BrAIn Lab.
Brain and Artificial Intelligence Lab.

SEOULTECH
SEOUL NATIONAL UNIVERSITY OF SCIENCE & TECHNOLOGY

Overcoming catastrophic forgetting in neural networks

James Kirkpatrick

2026.01.05

Introduction to Continual Learning

BrainLAB Journal Club
인공지능응용학과 허세민

Overview

00 Author & Journal

James Kirkpatrick

팔로우

Deepmind, University of Oxford, Imperial College, Max Planck Institute for Polymer Research
google.com의 이메일 확인됨

Physics solar cells reinforcment learning structural biology DFT

제목	인용	연도
Overcoming catastrophic forgetting in neural networks J Kirkpatrick, R Pascanu, N Rabinowitz, J Veness, G Desjardins, AA Rusu, ... Proceedings of the national academy of sciences 114 (13), 3521-3526	11292	2017
Improved protein structure prediction using potentials from deep learning AW Senior, R Evans, J Jumper, J Kirkpatrick, L Sifre, T Green, C Qin, ... Nature 577 (7792), 706-710	4165	2020
Progressive neural networks AA Rusu, NC Rabinowitz, G Desjardins, H Soyer, J Kirkpatrick, ... arXiv preprint arXiv:1606.04671	3878	2016
Policy distillation AA Rusu, SG Colmenarejo, C Gulcehre, G Desjardins, J Kirkpatrick, ... arXiv preprint arXiv:1511.06295	1004	2015

PNAS

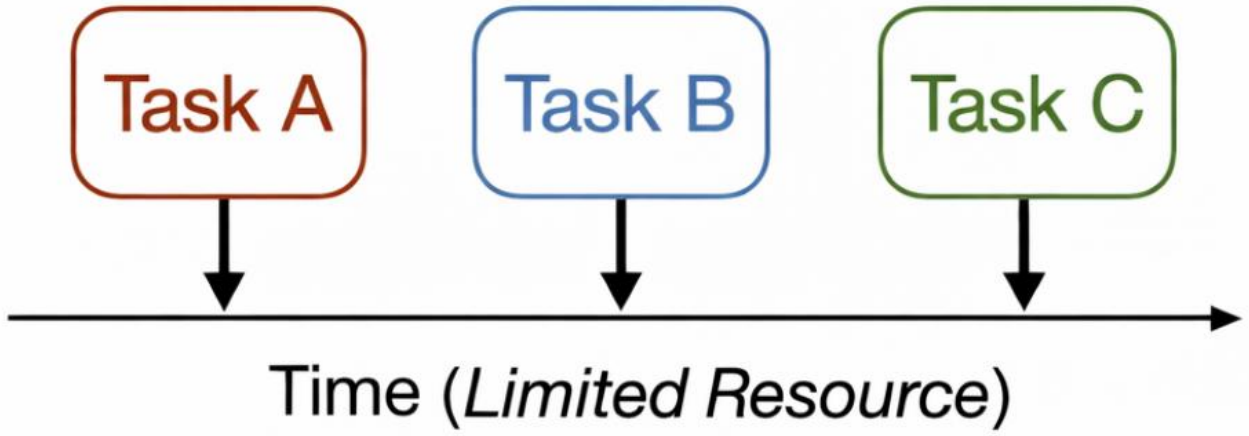
Continual Learning

The Necessity

✓ Dynamic World
- 전통적인 머신러닝: 고정된 데이터셋 가정 (i.i.d).
- 현실 세계: 데이터가 시간 흐름에 따라 순차적으로 입력됨 (Non-stationary).
  - *예: 계절이 바뀌는 CCTV, 새로운 상품이 추가되는 추천 시스템*

✓ Efficiency
- 데이터가 추가될 때마다 전체를 다시 학습하는 것은 시간/비용적으로 비효율적.
- 메모리 제약 및 프라이버시 문제로 과거 데이터를 저장할 수 없는 경우 발생.

✓ Goal: "Learn like humans" (인간처럼 과거의 지식 위에 새로운 지식을 쌓아가는 것)

