논문 2021-16-02

통신 실패에 강인한 분산 뉴럴 네트워크 분할 및 추론 정확도 개선 기법

(Communication Failure Resilient Improvement of Distributed Neural Network Partitioning and Inference Accuracy)

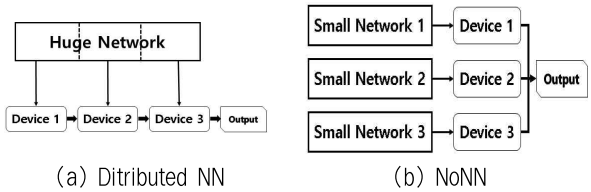
정 종 훈, 양 회 석*
(Jonghun Jeong, Hoeseok Yang)

Abstract : Recently, it is increasingly necessary to run high-end neural network applications with huge computation overhead on top of resource-constrained embedded systems, such as wearable devices. While the huge computational overhead can be alleviated by distributed neural networks running on multiple separate devices, existing distributed neural network techniques suffer from a large traffic between the devices; thus are very vulnerable to communication failures. These drawbacks make the distributed neural network techniques inapplicable to wearable devices, which are connected with each other through unstable and low data rate communication medium like human body communication. Therefore, in this paper, we propose a distributed neural network partitioning technique that is resilient to communication failures. Furthermore, we show that the proposed technique also improves the inference accuracy even in case of no communication failure, thanks to the improved network partitioning. We verify through comparative experiments with a real-life neural network application that the proposed technique outperforms the existing state-of-the-art distributed neural network technique in terms of accuracy and resiliency to communication failures.

Keywords : Internet-of-things device, Embedded system, Neural network, Distributed neural network, Neural network optimization

Ⅰ. 서 론

최근 다양한 사물인터넷 (Internet of Things) 장치의 확산과 함께 인체에 착용하는 다양한 웨어러블 디바이스가 개발되고 있다 [1]. 또한, 최근 인공지능 기술의 발전과 더불어 이러한 웨어러블 디바이스상에서도 뉴럴 네트워크 (Neural Network)기반 응용프로그램을 구동할 필요성이 대두되고 있다 [2]. 하지만 일반으로 웨어러블 디바이스에서 사용되는 ARM Cortex-M 시리즈의 임베디드 시스템은 가용 메모리가 약 500KB RAM과 1MB Flash 정도이므로 많은 자원을 필요로 하는 고성능 뉴럴 네트워크 구동이 제한된다. 이를 위해 최근 뉴럴 네트워크의 메모리와 연산량을 줄이는 다양한 최적화 연구 [3, 4]가 진행 중이지만 여전히 단일 디바이스만으로는 뉴럴 네트워크 구동이 제한되거나 낮은 정확도를 보인다. 하지만 통신 장치가 부착된 웨어러블 디바이스에서는 다수의 디바이스에 분산하여 병렬적으로 구동할 수 있다. 일반적으로 웨어러블 디바이스들은 블루투스나 Zigbee와 같은 근거리 통신 프로토콜을 사용하여 통신하나 군용



(a) Ditributed NN (b) NoNN

그림 1. 분산 뉴럴 네트워크 동작 개요도
Fig. 1. An Overview of Distributed Neural Network

제품이나 재난 환경 등의 특수한 상황에서는 사람의 몸을 매질로 하는 인체 통신을 사용하여 통신할 것으로 예측된다 [5]. 이러한 인체통신은 추가적인 기존 유무선 통신에 비하여 높은 편리성과 용이성을 가지나 낮은 데이터 전송률을 가지는 단점이 있다 [6]. 인체 통신과 같이 데이터 전송률이 낮고 불안정한 통신에 의존하는 경우, 웨어러블 디바이스에서 사용할 분산 뉴럴 네트워크 기법은 통신량이 적고 일부 디바이스에서 통신에 실패하여도 결과를 도출할 수 있도록 통신 실패에 강인해야 한다.

그림 1 (a)은 기존의 일반적인 분산 뉴럴 네트워크 [7]의 동작을 도식화한 것이다. 일반적인 뉴럴 네트워크는 단일 디바이스에서 구동하기 어렵기 때문에 레이어 단위로 여러 디바이스에 분산시킨다. 그러나 이러한 방식은 레이어 간 전달 데이터의 양이 커서 통신량과 통신 빈도가 높을 뿐 아

*Corresponding Author (hyang@ajou.ac.kr)
Received: Dec. 18, 2020, Revised: Feb. 6, 2021, Accepted: Feb. 10, 2021.
J.H. Jeong: Ajou University (M.S.)
H.S. Yang: Ajou University (Assoc. Prof.)
※ 본 연구는 방위사업청과 국방과학연구소가 지원하는 미래전투체계 네트워크기술 특화연구센터 사업의 일환으로 수행되었습니다 (UD190033ED).

