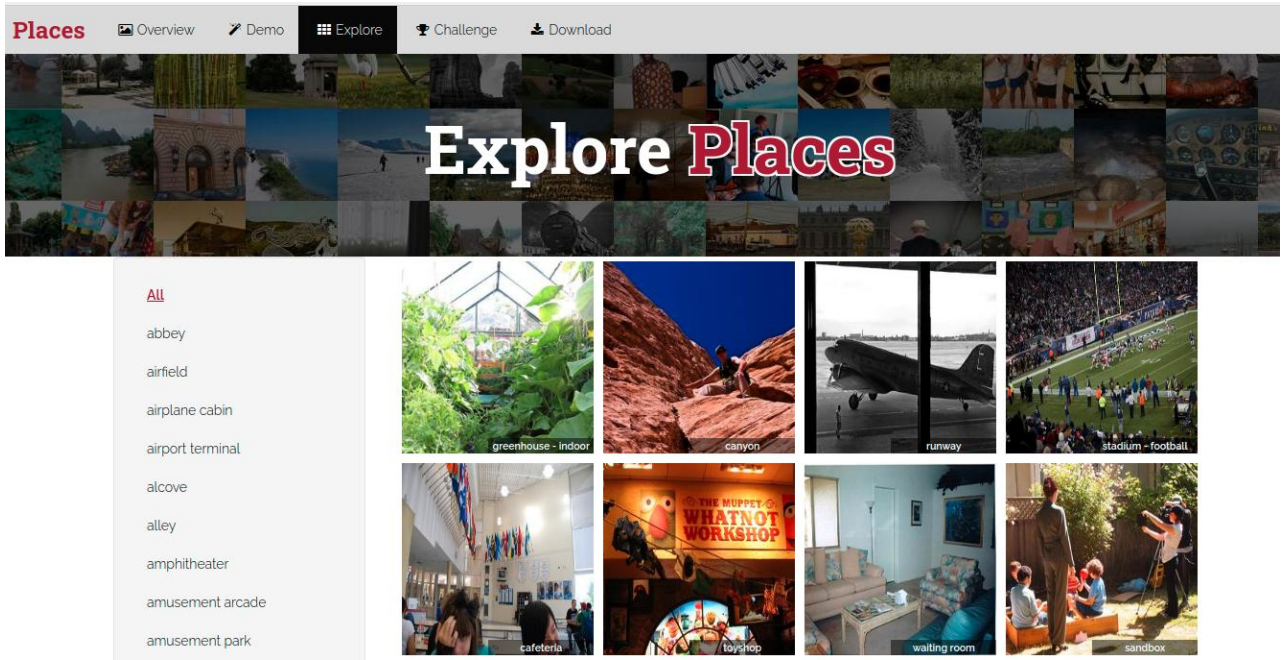
`https://www.image-net.org`

Microsoft-Coco



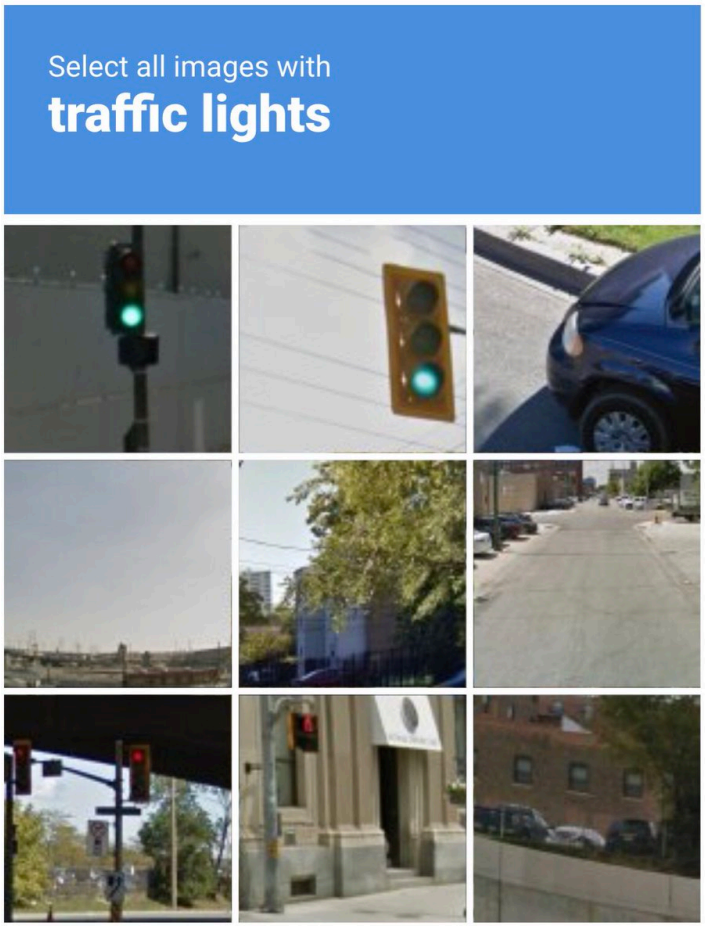
`https://cocodataset.org`

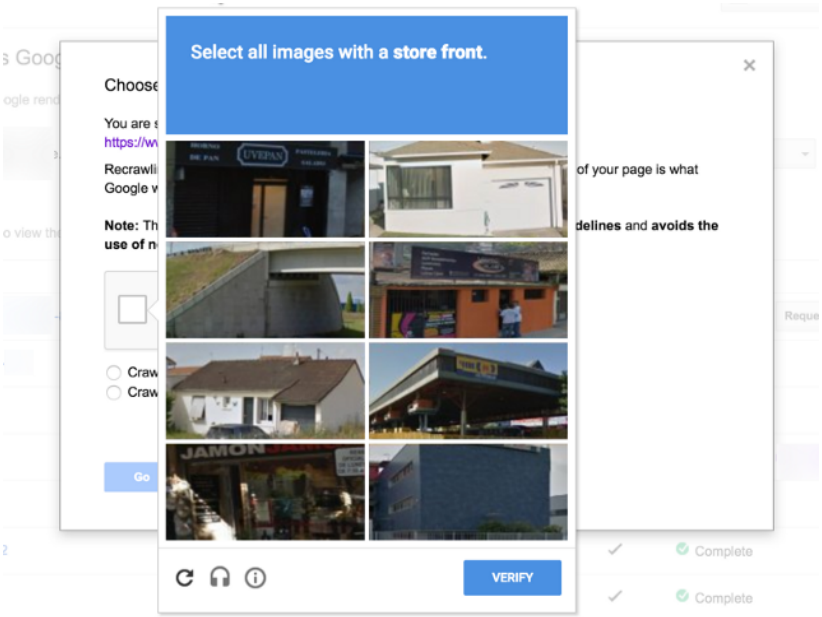
Scene Recognition

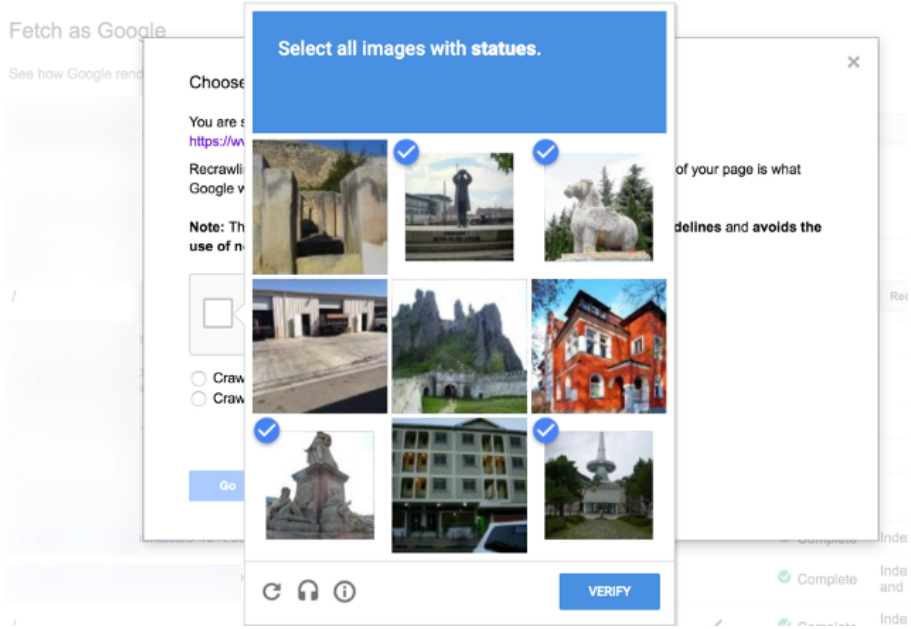


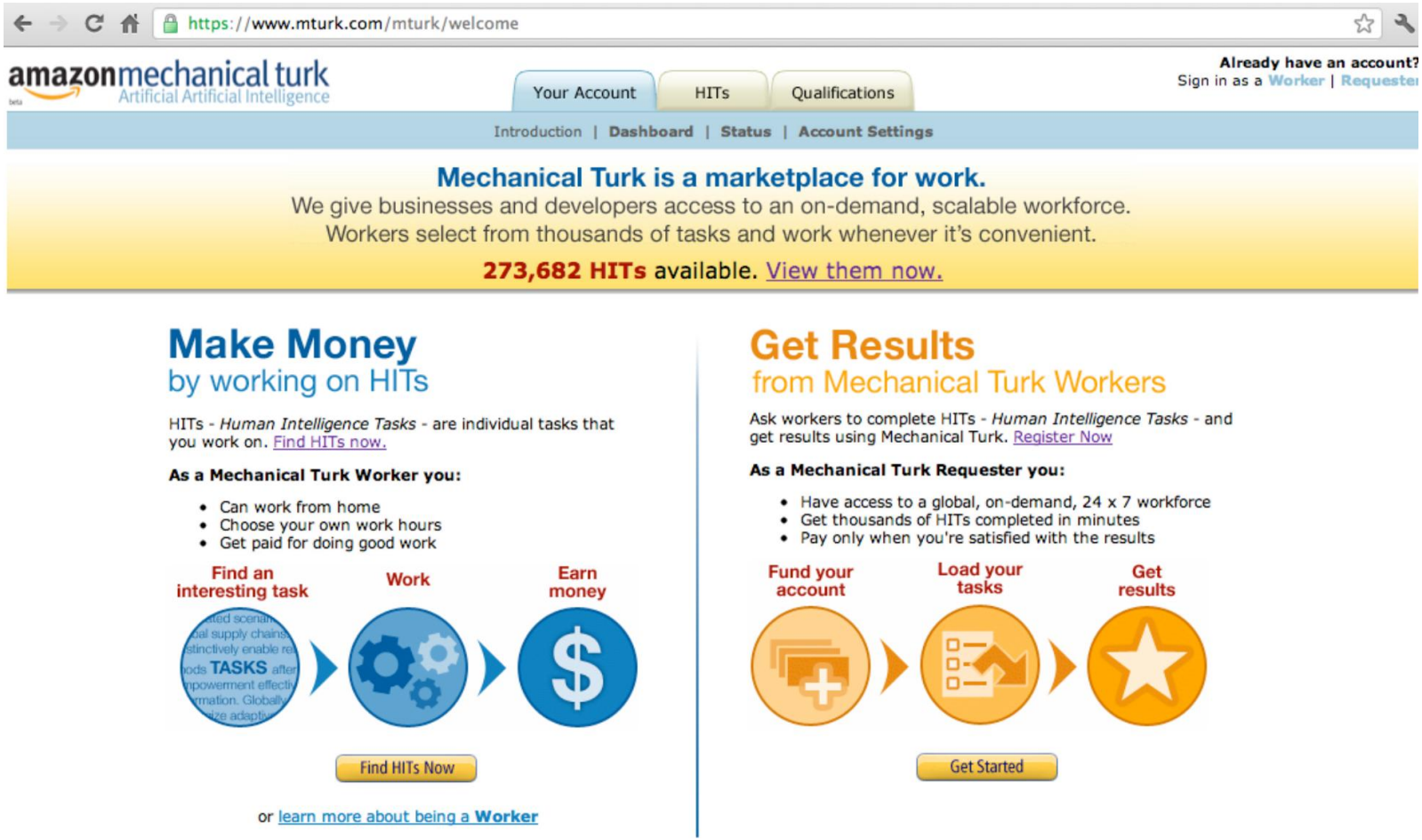
https://places2.csail.mit.edu

Data Labeling: Crowdsourcing


Select all images with
traffic lights


Select all images with a store front.
VERIFY


Select all images with statues.
VERIFY


https://www.mturk.com/mturk/welcome
amazonmechanical turk
Artificial Artificial Intelligence
Your Account
HITs
Qualifications
Already have an account?
Sign in as a Worker | Requester
Introduction | Dashboard | Status | Account Settings
Mechanical Turk is a marketplace for work.
We give businesses and developers access to an on-demand, scalable workforce.
Workers select from thousands of tasks and work whenever it's convenient.
273,682 HITs available. View them now.
Make Money
by working on HITs
HITs - Human Intelligence Tasks - are individual tasks that you work on. Find HITs now.
As a Mechanical Turk Worker you:
Can work from home
Choose your own work hours
Get paid for doing good work
Find an interesting task
Work
Earn money
Find HITs Now
or learn more about being a Worker
Get Results
from Mechanical Turk Workers
Ask workers to complete HITs - Human Intelligence Tasks - and get results using Mechanical Turk. Register Now
As a Mechanical Turk Requester you:
Have access to a global, on-demand, 24 x 7 workforce
Get thousands of HITs completed in minutes
Pay only when you're satisfied with the results
Fund your account
Load your tasks
Get results
Get Started

