CSE599J: Data-centric ML

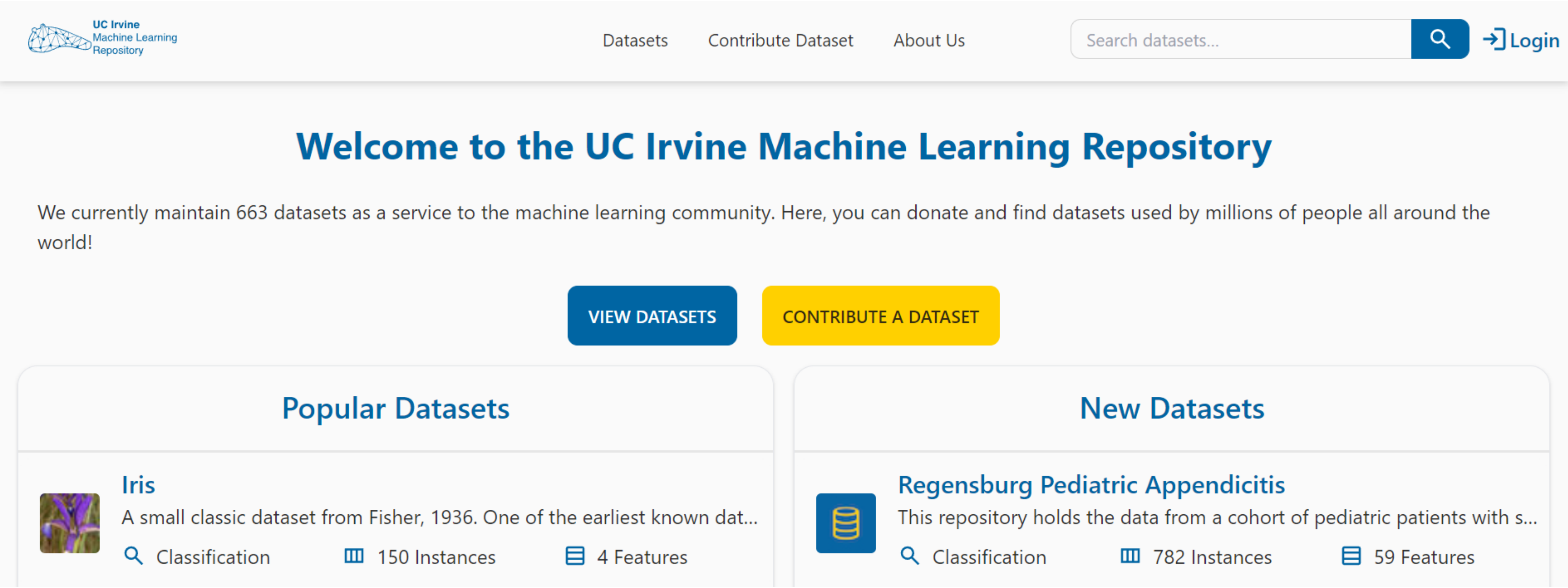
Jan 5, 2024

Pang Wei Koh
Akari Asai

[Husky images are from DALL-E.
Spelling errors are their fault.]

