# 8 XAI: Explainable AI

Systemtechnik BSc
FS 2025

## Aufgaben



### Lernziele

Nach dem Bearbeiten dieser Übungsserie ...

- kennen und verstehen Sie die Architektur des ResNet50 Modells.
  Aufgaben: 1
- können Sie sowohl die Aktivierungen ausgewählter Faltungsschichten auf bestimmte Inputs visualisieren, oder umgekehrt den Input eines Faltungsnetzwerkes so mittels Gradientenanstiegsverfahren verändern, dass die Aktivierungen einer von Ihnen gewählten Schicht maximiert wird.
  Aufgaben: 1
- Wissen Sie, dass Sie die hierarchisch-modulare Zerlegung des visuellen Raums, die ein Faltungsnetz erlernt, nicht dem entspricht, was der menschliche visuelle Kortex tut.
  Aufgaben: 1

### Explainable Artificial Intelligence (XAI)

Oft heisst es, Deep-Learning-Modelle seien »Blackboxes«, die Repräsentationen in einer Form erlernen, die schwer zu extrahieren und für Menschen kaum verständlich sind. Für einige Arten von Deep-Learning-Modellen mag das teilweise stimmen, auf CNNs trifft es jedoch definitiv nicht zu. Die von CNNs erlernten Repräsentationen sind in hohem Mass für Visualisierungen geeignet, und zwar vor allem deshalb, weil es sich um

Repräsentationen visueller Konzepte handelt. Seit 2013 wurde eine Vielzahl von Verfahren zur Visualisierung und Interpretation dieser Repräsentationen entwickelt. Adam Harley, Masterstudent an der Ryerson University, hat eine interaktive Visualisierung erstellt, die erklärt, wie ein neuronales Faltungsnetz, eine Art Programm für künstliche Intelligenz, das zur Analyse von Bildern verwendet wird, intern funktioniert.
In dieser Übungsserie geht es um die Darstellung der visuellen Muster, auf die verschiedene Filter reagieren sollen. Dazu dient das *Gradientenanstiegsverfahren* im Eingaberaum: Wenden Sie das Verfahren, ausgehend von einem leeren Bild, auf die Eingabewerte des CNNs an, um die Reaktion eines bestimmten Filters zu *maximieren*. Auf das resultierende Bild wird der ausgewählte Filter maximal reagieren. Das Verfahren ist unkompliziert: Sie erstellen eine *Verlustfunktion*, die den Wert eines bestimmten Filters in einem bestimmten 'Convolutional' Layer maximiert, und verwenden anschliessend das stochastische Gradientenabstiegsverfahren (stochastic gradient descent), um die Werte des Eingabebilds so anzupassen, dass die Aktivierung maximal wird.

- Die Idee besteht darin, ein bereits trainiertes Modell wiederzuverwenden, in diesem Fall das ResNet50-Modell, das aus mehreren Faltungsschichten (eigentlich Blöcken aus mehreren Faltungsschichten) besteht, gefolgt von einigen voll verknüpften / dichten Schichten und einer Softmax-Ausgabeschicht für die Klassifizierung. Die Softmax-Ausgabe verwenden wir nicht.
- Wir werden nur ein Teilmodell als Feature-Extraktor-Modell bis zu der Schicht verwenden, deren Aktivierungen wir visualisieren möchten, oder deren Aktivierungen wir durch einen bestimmten Input maximieren möchten.
- Keras stellt ein bereits auf dem ImageNet-Datensatz vortrainiertes `ResNet50V2` zur Verfügung.

## Aufgabe 1. Explainable AI (XAI)

(a) Laden des ResNet50V2-Modells (ohne Dense-Top-Schicht): Öffne das Notebook `XAI_ResNet50V2_TEMPLATE.ipynb`. Lade das vortrainierte ResNet50-Modell, ohne die dichte Ausgabeschicht. Verwende dafür `torchvision.models.resnet50` und setze `model.fc = nn.Identity()`.

