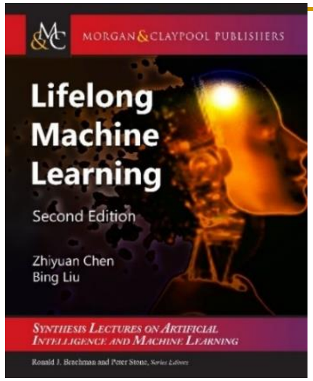
This talk was given at (1) NeurIPS AFM-2024 Workshop and (2) AIGC-2024 conference

Achieving Upper Bound Accuracy in Continual Learning



Bing Liu
Department of Computer Science
University of Illinois Chicago

Outline

- Continual learning and its key challenges
- Theory about class incremental learning (CIL)
- CIL using in-context learning
- Achieving CIL’s upper bound accuracy
- Summary

Lifelong or continual learning (CL)

(Thrun 1996, Silver et al 2013; Ruvolo and Eaton, 2013; Chen and Liu, 2018)

- Learn a sequence of tasks, T_1, T_2, …, T_N, … incrementally. Each task *t* has a training dataset $\mathcal{D}_t = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ in a neural network.
  - In supervised learning, a task is a set of classes to be learned.
  - **Incremental**: In learning a new task, we don't see the data of previous tasks
- **General challenges:**
  1. **Catastrophic forgetting (CF)**: Learning of the new task T_{N+1} should not result in accuracy degradation for any of the previous *N* tasks.
  2. **Knowledge transfer (KT)**: leveraging the knowledge learned from the previous tasks to learn the new task T_{N+1} better.

Two popular CL settings: TIL

- Task incremental learning (TIL): train a “separate” model for each task and task-id is provided during testing
  - **Example**: Task 1: learn to recognize different breeds of dogs. Task 2: learn to recognize different animals. Task 3: learn to recognize different types of fish.
  - Testing needs task information (e.g., task id).

Task incremental learning (TIL). TIL learns a sequence of tasks, $1, 2, ..., T$. Each task k has a training dataset $\mathcal{D}_k = \{((x_k^i, k), y_k^i)_{i=1}^{n_k}\}$, where n_k is the number of data samples in task $k \in \mathbf{T} = \{1, 2, ..., T\}$, and $x_k^i \in \mathbf{X}$ is an input sample and $y_k^i \in \mathbf{Y}_k \subset \mathbf{Y}$ is its class label. The goal of TIL is to construct a predictor $\boxed{f : \mathbf{X} \times \mathbf{T} \rightarrow \mathbf{Y}}$ to identify the class label $y \in \mathbf{Y}_k$ for (x, k) (the given test instance x from task k).

TIL has reached its upper bound accuracy

- The upper bound of TIL is multitask learning.
- Several methods can achieve forgetting free.
  - E.g., HAT (Serra et al. 2018) and SupSup (Worthsman et al., 2020)
    - **Parameter isolation**: finding a subnetwork for each task.
- In terms of knowledge transfer, it is reaching the upper bound (Ke et al, 2021; Ke et al, 2023).

- Serra, Suris, Miron, and Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. ICML-2018.
- Wortsman, Ramanujan, Liu, Kembhavi, Rastegari, Yosinski, and Farhadi. 2020. Supermasks in superposition. NeuriPS-2020.
- Ke, Liu, Xiong, Celikyilmaz, Li. Sub-network Discovery and Soft-masking for Continual Learning of Mixed Tasks. findings, EMNLP-2023), December 6 –10, 2023, Singapore.
- Ke, Liu, Ma, Hu Xu, Shu. Achieving Forgetting Prevention and Knowledge Transfer in Continual Learning. NeurIPS-2021.

Two popular CL settings: CIL

- Class incremental learning (CIL): produce a single model from all tasks and classify all classes during testing
  - **Example**: Task 1: learn to recognize pig and cat. Task 2: sheep. Task 3: chicken and dog. Task 4: horse and cow
  - Testing:



Class incremental learning (CIL). CIL learns a sequence of tasks, $1, 2, ..., T$. Each task k has a training dataset $\mathcal{D}_k = \{(x_k^i, y_k^i)_{i=1}^{n_k}\}$, where n_k is the number of data samples in task k, and $x_k^i \in \mathbf{X}$ is an input sample and $y_k^i \in \mathbf{Y}_k$ (the set of all classes of task k) is its class label. All $\mathbf{Y}_k$'s are disjoint ($\mathbf{Y}_k \cap \mathbf{Y}_{k'} = \emptyset, \forall k \neq k'$) and $\bigcup_{k=1}^{T} \mathbf{Y}_k = \mathbf{Y}$. The goal of CIL is to construct a single predictive function or classifier $\boxed{f : \mathbf{X} \rightarrow \mathbf{Y}}$ that can identify the class label y of each given test instance x.

Kim, Xiao, Konishi, Ke and Liu. A Theoretical Study on Solving Continual Learning. NeurIPS-2022, Nov. 28 - Dec. 9, 2022
van de Ven and Tolias Three scenarios for continual learning, *arXiv preprint arXiv*:1904.07734, 2019.

Additional challenge of CIL (Kim et al, 2022)

- CIL has another challenge of **inter-task class separation (ICS)**.
  - Since after learning each task, its data is no longer accessible, then how to establish the decision boundaries between the classes of the new task and those of old tasks?

- Question: What is the right way to solve CIL regardless what classification algorithm is used?

