HAW HAMBURG

MASTERTHESIS
Juri Zach

Entwicklung einer Qualitätsmetrik für Interpretationen von neuronalen Netzen

FAKULTÄT TECHNIK UND INFORMATIK
Department Informatik

Faculty of Computer Science and Engineering
Department Computer Science

HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN HAMBURG
Hamburg University of Applied Sciences

Juri Zach

Entwicklung einer Qualitätsmetrik für Interpretationen von neuronalen Netzen

Masterarbeit eingereicht im Rahmen der Masterprüfung
im Studiengang *Master of Science Informatik*
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer: Prof. Dr. Kai von Luck
Zweitgutachter: Prof. Dr. Tim Tiedemann

Eingereicht am: 03. Mai 2021

Juri Zach

Thema der Arbeit

Entwicklung einer Qualitätsmetrik für Interpretationen von neuronalen Netzen

Stichworte

neuronale Netze, erklärbare künstliche Intelligenz, Bayessche Optimierung, evolutionäre Algorithmen

Kurzzusammenfassung

Der Forschungsbereich der künstlichen Intelligenz hat in den letzten Jahren große Fortschritte gemacht, insbesondere im Bereich der neuronalen Netze. Der Einsatz von neuronalen Netzen in sicherheitskritischen Anwendungsbereichen birgt dennoch einige Risiken, da neuronale Netze *Black Box* Systeme sind. Das bedeutet, dass zum jetzigen Zeitpunkt weder das gelernte Wissen noch die Entscheidungsgrundlagen von neuronalen Netzen verstanden werden.

Der Forschungsbereich der erklärbaren künstlichen Intelligenz versucht, dieses Problem durch die Entwicklung von Interpretationsmethoden für neuronale Netze zu lösen. Doch obwohl auch hier große Fortschritte erzielt wurden, bleiben sowohl das gelernte Wissen von neuronalen Netzen als auch dessen Entscheidungsgrundlagen größtenteils unbekannt. Besonders vorangetrieben wurde die Entwicklung besserer Interpretationsmethoden durch bessere Regularisierungsmethoden. Leider gibt es in der wissenschaftlichen Gemeinschaft noch keinen genauen Konsens darüber, was Regularisierungsmethoden genau leisten sollen.

In dieser Arbeit werden die Anforderungen an Regularisierungsmethoden für Interpretationen von neuronalen Netzen definiert und eine Metrik entwickelt, mit der die Qualität einer Interpretation gemessen werden kann. Mithilfe dieser Qualitätsmetrik können Interpretationsmethoden für neuronale Netze erstmals automatisiert verbessert werden. Dies wird experimentell bewiesen, indem Interpretationsmethoden für zwei unterschiedlich komplexe Netze mithilfe der Bayesschen Optimierung und evolutionärer Algorithmen optimiert werden. Die hierdurch entwickelten Interpretationsmethoden liefern vor allem auf den tiefen und schwer zu interpretierenden Schichten der neuronalen Netze gute und für Menschen verständliche Ergebnisse.

Juri Zach

Title of Thesis

Development of a quality metric for neural network interpretations

Keywords

neural network, explainable artificial intelligence, Bayesian optimization, evolutionary algorithms

Abstract

The research field of artificial intelligence has made great progress in recent years, especially in the area of neural networks. The use of neural networks in safety-critical application areas nevertheless involves some risks since neural networks are *black box* systems. This means that at this point in time, neither the learned knowledge nor the decision-making basis of neural networks is understood. The research area of explainable artificial intelligence is trying to solve this problem by developing interpretation methods for neural networks. However, although much progress has been made in this area as well, both the learned knowledge of neural networks and its basis for decision making remain largely unknown. In particular, the development of better interpretation methods has been driven by better regularization methods. Unfortunately there is not yet a precise consensus in the scientific community on exactly what regularization methods should do.
This paper defines the requirements for regularization methods for neural network interpretations and develops a metric that can be used to measure the quality of an interpretation. With the help of this quality metric, interpretation methods for neural networks can be improved automatically for the first time. This is experimentally proven by optimizing interpretation methods for two differently complex networks using Bayesian optimization and evolutionary algorithms. The interpretation methods developed in this way deliver good results that are understandable for humans, even at the deep layers of the neural networks which are difficult to interpret.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Die Forschung im Bereich der künstlichen Intelligenz hat in den letzten Jahren große Fortschritte gemacht und findet sich zunehmend im alltäglichen Leben wieder. Bekannte Beispiele sind Facebook-Newsfeeds, Chatbots, Spracherkennung, digitales Marketing oder Einparkhilfen im Auto. Doch auch sicherheitskritische Anwendungen wie intelligente Roboter, medizinische Diagnosesysteme oder autonom fahrende Autos werden stark vorangetrieben. Für viele dieser Aufgaben werden tiefe neuronale Netze eingesetzt. Diese sind in der Lage, die hochdimensionalen Funktionen zu approximieren, mit denen komplexe Aufgaben wie Bild- und Spracherkennung oder das Steuern von Roboter Aktoren gelöst werden.
Allerdings hat diese Technologie im Bereich der Sicherheit und Testbarkeit starke Schwachstellen. Aufgrund der Komplexität ist es kaum möglich, nachzuvollziehen, was ein neuronales Netz gelernt hat und wie es seine Entscheidungen trifft.
Der Forschungsbereich der erklärbaren künstlichen Intelligenz befasst sich unter anderem mit der Interpretierbarkeit von neuronalen Netzen. Die Interpretation von neuronalen Netzen bezeichnet die Fähigkeit, das interne Wissen oder die Entscheidungsgrundlage eines neuronalen Netzes in verständlicher Weise einem Menschen zu erklären oder zu präsentieren [6].
Dies soll dabei helfen, die Sicherheit von neuronalen Netzen zu gewährleisten, das Vertrauen der Anwender zu gewinnen und neuronale Netze auf Werte zu validieren, die sich nicht als mathematische Funktion beschreiben lassen. Hierzu gehört zum Beispiel das Treffen ethisch korrekter Entscheidungen. Auch beim wissenschaftlichen Erkenntnisgewinn aus komplexen Systemen und großen Datenmengen kann die Interpretation von neuronalen Netzen weiterhelfen.
Obwohl in den letzten Jahren große Fortschritte bei der Interpretation neuronaler Netze gemacht wurden, bleiben sowohl das gelernte Wissen von neuronalen Netzen als auch deren Entscheidungsgrundlagen größtenteils unbekannt.

1.1 Motivation

Der Forschungsbereich zur Interpretation von neuronalen Netzen befasst sich einerseits mit theoretischen Überlegungen und andererseits mit praktischen Methoden. Die Forschung in beiden Bereichen finden aber oft getrennt voneinander statt.
In theoretischen Arbeiten wird genauer definiert, was Interpretationsmethoden sind sowie ihr potenzieller Nutzen und wie dieser getestet werden kann formuliert. In vielen praktischen Arbeiten der letzten Jahre wurde mithilfe von Regularisierungsmethoden die Entwicklung besserer Interpretationsmethoden stark vorangetrieben. Dennoch ist die Qualität moderner Interpretationsmethoden nicht ausreichend, als dass sie sich für praktische Anwendungen eignen.
Die Regularisierung von Interpretationsmethoden scheint das Potenzial zu haben, das Wissen von neuronalen Netzen auf menschlich verständliche Weise darzustellen. Meines Wissens gibt es in der wissenschaftlichen Gemeinschaft aber noch keinen genauen Konsens darüber, was Regularisierungsmethoden leisten sollen und wie dieses Ziel erreicht wird. Mithilfe einer klaren Zielsetzung kann die Erforschung besserer Regularisierungsmethoden fokussierter vorangetrieben und somit die Entschlüsselung neuronaler Netze ermöglicht werden.

1.2 Zielsetzung

In dieser Arbeit werden die Notwendigkeit der Regularisierung von Interpretationsmethoden genauer untersucht und die Anforderungen an Regularisierungsmethoden definiert. Um diese Anforderungen zu erfüllen, wird eine neuronale Sprache entwickelt, welche die Verteilung der internen Aktivierungen eines neuronalen Netzes approximiert. Mithilfe dieser Sprache können die Qualität einer Interpretation gemessen und somit die Regularisierungsmethoden gezielt verbessert werden.

1.3 Aufbau der Arbeit

In Kapitel 2 wird auf wichtige Forschungsergebnisse hingewiesen, auf denen diese Arbeit aufbaut, und das Forschungsfeld der Arbeit eingegrenzt. Die Notwendigkeit für Regularisierungsmethoden und ihre Anforderungen werden in Kapitel 3 definiert. In Kapitel 4

wird der Aufbau einer neuronalen Sprache zum Messen der Interpretationsqualität erklärt und in den Kapiteln 5 und 6 experimentell getestet.
In den Kapiteln 7 und 8 werden offene Fragen diskutiert und ein abschließendes Fazit gezogen.

2 Eine kurze Geschichte der Interpretation neuronaler Netze

Der Forschungsbereich zur Interpretation neuronaler Netze ist ein sehr junger Forschungsbereich, der erst mit dem Aufschwung der neuronalen Netze entstanden ist. Zur heutigen Zeit ist dieser Forschungszweig noch von geringer Bedeutung und findet kaum praktische Anwendung. Der Grund hierfür ist, dass für viele Anwendungen von neuronalen Netzen noch keine Interpretationen benötigt werden, weil die Konsequenzen bei Fehlern nicht signifikant oder die Anwendungen ausreichend erforscht sind und somit genug Vertrauen hergestellt wurde [6].
Laut Doshi-Velez und Kim [6] wird die Interpretation von neuronalen Netzen vor allem benötigt, wenn die Formulierung der Problemstellung unvollständig ist. In der Arbeit von Lipton [13] wird behauptet, Interpretation werde benötigt, wenn neuronale Netze auf Ziele angepasst werden, die sich nicht als mathematische Funktionen formulieren lassen.
Beide Definitionen beschreiben Wissenslücken in der Problemstellung. Im Folgenden werden einige Beispiele dafür aufgelistet:

- Wissenschaftliches Verständnis: Neuronale Netze werden häufig für Probleme eingesetzt, bei denen keine optimale Lösung bekannt ist, oder sogar das Problem selbst nicht vollständig verstanden wird. Dennoch können neuronale Netze das Problem lösen. Interpretationsmethoden können dabei helfen, den Lösungsweg des neuronalen Netzes zu verstehen und mögliche Fehler aufzudecken. Hierdurch kann sowohl das Verständnis des Problems verbessert als auch ein möglicher Lösungsweg aufgedeckt werden.

- Sicherheit: Bei sicherheitskritischen Anwendungen ist es für gewöhnlich nicht möglich, das neuronale Netz mit jedem möglichen Eingangswert zu testen. Durch ein

Verständnis des vom neuronalen Netz gelernten Wissens und seiner Entscheidungsgrundlagen kann die Sicherheit dennoch gewährleistet und mögliche Fehler und Schwächen im Vorfeld erkannt werden.

- Ethik: Für bestimmte Anwendungen wie beispielsweise Vorauswahlen von Bewerbungen oder Kreditanfragen ist es wichtig, faire und ethisch korrekte Entscheidungen zu treffen. Da Fairness und Ethik sich allerdings nicht als mathematische Funktion formulieren lassen, muss das neuronale Netz begründen, auf welcher Basis es seine Entscheidungen trifft, damit Fairness und Ethik aus menschlicher Sicht validiert werden können.

- Vertrauen: In einigen Anwendungsfeldern wie beispielsweise der medizinischen Diagnose können Fehler gravierende Folgen haben. Um dennoch neuronale Netze verwenden zu können, muss eine Vertrauensbasis zum Anwender geschaffen werden. Vertrauen beschreibt in diesem Kontext die Sicherheit, dass ein neuronales Netz in realen Szenarien gute Entscheidungen trifft, sowie eine Abschätzung über die Stärken und Schwächen des Netzes. Eine Möglichkeit, diese Vertrauensbasis zu schaffen, ist die Begründung von Entscheidungen.

Um neuronale Netze zu interpretieren, werden vorwiegend sogenannte *Post-hoc* (nachträgliche) Interpretationsmethoden verwendet. Diese sind in der Lage, Informationen aus einem fertig trainierten neuronalen Netz zu extrahieren. Es gibt auch Ansätze, bei denen der Aufbau eines neuronalen Netzes so verändert wird, dass es strukturell interpretierbar ist. Diese Ansätze werden hier allerdings nicht näher behandelt.
Die meisten Forschungen zur Interpretation von neuronalen Netzen befassen sich mit Visualisierungsmethoden [5] [7] [15] [18] [19] [20] [21] [23] [31] oder Bedeutungszuweisungen [8] [11] [23]. Während Bedeutungszuweisungen die Relevanz einzelner Eingangswerte für bestimmte Entscheidungen des neuronalen Netzes aufzeigen, werden Visualisierungsmethoden dazu verwendet, das interne Wissen, welches in den tiefen Schichten des neuronalen Netzes gespeichert ist, zu visualisieren. Der Großteil der wissenschaftlichen Arbeiten interpretiert neuronale Netze im Bildbereich, da hier die Interpretationen intuitiv als Bilder dargestellt werden können. Auch diese Arbeit befasst sich mit maschinellem Sehen. Die hier verwendeten Interpretationsmethoden können aber auch in anderen Domänen verwendet werden.
Zur Interpretation der neuronalen Netze werden in dieser, wie in vielen anderen Arbeiten, lernbasierte Interpretationsmethoden verwendet. Diese sind sehr vielfältig in ihrer

Anwendbarkeit und eignen sich für alle Arten von Lernmethoden, welche über das Gradientenverfahren optimiert werden.

2.1 Lernbasierte Interpretationsmethoden

Der Ausdruck *lernbasierte Interpretationsmethoden* beschreibt eine Menge von Interpretationsmethoden, bei denen Interpretationen mithilfe eines Optimierungsverfahren gelernt werden. Hierfür wird eine Verlustfunktion definiert, welche die Stärke eines bestimmten Effektes im neuronalen Netz misst. Die Verlustfunktion könnte zum Beispiel die Aktivierungsstärke eines einzelnen Neurons messen. Anschließend werden dem neuronalen Netz zufällige Eingangswerte übergeben. Diese Eingangswerte werden mithilfe der Fehlerrückführung darauf optimiert, den Verlustwert zu minimieren. Die optimierten Eingangswerte zeigen anschließend die Merkmale, mit denen der gewünschte Effekt im neuronalen Netz hervorgerufen wird.
Da bei lernbasierten Interpretationsmethoden die optimierten Eingangswerte die Interpretation darstellen, muss der Lernalgorithmus immer in derselben Domäne interpretiert werden, in der er arbeitet. Das bedeutet, dass die Interpretationen eines Bildklassifikators als Bilder dargestellt werden, während die Interpretationen einer Spracherkennung als Audiospur dargestellt werden.
Durch verschiedene Auslegungen der lernbasierten Interpretationsmethoden können beispielsweise der Zweck einzelner Neuronen dargestellt [7], Aktivierungs-Informationen welche in tiefen Schichten entstehen, reproduziert oder verstärkt [14] [15] oder Linearkombinationen verschiedener Neuronen interpretiert werden [5] [10] [20] [21].
Auch Bedeutungszuweisungen können über lernbasierte Interpretationsmethoden dargestellt werden [8], doch die spielen in dieser Arbeit nur eine untergeordnete Rolle.
Im Bildbereich sollen hierdurch Interpretationsbilder entstehen, welche Muster, Objekte, Lebewesen oder andere Dinge zeigen, die von einem menschlichen Betrachter erkannt und interpretiert werden können. Leider zeigen die optimierten Eingangsbilder häufig nur sehr verrauschte Muster, obwohl sie die gewünschten Effekte im neuronalen Netz erzeugen (siehe Abbildung 2.1 linkes Bild).
Diese Muster zeigen große Ähnlichkeiten zu den Mustern, mit denen feindliche Beispiele (engl.: adversarial examples) [19] [28] erstellt werden. Bei feindlichen Beispielen handelt es sich um Eingangsbilder, die so verändert werden, dass sie ein maschinelles Lernmodell zu einer falschen Vorhersage veranlassen. Um feindliche Beispiele zu erzeugen, werden

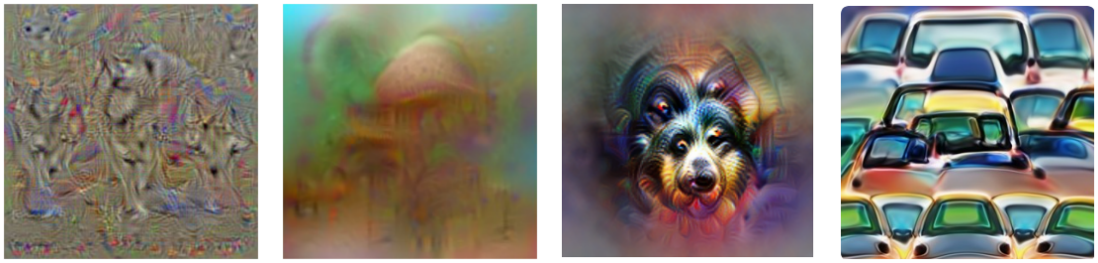

Abbildung 2.1: Beispiele von verschiedenen Interpretationen im Laufe der Forschung. Von links nach rechts: Simonyan et al. [23] (2013) Klassen-Aktivierung ohne Regularisierung, Mahendran und Vedaldi [14] (2014) Merkmalsumkehrung regularisiert durch totale Varianz, Olah et al. [20] (2017) Aktivierung eines tiefen Neurons regularisiert durch eine inverse Fourier-Transformation und Transformationsrobustheit, Mordvintsev et al. [16] (2018) Aktivierung eines tiefen Filters regularisiert mit einem *Compositional Pattern Producing Network*.