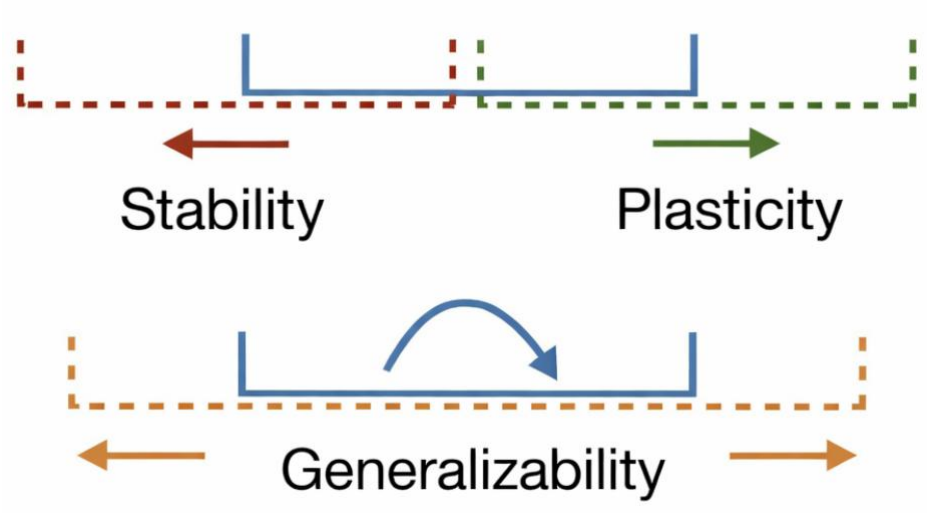


01 The Stability-Plasticity Dilemma

BrAIn Lab.
Brain and Artificial Intelligence Lab.

✓ **Definition**
- Plasticity (가소성): 새로운 지식을 빠르게 습득할 수 있는 능력.
  - *Too High?* → 이전 지식이 덮어씌워짐 (Old knowledge overwritten).
- Stability (안정성): 기존에 학습한 지식을 보존하려는 능력.
  - *Too High?* → 새로운 학습이 안 됨 (New learning inhibited).

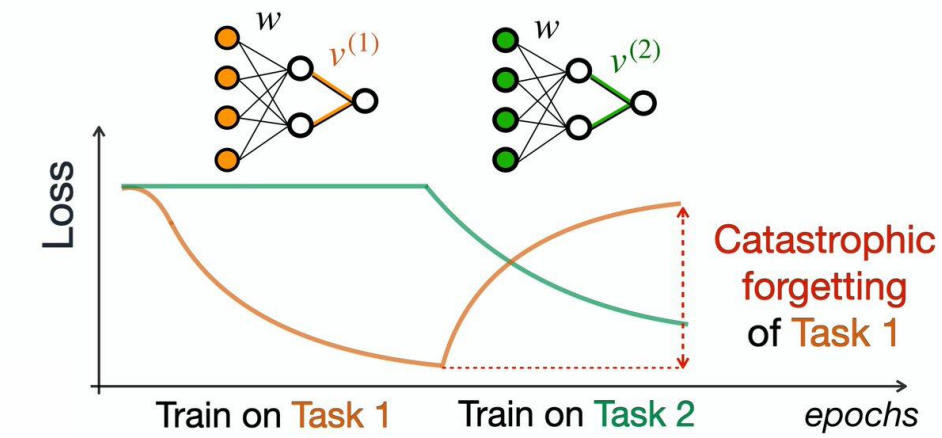
✓ **The Trade-off**
- 이상적인 AI는 이 둘 사이의 균형을 맞춰야 함.
- 실패 시 결과: **Catastrophic Forgetting (파국적 망각).**
  - *가소적(plasticity)이기만 하고 안정적(stability)이지 못하면 발생*



01 Catastrophic Forgetting

✓ **Definition**
- 새로운 태스크(Task B) 학습 시작 즉시, 이전 태스크(Task A) 성능이 급격히 떨어지는 현상
- 단순한 성능 저하가 아니라 이전 과제 성능이 거의 붕괴 수준으로 감소

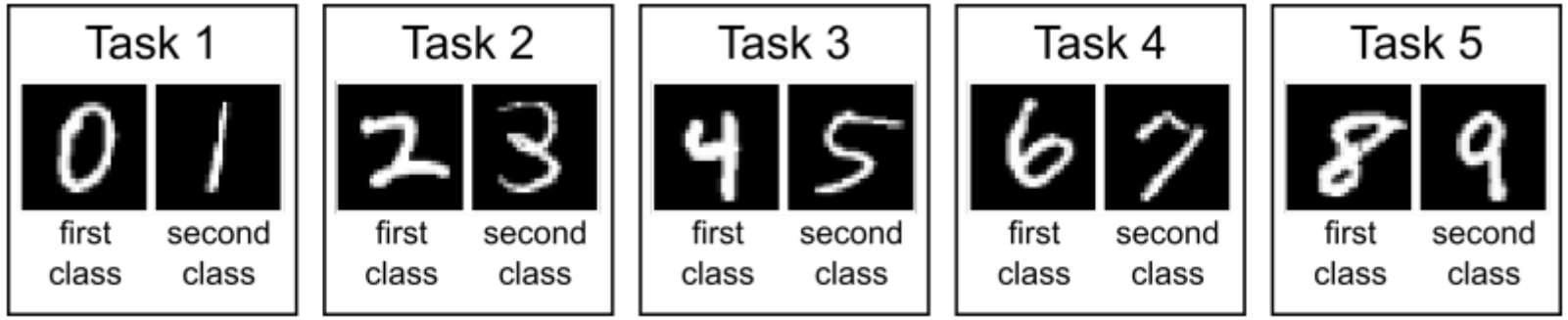
✓ **Cause**
- Shared Parameters
  - 동일한 네트워크 파라미터를 여러 task가 공유
- Non-i.i.d. Data Stream
  - 과거 데이터 없이, 새로운 데이터만 순차적으로 학습
- Gradient Interference
  - 새 task를 위한 업데이트가 이전 task에 중요했던 가중치를 덮어씀 (overwrite)



01 Three Scenarios

BrAIn Lab.
Brain and Artificial Intelligence Lab.

