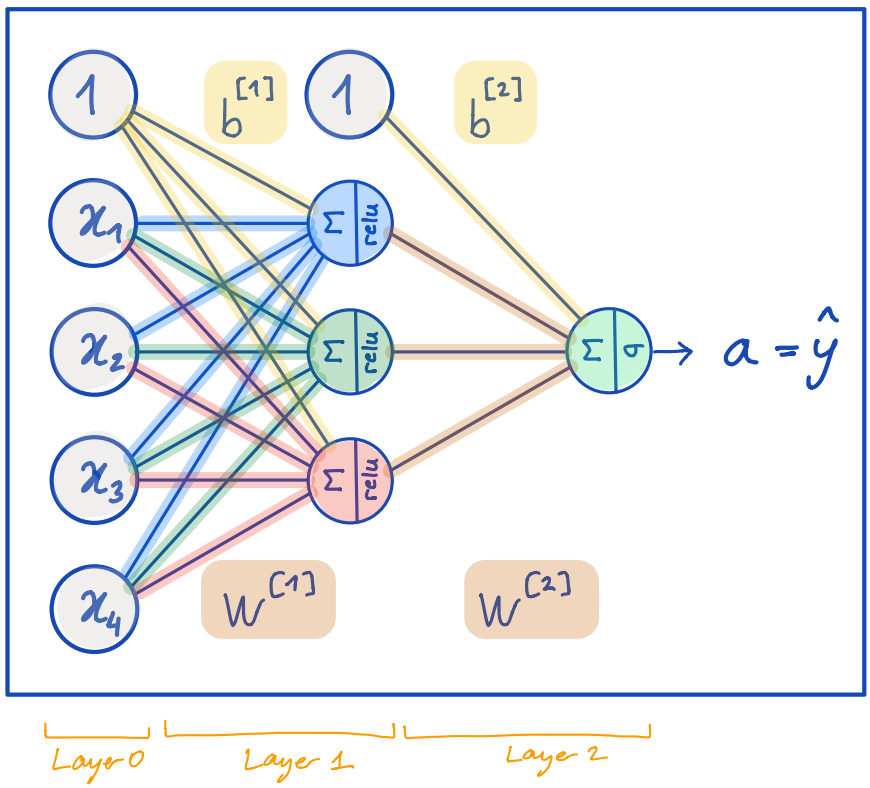
BINÄRE KLASSIFIKATION MIT DNN



$$\begin{bmatrix} \\ \\ \end{bmatrix} \cdot \begin{bmatrix} \\ \\ \\ \end{bmatrix} + \begin{bmatrix} \\ \\ \end{bmatrix} = \begin{bmatrix} \\ \\ \end{bmatrix} = z^{[1]}$$

$$\text{relu}\left(W^{[1]} \cdot x + b^{[1]}\right) = a^{[1]}$$

$$\begin{bmatrix} \\ \end{bmatrix} \cdot \begin{bmatrix} \\ \\ \\ \end{bmatrix} + \begin{bmatrix} \\ \end{bmatrix} = \begin{bmatrix} \\ \end{bmatrix} = z^{[2]}$$

$$\sigma\left(W^{[2]} \cdot a^{[1]} + b^{[2]}\right) = a^{[2]}$$

Broadcasting

$$\text{relu}\left(\begin{bmatrix} \\ \\ \end{bmatrix} \cdot \begin{bmatrix} (1) & (2) & (3) & \cdots & (6) & (7) \\ & & & & & \\ & & & & & \\ & & & & & \end{bmatrix} + \begin{bmatrix} \\ \\ \end{bmatrix}\right) = \begin{bmatrix} (1) & (2) & (3) & \cdots & (6) & (7) \\ & & & & & \\ & & & & & \\ & & & & & \end{bmatrix}$$

(3×4) (4×7) 3×7 (3×7)

$$\text{relu}\left(W^{[1]} \cdot X + b^{[1]}\right) = A^{[1]}$$

Broadcasting

$$\sigma\left(\begin{bmatrix} \end{bmatrix} \cdot \begin{bmatrix} & & & & & & \\ & & & & & & \\ & & & & & & \end{bmatrix} + \begin{bmatrix} \end{bmatrix}\right) = \begin{bmatrix} (1) & (2) & (3) & \cdots & (6) & (7) \\ & & & & & & \end{bmatrix}$$

(1×1) (3×7) (1×1) (1×7)

$$\sigma\left(W^{[2]} \cdot A^{[1]} + b^{[2]}\right) = A^{[2]} =: \hat{Y}$$

CROSS-ENTROPY LOSS

$A^{[2]} = \hat{Y} =$ [●●●●●●●] IST

$Y =$ [●●●●●●●] SOLL

LOSS = DISTANCE($\hat{Y}$, Y) = (1) (2) (3) ... (6) (7) [●●●●●●●]

