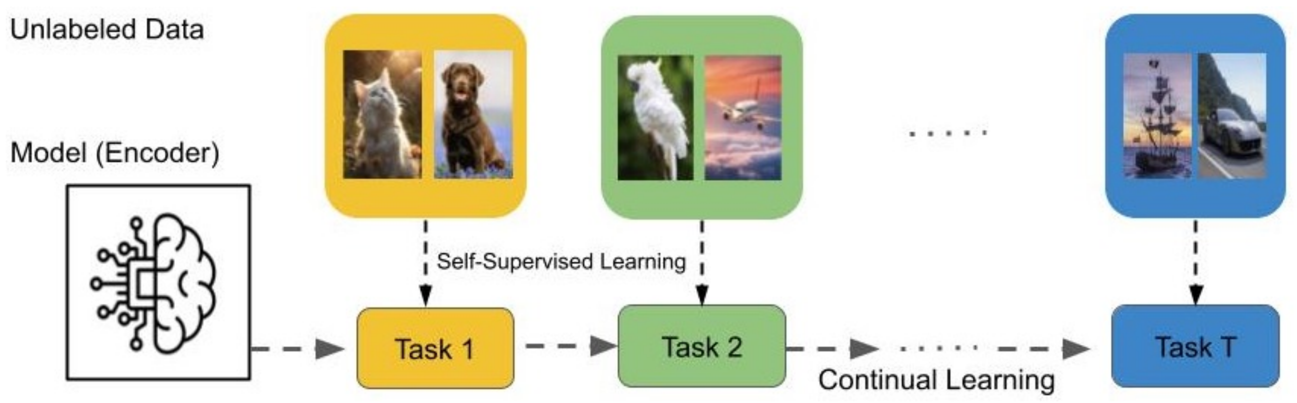
CroMo-Mixup: Augmenting Cross-Model Representations for Continual Self-Supervised Learning

USC Viterbi School of Engineering
Ming Hsieh Department of Electrical and Computer Engineering

Erum Mushtaq*, Duygu Nur Yaldiz*, Yavuz Bakman*, Jie Ding⋆, Chenyang Tao†, Dimitrios Dimitriadis†, Salman Avestimehr*

*University of Southern California, ⋆ University of Minnesota, † Amazon AI

Introduction

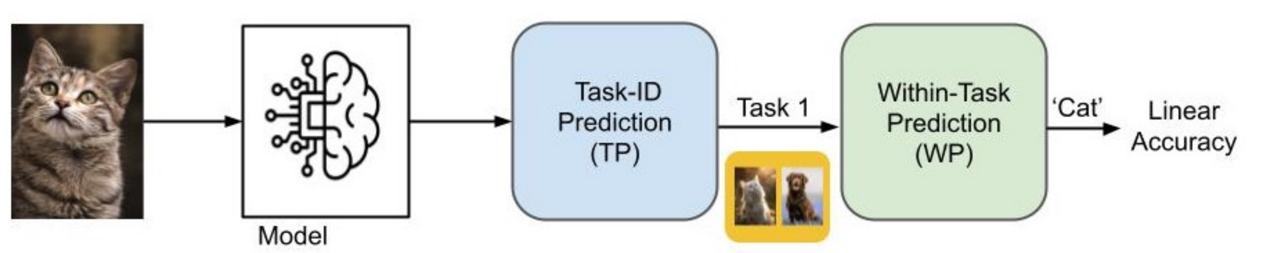


Continual Self-Supervised Learning (CSSL):

- ❑ Tasks arrive in an online manner
- ❑ Training data is unlabeled
- ❑ Classes are mutually exclusive across tasks

Evaluation:

❑ At the end of training, a linear classifier layer with the frozen encoder is trained on the labeled data.



