딥러닝 기초2

https://www.youtube.com/watch?v=02Cvzbd3Lck

Contents

CryptoCraft LAB

경사법

- 최적 매개변수= 손실함수의 최솟값을 찾는 것

But. 손실 함수 복잡 -> 기울기를 이용하여 손실함수의 최솟값을 찾는 것

= 경사법(경사 하강법)

- 기울기:

함수에서 모든 변수의 편미분을 벡터로 정리한 것

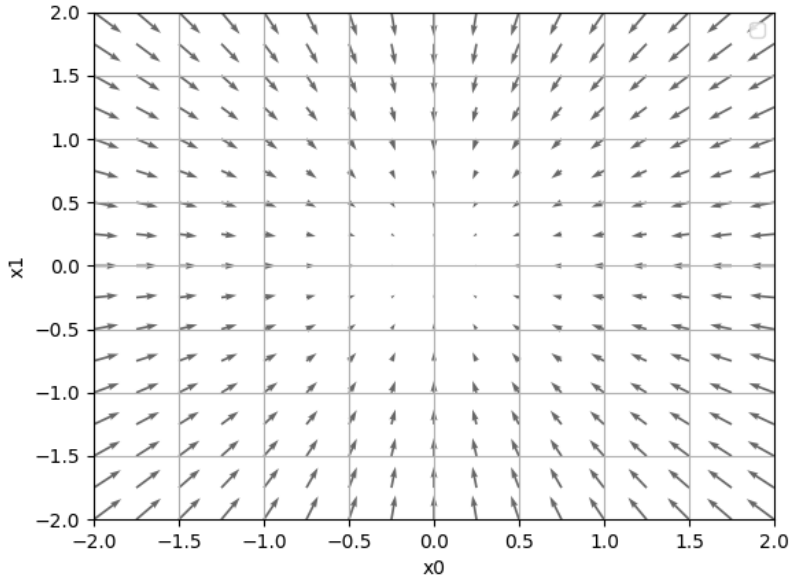
- 신경망에서의 기울기

: 가중치 매개변수에 관한 손실 함수의 기울기

$$W = \begin{pmatrix} w_{11} & w_{21} & w_{31} \\ w_{12} & w_{22} & w_{32} \end{pmatrix}$$

$\frac{\partial L}{\partial w_{11}}$ 은 W_{11} 을 조금 변경했을 때 손실 함수 L이 얼마나 변화하느냐

$$\frac{\partial L}{\partial W} = \begin{pmatrix} \frac{\partial L}{\partial w_{11}} & \frac{\partial L}{\partial w_{21}} & \frac{\partial L}{\partial w_{31}} \\ \frac{\partial L}{\partial w_{11}} & \frac{\partial L}{\partial w_{22}} & \frac{\partial L}{\partial w_{32}} \end{pmatrix}$$

