entropy

MDPI

Article

On the Difference between the Information Bottleneck and the Deep Information Bottleneck

Aleksander Wieczorek * **and Volker Roth**

Department of Mathematics and Computer Science, University of Basel, CH-4051 Basel, Switzerland; volker.roth@unibas.ch

* Correspondence: aleksander.wieczorek@unibas.ch

Received: 30 November 2019; Accepted: 16 January 2020; Published: 22 January 2020

Abstract: Combining the information bottleneck model with deep learning by replacing mutual information terms with deep neural nets has proven successful in areas ranging from generative modelling to interpreting deep neural networks. In this paper, we revisit the deep variational information bottleneck and the assumptions needed for its derivation. The two assumed properties of the data, X and Y, and their latent representation T, take the form of two Markov chains $T - X - Y$ and $X - T - Y$. Requiring both to hold during the optimisation process can be limiting for the set of potential joint distributions $P(X, Y, T)$. We, therefore, show how to circumvent this limitation by optimising a lower bound for the mutual information between T and Y: $I(T; Y)$, for which only the latter Markov chain has to be satisfied. The mutual information $I(T; Y)$ can be split into two non-negative parts. The first part is the lower bound for $I(T; Y)$, which is optimised in deep variational information bottleneck (DVIB) and cognate models in practice. The second part consists of two terms that measure how much the former requirement $T - X - Y$ is violated. Finally, we propose interpreting the family of information bottleneck models as directed graphical models, and show that in this framework, the original and deep information bottlenecks are special cases of a fundamental IB model.

Keywords: information bottleneck; Markov assumption; Markov chain; deep variational information bottleneck; conditional independence; mutual information

1. Introduction

Deep latent variable models, such as generative adversarial networks [1] and the variational autoencoder (VAE) [2], have attracted much interest in the last few years. They have been used in many application and formed a conceptual basis for a number of extensions. One of popular deep latent variable models is the deep variational information bottleneck (DVIB) [3]. Its foundational idea is that of applying deep neural networks to the information bottleneck (IB) model [4], which finds a sufficient statistic T of a given variable X while retaining side information about a variable Y.

The original IB model, as well as DVIB, assumes the Markov chain $T - X - Y$. Additionally, in the latter model, the Markov chain $X - T - Y$ appears by construction. The relationship between the two assumptions and how it influences the set of potential solutions have been neglected so far. In this paper, we clarify this relationship by showing that it is possible to lift the original IB assumption in the context of the deep variational information bottleneck. It can be achieved by optimising a lower bound on the mutual information between T and Y, which follows naturally from the model's construction. This explains why DVIB can optimise over a set of distributions which is not overly restrictive.

This paper is structured as follows. In Section 2 we describe the information bottleneck and deep variational information bottleneck models, along with their extensions. Section 3 introduces the lower bound on the mutual information which makes it possible to lift the original IB $T - X - Y$ assumption, and makes possible the interpretation of this bound. It also contains the specifications of IB as a directed graphical model. We provide concluding remarks in Section 4.

2. Related Work on the Deep Information Bottleneck Model

The information bottleneck was originally introduced in [4] as a compression technique in which a random variable X is compressed while preserving relevant information about another random variable Y. The problem was originally formulated using only information theory concepts. No analytical solution exists for the original formulation; however, an additional assumption that X and Y are jointly Gaussian distributed leads to a special case of the IB, the Gaussian information bottleneck, introduced in [5], where the optimal compression is also Gaussian distributed. The Gaussian information bottleneck has been further extended to sparse compression and to meta-Gaussian distributions (multivariate distributions with a Gaussian copula and arbitrary marginal densities) in [6]. The idea of applying deep neural networks to model the information common to X and T as well as Y and T has resulted in the formulation of the deep variational information bottleneck [3]. This model has been extended to account for invariance to monotonic transformations of the input variables in [7].

The information bottleneck method has also recently been applied to the analysis of deep neural networks in [8], by quantifying mutual information between the network layers and deriving an information theory limit on deep neural network efficiency. This has lead to attempts at explaining the behaviour of deep neural networks with the IB formalism [9,10].

We now proceed to formally define the IB and DVIB models.

