Applied Machine Learning!!!

W207 Section 9

Rasika Bhalerao
rasikabh@berkeley.edu

Aug 23: Welcome!
Nov 8 and 22: No classes

Schedule

Supervised learning methods

	Sync	Topic
2	Aug 30	Linear Regression / Gradient Descent
3	Sep 6	Feature Engineering Bonus: Naive Bayes
4	Sep 13	Logistic Regression
5	Sep 20	Multiclass classification / Eval Metrics Bonus: Reinforcement learning
6	Sep 27	Neural Networks
7	Oct 4	KNN, Decision Trees, Ensembles

Unsupervised learning methods

	Sync	Topic
8	Oct 11	KMeans and PCA
9	Oct 18	Text Embeddings Bonus: Language models
10	Oct 25	CNNs Bonus: GANs
11	Nov 1	EDA, Real data, Baselines, LDA
12	Nov 15	Fairness / Ethics
13	Nov 29	Fancy Neural Networks
14	Dec 6	Final Presentations

Assignment Schedule

Due Date	Assignment
Aug 28	HW1
Sep 4	HW2
Sep 11	HW3
Sep 18	HW4
Sep 25	HW5
Oct 2	HW6
Oct 16	Group project baseline
Oct 23	HW8
Nov 6	HW9
Nov 20	HW10
Dec 4	Final project notebook + presentation

Behavior expectations

- Healthy disagreement is expected
- Be mindful of one another’s schedules
- Be a good listener
- Have fun in a professional manner
- Share related real-world experience
- Ask questions when something is confusing
- Keep it 100 but be respectful
- Be open-minded to new ideas in the real world and when coding
- On time for group meetings

How are final projects going?

Guidelines:
https://docs.google.com/document/d/1R7mIHOtYXKU8vEQzw10uofb_iK3sgimw8iZLWSTzdgg/edit?usp=sharing

Why use a Convolutional layer?

Why put a Pooling layer after a Convolutional layer?

What is the difference between data preprocessing and augmentation?

View options (at the top) → Annotate

Async Practice Quiz Questions (vote!)

Position invariance refers to the fact that an object in different orientations remains the same object.	True	False
A convolutional filter is applied in exactly one location in an input image.	True	False
Max-pooling with Stride 2 applied to an input with shape (32, 32, 8) would produce an output with shape (16, 16, 4).	True	False
A CNN typically involves more operations than it has parameters because each filter is applied many times.	True	False
ImageNet annotations are guaranteed to be 100% correct.	True	False
The data augmentation strategy is to simply duplicate the training images.	True	False
It would be impossible to apply a Conv2D operator to text.	True	False

A quick note on Crowdsourcing

Reference: This talk by Adina Williams and Nikita Nangia

How ML researchers get labeled data

- Most ML research is not high-profile
- A lot of ML research requires collecting new datasets
    - Especially when applying ML to other fields (social media, biology, security, etc)
    - Or building benchmarks to measure progress
    - Or measuring bias in models (CrowS-Pairs and similar datasets)
- If you want to collect a labeled dataset, you will probably use crowdsourcing

Crowdsourcing

- “outsourcing a job traditionally performed by an employee to an undefined, generally large group of people” (Li et al., 2016)
  - It is a job, part of the “gig economy” (“short-term contract, piecemeal, or freelance work as opposed to permanent employment”)
  - This is making the source people
- Very different from Human Subjects Research, which is “systematic, scientific investigation involving human beings as research subjects”
  - We are studying the text data, not the people who collected it
  - Because of sample bias
  - Because ethically this is different (IRB needs to approve any human subjects research)
- It is subset of Collective Intelligence (“people collectively acting, often doing better than individual experts at solving some task”)
  - Other examples of Collective Intelligence include Stackoverflow, Reddit, Quora, WikiHow

