EBERHARD KARLS UNIVERSITÄT TÜBINGEN • ellis European Laboratory for Learning and Intelligent Systems • imprs-is • • UNIVERSITY OF CAMBRIDGE

A Practitioner's Guide to Continual Multimodal Pretraining

Motivation: Foundation models are awesome, but can get outdated!

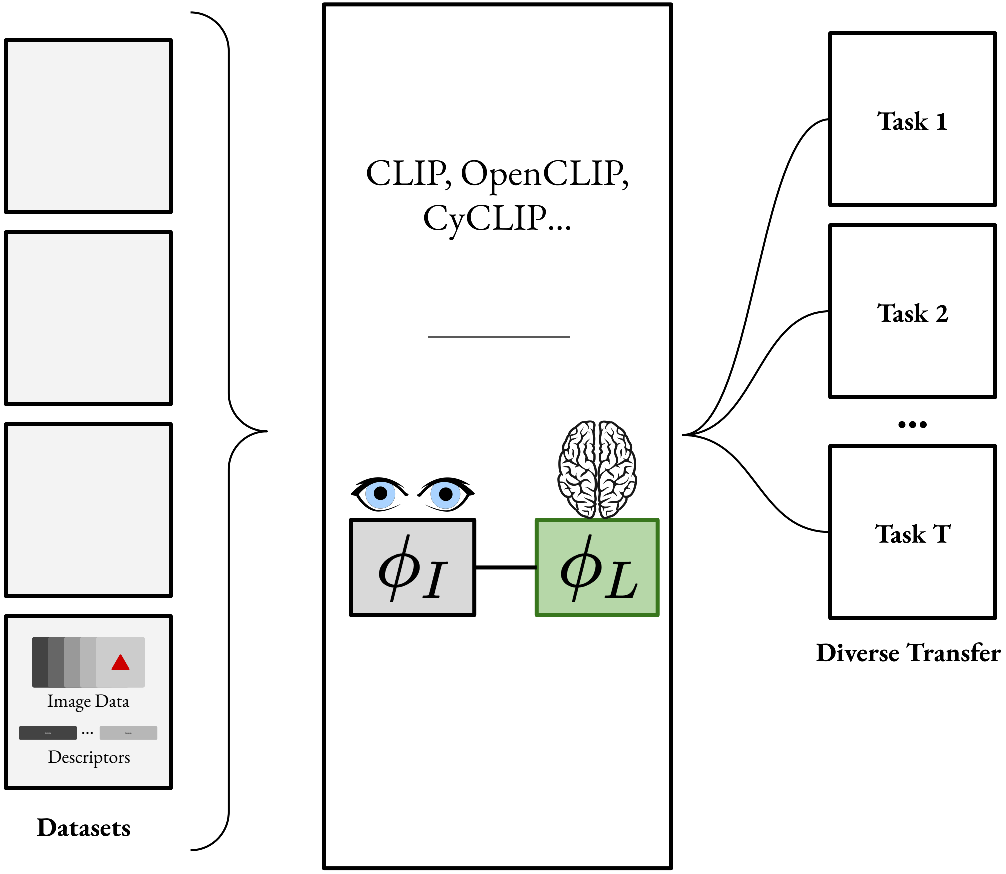


Image Data

Descriptors

Datasets

CLIP, OpenCLIP, CyCLIP...

ϕ_I

ϕ_L

Task 1

Task 2

...

Task T

Diverse Transfer

Large-scale Model Training on expansive data

Deploy across diverse tasks

- Zero-Shot Image Classification
- Any-Shot Retrieval
- Accelerate Cross-Modal Applications
- Guidance for Text-to-Image
- Transferable Insights

Motivation: Foundation models are awesome, but can get outdated!

