Multi-Klassen-Erkennung mit tiefen Lernarchitekturen des Typs VGG-19

Joshua Oppel, Martin Patrick Pauli, Martin Golz

Hochschule Schmalkalden, Fakultät Informatik, Blechhammer 4, 98573 Schmalkalden

Abstract

Tiefe Faltungsnetze gehören neben den tiefen rekurrenten Netzen zu den Methoden der computerbasierten Intelligenz, die in den letzten zehn Jahren zu bahnbrechenden Erfolgen in der Bild- und Audiosignalverarbeitung geführt haben. In diesem Beitrag werden Faltungsnetze vom Typ VGG-19 eingesetzt und es wird der Frage nachgegangen, wie gut computergrafisch generierte Objekte klassifiziert werden können, die in relativ hoher Anzahl und in verschiedenen Projektionen erzeugt wurden. Je Objektklasse wurden 9.072 Beispiele generiert und es wurde untersucht, wie stark die Klassifikationsgenauigkeit von der Anzahl der Objektklassen abhängen. An drei balancierten Datensätzen mit 19, 119 und 219 verschiedenen Objektklassen wurden Klassifikationsgenauigkeiten von 97,7%, 92,5% bzw. 86,7% erreicht. Es werden mögliche Gründe für diese Tendenz diskutiert.

1. Einleitung

Tiefe Faltungsnetze (CNN, convolutional neural networks) [1] erreichen seit dem Jahr 2012 bahnbrechende Erfolge in der Bildanalyse und Mustererkennung [2]. Aufgrund der tiefen hierarchischen Anordnung von adaptiven digitalen Filtern mit schwacher Nichtlinearität erreichen diese Netze hohe Klassifikationsgenauigkeiten, ohne dass besondere Maßnahmen in der Signalerfassung, der Datenvorverarbeitung oder der Merkmalsextraktion getroffen werden müssen. CNN verarbeiten Rohdaten und benötigen allerdings hohe Rechenkapazitäten und mächtige Stichproben, um ihr Potential voll entfalten zu können. Wichtige Anwendungsgebiete sind vor allem die Bild- und Videoverarbeitung, die Verarbeitung natürlicher Sprache, Gehirn-Computer-Schnittstellen und die Analyse finanzieller Zeitreihen. Die Erfolge wurden nicht nur wegen der besonderen Architektur und umfassenderen Spektrums an Trainingsalgorithmen ermöglicht, sondern insbesondere auch wegen der entstandenen sehr großen, öffentlich zugänglichen Bilddatenbanken, wie zum Beispiel die ImageNet-Daten [3], und wegen der Hochleistungs-Computersysteme, unter anderem begründet durch Fortschritte in der GPU-Technologie und durch groß angelegte verteilte Systeme. Ein Meilenstein in der Entwicklung tiefer Faltungsnetze war das Netz VGG-19, der Visual Geometry Group der Universität Oxford, das aus 19 neuronalen Schichten besteht [4].

2. Material und Methoden

Mit dem Echtzeit-3D-Betrachter LDView wurden Bilder von Klemm-Bausteinen des dänischen Spielzeugherstellers LEGO© computergrafisch generiert [5]. Diese quelloffene Software wird von einer Entwicklergemeinschaft stetig weiterentwickelt und enthielt im Jahre 2022 14.782 Modelle von Lego-Bausteinen. Es wurden pro Klasse (Bausteinart) 9.072 Trainingsbeispiele und 1.890 Validierungsbeispiele (82,8% / 17,2%) erzeugt. Die Beispiele unterscheiden sich bzgl. der Rotationsmatrix des Objektes. Es wurden Dre-

hungen um 7 fixierte Achsen und je Achse eine 360°-Drehung in 5°-Schritten ausgeführt. Dies führt zu einer Anzahl an Rotationen von 7·72=504. Zudem wurde die Richtung des Beleuchtungsvektors in 9 festgelegten Varianten verändert. Eine weitere Variation wurde durch Ein- und Ausschalten der Kontourlinien (Umrandung um das Modell) ausgeführt. Damit entstand pro Klasse eine Gesamtmenge von 7·72·2·9=9.072 Bildern. Mit zufällig gewählten Parametern der Rotationsmatrizen und des Beleuchtungsvektors wurden die 1.890 Elemente der Validierungsmenge generiert. Alle Bilder wurden als RGB-Tensor im Format 224×224×3 gespeichert. Im verlustfrei komprimierten Rastergrafik-Bildtyp PNG lag die Speichermenge bei nur 11,1 KB/Bild. Durch die Wahl von bis zu 219 verschiedenen Klassen und 9.072+1.890 Beispielen pro Klasse lag der Speicherumfang des Datensatzes bei 26,65GB. Alle Bilder hatten einen einfarbigen Hintergrund und enthielten keine weiteren Objekte. Hier wurden drei verschiedene Datensätze mit 19, 119 und 219 Klassen erzeugt, um den Einfluss der Klassenanzahl untersuchen zu können.