Kim, Xiao, Konishi, Ke and Liu. A Theoretical Study on Solving Continual Learning. NeurIPS-2022, Nov. 28 - Dec. 9, 2022..

Outline

- Continual learning and its key challenges
- Theory about class incremental learning (CIL)
- CIL using in-context learning
- Achieving CIL’s upper bound accuracy
- Summary

CIL decomposition and theoretical result (Kim et al, 2022)

- CIL problem can be decomposed into two subproblems: within-task prediction (WP) and task-id prediction (TP)

$$\mathbf{P}(x \in \mathbf{X}_{k_0,j_0}|D) = \sum_{k=1,\ldots,n} \mathbf{P}(x \in \mathbf{X}_{k,j_0}|x \in \mathbf{X}_k, D)\mathbf{P}(x \in \mathbf{X}_k|D)$$

$$= \mathbf{P}(\underbrace{x \in \mathbf{X}_{k_0,j_0}}_{\text{WP (i.e., TIL)}}|x \in \mathbf{X}_{k_0}, D)\mathbf{P}(\underbrace{x \in \mathbf{X}_{k_0}}_{\text{TP}}|D)$$

- **Theoretical results:** Good WP and TP (or OOD) are ***necessary*** and ***sufficient*** for good CIL.
  - TP and OOD bound each other.

Kim, Xiao, Konishi, Ke and Liu. A Theoretical Study on Solving Continual Learning. NeurIPS-2022, Nov. 28 - Dec. 9, 2022..

Intuition of the theory

- In learning a new class or task,
  - the system does not see the data of previous tasks, and
  - yet it needs to learn decision boundaries separating the classes of the current task and those of previous tasks,
- The only possible solution is
  - Each task is good at OOD detection.

- Based on this, we also proved that CIL is learnable (Kim et al, 2023)

Kim, Xiao, Konishi, and Liu. Learnability and Algorithm for Continual Learning, ICML-2023..

One proposed method (no pre-trained model)

(Kim et al, 2022)

- Theory-based methods outperform baselines by a large margin
  - (Kim et al 2022)
- No replay or pre-training
- Combination of
  - a TIL method to tackle CF
    - E.g., HAT and SupSup
  - a strong OOD detection
    - E.g., CSI.
- **HAT+CSI** and **Sup+CSI**

Method	M-5T	C10-5T	C100-10T	C100-20T	T-5T	T-10T
OWM	95.8±0.13	51.8±0.05	28.9±0.60	24.1±0.26	10.0±0.55	8.6±0.42
MUC	74.9±0.46	52.9±1.03	30.4±1.18	14.2±0.30	33.6±0.19	17.4±0.17
PASS†	76.6±1.67	47.3±0.98	33.0±0.58	25.0±0.69	28.4±0.51	19.1±0.46
LwF	85.5±3.11	54.7±1.18	45.3±0.75	44.3±0.46	32.2±0.50	24.3±0.26
iCaRL*	96.0±0.43	63.4±1.11	51.4±0.99	47.8±0.48	37.0±0.41	28.3±0.18
Mnemonics†*	96.3±0.36	64.1±1.47	51.0±0.34	47.6±0.74	37.1±0.46	28.5±0.72
BiC	94.1±0.65	61.4±1.74	52.9±0.64	48.9±0.54	41.7±0.74	33.8±0.40
DER++	95.3±0.69	66.0±1.20	53.7±1.30	46.6±1.44	35.8±0.77	30.5±0.47
Co^2L		65.6				
CCG	97.3	70.1				
HAT	81.9±3.74	62.7±1.45	41.1±0.93	25.6±0.51	38.5±1.85	29.8±0.65
HyperNet	56.6±4.85	53.4±2.19	30.2±1.54	18.7±1.10	7.9±0.69	5.3±0.50
Sup	70.1±1.51	62.4±1.45	44.6±0.44	34.7±0.30	41.8±1.50	36.5±0.36
PR-Ent	74.1	61.9	45.2			
HAT+CSI	94.4±0.26	87.8±0.71	63.3±1.00	54.6±0.92	45.7±0.26	47.1±0.18
Sup+CSI	80.7±2.71	86.0±0.41	65.1±0.39	60.2±0.51	48.9±0.25	45.7±0.76
HAT+CSI+c	96.9±0.30	88.0±0.48	65.2±0.71	58.0±0.45	51.7±0.37	47.6±0.32
Sup+CSI+c	81.0±2.30	87.3±0.37	65.2±0.37	60.5±0.64	49.2±0.28	46.2±0.53

Proposed method 2 (using a pre-trained model)

(Lin et al. 2024)

