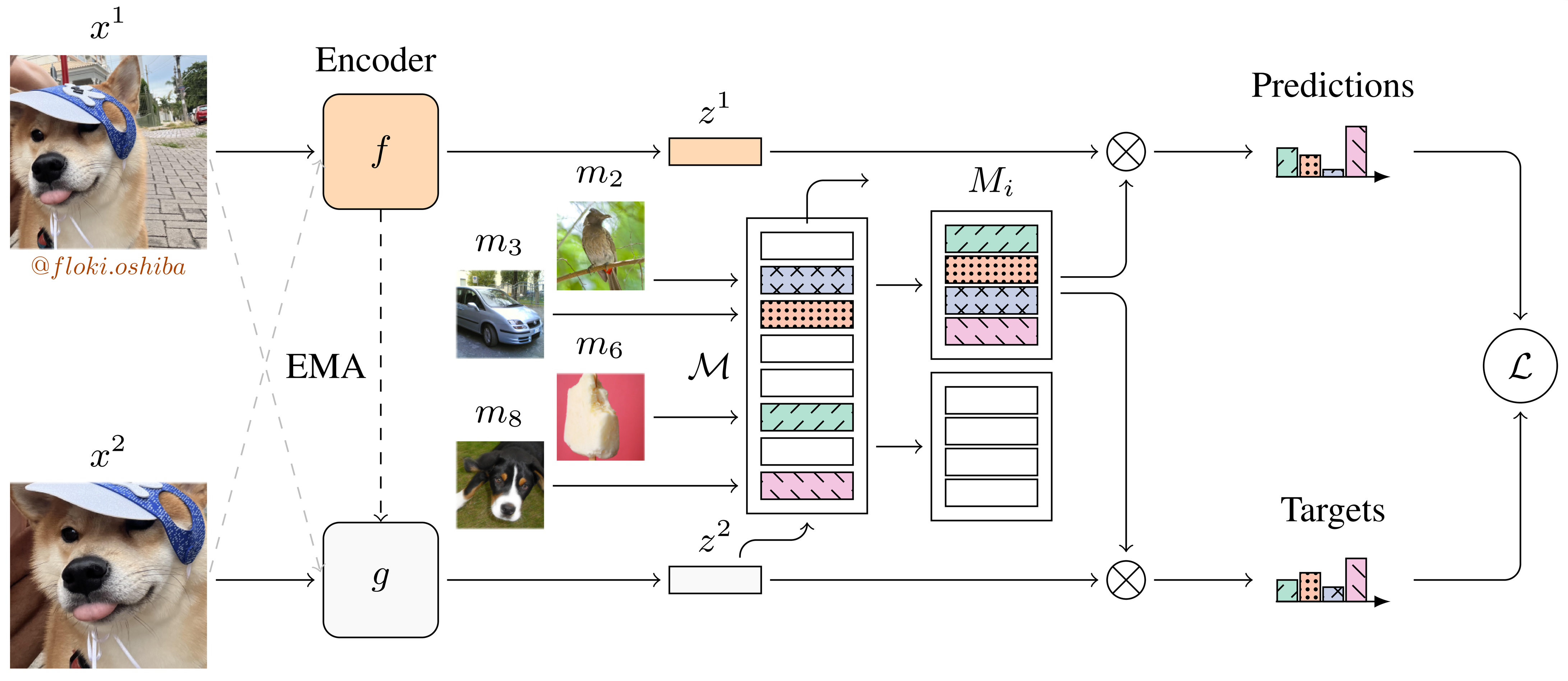
Learning from Memory: Non-Parametric Memory Augmented Self-Supervised Learning of Visual Features

Thalles Silva ✇ Helio Pedrini ✇ Adín Ramírez Rivera ❄

✇ University of Campinas ❄ University of Oslo

Methodology

Goal Improve training stability of clustering-based SSL methods.



Motivation

Exploring the role of memory for self-supervised learning.

- Memory plays a crucial role in learning.
- When learning a new concept, we constantly compare what we see with previous experiences to gain insights and create analogies.