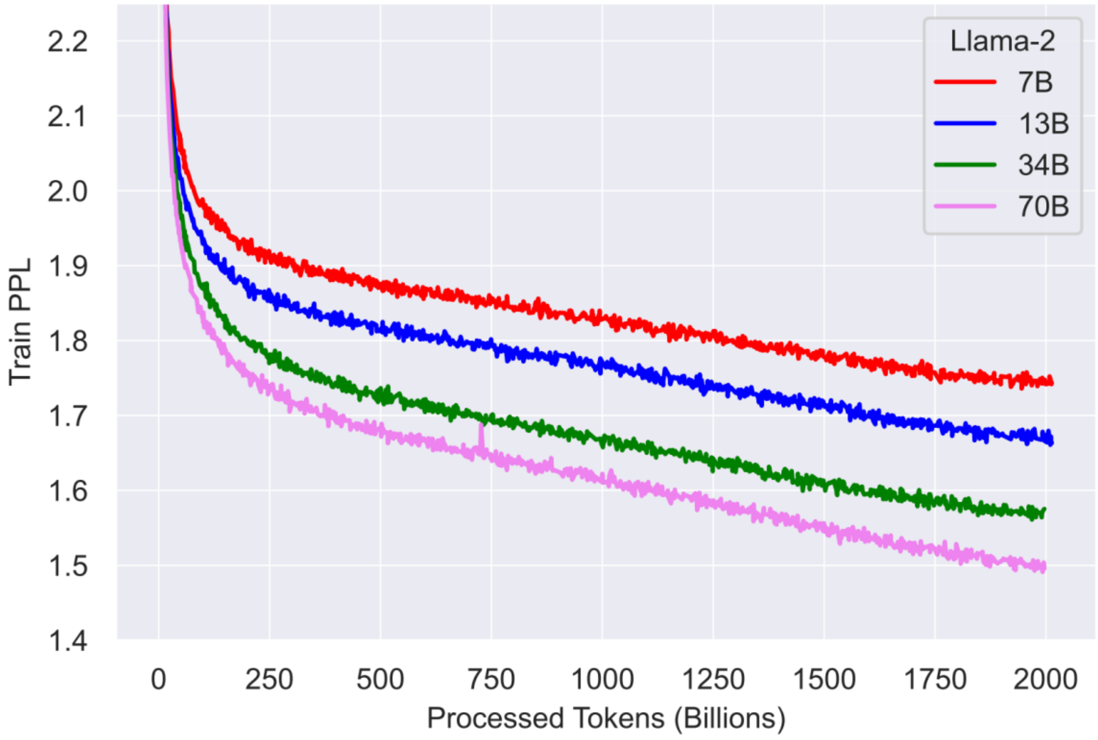
Why “data-centric ML”?

- Contrast with model-centric ML: fixed datasets, focus on models



Why "data-centric ML"?

- Contrast with model-centric ML: fixed datasets, focus on models
- Recent progress in ML has been powered by data



[Touvron et al., Llama 2]

Why “data-centric ML”?

- Contrast with model-centric ML: fixed datasets, focus on models
- Recent progress in ML has been powered by data
- What’s different now?
    - Scale of data
    - Sources of data (mostly scraped)
    - General purpose: more domains, more tasks
    - Generation vs. discrimination
- This class: survey of data-related topics in ML, biased towards recent papers

What we’re covering

1. Data for pretraining
   - Dataset construction
   - Scaling laws
   - Data filtering
   - Dataset composition
   - Biases in datasets
2. Data for tuning and evaluation
   - Generative evaluation
   - Data for alignment
   - Ambiguity and disagreement
3. Adapting to different distributions
   - Distribution shifts
   - Reweighting data
   - Domain adaptation
4. Linking model output to training data
   - Data attribution
   - Retrieval-based models
   - Memorization
5. Legal, ethical, and security issues
   - Copyright
   - Segregating data
   - Data security and robustness

What we’re not covering

- Dataset distillation
- Active learning
- Weak supervision
- Data valuation
- Everything else

Goals of this course

- Broad exposure to data-related topics
- Practice reading and discussing papers (reflections)
- Practice synthesizing and contextualizing research (presentations)
- More in-depth, hands-on experience in one topic (class project)

Format

- Each class has 2 assigned papers:
  - Everyone writes short reflections, due the day before class
- Each class has 2 assigned presenters:
  - Submit slides (with bibliography), due two days before class
- Course project:
  - Teams of up to 3 students
  - Proposal, due February 2
  - Project presentation, in class on March 8
  - Writeup, due March 11

Paper reflections

- Read each paper before class and answer 5 questions per paper
  - You can keep responses brief (1-3 sentences)
  - Pass/fail grading
- We’ll use these as a basis for discussion
- As you’re reading…
  - Be skeptical – don’t take at face value
  - But be open-minded and look for constructive takeaways

