글로벌 모델 보호를 위한 서브모델 기반 연합학습

윤태환 [1], 최봉준 [2]
[1] 숭실대학교 컴퓨터학과 석박통합과정
[2] 숭실대학교 컴퓨터학과 교수

dbs1045@soongsil.ac.kr, davidchoi@soongsil.ac.kr

Subnet based Federated Learning for Protecting Global Model

Tae-Hwan Yoon[1], Bong-Jun Choi[2]
[1, 2]Dept. of Computer Science and Engineering, Soong-Sil University

요 약

연합학습은 분산된 환경에서 데이터의 공유 없이 모델을 학습시킬 수 있는 방법이다. 그 중에서도 Fed-Avg 는 분산된 클라이언트의 파라미터의 평균으로 모델을 수집하고 반영한다. 이 방법을 통해 연합학습의 모델의 연구가 발전되었으며, 모델성능을 향상시키기 위한 연구들이 꾸준히 진행되어왔다. 기존의 연합학습은 데이터를 공유하지 않고 모델 파라미터만을 서버로 전송하는 방식을 채택하여 데이터의 노출을 최소화하였다. 그러나 지역학습을 위해 서버가 클라이언트들에게 모델을 공유해야 하기 때문에 모델의 노출은 불가피할 수밖에 없다. 특정 분야에서는 데이터 노출 뿐만 아니라 모델의 노출을 보호하는 것 또한 중요한 분야도 있다. 본 논문에서는 이런 문제를 해결하기 위해 서브모델을 활용한 연합학습 보간 방법을 제시하였다. 이 방법은 지식증류 방법 기반에서의 새로운 모델 학습 방법을 제시한다. 실험에서 기존의 모델이 노출되지 않으면서 노이즈에도 강건하게 모델을 학습시킬 수 있음을 보여주었다.

1. 서 론

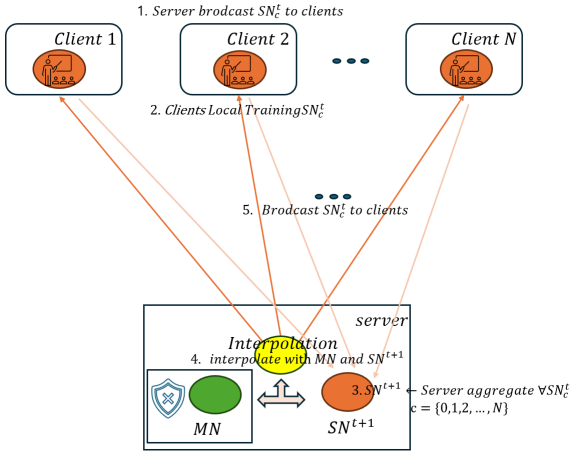
연합학습은 분산된 환경에서 데이터의 공유 없이 오직 모델 파라미터만을 공유 및 취합하여 학습하는 방법이다. 이런 과정에서 데이터의 보호는 이루어지지만 기존에 학습되었던 모델의 파라미터는 노출되어진다는 문제점이 있다 [1][2]. 때문에 악성 사용자에게 모델이 노출되어 악용될 가능성이 존재한다. 기존 논문 중에서 저자 Zhang et. al 은 Fine-tuning 과 지식증류 (Knowledge distillation) 기법을 통해 문제를 해결하였다. 그러나 이는 모델 일부를 클라이언트에게 노출한다는 점에서 한계점이 있다 [3]. 또한 같은 저자의 다른 논문에서는 동형암호를 통해 문제를 해결하고자 하였다, 하지만 동형암호는 수행 시간이 현저히 높아 실제 적용함에 있어서 한계점이 두드러진다 [4]. 본 논문에서는 서브모델 네트워킹(Sub-Net Networking)을 통한 모델 파라미터가 노출되는 문제의 해결책을 제시한다. 실험에서는 MNIST 와 WESAD(Wearable Stress Affect Detection) 데이터셋을 통해 기존의 연합학습 방식과 비교했을 때 비슷한 모델 성능을 보임과 동시에 모델 파라미터 보호가 가능함을 보여주었다. 또한 제시한 모델이 노이즈와 데이터포이즈닝 공격에 강건해짐을 보여주었다.

