사전 학습된 Vision Transformer 를 통한 Base Session 제거 및 효율적 학습 방법론

이정호, 박경문 교수님
경희대학교 컴퓨터공학부
e-mail : owiequhf@khu.ac.kr

Eliminating Base Session in Few-shot Class Incremental Learning with Pre-trained Vision Transformers

Jeong-Ho Lee, Gyeong-Moon Park
School of Computing
Kyung Hee University

Abstract

본 연구에서는 FSCIL Scenario 에서 Base Session 학습에 대한 의존도를 줄이고, 제한된 Data 로도 효과적으로 새로운 클래스를 학습할 수 있는 FSCIL 방법론을 제안한다. Pre-trained Vision Transformer 를 기반으로 Meta-training 을 활용하여 Model 의 일반화 성능을 극대화하고, Neural Collapse Loss 를 통해 학습의 효율성을 향상시킨다. 또한, FC Layer 에 Low-Rank Adaptation 을 적용하여 Catastrophic Forgetting 문제를 완화하였다.

제안된 방법은 CIFAR-100 Dataset 에서 전체 Data 의 4.167%만 사용하면서도 Baseline 대비 약 5%의 성능 향상을 의미하며, FSCIL 환경에서 새로운 Class를 학습하는 데 필요한 Data 와 자원을 효과적으로 줄일 수 있음을 입증한다. 이러한 결과는 Data 가 제한된 환경에서의 실질적인 응용 가능성을 제시하며, FSCIL 분야에서 새로운 방향성을 제공한다.

1. Introduction

Few-Shot Class Incremental Learning (FSCIL) 은 소량의 Data 로 새로운 Class 를 학습하면서도 기존 Class 에 대한 지식을 유지하는 학습 방식이다. 그러나 대부분의 FSCIL Model 은 대규모 Dataset 을 요구하는 Base Session 초기 학습 단계에 크게 의존하며, 이는 현실적인 응용 환경과는 괴리가 있다.

이러한 접근은 Data Privacy 와 보안 문제를 포함한 실질적인 제약 조건과 충돌하며, FSCIL 의 본질적인 목표인 '소량의 Data 로도 효율적인 학습'과 충돌한다.

따라서 FSCIL 분야에서 Base Session 학습 과정을 간소화하거나 완전히 제거하고, 초기 Data 가 제한된 환경에서도 효과적으로 작동할 수 있는 Model 개발이 필수적이다. 본 연구는 Base Session 의존도를 줄이면서도 새로운 Class 에 대해 신속히 적응하고, 기존 지식을 유지할 수 있는 학습 방법을 제안하고자 한다.

2. Related Work

Few-Shot Class Incremental Learning (FSCIL)

Few-Shot Class Incremental Learning (FSCIL) 은 새로운 Class 를 점진적으로 학습해야 하는 상황에서 소량의 Data 를 사용하고도 효율적으로 Model 을 학습시키는 Scenario 를 다룬다. 일반적으로 FSCIL Model 은 대규모 Dataset 을 사용하여 다수의 Class 를 학습하는 Base Session 과, 이후 새로운 Class Data 가 추가되는 Incremental Session 으로 구성된다. Base Session 은 다수의 Class 를 학습하는 초기 학습 단계로 대규모 Dataset 에 의존하며, Incremental Session 에서는 새로운 Class 를 학습하는 과정에서 Data 가 극도로 제한된다.

FSCIL 의 주요 도전 과제는 Catastrophic Forgetting 과 Overfitting 문제이다. Catastrophic Forgetting 은 순차적으로 새로운 Class 들을 학습하는 과정에서 이전에 학습했던 Class 정보를 잃어버리는 현상이다. Overfitting 은 새로운 Class 를 학습하는 동안 제한된 수의 학습 Data 에 과도하게 집중하여 전체적인 성능 저하를 초래하는 문제이다. 이러한 도전 과제는 FSCIL Model 개발에서 중요한 연구 주제로 다루어진다.

