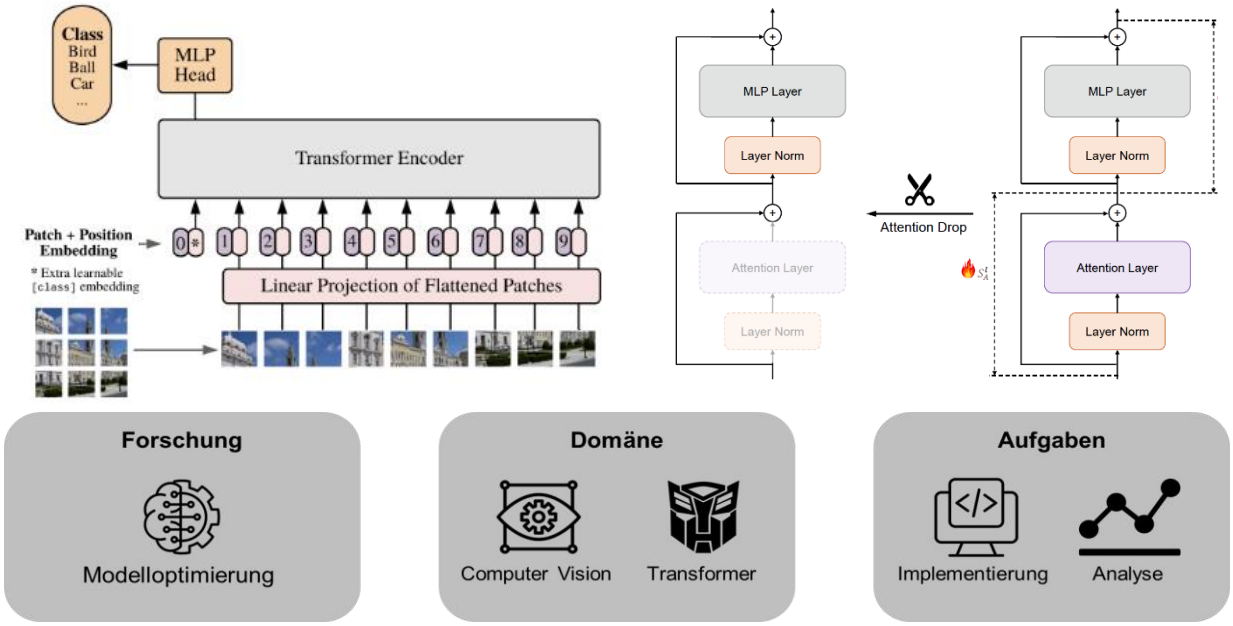
Institute for TMDT

BERGISCHE UNIVERSITÄT WUPPERTAL

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

Masterarbeit

Optimierung der Effizienz von großen Vision Transformer Modellen



Ausgangslage

Die Skalierung großer Sprachmodelle auf Transformer-Basis hat zwar vielversprechende Leistungen für verschiedene Aufgaben gezeigt, führt aber auch zu redundanten Architekturen, die eine Herausforderung für den Einsatz in der Praxis darstellen. Überraschenderweise wurde festgestellt, dass trotz der kritischen Rolle der Attention-Layer, die Transformer von anderen Architekturen unterscheiden, ein großer Teil dieser Layer eine übermäßig hohe Ähnlichkeit bzw. Redundanz aufweist und ohne Leistungseinbußen entfernt werden kann.

Problemstellung

Obwohl die Redundanz in LLMs in gewissem Maße untersucht ist, wurde die Variabilität der Redundanz über verschiedene Strukturen in Vision Transformern, wie MLP und Attention Layern, noch nicht systematisch untersucht. Darüber hinaus sind die Auswirkungen dieser Redundanzen auf die Modelleffizienz und die Leistungsabwägungen bei Computer Vision Aufgaben noch wenig bekannt, was eine genauere Untersuchung von auf diese Modelle zugeschnittene Pruning-Techniken erforderlich macht.

