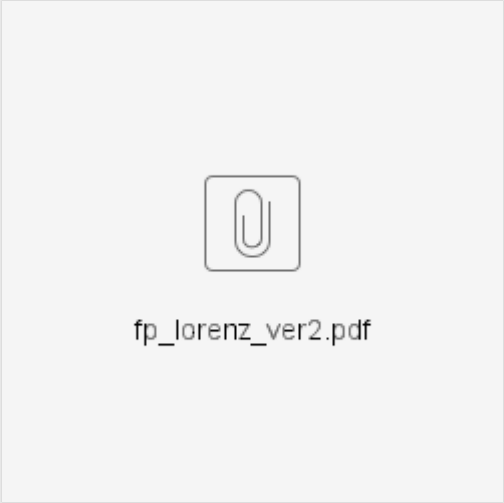
# Neural Network Online-Pruning: Accelerating Weighted Sum Calculation by Early Stopping

| | | |
|---|---|---|
| **Student:** | Julian Lorenz | **Abstract**: In this paper I propose a new method to shorten the weighted sum computation at each neuron in a neural network without requiring retraining. I sort the weighted sum computation order by the magnitude of the weights. If the activation function shows converging behavior, I stop the weighted sum computation early after it has passed a predetermined stopping threshold. I show how to find the stopping thresholds by statistical analysis of the weighted sum computation in a network. I also provide an experimental analysis on how the online-pruning method performs in comparison to the normal feed-forward computation. Using my approach, the MAC operations in the tested network can be reduced by 14.1%. This results in a speed improvement of 5.1% while achieving an average R2 score of 99.09%. |
| **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