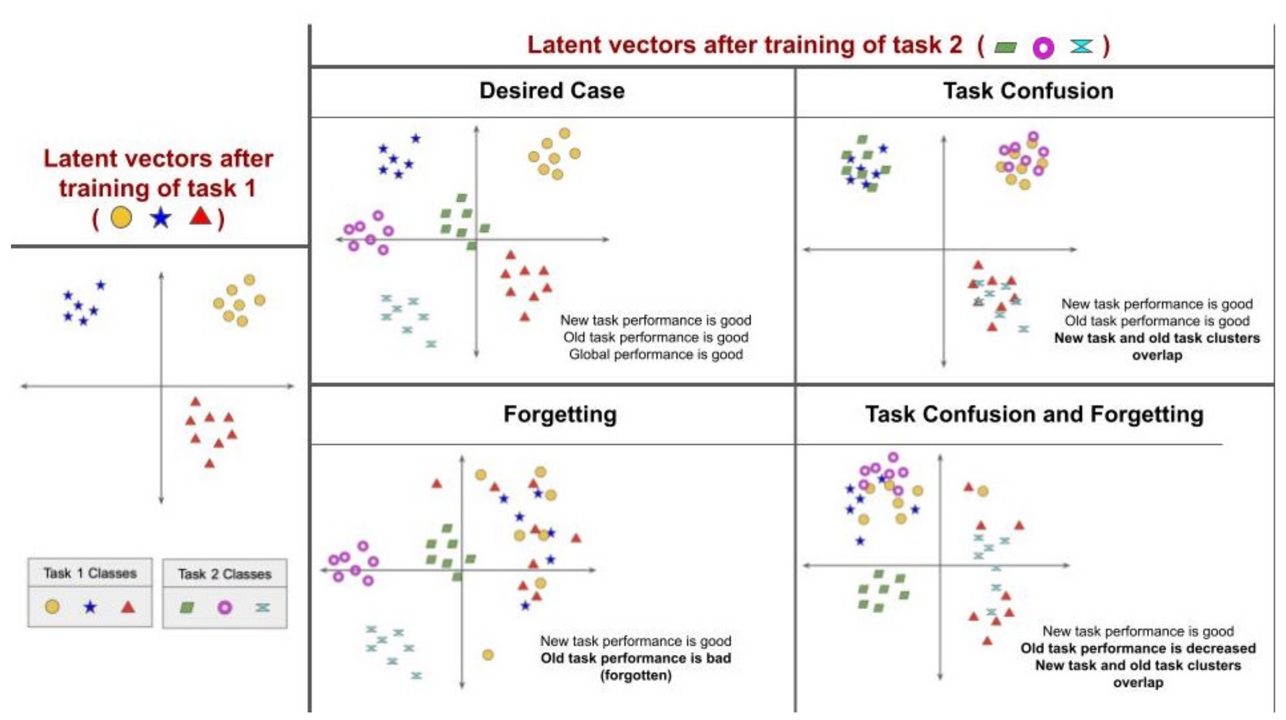
Linear Accuracy = TP * WP

Challenges

Continual Self-Supervised Learning suffers from two main challenges:

Catastrophic Forgetting: represents the significant loss of performance on previous tasks upon learning new ones.

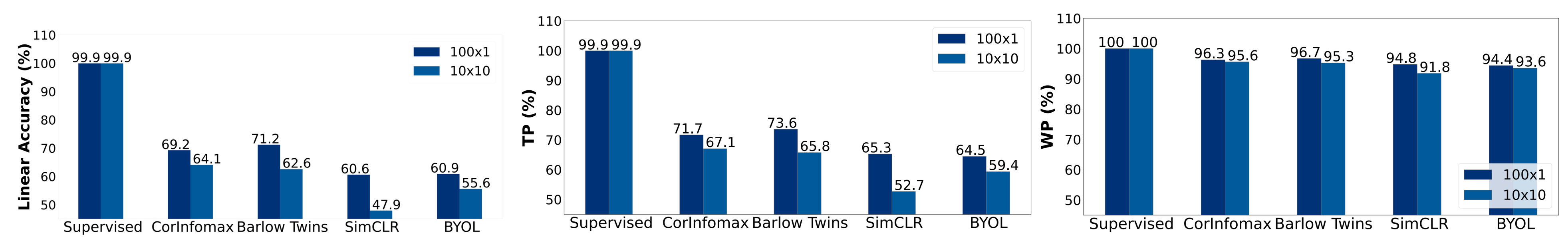
Task Confusion: represents the model failure to establish distinctive decision boundaries between different classes belonging to different tasks.



