ICLR The Thirteenth International Conference on Learning Representations

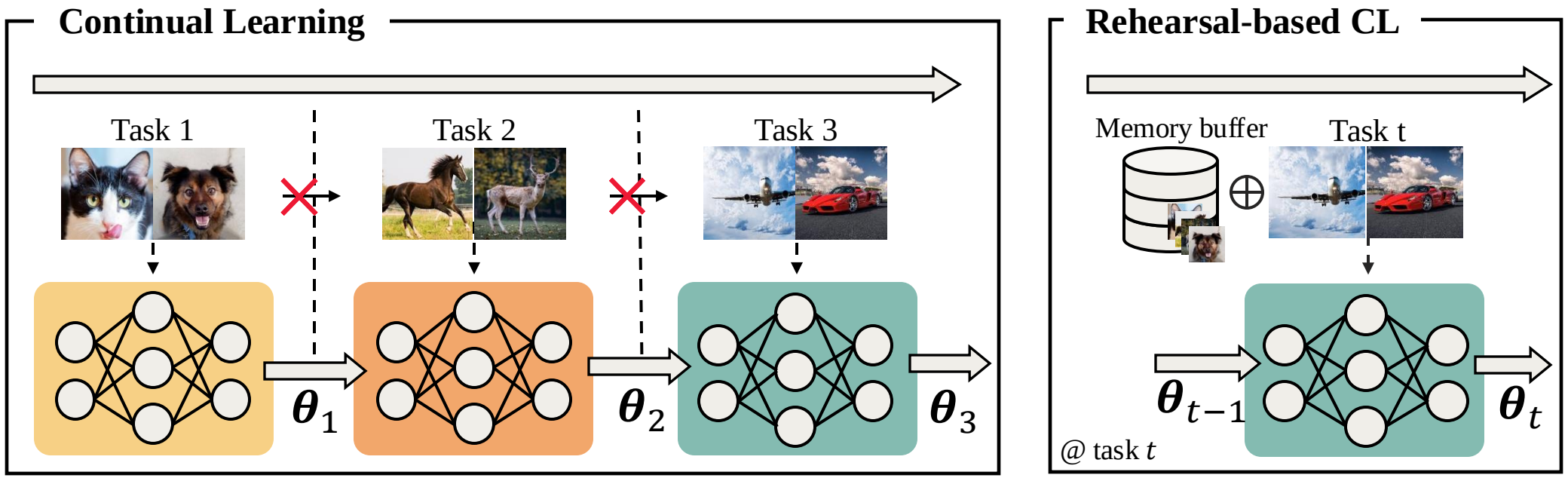
Active Learning for Continual Learning: Keeping the Past Alive in the Present

Jaehyun Park *Dongmin Park* *Jae-Gil Lee*

KAIST KRAFTON

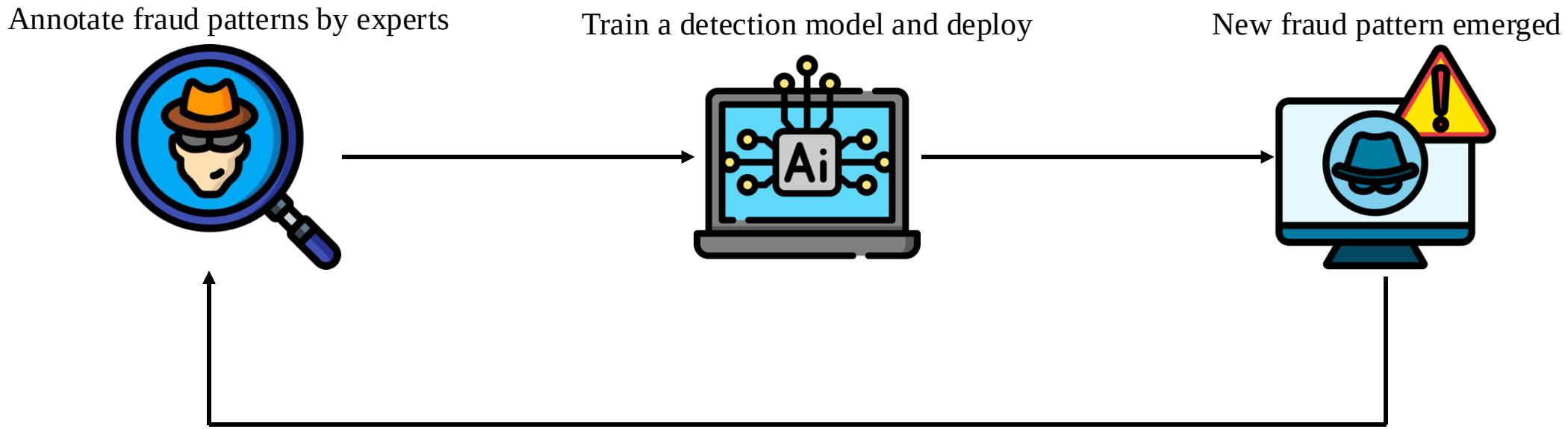
Continual Learning

- Realistic deep learning scenario to adapt models continuously on evolving data distribution to maintain knowledge of the past **without access to previous data**