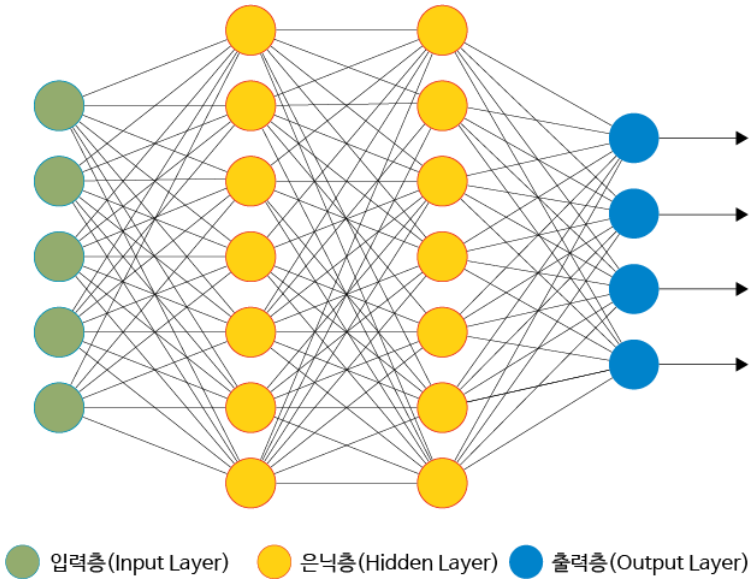


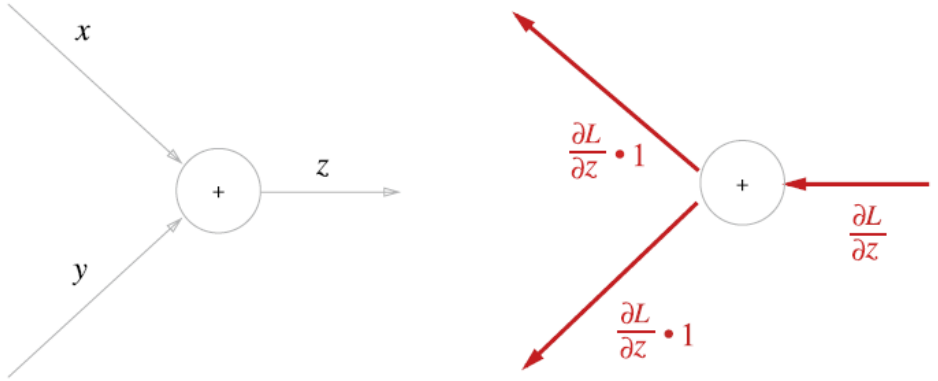
기울어진 방향이 꼭 최솟값 가리키는 것은 아님

그 방향으로 가야 함수 값 줄임.

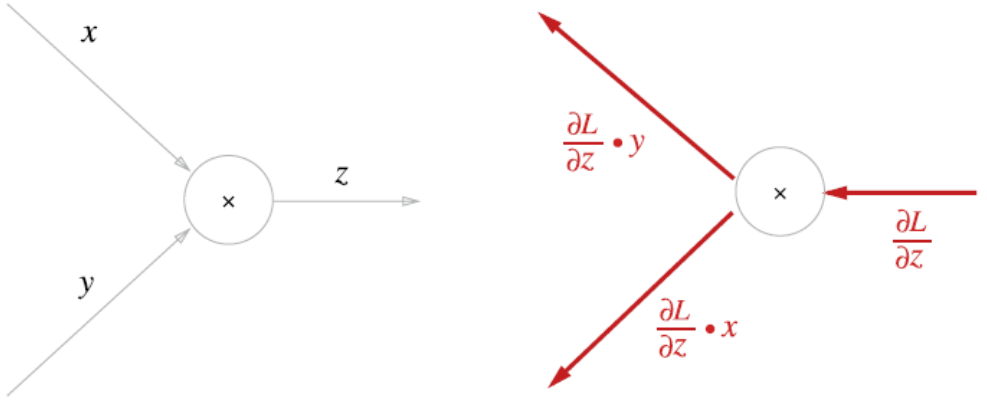
오차역전파

가중치 매개변수 기울기를 효율적으로 계산





덧셈 노드에서의 역전파



곱셈 노드에서의 역전파

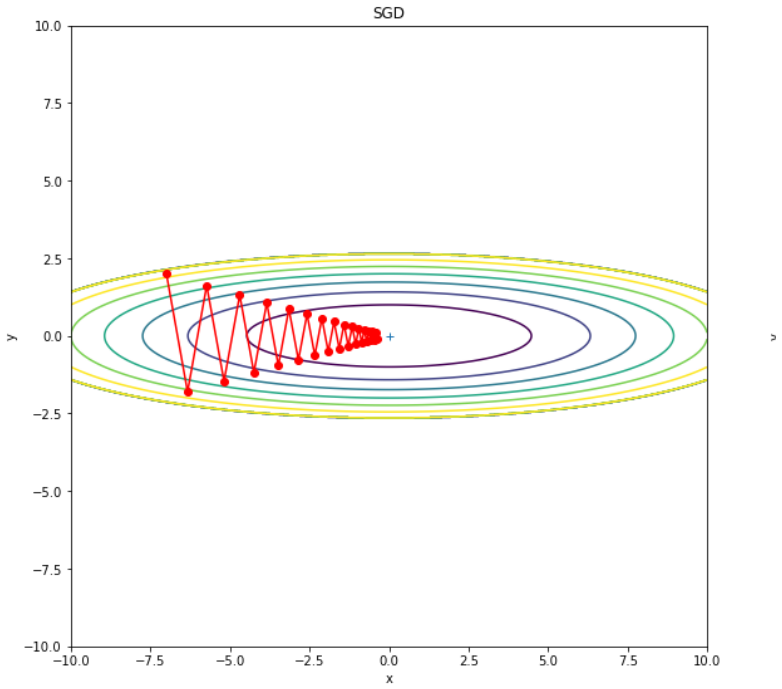
매개변수 최적화

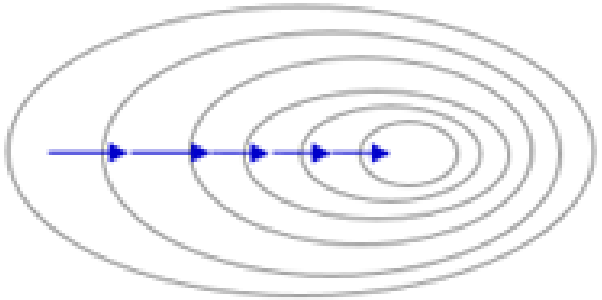
- SGD(확률적 경사하강법)

랜덤하게 추출한 데이터를 사용.

