|  |  |  |
| --- | --- | --- |
| **MA 3062** | | |
| Untersuchung und prototypische Umsetzung eines Lifelong Deep Neural Network Algorithmus | | |
| **Simon Kamm** | | |
|  | | |
|  | | |
|  | Prüfer: | Prof. Dr.-Ing. Michael Weyrich |
|  | Betreuer: | Benjamin Maschler, M.Sc. |
| Start: 29.04.2019 | | Abgabe: 29.10.2019 |
|  | |  |

**Dokument Versionsverwaltung**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Version | Autor | QS | Datum | Status | Änderungen |
| 0.1 | Kamm | Ms | 22.08.19 | in Bearb. | Erstellung |
| 1.0 | Kamm | Ms | 28.08.19 | Vorgelegt | Erste Version Evaluierungsprotokoll (ohne ImageNet) |

# Inhaltsverzeichnis

0 Inhaltsverzeichnis 2

1 Generelle Anmerkungen Evaluierungsprotokoll 3

2 Hyperparameter-Optimierung Modul B 5

2.1 Bezug zur Evaluierungsspezifikation 5

2.2 Evaluierungsprozedur 5

2.3 Ergebnis 6

2.4 Auswertung 8

3 Anzahl an Trainingsbilder pro Klasse 11

3.1 Bezug zur Evaluierungsspezifikation 11

3.2 Evaluierungsprozedur 11

3.3 Ergebnis 11

3.4 Auswertung 13

4 Finale Untersuchung Continual Learning 15

4.1 Bezug zur Evaluierungsspezifikation 15

4.2 Evaluierungsprozedur 15

4.3 Ergebnis 15

4.4 Auswertung 16

5 Finale Untersuchung Distributed Learning 18

5.1 Bezug zur Evaluierungsspezifikation 18

5.2 Evaluierungsprozedur 18

5.3 Ergebnis 18

5.4 Auswertung 19

6 Literaturverzeichnis 20

# Generelle Anmerkungen Evaluierungsprotokoll

In diesem Kapitel werden einige grundlegende Anmerkungen geführt. Diese werden zentral zu Beginn aufgeführt, um eine bessere Lesbarkeit im weiteren Verlauf zu gewährleisten und um mehrfach auftretende Änderungen nur einmal zu erwähnen.

Für ImageNet-10 wurden Bilder der Dimension 64x64x3 genutzt, da dieser Datensatz bereits lokal in dieser Konfiguration vorhanden war. Eine Nutzung von 224x224 Bildern hätte einen erheblichen Zusatzaufwand bedeutet. Der Test auf dem gesamten ImageNet-Datensatz wird wie in der Evaluierungsspezifikation beschrieben auf Bildern der Dimensionen 224x224x3 durchgeführt haben.

Aufgrund des vortrainierten MobileNet-v2 konnten nur spezielle Eingangsdimensionen für die Bilder genutzt werden. Die kleinste verfügbare Dimension ist dabei 96x96. Deshalb wurden kleinere Bilder (ImageNet-10 und MNIST) auf diese Dimensionen mithilfe der TensorFlow-Funktion *tf.image.resize* vergrößert. Zusätzlich müssen die Bilder im RGB-Format vorliegen. Für ImageNet-Bilder ist dies der Fall. MNIST-Bilder müssen jedoch mithilfe der TensorFlow-Funktion *tf.image.grayscale\_to\_rgb* umgewandelt werden. In Abbildung 1 ist ein Bild des MNIST-Datensatzes vor der Bild-Augmentation im Format 28x28x1 zu sehen. In diesem Format liegen die Bilder des Datensatz vor. Abbildung 1 zeigt die Zahl „7“.

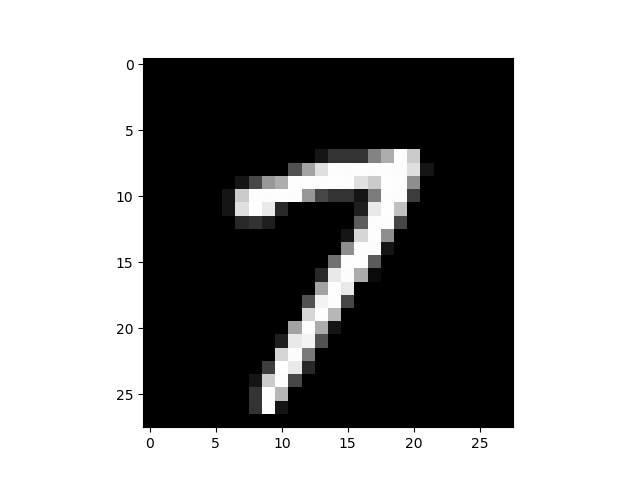


Abbildung : MNIST-Bild vor Bild-Augmentation

Abbildung 2 zeigt ein Bild des MNIST-Datensatzes nach der zuvor beschriebenen Bild-Augmentation. Es besitzt nun das Format 96x96x3. Jedoch ist zu sehen, dass das Bild in seiner Grundstruktur nicht wesentlich verändert wurde, wodurch durch diese Bild-Augmentation kein großer Einfluss auf Klassifikationsgenauigkeit erwartet wird. In diesem Beispielbild ist die Zahl „8“ zu sehen.

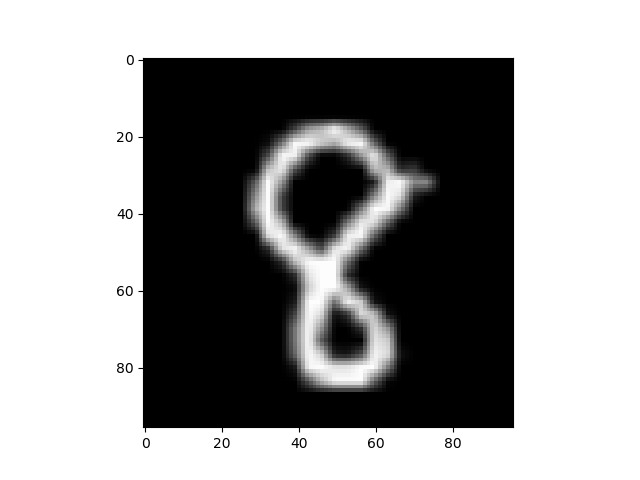


Abbildung : MNIST-Bild nach Augmentation

Zusätzlich wird die Bezeichnung des gesamten ImageNet-Datensatzes genauer definiert. Es wird der Datensatz der *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC) aus dem Jahr 2012 genutzt [1]. Als Test-Daten werden wie in der Literatur und für vergleichbare Modelle üblich die frei verfügbaren 50 Validationsbilder pro Klasse genutzt. In dem genutzten Datensatz gibt es für die 1000 Klassen insgesamt 1.281.167 Trainingsbilder. Die Anzahl an Trainingsbilder pro Klasse schwankt dabei zwischen 732 und 1300. Mit 50 Validationsbildern (hier später als Test-Bilder genutzt und bezeichnet) pro Klasse sind 50.000 Validationsbilder vorhanden. Weitere Informationen über die unterschiedlichen Klassen können in [1] nachgelesen werden.

# Hyperparameter-Optimierung Modul B

Neuronale Netzwerke besitzen eine Vielzahl an Hyperparametern, welche abhängig vom konkreten Anwendungsfall und den vorliegenden Daten unterschiedlich eingestellt werden können. Um eine möglichst sinnvolle Parametrierung des inkrementellen Klassifikators in Modul B zu gewährleisten, wurden zunächst die relevanten Parameter identifiziert und mithilfe einer Gitter-Suche die optimalen Werte für die folgenden Evaluierungsfälle zu finden. Als relevante Parameter der FuzzyARTMAP wurden die Lernrate und der Vigilance-Parameter identifiziert, da diese hauptsächlich das Training der FuzzyARTMAP und damit am Ende auch die Performanz im Testfall beeinflussen. Details zu den beiden Modell-Parametern sind in der Konzeption zu finden.

