POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Forcing latent space Disentanglement for enhanced model Explainability

Laurea Magistrale in Computer Science Engineering - Ingegneria Informatica

Author: Federico Romeo

Advisor: Prof. Giacomo Boracchi

Co-advisor: Loris Giulivi

Academic year: 2022-2023

1. Introduction

Deep learning models have established their dominance in the computer vision field, being employed for a variety of tasks, including representation learning. Despite their widespread use, their intricate nature and extensive parameterization are making deep neural networks essentially black boxes, posing challenges when interpreting their outputs. Recent deep learning trends aim at developing more interpretable models by working on the intermediate latent representations formed during training.
To this end, enabling a direct link between the data semantic factors of variation (FoV) and the latent representation could be beneficial to interpretability. This is the case of **disentanglement**, a property of latent representations in which a change in one latent dimension corresponds to a change in one single FoV, while being relatively invariant to changes in others [1]. In a disentangled representation, individual latent dimensions correspond to distinct and interpretable FoVs, allowing for a clearer control over the latent space, and accordingly over the output. As an example (see Figure 1), suppose to have a simple dataset of white dots: to fully describe it we just require the dot's x and y coordinates, since no other variabilities exist. We denote them as the **factors of variation (FoV)** of the data. Thus we only need two independent and interpretable latent dimensions.

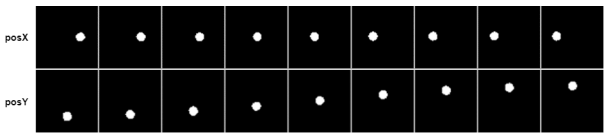

Figure 1: Example of a Latent Traversal on a disentangled representation. The two dimensions (rows) encode respectively the interpretable factors of variation $posX$ and $posY$.

Yet, achieving substantial disentanglement is a complex task, especially in unsupervised fashions where careful design choices and training strategies are required to effectively disentangle the data FoVs. With the increase of data complexity, models struggle to grasp the factors variability and to isolate them in dedicated latent dimensions. Past attempts at unsupervised learning of disentangled representations have shown promising results, but are mostly confined to biased toy datasets, missing real-world potential. For these reasons, when focusing on explainability of real-world datasets, unsupervised frameworks results being too weak, claiming the need of a driven supervision [6].

We propose SVAE, a supervised framework able to enhance the attribute-level disentanglement of a Beta Variational Autoencoder (β-VAE) [3] to develop interpretable models. Its applicability extends to real-world datasets, i.e. where the data generative process isn't trivially aligned to the data FoVs as in *DSprites* [7]. It requires a supervision made of paired images that share all but one FoV (see Figure 6); in this way the image pairs differ in a single aspect, or FoV, whose variance we force to be encoded in a dedicated dimension through a custom loss function.
After having assessed the soundness of our method on the benchmark dataset *DSprites*, to evaluate its effectiveness also on real-world datasets our contribution also includes the generation of the required paired supervision focusing on facial images. To this end we introduce a novel Semantic Facial Attribute Editing (SFAE) method able to perform fine-grained edits to modify single facial attributes, to generate image pairs with a single non-shared FoV (for example changing only the hair color).
Experiments on both datasets reached convincing results in term of attributes disentanglement.

2. Background & related works

Our work focuses on learning interpretable representations. Seeking latent space disentanglement can be of pivotal importance, where a one-to-one mapping connects semantic FoVs with latent dimensions. In this section we discuss previous works related to disentangled representations learning, categorizing them in:

- **unsupervised**: methods in which the latent space isn't constrained in any way;
- **supervised**: methods in which labeled observations guide different latent dimensions to encode specific data patterns.

In literature, many unsupervised methods tried to seek disentangled representations. Borrowing from the standard Variational Autoencoder (VAE), the ***β*VAE** [3] modified its objective by giving more weight to the Kullback-Leibler regularization term, adjusting it by a factor $\beta{>}1$:

$$\mathcal{L} = \underbrace{E_{q_\phi}(z|x)[\log p_\theta(x|z)]}_{\text{reconstruction loss}} - \beta \cdot \underbrace{D_{KL}(q_\phi(z|x)||p(z))}_{\text{KL divergence}}$$

where $q_\phi(z|x)$ and $p_\phi(x|z)$ respectively parametrize the encoder and decoder networks, being x the input and z the learnt latent vector. This led to an enhancement of the latent space disentanglement by stressing the independence among latent dimensions, leveraging the characteristics of the Isotropic Gaussian prior $p(z) \sim \mathcal{N}(0, I)$ which enforces zero correlations between dimensions due to its unit covariance matrix. **FactorVAE** [5] furtherly broke down the KL divergence term:

