Exposé zur Bachelorarbeit

“Explainable Artificial Intelligence” für Stammzellmodelle des Leigh-Syndroms

Maximilian Otto
Freie Universität Berlin
Bioinformatik
12. Mai 2022
Betreuer: Jakob J. Metzger

Einleitung:

In der biologischen Forschung gibt es zunehmende Anwendungsgebiete und Verwendungen künstlicher neuronaler Netze (KIs).
Diese angewandten Netze, deren allgemeine Architektur für eine Vielzahl von Fragestellungen verwendet werden kann, erzielen nach einer Anpassung auf teils sehr spezifische Probleme und Datensätze sehr gute Ergebnisse. Beispielsweise kann ein solches Netzwerk [1] mit relativ kurzer Trainingszeit sogar auf kleineren Datensätzen noch immer überraschend gute Prädiktionen liefern und die ihr gegebenen Bilder richtig einordnen.
Wenn es jedoch darum geht, herauszufinden, weshalb die KI in der Lage ist, solch gute Vorhersagen zu treffen, bzw. woran sie die Unterschiede und Merkmale innerhalb eines Datensatzes identifiziert, ist die Erklärung und Interpretation oftmals unintuitiv und komplex.
Um das Verständnis für solche Klassifikationen einfacher zu gestalten, arbeiten viele Forscher*innen daran, immer neue Algorithmen mit möglichst genauen und zugleich einfach zu interpretierenden Ergebnissen und Visualisierungen zu etablieren. Daher gibt es mittlerweile eine Bandbreite an Algorithmen und Verfahren, welche sich dem Gebiet der “Explainable Artificial Intelligence” zuordnen.
Neben vielen anderen Anwendungsmöglichkeiten, bietet sich daraus eine interdisziplinäre Kooperation auf biologischen und medizinischen Daten an, um bspw. anhand des Bildes der Retina eines Menschen Diabetes-assoziierte Krankheiten zu erkennen [2], oder die Areale eines Röntgenbildes zu lokalisieren, in denen Kariesläsionen vorhanden sind [3].

Projekt und Ziele:

Die AG Metzger am Max-Delbrück-Centrum für Molekulare Medizin steht in Kooperation mit der AG Prigione des Uniklinikum Düsseldorf, welche die Leitung des Projektes “CureMILS” koordiniert. Dabei geht es um die Untersuchung von maternal vererbten, mitochondrialen Krankheiten, insbesondere des Leigh-Syndroms.
Sie haben eine Reihe an *in vitro* Modellen aus Stammzellen gezüchtet, um zu untersuchen, wie genau sich diese Krankheit verhält und ob es einen Wirkstoff gibt, der die Symptome des Leigh-Syndroms reduzieren, und das mitochondriale Membranpotential ausreichend beeinflussen kann.
Dafür sollen im späteren Verlauf des Projekts mehrere tausend Medikamente an den Zellmodellen getestet und deren Wirkung analysiert werden. Für die Unterstützung der Analyse sollen auch neuronale Netze zum Einsatz kommen, welche bspw. die Zellkulturen als “gesund” und “erkrankt” klassifizieren sollen.
Neben dem Finden eines wirksamen Medikamentes, bleibt weiterhin die Frage offen, welche Unterschiede in den Zellkulturen erkennbar sind.
Ziel der Bachelorarbeit ist es, herauszufinden, ob die Unterschiede, die von der KI für die Entscheidung bei der Klassifizierung verwendet werden, auch aufzeigbar und erklärbar sind.

Damit wollen wir Biologen dabei unterstützen, Phänotypen mittels Techniken des Machine Learnings in der Analyse der Zellmodelle identifizieren zu können.

Problemstellung:

Viele der Ansätze zur Erklärung der Zugehörigkeit einer Klasse eines Bildes basieren auf visuellen Erklärungen, den sogenannten “Saliency Maps” [4] oder “Heat Maps”. Dabei wird der Fokus auf die für die Klassifizierung ausschlaggebenden Bildbereiche gelegt und der Einfluss jedes Areals oder Pixels visuell dargestellt.
Eine andere Methode ist das Extrahieren der wichtigen Merkmale jedes Bildes, anhand derer die Klassenzugehörigkeit vorhergesagt werden kann.
Diese Algorithmen und neuronale Netzwerke sind jedoch häufig auf alltägliche Probleme und Datensätze zugeschnitten (z.B. ImageNet, MNIST, CelebA, …).

Beispielsweise ist die Interpretation des Informationsgehalts der ausschlaggebenden Areale bei solchen Objekten oftmals recht Eindeutig.