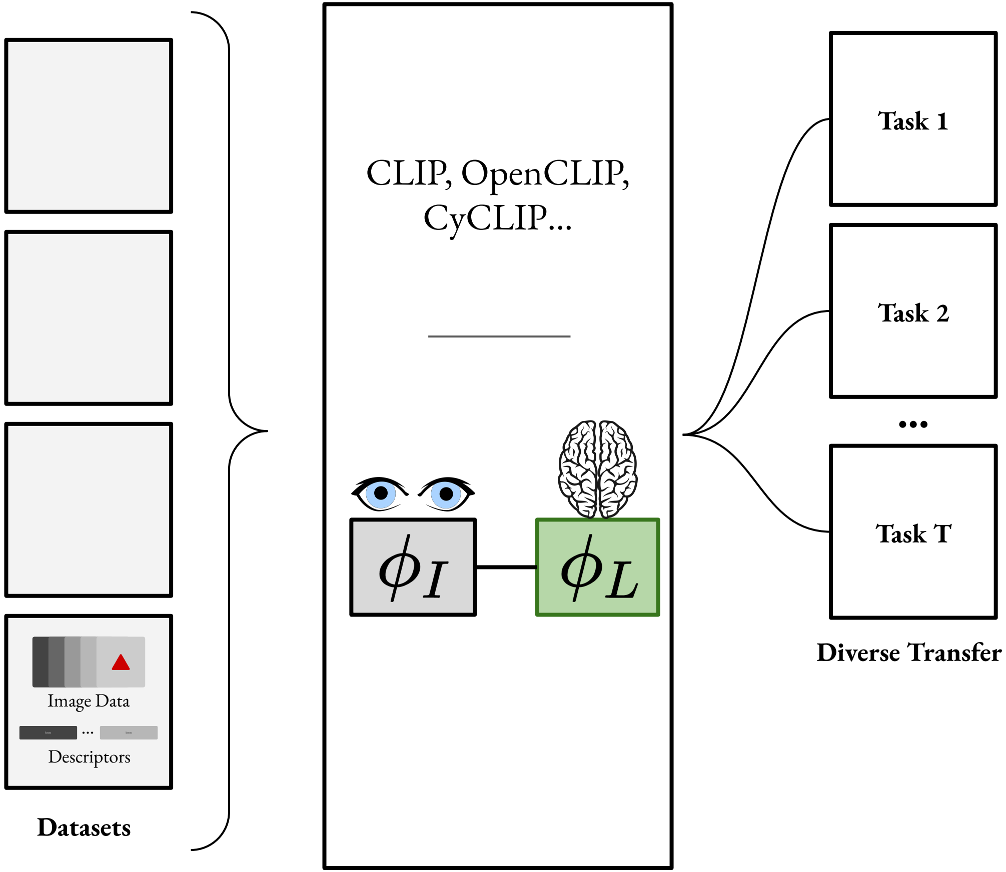


Image Data

Descriptors

Datasets

CLIP, OpenCLIP, CyCLIP...

ϕ_I

ϕ_L

Task 1

Task 2

...

Task T

Diverse Transfer

Large-scale Model Training on expansive data

Deploy across diverse tasks

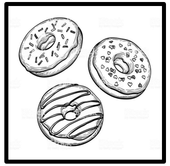
- Zero-Shot Image Classification
- Any-Shot Retrieval
- Accelerate Cross-Modal Applications
- Guidance for Text-to-Image
- Transferable Insights

But:

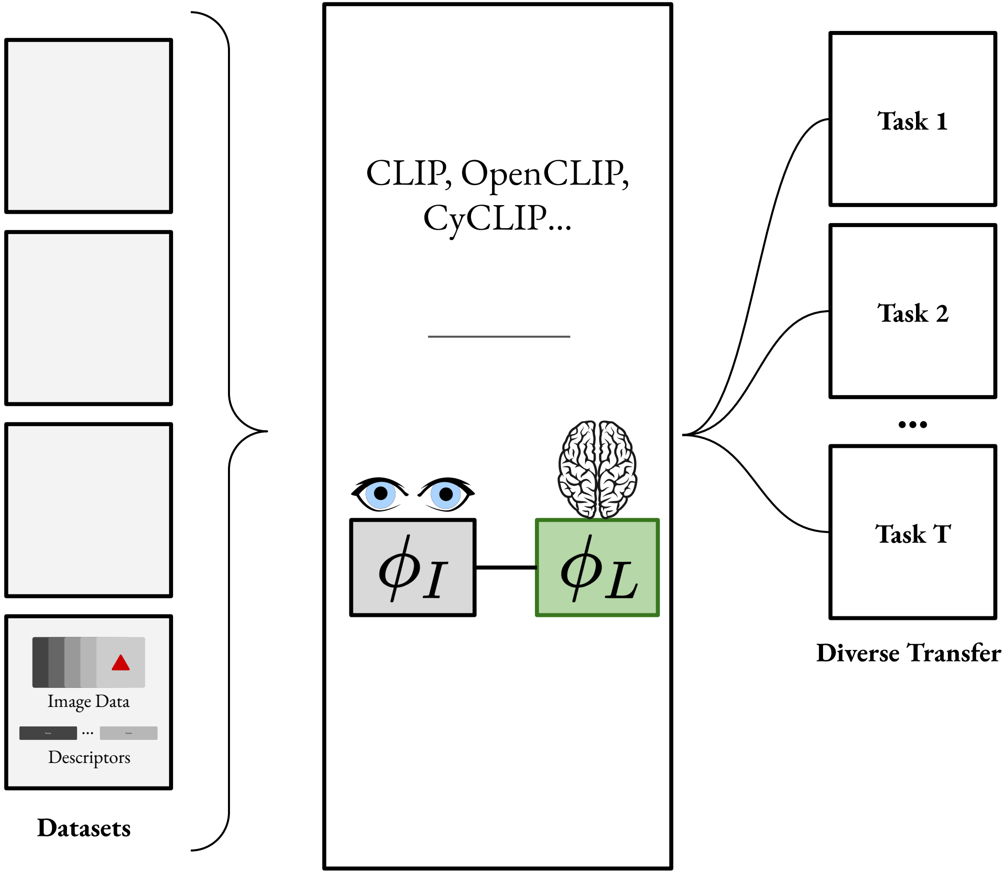
- Recency upper-bounded by dataset

Evergiven?