구분	Task-Incremental (Task-IL)	Domain-Incremental (Domain-IL)	Class-Incremental (Class-IL)
상황	태스크 ID가 주어짐.	태스크 ID 없음. 도메인만 바뀜.	태스크 ID 없음. 새로운 클래스 추가.
output	Multi-head (태스크별 별도 출력)	Single-head (출력 구조 동일)	Single-head (출력 구조 확장)
난이도	쉬움	중간	가장 어려움



01 CL vs. Other Paradigms

BrAIn Lab.
Brain and Artificial Intelligence Lab.

Methodology	Data Availability	Task Goal	Constraint	Fate of Old Knowledge
Offline/ Multi-task	All at once	Global Optimality	High Memory & Compute Cost	Preserved
Transfer Learning	Sequentail (Source → Target)	Adaptation	Low Data Regime	Forgotten
Online Learning	Stream (Instance by Instance)	Current Distribution	Real-time Update	Often decayed (Gradual forgetting)
Continual Learning	**Sequential (Task by Task)**	**Accumulation**	**No Access to Old Data**	**Must be Preserved**

Methods

02 3-Pillar Structure

BrAIn Lab.
Brain and Artificial Intelligence Lab.

구분	Regularization-based	Replay-based	Architecture-based
핵심 철학	**“Don’t chage too much”**	**“Remember the past”**	**“Devide and conquer”**
전략	중요 파라미터가 변하지 않도록 **Loss에 제약**을 추가	과거 데이터의 일부를 **저장**하거나 **생성**하여 함께 학습	태스크마다 전용 파라미터를 할당하거나 네트워크를 **확장**
대표 기법	**EWC(today)**, LwF, SI	Experience Replay (ER), GEM	Progressive Neural Nets
장점	메모리 효율적, 프라이버시 보호	SOTA	태스크 간 간섭이 없음
단점	태스크가 길어지면 용량 한계	메모리 비용 발생, 데이터 보안 이슈	모델 사이즈가 계속 커짐

02 Regularization-based Approach

BrAIn Lab.
Brain and Artificial Intelligence Lab.

✓ Mechanism
- 손실 함수에 규제항을 추가하여 급격한 변화를 막음.
- $\mathcal{L}_{total} = \mathcal{L}_{current} + \lambda \cdot \Omega(\theta_{old}, \theta_{new})$
  - $\mathcal{L}_{current}$: 새로운 태스크를 위한 손실
  - Ω : 과거 지식에서 멀어지는 것에 대한 벌점(Penalty)
  - λ : 기억 보존 강도 (Hyperparameter)

구분	**Weight Regularization**	**Function Regularization**
초점	**Internal Parameters** (θ)	**External Output** ($f(x)$)
방법	각 파라미터별 중요도 계산 후 제약	Knowledge Distillation 활용
대표 모델	**EWC**, SI, MAS	**LwF**

02 Regularization-based Approach

BrAIn Lab.
Brain and Artificial Intelligence Lab.

✓ Bayes' Rule
- $\log p(\theta \mid \mathcal{D}) = \log p(\mathcal{D} \mid \theta) + \log p(\theta) - \log p(\mathcal{D})$
  - $p(\theta \mid \mathcal{D})$: Posterior
  - $p(\mathcal{D} \mid \theta)$: Likelihood
  - $p(\theta)$: Prior
  - $p(\mathcal{D})$: Evidence

✓ Sequential Learning
- $\log p(\theta \mid \mathcal{D}_A, \mathcal{D}_B) = \underbrace{\log p(\mathcal{D}_B \mid \theta)}_{\text{Loss for Task B}} + \underbrace{\log p(\theta \mid \mathcal{D}_A)}_{\text{Knowledge from Task A}} - \text{const}$

✓ Intractability
- $\mathrm{p}(\theta \mid \mathcal{D}_A)$ 의 실제 분포를 정확히 계산하는 것은 **불가능**
- 해결책: **Laplace Approximation**
  - 복잡한 분포를 다루기 쉬운 **가우시안(Gaussian)으로 근사**

Elastic Weight Consolidation

03 Laplace Approximation

BrAIn Lab. Brain and Artificial Intelligence Lab.

