Robustheit von Modellen beim Maschinellen Lernen

Generalisierbarkeit jenseits der Testdaten

INNOVET
KI B³
Künstliche Intelligenz in die berufliche Bildung bringen

GEFÖRDERT VOM
INNOVET
Bundesministerium für Bildung und Forschung
bibb Bundesinstitut für Berufsbildung

Gefördert als InnoVET-Projekt aus Mitteln des Bundesministeriums für Bildung und Forschung.

Stichwort „Robustheit“

Overfitting

In gewissem Maß getestet durch Evaluierung auf Test- bzw. Validierungsdaten; Maßnahmen siehe vorige Folien

Overfitting

Durch die Vermeidung von Overfitting entstehen Modelle, die besser generalisieren und damit robuster sind.

Stichwort „Robustheit“

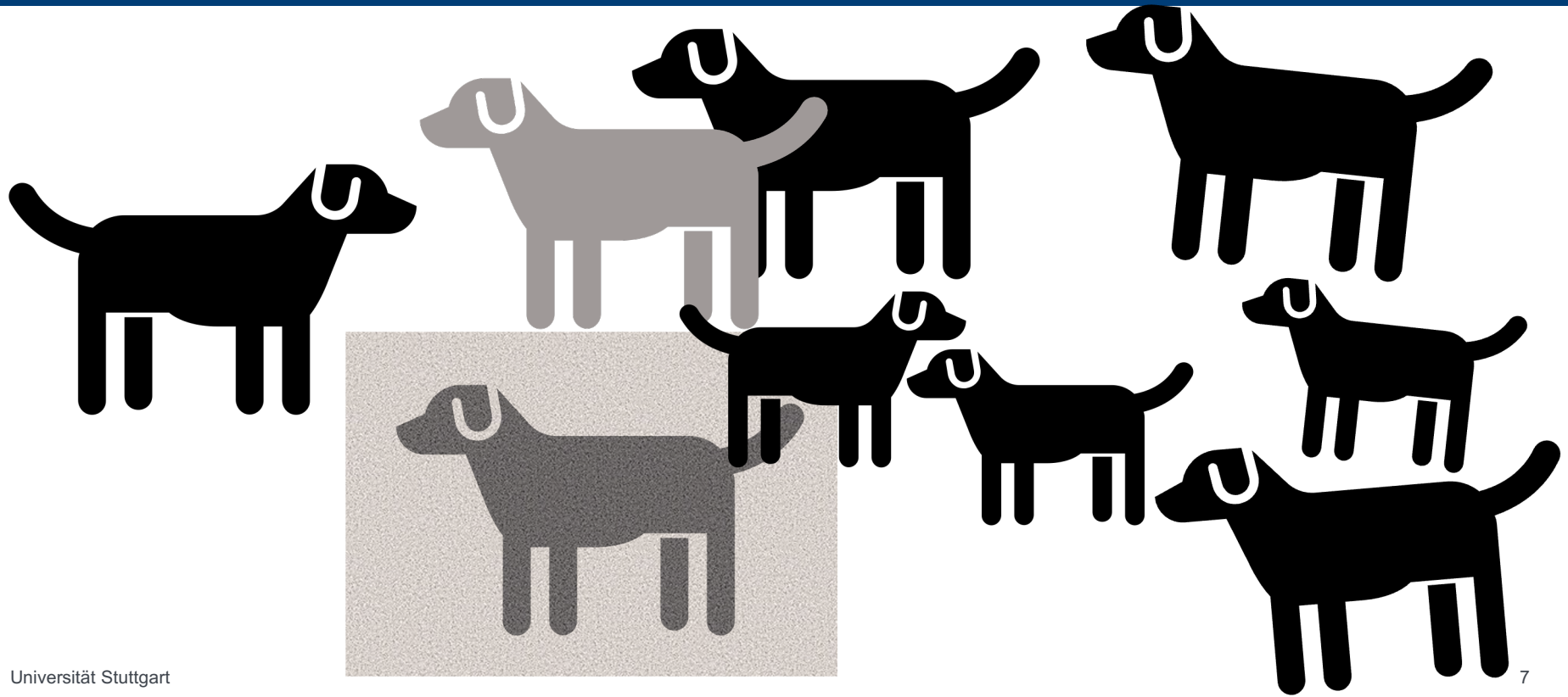
Overfitting

Mehr/variablere Trainingsdaten: Augmentation

In gewissem Maß getestet durch Evaluierung auf Test- bzw. Validierungsdaten; Maßnahmen siehe vorige Folien