hCaptcha

For Websites | Labeling Services | About | Docs | Login | Get hCaptcha

FROM THE BLOG: How hCaptcha Calculates Rewards

Cost-effective data labeling at scale

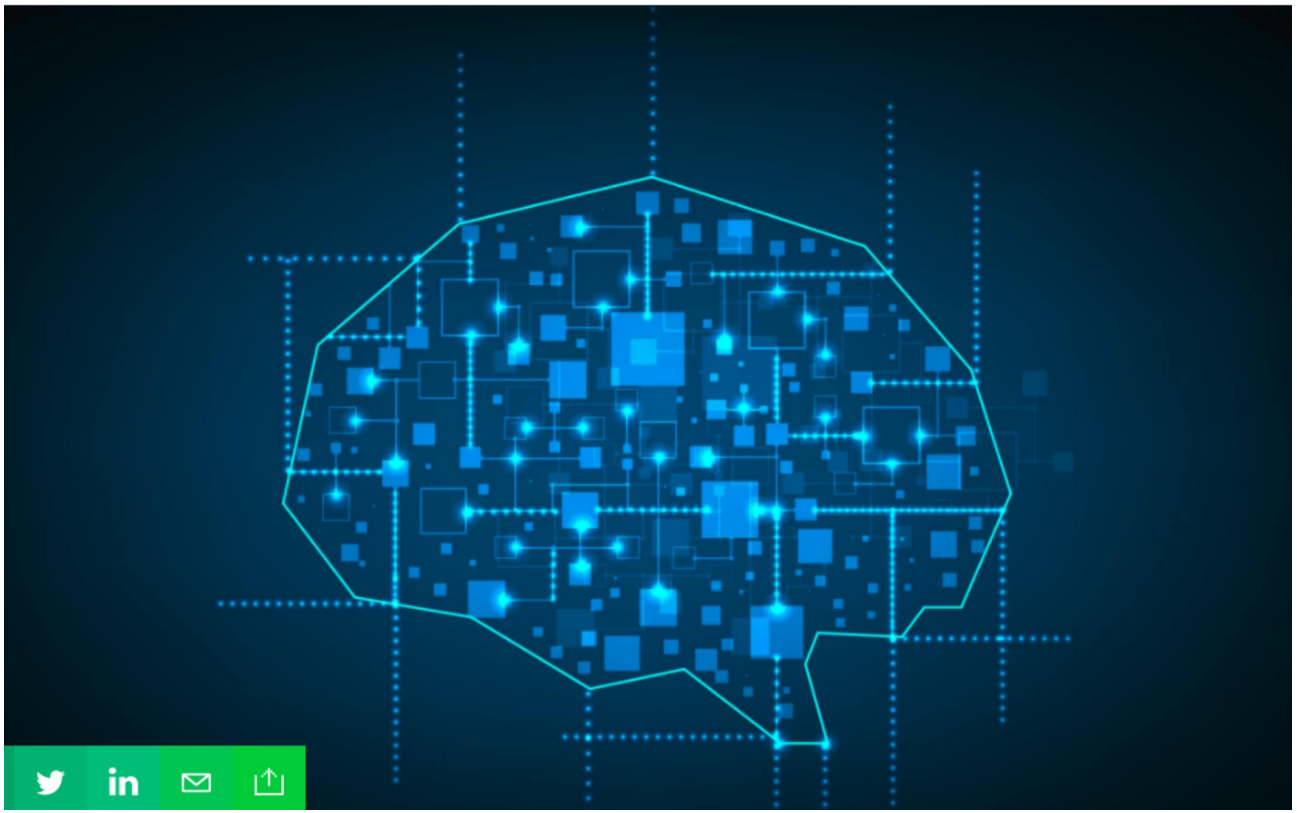
Do you have large datasets you need to understand at a human level? hCaptcha provides fast, cost-effective, and high quality data labeling for AI & machine learning companies among many others.

Request a labeling job

Appen acquires Figure Eight for up to $300M, bringing two data annotation companies together

