How does Continual Learning affect FLAVA?

Presented by

Jatin Karthik (313301)

27 July 2023

Supervisor:

Dr. Benny Stein

Supervisor:

Prof. Dr. Niels Landwehr

Dr. Ujjwal

Outline

Introduction

Foundation Models

➔ Models trained on a broad set of unlabeled data that can be used for different tasks, with minimal fine-tuning are termed as foundation models [1].

➔ Since they are trained on plethora of data, used for many applications.

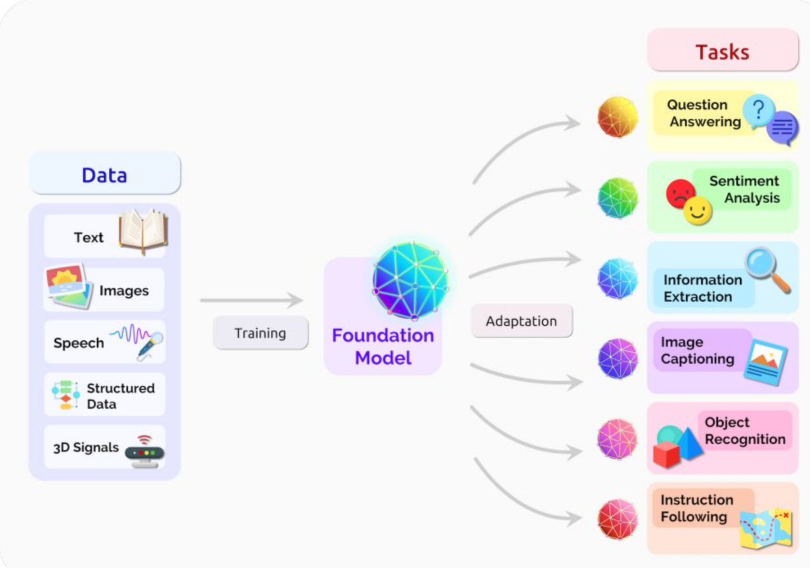
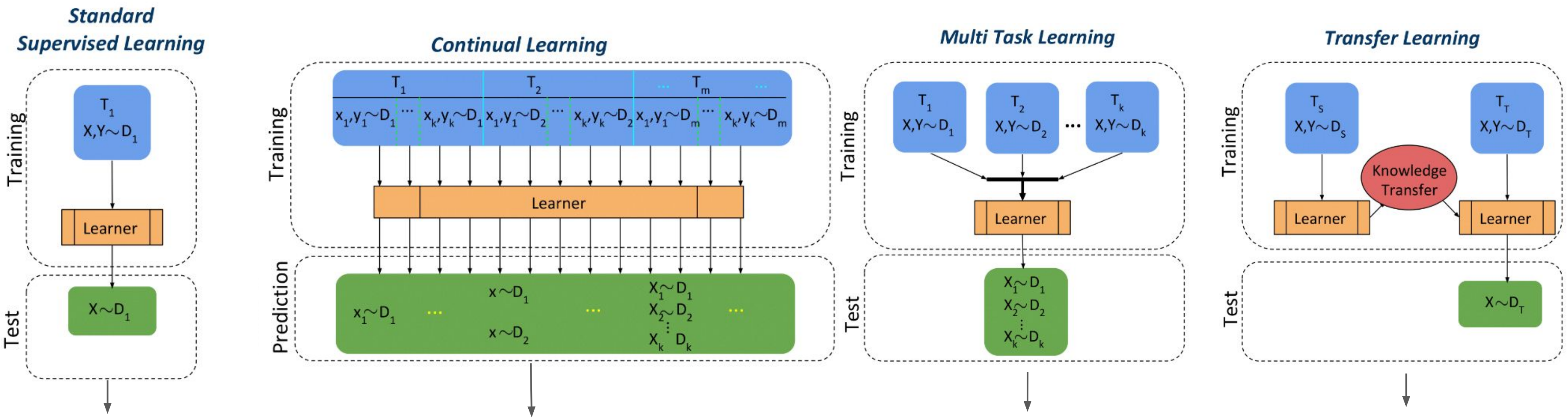


