TiC-CLIP: Continual Training of CLIP

Saurabh Garg[*] Mehrdad Farajtabar[†] Hadi Pouransari[†] Raviteja Vemulapalli[†]
Sachin Mehta[†] Oncel Tuzel[†] Vaishaal Shankar[†] Fartash Faghri[†]

[*]Carnegie Mellon University [†]Apple

[*]work done during an internship at Apple

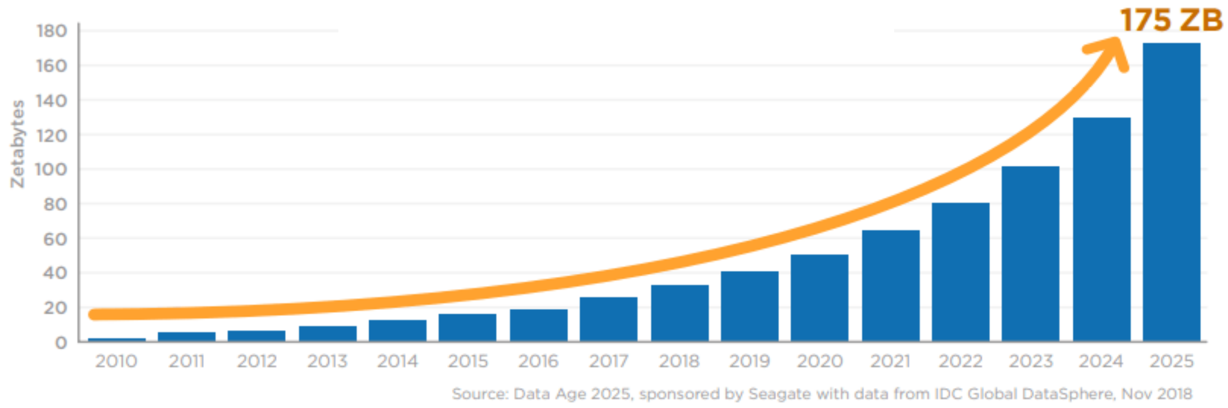
Training Large Scale Foundation Models is Expensive

- Multimodal models, e.g., CLIP, are trained at scale of several billion image-text pair data collected over 5-8 years

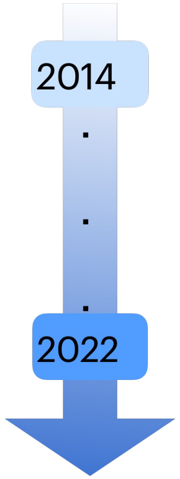
- Open CLIP ViT-G-14 model was trained for 240k A100 GPU hours which is approximately one month on 400 GPUs Schuhmann et al. (2022)

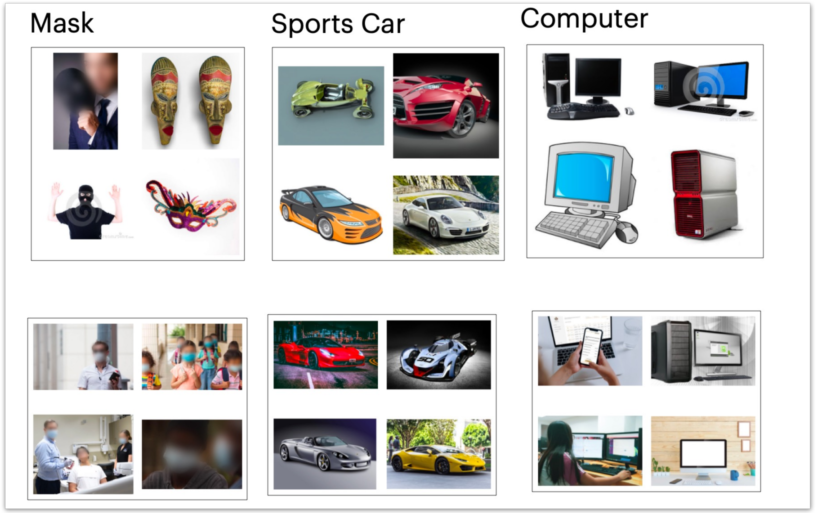
Schuhmann et al. Laion-5b: An open large-scale dataset for training next generation image-text models (2022)

