Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Methoden der Explainable AI (XAI)

Lukas Scholz | 24.06.2022 | HWR Berlin

Aspekte der XAI

- Verständnis der Modellstruktur
- Verständnis einzelner Komponenten (z.B. Parameter)
- Verständnis des Trainingsalgorithmus
- Analytische Aussagen zu Ergebnissen
- **Visuelle Aussagen zu Ergebnissen**
- Erklärungen anhand Beispielen

} Post-hoc Interpretability

Lipton (2016)

Visuelle Aussagen zu Ergebnissen

Prediction: Doctor

Prediction: Nurse

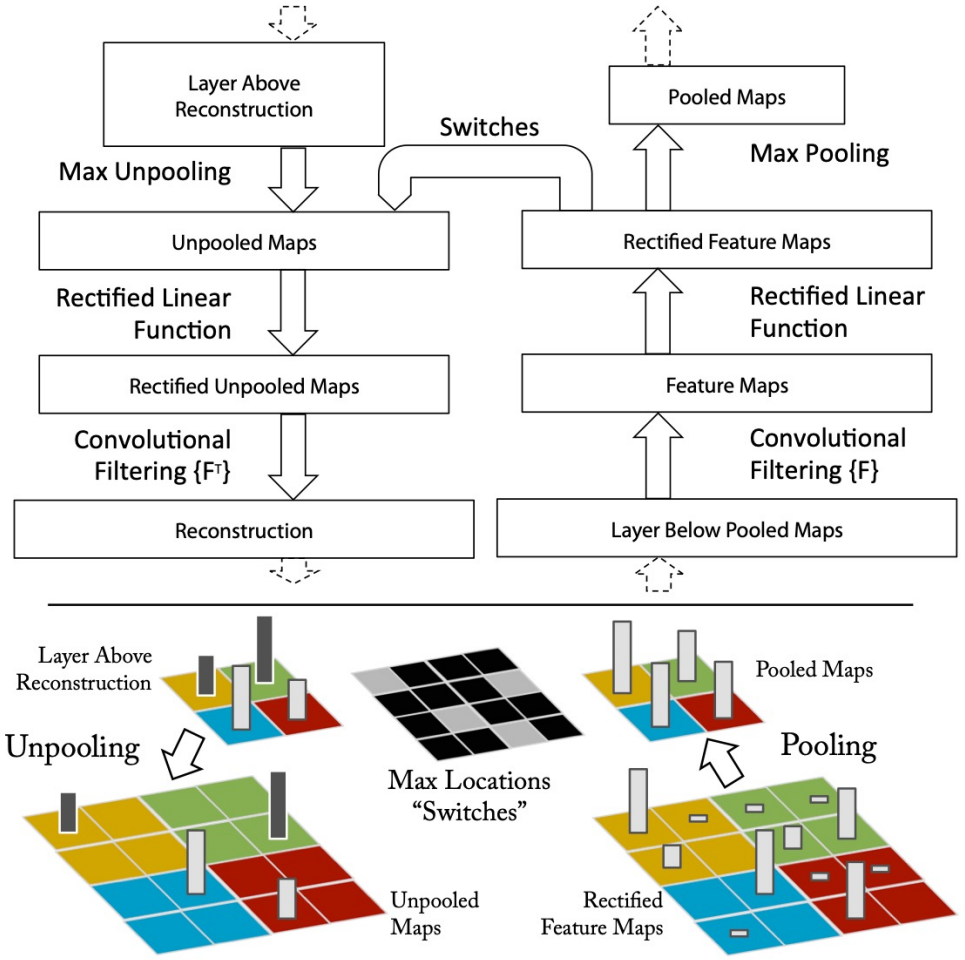
Visuelle Aussagen zu Ergebnissen

Wieso kann das Modell zwischen Doktor:innen und Krankenschwestern /-pflegern unterscheiden?

