실시간 분산 학습 서비스를 위한 분산 아키텍처와 패킷 특성

이정섭, 진의환, 이광훈, 김광순
연세대학교

{jungseop.lee, jinian, kwanghoon.lee, ks.kim}@yonsei.ac.kr

Distributed Architecture and Packet Characteristics for Real-Time Distributed Learning Services

Lee Jung Seop, Jin Eui Whan, Lee Kwanghoon, and Kim Kwang Soon
Yonsei Univ.

요 약

본 논문은 다양한 분산 학습 서비스를 지원하기 위한 일반화된 분산 학습 아키텍처를 제시하고 분산 학습 상황에서 대표적인 이미지 분류 모델(AlexNet, VGGNet, ResNet)에서의 데이터 크기를 유도하였다. 이를 통해 분산 기계학습 상황에서 패킷 크기의 집합이 넓은 영역에 걸쳐 있으므로 이를 서비스 별로 분류하여 각각의 서비스 별로 물리계층에서 패킷 요구조건을 지원할 수 있는 시스템을 설계해야 할 필요성을 제시하였다.

Ⅰ. 서 론

현재 새롭게 주목받고 있는 기술로서 기계학습이 대두되고 있다. 이러한 기계학습 기술은 V2X (vehicle-to-everything), teleoperation, UAV (unmanned aerial vehicle)등에 이용되고 있다[1]. 하지만 기계학습 기술은 인풋 데이터의 크기가 증가하고 단말에서의 컴퓨팅 파워의 한계로 인하여 모든 연산을 단말에서 처리하기에는 어려움이 존재한다.

기존의 방식이 갖는 문제점은 다음과 같다[2]. 1) 단말에서의 컴퓨팅 한계로 인해 고도의 기계학습 연산을 수행하기 어려우며 2) 간단한 기계학습 모델이 단말에서 수행될 경우 고도화된 연산을 하기 어려우며 3) 입력 센싱데이터를 서버의 대형 신경망 모델로 전송하는 과정에서의 통신 비용 및 대기시간 문제와 개인정보 보호의 문제가 존재한다. 즉 정확한 연산과 방대한 양의 데이터를 처리하기 위해서는 신경망을 적절히 나누어 처리해야 한다[3,4] 이러한 구조의 장점은 대규모 기계학습 작업을 수행할 때 서버가 로컬 노드의 판단을 지원할 수 있으며 시스템 확장성이 높다. 또한 분산 학습은 데이터 보안 및 짧은 응답시간에서 장점을 갖는다. 따라서 이러한 문제를 해결하기 위해 기계학습을 분산적으로 사용하는 것이 필수적이다.

분산 기계 학습은 V2X, teleoperation, UAV 등 다양한 서비스에 응용될 수 있지만 이러한 서비스는 높은 QoS (quality of services) 특성을 갖고 있기 때문에 통신의 측면에서 고려되어야 할 점이 있다. 중간 레이어 데이터를 양자화 및 인코딩하는 과정이 필요하며 QoS 를 만족하기 위한 고신뢰 저지연 통신 기술이 고려되어야 한다[5].

하지만 이러한 기술적 문제에 앞서 분산 기계학습의 구조와 이러한 구조에서 패킷의 특성을 도출하는 것이 중요하다. 따라서 본 논문에서는 분산 기계학습을 지원하기 위한 고신뢰 저지연 통신기술의 선행연구로서 분산 기계학습의 일반화된 구조를 제시하고 이러한 구조에서의 패킷 특성을 도출한다.

Ⅱ. 본론

본 논문에서 다양한 논문들에서 제시하고 있는 분산 기계학습의 아키텍처를 일반화 할 수 있는 일반화된 구조를 제시한다.

분산 기계학습을 통한 다양한 서비스를 제공하기 위해 단말은 비디오, 음성, 센서정보들을 받아 이를 전처리(preprocessing) 과정을 거친다[6]. 이렇게 처리된 다중 인풋 데이터들은 local DNN(deep neural network)라 불리는 인공 신경망에서 처리된다[7]. 이렇게 처리된 데이터가 판단을 하기에 충분하지 않다면 단말은 이 데이터를 엣지 서버로 전송한다. 하지만 raw 데이터를 전부 전송하는 것이 아니라 distillation 같은 처리 과정이 필요하다[8]. 이렇게 처리된 데이터를 인코딩 한 후 엣지 서버로 전송하고 엣지 서버에서는 여러 노드들이 보내온 정보를 취합하여 global DNN 이라 불리는 네트워크를 통해 처리하고 이를 통해 단말의 학습 및 추론에 도움을 준다. 이러한 구조의 모습은 다음 그림과 같다.

