Edge-Host Partitioning of Deep Neural Networks with Feature Space Encoding for Resource-Constrained Internet-of-Things Platforms

J. H. Ko, T. Na, M. F. Amir and S. Mukhopadhyay

2018 15th IEEE International Conference
on Advanced Video and Signal Based Surveillance (AVSS),
Auckland, New Zealand, 2018

Abstract

- 엣지와 호스트 사이에서 DNN 추론을 분할
- DNN의 중간 계층 피처를 인코딩하고

  그 것을 호스트에 전송
- 무손실, 손실 인코딩을 통해 엣지 플랫폼의 성능과 에너지 효율을 상승

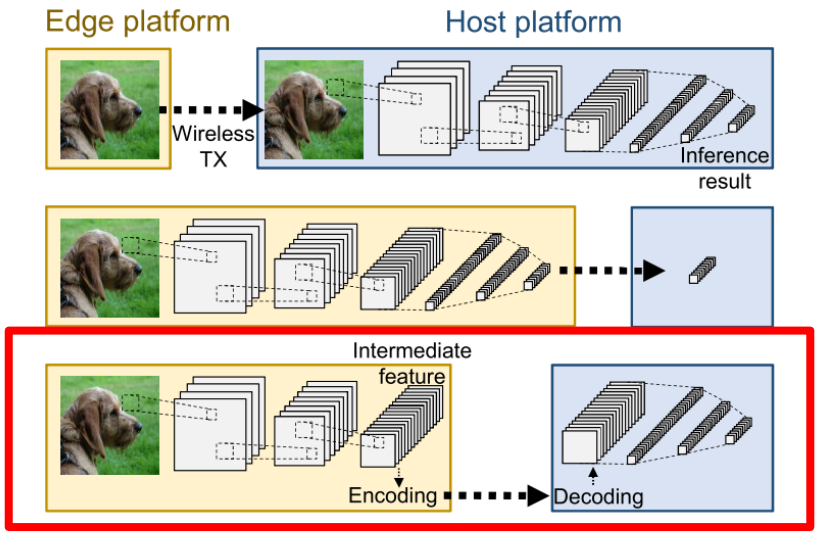
Introduction

- IoT 장치가 널리 사용되면서
  고품질 센서를 갖춘 수많은 엣지 장치에서 클라우드로 데이터를 전송
- 이미지 센서와 경량 프로세서로 구성된 엣지 장치가 널리 사용
- 이미지 데이터를 처리하는데 DNN이 강력한 성능을 보임

Introduction

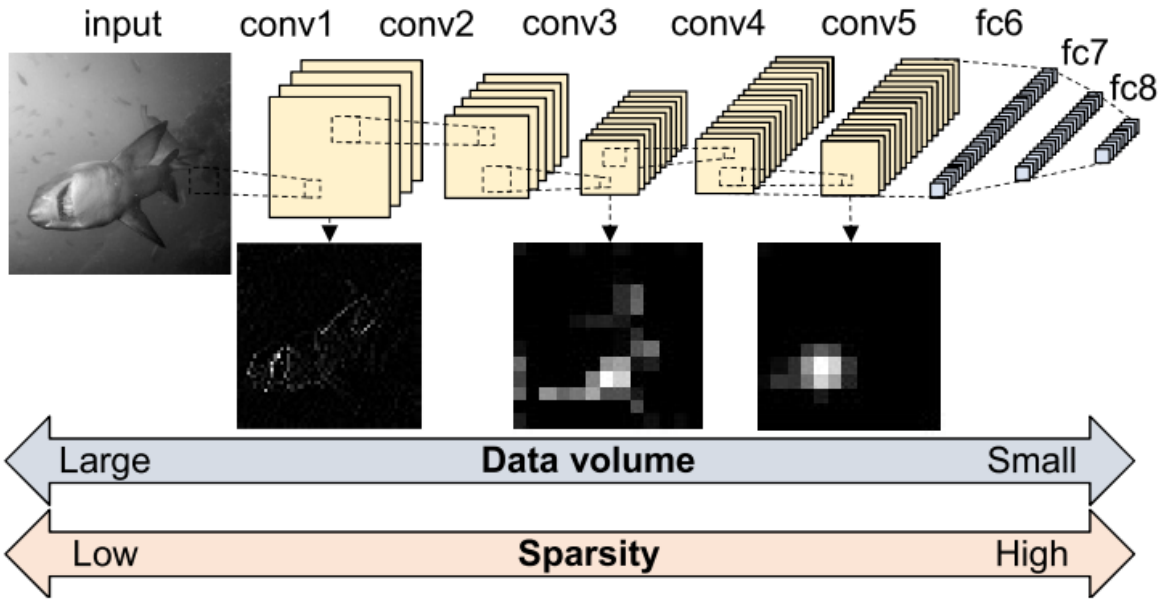
- IoT 환경에서 DNN을 활용하는 방식
  - 엣지 플랫폼에서 독립적으로 학습 추론
  - 전송받은 데이터를 호스트에서 학습 추론

