Projektreport Hautkrebserkennung

Gruppe Cluster-Busters

Thema und Motivation

Die Coronakriese hat gezeigt wie wichtig Machine Learning auch im Gesundheitsbereich sein kann. So wurde beispielsweise die Ausbreitung des Virus schon im Dezember 2019 von einem Algorithmus vorhergesagt. Neben der Prognose der Ausbreitung lassen sich aber auch Anwendungen zur Krankheitserkennung entwickeln und dieser Aufgabe haben wir uns in dieser Arbeit gestellt.

Das Projekt befasst sich mit der Analyse von 10.000 Bilddateien mehrerer Hautkrebsarten. Ziel soll es sein ein Modell zu erstellen, das anhand eines Bildes möglichst genau den darauf abgebildeten Hautkrebstyp vorhersagt.

Related Work

Der ausgewählte Datensatz war Grundlage für einen 2018 durchgeführten Wettbewerb zur Hautkrebserkennung unter dem Namen HAM10000 ("Human Against Machine with 10000 training images") [1]

Aufgrund der Popularität dieser Challenge exisiteren hierzu bereits eine Reihe anderer wissenschaftlicher Arbeiten. Besonders hervorzuheben sind die Arbeiten von Tschandl, die sich mit der Erstellung des Datensatzes befasst[2], und Codella, die sich dem Wettbewerb und der Ergebnisinterpretation widmet.[3]

Wirtschaftlicher Kontext

Da die Aufgabenstellung nicht aus dem betriebswirtschaftlichen Bereich kommt und keine Umsatzzahlen o.ä. analysiert oder vorausgesagt werden, kann der wirtschaftliche Zusammenhang nur oberflächlich betrachtet werden.

Als Einsparpotential im Gesungheitswesen lassen sich prinzipiell drei Punkte ausmachen:

- Vermeidung von falsch-positiv Klassifikationen und folgenden Laborkosten
- Vermeidung von falsch-negativ Klassifikationen und folgenden Kranheitskosten
- Langfristig kürzere Ausbildungszeit der Ärtzte

Die genauen Auswirkungen und Beträge sind jedoch für den Laien schwer abschätzbar und müssten in einer Folgestudie ermittelt werden.

---

[1] https://challenge2018.isic-archive.com/ [Abgerufen am 27.06.2021)
[2] Tschandl, P./Rosendahl, C./Kittler, H., 2018.
[3] Codella, N. et al., 2019.

Verwendete Technologien und Bibliotheken

Da sich andere Methoden wie Lineare Regression oder Random Forests nur schlecht zur Klassifizierung von Bildern eignen und sich für diese Problemstellung sogenannte Convolutional Neural Netwoks, kurz CNN durchgesetzt haben, wurden diese als Methode ausgewählt.

Bei CNN handelt es sich um Neuronale Netze, bei denen mithilfe von mehren Layern die (Bild-)Informationen verarbeitet werden.

Folgende Bibliotheken wurden verwendet:

- Pandas
- Numpy
- Daten
- Matplotlib
- Keras
- Sklearn
- hyperas/hyperopt

Ergebnisse

Nach anfänglichen guten Ergebnissen ließ sich schnell eine Stagnation der Accuracy auf den Trainingsdaten im Bereich von ca. 70-75% feststellen. Mehrfache Änderungen der Anzahl der Epochen, Anzahl der Layer sowie Over- und Undersampling erzielten keine oder keine nennenswerten Ergebnisse.

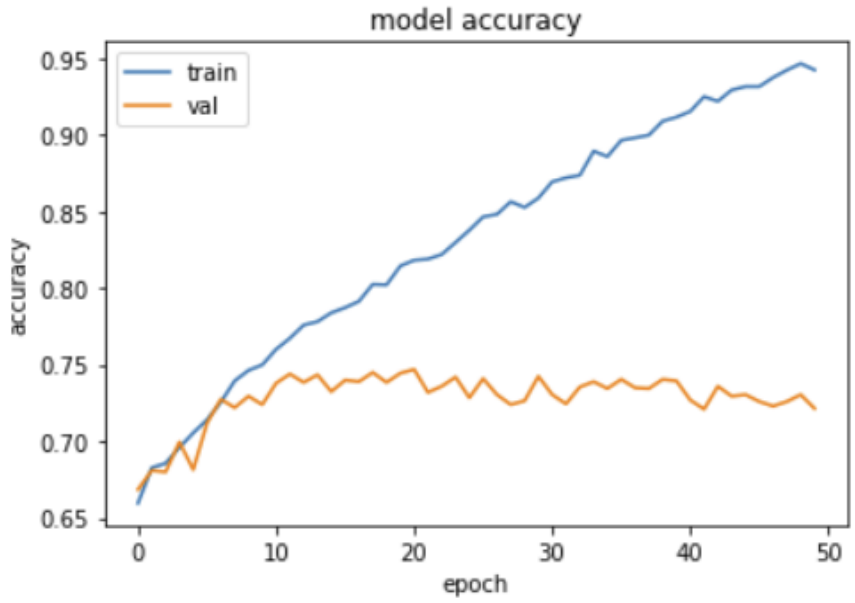

Abbildung :1 Abweichung von Genauigkeit auf Test- und Trainingsdaten, Stagnation bei ca. 73%

Nach dem Experimentieren mit verschiedenen Kombinationen konnte erst bei einer hohen Anzahl der Layer in Verbindung mit Over- und Undersampling eine Steigerung der Acccuracy festgestellt werden.

