Blatt 10

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

16. Juli 2019

1 Transferfunktionen

1. a)

$$\text{zz: } \operatorname{sig}(x) = \frac{1 + \tanh\frac{x}{2}}{2}$$

$$\begin{aligned}
&= \frac{1}{2}\left(1 + \frac{\exp\left(\frac{x}{2}\right) - \exp\left(\frac{-x}{2}\right)}{\exp\left(\frac{x}{2}\right) + \exp\left(-\frac{x}{2}\right)}\right) \\
&= \frac{1}{2}\left(\frac{\exp\left(\frac{x}{2}\right) + \exp\left(\frac{-x}{2}\right) + \exp\left(\frac{x}{2}\right) - \exp\left(\frac{-x}{2}\right)}{\exp\left(\frac{x}{2}\right) + \exp\left(\frac{-x}{2}\right)}\right) \\
&= \frac{1}{2}\left(\frac{2\exp\left(\frac{x}{2}\right)}{\exp\left(\frac{x}{2}\right) + \exp\left(\frac{-x}{2}\right)}\right) = \frac{\exp\left(\frac{x}{2}\right)}{\exp\left(\frac{x}{2}\right) + \exp\left(\frac{-x}{2}\right)} \\
&= \frac{1}{1 + \exp\left(\frac{-x}{2} + \frac{-x}{2}\right)} \\
&= \frac{1}{1 + e^{-x}}
\end{aligned}$$

b) Vorzeichen der Gewichte:

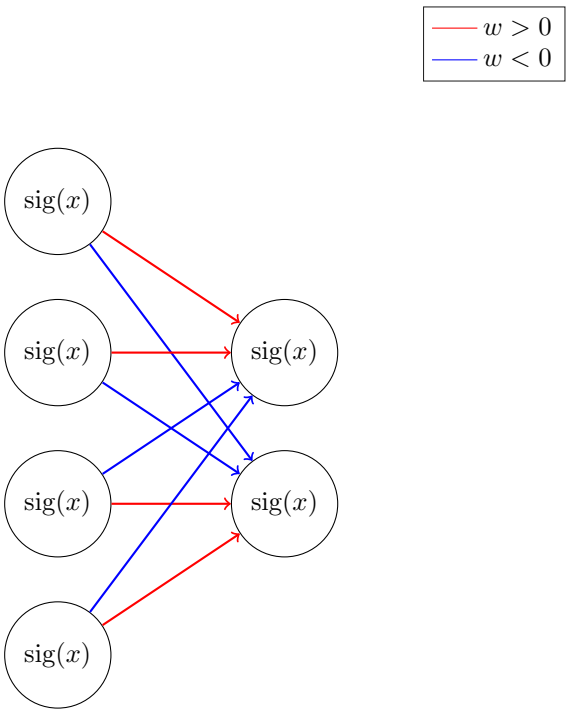

Abbildung 1: Tanh-Neuronen

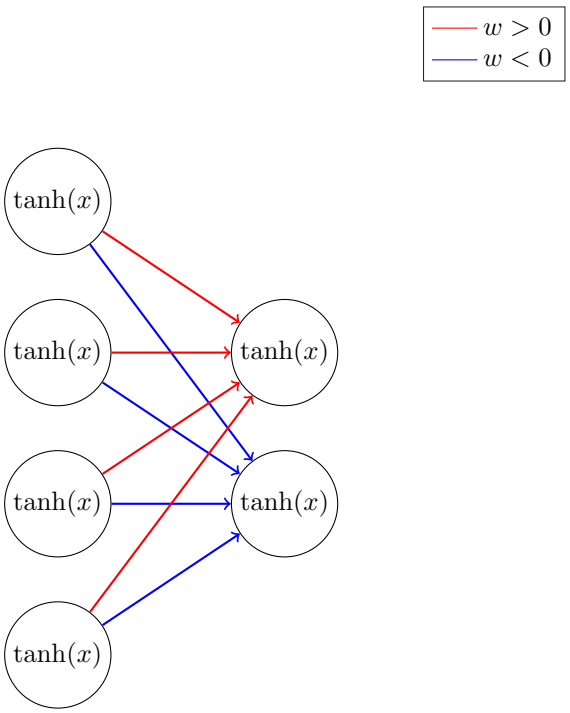

Abbildung 2: Sigmoid-Neuronen

2. a) Unter der Verwendung von *Rectified Linear Units* ist für dentritische Potential $u < 0$ die Ausgabe des Neurons $y = 0$ sowie der Gradient $\frac{\partial E}{\partial w} = 0$. Dadurch kommt es bei dem Neuron zu einem *Learning Slowdown*. Dieses Verhalten propagiert sich außerdem durch vorhergehende Schichten im Netzwerk.