Paper presentations

- Assume everyone has read the papers
- 20-30min presentation:
  - Contextualize the work (by reading other papers; include bib)
  - Do deep dives into one key aspect of each paper
  - Be interactive!
  - Present these jointly or separately; papers can be in either order
  - Presenters should work together; each should speak ~50%
- Some students will be assigned to present twice
- Submit slides in advance. After presentation, we'll upload slides.
- Submit the form by end of today

Project

- Teams of up to 3 students
- Related to “data-centric ML”
- Does <u>not</u> need to be publishable research (yet)
- Think of it as “what would be an interesting blog post?”
- We will be liberal with grading
- Final writeup should <u>not</u> just be a proposal
- Proposal is due February 2 – let us know by end of Jan 12 if you want help finding teammates

Grading

- Papers (50%)
  - Class participation (17%)
  - Paper reflections (17%)
  - Paper presentation (16%)

- Project (50%)
  - Proposal (5%)
  - Project presentation (10%)
  - Writeup (35%)

Discussion etiquette

- Be candid but professional
- Talk about the paper not the authors
  - Imagine the authors are in the room (they well might be) and you're giving honest feedback on a paper draft
- Keep discussions safe
  - Outside of class, you can talk about the substance, but
  - Nothing we discuss in class should be attributed to anyone

Introductions

- Name, year, advisors (if applicable), what you’re interested in

Today’s topic: Dataset construction

Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus

Jesse Dodge, Maarten Sap, Ana Marasovic, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, Matt Gardner