Throughout this paper, we adopt the following notation. Define the Kullback–Leibler divergence (KL divergence) between two (discrete or continuous) probability distributions P and Q as $D_{KL}(P(X) \,||\, Q(X)) = \mathbb{E}_{P(X)} \log \frac{P(X)}{Q(X)}$. Note that the KL divergence is always non-negative. The mutual information between X and Y is defined as

$$I(X;Y) = D_{KL}\Big(P(X,Y) \,||\, P(X)P(Y)\Big). \tag{1}$$

Since the KL divergence is not symmetric, the divergence between the product of the marginals and the joint distribution has also been defined as the lautum information [11]:

$$L(X;Y) = D_{KL}\Big(P(X)P(Y) \,||\, P(X,Y)\Big). \tag{2}$$

Both quantities have conditional counterparts:

$$\begin{aligned} I(X;Y|Z) &= D_{KL}\Big(P(X,Y,Z) \,||\, P(X|Z)P(Y|Z)P(Z)\Big) \\ &= \mathbb{E}_{P(Z)} D_{KL}\Big(P(X,Y|Z) \,||\, P(X|Z)P(Y|Z)\Big), \\ L(X;Y|Z) &= D_{KL}\Big(P(X|Z)P(Y|Z)P(Z) \,||\, P(X,Y,Z)\Big) \\ &= \mathbb{E}_{P(Z)} D_{KL}\Big(P(X|Z)P(Y|Z) \,||\, P(X,Y|Z)\Big). \end{aligned} \tag{3}$$

Let $H[X] = -\mathbb{E}_{P(X)}[\log P(X)]$ denote entropy for discrete and differential entropy for continuous X. Analogously, $H[P(X|Y)] = -\mathbb{E}_{P(X,Y)}[\log P(X|Y)]$ denotes conditional entropy for discrete and conditional differential entropy for continuous X and Y.

2.1. Information Bottleneck

Given two random vectors X and Y, the information bottleneck method [4] searches for a third random vector T, which, while compressing X, preserves information contained in Y. The resulting variational problem is defined as follows:

$$\min_{P(T|X)} I(X;T) - \beta I(T;Y), \tag{4}$$

where β is a parameter defining the trade-off between compression of X and preservation of Y. The solution is the optimal conditional distribution of $T|X$. No analytical solution exists for the general IB problem defined by Equation (4); however, for discrete X and Y, a numerical approximation of the optimal distribution T can be found with the Blahut–Arimoto algorithm for rate-distortion function calculation [4]. Note that the assumed property $T - X - Y$ of the solution is used in the derivation of the model.

2.1.1. Gaussian Information Bottleneck

For Gaussian distributed (X, Y), let the partitioning of the joint covariance matrix be denoted as follows:

$$(X,Y) \sim \mathcal{N}\left(0, \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}{}^\top & \Sigma_Y \end{pmatrix}\right). \tag{5}$$

The assumption that X and Y are jointly Gaussian distributed leads to the Gaussian information bottleneck [5] where the solution T of Equation (4) is also Gaussian distributed. T is then a noisy linear projection of X; i.e., $T = AX + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \Sigma_\epsilon)$ is independent of X. This means that $T \sim \mathcal{N}(0, A\Sigma_X A^\top + \Sigma_\epsilon)$. The IB optimisation problem defined in Equation (4) becomes an optimisation problem over the matrix A and noise covariance matrix Σ_ϵ:

$$\min_{A,\Sigma_\epsilon} I(X; AX + \epsilon) - \beta I(AX + \epsilon; Y). \tag{6}$$

Recall that for n-dimensional Gaussian distributed random variables, entropy, and hence mutual information, have the following form: $I(X;Y) = H(X) - H(X|Y) = \frac{1}{2}\log\left((2\pi e)^n |\Sigma_X|\right) - \frac{1}{2}\log\left((2\pi e)^n |\Sigma_{X|Y}|\right)$, where Σ_X and $\Sigma_{X|Y}$ denote covariance matrices of X and $X|Y$, respectively. The notation $|M|$ is used for the determinant of a matrix M. The Gaussian information bottleneck problem has an analytical solution, given in [5]: for a fixed β, Equation (6) is optimised by $\Sigma_\epsilon = I$ and A having an analytical form depending on Σ_X and eigenpairs of $\Sigma_{X|Y}\Sigma_Y^{-1}$. Here again, the $T - X - Y$ assumption is used in the derivation of the solution.