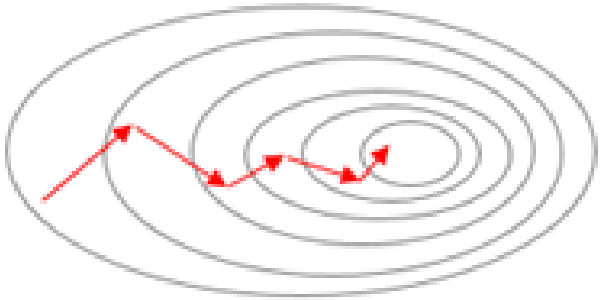
$$W = W - \eta \frac{\partial L}{\partial W}$$

(W:가중치 매개변수, η: 학습률 $\frac{\partial L}{\partial W}$ 기울기)





경사 하강법



SGD

단점: 비등방성 함수(기울기가 달라지는 함수) 에서는 탐색 경로가 비효율적

매개변수 최적화

- 모멘텀

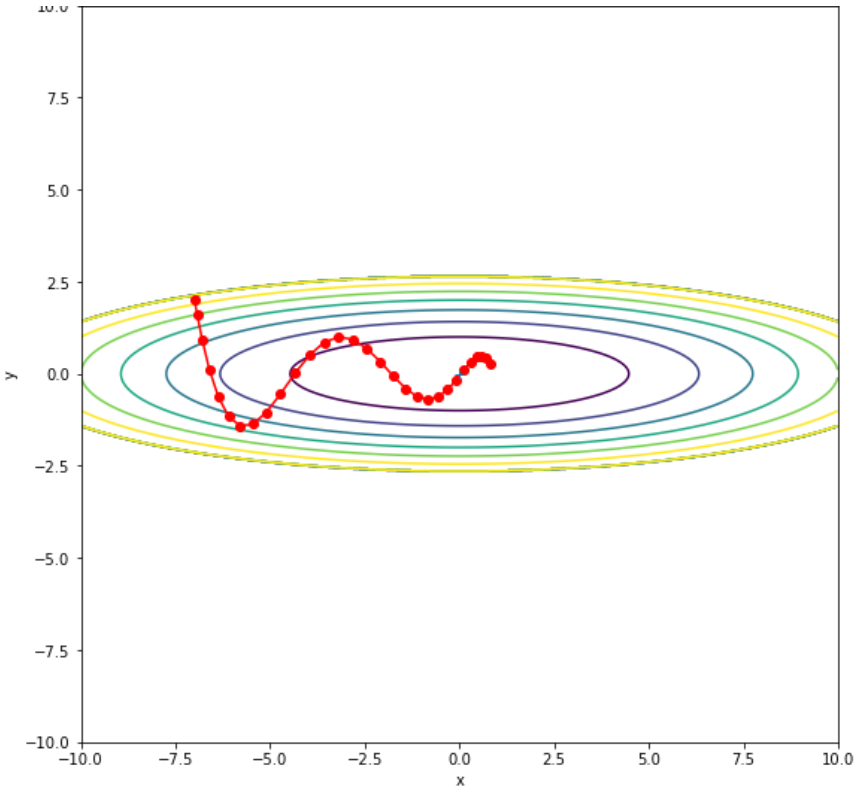
  기울기에서 속도의 개념이 추가

  속도가 크게 나올수록 기울기가 크게 업데이트 되어 확률적 경사하강법이 가지는 단점을 보완

  한 방향으로 일정하게 가속

$$v = av - \eta \frac{\partial L}{\partial W}$$
$$W = W + v$$

(W:가중치 매개변수, η: 학습률 ,
$\frac{\partial L}{\partial W}$:기울기,v:속도)



매개변수 최적화

- AdaGrad
  - 개별 매개변수에 적응적으로 학습률을 조정하면서 학습 진행
  - 학습률 감소 : 학습을 진행하면서 학습률을 점차 줄여가는 방법 => 학습률을 낮추는 간단한 방법= 매개변수 '전체'의 학습률 값을 일괄적으로 낮추는것