- 엣지와 호스트가 조합된 환경을 제안
  - 중간 계층 피쳐를 인코딩하여 데이터 볼륨을 최소화
  - 정확도를 높이기 위해 분할 재교육



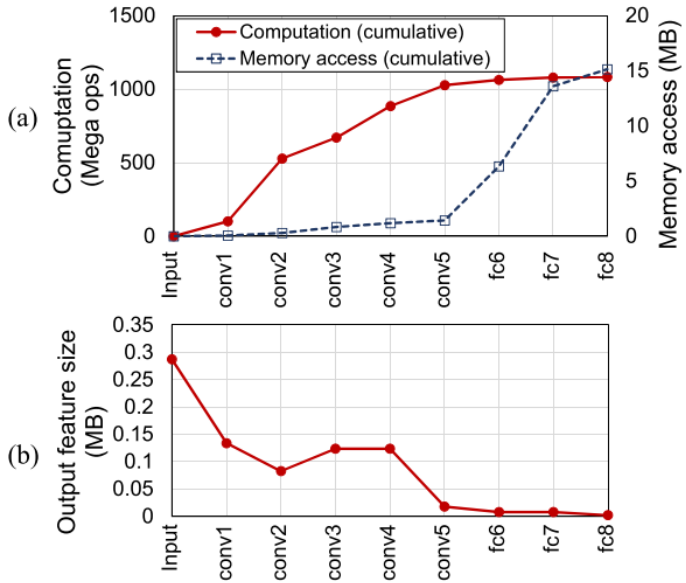
Partitioning of inference with encoding

- DNN을 데이터 인코딩 파이프라인으로 간주
  - 엣지 플랫폼은 로우 영상 데이터를 전 처리하는 인코딩 엔진으로 간주



Partitioning of inference with encoding

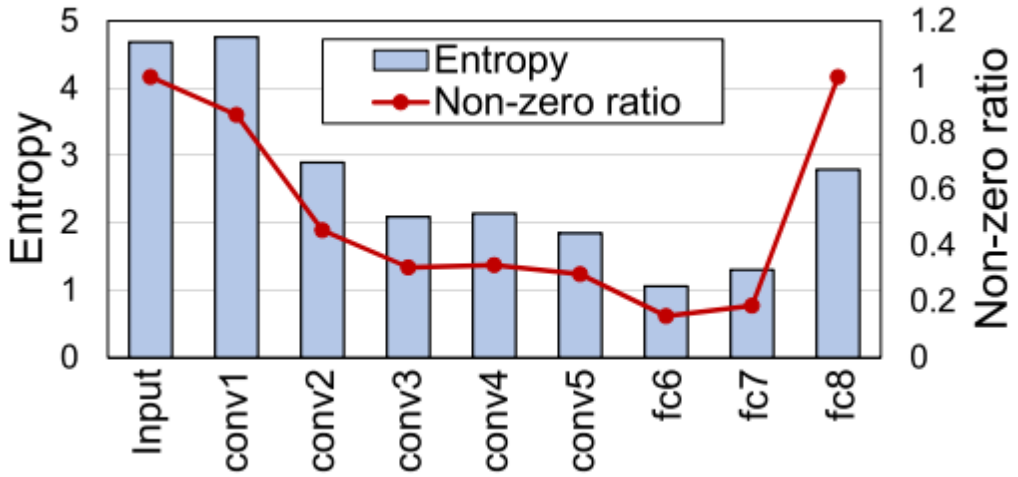
- 신경망 전체를 엣지 장치에서 처리하는 것은 부하가 큼
- AlexNet의 입력 피처 크기를 고려할 때
  30fps를 만족하기 위해서는 70Mbps의 대역폭이 필요
- Conv5 에서 신경망을 분할
  - Fc6부터 메모리 액세스가 증가
  - Conv layer는 커널 크기가 작아
    제한된 저장 공간을 가진
    엣지 장치에서 처리하기 좋음



