Differentiation

Chain Rule
(연쇄법칙)

소프트웨어 꼰대 강의

노기섭 교수

(kafa46@cju.ac.kr)

Course Overview

Topic	Contents
01. Orientation 오리엔테이션	Course introduction, motivations, final objectives 과정 소개, 동기부여, 최종 목표
02. Learning in deeplearning 딥러닝 학습	How does the deeplearing learns knowledge from data 어떻게 딥러닝은 데이터로부터 지식을 배우는가?
03. Principle of differentiation 미분의 원리	Basics of differentiation (concepts, notation, operations) 미분 기본지식 (개념, 표기, 연산)
04. Partial differentiation 편미분	Concept & operation of partial differenciation 편미분 개념, 연산
05. Gradient descent 경사 하강법	Concept, interpretation and learning in gradient descent 경사하강 알고리즘 개념, 해석 및 학습
06. Chain rule 연쇄법칙	Concept & operation of chain rule 연쇄법칙 개념 및 연산
07. Matrix differentiation 행렬미분	Partial differentiation in linear system 선형시스템에서의 편미분
08. Back propagation 역전파 학습	The mechanism of back propagation 역전파 학습의 작동 방법
09. Gradient vanishing 기울기 소실	Quick overview on activation function, cause root of gradient vanishing and its counter-measure 활성함수 간단 소개, 기울기 소실 근본원인과 대책

딥러닝에서의 Chain Rule

머리가 띵~~~ 해지는
수식이 등장합니다.

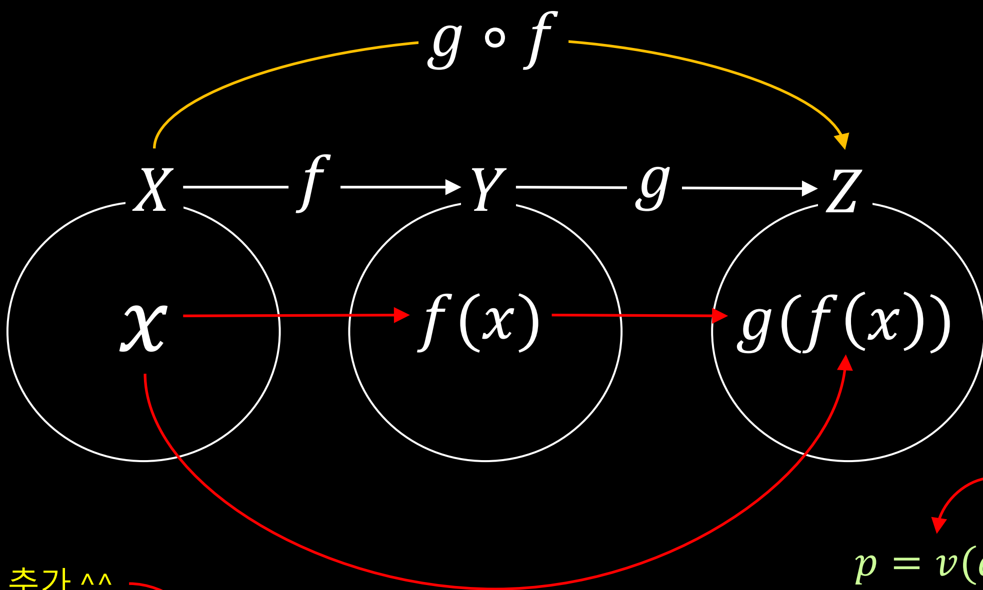
이미지 출처:
https://www.flaticon.com/kr/free-icon/horror_12355892
저작자: Flaticon

사실, 편미분을 안다면
매우 간단한 계산입니다.

사실... 여러분이 직접 계산할 일은 없습니다.
컴퓨터(프레임워크, 예: Pytorch, Tensorflow 등)가 해줍니다.

Chain Rule이 어떻게 쓰이는지 이해하는 것이 중요합니다 ^^.

