Adversarial Disentanglement with Grouped Observations

Jozsef Nemeth
Ultinous, Hungary
jnemeth@ultinous.com

Abstract

We consider the disentanglement of the representations of the relevant attributes of the data (*content*) from all other factors of variations (*style*) using Variational Autoencoders. Some recent works addressed this problem by utilizing grouped observations, where the content attributes are assumed to be common within each group, while there is no any supervised information on the style factors. In many cases, however, these methods fail to prevent the models from using the style variables to encode content related features as well. This work supplements these algorithms with a method that eliminates the content information in the style representations. For that purpose the training objective is augmented to minimize an appropriately defined mutual information term in an adversarial way. Experimental results and comparisons on image datasets show that the resulting method can efficiently separate the content and style related attributes and generalizes to unseen data.

1 Introduction

In the field of representation learning (Bengio, Courville, and Vincent 2013), autoencoder based approaches (Tschannen, Bachem, and Lucic 2018) are among the most effective methods to learn compact and meaningful representations even without any supervision. Such representations then can be used to solve downstream tasks like classification or clustering efficiently. Variational Autoencoders (VAEs) (Kingma and Welling 2014) attracted probably the most attention in recent years. By employing stochastic variational inference (Zhang et al. 2019), VAEs can learn intractable posterior distributions of the latent variables.

Disentangled representation learning (Desjardins, Courville, and Bengio 2012; Kumar, Sattigeri, and Balakrishnan 2018; Achille and Soatto 2018; Esmaeili et al. 2019; Xiang and Li 2019) aims to assign the different factors of variations to different dimensions of the representation vectors. This problem has been heavily studied in recent years, for example, it has been proved that learning disentangled representations without any inductive bias or supervision is theoretically impossible (Locatello et al. 2019). At the same time, it has been observed that using increased regularization weight, the VAE model learns disentangled latent spaces (Higgins et al. 2017).

In this work, we consider the problem of learning a disentangled representation in which the relevant attributes of the observations (which determine some kind of identity or class of a given sample) and the other irrelevant factors (like pose, lighting conditions, *etc.*) are separated. Following the notations of (Bouchacourt, Tomioka, and Nowozin 2018), we refer to the relevant attributes and the other factors as *content* and *style*, respectively. Recently, grouped observations based methods have been proposed for content-style disentanglement (Bouchacourt, Tomioka, and Nowozin 2018; Hosoya 2019). Grouping data elements that represent the same content is a way to induce weak supervision, and the main motivation behind this approach is that collecting and cleaning such datasets is less expensive. For example, in case of classification based methods it is expected that different classes represent different identities. At the same time, in case of learning from grouped data, we only expect that members of a given group share the same content, whereas different groups are not required to represent different classes. These methods (Bouchacourt, Tomioka, and Nowozin 2018; Hosoya 2019) share the idea of accumulating the content representation of the individual group members into a single common group-level content representation, while they learn individual style representations for each of the group members. However, these methods have no explicit mechanism to prevent the model from using the style variable to represent not only style attributes but also certain amount of content information. For example, the number of dimensions for the style variable had to be adjusted for the style variability, otherwise the model used the style variable to learn content related aspects of the data (Hosoya 2019). Our experiments also confirm this observation.

We show that the results of these methods can be significantly improved by suppressing the content information in the style variable. More specifically, the proposed method trains a neural network to estimate the mutual information (Belghazi et al. 2018) between the style representations and the observations corresponding to the same groups whereas the encoder network is trained in an adversarial manner. Exper-

iments prove that this enhanced training process provides more useful representations. The improvements are the most significant for small group sizes, which makes the proposed method more applicable in practice.

This paper is organized as follows. First, we briefly overview the related literature and the grouped observations based methods (Bouchacourt, Tomioka, and Nowozin 2018; Hosoya 2019) in Section 2. Then Section 3 introduces our adversarial disentanglement approach. Experimental results and comparisons can be found in Section 4 and in the supplementary material. The source code of the experiments and supplementary material are available online[1].

2 Related Work

The different variants of autoencoders are powerful tools for learning efficient data representations (Tschannen, Bachem, and Lucic 2018). Among them Variational Autoencoders (VAEs) (Kingma and Welling 2014) are the most heavily studied approaches. VAEs were developed to model probabilistic latent representations based on the variational inference principle. It has also been shown that by slightly modifying the VAEs objective function, the models tend to disentangle the latent space (Higgins et al. 2017; Burgess et al. 2017). This phenomenon was further investigated by Kim and Mnih (2018) leading to the FactorVAE method. Chen *et al.* (2018) also studied the disentangling properties of VAEs. They proposed to split the regularization term up to four parts and assign different weights to each term. Another group of methods uses structured latent spaces to learn efficient data representations (Grathwohl and Wilson 2016; Maaløe et al. 2019; Ranganath, Tran, and Blei 2016; Klys, Snell, and Zemel 2018). It has also been demonstrated that VAEs are capable to separate the continuous and discrete generative factors (Dupont 2018), furthermore, disentanglement can also be achieved in the semi-supervised setting (Kingma et al. 2014; Siddharth et al. 2017).