Bilder mit einem speziellen Pixel-Muster überlagert, meist so schwach, dass die Veränderung für einen Menschen nicht zu erkennen ist. Diese Pixel-Muster haben keine oder nur sehr wenig Bedeutung für einen Menschen, aber einen sehr starken Effekt auf die Vorhersage des neuronalen Netzes.
Feindliche Beispiele sind zwar relevant, um die Angreifbarkeit von neuronalen Netzen zu beurteilen, sie eignen sich allerdings nicht für die Interpretation, da sie für menschliche Betrachter keine Bedeutung haben. Um dieses Problem zu beheben, werden Regularisierungsmethoden verwendet.

2.2 Regularisierungsmethoden

Regularisierungsmethoden für Interpretationsmethoden sind ein wichtiger Bestandteil der aktuellen Forschung. In der Arbeit von Olah et al. [20] wird zwischen drei verschiedenen Varianten der Regularisierung unterschieden:

Bestrafen von hohen Frequenzen: Da vor allem hohe Frequenzen[1], also sich stark verändernde Farbwerte, in den Interpretationen auffallen, ist es naheliegend, diese direkt zu unterbinden. Dies kann explizit durch das Bestrafen der Varianz beieinander liegender Pixel geschehen (totale Varianz) [14], oder implizit durch das (digitale) Verwischen des Interpretationsbildes bei jedem Optimierungsschritt [19]. Leider wirken sich diese Regularisierungstechniken auch auf normale Kanten[2] aus, welche in Interpretationsbildern durchaus vorkommen können. Ein Beispiel der Regularisierung mit totaler Varianz ist in der Abbildung 2.1 (Mitte-links) zu sehen.

Transformationsrobustheit: Bei dieser Regularisierungtechnik werden die Interpretationsbilder bei jedem Optimierungsschritt zufällig verändert. Die Veränderungen erfolgen durch Bildmanipulationen wie Verschiebung, Skalierung und Rotation. Hierdurch können nur Interpretationsbilder entstehen, welche auch bei leichten Transformationen das neuronale Netz wie vorgesehen beeinflussen. Diese Art der Regularisierung lässt sich besonders gut mit anderen Regularisierungsmethoden kombinieren.

Vorbedingungen Bei der Verwendung von Vorbedingungen wird im Lernprozess der Interpretation nicht das Interpretationsbild selbst optimiert, sondern die Parameter einer Vorbedingung, welche das Interpretationsbild erzeugt (siehe Abbildung 2.2). Als starke Vorbedingungen werden neuronale Netze wie *Compositional Pattern Producing Networks* (CPPN) [25] oder *Deep Generator Networks* (DGN) [17] verwendet. Ein Beispiel der Regularisierung mit CPPNs ist in der Abbildung 2.1 (rechtes Bild) zu sehen. Besonders vortrainierte Vorbedingungen wie DGNs erzeugen sehr fotorealistische Interpretationen. Bei diesen Interpretationen ist allerdings nicht ersichtlich, welche der visualisierten Informationen aus der Vorbedingung und welche aus dem zu interpretierenden neuronalen Netz erzeugt wurden.
Alternativ können schwächere Vorbedingungen verwendet werden, welche das Interpretationsbild nicht allzu stark einschränken. Als sehr gut hat sich die inverse Fourier-Transformation zusammen mit der Transformationsrobustheit bewährt, welche in [20] verwendet wird (siehe Abbildung 2.1 Mitte-rechts). Die inverse Fourier-Transformation

[1] Bilder werden für gewöhnlich im Bildbereich, zum Beispiel durch die Farbwerte Rot, Gelb und Blau dargestellt. Mithilfe der Fourier-Transformation können Bilder aber auch im Frequenzbereich modelliert werden. Bei dieser Art der Modellierung beschreibt jede Frequenz die Veränderung der Farbwerte im Bild. Während niedrige Frequenzen einen kontinuierlichen Farbübergang erzeugen, beschreiben hohe Frequenzen sehr starke Farbübergänge, bei denen sich die Farbwerte benachbarter Pixel bereits deutlich unterscheiden.

[2] Als Kanten werden in einem Bild starke Helligkeits- oder Farbveränderungen bezeichnet.

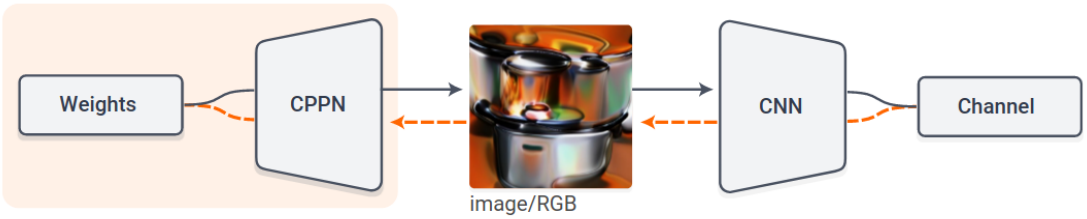

Abbildung 2.2: Interpretation mit einem CPPN als Vorbedingung [16]. In der Vorwärtsrichtung (schwarze Pfeile) wird auf Basis der Gewichte (weights) des CPPNs ein Bild erzeugt. Dieses wird als Eingangsbild des Faltungsnetzes (CNN) verwendet, um einen bestimmten Filter (channel) zu beeinflussen. Anschließend werden mithilfe der Fehlerrückführung (orange Pfeile) die CPPN-Gewichte angepasst, um den gewünschten Effekt im Faltungsnetz zu verstärken.

verschiebt das Optimierungsproblem vom Bildbereich in den Frequenzbereich. Dies vereinfacht das Optimierungsproblem erheblich, da die zu optimierenden Parameter im Frequenzbereich weniger miteinander korrelieren[3].

2.3 Was wird visualisiert?

Während am Anfang der Interpretationsforschung noch einzelne Neuronen des neuronalen Netzes untersucht wurden [7], stellte sich im Laufe der Zeit heraus, dass vor allem Kombinationen von mehreren Neuronen von Bedeutung sind [5] [10] [20] [21]. In der Arbeit von Kim et al. [10] wird gezeigt, dass sich Konzepte wie beispielsweise die Hunderasse Corgi oder sibirischer Husky in den tiefen Schichten eines neuronalen Netz wiederfinden. Diese Konzepte werden als Linearkombination neuronaler Aktivierungen repräsentiert, welche auch als *Konzept-Aktivierungsvektor* bezeichnet werden.
Die Theorie, dass für Menschen verständliche Konzepte in den tiefen Schichten eines neuronalen Netzes als Aktivierungvektoren dargestellt werden, ist ein wichtiger Grundbaustein dieser Arbeit und wird dazu verwendet, eine neuronale Sprache und damit die Qualitätsmetrik für Interpretationsmethoden zu entwickeln.

[3] In einer für den Bildbereich typischen RGB-Darstellung korrelieren die einzelnen Pixelwerte stark miteinander, zum Beispiel bei einer Veränderung der Helligkeit, von der alle Farbwerte im gleichen Maße beeinflusst werden.

3 Das Ziel der Regularisierung

Die Regularisierung von Interpretationsmethoden wird in vielen wissenschaftlichen Arbeiten [5] [8] [14] [16] [17] [20] [21] [29] erforscht oder verwendet. Hierbei wird die Regularisierung übereinstimmend dafür verwendet, bedeutungsvollere Interpretationen zu erzeugen. Konkret geht es oft um die Beseitigung hochfrequenter Muster [14] [20], welche das neuronale Netz stark beeinflussen, jedoch für Menschen keine Bedeutung haben.
In dieser Arbeit soll der Grund, weshalb Regularisierungsmethoden benötigt werden, und was sie leisten sollen genauer definiert werden.
Das Problem bei der Interpretation von neuronalen Netzen ist, dass sowohl der Mensch als auch neuronale Netze eine Vielzahl von Eingangswerten verarbeiten können. Besonders im Bildbereich kann ein durchschnittlicher Mensch eine unbestimmbar große Anzahl von Mustern erkennen und deuten.
Ein gut trainiertes neuronales Netz kann in der Regel alle Daten deuten, die im Trainingsdatensatz enthalten sind oder diesem ähneln. Wie die Forschungen zu feindlichen Beispielen [19] [28] und Ergebnisse der Interpretation neuronaler Netze zeigen, können neuronale Netze aber auch noch viele weitere Eingangsdaten deuten, die keine Ähnlichkeit mit realen Daten haben.
Damit Interpretationen von Nutzen sind, müssen diese sowohl für das neuronale Netz als auch für Menschen von Bedeutung sein. Hieraus ergibt sich eine konkrete Anforderung an Regularisierungsmethoden:

Eine gute Regularisierung soll die Interpretationsmethode so einschränken, dass die hiermit erzeugten Interpretationen nur Konzepte[1] zeigen, welche sowohl vom neuronalen Netz als auch vom Menschen verstanden werden.

Wie diese Anforderung genau umgesetzt werden kann, wird im nächsten Kapitel erläutert.

[1] Als Konzept wird in dieser Arbeit die abstrakte Idee eines Objektes bezeichnet. Beispiel Konzepte aus dem Bereich des (maschinellen) Sehens sind: Augen, Streifen, Fell, Licht und viele mehr.

4 Die Sprache der neuronalen Netze

In diesem Kapitel wird erklärt, wie die in Kapitel 3 definierten Anforderungen an Regularisierungsmethoden praktisch umgesetzt werden können. Hierfür wird eine Methode entwickelt, um diejenigen Konzepte zu finden, die sowohl vom Menschen als auch vom neuronalen Netz verstanden werden. Auf Basis dieser Konzepte wird anschließend die Qualitätsmetrik für Interpretationen von neuronalen Netzen entwickelt.

4.1 Die Basis der neuronalen Sprache

Da der Begriff *Konzept* nicht strikt definiert ist und auch keine exakte Abgrenzung zwischen verschiedenen Konzepten möglich ist, kann diese Aufgabe nicht analytisch gelöst werden. Wie auch bei anderen typischen Aufgabenstellungen der künstlichen Intelligenz kann die Lösung jedoch mithilfe einer größeren Datenmenge beschrieben werden.
Da der Trainingsdatensatz des zu interpretierenden neuronalen Netzes, menschlich erstellt wurde und das gesamte Wissen des neuronalen Netzes ausschließlich aus dem Trainingsdatensatz extrahiert wurde, enthält dieser nur Konzepte, die sowohl vom Menschen als auch vom neuronalen Netz verstanden werden. Dies macht den Trainingsdatensatz zu einer guten Basis für die Entwicklung einer gemeinsamen Sprache zwischen Mensch und neuronalem Netz.
Basierend auf diesem Ansatz kann die Anforderung an Regularisierungsmethoden aus Kapitel 3 umformuliert werden:

Eine gute Regularisierung soll die Interpretationsmethode so einschränken, dass die hiermit erzeugten Interpretationen nur Konzepte zeigen, die auch im Trainingsdatensatz vorkommen.

Hiermit ist nicht gemeint, dass die Interpretationen genauso aussehen sollen wie einzelne Trainingsbeispiele, sondern dass sich jedes in einer Interpretation vorkommende Konzept

auch irgendwo in den Trainingsdaten wiederfindet. Wichtig ist aber, dass der Trainingsdatensatz nicht zum Erstellen der Interpretation verwendet werden darf, da hierbei nicht bekannt ist, welche Informationen aus dem Trainingsdatensatz und welche aus dem zu interpretierenden neuronalen Netz stammen.
Aus dieser Anforderung ergeben sich zwei Schwierigkeiten. Zum einen muss eine Metrik gefunden werden, welche die Ähnlichkeit von abstrakten Konzepten errechnen kann. Zusätzlich muss eine effiziente Methode entwickelt werden, um aus allen in den Trainingsdaten vorkommenden Konzepten dasjenige herauszusuchen, welches einem Konzept aus einer Interpretation am ähnlichsten ist.

4.2 Die Suche nach gemeinsamen Konzepten

In der Bildverarbeitung gibt es verschiedene Metriken zum Berechnen der Ähnlichkeit von zwei Bildern. Dazu gehören unter anderem der *root-mean-square error* (RMSE), *Peak signal-to-noise ratio* (PSNR), *Structural similarity* (SSIM) [34], *Feature-based similarity index* (FSIM) [32] und *Information theoretic-based Statistic Similarity Measure* (ISSM) [4] Algorithmus.
Keine der mir bekannten Metriken eignet sich allerdings dafür, den semantischen Kontext von Bildern und somit Konzepten, die in Bildern vorkommen, miteinander zu vergleichen. Somit stellt sich die Frage, wie verschiedene Konzepte definiert, voneinander differenziert und dessen Ähnlichkeit bestimmt werden können.
Um dieses Problem zu lösen, wird auf das gelernte Wissen des zu interpretierenden neuronalen Netzes zurückgegriffen. Im Trainingsprozess hat das neuronale Netz bereits eine Vielzahl von verschiedenen Konzepten gelernt und die benötigten Informationen in seinen Gewichten gespeichert. In den verschiedenen Schichten des neuronalen Netzes werden die Konzepte als Linearkombination von neuronalen Aktivierungen dargestellt. Diese werden im folgenden Text als Aktivierungsvektor bezeichnet. Es wird angenommen, dass die Komplexität der Konzepte mit zunehmender Tiefe im neuronalen Netz ansteigt.
Um herauszufinden, ob die verschiedenen Konzepte einer Interpretation ähnlich zu denen aus dem Trainingsdatensatz sind, können die Aktivierungsvektoren der tiefen Schichten des neuronalen Netzes miteinander verglichen werden. Wird beispielsweise das Bild eines Zebras als Eingang eines Klassifizierungs-Netzes verwendet (siehe Abbildung 4.1 oberer Teil), entstehen in den verschiedenen Schichten des Netzes Aktivierungsvektoren, welche einzelne, im Bild vorkommende Konzepte repräsentieren. In den ersten Schichten werden