	C10-5T		C100-10T		C100-20T		T-5T		T-10T		Average	
	Last	AIA	Last	AIA	Last	AIA	Last	AIA	Last	AIA	Last	AIA
Non-CL	95.79$^{\pm0.15}$	97.01$^{\pm0.14}$	82.76$^{\pm0.22}$	87.20$^{\pm0.29}$	82.76$^{\pm0.22}$	87.53$^{\pm0.31}$	72.52$^{\pm0.41}$	77.03$^{\pm0.47}$	72.52$^{\pm0.41}$	77.03$^{\pm0.41}$	81.27	85.16
OWM	41.69$^{\pm6.34}$	56.00$^{\pm3.46}$	21.39$^{\pm3.18}$	40.10$^{\pm1.86}$	16.98$^{\pm4.44}$	32.58$^{\pm1.58}$	24.55$^{\pm2.48}$	45.18$^{\pm0.33}$	17.52$^{\pm3.45}$	35.75$^{\pm2.21}$	24.43	41.92
ADAM	83.92$^{\pm0.51}$	90.33$^{\pm0.42}$	61.21$^{\pm0.36}$	72.55$^{\pm0.41}$	58.99$^{\pm0.61}$	70.89$^{\pm0.51}$	50.11$^{\pm0.46}$	61.85$^{\pm0.51}$	49.68$^{\pm0.40}$	61.44$^{\pm0.44}$	60.78	71.41
PASS	86.21$^{\pm1.10}$	89.03$^{\pm7.13}$	68.90$^{\pm0.94}$	77.01$^{\pm2.44}$	66.77$^{\pm1.18}$	76.42$^{\pm1.23}$	61.03$^{\pm0.38}$	67.12$^{\pm6.26}$	58.34$^{\pm0.42}$	67.33$^{\pm3.63}$	68.25	75.38
HAT$_{CIL}$	82.40$^{\pm0.12}$	91.06$^{\pm0.36}$	62.91$^{\pm0.24}$	73.99$^{\pm0.86}$	59.54$^{\pm0.41}$	69.12$^{\pm1.06}$	59.22$^{\pm0.10}$	69.38$^{\pm1.14}$	54.03$^{\pm0.21}$	65.63$^{\pm1.64}$	63.62	73.84
SLDA	88.64$^{\pm0.05}$	93.54$^{\pm0.66}$	67.82$^{\pm0.05}$	77.72$^{\pm0.58}$	67.80$^{\pm0.05}$	78.51$^{\pm0.58}$	57.93$^{\pm0.05}$	66.03$^{\pm1.35}$	57.93$^{\pm0.06}$	67.39$^{\pm1.81}$	68.02	76.64
L2P	73.59$^{\pm4.15}$	84.60$^{\pm2.28}$	61.72$^{\pm0.81}$	72.88$^{\pm1.18}$	53.84$^{\pm1.59}$	66.52$^{\pm1.61}$	59.12$^{\pm0.96}$	67.81$^{\pm1.25}$	54.09$^{\pm1.14}$	64.59$^{\pm1.59}$	60.47	71.28
iCaRL	87.55$^{\pm0.99}$	89.74$^{\pm6.63}$	68.90$^{\pm0.47}$	76.50$^{\pm3.56}$	69.15$^{\pm0.99}$	77.06$^{\pm2.36}$	53.13$^{\pm1.04}$	61.36$^{\pm6.21}$	51.88$^{\pm2.36}$	63.56$^{\pm3.08}$	66.12	73.64
A-GEM	56.33$^{\pm7.77}$	68.19$^{\pm3.24}$	25.21$^{\pm4.00}$	43.83$^{\pm0.69}$	21.99$^{\pm4.01}$	35.97$^{\pm1.15}$	30.53$^{\pm3.99}$	49.26$^{\pm0.64}$	21.90$^{\pm5.52}$	39.58$^{\pm3.32}$	31.19	47.37
EEIL	82.34$^{\pm3.13}$	90.50$^{\pm0.72}$	68.08$^{\pm0.51}$	81.10$^{\pm0.37}$	63.79$^{\pm0.66}$	79.54$^{\pm0.69}$	53.34$^{\pm0.54}$	66.63$^{\pm0.40}$	50.38$^{\pm0.97}$	66.54$^{\pm0.61}$	63.59	76.86
GD	89.16$^{\pm0.53}$	94.22$^{\pm0.75}$	64.36$^{\pm0.57}$	80.51$^{\pm0.57}$	60.10$^{\pm0.74}$	78.43$^{\pm0.76}$	53.01$^{\pm0.97}$	67.51$^{\pm0.38}$	42.48$^{\pm2.53}$	63.91$^{\pm0.40}$	61.82	76.92
DER++	84.63$^{\pm2.91}$	89.01$^{\pm6.29}$	69.73$^{\pm0.99}$	80.64$^{\pm2.74}$	70.03$^{\pm1.46}$	81.72$^{\pm1.76}$	55.84$^{\pm2.21}$	66.55$^{\pm3.73}$	54.20$^{\pm3.28}$	67.14$^{\pm1.40}$	66.89	77.01
HAL	84.38$^{\pm2.70}$	87.00$^{\pm7.27}$	67.17$^{\pm1.50}$	77.42$^{\pm2.73}$	67.37$^{\pm1.45}$	77.85$^{\pm1.71}$	52.80$^{\pm2.37}$	65.31$^{\pm3.68}$	55.25$^{\pm3.60}$	64.48$^{\pm1.45}$	65.39	74.41
DER	86.79$^{\pm1.20}$	92.83$^{\pm1.10}$	73.30$^{\pm0.58}$	82.89$^{\pm0.45}$	72.00$^{\pm0.57}$	82.79$^{\pm0.76}$	59.57$^{\pm0.89}$	70.32$^{\pm0.57}$	57.18$^{\pm1.40}$	70.21$^{\pm0.86}$	69.77	79.81
FOSTER	86.09$^{\pm0.38}$	91.54$^{\pm0.65}$	71.69$^{\pm0.24}$	81.16$^{\pm0.39}$	72.91$^{\pm0.45}$	83.02$^{\pm0.86}$	54.44$^{\pm0.28}$	69.95$^{\pm0.28}$	55.70$^{\pm0.40}$	70.00$^{\pm0.26}$	68.17	79.13
BEEF	87.10$^{\pm1.38}$	93.10$^{\pm1.21}$	72.09$^{\pm0.33}$	81.91$^{\pm0.58}$	71.88$^{\pm0.54}$	81.45$^{\pm0.74}$	61.41$^{\pm0.38}$	71.21$^{\pm0.57}$	58.16$^{\pm0.60}$	71.16$^{\pm0.82}$	70.13	79.77
MORE	89.16$^{\pm0.96}$	94.23$^{\pm0.82}$	70.23$^{\pm2.27}$	81.24$^{\pm1.24}$	70.53$^{\pm1.09}$	81.59$^{\pm0.98}$	64.97$^{\pm1.28}$	74.03$^{\pm1.61}$	63.06$^{\pm1.26}$	72.74$^{\pm1.04}$	71.59	80.77
ROW	90.97$^{\pm0.19}$	94.45$^{\pm0.21}$	74.72$^{\pm0.48}$	82.87$^{\pm0.41}$	74.60$^{\pm0.12}$	83.12$^{\pm0.23}$	65.11$^{\pm1.97}$	74.16$^{\pm1.34}$	63.21$^{\pm2.53}$	72.91$^{\pm2.12}$	73.72	81.50
TPL (ours)	**92.33**$^{\pm0.32}$	**95.11**$^{\pm0.44}$	**76.53**$^{\pm0.27}$	**84.10**$^{\pm0.34}$	**76.34**$^{\pm0.38}$	**84.46**$^{\pm0.28}$	**68.64**$^{\pm0.44}$	**76.77**$^{\pm0.23}$	**67.20**$^{\pm0.51}$	**75.72**$^{\pm0.37}$	**76.21**	**83.23**
Non-CL$_{PFI}$	96.90$^{\pm0.07}$	97.96$^{\pm0.05}$	83.61$^{\pm0.33}$	89.72$^{\pm0.10}$	83.61$^{\pm0.33}$	88.89$^{\pm0.06}$	85.55$^{\pm0.07}$	88.26$^{\pm0.08}$	85.71$^{\pm0.14}$	88.66$^{\pm0.01}$	87.08	90.70
TPL$_{PFI}$	94.86$^{\pm0.02}$	96.89$^{\pm0.02}$	82.43$^{\pm0.12}$	88.28$^{\pm0.17}$	80.86$^{\pm0.07}$	87.32$^{\pm0.07}$	84.06$^{\pm0.11}$	87.19$^{\pm0.11}$	83.87$^{\pm0.07}$	87.40$^{\pm0.16}$	85.22	89.42