2. 본 론

기존의 Fed-Avg 의 경우 글로벌 모델을 클라이언트에게 공유하여, 연합학습을 수행하고 모델을 취합하여 평균으로 학습하는 구조이다. 이는 글로벌 모델의 노출과 데이터 포이즈닝 공격에 매우 취약하다. 본 논문에서 이에 대한 솔루션으로 제안하는 서브넷 네트워킹기법이란 기존의 연합학습에서 일반적인 연합학습을 수행하는 서브모델을 두고 기존의 학습이 되어 있는 노출이 되지 않았으면 하는 모델을 메인 모델로 설정하여 서브넷과 메인 모델들이 가지고 있는 파라미터들을 서버단에서 특정 보간법을 통해 학습하는 방법을 의미한다. 아래 (표 1) 에서c는 특정 클라이언트를 지칭하는 인덱스이고, SN은 서브모델을 MN은 글로벌 모델 즉 메인 모델을 의미한다. $\boldsymbol{SN^0}$ 은 초기 모델로 설정되며 본 논문에서 0 으로 설정하였다, MN은 오픈모델 혹은 미리 학습된모델을 사용할 수 있다.

표 1: 제안하는 아키텍처 심볼설명 테이블

$SN^0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ & \vdots & \end{bmatrix}$	Initializing Sub Model to zero
$c = \{1,2,\dots,N\}$	Specific client index
$MN = Train(SN)$ or Open Model	Global Model protected by Fed-SubNN



(그림 1): 제안된 Fed-Sub Net Networking 아키텍처

(그림 1)에서 서브모델 보간 연합학습 구조를 보여준다. 알고리즘의 순서는 다음과 같이 (1) 서버에서 초기모델 SN_0^t를 클라이언트들에게 공유한다. (2) 각 클라이언트에서 받은 모델을 가지고 학습데이터셋을 학습시킨다. (3) 서버는 클라이언트들이 학습을 마친 모델을 수집하여 평균 메소드를 통해 취합한다. (4) 글로벌 모델의 피처를 특정 보간법을 통해 서브모델을 학습시킨다. (5) 보간법을 통해 발생한 모델을 클라이언트들에게 다시 공유한다. (2)~(5) 과정을 라운드 수만큼 반복한다. 기존의 지식증류 방식에서 글로벌 모델에 좋은 파라미터를 전달해주는 방식을 기반으로, 본 논문에서는 좋은 파라미터를 가진 메인 모델을 선정하여 라운드마다 보간법을 통해 좋은 파라미터의 피처를 학습시키는 방식을 채택하였다[5][6]. 구체적인 서브모델 네트워킹 연합학습의 알고리즘은 다음과 같다.

표 2: Sub-Net Networking 연합학습 알고리즘.

	Fed-Sub Net Networking Algorithm
1	Init: t is a time step in round, c is a specific client index, MN is a main model protected by sub model interpolation method, SN is a sub-net to be trained and distributed. Ω is a weighted parameter for interpolation method. L is a learning rate. O is an objective function. CS is a cosine similarity function. DV is $SN^{t+1} - MN$ vectors.
2	Input: MN, SN_c^0, Ω, L, O
3	Output: SN^{t+1}
4	Clients do:
5	For epoch in Epochs:
6	$SN_c^{t+1} \leftarrow \boldsymbol{Local\ Training}(SN_c^t)$
7	return SN_c^{t+1}
8	Server do:
9	For t to {1,2,…,Round}:
10	$SN^{t+1} \leftarrow \boldsymbol{Aggregation}(\forall SN_c^{t+1})$
11	$SN^{t+1} \leftarrow \boldsymbol{Interpolation}(MN, SN^{t+1}, \Omega)$
12	Broadcast SN^{t+1}
13	$\boldsymbol{Local\ Training}$ (A client's SN_c^t):
14	For b to {1,2,…,Batch}:
15	$SN_c^{t+1} \leftarrow SN_c^t - L\nabla O(SN_c^t; b)$
16	return SN_c^{t+1}
17	$\boldsymbol{Aggregation}(\forall SN_c^{t+1})$
18	return $Average(\forall SN_c^{t+1})$ $(c = \{1,2,\dots,N\})$
19	$\boldsymbol{Interpolation}(MN,\ SN^{t+1},\ \Omega)$
20	$SN^{t+1} \leftarrow SN^{t+1} - \Omega \times DV \times (1 - \lvert CS(MN, SN^{t+1})\rvert)$
21	return SN^{t+1}

