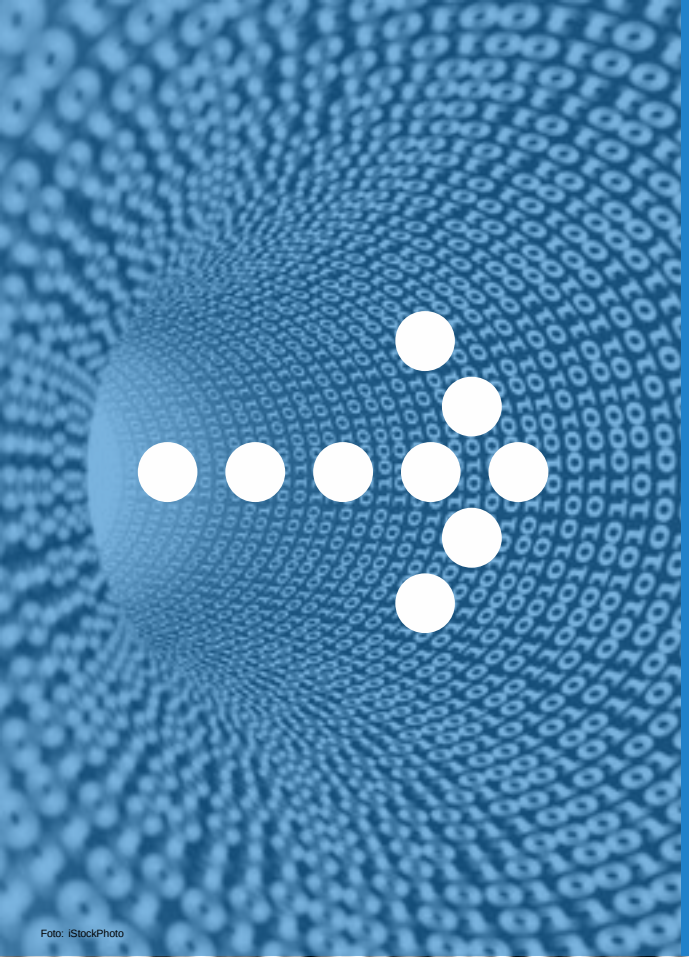


Foto: iStockPhoto

Hur kan AI-baserade analyser göras begripliga för människan?

Peter Svenmarck, Erik Wachtmeister

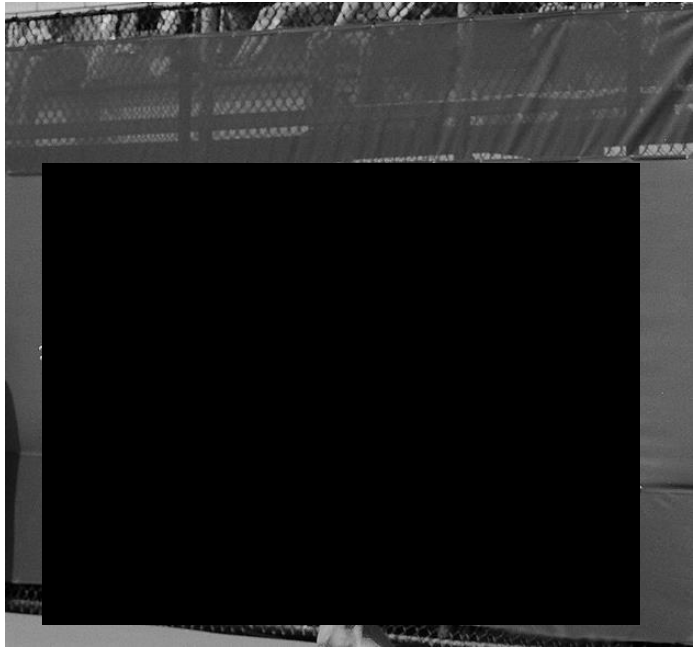
2026-05-07

Objektdetektion och klassificering av aktivitet

TennisSwing

Bild: Julius Motal, Creative Commons Attribution 2.0 Generic

Objektdetektion (inkl. klassificering)



?

Bild: Julius Motal, Creative Commons Attribution 2.0 Generic

Varför XAI? Några exempel

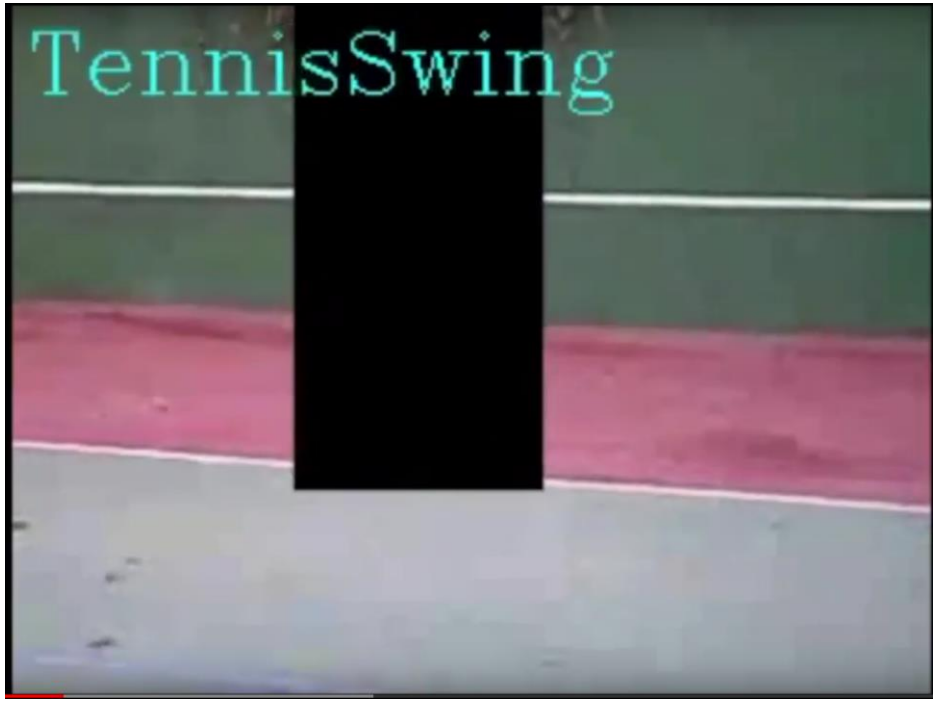


Bild: Yun He et al. (2016). Human Action Recognition without Human.

Varför XAI? Några exempel

Bild: Bojarski m.fl. (2018). VisualBackProp: efficient visualization of CNNs for autonomous driving

Saliency maps för självkörande bil

FOI

XAI och djupinlärning: Hur?