## Bezug zur Evaluierungsspezifikation

Die in der Evaluierungsspezifikation beschriebenen Fälle des kontinuierlichen Lernens auf einem Endgerät für Split-MNIST und ImageNet-10 werden genutzt, um die oben beschriebenen Hyperparameter zu optimieren und deren Einfluss zu untersuchen. Diese beiden Datensätze erlauben es aufgrund ihres begrenzten Umfangs (lediglich 10 Klassen) viele Tests in kurzer Zeit durchzuführen, was für die Gitter-Suche besonders relevant ist, da dort viele Tests für eine Abdeckung des Gitters erforderlich sind.

## Evaluierungsprozedur

Für die Hyperparameter-Optimierung wird eine zwei-dimensionale Gitter-Suche mit den Parametern und durchgeführt. Dafür werden diese beiden Parameter jeweils im Bereich in -er Schritten erhöht. Es werden 5 Wiederholungen pro mögliche Kombination durchgeführt, um eine statistische Aussagekraft zu erhalten. Das Vorgehen resultiert in jeweils 500 Tests für Split-MNIST und ImageNet-10. Die weiteren Parameter werden auf fixe Werte eingestellt, welche im Folgenden aufgelistet sind:

* modul\_b\_epsilon = 0,001
* modul\_b\_s = 1,05
* train\_img\_per\_class = 20
* test\_img\_per\_class = 100

Auf der Basis der erzielten Ergebnisse werden und ausgewählt.

Zu erwähnen ist, dass die Gruppen für Split-MNIST in einer festen Reihenfolge gezogen wurden (wie in der Evaluierungsspezifikation beschrieben). Bei ImageNet-10 wird die Reihenfolge der Klassen zufällig bestimmt, um eine zusätzliche Varianz der Trainingsdaten mit einzubeziehen.

Für beide Datensätze werden für die Bestimmung der besten Parameter in jeder Wiederholung zufällig die benötigte Anzahl an Trainingsbilder gezogen. Die ermittelte Genauigkeit wird auf Validationsdaten aus dem Trainings-Datensatz bestimmt. Diese Bilder werden nicht für das Training verwendet. Durch die Wiederholung und das zufällige Ziehen der Trainings- und Validations-Samples findet eine Kreuz-Validation der Ergebnisse statt. Die Test-Datensätze werden dann im späteren Verlauf der Arbeit zur finalen Performanz-Bewertung eingesetzt.

Zur Umsetzung des Tests müssen lediglich die Parameter gesetzt werden und die korrekten Parameter in den jeweiligen *for-*Schleifen eingesetzt werden. Dann kann der beschriebene SW-Prototyp ausgeführt werden und die *Main*-Funktion iteriert automatisch über alle Testfälle und sichert die Ergebnisse in den passenden Ordnern ab.

Nachdem alle Tests durchlaufen sind, können mithilfe des zusätzlichen Auswerte-Skripts (*Evaluation\_Results.py*) die gewünschten Metriken und Graphen für die verschiedenen Fälle erstellt werden.

## Ergebnis

Die Ergebnisse werden simultan für Split-MNIST und ImageNet-10 bewertet.

Es wird erwartet, dass für sehr kleine und große Werte von (Werte nahe 0 und 1) schlechte Ergebnisse erzielt werden, da bei diesen Fällen kaum ein Lernen stattfindet. Bei werden neue Trainingsdaten nicht für die Generalisierung genutzt und bei wird lediglich das neue Trainingssample genutzt und die alte Repräsentation damit überschrieben. Hier muss ein Wert gefunden werden, der es erlaubt altes, vorhandenes Wissen mit neuen Samples zu verbinden um eine möglichst ideale und generalisierte Repräsentation dieser Klasse zu bilden.

Bei wird erwartet, dass ab einem gewissen Wert kaum eine Änderung mehr zu erkennen ist, da die Ähnlichkeit zwischen zwei Samples einer Klasse selten nahe 1 liegt. Dieser Schwellwert für , ab welchem es kaum Unterschiede in der Performanz gibt, wird für die unterschiedlichen Datensätze unterschiedlich erwartet. Ab diesem Schwellwert kann gesagt werden, dass jedes Sample, das im Training gesehen wird, als neue Repräsentation angelegt wird. Dies erfordert einen hohen Speicherbedarf und zudem sind die erhaltenen Repräsentationen nicht generalisiert, wodurch auf späteren, abweichenden Test-Daten keine guten Generalisierungsfähigkeiten erwartet werden. Deshalb muss ein Wert für unterhalb dieser Schwelle gefunden werden. Ein zu kleiner Wert für führt dazu, dass die einzelnen Klassen durch sehr wenige (im Extremfall durch eine) Repräsentationen dargestellt werden. Je nach Klasse kann dass positiv sein (wenn alle Samples der Klasse sehr ähnlich aussehen), aber im generellen ist das nicht wünschenswert, da dadurch abweichende Samples der Klasse (zum Beispiel eine verdrehte Zahl) nicht gut erkannt werden können. Somit muss für ein guter Mittelwert zwischen zu vielen und zu wenigen Repräsentationen gewählt werden.

In Abbildung 3 sind die Ergebnisse der beschriebenen Gitter-Sucher zur Hyperparameter-Optimierung von Modul B auf Basis des Split-MNIST Datensatzes dargestellt. Entlang der x-Achse sind die unterschiedlichen Werte von zu sehen und entlang der y-Achse sind die Werte für aufgetragen. In den jeweiligen Feldern ist der Mittelwert der Klassifikationsgenauigkeit aus den fünf Wiederholungen für diese Parameterkombination eingetragen.

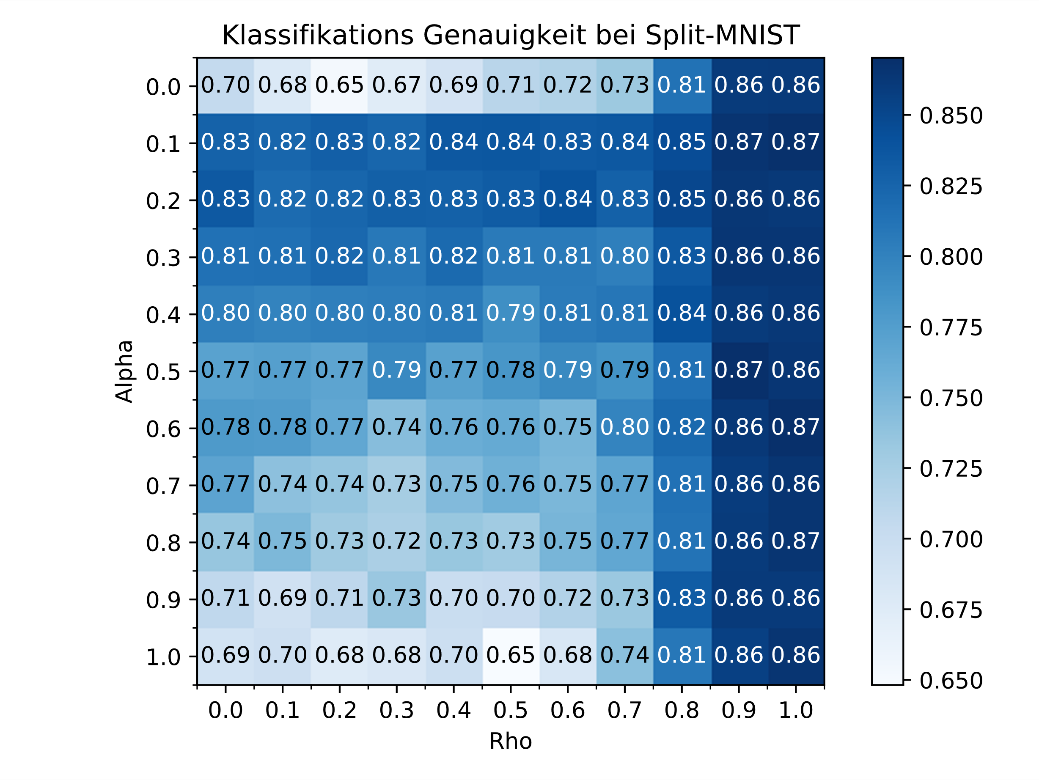


Abbildung : Ergebnisse der Gitter-Suche für und auf Basis des Split-MNIST Datensatzes

In Abbildung 4 sind die Ergebnisse für dieselben Testfälle auf Basis des ImageNet-10 Datensatzes dargestellt. Die Darstellungsart ist dabei identisch wie bereits für Abbildung 3 beschrieben.

