

# 핸즈온 머신러닝

## Chapter 11.6: 연습문제

Presentation: **Jeiyoon Park**  
6<sup>th</sup> Generation, TAVE

# Outline

1. 연습 문제
2. 연습 과제

# Outline

1. 연습 문제
2. 연습 과제

## 11.6: 연습문제

1. He 초기화를 사용하여 무작위로 선택한 값이라면 모든 가중치를 같은 값으로 초기화 해도 괜찮을까요?

## 11.6: 연습문제

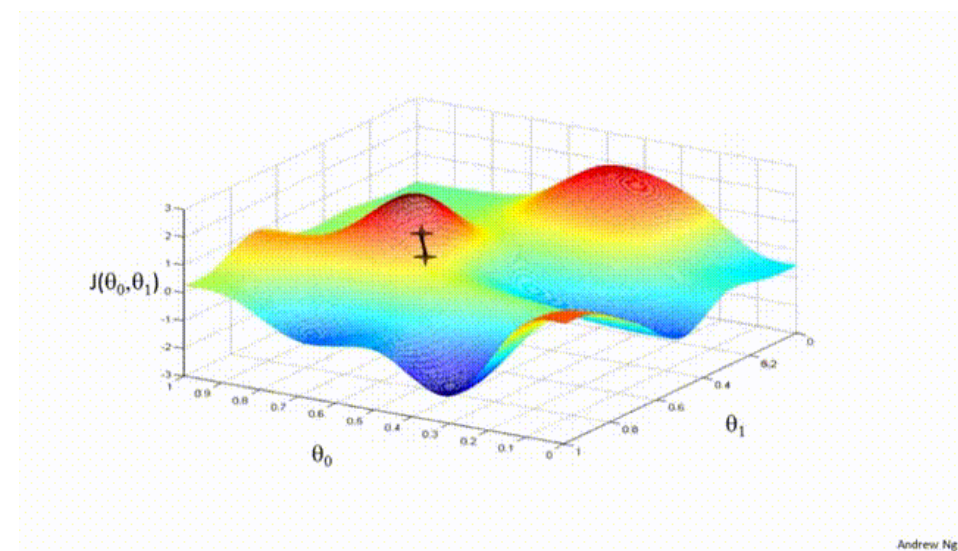
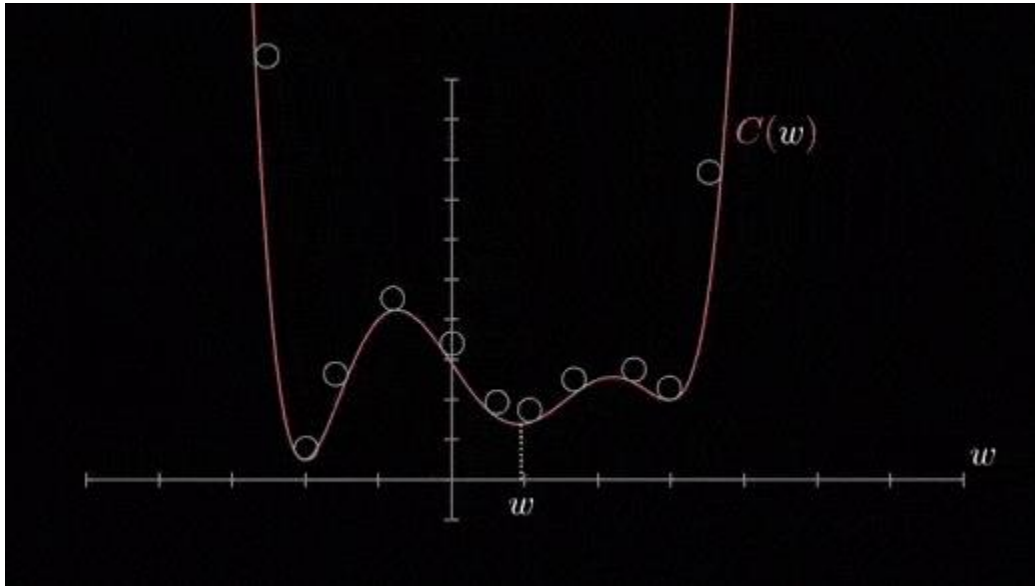
1. He 초기화를 사용하여 무작위로 선택한 값이라면 모든 가중치를 같은 값으로 초기화 해도 괜찮을까요?