✓ Taylor Expansion

- $\log p(\theta) \approx \log p(\theta^*) + \underbrace{\nabla_\theta \log p(\theta^*)}_{\text{First Order}} (\theta - \theta^*) + \frac{1}{2} (\theta - \theta^*)^T \underbrace{\nabla_\theta^2 \log p(\theta^*)}_{\text{Second Order}} (\theta - \theta^*)$

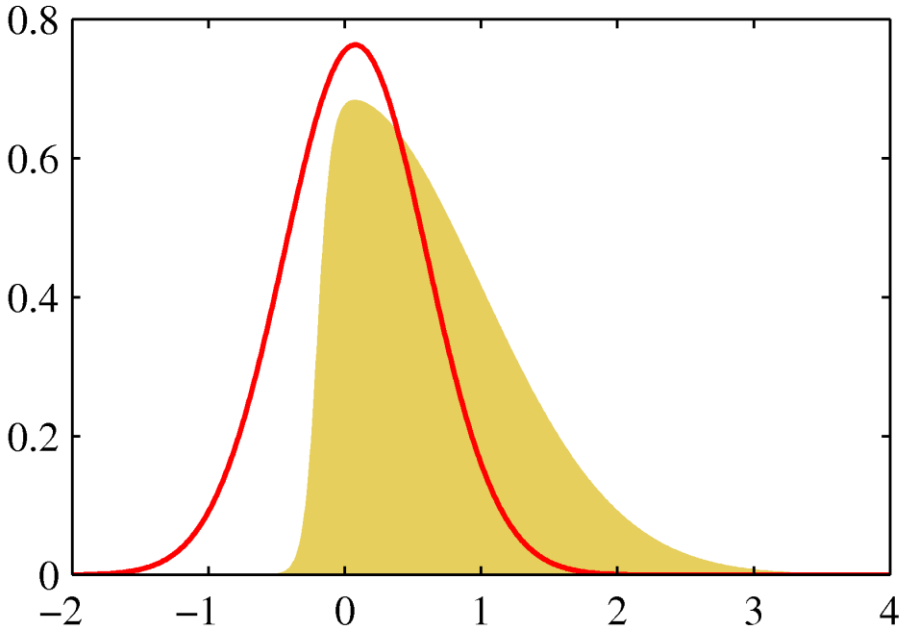
✓ Definition of "Score"

- $S(\theta) = \nabla_\theta \log p(\mathcal{D} \mid \theta)$
- θ^* 는 이미 Task A를 학습하여 찾은 **최적점(Local Maxima)**임.
- $\therefore$ Score at $\theta^* = 0$

✓ Hessian Matrix

- $\log p(\theta) \approx \text{const} - \frac{1}{2}(\theta - \theta^*)^T H(\theta - \theta^*)$

※ $\mathrm{f(x)} = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left[-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu)\right]$



03 Fisher Information Matrix

BrAIn Lab.
Brain and Artificial Intelligence Lab.

✓ Recall Score
- 정의: $S(\theta) = \nabla_\theta \log p(\mathcal{D} \mid \theta)$
- 의미: 파라미터 θ를 살짝 건드렸을 때, 확률 분포가 얼마나 민감하게 반응하는가? (기울기)

✓ Definition of Fisher Information
- 정의: "Covariance of Score"
- $F = E_{x\sim p}[S(\theta)S(\theta)^T]$
- 해석:
  - 기울기의 분산이 크다 → 데이터에 따라 기울기가 요동친다
    → **이 파라미터는 모델의 결정에 아주 중요한 역할을 한다.**

✓ Connection to Curvature
- High Fisher ($F \uparrow$):
  - Score의 분산이 큼 → Loss곡선이 뾰족함
  - 의미: "조금만 건드려도 망가짐" → 중요한 파라미터
- Low Fisher ($F \downarrow$):
  - Score의 분산이 작음 → Loss 곡선이 평평함
  - 의미: "건드려도 별 차이 없음" → 덜 중요한 파라미터

03 Fisher Information Matrix

BrAIn Lab.
Brain and Artificial Intelligence Lab.

✓ Fisher Information Matrix
- "Variance of the Score" $Var(S) = E[S^2]$

✓ Practical Calculation Steps
1. Task A의 데이터(x, y)를 모델에 넣고 Loss를 구한다.
2. 각 데이터 샘플마다 파라미터의 Gradient를 계산하고 제곱한다.
3. 모든 데이터에 대해 평균을 낸다.

