



**Hochschule für Technik  
und Wirtschaft Berlin**

**University of Applied Sciences**

Untersuchung von Image Colorization Methoden anhand  
Convolutional Neuronal Networks

**Abschlussarbeit**

zur Erlangung des akademischen Grades

**Bachelor of Science (B.Sc.)**

an der

Hochschule für Technik und Wirtschaft Berlin  
Fachbereich IV: Informatik, Kommunikation und Wirtschaft  
Studiengang Angewandte Informatik

1. Prüfer: Prof. Dr. Christin Schmidt
2. Prüfer: M.Sc. Patrick Baumann

Eingereicht von: Adrian Saiz Ferri  
Immatrikulationsnummer: s0554249  
Eingereicht am: XX.XX.2020

# **Abstract**

# Inhaltsverzeichnis

<b>1 Einleitung</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Zielsetzung . . . . .	1
1.3 Vorgehensweise und Aufbau der Arbeit . . . . .	2
<b>2 Grundlagen</b>	<b>3</b>
2.1 Neuronale Netze . . . . .	3
2.1.1 Feedforward Neural Network . . . . .	4
2.1.2 Fully-connected Neural Network . . . . .	4
2.1.3 Aktivierungsfunktionen . . . . .	7
2.1.4 Loss Functions . . . . .	10
2.1.5 Optimierungsalgorithmen . . . . .	11
2.1.6 Backpropagation . . . . .	13
2.2 Convolutional Neural Networks . . . . .	14
2.2.1 Transposed Convolution . . . . .	16
2.3 Autoencoder . . . . .	17
2.4 <i>Lab</i> -Farbraum . . . . .	18
2.5 Image Colorization Methoden . . . . .	18
2.6 Verwandte Arbeiten . . . . .	19
<b>3 Konzeption</b>	<b>20</b>
3.1 Image Colorization als Multimodales Problem . . . . .	20
3.2 Farbraum . . . . .	20
3.3 Binning . . . . .	21
3.4 Netzwerkarchitektur . . . . .	23
3.4.1 U-net . . . . .	24
3.5 Datensätze . . . . .	25
3.6 Image Preprocessing und Augmentation . . . . .	27
3.7 Tools . . . . .	28

<b>4 Implementierung</b>	<b>29</b>
4.1 Binning . . . . .	29
4.2 Datensätze . . . . .	31
4.3 Netzwerkarchitektur . . . . .	31
4.3.1 ConvBlock . . . . .	31
4.3.2 U-Net . . . . .	32
<b>5 Test</b>	<b>34</b>
5.1 Spiel-Datensatz Training . . . . .	34
5.2 CIFAR-100 Subset Training . . . . .	35
5.3 Landscape Datensatz Training . . . . .	38
<b>6 Evaluation</b>	<b>40</b>
6.1 Evaluationsmetrik . . . . .	40
6.2 Evaluation des Spiel-Datensatzes . . . . .	40
6.3 Evaluation des CIFAR-100 Subsets . . . . .	41
6.4 Evaluation des Landscape Datensatzes . . . . .	43
<b>7 Fazit</b>	<b>45</b>
7.1 Zusammenfassung . . . . .	45
7.2 Kritischer Rückblick . . . . .	45
7.3 Ausblick . . . . .	46
<b>Abbildungsverzeichnis</b>	<b>I</b>
<b>Source Code Content</b>	<b>III</b>
<b>Glossar</b>	<b>IV</b>
<b>Abkürzungsverzeichnis</b>	<b>V</b>
<b>Literaturverzeichnis</b>	<b>VI</b>
<b>Onlinereferenzen</b>	<b>VII</b>
<b>Bildreferenzen</b>	<b>VIII</b>
<b>Anhang A</b>	<b>IX</b>
<b>Eigenständigkeitserklärung</b>	<b>X</b>

# **Kapitel 1**

## **Einleitung**

Das Kapitel Einleitung verschafft einen Überblick über die Motivation, die angestrebte Zielsetzung, das genaue Vorgehen und den Aufbau der Arbeit.

### **1.1 Motivation**

Jeder hat sich sicherlich gefragt, vor allem wenn es um Familienbilder geht, wie ein Graustufenbild in Farbe aussehen würde. Es wäre faszinierend zu sehen wie die Welt von damals in Farbe aussehen würde, etwas dass mich seit jeher fasziniert hat. Ein Graustufenbild kann interaktiv von einem Menschen gefärbt werden, sodass Farben möglichst akkurat vergeben werden. Wenn jedoch mehrere Tausend Bilder zu bearbeiten sind, würde das einige Zeit in Anspruch nehmen. Dieses Problem kann mit Deep Learning gelöst werden, indem ein Algorithmus selbstständig und möglichst realistisch ein Graustufenbild färbt. Der Prozess der Einfärbung eines Bildes ist ein aktives Forschungsgebiet im Deep Learning. Es gibt bereits Methoden, basierend auf Convolutional Neuronal Networks, die sehr realistische Ergebnisse liefern.

### **1.2 Zielsetzung**

Das Ziel dieser Arbeit ist die bestehenden Methoden zu untersuchen und zu vergleichen. Es wird eine Methode implementiert, um die Ergebnisse mit den bestehenden Methoden zu vergleichen. Der Fokus wird auf Methoden, die das Problem von Image Colorization

als ein Multimodales Problem behandeln, gelegt und implementiert auf Grund dessen Klassifikationsmethoden.

### 1.3 Vorgehensweise und Aufbau der Arbeit

Die vorliegende Arbeit lässt sich in fünf Hauptkapitel aufteilen. Zu Beginn wird eine ausführliche Erläuterung der Grundlagen gegeben, um die Methoden und Techniken der Arbeit zu verstehen. Anschließend werden die Konzepte, Techniken und Methoden präsentiert. Um auf die bevorzugte Methode aufbauen zu können werden einige Datensätze gebraucht, diese werden bei den Konzepten präsentiert. Nachdem auch diese erläutert wurden, wird die Implementierung erklärt. Darauf folgend werden zahlreiche Tests mit verschiedenen Hyperparametern durchgeführt. Abschließend werden die Tests evaluiert und mit anderen Methoden verglichen.

# Kapitel 2

## Grundlagen

Dieses Kapitel verschafft einen Überblick über die benötigten theoretische Grundlagen, um die Methoden dieser Arbeit zu verstehen. Zunächst wird eine Einführung in Neuronale Netzwerke gegeben, anschließend werden einzelne Bestandteile und Varianten von Neuronalen Netzwerken erklärt. Als nächstes wird der “Lab-Farbraum” kurz erläutert. Abschließend wird ein Überblick über verwandte Arbeiten gegeben.

### 2.1 Neuronale Netze

Künstliche Neuronale Netze sind vom menschlichen Gehirn inspiriert und werden für Künstliche Intelligenz und Maschinelles Lernen angewendet. Genutzt werden sie für überwachtes und unüberwachtes lernen. In der vorliegenden Arbeit werden nur Methoden des überwachten lernens angewendet. Beim überwachten lernen sind die Datensätze gelabelt, sodass der Output des Neuronalen Netzes mit den richtigen Ergebnissen verglichen werden kann.

Neuronale Netze bestehen aus Neuronen, auch “Units” genannt, die schichtenweise in “Layers” (Schichten) angeordnet sind. Beginnend mit der Eingabeschicht (Input Layer) fließen Informationen über eine oder mehrere Zwischenschichten (Hidden Layers) bis hin zur Ausgabeschicht (Output Layer). Dabei ist der Output des einen Neurons der Input des nächsten. [Moe18]

### 2.1.1 Feedforward Neural Network

Das Ziel von einem Feedforward Neural Network ist die Annäherung an eine Funktion  $f^*$ . Ein Feedforward Neural Network definiert eine Abbildung  $y = f(x; W)$  wobei  $x$  der Input ist und  $W$  die lernbaren Parameter sind (auch Gewichte genannt). [GBC16, S. 164-223]

Diese Netzwerkarchitektur trägt den Namen “feedforward” weil der Informationsfluss von dem Input Layer, über die Hidden Layers bis zum Output Layer in eine Richtung weitergereicht wird.

Feedforward Neural Networks werden als eine Kette von Funktionen dargestellt. So, kann man die Funktionen  $f^{(1)}, f^{(2)}, f^{(3)}$  in Form einer Kette verbinden, um  $f(\mathbf{x}) = f^{(3)}(f^{(2)}(f^{(1)}(\mathbf{x})))$  zu bekommen. Diese Kettenstrukturen sind die am häufigsten genutzte Struktur bei Neuronalen Netzwerken. Im genannten Beispiel ist  $f^{(1)}$  die erste Layer,  $f^{(2)}$  die zweite und  $f^{(3)}$  die Output Layer von diesem Netzwerk. Die Länge dieser Kette definiert die Tiefe des Netzwerks. Je tiefer ein Netzwerk ist, desto mehr erlernbare Parameter besitzt es und braucht somit eine erhöhte Rechenleistung, um trainiert zu werden. In der Praxis sind die Netzwerke sehr tief, daher der Begriff Deep Learning.

Während dem Training werden die Gewichte von  $f(x)$  verstellt, um  $f^*(x)$  zu erhalten. Jedes Trainingsbeispiel  $x$  ist mit einem Label  $y = f^*(x)$  versehen. Die Trainingsbeispiele geben genau vor, was die Output Layer generieren soll. Die Output Layer soll Werte generieren, die nah an  $y$  liegen. Das Verhalten der Hidden Layers wird nicht durch die Trainingsbeispiele festgelegt, sondern der Lernalgorithmus soll selbst definieren, wie diese Layers verwendet werden, um die beste Annäherung von  $f^*(x)$  zu erzielen.

### 2.1.2 Fully-connected Neural Network

Fully-connected Neural Networks sind die am häufigsten vorkommende Art von Neuronalen Netzen. In dieser Netzwerkarchitektur sind alle Neuronen eines Layers mit allen Neuronen des vorherigen und des nächsten Layers verbunden. Neuronen die sich im selben Layer befinden, sind jedoch nicht miteinander verbunden. [Fei17a]

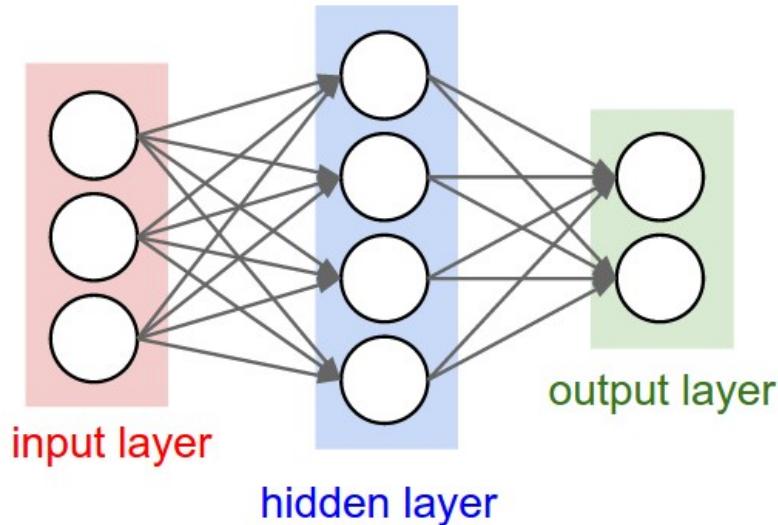


Abbildung 2.1: Fully-connected Neural Network mit 2 Layers (eine Hidden Layer mit 4 Neuronen und eine Output Layer mit 2 Neuronen) [Fei20a]

Einer der wichtigsten Gründe für die Anordnung von Neuronalen Netzen in Layers ist, dass so eine Struktur anhand von Matrix Multiplikationen berechnet werden kann. Die Abbildung 2.1 stellt ein Netzwerk mit 3 Inputs  $x$ , einer Hidden Layer mit 4 Neuronen und einer Output Layer mit 2 Neuronen dar. Die Kreise repräsentieren die Neuronen und einen Bias Wert  $b$ , die Pfeile stellen die Gewichte  $w$  dar.

$$f(x) = w * x + b \quad (2.1)$$

Nach jedem Hidden Layer läuft der Output durch eine Aktivierungsfunktion  $\sigma$  die unter Kapitel 2.1.3 erklärt wird. Dadurch wird die vorherige Formel um  $\sigma$  erweitert:

$$f(x) = \sigma(w * x + b) \quad (2.2)$$

### Forward Pass

Der Forward Pass von einem Neuronalen Netz wird anhand von Matrizen Multiplikationen berechnet. Um dies zu veranschaulichen wird es anhand eines Beispiels erklärt.

Ausgehend von einem Netzwerk mit 3 Inputs, einer Hidden Layer mit 2 Neuronen und einem Output Neuron, ergeben sich folgende Beispielwerte:

$$X = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix} \quad W = \begin{pmatrix} 10 & 20 \\ -20 & -40 \\ 20 & 0 \\ -40 & 0 \end{pmatrix} \quad W_{out} = \begin{pmatrix} 20 \\ 40 \\ -40 \end{pmatrix} \quad (2.3)$$

$X$  sind die Inputs,  $W$  die Gewichte des Hidden Layers und  $W_{out}$  die Gewichte des Output Layers. Die erste Spalte aus dem Input  $X$  und die ersten Zeilen aus beide Gewichtsmatrizen  $W$  und  $W_{out}$  sind die Werte für den Bias. Diese Anordnung des Bias Wertes ermöglicht die Berechnung durch eine einzige Matrix Multiplikation. Als Aktivierungsfunktion wird ReLU [NH10] verwendet:

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases} \quad (2.4)$$

Im ersten Schritt durchläuft der Input die Hidden Layer  $f(X \times W)$ :

$$f \left( \begin{pmatrix} 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix} \times \begin{pmatrix} 10 & 20 \\ -20 & -40 \\ 20 & 0 \\ -40 & 0 \end{pmatrix} \right) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{pmatrix} \quad (2.5)$$

Im zweiten Schritt wird der Output von der vorherigen Multiplikation mit den Gewichten des Output Layers multipliziert:

$$f \left( \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \times \begin{pmatrix} 20 \\ 40 \\ -40 \end{pmatrix} \right) = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad (2.6)$$

### 2.1.3 Aktivierungsfunktionen

Eine Aktivierungsfunktion definiert die Aktivierungsrate von einem Neuron. Es gibt verschiedene Aktivierungsfunktionen:

#### Sigmoid

Sigmoid ist eine nicht lineare Funktion welche die Werte in einem Wertebereich von  $[0, 1]$  bringt. Große negative Werte werden annähernd 0 und große positive Werte werden annähernd 1. Sigmoid hat einige Nachteile, so neigt es dazu den Gradienten verschwinden zu lassen und die Outputs sind nicht Null zentriert. Sigmoid wird definiert als:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (2.7)$$

wobei  $x$  ein Input ist.