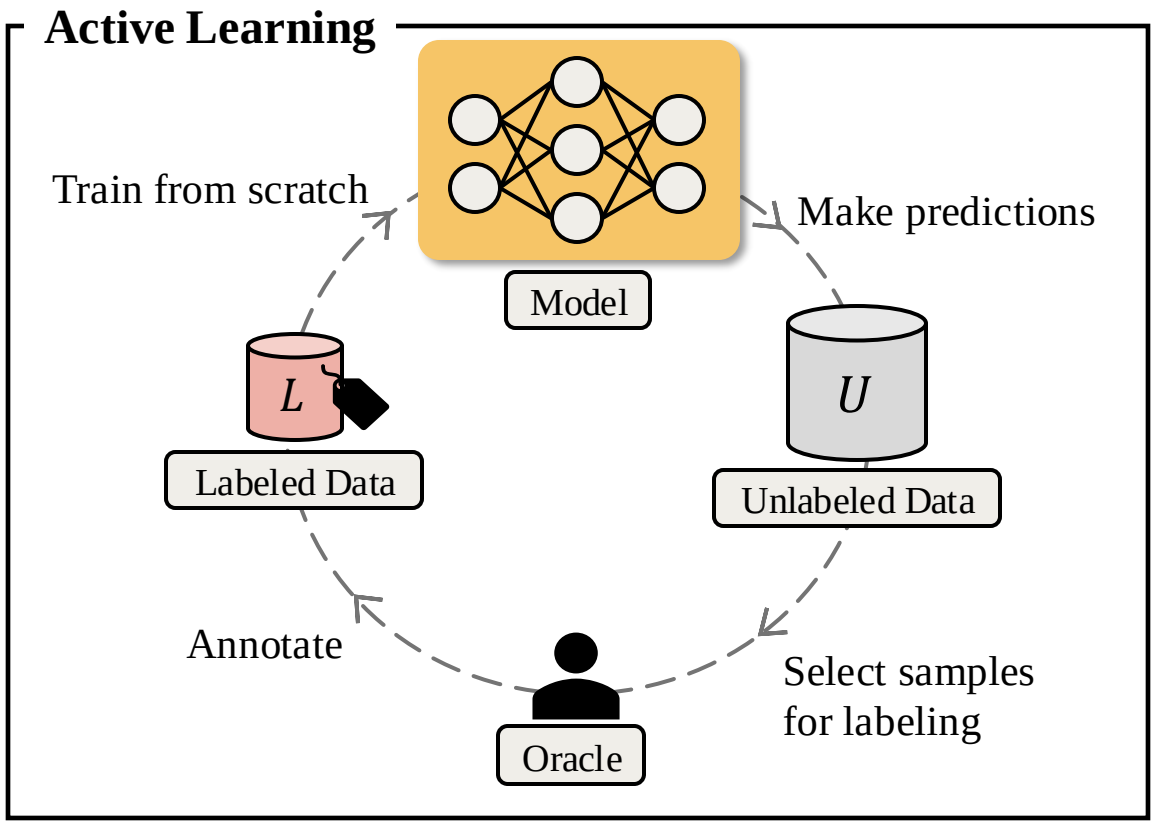
Problem of Continual Learning

- Assuming that the evolving data distributions are **fully** labeled is inaccurate
  - Continuously requesting annotation from experts to recognize new fraud patterns is necessary for fraud detection systems[1]
- Devising a method to mitigate the **limited labeling budget** in CL scenarios is necessary



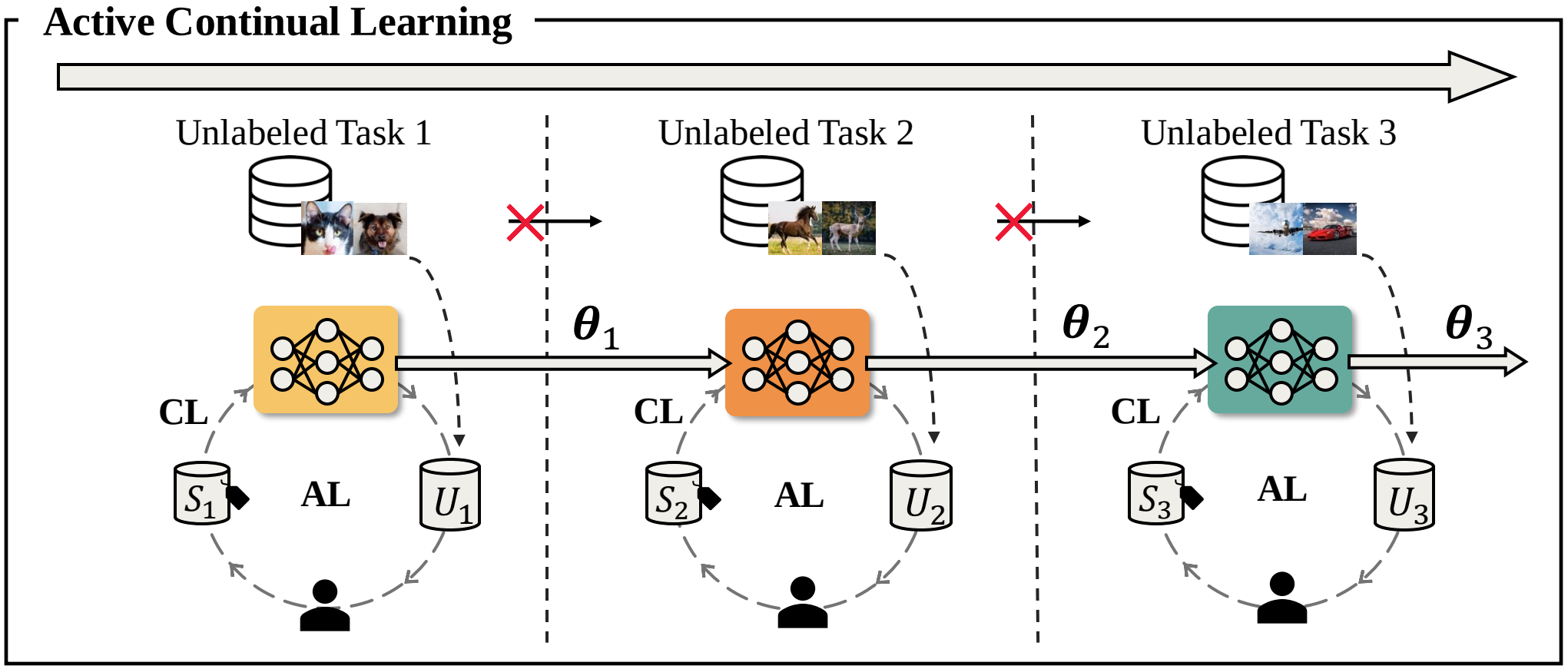
Active Learning (AL)