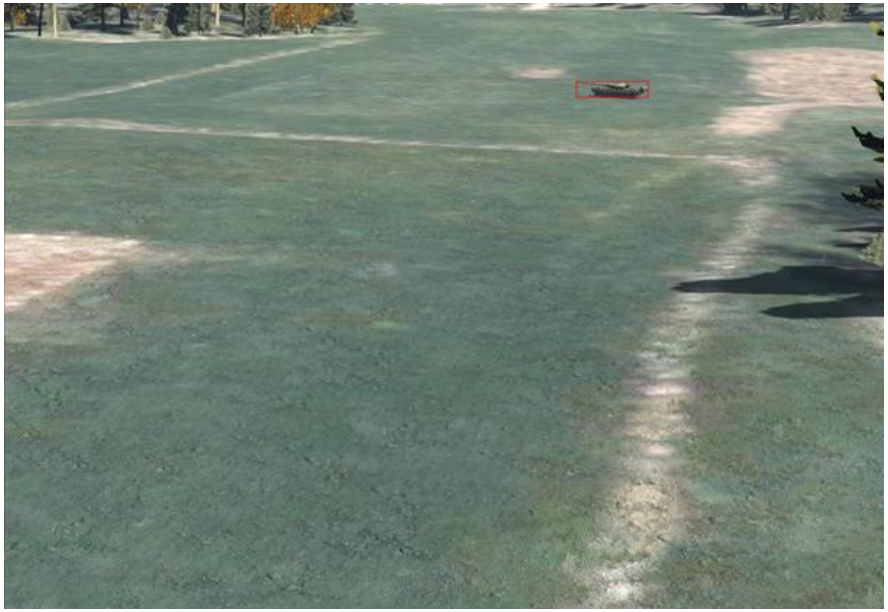
- Modellspecifika och modelloberoende tekniker
- Lokala, globala och hybridtekniker
- Vad är en användbar förklaring?
  - Vilken XAI-teknik ska man välja?
  - XAI-tekniker föreslås ofta men utvärderas sällan av användare

Bilder: Lapuschkin m.fl. (2019). Unmasking Clever Hans predictors and assessing what machines really learn

FOI

XAI-experiment för måligenkänning

Utvärderade om DNN-klassificering och RISE saliency maps förbättrar måligenkänningen av UAV-bilder på låg höjd från Virtual Battle Space (VBS) 3



För- och nackdelar med saliency maps

- Förbättrar användarnas förståelse av modellens prediktioner
- Visar var användarna ska titta men inga detaljer om vad de ska titta på
- Inte tillräckligt distinkta för att upptäcka felaktiga prediktioner
- Oklart om de minskar tendensen att lita på automatiska beslutsstöd

Förutsättningar för experimentet (1)

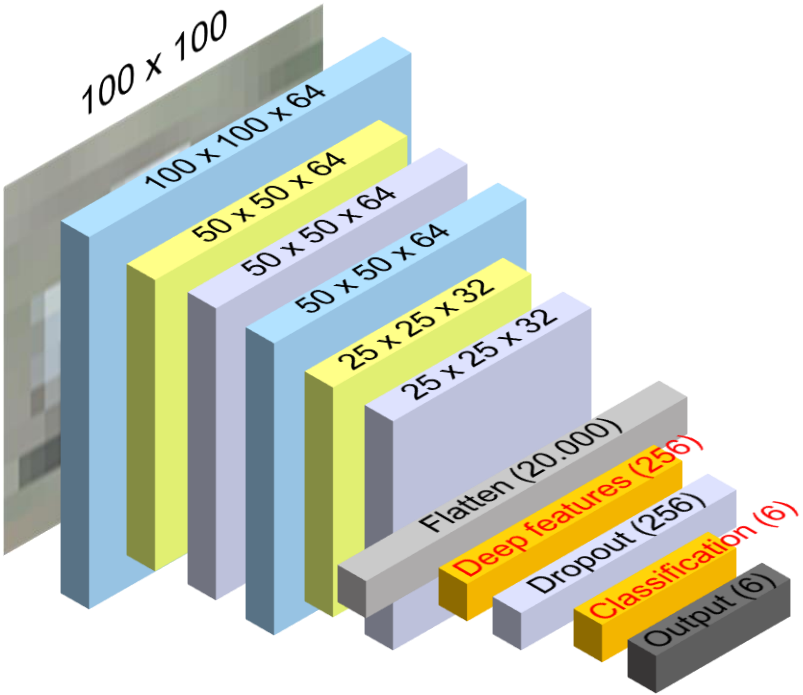
- Sex fordonsklasser
  - T-72, BMP-2, BTR-80, 2S3, ZSU-23 och Toyota Hilux
  - Olika synvinklar och platser i bilden
- Tre typer av stöd
  - Utan DNN-klassificerare, med DNN-klassificerare, med DNN-klassificerare och RISE saliency maps
- Två DNN korrekthet
  - Korrekt och inkorrekt klassificering
- Två upplösningar
  - Låg och medel

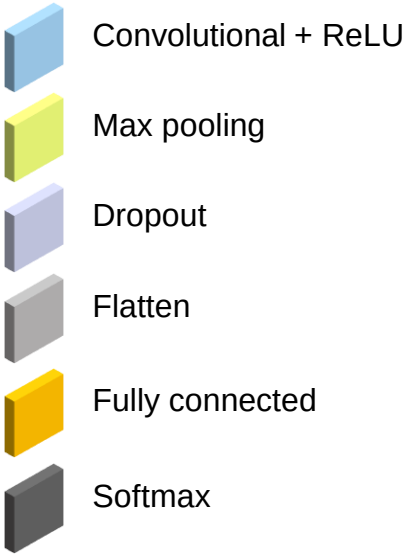
FOI

Förutsättningar för experimentet (2)

- Mått
  - Noggrannhet, responstid, överenstämmelse med DNN-klassificering (förtroende)
- Frågeformulär
  - Mental arbetsbelastning, tillit och nöjdhet
- Förstudier med FOI-personal för att hitta lämplig svårighetsgrad
- Deltagare var 16 studenter

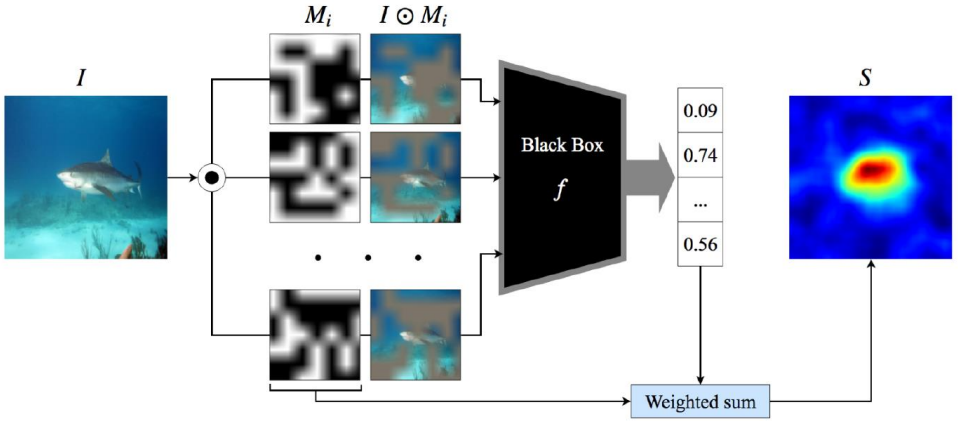
DNN-klassificeraren





RISE saliency maps

- Modelloberoende teknik
- Använder flera små slumpmässiga binära masker som förstoras till bildstorleken
- Beräknar vilka områden som har störst påverkan på klassificeringen



Fördelar med RISE saliency maps

- Använder särskiljande särdrag
- Använder särdrag som är viktiga för klassificering
- Särskiljer mellan klasser
- Små effekter av oviktiga variationer