- New domains and semantic concepts



Motivation: Foundation models are awesome, but can get outdated!



Large-scale Model Training on expansive data

Deploy across diverse tasks

- Zero-Shot Image Classification
- Any-Shot Retrieval
- Accelerate Cross-Modal Applications
- Guidance for Text-to-Image
- Transferable Insights

But:

- Recency upper-bounded by dataset
- New domains and semantic concepts

Visual/Semantic Understanding needs to adapt!

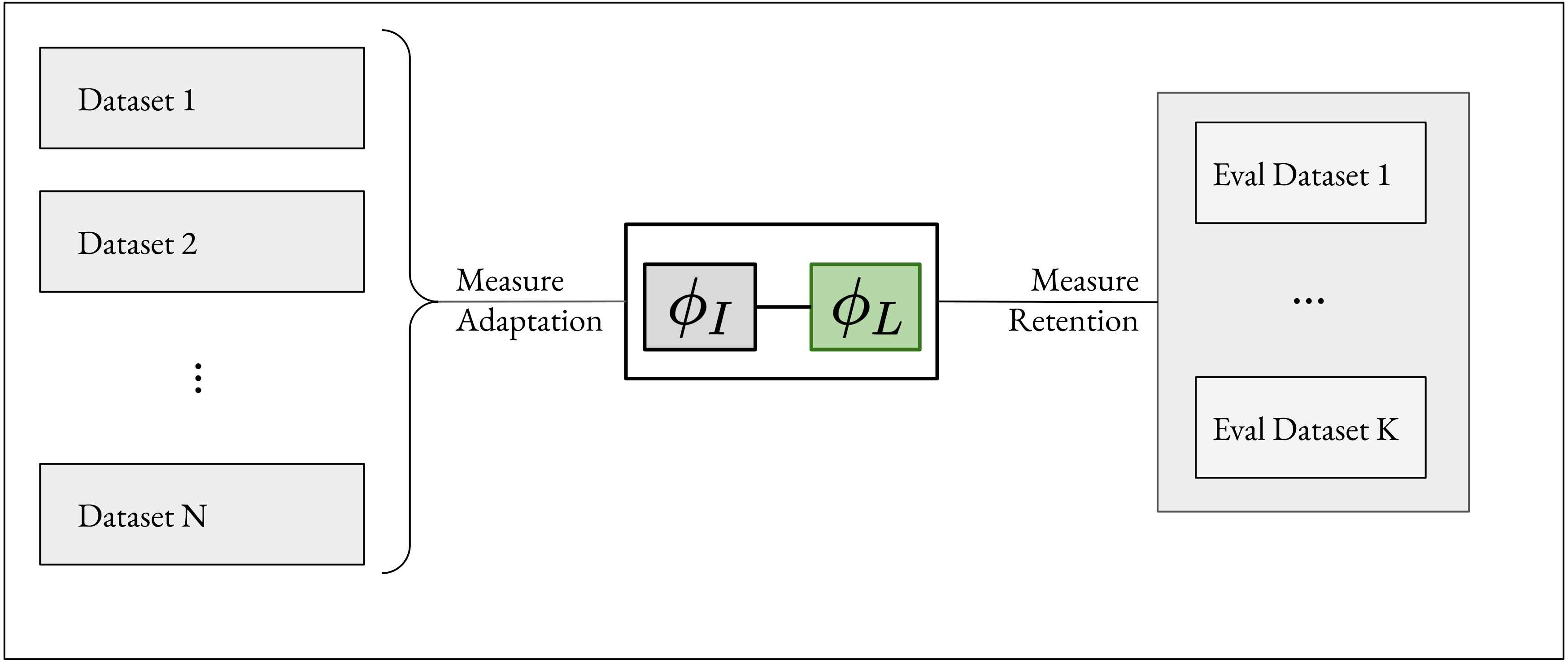
Can't retrain on bigger and bigger datasets!

How to continual pretrain across long update horizons?