But data is continuously increasing and evolving



#webpages continuously increases





Concepts evolve over time

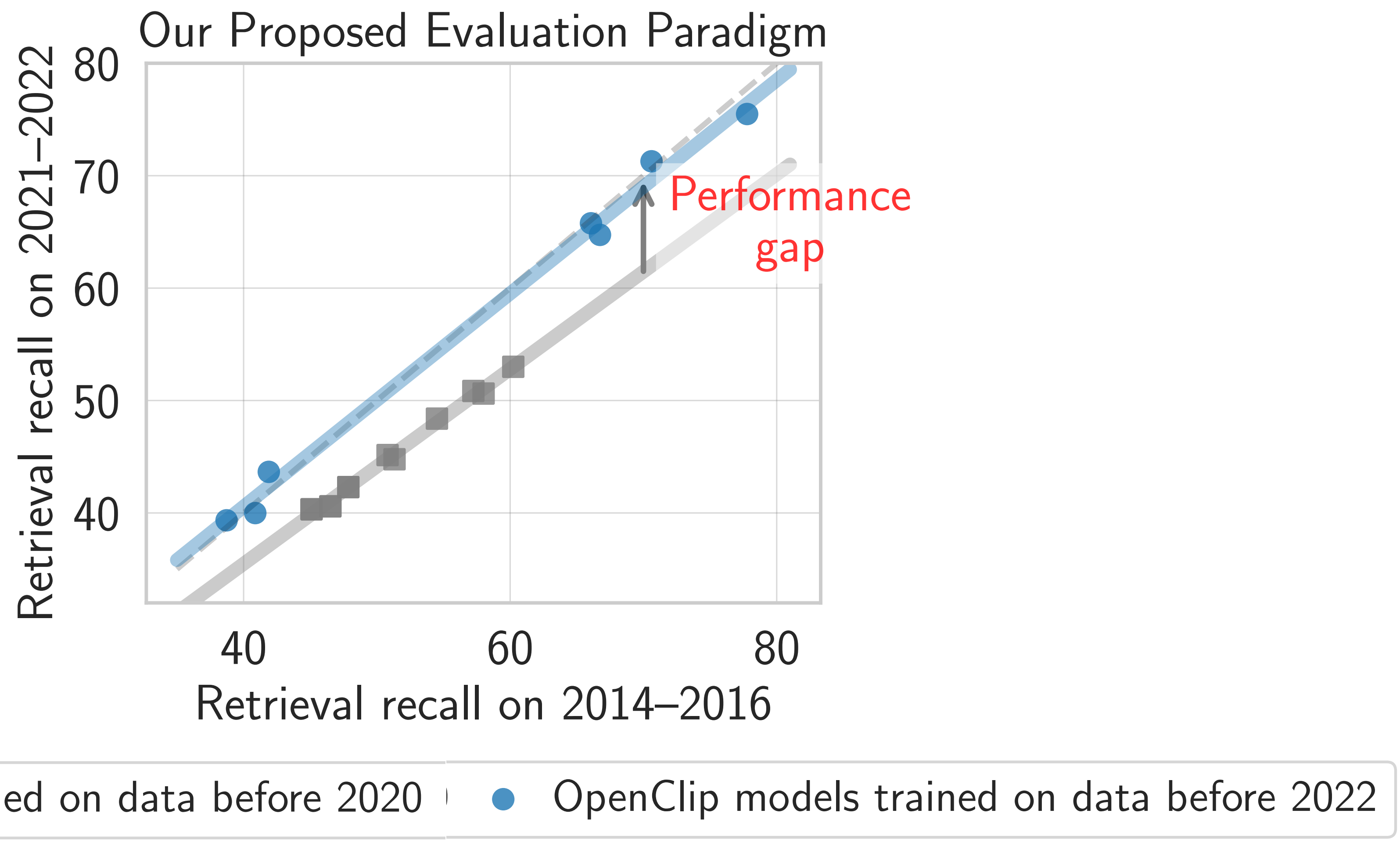
Training from scratch is not computationally feasible

How to continuously update these models as data distributions evolves over time?

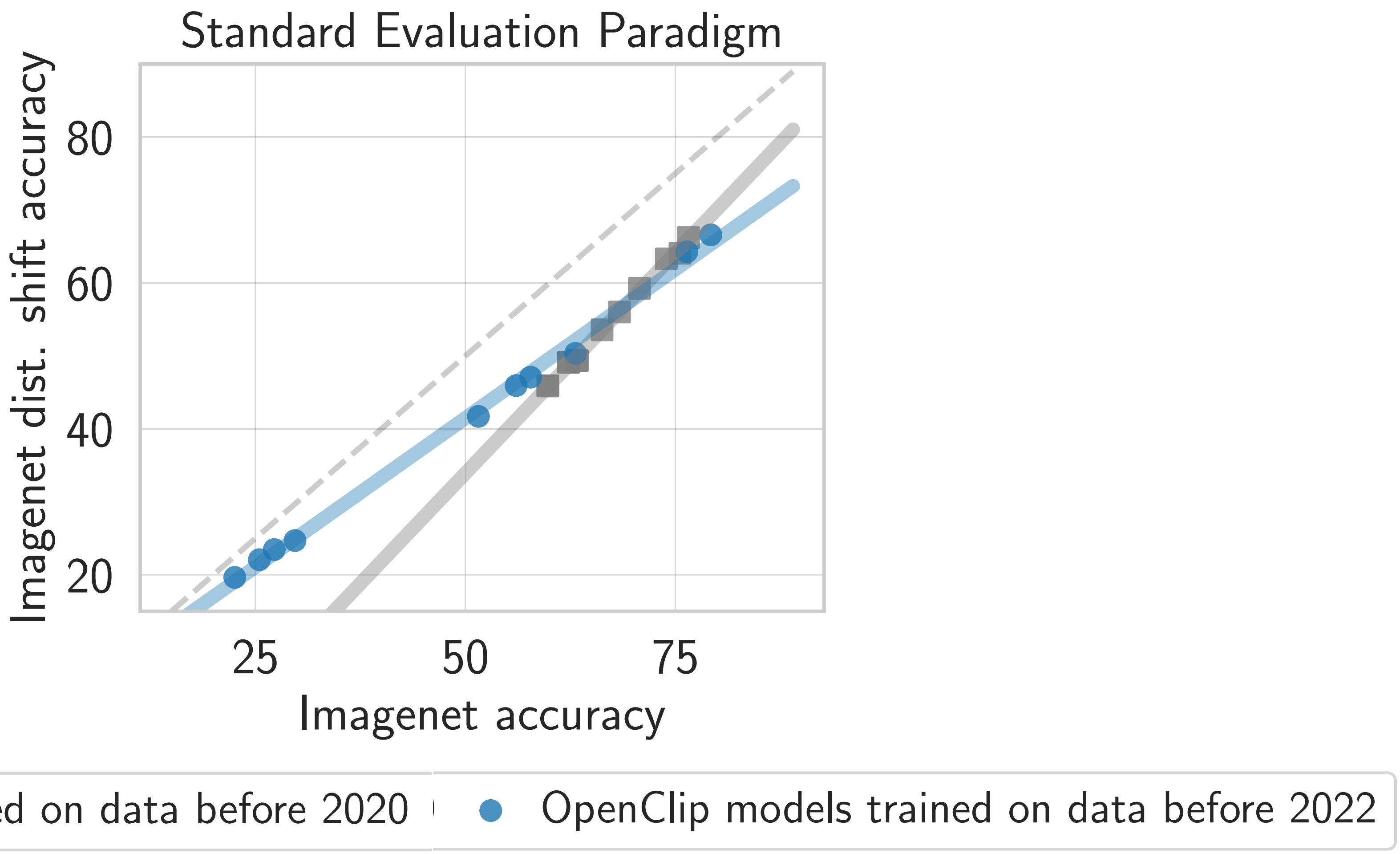
How to continuously update these models as data distributions evolves over time?

Do temporal data distribution shifts matter? Is there a need to continually to train a model?

