Praktikumsbericht

DLO Deep Learning und Objekterkennung
Übungsblatt 4

Modul von Prof. Dr. Jan Salmen
Tutor: Luca Uckermann, Matthias Bullert, Jinxin Eisenhut

Technology
Arts Sciences
TH Köln

Namen: David Mertens (11135340), Arthur Kehrwald (11135125)
Datum: 16. Mai 2025

Aufgabenstellung

In dieser Untersuchung wurde die Sensitivität eines trainierten MNIST-Klassifikationsnetzwerks gegenüber gezielten Pixelveränderungen analysiert. Dazu wurde ein gut funktionierendes Modell trainiert und anschließend auf korrekt klassifizierte Trainingsbilder angewendet. Für jedes dieser Bilder wurde systematisch jeder einzelne Pixel invertiert, um zu überprüfen, ob dadurch eine Fehlklassifikation ausgelöst wird. Ausgewertet wurde, welche Bilder und Bildbereiche besonders instabil reagieren, sowie welche Klassen am häufigsten durch minimale Veränderungen verwechselt werden.

Herangehensweise

Jedes Pixel jedes Bildes im Testdatensatz wurde systematisch auf seine Auswirkung auf die Klassifikation untersucht. Dazu wurde der ursprüngliche Pixelwert $x \in [0, 1]$ modifiziert, indem er zunächst invertiert und anschließend auf einen der Extremwerte $\{0, 1\}$ gerundet wurde:

$$x' = \begin{cases} 0, & \text{wenn } 1 - x > 0{,}5 \\ 1, & \text{wenn } 1 - x \leq 0{,}5 \end{cases}$$

Dieses gezielte Clipping auf die Extremwerte 0 und 1 verstärkt den Einfluss des jeweiligen Pixels auf das Netz deutlich. Um die resultierende Instabilität im Hinblick auf die Architektur zu analysieren, wurde das Experiment mit zwei unterschiedlich trainierten Netzwerken durchgeführt, die jeweils eine Testgenauigkeit von 97 % auf dem MNIST-Datensatz erreichen.

Beobachtungen

Insgesamt wurden ca. 95,6% der Trainingsbilder richtig und robust klassifiziert. Das heißt unser Netz erkennt in diesen Bildern die korrekte Zahl und lässt sich durch keine Invertierung eines einzelnen Pixels von diesem Urteil abbringen. 3,1% wurden von vorneherein nicht richtig klassifiziert und daher außen vor gelassen. Weitere 1,3% wurden zwar ursprünglich richtig erkannt, aber dann in mindestens einem der veränderten Bilder einer falschen Klasse zugeordnet. Das sind die bei der Untersuchung der Sensitivität interessanten Bilder.
Die folgenden Abbildungen geben einen Überblick über die Stabilität dieser nicht robusten Bilder. Besonders helle Bereiche kennzeichnen dabei instabile Regionen – also Bildbereiche, deren Veränderung besonders leicht zu einer anderen Vorhersage des Netzes führt. Zunächst wird das Ergebnis eines einzelnen Testdurchlaufs analysiert. Anschließend werden Auffälligkeiten zwischen zwei verschiedenen Netzwerken verglichen.
Ein erster markanter Befund zeigt sich in Abbildung 4: Die einzelnen Ziffernklassen weisen deutlich unterschiedliche Verwechslungsanfälligkeiten auf. So werden die Ziffern *0* und *1* von diesem Netz nahezu nie mit anderen Ziffern verwechselt, während beispielsweise die Klassen *8* und *9* besonders häufig falsch klassifiziert werden. Diese Klassen sind demnach instabiler.

In der Abbildung 2 lassen sich darüber hinaus instabile Bildbereiche lokalisieren. Nahezu jede Ziffer zeigt mindestens eine Region mit besonders hoher Helligkeit, was auf eine

besondere Sensitivität gegenüber lokalen Änderungen hindeutet.