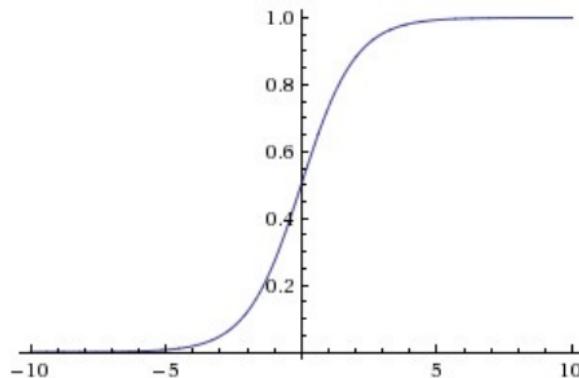


Abbildung 2.2: Sigmoid Aktivierungsfunktion [Fei20b]

#### Tanh

Die Tanh Aktivierungsfunktion bringt Werte in einen Wertebereich von  $[-1, 1]$ . Es ist eine skalierte Sigmoid ( $\sigma$ ) Funktion,  $tanh(x) = 2\sigma(2x) - 1$ . Die Nachteile von Tanh ähneln den von Sigmoid, wobei jedoch der Output Null zentriert ist.

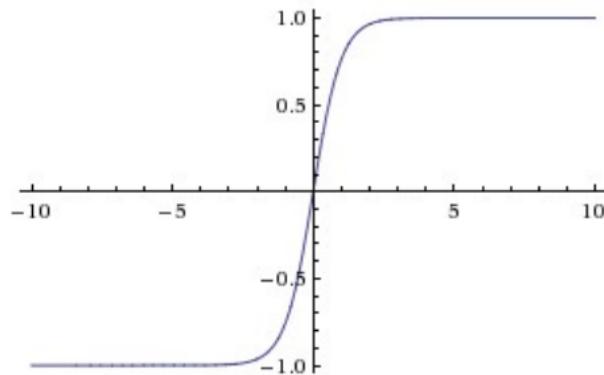


Abbildung 2.3: Tanh Aktivierungsfunktion [Fei20b]

## ReLU

Die Rectified Linear Unit konvertiert alle negativen Werte zu 0 und alle positiven Werte behalten ihre Identität. Diese Aktivierungsfunktion wurde für die Netzwerke in dieser Arbeit verwendet da, sie Vorteile gegenüber Sigmoid zeigt. Einer der Vorteile ist, dass die mathematische Auswertung der Funktion unkompliziert ist. Außerdem beschleunigt sie die Konvergenz des Stochastischen Gradientenabstiegsverfahrens im Vergleich zu Sigmoid. ReLU ist definiert als:

$$f(x) = \max(0, x) \quad (2.8)$$

wobei  $x$  ein Input ist.

Neuronen die ReLU als Aktivierungsfunktion verwenden können während des Trainings “sterben”. Zum Beispiel, wenn der Gradient in einem Neuron zu groß ist, kann dieser zu einem Update der Gewichte führen, wodurch das Neuron nie wieder aktiviert werden kann. Mit einer korrekten Einstellung der Lernrate kann das vermieden werden. [Fei17a]

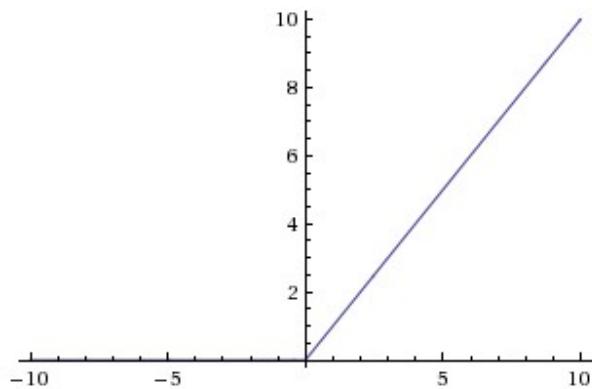


Abbildung 2.4: Rectified Linear Unit (ReLU) [Fei20b]

### Leaky ReLU

Leaky ReLU ist eine Variante der ReLU Aktivierungsfunktion, die versucht das Problem der “sterbenden” Neuronen zu minimieren. Anstatt alle negativen Werte zu Null zu konvertieren, werden die Werte mit einer Konstanten multipliziert. Die Funktion wird zu  $f(x) = 1(x < 0)(\alpha x) + 1(x \geq 0)(x)$ , wobei  $\alpha$  eine Konstante mit geringerem Wert ist, zum Beispiel 0.001.

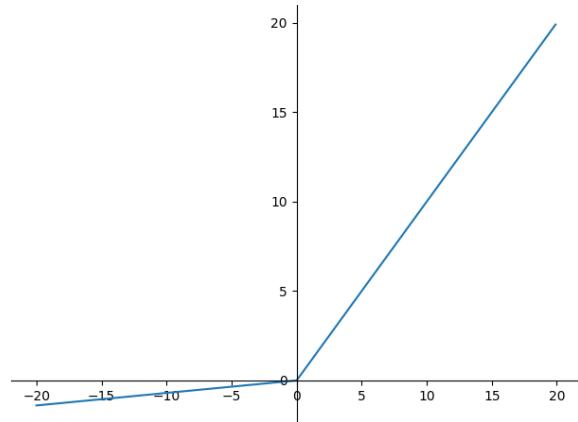


Abbildung 2.5: Leaky ReLU [ccs20]

### 2.1.4 Loss Functions

Die Loss Function (Kostenfunktion) dient zur Feststellung der Fehler (Loss) zwischen dem Output von einem Modell und dem vorgesehenen Zielwerten. Das Ziel von Neuronalen Netzen ist es den Loss zu minimieren. Wenn der Loss gleich Null ist, heißt dass  $y = \hat{y}$ . Es gibt verschiedene Arten von Loss Functions. Im Rahmen dieser Arbeit werden Loss Functions bezogen auf Regressions- und Klassifizierungsprobleme behandelt.

#### Mean Square Error Loss

Mean Square Error (MSE) Loss misst den mittleren quadratischen Fehler und ist definiert als:

$$J = \frac{1}{N} \sum (y - \hat{y})^2 \quad (2.9)$$

wobei,  $J$  der Loss ist,  $N$  die Anzahl der Klassen,  $y$  die korrekte Klasse (Ground Truth) und  $\hat{y}$  die vorhergesagte Klasse ist.

#### Cross Entropy Loss

Der Cross Entropy Loss wird bei Klassifizierungsproblemen verwendet. Es wird unterschieden zwischen, Binär und Multiclass Cross Entropy Loss. Bei dem Multiclass Cross Entropy Loss wird ein Vektor mit einer Wahrscheinlichkeitsverteilung  $x \in [0, 1]$  ausgewertet, wenn die korrekte Klasse eine 1 besitzt ist der Loss 0. Dabei gilt: je weniger Wahrscheinlichkeit die korrekte Klasse besitzt, desto höher wird der Loss sein. Der Multiclass Cross Entropy Loss ist definiert als:

$$J = -\frac{1}{N} \left( \sum_{i=0}^N y_i * \log(\hat{y}_i) \right) \quad (2.10)$$

wobei,  $J$  der Loss ist,  $N$  die Anzahl der Klassen,  $y$  die Korrekte Klasse (Ground Truth) und  $\hat{y}$  die vorhergesagte Klasse.

### Weighted Cross Entropy Loss

Bei dem Weighted Cross Entropy Loss werden die Klassen gewichtet bevor der Loss berechnet wird. Das ist zum Beispiel nützlich, um Klassen mit einer niedrigen Wahrscheinlichkeit zu bevorzugen.

### 2.1.5 Optimierungsalgorithmen

Das Ziel von Optimierungsalgorithmen ist eine Kombination von Gewichten  $W$  zu finden, die die Loss Function minimieren. Es gibt diverse relevante Optimierungsalgorithmen. In der vorliegenden Arbeit werden Gradient Descent und Adam verwendet.

#### Gradient Descent

Gradient Descent (Gradientenabstiegsverfahren) ist ein iteratives Verfahren, um bei einer Funktion das Minimum (oder das Maximum) zu finden. Mit Hilfe von partiellen Ableitungen kann der Gradient von einer Funktion berechnet werden. Ein Gradient ist, im Fall von Neuronalen Netzen, ein Vektor, der zum höchsten Punkt der Loss Function zeigt. Wird der negative Gradient genommen, zeigt dieser zum tiefsten Punkt. Bei jeder Kombination von Gewichten wird der Gradient berechnet und mit einer bestimmten Lernrate  $\alpha$  multipliziert, anschließend werden alle Gewichte aktualisiert. Die Lernrate definiert die Größe der Schritte in Richtung Minimum. Die Update Regel für die Gewichte ist definiert als:

$$w_{x+1} = w_x - \alpha * \nabla J(w_x) \quad (2.11)$$

wobei,  $w_{x+1}$  die aktualisierten Gewichte sind,  $w_x$  die vorherigen Gewichte,  $\alpha$  die Lernrate und  $\nabla J(w_x)$  der Gradient. Die Update Regel für den Bias sieht identisch aus.

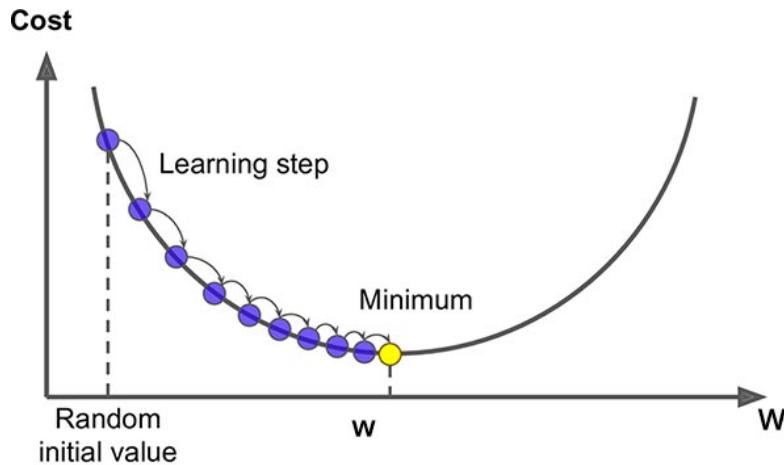


Abbildung 2.6: Gradient descent visualisiert [Bha18]

Es gibt verschiedene Arten von Gradient Descent: Gradient Descent, Mini-Batch Gradient Descent und Stochastic Gradient Descent. Beim normalen Gradient Descent werden die Gradienten im Bezug zu dem gesamten Datensatz berechnet und damit das Update durchgeführt. Der Mini-Batch Gradient Descent berechnet die Gradienten im Bezug zu einem kleinen Teil des Datensatzes und führt ein Update für alle Parameter durch. Der Stochastic Gradient Descent berechnet den Gradient bezogen auf ein einziges Element des Datensatzes und führt einen Update für alle Parameter durch.

Um die Konvergenz Richtung Minimum zu beschleunigen wurde der Gradient Descent mit Momentum entwickelt. Bei diesem Ansatz wird ein Geschwindigkeitsparameter zu der Update Regel hinzugefügt, der alle vorherigen Updates akkumuliert. Das ermöglicht die schnellere Konvergenz mit jedem Schritt. Die neue Update Regel ist definiert als:

$$\begin{aligned} v_{t+1} &= \rho v_t - \alpha * \nabla J(w_x) \\ w_{x+1} &= w_x + v_{t+1} \end{aligned} \tag{2.12}$$

wobei  $v_{t+1}$  der nächste Geschwindigkeitsparameter ist,  $v_t$  der aktuelle Geschwindigkeitsparameter und  $\rho$  ein Reibungsparameter (typisch 0.9) zur Regulierung.

### Adam

Adam steht für “Adaptive Moment Estimation Algorithm” und ist ein Optimierungsalgorithmus, der eine angepasste Lernrate für die verschiedenen Parameter berechnet [KB14].

Adam ist der bevorzugte Optimierungsalgorithmus für die vorliegende Arbeit. Adam kombiniert die Ansätze von AdaGrad [DHS11] und RMSProp. AdaGrad ist eine verbesserte Version von Gradient Descent, der eine angepasste Lernrate für die verschiedenen Parameter einführt.

### 2.1.6 Backpropagation

Neuronale Netze lernen indem der Loss minimiert wird. Wie in der vorherigen Sektionen erläutert, bestimmt die Loss Function die Fehlerrate von einem Neuronalen Netz. Dieser Loss kann mit Hilfe von einem Optimierungsalgorithmus reduziert werden. Backpropagation ermöglicht eine effiziente Berechnung der Gradienten in einem neuronalen Netzwerk [15]. Mit Hilfe der Kettenregel kann eine komplexe Loss Function in kleinere Unterfunktionen zerlegt werden, um Lokal die Ableitung zu berechnen. Das ermöglicht eine unkomplizierte Berechnung des Gradienten.

Als Beispiel wird die folgende Sigmoid Funktion in Unterfunktionen zerlegt:

$$f(w, x) = \frac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2)}} \quad (2.13)$$

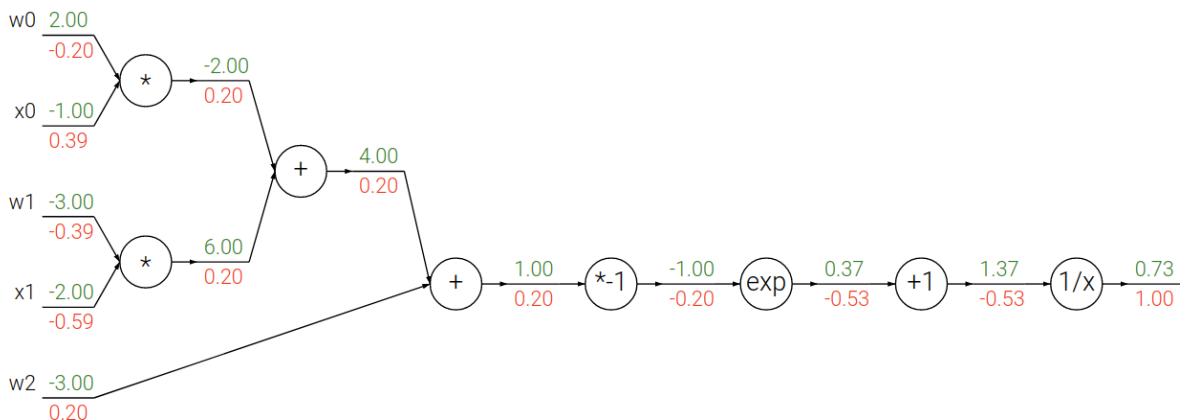


Abbildung 2.7: Backpropagation Beispiel anhand einer 2D Neuron mit der Aktivierungsfunktion Sigmoid [Fei20c]

Auf der Abbildung 2.7 stellen  $[w_0, w_1, w_2]$  die Gewichte und  $[x_0, x_1]$  die Inputs des Neurons dar. Um es unkompliziert zu halten wird die obere Funktion als eine beliebige Funktion,

die Inputs  $(w, x)$  entgegennimmt und eine einzelne Zahl als Output hat, visualisiert werden. Die grünen Zahlen repräsentieren die Ergebnisse aus dem Forward Pass und die roten Zahlen den zurück propagierten Loss. Jeder Knoten ist fähig ein Output und der lokale Gradient von dem Output im Bezug auf den Input zu berechnen, ohne die komplette Funktion kennen zu müssen [Fei17b].