Saliency Maps

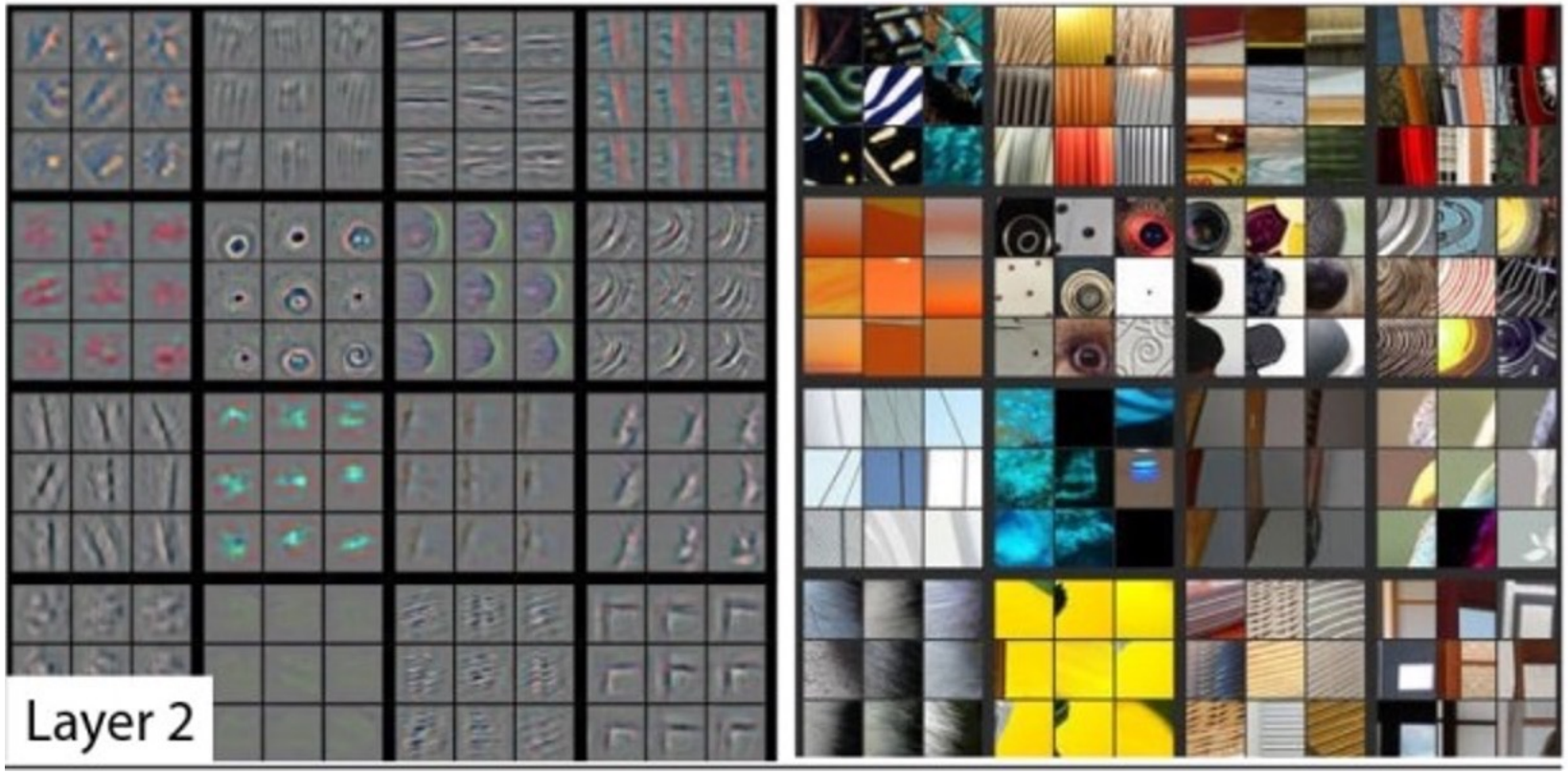
- Heben Regionen hervor die für Modell von Bedeutung sind
- “Einblick” in den Entscheidungsprozess des Modells
- Unterscheiden sich im Informationsgehalt und Aufwand der Implementierung

Saliency Maps - DeconvNet



[Zeiler and Fergus (2013)]

