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) Interpretabiliy: 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; Karaletos 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 [1]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 (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.
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., \( p(z|x) \). 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 \( \mathcal{D} \) 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., \( q^* = \arg\min_{q \in \mathcal{D}} \mathbb{KL}(q(z)||p(z|x)) \) (Blei et al., 2017). The idea of amortized inference (Kingma & Welling, 2013; Stuhlmüller et al., 2013; Gershman & Goodman, 2014; Rezende et al., 2014) 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

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.
We adopt a simpler alternative of matching the moments of the two distributions. In particular, we refer to equation 1 (Welling, 2013). However, this behavior does not carry over to more complex datasets (Aubry et al., 2012) is also an option, however it has its own challenges when combined with stochastic optimization (Dzugaite et al., 2015; Li et al., 2015).

2.2 INFERRING DISENTANGLED LATENTS

Although the generative model starts with a disentangled prior, our main objective is to infer disentangled latents which are 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,

\[ q_\phi(z) = \int q_\phi(z|x)p(x)dx, \]

which we will subsequently refer to as the inferred prior. For inferring disentangled factors, this should be factorizable along the dimensions, i.e., \( q_\phi(z) = \prod_i q_i(z_i) \), or equivalently \( q_{ij}(z_i|z_j) = q_i(z_i), \forall i, j \). This can be achieved by minimizing a suitable distance between the inferred prior \( q_\phi(z) \) and the disentangled generative prior \( p(z) \). If we take KL-divergence as our choice of distance, by relying on its pairwise convexity (i.e., \( \text{KL}(\lambda p_1 + (1 - \lambda)p_2 || \lambda q_1 + (1 - \lambda)q_2) \leq \text{KL}(p_1 || q_1) + (1 - \lambda)\text{KL}(p_2 || q_2) \) (Van Erven & Harremos, 2014), we can show that this distance is upper bounded by the objective of the variational inference (the ELBO in Eq. (1)):

\[
\text{KL}(q_\phi(z)||p(z)) = \text{KL}(E_x q_\phi(z|x)||E_x p(z|x)) \leq E_x \text{KL}(q_\phi(z|x)||p(z|x)).
\]

Hence, variational posterior inference of latent variables with disentangled prior naturally encourages inferring factors that are 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; Karaletsos et al., 2015). This can be due to non-convexity of the ELBO objective which prevents us from achieving the global minimum of \( E_x \text{KL}(q_\phi(z|x)||p(z|x)) = 0 \) (which would imply \( \text{KL}(q_\phi(z)||p(z)) = 0 \)). In other words, maximizing the ELBO (Eq. (1)) might also result in reducing the value of \( \text{KL}(q_\phi(z)||p(z)) \), however, due to the non-convexity of the loss surface of the ELBO, the gap between \( \text{KL}(q_\phi(z)||p(z)) \) and \( E_x \text{KL}(q_\phi(z|x)||p(z|x)) \) could be large at the stationary point of convergence. Hence, minimizing \( \text{KL}(q_\phi(z)||p(z)) \) explicitly will give us better control on the disentanglement. This motivates us to add \( \text{KL}(q_\phi(z)||p(z)) \) as part of the objective to encourage disentanglement during inference, i.e.,

\[
\max_{q_\phi} E_x \left[ E_{z \sim q_\phi(z|x)} \left[ \log p_\theta(x|z) \right] - \text{KL}(q_\phi(z|x)||p(z)) \right] - \lambda \text{KL}(q_\phi(z)||p(z)),
\]

where \( \lambda \) controls its contribution to the overall objective.

Optimizing directly is not tractable due to the presence of \( \text{KL}(q_\phi(z)||p(z)) \) 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 \( q_\phi(z) \) and \( p(z) \) to estimate a lower bound to \( \text{KL}(q_\phi(z)||p(z)) \). However, this would involve optimizing for a third set of parameters \( \psi \) for the KL-divergence estimator, and would also change the optimization to a saddle-point (min-max) problem which has its own challenges (e.g., gradient vanishing as encountered in training generative adversarial networks with KL or JS divergences (Goodfellow et al., 2014; Arjovsky & Bottou, 2017)). Replacing \( \text{KL}(q_\phi(z)||p(z)) \) with another suitable distance between \( q_\phi(z) \) and \( p(z) \) (e.g., integral probability metrics like Wasserstein distance (Sriperumbudur et al., 2009)) might alleviate some of these issues (Arjovsky et al., 2017) but will still involve complicating the optimization to a saddle point problem in three set of parameters. It should also be noted that using these variational forms of the distances will still leave us with an approximation to the actual distance.

