Linear and Logarithmic Quantization Approaches for Efficient Inference with Deep Neural Networks

Student:	Constantin Berger	**Abstract**: Quantization enables efficient processing of Deep Neural Networks. In this work, the methods of Linear and Logarithmic Quantization are discussed. These methodologies are applied to a Deep Neural Network for controlling an autonomous drone. The trade-off between reduction of computational complexity and loss of accuracy is the main subject of this investigation. Moreover, I propose an approach to overcome the limitations of logarithmic quantization, which requires the specific handling of negative values. This is achieved by storing the sign of the non-quantized value in the sign-bit of the fixed-point value representation after Quantization. This approach allows the application of Logarithmic Quantization to Neural Networks with positive and negative weights. The results show that the given hardware does not allow for significant performance improvements.
Email:	-	
Status:	FINISHED	
Supervisor:	Matthias Kissel	

Files

Documentation

-

Workflow

Start

- [ ] Topic specification
- [ ] Definition of work packages
- [ ] Composition of a project proposal and time plan
- [ ] Project Talk with Prof. Diepold
- [ ] Registration of the thesis
- [ ] Creation of a wiki page (supervisor)
- [ ] Creation of a gitlab repository or branch
- [ ] Access to lab and computers

Finalization

- [ ] Check code base and data
- [ ] Check documentation

- [ ] Provide an example notebook that describes the workflow/usage of your code (in your repo)
- [ ] Proof read written composition
- [ ] Rehearsal presentation
- [ ] Submission of written composition
- [ ] Submission of presentation
- [ ] Recording of presentation / Presentation in group meeting
- [ ] Final Presentation