DOI: 10.3217/978-3-85125-954-4-19

In Defense of Information Plane Analysis

Mina Basirat
Graz University of Technology
mina.basirat@ist.tugraz.at

Bernhard C. Geiger
Know-Center GmbH
geiger@ieee.org

Peter M. Roth
Vetmeduni Vienna, TU Munich
peter.m.roth@vetmeduni.ac.at

Abstract

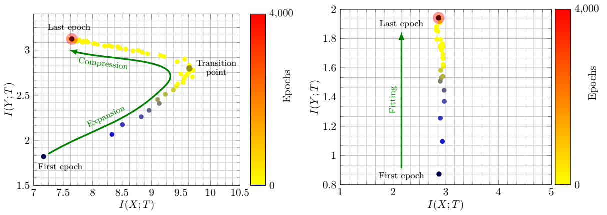
In this paper, we tackle the problem of analyzing neural network training via information plane analysis. The key idea is to describe the mutual information between the input and a hidden layer and a hidden layer and the target over time. Even though this is a reasonable approach, previous works showed inconsistent or even contradicting interpretations. Since the mutual information cannot be computed analytically, the authors applied different kinds of estimators, often not describing the mutual information very well. Taking these findings into account, we want to show that despite this theoretical limitation information planes allow at least for a geometric interpretation. Thus, enabling us to analyze different aspects of neural network learning for real-world problems.

1. Introduction and Problem Statement

One prominent approach to analyze neural network training is information plane (IP) analysis [7]. Building on the idea of the information bottleneck principle [8], the main idea is to describe and analyze the mutual information between the layers of a neural network over time. In particular, we are interested in the plane described by the mutual information $I(X;T)$ between the input X and the activation values of a hidden layer T and the mutual information $I(Y;T)$ between T and the target variable Y.

This is illustrated in Fig. 1 for two examples. From Fig. 1a, two phases can been observed, cf. [7]: first, a phase in which both $I(X;T)$ (expansion) and $I(Y;T)$ (fitting) are increasing, and, second, a compression phase during which $I(X;T)$ is decreasing again, whereas $I(Y;T)$ is increasing only slightly. The compression phase was interpreted as the hidden layer T discarding irrelevant information about the input X, and was causally connected to generalization. In contrast, Fig. 1b shows only fitting as an increase of $I(Y;T)$.

Even though these examples show that IPs appear to be an appealing way to analyze learning behaviors of NNs, we are facing the problem that the literature on IP analysis reports partially contradicting interpretations, cf. [2, 6, 7]. This, however, can be explained by the fact that the mutual information can often not be computed analytically, and different kinds of estimations for the mutual information terms $I(X;T)$ and $I(Y;T)$ are applied. Thus, similar to recent findings [2, 3], we would like to demonstrate that IPs represent geometric rather than information-theoretic phenomena. In this way, we are still able to use this technique to analyze NN training if the estimates are interpreted correctly.



(a) Expansion + compression.

(b) Fitting only.

Figure 1. Information planes reveal different behavior during neural network training.

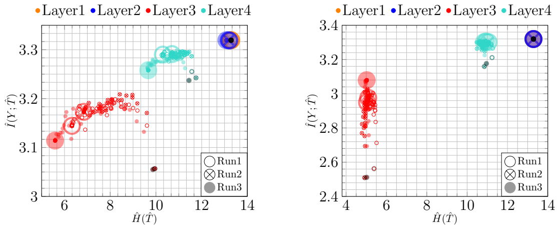
2. Geometric Interpretation of Information Planes

To this end, we create an IP from the plugin estimates for mutual information between the uniformly discretized activation value $\hat{T}$ and the network input X or class label Y, respectively. Introducing both fixed and adaptive binning schemes for obtaining $\hat{T}$, we get the estimators $\hat{I}(Y;\hat{T})$ and $\hat{I}(X;\hat{T}) = H(\hat{T})$. In this way, we argue that the correct interpretation of $H(\hat{T})$ yields an insight into the geometric compression of the activation T, both in absolute (e.g., describing the diameter of the set of all activations of a dataset) and relative (e.g., clustering of activations of a dataset) terms. To allow for a more intuitive interpretation, we additionally show a 2D visualization of latent space. For more details on the theoretical background and the applied binning approaches, we would like to refer to [1].

3. Illustrative Results

In the following, we show an example demonstrating that information planes can be a valuable tool to analyze and interpret neural network learning. To this end, we train a bottleneck network (*100-100-2-100*) for the well-known *MNIST* dataset [4] and *Brightness MNIST* (*BMNIST*) [5], a modified version of *MNIST*, where the illumination of the images has been increased. In this way, the contrast of the images is decreased and, thus, the classes are pushed closer together in the image space. The bottleneck model was chosen to make the geometric interpretation more apparent. In fact, in both cases, we finally obtain a similar classification result in terms of accuracy: 96.62% for *MNIST* and 95.25% for *BMNIST*. However, when looking at the corresponding information planes in Fig. 2 (*MNIST*) and Fig. 3 (*BMNIST*) reveals that the learning behavior is different. To make the temporal character of the trajectories more apparent, the first and the last epoch are highlighted by a black point and a large circle, respectively.

