Explainable AI mit Bilddaten

Unterscheidung von Pembroke und Cardigan Welsh Corgis

Lukas, Janik, Robin, Felix
DHBW - ExAI Projekt

Robin: Einführung

Agenda

- Datensatz-Auswahl & Modelltraining: Robin
- XAI Verfahren: Janik
- Demo: Felix
- Analyse & Kritische Diskussion: Lukas

Zielsetzung

Problemstellung

Unterscheidung der Corgi-Rassen **Pembroke** und **Cardigan** bei Mischlingen mit Hilfe eines CNN.

Untersuchung

Beobachtung des Verhaltens bei Input von Mischlingen
Ziel: Erklärbare Entscheidungen durch XAI-Methoden

Datensatz-Auswahl & Preprocessing

Datensatz

- **Stanford Dogs Dataset**
- 120 Hunderassen, über 20.000 Bilder
- Verwendung der Klassen: Pembroke & Cardigan Welsh Corgi
- **Link zum Datensatz**

Datenvorverarbeitung

- Train/Val-Split im Verhältnis 80%/20%
- Bildtransformationen für Trainingsdaten:
    - Resize auf 224x224px
    - Random Horizontal Flip
    - Random Rotation (10°)
    - Color Jitter (Helligkeit, Kontrast, Sättigung)
- Für Validierungsdaten: nur Resize und Normalisierung

Modelltraining

Modellauswahl

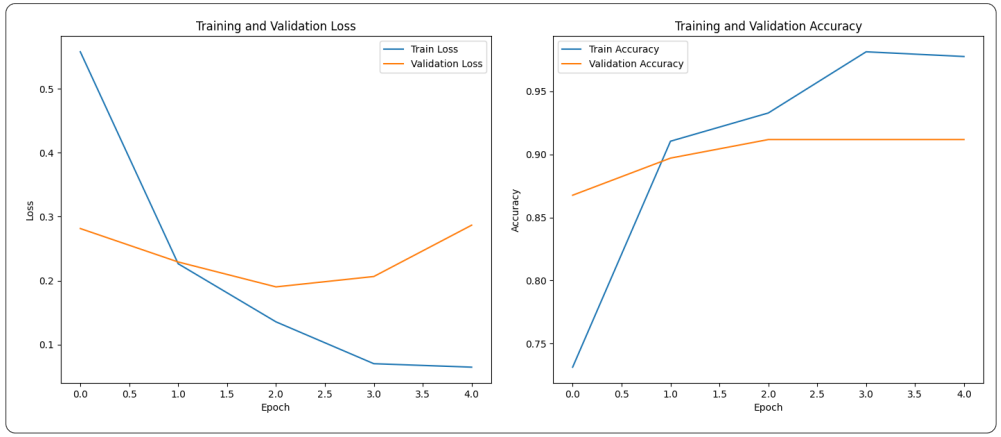
- Verwendung eines Convolutional Neural Networks (CNN) für Bilddaten
- Begründung: CNNs sind spezialisiert auf die Extraktion von Merkmalen aus Bildern
- Transfer Learning mit ResNet50, vortrainiert auf ImageNet
- Fine-Tuning auf Corgis (Pembroke&Cardigan)
- Nur letzte Layer ersetzt: Dense Layer für 2-Klassen Klassifikation

Modellarchitektur & Training

- Architektur: ResNet50 mit 50 Layern, Dense Layer für 2-Klassen Klassifikation
- Hyperparameter: Learning Rate 0.001, Batch Size 32
- Aktivierungsfunktionen: ReLU in den versteckten Schichten
- Training: CrossEntropyLoss als Verlustfunktion, Adam Optimizer
- Transfer Learning: Nur letzte Schichten trainiert
- Zunächst 10 Epochen für Feinabstimmung, optimiert auf 3 Epochen.