Fordonsklasser

T-72

BMP-2

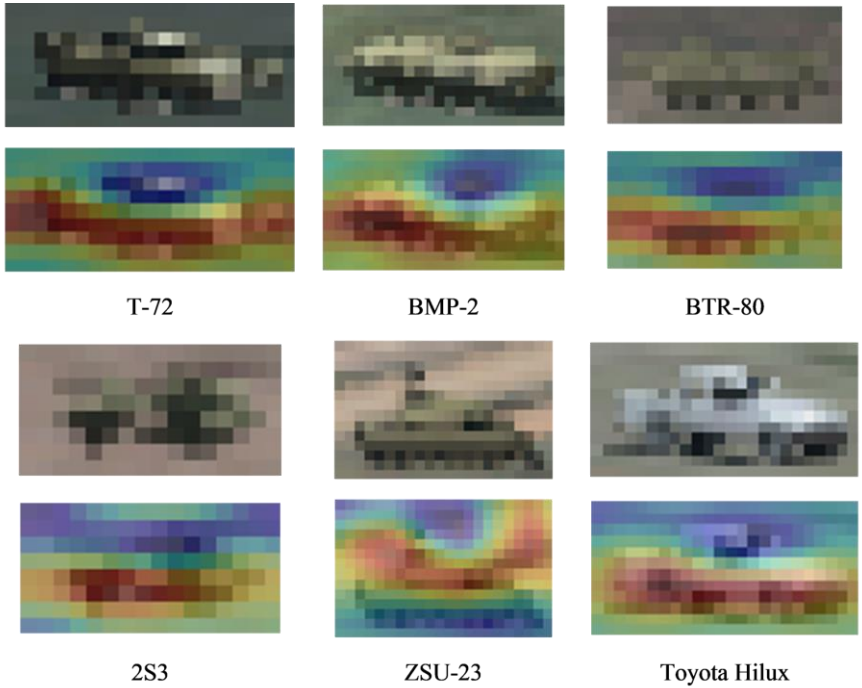
BTR-80

2S3

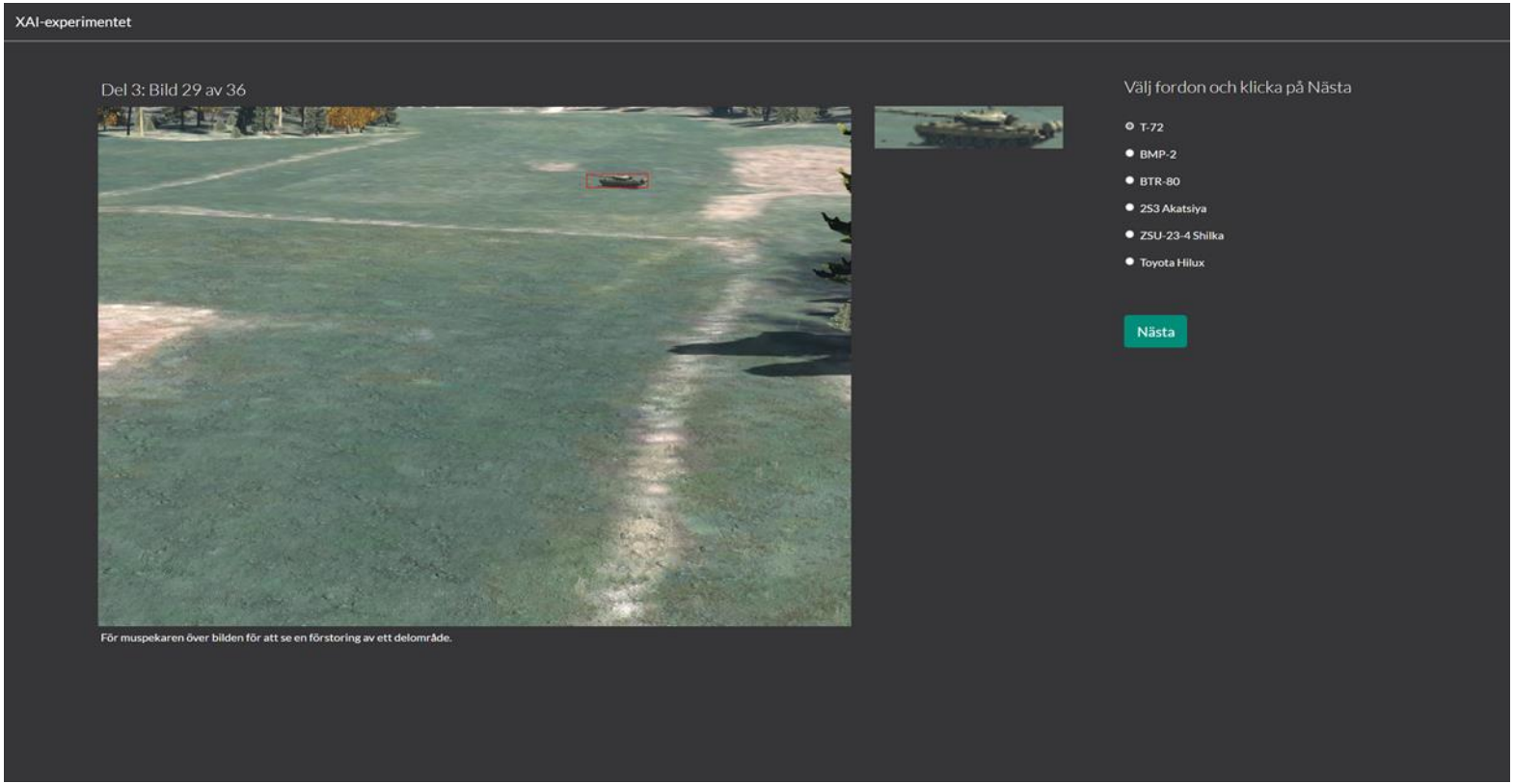
ZSU-23

Toyota Hilux

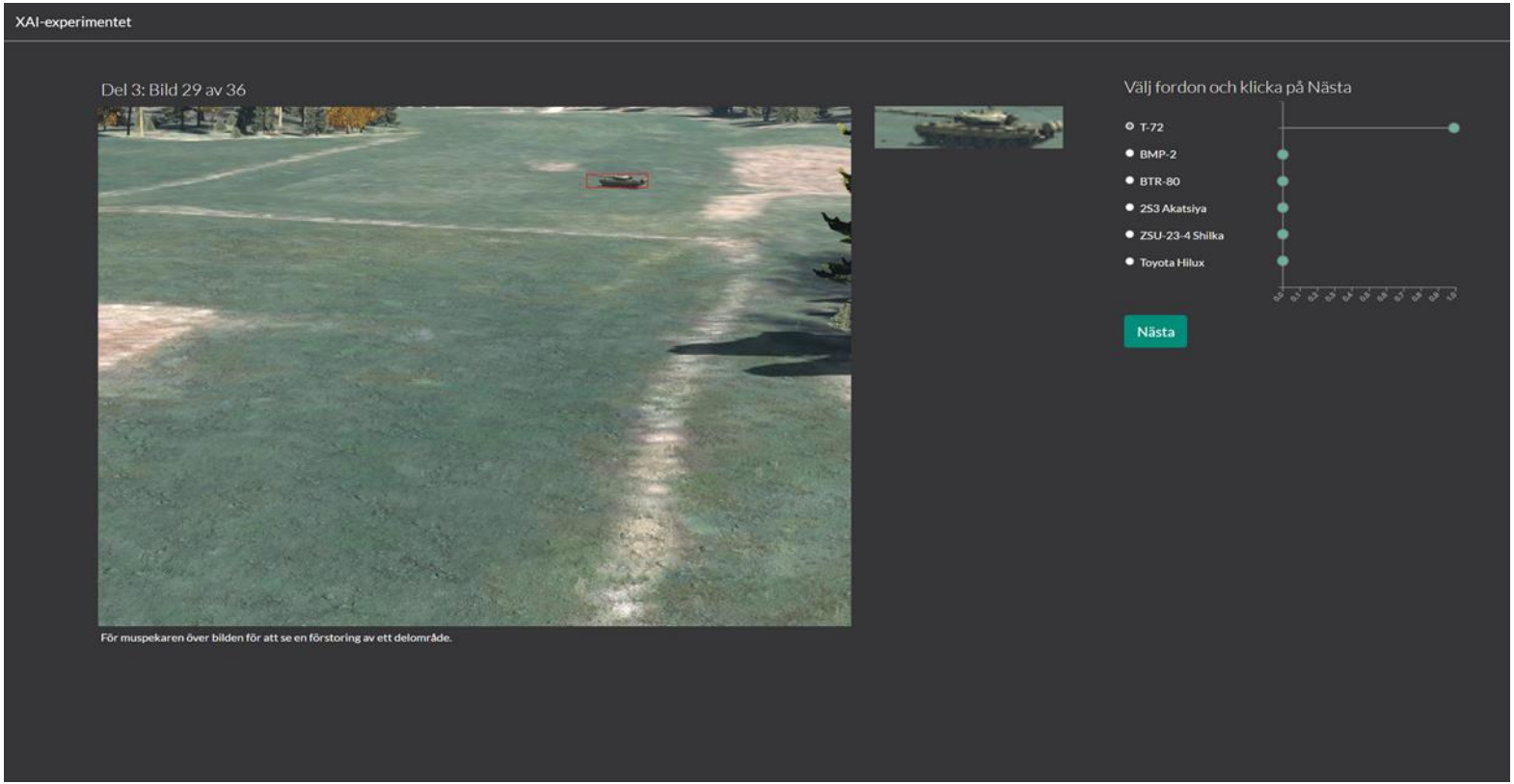
RISE saliency maps för fordonsklasser



