Bernhard Geiger

Information Theory for Sustainable Deep Learning

Information-theoretic analyses of neural networks not only open the black box of deep learning but could also make it environmentally friendly.

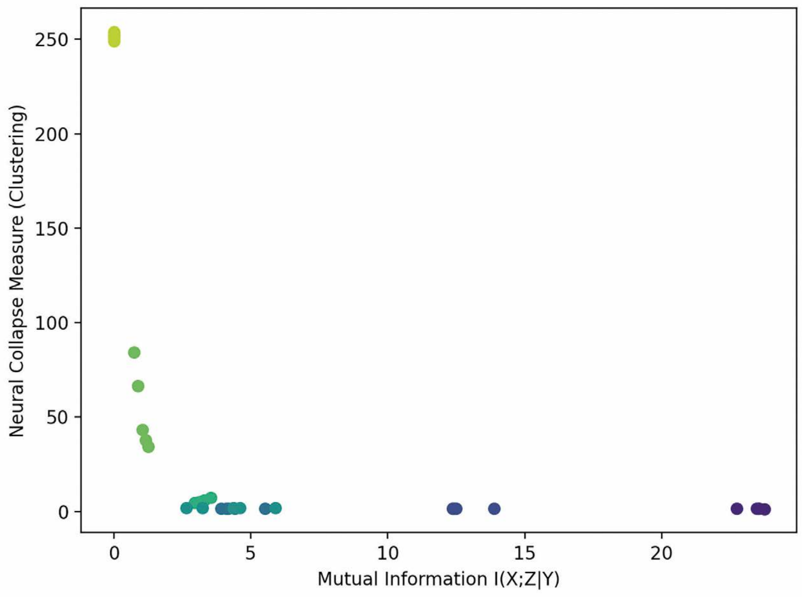
Neural networks (NNs) are undoubtedly the cornerstones of contemporary AI, with convolutional, recurrent, and transformer architectures helping us solve a myriad of everyday tasks. Especially large-scale architectures, as are common in foundation models or large language models, consume a substantial amount of energy during operation, resulting in a nonnegligible environmental footprint of AI. At the same time, the inner workings of deep learning architectures are still not fully understood, effectively limiting their trustworthiness.

There are two popular hypotheses in machine learning theory that could help address both shortcomings. On the one hand, the information bottleneck hypothesis claims that latent representations which are compressed, i.e., which are minimum sufficient statistics, reduce the chance that the NN overfits. Further, it was claimed (and disputed) that such compression happens inherently due to stochastic optimization of the NN parameters. The lottery ticket hypothesis, on the other hand, explains why overparameterized NNs are successful: redundancy in the network architecture increases the chance that parts of the network are initialized with weights that facilitate successful training.

Both hypotheses try to explain generalization performance with notions that have information-theoretic quantities as counterparts: compression and redundancy. While compression can be quantified by the mutual information between the input to the NN and the learned latent representation, the redundancy of this latent representation is obtained via partial information decomposition (PID). Furthermore, both hypotheses have implications for energy-efficient deep learning. Redundant parts of an NN architecture can be pruned without affecting performance, thus reducing inference costs. Simultaneously, monitoring and enforcing compression during training could improve the effectiveness of early stopping, effectively reducing training complexity.

While there is mathematical proof of the lottery ticket hypothesis, the information bottleneck hypothesis is subject to an ongoing debate. Specifically, it is not clear whether information-theoretic compression is different from geometric compression (in the sense of clustering) and how these two forms of compression are connected to generalization. Preliminary results obtained together with Linara Adilova, a visiting researcher within project ENFIELD, indicate that clustering and information-theoretic compression are negatively correlated under certain circumstances (see Figure 1).

Figure 1: Neural collapse vs. mutual information of differently regularized latent representations at the end of training. Clustered representations (low neural collapse measure) are linked with large mutual information, while information-theoretically compressed representations (low mutual information) appear not to cluster. Results were obtained by training a stochastic DenseNet121 architecture on CIFAR-100 via minimizing the conditional entropy bottleneck functional. Colors correspond to different regularization weights. Source: Dr. Linara Adilova, CC-BY-SA



For the class of stochastic NNs considered in this analysis, clustering proved to be more important for generalization performance than information-theoretic compression, thus presenting a piece of negative evidence for the information bottleneck hypothesis. Regarding PID, previous work has argued that early layers are characterized by high amounts of synergy (i.e., information from several neural units must be combined to retrieve information about the target), while deeper layers have increasing amounts of redundancy.

