Using Information Theory to Enhance the Understanding of Deep Neural Networks

Chuanwen Dong
Chalmers University of Technology
Department of Technology Management and Economics
Gothenburg, Sweden
chuanwen.dong@chalmers.se

Anver Hisham
Chalmers University of Technology
Department of Electrical Engineering
Gothenburg, Sweden
anver@chalmers.se

***Abstract*—Deep neural networks (DNNs) have demonstrated impressive success over the past couple of years. However, even with their outstanding (sometimes better than human) performances in a number of applications, little is known about their internal mechanisms and theoretical basis. Rigorous theoretical understandings are urgently needed to obtain knowledge and enhance control of DNNs. Currently, there is a heated debate in the usefulness of information theory, especially information bottleneck theory to explain DNNs, and we join this debate with our own data set and DNN. We validate the algorithm proposed in the mainstream literature and find that, the mutual information across layers of a neural network indeed increases, but the information compression process (the mutual information first increase and then decrease) cannot be observed with any kind of activation functions. However, we argue that the current mainstream literature has a high potential to over-fit the neural network with too many epochs ($> 10,000$). We introduce regularization in the training process and indeed observe the information compression process. As a result, we believe it is important to study the topic with solid understanding of the theory and the neural network at the same time.**

***Index Terms*—Deep neural network, information theory, information bottleneck theory, deep learning, mutual information**

I. Introduction

In recent years, deep neural networks (DNNs) have demonstrated impressive success in real-world tasks in supervised learning [Krizhevsky et al., 2012], unsupervised learning [Goodfellow et al., 2014], and reinforcement learning [Silver et al., 2016]. However, there is a lack of theoretical understanding of the DNNs and they are therefore oftentimes regarded as mysterious "black-boxes" [Alain and Bengio, 2016]. This prevents us from validating their performance limit and improving relevant algorithms, and most importantly, obtaining knowledge from information

One of the latest theoretical development of DNNs, among others, is the use of information theory to understand the information flow across the layers [Shwartz-Ziv and Tishby, 2017]. They propose to quantify the mutual information across different layers of a DNN and, on the basis of this, find that the goal of a DNN is to optimize the information bottleneck trade-off between compression and prediction for each layers.

Even though it is exciting to initiate the use of information theory to understand DNNs, their work has been criticized. [Saxe et al., 2018], among others, have validated their work and find that the result cannot be generalized, because the compression process cannot be observed when the activation function is changed from tanh to ReLu.

In this paper, we investigate the usefulness of information theory in explaining DNNs. To be more specific, we validate the work from [Shwartz-Ziv and Tishby, 2017] and [Saxe et al., 2018] respectively using a different data set and a different DNN in order to obtain some more generalized understandings. We are especially interested in answering the following research questions:

1) How does information flow across the different layers of a DNN?
2) To what extend can the information bottleneck theory be used to understanding of DNNs?
3) How sensitive are the results with respect to parameters/configurations of the neural network?

In Section II, we explain the information theory and information bottle neck theory. In Section III, we use a different data set and DNN to validate the usefulness of information theory in understanding the DNN and visualize the results. In Section IV, we discuss our findings and summarize the article.

II. The information theory used in understanding deep neural networks

In [Tishby and Zaslavsky, 2015], authors suggest to study the deep neural network in a so-called information plain focusing on the mutual information across layers. There are three kinds of layers in a neural network, namely the input layer X, the output layer Y, and the hidden layers T. When data is pushed into the neural network, the mutual information across layers can be calculated, and can be used to quantify the information flow across the neural network.

In a follow-up article, [Shwartz-Ziv and Tishby, 2017] suggest to make use of the mutual information and analyze it in the lens of the information bottleneck theory proposed by [Tishby et al., 1999] to enhance the learning of the neural network.