니라, 하나의 추론 결과를 얻기 위해서 모든 디바이스의 결과가 필요하므로 하나의 디바이스에서라도 통신이 실패할 경우 결론 도출이 불가능하다는 단점이 있다.

이러한 단점을 보완하기 위해 그림 1 (b)와 같은 NoNN (Network of Neural Network) [8]이라는 분산 뉴럴 네트워크 기법이 제안되었다. NoNN의 작은 네트워크들은 모두 독립적인 네트워크이며, 이 독립적인 네트워크들 각각의 최종 결과들을 concatenate 하여 최종 결론을 도출한다. 또한, 작은 네트워크들은 각각의 독립적인 네트워크로 독자적인 결론 도출이 가능하지만, 상대적으로 별개 네트워크의 정확도는 많이 떨어져 통신 실패의 경우 추론 정확도 열화가 크다는 단점이 있다. 우리는 NoNN과 같이 병렬로 분산된 형태의 뉴럴 네트워크의 문제점 극복과 성능 개선을 목표로 한다. 따라서 본 논문에서는 1) 네트워크를 균일하게 분산시킬 수 있는 새로운 분산 기법과 2) 분산 뉴럴 네트워크에서 최적의 성능을 낼 방법에 대해 서술한다.

본 논문의 나머지 구성은 다음과 같다. 2장에서는 본 논문에 대한 배경지식을 제시하고 3장에서는 분산 뉴럴 네트워크 성능 개선 방안을 제시한다. 그리고 4장에서는 실험을 통해 성능을 비교 분석하고 5장에서 결론을 내린다.

Ⅱ. 배경지식

본 장에서는 뉴럴 네트워크 분산 뉴럴네트워크의 기법인 NoNN (Network of Neural Network)를 소개하고 이에 앞서 NoNN에서 사용하는 최적화 기법인 지식증류 (Knowledge Distilation) [8]에 대해 소개한다.

1. 지식증류 (Knowledge Distilation)

지식증류 (KD)은 메모리와 연산량이 큰 고성능의 네트워크 (Teacher Network)를 이용하여 상대적으로 작은 네트워크 (Student Network)를 훈련시키는 최적화 기법이다. KD를 이용하여 생성한 student 네트워크는 일반적인 훈련을 통해 훈련한 네트워크보다 뛰어난 성능을 보인다.

지식증류에서 teacher 네트워크가 student 네트워크를 훈련시키는 방식에는 여러 가지가 있지만 본 논문에서는 아래 2가지 기법을 사용한다.

1.1 Soft Label [9]

일반적인 뉴럴 네트워크의 훈련은 Dataset의 입력을 x, Label을 y, 네트워크의 추론 값을 P_s라고 할 때 수식 (1)과 같이 정의된 loss (Hard Label Loss)$L(x)$를 최소화하도록 진행된다.

$$L(x) = H(y, P_s). \tag{1}$$

이때, H는 cross entropy를 의미하며 이러한 방식으로 훈련된 네트워크는 입력 data간의 유사성을 무시하여 좋지 않은 성능을 보인다.

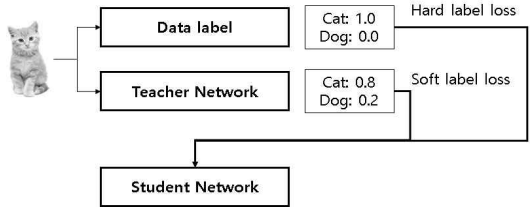


그림 2. Soft label을 이용한 KD 기법
Fig. 2. KD Technique Using Soft Label

Soft label은 그림 2와 같이 teacher 네트워크를 이용하여 student 네트워크를 훈련시키는 기법 중 한 가지로 정해진 teacher 네트워크의 softmax 이전 결과 (Soft Label)를 이용하여 훈련을 진행한다.

Teacher 네트워크의 softmax 이전 결과를 P_t^τ, student 네트워크의 softmax 이전 결과를 P_s^τ라고 할 때, Soft Label Loss $L^{SL}(x)$는 아래 수식 (2)와 같이 정의된다.

$$L^{SL}(x) = H(P_t^\tau, P_s^\tau). \tag{2}$$

일반적인 지식증류에서는 Hard Label Loss와 Soft Label Loss를 같이 고려한 L^{KD}를 수식 (3)과 같이 정의하여 훈련을 진행한다.

$$L^{KD}(x) = (1-\alpha)L(x) + \alpha L^{SL}(x). \tag{3}$$

Hard Label Loss와 Soft Label Loss에 대한 가중치를 α를 이용하여 조절할 수 있다.