We propose:

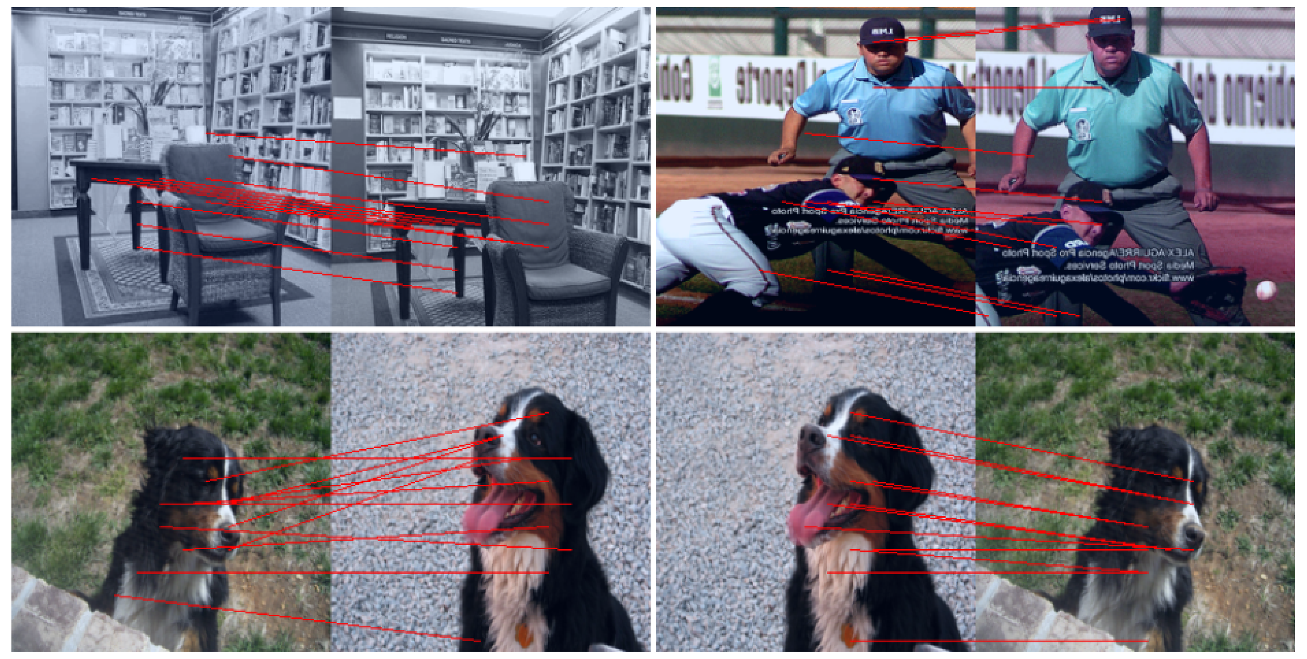
- An SSL method augmented with a non-parametric **memory**, $\mathcal{M}$, component to store representations from previously seen concepts.
- The memory is used to perform **multiple comparison-based tasks**.
  - Contrast the current image views against recollected representations from other images in memory.

Learning by Remembering!

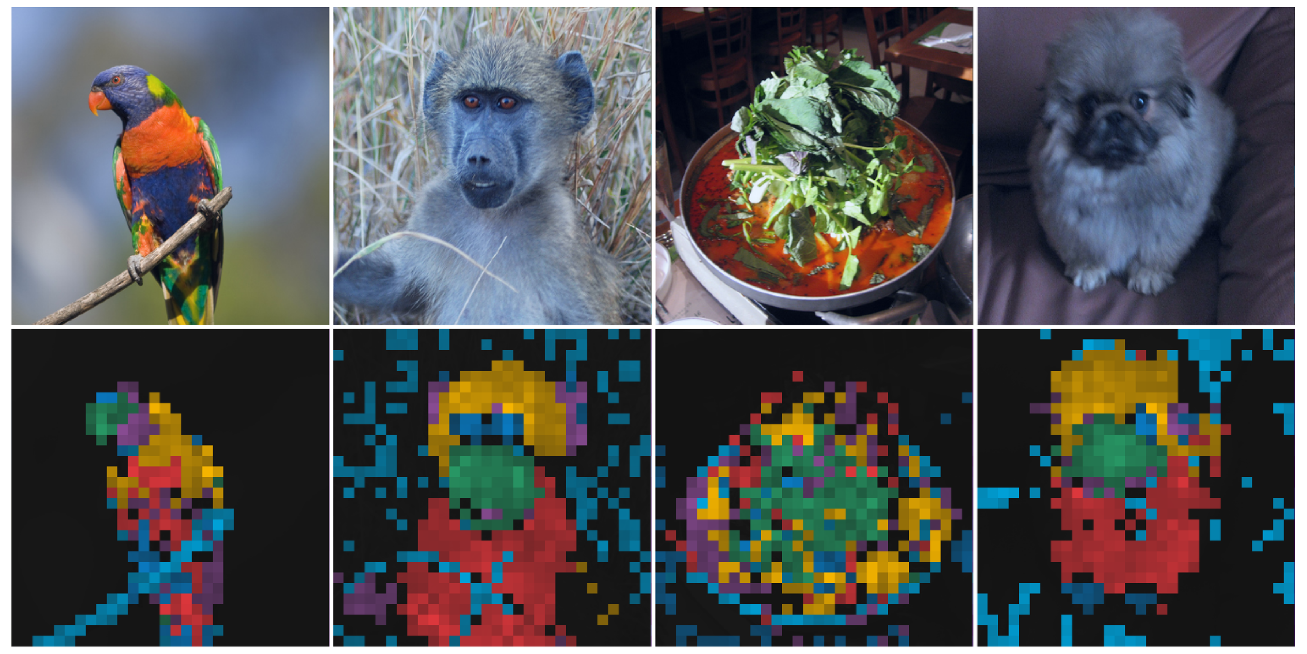
Follow the standard SSL pipeline.