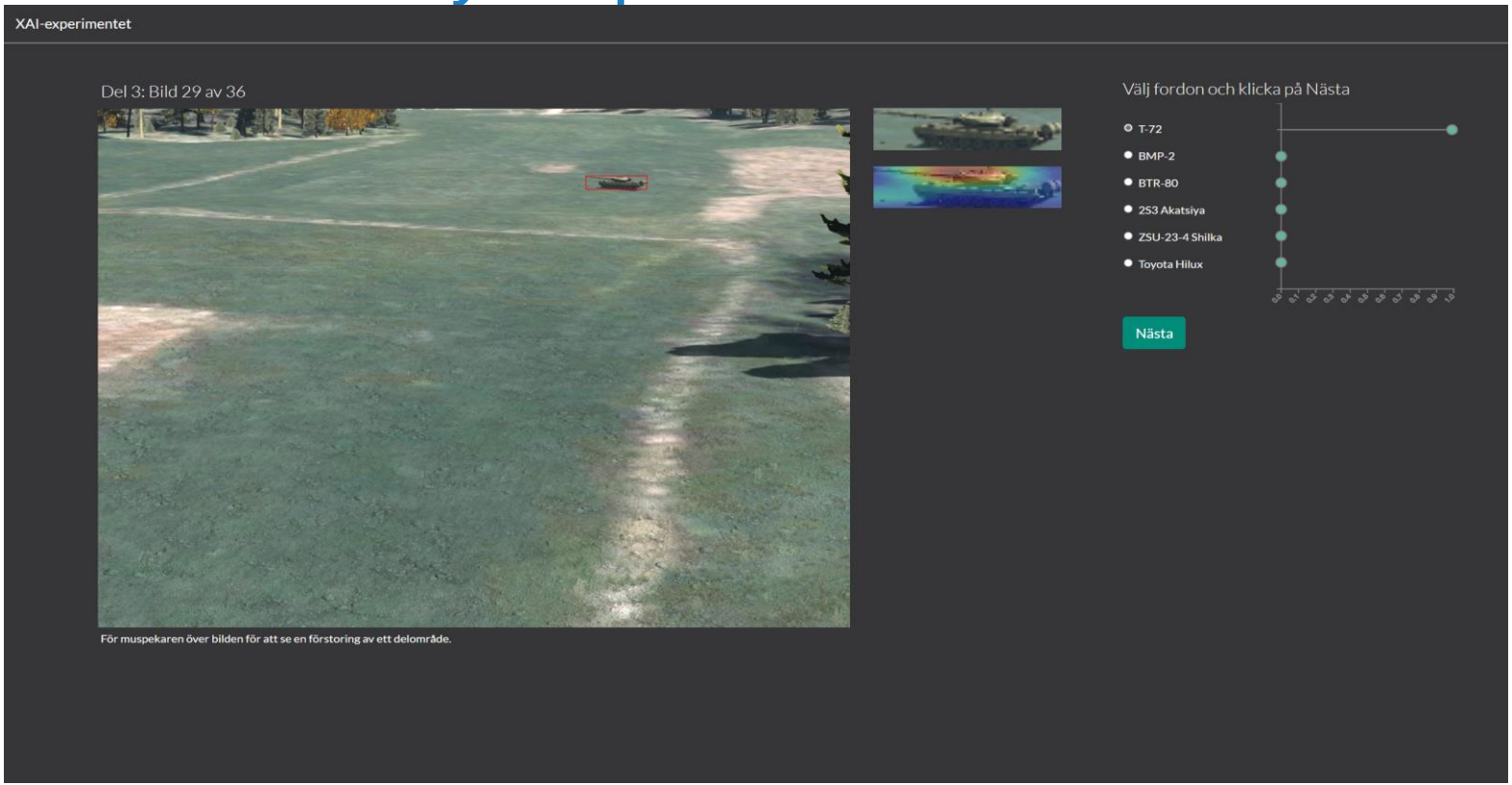
Exempeluppgift utan DNN-klassificerare



Exempeluppgift med DNN-klassificerare



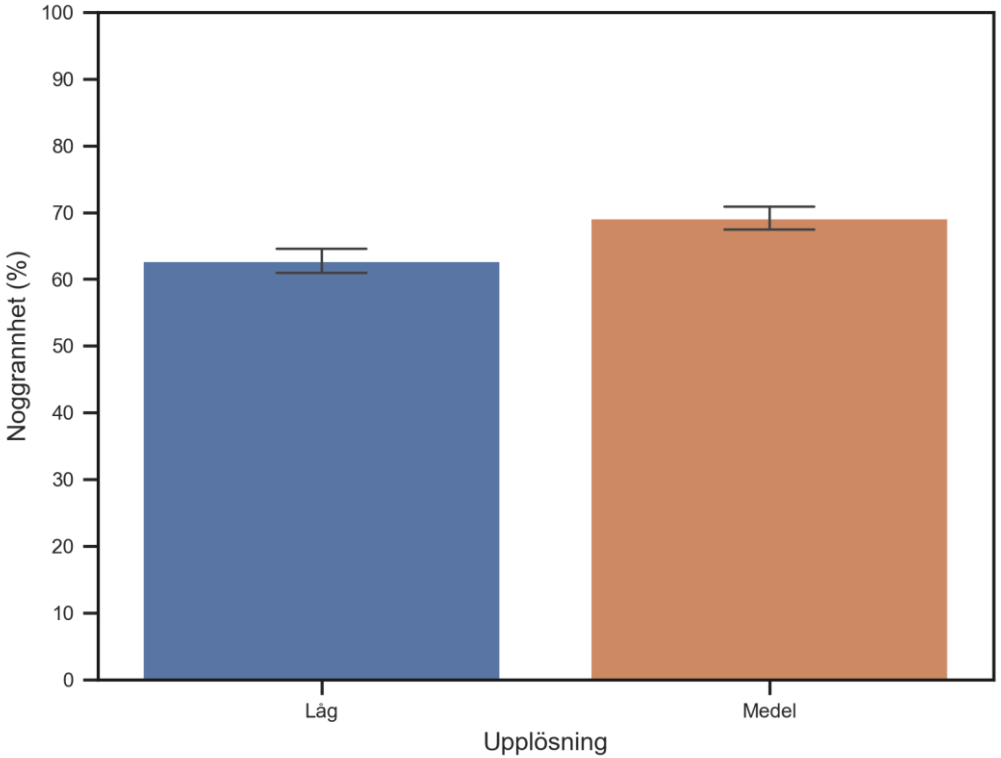
Exempeluppgift med DNN-klassificerare och RISE saliency maps



