Distributed neural networks for Internet of Things: the Big-Little approach

E. De Coninck *et al.*,
"Distributed neural networks for Internet of Things: the Big-Little approach,"
in *2nd EAI International Conference on Software Defined Wireless Networks and Cognitive Technologies for IoT*, 2015

Introduction

- IoT는 모든 종류의 장치들을 인터넷으로 서로 연결해 상호작용하는 인기 있는 패러다임

- IoT를 위한 어플리케이션을 만들기 위해서는 모든 연결된 장치로부터 오는 많은 데이터들이 의미 있는 정보로 처리되고 분석되어야 함

- 현재 막대한 연산 능력과 확장성을 가진 클라우드 컴퓨팅이 막대한 데이터 처리를 위해 이용되고 있음

Introduction

- 클라우드 컴퓨팅은 네트워크 트래픽과 레이턴시, 프라이버시 문제로 인해 무조건적인 해법이 될 수 없음

- 따라서 로컬 영역에서 데이터를 처리하는 방법이 필요

Introduction

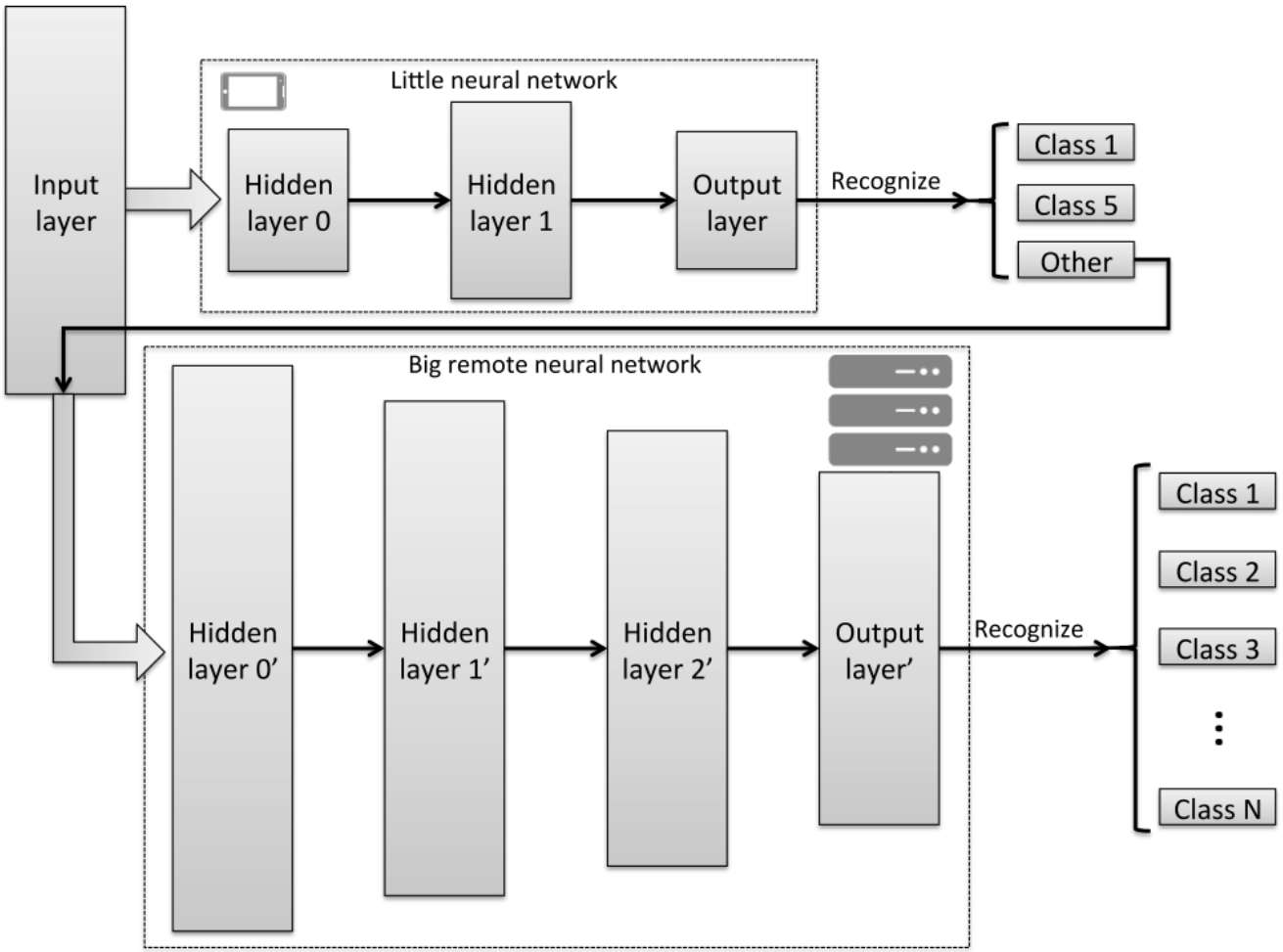
- IoT 응용에서 중요한 처리는
  시스템의 현재 상태를 판단하고 후속 동작을 추측하는 것이다.

