JOINT DISENTANGLEMENT OF LABELS AND THEIR FEATURES WITH VAE

Kaifeng Zou, Sylvain Faisan, Fabrice Heitz, Sebastien Valette

JOINT DISENTANGLEMENT OF LABELS AND THEIR FEATURES WITH VAE

Kaifeng Zou[⋆]*, Sylvain Faisan*[⋆]*, Fabrice Heitz*[⋆]*, Sébastien Valette*[†]

[⋆] ICube Laboratory, University of Strasbourg, France
[†] CREATIS, INSA-Lyon, Lyon, France

ABSTRACT

Most of previous semi-supervised methods that seek to obtain disentangled representations using variational autoencoders divide the latent representation into two components: the non-interpretable part and the disentangled part that explicitly models the factors of interest. With such models, features associated with high-level factors are not explicitly modeled, and they can either be lost, or at best entangled in the other latent variables, thus leading to bad disentanglement properties. To address this problem, we propose a novel conditional dependency structure where both the labels and their features belong to the latent space. We show using the CelebA dataset that the proposed model can learn meaningful representations, and we provide quantitative and qualitative comparisons with other approaches that show the effectiveness of the proposed method.

***Index Terms*—** disentangled representation, variational autoencoder

1. INTRODUCTION

It is a key challenge to learn disentangled representations where variables of interest are independently and explicitly encoded [1]. These representations allow to manipulate data by modifying high level factors (e.g. removing or adding glasses to a person's face). Probabilistic generative models, such as Variational Autoencoders (VAE) [2] are popular to learn such representations in the unsupervised [3, 4, 5], (semi-)supervised [6, 7], and in the weakly-supervised [8] cases. We focus hereafter on the semi-supervised case because supervision yields better disentangled models [9].

Most previous works [10, 6, 7, 11] divide the latent representation into two components: the non-interpretable part and the disentangled part corresponding to variables that explicitly model the factors of interest. Each factor of interest is therefore associated to a latent variable of the same type. As an example, if the label of interest refers to the glasses (1 when the subject is wearing glasses, 0 otherwise), there will be a categorical variable in the latent space that encodes the presence or absence of glasses. However, this variable does not allow to model the features of the glasses (e.g. shape/size/color of the glasses), that can be either lost, or at best entangled in the other latent variables.

To our knowledge, only [12] proposed to address this problem. In [12], a feature is associated with each high level factor. Moreover, the latent space no longer contains the labels but their features (the label is used to condition its associated feature). We propose here a novel conditional dependency structure in which the latent space contains both the labels and their features.Finally, we use an original architecture to build the decoder of the VAE. We show that AdaIN [13] improves the quality of the reconstructed images and that the use of learnable tokens [14, 15] improves disentanglement properties of the model.

Finally, note that generative adversarial networks can also be used to obtain disentangled representations: the methods proposed in [16] and [17] also allow to manipulate the features related to high level factors. This is achieved by swapping attributes between pairs of images. However, these methods are only able to accomplish a small number of the tasks that can be performed with VAE-based methods. As [16] and [17], the proposed method can also swap the high level factors and the related features of two images. However, (i) it allows also to generate new images by sampling from the model (without any other input or with high level factors only), (ii) it allows also to modify the high level factors and the associated features for a single image (by sampling), (iii) it provides finally a classifier that estimates the high level factors. Note also that the methods of [16] and [17] are fully supervised whereas the proposed method handles arbitrary supervision rates.

2. DISENTANGLEMENT OF LABELS AND THEIR FEATURES FROM OTHER LATENT VARIABLES

2.1. Conditional dependency structure

For the sake of simplicity, we consider here that a unique label (high level factor) is provided for an image. The extension to several labels is straightforward. For the illustration, we consider the binary case where the label is 1 (e.g. if the subject is wearing glasses), or 0 otherwise.

