Uncertainty-based Continual Learning with Adaptive Regularization

Hongjoon Ahn* [1], Sungmin Cha* [2], Donggyu Lee [2], and Taesup Moon [1,2]

Department of Artificial Intelligence[1], Department of Electrical and Computer Engineering[2]

Sungkyunkwan University

{hong0805, csm9493, ldk308, tsmoon}@skku.edu

Project page

Regularization-based Continual Learning

- **Regularization** prevents **catastrophic forgetting** by penalizing large updates of the important parameters for previous tasks

Current task loss

$$\underbrace{-\log p(Y_t|X_t,w_t)}_{\text{Current task loss}} + \|\boxed{\Omega_{t-1}} \odot (w_t - w_{t-1})\|_2^2$$

Regularization penalty for parameters

- **Caveats:**
  - **Large** memory cost (e.g., EWC, SI, Riemannian-walk, etc.)
  - Regularization penalty **not learnable**
  - **No** mechanism for gracefully forgetting

Bayesian Online Learning

- **A fresh interpretation of the KL-term in the ELBO**

$$\mathcal{F}(D_t,\boldsymbol{\theta}_t) = \mathbb{E}_{q(\mathcal{W}|\boldsymbol{\theta}_t)}[-\log p(D_t|\mathcal{W})] + D_{KL}(q(\mathcal{W}|\boldsymbol{\theta}_t)||q(\mathcal{W}|\boldsymbol{\theta}_{t-1}))$$

- For $q(\mathcal{W}|\boldsymbol{\theta}) = \prod_i \mathcal{N}(w_i|\mu_i,\sigma_i)$ $\quad \boldsymbol{\theta}_t^{(l)} = (\boldsymbol{\mu}_t^{(l)}, \boldsymbol{\sigma}_t^{(l)})$:

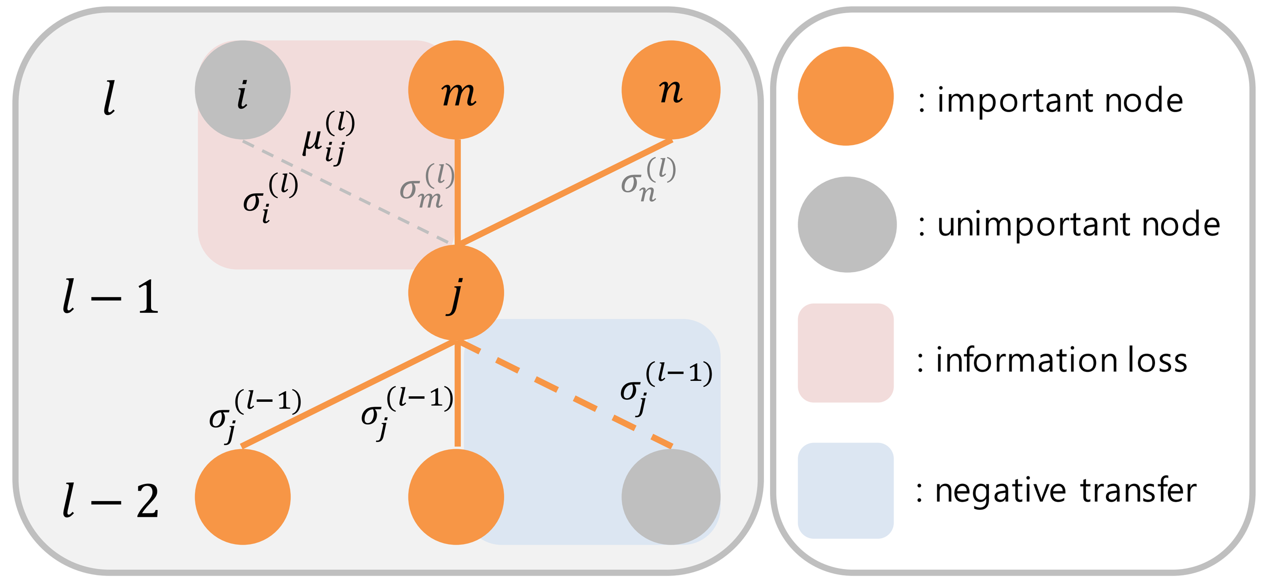
$$\frac{1}{2}\sum_{l=1}^{L}\Bigg[\underbrace{\Bigg\|\frac{\boldsymbol{\mu}_t^{(l)}-\boldsymbol{\mu}_{t-1}^{(l)}}{\boldsymbol{\sigma}_{t-1}^{(l)}}\Bigg\|_2^2}_{(a)} + \underbrace{\mathbf{1}^\top\Bigg\{\Bigg(\frac{\boldsymbol{\sigma}_t^{(l)}}{\boldsymbol{\sigma}_{t-1}^{(l)}}\Bigg)^2 - \log\Bigg(\frac{\boldsymbol{\sigma}_t^{(l)}}{\boldsymbol{\sigma}_{t-1}^{(l)}}\Bigg)^2\Bigg\}}_{(b)}\Bigg]$$

