Learning Disentangled Representations via Mutual Information Estimation

Eduardo Hugo Sanchez[1,2], Mathieu Serrurier[1,2], and Mathias Ortner[3]

[1] IRT Saint Exupéry, Toulouse, France
{eduardo.sanchez, mathieu.serrurier}@irt-saintexupery.com
[2] IRIT, Université Toulouse III - Paul Sabatier, Toulouse, France
[3] Airbus, Toulouse, France
mathias.ortner@airbus.com

Abstract. In this paper, we investigate the problem of learning disentangled representations. Given a pair of images sharing some attributes, we aim to create a low-dimensional representation which is split into two parts: a shared representation that captures the common information between the images and an exclusive representation that contains the specific information of each image. To address this issue, we propose a model based on mutual information estimation without relying on image reconstruction or image generation. Mutual information maximization is performed to capture the attributes of data in the shared and exclusive representations while we minimize the mutual information between the shared and exclusive representation to enforce representation disentanglement. We show that these representations are useful to perform downstream tasks such as image classification and image retrieval based on the shared or exclusive component. Moreover, classification results show that our model outperforms the state-of-the-art models based on VAE/GAN approaches in representation disentanglement.

Keywords: Representation learning, representation disentanglement, mutual information maximization and minimization

1 Introduction

Deep learning success involves supervised learning where massive amounts of labeled data are used to learn useful representations from raw data. As labeled data is not always accessible, unsupervised learning algorithms have been proposed to learn useful data representations easily transferable for downstream tasks. A desirable property of these algorithms is to perform dimensionality reduction while keeping the most important attributes of data. For instance, methods based on deep neural networks have been proposed using autoencoder approaches [15, 20, 21] or generative models [1, 8, 12, 22, 25, 30]. Nevertheless, learning high-dimensional data can be challenging. Autoencoders present difficulties to deal with multimodal data distributions and generative models rely on computationally demanding models [11, 19, 29] which are particularly complicated to train.

Recent work has focused on mutual information estimation and maximization to perform representation learning [2, 16, 27, 28]. As mutual information maximization is shown to be effective to capture the salient attributes of data, another desirable property is to be able to disentangle these attributes. For instance, it could be useful to remove some attributes of data that are not relevant for a given task such as illumination conditions in object recognition.

In particular, we are interested in learning representations of data that shares some attributes. Learning a representation that separates the common data attributes from the remaining data attributes could be useful in multiple situations. For example, capturing the common information from multiple face images could be advantageous to perform pose-invariant face recognition [33]. Similarly, learning representations containing the common information across satellite image time series is useful for image classification and segmentation [32].

In this paper, we propose a method to learn disentangled representations based on mutual information estimation. Given an image pair (typically from different domains), we aim to disentangle the representation of these images into two parts: a shared representation that captures the common information between images and an exclusive representation that contains the specific information of each image. An example is shown in Figure 1. To capture the common information, we propose a novel method called *crossed mutual information estimation and maximization*. Additionally, we propose an adversarial objective to minimize the mutual information between the shared and exclusive representations in order to achieve representation disentanglement. The following contributions are made in this work:

- Based on mutual information estimation (see Section 3), we propose a method to learn disentangled representations without relying on more costly image reconstruction or image generation models.
- In Section 4, we present a novel training procedure which is divided into two stages. First, the shared representation is learned via *crossed mutual information estimation and maximization*. Secondly, mutual information maximization is performed to learn the exclusive representation while minimizing the mutual information between the shared and exclusive representations. We introduce an adversarial objective to minimize the mutual information as the method based on statistics networks described in Section 3 is not suitable for this purpose.
- In Section 5, we perform several experiments on two synthetic datasets: a) colored-MNIST [23]; b) 3D Shapes [5] and two real datasets: c) IAM Handwriting [26]; d) Sentinel-2 [9]. We show that the obtained representations are useful at image classification and image retrieval outperforming the state-of-the-art models based on VAE/GAN approaches in representation disentanglement. We perform an ablation study to analyze the components of our model. We also show the effectiveness of the proposed adversarial objective in representation disentanglement via a sensitivity analysis. In Section 6, we show the conclusions of our work.

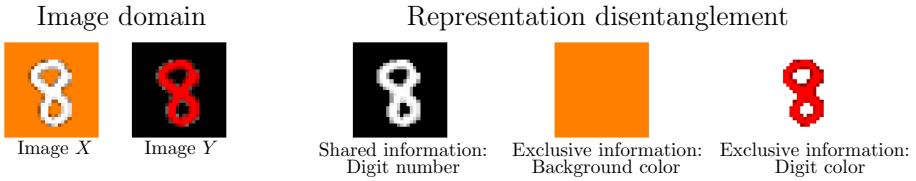


Fig. 1: Representation disentanglement example. Given images X and Y on the left, our model aims to learn a representation space where the image information is split into the shared information (digit number) and the exclusive information (background/digit color) on the right.

2 Related work

Generative adversarial networks (GANs) The GAN model [12, 13] can be thought of as an adversarial game between two players: the generator and the discriminator. In this setting, the generator aims to produce samples that look like drawn from the data distribution $\mathbb{P}_{data}$ while the discriminator receives samples from the generator and the dataset to determine their source (dataset samples from $\mathbb{P}_{data}$ or generated samples from $\mathbb{P}_{gen}$). The generator is trained to fool the discriminator by learning a distribution $\mathbb{P}_{gen}$ that converges to $\mathbb{P}_{data}$.

