

Bachelorarbeit

Titel der Bachelorarbeit: Image Analysis for Food Safety and Health

Name des Autors: TEMKENG Thibaut

Datum der Abgabe: Ende Oktober

Betreuung: Name der Betreuer/ des Betreuers: Shou Liu

Fakultät für Embedded Intelligence for Health Care and Wellbeing

Inhaltsverzeichnis

1	Einleitung	4
2	Grundlagen	5
2.1	Neuron	5
2.2	Merkmalskarten(Feature-Maps)	5
2.3	Filters	6
2.4	Entwicklung von Künstlichen Neuronalen Netzen	6
3	Convolutional Neural Network	6
3.1	Feedforward	6
3.1.1	Input Layer	6
3.1.2	Faltungsschicht	6
3.1.3	Aktivierungsfunktion	8
3.1.4	Pooling Layer	11
3.1.5	Multi-layer Perzeptron (Fully Connected Layer)	12
3.2	Backforward	13
3.2.1	Fehlerfunktion	13
3.2.2	Gradient	14
3.2.3	Lernrate	14
3.2.4	Gradientenabstiegsverfahren	14
4	Kompression von Tiefe neuronale Netze (DNN)	17
4.1	Pruning Network	17
4.2	Quantisierung von neuronalen Netzwerken	21
4.3	Huffman Codierung	25
5	Experiment	25
5.1	Lernrate	25
5.2	Besondere CNNs	25
5.2.1	AlexNet	25
5.2.2	SqueezeNet	26
5.2.3	Xception: Extreme Inception	28
5.2.4	MobileNet	29
5.3	Algorithmen zur Optimierung des Gradientenabstiegsverfahren: Optimizer	29
5.3.1	Adaptive Gradient Algorithm (AdaGrad)	29
5.3.2	Root Mean Square Propagation(RMSProp)	30
5.3.3	Adaptive Moment Estimation(Adam)	31
5.4	Problem beim Training von Convolutional neuronale Netzwerke	31
5.4.1	Overfitting	31
5.4.2	Dataset	36
6	Abkürzungsverzeichnis	36

Abbildungsverzeichnis

1	Funktionsweise eines künstlichen Neurons	5
3	Faltungsoperation mit einem 3×3 -Filter und Schrittgröße = 3	9
4	Binäre Treppenfunktion	10
6	Lineare Funktion	10
8	Logistische Aktivierungsfunktion: <i>sigmoid(x)</i>	11
10	Tangens Hyperbolicus.	11
12	ReLU Aktivierungsfunktion	12
14	Leaky ReLU Funktion	12
16	Funktionsweise der Pooling-Sicht mit <i>Pooling_size</i> = (2, 2) und <i>Stride</i> = 2	13
17	Darstellung eines neuronalen Netzes	13
18	Ablauf der Backpropagation	16
19	Netzbeschneidung während des Trainings	19
20	Einfluss der Intensität des Filters auf Feature-Map	20
21	Das Beschneiden eines Filters führt zum Entfernen der entsprechenden Feature-Map und der zugehörigen Kernel in der nächsten Ebene[6]. . . .	21
22	Parameter Approximieren durch Matrixfaktorisierung	22
23	AlexNet Architektur zum Bild	26
24	SqueezeNet Architekturen [19]	27
25	fire_module [19]	28
26	Xception Architektur	29
27	MobileNet Architektur	30
28	Anwendung von <i>ImageDataAugmentation</i>	32
29	Neuronales Netz mit Dropout ausgestattet [3].	33
30	Erhöhung des Trainingsdaten durch Dropout	34

1 Einleitung

2 Grundlagen

2.1 Neuron

Ein künstliches Neuron[17] ist eine mathematische Funktion, die das biologische Neuron nachbildet. Künstliche Neuronen sind elementare Einheiten in einem Künstliches neuronales Netz (KNN). Das künstliche Neuron empfängt einen oder mehrere Inputs und bildet sie auf einen Output ab. Normalerweise wird jeder Eingabe x_i separat mit einem Gewicht w_i multipliziert und danach aufsummiert und zum Schluss wird die Summe durch eine Funktion geleitet, die als Aktivierungs- oder Übertragungsfunktion bekannt ist. Eine schematische Darstellung eines künstlichen Neurons ist in Abbildung 1 zu sehen.

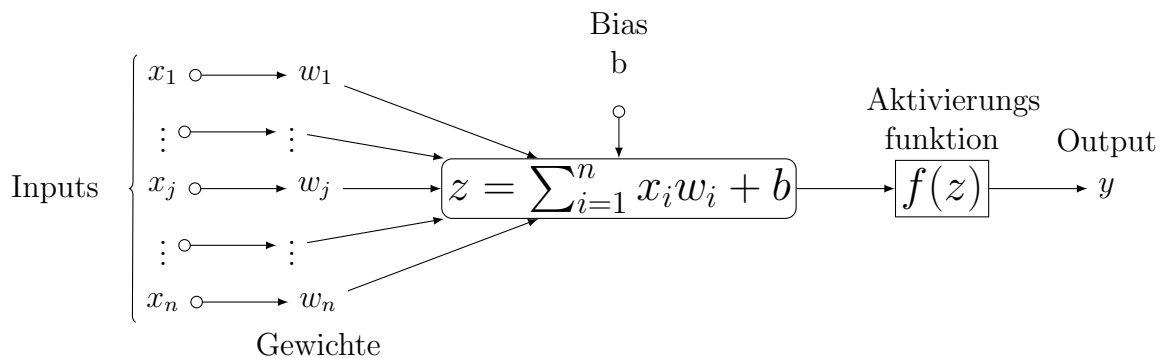


Abbildung 1: Funktionsweise eines künstlichen Neurons

Künstliche Neurone können aufgestapelt werden, um eine Schicht (*Layer*) zu bilden. Ein KNN besteht aus einer oder mehreren Schichten und je nach seiner Position in einem neuronalen Netz (NN) wird eine Schicht anders genannt: Eingangsschicht (*Input Layer*) bzw. Ausgabeschicht (*Output Layer*), wenn das Layer die Eingangsdaten bzw. Ausgabedaten des neuronalen Netzes darstellt und versteckte Schicht (*Hidden Layer*), wenn es keine Eingang- oder Ausgabeschicht ist. Ein kurzer Überblick über die Darstellung von KNNs kann sich in Abbildung 17 verschafft werden.

2.2 Merkmalskarten (Feature-Maps)

Die Ausgabe einer Schicht wird als Aktivierungskarten oder Feature-Map(s) bezeichnet. Die Anzahl von Feature-Maps in einer Schicht ist gleich der Anzahl der Ausgabekanäle (*output channels*) bzw. Tiefe (*depth*) dieser Schicht. In einer Faltungsschicht ist ein Feature-Map immer zweidimensional, während die Dimension eines Feature-Maps in einem Fully Connected Layer (FCL) nur von der des Inputs abhängt, genauer gesagt, für ein n -dimensionales Inputs ist ein Feature-Map $(n - 1)$ -dimensional. Wie die Feature-Maps berechnet werden, hängt sehr vom Schichttyp ab. Es wird beispielsweise in einer Faltungsschicht die Faltungsoperation mehrmals auf die Eingabe angewendet und jedes Mal wird ein neues Feature-Map erhalten.

2.3 Filters

Filter ermöglicht die Extraktion von Features

-

2.4 Entwicklung von Künstlichen Neuronalen Netzen

3 Convolutional Neural Network

3.1 Feedforward

3.1.1 Input Layer

Die Eingangsschicht stellt die Eingangsdaten dar. Hier müssen die Eingangsdaten dreidimensional sein. Also die Eingangsdaten von Convolutional Neural Network (CNN) haben immer die folgende Form $W \times H \times D$ wobei (W, H) der räumlichen Dimension und D die Tiefe der Daten entspricht. Z.B $100 \times 100 \times 3$ für ein RGB-Bild und $224 \times 224 \times 1$ für ein Graustufenbild.

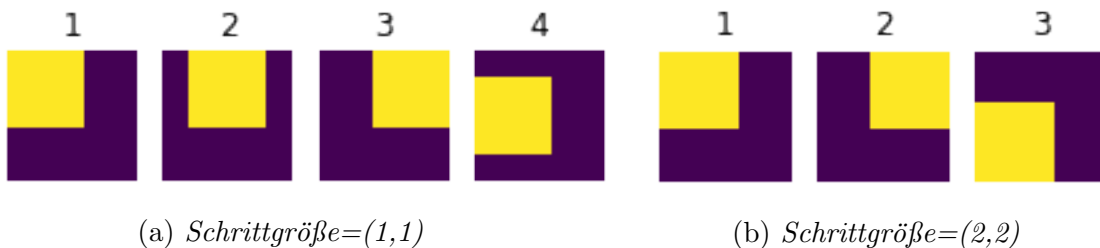
3.1.2 Faltungsschicht

Die wichtigste Sicht bzw: der Hauptschicht in einem CNN ist die Faltungsschicht (*Convolution Layer*). Die Eingabedaten eines CNN besteht auf jedenfall aus wichtigen und unwichtigen Informationen. Als wichtige und unwichtige Informationen haben wir z.B die starke Präsenz der weißen Farbe in einer Muschelsuppe bzw. die Präsenz eines Menschen, wenn man verschiedenen Ernährungsklasse klassifizieren möchte. Während des Trainings eines CNN wird versucht, diese relevanten Informationen aus den Daten zu entnehmen und die irrelevanten auszuschließen. Alles was ein CNN aus den Eingabedaten nutzt, um die Daten zu bestimmter Klassen zuzuordnen, wird als Feature bezeichnet. Das einzige bzw. das Hauptziel einer Faltungsschicht besteht darin, die Features aus seinen Inputdaten herauszuziehen. Dass ein Faltungsschicht in der Lage ist, Features selbst zu extrahieren, ohne dass man sie hinweist, was wichtig ist und was nicht, ist wirklich beeindruckend, Aber die Art und Weise, wie sie es tut, ist noch beeindruckender.

In einem Convolutional Layer (ConvL) wird die sogenannte Faltungsoperation (*convolution operation*) durchgeführt, dabei wird das komponentenweises Produkt (*Hadamard-Product*) zwischen einem kleinen Bereich der Eingabedaten und einem Kernel durchgeführt und dann die ganze aufsummiert. Eine Illustration der Faltungsoperation ist in der Abbildung 3a zu sehen. Die Resultante der Faltungsoperation wird die Aktivierung des Neurons genannt. Sollte diese Aktivierung des Neurons null sein, dann sagt man, dass das Neuron "nicht aktiv" ist und sonst ist es *aktiv*. Die null Aktivierung bedeutet, dass das vom Filter gesuchte Feature nicht gefunden wurde. Je stärker oder größer die Aktivierung eines Neurons ist, desto höher ist die Wahrscheinlichkeit, dass das vom Filter gesuchte Feature gefunden ist (siehe Abbildung 3). um diese Faltungsoperation durchzuführen, muss einigen Hyperparameter vordefiniert sein, und zwar:

- **Anzahl und Größe von Filtern**

Die Anzahl an Filtern gibt nicht nur an, wie viele Filter in dem ConvL verwendet werden, sondern auch, wie oft die Faltungsoperation durchgeführt wird, d.h. die Anzahl der Feature-Maps oder die Tiefe der Schicht. Anstatt das ganze Bild zu betrachten, wenn man auf der Suche nach einem Feature ist, was mit der Tatsache gleichbedeutend ist, dass jedes Neuron der Schicht mit allen Neuronen der vorherigen verbunden ist, wird nur einen lokalen Bereich des Bildes betrachtet, wir verbinden also jedes Neuron nur mit einem lokalen Bereich der Bild. Dieser lokaler Bereich wird als Empfangsfeld(*receptive field*) des Neurons bezeichnet und entspricht der Filtergröße. Es ist zu beachten, dass die Filtergröße nur der räumlichen Dimension des Filters entspricht. Die Tiefe des Filters ist gleich die vom Input, also für ein $(100, 100, 3)$ Bild haben alle Filter die Tiefe 3. Die Verwendung von solchen kleinen Filtern ist die Hauptidee hinter einer Faltungsschicht. Filter haben im Allgemeinen eine kleine räumliche Dimension wie z.B. 2×2 , 3×3 oder 5×5 , sonst verliert man einen großen Vorteil von ConvL, der darin besteht, die Speicheranforderung deutlich zu reduzieren, indem es die Gewichte verteilt.



- **Schrittgröße(Stride)**

Da ein Filter nur einen kleinen Bereich des Bilds wahrnehmen kann, wird eine Schrittgröße verwendet, um die Bewegung des Filters auf dem Bild zu steuern. Das Filter wird über das Bild von links nach rechts, von oben nach unten bewegt. Sei $S := (n, m)$ die Schrittgröße, dann wird das Filter von n Pixeln nach rechts für jede horizontale Bewegung des Filters und m Pixeln nach unten für jede vertikale Bewegung des Filters bewegt (siehe Abbildung 2a und 2b). Wie die folgende Tabelle zeigt, hängt die räumliche Dimension des Outputs sehr von der Schrittgröße ab.

- **Padding**

Mit einem großen Filter lernt man in der Regel "mehr" als mit einem kleinen. Aber wie in Tabelle 1 zu sehen, habe man eine Reduktion der Dimension, wenn ein Filter mit Filtergröße > 1 angewendet wird und um die Dimension zu behalten, wird vor der Anwendung der Filter auf das Bild eine Füllung(*padding*) an den Rändern des Bilds gemacht. Diese Füllung ermöglicht erstens den Entwurf immer tiefer Netzwerke, denn es gibt nach jedem ConvL einen kleinen Dimensionsverlust und zweitens, dass die Information an Rändern nicht zu schnell verschwinden werden.