Closed form

- **Term (a)**: regularization for $\boldsymbol{\mu}_t^{(l)}$ (mean parameter)
  - $\boldsymbol{\sigma}_{t-1}^{(l)}$: ***Uncertainty*** measure for $\boldsymbol{\mu}_{t-1}^{(l)}$
  - A parameter with **High**/**Low** uncertainty gets **Weak**/**Strong** regularization!
- **Term (b)**: regularization for $\boldsymbol{\sigma}_t^{(l)}$ (std of parameter)
  - Enforces the uncertainty to **stay the same**!
- Cf.) **VCL**: Uses the same ELBO and variational inference
  - **Huge memory cost:** Requires twice the memory to store $\boldsymbol{\sigma}_t^{(l)}$
  - **Multiple number of samplings:** Slow, No RL results

Uncertainty-based Continual Learning (UCL)

- **Information loss** and **negative transfer** cause catastrophic forgetting



Summary of main contributions

- Define the **uncertainty of a node** (**tied** $\boldsymbol{\sigma}$ of incoming weights) → Reduces the # of parameters
- Devise novel loss terms to prevent **information loss** and **negative transfer** via adaptive regularization
- Introduce a novel loss term to induce **gracefully forgetting**

- **Final loss function for UCL**

High regularization strengths on all connected weights of important (certain) nodes
→ Prevent **negative transfer** and **information loss**

$$\boldsymbol{\Lambda}_{ij}^{(l)} \triangleq \max\Bigg\{\frac{\sigma_{\text{init}}^{(l)}}{\sigma_{t-1,i}^{(l)}}, \frac{\sigma_{\text{init}}^{(l-1)}}{\sigma_{t-1,j}^{(l-1)}}\Bigg\}$$

$$-\log p(D_t|\mathcal{W}) + \sum_{l=1}^{L}\Bigg[\Bigg(\frac{1}{2}\Big\|\boldsymbol{\Lambda}^{(l)}\odot(\boldsymbol{\mu}_t^{(l)}-\boldsymbol{\mu}_{t-1}^{(l)})\Big\|_2^2 + (\sigma_{\text{init}}^{(l)})^2\Bigg\|\Bigg(\frac{\boldsymbol{\mu}_{t-1}^{(l)}}{\boldsymbol{\sigma}_{t-1}^{(l)}}\Bigg)^2\odot(\boldsymbol{\mu}_t^{(l)}-\boldsymbol{\mu}_{t-1}^{(l)})\Bigg\|_1\Bigg) \quad (5)$$

$$+\frac{\beta}{2}\mathbf{1}^\top\Bigg\{\Bigg(\frac{\boldsymbol{\sigma}_t^{(l)}}{\boldsymbol{\sigma}_{t-1}^{(l)}}\Bigg)^2 - \log\Bigg(\frac{\boldsymbol{\sigma}_t^{(l)}}{\boldsymbol{\sigma}_{t-1}^{(l)}}\Bigg)^2 \underset{(6)}{+} (\boldsymbol{\sigma}_t^{(l)})^2 - \log(\boldsymbol{\sigma}_t^{(l)})^2\Bigg\}\Bigg], \quad (7)$$