## 11.6: 연습문제

1. He 초기화를 사용하여 무작위로 선택한 값이라면 모든 가중치를 같은 값으로 초기화 해도 괜찮을까요? **No**

- 모든 가중치는 무작위로 & 독립적으로 샘플링 되어야 함.
- 층마다 모든 뉴런이 항상 같은 초기값을 가지면 이는 층마다 하나의 뉴런이 있는 것과 같음. 따라서 수렴도 오래걸리고 절대 좋은 성능이 나오지 않음



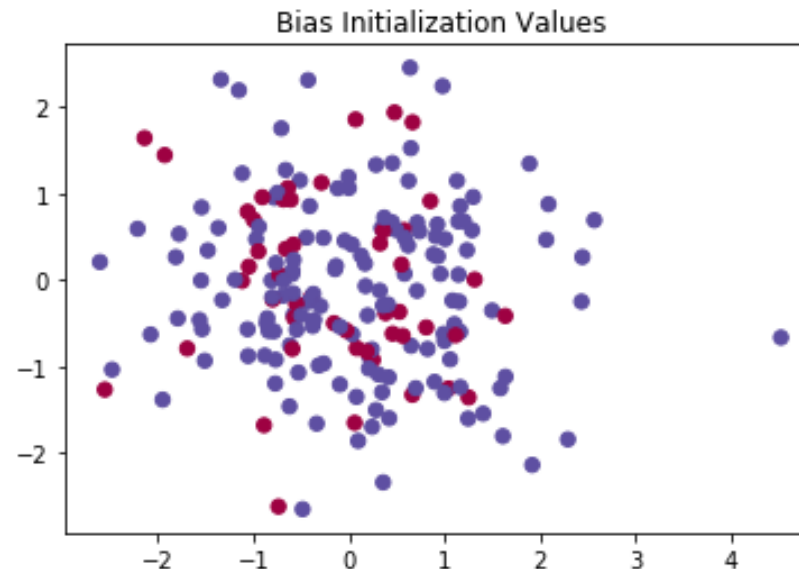
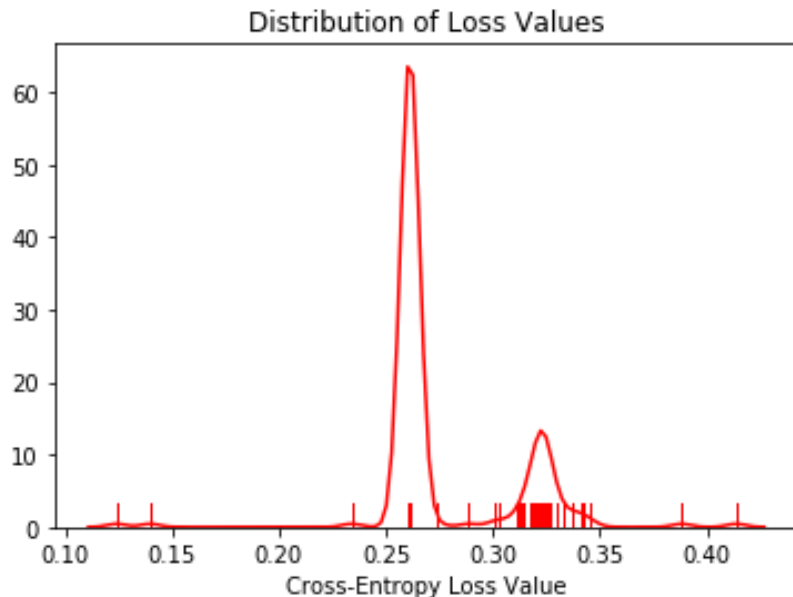
## 11.6: 연습문제

2. 편향 (bias)을 0으로 초기화 해도 괜찮을까요?

# 11.6: 연습문제

## 2. 편향 (bias)을 0으로 초기화 해도 괜찮을까요?

- 초기 연구에서는 ReLU 때문에 시작부터 동작하지 않는 죽은 뉴런(dead neuron)들을 방지하기 위해 편향치로 0.01, 0.1 정도의 값을 사용했었지만,
- 훈련 기법이 발전한 지금은 거의 언제나 0으로 초기화함.