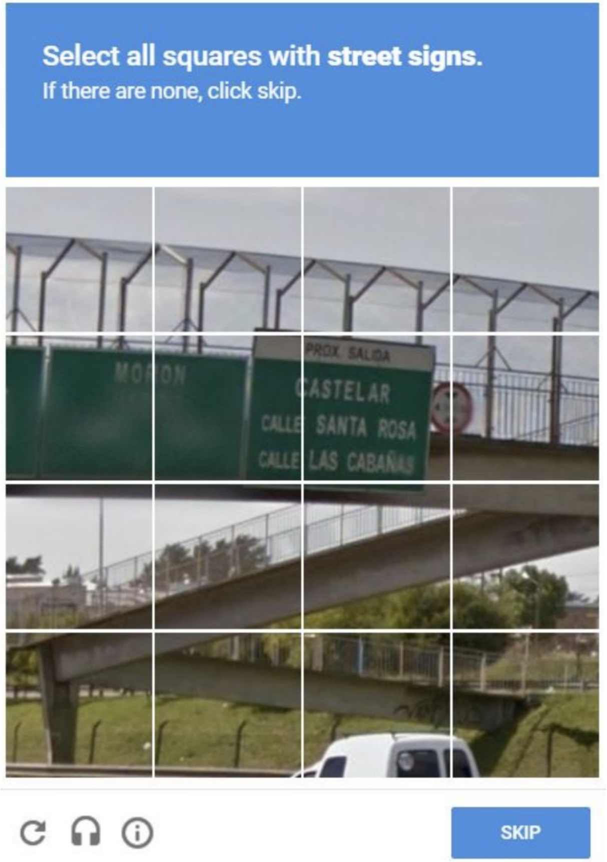
Some NLP tasks that need crowdsourced data

- Translation
- Textual Entailment / Natural Language Inference
- Question Answering
- Sentiment Analysis
- Image Captioning
- Word Similarity
- Word Sense Disambiguation
- Treebanking
- Summarization

https://drive.google.com/file/d/1qqLmFvCe-6jcKdXjcO0-SOxBHHXkWvYR/view

Where to do crowdsourcing

- Paid (usual)
  - Most popular platforms: Amazon Mechanical Turk and Figure Eight (previously known as Crowdflower)
- Unpaid (unusual)
  - Citizen Science
    - Voluntary (to progress towards a common goal for the betterment of humanity)
    - https://www.citizenscience.gov/catalog, https://www.zooniverse.org/projects
  - Gamification
    - Collect data from game interaction
    - Duolingo
  - Unpaid crowdsourcing
    - Involuntary, brief tasks embedded in normal life
    - reCAPTCHA



Select all squares with street signs.

If there are none, click skip.

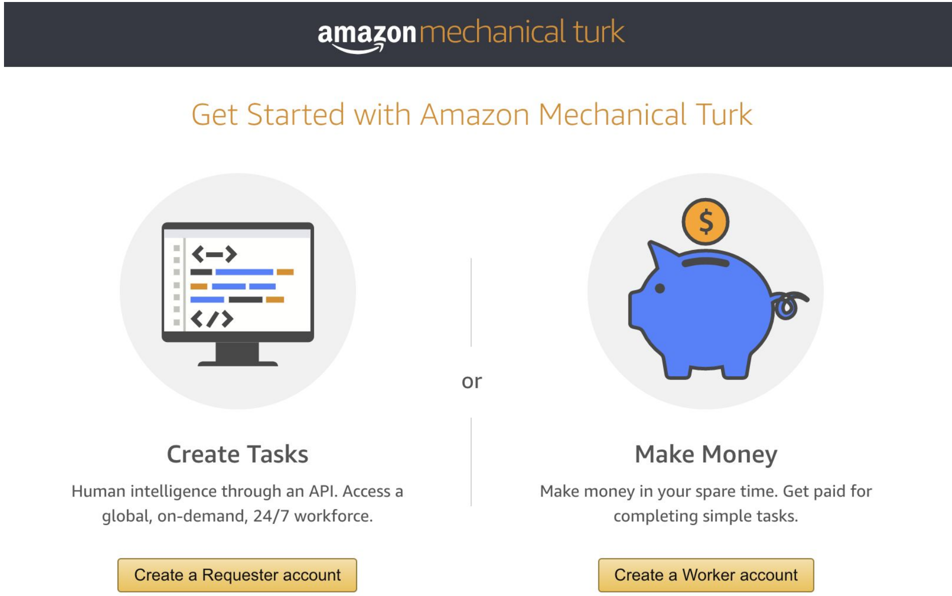
https://drive.google.com/file/d/1qqLmFvCe-6jcKdXjcO0-SOxBHHXkWvYR/view

