INTERPRETABLE REPRESENTATION LEARNING ON NATURAL IMAGE DATASETS VIA RECONSTRUCTION IN VISUAL-SEMANTIC EMBEDDING SPACE

Nao Nakagawa, Ren Togo, Takahiro Ogawa, Miki Haseyama (Hokkaido University, Japan)

Paper ID: 1517

1. INTRODUCTION

1.1 Representation Learning

The data representation strongly affects the performance of machine learning [1].

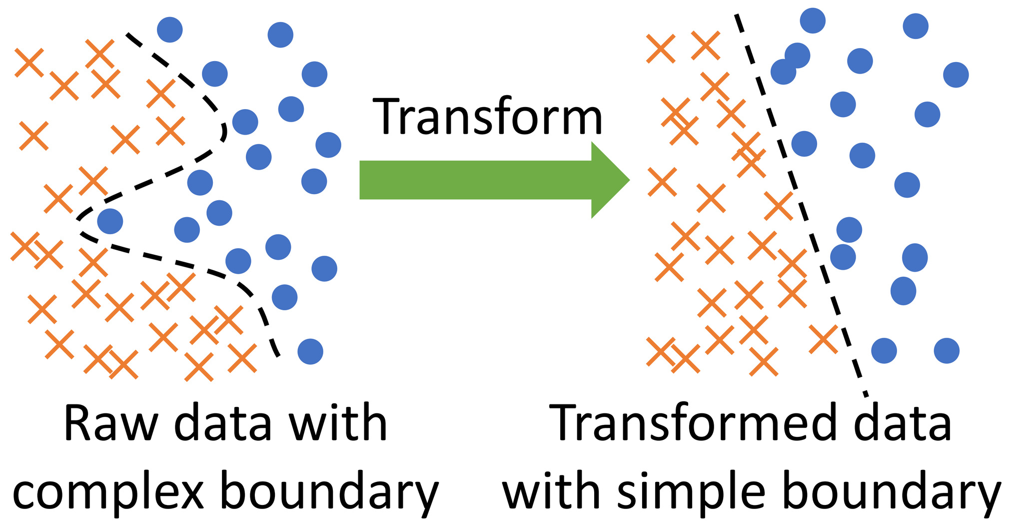


Figure: classification into two classes × and ●.

In particular, **disentangled representation learning (DRL)** have attracted much attention in the field of representation learning [15-20].

1.2 Disentangled Representation Learning

DRL aims to obtain **disjoint, independent latent variables** corresponding to semantically meaningful factors of variation by unsupervised learning [1, 2, 5, 6].

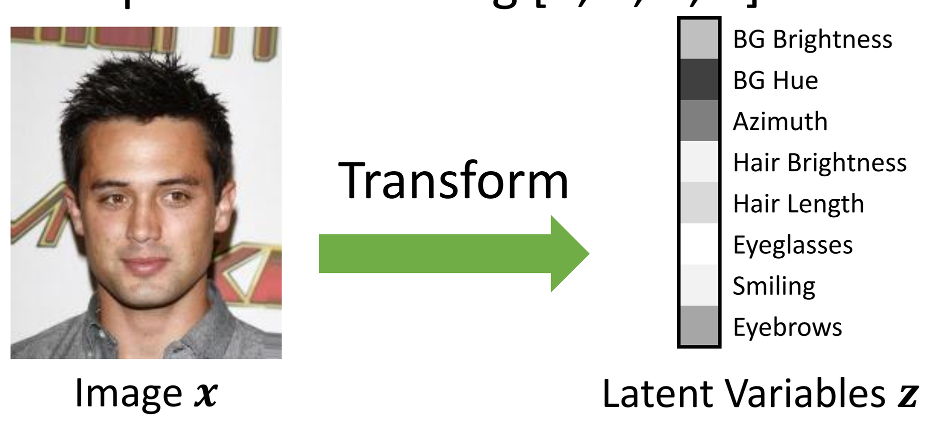


Figure: a diagram of DRL.

The most popular form is a deep generative model based on **Variational Autoencoder (VAE)** [4], which has an explicit constraint to infer independent latent variables.

1.3 Weakly-Supervised Disentanglement

✗ Unsupervised generative models cannot distinguish representations with the identical distribution [15].