Figure 1: Basic Layout of Foundational model (Source [1])

Continual Learning



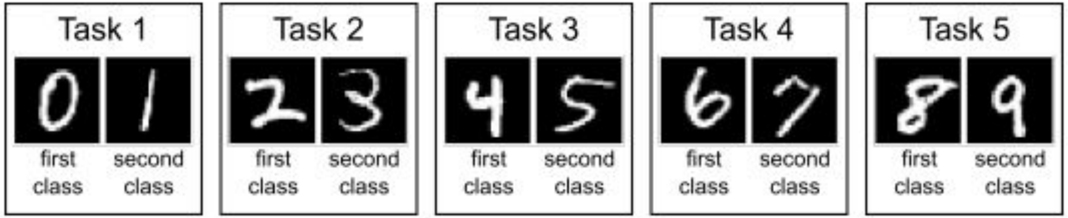
One Task & Data available at the same time.

- Multiple Tasks
- Data arrives incrementally
- Goal: all tasks

- Multiple Tasks
- Data available at the same time
- Goal: all tasks

- Multiple Tasks
- Data arrives incrementally
- Goal: last task

Continual Learning Settings



Task-IL	With task given, is it the 1st or 2nd class? (e.g., 0 or 1)
Domain-IL	With task unknown, is it a 1st or 2nd class? (e.g., in $[0, 2, 4, 6, 8]$ or in $[1, 3, 5, 7, 9]$)
Class-IL	With task unknown, which digit is it? (i.e., choice from 0 to 9)

- Task-Incremental Learning (TIL): Tasks have disjoint data label spaces. Task identities are provided in both training and testing.
- Domain-Incremental Learning (DIL): Tasks have the same data label space but different input distributions. Task identities are not required.
- Class-Incremental Learning (CIL): Tasks have disjoint data label spaces. Task identities are only provided in training.