Imagegröße	Stride	Filtergröße	Output (räumliche Dim.)	Mit Padding
(100, 100, 3)	(1, 1)	(1, 1)	(100, 100)	(100, 100)
(100, 100, 3)	(1, 1)	(3, 3)	(98, 98)	(100, 100)
(100, 100, 3)	(2, 2)	(3, 3)	(49, 49)	(50, 50)
(100, 100, 3)	(1, 1)	(4, 4)	(97, 97)	(100, 100)
(100, 100, 3)	(1, 1)	(10, 10)	(91, 91)	(100, 100)

Tabelle 1: Auswirkung von Schrittgröße, Filtergröße und Padding auf Output eines (100, 100, 3) Bild.

In einem CNN mit mehreren ConvLs kümmern sich die ersten ConvLs um das Erlernen von einfachen Merkmale wie Winkel, Kanten oder Linien und je tiefer das CNN ist, desto komplexer sind die extrahierten Merkmale, also die früher entdeckten Features werden kombiniert, um komplexe Features wie .Eine Vorverarbeitung der Netzeingangsdaten zur Extraktion der relevanten Information nicht mehr nötig.

3.1.3 Aktivierungsfunktion

Das neuronale Netzwerk wird während dem Training mit sehr vielen Daten gespeist und das sollte in der Lage sein, aus diesen Daten zwischen relevanten und irrelevanten Informationen Unterschied zu machen. Die Aktivierungsfunktion auch Transferfunktion oder Aktivitätsfunktion genannt, hilft dem NN bei der Durchführung dieser Trennung. Es gibt sehr viele Aktivierungsfunktionen und in folgenden werden wir sehen, dass eine Aktivierungsfunktion je nach zu lösende Aufgaben vorzuziehen ist.

$$\begin{cases} Y = f(\Sigma(\text{Gewicht} * \text{Input} + \text{Bias})) \\ f := \text{Aktivierungsfunktion} \end{cases}$$

Binäre Treppenfunktion ist extrem einfach, siehe Abbildung 4, definiert als $f(x) = \begin{cases} 1, & \text{if } x \geq a \text{ (a:= Schwellenwert)} \\ 0, & \text{sonst} \end{cases}$. Sie ist für binäre Probleme geeignet, also Probleme wo man mit *ja* oder *nein* antworten sollte. Sie kann leider nicht mehr angewendet werden, wenn es mehr als zwei Klassen klassifiziert werden soll oder wenn das Optimierungsverfahren gradientenbasierend ist, denn Gradient immer null.

Lineare Funktion ist definiert als $f(x) = ax$, $f'(x) = a$, siehe Abbildung 6. Sie ist monoton, null zentriert und differenzierbar. Es ist jetzt möglich, nicht mehr nur binäre Probleme zu lösen und mit gradientenbasierenden Optimierungsverfahren während der Backpropagation Parameter anzupassen, denn Gradient nicht mehr null, also sie ist besser als binäre Funktion. Nutzt ein mehrschichtiges Netz die lineare Aktivierungsfunktion, so kann es auf ein einschichtiges Netz überführt werden und mit einem einschichtigen Netz können komplexe Probleme nicht gelöst werden. Außerdem ist der Gradient konstant. Der Netzfehler wird also nach einigen Epochen nicht mehr minimiert und das Netz wird immer das Gleiche vorhersagen.

(a) Matrixdarstellung

0	0	3	3	3	0	0	0	0
0	2	0	3	2	0	0	2	0
2	0	2	2	0	2	2	0	2
1	1	1	1	1	1	1	1	1
1	0	1	0	2	0	1	0	1
0	0	1	2	0	2	0	0	1
1	1	1	1	1	1	1	3	1
3	0	3	1	2	0	3	0	3
0	3	0	2	0	2	0	0	0

(9,9)Bild

\times

0	0	0
0	1	0
1	0	1

 $=$

6	6	6
1	6	1
0	6	0

(3,3)Feature-Map

Filter

(b) Pixeldarstellung

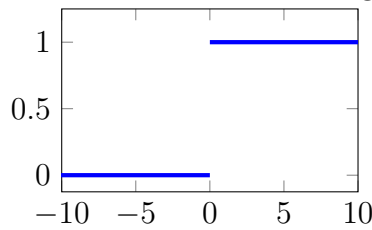


Abbildung 3: Faltungsoperation mit einem 3×3 -Filter und Schrittgröße = 3

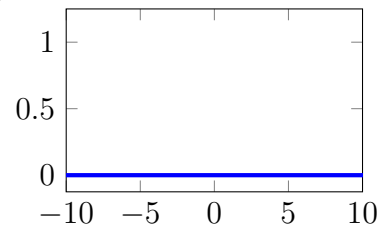
Logistische Funktion ist definiert als $f(x) = \frac{1}{1+\exp(-x)}$, $f'(x) = \frac{\exp(x)}{(1+\exp(x))^2}$, siehe Abbildung 8. Sie ist differenzierbar, monoton, nicht linear und nicht null zentriert (hier nur positive Werte). Zwischen $[-3, +3]$ ist der Gradient sehr hoch. Kleine Änderung in der Netzininput führt also zu einer großen Änderung der Netzausgabe. Diese Eigenschaft ist bei Klassifikationsproblemen sehr erwünscht. Die Ableitung ist glatt und von Netzininput abhängig. Parameter werden während der Backpropagation je nach Netzininput angepasst. Außerhalb von $[-3, 3]$ ist der Gradient fast gleich null, daher ist dort eine Verbesserung der Netzleistung fast nicht mehr möglich. Dieses Problem wird Verschwinden des Gradienten (*vanishing gradient problem*) genannt. Außerdem konvergiert das Optimierungsverfahren sehr langsam und ist wegen der exponentiellen (e^x) Berechnung rechenintensiv.

Tangens Hyperbolicus ist definiert als $\tanh := 2\text{sigmoid}(x) - 1$, siehe Abbildung 10. Außer dass sie null zentriert ist, hat sie die gleichen Vor- und Nachteile wie die Sigmoid

Abbildung 4: Binäre Treppenfunktion

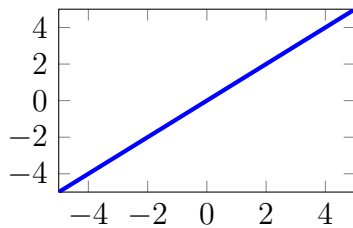
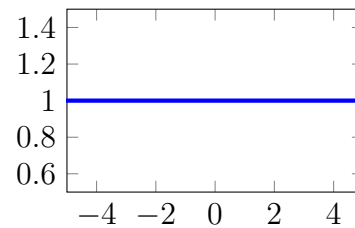


(a) Binäre Treppenfunktion



(b) Ableitung Binäre Treppenfunktion

Abbildung 6: Lineare Funktion


(a) Lineare Funktion: $f(x) = x$

(b) Ableitung Lineare Funktion: $f'(x) = 1$

Funktion. **Sättigung fehlt noch**

Rectified Linear Unit (ReLU) ist definiert als $f(x) = \max(x, 0)$, siehe Abbildung 12. Sie ist sehr leicht zu berechnen. Es gibt keine Sättigung wie bei *Sigmoid* und *tanh*. Sie ist nicht linear, deshalb kann der Fehler schneller propagiert werden. Ein größter Vorteil der ReLU-Funktion ist, dass nicht alle Neurone gleichzeitig aktiviert sind, negative Eingangswerte werden zu null, daher hat die Ausgabe von Neuronen mit negativen Eingangswerten keinen Einfluss auf die Schichtausgabe, diese Neurone sind einfach nicht aktiv. Das Netz wird also spärlich und effizienter und wir haben eine Verbesserung der Rechenleistung. Es gibt keine Parameteranpassungen, wenn die Eingangswerte negativ sind, denn der Gradient ist dort null. Je nachdem wie die Bias initialisiert sind, werden mehrere Neurone getötet, also nie aktiviert und ReLU ist leider nicht null zentriert. **ReLU haben die wünschenswerte Eigenschaft, dass sie keine Eingangsnormalisierung benötigen, um eine Sättigung zu verhindern.**

Leaky ReLU Funktion ist definiert als $f(x) = \begin{cases} x, & \text{if } x > 0 \\ 0.01x, & \text{sonst} \end{cases}$, siehe Abbildung

14. Sie funktioniert genauso wie die ReLU-Funktion, außer dass sie das Problem der toten Neurone löst und sie null zentriert. Es gibt somit immer eine Verbesserung der Netzleistung, solange das Netz trainiert wird. Wenn das Problem von Leaky ReLU nicht gut gelöst wird, wird empfohlen, die *Parametric ReLU* (PReLU) Aktivierungsfunktion zu verwenden, die während des Trainings selbst lernt, das Problem der toten Neurone zu lösen.

Softmax ist definiert als $f(x_1, x_2, \dots, x_n) = \frac{(e^{x_1}, e^{x_2}, \dots, e^{x_n})}{\sum_{i=1}^n e^{x_i}}$. Die Softmax-Funktion würde die Ausgänge für jede Klasse zwischen null und eins ausdrücken und auch durch die Summe der Ausgänge teilen. Dies gibt im Wesentlichen die Wahrscheinlichkeit

Abbildung 8: Logistische Aktivierungsfunktion: $\text{sigmoid}(x)$.

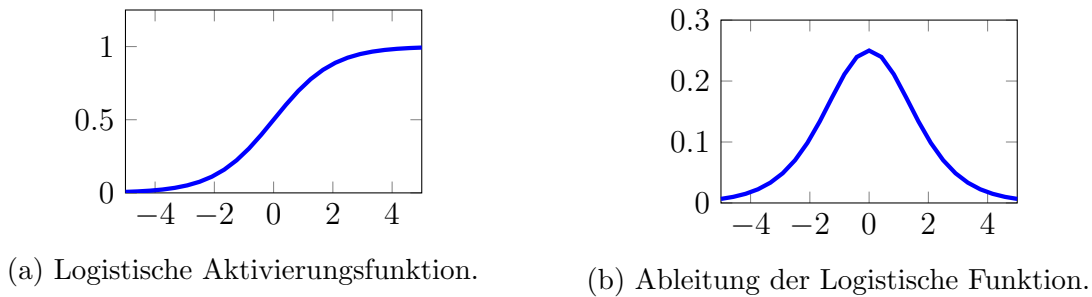
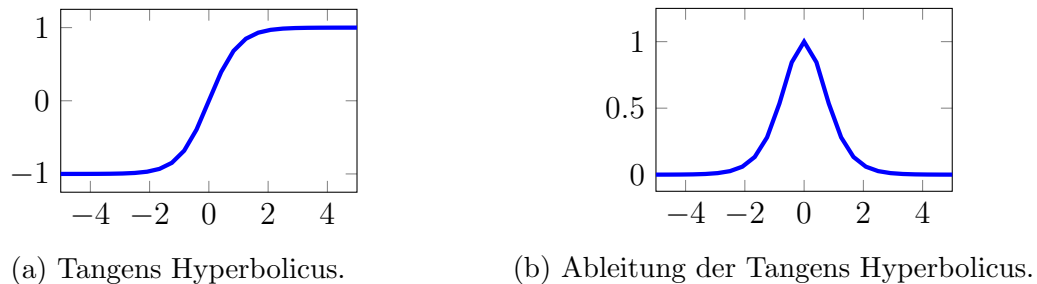


Abbildung 10: Tangens Hyperbolicus.



an, dass sich der Input in einer bestimmten Klasse befindet.

In allgemein wird die ReLU aufgrund des Problems der toten Neurone nur in versteckte Schichten und die Softmax-Funktion bei Klassifikationsproblemen und Sigmoid-Funktion bei Regressionsproblemen in Ausgabeschicht verwenden.

3.1.4 Pooling Layer

Die Funktionsweise von Pooling-Schichten ist sehr ähnlich zu der von ConvLs. Das Filter wird über die Inputdaten bewegen und dabei anstatt die Faltungsoperation durchzuführen, werden die Inputdaten Blockweise zusammengefasst. Ein Pooling-Layer besitzt nur ein Filter, das nicht wie in ConvLs lernbar ist, sondern gibt nur an, wie groß der Block, der zusammengefasst wird, sein muss. Als Standard werden ein 2×2 Filter und eine 2×2 Schrittgröße verwendet, was die Dimension der Inputdaten um die Hälfte reduziert. Interessanter dabei ist, dass die wichtigen Informationen oder Muster nach der Pooling-Layer vorhanden bleiben und damit haben wir nicht nur eine Erhöhung der Rechengeschwindigkeit, sondern auch eine wesentliche Reduzierung der Netzparameter, was die Wahrscheinlichkeit einer Modelüberanpassung (*Overfitting*) reduziert. Pooling-Schichten sind etwa invariant gegenüber kleiner Veränderung wie Parallelverschiebung[4].

