Aufgabe: Wie viele Neuronen sind nötig, damit ein neuronales Netzwerk den Text von Genesis 1-3 in sumerischer Keilschrift entziffern kann? Erläutern Sie anhand eines einfachen Beispiels, wie Text- und Spracherkennung basierend auf neuronalen Netzwerken funktionieren.

Lösung: Der Text Genesis 1-3 besteht aus insgesamt 50 Silben, von denen 35 unterschiedlich sind.

1	AM	8	UND	12	WAR	23	GOT	5	GOTT
2	AN	11	DIE	16	FIN	24	TES	32	SPRACH
3	FANG	9	ER	17	STER	25	SCHWEB	15	ES
4	SCHUF	10	DE	18	AUF	26	TE	33	WER
5	GOTT	12	WAR	19	DER	27	Ü	10	DE
6	HIM	13	WÜST	20	TIE	28	BER	34	LICHT
7	MEL	8	UND	21	FE	29	DEM	8	UND
8	UND	14	LEER	8	UND	30	WAS	15	ES
9	ER	8	UND	19	DER	31	SER	35	WARD
10	DE	15	ES	22	GEIST	8	UND	34	LICHT

Abbildung 1. Silbenzerlegung des Textes der drei ersten Sätze der Genesis

Wir brauchen also mindestens 6 Bit, mit denen wir bis zu $2^6 = 64$ Fälle binär unterscheiden können. Wie viele Silben die deutsche Sprache enthält, ist unbekannt. Da die Sumerer die binäre Schreibweise nicht kannten, brauchten sie für jede Silbe ein eigenes Symbol. Damit ein neuronales Netzwerk die lateinische Buchstabenschrift mit 27 Buchstaben und 10 arabischen Ziffern lesen kann, bedarf es ebenfalls nur 6 Bit, wenn man auf Umlaute verzichten will, wobei man noch zwischen Groß- und Kleinbuchstaben wählen kann. Die Silben- oder Buchstabenerkennung sagt allerdings noch nichts darüber aus, ob man den Text auch verstanden hat. Zusätzlich zur Buchstabenerkennung muß das Gehirn auch noch die Wortbedeutung anhand eines Wort-Bild-Vergleichs kennen und damit den kompletten Wortschatz. Trotzdem darf die Leistung der Sumerer weg von der Bilder- hin zur Silbenschrift nicht unterschätzt werden, zumal man durch unterschiedliche Silbenfolgen anhand von Zeichen, die ursprünglich eine völlig andere Bedeutung besaßen, plötzlich ganz neue Sätze schreiben konnte.

Der gesamte Text besteht aus 157 Buchstaben und 38 Leerzeichen, also aus insgesamt 196 Zeichen, Satzzeichen nicht eingerechnet. Insofern ist die Silbenschrift effizienter, weil sie weniger Erkennungsvorgänge benötigt. Die neuronalen Trainingsmuster für die 35 Silben einschließlich des Leerzeichens sind in Tabelle 1 und 2 dargestellt.

Gemäß dem Pascalschen Dreieck reichen 5 Bit gerade nicht mehr aus, um alle 35 Symbole darzustellen, denn $1+5+10+10+5+1=32$. Wir benötigen daher eine Kodierung mit 6 Bit. Damit lassen sich neben dem Leerzeichen mit gar keinem Abdruck 6 weitere Zeichen mit einem „Loch“ im Ton, 15 mit zwei, 20 mit drei und entsprechend wieder 15 mit vier bis hin zu 6 mit fünf und einem noch verbliebenen mit sechs Abdrücken darstellen, was in der Summe $1+6+15+20+15+6+1=64$ Zeichen ergibt, von denen wir nicht einmal die 20 dreiwertigen voll ausschöpfen können. Die Sumerer ritzten mit einem Griffel Löcher in weiche Tontäfelchen,

die sie anschließend trocknen ließen oder brannten. Wir verwenden hier als Löcher schwarze Kreise,[1] die in einem Muster aus 6 Feldern angeordnet sind.[2]