Overview

Goal: Understand Continual Multimodal Pretraining

FoMo-in-Flux | Method Perspective | Pipeline Design | Data Perspective

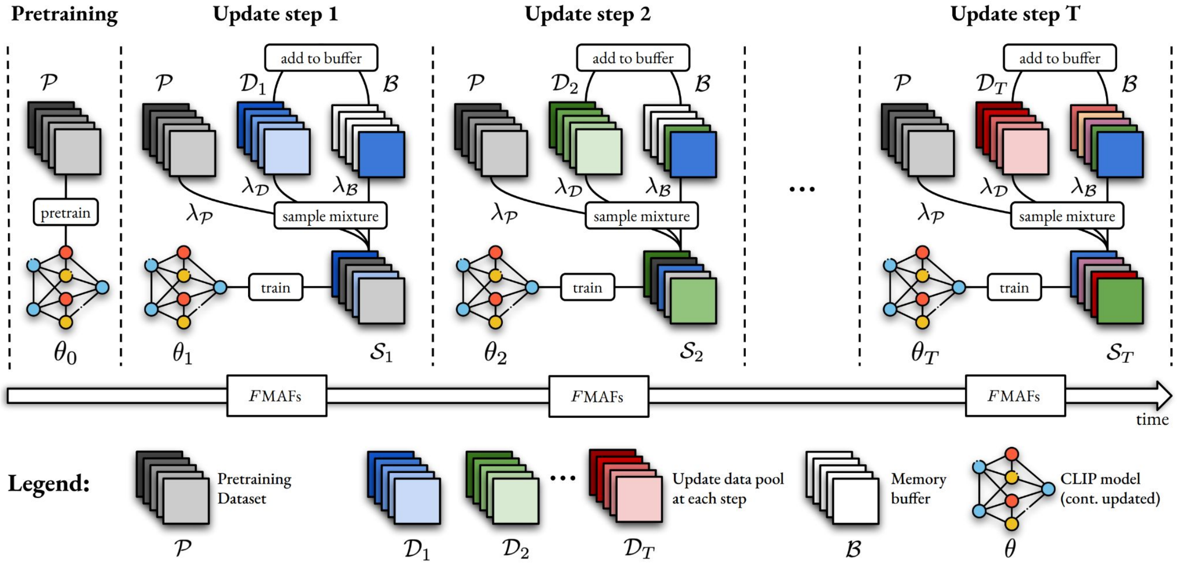


Overview

Goal: Understand Continual Multimodal Pretraining

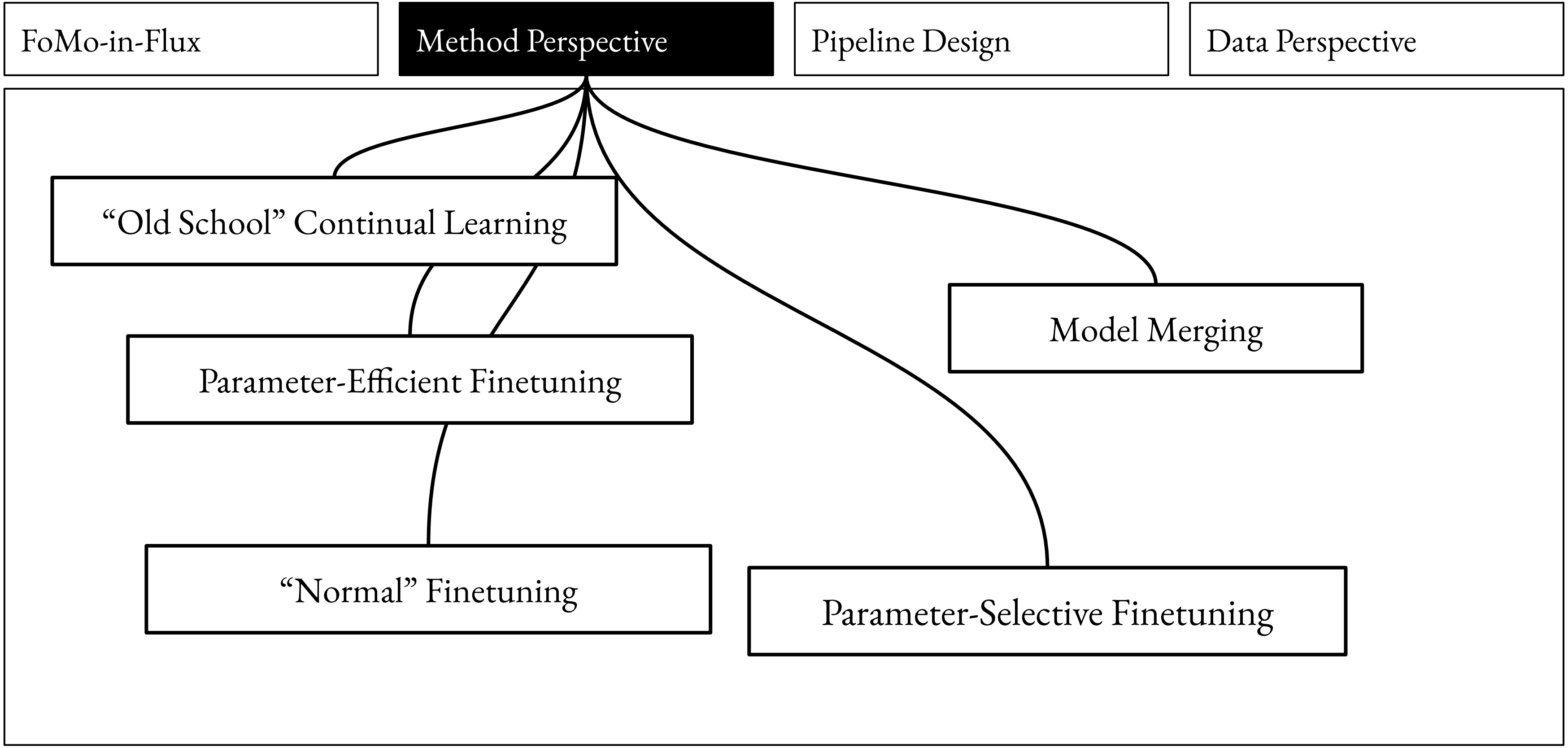
FoMo-in-Flux | Method Perspective | Pipeline Design | Data Perspective