기존 Hard Label만 이용한 훈련과 달리 Soft Label을 고려한 훈련은 입력data 간의 유사성을 해치지 않으며 teacher 네트워크의 지식을 전수받아 일반적인 훈련으로 훈련된 모델보다 더 좋은 성능을 보인다.

1.2 Attention Transfer [10]

Attention Transfer (AT)는 teacher 네트워크와 student 네트워크의 마지막 convolution layer (fconv)의 output feature map을 비교하여 훈련하는 방식이다. AT의 loss L^{AT}는 수식 (4)와 같이 정의된다.

$$L^{AT}(x) = \left\| \frac{Q_s(x)}{\| Q_s(x) \|_2} - \frac{Q_t(x)}{\| Q_t(x) \|_2} \right\|_2. \tag{4}$$

수식에서 Q_s와 Q_t는 각각 student 네트워크와 teacher 네트워크의 fconv의 output feature map이다. 단, AT를 적용하기 위해서는 Q_s와 Q_t의 크기가 동일해야 한다. AT방식은 Soft Label 방식과 다르게 teacher 네트워크의 중간결과를 이용하여 훈련을 진행하여보다 수월하게 teacher 네트워크의 지식을 훈련할 수 있다는 장점을 가진다.

2. NoNN (Network of Neural Network)

NoNN은 자원사용량이 많은 고성능의 teacher 네트워크

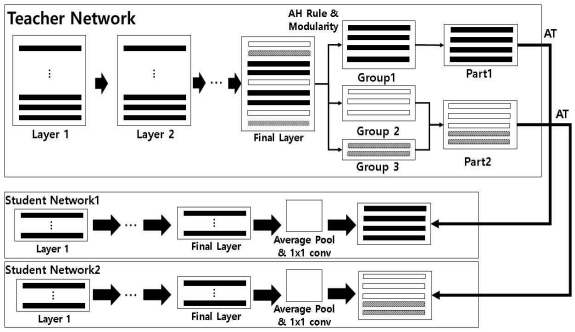


그림 3. NoNN 생성 개요도
Fig. 3. An Overview of NoNN Generation

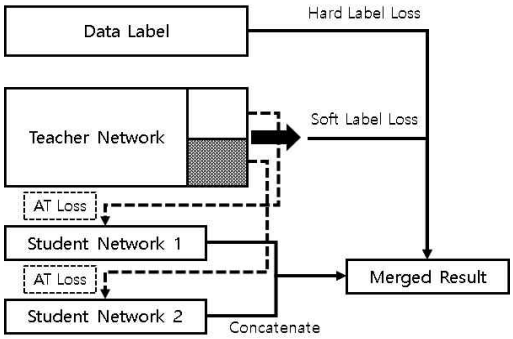


그림 4. NoNN 훈련 과정 개요도
Fig. 4. An Overview of NoNN's Training

를 다수의 자원제약 임베디드 시스템에서 병렬 분산처리하도록 하는 경량화 기법이다.

NoNN의 생성은 그림 3과 같이 teacher 네트워크를 병렬 분산처리하기 위해 fconv를 분산할 디바이스의 수만큼 분할한 뒤, 하나의 분할된 part가 하나의 student를 Attention Transfer로 훈련한다.

NoNN의 훈련은 각각의 student 네트워크가 매칭되는 part와 유사한 동작을 하는 동시에 concatenate 한 최종 결과가 Dataset Label에 맞도록 훈련을 진행해야 한다. 따라서 NoNN의 훈련에서는 위에서 언급한 Hard Label Loss, Soft Label Loss, AT Loss가 반영되어야 하므로 NoNN의 loss (L^{NoNN})는 수식 (5)로 정의된다.

$$L^{NoNN}(x) = L^{KD}(x) + \beta \sum_{i=1}^{P} L_i^{AT}(x). \quad (5)$$

수식 (5)에서 β는 AT Loss를 조절하는 가중치, P는 생성한 student 네트워크의 수, $L_i^{AT}(x)$는 teacher 네트워크의 i번째 part와 매칭되는 student 네트워크를 비교한 AT Loss의 값을 의미한다. NoNN의 훈련과정을 도식화하면 그림 4와 같은 방식으로 훈련이 진행된다.

NoNN에서는 모든 student에 지식을 고르게 분할하기 위해 Activation Hub (AH) rule을 제안하고 이를 이용하여 teacher 네트워크를 분할한다.

