CUAI 스터디 CS224n1(NLP)팀

Lecture 3 : Neural net learning: Gradients by hand (matrix calculus)
and algorithmically (the backpropagation algorithm)

2022.03.17

발표자 : 김서린

INDEX

- Introduction
- Matrix calculus
- Backpropagation

1. Introduction

1. Introduction

Named Entity Recognition (NER)

Last	night,	Paris	Hilton	wowed	in	a	sequin	gown.
		PER	PER					

Samuel	Quinn	was	arrested	in	the	Hilton	Hotel	in	Paris	in	April	1989.
PER	PER					LOC	LOC		LOC		DATE	DATE

Text 보고 단어 labeling 하는 것이 목표 (사람, 장소, 물건, 날짜, 시간 등으로)
문맥을 확인해야 함

1. Introduction

Window classification using binary logistic classifier

the museums in Paris are amazing to see .

$$X_{window} = [\ x_{museums} \quad x_{in} \quad x_{Paris} \quad x_{are} \quad x_{amazing}\ ]^T$$

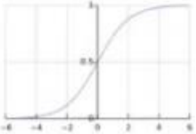
Window Classification

앞뒤 단어들을 포함한 word vector를 neural network layer의 input으로 -> 문맥 고려!

1. Introduction

$$J_t(\theta) = \sigma(s) = \frac{1}{1+e^{-s}}$$

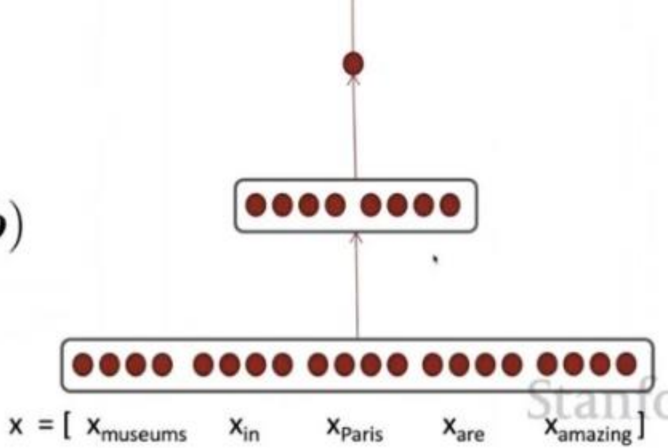
predicted model probability of class



$$s = \boldsymbol{u}^T \boldsymbol{h}$$

$$\boldsymbol{h} = f(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})$$

$$\boldsymbol{x} \quad \text{(input)}$$



1. Introduction

Stochastic Gradient Descent

$$\theta^{new} = \theta^{old} - \alpha \nabla_{\theta} J(\theta)$$

$\boxed{\alpha = \textit{step size or learning rate}}$

비용함수 미분 방법?

1. By hand (matrix calculus)
2. Algorithmically (backpropagation)

2. Matrix Calculus

2. Matrix calculus

Jacobian Matrix: Generalization of the Gradient

$$f(\boldsymbol{x}) = f(x_1, x_2, \ldots, x_n) \longrightarrow \boldsymbol{f}(\boldsymbol{x}) = [f_1(x_1, x_2, \ldots, x_n), \ldots, f_m(x_1, x_2, \ldots, x_n)]$$

$$\frac{\partial f}{\partial \boldsymbol{x}} = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n}\right] \longrightarrow \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \qquad \boxed{\left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\right)_{ij} = \frac{\partial f_i}{\partial x_j}}$$

Jacobian Matrix는 모든 함수, 변수에 대한 편미분 결과 조합의 mxn 행렬

2. Matrix calculus

Chain Rule

$$z = 3y$$
$$y = x^2$$
$$\frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx} = (3)(2x) = 6x$$

$$\boldsymbol{h} = f(\boldsymbol{z})$$
$$\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}$$
$$\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{z}}\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} = \ldots$$

x는 input vector, f는 sigmoid function

2. Matrix calculus

Example Jacobian: Elementwise activation Function

$$\boldsymbol{h} = f(\boldsymbol{z}), \text{what is } \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{z}}? \qquad \boldsymbol{h}, \boldsymbol{z} \in \mathbb{R}^n$$

$$h_i = f(z_i)$$

함수는 n개의 output과 n개의 input을 가지고 있으므로 -> nxn Jacobian

$$\left(\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{z}}\right)_{ij} = \frac{\partial h_i}{\partial z_j} = \frac{\partial}{\partial z_j} f(z_i) \qquad \text{definition of Jacobian}$$

$$= \begin{cases} f'(z_i) & \text{if } i = j \\ 0 & \text{if otherwise} \end{cases} \qquad \text{regular 1-variable derivative}$$

$$\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{z}} = \begin{pmatrix} f'(z_1) & & 0 \\ & \ddots & \\ 0 & & f'(z_n) \end{pmatrix} = \text{diag}(\boldsymbol{f}'(\boldsymbol{z}))$$

