# Information Bottleneck Method and Deep Learning

Daniel Kunin

December 16, 2017

## 1 Introduction

For my STATS 385 paper I chose to review Alex Alemi et al.'s recently published (July, 2017) paper "Deep Variational Information Bottleneck" in the context of recent work around the Information Bottleneck Method. The Information Bottleneck (IB) Method is an Information Theory based approach to understanding deep learning that has recently garnered more attention. Alex Alemi et al.'s recent paper has addressed some of the greatest challenges to the IB method, significantly improving the practicality of the theory, but at the same time has failed to adequately set up future experiments. Before diving into the paper I will first summarize the context of the current discourse on the Information Bottleneck Method and its application to deep learning.

### 1.1 The IB Method

The Information Bottleneck Method was first termed by Naftali Tishby et al. in their 1999 paper "The Information Bottleneck Method" [3]. The basic idea of the method is that we can view a supervised learning task as a constrained compression problem where an algorithm must balance maximally compressing the representation of the input features, while preserving the information relevant to the prediction of the output features. The authors describe this process as squeezing "the information that $X$ provides about $Y$ through a bottleneck". They further formalize this mathematically in an information theoretic manner such that they can determine theoretical bounds on the trade off between compression and prediction [3]. While their theory is well written and beautifully simple it took nearly one and a half decades before the importance of their work had significant traction in the machine learning community.

### 1.2 IB Method and Deep Learning

In 2015 Naftali Tishby and Noga Zaslavsky proposed exploring Deep Neural Networks (DNNs) through the lens of the IB Method. They noticed that the layered architecture of a DNN fits nicely into the mutual information framework of the IB method [4]. In 2017 Rafid Schwartz-Ziv and Naftali Tishby published a follow up paper visualizing DNNs in the information plane, a two-dimensional space where the x-axis represents the mutual information between the input $X$ and the representation of these features $T$ and the y-axis represents the mutual information between the representation $T$ and the output $Y$. Startlingly they observed two distinct phases to a DNN during training. The first phase is characterized by gradients with a large mean and low standard error, which they termed the drift phase. The second phase is characterized by gradients with small mean and large standard error, which they termed the diffusion phase [2]. Three months after Schwartz-Ziv and Tishby published their surprising results, Alex Alemi, Ian Fisher, Joshua Dillon and Kevin Murphy, all researchers at Google, published the "Deep Variational Information Bottleneck".

# 2 Deep Variational Information Bottleneck

This paper can be summarized into two major sections: the development of a novel approach to estimating mutual information and experiments using this approach as an objective function for training a neural network [1].

## 2.1 The VIB Method

In the first section of the paper the authors demonstrate how through common information theory inequalities such as the non-negativity of the Kullback Leibler divergence and variational approximations they can construct lower bounds on both terms of the IB objective function (i.e. the compression and predicting terms). They combine these lower bounds to construct the following lower bound on the IB objective function [1]:

$$L \approx \frac{1}{N}\sum_{n=1}^{N}[\int dz p(z|x_n)\log q(y_n|z) - \beta p(z|x_n)\log\frac{p(z|x_n)}{r(z)}]$$

## 2.2 Experimental Performance

In the second section of the paper the authors use this lower bound to construct an objective function which they train stochastic neural networks with. They then evaluate these networks under a variety of conditions and datasets including MNIST and ImageNet. They also have a strong focus on measuring the adversarial robustness of their networks under a variety of attacks. Their main results were that their VIB-trained networks, while not "state of the art" in classification performance did demonstrate improved resistance to adversarial attack [1].

# 3 Critique and Comments

I think the strongest aspect of Alemi et al.'s paper is the work they did on the VIB method to provide an effective and robust way for estimating mutual information. The authors were absolutely correct to notice that "the main drawback to the IB principle is that computing mutual information is, in general, computationally challenging" and only under strong assumptions is it tractable [1]. The work they did to construct a lower bound of the the IB objective function with an autoencorder is crucial to developing new methods of mutual information estimation that are more feasible and do not rely on probability distribution estimation. However, instead of creating and sharing an estimation tool that the community could use they instead used this lower bound to construct an objective function that they trained stochastic neural networks with. In many ways this is a significant divergence from previous work done using the IB Method. This paper is more an algorithm development paper than an algorithm explanation paper. The general tone of most of the previous work done in the Information Bottleneck Method has been to use the method as a tool for understanding how existing architectures and learning algorithms work. In this paper, Alemi et al. instead focused on how the method could be used to create improved architectures. While, this is an important next step to demonstrating the usefulness of the method I think prematurely applying the IB Method to train neural networks does not alleviate the concerns that these algorithms are "black boxes". I would of rather seen Alemi et al. use their VIB method to create a mutual information estimation tool that they could use to build off of or confirm the startling work of Schwartz-Ziv and Tishby on the phase transitions of DNNs. At the very least I wish that the authors provided their code open sourced to the community.

I personally cannot talk much about the work the authors did on measuring the adversarial robustness as this is an area of deep learning I am not very familiar with. That being said I do appreciate the thoroughness of their documentation and experimental setup. And believe that it is important to evaluate the performance of an algorithm with a broader range of measures than just testing accuracy.

## 4 Publication and Future Work

As a reviewer I would absolutely recommend that Alemi et al.'s paper "Deep Variational Information Bottleneck" be submitted for publication. That being said I actually think it should be split into two papers. I think that their is enough content in each of the two sections (VIB method and adversarial robustness) for its own paper. I also think this would highlight the important of the VIB method as a new tool for mutual information estimation while also providing new insight on how to use the IB method for more practical applications. I think that the VIB method and the construction of efficient mutual information estimators will be essential to the future of the Information Bottleneck Method as a "Theory of Deep Learning"!

## References

[1] Alex Alemi, Ian Fischer, Josh Dillon, and Kevin Murphy. Deep Variational Information Bottleneck. ICLR 2017. URL https://arxiv.org/abs/1612.00410.

[2] Ravid Shwartz-Ziv Naftali Tishby. Opening the black box of deep neural networks via information. arXiv: 1803.00810

[3] Tishby, Naftali, et al. The Information Bottleneck Method 1999 arXiv:physics/0004057v1

[4] Tishby, Naftali, and Noga Zaslavsky. Deep Learning and the Information Bottleneck Principle. [1503.02406] Deep Learning and the Information Bottleneck Principle, 9 Mar. 2015, arxiv.org/abs/1503.02406.