## 11.6: 연습문제

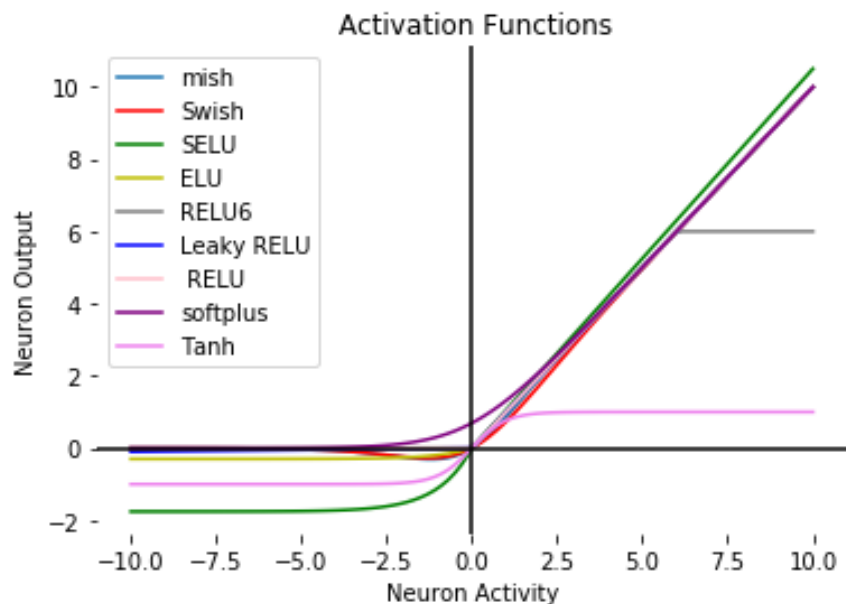
3. ReLU보다 SELU 활성화 함수가 나은 세 가지는 무엇인가요?



## 11.6: 연습문제

### 3. ReLU보다 SELU 활성화 함수가 나은 세 가지는 무엇인가요?

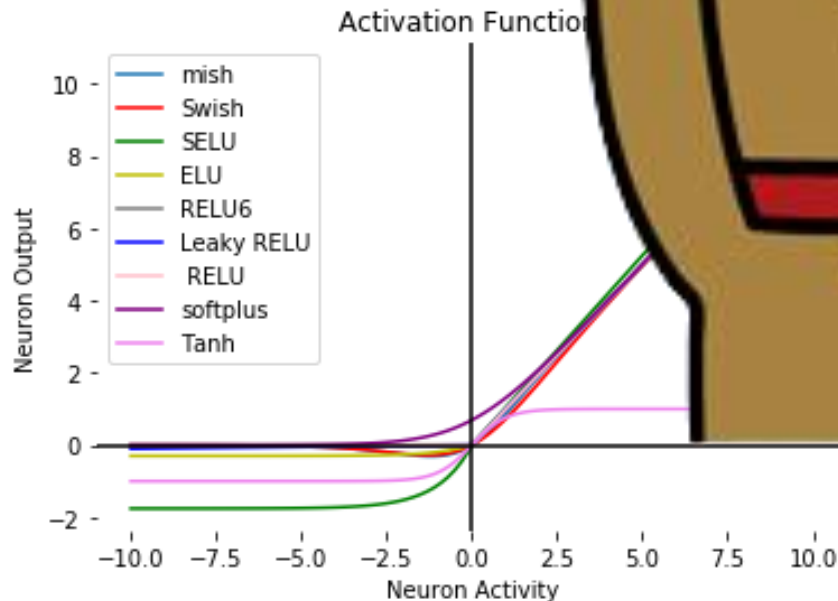
- 음수 출력 가능 -> 그래디언트 소실 문제 완화시켜줌
- 미분값이 항상 0이 아님 -> ReLU에서 발생하는 죽은 뉴런 효과를 줄여줌
- 모델이 자기 정규화가 되어 그래디언트 폭주나 소실문제 완화시켜줌



## 11.6: 연습문제

### 3. ReLU보다 SELU 활성화 함수가 나은 세 가지는 무엇인가요?

- 음수 출력 가능 -> 그래디언트 소실 문제 완화시켜줌
- 미분값이 항상 0이 아니므로 죽은 뉴런 효과를 줄여줌
- 모델이 자기 정규화가 잘 일어나 소실문제 완화시켜줌