첨자 i, j 가 같아야 미분값이 존재, 다르면 0으로 없어짐

2. Matrix calculus

Example Jacobian: Elementwise activation Function

$$\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{z}} = \begin{pmatrix} f'(z_1) & & 0 \\ & \ddots & \\ 0 & & f'(z_n) \end{pmatrix} = \text{diag}(\boldsymbol{f}'(\boldsymbol{z}))$$

$$\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{W}\boldsymbol{x}+\boldsymbol{b}) = \boldsymbol{W}$$

$$\frac{\partial}{\partial \boldsymbol{b}}(\boldsymbol{W}\boldsymbol{x}+\boldsymbol{b}) = \boldsymbol{I} \quad \text{(Identity matrix)}$$

$$\frac{\partial}{\partial \boldsymbol{u}}(\boldsymbol{u}^T\boldsymbol{h}) = \boldsymbol{h}^T$$

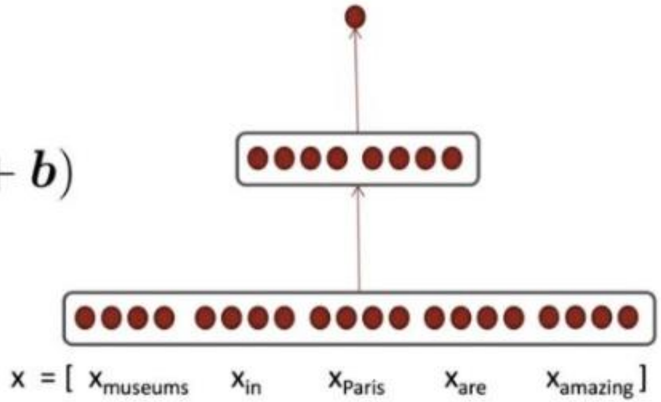
2. Matrix calculus

Back to our Neural Net!

$$s = \boldsymbol{u}^T \boldsymbol{h}$$

$$\boldsymbol{h} = f(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})$$

$$\boldsymbol{x} \quad \text{(input)}$$



Let's find $\dfrac{\partial s}{\partial \boldsymbol{b}}$

2. Matrix calculus

1) Break up equations into simple pieces

$$s = \boldsymbol{u}^T \boldsymbol{h} \qquad\qquad s = \boldsymbol{u}^T \boldsymbol{h}$$

$$\boldsymbol{h} = f(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}) \qquad\qquad \boldsymbol{h} = f(\boldsymbol{z})$$

$$\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}$$

$$\boldsymbol{x} \quad \text{(input)} \qquad\qquad \boldsymbol{x} \quad \text{(input)}$$

h = f(Wx+b) 가 아직 복잡 -> z = Wx+b 로 치환

2. Matrix calculus

2) Apply the chain rule

$$\boxed{s = \boldsymbol{u}^T \boldsymbol{h}}$$
$$\boxed{\boldsymbol{h} = f(\boldsymbol{z})}$$
$$\boxed{\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}}$$
$$\boldsymbol{x} \quad \text{(input)}$$

$$\frac{\partial s}{\partial \boldsymbol{b}} = \boxed{\frac{\partial s}{\partial \boldsymbol{h}}}\boxed{\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{z}}}\boxed{\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{b}}}$$

2. Matrix calculus

2) Apply the chain rule

$$\boxed{s = \boldsymbol{u}^T \boldsymbol{h}}$$
$$\boxed{\boldsymbol{h} = f(\boldsymbol{z})}$$
$$\boxed{\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}}$$
$$\boldsymbol{x} \quad \text{(input)}$$

$$\frac{\partial s}{\partial \boldsymbol{b}} = \boxed{\frac{\partial s}{\partial \boldsymbol{h}}}\boxed{\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{z}}}\boxed{\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{b}}}$$
$$= \boldsymbol{u}^T \operatorname{diag}(f'(\boldsymbol{z}))\boldsymbol{I}$$
$$= \boldsymbol{u}^T \circ f'(\boldsymbol{z})$$