Saliency Maps - DeconvNet



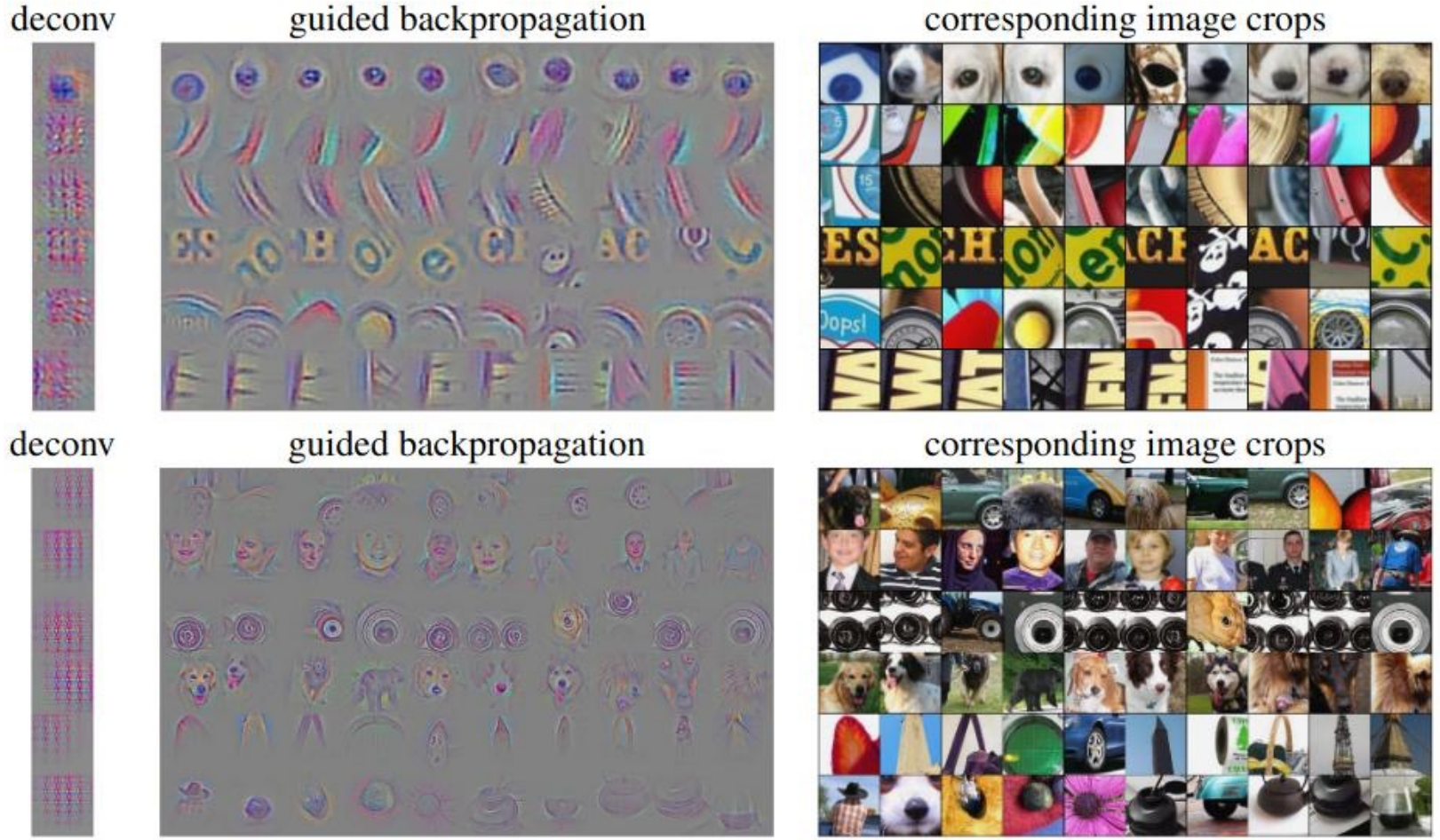
- Geringer Informationsgehalt
- Darstellungen pro Layer des Modells

→ viele einzelne Darstellungen; kein gesamter Eindruck

- Aufwendige Implementierung (DeconvNet als eigenständiges Modell)

Zeiler and Fergus (2013)

Saliency Maps – Guided Backprop



- Höhere Auflösung als reines DeconvNet
- Weiterhin Darstellungen pro Layer
- Aufwendige Implementierung, da DeconvNet verwendet wird

Springenberg et. al. (2014)