- $\mathcal{L}(\bullet, \bullet) = \mathcal{L}(\hat{y}, y) = -y \cdot \log(\hat{y}) - (1-y) \cdot \log(1-\hat{y}) =$ ●

- $\mathcal{L}($●●●●●●●, ●●●●●●●$)$

  $= \mathcal{L}(\hat{Y}, Y) = -Y * \log(\hat{Y}) - (1-Y) * \log(1-\hat{Y})$

  $=$ [●●●●●●●]

- elementweises Multiplizieren

  $Y * \log(\hat{Y}) =$ [●●●●●●●] $* \log($[●●●●●●●]$)$

  $(1 \times 7) \quad = \quad (1 \times 7) \quad * \quad (1 \times 7)$

KOSTENFUNKTION

$$J(\theta) = \frac{1}{m} \cdot np.sum\left(\mathcal{L}(\hat{Y}, Y)\right)$$

$$= \frac{1}{7} \cdot np.sum\left([\bullet\bullet\bullet\bullet\bullet\bullet\bullet]\right)$$

Summe über alle Datenpunkte

BATCH GRADIEN DESCENT

DER GRADIENT

$$\theta = [\quad]^T \quad \text{(Parameter)}$$

$$\frac{\partial J}{\partial \theta} = d\theta = [\quad]^T \quad \text{(für aktuelles } \theta \text{ auswerten)}$$

$$= \nabla J = \left[\frac{\partial J}{\partial w_{11}^{[1]}} \cdots \frac{\partial J}{\partial w_{34}^{[1]}} \quad \frac{\partial J}{\partial b_1^{[1]}} \cdots \frac{\partial J}{\partial b_3^{[1]}} \quad \frac{\partial J}{\partial w_{11}^{[2]}} \cdots \frac{\partial J}{\partial w_{13}^{[2]}} \quad \frac{\partial J}{\partial b_{11}^{[2]}} \right]^T \quad \text{(for current } w\text{)}$$

Richtung des steilsten Aufstiegs ! (Gehe in entgegengesetzte Richtung)!

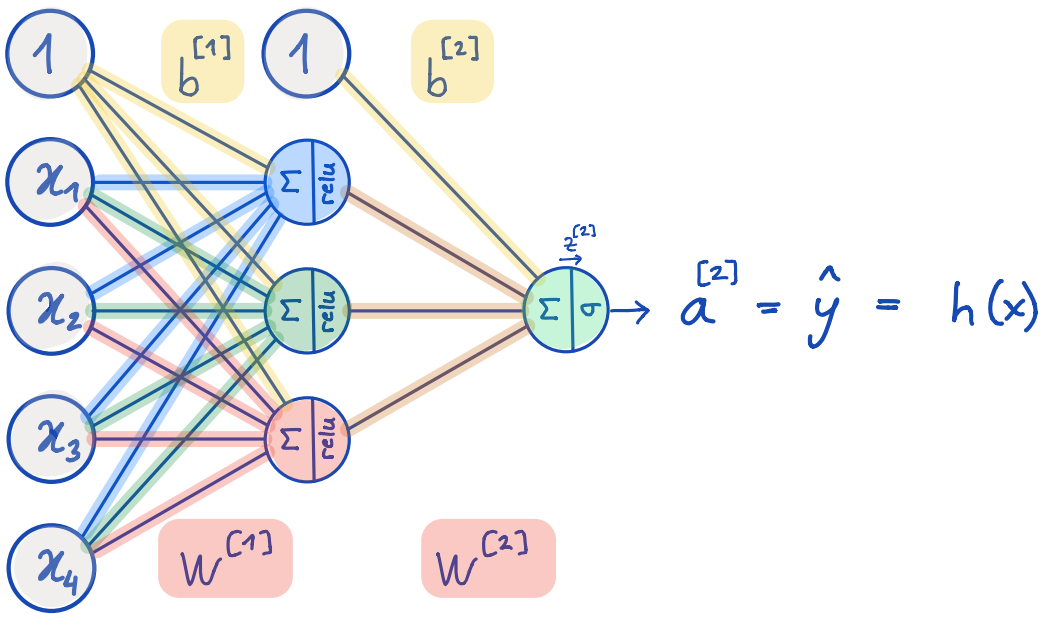
$$\theta := \theta - \alpha \cdot \nabla J$$

Gewichtsaktualisierungen

canan.yildiz@tau.edu.tr

• Backpropagation – Multiple Layers

2-Layer Network (Graph Representation)



$$h(x) = \sigma\left[W^{[2]} \cdot \text{relu}\left(W^{[1]} \cdot x + b^{[1]}\right) + b^{[2]}\right] = a^{[2]}$$

(Ausgabe als Funktion von x)

$$J = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(a^{(i)}, y^{(i)})$$

$$\mathcal{L} = -\left(y \cdot \log(a) + (1-y) \cdot \log(1-a)\right) \quad (a = a^{[2]})$$