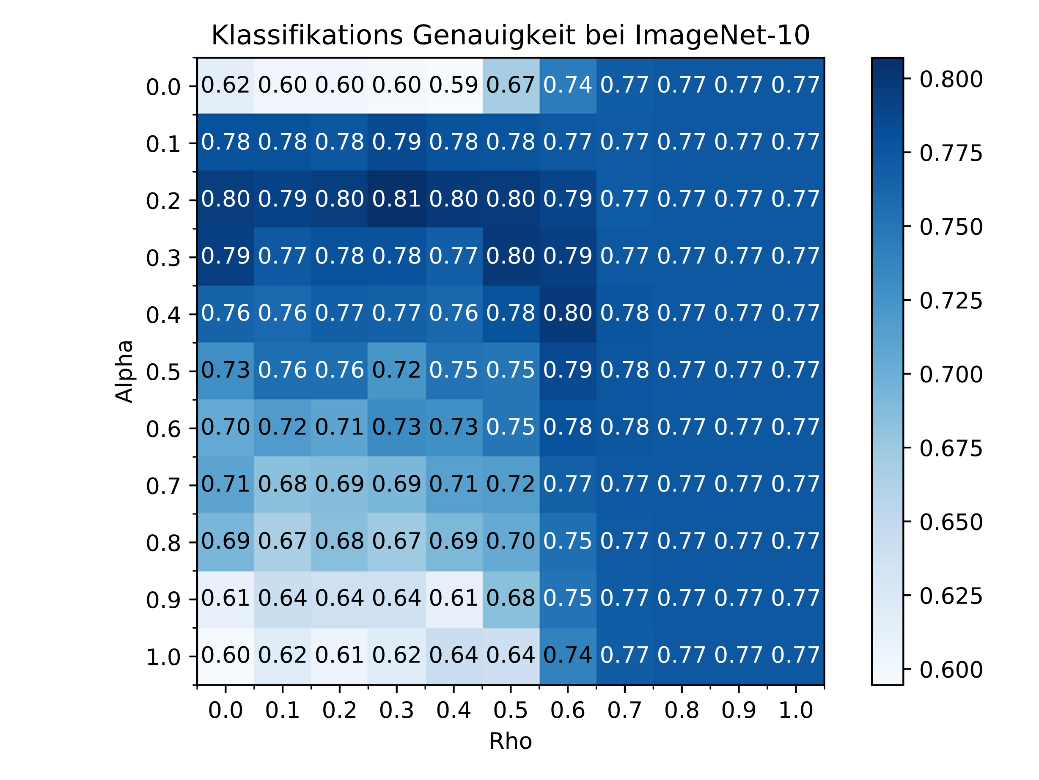


Abbildung : Ergebnisse der Gitter-Sucher für und auf Basis des ImageNet-10 Datensatzes

In Kapitel 2.4 werden die erzielten Ergebnisse nachfolgend ausgewertet und die passenden Hyperparameter für Modul B im Rahmen dieser Arbeit ausgewählt.

## Auswertung

Die Ergebnisse können für beide Parameter getrennt bewertet werden, da es keine sichtbare Korrelation der beiden Parameter gibt.

Zunächst werden die Ergebnisse für den Parameter bewertet. Hier ist das erwartete Verhalten zu erkennen. Bei beiden Datensätzen sind für hohe Werte von (>0,8) schlechte Ergebnisse zu sehen, da in diesem Fall lediglich die neuen Daten verwendet werden, um die Repräsentation zu bilden. Bei sind ebenfalls schlechte Ergebnisse zu beobachten, weil nur das erste gesehene Sample genutzt wird, um die Repräsentation zu bilden. Anhand der in Abbildung 3 und Abbildung 4 dargestellten Ergebnisse lassen sich für und die besten Ergebnisse erzielen. Für Split-MNIST variieren die besten Ergebnisse zwischen diesen beiden Werten von , abhängig vom Wert für . Bei ImageNet-10 können jedoch die besten Ergebnisse mit erzielt werden, weshalb dieser Wert im weiteren Verlauf der Arbeit genutzt wird.

Der Parameter ist maßgeblich dafür relevant, wie viele Repräsentationen angelegt werden. Damit beeinflusst direkt, wie groß der Speicherbedarf des FuzzyARTMAP-Netzwerks ist. Deshalb muss bei diesem Parameter auf einen Trade-Off zwischen Performanz und Speicherbedarf geachtet werden. Das erwartete Verhalten, bei dem ab einem gewissen Schwellwert kaum Veränderung in der Klassifikationsgenauigkeit beobachtet werden kann, kann durch die Versuche bestätigt werden. Bei Split-MNIST ist dieser Schwellwert bei ca. zu sehen, während bei ImageNet-10 dieser bereits bei zu beobachten ist. Dieser Unterschied kann mit der Komplexität der Bilder erklärt werden. Aufgrund der komplexen Bilder von ImageNet-10 sind die Ähnlichkeiten der einzelnen Bilder zueinander geringer, wodurch bereits bei niedrigeren Werten von nahezu jedes Trainingssample als Repräsentation angelegt wird. Bei Split-MNIST sind die besten Ergebnisse mit den Werten und zu beobachten, also in den Fällen wo jedes Trainingssample als Repräsentation abgelegt wird. Bei ImageNet-10 sind die besten Ergebnisse für zu sehen.

Für Split-MNIST zeigt Abbildung 3 das erwartete Ergebnis für . Es wird die beste Klassifikationsgenauigkeit (87%) erreicht, da sehr viele Repräsentationen pro Klasse angelegt werden. Für ImageNet-10 ist dieses Verhalten nicht zu sehen. Dort werden die besten Ergebnisse (80-81%) für erreicht. Bei ist die Performanz wie erwartet stabil auf einem Niveau, allerdings nur bei ca. 77% Klassifikationsgenauigkeit. Ein Grund dafür könnte in den Bildern des ImageNet-Datensatzes liegen. Die Objekte sind in realer Umgebung zu sehen, und der Hintergrund/die Umgebung ist nicht schwarz wie bei MNIST. Dadurch kann es vorkommen, dass der Feature-Extrahierer (Modul A) unrelevante Features aus dem Hintergrund extrahiert (zum Beispiel starkes Vorkommen der Farbe blau im Hintergrund bei Flugzeugen aufgrund des Himmels). Diese extrahierten Features geben jedoch keine Information über das Objekt. So können Features von Flugzeugen gleich denen von Vögeln sein, wenn beide im Himmel mit blauem Hintergrund abgebildet sind. Deshalb können sich einzelne, nicht generalisierte Repräsentationen, wie es für der Fall ist, den Validationsbildern von anderen Klassen ähneln, da einzelne Bilder der unterschiedlichen Klassen einen ähnlichen Feature-Vektor besitzen können aufgrund nicht relevanter Features.

Die genannten Schwellen spiegeln sich direkt im Speicherbedarf von Modul B wider. Ab dem genannten Schwellwert steigt der Speicherbedarf stark an, da viele Repräsentationen angelegt werden. Der mittlere Speicherbedarf von Modul B für Split-MNIST und ImageNet-10 ist in Abbildung 5 und Abbildung 6 graphisch dargestellt.