합성 함수의 미분



$$y = f(x)$$

$$z = g(y) = g(f(x))$$

$$(g \circ f)' = g'\big(f(x)\big) \cdot f'(x) = \frac{dz}{dy} \cdot \frac{dy}{dx}$$

추가 ^^

$$x = u(p) \qquad y = f(x)$$

$$z = g(y) = g\big(f(x)\big) = g(f(u(p)))$$

$$(g \circ f \circ u)' = (g' \circ f \circ u) \cdot (f \circ u)'$$

$$= (g' \circ f \circ u) \cdot f' \cdot u'$$

$$= \frac{dz}{dy} \cdot \frac{dy}{dx} \cdot \frac{dx}{dp}$$

추가 ^^

$$p = v(q) \qquad x = u(p) \qquad y = f(x)$$

$$z = g(y) = g\big(f(x)\big) = g(f(u(p))) = g(f(u(v(q))))$$

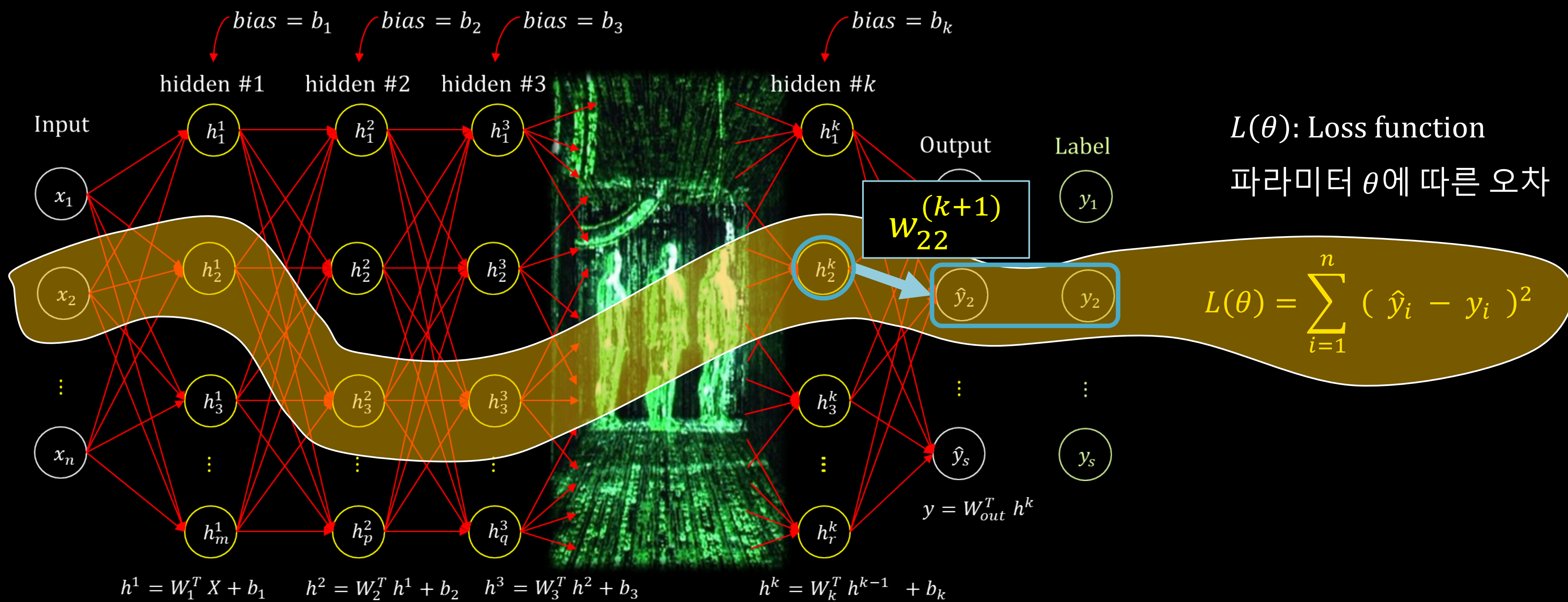
$$(g \circ f \circ u \circ v)' = (g' \circ f \circ u \circ v) \cdot (f \circ u \circ v)'$$

