다중사용자 접속환경을 통한 베이시안 연합학습 알고리즘

박찬호, 이남윤
포항공과대학교

{chanho26, nylee}@postech.ac.kr

요 약

본 논문은 다중 사용자 접속환경 (MAC)에서 연합학습의 학습 정확도 및 효율을 높이기 위하여, 부호 확률적 경사하강법 (signSGD) 기반으로 공기 중 연산 (AirComp) 기술을 이용한 새로운 학습 알고리즘을 고안하였다. 제안하는 알고리즘은 각 사용자의 1-비트 채널 정보만을 feedback 하여 효율을 높이고, 학습에 사용되는 sum-gradient 를 사용자의 gradient 분포를 이용해 Bayesian 관점에서 추정한다. 이미지 데이터를 통한 시뮬레이션을 통해, 제안하는 알고리즘이 관련 연구의 다른 알고리즘보다 더 높은 정확도를 얻는 것을 확인하였다.

Ⅰ. 서 론

연합학습은 다수의 말단 연산 장치들과 중앙 서버가 학습에 필요한 정보를 공유하여 모델을 학습하는 기계학습의 한 부류이다. 연합학습은 장치들에 존재하는 데이터가 이질적이어도 서버와 데이터를 공유하지 않고 학습하여 보안 관점에서 뛰어나다. 중앙 서버에서는 장치들이 보낸 정보를 적절히 합하여 최적화 기법을 이용해 모델을 학습한다 [1]. 서버와 장치는 무선 네트워크를 통해 정보를 공유하는데, 학습에 사용되는 모델이 가지는 매개변수가 일반적으로 10^5 개를 넘어 굉장히 큰 통신 비용을 요구한다. 효율적인 학습을 위해 통신 비용을 줄일 수 있는 많은 기술들이 연구 중에 있다 [2]-[3]. 연합학습의 효율을 높일 수 있는 한 예시로 AirComp 기술이 있다. AirComp 란 전자기파들에 대해 중첩의 원리가 적용되는 것을 기반으로, 통신에 참여하는 모든 사용자들이 같은 통신 자원을 사용하여 신호를 동시에 송신하는 통신 기술이다. AirComp 는 연합학습에 매우 적합한 통신 기술로, 사용자들이 보내는 모든 신호가 더해지기 때문에 서버에서 사용자의 신호를 추정할 수 없다. 또한 연합학습은 각 장치들이 보낸 정보를 합하여 학습하기 때문에 이 관점에서도 굉장히 적합하다. AirComp 기술을 이용하여 연합학습을 진행한 이전 연구로는, truncated channel-inversion precoding 기법을 적용하여 연합학습을 진행하는 OBDA 시스템이 있다 [4].

본 논문에서는 AirComp 통신 기술을 사용하여 signSGD 최적화 기법을 기반으로 한 새로운 연합학습 알고리즘을 제안한다. AirComp 기술과 signSGD 최적화 기법을 사용하여 제안하는 알고리즘의 통신 비용을 크게 줄일 수 있도록 설계하였다. 각 말단 장치가 받는 채널 정보의 1-비트만을 precoding 에 사용하며, 1-비트만으로도 기존 연구들의 성능을 만족하는 것을 보여준다. 추가하여, 연산된 gradient 의 분포를 이용해 Bayesian 관점에서 sum-gradient 를 추정한다. 본 알고리즘을 이용하여 이미지 분류를 목적으로 학습된 모델의 정확도를 다른 알고리즘들과 비교한다.

Ⅱ. 본론

본 논문에서 사용되는 연합학습 시스템은 중앙 서버와 K 대의 말단 장치들로 구성 되어있다. 각 말단 장치는 자신만의 이질적인 데이터셋 $\mathcal{D}_k$ 을 가지고, 데이터셋에는 총 N_k 개의 데이터 $\mathbf{A}_k$ 와 해당 데이터의 라벨 b_k 이 있다. 서버와 장치들은 학습 모델에 사용되는 매개변수 벡터 $\mathbf{w}$ 를 공유한다. 연합학습의 목적은 모델의 정확도를 높이는 것으로, 학습을 통해 계산되는 loss 를 최소화하여 목적을 달성할 수 있다. 각 장치의 데이터와 모델 매개변수를 통해 정의되는 local loss function $f_k(\mathbf{w})$ 은 다음과 같이 표현할 수 있다.

$$f_k(\mathbf{w}) = \frac{1}{N_k}\sum_{i=1}^{N_k} \ell\left(\mathbf{A}_k^i, b_k^i; \mathbf{w}\right) \quad (1)$$

Global loss function $F(\mathbf{w})$ 은 장치들이 가지는 모든 데이터에 대한 loss 를 구하는 함수로, local loss function 의 weighted sum 형태로 구할 수 있다.

