Learning Disentangled Representations via Mutual Information Estimation on the Shapes3D dataset

Lorenzo Busellato, VR472249

***Abstract*—In this work, the method presented by Sanchez et al. on learning disentangled representations using mutual information estimation is replicated, applying it to the Shapes3D dataset. The objective is to reproduce their methodology for the creation of low-dimensional representations of image pairs that present a number of shared features and a number of image-exclusive features. Following the proposed model, the mutual information between the images is maximized in order to capture the shared and exclusive representations of the images, while minimizing mutual information between them to ensure disentanglement. The approach is validated through experimentation on the Shapes3D dataset, aiming to achieve a performance comparable to the one reported by Sanchez et al, demonstrating the utility and effectiveness of the method.**

I. Introduction

In the field of machine learning, deep learning has made possible the extraction of meaningful representations of extensive collections of raw data. Traditional approaches in the field rely on supervised learning for the generation of the extensive labeled datasets required by the models. However, the process of data-labeling can be expensive and time-consuming, prompting the development of unsupervised learning algorithms that can autonomously handle the labeling of the raw data. One of the desirable properties of these approaches is the performing of dimensionality reduction, while preserving the most representative attributes of the data.

Among these techniques, mutual information estimation has emerged as a powerful tool for unsupervised representation learning. By maximizing mutual information, this method captures the salient attributes of the data. Furthermore, disentangling the learned representations into shared and exclusive components can be highly beneficial for a range of applications, such as image classification and retrieval.

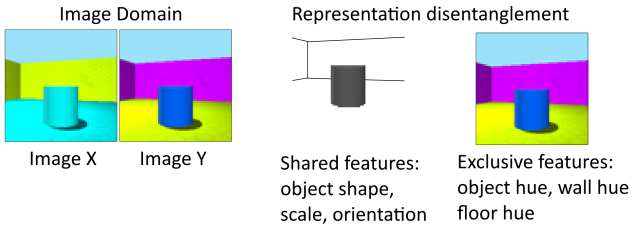

Fig. 1. Example of representation disentanglement.

In this work, the work of Sanchez *et al.* [1] on learning disentangled representations is replicated. The goal is to reproduce the methodology they proposed, which focuses on separating the shared attributes of image pairs from the exclusive attributes that characterize each image. Their approach uses a model based on mutual information estimation to achieve representation disentanglement without resorting to image reconstruction or generation.

A two-stage training procedure is employed to learn these disentangled representations. In the first stage, the common information between the images is captured to form a shared representation. In the second stage, the exclusive representation is learned by identifying the information specific to each image, ensuring disentanglement between the exclusive and shared components. The cross mutual information estimation and maximization method described in the paper is employed.

The method is validated on the Shapes3D dataset, aiming to replicate the findings of Sanchez *et al.*, demonstrating that the model effectively disentangles representations. By validating this approach, a confirmation can be given to the validity and effectiveness of their proposed method.

The code implementation is a modification of an available mutual information estimation for learning disentangled representations repository [2]. The adapted code is available in a public repository [3].

This document is structured as follows. The background knowledge on mutual information estimation and its use in the stated problem is given in section II. The implementation details of the code adapted to work with the Shapes3D dataset are illustrated in section III. The experimental setup is presented in section IV. The collected results and a discussion on them are given in section V. The conclusions for this work is given in section VI.

II. Background

A. Mutual Information Estimation

Mutual Information Estimation (MI) is a measure of the amount of information obtained about one random variable through the observation of another, thus capturing their cross-dependency. For random variables X and Z, MI is defined as:

$$\mathrm{MI}(X,Z) = \int\int p(x,z)\log\left(\frac{p(x,z)}{(p(x)p(z)}\right)dxdz \quad (1)$$

The estimation of MI for high-dimensional data is however computationally expensive. Therefore, neural network-based approaches such as Deep InfoMax (DIM) are employed. DIM uses a deep neural network to estimate and maximize MI, allowing for the handling of complex data distributions. The estimator is defined as:

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

where T_θ is a deep neural network with parameters θ called *statistics network*.

An objective function, that is based on the estimation and maximization of the mutual information between an image X and its representation Z, is defined as:

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

where ϕ are the parameters of the deep neural network that extracts Z from X and E_ϕ is the network's encoder.

Additionally, a local definition of the objective function can be defined:

$$L^{local}_{\theta,\phi}(X,Z) = \hat{I}^{(JSD)}_{\phi}(C_\varphi(X),Z) \tag{4}$$

where $C_\varphi(X)$ is a feature map obtained by applying the encoder to local image patches.

III. Method

Given two images X and Y, their representations are $R_X = (S_X, E_X)$ and $R_Y = (S_Y, E_Y)$, where S and E denote the shared and esclusive parts of the representations, respectively.

