Variational Bi-domain Triplet Autoencoder

Rita Kuznetsova
Moscow Institute of Physics and Technology
Antiplagiat Company
Moscow
rita.kuznetsova@phystech.edu

Oleg Bakhteev*
Moscow Institute of Physics and Technology
Antiplagiat Company
Moscow
bakhteev@phystech.edu

ABSTRACT

We investigate deep generative models that allow us to use training data from one domain to build a model for another domain. We consider domains to have similar structure (texts, images). We propose the Variational Bi-domain Triplet Autoencoder (VBTA) that learns a joint distribution of objects from different domains. There are many cases when obtaining any supervision (e.g. paired data) is difficult or ambiguous. For such cases we can seek a method that is able to the information about data relation and structure from the latent space. We extend the VBTAs objective function by the relative constraints or triplets that are sampled from the shared latent space across domains. In other words, we combine the *deep generative model* with a *metric learning ideas* in order to improve the final objective with the triplets information. We demonstrate the performance of the VBTA model on different tasks: bi-directional image generation and image-to-image translation. We also provide the qualitative analysis. We show that VBTA model is comparable with some of the existing generative models. We also show that it outperformes some of these methods methods.

KEYWORDS

Generative Models, Variational Inference, Representation Learning, Semi-Supervised Learning, Transfer Learning

1 INTRODUCTION

Learning distributed representations from data is one of the most challenging task in many machine learning problems. Recent advances in probabilistic deep generative models allow us to specify a model as joint probability distribution over the data and latent variable consider the representations as samples from the posterior distribution on latent variables given data.

Variational autoencoders (VAEs) [9] estimate the data using variational inference with a few assumptions about data distribution and approximate posterior distribution. They make it possible to use latent variables as our learned representation.

Inspired by works [6], [8], [19] we propose Variational Bi-domain Triplet Autoencoder (VBTA) that learns a joint distribution of objects from different domains $\mathbf{x}$ and $\mathbf{y}$ having a similar structure (e. g. texts, images). VBTA allows using distributed representations as samples from shared latent space $\mathbf{z}$ that captures characteristics from both domains. In Section (3) similar to [12] we make assumptions about shared-latent space, in which the paired objects (images, sentences) from different domains are close to each other. In Section (4), similar to [19] and [20] we define the joint probability as $p(\mathbf{x}, \mathbf{y}; \theta) = \int_{\mathbf{z}} p_{\theta_\mathbf{x}}(\mathbf{x}|\mathbf{z}) p_{\theta_\mathbf{y}}(\mathbf{y}|\mathbf{z}) p(\mathbf{z}) d\mathbf{z}$. But, unlike as in these works, our domains $\mathbf{x}$ and $\mathbf{y}$ have similar structures and dimensions, and we suppose approximate posterior distributions will be represented in form of $q_\phi(\mathbf{z}|\mathbf{x})$ and $q_\phi(\mathbf{z}|\mathbf{y})$. The proposed model builds the joint probability $p(\mathbf{x}, \mathbf{y})$ of domains $\mathbf{x}$ and $\mathbf{y}$ that are conditioned independently on latent variable $\mathbf{z}$ (joint representation in the shared latent space).

Like [6] we propose to use relative constraints or learning triplets $\mathbf{t}$ to help our model catch domain characteristics and similarity between domains better. We sample these triplets from the shared latent space. Our joint probability takes the form of:

$$p(\mathbf{x}, \mathbf{y}, \mathbf{t}) = \int_{\mathbf{z}} p_{\theta_\mathbf{x}}(\mathbf{x}|\mathbf{z}) p_{\theta_\mathbf{y}}(\mathbf{y}|\mathbf{z}) p(\mathbf{t}_{i,j,k}|\mathbf{z}_{ijk}) p(\mathbf{z}) d\mathbf{z}.$$

We argue that the use of this implicit knowledge about the data provides slight regularization of the proposed model and improve the performance. We sample negative triplets' examples by using Jensen Shannon divergence as distance function between distributions during training and we suppose that on each training epoch the information from the triplets regularizes our objective.