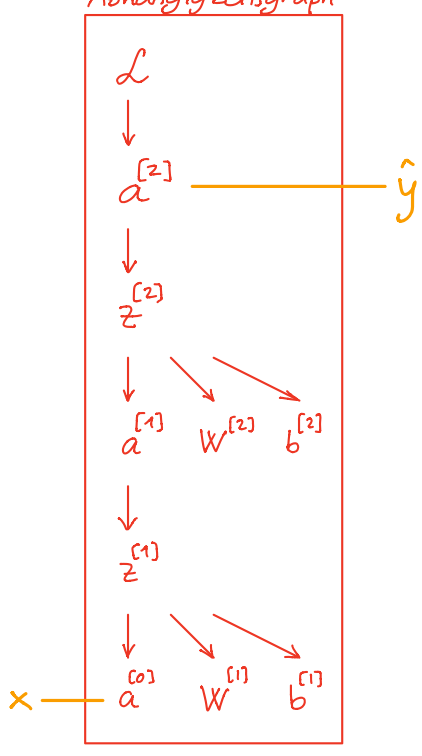
$$a^{[2]} = \sigma(z^{[2]})$$

$$z^{[2]} = W^{[2]} \cdot a^{[1]} + b^{[2]}$$

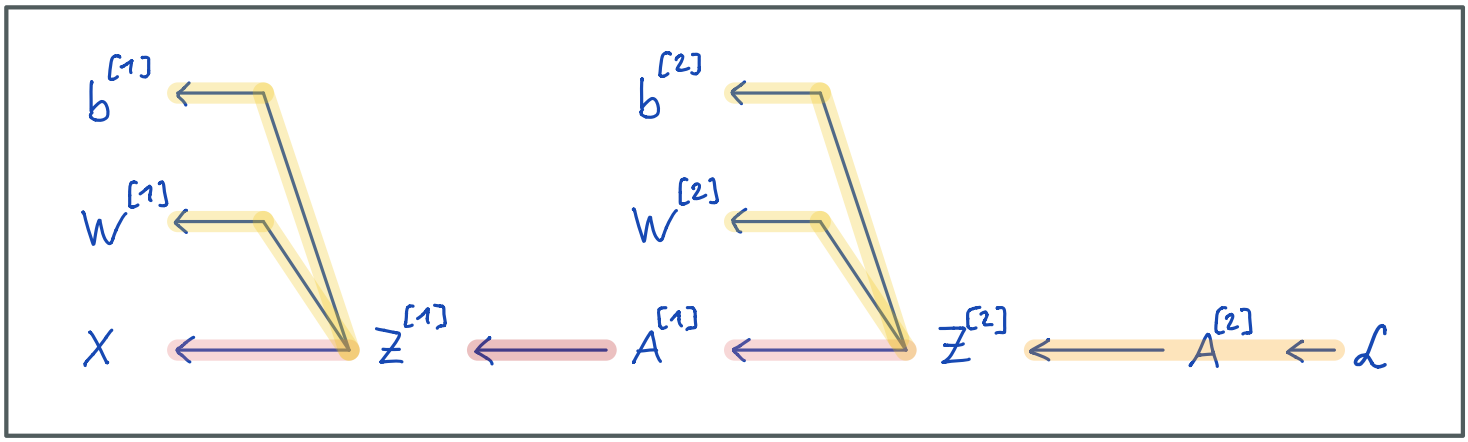
$$a^{[1]} = \text{ReLU}(z^{[1]})$$

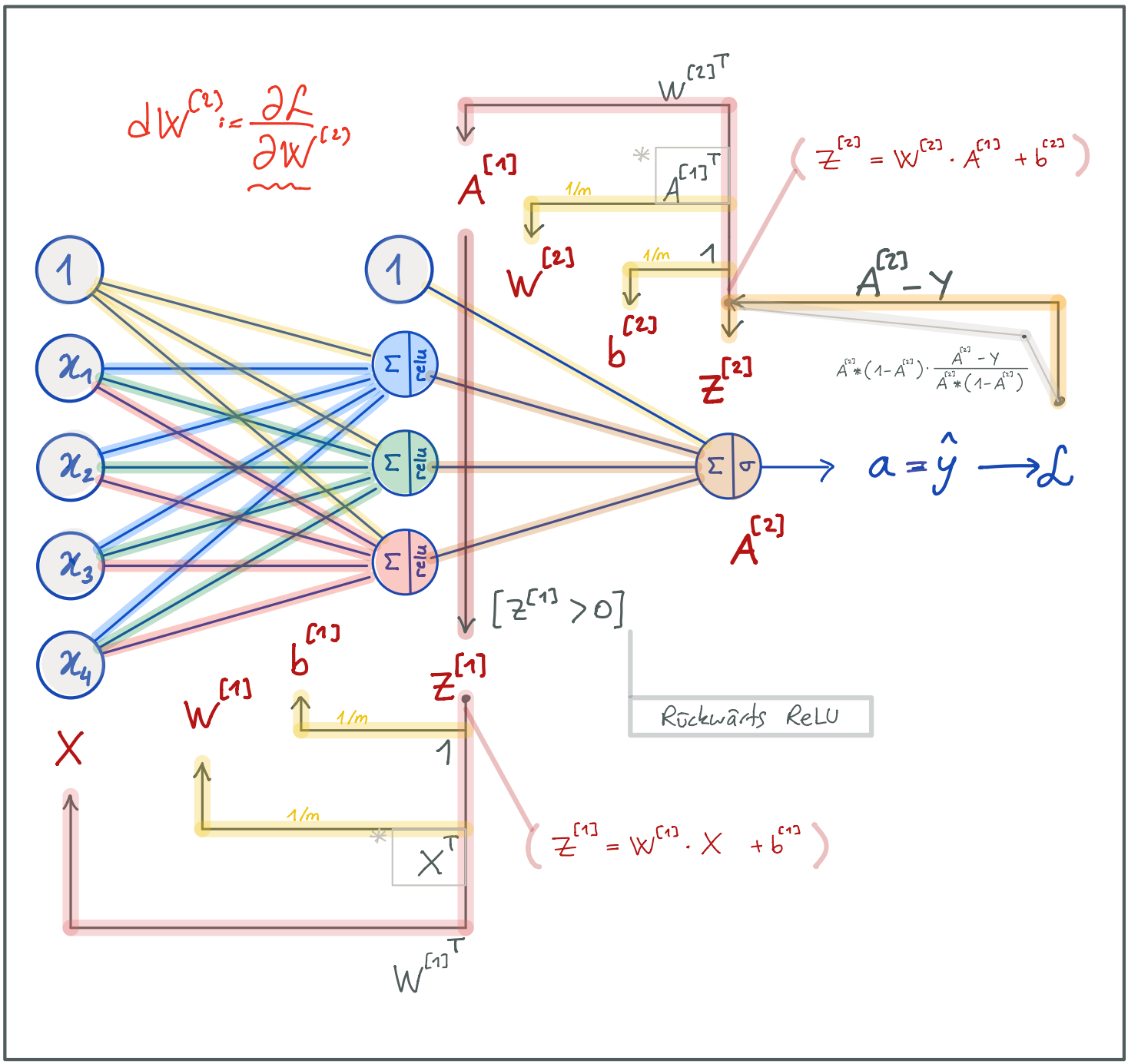
$$z^{[1]} = W^{[1]} \cdot a^{[0]} + b^{[1]}$$

Abhängigkeitsgraph



BACKWARD PASS





* : The right term to be multiplied with $dZ^{[2]}$, but not $\partial Z/\partial W$! See notes below!

• DIE ABLEITUNGEN

① $\frac{\partial \mathcal{L}}{\partial a}$ für $\boxed{\mathcal{L}(a,y) = -y \cdot \log_e(a) - (1-y) \cdot \log(1-a)}$

- $\frac{\partial \mathcal{L}}{\partial a} = \frac{-y}{a} - \frac{(1-y)}{(1-a)} \cdot (-1) = \boxed{\frac{-y}{a} + \frac{(1-y)}{(1-a)}}$

$$\frac{-y(1-a) + a(1-y)}{a(1-a)} = \frac{-y + ya + a - ay}{a(1-a)} = \boxed{\frac{a-y}{a(1-a)}}$$

② $\frac{da}{dz}$ für $\boxed{a = \sigma(z) = \frac{1}{1+e^{-z}}} \Rightarrow \frac{1}{a} = 1 + e^{-z}$

- $\frac{da}{dz} = \frac{d}{dz}(1+e^{-z})^{-1} = -1 \cdot (1+e^{-z})^{-2} \cdot (-e^{-z})$

$$= -1 \cdot a^2 \cdot \left(1 - \frac{1}{a}\right) \qquad A * (1-A)$$

$$= -a^2 \cdot \left(\frac{a-1}{a}\right) = \boxed{a(1-a)}$$

- $\frac{\partial \mathcal{L}}{\partial z} = \frac{\partial \mathcal{L}}{\partial a} \cdot \frac{\partial a}{\partial z} = \frac{a-y}{a(1-a)} \cdot a(1-a) = \boxed{a-y}$ $\qquad A^{[2]} - Y$