Precise control over data stream & high conceptual coverage



Overview

Goal: Understand Continual Multimodal Pretraining



FoMo-in-Flux

Method Perspective

Pipeline Design

Data Perspective

“Old School” Continual Learning

Parameter-Efficient Finetuning

“Normal” Finetuning

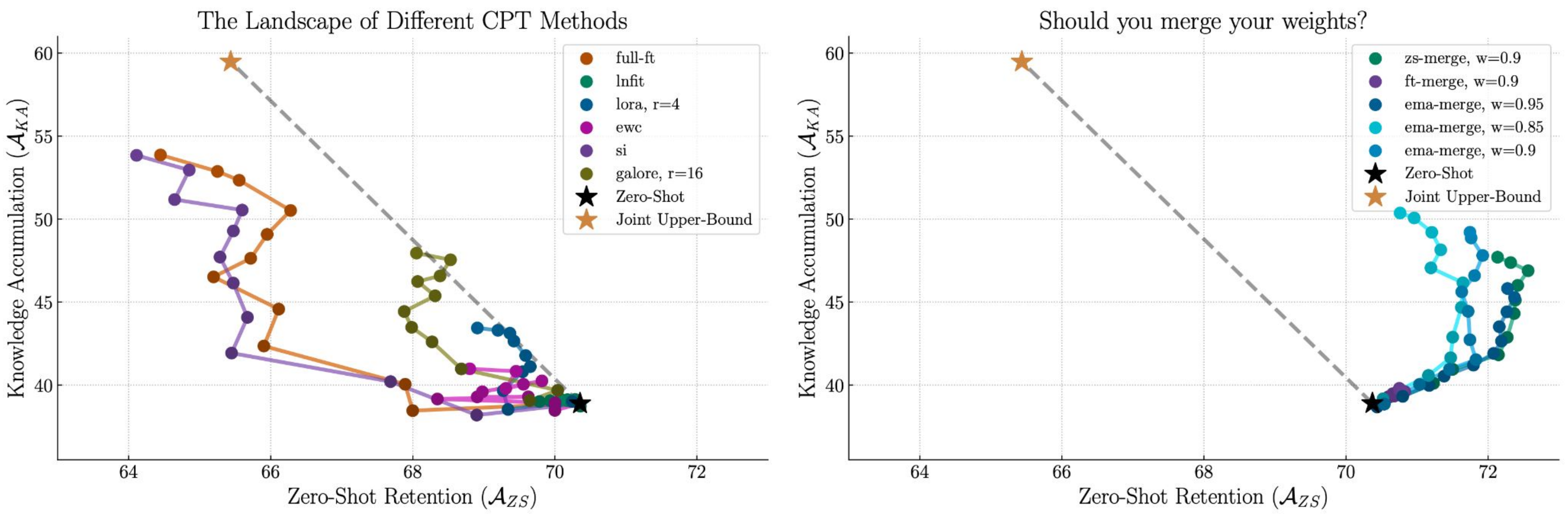
Model Merging

Parameter-Selective Finetuning

Overview

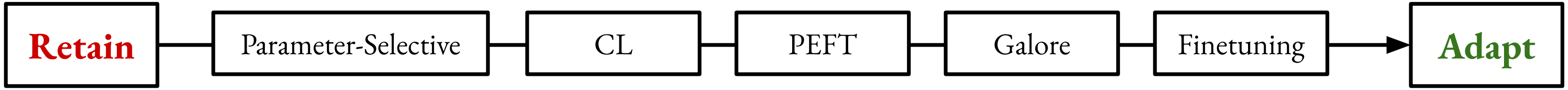
Goal: Understand Continual Multimodal Pretraining

FoMo-in-Flux | **Method Perspective** | Pipeline Design | Data Perspective



The Landscape of Different CPT Methods

Should you merge your weights?