We use the approximate posterior in the form of $q_\phi(\mathbf{z}|\mathbf{x})$ and $q_\phi(\mathbf{z}|\mathbf{y})$ because we want to solve the translation tasks — in images and languages. If we have a mapping between domains $f : \mathbf{x} \to \mathbf{y}$ and inverse mapping $g : \mathbf{y} \to \mathbf{x}$, then f and g should be inverse of each other. We want $g(f(\mathbf{x})) \approx \hat{\mathbf{x}}$ and $f(g(\mathbf{y})) \approx \hat{\mathbf{y}}$, where $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ are reconstructed input. At [26] these conditions are called *cycle consistency loss*.

It is worth to be mentioning that either in image-to-image translation or in machine translation tasks paired (or parallel) data is not always in sufficient quantities and obtaining such data can be difficult and in some tasks, like artistic style transfer, quite ambiguous. So we argue that the proposed model can translate between domains with slight supervision provided by triplets.

In Section (5) we describe the results on several different datasets and different tasks. The first dataset is MNIST [11], the second dataset is CelebA [13]. We show that our method is comparable with the previous methods on these datasets. We also show that it outperformes some of these methods methods.

The main contributions of this paper are the following:

- We introduce the Variational Bi-domain Triplet Autoencoder (VBTA) — new extension of variational autoencoder that trains a joint distribution of objects in different domains.

*Equal contribution.

KDD'18 Deep Learning Day, August 2018, London, UK

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

- We extend the VBTA objective with the learning triplets and propose negative sampling method that samples from the shared latent space purely unsupervised during training. We show that model with the triplets information improves the quality.
- We demonstrate the performance of the proposed model on different tasks such as bi-directional image generation, image-to-image translation, even on unpaired data and comparison with some previous methods. We also provide the qualitative analysis.

2 RELATED WORK

In this Section we consider some previous works that are close to ours, both in theoretical and practical sense.

Deep Generative Models. Various Deep Generative Models were proposed recently for many deep architectures. [9] introduced Variational Autoencoder, where it is assumed that the data is generated using some latent continuous random variable $\mathbf{z}$. In paper [8] extended the approach for semi-supervised settings. [2] presented a Recurrent Latent Variable Model for Sequential Data. [10] presented Deep Convolution Inverse Graphics Network and [4] proposed Generative Adversarial Nets.

Joint Models. Several works investigate joint models based on variational autoencoders in the similar way but in different training settings and tasks. VCCA objective was presented by [21] for multi-view representation learning. [19] introduced JMVAE model to represent different modalities, that are independently conditioned on joint representation. Also, the sampling process from $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$ and $q_\phi(\mathbf{z}|\mathbf{x})$ was showed, when $\mathbf{x}$ and $\mathbf{y}$ were different modalities. [20] presented an extension of joint VAE for multimodal setting and introduced the TELBO objective. However, [19] and [20] considered the task for modalities with different kind of structures (e.g. images and text attributes for this images).

Triplet learning. Many works investigate the metric learning approach, see [1], especially constructing the objective with the learning triplets: $\mathcal{T} = (x_i, x_j, x_k)$, where x_i should be more similar to x_j than to x_k in the sense of some distance function. [6] proposed the OPBN model with the VAEs objective extension with triplets. [15] sampled the triplets that are close to each other by Hamming distance. [23] sampled triplets from the training batches using combination of some strategies. The triplet loss for face recognition has been introduced by the paper [16]. They describe a new approach for training face embeddings using online triplet mining with different strategies.

Distributed representation learning. [14] demonstrated the potential of distributed representations for crosslingual case. In works [18, 25] bilingual autoencoder was demonstrated. Recent works by [17, 22] described the Variational Autoencoder for distributed representation learning, where variational distribution depends on both domains (languages) $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$.

