

연합학습 기반 오토인코더 모델을 이용한 네트워크 이상탐지 시스템에 관한 연구

계효선, 권민혜
승실대학교

ksh8301000@soongsil.ac.kr, minhae@ssu.ac.kr

Federated Learning Based Stacked Autoencoder for Anomaly Detection on Networks

Hyoseon Kye, Minhae Kwon
Soongsil University

요 약

네트워크 공격을 사전에 탐지하는 동시에 고성능 컴퓨팅의 비용 및 개인정보 보호의 문제를 해결하여 네트워크 상에서 안정성을 확보하는 것의 중요성이 높아지고 있다. 본 논문에서는 연합학습을 기반으로 오토인코더를 사용한 이상탐지 시스템을 제안한다. 제안하는 시스템의 성능을 평가한 결과, 본 제안 시스템이 기존의 중앙학습 방식과 유사하게 우수한 성능을 보이는 것을 확인하였다.

I. 서 론

네트워크 보안 강화를 위해 보편적인 정상 트래픽의 패턴을 기준으로 악성 트래픽을 탐지하는 이상탐지에 대한 연구가 활발히 이루어지고 있다. 기존에는 중앙학습(centralized learning; CL) 방식으로 머신러닝 모델을 구축하고 있으나, 개인정보 보호와 서버 운영 비용에 대한 한계점을 극복하고자 연합학습이 제안되었다[1, 2]. 본 연구에서는 연합학습을 기반으로 stacked autoencoder를 활용한 네트워크 이상탐지 시스템을 제안한다.

II. 연합학습 및 오토인코더 기반 이상탐지 연구

II.1. 연합학습의 정의

연합학습은 중앙 서버의 글로벌 모델을 클라이언트의 로컬 학습을 통해 학습하는 방식이다. 클라이언트에서는 전송받은 글로벌 모델의 파라미터 정보를 클라이언트의 로컬 데이터로 학습하여 중앙 서버에 업데이트하고, 중앙 서버에서는 업데이트된 정보들을 취합하여 글로벌 모델을 갱신하는 방식으로 학습이 이루어진다. 이때, 서버와 클라이언트는 모델의 파라미터 정보를 communication round 마다 주고받으며 학습을 진행한다.

연합학습의 클라이언트의 수는 K 로, 클라이언트의 로컬 데이터는 $X_k (1 \leq k \leq K)$ 로 정의한다. 로컬 데이터 X_k 는 N 개의 데이터의 샘플 수와 M 개의 데이터의 차원(feature)의 수로 구성되어 $N \times M$ 차원 ($X_k \in \mathbb{R}^{N \times M}$)으로 이루어진다. 데이터 샘플 $x_{k,n}$ 의 인덱스 집합은 $P_k = [x_{k,1}, x_{k,2}, \dots, x_{k,N}]$ 로 정의하며 $N = |P_k|$ 이다. 중앙 서버는 매 communication round 마다 학습에 참여할 클라이언트를 uniform random으로 선택한다. 이때, 선택되는 클라이언트의 비율을 $C (0 < C \leq 1)$ 로 정의하고, 매 communication round 마다 참여하는 클라이언트의 개수는 $K \cdot C$ 로 표현한다. 중앙 서버는 연합학습을 시작하기 위해 초기화된 글로벌 모델을 생성하고, 학습에 참여할 클라이언트를 선택하여 글로벌 모델의 파라미터를 전송한다. 글로벌 모델의 파라미터는 d 차원을 가지는 $w (w \in \mathbb{R}^d)$ 로 정의한다. 선택된 클라이언트들은 각자 보유하고 있는 로컬 데이터 X_k 를 이용하여 글로벌 모델의 파라미

터 w 를 기반으로 학습을 수행한다. 모델의 w 는 입력값인 데이터 샘플 $x_{k,n}$ 과 예측값의 차이인 loss를 기반으로 역전파를 통해 업데이트 된다. 로컬 모델의 loss function은 $F_k(w)$ 로 정의되며, $x_{k,n}$ 의 loss function은 $f_{k,n}(w) = \ell(x_{k,n}; w)$ 로 정의한다. 로컬 학습에서 모델을 최적화하기 위해 $F_k(w)$ 를 최소화할 수 있는 w 를 찾는 것을 목표로 하며 이는 (1)과 같이 표현할 수 있다.

$$\min_{w \in \mathbb{R}^d} F_k(w) \quad \text{where} \quad F_k(w) = \frac{1}{N} \sum_{n \in P_k} f_{k,n}(w) \quad (1)$$

t 번째 communication round에서 클라이언트로부터 업데이트된 모델 파라미터 w_t^k 를 중앙 서버로 전송한다. 중앙 서버는 업데이트된 파라미터 w_t^k 를 (2)와 같이 평균값으로 취합하고 글로벌 모델 파라미터 w_{t+1} 을 업데이트하여 글로벌 모델을 갱신한다.

$$w_{t+1} = \frac{1}{K} \sum_{k=1}^K w_t^k \quad (2)$$

연합학습은 글로벌 모델이 수렴할 때까지 이 과정을 반복하면서 학습을 진행한다.

