entropy

MDPI

Editorial

Information Bottleneck: Theory and Applications in Deep Learning

Bernhard C. Geiger [1,*,†] **and Gernot Kubin** [2,†]

1 Know-Center GmbH, Inffeldgasse 13/6, 8010 Graz, Austria
2 Signal Processing and Speech Communication Laboratory, Graz University of Technology, Inffeldgasse 16c, 8010 Graz, Austria; g.kubin@ieee.org
* Correspondence: geiger@ieee.org
† These authors contributed equally to this work.

Received: 2 December 2020; Accepted: 9 December 2020; Published: 14 December 2020

check for updates

Keywords: information bottleneck; deep learning; neural networks

The information bottleneck (IB) framework, proposed in [1], describes the problem of representing an observation X in a lossy manner, such that its representation T is informative of a relevance variable Y. Mathematically, the IB problem aims to find a lossy compression scheme described by a conditional distribution $P_{T|X}$ that is a minimizer of the following functional:

$$\min_{P_{T|X}} \Big(I(X;T) - \beta I(Y;T) \Big) \tag{1}$$

where the minimization is performed over a well-defined feasible set.

The IB framework has received significant attention in information theory and machine learning; cf. [2,3]. Recently, the IB framework has also gained popularity in the analysis and design of neural networks (NNs): The framework has been proposed to investigate the stochastic optimization of NN parameters with information–theoretic quantities, e.g., [4,5], and the IB functional was used as a cost function for NN training [6,7].

Based on this increased attention, this Special Issue aims to investigate the properties of the IB functional in this new context and to propose learning mechanisms inspired by the IB framework. More specifically, we invited authors to submit manuscripts that provide novel insight into the properties of the IB functional that apply the IB principle for training deep, i.e., multi-layer machine learning structures such as NNs and that investigate the learning behavior of NNs using the IB framework. To cover the breadth of the current literature, we also solicited manuscripts that discuss frameworks inspired by the IB principle, but that depart from them in a well-motivated manner.

In the remainder of this Editorial, we provide a brief summary of the papers in this Special Issue, in order of their appearance.

- Kunze et al. show that maximizing the evidence lower bound with a factorized Gaussian approximate posterior effectively limits mutual information between the available data and the learned parameters [8]. The effect of this tunable "model capacity" is validated in supervised and unsupervised settings, illustrating intuitive connections with overfitting, the NN architecture, and the dataset size;
- Wu et al. investigate the learnability within the IB framework. They show that if the parameter β in (1) falls below a certain threshold β_0, then a trivial representation T that is independent of X and Y minimizes the IB functional [9]. This threshold depends on the joint distribution of X and Y, and the authors propose an algorithm to estimate β_0 for a given dataset;