Weitere Verbesserungen ergaben sich durch iteratives Manipulieren der Hyperparameter wie der Dropout Rate, Art und Anzahl der Layer und Anzahl der Nodes.

Bei dieser Vorgehensweise sind wir auf *hyperas* gestoßen, eine Bibliothek die diese Arbeitsschritte automatisiert und die besten gefundenen Hyperparamater ausgibt.

Schlussendlich konnten wir damit eine Genauigkeit von knapp 90% auf den Testdatensätzen erreichen.

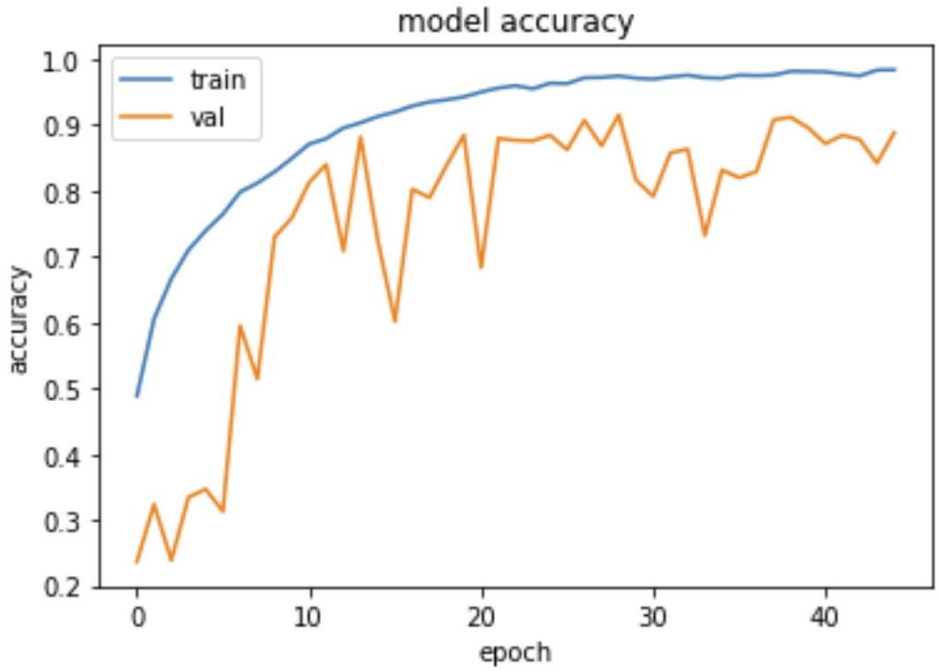

Abbildung 2: Erreichte Genauigkeit mit dem finalen Modell

Kritische Bewertung der Ergebnisse

Die höchste im *HAM-10000* Wettbewerb erreichte Genauigkeit betrug 88,5 %.

Allerdings ist hier anzumerken, dass die Algorithmen der Challenge-Teilnehmer auf einem weiteren, nicht-öffentlichen Datenset getestet wurden und eine andere Ergebebnisgewichtung haben.[4] Die Ergebnisse sind daher nur bedingt vergleichbar.

Auch wenn die erzielten Resultate im Vergleich mit anderen Wettbewerbsteilnehmern gut abschneiden, besteht durchaus noch Verbesserungspotential.

Einerseits mussten natürlich Kompromisse eingegangen werden hinsichtlich der Anzahl der Epochen und der Berechnungszeit, andererseits

Wichtig ist auch anzumerken, dass durch den Blackbox-Charakter der Neuronalen Netze prinzipiell nicht ausgeschlossen werden kann, dass bessere Ergebnisse erzielbar sind. Genausowenig kann die berechnung im Detail nachvollzogen werden.

Bei der Bearbeitung der Aufgabe ist ebenfalls aufgefallen, dass die Bilddateien wahrscheinlich ausschließlich von hellhäutigen Patienten stammen. Dies könnte zum einen der besseren Erkennbarkeit geschuldet sein, könnte aber auch einen racial Bias verursachen.

[4] ISIC 2018 Winners, 08.08.2018.

Anmerkungen zum Quellcode im Anhang

Die aufsummierte Berechnungszeit aller im Jupyter Notebook getesteten und dargestellten Modelle beträgt in etwa 55 Stunden, daher empfehlen wir nur einzelne Modelle zu starten.

Die Trainierten Modelldaten waren leider zu groß für einen Github-Upload oder E-Mail-Versand und können unter dem im Readme-genannten Link abgerufen werden.

Literaturverzeichnis

Codella, N./Rotemberg, V./Tschandl, P./Celebi, M. E./Dusza, S./Gutman, D./Helba, B./Kalloo, A./Liopyris, K./Marchetti, M./others (2019): Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic), in: arXiv preprint arXiv:1902.03368, 2019.

(2018): ISIC 2018 Winners. MetaOptima Wins the ISIC 2018 Disease Classification Competition!, https://www.dermengine.com/blog/metaoptima-isic-2018-skin-disease-classification-artificial-intelligence#:~:text=MetaOptima%20Wins%20the%20ISIC%202018%20Disease%20Classification%20Competition!,-by%20The%20DermEngine&text=We%20are%20excited%20to%20announce,Lesion%20Analysis%20Towards%20Melanoma%20Detection!, am 10.7.2021.

Tschandl, P./Rosendahl, C./Kittler, H. (2018): The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, https://resolver.obvsg.at/urn:nbn:at:at-ubmuw:3-12725.