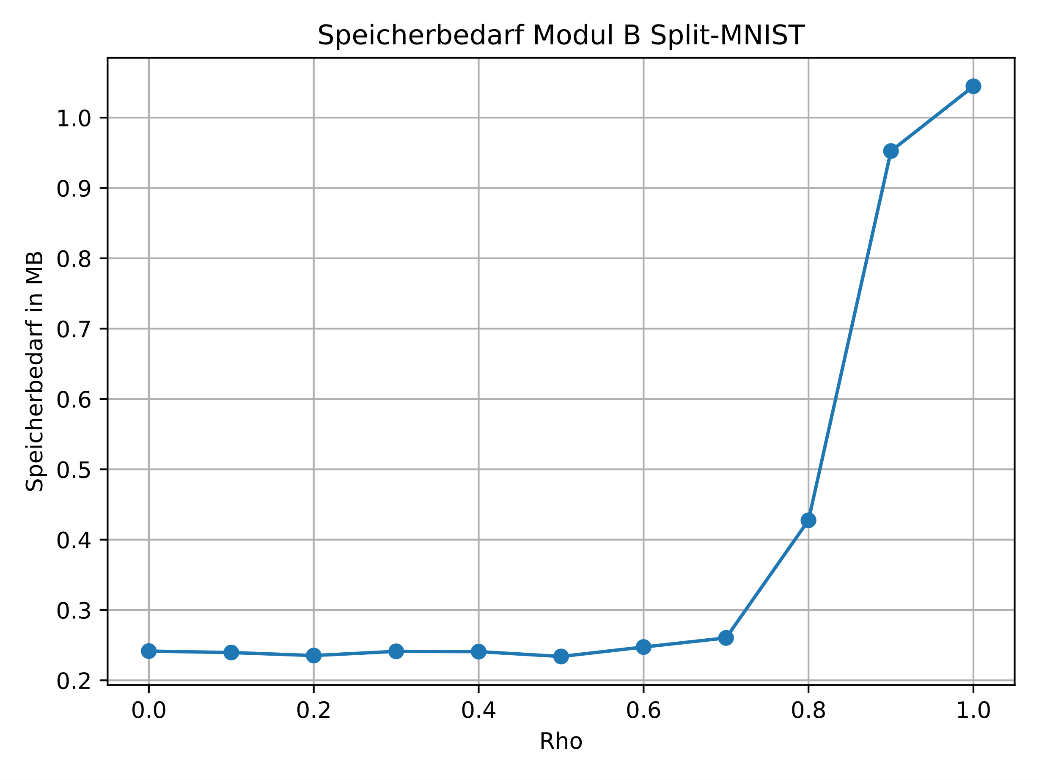


Abbildung : Speicherbedarf von Modul B in Abhängigkeit von für Split-MNIST

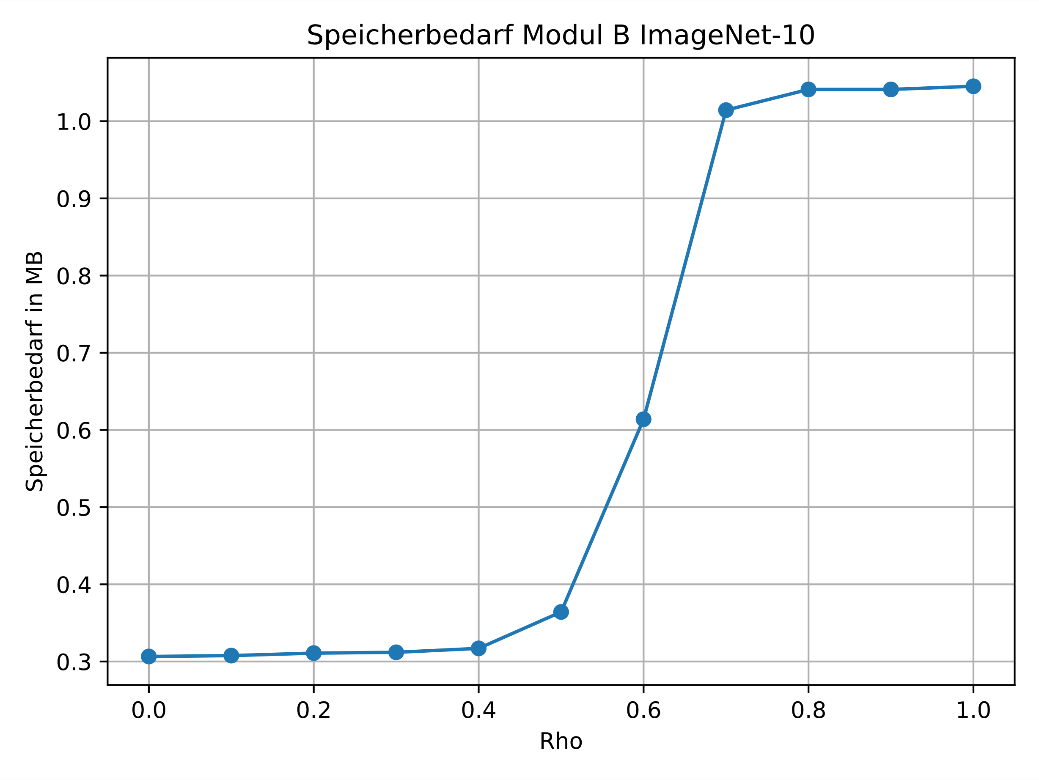


Abbildung : Speicherbedarf von Modul B in Abhängigkeit von für ImageNet-10

In diesen Abbildungen ist zu sehen, dass der Speicherbedarf ab dem Schwellwert (~0,9 für Split-MNIST und ~0,7 für ImageNet-10) deutlich zunimmt auf das 3- bist 4-fache des davor benötigten Speicherbedarfs.

Aufgrund der geringen Anzahl an Trainingsbildern (20 pro Klasse) ist der Speicherbedarf mit ca. 1MB immer noch gering. Bei späteren Anwendungen mit mehr Klassen oder Trainingsbildern kann der Speicherbedarf schnell sehr groß werden. Aufgrund dieser Ergebnisse wird für der Wert 0,5 gewählt, da damit auf beiden Datensätzen eine gute Klassifikationsgenauigkeit erzielt werden kann und dennoch der Speicherbedarf mit ca. 250 KB für Split-MNIST und ca. 350 KB für ImageNet-10 gering ist und er auch für weitere Anwendungen (z.B. gesamter ImageNet-Datensatz) in einem akzeptablen Bereich erwartet wird.

# Anzahl an Trainingsbilder pro Klasse

Schnell lernende inkrementelle Klassifikatoren sollen auf Basis von wenigen Beispieldaten einer Klasse diese erlernen können. Ein wichtiges Indiz ist dabei die Klassifikationsgenauigkeit in Abhängigkeit der gesehenen Anzahl an Samples pro Klasse. Damit kann untersucht werden, ab wie vielen Trainingssamples ein Maximum erreicht werden kann und wie sich der Performanz-Gewinn durch eine höhere Anzahl an Trainingsbilder verhält.

## Bezug zur Evaluierungsspezifikation

Wie in den Kriterien der Evaluierungsspezifikation dargestellt, wird dieser Test genutzt, um eine Aussage über die Fähigkeit des Algorithmus zu erhalten, wie schnell neue Klassen erlernt werden können.

## Evaluierungsprozedur

Es werden die zuvor ermittelten Hyperparameter von Modul B verwendet. Mit diesen Parametern wird als weitere Untersuchung die Anzahl an Trainingsbildern variiert. Die Anzahl der Trainingsbilder wird dabei auf folgende Werte gesetzt: . Es werden pro Anzahl an Trainingsbildern 5 Wiederholungen durchgeführt, um einen aussagekräftigen statistischen Mittelwert und die Standardabweichung bilden zu können. Die weiteren einstellbaren Parameter sowie der Ablauf sind wie in Kapitel 2.2 beschrieben.