Je nachdem wie die Blöcke in Pooling Layer (Pool) zusammengefasst werden, haben die Pools unterschiedliche Namen. Werden die Werte eines Blockes durch den Maximalwert des Blocks, dann sprechen wir von Max-Pooling-Layer (siehe Abbildung 16 links), wenn sie durch den Mittelwert des Blocks ersetzt, wird von Average-Pooling-Layer (siehe Abbildung 16 rechts) und wenn die Filtergröße gleich die räumliche Dimension der Ein-

Abbildung 12: ReLU Aktivierungsfunktion

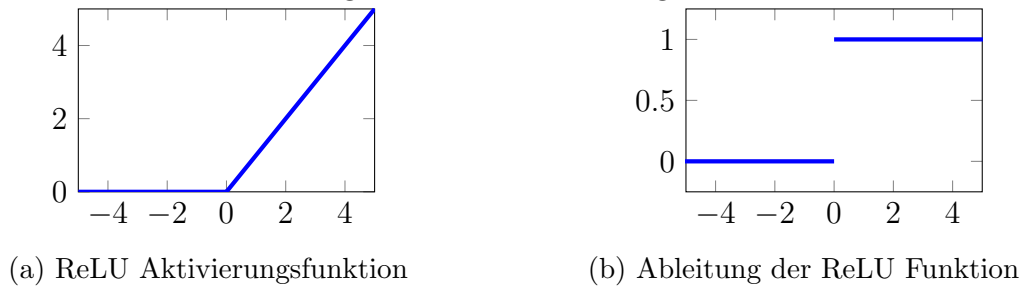
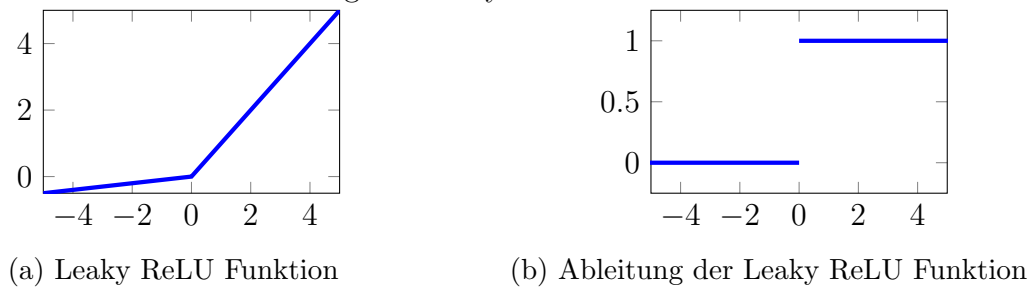


Abbildung 14: Leaky ReLU Funktion



gangdaten ist, sprechen wir von Global-Max-Pooling-Layer und Global-Average-Pooling-Layer, es wird also alle Neurone in einem Kanal zu einem Neuron, die Ausgabedimension solche Schicht entspricht der Anzahl der Kanäle bzw. Tiefe der Inputdaten.

Das Global-Pooling-Layer wird sehr oft angewendet, um das Vorhandensein von Merkmale in Daten aggressiv zusammenzufassen. Es wird auch manchmal in Modellen als Alternative zur Flatten-Schicht, die mehrdimensionale Daten zu eindimensionale umwandelt, beim Übergang von ConvLs zu einem FCL verwendet.

3.1.5 Multi-layer Perzeptron (Fully Connected Layer)

Nachdem die relevanten lokalen Merkmale durch die Wiederholung von Conv und Pooling-Schichten extrahiert werden sind, wenden sie in einem FCL kombiniert, um das Ergebnis jeder Klasse zu berechnen. Die FCLs des CNN ermöglichen, Informationssignale zwischen jeder Eingangsdimension und jeder Ausgangsklasse zu mischen, so dass die Entscheidung auf dem gesamten Bild basieren kann und ihm eine Klasse zugewiesen werden kann. Die FCLs funktionieren eigentlich genau wie ConvLs, außer dass jedes Neuron in FCL mit allen Neuronen und nicht mit einem kleinen Bereich von Neuronen im vorherigen Layer verbunden ist. Ein NN mit nur FCLs sieht wie in Abbildung 17 aus.

Aufgrund der hohen Anzahl von Verbindungen zwischen Neuronen in einem FCL wird viel Speicher und Rechenleistung benötigt und verlangsamt auch das Training, es ist auch einer der Gründe, weshalb die FCLs meist nur in der letzten Schicht von CNN zur Klassifizierung verwendet werden und die Anzahl der Neuronen in letzter Schicht entspricht der Anzahl von Klassen.

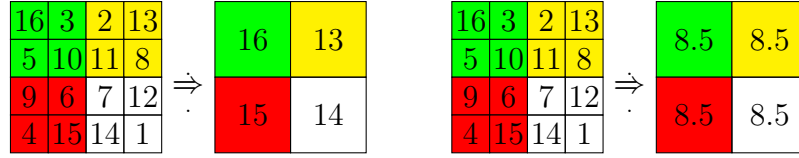


Abbildung 16: Funktionsweise der Pooling-Sicht mit Pooling_size= (2, 2) und Stride = 2

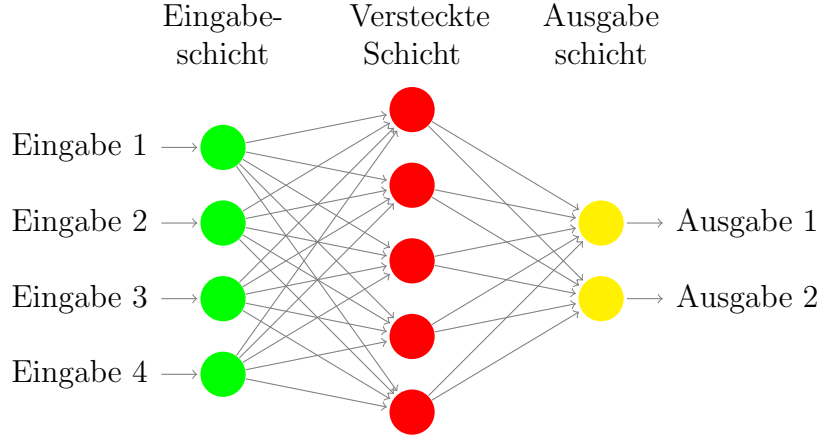


Abbildung 17: Darstellung eines neuronalen Netzes

3.2 Backforward

3.2.1 Fehlerfunktion

Das Training von CNNs besteht darin, den vom CNN begangenen Fehler zu korrigieren bzw. zu minimieren, daher wird es sehr oft als ein Optimierungsverfahren betrachtet. Wie gut die Vorhersage des neuronalen Netzes gerade ist, wird durch eine Fehlerfunktion auch Kostenfunktion genannt quantifiziert oder angegeben. Die Fehlerfunktion bringt die Ausgabewerte des NNs mit den gewünschten Werten in Zusammenhang. Sie ist ein nicht-negativer Wert und je kleiner dieser Wert wird, desto besser ist die Übereinstimmung des CNNs. Es wird im Laufe des Trainings von CNNs versucht, diese Kostenfunktion mit Gradientenbasierten Verfahren zu minimieren (siehe Absatz ??)

Die meisten benutzten Kostenfunktionen sind die Kreuzentropie (*cross-entropy*, Gleichung 3.2) (CE) und die mittlere quadratische Fehler (*mean squared error*, Gleichung 3.1) (MSE).

$$MSE(Y, \hat{Y}) = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (3.1)$$

$$CE(Y, \hat{Y}) = - \frac{1}{n} \sum_{i=1}^n (Y_i \log(\hat{Y}_i) + (1 - Y_i) \log(1 - \hat{Y}_i)) \quad (3.2)$$

$Y := \{Y_1, \dots, Y_n\}$:Die tatsächlichen Werte

$\hat{Y} := \{\hat{Y}_1, \dots, \hat{Y}_n\}$:Die Ausgabewerte des neuronalen Netzes

Im Gegenteil zu CE Fehlerfunktionen, die sich nur auf Wahrscheinlichkeitsverteilungen anwenden lassen, können die MSE auf beliebige Werte angewendet werden. Nach [9, Pavel et al] ermöglicht die CE-Verlustfunktion ein besseres Finden lokaler Optima als die MSE-Verlustfunktion und das soll daran liegen, dass das Training des MSE Systems schneller in einem schlechten lokalen Optimum stecken bleibt, in dem der Gradient fast null ist und damit keine weitere Reduzierung der Netzfehler ermöglicht. Im Allgemeinen ist die CE Kostenfunktion für die Klassifikationsprobleme und die MSE Fehlerfunktion für die lineare Regression-Probleme besser.

3.2.2 Gradient

Der Gradient einer Funktion ist die erste Ableitung einer Funktion und in mehrdimensionalem Raum ist der Gradient einer Funktion der Vektor, dessen Einträge die ersten partiellen Ableitungen der Funktion sind. Der Gradient an einem Punkt gibt die Richtung der steilsten Anstieg der Funktion an diesem Punkt. Also da wir die Kostenfunktion minimieren möchten, sollen wir lieber immer in die Gegenrichtung des Gradienten gehen. In auf dem Gradienten basierten Optimierungsverfahren wird der Gradient benutzt, um die lokalen oder globalen Extremwerte (hier das Minimum) zu erreichen. Da wir jetzt die Richtung des Minimums herausgefunden haben, bleibt noch zu bestimmen, wie wir in diese Richtung gehen sollen.

3.2.3 Lernrate

Die Lernrate oder Schrittweite beim maschinellen Lernen ist ein Hyperparameter, der bestimmt, inwieweit neue gewonnene Informationen alte Informationen überschreiben sollen [12], in anderen Worten wie schnell wir ans Ziel kommen. Je nachdem wie die Lernrate gesetzt wird, werden bestimmte Verhalten beobachtet (Siehe Absatz 5.1) und sie nimmt sehr oft Werte zwischen 0.0001 und 0.5: Die Lernrate muss allerdings im Intervall $]0, 1[$ Werte annehmen, sonst ist das Verhalten des NN nicht vorhersehbar bzw. konvergiert das Verfahren einfach nicht. Für jeden Punkt x aus dem Parameterraum gibt es eine optimale Lernrate $\eta_{opt}(x)$, sodass das globale oder lokale Minimum sofort nach der Parameteranpassung erreicht wird. Da $\eta_{opt}(x)$ am Trainingsanfang leider nicht bekannt ist, wird die Lernrate in die Praxis vom Programmierer basiert auf seine Kenntnisse mit NNs oder einfach zufällig gesetzt.

3.2.4 Gradientenabstiegsverfahren

Aktuelle leistungsfähige DNN bestehen fast immer aus Millionen Variablen (lernbarer Parameter). Wir können uns ein DNN als eine Gleichung mit Millionen von Variablen (lernbare Parameter) vorstellen, die wir lösen möchten. Mit Hilfe der Daten wollen wir uns in einem

Raum, dessen Dimension größer als eine Million ist, bewegen, um die optimalen Parameter (Parameter, die die Daten korrekt abbilden) zu finden. Aufgrund der unendlichen Anzahl von Punkten im solchen Räume wäre es nicht sinnvoll, einen Punkt zufällig auszuwählen, dann überprüfen, ob er optimal ist und, wenn nicht, nochmals einen anderen Punkt zufällig auszuwählen. Genauer zu diesem Zeitpunkt kommen Gradientenabstiegsverfahren zum Einsatz. Die Gradientenabstiegsverfahren sind Verfahren, die auf dem Gradient basieren, um Optimierungsprobleme zu lösen. Hier wird Gradientenabstiegsverfahren verwendet, die optimalen Parametern zu finden oder anzunähern.

Ablauf eines Gradientenverfahrens im DNN

Das Gradientenabstiegsverfahren kann in drei Hauptschritte aufgeteilt werden. Die Abbildung 18 stellt das Backpropagation-Verfahren bildlich dar. Beim ersten Schritt wird ein zufälliger Punkt aus dem Parameterraum ausgewählt und davon ausgehend wird der Parameterraum exploriert. Das entspricht der Netzparameterinitialisierung am Trainingsanfang. Der zweite Schritt besteht darin, die Abstiegsrichtung zu bestimmen. Dazu werden zuerst die Eingangsdaten in das NN eingespeist (*Forwardpropagation*), danach wird der Fehler zwischen den Netzvorhersagen und den korrekten Werten berechnet und ein Fehler gibt es (fast) immer, denn die Initialisierung wird zufällig gemacht und die Wahrscheinlichkeit, dass wir von Anfang an die optimalen Werte finden, ist verschwindend klein und zuletzt wird der Gradient (3.2.2) der Kostenfunktion in abhängig von den gegebenen Eingangsdaten und den erwarteten Werten berechnet. Beim letzten Schritt wird die Schrittweite bestimmt. Die Lernrate (3.2.3) oder die Schrittweite wird vor Trainingsbeginn festgelegt oder während des Trainings abhängig von aktuellem Netzzustand allmählich adaptiert.

Pfeile in Abbildung 18 weisen nur den Prozessablauf hin.

Variante des Gradientenverfahrens

Bisher existiert drei Variante des Gradientenabstiegsverfahren, die sich nur durch die Größe der Daten, die sie verwendet, um den Gradienten der Kostenfunktion berechnet, unterscheidet. Zum Aktualisierung der Netzwerkparameter nutzen sie jeweils die Gleichung (3.3).

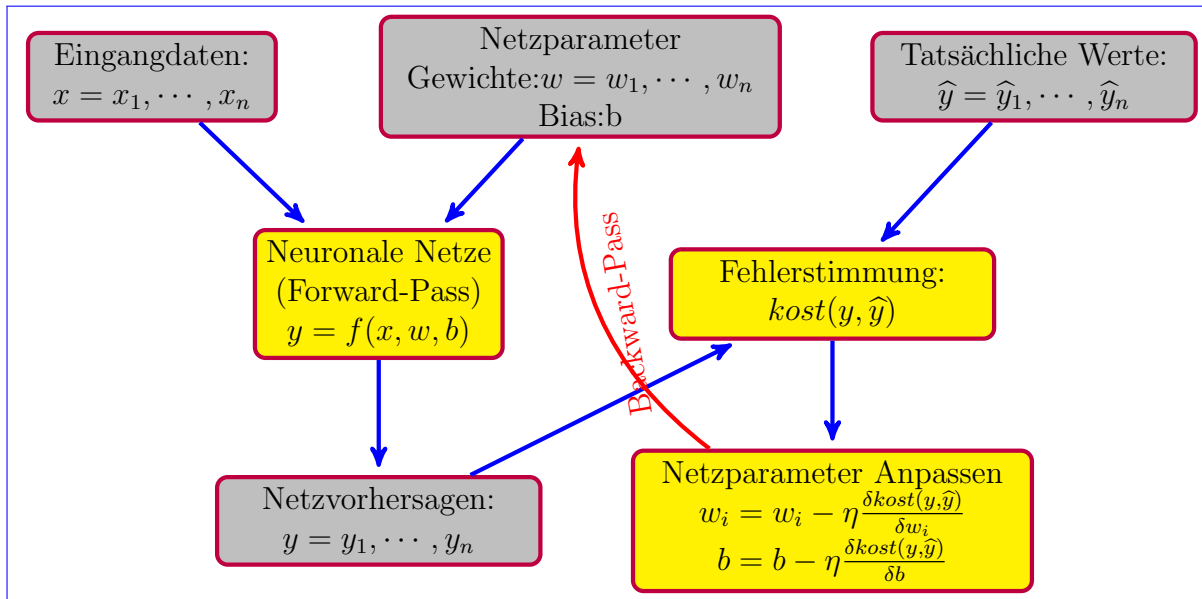
$$\theta_{t+1} = \theta_t - \eta g_t, \quad g_t = \frac{\delta E}{\delta \theta_t} \quad (3.3)$$

η : Lernrate

E : Die Fehlerfunktion

θ_t : Netzwerkparameter zum Zeitpunkt t

Stochastic Gradient Descent (SGD): Bei SGD wird jeweils ein Element bzw. Sample aus der Trainingsmenge durch das NN durchlaufen und den jeweiligen Gradienten berechnen, um die Netzwerkparameter zu aktualisieren. Diese Methode wird sehr oft online Training genannt, denn jedes Sample aktualisiert das Netzwerk. SGD verwendet geringer Speicherplatz und die Iterationen sind schnell durchführbar. Zusätzlich kann die



$f(x, w, b)$: Netzfunktion. $kost(y, \hat{y})$: Kostenfunktion.
 η : Lernrate $\frac{\delta kost(y, \hat{y})}{\delta w_i}$: Ableitung der Kostenfunktion abhängig von w_i .