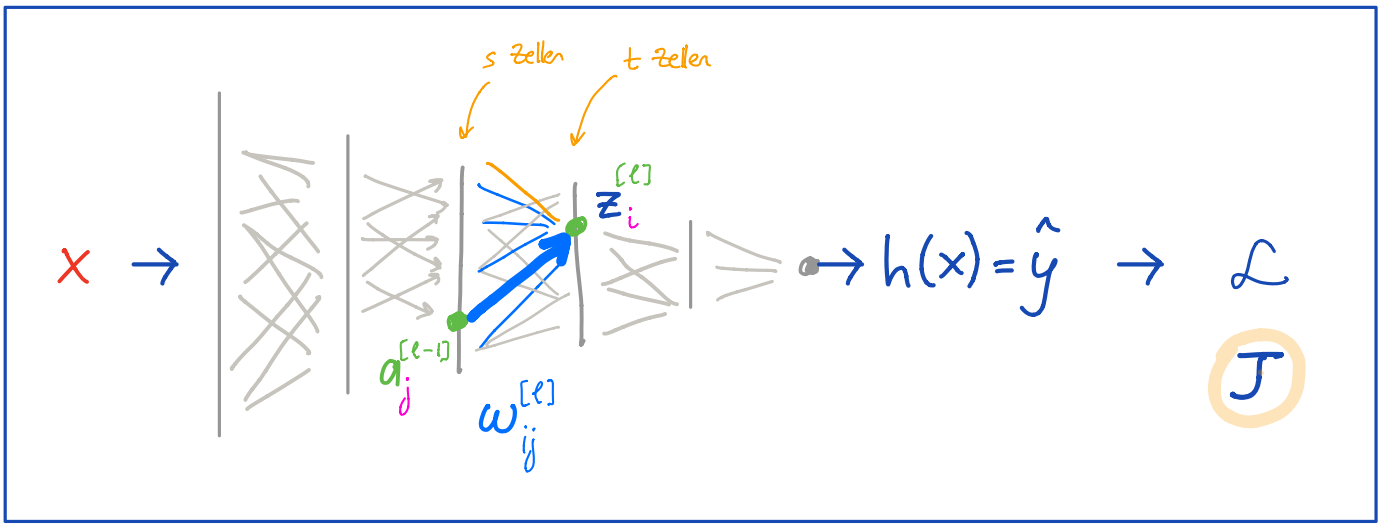
- $\frac{\partial \mathcal{L}}{\partial z^{[2]}} = \boxed{A^{[2]} - Y}$

This holds, as long as $A^{[2]} = \sigma(z^{[2]})$

$= [\bullet\bullet\bullet\bullet\bullet\bullet] - [\bullet\bullet\bullet\bullet\bullet\bullet] = [\bullet\bullet\bullet\bullet\bullet\bullet]$

canan.yildiz@tau.edu.tr

③ $\frac{\partial z}{\partial a}$ für $z = Wa + b$



$$\begin{bmatrix} w_{11} & \cdots & w_{1j} & \cdots & w_{1s} \\ \vdots & & \vdots & & \vdots \\ w_{i1} & \cdots & w_{ij} & \cdots & w_{is} \\ \vdots & & \vdots & & \vdots \\ w_{t1} & \cdots & w_{tj} & \cdots & w_{ts} \end{bmatrix} \cdot \begin{bmatrix} a_1^{[l-1]} \\ \vdots \\ a_j^{[l-1]} \\ \vdots \\ a_s^{[l-1]} \end{bmatrix} + \begin{bmatrix} b_1 \\ \vdots \\ b_i \\ \vdots \\ b_t \end{bmatrix} = \begin{bmatrix} z_1^{[l]} \\ \vdots \\ z_i^{[l]} \\ \vdots \\ z_t^{[l]} \end{bmatrix}$$

$$W^{[l]} \cdot a^{[l-1]} + b^{[l]} = z^{[l]}$$

$$\frac{\partial z_i^{[l]}}{\partial a_j^{[l-1]}} = w_{ij}^{[l]}$$

$$z_1^{[l]} = w_{11}^{[l]} \cdot a_1^{[l-1]} + \cdots + w_{1j}^{[l]} \cdot a_j^{[l-1]} + \cdots + w_{1s}^{[l]} \cdot a_s^{[l-1]} + b_1^{[l]}$$

$$\vdots$$

$$z_i^{[l]} = w_{i1}^{[l]} \cdot a_1^{[l-1]} + \cdots + w_{ij}^{[l]} \cdot a_j^{[l-1]} + \cdots + w_{is}^{[l]} \cdot a_s^{[l-1]} + b_i^{[l]}$$

$$\vdots$$

$$z_t^{[l]} = w_{t1}^{[l]} \cdot a_1^{[l-1]} + \cdots + w_{tj}^{[l]} \cdot a_j^{[l-1]} + \cdots + w_{ts}^{[l]} \cdot a_s^{[l-1]} + b_t^{[l]}$$

