Navigating Hurdles for Continual Learning in the Wild

Ameya Prabhu

Motivations for This Talk

- Quite Experimental! Need feedback.

Motivations for This Talk

- Quite Experimental! Need feedback.

<u>Backstory</u>

- Gave my first broader talk recently at Computer Vision Talks
    - Title: "Computationally Budgeted Continual Learning"
    - Link: drimpossible.github.io/talks_and_blogs/
- Asked to elaborate on "How I Look at Problems?"
    - Me: "How Phil, Adel, Ozan/Vladlen, Anoop, Maneesh taught me"
    - This talk!

Outline

My Thought Process

- [Methods] Beware: Complex Methods (5 mins)
- [Evaluation] Ask: Why Evaluate X? (15 mins)
- [Problem Setup] Target the Most Pressing Problems first.. (15 mins)

Outline

My Thought Process

- **[Methods]** Beware: Complex Methods (5mins)
- [Evaluation] Ask Why Evaluate X?
- [Data] Target the Most Pressing Problems first..

Part 1 [Brief]

My Way of Thinking about Methods

Continual Learning Research as Treatment Effects

- Come up with a novel idea A

Slides from: Helping or Hurting? Great Ideas and Why They Don't Matter, Chris Russell (TVG20 Talk @ Oxford)

Continual Learning Research as Treatment Effects

- Come up with a novel idea A
- Add it to a state-of-the-art pipeline P
- And evaluate

Slides from: Helping or Hurting? Great Ideas and Why They Don't Matter, Chris Russell (TVG20 Talk @ Oxford)

Continual Learning Research as Treatment Effects

- Come up with a novel idea A
- Add it to a state-of-the-art pipeline P
- And evaluate
- If E(P + A) > E(P)
- Write a paper => reviewers recognize how smart we are => publish

Slides from: Helping or Hurting? Great Ideas and Why They Don't Matter, Chris Russell (TVG20 Talk @ Oxford)

Continual Learning Research as Treatment Effects

- Come up with a novel idea A
- Add it to a state-of-the-art pipeline P
- And evaluate
- If E(P + A) > E(P)
- Write a paper => reviewers recognize how smart we are => publish

^A Large Fraction of Continual Learning Papers

Except.. This is Not How It Works!

- Come up with a novel idea A
- Add it to the state-of-the-art pipeline P

Slides from: Helping or Hurting? Great Ideas and Why They Don't Matter, Chris Russell (TVG20 Talk @ Oxford)

Except.. This is Not How It Works!

- Come up with a novel idea A.
- Add it to the state-of-the-art pipeline P
- And it breaks!

Slides from: Helping or Hurting? Great Ideas and Why They Don't Matter, Chris Russell (TVG20 Talk @ Oxford)

Except.. This is Not How It Works!

- Come up with a novel idea A.
- Add it to the state-of-the-art pipeline P
- And it breaks!
- Fix it with incremental changes i1 + i2 +…
- And Evaluate E(P + A + i1 + i2 +…) > E(P)
- Write a paper => reviewers recognize how smart we are => publish

Slides from: Helping or Hurting? Great Ideas and Why They Don't Matter, Chris Russell (TVG20 Talk @ Oxford)

Except.. This is Not How It Works!

- Come up with a novel idea A.
- Add it to the state-of-the-art pipeline P
- And it breaks!
- Fix it with incremental changes i1 + i2 +…
- And Evaluate E(P + A + i1 + i2 +…) > E(P)
- Write a paper => reviewers recognize how smart we are => publish

<u>The Problem</u>

A was the novel idea! But all performance increases came from i1 + i2 +…

Slides from: Helping or Hurting? Great Ideas and Why They Don't Matter, Chris Russell (TVG20 Talk @ Oxford)

Except.. This is Not How It Works!

- Come up with a novel idea A.
- Add it to the state-of-the-art pipeline P
- And it breaks!
- Fix it with incremental changes i1 + i2 +…
- And Evaluate E(P + A + i1 + i2 +…) > E(P)
- Write a paper => reviewers recognize how smart we are => publish

The Problem

A was the novel idea! But all performance increases came from i1 + i2 +...

Solution: Ablate A?

Slides from: Helping or Hurting? Great Ideas and Why They Don't Matter, Chris Russell (TVG20 Talk @ Oxford)

Ablations Are Not Enough!

The problem:

- We picked and tuned i1 + i2 +... to make P + A work
- Of course, when we remove A => Something degrades
- Doesn't mean A is needed!

Hence, my skepticism

Slides from: Helping or Hurting? Great Ideas and Why They Don't Matter, Chris Russell (TVG20 Talk @ Oxford)

Preliminary Takeaways

- Novelty is Overrated!
  - Be radically more skeptical of complex approaches: Justify extra burden!

Slides from: Helping or Hurting? Great Ideas and Why They Don't Matter, Chris Russell (TVG20 Talk @ Oxford)

Preliminary Takeaways

- Novelty is Overrated!
    - Be radically more skeptical of complex systems: Justify extra burden!

- Best Practices remain Stable!
    - Incremental "tricks" which generalize drive the field forward!

Slides from: Helping or Hurting? Great Ideas and Why They Don't Matter, Chris Russell (TVG20 Talk @ Oxford)

Preliminary Takeaways

- Novelty is Overrated!
  - Be radically more skeptical of complex systems: Justify extra burden!

- Best Practices remain Stable!
  - Incremental "tricks" which generalize drive the field forward!

So is simplicity better? Yes, but..
We need to go one important step ahead

Slides from: Helping or Hurting? Great Ideas and Why They Don't Matter, Chris Russell (TVG20 Talk @ Oxford)

Hypothesis Testing: Methods as Interventions

Form Hypothesis, Give an Intervention to test it: Benchmark & See Results

Slides from: Helping or Hurting? Great Ideas and Why They Don't Matter, Chris Russell (TVG20 Talk @ Oxford)

Hypothesis Testing: Methods as Interventions

Form Hypothesis, Give an Intervention to test it: Benchmark & See Results

- GDumb (ECCV 2020):
  - Methods don't use online stream
  - Performance degrades! Test on longer time horizons, best: Reset (B2/B3)
- BudgetCL (CVPR 2023):
  - (i) Methods focus on memory rather than compute! Bad!

Slides from: Helping or Hurting? Great Ideas and Why They Don't Matter, Chris Russell (TVG20 Talk @ Oxford)

Hypothesis Testing: Methods as Interventions

Form Hypothesis, Give an Intervention to test it: Benchmark & See Results

- GDumb (ECCV 2020):
  - Methods don't use online stream
  - Performance degrades! Test on longer time horizons, best: Reset (B2/B3)
- BudgetCL (CVPR 2023):
  - (i) Methods focus on memory rather than compute! Do they generalize?

Simplicity comes for free, as a by-product of a focused question

Slides from: Helping or Hurting? Great Ideas and Why They Don't Matter, Chris Russell (TVG20 Talk @ Oxford)

Hypothesis Testing: Methods as Interventions

Form Hypothesis, Give an Intervention to test it: Benchmark & See Results

- Simplicity comes for free, as a by-product of a focused question

Excitement about Pre-Registration

- Pre-registration promising to encourage this style of research!
- Excited to see what happens in CLAI Unconference!

Slides from: Helping or Hurting? Great Ideas and Why They Don't Matter, Chris Russell (TVG20 Talk @ Oxford)

Hypothesis Testing: Methods as Interventions

Form Hypothesis, Give an Intervention to test it: Benchmark & See Results

- Simplicity comes for free, as a by-product of a focused question

Excitement about Pre-Registration

- Pre-registration promising to encourage this style of research!
- Excited to see what happens in CLAI Unconference!

Let's Extend this thinking to Metrics and Data..

Outline

My Thought Process

- [Methods] Beware: Complex Methods (5 mins)
- **[Evaluation]** Ask: Why Evaluate X? (15 mins)
- [Data] Target the Most Pressing Problems first.. (15 mins)

Part 2

My Way of Thinking about Metrics

Continual Learning Research as Treatment Effects

- Have a novel idea A
- Add it to a state-of-the-art pipeline P
- And evaluate
- If E(P + A)>E(P)
- Write a paper, the reviewers recognize how smart we are, and publish

Slides from: Helping or Hurting? Great Ideas and Why They Don't Matter, Chris Russell (TVG20 Talk @ Oxford)

Chronic Over-Reporting of Metrics

Case Study: Online Continual Learning

Report a whole bunch of metrics **E**:

- Avg. Accuracy **and** Forgetting **and** Anytime Accuracy **and** LP Accuracy

.. and just say we are better!