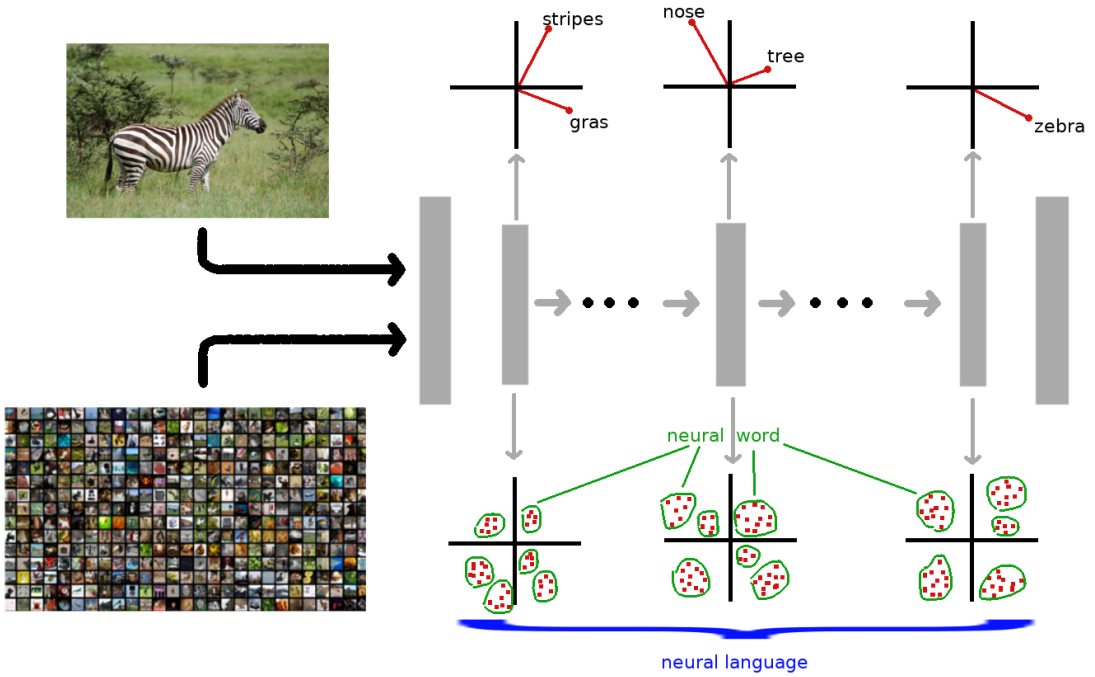

Abbildung 4.1: Diese Abbildung zeigt die Entwicklung der neuronalen Sprache. Bei einem einzelnen Bild repräsentieren die internen Aktivierungsvektoren des neuronalen Netzes die Merkmale, welche im Bild erkannt werden. Indem alle durch den Trainingsdatensatz entstehenden Aktivierungsvektoren geclustert werden, kann die typische Verteilung der Aktivierungsvektoren approximiert werden. Jedes Cluster wird als neuronales Wort bezeichnet. Alle Cluster zusammen ergeben die neuronale Sprache.

einfache Konzepte wie die Streifen des Zebras oder das Gras im Hintergrund erkannt. Diese werden in tieferen Schichten zu komplexen Konzepten wie beispielsweise der Schnauze des Zebras oder den Bäumen zusammengesetzt. In der letzten tiefen Schicht entsteht eine Repräsentation des Zebras, auf dessen Basis die Klassifizierung durchgeführt wird.
Um alle im Trainingsdatensatz vorkommenden Konzepte zu finden, werden die internen Aktivierungsvektoren für jedes Bild des Trainingsdatensatzes errechnet und gespeichert. Wird nun ein Interpretationsbild als Eingang des neuronalen Netzes verwendet, kann überprüft werden, ob die hierdurch entstehenden internen Aktivierungsvektoren denen des Trainingsdatensatzes ähneln. Findet sich für jeden Aktivierungsvektor der Interpretation ein sehr ähnlicher Aktivierungsvektor aus den Trainingsdaten, so ist dies eine gute Interpretation, da sie nur aus Konzepten besteht, die auch im Trainingsdatensatz vorkommen.

Da es allerdings sehr rechenintensiv ist, alle Aktivierungsvektoren einer einzelnen Interpretation mit allen Aktivierungsvektoren des gesamten Trainingsdatensatzes zu vergleichen, wird die Verteilung der Aktivierungsvektoren im neuronalen Netz approximiert.
Hierfür werden für jede tiefe Schicht des neuronalen Netzes die durch die Trainingsdaten entstehenden Aktivierungsvektoren errechnet (siehe Abbildung 4.1 unterer Teil). Diese werden, abhängig von ihrer Entfernung zueinander, in verschiedene Cluster unterteilt. Werden beispielsweise verschiedene Arten von Graslandschaften durch ähnliche Aktivierungsvektoren in einer der vorderen Schichten des neuronalen Netzes repräsentiert, können diese aufgrund ihres geringen Abstands zueinander zu einem Cluster zusammengefasst werden, welches das generelle Konzept Gras darstellt.
Angelehnt an die natürliche Sprache, bei der ein Wort eine Vielzahl von ähnlichen Konzepten beschreibt, wird jedes dieser Aktivierungsvektor-Cluster als *neuronales Wort* bezeichnet. Alle im Netz vorkommenden neuronalen Wörter ergeben die *neuronale Sprache*.

4.3 Die Qualität der Interpretationen

Die neuronale Sprache ermöglicht es nun, die Qualität einer Interpretation zu messen. Eine gute Interpretation besteht ausschließlich aus Wörtern der neuronalen Sprache. Dies kann gemessen werden, indem die Konzepte (beziehungsweise die Aktivierungsvektoren) einer Interpretation den Wörtern der neuronalen Sprache zugeordnet werden und überprüft wird, inwieweit sie diesen ähneln.
Da diese Qualitätsmetrik die Qualität einer Interpretation aus der Sicht des neuronalen Netzes darstellt, heißt es nicht zwangsläufig, dass die Qualitätsmetrik mit dem Qualitätsempfinden eines Menschen korreliert. Bei einem sorgfältig zusammengestellten Trainingsdatensatz und einem gut trainierten Netz sollte dies aber der Fall sein.
Im nächsten Kapitel wird experimentell gezeigt, wie Regularisierungsmethoden für Interpretationsmethoden auf Basis dieser Qualitätsmetrik optimiert werden können.

5 Experiment 1: Optimierung von Interpretationsmethoden für MNIST-Daten

Durch die Entwicklung der Qualitätsmetrik für Interpretationen von neuronalen Netzen aus Kapitel 4 ist es nun erstmals möglich, Interpretationsmethoden automatisiert zu verbessern. Anstatt die Regularisierung der Interpretationsmethoden manuell auszuwählen und zu parametrisieren, können automatisierte Suchmethoden verwendet werden, um eine geeignete Regularisierung für das zu interpretierende neuronale Netz zu finden.
In diesem Kapitel wird die Suche nach einer geeigneten Regularisierungsmethode experimentell durchgeführt. Hierbei geht es noch nicht darum, ein komplexes Regularisierungsproblem zu lösen, sondern nur darum, die Qualitätsmetrik für Interpretationen auf einem leichten Datensatz und einem nicht allzu komplexen neuronalen Netz zu testen.

5.1 Material / Methoden

In diesem Abschnitt werden die für das Experiment verwendeten Daten und Algorithmen beschrieben.

Der Datensatz:
Für dieses Experiment wird der Datensatz *MNIST* [12] verwendet. Dieser enthält 60.000 Trainingsdaten und 10.000 Testdaten mit Abbildungen der handgeschriebenen Zahlen von null bis neun. Die Klassifizierung des MNIST-Datensatzes mit neuronalen Netzen ist eine einfache Aufgabe, weshalb erwartet wird, dass auch die Interpretationen aus leicht zu verstehenden Konzepten zusammengesetzt werden.

Das neuronale Netz:
Die Qualitätsmetrik für Interpretationen von neuronalen Netzen geht davon aus, dass

die Merkmale, welche das Netz aus den Trainingsdaten erlernt, dieselben sind, die auch ein Mensch intuitiv zum Lösen der Aufgabe verwendet. Ob dies tatsächlich der Fall ist, ist schwer nachzuweisen, da sowohl künstliche neuronale Netze als auch das menschliche Gehirn noch zu großen Teilen unerforscht sind. Für dieses Experiment wird davon ausgegangen, dass ein neuronales Netz, welches die Klassifikationsaufgabe gut löst, wahrscheinlich auch viele Merkmale benutzt, mit denen ein Mensch die Aufgabe lösen würde. Aus diesem Grund wurde die Netzarchitektur von Ghouzam [2] übernommen, die besonders gute Ergebnisse auf dem MNIST-Datensatz erzielt. Aufgrund seiner Aufgabe wird dieses neuronale Netz im folgenden Text als Zahlen-Netz bezeichnet. Die Architektur des Zahlen-Netzes wird im Anhang A.1 beschrieben.

Der Cluster-Algorithmus zum Erstellen der neuronalen Sprache:
Um die Sprache von neuronalen Netzen zu approximieren, wird ein geeigneter Cluster-Algorithmus ausgewählt. Basierend auf der Grundannahme, dass sich Konzepte innerhalb des neuronalen Netzes umso ähnlicher sind, desto geringer der Abstand ihrer Aktivierungsvektoren ist, müssen die Cluster auf einer Abstandsmetrik basieren. Da die Sprache des neuronalen Netzes mithilfe des Trainingsdatensatzes approximiert wird, sollte der Cluster-Algorithmus auch für große Datenmengen geeignet sein.
Für besonders große Datensätze, die nicht vollständig im Computerspeicher geladen werden können, oder bei der Verwendung von Daten-Vermehrungsmethoden sollte der Cluster-Algorithmus ein Training im Mini-Batch-Verfahren ermöglichen.
Ein Algorithmus, der all diese Anforderungen erfüllt, ist der BIRCH Cluster-Algorithmus [33]. BIRCH clustert Datenpunkte in einer Baum-Datenstruktur auf Basis der euklidischen Distanz. Hierbei wird jedes Cluster durch den Mittelpunkt der zum Cluster gehörenden Datenpunkte beschrieben, wodurch die Aktivierungsvektoren, welche zum Erstellen der neuronalen Sprache benötigt werden, nicht gespeichert werden müssen. Hierdurch ist der Speicherverbrauch einer neuronalen Sprache deutlich geringer als der Speicherverbrauch der Aktivierungsvektoren, aus denen die Sprache erzeugt wurde. Der wichtigste Parameter des BIRCH-Algorithmus ist der Radius eines Clusters, auch Schwellenwert (*eng. threshold*) genannt, der bestimmt, bis zu welcher Entfernung ein Datenpunkt noch zu einem Cluster gehört. Dieses Maß entspricht sinngemäß der minimalen semantischen Ähnlichkeit verschiedener Entitäten desselben neuronalen Wortes.
Damit die Qualität der Interpretationen so genau wie möglich gemessen werden kann, muss auch die Approximation der neuronalen Sprache möglichst genau sein. Ein kleinerer Schwellenwert wirkt sich positiv auf die Genauigkeit der Approximation der neuronalen Sprache aus, indem die Baum-Datenstruktur des BIRCH-Algorithmus vergrößert wird.

Hierdurch steigen aber auch die benötigten Speicher- und Rechenkapazitäten. Da unbekannt ist, wie gut sich die internen Aktivierungen eines neuronalen Netzes mit dem BIRCH-Algorithmus approximieren lassen, wird der Schwellenwert (beziehungsweise der Abstand zwischen verschiedenen neuronalen Wörtern) so klein gewählt, wie es mit den verfügbaren Speicher- und Rechenkapazitäten zu vereinbaren ist.

Die Qualitätsmetrik:
Um die Qualität einer Interpretation zu berechnen, wird für jedes Konzept k der Interpretation der Abstand zum nächstgelegenen neuronalen Wort benötigt. Das nächstgelegene neuronale Wort w^* kann mithilfe des Cluster-Algorithmus ermittelt werden. Da sowohl das Konzept k als auch das neuronale Wort w^* als Aktivierungsvektor dargestellt werden, kann der Abstand mithilfe der normierten Vektordifferenz errechnet werden.

$$\Delta(k, w^*) = |k - w^*| \tag{5.1}$$

Der Abstand der Konzepte zu den nächstgelegenen neuronalen Wörtern wird hierbei für jede Schicht des neuronalen Netzes separat berechnet.
Da die Konzept-Aktivierungsvektoren in jeder Schicht andere Wertebereiche haben, können sich auch die Abstände zwischen Interpretations-Konzept und nächstgelegenem neuronalen Wort stark unterscheiden, unabhängig davon, wie ähnlich sich die Konzepte sind. Um die Qualitätsmetrik in einen verständlichen Zahlenbereich zu verschieben, werden die Testdaten des MNIST-Datensatzes als Referenzdaten verwendet.
Sorgfältig ausgewählte Testdaten werden als Messlatte für hochqualitative Interpretationen gesehen, da sie dieselben Konzepte zeigen, die auch im Trainingsdatensatz vorkommen und sowohl vom neuronalen Netz als auch vom Menschen verstanden werden.
Die Qualität einer Interpretation in der Schicht l ergibt sich aus dem Quotienten des mittleren Abstands der Interpretations-Konzepte zur neuronalen Sprache $\overline{\Delta I_l}$ und dem mittleren Abstand der Referenzdaten-Konzepte zur neuronalen Sprache $\overline{\Delta R_l}$.

$$Q_l = \frac{\overline{\Delta R_l}}{\overline{\Delta I_l}} \tag{5.2}$$

Die gesamte Qualität der Interpretation ergibt sich aus dem Mittelwert der Interpretationsqualität aller Schichten.
Die Werte der Qualitätsmetrik können wie folgt gedeutet werden: Eine Qualität von eins ist perfekt und bedeutet, dass die Interpretation den Trainingsdaten genauso ähnlich ist

wie die Testdaten. Je kleiner die Interpretationsqualität ist, desto unähnlicher ist die Interpretation dem Trainingsdatensatz. Der niedrigste Wert geht gegen null, wobei null nicht zu erreichen ist. Zwar kann die Qualitätsmetrik auch Werte größer als eins annehmen, wenn die Interpretation den Trainingsdaten ähnlicher ist als die Referenzdaten, davon wird allerdings nicht ausgegangen.

Die Optimierung der Interpretationsmethode:
In diesem Experiment wird versucht, eine gute Regularisierung für die Interpretationsmethode der Merkmalsumkehrung [14] zu finden. Bei der Merkmalsumkehrung wird versucht, die internen Aktivierungen einer tiefen Schicht des neuronalen Netzes, welche durch ein bestimmtes Eingangsbild entstehen, zu rekonstruieren. Hierfür wird ein Eingangsbild erlernt, welches dieselben Aktivierungen erzeugt. Das rekonstruierte Bild soll darstellen, welche Informationen dem Netz über das Originalbild in der besagten Schicht vorliegen. Indem die internen Aktivierungen aus Bildern des Trainingsdatensatzes erzeugt werden wird sichergestellt, dass die in der Interpretation vorkommenden Konzepte auch im Trainingsdatensatz vorkommen und somit die Qualitätsmetrik angewendet werden kann.
Um eine gute Regularisierung zu finden, wird die letzte tiefe, voll vernetzte Schicht des Zahlen-Netzes ausgewählt. Diese Schicht beinhaltet alle Informationen, auf dessen Basis die Klassifizierung durchgeführt wird. Zudem ist sie aufgrund ihrer Tiefe und der damit einhergehenden Abstraktion der Informationen schwer zu interpretieren. Die Tatsache, dass die Schicht voll vernetzt ist, erschwert ihre Interpretierbarkeit zusätzlich, da bei voll vernetzten Schichten die räumliche Dimension und somit die Einschränkung, eine im Raum verteilte Komposition von Konzepten darzustellen, aufgelöst wird.
Vor allem eignet sich die Schicht aber, da bei ihrer Interpretation alle tiefen Schichten des neuronalen Netz beteiligt sind. Eine qualitative Interpretation der letzten tiefen Schicht muss demnach auf jeder tiefen Schicht des neuronalen Netzes realistische Aktivierungen erzeugen.
Aufgrund früherer Erfahrungswerte [30] wird eine inverse Fourier-Transformation als Vorbedingung der Interpretationsmethode in Kombination mit Transformationsrobustheit und totaler Varianz als Regularisierung verwendet. Durch die Vorbedingung der inversen Fourier-Transformation wird der Optimierungsraum vom Farb- in den Frequenzbereich verschoben. Hierdurch wird bei der Optimierung die Korrelation, welche zwischen RGB Pixeln besteht, aufgelöst und somit das Optimierungsproblem stark vereinfacht.
Mithilfe der Transformationsrobustheit werden nur Interpretationen zugelassen, die auch bei leichten Transformationen das neuronale Netz wie vorgesehen beeinflussen. Als Transformationen werden Verschiebungen, Skalierungen oder Rotationen des Interpretations-

bildes verwendet.
Um hochfrequente Muster in den Interpretationsbildern weiter zu unterdrücken, wird zusätzlich noch die Metrik der totalen Varianz angewendet.
Sowohl die totale Varianz als auch die verschiedenen Transformationen verfügen über Parameter, um die Stärke ihres Effektes zu bestimmen. Diese Parametrisierung wird mithilfe einer Bayesschen Optimierung auf Basis der Interpretationsqualität optimiert. Bei der Bayesschen Optimierung wird ein Modell erstellt, welches den Zusammenhang zwischen verschiedenen Parametern und einer Qualitätsmetrik modelliert. Auf Basis dieses Modells werden neue Parameter ausprobiert und die Ergebnisse dazu genutzt, das Modell zu verbessern. Hierdurch konvergiert die Bayesschen Optimierung deutlich schneller als zufälliges Ausprobieren oder eine Rastersuche, ist dabei aber unkomplizierter und weniger rechenintensiv als komplexe Algorithmen wie zum Beispiel die evolutionäre Optimierung. Die Parameter der einzelnen Transformationen und der totalen Varianz können bei der Optimierung sowohl verstärkt als auch ganz abgeschaltet werden.

