**Ruprecht-Karls-Universität Heidelberg**

**Medizinische Fakultät Mannheim**

**Dissertations-Kurzfassung**

# Efficient Deep Learning at Inference Time for Gram Stained Image Classification

Autor: Hee Kim

Institut / Klinik: Medizinische Statistik, Biomathematik und Informationsverarbeitung

Doktorvater: Prof. Dr. T. Ganslandt

Deep learning (DL) and artificial intelligence (AI) are woven into the fabric of our daily lives, and they also hold/have shown promise in the medical domain. Despite numerous studies published in the last decade regarding AI application in medicine, DL models have yet to be widely implemented in daily clinical practice on a large scale. In the face of numerous obstacles on the path to a thriving healthcare AI landscape, this dissertation focuses specifically on technical issues related to constrained hardware resources. To address this problem, in this doctoral thesis, I investigated and demonstrated optimal DL techniques based on the use case of Gram-stain analysis for microorganism identification.

Efficient DL techniques such as transfer learning, pruning and quantization can be employed during model training and deployment strategies should be considered in advance. Particularly, I advocate for applying transfer learning to pre-trained models as feature extractors, as opposed to introducing novel model architectures. For Gram-stain classification, DL models could be compressed and test-time performance could be accelerated without compromising test accuracy or loss. While pruning contributed to the reduction in model size by 15×, quantizing the bit representation from 32-bit to 8-bit led to accelerated inference times by 3×. Taking into the quantization configuration, the findings demonstrated that quantization per channel outperformed tensor-wise quantization for the majority of DL models. This outcome contradicts conventional assumptions, however, intensive quantization may potentially hinder the generalization of DL models. Therefore, the most optimal configuration of DL models should be empirically determined depending on the custom task and data. In the majority of setups, vision transformers (VT) exhibited superior model performance compared to convolutional neural networks (CNN). Notably, among these configurations, DeiT tiny emerged as the fastest VT model in int8 configuration, processing six images per second.

By harnessing the investigated efficient DL techniques including transfer learning, pruning and quantization, this doctoral research might provide valuable insights for AI researchers to accelerate the pace of innovation in the medical domain and pave the way for the seamless integration of AI into everyday healthcare practices.