Bad!

Metrics as Measuring Progress on some Objective

State an Objective, propose/use a metric to measures progress, Benchmark!

Metrics as Measuring Progress on some Objective

State an Objective, propose/use a metric to measures progress, Benchmark!

<u>Case Study</u>: Online Continual Learning

Objective: I want to rapidly adapt to incoming data! (Goal of all online systems!)

Metrics as Measuring Progress on some Objective

State an Objective, propose/use a metric which measures that, Benchmark!

Case Study: Online Continual Learning

Objective: I want to rapidly adapt to incoming data! (Goal of all online systems!)

Relooking metrics in online continual learning:

Avg. Accuracy **and** Forgetting **and** Anytime Accuracy **and** LP Accuracy

None of them measure rapid adaptation! They measure forgetting!

Metrics as Measuring Progress on some Objective

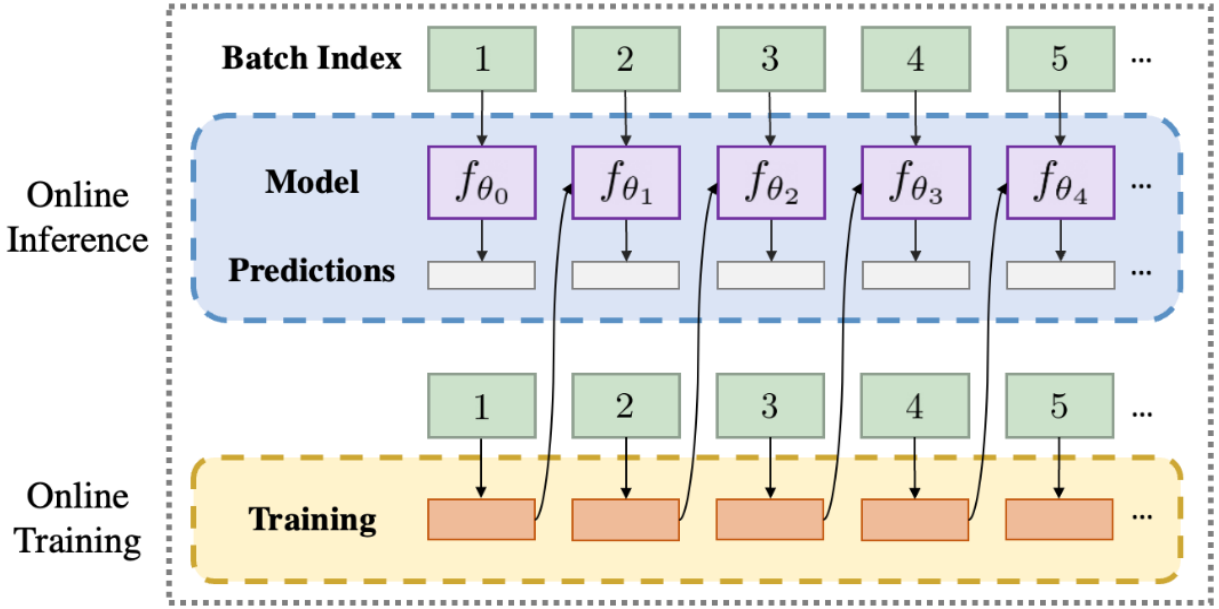
State a Objective, propose/use a metric which measures that, Benchmark!

<u>Case Study</u>: Online Continual Learning

Objective: I want to rapidly adapt to incoming data! (Goal of all online systems!)

<u>Trad. Online Learning uses Online accuracy</u>

- Measure of the model's performance on the next unseen sample/batch.



Free from Memory Limits: Prabhu et. al., “Online Continual Learning Without the Storage Constraint” Arxiv.

With Memory Limits: “Real-Time Evaluation in Online Continual Learning: A New Hope” Ghunaim et al, CVPR23.

Metrics as Measuring Progress on some Objective

State a Objective, propose/use a metric which measures that, Benchmark!

Case Study: Online Continual Learning

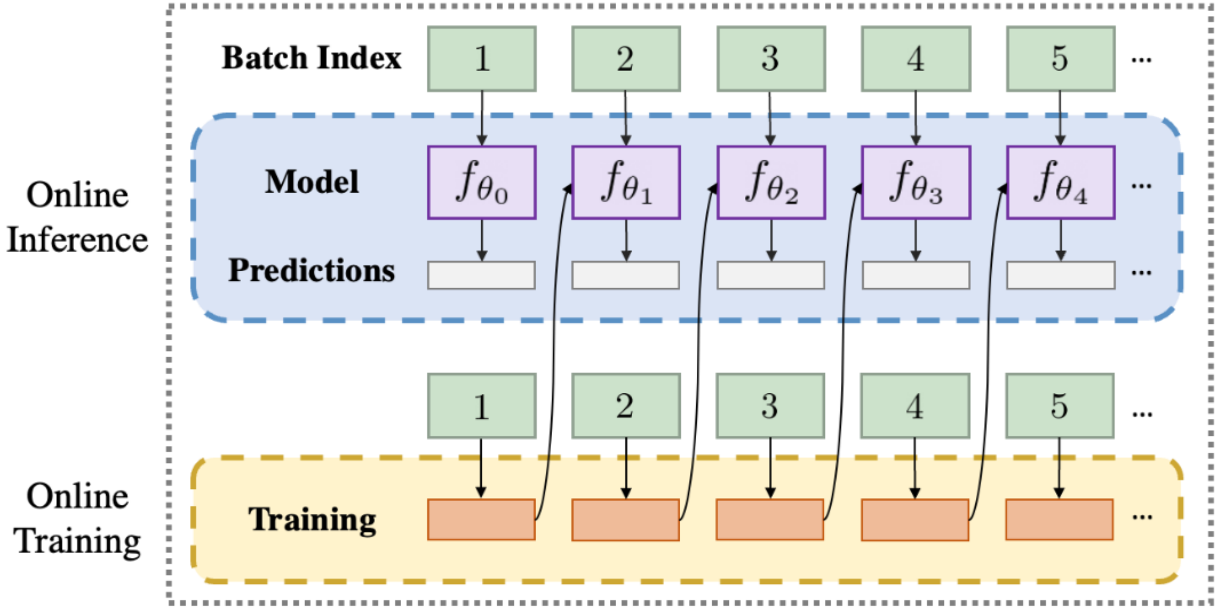
Objective: I want to rapidly adapt to incoming data! (Goal of all online systems!)

Trad. Online Learning uses Online accuracy

- Measure of the model's performance on the next unseen sample/batch.

Impractical in class-incremental OCL Setups!

- Next samples are by design same class

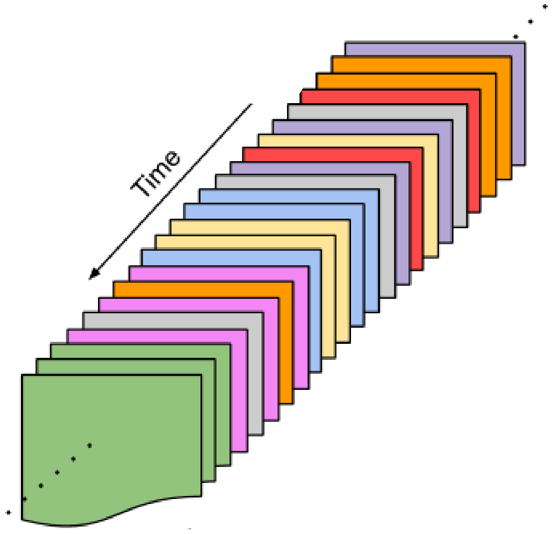


Free from Memory Limits: Prabhu et. al., “Online Continual Learning Without the Storage Constraint” Arxiv.

With Memory Limits: “Real-Time Evaluation in Online Continual Learning: A New Hope” Ghunaim et al, CVPR23.

Studying Data Streams: Measuring Adaptation

Online CL



Online accuracy

Measure of the model's performance on the next unseen sample/batch.