We adopt a simpler 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 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 (Dzugaite et al., 2015; Li et al., 2015).
\( z \sim q_\phi(z) \) if \( p(z) \) is \( N(0, I) \). Let us denote \( \text{Cov}_{q_\phi(z)}(z) := \mathbb{E}_{q(z)} \left[ (z - \mathbb{E}_{q(z)}(z))(z - \mathbb{E}_{q(z)}(z))^\top \right] \).

By the law of total covariance, the covariance of \( z \sim q_\phi(z) \) is given by

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

where \( \mathbb{E}_{q_\phi(z|x)}(z) \) and \( \text{Cov}_{q_\phi(z|x)}(z) \) 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. In this case Eq. (5) reduces to

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

which we want to be close to an identity matrix. For simplicity, we choose entry-wise squared \( \ell_2 \)-norm as the measure of proximity. However, as the entanglement is mainly reflected in the off-diagonal entries of this matrix, we opt for two separate hyperparameters controlling the relative importance of the loss on the diagonal and off-diagonal entries. This gives rise to the following optimization problem for inference:

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

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.

### 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:

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

where \( \beta \) is taken to be greater than 1. Higher \( \beta \) is argued to encourage disentanglement at the cost of reconstruction error (the likelihood term in the ELBO). Authors report empirical results with \( \beta \) ranging from 4 to 250 depending on the dataset. As already mentioned, most VAE models proposed in the literature, including \( \beta \)-VAE, work with \( N(0, I) \) as the prior \( p(z) \) and \( N(\mu_\phi(x), \Sigma_\phi(x)) \) with diagonal \( \Sigma_\phi(x) \) as the approximate posterior \( q_\phi(z|x) \). This reduces the objective (7) to

\[
\max_{\theta, \phi} \mathbb{E}_x \left[ \mathbb{E}_{z \sim q_\phi(z|x)} \left[ \log p_\theta(x|z) \right] - \frac{\beta}{2} \left( \sum_i \left[ \Sigma_\phi(x) \right]_{ii} - \ln \left[ \Sigma_\phi(x) \right]_{ii} + \| \mu_\phi(x) \|^2_2 \right) \right].
\]

For high values of \( \beta \), \( \beta \)-VAE would try to pull \( \mu_\phi(x) \) towards zero and \( \Sigma_\phi(x) \) 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 generated samples which is worse than VAE (\( \beta = 1 \)), particularly for high values of \( \beta \). Our proposed method does not have excessive 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 \( \text{Cov}_{q_\phi(z|x)}(z) \) in Eq. (5) towards the identity matrix: \( \beta \)-VAE attempts to do it by making \( \text{Cov}_{q_\phi(z|x)}(z) \) 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 \( \text{Cov}_{q_\phi(z|x)}(z) \) (marginalizing over the observations \( x \)) which retains the sensitivity of \( q_\phi(z|x) \) to the conditioned-upon-observation.

### 3 Experiments

We evaluate our proposed method, referred as DIP-VAE subsequently (Disentangled Inferred Prior), on three datasets – (i) CelebA (Liu et al., 2015): It consists of 202,599 RGB face images of celebrities. We use 64 \times 64 \times 3 cropped images as used in several earlier works, using 90% for training and 10% for test. (ii) 3D Chairs (Aubry et al., 2014): It consists of 1393 chair CAD models, with each model rendered from 31 azimuth angles and 2 elevation angles. Following earlier work (Yang et al., 2015; Dosovitskiy et al., 2015) that ignores near-duplicates, we use a subset of 809
Table 1: Disentanglement metric score and reconstruction error (per pixel) on the test sets for 2D Shapes and CelebA ($\beta_1 = 4, \beta_2 = 20$ for 2D Shapes, and $\beta_1 = 4, \beta_2 = 8$ for CelebA)

<table>
<thead>
<tr>
<th>Method</th>
<th>2D Shapes</th>
<th></th>
<th></th>
<th>CelebA</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Metric</td>
<td>Reconst. error</td>
<td>Metric</td>
<td>Reconst. error</td>
<td></td>
</tr>
<tr>
<td>VAE</td>
<td>81.3</td>
<td>0.0017</td>
<td>7.5</td>
<td>0.0876</td>
<td></td>
</tr>
<tr>
<td>$\beta$-VAE ($\beta = \beta_1$)</td>
<td>80.7</td>
<td>0.0031</td>
<td>8.1</td>
<td>0.0937</td>
<td></td>
</tr>
<tr>
<td>$\beta$-VAE ($\beta = \beta_2$)</td>
<td>88.0</td>
<td>0.0076</td>
<td>7.1</td>
<td>0.1065</td>
<td></td>
</tr>
<tr>
<td>DIP-VAE</td>
<td>98.1</td>
<td>0.0018</td>
<td>11.31</td>
<td>0.0911</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: Disentanglement metric score ([Higgins et al., 2017]) as a function of reconstruction error for $\beta$-VAE and the proposed DIP-VAE (left: CelebA, right: 2D Shapes). The plots are generated by varying $\beta$ for $\beta$-VAE and $\lambda_{od}$ for DIP-VAE with $\lambda_d$ set to 10$\lambda_{od}$.