- Ngyuen and Choi argue that every layer in a feedforward NN should be optimized w.r.t. the IB functional (1) separately, with the parameter β adapted to the layer index [10]. Proposing a cost function for this multiobjective optimization problem, a computable variational bound, and a greedy optimization procedure, they achieve superior accuracy and adversarial robustness in stochastic binary NNs;
- Kolchinsky et al. propose an NN-based implementation of the IB problem, i.e., the compression scheme $P_{T|X}$ and conditional distribution $P_{Y|T}$ are parameterized by NNs [11]. Acknowledging the issues in [12], these NNs are trained to minimize an upper bound on $(I(X;T))^2 - \beta I(Y;T)$, combining variational and non-parametric approaches for bounding. Their experiments yield a better trade-off between $I(X;T)$ and $I(Y;T)$ and more meaningful latent representations in the bottleneck layer than a corresponding reformulation of [6];
- Tegmark and Wu investigate binary classification from real-valued observations [13]. They show that the observations can be compressed to a discrete representation T parameterized by β in such a way that the Pareto frontier of (1) is swept, essentially characterizing the binary classification problem. The authors further show that the corner points of this Pareto frontier, corresponding to a maximization of $I(Y;T)$ for a given alphabet size of T, can be computed without multiobjective optimization;
- Rodríguez Gálvez et al. discuss the scenario in which the target Y is a deterministic function of X in [14]. In this case, it is known that sweeping the paramter β in (1) is not sufficient to sweep the Pareto frontier of optimal $(I(X;T), I(Y;T))$ pairs [12]. The authors show that this shortcoming can be removed by optimizing $u(I(X;T)) - \beta_u I(Y;T)$ instead, where u is a strictly convex function. Furthermore, the authors demonstrate that the particular choice of the strictly convex function u helps to obtain a desired value of $I(X;T)$ over a wide range of parameters β_u;
- Franzese and Visintin propose using the IB functional as a cost function to train ensembles of decision trees for classification [15]. The authors show that these ensembles perform similarly to bagged trees, while they outperform the naive Bayes and k-nearest neighbor classifiers;
- Jónsson et al. [16] investigate the learning behavior of a high-dimensional VGG-16 convolutional NN in the information plane. Using MINE [17] to estimate $I(X;T)$ and $I(Y;T)$ throughout training, the authors observed a separate compression phase, during which the estimate of $I(X;T)$ decreases, thus aligning with [4]. The authors further show that regularizing NN training via an MINE-based estimate of the compression term $I(X;T)$ yields improved classification performance;
- Voloshynovskiy et al. propose an IB-based framework for semi-supervised classification, considering variational bounds both with learned and hand-crafted marginal distributions and achieving competitive performance [18]. A close investigation of their cost function yields improved insight into previously proposed approaches to semi-supervised classification;
- Fischer formulates the principle of minimum necessary information and derives from it the conditional entropy bottleneck functional [19]. This functional is mathematically equivalent to the IB functional, but uses the chain rule of mutual information to replace $I(X;T)$ in (1) by $I(X;T|Y)$. This results in different variational bounds, which are shown to yield better classification accuracy, improved robustness to adversarial examples, and stronger out-of-distribution detection than deterministic models or models based on variational approximations of (1), cf. [6];
- Fischer and Alemi provide additional empirical evidence for the claims in [19]. Specifically, they show that optimizing the proposed variational bounds leads to improved robustness against targeted and untargeted projected gradient descent attacks and to common corruptions (cf. [20]) of the ImageNet data [21]. Furthermore, the authors indicate that the conditional entropy bottleneck functional yields improved calibration for both clean and corrupted test data;
- Geiger and Fischer investigate the variational bounds proposed in [6,19]. While the underlying IB and conditional entropy bottleneck functionals are equivalent, the authors show that the variational bounds are not; these bounds are generally unordered, but an ordering can be enforced by restricting the feasible sets appropriately [22]. Their analysis is valid for general optimization and does not rely on the assumption that the variational bounds are implemented using NNs.

We thank all the authors for their excellent contributions and timely submission of their works. We are looking forward to many future developments that will build on the current bounty of insightful results and that will make machine learning better explainable.

Funding: The work of Bernhard C. Geiger was supported by the iDev40 project and by the COMET programs within the Know-Center and the K2 Center "Integrated Computational Material, Process and Product Engineering (IC-MPPE)" (Project No 859480). The iDev40 project has received funding from the ECSEL Joint Undertaking (JU) under Grant Agreement No 783163. The JU receives support from the European Union's Horizon 2020 research and innovation program. It is co-funded by the consortium members, grants from Austria, Germany, Belgium, Italy, Spain, and Romania. The COMET programs are supported by the Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology, the Austrian Federal Ministry of Digital and Economic Affairs, and by the States of Styria, Upper Austria, and the Tyrol. COMET is managed by the Austrian Research Promotion Agency FFG.

