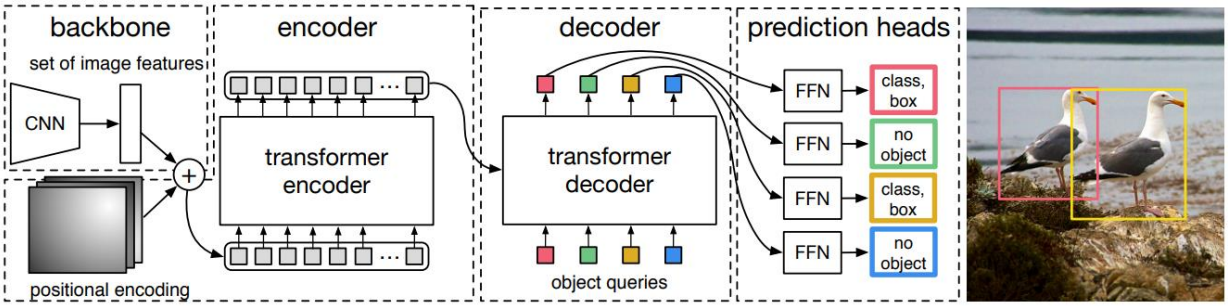
Institute for TMDT

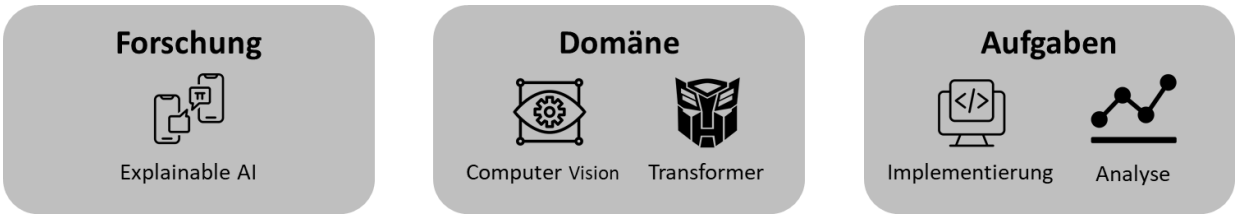
BERGISCHE UNIVERSITÄT WUPPERTAL

----- English version below -----

Ausschreibung Bachelor- / Masterarbeit

Interpretierbarkeit und Transparenz für Vision Transformer Modelle





Ausgangslage

Künstliche neuronale Netze sind seit etwa einem Jahrzehnt die erste Wahl von lernenden Modellen für die Verarbeitung von Bildern, Sprache und Text. Seit ihrer Erfindung im Jahr 2017 haben Transformer-Modelle Spitzenplätze auf Benchmarkdatensätzen in allen Bereichen des Deep Learnings belegt, von der Verarbeitung natürlicher Sprache (NLP) über Computer Vision (CV) bis hin zur Zeitreihenanalyse.

Problemstellung

Trotz der Fortschritte die durch Vision Transformer im Bereich Computer Vision erzielt werden konnten, gibt es ein entscheidendes Problem mit diesen Modellen: Ihre mangelnde Transparenz. Die führt vor allem bei unerwartet auftretenden Fehlern der Modelle dazu, dass es mit erheblichem Zeitaufwand verbunden ist die Ursachen zu finden. Auf ähnliche Problemstellungen stieß man in den Neurowissenschaften bei der Untersuchung biologischer Nervensysteme und entwickelte die Methodik der Ablationsstudien um sie anzugehen.

Vorgehensweise und Erwartete Ergebnisse

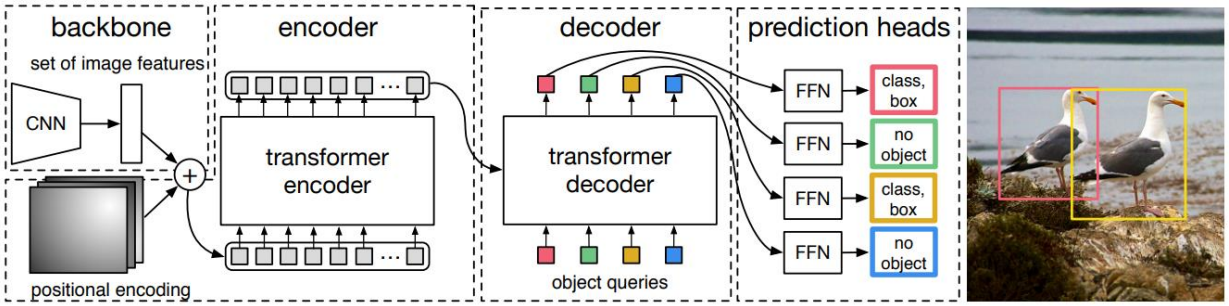
Der Fokus dieser Arbeit soll darauf liegen die Interpretierbarkeit von Vision Transformer Modelle durch die Adaption der neurowissenschaftlichen Methodik der Ablationsstudien zu erhöhen. Dazu erfolgt zunächst eine Einarbeitung in das Deep Learning Framework Pytorch und die Thematik der Vision Transformer. Darauf folgt die Erarbeitung von Konzepten wie Ablationen an Transformer Modellen umgesetzt werden können und die Erprobung der entwickelten Ideen an vortrainierten Modellen.

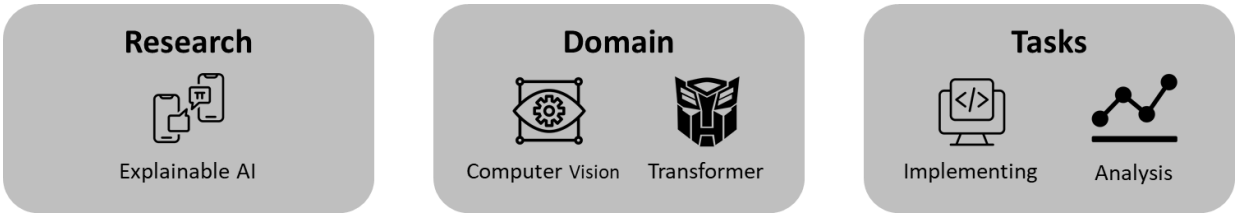
Ansprechpartner

Nils Hütten | **E-Mail:** nhuetten@uni-wuppertal.de

Bachelor- / Master thesis

Interpretability and Transparency of Vision Transformer Models





Initial Situation

Artificial neural networks have been the first choice of learning models for processing images, speech, and text for about a decade. Since their invention in 2017 Transformers have taken top spots on benchmarks in all areas of Deep Learning from natural language processing (NLP) over computer vision (CV) to time series analysis.

Problem Definition

Despite the progress that was induced through vision transformer in computer vision, there is a crucial problem with these models: Their lack of transparency, which makes finding the cause of unexpected errors very time consuming. Similar problems were encountered in the neurosciences when studying biological nervous systems. The focus of this work is to investigate vision transformer models by adapting the neuroscientific methodology of ablation studies.

Procedure and Expected Results

The focus of this work is to increase the interpretability of vision transformer models by adapting the neuroscientific methodology of ablation studies. To do so, the thesis includes the familiarization with the deep learning framework Pytorch and the topic of vision transformers. This is followed by the development of concepts how ablations can be implemented on Transformer models and the testing of the developed ideas on pre-trained models.

Contact Person

Nils Hütten| **E-Mail:** nhuetten@uni-wuppertal.de