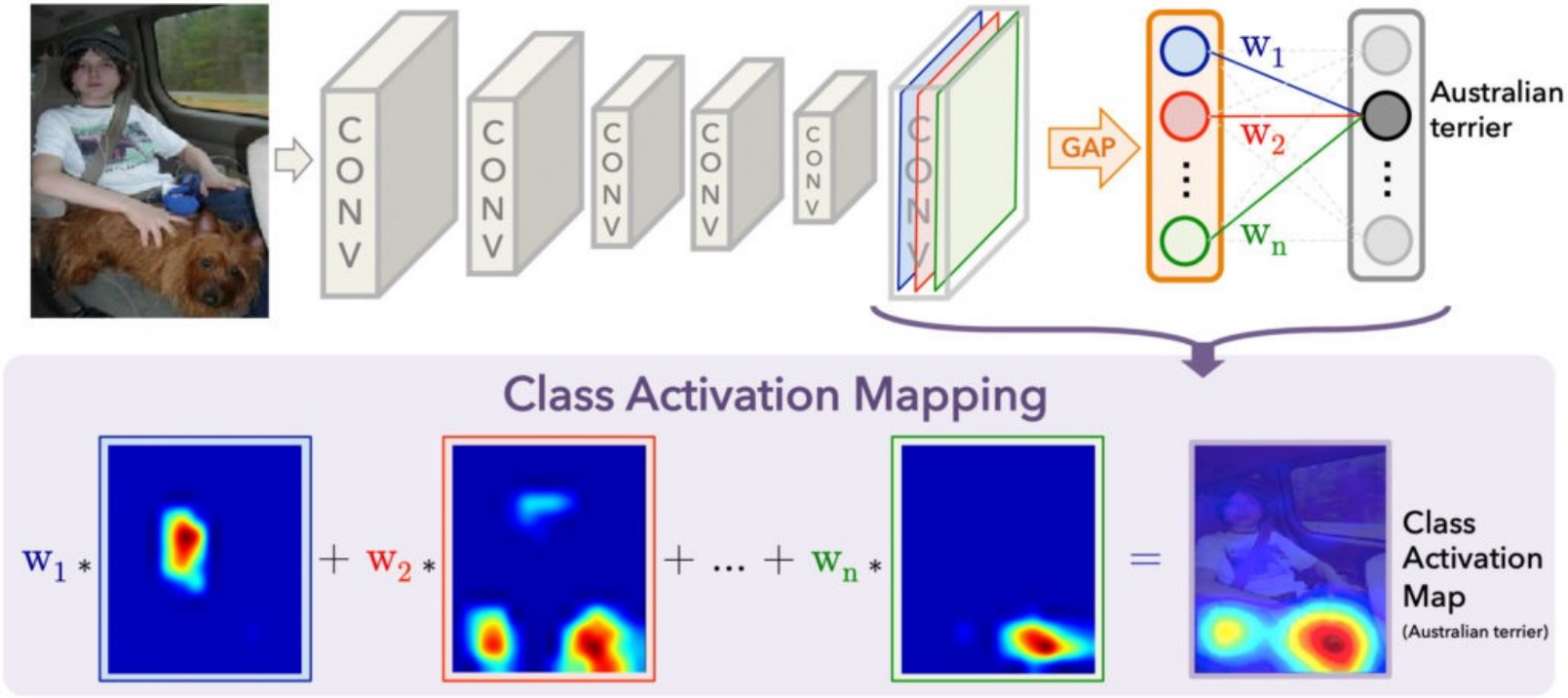
Saliency Maps – Class Activation Map

Zhou et. al. (2015)

- Heatmap bringt Wertung in die Analyse (Rot = Bedeutend; Blau = Unbedeutend)
- Darstellung für alle Layer des Modells
- Einfache Implementation

Alle Features des Modells werden betrachtet!