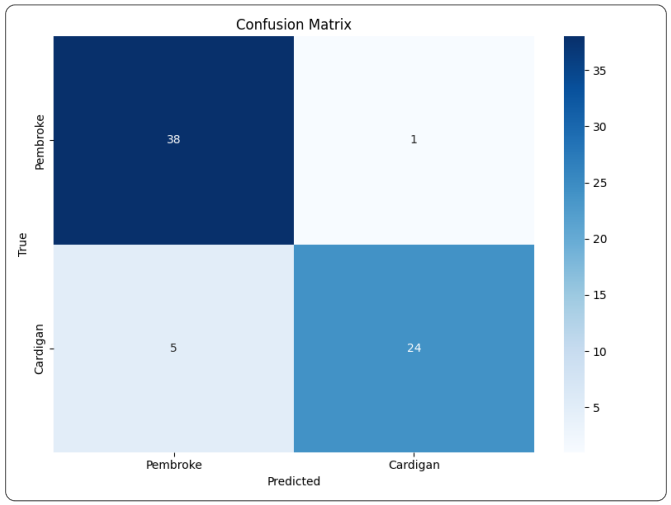
Modell-Ergebnisse

- Accuracy: >90% auf Validierungsdaten
- Transfer Learning mit Fine-Tuning der letzten Layer (layer4)
- Früher Trainingsabbruch durch Early Stopping (patience=5)
- Adaptive Lernrate mit ReduceLROnPlateau Scheduler
- Beste Ergebnisse bei Bildern mit klaren rassetypischen Merkmalen



Konstante Abnahme des Trainingsverlusts und Zunahme der Genauigkeit ohne Überanpassung. Die Validierungskurve stabilisiert sich bei etwa 93% Genauigkeit.

Modell-Ergebnisse: Confusion Matrix



Die Konfusionsmatrix zeigt: 38 korrekte Pembroke- und 24 korrekte Cardigan-Vorhersagen. Nur 6 Fehler insgesamt, hauptsächlich Cardigans als Pembrokes klassifiziert.

Ethische Betrachtungen

Ethik & Verantwortung

- Transparenz als Voraussetzung für verantwortungsvolle KI-Systeme
- Vermeidung von Biases durch erklärbare Entscheidungsprozesse
- Datenschutz bei der Erfassung und Verarbeitung von Bilddaten
- Berücksichtigung ethischer Aspekte bei der Entwicklung von XAI-Methoden
- Förderung des Vertrauens in KI-Systeme durch transparente Entscheidungen
- Verantwortlicher Einsatz von KI in sensiblen Anwendungsbereichen

Janik: XAI Verfahren

XAI-Verfahren im Überblick

1. **Contrastive Grad-CAM**: Visualisiert Unterschiede zwischen Klassen
2. **Layerwise Relevance Propagation (LRP)**: Liefert tiefere Einsicht auf Pixelebene

Beide Methoden erlauben es, die Entscheidungen des Modells nachzuvollziehen

Direkte Vergleichsmöglichkeit der Erklärungsansätze bei verschiedenen Bildtypen (reinrassig vs. Mischlinge)

Grad-CAM: Technische Details

- Verwendet Gradienten der letzten Convolutional Layer
- Target Layer: layer4 von ResNet50
- Berechnet gewichtete Aktivierungskarten
- Implementierung mit PyTorch Hooks für Forward/Backward Pass

Code-Snippet: Grad-CAM

Teil 1: ➡️ & ⬅️ Pass Gradienten

```
def __call__(self, input_tensor, target_class=None):
    input_tensor = input_tensor.to(device)
    # Reset gradients
    self.model.zero_grad()

    # => Forward pass
    output = self.model(input_tensor)

    if target_class is None:
        # ..use predicted class..
        target_class = torch.argmax(output, dim=1).item()
```

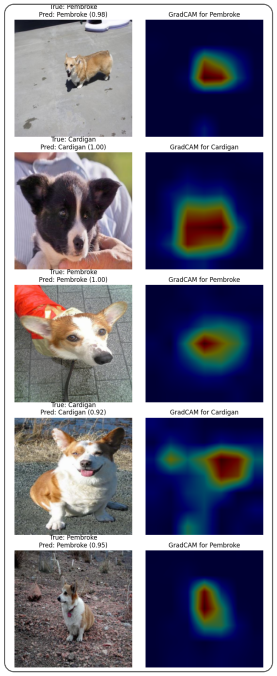
Code-Snippet: Grad-CAM

Teil 2: Aktive-Gewichte & Heatmap

```
# Get the target layer
target_layer = model.layer4

# Create GradCAM instance
grad_cam = GradCAM(model, target_layer)
cam = grad_cam(img_tensor, target_class)

cam_resized = cv2.resize(cam, (img_np.shape[1], img_np.shape[0]))
# ..to input image size

# Convert to heatmap
heatmap = cv2.applyColorMap(np.uint8(255 * cam_resized), cv2.COLORMAP_JET)
```

Visualisierung: Grad-CAM

Heatmap zeigt die für die Klassifikation relevanten Regionen:



- Rote/gelbe Bereiche zeigen Hauptaufmerksamkeit des Modells
- Bei Pembroke: Fokus auf Kopfform, Ohren und kurzen Schwanz
- Bei Cardigan: Fokus auf größere Ohren und längeren Schwanz
- Unterschiedliche Färbungsmuster beeinflussen Entscheidungsfindung