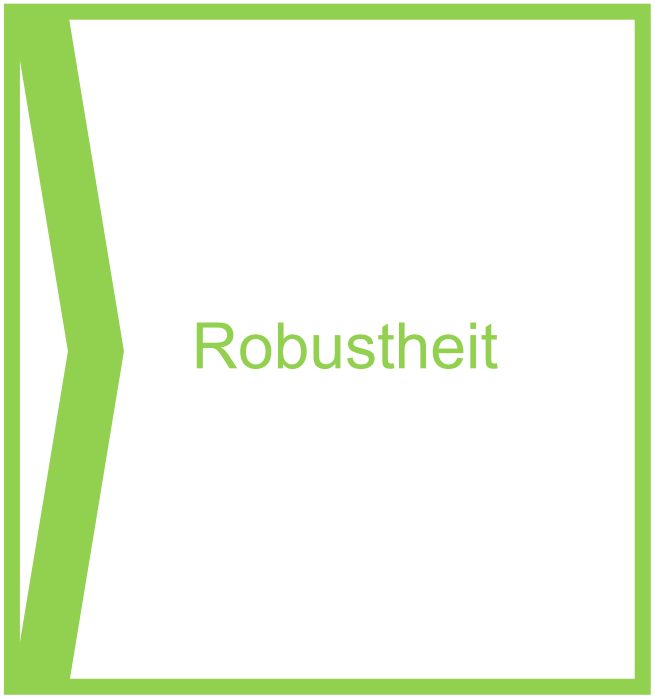
Data Augmentation

Aus 1 mach viele



Data Augmentation

- V.a. bei Klassifikationsproblemen
- Künstliche Datenvermehrung, insbesondere bei Bilddaten (aber nicht nur)
- Erzeuge aus den Trainingsdaten durch kleine Veränderungen der Features noch mehr Trainingsdaten
  - Idee: die Klasse bleibt dieselbe
- Ergibt **mehr** und **variablere** Daten
- Kann auch genutzt werden, um gleichmäßigere Klassenverteilung zu erreichen



Mögliche Transformationen bei Bilddaten

- Spiegeln, verkleinern/vergrößern, drehen, Objekte im Bild verschieben
- Farbverteilungen verändern, z.B. durch Anpassung von Helligkeit, Kontrast, Sättigung
- Rauschen (z.B. Farbwerte zufällig leicht verändern, oder zufällig ausgewählte Pixel weiß oder schwarz machen)

Data Augmentation (auf Deutsch in etwa „Datenanreicherung“) ist die künstliche Erzeugung von Trainingsdaten durch Anwendung von Transformationen auf die ursprünglichen Trainingsdaten.

Dadurch entstehen nicht nur mehr, sondern auch diversere Trainingsdaten. Data Augmentation verbessert so potentiell die Robustheit der darauf trainierten Modelle.

Stichwort „Robustheit“

Overfitting

Mehr/variablere Trainingsdaten: Augmentation

Mehrere/variablere Modelle: Ensembles

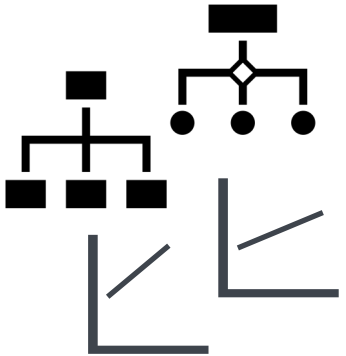
In gewissem Maß getestet durch Evaluierung auf Test- bzw. Validierungsdaten; Maßnahmen siehe vorige Folien

Ensembles

Was ist eigentlich ein Ensemble?