A. Mutual information

As a classical term in the classical information theory, the mutual information quantifies the expected relevant bits that

the input variable X contains about the label Y:

$$\begin{aligned} I(X;Y) &= D_{KL}[p(x,y)||p(x)p(y)] \\ &= \sum_{x\in X, y\in Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x\in X, y\in Y} p(x,y) log \frac{p(x|y)}{p(x)} \\ &= H(X) - H(X|Y), \end{aligned} \tag{1}$$

where $D_{KL}[p||q]$ is the Kullback-Liebler divergence of the distributions p and q, and $H(X)$ and $H(X|Y)$ are the entropy of X and conditional entropy of X with respect to Y, respectively.

In a neural network, the mutual information of any two layers, e.g., the input layer X and any hidden layer T, can be calculated using (1). As suggested by [Shwartz-Ziv and Tishby, 2017], the optimal learning problem of a neural network can be seen as the construction of an optimal encoder of the input information with respect to the relevant information in input about output labels.

In our simulations, we used the tight upper bound in Equation (7) in [Saxe et al., 2018] to compute mutual information between input layer X and a hidden layer T:

$$I(T;X) = \frac{1}{N}\sum_i \log N \sum_j \exp \frac{-||h_i - h_j||_2^2}{2\sigma^2}, \tag{2}$$

where N is the number of training samples. The vector h_i and h_j are activation outputs of layer i and j respectively, after adding noise. The added noise variance is σ^2 and mean is zero.

Similarly, we used the tight upper bound in equation (10) in [Saxe et al., 2018] to compute mutual information between a hidden layer T and output layer Y,

$$\begin{aligned} I(T;Y) = &\frac{1}{N}\sum_i \log N \sum_j \exp \frac{-||h_i - h_j||_2^2}{2\sigma^2} \\ &- \frac{1}{N}\sum_{c=1}^{C}\sum_i \log N_c \sum_{j:Y_j=c} \exp \frac{-||h_i - h_j||_2^2}{2\sigma^2}, \end{aligned} \tag{3}$$

where C is the total number of output labels and N_c denotes the number of data samples with output label c,

B. The information bottleneck theory

With the mutual information calculated in Equation (1), quantitative analysis can be further applied. One of the method is to use of the information bottleneck theory proposed by [Tishby et al., 1999]. The theory offers a quantitative framework to find the optimal trade-off between compression of X and the prediction of Y.

The framework can be briefly explain as follows. Define $t \in T$ the compressed representation of $x \in X$, and $p(t|x)$ a mapping between the two variables. In a Markov chain: $Y \longrightarrow X \longrightarrow T$, the information bottleneck theory can be formulated by the following equation:

$$min_{p(t|x),p(y|t),p(t)}\{I(X;T) - \beta I(T;Y)\}, \tag{4}$$

where the Lagrange multiplier β represents the level of relevant information captured by the representation T.

Their theory can be explained in Fig. 1 from [Shwartz-Ziv and Tishby, 2017]. When training a neural network, two phases happen sequentially: In the first initial fitting phase, the mutual information increase so that the neural network accumulates information from the data; in the second compression phase, information is push out of the neural network, suggesting a generalization performance of the neural network. The green dots represent the tipping-point where the two phases are separated, where the mutual information increases before that point and decreases after it. It has been hypothesized that this compression phase occurs due to the random diffusion-like behavior of stochastic gradient descent. This observation can be generalized to understand how the neural network learn information during its training process.

However, [Saxe et al., 2018] argue that this observation cannot be generalized because it does not exist with ReLu activation functions. They argue that the compression observed by [Shwartz-Ziv and Tishby, 2017] are primarily due to the double-saturating tanh activation function used in their numerical test.

