

[9차시 수업]

1. [Deep Learning] pre-training 과 fine-tuning (파인튜닝)

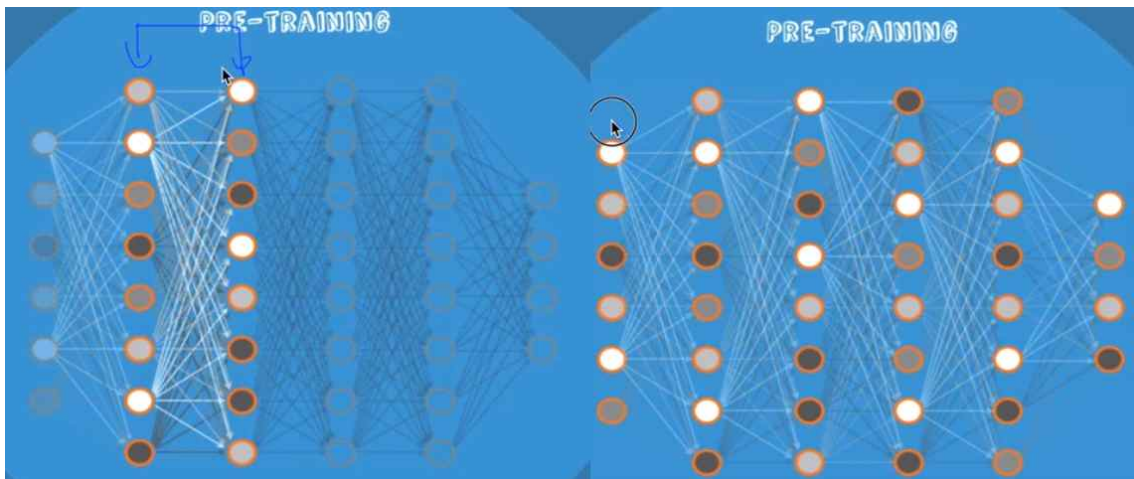
1) Pre Training 이란?

선행학습 = 사전훈련 = 전처리과정

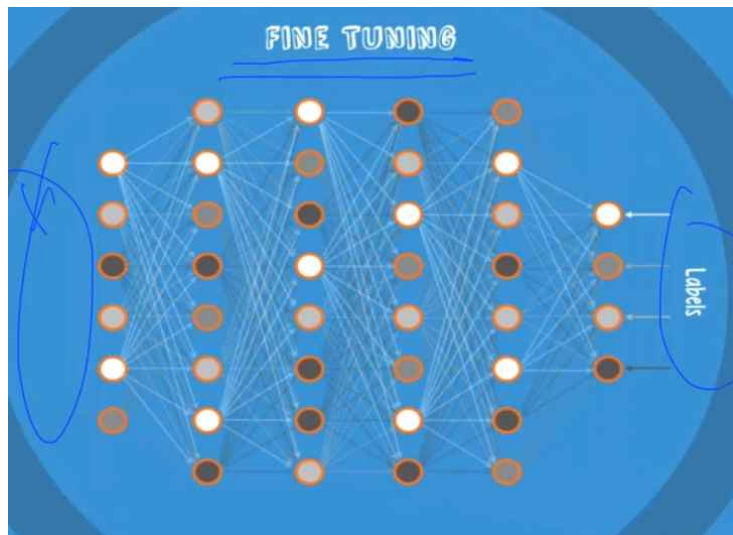
- Multi Layered Perceptron (MLP)에서 Weight와 Bias를 잘 초기화 시키는 방법이다.
- 이러한 Pre-training을 통해서 효과적으로 layer를 쌓아서 여러개의 hidden layer도 효율적으로 훈련 할 수 있다.
- unsupervised learning이 가능하기 때문에 레이블 되지 않은 큰 데이터를 넣어 훈련 시킬 수 있다는 점을 장점
- Drop-out, Mini-batch 방식을 사용하여 Pre-training을 생략하기도 한다.

2. Fine Tuning 이란?

- 기존에 학습되어져 있는 모델을 기반으로 아키텍처를 새로운 목적(나의 이미지 데이터에 맞게)변형하고 이미 학습된 모델 Weights로 부터 학습을 업데이트하는 방법
 - 모델의 파라미터를 미세하게 조정하는 행위
 - 파인튜닝을 했다고 말하려면 기존에 학습이 된 레이어에 내 데이터를 추가로 학습시켜 파라미터를 업데이트 해야 한다.
 - PRE - Training에서는 label도 필요없다. x값만 있으면 된다.
- 다음과 같이 2개의 layer씩 짝지어 weight들을 결정해준다. 이렇게 다 결정해 준 값을 초기화 값으로 사용한다.



- Fine Tunning에서는 x 데이터와 labels을 놓고 학습시킨다. (이미 weight이 잘 학습되어 있어 tuning이라고 함)



3. 가중치 초기화 (Weight Initialization)

딥러닝 학습에 있어 초기 가중치 설정은 매우 중요한 역할을 한다.

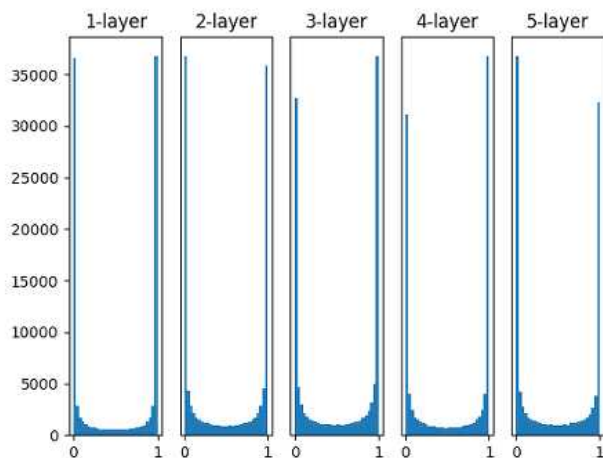
초기값 설정을 잘못해 문제가 발생하는 경우들을 살펴보자.

