ML4 B&I: Introduction to Machine Learning

Lecture 8- Fairness in ML

Murat A. Erdogdu

Vector Institute, Fall 2022

Overview of topics

- Environmental impacts
- Financial costs
- Fairness

- Lecture based on:

  [BGMS] On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? by Bender, Gebru, McMillan-Major, Shmitchell

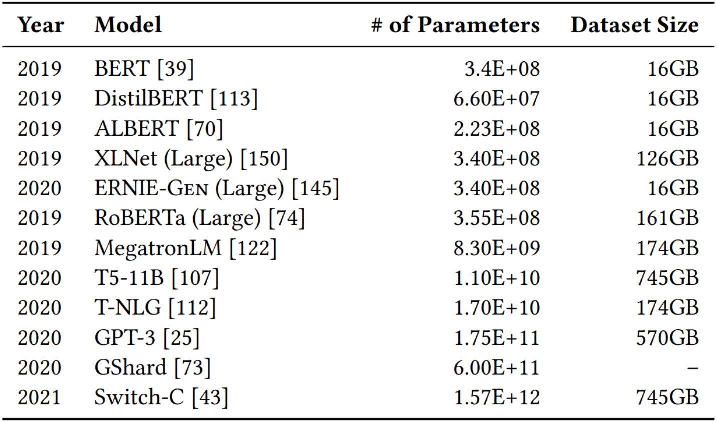
  [SGM] Energy and Policy Considerations for Deep Learning in NLP by Strubell, Ganesh, McCallum

More layers + More data = Better (?)

- One of the biggest trends in ML (specifically in NLP) has been the increasing size of models as measured by the number of parameters and size of training data.
  - BERT and variants, GPT-2, T-NLG, GPT-3, Switch-C
- The amount of compute used to train the largest deep learning models has increased 300,000x in 6 years.
- Are there any potential risks associated with these models?

Large Language Models

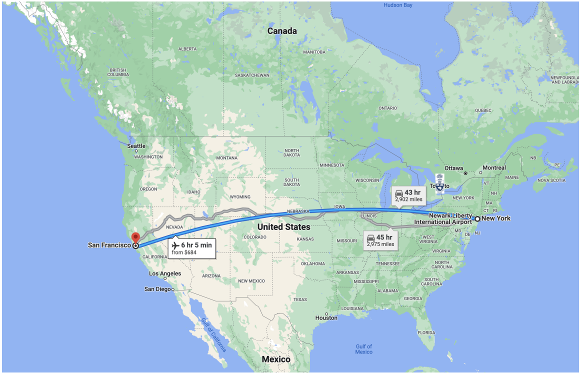
Bender and Gebru, et al.



Year	Model	# of Parameters	Dataset Size
2019	BERT [39]	3.4E+08	16GB
2019	DistilBERT [113]	6.60E+07	16GB
2019	ALBERT [70]	2.23E+08	16GB
2019	XLNet (Large) [150]	3.40E+08	126GB
2020	ERNIE-Gen (Large) [145]	3.40E+08	16GB
2019	RoBERTa (Large) [74]	3.55E+08	161GB
2019	MegatronLM [122]	8.30E+09	174GB
2020	T5-11B [107]	1.10E+10	745GB
2020	T-NLG [112]	1.70E+10	174GB
2020	GPT-3 [25]	1.75E+11	570GB
2020	GShard [73]	6.00E+11	–
2021	Switch-C [43]	1.57E+12	745GB

Table 1: Overview of recent large language models

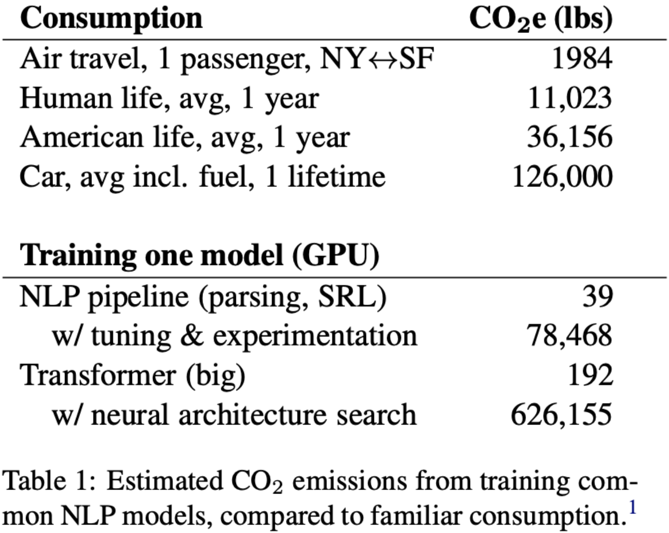
CO2 consumption



Training a BERT model on GPUs was estimated to require as much energy as a trans-American flight.