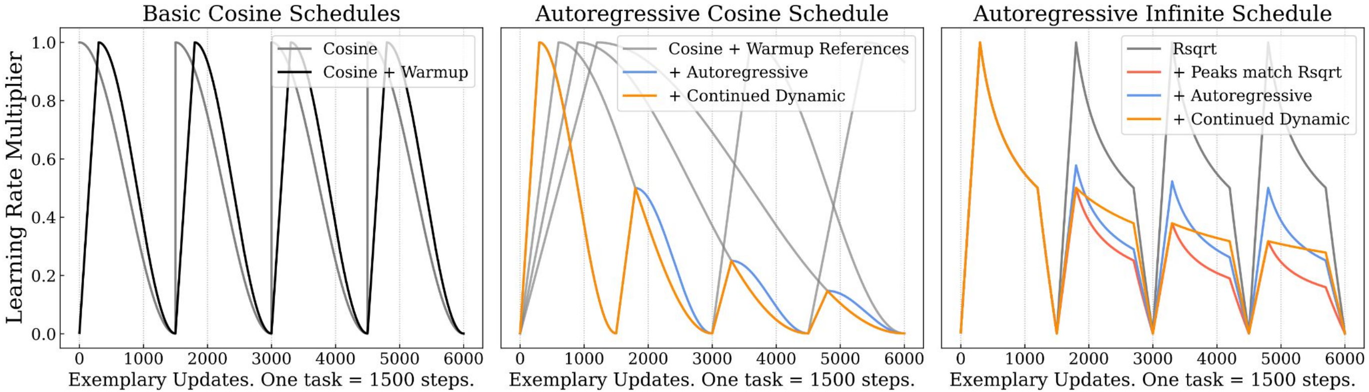


Retain — Parameter-Selective — CL — PEFT — Galore — Finetuning → **Adapt**

Overview

Goal: Understand Continual Multimodal Pretraining

FoMo-in-Flux | Method Perspective | **Pipeline Design** | Data Perspective



Learning rates & schedules matter → Account for meta update steps!

Overview

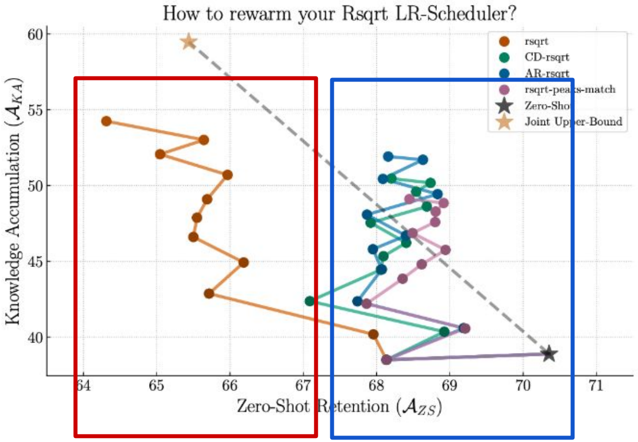
Goal: Understand Continual Multimodal Pretraining