How to get crowdsourced data

- Use Amazon Mechanical Turk or Appen (previously known as Figure Eight (previously known as Crowdflower))
- Ask workers to do “simple and repetitive tasks” [AMT website]
- You are now “managing a temporary workforce” [AMT website]

appen

figure eight



amazon mechanical turk

Get Started with Amazon Mechanical Turk

Create Tasks

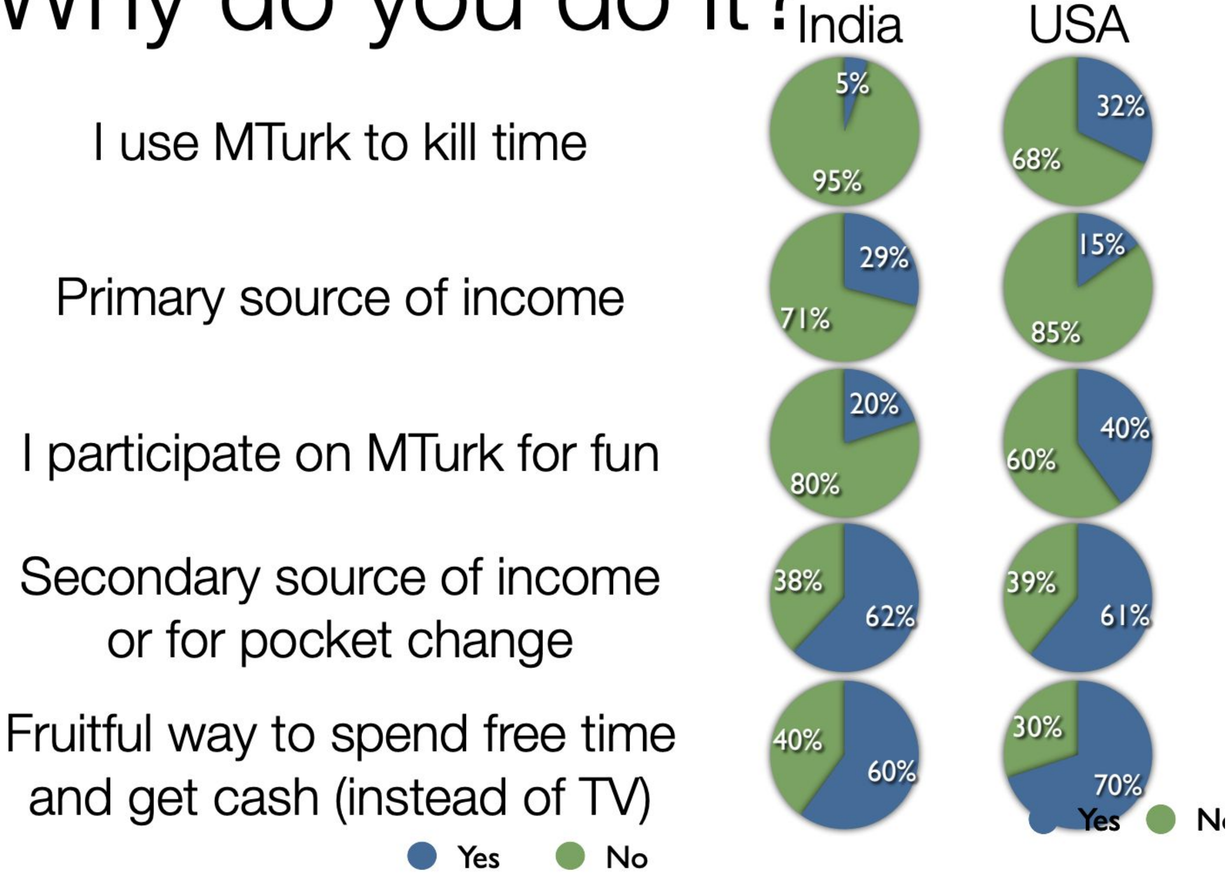
Human intelligence through an API. Access a global, on-demand, 24/7 workforce.

or

Make Money

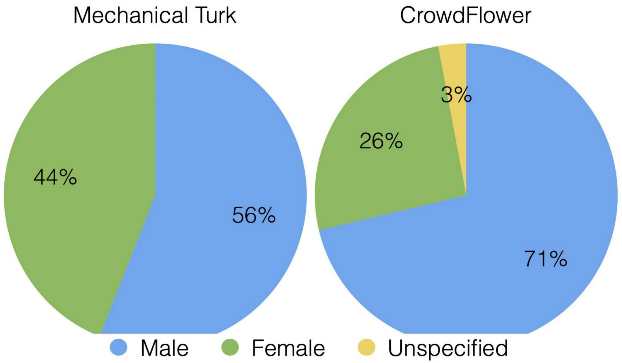
Make money in your spare time. Get paid for completing simple tasks.