Class Activation Maps



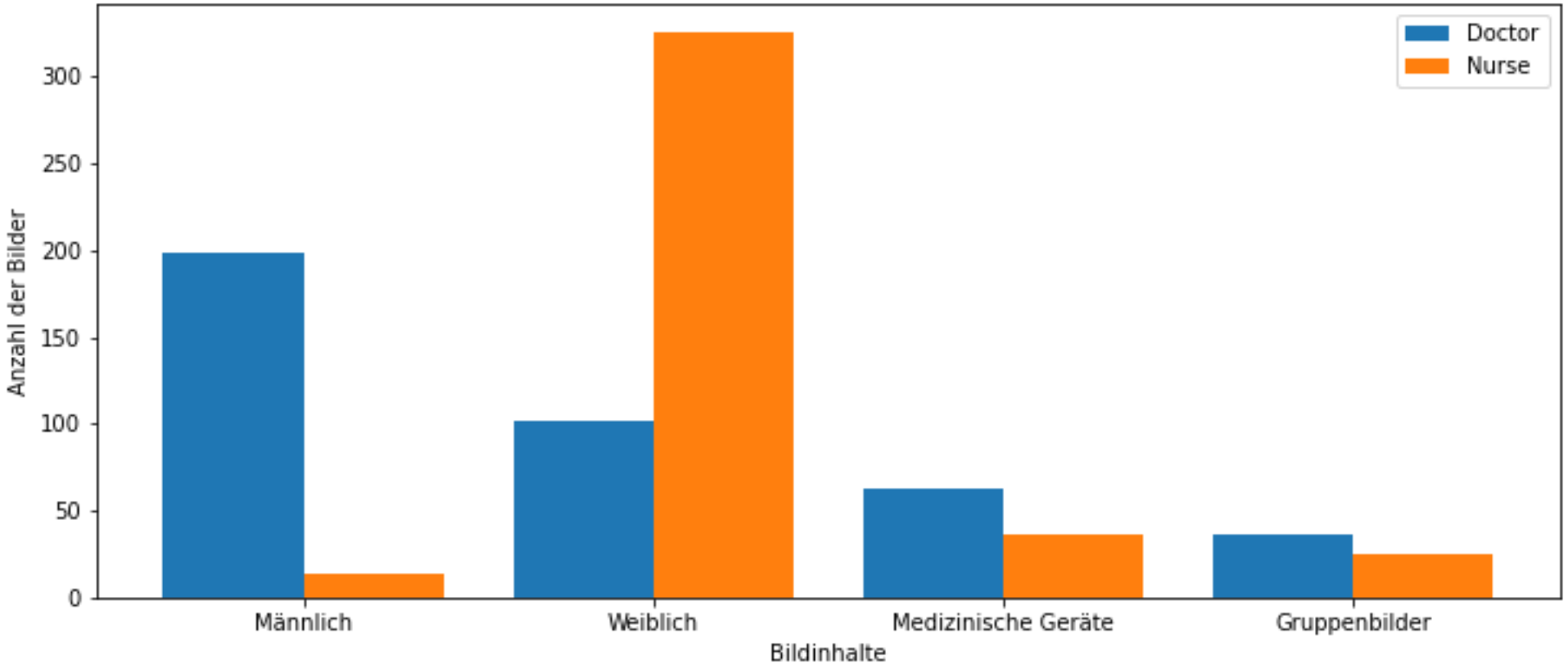
Zhou et. al. (2015)

Class Activation Maps

Category	Image	GradCAM
Dog		
Cat		

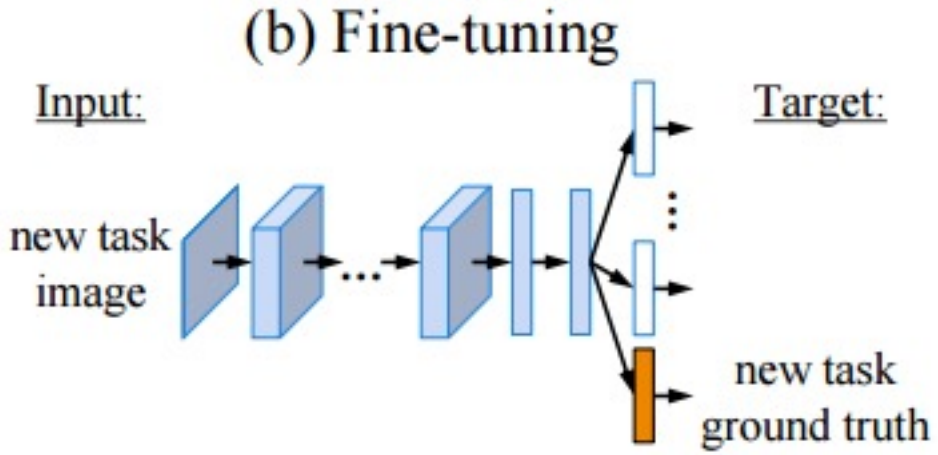
[Jacob Gildenblat and contributors (2021)]