Mutual information Recent work has focused on mutual information estimation and maximization as a means to perform representation learning. Since the mutual information is notoriously hard to compute for high-dimensional variables, some estimators based on deep neural networks have been proposed. Belghazi et al. [2] propose a mutual information estimator which is based on the Donsker-Varadhan representation of the Kullback-Leibler divergence. Instead, Hjelm et al. [16] propose an objective function based on the Jensen-Shannon divergence called Deep InfoMax. Similarly, Ozair et al. [28] use the Wasserstein divergence. Mutual information maximization based methods learn representations without training decoder functions that go back into the image domain which is the prevalent paradigm in representation learning.

Representation disentanglement Disentangling data attributes can be useful for several tasks that require knowledge of these attributes. Creating representations where each dimension is independent and corresponds to a particular attribute have been proposed using VAE variants [15, 20] and GAN-based models [7]. Another definition of disentangled representation is presented by image-to-image translation models [34, 6, 24, 17, 18, 31, 3, 11] where the goal is to separate the content and style of images. For instance, consider a collection of data grouped by a shared attribute (e.g. face images grouped by identity). These disentanglement models aim to create a representation domain that captures the shared information (e.g. identity) and the exclusive information (e.g. pose) sep-

arately. In contrast to models requiring some supervision to perform disentanglement [34, 24, 17], weakly-supervised learning models have been developed to reduce label cost. In order to perform content and style disentanglement, Jha et al. [18] use a cycle-consistency constraint combined with the VAE framework [21] and Bouchacourt et al. [3] extend the VAE framework for grouped observations. More related to our work, Gonzalez-Garcia et al. [11] have recently proposed a model based on VAE-GAN image translators, cross-domain autoencoders and gradient reversal layers [10] to disentangle the attributes of paired images into shared and exclusive representations. A similar approach is proposed by Sanchez et al. [32] to separate the spatial and temporal information of image time series.

In this work, we aim to learn disentangled representations of paired data by splitting the representation into a shared part and an exclusive part. We propose a model based on mutual information estimation to perform representation learning using the method of Hjelm et al. [16] instead of generative or autoencoding models. Additionally, we introduce an adversarial objective [12] to disentangle the information contained in the shared and exclusive representations which is more effective than the gradient reversal layers [10]. We compare our model with the models proposed by Jha et al. [18] and Gonzalez-Garcia et al. [11] on the synthetic datasets and the generative models [21, 32] on the real datasets. We show that we achieve better results in representation disentanglement.

3 Mutual information

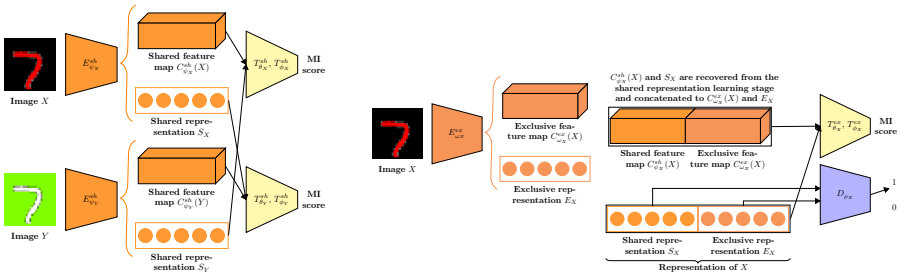
Let $X \in \mathcal{X}$ and $Z \in \mathcal{Z}$ be two random variables. Assuming that $p(x, z)$ is the joint probability density function of X and Z and that $p(x)$ and $p(z)$ are the corresponding marginal probability density functions, the mutual information between X and Z can be expressed as follows

$$I(X, Z) = \int_{\mathcal{X}} \int_{\mathcal{Z}} p(x, z) \log \left(\frac{p(x, z)}{p(x)p(z)} \right) dxdz \tag{1}$$

From Equation 1, the mutual information $I(X, Z)$ can be written as the Kullback-Leibler divergence between the joint probability distribution $\mathbb{P}_{XZ}$ and the product of the marginal distributions $\mathbb{P}_X\mathbb{P}_Z$, i.e. $I(X, Z) = \mathrm{D}_{KL}\left(\mathbb{P}_{XZ} \parallel \mathbb{P}_X\mathbb{P}_Z\right)$. In this work, we use the mutual information estimator Deep InfoMax [16] where the objective function is based on the Jensen-Shannon divergence instead, i.e. $I^{(\mathrm{JSD})}(X, Z) = \mathrm{D}_{JS}\left(\mathbb{P}_{XZ} \parallel \mathbb{P}_X\mathbb{P}_Z\right)$. We employ this method since it proves to be stable and we are not interested in the precise value of mutual information but in maximizing it. The estimator is shown in Equation 2 where $T_\theta : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ is a deep neural network of parameters θ called the *statistics network*.

$$\hat{I}^{(\mathrm{JSD})}_\theta(X, Z) = \mathbb{E}_{p(x,z)}\left[-\log\left(1+e^{-T_\theta(x,z)}\right)\right] - \mathbb{E}_{p(x)p(z)}\left[\log\left(1+e^{T_\theta(x,z)}\right)\right] \tag{2}$$

Hjelm et al. [16] propose an objective function based on the estimation and maximization of the mutual information between an image $X \in \mathcal{X}$ and its feature



(a) (b)

Fig. 2: Model overview. a) First, the shared representation is learned. Images X and Y are passed through the shared representation encoders to extract the representations S_X and S_Y. The statistics networks maximize the mutual information between the image X and the representation S_Y (and between Y and S_X); b) Then, the exclusive representation is learned. The image X is passed through the exclusive representation encoder to obtain the representation E_X. The statistics networks maximize the mutual information between the image X and its representation $R_X = [S_X, E_X]$ while the discriminator minimize the mutual information between representations S_X and E_X. The same operation is performed to learn E_Y. Best viewed in color and zoom-in.