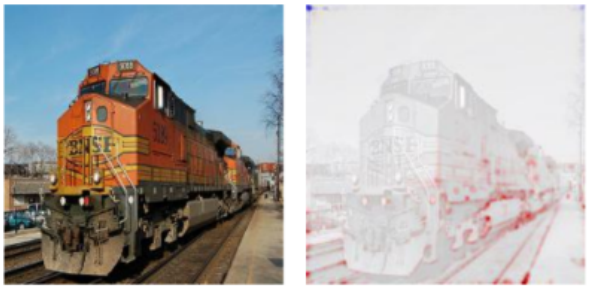


Figure 2: Example of a saliency map explanation of a True Positive (TP) image for the label *“train”*. It highlights the contours of the lines below the train. A possible interpretation is that the CNN has learned to recognise trains when rails are present.

(Abb. 1 - [5])

Ebenso lässt sich auch darauf schliessen, weshalb ggf. eine falsche Klassifizierung stattfand.

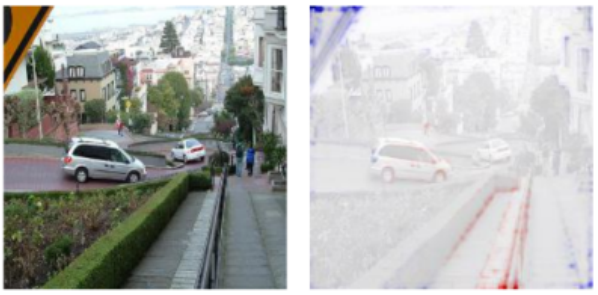
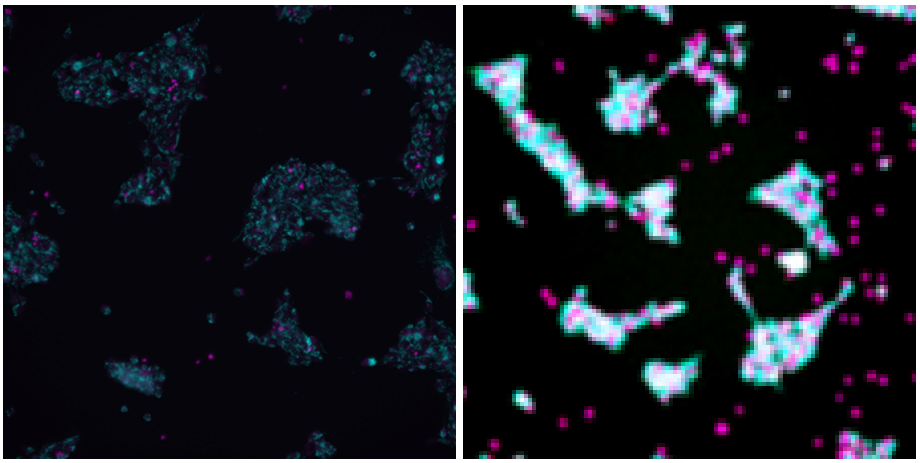


Figure 3: Example of a saliency map explanation of a False Positive (FP) image for the label *“train”*. A possible interpretation is that edges in the lower part appeared similar to rails, which could explain this error.

(Abb. 2 - [5])

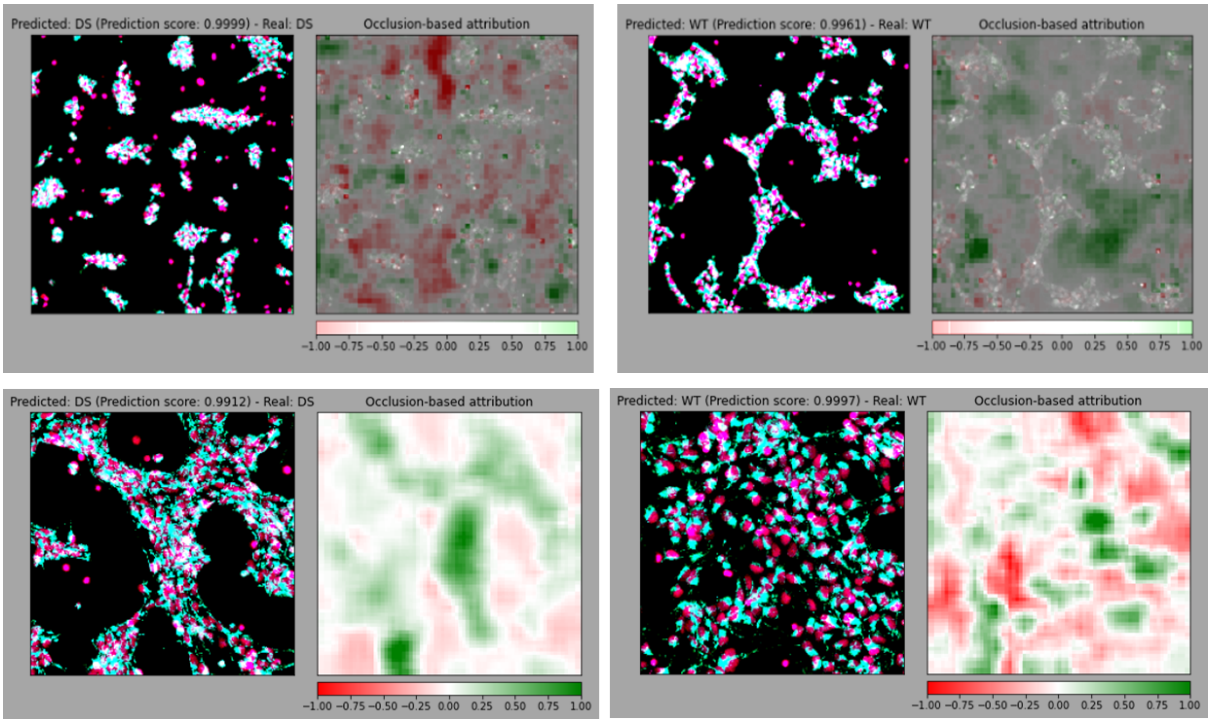
Die Verwendung von Netzwerken, welche zuvor bereits auf sehr großen Datenmengen trainiert wurden und daher bereits eine gewisse Generalisierung unterschiedlicher Strukturen mit sich bringen, ermöglicht es, einen deutlich kleineren Datensatz ähnlicher Problematik anwendbar zu machen.