Lin, Shao, Qian, Pan, Guo, and Liu. Class Incremental Learning Via Likelihood Ratio Based Task Prediction. ICLR-2024

Graph Class Incremental Learning (Niu et al, NeurIPS-2024)

Methods	Data Replay	CoraFull AA/%↑	CoraFull AF/%↑	Arixv AA/%↑	Arixv AF/%↑	Reddit AA/%↑	Reddit AF/%↑	Products AA/%↑	Products AF/%↑
Fine-tune	×	3.5±0.5	-95.2±0.5	4.9±0.0	-89.7±0.4	5.9±1.2	-97.9±3.3	7.6±0.7	-88.7±0.8
Joint	×	81.2±0.4	-	51.3±0.5	-	97.1±0.1	-	71.5±0.1	-
EWC	×	52.6±8.2	-38.5±12.1	8.5±1.0	-69.5±8.0	10.3±11.6	-33.2±26.1	23.8±3.8	-21.7±7.5
MAS	×	6.5±1.5	-92.3±1.5	4.8±0.4	-72.2±4.1	9.2±14.5	-23.1±28.2	16.7±4.8	-57.0±31.9
GEM	×	8.4±1.1	-88.4±1.4	4.9±0.0	-89.8±0.3	11.5±5.5	-92.4±5.9	4.5±1.3	-94.7±0.4
LwF	×	33.4±1.6	-59.6±2.2	9.9±12.1	-43.6±11.9	86.6±1.1	-9.2±1.1	48.2±1.6	-18.6±1.6
TWP	×	62.6±2.2	-30.6±4.3	6.7±1.5	-50.6±13.2	8.0±5.2	-18.8±9.0	14.1±4.0	-11.4±2.0
ERGNN	✓	34.5±4.4	-61.6±4.3	21.5±5.4	-70.0±5.5	82.7±0.4	-17.3±0.4	48.3±1.2	-45.7±1.3
SSM-uniform	✓	73.0±0.3	-14.8±0.5	47.1±0.5	-11.7±1.5	94.3±0.1	-1.4±0.1	62.0±1.6	-9.9±1.3
SSM-degree	✓	75.4±0.1	-9.7±0.0	48.3±0.5	-10.7±0.3	94.4±0.0	-1.3±0.0	63.3±0.1	-9.6±0.3
SEM-curvature	✓	77.7±0.8	-10.0±1.2	49.9±0.6	-8.4±1.3	96.3±0.1	-0.6±0.1	65.1±1.0	-9.5±0.8
CaT	✓	80.4±0.5	-5.3±0.4	48.2±0.4	-12.6±0.7	97.3±0.1	-0.4±0.0	70.3±0.9	-4.5±0.8
DeLoMe	✓	81.0±0.2	-3.3±0.3	50.6±0.3	5.1±0.4	97.4±0.1	-0.1±0.1	67.5±0.7	-17.3±0.3
OODCIL	✓	71.3±0.5	-1.1±0.1	19.3±1.4	-1.0±0.4	79.3±0.8	-0.1±0.0	41.6±0.9	-1.6±0.4
TPP (Ours)	×	**93.4±0.4**	**0.0±0.0**	**85.4±0.1**	**0.0±0.0**	**99.5±0.0**	**0.0±0.0**	**94.0±0.5**	**0.0±0.0**
Oracle Model	×	95.5±0.2	-	90.3±0.4	-	99.5±0.0	-	95.3±0.8	-

