부분적 모델 파라미터 손실에 따른 연합학습 성능 저하를 완화하기 위한 근사 데이터 전송 전략

권정민, 박형곤*

이화여자대학교

jungmin.kwon@ewha.ac.kr, *hyunggon.park@ewha.ac.kr

Approximate Data Transmission Strategy to Mitigate Performance Degradation of Federated Learning with Partial Model Parameter Loss

Jungmin Kwon, Hyunggon Park*

Ewha Womans University

요 약

본 논문에서는 UDP를 사용하는 FL 시스템에서 모델 파라미터의 부분적 손실로 인하여 발생하는 성능 저하 문제를 해결하기 위해, 저 랭크 근사 방법과 체계적 네트워크 코딩 방법을 결합한 안정적인 근사적 데이터 전송 전략을 제안한다. 제안한 알고리즘의 성능을 이론적으로 분석하기 위해 최적 모델 근사 실패 확률을 정리하였다. 이를 통해 제안한 알고리즘이 기존의 부분 손실을 겪는 모델 파라미터를 이용한 모델 학습에 비해 낮은 근사 실패 확률을 보이는 것을 확인하였다. 또한, 다양한 실험 결과를 통해 제안한 알고리즘이 패킷 손실로 인한 성능 저하를 크게 완화함을 확인하였다.

Ⅰ. 서 론

인공지능 기술의 상용화가 진행됨에 따라, 각 개인의 데이터를 안전하게 보호하며 효과적으로 학습을 실현할 수 있는 FL(Federated Learning) 기술이 점점 주목을 받고 있다. 그러나 FL 기술이 현실에서 구현되기까지는 불안정한 네트워크 환경에서 비롯되는 성능의 일관성 문제와 높은 통신비용과 같은 여러 도전 과제에 직면해 있다. 이러한 문제를 해결하기 위해, UDP(User Datagram Protocol)를 기반으로 모델 파라미터를 전송하는 새로운 FL 연구가 진행되었다 [1]. 그러나 UDP를 활용한 FL 시스템은 패킷 손실되이 발생하였을 경우 모델 파라미터의 부분적 손실로 인해 심각한 성능 저하가 초래될 수 있다. 따라서 본 논문에서는 UDP 기반의 FL 시스템에서 불안정한 네트워크 환경에도 불구하고 모델의 정확도와 성능을 안정적으로 유지하기 위하여, 저 랭크 근사와 체계적 네트워크 코딩(Systematic Network Coding, SysNC)을 결합한 근사적 데이터 전송 방법을 제안한다. 또한, 이론적 분석과 실험을 통해 제안한 방법의 우수 성을 확인한다.

Ⅱ. FL을 위한 안정적인 근사적 데이터 전송 전략

본 논문에서는 하나의 글로벌 서버와 여러 로컬 기기로 구성된 중앙 집권형 FL 구조를 고려한다. FL 과정은 1) 로컬 모델 학습, 2) 로컬 모델 전송, 3) 모델 통합, 4) 글로벌 모델 전송, 5) 로컬 모델 업데이트의 다섯 단계로 크게 구분되며 이 일련의 과정을 반복하게 된다. 이때 제안하는 근사적 데이터 전송전략은 모델 전송 단계에 적용된다. 전체적인 알고리즘 구성과 흐름은 그림 1과 같다.

Ⅱ.1. 모델 파라미터의 저 랭크 근사화

전달하는 모델 파라미터가 $\boldsymbol{W}$이고 $\boldsymbol{W}$의 랭크가 r 일 때, SVD(Singular Value Decomposition)를 이용하여 $\boldsymbol{W}=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T$ 로 모델 파라미터를 분해한다. 여기서 $\boldsymbol{U}$와 $\boldsymbol{V}$는 $\boldsymbol{W}$를 구성하는 고유벡터 행렬로 r개의 열벡터로 구성되어있으며, $\boldsymbol{\Lambda}$는 특이 값 행렬로 대각 행렬이며 각 대각 원소는 내림차순으로 정리되어있다. 이후, 상위 k개의 특이 값과 그에 대응 되는 고유벡터의 열벡터를 선별한다.

