

BACHELORARBEIT

Contrastive Learning mit Stable Diffusion-basierter Datenaugmentation

Verbesserung der Bildklassifikation
durch synthetische Daten

vorgelegt am 16. September 2024
Paul Hofmann

Erstprüferin: Prof. Dr. Larissa Putzar
Zweitprüfer: Prof. Dr. Jan Neuhöfer

**HOCHSCHULE FÜR ANGEWANDTE
WISSENSCHAFTEN HAMBURG**
Department Medientechnik
Finkenau 35
22081 Hamburg

Zusammenfassung

Der Arbeit beginnt mit einer kurzen Beschreibung ihrer zentralen Inhalte, in der die Thematik und die wesentlichen Resultate skizziert werden. Diese Beschreibung muss sowohl in deutscher als auch in englischer Sprache vorliegen und sollte eine Länge von etwa 150 bis 250 Wörtern haben. Beide Versionen zusammen sollten nicht mehr als eine Seite umfassen. Die Zusammenfassung dient u. a. der inhaltlichen Verortung im Bibliothekskatalog.

Abstract

The thesis begins with a brief summary of its main contents, outlining the subject matter and the essential findings. This summary must be provided in German and in English and should range from 150 to 250 words in length. Both versions combined should not comprise more than one page. Among other things, the abstract is used for library classification.

Inhaltsverzeichnis

Abbildungsverzeichnis	III
Tabellenverzeichnis	IV
1 Einleitung	1
1.1 Motivation	1
1.2 Zielsetzung	1
1.3 Aufbau der Arbeit	1
2 Theoretische Grundlagen	2
2.1 Maschinelles Lernen	2
2.1.1 Überwachtes und unüberwachtes Lernen	3
2.1.2 Deep Learning	3
2.1.3 Neuronale Netze	4
2.1.4 Out-of-Distribution Daten	6
2.1.5 Datenaugmentation und Generalisierung	7
2.2 Synthetische Daten	7
2.2.1 Variational Autoencoder	8
2.2.2 Generative Adversarial Networks	9
2.2.3 Diffusion Models	10
2.3 Semantische Datenaugmentation mit DA-Fusion	11
2.4 Robuste Datenrepräsentation durch Contrastive Learning	12
2.4.1 Unsupervised Contrastive Learning	13
2.4.2 Supervised Contrastive Learning	14
2.5 Forschungslücke	14
2.5.1 Herausforderungen bei der Generierung synthetischer Daten	14
2.5.2 Synthetische Daten als negativ-Beispiele im Contrastive Learning . .	14
2.5.3 Integration von DA-Fusion und Supervised Contrastive Learning . .	14
3 Methodisches Vorgehen	15
3.1 Forschungsfragen und Hypothesen	15
3.2 Datensatz	16
3.2.1 EIBA	16

3.2.2	Teildatensatz	16
3.2.3	Vorverarbeitung	17
3.3	Implementierung	17
3.3.1	DA-Fusion	17
3.3.2	Supervised Contrastive Learning	17
3.4	Synthetische Datengenerierung mit DA-Fusion	18
3.5	Trainings- und Testdurchläufe mit Supervised Contrastive Learning	18
3.6	Evaluationsmethoden und Metriken	18
4	Ergebnisse	19
4.1	Die generierten synthetischen Daten	19
4.1.1	In-Distribution	19
4.1.2	Near Out-of-Distribution	19
4.2	Trainings- und Testergebnisse mit Supervised Contrastive Leraning	20
4.2.1	Contrastive Pre-Training	20
4.2.2	Lineare Klassifikation	20
4.3	Vergleich der Ergebnisse mit und ohne In-Distribution-Augmentationen	20
4.4	Vergleich der Ergebnisse mit und ohne Near Out-of-Distribution-Augmentationen als Hard Negatives	20
5	Diskussion	21
5.1	Eignung von DA-Fusion für die synthetische Datengenerierung	21
5.2	Wirksamkeit von synthetischen Near Out-of-Distribution-Daten im Supervised Contrastive Learning	21
6	Fazit	22
6.1	Zusammenfassung der wichtigsten Erkenntnisse	22
6.2	Beantwortung der Forschungsfragen	22
6.3	Ausblick und potenzielle Weiterentwicklungen	22
Literatur		23
Anhang		24

Abbildungsverzeichnis

2.1	Darstellung eines einfachen neuronalen Netzes (O'Shea & Nash, 2015).	4
2.2	Das McCulloch-Pitts-Modell eines Neurons.	5
2.3	Beispiele für Datenaugmentationstechniken.	7
2.4	Überblick über die GAN-Struktur. Quelle: Google for Developers	9
2.5	Vergleich zwischen semantischen Augmentationen aus Baseline-Methode und DA-Fusion (Trabucco et al., 2023).	11
4.1	Beispieltext	19

Tabellenverzeichnis

1 Einleitung

1.1 Motivation

...

1.2 Zielsetzung

...

1.3 Aufbau der Arbeit

...

2 Theoretische Grundlagen

Im folgenden Kapitel werden die theoretischen Grundlagen des maschinellen Lernens und der verwendeten Modelle erläutert. Es wird auf die Konzepte des maschinellen Lernens, insbesondere des überwachten und unüberwachten Lernens, des Deep Learnings und der neuronalen Netze eingegangen. Anschließend wird die Funktionsweise von Diffusion-Modellen, insbesondere Stable Diffusion und DA-Fusion, sowie von Contrastive Learning und Supervised Contrastive Learning beschrieben. Zuletzt wird die bestehende Forschungslücke und die in dieser Arbeit thematisierte Integration von DA-Fusion und Supervised Contrastive Learning diskutiert.

2.1 Maschinelles Lernen

Die ersten großen Durchbrüche in der künstlichen Intelligenz (KI) kamen im Bezug auf Aufgaben, die für Menschen intellektuell eine große Herausforderung darstellten, die aber von Computern relativ einfach zu lösen waren, da sie als Liste formaler, mathematischer Regeln beschrieben werden konnten. Die große Schwierigkeit lag hingegen in den Aufgaben, die für Menschen relativ einfach und intuitiv sind, welche sich aber nur schwer formal beschreiben lassen. Hierunter fallen z.B. die Spracherkennung, oder Objekterkennung (Goodfellow et al., 2016).