Image-to-image translation. In our work we also consider image-to-image translation problem, where the goal of which to learn a mapping between an image from one domain to an image from another. The most common approach for this task is GAN modification [5] using Cycle-Consistent Adversarial Networks [26], DualGANs [24], Coupled GANs [12], Triangle GANs [3].

3 ASSUMPTIONS

Consider dataset $(\mathbf{X},\mathbf{Y}) = \{\mathbf{x},\mathbf{y}\}_{n=1}^N$ consisting of N *i.i.d.* objects from different domains. We assume that these objects are generated independently by the random process using the same latent variable $\mathbf{z}$.

We make an assumption that for each pair $(\mathbf{x},\mathbf{y})$ there exists a shared latent space variable $\mathbf{z}$, from which we can reconstruct both $\mathbf{x}$ and $\mathbf{y}$. Latent space variable $\mathbf{z}$ is built from the domain space variables $\mathbf{z}_x, \mathbf{z}_y$ according to equations:

$$\mathbf{z} = E(\mathbf{z}_x) = E\left(E_x(\mathbf{x})\right),$$

$$\mathbf{z} = E(\mathbf{z}_y) = E\left(E_y(\mathbf{y})\right),$$

where $\mathbf{z}_x$ and $\mathbf{z}_y$ are produced from $\mathbf{x}$ and $\mathbf{y}$ accordingly:

$$\mathbf{z}_x = E_x(\mathbf{x}),$$

$$\mathbf{z}_y = E_y(\mathbf{y}).$$

We define a shared intermediate variable $\mathbf{h}$, which is used to obtain corresponding domain variables $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$ from $\mathbf{y}$, $\mathbf{x}$ through $\mathbf{z}$:

$$\mathbf{h} = D(\mathbf{z}) = D\left(E(E_x(\mathbf{x}))\right),$$

$$\hat{\mathbf{y}} = D_y(\mathbf{h}) = D_y\left(D(E(E_x(\mathbf{x})))\right) = f(\mathbf{x}) \approx \mathbf{y},$$

$$\hat{\mathbf{x}} = D_x(\mathbf{h}) = D_x\left(D(E(E_y(\mathbf{x})))\right) = g(\mathbf{y}) \approx \mathbf{x}.$$

As it was mentioned in paper [26], the necessary condition for f and g to exist is the cycle-consistency constraint. That is, the proposed assumptions requires the cycle-consistency assumption.

The following diagram on Figure 1 presents VBTA generative process. Objects $\mathbf{z}_i$, $\mathbf{z}_i$ and $\mathbf{z}_k$ forme triplet.

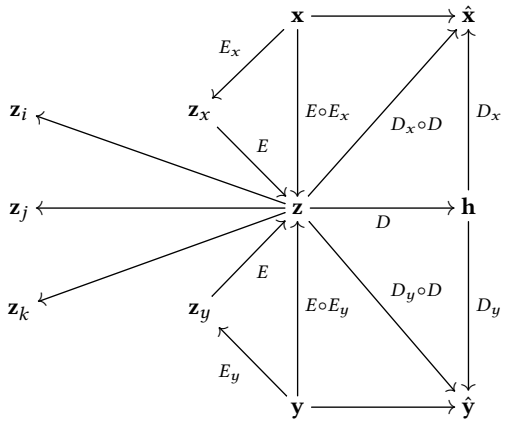

Figure 1: VBTA generative process

4 METHODS

4.1 Variational Autoencoder

Let the objects from $\mathbf{X} = \{\mathbf{x}\}_{i=1}^N$ be generated by condition on a latent variable $\mathbf{z} \sim p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$: $\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{z})$. The VAEs objective is to maximize the marginal likelihood:

$$\log p_\theta(\mathbf{x}) \geq -\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})) + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x}|\mathbf{z}), \quad (1)$$