Useful Jacobians from previous slide

$$\frac{\partial}{\partial \boldsymbol{u}}(\boldsymbol{u}^T\boldsymbol{h}) = \boldsymbol{h}^T$$
$$\frac{\partial}{\partial \boldsymbol{z}}(f(\boldsymbol{z})) = \operatorname{diag}(f'(\boldsymbol{z}))$$
$$\frac{\partial}{\partial \boldsymbol{b}}(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}) = \boldsymbol{I}$$

2. Matrix calculus

Re-using Computation

$$\frac{\partial s}{\partial \boldsymbol{W}} = \frac{\partial s}{\partial \boldsymbol{h}} \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{z}} \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{W}}$$
$$\frac{\partial s}{\partial \boldsymbol{b}} = \frac{\partial s}{\partial \boldsymbol{h}} \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{z}} \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{b}}$$

$$\frac{\partial s}{\partial \boldsymbol{W}} = \boldsymbol{\delta} \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{W}}$$
$$\frac{\partial s}{\partial \boldsymbol{b}} = \boldsymbol{\delta} \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{b}} = \boldsymbol{\delta}$$
$$\boldsymbol{\delta} = \frac{\partial s}{\partial \boldsymbol{h}} \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{z}} = \boldsymbol{u}^T \circ f'(\boldsymbol{z})$$

δ is the local error signal

2. Matrix calculus

Deriving local input gradient in backprop

$$\frac{\partial s}{\partial \boldsymbol{W}} = \boldsymbol{\delta}\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{W}} = \boldsymbol{\delta}\frac{\partial}{\partial \boldsymbol{W}}(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})$$

$$\frac{\partial z_i}{\partial W_{ij}} = \frac{\partial}{\partial W_{ij}}\boldsymbol{W}_{i\cdot}\boldsymbol{x} + b_i$$
$$= \frac{\partial}{\partial W_{ij}}\sum_{k=1}^{d} W_{ik}x_k = x_j$$

$$\begin{array}{ccc} \dfrac{\partial s}{\partial \boldsymbol{W}} & = & \boldsymbol{\delta}^T \quad \boldsymbol{x}^T \\ [n \times m] & & [n \times 1][1 \times m] \end{array}$$

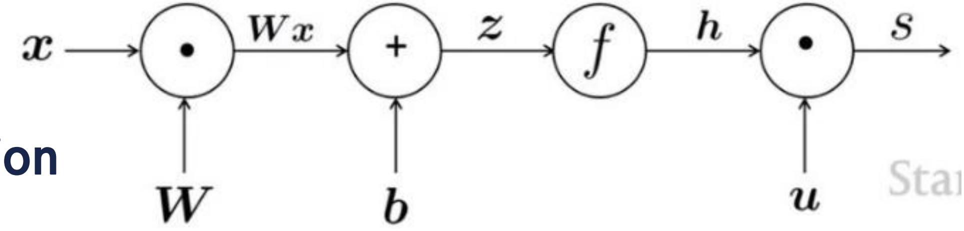
$$= \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_n \end{bmatrix} [x_1, \ldots, x_m] = \begin{bmatrix} \delta_1 x_1 & \ldots & \delta_1 x_m \\ \vdots & \ddots & \vdots \\ \delta_n x_1 & \ldots & \delta_n x_m \end{bmatrix}$$

