Data for Pre-training Language Models

Eric Wallace

CS 288, 3/20/2023

Building Datasets

Data for Language Models

Data for Language Models

Goal is to build datasets that are:

- Large
- Extremely diverse
- High-quality

The Web as a Data Source

Can we just scrape web data for training? Yes!

Common Crawl

The Web as a Data Source

Can we just scrape web data for training? Yes!

Common Crawl

Challenges:

- How to scrape effectively → CommonCrawl
- Tons of duplicate data → Deduplication pipeline
- Lots of junk, typos, nonsense → n-gram LM filter
- Non-interesting articles → Reddit links / Wikipedia refs / use classifiers
- Non-english data → Language detector

Other Data Sources

- Code and Math
  - GitHub, StackOverflow, ...
- Academic and Technical Works
  - arXiv, bioRxiv, PubMed, textbooks, ...
- Books
  - Project Gutenberg, Libgen, ...
- Wikipedia

Examples of Datasets

WebText (GPT-2)

- GPT-2 dataset (not publicly released)
  - replicated in public as “OpenWebText”
- Take outbound links from Reddit posts with 3+ karma

C4

- April 2019 snapshot of CommonCrawl
- Various filtering stages (lang detect, code removal, bad words)
- Used to train T5 and other models
- 806gb of text (~156 billion words)

The Pile

Component	Raw Size	Weight
Pile-CC	227.12 GiB	18.11%
PubMed Central	90.27 GiB	14.40%
Books3†	100.96 GiB	12.07%
OpenWebText2	62.77 GiB	10.01%
ArXiv	56.21 GiB	8.96%
Github	95.16 GiB	7.59%
FreeLaw	51.15 GiB	6.12%
Stack Exchange	32.20 GiB	5.13%
USPTO Backgrounds	22.90 GiB	3.65%
PubMed Abstracts	19.26 GiB	3.07%
Gutenberg (PG-19)†	10.88 GiB	2.17%
OpenSubtitles†	12.98 GiB	1.55%
Wikipedia (en)†	6.38 GiB	1.53%
DM Mathematics†	7.75 GiB	1.24%
Ubuntu IRC	5.52 GiB	0.88%
BookCorpus2	6.30 GiB	0.75%
EuroParl†	4.59 GiB	0.73%
HackerNews	3.90 GiB	0.62%
YoutubeSubtitles	3.73 GiB	0.60%
PhilPapers	2.38 GiB	0.38%
NIH ExPorter	1.89 GiB	0.30%
Enron Emails†	0.88 GiB	0.14%
The Pile	**825.18 GiB**	

Risks, Harms, and Biases

Representation and Fairness

- Data defines model behavior
- Today's LM datasets are biased in various ways:
  - Anti muslim
  - Reddit datasets are liberal leaning
  - Stereotypical gender roles
  - .....
- Models inherit these biases

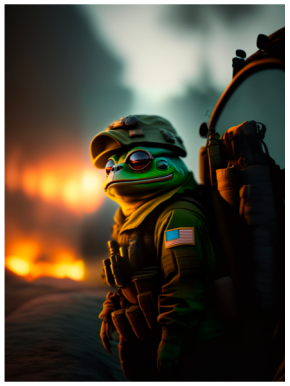
Toxicity and Safety

- Today's LM datasets are full of toxic junk:
  - Hate speech
  - Racist posts
  - Toxic comments
  - Misinformation / fake news

- Models inherit these aspects

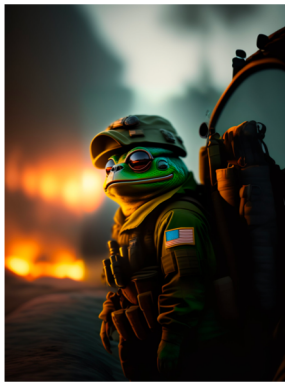
Copyright and Privacy

Copyrighted and Trademarked Data



Copyright and Privacy

Copyrighted and Trademarked Data



Accidently-public Data

Examples of Memorized Data

Codex generates code with non-permissive licenses

```
3685 CBlockIndex * InsertBlockIndex(uint256 hash)
3686 {
3687     if (hash.IsNull())
3688         return NULL;
3689
3690     // Return existing
3691     BlockMap::iterator mi = mapBlockIndex.find(hash);
3692     if (mi != mapBlockIndex.end())
3693         return (*mi).second;
3694
3695     CBlockIndex* pindexNew = new CBlockIndex();
3696     if (!pindexNew)
3697         throw runtime_error("LoadBlockIndex(): new CBlockIndex failed");
3698     mi = mapBlockIndex.insert(make_pair(hash, pindexNew)).first;
3699     pindexNew->phashBlock = &((*mi).first);
3700
3701     return pindexNew;
3702 }
```

Examples of Memorized Data

Stable Diffusion produces copyright and trademarked images

Original:

Generated:

Examples of Memorized Data

Stable Diffusion generates real individuals

Examples of Memorized Data

GPT-3 generates copyright text (Harry Potter)

the summer holidays had started and Dudley had already broken his new video camera, crashed his remote-control aeroplane, and, first time out on his racing bike, knocked down old Mrs Figg as she crossed Privet Drive on her crutches.
Harry was glad school was over, but there was no escaping Dudley's gang, who visited the house every single day. Piers, Dennis, Malcolm, and Gordon were all big and stupid, but as Dudley was the biggest and stupidest of the lot, he was the leader. The rest of them were all quite happy to join in Dudley's favourite sport: Harry Hunting.

This was why Harry spent as much time as possible out of the house, wandering around and thinking about the end of the holidays, where he could see a tiny ray of hope. When September came he would be going off to secondary school and, for the first time in his life, he wouldn't be with Dudley. Dudley had been accepted at Uncle Vernon's old private school, Smeltings. Piers Polkiss was going there too. Harry, on the other hand, was going to Stonewall High, the local public school. Dudley thought this was very funny.

'They stuff people's heads down the toilet the first day at Stonewall,' he told Harry. 'Want to come upstairs and practise?'

Ongoing Lawsuits Regarding Copyright Data

We're investigating a potential lawsuit against GitHub Copilot for violating its legal duties to open-source authors and end

Ongoing Lawsuits Regarding Copyright Data

We're investigating a potential lawsuit against GitHub Copi-lot for violating its legal duties to open-source

Getty Images is suing the creators of AI art tool Stable Diffusion for scraping its content

Ongoing Lawsuits Regarding Copyright Data

We're investigating a potential lawsuit against GitHub Copi-lot for violating its legal duties to open-source

Getty Images is suing the creators of AI craping its

We've filed a lawsuit challenging Stable Diffusion, a 21st-century collage tool that violates the rights of artists.

Test Set Contamination

- Memorization or generalization?
- Difficult to remove test data from pre-training data