Since most of these previous results have been shown in comparably narrow settings, the general picture has not evolved yet. Thus, in our FWF project "Information Planes and Decomposition" (05.2025-04.2029) we will work towards establishing an in-depth understanding of the interplay between compression, redundancy, and generalization for a wide range of NN architectures, including binarized, Bayesian, and stochastic NNs with convolutional, recurrent, or transformer-based units (Figure 2). We will furthermore investigate how insights from our theoretical analyses can motivate novel approaches to reduce the environmental footprint of deep learning (e.g., via regularization schemes facilitating subsequent pruning of redundant NN units).

A major scientific challenge is not only the breadth of NN types that we plan to cover, but also the estimation of mutual information (which is also involved in PID) from high-dimensional latent representations. Prior work has shown that the qualitative behavior of mutual information estimates is heavily influenced by the properties of the estimator, sometimes even more so than by the latent representations themselves. A substantial part of the project work will thus be targeted at understanding the properties and limitations of existing mutual information estimators and adapting them to the deep learning setting.

Further reading: Geiger, "On Information Plane Analyses of Neural Network Classifiers -- A Review", TNNLS, arXiv:2003.09671; Adilova, Geiger, & Fischer, "Information Plane Analysis for Dropout Neural Networks", ICLR, arXiv:2303.00596.

Project ENFIELD (No. 101120657) is funded by the European Union's HORIZON Research and Innovation Programme. Know Center Research GmbH co-leads the task on Green AI.

The project "Information Planes and Decompositions" (No. PAT7753623) is funded by the FWF and is headed by the PI Bernhard Geiger and by Franz Pernkopf. Two PhD positions are funded by this project. Interested candidates are invited to reach out to geiger@tugraz.at.

Source: Caroline Wagner

Bernhard Geiger is an assistant professor at the Signal Processing and Speech Communication Laboratory and holds the tenure track position for hybrid physics-based and data-driven modeling and simulation of complex mechanical or electrical engineering systems. He is also key researcher and research area manager at Know Center Research GmbH.

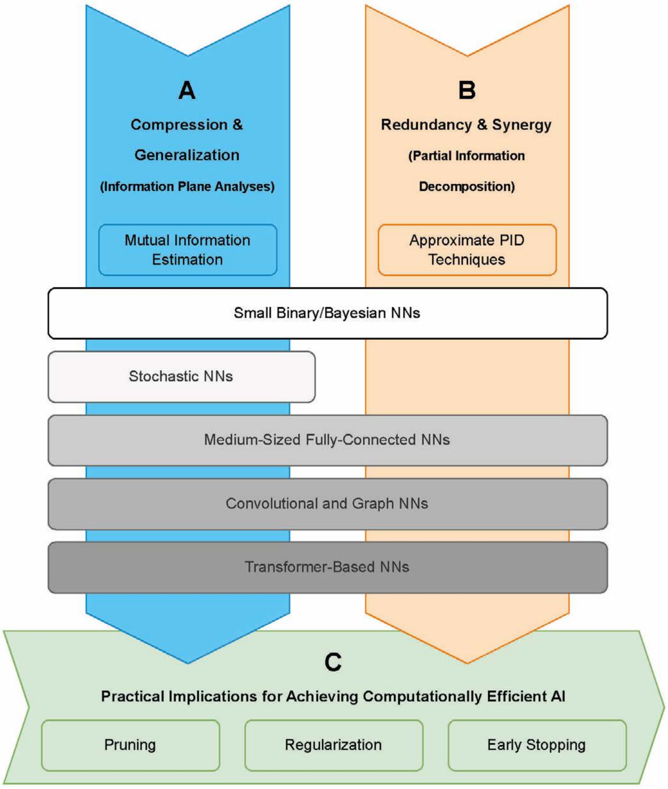

A
Compression & Generalization
(Information Plane Analyses)
Mutual Information Estimation
B
Redundancy & Synergy
(Partial Information Decomposition)
Approximate PID Techniques
Small Binary/Bayesian NNs
Stochastic NNs
Medium-Sized Fully-Connected NNs
Convolutional and Graph NNs
Transformer-Based NNs
C
Practical Implications for Achieving Computationally Efficient AI
Pruning
Regularization
Early Stopping

Figure 2: Overview of FWF project "Information Planes and Decompositions". We will work towards a comprehensive picture of the interplay between compression, redundancy, and generalization, and how these insights can be used to make AI greener.

Source: Bernhard Geiger, CC-BY-SA