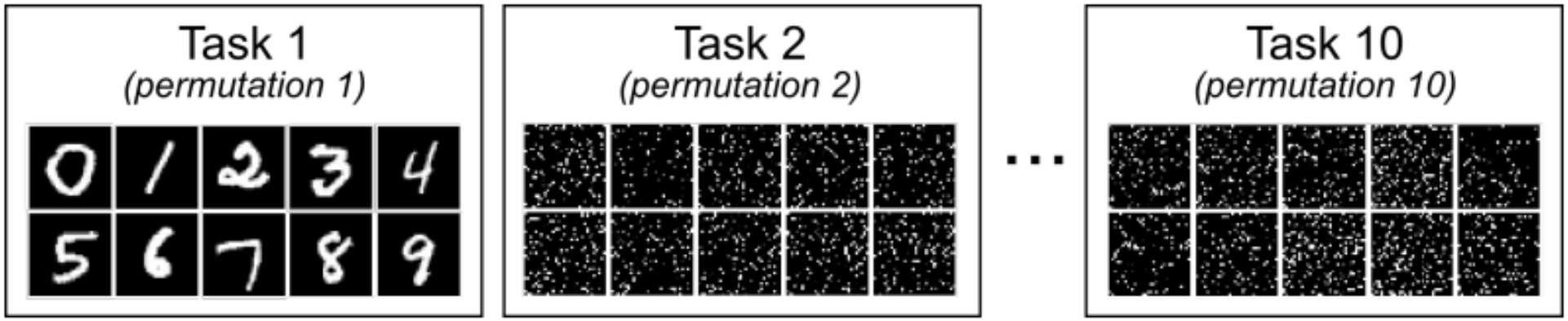
※ Diagonal만 사용함!

```
fisher_accum = {n: torch.zeros_like(p) for n, p in current_means.items()}

pbar = tqdm(self.dataloader, desc="Fisher Comp", leave=False)
for data, target in pbar:
    data, target = data.to(self.device), target.to(self.device)
    self.model.zero_grad()
    output = self.model(data)
    loss = F.cross_entropy(output, target)
    loss.backward()

    for n, p in self.model.named_parameters():
        if p.grad is not None:
            fisher_accum[n] += p.grad.data ** 2

num_samples = len(self.dataloader.dataset)
for n in fisher_accum:
    current_fisher[n] = fisher_accum[n] / num_samples
```