$$D_{KL}(q_\phi(z|x)||p(z)) = \underbrace{I(x;z)}_{\text{MI}} + \underbrace{D_{KL}(q(z)||p(z))}_{\text{prior KL divergence}}$$

where $I(x;z)$ represents the mutual information between x and z. Thus in βVAE increasing β was leading to better disentanglement but worsening the reconstruction quality, due to the penalization on the Mutual Information term.
The achievement of effective disentanglement in these unsupervised methods has to be attributed mainly to the exploitation of the inherent biases of the toy datasets they've been trained on. This is because those toy datasets have an underlying generative model which is precisely aligned with the underlying data FoVs, thus making the framework not generalizable. So as claimed in [6], it is essentially impossible for disentangled representation learning to capture the desired properties without exploiting inductive biases or adding an explicit supervision.
Some works employed supervised group-based disentanglement methods in which paired observations with a single shared FoV are shown at training time, helping the model to discern and separate FoVs in different latent dimensions given the pair's characteristics. Among these methods for group-disentangled representations learning, **MLVAE** [2] employs a product of approximate target posteriors, whereas **GVAE** [4] uses an empirical average of the parameters of the approximate target posteriors to disentangle the FoVs into separate latent dimensions.
Our method builds up from the βVAE objective function, with an additional loss term to enhance the latent space disentanglement starting from group-based observations. Differently from the supervised works mentioned above that employ paired observations that have one single *shared* FoV, we employ pairs with one single *non-shared* FoV as shown in Figure 6 and 8. Also, those methods works by fixing as FoVs just the content and the style of the images, while our framework can accomodate various and more detailed FoVs as described in Section 3.3.

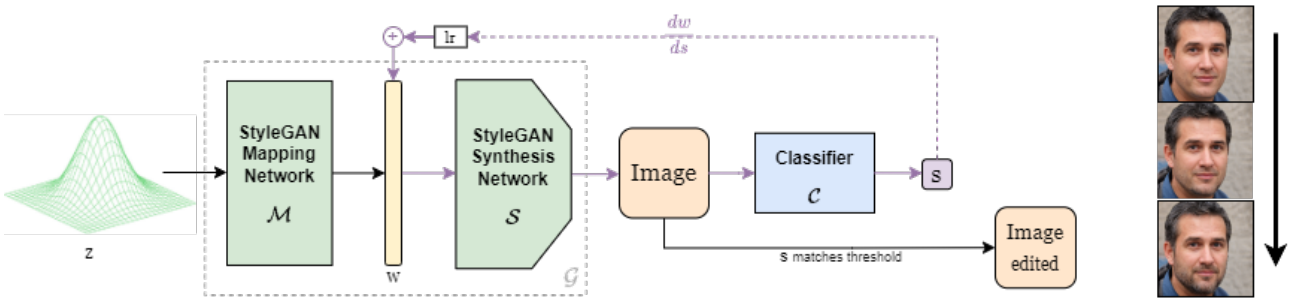

Figure 2: Architecture of our *E2Editor* framework to perform Semantic Facial Attribute Editing.

3. Methods

3.1. Problem Formulation

In this thesis, we tackle the problem of generating an interpretable model by leveraging on the disentanglement property of its latent space. The reference model for disentangled representation learning is the Beta Variational Autoencoder [3] described in Section 2, composed of:

- an encoder $q_\phi(z|x)$, that maps the input x to a latent gaussian distribution with mean and variance $\{\mu, \sigma\} = \{[\mu_1, \sigma_1], ..., [\mu_n, \sigma_n]\}$;
- a decoder $p_\phi(x|z)$ that reconstruct the input in $\tilde{x}$ after sampling the actual latent vector $z \sim \mathcal{N}(\mu, \sigma) = \{z_1, ..., z_n\}$;

being n the dimension of the latent space.
If we denote as $\mathcal{F} = \{f_1, ..., f_n\}$ the data factors of variations, $z = \{z_1, ..., z_n\}$ is said to be perfectly disentangled when there exists a one-to-one mapping between $\mathcal{F}$ and the latent vector z, i.e. where a change in each latent dimension z_i affects only a factor f_i. Since with the increase of data complexity unsupervised methods struggle to build efficient and non-redundant latent representation, our goal is to provide a supervision to enhance the model's disentanglement, by guiding it in encoding each semantic data FoV f_i in a dedicated dimension z_i of z.
In the following, we discuss at first a proposed method **E2Editor** to generate the pairwise required supervision in a selected real-world dataset, and then our **SVAE** framework to enhance the latent space disentanglement.