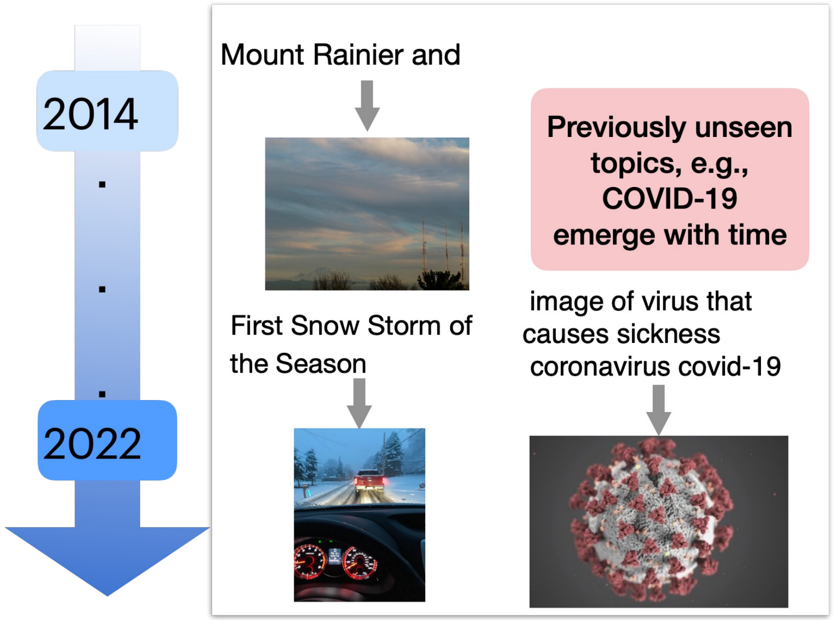
Performance of OpenAI models drop



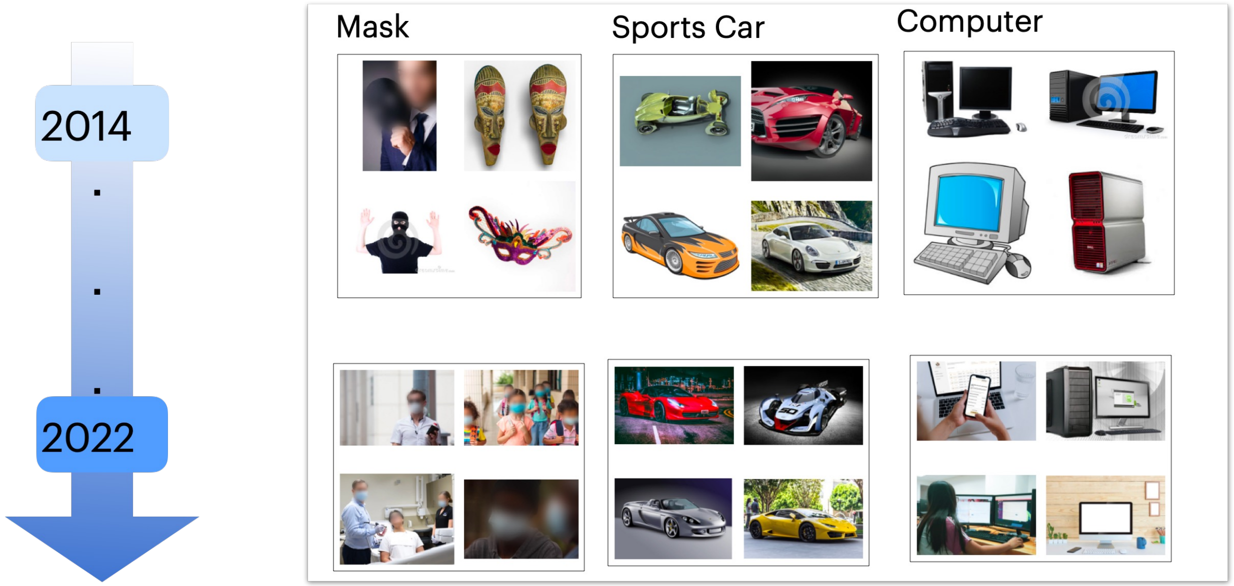
Standard Benchmarks do not capture differences



What changes? Dynamic Evaluation Benchmarks



Retrieval Task



Classification Task

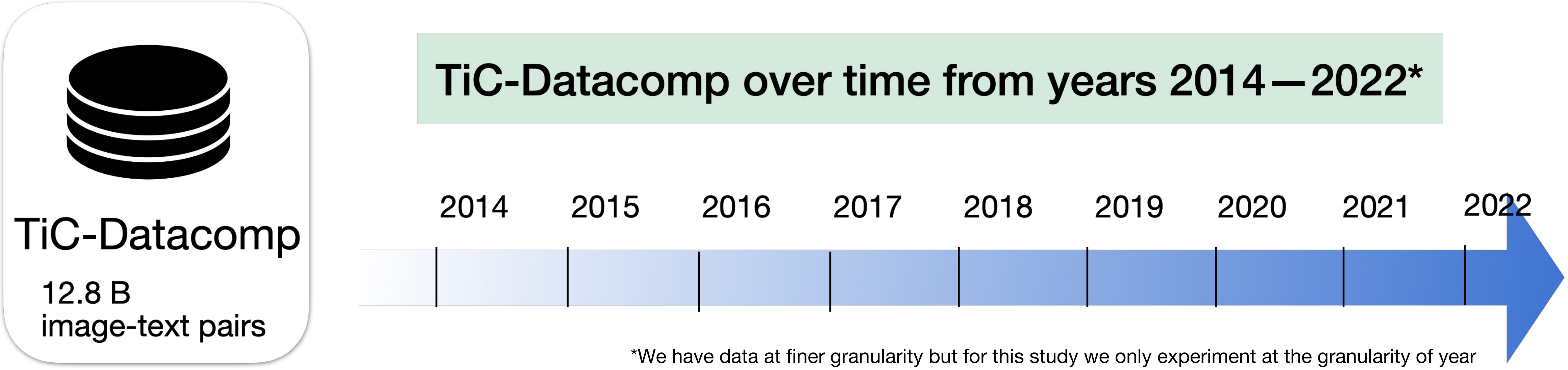
How to continuously update these models as data distributions evolves over time?

