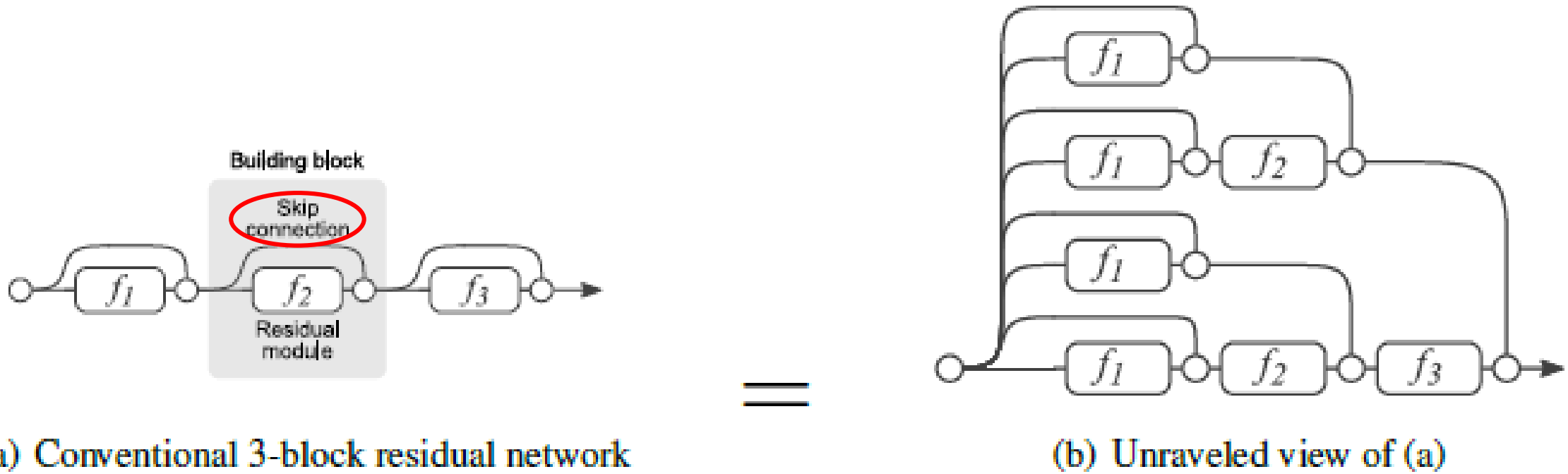
Residual Networks are Exponential Ensembles of Relatively Shallow Networks

2016. May

Cornell univ.

Residual Network 의 구조



(a) Conventional 3-block residual network

(b) Unraveled view of (a)

- 각 node 마다 두갈래길 거침 → 경로의 수가 exponentially 증가
- 2^n 개의 network들이 참여해서 output 내는 Ensemble 효과

