A comparison of sparse representations of low-dimensional data

Despite the great prevalence of linear layers in modern neural networks, studies have shown that up to 95% of their parameters are redundant (Denil et al. 2013, Gong et al. 2014, Sainath et al. 2013) (1., 2., 3.), which is extremely wasteful of both computer memory and processing, as well as presenting a dangerous potential for model overfitting, which can reduce performance of the model as well as pose privacy risks (Nasr et al. 2018, Klas Leino & Matt Fredrikson 2019) (4.) and produce unexpected performance hits on edge cases. To combat this, several Structured Efficient Linear Layers have been proposed, including ACDC (Moczulki et al. 2016) (5.), Fast Random Projections (Ailon & Chazelle 2009) (6.), and simpler methods such as linear autoencoders, dilated convolutions, pooling layers, Johnsonn Lindenstrauss, and so on.

This project will focus on very low-dimensional data for a variety of reasons including ease of computation, ease of visualization, and the author's curiosity. Various sparse representations of the data will be compared and the performances and costs will be presented. It is noteworthy that the networks will be very simple and not optimized for the data at hand, so poor performances are to be expected, however the author considers this acceptable because only the relative representational efficiencies are to be considered here, and not the absolute success of the model.

The data that will be considered here will only be the daily/hourly stock prices of the S&P 500 ETF over the course of 20 years (Kaggle dataset available at https://www.kaggle.com/pdquant/sp500-daily-19862018), with the goal of learning patterns that will help it predict future ETF prices from past prices alone. This project is also interesting for the purpose of learning more about the most important factors to consider in stock trading, and the return on investment that would be earned by an investor following this model's predictions will be presented, with any net positive gain to be considered a successful investment.

The deliverables of the project will consist of a practical part consisting of code and accuracy results on the considered data upon trying out the various architectures on it, and a theoretical part consisting of a summary of the following papers:

- Predicting Parameters in Deep Learning, by Misha Denil et. al (2013)
- ACDC: A Structured Efficient Linear Layer, by Moczulski et. Al (2016)
- Sparse Linear Networks with a Fixed Butterfly Structure: Theory and Practice, by Ailon et. al (2021)

Bibliography:

1. Denil, Misha, Shakibi, Babak, Dinh, Laurent, Ranzato, Marc'Aurelio, and de Freitas, Nando. Predicting parameters in deep learning. In NIPS, pp. 2148–2156, 2013.
2. Gong, Yunchao, Liu, Liu, Yang, Ming, and Bourdev, Lubomir. Compressing deep convolutional networks using vector quantization
3. Sainath, Tara N., Kingsbury, Brian, Sindhwani, Vikas, Arisoy, Ebru, and Ramabhadran, Bhuvana. Low-rank matrix factorization for deep neural network training with high-dimensional output targets. In ICASSP, pp. 6655–6659, 2013.
4. Klas Leino, Matt Fredrikson. Stolen Memories: Leveraging Model Memorization for Calibrated White-Box Membership Inference, 2019
5. Marcin Moczulski, Misha Denil, Jeremy Appleyard, Nano de Freitas. ACDC: A Structured Efficient Linear Layer. 2016
6. Ailon, Nir and Chazelle, Bernard. The Fast Johnson Lindenstrauss Transform and approximate nearest neighbors. SIAM Journal on Computing, 39(1):302–322, 2009.