… Kleidungs-stücken

Zusammenstellung von …



… Machine Learning-Modellen

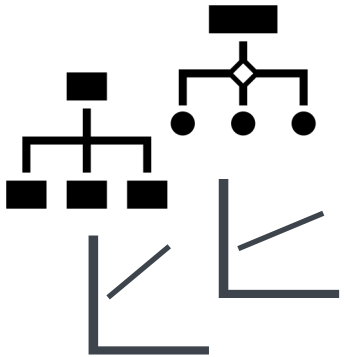
… Bauwerken

… Schauspieler*innen

… Musiker*innen

Idee von Ensembles

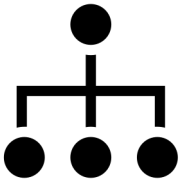
Mehrere **unterschiedliche** einfache Modelle



Vorhersagen durch „Abstimmung“ kombinieren

Ergebnis häufig besser als das Ergebnis des besten Modells

Verschiedene Modelle kombinieren

Bagging	Boosting	Stacking
		
Erzeuge n verschiedene Mengen von Trainingsdaten durch zufälliges Ziehen von je m Exemplaren aus den Daten Trainiere n Modelle auf je m Datenpunkten Mehrheitsentscheidung bzw. Mittelwert als Ergebnis	Trainiere nacheinander n Modelle Nach jedem Training die Trainingsdaten höher gewichten, für die die bisherigen Modelle Fehler machen Gewichte jedes Modell nach Performanz Gewichtete Mehrheitsentscheidung bzw. gewichteter Mittelwert als Ergebnis	Trainiere n Modelle verschiedener Art Trainiere ein Modell, das lernt, die Entscheidungen der n Modelle zu kombinieren

Beim Einsatz von Ensemble-Techniken werden mehrere möglichst diverse Modelle trainiert. Das Ergebnis ergibt sich dann durch Kombination der Ergebnisse der Modelle.

Durch Ensembles wird Overfitting vermieden.

Bemerkung

- Ensembles kamen bereits mit den einfachen Modellen auf (z.B. „Random Forests“ als Kombination vieler „Random Tree“ Entscheidungsbäume, die nur auf zufällig ausgewählte Features zugreifen dürfen)
- Funktioniert prinzipiell mit allen Methoden, bei denen sich zufallsbedingte Unterschiede im finalen Modell ergeben
- Also auch mit neuronalen Netzen
  - Hier können die Modelle des Ensembles sogar zu einem großen Modell kombiniert werden