## 2.2 Convolutional Neural Networks

Convolutional Neural Networks (CNN) sind eine besondere Form von künstlichen neuronalen Netzwerken, die unter anderen speziell für die Verarbeitung von Bildern vorgesehen sind [Dip19].

Im Gegensatz zu traditionellen neuronalen Netzwerken, die einen Vektor als Input nehmen, nehmen Convolutional Neural Networks ein 3D Volumen als Input ( $W \times H \times C$ , hierbei ist  $W$  die Breite,  $H$  die Höhe und  $C$  sind die Farbkanäle).

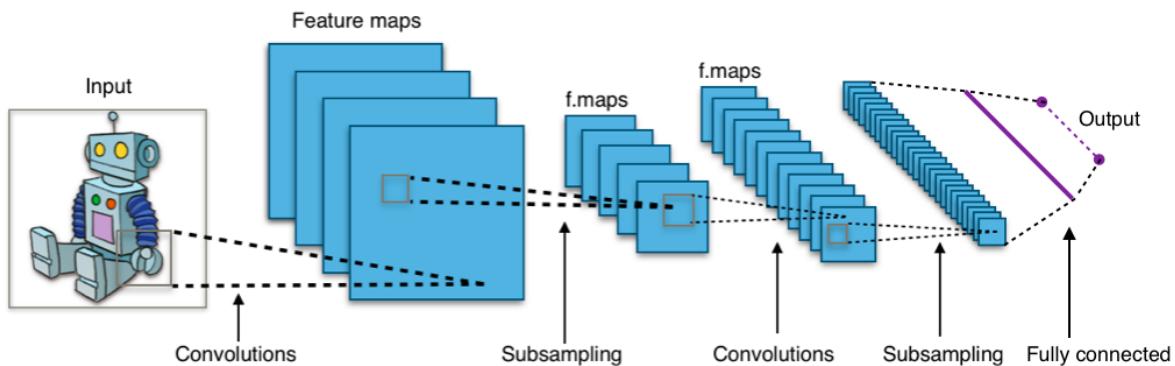


Abbildung 2.8: Typische Struktur von einem Convolutional Neural Network [Com15]

CNNs bestehen in der Regel aus 2 Formen von Layers, der Convolutional Layer und der Pooling Layer.

Die Convolutional Layer besteht aus mehreren hintereinander geschalteten 3 dimensionalen Filtern, auch Kernel genannt ( $W \times H \times D$ , wobei  $D$  die Tiefe der Feature Maps darstellt), die während dem Forward pass mit einer festgelegten Schrittweite (Stride), über das Bild geschoben werden. Mit dem sogenannten Padding wird das Verhalten an den Rändern festgelegt. An jeder Stelle wird eine Matrix Multiplikation zwischen den Filter und die aktuelle Position auf dem Bild durchgeführt. Als Output wird eine 2 dimensionale Feature Map generiert. Die Größe dieser Feature Map ist abhängig von der Größe des

Filters, dem Padding und vor allem dem Stride. Ein Stride von 2 bei einer Filter Größe von  $2 \times 2$  führt beispielsweise pro Filter zu einer Halbierung der Größe der Feature Map im Vergleich zum Input Volumen [Bec19]. Ein Stride von 1 bei einem  $3 \times 3$  Filter mit Padding 1 führt zu einer Feature Map mit der gleichen Größe wie dem Input Volumen.

Die Filter erkennen in den ersten Ebenen einfache Strukturen wie Linien, Farben oder Kanten. In den daran folgenden Ebenen lernt das CNN Kombinationen aus diesen Strukturen wie Formen oder Kurven. In den tieferen Layers werden komplexere Strukturen und Objekte identifiziert.

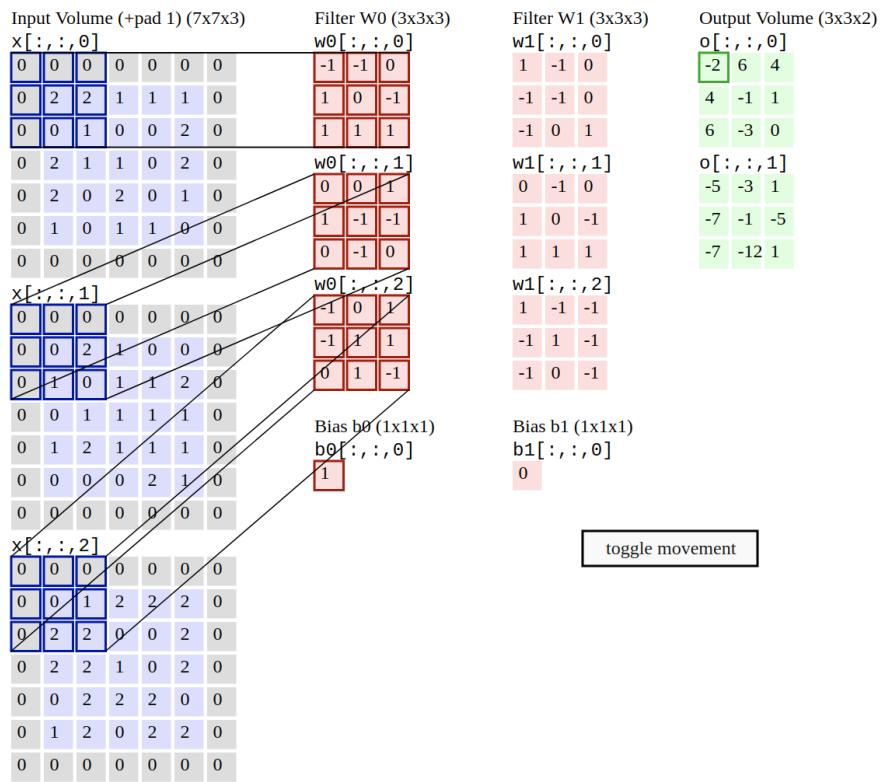


Abbildung 2.9: Beispiel eines Forward pass von einer Convolutional Layer mit einem  $7 \times 7 \times 3$  Input Volumen, zwei  $3 \times 3 \times 3$  Filtern, Padding 1 und Stride 2. [Fei20d]

Die Pooling Layer dient zur Reduktion der Dimensionen von einem Input Volumen und somit den Parametern vom Netzwerk. Es gibt verschiedene Pooling Operationen die angewendet werden können, wie zum Beispiel Maximum Pooling, Minimum Pooling, oder Average Pooling. Im Rahmen dieser Arbeit wird Maximum Pooling (auch Max Pooling genannt) verwendet.

Eine Max Pooling Layer aggregiert die Aktivierungsmatrizen von Convolutional Layers in dem nur die höchste Zahl eines Filters weitergegeben wird. So wird bei einem  $2 \times 2$  Filter von 4 Zahlen nur eine Zahl weitergegeben. Damit wird einer Reduktion der Dimensionen erreicht.

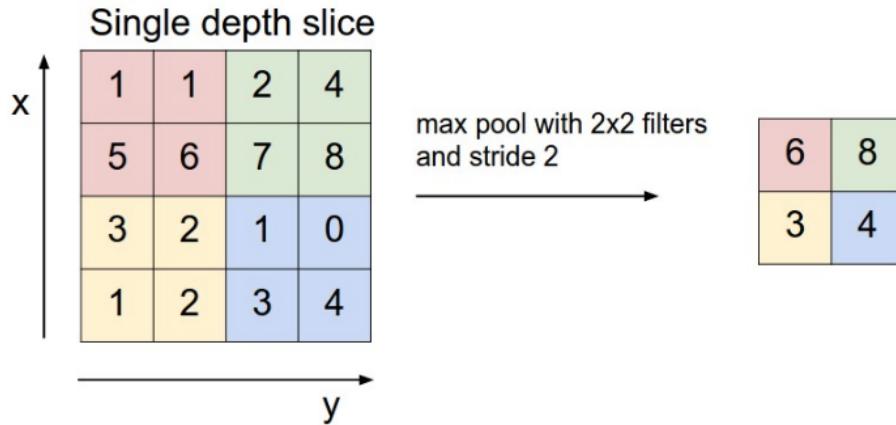


Abbildung 2.10: Max pooling Operation mit  $2 \times 2$  Filtern und Stride 2 [Fei20d]

### 2.2.1 Transposed Convolution

Im Gegensatz zu einer Pooling Layer ermöglicht eine Transposed Convolutional Layer, die Dimensionen von einem Volumen zu vergrößern. Die funktionsweise einer Transposed Convolution wird anhand von einem Beispiel erklärt.

Ausgehend von einer  $2 \times 2$  Input Matrix, die auf  $3 \times 3$  vergrößert werden soll, ein  $2 \times 2$  Filter, Null Padding und Stride 1, ergibt sich der folgenden Output.

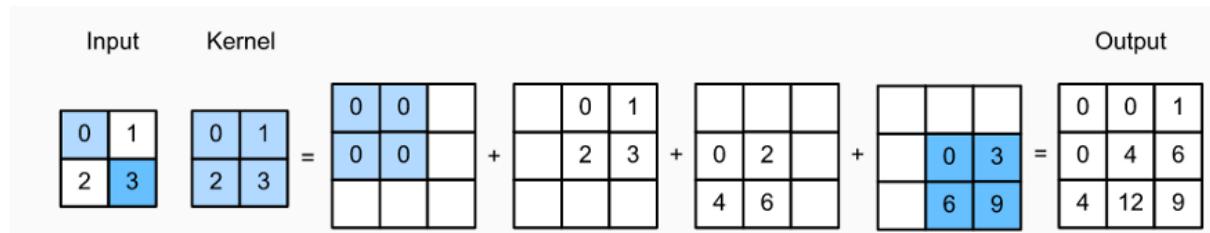


Abbildung 2.11: Die komplette Transposed Convolution Operation [Zha20]

Jede Zahl in der Input Matrix wird mit jeder Zahl in den Filtern multipliziert. Daraus ergibt sich an jeder Position in der Input Matrix einer  $2 \times 2$  Matrix. Die sich überlappenden

Zahlen auf der Output Matrix werden addiert. Daraus ergibt sich eine  $3 \times 3$  Output Matrix.

## 2.3 Autoencoder

Ein Autoencoder ist ein neuronales Netz, welches versucht die Eingangsinformationen zu komprimieren und mit den reduzierten Informationen im Ausgang wieder korrekt nachzubilden [Ngu20]. Die Komprimierung und Rekonstruktion des Inputs läuft in zwei Schritten ab, weshalb das Netz in zwei Teile betrachtet werden kann.

### Encoder

Der Encoder reduziert die Dimensionen von einem Input und somit werden die wichtigsten Features in einer reduzierter Dimension komprimiert. In einem neuronalen Netz wird diese Komprimierung durch die Hidden Layers erreicht. Der Encoder ist definiert als:

$$h = f(x) \quad (2.14)$$

wobei  $x$  ein Input ist,  $f$  der Encoder und  $h$  die Kodierten Features von  $x$ .

### Decoder

Der Decoder ist zuständig für die Rekonstruktion von  $x$  anhand  $h$  und ist definiert als:

$$\hat{x} = g(h) \quad (2.15)$$

wobei,  $\hat{x}$  der rekonstruierte Input ist,  $g$  der Decoder und  $h$  die Kodierten Features.

In der vorliegenden Arbeit wird eine Variante von Autoencoder, basierend auf Convolutional Neural Networks, eingesetzt.

## 2.4 Lab-Farbraum

Der *Lab*-Farbraum (auch CIELAB-Farbraum genannt) ist ein Farbraum definiert bei der internationalen Beleuchtungskommission (CIE) in 1976. Farben werden mit drei Werten beschrieben. „*L*“ (Lightness) definiert die Helligkeit. Die Werte liegen zwischen 0 und 100. „*a*“ gibt die Farbart und Farbintensität zwischen Grün und Rot an und „*b*“ gibt die Farbart und Farbintensität zwischen Blau und Gelb wieder. Die Werte für „*a*“ und „*b*“ liegen zwischen -128 und 127.

In der vorliegenden Arbeit wird der *Lab*-Farbraum verwendet, da es unkompliziert ist, den „*L*“ Kanal von beiden Farbkanälen „*a*“ und „*b*“ zu trennen. Außerdem bildet der *Lab*-Farbraum das menschliche Sehvermögen besser ab, als der RGB-Farbraum<sup>1</sup>.

## 2.5 Image Colorization Methoden

Der Prozess von Image Colorization kann manuell oder automatisch erfolgen. Zu den manuellen Methoden zählen die analoge Einfärbung eines Bildes durch einen Menschen bis hin zur digitalen Bearbeitung mit einem dafür vorgesehenem Programm. Für die automatischen Methoden wird öfters menschlicher Input benötigt, um zum Beispiel Bereiche von einem Graustufenbild mit Farbstichen zu markieren, die dann automatisch von einem Algorithmus über das Bild propagiert werden. Aktuelle automatische Methoden nutzen die Vorteile von Convolutional Neural Networks, um diesen Prozess effizienter und performanter zu gestalten.

Um ein CNN, der Bilder automatisch einfärbt, zu trainieren, werden das Original Bild und das Graustufenbild benötigt. Das Graustufenbild wird in den CNN eingespeist und dieser wird versuchen die Farbkanal Werte vorherzusagen. Anschließend werden die Werte von jedem Pixel mit jedem Pixel aus dem Original Bild verglichen. Dieser Prozess wird iterativ wiederholt, bis die erzeugten Werten einen niedrigen Loss Wert haben. In den nächsten Kapiteln wird dieser Vorgehensweise näher erläutert.

---

<sup>1</sup>RGB steht für Red, Green und Blue, die 3 Farbkanäle des Farbraums

## 2.6 Verwandte Arbeiten

Vor der Erstellung dieser Arbeit wurden zahlreiche automatische Methoden von Image Colorization bereits untersucht. Frühere Methoden waren stark an menschliches Input gebunden. Die Methode von Levin et al. verwendet Farbstiche auf dem Graustufenbild, die automatisch von einem Algorithmus über das gesamte Bild propagiert werden [LLW04].

Der Fokus dieser Arbeit ist auf voll automatische Image Colorization Methoden gesetzt. Konservative Methoden, die Convolutional Neural Networks verwenden, versuchen die Farben von dem originalen Bild wiederherzustellen, indem die Loss Function die Distanz der vorhergesagten Farben zu den realen Farben berechnet. Diese Methoden liefern in der Regel entsättigte und blasses Bilder wie bei [Özb19]. Einer der Gründe für diese Ergebnisse ist, dass die Modelle nicht richtig lernen. So können beispielsweise Äpfel verschiedene Farben wie Rot oder Grün haben. Wenn das Netzwerk mit einem MSE Loss trainiert wird und der Datensatz die gleiche oder annähernd gleiche Anzahl an grünen und roten Äpfeln hat, wird der Output bei einem Apfel eine Farbe zwischen Rot und Grün sein, was ein entsättigtes Bild erzeugen wird.