Abbildung 6 schließlich zeigt eine farbcodierte Übersicht darüber, mit welchen Klassen diese Verwechslungen jeweils am häufigsten auftreten. Einfach zu erklärende Bereiche sind zum Beispiel der untere Bildrand bei Bildern der Klasse 4. Bei einer Invertierung kann hier schnell eine Verwechslung mit der 9 passieren.

Vergleich zweier Netze

Beim direkten Vergleich der Heatmaps der beiden unterschiedlich trainierten Netzwerke (Abbildungen 2 und 3) zeigen sich auffällige Gemeinsamkeiten in den entstehenden Mustern. Dies deutet darauf hin, dass bestimmte Bildbereiche unabhängig vom jeweiligen Netzwerk weiterhin besonders anfällig für Fehlklassifikationen sind. Betrachtet man hingegen die Balkendiagramme zur Klasseninstabilität (Abbildungen 4 und 5), wird deutlich, dass sich die Häufigkeit der Verwechslungen zwischen den beiden Netzwerken zum Teil erheblich unterscheidet. Zwar treten ähnliche Fehler auf, diese variieren jedoch in ihrer Verteilung und Ausprägung.

Ein solches Verhalten lässt sich potenziell nutzen, indem man mehrere Netzwerke parallel trainiert und die finale Vorhersage durch ein Mehrheitsvotum trifft. Dabei könnte die Entscheidung zusätzlich durch die individuelle Klassensicherheit der einzelnen Modelle gewichtet werden, um besonders bei instabilen Ziffern die Gesamtaussage zu verbessern.

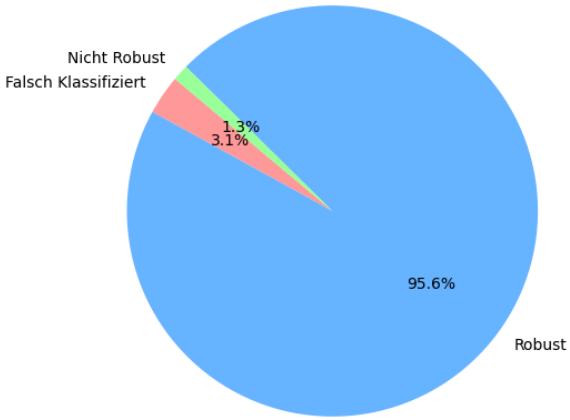

Abbildung 1: Das Verhältnis zwischen falsch klassifizierten, robust richtig, und nicht robust richtig klassifizierten Trainingsbildern.

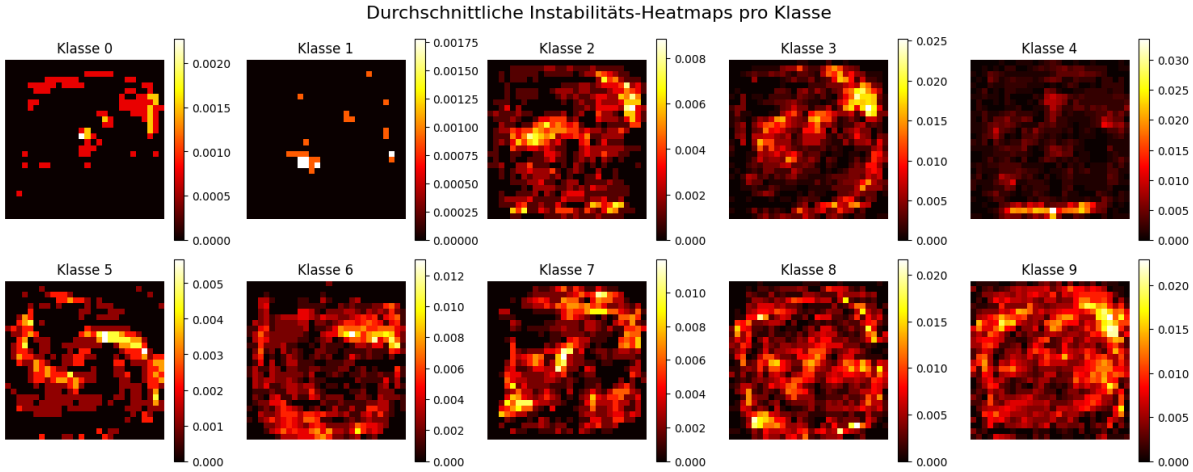

Abbildung 2: Diese Heatmap zeigt die besonders Instabilen Bereiche. (Entstanden durch Netz A)