Stichwort „Robustheit“

Konzeptverschiebung

Overfitting

Mehr/variablere Trainingsdaten: Augmentation

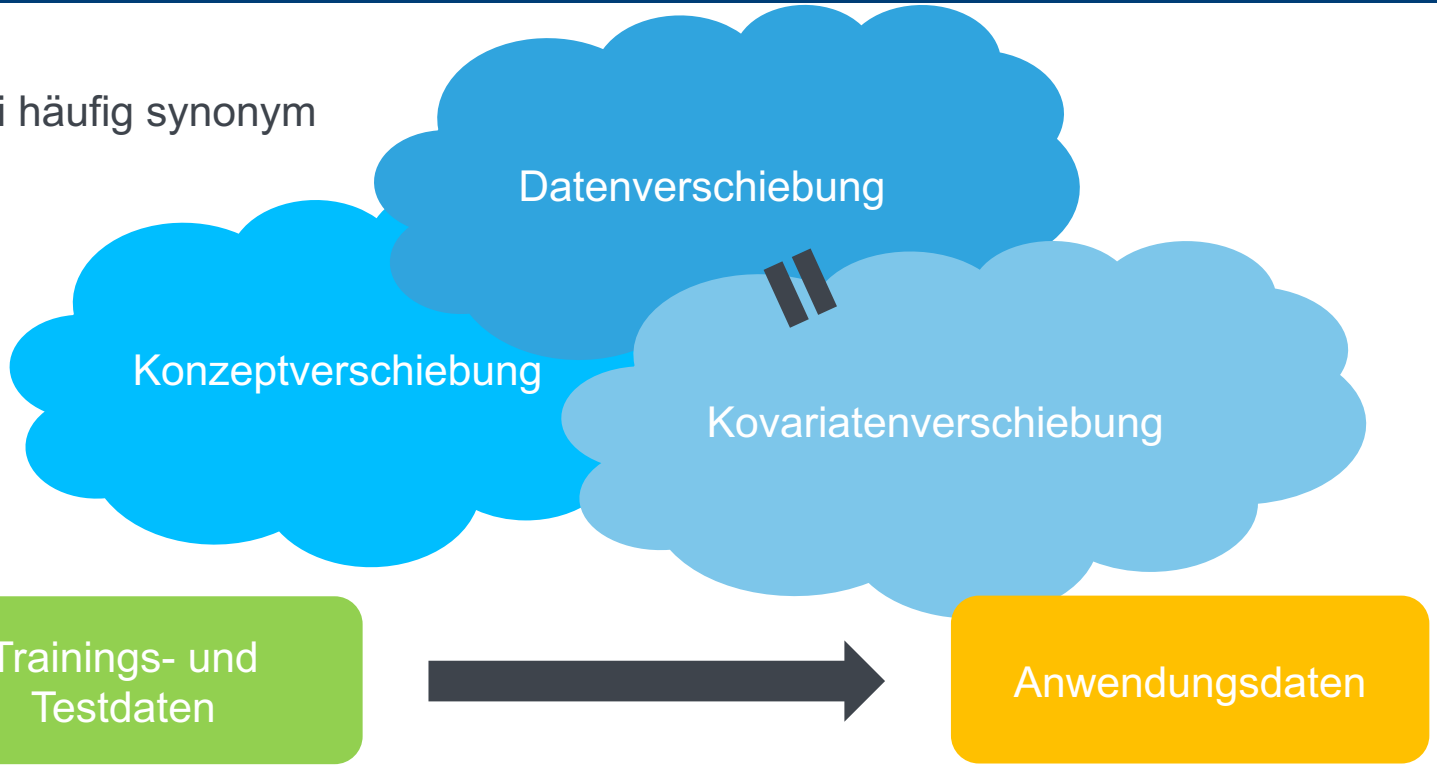
Mehrere/variablere Modelle: Ensembles

In gewissem Maß getestet durch Evaluierung auf Test- bzw. Validierungsdaten; Maßnahmen siehe vorige Folien

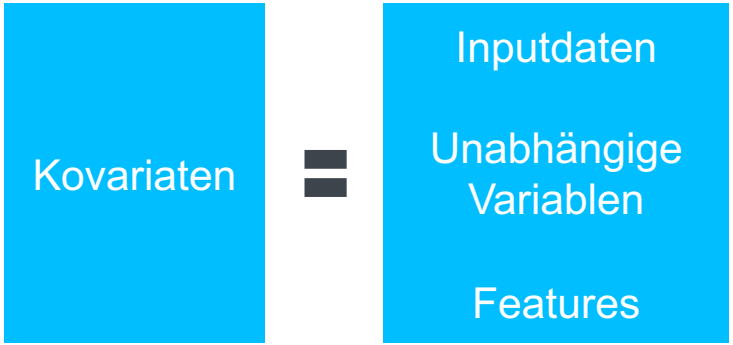
Konzeptverschiebung

Konzeptverschiebung

Alle drei häufig synonym



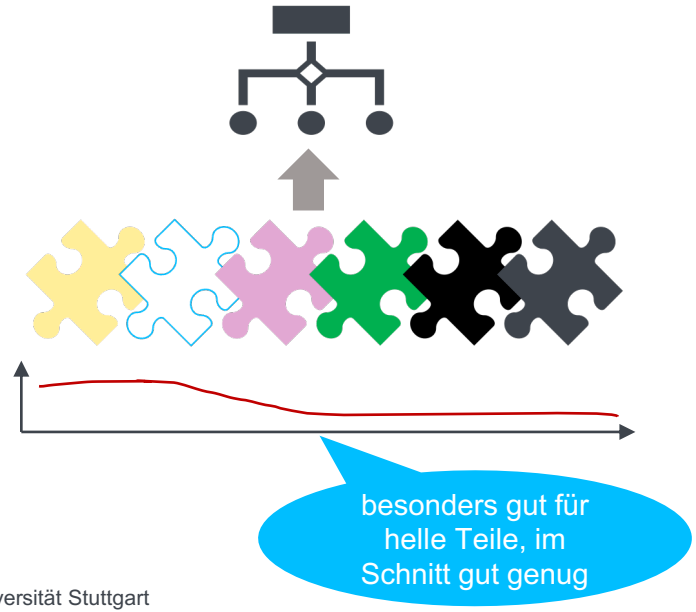
Kovariatenverschiebung, Datenverschiebung