Hard to do in class-incremental setups

- Next samples are by design same class

Better Datasets

Continual geoLOCalization (Cai et. al., 2021)

- *Geolocation at scale*
- *713 classes, 39M images*
- *Simulates images arriving on a Flickr server.*

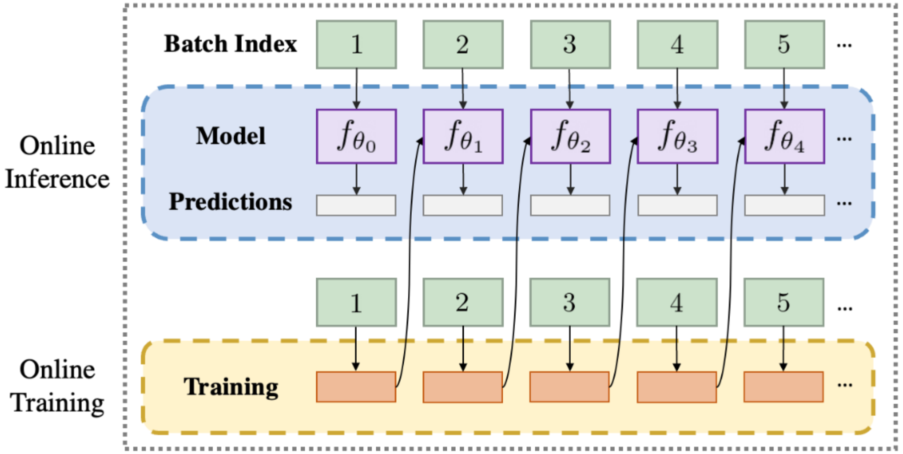
Continual Google Land Marks V2 (Prabhu et. al., 2023)

- *Long-tailed landmark classification*
- *10,788 classes, 450K images*
- *Simulates arrival on a Wikimedia Commons server.*

Free from Memory Limits: Prabhu et. al., "Online Continual Learning Without the Storage Constraint" Arxiv.

With Memory Limits: "Real-Time Evaluation in Online Continual Learning: A New Hope" Ghunaim et al, CVPR23.

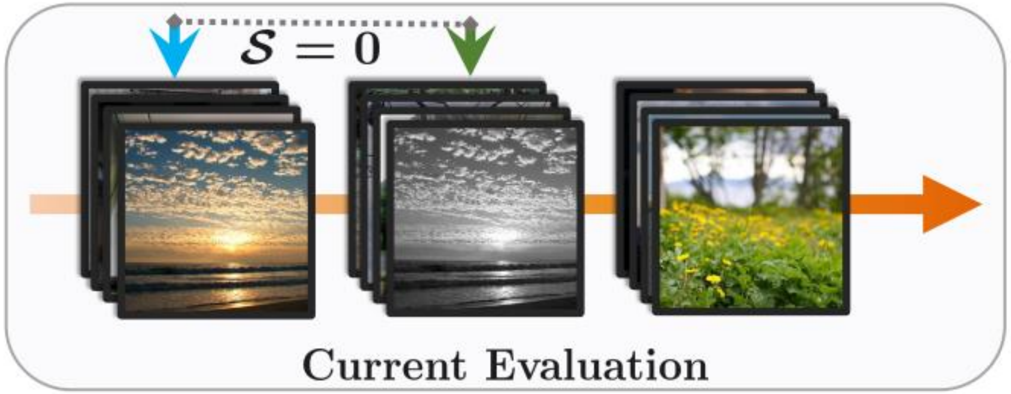
Insight: Metric Does not Evaluate Adaptation Well!



Online accuracy

Measures model's performance on the next unseen sample/batch.

Finding: The stream labels are correlated in natural streams!

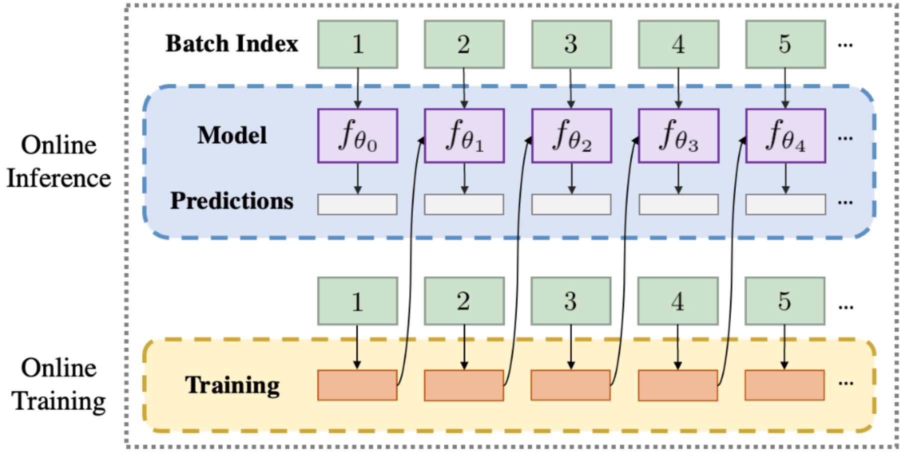


Hammoud et. al., Rapid Adaptation in Online Continual Learning: Are We Evaluating It Right?, ICCV' 23

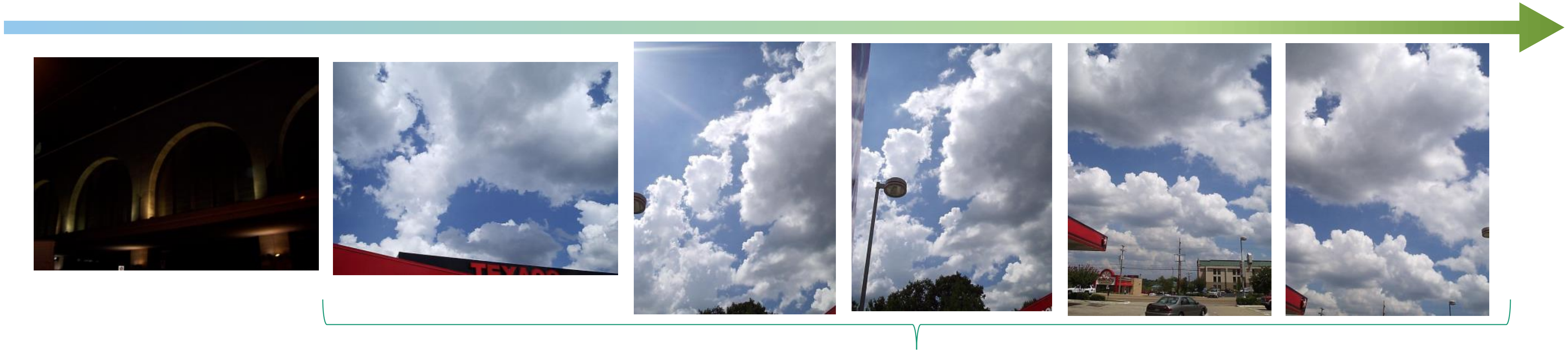
Insight: Incorrect Evaluation of Adaptation

Online accuracy

Measures model's performance on the next unseen sample/batch.



Let us look at Real CLOC Samples!

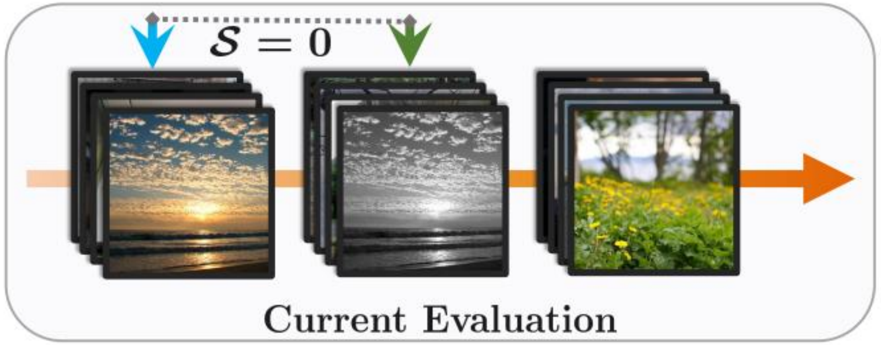


Same label !

Hammoud et. al., Rapid Adaptation in Online Continual Learning: Are We Evaluating It Right?, ICCV' 23

Insight+: Why are Correlated Samples Important?

Finding: The stream labels are correlated in natural streams!



Why is it important?