representation $Z \in \mathcal{Z}$ which is called *global mutual information*. The feature representation Z is extracted by a deep neural network of parameters ψ, $E_\psi : \mathcal{X} \to \mathcal{Z}$. Equation 3 displays the global mutual information objective.

$$\mathbf{L}^{\text{global}}_{\theta,\psi}(X, Z) = \hat{I}^{(\text{JSD})}_{\theta}(X, Z) \tag{3}$$

Additionally, Hjelm et al. [16] propose to maximize the mutual information between local patches of the image X represented by a feature map $C_\psi(X)$ of the encoder $E_\psi = f_\psi \circ C_\psi$ and the feature representation Z which is called *local mutual information*. Equation 4 shows the local mutual information objective.

$$\mathbf{L}^{\text{local}}_{\phi,\psi}(X, Z) = \sum_i \hat{I}^{(\text{JSD})}_{\phi}(C^{(i)}_{\psi}(X), Z) \tag{4}$$

4 Method

Let X and Y be two images belonging to the domains $\mathcal{X}$ and $\mathcal{Y}$ respectively. Let $R_X \in \mathcal{R}_\mathcal{X}$ and $R_Y \in \mathcal{R}_\mathcal{Y}$ be the corresponding representations for each image. The representation is split into two parts: the shared representations S_X and S_Y which contain the common information between the images X and Y and the exclusive representations E_X and E_Y which contain the specific information of each image. Therefore the representation of image X can be written as $R_X{=}[S_X, E_X]$. Similarly, we can write $R_Y{=}[S_Y, E_Y]$ for image Y. For instance, let us consider

the images shown in Figure 1. In this case, the shared representations S_X and S_Y contain the digit number information while the exclusive representations E_X and E_Y correspond to the background and digit color information.

To address this representation disentanglement, we propose a training procedure which is split into two stages. We think that a natural way to learn these disentangled representations can be done via an incremental approach. The first stage learns the common information between images and creates a shared representation (see Section 4.1). Knowing the common information, it is easy then to identify the specific information of each image. Therefore, using this learned shared representation, a second stage is performed to learn the exclusive representation (see Section 4.2) which captures the remaining information that is missing in the shared representation. The model overview is shown in Figure 2.

4.1 Shared representation learning

Let $E^{\text{sh}}_{\psi_X} : \mathcal{X} \to \mathcal{S}_\mathcal{X}$ and $E^{\text{sh}}_{\psi_Y} : \mathcal{Y} \to \mathcal{S}_\mathcal{Y}$ be the encoder functions to extract the shared representations S_X and S_Y from images X and Y, respectively. We estimate and maximize the mutual information between the images and their shared representations via Equations 3 and 4 using the global statistics networks $T^{\text{sh}}_{\theta_X}$ and $T^{\text{sh}}_{\theta_Y}$ and the local statistics networks $T^{\text{sh}}_{\phi_X}$ and $T^{\text{sh}}_{\phi_Y}$. In constrast to Deep InfoMax [16], to enforce to learn only the common information between images X and Y, we swap the shared representations to compute the *crossed mutual information* as shown in Equation 5 where global and local mutual information terms are weighted by constant coefficients α^{sh} and β^{sh}. Swapping the shared representations is a key element of the proposed method as it enforces to remove the exclusive information of each image (see Section 5.3).

$$\begin{aligned}\mathbf{L}^{\text{sh}}_{MI} = \alpha^{\text{sh}}(&\mathbf{L}^{\text{global}}_{\theta_X,\psi_Y}(X, S_Y) + \mathbf{L}^{\text{global}}_{\theta_Y,\psi_X}(Y, S_X)) \\ &+ \beta^{\text{sh}}\left(\mathbf{L}^{\text{local}}_{\phi_X,\psi_Y}(X, S_Y) + \mathbf{L}^{\text{local}}_{\phi_Y,\psi_X}(Y, S_X)\right)\end{aligned} \tag{5}$$

Additionally, images X and Y must have identical shared representations, i.e. $S_X = S_Y$. A simple solution is to minimize the L_1 distance between their shared representations as follows

$$\mathbf{L}_1 = \mathbb{E}_{p(s_x,s_y)}\left[|S_X - S_Y|\right] \tag{6}$$

The objective function to learn the shared representations is a linear combination of the previous terms as shown in Equation 7, where γ is a constant coefficient.

$$\max_{\{\psi,\theta,\phi\}_{X,Y}} \mathcal{L}^{\text{shared}} = \mathbf{L}^{\text{sh}}_{MI} - \gamma\mathbf{L}_1 \tag{7}$$

4.2 Exclusive representation learning

So far, our model is able to extract the shared representations S_X and S_Y. Let $E^{\text{ex}}_{\omega_X} : \mathcal{X} \to \mathcal{E}_\mathcal{X}$ and $E^{\text{ex}}_{\omega_Y} : \mathcal{Y} \to \mathcal{E}_\mathcal{Y}$ be the encoder functions to extract the exclusive representations E_X and E_Y from images X and Y, respectively. To

learn these representations, we estimate and maximize the mutual information between the image X and its corresponding representation R_X which is composed of the shared and exclusive representations i.e. R_X=$[S_X, E_X]$. The same operation is performed between the image Y and R_Y=$[S_Y, E_Y]$ as shown in Equation 8 where α^{ex} and β^{ex} are constant coefficients. Mutual information is computed by the global statistics networks $T^{\text{ex}}_{\theta_X}$ and $T^{\text{ex}}_{\theta_Y}$ and the local statistics networks $T^{\text{ex}}_{\phi_X}$ and $T^{\text{ex}}_{\phi_Y}$. Since the shared representation remains fixed, we enforce the exclusive representation to include the information which is specific to the image and is not captured by the shared representation.

