Multi-prize Lottery Ticket Hypothesis 기반 연합학습에 관한 연구

이상민, 장혜령
동국대학교

leesm.exe@dgu.ac.kr, hyeryung.jang@dgu.ac.kr

Accurate and Computation-efficient Federated Learning with Multi-prize Lottery Ticket Hypothesis.

Sangmin Lee, Hyeryung Jang
Dongguk Univ.

요 약

연합학습(Federated Learning)은 인공지능 학습 과정에서, 각 단말에서 수집된 데이터를 서버에 공유하지 않고 데이터를 보유한 기기에서 학습을 진행한 후 서버에서 결과를 취합하여 범용 모델을 학습하는 기법이다. 연합학습은 개인정보를 침해하지 않으므로 모바일이나 사물인터넷과 같이 풍부하고 분석 가치가 높은 사용자의 데이터를 안전하게 학습할 수 있다. 다만 연산학습에서 학습을 수행하는 기기들은 연산 능력이 제한적인 경우가 많아 복잡한 모델을 학습하기 어렵다. 본 논문에서는 모델을 압축하고 가지치기 하는 Multi-prize Lottery Ticket Hypothesis 를 사용하여 연산 효율적이면서도 정확도를 유지하는 연합학습 기법을 제안한다. 특히 학습 데이터가 non-iid 로 분포한 상황에서 기존 방법들 대비 우수한 정확도를 가지면서도 80%가량 적은 가중치를 갖는 희소 신경망을 학습할 수 있음을 확인하였다.

1. 서 론

모바일이나 사물인터넷(IoT: Internet of Things)은 최근 스마트홈과 같은 미래지향 서비스를 제공하기 위한 핵심기술로 주목받고 있다. 특히 사용자 만족도를 높이기 위해 사용자의 데이터를 분석하는 인공지능 알고리즘이 요구되는 반면, 이러한 데이터의 사용은 개인정보를 침해할 우려가 있어 실질적인 적용이 어렵다. 연합학습(Federated Learning)은 사용자의 데이터를 수집하지 않고 데이터를 보유한 다수의 기기에서 각자 학습을 진행한 후 학습 결과만을 공유하여 학습을 수행하는 기법이다[1]. 이는 개인정보를 침해하지 않고 학습을 가능케 하지만 IoT 혹은 기기에서 학습이 이루어지므로 부족한 연산 능력으로 인한 한계점을 가진다.

이를 해결하기 위해 학습 모델의 복잡성을 낮추는 두 가지 방법이 주로 연구되었다. 첫째로 모델의 가중치를 32 비트에서 1 비트로 압축하는 이진 신경망(BNN: Binary Neural Network)에 관한 연구가 대표적이며[2], 둘째로 중요도가 낮은 가중치를 가지치기 하여 희소 신경망을 구성하되 높은 정확도를 얻는 Lottery Ticket Hypothesis (LTH)이 대표적이다[3]. 또한 이 두 방법을 융합하여 모델을 더욱 간소화하면서도 높은 정확도를 유지하는 Multi-prize Lottery Ticket Hypothesis (MPT)가 최근 제시되었다[4]. MPT 를 통해 약 80%의 가중치를 가지치기한 1 비트 이진 신경망에서도 기존 32 비트 신경망과 유사한 정확도를 달성할 수 있어 BNN 과 LTH 단일기법보다 높은 효과를 얻을 수 있다.

본 논문은 연합학습의 학습모델로 MPT 를 사용하여 적은 연산과 메모리로 학습과 추론이 가능한 기법을 제시하고자 한다. 특히 실험으로 기존 연합학습 기법 대비 성능 향상을 확인하였으며, 하드웨어 제약이 많은 IoT 기기에서 인공지능의 실질적인 활용성을 높일 수 있을 것으로 기대한다.

2. 본론

2.1 FL: Federated Learning

연합학습은 학습 데이터를 서버에 모으지 않고, 학습 데이터를 보유한 클라이언트에서 로컬 학습을 수행하고, 학습 결과만을 서버에서 취합하여 범용 학습 모델을 얻는 방법이다. 대표적인 기법은 FedAvg[1]로, 각 클라이언트에서 학습된 가중치 값을 서버에 보내고, 서버에서는 평균 값을 계산하여 향상된 모델을 얻는 기법이다. 이는 클라이언트에서 학습을 위해 많은 연산량을 수행해야 하는 한계점을 가진다.