3.2. E2Editor

E2Editor is our novel technique for Semantic Facial Attribute Editing (SFAE), aimed at altering a single facial attribute while preserving others. It combines an Attribute Classifier $\mathcal{C}$, with a StyleGAN generator $\mathcal{G}$, which can be decomposed in a Mapping Network $\mathcal{M}$ and a Synthesis Network $\mathcal{S}$. Both networks are pretrained on *CelebA*. Our intuition is that we can enhance the presence of an attribute a (*e.g. Beard*) on a facial image by iteratively adjusting it by a step in the gradient direction of its misclassification error on a. The method's workflow (Figure 2) starts by sampling an initial base facial image:

$$I_{initial} = \mathcal{S}(w) = \mathcal{S}(\mathcal{M}(z)) \qquad \text{with } z = \mathcal{N}(0, 1)$$

storing the base vector w.
The core of the method lies in an iterative editing loop that halts when the classification score s_a of the target attribute a reaches a threshold t. At each iteration i, the vector w is adjusted by a step in the gradient direction $\nabla w = \frac{dw}{ds_a}$ weighted by a learning rate lr, thereby enhancing the presence of a:

$$w_i = w_{i-1} + lr \cdot \frac{\nabla w}{\|\nabla w\|}$$

This adjustment of w affects the re-generated image $I_{edited} = \mathcal{S}(w)$, whose attribute score s_a is re-evaluated by the classifier $\mathcal{C}$. To compute the gradient direction, the loss to be minimized is:

$$\mathcal{L} = \lambda \cdot \boldsymbol{\mathcal{L}_{tg}} + (\mathbf{1} - \lambda) \cdot \boldsymbol{\mathcal{L}_{oth}}$$

$$\begin{cases} \boldsymbol{\mathcal{L}tg} = -y_{tg} \log(p) - (1 - y_{tg}) \log(1 - p) \\ \boldsymbol{\mathcal{L}oth} = -y_{oth} \log(p) - (1 - y_{oth}) \log(1 - p) \end{cases}$$

in which both the losses defines a binary cross-entropy, whose goals are respectively bringing the edit in the direction of changing the target attribute a, and keeping the *others* non-target attributes unchanged, by maintaining them at their initial classification score. The parameter λ plays a crucial role: it manages the trade-off between the disentanglement precision of the

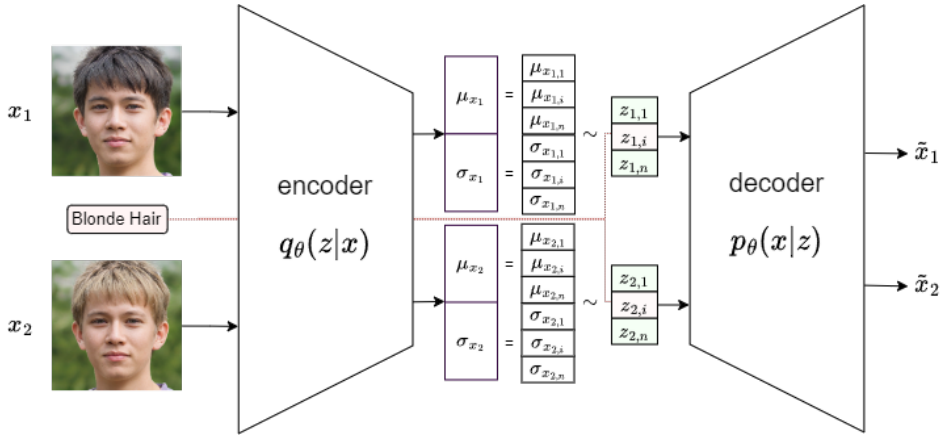

Figure 4: Our SVAE framework to enhance the latent space disentanglement with paired observations. Input images x_1 and x_2 have one non-shared factor f (*Blonde Hair* here), whose latent dimensions z_{1_i} and z_{2_i} have to be pushed apart, forcing in that dimension the encoding of factor f.

edit and step magnitude, as is shown Figure 3. Lower λ values give better disentangled edits.

(a) Base image. (b) edit, λ=1 (c) edit, λ=0.1

Figure 3: The λ trade-off explained when editing on *Blonde_Hair*: the edit (3b) with λ=1 unwantedly modifies also *Beard*, whereas the edit (3c) with λ=0.1 keeps the *Beard* desirably unchanged.