3. Backpropagation

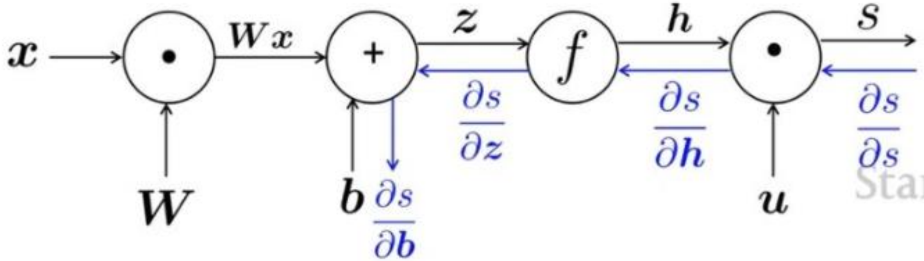
3. Backpropagation

Forward & Back Propagation

Forward propagation

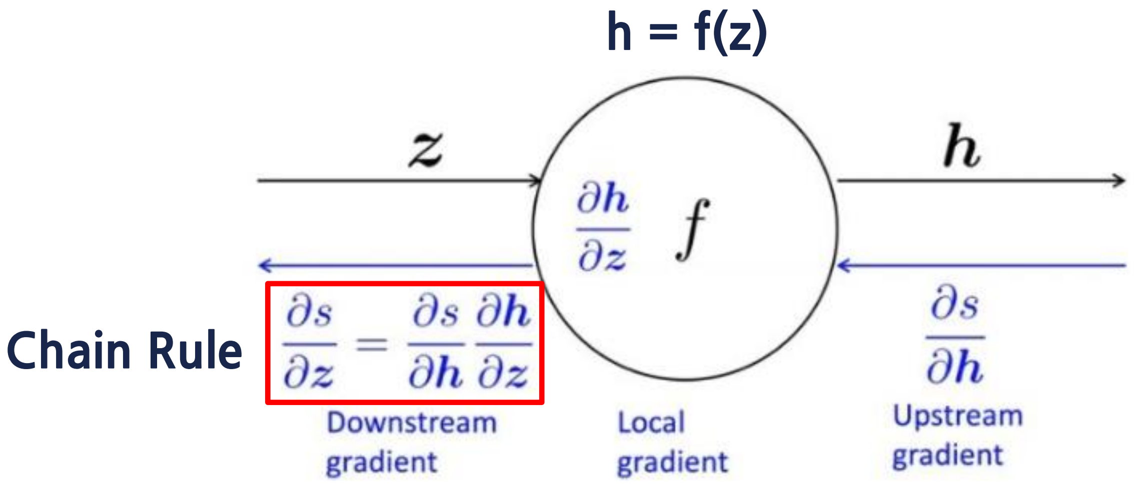


Back propagation



3. Backpropagation

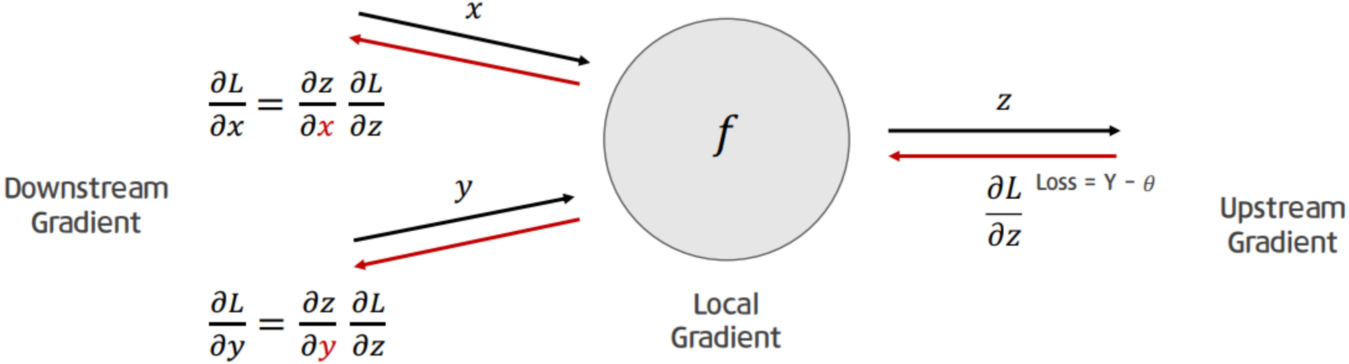
Backpropagation: Single Node



Downstream gradient = upstream gradient x local gradient

3. Backpropagation

역전파 분해



$$\frac{\partial L}{\partial x} = \frac{\partial z}{\partial x}\frac{\partial L}{\partial z}$$

$$\frac{\partial L}{\partial y} = \frac{\partial z}{\partial y}\frac{\partial L}{\partial z}$$