- Select samples under a **labeling budget** to maximize performance by defining an informativeness measurement to assess each sample's importance



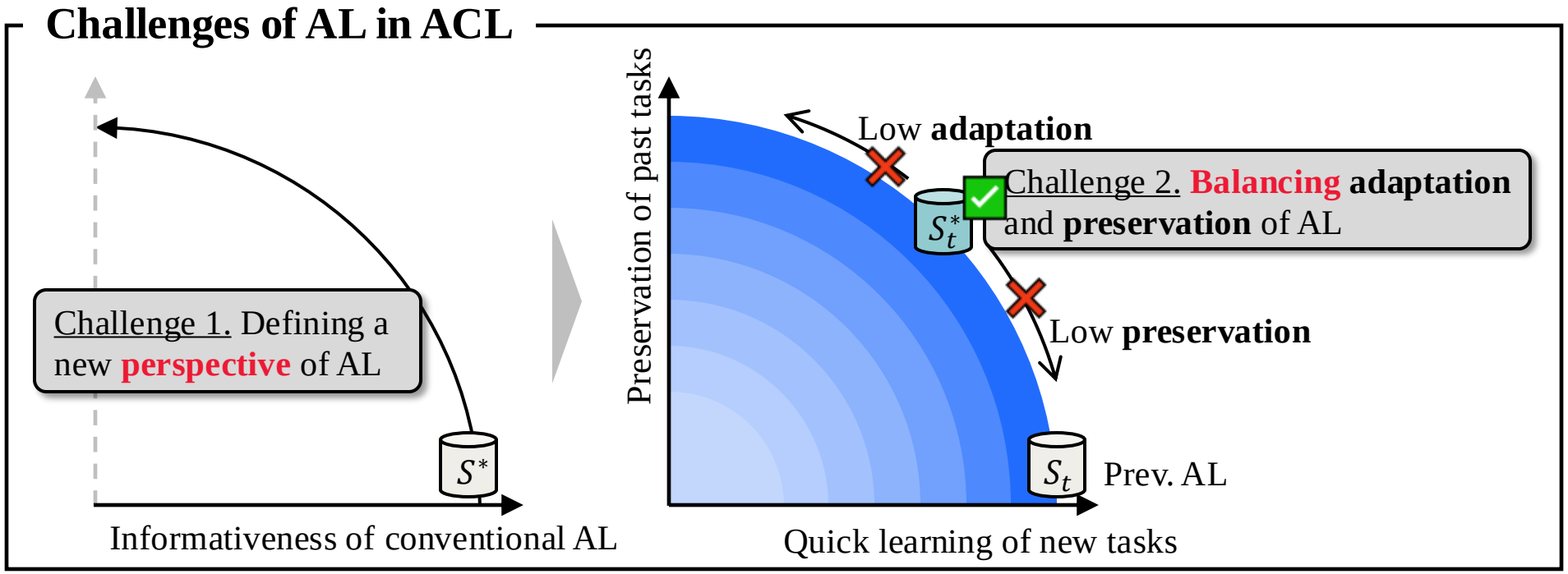
Active Continual Learning (ACL)

- A key problem for effectively mitigating the labeling budget, by querying the most important examples at **each CL task** that maximize the model's performance