LRP: Technische Details

- Propagiert Vorhersagen rückwärts durch das Netzwerk
- Berechnet Beiträge jedes Pixels zum finalen Output

Code-Snippet: LRP

Teil 1: Relevanz-Gradienten in ➡️ & ⬅️ Pass

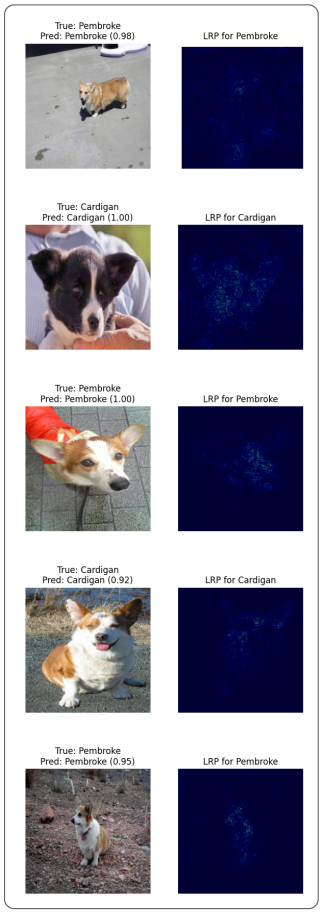
```
def __call__(self, input_tensor, target_class=None):
    # Make a detached copy of the input
    input_copy = input_tensor.clone().detach().to(device)
    input_copy.requires_grad = True

    # Forward pass
    self.model.zero_grad()
    output = self.model(input_copy)

    if target_class is None:
        # ..use predicted class
        target_class = torch.argmax(output, dim=1).item()
```

Code-Snippet: LRP

Teil 2: Normalisierte Relevanz-Heatmap

```
# Create LRP instance
lrp = LRP(model)

try:
    # Generate relevance map
    relevance_map = lrp(img_tensor, target_class)

    if relevance_map is None:
        # Return a blank heatmap if LRP fails
        relevance_map = np.zeros((img_np.shape[0], img_np.shape[1]))
        visualization = img_np.copy()
        return visualization, relevance_map
```

Visualisierung: LRP



Detaillierte Pixel-Relevanzverteilung zur finalen Entscheidung

Kritische Betrachtung der XAI-Methoden

- XAI-Visualisierungen bieten Erklärungen, aber keine kausalen Zusammenhänge
- Subjektivität in der Interpretation der Visualisierungen
- GradCAM: Fokus auf letzte Layer könnte wichtige frühe Features übersehen
- LRP: Höhere Komplexität erschwert intuitive Interpretation
- Balance zwischen Erklärbarkeit und technischer Tiefe ist herausfordernd

Felix:Demo zu Mischlingen vs Reinrassen

LIVE-DEMO

Vergleich der XAI-Methoden bei Rassenmerkmalen

Pembroke Welsh Corgi

- Fokus auf fuchsartige Kopfform
- Hervorhebung der aufrechten, spitzen Ohren
- Aktivierung bei bestimmten Fellmustern
- Kurzer oder fehlender Schwanz

Cardigan Welsh Corgi

- Deutliche Aktivierung am langen Schwanz
- Hervorhebung der größeren, runderen Ohren
- Fokus auf breiteren Körperbau
- Activation bei dunkleren Fellfarben

Anwendungsfall: Mischling-Erkennung

- Experiment: Bewertung von Mischlings-Bildern beider Rassen
- Beobachtung: Konfidenz des Modells sinkt bei gemischten Merkmalen (oft unter 75%)
- GradCAM: aktiviert Regionen beider Rassen gleichzeitig
- LRP: zeigt konfliktäre Pixel-Aktivierungen für beide Klassen
- XAI ermöglicht transparenten Einblick in Modell-Unsicherheit
- Ermöglicht besseres Verständnis von Entscheidungsgrenzen im Modell

Lukas: Analyse & Kritische Diskussion

Stärken & Grenzen des Ansatzes

- **Stärken:**
  - Hohe Klassifikationsgenauigkeit (>90%)
  - Transparente Entscheidungsprozesse durch XAI
  - Effiziente Nutzung von Transfer Learning
- **Grenzen:**
  - Eingeschränkte Generalisierbarkeit bei untypischen Bildaufnahmen
  - Abhängigkeit von der Qualität des Trainingsdatensatzes
  - Interpretationsaufwand bei XAI-Methoden