Hypoteser

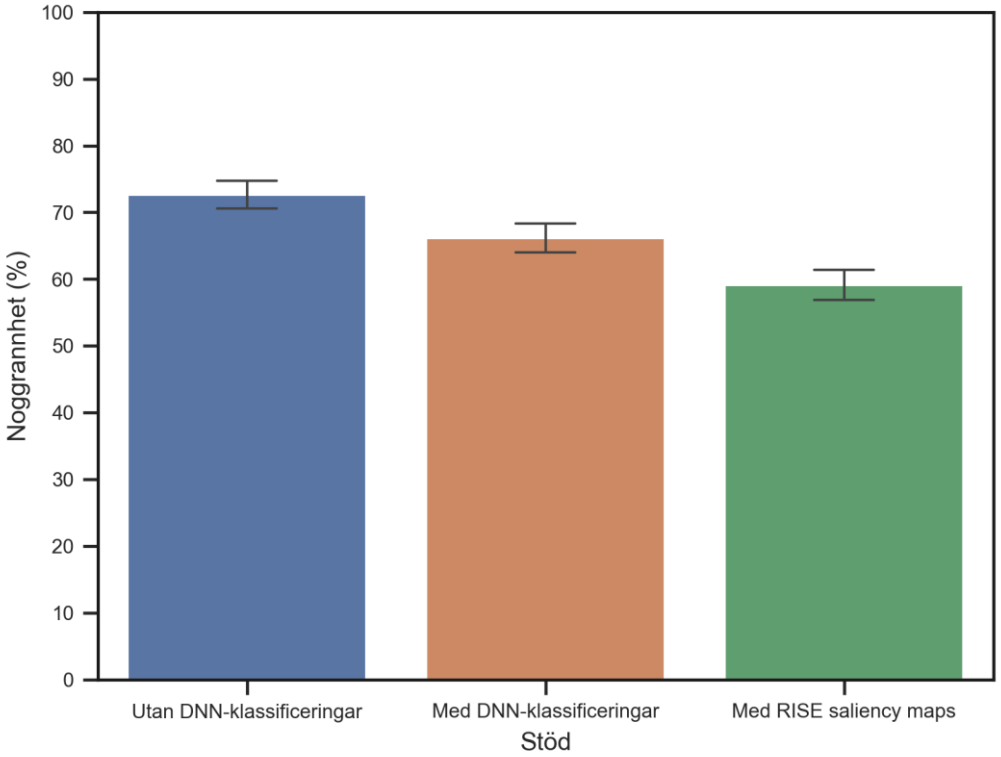
1. Högre noggrannhet med högre upplösning
2. Högre noggrannhet med stöd av DNN-klassificering
3. Högre noggrannhet med stöd av RISE saliency maps
4. Högre noggrannhet med stöd av RISE salency maps när DNN-klassificeringen är inkorrekt

Noggrannhet för målidentifiering



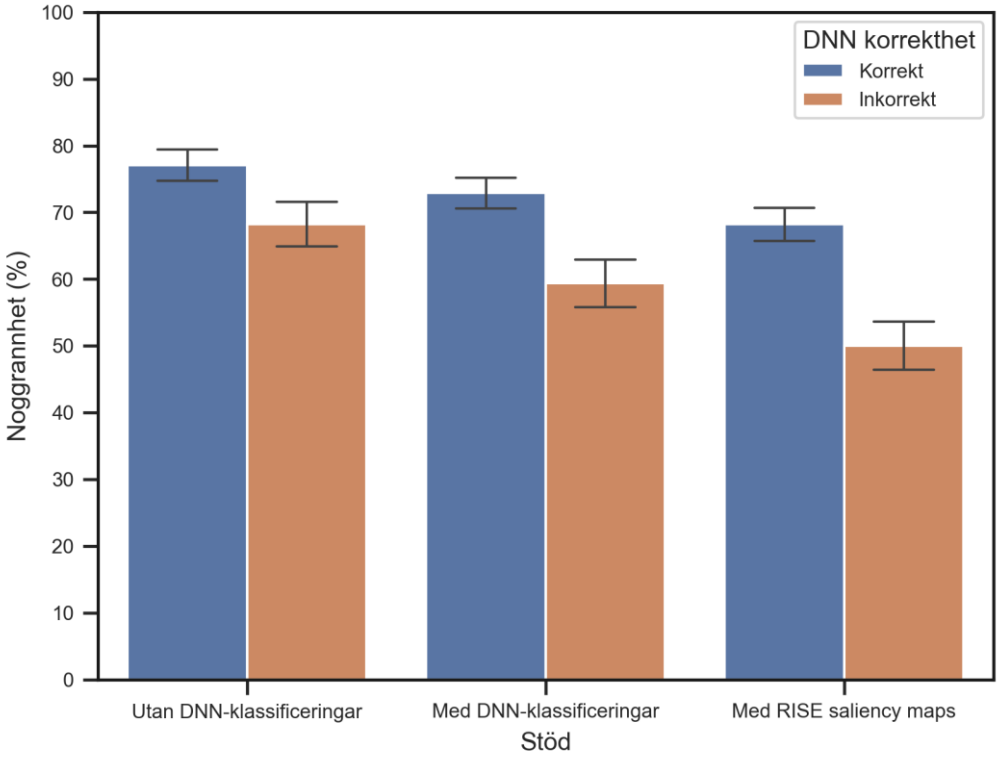
Enligt förväntningarna är noggrannheten högre med högre upplösning

Noggrannhet för målidentifiering



Tvärtemot förväntningarna minskar noggrannheten med stöd av DNN-klassificeringar och RISE saliency maps

Noggrannhet för målidentifiering



Tvärtemot förväntningarna minskar noggrannheten med RISE saliency maps när klassificeringen är inkorrekt

