EXHIBIT W

danny Jun 8, 2021 3:28 PM | Edited on Jun 8, 2021 3:48 PM

Some initial experiments on text data quality/transfer. I thought it was strange that on our current single best eval metric: language-combined-evals training on just pile and books3 did surprisingly well and training on just enwikipedia did openwebtext2 did surprisingly bad. How different is pile from our canonical datasets?

These are all 8-layer models trained on one distribution and evaluated on several others.

Models aren't fully trained yet.

Deleted on 06/08/2024 1:17 AM

Screen Shot 2021-06-08 260 KB

danny Replied on Jun 8, 2021 3:29 PM

language-combined-evals description
"Raw evaluation data we're using as a quick proxy for performance, combining "
"hellaswag, triviaqa, squad, winogrande, lambada, openbookqa, and arc [redacted]

danny Replied on Jun 8, 2021 3:30 PM

Maybe it's a formatting issue with wikipedia, or maybe wikipedia is very helpful mixed in, but not great to train on by itself because of sufficiently different formatting or something

danny Replied on Jun 8, 2021 3:32 PM

I was also surprised that books3 had much better transfer to python than webtext or wikipedia

Deleted on 06/08/2024 1:17 AM

Screen Shot 2021-06-08 284 KB

thinking_face 3

danny Replied on Jun 8, 2021 3:35 PM | Edited on Jun 8, 2021 3:37 PM

More excited for libgen after seeing how well books3 is doing in some ways.
I think it'd be helpful to split fiction and non-fiction, so I can study the quality/transfer of those two datasets seperately. They feel quite different.

2 2

danny Replied on Jun 8, 2021 3:40 PM

Another view on wikipedia and webtext not doing great

Deleted on 06/08/2024 1:17 AM

Screen Shot 2021-06-08 245 KB

danny Replied on Jun 8, 2021 3:48 PM

Next experiments I plan on doing are

1. A few mixtures. I wonder how much of the benefit for pile and canonical-8k are coming from having at least some small fraction of openwebtext2, books3, and enwikipedia.
2. Size scan.

@jared @ben

jared Replied on Jun 8, 2021 4:05 PM

Super interesting!! Thanks for looking at this.

jared Replied on Jun 8, 2021 4:10 PM

I don't have a strong prior belief here. I have a personal preference for books so I'm amused that they're doing well. The pile has a lot of academic papers and stuff in it as well. You can look up the paper to see the exact mixture.

danny Replied on Jun 8, 2021 4:17 PM | Edited on Jun 8, 2021 4:18 PM

Yeah, I was surprised by how academic the text in the pile is.

Deleted on 06/08/2024 1:17 AM



danny Replied on Jun 8, 2021 4:23 PM

It looks like canonical-8k is [redacted]books, [redacted]webtext[redacted]wikipedia, [redacted]pile. So makes sense why pile and canonical are so similar.

[redacted]

```
# Make sure we only get 1 copy of each of these.
["khan-math", 1e-8],
["hendrycks-math", 1e-8],
]
PILE = [
["mix", 100],
# Make sure we only get 1 copy of each of these.
["khan-math", 1e-8],
["hendrycks-math", 1e-8],
] https://github.com/anthropics[redacted]
```

+1 1

ben Replied on Jun 8, 2021 4:27 PM | Edited on Jun 8, 2021 4:28 PM

Are there any stats on what's in books3? Very curious why it's doing better on python. Could be worth just looking at 1000 shuffled titles and getting a rough feel. Maybe it's just line wrapping and indentation matching 😅

danny Replied on Jun 8, 2021 4:38 PM

I haven't found stats, but from what I can tell books3 is all of bibliotik which seems like a private torrent network around booklike things. in plain text form. I think it includes magazines, textbooks, and niche text people on the internet would want...

People on the internet like programming textbooks, so that's my guess at what makes books3 good at programming.

https://github.com/soskek/bookcorpus/issues/27#issuecomment-716104208

danny Replied on Jun 8, 2021 4:42 PM

It seems like the-eye is a group that works with the pile

https://the-eye.eu/public/Books/

I think these directories give you some idea of the kind of distribution books3 is, but don't know anything about the relative size.

There are some big top level categories like:

Poetry, SciFi_fantasy, History, Manuals, Medical

But there's also some much more specific top level categories that 4chan_pol_Archives, Asimov, Playboy, Radical Feminist Literature

B **ben** Replied on Jun 8, 2021 5:04 PM

Yea I think just glancing at the titles will be more informative for that reason. I can get you the bash command after I land in sf

+1 2

J **jared** Replied on Jun 9, 2021 12:33 PM

It's interesting that the curves cross here, rather than maintaining the same rankings throughout

D **danny** Replied on Jun 9, 2021 1:34 PM | Edited on Jun 9, 2021 1:34 PM

Yeah, the crossings of various curves makes me less sure what to expect when scaled up to larger models.

1