The Analysis of Deep Neural Networks by Information Theory: From Explainability to Generalization

Shujian Yu[1,2]

[1] Department of Computer Science, Vrije Universiteit Amsterdam
[2] Department of Physics and Technology, UiT - The Arctic University of Norway
yusj9011@gmail.com

Abstract

Despite their great success in many artificial intelligence tasks, deep neural networks (DNNs) still suffer from a few limitations, such as poor generalization behavior for out-of-distribution (OOD) data and the "black-box" nature. Information theory offers fresh insights to solve these challenges. In this short paper, we briefly review the recent developments in this area, and highlight our contributions.

Measure of Information from Samples

One important bridge between traditional information theory and real-world machine learning problems is the information-theoretic quantities (such as entropy, mutual information, and divergence) estimation from samples, which is actually extremely hard in high-dimensional space.

In an effort to devise estimators that scale to present-day's machine learning problems, most recent work on estimating mutual information has focused on variational lower bounds that can be parameterized, for instance using neural networks (Belghazi et al. 2018). Alternatively, it is feasible to quantify the information (or interactions) in variables directly from samples by just using the eigenspectrum of a symmetric positive semidefinite (SPS) matrix constructed from data, avoiding the necessity of density estimation and the training of neural networks (Yu et al. 2019, 2021b). The computational complexity of the developed matrix-based entropy functional can be reduced from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^2)$ (n is the number of samples or mini-batch size) using techniques like randomized numerical linear algebra, with negligible loss in accuracy (Dong et al. 2022).

The Explainability and Generalization of DNNs from Information-Theoretic Perspective

For DNNs, the Information Bottleneck (IB) principle and the Rate-Distortion framework provide mathematically well-founded methods to understand the inner mechanisms of DNNs, especially in the training phase (Shwartz-Ziv and Tishby 2017). The matrix-based entropy functional enables us to systematically analyze the flow of information inside large-scale DNNs and discover more fundamental properties behind the training (Yu and Principe 2019).

Information theory is also crucially important to the generalization. Theoretically, it helps us determine tighter lower bounds for the generalization error. Practically, it can be used to design robust loss functions under distributional shift (Yu et al. 2021a), and to quantify the relatedness (Yu et al. 2021b) or transferability between domains or tasks.

The theoretical results can be transferred to design novel graph neural networks (GNNs) for brain network analysis, and thereby improving both the explainability and generalization of existing GNN-based diagnosis models for mental disorders (Yu et al. 2022).

References

Belghazi, M. I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; and Hjelm, D. 2018. Mutual information neural estimation. In *International conference on machine learning*, 531–540. PMLR.

Dong, Y.; Gong, T.; Yu, S.; and Li, C. 2022. Optimal Randomized Approximations for Matrix based Renyi's Entropy. *arXiv preprint arXiv:2205.07426*.

Shwartz-Ziv, R.; and Tishby, N. 2017. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*.

Yu, S.; Alesiani, F.; Yin, W.; Jenssen, R.; and Principe, J. C. 2022. Principle of Relevant Information for Graph Sparsification. In *The 38th Conference on Uncertainty in Artificial Intelligence*, volume 180, 2331–2341. PMLR.

Yu, S.; Alesiani, F.; Yu, X.; Jenssen, R.; and Principe, J. 2021a. Measuring dependence with matrix-based entropy functional. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 10781–10789.

Yu, S.; Giraldo, L. G. S.; Jenssen, R.; and Principe, J. C. 2019. Multivariate Extension of Matrix-Based Rényi's α-order Entropy Functional. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11): 2960–2966.

Yu, S.; and Principe, J. C. 2019. Understanding Autoencoders with Information Theoretic Concepts. *Neural Networks*, 117: 104–123.

Yu, S.; Shaker, A.; Alesiani, F.; and Principe, J. 2021b. Measuring the discrepancy between conditional distributions: methods, properties and applications. In *International Joint Conferences on Artificial Intelligence*, 2777–2784.