Benchmarks for continual training of CLIP

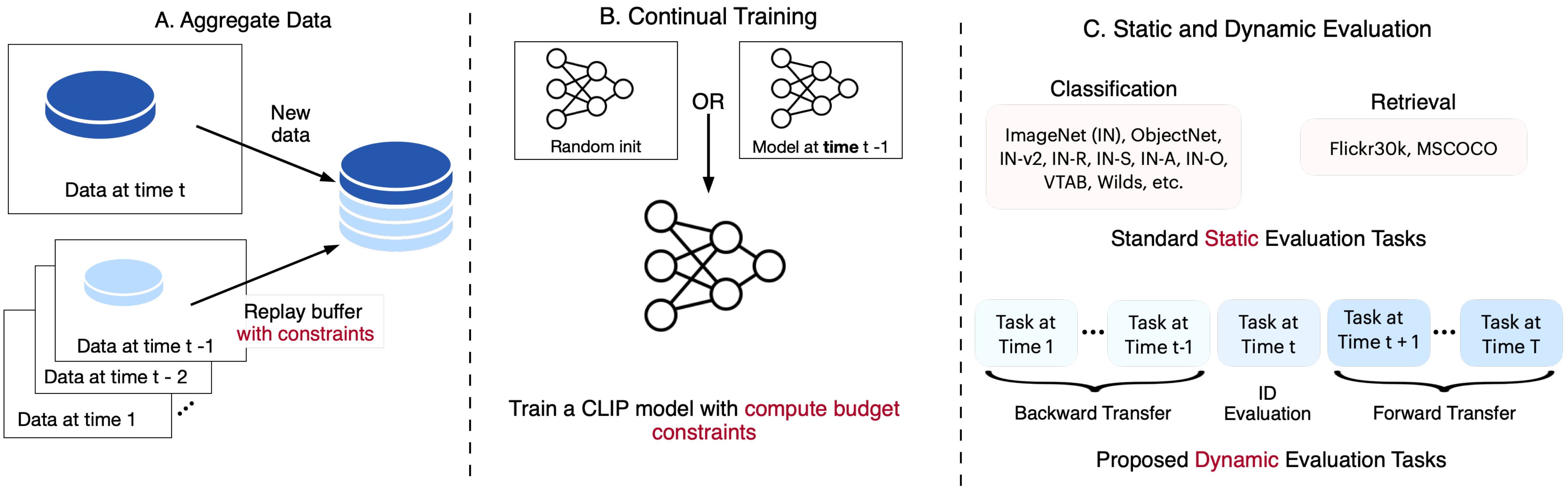
- We create benchmarks by augmenting time information to existing datasets.

Benchmarks for continual training of CLIP

- We create benchmarks by augmenting time information to existing datasets.