Partitioning of inference with encoding

- 레이어가 깊어질 수록 zero ratio가 증가하고 entropy가 감소

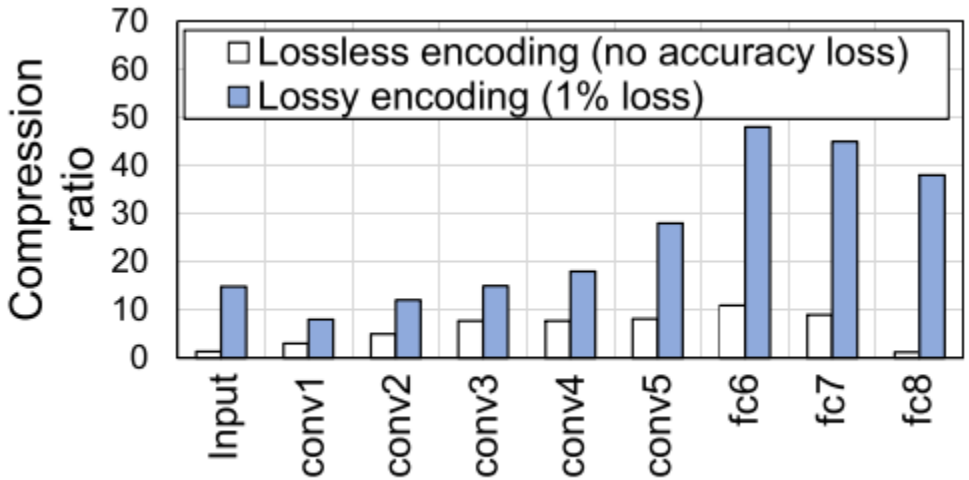
- 정확도 손실을 최소화하면서 인코딩의 이점을 활용하기 위해
  무손실 인코딩, 손실 인코딩을 비교



Partitioning of inference with encoding

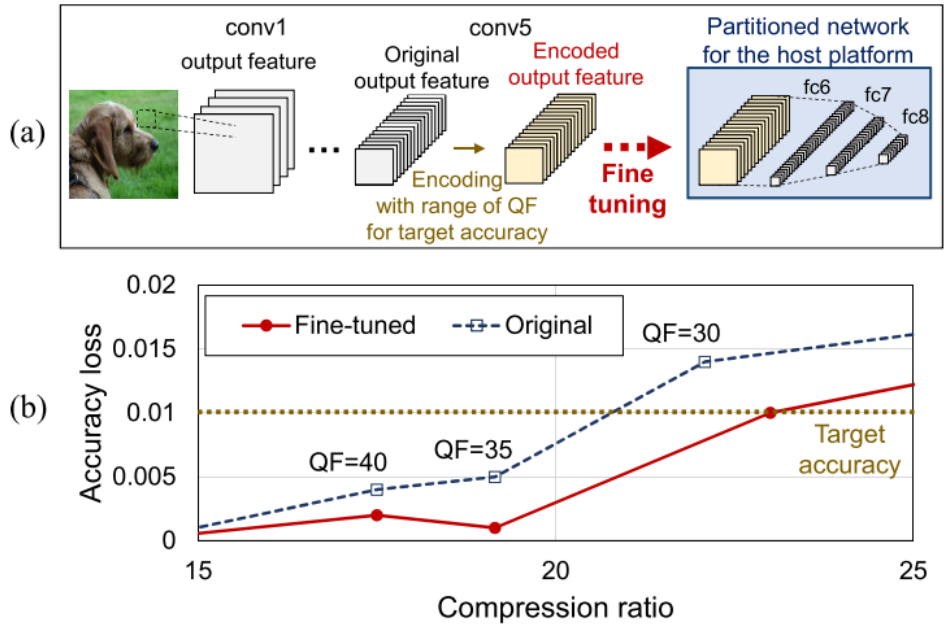
- Lossless encoding
  - 피처의 희소성을 이용해 호프만 인코딩 기법을 활용
  - 정확도 손실 없이 3~10배 데이터 압축 가능