1.- Create views from an image using random augmentations.
2.- Define **two encoder streams in a teacher-student setup**, where each stream consumes a different view.
3.- Pass the features to student and teacher encoders and **receive individual vector representations.**
4.- Sample a random memory block M_i and compare the views currently seen with the ones in the memory block.
5.- Take the resulting **probability distribution** relating the views to the concepts in the block and optimize for consistency.
6.- Update the memory with the current view's representation.

Sparse Feature Correspondence



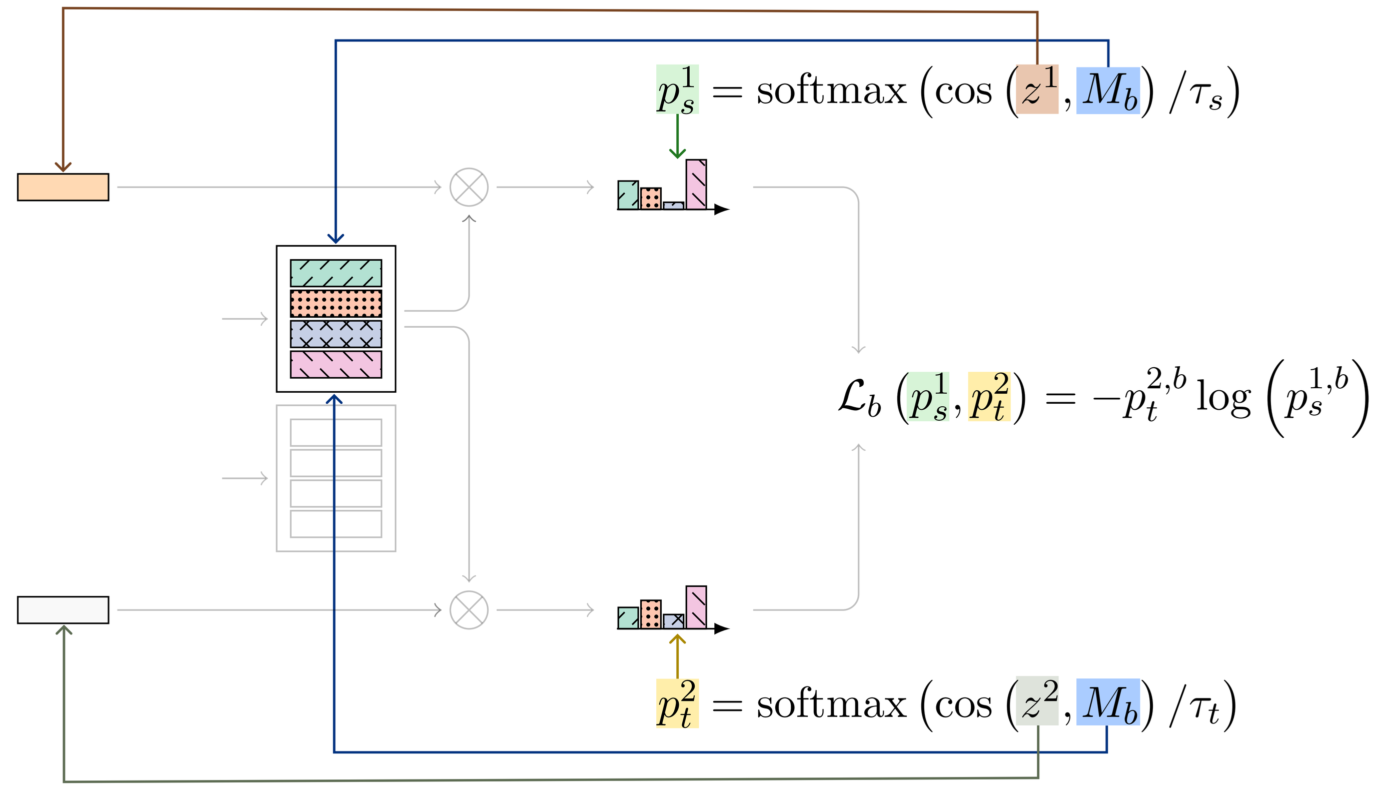
Visualizing Self-Attention Maps



Optimization Task

- Optimizing over random memory blocks regularizes training and naturally avoids mode collapse—**no need for extra regularizers**.

In math, you minimize this!



Results

Transfer learning (k-NN)

		Pets	Flowers	Aircraft	Cars	Country	Food	STL	GTSRB	Avg @k			
Methods	Epo.	results for $k=20$								10	20	100	200
MAE	800	19.4	16.9	9.7	6.0	5.0	11.9	64.6	27.6	20.9	20.1	16.9	15.2
MoCo-v3	300	83.8	70.2	27.4	22.4	14.3	64.5	97.5	56.1	55.3	54.5	52.2	51.3
DINO	800	90.1	84.6	38.5	32.7	**15.9**	70.7	98.9	64.7	62.0	62.0	60.8	60.2
iBOT	800	89.2	83.4	33.7	28.8	15.7	**72.6**	**99.0**	63.0	60.8	60.7	59.5	58.8
Ours	800	**91.6**	**84.6**	**41.1**	**33.3**	15.7	72.5	98.8	**69.3**	**63.3**	**63.4**	**62.4**	**61.8**