Abbildung 1: Beispiele der generierten Bilder von Klemm-Bausteinen, wie sie für das Training verwendet wurden.

Die so generierten RGB-Tensoren waren die Eingangsdaten des Faltungsnetzes vom Typ VGG-19 [4]. Dieses Netz ist dadurch charakterisiert, dass es in allen Faltungsschichten mit sehr kleinen Kernmatrix-Formatparametern von 3x3 arbeitet. Nach zwei oder vier hintereinander angeordneten Faltungsschichten liegt eine Bündelungsschicht (maximum pooling), die die Breite und Höhe der nächsten Faltungsschicht halbiert. Im Gegenzug wird die Tiefe der Schicht verdoppelt. Durch dieses Prinzip hat bspw. die erste Faltungsschicht ein Format von 224x224x64. Nach der ersten, zweiten und dritten Bündelung beträgt das Format 112x112x128 bzw. 56x56x256 bzw. 28x28x512; nach der fünften Bündelung liegt es bei 7x7x512, da die Autoren festgelegt hatten, ab der vierten und fünften Bündelung die Tiefe nicht mehr zu verdoppeln, sondern konstant zu lassen. Danach folgen drei vollständig verbundene Schichten mit 4096, 4096 und 1000 Neuronen. Die letzte Zahl war durch den Datensatz vorgegeben, da er 1000 Klassen enthielt. Hier wurde dieser Parameter abgeändert auf die jeweils gegebene Klassenanzahl (19, 119, 219) und das Netz wurde nachtrainiert nach dem Prinzip des Transfer-Lernens. Um die erforderliche Trainingszeit zu reduzieren, wurde eine zwanzigste Faltungsschicht eingefügt mit anschließender Bündelung (average pooling), sodass die nachfolgenden vollständig verbundenen Schichten auf je 256 Neuronen begrenzt werden konnten und die letzte Schicht enthielt für jede Klasse ein Neuron, um mit der Soft-

Max-Aktivierungsfunktion [6] die klassenbedingte Wahrscheinlichkeitsschätzung ausgeben zu können. Die Anzahl der zu trainierenden Parameter lag bei ≈1,23·10^8, während sie beim ursprünglichen VGG19-Netz bei ≈1,44·10^8 lag.

Für die numerischen Untersuchungen diente ein moderner Arbeitsplatz-Computer mit GPU-Subsystem (NVIDIA RTX 2070). Die Bildgenerierung erforderte ca. 23 Tage Rechenzeit, während das Nachlernen und die Validierung ca. 15 bzw. 0,5 Tage benötigte.

3. Ergebnisse

Die Objekte einiger Klassen wiesen nur geringfügige Unterschiede in ihrer Form auf. Einige waren sogar identisch und unterschieden sich nur durch einen Aufdruck. Diese Objekte waren relativ flach. Solche Objektklassen lassen sich nur schwer unterscheiden.

Die Software LDView zum Erstellen fotorealistischer Bilder der Objekte aus unterschiedlichen Perspektiven stellte sich als nicht gänzlich fehlerfrei heraus. Aufgrund einer mangelhaften Clipping-Prozedur konnte es in seltenen Fällen vorkommen, dass die Kamera innerhalb des Objektes lag und somit die generierten Bilder unrealistisch waren.

Für den ersten Datensatz (19 Klassen) wurden mittlere Top-1- und Top-5-Genauigkeiten von 97,71% bzw. 99,99% an der Validierungsmenge erreicht. Die Klasse mit der geringsten Top-1-Genauigkeit lag bei 89,26%. Für den zweiten Datensatz (119 Klassen) wurden mittlere Top-1- und Top-5-Genauigkeiten von 86,38% bzw. 98,28% erzielt. Beim dritten Datensatz (219 Klassen) lag die Top-1-Genauigkeit bei 86,68%.