Pre-trained Networks in FSCIL

FSCIL 에서 좋은 성능을 보이는 Model 대부분은 Pre-trained Network, 특히 Vision Transformer 와 같은 강력한 특성 추출기를 활용한다. Pre-trained Network 는 대규모 Dataset 에서 학습되어 일반화된 특징을 효과적으로 추출할 수 있는 기능을 제공하며, Base Session 과 Incremental Session 모두에서 적은 양의 Data 로 학습이 가능하도록 지원한다.

Pre-trained Network 의 주요 장점은 고정된 저수준 특징이 Incremental Session 에서도 유지되어 기존 Class 에 대한 지식 손실이 최소화된다는 점이다. 이러한 특성은 Base Session 없이도 Pre-trained Network 를 FSCIL Scenario 에 빠르게 적응시키고, 학습된 지식을 Incremental Session 으로 전이할 수 있도록 한다.

Meta-Learning for FSCIL

Meta-learning 은 다양한 작업에 Model 이 빠르게 적응할 수 있도록 학습하는 방법으로, "학습하는 방법을 학습하는" 접근 방식을 따른다. Episode 기반 학습은 소량의 Data 를 활용하는 Meta-learning 의 대표적인 방식으로, 각 Episode 는 소량의 Data 로 이루어진 Support Set 과 Query Set 으로 구성된다. Support Set 은 각 Class 의 Prototype 을 생성하는 데 사용되며, Query Set 은 Model 의 학습을 위한 손실을 계산하는 데 활용된다.

Meta-learning 은 소량의 Data 만 사용할 수 있는 환경에서 Model 의 일반화 능력을 극대화하는 데 목적이 있다. 특히 Pre-trained Network 를 적은 Data 로 Meta-training 하여 FSCIL Scenario 에 빠르게 적응할 수 있도록 한다.

Neural Collapse

Neural Collapse (NC)는 Deep Neural Network 의 학습이 진행되면서, 같은 Class 의 특징들이 단일 점으로 수렴하는 현상을 설명하는 이론이다. 이 현상은 Class 간 구별을 극대화하기 위해 Network 가 각 Class 의 특징을 조정하기 때문이다. NC 이론에 따르면 Class 간의 대표 점들이 이상적인 기하학적 구조 (Simplex Euclidean Tight Frame)를 이루게 되며, 이는 분류의 효율성을 극대화한다.

Low-Rank Adaptation

Low-Rank Adaptation (LoRA)는 저차원 행렬을 이용하여 가중치를 Update 함으로써 학습에 필요한 Parameter 수를 줄이고, 학습에 소요되는 시간과 자원을 절감하는 기법이다. FSCIL 에서는 LoRA 가 일부 Parameter 만 학습한다는 사실을 이용하여 기존 지식을 유지하고 Catastrophic Forgetting 을 방지하기 위한 의도로 사용되었다. 특히 Incremental Session 에서 LoRA 는 FC Layer 에 적용되어 사전 학습된 가중치를 고정하고 저차원 행렬을 학습함으로써 효율적으로 새로운 Class 를 학습하도록 돕는다.

3. Method

Modulation for FSCIL Challenges

FSCIL Scenario 에서 해결해야 할 주요 도전 과제는 Catastrophic Forgetting 과 Overfitting 문제이다. Catastrophic Forgetting 은 Incremental Session 에서 새로운 Class 학습 시 기존 Class 지식을 잃어버리는 현상으로, Model 의 성능 저하를 초래한다. Overfitting 은 소량의 Data 에 과도하게 집중하여 Model 의 일반화 성능이 저하되는 문제이다. Catastrophic Forgetting 과 Overfitting 문제를 해결함과 동시에 Data 가 제한적인 환경에서도 Pre-trained Network 가 빠르게 FSCIL Scenario 에 적응할 수 있는 방법을 모색해야 한다.

Pre-trained Knowledge Tuning

PriViLege Model 에서는 Pre-trained Vision Transformer Model 을 Domain 에

맞게 Fine-tuning 하기 위한 방법으로 하위 2 개의 Layer 만 학습하고 다른 Layer 는 Freezing 하여 Catastrophic Forgetting 을 방지했다. 많은 Layer 를 Freezing 했기 때문에 Incremental Session 에서의 성능이 부족할 수 있는데, B Prompt 와 VL Prompt 를 이용하여 이를 방지했다.