With this method, we have been able to generate a "facial-wise disentangled" dataset, *DFaces*, made of image couples that vary in a single facial attribute. The idea is to provide those paired observation to the SVAE, in order to drive it in encoding the non-shared facial attributes within the couples in different latent dimensions.

3.3. Supervised VAE

To achieve latent space disentanglement on non-toy datasets, i.e. when the data generative process isn't trivially aligned with the data FoVs, an explicit supervision is required [6]. We propose SVAE (Figure 4), a framework that employs a set of paired observations $(x_1, x_2)_f$ that are the output of an ideal generative process $\mathcal{GP}$, whose FoVs are kept unaltered except for a single one f. When dealing with real-world datasets, the $\mathcal{GP}$ is never trivial, thus it has to be parametrized, and so the FoVs. In Section 3.2 we show how can it be done in the use-case of facial images. With this pairwise supervision, we aim to force the model to encode in a dedicated latent dimension the variance of the target FoV, grasped from the difference in the images' pair, for each FoV. See Figure 5 as an example on the toy dataset *DSprites*: here the couple differs only in the *shape* FoV, that is constrained to be encoded in the third latent dimension.

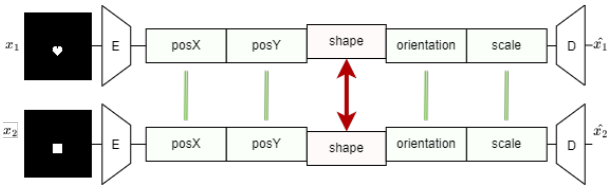

Figure 5: Visual effect of our pair loss on the latent space: it pushes away the non-shared target dimension while it keeps close the others.

So in addition to the standard β-VAE [3] loss function described in Section 2, we add a pair loss term that acts as a disentanglement loss:

$$\mathcal{L}_{pair} = -(z_{1_d} - z_{2_d})^2 + \frac{1}{N-1}\sum_{k \neq d}^{N}(z_{1_k} - z_{2_k})^2$$

where N is the number of FoVs to be encoded, d is the index of the non-shared FoV in the pair $(x_1, x_2)_f$, and z_1,z_2 are the sampled latent representation after the reparametrization trick:

$$z_1 = \mu_1 + \epsilon \cdot \sigma_1 \qquad z_2 = \mu_2 + \epsilon \cdot \sigma_2$$

The heuristic for selecting the latent dimension d responsible for the encoding of the target FoV is to choose the one with the highest KL Divergence. With this additional loss, we expect the model to grasp the difference between each

given pair and to be able to generalize and isolate the variance of the given non-shared factor to a single latent dimension. Once trained the model, the desiderata is visualizing that the latent traversals over each target dimension moves ideally only the correspondent semantic FoV, like it happens the example of Figure 1.

4. Experiments & Results

4.1. Datasets

The main dataset used in the literature when dealing with disentanglement is **DSprites** [7], which was purposefully built for the evaluation of disentangled representation. It's a toy dataset, where each image essentially represents a simple white shape on a black background. It has been generated from five explicit FoVs, namely *shape, scale, orientation, posX, posY*, that coincide with the data generative factors, as they exactly define the image appearance. The other dataset we focused on is **CelebA**, a large-scale face attributes dataset with more than 200K facial images, each with 40 binary attribute annotations.
Since our SVAE requires a specific supervision, we modified the above datasets to generate:

- **SDSprites**: a pairwise selection of the *DSprites* toy dataset, to assess our method soundness.
- **DFaces**: paired facial images varying in a single facial attribute (FoV), generated with our *E2Editor*, that contains networks pretrained on the real-world dataset *CelebA*.

4.2. Metrics

To evaluate the disentanglement level of our results, the three most popular metrics in literature have been tested, namely the *z-diff* metric [3], *z-min variance* metric [5] and the *DCI* metric. They respectively monitor the separation of means, variance, and mutual information.

4.3. Experiments on SDSprites

We first generated **SDSprites**, composed of paired images from *DSprites* in which a single FoV varies. Samples are shown in Figure 6.

Figure 6: Paired samples from SDSprites, varying respectively in *shape, scale, orientation, posX, posY*.

For this experiment we set the *latent dimension* to 5, equal to the number of FoVs; it's a wide enough bottleneck to allow for a valid reconstruction, given the triviality of the dataset. As we can see from the latent traversal of Figure 7, the FoVs reach a valid disentanglement being encoded in separate independent dimensions.