## 11.6: 연습문제

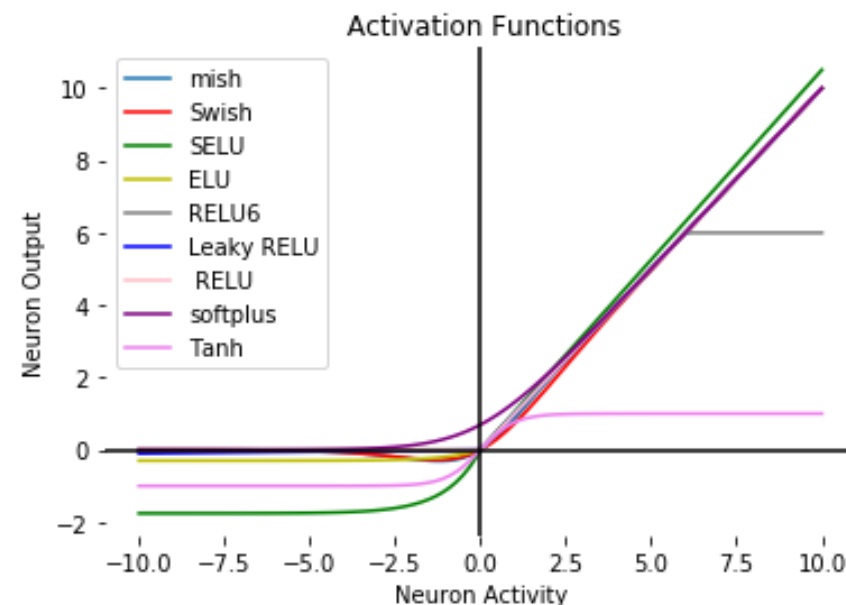
4. 언제 SELU, LeakyReLU, ReLU, tanh, 로지스틱, 소프트맥스와 같은 활성화 함수를 사용해야 하나요?



## 11.6: 연습문제

4. 언제 SELU, LeakyReLU, ReLU, tanh, 로지스틱, 소프트맥스와 같은 활성화 함수를 사용해야 하나요?

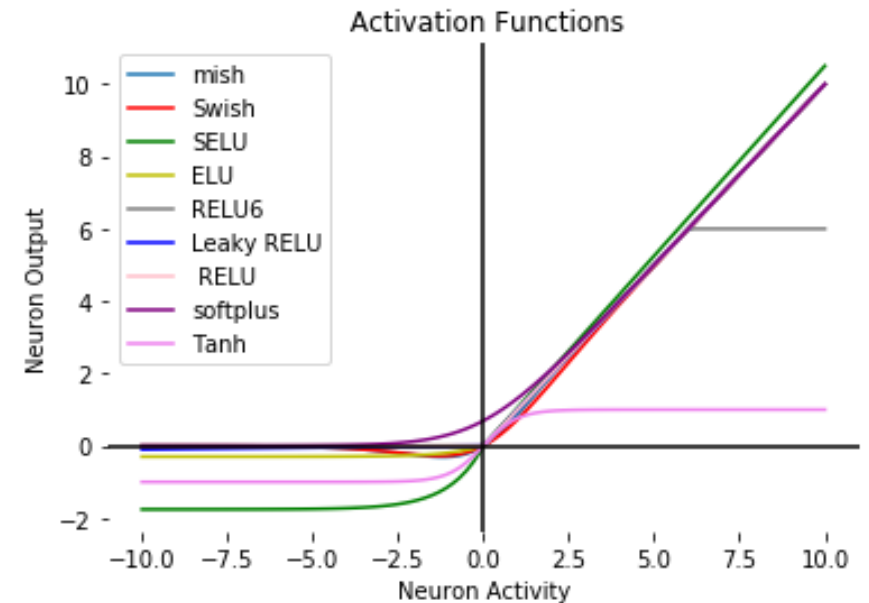
- SELU: 기본적으로 좋음
- LeakyReLU: 가능한 빠른 신경망을 원하면
- ReLU: 간단해서 선호
- tanh: -1 to 1 값을 가져야 하는 출력층
- logistic: hidden layer에서는 거의 사용하지 않음 (예외: Auto encoder)
- softmax: exclusive class들의 확률 출력층에 사용



## 11.6: 연습문제

4. 언제 SELU, LeakyReLU, ReLU, tanh, 로지스틱, 소프트맥스와 같은 활성화 함수를 사용해야 하나요?