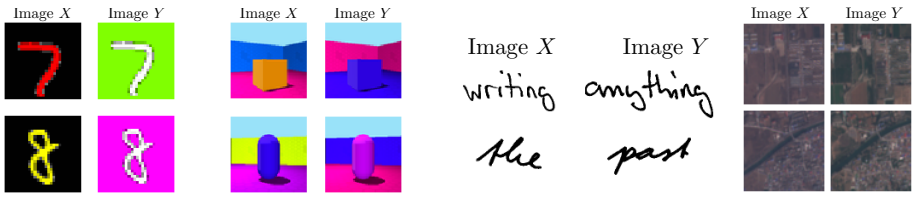
$$\begin{aligned}\mathbf{L}^{\text{ex}}_{MI} = \alpha^{\text{ex}}(&\mathbf{L}^{\text{global}}_{\theta_X,\omega_X}(X, R_X)+\mathbf{L}^{\text{global}}_{\theta_Y,\omega_Y}(Y, R_Y)) \\ &+ \beta^{\text{ex}}\left(\mathbf{L}^{\text{local}}_{\phi_X,\omega_X}(X, R_X)+\mathbf{L}^{\text{local}}_{\phi_Y,\omega_Y}(Y, R_Y)\right)\end{aligned} \tag{8}$$

On the other hand, the representation E_X must not contain information captured by the representation S_X when maximizing the mutual information between X and R_X. Therefore, the mutual information between E_X and S_X must be minimized. While mutual information estimation and maximization via Equation 2 works well, using statistics networks fails to converge when performing mutual information estimation and minimization. It is straightforward to see that minimizing Equation 2 makes the statistics networks diverge. Therefore, we propose to minimize the mutual information between S_X and E_X (i.e. $I(S_X, E_X)$) via a different implementation of Equation 2 using an adversarial objective [12] as shown in Equation 9. Minimizing $I(S_X, E_X)$ is equivalent to minimizing $\mathrm{D}_{JS}\left(\mathbb{P}_{S_X E_X} \parallel \mathbb{P}_{S_X}\mathbb{P}_{E_X}\right)$ which can be achieved in an adversarial manner. Therefore, a discriminator D_{ρ_X} defined by a neural network of parameters ρ_X is trained to classify representations drawn from $\mathbb{P}_{S_X E_X}$ as fake samples and representations drawn from $\mathbb{P}_{S_X}\mathbb{P}_{E_X}$ as real samples. Samples from $\mathbb{P}_{S_X E_X}$ are obtained by passing the image X through the encoders $E^{\text{sh}}_{\psi_X}$ and $E^{\text{ex}}_{\omega_X}$ to extract (S_X, E_X). Samples from $\mathbb{P}_{S_X}\mathbb{P}_{E_X}$ are obtained by shuffling the exclusive representations of a batch of samples from $\mathbb{P}_{S_X E_X}$. The encoder function $E^{\text{ex}}_{\omega_X}$ strives to generate exclusive representations E_X that combined with S_X look like drawn from $\mathbb{P}_{S_X}\mathbb{P}_{E_X}$. By minimizing Equation 9, we minimize the Jensen-Shannon divergence $\mathrm{D}_{JS}\left(\mathbb{P}_{S_X E_X} \parallel \mathbb{P}_{S_X}\mathbb{P}_{E_X}\right)$ and thus the mutual information between E_X and S_X is minimized. A similar procedure to generate samples of the product of the marginal distributions from samples of the joint probability distribution is proposed in [4, 20]. In these models, an adversarial objective is used to make each dimension independent of the remaining dimensions of the representation. Instead, we use an adversarial objective to make the dimensions of the shared part independent of the dimensions of the exclusive part.

$$\mathbf{L}^{X}_{\text{adv}} = \mathbb{E}_{p(s_x)p(e_x)}\left[\log D_{\rho_X}(S_X, E_X)\right] + \mathbb{E}_{p(s_x,e_x)}\left[\log\left(1 - D_{\rho_X}(S_X, E_X)\right)\right] \tag{9}$$

Equation 10 shows the objective function to learn the exclusive representation which is a linear combination of the previous terms where λ_{adv} is a constant coefficient.

$$\max_{\{\omega,\theta,\phi\}_{X,Y}} \min_{\{\rho\}_{X,Y}} \mathcal{L}^{\text{ex}} = \mathbf{L}^{\text{ex}}_{MI} - \lambda_{\text{adv}}(\mathbf{L}^{X}_{\text{adv}} + \mathbf{L}^{Y}_{\text{adv}}) \tag{10}$$



(a) (b) (c) (d)

Fig. 3: Image pair samples (best viewed in color). (a) Colored-MNIST; (b) 3D Shapes; (c) IAM; (d) Sentinel-2.

4.3 Implementation details

Concerning the model architecture, we use DCGAN-like encoders [30], statistics networks used by Deep InfoMax [16] and a discriminator defined by a fully-connected network with 3 layers. Every network is trained from scratch using batches of 64 image pairs. We use Adam optimizer with a learning rate value of 0.0001. Concerning the loss coefficients, we use $\alpha^{\text{sh}} = \alpha^{\text{ex}} = 0.5$, $\beta^{\text{sh}} = \beta^{\text{ex}} = 1.0$, $\gamma = 0.1$. The coefficient λ_{adv} is analyzed in Section 5.3. The training algorithm is executed on a NVIDIA Tesla P100. More details about the architecture, hyper-parameters and optimizer are provided in the supplementary material section.