Aus diesem Grund betrachtet die vorliegende Arbeit das Problem von Image Colorization als ein Multimodales Problem, da gleiche Objekte verschiedene Farben einnehmen können. Die vorliegende Arbeit orientiert sich an die Methoden von Zhang et al. und Billaut et al., die ähnliche Ansätze für Image Colorization vorschlagen. Sie betrachten das Problem als ein Klassifizierungsproblem und trainieren eine CNN mit der Cross Entropy und Weighted Cross Entropy Loss. Der Output des Netzwerks ist eine Wahrscheinlichkeitsverteilung über die möglichen Bins für jeden Pixel, die im Kapitel 3.3 erläutert werden.

Beide Ansätze verwenden “Color Bins” die es unkompliziert ermöglichen die Farben von jedem Pixel zu klassifizieren. Es wird im nächsten Kapitel weiter auf diese Methode eingegangen.

# Kapitel 3

## Konzeption

Dieses Kapitel beschreibt alle notwendigen Schritte für die Konzeption der angewandten Methode, außerdem werden die Datensätze und die verwendeten Tools präsentiert.

### 3.1 Image Colorization als Multimodales Problem

Konventionelle automatische Methoden zielen darauf ab, die Farben für ein generiertes Bild so nah wie möglich an das originale Bild vorherzusagen. Diese Methoden verwenden ein MSE Loss, das Vorhersagen, die weit von den originalen Farbwerten entfernt liegen, stärker bestraft, als Farbwerte die dichter an den originalen Farbwerten liegen. Das führt, wie bei 2.6 beschrieben zu entsättigten Bildern. Die Gründe für diese Ergebnisse lassen sich dadurch erklären dass verschiedene Objekte verschiedene Farben besitzen können.

Die gewählte Methode dieser Arbeit berücksichtigt die multimodaleität von Image Colorization und behandelt das Problem als Klassifikation.

### 3.2 Farbraum

Der Standard Farbraum der Bilder für die Methode dieser Arbeit ist der RGB<sup>1</sup>-Farbraum. Der RGB-Farbraum hat einige Nachteile, die es kompliziert machen mit diesem Farbraum zu arbeiten. Einer der Nachteile ist, dass das Graustufenbild sich schwer von den Farbkanälen trennen lässt. Ein anderer Nachteil ist, dass die Farbinformationen in drei Farbkanäle

---

<sup>1</sup>Rot, Grün, Blau

kodiert sind, was die Komplexität eines Modells erhöht. Aus diesem Grund werden die Bilder für das Preprocessing und die Methode in den Lab-Farbraum konvertiert. Für die Darstellung der Ergebnisse werden die Bilder von dem Lab-Farbraum in den RGB-Farbraum umgewandelt.

Im Lab-Farbraum können die Farbkanäle “ab” problemlos vom Belichtungskanal “L” getrennt werden. Der Belichtungskanal “L” enthält das Graustufenbild, das in den CNN eingespeist wird.



Abbildung 3.1: Original Bild in RGB oben links, Belichtungskanal “L” oben rechts, Farbkanal “a” unten links und Farbkanal “b” unten rechts.

### 3.3 Binning

Binning ist eine Technik, die für die Bildverarbeitung verwendet wird. Binning wird, im Kontext von Image Colorization, als Eingruppierung von naheliegenden Farben definiert. Die Farben werden in gleich große Intervalle aufgeteilt. Diese Intervalle bezeichnet man im englischen als “Bins”. Jedes dieser Intervalle wird durch einen Bin Index repräsentiert, somit reduziert sich die Anzahl der Klassen die vorhergesagt werden können.

Als Beispiel für die Veranschaulichung wird der normalisierte Lab-Farbraum in 36 gleich große Bins unterteilt. Da die Farbinformationen in den “ab” Farbkanälen kodiert sind,

werden nur diese 2 Farbkanäle in Bins klassifiziert. Auf der Abbildung 3.2 ist der Farbkanal "a" auf der x-Achse und der Farbkanal "b" auf der y-Achse abgebildet. Die Quadrate repräsentieren die Bins. Die obere Zahl in den Bins symbolisiert die  $xy$  Koordinaten auf dem Grid, die untere, unterstrichene Zahl symbolisiert den Bin Index. Die  $xy$  Koordinaten sind Bedeutsam für die Berechnung der Bins.

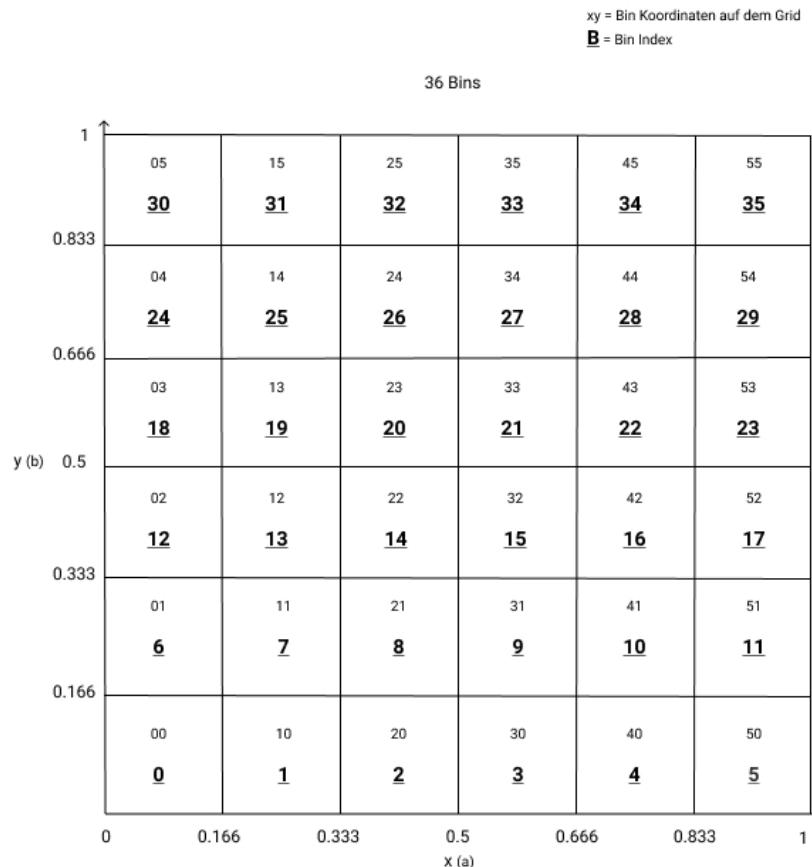


Abbildung 3.2: Grid mit 36 bins. Die x-Achse bildet die Werte des Farbkanals "a" und die y-Achse die Werte des Farbkanals "b" ab.

Für die Methoden dieser Arbeit wurden nur symmetrische Grids verwendet, so ergibt zum Beispiel ein  $6 \times 6$  Grid 36 Bins und ein  $18 \times 18$  Grid 324 Bins.

Die Umwandlung von Bin zu Farbe muss für die Ergebnisse des Netzwerks ebenfalls vorgegeben sein. Für dieses Problem werden vor dem Training die Farben von jedem Pixel aus den Trainingsbildern in Bins klassifiziert. Jeder Bin wird durch eine Liste mit je zwei Listen, für den jeweiligen Farbkanal, repräsentiert. Jeder Pixelwert wird zur entsprechenden Bin Liste hinzugefügt. Anschließend wird der Modus und der Durchschnitt jedes Farbkanals, der unter jedem Bin klassifizierten Farben, ausgerechnet. Am Ende beinhaltet jeder Bin

eine Liste mit je zwei Werten, auf dem Nulltem Index der Wert für den Farbkanal “a” und auf dem ersten Index der Wert für den Farbkanal “b”. Es werden zwei separate Dateien erzeugt, eins für den Modus und eins für den Durchschnitt.

Für die Umwandlung von Bin zu Farbe werden die jeweiligen Werte für jeden Farbkanal eines Bins von der jeweiligen Datei abgeguckt. Es kann zwischen dem Durchschnitt und dem Modus ausgewählt werden. Der Durchschnitt verhält sich ähnlich wie ein Modell trainiert mit dem MSE Loss und der Modus würde Ergebnisse mit einem rötlichen Farbton liefern. Als Lösung für dieses Problem schlägt Zhang et al. den “annealed mean” vor [ZIE16]. Der “annealed mean” versucht einen Kompromiss zwischen dem Durchschnitt und dem Modus zu finden. Dieser Kompromiss wird durch einen Temperatur Parameter ( $T$ ) reguliert. In der vorliegenden Arbeit wird eine von den “annealed mean” inspirierten Methoden implementiert. Die Methode wird wie folgt implementiert:

$$\begin{aligned} D &= K_{mode} - K_{mean} \\ \hat{y}_K &= K_{mode} - (D * T) \end{aligned} \tag{3.1}$$

wobei  $K$  ein Farbkanal ist,  $D$  die Distanz zwischen Durchschnitt und Modus,  $T$  ein Temperaturwert zwischen 0 und 1 und  $\hat{y}_k$  die endgültige Farbe für den Farbkanal  $K$  ist. Ein Temperaturwert von 1 würde den Durchschnitt ergeben, wohingegen ein Temperaturwert von 0, den Modus nicht verändert.

## 3.4 Netzwerkarchitektur

Die Netzwerkarchitektur ist ein wichtiger Faktor der u.a. die Ergebnisse beeinflusst. Um die Methoden zu vergleichen ist es wichtig, ein leichtes Convolutional Neural Network, das wenige Parameter besitzt, schnell zu trainieren ist und gute Ergebnisse liefert.

Das Ziel von dem Netzwerk ist es, ein Graustufenbild als Input zu bekommen und eine Wahrscheinlichkeitsverteilung für alle Bins per Pixel vorherzusagen. Das Output Volumen hat die Dimensionen  $W_{Input} \times H_{Input} \times n\_bins$ , wobei  $W$  und  $H$  die Breite und Höhe des Bildes sind und  $n\_bins$  eine Wahrscheinlichkeitsverteilung für alle Bins pro Pixel ist. Dieser Ansatz wird auch bei Image Segmentation Probleme genutzt, wobei ein Bild in das Netzwerk eingespeist wird und eine Segmentation map, mit einer Klasse per Pixel, als Output erzeugt wird. In der Regel wird jeder Klasse eine bestimmte Farbe zugeordnet,

was die Objekte klassifiziert und trennt. In dem Fall von Image Colorization bekommt jedes Pixel in dem Output Volumen eine Wahrscheinlichkeitsverteilung für alle Bins, die in eine Farbe umgewandelt wird.

Die Methode von Zhang et al. [ZIE16] verwenden ein CNN, das ein Graustufenbild als Input entgegennimmt und ein Volumen mit einer Wahrscheinlichkeitsverteilung für alle Bins per Pixel generiert. Diese Netzwerkarchitektur besteht aus Blöcken mit jeweils 2 oder 3 Convolutional und ReLU Layers, gefolgt von einer Batch Normalization Layer. Batch Normalization ist eine Regularisierungstechnik, die die Werte in einer Hidden Layer normalisiert, bevor sie in die nächsten Layer weitergereicht werden. Das Netzwerk hat keine Pooling Layers, alle Änderungen in der Auflösung werden durch Downsampling oder Upsampling zwischen Blöcken erreicht.

Diese Netzwerkarchitektur ist sehr herausfordernd für die GPU<sup>2</sup>, was die Batch Größe und Trainingszeit stark beeinflusst.

Aus diesem Grund orientiert sich die Netzwerkarchitektur dieser Arbeit an der Netzwerkarchitektur von Billaut et al. [BRT18]. Sie verwenden eine angepasste Version von einem U-net Convolutional Neural Network [RFB15].

### 3.4.1 U-net

Ein U-net ist ein Autoencoder mit Skip Connections und Transposed Convolutions als Upsampling Methode das bei Image Segmentation angewendet wird. Im Vergleich zu konventionellen Autoencoder können der Encoder und Decoder nicht getrennt voneinander verwendet werden. Die Skip Connections ermöglichen fein-granuläre Details in dem Output Volumen wiederherzustellen und helfen mit dem Vanishing Gradient Problem während Backpropagation. Skip Connections konkatenieren bestimmte Layers aus dem Encoder mit Layers aus dem Decoder, mit den gleichen Dimensionen.

Ein U-net besteht, wie ein Autoencoder, aus einem Encoder und Decoder Teil. Der Encoder besteht aus sogenannten “ConvBlocks”. ConvBlocks bestehen aus 2 Convolutional Layers gefolgt von Batch Normalization und ReLU. Den ConvBlocks folgt eine Pooling Layer, die die Dimensionen des Volumens verringert. Der Decoder besteht aus ConvBlocks gefolgt von Transposed Convolutions, die die Dimensionen von dem Volumen wieder vergrößern.

---

<sup>2</sup>Graphics Processing Unit

Die letzte Layer ist ein  $1 \times 1$  Convolutional Layer, die das Output Volumen generiert. Skip Connections konkatenieren ConvBlocks aus dem Encoder mit Transposed Convolutions aus dem Decoder mit der gleichen Dimensionen.

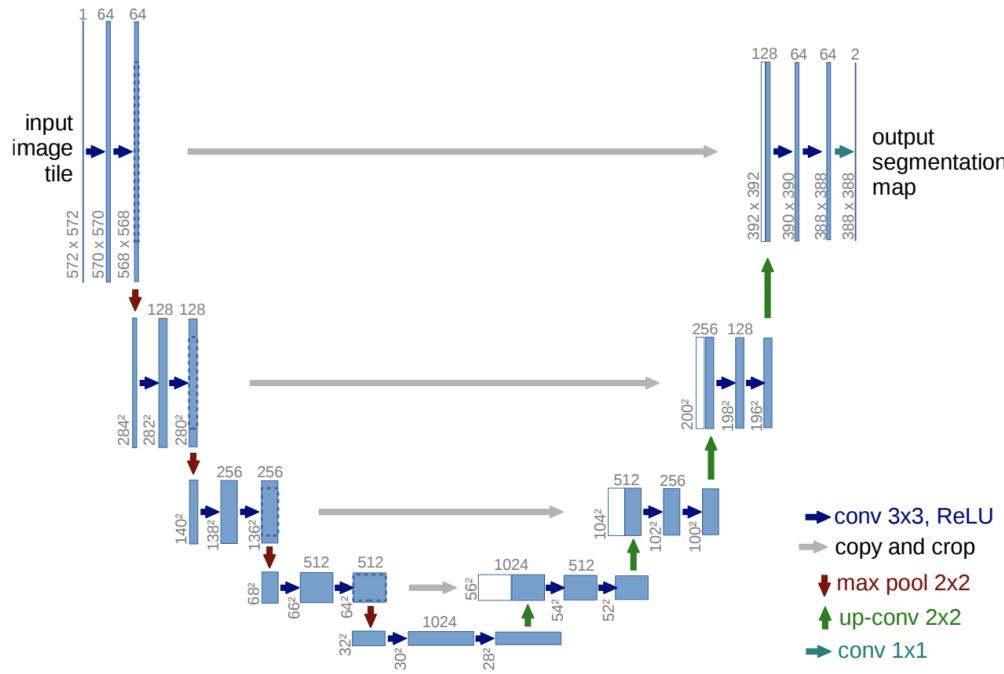


Abbildung 3.3: U-net Architektur (Beispiel für  $32 \times 32$  Pixels in der niedrigsten Auflösung). Jede blaue Box entspricht einer multi-Kanal Feature Map. Die Tiefe der Feature Maps ist gekennzeichnet durch die Zahl über der Box. Die Breite und Höhe ist durch die Zahl unten links erkennbar. Die weißen Boxen repräsentieren die kopierten Feature Maps. Die Pfeile bestimmen die verschiedenen Operationen. [RFB15]

### 3.5 Datensätze

Um das Netzwerk zu trainieren werden bedeutsame Bilder mit der gleiche Thematik gebraucht. Ein Vorteil von Image Colorization ist, dass jedes Bild für das trainieren verwendet werden kann, da nur die Graustufen Version davon gebraucht wird.