- SELU: 기본적으로 좋음
- LeakyReLU: 가능한 빠른
- ReLU: 간단해서 선호
- tanh: -1 to 1 값을 가져
- logistic: hidden layer에
- (예외: Auto encoder)
- softmax: exclusive class를



## 11.6: 연습문제

5. SGD를 사용할때 momentum 하이퍼파라미터를 너무 1에 가깝게 하면(e.g. 0.99999) 어떤 일이 일어날까요?

## 11.6: 연습문제

5. SGD를 사용할때 momentum 하이퍼파라미터를 너무 1에 가깝게 하면(e.g. 0.99999) 어떤 일이 일어날까요?

1) Momentum?

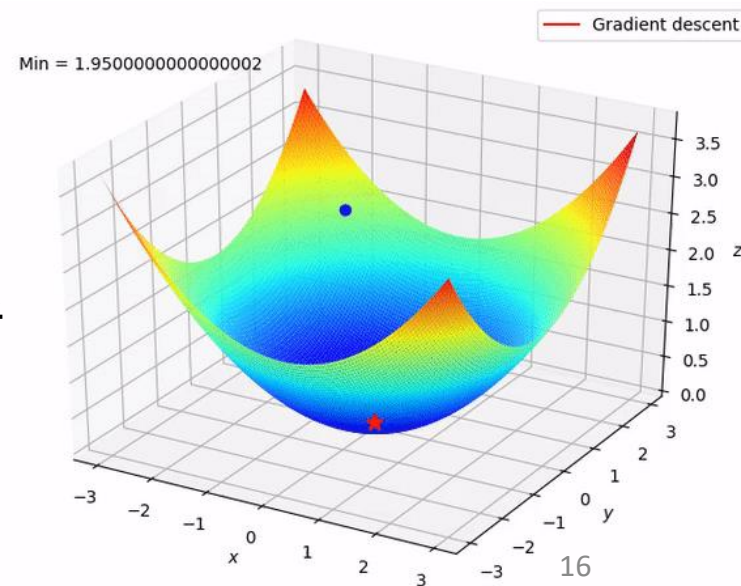
- 말 그대로 Gradient Descent를 통해 이동하는 과정에 '관성'을 주는 것

$$v_t = \gamma v_{t-1} + \eta \nabla_{\theta} J(\theta)$$

$$\theta = \theta - v_t$$

- 여기서  $\gamma$  는 momentum을 얼마나 줄 것인지에 대한 변수

- 즉, 이전 time step의 값을 얼마나 추가적으로 반영할지 결정



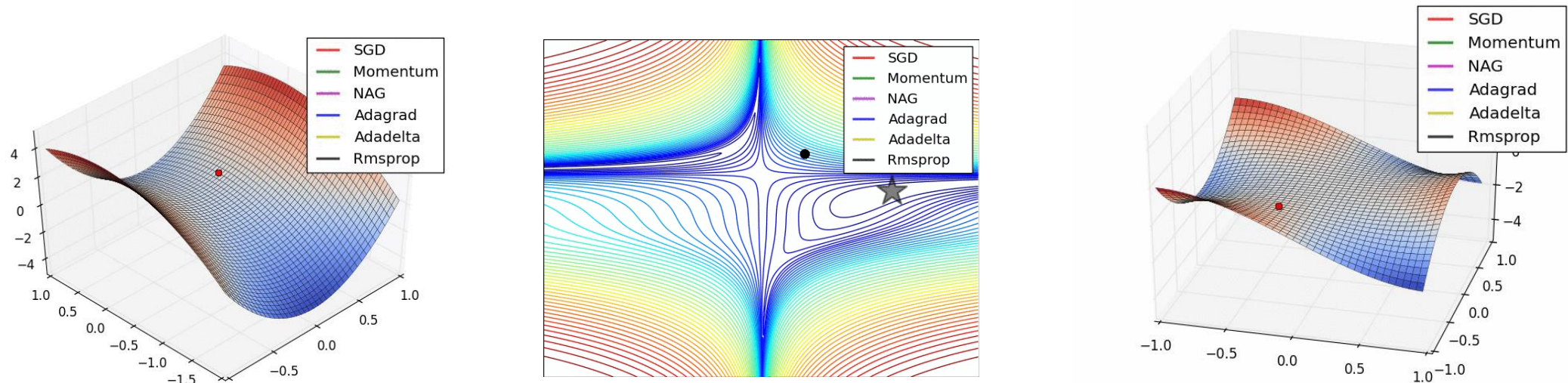


## 11.6: 연습문제

5. SGD를 사용할때 momentum 하이퍼파라미터를 너무 1에 가깝게 하면(e.g. 0.99999) 어떤 일이 일어날까요?

### 2) 문제점

- 지나치게 높은 관성을 주게되면 수렴전에 여러번 진동하게 되어 수렴에 훨씬 오래 걸리게 됨 (oscillation)





## 11.6: 연습문제

6. 희소 모델을 만들 수 있는 세 가지 방법은 무엇인가요?