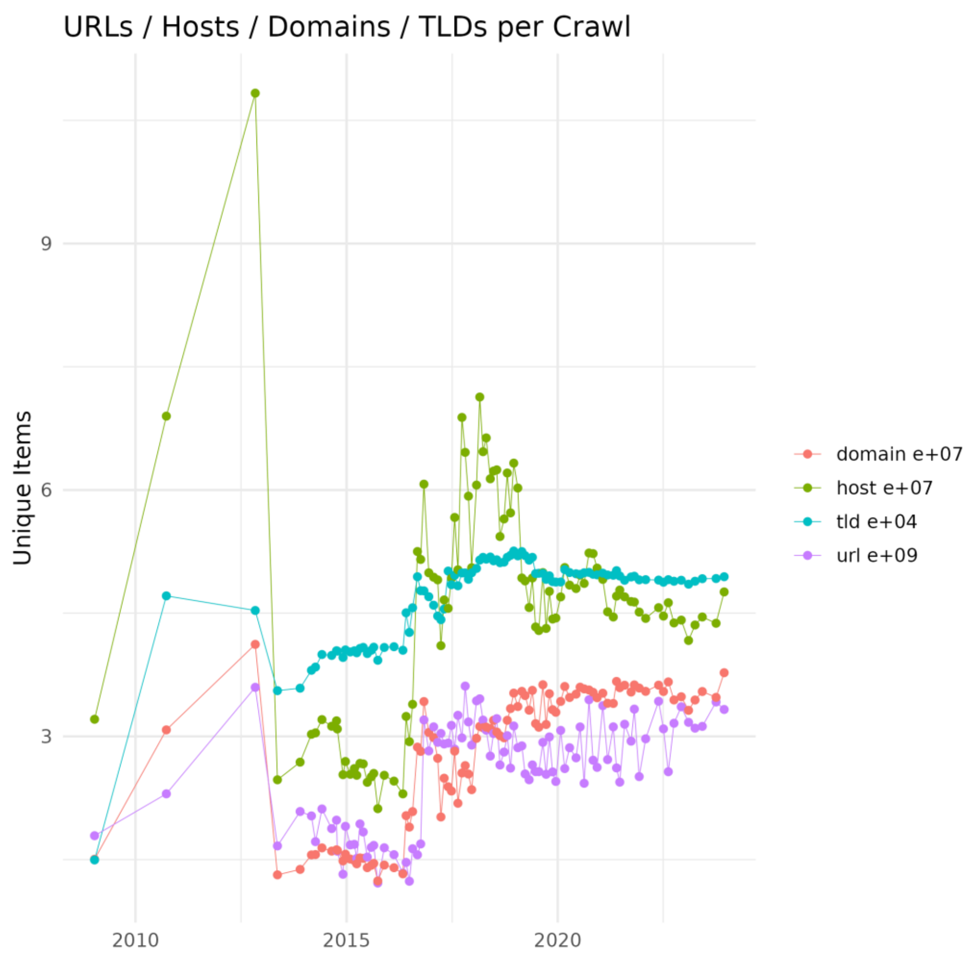
A brief history of web text datasets

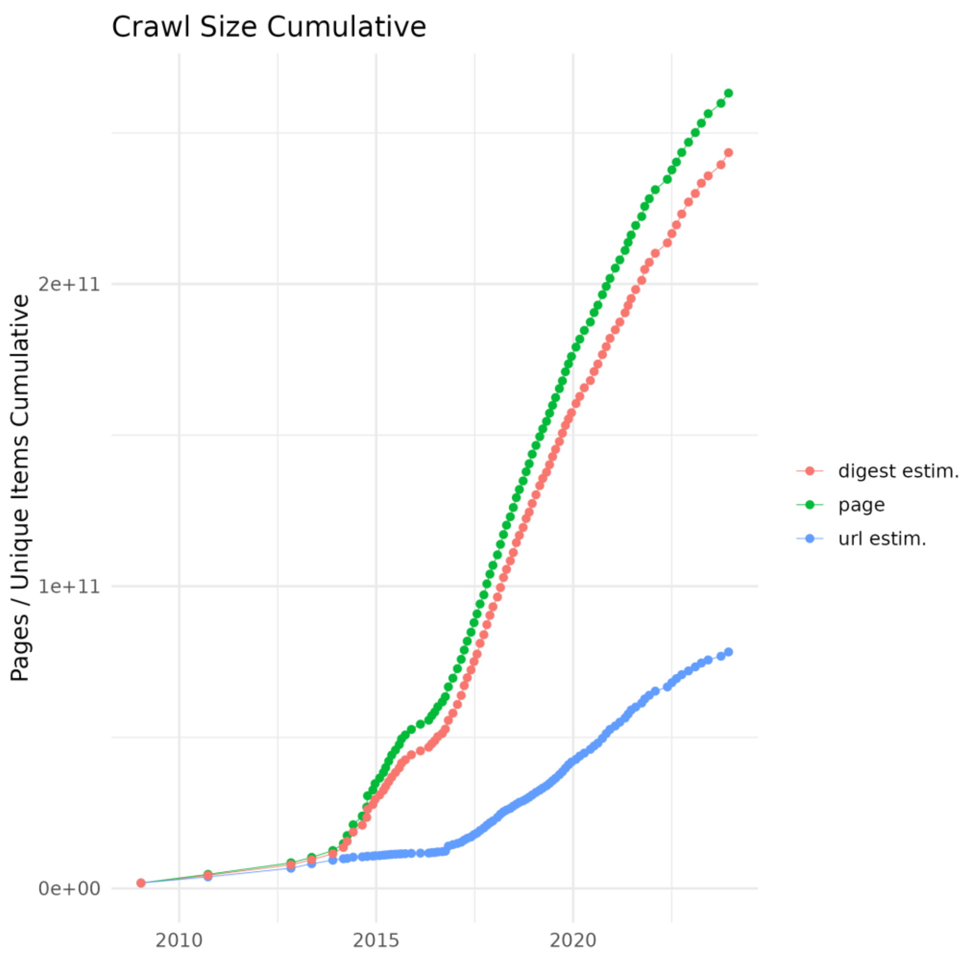
- Transformers (Vaswani et al., 2017)
  - WMT 2014 English-French dataset: 36M sentences
- BERT (Devlin et al., 2018)
  - BooksCorpus (800M words) + English Wiki (2,500M words, ~6M pages)
- GPT-2 (Radford et al., 2019)
  - WebText (8M docs): all outbound links from Reddit with 3+ karma
- C4 / T5 (Raffel et al., 2020)
  - Common Crawl web scrape

Data set	Size	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ C4	745GB	83.28	**19.24**	80.88	71.36	**26.98**	**39.82**	**27.65**
C4, unfiltered	6.1TB	81.46	19.14	78.78	68.04	26.55	39.34	27.21
RealNews-like	35GB	**83.83**	**19.23**	80.39	72.38	**26.75**	**39.90**	**27.48**
WebText-like	17GB	**84.03**	**19.31**	**81.42**	71.40	**26.80**	**39.74**	**27.59**
Wikipedia	16GB	81.85	**19.31**	81.29	68.01	**26.94**	39.69	**27.67**
Wikipedia + TBC	20GB	83.65	**19.28**	**82.08**	**73.24**	**26.77**	39.63	**27.57**

Table 8: Performance resulting from pre-training on different data sets. The first four variants are based on our new C4 data set.