Let x be an image, y its label, u the features related to

This work was funded by the TOPACS ANR-19-CE45-0015 project of the French National Research Agency (ANR).

label y, and z the other latent variables that are supposed to carry no information on y and u. The latent space is formally composed of y, z and u. The generative process is inspired by the previous work of [7], except that no feature u is defined in [7]. It writes:

$$p_\theta(x, y, z, u) = p_\theta(x|y, z, u)p(u|y)p(y)p(z), \qquad (1)$$

where θ stands for the parameters of the decoder. A weak prior is defined over z and y: z follows a zero-centered multivariate normal distribution with unit variance ($p(z) = \mathcal{N}(z; 0, I)$) and y follows a uniform discrete distribution. $p_\theta(x|y, z, u)$ is modelled as a Gaussian distribution whose mean is computed by a neural network (the decoder d_θ of parameter θ) that takes as input y, z and u. We have:

$$p_\theta(x|y, z, u) = \mathcal{N}(x; d_\theta(y, z, u), vI), \qquad (2)$$

where v is a hyperparameter. Finally, a special care is needed to model $p(u|y)$. In our application, the feature vector u encodes the shape/size/color of the glasses. So as to favor disentanglement properties of the model, the two prior distributions (one for each possible value of y) differ from each other. Two different approaches denoted as PA1 and PA2 (proposed approach 1 and 2) are considered:

- Case $y = 1$ (glasses). For both approaches, $p(u|y = 1)$ is a zero-centered multivariate normal distribution with unit variance.
- Case $y = 0$ (no glasses). For PA1, $p(u|y = 0)$ is a multidimensional Dirac delta function, enforcing the components of u to be zero. For PA2, it is a zero-centered multivariate normal distribution with a variance equal to the identity matrix multiplied by 0.1, favoring the components of u to be close to 0.

PA1 seems to be a better choice since images with no glasses should all have the same value of u. We use PA2 to show that the proposed modeling may work with a less informative prior.

The posterior $p_\theta(y, z, u|x)$ is approximated by $q_\phi(y, z, u|x)$ which can be factorized as:

$$q_\phi(y, z, u|x) = q_\phi(y|x)q_\phi(z|x, y)q_\phi(u|x, y), \qquad (3)$$

where ϕ stands for parameters of the encoder. In Eq. 3, we assume that z and u are independent conditionally to x and y. The distribution $q_\phi(y|x)$ is a discrete distribution whose probabilities are provided by the softmax layer (See Fig. 1). The distribution $q_\phi(z|x, y)$ is defined as a Gaussian distribution whose mean (resp. covariance matrix) is given by the encoder. For PA1 (case y=1 only) and for PA2, the distribution $q_\phi(u|x, y)$ is defined in the same way as $q_\phi(z|x, y)$. For PA1 (case y=0), $q_\phi(u|x, y = 0)$ is modeled as a multidimensional dirac delta function. Indeed, in this case, the prior distribution $p(u|y = 0)$ tells us that u is the null vector.

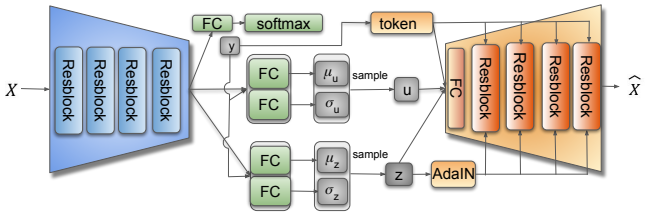

Fig. 1. Model architecture. FC stands for fully connected layer. For testing, if y is not known, y is set to the most likely label (based on the output of the softmax layer that represents $q_\phi(y|x)$). z and u are set to μ_z and μ_u. For training (Section 2.2), if y is not known, y is sampled from $q_\phi(y|x)$ using a Gumbel-softmax relaxation. z and u are sampled from $q_\phi(z|x, y)$ and $q_\phi(u|x, y)$.

The proposed architecture is depicted in Fig. 1. We use AdaIN [13] as a normalization method: it enables the information carried by z to be transferred to each layer of the decoder (through a fully connected layer that is not shown in Fig. 1). Moreover, we use one set of learnable tokens [14, 15] per class. The set is then selected according to the value of y. Each set is composed of five tokens (one scalar and four images that are associated each one to a residual block of the decoder). The first one (the scalar) is concatenated to u and z to feed the first fully connected layer of the decoder. Then, for each token (an image), we concatenate the token and the input of its associated residual block along the channel dimension. It allows the information provided by y to be transferred to each input of the residual block.

Finally, for PA1, u is multiplied by y. It enables to constrain u to be a null vector if y is 0, and not to modify its value otherwise ($y = 1$).