FLAVA Architecture

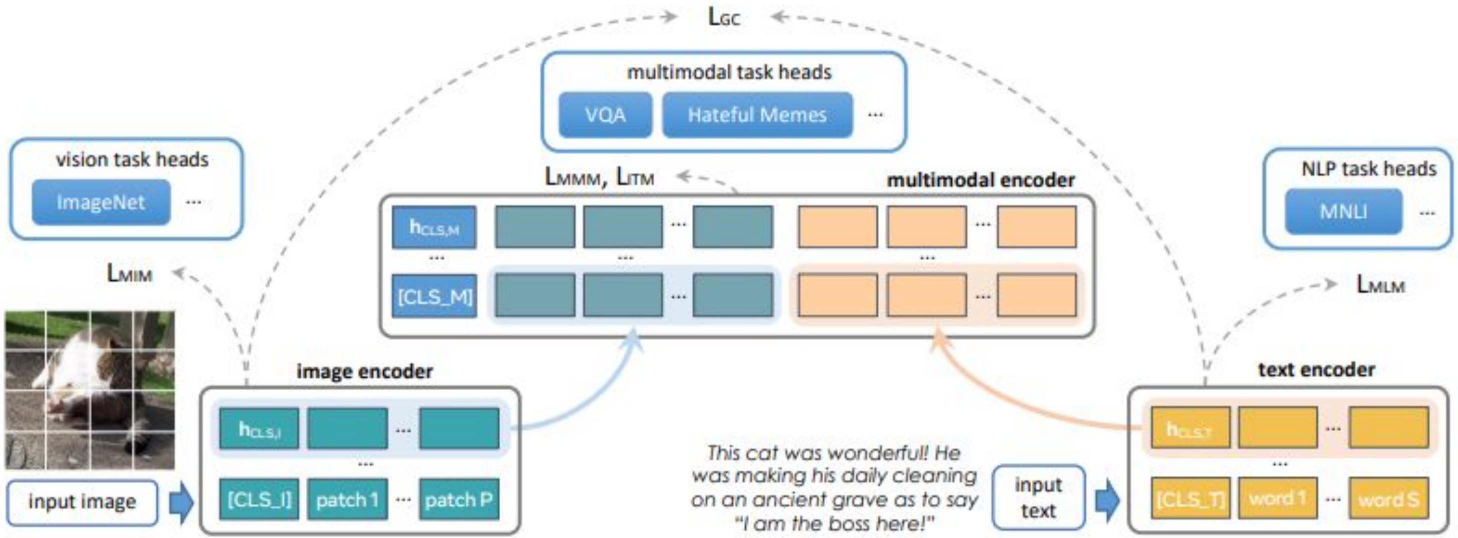


Figure 2: FLAVA Model Architecture

Foundational Model Comparison

Foundational Models	(Data size on image classification task)	No. of Parameters (all wrt image classification task)	Training Time	Inference Time
CLIP	400M	~ 3.6B	32 epochs on 400M data on ImageNet took 5600 GPU days	
Visual GPT-3		175B+	More than 7500 GPU days	
FLORENCE	900M	893M	10 days to train on 512 NVIDIA-A100 GPUs with 40GB memory per GPU	
FLIP	340M	2B - 12.8B	Similar setup as CLIP takes 2000 GPU days	
FLAVA	70M	350M	Less than CLIP and FLIP but exact time not mentioned	

Motivation

- Continual Learning helps deal with ever evolving data and tackle data drift and catastrophic forgetting.
- With the advent of foundation models, it is quintessential to check how these large models react to new data without forgetting; since retraining them is not feasible and fine tuning them is expensive. Motivation for the topic is to check how FLAVA responds to different CL methods.

Problem Formulation

- Consider FLAVA model architecture as F. Since FLAVA has two encoders i.e F(image) and F(text).
- Suppose we get a sequence of tasks $D=\{D_1, D_2, \dots, D_T\} \in$ Image classification task category data
- Then, supposing these tasks D_i are passed to F(image); then how does the task category accuracy or metrics change with and without CL?

Related Work and State-of-the-art work

Continual Learning (CL)

- Regularization based
  - Parameter Regularization
    - EWC
    - SI
  - Functional Regularization
    - LwF
    - FROMP
- Replay based
  - Experience Replay
  - Deep Generative Replay
- Architecture/Context based
  - Context dependent gating
  - iCARL

Related Work and State-of-the-art work

CL with CLIP (Vision Language Model)

- Extensive study of various CL methods on CLIP model. [2]
- CLIP has been evaluated on Zero-shot evaluation. [4]

CL on other Pretrained Models

- Visual Question-Answering (VQA) has been extensively researched on ViLT, VAuLT, ALBEF and FLAVA models on mostly Replay based methods. [3]
- Comparative analysis on 4 LLM's like BERT, GPT2 etc. on 4 CL methods in 2 incremental settings has been studied in [5].

Related Work and State-of-the-art work

CL with CLIP (Vision Language Model)

- Extensive study of various CL methods on CLIP model. [2]
- CLIP has been evaluated on Zero-shot evaluation. [4]

CL on other Pretrained Models

- Visual Question-Answering (VQA) has been extensively researched on ViLT, VAuLT, ALBEF and FLAVA models on mostly Replay based methods. [3]

