# Discovering Interpretable Representations for Both Deep Generative and Discriminative Models

**Tameem Adel** [1 2] **Zoubin Ghahramani** [1 2 3] **Adrian Weller** [1 2 4]

## Abstract

Interpretability of representations in both deep generative and discriminative models is highly desirable. Current methods jointly optimize an objective combining accuracy and interpretability. However, this may reduce accuracy, and is not applicable to already trained models. We propose two interpretability frameworks. First, we provide an interpretable lens for an existing model. We use a generative model which takes as input the representation in an existing (generative or discriminative) model, weakly supervised by limited side information. Applying a flexible and invertible transformation to the input leads to an interpretable representation with no loss in accuracy. We extend the approach using an active learning strategy to choose the most useful side information to obtain, allowing a human to guide what "interpretable" means. Our second framework relies on joint optimization for a representation which is both maximally informative about the side information and maximally compressive about the non-interpretable data factors. This leads to a novel perspective on the relationship between compression and regularization. We also propose a new interpretability evaluation metric based on our framework. Empirically, we achieve state-of-the-art results on three datasets using the two proposed algorithms.

## 1. Introduction

Learning interpretable data representations is becoming ever more important as machine learning models grow in size and complexity, and as applications reach critical social, economic and public health domains. In addition to understanding the model and gaining insights about the corresponding application, learning interpretable data representations can improve the model's generalization power. As such, representations that can efficiently be used across a variety of tasks are those comprising disentangled latent variables such that each variable corresponds to a salient or meaningful data attribute (Bengio et al., 2013; Bengio, 2009).

Generative modeling constitutes one of the most influential approaches to representation learning. Generative models seek to infer the data-generating latent space, which implies capturing to some extent the salient characteristics of such data. This motivates the current consensus in the field that generative models can potentially provide interpretable and disentangled data representations (Kingma et al., 2014; Chen et al., 2016; Desjardins et al., 2012; Higgins et al., 2017; Kulkarni et al., 2015). Deep generative models grant further flexibility to the learning and inference procedures via utilizing neural networks for parameter estimation.

Variational autoencoders (VAEs, Kingma & Welling, 2014; Kingma et al., 2014) and generative adversarial networks (GANs, Goodfellow et al., 2014; Goodfellow, 2016) are considered two important models. The original versions of these two algorithms solely optimize for data reconstruction fidelity. Several works that followed (Kulkarni et al., 2015; Hsu et al., 2017; Chen et al., 2016; Siddharth et al., 2017; Higgins et al., 2017) instead jointly optimize the latent representation for both reconstruction fidelity and interpretability. Further elaboration on related works is provided in Section 9 of the Appendix.

Optimizing for interpretability can be costly. It is evident that a latent representation optimized solely for reconstruction fidelity can typically be better at fitting the data than one optimized for both data reconstruction fidelity and interpretability. We propose an algorithm that aims at providing an interpretable representation without losing ground on reconstruction fidelity. The first introduced algorithm can be seen as a generalization of variational autoencoders that can be applied as an interpretable lens on top of an existing model of many types. Such an existing model is the input to the algorithm, along with a limited amount of side information denoting salient data attributes. The resulting

[1]University of Cambridge, UK [2]Leverhulme CFI, Cambridge, UK [3]Uber AI Labs, USA [4]The Alan Turing Institute, UK. Correspondence to: Tameem Adel <tah47@cam.ac.uk>.

*Proceedings of the $35^{th}$ International Conference on Machine Learning*, Stockholm, Sweden, PMLR 80, 2018. 

latent space is assumed to generate both the input representation and the side information (Figure 1). The dependence between the resulting and input representations is modeled by a flexible and invertible transformation. Since we can always recover the input representation, reconstruction fidelity is not at risk of being sacrificed. We define interpretability in this work as *a simple relationship to something we can understand*. Consequently, the dependence between the latent representation and the side information is optimized to be simple and linear, i.e. interpretable.

Having humans in the interpretability loop is crucial for its success. Thus, we propose an active learning strategy that defines a mutual information based criterion upon which side information of chosen data points is obtained. We also propose a metric to evaluate the degree of interpretability of a representation. Even though there are numerous works in the literature aiming at inferring interpretable latent spaces (Chen et al., 2016; Higgins et al., 2017; Kulkarni et al., 2015; Siddharth et al., 2017), there is a striking shortage of relevant metrics. As opposed to our more widely applicable metric, the one provided in Higgins et al. (2017) can only be applied when it is feasible to generate artificial data samples from controlled interpretable factors.

Consistent with the current trend in deep interpretable generative models, our second proposed interpretability framework is based on joint optimization for data reconstruction and interpretability. Through this method, we provide a novel perspective linking the notions of compression and regularization. We prove that a model which is maximally compressive about non-interpretable data factors, and which aims at fitting the data, is equivalent to a model fitting the data with further regularization constraints. We subsequently analyze the performance of both algorithms.

We make the following contributions: 1) We propose an interpretability framework that can be applied as a lens on an existing model of many types. To the best of our knowledge, the flexibility provided by explaining an existing model without having to redo the data fitting and without affecting accuracy represents a new direction in interpretable models (Section 2); 2) We propose an active learning methodology which bases the acquisition function on having high mutual information with interpretable data attributes (Section 3); 3) We propose a quantitative metric to evaluate the degree of interpretability of an inferred latent representation (Section 4); 4) We propose another interpretability framework jointly optimized for reconstruction and interpretability (Section 5). A novel analogy between data compression and regularization is derived under this framework; 5) Our qualitative and quantitative state-of-the-art results on three datasets demonstrate the effectiveness of the two proposed methods (Section 6).

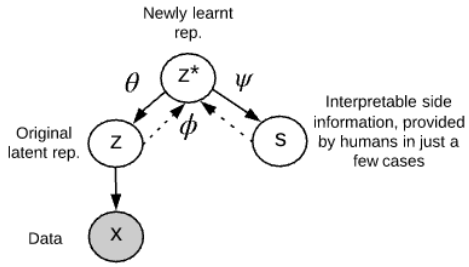



*Figure 1.* The proposed interpretable lens variable model (ILVM). The latent space $\mathbf{z}$ has been inferred (and optimized for data reconstruction) via a VAE prior to this procedure. The goal of ILVM is to provide an interpretable lens on top of $\mathbf{z}$. The latent space $\mathbf{z}^*$ is the space optimized for interpretability. Normalizing flows, which are a series of invertible transformations, are used to transform the density of $\mathbf{z}$, whereas $\mathbf{z}^*$ is also optimized for interpretability via having a simple linear, i.e. interpretable, dependence with $\mathbf{s}$. The generative parameters are $\boldsymbol{\theta}$ and $\psi$, whereas $\boldsymbol{\phi}$ are the parameters of the recognition model.