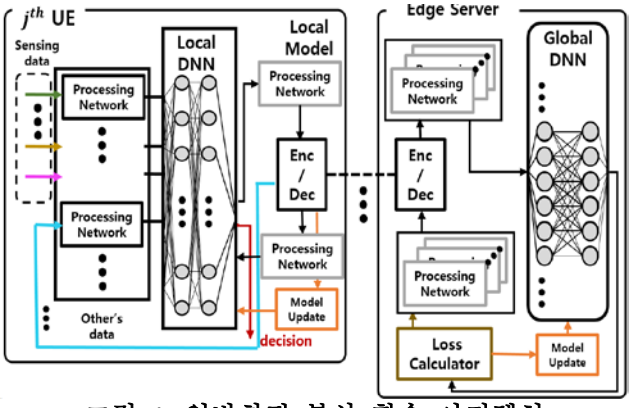

그림 1. 일반화된 분산 학습 아키텍처

다음으로 이러한 구조를 갖는 분산 기계학습에서 중간 레이어(intermediate layer)의 출력 값을 도출하기 위해 대표적인 이미지 분류의 예제로써 AlexNet, VGGNet, ResNet 과 같은 모델의 출력 값을 살펴 본다[9-11]. 3 가지 모델 중 224x224 이미지 인풋과 16 개의 레이어를 갖는 VGGNet 은 13 개의 컨볼루션 레이어와 5 개의 pooling 레이어, 3 개의 fully connected 레이어를 갖고 있다. 이때 마지막 pooling 레이어에서의 데이터는 대략 25K 정도 이고 일반적인 single precision 가정한다면 중간 레이어에서의 데이터 크기의 오더는 100Kbyte 정도로 유도해 볼 수 있다. 이러한 방법론을 통해 다른 모델에서의 데이터 량 유도 할 수 있다. 또한 인풋 이미지가 8K 혹은 그 이상 증가 하고 [5]과 같이 10% 정도의 압축률을 가정하면 다음과 같은 데이터 량을 유도해 볼 수 있다.

224x224 image	AlexNet	VGGNet	ResNet
	1~50K	1~100K	1~50K
State of the art (8K)	100K~10M (10% compression ratio)		
Future (10~100 times)	1M~1000M (10% compression ratio)		

표 1. 이미지 분류 모델의 패킷 특징

논문 [12]의 경우처럼 5G 에서는 고신뢰 저지연 특성을 갖는 패킷을 잘 분류하고 이를 지원할 수 있는 통신 기술이 필요하다. 분산 기계 학습 상황에서도 마찬가지로 넓은 영역에 패킷 특성이 분포하고 있으므로 이를 잘 분류하여 물리 계층에서 패킷 요구조건을 지원할 수 있는 통신 기술에 대한 연구가 필요하다.

Ⅲ. 결론

본 논문은 분산기계학습을 지원하기 위해 기존에 제시되고 있는 다양한 분산 기계학습 예제를 통해 일반화된 분산 기계학습 아키텍처를 제시하였다. 또한 분산 학습 상황에서 패킷 특징을 도출하기 위해 기존의 대표적인 이미지 분류 모델(AlexNet, VGGNet, ResNet)의 중간 레이어에서의 데이터 크기를 유도해 보고 몇가지 가정을 통해 더 큰 이미지 처리 상황에서의 데이터 크기를 유도해 보았다. 이러한 과정을 통해 분산 기계학습 상황에서의 데이터 크기는 넓은 영역에 분포해 있다. 따라서 다양한 분산 기계 학습의 서비스 별로 패킷의 특성을 분류하여 이를 지원하기 위한 물리 계층의 통신 기술이 필요함을 제시하였다.

ACKNOWLEDGMENT

이 논문은 2019 년도 정부(교육과학기술부)의 재원으로

한국연구재단의 지원을 받아 수행됨(No.2019R1A2C2007982)

참 고 문 헌

[1] A. I. Maqueda, *et al.*, “Event-based vision meets deep learning on steering prediction for selfdriving cars,” in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition,* Jun. 2018, pp. 5419- 5427.

[2] S. Teerapittayanon *et al.*, “Distributed deep neural networks over the cloud, the edge and end devices,” in *Proc. IEEE 37th Int. Conf. Distributed Computing Systems (ICDCS),* Jun. 2017, pp. 328- 339.

[3] K. Skala *et al.*, “Scalable distributed computing hierarchy: Cloud, fog and dew computing,” *Open Journal of Cloud Computing (OJCC),* no. 2(1), pp. 16- 24, 2015.

[4] Y. Kang *et al.*, “Neurosurgeon: Collaborative intelligence between the cloud and mobile edge,” *SIGARCH Comput. Archit. News*, vol. 45, no. 1, pp. 615-629, Apr. 2017.

[5] J. H. Ko *et al.*, “Edge-host Partitioning of Deep Neural Networks with Feature Space Encoding for Resource-Constrained Internet-of-Things Platforms.” *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Auckland, New Zealand, 2018, pp. 1-6

[6] O. Maksymiv *et al.*, "Deep convolutional network for detecting probable emergency situations," *2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP)*, Lviv, 2016, pp. 199-202.

[7] S. Teerapittayanon, B. McDanel, and H. T. Kung, “Distributed deep neural networks over the cloud, the edge and end devices,” in *Proc. IEEE 37th Int. Conf. Distributed Computing Systems (ICDCS)*, Jun. 2017, pp. 328- 339.

[8] J. Park *et al.*, "Wireless Network Intelligence at the Edge," in *Proceedings of the IEEE*, vol. 107, no. 11, pp. 2204-2239, Nov. 2019.

[9] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[10] Simonyan, Karen & Zisserman, Andrew. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 1409.1556.

[11] K. He, *et al.*, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2016.

[12] K. S. Kim et al., "Ultrareliable and Low-Latency Communication Techniques for Tactile Internet Services," in Proceedings of the IEEE, vol. 107, no. 2, pp. 376-393, Feb. 2019.