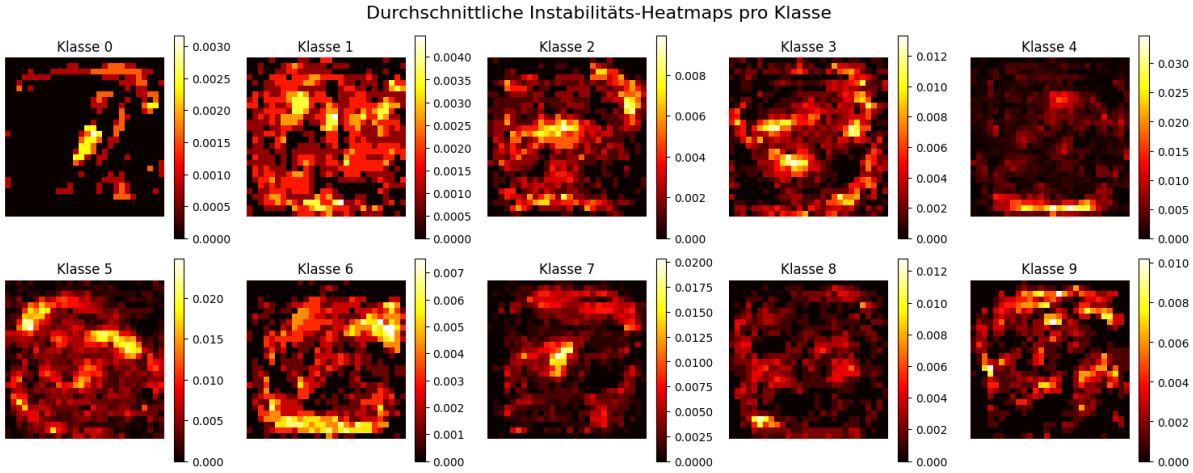

Abbildung 3: Diese Heatmap zeigt die besonders Instabilen Bereiche. (Entstanden durch Netz B)