## 11.6: 연습문제

6. 희소 모델을 만들 수 있는 세 가지 방법은 무엇인가요?

1) 희소 모델이란?

- 대부분의 가중치가 0인 모델 (sparse model)

2) 방법

- 훈련 후 작은 가중치를 0으로
- Regularization
- Model Optimization Toolkit (at Tensorflow)

## 11.6: 연습문제

6. 희소 모델을 만들 수 있는 세 가지 방법은 무엇인가요?

1) 희소 모델이란?

- 대부분의 가중치가 0인

2) 방법

- 훈련 후 작은 가중치를

- Regularization

- Model Optimization Tool



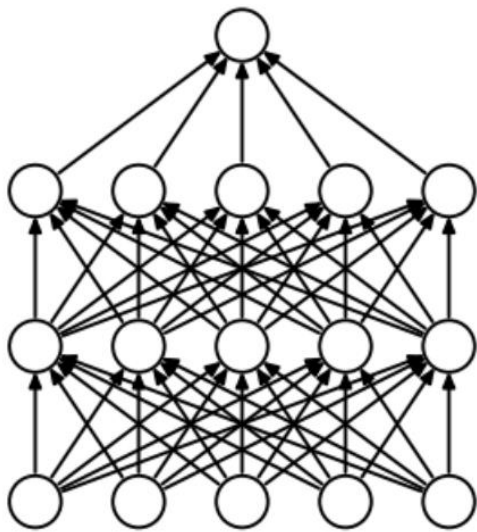
## 11.6: 연습문제

7. Dropout이 훈련 속도를 느리게 만드나요? Inference도 느리게 만드나요? MC 드롭아웃은 어떤가요?

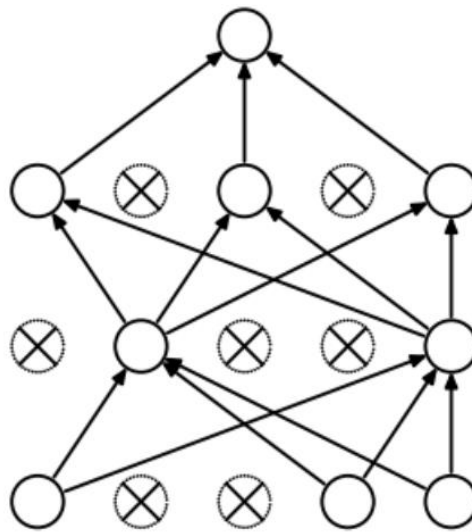
## 11.6: 연습문제

7. Dropout이 훈련 속도를 느리게 만드나요? Inference도 느리게 만드나요? MC 드롭아웃은 어떤가요?

- 1) 보통 dropout(e.g. 0.5)은 훈련속도를 느리게 만드는데 맞음.
- 2) Dropout은 학습시에만 적용하므로 inference 속도에 영향을 미치지 않음



(a) Standard Neural Net



(b) After applying dropout.

# 11.6: 연습문제

7. Dropout이 훈련 속도를 느리게 만드나요? Inference도 느리게 만드나요? MC 드롭아웃은 어떤가요?

3) MC dropout?

- [Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning \(Gal et al., ICML 2016\)](#)

- MC dropout (Monte Carlo dropout): Evaluation에서도 dropout을 사용하고 랜덤하게 적용하기 때문에 매 시행마다 결과가 조금 다름

- 이때, 모델은 uncertainty를 estimation할 수 있는 특성이 생김. 즉, 만약 학습에 사용된 data라면 몇번을 집어넣던지 표준편차가 낮을거고 그렇지 않은 데이터라면 표준편차가 매우 클거임

# 11.6: 연습문제

7. Dropout이 훈련 속도를 느리게 만드나요? Inference도 느리게 만드나요? MC 드롭아웃은 어떤가요?

3) MC dropout?