Generative Adversarial Networks (Goodfellow et al. 2014) are probabilistic methods designed to model high dimensional data distributions. They can be extended to representation learning, for example, the InfoGAN model (Chen et al. 2016) can learn disentangled representations by maximizing the mutual information between the generated data and a subset of latent variables. Another GAN based model has been proposed recently which can be used to control the content and the style of the generated images (Chen, Denoyer, and Artières 2018). During training, the generator aims to produce image pairs from shared content vectors but independent style vectors, while the discriminator tries to distinguish these fake pairs from real image pairs. The generator can compete with the discriminator only by producing image pairs containing the same object. Donahue *et al.* (2018) proposed a similar method although their approach to the discriminator differs.

Adversarial training to disentangle content and style attributes has been proposed recently (Mathieu et al. 2016). The basic idea of this work is to pair the content variable of a data sample with the style variable of another sample from the same group during training. The resulting representation should be suitable to recover the the latter sample using the decoder. This approach is similar in spirit with the grouped observations based methods (Bouchacourt, Tomioka, and Nowozin 2018; Hosoya 2019), on which the current work is based. The main difference is that the latter methods do not require access to class labels. Kulkarni *et al.*(2015) proposed a method to disentangle the generative factors and create a compact and interpretable representation. The algorithm encourages the different dimensions of the latent variable to represent specific attributes of the images. Recently, Wu *et al.* (2019) proposed an architecture to disentangle geometry and style information using prior knowledge on structure. Other GAN and autoencoder based generative models were successfully applied to face frontalization (Yin et al. 2017) and recognition (Liu et al. 2018). Learning representations that are invariant to specific factors is also highly related to the current work (Jaiswal et al. 2018). It has recently been shown that invariance can be obtained without adversarial training as well (Moyer et al. 2018).

2.1 Multi-Level Variational Autoencoders

To separate the content and style representations Bouchacourt, Tomioka, and Nowozin (2018) introduced the Multi-Level Variational Autoencoder (MLVAE) which utilizes a dataset of grouped observations. In this section, we briefly overview this approach. Let the training dataset consist of N observation groups $\mathbf{x}^n = \left\{x_1^n, \ldots, x_{K_n}^n\right\}$ for $n = 1, \ldots, N$, where $K_n = |\mathbf{x}^n|$ denote the number of individual observations within the group and $x_i^n \in \mathbb{R}^d$ is the *ith* member of the *nth* group. The group index n is omitted sometimes for the sake of simplicity and in this work we consider only datasets with $K_n = K$, *i.e.*, all the groups in the dataset are of the same size. The underlying empirical data distribution represented by the dataset is denoted by $p_{\mathcal{D}}(\mathbf{x})$. We also define the distribution of individual observations $p_{\mathcal{D}}(x)$. Note that sampling from $p_{\mathcal{D}}(x)$ can be performed by first sampling a group from $p_{\mathcal{D}}(\mathbf{x})$ then choosing a member uniformly at random. We assume that the members in a given group share some content attributes that we want to capture by the latent variable $c \in \mathbb{R}^{d_c}$, while the individual group members have style attributes represented by the latent variables $\mathbf{s} = \left\{s_i \in \mathbb{R}^{d_s} : i = 1, \ldots, K\right\}$. The Evidence Lower Bound (ELBO) for a group (Bouchacourt, Tomioka, and Nowozin 2018) is given by:

$$\begin{aligned}\log p_\theta(\mathbf{x}) \geq \mathbb{E}_{q_\phi(c,\mathbf{s}|\mathbf{x})} \sum_{i=1}^{K} \log p_\theta(x_i|c, s_i) \\ - \sum_{i=1}^{K} KL(q_\phi(s_i|x_i)|p(s_i)) \\ - KL(q_\phi(c|\mathbf{x})|p(c)) = \mathcal{L}_{\theta,\phi}(\mathbf{x}), \qquad (1)\end{aligned}$$

where $p(s_i)$ and $p(c)$ are priors on the latent variables, while θ is the parameter of the generative model and ϕ is the variational parameter of both the content and style variables. The training process maximizes this lower bound for the whole

[1] https://github.com/jonemeth/aaai20

dataset:

$$\mathcal{L}_{\theta,\phi}(\mathcal{D}) = \frac{1}{N}\sum_{n=1}^{N}\mathcal{L}_{\theta,\phi}(\mathbf{x}^n). \quad (2)$$

As usually in case of Variational Autoencoders, both the generative model p_θ and the approximate inference model q_ϕ are realized by neural networks P_θ and Q_ϕ with parameters θ and ϕ, respectively. In the experiments presented in this paper the encoder neural network estimated the expected value and log-variance of the Gaussian approximate posterior distribution and we used standard normal priors everywhere. Furthermore, we considered Bernoulli likelihood for binary images and Gaussian likelihood (with fixed variance of 1.0) for color images.

Content Accumulation For the MLVAE method, it was proposed to define the single content posterior for a given group based on the content encodings obtained from the individual observations withing that group. More specifically, the product of the posterior density functions provided by the encoder was considered:

$$q_\phi(c|\mathbf{x}) = \frac{1}{Z}\prod_{i=1}^{K} q_\phi(c|x_i), \quad (3)$$

where Z is a normalization constant. Although theoretically other approaches are also feasible (such as accumulating as a Gaussian mixture), in our experiments we focus on Eq. (3). In (Hosoya 2019), another accumulation method has been proposed. Our experiments in the supplemental show that their simplified accumulation approach slightly improves the results, but the adversarial method described in the next section is still important to obtain significantly better disentanglement.