## 2. An Interpretable Lens on an Existing Model

An important goal of deep generative models is to fit the data with high fidelity. Our principal aim in this proposed generative model is to learn an interpretable representation without degrading the reconstruction fidelity. We refer to it as the Interpretable Lens Variable Model (ILVM). A diagram of the model is displayed in Figure 1. Separating the reconstruction fidelity and interpretability optimization procedures permits the latter to improve the interpretability of different types of existing models, e.g. another latent representation already optimized for reconstruction fidelity, a hidden layer of a neural network optimized for classification, etc. Moreover, since the transformation from the input representation $\mathbf{z}$ to the resulting representation $\mathbf{z}^*$ is invertible, we can recover $\mathbf{z}$.

We assume that the input representation $\mathbf{z}$ has been inferred prior to the ILVM procedure, and that a standard VAE (Kingma & Welling, 2014) has been used to infer $\mathbf{z}$. Up to this point we have not lost any ground on data reconstruction, but we have also not gained anything yet in terms of interpretability. Transforming $\mathbf{z}$ into $\mathbf{z}^*$ via a powerful (nonlinear) and invertible transformation, so that $\mathbf{z}^*$ is interpretable, can get the best of both worlds. Thus, the main aims of the remaining steps are to ensure that: i) the latent space $\mathbf{z}^*$ is optimized for interpretability. This is achieved via a flexible transformation from $\mathbf{z}$ to $\mathbf{z}^*$ that can capture the nonlinear dependencies among the factors of $\mathbf{z}$; ii) the transformation is invertible.

A major issue is how to algorithmically optimize $\mathbf{z}^*$ for interpretability. We assume the existence of limited side information in the form of salient attributes observed for

a small subset of the data instances. A linear, i.e. simple, relationship between $\mathbf{z}^*$ and the side information $\mathbf{s}$ signifies an acceptable degree of interpretability of $\mathbf{z}^*$.

In order to transform the density of $\mathbf{z}$ into $\mathbf{z}^*$, we use normalizing flows (Tabak & Turner, 2013; Tabak & Vanden-Eijnden, 2010; Rezende & Mohamed, 2015; Kingma et al., 2016), which are powerful and invertible transformations used to construct flexible posterior distributions through a series of mappings. A normalizing flow (NF) transformation applies a chain of invertible parameterized transformations, $\mathbf{f_t}$, $\mathbf{t} = \mathbf{1}, \ldots, \mathbf{T}$, to its input ($\mathbf{z}$ in this case) such that the outcome of the last iteration, $\mathbf{z}^* = \mathbf{z_T}$ has a more flexible distribution that in our algorithm is optimized for interpretability. The transformation from $\mathbf{z}$ to $\mathbf{z}^*$ goes through $\mathbf{T}$ steps, where each step is indexed by $\mathbf{t}$ and where $\mathbf{z_0}$ is an initial random variable with density $\mathbf{q_0}(\mathbf{z_0})$ that is successively transformed, along with the old representation $\mathbf{z}$, through a chain of transformations $\mathbf{f_1}, \ldots, \mathbf{f_T}$ (Rezende & Mohamed, 2015):

$$\mathbf{z} \sim \mathbf{q}(\mathbf{z}|\mathbf{x}), \tag{1}$$

$$\mathbf{z_t} = \mathbf{f_t}(\mathbf{z_{t-1}}, \mathbf{z}) \quad \forall \mathbf{t} = \mathbf{1} \ldots \mathbf{T}, \quad \mathbf{z}^* = \mathbf{z_T}. \tag{2}$$

The probability density function of the final latent representation, $\mathbf{z}^* = \mathbf{z_T}$, can be computed provided that the determinant of the Jacobian of each of the transformations, $\mathbf{det}(\mathbf{f_t})$, can be computed. Assuming that $\mathbf{q}$ refers to the estimated approximation of a ground truth distribution (where the approximation is via normalizing flows except for $\mathbf{z} \sim \mathbf{q}(\mathbf{z}|\mathbf{x})$ for which a standard VAE is used), the probability density of $\mathbf{q}(\mathbf{z}^*|\mathbf{z})$ can be expressed as follows:

$$\log \mathbf{q_T}(\mathbf{z_T}|\mathbf{z}) = \log \mathbf{q_0}(\mathbf{z_0}|\mathbf{z}) - \sum_{\mathbf{t=1}}^{\mathbf{T}} \log \det \left| \frac{d\mathbf{z_t}}{d\mathbf{z_{t-1}}} \right|,$$

$$\mathbf{z}^* = \mathbf{z_T}. \tag{3}$$

Several types of normalizing flows can be utilized; we opt for planar flows. We empirically tried radial flows but they did not improve the results. Planar flows have the form:

$$\mathbf{f_t}(\mathbf{z_{t-1}}) = \mathbf{z_{t-1}} + \mathbf{u}\mathbf{h}(\mathbf{w^T}\mathbf{z_{t-1}} + \mathbf{b}), \tag{4}$$

where $\mathbf{u}$ and $\mathbf{w}$ are vectors, $\mathbf{b}$ is a scalar, $\mathbf{w^T}$ refers to the transpose of $\mathbf{w}$, and $\mathbf{h}$ is a nonlinearity. Each map from $\mathbf{z}$ to $\mathbf{z}^*$ has the form given in (4). Thus:

$$\mathbf{z}^* = \mathbf{f_T} \circ \mathbf{f_{T-1}} \circ \ldots \circ \mathbf{f_1}(\mathbf{z}). \tag{5}$$