A Blind Classifier

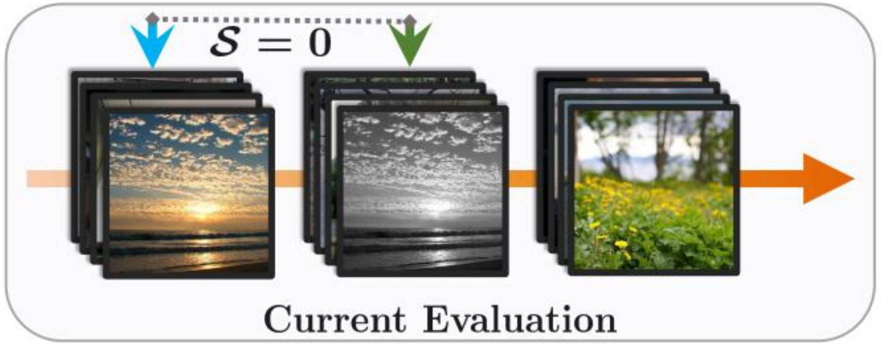
*Blind Classifier: A model that predicts the mode of the last **K** samples seen without access to the input images.*

Hammoud et. al., Rapid Adaptation in Online Continual Learning: Are We Evaluating It Right?, ICCV' 23

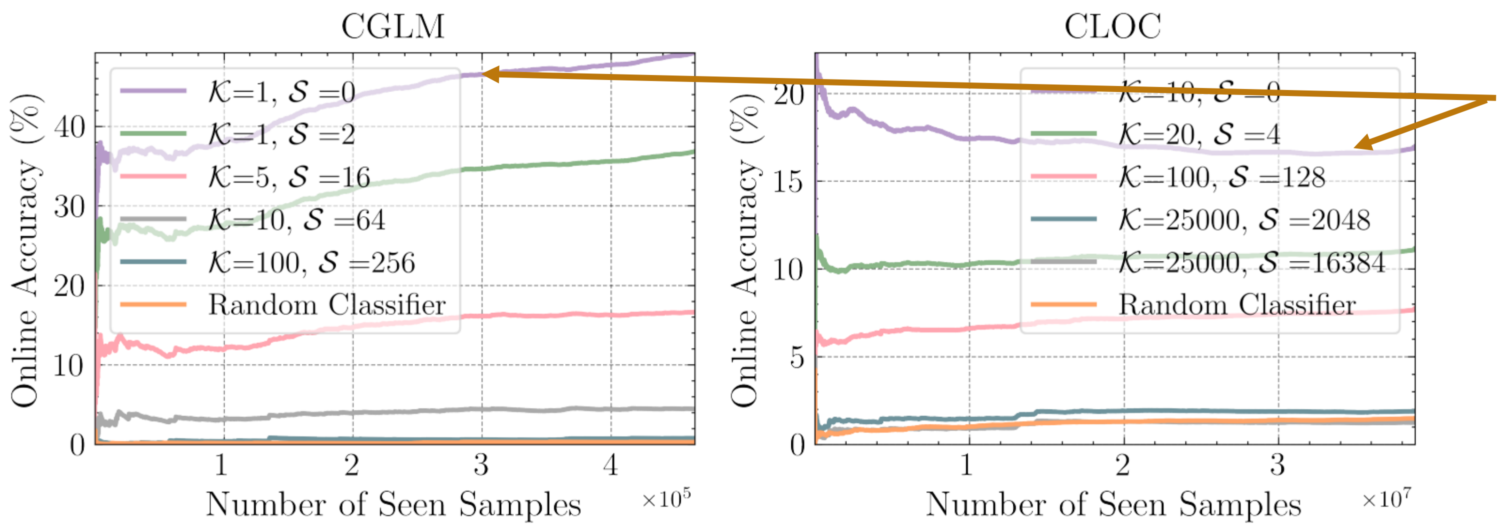
Insight+: Why are Correlated Samples Important?

Finding: The stream labels are correlated in natural streams!

Why is it important?



A Blind Classifier has Great Performance



Competitive accuracies to OCL methods!

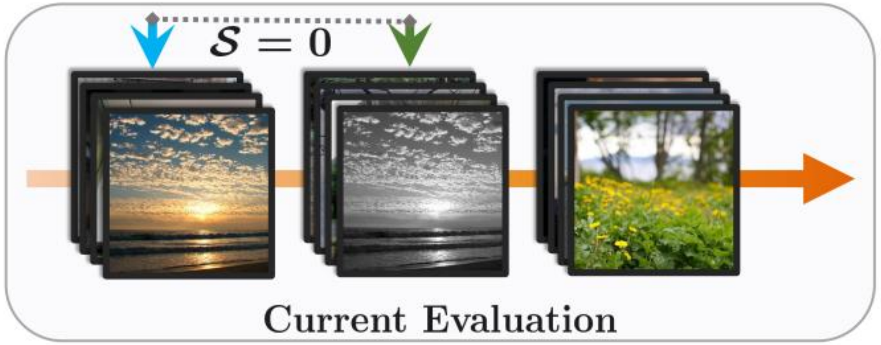
It's unambiguously a bad model!

Hammoud et. al., Rapid Adaptation in Online Continual Learning: Are We Evaluating It Right?, ICCV' 23

OverAdapt: Hacking the Label Correlations

Finding: The stream labels are correlated in natural streams!

A Blind Classifier has Great Performance



OverAdapt: A simple and clearly *wrong* baseline!

A model made to overfit to the latest data by:

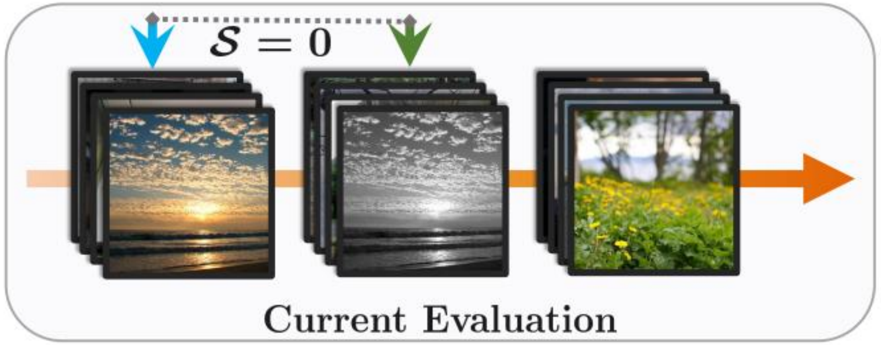
1. Adopting FIFO sampling to select training samples: The Bad Design Component in Cai et. al. 2021
2. Fix a pretrained ResNet50 backbone and train linear layer only

Hammoud et. al., Rapid Adaptation in Online Continual Learning: Are We Evaluating It Right?, ICCV' 23

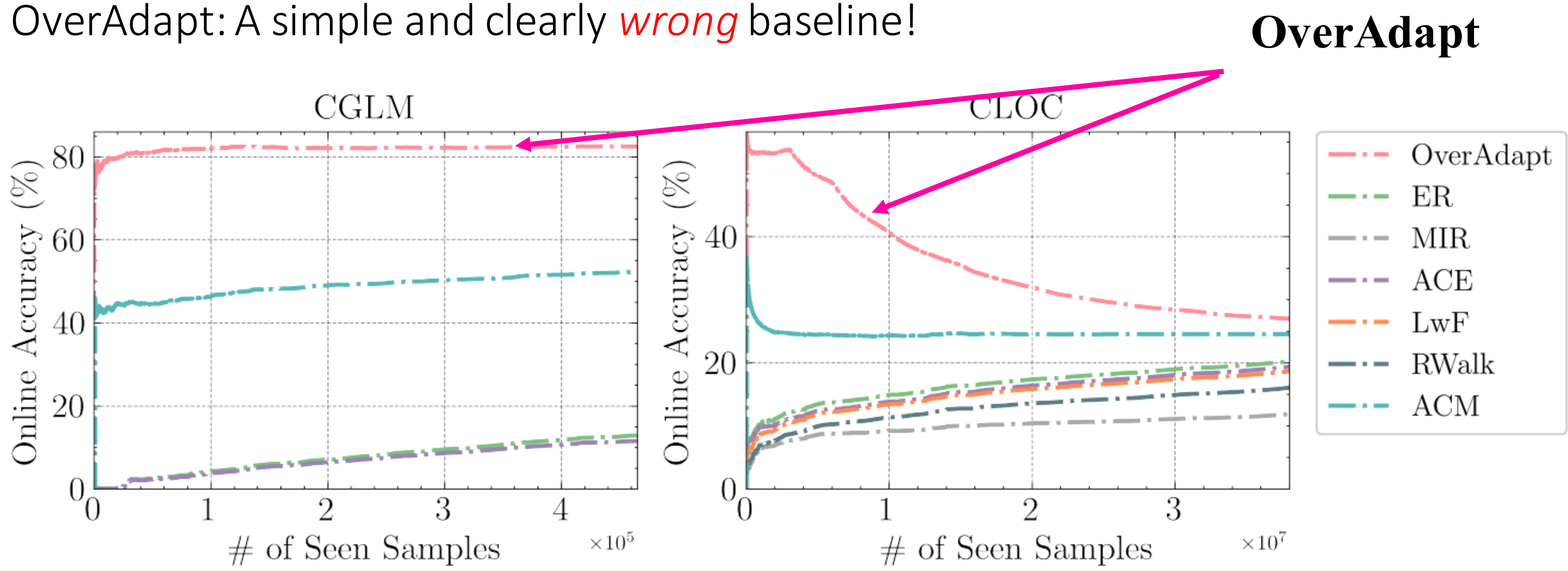
OverAdapt: Hacking the Label Correlations

Finding: The stream labels are correlated in natural streams!