3 Adversarial Disentanglement

From the point of view of disentangling the latent space, the accumulation of the content posteriors over group members effectively discourages the encoder from storing style related information in the content variables. This is because the decoder reconstructs the inputs from the accumulated contents and individual styles. On the other hand, there is no explicit mechanism that would prevent the model from encoding content in the style variable s. For example, in case of the Chairs dataset (Yang et al. 2015), in which the only factors of style variations are two rotation angles, it is important to choose a low number of dimensions for s (*e.g.*, $d_s = 2$) to prevent the model from storing content in it (Hosoya 2019).

Here we argue that the reason why maximizing the Group-ELBO in Eq. (1) can result in disentanglement (at least to some extent) is the imbalanced regularization approach. Informally speaking, Eq. (1) puts more weight on regularizing the style variable (which in turn minimizes its information content about the data (Kim and Mnih 2018)) and thus encourages the encoder to use the content variable as well. As experimental results in Section 4 and in the supplementary material show, larger group size leads to better disentanglement, however, for small group sizes the algorithm ignores the content variable c, and uses only s to represent all aspects of the data. One might expect to obtain better results by increasing the regularization on s and thus encouraging the model to use the content vector c. Experiments presented in Section 4 show that while this approach increases the disentanglement indeed, it also harms the overall performance of the method. This is because increased regularization reduces the capacity of s and weaker style encoding affects the quality of the content representation as well.

In this section, we show how to minimize only the content related information in the style representation. The group members are assumed to share the same content, but the style attributes are independent. Therefore, we expect that the style representation of a given sample can not be inferred from other data samples within the same group. The content information encoded in s can be measured as the mutual information between s and other data samples in the given group. To formulate this concept, let us first define:

$$r_\phi(s|x) = r_\phi(s|x = x_i^n) \equiv \frac{1}{K-1}\sum_{j=1,j\neq i}^{K} q_\phi(s|x_j^n) \quad (4)$$

and

$$r_\phi(x) \equiv p_\mathcal{D}(x), \quad (5)$$

where $p_\mathcal{D}(x)$ is the underlying distribution of individual observations. The joint distribution of the observations and style representations of other group members is $r_\phi(x,s) = r_\phi(x)r_\phi(s|x)$, furthermore, it can be easily shown that $r_\phi(s) = q_\phi(s)$ (see the supplementary material). Moreover let us denote the factorized distribution $r_\phi(x)r_\phi(s)$ by $\bar{r}_\phi(x,s)$. We can express the mutual information between data samples and the style representations of other group members as the mutual information between s and x w.r.t. the distribution r_ϕ:

$$\begin{aligned} I_{r_\phi}(x;s) &= KL(r_\phi(x,s)\|\bar{r}_\phi(x,s)) \\ &= \mathbb{E}_{r_\phi(x,s)}\ \log\frac{r_\phi(x,s)}{\bar{r}_\phi(x,s)}. \end{aligned} \quad (6)$$

The goal of the proposed method is to maximize the lower bound in Eq. (2), while keeping $I_{r_\phi}(x;s)$ low enough to make s invariant to content and thus force the encoder to use only the latent vector c to represent content attributes. Instead of handling the problem as a constrained optimization task, we propose to take its Lagrangian relaxation by penalizing the mutual information, which leads to the following training objective:

$$\underset{\theta,\phi}{\text{maximize}} \quad \mathcal{L}_{\theta,\phi}(\mathcal{D}) - \lambda I_{r_\phi}(x;s), \quad (7)$$

where the parameter λ controls the weight of the mutual information term. We can estimate the mutual information term in Eq. (7) using a parametric neural estimator (Belghazi et al. 2018) as it is possible to generate samples from both $r_\phi(x,s)$ and $\bar{r}_\phi(x,s)$. For example, to generate a sample from $r_\phi(x,s)$ we can first choose a group $\mathbf{x}^n$ and pick two members x_i^n and x_j^n $(i \neq j)$ uniformly at random. Then let $x = x_i^n$ and $s \sim q_\phi(s|x_j^n)$. The samples from $\bar{r}_\phi(x,s)$ are pairs of independent points from $r_\phi(x) = p_\mathcal{D}(x)$ and $r_\phi(s) = q_\phi(s)$. Thus to generate a sample, we can first pick

Algorithm 1: Generating samples from $r_\phi(x,s)$ and $\bar{r}_\phi(x,s)$

Input : $B = \{\mathbf{x}^b : b = 1,\dots,N_B\}$.

$R \leftarrow \emptyset$, $\bar{R} \leftarrow \emptyset$.
foreach $\mathbf{x}^b \in B$ **do**
  Choose i and j ($i \neq j$) uniformly at random from $\{1,\dots,K\}$.
  Sample $s_j^b \sim q_\phi(s|x_j^b)$.
  $R \leftarrow R \cup \{(x_i^b, s_j^b)\}$.
  $m \leftarrow 1 + (b \mod N_B)$ `# Index of next group`
  Choose k uniformly at random from $\{1,\dots,K\}$.
  Choose l uniformly at random from $\{1,\dots,K\}$.
  Sample $s_l^m \sim q_\phi(s|x_l^m)$.
  $\bar{R} \leftarrow \bar{R} \cup \{(x_k^b, s_l^m)\}$.
end
return R, $\bar{R}$

two observations $x_i^n \sim p_\mathcal{D}(x)$ and $x_j^m \sim p_\mathcal{D}(x)$, then choose $x = x_i^n$ and $s \sim q_\phi(s|x_j^m)$.