- 이 분야에서 가장 유망한 기술은 DNN

- 논문은 IoT의 제한된 환경에서 신경망 추론을 하기 위한
  Big-Little 아키텍처를 제안

- 특정 분류 문제에 대해 대규모의 신경망과
  클래스의 일부만을 분류하는 작은 신경망을 함께 사용

Big-Little Neural Network Architecture



Big-Little Neural Network Architecture

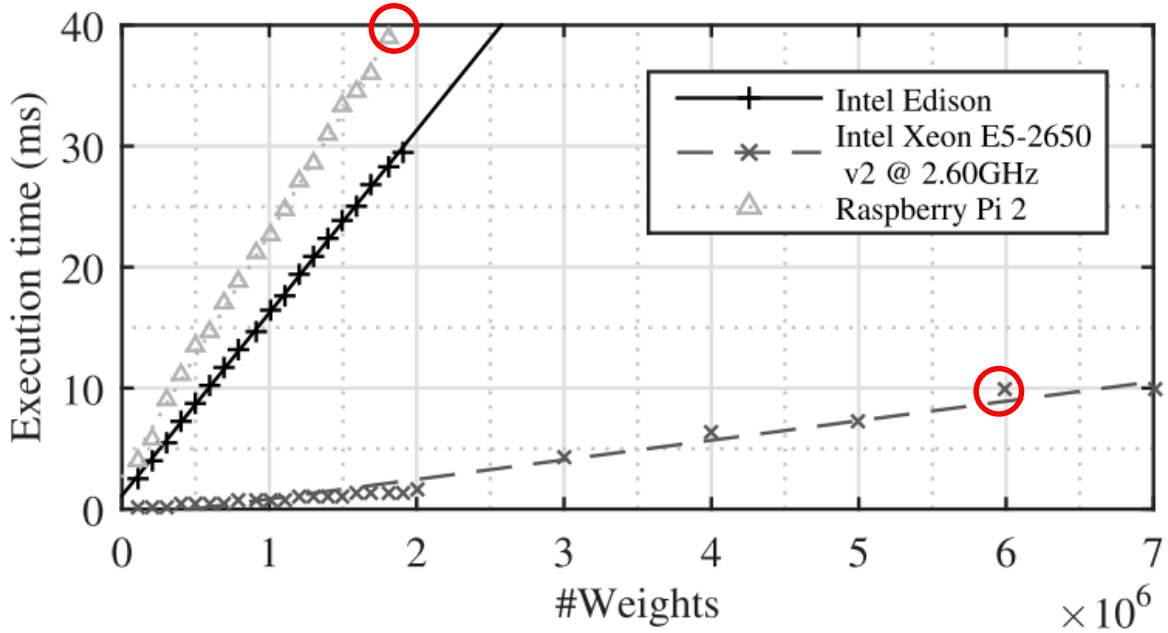
- **Big-Little 신경망 구조**
  - 작은 신경망은 출력 클래스의 일부만을 분류
  - 작은 신경망이 입력 데이터를 분류할 수 없다면
    클라우드에서 큰 신경망을 통해 분류 작업을 계속 진행함