FoMo-in-Flux

Method Perspective

Pipeline Design

Data Perspective

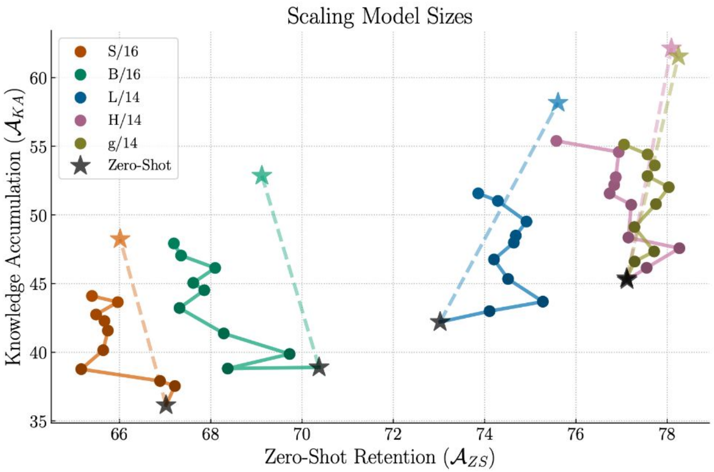


Base Scheduler Meta Schedules

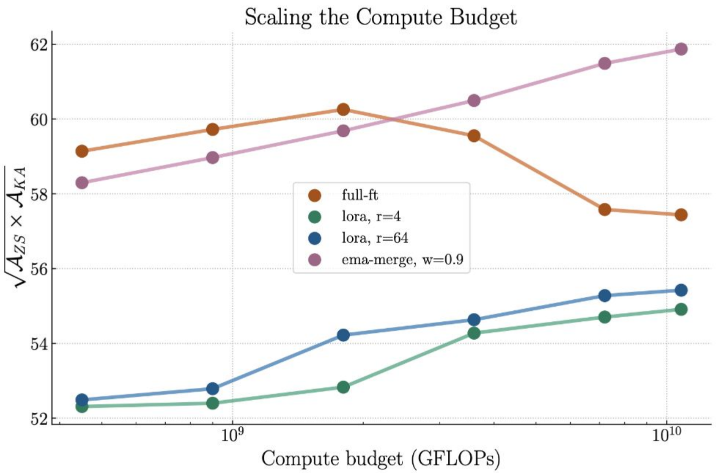
Overview

Goal: Understand Continual Multimodal Pretraining

FoMo-in-Flux | Method Perspective | **Pipeline Design** | Data Perspective



Larger models:
Easier to incorporate new knowledge without overwriting existing knowledge!

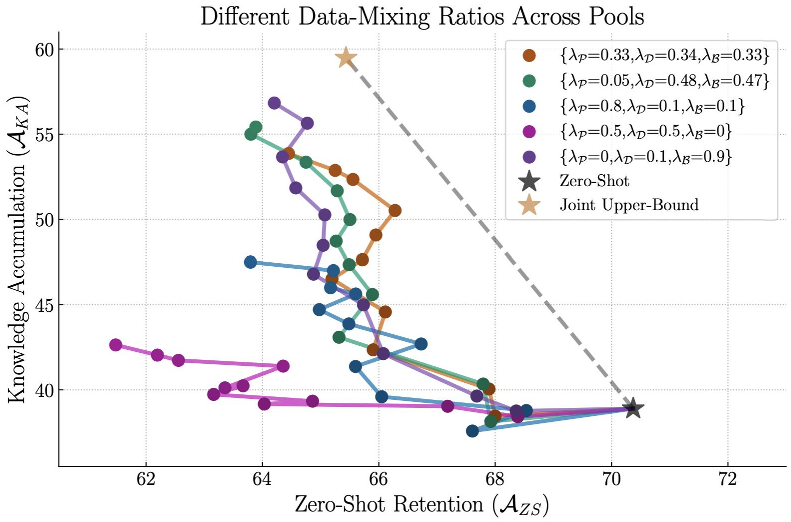


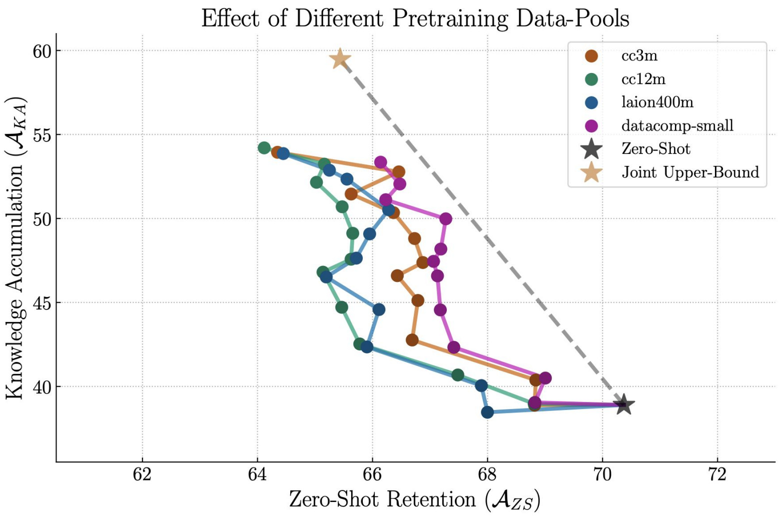
Higher compute allocation / update: Much more favourable scaling behaviour for model merging techniques!