For the proposed framework, there is a generative model with parameters $\boldsymbol{\theta}$ and $\psi$ where $\mathbf{z}^*$ is assumed to generate both $\mathbf{s}$ and $\mathbf{z}$. We also estimate the variational parameters $\boldsymbol{\phi}$ of the distributions of the recognition (inference) model approximating the true posterior, e.g. $\mathbf{q_\phi}(\mathbf{z}^*|\mathbf{z}, \mathbf{s})$. We follow the standard variational principle and derive the evidence lower bound (ELBO) of the proposed variational objective for the model in Figure 1. For data points which have observed side information $\mathbf{s}$, the marginal likelihood of a point according to the proposed model is as follows:

$$\begin{aligned}
&\log \mathbf{p_\theta}(\mathbf{z}, \mathbf{s}) = \log \int_{\mathbf{z}^*} \mathbf{p_\theta}(\mathbf{z}, \mathbf{s}, \mathbf{z}^*)\, d\mathbf{z}^* \\
&= \log \int_{\mathbf{z}^*} \mathbf{p}(\mathbf{z}^*)\mathbf{p_\theta}(\mathbf{z}|\mathbf{z}^*)\mathbf{p}_\psi(\mathbf{s}|\mathbf{z}^*) \frac{\mathbf{q_\phi}(\mathbf{z}^*|\mathbf{z}, \mathbf{s})}{\mathbf{q_\phi}(\mathbf{z}^*|\mathbf{z}, \mathbf{s})}\, d\mathbf{z}^* \\
&= \log \mathbb{E}_{\mathbf{q_\phi}(\mathbf{z}^*|\mathbf{z},\mathbf{s})}[\mathbf{p}(\mathbf{z}^*)\mathbf{p_\theta}(\mathbf{z}|\mathbf{z}^*)\mathbf{p}_\psi(\mathbf{s}|\mathbf{z}^*)/\mathbf{q_\phi}(\mathbf{z}^*|\mathbf{z}, \mathbf{s})] \\
&\geq \mathbb{E}_{\mathbf{q_\phi}(\mathbf{z}^*|\mathbf{z},\mathbf{s})}[\log \mathbf{p}(\mathbf{z}^*) + \log \mathbf{p_\theta}(\mathbf{z}|\mathbf{z}^*) + \log \mathbf{p}_\psi(\mathbf{s}|\mathbf{z}^*) \\
&\quad - \log \mathbf{q_\phi}(\mathbf{z}^*|\mathbf{z}, \mathbf{s})].
\end{aligned} \tag{6}$$

Using (3), properties of the normalizing flows and the fact that $\mathbf{q_\phi}(\mathbf{z}^*|\mathbf{z}, \mathbf{s}) = \mathbf{q_T}(\mathbf{z_T}|\mathbf{z}, \mathbf{s})$, then from (6) we obtain:

$$\begin{aligned}
\log \mathbf{p_\theta}(\mathbf{z}, \mathbf{s}) \geq \mathbb{E}_{\mathbf{q_0}(\mathbf{z_0})}[&\log \mathbf{p}(\mathbf{z_T}) + \log \mathbf{p_\theta}(\mathbf{z}|\mathbf{z_T}) \\
&+ \log \mathbf{p}_\psi(\mathbf{s}|\mathbf{z_T}) - \mathbf{q_0}(\mathbf{z_0}|\mathbf{z}, \mathbf{s}) + \sum_{\mathbf{t=1}}^{\mathbf{T}} \log \det \left| \frac{d\mathbf{z_t}}{d\mathbf{z_{t-1}}} \right|].
\end{aligned} \tag{7}$$

For data points without observed side information $\mathbf{s}$, $\mathbf{s}$ is treated as a latent variable. The corresponding ELBO is:

$$\begin{aligned}
&\log \mathbf{p_\theta}(\mathbf{z}) = \log \int_{\mathbf{s},\mathbf{z}^*} \mathbf{p_\theta}(\mathbf{z}, \mathbf{s}, \mathbf{z}^*)\, d\mathbf{s}\, d\mathbf{z}^* \\
&= \log \mathbb{E}_{\mathbf{q_\phi}(\mathbf{s},\mathbf{z}^*|\mathbf{z})}[\mathbf{p}(\mathbf{z}^*)\mathbf{p_\theta}(\mathbf{z}|\mathbf{z}^*)\mathbf{p}_\psi(\mathbf{s}|\mathbf{z}^*)/\mathbf{q_\phi}(\mathbf{s}, \mathbf{z}^*|\mathbf{z})] \\
&\geq \mathbb{E}_{\mathbf{q_\phi}(\mathbf{s},\mathbf{z}^*|\mathbf{z})}[\log \mathbf{p}(\mathbf{z}^*) + \log \mathbf{p_\theta}(\mathbf{z}|\mathbf{z}^*) + \log \mathbf{p}_\psi(\mathbf{s}|\mathbf{z}^*) \\
&\quad - \log \mathbf{q_\phi}(\mathbf{s}, \mathbf{z}^*|\mathbf{z})].
\end{aligned} \tag{8}$$

The inequality in (6) and (8) is due to Jensen's inequality. We restrict the probability distribution $\mathbf{p}_\psi(\mathbf{s}|\mathbf{z}^*)$ to express a simple linear, i.e. interpretable, relationship between $\mathbf{s}$ and $\mathbf{z}^*$. For a $k$-dimensional $\mathbf{z}^*$, the distribution $\mathbf{p}_\psi(\mathbf{s}|\mathbf{z}^*)$ depends on $\sum_{\mathbf{j}=1}^{k} \psi_\mathbf{j}\mathbf{z}^*_\mathbf{j} + \psi_0$. Hence, the parameters $\psi$ include $\psi_0, \psi_1, \ldots, \psi_k$.

Similar to several recent frameworks, the inference cost can be amortised via the parameters of a recognition network (Kingma & Welling, 2014; Rezende et al., 2014) instead of repeating the E-step for each data point, as in the variational EM algorithm. The recognition network is used to build a mapping from the model input $\mathbf{z}$ (and $\mathbf{s}$) to the parameters of the initial density $\mathbf{q_0}$ and all the other parameters of the normalizing flow, up until $\mathbf{q_\phi} = \mathbf{q_T}$. The key steps of the algorithm are shown in Algorithm 1 in the Appendix.

### 2.1. An Interpretable Lens for Existing Deep Models

The model input $\mathbf{z}$ in Figure 1 need not be an output of another deep generative model. It can as well be a hidden layer of a discriminative neural network. Interpretable insights into a hidden layer during classification is a potential gain. Only (1) changes whereas all the equations from (2) to (8) remain the same. We show an example of preliminary results in Figure 7 and Section 6.1.3.

## 3. Interactive Interpretability via Active Learning

In this section, we address an issue that we believe is important for the success of interpretable models, which is how to improve interpretability via an interactive procedure involving humans. Interpretability in machine learning refers to the ability to explain learnt information in understandable terms to humans. Hence, the participation of humans in evaluating and enhancing interpretability is fundamental.

We aim at enhancing interpretability through an active learning approach. Out of the pool of data points with hidden side information $\mathbf{s}$, a data point is chosen and its side information is revealed by an expert. We propose a mutual information based objective for choosing the next most useful data point. Our method is similar to Bayesian Active Learning by Disagreement (BALD) (Houlsby et al., 2011).