Maschinelles Lernen (ML) beschreibt den Ansatz, Computer mit der Fähigkeit auszustatten, selbstständig Wissen aus Erfahrung zu generieren, indem Muster und Konzepte aus rohen Daten erlernt werden. So kann ein Computerprogramm auf Basis von Beispielen lernen, wie es eine bestimmte Aufgabe lösen soll, ohne dass ihm explizit Regeln oder Algorithmen vorgegeben werden.

Eine allgemeine Definition für maschinelles Lernen bietet (Mitchell, 1997):

Ein Computerprogramm soll aus Erfahrung E in Bezug auf eine Klasse von Aufgaben T und Leistungsmaß P lernen, wenn sich seine Leistung bei Aufgaben T , gemessen durch P , mit Erfahrung E verbessert.

Die Erfahrung E besteht dabei aus einer Menge von Trainingsdaten, die etwa aus Eingabe-Ausgabe-Paaren bestehen. Die Aufgaben T können sehr vielfältig sein, von einfachen Klassifikations- und Regressionsaufgaben bis hin zu komplexen Problemen wie Spracherkennung oder autonomen Fahren. Das Leistungsmaß P gibt an, wie gut das Modell die Aufgaben T löst, und kann z.B. die Genauigkeit (engl. *accuracy*) einer Klassifikation oder die mittlere quadratische Abweichung bei einer Regression sein.

2.1.1 Überwachtes und unüberwachtes Lernen

Wie genau Wissen aus Erfahrung bzw. aus Rohdaten generiert wird hängt vom gewählten Verfahren ab. Im Maschinellen Lernen gibt es dabei verschiedene Paradigmen, wobei die wichtigsten das überwachte (engl. *supervised*) und das unüberwachte (engl. *unsupervised*) Lernen sind.

Beim überwachten Lernen wird das Modell mit einem vollständig annotierten Datensatz trainiert. Das heißt, jeder Datenpunkt ist mit einem Klassenlabel versehen, sodass Eingabe-Ausgabe-Paare entstehen. Das Ziel ist es, eine Funktion zu lernen, die Eingaben auf die entsprechenden Ausgaben abbildet. Beispiele für überwachtes Lernen sind Klassifikations- und Regressionsaufgaben. Ein typisches Beispiel ist die Bilderkennung, bei der ein Modell darauf trainiert wird, Bilder von Katzen und Hunden zu unterscheiden. (<empty citation>)

Im Gegensatz dazu arbeitet unüberwachtes Lernen mit unbeschrifteten Daten; es gibt also keine vorgegebenen Ausgaben. Stattdessen wird versucht, ein Modell zu befähigen, eigenständig Muster und Strukturen in den Daten zu erkennen und z.B. nützliche Repräsentationen der Eingangsdaten zu erlernen. Zu den häufigsten Methoden des unüberwachten Lernens gehören Clustering- und Assoziationsalgorithmen. Ein Beispiel ist die Segmentierung von Kunden in verschiedene Gruppen basierend auf ihrem Kaufverhalten. (<empty citation>)

In der Praxis werden oft auch hybride Ansätze genutzt, wie das semi-überwachte Lernen, bei dem eine Kombination aus beschrifteten und unbeschrifteten Daten verwendet wird, oder das selbstüberwachte Lernen, bei dem das Modell eigenständig Teile der Daten zur Erzeugung von Überwachungssignalen verwendet, anstatt sich auf externe, von Menschen bereitgestellte Labels zu verlassen. (<empty citation>)

2.1.2 Deep Learning

Das Wissen, das ein Modell aus den Trainingsdaten lernt, wird in Form von Merkmalen (engl. *features*) repräsentiert. Diese Merkmale können einfache Konzepte wie Kanten oder Farben sein, oder komplexere Konzepte wie Gesichter oder Objekte. Unter Deep Learning versteht man eine tiefe, hierarchische Vernetzung dieser Konzepte, sodass komplexere Konzepte auf

simpleren Konzepten aufbauen können. Visuell veranschaulicht entsteht ein Graph mit vielen Ebenen (Goodfellow et al., 2016). Deep Learning ist daher eine spezialisierte Unterkategorie des maschinellen Lernens, in der künstlichen neuronale Netzen mit mehreren Schichten verwendet werden, um eine hierarchische Repräsentation von Daten zu ermöglichen. Jede Schicht transformiert die Eingabedaten in eine etwas abstraktere Darstellung.

Deep Learning hat in den letzten Jahren erhebliche Fortschritte gemacht und findet Anwendung in Bereichen wie Bild- und Spracherkennung, autonomen Fahrzeugen und vielen anderen. Die Popularität von Deep Learning ist auf mehrere Faktoren zurückzuführen, darunter die Verfügbarkeit großer Datensätze, die Leistungsfähigkeit moderner Hardware und die Entwicklung effizienterer Algorithmen. (Zhou, 2021)

2.1.3 Neuronale Netze

Während die rasante Entwicklung von Deep Learning vor allem in den vergangenen Jahren spürbar geworden ist, sind die zugrundeliegenden Algorithmen und Konzepte schon seit Jahrzehnten bekannt (Zhou, 2021). Dabei bildet das künstliche neuronale Netz (KNN) die Grundlage der allermeisten Deep-Learning-Modelle. Es ist inspiriert von der Struktur und Funktionsweise des menschlichen Gehirns und besteht aus einer Vielzahl von miteinander verbundenen Knoten (Neuronen), die in Schichten organisiert sind. Die Struktur eines neuronalen Netzes besteht aus einer Eingabeschicht, einer oder mehreren verdeckten Schichten (engl. *hidden layers*) und einer Ausgabeschicht, wie in Abbildung 2.1 dargestellt.

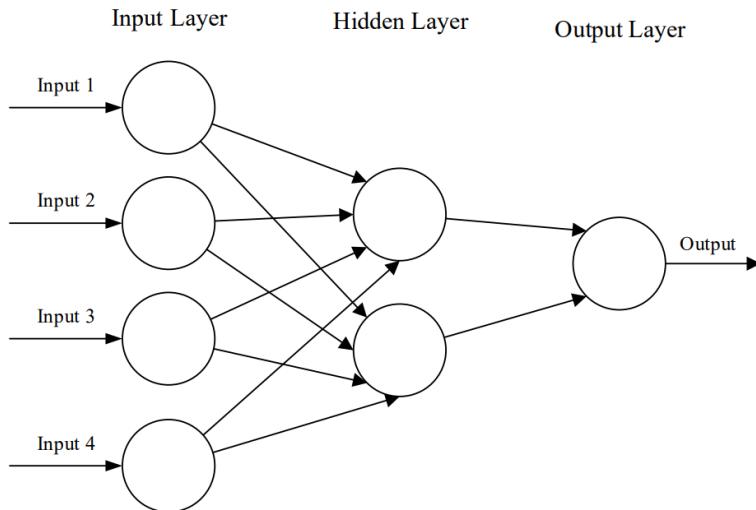


Abbildung 2.1: Darstellung eines einfachen neuronalen Netzes (O'Shea & Nash, 2015).