- **출력 클래스의 수를 제한함으로써**
  원하는 분류 정확도를 유지하면서 신경망의 크기를 작게 유지할 수 있음

- **작은 신경망을 가진 장치는 우선순위가 높은 클래스만을 분류하여**
  중대한 상황에서 매우 빠른 속도로 응답할 수 있음

Evaluation

name	architecture	CPU	RAM
Raspberry Pi 2	ARM	Cortex-A7 (quad-core @ 900 MHz)	1 GB
Intel Edison	x86	Intel Atom (dual-core @ 500 MHz)	1 GB
Generic server	x86	2x Intel Xeon E5-2650v2 (8-core @ 2.60 GHz)	48 GB



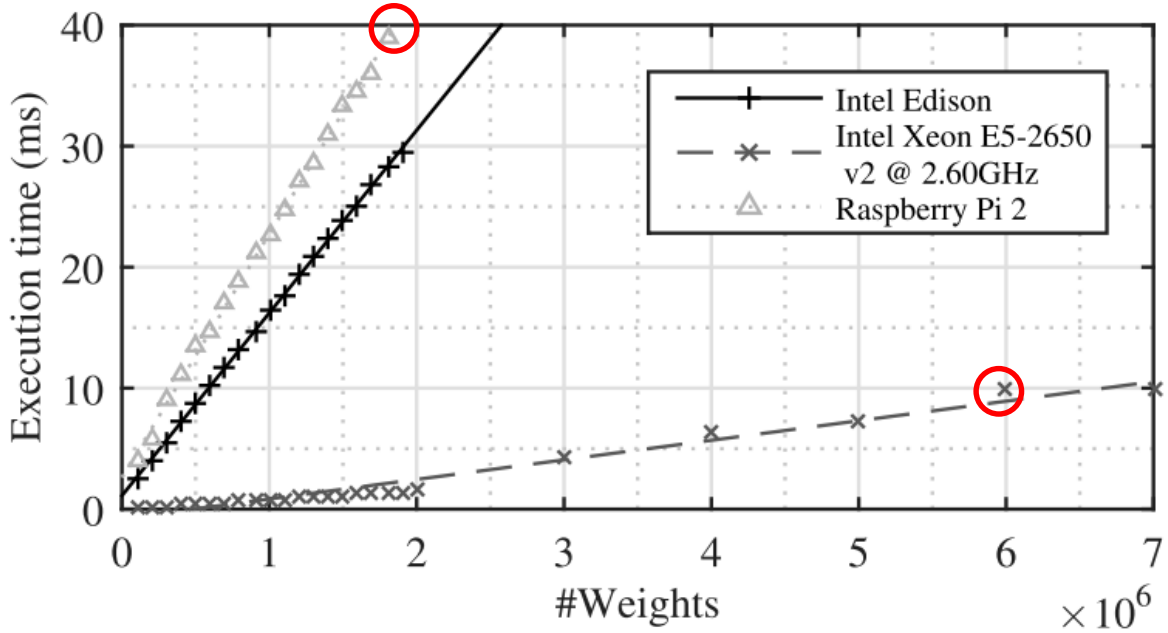
가정
최대 응답시간: 40ms
단일 링크 레이턴시: 15ms

에지 장치인 RPI에 제공된 연산 시간은 **40ms**

서버에게 주어진 연산 시간은 **10ms**

Evaluation

name	architecture	CPU	RAM
Raspberry Pi 2	ARM	Cortex-A7 (quad-core @ 900 MHz)	1 GB
Intel Edison	x86	Intel Atom (dual-core @ 500 MHz)	1 GB
Generic server	x86	2x Intel Xeon E5-2650v2 (8-core @ 2.60 GHz)	48 GB