We propose an active learning methodology for the `ILVM` model. The proposed acquisition function notes that the most suitable data point to choose as the point whose hidden side information should be obtained, is the point with index $\mathbf{j}$ that maximizes the following:

$$\begin{aligned}\hat{\mathbf{j}} &= \operatorname{argmax}_{\mathbf{j}} \mathbf{I}(\mathbf{s_j}, \psi) = \mathbf{H}(\mathbf{s_j}) - \mathbb{E}_{\mathbf{q_\phi}(\mathbf{z}^*|\mathbf{s})}[\mathbf{H}(\mathbf{s_j}|\mathbf{z}^*_\mathbf{j})] \\ &= -\int \mathbf{p}(\mathbf{s_j}) \log \mathbf{p}(\mathbf{s_j})\, d\mathbf{s} \\ &\quad + \mathbb{E}_{\mathbf{q_\phi}(\mathbf{z}^*|\mathbf{s})}\left[\int \mathbf{p}_\psi(\mathbf{s_j}|\mathbf{z}^*) \log \mathbf{p}_\psi(\mathbf{s_j}|\mathbf{z}^*)\, d\mathbf{s}\right]. \quad (9)\end{aligned}$$

Since $\mathbf{p}(\mathbf{s_j}) = \int \mathbf{p}_\psi(\mathbf{s_j}|\mathbf{z}^*)\mathbf{p}(\mathbf{z}^*)\, d\mathbf{z}^*$, from (9) we obtain:

$$\begin{aligned}\hat{\mathbf{j}} &= \operatorname{argmax}_{\mathbf{j}} \mathbf{I}(\mathbf{s_j}, \psi) \\ &= -\int\int \mathbf{p}_\psi(\mathbf{s_j}|\mathbf{z}^*)\mathbf{p}(\mathbf{z}^*) d\mathbf{z}^* \ \log\Big(\int \mathbf{p}_\psi(\mathbf{s_j}|\mathbf{z}^*)\mathbf{p}(\mathbf{z}^*) d\mathbf{z}^*\Big)\, d\mathbf{s} \\ &\quad + \mathbb{E}_{\mathbf{q_\phi}(\mathbf{z}^*|\mathbf{s})}[\int \mathbf{p}_\psi(\mathbf{s_j}|\mathbf{z}^*_\mathbf{j}) \log \mathbf{p}_\psi(\mathbf{s_j}|\mathbf{z}^*_\mathbf{j}) d\,\mathbf{s}]. \quad (10)\end{aligned}$$

The reasoning behind the proposed active learning methodology is to choose the point possessing side information about which the model is most uncertain (maximized $\mathbf{H}(\mathbf{s_j})$) but in which the individual settings of the founding latent space $\mathbf{z}^*$ are confident (minimized $\mathbb{E}_{\mathbf{q_\phi}(\mathbf{z}^*|\mathbf{s})}[\mathbf{H}(\mathbf{s_j}|\mathbf{z}^*_\mathbf{j})]$). This can be seen as if we choose the point whose side information values are being disagreed about the most by the individual values of the latent space $\mathbf{z}^*$. The outcome of this procedure is a tractable estimator of $\hat{\mathbf{j}}$, the index of the next data point for which obtaining the side information $\mathbf{s_j}$ maximally enhances interpretability.

## 4. Evaluation of Interpretable Latent Variable Models

Despite the growing number of works on interpretability, there is no consensus on a good evaluation metric. To the best of our knowledge, the relevant metric that can be used in evaluating interpretable latent models is the one proposed in Higgins et al. (2017). However, it is conditioned on the availability of artificial samples generated by a simulator and on the ability to generate these samples from controlled interpretable factors. Moreover, the controlled factors cannot be discovered nor measured in cases when they are not statistically independent. We propose a metric that follows naturally from our modeling assumptions and that does not suffer from the aforementioned issues.

The gist of the proposed metric is based on the earlier notion that interpretability refers to *a simple relationship to something we understand*. In this sense, a latent space is (more) interpretable if it manages to explain the relationship to salient attributes (more) simply. The proposed metric evaluates the interpretability degree of a latent space based on a test sample $\mathbf{x_t}$ containing observed side information. Assuming that the resulting $\mathbf{z}^*$ consists of $k$ dimensions, our metric consists of two straightforward steps for each dimension $\mathbf{j}$ of the side information:

(i) Out of the $k$ latent dimensions, select the dimension $\mathbf{z}^*_\mathbf{i}$ which is maximally informative about the $\mathbf{j}^{\text{th}}$ side information factor, $\mathbf{s_j}$, i.e. select $\mathbf{z}^*_\mathbf{i}$ such that: $\mathbf{i} = \operatorname{argmax}_{\mathbf{i}} \mathbf{I}(\mathbf{s_j}, \mathbf{z}^*_\mathbf{i}|\mathbf{x_t})$.

(ii) Evaluate how interpretable $\mathbf{z}^*_\mathbf{i}$ is with respect to $\mathbf{s_j}$ by measuring $\mathbf{p}(\mathbf{s_j}|\mathbf{z}^*_\mathbf{i})$ where $\mathbf{p}$ is a simple, i.e. linear, probabilistic relationship. By summing the logarithms of the resulting probabilities corresponding to every test sample point for a dimension $\mathbf{j}$ of the side information, we get an interpretability score of the examined latent space with respect to a salient attribute.

Finally by aggregating the scores over all the dimensions of the side information $\mathbf{s}$, we obtain the interpretability score.

## 5. Joint Optimization of the Latent Variable Model

We propose another method where the two objectives, interpretability and reconstruction fidelity, are jointly optimized in the procedure of learning a latent variable model. We refer to this model as the Jointly Learnt Variable Model (`JLVM`). The method is based on the information bottleneck concept (Tishby et al., 1999). The intuition is that the latent representation $\mathbf{z}^*$ can be jointly optimized for both reconstruction fidelity and interpretability if the objective is to make $\mathbf{z}^*$ maximally expressive about the side information $\mathbf{s}$ while being maximally compressive about the data $\mathbf{x}$. We prove that being maximally compressive about the input for the sake of interpretability is analogous to further regularizing the data fitting procedure.

In contrast with the approach proposed in Section 2, where the latter concentrates on not losing reconstruction fidelity, the algorithm proposed in this section trades off interpretability and data reconstruction. On the one hand, there is a risk that this might lead to some loss in terms of reconstruction fidelity. On the other hand, $\mathbf{z}^*$ has more freedom to learn an interpretable representation from the whole data $\mathbf{x}$ without being confined by optimizing interpretability only through $\mathbf{z}$.