Um die Methode zu prüfen werden 3 Datensätze benutzt, die verschiedene Auflösungen und Themen beinhalten.

Als erstes wird ein Spiel-Datensatz von  $32 \times 32$  Bildern generiert. Dieser setzt sich aus 3 geometrischen Objekten und 3 Farben pro Objekt zusammen. Die geometrischen Objekte

sind ein Rechteck, ein Kreis und ein Dreieck pro Bild, die in jeweils eine der 3 Farben eingefärbt sind. Die Bilder haben einen einheitlichen schwarzen Hintergrund.

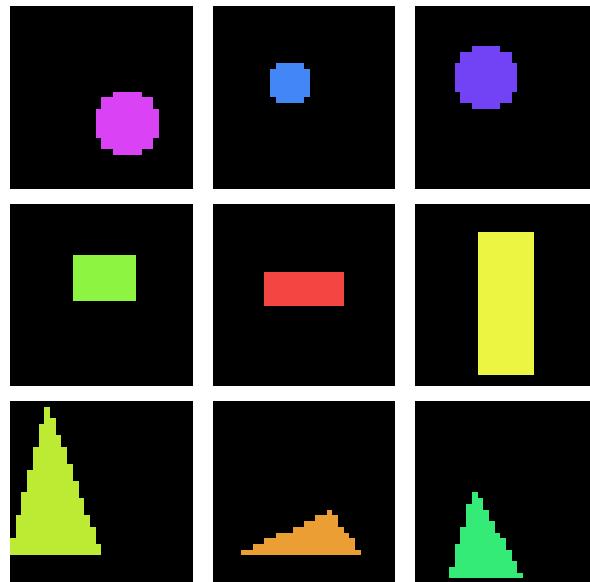


Abbildung 3.4: Beispiel von Trainingsbildern mit einer der möglichen Farbe pro Klasse

Der Datensatz besteht aus 2664 Bildern für das Training und 324 Bildern für die Validierung. Dieser Datensatz dient als Beweis für das funktionieren der Methode und hilft bei der Entwicklung und Optimierung des Netzwerks.

Als zweites werden 12 Klassen des CIFAR-100<sup>3</sup> Datensatzes verwendet. CIFAR-100 ist ein öffentlich verfügbarer Datensatz, der von Krizhevsky et al. erstellt wurde. Der Datensatz setzt sich aus 100 Klassen mit jeweils 600 Bildern pro Klasse zusammen, außerdem sind die 100 Klassen in 20 Superklassen gruppiert. Der Datensatz wird in 50000 Bildern für das Training und 10000 Bildern für die Validierung aufgeteilt. Einige Beispiele für Klassen sind *apples*, *palm* oder *bee*. Die Bilder haben ebenfalls einer Auflösung von  $32 \times 32$ .

Da es äußerst aufwändig wäre ein Modell, dass jedes Objekt aus dem CIFAR-100 Datensatz richtig erkennt und einfärbt, von Null auf zu trainieren, werden nur bestimmte Klassen für das Training verwendet. Diese Klassen sind: *apples*, *sunflower*, *rose*, *cloud*, *maple\_tree*, *oak\_tree*, *pine\_tree*, *willow\_tree*, *palm\_tree*, *mountain*, *forest* und *sea*. Alle Klassen haben Gemeinsamkeiten, was das Erlernen von Merkmalen erleichtert im Gegensatz

<sup>3</sup>CIFAR-100: <https://www.cs.toronto.edu/~kriz/cifar.html>

zu einem sehr allgemeinen Datensatz.

Abschließend wird ein komplexerer und hochauflösender Datensatz verwendet, der Naturbezogene Bilder enthält. Ziel ist es, ein Netzwerk zu trainieren, dass Bilder aus der Natur einfärben kann. Dieses besteht aus 3 Datensätzen von Kaggle<sup>4</sup> und GitHub<sup>5</sup>. Der erste ist der “Landscape Pictures”<sup>6</sup> Datensatz von Arnaud Rougetet, von diesem Datensatz werden alle Bilder verwendet. Der zweite Datensatz ist “Landscape Classification”<sup>7</sup> von Huseyb Guliyev, hiervon werden nur die Klassen *forest*, *glacier*, *mountain* und *sea* verwendet. Das letzte Datensatz ist der “Landscapes dataset”<sup>8</sup> von ml5js auf GitHub. Von diesem Datensatz wurden die Klassen *field*, *forest*, *lake*, *mountain* und *road* verwendet. Desweiteren werden alle Bilder entfernt, die keine öffentliche Lizenz haben.

Der komplette Datensatz besteht aus 8 Klassen und hat insgesamt 12479 Bilder, 10120 für das Training und 2359 für das Testen. Die 8 Klassen sind: *field*, *forest*, *glacier*, *lake*, *mountain*, *road*, *sea* und “ohne Kategorie”. Die Klasse “ohne Kategorie” beinhaltet die Bilder aus dem “Landscape Pictures” Datensatz. Die Klassen sind für das Training und die Methode nicht relevant, da jedes Pixel in Bins klassifiziert wird. Der finale Datensatz wird für der Rest der Arbeit “Landscape Datensatz” genannt.

## 3.6 Image Preprocessing und Augmentation

Für das optimale Training und die beste Ergebnisse werden die Bilder vorverarbeitet. Außerdem werden Techniken von Image Augmentation angewendet. Der Spiel-Datensatz und die 12 Klassen von CIFAR-100 werden für das Training mit einer Wahrscheinlichkeit von 50% horizontal gespiegelt.

Da der Landscape Datensatz Bilder mit verschiedenen Auflösungen beinhaltet, werden alle Bilder auf  $128 \times 128$  angepasst. Das reduziert die Trainingszeit und die Komplexität des Datensatzes. Für das Training werden pro Bild 4 zusätzliche augmentierte Bilder generiert. Zunächst werden die Bilder zufälligerweise um  $\pm 30$  Grad rotiert, anschließend wird die Größe der Bilder auf einen Wert zwischen 0 und 30% geändert. Nachdem dieser Schritt abgeschlossen ist, werden die Bilder horizontal gespiegelt und abschließend nochmal

---

<sup>4</sup>Kaggle: <https://www.kaggle.com/>

<sup>5</sup>Github: <https://github.com/>

<sup>6</sup><https://www.kaggle.com/arnaud58/landscape-pictures>

<sup>7</sup><https://www.kaggle.com/huseynguliyev/landscape-classification?>

<sup>8</sup><https://github.com/ml5js/ml5-data-and-models/tree/master/datasets/images/landscapes>

Vertikal gespiegelt. Nach der Image Augmentation besteht der Trainings Datensatz aus 50600 Bildern.

## 3.7 Tools

Um die Methode zu realisieren werden einige Tools genutzt. Für die Implementierung wird das Framework PyTorch<sup>9</sup> verwendet. PyTorch ist ein Open-Source Framework basierend auf Python für Machine Learning und Deep Learning. Es wurde vom Facebook AI Research Team entwickelt und erschien im Jahr 2016. Zum Zeitpunkt der Verfassung dieser Arbeit ist die Version 1.6.0 die aktuellste. Für die Farbraum Konvertierung wird die “scikit-image” Bibliothek eingesetzt und für die Image Augmentation wurde die Bibliothek “imgaug” ausgewählt. Des weiteren werden Hilfsbibliotheken wie “numpy” und “matplotlib” angewendet.

Das Trainieren von den Modellen wird, aufgrund des hohen Rechenaufwands, auf zwei verschiedenen Plattformen durchgeführt. Die Modelle mit den  $32 \times 32$  Bildern werden auf Google Colab<sup>10</sup> trainiert. Google Colab ist eine Plattform von Google, die es ermöglicht Experimente im Browser mit einer Hochleistungsgrafikkarte (Nvidia Tesla P100) kostenlos umzusetzen. Für das Modell mit den  $128 \times 128$  Bildern wird das Curious Containers (CC) Framework<sup>11</sup> benutzt. Curious Containers ermöglicht eine gleichzeitige Durchführung von verschiedenen Experimenten in einem Cluster von Hochleistungsrechnern.

---

<sup>9</sup><https://pytorch.org/>

<sup>10</sup><https://colab.research.google.com/>

<sup>11</sup><https://www.curious-containers.cc/>

# Kapitel 4

## Implementierung

In diesem Kapitel wird die Implementierung der Methode näher erläutert. Bei der Implementierung werden alle Konzepte aus dem Kapitel Konzeption angewendet.

### 4.1 Binning

Das Binning ist eine wichtige Komponente der Methode, die mit Hilfe der “numpy” Bibliothek für normalisierte Lab-Bilder implementiert wurde. Als erstes wird mit Hilfe der Wurzel anhand der Anzahl von Bins ( $n\_bins$ ) die Breite ( $W$ ) und Höhe ( $H$ ) des Grids, das auf der Abbildung 3.2 zu sehen ist, berechnet. Es wird angenommen dass  $W = H$  ist. Nachdem die Breite des Grids berechnet wurde, wird der Intervall von  $[0, 1]$  in  $W$  gleich große Intervalle aufgeteilt. Als nächstes wird mit Hilfe der *digitize* Funktion der Intervall Index der jeweiligen  $a, b$  Farbkanal Wert von jedem Pixel kalkuliert. Abschließend werden beide Indices zu einem Bin Index auf dem Grid umgewandelt. Der Output ist ein kodiertes Bild mit einem Bin Index per Pixel.

```

1      # a, b sind die Koordinaten der Farbkanäle auf dem Grid
2      def calculate_bin(a, b, width):
3          return (width * b) + a
4
5      def encode_bins(ab_image, n_bins):
6          W = np.sqrt(n_bins).astype(int)
7
8          # Intervall in gleich große Intervalle aufteilen
9          interval = np.linspace(0, 1, W+1)
10
11         # Indices für jeweils a, b Kanäle berechnen
12         indices = np.digitize(ab_image, interval) - 1
13
14         # Bin Index berechnen
15         bins = np.vectorize(calculate_bin)(indices[:, :, 0], indices[:, :, 1], W)
16
17     return bins

```

Code snippet 4.1: Binning eines normalisierten Lab Bildes

Für die Umwandlung werden vor dem Training alle Farben von den Trainingsbildern in Bins klassifiziert. Es wird ein Python Dictionary mit Bin Indices als Key und ein 2-Dimensionales Array mit einem Array pro Farbkanal als Value erzeugt.

```
1      bin_colors = {i: [[], []] for i in range(n_bins)}
```

Code snippet 4.2: Leere Dictionary Erzeugung für  $n\_bins$ 

Anschließend werden die Farben von jedem Farbkanal zu dem jeweiligen Bin Array zugeordnet. Als letztes wird der Durchschnitt pro Bin Index und Farbkanal berechnet. Das Dictionary beinhaltet abschließend ein Array mit 2 einzelnen Werten für den jeweiligen Farbkanal pro Bin Index. Somit kann unkompliziert ein Bin in eine Farbe umgewandelt werden.

```
1      a_color = bin_colors[bin_index][0]
2      b_color = bin_colors[bin_index][1]
```

Code snippet 4.3: “Bin-zu-Farbe” Umwandlung

Der gleiche Vorgang wird für den Modus angewendet. Um einen Kompromiss zwischen Durchschnitt und Modus zu finden, werden der Durchschnitt und der Modus mit einem Temperatur Wert interpoliert.

```

1     a_distance = a_mode - a_mean
2     b_distance = b_mode - b_mean
3
4     a = a_mode - (a_distance * T)
5     b = b_mode - (b_distance * T)
```

Code snippet 4.4: “Bin-zu-Farbe” Berechnung mit einem Temperaturwert

## 4.2 Datensätze

Die Datensätze werden mit Hilfe der “torchvision” Bibliothek von PyTorch importiert und transformiert. Für das Importieren wurde ein “ImageFolder” implementiert, der die Bilder importiert, transformiert, normalisiert und die einzelnen Pixel in Bins klassifiziert. Der Output der “ImageFolder” sind das Graustufenbild, das Bild mit den “ab” Farbkanälen und das in Bins kodierte Bild.

Mit Hilfe eines “DataLoaders” werden die Datensätze in Batches aufgeteilt und gemischt.

## 4.3 Netzwerkarchitektur

Für diese Arbeit wurden 2 U-Nets mit verschiedenen Größen verwendet. Ein U-Net für  $32 \times 32$  Input Bilder und ein U-Net für  $128 \times 128$  Input Bilder. Außerdem wurde das U-Net für  $32 \times 32$  Bilder angepasst, damit es mit einem MSE Loss verwendet werden kann. Bei der Anpassung wurde das Output Volumen zu  $W_{Input} \times H_{Input} \times 2$  geändert, wobei das Netzwerk direkt die Werte für die “ab” Farbkanäle vorhersagt. Bei der Verwendung von einem MSE Loss wird das Binning nicht angewendet.

### 4.3.1 ConvBlock

Das Kernstück eines U-Nets ist der sogenannte ConvBlock. Ein ConvBlock beinhaltet zwei hintereinander geschaltete Blöcke, die wiederum aus einer Convolutional Layer

mit  $3 \times 3$  Filtern und Stride 1, gefolgt von ReLU und Batch Normalization bestehen. Die Convolutional Layers verwenden Padding, um die Dimensionen des Inputs nicht zu verändern. Die Layers wurden mit den PyTorch Klassen *Conv2d*, *BatchNorm2d* und der Funktion *relu* implementiert. Der ConvBlock wird mit einer Klasse implementiert, die wiederrum von der Oberklasse *torch.nn.Module* erbt.