2.1.2. Sparse Gaussian Information Bottleneck

Sparsity of the compression in the Gaussian IB can be ensured by requiring the projection matrix A to be diagonal; i.e., $A = \mathrm{diag}(a_1, \ldots, a_n)$. It has been shown in [6] that since $\log|A\Sigma A^\top + I| = \log|\Sigma A^\top A + I|$ for any positive definite Σ and symmetric A, the sparsity requirement simplifies Equation (6) to minimisation over diagonal matrices with positive entries $D = A^\top A = \mathrm{diag}({a_1}^2, \ldots, {a_n}^2) = \mathrm{diag}(d_1, \ldots, d_n)$. I.e.:

$$\min_{D=\mathrm{diag}(d_1,\ldots,d_n)} I(X; AX + \epsilon) - \beta I(AX + \epsilon; Y) \tag{7}$$

with $d_i = {a_i}^2$ and $\epsilon \sim \mathcal{N}(0, I)$ independent of X.

2.2. Deep Variational Information Bottleneck

The deep variational information bottleneck [3] is a variational approach to the problem defined in Equation (4). The main idea is to parametrise the conditionals $P(T|X)$ and $P(Y|T)$ with neural networks so that the two mutual informations in Equation (4) can be directly recovered from two deep neural nets. To this end, one can express the mutual informations as follows:

$$\begin{aligned} I(X;T) &= D_{KL}\left(P(T|X)P(X)\|P(T)P(X)\right) \\ &= \int P_\Phi(T|X)P(X)\log\frac{P(T|X)}{P(T)}\,\mathrm{d}x\,\mathrm{d}t \\ &= \mathbb{E}_{P(X)}D_{KL}\left(P(T|X)\|P(T)\right) \end{aligned} \tag{8}$$

$$\begin{aligned} I(T;Y) &= D_{KL}\left(\left[\int P(T|Y,X)P(Y,X)\,\mathrm{d}x\right]\|P(T)P(Y)\right) \\ &= \int P(T|X,Y)P(X,Y)\log\frac{P(Y|T)P(T)}{P(T)P(Y)}\,\mathrm{d}t\,\mathrm{d}x\,\mathrm{d}y \\ &= \mathbb{E}_{P(X,Y)}\left[\int P(T|X,Y)\log P(Y|T)\,\mathrm{d}t\right] \\ &\quad - \mathbb{E}_{P(X,Y)}\left[\log P(Y)\int P(T|X,Y)\,\mathrm{d}t\right] \\ &= \mathbb{E}_{P(X,Y)}\mathbb{E}_{P(T|X,Y)}\log P(Y|T) + H(Y), \\ &= \mathbb{E}_{P(X,Y)}\mathbb{E}_{P(T|X)}\log P(Y|T) + H(Y), \end{aligned} \tag{9}$$

where the last equality in Equation (9) follows from the Markov assumption $T - X - Y$ in the information bottleneck model: $P(T|X,Y) = P(T|X)$. The conditional $Y|T$ is computed by sampling from the latent representation T as in the variational autoencoder [2]. Note that this form of the DVIB makes sure that one is only required to sample from the data distribution $P(X,Y)$, the variational decoder $P_\theta(Y|T)$, and the stochastic encoder $P_\phi(T|X)$—implemented as deep neural networks parametrised by θ and ϕ, respectively. In the latter, T depends only on X because of the $T - X - Y$ assumption.

