Praktikumsbericht

DLO Deep Learning und Objekterkennung
Übungsblatt 5

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: 23. Mai 2025

Aufgabenstellung

Das Ziel ist die Klassifizierung von Bildern auf denen meistens eine Hand in einer von drei Posen (Schere, Stein, Papier) abgebildet ist. Eine vierte Kategorie enthält Bilder, die keiner der drei Klassen zuzuordnen sind, weil sie Hände in anderen Posen oder gar keine Hand beinhalten. Dazu soll ein Faltungsnetzwerk konzipiert und trainiert werden.

Zielsetzung

Die Trainingsdaten beinhalten einige potenziell problematische Bilder. Einige Bilder sind stark verzerrt, weil sie im Hochformat aufgenommen wurden. In diesen Fällen ist die Hand dann auch vertikal abgebildet anstatt wie sonst horizontal. Außerdem gibt es einige tätowierte Hände und Bilder mit Schlagschatten. Diese Formen sind für die Klassifizierung nicht relevant und könnten zu Fehlern führen. Wir schätzen, dass eine Genauigkeit von 95% erreichbar ist.

Vorbereitung der Daten

Die Skalierung der Bilder stellte einen entscheidenden Schritt in der Datenvorbereitung dar. Große Bildauflösungen benötigen mehr VRAM und führen zu einer langsameren Trainingsgeschwindigkeit. Kleinere Bildgrößen hingegen beeinträchtigen die Genauigkeit der Netzergebnisse. Im finalen Modell, das in den Tests die besten Resultate erzielte, wurden die Bilder auf 384×384 Pixel skaliert. In Kombination mit unserem Netzwerk und einer Batch-Größe von 32 Bildern wird der auf der verwendeten RTX 3070 verfügbare VRAM von 8GB optimal ausgenutzt. Die Kombination der drei Bildkanäle auf einen Graukanal ermöglicht zudem auch eine höhere Auflösung. Um die Ergebnisse zu evaluieren, wurde das Netz sowohl mit RGB als auch Grauwerten evaluiert.

Struktur

Unser Netz besteht aus zwei Teilen. Der erste Teil soll zunächst in kleineren und dann in größeren Bildbereichen abstrakte Muster erkennen, die im zweiten Teil zur Klassifizierung herangezogen werden. Zur Mustererkennung verwenden wir sechs Faltungsschichten. Auf jede Faltungsschicht folgt eine nicht-lineare Aktivierungsfunktion, damit die beiden Faltungsschichten nicht effektiv zu einer zusammenfallen. Danach wird Max Pooling angewendet, um im jeweils nächsten Schritt Muster in größeren Bildbereichen zu erkennen und die Datenmenge zu reduzieren, damit Training und Inferenz nicht zu lange dauern.
Der zweite Teil zur Klassifzierung rechnet zunächst die hochdimensionalen Ausgangsdaten aus dem ersten Teil durch Apaptive Average Pooling auf eine einzige Dimension runter. Darauf folgen zwei vollständig verbundene lineare Schichten jeweils mit einer ReLU Aktivierungsfunktion.

Beobachtungen

Das Netz erreicht eine Genauigkeit von 96% auf dem Validierungsdatensatz.

