Bayesian Model Reduction 기반 환경 적응적 무선 연학학습

김주영, 홍준표
홍익대학교

collala1234@g.hongik.ac.kr, jp_hong@hongik.ac.kr

Bayesian Model Reduction based Adaptive Wireless Federated Learning

Ju Young Kim, Jun-Pyo Hong
Hongik Univ.

요 약

본 논문은 무선 엣지 네트워크에서의 연학학습에서 희소하고 상이한 로컬 데이터 분포와 제한적인 통신 자원으로 인한 과적합, 학습 지연, 수렴 불안정성 문제를 완화하기 위한 해결책으로 Bayesian model reduction(BMR) 기반 적응적 연합학습(FL: Federated learning) 기법을 제안한다. 즉, 제안 기법은 Bayesian 접근 기반의 지역학습 및 종합(aggregation)을 통해 학습 불안정성을 완화하고, BMR 을 통한 모델 경량화를 수행함으로써 Bayesian 접근법에 따른 추가적은 통신 부하를 효과적으로 줄일 수 있다.

Ⅰ. 서 론

무선 엣지 네트워크 환경에서의 연합학습에서는 통신 병목이 학습 속도를 결정하는 핵심 요인이며, 학습 결과를 확률 분포 형태로 나타내는 Bayesian 연합학습(FL: Federated learning)에서는 전송량 증가로 인해 학습지연이 더욱 악화될 수 있다 [1, 2, 3]. 이를 완화하기 위한 효과적인 전략으로 모델 경량화를 고려할 수 있으며, 모델 파라미터의 불확실성 정보에 기반한 모델 경량화가 중앙 학습 환경하에서 폭넓게 연구되어 왔다. 하지만 이들 대부분은 SNR 혹은 SNP 와 같은 경험적(heuristic) 지표에 의존하기 때문에, 그 성능이 임계값 선택에 따라 민감하고, pruning 이 학습 목적함수(ELBO/VFE) 최소화와의 연관성을 찾기 어려운 단점을 갖는다. 이론에 기반한 모델 경량화 기법인 Bayesian Model Reduction (BMR)을 신경망에 적용이 최근 연구되어, 기존 지표보다 효과적인 경량화가 가능함을 보였다. 이에 본 논문은 BMR 과 결합된 구조의 새로운 FL 프레임워크를 제안해 학습 불안정성 완화는 물론 통신 자원이 제한된 환경에서도 Bayesian 접근 도입으로 인한 지나친 통신 지연이 발생하지 않도록 한다.

Ⅱ. 본 론

BMR 은 큰 베이지안 모델에서 일부 파라미터를 제거해 만든 하위모델들을 재평가 없이 빠르게 비교하는 기법이다. 이는 각 파라미터 제거에 따른 Variational Free Energy(VFE) 변화량, ΔF 를 효율적으로 계산할 수 있게 한다. VFE 는 변분추론(variational inference)기반의 베이지안 학습에서 목적함수로써, 결국 BMR 을 통해 베이지안 학습과 모델 경량화는 동일한 목표를 갖게 된다 [3]. 제안하는 FL 프레임워크에서는 단말의 지역학습 후 얻어진 지역 posterior 분포를 기반으로 BMR 을 적용해 경량화된 모델을 구성하고 이를 전송함으로써, 성능 열화 없이 통신 부하를 효과적으로 줄일 수 있다.

본 연구에서 고려하는 무선 엣지 네트워크는 서버 역할을 수행하는 하나의 기지국과 K 개의 단말로 구성되어 있으며, 각 단말 $k \in \{1, 2, \dots, K\}$ 는 지역 데이터셋 D_k 를 갖는다. 이와 같은 환경에서 제한된 통신 자원으로 연합학습을 통해 모델 파라미터 $\mathbf{w} \in \mathbb{R}^d$ 에 대한 전역 posterior 분포 $p(\mathbf{w}|\{D_k\}_{k\in K})$ 를 도출하는 것을 목표로 한다. 현실적 제약을 고려해 posterior 분포는 변분 추론(variational inference)을 기반으로 mean-field Gaussian 분포로 근사한다. 따라서 posterior 는 평균과 표준편차로 구성된 변분 파라미터(variational parameters) $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}) \in \mathbb{R}^{2d}$ 로 표현된 수 있다. 제안 기법의 구체적인 동작은 다음과 같다.

① *전역 posterior 배포:* 훈련 라운드 t 가 시작되면, 서버는 전역 posterior 분포 $q_{\boldsymbol{\theta}_t}(\mathbf{w})$에 대한 정보를 모든 단말들에게 전송한다.

② *변분 추론 기반 지역 학습:* 각 단말 k는 지역 posterior 분포와의 KL divergence 를 최소화할 수 있는 변분 파라미터를 다음 최적화 문제를 풀어 도출한다

$$\boldsymbol{\theta}_{t,k} = \operatorname{argmin}_{\boldsymbol{\theta}} \mathrm{KL}[q_{\boldsymbol{\theta}}(\mathbf{w}) \| p(\mathbf{w}|\mathcal{D}_k)] = \operatorname{argmin}_{\boldsymbol{\theta}} \mathrm{KL}[q_{\boldsymbol{\theta}}(\mathbf{w}) \| p(\mathbf{w})] - \mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{w})}[\log p(\mathcal{D}_k|\mathbf{w})]. \quad (2)$$