$$= (g' \circ f \circ u \circ v) \cdot (f' \circ u \circ v) \cdot (u \circ v)'$$

$$= (g' \circ f \circ u \circ v) \cdot (f' \circ u \circ v) \cdot u' \cdot v'$$

$$= \frac{dz}{dy} \cdot \frac{dy}{dx} \cdot \frac{dx}{dp} \cdot \frac{dp}{dq}$$

Chain Rule in Deeplearning (1/4) → 손가락으로 해보는 미분

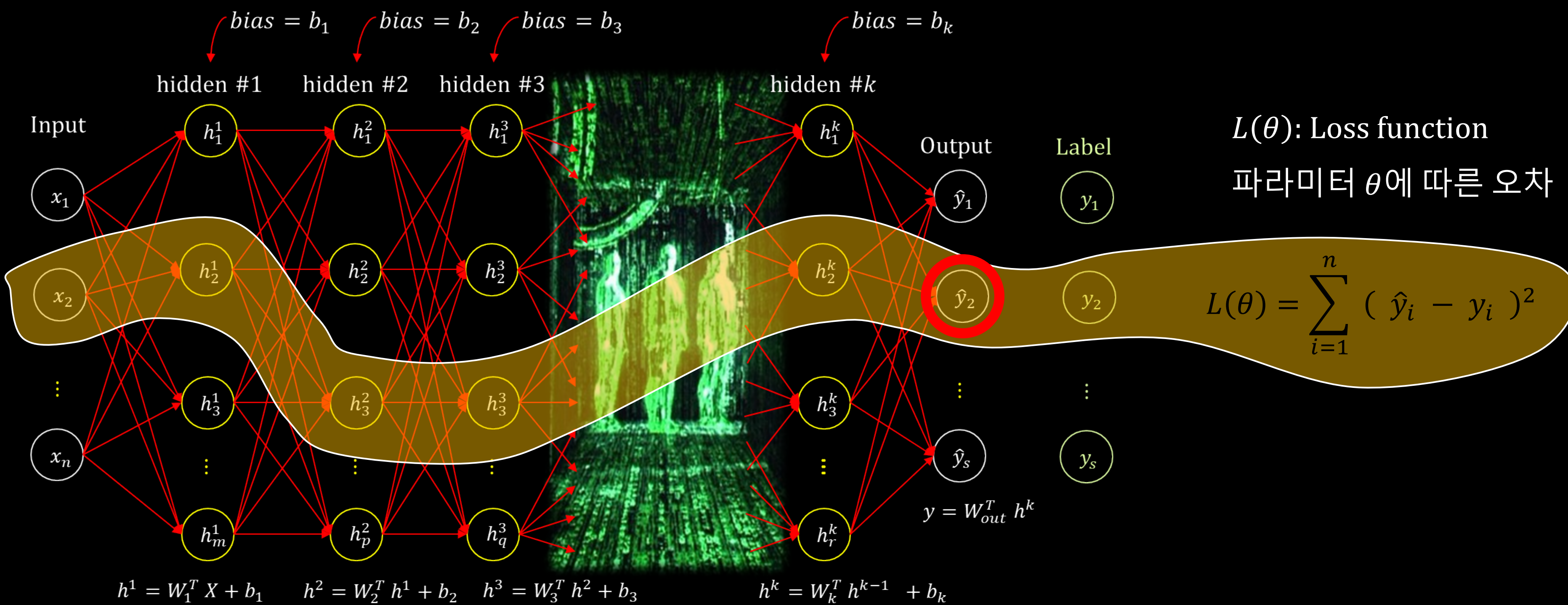


$w_{22}^{(k+1)}$ → 이 값을 업데이트 하려면?

$L(\theta)$를 $w_{22}^{(k+1)}$에 대하여 편미분 해본다.

$$\frac{\partial L(\theta)}{\partial w_{22}^{(k+1)}} = \frac{\partial L(\theta)}{\partial \hat{y}_2} \cdot \frac{\partial \hat{y}_2}{\partial w_{22}^{(k+1)}}$$

Chain Rule in Deeplearning (2/4): $\frac{\partial L(\theta)}{\partial \hat{y}_2}$ 손으로 구해보기



$\hat{y}_2 \rightarrow$ 이 값을 업데이트 하려면?

$L(\theta)$를 $\hat{y}_2$에 대하여 편미분한다.