Abbildung 18: Ablauf der Backpropagation

Konvergenz für großen Datensatz wegen der ständigen Aktualisierung der Netzwerkparameter beschleunigen. Diese ständigen Aktualisierung hat die Schwankung der Schritte in Richtung der Minima zur Folge, was die Anzahl der Iteration bis zum Erreichen des Minimums deutlich ansteigt und dabei helfen kann, aus einem unerwünschten lokalen Minimum zu entkommen. Ein großer Nachteil dieses Verfahren ist der Verlust der parallelen Ausführung, es kann jeweils nur ein Sample ins NN eingespeist werden.

Batch Gradient Descent (BGD): BGD funktioniert genauso wie SGD, außer dass der ganze Datensatz statt jeweils ein Element aus dem Datensatz zur Netzwerkparameteraktualisierung genutzt wird. Jetzt kann das Verfahren einfach parallel ausgeführt werden, was den Verarbeitungsprozess des Datensatzes stark beschleunigt. BGD weist weniger Schwankungen in Richtung des Minimums der Kostenfunktion als SGD auf, was das Gradientenabstiegverfahren stabiler macht. Außerdem ist das BGD recheneffizienter als das SGD, denn nicht alle Ressourcen werden für die Verarbeitung eines Samples, sondern für den ganzen Datensatz verwendet. BGD ist leider sehr langsam, denn die Verarbeitung des ganzen Datensatz kann lange dauern und es ist nicht immer anwendbar, denn sehr große Datensätze lassen sich nicht im Speicher einspeichern.

Mini-batch Stochastic Gradient Descent (MSGD): MSGD ist eine Mischung aus SGD und BGD. Dabei wird der Datensatz in kleine Mengen (*Mini-Batch* oder *Batch*)

möglicherweise gleicher Größe aufgeteilt. Je nachdem wie man die Batch-Größe setzt, erhalten wir SGD oder BGD wieder. Das Training wird Batchweise durchgeführt, d.h. es wird jeweils ein Batch durch das NN propagiert, der Verlust jedes Sample im Batch wird berechnet und dann deren Durchschnitt benutzt, um die Netzwerkparameter zu anzupassen. MSGD verwendet den Speicherplatz effizienter und kann von Parallelen Ausführung profitieren. Noch dazu konvergiert MSGD schneller und ist stabiler. In die Praxis wird fast immer das MSGD Verfahren bevorzugt.

Zum besserer Anwendung der Gradientenabstiegsverfahren wurden mehrere Optimierte Lernverfahren entwickelt. Im folgenden wird ein kurzer Einblick über die bekanntesten Lernverfahren (*Optimizer*) gegeben.

Alle heutige Optimizer haben SGD als Vorfahren und der Hauptnachteil von SGD ist , dass es die gleiche Lernrate für die Anpassung aller Netzwerkparameter verwendet und diese Lernrate wird auch während des Trainings nie geändert.

4 Kompression von DNN

Die neueren maschinellen Lernmethoden verwenden immer tiefer neuronale Netze wie z.B *Xception(134 Layers)*, *MobileNetV2(157 Layers)*, *InceptionResNetV2(782 Layers)*, um Ergebnisse auf dem neuesten Stand der Technik zu erzielen. Aber die Verwendung von sehr tiefer NNs bringt mit sich nicht nur eine deutliche Verbesserung der Modellleistung, sondern auch einen bedeutenden Bedarf an Rechenleistung und an Speicherplatz, was der Einsatz solcher Modelle auf Echtzeitsystemen mit begrenzten Hardware-Ressourcen schwierig macht. Es wurden bisher mehrere Ansätze untersucht, um die dem NN zugewiesenen Ressourcen effizienter zu nutzen:

- Die Modellbeschneidung (*Network pruning*), die die redundanten und die nicht relevanten Verbindungen zwischen Neuronen entfernt.
- Die Destillation von NNs, die ermöglicht, die großen Modellen in kleinen zu komprimieren.
- Die Quantisierung von NN, die für die Darstellung von einzelmem Netzparameter weniger als 32 Bits nutzt.
- Huffman-Codierung, die eine komprimierte Darstellung des Netzwerks ermöglicht.

Im folgenden werden nur die Beschneidung, die Quantisierung von NN und die Anwendung von Huffman auf NN mehr eingegangen werden.

4.1 Pruning Network

Wie oben schon erwähnt, wird beim *Pruning* neuronaler Netzwerke versucht, unwichtige oder redundante Verbindungen oder komplette Neuronen aus dem Netzwerk zu entfernen, um ein Netz mit möglichst geringer Komplexität zu erhalten. Mit unwichtigen Verbindungen werden die Parameter (Gewichte und Bias) gemeint, die fast null sind,

denn Parameter mit Nullwert haben keinen Einfluss auf das Output des Neurons, sie sind einfach überflüssig. Während oder nach dem Training gibt es mehrere Parameter, die nicht wirklich oder nicht zu viel zum Neuronenergebnis beitragen, obwohl sie keinen Nullwert haben, deshalb ist es zum Reduzieren der Netzwerkdicke notwendig, anderen Maßstäbe als den Nullwert anzulegen, um Verbindungen zu entfernen.

Das Pruning-Verfahren bietet einige Vorteile wie Reduzierung der Speicher- und Hardwarekosten, die Trainingsbeschleunigung, die schnellere Antwortzeit und das Verringern der Wahrscheinlichkeit der Overfitting(5.4.1). Es ist sehr wichtig zu beachten, dass die Anwendung vom Pruning-Verfahren auf ein NN nur Sinn macht, wenn das NN teilweise oder komplett trainiert ist, sonst macht das Pruning nur eine Reduktion der Anzahl der Netzwerkparameter.

Es gibt zwei Hauptszenarien für das Pruning von NN. Die erste besteht darin, die irrelevanten Verbindungen in einem komplett trainierten NN zu entfernen. Mit komplett trainierten NN wird gemeint, dass die erwünschte Genauigkeit schon erreicht ist. Im zweiten Szenario wird Pruning während des Trainings durchgeführt, das wird sehr oft als *iteratives Pruning* bezeichnet. Dabei wird vor Trainingsbeginn bestimmte Dinge festgelegt, wie z.B. ab wann wird das Netzwerk beschnitten und wie oft es durchgeführt werden soll. Das erste Szenario ist einfacher anzuwenden, denn man muss nur darauf warten, bis es keine Verbesserung der Genauigkeit des NN mehr gibt und dann Pruning auf das NN anwenden, aber damit verliert man großen Vorteile des Pruning, die die Beschleunigung des Trainings und Reduzierung der Overfitting-Wahrscheinlichkeit sind. Aus diesen Gründen wird in der Praxis das zweite Szenario bevorzugt, aber die richtigen Hyperparameter zu finden, ist kompliziert. Eine zu früh Beschneidung kann z.B. die Genauigkeit des NN zu sehr verschlechtern, denn es kann sein, dass eine Verbindung, die nur nach einer späteren Gewichts Anpassung zum Ergebnis eines Neurons hätte beigetragen können, entfernt würde, aber es bietet eine Beschleunigung des Trainings. Was eine zu spät Beschneidung angeht, haben wir fast die gleichen Nachteile wie im ersten Szenario und eine sehr häufiges Pruning kann das Training auch verlangsamen. Das ganze Prozess muss sehr oft leider mehrmals angewendet, um die richtigen Hyperparameter zu finden, sodass es sich wirklich nicht mehr lohnt. Ein graphischer Ablauf der Beschneidung ist in der Abbildung 19 zu sehen. Wenn wir zu viel auf einmal schneiden, könnte das Netzwerk so sehr beschädigt werden, dass es nicht mehr wiederhergestellt werden kann.

Die Hauptarbeit bei Pruning-Verfahren ist sicherlich, die guten Kriterien für die Bewertung der Wichtigkeit von Parametern zu finden und es ist vielleicht einer der Gründe, warum Pruning-Verfahren bisher nicht so populär ist, obwohl es immer noch zu schwierig ist, tiefe NNs beispielsweise auf mobile Geräten mit eingeschränkten Ressourcen durchzuführen.

Bisher gibt es viele Kriterien für die Bewertung der Wichtigkeit eines Parameters, die sich miteinander unterscheiden und je nach Anwendung Vor- und Nachteile aufweisen. Im Folgenden werden einige Kriterien vorgestellt.

- **Schwellenwert**

Für die Bewertung der Wichtigkeit eines Parameters verwenden [5, Han et al] einen Schwellenwert θ , also alle Parameter mit einem Wert in $[-\theta, \theta]$ werden eli-

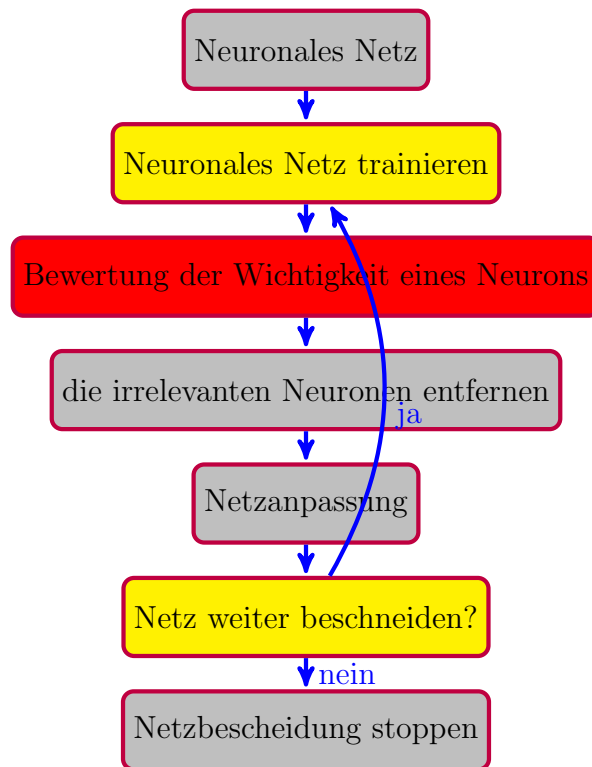


Abbildung 19: Netzbeschneidung während des Trainings

minierte. So ein Kriterium macht das Pruning schneller und effizienter, weil es sehr einfach ist, die zu eliminierenden Parameter zu finden. Da nach dem Pruning immer ein Genauigkeitsverlust auftritt, muss das NN erneut trainiert werden, um seine Genauigkeit wiederherzustellen. Aber dieser Ansatz hat viele Nachteile. Erstens muss der optimale Schwellenwert gefunden werden und dafür muss der ganze Prozess (Schwellenwert auswählen, NN beschneiden, Verlust vergleichen) mehrmals wiederholt werden, um nur den optimalen Schwellenwert zu finden, deshalb können wir nie sicher sein, dass wir den optimalen Schwellenwert gefunden haben. Zweitens wird eine große Reduktion der Rechenkosten nur in FCLs und nicht in ConvLs beobachtet [6]. Die modernen Architekturen von CNNs können also bezüglich der Rechenkosten aus dem Verfahren keinen Vorteil ziehen, denn sie bestehen am meisten aus Faltungsschichten. Drittens muss spärliche Bibliotheken oder spezielle Hardware verwendet werden, damit die Bewertung des beschnittenen NN effektiv wird.

- **Filter und Feature-Map Pruning**

Anstatt die Gewichte in einem Filter elementweise zu entfernen, schlagen [6, Li et al] vor, das ganze Filter und die entsprechende Feature-Map zu entfernen. Dieses Vorgehen bringt mit sich im Vergleich mit dem oben vorgestellten Ansatz mehrere Vorteile. Es werden z.B. keine Bibliotheken, die eine Beschleunigung durch spärliche Operationen über CNNs ermöglichen, mehr benötigt. Noch dazu werden die

Rechenkosten deutlich reduziert. Bei diesem Ansatz werden die weniger nützlichen Filter aus einem vollständig trainierten CNN entfernen. Zur Auswahl der zu entfernenden Filter wird zuerst die l_1 -Norm(4.1) jedes Filter im NN berechnet und dann werden die m Filter mit der kleinsten l_1 -Norm in jeder Schicht entfernen, wobei m ein Hyperparameter ist, der die Anzahl der zu löschenden Filter angibt.

$$\|F\|_1 = \sum_{i=1}^h \sum_{j=1}^w |F_{i,j}| \quad (4.1)$$

$(w, h) :=$ Breite und Höhe des Filters

Der Grund, warum nur die Filter mit einer kleineren l_1 -Norm entfernt werden, liegt daran, dass sie neigen dazu, Feature-Maps mit geringen Aktivierungen im Vergleich zu den anderen Filtern in der selben Schicht zu erzeugen und das kann gut in der Abbildung 20 festgestellt werden. Wobei das Feature-Map links aus der Abbildung 20 ist das selbe mit dem aus Abbildung 3b und das Feature-Map rechts erhalten wir, indem wir das Filter aus Abb. 3a mit drei multiplizieren.