③ *지역 모델 경량화:* BMR 을 기반으로 i번째 모델 파라미터 w_i가 제거되었을 때, VFE 의 변화량을 기반으로 해당 파라미터의 중요도를 판별해 제거해 기지국으로 전송한다. 즉, i 번째 모델 파라미터 w_i에 대한 VFE 변화량은 다음과 같다.

$$\Delta F_{t,k,i} = \log \int q_{\theta_{t,k,i}}(w) \frac{\tilde{p}(w)}{q_{\theta_{t,i}}(w)} dw, \quad (3)$$

여기서 $\theta_{t,k,i} = (\mu_{t,k,i}, \sigma_{t,k,i})$ 는 i 번째 지역 모델 파라미터의 평균과 표준편차, $\theta_{t,i} = (\mu_{t,i}, \sigma_{t,i})$ 는 i 번째 전역 모델 파라미터의 평균과 표준편차, $\tilde{p}(w) = \mathcal{N}(w|0, \epsilon)$, $\epsilon \approx 0$.

④ *지역 posterior 분포 합성을 통한 전역 posterior 갱신:* 경량화된 지역 posterior 분포들의 곱연산을 통해 전역 posterior 분포를 갱신한다. 이와 같은 곱연산은 다수의 확률분포를 대표하는 하나의 나타낼 때, 정보 손실을 최소화 할 수 있는 방법이다. 즉, 갱신된 전역 posterior 분포는 다음과 같다.

$$q_{\boldsymbol{\theta}_{t+1}(\theta)} = C \prod_{k \in \mathcal{K}} q_{\tilde{\boldsymbol{\theta}}_{t,k}}(\mathbf{w}) \quad (4)$$

여기서 C는 분포의 적분값을 1 로 만들기 위한 정규화 상수이고, $\tilde{\boldsymbol{\theta}}_{t,k} \in \mathbb{R}^{2d}$ 단말 k 로부터 수신한 경량화 변분 파라미터를 나타낸다.

⑤ *수렴 조건 확인 및 반복:* 전역 posterior 분포가 수렴 조건을 만족하는지 확인하고, 수렴하지 않았다면, 훈련 라운드 $t+1$ 을 시작하고 단계 ①~④를 수행한다.

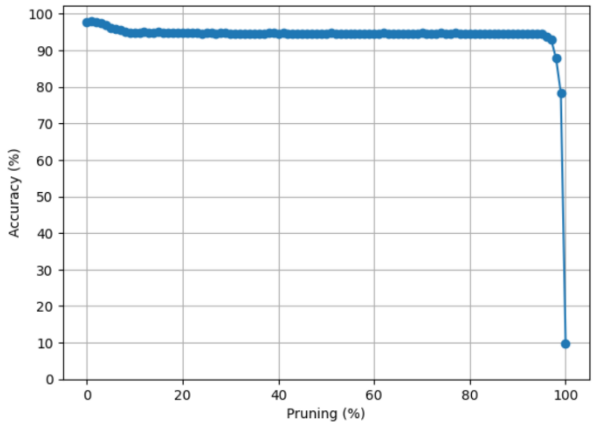

그림 1. 모델 경량화 비율에 따른 성능 변화

제안 기법에 대한 구현 및 검증에 앞서, 보다 간단한 중앙학습 환경에서 BMR 접근법의 효용성을 시험해 보았다. 4 개의 완전연결층(784*224*672*25*10)으로 구성된 MLP(Multy Layer Perceptrone)을 모델로 사용하여, MNIST 분류 문제에서의 성능 이득을 살펴보았다. 그림 1 은 VFE 변화율 기준으로 한 모델 파라미터 제거 비율 변화에 따른 분류 정확도를 실험한 결과이다. 제거 비율이 증가함에 따라 성능이 감소하지만 낮은 중요도의 파라미터를 우선적으로 제거하기 때문에 이에 따른 성능 저하는 미미함을 확인할 수 있었다. 특히, 전체 파라미터 중 97%를 제거한 경우에도 정확도 저하가 약 3%에 불과함을 확인하였다. 이는 BMR 기반 대규모 파라미터 감축이 성능 저하 없이 가능함을 시사하며 Bayesian FL 에서의 과도한 전송량으로 인한 학습 지연을 완화할 수 있는 유효한 대안이 될 수 있음을 의미한다.

Ⅲ. 결 론

본 연구에서는 BMR 을 통해 Bayesian FL 의 통신 부하를 효과적으로 낮출 수 있는 기법에 대해 제안하였다. 현재까지는 제안 기법에 대한 가능성만 검증한 상태로, 추가적인 연구 진행을 통해 연합학습 환경에서의 수렴 속도 및 정확도 측면에서의 성능을 분석할 계획이다.

ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (RS-2024-00464570)

참 고 문 헌

[1] H. Brendan McMahan et al., “Communication-Efficient Learning of Deep Networks from Decentralized Data,” in *Proc. AISTATS*, 2017.

[2] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra,“Weight Uncertainty in Neural Networks,” in *Proc. 32nd International Conference on Machine Learning (ICML)*, Lille, France, Jul. 2015, pp. 1613–1622.

[3] J.-P. Hong, H. Seo, and K. Lee, “Distribution-level AirComp for wireless federated learning under data scarcity and heterogeneity,” arXiv:2506.06090, 2025.

[4] J. Beckers, B. van Erp, Z. Zhao, K. Kondrashov, and A. de Vries, “Principled pruning of Bayesian neural networks through variational free energy minimization,” *IEEE Open J. Signal Process.*, vol. 5, pp. 195–203, 2024