$\hat{y}_2^2 = 0.5, y_2 = 1.0$ 이라고 가정

(최초 forward 과정에서 이미 계산된 값)

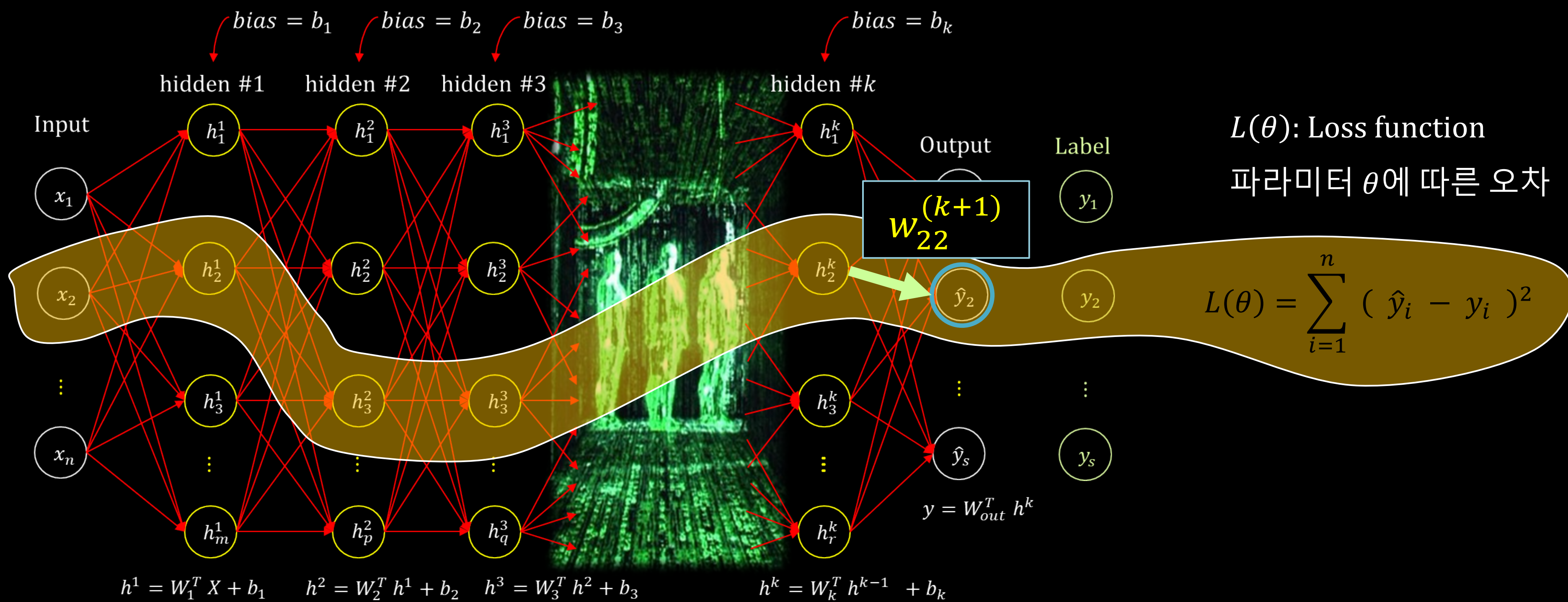
$$L(\theta) = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = (\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2 + \cdots + (\hat{y}_n - y_n)^2$$

$$\frac{\partial L(\theta)}{\partial \hat{y}_2} = \frac{\partial}{\partial \hat{y}_2} \cancel{(\hat{y}_1 - y_1)^2} + (\hat{y}_2 - y_2)^2 + \cdots + \cancel{(\hat{y}_n - y_n)^2} = \frac{\partial}{\partial \hat{y}_2} (\hat{y}_2 - y_2)^2$$

$$= \frac{\partial}{\partial \hat{y}_2} \hat{y}_2^2 - 2\hat{y}_2 y_2 + \cancel{y_2^2} = 2\hat{y}_2 - 2y_2$$

$$\frac{\partial L(\theta)}{\partial \hat{y}_2} = 2(0.5 - 1.0) = -1.0$$

Chain Rule in Deeplearning (3/4): $\frac{\partial \hat{y}_2}{\partial w_{22}^{(k+1)}}$ 손으로 구해보기