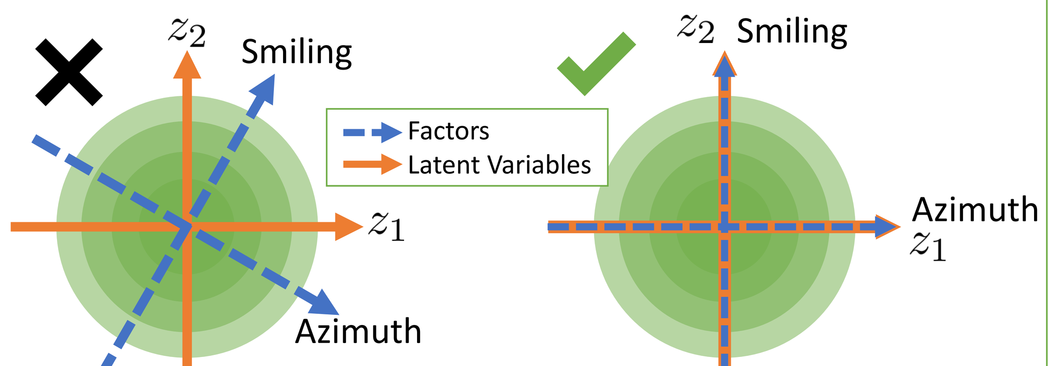


Figure: an entangled representation (left) and a disentangled one (right) in the same distribution.

Our Approach in this paper:
Learning an unsupervised VAE-based generative model where each latent variable has a word explaining its representation

2. PROPOSED DISENTANGLEMENT METHOD

2.1 Networks and Learning

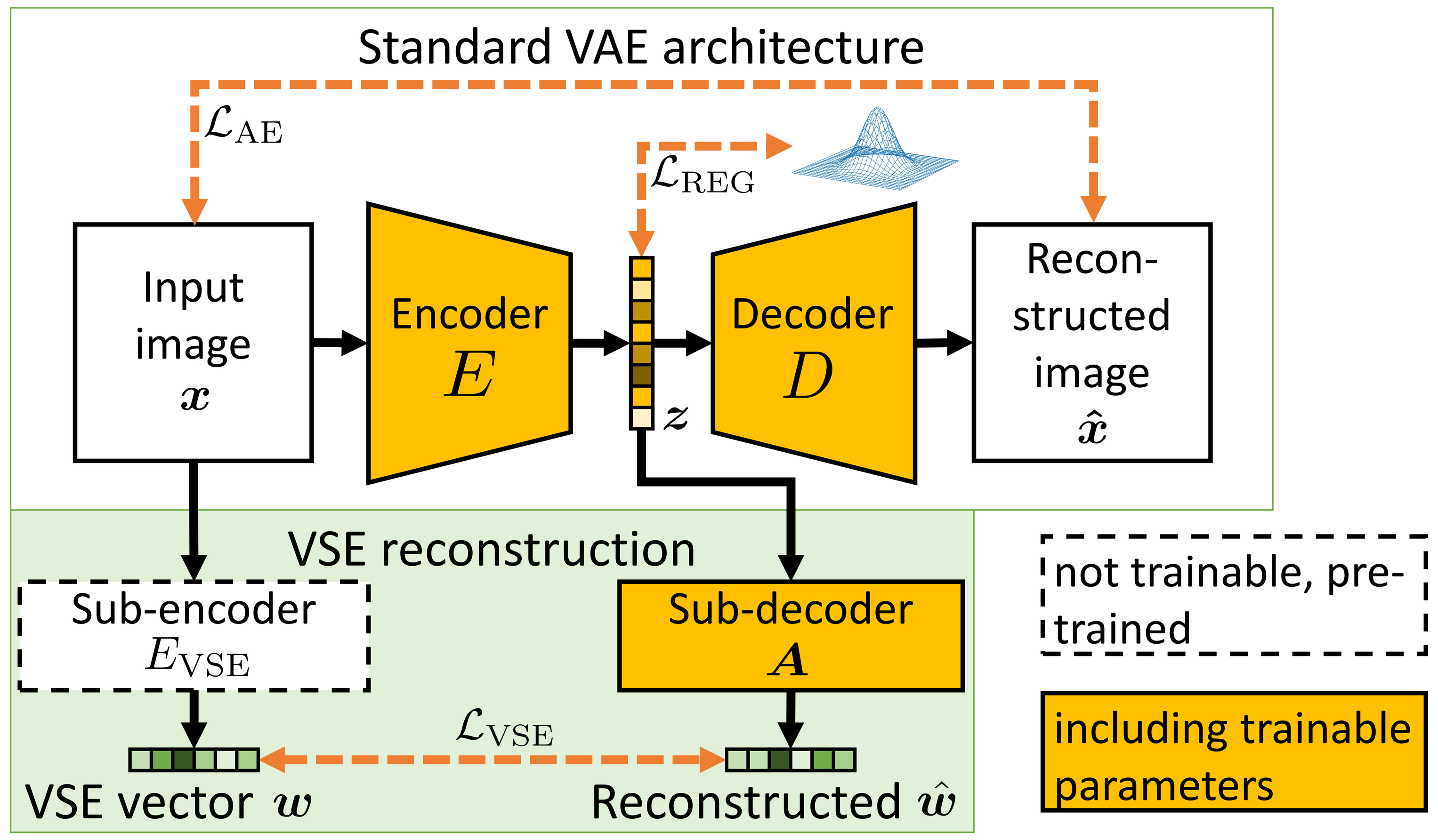


Figure: an overview of our VAE-based model.