2.2. Parameter optimization

If y is known, the optimization of $\log p(x, y)$ can be achieved by maximizing the ELBO (Evidence Lower BOund), that writes: $E_{z,u\sim q_\phi(z,u|x,y)} log(p_\theta(x, y, u, z)/q_\phi(z, u|x, y))$. As in [7], we add a classification loss $\alpha \log q_\phi(y|x)$ to the ELBO term. By using Eq. 1 and 3, we obtain the following criterion (it is divided by the number of pixels N):

$$\begin{array}{ll} \beta E_{z\sim q_\phi(z|x,y)} \left[\log(p(z)) - \log(q_\phi(z|x, y))\right] & + \\ \beta E_{u\sim q_\phi(u|x,y)} \left[\log(p(u|y)) - \log(q_\phi(u|x, y))\right] & + \\ \frac{1}{N} E_{z,u\sim q_\phi(z,u|x,y)} \left[\log(p_\theta(x|y, z, u))\right] & + \\ \beta \log(p(y)) + \alpha \log q_\phi(y|x) & \end{array} \qquad (4)$$

The two first terms are a Kullback–Leibler divergence which can be computed analytically since the distributions are Gaussian except for the second term in the case of PA1 with y=0. In this case, it vanishes to 0 since both distributions are equal. Then, the third term is approximated by a Monte Carlo estimate: we use the SGVB estimator and the reparam-

eterization trick [2] (with the notation of Fig. 1, we have: $(z, u) = (\mu_z, \mu_u) + (\sigma_z, \sigma_u) \odot \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$). The fourth term corresponds to the prior of the label y, that has been set to 1/2. Without loss of generality, the variance v (Eq. 2) is set to 1 to compute the third term of Eq. 4 and the other terms of the ELBO are weigthed by a factor β. Consequently, two hyperparameters have to be set: α and β.

If y is not known (semi-supervised case), it has to be treated as a latent variable. Marginalization can be performed [7]. We sample y, as in [6], from the discrete distribution $q_\phi(y|x)$ using a Gumbel-softmax relaxation.

3. EXPERIMENT

We experiment on the CelebA dataset [18], with an image size of 128x128. The glasses label has been selected because it leaves little room for subjectivity. The hyperparameters of the methods have been set by using a cross-validation strategy on the training set. Concerning the criterion (Eq. 4), α has been set to 1 (the value of α has little influence on the results) and β to $1e-4$. Since the second term of the ELBO (for y=1) leads to degrade the results, its weight (for y=1) has been divided by 100 (for both PA1 and PA2). We use the Adam optimizer [19] with a learning rate equal to $1e-4$ and a batch size of 32. The sizes of z and u are set to 100 and 16 respectively. The supervision rate has been set to 0.2.

We compare our method with CCVAE [12] and with the model M2 of [7]. Two different architectures are used for CCVAE. We first use the implementation of the authors. Since it is adapted to the processing of images of size 64x64, the sizes of the input/output layers have been modified accordingly. This method is denoted as CCVAE. For the second method denoted as CCVAE2, we adapt the architecture of our model to the conditional dependency structure of CCVAE. As an example, y is no longer part of the latent space in CCVAE so that it is no more used for estimating the reconstruction. In the same way, we adapt the architecture of the proposed model to the conditional dependency structure of M2. It leads to the removal of u from the modeling. For all models, the size of the latent space that models the attributes of the face has been set to 100.

As mentioned in the introduction section, the VAE-based methods allow to accomplish several tasks. For the sake of simplicity, we will only consider three different tasks for comparison purposes: the classification task, the reconstruction task, and the exchange of high level factors and of the related features (if they exist) between two images. The latter task enables us to clearly observe the disentanglement capability of the model. Indeed, it enables to check that the model disentangles not only the label but also the features (of the glasses) from the face attributes z. To evaluate the quality of the reconstructed images, we use Learned Perceptual Image Patch Similarity (LPIPS) [20] that computes perceptual difference between two images. The disentangled ability of the model

Table 1. Quantitatives results in terms of (i) success rates for removing and adding glasses (SR(-) and SR(+)), (ii) LPIPS between the original images and the reconstructed ones, and (iii) balanced classification accuracy (BCA).