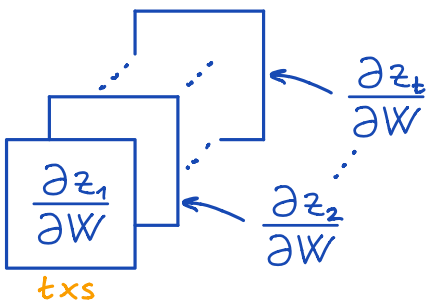
- Zwei Layout Optionen für $\partial \vec{z} / \partial \vec{a}$:
  - Numerator (Zähler) Layout
  - Denominator (Nenner) Layout

$$\circ \frac{\partial z}{\partial a} = \begin{bmatrix} \frac{\partial z_1}{\partial a_1} & \frac{\partial z_2}{\partial a_1} & \frac{\partial z_t}{\partial a_1} \\ \frac{\partial z_1}{\partial a_2} & \frac{\partial z_2}{\partial a_2} & \frac{\partial z_t}{\partial a_2} \\ \vdots & \vdots & \vdots \\ \frac{\partial z_1}{\partial a_s} & \frac{\partial z_2}{\partial a_s} & \frac{\partial z_t}{\partial a_s} \end{bmatrix}_{(s \times t)} = W^{(l)^T}$$

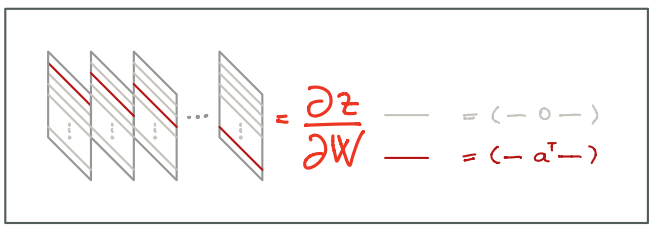
(Hessian Formulation oder "Denominator Layout")

---

④ $\frac{\partial z}{\partial W}$ für $z = Wa + b$



Vektor-nach-Matrix Ableitung: 3D-Tensor $(t \times s \times t)$



- $dW := \frac{\partial L}{\partial W}$ ist was wir eigentlich brauchen

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial z} \cdot \frac{\partial z}{\partial W}$$

[für Gewichtsaktualisierung $W := W - \alpha \cdot dW$]

↑

Mit einem Trick können wir diese Ableitungen einfacher berechnen. (ohne $\partial z / \partial W$ explizit auszurechnen).

- IDEE: Berechne den Gradienten für W Zeile-für-Zeile.

Bemerkung: Erste Zeile von W beeinflusst nur z_1!

Frage: Was wäre, wenn Schicht l nur die eine Zelle z_1 hätte?

Sei $w_{1\cdot}$ die erste Zeile von W. ($w_{i\cdot}$: i-te Zeile von W)

$$\frac{\partial \mathcal{L}}{\partial w_{1\cdot}} = dw_{1\cdot} = \frac{\partial \mathcal{L}}{\partial z_1} \cdot \frac{\partial z_1}{\partial w_{1\cdot}} = \frac{\partial \mathcal{L}}{\partial z_1} \cdot a^T \qquad \frac{\partial z_2}{\partial w_{1\cdot}} = 0$$

$$\frac{\partial \mathcal{L}}{\partial w_{2\cdot}} = dw_{2\cdot} = \frac{\partial \mathcal{L}}{\partial z_2} \cdot \frac{\partial z_2}{\partial w_{2\cdot}} = \frac{\partial \mathcal{L}}{\partial z_2} \cdot a^T$$

$$\vdots$$

$$\frac{\partial \mathcal{L}}{\partial w_{t\cdot}} = dw_{t\cdot} = \frac{\partial \mathcal{L}}{\partial z_t} \cdot \frac{\partial z_t}{\partial w_{t\cdot}} = \frac{\partial \mathcal{L}}{\partial z_t} \cdot a^T$$

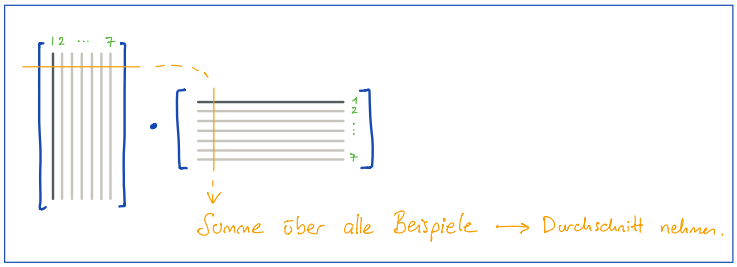
$$\Rightarrow dW = \begin{bmatrix} \text{—} & dz_1 \cdot a^T & \text{—} \\ \text{—} & dz_2 \cdot a^T & \text{—} \\ & \vdots & \\ \text{—} & dz_t \cdot a^T & \text{—} \end{bmatrix} = \begin{bmatrix} dz_1 \\ dz_2 \\ \vdots \\ dz_t \end{bmatrix} \cdot \begin{bmatrix} \text{—} & a^T & \text{—} \end{bmatrix}$$

$$dW = dz \cdot a^T$$

- Frage: Was passiert, wenn wir mehrere Beispiele haben? ($m > 1$)

$$dW = \frac{1}{m} dZ \cdot A^T$$

[ℓ] [ℓ] [ℓ-1]

