Variational Inference of Disentangled Latent Concepts from Unlabeled Observations

Anonymous authors
Paper under double-blind review

Abstract

Disentangled representations, where the higher level data generative factors are reflected in disjoint latent dimensions, offer several benefits such as ease of deriving invariant representations, transferability to other tasks, interpretability, etc. We consider the problem of unsupervised learning of disentangled representations from large pool of unlabeled observations, and propose a variational inference based approach to infer disentangled latent factors. We introduce a regularizer on the expectation of the approximate posterior over observed data that encourages the disentanglement. We evaluate the proposed approach using several quantitative metrics and empirically observe significant gains over existing methods in terms of both disentanglement and data likelihood (reconstruction quality).

1 Introduction

Feature representations of the observed raw data play a crucial role in the success of machine learning algorithms. Effective representations should be able to capture the underlying (abstract or high-level) latent generative factors that are relevant for the end task while ignoring the inconsequential or nuisance factors. Disentangled feature representations have the property that the generative factors are revealed in disjoint subsets of the feature dimensions, such that a change in a single generative factor causes a highly sparse change in the representation. Disentangled representations offer several advantages – (i) Invariance: it is easier to derive representations that are invariant to nuisance factors by simply marginalizing over the corresponding dimensions, (ii) Transferability: they are arguably more suitable for transfer learning as most of the key underlying generative factors appear segregated along feature dimensions, (iii) Interpretability: a human expert may be able to assign meanings to the dimensions, (iv) Conditioning and intervention: they allow for interpretable conditioning and/or intervention over a subset of the latents and observe the effects on other nodes in the graph. Indeed, the importance of learning disentangled representations has been argued in several recent works (Bengio et al., 2013; Lake et al., 2016; Ridgeway, 2016).

Recognizing the significance of disentangled representations, several attempts have been made in this direction in the past (Ridgeway, 2016). Much of the earlier work assumes some sort of supervision in terms of: (i) partial or full access to the generative factors per instance (Reed et al., 2014; Yang et al., 2015; Kulkarni et al., 2015; Karaletsos et al., 2015), (ii) knowledge about the nature of generative factors (e.g., translation, rotation, etc.) (Hinton et al., 2011; Cohen & Welling, 2014), (iii) knowledge about the changes in the generative factors across observations (e.g., sparse changes in consecutive frames of a video) (Goroshin et al., 2015; Whitney et al., 2016; Fraccaro et al., 2017; Denton & Birodkar, 2017; Hsu et al., 2017), (iv) knowledge of a complementary signal to infer representations that are conditionally independent of it (Cheung et al., 2014; Mathieu et al., 2016; Siddharth et al., 2017). However, in most real scenarios, we only have access to raw observations without any supervision about the generative factors. It is a challenging problem and many of the earlier attempts have not been able to scale well for realistic settings (Schmidhuber, 1992; Desjardins et al., 2012; Cohen & Welling, 2015) (see also, Higgins et al. (2017)).

Recently, Chen et al. (2016) proposed an approach to learn a generative model with disentangled factors based on Generative Adversarial Networks (GAN) (Goodfellow et al., 2014), however

---

1The representation itself can still be entangled in rest of the generative factors.
implicit generative models like GANs lack an effective inference mechanism, which hinders its applicability to the problem of learning disentangled representations. More recently, Higgins et al. (2017) proposed an approach based on Variational AutoEncoder (VAE) Kingma & Welling (2013) for inferring disentangled factors. The inferred latents using their method (termed as $\beta$-VAE) are empirically shown to have better disentangling properties, however the method deviates from the basic principles of variational inference, creating increased tension between observed data likelihood and disentanglement. This in turn leads to poor quality of generated samples as observed in (Higgins et al. 2017).

In this work, we propose a principled approach for inference of disentangled latent factors based on the popular and scalable framework of amortized variational inference (Kingma & Welling 2013; Stuhlmüller et al., 2013; Gershman & Goodman, 2014; Rezende et al., 2014) powered by stochastic optimization (Hoffman et al., 2013; Kingma & Welling, 2013; Rezende et al., 2014). Disentanglement is encouraged by introducing a regularizer over the induced inferred prior. Unlike $\beta$-VAE (Higgins et al. 2017), our approach does not introduce any extra conflict between disentanglement of the latents and the observed data likelihood, which is reflected in the overall quality of the generated samples that matches the VAE and is much better than $\beta$-VAE. This does not come at the cost of higher entanglement and our approach also outperforms $\beta$-VAE in disentangling the latents as measured by various quantitative metrics.