Visual-Semantic Embedding (VSE) [22]
Visual and semantic contents are embedded in the same space.

Loss Function

$$\mathcal{L} = \mathcal{L}_{\mathrm{AE}} + \beta\mathcal{L}_{\mathrm{REG}} + \gamma\mathcal{L}_{\mathrm{VSE}}$$

- β encourages the independence of latent variables.
- γ encourages the reconstruction in the VSE space, which supports the semantic disentanglement and the explanation.

✓ We introduce the semantic information into a VAE-based deep generative model via the VSE reconstruction.

2.2 Explanation by Additive Compositionality: Word Embedding and Latent Units

The basis vector $\boldsymbol{a}_i$ of the linear sub-decoder $\boldsymbol{A}$ can be interpreted as the meaning of the latent representation z_i by finding a word with the highest cosine similarity between its embedding and $\boldsymbol{a}_i$.

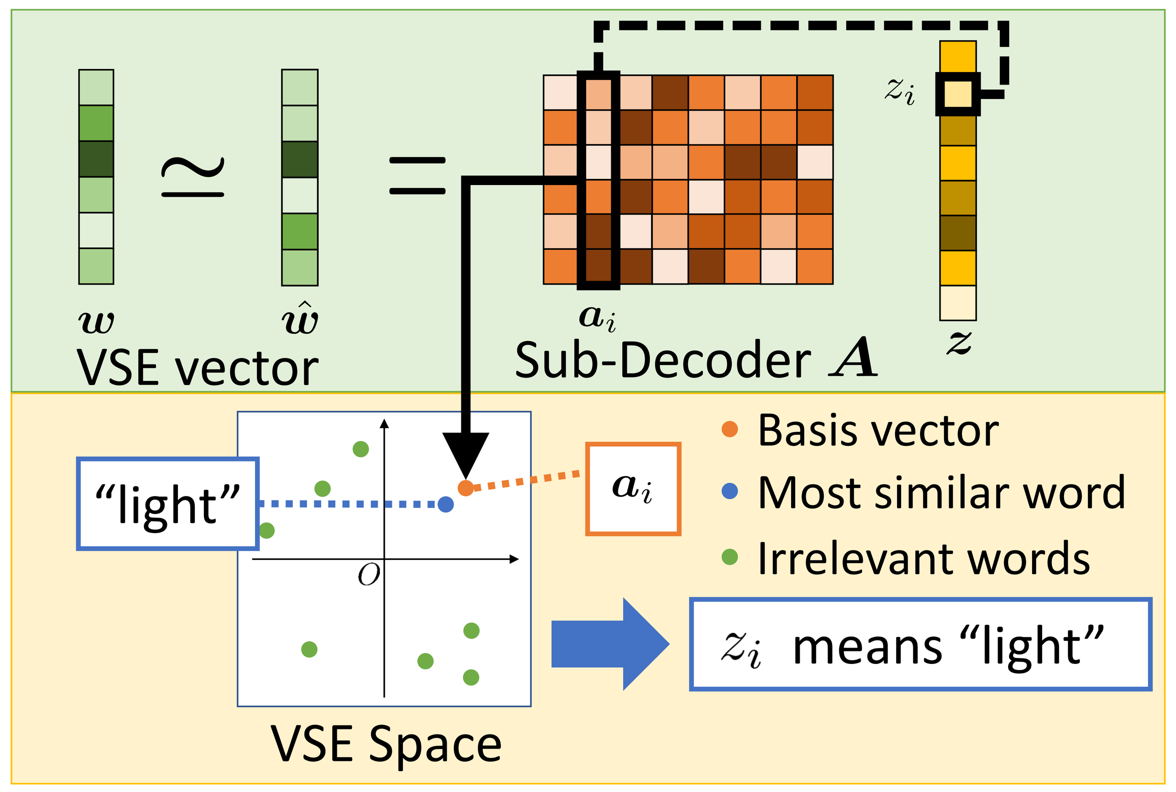


Figure: explanation of learned latent representations by our model.

The meanings of the independent latent variables are superposed using the additive compositionality of word embeddings.

$$\text{e.g.,}\quad \boldsymbol{w}_{\mathrm{king}} \simeq \boldsymbol{w}_{\mathrm{man}} + \boldsymbol{w}_{\mathrm{royal}}$$

✓ Our model can explain the obtained latent representations to perform unsupervised DRL along the explained words.

