



Searching Efficient HAR Neural Networks

CAU Artificial Intelligence LAB

임원선

costor107@gmail.com

6th August, 2021

1. 기존 주장

문제 정의) HAR이 주로 동작하는 모바일 환경에 최적화된 효율적인 모델 구조 탐색

✓ Differentiable-NAS(DARTS)기반의 Accuracy와 Latency를 최적화하는 방식에서의 1 Phase 최적화의 문제점

원인 파악)

1. $\text{Loss} = \text{Cross-Entropy} + \lambda E[\text{latency}]$ 에서 각 term의 scalability가 다름
 - > Latency term이 dominant하면서 구조가 결정되는 경향
 - > 오히려 효율적인 모델 구성에 장점 가능
 - > 또한, 단순히 각 term의 scale을 조절하면 해결 가능한 문제

1. 기존 주장

- ✓ Differentiable-NAS(DARTS)기반의 Accuracy와 Latency를 최적화하는 방식에서의 1 Phase 최적화의 문제점

원인 파악)

- 2. DARTS기반 최적화 방식의 구조적 문제 : α 가 성능 지표를 보장하지 않음

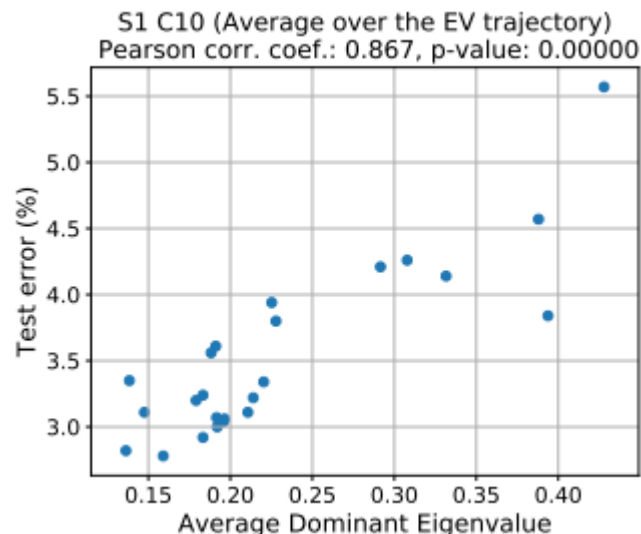
$$\min_{\alpha} \mathcal{L}_{val}(w^*(\alpha), \alpha) \quad \text{--- (1) outer optimization}$$

$$\text{s.t. } w^*(\alpha) = \operatorname{argmin}_w \mathcal{L}_{train}(w, \alpha) \quad \text{--- (2) inner optimization}$$

식(1)을 최적화 하기 전에 식(2)가 최적화 되어야 하는데,
계산비용이 크게 들기 때문에 inner optimization을
식(3)Approximation을 통해 계산 비용을 감소

$$\begin{aligned} & \nabla_{\alpha} \mathcal{L}_{val}(w^*(\alpha), \alpha) \\ & \approx \nabla_{\alpha} \mathcal{L}_{val}(w - \xi \nabla_w \mathcal{L}_{train}(w, \alpha), \alpha) \end{aligned} \quad \text{--- (3)}$$

여기서 구조 파라미터 α 가 업데이트할 때 사용하는 $\nabla_{\alpha}^2 \mathcal{L}_{valid}$ 에 대한 Hessian Matrix의 eigenvalues의 크기에 따라 생성된 구조의 Test error가 증가한다는 상관 관계를 증명한 [연구 결과 보고](#)^[1]



1. 기존 주장

✓ Differentiable-NAS(DARTS)기반의 Accuracy와 Latency를 최적화하는 방식에서의 1 Phase 최적화의 문제점

원인 파악)