Da einerseits die Qualität der Interpretationen verbessert werden, aber andererseits auch sichergestellt werden soll, dass die Interpretationsmethoden das neuronale Netz auf die gewünschte Weise beeinflussen, muss die Bayessche Optimierung beide Eigenschaften optimieren.
Jede lernbasierte Interpretationsmethode beinhaltet eine Verlustfunktion, die misst, wie stark ein gewünschter Effekt im neuronalen Netz auftritt. Hierdurch kann die Beeinflussung des Netzes durch den Verlustwert einer Interpretation ausgedrückt werden. Abhängig von der Art der Interpretation und der Schicht des Netzes, in der interpretiert wird, kann der Verlustwert sehr unterschiedliche Werte annehmen.
Um den Verlustwert mit der Qualitätsmetrik zu kombinieren, muss dieser in einen festen Wertebereich verschoben werden. Hierfür wird, wie auch bei der Qualitätsmetrik, ein Referenzwert benötigt. Als Messlatte für die Interpretation, welche das neuronale Netz am genausten beeinflusst, wird die Interpretation ohne Regularisierung genommen. Die relative Genauigkeit einer Interpretation wird als Quotient des Referenzverlustwerts, also des Interpretations-Verlustwertes ohne Regularisierung, und des Verlustwertes der regularisierten Interpretation berechnet.

$$\text{Genauigkeit} = \frac{\text{Referenz Verlustwert}}{\text{Verlustwert}} \tag{5.3}$$

Die beste relative Genauigkeit einer regularisierten Interpretation ist somit eine mit dem Wert eins, da hier das Netz genauso gut beeinflusst wird wie ohne Regularisierung.

Nun nehmen sowohl die Qualität als auch die Genauigkeit der Interpretation Werte zwischen null und eins an. Um beide Werte zu optimieren, wird der Nutzen einer Interpretation definiert, der wie folgt errechnet wird:

$$\text{Nutzen} = \frac{\text{Qualität} + \text{Genauigkeit}}{2} \tag{5.4}$$

Für dieses Experiment wird zur Errechnung des Nutzens die Qualität und Genauigkeit gleich gewichtet. Je nach Präferenz kann aber auch eine andere Gewichtung vorgenommen werden.
Um die Bayessche Optimierung durchzuführen, wird die Entwicklerplattform *Weights & Biases* verwendet und versucht, eine Regularisierung zu finden, die den Nutzen der Interpretationsmethode maximiert.

5.2 Durchführung und Ergebnisse

In diesem Abschnitt werden die Durchführung und die Ergebnisse des Experiments beschrieben. Die Durchführung des Experiments besteht aus dem Erstellen der neuronalen Sprache, der automatisierten Suche nach einer zufriedenstellenden Regularisierung und der Evaluation der Qualitätsmetrik durch die Darstellung von verschiedenen Interpretationen, und dem Abgleich ihrer Qualität mit dem menschlichen Qualitätsempfinden.

Das Erstellen der neuronalen Sprache:
Die neuronale Sprache für das Zahlen-Netz wird für jede Schicht des neuronalen Netzes approximiert. Um zu beeinflussen, wie viele Cluster und somit neuronale Wörter der BIRCH-Algorithmus erzeugt, muss ein Schwellenwert gewählt werden, der bestimmt, ab welchem Abstand zwei Aktivierungsvektoren zu unterschiedlichen neuronalen Wörtern zählen. Geeignete Schwellenwerte werden mithilfe des Versuch-und-Irrtum-Prinzips gewählt. Die Schwellenwerte und die hieraus resultierende Anzahl der neuronalen Wörter pro Schicht des Zahlen-Netzes werden in der folgenden Tabelle 5.1 dargestellt.

Name der Schicht	Schwellenwert	Anzahl neuronale Wörter
conv1	0.3	2244
conv2	0.4	2395
conv3	0.5	3677
conv4	0.5	12807
dense1	1.0	45187

Tabelle 5.1: In dieser Tabelle wird für jede Schicht des Zahlen-Netzes der verwendete Schwellenwert des BIRCH-Algorithmus und die aus dem Cluster-Verfahren resultierende Anzahl der neuronalen Wörter angegeben.

Die Optimierung der Interpretationsmethode:
Um eine gute Regularisierung zu finden, werden die Parameter der Transformationsrobustheit und der totalen Varianz darauf optimiert, nützliche, also sowohl genaue als auch qualitative, Interpretationen zu erzeugen. Als mögliche Transformationen werden in ebendieser Reihenfolge eine Verschiebung, Rotation, Skalierung und eine weitere Verschiebung gewählt. Die möglichen Werte der Parameter werden in der folgenden Tabelle 5.2 aufgelistet.

Regularisierung	Mögliche Werte
Verschiebung 1	0 - 6 Pixel
Rotation	0 - 20°
Skalierung	0.8 - 1.2
Verschiebung 2	0 - 6 Pixel
totale Varianz	beta: 4 - 16 oder keine TV

Tabelle 5.2: In dieser Tabelle finden sich die möglichen Konfigurationswerte für die Transformationsrobustheit und die totale Varianz, welche bei der Optimierung der Interpretationsmethode verwendet werden.

In der Abbildung 5.1 ist die Nützlichkeit der verschiedenen Interpretationsmethoden, relativ zu der Zeit, in der sie von der Bayesschen Optimierung ausprobiert wurden, abgebildet. Während die ersten Interpretationsmethoden noch keinen sonderlich großen Nutzen aufweisen, werden schnell geeignetere Regularisierungen gefunden, die den Nutzen der Interpretation verbessern.
Die Abbildung 5.2 zeigt den Nutzen der verschiedenen Interpretationsmethoden in Abhängigkeit von den Regularisierungs-Parametern. Hier ist zu erkennen, dass die nützlichs-

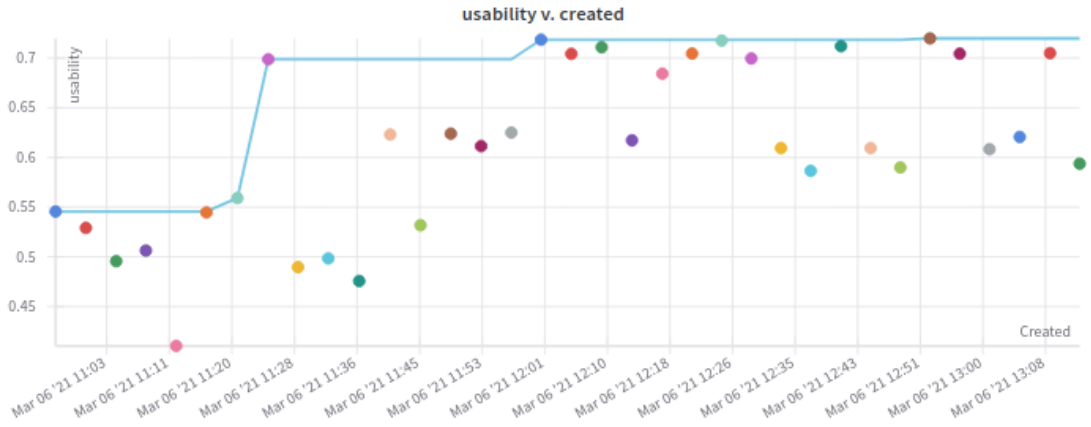

Abbildung 5.1: Darstellung des Nutzens (usability) der verschiedenen Interpretationen in Relation zu der Zeit, in der sie ausprobiert wurden. Hier ist zu sehen, dass die ersten von der Bayesschen Optimierung ausprobierten Interpretationsmethoden noch keinen sonderlich großen Nutzen aufweisen. Im Laufe der Optimierung verbessern sich die Interpretationsmethoden, bis nach etwa der Hälfte der Zeit eine der besten Regularisierungen gefunden wird, wodurch sich der Nutzen der Interpretationsmethode im weiteren Verlauf kaum noch verbessern lässt.

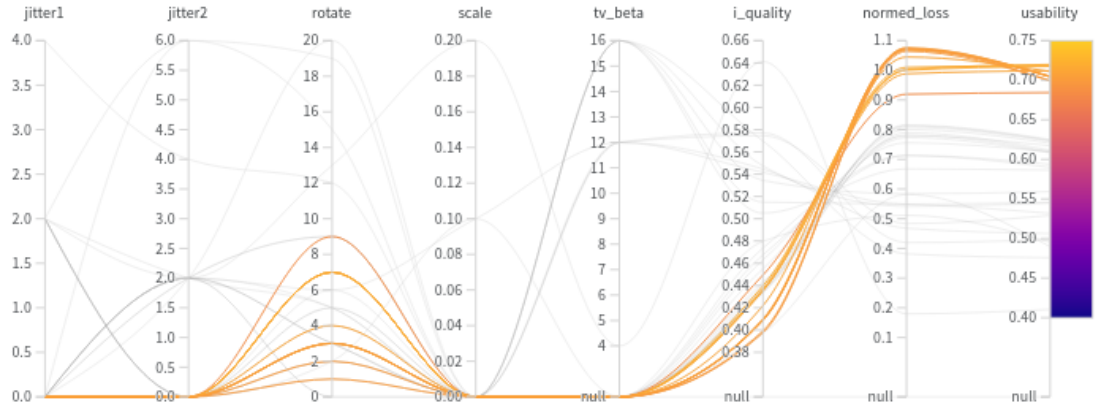

Abbildung 5.2: Darstellung des Nutzens (usability) einer Interpretationsmethode in Abhängigkeit von den gewählten Regularisierungs-Parametern (die inverse Fourier-Transformation wird nicht mit abgebildet, da diese nicht parametrisierbar ist). Hier ist zu sehen, dass die nützlichsten Regularisierungsmethoden für MNIST nur aus der inversen Fourier-Transformation und einer zufälligen Rotation bestehen. Der hohe Nutzen ergibt sich aus einer niedrigen Qualität (i_quality) und einer sehr hohen Genauigkeit (normed_loss).

Regularisierung	I1	I2	I3	I4	I5
Mit inverser FFT	nein	ja	ja	ja	ja
Verschiebung 1 in [Pixel]	-	8	-	4	4
Rotation [°]	-	-	10	10	10
Skalierung	-	-	-	0.9-1.1	0.9-1.1
Verschiebung 2 [Pixel]	-	-	-	2	2
Totale Varianz beta	-	-	-	-	10

Tabelle 5.3: Regularisierungs-Parameter der verschiedenen Interpretationsmethoden.

ten Regularisierungsmethoden für MNIST nur aus der inversen Fourier-Transformation und einer zufälligen Rotation bestehen. Der höchste Nutzen wird mit einer maximalen Rotation von 7° erreicht, doch auch Rotationen mit etwas niedrigeren oder höheren Werten erreichen einen hohen Nutzen. Eine genauere Analyse der Werte zeigt, dass sich der hohe Nutzen aus einer niedrigen Qualität, aber einer sehr hohen Genauigkeit der Interpretationen ergibt. Auf weitere Vor- und Nachteile verschiedener Regularisierungen wird im folgenden Abschnitt weiter eingegangen.

Abgleich der Interpretationsqualität mit menschlicher Intuition:
Die in diesem Experiment verwendete Interpretationsqualität wird aus der Perspektive des Zahlen-Netzes errechnet. Um zu überprüfen, ob die Interpretationsqualität auch mit menschlicher Intuition korreliert, werden mit fünf verschiedenen Regularisierungsmethoden Interpretationen erzeugt und ihre Interpretationsergebnisse, ihre Genauigkeit und Qualität dargestellt. Die hierfür verwendeten Regularisierungsmethoden *I1* bis *I5* sind in der Tabelle 5.3 aufgelistet. Diese werden anhand von Merkmalsumkehrungen auf der Schicht *dense1* des Zahlen-Netzes und durch Klassen-Aktivierungen getestet. In den Abbildungen 5.3 und 5.4 werden die verschiedenen Interpretationsergebnisse abgebildet. Die Genauigkeit und die Interpretationsqualität sind in den Tabellen 5.4 und 5.5 aufgelistet. Bei der Merkmalsumkehrung wird zur Berechnung der Interpretationsqualität die letzte Schicht des neuronalen Netzes nicht mit eingerechnet, da die Merkmalsumkehrung versucht, einen Aktivierungsvektor, der durch reale Daten entstanden ist, auf der letzten Schicht zu reproduzieren. Hierdurch wird die Qualität der letzten Schicht direkt durch den Interpretations-Algorithmus optimiert, wodurch Interpretationsmethoden mit einer hohen Genauigkeit auch gleichzeitig eine hohe Qualität auf der letzten Schicht erreichen.
Die Qualität der Interpretationen aus den Abbildungen 5.3 und 5.4 wird im folgenden Text mit der subjektiven Interpretationsqualität verglichen. Bei der subjektiven Qualität handelt es sich um die intuitiv bewertete Interpretationsqualität aus der Sicht des Autors. Hiermit wird natürlich nur verglichen, ob die errechnete Qualität mit der Intuition

eines einzelnen Menschen korreliert, dessen Meinung zudem noch durch den Wunsch nach guten Ergebnissen verfälscht wird. Eine größer angelegte und statistisch aussagekräftige (psychologische) Studie würde allerdings den Umfang dieser Arbeit überschreiten.