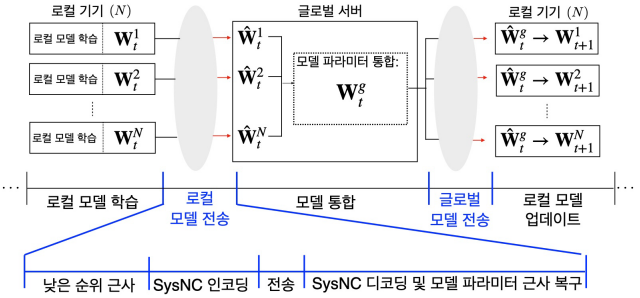

그림 1. FL의 구조와 제안하는 데이터 전송 방법 프로세스 모식도

Ⅱ.2. SysNC를 이용한 데이터 인코딩 및 전송

상위 k개의 특이 값을 기반으로 선별된 행렬들을 안정적으로 전달하기 위하여 SysNC 기법을 이용해 인코딩 된 데이터 $\boldsymbol{Y}=\boldsymbol{C}\times\boldsymbol{Z}$를 생성한다. 여기서 $\boldsymbol{Z}$는 선별된 행렬을 나타내고, $\times$는 행렬 곱을,그리고 $\boldsymbol{C}=[\boldsymbol{c}_1,\ \cdots,\boldsymbol{c}_M]$는 코딩계수 행렬을 나타내며 열벡터 $\boldsymbol{c}_i$로 구성되어있다고 하자. 이때, $[\boldsymbol{c}_1,\ \cdots,\boldsymbol{c}_k]=\boldsymbol{I}_k$를 만족하며 여기서 $\boldsymbol{I}_k$행렬은 k개의 열벡터로 구성된 항등행렬을 나타낸다. 이와 달리, $[\boldsymbol{c}_{k+1},\ \cdots,\boldsymbol{c}_M]$ 행렬은 모든 원소가 특정 수의 범위 안에서 임의로 결정된 값을 갖는다. 이후, 새롭게 인코딩 된 행렬 $\boldsymbol{Y}$와 코딩계수 $\boldsymbol{C}$의 각 행벡터를 패킷화하여 전달하고, 패킷 손실 확률 p를 갖는 불안전한 네트워크를 통해 전달된다.

Ⅱ.3. 데이터 디코딩 및 모델 파라미터 근사화

수신 단에서는 인코딩 된 데이터 $\boldsymbol{Y}$의 행벡터가 일부분 손실된 $\hat{\boldsymbol{Y}}$와 $\hat{\boldsymbol{C}}$를 수신하여 행렬 $\hat{\boldsymbol{Z}}=\hat{\boldsymbol{C}}^{-1}\times\hat{\boldsymbol{Y}}$를 복구한다. 이때 패킷 손실 확률 p가 $p<k/M$ 일 경우에는 $\hat{\boldsymbol{Z}}=\boldsymbol{Z}$를 만족하며, 반대의 경우에는 $\hat{\boldsymbol{Z}}=\tilde{\boldsymbol{Z}}$로 원 $\boldsymbol{Z}$ 행렬의 부분집합에 해당하는 행렬 $\tilde{\boldsymbol{Z}}$와 같아진다. 이렇게 수신된 행렬 $\hat{\boldsymbol{U}},\hat{\boldsymbol{\Lambda}},\hat{\boldsymbol{V}}$을 기반으로 최종적으로 수신 단에서 근사된 모델 파라미터 $\hat{\boldsymbol{W}}=\hat{\boldsymbol{U}}\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{V}}^T$ 얻고 이를 모델 통합 또는 업데이트 단계에 사용한다.

Ⅲ. 최적 모델 근사 실패 확률 분석

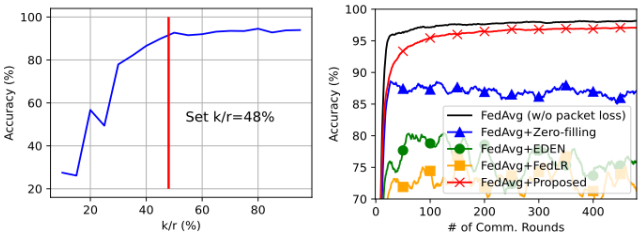

그림 2. MNIST 데이터셋에 대한 정확도 결과. (a)는 선별하는 상위 특이 값의 개수를 변화하였을 때의 정확도, (b)는 통합된 모델 파라미터를 전달받았을 때 시간에 따른 정확도를 보여준다.

