

위해 auxiliary classifier를 사용하였다. 모든 layer를 통과한 output에 대해서만 역전파를 계산하는 것이 아니라, 중간에 2개의 auxiliary classifier를 사용하여 중간 과정의 output에 대해서도 역전파를 계산할 수 있게 해준다. 중간 과정 output의 역전파 계산 과정에서는 gradient가 아직 덜 작아져서 0에 가깝지 않기 때문에 정상적으로 초반 layer에 대해서 가중치 업데이트를 진행할 수 있게 해준다. (GoogLeNet 모델에서 도출되는 최종 output에는 auxiliary classifier를 통한 중간 output이 사용되지 않는다.)

D. ResNet

ResNet은 VGG-19를 토대로 convolution layer를 더 쌓아서 만든 모델이다.

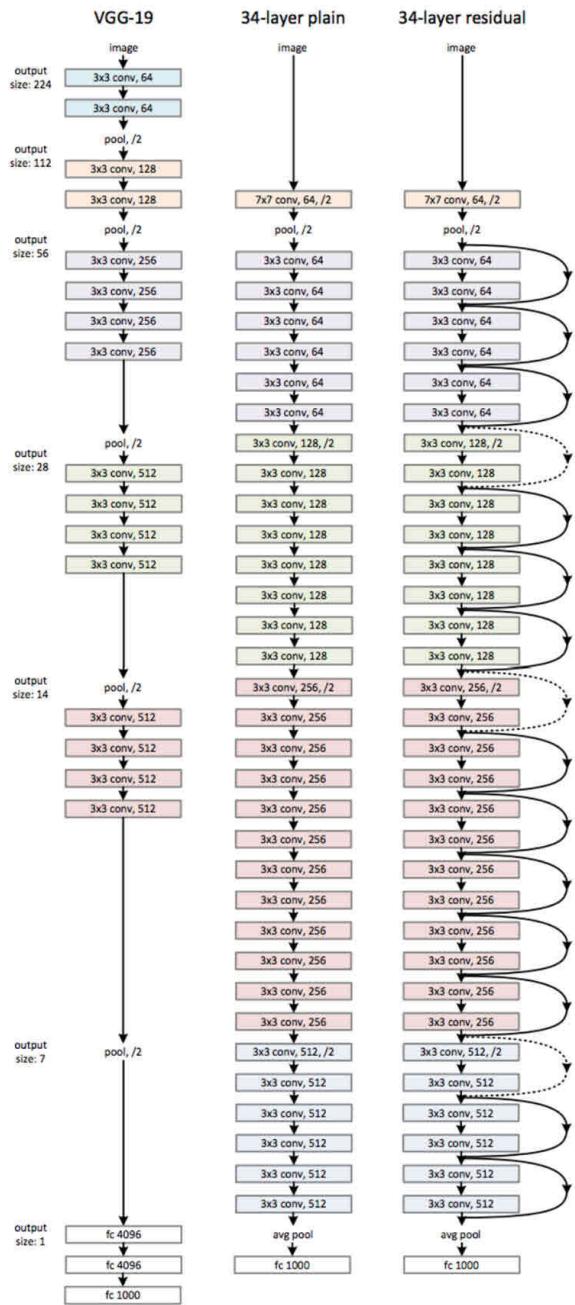


Fig. 6. ResNet의 구조

Figure 6의 34-layer plain의 구조가 VGG-19에서 con-

volution layer를 추가한 것인데, 실제 ResNet의 구조는 34-layer residual의 형태이다. ResNet에 사용된 residual block은 layer를 깊게 쌓으면서 발생하는 gradient vanishing을 방지하는 역할을 하게 된다. Residual block은 convolution layer를 통과한 결과값에 input 값인 x 를 더해주는 과정을 의미한다.

$$H(x) = F(x) + x \quad F(x) = H(x) - x$$

위와 같은 식에서 $H(x)$ 는 input x 를 넣었을 때 출력되는 결과를 의미하고, $F(x)$ 는 convolution layer를 통과한 중간 결과값을 의미한다. 이 때 $H(x)$ 를 x 와 같게 해주는, 즉 $F(x)$ 를 0에 가깝게 만드는 identity mapping을 적용하고자 residual block을 사용하였고, $F(x)$ 는 곧 $H(x)$ 와 x 의 차이인 잔차가 되기 때문에 residual block이라고 부르게 되었다.

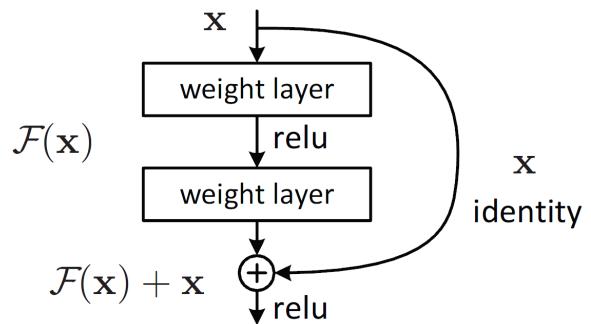


Fig. 7. Residual Block

ResNet에서 역전파를 계산할 때, $H(x)$ 에 대해 미분하여 계산하게 된다. $H(x)$ 는 $F(x) + x$ 이고 $F(x) + x$ 를 미분하면 어떤 값($f'(x)$) + 1이 되어서 1이 항상 남게되기 때문에 gradient vanishing을 해결할 수 있다.

E. Vision Transformer

Deep Learning을 활용한 이미지 분류 연구에서 CNN 기반 모델이 가장 많이 활용되었다. AlexNet 이후로 다양한 모델과 개선 방향들이 제시 되었고 높은 성능 평가 지수를 나타내었다. 하지만 Transformer 기반 모델인 Vision Transformer(ViT)의 발표 이후 ImageNet / ImageNet-ReaL / CIFAR-10 / VTAB에서 최상의 성능을 달성하였다. CNN 계열의 모델은 convolution filter size를 이용하여 학습하지만, Transformer 기반 모델인 Vision Transformer는 Transformer의 self-attention과 self-embedding을 활용한다. ViT의 구조는 figure 8과 같다.

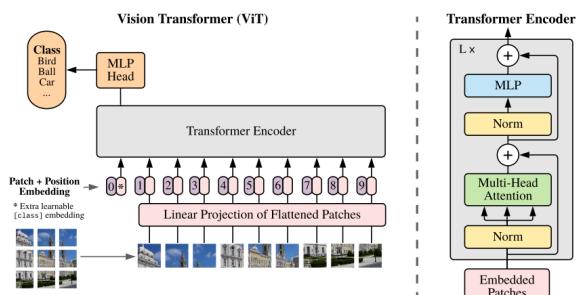


Fig. 8. Vision Transformer 구조

Input 이미지를 고정된 크기의 패치(patch)로 분할한 뒤, 각각의 패치를 선형적으로 임베딩(linear embedding)