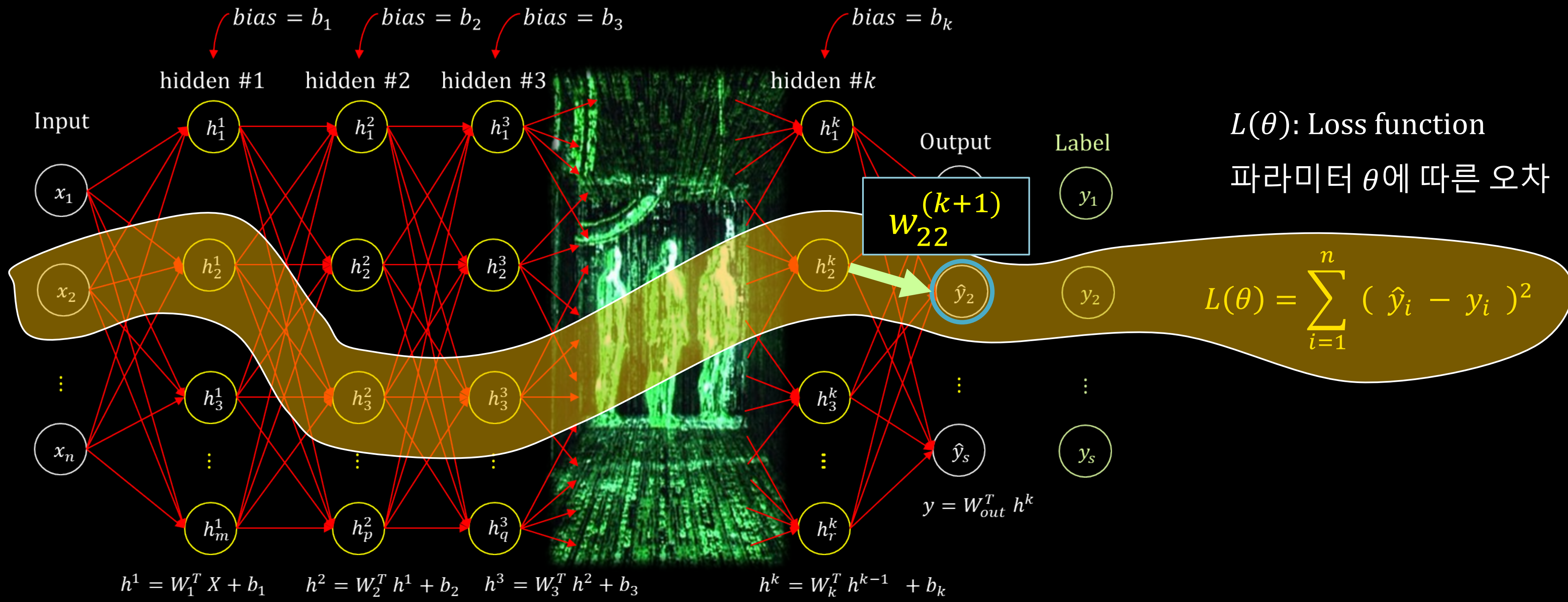
$$\frac{\partial \hat{y}_2}{\partial w_{22}^{(k+1)}} = \frac{\partial}{\partial w_{22}^{(k+1)}} \cancel{h_1^k w_{12}^{(k+1)}} + h_2^k w_{22}^{(k+1)} + \cdots + \cancel{h_r^k w_{r2}^{(k+1)}} = h_2^k$$

$h_2^k = 0.1$ 이라고 가정하면,

(최초 forward 과정에서 이미 계산된 값)