2.2.1. Deep Copula Information Bottleneck

The authors of [3] argue that the entropy term $H(Y)$ in the last line of Equation (9) can be omitted, as Y is a constant. It has, however, been pointed out [7] that the IB solution should be invariant to monotonic transformations of both X and Y, since the problem is defined only in terms of mutual information which exhibits such invariance (i.e., $I(X;T) = I(f(X);T)$ for an invertible f). The term remaining in Equation (9) after leaving out $H(Y)$ does not have this property. Furthermore, problems limiting the DVIB when specifying marginal distributions of $T|X$ and $Y|T$ in Equations (8) and (9) have been identified [7]. These considerations have lead to the formulation of the deep copula information bottleneck, where the data are subject to the following transformation $\tilde{X} = \Phi^{-1}(\hat{F}(X))$, where Φ and $\hat{F}$ are the Gaussian and empirical cumulative distribution functions, respectively. This transformation makes them depend only on their copula and not on the marginals. This has also been shown to result in superior interpretability and disentanglement of the latent space T.

2.3. Bounds on Mutual Information in Deep Latent Variable Models

The deep information bottleneck model can be thought of as an extension of the VAE. Indeed, one can incorporate a variational approximation $Q(Y|T)$ of the posterior $P(Y|T)$ to Equation (9) and by $D_{KL}(Q(Y|T)\;||\;P(Y|T)) \geq 0$ obtain $I(T;Y) \geq \mathbb{E}_{P(X,Y)}\mathbb{E}_{P(T|X)}\log Q(Y|T) + H(Y)$ [3]. A number of other

bounds and approximations of mutual information have been considered in the literature. Many of them are motivated by obtaining a better representation of the latent space T. The article [12] considers different encoding distributions $Q(T|X)$ and derives a common bound for $I(X;T)$ on the rate-distortion plane. The authors subsequently extend this bound to the case where it is independent of the sample, which makes it possible to compare VAE and generative adversarial networks [13].

The authors of [14] use a Gaussian relaxation of the mutual information terms in the information bottleneck to bound them from below. They then proceed to compare the resulting method to canonical correlation analysis [15].

Extensions of generative models with an explicit regularisation in the form of a mutual information term have been proposed [16,17]. In the latter, an explicit lower bound on the mutual information between the latent space T and the generator network is derived.

Similarly, implicit regularisation of generative models in the form of dropout has been shown to be equivalent to the deep information bottleneck model [18,19]. The authors also mention that both Markov properties should hold in the IB solution, and note that $T - X - Y$ is enforced by construction, while $X - T - Y$ is only approximated by the optimal joint distribution of X, Y, and T. They do not, however, analyse the impact of both Markov assumptions and the relationship between them.

3. The Difference between Information Bottleneck Models

In this section, we focus on the difference between the original and deep IB models. First, we examine how the different Markov assumptions lead to different forms that the $I(Y;T)$ term admits. In Section 3.2, we consider both models and show that describing them as directed graphical models makes it possible to elucidate a fundamental property shared by all IB models. We then proceed to summarise the comparison in Section 3.3.

3.1. Clarifying the Discrepancy between the Assumptions in IB and DVIB

3.1.1. Motivation

The derivation of the deep variational information bottleneck model described in Section 2.2 uses the Markov assumption $T - X - Y$ (last line of Equation (9), Figure 1a). At the same time, by construction, the model adheres to the data generating process described by the following structural equations (η_T, η_Y are noise terms independent of X and T, respectively):

$$\begin{aligned} T &= f_T(X, \eta_T), \\ Y &= f_Y(T, \eta_Y). \end{aligned} \tag{10}$$

This implies that the Markov chain $X - T - Y$ is satisfied in the model, too (Figure 1b). Requiring that both Markov chains hold in the resulting joint distribution $P(X, Y, T)$ can be overly restrictive (note that no directed acyclic graph with three vertices to which such a distribution is faithful exists). Thus, the question of whether the $T - X - Y$ property in DVIB can be lifted arises. In what follows, we show that it is indeed possible.

Recall from Section 2.2 that the DVIB model relies on sampling only from the data $P(X, Y)$, encoder $P(T|X)$, and decoder $P(Y|T)$. Therefore, for optimising the latent IB, we want to avoid specifying the full conditional $P(T|X, Y)$, since this would require us to explicitly model the joint influence of both X and Y on T (which might be a complex distribution). We now proceed to show how to bound $I(T;Y)$ in a way that only involves sampling from the encoder $P(T|X)$ and circumvents modelling $P(T|X, Y)$ without using the $T - X - Y$ assumption.

(a) The original IB assumption. **(b)** The DVIB assumption.