Die einzelnen Neuronen, auf dem diese Netze aufbauen, sind eine mathematische Modellierung des biologischen Neurons, das erstmals 1943 von Warren McCulloch und Walter Pitts vorgestellt wurde (Zhou, 2021). Jedes Neuron empfängt eine Reihe von Eingaben $x_1 \dots n$, entweder von externen Quellen oder von den Ausgaben anderer Neuronen. Für jede dieser Eingaben gibt es zugehörige Gewichtungen (engl. *weights*) $w_{1j \dots nj}$, welche die Stärke und Richtung (positiv oder negativ) des Einflusses der jeweiligen Eingaben auf das Neuron bestimmen.

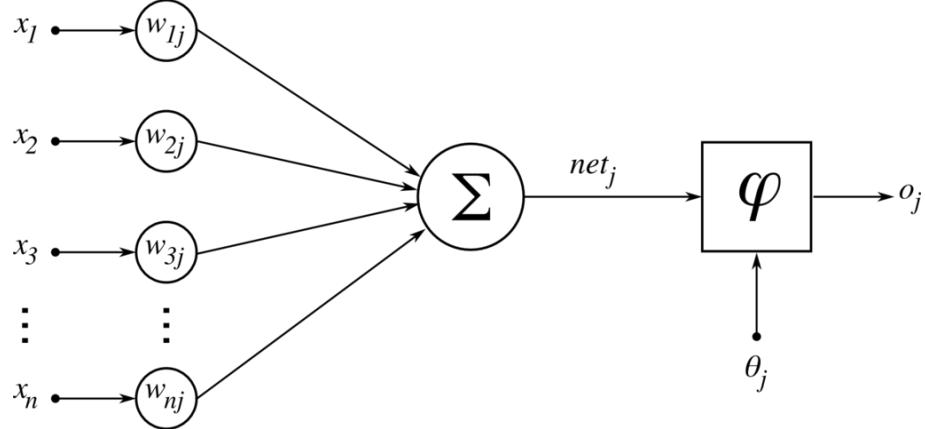


Abbildung 2.2: Das McCulloch-Pitts-Modell eines Neurons.

Das Neuron berechnet dann die gewichtete Summe aller Eingaben und falls ein bestimmter Schwellenwert (engl. *bias*) überschritten wurde, wird das Neuron aktiviert:

$$o_j = \phi \left(\sum_{i=1}^n w_i x_i - \theta_j \right) \quad (2.1)$$

Die Aktivierungsfunktion ϕ kann dabei unterschiedlich gewählt werden, um die Ausgabe des Neurons zu modellieren. Eine simples Beispiel ist die sogenannte Schwellenwertfunktion (engl. *step function*), die den Wert 1 zurückgibt, wenn die gewichtete Summe größer als der Schwellenwert ist, sonst 0. Eine häufig verwendete Aktivierungsfunktion ist jedoch die sogenannte Sigmoid-Funktion, die kontinuierlich und differenzierbar ist und somit die Optimierung des Netzwerks vereinfacht:

$$\text{Sigmoid}(x) = \frac{1}{1 + e^{-x}} \quad (2.2)$$

Während die Sigmoid-Funktion allerdings nur für binäre Klassifikationen geeignet ist, wird für die Klassifikation von mehreren Klassen die Softmax-Funktion verwendet, die die Wahrscheinlichkeitsverteilung über alle Klassen berechnet:

$$\text{Softmax}(x)_i = \frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}} \quad (2.3)$$

Im Training fließen die Eingabedaten in einer Vorwärtsausbreitung (engl. *forward propagation*) durch das Netzwerk, um die Ausgabe zu berechnen. Es wird dann eine geeignete Verlustfunktion (engl. *loss function*) angewendet, um den Fehler des Modells zu berechnen. Das Ziel des Trainings ist es, die Gewichtungen der Neuronen so anzupassen, dass der Fehler minimiert wird.

Diese Optimierung geschieht durch eine Rückwärtsausbreitung (engl. *backpropagation*), welche den berechneten Fehler rückwärts durch das Netz propagiert, um die Gewichte und Schwellenwerte um einen geringen Wert in die Richtung anzupassen, die den Fehler minimieren würde. Dieser Prozess wird in der Regel mit dem Stochastic Gradient Descent (SGD) Algorithmus durchgeführt, der die Gewichtungen iterativ anpasst, um den Fehler zu minimieren.

2.1.4 Out-of-Distribution Daten

Wenn ein KI-Modell mit Daten konfrontiert wird, die außerhalb des Bereichs liegen, den es während des Trainings gesehen hat, spricht man von Out-of-Distribution (OOD) Daten. Es handelt sich also um Datenpunkte oder Muster, die sich signifikant von den Trainingsdaten unterscheiden. Dies kann zu Problemen führen, da das Modell möglicherweise nicht in der Lage ist, angemessene Vorhersagen oder Entscheidungen für diese Daten zu treffen. Stattdessen werden falsche Vorhersagen mit übermäßigem Vertrauen getroffen.

Die Erkennung von OOD-Daten ist ein wichtiges Forschungsgebiet im maschinellen Lernen, da sie dazu beitragen kann, die Zuverlässigkeit und Sicherheit von KI-Systemen zu verbessern. Idealerweise sollte ein neuronales Netz höhere Softmax-Wahrscheinlichkeiten für In-Distribution-Daten und niedrigere Wahrscheinlichkeiten für OOD-Daten ausgeben. Durch Festlegen eines Schwellenwerts für diese Wahrscheinlichkeiten können Instanzen unterhalb des Schwellenwerts frühzeitig als OOD-Instanzen erkannt und entsprechend behandelt werden. In der Praxis kommt dieser Ansatz jedoch oft an seine Grenzen, da die Softmax-Wahrscheinlichkeiten nicht immer zuverlässig sind und das Modell auch für OOD-Daten hohe Wahrscheinlichkeiten ausgeben kann. Daher werden alternative Ansätze verwendet, wie etwa das Training eines binären Klassifikationsmodells zur Unterscheidung von In-Distribution und OOD-Daten.