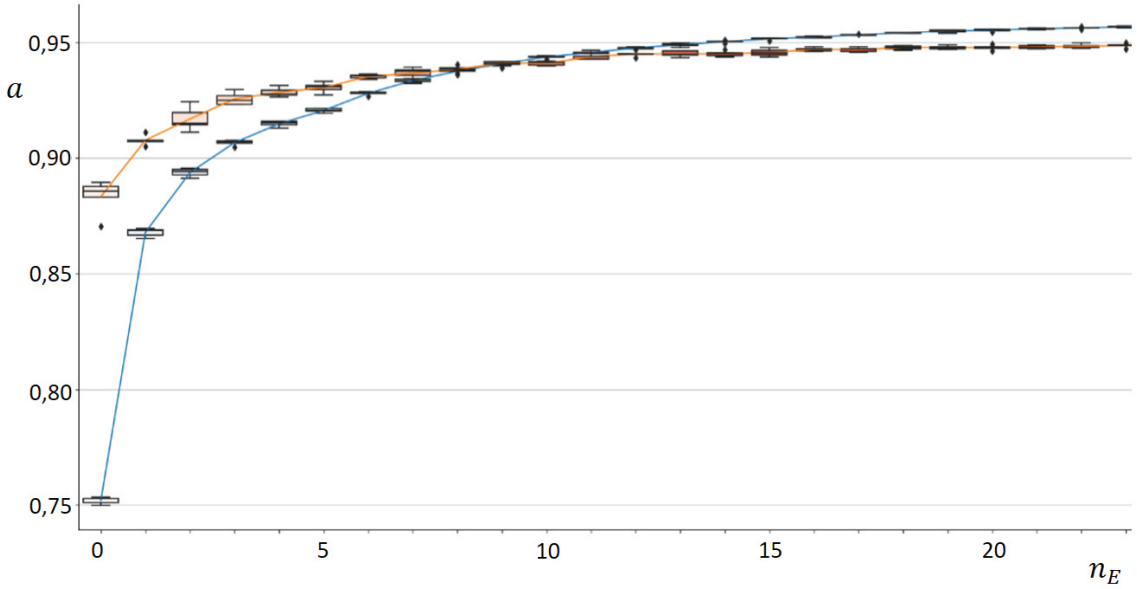

Abbildung 2: Perzentile der Klassifikationsgenauigkeiten an Trainings- (blau) und Validierungsmengen (orange) über der Anzahl der Trainingsepochen für den größten Datensatz mit 19 Objektklassen und ca. 1,9 und 0,41 Millionen Trainings- bzw. Validierungsbeispielen.

Die Zwischenergebnisse der fünffachen Kreuzvalidierung im Trainingsverlauf wurden für den ersten Datensatz gespeichert, sodass die Perzentile der Genauigkeitsverteilung als Kastengrafik über der Epochenanzahl dargestellt werden können (Abbildung 1). Die Kästen werden durch das untere und obere Quartil begrenzt und enthalten das mittlere

Quartil (Median). Während die Trainingsgenauigkeit anfangs geringer ist als die Validierungsgenauigkeit, kommt es ab Epoche 9 zu umgekehrter Relation beider. Das Training wurde nach 23 Epochen abgebrochen, obwohl die Trainingsgenauigkeit noch eine leicht steigende Tendenz hat, die aufgrund der sehr geringen Streuung noch signifikant ist. Aber die Validierungsgenauigkeit befindet sich seit Epoche 16 auf einem Plateau und es sind keine signifikanten Steigerungen mehr nachweisbar.

4. Diskussion

Mit erhöhter Anzahl der Klassen von 19 auf 119 und auf 219 verringerte sich die Validierungsgenauigkeit, obwohl der Stichprobenumfang proportional anwuchs, denn stets wurden 9.072 Beispiele pro Klasse generiert. Die Ursache scheint in der begrenzten Kapazität der Hypothesenklasse des verwendeten Faltungsnetzes zu liegen. Offenbar waren die von uns vorgenommenen Modifikationen durch Hinzufügen einer weiteren Faltungsschicht mit Bündelung und die sich ergebende Verringerung der Größe der vollständig verbundenen Schichten zu kräftig, um die erhöhte Diversität der Trainingsmenge beherrschen zu können. Diese ersten Untersuchungen müssen fortgesetzt werden und es sollte ermittelt werden, welche Architekturänderungen die Validierungsgenauigkeit bei hoher Klassenanzahl signifikant anheben. Dazu sind allerdings erheblich größere Berechnungs-Ressourcen erforderlich.

Quellen

[1] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

[2] Hinton GE, Krizhevsky A, & Sutskever I (2012) Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25(1):1106-14

[3] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. Proc. IEEE Conf Computer Vision Pattern Recognition: 248-255

[4] Simonyan K & Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint:1409.1556

[5] Ldraw.org; Parts & Tools: Latest Parts, 2022.
online verfügbar: https://www.ldraw.org/parts/latest-parts.html

[6] Aggarwal, C. C. (2018). Neural networks and deep learning: A textbook. Springer.