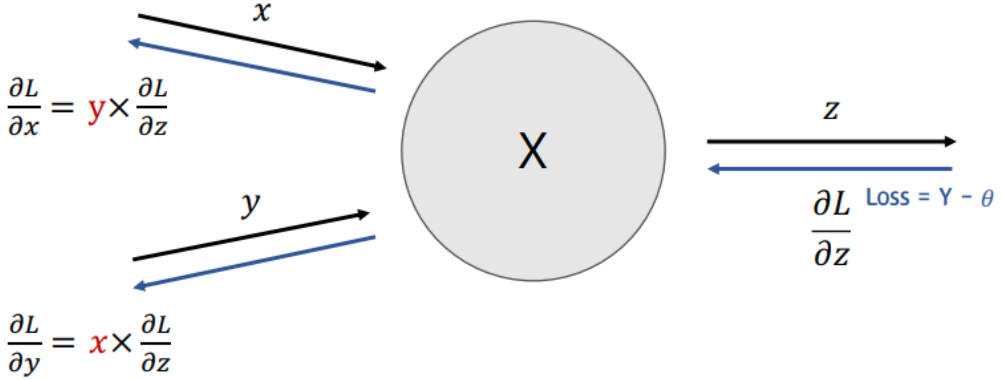
$$\frac{\partial L}{\partial z}$$

3. Backpropagation

역전파 분해

곱셈의 역전파

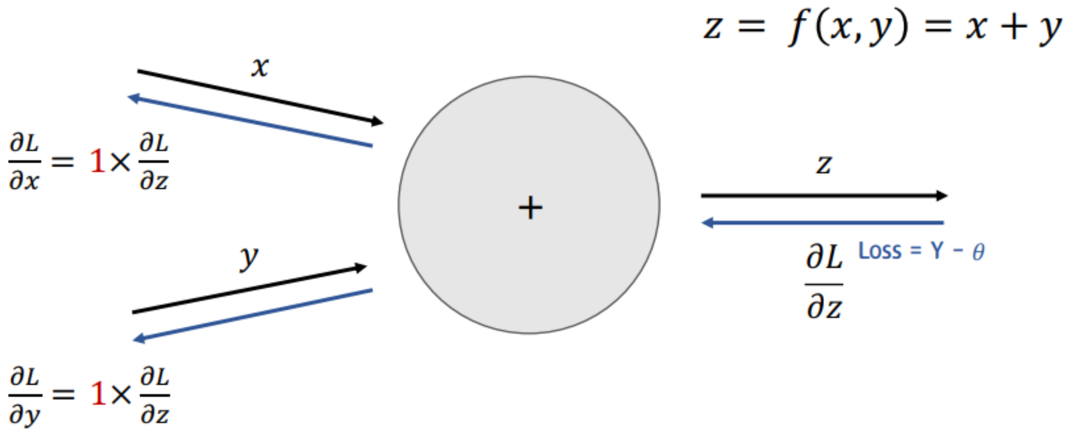
$$z = f(x, y) = xy$$



3. Backpropagation

역전파 분해

덧셈의 역전파

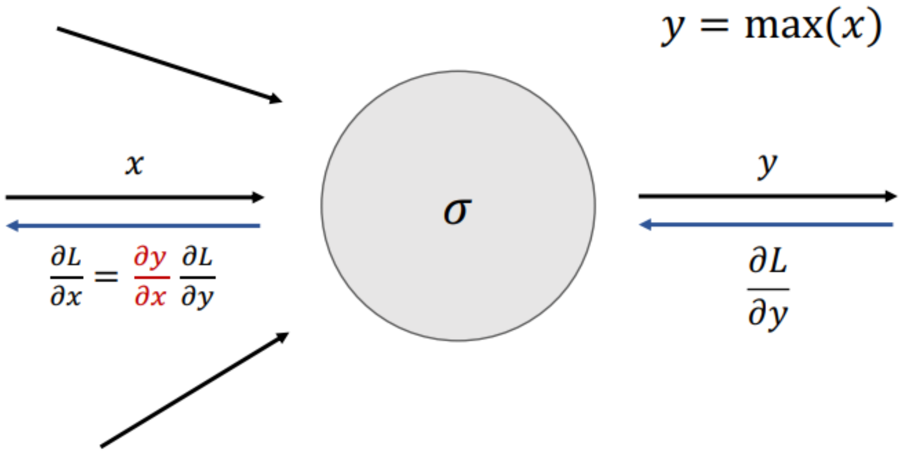


3. Backpropagation

역전파 분해

max 역전파

$$\frac{\partial y}{\partial x} = \begin{cases} 1 \ if \ x \ is \ max \\ 0 \ \ otherwise \end{cases}$$