   b) $\tanh(x)$ erfordert aufgrund der komplexen Berechnung von Winkelfunktionen einen hohen Rechenaufwand. Die Performance des Lernvorgang ist daher wahrscheinlich erheblich besser, wenn eine weniger komplexe Funktion wie $\mathrm{ReLU}(x)$ verwendet wird.

   c) Für $u_i < 0$ gilt:

$$\mathrm{ReLU}(u_i) = y_i^{(1)} = 0 \;\Rightarrow\; \frac{\partial E}{\partial w_{ij}} = \delta_j^{(2)} y_i^{(1)} = 0$$

   In der Regel ist der rückwärts propagierte Fehler $\delta_i^{(k)} \propto f'(u_i^{(k+1)})$, wonach in unserem Fall ($u_i < 0$) und $\mathrm{ReLU}'(u_i^{(k+1)}) = 0$ auch Lernschritte vorheriger Neuronen blockiert werden.
   Für Leaky-ReLU(x) als Transferfunktion treten diese Probleme nicht

auf, da für den Fall $u_i < 0$ gilt:

$$\text{Leaky-ReLU}(u_i) \neq 0 \quad \wedge \quad \text{Leaky-ReLU}'(u_i) \neq 0$$

3. a) Leaky-ReLU(x) ist in einer kleinen Umgebung von $x = 0$ stark asymmetrisch. Somit ist die Klassifiezierung im Neuron wesentlich fehleranfälliger gegenüber einem Rauschen.
Außerdem gilt

$$\lim_{x \to \infty} \text{Leaky-ReLU}(x) = -\infty$$

Extremwerte bekommen also eine hohe Gewichtung und es kann zu negativen *exploding Gradients* kommen.

b) i) Die axonalen Potentiale haben direkten Einfluss auf den Gradienten der Errorfunktion und dem damit verbundenen Lernverhalten. Eine Gleichverteilung der axonalen Potentiale nach ähnlichen Verteilungparametern in jeder Schicht führt also auch zu ähnlichem Lernverhalten über alle Schichten hinweg. In diesem Beispiel mit unterschiedlichen Verteilungsparametern wird die $l+1$ Schicht in jeder Epoche mit unterschiedlichen Eingangsdaten trainiert.

ii) Für $u_i > 0$ gilt $\text{SELU}(x) = x$, die Verteilung bleibt also unverändert.
Für $u_i < 0$ folgt:

$$\lim_{x \to -\infty} \text{SELU}(x) = const. \approx -1.76$$
$$\Rightarrow \text{SELU}(x) \in [-1.76, 0]$$

Die Werte werden also von ursprünglich $u \in [-\infty, 0]$ auf den Wertebereich von SELU(x) skaliert.

$$\Rightarrow \quad \sigma \downarrow \quad \wedge \quad \mu \uparrow$$

iii) Ein kleiner Wertebereich von $u \in [-1.25, 0]$ wird durch SELU(x) getreckt, da

$$\text{SELU}(u) < u \quad \forall u \in [-1.25, u]$$

$$\Rightarrow \quad \sigma \uparrow \quad \wedge \quad \mu \downarrow$$

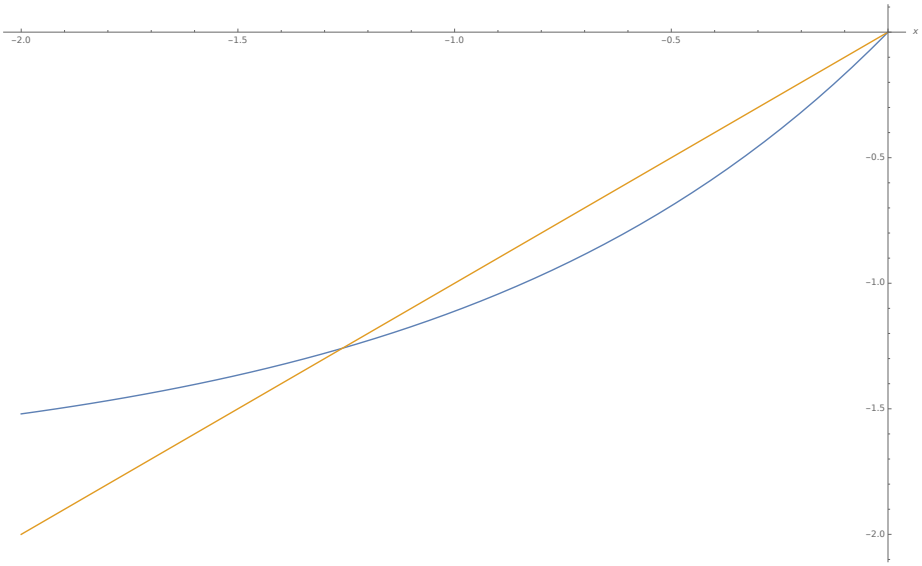

Abbildung 3: Überhang der SELU-Funktion für betragsmäßig kleine dendritische Potentiale