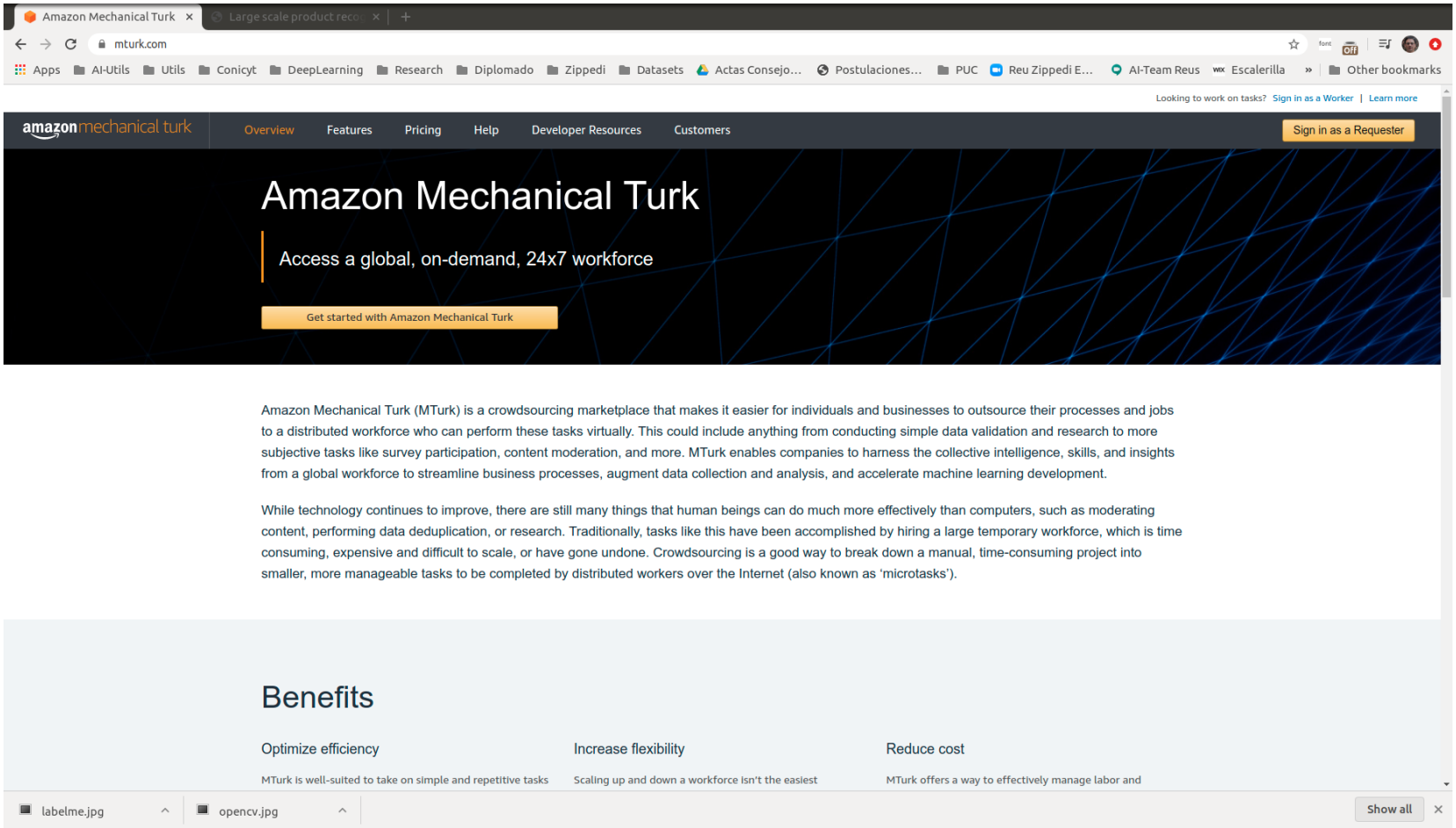
Anthony Ha @anthonyha / 4 hours ago

Comment



10/03/2019

AMT Example

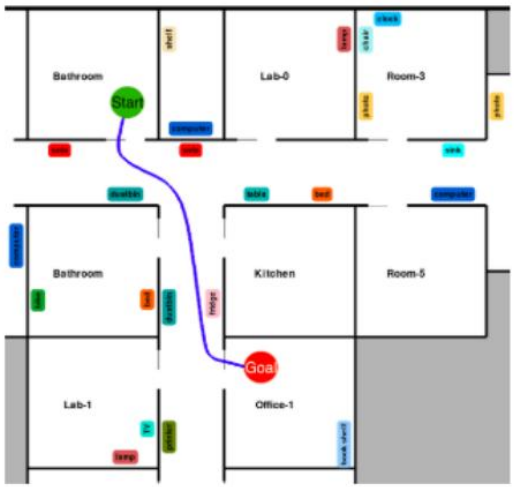

Amazon Mechanical Turk
mturk.com
Looking to work on tasks? Sign in as a Worker | Learn more
amazon mechanical turk
Overview
Features
Pricing
Help
Developer Resources
Customers
Sign in as a Requester
Amazon Mechanical Turk
Access a global, on-demand, 24x7 workforce
Get started with Amazon Mechanical Turk
Amazon Mechanical Turk (MTurk) is a crowdsourcing marketplace that makes it easier for individuals and businesses to outsource their processes and jobs to a distributed workforce who can perform these tasks virtually. This could include anything from conducting simple data validation and research to more subjective tasks like survey participation, content moderation, and more. MTurk enables companies to harness the collective intelligence, skills, and insights from a global workforce to streamline business processes, augment data collection and analysis, and accelerate machine learning development.
While technology continues to improve, there are still many things that human beings can do much more effectively than computers, such as moderating content, performing data deduplication, or research. Traditionally, tasks like this have been accomplished by hiring a large temporary workforce, which is time consuming, expensive and difficult to scale, or have gone undone. Crowdsourcing is a good way to break down a manual, time-consuming project into smaller, more manageable tasks to be completed by distributed workers over the Internet (also known as 'microtasks').
Benefits
Optimize efficiency
MTurk is well-suited to take on simple and repetitive tasks
Increase flexibility
Scaling up and down a workforce isn't the easiest
Reduce cost
MTurk offers a way to effectively manage labor and
labelme.jpg
opencv.jpg
Show all

AMT Example

HIT Instructions

You will be given a map of an indoor environment below. In the box to the right of the map, you want to write a verbal description of how to get from the **starting point** **(green circle)** to the **end goal** **(red circle)** in English. This verbal description should be written in a way that another person could read it and navigate from the start position to the goal without the map, and just by looking at landmarks in the environment. These landmarks can be places (like kitchens, bathrooms, offices, etc.) or objects (doors, shelfs, bikes, etc.). Please make sure that your navigation instructions are not ambiguous and match the route in the map.

