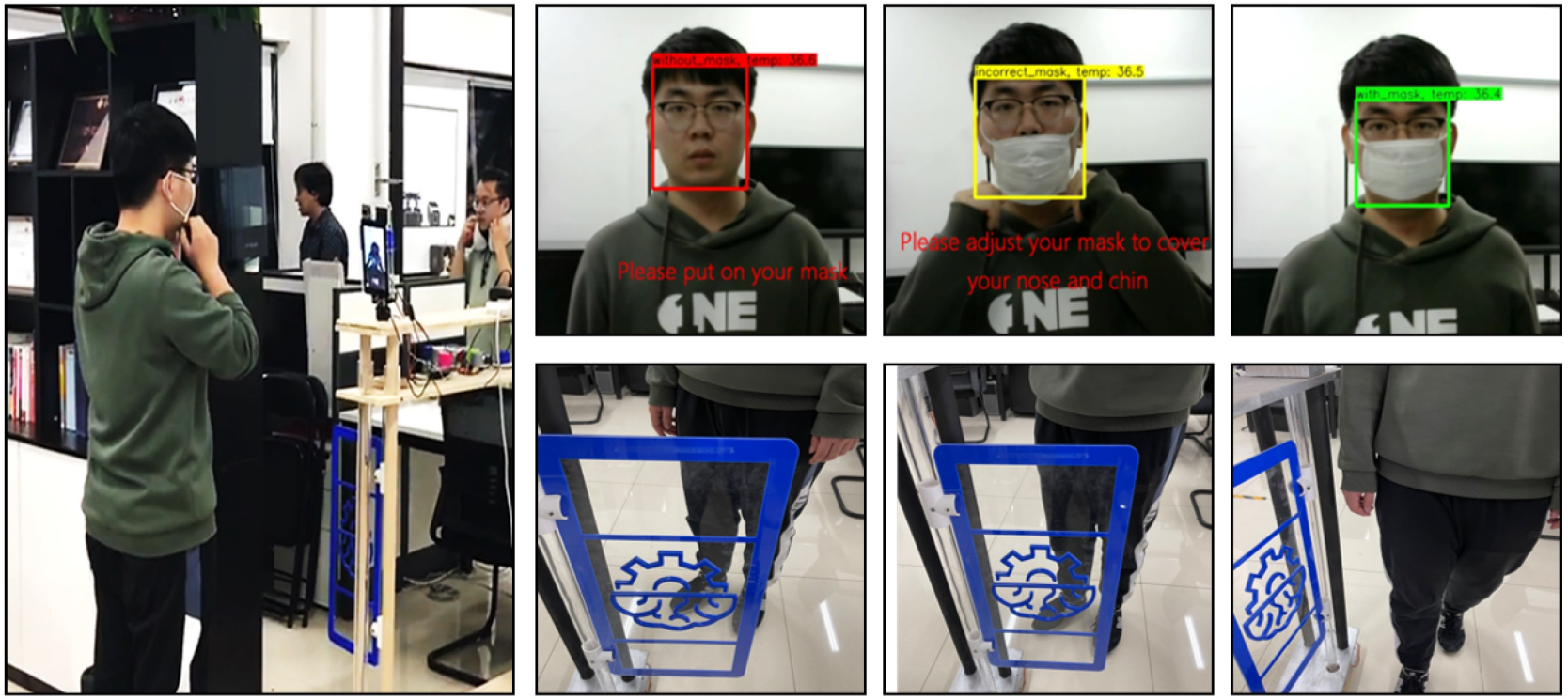
Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Explainable AI – Analyse und Realisierung zur Erklärbarkeit von Computer-Vision-Modellen