Why do you do it?



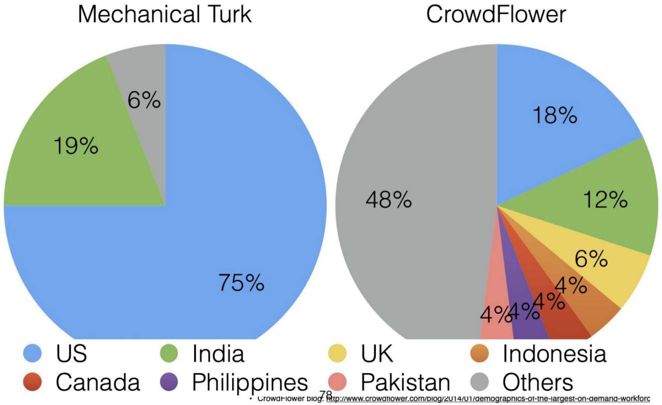
http://crowdsourcing-class.org/slides/crowd-workers.pdf

Demographic comparisons (the study they came from)



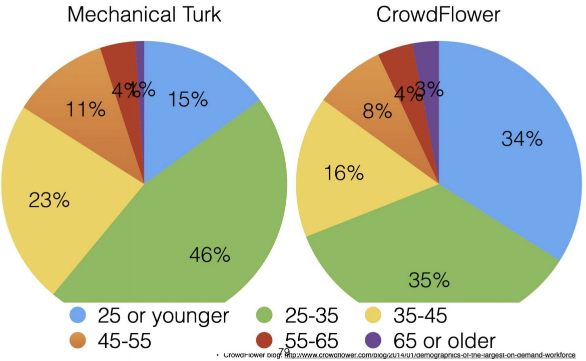
- Panos Ipeirotis's site: http://demographics.mturk-tracker.com
- CrowdFlower blog: http://www.crowdflower.com/blog/2014/01/demographics-of-the-largest-on-demand-workforce

77



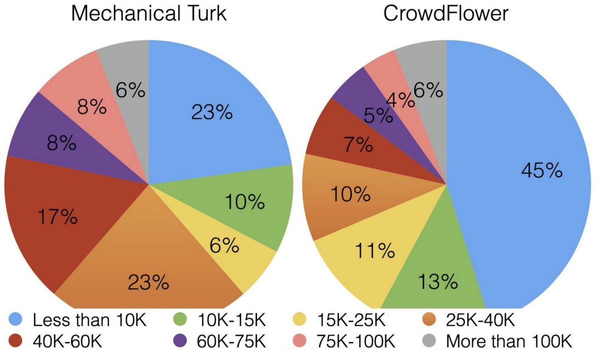
78

- CrowdFlower blog: http://www.crowdflower.com/blog/2014/01/demographics-of-the-largest-on-demand-workforc



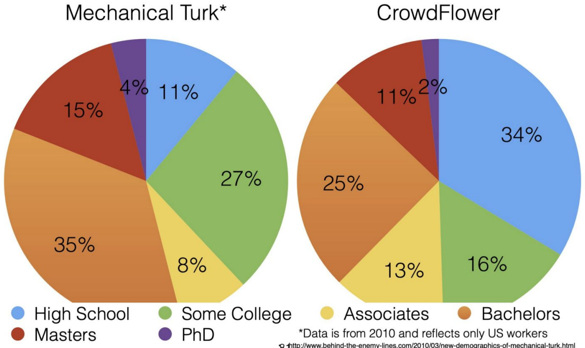
79

- CrowdFlower blog: http://www.crowdflower.com/blog/2014/01/demographics-of-the-largest-on-demand-workforce



80

- http://www.behind-the-enemy-lines.com/2010/03/new-demographics-of-mechanical-turk.html
- http://www.crowdflower.com/blog/2014/01/demographics-of-the-largest-on-demand-workforce