An example map and 2 possible verbal descriptions for the route in blue are provided below. Both of these descriptions are valid.



Start place: Bathroom
End place: Office-1

Example route 1:

"Go out the bathroom and turn left. Turn right at the first corner you see and follow the corridor. Pass a dustbin on the right, a fridge on the left, and enter the office 1 on your left."

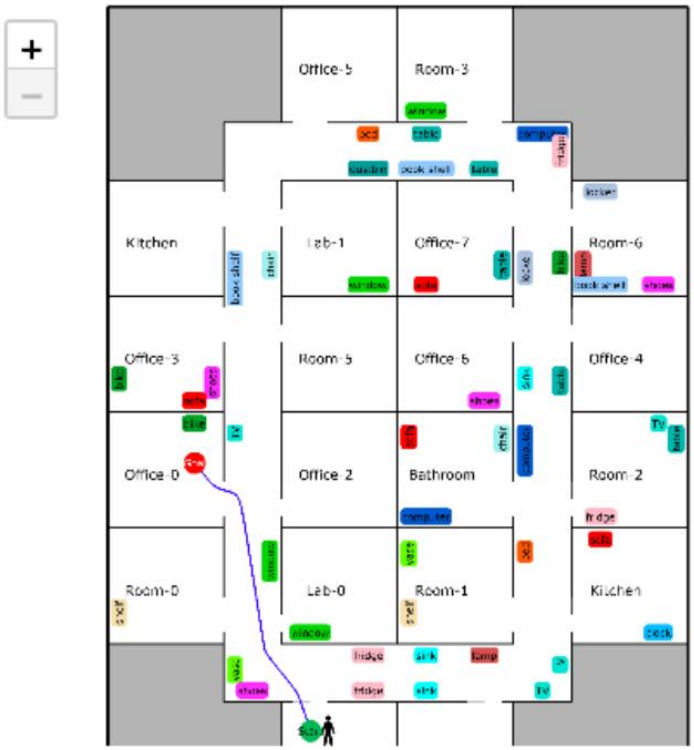
Example route 2:

"Go out, turn left, and turn right at the next corridor when you see the sofa. Enter office 1 on your left after passing a dustbin."

The map with the route that you should describe for this HIT is shown after these instructions.

AMT Example

You can zoom into the map by scrolling or pressing the + and - buttons. You can also displace the image to look at another part of the map by clicking, holding the mouse, and dragging in the desired direction