$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{A,i})^2$$

Experiments

04 Permuted MNIST

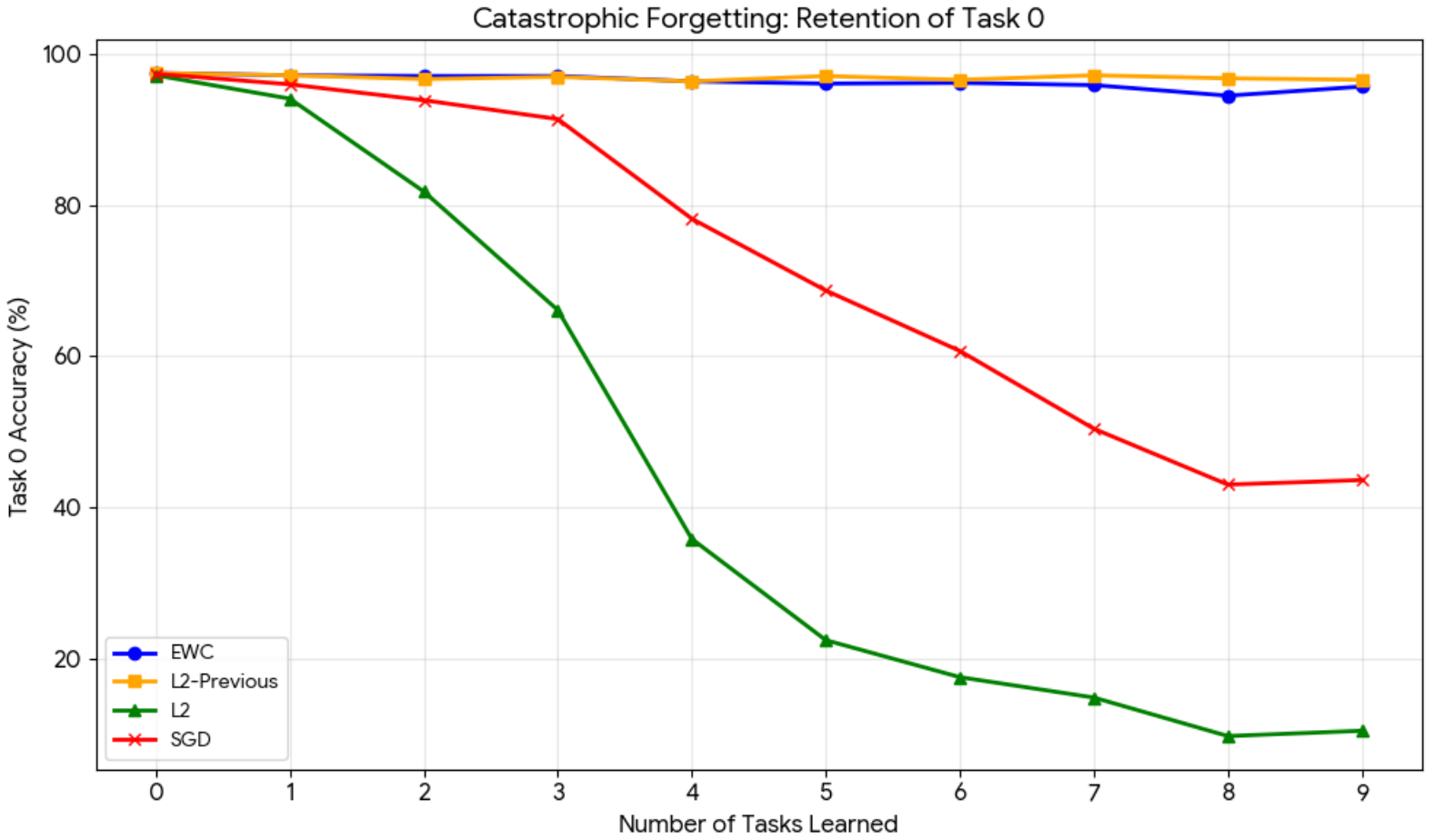


- Full-connected network
  - Input 784 → Hidden 400-400 → Output 10

- Baselines
  - SGD
  - L2 Regularization
  - L2 to Previous Weights ($F = I$)
  - EWC

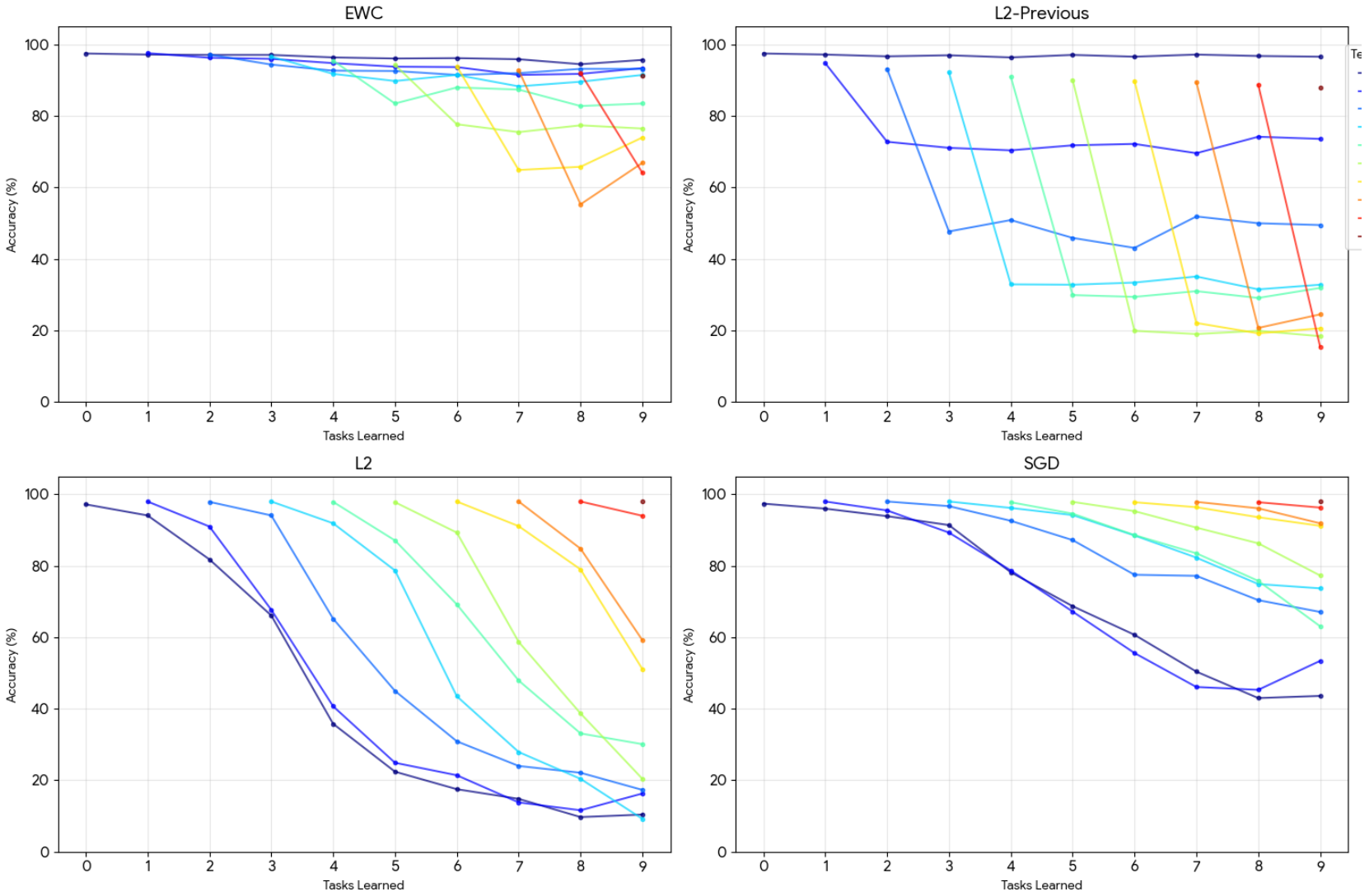
04 Evaluation of Catastrophic Forgetting