Bei der Betrachtung der Interpretationsergebnisse der Merkmalsumkehrung (siehe Abbildung 5.3), Klassen-Aktivierung (siehe Abbildung 5.4) und der dazugehörigen Werte für Genauigkeit und Qualität (siehe Tabelle 5.4 und 5.5) sind deutliche Unterschiede zwischen den verschiedenen Interpretationsmethoden *I1* bis *I5* zu erkennen.
Die Ergebnisse der Interpretationsmethode *I1* wurde ohne Regularisierung erzeugt. Hierdurch erreichen diese die höchste Genauigkeit und niedrigste Qualität von allen Interpretationsmethoden. Hierbei fällt auf, dass die Qualität der ersten Schicht *conv1* und der letzten Schicht *dense1* (bei der Klassen-Aktivierung) im Verhältnis zu den anderen Interpretationsmethoden besonders niedrig ist. Insgesamt stimmt die errechnete Interpretationsqualität mit der subjektiven Qualität überein.
Die Interpretationsmethode *I2* wurde durch eine inverse Fourier-Transformation kombiniert mit einer Verschiebung regularisiert. Sowohl bei der Merkmalsumkehrung als auch bei der Klassen-Aktivierung erzeugt dies Interpretationen mit einer niedrigen Genauigkeit, ähnlich wie bei den Interpretationsmethoden *I4* und *I5*. Bei der Merkmalsumkehrung liefert die Interpretationsmethode *I2* die zweitniedrigste Interpretationsqualität, nur knapp vor der Interpretationsmethode *I1*. Dies deckt sich mit der subjektiven Qualität. Bei der Klassen-Aktivierung liegt die Interpretationsqualität hingegen im mittleren Bereich, ungefähr so gut wie bei der Interpretationsmethode *I4*, was sich nicht mit der subjektiven Qualität deckt. Hierbei fällt auf, dass die Interpretationsqualität der Klassen-Aktivierung von *I2* auf der letzten Schicht *dense1* überdurchschnittlich hoch ist.
Die Interpretationsmethode *I3* wurde durch eine inverse Fourier-Transformation kombiniert mit einer Rotation regularisiert. Dies erzeugt Interpretationsergebnisse mit einer sehr hohen Genauigkeit, besonders bei der Klassen-Aktivierung. Bei der Merkmalsumkehrung erreicht die Interpretationsmethode *I3* eine mittlere Qualität, während die Qualität der Klassen-Aktivierungen im unteren Bereich liegt, was auch an der letzten Schicht *dense1* liegt, die eine besonders niedrige Qualität hat. Subjektiv gesehen ist die Qualität der Interpretationsmethode *I3* leicht besser als die Qualität von *I2*. Dieses Ergebnis bestätigt sich bei der Merkmalsumkehrung, nicht aber bei der Klassen-Aktivierung.
Die Interpretationsmethoden *I4* und *I5* sind sehr ähnlich aufgebaut und werden beide durch eine inverse Fourier-Transformation, eine Verschiebung, eine Rotation, eine Skalierung und eine weitere Verschiebung regularisiert. Bei der Interpretationsmethode *I5* kommt zusätzlich noch eine Regularisierung durch die totale Varianz Metrik hinzu. Beide

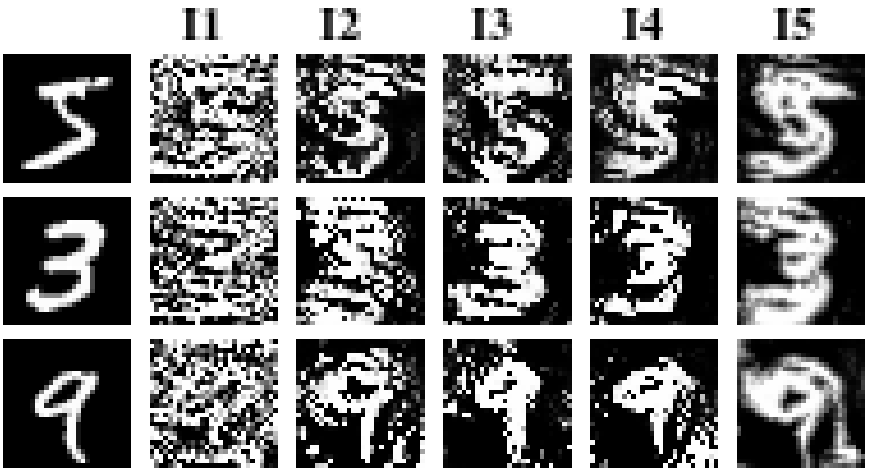

Abbildung 5.3: Interpretationsergebnisse mit verschiedenen Regularisierungsmethoden. Links sind die Originalbilder zu sehen, in den Spalten die Ergebnisse der verschiedenen Interpretationsmethoden. Die Zeilen zeigen verschiedene Beispiele der Merkmalsumkehrung.

Interpretation		Genauigkeit	Qualität	Qualität der Schichten			
				conv1	conv2	conv3	conv4
I1	(5)	0.706	0.309	0.167	0.329	0.344	0.397
	(3)	0.642	0.302	0.162	0.316	0.339	0.391
	(9)	0.877	0.349	0.164	0.352	0.416	0.465
I2	(5)	0.292	0.381	0.227	0.394	0.430	0.474
	(3)	0.284	0.328	0.199	0.346	0.370	0.395
	(9)	0.403	0.387	0.212	0.364	0.455	0.519
I3	(5)	0.578	0.378	0.206	0.399	0.431	0.476
	(3)	0.565	0.409	0.275	0.422	0.443	0.497
	(9)	0.716	0.492	0.284	0.478	0.588	0.620
I4	(5)	0.247	0.441	0.258	0.473	0.498	0.537
	(3)	0.411	0.424	0.280	0.427	0.458	0.530
	(9)	0.442	0.478	0.269	0.453	0.577	0.614
I5	(5)	0.278	0.527	0.350	0.554	0.584	0.626
	(3)	0.255	0.507	0.376	0.538	0.546	0.569
	(9)	0.395	0.567	0.373	0.588	0.646	0.659

Tabelle 5.4: Genauigkeit und Interpretationsqualität der Merkmalsumkehrungen mit verschieden Regularisierungsmethoden. Die Parameter der Regularisierung für I1 bis I5 finden sich in der Tabelle 5.3.

Klassen-Aktivierungen mit verschieden Regularisierungen:

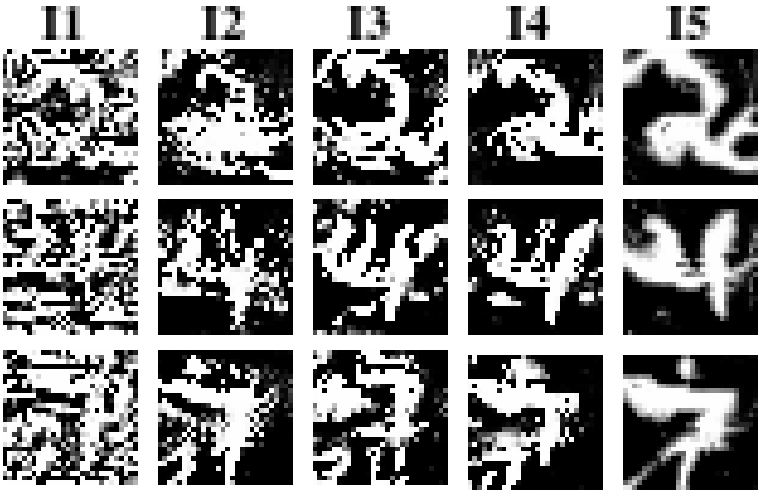

Abbildung 5.4: Interpretationsergebnisse mit verschiedenen Regularisierungsmethoden. In den Spalten werden die Ergebnisse der verschiedenen Interpretationsmethoden gezeigt. Die Zeilen zeigen verschiedene Beispiele der Klassen-Aktivierung.

Interpretation		Genauigkeit	Qualität	Qualität der Schichten				
				conv1	conv2	conv3	conv4	dense1
I1	(2)	1.198	0.237	0.176	0.308	0.316	0.340	0.043
	(4)	1.284	0.214	0.159	0.269	0.273	0.303	0.068
	(7)	0.965	0.227	0.164	0.287	0.302	0.326	0.055
I2	(2)	0.315	0.373	0.211	0.357	0.408	0.449	0.441
	(4)	0.359	0.441	0.220	0.384	0.497	0.543	0.562
	(7)	0.353	0.379	0.237	0.394	0.464	0.482	0.317
I3	(2)	1.046	0.270	0.214	0.328	0.369	0.388	0.050
	(4)	1.147	0.317	0.244	0.363	0.435	0.457	0.084
	(7)	0.771	0.279	0.227	0.347	0.363	0.384	0.073
I4	(2)	0.500	0.379	0.272	0.422	0.477	0.513	0.211
	(4)	0.751	0.372	0.272	0.403	0.490	0.513	0.180
	(7)	0.407	0.383	0.278	0.414	0.472	0.500	0.253
I5	(2)	0.450	0.488	0.396	0.556	0.621	0.626	0.241
	(4)	0.618	0.539	0.414	0.615	0.647	0.653	0.364
	(7)	0.365	0.539	0.458	0.595	0.649	0.636	0.355

Tabelle 5.5: Genauigkeit und Interpretationsqualität von Klassen-Aktivierungen mit verschieden Regularisierungsmethoden. Die Parameter der Regularisierung für I1 bis I5 finden sich in der Tabelle 5.3.

Interpretationsmethoden erreichen eine niedrige Genauigkeit und hohe Qualität, wobei die Qualität der Interpretationsmethode *I5* noch ein wenig besser ist. Auffällig ist, dass beide Interpretationsmethoden auf allen Schichten eine verhältnismäßig hohe Qualität erzielen. Der Haupt-Qualitätsunterschied zwischen den beiden Interpretationsmethoden liegt in den ersten beiden Schichten, auf denen die Interpretationsmethode *I5* etwas besser abschneidet. Die errechneten Qualitäts-Ergebnisse decken sich mit der subjektiven Qualität.

6 Experiment 2: Optimierung von Interpretationsmethoden für ImageNet-Daten

Das erste Experiment aus Kapitel 5 hat gezeigt, dass die Qualitätsmetrik mit einem einfachen Datensatz und neuronalen Netz funktioniert und größtenteils mit der Intuition des Autors korreliert. Allerdings ist der MNIST-Datensatz ein sehr einfacher Datensatz, und auch die Regularisierung der Interpretationsmethode hätte ohne einen Suchalgorithmus parametrisiert werden können.
In diesem Abschnitt werden die im vorherigen Experiment erprobten Techniken auf ein deutlich schwierigeres Problem angewendet. Hiermit soll überprüft werden, ob sich die Qualitätsmetrik auch für komplexe Datensätze und neuronale Netze anwenden lässt.

6.1 Material / Methoden

In diesem Abschnitt werden die für das Experiment verwendeten Daten und Algorithmen beschrieben.

Der Datensatz:
Für dieses Experiment wird der Datensatz ImageNet [22] verwendet. ImageNet ist einer der größten und bekanntesten Bilddatensätze. Inzwischen besteht dieser aus 14 Millionen Bildern und unterscheidet zwischen 21841 Klassen [1].
Da die hier verwendeten Algorithmen sehr rechenintensiv sind, wird nur ein kleiner Teil des ImageNet-Datensatzes verwendet. Dies erleichtert das Erstellen der neuronalen Sprache, verringert aber nicht die Komplexität der Aufgabe. Aus ästhetischen Gründen werden alle Klassen von Tieren (Meerestiere ausgenommen) verwendet. Hierdurch entsteht eine Teilmenge des ImageNet-Trainingsdatensatzes mit insgesamt 300800 Bildern und ein Validierungsdatensatz von 11520 Bildern mit insgesamt 232 Klassen.

Das neuronale Netz:
In vielen wissenschaftlichen Arbeiten über die Interpretation von neuronalen Netzen [5] [15] [20] [21] wird das GoogleLeNet (auch InceptionV1-Netz genannt) [27] verwendet. In einer früheren Arbeit [30] hat sich allerdings herausgestellt, dass sich das InveptionV1-Netz deutlich leichter interpretieren lässt als viele andere der größeren neuronalen Netze. Dadurch wären die komplexen Regularisierungsmethoden, mit denen die Qualitätsmetrik in diesem Experiment evaluiert werden soll, für die Interpretation des InveptionV1-Netzes nicht gerechtfertigt. Um zu überprüfen, ob sich durch die Optimierung der Regularisierungsmethoden auch andere, schwer zu interpretierende neuronale Netze mit hoher Qualität interpretieren lassen, wird in diesem Experiment das VGG16-Netz [24] verwendet. Im Vergleich zu moderneren Netzen, welche für die ImageNet-Klassifikationsaufgabe verwendet werden, ist das VGG16-Netz verhältnismäßig klein, wodurch die für das Experiment benötigten Rechenkapazitäten in einem vertretbaren Rahmen bleiben. Zudem legt eine Arbeit von Gatys et al. [9] zum Thema Style-Transfer nahe, dass VGG-Netze geeignete Merkmale zum Erstellen einer neuronalen Sprache erlernen können.
In der Arbeit von Gatys et al. wird der Style eines Bilds mit der Semantik eines anderen Bilds kombiniert. Sowohl der Style als auch die Semantik der Bilder werden mithilfe der internen Repräsentationen eines VGG19-Netzes beschrieben. Hierdurch lassen sich komplett neue Bilder erstellen, die sowohl den Style als auch die Semantik beinhalten. Dies lässt vermuten, dass VGG-Netze ausreichend Konzepte erlernen können, um natürliche Bilder, wie sie im ImageNet-Datensatz vorkommen, zu beschreiben.
Die Architektur des VGG16-Netzes wird im Anhang A.2 beschrieben. Für dieses Experiment wird ein vortrainiertes VGG16-Netz aus der *Tensorflow* Programmbibliothek [3] verwendet.

Cluster Methode zum Erstellen der neuronalen Sprache und Qualitätsmetrik:
Für das Erstellen der neuronalen Sprache und der Qualitätsmetrik werden dieselben Algorithmen verwendet wie im vorherigen Experiment (siehe Kapitel 5.1). Die Sprache des neuronalen Netzes wird mithilfe des BIRCH-Algorithmus auf dem Trainingsdatensatz erstellt. Die Validierungsdaten werden als Referenzdaten für Interpretationsqualität verwendet.

Die Optimierung der Interpretationsmethode:
Wie auch im vorherigen Experiment wird als Interpretationsmethode die Merkmalsumkehrung [14] auf der letzten tiefen Schicht des neuronalen Netzes durchgeführt. Warum sich diese Methode besonders gut für die Optimierung von Regularisierungen eignet, wird im Abschnitt 5.1 erklärt.

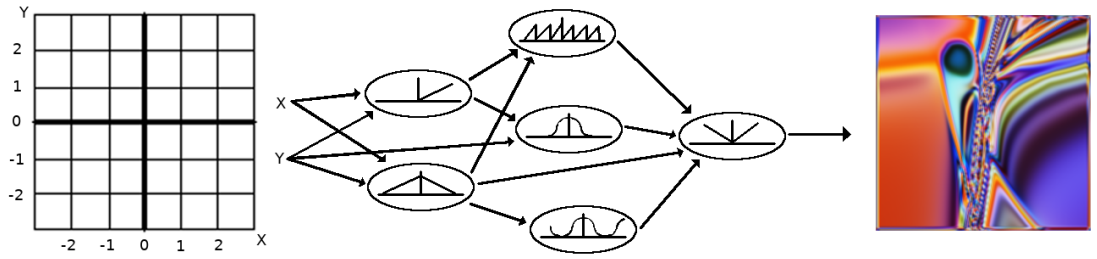

Abbildung 6.1: Grafische Darstellung eines CPPN. Der Eingangswert eines CPPNs ist eine kartesische Koordinate im Raum. Abhängig von den Aktivierungsfunktionen und den Gewichten des CPPNs wird dieser Koordinate ein Farbwert zugeordnet. Indem viele beieinanderliegende Koordinaten visualisiert werden, ergibt sich ein Muster.

Aufgrund der Komplexität der hier verwendeten Daten und des neuronalen Netzes reichen einfache Regularisierungsmethoden wie die inverse Fourier-Transformation kombiniert mit Transformationsrobustheit nicht mehr aus, um qualitative Interpretationen zu erzeugen.
In der Arbeit von Mordvintsev et al. [16] wurden qualitativ wirkende Interpretationen mithilfe von *Compositional Pattern Producing Networks* (CPPN) [25] auf einem InceptionV1-Netz [27] erzeugt.
CPPNs sind eine spezielle Art von neuronalen Netzen, die zum Erzeugen von symmetrischen und sich wiederholenden Mustern eingesetzt werden. Eine grafische Darstellung der Funktionsweise eines CPPNs wird in der Abbildung 6.1 gezeigt.
Vom Aufbau ähneln CPPNs herkömmlichen neuronalen Netzen, nur dass diese nicht nur eine Art von Aktivierungsfunktion (die meisten neuronalen Netze verwenden RELU), sondern eine Vielzahl von verschiedenen Aktivierungsfunktionen verwenden.
Der Eingang eines CPPNs besteht aus einer kartesischen Koordinate (für 2D Bilder x und y), welche die Position im Bild beschreibt. Abhängig von den Gewichten und Aktivierungsfunktionen des CPPNs wird für jede Koordinate ein bestimmter Farb- oder Helligkeitswert ausgegeben. Je nach Komplexität des CPPNs können hierdurch einfache Muster, aber auch komplexe Bilder dargestellt werden.
In diesem Experiment werden CPPNs als Vorbedingung der Interpretationsmethode verwendet. Im Lernprozess werden die Gewichte des CPPNs optimiert, wodurch ein Interpretationsbild erzeugt wird, welches das zu interpretierende neuronale Netz auf die gewünschte Weise beeinflusst. Da mit einem CPPN nicht jede Art von Bild erzeugt werden kann, sollten die hochfrequenten Muster, welche bei Interpretationsmethoden ohne

Regularisierung entstehen, unterdrückt werden.
Mithilfe des Verfahrens *NeuroEvolution of Augmenting Topologies* (NEAT) [26] werden in diesem Experiment CPPNs gezüchtet, die eine möglichst gute Regularisierung der Interpretationsmethode bewirken.
NEAT ist ein von der biologischen Evolution inspirierter Algorithmus, der zum Erstellen von neuronalen Netzen entwickelt wurde. Hierfür wird die Topologie der neuronalen Netze mithilfe eines Genoms beschrieben. Durch Reproduktion und Mutation verändern sich die Genome und werden mithilfe von künstlicher Selektion stetig verbessert.
Indem der Aufbau von CPPNs durch ein Genom beschrieben wird, können sich die CPPNs evolutionär weiterentwickeln, um die Regularisierung der Interpretationsmethode zu verbessern. Je besser ein CPPN die Interpretationsmethode regularisiert, desto größer ist die Chance, dass es überlebt und sich reproduzieren kann. Die Selektion wird auf Basis der Nützlichkeit der vom CPPN regularisierten Interpretationsmethode durchführt.
Um die Entwicklung der CPPNs zu beschleunigen, wird als Basis der evolutionären Entwicklung das CPPN aus der Arbeit von Mordvintsev et al. [16] verwendet, welches bereits gute Interpretationsbilder für das InceptionV1-Netz erzeugen kann.