(표 2)에서 2~5 줄까지 클라이언트가 학습하여 서버로 학습한 모델을 보내주는 코드이다. 다음으로 6~10 줄까지 서버가 클라이언트들이 보낸 모델들을 평균으로 취합하고, 11 줄에서 다음 방정식(1)을 통한 보간법을 수행하고 결과를 클라이언트들에게 다시 보내주는 코드이다.

$$SN^{t+1} \leftarrow SN^{t+1} - \Omega \times DV \times (1 - |CS(MN, SN^{t+1})|) \quad (1)$$

방정식(1)에서 Ω는 메인 모델에 닮아가는 학습률이며, $|CS(MN, SN^{t+1})|$ 은 MN 과 SN^{t+1} 의 Cosine Similarity Vector 를 의미한다.

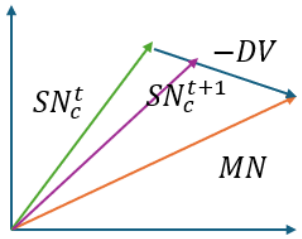
$$DV = SN^{t+1} - MN \quad (2)$$

방정식(2)에서 DV는 SN^{t+1}과 MN의 차를 의미한다. 모델의 각 레이어별로 닮은 확률을 나타내어 MN과 멀어지거나 가까워질수록 1 에 가까워진다, Cosine Similarity 는 데이터의 수치가 가까울수록 1 의 수치를 나타내며 멀어질수록 -1 의 값을 나타내기 때문이다. 이런 보간법을 통해 메인 모델을 닮아 가는 학습방식을 수행할 수 있으며 동시에 클라이언트가 가지고 있는 새로운 피처를 통한 학습이 가능하다.

표 3: 보간법 변수설명 테이블

t	Time step (round)
Ω	learning rate [0,1]

(표 3)에서 t 는 특정라운드의 스텝을 의미한다. Ω는 보간법에서 학습률 역할을 수행한다. 이런 보간법은 기존의 방식 보다 서버의 메모리 부하가 증가한다는 한계점이 존재한다. 때문에 LLM 같은 파라미터 크기가 매우 큰 모델의 경우 활용되기 어렵다. 그러나 모델 크기가 작은 경우에는 특정 모델이 노출되지 않아 보호됨과 동시에 클라이언트들 또한 학습에 참여가 가능하다.



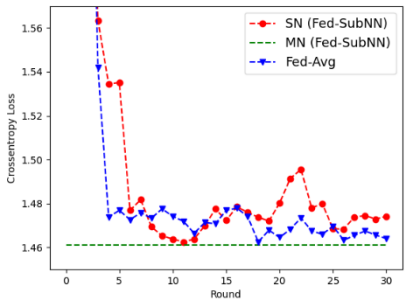
(그림 2): 서브모델 보간법

(그림 2) 에서는 보간법에 의한 서브 모델 학습을 시각적으로 보여준다. 이런 학습법을 통해 학습된 서브 모델(SN_c^{t+1})은 점점 보호하고자 하는 모델을 닮아 가게 학습된다. 그러나 같은 모델이 될 확률이 없지 않다. 때문에 방정식(2)에서 Cosine Similarity 를 활용하여 보간법에 가중치를 주어 닮을 확률을 최소화한다. 이를 통해 기존의 모델은 보호함과 동시에, 비슷한 성능의 모델을 클라이언트들이 학습을 할 수 있게 되었다. 다음(표 4)은 MNIST 데이터셋에서 ResNet 을 가지고 분류 과업을 수행했을 때의 결과이다.