본 연구에서 제안한 근사적 모델 파라미터 전송 방법을 이용하여 로컬 모델 학습을 진행할 때 최적 모델을 근사하는데 실패할 확률을 분석한다. 최적 모델 파라미터는 $\boldsymbol{W}^{*}=[\boldsymbol{w}_1^{*},\cdots,\boldsymbol{w}^{*}_r]$로 r개의 열벡터 $\boldsymbol{w}_i^{*}$로 구성되어있고 각 열벡터는 q개의 원소로 구성되어있다고 하자. 패킷 손실로 인해 부분적으로 손실된 모델 파라미터를 $\widetilde{\boldsymbol{W}}=[\widetilde{\boldsymbol{w}}_1,\cdots,\widetilde{\boldsymbol{w}}_r]$로, 제안한 알고리즘의 파라미터는 $\widehat{\boldsymbol{W}}=[\widehat{\boldsymbol{w}}_1,\cdots,\widehat{\boldsymbol{w}}_r]$로 구성하고, $\boldsymbol{x}$는 모델의 입력 데이터라고 할 때, 근사 실패 확률은 다음과 같이 정리된다.

Theorem 1. 상수 $\alpha,\beta>0$이고 L는 립시츠 상수일 때, 최적 모델의 결과 값과 제안한 알고리즘을 통해 계산된 모델의 결과 값의 오차가 ϵ보다 클 확률은 다음과 같은 상한을 갖으며 이는 $\widetilde{\boldsymbol{W}}$를 사용하였을 때 근사 실패 확률보다 낮은 값을 보인다.

$$\boldsymbol{P}\left[\left|\sum_{i=1}^{r}(w_i^{*}-\widehat{w}_i)x\right|>\frac{\epsilon}{L}\right]\leq\frac{\alpha L}{\epsilon}e^{(-\beta q)}(1-p)^{rq} \tag{2}$$

proof. $\alpha=\max_{i\leq q}\alpha_i$이고 $\beta=\max_{i\leq q}\beta_i$를 만족하며, $|\bullet|$ 표기는 절대값을 나타낸다고 하자. 일반적인 모델 파라미터 $\boldsymbol{W}$에 대하여 각 행이 이항 분포에 따라 p의 확률로 무작위 손실되어 불완전한 $\widetilde{\boldsymbol{W}}$이 되었을 경우 손실된 원소는 0으로 채우는 경우를 고려하면, 이에 대한 최적 모델 근사 실패 확률은 다음과 같이 정리된다 [2].

$$\boldsymbol{P}\left[\left|\sum_{i=1}^{r}(w_i^{*}-\widetilde{w}_i)x\right|>\frac{\epsilon}{L}\right]\leq\sum_{i=0}^{q}\frac{\alpha_i L}{\epsilon}e^{(-i\beta_i)}\binom{rq}{ri}(1-p)^{ri}p^{r(q-i)} \tag{3}$$

이때, 제안한 알고리즘의 경우 부분적으로 잃어버린 행렬의 원소는 언제나 저랭크 근사를 통해 복원되기 때문에 0이 아닌 값을 갖게 된다 [3]. 즉 제안한 알고리즘은 $r(q-i)=0$를 만족하는 대한 경우만 실패 확률로 고려되어 최종적으로 (2)와 같은 상한을 갖으며 언제나 부분 손실된 모델 파라미터를 통해 학습을 진행하는 경우보다 근사 실패 확률이 낮음을 알 수 있다. □

Ⅳ. 실 험

본 실험에서는 하나의 글로벌 서버와 100개의 로컬 디바이스로 구성된 네트워크를 고려하며, 모델 통합에 참여하는 로컬 디바이스의 비율은 0.2, 패킷 손실률은 $p=0.1$로 설정한다. 각 로컬 모델은 DNN(Deep Neural Network) 모델을 이용하며 DNN 모델은 2개의 hidden layer로 각각 100개의 hidden node로 구성한다. 본 실험에서는 MNIST와 CIFAR-10 데이터 셋을 사용하여 이미지 분류 작업을 수행하며, 이때 각 데이터 셋들은 non-IID(Independent and Identically Distribution)에 따라 분포되어있다고 가정한다. 기본 학습율은 0.001로 설정하고, 로컬 epoch는 20, 배치 크기는 10으로 설정한다. 본 실험에서는 불안정한 네트워크에서 모델 파라미터를 전달하는 비교 방법으로, 손실된 모델 파라미터를 채우는 0으로 채우는 zero-filling 기법과, 모델 파라미터를 압축한 후 UDP 환경에서 전송하는 방법으로 제안된 EDEN[2] 기법, 저랭크 근사만을 이용하여 모델 파라미터를 전달하는 FedLR[6]을 이용하여 FL 정확도 성능 저하를 비교 분석한다. 이때 모델 통합에 적용되는 FL 알고리즘은 FedAvg, FedNTD[4], FML[5], FedProx[6]을 사용하여 성능 분석을 진행한다.