B Prompt 는 Layer 를 Fine-tuning 하며 Domain 의 일반적인 정보를 학습하게 되며 VL Prompt 는 이전 Session 에서 학습했던 정보를 다음 Session 에 전이하는 역할을 한다.

B Prompt 의 느린 적용 속도로 인해 새로운 학습법이 필요했는데 Prompt Modulation 이 그것이다. Head Specific Prompt 는 Head 별로 추출된 정보들의 관계를 학습하고, Generic Prompt 는 넓은 범위의 일반적인 특징을 학습하는데 중점을 둔다.

Base Session 의 학습 과정이 Fine-tuning 일 때는 대규모의 Data 로 저수준의 일반적인 Feature 들을 학습하기 위해 하위 2 개의 Layer 를 학습했지만, Episode 기반 학습의 경우 Few-shot 으로 주어지는 Data 를 학습하기 때문에 상위 2 개의 Layer 를 추가적으로 학습하여 성능을 향상시켰다.

Energy-Based Divergence Loss (ED Loss)

Vision Token 과 CLS Token 은 동일한 목표를 두고 학습한다. 때문에, 학습이 진행될수록 Vision Token 과 CLS Token 의 Output 이 비슷해지게 되어 Vision Token 의 개별 학습 능력이 저해된다. ED Loss 는 두 Token 의 출력이 점차 유사해지는 현상을 방지하고, Vision Token 이 독립적인 분류 능력을 유지하도록 유도한다.

각 Class c_j 에 대해, CLS Token 의 평균 출력 Vector 로 구성된 Prototype 분류기 ψ 를 생성한다.

$$proto_{c_j} = \frac{1}{N_{c_j}} \sum_{k=1}^{N_{c_j}} f_k^{cls}$$

$$\psi = \left[proto_{c_1}; proto_{c_2}; \dots; proto_{|C_0|} \right]$$

여기서 N_{c_j} 는 Class c_j 의 Sample 수, f_k^{cls}는 CLS Token 의 출력 Vector 이다.

Prototype 분류기를 사용해 Vision Token 과 CLS Token 에 대한 Logit y_i^{vis}, y_i^{cls} 를 계산한다.

ED Loss 는 아래의 수식으로 정의된다.

$$L_{ED} = \log\left(\frac{L_{CE}(y_i^{vis}, y_i) + L_{CE}(y_i^{cls}, y_i)}{L_{CE}\left(\delta(y_i^{vis}), \delta(y_i^{cls})\right) + 1} \right)$$

여기서 L_{CE} 는 Cross Entropy 손실, L_{KL} 은 Kullback-Leibler 발산 손실, $\delta(\cdot)$는 Softmax 함수이다.

ED Loss 는 두 Token 간의 확률 분포는 비슷하게 만들면서, 특징 공간 상에서의 위치는 유지하도록 하여 분류 성능은 향상시키면서 고유의 특징을 유지하도록 유도한다.

Semantic Knowledge Distillation Loss (SKD Loss)

SKD Loss 는 외부 정보를 활용하여 제한된 Sample Data 로 새로운 Class 를 더 잘 학습할 수 있도록 지원하는 방법이다. PriViLege Model 에서는 BERT 와 같은 사전 학습된 언어 Model 을 사용하여 Class 이름의 Text 특징을 추출하고 언어 Token 과 유사해지도록 학습하는 방식으로 SKD Loss 를 정의한다.

Class 이름 $word_{c_i}$ 를 "a photo of [class]" 형태로 변환하고, 사전 학습된 BERT Model f_ϕ 를 통해 언어 Embedding $w_{c_i} = f_\phi(word_{c_i})$ 을 생성한다.

VL-Prompt 의 언어 Token 에서 생성된 특징 f_i^{lang} 를 사용한다.