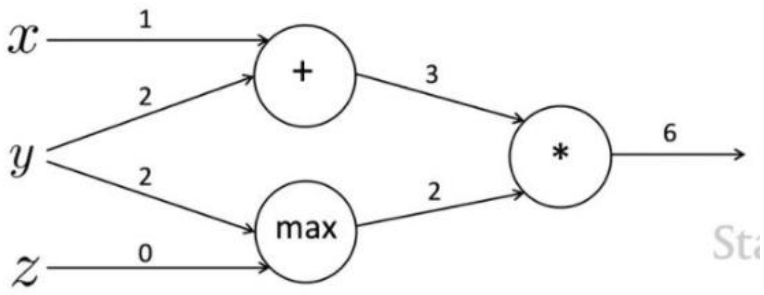
3. Backpropagation

example

$$f(x, y, z) = (x + y)\max(y, z)$$
$$x = 1, y = 2, z = 0$$

Forward prop steps

$$a = x + y$$
$$b = \max(y, z)$$
$$f = ab$$



3. Backpropagation

example

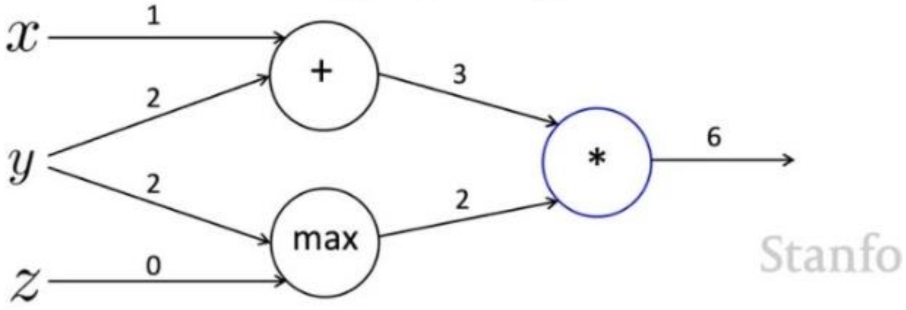
$$f(x, y, z) = (x + y)\max(y, z)$$
$$x = 1, y = 2, z = 0$$

Forward prop steps

$$a = x + y$$
$$b = \max(y, z)$$
$$f = ab$$

Local gradients

$$\frac{\partial a}{\partial x} = 1 \quad \frac{\partial a}{\partial y} = 1$$
$$\frac{\partial b}{\partial y} = \mathbf{1}(y > z) = 1 \quad \frac{\partial b}{\partial z} = \mathbf{1}(z > y) = 0$$
$$\frac{\partial f}{\partial a} = b = 2 \quad \frac{\partial f}{\partial b} = a = 3$$



3. Backpropagation

example

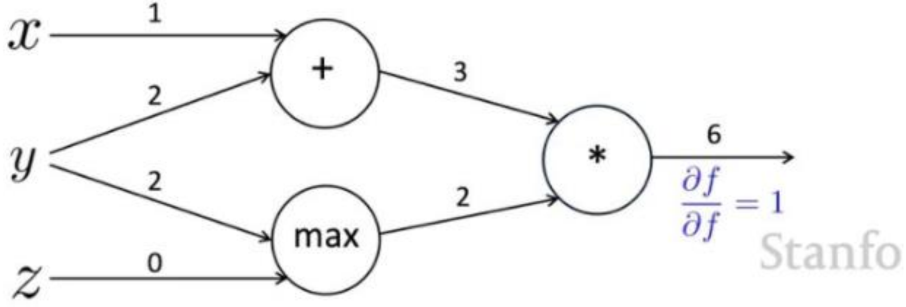
$$f(x, y, z) = (x + y)\max(y, z)$$
$$x = 1, y = 2, z = 0$$

Forward prop steps

$$a = x + y$$
$$b = \max(y, z)$$
$$f = ab$$

Local gradients

$$\frac{\partial a}{\partial x} = 1 \quad \frac{\partial a}{\partial y} = 1$$
$$\frac{\partial b}{\partial y} = \mathbf{1}(y > z) = 1 \quad \frac{\partial b}{\partial z} = \mathbf{1}(z > y) = 0$$
$$\frac{\partial f}{\partial a} = b = 2 \quad \frac{\partial f}{\partial b} = a = 3$$



3. Backpropagation

example

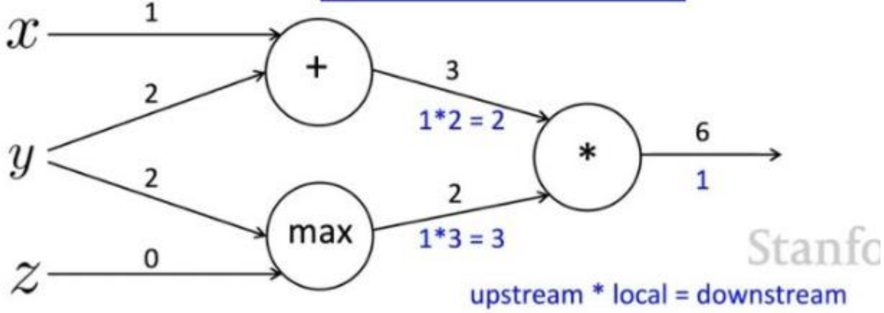
$$f(x, y, z) = (x + y)\max(y, z)$$
$$x = 1, y = 2, z = 0$$

Forward prop steps

$$a = x + y$$
$$b = \max(y, z)$$
$$f = ab$$

Local gradients

$$\frac{\partial a}{\partial x} = 1 \quad \frac{\partial a}{\partial y} = 1$$
$$\frac{\partial b}{\partial y} = \mathbf{1}(y > z) = 1 \quad \frac{\partial b}{\partial z} = \mathbf{1}(z > y) = 0$$
$$\boxed{\frac{\partial f}{\partial a} = b = 2 \quad \frac{\partial f}{\partial b} = a = 3}$$