Experiment Protocol for Continual Learning

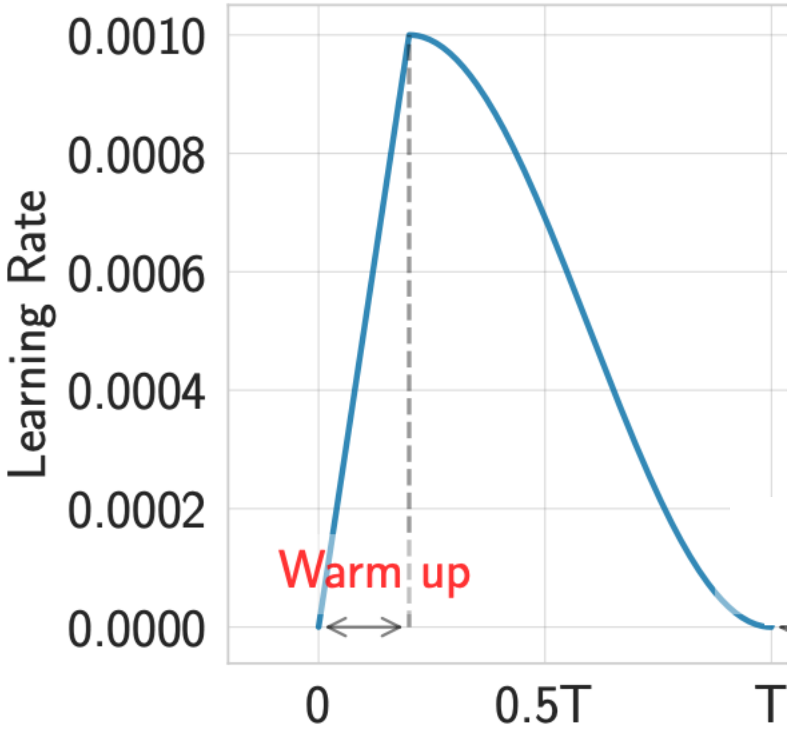


Oracle method uses t*C compute at every time step t

Baselines use C compute at every time step t

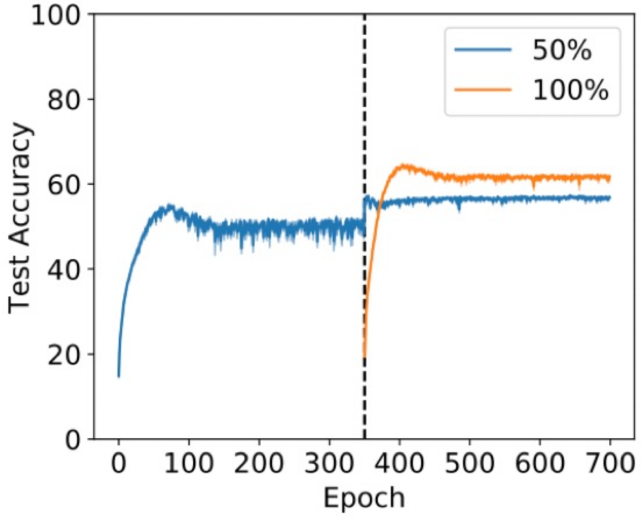
Challenges in Continual Training

- Unclear how to schedule learning rates for subsequent runs



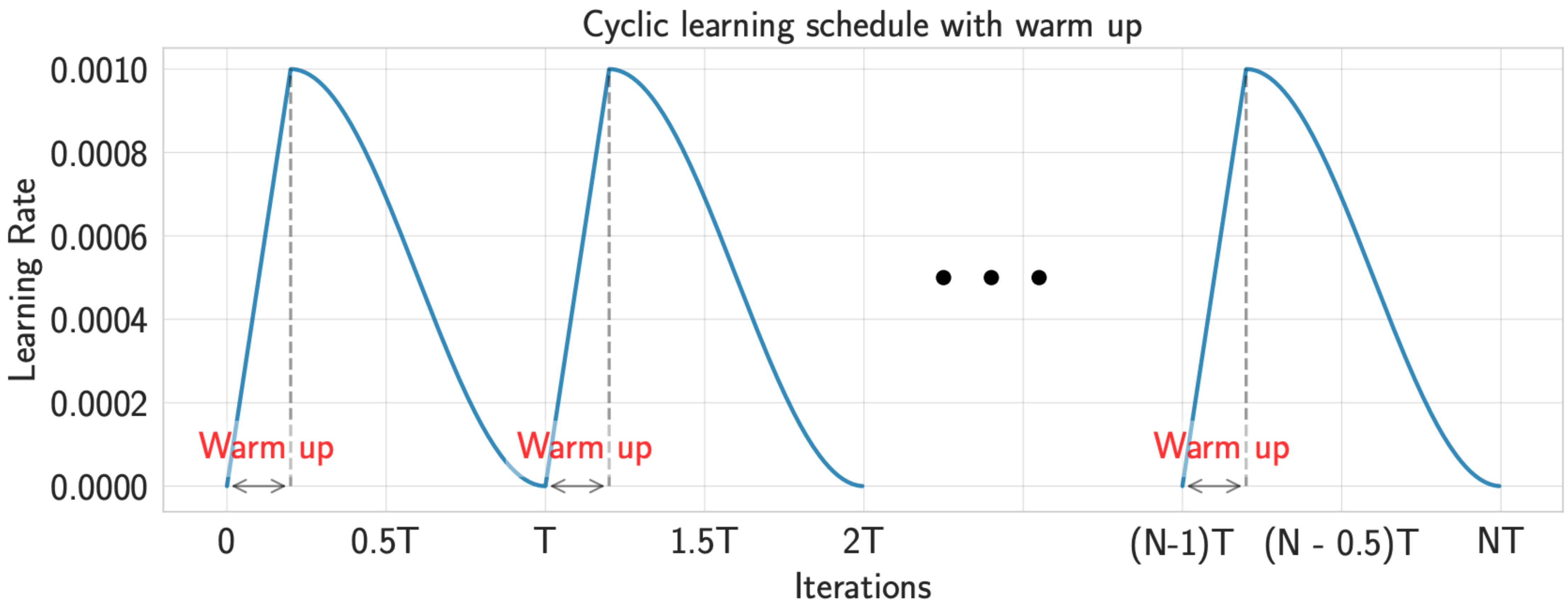
Challenges in Continual Training

- Unclear how to schedule learning rates for subsequent runs
- **Common wisdom:** Start training from scratch instead of using previous models
- Rationale: Loss of plasticity Ash and Adams 2020

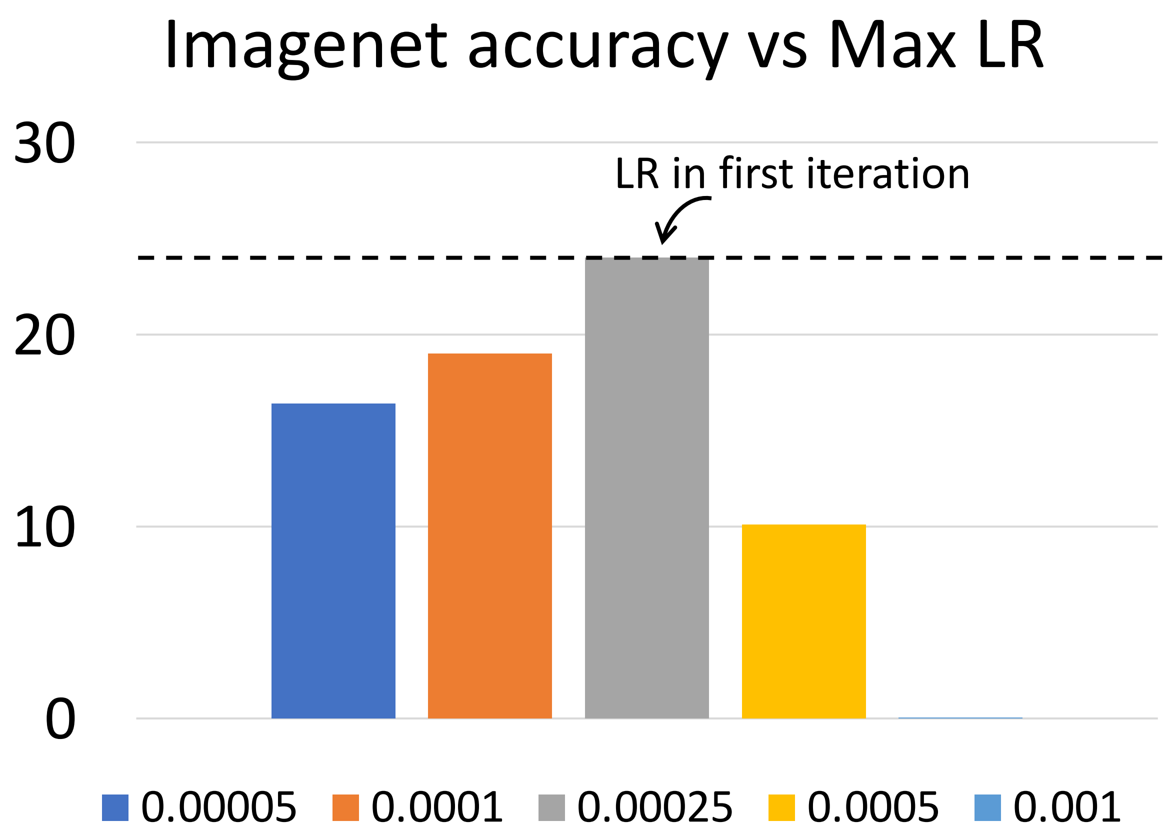


Ash and Adams. On Warm-Starting Neural Network Training (NeurIPS 2020)

How to schedule learning rate?