2 FORMULATION

We start with a generative model of the observed data that first samples a latent variable $z \sim p(z)$, and an observation is generated by sampling from $p_\theta(x|z)$. The joint density of latents and observations is denoted as $p_\theta(x,z) = p(z)p_\theta(x|z)$. The problem of inference is to compute the posterior of the latents conditioned on the observations, i.e., $p_\theta(z|x) = \frac{p_\theta(x,z)}{\int p_\theta(x,z)dz}$. We assume that we are given a finite set of samples (observations) from the true data distribution $p(x)$. In most practical scenarios involving high dimensional and complex data, this computation is intractable and calls for approximate inference. Variational inference takes an optimization based approach to this, positing a family $D$ of approximate densities over the latents and reducing the approximate inference problem to finding a member density that minimizes the Kullback-Leibler divergence to the true posterior, i.e., $q_\phi^* = \min_{q \in D} \text{KL}(q(z) || p_\theta(z|x))$ (Blei et al., 2017). The idea of amortized inference (Kingma & Welling, 2013; Stuhlmüller et al., 2013; Gershman & Goodman, 2014; Rezende et al., 2014) is to explicitly share information across inferences made for each observation. One successful way of achieving this for variational inference is to have a so-called recognition model, parameterized by $\phi$, that encodes an inverse map from the observations to the approximate posteriors (also referred as variational autoencoder or VAE) (Kingma & Welling, 2013; Rezende et al., 2014). The recognition model parameters are learned by optimizing the problem $\min_\phi \mathbb{E}_x \text{KL}(q_\phi(z|x)||p_\theta(z|x))$, where the outer expectation is over the true data distribution $p(x)$ which we have samples from. This can be shown as equivalent to maximizing what is termed as evidence lower bound (ELBO):

$$
\arg \min_{\theta,\phi} \mathbb{E}_x \text{KL}(q_\phi(z|x)||p_\theta(z|x)) = \arg \max_{\theta,\phi} \mathbb{E}_x \left[ \mathbb{E}_{z \sim q_\phi(z|x)} \left[ \log p_\theta(x|z) \right] - \text{KL}(q_\phi(z|x)||p(z)) \right]
$$

(1)

The ELBO (the objective at the right side of Eq. 1) lower bounds the log-likelihood of observed data, and the gap vanishes at the global optimum. Often, the density forms of $p(z)$ and $q_\phi(z|x)$ are chosen such that their KL-divergence can be written analytically in a closed-form expression (e.g., $p(z)$ is $N(0, I)$ and $q_\phi(z|x)$ is $q_\phi(z|x) = N(\mu_\phi(x), \Sigma_\phi(x))$ (Kingma & Welling, 2013). In such cases, the ELBO can be efficiently optimized (to a stationary point) using stochastic first order methods where both expectations are estimated using mini-batches. Further, in cases when $q_\phi(\cdot)$ can be written as a continuous transformation of a fixed base distribution (e.g., the standard normal distribution), a low variance estimate of the gradient over $\phi$ can be obtained by coordinate transformation (also referred as reparameterization) (Fu, 2006; Kingma & Welling, 2013; Rezende et al., 2014).
2.1 GENERATIVE STORY: DISENTANGLED PRIOR

Most VAE based generative models for real datasets (e.g., text, images, etc.) already work with a relatively simple and disentangled prior $p(z)$ having no interaction among the latent dimensions (e.g., the standard Gaussian $N(0, I)$) (Bowman et al., 2015; Miao et al., 2016; Hou et al., 2017; Zhao et al., 2017). The complexity of the observed data is absorbed in the conditional distribution $p_{θ}(x|z)$ which encodes the interactions among the latents. Hence, as far as the generative modeling is concerned, disentangled prior sets us in the right direction.