## Ergebnis

Es wird erwartet, dass mit mehr Trainingsbildern pro Klasse bessere Ergebnisse erzielt werden können, da mehr und besser generalisierte Repräsentationen für die einzelnen Klassen angelegt werden können. Mit steigender Anzahl an Trainingsbildern pro Klasse wird zudem erwartet, dass die Varianz der Klassifikationsgenauigkeit abnimmt, und damit stabilere Ergebnisse erzielt werden können. Die Annahme ist, dass bei wenigen Trainingsbildern die spätere Klassifikationsgenauigkeit stark von der Auswahl dieser wenigen Trainingsbilder abhängig ist. Bei steigender Anzahl an Trainingsbilder sinkt die Abhängigkeit von den einzelnen Trainingsbilder.

Zudem ist die Erwartung, dass der Speicherbedarf mit der Anzahl an Trainingsbildern steigt, da mehr Repräsentationen durch die höhere Zahl an gesehenen Trainingsdaten angelegt werden.

Die Ergebnisse für die Klassifikationsgenauigkeit werden in einem Balkendiagramm dargestellt. Die schwarzen Linien stellen dabei die Standardabweichung um den Mittelwert dar. Aufgrund der großen Unterschiede zwischen den einzelnen Werten für die Anzahl an Trainingsbilder wird die x-Achse in logarithmischer Skala dargestellt.

In Abbildung 7 ist die Klassifikationsgenauigkeit für Split-MNIST dargestellt.

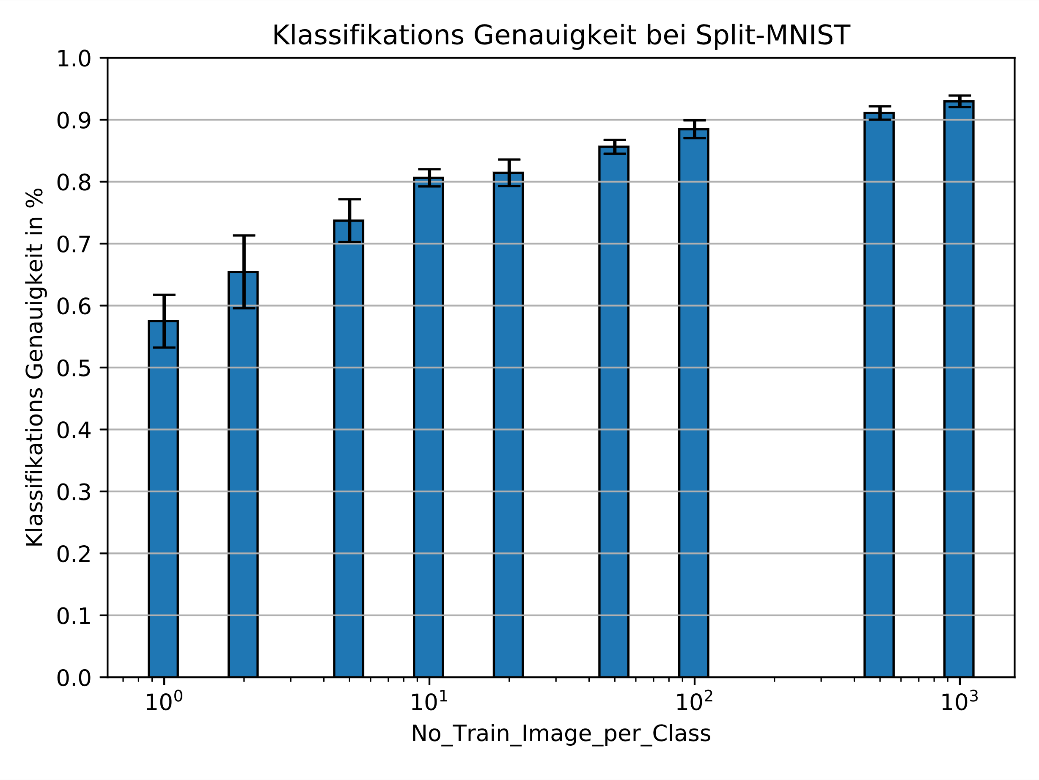


Abbildung : Klassifikationsgenauigkeit über die Anzahl an Trainingsbildern Split-MNIST

Abbildung 8 stellt die Klassifikationsgenauigkeit über die Anzahl an Trainingsbildern für ImageNet-10 dar.

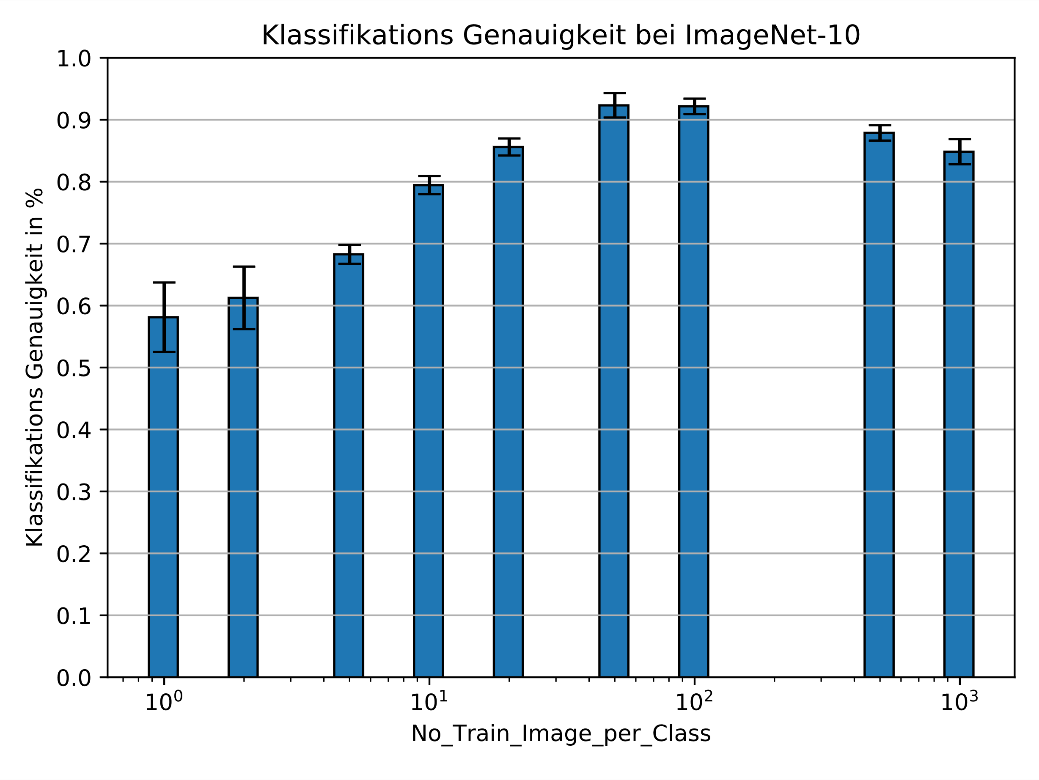


Abbildung : Klassifikationsgenauigkeit über die Anzahl an Trainingsbildern ImageNet-10

Zusätzlich wird der mittlere Speicherbedarf von Modul B für die jeweiligen Datensätze in Abhängigkeit von der Anzahl an Trainingsbildern in Abbildung 9 für Split-MNIST und in Abbildung 10 für ImageNet-10 dargestellt.