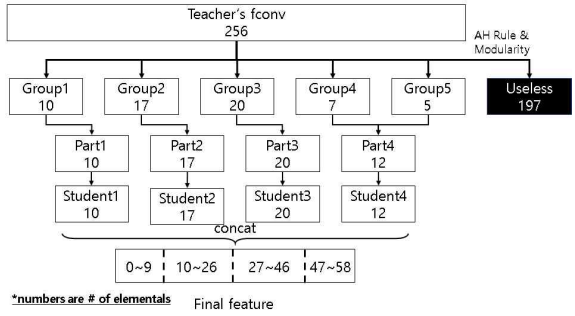


그림 5. NoNN 생성 예시
Fig. 5. An example of NoNN generation

AH rule이란 유사한 역할을 하는 teacher 네트워크 fconv의 필터들을 강제적으로 다른 part로 나누기 위한 규칙이다. i번째 필터의 출력평균을 a_i, j번째 필터의 출력평균을 a_j라고 할 때 AH 규칙은 수식 (6)로 표현된다.

$$AH(i,j) = \sum a_i a_j |a_i - a_j|. \quad (6)$$

따라서 AH rule에 의해 동일한 입력에 대해 비슷한 값을 출력하는 필터들은 작은 AH 값을 가지게 되고 동일한 입력에 대해 다른 값을 출력하는 필터들은 높은 AH값을 가지게 된다.

분산 네트워크의 효율적인 분할을 위하여 각 필터들을 정점 (Vertex)으로 하고, 그 필터 간의 간선 (Edge)들의 가중치를 AH값으로 가지는 그래프 형태의 자료구조를 생성할 수 있다. 이를 소셜 네트워크 등을 분석하는 네트워크 사이언스에서 활용되는 분할 알고리즘 [11]을 활용하여 별개의 그룹으로 분리할 수 있는데, 이를 Community Structure라고 한다. NoNN에서는 이 분할 알고리즘에 사용자 설정 resolution 변수 γ를 추가하여 teacher 네트워크를 분할하는 동시에 결과에 큰 영향을 끼치지 않은 불필요한 부분을 제거하도록 수정하였다. 이렇게 나누어진 필터들의 집합체를 group이라고 하며 네트워크 모델마다 group의 수와 각 group의 크기가 다르므로 student 네트워크와 수를 맞추기 위해 그림 5와 같이 group들을 합쳐 새로운 집합체인 part를 생성하며 이때 part는 모두 비슷한 수의 필터로 구성되도록 생성한다.

이렇게 생성된 분산 뉴럴 네트워크는 통신량이 적고 일부 디바이스에서 통신에 실패하여도 추론이 가능하지만 통신 실패에 따른 추론 정확도 열화가 크다는 단점을 가진다. 본 논문에서는 분산 뉴럴 네트워크 성능 개선을 위한 방안을 제시한다.

Ⅲ. 분산 뉴럴 네트워크 성능 최적화

본 장에서는 분산 뉴럴 네트워크의 성능 최적화를 위해 통신량 조절이 가능하고 통신 실패에 강인한 새로운 분할기법을 제안한다.

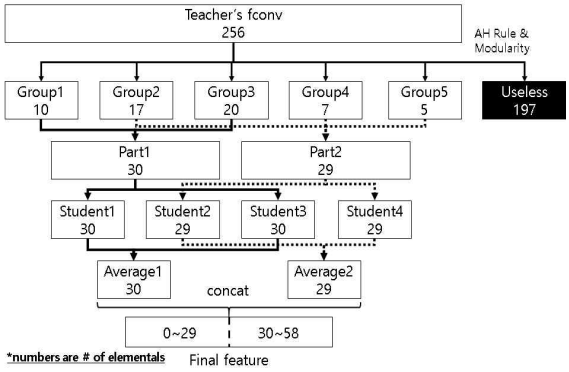


그림 6. Average 기법을 이용한 분산 네트워크 생성 예시
Fig. 6. Generation Example of Distributed Network Using Average Technique

1. Average 기법 [12]

기존 NoNN에서는 통신 실패에 강인하도록 AH rule [8]을 이용한 분할기법을 사용하여 student 네트워크를 생성하였다. 하지만 해당 논문에서 일부 디바이스에서 통신 실패를 가정한 실험을 진행한 결과 추론 성능 열화가 심한 것을 확인할 수 있다. 이를 개선하기 위해 그림 6과 같이 하나의 part에서 다수의 student를 생성하고 추론 시 동일한 part에서 파생된 student끼리 평균을 구해 결과를 도출하는 Average 기법을 사용하였다. Average 기법은 part 간 중복성을 강화하는 기법으로 동일한 part로부터 파생된 student 중 하나라도 통신에 성공하면 손실되는 데이터가 매우 적음으로 통신 실패에 강인한 분산 뉴럴 네트워크 생성이 가능하다.