2.2 INFERRING DISENTANGLED LATENTS

Although the generative model starts with a disentangled prior, our main objective is to infer disentangled latents which are potentially conducive for various goals mentioned in Sec. I (e.g., invariance, transferability, interpretability). To this end, we consider the density over the inferred latents induced by the approximate posterior inference mechanism,

$$q_{θ}(z) = \int q_{θ}(z|x)p(x)dx,$$

which we will subsequently refer to as the \textit{inferred prior} or \textit{expected variational posterior} ($p(x)$ is the true data distribution that we have only samples from). For inferring disentangled factors, this should be factorizable along the dimensions, i.e., $q_{θ}(z) = \prod_{i} q_{θi}(z_{i})$, or equivalently $q_{θij}(z_{i}|z_{j}) = q_{i}(z_{i}), \forall i, j$. This can be achieved by minimizing a suitable distance between the inferred prior $q_{θ}(z)$ and the disentangled generative prior $p(z)$. We can also define \textit{expected posterior} as $p_{θ}(z) = \int p_{θ}(z|x)p(x)dx$. If we take KL-divergence as our choice of distance, by relying on its pairwise convexity, i.e., $KL(λp_{1} + (1 − λ)p_{2}∥λq_{1} + (1 − λ)q_{2}) ≤ KL(p_{1}∥q_{1}) + (1 − λ)KL(p_{2}∥q_{2})$ (Van Erven & Harremos, 2014), we can show that the distance between $q_{θ}(z)$ and $p_{θ}(z)$ is bounded by the objective of the variational inference:

$$KL(q_{θ}(z)||p_{θ}(z)) = KL(E_{x~p(x)}q_{θ}(z|x)||E_{x~p(x)}p_{θ}(z|x)) ≤ E_{x~p(x)}KL(q_{θ}(z|x)||p_{θ}(z|x)).$$

In general, the prior $p(z)$ and expected posterior $p_{θ}(z)$ will be different, although they may be close (they will be same when $p_{θ}(x) = \int p_{θ}(x|z)p(z)dz$ is equal to $p(x)$). Hence, variational posterior inference of latent variables with disentangled prior naturally encourages inferring factors that are close to being disentangled. We think this is the reason that the original VAE (Eq. (1)) has also been observed to exhibit some disentangling behavior on simple datasets such as MNIST (Kingma & Welling, 2013). However, this behavior does not carry over to more complex datasets (Aubry et al., 2014; Liu et al., 2015; Higgins et al., 2017), unless extra supervision on the generative factors is provided (Kulkarni et al., 2015; Karatzos et al., 2015). This can be due to: (i) $p(x)$ and $p_{θ}(x)$ being far apart which in turn causes $p(z)$ and $p_{θ}(z)$ being far apart, and (ii) the non-convexity of the ELBO objective which prevents us from achieving the global minimum of $E_{x}KL(q_{θ}(z|x)||p_{θ}(z|x))$ (which is 0 and implies $KL(q_{θ}(z)||p_{θ}(z)) = 0$). In other words, maximizing the ELBO (Eq. (1)) might also result in reducing the value of $KL(q_{θ}(z)||p(z))$, however, due to the aforementioned reasons, the gap between $KL(q_{θ}(z)||p(z))$ and $E_{x}KL(q_{θ}(z|x)||p_{θ}(z|x))$ could be large at the stationary point of convergence. Hence, minimizing $KL(q_{θ}(z)||p(z))$ or any other suitable distance $D(q_{θ}(z), p(z))$ explicitily will give us better control on the disentanglement. This motivates us to add $D(q_{θ}(z)||p(z))$ as part of the objective to encourage disentanglement during inference, i.e.,

$$\max_{θ, φ} E_{x~p(x)}[\log p_{θ}(x|z) − KL(q_{θ}(z|x)||p(z))] − λ D(q_{θ}(z)||p(z)),$$

where $λ$ controls its contribution to the overall objective. We refer to this as DIP-VAE (for Disentangled Inferred Prior) subsequently.