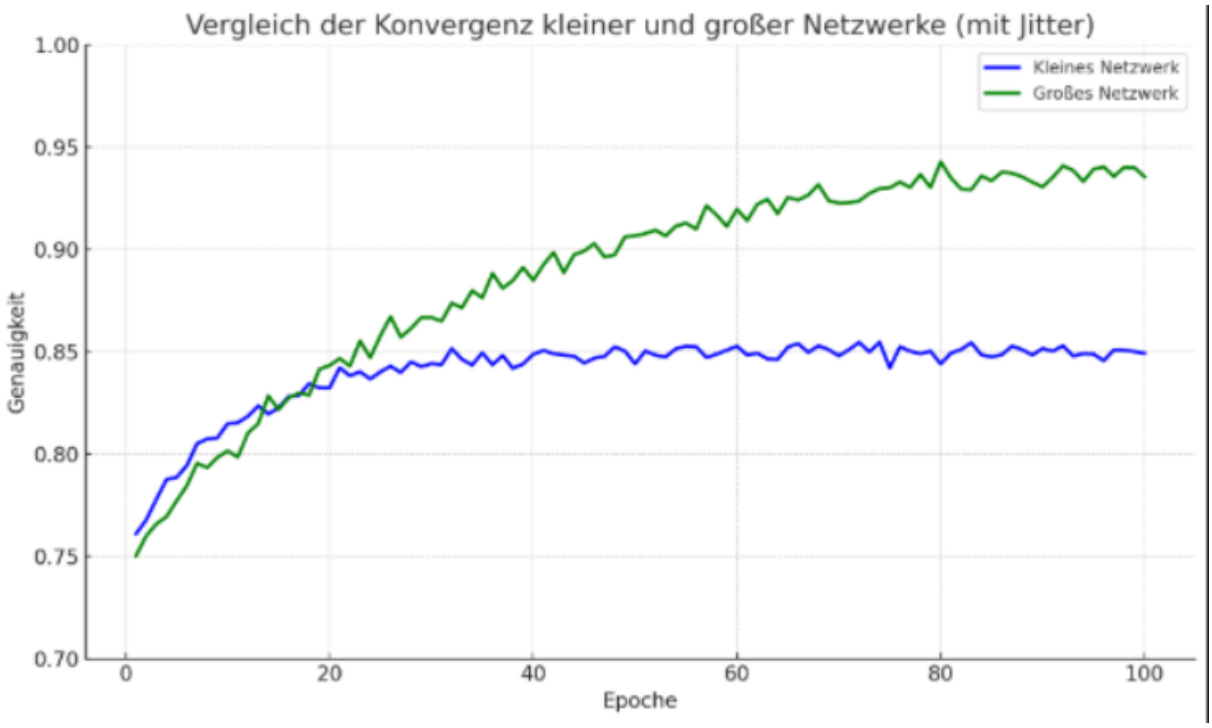


Abbildung 1: Vergleich zweier Netze mit unterschiedlicher Größe

Netzgröße

Bei der Evaluation der Netzarchitektur erweist sich insbesondere der Vergleich der Ergebnisse auf dem Trainings- und dem Validierungsdatensatz als aufschlussreich. Es lässt sich grundsätzlich beobachten, dass größere Netzwerke tendenziell eine höhere Genauigkeit auf dem Trainingsdatensatz erzielen. Dieses Verhalten ist auf die gesteigerte Modellkapazität zurückzuführen, die eine genauere Anpassung an die Trainingsdaten erlaubt. Allerdings führt eine zunehmende Netzgröße nicht zwangsläufig zu einer verbesserten Leistung auf dem Validierungsdatensatz. Ab einem bestimmten Punkt kann die Genauigkeit auf den Validierungsdaten sogar wieder abnehmen – ein typisches Indiz für Overfitting. Besonders die Größe des vollständig verbundenen (fully connected) Layers spielt hierbei eine zentrale Rolle, da dieser die im Netzwerk extrahierten Merkmale abstrahieren und generalisieren soll. Eine zu hohe Kapazität in diesem Bereich kann dazu führen, dass das Modell nicht gut generalisiert.

Transformationen

Bild Input Größe

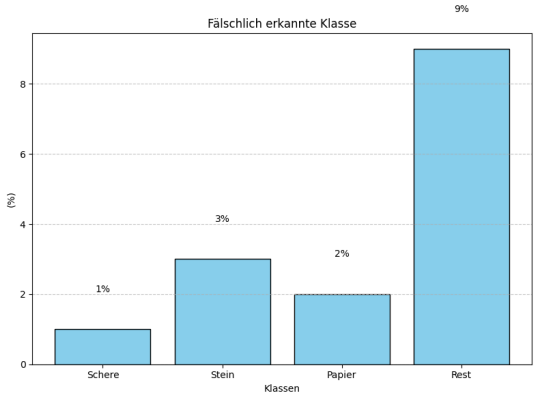
Die Veränderung der Bildgröße hat einen deutlichen Einfluss auf das Trainingsverhalten des Netzwerks. Grundsätzlich gilt: Je kleiner die Eingabebilder, desto geringer kann die Komplexität des Netzwerks ausfallen, und desto schneller konvergiert die Genauigkeit gegen ein stabiles Leistungsniveau. Ab einem gewissen Punkt ist eine Fortsetzung des Trainings nicht mehr zielführend, da keine signifikante Leistungssteigerung mehr erzielt wird.