- roundtrip
- without hyperparameter tuning

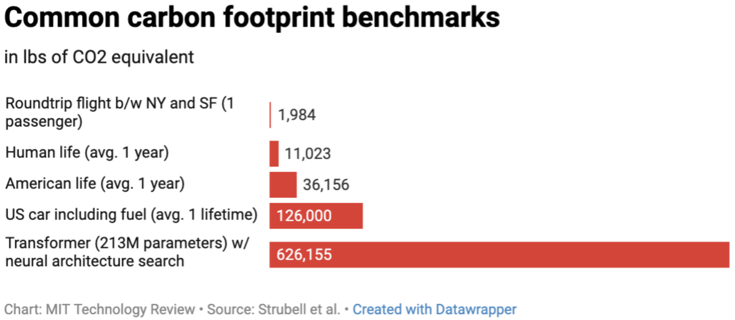
CO2 emissions



Consumption	**CO_2e (lbs)**
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

CO2 emissions



Common carbon footprint benchmarks

in lbs of CO2 equivalent

Chart: MIT Technology Review • Source: Strubell et al. • Created with Datawrapper

- Most of these large language models require hyperparameter tuning, retraining, calibration etc.
- The real environmental impact may be much larger than estimated!

What should we do?

- Time to retrain and sensitivity to hyperparameters should be reported for NLP machine learning models.
- Researchers should prioritize developing efficient models and hardware.
- Different performance metrics should be considered when comparing models.

Financial cost and fair play

- Access to compute has become critical in applied ML research!
- So et al. (2019) report that NAS achieves a new state-of-the-art BLEU score of 29.7 for English to German machine translation.
    - an increase of just 0.1 BLEU compared to the previous state-of-the-art at the cost of at least $150k in on-demand compute time and non-trivial carbon emissions.

Financial cost and fair play

- Most papers from ACL, NeurIPS, and CVPR claim accuracy improvements alone as primary contributions to the field.
- Very little focused on measures of efficiency as primary contributions.
- Optimization research is an exception.
- Access to science has never been fair. But this is adding another dimension to it.
- Academic researchers need equitable access to computational resources.

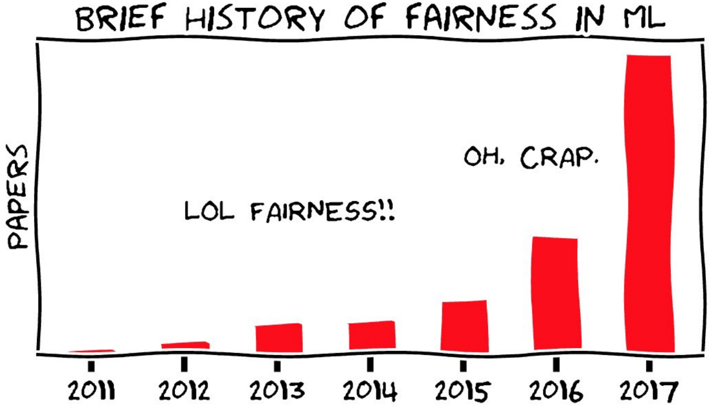
Unfathomable Training Data

- Training data may often have characteristics resulting in ML models that encode stereotypical and derogatory associations
  - along gender, race, ethnicity, and disability.
- Large data sets do not guarantee diversity:
  - the voices of people with a hegemonic viewpoint are overrepresented.
  - models trained on these datasets further amplify biases and harms
- GPT-2's training data is sourced from Reddit:
  - 67% of Reddit users in the USA are men, and 64% between ages 18 and 29.
  - Only 8.8–15% of Wikipedians are women.

A few applications

- As ML starts to be applied to critical applications involving humans, the field is wrestling with the societal impacts
  - **Security:** what if an attacker tries to poison the training data, fool the system with malicious inputs, “steal” the model, etc.?
  - **Privacy:** avoid leaking (much) information about the data the system was trained on (e.g. medical diagnosis)
  - **Fairness:** ensure that the system doesn’t somehow disadvantage particular individuals or groups
  - **Transparency:** be able to understand why one decision was made rather than another
  - **Accountability:** an outside auditor should be able to verify that the system is functioning as intended
- If some of these definitions sound vague, that’s because formalizing them is half the challenge!