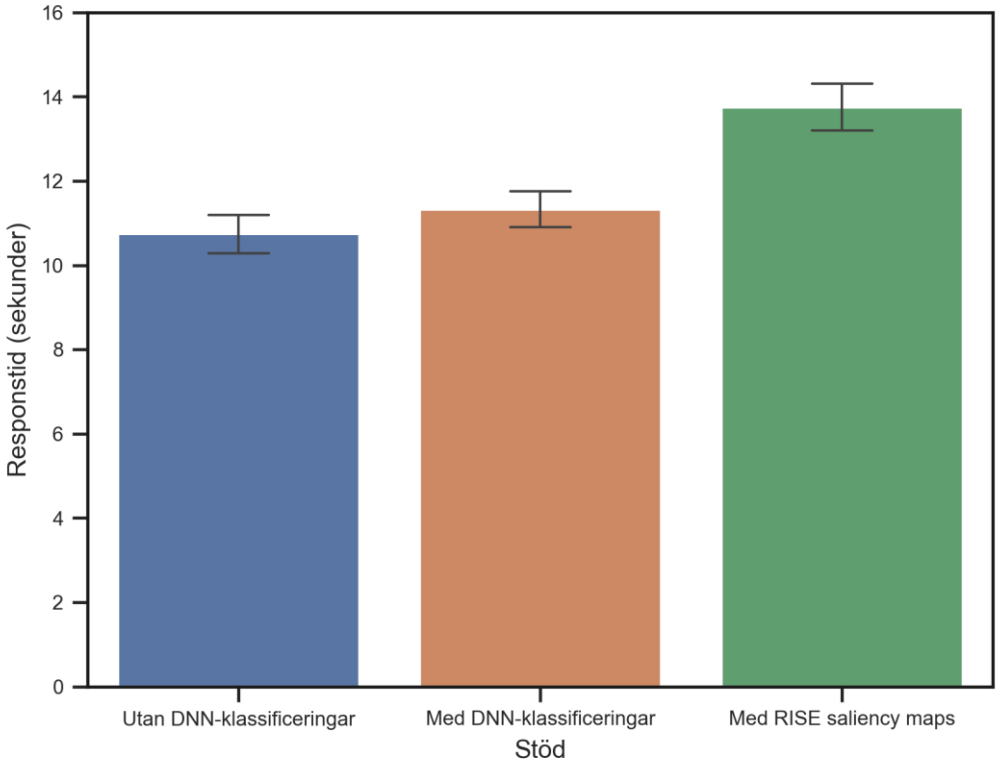
FOI

Noggrannhet för målidentifiering

Orsaker till minskad noggrannhet med RISE saliency maps

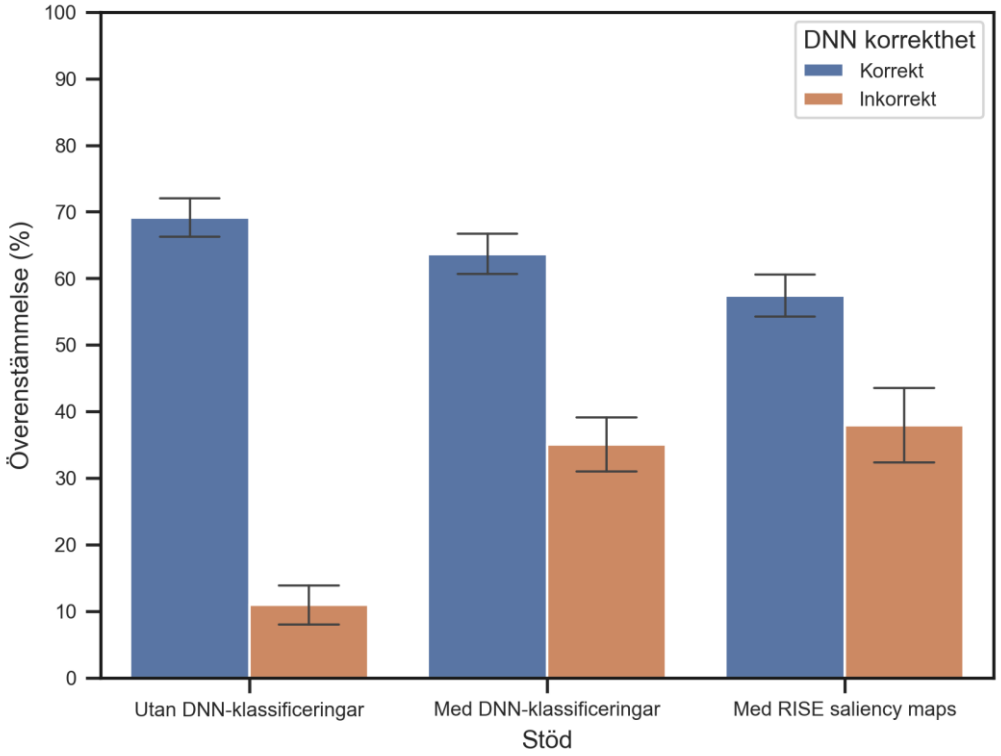
- Minskad noggrannhet när klassificeringen är inkorrekt
- Oproportionerligt minskad noggrannhet för T-72

Responstid för målidentifiering



Deltagarna försöker använda RISE saliency maps eftersom responstiden ökar

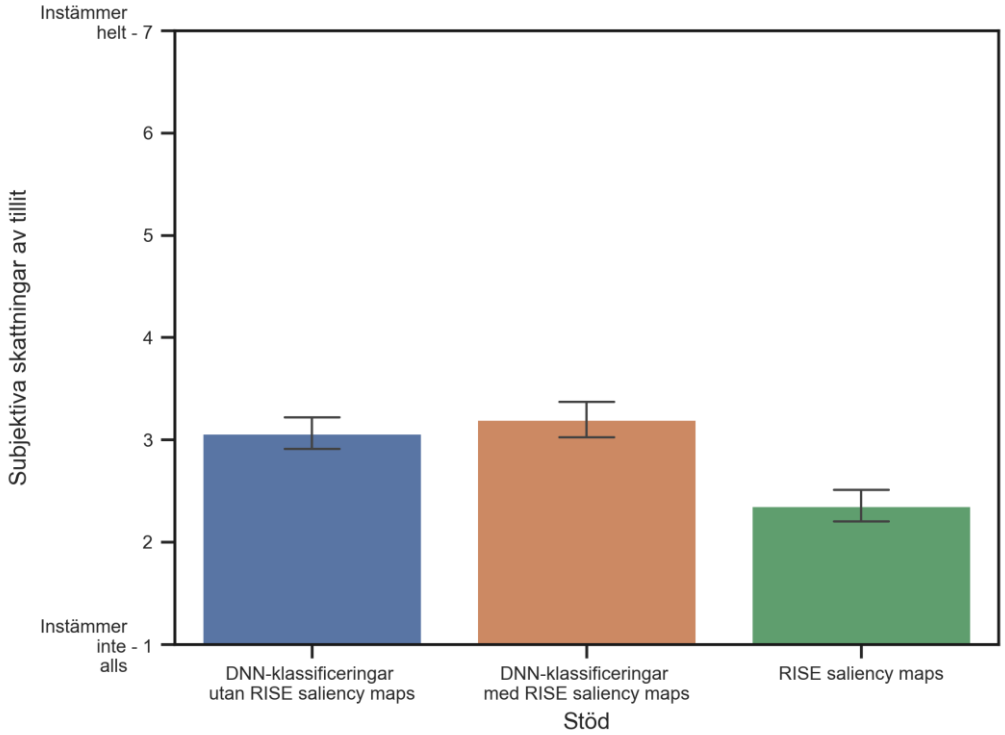
Överenstämmelse med DNN-klassificering



Deltagarna har svårt att bedöma DNN-klassificeringens tillförlitlighet

- Tvärtemot förväntningarna minskar överenstämmelsen med ökat stöd när DNN-klassificeringen är korrekt
- Tvärtemot förväntningarna ökar överenstämmelsen med ökat stöd när DNN-klassificeringen är inkorrekt

