KI für Alle 2: Verstehen, Bewerten, Reflektieren

Themenblock Bildklassifikation und Bildsegmentierung: 07_05Evaluation_Metriken

Wie gut ist mein Modell? Bewertung von Bildklassifikationsmodellen

Erarbeitet von
Dr. Ludmila Himmelspach

Lernziele 1
Inhalt 2
Einstieg 2
Evaluationsablauf von Bildklassifikationsmodellen 2
Bewertung von Bildklassifikationsmodellen im medizinischen Bereich 2
Evaluationsergebnisse eines beispielhaften CNNs 3
Take-Home Message 5
Quellen 5
Disclaimer 6

Lernziele

- Du kannst erklären, wie die Evaluation eines Bildklassifikationssystems im Allgemeinen abläuft
- Du kannst erklären, worauf man bei der Evaluation eines Klassifikationssystems im medizinischen Bereich achten soll

Projekt ai4all „KI für Alle"
HeiCAD –Heine Center for Artificial Intelligence and Data Science

Inhalt

Einstieg

Ein Bildklassifikationssystem muss bestimmte Qualitätsanforderungen aufweisen, damit es im laufenden Betrieb eingesetzt werden kann. Bei der Bewertung solcher Systeme muss vor allem darauf geachtet werden, inwieweit es in der Lage ist, ein gegebenes Problem zu lösen. In diesem Video erfährst Du, wie die Evaluation von maschinellen Bildklassifikationssystemen im Allgemeinen abläuft. Außerdem gehen wir kurz darauf ein, worauf man bei der Bewertung der Klassifikationsmodelle im medizinischen Bereich achten soll. Zum Schluss schauen wir uns noch die Evaluationsergebnisse eines beispielhaften CNNs an, das wir auf einem öffentlich verfügbaren Röntgendatensatz trainiert haben.

Evaluationsablauf von Bildklassifikationsmodellen

Der Ablauf der Evaluation eines Bildklassifikationsmodells sieht wie folgt aus: Zuerst werden die Vorhersagen für die Testbilder mit Hilfe des trainierten CNN-Modells erstellt, d.h. für den Teil der annotierten Bilder, die fürs Training des neuronalen Netzes nicht verwendet wurden. Danach wird geprüft, inwieweit die vom CNN vorhergesagten Klassenzugehörigkeiten mit den Annotationen von Expert*innenen übereinstimmen.

Wenn der annotierte Korpus relativ klein ist, was zum Beispiel im medizinischen Bereich oft der Fall ist, können alle Bilder aus dem Korpus mittels der Kreuzvalidierung bzw. der verschachtelten Kreuzvalidierung zum Trainieren des Klassifikationsmodells verwendet werden. Dann findet die Evaluation anhand der Vorhersagen auf den Validierungsdaten statt.

Bewertung von Bildklassifikationsmodellen im medizinischen Bereich

Die Bewertung der Modelle für Bildklassifikation läuft genauso wie die Bewertung anderer Klassifikationsmodelle ab. Man kann dafür auch die üblichen Evaluationsmetriken wie Accuracy, Sensitivität, Spezifität und F-Score einsetzen. Allerdings haben diese Evaluationsmetriken bei der Bewertung der Klassifikationssysteme im medizinischen Bereich unterschiedliche Gewichtungen, weil unterschiedliche Arten von Klassifikationsfehlern verschiedene Folgen für Patient*innen haben können.

Wenn zum Beispiel eine Röntgenaufnahme mit Lungenentzündung der falschen Klasse zuordnet wird, wird der entsprechende Patient oder die Patientin nicht rechtzeitig behandelt. Als Folge dessen kann sich der Erreger unter Umständen im gesamten Körper ausbreiten. Wird eine Röntgenaufnahme ohne Lungenentzündung als eine mit Lungenentzündung eingeordnet, bekommt der Patient oder die Patientin unter Umständen eine Medikation, die er*sie nicht braucht. So wäre eine hohe *Sensitivität*, die der Trefferquote der Erkrankten entspricht, für ein System zur Erkennung einer schnell

voranschreitenden Erkrankung wie Lungenentzündung wichtiger, als die *Spezifität*, die der Trefferquote der Gesunden entspricht.

Bei der Erkennung von langsam voranschreitenden Erkrankungen, wie zum Beispiel Zahnkaries, deren Diagnostik normalerweise in 6- bis 12-monatigen Abständen durchgeführt wird, ist dagegen eine hohe *Spezifität* wichtiger als *Sensitivität*. Denn eine übersehene Karies, die möglicherweise bei der nächsten Untersuchung entdeckt und behandelt wird, wäre weniger schlimm als eine invasive Therapie eines gesunden Zahns.