Model	ViLT			VAuLT			FLAVA			ALBEF		
	Acc	BWT	FWT	Acc	BWT	FWT	Acc	BWT	FWT	Acc	BWT	FWT
Sequential	26.82±2.29	-42.49±3.41	-0.05±0.06	26.21±4.79	-42.58±5.64	-0.14±0.24	26.61±2.78	-34.31±2.36	-0.02±0.40	45.71±3.43	-27.13±3.69	9.77±2.83
ER	54.15±1.36	-12.38±1.82	**0.03±0.06**	**51.51±0.91**	**-12.67±1.07**	0.08±0.12	44.52±0.80	**-10.53±1.14**	-0.11±0.18	**60.79±0.54**	**-9.77±1.02**	11.90±2.39
DER	51.42±1.71	-12.56±1.58	-0.15±0.38	49.35±1.29	-14.86±1.63	-0.15±0.41	**44.82±1.09**	-10.91±2.75	-0.07±0.16	51.49±2.08	-21.18±2.29	**12.48±1.26**
DERPP	**54.21±1.31**	**-12.34±1.70**	0.00±0.28	51.30±1.03	-13.08±1.12	**0.23±0.31**	44.52±1.57	-11.30±2.64	-0.06±0.33	59.84±0.97	-10.89±1.27	11.91±3.14
EWC	26.94±2.75	-42.86±3.96	0.03±0.06	25.43±3.79	-43.64±4.02	-0.03±0.60	25.83±4.89	-34.09±5.87	**-0.01±0.18**	46.57±5.82	-27.67±7.07	9.62±3.53

Figure 4: Table shows comparison of different VLMs on several CL methods showing scope of experimentation (Source[3])