Bei der Handhabung biologischer Daten gibt es jedoch mehrere Probleme zugleich.
Eines davon ist, dass oftmals nur sehr geringe Datenmengen verfügbar sind, auf die man ein neuronales Netz trainieren kann, da diese sehr spezifisch und zeitaufwendig generiert sind.
Ein anderes gravierendes Problem ist, dass die Daten oftmals weder eindeutig, noch für das menschliche Auge gut unterscheidbar sind, was die Interpretation der Daten zusätzlich erschwert. Ebenso befindet sich das Objekt nicht im Zentrum des Bildes, sondern füllt es mit seinen nicht ganz klar von einander separierbaren Merkmalen gänzlich aus.
Vortranierten Netze bieten somit auf Grund der neu erforderlichen Spezialisierung auf diesen Datensatz lediglich einen kleinen Vorteil.
In diesem Fall handelt es sich um induzierte pluripotente Stammzellen (iPSCs) und zwei fluoreszierenden Markern (DAPI, TMRM), für die eine Klassifizierung zwischen gesunden Kulturen verschiedener Zelllinien, und am Leigh-Syndrom erkrankter Zelllinien mit einer Genauigkeit von über 97% erfolgte.



(Abb. 3 - Rohdaten eines Ausschnitts einer gesunden Zellkultur (l.), und einer auf 224x224 Pixel herunterskalierten, normalisierten und erkrankten Zellkultur (r.))

Trotz guter Ergebnisse in der Klassifikation des relativ kleinen Datensatzes durch die KI, ist die Interpretation mittels “Occlusion-based Attribution” und “Integrated Gradients” schwierig. Ein eindeutiger, allgemeiner Unterschied zwischen den beiden Klassen ist nur schwer erkennbar und erfordert weitere Untersuchungen oder andere Methoden, um den Interessenbereich feiner einzugrenzen (bspw. durch IBA [6]) und dann einen statistischen Unterschied feststellen zu können.



(Abb. 4 - “Occlusion-based Attribution”-Beispiele unserer Daten;. Links: Leigh-Syndrome, rechts: Wildtyp,)

Methoden:

Für die weitere Analyse des Datensatzes, welcher drei Kategorien enthält (Wildtyp, behandelt, und erkrankt), sind die bisher angewandten Methoden nicht ausreichend, um direkt innerhalb eines Bildes eine eindeutige Fläche bestimmen zu können, die für die Entscheidung über die jeweilige Kategorie ausschlaggebend ist.
Daher ist es für uns interessant, die kleinstmögliche Änderung an einem solchen Bild zu finden, die für einen Wechsel der zuvor zugeordneten Klasse sorgt.
Da in einem kurzen Experiment während meines Praktikums festgestellt wurde, dass simple Architekturen von klassischen Autoencodern für diesen Anwendungsfall nicht ausreichend sind und die Ergebnisse zu viele Artefakte enthalten, müssen weitere Architekturen [7, 8, 9, 10] und Methoden evaluiert werden.
Dafür machen wir uns invertierbare neuronale Netze [8] zunutze, die nicht nur in der Lage sind, zu klassifizieren, sondern anschließend auch wieder Bilder zu rekonstruieren.
Um die Änderung zwischen den Klassen aufzuzeigen, werden sogenannte “Counterfactuals” [9] generiert. Mit diesen wird evaluiert, ob es eine komprimierte Form (sog. “Latent Space”) gibt, welche alle wichtigen Merkmale enthält, um daraus wieder ein Bild einer solchen Zellkultur zu rekonstruieren. Dieser Latent Space kann dann in der Theorie so verändert, bzw. verschoben werden, dass ein neues Bild konstruiert wird, welches noch immer dem Original ähnlich sieht, jedoch bei einer Klassifikation durch eine KI in die jeweils andere Klasse eingeordnet wird, da die Hauptmerkmale des Bildes sich verändert haben.
Die Erwartung ist, dass durch eine schrittweise Traversierung eines Bildes zur anderen Klasse die Unterschiede zwischen den gesunden und kranken Zellkulturen sichtbar werden. Ziel ist es, solch eine Verschiebung über versch. Architekturen zu evaluieren und eine aussagekräftige Visualisierung leicht zugänglich zu machen. Somit sollte der Fokus auf wenige Merkmale eingeschränkt werden können.