Optimizing (4) directly is not tractable if $D(·, ·)$ is taken to be the KL-divergence $KL(q_{θ}(z)||p(z))$, which does not have a closed-form expression. One possibility is use the variational formulation of the KL-divergence (Nguyen et al., 2010; Nowozin et al., 2016) that needs only samples from $q_{θ}(z)$ and $p(z)$ to estimate a lower bound to $KL(q_{θ}(z)||p(z))$. However, this would involve optimizing for a third set of parameters $ψ$ for the KL-divergence estimator, and would also change the optimization to a saddle-point (min-max) problem which has its own optimization challenges (e.g., gradient vanishing as encountered in training generative adversarial networks with KL or Jensen-Shannon divergence).
We adopt a simpler yet effective alternative of matching the moments of the two distributions. In particular, we match the covariance of the two distributions which will amount to decorrelating the dimensions of $z \sim q_\phi(z)$ if $p(z)$ is $N(0, I)$. Let us denote $\text{Cov}_{q(z)}[z] := \mathbb{E}_{q(z)}[(z - \mathbb{E}_{q(z)}[z])(z - \mathbb{E}_{q(z)}[z])^\top]$. By the law of total covariance, the covariance of $z \sim q_\phi(z)$ is given by

$$\text{Cov}_{q_\phi(z)}[z] = \mathbb{E}_{p(x)}\text{Cov}_{q_\phi(x|z)}[z] + \text{Cov}_{p(x)}\left(\mathbb{E}_{q_\phi(x|z)}[z]\right), \tag{5}$$

where $\mathbb{E}_{q_\phi(x|z)}[z]$ and $\text{Cov}_{q_\phi(x|z)}[z]$ are random variables that are functions of the random variable $x$ ($z$ is marginalized over). Most existing work on the VAE models uses $q_\phi(z|x)$ having the form $N(\mu_\phi(x), \Sigma_\phi(x))$, where $\mu_\phi(x)$ and $\Sigma_\phi(x)$ are the outputs of a deep neural net parameterized by $\phi$. In this case Eq. (5) reduces to $\text{Cov}_{q_\phi(x|z)}[z] = \mathbb{E}_{p(x)}[\Sigma_\phi(x)] + \text{Cov}_{p(x)}[\mu_\phi(x)]$, which we want to be close to an identity matrix. For simplicity, we choose entry-wise squared $l_2$-norm as the measure of proximity. Further, $\Sigma_\phi(x)$ is commonly taken to be a diagonal matrix which means that cross-correlations (off-diagonals) between the latents are due to only $\text{Cov}_{p(x)}[\mu_\phi(x)]$. This suggests two possible options for the disentangling regularizer: (i) regularizing only $\text{Cov}_{p(x)}[\mu_\phi(x)]$ which we refer as DIP-VAE-I, (ii) regularizing $\text{Cov}_{q_\phi(x)}[z]$ which we refer as DIP-VAE-II. Penalizing just the off-diagonals in both cases will lead to lowering the diagonal entries of $\text{Cov}_{p(x)}[\mu_\phi(x)]$ as the $i$th and $j$th diagonal is really a derived attribute obtained by multiplying the square-roots of $i$th and $j$th diagonals (for each example $x \sim p(x)$, followed by averaging over all examples). This can be compensated in DIP-VAE-I by a regularizer on the diagonal entries of $\text{Cov}_{p(x)}[\mu_\phi(x)]$ which pulls these towards 1. We opt for two separate hyperparameters controlling the relative importance of the loss on the diagonal and off-diagonal entries as follows:

$$\max_{\theta, \phi} \text{ELBO}(\theta, \phi) - \lambda_{od} \sum_{i \neq j} \left[ \text{Cov}_{p(x)}[\mu_\phi(x)] \right]_{ij}^2 - \lambda_d \sum_i \left( \left[ \text{Cov}_{p(x)}[\mu_\phi(x)] \right]_{ii} - 1 \right)^2. \tag{6}$$

The regularization terms involving $\text{Cov}_{p(x)}[\mu_\phi(x)]$ in the above objective (6) can be efficiently optimized using SGD. We maintain a running estimate of $\text{Cov}_{p(x)}[\mu_\phi(x)]$ which is updated with every minibatch of $x \sim p(x)$. The gradient for the current minibatch can be computed by treating the previous estimate of $\text{Cov}_{p(x)}[\mu_\phi(x)]$ as constant.