- 수학적 modelling

$$y_{i+1} \equiv f_{i+1}(y_i) + y_i$$

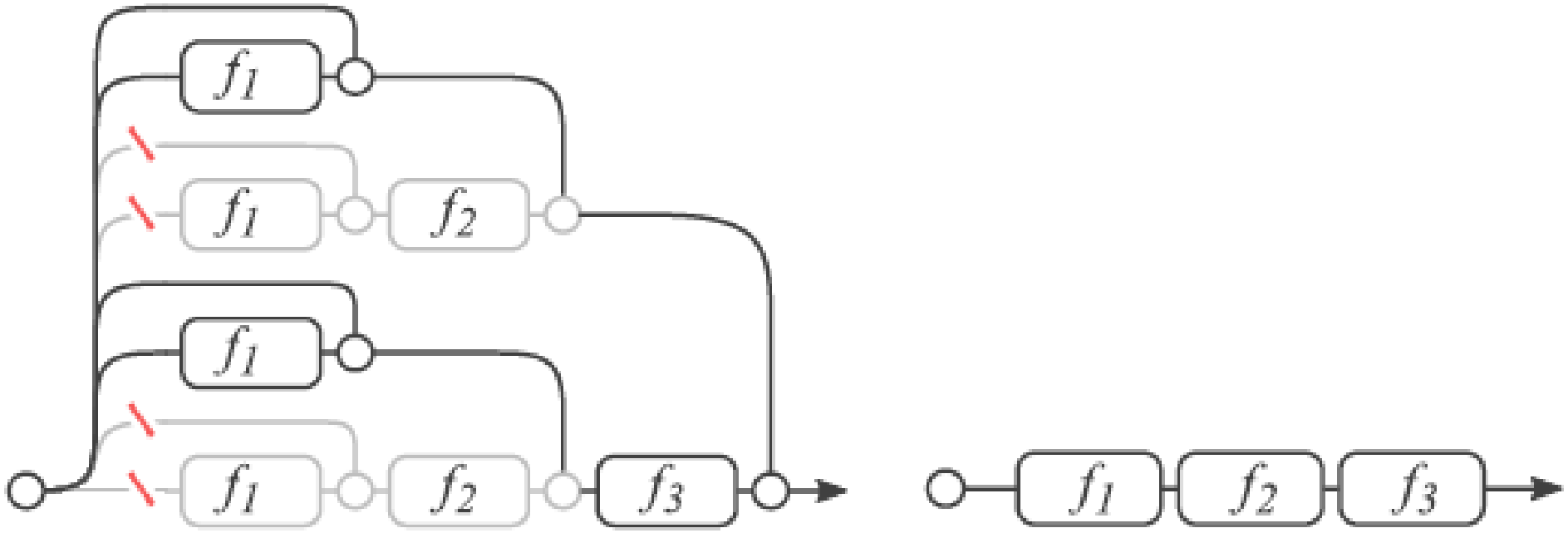
$$f_i(x) \equiv W_i \cdot \sigma(B(W_i' \cdot \sigma(B(x))))$$

(W=convolution , σ=ReLU , B=batch normalization)

- 기존보다 100배 이상의 깊이로 network 쌓아도 잘 작동

: Vanishing gradient 문제를 해결한 건 아님.

: *Shallow* network 들의 ensemble 로 우회함!