Im Gegensatz dazu benötigen größere Bildgrößen (mit größeren Netzwerken) in der Regel deutlich mehr Trainingszeit, um gegen ein vergleichbares Konvergenzniveau zu laufen.

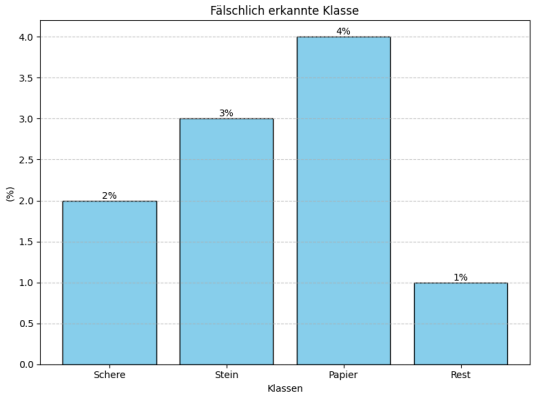
Dieses liegt jedoch typischerweise höher, sodass bei ausreichend langer Trainingsdauer eine bessere Generalisierungsleistung erreicht werden kann. Dieses Verhalten lässt sich auch in der Abbildung 1 erkennen.

Graustufen

Durch das Transformieren der Bilder in Graustufen kann mit gleicher VRAM-Ausnutzung ein größerer Bildausschnitt trainiert werden. Dies könnte zu besseren Ergebnissen führen, da viele wichtige Bildinformationen auf der Strukturebene und nicht auf der Farbebene liegen. Bei der Betrachtung der Verwechslungen zwischen den einzelnen Klassen fällt ein interessantes Muster auf. Dies ist insbesondere beim Vergleich zwischen einem Netz, das mit Graustufen trainiert wurde, und einem, das mit Farbwerten trainiert wurde, erkennbar (Abbildung 3). Das Netz, welches auf Graustufen trainiert wurde, verwechselt besonders oft die Klasse „Rest". Dies lässt sich damit erklären, dass Farbinformationen vor allem für die Erkennung dieser Klasse bzw. der anderen Klassen nützlich sind.



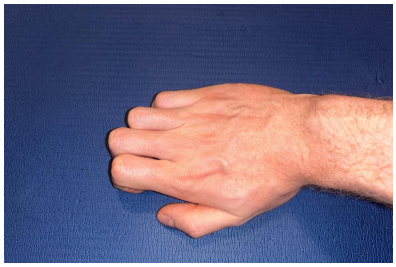
(a) Graustufen



(b) RGB-Werte

Abbildung 2: Vergleich der Verwechslungen zwischen Klassen bei einem Netzwerk, das mit Grauwerten trainiert wurde, und einem, das mit Farbwerten trainiert wurde

Fehler



(a) Klassifizierung: Stein, Label: Rest

(b) Klassifizierung: Papier, Label: Schere

(c) Klassifizierung: Papier, Label: Rest

Abbildung 3: Drei Beispiele, die falsch klassifiziert wurden.

Viele der Fehler, die unser Netz macht, sind nachvollziehbar. Einige Bilder (z.B Abbildung 3a) sind kaum eindeutig klassifizierbar. In anderen Bildern ist die Hand unvollständig

oder am Bildrand (z.B Abbildung 3b). Auf den Bildern, auf denen keine Hände zu sehen sind, gibt es einige Gegenstände (z.B. Abbildung 3c), die ungefähr die Form einer Hand haben und vermutlich daher als solche erkannt werden.

Fazit

Faltungsnetze eignen sich gut zur Klassifizierung von Bildern. Dabei schneiden Netze mit einer hohen Kapazität, also tiefe Netze mit mehrdimensionalen Faltungsmatrizen, besonders gut ab, aber nur wenn sie stark genug regularisiert werden, um Überanpassung zu vermeiden. Das Training solcher Netze ist allerdings rechenintensiv.