표1. 글로벌 epoch가 500일 때 모델 정확도 성능 (%) 결과 비교 분석

	알고리즘	Vanilla	Zero -filling	EDEN	FedLR	Proposed
MNIST	FedAvg	98.1	86.1	75.3	74.2	**97.0**
	FedNTD	75.2	69.1	67.2	66.2	**84.3**
	FML	98.6	91.8	9.5	79.9	**97.7**
	FedProx	95.5	65.5	71.2	70.0	**80.7**
CIFAR-10	FedAvg	74.3	67.5	0	60.5	**70.1**
	FedNTD	74.2	47.8	50.2	53.7	**65.3**
	FML	79.6	72.8	11.8	67.1	**79.1**
	FedProx	72.5	45.9	50.7	51.6	**70.4**

그림 2(a)는 제안한 알고리즘의 상위 k개를 조정하면서 관찰되는 전체 평균 정확도를 나타낸 것으로 상위 k개의 개수가 증가할수록 정확도는 수렴하는 것을 확인할 수 있다. 따라서 본 실험에서는 수렴이 시작되는 지점인 $k/r=48\,\%$를 만족하는 k값을 이용하여 실험을 진행하였다. 그림 2(b)에서는 시간에 따라 관찰되는 정확도 성능을 나타낸 것으로 FedAvg를 이용하여 모델 통합을 하였을 때 다른 파라미터 복구 방법과 비교하여 성능 저하를 크게 방지한 것을 확인할 수 있다. 이 결과는 표 1에서 다른 알고리즘과 데이터 셋을 고려하였을 때에도 항상 우수한 성능을 보이는 것을 확인할 수 있다.

Ⅴ. 결 론

본 논문에서는 불안정한 네트워크 환경에서 모델 파라미터의 부분 손실로 인하여 발생하는 FL 성능 저하를 효과적으로 방지하는 알고리즘을 제안하였다. 제안한 전송 방법은 불안정한 네트워크에서 최적 모델에 근사 실패할 확률을 이론적으로 정리하였고, 실험 결과를 통해서 다른 접근 방식과 비교하여 성능 저하를 크게 방지하며 최적 모델에 접근하는 것을 확인하였다.

ACKNOWLEDGMENT

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2021-0-00739, 분산/협력 AI 기반 5G+ 네트워크 데이터 분석 기능 및 제어 기술 개발, RS-2024-00344830-(총괄10-세부3) 6G 네트워크 구조/산업융합 표준기술 개발 및 표준화)

참 고 문 헌

[1] S. Vargaftik, et al., "EDEN: Communication -efficient and robust distributed mean estimation for federated learning," in ICML, pp. 21984-22014, 2022.

[2] A. Gadhikar, S. Mukherjee, and R. Burkholz, "Why random pruning is all we need to start sparse," ICML, 2023.

[3] Z. Bao, X. Ding, and K. Wang, "Singular vector and singular subspace distribution for the matrix denoising model," Ann. Stat., 49(1), pp. 370-392, 2021.

[6] H. Zhou, J. Cheng, X. Wang, and B. Jin, "Low rank communication for federated learning," in DSFAA Int. Workshops, pp. 1-16, 2020.

[7] G. Lee, M. Jeong, Y. Shin, S. Bae, and S.-Y. Yun, "Preservation of the global knowledge by not-true distillation in federated learning," in NeurIPS, pp. 38461-38474, 2022.

[8] T. Shen, et al., "Federated mutual learning," arXiv preprint arXiv:2006.16765, 2020.

[9] T. Li, et al., "Federated optimization in heterogeneous networks," in Proc. Mach. Learn. Res, 2, pp. 429-450, 2020.