5 Experiments

5.1 Datasets

We perform representation disentanglement on the following datasets: a) **Colored-MNIST**: Similarly to Gonzalez-Garcia [11], we use a colored version of the MNIST dataset [23]. The colored background MNIST dataset (MNIST-CB) is generated by modifying the color of the background and the colored digit MNIST dataset (MNIST-CD) is generated by modifying the digit color. The background/digit color is randomly selected from a set of 12 colors. Two images with the same digit are sampled from MNIST-CB and MNIST-CD to create an image pair; b) **3D Shapes**: The 3D Shapes dataset [5] is composed of 480000 images of 64×64×3 pixels. Each image corresponds to a 3D object in a room with six factors of variation: floor color, wall color, object color, object scale, object shape and scene orientation. These factors of variation have 10, 10, 10, 8, 4 and 15 possible values respectively. We create a new dataset which consists of image pairs where the object scale, object shape and scene orientation are the same for both images while the floor color, wall color and object color are randomly selected; c) **IAM**: The IAM dataset [26] is composed of forms of handwritten English text. Words contained in the forms are isolated and labeled which can be used to train models to perform handwritten text recognition or writer identification. To train our model we select a subset of 6711 images of 64×256×1 pixels corresponding to the top 50 writers. Our dataset is composed

of image pairs where both images correspond to words written by the same person; d) **Sentinel-2**: Similarly to [32], we create a dataset composed of optical images of size 64×64 from the Sentinel-2 mission [9]. A 100GB dataset is created by selecting several regions of interest on the Earth's surface. Image pairs are created by selecting images from the same region but acquired at different times. Further details about the dataset creation can be found on the supplementary material. Some dataset image examples are shown in Figure 3. For all the datasets, we train our model to learn a shared representation of size 64. An exclusive representation of size 8, 64 and 64 is respectively learned for the colored-MNIST, 3D Shapes and IAM datasets. During training, when data comes from a single domain the number of networks involved can be halved by sharing weights (i.e. $\psi_X{=}\psi_Y$, $\theta_X{=}\theta_Y$, etc). For example, the reported results for the 3D Shapes, Sentinel-2 and IAM datasets are obtained using 3 networks (shared representation encoder, global and local statistics networks) to learn the shared representation and 4 networks (discriminator, exclusive representation encoder, global and local statistics networks) to learn the exclusive representation.

5.2 Representation disentanglement evaluation

To evaluate the learned representations, we perform several classification experiments. A classifier trained on the shared representation should be good for classifying the shared attributes of the image as the shared representation only contains the common information while it should achieve a performance close to random for classifying the exclusive attributes of the image. An analogous case occurs when performing classification using the exclusive representation. We use a simple architecture composed of 2 hidden fully-connected layers of few neurons to implement the classifier (more details in the supplementary material).

In the colored-MNIST dataset case, a classifier trained on the shared representation must perform well at digit number classification while the accuracy must be close to 8.33% (random decision between 12 colors) at background/digit color classification since no exclusive information is included in the shared representation. Similarly, using the exclusive representations to train a classifier, we expect the classifier to predict correctly the background/digit color while achieving a digit number accuracy close to 10% (random decision between 10 digits) as the exclusive representations contains no digit number information. Results are shown in Tables 1 and 2. We note that the learned representations by our model achieve the expected behavior. The same experiment is performed using the learned representations from the 3D Shapes dataset. A classifier trained on the shared representation must correctly classify the object scale, object shape and scene orientation while the accuracy must be close to random for the floor, wall and object colors (10%, random decision between 10 colors). Differently, a classifier trained on the exclusive representation must correctly classify the floor, wall and object colors while it must achieve a performance close to random to classify the object scale (12.50%, random decision between 8 scales), object shape (25%, random decision between 4 shapes) and scene orientation

Table 1: Background color and digit number accuracy using the shared representation S_X and the exclusive representation E_X for classification.

Feature	Background color	Digit number	Distance to ideal
Ideal S_X	8.33%	100.00%	0.0000
S_X (ours)	**8.22%**	**94.48%**	**0.0563**
S_X ([11])	99.56%	95.42%	0.9581
S_X ([18])	97.45%	88.15%	1.0097
Ideal E_X	100.00%	10.00%	0.0000
E_X (ours)	**99.99%**	**13.20%**	**0.0321**
E_X ([11])	99.99%	71.63%	0.6164
E_X ([18])	95.83%	21.90%	0.1607

Table 2: Digit color and number accuracy using the shared representation S_Y and the exclusive representations E_Y for classification.

Feature	Digit color	Digit number	Distance to ideal
Ideal S_Y	8.33%	100.00%	0.0000
S_Y (ours)	**8.83%**	**94.27%**	**0.0623**
S_Y ([11])	29.81%	95.06%	0.2641
S_Y ([18])	8.62%	88.15%	0.1214
Ideal E_Y	100.00%	10.00%	0.0000
E_Y (ours)	**99.92%**	**13.75%**	**0.0383**
E_Y ([11])	99.83%	74.54%	0.6471
E_Y ([18])	8.46%	21.90%	1.0304

Table 3: Accuracy on the 3D Shapes factors using the disentangled representations S_X and E_X for classification.