For DIP-VAE-II, we have the following optimization problem:

$$\max_{\theta, \phi} \text{ELBO}(\theta, \phi) - \lambda_{od} \sum_{i \neq j} \left[ \text{Cov}_{q_\phi(x|z)}[z] \right]_{ij}^2 - \lambda_d \sum_i \left( \left[ \text{Cov}_{q_\phi(x|z)}[z] \right]_{ii} - 1 \right)^2. \tag{7}$$

As discussed earlier, the term $\mathbb{E}_{p(x)}\text{Cov}_{q_\phi(x|z)}[z]$ contributes only to the diagonals of $\text{Cov}_{q_\phi(x|z)}[z]$. Penalizing the off-diagonals of $\text{Cov}_{p(x)}[\mu_\phi(x)]$ in the objective (7) will contribute to reduction in the magnitude of its diagonals as discussed earlier. As the regularizer on the diagonals is not directly on $\text{Cov}_{p(x)}[\mu_\phi(x)]$, unlike DIP-VAE-I, it will be not be able to keep $\left[ \text{Cov}_{p(x)}[\mu_\phi(x)] \right]_{ii}$ close to 1: the reduction in $\left[ \text{Cov}_{p(x)}[\mu_\phi(x)] \right]_{ii}$ will be accompanied by increase in $\left[ \mathbb{E}_{p(x)}[\Sigma_\phi(x)] \right]_{ii}$ such that their sum remains close to 1. In datasets where the number of generative factors is less than the latent dimension, DIP-VAE-II is more suitable than DIP-VAE-I as keeping all dimensions active might result in splitting of an attribute across multiple dimensions, hurting the goal of disentanglement.

### 2.3 Comparison with $\beta$-VAE

Recently proposed $\beta$-VAE (Higgins et al. 2017) proposes to modify the ELBO by upweighting the KL($q_\phi(z|x)$ || $p(z)$) term in order to encourage the inference of disentangled factors:

$$\max_{\theta, \phi} \mathbb{E}_x \left[ \log p_0(x|z) - \beta \text{KL}(q_\phi(z|x)||p(z)) \right], \tag{8}$$

Nonparametric distances like maximum mean discrepancy (MMD) with a characteristic kernel (Gretton et al. 2012) is also an option, however it has its own challenges when combined with stochastic optimization (Druge et al. 2015; Li et al. 2015).
We observe in our experiments that the Z-diff score (Higgins et al., 2017) is not correlated well with the qualitative disentanglement at the decoder's output as seen in the latent traversal plots, that is better aligned with the qualitative disentanglement Attribute Predictability (SAP) score. This average difference vector is assigned the label of $y$ on other latents. A high SAP score indicates that each generative factor is primarily captured in only one generative factor is not enough as it does not rule out the possibility of the factor being captured by other latents. A low capacity multiclass classifier is then trained on these vectors to predict the identities of the corresponding generative factors. Accuracy of this classifier on the difference vectors (i) observed in the latent traversals and also does not involve training any classifier. It is computed for test set is taken to be a measure of disentanglement. We evaluate the proposed method on this metric and refer to this as Z-diff score subsequently.

For high values of $\beta$, $\beta$-VAE would try to pull $\mu_i(x)$ towards zero and $\Sigma_i(x)$ towards the identity matrix (as the minimum of $x - \ln x$ for $x > 0$ is at $x = 1$), thus making the approximate posterior $q_\phi(z|x)$ insensitive to the observations. This is also reflected in the quality of the reconstructed samples which is worse than VAE ($\beta = 1$), particularly for high values of $\beta$. Our proposed method does not have such increased tension between the likelihood term and the disentanglement objective, and the sample quality with our method is on par with the VAE.

Finally, we note that both $\beta$-VAE and our proposed method encourage disentanglement of inferred factors by pulling $\text{Cov}_{q_\phi(z)}(z)$ in Eq. (3) towards the identity matrix: $\beta$-VAE attempts to do it by making $\text{Cov}_{q_\phi(z)}(z)$ close to $I$ and $\mathbb{E}_{q_\phi(z|x)}(z)$ close to $0$ individually for all observations $x$, while the proposed method directly works on $\text{Cov}_{q_\phi(z)}(z)$ (marginalizing over the observations $x$) which retains the sensitivity of $q_\phi(z|x)$ to the conditioned-upon observation.

