Convergence of DNNs with mutual-information-based regularization

Hlynur Jónsson
IBM Research - Zurich
8803 Rüschlikon, Switzerland
hlynur4@gmail.com

Giovanni Cherubini
IBM Research - Zurich
8803 Rüschlikon, Switzerland
cbi@zurich.ibm.com

Evangelos Eleftheriou
IBM Research - Zurich
8803 Rüschlikon, Switzerland
ele@zurich.ibm.com

Abstract

Information theory concepts are leveraged with the goal of better understanding and improving Deep Neural Networks (DNNs). The information plane of neural networks describes the behavior during training of the mutual information at various depths between input/output and hidden-layer variables. Previous analysis revealed that most of the training epochs are spent on compressing the input. The estimation of mutual information is nontrivial for high-dimensional continuous random variables. Therefore, the computation of the mutual information for DNNs and its visualization on the information plane mostly focused on low-complexity fully-connected networks. In fact, even the existence of the compression phase in complex DNNs has been questioned and viewed as an open problem. In this paper, we present the convergence of mutual information on the information plane for a high-dimensional VGG-16 Convolutional Neural Network (CNN) by resorting to Mutual Information Neural Estimation (MINE), thus confirming and extending the results obtained with low-dimensional fully-connected networks. Furthermore, we demonstrate the benefits of regularizing a network, especially for a large number of training epochs, by adopting mutual information estimates as additional terms in the loss function characteristic of the network. Experimental results show that the regularization stabilizes the test accuracy and significantly reduces its variance.

1 Introduction

Deep Neural Networks (DNNs) have revolutionized several application domains of machine learning, including computer vision, natural language processing and recommender systems. Despite their success, the internal learning process of these networks is still an active field of research. One of the goals of this paper is to leverage information theoretical concepts to analyze and further improve DNNs. The analysis of DNNs through the information plane, i.e., the plane of mutual information values that each layer preserves at various learning stages on the input and the output random variables, was proposed in [1,2]. Previous approaches for the visualization of the information plane applied non-parametric estimation methods that do not work well with high dimensional data [2,3,4], as in this case the estimation of mutual information is nontrivial. The information plane for small fully-connected networks was visualized in [2]. The results in [2] suggested that most of the training epochs of a DNN, the "compression phase", are spent on compressing the input variables. The existence of the compression phase was later questioned in [4] for different activation functions. Our

4th workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada.

focus is on Convolutional Neural Networks (CNNs) with high complexity. In this paper, we present the convergence of mutual information on the information plane for a high-dimensional VGG-16 CNN [5] by resorting to Mutual Information Neural Estimation (MINE) [6]. The compression phase is evident from the obtained results, which confirm and extend the results previously found with low-dimensional fully-connected networks.

Furthermore, we consider DNNs with mutual-information-based regularization. The use of the mutual information between the input and a hidden layer of a DNN as a regularizer was suggested in [6,7,8]. The idea is based on the Information Bottleneck (IB) approach, which provides a maximally compressed version of the input random variable, while still retaining as much information as possible on a relevant random variable. Here we compare the accuracy achieved by a VGG-16 CNN, using well-known regularization techniques, such as dropout, batch normalization and data augmentation, with that of a VGG-16 network applying mutual-information-based regularization, by resorting either to MINE or VIB, and demonstrate the advantages of mutual-information-based regularization, especially for a large number of training epochs.