Sample only once

Modification of **Term (b)**

Enables the uncertainty of a node grow again
→ Intend **gracefully forgetting**

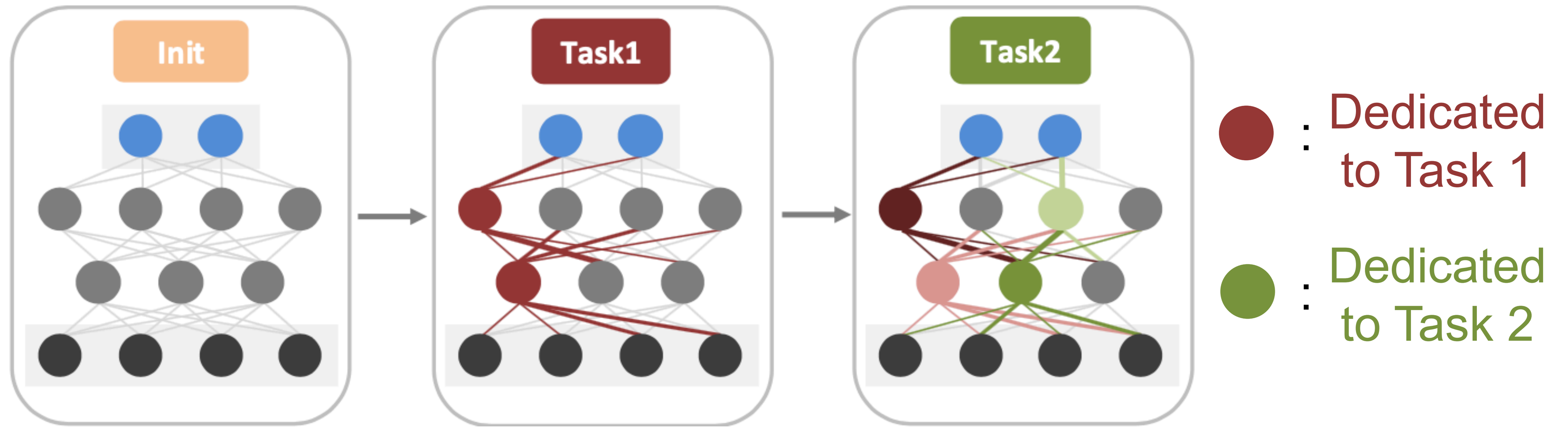
Modification of **Term (a)**

Freeze the important weights
→ Prevent **negative transfer**

- **Illustration of the regularization mechanism of UCL**

1. Randomly initialized weights

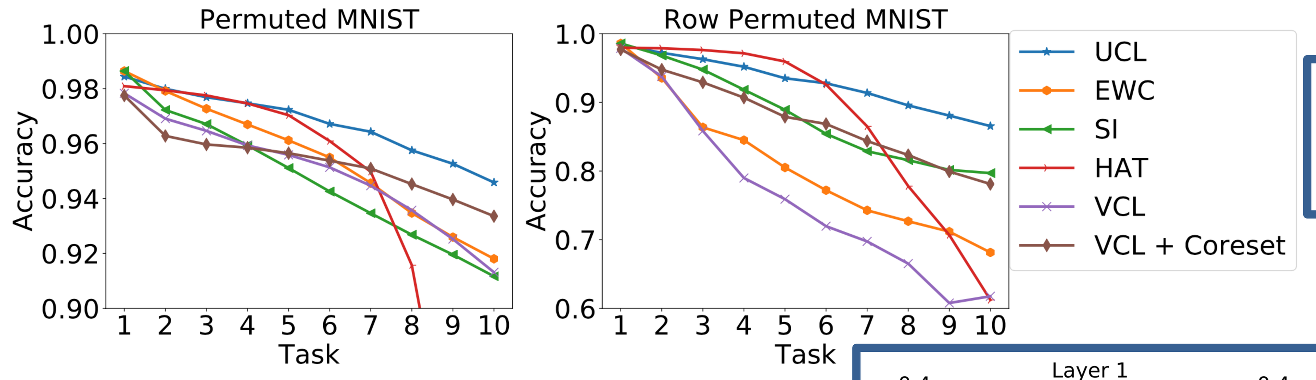