Feature	Floor color	Wall color	Object color	Object scale	Object shape	Scene orientation	Distance to ideal
Ideal S_X	10.00%	10.00%	10.00%	100.00%	100.00%	100.00%	0.0000
S_X (ours)	**9.96%**	**10.08%**	**9.95%**	**99.99%**	**99.99%**	**99.99%**	**0.0020**
S_X ([11])	99.92%	99.81%	96.67%	99.99%	99.99%	99.99%	2.6643
S_X ([18])	95.80%	98.30%	93.07%	97.77%	99.78%	97.39%	2.6223
Ideal E_X	100.00%	100.00%	100.00%	12.50%	25.00%	6.66%	0.0000
E_X (ours)	**95.10%**	**99.79%**	**96.17%**	**17.25%**	**30.73%**	**6.79%**	**0.1955**
E_X ([11])	99.99%	99.99%	99.94%	99.06%	99.98%	99.81%	2.5477
E_X ([18])	99.43%	99.72%	99.28%	43.30%	63.65%	20.99%	0.8535

(6.66%, random decision between 15 orientations). Accuracy results using the shared and exclusive representations are shown in Table 3.

For the colored-MNIST and 3D Shapes datasets, we compare our representations to the representations obtained from the models proposed by Jha et al. [18] and Gonzalez-Garcia et al. [11] using their code. In their models, even though the exclusive factors at image generation are controlled by the exclusive representation, the classification experiment shows that representation disentanglement is not correctly performed as the shared representation contains exclusive information and vice versa. In all the cases, the representations of our model are much closer in terms of accuracy to the ideal disentangled representations than the representations from the models of [18, 11]. We compute the distance to the ideal representation as the L_1 distance between the accuracies on data attributes. As representations obtained from generative models are determined by an objective function defined in the image domain, disentanglement constraints are not explicitly defined in the representation domain. Therefore, representation disentanglement is deficiently achieved in generative models. Moreover, our model is

Table 4: Writer and word accuracy.

Feature	Writer	Word
Ideal feature S_X	100.00%	$\sim$ 1.00%
Ideal feature E_X	$\sim$ 2.00%	100.00%
Feature S_X (ours)	**61.64%**	**9.94%**
Feature E_X (ours)	**10.80%**	**20.88%**
Feature f_X ([21])	13.77%	20.30%

Table 5: Writer and word accuracy using N nearest neighbors.

Feature	Writer	Word
Feature S_X $(N = 1)$	62.65%	15.78%
Feature S_X $(N = 5)$	64.06%	12.96%
Feature E_X $(N = 1)$	19.68%	19.84%
Feature E_X $(N = 5)$	16.87%	19.69%

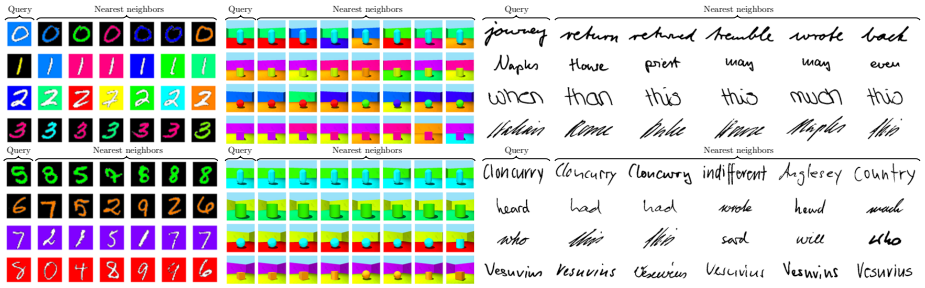


Fig. 4: Image retrieval on the colored-MNIST, 3D Shapes and IAM datasets (best viewed in color and zoom-in). Retrieved images using the shared representations (on the top) and the exclusive representations (on the bottom).

less computationally demanding as it does not require decoder functions to go back into the image domain. Training our model on the colored-MNIST dataset takes 20 min/epoch while the model of [11] takes 115 min/epoch. Additionally, our mutual information approach is more stable during training without requiring excessive hyperparameter tuning as models based on image generation.

For the IAM dataset, as the shared representation must capture the writer style, it must be useful to perform writer recognition while the exclusive representation must be useful to perform word classification. Accuracy results based on these representations can be seen in Table 4. Reasonable results are obtained at writer recognition while less satisfactory results are obtained at word classification as it is a more difficult task. To provide a comparison, we use the latent representation of size 128 learned by a VAE model [21] (as the models of [18, 11] fail to converge) to train a classifier for the mentioned classification tasks. Table 4 shows that the shared representation outperforms the VAE representation for writer recognition and the exclusive representation achieves a similar performance for word classification.

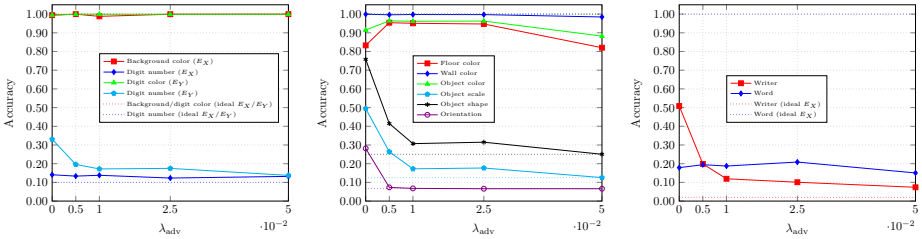
Additionally, we perform image retrieval experiments using the learned representations. In the colored-MNIST dataset, using the shared representation of a query image retrieves images containing the same digit independently of the background/digit color. In contrast, using the exclusive representation of a query image retrieves images corresponding to the same background/digit color independently of the digit number. A similar case occurs for the 3D Shapes dataset.

Table 6: MNIST ablation study. Accuracy using the representation S_X.