The first stage of the learning, described in section III-A, deals with constructing a shared representation. With this common description, it becomes easy to recover the exclusive representation of each image, process described in section III-B.

A. Shared representation learning

Let $E^{sh}_{\varphi_X}$ and $E^{sh}_{\varphi_Y}$ be the encoders that extract the shared representations S_X and S_Y, respectively. The goal is to maximize their mutual information.

The global and local mutual information terms are combined into a singular term, weighed by constants α^{sh} and β^{sh}:

$$\begin{aligned} L^{sh}_{MI} = \alpha^{sh}(L^{global}_{\theta_X,\varphi_Y}(X,S_Y) + L^{global}_{\theta_Y,\varphi_X}(Y,S_X)) \\ +\beta^{sh}(L^{local}_{\phi_X,\varphi_Y}(X,S_Y) + L^{local}_{\phi_Y,\varphi_X}(Y,S_X)) \end{aligned} \tag{5}$$

a key observation is that, compared to DIM, the representations appear into the function in a switched order. This is done in order to force the removal of the exclusive information of each image, leaving behind only the shared one. An additional constraint is given by the fact that S_X must be equal to S_Y, which can be achieved by minimizing their distance:

$$L_1 = \mathbb{E}_{p(s_x,s_y)}[\mid S_X - S_Y \mid] \tag{6}$$

Therefore, the loss function to be minimized to learn the shared representations becomes:

$$\max_{\{\varphi,\theta,\phi\}_{X,Y}} \mathcal{L}^{shared} = L^{sh}_{MI} - \gamma L_1 \tag{7}$$

where γ is a constant coefficient.

B. Exclusive representation learning

Let $E^{ex}_{\omega_X}$ and $E^{ex}_{\omega_Y}$ be the encoders that extract the exclusive representations E_X and E_Y. The goal is to maximize the mutual information between the images and their full representations R_X and R_Y.

In a similar fashion to the definition of the objective function for the shared representation learning step, in the exclusive representation learning the objective function is defined as:

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

To make sure that the exclusive representation E_X does not contain information already captured by the shared representation S_X, their respective mutual information must be minimized. This is done by introducing an adversarial objective function defined as follows:

$$\begin{aligned} L^{X}_{adv} = \mathbb{E}_{p(s_x)p(e_x)}[\log D_{\rho_x}(S_X,E_X)] \\ +\mathbb{E}_{p(s_x,e_x)}[\log(1 - D_{\rho_x}(S_X,E_X))] \end{aligned} \tag{9}$$

where D_{ρ_x} is a neural network that acts as a discriminator, classifying samples from $\mathbb{P}_{S_X E_X}$ as fake samples, and samples from $\mathbb{P}_{S_X}\mathbb{P}_{E_X}$ as real samples. Samples from $\mathbb{P}_{S_X E_X}$ come from feeding X to the encoders $E^{sh}_{\varphi_X}$ and $E^{ex}_{\omega_X}$, while samples $\mathbb{P}_{S_X}\mathbb{P}_{E_X}$ are obtained by shuffling the exclusive representations of a batch of samples from $\mathbb{P}_{S_X E_X}$. By minimizing equation 9 the statistical divergence, i.e. loss of information, between the two distributions is minimized, and so is the mutual information between S_X and E_X.

The cumulative loss for the learning of exclusive features becomes:

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

where λ_{adv} is a constant coefficient.

IV. Experimental validation

A. Dataset

The 3D shapes dataset [4] contains 480000 images of 64x64 pixel and 3 color channels. Each image represents a 3D object placed in a room. The features are the object's scale (8 linearly spaced values in the range $[0.75, 1.25]$), orientation (15 linearly spaced values in the range $[-30, 30]$) and shape (0 through 3 for, respectively: box, sphere, cylinder, pill-shape), the object's hue (10 linearly spaced values in the range $[0, 1]$), the wall's hue (10 linearly spaced values in the range $[0, 1]$) and the floor's hue (10 linearly spaced values in the range $[0, 1]$). Each image in the dataset is procedurally generated by spanning the ranges of the aforementioned features. The object's shape, scale and orientation are treated as shared features, while the hues are treated as exclusive features. A new dataset is created by generating image pairs from the original dataset, that have the same shared features but different exclusive features. The new dataset consists of 24000 unique image pairs.

B. Network architectures

In this section the network architectures that implement the previously described method are defined.