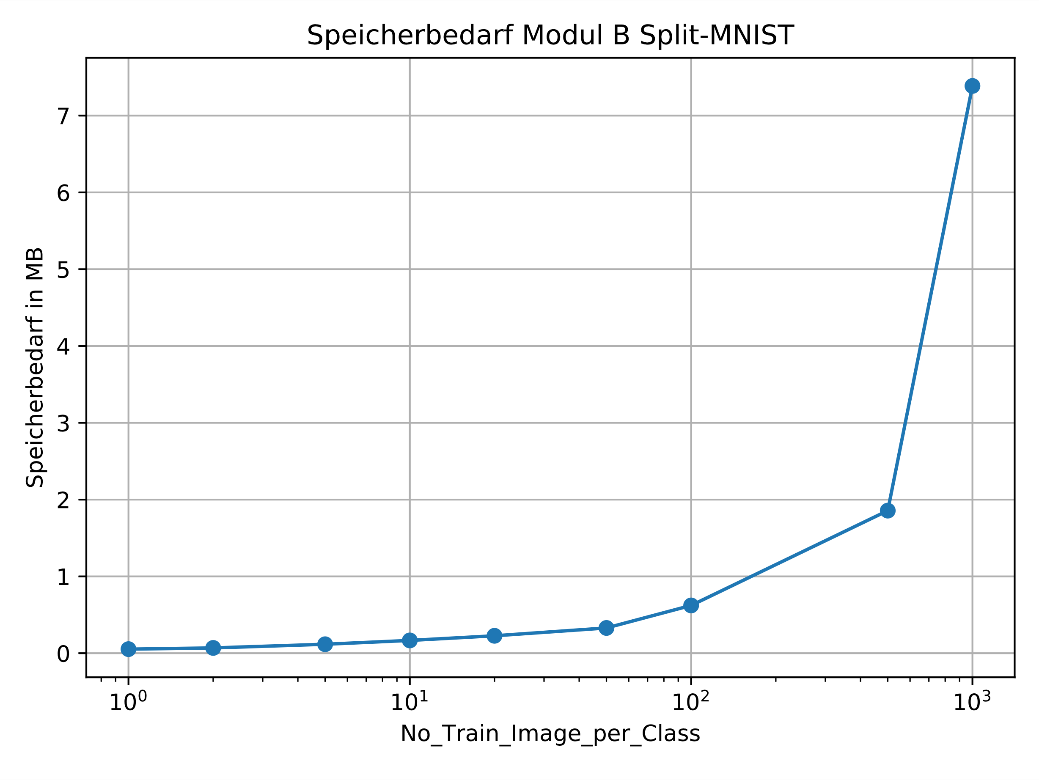


Abbildung : Speicherbedarf Modul B über die Anzahl an Trainingsbildern Split-MNIST

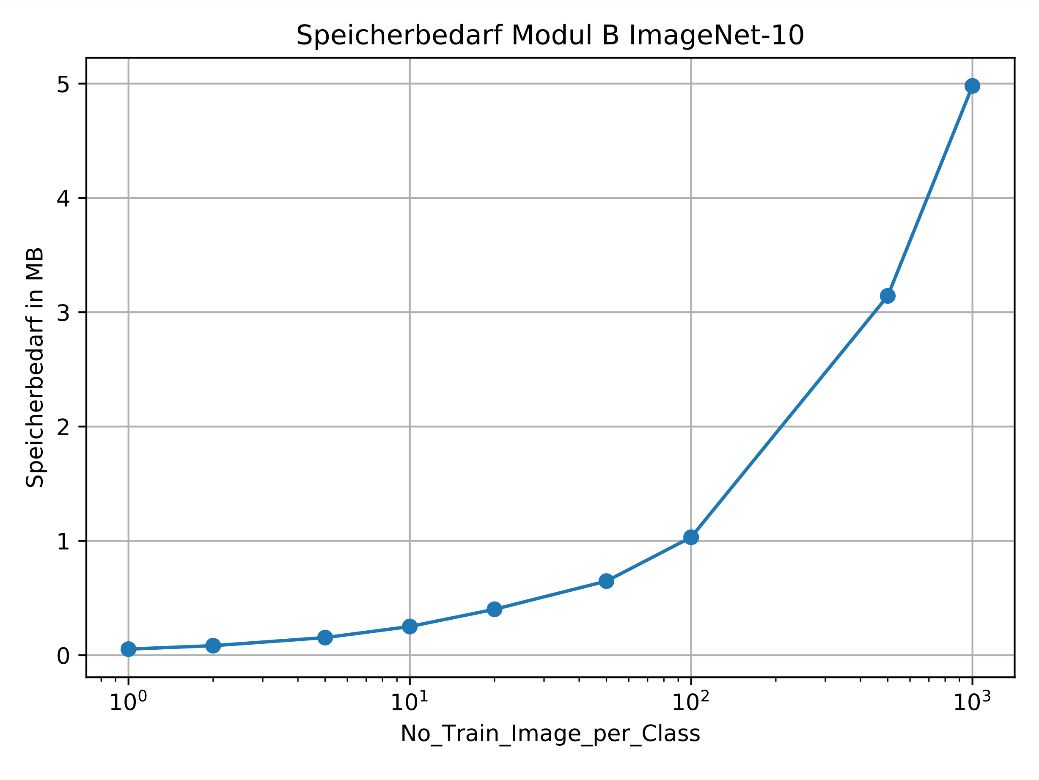


Abbildung : Speicherbedarf Modul B über die Anzahl an Trainingsbildern ImageNet-10

Die Bewertung der gezeigten Ergebnisse findet im folgenden Abschnitt 3.4 statt.

## Auswertung

Die Annahme, dass mit einer geringen Anzahl an Trainingsbildern eine hohe Varianz auftritt, kann auf Basis der erzielten Ergebnisse bestätigt werden. Für Split-MNIST (Abbildung 7) beträgt die Standardabweichung bei einem beziehungsweise zwei Trainingsbildern pro Klasse 4,3 und 5,9 Prozentpunkte. Mit 500 und 1000 Trainingsbildern pro Klasse beträgt die errechnete Standardabweichung für 5 Wiederholungen lediglich 1,1 und 0,9 Prozentpunkte.

Bei ImageNet-10 ist ebenfalls eine höhere Standardabweichung bei einer geringen Anzahl Trainingsbildern zu sehen mit 5,6 und 5,1 Prozentpunkten bei einem beziehungsweise zwei Trainingsbildern pro Klasse. Die geringste Abweichung ist für ImageNet-10 mit 100 und 500 Trainingsbildern pro Klasse zu erreichen, mit jeweils 1,2 Prozentpunkten.

Wenn nun die Klassifikationsgenauigkeit an sich betrachtet wird, kann für Split-MNIST gesagt werden, dass mit mehr Trainingsbildern pro Klasse die Genauigkeit besser wird. Bei 1000 Trainingsbildern pro Klasse wird im Mittel über fünf Wiederholungen eine Klassifikationsgenauigkeit von 92,96% bei einer Standardabweichung von +/-0,94 erreicht. Mit 500 Trainingsbildern pro Klasse kann ebenfalls bereits eine Klassifikationsgenauigkeit von 91,08% +/- 1,1 erreicht werden.

Da zusätzlich der Speicherbedarf in Abbildung 9 in Betracht gezogen wird, werden im weiteren Verlauf der Arbeit 500 Trainingsbilder pro Klasse für Split-MNIST genutzt. Der Speicherbedarf steigt zwischen 500 und 1000 Trainingsbildern pro Klasse von ca. 1,9 MB auf ca. 7,2 MB an. Da Speicherbedarf in dieser Arbeit eine Rolle spielt, kann die leicht geringere Genauigkeit (ca. 1 Prozentpunkt) akzeptiert werden, wenn dafür weniger als ein Drittel an Speicher benötigt wird.

Für ImageNet-10 lässt sich ein anderes Verhalten der Klassifikationsgenauigkeit über der Anzahl an Trainingsbildern beobachten. Die Genauigkeit steigt zunächst bis 100 Trainingsbildern pro Klasse an, jedoch fällt die Genauigkeit im Gegensatz zu Split-MNIST daraufhin wieder ab. Ein möglicher Grund könnte der bereits beschrieben Fall von zu vielen Repräsentation sein (siehe die Auswertung des Parameters in Kapitel 2.4 für Details). Die beste Klassifikationsgenauigkeit lässt sich mit 50 und 100 Trainingsbildern pro Klasse erzielen, mit 92,31% +/- 1,9 und 92,16% +/- 1,2. Aufgrund der minimal geringeren Varianz und damit dem besseren „Worst-Case“ Ergebnis (90,96% gegen 90,41%) werden 100 Trainingsbilder pro Klasse für weitere Untersuchungen mit dem ImageNet-10 Datensatz genutzt. Auch der Speicherbedarf ist in diesem Bereich noch akzeptabel, mit ca. 1 MB Speicherbedarf von Modul B für diese Anzahl an Trainingsbildern (Abbildung 10). Mit 50 Trainingsbildern pro Klasse läge der Speicherbedarf bei ca. 0,65 MB und somit nicht wesentlich geringer.