Feature	Background color	Digit number	Distance to ideal
Ideal S_X	8.33%	100.00%	0.0000
Baseline	**8.22%**	**94.48%**	**0.0563**
Non-SSR	99.99%	89.57%	1.0209
$\gamma = 0$	8.49%	92.36%	0.0780
$\alpha^{\text{sh}} = 0$	11.11%	94.83%	0.0795
$\beta^{\text{sh}} = 0$	8.51%	80.59%	0.1958

Table 7: IAM ablation study. Accuracy using the representation S_X.

Method	Word	Writer	Distance to ideal
Ideal S_X	$\sim 1.00\%$	100.00%	0.0000
Baseline	**9.94%**	**61.64%**	**0.4730**
Non-SSR	20.88%	58.94%	0.6094
$\gamma = 0$	10.51%	55.39%	0.5412
$\alpha^{\text{sh}} = 0$	11.36%	61.50%	0.4886
$\beta^{\text{sh}} = 0$	13.63%	50.28%	0.6235



(a) (b) (c)

Fig. 5: Different values of λ_{adv} are used to learn the exclusive representation. Results are plotted in terms of factor accuracy as a function of λ_{adv}. Solid curves correspond to the obtained values and dotted curves correspond to the expected behavior of an ideal exclusive representation (best viewed in color). (a) Colored-MNIST; (b) 3D Shapes; (c) IAM datasets.

In the IAM dataset, using the shared representations retrieves words written by the same person or similar style. While using the exclusive representation seems to retrieve images corresponding to the same word. Some image retrieval examples using the shared and exclusive representations are shown in Figure 4. As image retrieval is useful for clustering attributes, we also perform writer and word recognition on the IAM dataset using $N \in \{1, 5\}$ nearest neighbors based on the disentangled representations. We achieve similar results to those obtained using a neural network classifier as shown in Table 5.

5.3 Analysis of the objective function

Ablation study To evaluate the contribution of each element of the model during the shared representation learning, we remove it and observe the impact on the classification accuracy on the data attributes. As described in Section 4.3, our baseline setting is the following: $\alpha^{\text{sh}} = 0.5$, $\beta^{\text{sh}} = 1.0$, $\gamma = 0.1$ and swapped shared representations S_X/S_Y (SSR). We perform the ablation study and show the results for the colored-MNIST and IAM datasets in Tables 6 and 7. Swapping the shared representations plays a crucial role in representation

disentanglement avoiding these representations to capture exclusive information. When the shared representations are not swapped (non-SSR), the accuracy on exclusive attributes considerably increases meaning the presence of exclusive information in the shared representations. Removing the L_1 distance between S_X and S_Y ($\gamma = 0$) slightly reduces the accuracy on shared attributes. Removing the global mutual information term ($\alpha^{\text{sh}} = 0$) slightly increases the presence of exclusive information in the shared representation. Finally, using the local mutual information term is important to capture the shared information as the accuracy on shared attributes considerably decreases when setting $\beta^{\text{sh}} = 0$. Similar results are obtained by setting $\alpha^{\text{ex}} = 0$ or $\beta^{\text{ex}} = 0$ during the exclusive representation learning. In general, all loss terms lead to an improvement in representation disentanglement.

Sensitivity analysis As the parameter λ_{adv} weights the term that minimizes the mutual information between the shared and exclusive representations, we empirically investigate the impact of this parameter on the information captured by the exclusive representation. In order to train our model, we use different values of $\lambda_{\text{adv}} \in \{0.0, 0.005, 0.010, 0.025, 0.05\}$. Then, exclusive representations are used to perform classification on the attributes of data. Results in terms of accuracy as a function of λ_{adv} are shown in Figure 5. For $\lambda_{\text{adv}} = 0.0$ no representation disentanglement is performed, then the exclusive representation contains shared information and achieves a classification performance higher than random for the shared attributes of data. While increasing the value of λ_{adv} the exclusive representation behavior (solid curves) converges to the expected behavior (dotted curves). However, values higher than 0.025 decrease the performance classification on exclusive attributes of data.

5.4 Satellite applications

We show that our model is particularly useful when large amounts of unlabeled data are available and labels are scarce as in the case of satellite data. We train our model to learn the shared representations of our Sentinel-2 dataset which contains 100GB of unlabeled data. Then, a classifier is trained on the EuroSAT dataset [14] (27000 Sentinel-2 images of size 64×64 labeled in 10 classes) using the learned representations of our model as inputs. Using the shared representation makes the classifier robust to time-related conditions (seasonal changes, atmospheric conditions, etc.). We achieve an accuracy of 93.11% outperforming the performance obtained using the representations of the VAE model [21] (87.64%), the BicycleGAN model [35] (87.59%) and the VAE-GAN model proposed by Sanchez et al. [32] (92.38%).

As another interesting application, we found that Equation 5 could be used to measure the similarity between the center pixels of image patches X and Y in terms of mutual information. Some examples are shown in Figure 6. As can be seen, using this similarity measure we are able to distinguish the river, urban regions and agricultural areas. We think this could be useful for further applications such as unsupervised image segmentation and object detection.