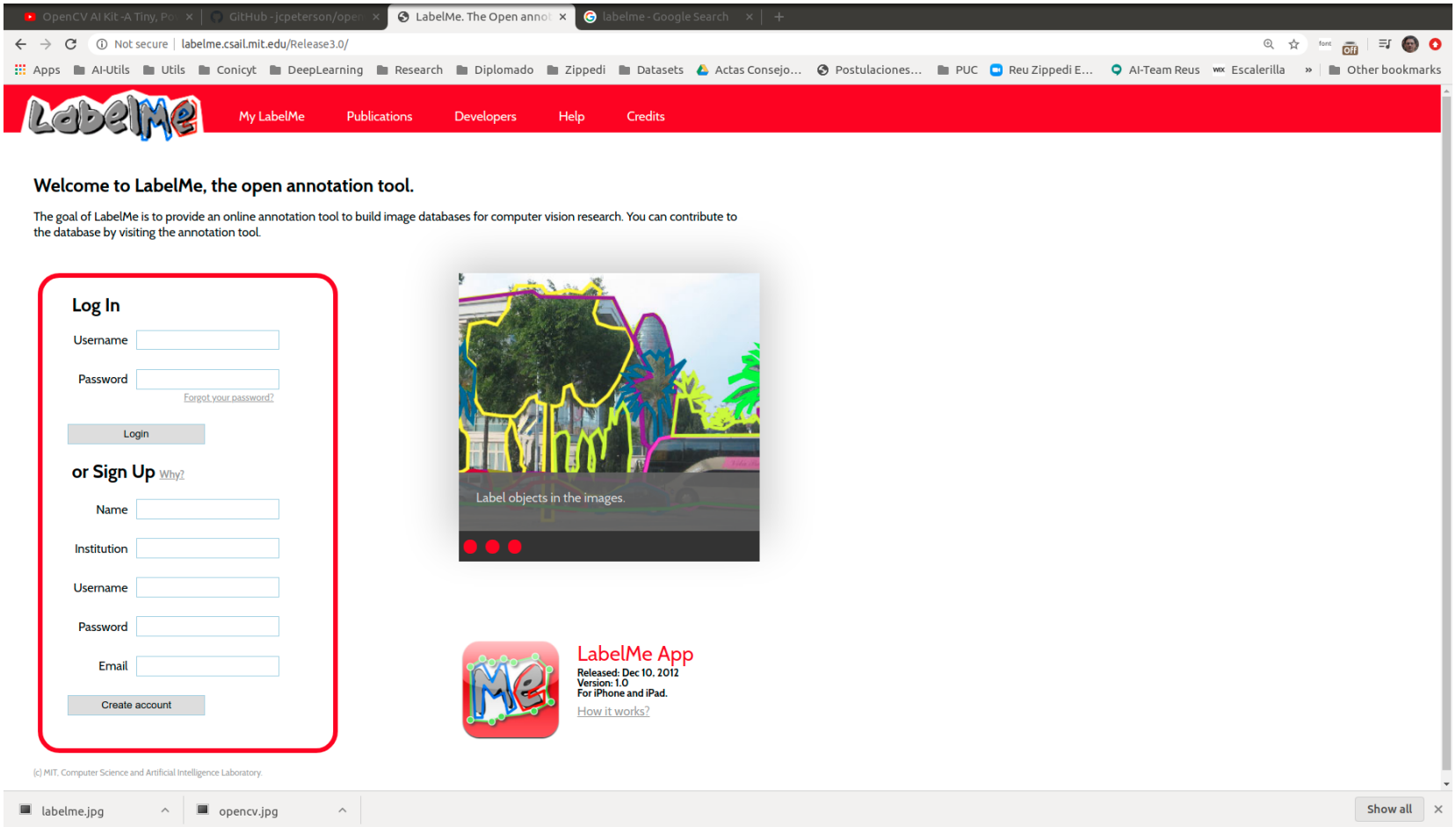
Start Place

End Place

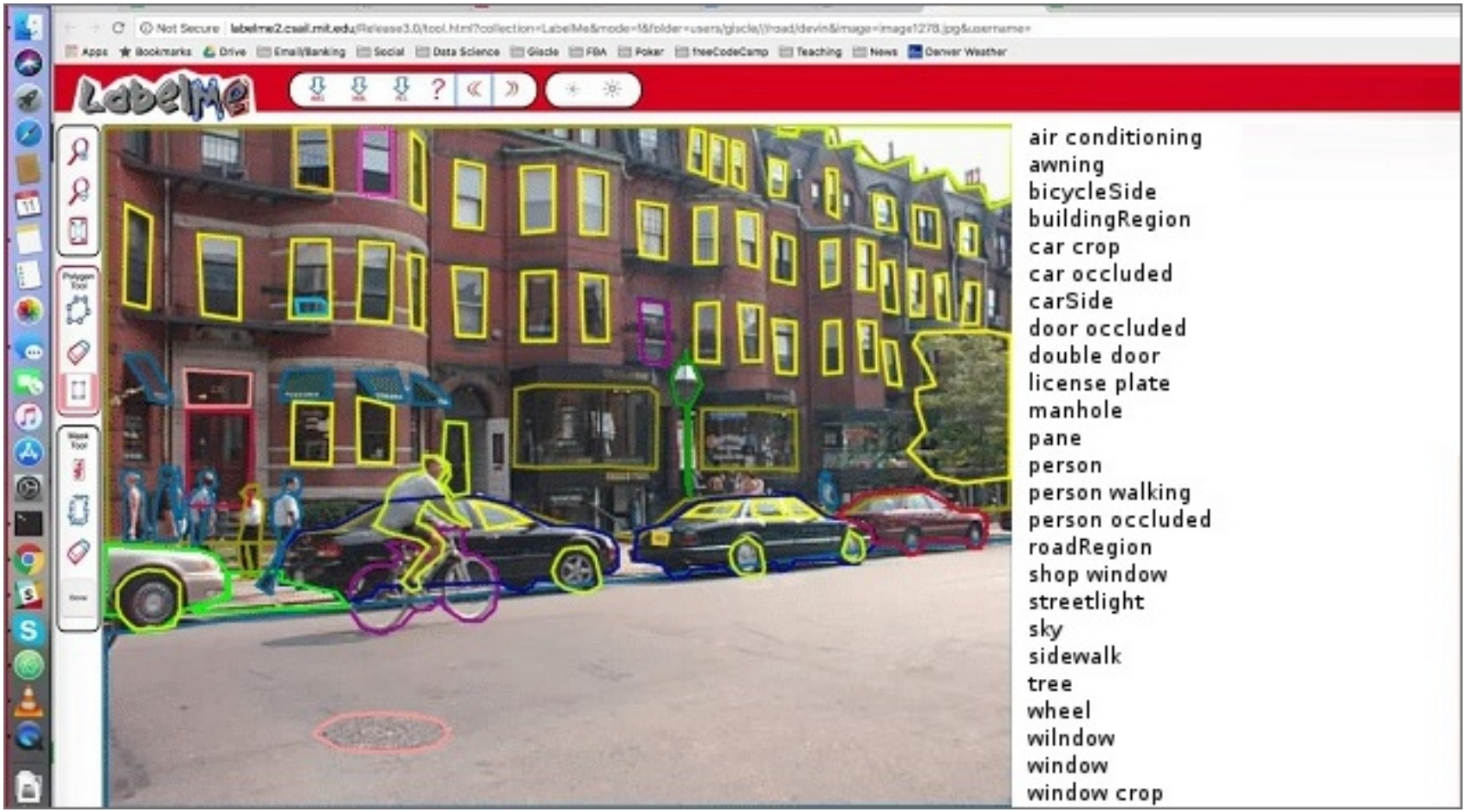
Directions

Data Labeling: Manual