표 4: MNIST classification

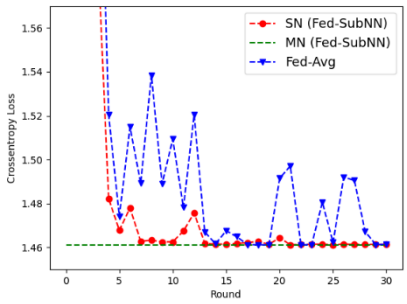
Model		Loss	Accuracy	F1-Score
Fed-Avg		1.4612	97.61%	97.48%
Fed-SubNN	*MN*	1.4612	98.86%	98.78%
	SN	**1.4612**	**95.66%**	**95.57**

(표 4)에서 라운드 수는 30 Round 로 고정이며, 클라이언트수는 5 명으로 설정되었다. Main-Net 은 실험에 설정한 메인모델이며 미리 중앙학습된 모델이다. Fed-Avg 는 기존의 연합학습 모델이며, Fed-SubNN 은 본논문에서 제시한 모델이다. Fed-SubNN 보다 Fed-Avg 이 Accuracy 와 F1-Score 에서 더 좋은 성능을 가짐을 보여준다 [1]. 그러나 Fed-Avg 는 기존에 글로벌 모델 노출과 데이터 포이즈닝 공격에 취약하다는 한계점이 있다. 반면에 Fed-SubNN 은 글로벌 모델 노출을 최소화하였으며, 성능은 Fed-Avg 보다 정확도가 1~2%정도 낮은 차이를 보였다. 다음은 본 논문에서 제안하는 Fed-SubNN 알고리즘을 통한 MNIST classification 과업에서 SN 과 MN및 Fed-Avg 모델의 글로벌 모델을 가지고 라운드별 성능을 보여주는 그래프이다. (그림 3)에서. Fed-SubNN 의 SN 성능과 Fed-Avg 의 학습 성능이 비슷함을 보여준다. 때문에 MNIST 데이터에서도 MN을 보호함과 동시에 클라이언트가 학습에 참여하여 Fed-Avg 와 비슷한 성능의 모델학습이 가능함을 보여준다.



(그림 3): Fed-SubNN 에서 MN과SN, Fed-Avg 에서 글로벌 모델의 성능

(그림 4)는 MNIST classification 에서 가우시안노이즈를 더해보았을 때 round 별 성능 비교를 보여준다.



(그림 4): Noise 가 더해진 MNIST 데이터셋으로 학습했을 때 성능 비교

(그림 4)에서 클라이언트가 가지고 있는 MNIST 데이터에 노이즈가 더해진 경우 Fed-Avg 에서는 모델 학습이 불안정하게 요동치는 그래프를 보여준다. 반면에 Fed-SubNN 에서는 좀더 안정적인 학습이 되는 그래프를 보여준다. 노이즈가 커질 때 메인모델과 거리가 멀어지고 그만큼 보완이 되기 때문에 안정적인 학습이 가능하다. 이를 통해 데이터포이즈닝 공격이 수행되어질 때와 데이터가 이질적인 상황에서도 안정적으로 학습이 가능하다.

표 5: MNIST classification (Noise Model)

Model		Loss	Accuracy	F1-Score
Fed-Avg		1.5236	97.93%	97.86%
Fed-SubNN	*MN*	1.4612	98.86%	98.78%
	SN	**1.4615**	**98.38%**	**98.40**

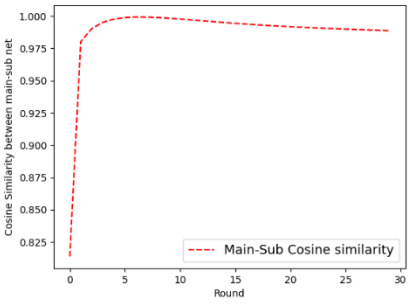
(표 5)는 노이즈 모델들의 성능을 보여준다. Fed-Avg

는 노이즈의 영향을 많이 받아 성능이 다소 떨어진 것을 보여준다. 반면에 Fed-SubNN 은 노이즈의 영향이 보완 학습되어 성능이 많이 개선됨을 보여주었다. 다음으로는 WESAD 데이터 셋에서 기존의 Fed-Avg 와 본 논문에서 제시하는 Fed-Sub Net Networking(Fed-SubNN)의 성능을 비교하였다.