SKD Loss 는 아래와 같이 정의된다.

$$L_{SKD} = L_{KD}(f_i^{lang}, w_{c_i}) + \omega \cdot L_{CE}(y_i^{lang}, y_i)$$

여기서 L_{KD} 는 Knowledge Distillation 손실, L_{CE} 는 Cross Entropy 손실, ω 는 가중치 조정 계수이며 모든 실험군에서 0.2 를 사용했다.

SKD Loss 는 언어 Embedding 에서 추출한 정보를 시각적 특징과 통합하여, Model 이 새로운 Class 에 더 적응할 수 있도록 유도한다.

Neural Collapse Loss (NC Loss)

Neural Collapse (NC) 현상은 Deep Neural Netwrok 학습이 완료될 때 각 Class 의 특징 Vector 가 단일 점으로 수렴하고, Class 간 Prototype 이 Simplex Euclidean Tight Frame 을 형성하는 기하학적 구조로 정렬되는 현상을 나타낸다. NC Loss 는 이러한 구조를 FSCIL Scenario 에 적용하여 Class 간 구별성을 극대화하고 Incremental Session 에서도 성능 저하를 방지하도록 설계되었다.

NC 이론에 따르면, N 개의 Class 에 대해 $\boldsymbol{W}=[\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_N]$ 는 다음 수식을 만족해야 한다.

$$\boldsymbol{w}^T\boldsymbol{w} = \frac{N}{N-1}\boldsymbol{I}_N - \frac{1}{N-1}\mathbf{1}_N\mathbf{1}_N^T$$

여기서 $\boldsymbol{I}_N$은 $N \times N$ 단위 행렬이고, $\mathbf{1}_N$은 $N \times N$ 모든 요소가 1 인 행렬이다.

Prototype $\boldsymbol{p}_i$ 와 Query Vector $\boldsymbol{q}$ 간의 Cosine 유사도는 다음과 같이 정의된다.

$$\cos(\boldsymbol{q}, \boldsymbol{p}_i) = \frac{\boldsymbol{q} \cdot \boldsymbol{p}_i}{\|\boldsymbol{q}\|\|\boldsymbol{p}_i\|}$$

NC Loss 는 Cosine 유사도가 각 Class 간 이론적으로 최적의 각도를 만족하도록 설계되었다. 손실 함수는 다음과 같다.

$$L_{NC}(\boldsymbol{x}_i, y_i) = -\log\frac{e^{s\cdot\cos(\boldsymbol{q},\boldsymbol{p}_i)}}{e^{s\cdot\cos(\boldsymbol{q},\boldsymbol{p}_i)} + \sum_{j\neq y_i} e^{s\cdot\phi(\cos(\boldsymbol{q},\boldsymbol{p}_i))}}$$

여기서 $\phi(\cdot)$는 NC-Softabs 활성 함수로, 다음과 같이 정의된다.

$$\phi(\alpha) = \frac{1}{1+e^{-\beta\left(\alpha+\frac{1}{N-1}-0.5\right)}} + \frac{1}{1+e^{-\beta\left(-\alpha-\frac{1}{N-1}-0.5\right)}}$$

s 는 Scaling Parameter 이며, β 는 함수의 강성을 조절하는 Parameter 이며 모든 실험군에서 10 을 사용했다.