Vorgehensweise und Erwartete Ergebnisse

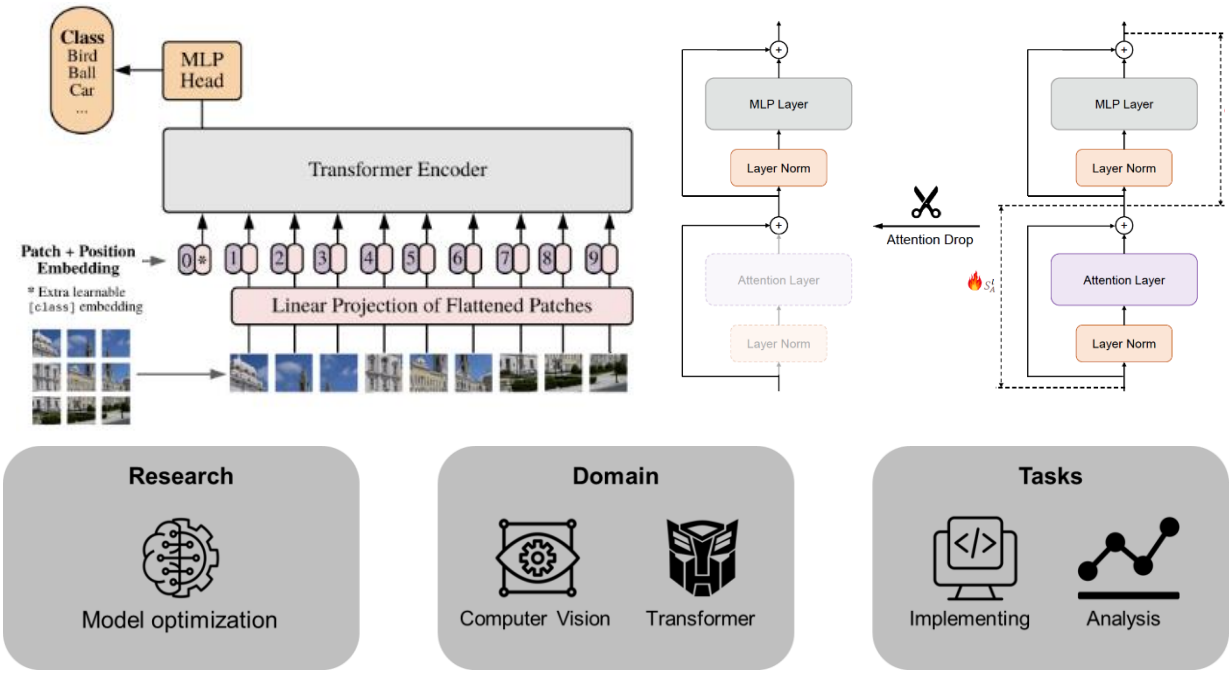
Der Schwerpunkt dieser Arbeit liegt auf der Untersuchung der Möglichkeit, die Effizienz von Vision-Transformer-Modellen zu verbessern, indem die für LLMs entwickelten Pruning-Strategien angepasst und möglicherweise erweitert werden. Zu diesem Zweck umfasst die Arbeit die Einarbeitung in das Deep Learning Framework Pytorch und die Thematik der Vision Transformer Modelle. Anschließend werden die für LLMs entwickelten Pruning-Konzepte aus der Literatur validiert und ihre Übertragbarkeit für (große) Computer Vision Modelle evaluiert.

Ansprechpartner

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

Master thesis

Optimizing the efficiency of large vision transformer models



Initial Situation

While scaling Transformer-based large language models has demonstrated promising performance across various tasks, it also introduces redundant architectures, posing efficiency challenges for real-world deployment. Surprisingly, despite the critical role of attention layers in distinguishing transformers from other architectures, it was found that a large portion of these layers exhibit excessively high similarity and can be pruned without degrading performance.

Problem Definition

Despite some recognition of redundancy in LLMs, the variability of redundancy across different structures in vision transformers, such as MLP and attention layers, has not yet been systematically investigated. Furthermore, the impact of these redundancies on model efficiency and performance trade-offs in vision-specific tasks remains poorly understood, necessitating a closer examination of pruning techniques tailored to these models.

Procedure and Expected Results

The focus of this work is to explore the possibility of improvements to the efficiency of vision transformer models by adapting, and possibly extending, the pruning strategies developed for LLMs. To do so, the thesis includes the familiarization with the deep learning framework Pytorch and the topic of vision transformers models. This is followed by validating pruning concepts developed for LLMs from literature and evaluating their feasibility for (large) vision models.

Contact Person

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