With a parameter $\beta$ regulating the tradeoff between compression of $\mathbf{x}$ and expressiveness about $\mathbf{s}$, the notion of information bottleneck among $\mathbf{x}$, $\mathbf{s}$ and $\mathbf{z}^*$ can be defined as follows:

$$\mathbf{IB}(\mathbf{z}^*, \mathbf{x}, \mathbf{s}) = \mathbf{I}(\mathbf{z}^*, \mathbf{s}) - \beta \mathbf{I}(\mathbf{z}^*, \mathbf{x}). \tag{11}$$

We therefore consider maximizing (11) to be a proxy for having an interpretable $\mathbf{z}^*$. First, assume that $\mathbf{x}$ is generated by $\mathbf{z}^*$ and that $\mathbf{z}^*$ is dependent on $\mathbf{s}$, i.e. $\mathbf{p}(\mathbf{x}, \mathbf{s}, \mathbf{z}^*) = \mathbf{p}(\mathbf{s})\mathbf{p}(\mathbf{z}^*|\mathbf{s})\mathbf{p}(\mathbf{x}|\mathbf{z}^*)$. Recall that $\mathbf{z}^*$ should as well fit the data and for that we use a VAE consisting of a recognition model and a generative model. Let's refer to the parameters of the generative and recognition models as $\boldsymbol{\omega}$ and $\boldsymbol{\alpha}$, respectively. Also, note that an interpretable model will optimize the dependence between the latent space $\mathbf{z}^*$ and the side information $\mathbf{s}$ to express a simple, i.e. linear, relationship. In addition to the interpretability objective in (11) and for the sake of data fitting, the latent space $\mathbf{z}^*$ is also constrained by the variational objective described as follows:

$$\begin{aligned}
\log \mathbf{p}_{\boldsymbol{\omega}}(\mathbf{x}) &= \log \int_{\mathbf{z}^*} \mathbf{p}_{\boldsymbol{\omega}}(\mathbf{x}, \mathbf{z}^*)\, d\mathbf{z}^* \\
&\geq \mathbb{E}_{\mathbf{q}_{\boldsymbol{\alpha}}(\mathbf{z}^*|\mathbf{x})}[\log \mathbf{p}_{\boldsymbol{\omega}}(\mathbf{x}, \mathbf{z}^*) - \log \mathbf{q}_{\boldsymbol{\alpha}}(\mathbf{z}^*|\mathbf{x})] \qquad (12) \\
&= \mathbb{E}_{\mathbf{q}_{\boldsymbol{\alpha}}(\mathbf{z}^*|\mathbf{x})}[\log \mathbf{p}_{\boldsymbol{\omega}}(\mathbf{x}|\mathbf{z}^*)] - \mathbb{KL}(\mathbf{q}_{\boldsymbol{\alpha}}(\mathbf{z}^*|\mathbf{x}) \| \mathbf{p}_{\boldsymbol{\omega}}(\mathbf{z}^*)).
\end{aligned}$$

From (11) and (12) through a proof provided (along with other details) in Section 11 of the Appendix, the overall objective of the JLVM model can be lower bounded by:

$$\begin{aligned}
&max_{\boldsymbol{\omega}, \boldsymbol{\alpha}} \frac{1}{\mathbf{N}} \sum_{1}^{\mathbf{N}} E_{\mathbf{q}_{\boldsymbol{\alpha}}(\mathbf{z}^*|\mathbf{x})}[\log \mathbf{p}_{\boldsymbol{\omega}}(\mathbf{x}|\mathbf{z}^*)] \\
&- \frac{1}{\mathbf{N}} \sum_{1}^{\mathbf{N}} [\mathbb{KL}(\mathbf{q}_{\boldsymbol{\alpha}}(\mathbf{z}^*|\mathbf{x}) \| \mathbf{p}_{\boldsymbol{\omega}}(\mathbf{z}^*)) + \beta \mathbb{KL}[\mathbf{p}_{\boldsymbol{\omega}}(\mathbf{z}^*|\mathbf{x}) \| \mathbf{q}_{\boldsymbol{\alpha}}(\mathbf{z}^*)]] \\
&+ \int \mathbf{p}(\mathbf{s}) \int \mathbf{p}_{\boldsymbol{\omega}}(\mathbf{z}^*|\mathbf{s}) \log \mathbf{q}_{\boldsymbol{\alpha}}(\mathbf{s}|\mathbf{z}^*)\; d\mathbf{s}\; d\mathbf{z}^* \qquad (13)
\end{aligned}$$

The second KL-divergence originates from the compression of $\mathbf{x}$ (second term in (11)) and adds up further regularization to the data fitting process (described by the first line of (13) and first KL-divergence). This comes in line with the established intuition supported by works showing that the two objectives (compression and regularization) are almost perfectly aligned. This includes early works like Nowlan & Hinton (1992); Hinton & Camp (1993) as well as recent works like Ullrich et al. (2017).

# 6. Experiments

We qualitatively and quantitatively evaluate the proposed frameworks on three datasets: MNIST, SVHN and Chairs. Details of the datasets and experiments are provided in Sections 10 and 12 of the Appendix, respectively. We begin with a qualitative assessment.

## 6.1. Qualitative Evaluation

### 6.1.1. Latent variable manipulation

In this section, and for the sake of simplicity and clarity, we perform experiments using a single dimension per side information attribute. However, it is important to note that the method allows for more than one latent dimension to map to the same side information. In Figures 2 and 3, we display the results of performing both ILVM and JLVM on MNIST. We have got two side information attributes, one representing the digit identity and another representing thickness. In every row of Figures 2 and 3, the leftmost image represents an example image from MNIST belonging to a certain thickness level and with a certain identity (digit label). The latent dimension of $\mathbf{z}^*$ maximally informative about thickness is kept fixed whereas the latent dimension maximally informative about the digit identity is varied to produce the rest of the images throughout the same row. Images resulting from the latent dimension representing the digit identity are originally unordered but we order them for visualization purposes. For each row, the same process is repeated with an original MNIST image of a different thickness level. As can be seen, the generated images in each row cover all digit identities. The change in the digit identities keeps the thickness level per row almost the same (a sign of a good degree of disentanglement) with ILVM. The same happens with JLVM but the thickness level is less clearly kept. For example, the digits '2' and '3' in the third row of Figure 2 are more expressive of their thickness level than the corresponding images in Figure 3.