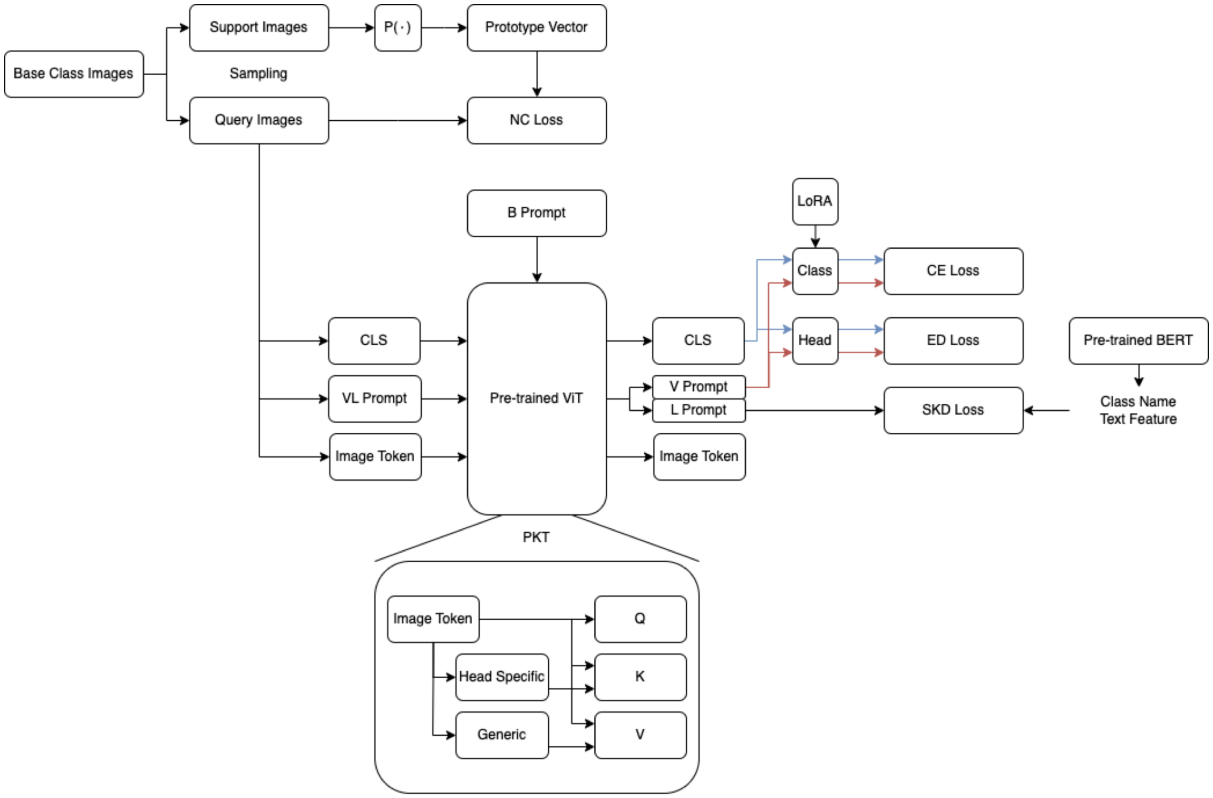

그림 1 Model 의 전체 구조이다. Base Session 에서는 기존 Model 의 학습 방식을 Meta-training 으로 전환하여 적은 Data 에도 빠르게 적응할 수 있도록 했다. L_NC를 적용하여 추후 Incremental Session 학습이 진행되어도 빠르게 ETF 구조를 회복할 수 있도록 유도했다. FC Layer 에는 LoRA 를 적용하여 Catastrophic Forgetting 이 최대한 방지되도록 유도했다.

이 손실 함수는 각 Class 의 특징 Vector 간 각도가 NC 이론에 의해 정의된 최적값 $-\frac{1}{N-1}$ 를 유지하도록 유도한다. 이를 통해 같은 Class 와의 유사도는 1, 다른 Class 와의 유사도는 $-\frac{1}{N-1}$ 이 되도록 유도하여 효율적으로 학습한다.

Base Session 학습의 최종 Loss 는 다음과 같다.

$$L_{Base} = L_{CE}(\hat{y}_i, y_i) + \alpha \cdot L_{ED} + \beta \cdot L_{SKD} + \gamma \cdot L_{NC}$$

α, β, γ는 ED Loss, SKD Loss, NC Loss 의 가중치 조정 계수이며, 모든 실험군에서 α, β는 0.1, γ는 1을 사용했다.

Incremental Session 학습의 최종 Loss 는 다음과 같이 정의된다.

$$L_{base} = L_{CE}(\hat{y}_i, y_i) + \beta \cdot L_{SKD}$$

Model Architecture (그림 1)

제안된 Model 은 PriViLege Model을 기반으로 Meta-Training, Neural Collapse Loss, LoRA를 조합하여 설계되었다. Meta-Training 과 Neural Collapse Loss 는 소량의 Data 로도 Model 의 빠른 적응을 가능하게 하며, LoRA는 Catastrophic Forgetting 문제를 최소화한다. 이러한 설계는 FSCIL Scenario 에서 기존 지식을 보존하면서도 새로운 Class 를 효과적으로

학습할 수 있도록 돕는다.