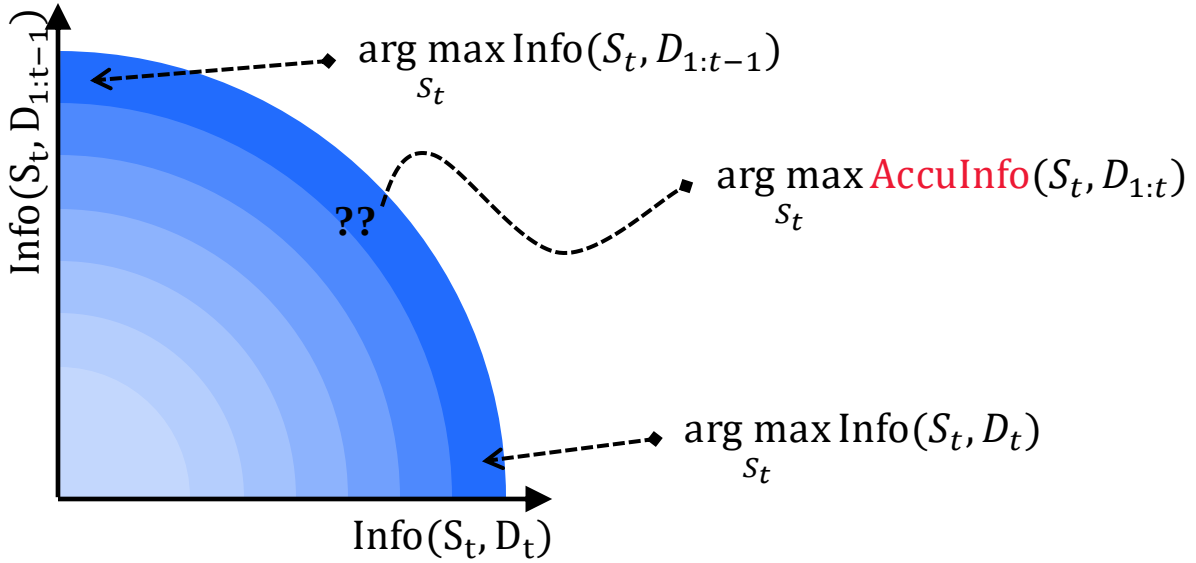
Challenges

- Conventional AL algorithms are prone to catastrophic forgetting!
  - AL does not assume distribution shifts
  - Focuses on quickly learning new information, leading to loss of past information
- *Find unlabeled samples in the **new** task that preserves the knowledge of the **past***



Idea. Accumulated Informativeness

- $\mathrm{Info}(\mathrm{S_t}; D) = \mathbb{E}_{(x,y)\sim\mathcal{A}(D)}\left[p(y|\mathbf{x}; \widehat{\boldsymbol{\theta}}_t)\right] \; s.t. \; \widehat{\boldsymbol{\theta}}_t = \arg\min_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{CL}}\left(\boldsymbol{\theta}; \boldsymbol{\theta}_{t-1}, \mathcal{A}(S_t)\right)$
  - Expected likelihood of the model trained by S_t over D
  - $\mathrm{Info}(S_t, D_t)$: Informativeness of S_t w.r.t. the **new** task
  - $\mathrm{Info}(S_t, D_{1:t-1})$: Informativeness of S_t w.r.t. the **past** tasks
- **Accumulated Informativeness**: An arbitrary combination of the two $\mathrm{Info}(\cdot)$
  - $\mathrm{AccuInfo}(\mathrm{S_t}, \mathrm{D_{1:t}}) = f\left(\mathrm{Info}(\mathrm{S_t}; \mathrm{D_t}), \mathrm{Info}(\mathrm{S_t}; \mathrm{D_{1:t-1}})\right)$



Accumulated Info. and Fisher-based ACL

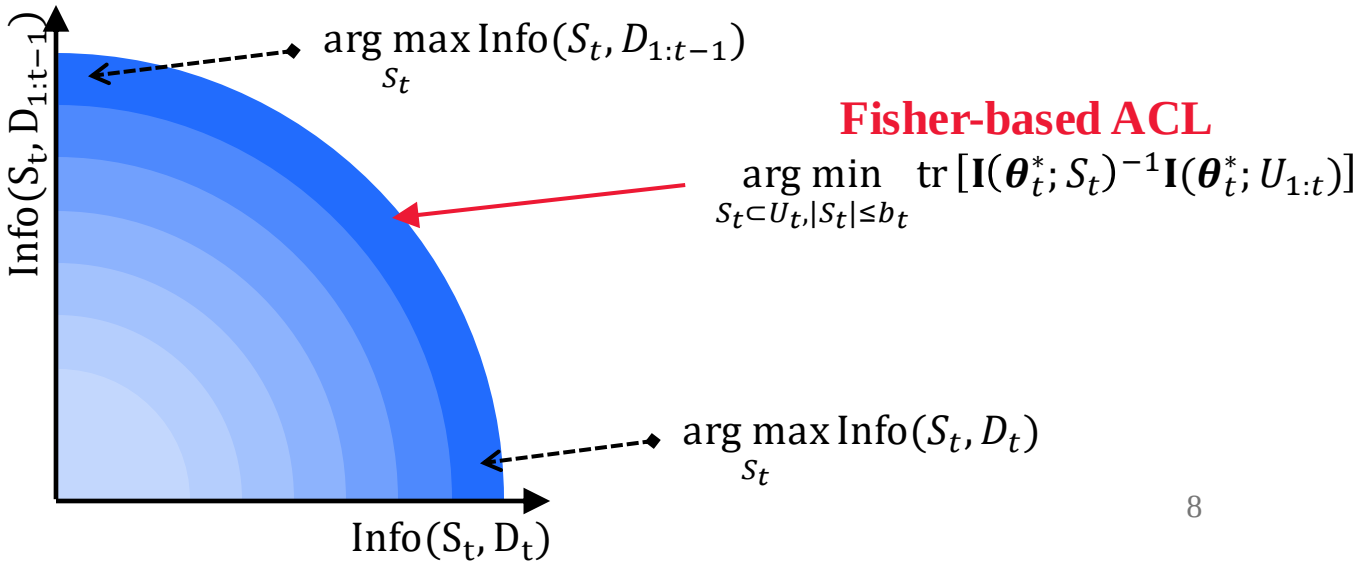
- **Fisher-based ACL** offers a practical form of the arbitrary combination for **accumulated informativeness**, achieving an optimal balance between **adaptation** and **preservation**