Proposed Approach

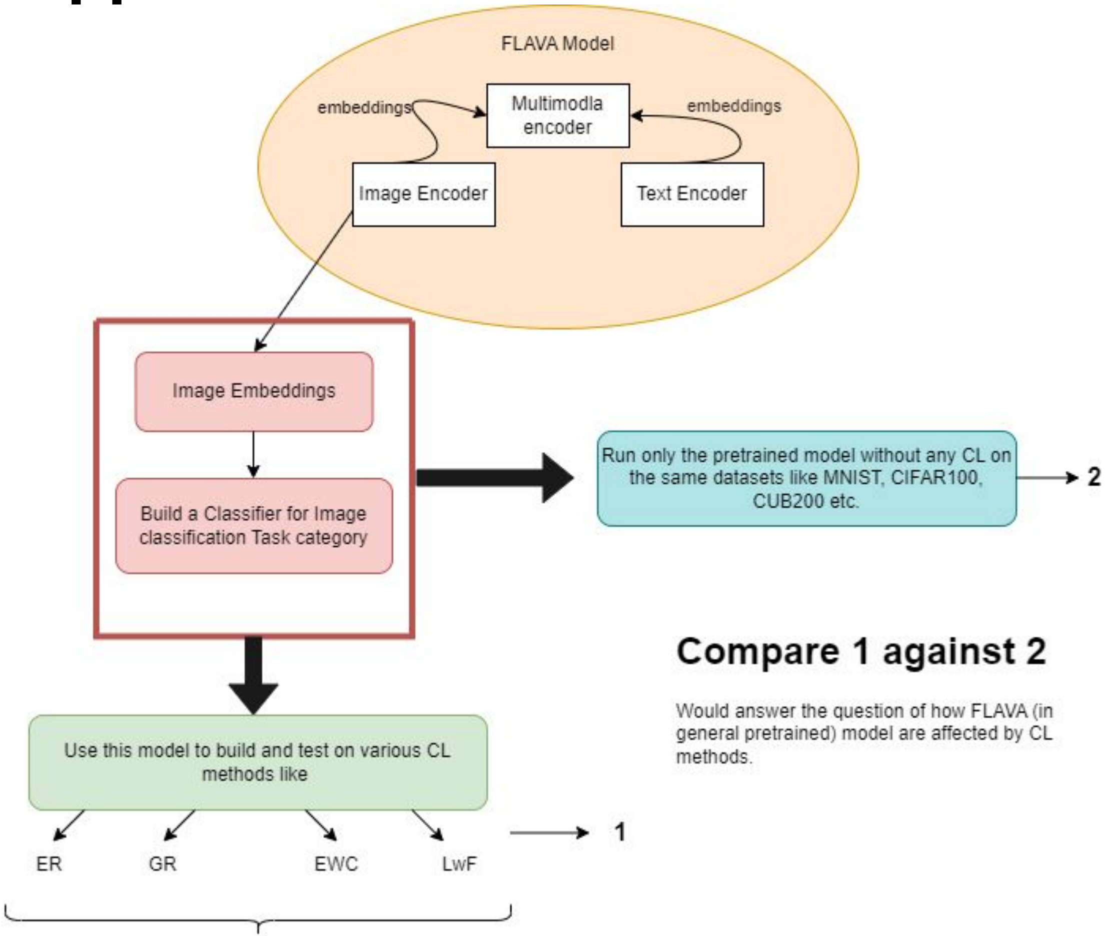
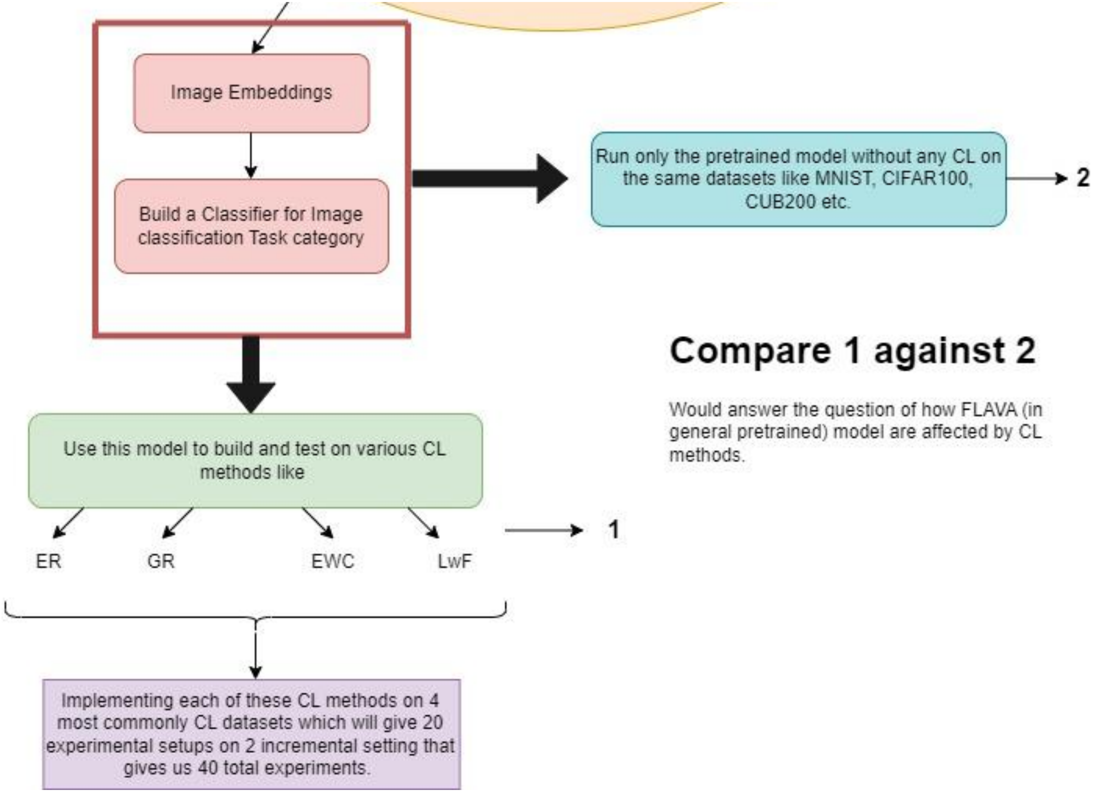


Figure 4: Proposed Approach

Compare 1 against 2

Would answer the question of how FLAVA (in general pretrained) model are affected by CL methods.

Proposed Approach



Compare 1 against 2

Would answer the question of how FLAVA (in general pretrained) model are affected by CL methods.