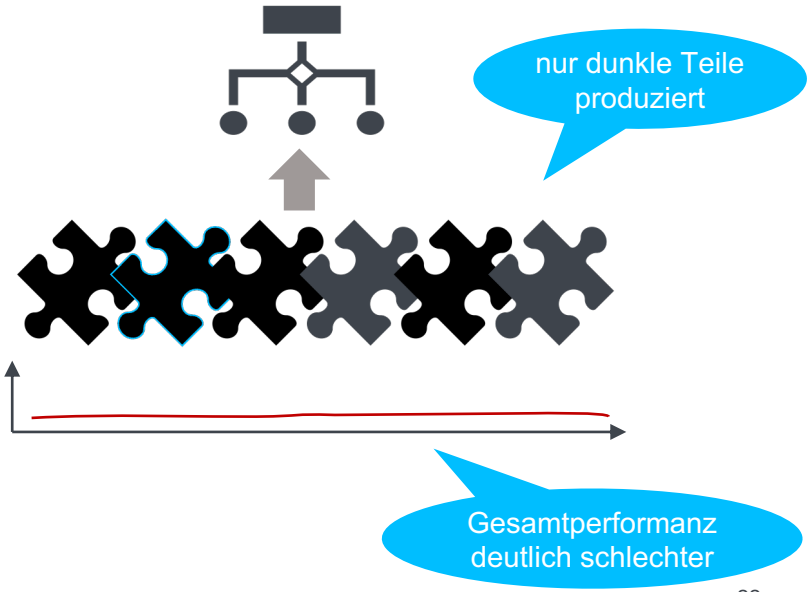
- „Verschiebung“
- Verteilung ändert sich
  - Häufigkeit der Werte, Häufigkeit von Kombinationen von Werten
- Unproblematisch, wenn Modell für alle Bereiche gleich gut