Overview: Fairness



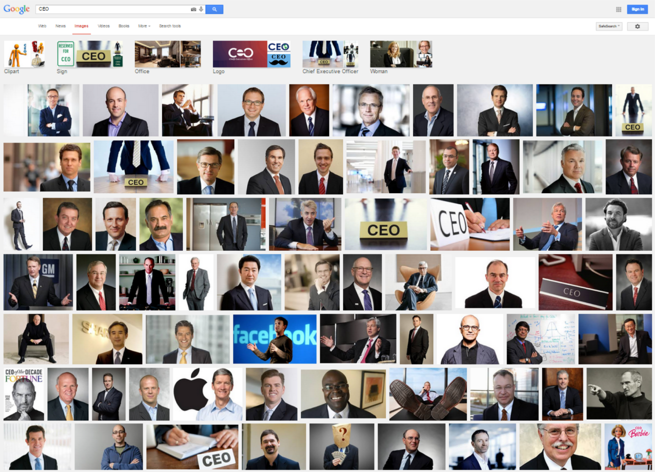
Credit: Moritz Hardt

Overview: Fairness

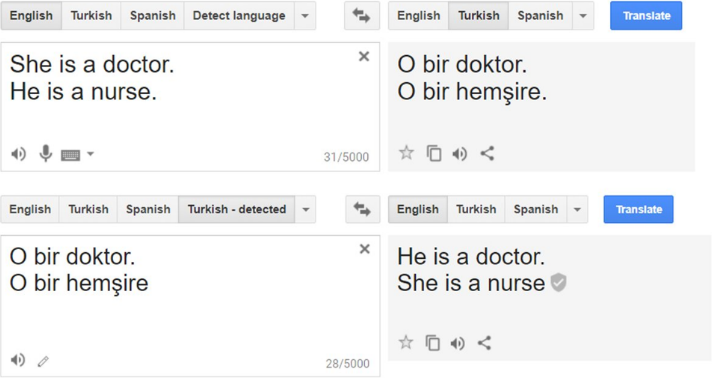
FAIRNESS IN AUTOMATED DECISIONS

Overview: Fairness

SUBTLER BIAS



Overview: Fairness



She is a doctor.
He is a nurse.

O bir doktor.
O bir hemşire.

O bir doktor.
O bir hemşire

He is a doctor.
She is a nurse

Turkish has gender neutral pronouns.

An example from FairML book

- **Self-fulfilling predictions:** Suppose a predictive policing system determines certain areas of a city to be at high risk for crime.
    - More police officers might be deployed to such areas.
    - Officers in areas predicted to be high risk might be subtly lowering their threshold for stopping, searching, or arresting people—perhaps even unconsciously.
    - Either way, the prediction will appear to be validated, even if it had been made purely based on data biases.
- A 2016 paper analyzed a predictive policing algorithm by PredPol.
    - By the Oakland police records, they found that Black people would be targeted by PredPol for drug crimes at roughly twice the rate of whites, even though two groups have roughly equal drug use.
    - This initial bias is amplified by a feedback loop, with policing increasingly concentrated on targeted areas.
    - This is despite the fact that the PredPol algorithm does not explicitly take demographics into account.

Fairness Summary

- Fairness is a challenging issue to address
  - Not something you can just measure on a validation set
  - Philosophers and lawyers have been trying to define it for thousands of years
  - Different notions are incompatible. Need to carefully consider the particular problem.
    - individual vs. group
- Explosion of interest in ML over the last few years
- Conference on Fairness, Accountability, and Transparency (FAT*)
- Textbook: `https://fairmlbook.org/`

Fair ML

- Increasing the environmental and financial costs of large ML models doubly punishes marginalized communities.
  - They are least likely to benefit from the progress achieved.
  - and most likely to be harmed by negative environmental consequences.
- Large datasets based on texts from the Internet overrepresent hegemonic viewpoints and encode biases damaging marginalized populations.

Algorithmic Fairness

- These points shouldn't reduce hype around ML, rather to encourage
  - reproducibility
  - new research that reduce their negative impact
  - models that do not necessarily depend on having large number of parameters
- Instead of focusing on increasing size in ML as the primary driver of increased performance
  - need methods that avoid some of these risks while still keeping the benefits.

Closing Remarks

Continuing with machine learning

- Videos & papers from top ML conferences (NeurIPS, ICML, ICLR, UAI)
- Try to reproduce results from papers
  - If they've released code, you can use that as a guide if you get stuck
- Lots of excellent free resources available online!
- Keep in touch!