2.2 MPT: Multi-prize Lottery Ticket Hypothesis

MPT[4]는 충분히 거대한(over-parameterized) 신경망이라면, 가중치의 학습 과정 없이 본래 신경망보다 높은 정확도를 달성하는 1 비트 이진 신경망(BNN)을 찾을 수 있다는 이론이다. MPT 는 약 80%의 가중치를 가지치기한 1 비트 부분 신경망에서도 본래 모델의 정확도를 유지할 수 있다. 최적의 희소 신경망을 찾기 위해 가중치의 중요도 (가지치기 점수: pruning score)를 학습하는 과정을 필요로 한다.

2.3 MPT-FL 학습 알고리즘

본 논문에서는 높은 정확도를 가지면서도 연산 효율적인 MPT 를 연합학습에 사용하는 MPT-FL 기법을 제안한다(표 1 참고). 구체적으로, 각 클라이언트에서는 MPT 에 따라 학습을 수행한다. MPT 는 모델의 1 비트 가중치를 직접 학습하지 않고, 초기값을 그대로 유지하며 각 가중치의 pruning score 를 학습한 후 절댓값 크기를 기준으로 원하는 비율 (pruning ratio)만큼 가지치기 하여 희소 신경망을 생성한다. 생성된 희소 신경망은 학습 데이터를 통해 손실을 계산할 수 있어, 손실을

최소화하는 방향으로 pruning score 를 학습한다. 각 클라이언트에서 학습된 pruning score 를 서버에서 평균 취합하여 다시 클라이언트에 배포하는 과정을 반복하여 최종적으로 모델의 크기가 작으면서도 본래 신경망의 높은 정확도를 유지하는 범용 모델을 얻을 수 있다.

표 1. MPT-FL 알고리즘

Algorithm 1 MPT-FL

Input: Pruning ratio P, **Loss function** L

Sever Executes:

Initialize global weight W_g of network depth l and of width $\{k_j\}_{j=1}^{l}$
$B_0 \leftarrow sign(W_g)$ //Binarize the global initial model
Initialize all clients model with B_0
Initialize Layerwise global pruning scores S_g
for each round $t = 1, 2, \ldots$ **do**
  Randomly select K clients, indexed by k.
  for each client $k \in K$ **do**
    $S_k^{t+1} \leftarrow \text{ClientUpdate}(B_0, S_g^t, W_g)$
  end for
  $S_g^{t+1} \leftarrow$ (aggregate pruning score $\{S_k^{t+1}\}$)
end for

ClientUpdate(B_0, S_g^t, W_g)**:**

$S_k \leftarrow S_g^t$
for each local epoch n from 1 to E **do**
  $\{\tau(i)\}_{i=1}^{k_j} \leftarrow$ Sorting of indices $\{i\}_{i=1}^{k_j}$ s.t. $|S_k^{\tau(i)}| \leq |S_k^{\tau(i+1)}|$
  $M \leftarrow \mathbb{1}_{\{\tau(i) \geq \lfloor k_j P/100 \rfloor\}}(i)$ // Make layerwise pruning mask
  $\alpha \leftarrow \|M \odot W_g\|_1 \,/\, \|M\|_1$ // Calulate amplitude gain at layer l
  $S_k \leftarrow S_k - \eta \nabla_{S_k} L(\{\alpha(M \odot B_0)\})$ // Update layerwise pruning scores
end for

return S_k to server

2.4 실험 결과

실험 환경. 100 명의 클라이언트를 설정하며 라운드마다 무작위로 10 명을 선정하여 학습한다. 모델은 MPT 논문의 Conv-4 를 사용하였으며, 4 개의 Conv2D layer 와 3 개의 FC layer 로 구성되어 총 1,932,352 개의 파라미터를 가진다.

데이터. 연합학습을 위해 MNIST 손 글씨 숫자 데이터를 변형하여 사용한다. 여러 기기가 서로 다른 데이터를 보유하는 non-iid 상황에 부합하도록 각 클라이언트가 2 개의 클래스에 대해서 각 20 개씩의 데이터를 갖는 환경을 구성하였다. 학습 데이터 4000 개를 100 명에게 분배하며, 모델의 검증을 위해 서버 모델에 전체 테스트 데이터를 사용해 검증한다.

비교 기법. MPT-FL 의 성능을 평가하기 위해 4 개의 비교 모델, DNN, MPT[4], FedAvg[1], BNN-FL 을 선정하였다. DNN 과 MPT 는 연합학습하지 않고 전체 데이터로 학습한 standalone 모델이며, FedAvg 는 가장 기본이 되는 32 비트 연합학습 모델이다. BNN-FL 은 연합학습에 이진 신경망을 사용하여 구현한 기본적인 모델이다. BNN-FL 과 MPT-FL 은 신경망의 가중치만 이진화한 경우(1/32)와 가중치와 입력을 모두 이진화한 경우(1/1)를 모두 확인하였다.