6.2 Durchführung und Ergebnisse

In diesem Abschnitt werden die Durchführung und die Ergebnisse des Experiments beschrieben. Wie auch im vorherigen Experiment (siehe Kapitel 5.2) besteht die Durchführung aus dem Erstellen der neuronalen Sprache, der automatisierten Suche nach einer guten Regularisierung und der Evaluation der Qualitätsmetrik durch die Darstellung von verschiedenen Interpretationen und dem Abgleich ihrer Qualität mit dem menschlichen Qualitätsempfinden.

Erstellen der neuronalen Sprache:
Die neuronale Sprache für das VGG16-Netz wird auf insgesamt sechs Schichten erstellt. Vier davon sind Faltungsschichten, die in regelmäßigen Abständen im Netz ausgewählt wurden, um unterschiedliche Abstraktionsgrade von Merkmalen zu erkennen. Zusätzlich wird die neuronale Sprache auf den letzten beiden tiefen und voll vernetzten Schichten des Netzes erstellt.
Geeignete Schwellenwerte des BIRCH-Algorithmus werden mithilfe des Versuch-und-Irrtum-Prinzips gewählt. Die Schwellenwerte und die hieraus resultierende Anzahl der

neuronalen Wörter pro Schicht des VGG16-Netzes werden in der folgenden Tabelle 6.1 dargestellt.

Name der Schicht	Schwellenwert	Anzahl neuronale Wörter
block2 conv1	1600	127547
block3 conv1	6800	132382
block4 conv1	11000	126350
block5 conv1	2500	117777
fc1	500	129123
fc2	120	116645

Tabelle 6.1: In dieser Tabelle wird für ausgewählte Schichten des VGG16-Netzes der verwendete Schwellenwert des BIRCH-Algorithmus und die aus dem Cluster-Verfahren resultierende Anzahl der neuronalen Wörter angegeben.

Da in diesem Experiment sowohl das neuronale Netz als auch der Trainingsdatensatz deutlich größer sind als im vorherigen Experiment, kann die neuronale Sprache aufgrund des hohen Speicher-Verbrauchs nicht auf jeder Schicht des Netzes approximiert werden. Dennoch benötigt die Approximation der neuronalen Sprache insgesamt 47.8 Gigabyte Speicherplatz im Arbeitsspeicher eines Computers.

Die Optimierung der Interpretationsmethode:
Um ein CPPN zu finden, welches die Merkmalsumkehrung besonders gut regularisiert, werden mithilfe des NEAT-Algorithmus verschiedene CPPNs gezüchtet und auf einen hohen Nutzen optimiert.
In der Abbildung 6.2 wird die mittlere und maximale Nützlichkeit, Qualität und Genauigkeit der Interpretationsmethoden abhängig von der Generation der evolutionären Entwicklung dargestellt. Im Bild ganz links ist zu sehen, dass sich die Nützlichkeit der Interpretationsmethoden im Laufe der Evolution deutlich verbessert. Der durchschnittliche Nutzen der Interpretationsmethoden (blaue Kurve) verschlechtert sich zwar leicht in den ersten Generationen, steigt anschließend aber an und erreicht einen Wert knapp über dem Startwert. Am maximalen Nutzen (orange Kurve) ist gut zu sehen, dass der Evolutionsprozess einige CPPNs hervorbringt, dessen Nutzen deutlich über dem Durchschnitt liegt. Diese Interpretationsmethoden verbessern sich deutlich in beinahe jeder neuen Generation.
Bei einer genaueren Betrachtung der Qualität und Genauigkeit ist zu erkennen, dass die mittlere Qualität anfangs ganz leicht ansteigt und danach auf einem Wert nahe eins (also

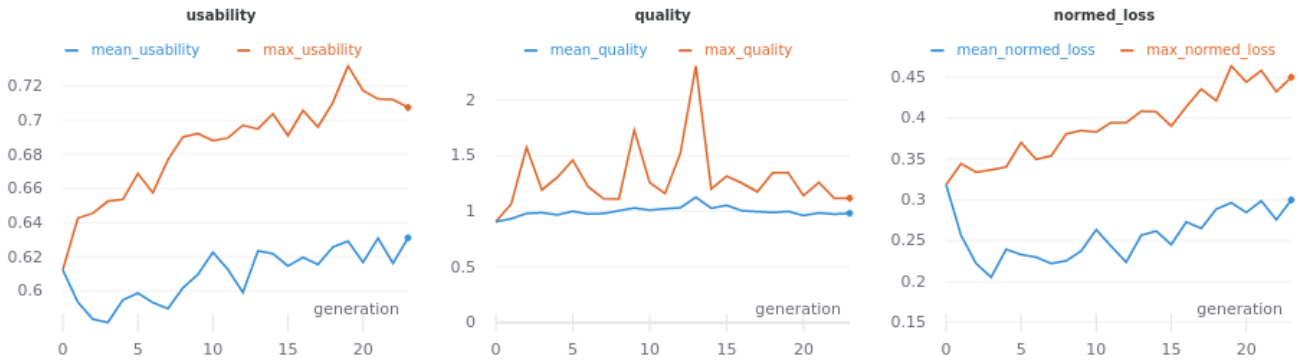

Abbildung 6.2: In dieser Abbildung werden die Nützlichkeit (usability), Qualität (quality) und Genauigkeit der Interpretationsmethoden für jede Generation des evolutionären Prozesses dargestellt. Die blaue Kurve zeigt den Mittelwert über alle Interpretationsmethoden und die orange Kurve den Maximalwert.

dem Optimalwert) stagniert. Sowohl die Graphen der mittleren als auch der maximalen Genauigkeit ähneln im Verlauf (abgesehen vom verändertem Wertebereich) stark denen der Nützlichkeit. Demnach liegt der gesteigerte Nutzen der Interpretationsmethoden fast ausschließlich an einer gesteigerten Genauigkeit.

Stark auffällig ist die Kurve der maximalen Qualität. Hier werden Werte bis zu 2.3 erreicht. Da für die Normierung der Qualität Validierungsdaten verwendet wurden, welche als hochqualitativ angesehen werden, sollten die Werte für die Interpretationsqualität maximal eins erreichen (in Ausnahmefällen etwas höher).

In der Abbildung 6.3 werden einige Bilder von Interpretationsmethoden gezeigt, welche eine besonders gute Interpretationsqualität, mit Werten deutlich über eins liefern. Hierbei fällt auf, dass diese Interpretationsbilder nur einige wenige Konzepte wie einfarbige Flächen oder Farbübergänge zeigen. Zudem ist die Genauigkeit der Interpretationsbilder besonders niedrig.

Bei der Berechnung des Nutzens einer Interpretationsmethode ist eine maximale Interpretationsqualität von eins möglich. Die Interpretationsmethoden mit einer Interpretationsqualität deutlich über eins haben ausnahmslos immer eine sehr niedrige Genauigkeit und somit auch einen niedrigen Nutzen. Hierdurch wird der evolutionäre Prozess nicht von diesen Interpretationsmethoden beeinflusst, da sie sich aufgrund ihres geringen Nutzens nicht fortpflanzen können und somit schnell aussterben.

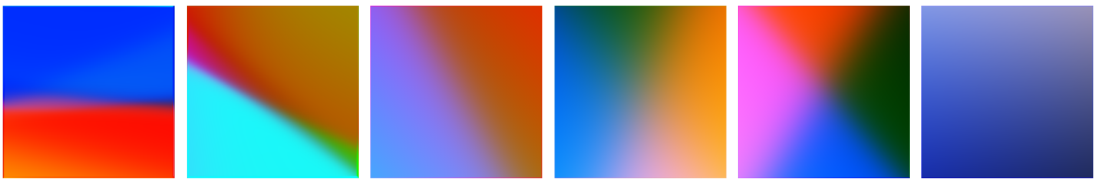

Abbildung 6.3: Beispiel Interpretationsbilder mit besonders hoher Qualität. Diese Bilder erreichen Qualitätswerte deutlich über eins, von denen angenommen wurde, dass sie im Experiment nicht vorkommen können. Auffällig ist, dass die Interpretationsbilder nur einige wenige Konzepte wie einfarbige Flächen oder Farbübergänge zeigen und eine besonders niedrige Genauigkeit erreichen.

Abgleich der Interpretationsqualität mit menschlicher Intuition:
Wie auch im vorherigen Experiment wird im folgenden Text untersucht, ob es bei den Interpretationsergebnissen eine Korrelation zwischen errechneter Qualität und subjektiver Qualität gibt. Um die errechnete Qualität mit der subjektiven Qualität zu vergleichen, werden die Ergebnisse von fünf verschiedenen Interpretationsmethoden gegenübergestellt. Die fünf Interpretationsmethoden sind:

I1: Ohne Regularisierung.

I2: Regularisiert durch die inverse Fourier-Transformation kombiniert mit Transformationsrobustheit.

I3: Regularisiert durch das CPPN, welches in der Arbeit von Mordvintsev et al. [16] verwendet wurde.

I4: Regularisiert durch das CPPN, welches im evolutionären Prozess den höchsten Nutzen erreicht hat.

I5: Regularisiert durch ein CPPN mit überdurchschnittlich hoher Qualität und einer guten Genauigkeit.

Die verschiedenen Interpretationsmethoden werden anhand der Merkmalsumkehrung auf der Schicht *fc2* des VGG16-Netzes und durch Klassen-Aktivierungen getestet. In den Abbildungen 6.4 und 6.5 werden die verschiedenen Interpretationsergebnisse abgebildet. Die Genauigkeit und die Interpretationsqualität sind in den Tabellen 6.2 und 6.3 aufgelistet. Wie auch beim vorherigen Experiment wird bei der Merkmalsumkehrung zur Berechnung der Interpretationsqualität die letzte Schicht nicht mit eingerechnet.

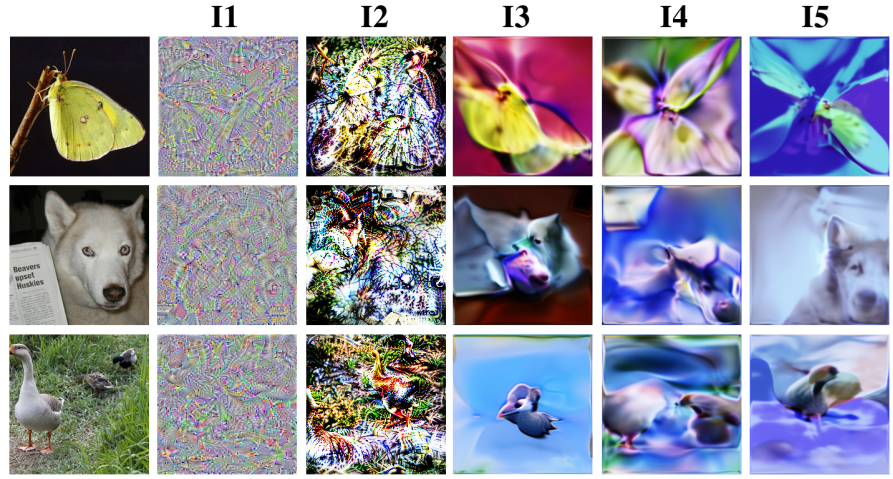

Abbildung 6.4: Interpretationsergebnisse mit verschiedenen Regularisierungsmethoden. Links sind die Originalbilder zu sehen, in den Spalten die Ergebnisse der verschiedenen Interpretationsmethoden. Die Zeilen zeigen verschiedene Beispiele der Merkmalsumkehrung.

Interpretation		Genauigkeit	Qualität	Qualität der Schichten				
				block2 conv1	block3 conv1	block4 conv1	block5 conv1	fc1
I1	(Schmetterling)	1.006	0.931	0.522	0.577	0.6275	0.664	2.264
	(Hund)	1.119	0.703	0.559	0.624	0.671	0.716	0.944
	(Gans)	0.817	0.731	0.500	0.568	0.588	0.566	1.432
I2	(Schmetterling)	0.217	0.588	0.285	0.323	0.395	0.541	1.398
	(Hund)	0.184	0.477	0.293	0.332	0.399	0.540	0.820
	(Gans)	0.189	0.457	0.278	0.325	0.385	0.456	0.842
I3	(Schmetterling)	0.356	1.113	0.918	0.941	0.942	0.943	1.819
	(Hund)	0.317	1.035	1.125	1.095	1.035	0.934	0.986
	(Gans)	0.188	1.104	1.248	1.275	1.212	0.998	0.788
I4	(Schmetterling)	0.401	1.038	0.900	0.891	0.860	0.844	1.694
	(Hund)	0.313	0.882	0.857	0.881	0.875	0.869	0.930
	(Gans)	0.370	0.947	0.863	0.878	0.889	0.832	1.271
I5	(Schmetterling)	0.262	0.920	0.719	0.780	0.812	0.903	1.384
	(Hund)	0.350	1.449	1.556	1.714	1.624	1.378	0.972
	(Gans)	0.110	0.957	0.939	0.996	0.983	0.963	0.907

Tabelle 6.2: Genauigkeit und Interpretationsqualität der Merkmalsumkehrungen mit verschiedenen Regularisierungsmethoden.

Abbildung 6.5: Interpretationsergebnisse mit verschiedenen Regularisierungsmethoden. In den Spalten werden die Ergebnisse der verschiedenen Interpretationsmethoden gezeigt. Die Zeilen zeigen verschiedene Beispiele der Klassen-Aktivierung.

Interpretation		Genauigkeit	Qualität	Qualität der Schichten					
				block2 conv1	block3 conv1	block4 conv1	block5 conv1	fc1	fc2
I1	(Rotkehlchen)	1.022	0.132	0.208	0.234	0.192	0.128	0.021	0.009
	(Libelle)	1.164	0.124	0.204	0.228	0.181	0.107	0.015	0.008
	(Elefant)	1.138	0.130	0.203	0.232	0.192	0.126	0.018	0.009
I2	(Rotkehlchen)	1.042	0.157	0.256	0.279	0.234	0.149	0.019	0.008
	(Libelle)	1.174	0.138	0.224	0.243	0.207	0.128	0.018	0.010
	(Elefant)	1.257	0.146	0.229	0.246	0.217	0.148	0.024	0.012
I3	(Rotkehlchen)	4.314	0.320	0.447	0.465	0.453	0.391	0.113	0.050
	(Libelle)	5.443	0.365	0.579	0.554	0.490	0.394	0.108	0.063
	(Elefant)	7.299	0.372	0.464	0.493	0.511	0.491	0.166	0.105
I4	(Rotkehlchen)	4.540	0.350	0.492	0.496	0.500	0.450	0.114	0.045
	(Libelle)	3.867	0.276	0.403	0.415	0.389	0.317	0.084	0.047
	(Elefant)	4.817	0.326	0.430	0.462	0.469	0.400	0.125	0.069
I5	(Rotkehlchen)	4.541	0.341	0.451	0.495	0.495	0.438	0.116	0.052
	(Libelle)	4.653	0.350	0.493	0.529	0.511	0.402	0.108	0.056
	(Elefant)	6.081	0.422	0.533	0.605	0.610	0.527	0.178	0.085

Tabelle 6.3: Genauigkeit und Interpretationsqualität von Klassen-Aktivierungen mit verschiedenen Regularisierungsmethoden.