Low-shot classification on ImageNet-1M

Method	Arch	Protocol	1%	10%
DINO	VIT-B	k-NN	62.5	70.1
iBOT	VIT-B	k-NN	66.3	72.9
Ours	VIT-B	k-NN	**68.8**	**74.1**
DINO	VIT-B	Linear	66.2	74.2
iBOT	VIT-B	Linear	68.2	75.7
Ours	VIT-B	Linear	**70.4**	**76.4**
DINO	VIT-B	LogReg	67.1	74.2
iBOT	VIT-B	LogReg	69.6	75.9
Ours	VIT-B	LogReg	**71.3**	**76.3**

Image retrieval

			$\mathcal{R}$Ox		$\mathcal{R}$Par	
Method	Arch	Epo.	M	H	M	H
Sup	RN101	100	49.8	18.5	74.0	52.1
MoCo-v3	ViT-S	300	21.7	5.1	38.9	13.1
DINO	ViT-S	800	37.2	13.9	63.1	34.4
iBOT	ViT-S	800	36.6	13.0	61.5	34.1
Ours	ViT-S	800	**38.5**	**15.9**	**63.4**	**34.8**
MoCo-v3	ViT-B	300	30.5	8.6	54.3	23.5
DINO	ViT-B	400	37.4	13.7	63.5	35.6
iBOT	ViT-B	400	36.8	**14.3**	64.1	36.6
Ours	ViT-B	400	**39.3**	14.1	**65.8**	**38.1**

Lower-shot and long-tailed

	# images per class			ImNet-LT
	1	2	4	top-1
MoCo-v3	37.7± 0.3	47.8± 0.6	54.8± 0.2	56.7
DINO	39.2± 0.4	49.3± 0.8	57.6± 0.4	63.7
iBOT	42.2± 0.7	52.8± 0.3	60.6± 0.3	66.2
Ours	**44.8± 0.4**	**56.3± 0.3**	**63.8± 0.2**	**67.9**

Copy detection

Method	Arch	Epo.	mAP
DINO	VIT-S	800	**85.7**
iBOT	VIT-S	800	83.7
Ours	VIT-S	800	85.5
DINO	VIT-B	400	86.8
iBOT	VIT-B	400	84.2
Ours	VIT-B	400	**87.6**

The Forty-First International Conference on Machine Learning

Acknowledgements: Sigma2 NN8104K, RCN 309439, & CAPES 001.