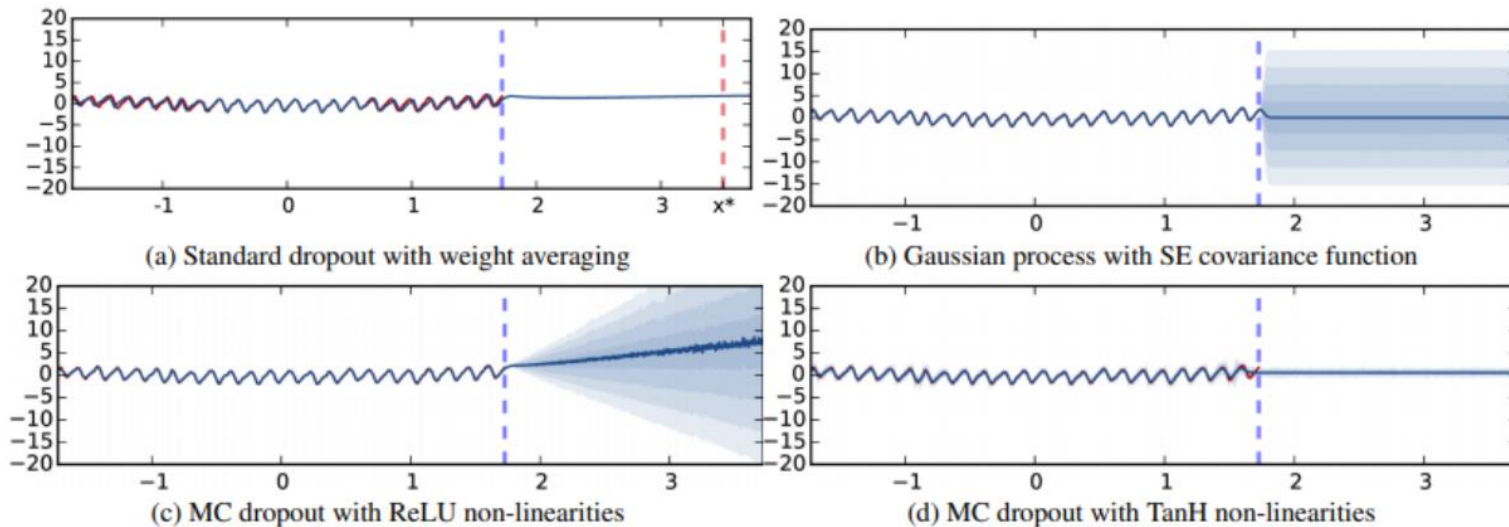


Figure 2. Predictive mean and uncertainties on the Mauna Loa CO<sub>2</sub> concentrations dataset, for various models. In red is the observed function (left of the dashed blue line); in blue is the predictive mean plus/minus two standard deviations (8 for fig. 2d). Different shades of blue represent half a standard deviation. Marked with a dashed red line is a point far away from the data: standard dropout confidently predicts an insensible value for the point; the other models predict insensible values as well but with the additional information that the models are uncertain about their predictions.

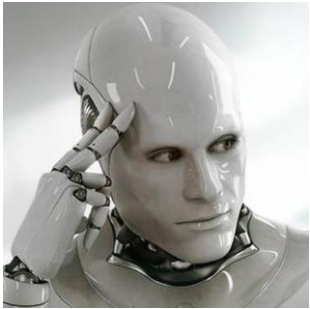


# Outline

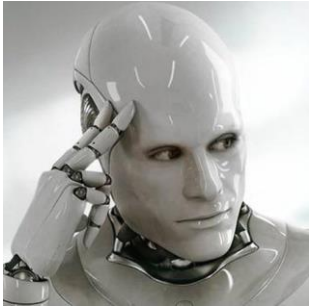
1. 연습 문제
2. 연습 과제

## 11.6: 연습과제: 링크

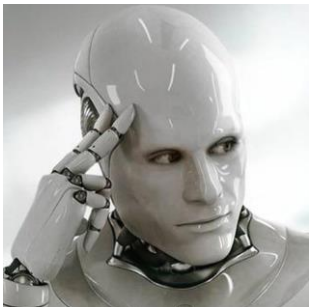
8. CIFAR10 이미지 데이터셋에 심층 신경망을 훈련해보세요.



CIFAR 10



CIFAR 10



CIFAR 10

# Thank you

 : <https://jeiyoongithub.io/>

 : <https://www.instagram.com/cloudwantsasnack/>