Niu, Pang, Chen, Liu. Replay-and-Forget-Free Graph Class-Incremental Learning: A Task Profiling and Prompting Approach. NeurIPS-2024

Outline

- Continual learning and its key challenges
- Theory about class incremental learning (CIL)
- CIL using in-context learning
- Achieving CIL’s upper bound accuracy
- Summary

CIL using in-context learning: Naïve approach

- When a task arrives, we can simply add new classes and their training samples to the prompt.
- This approach does not work due to the LLM token limit
  - Long context LLMs don’t work well for CL.

(1). Incremental summarization (Qiu et al, Coling-2025)

- Online or stream continual learning
- Training: Each class is represented by a summary that is incrementally updated as new samples arrive
- Testing: for each test instance x, we
  - Divide classes learned so far into chunks such that each chunk is within the token limit of the LLM
  - Prompt LLM to generate confidence that *x* belongs to each class in a chunk,
  - Get the top *k* classes with the highest confidences from all chunks
  - Prompt the LLM again with only the resulting *k* classes,
    - Select the class with the highest confidence for x.

Results

- (Qiu et al, Coling-2025)

	CIS (Llama)		Joint (Llama)				CL Baselines					
	3/4-Blurry	4/3-Blurry	Zero-shot	Prompting	Fine-tuning		EWC		LAMOL		VAG	
	7 samples	7 samples	no sample	7 samples	7 samples	full data	7 samples	full data	7 samples	full data	7 samples	full data
Banking-77	78.78 ±1.68	79.23 ±2.50	50.22 ±0.00	87.92 ±0.60	69.39 ±0.17	91.19 ±0.08	2.14 ±0.35	9.09 ±0.84	3.50 ±0.04	33.43 ±0.18	36.25 ±3.80	55.19 ±1.54
CLINC-80	91.51 ±4.35	90.40 ±5.46	80.67 ±0.00	95.10 ±2.51	91.18 ±0.46	97.92 ±0.06	1.14 ±0.33	8.26 ±0.76	17.60 ±0.19	52.20 ±0.09	64.75 ±0.69	80.68 ±0.72
DBpedia-14	92.07 ±1.07	92.26 ±0.76	93.36 ±0.00	90.50 ±0.40	93.74 ±0.11	99.00 ±0.00	6.55 ±0.73	23.14 ±1.55	0.70 ±0.14	28.61 ±0.02	55.36 ±3.30	56.58 ±1.22
Reuters-14	83.97 ±1.08	84.61 ±1.24	92.55 ±0.48	77.82 ±2.99	82.64 ±0.33	92.55 ±0.48	7.70 ±0.70	12.79 ±0.14	0.95 ±0.07	29.93 ±0.17	44.08 ±0.27	58.71 ±1.92

(2). ICL with the help of an external learner

- Employ an **external continual learner** (ECL) that has no forgetting, but inaccurate (Momeni et al, 2025a)
  - Training: ECL uses only the features from the LLM, no parameter updating
    - Generating tags for training examples using ICL
      - Compute a mean of tag embeddings for each class and a shared covariance matrix of the embeddings for all classes
  - Testing: apply ***linear discriminant analysis*** (LDA),
    - (1) Given a test sample, ECL identifies the top-*k* candidate classes
    - (2) Summaries of the top-*k* classes are used by ICL for final classification.

Results (Momeni et al, 2025a)

Dataset	#Tasks	Fine-tuning based Methods: Vanilla	EWC	L2P	LAMOL	VAG	INCA	JOINT
CLINC	10	51.27 ±1.26	54.22 ±1.14	52.53 ±1.72	58.42 ±0.84	76.42 ±0.90	**94.40**	97.60
Banking	7	27.77 ±2.46	29.10 ±1.78	25.78 ±1.21	42.60 ±1.36	59.34 ±1.28	**84.90**	92.50
DBpedia	7	39.02 ±2.68	40.30 ±2.89	42.84 ±5.47	48.61 ±1.82	65.40 ±1.52	**84.20**	95.70
HWU	8	38.38 ±4.01	42.72 ±2.62	28.77 ±3.18	44.85 ±1.57	56.88 ±1.22	**86.61**	90.43

- Table 1: LLM is Mistral 7B
- Table 2: with or without ECL

Model	CLINC	Banking	DBpedia	HWU
Mistral-7B	94.40%	84.90%	84.20%	86.61%
Llama3-8B	95.73%	84.30%	87.60%	87.45%
Gemini 1.5 flash	95.32%	86.15%	91.63%	89.22%
	Without ECL			
Mistral-7B	86.93%	65.90%	65.30%	81.04%
Llama3-8B	83.73%	77.80%	72.70%	83.27%
Gemini 1.5 flash	93.86%	83.52%	79.64%	87.27%
LongAlpaca-7B	45.87%	33.20%	24.90%	35.97%
LongAlpaca-13B	51.20%	63.60%	59.10%	62.83%
LongLlama-3B	62.00%	52.80%	38.90%	58.46%
LongLlama-7B	84.67%	73.10%	61.00%	77.88%

Momeni, Mazumder, Ke and Liu. In-context Continual Learning Assisted by an External Continual Learner. COLING-2025

Not a good idea

- Weird idea: Generating tags from training samples and then getting their embeddings to compute mean and covariance.
  - We did this because we originally want to do **retrieval-augmented CL**
  - **Retrieval** uses TF-IDF to obtain the top k classes. Each class is represented by a set of tags generated from its training documents.
    - Sadly, nobody liked the idea. The paper got rejected multiple times.
- Why not extracting features of training samples directly from an LLM to compute the mean and covariance?
  - This did wonders!