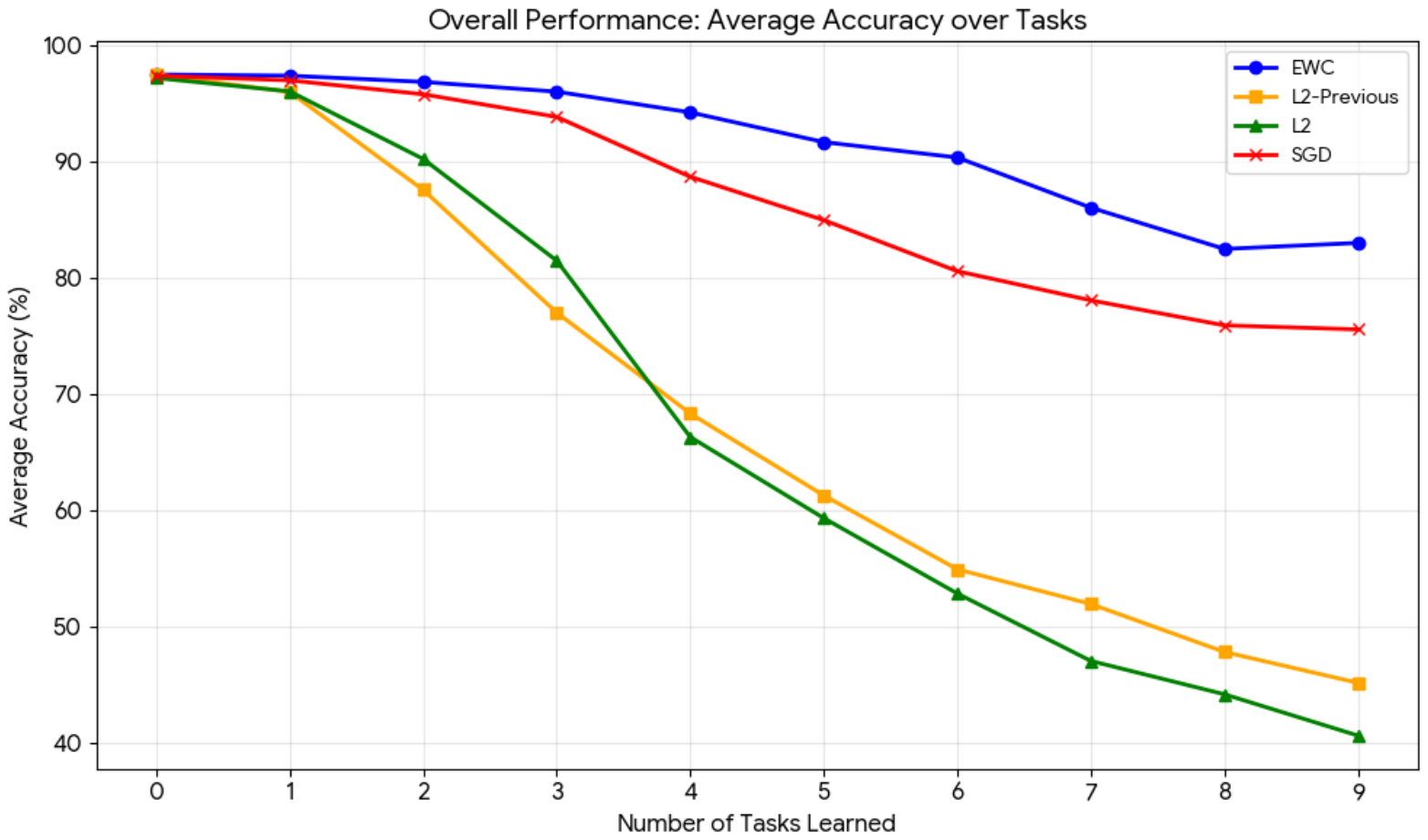
BrAIn Lab.
Brain and Artificial Intelligence Lab.