Quelle [1]

Aufgrund verschiedener Randfälle in Datenbasis kann man eigentlich von keinem Klassifikationssystem eine hundertprozentige Erfolgsquote erwarten. Bei der Wahl der Evaluationsmaße für die Bewertung der Klassifikationsmodelle sollte man allerdings die Konsequenzen der Klassifikationsfehler für den jeweiligen Anwendungsbereich berücksichtigen.

Evaluationsergebnisse eines beispielhaften CNNs

Für die automatische Erkennung der Lungenentzündung in Röntgenaufnahmen haben wir beispielhaft das VGG16 Net als Klassifikationsmodell nachimplementiert. Das VGG16 Net wurde bereits 2015 von Karen Simonyan und Andrew Zisserman vorgestellt.

Quelle [2]

Wie Du in der Abbildung sehen kannst, hat VGG16 Net eine relativ einfache Architektur. Es besteht aus 13 Convolutional und drei vollständig verbundenen Schichten. Auf jeden zweiten und später auf jeden dritten Convolutional Layer folgt jeweils ein Max-Pooling Layer. Die ersten beiden Convolutional Layer haben 64 Filter mit 3 x 3 großen Wahrnehmungsfeldern. Nach den ersten drei Pooling Layern verdoppelt sich die Filteranzahl in darauffolgenden Convolutional Layern.

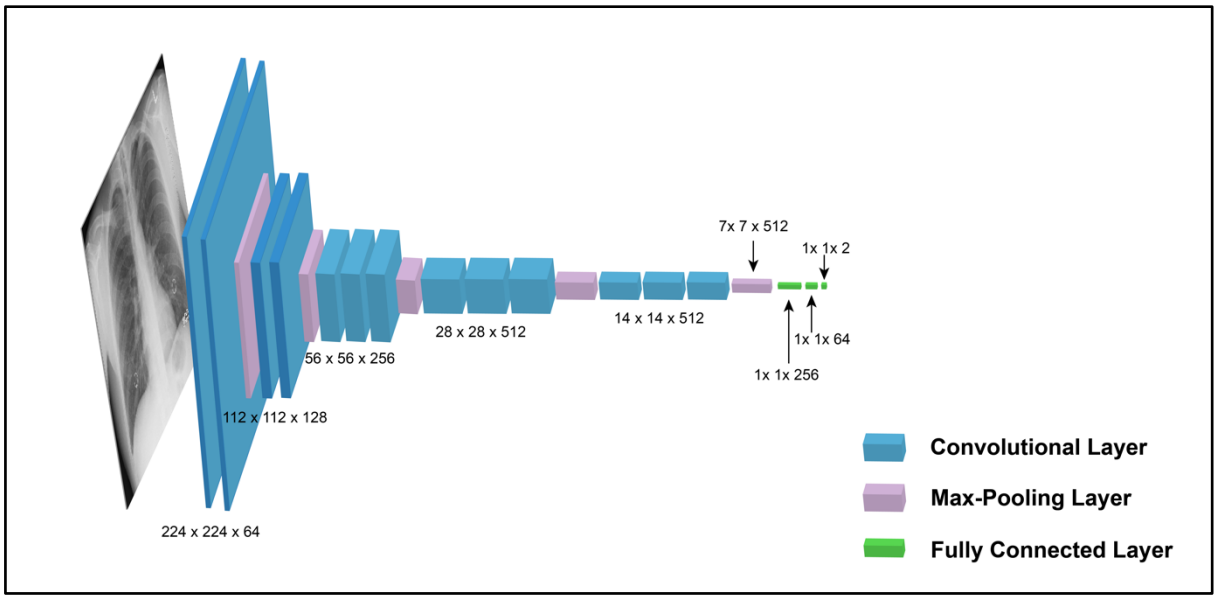


Abbildung 1: Die Architektur des angepassten VGG16 Nets.