The scale of the Common Crawl





[https://commoncrawl.github.io/cc-crawl-statistics/plots/crawlsize]

C4 (Colossal Clean Crawled Corpus)

- The C4 / T5 (Raffel et al., 2020) paper:
  - Introduced C4
  - Analyzed model design choices -> T5
  - Studied filtering C4 (see previous table)
  - Didn't include downloadable link
  - What really was included in C4?

language, placeholder text, source code, etc.). To address these issues, we used the following heuristics for cleaning up Common Crawl's web extracted text:

- We only retained lines that ended in a terminal punctuation mark (i.e. a period, exclamation mark, question mark, or end quotation mark).
- We discarded any page with fewer than 5 sentences and only retained lines that contained at least 3 words.
- We removed any page that contained any word on the "List of Dirty, Naughty, Obscene or Otherwise Bad Words".[6]
- Many of the scraped pages contained warnings stating that Javascript should be enabled so we removed any line with the word Javascript.
- Some pages had placeholder "lorem ipsum" text; we removed any page where the phrase "lorem ipsum" appeared.
- Some pages inadvertently contained code. Since the curly bracket "{" appears in many programming languages (such as Javascript, widely used on the web) but not in natural text, we removed any pages that contained a curly bracket.
- To deduplicate the data set, we discarded all but one of any three-sentence span occurring more than once in the data set.

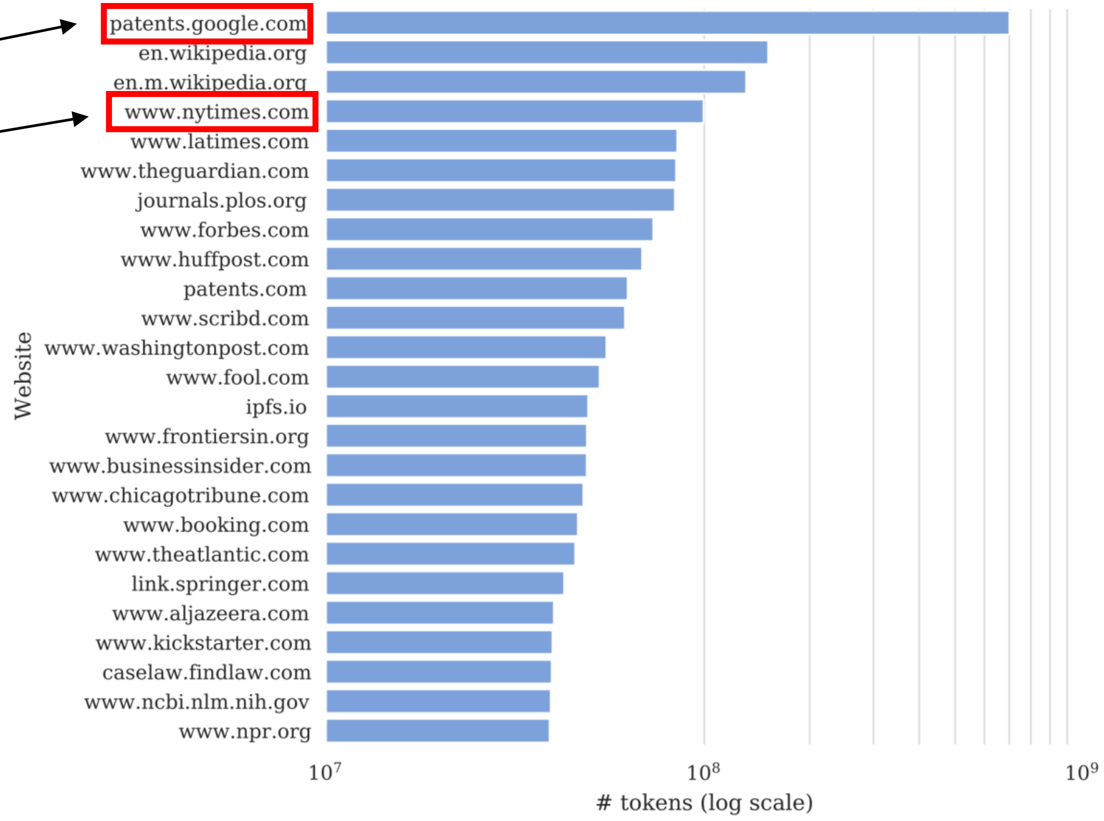
Additionally, since most of our downstream tasks are focused on English-language text, we used `langdetect`[7] to filter out any pages that were not classified as English with a probability of at least 0.99. Our heuristics are inspired by past work on using Common