Bei der Betrachtung der Interpretationsergebnisse der Merkmalsumkehrung (siehe Abbildung 6.4), Klassen-Aktivierung (siehe Abbildung 6.5) und der dazugehörigen Werte für Genauigkeit und Qualität (siehe Tabelle 6.2 und 6.3) sind deutliche Unterschiede zwischen den Interpretationsmethoden *I1* und *I2* und den Interpretationsmethoden *I3* bis *I5* zu erkennen.
Auf den Interpretationsbildern ist deutlich zu sehen, dass die subjektive Qualität der Interpretationsbilder, welche mithilfe von CPPNs erzeugt wurden (*I3* bis *I5*), deutlich höher ist als die Qualität der Interpretationsbilder, die ohne CPPNs erzeugt wurden (*I1* und *I2*). Die Interpretationsbilder von *I1* und *I2* zeigen eine Vielzahl von abstrakten Formen, die sich vor allem durch starke Veränderungen in den Farbwerten auszeichnen. Nahe beieinanderliegende Pixel haben sehr unterschiedliche Farbwerte, wodurch in den Bildern nur wenige einfarbige Flächen und Kanten vorkommen.
Die Interpretationsbilder von *I3* bis *I5* sind deutlich schlichter gehalten. Es gibt größere einfarbige Flächen und es werden nur ein oder zwei Objekte abgebildet, die sich in ihrer Komplexität stark vom monotonen Hintergrund abheben. Hierdurch ähneln die Interpretationsbilder von *I3* bis *I5* deutlich mehr den realen Bildern aus dem Trainingsdatensatz. Dieser Eindruck spiegelt sich auch in der errechneten Interpretationsqualität wider. Diese ist sowohl bei der Merkmalsumkehrung als auch bei der Klassen-Aktivierung von *I1* und *I2* deutlich niedriger als bei den Interpretationsmethoden *I3*, *I4* und *I5*.
Eine Ausnahme sind hierbei die Ergebnisse der Merkmalsumkehrung von *I1*. Diese erreichen eine vergleichsweise hohe errechnete Qualität, trotz ihrer offensichtlich niedrigen subjektiven Qualität. Dies liegt einerseits an der hohen Interpretationsqualität der letzten Schicht. Andrerseits fällt auf, dass sich die Merkmalsumkehrungen von *I1* sich vor allem durch sehr körnige, verrauschte Bilder auszeichnen. Dies sind Merkmale, die wahrscheinlich in den ersten Schichten des neuronalen Netzes erkannt werden. Da die neuronale Sprache aber erst ab dem zweiten Faltungs-Block approximiert wurde, ist es möglich, dass diese Merkmale nicht in die Interpretationsqualität mit einfließen.
Die subjektive Qualität der Interpretationsbilder von *I2* scheint etwas besser zu sein als die von der Interpretationsmethode *I1*. Dies deckt sich mit der errechneten Qualität der Klassen-Aktivierung, nicht aber mit der Merkmalsumkehrung.
Die subjektive Qualität der Interpretationsmethoden *I3* bis *I5* ist sehr schwer miteinander zu vergleichen. Dennoch wird ein Versuch unternommen, indem intuitiv diejenigen Bilder ausgewählt werden, welche realen Bildern am ähnlichsten sehen, und überprüft wird, ob diese auch die höchste errechnete Qualität haben.
Die Interpretationsbilder mit der besten Qualität sind meiner Meinung nach bei der Merkmalsumkehrung die Gans der Interpretationsmethode *I3* und die Hunde der Inter-

pretationsmethoden *I3* und *I5*. Bei der Klassen-Aktivierung hat der Elefant der Interpretationsmethode *I5* eine auffällig hohe Qualität. Diese Beobachtungen decken sich mit den Werten der errechneten Qualität. Hierbei wurde aber die Interpretationsqualität der Merkmalsumkehrung und der Klassen-Aktivierung nicht direkt miteinander verglichen.
Bei einem direkten Vergleich der Interpretationsqualität zwischen Merkmalsumkehrung und Klassen-Aktivierung sind deutliche Unterschiede in den Wertebereichen zu erkennen. Bei der Merkmalsumkehrung werden Qualitätswerte im Bereich von ungefähr 0.45 bis 1.45 erreicht. Bei der Klassen-Aktivierung werden deutlich niedrigere Werte von ungefähr 0.12 bis 4.3 erreicht. Dennoch haben die verschiedenen Interpretationsmethoden der Merkmalsumkehrung und der Klassen-Aktivierung eine ähnliche subjektive Qualität. Weitere Aussagen über die Interpretationsqualität können aufgrund der ungenauen Validierungsmethode (Vergleich der subjektiven und errechneten Qualität auf Basis einer Person) nicht getroffen werden.
Bei einem Vergleich der Genauigkeit der verschiedenen Interpretationsmethoden fällt auf, dass bei der Merkmalsumkehrung die Interpretationsmethode *I1* (zusammen mit der Interpretationsmethode *I2*) wie erwartet die höchste Genauigkeit erreicht, da das Optimierungsverfahren nicht durch eine Regularisierung eingeschränkt wird. Bei der Klassen-Aktivierung verhält es sich genau andersherum. Auch hier erreichen die Interpretationsmethoden *I1* und *I2* perfekte Werte nahe eins. Die stark regulierten Interpretationsmethoden *I3* bis *I5* schaffen es aber, das Klassen-Neuron deutlich stärker zu aktivieren, wodurch sie Genauigkeiten mit Werten deutlich über eins erreichen.

7 Diskussion

In diesem Kapitel werden die Ergebnisse der beiden Experimente aus Kapitel 5 und 6 diskutiert.

7.1 Die Optimierung der Transformationsrobustheit

Die Ergebnisse des ersten Experiments (siehe Kapitel 5) haben gezeigt, dass die Qualitätsmetrik die automatisierte Suche nach einer guten Regularisierung für Interpretationen des Zahlen-Netzes ermöglicht. Im Abschnitt 5.2 wurde als nützlichste Regularisierung eine inverse Fourier-Transformation mit einer zufälligen Rotation ermittelt. Diese Art der Regularisierung erzeugt Interpretationen von mittlerer Qualität, aber hoher Genauigkeit. Andere Varianten der Regularisierung erreichen zwar eine höhere Qualität, aber nur auf Kosten einer deutlich geringeren Genauigkeit.
Bei einer Betrachtung des Wertebereichs von Interpretationsqualitäten verschiedener Schichten fällt auf, dass sich die verschiedenen Regularisierungsmethoden oft auf bestimmte Schichten auswirken. Die Interpretationsmethode *I1* ohne Regularisierung hat eine besonders niedrige Interpretationsqualität auf der ersten Schicht *conv1*. Die Interpretationsergebnisse von *I1* (siehe Abbildung 5.3 und 5.4) zeigen Bilder mit sehr vielen, einzeln im Bild verteilten weißen Pixeln auf schwarzem Hintergrund und diffusen Kanten. Simple Merkmale wie glatte Kanten und gleichfarbige Flächen werden wahrscheinlich in den ersten Schichten des neuronalen Netzes erkannt. Der Vergleich der Interpretationsmethoden *I4* und *I5* zeigt, dass die Regularisierung durch totale Varianz die Interpretationsqualität auf den ersten beiden Schichten verbessert. Auf den Interpretationsbildern ist gut zu erkennen, dass die Regularisierung mit totaler Varianz die Kanten auf den Interpretationsbildern glättet und einzelne weiße Pixel auf dem schwarzen Hintergrund unterdrückt.

Bei einem Vergleich zwischen errechneter und subjektiver Interpretationsqualität kann eine Korrelation der Qualitätsmetrik mit menschlicher Intuition bestätigt werden. Der

einzige Ausreißer ist die errechnete Interpretationsqualität der Methode *I2*, welche bei der Klassen-Aktivierung unerwartet hoch ist (siehe Tabelle 5.5). Diese Interpretationsmethode erzeugt vor allem auf den tieferen Schichten des Zahlen-Netzes eine gute Interpretationsqualität. Da allerdings nicht bekannt ist, welche Merkmale in diesen Schichten erkannt werden, bleibt auch unklar, welche Teile der Interpretationen eine hohe Qualität aufweisen. Wahrscheinlich ist, dass abstrakte Merkmale wie die grobe Form, Größe und Ausrichtung der Interpretationsbilder den Zahlen des Trainingsdatensatzes ähnelt und somit eine hohe Interpretationsqualität erzeugen.

7.2 Evolutionäre Entwicklung von CPPNs

Die Ergebnisse des zweiten Experiments (siehe Kapitel 6) haben gezeigt, dass die Qualitätsmetrik auch bei komplexen Daten, tiefen neuronalen Netzen und starken Regularisierungen die Optimierung der Regularisierung ermöglicht. Zudem wird deutlich, dass sich CPPNs sehr gut dazu eignen, neuronale Netze zu interpretieren. Bei der Merkmalsumkehrung erreichen alle Ergebnisse der Interpretationsmethoden mit CPPN (*I3* bis *I5*) Interpretationsqualitäten nahe eins (siehe Tabelle 6.2). Somit ist die Qualität ungefähr so gut wie die der Validierungsdaten, welche als Referenzqualität verwendet wurden.
Bei der Klassen-Aktivierung erreichen Interpretationsmethoden mit CPPNs deutlich niedrigere Interpretationsqualitäten als bei der Merkmalsumkehrung. Dies könnte daran liegen, dass bei der Merkmalsumkehrung die Merkmale eines realen Bildes reproduziert werden. Diese Merkmale werden durch einen Aktivierungsvektor beschrieben, der eine Vielzahl von Bildinformationen des realen Bilds beinhaltet. Hierdurch werden bei der Optimierung des Interpretationsbilds Informationen verwendet, die ein reales Aussehen vorprägen. Im Vergleich dazu enthält ein einzelnes Klassen-Neuron kaum Informationen und kann durch eine Vielzahl von verschiedenen Eingangsbildern aktiviert werden.
Im ersten Experiment ist dieser Effekt nicht aufgefallen, da die MNIST-Daten sehr viel primitiver sind. Hier wird ein Klassen-Neuron immer durch eine mittig ausgerichtete, weiße Zahl auf schwarzem Hintergrund aktiviert, wodurch die Möglichkeiten zur Aktivierung eines Klassen-Neurons verhältnismäßig beschränkt bleiben.

Ein bedeutendes Ergebnis des zweiten Experiments ist, dass Interpretationsmethoden die mit einem CPPN regularisiert werden, Klassen-Neuronen deutlich stärker aktivieren können als Interpretationsmethoden ohne Regularisierung (siehe Tabelle 6.3). Die Interpretation von Klassen-Neuronen ist besonders schwierig, da diese kaum Informationen

enthalten, die eine Optimierung erleichtern und real aussehende Interpretationen vorprägen. Hierdurch entstehen oft unverständliche Interpretationsbilder. Die Ergebnisse aus dem zweiten Experiment legen jedoch nahe, dass mit CPPNs besonders gute Klassen-Aktivierungen gefunden werden können, die sowohl das Klassen-Neuron stark aktivieren als auch intuitiv zu verstehen sind.
Das Erstellen von aussagekräftigen Klassen-Aktivierungen wäre ein großer Fortschritt für die Interpretation von neuronalen Netzen, da diese intuitiv dazu verwendet werden können, neuronale Netze zu testen und Wissen aus einem neuronalen Netz zu generieren. Wird die Klassen-Aktivierung beispielsweise auf ein Klassifikator-Netz angewendet, welches gelernt hat, zwischen verschiedenen Hunderassen zu unterscheiden, könnte es sein, dass die Klassen-Aktivierung des Huskys nur Schnee zeigt. In diesen Fall sollte beim Trainingsdatensatz überprüft werden, ob alle Husky-Bilder im Schnee aufgenommen wurden. Das neuronale Netz könnte gelernt haben, dass Schnee ein Unterscheidungsmerkmal zwischen Huskys und anderen Hunderassen ist.
Es könnte auch sein, dass bei der Klassen-Aktivierung des Huskys spitze Hundeohren hervorgehoben werden. Dieser Interpretation kann die Information entnommen werden, das spitze Ohren ein Unterscheidungsmerkmal zwischen Huskys und anderen Hunderassen sind. Anschließend sollte aber überprüft werden, ob beispielsweise Schäferhunde nicht auch als Huskys erkannt werden.

Eine weitere Erkenntnis aus dem zweiten Experiment ist, dass zwei Grundannahmen, welche in der Vorbereitung der Experimente getroffen wurden, falsch sind. Es wurde erwartet, dass die höchste Interpretationsqualität durch Test oder Validierungsdaten erreicht wird. Im evolutionären Prozess hat sich aber gezeigt, dass sich die errechnete Interpretationsqualität noch deutlich verbessert, wenn die Interpretationsbilder sehr simpel gestaltet werden und nur wenige, primitive Konzepte wie einfarbige Flächen und gerade Farbübergänge zeigen (siehe Abbildung 6.3). Diese Interpretationsbilder beeinflussen das Netz allerdings nicht auf die gewünschte Weise und sind dadurch für eine Interpretation unbrauchbar.
Die Annahme, dass Interpretationsmethoden ohne Regularisierung die höchste Genauigkeit erreichen, konnte im ersten Experiment sowie im zweiten Experiment bei der Merkmalsumkehrung bestätigt werden. Bei der Klassen-Aktivierung aus dem zweiten Experiment verhält es sich allerdings genau umgekehrt. Hier erzielen die Interpretationsmethoden die höchste Genauigkeit, welche am stärksten regularisiert wurden.

7.3 Allgemeine Diskussion

In beiden Experimenten konnte die Korrelation zwischen errechneter und subjektiver Interpretationsqualität belegt werden. Dieses Ergebnis ist allerdings sehr ungenau, da je Experiment nur 30 Interpretationsergebnisse von nur einem Menschen bewertet wurden. Zudem gestaltet sich der Vergleich zwischen errechneter und subjektiver Qualität als schwierig, da weder beim neuronalen Netz noch beim Menschen bekannt ist, wie stark bestimmte Merkmale eines Interpretationsbildes gewichtet werden. Um tatsächlich herauszufinden, ob die errechnete Qualität mit der subjektiven Qualität korreliert, müsste ein größer angelegtes (psychologisches) Experiment mit vielen Probanden durchgeführt werden.
Deutlich interessanter ist allerdings die Frage, ob Interpretationsmethoden mit einem hohen errechneten Nutzen auch wirklich nützlich sind in dem Sinne, dass sie bei einer konkreten Aufgabenstellung dazu verwendet werden können, das gelernte Wissen oder die Entscheidungsgrundlagen von neuronalen Netzen einem Menschen auf verständliche Weise zu präsentieren.

Im ersten Experiment konnte der Effekt beobachtet werden, dass sich einige Regularisierungsmethoden nur auf die Interpretationsqualität von einzelnen Schichten des neuronalen Netzes auswirken. Die Regularisierung mit totaler Varianz glättet beispielsweise Kanten. Hierdurch wird die Interpretationsqualität auf den vorderen Schichten verbessert, auf denen diese simplen Merkmale erkannt werden.
Im zweiten Experiment konnte dieser Effekt zwar nicht direkt beobachtet werden, allerdings erklärt er die gute Interpretationsqualität der Merkmalsumkehrungen von *I1* (siehe Tabelle 6.2). Die Interpretationsbilder der Merkmalsumkehrung von *I1* (siehe Abbildung 6.4) sind sehr körnig und verrauscht. Einfache Merkmale wie simple Muster, einfarbige Flächen oder Kanten werden wahrscheinlich in den ersten Schichten des VGG16-Netzes erkannt. Da die neuronale Sprache aber nicht auf den ersten Schichten des Netzes approximiert wurde, kann die starke Körnung der Bilder oder fehlende Kanten und fehlende einfarbige Flächen auch nicht in die Interpretationsqualität mit einfließen.
Der Effekt, dass die verschiedenen Schichten eines neuronalen Netzes auf verschiedene Bildmerkmale schauen, kann genutzt werden, um zu bestimmen, welche Merkmale für eine hochqualitative Interpretation besonders wichtig sind. Dies kann durch eine einfache Gewichtung der Interpretationsqualität einzelner Schichten umgesetzt werden. Leider ist für gewöhnlich nicht bekannt, wo im neuronalen Netz bestimmte Merkmale erkannt werden.