TABLE I
ENCODER ARCHITECTURE

ID	Layer	Kernel	Stride	Activation	Normalization
Input0	Input	-	-	None	None
Conv0	CNN	4 x 4	1 x 1	LeakyReLU	None
Conv1	CNN	4 x 4	2 x 2	LeakyReLU	BatchNorm
Conv2	CNN	4 x 4	2 x 2	LeakyReLU	BatchNorm
Conv3	CNN	4 x 4	2 x 2	LeakyReLU	BatchNorm
Flat0	Flatten	-	-	None	None
Output0	Dense	-	-	None	None

Table I reports the architecture for the shared and exclusive representation encoders, $E^{sh}_{\varphi_X}$ and $E^{ex}_{\omega_X}$. The input is a batch of image pairs, from which a vector of 64 elements is extracted, representing either the shared or exclusive representations.

TABLE II
GLOBAL STATISTICS NETWORK ARCHITECTURE

ID	Layer	Kernel	Stride	Activation	Normalization
GInput0	Input	-	-	None	None
GConv0	CNN	3 x 3	1 x 1	ReLU	None
GConv1	CNN	3 x 3	1 x 1	None	None
GFlat0	Flatten	-	-	None	None
GInput1	Input	-	-	None	None
GConcat0	Concatenate GInput0 + GInput1	-	-	None	None
GDense0	Dense	-	-	ReLU	None
GDense1	Dense	-	-	ReLU	None
GOutput0	Dense	-	-	None	None

Table II reports the architecture for the Global Statistics Network. GInput0 is the output of the Conv3 layer of the encoder, while GInput1 is the output of the Output0 layer. In the case of shared representation learning, the input is a $5\times5\times512$ tensor, while in the case of exclusive representation learning the input is a $5 \times 5 \times 1024$ tensor. From the input, a scalar value representing Global Mutual Information is extracted.

TABLE III
LOCAL STATISTICS NETWORK ARCHITECTURE

ID	Layer	Kernel	Stride	Activation	Normalization
LInput0	Input	-	-	None	None
LInput1	Input	-	-	None	None
LConcat0	Concatenate LInput0 + Tiled LInput1	-	-	None	None
LConv0	CNN	1 x 1	1 x 1	ReLU	None
LConv1	CNN	1 x 1	1 x 1	ReLU	None
LOutput0	CNN	1 x 1	1 x 1	None	None

Table III reports the architecture for the Local Statistics Network. The inputs are the same to the ones described for the Global Statistics Network. In this case however, the output of layer Output0 of the encoder, which is LInput1, needs to be tiled in order to be concatenated with LInput0, which does not get flattened. The output of the network is a $5\times5\times1$ map of Local Mutual Information.

Table IV reports the architecture for the discriminator. The inputs are the shared or exclusive representations produced by the encoders, while the output is the classification between real and fake data.

TABLE IV
DISCRIMINATOR NETWORK ARCHITECTURE

ID	Layer	Kernel	Stride	Activation	Normalization
DInput0	Input	-	-	None	None
DInput1	Input	-	-	None	None
DConcat0	Concatenate DInput0 + DInput1	-	-	None	None
DDense0	Dense	-	-	ReLU	None
DDense1	Dense	-	-	ReLU	None
DOutput0	Dense	-	-	None	None

TABLE V
CLASSIFIER NETWORK ARCHITECTURE

ID	Layer	Kernel	Stride	Activation	Normalization
CInput0	Input	-	-	None	None
CDense0	Dense	-	-	ReLu	BatchNorm
CDense0	Dense	-	-	ReLu	BatchNorm
COutput0	Dense	-	-	Softmax	None

Table V reports the architecture for the Classifier. It takes as input the 32-long vectors of feature representations, and passes them through a series of dense layers. Each layer has a number of neurons equal to the number of classification labels (4 for shape, 15 for orientation, 8 for scale, 10 for hue).

C. Shared representation learning

The encoders, global and local statistics networks, and classifier were trained using the Adam optimizer with learning rate $\lambda = 0.0001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ over 50000 iterations. The coefficients for the loss functions were $\alpha^{sh} = 0.5, \beta^{sh} = 1, \gamma = 0.1$. From the dataset, random batches of 64 image pairs were drawn at each iteration. The resulting shared representation had dimension 64.

D. Exclusive representation learning

The encoders, global and local statistics networks, and classifier were trained using the Adam optimizer with learning rate $\lambda = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ over 30000 iterations. The coefficients for the loss functions were $\alpha^{sh} = 0.5, \beta^{sh} = 1, \gamma = 0.01$. From the dataset, random batches of 64 image pairs were drawn at each iteration. The resulting exclusive representation had dimension 64.

V. RESULTS AND DISCUSSION

Table V presents the mean values for accuracy, computed on the recorded values after convergence. By looking at the learning curves reported in figure 2, convergence can be said to be reached after 20000 iterations. As such, accuracy was computed as the mean of the accuracy values of the last 10000 iterations. As expected, high accuracy is obtained for the shared/exclusive features when classified with classifiers trained on shared/exclusive representations, while a low, but coherent with a random choice, accuracy is obtained by using classifiers trained on exclusive/shared representations.