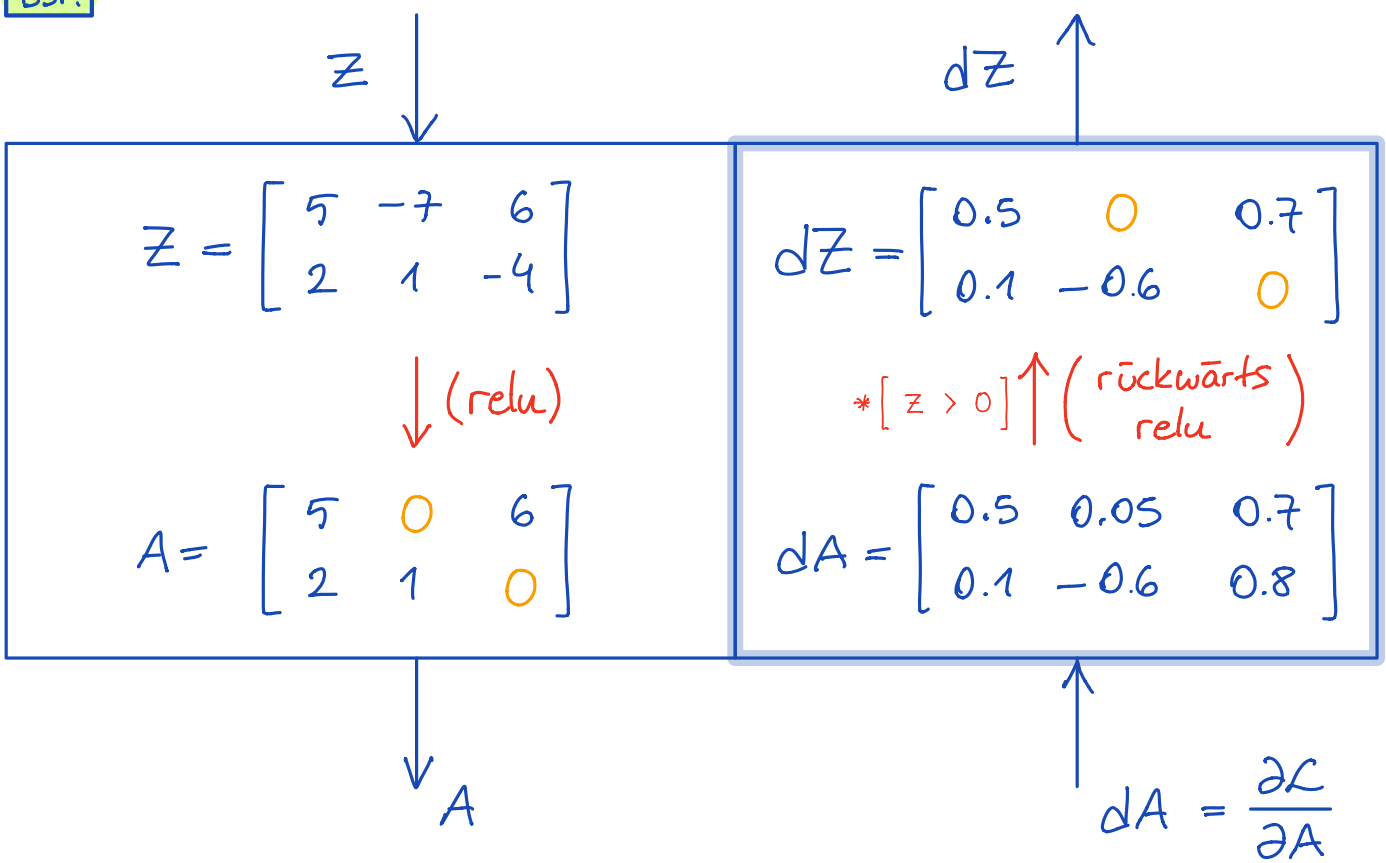


④ $\frac{dA}{dZ}$ für $A = relu(Z)$

- $\frac{dA}{dZ} = [\, Z > 0 \,] = [\, \text{True if } (Z > 0) \text{ else False} \,]$

$= [\, 1 \text{ if } (Z > 0) \text{ else } 0 \,]$

BSP.



$$\Rightarrow \frac{dA}{dZ} = [\, Z > 0 \,] = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

THE BACKPROP MAP

$\mathcal{L}$

$A^{[2]}$

$$dA^{[2]} := \frac{\partial \mathcal{L}}{\partial A^{[2]}} = \frac{A^{[2]} - Y}{A^{[2]} * (1 - A^{[2]})} = -Y/A^{[2]} + (1-Y)/(1-A^{[2]})$$

(1x7)

element-wise division

$A^{[2]} * (1 - A^{[2]})$

$Z^{[2]}$

$$dZ^{[2]} := \frac{\partial \mathcal{L}}{\partial Z^{[2]}} = dA^{[2]} * A^{[2]} * (1 - A^{[2]}) = \boxed{A^{[2]} - Y}$$

