Blatt 7

Luca Krüger, Jonas Otto, Jonas Merkle (Gruppe R)

25. Juni 2019

1 Learning Slowdown

1. Einfluss der Parameter im zweiten Neuron:

   a) Initiales Gewicht $w_2 = 0.6$ und Bias $b_2 = 0.9$

   $$\begin{pmatrix} \frac{\partial E}{\partial w_2} \\ \frac{\partial E}{\partial b_2} \end{pmatrix} = \begin{pmatrix} 0.243877 \\ 0.243877 \end{pmatrix}$$

   b) Initiales Gewicht $w_2 = 2$ und Bias $b_2 = 2$

   $$\begin{pmatrix} \frac{\partial E}{\partial w_2} \\ \frac{\partial E}{\partial b_2} \end{pmatrix} = \begin{pmatrix} 0.03469 \\ 0.03469 \end{pmatrix}$$

   c) Das Gewicht w_2 und Bias b_2 stehen mit negativem Vorzeichen in der Exponentialfunktion.

   $$\lim_{w_2, b_2 \to \infty} \nabla E = 0 \quad \lim_{w_2, b_2 \to 0} \nabla E \propto y_1 T_1 + c$$

2. a) Einfluss der Parameter im ersten Neuron:

   $$\begin{aligned} \frac{\partial E}{\partial b_1} &= \underbrace{\frac{\partial E}{\partial y_2} \frac{\partial y_2}{\partial u_2}}_{\text{alt}} \frac{\partial u_2}{\partial y_1} \frac{\partial y_1}{\partial u_1} \frac{\partial u_1}{\partial b_1} \\ &= -2(T - y_2) \cdot f'(u_2) \cdot w_2 \cdot f'(u_1) \end{aligned}$$

   $$\frac{\partial u_2}{\partial y_1} = w_2 \qquad \frac{\partial u_1}{\partial b_1} = 1 = const.$$

   Problematisch ist wieder:

   $$\frac{\partial y_1}{\partial u_1} = f'(u_1) = f(u_1)(1 - f(u_1)) = f(u_1) \underbrace{(1 - \frac{1}{1 + e^{-u_1}})}_{\to 0 \text{ für } u_1 \to \infty}$$

   $\Rightarrow$ das Problem verstärkt sich mit dem ersten Neuron.

b) Das Problem entsteht aus der Mehrfachanwendung der Kettenregel und verstärkt sich mit jeder zusätzlichen Zwischenschicht.

c) Der Gradient Descent Algorithmus zur Bestimmung eines lokalen Minimums ist im jeweiligen Iterationsschritt direkt abhängig vom Gradienten der Errorfunktion. Die Suche nach einem lokalen Minimum erfolgt also in Schritten, wobei die Schrittweite $d \propto \nabla E$ dem Gradienten der Errorfunktion.

3. a)

$$\frac{\partial E}{\partial b_2} = \frac{1}{4}$$

$$\frac{\partial E}{\partial b_1} = 12.5$$

b) In jedem Schritt kommt durch weiteres Ableiten der Faktor $f'(0)^2 * 100 = 12.5$ hinzu.

c) (Divergenz) Der Gradient divergiert. Dies führt zu Problemen wenn der Fehler bereits gering ist, aber der noch sehr große Gradient eine Konvergenz im Minimum verhindert.

4. In der Ausgabeschicht ist die Ableitung der Cross-Entropy Funktion E_c nicht mehr direkt abhängig von $f'(x)$. Auch die Zwischenschichten werden weniger durch $f'(x)$ beeinflusst. Das Problem aus b) wird dadurch reduziert.

2 Flat vs. Deep Networks

1. a) Simulation für $x = (0, 0)$ und $y = (1, 1)$. Netzwerkausgabe:

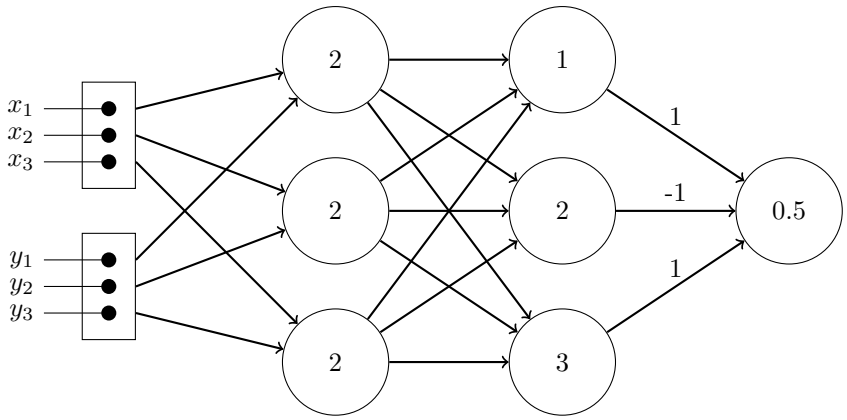
$$\begin{aligned} u_{11} = 1 < 2 \qquad & u_{12} = 1 < 2 \\ y_{11} = 0 \qquad & y_{12} = 0 \\ u_{21} = 0 \qquad & u_{22} = 0 \\ y_{21} = f(x, y) &= 0 \end{aligned}$$

Ergebnis aus Gleichung 4:

$$f(x, y) = \langle x, y \rangle \bmod 2 = 0$$

$\implies$ Netzwerkausgabe korrekt

b) Die Zahlen in den Neuronen stellen jeweils den Schwellwert dar, unbeschriftete Pfeile kennzeichnen Gewicht 1.



c) Die Neuronen in der ersten Schicht realisieren ein logisches `AND` der einzelnen Komponenten der beiden Eingangsvektoren. Die zweite Schicht „zählt “, welche Neuronen aus der ersten Schickt aktiv sind indem jedes Neuron ein logisches `OR` implementiert.