Relevanz



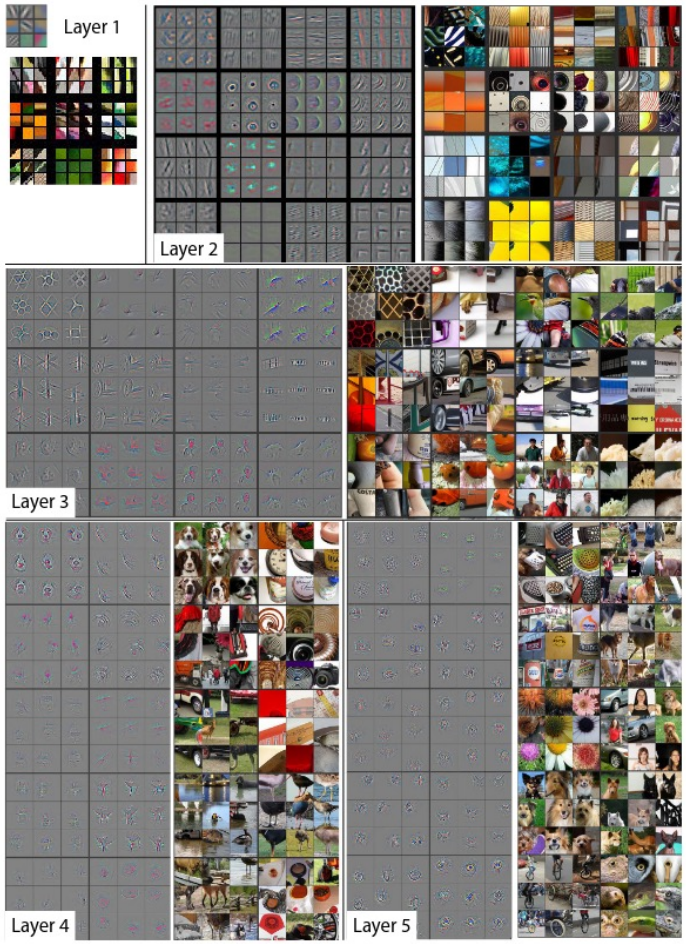
(a) (b) (c) (d)

Quelle: Jiang et al. (2021)

Analyse von XAI-Methoden

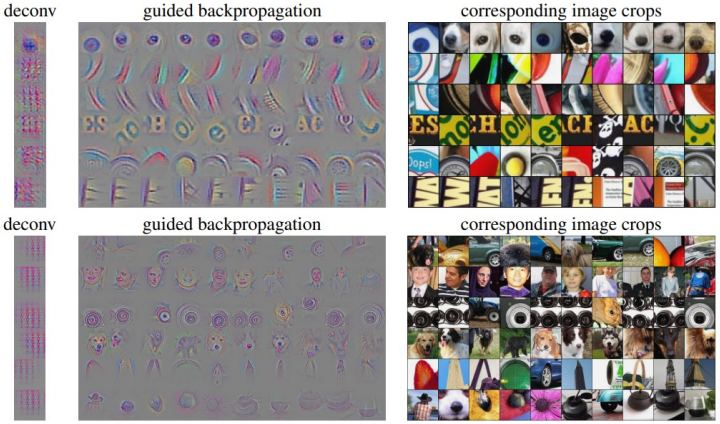
Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Deconvolutional Networks



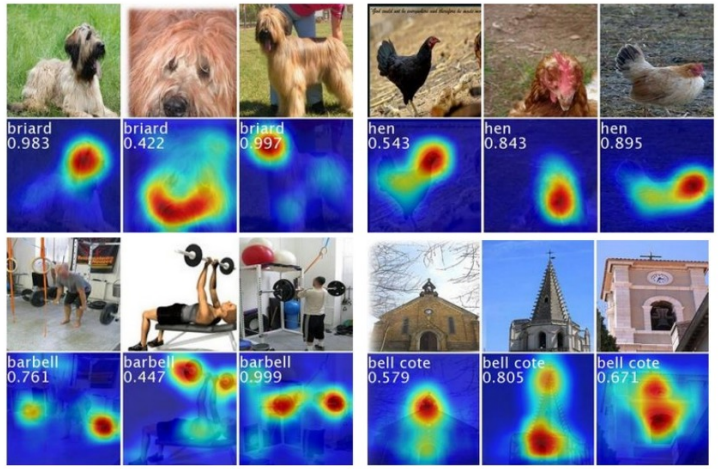
Quelle: Zeiler and Fengus (2013)