$$S_t^* = \underset{S_t \subset U_t, |S_t| \le b_t}{\arg\min} \ \mathrm{tr}\left[\mathbf{I}(\boldsymbol{\theta}_t^*; S_t)^{-1}\mathbf{I}(\boldsymbol{\theta}_t^*; U_{1:t})\right]$$

$$\approx \underset{S_t \subset U_t, |S_t| \le b_t}{\arg\min} \ \lambda \cdot \overbrace{\mathrm{tr}\left[\mathbf{I}(\boldsymbol{\theta}_t^*; S_t)^{-1}\mathbf{I}(\boldsymbol{\theta}_t^*; M_t)\right]}^{\text{Past info.}} + (1-\lambda) \cdot \overbrace{\mathrm{tr}\left[\mathbf{I}(\boldsymbol{\theta}_t^*; S_t)^{-1}\mathbf{I}(\boldsymbol{\theta}_t^*; U_t)\right]}^{\text{New info.}}, \lambda = \frac{|U_{1:t-1}|}{|U_{1:t}|}$$

$$= \underset{S_t \subset U_t, |S_t| \le b_t}{\arg\min} \ \mathrm{tr}\left[\mathbf{I}(\boldsymbol{\theta}_t^*; S_t)^{-1}\mathbf{I}(\boldsymbol{\theta}_t^*; U_t, M_t)\right]$$

$$= \underset{S_t \subset U_t, |S_t| \le b_t}{\arg\min} \ -\mathrm{AccuInfo}(S_t, U_t, M_t)$$



Fisher Information Embedding (FIE)

- Fisher-based ACL is infeasible for large-scale data due to its heavy computation
- Propose the **Fisher information embedding**, the **diagonal** component of the FIM
  - $\mathbf{f}(\boldsymbol{\theta}_t; \mathbf{x}) = \sum_{y \in C} p(y|\mathbf{x}; \boldsymbol{\theta}_t)\left[\nabla_{\boldsymbol{\theta}_t} \log p(y|\mathbf{x}; \boldsymbol{\theta}_t)^2\right] \in \mathbb{R}^{|\boldsymbol{\theta}_t|}$

Fisher Information Embedding

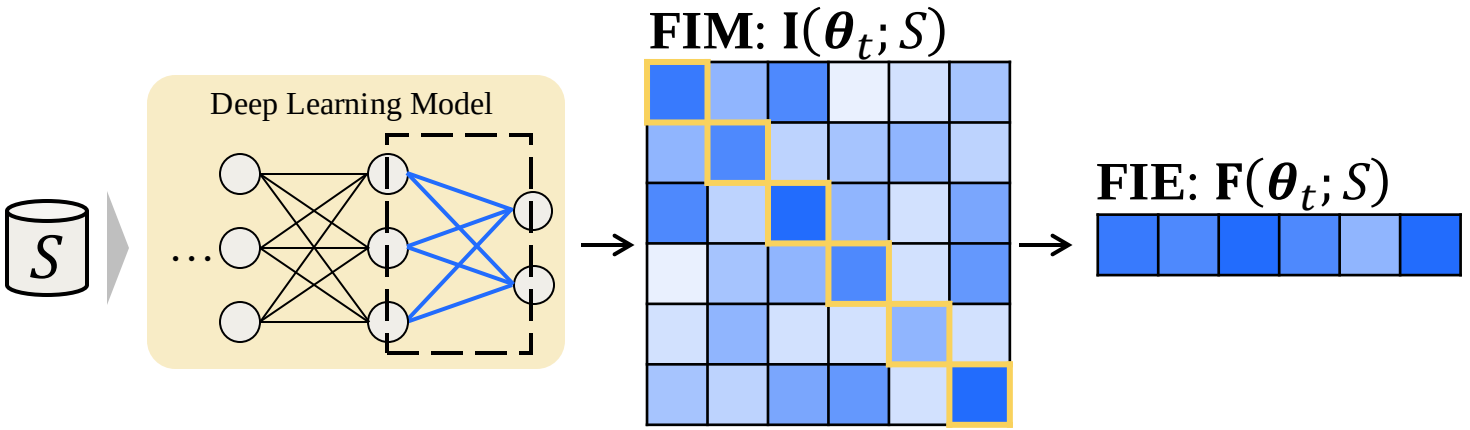
Sample (Thrm. 4.3.2.)

$$\mathbf{f}(\boldsymbol{\theta}_t; \mathbf{x})_{(k,i)} = p_k(1 - p_k)\mathbf{h}(\boldsymbol{\theta}_t; \mathbf{x})_i^2$$

k: class, i: embedding

Subset

$$\mathbf{F}(\boldsymbol{\theta}_t; S) = \frac{1}{|S|}\sum_{\mathbf{x} \in S} \mathbf{f}(\boldsymbol{\theta}_t; \mathbf{x})$$