Beispiel: automatische Qualitätskontrolle von farbigen Kunststoffteilen

Training und Test

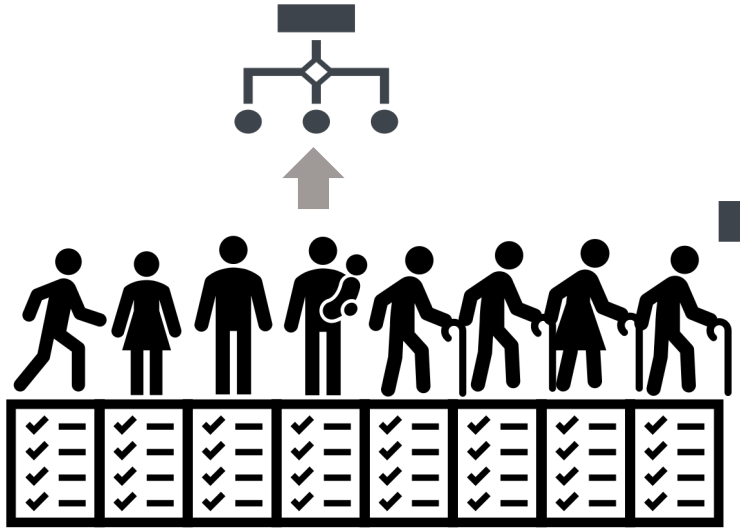


Anwendung 1 Jahr später



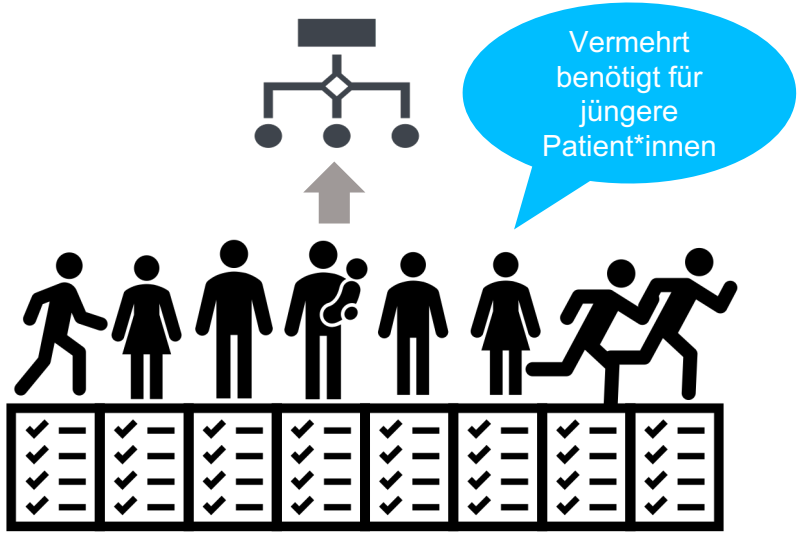
Beispiel: Risikofaktor für eine Krankheit

Training und Test



besonders gut für Ältere (dort häufiger)

Anwendung 10 Jahre später



Gesamtperformanz deutlich schlechter

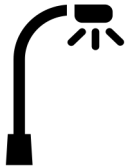
Von Kovariatenverschiebung oder Datenverschiebung spricht man, wenn sich die Verteilung der Inputdaten in ein Modell verändert.

Solche Veränderungen treten häufig im Laufe der realen Anwendung auf, oder sogar schon beim Übergang von Trainings- und Testphase zur Anwendung.

Konzeptverschiebung

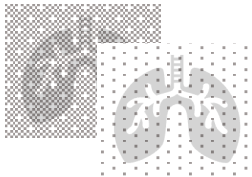
- Abhängigkeit zwischen Input und Output ändert sich
- Veränderte Bedeutung der Features
  oder sogar bisher nicht erfasste Features

andere Bildqualität bei Bilderkennung


Beleuchtung


andere Kamera


Komprimierung


besseres Gerät


anderer Winkel

Immobilienbewertung


neues Viertel
in Mode


Ölheizung vs.
Wärmepumpe

Kfz-Bewertung


Benzin vs. Diesel
vs. E-Motor

Attraktivität


veränderte
Normen

Von Konzeptverschiebung spricht man, wenn sich die Abhängigkeit zwischen Input und erwünschtem Output ändert.

Ursachen können sowohl Änderungen in der Erhebung der Inputdaten sein als auch externe, z.B. gesellschaftliche, Umstände, die die Bedeutung der Features verändern.

Auch solche Veränderungen treten häufig im Laufe der realen Anwendung auf.

Maßnahmen im Umgang mit Konzeptverschiebung

Bemerken

- Laufende Kontrolle der Verteilung der Inputdaten
- Laufende Bewertung des Modells: bleibt die Performanz hoch?

nein?

Beheben

- Mit aktuellen Daten neu trainieren
- Transfer-Learning oder Fine-Tuning mit aktuellen Daten

Stichwort „Robustheit“

Konzeptverschiebung

Overfitting

Neu trainieren, Fine Tuning, Transfer Learning

Beobachten: Laufende Evaluierung im Betrieb

Mehr/variablere Trainingsdaten: Augmentation

Mehrere/variablere Modelle: Ensembles

In gewissem Maß getestet durch Evaluierung auf Test- bzw. Validierungsdaten; Maßnahmen siehe vorige Folien

Stichwort „Robustheit“

Adversarial Examples

Konzeptverschiebung

Overfitting

Neu trainieren, Fine Tuning, Transfer Learning

Beobachten: Laufende Evaluierung im Betrieb

Mehr/variablere Trainingsdaten: Augmentation

Mehrere/variablere Modelle: Ensembles

In gewissem Maß getestet durch Evaluierung auf Test- bzw. Validierungsdaten; Maßnahmen siehe vorige Folien

Adversarial Attacks

Adversarial Attacks – „Feindliche Angriffe“

Poisoning („Vergiftung“)