2. Activation-based Partition

Average 기법을 이용하여 NoNN의 통신 실패로 인한 추론 성능 열화를 일부 개선하였지만, 여전히 AH 기법을 이용한 분할 기법은 group의 크기와 개수가 모델마다 다르므로 part의 크기를 조절하기 힘들다. 이러한 문제점은 part간의 불균형으로 이어져 특정 디바이스가 통신에 실패할 경우 추론 성능이 크게 감소할 수 있다. 본 논문에서는 이를 개선하기 위해 AH rule을 이용한 분할 기법이 아닌 teacher 네트워크 fconv의 output feature map (Activation)을 이용한 분할 기법을 제안한다.

제안하는 기법은 우선, teacher 네트워크에 모든 train dataset (단, 모든 label에 대한 입력데이터의 수는 균일해야 한다)을 입력으로 넣은 뒤 평균 activation값을 구한다. 그 후, activation을 내림차순으로 sorting한 배열을 S, 생성할 part의 수를 p, 각 part의 크기를 n이라고 할 때 i번째 part P_i는 수식 (7)과 같이 activation 값에 따라 교차하여 생성한다.

$$P_i = \sum_{j=0}^{n-1} S_{p \times j + i} \,. \qquad (7)$$

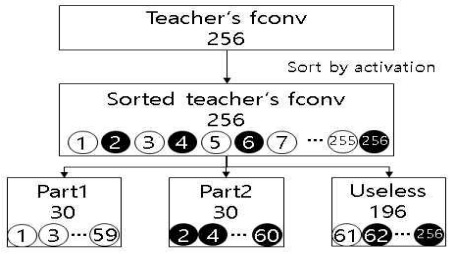


그림 7. Activation-based Partition 예시
Fig. 7. An example of Activtion Partition

그림 7은 part를 2개, part의 크기를 30으로 설정하였을 때의 생성 예시이다. 그림 7과 같이 평균 activation이 61번째로 큰 필터부터는 결과에 큰 영향을 미치지 않는다고 판단하여 모두 useless로 처리한다.

제안하는 Activation-based Partition 기법은 fconv에서 반응성이 활발한 필터들을 이용하고 반응성이 떨어지는 필터들을 useless로 처리하기 때문에 teacher 네트워크의 지식을 거의 손실하지 않으며 part의 수와 part의 크기를 자유롭게 조절할 수 있다는 장점을 가진다.

Ⅳ. 실 험

본 장에서는 분산 뉴럴 네트워크의 성능 최적화를 위해 제안하는 기법의 개선을 실험을 통해 검증하며 그 외의 성능 개선을 위한 실험을 진행한다.

1. 실험환경

실험은 Cifar-10과 Cifar-100 데이터 셋을 학습한 이미지 분류 어플리케이션을 이용하였다. 모든 실험은 8개의 student 네트워크를 생성하여 실험을 진행하였다.

2. Average 기법 성능 검증

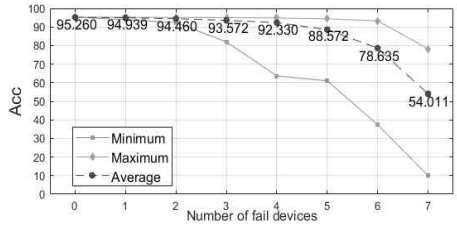
Average 기법이 실제로 통신 실패에 강인한지 성능 검증을 위해 통신 실패를 가정하여 일부 student의 출력값을 0으로 설정하는 실험을 구성하였다. Teacher와 student 네트워크는 표 1, 2와 같이 NoNN논문과 동일한 Wide Residual Network (WRN) [13] 네트워크를 사용하여 실험을 진행하

표 1. WRN Teacher 네트워크 모델
Table 1. Teacher Network Model

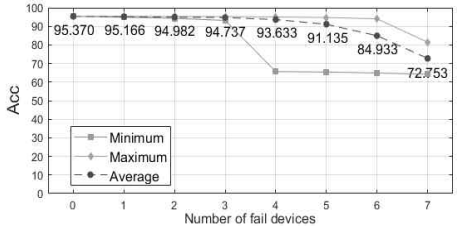
Dataset	Network	# of parameter	Accuracy
Cifar 10	WRN40-4	8.9M	95.82%
Cifar 100	WRN28-10	36.5M	80.88%