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0
0	1	0	0	0	0	1	0	0	0	0	1	1	1	1	0	0	0
0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	1	1	1
0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	1	0	0
0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	1	0
0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	1

Tabelle 1. Die ersten 18 Silben des Textes in willkürlichen Binärmustern, die Keilschriftsymbolen entsprechen

In Abb. 2 ist das von uns verwendete Keilschriftalphabet angegeben, welches die Objekterkennung recht einfach macht. Wir übersetzen nun den Text Genesis 1-3 in Keilschriftzeichen:

„Am Anfang schuf Gott Himmel und Erde. Und die Erde war wüst und leer, und es war finster auf der Tiefe; und der Geist Gottes schwebte über dem Wasser. Und Gott sprach: Es werde Licht! Und es ward Licht."

19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0
0	0	0	1	1	1	1	0	0	0	0	0	0	1	1	1	1	0
0	0	0	1	0	0	0	1	1	1	0	0	0	1	1	1	0	0
1	1	0	0	1	0	0	1	0	0	1	1	0	1	0	0	1	0
1	0	1	0	0	1	0	0	1	0	1	0	1	0	1	0	1	0
0	1	1	0	0	0	1	0	0	1	0	1	1	0	0	1	0	0

Tabelle 2. Weitere 18 Silben des Textes zur Übersetzung der Keilschriftsymbole

Die komplette Tontafel oder eine Seite des tönernen Buches liest sich dann wie in Abb. 3 dargestellt. Um den kompletten Text einer Sprache in Form von Silben anzugeben, brauchen wir natürlich noch einige weitere Neuronen, jedoch bleibt deren Zahl überschaubar. Die Darstellung hat den Vorteil, daß sie von einem neuronalen Netzwerk anhand einer optischen Sichtverbindung leicht gelesen werden kann.[3] Dennoch kommt man mit weitaus weniger Zeichen aus als beispielsweise in der chinesischen Bilderschrift. Die Buchstabenschrift erfordert die geringste Zahl von Symbolen, die man sich merken muß und ist daher am fortschrittlichsten. Bis es zum ersten Buchstabenalphabet kam, das von den Phöniziern entwickelt wurde, von denen es schließlich die Griechen und später die Lateiner übernahmen, dauerte es allerdings noch einige Jahrtausende.

[1] Die den Schattenwurf repräsentieren sollen

[2] Die wirkliche Keilschrift ist natürlich bedeutet komplexer, weil nicht nur Löcher, sondern auch federähnliche Symbole geritzt wurden, die einfach nur einprägsamer sind. Das spielt für uns jedoch keine Rolle.

[3] Man kann den Wert der Digitalisierung daher nicht hoch genug einschätzen.

In Abb. 4 ist schließlich das dazugehörige neuronale Netzwerk dargestellt, welches die sumerische Keilschrift lesen kann.[4] Jedes der 6 Eingangsneuronen fragt ausschließlich das Eingangssignal seines eigenen Bits ab.[5] Das ist geometrisch recht einfach, falls das Tontäfelchen nicht Kopf steht. Wenn z.B. Bit 5 auf 1 gesetzt ist, feuert Neuron 5 und schickt sein Signal an alle Ausgangsneuronen. Aufgrund des vorher trainierten Netzes wird die Silbe zweifelsfrei identifiziert. Das neuronale Ausgangsmuster bekommt also zeitgleich die eingehenden Signale übermittelt und kann sie via Printbefehl niederschreiben oder zurück in Sprache verwandeln.