### 3 Quantifying disentanglement: SAP Score

[Higgins et al. (2017)] propose a metric to evaluate the disentanglement performance of the inference mechanism, assuming that the ground truth generative factors are available. It works by first sampling a generative factor $y$, followed by sampling $L$ pairs of examples such that for each pair, the sampled generative factor takes the same value. Given the inferred $z_x := \mu_i(x)$ for each example $x$, they compute the absolute difference of these vectors for each pair, followed by averaging these difference vectors. This average difference vector is assigned the label of $y$. By sampling $n$ such minibatches of $L$ pairs, we get $n$ such averaged difference vectors for the factor $y$. This process is repeated for all generative factors. A low capacity multiclass classifier is then trained on these vectors to predict the identities of the corresponding generative factors. Accuracy of this classifier on the difference vectors for test set is taken to be a measure of disentanglement. We evaluate the proposed method on this metric and refer to this as Z-diff score subsequently.

We observe in our experiments that the Z-diff score ([Higgins et al. 2017]) is not correlated well with the qualitative disentanglement at the decoder’s output as seen in the latent traversal plots (obtained by varying only one latent while keeping the other latents fixed). It also depends on the multiclass classifier used to obtain the score. We propose a new metric, referred to as Separated Attribute Predictability (SAP) score, that is better aligned with the qualitative disentanglement observed in the latent traversals and also does not involve training any classifier. It is computed as follows: (i) We first construct a $d \times k$ score matrix $S$ (for $d$ latents and $k$ generative factors) whose $ij$'th entry is the linear regression or classification score (depending on the generative factor type) of predicting $j$'th factor using only $i$'th latent $[\mu_i(x)]_j$. For regression, we take this to be the $R^2$ score obtained with fitting a line (slope and intercept) that minimizes the linear regression error (for the test examples). The $R^2$ score is given by $rac{\text{Cov}(\mu_i(x), y_j)}{\sigma_{\mu_i(x)} \sigma_{y_j}}^2$ and ranges from 0 to 1, with a score of 1 indicating that a linear function of the $i$'th inferred latent explains all variability in the $j$'th generative factor. For classification, we fit one or more thresholds (real numbers) directly on $i$'th inferred latents for the test examples that minimize the balanced classification errors, and take $S_{ij}$ to be the balanced classification accuracy of the $j$'th generative factor. For inactive latent dimensions (having $\sigma_{\mu_i(x)} = [\text{Cov}_{\mu_i(x)}^2][\mu_i(x)]_{ii}$ close to 0), we take $S_{ij}$ to be 0. (ii) For each column of the score matrix $S$ which corresponds to a generative factor, we take the difference of top two entries (corresponding to top two most predictive latent dimensions), and then take the mean of these differences as the final SAP score. Considering just the top scoring latent dimension for each generative factor is not enough as it does not rule out the possibility of the factor being captured by other latents. A high SAP score indicates that each generative factor is primarily captured in only one
We evaluate our proposed method, DIP-V AE, on three datasets – (i) CelebA (Higgins et al., 2017) for CelebA, 3D Chairs, 2D Shapes datasets. To be consistent with the evaluations in (Higgins et al., 2017), we use the same (i) V AE (Kingma & Welling, 2013; Rezende et al., 2014), and (ii) the recently proposed x-position (32 values), scale (6 values) and rotation (40 values). We consider two baseline methods for the task of unsupervised inference of disentangled factors: (i) V AE (Kingma & Welling 2013; Rezende et al. 2014), and (ii) the recently proposed β-VAE (Higgins et al. 2017). To be consistent with the evaluations in (Higgins et al., 2017), we use the same CNN network architectures (for our encoder and decoder), and same latent dimensions as used in (Higgins et al., 2017) for CelebA, 3D Chairs, 2D Shapes datasets.

**Hyperparameters.** For the proposed DIP-V AE-I, in all our experiments we vary \( \lambda_{od} \) in the set \( \{1, 2, 5, 10, 20, 50, 100, 500\} \) while fixing \( \lambda_d = 10\lambda_{od} \) for 2D Shapes and 3D Chairs, and \( \lambda_d = 50\lambda_{od} \) for CelebA. For DIP-V AE-II, we fix \( \lambda_{od} = \lambda_d \). For β-VAE, we experiment with \( \beta = \{1, 2, 4, 8, 16, 25, 32, 64, 100, 128, 200, 256\} \) (where \( \beta = 1 \) corresponds to the V AE). For both CelebA and 2D Shapes, we show the results in terms of the Z-diff score (Higgins et al. 2017), the proposed SAP score, and reconstruction error. For 3D Chairs data, only two ground truth generative factors are available and the quantitative scores for these are saturated near the peak values, hence we show only the latent traversal plots which we based on our subjective evaluation of the reconstruction quality and disentanglement (shown in Appendix).