- Manipulation der Trainingsdaten eines Modells

Datensicherheit

Evasion („Vermeidung“)

- Manipulation der Inputdaten bei der Anwendung des Modells
- kaum bemerkbare Veränderung des Inputs, bis falsche Klassifikation erreicht
- “Adversarial Example“ („feindliches Beispiel“)

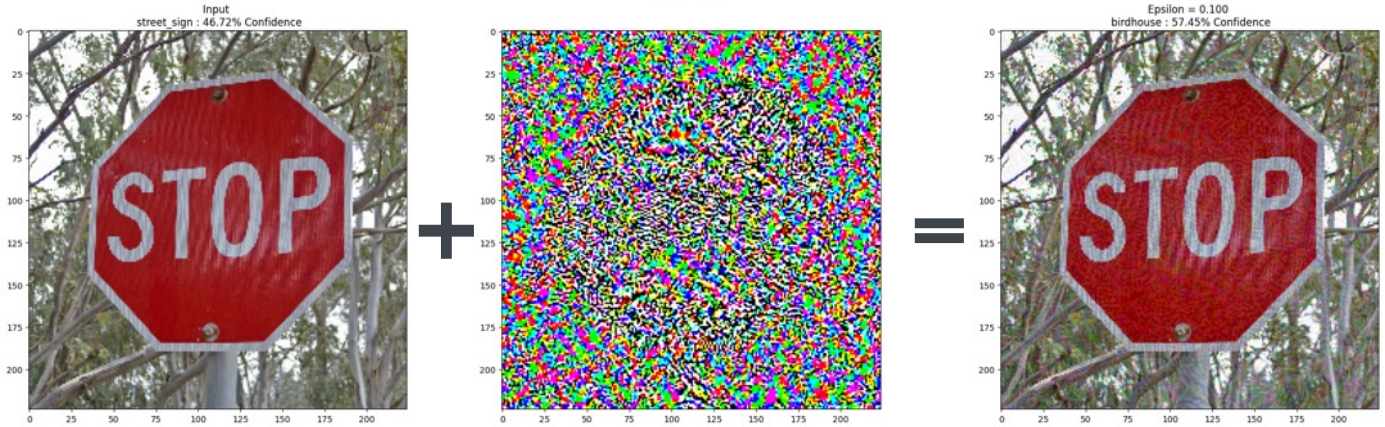
Robustere Modelle

Vertraulichkeit

- Ungewollter Zugriff auf das Modell oder Informationen im Modell
- Z.B. auch Rückschlüsse auf vertrauliche Trainingsdaten

Datensicherheit

Beispiele für Adversarial Examples



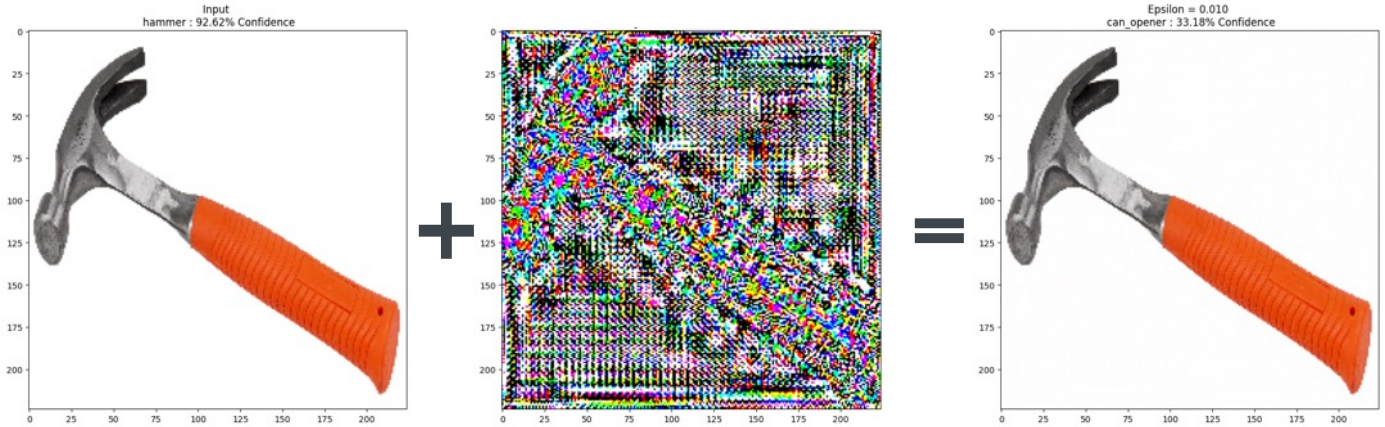
Klassifikation mit MobileNet V2: Straßenschild