$$F(\mathbf{w}) = \sum_{k=1}^{K} \frac{N_k}{N} f_k(\mathbf{w}) \quad (2)$$

모든 말단 장치들은 loss 를 최소화하는 $\mathbf{w}$ 를 찾기 위해 local loss function 에 대한 gradient $\nabla f_k(\mathbf{w})$ 을 계산한다. 서버에 송신할 때 사용되는 통신 비용을 줄이기 위해 gradient 를 1 비트로 quantize 하고, 이후 말단 장치들은 quantize 된 gradient $\mathbf{x}_k$ 를 무선 네트워크를 통해 송신한다. 중앙 서버에서는 장치들이 보낸 모든 신호들을 받은 뒤 모든 장치들에 대한 sum-gradient $\mathbf{g}_\Sigma$ 를 추정한다. 마지막으로 signSGD 최적화 기법을 기반으로 sum-gradient 를 이용하여 학습하는 모델 매개변수를 새로 갱신한다 [3]. t 번째 round 에 진행된다고 할 때 갱신되는 모델 매개변수는 다음과 같이 쓸 수 있다.

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta^t \mathrm{sign}(\mathbf{g}_\Sigma^t) \quad (3)$$

서버에서는 갱신된 $\mathbf{w}$ 를 모든 말단 장치들에게 다시 보내며, 이 과정을 학습 loss 가 일정 값 이하로 줄어들 때까지 반복한다.

서버와 장치들이 통신하는 시스템을 본 논문에서는 주파수 분할 다중화 방식 (FDD) 시스템으로 구상하였고, 효율적인 연합학습 시스템을 위해 AirComp 통신 기술을 사용하였다. 하향링크 (downlink) 통신은 서버가 높은 전송 power 를 가지고 있기에 말단 장치들이 완벽히 decoding 가능한 것으로 가정하지만, 상향링크 (uplink) 통신은 그렇지 않다. 무선 네트워크를 이용하면서 채널 fading 과 noise 로 인해 송신된 신호가 변형되는데, 문제 해결을 위해 precoding 기법을 사용한다. t번째 round 에 k 번째 장치가 받는 채널 계수 $h_k^t \sim \mathcal{N}(0,1)$, i 번째로 송신되는 gradient $x_{k,i}^t$, 대응되는 noise $n_i^t \sim \mathcal{N}(0,\sigma_n^2)$ 에 대하여 p_k^t로 precoding 된다고 할 때 서버가 받게 되는 신호 y_i^t는 다음과 같다.

$$y_i^t = \sum_{k=1}^{K} h_k^t p_k^t x_{k,i}^t + n_i^t \quad (4)$$

모든 장치들은 gradient 를 송신하는데 최대 P 의 power 를 사용할 수 있다고 설정하였다.

제안하는 연합학습 알고리즘은 기존의 AirComp 를 이용한 연합학습 방법들보다 더 효율적이고 높은 정확도를 얻기 위하여 새로운 precoding 기법을 제시한다. 각 장치에 대응되는 채널 정보의 1 비트만을 precoding 에 사용하여 채널 정보에 대한 feedback 양을 줄인다. 또한, gradient 가 1 비트로 quantize 되기 때문에 gradient 의 부호만 보존되면 학습 성능에 영향이 없다는 직관 하에, 제안하는 precoding 방식은 gradient 부호 유지가 가능해 굉장히 획기적이다. 따라서 서버가 받게 되는 신호는 다음과 같이 바꿔서 표현할 수 있다.

$$y_i^t = P\sum_{k=1}^{K}|h_k^t|x_{k,i}^t + n_i^t \quad (5)$$

알고리즘에 추가된 다른 기술은 각 장치들의 gradient 분포를 추정하여 sum-gradient 추정에 이용하는 것이다. 이미지 데이터에 대한 gradient 분포 추정 연구 결과를 이용하여 Gaussian 혹은 Laplace 분포를 따르는 것으로 가정하고, 각 말단 장치에서는 gradient 에 적합한 moment 값들을 추정한다 [5]. 이 정보를 gradient 와 함께 서버에 송신하여 최소 평균제곱오차 (MMSE) 관점에서 sum-gradient 를 추정하는데 사용한다. 이 기술을 사용함으로써 말단 장치들이 이질적으로 데이터를 가지더라도 각 장치들의 gradient 분포를 이용하여 sum-gradient 를 정확히 추정하도록 한다.

제안하는 알고리즘은 위의 기법들을 활용하여 장치들이 이질적으로 데이터를 가지더라도 적은 채널 정보만으로 기존 방식들보다 더 좋은 학습 정확도를 가지는 것을 목표로 한다.