표 2. WRN Student 네트워크 모델
Table 2. Student Network Model

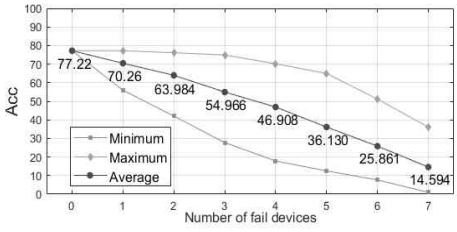
Dataset	Network	Teacher	# of parameter (each student)
Cifar 10	WRNbased [8]	WRN40-4	0.43M
Cifar 100	WRNbased [8]	WRN28-10	0.82M



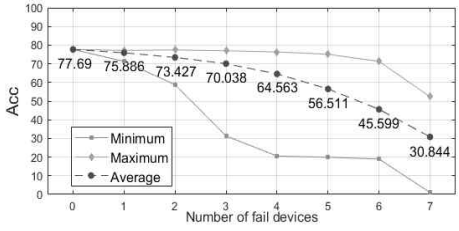
a. NoNN (Cifar-10)



b. Average (Cifar-10)



c. NoNN (Cifar-100)



d. Average (Cifar-100)

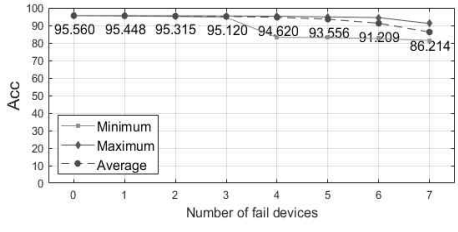
그림 8. NoNN과 Average 기법 통신 실패 실험 결과
Fig. 8. Communication Error Test Result of NoNN and Average Method

였다. 이때, Average 기법을 이용한 네트워크는 2개의 part를 생성하여 각 part 당 4개의 student를 생성하도록 하였고 어떤 디바이스에서 통신에 실패하는지에 따라 정확도의 차이가 크기 때문에 8개의 student에서 일어날 수 있는 모든 통신 실패 경우의 수에 대해 실험을 진행하여 평균 정확도, 최선의 경우 정확도, 최악의 경우 정확도를 비교하였다.

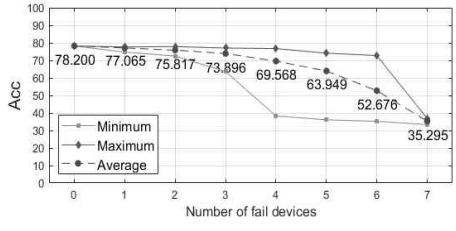
실험 결과, 그림 8과 같이 Cifar-10에서 통신 실패가 3개 발생하는 경우 정확도가 평균 1.164% 개선을 보이며 Cifar-100에서는 통신 실패가 3개 발생하는 경우 정확도가 평균 15.072% 개선을 보여 Average 기법이 통신 실패에 강인한 것을 확인할 수 있었다.

3. Activation-based Partition 기법 성능 검증

제안하는 Activation-based partition 기법의 성능검증을 위해 위의 실험과 동일한 방식의 실험을 진행하였으며 Activation-based partition으로 생성한 part의 개수는 2개



a. Activation-based Partition & Average (Cifar-10)



b. Activation-based Partition & Average (Cifar-100)

그림 9. Activation-based Partition 기법 통신 실패 실험 결과
Fig. 9. Communication Error Test Result of Activation-based Partition Method

크기는 30으로 설정하였고 Average 기법을 적용하여 각 part 당 4개의 student를 생성하도록 하였다. 실험 결과 그림 9와 같이 기존 NoNN이나 Average 기법만 이용한 결과보다 전반적으로 좋은 성능을 보이며 특히, 통신 실패가 7개 발생하는 경우 Average만 사용하는 경우보다 정확도가 Cifar-10에서는 평균 13.457%, Cifar-100에서는 4.456% 개선을 확인할 수 있다.

4. 네트워크 모델

NoNN논문에서 실험한 Resnet 계열인 WRN 네트워크 이외에도 일반적으로 저성능 하드웨어에서 사용되는 VGG (Visual Geometry Group)네트워크 [14]를 이용한 분산 뉴럴 네트워크를 생성하고 성능을 비교하였다.

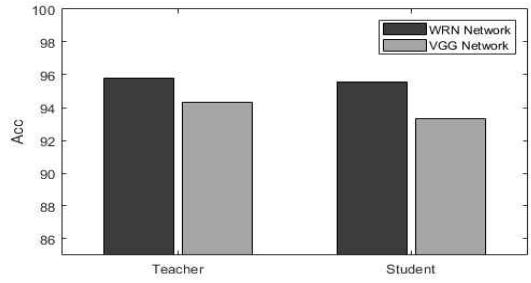
WRN 네트워크의 teacher와 student 네트워크는 표 1, 2와 동일하게 진행하였고 VGG 네트워크의 teacher와 student 네트워크는 표 3, 4와 같이 진행하였다. 네트워크의 분산은 제안하는 Activation-based Partition & Average를 이용하였으며 모두 30의 크기를 가지는 part 2개로 8개의 student를 생성하였다.