3. Backpropagation

example

$$f(x, y, z) = (x + y)\max(y, z)$$
$$x = 1, y = 2, z = 0$$

Forward prop steps

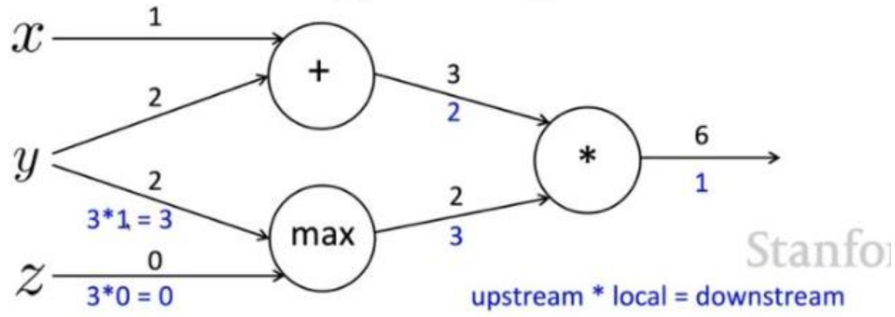
$$a = x + y$$
$$b = \max(y, z)$$
$$f = ab$$

Local gradients

$$\frac{\partial a}{\partial x} = 1 \quad \frac{\partial a}{\partial y} = 1$$

$$\boxed{\frac{\partial b}{\partial y} = \mathbf{1}(y > z) = 1 \quad \frac{\partial b}{\partial z} = \mathbf{1}(z > y) = 0}$$

$$\frac{\partial f}{\partial a} = b = 2 \quad \frac{\partial f}{\partial b} = a = 3$$



3. Backpropagation

example

$$\boxed{\begin{aligned} &f(x, y, z) = (x + y)\max(y, z) \\ &x = 1, y = 2, z = 0 \end{aligned}}$$

Forward prop steps

$$a = x + y$$

$$b = \max(y, z)$$

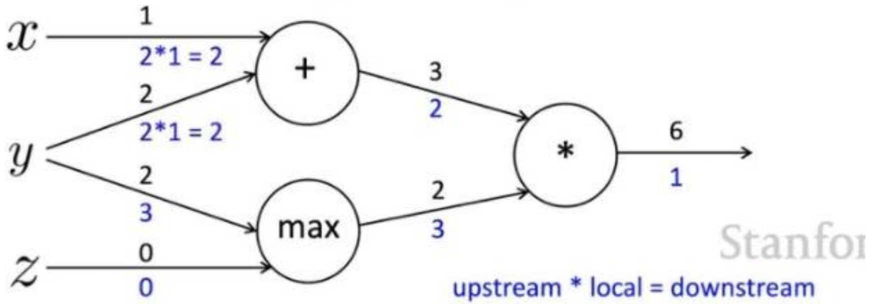
$$f = ab$$

Local gradients

$$\boxed{\frac{\partial a}{\partial x} = 1 \quad \frac{\partial a}{\partial y} = 1}$$

$$\frac{\partial b}{\partial y} = \mathbf{1}(y > z) = 1 \quad \frac{\partial b}{\partial z} = \mathbf{1}(z > y) = 0$$

$$\frac{\partial f}{\partial a} = b = 2 \quad \frac{\partial f}{\partial b} = a = 3$$