As usually in practice, we train our models using mini-batch based stochastic optimization methods, thus it is the most efficient to draw samples from $r_\phi(x,s)$ and $\bar{r}_\phi(x,s)$ based on the same mini-batches of groups that we use to maximize Eq. (2). The pseudo-code of our mini-batch based sampling approach is described in Algorithm 1. We note that to generate a sample from $\bar{r}_\phi(x,s)$ we never choose x and s from the same group otherwise such pairs would be over-represented. For the same reason, to generate samples from $r_\phi(x,s)$ a given data point x is never paired with a style vector sampled from its style posterior $q_\phi(s|x)$.

Since samples from $r_\phi(x,s)$ and $\bar{r}_\phi(x,s)$ are accessible, it is possible to estimate $I_{r_\phi}(x;s)$ as:

$$I_{r_\phi}(x;s) \geq \sup_\psi \big\{\mathbb{E}_r[T_\psi(x,s)] - \log(\mathbb{E}_{\bar{r}}[e^{T_\psi(x,s)}])\big\}, \quad (8)$$

where T_ψ is a neural network parameterized by ψ, see (Belghazi et al. 2018; Nguyen, Wainwright, and Jordan 2010). Let R and $\bar{R}$ be sets of samples from $r_\phi(x,s)$ and $\bar{r}_\phi(x,s)$, respectively. We train T_ψ to maximize

$$\mathcal{L}_\psi(R,\bar{R}) = \sum_{(x,s)\in R} \frac{T_\psi(x,s)}{|R|} - \log \sum_{(x,s)\in\bar{R}} \frac{e^{T_\psi(x,s)}}{|\bar{R}|}, \quad (9)$$

and use this approximation to minimize the mutual information in Eq. (7) in an adversarial manner. To achieve a sufficiently small mutual information, we adaptively adjust the weight of the mutual information term in Eq. (7) in each iteration of the training process:

$$\lambda \leftarrow \lambda + \alpha\big(\mathcal{L}_\psi(R,\bar{R})/I^\star - 1\big), \quad (10)$$

where $I^\star$ is the target value for the mutual information (we set $I^\star = 0.2$ in our experiments), $\mathcal{L}_\psi(R,\bar{R})$ is evaluated based on the current training mini-batch, while α is a step-size (we used $\alpha = 0.1$ in our experiments). This update rule increases (decreases) the weight λ if the current estimation of the mutual information is higher (lower) than the target value $I^\star$.

Algorithm 2: MLVAE with Adversarial Disentanglement

Input : Dataset $\{\mathbf{x}^n : n = 1,\dots,N\}$, mini-batch size N_B, number of training iterations it, number of T_ψ update steps it_T, optimizers $g_{\theta,\phi}$ and g_ψ.

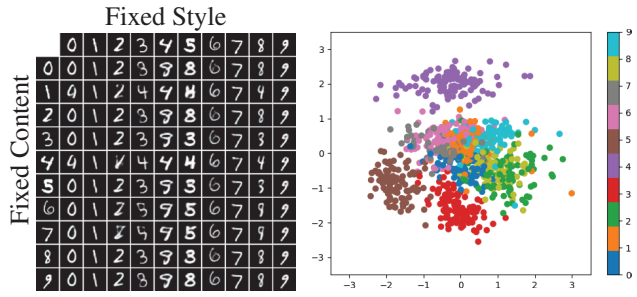
Initialize θ,ϕ, ψ, and λ.
repeat it **times**
  Sample a mini-batch of observation groups $B = \{\mathbf{x}^b : b = 1,\dots,N_B\}$.
  Use Algorithm 1 with B as input to obtain two sets R and $\bar{R}$ containing samples from $r_\phi(x,s)$ and $\bar{r}_\phi(x,s)$, respectively.
  $\theta, \phi \leftarrow g_{\theta,\phi}\big(-\mathcal{L}_{\theta,\phi}(B) + \lambda\mathcal{L}_\psi(R,\bar{R})\big)$.
  Update λ using Eq. (10).
  repeat it_T **times**
    Sample a mini-batch of groups $B = \{\mathbf{x}^b : b = 1,\dots,N_B\}$.
    Use Algorithm 1 with B as input to obtain two sets R and $\bar{R}$ containing samples from $r_\phi(x,s)$ and $\bar{r}_\phi(x,s)$, respectively.
    $\psi \leftarrow g_\psi\big(-\mathcal{L}_\psi(R,\bar{R})\big)$.
  end
end

As experimental results prove, the method described in this section makes the style variable s less informative about the content, thus forcing the method to learn more effective content representation in c. We refer to MLVAE improved with our adversarial disentanglement method as MLVAE-AD. The pseudo code in Algorithm 2 shows the main steps of the proposed algorithm.

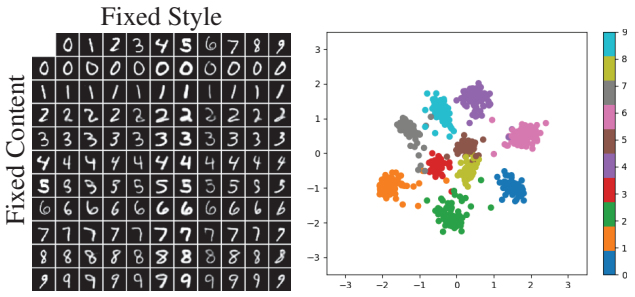
4 Experiments

In our experiments we focused on testing the effect of the proposed adversarial disentanglement approach. We used three image datasets that exhibit different variability in both content and style. To evaluate the disentanglement in the learned representations simple classifiers were trained on both the content and style mean variables. For that purpose, each of the datasets had been split into three parts. For a given dataset, the first set was used to form the groups for training the models, the classifiers were trained on the second, while the classification performances were evaluated on the last part. Herein, we present the classification results obtained by standard SVM. In the supplemental, evaluation using linear regression classifier shows that such a simpler method can also easily distinguish different classes in the content representations, but performs weaker in discovering content information in style encodings.