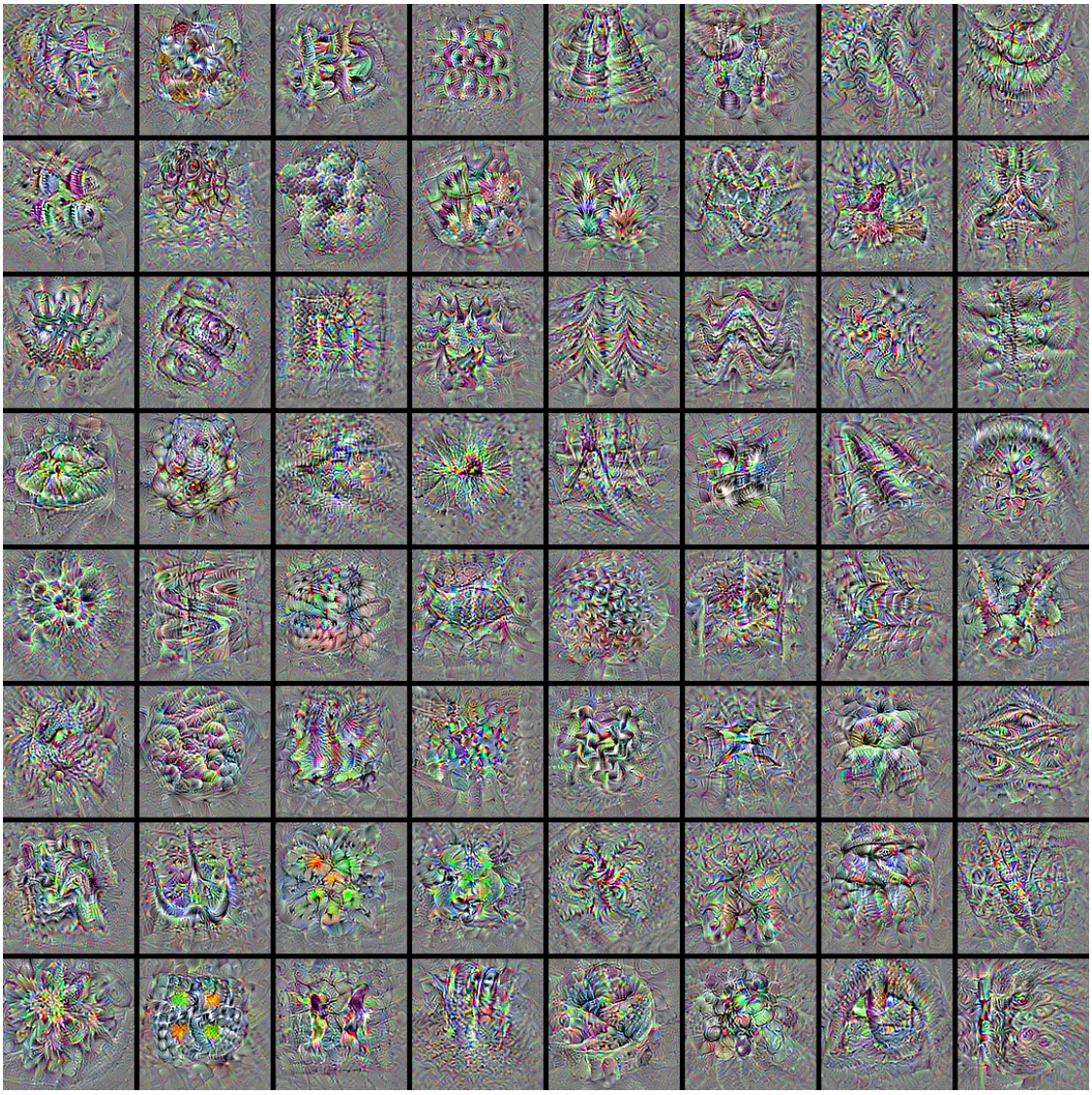

Abbildung 1:

(b) Modell zur Merkmalsextraktion aufbauen: Baue ein LightningModule, das beim Forward-Pass die Aktivierungen einer bestimmten Zielschicht extrahiert.

(c) Einrichten des Gradientenanstiegsprozesses: Implementiere eine Funktion, die die mittlere Aktivierung eines bestimmten Filters aus der Zielschicht berechnet – ohne Randpixel.

(d) End-to-End-Filtervisualisierungsschleife aufbauen: Visualisiere, welche Muster einen bestimmten Filter aktivieren. Dazu:

- Initialisiere ein zufälliges Bild im Bereich [-0.125, +0.125]
- Wende 30 Schritte Gradientenanstieg an
- Skaliere das Ergebnis zurück in den Bereich [0, 255] zur Darstellung

(e) Visualisierung der ersten 64 Filter in der Zielebene

(f) Interpretation der Aktivierungen: Was erkennen die CNN-Filter? Wie verändern sich die erkannten Muster in tieferen Schichten? Was kann man über das gelernte Repräsentationsvermögen des Netzwerks sagen?