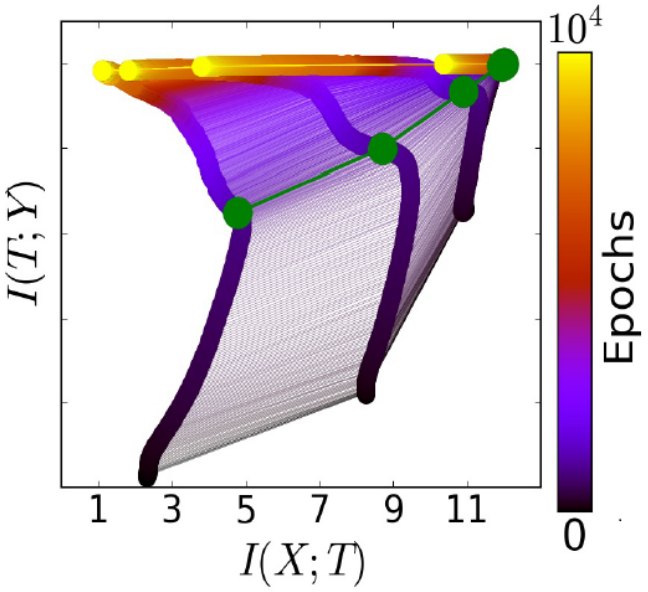

Fig. 1. The mutual information first increases and then decreases after a certain number of epochs (at the green dots), indicating that the information is compressed after the green dots [Shwartz-Ziv and Tishby, 2017].

III. Numerical experiments

In this section, we investigate the information theory in a neural network. We are interested in the usefulness of information bottleneck theory to explain the mechanism of the neural network, and especially the compression process.

We use the classical Pokemon data available from Kaggle (https://www.kaggle.com/rounakbanik/pokemon). The data set

contains about 293,000 Pokemon sightings (historical appearances of Pokemon in the Pokemon Go game). For this experiment we took 4 input features; longitude, latitude, temperature, and pressure of the Pokemon. The output is the prediction of the 3 classes of Pokemon; Diglett, Seel, and Tauros.

There are mainly two reasons for using Pokemon dataset: 1) It is a classic and well-known data set used in deep neural network applications, 2) we already have an experience of the data set and the neural network via the previous homework. The neural network is structured with 3 hidden layers, with $4 \times 16 \times 16 \times 8 \times 3$ neurons and *relu* activation functions.

A. The mutual information across layers

We follow the algorithm of [Shwartz-Ziv and Tishby, 2017] and plot the mutual information across all 3 hidden layers (Fig. 2). There are two main observations: 1) Mutual information increases with respect to the number of epochs, and 2) There is no compression phase as mentioned in [Shwartz-Ziv and Tishby, 2017], where the mutual information begin to decrease at a certain number of epochs. .

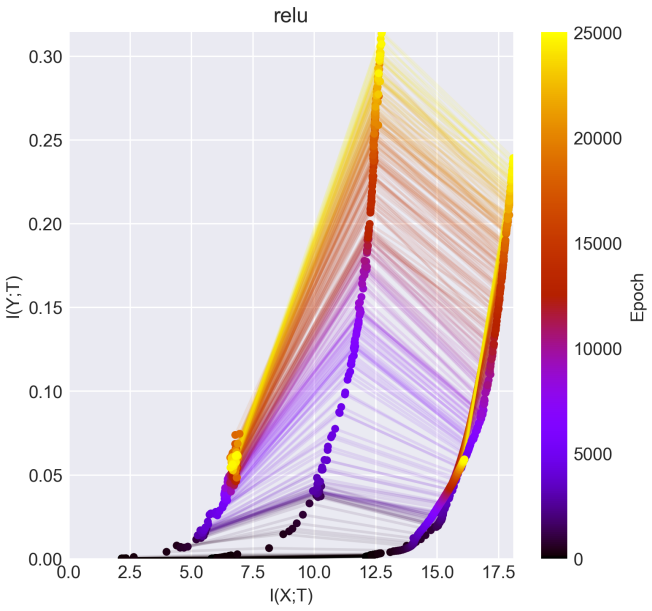

Fig. 2. The Pokeman data set is trained for *25,000* epochs in a deep neural network with four hidden layers structured $4 \times 16 \times 16 \times 8 \times 3$ and *relu* activation functions. When the number of epoch increases, both the mutual information $I(X,T)$ and $I(Y,T)$ (almost always) increase.

B. The impact of activation functions

As noted by [Saxe et al., 2018], the compression phase is actually formed from the activation function, but not the information flow inside the neural network itself. It is argued that the double-sided saturating nonlinear function such as tanh will yield a compression phase, but single-sided linear saturating activation functions such as relu will not. As a result, we test the tanh function and present the result in Fig. 3.