(a) Deleting f_2 from unraveled view

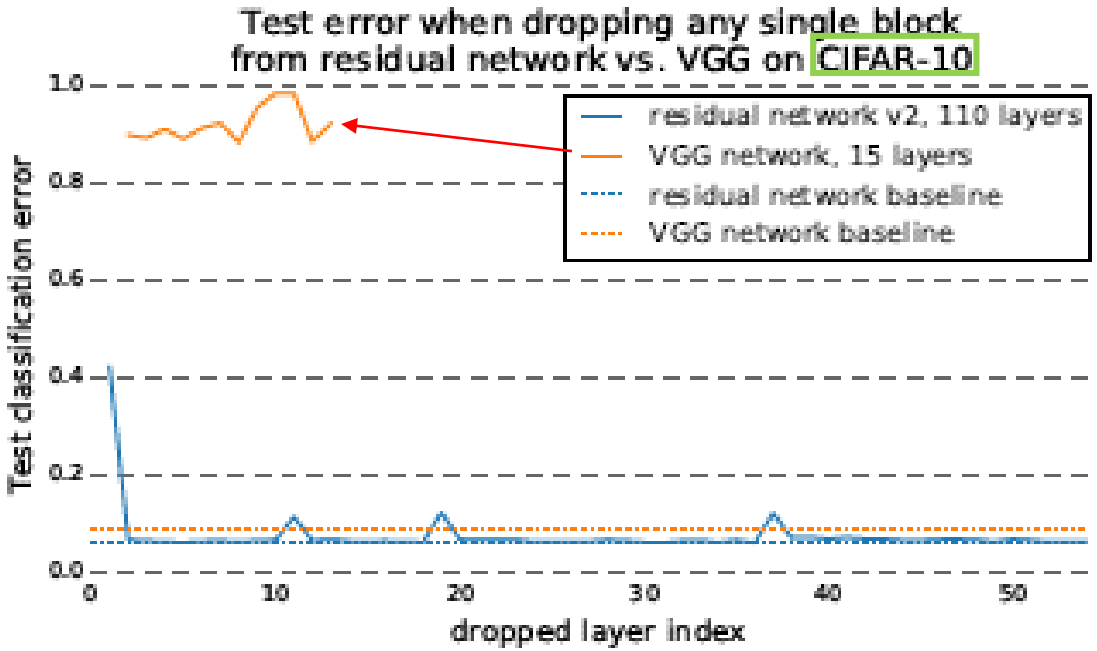
(b) Ordinary feedforward network

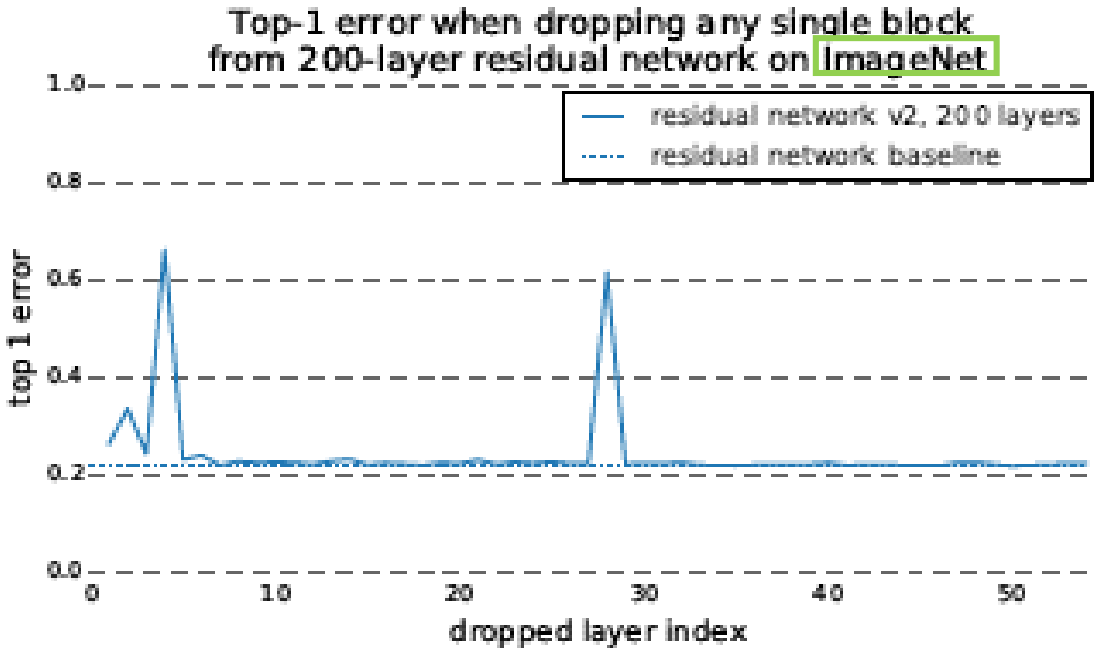
(a) 한 layer(f_2) 가 삭제되면 effective path 개수는 2^n → 2^{n-1}

(b) VGG or AlexNet 에서는 viable path 가 유일하게 존재함

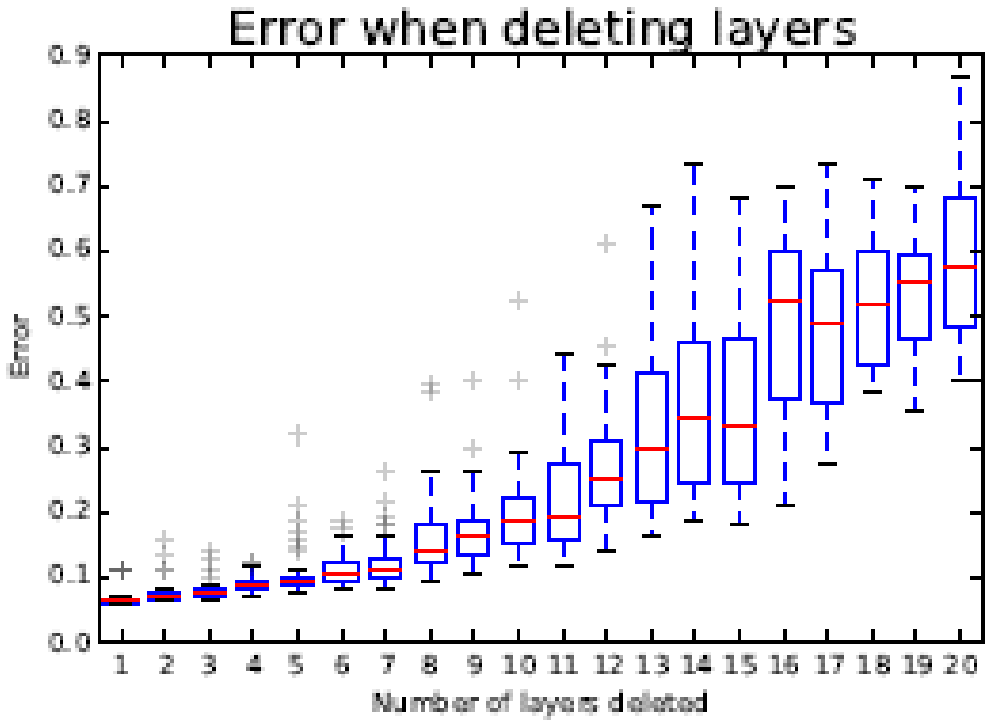
→ 어느 한 layer 만 삭제해도 전체가 끊김 (multiplicity = 1)