Fisher-Optimality-Preserving Properties

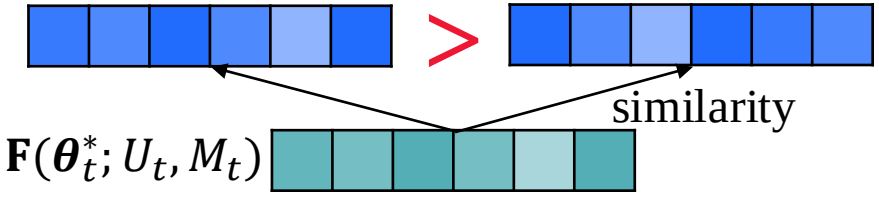
Approximation of Fisher-based ACL

$$S_t^* = \underset{S_t \subset U_t, |S_t| \le b_t}{\arg\min} \ \mathrm{tr}\left[\mathbf{I}(\boldsymbol{\theta}_t; S_t)^{-1} \mathbf{I}(\boldsymbol{\theta}_t; U_t, M_t)\right]$$

$$\xrightarrow{\text{diagonalize}} \underset{S_t \subset U_t, |S_t| \le b_t}{\arg\min} \ \mathbf{F}(\boldsymbol{\theta}_t^*; U_t, M_t) \oslash \mathbf{F}(\boldsymbol{\theta}_t^*; S_t)$$

- Diagonalizing FIM allows two properties of $\mathbf{F}(\boldsymbol{\theta}_t^*; S_t^*)$ to be found:
  - **Property 1.** Position-Wise Optimality:
    - Overall high magnitude of information is advantageous

  - **Property 2.** Distribution-Wise Optimality (Thrm. 4.3.):
    - Aligning the information distribution with the **target** FIE is beneficial when magnitude is the same



Putting Them All Together: AccuACL

- Defining sample-wise scoring metric based on the two properties
  - **Property 1.** Position-Wise Optimality:
    - Magnitude score $\mathcal{M}(\boldsymbol{\theta}_t, \mathbf{x}) = \|\mathbf{f}(\boldsymbol{\theta}_t; \mathbf{x})\|_2$
  - **Property 2.** Distribution-Wise Optimality:
    - Distribution score $\mathcal{D}(\boldsymbol{\theta}_t, \mathbf{x}, M_t, U_t) = \exp(-\mathrm{D}_{\mathrm{JS}}(\sigma(\boldsymbol{f}(\boldsymbol{\theta}_t; \mathbf{x}))||\sigma(\boldsymbol{F}(\boldsymbol{\theta}_t; M_t, U_t))))$
- Overly-sample the subset that ranks highest with $\mathcal{D}$, then further narrow it down with $\mathcal{M}$

Overall Performance (1/2)

