ON THE DANGERS OF STOCHASTiC PARROTS: CAN LANGUAGE MODELS BE TOO BiG

Bender et al. 2021

Created by
Anupreet Singh

Sept 16, 2025

WHAT IS A LANGUAGE MODEL?

- Systems trained to predict sequence of words
- Trained on Enormous Text(Common Crawl of Websites)
- Have Many Parameters(Internal Weights) for Capturing more Complex Patterns

OVER THE YEARS

EARLY MODELS- N GRAMS

- Predict the next words based on the previous n words
- Required Massive Text Corpora(Ex-1.8Trillion n-grams for an Machine Translation Models)
- Performance Plateaued(Couldn’t Capture Long-range context)

WORD EMBEDDING

- Representation of Words in Vector Form
- Techniques: Word2Vec, GloVe, context2vec/ELMo
- Advantage: Captured Semantic Similarity(Eg- King-man+woman ≈ Queen
- Eg- For SRL, a model trained with Elmo reached similar F1 score:

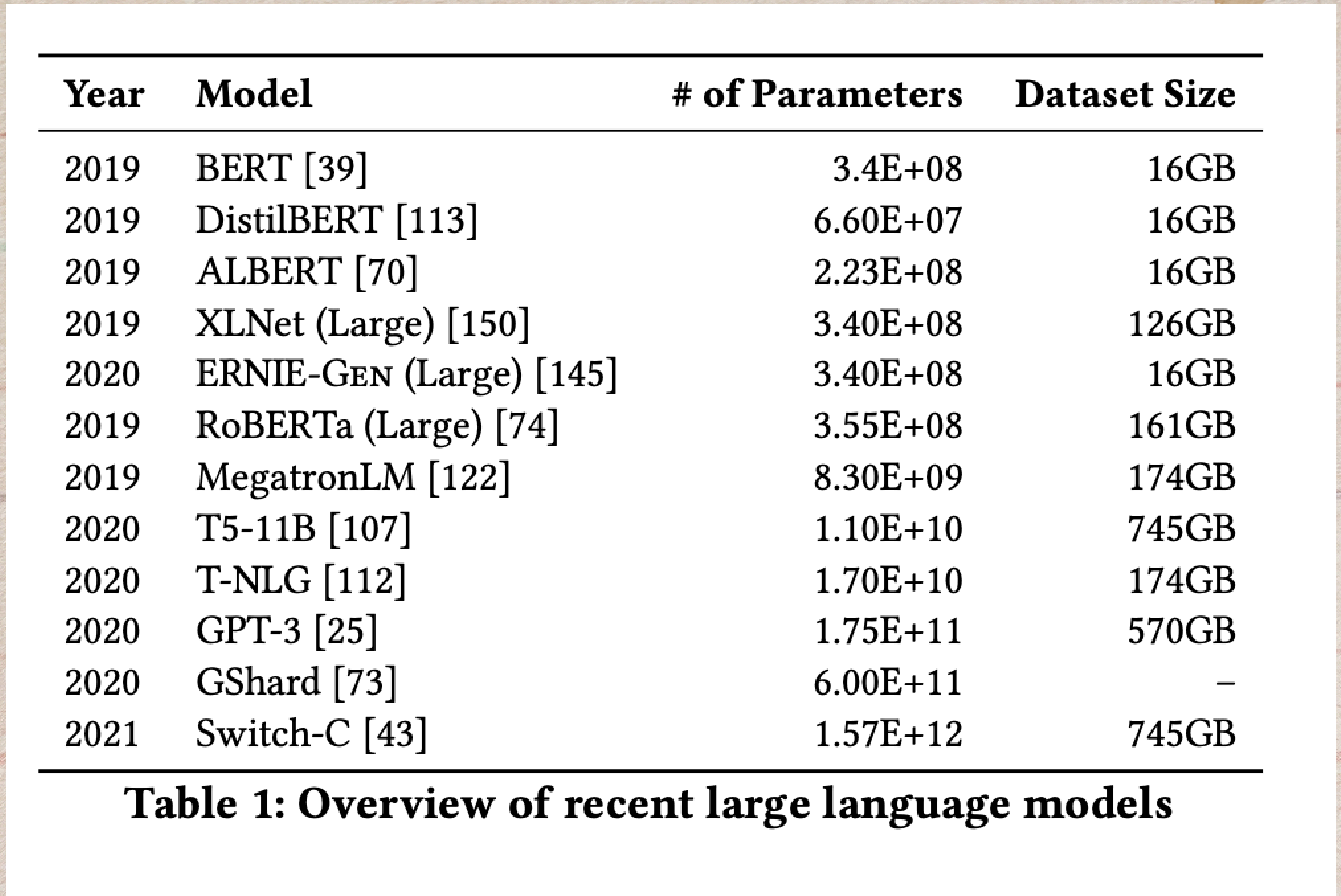
  10 epochs vs 486 epochs

  1% vs 10% of Training data

TRANSFORMERS

- Introduced by Vaswani et al.(2017)
- Used Self-Attention to Capture Contextual Relationship between Tokens
- By Transforming inital Embedding to Key, Query, Value Vectors
- Models like BERT, GPT, T5 achieved big performance jumps by Training on Huge Datasets
- “Bigger is Better” Trend Started

SIZE INCREASE



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

Bender et al 2021

WHAT'S THE PROBLEM?

ENVIRONMENTAL AND FINANCIAL COSTS

- Training an LLM requires massive amounts of Energy
- Training a Basic version of BERT uses as much energy as a Trans-American Flight
- Financial cost of training and Improving is high
- Example: 0.1% improvement in translation Accuracy costs ≈ 150,000 USD

ENVIRONMENTAL AND FINANCIAL COSTS

- Most of the computing Power comes from Non-Renewable sources
- One Estimate suggests training one LLM emits 284 Tons of CO2(5 Times a Human)
- Building ever Large LLMs contributes to carbon emission and Climate Change
-

ENVIRONMENTAL AND FINANCIAL COSTS

- These Costs and Harms aren’t borne equally
- The Environmental Damage hits Marginalized communities the hardest, yet those communities don’t often benefit from the resulting technology
- Example: The people of Maldives and Sudan are hit first by drastic floods, yet their is no LLM produced for Dhivehi and or Sudanese Arabic

TRAINING DATA ISSUES: DIVERSITY

- Huge LLM's are trained on Internet Data(WebCrawl, Wikipedia, Reddit)
- While the Internet is vast, "Size doesn't guarantee Diversity"
- Views of Young Westen Males(Reddit- 67% men, 64% are 18-29 Year old in USA)