TABLE VI
ACCURACY OF THE CLASSIFIERS AT CONVERGENCE

		Floor color	Wall color	Object color	Object shape	Object scale	Scene orientation
S_x	Sanchez et al.	9.96%	10.08%	9.95%	99.99%	99.99%	99.99%
	Mine	10.28%	10.39%	10.27%	99.99%	99.44%	97.78%
	Difference	0.32%	0.31%	0.32%	-	0.55%	2.21%
E_x	Sanchez et al.	95.10%	99.79%	96.17%	30.73%	17.25%	6.79%
	Mine	77.74%	95.52%	81.19%	33.78%	16.98%	7.92%
	Difference	18.64%	4.27%	15.02%	3.05%	0.27%	1.27%

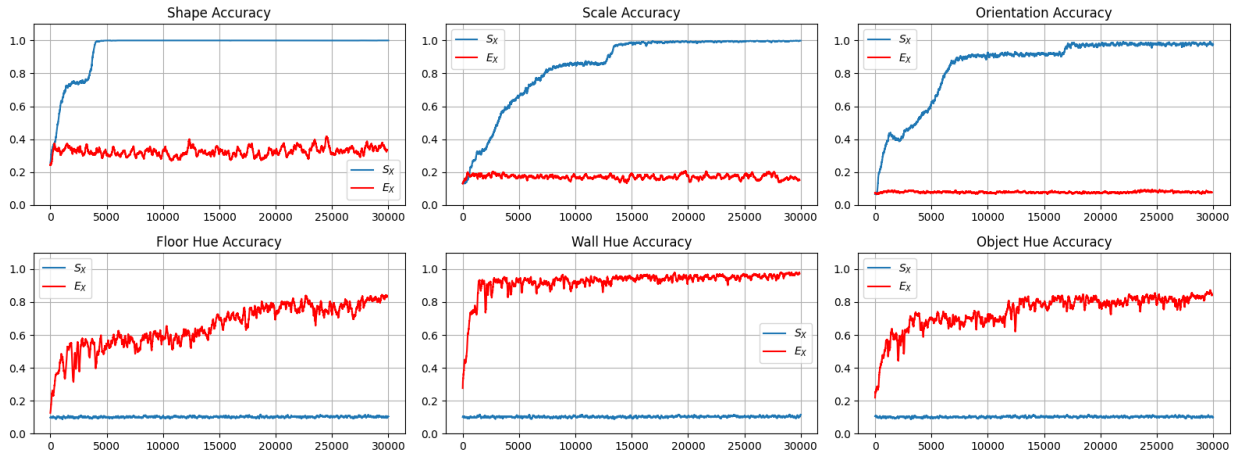

Fig. 2. Learning curves for the twelve classifiers. For comparison purposes, only the first 30000 iterations are shown for the shared feature classifiers.

Regarding classification with shared representations, high accuracy is achieved for the shared features. The accuracy for the shared features is very close to the one reported by Sanchez *et al.*

For classification with exclusive features, an accuracy consistent with the one reported by Sanchez *et al.* is found for the shared features (i.e. object shape and scale and scene orientation). A high discrepancy is found for the exclusive features, namely floor and object hue, while the wall hue's accuracy is closer to the reported values.

VI. CONCLUSIONS

This project almost successfully replicated the methodology proposed by Sanchez *et al.* for learning disentangled representations via mutual information estimation on the Shapes3D dataset. The approach effectively captures shared and exclusive features in image pairs, validating the utility and effectiveness of mutual information as a tool for unsupervised representation learning.

The experimental results demonstrated high accuracy for the shared/exclusive features when classified using the shared/exclusive representations, although some deviations from the original study were noted. These can be explained by the reduced training time and an excessive subsampling of the dataset. Furthermore, the classification of shared/exclusive features using the exclusive/shared features produced results consistent with random choice between the feature classes, confirming the effective disentanglement of image information into its shared and exclusive components.

REFERENCES

[1] E. H. Sanchez, M. Serrurier, and M. Ortner, "Learning disentangled representations via mutual information estimation," 2019. [Online]. Available: https://arxiv.org/abs/1912.03915
[2] M. Zouitine, "Learning-disentangled-representations-via-mutual-information-estimation," https://github.com/MehdiZouitine/Learning-Disentangled-Representations-via-Mutual-Information-Estimation.
[3] L. Busellato, "dl_project," https://github.com/lbusellato/dl_project.
[4] C. Burgess and H. Kim, "3d shapes dataset," https://github.com/deepmind/3dshapes-dataset/, 2018.