Abbildung 20: Einfluss der Intensität des Filters auf Feature-Map

Noch interessanter an diesem Verfahren ist, dass das Pruning und das neue Training des Netzwerks auf einmal auf mehrere Schichten durchgeführt werden, was die Beschneidungszeit noch weiter beschleunigt und noch effizienter ist, wenn es um sehr tiefe Netzwerke wie *InceptionResNet* oder *GoogLeNet* angeht. Zur Beschneidung von Filter auf mehrere Schichten kann man die zu entfernenden Filter auf jeder Schicht entweder unabhängig von anderen Schicht oder nicht auswählen. Sollte man die zu löschenden Filter auf jeder Schicht unabhängig von anderen Schichten bestimmen, so muss man mit höheren Rechenkosten rechnen, denn es werden Filter, die in vorherigen Schichten schon ausgenommen wurden, in der Berechnung der Summe der absoluten Gewichte noch miteinbezogen. Nach Li et al [6] ist die zweite Strategie nicht global optimal und kann trotzdem unter bestimmten Umständen zu besseren Ergebnissen führen. Die Abbildung 21 zeigt, wie die Entfernung eines Filters die Sichtausgabe beeinflusst.

- **Automatische Beschneidung**

Die bisher beschriebenen Kriterien oder die meisten von ihnen fordern, dass das beschnittene Netzwerk noch trainiert wird, um die Endgewichte für die verbleibenden spärlichen Verbindungen zu lernen. Das ganze macht das Pruning-Verfahren

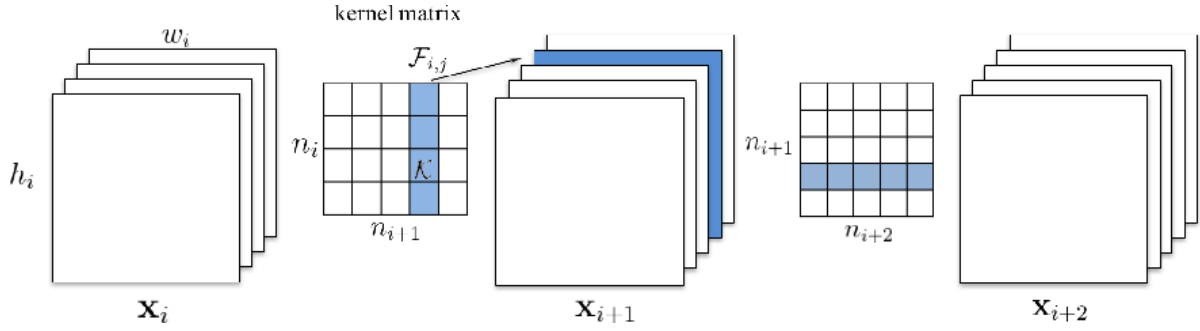


Abbildung 21: Das Beschneiden eines Filters führt zum Entfernen der entsprechenden Feature-Map und der zugehörigen Kernel in der nächsten Ebene[6].

noch schwieriger, denn es wird nur dann unterbrochen, wenn der Genauigkeitsverlust wirklich groß ist, was zu viel Zeit in Anspruch nimmt. Um dieses Problem zu lösen, schlagen *Manessi et al*[7] eine differenzierbare Technik vor, die es ermöglicht, das Pruning durchzuführen, wenn die Gewichte des Netzwerks während der Trainingsphase angepasst werden und die optimalen Schwellwerte automatisch zu suchen. Dieses Verfahren ist inspiriert von dem von Han et al[5] und löst viele Probleme, die das Verfahren von *Han et al* nicht lösen könnten oder wegen der Art und Weise, wie sie das Pruning durchführen, entstanden. Erstens werden die Schwellwerte während des Trainings wie die Gewichte auch gelernt, es kann also eine große Menge von unterschiedlichen Schwellwerten ausprobiert werden und anstatt der selbe Schwellwert auf alle Layers anzuwenden, wird jeder Schicht eine Schwelle zugewiesen. Da Nicht alle Layers gleich empfindlich gegenüber dem Pruning sind und die Verwendung eine selbe Schwelle auf jede Schicht davon ausgeht, dass sie sind, muss dieses Verfahren optimaler. Zweitens, anstatt die Trainingszeit zu erhöhen, weil auch neue Parameter gelernt werden müssen, wird sie reduziert, das liegt daran, dass die Netzwerkparameter irgendwann mehr oder weniger spärlich werden, was Feedforward- und Backpropagationzeit erheblich reduziert. Es ist wichtig zu betonen, dass das Netzwerk nur einmal trainiert werden muss, um eine selbe bzw. bessere Genauigkeit wie bzw. als die anderen Methoden zu erzielen.

Aufgrund seiner Automatisierung und Effizienz wird das Pruning in Zukunft sicherlich immer häufiger eingesetzt werden.

4.2 Quantisierung von neuronalen Netzwerken

Im Allgemein wird von Quantisierung gesprochen, wenn es versucht wird, sich den Werten einer großen Menge mit denen einer kleineren Menge zu nähern. Im Maschine Lernen (ML) ist die Quantisierung der Prozess der Transformation eines ML-Programms in eine approximierte Darstellung mit verfügbaren Operationen mit geringerer Genauigkeit.

Methoden

- Matrixfaktorisierung

- Vektorquantisierung
- low-bit Quantisierung

Matrixfaktorisierung

Die Matrixfaktorisierung versucht, eine Matrix durch die Singulärwertzerlegung (SVD) zu approximieren. Die Singulärwertzerlegung zur Annäherung an eine Matrix $A \in \mathbb{R}^{m \times n}$ verwendet eine Matrix $B \in \mathbb{R}^{m \times n}$ vom Rang $k \leq \min(m, n)$, siehe Gleichung (4.2), wobei U und V orthogonale Matrizen und S eine Diagonalmatrix, deren Diagonaleinträge Singulärwerte heißen und bei Betrachtung von oben nach unten monoton abnehmen. Sollte man B direkt speichern, so hat man keine Kompression gemacht, sondern nur Informationen und Rechenzeiten verschwendet, denn die Matrizen A und B haben die gleiche Dimension, also liegt der Hauptvorteil von der Matrixfaktorisierung darin, wie die Matrix B gespeichert wird.

$$\begin{aligned} A &= USV^T \\ B &= \hat{U}\hat{S}\hat{V}^T = U[:, :k]S[:, :k]V[:, :k]^T \\ \|A - B\|_F &= \sqrt{\sum_{i=1}^m \sum_{j=1}^n (A - B)_{i,j}^2} := \text{Frobenius Norm} \end{aligned} \quad (4.2)$$

Eigentlich werden \hat{U} , \hat{S} und \hat{V} statt B gespeichert, damit wird $mk + kk + kn$ statt mn Speicherplatz verwendet und da \hat{S} Diagonalmatrix ist, kann auch einfach nur die Diagonaleinträge gespeichert werden, was zu einer Kompressionsrate von $\frac{mn}{k(m+n+1)}$ führt. Die Verwendung von Matrizen von kleinen Rang hebt die Tatsache hervor, dass es redundante Parameter in NN-Parametern gibt. Diese Redundanzeigenschaft wird ausgenutzt, um die Geschwindigkeit zu erhöhen [8, Denton et al]. Im Sinne der Frobeniusnorm ist B die bestmögliche Approximation von A mit Singulärwertzerlegung durch eine Matrix von Rang k . Angenommen bildet die zu approximierenden NN-Parameter ein Bild, dann zeigt die Abbildung 22, wie man die NN-Parameter approximieren kann.



SW: Singulärwerte

Abbildung 22: Parameter Approximieren durch Matrixfaktorisierung

Wie man in der Abbildung 22 sieht, reichen relativ wenige Singulärwerte (ca. 5) schon aus, um die groben Konturen zu erkennen. Details erkennt man bei ca. 10 Singulärwerten schon. Generell wird das Bild mit mehr Singulärwerten schärfer.

Quantisierung mit weniger Bits (Low-bit Quantization)

Im Bereich des *Deep Learning* ist das Standard numerische Format für Forschung und Einsatz bisher 32-Bit Fließkommazahlen oder FP32, denn es bietet eine bessere Genauigkeit, aber die anderen Formate wie 8-, 4-, 2- oder 1-Bits werden auch verwendet, obwohl sie mehr oder weniger einen Verlust an Genauigkeit aufweisen. Die Verwendung von weniger genauen numerischen Formaten hat nicht nur einen kleinen Verlust der Netzleistung zur Folge, sondern auch die Verwendung von deutlich reduzierter Bandbreite und Speicherplatz. Noch dazu beschleunigt die Quantisierung die Berechnungen, denn die ganzzahlige Berechnung zum Beispiel ist schneller als die Fließkommaberechnung.

Bei Quantisierung mit wenigen Bits geht es mehr oder weniger um die Abbildung eines großen Bereiches auf einen kleinen und dazu werden zwei Hauptwerte benötigt: der dynamische Bereich des Tensors und ein Skalierungsfaktor. Angenommen haben wir einen dynamischen Bereich $[0, 500]$ und einen Skalierungsfaktor 5, dann ergibt sich der neue Bereich $[0, 100]$, es werden also Werte zwischen $[5k, 5(k+1)]$ oder $[5k - 0.5, 5k + 0.5]$ auf $5k$ abgebildet. Es ist sinnvoller, den Skalierungsfaktor unter Berücksichtigung der Anzahl und der Verteilung der Werte im dynamischen Bereich des Tensors auszuwählen, sonst können zu viel Informationen verloren gehen.

Sobald ein anderes Format mit weniger Bits als FP32 verwendet wird, spricht man in ML von Quantisierung. Als Hauptvorteile von Quantisierung haben wir:

- Reduzierung des Speicherplatzbedarfs.
 - 16 bit float: 2x Reduzierung der Modellparameter
 - 8 bit integer: 4 Reduzierung der Modellparameter
 - 4 bit integer: 8x Reduzierung der Modellparameter
- Berechnungen schneller durchführen.
 - Integer Operationen sind schneller
- Reduzierung der benötigten Energie
 - Integer Operationen verbrauchen weniger Energie
- Reduktion der Bandbreite
 - kleinere Modelle und dynamische Werte reduzieren die Bandbreite.

Im Allgemeinen wird ein Skalierungsfaktor für jeden Tensor jeder Schicht berechnet und dieser kann offline oder online gemacht werden. Bei der *Offline* Berechnung werden vor der Bereitstellung des Modells einige Statistiken gesammelt, entweder während des Trainings oder durch die Ausführung einiger Epochen auf dem trainierten FP32-Modell und basierend auf diesen Statistiken werden die verschiedenen Skalierungsfaktoren berechnet und nach der Bereitstellung des Modells festgelegt. Durch die Anwendung dieser

Methode läuft man die Gefahr, dass zur Laufzeit die Werte, die außerhalb der zuvor beobachteten Bereiche auftreten, abgeschnitten werden, was zu einer Verschlechterung der Genauigkeit führen kann. Bei der *online* werden die *Min/Max*-Werte für jeden Tensor dynamisch zur Laufzeit berechnet. Bei dieser Methode kann es nicht zu einer Beschneidung kommen, jedoch können die zusätzlichen Rechenressourcen, die zur Berechnung der Min/Max-Werte zur Laufzeit benötigt werden, unerschwinglich.[15]

Es gibt zwei Arten und Weisen, wie die Quantisierung durchgeführt wird. Die erste ist das vollständige Training eines Modells mit einer gewünschten niedrigeren Bit-Genauigkeit (kleiner als 32 Bits). Die Quantisierung mit sehr geringer Genauigkeit ermöglicht ein potenziell schnelles Training und Inferenz, aber der Hauptproblem mit diesem Ansatz ist, dass Netzparameter nur bestimmte Werte annehmen können, so ist die Aktualisierung der Netzparameter bzw. das Backpropagation nicht mehr wohldefiniert. Das zweite Szenario quantisiert ein trainiertes FP32-Netzwerks mit einer geringeren Bit-Genauigkeit ohne vollständiges Training. Eine aggressive Quantisierung hat im Allgemeinen einen negativen Einfluss auf die Netzleistung und um diesen Leistungsabfall zu überwinden, wird sehr oft auf Methoden wie das erneuerte Training des Netz nach der Quantisierung, die gleichzeitige Verwendung von verschiedenen Formaten oder die uneinheitliche Quantisierung zurückgegriffen.

Vor kurzem haben [16, Yoni et al] einen neue Ansatz für die Quantisierung vorgeschlagen, der das lineare Quantisierungsproblem als *Minimum Mean Squared Error* (MMSE) Problem löst und der nicht nur die 4-Bit (INT4) Quantisierung von schon trainierten Modellen ohne ein neues Training des Modells, sondern auch die Einsparung von Chipfläche (*chip area*) ermöglicht. Obwohl diese Methode minimalen Verlust der Genauigkeit (*accuracy*) aufweist, liefert sie Ergebnisse auf dem neuesten Stand der Technik und nach [16] weist dieser Ansatz den geringeren Genauigkeitsverlust als alle Quantisierungsverfahren auf.

kommt noch dazu

- Der Speicherbedarf für ein Low-Bit-Präzisionsmodell kann ebenso wie der Stromverbrauch erheblich reduziert werden.
- nutzen mehrere Tensor um eine bessere Approximation von Tensor die gegenüber der Quantisierung sehr empfindlich sind statt nur eine Tensor
- Reduzierte Float Darstellung (32 \rightarrow 16)
- Hybrid Quantisierung: Mischung von 8 bit Integer und 32 bit float Bits
- 8 oder 16 bit Integer Quantisierung

4.3 Huffman Codierung

Die Huffman-Codierung ermöglicht eine verlustfreie Datenkompression, indem sie jeder einzelnen Dateneinheit eine unterschiedlich lange Folge von Bits zuordnet. Daraus folgt, dass eine gute Möglichkeit zur besseren Verwaltung der dem Modell zugeordneten Ressourcen ist: Erstmal das Netzwerk zu beschneiden, dann zu quantisieren und am Ende die Huffman-Codierung durchzuführen.