II.2. Stacked Autoencoder기반 이상탐지

본 논문에서는 연합학습 기반의 중앙 서버 및 클라이언트 모델로 stacked autoencoder를 사용하여 이상탐지 시스템을 구성하였다. Stacked autoencoder는 여러 개의 은닉층(hidden layer)으로 구성된 심층 신경망(deep neural network)형태의 오토인코더이다[3]. 오토인코더는 비지도 학습 방식으로 데이터를 학습하고 재구성하기 위한 모델로, 인코더(encoder)와 디코더(decoder)로 구성된다. 인코더와 디코더의 층은 $L (1 \leq i \leq L)$ 로 동일하며 인코더는 $g = g_L \circ \dots \circ g_1$, 디코더는 $f = f_1 \circ \dots \circ f_L$ 로 정의한다. 오토인코더 A 를 사용한 클라이언트의 데이터 샘플 $x_{k,n}$ 에 대한 출력값은 $A(x_{k,n}) = f \circ g(x_{k,n})$ 으로 정의할 수 있다. 인코더는 데이터 차원의 수를 입력 노드의 수로 설정하여 데이터를 입력층(input layer) g_1 으로 입력받아 학습을 진행한다. 인코더 g 는 입력 데이터의 차원을 축소시켜 의미 있는 정보만을 가지고 있는 잠재 공간(latent vector)을 형성한다. 디코더는 잠재 공간을 이용

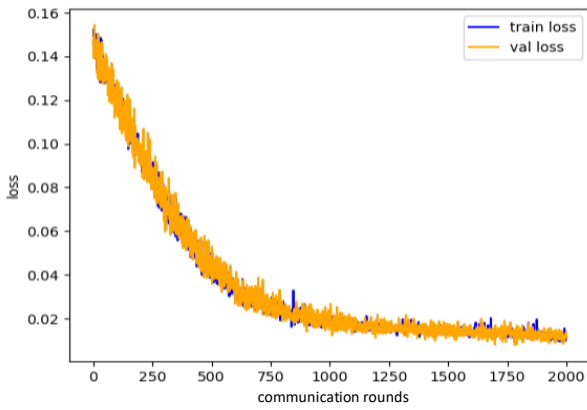


그림 1. Communication rounds 에 따른 loss 변화

하여 축소된 데이터의 차원을 복원시킨 복원 데이터(reconstruction data) $\hat{x}_{k,n} = A(x_{k,n})$ 를 생성한다. 학습은 원형 데이터와 복원 데이터의 차이를 최소화하는 방향으로 이루어진다. 이를 위해 평균 제곱 오차(Mean Squared Error; MSE)를 loss function으로 사용하여 역전파를 진행하며, 이를 통해 수식(1)은 (3)과 같이 표현된다.

$$F_k(w) = \frac{1}{N} \sum_{n \in P_k} \ell(x_{k,n}; w) \\ = \frac{1}{N} \sum_{n=1}^N (\hat{x}_{k,n} - x_{k,n})^2 \quad (3)$$

오토인코더 기반의 이상탐지를 위해 정상 데이터를 오토인코더로 학습시킨다. 복원 데이터와 원형 데이터의 복원 오차(reconstruction error)는 ϵ 로 정의한다. 정상 데이터의 복원 오차를 기반으로 설정한 특정 임계값(threshold)을 h 로 표현한다. 임계값 h 를 설정하기 위한 복원 오차는 (4)와 같이 평균 절대 오차(Mean Absolute Error; MAE)를 이용하여 계산한다.

$$\epsilon = \frac{1}{N} \sum_{n=1}^N |\hat{x}_{k,n} - x_{k,n}| \quad (4)$$

이상탐지를 수행하는 과정에서 입력 데이터의 복원 오차가 $\epsilon \geq h$ 이면 이를 비정상으로 간주하여 이상을 탐지한다.