Model	SR(-)	SR(+)	LPIPS	BCA
CCVAE	**99.38%**	19.37%	0.4414	95.45%
CCVAE2	95.52%	47.68%	0.2549	94.26%
M2	80.34%	33.03%	0.2564	96.13%
PA1	96.31%	59.85%	0.2484	96.55%
PA2	94.98%	**64.25%**	**0.2416**	**97.09%**

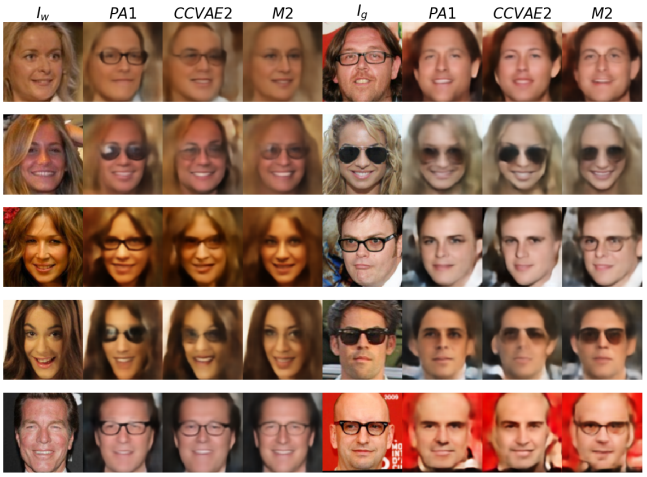

Fig. 2. Attribute swapping for 5 pairs of images using PA1, CCVAE2 and M2. For each line, the second, third and fourth image should be I_w with the glasses of I_g. The three rightmost images should be I_g, but without the glasses.

is evaluated by computing the success rate of swapping. To this end, we select random pairs of images composed of one image with glasses (I_g) and one image without (I_w). Their values of y and u are then exchanged. We consider that the glasses are correctly removed from I_g (resp. added to I_w) if the reconstruction (after the attribute swapping) is classified as $y = 0$ (resp. $y = 1$) with an independent classifier based on ResNet 50. We denote by SR(-) (resp. SR(+)) the success rates for removing (resp. adding) glasses. Results obtained with the different approaches are shown in Tab. 1. Since SR(+) is not a perfect evaluation criterion for measuring disentanglement properties of the models (it does not check that the glasses added to I_w are those of I_g), Fig. 2 presents swapping results for 6 pairs of images in the case of PA1 (PA2 provides similar results), M2 and CCVAE2.

First, all methods obtain good classification accuracy (BCA) despite a supervision rate equal to 0.2.

Regarding the quality of the generated images (LPIPS), all methods, except CCVAE achieve very similar results. This is mainly due to the fact that the decoders of all the methods (ex-

cept CCVAE) are very similar. Moreover, we observed that the removal of AdaIN leads to a substantial increase of LPIPS (without modifying significantly the other evaluation criteria). As an example, the removal of AdaIN for PA1 brings the LPIPS criterion from 0.2484 to 0.3059. This clearly highlights the benefit of AdaIN: it allows to improve the reconstruction of the images by transferring to each layer of the decoder information carried by z.

With respect to the success rates of swapping (SR(-) and SR(+)), results obtained with M2 are not very satisfactory, thus illustrating the importance to model the features related to the label. Since the glasses (for M2) are actually well-reconstructed without any label/feature swapping, their features may be entangled in the other variables z of the latent space. This makes the addition of glasses difficult because modifying y is not enough: other variables of the latent space have to be modified to define some proper features of the glasses to be added. Conversely, it appears that the removal of the glasses is simpler (SR(-)>SR(+)) insofar as modifying the label y is enough.

For CCVAE, results obtained for SR(+) and SR(-) do not inform us about the disentanglement properties of the model because the generated images are actually so blurred that it is most of the time difficult to observe the glasses.

Results obtained with CCVAE2 are better than those obtained with M2 in terms of SR(+) and SR(-), illustrating the interest of modeling the features related to a label. However, we observe Fig. 2 that the features of glasses cannot be transferred to other images. This means that two images with the same values of u do not exhibit the same glasses. Since the modification of u still leads to the modify the features of the glasses, the features of the glasses are partially entangled in the other latent variables with CCVAE2.

Finally, the proposed approaches (PA1 and PA2) achieve a success rate for adding glasses that is superior to those obtained with other models as well as a very high success rate for removing glasses. Moreover, we can observe (Fig. 3) that the proposed model (PA1) correctly extracts the features of the glasses from the image I_g and is able to reconstruct them reasonably well on another image, which shows that the label as well as the features of the glasses have been properly disentangled from the attributes of the faces. Similar results are obtained for PA2. Note that the results presented in Fig. 2 cannot be considered as representative: $SR(+)$ is about 60 % for PA1 but PA1 obtains good results for all pairs of images of Fig 2. The proposed methods achieve actually very good results (the glasses added to I_w match those of I_g and the glasses are correctly removed from I_g) for many pairs of images. However, such results are extremely rare with CCVAE2 and M2. These results show the relevance of the proposed conditional dependency structure, and in particular the benefit of y being in the latent space. We have also noted that the tokens favor the disentanglement properties of the model by allowing the information provided by y to be transferred to each input of the residual block of the decoder. As an example, for PA1, the removal of the tokens (y is then just used as an input of the fully connected layer of the decoder) brings SR(+) down from 59.85% to 47.23% (SR(-) is not modified significantly).

