재연 기반 연속학습에서 차등 개인정보 보호 기법을 활용한 본보기 데이터 구성에 따른 성능 분석

강민준, 엄정민, 이재구*
국민대학교

*jaekoo@kookmin.ac.kr

Performance Analysis of Exemplar Data Condition with Differential Privacy in Replay-based Continual Learning

Minjun Kang, Jungmin Eom, Jaekoo Lee*
College of Computer Science, Kookmin University

요 약

심층 신경망 모델은 새로운 과업을 학습시키는 경우 기존 과업에서 심각한 성능 저하를 보인다. 이를 치명적 망각 현상(Catastrophic Forgetting)이라고 하며 해당 현상을 해결하는 것을 목표로 연속 학습 방법론이 지속적으로 연구되었다. 재연(Replay) 기반 연속 학습 방법론은 이전 과업의 데이터 일부를 메모리에 저장하여 현재 과업에 본보기 데이터(Exemplar)로써 활용하는 방식으로 효과적으로 연속 학습을 달성한다. 그런데, 이러한 연속 학습 방법론은 데이터의 개인정보 보호가 필요한 상황에서 본보기 데이터 구성에 문제가 발생한다. 본 논문에서는 본보기 데이터 구성에 차등 개인정보 보호(Differential Privacy) 기반 라플라스 메커니즘(Laplace Mechanism) 잡음을 도입하여, 이러한 개인정보 보호 방법이 연속 학습에 미치는 영향을 살펴보았다.

Ⅰ. 서 론

현재의 심층 신경망은 컴퓨터 비전, 자연어 처리 등 다양한 과업에서 뛰어난 성능을 보인다. 하지만, 현재 심층 신경망 모델은 현실에서 흔히 발생하는 과업이 변화하는 환경에 대응하는 것에 어려움을 겪는다. 또한, 현실 세계에 의료, 금융 등 민감한 개인정보 데이터를 다루는 분야에서 심층 신경망을 활용하기 위해서는 개인정보 보호를 고려해야 한다.

심층 신경망 모델이 새로운 과업을 학습함에 따라 이미 학습한 과업의 성능이 현저하게 떨어지는 문제를 치명적 망각(Catastrophic Forgetting)[1]이라고 하며, 이러한 현상을 해결하기 위해 연속 학습(Continual Learning) 연구가 진행되고 있다. 연속 학습은 이전 과업의 지식을 잊지 않으면서 새로운 과업을 학습하는 것을 목표로 한다. 연속 학습은 일반적으로 이전 과업의 데이터에는 접근하지 못하거나 일부만 이용할 수 있는 상황을 가정한다. 연속 학습의 다양한 전략 중에서 재연(Replay) 방법은 학습한 데이터의 일부를 본보기 데이터(Exemplar)로 메모리에 저장하여 새로운 과업 학습에 활용하는 방식으로 간단하면서 뛰어난 성능을 보인다. 대표적인 재연 기반 연속 학습 기법인 iCaRL[2]은 Herding[3] 기반 본보기 데이터 구성과 Nearest-Mean-of-Exemplars 분류 기법[2]을 통해 뛰어난 성능을 달성하였다.

한편, 개인정보 보호는 심층 신경망 학습 및 활용에 중요한 고려 사항 중 하나이다. 심층 신경망 학습에 민감한 학습 데이터의 노출을 방지하려는 다양한 연구가 있으며, 차등 개인정보 보호(Differential Privacy)[4]는 최우선으로 고려되는 유망한 기법이다. 차등 개인정보 보호[4]는 원본 데이터 혹은 임의의 알고리즘 $\mathcal{F}$의 출력값에 무작위성 잡음을 삽입함으로써 개별 예제에 대한 민감한 정보를 보호한다. 무작위 알고리즘 $\mathcal{F}$가 레코드가 하나만 다른 모든 이웃 데이터베이스 D와 D'에 대해 아래의 수식 (1)을 만족하면 ε-차분 개인정보 보호[4]가 성립한다.

$$P\ [\ \mathcal{F}(D)\ \in\ R] \le e^{\varepsilon}\ *\ P\ [\ \mathcal{F}(D')\ \in\ R] \quad (1)$$

여기서 ε 은 개인정보 보호 정도를 조정하는 변수이며, ε이 작을수록 개인정보 보호가 강해진다.