Insgesamt hat die Anzahl an Trainingsbildern pro Klasse einen, wie erwartet, direkten Einfluss auf die spätere Klassifikationsgenauigkeit. Bereits mit wenigen Trainingsbildern pro Klasse (z.B. 10) können jedoch akzeptable Klassifikationsgenauigkeiten von ca. 80% für die genutzten Validationsdaten von Split-MNIST und ImageNet-10 erreicht werden. Generell kann gesagt werden, dass mit steigender Anzahl an Trainingsbilder die Klassifikationsgenauigkeit ebenso wie der Speicherbedarf des inkrementellen Klassifikator weiter zunimmt. Je nach Anwendungsfall kann eine geringere Genauigkeit infolge von weniger Trainingsdaten akzeptiert werden, wenn z.B. die Erzeugung von Trainingsdaten sehr kosten- und zeitaufwändig ist.

Mit diesem Testfall kann gezeigt werden, dass der hier untersuchte Algorithmus für spätere Anwendungen, bei denen wenigen Daten pro Klasse verfügbar sind, geeignet ist, da er bereits mit wenigen Trainingsbildern pro Klasse gute Ergebnisse erzielen kann.

# Finale Untersuchung Continual Learning

Für eine abschließende Bewertung und Einordnung des Potenzials des L DNN Algorithmus werden auf Basis der zuvor untersuchten Hyperparameter finale Tests für das kontinuierliche Lernen durchgeführt. Dabei wird eine größere Anzahl an Wiederholungen mit festen Parametern durchgeführt. Zudem werden die finalen Ergebnisse auf Basis der Testdaten der jeweiligen Datensätze ermittelt, nachdem zuvor für die Untersuchungen der Hyperparameter eine Kreuzvalidation mithilfe von Trainings- und Validationsdaten durchgeführt wurde.

## Bezug zur Evaluierungsspezifikation

Die in Kapitel 4 der Evaluierungsspezifikation genannten Testfälle werden nun durchgeführt. Es werden die Testdaten genutzt, nachdem zuvor für die Hyperparameteroptimierung Validationsdaten eingesetzt wurden, um eine korrekte Evaluierung und Optimierung der Parameter zu gewährleisten. Für die Erstellung der folgenden Metriken werden Daten genutzt, die das Netzwerk bisher noch nicht gesehen hat. Dadurch kann eine Optimierung und *Overfitting* des Netzwerks auf Testdaten verhindert werden.

## Evaluierungsprozedur

Es werden die Testfälle des kontinuierlichen Lernens auf einem Gerät durchgeführt. Die Parameter werden auf Basis der vorherigen Testfälle ausgewählt und sind im Folgenden nochmals aufgelistet. Als Testdaten werden alle verfügbaren Testbilder der jeweiligen Datensätze genutzt.

* modul\_b\_epsilon = 0,001
* modul\_b\_s = 1,05
* modul\_b\_alpha = 0,2
* modul\_b\_rho = 0,5
* train\_img\_per\_class = 100 (ImageNet-10), 500 (Split-MNIST)
* test\_img\_per\_class = 50 (ImageNet-10), 1000 (Split-MNIST)

Es werden 10 Wiederholungen pro Datensatz durchgeführt.

Zusätzlich zu den bisher untersuchten Datensätzen wird für die finale Bewertung des Algorithmus auch auf dem gesamten ImageNet-Datensatz mit 1000 Klassen getestet. Dafür haben die Parameter auf dieselben Werte wie für ImageNet-10.

## Ergebnis

In diesem Abschnitt werden die Ergebnisse dargestellt. Dafür wird für Split-MNIST eine Tabelle mit Ergebnissen anderer *Continual Learning*-Verfahren aus der Literatur sowie aktueller Top-Ergebnisse mit traditionellen Deep Learning Ansätzen angelegt.

Für ImageNet-10 sind keine weiteren Ergebnisse von *Continual Learning*-Verfahren bekannt. Dieser Datensatz diente lediglich der Überprüfung des Potenzials auf komplexeren Eingangsdaten. Der Vergleich zu anderen Algorithmus und damit auch die Einordnung findet auf Basis des gesamten ImageNet-Datensatzes statt (1000 Klassen).

In Tabelle 1 sind die Ergebnisse für den Split-MNIST Datensatz dargestellt. Die erste Zeile ist dabei der hier untersuchte Algorithmus. Bei den weiteren Algorithmen ist angegeben, von welcher Quelle die genannte Klassifikationsgenauigkeit stammt. Die Ergebnisse gelten für den Fall des inkrementellen Klassen Lernens.

Tabelle : Klassifikationsgenauigkeit verschiedener Algorithmen auf Split-MNIST

|  |  |
| --- | --- |
| Algorithmus | Klassifikationsgenauigkeit in % |
| L DNN Algorithmus | 86,94 +/- 1,29 |
| Deep Generativ Replay (DGR) [2] | 91,24 +/- 0,33 |
| Elastic Weight Consolidation (EWC) [3] | 19,90 +/- 0,05 |
| Synaptic Intelligence (SI) [3] | 20,04 +/- 0,08 |
| Multi-Layer Perceptron (MLP) – inkrementell trainiert [3] | 19,90 +/- 0,02 |
| MLP – offline trainiert [3] | 97,93 +/- 0,04 |

EWC und SI stellen dabei typische Methoden des kontinuierlichen Lernens dar. Diese Methoden speichern keine Trainingsdaten und nutzen keine gespeicherten Repräsentationen zum Training. Dies wird in der Literatur auch Replay oder Rehearsal genannt. DGR nutzt diese Methode, in dem es komprimierte Repräsentationen der Trainingsdaten abspeichert. Wenn neue Klassen hinzukommen, werden aus den gespeicherten Komprimierungen der alten Klassen sowie mithilfe eines erlernten generativen Moduls Trainingsbilder dieser Klassen erzeugt (Generative) und in die neuen Trainingsdaten eingebracht. Das *inkrementell trainierte MLP* kann als untere Grenze gesehen werden, da hier *Catastrophic Forgetting* durch einfaches Anwenden des *Backpropagation*-Algorithmus auftritt. Das *offline trainierte MLP* wurde mit allen Klassen offline trainiert, und kann als obere Grenze angesehen werden.

Für ImageNet-10 wird eine gemittelte Klassifikationsgenauigkeit von erreicht. Vergleichbare Ergebnisse für diesen Anwendungsfall sind in der Literatur nicht zu finden.

## Auswertung

Für Split-MNIST können die Ergebnisse mit anderen *Continual Learning* Algorithmen verglichen werden, da es viele Untersuchungen auf Basis dieses Datensatzes gibt. Im Vergleich zu den klassischen Methoden wie EWC und SI kann der L DNN Algorithmus deutlich bessere Ergebnisse für das hier untersuchte inkrementelle Klassen Lernen erzielen. Algorithmen, die mit generativen Methoden arbeiten (wie DGR), erreichen eine bessere Klassifikationsgenauigkeit für diesen Anwendungsfall. Allerdings besitzen sie auch eine erhöhte Komplexität während des Trainings. Denn bei diesen Modellen muss zusätzlich zu dem inkrementellen Klassifikator ein generatives Modell trainiert werden, welches ausgewählte Trainingsdaten komprimiert und aus den komprimierten Darstellungen wiederherstellt (z.B. mithilfe eines *Variational Auto-Encoder*). Zudem müssen diese komprimierten Darstellungen abgespeichert werden. All diese Punkte ermöglichen generativen Modellen eine bessere Klassifikationsgenauigkeit, jedoch sind diese Modelle aus den genannten Gründen (noch) nicht für den Einsatz auf einem mobilen Endgerät geeignet, was in dieser Arbeit ein wichtiger Auswahlpunkt für den Algorithmus war. Insgesamt ist festzuhalten, dass eine Genauigkeit von ca. 87% auf Split-MNIST ist ein sehr gutes Resultat für das inkrementelle Klassenlernen ohne Replay/Rehearsal.