- Lossy encoding
  - 중간 피처에 JPEG 인코딩을 적용
  - 1%의 정확도 손실을 통해 5~50배 데이터 압축 가능



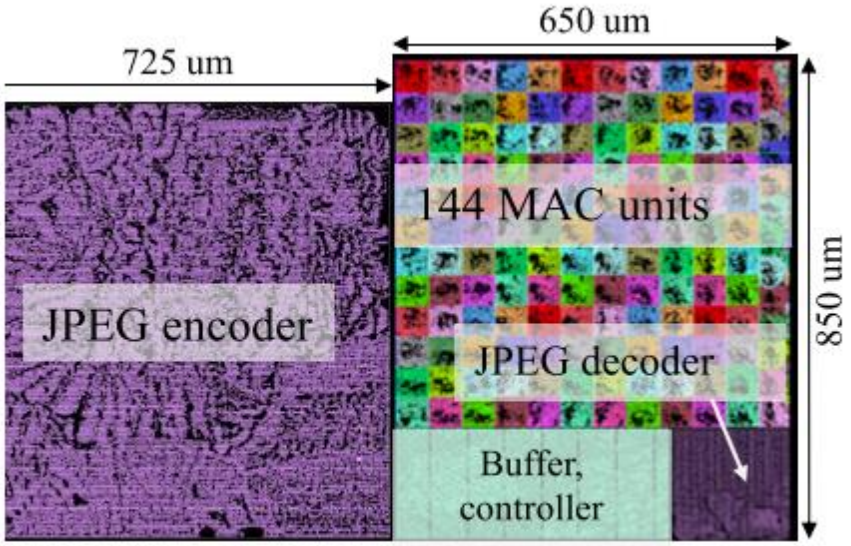
Partitioning of inference with encoding

- 미세 조정
  - 엣지 계층은 고정하고 호스트 계층에서 미세 조정
  - 미세 조정을 통해 압축률을 11% 향상시킬 수 있음



Simulation Results

- 144개의 16비트 MAC 유닛 어레이, JPEG 인코더로 구성된 추론 엔진
- 2Mbps로 대역폭 제한



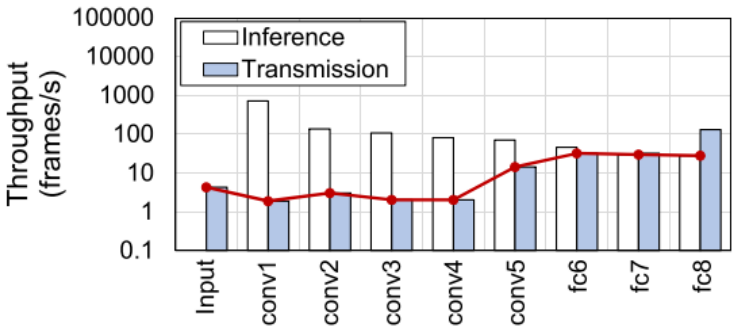
(a)

