다양한 컴퓨팅 환경에 강인한 분산형 signSGD 알고리즘

박찬호, 이남윤*
포항공과대학교, *고려대학교

chanho26@postech.ac.kr, *namyoon@korea.ac.kr

요 약

본 논문은 부호 확률적 경사 하강법 및 다수결 부호 집계 (signSGD-MV)을 부호 이론의 시선으로 접근하여, 최대우도법 (MLE) 관점에서 최적의 분산 학습 알고리즘인 IT-signSGD 를 제안한다. IT-signSGD 는 그래디언트의 부호 정보만을 요구하는 signSGD 의 뛰어난 통신 효율을 따르면서도, 로그우도비 (LLR) 가중치를 활용하여 학습 참여자들의 연산 능력이 다를 때의 학습 성능을 향상시킨다. 본 논문의 이론적 분석 및 실험 결과들이 IT-signSGD 의 성능을 입증한다.

Ⅰ. 서론

본 논문은 분산 학습의 고질적 문제인 통신 비용 및 사용자들의 이질적 연산 문제를 동시에 고려하는 IT-signSGD 알고리즘을 제안한다. 분산 학습 시스템에서 공유되는 그래디언트 정보에 대한 통신 비용은 학습 모델 크기에 비례하는데, 학습에 사용되는 신경망 모델의 크기가 기하급수적으로 증가하여 엄청난 통신 비용이 요구된다. 그 대안으로 각 그래디언트 성분의 부호 정보만을 활용하여 통신 비용을 기존 대비 32 배 줄일 수 있는 signSGD-MV [1] 알고리즘이 개발되었지만, 사용자들의 연산 능력에 차이가 존재하는 현실적인 제약 하에 signSGD-MV 는 잘 동작하지 않는 한계를 가진다. IT-signSGD 는 기존의 다수결 부호 집계에 LLR 가중치를 추가적으로 활용하여, 이질적 연산 환경에 강인한 특성을 지닌다. 또한 해당 집계 방식은 MLE 관점에서 최적이라는 이론적 가치도 지니며, 하기될 실험 결과들은 IT-signSGD 의 우월한 성능을 입증한다.

Ⅱ. 본론

본 논문에서 구상하는 분산 학습 시스템은 중앙 서버와 M 명의 사용자들로 구성된다. 사용자 $m \in [M]$ 은 각자의 데이터셋 $\mathcal{D}_m$ 을 보유하고 있다. 학습 모델(인공 신경망)의 매개변수는 $\mathbf{x} = [x_1, \cdots, x_N] \in \mathbb{R}^N$ 으로 표현한다. 분산 학습의 목표는 학습 목표에 적합한 손실 함수 $F(\cdot)$를 바탕으로 목적 함수 (1)을 최소화하는 것이다.

$$\min_{\mathbf{x}\in\mathbb{R}^N} f(\mathbf{x}) \coloneqq \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{\mathbf{d}\sim\mathcal{D}_m}[F(\mathbf{x};\mathbf{d})]. \quad (1)$$

SGD 최적화 방법을 따르면, 사용자 m은 매 학습 라운드 $t \in [T]$ 마다 각자의 데이터셋 내에서 무작위로 B_m 개의 데이터 샘플들 $\mathcal{B}_m^t \subset \mathcal{D}_m$ 을 선택한다. 이 데이터 샘플들을 미니배치라고 하며, 이를 이용하여 매개변수 $\mathbf{x}^t$에 대한 그래디언트를 (2)와 같이 연산한다.

$$\mathbf{g}_m^t = \frac{1}{B_m}\sum_{\mathbf{d}\in\mathcal{B}_m^t} \nabla F(\mathbf{x}^t;\mathbf{d}) \in \mathbb{R}^N. \quad (2)$$

이후 signSGD 최적화 방법을 따라 사용자들은 각 그래디언트 성분의 부호 정보 $Y_{m,n}^t = \text{sign}(g_{m,n}^t)$ 만을 서버에 전송한다. 중앙 서버는 사용자들로부터 받은 부호 정보들을 모아 LLR 가중치 다수결 집계 (3)을 진행한다.

$$\hat{U}_n^t = \text{sign}\left(\sum_{m=1}^{M} \log\frac{1-\hat{p}_{m,n}^t}{\hat{p}_{m,n}^t}\cdot Y_{m,n}^t\right). \quad (3)$$

이 때 가중치에 활용되는 교차 확률 (4)는 이전 학습 라운드 동안 집계된 부호 결과와 사용자가 보내온 부호가 다른 횟수를 세어 그 비율로 추정한다.

$$\hat{p}_{m,n}^t = \frac{\sum_{i=1}^{t}\mathbf{1}[Y_{m,n}^t \neq \hat{U}_n^t]}{t}. \quad (4)$$

다만, 학습 초기에는 오차가 존재하는 샘플 수가 부족하여 가중치가 다소 불안정할 수 있어, 초기 구간에는 signSGD-MV 와 같이 가중치를 활용하지 않는다. 추정한 이후 LLR 가중치를 업데이트하며, 서버는 모든 성분에 대해 집계한 부호들을 다시 사용자들에게 전송한다. 최종적으로, 각 사용자는 모델 매개변수를 (5)와 같이 업데이트한다.

$$\mathrm{x}_n^{t+1} = \mathrm{x}_n^t - \delta\cdot\hat{U}_n^t, \qquad \forall n \in [N]. \quad (5)$$

이 전체 학습 과정은 모델이 수렴할 때까지 반복된다.

Ⅲ. 결론

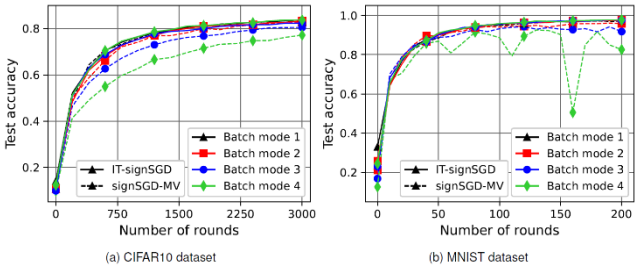

그림 1. 배치 모드에 따른 IT-signSGD 와 signSGD-MV 의 학습 정확도 비교 그래프

제안하는 IT-signSGD 알고리즘의 학습 성능을 확인하기 위해 MNIST, CIFAR10 데이터셋을 분류하는 모델 학습 시뮬레이션을 진행하였다. 총 $M = 15$ 명의 사용자가 학습에 참여하며, 이질적 연산 시스템을 구상하기 위해 작은 $B_m = 4$을 사용하는 사용자 수를 0, 9, 12, 14 명으로 늘려가며 배치 모드 1~4 를 구상하였다. 그림 1 은 각 배치 모드에 따라 signSGD-MV 와의 학습 정확도를 비교한 그래프로, 사용자들의 연산 능력이 이질적으로 변할수록 IT-signSGD 가 더 우월한 성능을 달성함을 확인할 수 있다.

ACKNOWLEDGMENT

이 논문은 2024 년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No. 2021-0-00467, 지능형 6G 무선 액세스 시스템).

참 고 문 헌

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