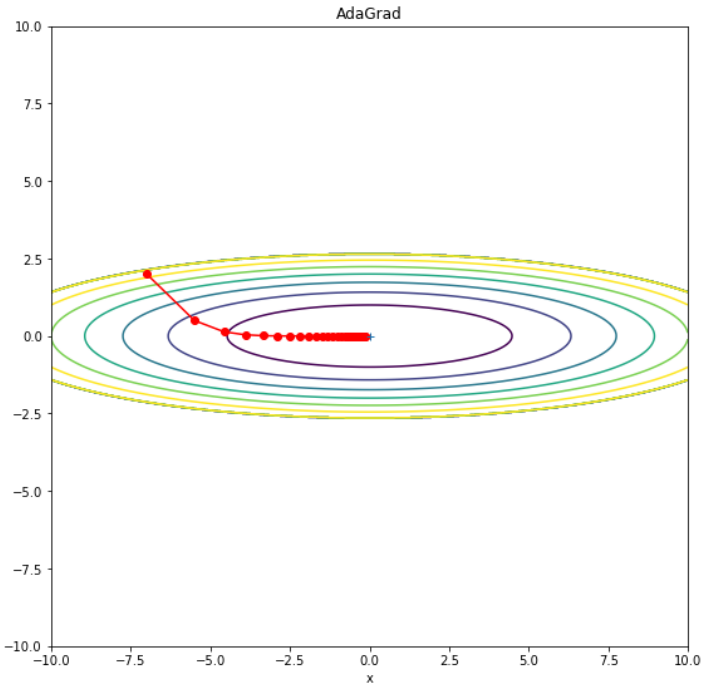
$$h = h + \frac{\partial L}{\partial W} \odot \frac{\partial L}{\partial W}$$

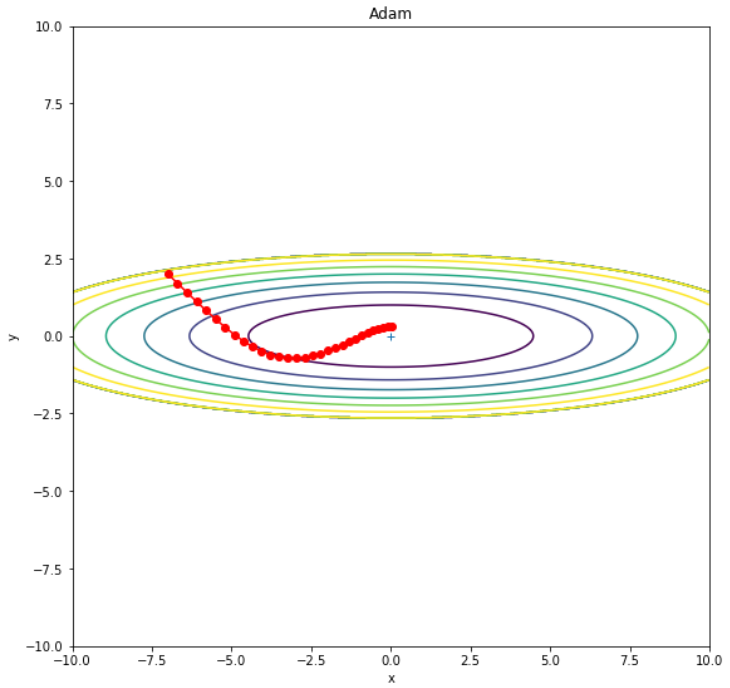
$$W = W - \eta \frac{1}{\sqrt{h}} \frac{\partial L}{\partial W}$$

(W=가중치 매개변수, $\frac{\partial L}{\partial W}$: 기울기 , η: 학습률 , $\odot$: 행렬의 원소별 곱셈) ,