$W^{[2]T}$ $\quad$ $A^{[1]T}$ * $\quad$ 1 $\quad$ $Z^{[2]} = W^{[2]} \cdot A^{[1]} + b^{[2]}$

$A^{[1]}$

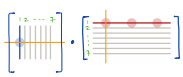
$$dA^{[1]} = W^{[2]T} \cdot dZ^{[2]}$$

(3x7) = (3x1) · (1x7)

$W^{[2]}$

$$dW^{[2]} = \frac{1}{m} \cdot dZ^{[2]} \cdot A^{[1]T}$$

(1x3) = (1x7) · (7x3)



$b^{[2]}$

$$db^{[2]} = \frac{1}{m} \cdot \text{np.sum}(dZ^{[2]})$$

(1x1) = (1x7), axis = 1

$[Z^{[1]} > 0]$

$Z^{[1]}$

$$dZ^{[1]} := \frac{\partial \mathcal{L}}{\partial Z^{[1]}} = dA^{[1]} * [Z^{[1]} > 0] \text{ (element-wise)}$$

(3x7) $\quad$ (3x7)

$W^{[1]T}$ $\quad$ X^T * $\quad$ 1

X

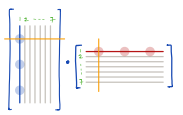
$$dX = W^{[1]T} \cdot dZ^{[1]}$$

(4x7) = (4x3) · (3x7)

$W^{[1]}$

$$dW^{[1]} = \frac{1}{m} \cdot dZ^{[1]} \cdot X^T$$

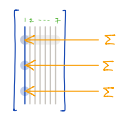
(3x4) = (3x7) · (7x4)



$b^{[1]}$

$$db^{[1]} = \frac{1}{m} \cdot \text{np.sum}(dZ^{[1]})$$

(3x1) = (3x7), axis = 1

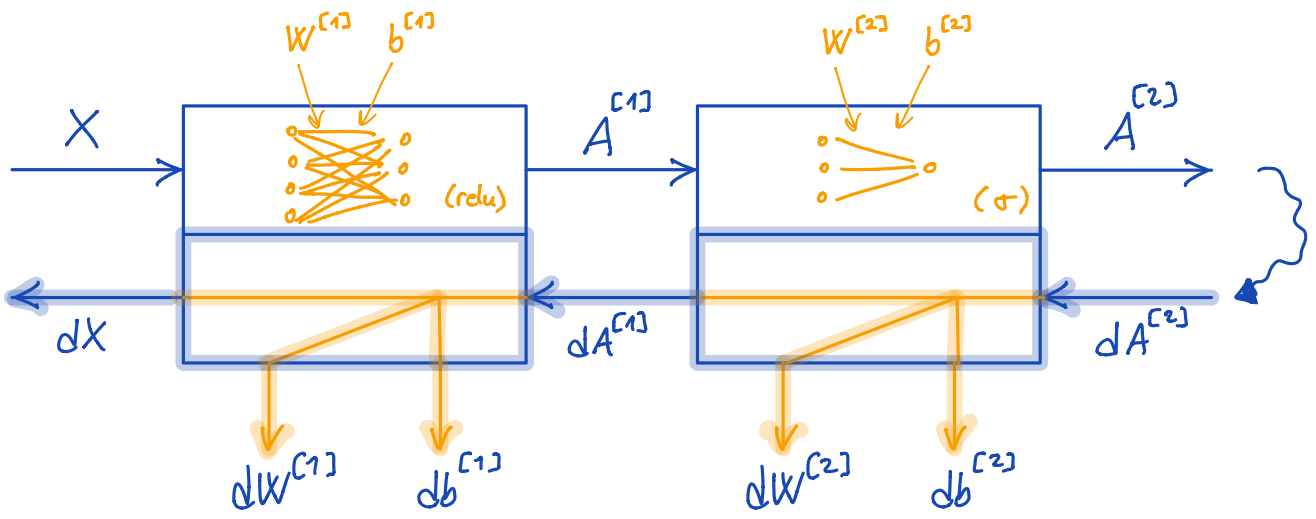


* Nicht gleich $\partial Z / \partial W$, aber der richtige Term in der "Kette". Siehe Notizen unten.

canan.yildiz@tau.edu.tr

• PROZESSABLAUF IM ÜBERBLICK

FORWÄRTSLAUF



RÜCKWÄRTSLAUF

GEWICHTSAKTUALISIERUNGEN

$$\begin{bmatrix} W^{[1]} \\ b^{[1]} \\ W^{[2]} \\ b^{[2]} \end{bmatrix}_{neu} := \begin{bmatrix} W^{[1]} \\ b^{[1]} \\ W^{[2]} \\ b^{[2]} \end{bmatrix}_{alt} - \alpha \cdot \begin{bmatrix} dW^{[1]} \\ db^{[1]} \\ dW^{[2]} \\ db^{[2]} \end{bmatrix}$$