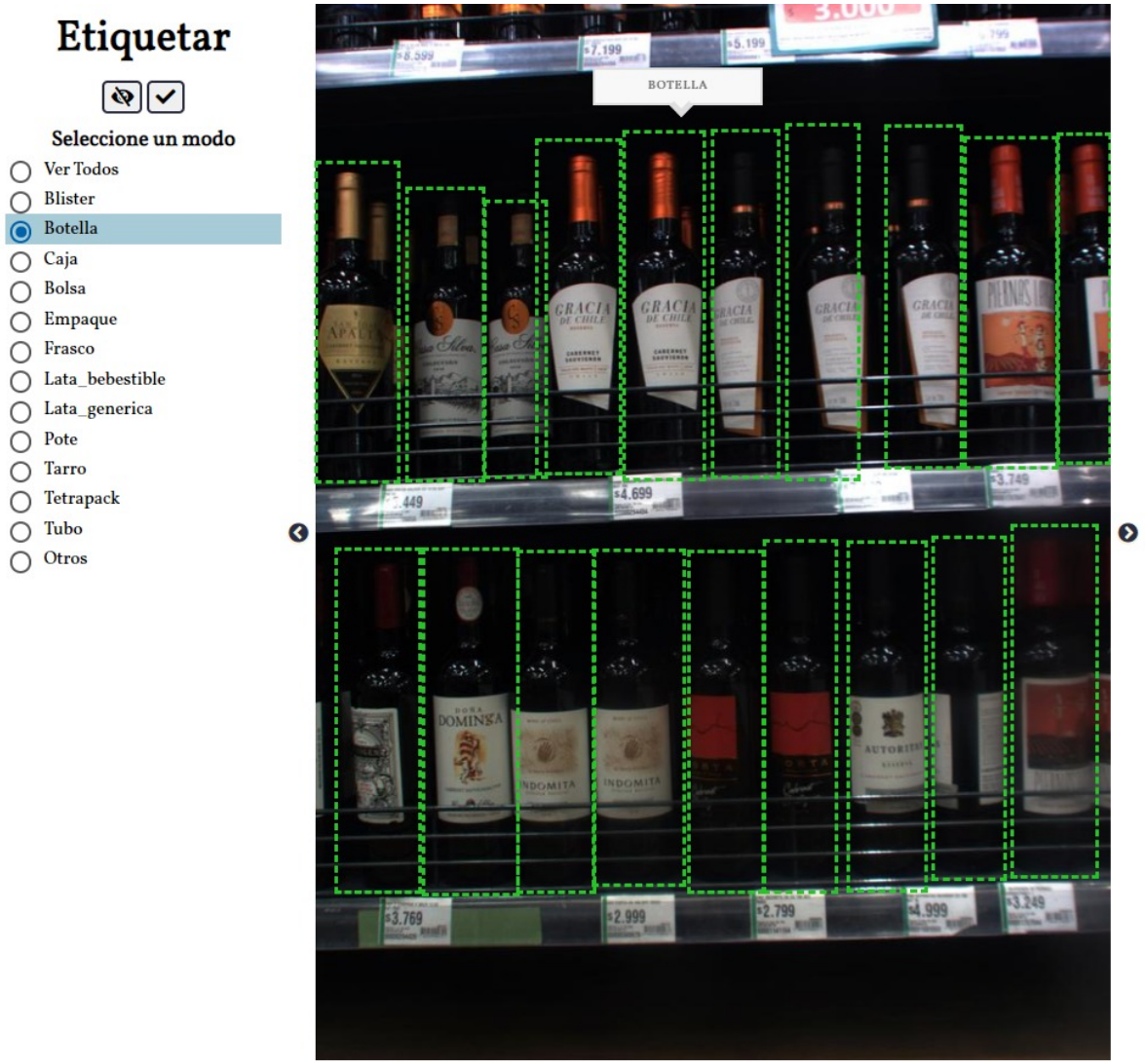
Tools to Label



Tools to Label



Create Your Own Tool



Create Your Own Tool