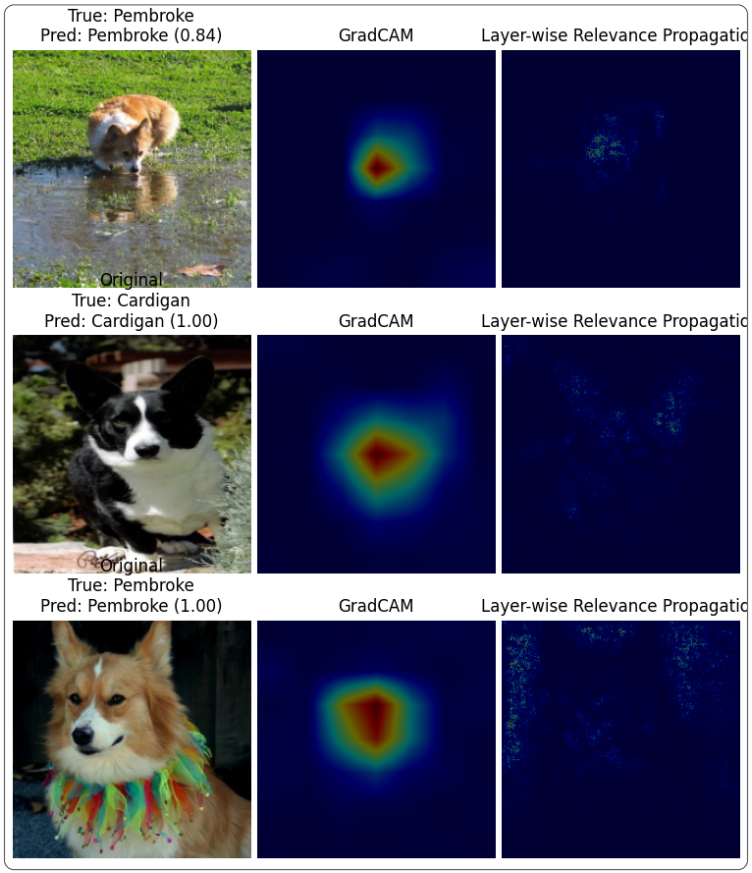
Zusammenfassung der XAI-Erkenntnisse

- Beide XAI-Methoden zeigen, dass das Modell tatsächlich die rassetypischen Merkmale erkennt
- Rasseunterschiede werden primär anhand anatomischer Features erkannt:
    - Schwanz (lang vs. kurz/fehlend)
    - Ohren (groß/rund vs. spitz/aufrecht)
    - Körperbau (breiter vs. schlanker)
- Bei Mischlingen: XAI offenbart die "Unsicherheit" des Modells visuell
- Direkter Vergleich zeigt komplementäre Stärken der Methoden: GradCAM (Übersicht) und LRP (Detail)

Vergleich XAI-Methoden: Gesamtüberblick

Kriterium	Grad-CAM	LRP
Interpretierbarkeit	hoch	hoch
Modellabhängigkeit	nur CNN	flexibel
Genauigkeit	grob	fein
Berechnungskosten	gering	hoch
Anwendung	schnell	detailliert

Vergleich XAI-Methoden: Ergebnisse



Direkter Vergleich: GradCAM (links) zeigt grobe Fokusregionen, während LRP (rechts) pixelgenaue Merkmalszuordnung ermöglicht.

Praktische Anwendungsfälle

- **Tiermedizinische Diagnose:** Identifikation von Anomalien in Tierbildern
- **Zuchtanalyse:** Objektive Bewertung von Rassemerkmalen
- **Bildsuche:** Verbesserung von Suchergebnissen durch merkmalsbasierte Ähnlichkeiten
- **Qualitätssicherung:** Überprüfung der Modellentscheidungen in sicherheitskritischen Anwendungen
- **Bildungsbereich:** Visuelle Darstellung von Merkmalsunterschieden für Lernzwecke

Ausblick

- Integration weiterer XAI-Methoden (SHAP, Integrated Gradients)
- Erweiterung auf komplexere Klassifikationsaufgaben
- Quantitative Bewertung der XAI-Ergebnisse
- Verbesserung der Modellrobustheit durch XAI-Feedback

Beantwortung unserer Fragestellung

- Konnten wir Pembroke und Cardigan Corgis mit unserem CNN unterscheiden?
- Wie hat sich das Modell bei Mischlingen verhalten?
- Haben die XAI-Methoden uns geholfen zu verstehen, wie das Modell entscheidet?
- Welche Rassemerkmale waren am wichtigsten für die Klassifikation?
- Ist unser Ansatz für ähnliche Probleme anwendbar?

Vielen Dank!

Fragen?

Projektteam: Lukas, Janik, Robin, Felix

Code und Präsentation: **https://github.com/mausio/ExAI**