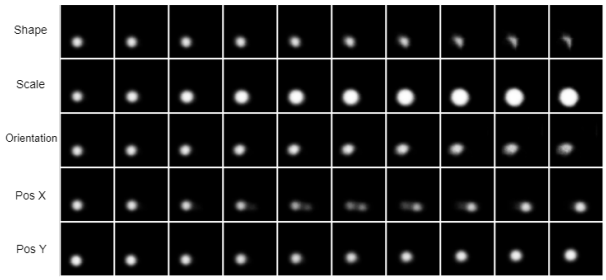

Figure 7: Latent Traversal of SVAE trained on *SDSprites*. Given the same sampled image of the first column, each row displays a traversal on a different dimension, that semantically encodes the FoV on the left.

Comparing SVAE (trained for 3k epochs) with the state-of-the-art methods (each trained for 300k epochs in Google's Official Disentanglement Library) [6], we show competitive results given the significantly shorter training duration, highlighting the efficiency and practicality of our approach for disentanglement representation learning. Results are shown in Table 1.

	z-diff	z-min	DCI
βVAE	0.823	0.660	0.186
FactorVAE	0.853	0.750	0.256
MLVAE	0.896	0.701	0.294
GVAE	**0.923**	**0.847**	**0.479**
SVAE	0.652	0.711	0.313

Table 1: Disentanglement metrics comparison.

4.4. Experiments on DFaces

Since our focus wants to be real-world datasets, we needed a tailored supervision on that domain too. With the *E2Editor* method described in Section 3.2, we managed to create a novel dataset, **DFaces**. This dataset comprises 5883 image pairs for each of the 17 chosen attributes, namely A, where each pair varies in a single facial attribute. The facial attributes have been treated as the data FoVs, as they are the finest semantic factors that define the image aspect.

Figure 8: Paired samples from *DFaces* dataset; side by side images have one non-shared attribute, among the 17/40 chosen from *CelebA*.

The more this dataset is precise in its edits, the more disentangled our VAE can be. To evaluate the level of "facial disentanglement" within each couple, we defined a metric that keeps into account the variations of the non-target attributes scores when editing on a target one (the lower the better). Given N generated base images, we performed $N \cdot A$ edits, one for each attribute a on every base image; we then took the average across the MSE over the scores on the non-target attributes between the base and edited image:

$$v_A = \frac{1}{A}\sum_{a_{tg}}^{A} \underbrace{\frac{1}{N}\sum_{i}^{N}\sum_{a\neq a_{tg}}^{A} |\widetilde{s}_a - s_a|^2}_{\text{MSE editing on } a_{tg}}$$

where $\widetilde{s}_a$ and s_a are respectively the classifier $\mathcal{C}$ score of attribute a of the edited image on a_{tg} and of the base image. We improve the state-of-the-art result of *InterFaceGAN* [8] when setting $\alpha = 0.1$. Results are shown in Table 2.

	Avg Sum of MSE
E2Editor (α=1)	0.120156
InterFaceGAN	0.074868
E2Editor (α=0.1)	**0.057652**

Table 2: SFAE methods quality comparison.

Figure 9: Latent Traversals of SVAE trained on *DFaces*. Given the sampled images of the first row, each row displays the decoded image when modified on the same dimension, that encode respectively *Blonde_Hair, Pointy_Nose, Rosy_Cheeks, Bald*.

Finally, we have been able to feed these pairs to our SVAE, to generate an explainable model disentangled over facial attributes. Here the *latent dimension* has been set to 128, but only the forced 17 dimensions carry an interpretable semantic FoV. In the rows of Figure 9 we showcase the most significant latent traversals.

5. Conclusions

In summary, this work addressed the problem of generating explainable models by leveraging on the forced latent space disentanglement. The proposed solution consists of a novel Variational framework in which paired observations differing in a single factor of variation induce the model into encoding that non-shared attribute in a dedicated latent dimension. To test our method's effectiveness on real-world scenarios, we developed a novel Semantic Facial Attribute Method to generate a tailored dataset. The gathered results show the viability of the suggested strategy and its competitiveness compared to the main solutions in the literature.

References

[1] Yoshua Bengio. Representation learning: A review and new perspectives. 2013.

[2] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Mlvae: Learning disentangled representations from grouped observations, 2017.

[3] Irina Higgins, Loïc Matthey, Arka Pal, and Christopher P. Burgess. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

[4] Haruo Hosoya. Group-based learning of disentangled representations with generalizability for novel contents, 2021.

[5] Hyunjik Kim and Andriy Mnih. Disentangling by factorising, 2019.

[6] Francesco Locatello, Stefan Bauer, Mario Lucic, and Gunnar Rätsch. Challenging common assumptions in the unsupervised learning of disentangled representations, 2019.

[7] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites, 2017.

[8] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing, 2020.