Figure 1. Markov assumptions for the information bottleneck and the deep information bottleneck.

3.1.2. Bound Derivation

First, adopt the mutual information $I(T;Y)$ from the penultimate line of Equation (9) (i.e., without assuming the $T-X-Y$ property):

$$I(T;Y) = \mathbb{E}_{P(X,Y)}\mathbb{E}_{P(T|X,Y)} \log P(Y|T) + H(Y) \tag{11}$$

Now, rewrite Equation (11) using $X-T-Y$ (i.e., $X \perp\!\!\!\perp Y \mid T$: X and Y are conditionally independent given T):

$$\begin{aligned} I(T;Y) &= \mathbb{E}_{P(X,Y)}\mathbb{E}_{P(T|X,Y)} \log P(Y|T) + H(Y) \\ &= \mathbb{E}_{P(X)}\mathbb{E}_{P(Y|X)}\mathbb{E}_{P(T|X,Y)} \log P(Y|T,X) + H(Y). \end{aligned} \tag{12}$$

Focusing on $\mathbb{E}_{P(Y|X)}\mathbb{E}_{P(T|X,Y)} \log P(Y|T,X)$ in Equation (12), we obtain:

$$\begin{aligned} &\mathbb{E}_{P(Y|X)}\mathbb{E}_{P(T|X,Y)} \log P(Y|T,X) = \int\int P(T,Y|X) \log P(T,Y|X) \mathrm{d}t\mathrm{d}y \\ &= \int\int P(T,Y|X) \log \frac{P(Y|X)P(T,Y|X)}{P(Y|X)P(T|X)} \mathrm{d}y\mathrm{d}t \\ &= D_{KL}\Big(P(Y,T|X) \| P(Y|X)P(T|X)\Big) + \int\int P(T,Y|X) \log P(Y|X) \mathrm{d}t\mathrm{d}y \\ &= D_{KL}\Big(P(Y,T|X) \| P(Y|X)P(T|X)\Big) + \int P(Y|X) \log P(Y|X) \mathrm{d}y \\ &= D_{KL}\Big(P(Y,T|X) \| P(Y|X)P(T|X)\Big) + \int\int P(Y|X)P(T|X) \log P(T|X) \mathrm{d}t\mathrm{d}y \\ &= D_{KL}\Big(P(Y,T|X) \| P(Y|X)P(T|X)\Big) \\ &\qquad + \int\int P(Y|X)P(T|X) \log \frac{P(T|X)P(Y|X)P(T,Y|X)}{P(T,Y|X)P(T|X)} \mathrm{d}t\mathrm{d}y \\ &= D_{KL}\Big(P(Y,T|X) \| P(Y|X)P(T|X)\Big) + D_{KL}\Big(P(Y|X)P(T|X) \| P(Y,T|X)\Big) \\ &\qquad + \mathbb{E}_{P(T|X)P(Y|X)} \log P(Y|T,X) \\ &\geq \mathbb{E}_{P(T|X)P(Y|X)} \log P(Y|T,X). \end{aligned} \tag{13}$$

Plugging Equation (13) into Equation (12), i.e., averaging over X, and using $X \perp\!\!\!\perp Y \mid T$ again, we arrive at:

$$\begin{aligned} I(T;Y) &= \mathbb{E}_{P(X)}\mathbb{E}_{P(Y|X)}\mathbb{E}_{P(T|X,Y)} \log P(Y|T) + H(Y) \\ &= \mathbb{E}_{P(X)}\mathbb{E}_{P(Y|X)}\mathbb{E}_{P(T|X)} \log P(Y|T) \\ &\quad + \mathbb{E}_{P(X)} D_{KL}(P(Y,T|X) \| P(Y|X)P(T|X)) \\ &\quad + \mathbb{E}_{P(X)} D_{KL}(P(Y|X)P(T|X) \| P(Y,T|X)) + H(Y) \\ &= \mathbb{E}_{P(X)}\mathbb{E}_{P(Y|X)}\mathbb{E}_{P(T|X)} \log P(Y|T) \\ &\quad + I(Y;T|X) + L(Y;T|X) + H(Y) \\ &\geq \mathbb{E}_{P(X)}\mathbb{E}_{P(Y|X)}\mathbb{E}_{P(T|X)} \log P(Y|T) + H(Y). \end{aligned} \tag{14}$$