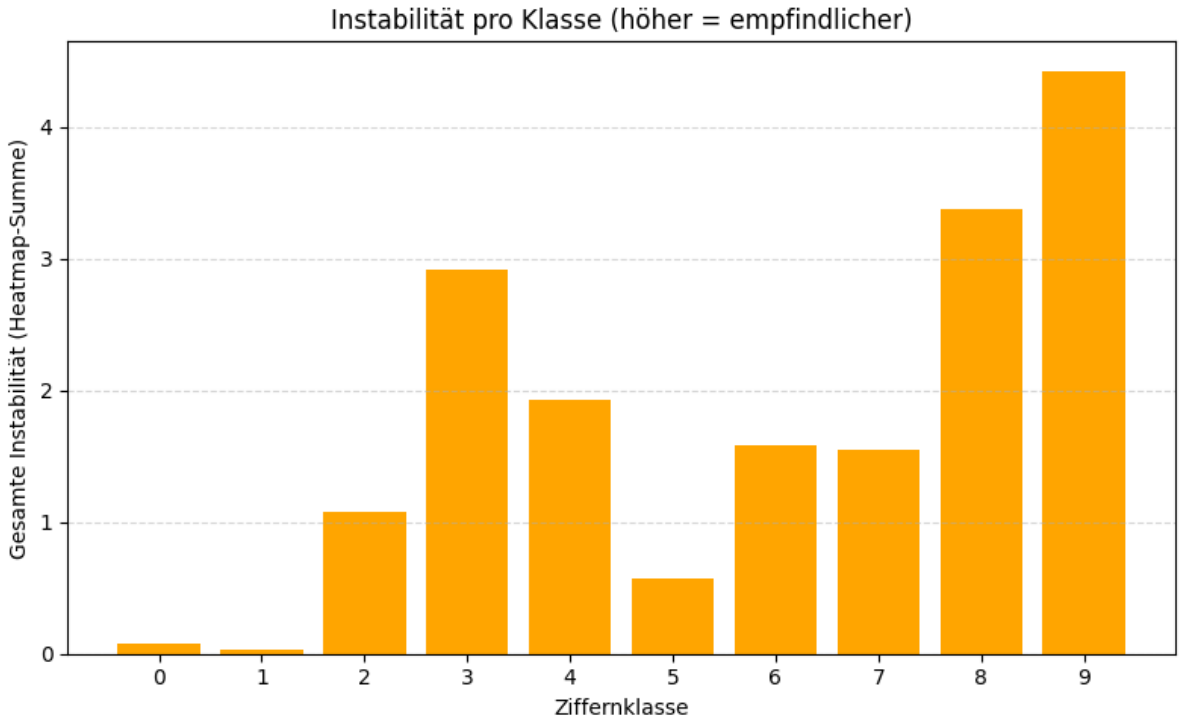

Abbildung 4: Instabilität pro Klasse. (Entstanden durch Netz A)

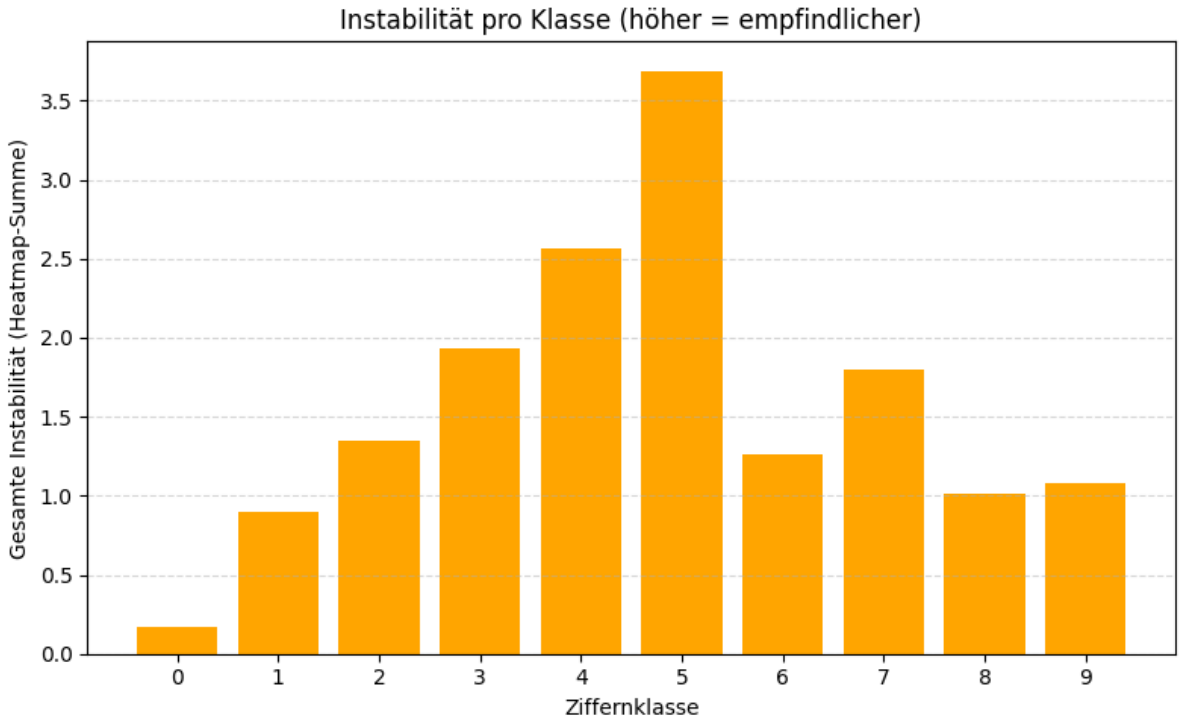

Abbildung 5: Instabilität pro Klasse. (Entstanden durch Netz B)