*Data is from 2010 and reflects only US workers

81

- http://www.behind-the-enemy-lines.com/2010/03/new-demographics-of-mechanical-turk.html
- http://www.crowdflower.com/blog/2014/01/demographics-of-the-largest-on-demand-workforce

https://drive.google.com/file/d/1qqLmFvCe-6jcKdXjcO0-SOxBHHXkWvYR/view

Concerns

Requesters (us)

- Are workers doing their best work?
- It is possible to cheat (click randomly, have a script to click randomly)
- We don’t get to judge workers’ skill in advance
- It is hard to automatically validate labels

Crowdworkers (workers)

- Not paid enough
  - Time spent searching for tasks is unpaid
  - Rejected labels are unpaid
  - Pay is already low by itself
- Gig workers have no employment stability / benefits
- Requesters can decide your work wasn’t good enough (without giving a reason) and refuse to pay you
- The work could be unethical, stressful, or triggering

From Nikita Nangia’s slide

How to get good crowdsourced labels

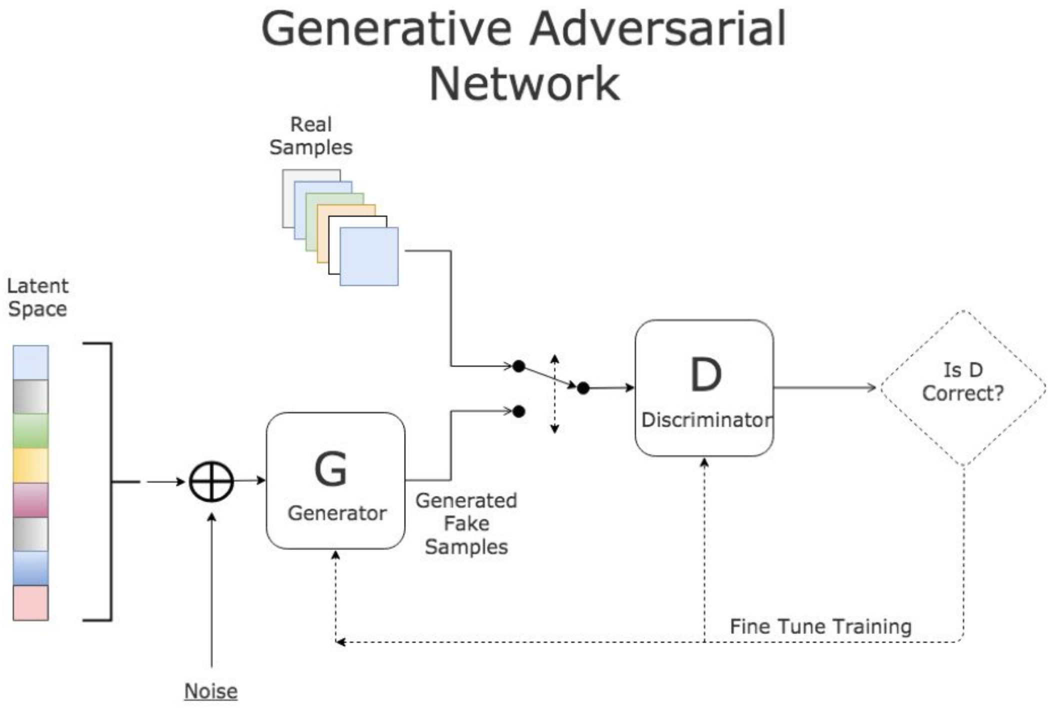
- “Market for Lemons”: requesters cannot judge quality of labels, so they pay them all average price, incentivizing workers to do poorer quality work since they get paid poorly either way (speed through as many tasks as possible)
- “Gold Labels”: requesters label a percent of the data themselves, and sprinkle them in the crowdsource task. If a worker gets too many of these “Catch Trials” wrong, reject them.
- Make sure to specify the instructions clearly, with examples and an FAQ
  - Answer their questions via email
- Pay them well. Use scripts to make sure you pay at least min wage
- Workers are also communicating on their forums and Turkopticon

Preview of homework 10

Generative Adversarial Networks (GANs)