가정
최대 응답시간: 40ms
단일 링크 레이턴시: 15ms

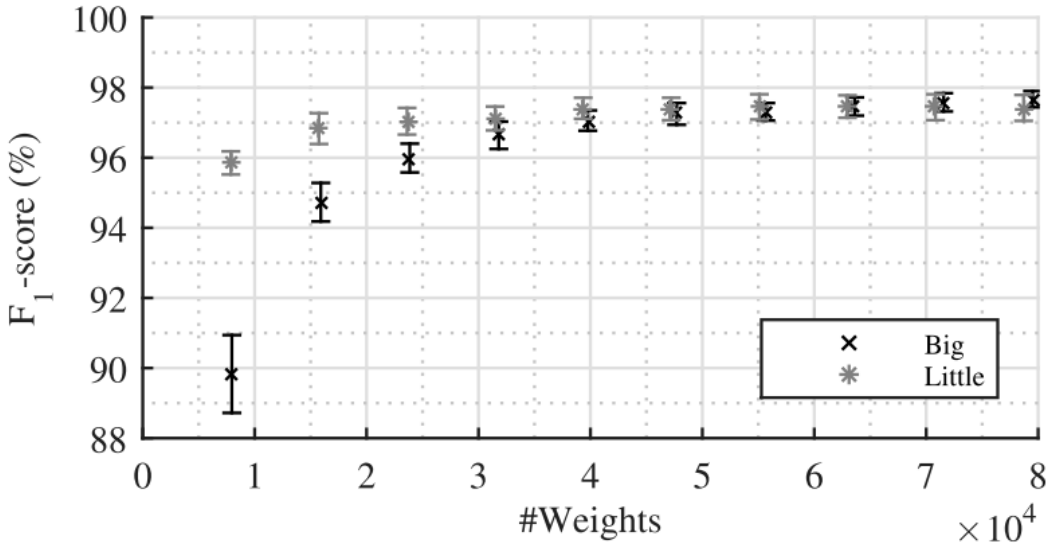
RPI는 $2x10^6$ 만큼의 weights를 사용할 수 있음

서버는 $6x10^6$ 만큼의 weights를 사용할 수 있음

빠른 응답을 위해서는 실제 연산 시간보다 레이턴시가 더 중요하다.

Evaluation

- 분류 클래스의 수가 다른 신경망 간의 정확도를 비교하기 때문에 F_1-score를 척도로 사용



- 작은 신경망은 레이어가 적지만
  출력 클래스 또한 적기 때문에 높은 성능을 보여줌

Evaluation

ID	architecture (number of neurons in each layer)	weights	process time for one sample [ms]
1	784, 1 000, 500, 2	1 250 502	0.98
2	784, 1 000, 500, 10	1 254 510	1.00
3	784, 2 500, 2 000, 1 500, 1 000, 500, 10 [5]	11 972 510	16.42

서버에서 실행한 결과

ID	priority class	test error for best validation [%]	priority class [%] recall	specificity	precision	F_1-score
1	1	NA	98.94	99.92	99.38	99.16
	8	NA	96.41	99.76	97.71	97.05
3	1	2.11	99.12	99.90	99.21	99.16
	8	2.11	96.82	99.70	97.22	97.02

1번 모델을 RPI에서 실행시키면 한 샘플을 처리하는데 **30ms**가 걸림

이와 비슷한 성능의 다중 클래스 모델을 서버에서 처리할 경우
16.42 + 15 x 2 = **46.42ms**가 걸림

Conclusion

- Big-Little 신경망 구조를 이용해

  전체 네트워크의 응답시간을 줄이면서

  성능도 동시에 확보함

- 신경망을 실행하는 것은 비용이 많이 들지 않을 수 있다

- 노드간 레이턴시에 대해 더욱 고려해야 함