3. EXPERIMENTAL RESULTS

3.1 Experimental Settings

Datasets

- CelebA [23]: 202,599 face images with 40 attribute labels (training images: 200,551, test images: 2,048)
- Stanford Cars [24]: 16,185 automobile images with 196 class labels (training images: 8,144, test images: 8,041)

Network Architecture Settings

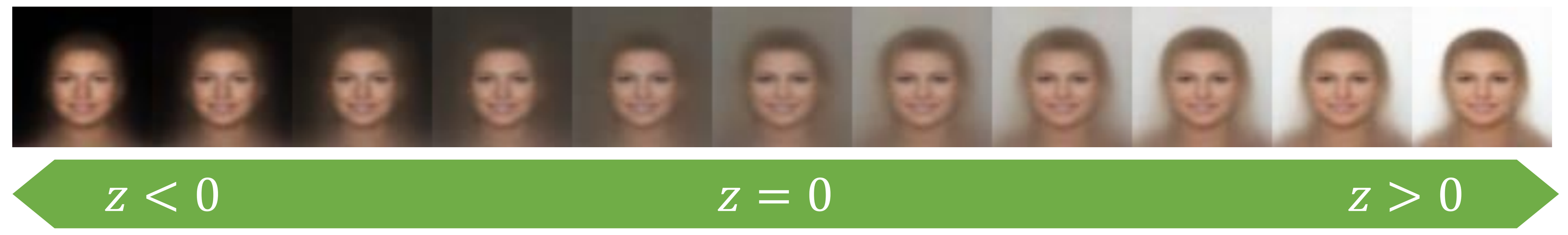
- Num. of latent variables: $N = 32$ → the same settings as [2]
- Sub-Encoder: the pre-trained VSE image encoder [22]
- Hyperparameters: $\beta = 1, \gamma = 10$

Compared Methods

- VAE [4]
- β-VAE [2]: $\beta = 10$
- CCβ-VAE [11]: $\beta = 10$
- β-TCVAE [6]: $(\alpha, \beta, \gamma) = (1,10,1)$
- FactorVAE [12]: $\gamma = 10$
- DIP-VAE-I [13]: $\lambda_{od} = 4, \lambda_d = 200$
- DIP-VAE-II [13]: $\lambda_{od} = 80, \lambda_d = 40$

3.2 Qualitative Evaluation

(Dataset: CelebA [23])



Words describing the − direction	**Words describing the + direction**
"dark" (sim: -0.38892)	"fishing" (sim: 0.31756)
"night" (sim: -0.31997)	"parasail" (sim: 0.29918)
"spraying" (sim: -0.27426)	"Oatmeal" (sim: 0.29678)

Figure: An example of latent traversal with the top-3 similar words (sim: cosine similarity with the basis vector of the latent variable)

3.3 Quantitative Evaluations

↑: higher is better. ↓: lower is better.

Table: Evaluations of obtained representations in disentanglement and transferability.

Dataset	CelebA			Stanford Cars			CelebA
Metric	WINDIN ↑	RMIG ↑	JEM-MIG ↓	WINDIN ↑	RMIG ↑	JEM-MIG ↓	Transfer Learning Error↓
VAE [4]	0.0353	0.0462	0.727	0.0367	0.0030	1.302	16.15% ± 0.32
β-VAE [2]	0.0563	0.0267	0.851	0.0520	0.0034	1.380	18.06% ± 0.30
CCβ-VAE [11]	0.0382	0.0465	0.635	0.0367	0.0031	1.022	16.61% ± 0.32
β-TCVAE [6]	0.0661	0.0269	0.996	0.0941	0.0038	1.389	18.18% ± 0.36
FactorVAE [12]	0.0352	0.0520	0.376	0.0360	0.0035	0.991	16.94% ± 0.27
DIP-VAE-I [13]	0.0336	0.0205	0.730	0.0333	0.0032	1.312	16.78% ± 0.30
DIP-VAE-II [13]	0.0358	0.0178	0.445	0.0330	0.0009	0.913	17.81% ± 0.64
Ours	0.0394	0.0506	0.714	0.0342	0.0030	1.258	**16.01% ± 0.24**
Ours + β-TCVAE	**0.0848**	0.0256	0.985	**0.0965**	**0.0038**	1.386	-
Ours + FactorVAE	0.0336	**0.0588**	**0.247**	0.0360	0.0033	**0.903**	-

✓ The effectiveness of our methods has been demonstrated in disentanglement and transferability over other existing VAE-based DRL methods.

Acknowledgement: This work was partly supported by JSPS KAKENHI Grant Number JP21H03456.