Ⅲ. 결론

제안하는 연합학습의 학습 정확도를 확인하기 위해 MNIST, CIFAR10 데이터셋을 분류하는 모델을 학습하는 시뮬레이션을 진행하였다. 모든 데이터셋에 대해 장치들이 동질적 혹은 이질적으로 데이터를 가지는 두 경우로 나누어 실험하였다. 학습에 사용한 모델은 MNIST 와 CIFAR10 데이터셋에 대해서 각각 CNN 과 ResNet44 모델로 학습하였다. 연합학습에 참여하는 말단 장치는 총 100 대로, 연합학습의 한 round 가 진행될 때 100 대 중 10 대의 장치를 무작위로 뽑아 해당한 장치의 데이터로만 학습을 진행한다. 각 장치는 1km 반경의 COST-231 HATA 모델링에 따라 pathloss 를 받고, 사용 가능한 최대 power P는 1 로 설정하였다.

기존에 존재했던 AirComp 기법을 사용한 연합학습 방법과의 비교를 위해 사용한 알고리즘은 [4]의 OBDA 이다. OBDA 는 말단 장치에서 truncated channel-inversion precoding 기법을 활용하며, 서버에서는 받은 모든 gradient 들에 대해 majority-voting 기반으로 학습에 사용할 최종 gradient 를 유추한다.

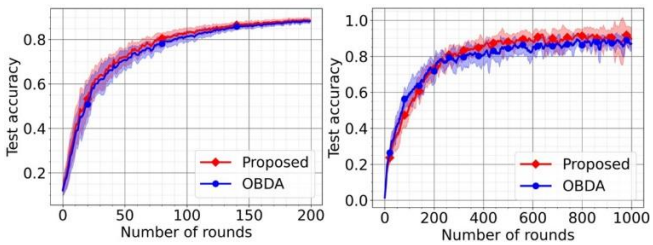

그림 1. MNIST 데이터셋에 대해 제안한 알고리즘과 OBDA 알고리즘으로 학습시킨 모델의 학습 정확도 비교

그림 1 과 2 는 각각 MNIST 와 CIFAR10 데이터셋으로 학습하여 얻은 학습 정확도 그래프이다. 왼쪽과 오른쪽 그래프는 데이터가 동질적 혹은 이질적으로 장치들에 나눠진 경우에 대한 학습 정확도 결과이다. MNIST 데이터셋에 대해서는 데이터가 나눠진 방식에 따라 round 차이는 있지만 90%가 넘는 정확도를 보였다. CIFAR10 데이터셋에서는 5000 round 동안 동질적, 이질적으로 나눠진 경우 각각 약 70 %, 65%의 정확도를 보였다. 두 데이터셋 모두 동질적으로 나눠진 경우 제안하는 알고리즘이 OBDA 에 비해 1% 정도 높았고, 이질적으로 나눠진 경우 약 3.5% 높은 것을 확인하였다.

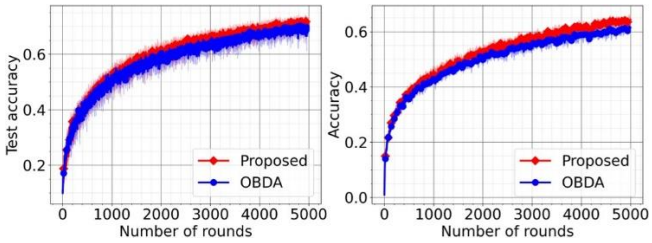

그림 2. CIFAR10 데이터셋에 대해 제안한 알고리즘과 OBDA 알고리즘으로 학습시킨 모델의 학습 정확도 비교

결과적으로, MNIST 와 CIFAR10 데이터셋 모두 데이터가 동질적으로 나누어진 경우 제안하는 알고리즘이 OBDA 와 비슷하거나 조금 높은 학습 정확도를 보이는 것을 확인하였다. 이질적으로 나눠진 경우에는 제안하는 알고리즘이 각 장치의 gradient 분포를 이용하였기 때문에 OBDA 에 비해 더 높은 정확도를 보이는 것을 확인하였다.

참 고 문 헌

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in Artif. Intell. And Statist. PMLR, 2017, pp. 1273-1282.

[2] S. Shi, Q. Wang, K. Zhao, Z. Tang, Y. Wang, X. Huang, and X. Chu, “A distributed synchronous SGD algorithm with global Top-k sparsification for slow bandwidth networks,” in 2019 IEEE 39th int. Conf. Distrib. Comput. Systems. (ICDCS). IEEE, 2019, pp. 2238-2247.

[3] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, “signSGD: Compressed optimization for non-convex problems,” in Int. Conf. Mach. Learn. (ICML). PLMR, 2018, pp. 560-569.

[4] G. Zhu, Y. Du, D. Gündüz, and K. Huang, “One-bit over-the-air aggregation for communication-efficient federated edge learning: Design and convergence analysis,” IEEE Trans. Wireless. Commun., 2020.

[5] S. Lee, C. Park, S.-N. Hong, Y. C. Eldar, and N. Lee, “Bayesian federated learning over wireless networks,” arXiv preprint arXiv:2012.15486, 2020.