3. Backpropagation

example

$$f(x, y, z) = (x + y)\max(y, z)$$
$$x = 1, y = 2, z = 0$$

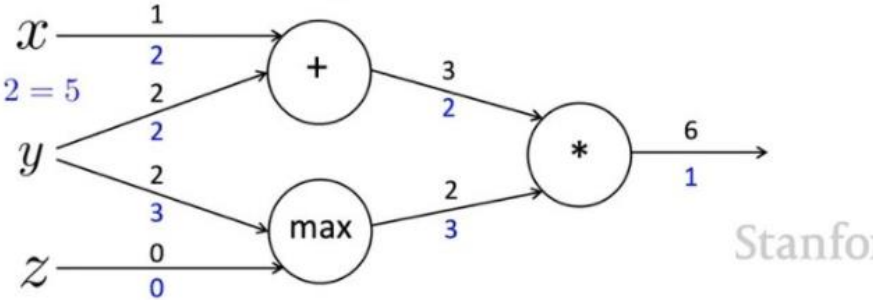
Forward prop steps

$$a = x + y$$
$$b = \max(y, z)$$
$$f = ab$$

Local gradients

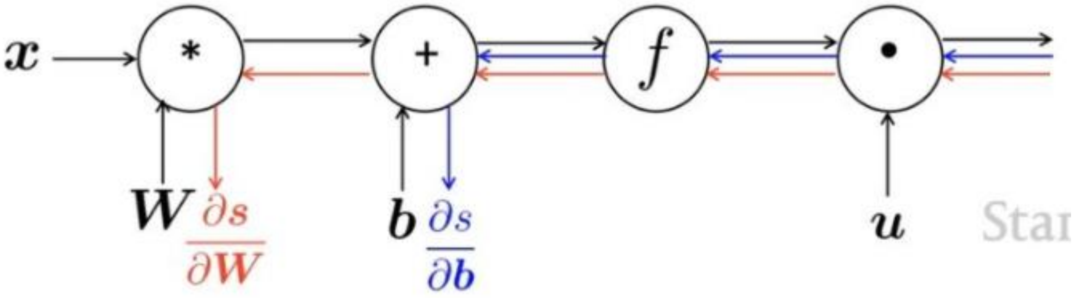
$$\frac{\partial a}{\partial x} = 1 \quad \frac{\partial a}{\partial y} = 1$$
$$\frac{\partial b}{\partial y} = \mathbf{1}(y > z) = 1 \quad \frac{\partial b}{\partial z} = \mathbf{1}(z > y) = 0$$
$$\frac{\partial f}{\partial a} = b = 2 \quad \frac{\partial f}{\partial b} = a = 3$$

$$\frac{\partial f}{\partial x} = 2$$
$$\frac{\partial f}{\partial y} = 3 + 2 = 5$$
$$\frac{\partial f}{\partial z} = 0$$



3. Backpropagation

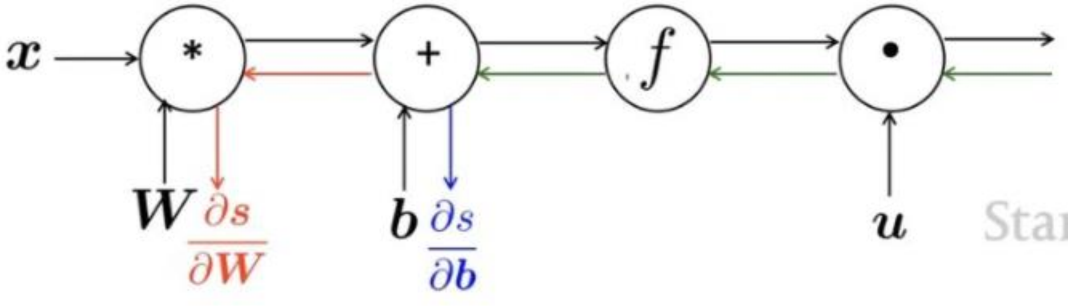
Efficiency: compute all gradients at once



ds/db 계산 후 ds/dW 계산하면 비효율
(앞서 나왔던 chain rule의 공통부분 중복계산)

3. Backpropagation

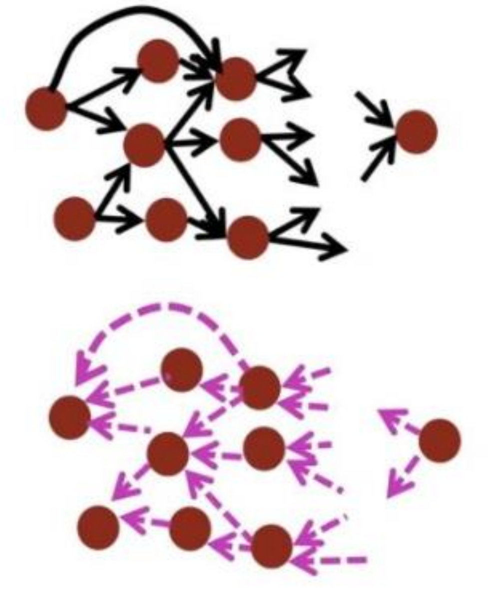
Efficiency: compute all gradients at once



계산 한꺼번에 하면 효율적임

3. Backpropagation

Back-Prop in General Computation Graph



Forward propagation
방향 맞추어, topological sort로 정렬한 뒤 노드 방문

Back propagation
Output gradient = 1로 설정하고 시작
역방향으로 방문하며 local gradient 계산

Forward/Backward 시간복잡도 동일

Tensorflow, PyTorch에 잘 구현되어 있음

감사합니다☺