2.1.5 Datenaugmentation und Generalisierung

Datenaugmentation ist ein wichtiger Schritt im Training von neuronalen Netzen, insbesondere bei begrenzten Datensätzen. Sie bezieht sich auf die künstliche Erweiterung des Trainingsdatensatzes durch Anwenden von Transformationen auf die vorhandenen Daten. Diese Transformationen können z.B. Rotation, Skalierung, Verschiebung, Spiegelung, Helligkeitsanpassung oder Rauschen sein. Das Ziel der Datenaugmentation ist es, das Modell robuster gegenüber Variationen in den Eingabedaten zu machen und die Generalisierungsfähigkeit zu verbessern.

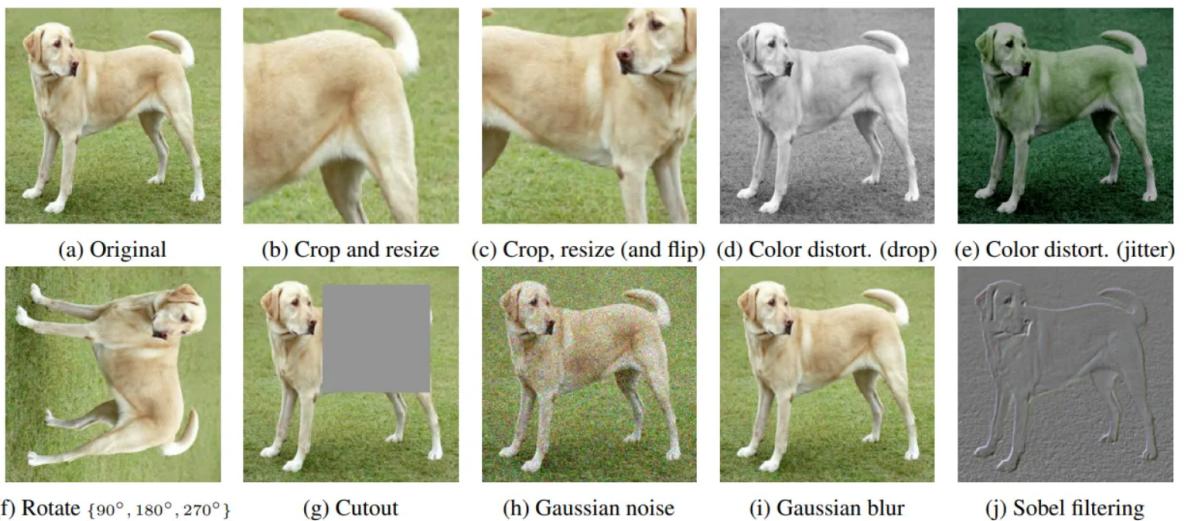


Abbildung 2.3: Beispiele für Datenaugmentationstechniken. (<empty citation>)

2.2 Synthetische Daten

Während die Verfügbarkeit großer Datensätze für das Training von neuronalen Netzen ein entscheidender Faktor für den Erfolg von Deep Learning-Modellen ist, ist es oft schwierig, solche Datensätze zu sammeln, insbesondere in Domänen wie der Medizin oder der Robotik, wo die Daten rar und teuer sind (<empty citation>). In solchen Fällen können synthetische Daten eine nützliche Alternative oder Ergänzung zu echten Daten sein.

Synthetische Daten sind künstlich erzeugte Daten, welche die zugrundeliegenden Muster der realen Daten nachahmen. Sie können durch Simulation, Generierung oder Transformation von echten Daten erstellt werden.

...

2.2.1 Variational Autoencoder

Ein Autoencoder ist eine spezielle Art von KI-Modell, das darauf ausgelegt ist, Daten effizient zu komprimieren und dann wieder zu rekonstruieren. Es besteht aus zwei Hauptkomponenten: (Foster, 2020)

- einem **Encoder**-Netzwerk, das hochdimensionale Eingabedaten in einem niederdimensionalen Darstellungsvektor komprimiert, und
- einem **Decoder**-Netzwerk, das einen gegebenen Darstellungsvektor zurück in den ursprünglichen hochdimensionalen Raum umwandelt

Der Darstellungsvektor ist eine Kompression des Originalbilds in einen niedriger dimensionalen latenten Raum, wodurch es sich beim Autoencoder um eine Form des *Representation Learning* handelt.

Das Training eines Autoencoders erfolgt durch Minimierung des Rekonstruktionsfehlers, der die Differenz zwischen den ursprünglichen Eingabedaten und den rekonstruierten Ausgaben beschreibt. Eine gängige Verlustfunktion hierfür ist der *Mean Squared Error* (MSE):

$$\text{Loss} = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{x}_i)^2 \quad (2.4)$$

Ein besonders interessantes Versprechen des Autoencoders ist, dass man theoretisch durch die Wahl eines beliebigen Punkts im latenten Raum neue Bilder erzeugen kann, indem man diesen Punkt durch den Decoder schickt, da der Decoder gelernt hat, wie man Punkte im latenten Raum in realistische Bilder umwandelt. (Foster, 2020) In der herkömmlichen Form hat der Autoencoder in Bezug auf diese Aufgabe allerdings einige Schwachstellen.

Der **Variational Autoencoder** (VAE) adressiert diese Schwachstellen und verwendet probabilistische Methoden, um die Datenverteilung im latenten Raum zu modellieren; Anstatt einen einzelnen, festen Punkt im latenten Raum für jede Eingabe zu lernen, wird eine Verteilung gelernt, aus der die latenten Variablen für jede Eingabe stammen. Dadurch entsteht ein strukturierter und kontinuierlicher latenter Raum, der es ermöglicht, neue, realistische Daten zu generieren.

...

2.2.2 Generative Adversarial Networks

Ein Generative Adversarial Network (GAN) ist ein KI-Modell, das in (<empty citation>) vorgestellt wurde. GANs bestehen aus zwei neuralen Netzwerken, die gegeneinander antreten, um realistische synthetische Daten zu erzeugen. Diese Technologie hat sich als äußerst mächtig in der Bild- und Datengenerierung erwiesen.