3.1.3. Interpretation

According to Equation (14), the mutual information $I(T;Y)$ consists of three terms: its lower bound $\mathbb{E}_{P(X)}\mathbb{E}_{P(Y|X)}\mathbb{E}_{P(T|X)} \log P(Y|T)$ which is actually optimised in DVIB and its extensions, $I(Y;T|X) + L(Y;T|X)$ which are 0 when both Markov assumptions are satisfied, and the entropy term $H(Y)$.

Equation (14 shows how to bound the mutual information term $I(T;Y)$ in the IB model (Equation (4)) so that the value of the bound depends on the data $P(X,Y)$ and marginals $T|X$, $Y|T$ without using the Markov assumption $T - X - Y$. If we again implement the marginal distributions as deep neural nets $P_\phi(T|X)$ and $P_\theta(Y|T)$, (Equation (14) provides the lower bound which is actually optimised in DVIB (Equation (9)). By training the networks, we find parameters ϕ and θ in $P_\phi(T|X)$ and $P_\theta(Y|T)$ such that both $I(Y;T|X)$ and $L(Y;T|X)$ are close to zero. The terms $I(Y;T|X)$ and $L(Y;T|X)$ can thus be interpreted as a measure of how much the original IB assumption $T - X - Y$ is violated during the training of the model that implements $X - T - Y$ by construction.

The difference between the original IB and DVIB is that in the former, $T - X - Y$ is used to derive the general form of the solution T, while $X - T - Y$ is approximated as closely as possible by T (as noted in [18]. In the latter, $X - T - Y$ is forced by construction, and $T - X - Y$ is approximated by optimising the lower bound given by Equation (14). The "distance" to a distribution satisfying both assumptions is measured by the tightness of the bound.

3.2. The Original IB Assumption Revisited

3.2.1. Motivation for Conditional Independence Assumptions in Information Bottleneck Models

In the original formulation of the information bottleneck (Section 2.1 and Equation (4)), given by $\min_{P(T|X)} I(X;T) - \beta I(T;Y)$, one optimises over $P(T|X)$ while disregarding any dependence on Y. This suggests that the defining feature of the IB model is the absence of a direct functional dependence of T on Y. This can be achieved, e.g., by the first structural equation in Equation (10):

$$T = f_T(X, \eta_T). \tag{15}$$

That means any influence of Y on T must go through X. Note that this is implied by the original IB assumption $T - X - Y$, but not the other way around. In particular, the model given by $X - T - Y$ can also be parametrized such that there is no direct dependence of T on Y, as in, e.g., Equation (10). This means that DVIB, despite optimising a lower bound on the IB, implements the defining feature of IB as well.

3.2.2. Information Bottleneck as a Directed Graphical Model

The above discussion leads to the conclusion that the IB assumptions might also be described by directed graphical models. Such models encode conditional independence relations with d-separation (for the definition and examples of d-separation in directed acyclic graphs, see [20] or [21] (Chapters 1.2.3 and 11.1.2)). In particular, any pair of variables d-separated by Z is conditionally independent given Z. The arrows of the directed acyclic graph (DAG) are assumed to correspond to the data generating process described by a set of structural equation (as in Equation (10)). Therefore, the following probability factorisation and data generating process hold for a DAG model:

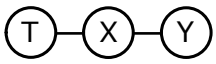
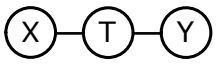
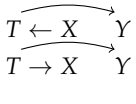
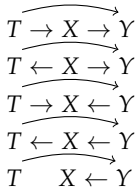
$$P(X_1, X_2, \ldots, X_n) = \prod_i P(X_i | pa(X_i)) \tag{16}$$

$$X_i = f_i(pa(X_i), U_i), \tag{17}$$

where $pa(X_i)$ stands for the set of direct parents of X_i and U_i are exogenous noise variables.