4. i) Vergleich der verschiedenen Transferfunktionen in der Anwendung:

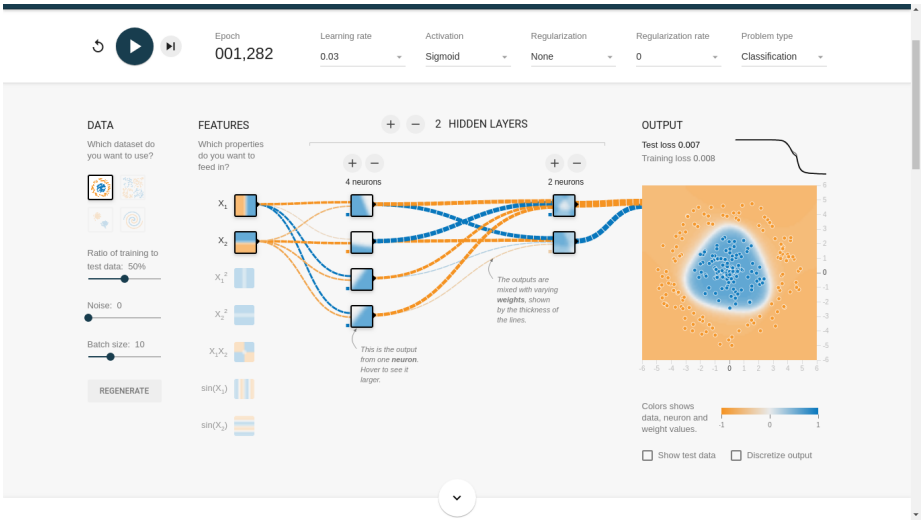

Abbildung 4: Transferfunktion sig

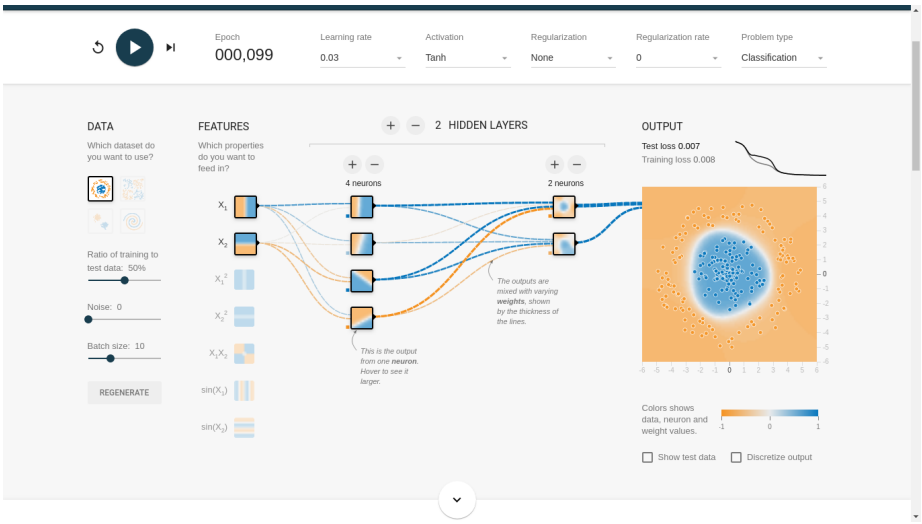

Abbildung 5: Transferfunktion $\tanh(x)$

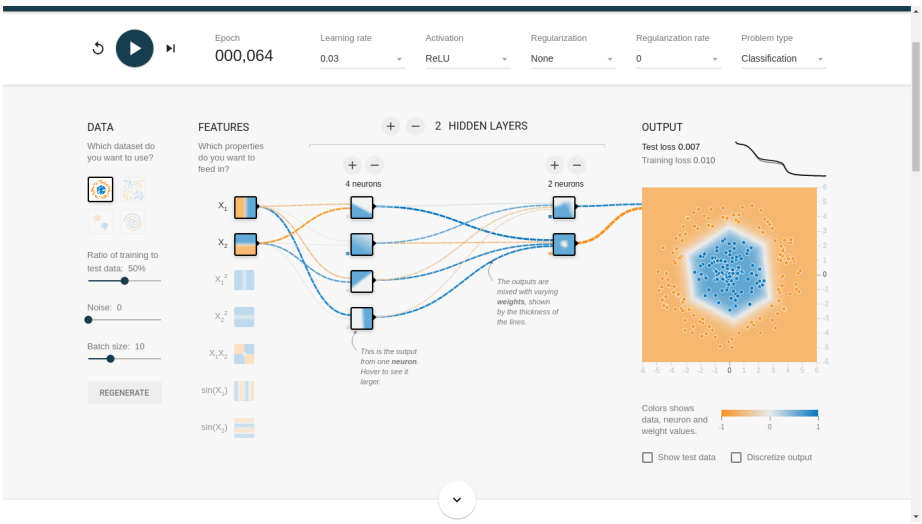

Abbildung 6: Transferfunktion ReLU

ii) In folgendem Beispiel leiden in der vorletzten Schicht das erste und dritte Neuron unter dem Effekt von *dying*ReLU. Dadurch, dass die Beiden Neuronen kurz hinter der Ausgabeschicht liegen, kommt auch der Lernvorgang in den vorhergehenden Schichten durch diesen Effekt zum erliegen. Das lässt sich auch an dem nahezu konstanten Fehler in der Ausgabeschicht beobachten.

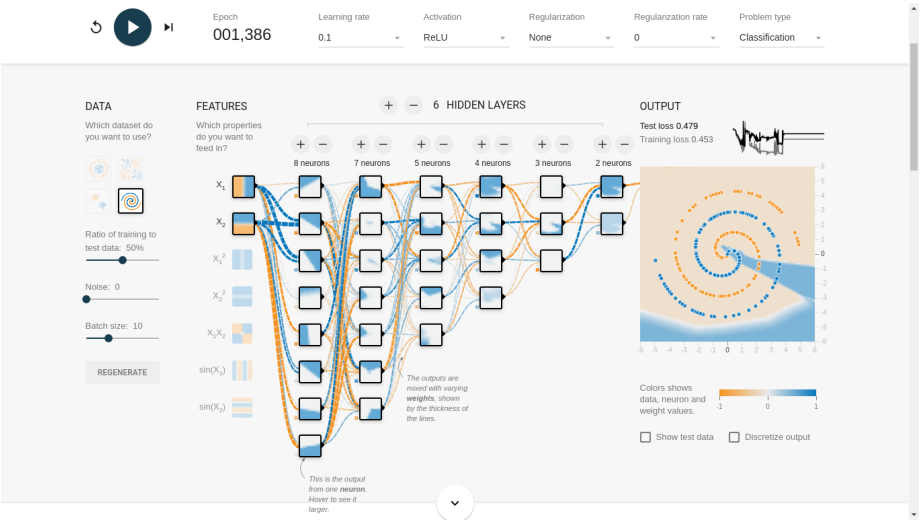

Abbildung 7: Dying ReLU in einem komplexeren Netzwerk

5. a) i. Bei der SELU Aktivierungsfunktion ist deutlich zu erkennen, dass die Standardabweichung der Aktivierungen für tiefere Schichten gegen 1 geht. Für den Gradienten ist zu erkennen, dass der Mittelwert während der Rückwärtsphase gegen 0 geht.