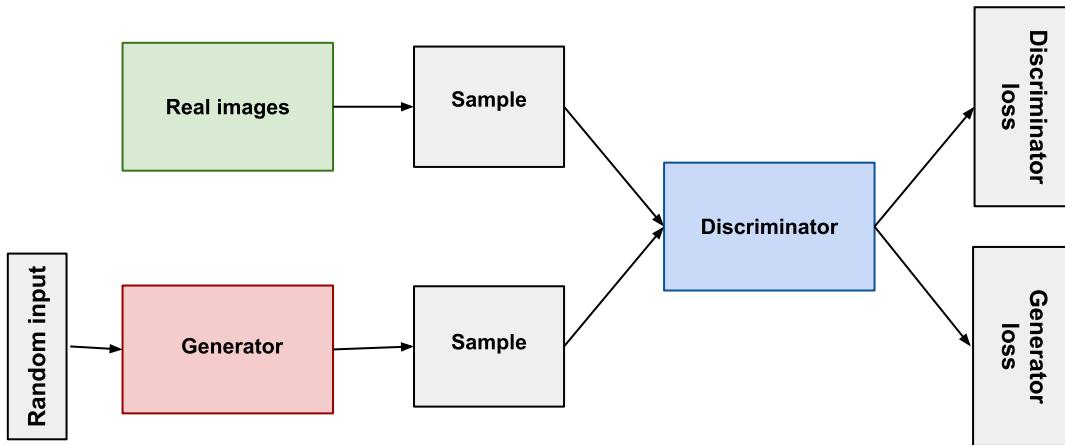


Abbildung 2.4: Überblick über die GAN-Struktur. Quelle: Google for Developers

Die Architektur eines GANs besteht aus zwei Hauptkomponenten:

- **Generator:** Das generative Netzwerk nimmt Zufallsrauschen als Eingabe und erzeugt daraus Daten, die möglichst realistisch wirken sollen. Der Generator versucht, die wahre Datenverteilung zu imitieren und realistische Beispiele zu erstellen.
- **Diskriminator:** Das diskriminative Netzwerk erhält sowohl echte Daten aus dem Trainingsdatensatz als auch die vom Generator erzeugten Daten. Seine Aufgabe ist es, zwischen echten und künstlichen Daten zu unterscheiden. Der Diskriminatior gibt eine Wahrscheinlichkeit aus, dass die Eingabedaten echt sind.

Der Trainingsprozess eines GANs ist als minimax-Spiel zwischen dem Generator und dem Diskriminatior formuliert:

1. **Diskriminatior-Training:** Der Diskriminatior wird trainiert, um echte Daten von generierten Daten zu unterscheiden. Dies geschieht durch eine binäre Klassifikation („echt“ oder „synthetisch“). Der Diskriminatior passt seine Gewichte an, um die Unterscheidung zu verbessern.

2. **Generator-Training:** Der Generator wird trainiert, um den Diskriminatator zu täuschen. Dies geschieht, indem der Generator seine erzeugten Daten durch den Diskriminatator laufen lässt und die Rückmeldung (Gradienten) des Diskriminators verwendet, um seine eigenen Parameter zu optimieren. Ziel ist es, den Diskriminatator zu überlisten, sodass er die generierten Daten als echt klassifiziert.

...

2.2.3 Diffusion Models

In den letzten Jahren haben sogenannte Diffusion Models für einen enormen Sprung in der Bildgenerierung, insbesondere der Text-to-Image (T2I) Generierung, gesorgt. Das Konzept der Diffusion soll hier deshalb einmal genauer erläutert werden.

Unter Diffusion versteht man den Prozess der langsamen Vermischung von Partikeln oder Informationen über die Zeit. So beschreibt die Diffusionsgleichung in der Physik die zeitliche Entwicklung der Dichte von Teilchen, die sich zufällig bewegen. Im maschinellen Lernen fand das Konzept erstmals in (<empty citation>) Anwendung. Es entstand eine neue Klasse von generativen Deep Learning-Modellen, die Diffusion Models, welche im Trainingsprozess schrittweise die Struktur der Eingabedaten durch das Hinzufügen von Rauschen auflösen und anschließend darauf trainiert werden, das ursprüngliche Bild aus dem verrauschten Bild zu rekonstruieren.

Das Training von Diffusion Models teilt sich somit in zwei Phasen auf, die Vorwärts- und die Rückwärtsdiffusion. Beide Phasen ...

$$q(x^{(0\dots T)}) = q(x^{(0)}) \prod_{t=1}^T q(x^{(t)}|x^{(t-1)}) \quad (2.5)$$

In der Rückwärtsdiffusion geht das Bild durch ein Modell, welches selbst nur Rauschen generiert. Es ist die Vorhersage darüber, welches Rauschen vom Eingang entfernt werden soll, um das gewünschte entrauschte Bild zu erhalten.

$$p(x^{(0\dots T)}) = p(x^{(T)}) \prod_{t=1}^T p(x^{(t-1)}|x^{(t)}) \quad (2.6)$$

Anders als bei GANs gibt es keine direkten adversarialen Optimierungsmechanismen, die zu einem Ungleichgewicht führen können. Es wird stattdessen explizit die Wahrscheinlichkeitsdichte zwischen den realen Daten und den erzeugten Daten minimiert, was zu einem robusteren und stabileren Trainingsprozess führt.

Eine entscheidende Weiterentwicklung der Diffusion Models war die Text-Konditionierung des generativen Prozesses. In (<empty citation>) wurde DALL-E vorgestellt, ein Modell, das es ermöglicht, Bilder aus Textbeschreibungen zu generieren. DALL-E ...

...

2.3 Semantische Datenaugmentation mit DA-Fusion

In (Trabucco et al., 2023) wird DA-Fusion, eine flexible, auf Stable Diffusion basierende Methode zur Datenaugmentation vorgestellt, die für diese Arbeit von besonderem Interesse ist. Der traditionelle Ansatz der Datenaugmentation, wie in Abschnitt ?? beschrieben, hat sich als effektiv erwiesen, um die Generalisierungsfähigkeit von Modellen zu verbessern. Allerdings erfordert dieser Ansatz auch eine gute Intuition in Bezug auf den verwendeten Datensatz, um zu vermeiden, dass Transformationen gewählt werden, durch die Informationen verloren gehen, die für die Aufgabe des zu trainierenden Modells wichtig sind. Wenn beispielsweise Farbinformationen für die Klassifizierung von Blumen wichtig sind, könnte die Datenaugmentation durch zufällige Farbänderungen die Leistung des Modells verschlechtern. Ein weiteres Beispiel sind Objekte, die klein im Bild sind und durch zufällige Ausschnitte des Bildes aus der Sicht des Modells verschwinden können. DA-Fusion hingegen nutzt das Wissen eines vortrainierten Diffusion Models, um den Bildinhalt semantisch zu verstehen und automatisch neue, realistische Variationen zu generieren.