- AccuACL consistently outperforms AL baselines for different CL strategies

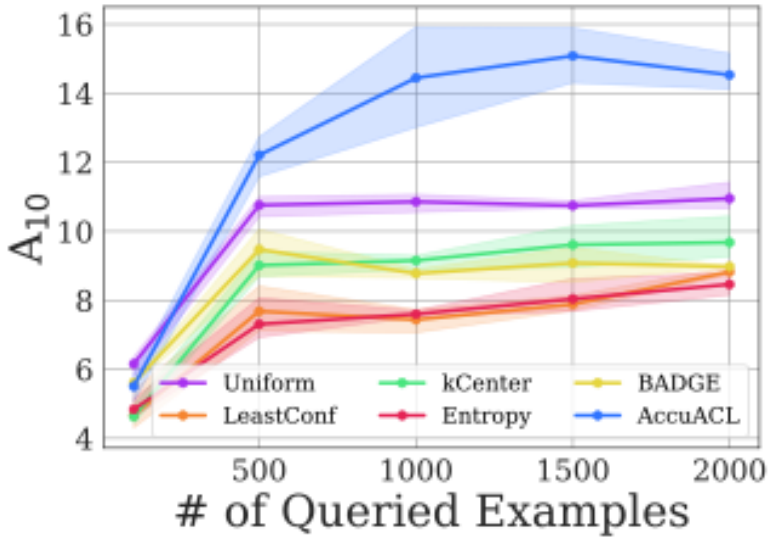
Continual Learning	Active Learning	SplitCIFAR10				SplitCIFAR100				SplitTinyImageNet			
		M=100		M=200		M=500		M=1000		M=2000		M=5000	
		$A_5(\uparrow)$	$F_5(\downarrow)$	$A_5(\uparrow)$	$F_5(\downarrow)$	$A_{10}(\uparrow)$	$F_{10}(\downarrow)$	$A_{10}(\uparrow)$	$F_{10}(\downarrow)$	$A_{10}(\uparrow)$	$F_{10}(\downarrow)$	$A_{10}(\uparrow)$	$F_{10}(\downarrow)$
ER	Full	20.1±0.6	93.3±1.2	26.3±3.5	85.9±4.4	12.6±0.1	75.0±0.7	17.9±0.3	68.5±0.6	7.7±0.1	60.4±0.6	11.5±0.2	54.1±0.5
	Uniform	20.4±3.0	77.1±4.8	26.7±3.0	**67.2**±5.2	10.9±0.4	63.7±0.6	16.7±0.4	56.7±0.3	6.8±0.2	45.3±0.3	8.9±0.3	42.5±0.6
	Entropy	19.7±1.2	81.7±0.7	23.6±1.8	76.5±2.3	8.5±0.3	64.7±0.6	11.0±0.4	61.8±0.5	4.7±0.1	44.1±0.2	5.5±0.2	42.8±0.2
	LeastConf	19.9±1.4	81.0±0.6	22.5±1.7	76.4±0.8	8.8±0.1	66.3±0.2	11.3±0.3	63.5±0.1	4.8±0.3	44.4±0.7	5.7±0.1	42.7±0.7
	kCenter	19.4±1.2	**76.2**±1.5	23.4±2.1	71.8±1.6	9.7±0.7	66.1±0.2	14.7±0.7	60.0±0.2	6.0±0.3	46.1±0.7	7.4±0.3	44.6±0.4
	BADGE	19.4±1.7	81.0±1.7	25.1±1.5	73.5±1.4	9.0±0.0	66.4±0.3	12.5±0.3	62.2±0.7	5.8±0.2	45.5±0.3	7.2±0.1	43.0±0.8
	BAIT	18.4	82.3	23.3	76.5	*	*	*	*	*	*	*	*
	AccuACL	**20.7**±1.0	77.9±1.2	**26.9**±0.2	70.3±0.1	**14.1**±0.7	**55.8**±0.9	**22.0**±1.1	**44.5**±1.5	**7.3**±0.0	**41.9**±0.3	**10.5**±1.0	**37.5**±1.0
GSS	Full	22.9±0.3	88.9±0.6	27.8±2.6	82.0±3.4	10.1±0.6	67.9±0.5	10.8±0.7	67.3±1.2	7.2±0.3	54.5±0.4	8.0±0.4	53.0±1.2
	Uniform	19.7±1.0	76.7±2.8	23.6±1.9	71.7±2.1	7.9±0.4	57.6±1.6	7.9±0.3	57.4±0.3	5.3±0.1	42.0±0.2	5.3±0.2	42.1±0.2
	Entropy	18.0±0.6	75.4±3.1	17.1±1.5	76.4±3.0	7.0±0.3	57.2±0.6	7.3±0.3	56.1±0.2	4.0±0.2	**39.1**±0.8	4.3±0.2	40.0±0.1
	LeastConf	18.4±1.4	77.8±3.3	20.6±1.6	72.0±5.5	7.1±0.1	58.2±0.7	7.2±0.2	57.1±0.3	3.9±0.2	40.0±1.3	4.3±0.3	39.8±1.7
	kCenter	19.1±0.6	77.8±1.7	19.6±0.8	75.1±3.3	7.1±0.5	59.3±1.2	7.5±0.6	56.2±4.6	5.1±0.2	41.2±1.2	5.1±0.3	42.1±0.4
	BADGE	18.6±0.9	78.6±1.9	20.6±1.7	74.1±5.8	7.7±0.5	57.7±0.6	7.4±0.7	57.5±2.0	4.5±0.3	40.4±0.7	4.5±0.2	41.3±0.7
	BAIT	17.5	81.8	16.6	76.8	*	*	*	*	*	*	*	*
	AccuACL	**26.5**±0.7	**68.2**±2.2	**30.0**±0.6	**61.3**±1.8	**8.4**±0.4	**53.7**±1.7	**8.4**±0.4	**54.3**±1.0	**5.7**±0.4	39.8±1.0	**5.8**±0.2	**38.4**±1.7

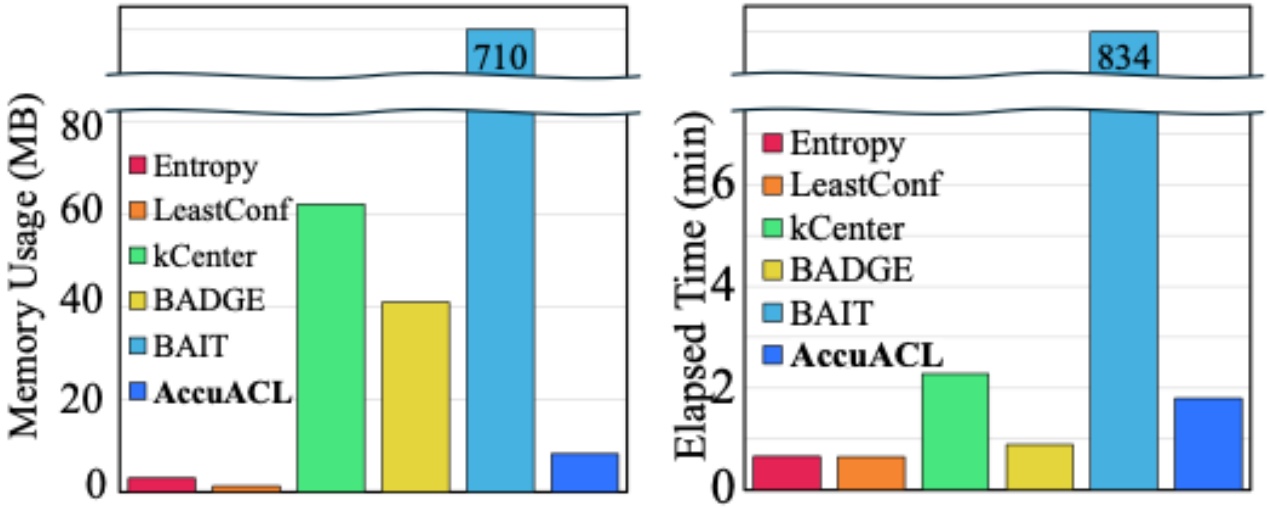
Overall Performance (2/2)