실험적 접근

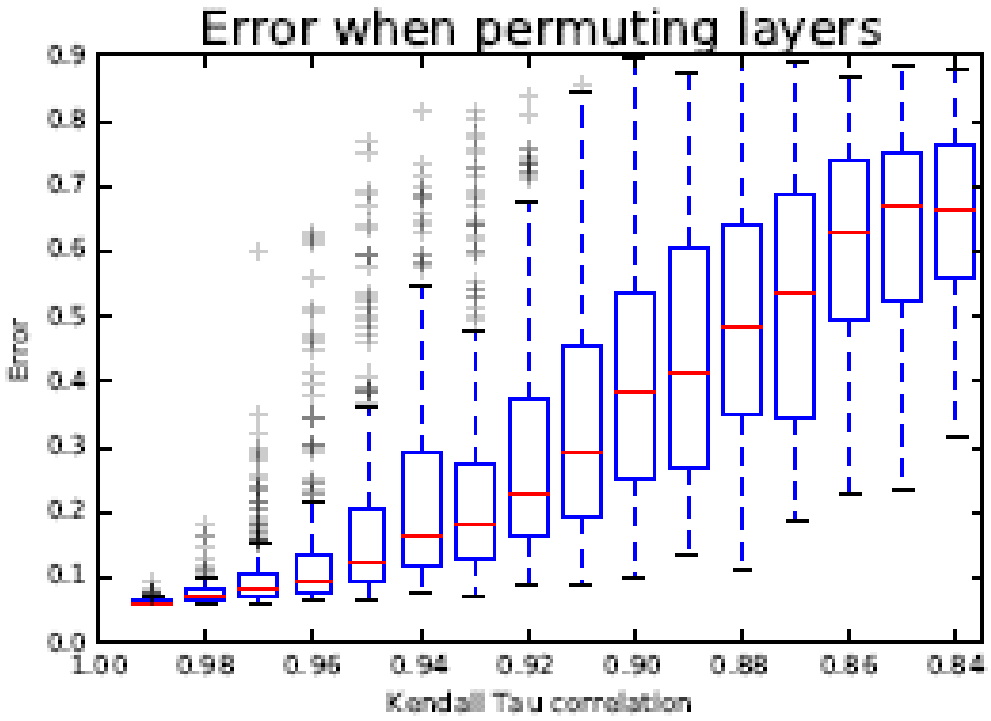




- Error peaks on initial/downsampling modules – Why ??