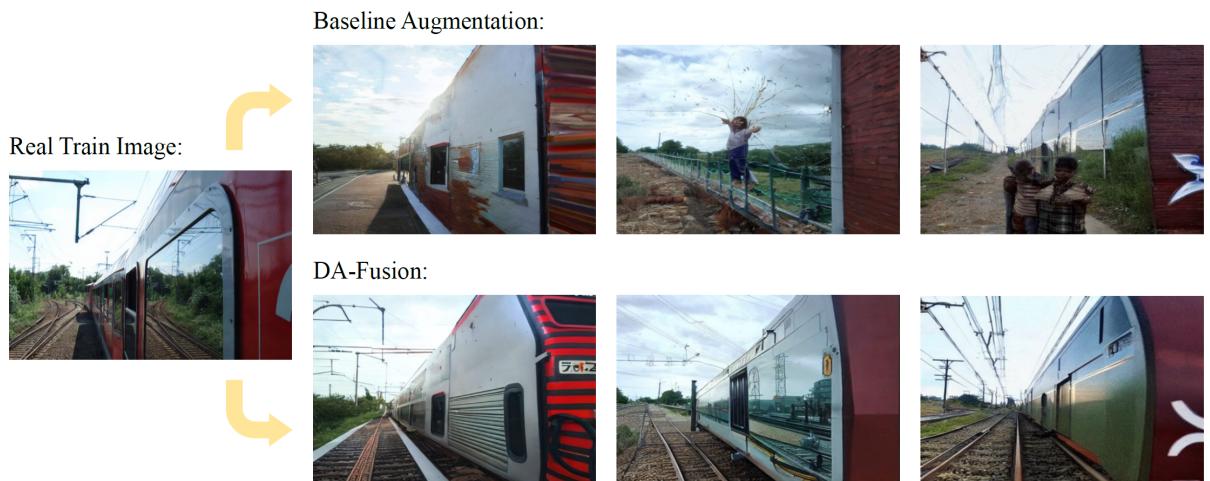


Abbildung 2.5: Vergleich zwischen semantischen Augmentationen aus Baseline-Methode und DA-Fusion (Trabucco et al., 2023).

Es wird zunächst die Methode Textual Inversion (<empty citation>) angewendet, um ein vortrainiertes Stable Diffusion-Modell auf den gegebenen Datensatz feinabzustimmen. Dazu

wird für jedes neue Konzept bzw. für jede neue Klasse ein Pseudo-Token y in das Modell integriert, der unter Verwendung von Trainings-Prompts wie „a photo of a $< y >$ “ und den zugehörigen Bilddaten trainiert wird.

Anschließend können aus den echten Bildern neue Augmentationen generiert werden, indem den Bildern eine geringe Menge an Rauschen hinzugefügt wird, welches dann wieder durch das feinabgestimmte Modell entfernt werden soll. Hier kommen wieder die selben Text-Prompts zum Einsatz. Auf diese Weise bleibt die grundlegende Struktur der Bilder erhalten, während gleichzeitig neue, semantisch sinnvolle Variationen und Details erzeugt werden.

Ein Vorteil von DA-Fusion ist die Möglichkeit, den Grad der Augmentation durch die Wahl des Insertion Timesteps zu steuern. Der Insertion Timestep bestimmt, wie weit in den Diffusionsprozess das Bild eingefügt wird und wie stark es dafür vorher verrauscht werden muss. Ein niedriger Timestep führt zu stärkeren Augmentationen, während ein hoher Timestep subtilere Variationen erzeugt.

2.4 Robuste Datenrepräsentation durch Contrastive Learning

Neben den vorgestellten Methoden zur Vervielfältigung der Trainingsdaten soll nun auch das Contrastive Learning als effektive Methode zur Verbesserung der Generalisierungsfähigkeit und Robustheit von Modellen vorgestellt werden. Die Methode zielt darauf ab, ähnliche Beispiele im Datensatz zu gruppieren und unähnliche Beispiele voneinander zu trennen. Durch die Kontrastierung der Daten wird eine Art von Supervision erzeugt, die es dem Modell ermöglicht, nützliche Repräsentationen der Eingabedaten zu lernen.

Contrastive Learning kommt ursprünglich aus dem unüberwachten Lernen. Da die Annotation von Daten sehr aufwendig sein kann, insbesondere in Domänen, in denen Expertenwissen erforderlich ist, hat sich das Contrastive Learning als vielversprechende Alternative mit äußerst starker Generalisierungsfähigkeit und Robustheit gegenüber Adversarial Attacks erwiesen (Liu, 2021).

Mittlerweile gibt es vermehrt Ansätze, Contrastive Learning auch im überwachten Setting anzuwenden, um die Repräsentationen von Daten zu verbessern. Während das Modell im unüberwachten Setting lernt, zwischen einzelnen Instanzen zu unterscheiden, werden im überwachten Setting die Klassenzugehörigkeit der Beispiele berücksichtigt.

In den folgenden Abschnitten wird genauer auf die Funktionsweise von sowohl unüberwachten als auch überwachten Varianten des Contrastive Learning eingegangen, die in den letzten Jahren vielversprechende Ergebnisse erzielt haben.

2.4.1 Unsupervised Contrastive Learning

Ob unüberwacht oder überwacht: Im Contrastive Learning soll die Distanz ähnlicher Beispiele in einem Repräsentationsraum minimiert und die Distanz unähnlicher Beispiele maximiert werden. Dazu bildet das Modell kontrastive Paare aus einem Anchor-Beispiel und verschiedenen positiv- oder negativ-Beispielen. Je nach Methode kann vor allem die Anzahl der positiv- und negativ-Beispiele pro Anchor variieren, wodurch sich auch die jeweiligen Loss-Funktionen unterscheiden.

