최근 연구들은 네트워크의 깊이가 중요한 요인이라는 것을 밝혔다. 하지만 단순히 레이어를 쌓는 것은 어느 순간부턴 성능을 개선하지 못한다는 문제가 발생했다.

그에 따라 "Is learning better networks as easy as stacking more layers?"라는 질문이 제시되었다.

이 문제를 답하는 것에는 두 가지 어려움이 존재한다. Vanishing/exploding gradients.

이 문제는 normalized initialization과 intermediate normalization에 의해 어느 정도 해소되었다. 이 두 가지를 통해 깊은 네트워크도 역전파를 사용하는 확률적 경사 하강법에서 수렴할 수 있게 되었다.

Normalized initialization: 정규화된 초기화, 신경망의 가중치를 적절한 범위로 초기화하는 법 (Xavier Initialization: 평균 0 분산 1 유지, He Initialization: ReLU 활성화 함수에 최적화)

Intermediate normalization: 중간 정규화된 초기화, 네트워크의 중간 단계에서 신호를 정규화하는 기법, 각 레이어의 출력을 정규화하여 Gradients Vanishing 문제를 방지, Batch Normalization(배치 정규화)

그런데 문제가 또 발생, 네트워크가 깊어지다보니 degradation problem이 발생함. 깊이가 깊어질수록 정확도는 정체되다가 급격히 감소함. 놀랍게도 이는 과적합 문제가 아니라 training error였음.

이론적으로 deeper model은 shallow model보다 성능이 떨어지면 안된다. 왜냐하면 추가된 layer들은 항등함수이기 때문이다. 다른 layer들은 이미 학습된 shallow model에서 그대로 복사되기에 이미 좋은 성능을 내고 있다면 deeper model에서도 그 성능이 유지되어야 한다. 하지만 실험에서는 deeper model의 성능이 떨어졌다. 이는 일반적인 신경망에서 layer가 직접 학습해야할 매핑이 복잡해져 최적화가 어려워졌기 때문이다.

해당 논문에서는 deep residual learning framework를 통해 이러한 문제를 해결했다.

기존 네트워크에서는 각 층이 특정 매핑 함수 H(x)를 직접 학습해야 한다.