A Blind Classifier has Great Performance



- OverAdapt: A simple and clearly *wrong* baseline!

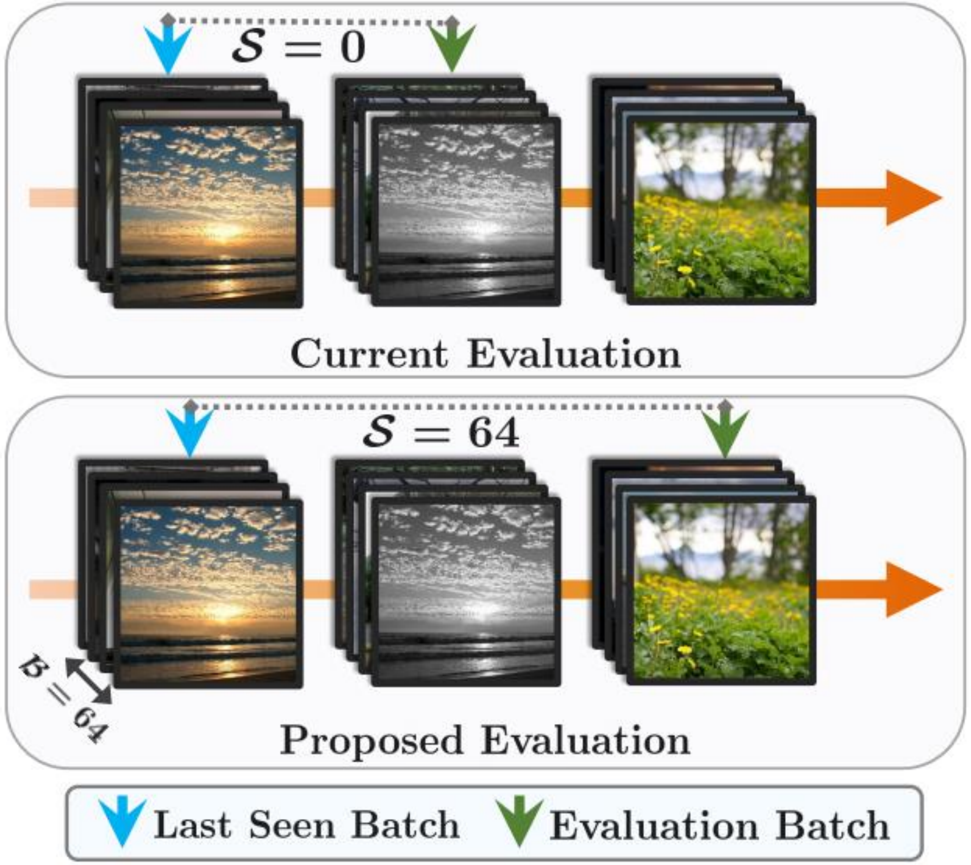


- Online Accuracy favours methods which overfit to incoming data!
- Incentivizes bad algorithms!

Hammoud et. al., Rapid Adaptation in Online Continual Learning: Are We Evaluating It Right?, ICCV' 23

Our Solution: Near-Future Accuracy

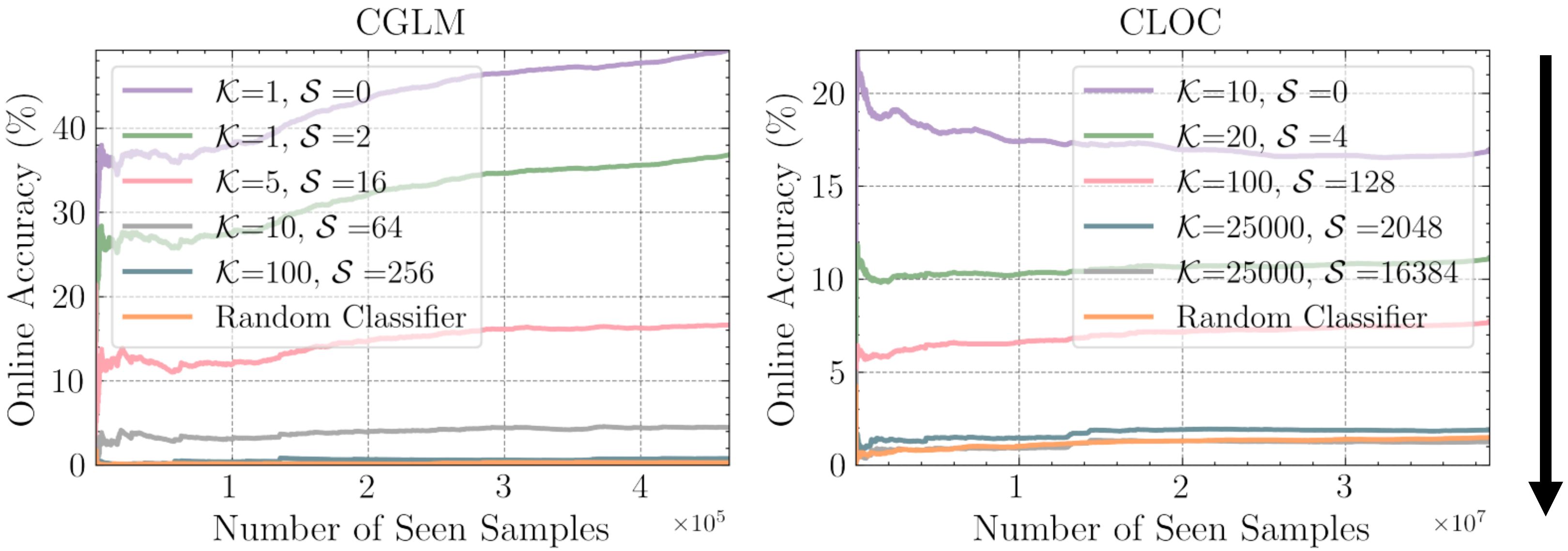
- Instead of measuring the accuracy on the immediate next batch how about we measure the accuracy on the next uncorrelated batch?
- **Question:** How do we estimate the batches to delay our evaluation with?



Hammoud et. al., Rapid Adaptation in Online Continual Learning: Are We Evaluating It Right?, ICCV' 23