Figure 4: Proposed Approach

Data Foundation

Datasets that I intend to use are splitMNIST, splitCIFAR10, splitCIFAR100 and CUB200.

splitCIFAR10-: the suggested train and test split, where each category has 500 training and 100 test data.

CUB200: CUB200 contains 5594 training (~30 per category) and 6194 test images for 200 different bird species.

Reasons:

- Most of CL literature use these datasets for comparison and they are open source.
- This would help create a level ground while evaluating results. Therefore, we can compare FLAVA on these datasets against FLAVA_with_CL.

Evaluation Metrics

1. UT- Accuracy: Updated model accuracy via CL method's accuracy.

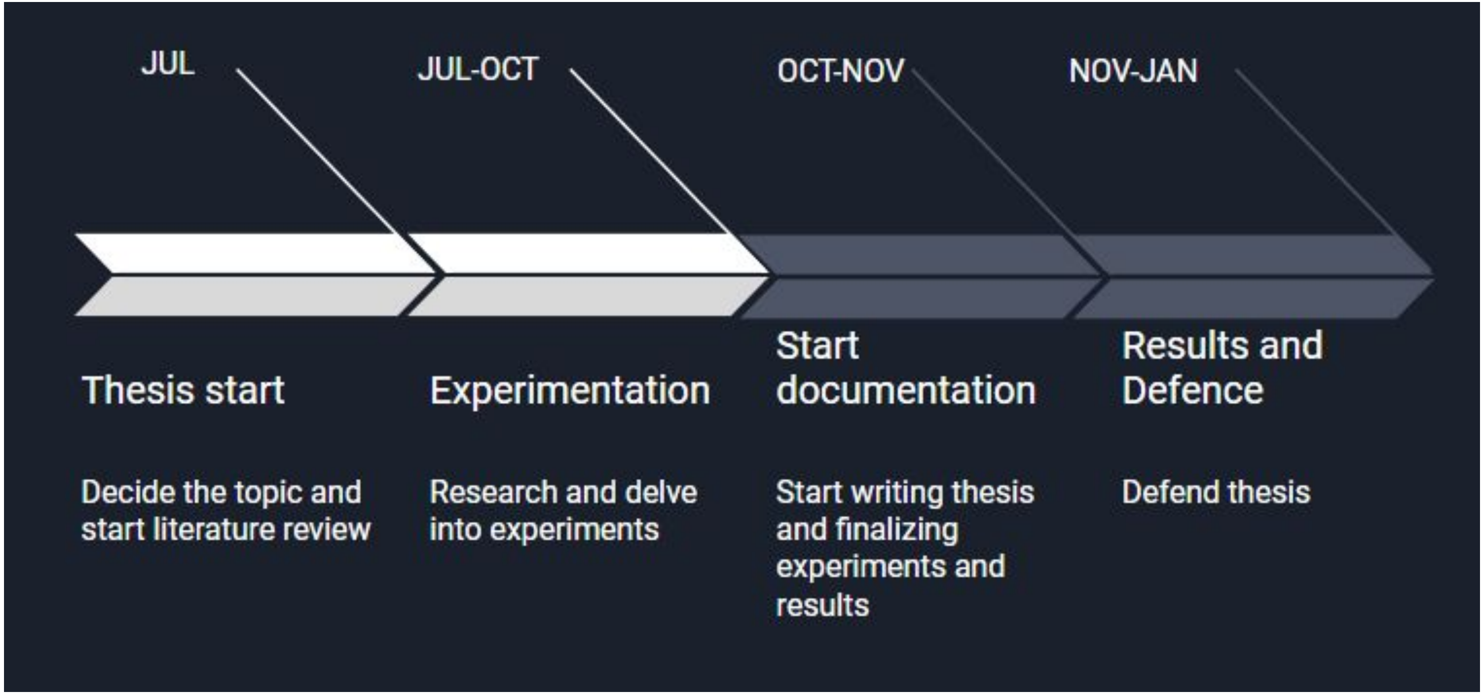
2. Forgetting measure: Here, $a_{j,i}$ is the performance of task t_i after training on task t_j.

