Research Internship

Teun Peeters
s1003465, teun.peeters@ru.nl

Data Science Group, ICIS
Supervisor: Dr. Twan van Laarhoven

Background

During my internship I will be implementing a method for estimating linearity of Deep Neural Networks (DNNs). Linearity is not a topic that has been extensively researched in DNNs, although some papers have examined effects of the degree of linearity. Serra et al. (2018) finds that the number of linear regions in a DNN with rectifier units can be counted and shows a correlation between the number of linear regions and model performance. Hu et al. (2020) in a similar method uses the linear regions to prevent overfitting. Inversely, Bouniot et al. (2023) examines non-linearity in DNNs, proposing a metric that quantifies the non-linearity of a transformation and shows that it can predict model performance.

I will furthermore attempt to predict layer-wise contribution using the linearity measure, based on the hypothesis that linearity influences model performance and can be measured per layer. This is related to layer-wise relevance propagation, which has been studied more extensively (Montavon et al. 2019).

Carlini and Wagner (2017) and Goodfellow et al. (2015) argue that local linearity in neural networks enables the use of adversarial examples. I will attempt to examine this effect and see if there is a correlation between global linearity and this local linearity. Considering the loss, local linearity in fact contributes to adversarial robustness (Qin et al. 2019). A similar effect is used in Sharpness-Aware Minimization to increase model performance and robustness (Foret et al. 2021).

Research activities

1. Implement the linearity measure over a simple fully connected model
2. Investigate the correlation between linearity and model performance
3. Implement a layer-wise linearity measure
4. Investigate the correlation between layer-wise linearity and model performance
5. Investigate the correlation between linearity and model robustness
6. Extend the linearity measure to other models, such as CNNs, LSTMs, and RNNs

Planning

The start date of the internship was 4 March 2024, and the end date is 12 July. During the internship I will work an average of 3 days per week.

Approval

This uploaded proposal was (not yet) approved by dr. Twan van Laarhoven.

References

Bouniot, Quentin et al. (Oct. 2023). *Understanding deep neural networks through the lens of their non-linearity.* arXiv:2310.11439 [cs, stat]. DOI: `10.48550/arXiv.2310.11439`. URL: `http://arxiv.org/abs/2310.11439` (visited on 05/03/2024).

Carlini, Nicholas and David Wagner (Mar. 2017). *Towards Evaluating the Robustness of Neural Networks.* arXiv:1608.04644 [cs]. DOI: 10.48550/arXiv.1608.04644. URL: http://arxiv.org/abs/1608.04644 (visited on 05/03/2024).

Foret, Pierre et al. (Apr. 2021). *Sharpness-Aware Minimization for Efficiently Improving Generalization.* arXiv:2010.01412 [cs, stat]. DOI: 10.48550/arXiv.2010.01412. URL: http://arxiv.org/abs/2010.01412 (visited on 05/03/2024).

Goodfellow, Ian J., Jonathon Shlens and Christian Szegedy (Mar. 2015). *Explaining and Harnessing Adversarial Examples.* arXiv:1412.6572 [cs, stat]. DOI: 10.48550/arXiv.1412.6572. URL: http://arxiv.org/abs/1412.6572 (visited on 05/03/2024).

Hu, Xia et al. (Aug. 2020). "Measuring Model Complexity of Neural Networks with Curve Activation Functions". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* KDD '20. New York, NY, USA: Association for Computing Machinery, pp. 1521–1531. ISBN: 978-1-4503-7998-4. DOI: 10.1145/3394486.3403203. URL: https://doi.org/10.1145/3394486.3403203 (visited on 05/03/2024).

Montavon, Grégoire et al. (2019). "Layer-Wise Relevance Propagation: An Overview". en. In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning.* Ed. by Wojciech Samek et al. Cham: Springer International Publishing, pp. 193–209. ISBN: 978-3-030-28954-6. DOI: 10.1007/978-3-030-28954-6_10. URL: https://doi.org/10.1007/978-3-030-28954-6_10 (visited on 05/03/2024).

Qin, Chongli et al. (Oct. 2019). *Adversarial Robustness through Local Linearization.* arXiv:1907.02610 [cs, stat]. DOI: 10.48550/arXiv.1907.02610. URL: http://arxiv.org/abs/1907.02610 (visited on 05/03/2024).

Serra, Thiago, Christian Tjandraatmadja and Srikumar Ramalingam (Sept. 2018). *Bounding and Counting Linear Regions of Deep Neural Networks.* arXiv:1711.02114 [cs, math, stat]. DOI: 10.48550/arXiv.1711.02114. URL: http://arxiv.org/abs/1711.02114 (visited on 05/03/2024).