where $q_\phi(\mathbf{z}|\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\sigma}^2_\phi(\mathbf{x}))$ is approximate distribution of posterior or encoder, $p_\theta(\mathbf{x}|\mathbf{z}) \sim \mathcal{N}(\boldsymbol{\mu}_\theta(\mathbf{z}), \boldsymbol{\sigma}^2_\theta(\mathbf{z}))$ is decoder. Both encoder and decoder are modelled by a neural network. To optimize the variational parameters θ and ϕ the reparametrization trick is used: $\mathbf{z} = \boldsymbol{\mu}_\phi(\mathbf{x}) + \boldsymbol{\sigma}^2_\phi(\mathbf{x}) \odot \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Therefore, gradients estimation with respect to θ and ϕ is $\nabla_{\theta,\phi} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x}|\mathbf{z}) =$
$= \mathbb{E}_{\mathcal{N}(\mathbf{0},\mathbf{I})} \nabla_{\theta,\phi} \log p_\theta(\mathbf{x}|\boldsymbol{\mu}_\phi(\mathbf{x}) + \boldsymbol{\sigma}^2_\phi(\mathbf{x}) \odot \boldsymbol{\epsilon})$.

4.2 Learning Triplets

Based on the metric learning approach and similar to [6] we extend our model by relative constraints or triplets:

$$\mathcal{T} = \{(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k) : d(\mathbf{z}_i, \mathbf{z}_j) < d(\mathbf{z}_i, \mathbf{z}_k)\}.$$

We define the conditional triplet likelihood in the following form:

$$p(t_{i,j,k} = True|i,j,k) = \\ = \int_{\mathbf{z}} p(\mathbf{t}_{i,j,k}|\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k) p(\mathbf{z}_i) p(\mathbf{z}_j) p(\mathbf{z}_k) d\mathbf{z}_i d\mathbf{z}_j d\mathbf{z}_k, \quad (2)$$

that was modelled by Bernoulli distribution over the states *True* and *False* parametrized with the use of softmax-function.

$$p(t_{i,j,k}|i,j,k) = \frac{e^{-d_{i,j}}}{e^{-d_{i,j}} + e^{-d_{i,k}}}, \quad (3)$$

with $d_{a,b} = \sum_{m=1}^M d^m_{a,b} = -\sum_{m=1}^M (\mathrm{JS}(p(\mathbf{z}^m_a) \parallel p(\mathbf{z}^m_b)))$, where $\mathbf{z} \in \mathbb{R}^m$,
$\mathrm{KL}(p(\mathbf{z}^m_a) \parallel p(\mathbf{z}^m_b)) = \int_{\mathbf{z}} p(\mathbf{z}_a) \log \frac{p(\mathbf{z}^m_a)}{p(\mathbf{z}^m_b)} d\mathbf{z}$.
For d the following approximation is used:

$$d_{a,b} = \sum_{m=1}^M d^m_{a,b} = \\ = -\sum_{m=1}^M \Big[\frac{1}{2}\mathrm{KL}(p(\mathbf{z}^m_a) \parallel p(\mathbf{z}^m_b)) + \frac{1}{2}\mathrm{KL}(p(\mathbf{z}^m_b) \parallel p(\mathbf{z}^m_a))\Big]. \quad (4)$$

4.3 Variational Bi-domain Triplet Autoencoder

Consider the dataset $(\mathbf{X}, \mathbf{Y}) = \{\mathbf{x}, \mathbf{y}\}_{n=1}^N$ consisting of N *i.i.d.* objects from different domains. We assume that these objects are generated independently by the random process the use of the same latent variable $\mathbf{z}$. We consider the following formulation as *bi-domain generative model* using triplet information:

$$p(\mathbf{x}, \mathbf{y}, \mathbf{t}) = \int_{\mathbf{z}} \prod_n p_{\theta_\mathbf{x}}(\mathbf{x}|\mathbf{z}) p_{\theta_\mathbf{y}}(\mathbf{y}|\mathbf{z}) p(\mathbf{t}_{i,j,k}|\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k) p(\mathbf{z}) d\mathbf{z} \quad (5)$$