Für eine Klassifizierung des Wirkungsgrades eines Medikamentes kann ein Bild von der zuvor auf gesunde und kranke Kulturen trainierten KI evaluiert werden. Das daraus resultierende Ergebnis, also die Wahrscheinlichkeit über eine Klassenzugehörigkeit, bietet nur begrenzt Einblick in die Effektivität eines Medikaments. Hierfür wäre die tatsächliche Distanz des komprimierten Bildes und seiner Merkmale im Vergleich zur Kontrollgruppe, sowie eine Visualisierung der Unterschiede innerhalb eines Bildes von ebenso großem Interesse. Mittels der Counterfactuals für generative Netze und Klassifikatoren sollte eine möglichst effiziente und robuste Architektur trainierbar sein, welche für diese beiden Anwendungszwecke geeignet ist.

Quellen:

[1] S. Xie, R. Girshick, P. Dollár, Z. Tu, und K. He, „Aggregated Residual Transformations for Deep Neural Networks“, *arXiv:1611.05431 [cs]*, Apr. 2017, Zugegriffen: 12. Mai 2022. [Online]. Verfügbar unter: http://arxiv.org/abs/1611.05431

[2] F. Saeed *u. a.*, „Diabetic Retinopathy Screening Using Custom-Designed Convolutional Neural Network“, *arXiv:2110.03877 [cs, eess]*, Okt. 2021, Zugegriffen: 12. Mai 2022. [Online]. Verfügbar unter: http://arxiv.org/abs/2110.03877

[3] Y. Bayraktar und E. Ayan, „Diagnosis of interproximal caries lesions with deep convolutional neural network in digital bitewing radiographs“, *Clin Oral Investig*, Bd. 26, Nr. 1, S. 623–632, Jan. 2022, doi: 10.1007/s00784-021-04040-1.

[4] K. Simonyan, A. Vedaldi, und A. Zisserman, „Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps“, *arXiv:1312.6034 [cs]*, Apr. 2014, Zugegriffen: 12. Mai 2022. [Online]. Verfügbar unter: http://arxiv.org/abs/1312.6034

[5] A. Alqaraawi, M. Schuessler, P. Weiß, E. Costanza, und N. Berthouze, „Evaluating Saliency Map Explanations for Convolutional Neural Networks: A User Study“, S. 10.

[6] K. Schulz, L. Sixt, F. Tombari, und T. Landgraf, „Restricting the Flow: Information Bottlenecks for Attribution“, *arXiv:2001.00396 [cs, stat]*, Mai 2020, Zugegriffen: 12. Mai 2022. [Online]. Verfügbar unter: http://arxiv.org/abs/2001.00396

[7] A. Sauer und A. Geiger, „COUNTERFACTUAL GENERATIVE NETWORKS“, S. 25, 2021.

[8] J. Behrmann, W. Grathwohl, R. T. Q. Chen, D. Duvenaud, und J.-H. Jacobsen, „Invertible Residual Networks“, *arXiv:1811.00995 [cs, stat]*, Mai 2019, Zugegriffen: 12. Mai 2022. [Online]. Verfügbar unter: http://arxiv.org/abs/1811.00995

[9] J. Behrmann, W. Grathwohl, R. T. Q. Chen, D. Duvenaud, und J.-H. Jacobsen, „Invertible Residual Networks“, *arXiv:1811.00995 [cs, stat]*, Mai 2019, Zugegriffen: 12. Mai 2022. [Online]. Verfügbar unter: http://arxiv.org/abs/1811.00995

[10] A. van den Oord, O. Vinyals, und K. Kavukcuoglu, „Neural Discrete Representation Learning“, *arXiv:1711.00937 [cs]*, Mai 2018, Zugegriffen: 12. Mai 2022. [Online]. Verfügbar unter: http://arxiv.org/abs/1711.00937