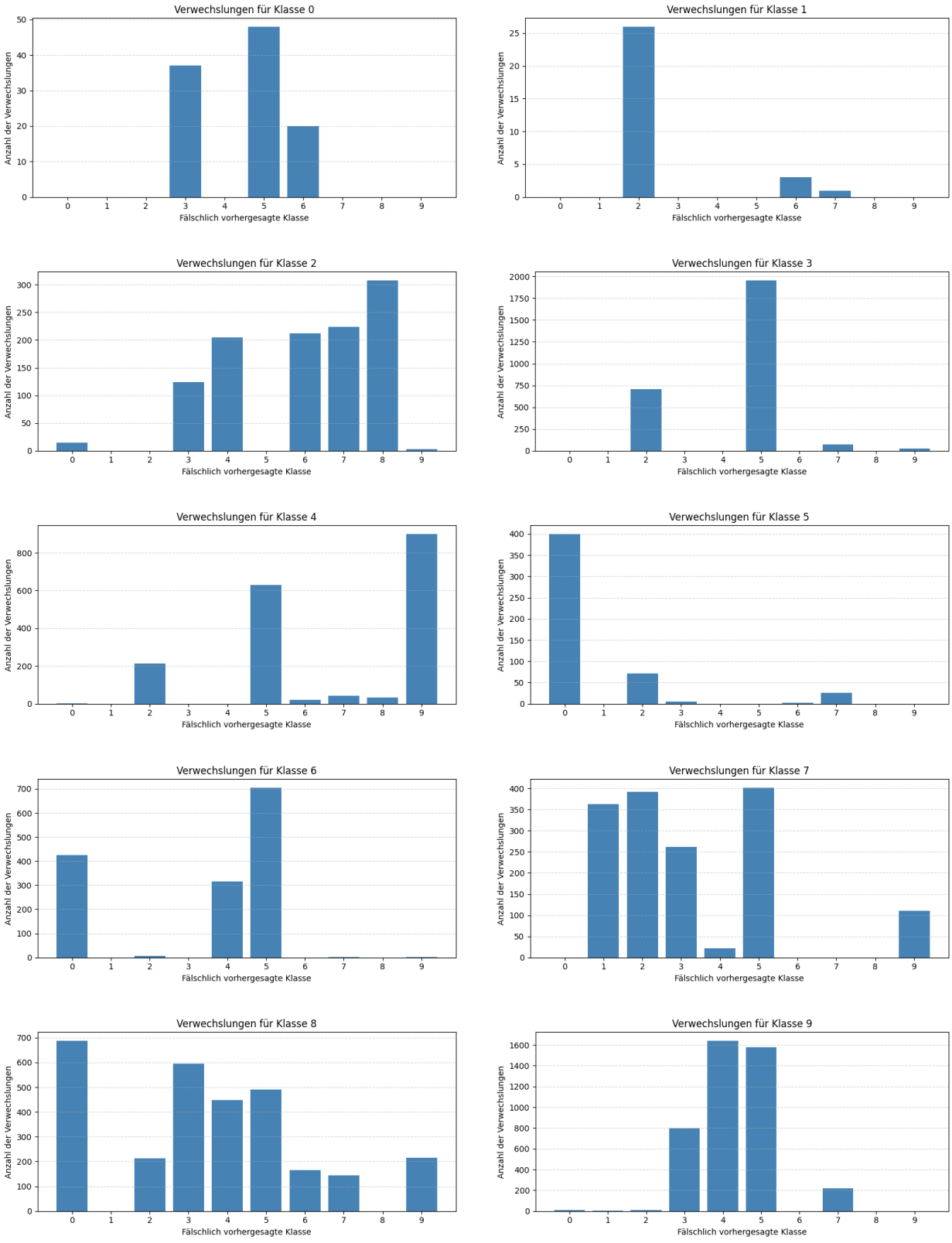

Abbildung 6: Verwechslungstabelle (Entstanden durch Netz A)