표 3. VGG Teacher 네트워크 모델
Table 3. VGG Teacher Network Model

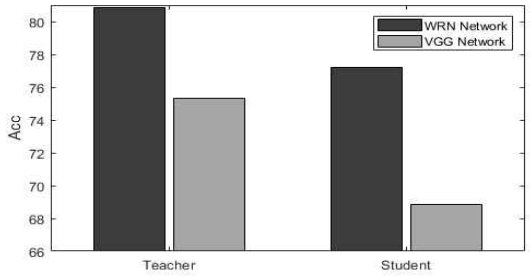
Dataset	Network	# of parameter	Accuracy
Cifar 10	VGG13	9.42M	94.33%
Cifar 100	VGG19	20.1M	75.34%

표 4. VGG Student 네트워크 모델
Table 4. VGG Student Network Model

Dataset	Network	Teacher	# of parameter (each student)
Cifar 10	VGG7	VGG13	1.1M
Cifar 100	VGG7	VGG19	1.1M



a. Cifar-10



b. Cifar-100

그림 10. WRN과 VGG모델 비교 실험 결과
Fig. 10. Accuracy Comparison of WRN and VGG

표 5. WRN Teacher 네트워크 모델 비교
Table 5. Comparison of WRN Teacher Networks

Dataset	Network	# of parameter	Accuracy
Cifar 10	WRN40-2	2.2M	95.17%
	WRN40-4	8.9M	95.82%
	WRN22-8	17.2M	96.28%
Cifar 100	WRN40-4	8.9M	78.65%
	WRN28-10	36.5M	80.88%
	WRN22-8	17.2M	80.52%

표 6. Cifar-10 실험 결과
Table 6. Experiment Result of Cifar-10

Network	Teacher	# of parameter (each student)	Accuracy
WRNbased	WRN40-2	0.43M	95.48%
WRNbased	WRN40-4	0.43M	95.56%
WRNbased	WRN22-8	0.43M	94.84%

표 7. Cifar-100 실험 결과
Table 7. Experiment Result of Cifar-100

Network	Teacher	# of parameter (each student)	Accuracy
WRNbased	WRN40-4	0.82M	78.72%
WRNbased	WRN22-8	0.82M	78.75%
WRNbased	WRN28-10	0.82M	78.20%

실험 결과, 그림 10과 같이 VGG네트워크가 더 큰 student 네트워크를 사용하지만, 네트워크를 분산할 경우 큰 성능 저하를 보여 제안하는 분산 뉴럴 네트워크에는 VGG 계열 네트워크보다 Resnet계열 네트워크가 더 효과적인 것을 확인하였다.

5. Teacher 네트워크에 따른 성능 변화

NoNN방식에서 student 네트워크의 크기는 임베디드 디바이스에서 동작 가능한 크기로 제한이 되어있지만, teacher 네트워크에 대한 제한은 없다. 하지만 teacher 네트워크의 크기도 분산 뉴럴 네트워크의 성능에 영향을 미치기 때문에 이를 확인하기 위한 실험을 진행하였다. 실험은 표 5와 같이 크기가 다른 teacher 네트워크를 Dataset별로 3개씩 사용하였고 student 네트워크는 표 2의 student 네트워크와 동일하게 사용하였다. 실험에 사용하는 네트워크는 모두 제안하는 Activation & Average 기법을 이용하여 30의 크기를 가진 2개의 part로 8개의 student를 생성하였다. 표 6, 7은 각각 Cifar-10, 100 dataset에 대한 실험 결과로 동일한 student 네트워크를 사용했지만 너무 작지도 크지도 않은 크기의 teacher를 이용하여 학습한 분산 뉴럴 네트워크의 성능이 더 좋은 것을 보여 teacher 네트워크의 선정이 분산 뉴럴 네트워크의 성능에 큰 영향을 끼치는 것을 확인하였다.

V. 결 론

임베디드 시스템을 위한 분산 뉴럴 네트워크는 통신량이 적고 통신 실패에 강인해야 한다. 하지만 기존에 연구된 NoNN 형태의 분산 뉴럴 네트워크는 통신량은 적지만 통신 실패의 경우 추론 성능 열화가 크다. 따라서 본 논문에서는 통신 실패에 강인한 Activation&Average 네트워크 분할 기법을 제안하고 실험으로써 제안하는 기법이 통신 실패에 강인함을 검증하였다. 또한 다양한 실험을 통해 분산 뉴럴 네트워크의 성능 최적화 방안에 대해 연구하였다.