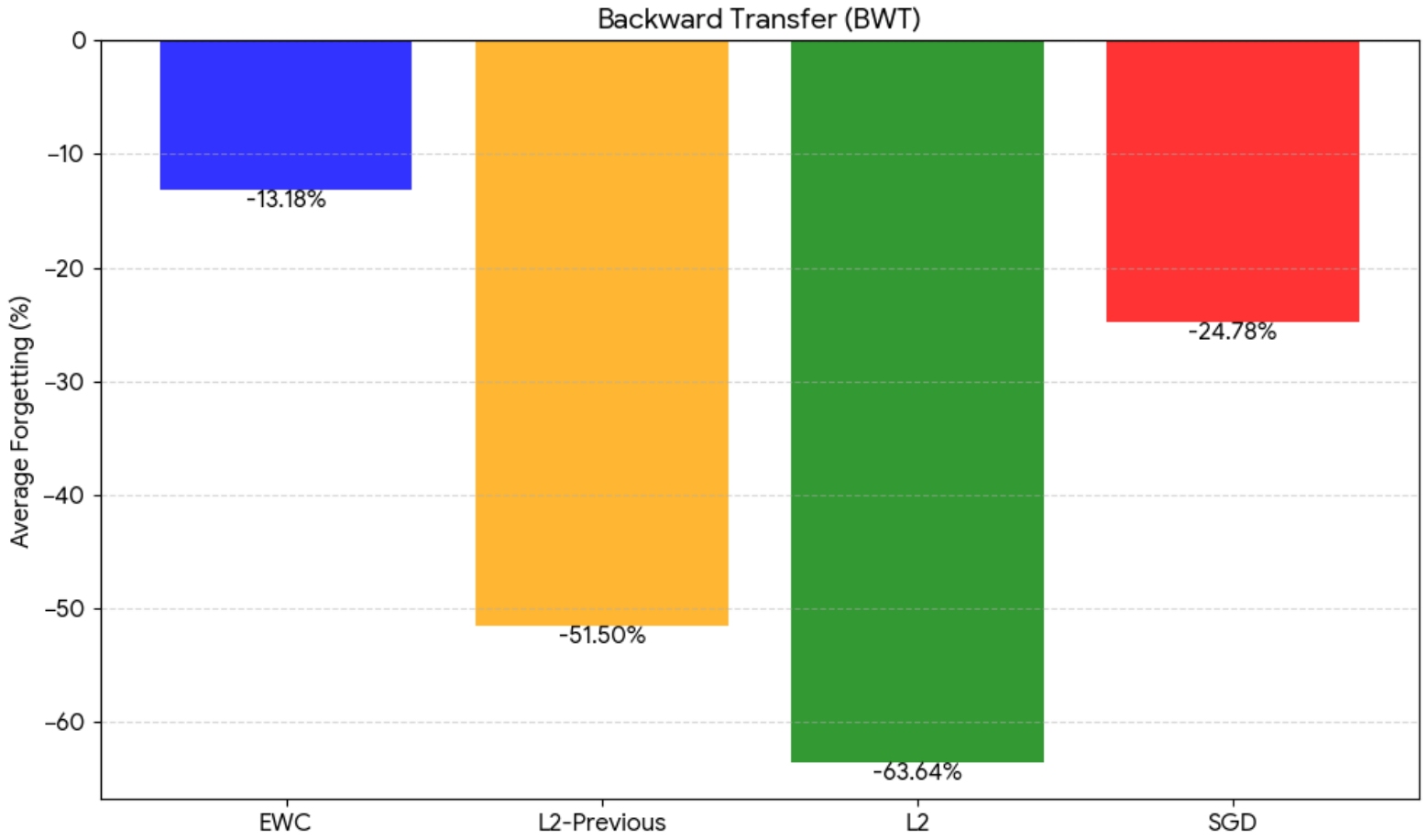
04 Overall Performance Comparision

BrAIn Lab.
Brain and Artificial Intelligence Lab.




Overall Performance: Average Accuracy over Tasks
Average Accuracy (%)
Number of Tasks Learned
EWC
L2-Previous
L2
SGD
100
90
80
70
60
50
40
0
1
2
3
4
5
6
7
8
9

04 Backward Transfer



Conclusion

05 Conclusion

BrAIn Lab.
Brain and Artificial Intelligence Lab.

✓ **Key Contributions**
- Bio-inspired Mechanism
- Replay-Free
- Selective Plasticity

✓ **Limitations & Future Work**
- Limitations:
  - Approximation: 파라미터 분포를 Diagonal Gaussian으로 단순하게 가정.
  - Rigidity: 태스크 수가 많아질수록 제약이 누적되어 학습이 뻣뻣해짐
- Future Directions:
  - Online EWC: 태스크 경계없이 실시간으로 중요도 갱신.
  - Generative Replay: 생성 모델과 결합하여 데이터 없이도 과거 경험 복원.

"Knowing What is important to remember is key to AGI."

Thank You

2026.01.05

BrainLAB Journal Club
인공지능응용학과 허세민