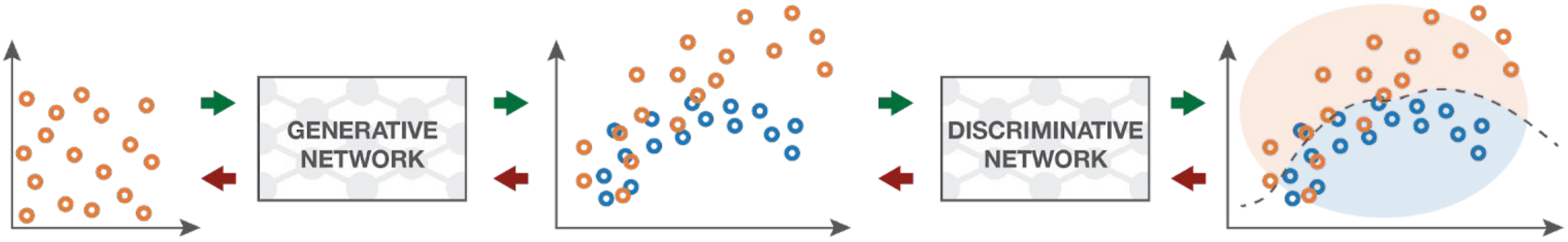
Train two networks, a discriminator and a generator

- The discriminator tries to tell real from generated images; gets rewarded for getting it right
- The generator tries to fool the discriminator; gets rewarded for the discriminator getting it wrong



GAN Training

🟩 Forward propagation (generation and classification) 🟥 Backward propagation (adversarial training)



Input random variables.

The generative network is trained to **maximise** the final classification error.

The generated distribution and the true distribution are not compared directly.

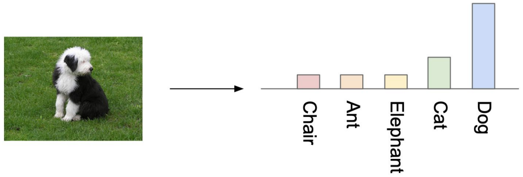
The discriminative network is trained to **minimise** the final classification error.

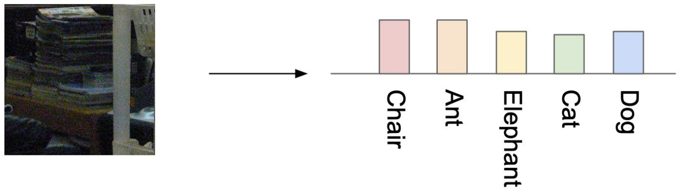
The classification error is the basis metric for the training of both networks.

https://towardsdatascience.com/understanding-generative-adversarial-networks-gans-cd6e4651a29

Inception Score

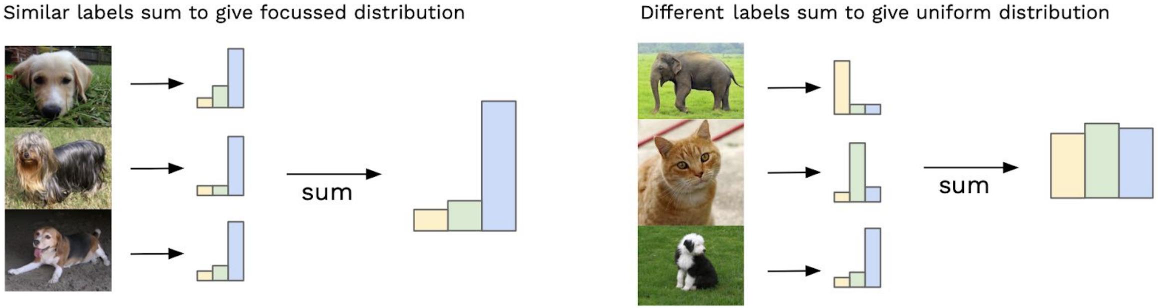
- How good is the GAN? → How realistic are the images?
- Two criteria to optimize:
  - Each image looks like something





  - Variety in images

- Measure: difference in these two distributions (KL-divergence) (0 to inf)


Similar labels sum to give focussed distribution

Different labels sum to give uniform distribution

https://medium.com/octavian-ai/a-simple-explanation-of-the-inception-score-372dff6a8c7a

BigGAN

Project done by an intern at Google DeepMind

Interesting GAN demo

https://ganbreeder.app/

(Note: On Sept. 28 OpenAI released an even better version called DALL-E 2 that uses GPT-3 and not a GAN: https://labs.openai.com/)