2 Mutual-information-based regularization of DNNs

The goal of any supervised learning method is to efficiently capture the relevant information about an input random variable, typically a label for classification tasks in the output random variable [1]. The Information Bottleneck (IB) method, first introduced in [9], finds a maximally compressed representation of an input random variable, X, which is a function of a relevant random variable, Y, such that it preserves as much information as possible on Y. Let us consider a DNN for image recognition, with input X, output $\hat{Y}$, and i-th hidden layer denoted by h_i. The classification task is related to the interpretation of an image that is generated from a relevant random variable Y. In case the hidden layer, h_i, only processes the output of the previous layer, h_{i-1}, the layers form a Markov chain of successive internal representations of the input. By applying the Data Processing Inequality (DPI) [10] to a DNN that consists of L layers we have $I(X;h_1) \geq I(X;h_2) \geq \ldots \geq I(X;h_L) \geq I(X;\hat{Y})$, where $I(A;B)$ denotes the mutual information of the random variables A and B. As mentioned in the Introduction, DNNs may be analyzed through the mutual information values on the information plane. However, estimating mutual information across layers in DNNs is a nontrivial task, as the outputs of the hidden layer neurons are continuous-valued high-dimensional random vectors.

A method for the estimation of mutual information in CNNs was recently proposed in [3], where a multivariate matrix-based Renyi's α-entropy method was considered for a LeNet-5 [11] network. However, the extension of the method to larger networks, such as the VGG-16 trained on CIFAR-10 that is considered in this paper, is not straightforward. An alternative approach in [6] proposes the estimation of the mutual information between high-dimensional continuous random variables by training a separate neural network with gradient descent through back-propagation. The approach, called Mutual Information Neural Estimation (MINE), utilizes a lower bound on the Donsker-Varadhan representation of the Kullback-Leibler divergence. The required distributions are learned through gradient descent, provided a sufficiently large number of samples are available for training. In order to visualize the information plane, two estimations are needed for each layer i in the network, namely those of $I(X;h_i)$ and $I(h_i;Y)$. Each of these estimations is parameterized by a separate deep neural network. As stated in [6], more training samples are needed as the complexity of MINE increases. Therefore, the aim here is to have a network capable of estimating mutual information sufficiently accurately while limiting the number of parameters in the network. To visualize the information plane, we estimate $I(X;h_i)$ and $I(h_i;Y)$ for each layer of the VGG-16. Therefore, we adopt a total of 32 networks for the visualization of the 16 information planes. The information planes corresponding to the 9-th and 16-th layer of the VGG-16 are shown in Fig. 1(a) and 1(b), respectively, where the mutual information estimates are expressed in bits. The compression phase is evident, which is consistent with previous work presented in [2], however for the first time shown in a CNN with a high complexity. One difference with respect to [2] is that for the VGG-16 network, the compression phase is evident earlier in the training process. We observe that $I(X;h_i)$ starts decreasing after the first VGG-16 epoch for layers h_i, $i > 7$, and continues to exhibit a decreasing trend until convergence. The estimation of $I(h_i;Y)$, for all layers h_i, $i > 2$, converges towards the upper bound equal to $log_2(10)$, which is the desired value of the mutual information.

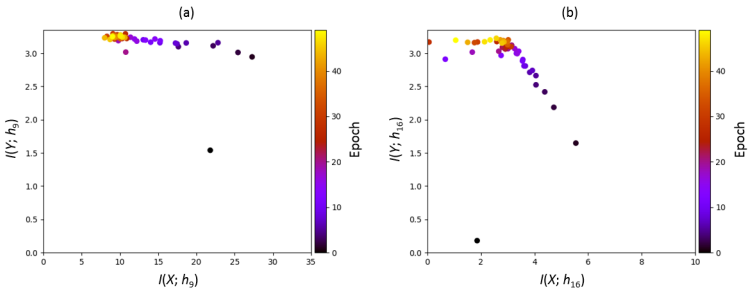

Figure 1: Information plane for (a) layer 9 and (b) layer 16 of VGG-16 trained on CIFAR-10.