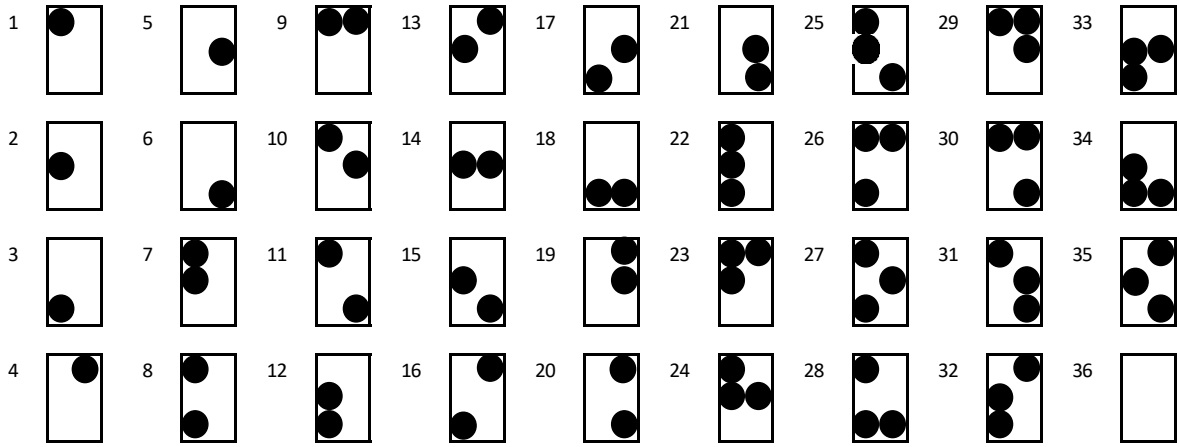

Abbildung 2. Die ersten 36 Keilschriftsilben

Jedes Eingangsneuron erhält seine analoge Eingangsinformation in Form von optischen oder akustischen Wellen mitgeteilt, aber immer nur vom gleichen Bildelement, für welches dieses Neuron zuständig ist.[6] Sämtliche Eingangsneuronen erhalten daher ihre Information stets gleichzeitig, da Licht sich „unendlich“ schnell ausbreitet.[7] Entweder das Neuron überschreitet die Reizschwelle wie im Falle von Abb. 4 das Neuron mit der Nummer 5 und feuert daraufhin oder es feuert nicht wie all die anderen Neuronen, deren Signal unter der Reizschwelle liegt. Erst die Ausgangsneuronen stellen aus den Informationen der Eingangsneuronen die Ausgangsinformation zusammen, die eine Kopie des eingehenden Signals ist. Da dem neuronalen Netzwerk die Bedeutung des Eingangssignals aber vorher durch Training bekanntgemacht worden ist, führt es den mit diesem Muster gekoppelten Druckbefehl instantan aus und schreibt die entsprechende Silbe nieder. Das Verfahren funktioniert allerdings nur, wenn die Eingangssignale innerhalb der Anstiegszeit der Amplituden gleichzeitig eintreffen, denn bei Verzögerungen können sich ganz andere Silbenbedeutungen ergeben. Bei einem technischen System muß daher ein Takt eingeführt werden, mit dem die Botschaften ausgelesen werden. An der Synchronisierung wird das Problem jedenfalls nicht scheitern.

Bei optischen Signalen werten die Eingangsneuronen pixelweise den Bildinhalt aus oder entsprechend die Ortsfrequenzen des jeweiligen Pixels. Den Ausgangsneuronen obliegt es, die Pi-

[4] Auch die echten Silbenzeichen der Keilschrift müssen nur darauf trainiert werden, erkannt zu werden.
[5] Dargestellt ist hier unser festgelegtes Keilschriftsymbol für „Gott“.
[6] Ähnlich wie auf unserer Netzhaut jede Sehzelle nur ihr eigenes Eingangssignal mißt
[7] Das gleiche gilt bedingt auch für Schallwellen.

xel wieder zu einem Bild und die Frequenzen wieder zu Tönen und Klängen zusammenzusetzen[8] und bekannte Bild- oder Tonobjekte zu erzeugen. Bei gesprochenen Silben kommen die Töne als Frequenzverläufe an. Je nachdem, in wie viele Intervalle man den Frequenzverlauf unterteilt, muß das zerlegte Signal an entsprechend viele Neuronen zur Auswertung weitergeleitet werden. Erst die Ausgangsneuronen setzen wie gesagt die Signale wieder zu Wörtern zusammen und können damit Sprache fortlaufend erkennen.