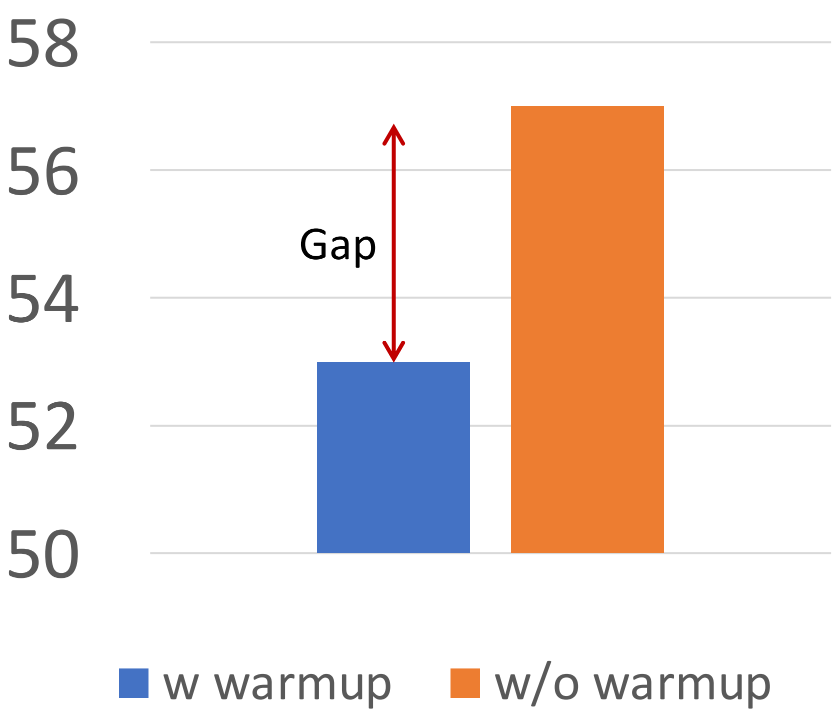
Do we need the same maximum learning rate?



If we decrease LR in subsequent iterations then zero shot IN accuracy drops

Increasing LR in subsequent iterations makes training unstable

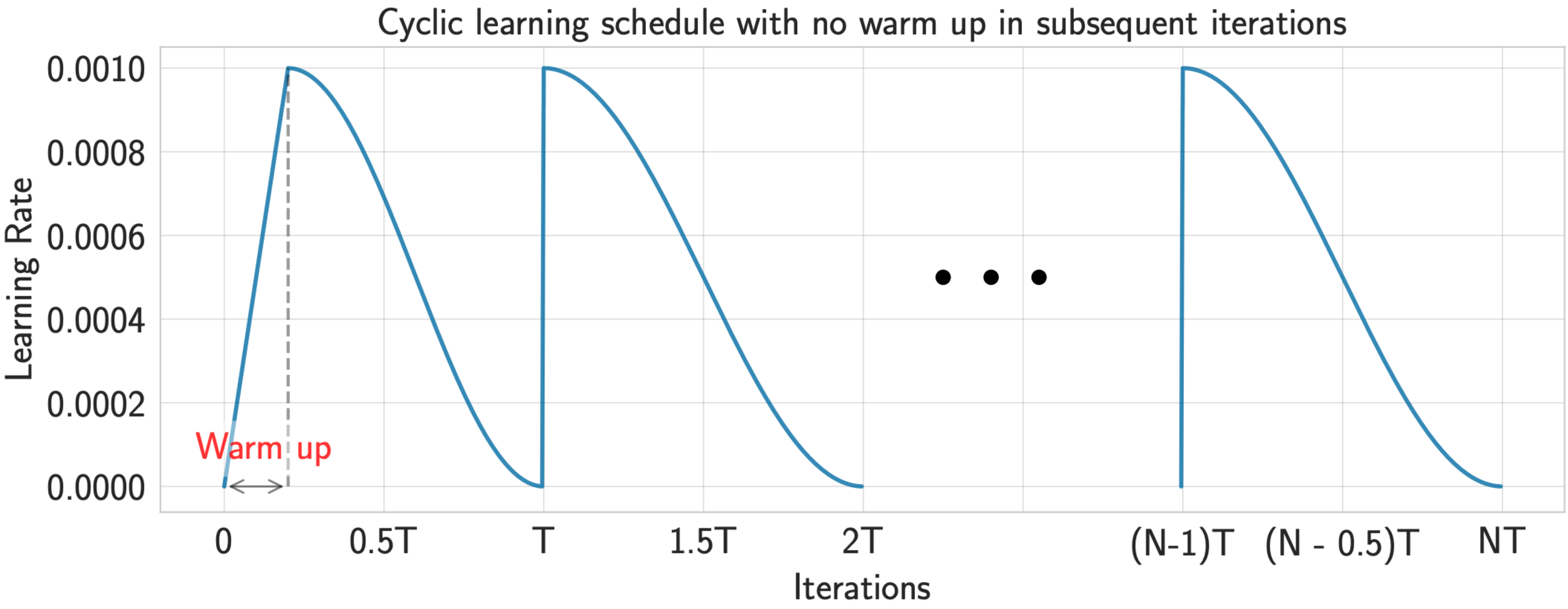
Do we need warmup in subsequent iterations?



