Abstract der fertigen Arbeit

Serverless computing is a new paradigm for the deployment of applications in the cloud. Its extension to the field of Machine Learning has been a topic of research for the past few years—especially in the serverless edge computing community. Seamless integration of Machine Learning models for usage in serverless computing environments is still a challenge. Selecting appropriate models for particular use cases is crucial for the performance of the whole system. In other domains, service selection and load balancing with quality-of-service consideration have been used to improve the performance of applications. However, current approaches to service selection and load balancing either do not take into account the specifics of Machine Learning models or the latency of edge nodes. So, developers of serverless applications currently have to manually select the best model for their use case or rely on subpar model selection approaches. Manually selecting the models limits the flexibility of the application and requires the developer to have knowledge about the underlying infrastructure. In this thesis, we propose a solution that automatically and transparently selects the best model for a given usecase provided in the underlying infrastructure of serverless platforms while taking into account Machine Learning and edge computing-specific concerns. Application developers then only have to provide the data and the desired performance traits (e.g., latency, accuracy, etc.) of the model. We evaluate our solution by comparing it to common baseline approaches for service selection and balancing. First, we conduct case studies to evaluate the performance of our solution in different use cases. Then, we compare the performance of our solution to the baseline approaches in a simulated environment. The results show that our solution outperforms the baseline approaches in most cases.