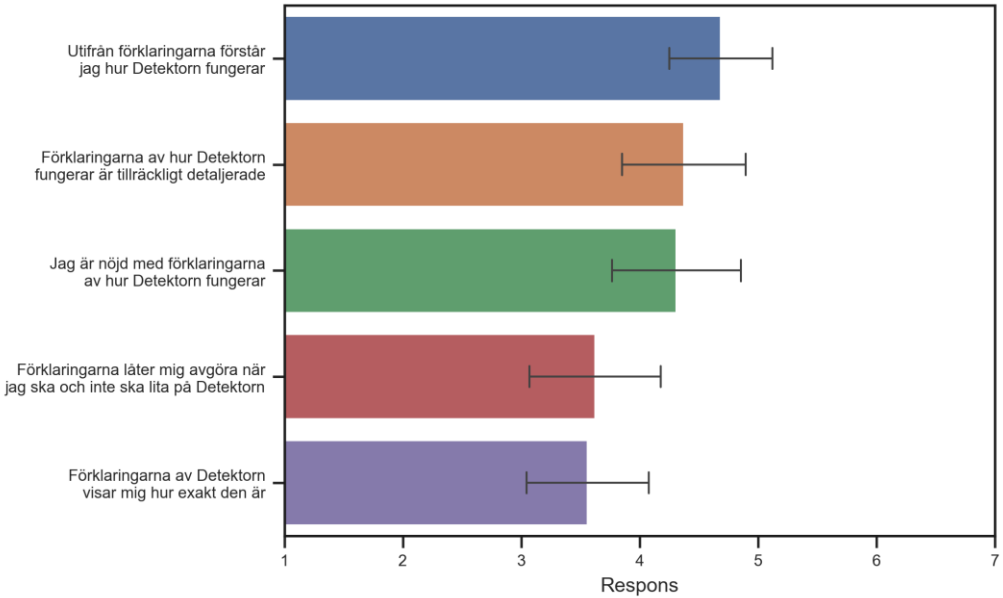
Tillit till stöd



Låg tillit till DNN-klassificering och RISE saliency maps

Särskilt låg skattning för påståendet ’Jag tycker om att använda RISE saliency maps för beslutsfattande’

Nöjdhet med RISE saliency maps



Varken nöjd eller missnöjd med RISE saliency maps

Sammanfattning av resultaten

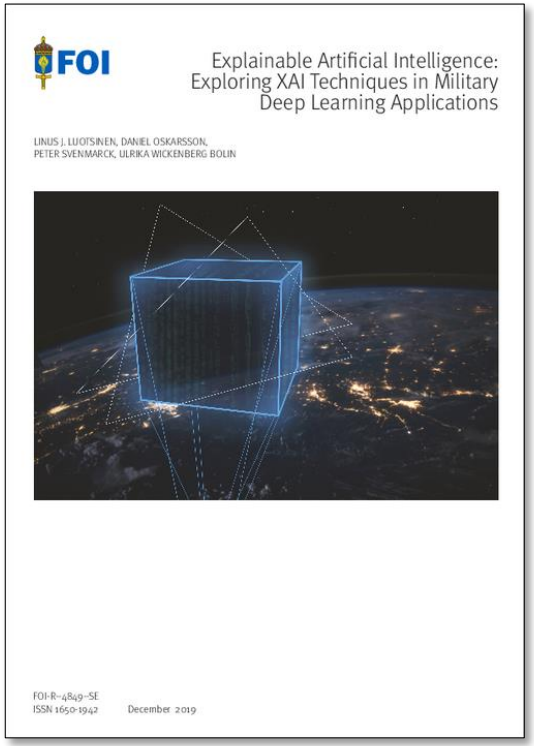
- Högre noggrannhet med högre upplösning
- Lägre noggrannheten med DNN-klassificering och RISE saliency maps
- Lägre noggrannhet när DNN-klassificeringen är inkorrekt
- Försöker använda DNN-klassificeringen och RISE saliency maps
- Svårt att bedöma tillförlitligheten för DNN-klassificeringen både utan och med RISE saliency maps
- Låg tillit till DNN-klassificeringen

Slutsatser

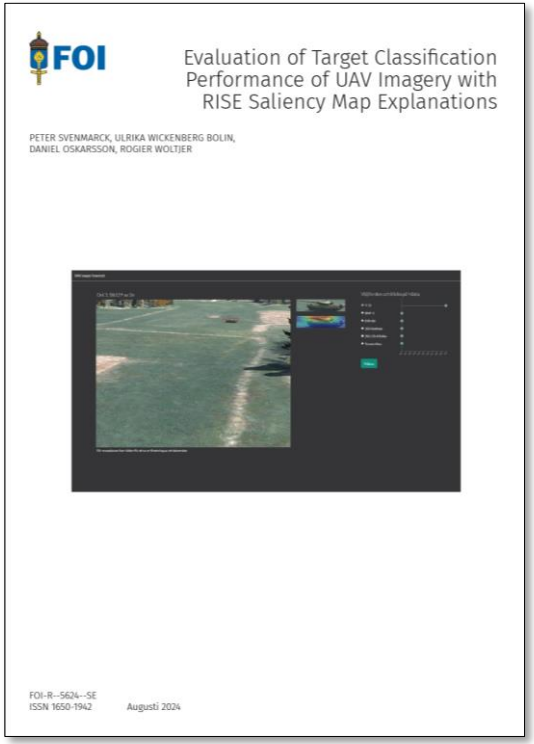
- Det är inte uppenbart hur XAI ska utformas för att förbättra måligenkänning
- Försämringen av måligenkänningen beror på:
  - Svårigheten att bedöma DNN-klassificeringens tillförlitlighet
  - Tendens att lita på automatiserade beslutsstöd

- Utvärdera andra XAI-metoder:
  - Konceptbaserade förklaringar
  - Liknande exempel
  - Liknande utsnitt
  - Kontrafaktiska
  - Generera textförklaringar