Interestingly, the result of a tanh function (Fig. 3) is very similar to that of a relu function (Fig. 2), except for some of the last epochs. However, this small difference happens after training the network for about 25,000 epochs. The network might be over-fitted and the practicality of the result is questionable.

Our finding is different from that from [Saxe et al., 2018, Figure 1], where the two activation functions have significantly different output plots. Our findings show that, no compression phase can be observed. As a comparison, [Saxe et al., 2018] shows an establishment of the compression phase when the tanh activation function is used.

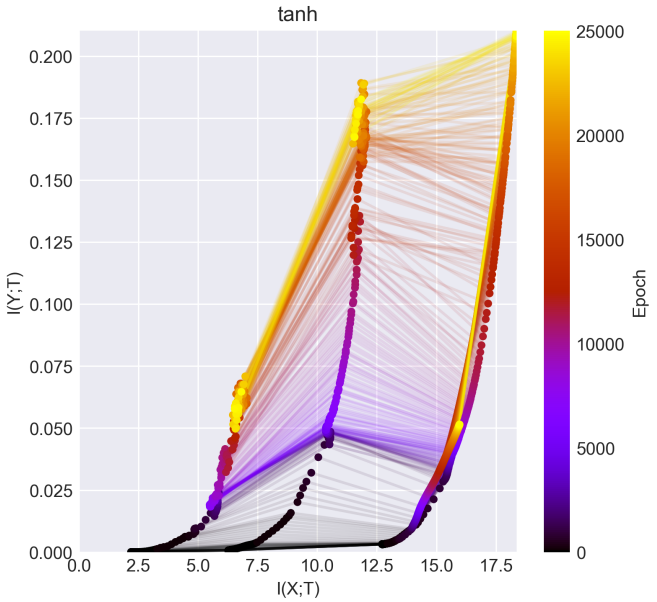

Fig. 3. The Pokeman data set is trained for *25,000* epochs in a deep neural network with four hidden layers structured $4 \times 16 \times 16 \times 8 \times 3$ and *tanh* activation functions. When the number of epoch increases, both the mutual information $I(X,T)$ and $I(Y,T)$ increase, except for some of the last epochs.

C. The impact of regularization

Now that we are always working with a large number of epochs (>10,000), over-fitting is very probably unavoidable. As a result, we use regularization (L2 norm, with factor 0.001) in our training and re-validate the results in Figure 4 for *relu* activation function, and in Figure 5 for *tanh* activation function respectively. Interesting, we indeed observe the compression process for $I(Y;T)$, where the mutual information increases during the initial epochs, and decreases eventually. We also observe a very small compression phase of $I(X;T)$. However, our finding is different from [Shwartz-Ziv and Tishby, 2017], where a large compression of $I(X;T)$ but a moderate compression of $I(Y;T)$ are observed. Our result is also different from [Saxe et al., 2018], where no compression is found with relu activation function.

IV. Summary and Discussion

The main contribution of this article is to validate the usefulness of information theory in explaining deep neural