Based on (3) we consider approximate posterior distributions to be $q_{\phi_\mathbf{x}}(\mathbf{z}|\mathbf{x})$, $q_{\phi_\mathbf{y}}(\mathbf{z}|\mathbf{y})$ and $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y})$ and estimate the lower bound of the log-likelihood as follows:

$$\mathcal{L}_{\mathrm{VBTA}}(\mathbf{x}, \mathbf{y}, \mathbf{t}) = \mathbb{E}_{q_{\phi_\mathbf{x}}(\mathbf{z}|\mathbf{x})}\Big[\log \frac{p(\mathbf{x}, \mathbf{z})}{q_{\phi_\mathbf{x}}(\mathbf{z}|\mathbf{x})}\Big] + \mathbb{E}_{q_{\phi_\mathbf{y}}(\mathbf{z}|\mathbf{y})}\Big[\log \frac{p(\mathbf{y}, \mathbf{z})}{q_{\phi_\mathbf{y}}(\mathbf{z}|\mathbf{y})}\Big] + \\ + \mathbb{E}_{q_{\phi_\mathbf{x}}(\mathbf{z}|\mathbf{x}), q_{\phi_\mathbf{y}}(\mathbf{z}|\mathbf{y})}\Big[\log \frac{p(\mathbf{t}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{t})}\Big] = \\ = -\alpha\Big[\mathrm{KL}(q_{\phi_\mathbf{x}(\mathbf{z}|\mathbf{x})}(\mathbf{z}|\mathbf{x}) \parallel p_{\theta_\mathbf{x}}(\mathbf{z})) + \mathrm{KL}(q_{\phi_\mathbf{y}(\mathbf{z}|\mathbf{y})}(\mathbf{z}|\mathbf{y}) \parallel p_{\theta_\mathbf{y}}(\mathbf{z}))\Big] + \\ + \beta\Big[\mathbb{E}_{q_{\phi_\mathbf{x}}(\mathbf{z}|\mathbf{x})}\big[\log p_{\theta_\mathbf{x}}(\mathbf{x}|\mathbf{z})\big] + \mathbb{E}_{q_{\phi_\mathbf{y}}(\mathbf{z}|\mathbf{y})}\big[\log p_{\theta_\mathbf{y}}(\mathbf{y}|\mathbf{z})\big]\Big] + \\ + \gamma \mathbb{E}_{q_{\phi_\mathbf{x}}(\mathbf{z}|\mathbf{x}), q_{\phi_\mathbf{y}}(\mathbf{z}|\mathbf{y})}\big[\log p(\mathbf{t}|\mathbf{z})\big]. \quad (6)$$

Here, both $q_{\phi_\mathbf{x}(\mathbf{z}|\mathbf{x})}(\mathbf{z}|\mathbf{x})$ and $q_{\phi_\mathbf{y}(\mathbf{z}|\mathbf{y})}(\mathbf{z}|\mathbf{y})$ are encoders, $p_{\theta_\mathbf{x}}(\mathbf{x}|\mathbf{z})$ and $p_{\theta_\mathbf{y}}(\mathbf{y}|\mathbf{z})$ are decoders, modeled by the deep neural networks. Similar to [12] our decoders and encoders use the common functions E and D, see (3). For detailed networks architecture, applying to each considering task see (5.3). The term $p(\mathbf{t}|\mathbf{z})$ is estimated by (4).

We apply the Stochastic Gradient Variational Bayes (SGVB) to (6) and optimize the variational parameters $\theta_\mathbf{x}$, $\theta_\mathbf{y}$, $\phi_\mathbf{x}$ and $\phi_\mathbf{y}$.

5 EXPERIMENTS

We present the results on an image-to-image translation task for two datasets: MNIST [11] and CelebA [13]. For the MNIST dataset we obtain the quantitative results and compared proposed method with GANs [3] and JMVAE [19]. For the CelebA dataset we provid an image-to-image translation considering CelebA as a set of two image domains: faces of men and women. We inspect the results of man-to-woman and woman-to-man translation.