$$\frac{\partial \hat{y}_2}{\partial w_{22}^{(k+1)}} = 0.1$$

Chain Rule in Deeplearning (4/4): 최종 업데이트 수행



$w_{22}^{(k+1)} \rightarrow$ 업데이트 수행

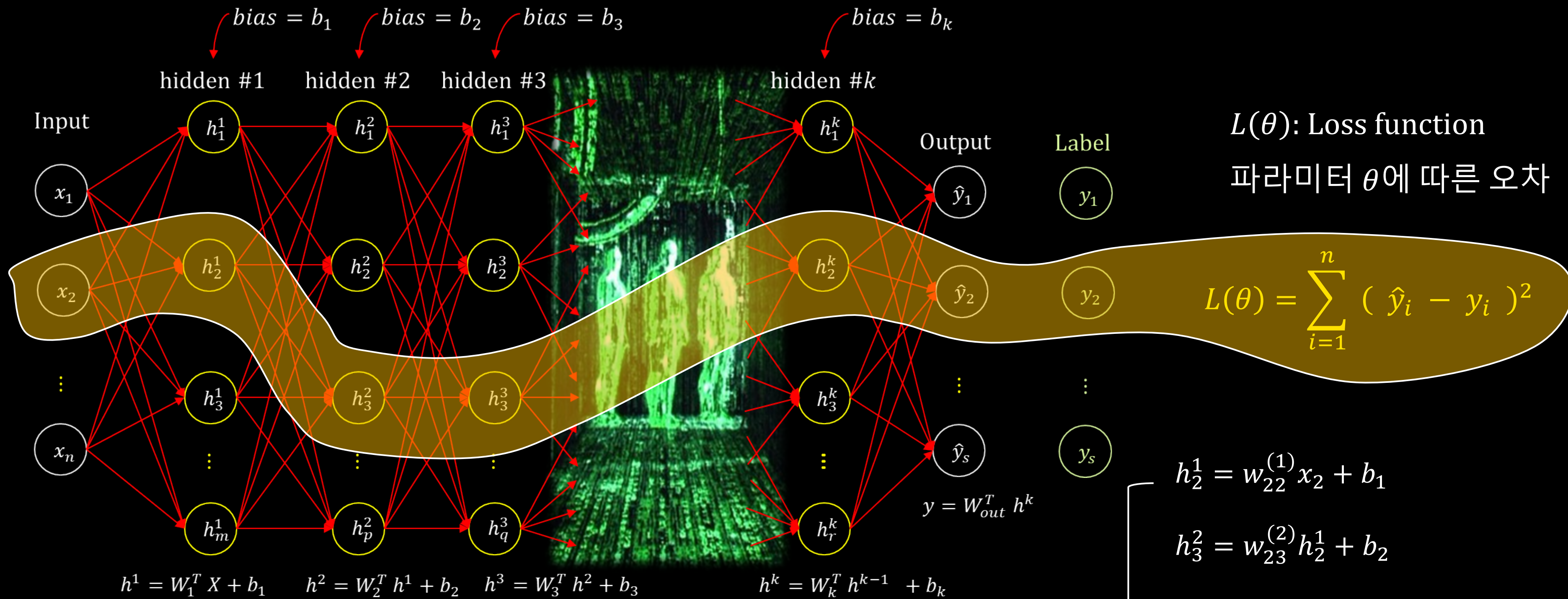
$$w_{22}^{(k+1)} = w_{22}^{(k+1)} - \alpha \frac{\partial L(\theta)}{w_{22}^{(k+1)}}, assume\ \alpha = 1.0$$

$$\frac{\partial L(\theta)}{\partial w_{22}^{(k+1)}} = \frac{\partial L(\theta)}{\partial \hat{y}_2} \cdot \frac{\partial \hat{y}_2}{\partial w_{22}^{(k+1)}}$$

$$= -1.0 \times 0.1 = -0.01$$

$$w_{22}^{(k+1)} = w_{22}^{(k+1)} - (1) \times (-0.01)$$

Chain Rule이 필요한 상황 - Simple Path



오차를 최소화 되도록
파라미터 θ를 바꿔줘야 한다!

How?

편미분 하면서
업데이트 하면 된다!

입력 x_i 부터 손실값 $L(\theta)$까지
하나의 경로만 생각해 볼까요 ^^?

$$h^1_2 = w^{(1)}_{22} x_2 + b_1$$

$$h^2_3 = w^{(2)}_{23} h^1_2 + b_2$$

$$h^3_3 = w^{(3)}_{33} h^2_3 + b_3$$

$$\vdots \qquad \vdots$$

$$h^k_2 = w^{(k)}_{i2} h^{k-1}_i + b_k$$

$$\hat{y}_2 = w^{(k+1)}_{22} h^k_2$$

$$L(\boxed{\theta_2}) = (\hat{y}_2 - y_2)^2$$

$$\theta_2 = \left\{ \begin{matrix} w^{(1)}_{22}, w^{(2)}_{23}, w^{(3)}_{33}, \cdots, w^{(k)}_{i2}, w^{(k+1)}_{22} \\ b_1, b_2, b_3, \cdots, b_k \end{matrix} \right\}$$

여기서 θ_2는?

교수님~~ 좀 더 멋있게… 하고 싶어요^^

딥러닝에 적용되는 Chain Rule을 정확히 이해하기 위해서는…

2가지를 더 공부해야 합니다 ^^.

1. 선형 시스템의 미분

다음 강의에서 설명해 드릴께요^^
(7강. 행렬 미분)

2. 일괄 미분 결과를 적용한
Back-propagation

다음다음 강의에서 설명해 드릴께요 ^^
(8강. Back-propagation)

수고하셨습니다 ..^^..