[1주차] 논문리뷰

1. Introduction

💡 논문에서 다루고 있는 주제가 무엇인지와 해당 주제의 필요성이 무엇인가

논문에서 제안하는 방법이 기존 방법의 문제점에 대응되도록 제안 되었는가

논문이 다루는 분야

- 딥러닝 / 컴퓨터 비전 / CNN

해당 task에서 기존 연구 한계점

- 레이어만 깊게 쌓으면 더 좋은 모델을 만들 수 있는 것인지에 대한 궁금증이 이어져오고 있었음.
- vanishing 문제로 너무 깊은 층에서 오히려 error율이 높았음. (accuracy 감소함) (vanishing/exploding gradient : sigmoid 같은 활성함수 사용 시 기울기가 0에 가까워져 학습이 잘 안되는 문제)

    - 네트워크 가중치값 초기에 적절히 설정하여 문제 해결하려는 시도 있었음.
    - 단순히 overfitting 때문이 아니라 training error을 증가시키는 문제=학습 자체가 잘 안된다는 것
- layer를 깊이 쌓기 위해서 identity mapping를 증가시켰을 때, 논리적으로는 (층을 쌓으면 더 좋다는 논리로는) 적어도 error가 더 커져서는 안되는데 모순이 발생 ???제대로 이해한 게 맞나???

논문의 contributions

- skip-connection을 취했고, 이는 역전파 계산 시 기울기 소실 문제를 완화해줌
- 훨씬 깊은 network(깊은 층의 network)를 학습할 수 있도록 해줌
- VGG에 비해 복잡도 낮고 성능은 더 좋음

  → 이미지 task에서 깊이는 표현력의 증대를 의미함. 깊은 층을 활용할 수 있게 되었으므로 표현력 증대에 기여한 것.
- object detection이나 semantic segmantation에 대해서도 좋은 성능을 보여주었음.

2. Related Work

Introduction에서 언급한 기존 연구들에 대해 어떻게 서술하는가

제안 방법의 차별성을 어떻게 표현하고 있는가

- residual한 technique들에 대해 기존부터 사용됐던 방법임을 밝히고 있음.
- shortcut connection에 대해 비슷한 갈래의 논문으로 “highway networks”를 언급하고 있음
    - ResNet처럼 깊은 네트워크 잘 학습하게 하기 위한 방법 제시하였음
    - gating function이용해서 shortcut 커넥션과 유사한 기능을 제공
- 다른 점
    - 일단 gate와 달리 resnet은 parameter-free임
    - 그리고 gate shortcut이 닫히면 non-residual function과 다를 바가 없어짐. ResNet은 always residual function을 학습함. (identity shortcut never closed ??? 이 부분이 잘 이해가 안감)

3. 제안 방법론

💡 Introduction에서 언급된 내용과 동일하게 작성되어 있는가

Introduction에서 언급한 제안 방법이 가지는 장점에 대한 근거가 있는가

제안 방법에 대한 설명이 구현 가능하도록 작성되어 있는가

Main Idea

- degradation problem을 해결하기 위해 deep residual learning framework를 제안함
  - Instead of hoping each few stacked layers directly fit a desired underlying mapping, we explicitly let these layers fit a residual mapping.
  - 의도하는 H(x)에 대한 직접적인 학습이 아니라, 문제를 바꾸어서 F(x) := H(x) - x 를 대신 학습하는 것
  - ex) identity mapping이 적절한 mapping이라고 한다면 그걸 학습하는 것보다 F(x)가 0인 걸 학습하는 방향이 더 쉽다는 것 ??? 왜 더 쉽지 ???
  - “shortcut connections” 결과인 F(x)에 x를 더해주는 것 (=skip connection)
    - 그냥 x를 더해주는 것이기 때문에 새로운 parameter가 등장하지도 않고, 복잡도가 증가하지도 않는다. 구현도 간단하다.
- H(x) 를 optimal한 mapping이라 할 때,
- 여러 개의 non-linear한 layer를 이용해서 점진적으로 복잡한 함수 학습
- H(x)-x를 학습하는 것 → 수학적으로는 별 차이가 없어보이지만 F(x)를 학습하는 것으로 인식하면 난이도가 훨씬 쉬워짐. (???왜???)
  - 기본적인 identity mapping을 매번 추가해줌(???입력값을 매번 추가해준다 뭐 그런건가?) - 얘가 답은 아닐지라도 방향성 제시에 도움이 됨.
  - 찾고자 하는 optimal한 function이 zero mapping보다는 identity mapping에 가깝다면 identity mapping을 제시해주는 게 더 쉽겠다.
  - 잔여한 정보(F(x))만 학습할 수 있도록 도와주는 것. 기존의 입력값을 더해줌으로서 방향성을 잘 잡아주는 것. ???근데 function을 거쳐나온 값이랑 기존값을 더하면 1을 넘을 일은 없나?
  - 이전 layer의 값을 보존하고 추가적으로 학습할 부분만 학습 - 이런 시스템이 아니면 매번 새로운 mapping을 만들어내야하는 거니까 더 어려운 과제가 되는 것임