Overview

Goal: Understand Continual Multimodal Pretraining

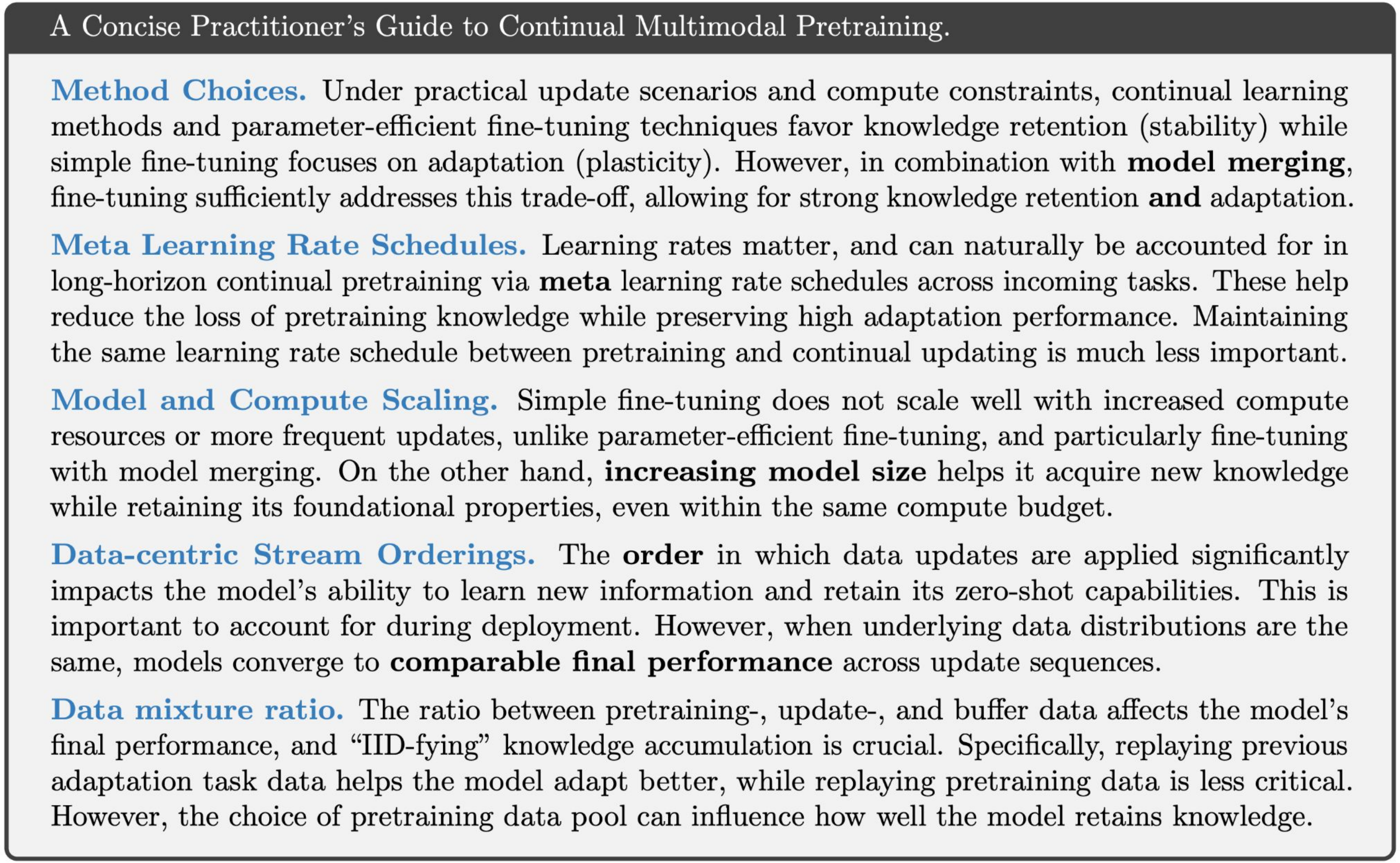
FoMo-in-Flux | Method Perspective | Pipeline Design | **Data Perspective**





- Replaying from pretraining data helps, but much less important than replay on new streamed data.
- How you replay from pretraining impacts trajectory.

Summary



A Concise Practitioner's Guide to Continual Multimodal Pretraining.

Method Choices. Under practical update scenarios and compute constraints, continual learning methods and parameter-efficient fine-tuning techniques favor knowledge retention (stability) while simple fine-tuning focuses on adaptation (plasticity). However, in combination with **model merging**, fine-tuning sufficiently addresses this trade-off, allowing for strong knowledge retention **and** adaptation.

Meta Learning Rate Schedules. Learning rates matter, and can naturally be accounted for in long-horizon continual pretraining via **meta** learning rate schedules across incoming tasks. These help reduce the loss of pretraining knowledge while preserving high adaptation performance. Maintaining the same learning rate schedule between pretraining and continual updating is much less important.

Model and Compute Scaling. Simple fine-tuning does not scale well with increased compute resources or more frequent updates, unlike parameter-efficient fine-tuning, and particularly fine-tuning with model merging. On the other hand, **increasing model size** helps it acquire new knowledge while retaining its foundational properties, even within the same compute budget.

Data-centric Stream Orderings. The **order** in which data updates are applied significantly impacts the model's ability to learn new information and retain its zero-shot capabilities. This is important to account for during deployment. However, when underlying data distributions are the same, models converge to **comparable final performance** across update sequences.

Data mixture ratio. The ratio between pretraining-, update-, and buffer data affects the model's final performance, and "IID-fying" knowledge accumulation is crucial. Specifically, replaying previous adaptation task data helps the model adapt better, while replaying pretraining data is less critical. However, the choice of pretraining data pool can influence how well the model retains knowledge.