Motivation

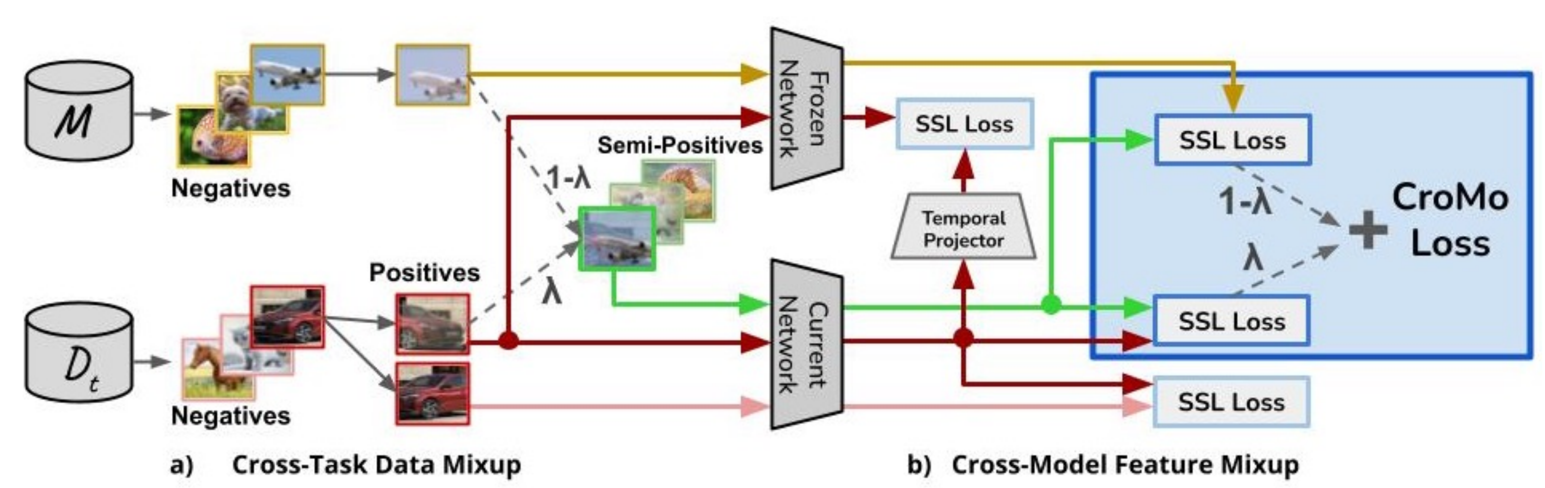
Hypothesis:

"The task confusion problem in Contrastive self-supervised learning (SSL) methods arises primarily from the inability to train the model with different classes belonging to different tasks concurrently."



⚠ Lesser Diversity of negative Samples $\xrightarrow{\text{Contrastive SSL}}$ Task Confusion

Proposed Method



a. Cross-Task Data Mixup:

$$x_{mix_{ij}} = \lambda x_{t,i} + (1-\lambda) x_{\mathcal{M},j}$$

b. Cross-Model Feature Mixup:

$$\mathcal{L}_{CroMo}(z_{mix_{ij}}, z_{t,i}, \bar{z}_{\mathcal{M},j}) = \lambda \cdot \mathcal{L}_{SSL}(z_{mix_{ij}}, z_{t,i}) + (1-\lambda) \cdot \mathcal{L}_{SSL}(z_{mix_{ij}}, \bar{z}_{\mathcal{M},j})$$

$$\boxed{\mathcal{L}_{total} = \mathcal{L}_{SSL}(z^1_{t,i}, z^2_{t,i}) + \zeta(\mathcal{L}_{SSL}(\bar{z}^1_{t,i}, h(z^1_{t,i})) + \mathcal{L}_{SSL}(\bar{z}^2_{t,i}, h(z^2_{t,i}))) + \mathcal{L}_{CroMo}(z^1_{mix_{ij}}, z^1_{t,i}, \bar{z}^1_{\mathcal{M},j}) + \mathcal{L}_{CroMo}(z^2_{mix_{ij}}, z^2_{t,i}, \bar{z}^2_{\mathcal{M},j})}$$

Total Loss Objective

Experiment Setup

Experiment Name	Total # of Classes	Classes per Task	Total # of Tasks	Samples saved/Task	Total # of Samples per Task
cifar10-Split2	10	5	2	500	25,000
cifar100-Split5	100	20	5	500	10,000
cifar100-Split10	100	10	10	100	5,000
tinyImageNet-Split10	200	20	10	100	50,000

Results

$$\text{Linear Accuracy (LA)} = \frac{\text{Total \# of Correct Classification of both Class and Task-ID Prediction Tasks}}{\text{Total \# of Test Samples}}$$

$$\text{Task-ID Prediction (TP)} = \frac{\text{Total \# of Correct Classification of both Class and Task-ID Prediction Tasks}}{\text{Total \# of Test Samples}}$$

$$\text{Within-Task Prediction (WP)} = \frac{\text{Total \# of Correct Classification of both Class and Task-ID Prediction Tasks}}{\text{Total \# of Correct Classification of Task-ID Prediction Task}}$$