5 Experiment

5.1 Lernrate

Wie in 3.2.3 gesehen, die Lernrate sagt uns, wie schnell wir ans Ziel kommen möchten.

- $\eta < \eta_{opt}$: So sind wir sichern, ein lokales oder globales Minimum zu erreichen. Aber die Anzahl der benötigten Iterationen bis zum Minimum steigt offensichtlich an und das Verfahren kann in einem unerwünschten lokalen Minimum stecken bleiben.
- $\eta > \eta_{opt}$: Hier wird die Anzahl der Iterationen zwar verringert, aber das Verfahren ist nicht stabil, denn **in der Nähe vom Minimum oder es** wird über das Minimum ständig hinausgegangen und es ist nicht mehr sichern, zum lokalen oder globalen Minimum zu gelangen.

In die Praxis gibt es Methoden und Funktionen, um die Lernrate während des Trainings anzupassen. z.B die *KERAS* Funktion *ReduceLROnPlateau*, die die Lernrate reduziert, wenn das NN nach einer bestimmten Anzahl von Epochen keine Verbesserung mehr aufweist.

5.2 Besondere CNNs

5.2.1 AlexNet

Wie in 2.4 schon erwähnt, wurde AlexNet von *Alex Krizhevsky* et al [18] im Jahr 2012 entwickelt. AlexNet hat 60 Millionen Parametern und etwas 650.000 Neuronen, besteht aus fünf ConvLs, von denen einige von Max-Pooling gefolgt sind, und drei FCLs (siehe Abbildung 23) [18]. Das Besondere an AlexNet ist aber nicht die Anzahl der Schichten, sondern der Einsatz bestimmter Methoden oder Techniken:

- **ReLU**: Vor AlexNet waren *tanh* und *sigmoid* die Standard Aktivierungsfunktionen, aber wegen ihrer Sättigung bei hohen oder sehr niedrigen Werten und der Tatsache, dass sie in diesen Bereichen eine Steigung nahe bei null haben, verlangsamten sie stark die Gewichtsanpassungen, was nicht der Fall bei *ReLU* ist, das eine Steigung gleich null nur bei negativen Werten und bei positiven höheren Werten eine Steigung nicht nahe bei null hat. Das Training von CNN wird durch *ReLU* um ein Vielfaches schneller als ein Äquivalent mit *tanh* beschleunigt [18].

- **Überlappendes Pooling:** Die normalen Pools funktioniert wie in 3.1.4 ($pool_size = stride$), aber die Überlappenden verwenden einfach ein $stride$ kleiner als das $pool_size$. Nach [18] verbessern die überlappenden Pool die Netzgenauigkeit und macht das Netz gegenüber Overfitting robuster.
- **Dropout und Data Augmentation :** Ein anderer Vorteil von AlexNet ist die Verwendung von Dropout in FCLs, das sich heute als die beste oder eine der besten Regulierungsmethoden erweist und der Vermehrung der Daten durch Spiegelung, die die Wahrscheinlichkeit von Overfitting reduziert.

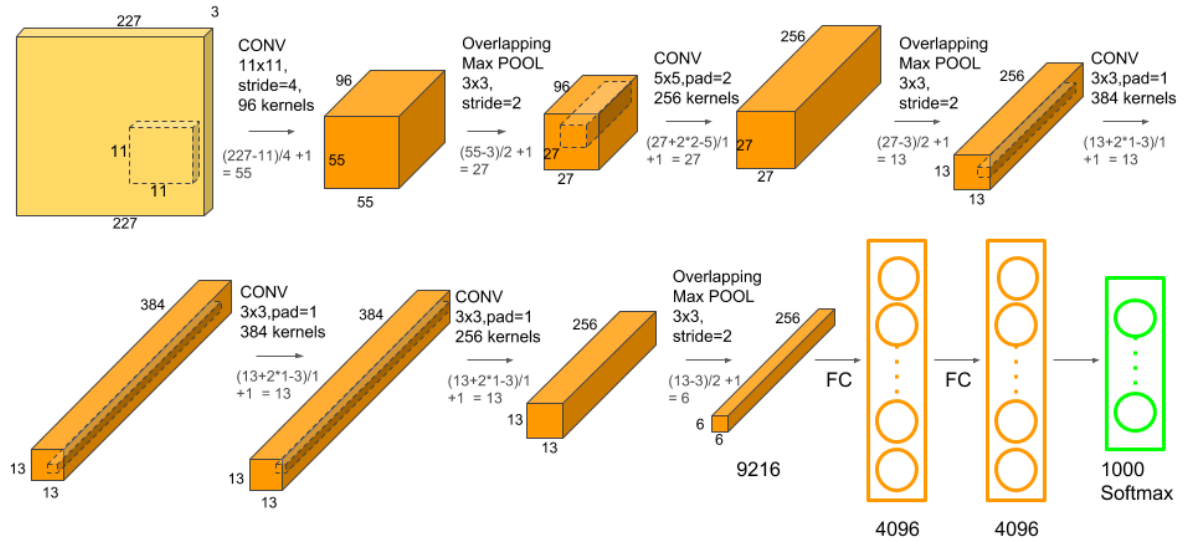


Abbildung 23: AlexNet Architektur zum Bild

AlexNet kann aufgrund seiner Größe ($\sim 240MB$) nicht immer auf Systemen mit begrenztem Speicherplatz eingesetzt werden, deshalb ist man seit AlexNet ständig auf der Suche nach neuen Architekturen, die weniger Parameter und Speicherplatz brauchen und gleichzeitig die Ergebnisse auf der Stand der Technik erreichen. Es ist sicherlich in diesem Zusammenhang, dass viele neue effiziente Modelle wie *SqueezeNet*, *Xception* und *MobileNet* entstanden sind.

5.2.2 SqueezeNet

SqueezeNet von [19, Song Han et al] erreicht die Genauigkeit von AlexNet mit 50x weniger Parametern und einer Größe kleiner als $0.5MB$. Die *SqueezeNet* Architektur und einige Variante davon können in Abbildung 24 entnehmen. Zum Erreichen einer solchen Architektur werden drei Hauptstrategien und ein spezieller Block verwendet. Bezüglich

5 Experiment

dieser Strategien wurde zuerst die meisten 3×3 Filter durch 1×1 Filter ersetzt, was die Parameteranzahl stark reduziert, denn ein 1×1 Filter nimmt $9\times$ weniger Parameter als ein 3×3 Filter in Anspruch, dann wird die Anzahl der Eingangskanäle auf 3×3 Filter verringern, denn die Tiefe eines Filters in einem ConvL ist gleich die Tiefe vom ConvL Input und ein Input mit einer großen Tiefe wird die Parameteranzahl deutlich zunehmen, obwohl Filter mit kleiner räumlicher Dimension benutzt werden. Die letzte Strategie besteht darin, die Reduzierung der räumlichen Dimension im Netzwerk zu verzögern, so werden Feature-Maps mit größerer Raumaufösung am Ende enthalten. [19, Song Han et al] sind der Meinung, dass größere Feature-Maps zu einer höheren Klassifizierungsgenauigkeit führen.

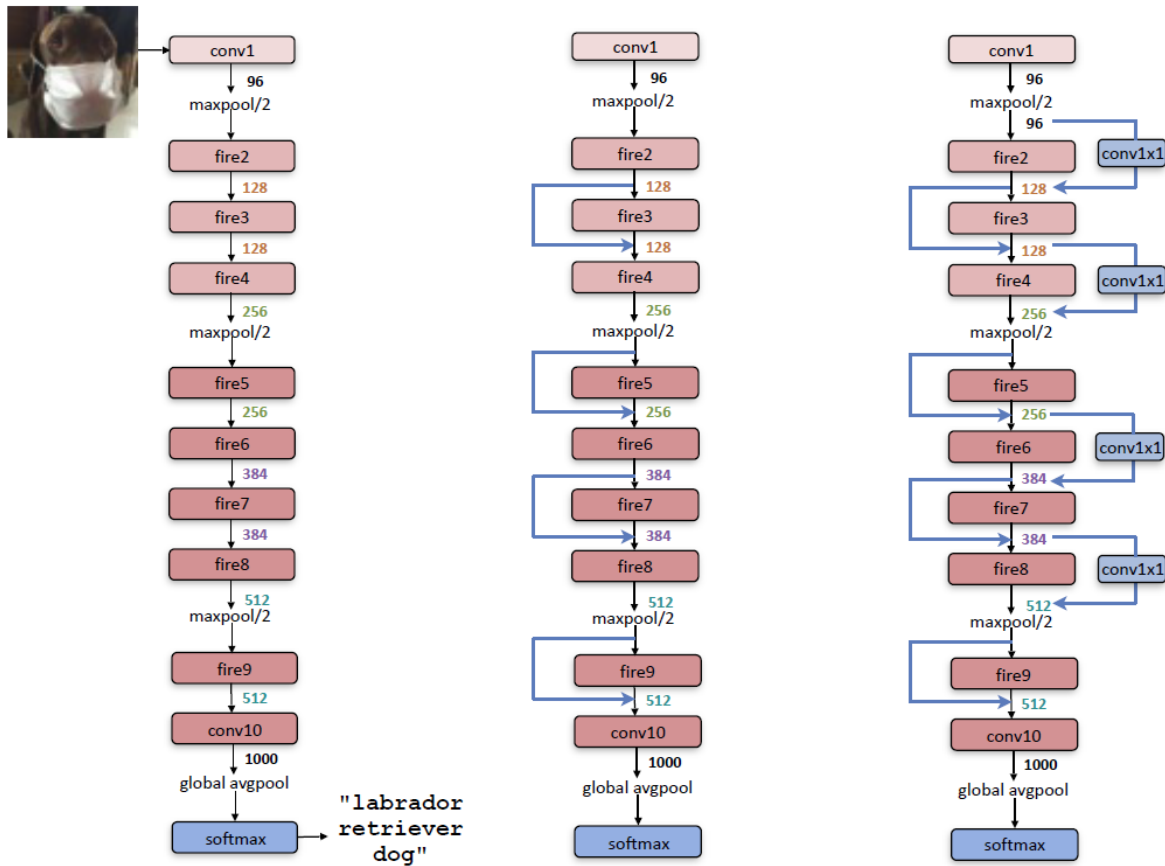


Abbildung 24: SqueezeNet Architekturen [19]

Um die Ziele der ersten und zweiten Strategien zu erreichen, wird ein neuer Baustein (*fire module*) (FM) eingeführt. Das FM besteht aus einem *Squeeze*-(mit 1×1 Filtern) und einem *Expand*(mit 1×1 und 3×3 Filtern)-Block (siehe Abbildung 25). Der *Squeeze* Block ermöglicht die Reduzierung der Dimensionalität, indem die Anzahl Feature-Maps verringert wird, während die wichtigsten Features erhalten bleiben und die Anwendung von *ReLU* nach dem *Squeeze* Block ermöglicht dem Netzwerk, komplexere Features zu lernen.

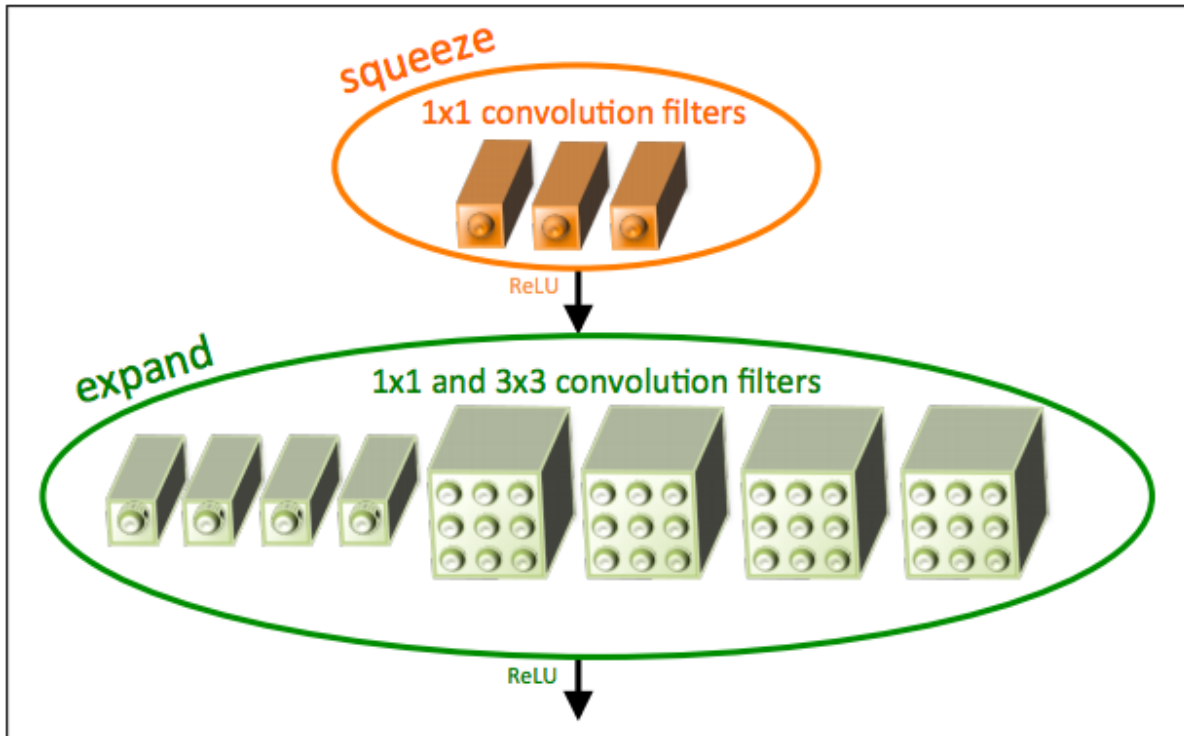


Abbildung 25: fire_module [19]

Was der *Expand* Block angeht, funktioniert der Block mit 1×1 Filtern genauso wie der *Squeeze* Block und der Block mit 3×3 Filtern ermöglicht der Erlangung von Empfangsfelder(*receptive field*), die in der Lage sind, lokale räumliche Informationen zu erfassen, was 1×1 Filter nicht schaffen können.