본 논문에서는 iCaRL[2] 기법과 차등 개인정보 보호[4] 기반 라플라스 메커니즘(Laplace Mechanism)[4] 기법을 응용하여 개인정보 보호를 고려한 재연 기반 연속 학습을 탐구하였다.

Ⅱ. 본 론

본 논문에서는 연속 학습의 다양한 학습 시나리오 중 현실적인 상황을 잘 반영하는 클래스 증분 시나리오(Class-incremental Scenario)를 고려한다. 모델은 추론 시에 과업 식별 정보를 제공받지 못하며, 지금까지 학습한 모든 클래스를 분류할 수 있어야 한다.

iCaRL[2]의 Herding[3] 기반 본보기 데이터 구성은 클래스별로 평균을 계산하고, 평균에서 가까운 데이터를 선택하여 클래스를 대표하는 본보기 데이터 집합을 구성하는 전략이다. Nearest-Mean-of-Exemplars 분류 기법[2]은 클래스별 본보기 데이터의 평균을 식별자로 설정하여 각 평균에 가장 가까운 클래스로 분류하는 전략이다.

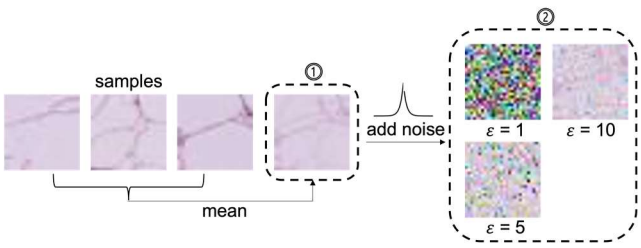

그림 1. 개인정보 보호 기법 단계별 적용

연속 학습의 학습 데이터가 민감한 정보를 담고 있다면, 본보기 데이터 구성에 개인정보 보호 문제가 발생한다. 따라서, [그림 1]에서 볼 수 있듯이 본보기 데이터 구성을 위한 두 단계의 개인정보 보호 기법을 설정하였다.

우선, 기초적인 개인정보 보호를 위해 과업 $\mathcal{T}_i$ 의 학습 이후 후보 본보기 데이터를 $\mathcal{N}$ 개 단위로 평균을 내어 본보기 데이터를 구성하였다. 이를 통해 해당 클래스의 통계적 특성을 유지하면서 원본 데이터를 저장하지 않도록 하였다. 다음으로, 더욱 엄밀한 개인정보 보호를 위해 ε-차분 개인정보 보호[4] 기반 라플라스 메커니즘[4] 잡음을 본보기 데이터 평균 과정에 삽입하였다. 또한, 본보기 데이터에 삽입한 잡음을 고려하여 Nearest-Mean-of-Exemplars 분류 기법[2]의 클래스별 본보기 데이터 평균을 과업마다 갱신하지 않고 고정하였다.

Ⅲ. 실 험

실험에 사용한 데이터 집합은 대표적 생물의학 사진 데이터 집합 중 하나인 MedMNIST[5]이며, 이 중 OrganAMNIST와 PathMNIST 데이터 집합을 사용하였다. OrganAMNIST는 신체 장기를 촬영한 3D CT 영상으로부터 얻은 2D 사진으로, 11개의 클래스로 구성된다. PathMNIST는 대장암 조직학 슬라이드 사진으로, 9개의 클래스로 구성된다. 전체 과업의 수는 $\mathcal{T}$ = 4이며, 본보기 데이터의 평균 단위는 $\mathcal{N}$ = 3이다. 과업별 학습 클래스 수는 각 데이터 집합 별로 [3, 3, 3, 2], [3, 2, 2, 2]로 구성하였다. 과업은 클래스 증분 시나리오에서 질병 분류 과업에 대한 성능을 평가하였으며, 모델의 정확도(Accuracy, %)와 망각(Forgetting, %)을 측정하였다. 심층 신경망 모델은 ResNet18[6]을 사용하였다. 본보기 데이터 집합의 크기는 현실적인 제약을 고려하여 전체 과업을 통틀어 200개로 유지하였다.