Outline

- Continual learning and its key challenges
- Theory about class incremental learning (CIL)
- CIL using in-context learning
- Achieving CIL’s upper bound accuracy
- Summary

KLDA: kernel linear discriminant analysis

(Momeni et al, 2025b)

- Using only ***large foundation models*** as feature extractors, no training.
  - The extracted features are kernelled using the RBF kernel and *random Fourier features*
  - Training / Learning
    - Compute **a feature mean for each class** and a **shared covariance matrix**
  - Testing
    - Using LDA

KLDA for CIL using text datasets

(Momeni et al, 2025b)

- LM = BART-base as VAG (Shao et al, ACL-2023) uses BART-base

	Method	**CLINC** (10-T)	**Banking** (7-T)	**DBpedia** (7-T)	**HWU** (8-T)
(upper bound)	Joint Fine-tuning	95.33 ±0.04	91.36 ±0.32	94.83 ±0.16	88.60 ±0.29
	Vanilla	42.06 ±1.53	31.80 ±1.20	43.45 ±2.54	30.95 ±3.37
	EWC	45.73 ±0.46	38.40 ±2.70	44.99 ±2.90	34.01 ±3.46
	KD	36.33 ±0.86	27.40 ±1.59	42.10 ±2.40	25.46 ±2.13
	L2P	30.66 ±2.46	31.45 ±0.55	23.52 ±1.54	24.04 ±0.88
	LAMOL	58.42 ±0.84	42.60 ±1.36	48.61 ±1.82	44.85 ±1.57
	VAG	76.42 ±0.90	59.34 ±1.28	65.40 ±1.52	56.88 ±1.22
	NCM	83.60 ±0.00	71.10 ±0.00	75.70 ±0.00	73.30 ±0.00
	LDA	93.71 ±0.00	89.09 ±0.00	93.42 ±0.00	86.41 ±0.00
	KLDA	95.90 ±0.68	92.23 ±0.32	94.13 ±0.32	87.27 ±1.39
	KLDA with Ensemble	**96.62** ±0.08	**93.03** ±0.06	**94.53** ±0.12	**89.78** ±0.09

Momeni, Mazumder, and Liu, Continual Learning Using a Kernel-Based Method over Foundation Models AAAI-2025

More results

(Momeni et al, 2025b)

- Using more LMs

Model	Dataset	KLDA-E	Joint (upper bound)
MiniLM 3 layers 384 dimensions	CLINC	94.53±0.00	93.20±0.16
	Banking	91.73±0.09	90.90±0.08
	DBpedia	86.83±0.17	87.43±0.16
	HWU	87.95±0.23	87.13±0.12
BERT-base 12 layers 768 dimensions	CLINC	94.98±0.31	94.56±0.04
	Banking	91.00±0.24	88.96±0.16
	DBpedia	95.40±0.08	95.03±0.09
	HWU	88.32±0.31	87.26±0.28
RoBERTa-large 24 layers 1024 dimensions	CLINC	96.31±0.06	95.96±0.30
	Banking	92.93±0.05	91.16±0.04
	DBpedia	94.60±0.08	94.99±0.21
	HWU	89.25±0.04	88.40±0.29
T5-3b 24 layers 1024 dimensions	CLINC	96.04±0.17	96.86±0.06
	Banking	93.77±0.05	92.30±0.10
	DBpedia	95.33±0.09	94.60±0.03
	HWU	89.31±0.27	90.30±0.10
Mistral-7b 32 layers 4096 dimensions	CLINC	97.13±0.11	97.60±0.11
	Banking	92.53±0.12	92.50±0.14
	DBpedia	96.00±0.08	95.70±0.07
	HWU	90.02±0.09	90.43±0.11

Momeni, Mazumder, and Liu, Continual Learning Using a Kernel-Based Method over Foundation Models. AAAI-2025

KLDA for CIL using image datasets

(Momeni et al, 2025b)

- **DINOv2:** a pre-trained model trained with self-supervision
  - Using a pre-trained ***foundation model*** trained using **supervised data** is problematic: information leak

Model	Dataset	KLDA	Joint (upper bound)
DINOv2-small 12 layers 384 dimensions	CIFAR10	97.00±0.07	97.02±0.09
	CIFAR100	84.21±0.08	85.52±0.17
	T-ImageNet	78.67±0.08	81.30±0.17
	Cars	81.94±0.11	81.88±0.23
DINOv2-base 12 layers 768 dimensions	CIFAR10	98.45±0.04	98.54±0.06
	CIFAR100	88.81±0.07	90.30±0.09
	T-ImageNet	83.18±0.11	86.43±0.14
	Cars	87.45±0.14	87.47±0.21

Momeni, Mazumder, and Liu, Continual Learning Using a Kernel-Based Method over Foundation Models .AAAI-2025

Note the gap: Last and upper bound in (Lin et al 2024)

Non-CL is **Joint** (upper bound)