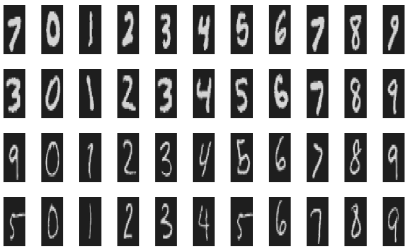


*Figure 2.* Results of ILVM on the MNIST dataset. The leftmost image represents an example from MNIST with certain thickness and with a certain digit label. To generate the remaining images of each row, the latent dimension maximally informative about the digit label is traversed while the one maximally informative about thickness is fixed. Row Images are reordered for visualization purposes. The latent space manages to cover and generate all the labels with similar thickness throughout each row.

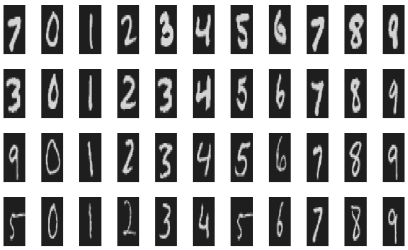
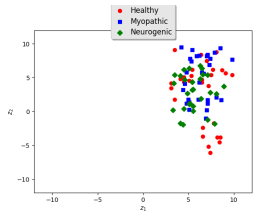
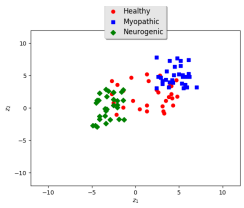


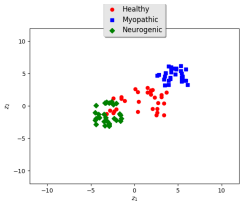
*Figure 3.* Results of JLVM on the MNIST dataset. Settings are equivalent to Figure 2. Results are also rather similar to those obtained by ILVM in Figure 2. However, ILVM is better at keeping the thickness degree almost fixed per row. For example, the digits '2' and '3' in the third row of Figure 2 are more representative of their thickness degree than the corresponding digit images here.




(a) A 2D latent space of a variational autoencoder (VAE).




(b) A 2D latent space of the ILVM model.




(c) A 2D latent space of the JLVM model.

*Figure 4.* 2D latent spaces inferred via a standard VAE framework, ILVM and JLVM. The latent spaces from ILVM and JLVM learn to better disentangle the points of different classes.

Similar experiments are performed on the Street View House Numbers (SVHN) data. For each row, an image from SVHN is chosen first, and then its latent space is traversed to produce the other images of the row. Such latent space is estimated by the two proposed algorithms, ILVM and JLVM, compared to two state-of-the-art algorithms, InfoGAN and $\beta$-VAE. In every row, one latent dimension is varied at a time while keeping all the other dimensions fixed. In Figure 5, the varied dimension denotes the lighting conditions of the image, whereas the varied dimension in Figure 1 in the Appendix denotes the saturation level. The spectra of lighting and saturation values generated by JLVM are considerably higher and clearer than those generated by InfoGAN and $\beta$-VAE, and slightly higher than those generated by ILVM. This demonstrates the ability of JLVM to adapt to challenging data conditions since SVHN is noisier than MNIST, does not have many variations of the same object and contains numerous images of lower resolution levels.

We also apply the proposed ILVM and JLVM models to the images of the 3D Chairs dataset (Aubry et al., 2014). Similar to the experiments on SVHN, an image is chosen from images of the 3D chairs, and then the corresponding latent space estimated via ILVM, JLVM, InfoGAN and $\beta$-VAE is traversed. The dimensions to be traversed one at a time (while keeping all the others fixed) in this case refer to azimuth and width level in Figures 6 here and 2 in the Appendix, respectively. Similar to the results on SVHN, the JLVM model manages to better understand and vary the azimuth and width features of the 3D Chairs. The ILVM model provides the second most interpretable latent representation with respect to the azimuth and width attributes.

#### 6.1.2. A 2D Latent Visualization

We perform an illustrative experiment based on data points from a small dataset with three labels (classes) (Adel et al., 2013). The three labels refer to diagnosis outcomes, which are: healthy, myopathic or neurogenic. The labels are given as the side information. In Figure 4, we plot and compare how a 2D latent space looks like when inferred via a standard VAE, ILVM and JLVM. The latent spaces resulting from ILVM and JLVM are more interpretable since they learn to better disentangle the points of different classes. The Myopathic and Healthy classes are even more disentangled with JLVM in Figure 4(c).

#### 6.1.3. Interpretable Lens on a Hidden Layer of a Neural Network

For the ILVM framework, an experiment is performed where the input $\mathbf{z}$ is a hidden layer of a neural network, instead of being a latent space inferred from a VAE. The dataset used is a variation of MNIST referred to as MNIST-rot (Larochelle et al., 2007) containing various images of rotated MNIST digits. We inspect the impact of learning the latent space $\mathbf{z}^*$ in Figure 7. The side information in this case is the orientation. We compare how a certain hidden layer looks like before and after applying ILVM. The first row shows three different images where the second and third images have a similar orientation, different from the first. In the second row, a certain hidden layer of the neural network is plotted after training. There are hardly any visual similarities reflecting the orientation analogy between the second and third plots in the second row. However, in the third row after performing ILVM, similarities between the second and third plots can be noticed, indicating that the interpretable orientation information is now taken into account. This preliminary result is a beginning of a

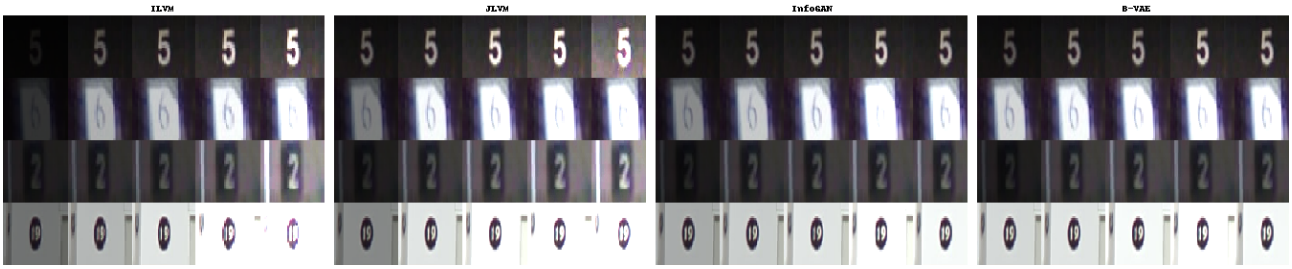



*Figure 5.* A comparison between the proposed frameworks ILVM and JLVM vs. InfoGAN and $\beta$-VAE applied to SVHN. Each row represents an experiment where the lighting condition of an input SVHN image is varied while the other latent dimensions are kept fixed. White margins separate the results of ILVM, JLVM, InfoGAN and $\beta$-VAE (from left to right). The spectra of lighting values generated by JLVM are higher than those generated by InfoGAN and $\beta$-VAE. Better viewed in color.