Near-Future Accuracy: Blind Classifier to Rescue

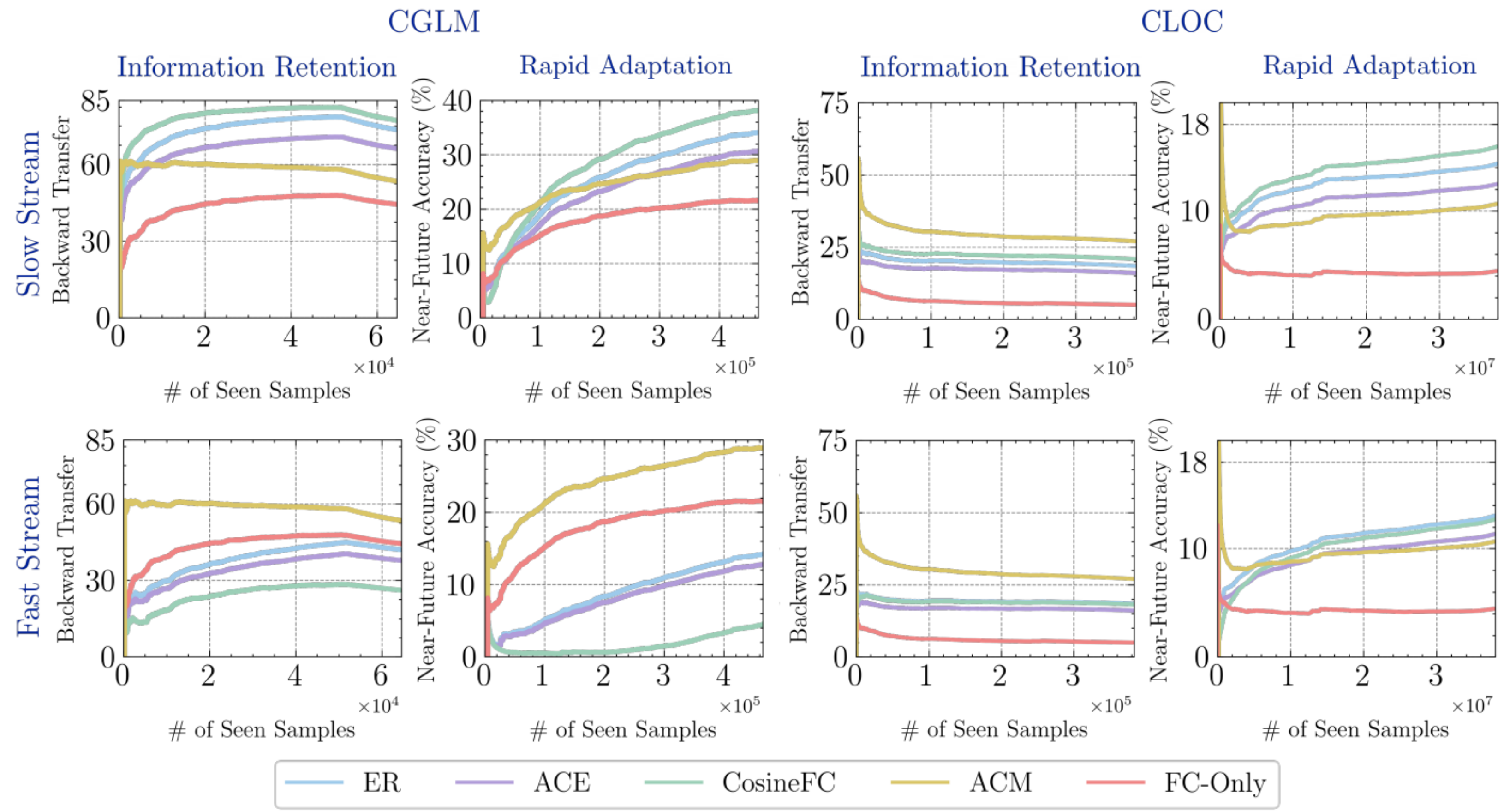
- **Question:** How to estimate the batches to delay our evaluation with?



Delay the Blind Classifier evaluation just until it converges to a random classifier!

Hammoud et. al., Rapid Adaptation in Online Continual Learning: Are We Evaluating It Right?, ICCV' 23

Near-Future Accuracy: Benchmark



Slow Stream (5/10 steps): *Use offline continual learning*

Fast Stream (1 step): Use ACM

Hammoud et. al., Rapid Adaptation in Online Continual Learning: Are We Evaluating It Right?, ICCV' 23

Outline

My Thought Process

- [Methods] Beware: Complex Methods (5 mins)
- [Evaluation] Ask: Why Evaluate X? (15 mins)
- [**Problem Setup**] Target the Most Pressing Problems First.. (15 mins)

Part 3
My Approach about Deciding Problem Setup

Thinking About the Problem Setup

Hypothesis testing frame hopefully clear by now!

Thinking About the Problem Setup

Problem Setup: How to choose assumptions while formulating a problem?

Assumptions help simplify and contextualize hard problems which..

- Allows principled approaches, far faster than hit-and-try grad student descent!
  - This is for the empirical/applied folks in the audience

Thinking About the Problem Setup

Problem Setup: How to choose assumptions while formulating a problem?

Assumptions help simplify and contextualize hard problems which

- Allows principled approaches, far faster than hit-and-try grad student descent!
  - This is for the empirical/applied folks in the audience
- Assumptions should reflect the real world
  - For the theoretical folks in the audience

Thinking About the Problem Setup

Problem Setup: How to choose assumptions while formulating a problem?

Great to see continual progression towards more realistic assumptions!

- Task-increment => Class-increment => Blurry-boundaries => Time-incremental
- Small-scale => Large-scale
- Closed-set => Open-set => Open-world => Actively Learned Open-world
- No memory => Tiny-memory => Memory-limited => Memory-unconstrained

..while imposing computational constraints!

Thinking About the Problem Setup

Problem Setup: How to choose assumptions while formulating a problem?

Great to see continual progression towards more realistic assumptions!

- Task-increment => Class-increment => Blurry-boundaries => Time-incremental
- Small-scale => Large-scale
- Closed-set => Open-set => Open-world => Actively Learned Open-world
- No memory => Tiny-memory => Memory-limited => Memory-unconstrained

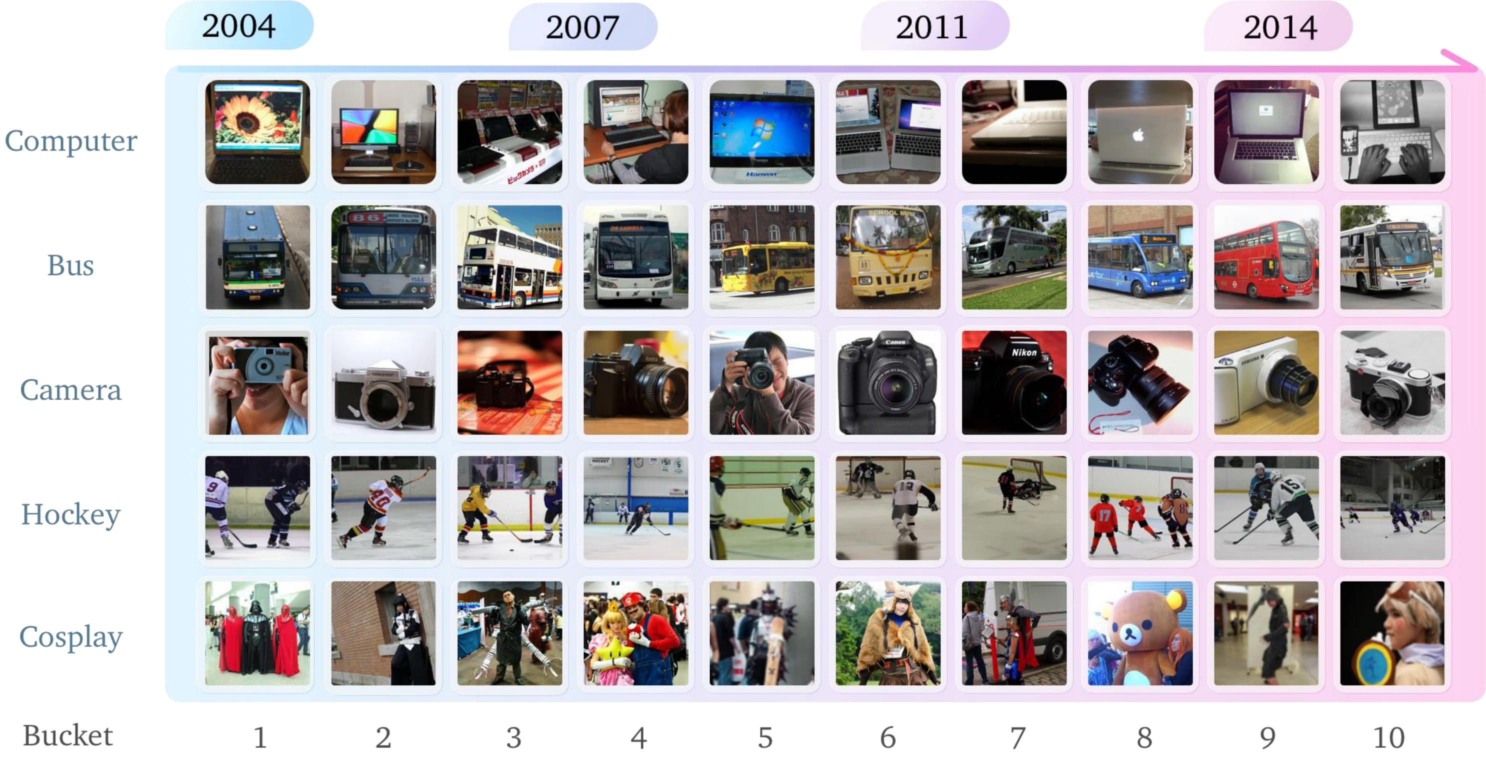
..while imposing computational constraints!

A new dimension today: Where does labeled data come from?

Studying Data Streams: Where is Training Data?