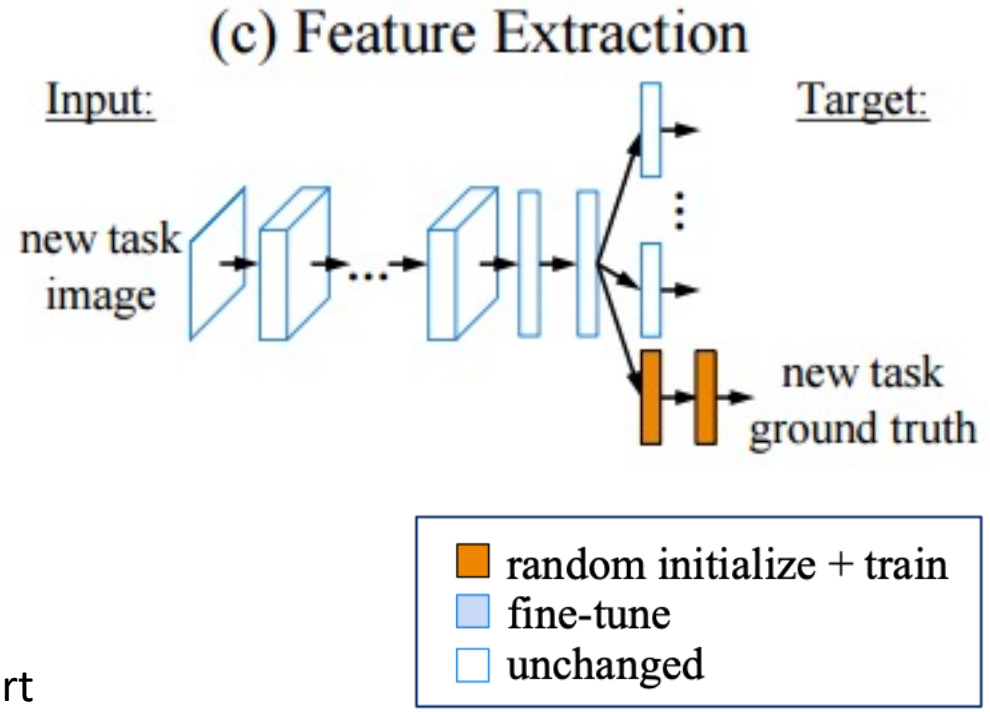
Erkennung von Bias in Datensätzen



- Datensatz enthält starken Geschlechterbias
- Medizinische Geräte = Stethoskope, etc.
- Gruppenbilder = Bilder mit mehreren Geschlechtern

Trainingsmethoden





Li and Hoiem (2016)

- Fine-tuning: Modell wird auf Datensatz neu trainiert
- Feature Extraction: Modell übernimmt vorerlernte Features und erweitert diese

Modelle mit Bias

a) Feature Extraction mit Bias - Doctor

b) Fine-tuning mit Bias - Doctor

c) Feature Extraction mit Bias - Nurse

d) Fine-tuning mit Bias - Nurse

- Feature Extraction hat die unterschiedlichen Arbeitsuniformen erkannt

→ Doktor:innen langärmlig; Krankenschwester/-pfleger kurzärmlig

- Fine-tuning unterliegt dem Bias innerhalb des Datensatzes

→ Klassifiziert anhand geschlechterspezifischen Merkmalen

Modelle ohne Bias

a) Feature Extraction ohne Bias - Doctor

b) Fine-tuning ohne Bias - Doctor

c) Feature Extraction ohne Bias - Nurse

d) Fine-tuning ohne Bias - Nurse

- Feature Extraction klassifiziert weiterhin anhand der Arbeitsuniform

- Fine-tuning klassifiziert nun auch anhand der Arbeitsuniform

→ Bias wurde durch ausgleichen der Klassen bereinigt

Vielen Dank für Ihre Aufmerksamkeit!

Lukas Scholz | 24.06.2022 | HWR Berlin