Acknowledgments: We would like to express our gratitude to the Editorial Assistants of Entropy for their help in organizing this Special Issue.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tishby, N.; Pereira, F.C.; Bialek, W. The Information Bottleneck Method. In Proceedings of the Allerton Conference on Communication, Control, and Computing, Monticello, IL, USA, 22–24 September 1999; pp. 368–377.
2. Zaidi, A.; Estella-Aguerri, I.; Shamai (Shitz), S. On the Information Bottleneck Problems: Models, Connections, Applications and Information Theoretic Views. *Entropy* **2020**, *22*, 151. [CrossRef] [PubMed]
3. Goldfeld, Z.; Polyanskiy, Y. The Information Bottleneck Problem and Its Applications in Machine Learning. *IEEE J. Sel. Areas Inf. Theory* **2020**, *1*, 19–38. [CrossRef]
4. Shwartz-Ziv, R.; Tishby, N. Opening the Black Box of Deep Neural Networks via Information. *arXiv* **2017**, arXiv:1703.00810.
5. Geiger, B.C. On Information Plane Analyses of Neural Network Classifiers—A Review. *arXiv* **2020**, arXiv:2003.09671.
6. Alemi, A.A.; Fischer, I.; Dillon, J.V.; Murphy, K. Deep Variational Information Bottleneck. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017 .
7. Achille, A.; Soatto, S. Information Dropout: Learning Optimal Representations Through Noisy Computation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 2897–2905. [CrossRef] [PubMed]
8. Kunze, J.; Kirsch, L.; Ritter, H.; Barber, D. Gaussian Mean Field Regularizes by Limiting Learned Information. *Entropy* **2019**, *21*, 758. [CrossRef] [PubMed]
9. Wu, T.; Fischer, I.; Chuang, I.L.; Tegmark, M. Learnability for the Information Bottleneck. *Entropy* **2019**, *21*, 924. [CrossRef]
10. Nguyen, T.T.; Choi, J. Markov Information Bottleneck to Improve Information Flow in Stochastic Neural Networks. *Entropy* **2019**, *21*, 976. [CrossRef]
11. Kolchinsky, A.; Tracey, B.D.; Wolpert, D.H. Nonlinear Information Bottleneck. *Entropy* **2019**, *21*, 1181. [CrossRef]
12. Kolchinsky, A.; Tracey, B.D.; Van Kuyk, S. Caveats for information bottleneck in deterministic scenarios. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
13. Tegmark, M.; Wu, T. Pareto-Optimal Data Compression for Binary Classification Tasks. *Entropy* **2020**, *22*, 7. [CrossRef] [PubMed]
14. Rodríguez Gálvez, B.; Thobaben, R.; Skoglund, M. The Convex Information Bottleneck Lagrangian. *Entropy* **2020**, *22*, 98. [CrossRef] [PubMed]
15. Franzese, G.; Visintin, M. Probabilistic Ensemble of Deep Information Networks. *Entropy* **2020**, *22*, 100. [CrossRef] [PubMed]
16. Jónsson, H.; Cherubini, G.; Eleftheriou, E. Convergence Behavior of DNNs with Mutual-Information-Based Regularization. *Entropy* **2020**, *22*, 727. [CrossRef] [PubMed]

17. Belghazi, M.I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; Hjelm, D. Mutual Information Neural Estimation. In Proceedings of the International Conference on Machine Learning (ICML), Stockholm, Sweden, 10–15 July 2018; pp. 531–540.
18. Voloshynovskiy, S.; Taran, O.; Kondah, M.; Holotyak, T.; Rezende, D. Variational Information Bottleneckfor Semi-Supervised Classification. *Entropy* **2020**, *22*, 943. [CrossRef] [PubMed]
19. Fischer, I. The Conditional Entropy Bottleneck. *Entropy* **2020**, *22*, 999. [CrossRef] [PubMed]
20. Hendrycks, D.; Dietterich, T. Benchmarking Neural Networks Robustness to Common Corruptions and Perturbations. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
21. Fischer, I.; Alemi, A.A. CEB Improves Model Robustness. *Entropy* **2020**, *22*, 1081. [CrossRef] [PubMed]
22. Geiger, B.C.; Fischer, I.S. A Comparison of Variational Bounds for the Information Bottleneck Functional. *Entropy* **2020**, *22*, 1229. [CrossRef] [PubMed]