$\frac{1}{\sqrt{h}}$ 으로 학습률 조정

- Adam :
  모멘텀 + AdaGrad
  편향 보정

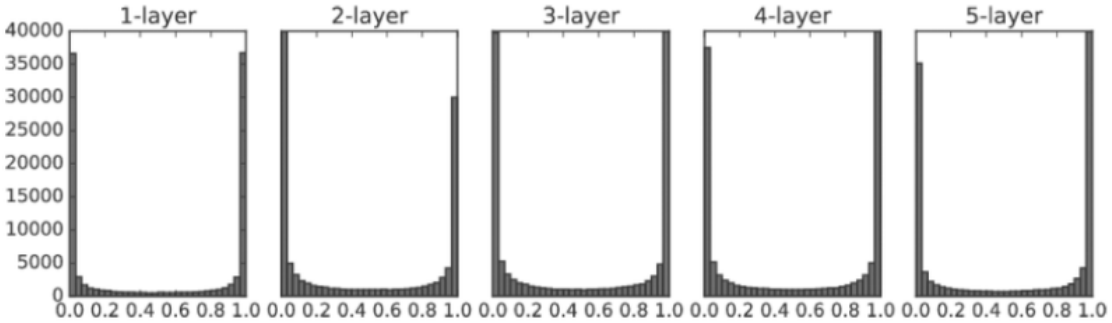




가중치 초기값

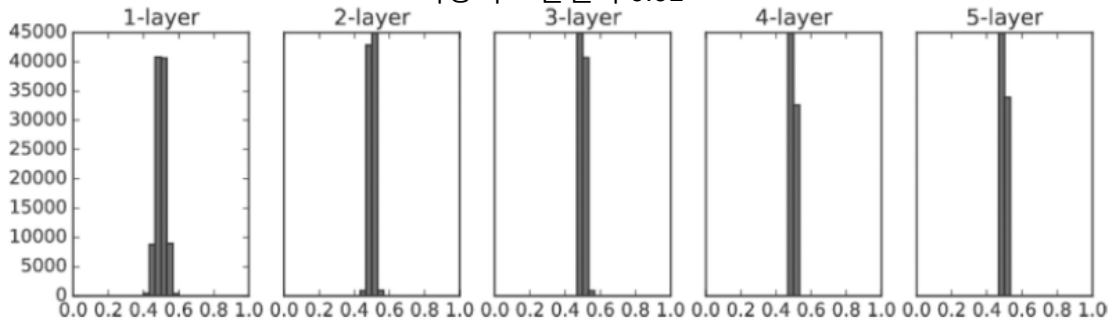
- 가중치 값을 작게해 오버피팅 억제
- " 가중치 초기값을 0" => 역전파 때 두번째 층 가중치가 모두 똑같이 갱신 -> 계속해서 같은 값 유지.
- 가중치 초깃값에 따라 활성화 값 분포 달라짐.
- 각 층의 활성화 값 적당히 고루 분포되어야함.(층과 층 사이에 적당하게 다양한 데이터 흐르게함.) -> 신경망 학습 효율적.

가중치 표준편차 1



기울기 소실

가중치 표준편차 0.01



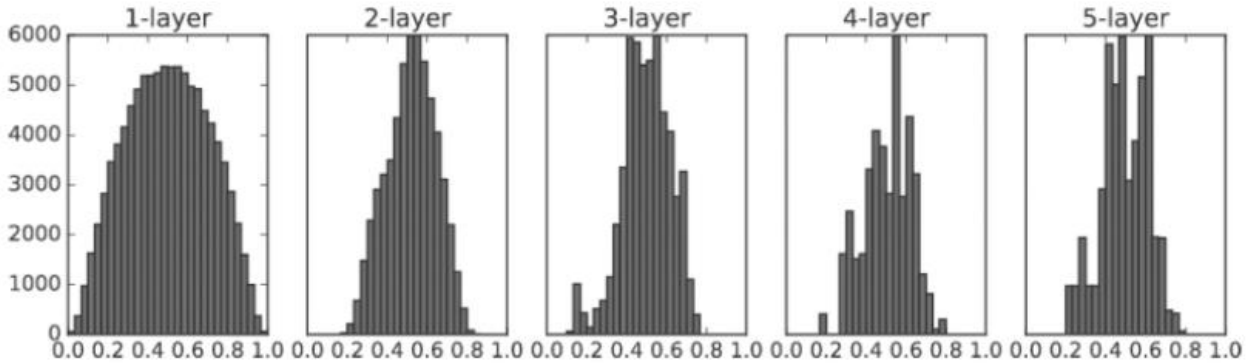
표현력 제한 = 다수의 뉴런이 같은 값을 가져 여러 개의 뉴런을 두는 의미가 없음

