Maximilian Armin Zenk
Dr. sc. hum.

Robustness of Medical Image Segmentation Algorithms in the Context of Federated Data

Fach/Einrichtung: Deutsches Krebsforschungszentrum (DKFZ)
Doktorvater: Prof. Dr. rer. nat. Klaus Hermann Maier-Hein

Bei der Einführung von auf künstlicher Intelligenz basierten Lösungen in der klinischen Praxis hat die Radiologie eine Vorreiterrolle, da der stetig wachsende Bedarf an bildbasierten Untersuchungen nicht von den verfügbaren Radiologen gedeckt werden kann. Die semantische Segmentierung ist eine zentrale Komponente von Bildanalyse-Pipelines und findet unter anderem Anwendung in der computergestützten Diagnose, der Planung von Strahlentherapien und der Überwachung von Krankheitsverläufen. Heutzutage können Deep Learning-Algorithmen verschiedene anatomische Strukturen automatisch segmentieren, mithilfe von entsprechend annotierten Trainingsdatensätzen. Diese Algorithmen können jedoch auch Fehler machen, insbesondere wenn sie auf Daten angewendet werden, die sich in ihren Eigenschaften von den Trainingsdaten unterscheiden. Die Diskrepanz zwischen den Eigenschaften der Trainings- und Testdaten wird als Distribution Shift bezeichnet und tritt häufig auf, wenn Modelle in neuen Krankenhäusern eingesetzt werden. Für diese Doktorarbeit wurden Benchmarks für Methoden entwickelt, die die Robustheit von Segmentierungsverfahren gegenüber solchen Distribution Shifts verbessern. Dabei wurden zwei komplementäre Ansätze untersucht: Methoden, die die Generalisierung auf Daten mit Distribution Shifts verbessern, sowie Methoden, die erkennen können, wann sie falsche Vorhersagen treffen (Fehlererkennung).
Die Benchmarking-Studie zu Generalisierung erfolgte in dieser Arbeit durch die Organisation eines internationalen Wettbewerbs (auch Challenge genannt). Solche Challenges gelten als Goldstandard in der medizinischen Bildanalyse für den Vergleich von Algorithmen, da sie standardisierte und faire Bedingungen für alle Teilnehmenden bieten. Obwohl jedes Jahr zahlreiche Wettbewerbe organisiert werden, basieren sie in der Regel auf Forschungsdatensätzen, die von wenigen Institutionen und Scannern stammen. Daher ist oft unklar, wie gut die Algorithmen auf multizentrische Daten mit größerer Diversität und realistischen Distribution Shifts generalisieren. Diese Arbeit führt das Konzept ein, föderierte Daten in Challenge-Settings zu nutzen. Solche Daten verlassen die Institution, in der sie erhoben wurden, nicht, was die Hürden für die Bereitstellung von Daten erheblich senkt. Für eine föderierte Evaluation werden die Segmentierungsalgorithmen an die Institutionen im Verbund geschickt, und deren Evaluierungsergebnisse werden zurückgemeldet, um die Robustheit der Modelle zu analysieren. Dieses Konzept wird in einer Challenge zur Segmentierung von Hirntumoren umgesetzt—der Federated Tumor Segmentation (FeTS) Challenge. Als erste ihrer Art offenbart und adressiert die FeTS Challenge einige praktische Herausforderungen der föderierten Evaluation, insbesondere den hohen organisatorischen Aufwand, die erschwerte Qualitätskontrolle von Annotationen im Vergleich zu konventionellen Challenges und die eingeschränkte Analysemöglichkeit aufgrund des fehlenden direkten Zugriffs auf die föderierten Daten. Gleichzeitig zeigt die Challenge aber auch das Potenzial föderierter Benchmarks, die Größe und Vielfalt der Testdatensätze erheblich zu steigern. Dies wird durch die FeTS Challenge exemplarisch demonstriert, bei der 32 internationale Institutionen insgesamt 2625 Fälle mit multiparametrischen Magnetresonanztomographie (MRT)-Scans beisteuerten. Die Evaluierung der 41 in

der Challenge eingereichten Segmentierungsmodelle auf diesen Testdaten zeigte, dass die Modelle im Durchschnitt gut generalisierten, aber auf Daten von 13 der 32 beteiligten Institutionen in Einzelfällen Fehler machten, die auf einen Mangel an Robustheit hinweisen.

Die Fehlererkennung ist für die Zuverlässigkeit von Segmentierungsmethoden in der Praxis von großer Bedeutung und wurde aus vielen Perspektiven untersucht, darunter Unsicherheitsabschätzung, Out-of-Distribution-Erkennung und Schätzung der Segmentierungsqualität. Der Fortschritt in diesem Forschungsbereich wird derzeit durch zwei Probleme behindert: Erstens unterscheiden sich die Evaluationsprotokolle der verschiedenen Ansätze, was einen direkten Vergleich der Methoden zur Fehlererkennung erschwert. Zweitens wurden neue Methoden bisher oft nur für ein Segmentierungsproblem (z. B. in einer anatomischen Region) getestet oder nicht hinsichtlich Distribution Shifts evaluiert, sodass ihre Anwendbarkeit auf ein breiteres Aufgabenspektrum unklar bleibt. Der zweite Teil dieser Arbeit adressiert diese Defizite durch die Entwicklung eines Evaluationsprotokolls basierend auf einer Risk-Coverage Analyse, welches den Vergleich aller relevanten Methoden der Fehlererkennung ermöglicht und Schwachstellen bisheriger Ansätze vermeidet. Ein Benchmark wurde entwickelt, der diese Evaluationsstrategie implementiert und verschiedene, diverse Methoden zur Fehlererkennung in Experimenten mit mehreren öffentlichen Datensätzen vergleicht, die realistische Distribution Shifts enthalten. Die Ergebnisse dieser Studie lieferten Erkenntnisse darüber, wie Unsicherheitswerte auf Pixel-Ebene effektiv zu einem Unsicherheitswert auf Bild-Ebene für die Fehlererkennung aggregiert werden können. Zudem wurde eine existierende, einfache Methode als starke Referenz für zukünftige Forschung identifiziert, da sie über mehrere Datensätze hinweg konsistent leistungsfähiger als komplexere Algorithmen war. Dank ihrer Flexibilität und Effizienz kann diese Methode leicht an neue Segmentierungsprobleme und praktische Anwendungen angepasst werden.

Zusammenfassend führte diese Dissertation groß angelegte Benchmarking-Studien durch, die modernste Generalisierungs- und Fehlererkennungsalgorithmen in realitätsnahen Szenarien testen. Die Experimente demonstrieren, wie multizentrische Daten sowohl in zentralisierter als auch in föderierter Form genutzt werden können, um die Robustheit gegenüber Distribution Shifts zu evaluieren. Dabei wurden häufige Fehlerquellen aufgedeckt und praxistaugliche Algorithmen identifiziert, die eine gute Generalisierung auf neue Krankenhäuser ermöglichen und außerdem signalisieren können, wenn Segmentierungen potenziell fehlerhaft sind. Der Code für beide Benchmarks wird der wissenschaftlichen Gemeinschaft zur Verfügung gestellt, um eine fundierte Vergleichbarkeit von Methoden zu ermöglichen und den Fortschritt in der robusten medizinischen Bildsegmentierung voranzutreiben.