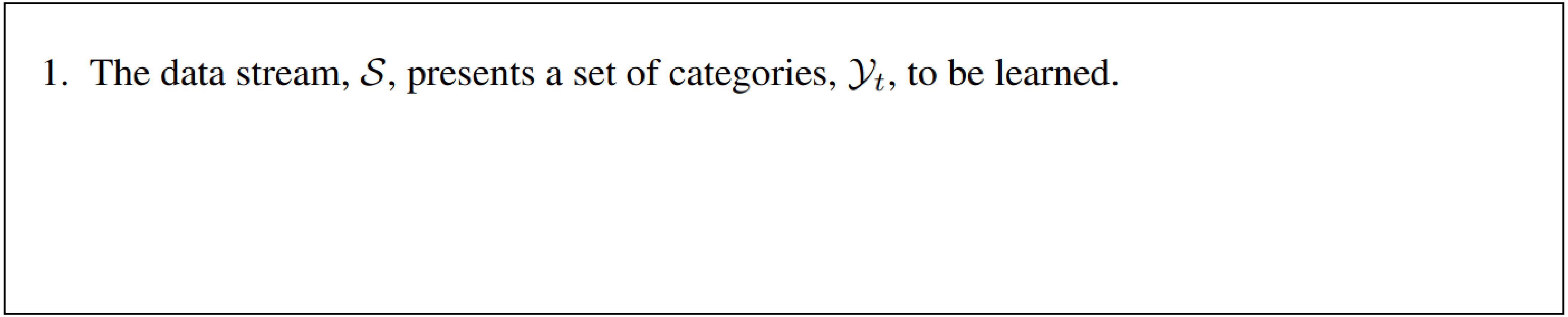
Continual Manual Annotation is costly and time-consuming — huge problem!

- CLEAR10 required $4500 for 30K



Alternate Task: Going from Categories to Classifier

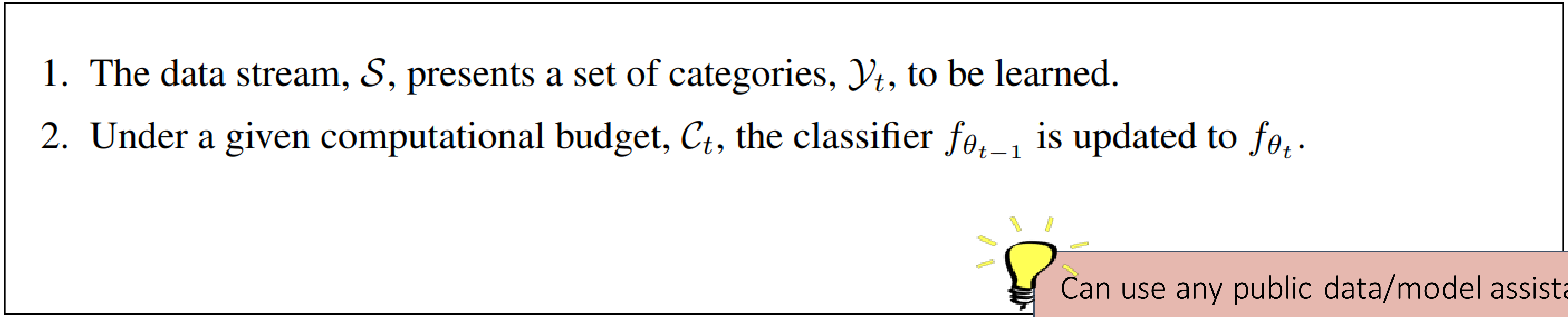
Task: Go from Category List to Trained Classifier within minutes *continually*



1. The data stream, $\mathcal{S}$, presents a set of categories, $\mathcal{Y}_t$, to be learned.

Alternate Task: Going from Categories to Classifier

Task: Go from Category List to Trained Classifier within minutes *continually*

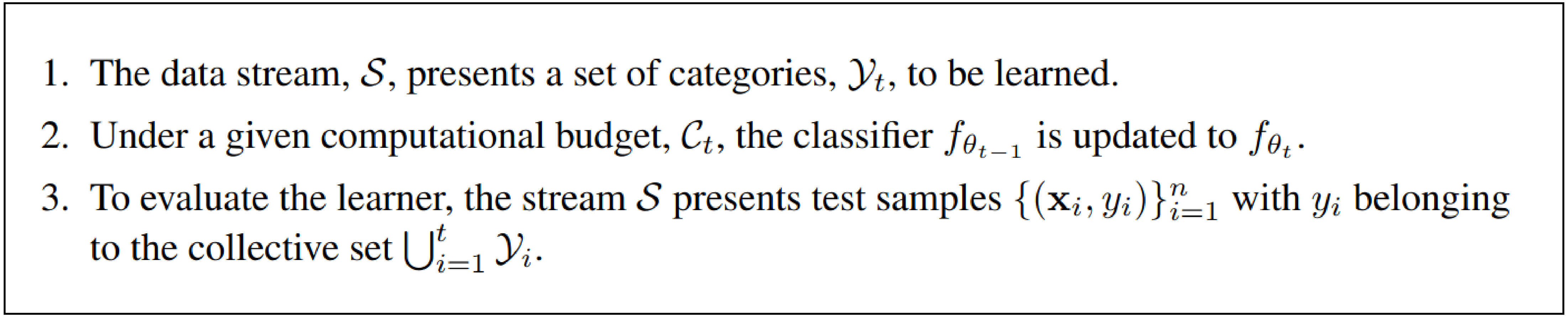


1. The data stream, $\mathcal{S}$, presents a set of categories, $\mathcal{Y}_t$, to be learned.
2. Under a given computational budget, $\mathcal{C}_t$, the classifier $f_{\theta_{t-1}}$ is updated to f_{θ_t}.

Can use any public data/model assistance to do this! E.g. GPT3, DALL-E, LAION5B

Alternate Task: Going from Categories to Classifier

Task: Go from Category List to Trained Classifier within minutes *continually*



1. The data stream, $\mathcal{S}$, presents a set of categories, $\mathcal{Y}_t$, to be learned.
2. Under a given computational budget, $\mathcal{C}_t$, the classifier $f_{\theta_{t-1}}$ is updated to f_{θ_t}.
3. To evaluate the learner, the stream $\mathcal{S}$ presents test samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ with y_i belonging to the collective set $\bigcup_{i=1}^{t} \mathcal{Y}_i$.

How to Get Around Manual Data Annotation?