Interpretationsmethoden wie DeepDream [15] oder die Merkmalsumkehrung [14] könnten dabei helfen herauszufinden in welchen Schichten eines neuronalen Netzes bestimmte Merkmale erkannt werden. Damit dies gelingt, müssen diese Interpretationsmethoden allerdings gut regularisiert werden. Hierdurch ergibt sich ein Optimierungs-Kreis: Indem qualitative Interpretationsmethoden entwickelt werden, kann das neuronale Netz besser verstanden werden. Durch ein besseres Verständnis des neuronalen Netzes kann aber auch die Suche nach guten Interpretationsmethoden verbessert werden, indem gezielt Schichten herausgesucht werden, auf denen die neuronale Sprache erzeugt wird, und diese entsprechend der Relevanz der Merkmale gewichtet werden.

Eine Schwäche des hier entwickelten Algorithmus ist der Wertebereich von Genauigkeit und Interpretationsqualität. Zum einen haben die Experimente gezeigt, dass sowohl bei der Genauigkeit als auch bei der Interpretationsqualität der geplante Wertebereich zwischen null und eins überschritten werden kann. Zudem sind auch Werte, die zwischen null und eins liegen, nicht sonderlich aussagekräftig. Dies liegt zum einen daran, dass nicht bekannt ist, welche konkrete Bedeutung eine bestimmte Entfernung zwischen zwei Aktivierungsvektoren hat. Zum anderen entsteht bei der Normierung der Werte (siehe Gleichung 5.2 und 5.3) eine Nichtlinearität, wodurch die Werte für Menschen weniger intuitiv werden.
Nur der Wert von eins lässt die konkrete Aussage zu, dass die Interpretationsqualität oder Genauigkeit gleich gut ist wie die Referenzqualität oder Referenzgenauigkeit.
Bei der Auswertung der Experimente war ein direkter Vergleich zwischen zwei Genauigkeiten oder Interpretationsqualitäten die einzige Möglichkeit der Auswertung. Aus einem einzelnen Wert für Nutzen, Qualität oder Genauigkeit (der ungleich eins ist) können kaum Schlüsse gezogen werden.
Um den Wertebereich von sowohl der Interpretationsqualität als auch der Genauigkeit zu verbessern, müssen die internen Aktivierungsvektoren von neuronalen Netzen besser verstanden werden. In dieser Arbeit wurden minimale Annahmen über die Verteilung der internen Aktivierungsvektoren gemacht. Um sicherzugehen, dass die Qualitätsmetrik dennoch funktioniert, wurde die neuronale Sprache so wenig wie möglich approximiert, wodurch diese vor allem im zweiten Experiment sehr viel Speicherplatz beanspruchte. Durch ein besseres Verständnis der internen Aktivierungsvektoren könnte die Approximation der neuronalen Sprache verbessert werden, zum Beispiel durch eine geeignetere Cluster-Methode als des BIRCH-Algorithmus, und somit die benötigten Rechenressourcen stark reduziert werden. Auch könnte es dabei helfen, die Berechnung der Ähnlichkeit

zweier Aktivierungsvektoren zu optimieren, wodurch der Wertebereich der Qualitätsmetrik und der Genauigkeit verbessert werden kann.

8 Fazit

Die hier entwickelte Methode zum Messen der Interpretationsqualität und die hieraus resultierenden Möglichkeiten zur automatisierten Optimierung von Interpretationsmethoden weisen noch einige Schwachstellen und viele Möglichkeiten der Verbesserung auf. Dennoch konnte in dieser Arbeit gezeigt werden, dass es möglich ist, die Qualität einer Interpretation zu messen. Die hieraus resultierenden Möglichkeiten zur automatisierten Verbesserung von Interpretationsmethoden sind sehr vielseitig und können auf eine Vielzahl von Regularisierungsmethoden angewendet werden. Mit einigen Verbesserungen, besonders beim Einsparen von Rechenressourcen, könnte die Qualitätsmetrik das Forschungsfeld zur Interpretation neuronaler Netze stark vorantreiben. Bewährte Regularisierungsmethoden können auf die verschiedenen Datensätze und neuronalen Netze angepasst werden. Zudem kann die Forschung nach neuen Regularisierungsmethoden fokussiert werden, indem explizit Methoden entwickelt werden, die den Nutzen der Interpretation maximieren.

Literaturverzeichnis

[1] *Image Net Webseite*. http://www.image-net.org/. – Accessed: 2021-02-02

[2] *Introduction to CNN Keras - 0.997 (top 6%)*. https://www.kaggle.com/yassineghouzam/introduction-to-cnn-keras-0-997-top-6. – Accessed: 2021-02-21

[3] *Tensorflow VGG16-Netz*. https://www.tensorflow.org/api_docs/python/tf/keras/applications/VGG16/. – Accessed: 2021-04-25

[4] Aljanabi, Mohammed A. ; Hussain, Zahir M. ; Shnain, Noor Abd A. ; Lu, Song F.: Design of a hybrid measure for image similarity: a statistical, algebraic, and information-theoretic approach. In: *European Journal of Remote Sensing* (2019)

[5] Carter, Shan ; Armstrong, Zan ; Schubert, Ludwig ; Johnson, Ian ; Olah, Chris: Activation Atlas. In: *Distill* (2019)

[6] Doshi-Velez, Finale ; Kim, Been: Towards A Rigorous Science of Interpretable Machine Learning. In: *arXiv:1702.08608 [stat.ML]* (2017)

[7] Erhan, Dumitru ; Bengio, Y. ; Courville, Aaron ; Vincent, Pascal: Visualizing Higher-Layer Features of a Deep Network. In: *Technical Report, Univeristé de Montréal* (2009)

[8] Fong, Ruth C. ; Vedaldi, Andrea: Interpretable Explanations of Black Boxes by Meaningful Perturbation. In: *2017 IEEE International Conference on Computer Vision (ICCV)* (2017)

[9] Gatys, Leon A. ; Ecker, Alexander S. ; Bethge, Matthias: A Neural Algorithm of Artistic Style. In: *arXiv:1508.06576 [cs.CV]* (2015)

[10] Kim, Been ; Wattenberg, Martin ; Gilmer, Justin ; Cai, Carrie ; Wexler, James ; Viegas, Fernanda ; sayres, Rory: Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In: *Proceedings of Machine Learning Research* (2018)

[11] Kindermans, Pieter-Jan ; Schütt, Kristof T. ; Alber, Maximilian ; Müller, Klaus-Robert ; Erhan, Dumitru ; Kim, Been ; Dähne, Sven: Learning how to explain neural networks: PatternNet and PatternAttribution. In: *arXiv:1705.05598 [stat.ML]* (2017)

[12] LeCun, Yann ; Cortes, Corinna ; Burges, CJ: MNIST handwritten digit database. In: *ATT Labs [Online]* (2010)

[13] Lipton, Zachary C.: The Mythos of Model Interpretability. In: *arXiv:1606.03490 [cs.LG]* (2016)

[14] Mahendran, Aravindh ; Vedaldi, Andrea: Understanding Deep Image Representations by Inverting Them. In: *arXiv:1412.0035 [cs.CV]* (2014)

[15] Mordvintsev, Alexander ; Olah, Christopher ; Tyka, Mike: Inceptionism: Going Deeper into Neural Networks. In: *Google AI Blog* (2015)

[16] Mordvintsev, Alexander ; Pezzotti, Nicola ; Schubert, Ludwig ; Olah, Chris: Differentiable Image Parameterizations. In: *Distill* (2018)

[17] Nguyen, Anh ; Dosovitskiy, Alexey ; Yosinski, Jason ; Brox, Thomas ; Clune, Jeff: Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In: *arXiv:1605.09304 [cs.NE]* (2016)

[18] Nguyen, Anh ; Yosinski, Jason ; Bengio, Yoshua ; Dosovitskiy, Alexey ; Clune, Jeff: Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space. In: *arXiv:1612.00005 [cs.CV]* (2016)

[19] Nguyen, Anh M. ; Yosinski, Jason ; Clune, Jeff: Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. In: *arXiv:1412.1897 [cs.CV]* (2014)

[20] Olah, Chris ; Mordvintsev, Alexander ; Schubert, Ludwig: Feature Visualization. In: *Distill* (2017)

[21] Olah, Chris ; Satyanarayan, Arvind ; Johnson, Ian ; Carter, Shan ; Schubert, Ludwig ; Ye, Katherine ; Mordvintsev, Alexander: The Building Blocks of Interpretability. In: *Distill* (2018)

[22] Russakovsky, Olga ; Deng, Jia ; Su, Hao ; Krause, Jonathan ; Satheesh, Sanjeev ; Ma, Sean ; Huang, Zhiheng ; Karpathy, Andrej ; Khosla, Aditya ; Bernstein, Michael ; Berg, Alexander C. ; Fei-Fei, Li: ImageNet Large Scale Visual Recognition Challenge. In: *International Journal of Computer Vision (IJCV)* (2015), Nr. 3, S. 211–252

[23] Simonyan, Karen ; Vedaldi, Andrea ; Zisserman, Andrew: Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In: *arXiv:1312.6034 [cs.CV]* (2013)

[24] Simonyan, Karen ; Zisserman, Andrew: Very Deep Convolutional Networks for Large-Scale Image Recognition. In: *arXiv:1409.1556 [cs.CV]* (2015)

[25] Stanley, Kenneth: Compositional pattern producing networks: A novel abstraction of development. In: *Genetic Programming and Evolvable Machines* (2007), S. 131–162

[26] Stanley, Kenneth O. ; Miikkulainen, Risto: Efficient Evolution Of Neural Network Topologies. In: *Proceedings of the Genetic and Evolutionary Computation Conference* (2002), S. 1757–1762

[27] Szegedy, Christian ; Liu, Wei ; Jia, Yangqing ; Sermanet, Pierre ; Reed, Scott ; Anguelov, Dragomir ; Erhan, Dumitru ; Vanhoucke, Vincent ; Rabinovich, Andrew: Going Deeper with Convolutions. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015)

[28] Szegedy, Christian ; Zaremba, Wojciech ; Sutskever, Ilya ; Bruna, Joan ; Erhan, Dumitru ; Goodfellow, Ian ; Fergus, Rob: Intriguing properties of neural networks. In: *arXiv:1312.6199 [cs.CV]* (2014)

[29] Ulyanov, Dmitry ; Vedaldi, Andrea ; Lempitsky, Victor S.: Deep Image Prior. In: *arXiv:1711.10925 [cs.CV]* (2017)

[30] Zach, Juri: Entwicklung einer Softwareumgebung zum Interpretieren von neuronalen Netzen. In: *Hochschule für Angewandte Wissenschaften Hamburg (HAW Hamburg)* (2020)

[31] Zeiler, Matthew D. ; Fergus, Rob: Visualizing and Understanding Convolutional Networks. In: *Computer Vision – ECCV 2014* (2014), S. 818–833

[32] Zhang, L. ; Zhang, L. ; Mou, X. ; Zhang, D.: FSIM: A Feature Similarity Index for Image Quality Assessment. In: *IEEE Transactions on Image Processing* (2011), S. 2378–2386

[33] Zhang, Tian ; Ramakrishnan, Raghu ; Livny, Miron: BIRCH: An Efficient Data Clustering Method for Very Large Databases. In: *Data Mining and Knowledge Discovery* (1996), S. 141–182

[34] Zhou Wang ; Bovik, A. C. ; Sheikh, H. R. ; Simoncelli, E. P.: Image quality assessment: from error visibility to structural similarity. In: *IEEE Transactions on Image Processing* (2004), S. 600–612

A Aufbau der neuronalen Netze

In diesem Abschnitt wird der Aufbau des Zahlen-Netzes und des VGG16-Netzes in Tabellenform dargestellt.

A.1 Aufbau des Zahlen-Netzes

Name	Schicht Type	Eigenschaften
conv1	Faltung	nFilter: 32, Kernelgröße: 5x5, Aktivierung: Relu
conv2	Faltung	nFilter: 32, Kernelgröße: 5x5, Aktivierung: Relu
pool1	Pooling	Max Pooling
drop1	Dropout	Dropout-Rate: 0.25
conv3	Faltung	nFilter: 64, Kernelgröße: 5x5, Aktivierung: Relu
conv4	Faltung	nFilter: 64, Kernelgröße: 5x5, Aktivierung: Relu
pool2	Pooling	Max Pooling
drop2	Dropout	Dropout-Rate: 0.25
flat	Flatten	
dense1	Dense	Neuronen: 256, Aktivierung: Relu
drop3	Dropout	Dropout-Rate: 0.5
out	Dense	Neuronen: 10, Aktivierung: Softmax

Tabelle A.1: Architektur des Zahlen-Netzes.

A.2 Aufbau des VGG16-Netzes

Name	Schicht Type	Eigenschaften
block1 conv1	Faltung	nFilter: 64, Kernelgröße: 3x3, Aktivierung: Relu
block1 conv2	Faltung	nFilter: 64, Kernelgröße: 3x3, Aktivierung: Relu
block1 pool	Pooling	Max Pooling
block2 conv1	Faltung	nFilter: 128, Kernelgröße: 3x3, Aktivierung: Relu
block2 conv2	Faltung	nFilter: 128, Kernelgröße: 3x3, Aktivierung: Relu
block2 pool	Pooling	Max Pooling
block3 conv1	Faltung	nFilter: 256, Kernelgröße: 3x3, Aktivierung: Relu
block3 conv2	Faltung	nFilter: 256, Kernelgröße: 3x3, Aktivierung: Relu
block3 conv3	Faltung	nFilter: 256, Kernelgröße: 3x3, Aktivierung: Relu
block3 pool	Pooling	Max Pooling
block4 conv1	Faltung	nFilter: 512, Kernelgröße: 3x3, Aktivierung: Relu
block4 conv2	Faltung	nFilter: 512, Kernelgröße: 3x3, Aktivierung: Relu
block4 conv3	Faltung	nFilter: 512, Kernelgröße: 3x3, Aktivierung: Relu
block4 pool	Pooling	Max Pooling
block5 conv1	Faltung	nFilter: 512, Kernelgröße: 3x3, Aktivierung: Relu
block5 conv2	Faltung	nFilter: 512, Kernelgröße: 3x3, Aktivierung: Relu
block5 conv3	Faltung	nFilter: 512, Kernelgröße: 3x3, Aktivierung: Relu
block5 pool	Pooling	Max Pooling
flat	Flatten	
fc1	Dense	Neuronen: 4096, Aktivierung: Relu
fc2	Dense	Neuronen: 4096, Aktivierung: Relu
predictions	Dense	Neuronen: 1000, Aktivierung: Softmax

Tabelle A.2: Architektur des VGG16-Netzes.

Erklärung zur selbstständigen Bearbeitung einer Abschlussarbeit

Gemäß der Allgemeinen Prüfungs- und Studienordnung ist zusammen mit der Abschlussarbeit eine schriftliche Erklärung abzugeben, in der der Studierende bestätigt, dass die Abschlussarbeit „— bei einer Gruppenarbeit die entsprechend gekennzeichneten Teile der Arbeit [(§ 18 Abs. 1 APSO-TI-BM bzw. § 21 Abs. 1 APSO-INGI)] — ohne fremde Hilfe selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel benutzt wurden. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich zu machen."

Quelle: § 16 Abs. 5 APSO-TI-BM bzw. § 15 Abs. 6 APSO-INGI

Erklärung zur selbstständigen Bearbeitung der Arbeit

Hiermit versichere ich,

Name: _____

Vorname: _____

dass ich die vorliegende Masterarbeit – bzw. bei einer Gruppenarbeit die entsprechend gekennzeichneten Teile der Arbeit – mit dem Thema:

Entwicklung einer Qualitätsmetrik für Interpretationen von neuronalen Netzen

ohne fremde Hilfe selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel benutzt habe. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich gemacht.

_____	_____	_____
Ort	Datum	Unterschrift im Original