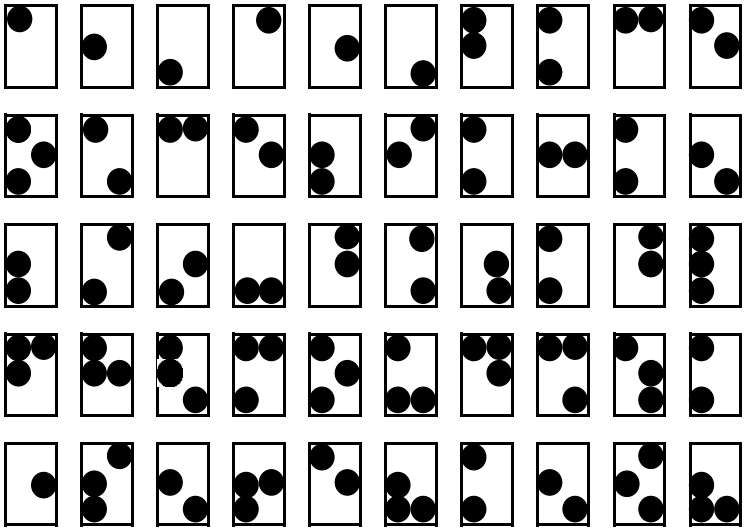

Abbildung 3. Tontäfelchen in sumerischer Keilschrift mit dem Text aus Genesis 1-3

Unsere Netzhaut besitzt ungefähr 130 Millionen Sehzellen. Etwas mehr als 100 Sehzellen werden zu einer sogenannten Schaltzelle zusammengefaßt, von der die elektrischen Impulse ausgehen, die ins Gehirn gelangen. Im Innenohr befinden sich analog nur etwa 30 bis 150 haarartige Fortsätze am oberen Ende der Stereozilien, welche die Nervenimpulse auslösen, die ans Gehirn weitergeleitet werden. Die Sprachwahrnehmbarkeit des Menschen liegt dabei zwischen 80 Hz und 12 kHz. Der Mensch kann somit etwa 400 000 Töne unterscheiden, und das in einem Bereich zwischen 16 Hz und 20 kHz Gesamthörfähigkeit. Bis ca. 500 Hz liegt das Unterscheidungsvermögen bei etwa 0,5 Hz.

Sei p die Proportion bzw. das Frequenzverhältnis

$$p = \frac{f_2}{f_1}.$$

Das logarithmische Intervallmaß i definiert sich dann durch

$$i = \frac{\lg p}{\lg 2} \cdot 1\,\text{Oktave} = 1200 \frac{\lg\left(f_2/f_1\right)}{\lg 2}\,\text{Cent} = 3986{,}3\,\text{Cent} \cdot \lg \frac{f_1 + \Delta f}{f_1}.$$

Daraus folgt ein Frequenzverhältnis von

[8] Auf die Schalldruckpegel gehen wir hier der Einfachheit halber nicht ein.

$$\frac{\Delta f}{f_1} = 10^{\frac{i}{3986{,}3\,\text{Cent}}} - 1.$$

Ein Mensch mit einem absoluten Gehör von 4 Cent kann demnach Frequenzen im unteren Frequenzbereich von 16 Hz mit 0,037 Hz auflösen, den oberen Bereich bei 20 kHz auf 45,5 Hz genau. Ein normaler[9] Mensch kann allerdings lediglich einen Halbton (100 Cent) unterscheiden, das entspricht einer Bandbreite von rund einem Hertz bei der Tiefstfrequenz von 16 Hz und 1,2 kHz bei der Höchstfrequenz von 20 kHz. Demgemäß ergibt sich für den Standardfall eine relative Frequenzauflösung von

$$\frac{\Delta f}{f_1} = 10^{\frac{100}{3986{,}3}} - 1 = 0{,}05974\,.$$

„Die Hörschwelle liegt zwischen 2000 Hz und 5000 Hz am niedrigsten, dort hört der Mensch am besten, hier treten auch die meisten Laute der gesprochenen Sprache auf.“ (Wikipedia)