(g) Spielen sie selber mit dem Netzwerk und den Parametern, damit Sie die CNNs besser verstehen.

## Epilog: Where's the intelligence?

Was «verstehen» sie also wirklich, die CNNs? Zwei Dinge: Erstens verstehen sie eine Zerlegung ihres visuellen Eingaberaums als hierarchisch-modulares Netz von Faltungsfiltern, und zweitens verstehen sie eine wahrscheinlichkeitstheoretische Zuordnung zwischen bestimmten Kombinationen dieser Filter und einem Satz beliebiger Bezeichnungen. Natürlich ist dies kein «Sehen» im menschlichen Sinne, und aus wissenschaftlicher Sicht bedeutet es sicherlich nicht, dass wir das Computersehen an diesem Punkt irgendwie gelöst haben. Glauben Sie nicht an den Hype; wir stehen lediglich auf der ersten Stufe einer sehr hohen Leiter.
Manche sagen, dass die hierarchisch-modulare Zerlegung des visuellen Raums, die ein Faltungsnetz erlernt, dem entspricht, was der menschliche visuelle Kortex tut. Das mag stimmen oder auch nicht, aber es gibt keine stichhaltigen Beweise für diese Annahme. Natürlich würde man erwarten, dass der visuelle Kortex etwas Ähnliches lernt, insofern dies eine «natürliche» Zerlegung unserer visuellen Welt darstellt (ähnlich wie die Fourier-Zerlegung eine «natürliche» Zerlegung eines periodischen Audiosignals wäre). Aber die genaue Art der Filter und der Hierarchie sowie der Prozess, durch den sie erlernt werden, haben höchstwahrscheinlich wenig mit unseren mickrigen Faltungsnetzen gemein. Der visuelle Kortex ist nicht von vornherein faltbar, und obwohl er in Schichten strukturiert ist, sind die Schichten selbst in kortikale Spalten gegliedert, deren genauer Zweck noch immer nicht ganz verstanden ist - ein Merkmal, das in unseren künstlichen Netzen nicht zu finden ist (obwohl Geoff Hinton daran arbeitet). Ausserdem geht es bei der visuellen Wahrnehmung um so viel mehr als nur um die Klassifizierung statischer Bilder - die menschliche Wahrnehmung ist grundsätzlich sequenziell und aktiv, nicht statisch und passiv, und sie ist eng mit der motorischen Steuerung (z. B. Augensakkaden) verflochten.
Denken Sie daran, wenn Sie das nächste Mal hören, wie ein VC oder ein namhafter CEO in den Nachrichten vor der existenziellen Bedrohung warnt, die von unseren jüngsten Fortschritten beim Deep Learning ausgeht. Wir verfügen heute über bessere Werkzeuge, um komplexe Informationsräume abzubilden, als wir es je zuvor getan haben, und das ist grossartig, aber letztendlich sind es Werkzeuge, keine Lebewesen, und nichts von dem, was sie tun, kann vernünftigerweise als «Denken» bezeichnet werden. Ein Smiley auf einen Stein zu malen, macht ihn nicht «glücklich», auch wenn Ihr Primaten-Neokortex

Ihnen das sagt.
Abgesehen davon ist die Visualisierung dessen, was Convnets lernen, ziemlich faszinierend - wer hätte gedacht, dass ein einfacher Gradientenabstieg mit einer vernünftigen Verlustfunktion über einen ausreichend grossen Datensatz ausreichen würde, um dieses wunderschöne hierarchisch-modulare Netzwerk von Mustern zu lernen, das einen komplexen visuellen Raum überraschend gut erklären kann. Deep Learning ist vielleicht keine Intelligenz im eigentlichen Sinne, aber es funktioniert immer noch wesentlich besser, als man noch vor ein paar Jahren hätte erwarten können. Wenn wir jetzt nur verstehen würden, warum... ;-)
Francois Chollet

## Lösungen

**Lösung 1.**
Sie finden die Lösung auf Moodle unter `ANN_U08_XAI_ResNet50V2_SOLUTION.ipynb`.