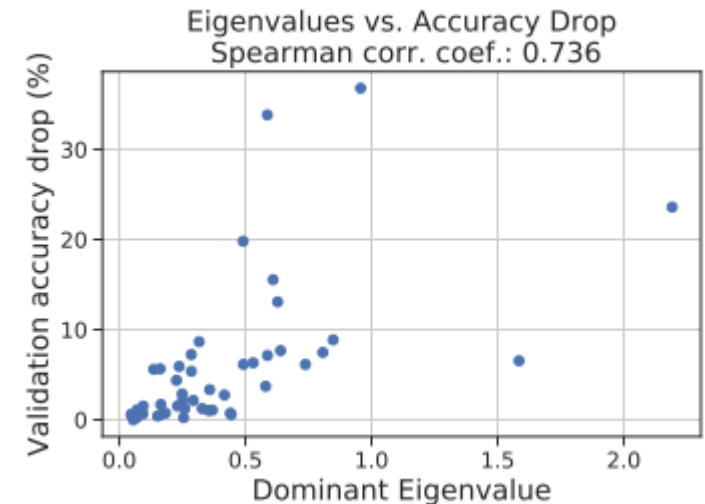
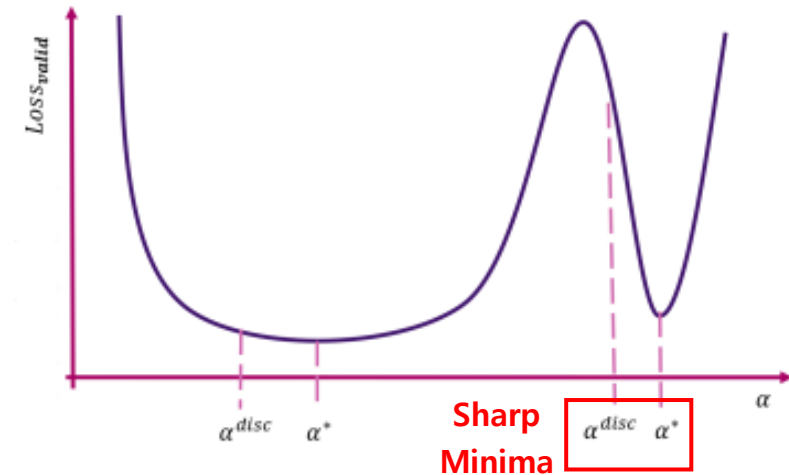
3. 최종 네트워크 구조 선택 과정에서 continuous space의 optimal 구조 파라미터 α^* 를 discretize하였을 때, α^{disc} 구조의 loss는 α^* 성능의 주위에 존재

-> α^* 가 Sharp minima에 있을 경우 α^{disc} 와의 loss 차이가 크게 나타남

-> Dominant eigenvalue를 α 의 discretize에 따른 Loss 변화량 을 확인한 결과에서도 클수록 loss 차이가 크게 나타남

Loss 변화량 : $\mathcal{L}_{valid}(\alpha^{disc}, w^*) - \mathcal{L}_{valid}(\alpha^*, w^*)$

따라서, α 가 정확도를 직접적으로 반영하는 지표가 되기 어려운 경우 존재



1. 기존 주장

✓ 기존 주장에 대한 고찰

- 결론적으로 현재 목표로 하는 효율적으로 동작하는 모델 탐색에 대한 원인으로는 작용하지 않고 모델의 generalization 문제
- Model Selection 과정에서 고려는 가능
- Latency와 같은 다른 Metric이 포함되었을 때의 영향은 더욱 알기 어려우므로 해당 문제는 다루지 않음