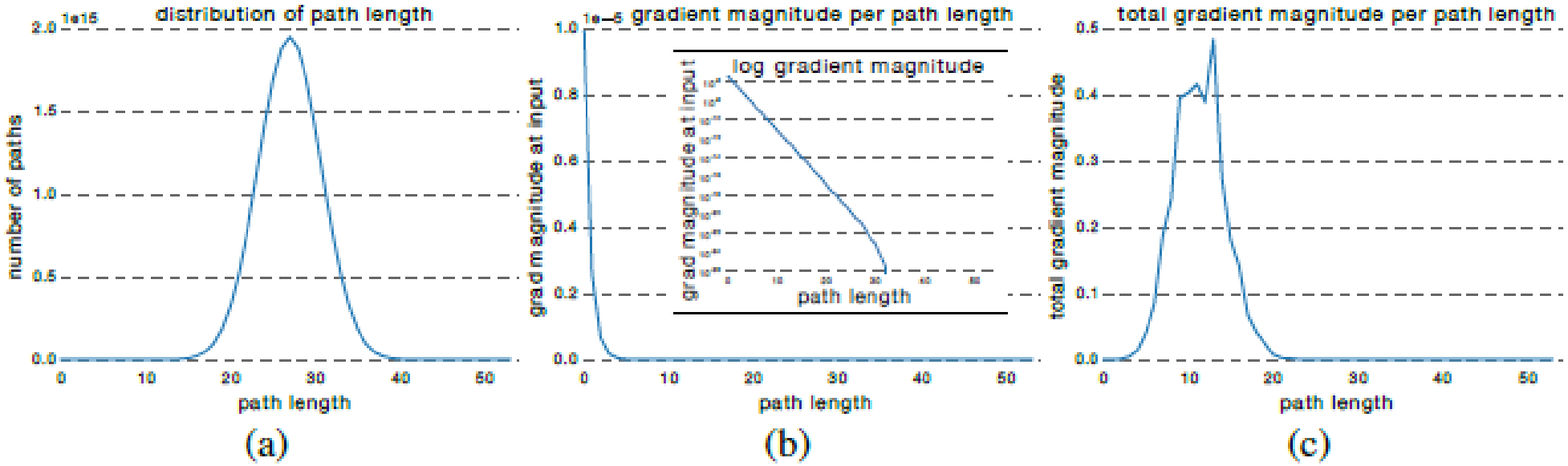


(a)



(b)

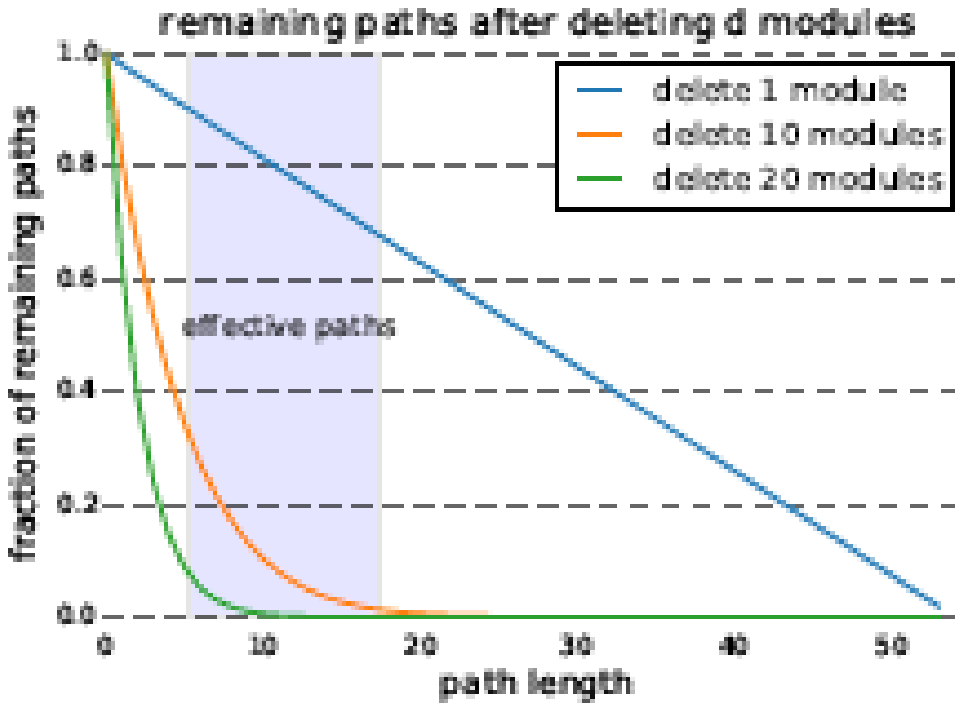
(a) deleting 에서나 (b) permuting(=shuffling) 에서나 동일하게 error rate 는 "완만하게" 증가 ➔ Ensemble 구조의 증거!