	C10-5T		C100-10T		C100-20T		T-5T		T-10T		Average	
	Last	AIA	Last	AIA	Last	AIA	Last	AIA	Last	AIA	Last	AIA
Non-CL	95.79$^{\pm0.15}$	97.01$^{\pm0.14}$	82.76$^{\pm0.22}$	87.20$^{\pm0.29}$	82.76$^{\pm0.22}$	87.53$^{\pm0.31}$	72.52$^{\pm0.41}$	77.03$^{\pm0.47}$	72.52$^{\pm0.41}$	77.03$^{\pm0.41}$	81.27	85.16
OWM	41.69$^{\pm6.34}$	56.00$^{\pm3.46}$	21.39$^{\pm3.18}$	40.10$^{\pm1.86}$	16.98$^{\pm4.44}$	32.58$^{\pm1.58}$	24.55$^{\pm2.48}$	45.18$^{\pm0.33}$	17.52$^{\pm3.45}$	35.75$^{\pm2.21}$	24.43	41.92
ADAM	83.92$^{\pm0.51}$	90.33$^{\pm0.42}$	61.21$^{\pm0.36}$	72.55$^{\pm0.41}$	58.99$^{\pm0.61}$	70.89$^{\pm0.51}$	50.11$^{\pm0.46}$	61.85$^{\pm0.51}$	49.68$^{\pm0.40}$	61.44$^{\pm0.44}$	60.78	71.41
PASS	86.21$^{\pm1.10}$	89.03$^{\pm7.13}$	68.90$^{\pm0.94}$	77.01$^{\pm2.44}$	66.77$^{\pm1.18}$	76.42$^{\pm1.23}$	61.03$^{\pm0.38}$	67.12$^{\pm6.26}$	58.34$^{\pm0.42}$	67.33$^{\pm3.63}$	68.25	75.38
HAT$_{CIL}$	82.40$^{\pm0.12}$	91.06$^{\pm0.36}$	62.91$^{\pm0.24}$	73.99$^{\pm0.86}$	59.54$^{\pm0.41}$	69.12$^{\pm1.06}$	59.22$^{\pm0.10}$	69.38$^{\pm1.14}$	54.03$^{\pm0.21}$	65.63$^{\pm1.64}$	63.62	73.84
SLDA	88.64$^{\pm0.05}$	93.54$^{\pm0.66}$	67.82$^{\pm0.05}$	77.72$^{\pm0.58}$	67.80$^{\pm0.05}$	78.51$^{\pm0.58}$	57.93$^{\pm0.05}$	66.03$^{\pm1.35}$	57.93$^{\pm0.06}$	67.39$^{\pm1.81}$	68.02	76.64
L2P	73.59$^{\pm4.15}$	84.60$^{\pm2.28}$	61.72$^{\pm0.81}$	72.88$^{\pm1.18}$	53.84$^{\pm1.59}$	66.52$^{\pm1.61}$	59.12$^{\pm0.96}$	67.81$^{\pm1.25}$	54.09$^{\pm1.14}$	64.59$^{\pm1.59}$	60.47	71.28
iCaRL	87.55$^{\pm0.99}$	89.74$^{\pm6.63}$	68.90$^{\pm0.47}$	76.50$^{\pm3.56}$	69.15$^{\pm0.99}$	77.06$^{\pm2.36}$	53.13$^{\pm1.04}$	61.36$^{\pm6.21}$	51.88$^{\pm2.36}$	63.56$^{\pm3.08}$	66.12	73.64
A-GEM	56.33$^{\pm7.77}$	68.19$^{\pm3.24}$	25.21$^{\pm4.00}$	43.83$^{\pm0.69}$	21.99$^{\pm4.01}$	35.97$^{\pm1.15}$	30.53$^{\pm3.99}$	49.26$^{\pm0.64}$	21.90$^{\pm5.52}$	39.58$^{\pm3.32}$	31.19	47.37
EEIL	82.34$^{\pm3.13}$	90.50$^{\pm0.72}$	68.08$^{\pm0.51}$	81.10$^{\pm0.37}$	63.79$^{\pm0.66}$	79.54$^{\pm0.69}$	53.34$^{\pm0.54}$	66.63$^{\pm0.40}$	50.38$^{\pm0.97}$	66.54$^{\pm0.61}$	63.59	76.86
GD	89.16$^{\pm0.53}$	94.22$^{\pm0.75}$	64.36$^{\pm0.57}$	80.51$^{\pm0.57}$	60.10$^{\pm0.74}$	78.43$^{\pm0.76}$	53.01$^{\pm0.97}$	67.51$^{\pm0.38}$	42.48$^{\pm2.53}$	63.91$^{\pm0.40}$	61.82	76.92
DER++	84.63$^{\pm2.91}$	89.01$^{\pm6.29}$	69.73$^{\pm0.99}$	80.64$^{\pm2.74}$	70.03$^{\pm1.46}$	81.72$^{\pm1.76}$	55.84$^{\pm2.21}$	66.55$^{\pm3.73}$	54.20$^{\pm3.28}$	67.14$^{\pm1.40}$	66.89	77.01
HAL	84.38$^{\pm2.70}$	87.00$^{\pm7.27}$	67.17$^{\pm1.50}$	77.42$^{\pm2.73}$	67.37$^{\pm1.45}$	77.85$^{\pm1.71}$	52.80$^{\pm2.37}$	65.31$^{\pm3.68}$	55.25$^{\pm3.60}$	64.48$^{\pm1.45}$	65.39	74.41
DER	86.79$^{\pm1.20}$	92.83$^{\pm1.10}$	73.30$^{\pm0.58}$	82.89$^{\pm0.45}$	72.00$^{\pm0.57}$	82.79$^{\pm0.76}$	59.57$^{\pm0.89}$	70.32$^{\pm0.57}$	57.18$^{\pm1.40}$	70.21$^{\pm0.86}$	69.77	79.81
FOSTER	86.09$^{\pm0.38}$	91.54$^{\pm0.65}$	71.69$^{\pm0.24}$	81.16$^{\pm0.39}$	72.91$^{\pm0.45}$	83.02$^{\pm0.86}$	54.44$^{\pm0.28}$	69.95$^{\pm0.28}$	55.70$^{\pm0.40}$	70.00$^{\pm0.26}$	68.17	79.13
BEEF	87.10$^{\pm1.38}$	93.10$^{\pm1.21}$	72.09$^{\pm0.33}$	81.91$^{\pm0.58}$	71.88$^{\pm0.54}$	81.45$^{\pm0.74}$	61.41$^{\pm0.38}$	71.21$^{\pm0.57}$	58.16$^{\pm0.60}$	71.16$^{\pm0.82}$	70.13	79.77
MORE	89.16$^{\pm0.96}$	94.23$^{\pm0.82}$	70.23$^{\pm2.27}$	81.24$^{\pm1.24}$	70.53$^{\pm1.09}$	81.59$^{\pm0.98}$	64.97$^{\pm1.28}$	74.03$^{\pm1.61}$	63.06$^{\pm1.26}$	72.74$^{\pm1.04}$	71.59	80.77
ROW	90.97$^{\pm0.19}$	94.45$^{\pm0.21}$	74.72$^{\pm0.48}$	82.87$^{\pm0.41}$	74.60$^{\pm0.12}$	83.12$^{\pm0.23}$	65.11$^{\pm1.97}$	74.16$^{\pm1.34}$	63.21$^{\pm2.53}$	72.91$^{\pm2.12}$	73.72	81.50
TPL (ours)	**92.33**$^{\pm0.32}$	**95.11**$^{\pm0.44}$	**76.53**$^{\pm0.27}$	**84.10**$^{\pm0.34}$	**76.34**$^{\pm0.38}$	**84.46**$^{\pm0.28}$	**68.64**$^{\pm0.44}$	**76.77**$^{\pm0.23}$	**67.20**$^{\pm0.51}$	**75.72**$^{\pm0.37}$	**76.21**	**83.23**