3. Result of **gracefully forgetting**



2. Small σ for important nodes

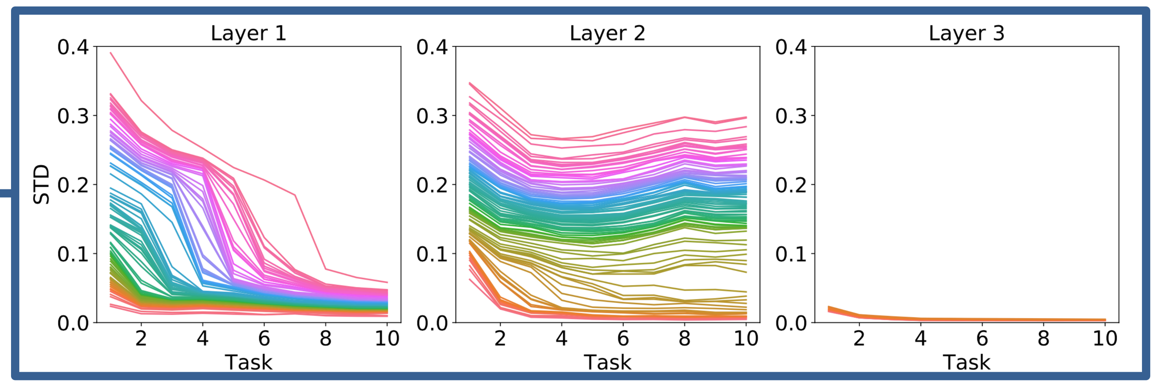
Experimental Results

- **Permuted MNIST & Row permuted MNIST (FCNN)**

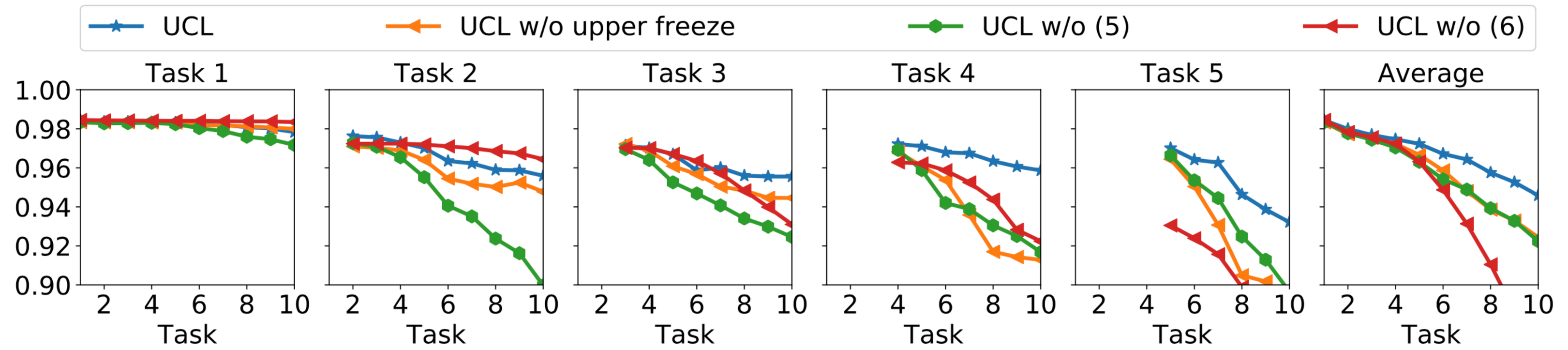


UCL achieves SOTA in various SL tasks!

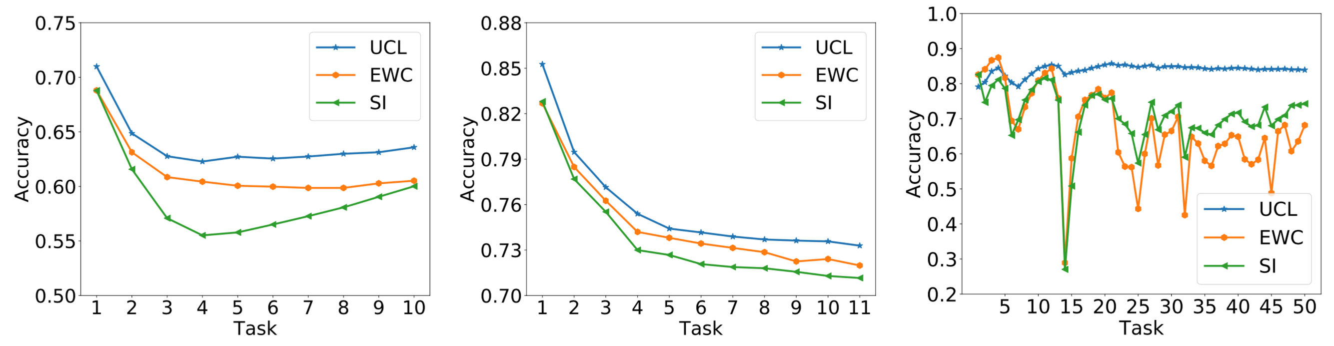
# of uncertain nodes changes as learning continues (gracefully forgetting)