The MNIST (LeCun et al. 1998) dataset is composed of only 10 classes of handwritten digits, while there is a wide range of variability in style. We split the set of 50000 training images into two parts. The models were trained on 45000



(a) MLVAE



(b) MLVAE-AD

Figure 1: Example results from the qualitative experiments on MNIST using $d_c = 2$, $d_s = 14$, and $K = 2$. The figure contains generated images obtained from swapped encodings (see text) and scatter plots for 2-dimensional content embeddings.

Table 1: Evaluation of the results obtained on the MNIST test set using different numbers of latent dimensions and group sizes. Results for $d_c = 2, d_s = 14$ and $d_c = 8, d_s = 8$ are shown in rows 1-3 and 4-6, respectively.

K	MLVAE			MLVAE-AD		
	$\mathcal{C}(c)$	$\mathcal{C}(s)$	$\mathcal{L}_{rec}$	$\mathcal{C}(c)$	$\mathcal{C}(s)$	$\mathcal{L}_{rec}$
2	65.0%	89.2%	75.2	97.6%	41.2%	78.2
5	92.9%	85.2%	76.0	97.3%	40.2%	80.6
10	94.1%	85.9%	75.7	96.3%	57.9%	79.5
2	84.6%	79.6%	77.4	98.4%	26.8%	79.6
5	97.3%	61.9%	81.5	98.5%	29.4%	85.3
10	97.6%	60.9%	83.5	98.0%	30.6%	86.0

samples, in case of a given group size we randomly formed 10000 groups from each of the 10 classes. The remaining 5000 images were used to train the classifiers, while the MNIST test set was used for evaluation. The original 28×28 pixels images were padded to 32×32 pixels to fit to the network architecture.

The Chairs (Yang et al. 2015) dataset contains rendered images of about one thousand different three-dimensional chair models, but the intra-class variability is very low as the images within a given class differ only by the view of the model. The version of this dataset that we used contains 64×64 pixel images of 809 different three-dimensional chair models rendered from 62 different views. In the experiments, our algorithm was trained on 659 randomly chosen classes, while the remaining 150 classes were used for evaluation. For a given group size, 100 random groups were formed from each of the 659 training classes. From each of the 150 test classes we used 31 random images to train the classifiers, while their accuracies were measured on the remaining 31 samples.

Finally, the VGGFace2 (Cao et al. 2018) face recognition dataset represents high variability in both content and style. It contains more than 3 million in-the-wild images of about 9000 identities. The training set consist of 8631 classes, for each class we formed 50 random groups for a given group size. From each of the 500 test classes, we randomly selected

Table 2: Results of MLVAE on MNIST ($d_c = 2$, $d_s = 14$, $K = 2$) with increased style regularization weight β.

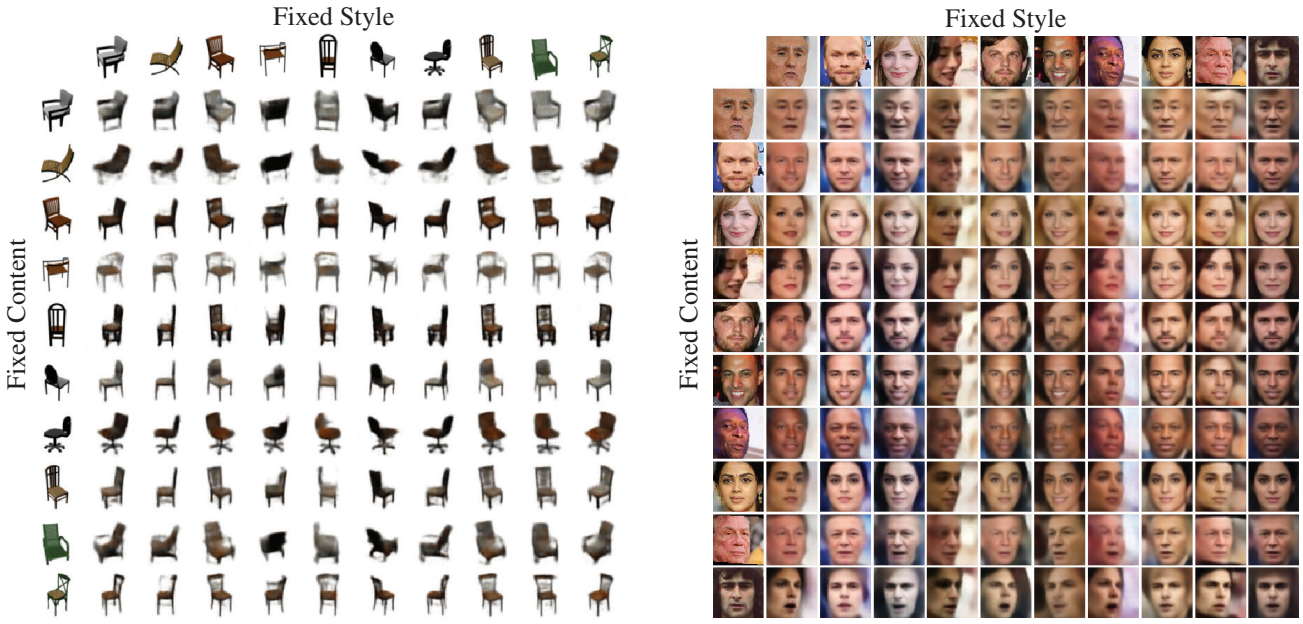
β	$\mathcal{C}(c)$	$\mathcal{C}(s)$	$\mathcal{L}_{rec}$
1.5	87.6%	77.4%	80.4
2.0	93.5%	55.5%	85.2
5.0	95.6%	30.9%	106.8
10.0	94.3%	20.3%	128.7
20.0	94.5%	29.2%	141.3