Das wohl prominenteste Beispiel für Contrastive Learning ist **SimCLR**, das in (Chen et al., 2020) vorgestellt wurde. SimCLR verwendet den NT-Xent Loss (Normalized Temperature-scaled Cross-Entropy Loss), der die Ähnlichkeiten zwischen allen Paaren im Batch (auch *Mini-Batch*) berücksichtigt, anstatt nur einzelne Triplets oder Paare. Jedes Beispiel wird dabei zweimal augmentiert, um zwei Ansichten zu erzeugen, welche als positives Paar für das jeweilige Beispiel dienen. Alle anderen Beispiele (bzw. dessen Ansichten) im Batch werden als negativ-Beispiele gesehen.

Die Eingabedaten werden durch ein Convolutional Neural Network (CNN) in eine hochdimensionale Repräsentation transformiert. Dieser Schritt wird auch als *Feature Extraction* bezeichnet. SimCLR verwendet anschließend einen sogenannten *Projection Head*, der die encodierten Repräsentationen weiter transformiert, um einen Repräsentationsraum zu erzeugen, der für die Unterscheidung der Beispiele geeignet ist. In diesem Representationsraum werden die Ähnlichkeitswerte der Paare berechnet, um den Loss zu bestimmen.

Dafür wird die Kosinus-Ähnlichkeit $s_{i,j}$ der Paare z_i und z_j berechnet:

$$s_{i,j} = \frac{z_i \cdot z_j}{\|z_i\| \cdot \|z_j\|} \quad (2.7)$$

Der Loss ergibt sich dann aus der Berechnung des Softmax über die Ähnlichkeiten aller Paare im Batch, skaliert mit einem Temperaturparameter, um die Unterscheidung zwischen positiven und negativen Paaren hervorzuheben.

Durch die Verwendung aggressiver Datenaugmentation zur Erzeugung der zwei Ansichten wird die Robustheit der Repräsentationen verbessert. Größere Batch Sizes und längere Trainingszeiten begünstigen die Lernfähigkeit des Modells. Besonders die Wahl von *Hard Negatives*, also von Paaren, welche ähnliche Konzepte darstellen, aber sehr unterschiedlich aussehen, hat sich als entscheidend für den Erfolg des Modells erwiesen.

2.4.2 Supervised Contrastive Learning

...

2.5 Forschungslücke

...

2.5.1 Herausforderungen bei der Generierung synthetischer Daten

...

2.5.2 Synthetische Daten als negativ-Beispiele im Contrastive Learning

...

2.5.3 Integration von DA-Fusion und Supervised Contrastive Learning

...

3 Methodisches Vorgehen

In diesem Kapitel wird das methodische Vorgehen der Arbeit beschrieben. Es wird auf die Forschungsfragen und Hypothesen eingegangen, der verwendete Datensatz vorgestellt und die Implementierung der Modelle DA-Fusion und Supervised Contrastive Learning erläutert. Anschließend wird die synthetische Datengenerierung mit DA-Fusion und die Trainings- und Testdurchläufe mit Supervised Contrastive Learning beschrieben. Abschließend werden die Evaluationsmethoden und Metriken vorgestellt, die zur Analyse der Ergebnisse verwendet werden.

3.1 Forschungsfragen und Hypothesen

Im vorherigen Kapitel wurden Forschungslücken identifiziert, die sich aus der Verwendung von DA-Fusion im Supervised Contrastive Learning und der Verwendung von Near OOD-Augmentationen für das Negative Sampling ergeben. Um diese Lücken zu schließen, werden die folgenden Forschungsfragen und Hypothesen formuliert:

Forschungsfrage 1: Kann DA-Fusion für den EIBA-Datensatz synthetische Augmentationen erzeugen, die die Generalisierungsfähigkeit im Supervised Contrastive Learning verbessern?

Durch Beantwortung dieser Frage soll festgestellt werden, ob sich DA-Fusion grundsätzlich eignet, um die Herausforderungen der synthetischen Datengenerierung in Anwendungsfällen wie dem EIBA-Datensatz zu bewältigen (genaueres zum Datensatz in Abschnitt 3.2.1). Dazu wird DA-Fusion auf „normale“ Weise verwendet, d.h. es werden synthetische In-Distribution Daten generiert, die die Repräsentationen der realen Daten verbessern sollen. Es wird untersucht, ob die Verwendung der Augmentationen im Supervised Contrastive Learning dazu beiträgt, die Leistung des Modells für zuvor ungesehene Daten zu verbessern.

Forschungsfrage 2: Trägt die Verwendung von Out-of-Distribution (OOD) Augmentationen im Supervised Contrastive Learning dazu bei, die Robustheit des Modells gegenüber OOD-Daten zu erhöhen und die Repräsentationen von In-Distribution-Daten zu verbessern?

Im Rahmen dieser Frage wird untersucht, ob Near OOD-Augmentationen –also synthetische Daten, welche aus den echten Objekten abgeleitet sind, diese aber nicht akkurat darstellen müssen –im Supervised Contrastive Learning einen Mehrwert bieten, indem sie

die Repräsentationen der In-Distribution-Daten verbessern und die Robustheit gegenüber OOD-Daten erhöhen. Dazu wird eine neue Negative Sampling-Strategie für das Supervised Contrastive Learning verwendet, die es ermöglicht, für jeden Anchor genau die Near OOD-Augmentationen als negativ-Beispiele heranzuziehen, die aus einem Beispiel der Anchor-Klasse generiert wurden. Es wird untersucht, ob so die Generalisierungsfähigkeit und die Robustheit gegenüber OOD-Daten noch weiter gesteigert werden kann.

3.2 Datensatz

Es soll zunächst genauer auf den verwendeten Datensatz eingegangen werden, um den untersuchten Anwendungsfall im Detail zu verstehen.

3.2.1 EIBA

Grundlage der Forschungsarbeit ist ein am Fraunhofer-IPK entstandener Datensatz von Gebrauchsgegenständen, darunter hauptsächlich Autoteile und Komponenten. Er wurde im Rahmen des Projekts “Sensorische Erfassung, automatisierte Identifikation und Bewertung von Altteilen anhand von Produktdaten sowie Informationen über bisherige Lieferungen” (EIBA) erstellt, das von 2019 bis 2023 lief und von der Circular Economy Solutions GmbH koordiniert und in Kooperation mit der Technischen Universität Berlin und der deutschen Akademie der Technikwissenschaften durchgeführt wurde. (<empty citation>)

Der Datensatz ist multimodal, d.h. er besteht aus verschiedenen Datenquellen, die unterschiedliche Informationen über die Gegenstände enthalten. Neben herkömmlichen RGB-Bildern aus verschiedenen Perspektiven und weiteren Bilddaten, wie z.B. Objektmasken, gibt es auch Metadaten, etwa das Gewicht, oder Beschreibungen der Objekte in natürlicher Sprache durch verschiedene Stichwörter („CarComponent“, „cylinder“, „rusty“, usw.).