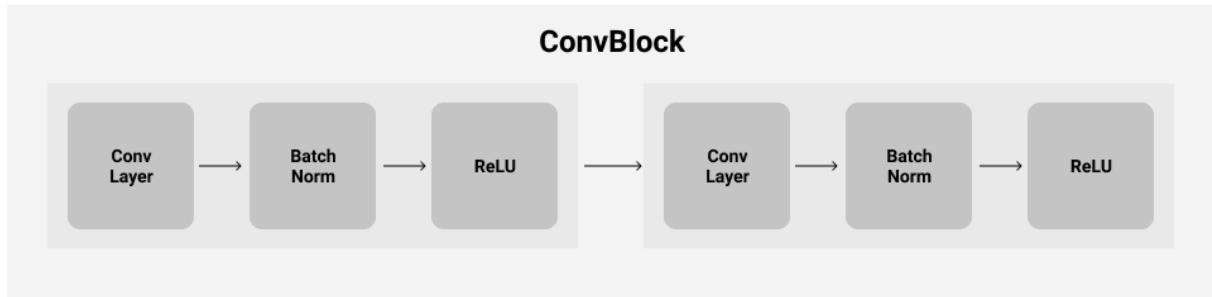
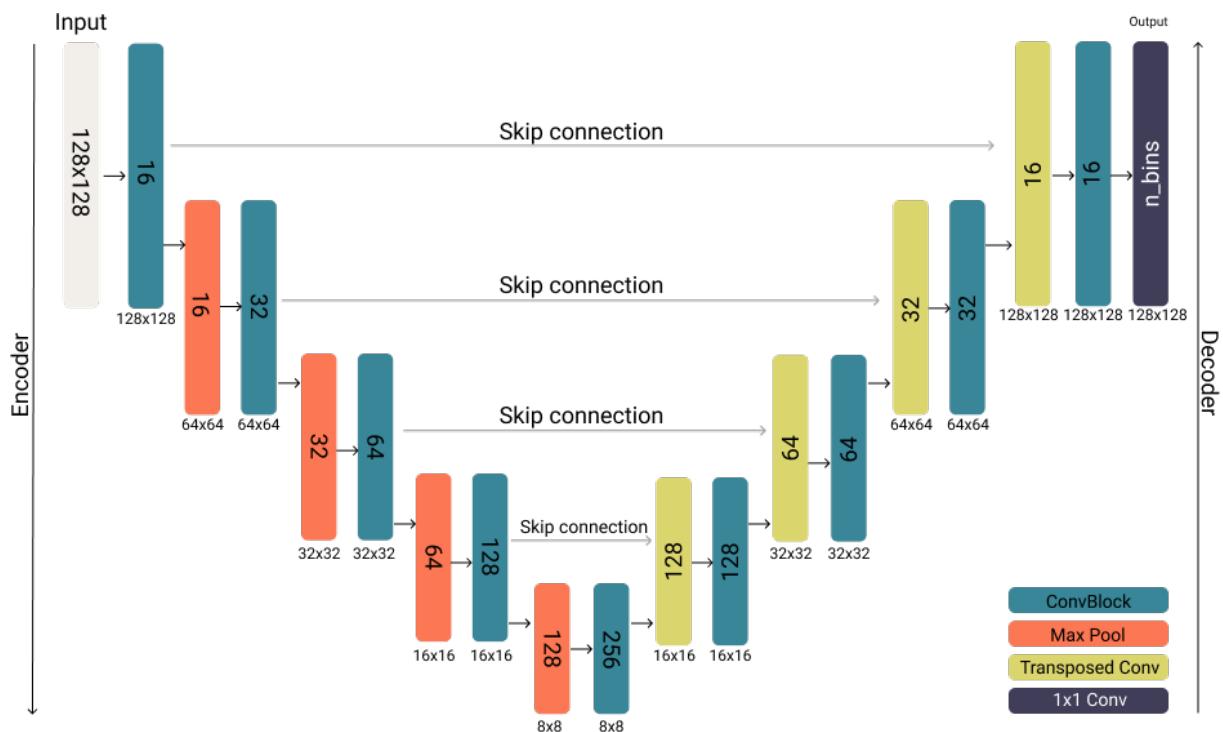


Abbildung 4.1: ConvBlock

### 4.3.2 U-Net

Das U-Net wird mit den Spezifikationen aus 3.4.1 implementiert. Dies geschieht mit einer Klasse, die von der Oberklasse *torch.nn.Module* erbt. Das U-Net verwendet neben den ConvBlocks, Max Pooling Layers mit einem  $2 \times 2$  Filter und Stride 2, um die Dimensionen in den Encoder zu halbieren. Die kleinste Größe der Feature Maps bei dem Encoder ist  $8 \times 8$ . Der Decoder verwendet ebenfalls, neben den ConvBlocks, Transposed Convolutions mit  $3 \times 3$  Filtern, Stride 2 und Padding 1, die die Größe der Feature Maps verdoppeln. Die letzte Layer ist eine Convolutional Layer mit  $1 \times 1$  Filter, Stride 1 und Padding 0, die die Wahrscheinlichkeitsverteilung pro Pixel generiert. Alle Transposed Convolutions werden mit den Outputs der ConvBlocks mit den gleichen Dimensionen aus dem Encoder konkateniert. Es werden zwei verschieden große U-Nets implementiert, eins für  $32 \times 32$  Bilder und eins für  $128 \times 128$  Bilder.

Die Max Pooling Layers wurden mit der *MaxPool2d* Klasse und die Transposed Convolutions mit der *ConvTranspose2d* implementiert.

Abbildung 4.2: U-Net Architektur für  $128 \times 128$  Input Bilder

# Kapitel 5

## Test

In diesem Kapitel werden Tests mit den drei Datensätzen durchgeführt. Darüber hinaus werden verschiedene Hyperparameter getestet.

### 5.1 Spiel-Datensatz Training

Als erstes wird die Methode auf dem Spiel-Datensatz angewendet. Hierbei soll geprüft werden, ob die Methode die erwarteten Ergebnisse liefert. Das U-Net wird für 100 Epochen mit Adam, einer Lernrate von 0.001 und der Cross Entropy Loss Function trainiert. Diese Einstellung zeigte die besten Ergebnisse. Zeitgleich wurden Tests mit 36 und mit 324 Bins durchgeführt, die jedoch keinen Unterschied bei den Ergebnissen verwiesen. Es kann davon ausgegangen werden, dass bei der niedrigen Anzahl an möglichen Farben, ein Unterschied zwischen 36 und 324 Bins nicht zu erkennen ist. Die unteren Ergebnisse wurden mit 324 Bins erstellt.

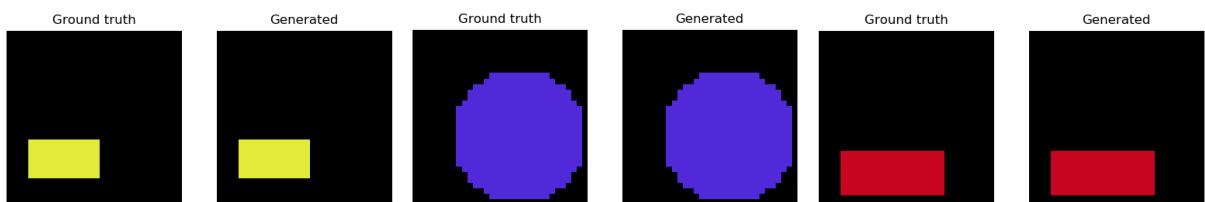


Abbildung 5.1: Beispiele von sehr guten Ergebnissen aus dem Spiel-Datensatz

Bei den oberen Ergebnissen wurden alle Pixel richtig klassifiziert, was bei der Größe des Datensatzes oft zu Overfitting deutet. Die folgenden Ergebnissen zeigen dass das Modell generalisiert und nicht overfitted hat.

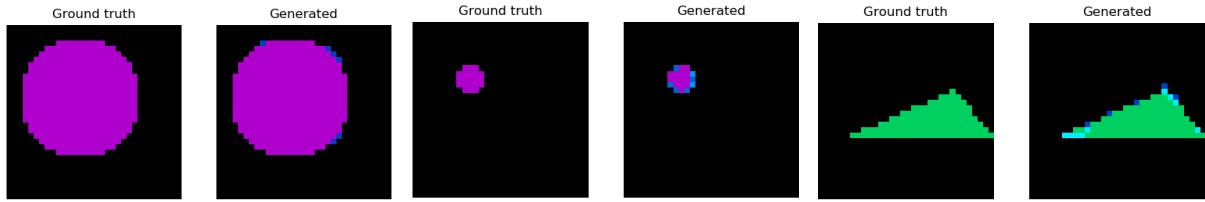


Abbildung 5.2: Beispiele von generalisierten Ergebnissen

Das Modell zeigte bei einigen Ergebnissen, Schwierigkeiten die Pixeln am Rand der geometrischen Formen richtig zu klassifizieren. Dies trifft besonders auf die Kreise und Dreiecke zu, wo die Ränder nicht aus glatten Linien bestehen. Des weiteren wurden die Farben mittels des Durchschnitts für jeden möglichen Bin, für alle Farben von jedem Trainingsbild rekonstruiert. Ein Unterschied zwischen dem Modus und dem Durchschnitt konnte nicht erkannt werden, da beide Werte gleich waren.

Die Ergebnisse bestätigen dass das Binning und die Methode funktionieren. Anschließend wurden Tests mit komplexeren Bildern des Subsets von CIFAR-100 durchgeführt.

## 5.2 CIFAR-100 Subset Training

Das Modell wurde auf 12 Klassen von CIFAR-100 über 100 Epochen mit Adam, einer Lernrate von 0.001 und der Cross Entropy Loss Function trainiert. Diese Einstellungen stellten sich als die beste Kombination von Hyperparametern heraus, jedoch wurde das Training vor Vollendung von 50 Epochen unterbrochen um Overfitting zu verhindern.

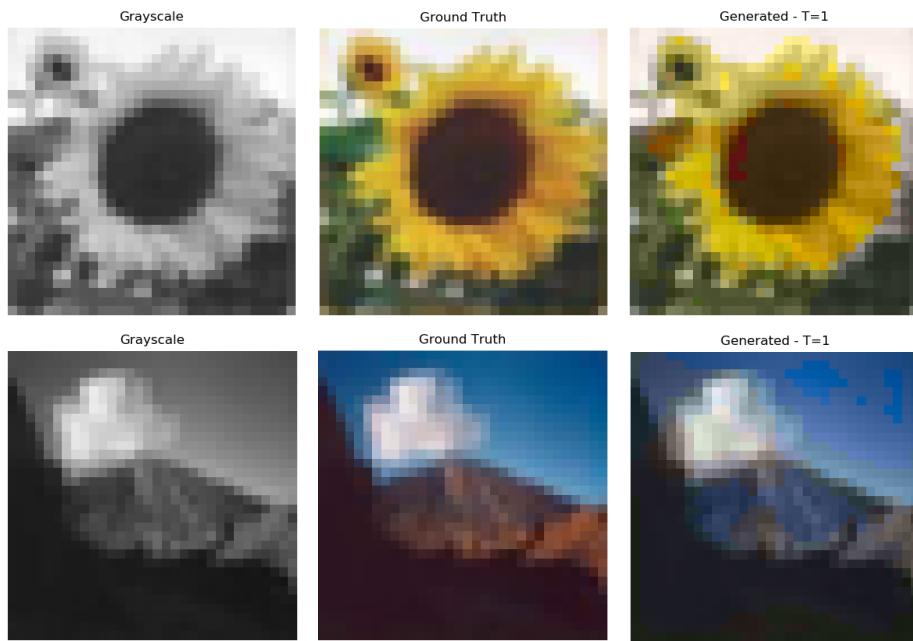


Abbildung 5.3: Beispiele von guten Ergebnissen aus dem Subset von CIFAR-100 mit 324 Bins. Die erste Spalte beinhaltet das Graustufenbild, die zweite Spalte beinhaltet das Original Bild und die letzte Spalte stellt das generierte Bild dar. Das generierte Bild wurde mit einer Temperatur von 1 erzeugt, was bedeutet, dass die rekonstruierten Farben den Durchschnitt aus jedem Bin repräsentieren.

Die Tests mit diesem Datensatz haben gezeigt, dass die Anzahl der Bins bei der Auswahl an möglichen Farben, die Ergebnisse beeinträchtigen. Eine Erhöhung der Trainingszeit zwischen 36 und 324 Bins war nicht zu erkennen.

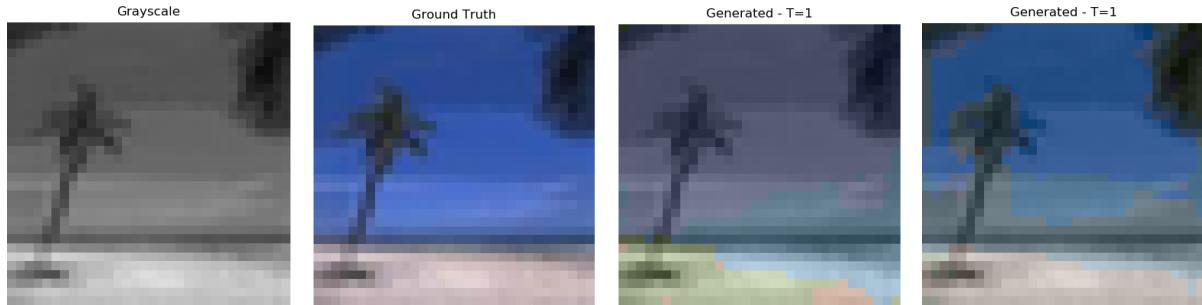


Abbildung 5.4: Einfluss der Anzahl der Bins auf die Ergebnisse. Das zweite Bild ist das original, das dritte Bild wurde mit 36 Bins und einem Temperaturwert von 1 generiert und das vierte mit 324 Bins und ebenfalls mit einem Temperaturwert von 1.

Um das Overfitting zu verhindern wurden zahlreiche Experimente mit verschiedenen Optimierer, Aktivierungsfunktionen und Lernraten durchgeführt. Ein Austausch von ReLU durch Tanh zeigte ein stabileres Trainingsverhalten, aber eine Verschlechterung der Validation Loss. Leaky ReLU zeigte ein ähnliches Verhalten wie ReLU, aber keine Verbesserung der Ergebnissen. Die Verwendung von RMSprop anstelle von Adam zeigte eine langsame Konvergenz Richtung Minimum. Abschließend wurde die Anzahl der Filter in den Convolutional Layers halbiert, was eine positive Wirkung auf das Training hatte.

Um die Performance der Klassifikation gegenüber der Regression zu messen, wurde ein Modell mit der MSE Loss Function trainiert. Dieses Modell wurde ebenfalls mit den gleichen Parametern wie das Klassifikationsmodell trainiert und hat vergleichbare Ergebnisse erreicht. Einige Ergebnisse zeigten blasse Stellen im Vergleich zu dem Klassifikationsmodell, der leuchtende Farben an den gleichen Stellen gezeigt hat.