50 images to train the classifiers and 10 other images for evaluation. The images were aligned based on facial keypoints provided with the dataset and resized to 64×64 pixels.

For the Chairs and VGGFace2 datasets the classes used to train the models were disjoint from the classes used for evaluation. Thus, experiments on these sets show the ability of the method to generalize to unseen contents.

To implement the encoder, the decoder, and T_ψ we used simple convolutional neural networks. Note that T_ψ is unusual in the sense that it takes a (higher dimensional) data point and a (lower dimensional) representation vector as input. To implement such a network, we first used a similar neural-network structure that we used for the encoder to map x into a lower dimensional vector. Then we concatenated the resulting vector with s and applied some more layers to obtain the output of T_ψ. Note that the neural network architectures and training hyper-parameters were not fine-tuned to obtain high quality results. Adam optimizer (Kingma and Ba 2015) was applied in all of the experiments. The description of the neural networks and the optimization parameters used for the different experiments can be found in the supplementary material.

For qualitative evaluation we created image plots with *swapped latents*. For such image grids, we encoded a few images from the test datasets and swapped the content and style vectors to obtain new representations. These representations were then fed into the decoder to generate the images. In such an image grid, the first row and the left-most column contain the input samples.



(a) Chairs (b) VGGFace2

Figure 2: Qualitative experiments on Chairs ($d_c = 32, d_s = 32, K = 2$) and VGGFace2 ($d_c = 64, d_s = 64, K = 2$) datasets: decoding outputs from swapped representations.

Table 3: Quantitative results on Chairs test set with $d_c = 32$, $d_s = 32$, and different group sizes.

K	MLVAE			MLVAE-AD		
	$\mathcal{C}(c)$	$\mathcal{C}(s)$	$\mathcal{L}_{rec}$	$\mathcal{C}(c)$	$\mathcal{C}(s)$	$\mathcal{L}_{rec}$
2	76.3%	74.5%	85.7	93.8%	8.6%	111.7
5	82.4%	72.3%	89.8	95.0%	15.4%	110.7
10	92.3%	60.9%	97.9	92.8%	25.7%	113.1

Table 4: Results on the test set of VGGFace2 with $d_c = 64$, $d_s = 64$, and different group sizes.

K	MLVAE			MLVAE-AD		
	$\mathcal{C}(c)$	$\mathcal{C}(s)$	$\mathcal{L}_{rec}$	$\mathcal{C}(c)$	$\mathcal{C}(s)$	$\mathcal{L}_{rec}$
2	21.4%	27.5%	209.4	39.0%	6.8%	260.9
5	31.7%	27.9%	212.7	36.8%	13.1%	251.0
10	39.6%	27.5%	215.5	41.1%	14.1%	249.9

4.1 Evaluation

When the representation is successfully disentangled, we expect to obtain high classification accuracy on content representations while low accuracy on style representations. In our tables, the classification accuracies obtained on the content and style representations are denoted by $\mathcal{C}(c)$ and $\mathcal{C}(s)$, respectively. We also present the average reconstruction errors (average per-sample negative log-likelihoods) obtained on the test sets which is denoted by $\mathcal{L}_{rec}$. A significant increase in reconstruction error would suggest that adversarial disentanglement weakens the representation capacity of the model. In our experiments, we focused on small group sizes ($K = 2, 5, 10$). For a given setting, the models were trained both with and without adversarial disentanglement.

In case of the MNIST handwritten digits, the focus of the experiments was on examining whether the proposed method is capable of forcing the model to represent the 10 classes on a very low-dimensional content vector ($d_c = 2$). We tested the algorithm for $d_c = 2, d_s = 14$ and for $d_c = 8, d_s = 8$ as well. Results in Table 1 show that the original MLVAE method obtains poor disentanglement when the group size is small. At the same time when adversarial training is applied, the method can separate the content and style attributes well. Figure 1 shows an example where MLVAE failed to use only c to represent the content, while MLVAE-AD formed a content distribution in which the 10 classes are well separated. Additional experiments on the MNIST dataset can be found in the supplementary material.