chair models in our experiments. We use the binary masks of the chairs as the observed data in our experiments following ([Higgins et al., 2017]). First 80% of the models are used for training and the rest are used for test. (iii) 2D Shapes ([Higgins et al., 2017]): This is a synthetic dataset of binary 2D shapes generated from the Cartesian product of the shape (heart, oval and square), $x$-position (32 values), $y$-position (32 values), scale (6 values) and rotation (40 values). We consider two baselines for the task of unsupervised inference of disentangled factors: (i) VAE ([Kingma & Welling, 2013; Rezende et al., 2014], and (ii) the recently proposed $\beta$-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-VAE, in all our experiments we vary $\lambda_d$ in the set $\{1, 2, 5, 10, 20, 50\}$ while fixing $\lambda_{od} = 10\lambda_{od}$. For $\beta$-VAE, we experiment with $\beta = \{1, 2, 4, 8, 16, 25, 32, 64, 100, 128, 200, 256\}$ (where $\beta = 1$ corresponds to VAE). For both CelebA and 2D Shapes, we show the results for the best performing models in terms of the disentanglement metric score which was introduced in ([Higgins et al., 2017]). For 3D Chairs data, this metric is at 100% for almost all models and we pick the models based on our subjective evaluation of the reconstruction quality and disentanglement.

**Disentanglement metric score and reconstruction error.** ([Higgins et al., 2017]) propose a metric to evaluate the disentanglement performance of the inference mechanism. It assumes 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 $z_x := \mu_y x$ 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 identity of the corresponding generative factor. 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. ([Higgins et al., 2017]) argue that this
To further analyze the latent correlations with the ground truth attributes, we pick one attribute at a time, and plot the correlations between inferred latents for DIP-V AE. Latent correlations for \( \beta \) increase further as \( \beta \) is increased. The reconstruction error for \( \beta \)-VAE gets worse as \( \beta \) is increased further.

**Metric captures the disentangled property of the inferred latents reasonably well.** Table 1 shows the disentanglement metric scores along with reconstruction error (which directly corresponds to the data likelihood) for the test sets for CelebA and 2D Shapes data. It is evident that the proposed DIP-VAE outperforms \( \beta \)-VAE both in terms of the disentanglement metric score and reconstruction error. Further we also show the plot of how the disentanglement metric changes with the reconstruction error as we vary the hyperparameter for both methods (\( \beta \) and \( \lambda \), respectively). It is clear that the proposed method gives much higher disentanglement metric score at little to no cost on the reconstruction error when compared VAE (\( \beta = 1 \)). The reconstruction error for \( \beta \)-VAE gets much worse as \( \beta \) is increased.

**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 \( w^k = \frac{1}{|\mathbf{x}_i : y^k_i = 1|} \sum_{\mathbf{x}_i : y^k_i = 1} \mu_\phi(x_i) - \frac{1}{|\mathbf{x}_i : y^k_i = 0|} \sum_{\mathbf{x}_i : y^k_i = 0} \mu_\phi(x_i) \) 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.

**Correlations in the inferred latent space.** We visualize the Pearson’s correlations between dimensions of the inferred latents \( \mu_\phi(x) \). We also visualize the correlations between inferred latent dimensions and the ground truth attributes which can be taken as a proxy for true generative factors. Tables 3 and 4 show these correlations for VAE, \( \beta \)-VAE and the proposed DIP-VAE. We observe less correlations between inferred latents for DIP-VAE. Latent correlations for \( \beta \)-VAE are even higher than those for VAE which seems to be going against the objective of disentanglement. This can be due to the fact that \( \beta \)-VAE gives less weight to the ELBO which implies the KL(\( q_\phi(z) || p(z) \)) does not get minimized as well (see Eq. 3). Indeed Eq. 3 indicates that \( \beta \)-VAE gives more importance to minimizing the \( \ell_2 \)-norm of individual posterior means \( \mu_\phi(x) \) and does not minimize their correlations as DIP-VAE.