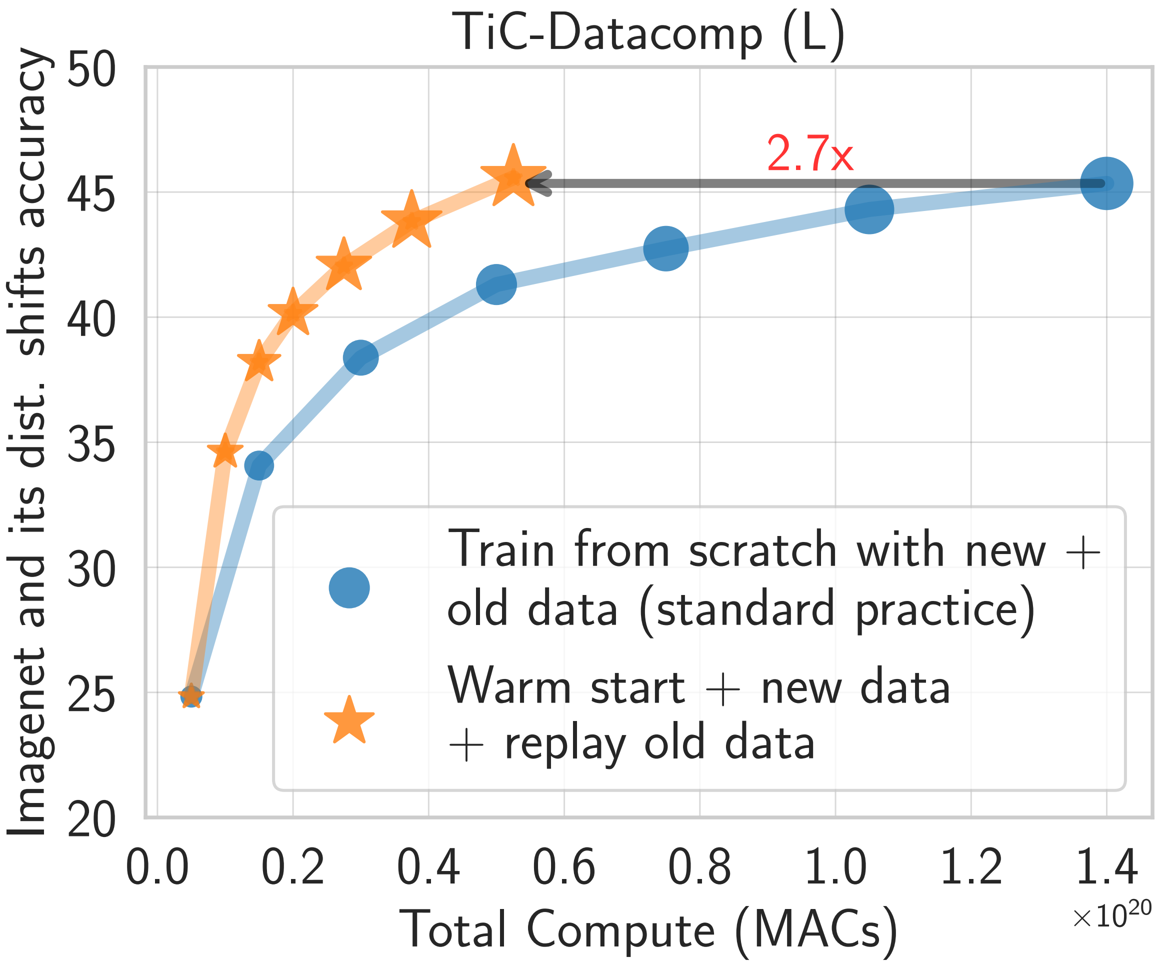
Removing warmup in subsequent iterations improve downstream performance

This gap often corresponds to 2-3x extra compute

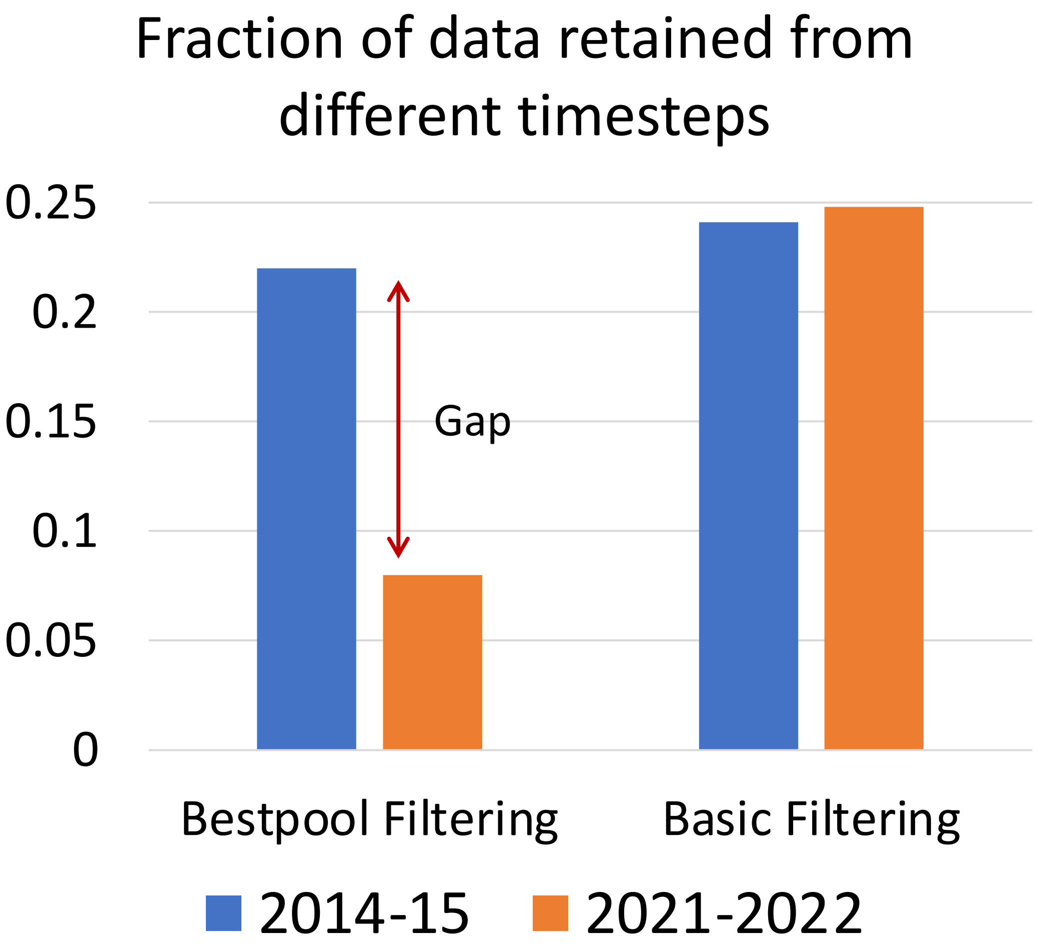
How to schedule learning rate?