Documenting Large Webtext Corpora (Dodge et al., 2021)

- One of the first efforts to study the composition of web text data
- Sources?

~10% machine-translated

ongoing lawsuit



Documenting Large Webtext Corpora (Dodge et al., 2021)

- Removing “offensive language”
  - In practice, list is mostly related to sexual/lewd content, not toxicity
  - Majority of excluded docs relate to science, medicine, legal, etc. topics?
  - Disproportionate effect on mentions of sexual orientation

README CC-BY-4.0 license

Our List of Dirty, Naughty, Obscene, and Otherwise Bad Words

With millions of images in our library and billions of user-submitted keywords, we work hard at Shutterstock to make sure that bad words don't show up in places they shouldn't. This repo contains a list of words that we use to filter results from our autocomplete server and recommendation engine.

Please add to it as you see fit (particularly in non-English languages) or use it to spice up your next game of Scrabble :)

DataComp: In search of the next generation of multimodal datasets

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, Ludwig Schmidt

A brief history of image-text datasets

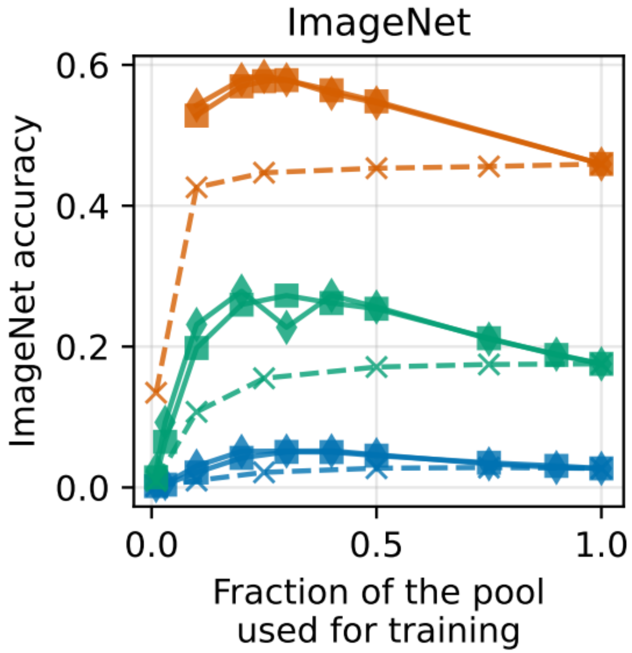
- Classically: labeled image datasets like ImageNet
- Conceptual Captions (Sharma et al., 2018)
  - 3.3M image-text pairs scraped from web (images + alt-text)
- ConVIRT (Zhang et al., 2020)
  - ~250k image-text pairs of chest (MIMIC-CXR) and bone (private) x-rays
- CLIP / WIT400M (Radford et al., 2021)
  - 400M image-text pairs, web search with hand-crafted queries
- ALIGN (Jia et al., 2021)
  - 1.8B unfiltered version of Conceptual Captions
- LAION-5B (Schuhmann et al., 2022)
  - 5.9B image-text pairs from Common Crawl, filtered with CLIP

What’s missing?