To further analyze the latent correlations with the ground truth attributes, we pick one attribute at a time, and pick the latent dimension that has highest correlation with that attribute. Then we plot the correlations of this latent dimension with rest of the attributes (picking top 4 correlations). This will indicate the entanglement of a latent dimension with the ground truth attributes. It should be noted that many attributes in CelebA are naturally entangled (e.g., lipstick, heavy-makeup and gender) which can be quite difficult to disentangle. These plots for two of the attributes are shown in Fig 2.

Finally we also show the generated samples by varying one latent dimension at a time for CelebA and 3D Chairs data for all three methods in Tables 5 and Tables 6.
Table 3: Correlations between inferred latents (top), and correlations between ground truth attributes and inferred latents (bottom) for 2D Shapes data (Higgins et al., 2017).

<table>
<thead>
<tr>
<th>Latents auto-correlations</th>
<th>Latents-attributes correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAE</td>
<td>$\beta$-VAE ($\beta = 4)$</td>
</tr>
</tbody>
</table>

Table 4: Correlations between inferred latents (top), and correlations between ground truth attributes and inferred latents (bottom) for CelebA dataset.

<table>
<thead>
<tr>
<th>Latents auto-correlations</th>
<th>Latents-attributes correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAE</td>
<td>$\beta$-VAE ($\beta = 8)$</td>
</tr>
</tbody>
</table>

Figure 2: Top attributes similar to Bangs and Rosy Cheeks attribute in CelebA dataset based on correlations with inferred latents for VAE, $\beta$-VAE and the proposed DIP-VAE, respectively.
Table 5: Qualitative results for disentanglement in CelebA dataset.

<table>
<thead>
<tr>
<th></th>
<th>VAE</th>
<th>$\beta$-VAE ($\beta = 4$)</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smile</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
</tr>
<tr>
<td>Azimuth</td>
<td><img src="image4" alt="Image" /></td>
<td><img src="image5" alt="Image" /></td>
<td><img src="image6" alt="Image" /></td>
</tr>
<tr>
<td>Hair color</td>
<td><img src="image7" alt="Image" /></td>
<td><img src="image8" alt="Image" /></td>
<td><img src="image9" alt="Image" /></td>
</tr>
</tbody>
</table>

Table 6: Qualitative results for disentanglement in Chairs dataset.

<table>
<thead>
<tr>
<th></th>
<th>VAE</th>
<th>$\beta$-VAE ($\beta = 4$)</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leg Shape</td>
<td><img src="image10" alt="Image" /></td>
<td><img src="image11" alt="Image" /></td>
<td><img src="image12" alt="Image" /></td>
</tr>
<tr>
<td>Rotation</td>
<td><img src="image13" alt="Image" /></td>
<td><img src="image14" alt="Image" /></td>
<td><img src="image15" alt="Image" /></td>
</tr>
<tr>
<td>Size</td>
<td><img src="image16" alt="Image" /></td>
<td><img src="image17" alt="Image" /></td>
<td><img src="image18" alt="Image" /></td>
</tr>
</tbody>
</table>
4 DISCUSSION

We proposed a principled variational framework with infer disentangled latents from unlabeled observations. Unlike $\beta$-VAE our variational objective does not have any conflict between the data log-likelihood and the disentanglement of the inferred latents, which is reflected in our empirical results that outperform $\beta$-VAE. Directions for future work include tackling the sampling bias in the generative process which makes the problem challenging (e.g., sampling the male gender makes it likely to sample beard), and effective use of disentangled representations in transfer learning.

REFERENCES


Appendix

A Correlations between inferred latent factors and ground truth attributes for CelebA
Under review as a conference paper at ICLR 2018

Eyeglasses

Goatee

Gray Hair
Under review as a conference paper at ICLR 2018
Under review as a conference paper at ICLR 2018
B  LABEL CORRELATIONS FOR CELEBA AND 2D SHAPES DATASETS

Figure 3: VAE latent correlations with attributes for CelebA dataset.
Figure 4: $\beta$-VAE ($\beta = 2$) latent correlations with attributes for CelebA dataset.
Figure 5: $\beta$-VAE ($\beta = 4$) latent correlations with attributes for CelebA dataset.
Figure 6: $\beta$-VAE ($\beta = 8$) latent correlations with attributes for CelebA dataset.
Figure 7: $\beta$-VAE ($\beta = 64$) latent correlations with attributes for CelebA dataset.
Figure 8: DIP-VAE latent correlations with attributes for CelebA dataset.
Figure 9: VAE latent correlations with attributes for 2D Shapes.
Figure 10: $\beta$-VAE ($\beta = 4$) latent correlations with attributes for 2D Shapes.
Figure 11: β-VAE (β = 20) latent correlations with attributes for 2D Shapes.
Figure 12: DIP-VAE latent correlations with attributes for 2D Shapes.