Summing up: Simple Baselines Show Promise



Implications for data filtering



Bestpool filtering (which using Imagenet for filtering) biases data filtering technique to prefer old data

More Findings in the Paper

- Diverse dataset sources: Tic-Redcaps, Tic-YFCC
- Different continual learning methods, e.g., EWC, LwF, etc.

Benchmark	**Method**
TIC-DataComp (M)	Sequential Patching Cumulative-Exp Cumulative-Equal Cumulative-All
	EWC ($\lambda_{EWC} = 1$)* LwF* Cumulative-All* Oracle**

Takeaways

Takeaway 1: Need to continually train CLIP models as performance drops on data from new time steps

Takeaway 2: First continual learning benchmark to train CLIP models at time evolving internet data; simple baselines show promise

lot of interesting questions to explore next …

- *impacts on generative VLMs like Llava?*
- *better LR schedules to improve efficiency*
- *downstream implications on Stable Diffusion models?*

Questions?

Code & Data

Paper

@saurabh_garg67

sgarg2@andrew.cmu.edu

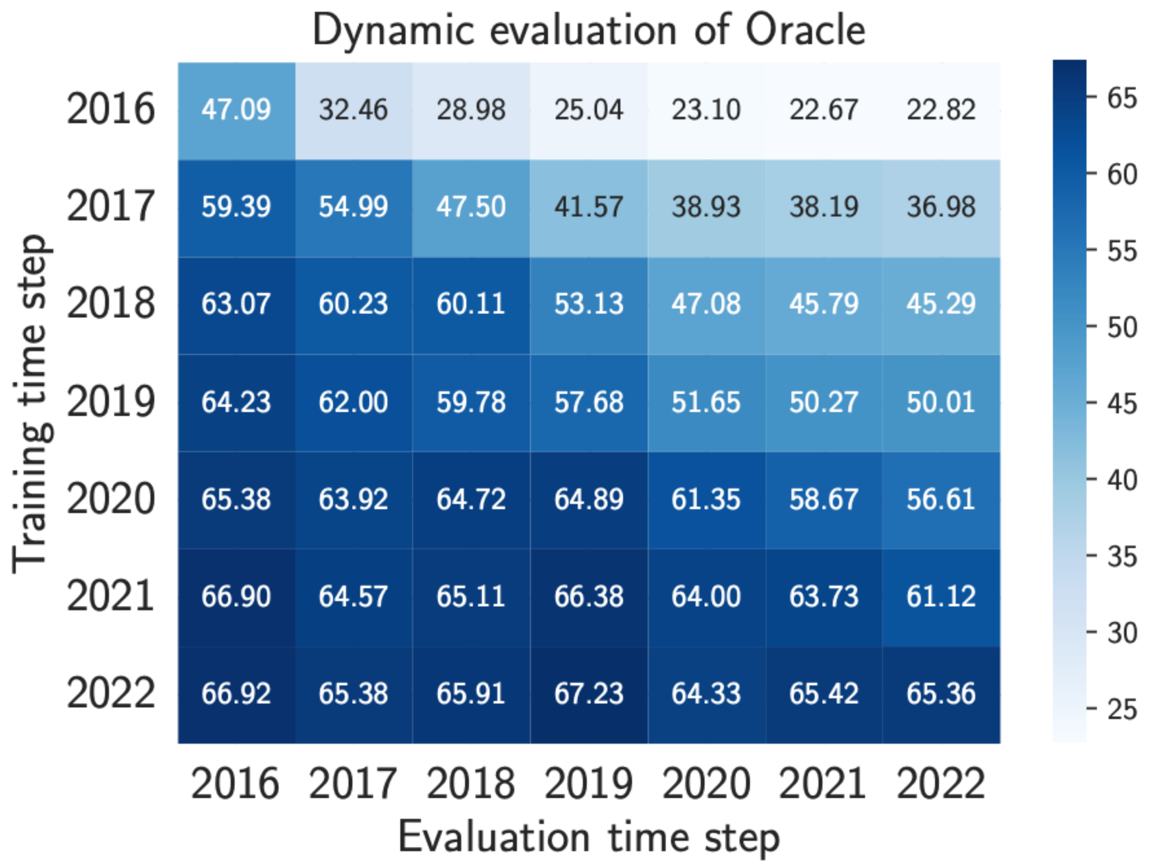
http://saurabhgarg1996.github.io/

More Findings in the Paper

- Diverse dataset sources: Tic-Redcaps, Tic-YFCC
- Different continual learning methods, e.g., EWC, LwF, etc.

Benchmark	Method	Compute (MACs)	Static Tasks: ImageNet	Static Tasks: ImageNet dist. shift	Static Tasks: Flickr30k	Static Tasks: Average over 28 datasets	Dynamic Retrieval Tasks: Backward Transfer	Dynamic Retrieval Tasks: ID Performance	Dynamic Retrieval Tasks: Forward Transfer
TIC-DataComp (M)	Sequential	3.0×10^{18}	19.2	16.4	16.4	15.0	25.7	26.4	14.9
	Patching	3.0×10^{18}	19.3	16.8	18.5	14.7	26.9	25.4	14.5
	Cumulative-Exp	3.0×10^{18}	22.1	18.4	20.4	16.7	31.7	27.1	**15.2**
	Cumulative-Equal	3.0×10^{18}	22.1	18.4	19.2	17.1	31.8	26.8	15.1
	Cumulative-All	3.0×10^{18}	24.0	20.2	20.9	17.9	33.8	26.4	15.1
	EWC ($\lambda_{EWC} = 1$)*	3.6×10^{18}	18.7	16.3	16.2	15.1	25.5	26.4	14.8
	LwF*	3.8×10^{18}	19.2	16.5	17.7	14.3	25.6	26.6	14.9
	Cumulative-All*	3.9×10^{18}	**30.0**	**25.0**	**28.6**	**22.3**	**36.7**	**28.3**	**15.5**
	Oracle**	1.2×10^{19}	25.5	21.2	23.3	19.0	34.9	27.8	**15.6**

More Findings in the Paper



Performance of Oracle on future time steps drops highlighting distribution shift in dataset.