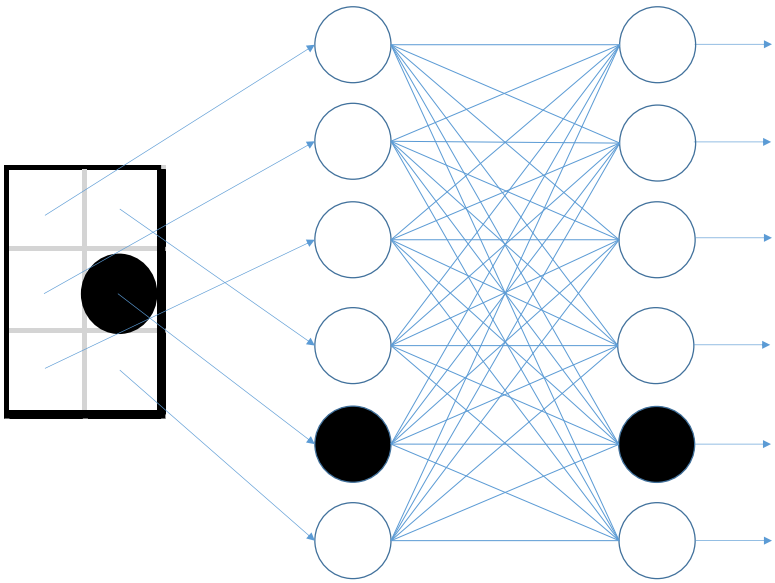

Abbildung 4. Neuronales Netzwerk zur Erkennung von Keilschriftzeichen

Das gesamte hörbare Frequenzintervall $\Delta f = f_{\max} - f_{\min} = 3000$ Hz muß dazu in eine Folge

$$\Delta f_k = 0{,}05974\, f_k, \quad f_{k+1} = (1 - 0{,}05974)\, f_k, \quad f_1 = f_{\max}$$

bzw.

$$f_k = (1 - 0{,}05974)^{k-1} f_{\max}$$

zerlegt werden, derart daß

[9] untrainierter

$$\Delta f = \sum_{k=1}^{n} \Delta f_k = \left[10^{\frac{100}{3986,3}} - 1 \right] \sum_{k=1}^{n} f_k = 0{,}05974 f_{\max} \sum_{k=1}^{n} (1 - 0{,}05974)^{k-1} = f_{\max} \left(1 - 0{,}9405^n \right).$$

Wegen

$$\frac{\Delta f}{f_{\max}} = \frac{f_{\max} - f_{\min}}{f_{\max}} = \frac{3}{5} \quad \text{ist} \quad \frac{f_{\min}}{f_{\max}} = \frac{2}{5}.$$

Bei einem realen Stimmumfang von 3 kHz werden demnach

$$n = \ln \frac{f_{\min} / f_{\max}}{\ln 0{,}9405} = \frac{\ln 0{,}4}{\ln 0{,}9405} \approx 15$$

Neuronen benötigt. Damit können $2^{15} \approx 32768$ Klang- bzw. Geräuschkombinationen erkannt werden. Das Frequenzspektrum läßt sich anhand einer Fourier-Transformation mit einem Spektrumanalysator ermitteln. Das „Time Delay Neural Network“ dient der Sprach- wie der Gesichtserkennung, hat seine Bedeutung aber mittlerweile an die Hidden Markov Models abgeben müssen. Erst mittels Deep Learning sind im Bereich der neuronalen Netzwerke für Sprach- und Texterkennung wieder neuere Ansätze in den Fokus gerückt. Dennoch steckt das neuronale Verfahren im großen ganzen noch in den Kinderschuhen.