4. Experiments

Experimental Settings

실험은 CIFAR-100 Dataset 을 활용하여 진행되었으며, 성능 평가는 Base Session 에서의 정확도 (A_{Base}), 마지막 Incremental Session 에서의 정확도 (A_{Last}), 그리고 모든 Session 에서의 정확도 평균 (A_{Avg}) 을 기준으로 수행되었다. 비교를 위해 Base Session 학습을 포함한 PriViLege Model 과 Base Session 없이 학습된 PriViLege Model, 두 가지 Baseline 설정을 사용하였다. Base Session 학습을 포함한 PriViLege Model 의 경우 평가의 형평성을 위해 비교군에서 Base Session 학습에 필요한 이미지의 수와 동일한 수의 이미지로 학습하였다.

모든 실험에서 ImageNet-21K 로 사전 학습된 ViT-B/16 을 Backbone Network 로 사용하였다. Optimizer 는 SGD 를 사용했으며, 학습률은 1e-4, Momentum 은 0.9, Weight Decay 는 5e-4 로 설정하였다. 학습은 RTX 3090 GPU 에서 수행되었으며, Batch Size 는 Baseline 의 경우 128, Meta-training 의 경우 5-way 1-shot 4-query 설정으로 100 Episode 를 학습하였다. Base Session 은 5 Epoch, Incremental Session 은 3 Epoch 동안 학습되었으며, Incremental Session 은 총 8 개로 구성되었다.

Main Experimental Results

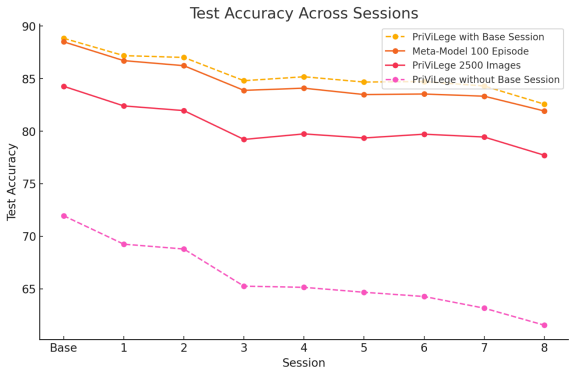

그림 2 성능 비교 그래프. 제안한 Model 이 모든 Session 에서 Baseline 보다 좋은 성능을 내고 있는 것을 확인할 수 있다.

본 연구에서 제안한 방법은 전체 Datset 의 4.167%만을 사용하여 학습되었음에도 불구하고 Baseline 대비 우수한 성능을 보였다. 각각의 정확도 차이는 4.233%, 4.21%, 4.209%로 나타났으며, 이는 같은 양의 Data 를 사용한 Baseline Model 대비 약 5.023%, 5.418%, 5.234%의 성능 향상을 보였다. 이러한 결과는 제한된 Data 환경에서도 제안된 방법이 효과적임을 입증한다.

Ablation Study

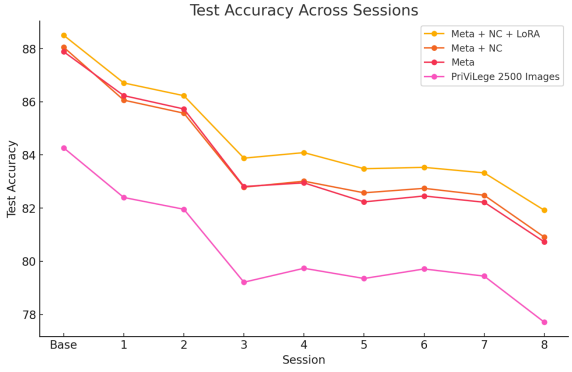

그림 3 적용 개념별 성능 비교 그래프. Meta-training 과 LoRA 를 적용했을 때는 큰 성능 변화가 있었지만, NC Loss 의 경우에는 성능 향상이 미미한 것을 확인할 수 있다.