Die Ergebnisse für ImageNet-10 können nicht mit anderen Algorithmen verglichen werden, jedoch kann damit geprüft werden, ob der Algorithmus auch auf komplexeren Eingangsdaten (hier 64x64 RGB-Bilder mit komplexen Klassen) funktioniert, bevor ein großer Test auf dem gesamten ImageNet-Datensatz durchgeführt wird. Mit einer finalen mittleren Klassifikationsgenauigkeit von 76,4% kann gesagt werden, dass der Algorithmus auch komplexere Klassen und Eingangsdaten korrekt klassifizieren kann.

# Finale Untersuchung Distributed Learning

Für eine abschließende Bewertung und Einordnung des Potenzials des L DNN Algorithmus werden auf Basis der zuvor untersuchten Hyperparameter finale Tests für das verteilte Lernen durchgeführt. Dabei wird eine größere Anzahl an Wiederholungen mit festen Parametern durchgeführt. Zudem werden die finalen Ergebnisse auf Basis der Testdaten der jeweiligen Datensätze ermittelt, nachdem zuvor für die Untersuchungen der Hyperparameter eine Kreuzvalidation mithilfe von Trainings- und Validationsdaten durchgeführt wurde.

## Bezug zur Evaluierungsspezifikation

Die in Kapitel 4 der Evaluierungsspezifikation genannten Testfälle für das verteilte Lernen werden nun durchgeführt. Es werden die Testdaten genutzt, nachdem zuvor für die Hyperparameteroptimierung Validationsdaten eingesetzt wurden, um eine korrekte Evaluierung und Optimierung der Parameter zu gewährleisten. Für die Erstellung der folgenden Metriken werden Daten genutzt, die das Netzwerk bisher noch nicht gesehen hat. Dadurch kann eine Optimierung und *Overfitting* des Netzwerks auf Testdaten verhindert werden.

## Evaluierungsprozedur

Die Testfälle des verteilten Lernens werden auf zwei Geräten durchgeführt. Die Parameter werden auf Basis der vorherigen Testfälle ausgewählt und sind im Folgenden nochmals aufgelistet. Als Testdaten werden alle verfügbaren Testbilder der jeweiligen Datensätze genutzt.

* modul\_b\_epsilon = 0,001
* modul\_b\_s = 1,05
* modul\_b\_alpha = 0,2
* modul\_b\_rho = 0,5
* train\_img\_per\_class = 100 (ImageNet-10), 500 (Split-MNIST)
* test\_img\_per\_class = 50 (ImageNet-10), 1000 (Split-MNIST)

Es werden 10 Wiederholungen pro Datensatz durchgeführt.

Zusätzlich zu den bisher untersuchten Datensätzen wird für die finale Bewertung dieser Test auch auf dem gesamten ImageNet-Datensatz mit 1000 Klassen durchgeführt. Dafür haben die Parameter dieselben Werte wie für ImageNet-10.

## Ergebnis

Die Ergebnisse der einzelnen Datensätze werden hier dargestellt. Dafür wird für den Fall des verteilten Lernens auf zwei Geräten die Klassifikationsgenauigkeit der einzelnen Geräte nach dem Erlernen ihrer verfügbaren Klassen angegeben. Zusätzlich wird die finale Genauigkeit des „verschmolzenen“ Netzwerks gegeben. Als Referenz dienen die Ergebnisse des kontinuierlichen Lernens auf einem Gerät, da im besten Fall durch das verteilte Lernen keine schlechteren Ergebnisse erzielt werden sollen.

Zunächst findet die Auswertung für Split-MNIST statt. Dabei wurden auf Gerät 1 die Gruppen 0/1, 2/3 und 4/5 trainiert, während auf Gerät 2 die Gruppen 6/7 und 8/9 trainiert wurden. Die gemittelten Ergebnisse aus 10 Läufen sowie deren Standard-Abweichung sind in Tabelle 2 gegeben.

Tabelle : Klassifikationsgenauigkeit des verteiltem L DNN Algorithmus auf Split-MNIST

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithmus | Klassifikations-genauigkeit Gerät 1 in % | Klassifikations-genauigkeit Gerät 2 in % | Klassifikations-genauigkeit final in % |
| L DNN Algorithmus – 1 Gerät | - | - | 86,94 +/- 1,29 |
| L DNN Algorithmus – 2 Geräte | 88,44 +/- 0,82 | 95,02 +/- 0,44 | 86,56 +/- 0,91 |

Dieselbe Evaluation wird auf Basis des ImageNet-10 Datensatzes durchgeführt. Die Ergebnisse dieser Untersuchung sind in Tabelle 3 zu sehen. Die beiden Geräte haben dieselbe Anzahl an Klassen (fünf) gesehen. Die jeweiligen Klassen, die auf den Geräten trainiert werden, sind ebenso wie die Reihenfolge des Trainings der Klassen zufällig gezogen.

Tabelle : Klassifikationsgenauigkeit des verteiltem L DNN Algorithmus auf ImageNet-10

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithmus | Klassifikations-genauigkeit Gerät 1 in % | Klassifikations-genauigkeit Gerät 2 in % | Klassifikations-genauigkeit final in % |
| L DNN Algorithmus – 1 Gerät | - | - | 76,4 +/- 1,2 |
| L DNN Algorithmus – 2 Geräte | 84,64 +/- 4,73 | 87,44 +/- 4,04 | 76,26 +/- 1,5 |

Die Auswertung der Ergebnisse erfolgt im folgenden Kapitel 5.4.

## Auswertung

Die Ergebnisse für Split-MNIST und ImageNet-10 zeigen, dass mit separatem Training auf mehreren Geräten (hier zwei) dieselbe Genauigkeit erreicht werden wie mit dem Training auf einem einzelnen Gerät. Die Ergebnisse in Tabelle 2 und Tabelle 3 verdeutlichen diese Aussage. Zudem kann gesagt werden, dass durch das „Verschmelzen“ von Wissen eine insgesamt schlechtere Genauigkeit auftritt. Die einzelnen Genauigkeiten von Gerät 1 und Gerät 2 auf ihren jeweiligen Testdaten sind besser als die finale Klassifikationsgenauigkeit. Dies hängt mit der komplexeren finalen Aufgabe zusammen, da dort statt 6 und 4 (für Split-MNIST) oder 5 und 5 Klassen (für ImageNet-10) nun 10 Klassen im Test-Datensatz vorkommen. Insgesamt kann auf Basis dieser Ergebnisse gesagt werden, dass das verteilte Lernen mit diesem Algorithmus funktioniert, da dort kein nennenswerter Performanz Verlust gegenüber dem zentralen Erlernen der Aufgabe auf einem Gerät auftritt. Somit eignet sich dieser Algorithmus zum Einsatz auf verteilten Systemen, die ihr Wissen ohne die jeweiligen Rohdaten austauschen müssen.

# Literaturverzeichnis

|  |  |
| --- | --- |
| [1] | O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg und L. Fei-Fei, „ImageNet Large Scale Visual Recognition Challenge,“ 30 Januar 2015. |
| [2] | Y.-c. Hsu, Y.-c. Liu und Z. Kira, „Re-evaluating Continual Learning Scenarios : A Categorization and Case for Strong Baselines,“ in *32nd Conference on Neural Information Processing Systems (NIPS2018)*, Montréal, 2018. |
| [3] | G. M. van de Ven und A. S. Tolias, „Three continual learning scenarios and a case for generative replay,“ in *International Conference on Learning Representations*, New Orleans, 2019. |