ii. Die Asymetrie der ReLU und ELU Funktionen ist am Stärksten bei den Aktivierungen der ersten Schichten ersichtlich. Besonders bei ReLU ist nicht nur der Mittelwert stark von 0 verschieden, auch der Wertebereich ist stark asymetrisch.

iii. Die Aktivierungen liegen in dem Bereich, in dem die Ableitung von tanh noch nicht gegen 0 geht.

iv. Die SELU Transferfunktion weist die größten Gradienten auf, folglich wird mit dieser Funktion am meisten gelernt.

b) 3D Plot

tanh Die Verteilung wird bei tieferen Schichten zunehmend schmaler, da die Funktionswerte des tanh zwischen 0 und 1 liegen, und die Funktionswerte von diesen Werten in immer kleinerem Wertebereich liegen.

ReLU Die Verteilung wird noch schmaler als bei tanh, da viele der Eingabewerte ($u < 0$) direkt auf 0 abgebildet werden.

ELU Auch hier treten die gleichen Effekte wie bei tanh und ReLU auf, zusätzlich ist besonders bis ca. Layer 10 die Asymetrie der Transferfunktion in der Verteilung zu erkennen.

SELU Hier werden die oben betrachteten normalisierenden Eigenschaften der SELU Funktion ersichtlich: Die Verteilung bleibt über die Layer hinweg breit und konzentriert sich nicht um 0.