Guided Backpropagation



Quelle: Springenberg et al. (2014)

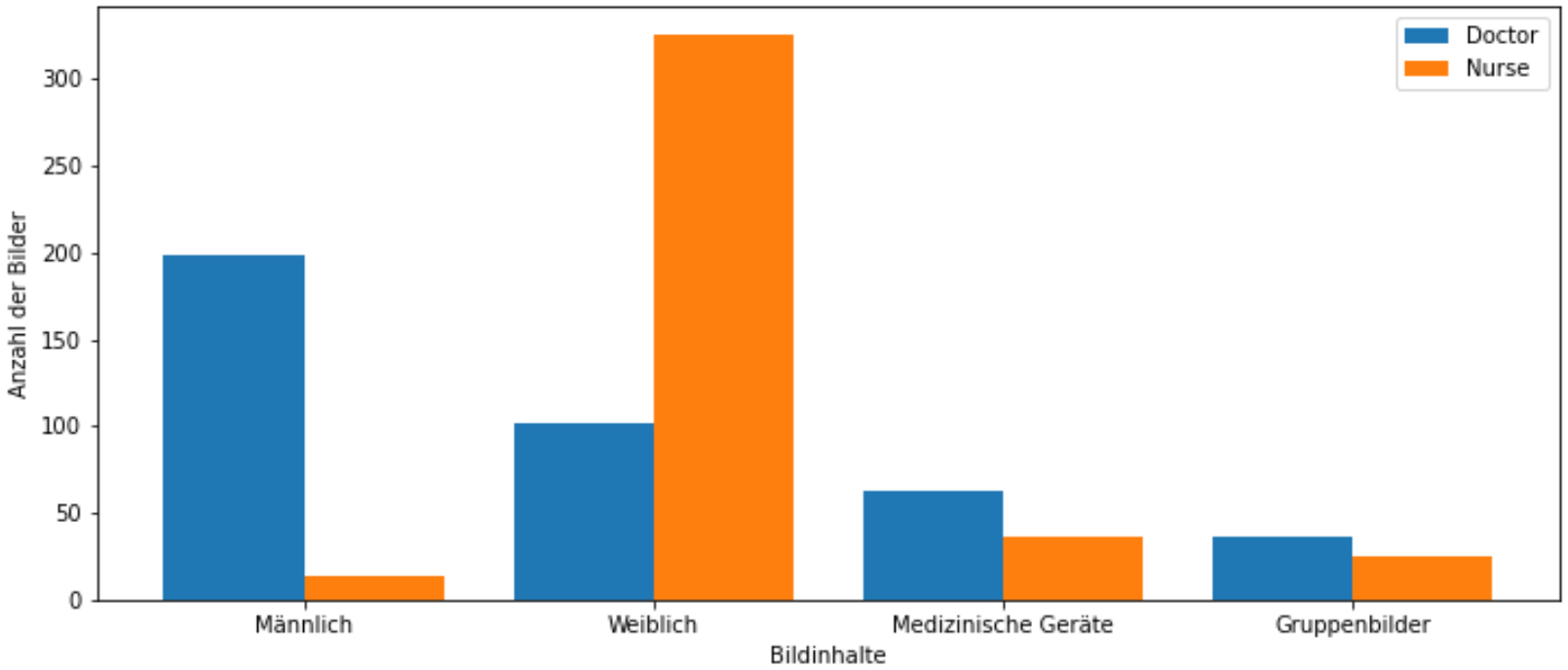
Class Activation Maps



Quelle: Zhou et al. (2015)