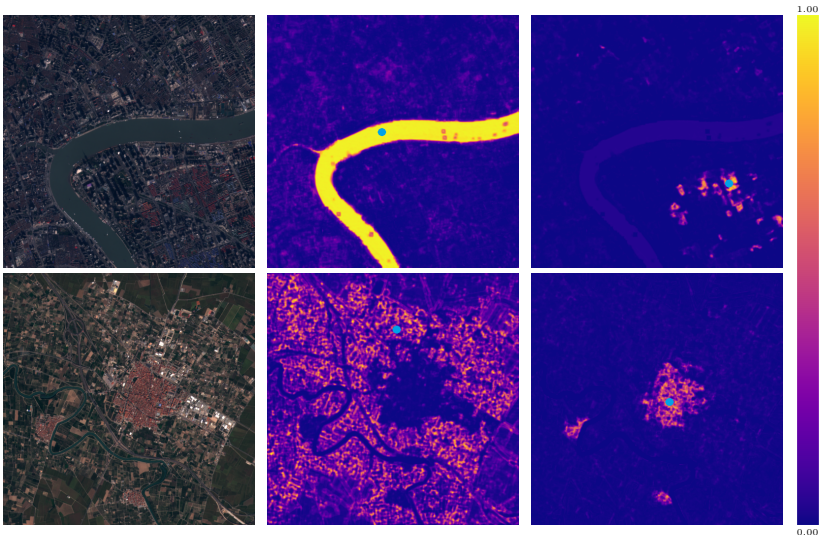


Fig. 6: Pixel similarity. The mutual information is computed between a given pixel (blue point) and the remaining image pixels via Equation 5.

6 Conclusions

We have proposed a novel method to perform representation disentanglement on paired images based on mutual information estimation using a two-stage training procedure. We have shown that our model is less computationally demanding and outperforms the state-of-the-art models [11, 18] to produce disentangled representations. We have performed an ablation study to demonstrate the usefulness of the key elements of our model (swapped shared representations, local and global statistics networks) and their impact on disentanglement. Additionally, we have empirically proven the disentangling capability of our model by analyzing the role of λ_{adv} during training. We have also demonstrated the benefits of our model on a challenging setting where large amounts of unlabeled paired data are available as in the Sentinel-2 case. We have shown that our model outperforms state-of-the-art models [21, 35, 32] relying on image reconstruction or image generation at image classification. We have also shown that the *crossed mutual information* objective could be useful for unsupervised image segmentation and object detection. Finally, we think that our model could be useful for image-to-image translation models to constrain the representations to separate content and style. We leave the development of such algorithm for future work.

Acknowledgments

We would like to thank the projects SYNAPSE and DEEL of the IRT Saint Exupéry for funding to conduct our experiments.

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning (2017)
2. Belghazi, M.I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., Hjelm, D.: Mutual information neural estimation. In: Proceedings of the 35th International Conference on Machine Learning (2018)
3. Bouchacourt, D., Tomioka, R., Nowozin, S.: Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
4. Brakel, P., Bengio, Y.: Learning independent features with adversarial nets for non-linear ica. arXiv preprint arXiv:1710.05050 (2017)
5. Burgess, C., Kim, H.: 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/ (2018)
6. Chen, M., Denoyer, L., Artières, T.: Multi-view data generation without view supervision (2018)
7. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: Advances in Neural Information Processing Systems (2016)
8. Donahue, J., Krähenbühl, P., Darrell, T.: Adversarial feature learning. In: International Conference on Learning Representations (2017)
9. Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., et al.: Sentinel-2: Esa's optical high-resolution mission for gmes operational services. Remote sensing of Environment **120**, 25–36 (2012)
10. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: Proceedings of the 32nd International Conference on Machine Learning (2015)
11. Gonzalez-Garcia, A., van de Weijer, J., Bengio, Y.: Image-to-image translation for cross-domain disentanglement. In: Advances in Neural Information Processing Systems (2018)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems (2014)
13. Goodfellow, I.J.: NIPS 2016 tutorial: Generative adversarial networks (2016)
14. Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. CoRR **abs/1709.00029** (2017), http://arxiv.org/abs/1709.00029
15. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. In: International Conference on Learning Representations (2017)
16. Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y.: Learning deep representations by mutual information estimation and maximization. In: International Conference on Learning Representations (2019)
17. Ilse, M., Tomczak, J.M., Louizos, C., Welling, M.: Diva: Domain invariant variational autoencoders. arXiv preprint arXiv:1905.10427 (2019)
18. Jha, A.H., Anand, S., Singh, M., Veeravasarapu, V.: Disentangling factors of variation with cycle-consistent variational auto-encoders. In: European Conference on Computer Vision (2018)

19. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
20. Kim, H., Mnih, A.: Disentangling by factorising. In: Proceedings of the 35th International Conference on Machine Learning (2018)
21. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: International Conference on Learning Representations (2014)
22. Larsen, A.B.L., SÃ¸nderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. In: Proceedings of The 33rd International Conference on Machine Learning (2016)
23. LeCun, Y., Cortes, C.: MNIST handwritten digit database (2010), http://yann.lecun.com/exdb/mnist/
24. Liu, Y.C., Yeh, Y.Y., Fu, T.C., Wang, S.D., Chiu, W.C., Frank Wang, Y.C.: Detach and adapt: Learning cross-domain disentangled deep representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
25. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Smolley, S.P.: Least squares generative adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV) (2017)
26. Marti, U.V., Bunke, H.: The IAM-database: an english sentence database for offline handwriting recognition. International Journal on Document Analysis and Recognition (2002)
27. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
28. Ozair, S., Lynch, C., Bengio, Y., Oord, A.v.d., Levine, S., Sermanet, P.: Wasserstein dependency measure for representation learning. arXiv preprint arXiv:1903.11780 (2019)
29. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
30. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: International Conference on Learning Representations (2016)
31. Sanakoyeu, A., Kotovenko, D., Lang, S., Ommer, B.: A style-aware content loss for real-time hd style transfer. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
32. Sanchez, E., Serrurier, M., Ortner, M.: Learning disentangled representations of satellite image time series. In: Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (2019)
33. Tran, L., Yin, X., Liu, X.: Disentangled representation learning gan for pose-invariant face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
34. Yang, J., Reed, S.E., Yang, M.H., Lee, H.: Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In: Advances in Neural Information Processing Systems (2015)
35. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. In: Advances in neural information processing systems (2017)