III. 제안하는 방식의 실험 설정 및 성능평가

III.1. 성능평가 실험 설정

본 논문에서 제안하는 네트워크 이상탐지 시스템은 연합학습을 기반으로 stacked autoencoder 모델을 적용하였다. 실험에서 중앙 서버 및 모든 클라이언트들이 동일한 stacked autoencoder 구조를 가지는 것을 전제로 한다. 연합학습을 위해 500개의 고정된 데이터 샘플을 가진 100개의 클라이언트를 생성하였다. 실험의 학습 및 성능 검증을 위해 NSL-KDD 데이터셋을 사용하였다[4]. NSL-KDD train 데이터셋을 6:4의 비율로 train set과 validation set으로 분리하여 제안하는 시스템의 학습과 검증을 진행하였다. 모델의 평가를 위해 NSL-KDD test 데이터셋을 사용하였다. 제안하는 시스템의 정량적인 평가를 위해 정상 및 비정상 데이터를 탐지한 비율을 토대로 수치를 나타내는 성능지표인 F1-score와 MCC를 사용하였고, 기존 중앙학습 방식과 성능을 비교하였다.

III.2 성능평가 결과

본 제안 시스템은 communication rounds에 따른 loss curve를 나타내는 그림1을 통해 overfitting 없이 loss가 수렴하면서 모델이 최적화되었음을 확인하였다. Loss가 수렴하기 시작하는 500번째 communication round 이후부터 AUC 또한 안정적으로 1에 가깝게 수렴하는

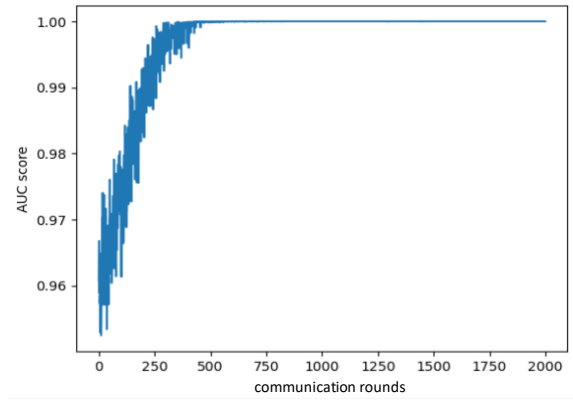


그림 2. Communication rounds 에 따른 AUC 변화

표 1. 연합학습(FL), 중앙학습(CL) NSL-KDD 성능평가

	F1-score	MCC
FL	0.9917	0.9747
CL	0.9999	0.9997

것을 그림 2를 통해 확인하였다. 표 1은 연합학습 FL과 중앙학습 CL의 성능을 나타낸 결과이다. CL은 F1-score와 MCC 모두 1에 가까운 높은 성능을 보였다. 이에 반해 FL은 100개의 클라이언트가 분산적으로 학습을 진행하였음에도 CL과 유사하게 높은 성능을 보이는 것을 확인하였다.

IV. 결론

본 논문은 연합학습을 기반으로 stacked autoencoder 모델을 적용한 네트워크 이상탐지 시스템을 제안한다. 제안한 시스템은 기존의 중앙학습 방식과 유사하게 우수한 성능을 보였다. 이를 통해 본 제안 시스템은 중앙학습 방식의 한계점인 고성능 컴퓨팅 파워의 비용 및 개인정보 보호 문제를 해결하면서 성공적으로 최적화된 글로벌 모델을 구축하여 이상탐지를 진행할 수 있음을 확인하였다. 본 연구 결과를 바탕으로 클라이언트에서 이상탐지 시스템을 직접 적용하여 서버 운영 비용 및 프라이버시 면에서 더 안정적인 네트워크 서비스를 제공할 수 있음을 확인하였다.

ACKNOWLEDGMENT

이 논문은 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT 연구센터지원사업(IITP-2021-2020-0-01602)과 방송통신산업기술개발사업(IITP-2021-0-00739)의 지원을 받아 수행된 연구임.

참 고 문 헌

- [1] 계효선, 권민혜, “모델 파라미터의 평균값을 이용한 연합지도학습 기반 네트워크 이상탐지 시스템,” 통신정보 합동학술대회, Apr. 2021.
- [2] B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. y. Arcas, “Communication-Efficient Learning of Deep Networks from Decentralized Data,” Artificial intelligence and statistics. PMLR, Apr. 2017.
- [3] A. M. Vartouni, S. S. Kashi, M. Teshnehlab, “An Anomaly Detection Method to Detect Web Attacks Using Stacked Auto-Encoder,” Iranian Joint Congress on Fuzzy and Intelligent Systems, pp. 131-134, Feb. 2018.
- [4] <https://www.unb.ca/cic/datasets/nsll.html>