- Ablation study for Permuted MNIST

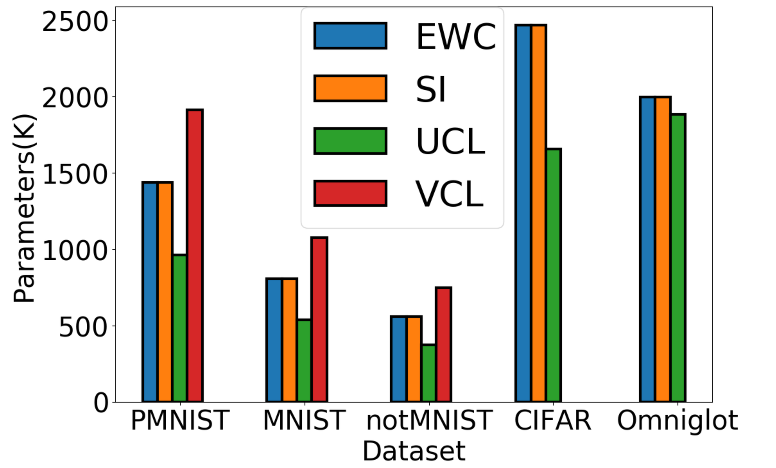


- **Vision datasets (Deep CNN)**

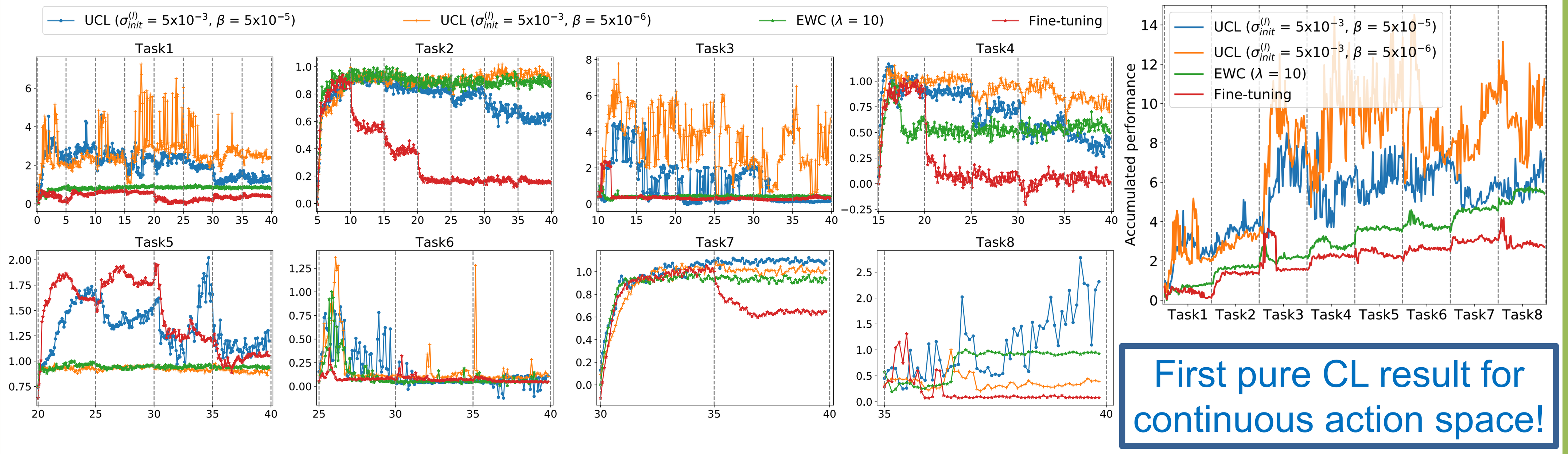


(a) Split CIFAR-100 (b) Split CIFAR-10/100 (c) Omniglot

- **# of parameters**



- **Roboschool RL tasks (FCNN, algorithm: PPO)**



First pure CL result for continuous action space!