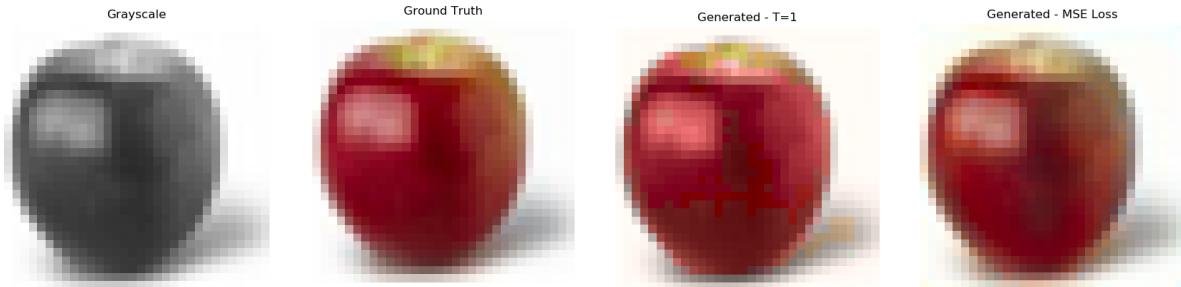


Abbildung 5.5: Vergleich von Klassifikation mit Binning gegenüber Regression. Das zweite Bild wurde mit 324 Bins und dem Cross Entropy Loss generiert. Das dritte Bild wurde ohne Binning und mit einem MSE Loss generiert.

### 5.3 Landscape Datensatz Training

Dieser Datensatz wurde anhand der Hyperparameter Optimierung des CIFAR-100 Subsets trainiert. Es wurde das größere Modell für  $128 \times 128$  Input Bildern angewendet. Das Modell wurde ebenfalls für 36 und 324 Bins, mit dem Adam Optimizer, eine Lernrate von 0.001 und der Cross Entropy Loss Function für 60 Epochen trainiert. Das Modell mit dem besten Validation Loss wurde gespeichert, um die Ergebnisse zu evaluieren. Außerdem wurde der Einfluss der Temperaturwerte und die Anzahl der Bins auf die Ergebnisse gemessen.

Das Modell tendierte bei diesem Datensatz ebenfalls zum Overfitting. Um das zu verhindern wurden die gleichen Techniken wie bei CIFAR-100 angewendet, was keine besseren Ergebnisse geliefert hat. Eine Halbierung der Anzahl der Filter bei den Convolutional Layers führte zu einer Verschlechterung des Validation Loss und half nicht bei Overfitting. Eine Änderung des Optimierers zu RMSprop und der Ersatz von ReLU durch Tanh zeigten ein stabileres Training, aber keine Verbesserung des Validation Loss. Das endgültige Modell wurde trainiert, bis der Validation Loss wieder stieg.



Abbildung 5.6: Ergebnisse mit 324 Bins. Die erste Spalte zeigt das Originale Bild, die zweite zeigt das generierte Bild mit einer Temperatur von 0, die dritte Spalte zeigt das generierte Bild mit einer Temperatur von 0.8 und die vierte Spalte zeigt das generierte Bild mit einer Temperatur von 1

Nach der Auswertung der Experimente wurden 324 Bins und ein Temperaturwert von 0.8 bis 1 bevorzugt, um die bestmöglichen Ergebnisse zu bekommen.

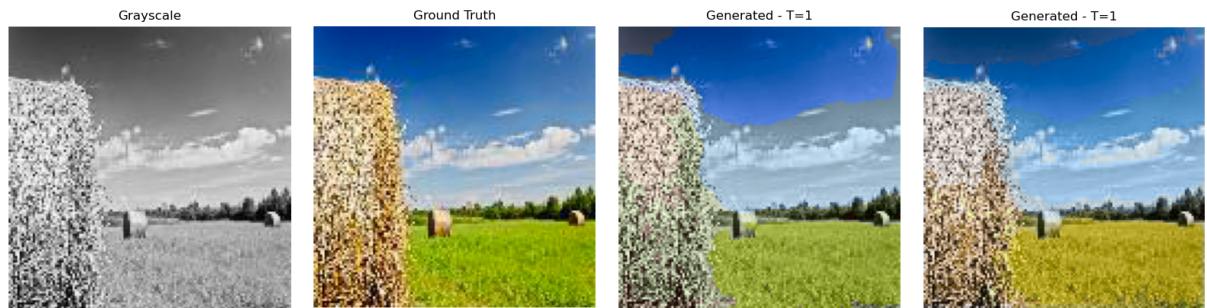


Abbildung 5.7: Ergebnisse mit 36 und 324 Bins. Die erste Spalte zeigt das Graustufenbild, die zweite zeigt das originale Bild, die dritte Spalte zeigt das mit 36 Bins generierte Bild mit einer Temperatur von 1 und die vierte Spalte zeigt das mit 324 Bins generierte Bild mit einer Temperatur von 1

# Kapitel 6

## Evaluation

Im Kapitel Evaluation werden die im Kapitel 5 vorgestellten Tests evaluiert. Außerdem wird die gewählte Methode ausgewertet und mit anderen Methoden verglichen.

### 6.1 Evaluationsmetrik

Für das Problem von Image Colorization existiert keine relevante Evaluationsmetrik, die die Farben von den Objekten auf einem Bild auswerten kann. Das während des Trainings angewendete Cross Entropy Loss ist nicht relevant für die Auswertung der Ergebnisse aus dem Test Datensatz. Aus diesem Grund wurde die Evaluation der Ergebnisse durch eine Menschliche Auswertung wie bei Zhang et al. und Billaut et al. durchgeführt.

### 6.2 Evaluation des Spiel-Datensatzes

Mit dem Spiel-Datensatz wurde die Funktionsweise der Methode bestätigt. Die von Billaut et al. vorgeschlagene Netzwerkarchitektur für Image colorization hat beeindruckende Ergebnisse nach wenigen Epochen erreicht.

Die gewählte Methode für das Binning funktionierte und hat ermöglicht, die originalen Farben wiederherzustellen. Die Anzahl an Bins war für diesen Datensatz nicht relevant da es nur 9 mögliche Farben gab.

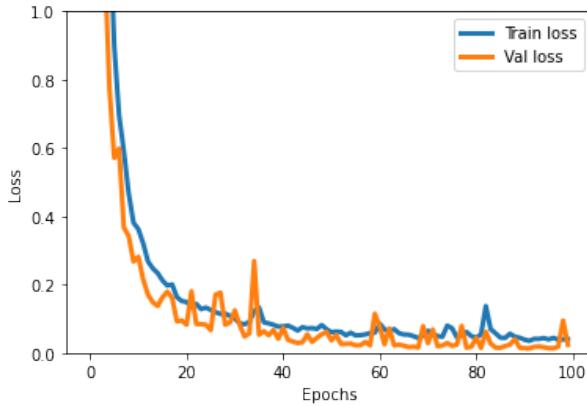


Abbildung 6.1: Training und Validation Loss Verlauf des Spiel-Datensatzes

### 6.3 Evaluation des CIFAR-100 Subsets

Die Ergebnissen aus dem CIFAR-100 Subset zeigten dass das Modell nach wenigen Epochen viele Merkmale lernen konnte. Ein wichtiger Faktor, der erwähnt werden muss, ist dass die Gewichte des Netzwerks zufällig initialisiert wurden und nicht vortrainiert waren.

In diesem Datensatz wurde die Auswirkung der Bin Anzahl gemessen. Die Nutzung von 36 Bins zeigte im Vergleich zu 324 eine Verschlechterung der Farben in der Vorhersage. Dies ist darauf zurückzuführen dass das Modell nur 36 mögliche Farben zu Verfügung hat. Eine Erhöhung der Bins auf 324, ermöglichte es dem Modell eine Auswahl an mehreren Farben zu treffen. Der Ansatz von Billaut et al. verwendet nur 32 Bins und erzielt ähnliche Ergebnisse wie die Methode dieser Arbeit mit 324 Bins. Dies wurde erreicht in dem die Pixel von jedem Trainingsbild vor dem Training in Bins klassifiziert wurden und daraus nur die am meisten vorkommenden 32 Bins ausgewählt wurden. Pixel die nicht in den gewählten 32 Bins klassifiziert werden konnten, wurden in das nächstliegende Bin zugeordnet [BRT18].

Die Methode mit einem MSE Loss liefert eine um ein Vielfaches größere Auswahl an Farben für die Vorhersage. Bei dieser Methode treten die im Kapitel 2.6 erwähnten Schwierigkeiten auf, wobei im Falle dieses Datensatzes, die Schwierigkeiten nur sehr latent ausgeprägt waren. Da die Klassifikationsmethode bessere Ergebnissen geliefert hat und der Fokus der Arbeit auf Klassifikationsmethoden gesetzt war, wurden alle Experimente des Landscape

Datensatzes mit der Klassifikationsmethode durchgeführt.

Der Verlauf von dem Training und Validation Loss deutete bei diesem Datensatz zu Overfitting, was bei der Größe des Datensatzes nicht auszuschließen war. Wie bei 5.2 beschrieben wurden das U-net und die Hyperparameter angepasst, um dieses zu verhindern.

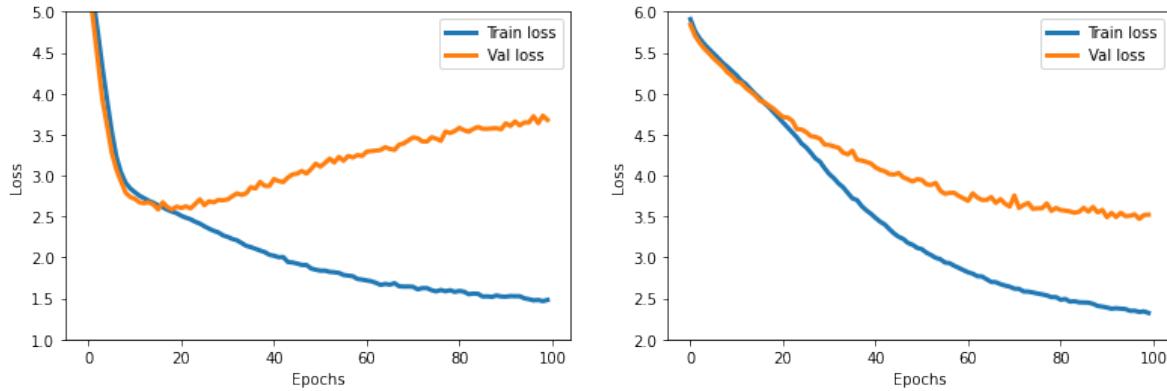


Abbildung 6.2: Overfitting auf dem CIFAR-100 Subset. **Links:** 100 Epochen mit Adam, ReLU und einer Lernrate von 0.001. **Rechts:** 100 Epochen mit Adam, ReLU und einer Lernrate von 0.0001.

Eine Anpassung der Lernrate führte nur zu einem langsameren Training. Eine Reduktion der lernbaren Parameter von 135684 auf 35748 zeigte eine deutliche Verbesserung der Performance des Modells und reduzierte das Overfitting.

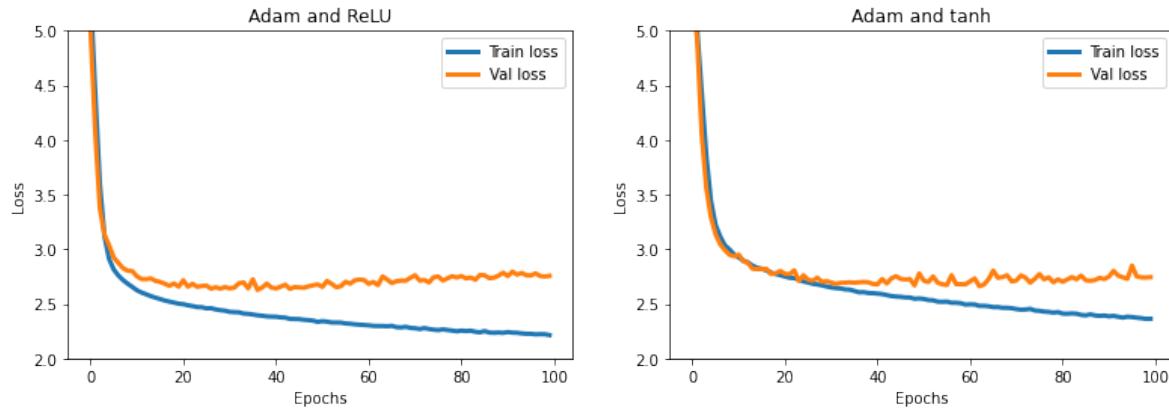


Abbildung 6.3: Loss Verlauf mit einer reduzierten Anzahl an Parametern und verschiedenen Aktivierungsfunktionen. **Links:** 100 Epochen mit Adam, ReLU und einer Lernrate von 0.001. **Rechts:** 100 Epochen mit Adam, Tanh und einer Lernrate von 0.001.

Der Grund für das Overfitting ist in diesem Fall auf die Größe des Subsets zurückzuführen. Um das zu prüfen wurde das Modell auf dem kompletten CIFAR-100 Datensatz für 150 Epochen trainiert, was einen guten Loss Verlauf zeigte. Andererseits, hat das Modell nichts relevantes gelernt und konnte kein Bild richtig einfärben, was bei der Anzahl der Klassen und der Anzahl der Epochen nichts Ungewöhnliches ist.

## 6.4 Evaluation des Landscape Datensatzes

Die Ergebnisse dieses Modells zeigen, dass die ausgewählte Methode mit wenigen Bildexemplaren und Epochen, sehr gute Ergebnisse erreichen kann. Die Anzahl der Bins für diesen Datensatz wurde durch die Ergebnisse der vorgeführten Experimenten auf 324 gesetzt, da diese Anzahl die beste Kombination aus Trainingszeit und Performance gezeigt hat.

Der Loss Verlauf der Experimente deutete bei diesem Datensatz ebenfalls nach wenige Epochen auf Overfitting hin. Nach Zahlreichen Experimenten konnte das Overfitting durch Anpassung der Hyperparameter nur minimiert werden. Für die Auswertung der Ergebnisse wurde das Modell mit dem besten Validation Loss verwendet.

Bei der Evaluation der Ergebnisse ist zu erkennen, dass schon nach den wenigen Epochen das Modell in der Lage ist die wichtigsten Entitäten wie den Himmel, Wolken, Wasser, Grass und Erde richtig einzufärben. Es fällt auf, dass bei einigen Bildern die Vorhersage des Modells realer aussieht, als das Original Bild, wie im zweiten Beispiel von 5.6 zu sehen ist. Im Vergleich zum Modell von Billaut et al. sehen die vorhergesagten Farben von diesem Modell sehr ähnlich aus. Das zeigt, dass die Performance der Methode ohne die Optimierungstechniken vergleichbar ist. Die Auswirkung des Temperaturwerts ist bei beiden Modellen ähnlich. Der Modus der Verteilung zeigt einen rötlichen Ton bei den Ergebnissen und der Durchschnitt zeigt bei einigen Fällen gesättigte Bilder.

Da die Datensatz Größe von Billaut et al. fast identisch zu der Datensatz Größe dieser Arbeit ist, könnte das Problem mit dem Overfitting auf die Qualität des Datensatzes zurückzuführen sein. Der Loss Verlauf zeigt kein Overfitting und das verwendete U-Net ist ähnlich zum U-Net der vorliegenden Arbeit, was ein Overfitting wegen der angewendeten Netzwerkarchitektur ausschließt.