Datensatz mit Geschlechter Bias

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



Ergebnisse-CNN

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

a) Feature Extraction mit Bias - Doctor

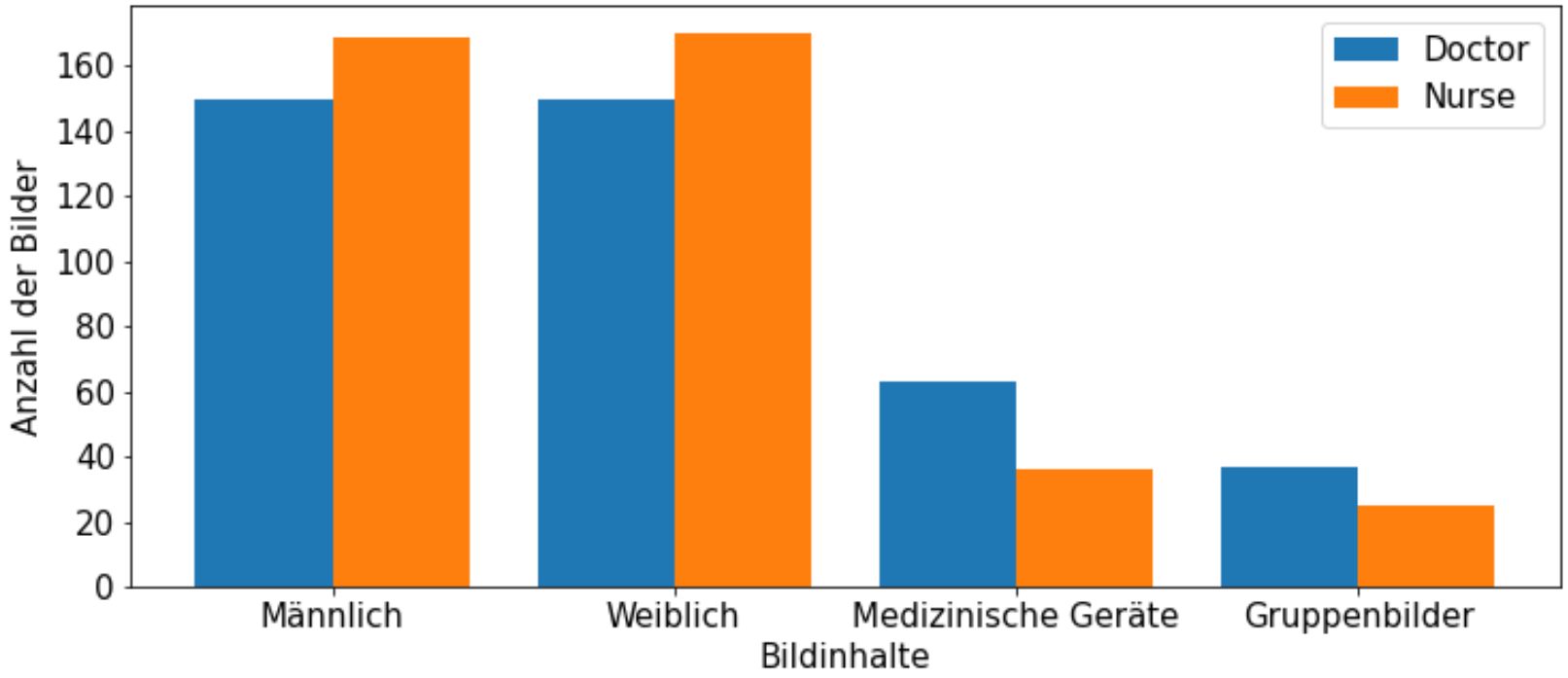
b) Fine-tuning mit Bias - Doctor

c) Feature Extraction mit Bias - Nurse

d) Fine-tuning mit Bias - Nurse

Datensatz ohne Geschlechter Bias

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



Ergebnisse-CNN

a) Feature Extraction ohne Bias - Doctor

b) Fine-tuning ohne Bias - Doctor

c) Feature Extraction ohne Bias - Nurse

d) Fine-tuning ohne Bias - Nurse

Fazit

- CAM helfen bei der Nachvollziehbarkeit
- Unterschiede nicht nur in Modellarchitektur, sondern auch Trainingsmethoden
- Nachvollziehbarkeit unterscheidet sich bei gleicher Accuracy
- Analyse von qualitativen Merkmalen ist wichtig