A Theoretical Study on Solving Continual Learning

Gyuhak Kim*, Changnan Xiao*, Tatsuya Konishi, Zixuan Ke, Bing Liu

(* equal contribution)

Key Results

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 $f : \mathbf{X} \rightarrow \mathbf{Y}$ that can identify the class label y of each given test instance x.

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 $f : \mathbf{X} \times \mathbf{T} \rightarrow \mathbf{Y}$ to identify the class label $y \in \mathbf{Y}_k$ for (x, k) (the given test instance x from task k).

- **Key results:** Identify/prove ***necessary*** & ***sufficient*** conditions for solving CIL.
- New CIL methods designed based on the theoretical results. They outperform strong baselines in both CIL and TIL settings by a large margin.

CIL Decomposition

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

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

$$= \mathbf{P}(x \in \mathbf{X}_{k_0,j_0}|x \in \mathbf{X}_{k_0}, D)\mathbf{P}(x \in \mathbf{X}_{k_0}|D)$$

WP (i.e., TIL) TP

Upper Bound of CIL Loss

- The loss of CIL is bounded by that of WP and TP

Theorem 1. *If* $H_{TP}(x) \leq \delta$ *and* $H_{WP}(x) \leq \epsilon$, *we have* $H_{CIL}(x) \leq \epsilon + \delta$.

- CIL improves with WP or TP

Upper Bound of CIL Loss

- TP and out-of-distribution (OOD) detection bound each other

Theorem 2. *i) If* $H_{TP}(x) \leq \delta$, *let* $\mathbf{P}'_k(x \in \mathbf{X}_k|D) = \mathbf{P}(x \in \mathbf{X}_k|D)$, *then* $H_{OOD,k}(x) \leq \delta, \forall\, k = 1, \ldots, T$. *ii) If* $H_{OOD,k}(x) \leq \delta_k, k = 1, \ldots, T$, *let* $\mathbf{P}(x \in \mathbf{X}_k|D) = \frac{\mathbf{P}'_k(x \in \mathbf{X}_k|D)}{\sum_k \mathbf{P}'_k(x \in \mathbf{X}_k|D)}$, *then* $H_{TP}(x) \leq (\sum_k \mathbf{1}_{x \in \mathbf{X}_k} e^{\delta_k})(\sum_k 1 - e^{-\delta_k})$, *where* $\mathbf{1}_{x \in \mathbf{X}_k}$ *is an indicator function.*

- The loss of CIL is bounded by that of WP and OOD

Theorem 3. *If* $H_{OOD,k}(x) \leq \delta_k,\ k = 1, \ldots, T$ *and* $H_{WP}(x) \leq \epsilon$, *we have*

$$H_{CIL}(x) \leq \epsilon + (\sum_k \mathbf{1}_{x \in \mathbf{X}_k} e^{\delta_k})(\sum_k 1 - e^{-\delta_k}),$$

where $\mathbf{1}_{x \in \mathbf{X}_k}$ *is an indicator function.*

Necessary Condition for CIL

- We just showed that good performances of WP and TP or (OOD) are *sufficient* to guarantee a good CIL

- This theorem shows that good performances of WP and TP (or OOD) are *necessary* for a good CIL

Theorem 4. *If* $H_{CIL}(x) \leq \eta$, *then there exist i) a WP, s.t.* $H_{WP}(x) \leq \eta$, *ii) a TP, s.t.* $H_{TP}(x) \leq \eta$, *and iii) an OOD detector for each task, s.t.* $H_{OOD,k} \leq \eta$, $k = 1, \ldots, T$.

Empirical Validation

- OOD methods improve both WP and OOD (or TP)

- We want to show that OOD improves CIL in two ways:
  - Post-processing CIL models with OOD detection (ODIN)
  - A TIL method + OOD detection (CSI)

Empirical Validation

- CIL accuracy increases and decreases by the OOD detection performance (AUC)

Method	OOD	AUC	CIL
OWM	Original	71.31	28.91
	ODIN	70.06	28.88
MUC	Original	72.69	30.42
	ODIN	72.53	29.79
PASS	Original	69.89	33.00
	ODIN	69.60	31.00
LwF	Original	88.30	45.26
	ODIN	87.11	51.82
BiC	Original	87.89	52.92
	ODIN	86.73	48.65
DER++	Original	85.99	53.71
	ODIN	88.21	55.29
HAT	Original	77.72	41.06
	ODIN	77.80	41.21
HyperNet	Original	71.82	30.23
	ODIN	72.32	30.83
Sup	Original	79.16	44.58
	ODIN	80.58	46.74

Proposed Methods

- Training existing TIL methods with a strong OOD detection method (CSI)
  - HAT + CSI
  - SupSup + CSI
- Better OOD method (CSI) results in better CIL

CL	OOD	C10-5T		C100-10T		C100-20T		T-5T		T-10T	
		AUC	CIL	AUC	CIL	AUC	CIL	AUC	CIL	AUC	CIL
HAT	ODIN	82.5	62.6	77.8	41.2	75.4	25.8	72.3	38.6	71.8	30.0
	CSI	91.2	87.8	84.5	63.3	86.5	54.6	76.5	45.7	78.5	47.1
Sup	ODIN	82.4	62.6	80.6	46.7	81.6	36.4	74.0	41.1	74.6	36.5
	CSI	91.6	86.0	86.8	65.1	88.3	60.2	77.1	48.9	79.4	45.7

Proposed Methods - Comparison (CIL)

- The proposed methods outperform the baselines by large margins

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^{2}L		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 Methods - Comparison (TIL)

- The proposed methods are also superior in TIL

Method	M-5T	C10-5T	C100-10T	C100-20T	T-5T	T-10T
DER++	99.7±0.08	92.0±0.54	84.0±9.43	86.6±9.44	57.4±1.31	60.0±0.74
HAT	99.9±0.02	96.7±0.18	84.0±0.23	85.0±0.98	61.2±0.72	63.8±0.41
Sup	99.6±0.01	96.6±0.21	87.9±0.27	91.6±0.15	64.3±0.24	68.4±0.22
HAT+CSI	99.9±0.00	98.7±0.06	92.0±0.37	94.3±0.06	68.4±0.16	72.4±0.21
Sup+CSI	99.0±0.08	98.7±0.07	93.0±0.13	95.3±0.20	65.9±0.25	74.1±0.28

Conclusion

- Showed **necessary** and **sufficient** condition for good CIL performances
- Agnostic to any specific implementation and applicable to any CIL problems (e.g., offline, online, blurry task, etc.)
- Provided a principled guidance on how to design CIL algorithms
- Proposed several CIL methods superior to strong baselines