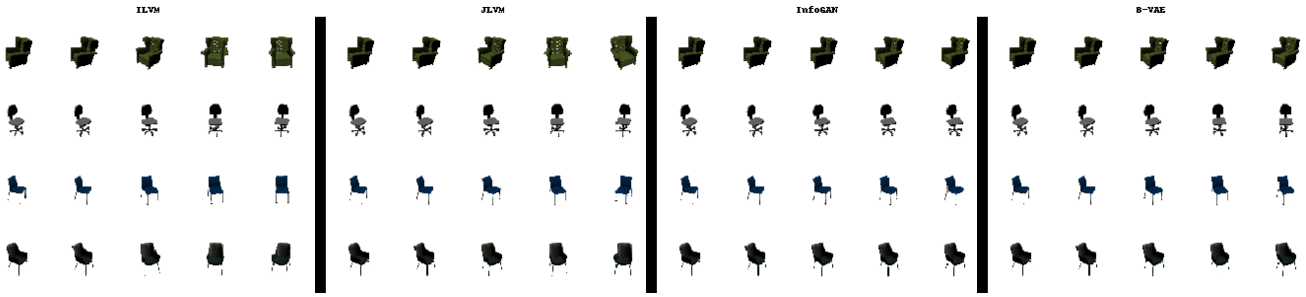



*Figure 6.* A comparison between ILVM, JLVM, InfoGAN and $\beta$-VAE on the 3D Chairs dataset. Each row represents an experiment where only the latent dimension maximally informative about the azimuth level of an SVHN image is varied. This is equivalent to rotating the chair in the original image. The JLVM model manages to disentangle and represent the chair's interpretable azimuth information better than the other models, as can be seen by the wide range of orientations covered by JLVM. The ILVM model comes second.

more comprehensive set of experiments applying ILVM to a broader range of deep models.

### 6.2. Quantitative Evaluation

#### 6.2.1. The Interpretability Metric

We compare ILVM and JLVM to the state-of-the-art algorithm referred to as InfoGAN (Chen et al., 2016). In order to apply the metric to the latter, the InfoGAN style variables are treated as the variables denoting the side information $\mathbf{s}$. The interpretability metric is not applicable to $\beta$-VAE (Higgins et al., 2017) since computing the joint likelihood of the latent space and the (dependent and independent) factors of variation is not feasible in their model. Results of the proposed metric on MNIST, SVHN and the 3D Chairs data are displayed in Table 1. JLVM outperforms the others on SVHN and Chairs. The conditions are favorable for JLVM when there are not enough variations of each object. When this is not the case, i.e. when the input latent space is representative enough, ILVM outperforms the competitors (on MNIST). A bold entry denotes that an algorithm is significantly better than its competitors. Significant results are identified using a paired t-test with p = 0.05. Side information used with a few of the MNIST images are the digit labels and thickness. Side information for SVHN is the lighting condition and saturation degree, and it comes in the form of azimuth and width for the 3D Chairs data.

*Table 1.* Results of the interpretability metric for 3 datasets, MNIST, SVHN and the 3D Chairs datasets. JLVM outperforms the others on SVHN and the Chairs data. In cases when there are not enough variations of each object, e.g. SVHN, JLVM performs better. When the input latent space is representative, the simple and fully generative ILVM outperforms the competitors (MNIST).

| | MNIST | SVHN | Chairs |
|---|---|---|---|
| **ILVM** | **95.2** $\pm$ **1.3** % | 85.7 $\pm$ 0.9 % | 87.4 $\pm$ 1.0 % |
| **JLVM** | 89.8 $\pm$ 0.9 % | **90.1** $\pm$ **1.1** % | **89.8** $\pm$ **1.5** % |
| InfoGAN | 83.3 $\pm$ 1.8 % | 83.9 $\pm$ 1.3 % | 85.2 $\pm$ 1.4 % |

#### 6.2.2. Active Learning

Based on the active learning strategy introduced in Section 3, the following experiments are performed on MNIST and SVHN using ILVM. We compare the introduced strategy to two other strategies. The first is random acquisition where each data point (whose side information is to be obtained) is drawn from a uniform distribution (Gal et al., 2017). The second strategy is to choose the data points which maximize the predictive entropy (Max En-

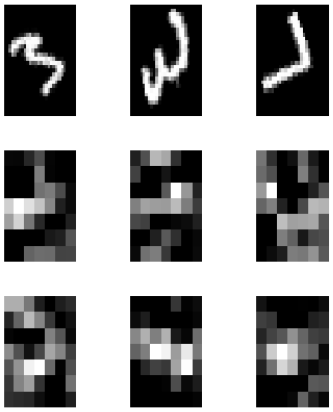


*Figure 7.* Application of ILVM as an interpretable lens on a hidden layer of a neural network. A comparison of how a certain hidden layer looks like before and after performing ILVM. The first row shows three different images where the second and third have a similar orientation, different from the first. In the second row, a certain hidden layer of the neural network is plotted after training. There are hardly any visual similarities reflecting the orientation analogy between the second and third plots. In the third row, plots of the corresponding $\mathbf{z}^*$ after performing ILVM show similarities between the second and third plots, indicating that the interpretable orientation information has been taken into account.

tropy) (Shannon, 1948):

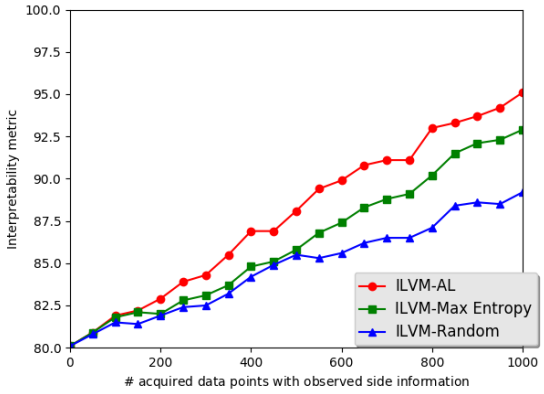
$$\hat{\mathbf{j}} = \text{argmax}_{\mathbf{j}} \mathbf{H}(\mathbf{s_j}) = -\int \mathbf{p}(\mathbf{s_j}) \log \mathbf{p}(\mathbf{s_j})\, d\,\mathbf{s} \qquad (14)$$

$$= -\iint \mathbf{p}_\psi(\mathbf{s_j}|\mathbf{z}^*)\mathbf{p}(\mathbf{z}^*) d\mathbf{z}^* \;\; \log(\int \mathbf{p}_\psi(\mathbf{s_j}|\mathbf{z}^*)\mathbf{p}(\mathbf{z}^*) d\mathbf{z}^*) d\mathbf{s}$$