As it was mentioned previously, one might expect that the original MLVAE method could obtain better disentanglement with increased regularization weight on the style posterior, as it would force the model to use the content vector. This would be a simple alternative to our adversarial approach. On the other hand, increased regularization decreases the mutual information between the latents and the data, thus it suppresses not only the content but also the style information in s. To show this, we assigned a weight β to the second term

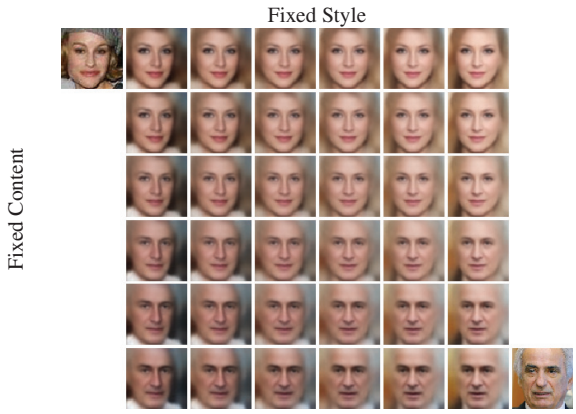


Figure 3: Latent traversals between encodings of test images at the upper-left and bottom-right. These results are from the experiment with $K = 2$.

of the MLVAE objective:

$$\mathbb{E}_{q_\phi(c,\mathbf{s}|\mathbf{x})} \sum_{i=1}^{K} \log p_\theta(x_i|c, s_i) - \beta \sum_{i=1}^{K} KL(q_\phi(s_i|x_i)|p(s_i)) - KL(q_\phi(c|\mathbf{x})|p(c)). \quad (11)$$

Results for different values of β can be found in Table 2. The table shows classification accuracies and average reconstruction errors measured on the MNIST test set. Comparing these results to the first row of Table 1, it can be seen that although increased regularization results in better disentanglement, it also weakens the overall capacity of the model (higher reconstruction errors) while the content representations are less effective compared to those of the proposed method.

In case of the Chairs dataset, the intra-class variability is restricted to the view (rotation) of the objects and this low style variability can be efficiently represented with a low dimensional style variable, *e.g.*, $d_s = 2$ (Hosoya 2019). Therefore, we tested the performance of our adversarial approach for a case when the style variable is strongly over-parameterized ($d_s = 32$). Experimental results presented in Table 3 and Figure 2a show that the proposed method efficiently eliminates content related information from s.

Separating the facial identity from other factors like facial expression, pose or illumination is a more challenging task. Quantitative results in Table 4 shows that adversarial disentanglement improves the results of the original MLVAE method while Figure 2b and Figure 3 demonstrate that the proposed method was able to separate the most important attributes corresponding to facial identity. Additional qualitative experiments can be found in the supplementary material.

4.2 Comparisons

To learn representations that are invariant to specific factors, an autoencoder based approach was recently proposed by Jaiswal et al. (2018). This method is similar to the current work as it is unsupervised on the considered nuisance factors and applies adversarial training to induce invariance, but differs in that it requires supervision on the target variables and in its non-probabilistic nature. We considered their experiment on a rotated version of MNIST (MNIST-ROT), using the same single hidden layer network architectures, except that our encoder outputs latent variances as well. Both the method and the classifier were trained on $\Theta \in \{0°, \pm 22.5°, \pm 45°\}$ with $d_c = 10, d_s = 20$, $K = 2$ and tested with rotations Θ, $\pm 55°$, and $\pm 65°$. The results and comparison to (Jaiswal et al. 2018) can be found in Table 5.

Table 5: Classification results ($\mathcal{C}(c)$) and comparison on MNIST-ROT. For the MLVAE and MLVAE-AD methods, means and standard deviations over 10 runs are presented.

	Jaiswal et al.	MLVAE	MLVAE-AD
Θ	**97.7%**	52.2% ($\pm$ 3.8)	95.1% ($\pm$ 0.2)
$\pm 55°$	85.6%	41.9% ($\pm$ 3.4)	**85.7%** ($\pm$ 0.6)
$\pm 65°$	69.6%	29.3% ($\pm$ 3.2)	**71.0%** ($\pm$ 0.9)

As a further experiment, we considered (Kingma et al. 2014). Although this method was proposed for semi-supervised learning, it can also be used in a supervised setting. For the MNIST dataset with all available training labels a test-set accuracy of 99.04% was reported. For comparison we employed the same two-hidden MLPs and set $d_c = 10$ and $d_s = 50$. Averaged over 10 runs, using our method we achieved a classification accuracy of 97.86% (± 0.13), while with the simplified content accumulation strategy of (Hosoya 2019) using our method we obtained 98.09% (± 0.08).

These results indicate that our adversarial method renders grouped observations based learning competitive to supervised methods that learn from explicit target labels.

5 Conclusion

This paper proposed an adversarial disentanglement approach to improve the results of grouped observations based methods. To eliminate content related information in the style variable, the algorithm minimizes the predictability of the style representation of a given data sample from other members within the same group. This idea was formulated as the minimization of an appropriately constructed mutual information term. We proposed to estimate this mutual information using a parametric neural estimator and train the encoder in an adversarial manner. Experiments and comparisons on image datasets demonstrated the efficacy of the proposed method in separating the content and style related attributes and its ability to generalize to unseen classes.

References

Achille, A., and Soatto, S. 2018. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research* 19(1):1947–1980.

Belghazi, M. I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; and Hjelm, D. 2018. Mutual information neural estimation. In *Proc. of ICML*.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation

learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(8):1798–1828.

Bouchacourt, D.; Tomioka, R.; and Nowozin, S. 2018. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Proc. of AAAI*.