Ein Vergleich mit den Ergebnissen von Zhang et al. wäre nur angemessen mit den Ergebnissen aus deren Klassifikationsmodell ohne Rebalancing. In diesem Fall sind die Resultate dieses Models nicht weit entfernt von deren Ergebnissen. Es ist zu erkennen, dass deren Ergebnisse ähnliche Merkmale mit den Ergebnissen dieser Arbeit vorweisen, wie z.B. die Ähnlichkeit der Vorhersagen zwischen Klassifikationsmethode und Regressionsmethode. Die Ergebnisse mit Class Rebalancing unterscheiden sich deutlicher von den Ergebnissen der Regressionsmethode. Da deren Datensatz über 1.5 Millionen Bilder beinhaltet, kann das Modell mehr Objekte auf den Bildern einfärben, was sich bei einigen Bildern z.B. mit Menschen, bemerkbar macht.

Im allgemeinen sind die Ergebnisse dieses Modells mit anderen Ergebnissen von Klassifikationsmethoden ohne Optimierungstechniken wie Class Rebalancing vergleichbar.

# Kapitel 7

## Fazit

### 7.1 Zusammenfassung

Das Ziel dieser Bachelorarbeit war es, die Methoden von Image Colorization anhand Convolutional Neural Networks zu untersuchen, insbesondere Klassifikationsmethoden. Durch die ausführliche Auseinandersetzung mit Klassifikationsmethoden und die Implementierung der verwendeten Techniken wie Binning, konnte eine Methode, basierend auf dem letzten Stand der Technik, implementiert werden. Diese Methode wurde angewandt und mit verschiedenen Datensätzen getestet, um die Ergebnisse zu untersuchen.

Die Ergebnisse dieser Arbeit haben gezeigt, dass die Methoden funktionieren und sogar mit weniger Bildern und Epochen gute Ergebnisse produzieren. Zahlreiche Experimente bestätigten, dass Klassifikationsmethoden bessere Ergebnisse als Regressionsmethoden liefern. Diese Differenz ist deutlicher bei Methoden die Optimierungstechniken, wie Class Rebalancing, anwenden.

### 7.2 Kritischer Rückblick

Der Fokus dieser Arbeit wurde auf die Untersuchung der Methoden gelegt. Bei der Erstellung, Implementierung und Durchführung der Methode wurde festgestellt, dass das Problem von Image Colorization sehr komplex ist. Obwohl die erreichten Ergebnissen ausreichend für den Vergleich mit anderen Methoden sind, wurde festgestellt dass ohne Optimierungstechniken eine Verbesserung der Qualität der Bilder nur schwer zu erreichen ist.

Die Auswahl eines balancierten Datensatzes ist ein wichtiger Faktor für die Qualität der Ergebnisse und kann Overfitting verhindern. Die Größe des Datensatzes ist auch entscheidend bei dem Training eines nicht vor-trainierten Modells.

Die Zielsetzung wurde erreicht in dem die Klassifikationsmethode bestätigt wurde, obwohl die Qualität der Ergebnisse durch weiteres Training und Optimierung verbessert werden kann.

### 7.3 Ausblick

Die implementierte Methode dieser Arbeit hat viel Potential und kann durch Anpassung der Techniken und Anwendung von Optimierungstechniken verbessert werden. Eine Anwendung auf andere und größere Datensätze wird mit Sicherheit bessere Ergebnisse produzieren. Die Methode auf Videos anzuwenden würde ebenfalls interessante Ergebnisse liefern.

Methoden aus dem Bereich der unüberwachten Lernens werden heutzutage angewendet um das Image Colorization Problem zu lösen. Generative adversarial Networks (GANs) werden in der Praxis öfter angewendet, als normale CNNs da diese viel bessere Ergebnisse erzeugen.

# Abbildungsverzeichnis

2.1	Fully-connected Neural Network mit 2 Layers (eine Hidden Layer mit 4 Neuronen und eine Output Layer mit 2 Neuronen) [Fei20a]	5
2.2	Sigmoid Aktivierungsfunktion [Fei20b]	7
2.3	Tanh Aktivierungsfunktion [Fei20b]	8
2.4	Rectified Linear Unit (ReLU) [Fei20b]	9
2.5	Leaky ReLU [ccs20]	9
2.6	Gradient descent visualisiert [Bha18]	12
2.7	Backpropagation Beispiel anhand einer 2D Neuron mit der Aktivierungsfunktion Sigmoid [Fei20c]	13
2.8	Typische Struktur von einem Convolutional Neural Network [Com15]	14
2.9	Beispiel eines Forward pass von einer Convolutional Layer mit einem $7 \times 7 \times 3$ Input Volumen, zwei $3 \times 3 \times 3$ Filtern, Padding 1 und Stride 2. [Fei20d]	15
2.10	Max pooling Operation mit $2 \times 2$ Filtern und Stride 2 [Fei20d]	16
2.11	Die komplette Transposed Convolution Operation [Zha20]	16
3.1	Original Bild in RGB oben links, Belichtungskanal "L" oben rechts, Farbkanal "a" unten links und Farbkanal "b" unten rechts.	21
3.2	Grid mit 36 bins. Die x-Achse bildet die Werte des Farbkanals "a" und die y-Achse die Werte des Farbkanals "b" ab.	22
3.3	U-net Architektur (Beispiel für $32 \times 32$ Pixels in der niedrigsten Auflösung). Jede blaue Box entspricht einer multi-Kanal Feature Map. Die Tiefe der Feature Maps ist gekennzeichnet durch die Zahl über der Box. Die Breite und Höhe ist durch die Zahl unten links erkennbar. Die weißen Boxen repräsentieren die kopierten Feature Maps. Die Pfeile bestimmen die verschiedenen Operationen. [RFB15]	25
3.4	Beispiel von Trainingsbildern mit einer der möglichen Farbe pro Klasse	26
4.1	ConvBlock	32
4.2	U-Net Architektur für $128 \times 128$ Input Bilder	33

5.1	Beispiele von sehr guten Ergebnissen aus dem Spiel-Datensatz . . . . .	34
5.2	Beispiele von generalisierten Ergebnissen . . . . .	35
5.3	Beispiele von guten Ergebnissen aus dem Subset von CIFAR-100 mit 324 Bins. Die erste Spalte beinhaltet das Graustufenbild, die zweite Spalte beinhaltet das Original Bild und die letzte Spalte stellt das generierte Bild dar. Das generierte Bild wurde mit einer Temperatur von 1 erzeugt, was bedeutet, dass die rekonstruierten Farben den Durchschnitt aus jedem Bin repräsentieren. . . . .	36
5.4	Einfluss der Anzahl der Bins auf die Ergebnisse. Das zweite Bild ist das original, das dritte Bild wurde mit 36 Bins und einem Temperaturwert von 1 generiert und das vierte mit 324 Bins und ebenfalls mit einem Tempe- raturwert von 1. . . . .	37
5.5	Vergleich von Klassifikation mit Binning gegenüber Regression. Das zweite Bild wurde mit 324 Bins und dem Cross Entropy Loss generiert. Das dritte Bild wurde ohne Binning und mit einem MSE Loss generiert. . . . .	38
5.6	Ergebnisse mit 324 Bins. Die erste Spalte zeigt das Originale Bild, die zweite zeigt das generierte Bild mit einer Temperatur von 0, die dritte Spalte zeigt das generierte Bild mit einer Temperatur von 0.8 und die vierte Spalte zeigt das generierte Bild mit einer Temperatur von 1 . . . . .	39
5.7	Ergebnisse mit 36 und 324 Bins. Die erste Spalte zeigt das Graustufenbild, die zweite zeigt das originale Bild, die dritte Spalte zeigt das mit 36 Bins generierte Bild mit einer Temperatur von 1 und die vierte Spalte zeigt das mit 324 Bins generierte Bild mit einer Temperatur von 1 . . . . .	39
6.1	Training und Validation Loss Verlauf des Spiel-Datensatzes . . . . .	41
6.2	Overfitting auf dem CIFAR-100 Subset. <b>Links:</b> 100 Epochen mit Adam, ReLU und einer Lernrate von 0.001. <b>Rechts:</b> 100 Epochen mit Adam, ReLU und einer Lernrate von 0.0001. . . . .	42
6.3	Loss Verlauf mit einer reduzierten Anzahl an Parametern und verschiedenen Aktivierungsfunktionen. <b>Links:</b> 100 Epochen mit Adam, ReLU und einer Lernrate von 0.001. <b>Rechts:</b> 100 Epochen mit Adam, Tanh und einer Lernrate von 0.001. . . . .	43

# Source Code Content

4.1	Binning eines normalisierten Lab Bildes . . . . .	30
4.2	Leere Dictionary Erzeugung für $n\_bins$ . . . . .	30
4.3	“Bin-zu-Farbe” Umwandlung . . . . .	30
4.4	“Bin-zu-Farbe” Berechnung mit einem Temperaturwert . . . . .	31

# Glossar

**Bin** Behälter. 18, 21

**Grid** Raster. 21, 22, 29, I

**Ground Truth** Zielwerte. 10

**Layer** Schicht. 3, 14, 15, I

**Loss Function** Kostenfunktion. 10, 11, 13, 18

**Overfitting** Überanpassung eines Modells auf die Trainings Datenpunkte. 34, 35, 37, 38, 41, 42, 43, 45, II

**Stride** Schrittweite einer Faltung bei einem CNN. 14, 15, 16, I

# Abkürzungsverzeichnis

**CIE** Internationale Beleuchtungskommission. 17

**CNN** Convolutional Neural Network (Gefaltetes Neuronales Netzwerk). 14, 23, 46, IV

# Literaturverzeichnis

- [BRT18] Vincent Billaut, Matthieu de Rochemonteix und Marc Thibault. *ColorUNet: A convolutional classification approach to colorization*. 2018. arXiv: 1811.03120 [cs.CV]. (Besucht am 01.08.2020).
- [DHS11] John Duchi, Elad Hazan und Yoram Singer. „Adaptive subgradient methods for online learning and stochastic optimization“. In: *Journal of Machine Learning Research* 12.Jul (2011), S. 2121–2159.
- [GBC16] Ian Goodfellow, Yoshua Bengio und Aaron Courville. *Deep Learning*. <http://www.deeplearningbook.org>. MIT Press, 2016.
- [KB14] Diederik P. Kingma und Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2014. arXiv: 1412.6980 [cs.LG].
- [LLW04] Anat Levin, Dani Lischinski und Yair Weiss. „Colorization Using Optimization“. In: *ACM Trans. Graph.* 23.3 (Aug. 2004), S. 689–694. ISSN: 0730-0301. DOI: 10.1145/1015706.1015780. URL: <https://doi.org/10.1145/1015706.1015780>.
- [NH10] Vinod Nair und Geoffrey E. Hinton. „Rectified Linear Units Improve Restricted Boltzmann Machines“. In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML’10. Haifa, Israel: Omnipress, 2010, S. 807–814. ISBN: 9781605589077. (Besucht am 27.07.2020).
- [Özb19] Gökhan Özbulak. *Image Colorization By Capsule Networks*. 2019. arXiv: 1908.08307 [eess.IV].
- [RFB15] Olaf Ronneberger, Philipp Fischer und Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV].
- [Zha20] Aston Zhang u. a. *Dive into Deep Learning*. <https://d2l.ai>. 2020.
- [ZIE16] Richard Zhang, Phillip Isola und Alexei A. Efros. *Colorful Image Colorization*. 2016. arXiv: 1603.08511 [cs.CV]. (Besucht am 01.08.2020).

# Onlinereferenzen

- [15] *Backpropagation*. 2015. URL: <http://www.inztitut.de/blog/glossar/backpropagation/> (besucht am 01.08.2020).
- [Bec19] Roland Becker. *Convolutional Neural Networks – Aufbau, Funktion und Anwendungsbiete*. 2019. URL: <https://jaai.de/convolutional-neural-networks-cnn-aufbau-funktion-und-anwendungsbiete-1691/> (besucht am 01.08.2020).
- [Dip19] Nico Litzel Dipl.-Ing. (FH) Stefan Luber. *Was ist ein Convolutional Neural Network?* 2019. URL: <https://www.bigdata-insider.de/was-ist-ein-convolutional-neural-network-a-801246/> (besucht am 01.08.2020).
- [Fei17a] Serena Yeung Fei-Fei Li Justin Johnson. *Neural Networks 1*. 2017. URL: <https://cs231n.github.io/neural-networks-1/> (besucht am 12.07.2020).
- [Fei17b] Serena Yeung Fei-Fei Li Justin Johnson. *Optimization 2*. 2017. URL: <https://cs231n.github.io/optimization-2/> (besucht am 01.08.2020).
- [Moe18] Julian Moeser. *Funktionsweise und Aufbau künstlicher neuronaler Netze*. 2018. URL: <https://jaai.de/kuenstliche-neuronale-netze-aufbau-funktion-291/> (besucht am 10.07.2020).
- [Ngu20] Hoang Tu Nguyen. *Einführung in die Welt der Autoencoder*. 2020. URL: [%5Curl%7Bhttps://data-science-blog.com/blog/2020/04/01/einfuehrung-in-die-welt-der-autoencoder/%7D](https://data-science-blog.com/blog/2020/04/01/einfuehrung-in-die-welt-der-autoencoder/) (besucht am 01.08.2020).

# Bildreferenzen

- [Bha18] Saugat Bhattacharai. *What is gradient descent in machine learning?* 2018. URL: <https://saugatbhattacharai.com.np/what-is-gradient-descent-in-machine-learning/> (besucht am 01.08.2020).
- [ccs20] ccs96307. 2020. URL: <https://clay-atlas.com/us/blog/2020/02/03/machine-learning-english-note-relu-function/> (besucht am 31.07.2020).
- [Com15] Wikimedia Commons. *Typical CNN architecture.* 2015. URL: [https://upload.wikimedia.org/wikipedia/commons/6/63/Typical\\_cnn.png](https://upload.wikimedia.org/wikipedia/commons/6/63/Typical_cnn.png) (besucht am 03.04.2018).
- [Fei20a] Serena Yeung Fei-Fei Li Justin Johnson. 2020. URL: <https://cs231n.github.io/neural-networks-1/> (besucht am 10.07.2020).
- [Fei20b] Serena Yeung Fei-Fei Li Justin Johnson. 2020. URL: <https://cs231n.github.io/neural-networks-1/> (besucht am 12.07.2020).
- [Fei20c] Serena Yeung Fei-Fei Li Justin Johnson. 2020. URL: <https://cs231n.github.io/optimization-2/> (besucht am 01.08.2020).
- [Fei20d] Serena Yeung Fei-Fei Li Justin Johnson. 2020. URL: <https://cs231n.github.io/convolutional-networks/> (besucht am 01.08.2020).

# Anhang A

TODO

# **Eigenständigkeitserklärung**

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel verfasst habe. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt.

Berlin, den XX.XX.2018

Adrian Saiz Ferri