Lin, Shao, Qian, Pan, Guo, and Liu. Class Incremental Learning Via Likelihood Ratio Based Task Prediction. ICLR-2024

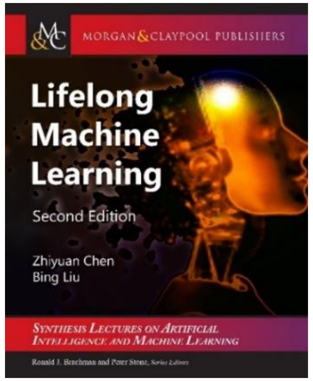
Outline

- Continual learning and its key challenges
- Theory about class incremental learning (CIL)
- CIL using in-context learning
- Enabling CIL to achieve joint training accuracy (upper bound)
- Summary

Summary

- Foundation models are critical for continual learning (CIL)
  - Eliminate *catastrophic forgetting* and *inter-task class separation* challenges
    - Help CIL achieve upper bound accuracy
- The new methods are theoretically justified
  - They are all good at OOD detection for each task/class
    - Summary represents a class only (Qiu et al 2025):
    - Mean and covariance represent the distribution of a class (Momeni et al 2025)
- Controversial questions?
  - **Does continual learning need to learn features?**
  - **Do humans learn features? Are they in-built?**

Thank You



Q&A

Students: Zhiyuan Chen (ex), Sepideh Esmaeilpour (ex), Zixuan Ke (ex), Gyuhak Kim (ex), Nianzu Ma (ex), Sahisnu Mazumder (ex), Saleh Momeni, Jade Qiu, Lei Shu (ex), Hu Xu (ex)

Collaborators: Wenpeng Hu, Scott Grigsby, Yiduo Guo, Tatsuya Konishi, Haowei Lin, Eric Robertson, Yijia Shao, Changnan Xiao.

Funding:


CoLLAs
4th Conference on Lifelong Learning Agents
Abstract: Feb 21
Paper Submission: Feb 26
Aug 11-14
@ UPenn, USA
FMC

CoLLAs

4th Conference on Lifelong Learning Agents

Abstract Deadline: Feb 21
Paper Submission: Feb 26

- Theory for continual/lifelong learning
- Continual learning paradigms (class-incremental, task incremental, domain incremental, curriculum learning, active learning, federated learning, online learning, meta-learning, few-shot learning, and other non-stationary learning paradigms)
- Challenges with non-stationary learning (loss of plasticity, catastrophic forgetting, policy collapse, unlearning, OOD generalization, distribution shift, etc.)
- Continual reinforcement learning (options, skill discovery, hierarchical RL, intrinsically motivated learning, multi-agent RL)
- Continual learning in LLMs (in-context learning, pre-training, model editing, fine-tuning, adaptation)
- Knowledge transfer (transfer learning, multi-task learning, domain adaptation, sim2real, meta-learning)
- Non-stationary Optimization
- Streaming learning, on-device, real-time learning
- Open-world learning, open-ended learning
- Neuroscience-inspired continual/lifelong learning
- Applications (control, robotics, healthcare, etc.)
- Datasets, benchmarks, evaluation, software libraries