Manual Annotation

IMAGENET

Stable Diffusion

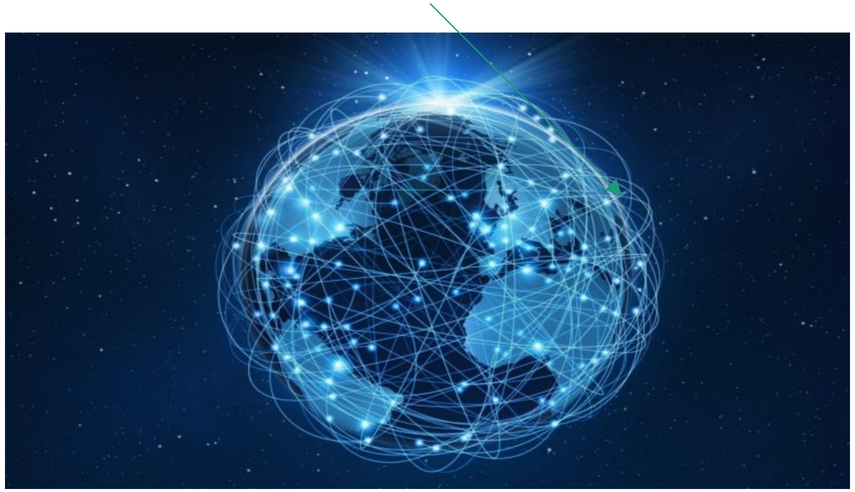
OpenAI

DALL-E

Retrieval from Source

LAION-5B

Large-scale Artificial Intelligence Open Network

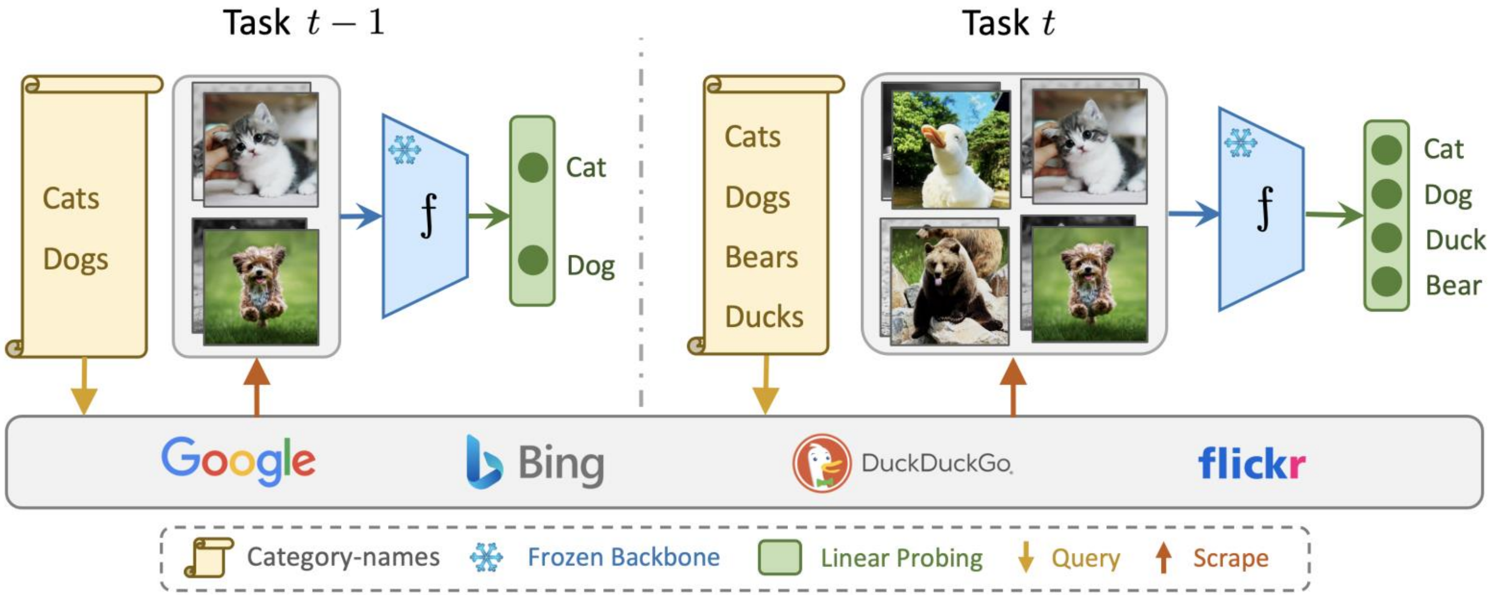


Internet: Billions of images uploaded **each day**

Static datasets are miniscule and out-of-date in comparison to the Internet!

Internet as a Huge, Continually Evolving Train Set

Idea: Download data from searching the Internet! [Fergus, et.al. 2005]



==> Go from Category List to Trained Classifier within minutes *continually!*

Internet as a Huge, Continually Evolving Train Set

UNIVERSITY OF OXFORD

Is it better than using Stable Diffusion or Retrieval From LAION5B?

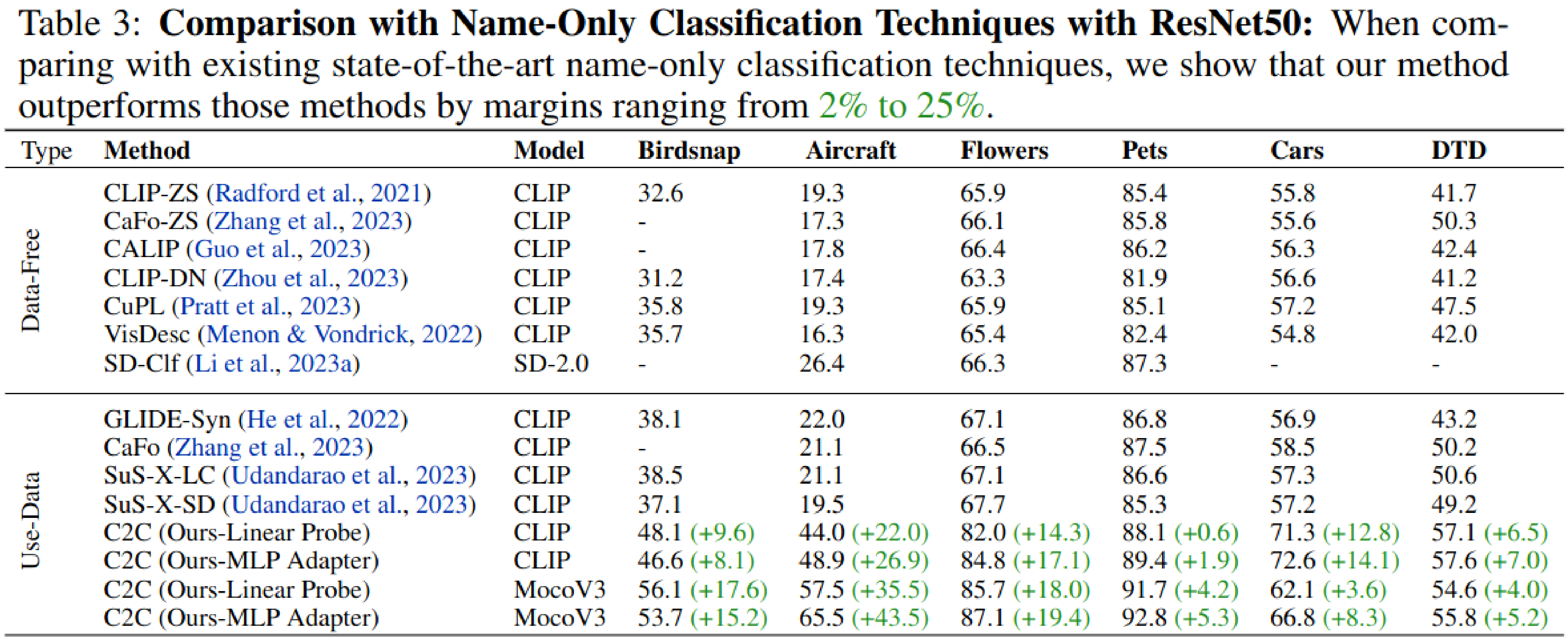


Table 3: **Comparison with Name-Only Classification Techniques with ResNet50:** When comparing with existing state-of-the-art name-only classification techniques, we show that our method outperforms those methods by margins ranging from 2% to 25%.

Type	**Method**	**Model**	**Birdsnap**	**Aircraft**	**Flowers**	**Pets**	**Cars**	**DTD**
Data-Free	CLIP-ZS (Radford et al., 2021)	CLIP	32.6	19.3	65.9	85.4	55.8	41.7
	CaFo-ZS (Zhang et al., 2023)	CLIP	-	17.3	66.1	85.8	55.6	50.3
	CALIP (Guo et al., 2023)	CLIP	-	17.8	66.4	86.2	56.3	42.4
	CLIP-DN (Zhou et al., 2023)	CLIP	31.2	17.4	63.3	81.9	56.6	41.2
	CuPL (Pratt et al., 2023)	CLIP	35.8	19.3	65.9	85.1	57.2	47.5
	VisDesc (Menon & Vondrick, 2022)	CLIP	35.7	16.3	65.4	82.4	54.8	42.0
	SD-Clf (Li et al., 2023a)	SD-2.0	-	26.4	66.3	87.3	-	-
Use-Data	GLIDE-Syn (He et al., 2022)	CLIP	38.1	22.0	67.1	86.8	56.9	43.2
	CaFo (Zhang et al., 2023)	CLIP	-	21.1	66.5	87.5	58.5	50.2
	SuS-X-LC (Udandarao et al., 2023)	CLIP	38.5	21.1	67.1	86.6	57.3	50.6
	SuS-X-SD (Udandarao et al., 2023)	CLIP	37.1	19.5	67.7	85.3	57.2	49.2
	C2C (Ours-Linear Probe)	CLIP	48.1 (+9.6)	44.0 (+22.0)	82.0 (+14.3)	88.1 (+0.6)	71.3 (+12.8)	57.1 (+6.5)
	C2C (Ours-MLP Adapter)	CLIP	46.6 (+8.1)	48.9 (+26.9)	84.8 (+17.1)	89.4 (+1.9)	72.6 (+14.1)	57.6 (+7.0)
	C2C (Ours-Linear Probe)	MocoV3	56.1 (+17.6)	57.5 (+35.5)	85.7 (+18.0)	91.7 (+4.2)	62.1 (+3.6)	54.6 (+4.0)
	C2C (Ours-MLP Adapter)	MocoV3	53.7 (+15.2)	65.5 (+43.5)	87.1 (+19.4)	92.8 (+5.3)	66.8 (+8.3)	55.8 (+5.2)

Better Prompts

SD/ LAION

Using Internet is Vastly Better!

Questions?

Anonymous Feedback

admonymous.co/bayesiankitten