3.2.2 Teildatensatz

Um die Rechenzeit zu reduzieren und die Experimente auf eine bestimmte Objektkategorie zu beschränken, wurde ein Teildatensatz des EIBA-Datensatzes verwendet. Dabei wurden zufällig 20 Klassen aus der super class „CarComponent“ ausgewählt. Es wurden außerdem nur die RGB-Bilder verwendet, allerdings kommen auch die Objektmasken im Pre-Processing der Daten zum Einsatz. ...

3.2.3 Vorverarbeitung

...

3.3 Implementierung

Für die Vorbereitung der in dieser Arbeit durchgeführten Experimente konnte sich größtenteils auf die Implementierung von DA-Fusion und Supervised Contrastive Learning aus den Quellen ... und ... gestützt werden. Beide Implementierungen sind in Python geschrieben und verwenden die Bibliothek PyTorch. ...

Dennoch mussten einige Anpassungen vorgenommen werden, um die Modelle auf den EIBA-Teildatensatz anzuwenden, und um die synthetischen Augmentationen aus DA-Fusion im Supervised Contrastive Learning zu verwenden. ...

3.3.1 DA-Fusion

In ...'s Implementierung von DA-Fusion wird zunächst mit Textual Inversion ein vortrainiertes Stable Diffusion-Modell fine-tuned, indem ein neuer Token für jede Klasse im Datensatz erlernt wird. Um anschließend die Augmentationen zu generieren, werden die Bilder des Datensatzes genommen, Rauschen hinzugefügt und unter Konditionierung auf den entsprechenden Token wiederhergestellt. Je nachdem, wie viel Rauschen hinzugefügt wurde, entstehen so mehr oder weniger stark veränderte Bilder, die als synthetische Daten verwendet werden können.

...

Die Implementierung von DA-Fusion kann weitgehend unverändert angewendet werden, um synthetische Daten für den EIBA-Teildatensatz zu generieren. Es muss lediglich eine eigene Klasse für den Datensatz erstellt werden, die die Bilder und Masken aus dem EIBA-Datensatz lädt und die Token für die Klassen bereitstellt. ...

3.3.2 Supervised Contrastive Learning

...'s Implementierung von Supervised Contrastive Learning beinhaltet drei Trainings-Skripte; eines für das Pre-Training der latenten Repräsentationen unter Verwendung der Supervised Contrastive Loss-Funktion, eines für die lineare Klassifikation der Repräsentationen und eines zum Training eines klassischen Klassifikator-Modells mit Cross Entropy Loss (zum Vergleich).

...

Auch hier muss eine eigene Klasse für den EIBA-Teildatensatz erstellt werden, die die Bilder und Masken lädt und die synthetischen Daten von DA-Fusion bereitstellt. Die Klasse muss nun auch Parameter bereitstellen, die die Verwendung der synthetischen Daten steuern, z.B. ob keine Augmentationen, ausschließlich positiv-Beispiele oder auch negativ-Beispiele verwendet werden sollen. ...

...

3.4 Synthetische Datengenerierung mit DA-Fusion

...

3.5 Trainings- und Testdurchläufe mit Supervised Contrastive Learning

...

3.6 Evaluationsmethoden und Metriken

...

4 Ergebnisse

Dieses Kapitel präsentiert die Ergebnisse der Arbeit. Es wird auf die generierten synthetischen Daten eingegangen und die Trainings- und Testergebnisse der Modelle beschrieben. Anschließend wird die Klassifikations-Performance der Modelle verglichen und die Out-of-Distribution-Detektion analysiert.

4.1 Die generierten synthetischen Daten

...

4.1.1 In-Distribution

...

4.1.2 Near Out-of-Distribution

...



Abbildung 4.1: Beispieltext

4.2 Trainings- und Testergebnisse mit Supervised Contrastive Leraning

...

4.2.1 Contrastive Pre-Training

...

4.2.2 Lineare Klassifikation

...

4.3 Vergleich der Ergebnisse mit und ohne In-Distribution-Augmentationen

...

4.4 Vergleich der Ergebnisse mit und ohne Near Out-of-Distribution-Augmentationen als Hard Negatives

...

5 Diskussion

...

5.1 Eignung von DA-Fusion für die synthetische Datengenerierung

...

5.2 Wirksamkeit von synthetischen Near Out-of-Distribution-Daten im Supervised Contrastive Learning

...

6 Fazit

...

6.1 Zusammenfassung der wichtigsten Erkenntnisse

...

6.2 Beantwortung der Forschungsfragen

...

6.3 Ausblick und potenzielle Weiterentwicklungen

...

Literatur

- Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations.
- Foster, D. (2020). *Generatives Deep Learning*. O'Reilly.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning* [<http://www.deeplearningbook.org>]. MIT Press.
- Liu, R. (2021). Understand and Improve Contrastive Learning Methods for Visual Representation: A Review. <https://arxiv.org/abs/2106.03259>
- Mitchell, T. M. (1997). *Machine Learning*. McGraw Hill.
- O'Shea, K., & Nash, R. (2015). An Introduction to Convolutional Neural Networks. <https://arxiv.org/abs/1511.08458>
- Trabucco, B., Doherty, K., Gurinas, M., & Salakhutdinov, R. (2023). Effective Data Augmentation With Diffusion Models.
- Zhou, Z.-H. (2021). *Machine Learning*. Springer.

Anhang

Hier beginnt der Anhang. Siehe die Anmerkungen zur Sinnhaftigkeit eines Anhangs in Abschnitt ?? auf Seite ??.

Der Anhang kann wie das eigentliche Dokument in Kapitel und Abschnitte unterteilt werden. Der Befehl \appendix sorgt im Wesentlichen nur für eine andere Nummerierung.

Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit mit dem Titel

Viele zufällige Zahlen

selbstständig und nur mit den angegebenen Hilfsmitteln verfasst habe. Alle Passagen, die ich wörtlich aus der Literatur oder aus anderen Quellen wie z. B. Internetseiten übernommen habe, habe ich deutlich als Zitat mit Angabe der Quelle kenntlich gemacht.

Hamburg, 21. Dezember 1940