가중치 초기값

- **Xavier 초깃값**

일반적인 딥러닝 프레임워크들이 표준적으로 이용중

앞 계층의 노드가 n개라면 표준편차가$\frac{1}{\sqrt{n}}$인 분포 사용



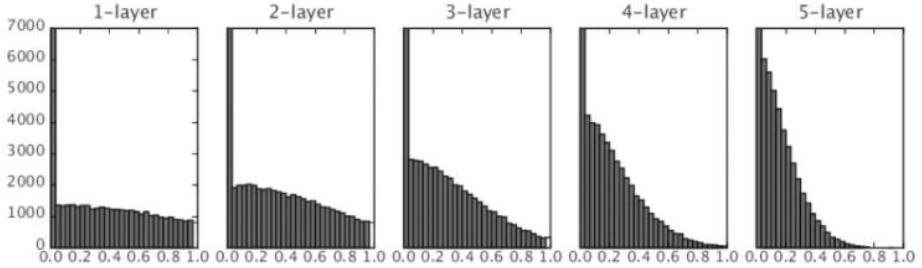
Xavier 초깃값은 활성화 함수가 선형인 것을 전제

sigmoid 함수와 tanh 함수는 좌우 대칭이라 '중앙 부근'이 선형인 함수

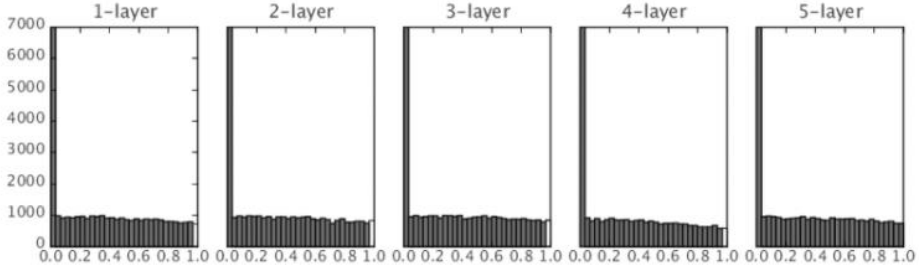
- **He 초깃값**

ReLU에 특화된 초깃값

앞 계층의 노드가 n개라면 표준편차가$\sqrt{\frac{2}{n}}$인 분포 사용



Xavier 초깃값을 사용한 경우



He 초깃값을 사용한 경우

오버피팅

•오버피팅 발생이유

- 매개변수가 많고 표현력이 높은 모델
- 훈련 데이터가 적음

•가중치 감소

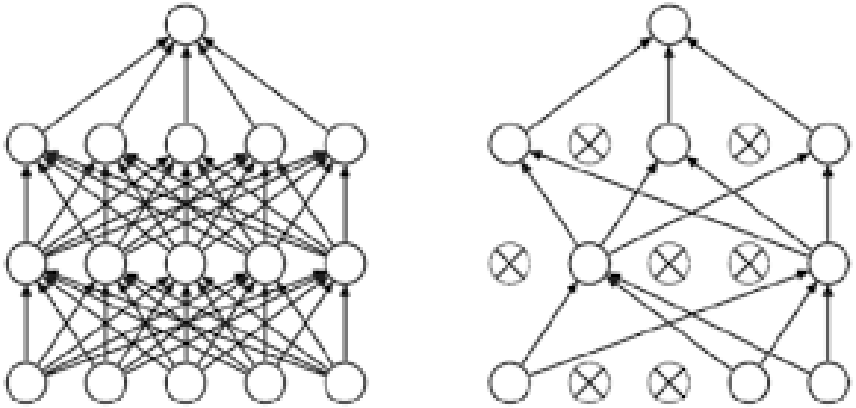
오버피팅 억제하는 기법 에는 가중치 감소 기법

(오버피팅이 가중치 매개변수 값이 커서 생기는 경우가 많음)

큰 가중치에 대해서는 패널티를 부과하여 오버피팅 억제

• 드롭아웃

학습 시 뉴런을 임의로 삭제하는 기법



(a) 일반 신경망 (b) 드롭아웃을 적용한 신경망

하이퍼 파라미터

- 가중치, 편향 - 파라미터
- Ex. 뉴런 수 , 배치 크기, 매개변수 갱신 시학습률과 가중치 감소 등 사람이 직접 조정

- 하이퍼 파라미터 검증 시 시험 데이터 사용 X
- => 시험 데이터를 사용하여 하이퍼 파라미터를 조정하면 하이퍼 파라미터 값이 시험 데이터에 오버피팅 됨. -> 다른 데이터에는 적용 x, 가능성 높음

- 검증 데이터(validation data) 필요

훈련 데이터(Training data)	검증 데이터(Validation data)	시험 데이터(Test data)
매개변수 학습	하이퍼 파라미터 성능 평가	신경망 범용 성능 평가

Q & A