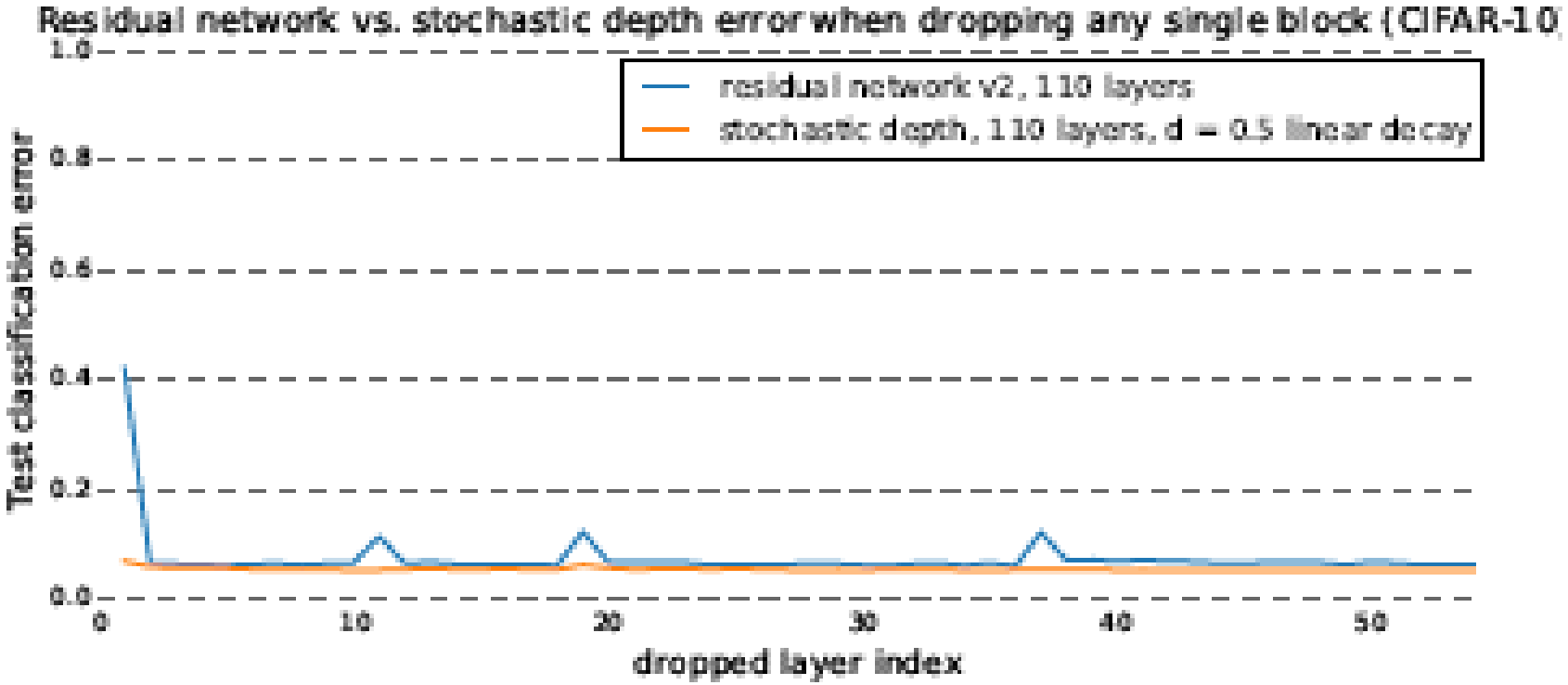
(a) Path length 는 Binomial distribution 따름 ~ B(n, ½)

(b) 각 경로가 만들어내는 Gradient 의 크기는
경로 길이에 따라 지수적으로 감소 (vanishing gradient 때문)

(c) =(a)x(b). 각 path length 빈도 고려하면 Gradient update 는 주로
5~17 modules 의 shallow networks 로부터 형성 (effective paths)



54 module network 에서 10개의 residual module 을 삭제해도
5~17 module 길이를 가진 effective paths 의 상당수가 유효함
→ Performance 는 조금만 감소!



- Stochastic depth training → Better performance !

  module들 중에서 random subset 선택해서 mini-batch training

  (Test 시 모든 path 이용 가능 ∴ multiplicity 유지됨)

결론

- Residual network 가 잘 작동하게 하는 것은 depth 가 아니라 ensemble 이다!

- Neural net 을 표현하기 위해, depth 와 width 외에 추가적으로 “Multiplicity” 라는 차원이 필요함 : 내재된 ensemble 의 size

- Multiplicity 의 개념과 작동방식에 대한 후속 연구가 필요!