Ablation Study 에서는 제안된 방법의 주요 구성 요소가 Model 성능에 미치는 영향을 분석하였다. Baseline Model 과 Meta-Training, NC Loss, LoRA 를 단계적으로 추가하며 성능 변화를 확인하였다. Meta-Training 을 적용했을 때 가장 큰 성능 향상을 보였으며, NC Loss 를 추가했을 때는 Prototype Vector 를 생성하기 위한 Data 부족으로 인해 성능 변화가 크지 않았다. 이는 Incremental Session에서 Data 를 더욱 효율적으로 사용하는 방법이 필요함을 시사한다.

Rank	A_{Base}	A_{Last}	A_{Avg}
Baseline	84.267	77.71	80.421
1	**88.5**	**81.92**	**84.63**
2	87.25	80.25	82.961
4	86.85	79.53	82.295
8	78.517	65.04	70.982
16	88.1	81.41	84.095

표 1 LoRA Rank 별 성능 비교 표. Rank 값이 1일 때 최고 성능이 보이고 Rank 가 커질수록 Catastrophic Forgetting 때문에 성능이 저하됐다고 판단된다.

LoRA 의 Rank 값을 변경하며 성능 변화를 분석하였다. Rank 값이 1일 때 모든 지표에서 최고의 성능을 보였으며, Rank 값이 커질수록 학습 Parameter 가 증가하여 Catastrophic Forgetting 이 심화되는 경향이 나타났다. 반면 Rank 값이 작아지면 Model 의 표현력이 감소하여 새로운 Class 를 학습하는 데 어려움을 겪었다. 실험 결과, Rank 값의 적절한 조정이 FSCIL Model 의 성능을 최적화하는 데 중요한 역할을 한다는 것을 확인할 수 있었다.

Rank 값이 16일 때 성능이 좋게 나오는 현상은 우연일 가능성이 가장 크고, Incremental Session 의 Data 분배가 고정적이지 않기 때문에 이 실험 설정에서 우연히 좋은 성능을 냈다고 판단된다.

Conclusion

본 연구에서는 Few-Shot Class Incremental Learning(FSCIL) 에서 Base Session 에 사용되는 Data 가 제한된 상황에서도 높은 성능을 달성할 수 있는 새로운 방법론을 제안하였다. 대규모 Dataset 을 요구하는 기존 방식과 달리, 본 연구는 Pre-trained Vision Transformer 와 Meta-Training을 결합하여 제한된 Data 로도 빠르게 학습할 수 있는 접근법을 설계하였다.

또한, Neural Collapse Loss (NC Loss) 를 활용하여 Incremental Session 에서 Class 간 이상적인 구조를 유지하였고, Low-Rank Adaptation 을 통해 Catastrophic Forgetting 문제를 완화하며 학습 효율성을 극대화하였다. 실험 결과, 제안된 방법은 CIFAR-100 Dataset 에서 Baseline Model 대비 모든 Session에서 우수한 성능을 보여, 데이터 사용량을 크게 줄이면서도 안정적인 학습이 가능함을 입증하였다.

추후 연구에서는 Base Session 을 완전히 제거하고 Incremental Session 에 적합한 Meta-Training 과 NC Loss 를 더욱 발전시켜, Base Session 없이도 효과적으로 새로운 Class 를 학습할 수 있는 FSCIL Model을 개발할 예정이다. 이러한 접근은 Data Privacy 와 보안 문제를 해결하고, 실제 응용 환경에서 더욱 실용적인 학습 방

식을 제공할 것이다.

References

[1] **Park, K.-H., Song, K., & Park, G.-M.** "Pre-trained Vision and Language Transformers Are Few-Shot Incremental Learners." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024

[2] **Ran, H., Li, W., Li, L., Tian, S., Ning, X., & Tiwari, P.** "Learning optimal inter-class margin adaptively for few-shot class-incremental learning via neural collapse-based meta-learning." *Information Processing and Management*, Vol. 61, No. 1, pp. 103664, 2024. https://doi.org/10.1016/j.ipm.2024.103664.