표 6: Stress vs Non-Stress Detection

Model		Loss	Accuracy	F1-Score
Fed-Avg		0.5738	77.75%	50.61%
Fed-SubNN	*MN*	0.5690	78.02%	50.02%
	SN	**0.5546**	**76.66%**	**44.23%**

(표 6)를 통해 LSTM 모델을 가지고 시계열 데이터인 WESAD 데이터셋에서 웨어러블 기기를 통해 가슴에서 측정한 다양한 메타데이터(ACC: Three-Axis Acceleration, EDA: Electrodermal Activity, Temp: Body Temperature)를 인풋으로 실험자가 스트레스를 받았는지 받지 않았는지를 추론하는 과업을 기반으로 Fed-Avg 와 우리가 제시하는 Fed-SubNN 의 성능을 비교해서 보여준다 [7]. 총 학습된 라운드 수는 모델별로 30Round 고정이다. 기존의 Fed-Avg 와 비교했을 때 정확도에서 1%감소가 되었음을 보여주었으며 이는 차이가 크지 않다. 그러나 F1-Score 에서는 6%정도의 감소를 보였다. 반면에 Loss 는 가장 적음을 보여주었다. 시계열 모델에서 특정레이어의 파라미터를 수정 및 보완하면 시계열적인 모델 파라미터의 피처가 다소 손상됨을 의미한다. 이를 통해 특정 구간에서 오버피팅됨을 유추할 수 있다. (그림 5)은 *SN*이 *MN*에 닮아가는 과정을 라운드별로 보여주는 그래프이다.



(그림 5): Fed-SubNN 에서 라운드마다 Main-Sub 모델 간의 Cosine Similarity 변동 그래프

닮음을 측정하는 지표로 Cosine Similarity 를 사용하였다. (그림 5)에서 서브모델이 메인모델과 매우 유사하지만 성능이 다름을 통해 서로 다른 모델임을 보여준다. 따라서 글로벌 모델의 노출은 최소화되며 보호되고 서브모델은 학습성능이 글로벌 모델과 비슷한 수준으로 학습이 가능함을 보여준다.

3. 결 론

기존의 연합학습은 글로벌 모델을 공유하여 클라이언트들의 학습을 진행하는 방식이다. 때문에 글로벌 모델이 노출된다는 문제점을 가지고 있다. 본 논문에서는 Fed-Sub Net Networking 기법을 통해 글로벌 모델을 노출하지 않으면서. 연합학습을 수행하는 기법을 제안하였다. 또한 메인모델과 서브모델의 지식증류 기반의 새로운 보간법을 제시하여 서브모델이 학습함과 동시에 메인모델의 좋은 피처들을 학습할 수 있게 되었다. 본 논문에서는 실험을 통해 WESAD, MNIST 데이터셋에서 기존의 연합학습(Fed-Avg)과 비교했을 때 노이즈가 없는 상황에서 비슷한 성능을 보여주었다. 반면에 노이즈가 있는 상황에서 제안한 Fed-SubNN 의 메소드를 통해 노이즈에 강건한 모델을 학습할 수 있게 되어 기존의 연합학습보다 강건함을 보여주었다.

사사문구

본 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2022R1A2C4001270). 또한, 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 융합보안핵심인재양성사업의 연구 결과로 수행되었음 (IITP-2024-RS-2024-00426853).

참고문헌

[1] McMahan, et al. "Communication-efficient learning of deep networks from decentralized data.", PMLR, 2017.

[2] Li, Tian, et al. "Federated learning: Challenges, methods, and future directions." *IEEE signal processing magazine* Vol 37. No 3, 50-60p, 2020.

[3] Zhang, Lin, et al. "Fine-tuning global model via data-free knowledge distillation for non-iid federated learning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[4] Zhang, Li, et al. "Homomorphic encryption-based privacy-preserving federated learning in IoT-enabled healthcare system." *IEEE Transactions on Network Science and Engineering* Vol 10, No 5, 2864-2880p, 2022.

[5] Zhu, et al. "Data-free knowledge distillation for heterogeneous federated learning." , PMLR, 2021.

[6] Li, Daliang, and Junpu Wang. "Fedmd: Heterogenous federated learning via model distillation." arXiv preprint arXiv:1910.03581, 2019.

[7] Schmidt, Philip, et al. "Introducing wesad, a multimodal dataset for wearable stress and affect detection." *Proceedings of the 20th ACM international conference on multimodal interaction*. 2018.