**Disentanglement scores and reconstruction error.** For the Z-diff score (Higgins et al. 2017), in all our experiments we use a one-vs-rest linear SVM with weight on the hinge loss \( C \) set to 0.01 and weight on the regularizer set to 1. Table 1 shows the Z-diff scores and the proposed SAP scores along with reconstruction error (which directly corresponds to the data likelihood) for the test sets of CelebA and 2D Shapes data. Further we also show the plots of how the Z-diff score...
Figure 1: Proposed Separated Atomic Predictability (SAP) score and the Z-diff disentanglement score (Higgins et al., 2017) as a function of average reconstruction error (per pixel) on the test set of 2D Shapes data for $\beta$-VAE and the proposed DIP-VAE. The plots are generated by varying $\beta$ for $\beta$-VAE, and $\lambda_{\text{od}}$ for DIP-VAE-I and DIP-VAE-II (the number next to each point is the value of these hyperparameters, respectively).

Table 2: Attribute classification accuracy on CelebA: A classifier $w^k = \frac{1}{|x_i: y_i^k = 1|} \sum_{x_i: y_i^k = 1} \mu(x_i) - \frac{1}{|x_i: y_i^k = 0|} \sum_{x_i: y_i^k = 0} \mu(x_i)$ is computed for every attribute $k$ using the training set and a bias is learned by minimizing the hinge loss. Accuracy on other attributes stays about same across all methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Arched Eyebrows</th>
<th>Attractive</th>
<th>Bangs</th>
<th>Black hair</th>
<th>Blond hair</th>
<th>Heavy makeup</th>
<th>Male</th>
<th>Mouth slightly open</th>
<th>No Beard</th>
<th>Wavy hair</th>
<th>Wearing hat</th>
<th>Wearing lipstick</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAE</td>
<td>71.8</td>
<td>73.0</td>
<td>89.8</td>
<td>78.0</td>
<td>88.9</td>
<td>79.6</td>
<td>83.9</td>
<td>76.3</td>
<td>87.3</td>
<td>70.2</td>
<td>95.8</td>
<td>83.0</td>
</tr>
<tr>
<td>$\beta = 2$</td>
<td>71.6</td>
<td>72.6</td>
<td>90.6</td>
<td>79.3</td>
<td>89.1</td>
<td>79.3</td>
<td>83.5</td>
<td>76.1</td>
<td>86.9</td>
<td>67.8</td>
<td>95.9</td>
<td>82.4</td>
</tr>
<tr>
<td>$\beta = 4$</td>
<td>71.6</td>
<td>72.6</td>
<td>90.0</td>
<td>76.6</td>
<td>88.9</td>
<td>77.8</td>
<td>82.3</td>
<td>75.7</td>
<td>85.3</td>
<td>66.8</td>
<td>95.8</td>
<td>80.6</td>
</tr>
<tr>
<td>$\beta = 8$</td>
<td>71.6</td>
<td>71.7</td>
<td>90.0</td>
<td>76.0</td>
<td>87.2</td>
<td>76.2</td>
<td>80.5</td>
<td>73.1</td>
<td>85.3</td>
<td>63.7</td>
<td>95.8</td>
<td>79.6</td>
</tr>
<tr>
<td>DIP-VAE-I</td>
<td>73.7</td>
<td>73.2</td>
<td>90.9</td>
<td>80.6</td>
<td>91.9</td>
<td>81.5</td>
<td>85.9</td>
<td>75.9</td>
<td>85.3</td>
<td>71.5</td>
<td>96.2</td>
<td>84.7</td>
</tr>
</tbody>
</table>

and the proposed SAP score change with the reconstruction error as we vary the hyperparameter for both methods ($\beta$ and $\lambda_{\text{od}}$, respectively) in Fig. 1 (for 2D Shapes data) and Fig. 2 (for CelebA data). The proposed DIP-VAE-I gives much higher Z-diff score at little to no cost on the reconstruction error when compared with VAE ($\beta = 1$) and $\beta$-VAE, for both 2D Shapes and CelebA datasets. However, we observe in the decoder’s output for single latent traversals (varying a single latent while keeping others fixed, shown in Fig. 3 and Fig. 4) that a high Z-diff score is not necessarily a good indicator of disentanglement. Indeed, for 2D Shapes data, DIP-VAE-I has a higher Z-diff score (98.7) and almost an order of magnitude lower reconstruction error than $\beta$-VAE for $\beta = 60$, 