Fig. 3. Multiple attribute swapping with our method. We add to the 3 images of the first column the beard associated with the image which is located on the same line on the rightmost and the glasses associated with the images of the first line.

Finally, the proposed method can easily be extended to the case of several high level factors. In the case of two factors, (y_1, u_1) and (y_2, u_2) can be considered as independent for the generative process. Then, for the variational approximation, we write $q_\phi(y_1, y_2, z, u_1, u_2|x)$ as $q_\phi(y_1|x)q_\phi(y_2|x)q_\phi(z|x, y_1, y_2)q_\phi(u_1|x, y_1)q_\phi(u_2|x, y_2)$. Results obtained with the glass and the beard labels are shown in Fig. 3. They illustrate that the proposed model allows to manipulate the attributes of beard and glasses separately.

4. CONCLUSION

The proposed approach compares favorably to other VAE-based approaches, thus showing the interest of modeling both the labels and their features in the latent space. Moreover, our experiments illustrate the benefit of using AdaIn and learnable tokens to build the decoder: the first one allows to improve the quality of the generated images while the second one favors disentanglement properties of the model. To further improve the quality of the generated images and in particular to obtain less blurry images, a perspective of this work could be to replace the Gaussian prior on z by a categorical distribution [21]. Better reconstruction may also favor a better disentanglement.

5. REFERENCES

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[2] Diederik P. Kingma and Max Welling, "Auto-Encoding Variational Bayes," in *2nd International Conference on Learning Representations, ICLR, Canada, Conference Track Proceedings*, 2014.

[3] Ricky T. Q. Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud, "Isolating sources of disentanglement in variational autoencoders," in *Advances in Neural Information Processing Systems*, 2018, vol. 31.

[4] I. Higgins, L. Matthey, A. Pal, Christopher P. Burgess, Xavier Glorot, M. Botvinick, S. Mohamed, and Alexander Lerchner, "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *ICLR*, 2017.

[5] Hyunjik Kim and Andriy Mnih, "Disentangling by factorising," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 2649–2658.

[6] N. Siddharth, Brooks Paige, Jan-Willem van de Meent, Alban Desmaison, Noah D. Goodman, Pushmeet Kohli, Frank Wood, and Philip H. S. Torr, "Learning disentangled representations with semi-supervised deep generative models," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

[7] D.P. Kingma, D.J. Rezende, S. Mohamed, and M. Welling, "Semi-supervised learning with deep generative models," in *Advances in Neural Information Processing Systems*, 2014.

[8] Adria Ruiz, Oriol Martinez, Xavier Binefa, and Jakob Verbeek, "Learning disentangled representations with reference-based variational autoencoders," 2019.

[9] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4114–4124.

[10] Lars Maaløe, Casper Kaae Sønderby, Søren Kaae Sønderby, and Ole Winther, "Auxiliary deep generative models," in *Proceedings of The 33rd International Conference on Machine Learning*, 2016, pp. 1445–1453.

[11] Kaifeng Zou, Sylvain Faisan, Fabrice Heitz, Marie Epain, Pierre Croisille, Laurent Fanton, and Sébastien Valette, "Disentangled representations: towards interpretation of sex determination from hip bone," *arXiv preprint arXiv:2112.09414*, 2021.

[12] Tom Joy, Sebastian Schmon, Philip Torr, N Siddharth, and Tom Rainforth, "Capturing label characteristics in vaes," in *International Conference on Learning Representations*, 2020.

[13] Xun Huang and Serge Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1501–1510.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

[15] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, and al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Ninth International Conference on Learning Representations*, 2021.

[16] Taihong Xiao, Jiapeng Hong, and Jinwen Ma, "Elegant: Exchanging latent encodings with gan for transferring multiple face attributes," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 168–184.

[17] Taihong Xiao, Jiapeng Hong, and Jinwen Ma, "Dna-gan: Learning disentangled representations from multi-attribute images," *arXiv preprint arXiv:1711.05415*, 2017.

[18] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[19] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[21] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.