- Clear that training datasets matter a lot for performance
- But no controlled experiments on dataset construction
- Prior works change many factors at once: different datasets, architectures, training objectives, evals, …

DataComp (Gadre et al., 2023)

- Goal: Facilitate systematic experimentation on training datasets
- Dataset size is a design choice -> Scaling laws (Jan 10)

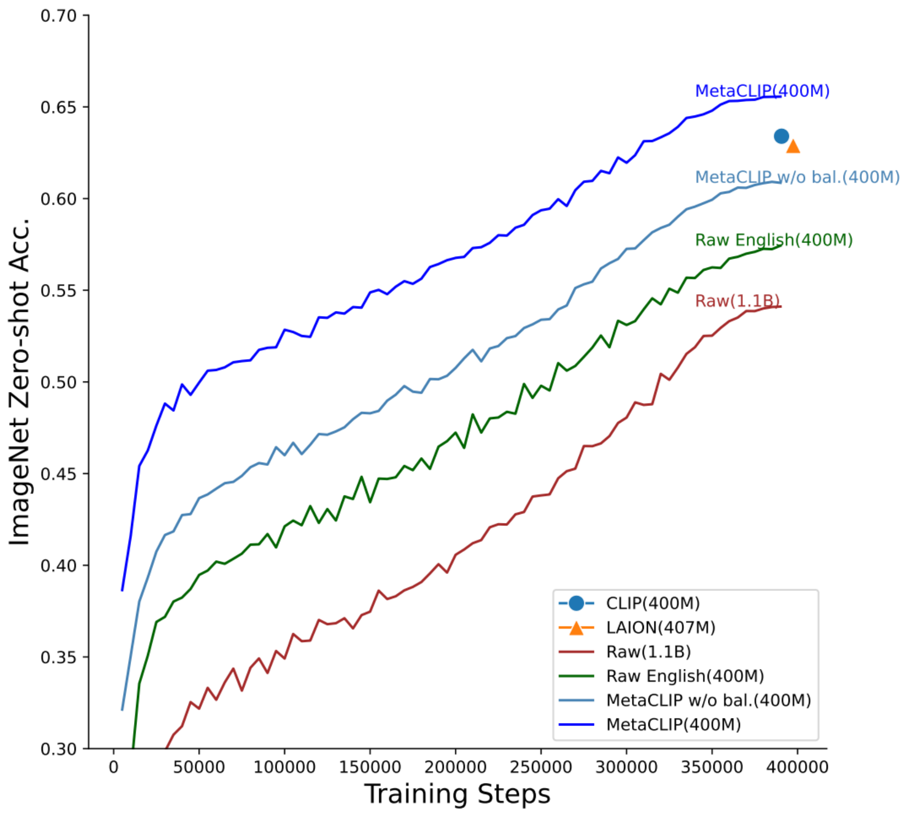


DataComp (Gadre et al., 2023)

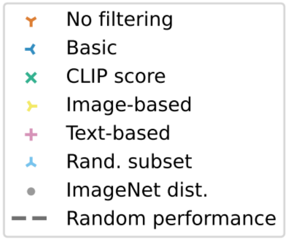
- Goal: Facilitate systematic experimentation on training datasets
- Dataset size is a design choice -> Scaling laws (Jan 10)
- CLIP and image filtering -> Data filtering (Jan 12) + dataset composition (Jan 17)
- Effects on fairness & representation -> Biases in datasets (Jan 19)