1) 초기값을 모두 0으로 설정한 경우

- 실제로 0으로 가중치를 초기화 한다면 모든 뉴런들이 같은 값을 나타낼 것이고, 역전파 과정에서 각 가중치의 update가 동일하게 이뤄질 것이다.
- 이러한 update는 학습을 진행해도 계속해서 발생할 것이며, 결국 제대로 학습하기 어려울 것이다.
- 동일한 update는 여러 층으로 나누는 의미를 상쇄시킨다.

2) 활성화 함수로 sigmoid 사용시 정규 분포 사용

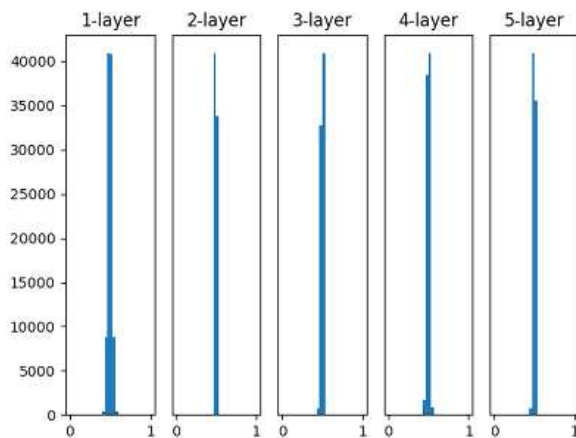
- sigmoid함수는 input의 절대값이 조금이라도 커지게 되면 미분값이 소실되는 문제가 발생한다.
- 이 경우에 평균 0이고 표준편차가 1인 정규분포를 따르도록 가중치를 랜덤하게 초기화 한다고 가정하자.



- 표준편차가 크기 때문에 학습을 반복할수록 가중치 값들이 0,1 로 치우치는 문제 발생한다.(Gradient Vanishing)
- Activation Function을 바꿈으로써 해결 할 수도 있겠지만, 가중치 초기화를 잘 설정함으로써 어느 정도 해결할 수 있다.

3) 2의 case에서 표준편차를 줄였을 경우

- 2의 문제를 확인하고 표준편차가 커 |x|값이 커지면서 기울기가 소실되는 문제를 확인했기 때문에, 표준편차를 줄여서 |x|값을 줄이려는 생각을 가지고 표준편차를 0.01로 설정한다고 가정했을 때 또 다른 문제발생



- 이렇게 표준편차를 적게 하면 층이 깊어질수록 가중치 값들이 중간 값인 0.5 부근에 몰리는 문제가 발생

■ 가중치 초기화의 또 다른 방법

1) Xavier Initialization

- 이전 노드와 다음 노드의 개수에 의존하는 방법
- Uniform 분포를 따르는 방법과 Normal분포를 따르는 두가지 방법이 사용된다.(Glorot & Bengio, AISTATS 2010)
- Xavier Normal Initialization

$$W \sim N(0, Var(W))$$

$$Var(W) = \sqrt{\frac{2}{n_{in} + n_{out}}}$$

(n_{in} : 이전 layer(input)의 노드 수, n_{out} : 다음 layer의 노드 수)

○ Xavier Uniform Initialization

$$W \sim U\left(-\sqrt{\frac{6}{n_{in} + n_{out}}}, +\sqrt{\frac{6}{n_{in} + n_{out}}}\right)$$

(n_{in} : 이전 layer(input)의 노드 수, n_{out} : 다음 layer의 노드 수)

- Xavier함수는 비선형함수(ex. sigmoid, tanh)에서 효과적인 결과를 보여준다.
- 하지만 ReLU함수에서 사용 시 출력 값이 0으로 수렴하게 되는 현상을 확인 할 수 있다.
따라서 ReLU함수에는 또 다른 초기화 방법을 사용해야 한다.

2) He Initialization

- ReLU를 활성화 함수로 사용 시 출력값이 0으로 수렴하는 비효율적인 결과의 해결방법
- 초기화 방법을 He initialization을 사용
- 정규분포와 균등분포 두가지 방법이 사용
- He Normal Initialization

$$W \sim N(0, Var(W))$$

$$Var(W) = \sqrt{\frac{2}{n_{in}}}$$

(n_{in} : 이전 layer(input)의 노드 수)

- He Uniform Initialization

$$W \sim U(-\sqrt{\frac{6}{n_{in}}}, +\sqrt{\frac{6}{n_{in}}})$$

(n_{in} : 이전 layer(input)의 노드 수)

4. Bias 초기화

가중치 초기화 뿐만 아니라 편향(bias) 초기값 또한 초기값 설정 또한 중요하다.

보통의 경우에는 Bias는 0으로 초기화 하는 것이 일반적이다. ReLU의 경우 0.01과 같은 작은 값으로 b를 초기화 하는 것이 좋다는 보고도 있지만 모든 경우는 아니라 일반적으로는 0으로 초기화 하는 것이 효율적이다.

5. Conclusion

- 초기값 설정에는

Sigmoid, tanh 경우 Xavier 초기화 방법이 효율적이다.

ReLU계의 활성화 함수 사용 시 He 초기화 방법이 효율적이다.

- 최근의 대부분의 모델에서는 He초기화를 주로 선택한다.
- 최근의 Deep CNN 모델들은 주로 Gaussian Distribution을 따르는 가중치 초기화 방법을 사용한다.