제안하는 분산 뉴럴 네트워크를 실제로 ARM Cortex M 시리즈를 사용하고 320KB RAM, 1MB Flash를 가지는 NUCLEO F746ZG 보드 2개를 이용하여 구동시킨 결과 정상적으로 동작함을 확인할 수 있었지만, 연산에 약 6초, 통신에 약 0.68초가 소요되었다. 이는 연산과 통신에 사용되는 자원관리에 대한 구현이 미흡하여 수행 시간이 오래 걸리는 것으로 분석된다. 따라서 추후 연구로는 이를 개선하고 수행 시간 최적화 연구를 진행한다면 더욱 효과적으로 임베디드 시스템에서 뉴럴 네트워크를 구동할 수 있을 것으로 기대된다.

References

[1] S.P. Heo, D.H. Noh, C.B. Bae, D.S. Kim.. "Trend of IoT-based Healthcare Service" IEMEK Journal of Embedded Systems and Applications Vol. 10, No. 4, pp. 221-230, 2015 (in Korean).

[2] J. Yuan, S. Yu "Privacy Preserving Back-propagation Neural Network Learning Made Practical with Cloud Computing." IEEE Transactions on Parallel and Distributed Systems Vol. 25, No. 1, pp. 212-221, 2013.

[3] L. Lai, N. Suda, V. Chandra, "Cmsis-nn: Efficient Neural Network Kernels for Arm Cortex-m Cpus." arXiv preprint arXiv:1801.06601, 2018.
[4] J.H. Jeong, D.S. Lee, H.S. Jung, H.S. Yang. "Automatic Convolution Neural Network Model Compression Framework for Resource-Constrained Embedded Systems.",Journal of Korean Institute of Information Scientists and Engineers, Vol. 47, No. 2, pp. 136-146, 2020 (in Korean).
[5] S. Kim, J.G. Ko, "IB-MAC: Transmission Latency-aware MAC for Electro-magnetic Intra-body Communications." Sensors Vol. 19, No. 2, pp. 341, 2019.
[6] S.W. Kang, H.I. Park, K.H. Park. "Trends of Human Body Communications." [ETRI] Electronics and Telecommunications Trends Vol. 28, No. 2, pp. 70-76, 2013 (in Korean).
[7] J. Mao, X. Chen, K.W. Nixon, C. Krieger, Y. Chen, "Modnn: Local Distributed Mobile Computing System for Deep Neural Network." Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017. IEEE, 2017.
[8] K. Bhardwaj, C.Y. Lin, A. Sartor, R. Marculescu, "Memory-and Communication-aware Model Compression for Distributed Deep Learning Inference on iot." ACM Transactions on Embedded Computing Systems (TECS) Vol. 18, No. 5s, pp. 1-22, 2019.
[9] G. Hinton, O. Vinyals, J. Dean, "Distilling the Knowledge in a Neural Network." arXiv preprint arXiv:1503.02531 ,2015.
[10] N. Komodakis, S. Zagoruyko, "Paying more Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer.", 2017.
[11] M.E.J. Newman, "Modularity and Community Structure in Networks." Proceedings of the national academy of sciences Vol. 103, No. 23, pp. 8577-8582, 2006.
[12] J.H. Jeong, D.S. Lee, H.S. Yang, "Optimization of Distributed Neural Network on Wearable Devices with Low Data Rate Communication", Proceedings of Korean Institute of Communications and Information Sciences Summer Conference, pp. 661-662, 2020 (in korean).
[13] S. Zagoruyko, N. Komodakis, "Wide Residual Networks." arXiv preprint arXiv:1605.07146, 2016.
[14] K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-scale Image Recognition." arXiv preprint arXiv:1409.1556, 2014.

Jonghun Jeong (정 종 훈)

2020 Electrical and Computer Engineering from Ajou University (B.S.)
2020~Artificial Intelligence Convergence Network from Ajou University (M.S.)

Field of Interests: Deep Learning for Embedded & Neural Network Optimization & Distributed Neural Network
Email: kkjjh223@ajou.ac.kr

Hoeseok Yang (양 회 석)

2003 Computer Science and Engineering from Seoul National University (B.S.)
2010 Computer Science and Engineering from Seoul National University (Ph.D.)
2014~Department of Electrical and Computer Engineering, Ajou University (Associate Professor)

Career:
2010~2014 Post-Doctoral Researcher, D-ITET, ETH Zürich
2016~ Editorial Committee, IeMeK

Field of Interests: HW/SW Codesign & Deep Learning for Embedded System & Non Volatile Memory and Embedded System Design
Email: hyang@ajou.ac.kr