Explainable Visual Question Answering (VQA)

Motivation

- Erklärbarkeit von Black-Box-KI-Modellen mithilfe eines generierten „Wärmebildes“
- Barrierefreiheit durch eine sprachliche Ausgabe einer zusätzlich generierten textuellen Erklärung

Technologien

Backend

python™

PyTorch

Frontend

Streamlit

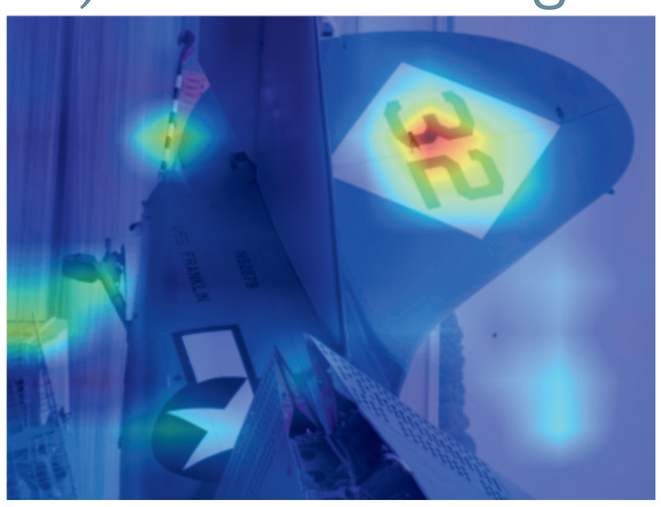
VQA-Modell

InstructBLIP

Overview

Visual Question Answering

1) Visuelle Erklärung



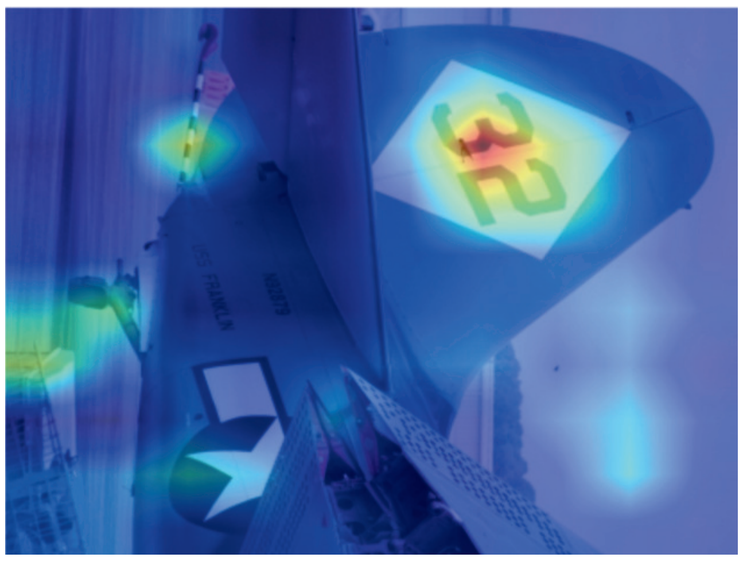
2) Textuelle Erklärung

The visible part of the sign shows the digits “3” and “2,” forming the number 32.

1) Visuelle Erklärung

Ziel: Herausfinden wo die KI hinschaut?

Ausgabe: Ein “Wärmebild” → warme Bereiche entsprechen wichtigen Bereichen

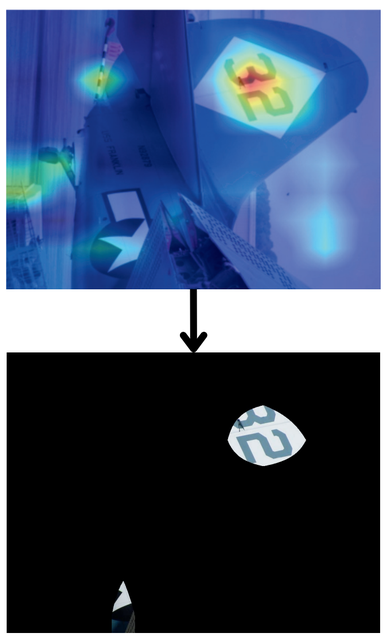


Umsetzung: Anwendung der SIDU-VQA-Methode

1. Bild in Raster unterteilen
2. Systematisches Ausblenden einzelner Bildausschnitte
3. Messung der Auswirkung auf das KI-Verständnis
   - Wenn ein wichtiges Quadrat abgedeckt wird → Die KI gibt eine veränderte Antwort
   - Wenn ein unwichtiges Quadrat abgedeckt wird → Die KI gibt immer noch dieselbe Antwort
4. Wichtige Quadrate werden im “Wärmebild” hervorgehoben

2) Textuelle Erklärung

1) Heatmap zu Maske



2) Objekte erkennen

Object Detection Model

↓

Erkennt vorhandene Objekte im maskierten Bild

3) Prompt bauen

Anweisung mit diesen Infos:
+ Frage zum Bild
+ Antwort von InstructBLIP
+ maskiertes Bild ODER erkannte Objekte

4) Erklärung generieren

LLM

↓

Textuelle Erklärung der InstructBLIP -Antwort