5.1 Datasets

We used MNIST dataset for toy problem of image-to-image translation. Similar to [3] we considered a transposition of this dataset as a second domain $\mathbf{y}$. We used 50,000 as training set and the remaining 10,000 as a test set.

CelebA consists of 202,599 face images with 40 binary attributes. In this work we considered this dataset as a union of two domains: faces of men $\mathbf{x}$ and faces of women $\mathbf{y}$. Similar to [19] we cropped and normalized the images and resized them to 64x64.

5.2 Sampling methods

For all the experiments we selected the negative (the closest object except paired) triplet examples from domain $\mathbf{y}$ with the minimal Jensen-Shannon divergence with to the corresponding objects from domain $\mathbf{x}$:

$$\mathbf{z}_k = \underset{\mathbf{z}_{i'} \in \mathcal{S}_b \setminus (\mathbf{z}_i, \mathbf{z}_j)}{\operatorname{argmin}} \mathrm{JS}(\mathbf{z}_i, \mathbf{z}_{i'}),$$

where $\mathcal{S}_b \in \mathcal{S}$ — current mini-batch, $\mathbf{z}_i$ and $\mathbf{z}_j$ — the paired objects from different domains. That is, we wanted to choose an example $\mathbf{z}_k$ that is similar to $\mathbf{z}_i$ according to the current model parameters.

For the MNIST dataset as a pair of similar images we used an image and its transposition:

$$\mathbf{z}_i, \mathbf{z}_j = \{q_{\phi_\mathbf{x}(\mathbf{z}|\mathbf{x})}(\mathbf{z}|\mathbf{x}), q_{\phi_\mathbf{y}(\mathbf{z}|\mathbf{y})}(\mathbf{z}|\mathbf{y}) : \mathbf{y} = \mathbf{x}^\mathrm{T}\}.$$

Since we did not have any paired men and women in CelebA dataset, we considered that the object $\mathbf{y}$ (women) is similar to object $\mathbf{x}$ (men) if they had the largest matching of their attributes.

5.3 Model Architecture

For the MNIST dataset we used one-layer network of 512 hidden units with ReLU for decoder D and encoders E_x, E_y. For the modeling shared encoder E and decoder D_x, D_y we used the linear mappings. The shared latent space dimension was set to 64.

For the classification evaluation we set $p_{\theta_x}(\mathbf{x}|\mathbf{z})$ and $p_{\theta_y}(\mathbf{y}|\mathbf{z})$ to be Gaussian distribution. For the comparison to JMVAE model we set $p_{\theta_x}(\mathbf{x}|\mathbf{z})$ and $p_{\theta_y}(\mathbf{y}|\mathbf{z})$ to be Bernoulli. We set model of JMVAE to the same configuration.

For CelebA we used encoders E_x, E_y with two convolution layers and a flattened layer with ReLU. For the shared encoder E and decoder D_x, D_y we used linear mapping into 64 hidden units. For the decoder D we used a network with one dense layer with 8192 units and a deconvolution layer. We considered $p_{\theta_x}(\mathbf{x}|\mathbf{z})$ and $p_{\theta_y}(\mathbf{y}|\mathbf{z})$ as a Gaussian distribution.

We used the Adam [7] optimization algorithm with a learning rate of 10^{-3} for the MNIST dataset and 10^{-4} for CelebA dataset. All the models were trained for 100 epochs with batch size set of 50.

6 RESULTS

6.1 Image-to-image translation for MNIST dataset

Following [3] we firstly evaluated our approach on MNIST-transpose, where the two image domains $\mathbf{x}$ and $\mathbf{y}$ are the MNIST images and their corresponding transposed ones. Similar to [3] we used the classifier that trained on MNIST images as a ground-truth evaluator. For all the transposed images we encoded them via our model encoder $E \circ E_{\mathbf{y}}$ and decoded via decoder $D_{\mathbf{x}} \circ D$. Then we sent them to the classifier. The results of the classification are shown in Table 1. We showed that the triplet information improves the quality significantly, see the last row, $n = 0$ and $n = 10$, where n is the number of objects used for triplets sampling. As we can see, our approach gives classification results comparable to the state-of-the-art GAN model results. The intermediate results of the proposed method are illustrated in Figure 2. Figure 3 shows PCA vizualization on MNIST dataset. The right Figure shows the projection of the translated version of MNIST-transpose projected using the same PCA model. As we can see, the translation function $f(\mathbf{x})$ preserves the latent information of the dataset.