Burgess, C. P.; Higgins, I.; Pal, A.; Matthey, L.; Watters, N.; Desjardins, G.; and Lerchner, A. 2017. Understanding disentangling in beta-vae. In *Learning Disentangled Representations: From Perception to Control Workshop*.

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*.

Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proc. of NIPS*.

Chen, R. T. Q.; Li, X.; Grosse, R.; and Duvenaud, D. 2018. Isolating sources of disentanglement in variational autoencoders. In *Proc. of NIPS*.

Chen, M.; Denoyer, L.; and Artières, T. 2018. Multi-view data generation without view supervision. In *Proc. of ICLR*.

Desjardins, G.; Courville, A.; and Bengio, Y. 2012. Disentangling factors of variation via generative entangling. *arXiv e-prints* abs/1210.5474.

Donahue, C.; Lipton, Z. C.; Balsubramani, A.; and McAuley, J. 2018. Semantically decomposing the latent spaces of generative adversarial networks. In *Proc. of ICLR*.

Dupont, E. 2018. Learning disentangled joint continuous and discrete representations. In *Proc. of NIPS*.

Esmaeili, B.; Wu, H.; Jain, S.; Bozkurt, A.; Siddharth, N.; Paige, B.; Brooks, D. H.; Dy, J.; and van de Meent, J.-W. 2019. Structured disentangled representations. In *Proc. of AISTATS*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Proc. of NIPS*.

Grathwohl, W., and Wilson, A. 2016. Disentangling space and time in video with hierarchical variational auto-encoders. *arXiv e-prints* 1612.04440.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. Beta-vae: Learning basic visual concepts with a constrained variational framework. In *Proc. of ICLR*.

Hosoya, H. 2019. Group-based learning of disentangled representations with generalizability for novel contents. In *Proc. of IJCAI*.

Jaiswal, A.; Wu, R. Y.; Abd-Almageed, W.; and Natarajan, P. 2018. Unsupervised adversarial invariance. In *Proc. of NIPS*.

Kim, H., and Mnih, A. 2018. Disentangling by factorising. In *Proc. of ICML*.

Kingma, D. P., and Ba, J. 2015. Adam: A method for stochastic optimization. In *Proc. of ICLR*.

Kingma, D. P., and Welling, M. 2014. Auto-encoding variational bayes. In *Proc. of ICLR*.

Kingma, D. P.; Rezende, D. J.; Mohamed, S.; and Welling, M. 2014. Semi-supervised learning with deep generative models. In *Proc. of NIPS*.

Klys, J.; Snell, J.; and Zemel, R. S. 2018. Learning latent subspaces in variational autoencoders. In *Proc of NIPS*.

Kulkarni, T. D.; Whitney, W. F.; Kohli, P.; and Tenenbaum, J. 2015. Deep convolutional inverse graphics network. In *Proc of NIPS*.

Kumar, A.; Sattigeri, P.; and Balakrishnan, A. 2018. Variational inference of disentangled latent concepts from unlabeled observations. In *Proc. of ICLR*.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–2324.

Liu, Y.; Wei, F.; Shao, J.; Sheng, L.; Yan, J.; and Wang, X. 2018. Exploring disentangled feature representation beyond face identification. In *Proc. of CVPR*.

Locatello, F.; Bauer, S.; Lučić, M.; Rätsch, G.; Gelly, S.; Schölkopf, B.; and Bachem, O. F. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proc. of ICML*.

Maaløe, L.; Fraccaro, M.; Liévin, V.; and Winther, O. 2019. Biva: A very deep hierarchy of latent variables for generative modeling. In *Proc. of NIPS*.

Mathieu, M. F.; Zhao, J. J.; Zhao, J.; Ramesh, A.; Sprechmann, P.; and LeCun, Y. 2016. Disentangling factors of variation in deep representation using adversarial training. In *Proc. of NIPS*.

Moyer, D.; Gao, S.; Brekelmans, R.; Galstyan, A.; and Steeg, G. V. 2018. Invariant representations without adversarial training. In *Proc of NIPS*.

Nguyen, X.; Wainwright, M. J.; and Jordan, M. I. 2010. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory* 56(11):5847–5861.

Ranganath, R.; Tran, D.; and Blei, D. 2016. Hierarchical variational models. In *Proc. of ICML*.

Siddharth, N.; Paige, B.; van de Meent, J.-W.; Desmaison, A.; Goodman, N.; Kohli, P.; Wood, F.; and Torr, P. 2017. Learning disentangled representations with semi-supervised deep generative models. In *Proc. of NIPS*.

Tschannen, M.; Bachem, O.; and Lucic, M. 2018. Recent advances in autoencoder-based representation learning. *arXiv e-prints* 1812.05069v1.

Wu, W.; Cao, K.; Li, C.; Qian, C.; and Loy, C. C. 2019. Disentangling content and style via unsupervised geometry distillation. In *Proc. of ICLR Workshop*.

Xiang, S., and Li, H. 2019. Disentangling style and content in anime illustrations. *arXiv e-prints* 1905.10742.

Yang, J.; Reed, S.; Yang, M.-H.; and Lee, H. 2015. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In *Proc. of NIPS*.

Yin, X.; Yu, X.; Sohn, K.; Liu, X.; and Chandraker, M. 2017. Towards large-pose face frontalization in the wild. In *Proc. of ICCV*.

Zhang, C.; Butepage, J.; Kjellstrom, H.; and Mandt, S. 2019. Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41(08):2008–2026.