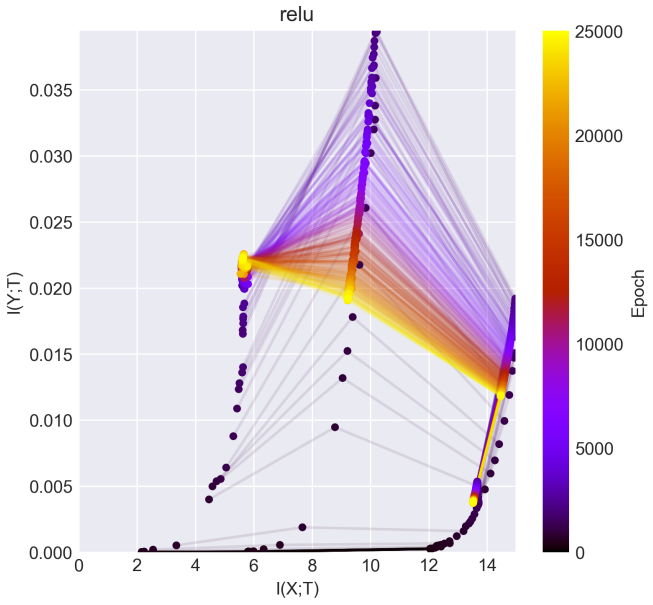

Fig. 4. The Pokemon data set is trained for *25,000* epochs in a deep neural network with three hidden layers structured $4 \times 16 \times 16 \times 8 \times 3$ and *relu* activation functions, and with L2 norm regularization. Information compression can be observed.

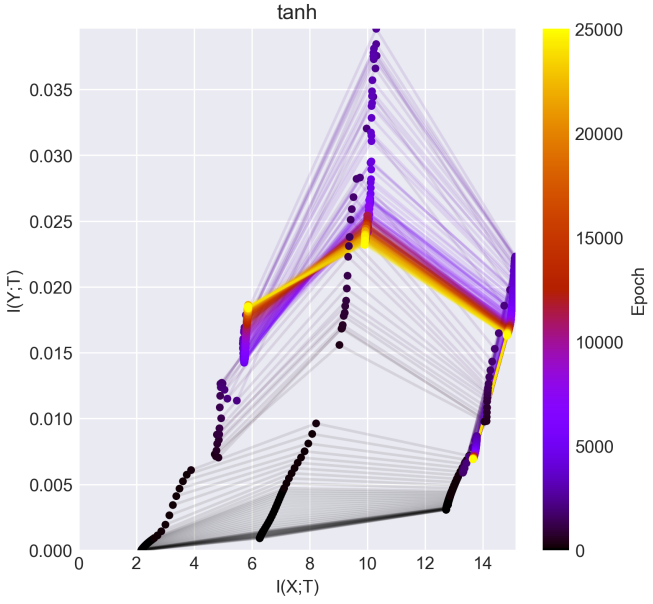

Fig. 5. The Pokemon data set is trained for *25,000* epochs in a deep neural network with three hidden layers structured $4 \times 16 \times 16 \times 8 \times 3$ and *tanh* activation functions, and with L2 norm regularization. Information compression can be observed

networks. We use a different data set with a different neural network. The main findings can be summarized as follows:

1) The mutual information across different layers indeed increases when the number of epochs increases.
2) We do not observe the compression phase mentioned in [Shwartz-Ziv and Tishby, 2017] using their algorithm, which aligns with [Saxe et al., 2018].
3) However, when we introduce regularization in the training (obviously there could be over-fitting due to the large number of epochs), the compression process can be shown with both relu and tanh activation functions, though with a different magnitude w.r.t $I(X;T)$ and $I(Y;T)$ compared to [Shwartz-Ziv and Tishby, 2017].

On the basis of the findings, we argue that information theory can indeed be used to support the understanding of neural networks. However, some additional configurations such as regularization needed to be implemented, and these configurations must be based on the understanding of the functionality of the neural network itself.

Taking our experiment as an example, we have used the neural network to hunt Pokemon and therefore have already accumulated solid knowledge on the set-up of the neural network itself. We already know beforehand that, a useful neural network can generally already provide good result after about 100 epochs. This experience has motivated us to question the use of huge number of epochs $(> 10,000)$ in the literature and the potential adverse impact on over-fitting. As a result, we have introduced regularization in the training, and finally find very different results from [Shwartz-Ziv and Tishby, 2017] and [Saxe et al., 2018].

As a summary, we believe that the use of information theory can support the understand of neural networks, but we are skeptical of the findings from current research. Future research need to combine a solid understanding of both the theory and the network.

REFERENCES

[Alain and Bengio, 2016] Alain, G. and Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J. Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advanceds in Neural Information Processing System*, pages 2672–2680.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advanceds in Neural Information Processing System*, pages 1097–1105.

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

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

[Silver et al., 2016] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

[Tishby et al., 1999] Tishby, N., Pereira, F. C., and Bialek, W. (1999). The information bottleneck method. In *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377.

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