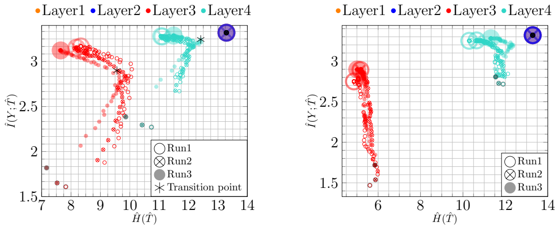
For *MNIST*, using adaptive binning (see Fig. 2b), we can recognize a fitting phase, i.e., $\hat{I}(Y;\hat{T})$ is increasing over time, indicating a growth of the class separability. In addition, using fixed binning (see Fig. 2a), we can recognize a geometric compression with an absolute scale for $\hat{H}(\hat{T})$ from the first to the last epoch for the last two layers. Indeed, as can be seen in Fig. 4, where we plot the two-dimensional latent space, the absolute scale reduces from approx. 47×69 to approx. 7×7 during training.



(a) Fixed binning.

(b) Adaptive binning.

Figure 2. IPs for *MNIST*: (a) fixed and (b) adaptive binning.

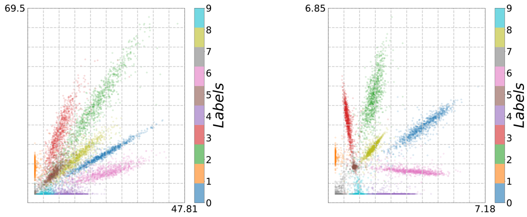


(a) Fixed binning.

(b) Adaptive binning.

Figure 3. IPs for *BMNIST*: (a) fixed and (b) adaptive binning.

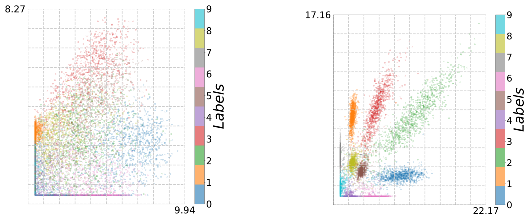
In contrast, for *BMNIST*, the IP analysis shown in Fig. 3 reveals that the learning behavior is different due to a different initial setting. Due to reduced contrast in the images, the classes are mapped to highly overlapping regions in the beginning (see Fig. 5a); for the original *MNIST* dataset, this is not the case (see Fig. 4a). Thus, during NN training the data points in the latent space have to be pushed apart according to their class label. In this way, we can recognize a fitting phase, i.e., increasing $\hat{I}(Y;\hat{T})$, for adaptive binning (see Fig. 3b) and an expansion phase for fixed binning (see Fig. 3a). Simultaneously, the data points are pushed apart and occupy a larger volume in the latent space (increased from 10×8 to 22×17), as can be seen in Fig. 5.



(a) First epoch.

(b) Last epoch.

Figure 4. 2D plots for *MNIST*: (a) first and (b) last epoch.



(a) First epoch.

(b) Last epoch.

Figure 5. 2D plots for *BMNIST*: (a) first and (b) last epoch.

4. Discussion and Conclusion

To overcome the known issues of IP analysis, we demonstrated that the IP represents geometric rather than information-theoretic effects, which we showed based on an illustrative example. To support these findings, we built a bottleneck architecture (i.e., using a two-dimensional layer), which allows us to directly relate the information covered by IPs to the geometric structure of the latent space. **For more technical details and a more thorough evaluation, we would like to refer to [1].**

Acknowledgment

This work was partially supported by the FFG Bridge project *SISDAL*, the BMBF (*International Future AI Lab "AI4EO"*), the project *iDev40*, and the Austrian COMET Program (*Know-Center*).

References

[1] Mina Basirat, Bernhard Geiger, and Peter M. Roth. A geometric perspective on information plane analysis. *Entropy*, 23(6):711, 2021.

[2] Bernhard Claus Geiger. On information plane analyses of neural network classifiers – a review. *arXiv:2003.09671*, 2020.

[3] Ziv Goldfeld, Ewout Van Den Berg, Kristjan Greenewald, Igor Melnyk, Nam Nguyen, Brian Kingsbury, and Yury Polyanskiy. Estimating information flow in deep neural networks. In *International Conference on Machine Learning*, pages 2299–2308, 2019.

[4] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[5] Norman Mu and Justin Gilmer. MNIST-C: A robustness benchmark for computer vision. *arXiv:1906.02337*, 2019.

[6] A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox. On the information bottleneck theory of deep learning. In *International Conference on Learning Representations*, 2018.

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

[8] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Allerton Conference on Communication, Control, and Computing*, pages 368–377, 1999.