...
Figure 2: The proposed SAP score and the Z-diff score \cite{Higgins2017} as a function of average reconstruction error (per pixel) on the test set of CelebA data for \( \beta \)-VAE and the proposed DIP-VAE. The plots are generated by varying \( \beta \) for \( \beta \)-VAE, and \( \lambda_{od} \) for DIP-VAE-I and DIP-VAE-II (the number next to each point is the value of these hyperparameters, respectively).

however comparing the latent traversals of \( \beta \)-VAE in Fig. 3 and DIP-VAE-I in Fig. 4 indicate a better disentanglement for \( \beta \)-VAE for \( \beta = 60 \) (though at the cost of much worse reconstruction where every generated sample looks like a hazy blob). On the other hand, we find the proposed SAP score to be correlated well with the qualitative disentanglement seen in the latent traversal plots. This is reflected in the higher SAP score of \( \beta \)-VAE for \( \beta = 60 \) than DIP-VAE-I. We also observe that for 2D Shapes data, DIP-VAE-II gives a much better trade-off between disentanglement (measured by the SAP score) and reconstruction error than both DIP-VAE-I and \( \beta \)-VAE, as shown quantitatively in Fig. 1 and qualitatively in the latent traversal plots in Fig. 3. The reason is that DIP-VAE-I enforces \([\text{Cov}_{\phi}(x)|\mu_\phi(x)]_{ii}\) to be close to 1 and this may affect the disentanglement adversely by splitting a generative factor across multiple latents for 2D Shapes where the generative factors are much less than the latent dimension. For real datasets having lots of factors with complex generative processes, such as CelebA, DIP-VAE-I is expected to work well which can be seen in Fig. 2 where DIP-VAE-I yields a much lower reconstruction error with a higher SAP score (as well as higher Z-diff scores).

**Binary attribute classification for CelebA.** We also experiment with predicting the binary attribute values for each test example in CelebA from the inferred \( \mu_\phi(x) \). For each attribute \( k \), we compute the attribute vector \( w^k = \frac{1}{|x: y_k^i = 1|} \sum_{x: y_k^i = 1} \mu_\phi(x) - \frac{1}{|x: y_k^i = 0|} \sum_{x: y_k^i = 0} \mu_\phi(x) \) from the training set, and project the \( \mu_\phi(x) \) along these vectors. A bias is learned on these scalars (by minimizing hinge loss) which is then used for classifying the test examples. Table 2 shows the results for the attribute which show the highest change across various methods (most other attribute accuracies do not change). The proposed DIP-VAE outperforms both VAE and \( \beta \)-VAE for most attributes. The performance of \( \beta \)-VAE gets worse as \( \beta \) is increased further.

5 Concluding Remarks

We proposed a principled variational framework to infer disentangled latents from unlabeled observations. Unlike \( \beta \)-VAE, our variational objective does not have any conflict between the data
Figure 3: Qualitative results for disentanglement in 2D Shapes dataset (Matthey et al., 2017). SAP scores, Z-diff scores and reconstruction errors for the methods (rows) can be read from Fig. 1.
log-likelihood and the disentanglement of the inferred latents, which is reflected in the empirical results. We also proposed the SAP disentanglement metric that is much better correlated with the qualitative disentanglement seen in the latent traversals than the Z-diff score [Higgins et al., 2017]. An interesting direction for future work is to take into account the sampling biases in the generative process, both natural (e.g., sampling the female gender makes it unlikely to sample beard for face images in CelebA) as well as artificial (e.g., a collection of face images that contain much more smiling faces for males than females misleading us to believe \( P(\text{gender, smile}) \neq P(\text{gender})P(\text{smile}) \)), which makes the problem challenging and also somewhat less well defined (at least in the case of natural biases). Effective use of disentangled representations for transfer learning is another interesting direction for future work.

REFERENCES


Appendix

A Latent traversals for 2D Shapes and Chairs dataset

Figure 5: Qualitative results for disentanglement in CelebA dataset.
Figure 6: Qualitative results for disentanglement in Chairs dataset.