Veränderung

Klassifikation mit MobileNet V2: Vogelhaus

Beispiele generiert mit Code von https://www.tensorflow.org/tutorials/generative/adversarial_fgsm

Beispiele für Adversarial Examples



Klassifikation mit MobileNet V2: Hammer

Veränderung

Klassifikation mit MobileNet V2: Dosenöffner

Beispiele generiert mit Code von https://www.tensorflow.org/tutorials/generative/adversarial_fgsm

Gefahr von Adversarial Examples

- Berühmtes Beispiel: mit Aufklebern beklebte echte Stoppschilder werden nicht mehr als Stoppschilder erkannt*

* z.B. berichtet in https://arxiv.org/pdf/1707.08945.pdf

Umgang mit Adversarial Examples

Finden von Adversarial Examples

- Adversarial Examples können systematisch erzeugt werden
- Es gibt diverse Algorithmen dafür
- Besonders leicht, wenn Zugriff auf die Gradienten möglich

Problematisch, wenn Außenstehende Input für Modelle liefern

Robustheit gegenüber Adversarial Examples

- Es gibt laufend neue Gegenmaßnamen
- z.B. schon beim Training Adversarial Examples suchen und in Trainingsdaten integrieren
- Gegenmaßnahmen werden durch neue Algorithmen wieder umgangen
- „Wettrüsten“
- Forschungsthema!

Adversarial Examples sind Inputs für Modelle, die so manipuliert wurden, dass statt des ursprünglich richtigen Ergebnis ein falsches Ergebnis vorhergesagt wird.

Dabei ist die Manipulation für Menschen nicht bemerkbar: Menschen würden weiterhin das korrekte Ergebnis vorhersagen.

Stichwort „Robustheit“

Adversarial Examples	Forschungsthema
Konzeptverschiebung	Neu trainieren, Fine Tuning, Transfer Learning
	Beobachten: Laufende Evaluierung im Betrieb
Overfitting	Mehr/variablere Trainingsdaten: Augmentation
	Mehrere/variablere Modelle: Ensembles
	In gewissem Maß getestet durch Evaluierung auf Test- bzw. Validierungsdaten; Maßnahmen siehe vorige Folien

INNOVET
KI B³
Künstliche Intelligenz in die berufliche Bildung bringen

Dr. Antje Schweitzer

Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung

Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung
Institut für Software Engineering

IHK Industrie- und Handelskammer Reutlingen
Reutlingen | Tübingen | Zollernalb

IHK Region Stuttgart

IHK Industrie- und Handelskammer Karlsruhe

LMU Ludwig-Maximilians-Universität München

GEFÖRDERT VOM
INNOVET
Bundesministerium für Bildung und Forschung
bibb Bundesinstitut für Berufsbildung

Gefördert als InnoVET-Projekt aus Mitteln des Bundesministeriums für Bildung und Forschung.

Lizenzbestimmungen

Quellenverzeichnis

Titelfoto: Sigmund (https://unsplash.com/de/@sigmund) auf Unsplash (https://unsplash.com/de/fotos/weisse-und-schwarze-bucher-auf-weissem-tisch-UorPU70_D60), lizenziert unter Unsplash-Lizenz (https://unsplash.com/license). Bildausschnitt verändert.

S. 32: Foto von Bidgee (https://commons.wikimedia.org/wiki/File:STOP_sign.jpg), Stoppschild in Australien, lizenziert unter CC-BY 3.0 (https://creativecommons.org/licenses/by/3.0).

S. 33: Foto von Evan-Amos, Hammer, Public domain, via Wikimedia Commons