XAI rapporter 2020-2024



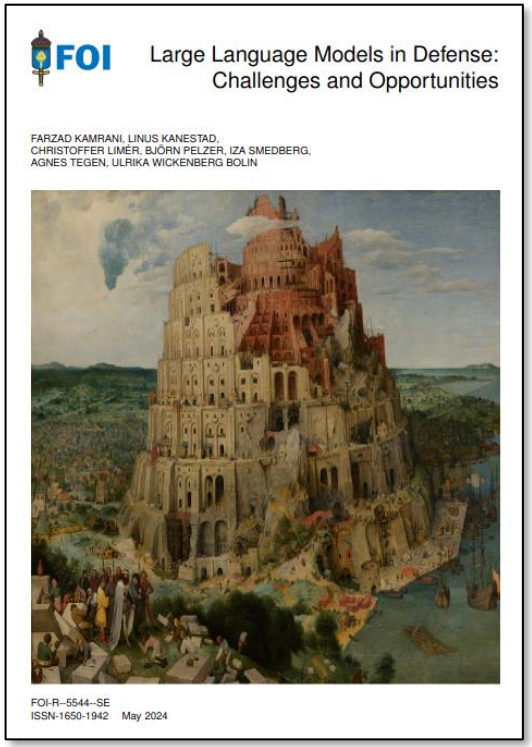
2020



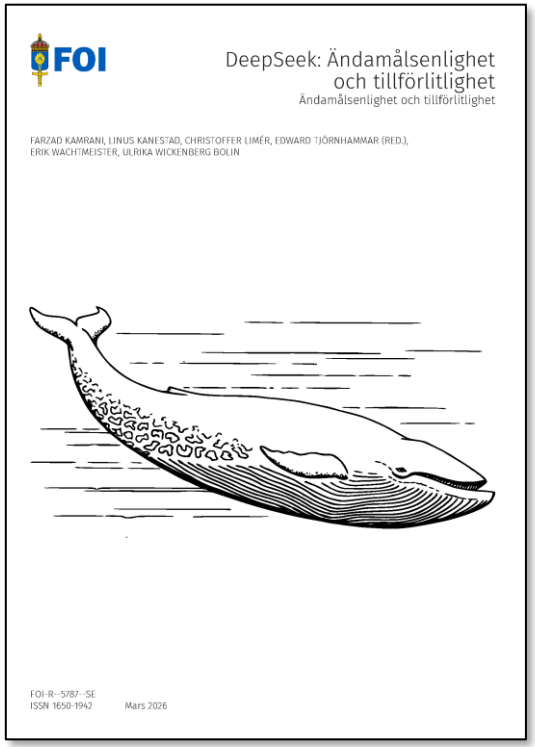
2024

FOI

Rapporter 2024-



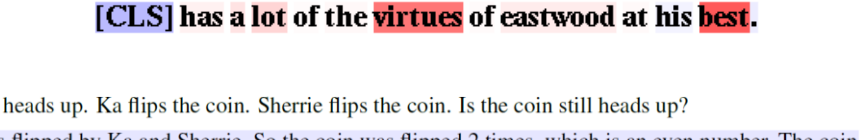
2024



2025

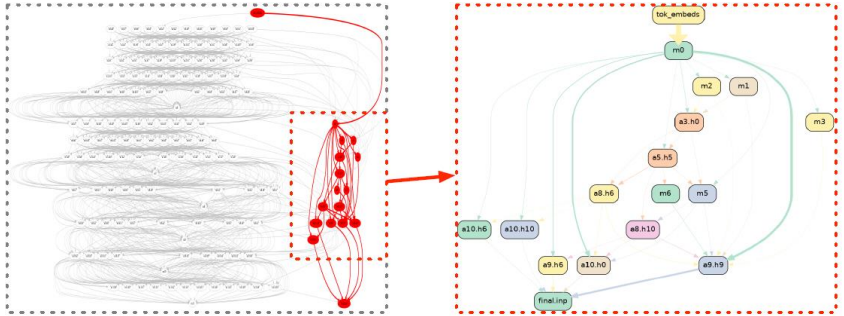
Exempel på XAI för LLM:er

- Feature attribution
- Chain of thought prompting
- Mechanistic interpretability
  - Circuits
  - Monosemanticity



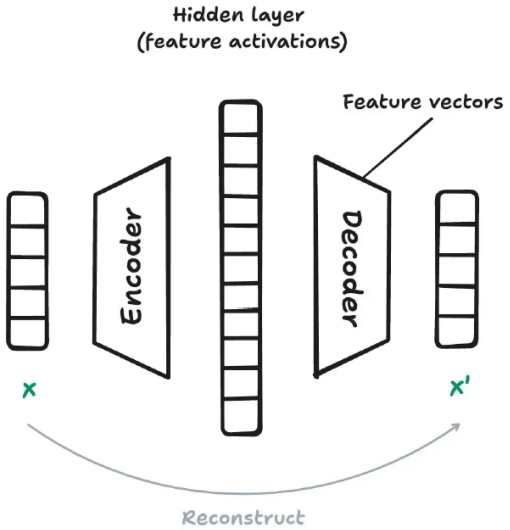
Q: Q: A coin is heads up. Ka flips the coin. Sherrie flips the coin. Is the coin still heads up?

A: The coin was flipped by Ka and Sherrie. So the coin was flipped 2 times, which is an even number. The coin started heads up, so after an even number of flips, it will still be heads up. So the answer is yes.



Monosemanticity

- LLM neuroner är polysemantiska
- Särdrag är en superposition av många LLM neuroner
- Extrahera tolkningsbara särdrag
  - Dictionary learning med sparse autoencoder för att särskilja särdrag



Monosemanticity exempel

Feature #34M/31164353 **Golden Gate Bridge** feature example

The feature activates strongly on English descriptions and associated concepts

in the Presidio at the end (that's the
huge park right next to the Golden Gate
bridge), perfect. But not all people

repainted, roughly, every dozen years."
"while across the country in san fran
cisco, the golden gate bridge was

it is a suspension bridge and has similar
coloring, it is often> compared to the
Golden Gate Bridge in San Francisco, US

They also activate in multiple other languages on the same concepts

ゴールデン・ゲート・ブリッジ、金門橋は、ア
メリカ西海岸のサンフランシスコ湾と太平洋が
接続するゴールデンゲート海

골든게이트 교 또는 금문교 는 미국 캘리포니아주
골든게이트 해협에 위치한 현수교이다. 골든게이
트 교는 캘리포니아주 샌프란시

мост золотые ворота — висячий мост
через пролив золотые ворота. он со
единяет город сан-франц

And on relevant images as well

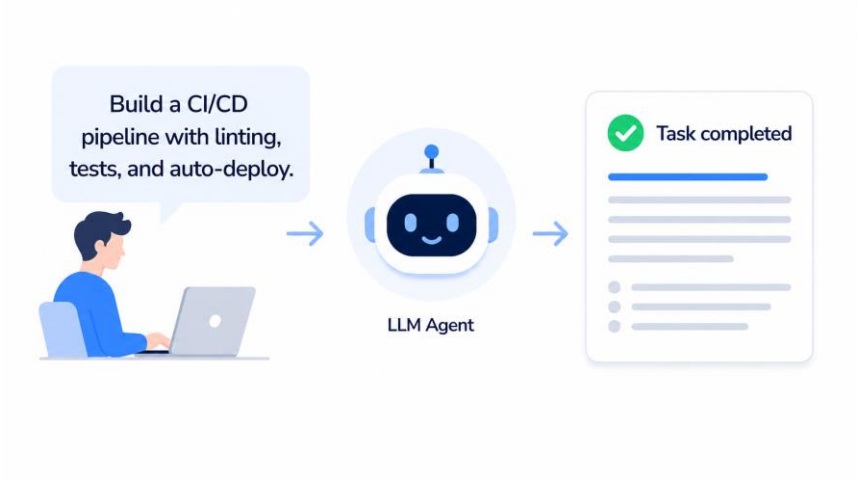
Bild: Templeton m.fl. (2024). Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet

LLM-baserade agenter

System som använder LLM med minne, planering, verktyg och kontrollfunktioner för att genomföra uppgifter i flera steg

Exempel:

- OpenAI Deep Research
- Claude Code
- OpenClaw



LLM-baserade multi-agent system (MAS)

System med flera LLM-baserade agenter som samarbetar för att genomföra uppgifter

Arkitekturer

- Sekventiell
- Hierarkisk
- Ensemble
- Peer-to-peer

Tillämpningar

- Parallella uppgifter
- Komplexa problem

Exempel

- Programutveckling
- Sjukvård
- Vetenskaplig forskning

FOI

MAS för- och nackdelar

Fördelar

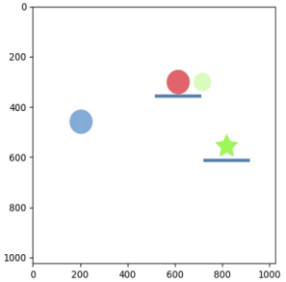
- Kan prestera bättre än enskilda agenter
- Mindre biasering

Nackdelar

- Grupptänkande
- Förstärker initiala fel
- LLM inte tränade för MAS

Explorativ studie av XMAS

- MAS för att skapa träningsscenarier
  - Stridsorder
- Test av genomförbarhet
  - MAS för att skapa fysikpussel
  - Planering och återmatning, designer, lösare, utvärderare
- XAI for MAS (XMAS)
  - Monosemanticity för semantisk analys av LLM agenterna
  - Analys av MAS dialoger
    - Grafanalys, tidsserieanalys, kausalanalys



Lästips

- Ali m.fl. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information fusion*, 99, 101805.

- ŞAHiN m.fl. (2025). Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning. *Neural computing and applications*, 37(2), 859-965.