✓ 효율적으로 동작하는 모델을 탐색하기 위한 방안 제시

2. 효율적으로 동작하는 모델을 탐색하기 위한 방안

✓ SuperNet 구성에서 Dynamic Pooling Layer를 적용

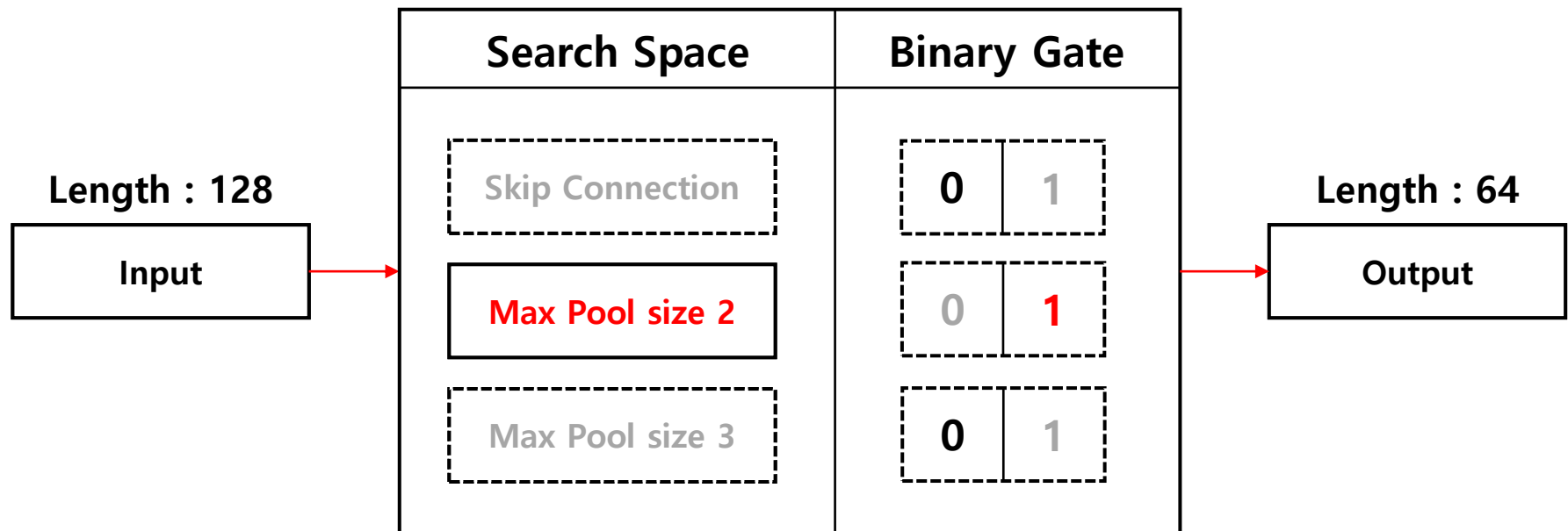
기존 연구) NAS SuperNet 구성에서는 고정된 Pooling(Reduction)층과 Pooling size

1. Pooling Layer를 탐색 가능하게 함으로써 탐색 과정에서 계산량과 연관되는 Feature map을 줄일 수 있는 가능성을 부여
2. HAR sensor 데이터의 1차원 특성상 넓은 인식 영역이 성능에도 좋은 영향을 보임
Ex) 실험에서 Cross-Entropy만을 Loss 함수로 설정한 후 탐색하였을 때에도 Standard Conv보다 Dilated Conv이 좋은 효과를 보임
3. 계산이 필요하지 않은 효율적인 연산자 Skip Connection을 탐색 공간에 같이 구성하여 효율성 유지하면서 정확도 하락 보정
4. Pooling size가 클수록 뒤쪽 layer에서 작은 kernel size를 사용 가능

2. 효율적으로 동작하는 모델을 탐색하기 위한 방안

✓ Dynamic Pooling Layer Overview

1. 같은 Layer에 서로 다른 Pooling size를 적용하면 각기 다른 Feature map 사이즈가 나오므로 Binary Gate를 통해 한 step 학습 시 한 개의 Pooling 연산자만 동작하도록 구성



A vertical bar on the left side of the slide, consisting of a red upper half and a grey lower half.

Thank you

참고 문헌

[1] Zela, Arber et al. “Understanding and Robustifying Differentiable Architecture Search.” ICLR, 2020.