- AccuACL consistently outperforms AL baselines for different CL strategies

Continual Learning	Active Learning	SplitCIFAR10				SplitCIFAR100				SplitTinyImageNet			
		M=100		M=200		M=500		M=1000		M=2000		M=5000	
		$A_5(\uparrow)$	$F_5(\downarrow)$	$A_5(\uparrow)$	$F_5(\downarrow)$	$A_{10}(\uparrow)$	$F_{10}(\downarrow)$	$A_{10}(\uparrow)$	$F_{10}(\downarrow)$	$A_{10}(\uparrow)$	$F_{10}(\downarrow)$	$A_{10}(\uparrow)$	$F_{10}(\downarrow)$
DER++	Full	40.0±1.1	68.6±1.3	48.7±1.1	57.5±1.1	30.6±1.2	51.1±0.9	40.1±1.4	38.0±1.0	10.3±0.3	55.3±1.1	19.6±0.1	32.2±0.2
	Uniform	39.2±0.4	49.0±1.5	49.6±1.2	31.9±2.1	27.6±0.9	38.9±1.4	35.9±0.7	21.5±0.5	11.3±0.2	29.0±0.4	15.2±0.1	10.7±0.3
	Entropy	32.3±0.6	62.9±0.2	47.5±4.2	38.6±4.7	21.3±0.7	48.6±1.0	31.7±0.2	27.5±0.8	8.1±0.1	29.7±1.1	13.1±0.3	9.7±0.4
	LeastConf	33.8±4.2	62.1±5.6	45.2±2.6	42.1±3.3	22.1±0.6	48.0±1.5	33.1±1.0	27.0±0.6	8.5±0.3	28.9±0.7	13.3±0.6	**9.5**±0.5
	kCenter	37.0±1.1	55.1±2.3	47.0±1.6	39.6±3.5	25.9±0.0	43.4±0.3	35.0±0.7	24.9±0.5	10.7±0.1	27.9±0.8	14.4±0.4	11.2±0.5
	BADGE	36.4±2.3	57.9±0.6	**51.0**±2.8	35.8±3.1	24.8±0.4	45.6±1.0	34.1±1.0	27.7±0.7	9.7±0.1	28.5±0.8	14.7±0.2	10.8±0.2
	BAIT	36.7	56.5	49.7	36.4	*	*	*	*	*	*	*	*
	AccuACL	**44.2**±4.6	**40.4**±6.1	50.1±2.6	**28.1**±1.8	**30.5**±0.2	**27.0**±0.4	**36.3**±0.4	**15.0**±0.5	**12.5**±0.4	**24.0**±0.8	**15.7**±0.6	11.4±0.3
ACE	Full	57.6±1.2	27.9±0.3	63.7±0.5	22.4±1.0	34.9±1.2	34.6±0.8	40.1±0.7	30.5±0.8	16.8±0.4	36.5±0.7	20.2±0.3	30.9±0.2
	Uniform	41.3±1.3	25.9±1.8	**49.6**±1.6	20.0±3.6	**28.4**±0.4	30.0±0.6	**34.2**±0.6	25.5±1.1	12.3±1.0	27.6±0.9	14.6±0.2	23.2±0.5
	Entropy	42.7±1.0	30.3±1.6	47.0±2.5	28.8±2.9	24.9±0.4	37.1±0.6	31.5±0.5	31.4±0.7	9.5±0.4	27.8±0.5	11.9±0.2	23.4±0.6
	LeastConf	41.8±2.4	32.5±1.1	47.4±1.6	27.8±2.4	25.7±0.4	35.7±0.3	30.9±1.0	31.7±0.6	9.9±0.3	28.5±0.2	11.8±0.2	23.9±0.6
	kCenter	36.8±0.6	33.0±2.2	43.2±3.0	29.6±4.6	27.4±0.6	34.0±0.7	33.1±0.7	28.8±0.9	11.4±0.2	29.4±0.5	13.7±0.3	25.3±0.1
	BADGE	41.3±2.4	32.3±1.9	47.8±0.9	26.7±0.7	26.5±0.3	35.9±0.4	33.4±0.7	30.3±0.7	11.1±0.4	28.5±0.5	13.5±0.3	24.5±0.5
	BAIT	41.2	33.3	48.0	28.3	*	*	*	*	*	*	*	*
	AccuACL	**43.7**±1.7	**20.8**±1.3	48.1±1.1	**17.7**±1.4	**28.4**±0.6	**26.1**±0.4	33.9±1.0	**20.9**±1.2	**13.4**±0.1	**24.9**±0.5	**16.1**±0.4	**20.5**±0.2

Experiments





- AccuACL shows superior performance for various labeling budget
- AccuACL is able to achieve SOTA ACL performance with reasonable complexity

Thank You!

Jaehyun Park *Dongmin Park* *Jae-Gil Lee*