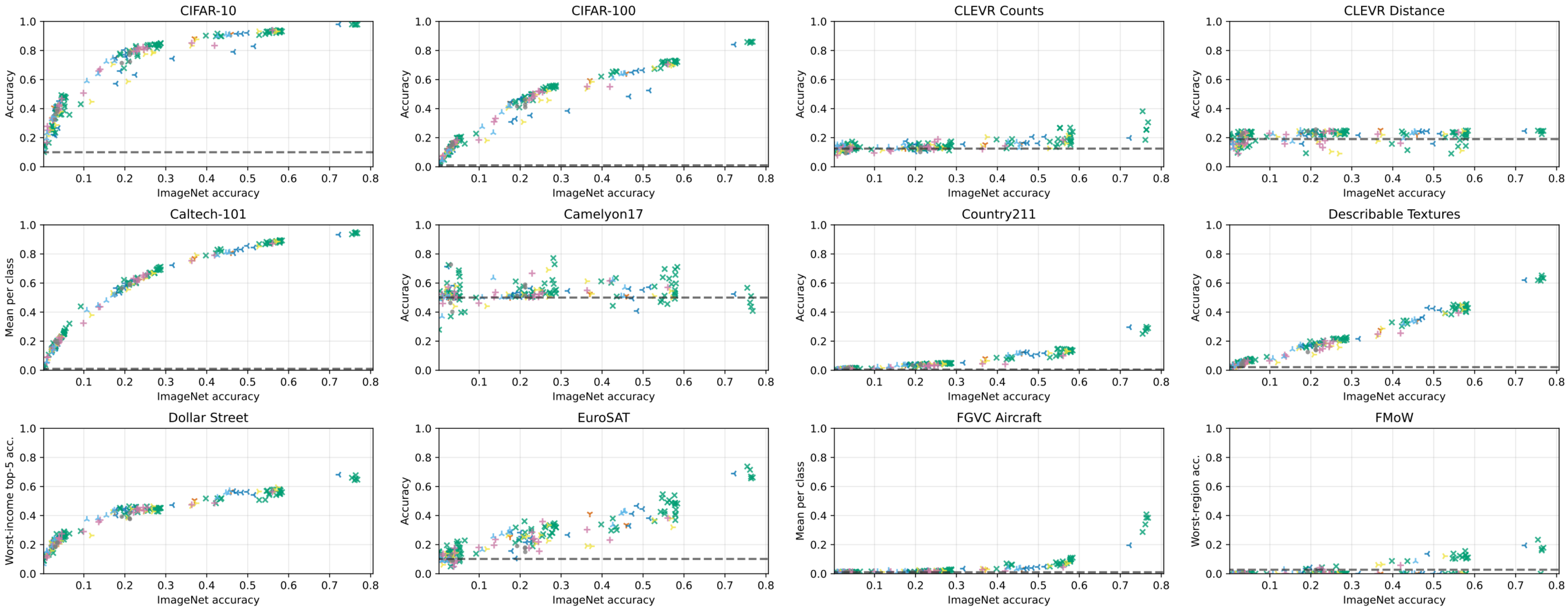
CLIP filtering

- MetaCLIP (Xu et al., 2023) set out to replicate the proprietary WIT dataset behind CLIP, *without* using CLIP filtering:



Non-ImageNet evals





Multilingual evals (h/t Gabriel)

- The best performing filters in DataComp were CLIP filtering ∩ image-based filtering
  - The English filtering used in LAION-2B hurt performance
- A contributor (Alex Visheratin) evaluated DataComp models on multilingual image retrieval datasets Crossmodal-3600 & XTD10
- Best-performing model was the one trained on unfiltered data
- None of the DataComp filters were explicitly English-centric

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
3. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.
4. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 21(1), 5485-5551.
5. Sharma, P., Ding, N., Goodman, S., & Soricut, R. (2018, July). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 2556-2565).
6. Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., & Langlotz, C. P. (2022, December). Contrastive learning of medical visual representations from paired images and text. In Machine Learning for Healthcare Conference (pp. 2-25). PMLR.
7. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.
8. Jia, C., Yang, Y., Xia, Y., Chen, Y. T., Parekh, Z., Pham, H., ... & Duerig, T. (2021, July). Scaling up visual and vision-language representation learning with noisy text supervision. In International conference on machine learning (pp. 4904-4916). PMLR.
9. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., ... & Jitsev, J. (2022). LAION-5B: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems, 35, 25278-25294.
10. Xu, H., Xie, S., Tan, X. E., Huang, P. Y., Howes, R., Sharma, V., ... & Feichtenhofer, C. (2023). Demystifying CLIP data. arXiv preprint arXiv:2309.16671.