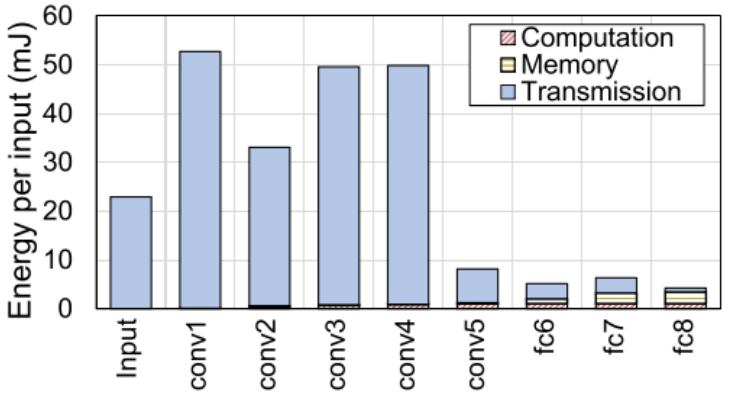
Module	Area (mm^2)	Power (mW)
JPEG encoder	0.526	3.5
JPEG decoder	0.032	0.88
MAC units	0.436	4.32
Buffer/controller	0.085	0.48
Total	**1.079**	**9.18**

(b)

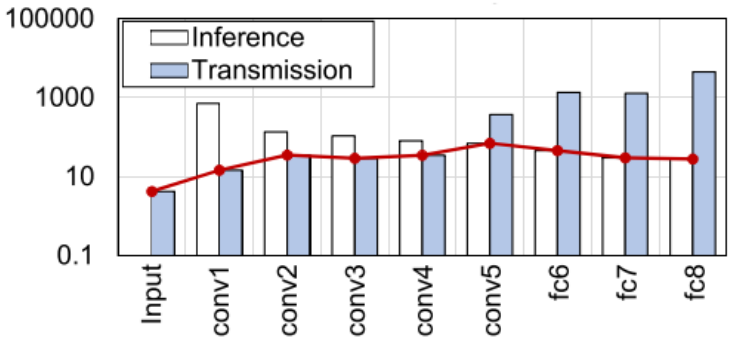
Simulation Results

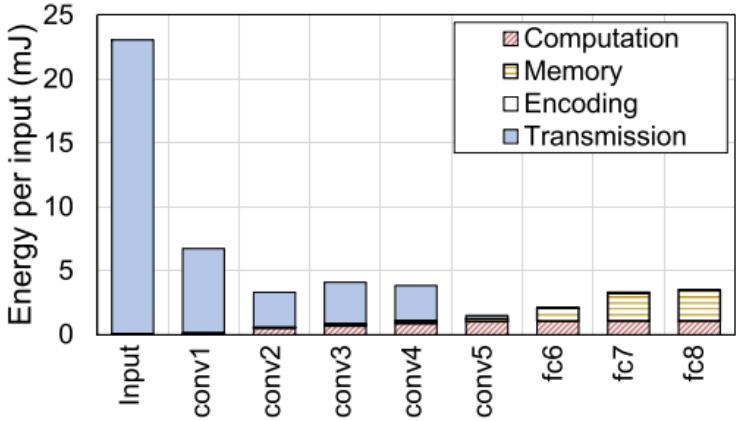
Without encoding





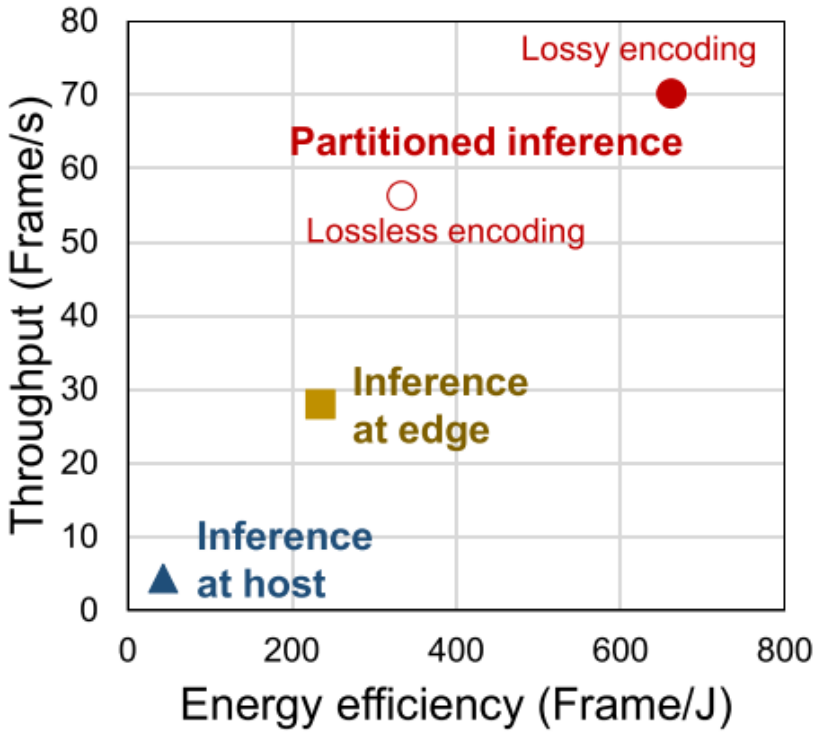
With lossy encoding





Simulation Results

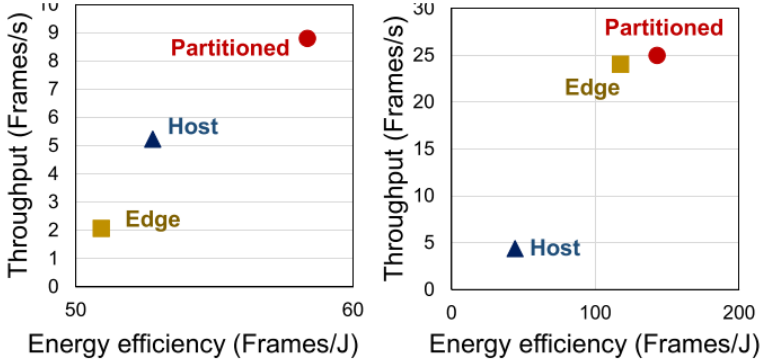
- 손실 인코딩을 이용한 파티셔닝 방법은
  - 호스트 추론보다 15.3배 높은 에너지 효율과 16.5배 높은 처리량 향상을 달성
  - 엣지 추론보다 2.3배 높은 에너지 효율과 2.5배 높은 처리량 향상을 달성



Effect of Nework Types

- VGG16
  - Fc 레이어에서 많은 메모리 접근이 요구되어 엣지 추론 성능이 크게 떨어짐
  - 파티셔닝 방법은 엣지 추론보다
    1.2배 높은 에너지 효율과 4.3배 높은 처리량 향상을 달성

- ResNet50
  - 하나의 Fc 레이어를 보유
  - 피처 크기가 작아 연산 요구량이 낮음
  - 파티셔닝 방법은
    엣지 추론보다 나은 성능을 보임



	Computation (Giga ops)		Memory (MB)	
	Conv	FC	Conv	FC
AlexNet	1.0	0.06	1.43	13.7
VGG-16	14.6	0.12	8.2	22.2
ResNet-50	3.5	0.002	22.1	0.7