		Barlow-Twins			CorInfoMax			SimCLR			BYOL		
	Method	LA(%)	WP(%)	TP(%)	LA(%)	WP(%)	TP(%)	LA(%)	WP(%)	TP(%)	LA(%)	WP(%)	TP(%)
CIFAR10-Split2	Offline	91.65	-	-	92.18	-	-	90.35	-	-	89.60	-	-
	Fine-tune	82.67	90.12	91.73	81.71	91.25	89.55	80.97	90.54	89.43	84.16	94.43	89.12
	ER	85.61	90.36	94.62	87.67	95.81	91.50	81.52	90.95	89.63	86.71	94.97	91.30
	CaSSLe	87.64	91.25	95.87	87.62	96.02	91.17	86.88	95.21	91.25	87.00	95.80	90.81
	CaSSLe+	86.81	90.30	95.91	87.58	95.51	91.70	87.52	95.54	91.61	87.81	96.01	91.49
	CroMo-Mixup*	87.56	92.11	94.69	86.00	93.54	91.96	84.18	92.70	90.81	89.27	96.28	92.72
	CroMo-Mixup	88.22	91.78	95.75	88.51	95.97	92.23	88.49	95.93	92.24	88.88	96.36	92.19
		(+0.6)		**(-0.1)**	**(+0.8)**		**(+0.7)**	**(+1.0)**	–	**(+0.6)**	**(+1.5)**	–	**(+0.7)**
CIFAR100-Split5	Offline	70.03	-	-	70.76	-	-	65.11	-	-	66.73	-	-
	Fine-tune	54.40	85.19	63.86	56.68	86.32	65.66	42.65	78.17	54.56	55.19	86.23	64.00
	ER	57.23	88.26	65.57	59.94	88.62	67.64	47.77	81.87	58.35	56.05	85.60	65.48
	CaSSLe	60.64	87.29	66.35	60.82	88.85	68.45	57.54	87.89	65.47	56.86	86.05	66.08
	CaSSLe+	61.25	88.50	69.03	60.26	88.74	67.92	59.48	88.26	67.39	57.35	87.28	65.71
	CroMo-Mixup*	63.94	91.40	69.88	62.32	88.46	70.39	59.15	87.66	67.48	59.60	88.11	67.64
	CroMo-Mixup	65.48	90.72	72.11	65.06	90.62	71.78	62.72	89.50	70.08	60.60	88.64	68.37
		(+4.2)		**(+3.1)**	**(+4.2)**		**(+3.3)**	**(+3.2)**		**(+2.7)**	**(+3.3)**	-	**(+2.7)**
CIFAR100-Split10	Offline	70.03	-	-	70.76	-	-	65.11	-	-	66.73		
	Fine-tune	51.12	92.01	55.56	50.66	91.58	55.32	39.02	86.61	45.04	49.63	92.23	53.81
	ER	52.81	92.59	56.98	56.99	93.98	60.64	44.83	89.98	49.88	52.32	92.70	56.40
	CaSSLe	56.59	93.93	60.25	56.35	93.95	59.98	53.60	93.39	57.38	52.77	93.22	56.61
	CaSSLe+	56.64	93.49	60.68	57.00	93.95	60.67	55.02	93.43	58.89	53.39	92.67	57.61
	CroMo-Mixup*	60.01	94.50	63.41	60.30	94.34	63.74	55.21	92.84	59.84	56.59	93.52	60.51
	CroMo-Mixup	62.48	95.10	65.70	61.66	94.91	64.97	58.84	94.66	62.18	56.97	93.35	61.03
		(+5.8)		**(+5.0)**	**(+4.7)**		**(+4.3)**	**(+3.8)**		**(+3.3)**	**(+3.6)**	-	**(+3.4)**
tinyImageNet-Split10	Offline	55.60	-	-	55.20	-	-	49.74	-	-	47.58	-	-
	Fine-tune	39.90	77.00	51.82	41.14	78.12	52.66	36.72	75.09	48.90	37.15	76.10	48.82
	ER	40.14	77.07	52.08	41.44	78.22	52.98	37.96	70.08	50.56	37.78	76.17	49.60
	CaSSLe	43.40	79.08	54.88	41.66	77.69	53.62	40.66	77.80	52.26	38.18	77.07	49.54
	CaSSLe+	42.64	78.90	54.04	43.86	79.46	55.20	41.74	78.58	53.12	40.24	78.72	51.12
	CroMo-Mixup*	45.70	80.54	56.74	46.74	81.32	57.48	41.02	78.43	52.30	43.19	79.45	54.36
	CroMo-Mixup	47.32	81.78	57.86	48.22	81.90	58.88	45.82	80.36	57.02	45.44	80.68	56.32
		(+3.7)		**(+3.8)**	**(+4.4)**		**(+3.6)**	**(+4.1)**		**(+3.9)**	**(+5.2)**		**(+4.2)**

References

[1]. Fini, Enrico, et al. "Self-supervised models are continual learners." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.