- building block
  - y = F(x, {Wi}) + x
  - F는 residual mapping을 의미함
  - x는 identity mapping 즉 shortcut connection을 의미함
  - y = F(x, {Wi}) + Wsx : x가 속한 input dimension과 결과 dimension인 F의 dimension을 맞춰주고자 Ws를 이용해서 linear projection함. 같은 차원으로 dimensino 맞춰줌.
    - but 그냥 identity mapping 이용했을 때에도 충분히 높은 성능 보여줄 수 있다. (Wsx가 아니라 x이용해도 높은 성능이라는 뜻이지???)
  - F가 single layer면 별 의미가 없고 여러 개 weight 값이 중첩된 형태여야 한다.
    ??? 어떤 꼴인지 잘 안그려짐
    ??? convolution layer를 하나씩 묶어서 진행하는 게 아니라 두 개 이상씩 묶어서 residual 방식을 진행해주라는 의미인가???

Contribution

- 결과적으로
  1. easy to optimize
  2. increased depth에서도 더 높은 accuracy
  3. CIFAR-10과 ImageNet 모두 ResNet 사용 시 성능 개선 (=특정 dataset에만 적용되는 이야기가 아니다. Object detection이나 Semantic segmentation에 대해서도 좋은 성능을 보였음)
  4. 앙상블 기법에 적용했을 때 좋은 accuracy 보임

4. 실험 및 결과

Introduction에서 언급한 제안 방법의 장점을 검증하기 위한 실험이 있는가

Dataset

- ImageNet / CIFAR-10 등의 Data Set 사용
  - ImageNet 2012년도 Dataset 이용함
  - 1.28 million training images / 1000 classes

- Object detection과 Segmentation 등도 진행

Baseline

- 비교 목적으로 기본적인 CNN 과 비교함 (Plain Network)
  - VGG 기법을 기본으로 따르고 있음
  - 3X3 filter 사용, filter개수 동일
  - feature map size가 절반이 되면 필터 개수를 2배로 함 (time complexcity 보존)
  - 별도의 pooling layer 사용 X, stride 값을 2로 설정해서 downsampling함
  - 마지막 단계에서 global average pooling 이용해서 1000-way fully connected layer 만듦(softmax 사용)

결과

- 본 모델은 일반적인 VGG 모델보다 더 적은 파라미터 사용하고 복잡도도 낮았다.
  - ??? 일반적인 VGG모델에 residual 블록을 추가한 꼴 아니었나? 파라미터 개수는 같은 거 아닌가?
- 그림에서 34 layers의 VGG와 residual box 가 가미된 34 layers VGG를 확인할 수 있음. (점선은 입력값 차원 조정해준 것)
- 크기를 바꿔가며 layer를 쌓아주고 있고, layer를 쌓는 중간중간에 크기에 따라서 각각 residual 기법을 3번, 4번, 6번, 3번 사용하였음
- plain Net에서는 layer가 깊어질수록 정확도가 떨어졌으나,
- Res Net에서는 layer가 깊어짐에 따라 정확도가 높아짐을 확인할 수 있었다.
- 달라진 건 shortcut connection이 추가된 것 뿐임에도 불구하고 성능 훨씬 개선
- training error 개선되었고, 일반화 성능도 좋았음. 그래프에서 수렴 속도 또한 더 빠름을 알 수 있음.