표 2 는 학습 중 달성한 최고 테스트 정확도와 최종 모델 크기, 그리고 pruning ratio 를 나타낸다. 연합학습 여부와 관계없이 MPT 기반 학습 기법 (MPT 1/32, MPT-FL 1/32)이 MPT 를 사용하지 않은 경우보다 높은 정확도를 달성함을 확인함으로써 MPT 이론을 실험을 통해 검증하였다.

또 다른 1 비트 모델인 BNN 을 사용한 BNN-FL 과 비교하여도 1/1 일 때와 1/32 일 때 모두 MPT-FL 이 평균적으로 0.79% 향상된 모습을 보여주었다. 특히, MPT-FL 은 높은 정확도를 유지하면서도 최종 모델 크기는 본래 DNN 모델의 크기인 7,729,408 바이트에서 241,544 바이트로 32 배 감소시켜 모델의 복잡도를 크게 낮추었다. 모델의 크기는 메모리 사용량과 전력 소모에 큰 연관성이 있어[2], MPT-FL 은 클라이언트의 메모리와 전력 소모량을 크게 줄일 수 있다.

그림 1 은 MPT-FL 의 가중치를 pruning ratio 에 따른 테스트 정확도를 나타낸다. MPT-1/32 은 모든 구간에서 FedAvg 를 능가하는 성능을 보여주며, MPT-1/1 은 pruning ratio 가 높아질수록 높은 정확도를 보여 FedAvg 와 비슷한 값을 가지게 된다.

표 2. 기법의 정확도, 모델 크기와 희소 비율의 비교.

Method	Test Acc(%)	Model Size(byte)	Sparsity(%)
DNN	99.34±0.04	7,729,408	0%
MPT 1/32	**99.46±0.04**	**241,544**	**80%**
FedAvg	97.48±0.10	7,729,408	0%
BNN-FL 1/1	96.53±0.25	**241,544**	0%
BNN-FL 1/32	97.27±0.08	**241,544**	0%
MPT-FL 1/1	97.47±0.22	**241,544**	**80%**
MPT-FL 1/32	**97.92±0.11**	**241,544**	**80%**

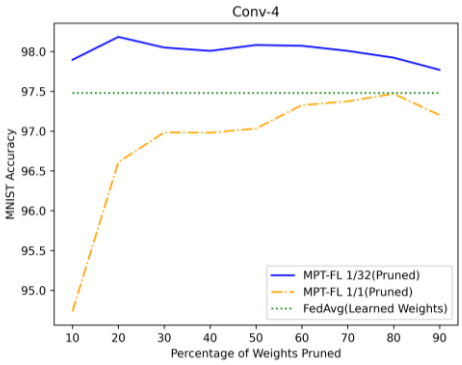

그림 1. 가지치기 비율에 따른 테스트 정확도 비교

3. 결론

본 논문은 연합학습의 모델을 간소화하기 위해 MPT 를 적용하였으며, 실험을 통해 효과를 확인하였다. 실험을 통해 MPT-FL 로 학습한 모델은 1 비트로 압축된 희소 신경망에서 FedAvg 보다 적은 연산으로, 비슷하거나 더 높은 정확도를 달성할 수 있음을 확인하였다. 이는 기존 연합학습 방법들이 복잡한 모델을 학습하면서 마주하는 제한점들을 해결하면서 연산 효율적이고 높은 정확도를 달성할 수 있게 하므로 활용도가 높을 것으로 기대된다.

ACKNOWLEDGMENT

본 연구는 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2021-0-02081, 5G-Advanced 네트워크 기계학습 기반 지능화,자동화 기술 표준개발)과 한국연구재단(No.2021R1F1A1063288)의 지원을 받아 수행되었음.

참 고 문 헌

[1] McMahan, Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." Artificial intelligence and statistics. PMLR, pp. 1273-1282, 2017.

[2] Courbariaux, Matthieu, et al. "Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1." arXiv preprint arXiv:1602.02830, 2016.

[3] Frankle, Jonathan, and Michael Carbin. "The lottery ticket hypothesis: Finding sparse, trainable neural networks." arXiv preprint arXiv:1803.03635, 2018.

[4] Diffenderfer, James, and Bhavya Kailkhura. "Multi-prize lottery ticket hypothesis: Finding accurate binary neural networks by pruning a randomly weighted network." arXiv preprint arXiv:2103.09377, 2021.