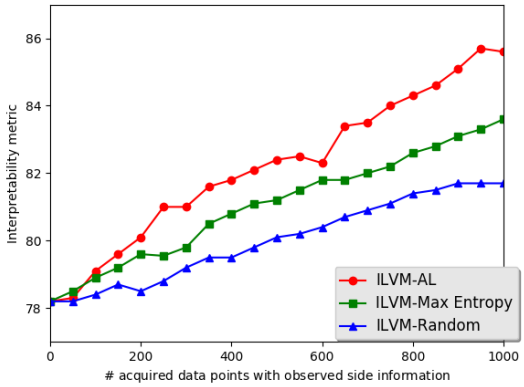
where $\mathbf{p}_\psi(\mathbf{s_j}|\mathbf{z}^*)$ is computed from the generative model and $\mathbf{p}(\mathbf{z}^*)$ is the prior. Results are displayed in Figure 8. All the values shown in the figure are tested on the same test set. The x-axis represents the training size. Along each curve, every new point represents the interpretability metric resulting from adding the side information of extra 50 training data points to the training data.

## 7. Conclusion

We have introduced two interpretability frameworks and a corresponding evaluation metric. The first framework opens a new direction since it can be applied to models of different types, including deep generative and discriminative models. In addition, it does not conflict with their original objective, be it reconstruction fidelity or classification accuracy. We have also introduced a strategy to bring human subjectivity into interpretability to yield interactive 'human-in-the-loop' interpretability. We believe this approach will be a fruitful line for future work.




(a) MNIST




(b) SVHN

*Figure 8.* Results of applying the proposed active learning strategy (ILVM-AL) vs. Random acquisition and Max Entropy.

The second proposed interpretability framework is consistent with the current trend of joint optimization for interpretability and reconstruction fidelity, but it sheds light on a newly derived relationship between compression and regularization. The introduced frameworks achieve state-of-the-art results on three datasets.

The prospect that other existing deep models, such as those related to reinforcement learning, do not have to be retrained to achieve interpretability suggests much interesting potential for future work.

There is also potential for imminent future work to further bridge the gap between our introduced joint optimization for reconstruction and interpretability, and posterior constrained Bayesian inference. Some prior works on the latter focus on imposing explicit hard constraints or expectation constraints like RegBayes (Zhu et al., 2014). Our contribution is to focus on the link between compression and regularization, and on relating this to interpretability.

## Acknowledgements

All the authors acknowledge support from the Leverhulme Trust, DeepMind and the AI Fund via the CFI. AW acknowledges support from the David MacKay Newton research fellowship at Darwin College and The Alan Turing Institute under EPSRC grant EP/N510129/1 & TU/B/000074.

## References

Adel, T., Urner, R., Smith, B., Stashuk, D., and Lizotte, D. Generative multiple-instance learning models for quantitative electromyography. *Uncertainty in Aritifical Intelligence (UAI)*, 2013.

Aubry, M., Maturana, D., Efros, A., Russell, B., and Sivic, J. Seeing 3D chairs: Exemplar part-based 2D-3D alignment using a large dataset of CAD models. pp. 3762–3769, 2014.

Bengio, Y. Learning deep architectures for AI. *Foundations and trends in Machine Learning*, 2:1–127, 2009.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (8), 2013.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems (NIPS)*, pp. 2172–2180, 2016.

Desjardins, G., Courville, A., and Bengio, Y. Disentangling factors of variation via generative entangling. *arXiv:1210.5474*, 2012.

Gal, Y., Islam, R., and Ghahramani, Z. Deep Bayesian active learning with image data. *International Conference on Machine Learning (ICML)*, 2017.

Goodfellow, I. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems (NIPS)*, pp. 2672–2680, 2014.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. Beta-VAE: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations (ICLR)*, 2017.

Hinton, G. and Camp, D. Van. Keeping the neural networks simple by minimizing the description length of the weights. *Conference on Learning Theory (COLT)*, pp. 5–13, 1993.

Houlsby, N., Huszar, F., Ghahramani, Z., and Lengyel, M. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Hsu, W., Zhang, Y., and Glass, J. Unsupervised learning of disentangled and interpretable representations from sequential data. *Advances in neural information processing systems (NIPS)*, pp. 1876–1887, 2017.

Kingma, D. and Welling, M. Auto-encoding variational Bayes. *International Conference on Learning Representations (ICLR)*, 2014.

Kingma, D., Rezende, D., Mohamed, S., and Welling, M. Semi-supervised learning with deep generative models. *Advances in neural information processing systems (NIPS)*, 28:3581–3589, 2014.

Kingma, D., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems (NIPS)*, 30, 2016.

Kulkarni, T., Whitney, W., Kohli, P., and Tenenbaum, J. Deep convolutional inverse graphics network. *Advances in neural information processing systems (NIPS)*, pp. 2539–2547, 2015.

Larochelle, H., Erhan, D., Courville, A., Bergstra, J., and Bengio, Y. An empirical evaluation of deep architectures on problems with many factors of variation. *International Conference on Machine Learning (ICML)*, 2007.

Nowlan, S. and Hinton, G. Simplifying neural networks by soft weight-sharing. *Neural computation*, 4:473–493, 1992.

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. *International Conference on Machine Learning (ICML)*, 32:1530–1538, 2015.

Rezende, D., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *International Conference on Machine Learning (ICML)*, 31, 2014.

Shannon, C. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

Siddharth, N., Paige, B., van den Meent, J., Demaison, A., Goodman, N., Kohli, P., Wood, F., and Torr, P. Learning disentangled representations with semi-supervised deep generative models. *Advances in neural information processing systems (NIPS)*, 2017.

Tabak, E. and Turner, C. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.

Tabak, E. and Vanden-Eijnden, E. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217–233, 2010.

Tishby, N., Pereira, F., and Biale, W. The information bottleneck method. *Allerton Conf. on Communication, Control, and Computing*, 37:368–377, 1999.

Ullrich, K., Meeds, E., and Welling, M. Soft weight-sharing for neural network compression. *International Conference on Learning Representations (ICLR)*, 2017.

Zhu, J., Chen, N., and Xing, E. Bayesian inference with posterior regularization and applications to infinite latent SVMs. *Journal of Machine Learning Research (JMLR)*, 15:1799–1847, 2014.