실험 결과는 [표 1]에 나타냈으며, [표 1]의 정확도와 망각은 마지막 과업 학습을 마친 후, 학습한 모든 과업에 대한 각 성능을 평균한 값이다. Baseline은 기초적인 개인정보 보호를 적용한 것으로 본보기 데이터를 $\mathcal{N}$ = 3개의 평균을 적용하여 구성하였다. 다음으로, 엄밀한 개인정보 보호를 위한 라플라스 메커니즘[4] 기반 잡음 삽입 기법은 개인정보 보호 정도와 성능과의 관계를 확인하기 위해 ε = 1, 5, 10으로 각각 실험하였다. Finetuning은 어떠한 연속 학습 기법도 적용하지 않고 매 과업을 학습한 것이다. [표 1]의 모든 실험 결과는 3개의 임의 초깃값을 사용하여 측정한 실험 결과를 평균한 수치이다.

표 1. 개인정보 보호 기법 적용 실험 결과

Dataset	OrganAMNIST		PathMNIST	
Metric	Acc*(↑)	Fg**(↓)	Acc(↑)	Fg(↓)
Baseline	63.08	40.11	54.03	49.15
ε = 1	57.55	46.11	47.19	60.09
ε = 5	62.93	40.09	50.42	55.77
ε = 10	64.62	37.88	51.03	54.28
Finetuning	27.66	76.16	23.90	70.76

***: Accuracy **: Forgetting**

실험 결과 [표 1]에서 보이듯이, 두 데이터 집합 모두 개인정보 보호 정도 조절 변수인 ε 값이 작을수록 모델의 정확도가 떨어지고 망각이 증가함을 보였다. 그러나, 개인정보 보호 강도가 가장 높은 ε = 1과 기초적인 개인정보 보호가 적용된 Baseline의 차이가 OrganAMNIST의 경우 5.53%p, PathMNIST의 경우 6.84%p로 큰 성능 하락이 발생하지 않았다. 반면, ε = 1과 연속 학습 기법이 적용되지 않은 Finetuning의 정확도 차이는 각각 29.89%p, 23.29%p로 큰 성능 차이를 보인다. 이는, 개인정보 보호 강도가 높아질수록 성능 열화가 발생하지만, 본보기 데이터 구성에 개인정보 보호 기법이 강하게 적용되어도 여전히 연속 학습이 효과적으로 달성된다는 것을 보여준다.

Ⅳ. 결 론

본 논문에서는 개인정보 보호가 필요한 재연 기반 연속 학습 환경에서 본보기 데이터 구성 문제를 해결하기 위해 개인정보 보호 기법을 적용하였다. 실험 결과, 개인정보 보호와 모델 성능 간에 적절한 균형을 요구하는 상충 관계를 확인하였으며, 높은 수준의 차등 개인정보 보호[4] 기반 라플라스 메커니즘[4] 잡음을 삽입하여도 연속 학습이 성공적으로 수행됨을 확인하였다.

ACKNOWLEDGMENT

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2022-0-00516,국가통계데이터에 적용 가능한 차등정보보호 개념을 도출하고 통계분석의 유용성을 보장해야 하는 문제 해결)

참 고 문 헌

[1] McCloskey, Michael, and Neal J. Cohen. "Catastrophic interference in connectionist networks: The sequential learning problem." *Psychology of learning and motivation*. Vol. 24. Academic Press, 1989. 109-165.

[2] Rebuffi, Sylvestre-Alvise, et al. "icarl: Incremental classifier and representation learning." *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2017.

[3] Welling, Max. "Herding dynamical weights to learn." *Proceedings of the 26th Annual International Conference on Machine Learning*. 2009.

[4] Dwork, Cynthia, et al. "Calibrating noise to sensitivity in private data analysis." *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Springer Berlin Heidelberg, 2006.

[5] Yang, Jiancheng, et al. "MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification." *Scientific Data* 10.1 (2023): 41

[6] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.