Let us now focus again on the motivation for the $T - X - Y$ assumption in Equation (4). It prevents the model from choosing a degenerate solution of $T = Y$ (in which case $I(X;T) = \text{const.}$ and $I(T;Y) = \infty$). Note, however, that while $T - X - Y$ is a sufficient condition for such a solution to be excluded (which justifies the correctness of the original IB), the necessary condition is that T cannot depend directly on Y. This means that the IB Markov assumption can be indeed reduced to requiring the absence of a direct arrow from Y to T in the underlying DAG. Note that this can be achieved in the undirected $X - T - Y$ model too. One thus wishes to avoid degenerate solutions which impair the bottleneck nature of T: it should contain information about both X and Y, the trade-off between them being steered by β. It is therefore necessary to exclude DAG structures which encode independence of X and T as well as Y and T. Such independences are achieved by collider structures (with two different variables pointing towards a common child) in DAGs; i.e., $T \to Y \leftarrow X$ and $T \to X \leftarrow Y$ (they lead to degenerate solutions of $I(X;T) = 0$ and $I(T;Y) = 0$, respectively). To sum up, the goal of asserting the conditional independence assumption in Equation (4) is to avoid degenerate solutions which impair the bottleneck nature of the representation T. When modelling the information bottleneck with DAG structures, one has to exclude the arrow $Y \to T$ and collider structures. A simple enumeration of the possible DAG models for the information bottleneck results in 10 distinct models listed in Table 1.

Table 1. Directed graphical models of the information bottleneck.

Defining Markov Assumption			**Other/None**
Admissible DAG models	$T \to X \to Y$ $T \leftarrow X \to Y$ $T \leftarrow X \leftarrow Y$		

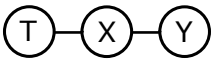
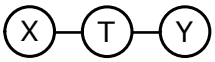
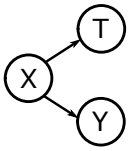
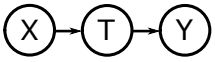
As can be seen, considering the information bottleneck as a directed graphical model (DAG) makes room for a family of models which fall into three broad categories, satisfying one of the two undirected Markov assumptions $T - X - Y$ or $X - T - Y$, as described in Section 3.1, or neither of them (see Table 1).

The difference between particular models lies in the necessity to specify different conditional distributions and parametrise them, which might lead to situations in which no joint distribution $P(X, Y, T)$ exists (which is likely to be the case in the third category). Focusing on the two first categories, we see that the former corresponds to the standard parametrisations of the information bottleneck and the Gaussian information bottleneck (see Section 2.1). In the latter, we see the deep information bottleneck (Equation (10)) as the first DAG. Note also that the second DAG satisfying the $X - T - Y$ assumption in Table 1 defines the probabilistic CCA model [22]. This is not surprising, since the solutions of CCA and the Gaussian information bottleneck use eigenvectors of the same matrix [5].

3.3. Comparing IB and DVIB Assumptions

The original and deep information bottleneck models differ by using different Markov assumptions (see Figure 1) in the derivation of the respective solutions. As demonstrated in Section 3.1, DVIB optimises a lower bound on the objective function of IB. The tightness of the bound measures to what extent the IB assumption (Figure 1a) is violated. As described in Section 3.2, characterising both models as directed graphical models results in two different DAGs for the IB and DVIB. Both models are summarised in Table 2.

Table 2. Comparison of the information bottleneck and deep variational information bottleneck.

	Information Bottleneck (IB)	**Deep Information Bottleneck (DVIB)**
Assumed Markov chain		
Possible set of structural equations	$T = f_T(X, \eta_T)$, $Y = f_Y(X, \eta_Y)$	$T = f_T(X, \eta_T)$, $Y = f_Y(T, \eta_Y)$
Corresponding DAG		
Optimised term corresponding to $I(T;Y)$	$\mathbb{E}_{P(X,Y)}\mathbb{E}_{P(T\|X)} \log P(Y\|T)$ $+ I(Y;T\|X) + L(Y;T\|X)$ $+ H(Y)$	$\mathbb{E}_{P(X,Y)}\mathbb{E}_{P(T\|X)} \log P(Y\|T)$ $+ H(Y)$

4. Conclusions

In this paper, we showed how to lift the information bottleneck's Markov assumption $T - X - Y$ in the context of the deep information bottleneck model, in which $X - T - Y$ holds by construction. This result explains why standard implementations of the deep information bottleneck can optimise over a larger amount of joint distributions $P(X, T, Y)$ while only specifying the marginal $T|X$. It is made possible by optimising the lower bound on the mutual information $I(T;Y)$ provided here, rather than the full mutual information. We also provided a description of the information bottleneck as a DAG model and showed that it is possible to identify a fundamental necessary feature of the IB in the language of directed graphical models. This property is satisfied for both the original and deep information bottlenecks.