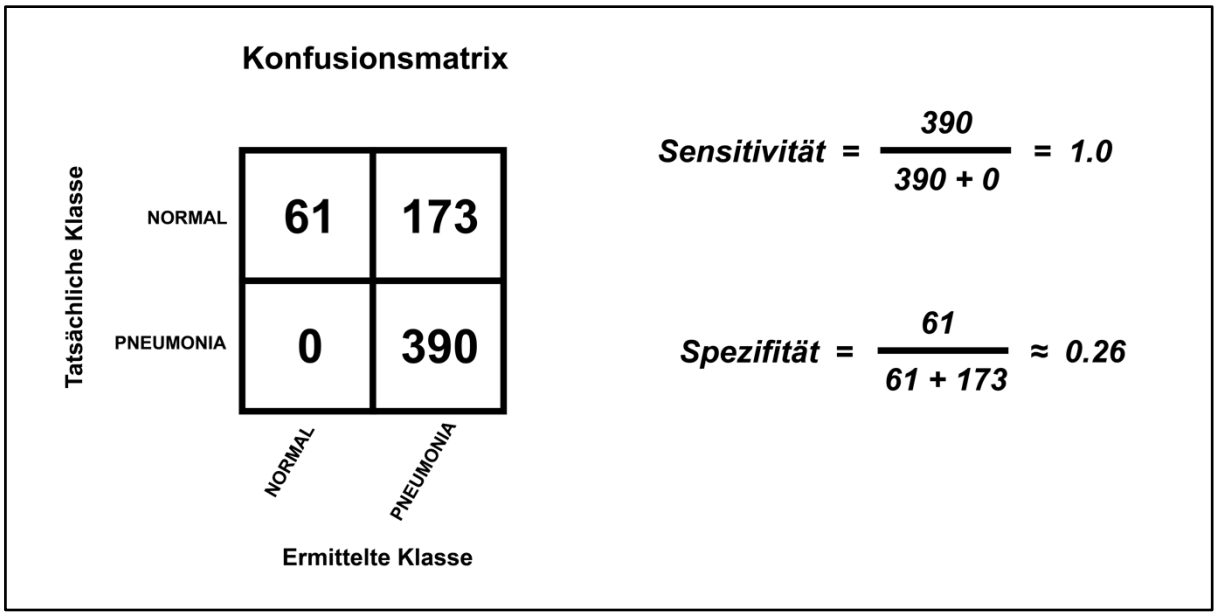
Anders als VGG16 Net, das für ein 1000-Klassen-Problem entwickelt wurde, besitzt die letzte vollständig verbundene Schicht unseres Netzes nur zwei Neuronen, die die geschätzten Wahrscheinlichkeiten für die Klassen „Lungenentzündung“ bzw. „keine Lungenentzündung“ berechnen.

Wir haben unser Convolution Neural Network auf dem öffentlich verfügbaren Datensatz mit 5216 Röntgenaufnahmen des Brustkorbs im Trainingssatz und 18 Röntgenbildern im Validierungssatz über 15 Epochen auf einem PC trainiert. Der Trainingssatz enthielt dabei 3875 Aufnahmen mit Lungenentzündung und 1341 Röntgenaufnahmen, die als „gesund“ gekennzeichnet wurden. Wie man sieht, ist die Klassenverteilung unausgewogen.

Quelle [3]

Um den Speicherplatz für die Feature Maps in Convolutional Layern zu reduzieren, wurden alle Röntgenaufnahmen im Rahmen der Vorverarbeitung auf eine Größe von 224 x 224 Pixeln verkleinert.

Wir haben das trainierte CNN-Modell anhand der 624 Röntgenaufnahmen des Testdatensatzes evaluiert. Die Ergebnisse sind in der Konfusionsmatrix zusammengefasst.



Als Erstes fällt auf, dass alle Röntgenaufnahmen mit Lungenentzündungen vom Modell richtig klassifiziert wurden. Das entspricht einer Sensitivität von 1,0, was bei einer schnell voranschreitenden Erkrankung erstmal als positiv zu bewerten ist. Allerdings wurden auch viele Röntgenaufnahmen, die von Expert*innen als „gesund“ gekennzeichnet wurden, der falschen Klasse zugeordnet. Der Wert für Spezifität liegt bei ungefähr 0,26, was natürlich viel zu niedrig ist. Der Grund dafür sollte in erster Linie in der Unausgewogenheit des Datensatzes vermutet werden, mit dem das Klassifikationsmodell trainiert wurde. Da der Anteil der Aufnahmen dort bei über 70 % lag, wurde dem Modell ein gewisses Bias antrainiert, die Röntgenaufnahmen der Klasse "Lungenentzündung" mit einer höheren Wahrscheinlichkeit zu zuordnen.

Natürlich eignet sich dieses Netz für den Einsatz in einem Krankenhaus nicht. Andererseits werden im Realfall die Trainingsbilder nicht so stark verkleinert und die CNNs auf leistungsstarken GPUs trainiert.

Take-Home Message

In diesem Video hast Du den allgemeinen Ablauf der Evaluation von maschinellen Bildklassifikationssystemen kennengelernt. Außerdem hast Du erfahren, worauf man bei der Bewertung der Klassifikationsmodelle im medizinischen Bereich achten soll.

Quellen

Quelle [1] Schwendicke, F., Krois, J. (2022, September). *Sensitivität und Spezifität: Was ist wann wichtig? KI in der Zahnarztpraxis – Teil 3.* ZM Online. https://www.zm-online.de/artikel/2022/die-strategie-der-investoren/sensitivitaet-und-spezifitaet-was-ist-wann-wichtig

Quelle [2] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, 1–14.

Quelle [3] Kermany, D., Zhang, K., Goldbaum, M. (2018), Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification, Mendeley Data, V2, doi: 10.17632/rscbjbr9sj.2

Disclaimer