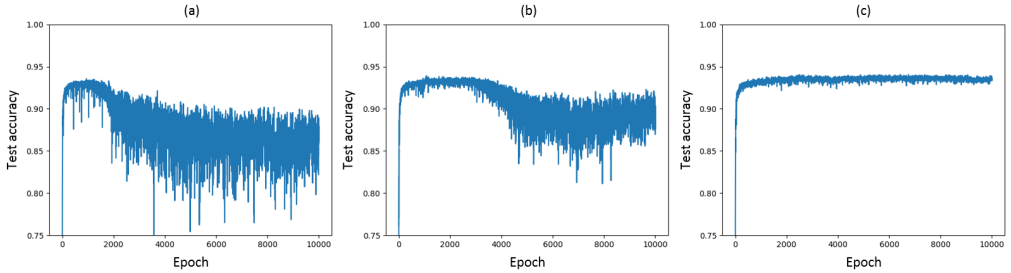

Figure 2: Test accuracies over 10000 epochs for VGG-16 trained on CIFAR-10 (a) without mutual-information-based regularizer, (b) with MINE-based regularizer and (c) with VIB-based regularizer.

Using MINE as a regularizer was proposed in [6] for a small fully-connected network trained on MNIST. As in [6], we also consider a MINE network to estimate the mutual information, however with a VGG-16 network of significantly higher complexity. Moreover, to further improve accuracy, the mutual information estimates of two layers were added as further regularization terms. The results comparing the test accuracies over 10000 epochs achieved by a VGG-16 network with the regularization techniques mentioned in the Introduction and a VGG-16 with MINE-based regularization on CIFAR-10 are shown in Fig. 2(a) and (b), respectively.

As an alternative mutual-information-based estimation method between consecutive layers in CNNs, a Variational Information Bottleneck (VIB) method was presented in [7]. This technique was used in [12] to reduce network complexity. VIB-based network regularization was proposed in [7] for fully-connected networks with low complexity. Here we extend this technique to CNNs with substantially higher complexity. Each hidden layer takes as input the previous hidden layer output, which typically contains some information not relevant for the classification task. By minimizing the mutual information between subsequent layers, a VIB-based regularizer reduces the amount of redundant information. The test accuracies over 10000 epochs achieved by a VGG-16 with VIB-based regularization on CIFAR-10 are shown in Fig. 2(c). The maximum test accuracy with the VIB-based regularizer is 94.1% and with the MINE-based regularizer is 93.9%, whereas a baseline accuracy for a VGG-16 network is measured as 93.25% in [13], which is similar to our results shown in Fig. 2(a). As can be seen, the two mutual-information-based regularization techniques presented in this paper not only stabilize the network, but also achieve higher accuracies.

3 Conclusion

Information theoretic concepts were adopted to analyze and improve high-complexity CNNs. We demonstrated the convergence of mutual information on the information plane and the existence of a compression phase for VGG-16, thus extending the results of [1] for fully-connected networks with low-complexity. Furthermore, our experiments highlighted the advantages of regularizing DNNs by mutual-information-based additional terms in the network loss function. Specifically, mutual-information-based regularization improves and stabilizes the test accuracy, significantly reduces its variance, and prevents the model from overfitting, especially for a large number of training epochs.

References

[1] N. Tishby and N. Zaslavsky. Deep learning and the information bottleneck principle. In 2015 IEEE Information Theory Workshop (ITW), pp. 1–5, 2015.

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

[3] S. Yu, R. Jenssen, and J. C. Principe. Understanding convolutional neural network training with information theory. arXiv:1804.06537, 2018.

[4] 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 Sixth International Conference on Learning Representations (ICLR), 2018.

[5] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014.

[6] M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm. Mine: mutual information neural estimation. arXiv:1801.04062, 2018.

[7] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy. Deep variational information bottleneck. arXiv:1612.00410, 2016.

[8] A. Achille and S. Soatto. Information dropout: Learning optimal representations through noisy computation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(12):2897–2905, 2018.

[9] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. arXiv preprint physics/0004057, 2000.

[10] T. M. Cover and J. A. Thomas. Elements of information theory. John Wiley & Sons, 2012.

[11] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

[12] B. Dai, C. Zhu, and D. Wipf. Compressing neural networks using the variational information bottleneck. arXiv:1802.10399, 2018.

[13] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf. Pruning filters for efficient convnets. arXiv:1608.08710v3, 2017.