Author Contributions: Conceptualization, A.W.; Supervision, V.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the Swiss National Science Foundation grant number CR32I2159682.

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

References

1. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2014; pp. 2672–2680.
2. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
3. Alemi, A.A.; Fischer, I.; Dillon, J.V.; Murphy, K. Deep variational information bottleneck. *arXiv* **2016**, arXiv:1612.00410.
4. Tishby, N.; Pereira, F.C.; Bialek, W. The information bottleneck method. *arXiv* **2000**, arXiv:physics/0004057.
5. Chechik, G.; Globerson, A.; Tishby, N.; Weiss, Y. Information bottleneck for Gaussian variables. *J. Mach. Learn. Res.* **2005**, *6*, 165–188.
6. Rey, M.; Roth, V.; Fuchs, T. Sparse meta-Gaussian information bottleneck. In Proceedings of the 31st International Conference on Machine Learning (ICML-14), Beijing, China, 21–26 June 2014; pp. 910–918.
7. Wieczorek, A.; Wieser, M.; Murezzan, D.; Roth, V. Learning Sparse Latent Representations with the Deep Copula Information Bottleneck. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
8. Tishby, N.; Zaslavsky, N. Deep learning and the information bottleneck principle. In Proceedings of the 2015 IEEE Information Theory Workshop, ITW, Jerusalem, Israel, 26 April–1 May 2015; pp. 1–5, doi:10.1109/ITW.2015.7133169.
9. Shwartz-Ziv, R.; Tishby, N. Opening the black box of deep neural networks via information. *arXiv* **2017**, arXiv:1703.00810.
10. Saxe, A.M.; Bansal, Y.; Dapello, J.; Advani, M.; Kolchinsky, A.; Tracey, B.D.; Cox, D.D. On the information bottleneck theory of deep learning. *J. Stat. Mech.-Theory Exp.* **2019**, *2019*, 124020.
11. Palomar, D.P.; Verdú, S. Lautum information. *IEEE Trans. Inf. Theory* **2008**, *54*, 964–975.
12. Alemi, A.A.; Poole, B.; Fischer, I.; Dillon, J.V.; Saurous, R.A.; Murphy, K. Fixing a broken ELBO. *arXiv* **2017**, arXiv:1711.00464.
13. Alemi, A.A.; Fischer, I. GILBO: One metric to measure them all. *arXiv* **2018**, arXiv:1802.04874.
14. Painsky, A.; Tishby, N. Gaussian lower bound for the information bottleneck limit. *J. Mach. Learn. Res.* **2017**, *18*, 7908–7936.
15. Hotelling, H. Relations between two sets of variates. In *Breakthroughs in Statistics*; Springer: Berlin, Germany, 1992; pp. 162–190.
16. Zhao, S.; Song, J.; Ermon, S. Infovae: Balancing learning and inference in variational autoencoders. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 5885–5892.
17. Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv* **2016**, arXiv:1606.03657.
18. Achille, A.; Soatto, S. Information dropout: Learning optimal representations through noisy computation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 2897–2905.
19. Achille, A.; Soatto, S. Emergence of invariance and disentanglement in deep representations. *J. Mach. Learn. Res.* **2018**, *19*, 1947–1980.
20. Lauritzen, S.L. *Graphical Models*; Clarendon Press: Oxford, UK, 1996; Volume 17.

21. Pearl, J. *Causality*; Cambridge University Press: Cambridge, UK, 2009.
22. Bach, F.R.; Jordan, M.I. *A Probabilistic Interpretation of Canonical Correlation Analysis*; University of California: Berkeley, CA, USA, 2005.