Resource-Constrained Classification Using a Cascade of Neural Network Layers

S. Leroux, S. Bohez, T. Verbelen,
B. Vankeirsbilck,P. Simoens and B. Dhoedt

2015 International Joint Conference on Neural Networks(IJCNN),
Killarney, 2015, pp. 1-7.

Introduction

- IoT 장치는 성능과 전력 공급에 대한 제한 사항이 있음.
- IoT 장치에서 훈련된 신경망에 대한 추론은 중요한 작업.
- DNN 모델에 대한 추론 속도를 높이는 방법으로
  신경망 레이어에 Cascade를 적용하는 것을 제안.
- Cascade를 이용하면 모든 레이어에서 신뢰도를 출력할 수 있음
- 임계값 이상의 신뢰도가 달성되면 더 깊은 레이어를 연산할 필요 없음.

Introduction

- Cascade를 이용하면 모든 레이어에서 신뢰도를 출력할 수 있음.
- 임계 값 이상의 신뢰도가 달성되면 더 깊은 레이어를 연산할 필요 없음.
- 엣지-클라우드 연산으로도 이용 가능.
  - 로컬에서 추론을 시작하고, 더 깊은 연산이 필요할 때만 서버에서 추가 연산.
  - 서버와 통신이 단절되어도 지속적인 추론 가능.
- 로컬 레이어에서 실시간 추론 결과를 제공하고,
  원격 레이어를 통해 더 신뢰도 높은 결과를 추후 제공할 수도 있음.

Architecture

- 레이어가 깊어질 수록 에러율이 감소.
- 레이어 증가에 따른 에러율 감소의 폭이 점점 줄어듬.

Network architecture	Test error rate	Reference
1-layer NN (no hidden layer)	12.00%	LeCun et al. [5]
2-layer NN (1 hidden layer)	4.70%	LeCun et al. [5]
3-layer NN (2 hidden layers)	3.05%	LeCun et al. [5]
4-layer NN (3 hidden layers)	1.00%	Salakhutdinov and Hinton [6]
6-layer NN (5 hidden layers)	0.35%	Ciresan et al. [1]

- 얕은 신경망으로도 충분한 신뢰도를 얻을 수 있음.

Architecture

- 레이어마다 출력 레이어를 추가해 신뢰도를 확인.

- 신뢰도가 임계값을 초과하면
  해당 레이어의 출력이 신경망의 출력으로 사용.

Algorithm 1 Propagating a sample through the network

```
1: procedure FPROP(x)
2:     i ← 0
3:     y ← output_layer_i(x)
4:     while confidence(y) < threshold and i < n do
5:         x ← hidden_layer_i(x)
6:         i ← i + 1
7:         y ← output_layer_i(x)
8:     return y
```

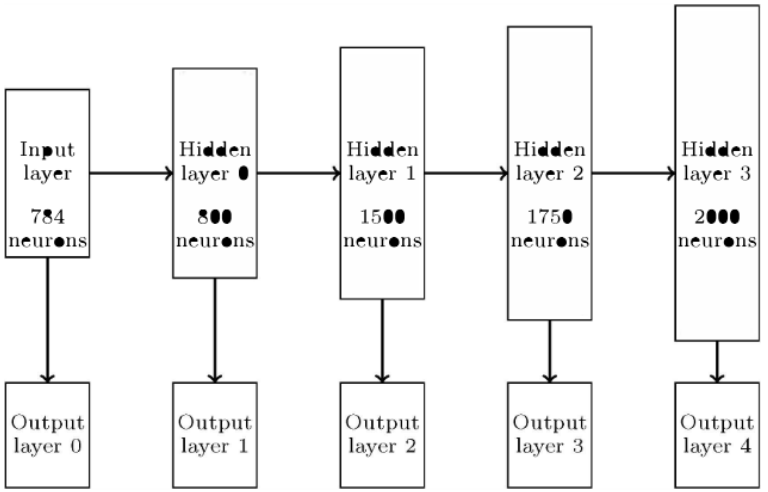
Threshold

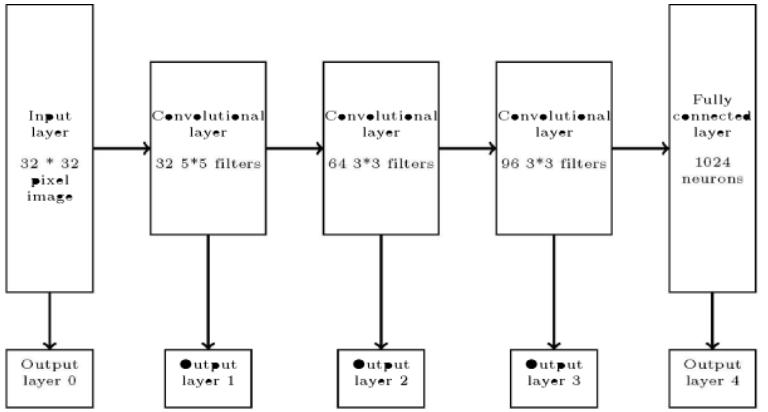
- 낮은 임계값을 선택하면 얕은 레이어에서 신뢰도가 낮은 값을 반환.
- 높은 임계값을 선택하면 깊은 레이어에서 신뢰도가 높은 값을 반환.
- 최적의 임계값은 정확도와 속도 중에서 중요하게 여기는 것에 따라 다름.

성능 평가

- MNIST
- CIFAR-10

- 기본 신경망 구성에
  각 레이어마다 출력 레이어를 추가
  (Softmax Classifier)





성능 평가

- MNIST
  - 레이어 수의 증가, 임계값의 증가로 인해 발생하는 처리 시간에 비해 성능 향상이 더딤.
  - 임계값이 0.999 이상이면 정확도는 증가하지 않지만 처리 시간은 증가함.

Number of hidden layers	Test error rate	Average time needed to process one test sample (ms)
0	7.61%	0.5
1	1.42%	0.9
2	0.79%	1.5
3	0.69%	2.6
4	0.64%	4.0

Threshold	Test error rate	Average time needed to process one test sample (ms)
0.9	1.36%	0.8
0.99	0.74%	1.3
0.999	0.64%	2.0
0.9999	0.64%	2.9
0.999 99	0.64%	4.1
0.999 999	0.64%	4.9
1	0.64%	5.2

성능 평가

- CIFAR-10

Output layer	Test error rate	Average time needed to process one test sample (ms)
0	63.27%	0.2
1	33.83%	3.0
2	28.56%	6.3
3	22.87%	8.2
4	15.33%	23.0

Threshold	Test error rate	Average time needed to process one test sample (ms)
0.9	17.57%	11.2
0.99	15.6 %	16.8
0.999	15.39%	19.9
0.9999	15.33%	21.6
1	15.33%	24.8

결론

- Cascade 방식에 대한 논문이 2015년에 발표되어
  사용한 모델이 VGG, Inception, ResNet에 비해 단순함.

- 복잡한 레이어 구조를 갖는 모델에서 어떻게 중간 결과를 출력해야 하고
  그 결과의 신뢰도가 충분한지를 확인한 논문을 찾아봐야 함.

- Cascade 방식을 분산처리 할 경우
  중간 결과 도출을 위해 이를 취합하는 장치가 별도로 필요할 것 같은데
  어떤 방식으로 설계해야 할지,
  실시간 처리가 가능한 레이턴시 내 가능한지 확인이 필요.