Wie es in Abbildung 24 zu sehen ist, kann eine Verbindung zwischen zwei Layers über eine oder mehrere Schichte überspringen. Solche Verbindungen werden *Residual connections* (RC) genannt und Netze mit RCs gehören zu der Klasse von Residual Netzwerke (mehr dazu siehe [20]). Wie die Abbildung ?? zeigt, liefert *SqueezeNet* mit Simple RCs als die anderen Varianten

5.2.3 Xception: Extreme Inception

Xception ist eine Variante von *InceptionV3* **Related work**

- Inception Family
- Depthwise separable convolutions
- Residual Connection
- Die Leistungssteigerungen sind nicht auf eine Kapazitätssteigerung zurückzuführen, sondern auf eine effizientere Nutzung der Modellparameter.

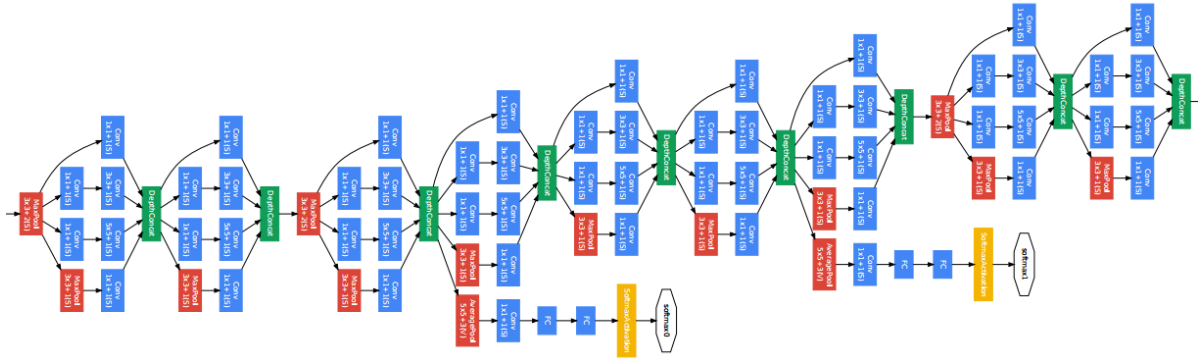


Abbildung 26: Xception Architektur

5.2.4 MobileNet

5.3 Algorithmen zur Optimierung des Gradientenabstiegsverfahren: Optimizer

Die Wahl des Optimierungsalgorithmus für ein CNN kann den Unterschied zwischen guten Ergebnissen in Minuten, Stunden und Tagen ausmachen.

5.3.1 Adaptive Gradient Algorithm (AdaGrad)

AdaGrad bietet während des Netztrainings nicht nur die Möglichkeit, die Lernrate zu verändern, sondern auch für jeden Parameter eine geeignete Lernrate zu finden. Die AdaGrad-Aktualisierungsregel ergibt sich aus der folgenden Formel:

$$\alpha_t = \sum_{i=1}^t (g_{i-1})^2 \quad \theta_{t+1} = \theta_t - \eta_t g_t \quad (5.1)$$

$$\eta_t = \frac{\eta}{\sqrt{\alpha_t} + \epsilon}$$

Voreingestellte Parameter(*KERAS*) : $\alpha_0 = 0.0$ $\eta = 0.001$ $\epsilon = 10^{-7}$

Dabei wird am Trainingsanfang eine Lernrate für jeden Parameter definiert und im Trainingsverlauf separat angepasst. Dieses Verfahren eignet sich gut für spärliche Daten, denn es gibt häufig auftretende Merkmale sehr niedrige Lernraten und seltene Merkmale hohe Lernraten, wobei die Intuition ist, dass jedes Mal, wenn eine seltene Eigenschaft gesehen wird, sollte der Lernende mehr aufpassen. Somit erleichtert die Anpassung das Auffinden und Identifizieren sehr voraussehbarer, aber vergleichsweise seltener Merkmale.[13]. Wie in der Gleichung (5.1) festzustellen, nach einer bestimmten Anzahl von Iterationen haben wir keine Verbesserung der Netzleistung, denn je größer t wird, desto kleiner η_t wird und irgendwann wird η_t so klein, dass $\eta_t g_t$ fast gleich null ist.

Table 1. MobileNet Body Architecture

Type / Stride	Filter Shape	Input Size
Conv / s2	$3 \times 3 \times 3 \times 32$	$224 \times 224 \times 3$
Conv dw / s1	$3 \times 3 \times 32$ dw	$112 \times 112 \times 32$
Conv / s1	$1 \times 1 \times 32 \times 64$	$112 \times 112 \times 32$
Conv dw / s2	$3 \times 3 \times 64$ dw	$112 \times 112 \times 64$
Conv / s1	$1 \times 1 \times 64 \times 128$	$56 \times 56 \times 64$
Conv dw / s1	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 128$	$56 \times 56 \times 128$
Conv dw / s2	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 256$	$28 \times 28 \times 128$
Conv dw / s1	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 256$	$28 \times 28 \times 256$
Conv dw / s2	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 512$	$14 \times 14 \times 256$
5×	Conv dw / s1	$3 \times 3 \times 512$ dw
	Conv / s1	$1 \times 1 \times 512 \times 512$
		$14 \times 14 \times 512$
Conv dw / s2	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
Conv / s1	$1 \times 1 \times 512 \times 1024$	$7 \times 7 \times 512$
Conv dw / s2	$3 \times 3 \times 1024$ dw	$7 \times 7 \times 1024$
Conv / s1	$1 \times 1 \times 1024 \times 1024$	$7 \times 7 \times 1024$
Avg Pool / s1	Pool 7×7	$7 \times 7 \times 1024$
FC / s1	1024×1000	$1 \times 1 \times 1024$
Softmax / s1	Classifier	$1 \times 1 \times 1000$

Abbildung 27: MobileNet Architektur

5.3.2 Root Mean Square Propagation(RMSProp)

RMSProp wie AdaGrad findet für jeden Parameter eine geeignete Lernrate und zur Anpassung der Netzparameter basiert der RMSProp Optimizer auf den Durchschnitt der aktuellen Größen der Gradienten statt auf der Summe der ersten Moment wie in AdaGrad. Da $E[g^2]_t$ nicht schneller als α_t (5.1) ansteigt, wird die radikal sinkenden Lernraten von Adagrad deutlich verlangsamt. Die Parameteranpassungen richten sich nach der folgenden Gleichung:

$$\begin{aligned}
 E[g^2]_t &= \alpha E[g^2]_{t-1} + (1 - \alpha) g_t^2 \\
 \theta_{t+1} &= \theta_t - \frac{\eta}{\sqrt{E[g^2]_t + \epsilon}} g_t, \quad \epsilon \approx 0
 \end{aligned}
 \tag{5.2}$$

Der RMSProp funktioniert besser bei Online- und nicht-stationären Problemen.

5.3.3 Adaptive Moment Estimation(Adam)

Der Adam[14] Optimizer ist auch ein adaptiver Algorithmus, der die ersten und zweiten Momente der Gradienten schätzt, um individuelle adaptive Lernraten für verschiedene Parameter zu berechnen. Adam weist die Hauptvorteile von AdaGrad, das mit spärlichen Gradienten gut funktioniert, und RMSProp, das einige Probleme von AdaGrad löst und das für nicht-konvexe Optimierung geeignet ist, auf. Wie die Parameteranpassung von Adam Optimizer genau funktioniert, ergibt sich aus der folgenden Gleichung:

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1 - \beta_1) g_t, & \hat{m}_t &= \frac{m_t}{1 - \beta_1^t} \\ v_t &= \beta_2 v_{t-1} + (1 - \beta_2) g_t^2, & \hat{v}_t &= \frac{v_t}{1 - \beta_2^t} \\ \theta_{t+1} &= \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t \end{aligned} \quad (5.3)$$

$$\text{Voreingestellte Parameter(KERAS)} : \quad \begin{array}{ll} \beta_1: 0.9 & \beta_2: 0.999 \\ \eta: 0.001 & \epsilon: 10^{-7} \end{array}$$

Zu weiteren Vorteilen der Nutzung von Adam gehört auch seine Einfachheit zur Implementierung, effizienter Nutzung des Speicherplatzes und seine Invarianz zur diagonalen Neuskalierung der Gradienten.

Kleines Experiment

5.4 Problem beim Training von Convolutional neuronalen Netzwerken

5.4.1 Overfitting

Wenn ein von der Maschine gelerntes Modell zu gut auf die Trainingsdaten abgestimmt ist und sehr schlechte Vorhersagen über Daten macht, die es bisher nicht gesehen hat, wird gesagt, dass das Modell an Überanpassung(Overfitting) leidet, anders gesagt, das Modell war nicht in der Lage, die relevanten Merkmale aus den Trainingsdaten zu verallgemeinern, sondern die ganzen Trainingsdaten auswendig zu lernen. Die irrelevanten Informationen aus den Trainingsdaten können z.B. die Position des Tellers in einem Bild sein, wenn man Essen klassifizieren sollte. Im folgenden werden einige Mittel vorgestellt, um mit Overfitting umzugehen.

Strategie gegen Overfitting

Data Augmentation Ein großer Datensatz ist entscheidend für die Leistung tiefer neuronaler Netze. Dass ein Datensatz groß oder ausreichend für das Training eines NNs ist, hängt nur von der Größe des NNs ab und da die NNs, die die besten Leistungen aufweisen, Millionen von Parametern haben, ist fast unmöglich für jedes Problem von Maschine Lernen ausreichende Daten zu finden. Anstatt immer neue Daten zur Verbesserung der Netzleistung zu sammeln, können wir die Leistung des Modells verbessern,

indem wir neue Daten von den bereits vorhandenen Daten aus erzeugen.

Die populären Techniken oder Transformationen zur Vermehrung des Datensatzes sind die horizontalen oder vertikalen Spiegelungen, Drehungen, Skalierungen, Zuschneiden, Parallelverschiebungen und die Gauß'sches Rauschen. Für diese Arbeit habe ich zwei Ansätze im Gebrauch gehabt, um den Datensatz zu erhöhen: Der erste Ansatz besteht darin vor dem Training neue Daten zu erzeugen. Dabei werden die oben erwähnten Techniken vor dem Training angewendet, um zum Trainingszeitpunkt und zur Testzeit einen großen Datensatz zu haben und die originalen Daten werden zur Validierung verwendet. Der zweite Ansatz besteht darin, zum Trainingszeitpunkt und zur Testzeit die neuen Daten zu erzeugen. Hier haben wir keinen Datensatz größer als den originalen, aber die Daten, die ins Netzwerk eingespeist werden, ändern sich ständig. Angenommen wir die Möglichkeit haben, alle diese Transformationen durchzuführen, dann kann es vorkommen, dass eine Photo während der ersten Epoche horizontal gespiegelt wird und während der zweiten um zwanzig Grad gedreht wird, umso weiter. Zur Implementierung des zweiten Ansatzes bietet *KERAS* Framework die Funktion *ImageDataGenerator*. Das interessanteste an *ImageDataGenerator* ist, dass es mehrere Transformationen gleichzeitig anwenden (siehe Abbildung 28).



Abbildung 28: Anwendung von *ImageDataAugmentation*

Je mehr Daten verfügbar sind, desto effektiver können die CNNs sein. Es ist also mehr als wichtig über eine große Datenmenge zu verfügen. Leider können die gesammelten Datensätze nicht alle mögliche Szenarios des reellen Lebens abdecken, deshalb ist es auch bedeutend, CNN mit zusätzlichen synthetisch modifizierten Daten zu trainieren. Die CNNs funktionieren glücklicherweise besser oder immer gut, solange nützliche Daten durch das Modell aus dem ursprünglichen Datensatz extrahiert werden können, selbst

wenn die erzeugten Daten von geringerer Qualität sind.

Dropout Künstliche Neurone sind von biologischen Neuronen inspiriert, aber die Beiden unterscheidet sich sehr voneinander und einer der wichtigen Unterschiede ist, dass biologische Neuronen unvollkommene Maschinen sind, die sehr oft nicht richtig funktioniert und das ist a priori nie den Fall bei künstlichen Neuronen. Wir könnten also glauben, dass KNN die biologische übertreffen könnten. Es sei denn, dass diese Funktionsstörung von biologischen Neuronen nicht eine Schwäche ist, sondern eher eine Stärke ist. Eine der verblüffenden Entdeckungen im Künstliche Intelligenz (KI) Bereich ist, dass es wünschenswert ist, künstliche Neuronen von Zeit zu Zeit zu Fehlfunktionen zu bringen [1]. [Jetzt können wir uns fragen, wie Dysfunktion von Neuronen die Performances CNNs verbessern kann.](#) Die zufällige Hinzufügen von Dysfunktionen in einer Schicht der CNN wird *Dropout* benannt und wurde von [2, Geoffrey E. et al] eingeführt.

Funktionsweise von Dropout

Genauer gesagt, Dropout bezeichnet die zeitliche zufällige Ausschaltung von Neuronen (versteckt und sichtbar) in einem NN [3]. Wie die Abbildung 29 zeigt, wenn ein Neuron zufällig aus dem NN entfernt wird, werden auch all seine ein- und ausgehenden Verbindungen entfernt. In einer Dropout-Schicht wird ein Neuron N unabhängig von anderen Neuronen mit einer Wahrscheinlichkeit p zurückgehalten, d.h. N wird mit einer Wahrscheinlichkeit von p nicht am Ergebnis der Schicht teilnehmen. Während der Testphase werden alle Verbindungen zurückgesetzt, die während des Trainings gelöscht wurden und die ausgehenden Verbindungen gelöschter Neurone mit p multipliziert.