$$F_{\mathcal{T}} = \frac{1}{\mathcal{T}-1}\sum_{i=1}^{\mathcal{T}-1} f_i^{\mathcal{T}}, \qquad f_i^{\mathcal{T}} = \max_{k\in\{1,\ldots,j-1\}} a_{k,i} - a_{j,i} \qquad \forall\ i<j$$

3. Forward Transfer: where b_i = test accuracy for task i at random initialization

$$FT_{\mathcal{T}} = \frac{1}{\mathcal{T}-1}\sum_{i=2}^{\mathcal{T}} a_{i-1,i} - b_i$$

Timeline


JUL
JUL-OCT
OCT-NOV
NOV-JAN
Thesis start
Decide the topic and start literature review
Experimentation
Research and delve into experiments
Start documentation
Start writing thesis and finalizing experiments and results
Results and Defence
Defend thesis

Goals

- Compare non-memory based CL methods against memory based ones (on Image Classification Task category only; on different incremental settings).

- Analyzing and comparing these two main categories of CL methods on FLAVA will help us decide upon whether a single frozen component like Image encoder of FLAVA is better performing with or without CL on zero shot evaluation.

- Identify the bottlenecks of CL in VLM’s (Vision Language Models) and specify the reasons of failure or success depending on the experiments.

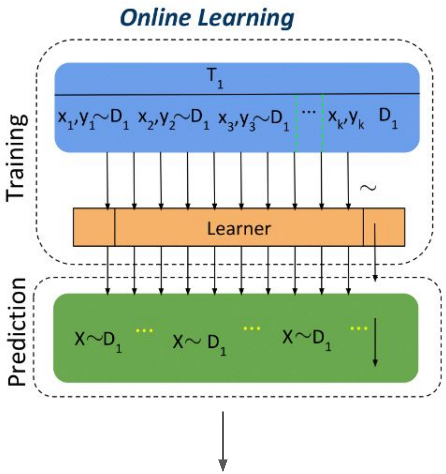
References

1. Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258, 2021.
2. Yuxuan Ding, Lingqiao Liu, Chunna Tian, Jingyuan Yang, and Haoxuan Ding. Don't stop learning: Towards continual learning for the clip model. arXiv preprint arXiv:2207.09248, 2022.
3. Yao Zhang, Haokun Chen, Ahmed Frikha, Yezi Yang, Denis Krompass, Gengyuan Zhang, Jindong Gu, and Volker Tresp. Cl-crossvqa: A continual learning benchmark for cross-domain visual question answering. arXiv preprint arXiv:2211.10567, 2022.
4. Thengane, Vishal, Salman Khan, Munawar Hayat, and Fahad Khan. "Clip model is an efficient continual learner." *arXiv preprint arXiv:2210.03114* (2022).
5. Wu, Tongtong, Massimo Caccia, Zhuang Li, Yuan-Fang Li, Guilin Qi, and Gholamreza Haffari. "Pretrained language model in continual learning: A comparative study." In *International Conference on Learning Representations*. 2021.

Thank You

Appendix slides

Online Learning



- One task
- Data arrives incrementally

More Details about CL methods:

1. Regularization based approach: Also known as prior-based approaches, these methods prevent significant updates by adding a penalty to encourage the model to stay close to its previous version.
    a. Parameter regularization: parameters important for past tasks are encouraged not to change too much when learning a new task.
    b. Functional regularization: The input-output mapping learnt previously is encouraged not to change too much at a particular set of inputs (anchor points) aka knowledge distillation.
2. Replay based approach: Current training data is complemented with data representative of past observations. The replayed data is sampled usually from a memory buffer or a generative model.

More Details about CL methods:

3. Architecture/ Context based approach: Also known as parameter-isolation methods, this family of algorithm alleviates forgetting by using different subset of parameters for fitting different tasks. A template is learnt for each class, and classification is performed based on which template is most suitable for sample to be classified.