- shortcut connection을 위해
  - (A) identity mapping을 사용할 지

    → zero padding을 이용해서 디멘션을 늘려주고 identity mapping 사용
  - (B) projection을 사용할 지

    → dimension 증가할 때만 projection 연산 사용

    - (C) 모든 shorcut에 대해 projection 사용

⇒ (C) table3에서 C의 성능이 가장 높게 나옴 but 필수라고 할 정도로 높은 성능개선을 보여준 것은 아님

- bottlenect Architectures : 복잡도를 증가시키지 않기 위해 효과적으로 사용 가능
  ??? 이 Arcitecture는 무엇이냐

- The parameter-free identity shortcuts are particularly important for the bottleneck architectures. → ??? 보틀넥인 경우는 identity 가 더 효과적이라고 하려는 것 같은데 일단 보틀넥 구조를 모르기에 왜 더 효과적이라고 하는지 이해를 못했음.
- layer 깊게 쌓고 앙상블까지 더했을 때 매우 좋은 성능을 보였다.
- ??? 근데 왜 복잡도가 낮아졌다고 하는걸까...복잡도는 왜...낮아진걸까???

- 저자들은 vanishing gradient 때문에 발생한 문제(optimaization difficulty)가 아니라고 주장함. forward나 backward signal vanish는 거의 이루어지지 않았다고 말하고 있음

  → Plain Net의 문제점은 exponentially low convergence rates 때문일거라고 말하고 있음. 수렴률이 기하급수적으로 낮아지는 문제 (???무슨 수렴률???)

  (convergence rate은 최적화 기법에서 등장하는 개념으로 수렴을 위해 필요한 epoch이나 수렴 난이도를 언급하고자 할 때 사용하는 척도)

- VGG에 비해 FLOPs 감소(딥러닝 모델에서 계산 복잡도를 나타내기 위한 척도)
- 입력단과 출력단의 dimension이 동일할 때는 identity mapping 사용할 수 있다고 말하고 있음
    - 그렇지 않다면

    1. 사이드에 패딩 붙여서 identitiy mapping 사용
    2. projection 연산 이용한 shortcut connection 이용
- ImageNet - 224X224 crop / horizontal flip ???
- ResNet의 또다른 특징 : 매 컨볼루션 layer 거칠 때마다 batch 정규화 이용했음
- 학습 진행 과정에서 learning rate도 점차적으로 줄여나갈 수 있도록 설계하였음

- CIFAR-10에 대해서도 마찬가지 분석 진행
- input 크기가 32X32로 작아져서 파라미터 수 줄인 ResNet 고안했음.
- 6n+2개 layer를 사용하였음

- 기존의 것과 비교했을 때 parameter의 수는 적지만 성능은 더 좋음을 확인할 수 있었음. (error percent 6.43%)
- data augmentation 등의 기법은 그대로 사용함
- weight decay = 0.0001 / momentum parameter = 0.9 / no dropout / 배치정규화, 가중치 초기화 / 학습 진행함에 따라 learning rate은 줄여나갔음.

⇒ Layer가 깊어질수록 좋은 성능이 나타나는 것을 확인할 수 있었음. ImageNet과 MNIST dataset 모두에 대해

- Response 값 확인 (???response 값이 무엇인가???)
- residual function이 non-resi인 것에 비해 좀 더 0에 가깝게 optimization 되었다??
- layer가 불필요한 수준으로 깊어지면 성능이 떨어질 수 있다. (overfitting)
- Object detection과 segmentaion에 대해서도 마찬가지로 test (VGG 뼈대로)

5. 결론 (배운점)

연구의 의의 및 한계점, 본인이 생각하는 좋았던/아쉬웠던 점 (배운점)

- 의의 : 깊은 신경망도 잘 학습할 수 있도록 해주었음
- 한계 : 여타 다른 모델처럼 overfitting에 대해서 최적의 방향을 제시한 건 아님(굳이 한계를 찾자면)
- 입력값을 더해주는 간단한 방식이 모델의 성능에 큰 영향을 미칠 수 있음을 알게 되었음.
- skip connection이 수학적으로는 기울기 소실 문제를 완화해주는 쪽으로 해석이 되었던 것 같은데, 연구자들이 그게 중요 포인트가 아니라고 한다면 수학적으로는 어떻게 해석해야 할지 궁금함.
- 활성함수를 relu로 사용해도 비슷한 문제가 발생하나?
- 시작은 기울기 소실 문제에서 시작한 것 같은데, 실험에서 과정을 분석하면서 기울기 소실 문제가 원인이 아니라 convergence가 문제라고 원인 규명하였음. 아닌가 기울기 소실 문제도 같이 해결한건가...?

6. 공부 참고 자료

- https://www.youtube.com/watch?v=_bIFagKJhks&t=30s

- https://www.youtube.com/watch?v=Yf_-bj2Zaq4
- https://www.youtube.com/watch?v=671BsKl8d0E
- 역전파 다시 깊이 학습하기