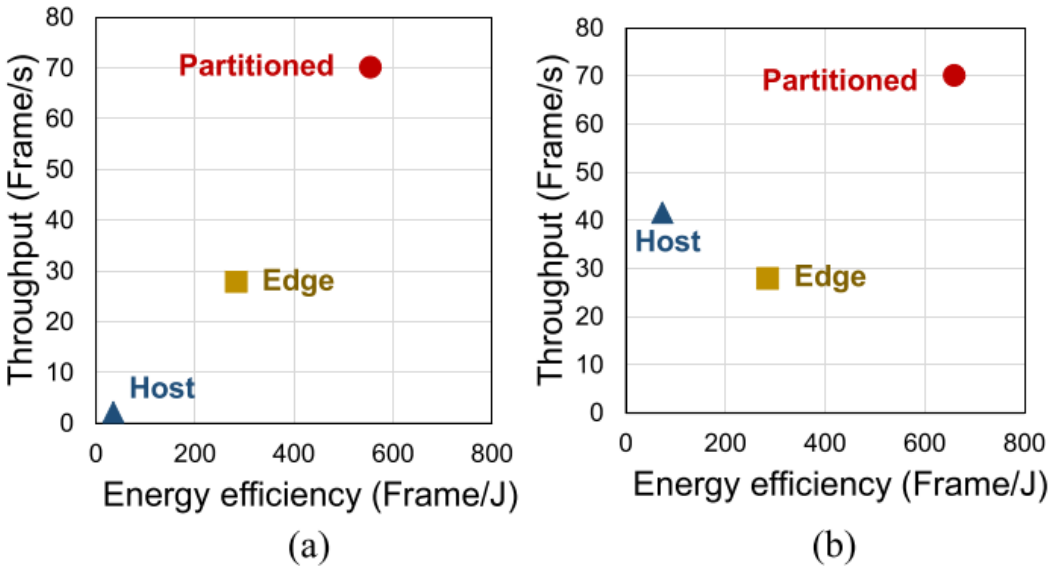
(c)

Effect of Transmission Channel

- 엣지 추론과 파티셔닝 방식은 네트워크 병목이 거의 없어

  성능 변화 또한 거의 없음

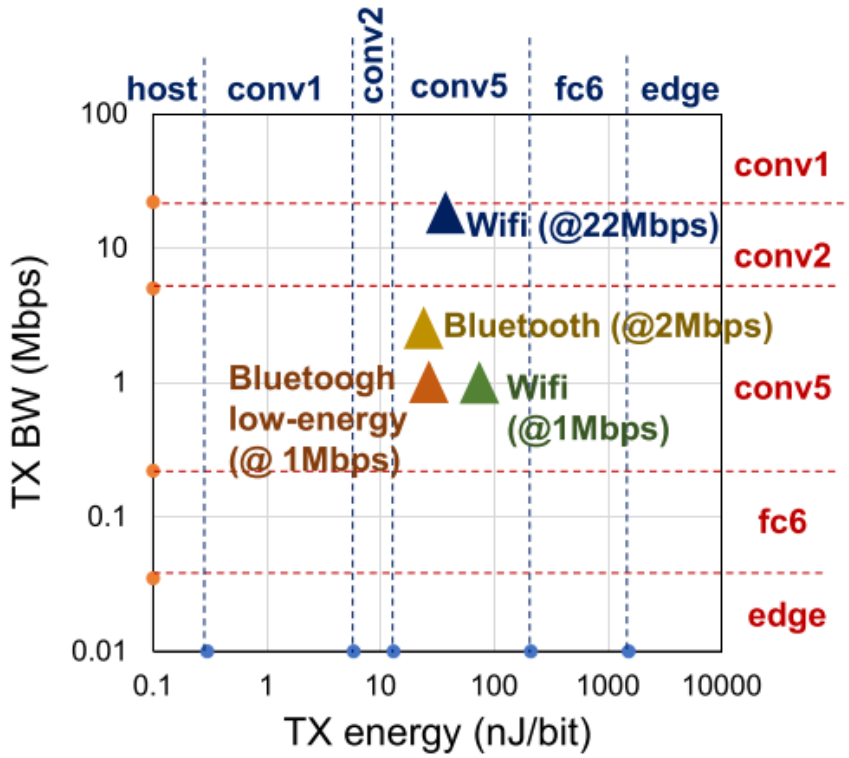
- 호스트 추론은 네트워크 대역폭에 따른 성능 차이가 큼



Effect of Transmission Channel

- 네트워크 대역폭과 에너지 소비량에 따라

  최적의 분할 계층이 달라짐



Conclusions

- 엣지 장치의 에너지 효율과 처리량을 향상시키기 위해

  DNN 추론 작업의 분할을 제안

- 중간 피쳐 인코딩과 분할된 네트워크의 미세 조정을 통해

  모델의 성능을 향상