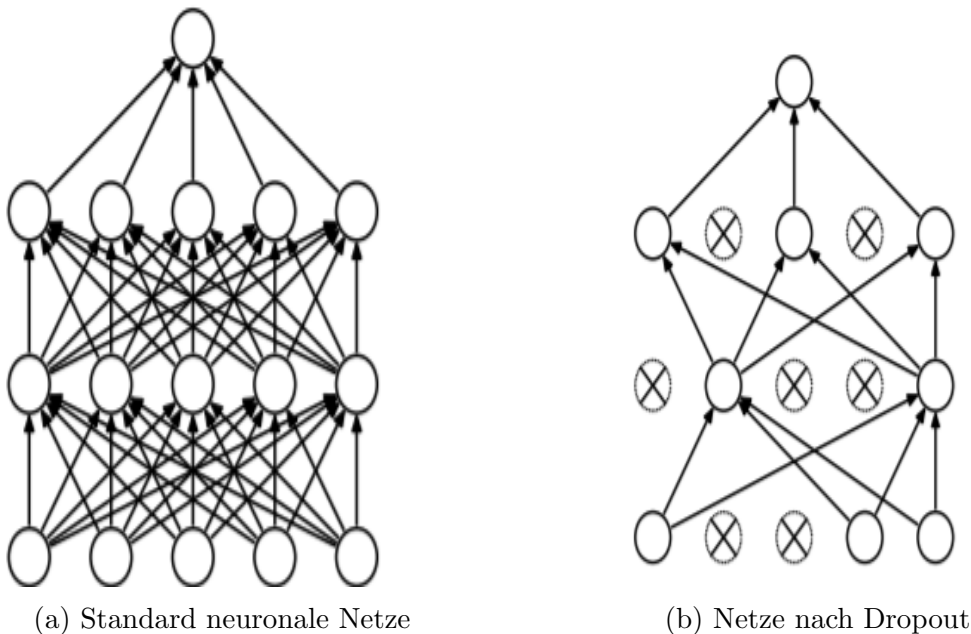


Abbildung 29: Neuronales Netz mit Dropout ausgestattet [3].

Verhinderung der Koadaptationen zwischen Neuronen

Während des Trainings können mehrere Neurone zur Minimierung der Fehlerfunktion so gut zusammenarbeiten, dass die Erkennung bestimmter Merkmale ohne diese Zusammenarbeit nicht mehr möglich ist, aber solche komplexe Koadaptationen können zu einer Überanpassung führen, denn diese komplexe Koadaptationen existieren nicht immer in Testdaten. Da die am Training teilnehmenden Neuronen nach dem Zufallsprinzip nach jeder Epoche ausgewählt werden, haben wir für jedes Training ein neues Modell, was die Neuronen zur Zusammenarbeit zwingt, ohne jedoch voneinander abhängig zu sein, anders gesagt, wird jedes Neuron unabhängig von anderen Neuronen die Muster korrekt lernen können.

Automatische Erhöhung von Training Daten und Regelung

Noch dazu führt die Ausschaltung von Neuronen, wie es in Abbildung 30 angezeigt ist, zu einer automatische Erzeugung neuer Trainingsdaten. Die verwendeten Daten in ausgedünnten Modellen sind also nur eine Abstraktion von echten Daten bzw. Rauschdaten und da wir für ein Netz mit n versteckten Einheiten, von denen jede fallen gelassen werden kann, 2^n mögliche Modelle haben, haben wir 2^n mögliche Abstraktion von unseren Daten und das sollte einer der Gründe sein, warum Dropout effektiver als andere rechnerisch kostengünstige Regler ist [3] und warum die Trainingszeit von NNs mit Dropout mindestens verdoppelt wird.

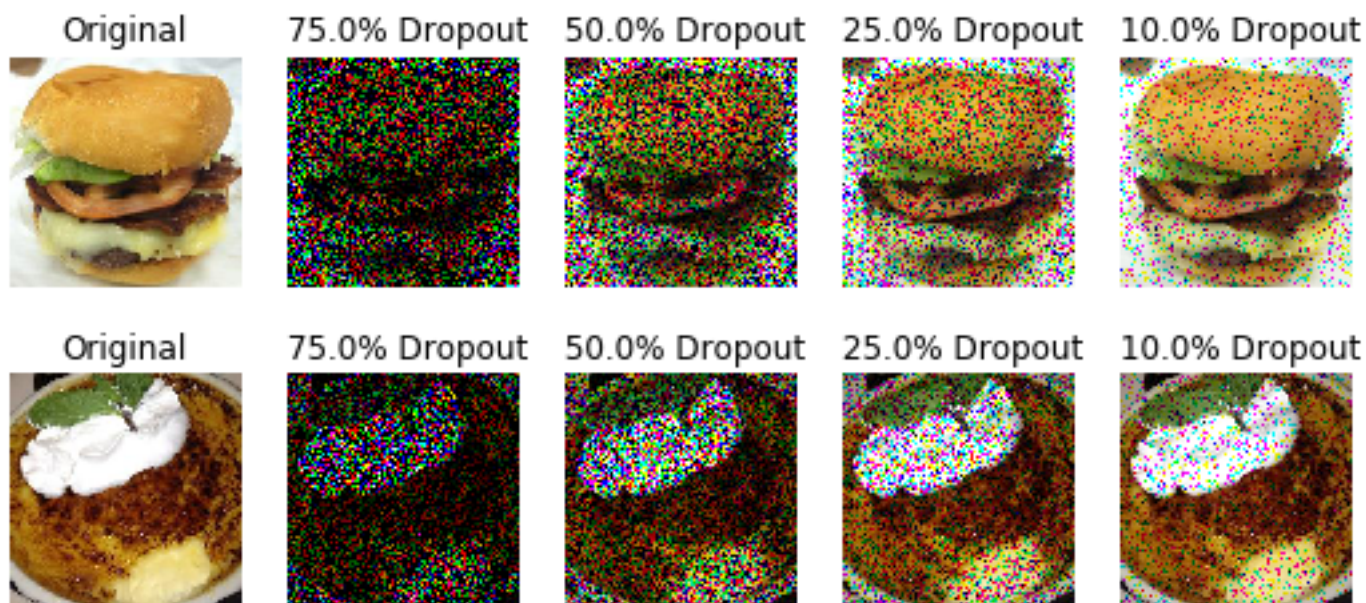


Abbildung 30: Erhöhung des Trainingsdaten durch Dropout

Da heutige CNNs Million von Neurons haben, wäre es unmöglich alle mögliche ausgedünnte Netzwerke zu trainieren, deshalb ist das Modell, das am Ende des Trainings erhalten wird, nur eine durchschnittliche Approximation aller mögliche Modelle, was schon gut, denn es gibt schlechte und gute Modelle.

Batch-Normalisierung Das Training tiefer neuronaler Netze ist sehr kompliziert und ein Grund dafür ist zum Beispiel die Tatsache, dass die Parameter einer Schicht während des Trainings tiefer neuronaler Netze immer unter der Annahme, dass sich die Parameter anderer Schichten nicht ändern, aktualisiert werden und da alle Schichten während des Updates geändert werden, verfolgt das Optimierungsverfahren ein Minimum, das sich ständig bewegt. Ein anderer Grund dafür ist die ständigen Veränderungen im Laufe des Trainings in die Verteilung des Netzeingangs, diese Veränderung wird von [11] als interne kovariante Verschiebung (*Internal Covariate Shift*) genannt. Zur Lösung dieser Probleme schlagen *LeCun et al*[?] vor dem Training das Netzeingangs zu normalisieren. Aber dieser Ansatz bringt nicht so viel, wenn das NN wirklich tief ist, denn nur der Netzeingang profitiert von der Normalisierung und die kleinen Veränderungen in versteckten Schichten werden sich immer mehr verstärken, je tiefer man das Netz durchläuft. Mit der Ausbreitung tiefer NNs dehnt Batch-Normalisierung (BN) [11] diese Idee der Datennormalisierung auf versteckte Schichten tiefer NNs aus. Bei der BN werden die Eingaben in einem Netzwerk standardisiert, die entweder auf die Aktivierungen einer vorherigen Schicht oder auf direkte Eingaben angewendet wird, so standardisiert, dass der Mittelwert in der Nähe von null liegt und die Standardabweichung in der Nähe von eins liegt. Die BN wird über Mini-Batches und nicht über den gesamten Trainingssatz durchgeführt, daher enthalten wir nur Näherungen an tatsächliche Werte der Standardabweichung und des Mittelwerts über das Trainingssatzes, aber wir gewinnen an Geschwindigkeit und an Speicherplatzverbrauch. Die Gleichung (5.4) gibt die formale Beschreibung des BN Algorithmus an.

Batch-Normalisierungstransformation, angewendet auf Aktivierung x über einen Mini-Batch

Input: Werte von x über einen Mini-Batch: $B = \{x_{1..m}\}$

Lernbare Parameter β, γ

Output: $\{y_i = BN_{\beta, \gamma}(x_i)\}$

$$\text{Mini-Batch Mittelwert : } \mu_\beta = \frac{1}{m} \sum_{i=1}^m x_i \quad (5.4a)$$

$$\text{Mini-Batch Standardabweichung : } \sigma_\beta^2 = \frac{1}{m} \sum_{i=1}^m (x_i - \mu_\beta)^2 \quad (5.4b)$$

$$\text{Normalisierung: } \hat{x}_i = \frac{x_i - \mu_\beta}{\sqrt{\sigma_\beta^2 + \epsilon}} \quad (5.4c)$$

$$\text{Skalierung und Verschiebung : } y_i = \gamma \hat{x}_i + \beta \equiv BN_{\gamma, \beta} \quad (5.4d)$$

Wenn $\gamma = \sqrt{\sigma_\beta^2 + \epsilon}$ und $\beta = \mu_\beta$, bekommen wir die gleiche Verteilung wie vor der Batch-Normalisierung, d.h die Eingabe war also schon normalisiert. Interessanterweise

kann das Netz während des Trainings eine bessere Verteilung als die erwünschte finden, denn γ und β sind lernbare Parameter.

Durch die BN kann zum einen eine hohe Lernrate verwendet, was in tiefer NNs ohne BN dazu führen kann, dass die Gradienten explodieren oder verschwinden und in schlechten lokalen Minima stecken bleiben. Die Verwendung einer höheren Lernrate ermöglicht eine schnellere Konvergenz. Zum anderen wird die interne kovariante Verschiebung geringer, was das Training beschleunigt, in einigen Fällen durch Halbierung der Epochen oder besser. Noch dazu wird das Netz durch die BN in gewissem Maße reguliert, daher wird die Verwendung von Dropout bzw. Regulierungstechnik reduziert oder sogar überflüssig und somit eine Verbesserung der Verallgemeinerungsgenauigkeit.

5.4.2 Dataset

6 Abkürzungsverzeichnis

KNN	Künstliches neuronales Netz
CNN	Convolutional Neural Network
KI	Künstliche Intelligenz
NN	neuronales Netz
ConvL	Convolutional Layer
FCL	Fully Connected Layer
PoolL	Pooling Layer
ILSVRC	Large Scale Visual Recognition Challenge
DNN	Tiefe neuronale Netze
ML	Maschine Lernen

Literatur

- [1] P. Kerlirzin, and F. Vallet: Robustness in Multilayer Perceptrons
- [2] Geoffrey E. Hinton and Nitish Srivastava and Alex Krizhevsky and Ilya Sutskever and Ruslan R. Salakhutdinov: Improving neural networks by preventing co-adaptation of feature detectors
- [3] Srivastava, Hinton, Krizhevsky, Sutskever and Salakhutdinov: Dropout: A Simple Way to Prevent Neural Networks from Overfitting

- [4] Ian Goodfellow, Yoshua Bengio, Aaron Courville: Adaptive Computation and Machine Learning series Page 342
- [5] Song Han, Huizi Mao, William J. Dally Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding
- [6] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, Hans Peter Graf Pruning Filters for Efficient ConvNets
- [7] Franco Manessi, Alessandro Rozza, Simone Bianco, Paolo Napoletano, Raimondo Schettini Automated Pruning for Deep Neural Network Compression
- [8] Emily Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun and Rob Fergus Exploiting Linear Structure Within Convolutional Networks for Efficient Evaluation
- [9] Pavel Golik, Patrick Doetsch, Hermann Ney Cross-Entropy vs. Squared Error Training: a Theoretical and Experimental Comparison
- [10] Neuronale Netze: Eine Einführung
- [11] Sergey Ioffe, Christian Szegedy Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift
- [12] Wikipedia: Learning Rate
- [13] John Duchi, Elad Hazan, Yoram Singer: Adaptive Subgradient Methods for Online Learning and Stochastic Optimization
- [14] Diederik P. Kingma, Jimmy Ba: Adam: A Method for Stochastic Optimization
- [15] Compressing Models: Quantization
- [16] Yoni Choukroun, Eli Kravchik, Fan Yang, Pavel Kisilev: Low-bit Quantization of Neural Networks for Efficient Inference
- [17] wikipedia: Artificial neuron
- [18] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton ImageNet Classification with Deep Convolutional Neural Networks
- [19] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, Kurt Keutzer: SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size
- [20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun: Deep Residual Learning for Image Recognition

Erklärung

Ich versichere wahrheitsgemäß, die Arbeit selbstständig verfasst, alle benutzten Hilfsmittel vollständig und genau angegeben und alles kenntlich gemacht zu haben, was aus Arbeiten anderer unverändert oder mit Abänderungen entnommen wurde, sowie die Satzung der Universität Augsburg zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet zu haben.

Ort, den Datum