- Twitter(Account of people issuing death threats persist, while the ones receiving them are suspended
- Older Adult in US and UK prefer Blogs for Anti-Ageist discourse, but such a niche community is less likely to be found by the crawler

TRAINING DATA ISSUES: STATIC DATA & CHANGE IN SOCIAL VIEWS

- Huge LLM's trained on a snapshot of text risk "value lock"
- Reinforcing less inclusive and outdated norms
- Example: A model trained in 2019 wouldn't have info about Black Lives Matter movement
- So it will align with Existing Regimes of Power
- Retraining trillion-parameter models often enough to keep up with evolving discourse is infeasible

TRAINING DATA ISSUES: ENCODING BIAS

- LMs replicate and even amplify stereotypes and negative associations from training data.
- Example: Disabilities → negative words
- Example: Mental illness → gun violence, homelessness, addiction

TRAINING DATA ISSUES: CURATION AND DOCUMENTATION

- Training on Huge Dataset without Curation encodes Hegemonic views and Harms Marginalized Groups
- Problem - “Documentation Debt”, Dataset so large it becomes infeasible to audit
- Solution: Budgeting for Curation at the start of a project

DO LM'S UNDERSTAND LANGUGAE?

- LLMs achieve high scores on language benchmarks
- Creates impression of genuine progress in Natural Language Understanding (NLU)
- In reality, models may exploit shallow patterns in form rather than grasping meaning
- Results = illusion of advancement → research effort potentially misdirected

STOCHASTIC PARROTS
FLUENT ≠ SENTIENT

- Randomly(but cleverly) stitches together words without understanding
- Because text is Fluent, Humans may impute meaning where there is None
- Could spread misinformation, offend users without “knowing” it
-

STOCHASTIC PARROTS RISKS

- Biased text dissemination(Nurse->Women, Engineer-> Men)
- Invisible harms in systems(Resume Screener might rank women Resume lower for Engineering Jobs)
- Malicious exploitation(Plaigarism)
- Mistranslation consequences(Arabic-Good Morning, English- Attack Him)
-

PATH FORWARD: CONSIDERATIONS

- Ethical, Environmental and Social Implication
- Weigh Energy Efficiency and Costs along with performance
- Data Curation over Quantity
- Document and be Transparent about the models Use

CONCLUSION

- Broaden Research Efforts into Understanding how LLM’s work
- Inclusivity: Other Languages and Marginalized Group
- Beyond “Bigger is Better” Step off the Gas and Think

THANK YOU