Table 1: Classification accuracy (%) on the MNIST-transpose dataset. The DiscoGAN, Triple GAN and Δ-GAN results are taken from [3]

Model	$n = 0$	$n = 10$	$n = 100$	$n = 1000$	All
DiscoGAN	-	-	-	-	15.00
	-	-	-	-	±0.20
Triple GAN	-	-	63.79	84.93	86.70
	-	-	±0.85	±1.63	±1.52
Δ-GAN	-	-	83.20	88.98	**93.34**
	-	-	±1.88	±1.50	**±1.46**
Proposed	17.07	59.09	**90.74**	**90.83**	90.78
	±3.59	±17.78	**±0.26**	**±0.34**	±0.31

As we can see, the performance of both our methods is comparable when the number of used labels is large. However, the model

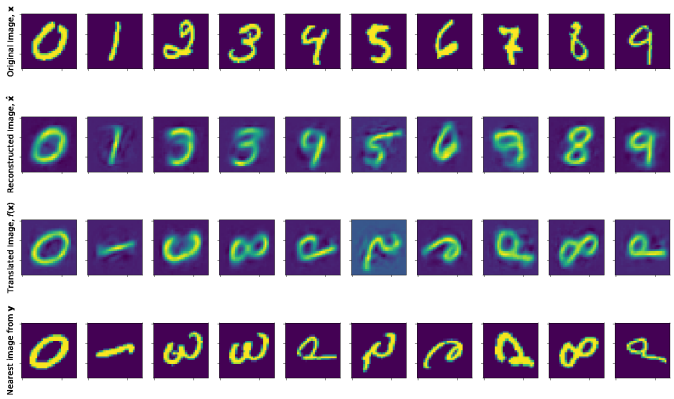

Figure 2: Intermediate results of training model for 10 epochs. As we can see, the digit "2" is purely reconstructed and similar to "3". Therefore the corresponding negative sample from domain y is chosen to distinguish them.

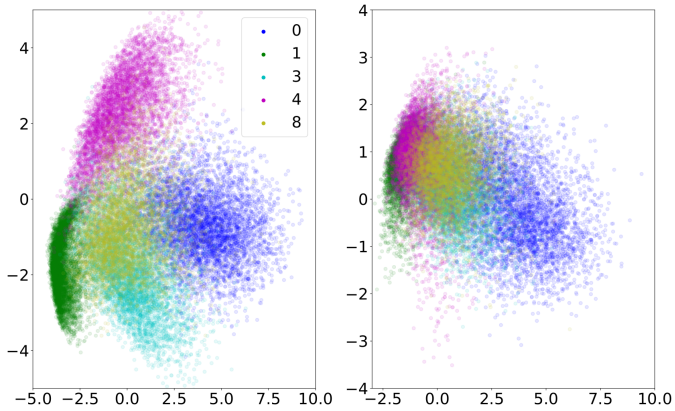

Figure 3: PCA projection of the dataset y (left) and the translation of Y, i.e. $g(\mathrm{y})$ (right). In both cases the PCA model was optimized only using the dataset y.

with shared decoder gives us good quality with a few examples. We argue that this is connected with the simplicity of translation between MNIST and its transposed version.

We further evaluated the marginal log-likelihood of our model on binarized versions of MNIST and MNIST-transpose and compared it with two models: variational autoencoder and JMVAE. The results are listed in Table 2. The results show that the marginal log-likelihood of our model is slightly better than the likelihood of the variational autoencoder, because the latent space was shared among two domains.

6.2 Qualitative results for CelebA dataset

In this Section we confirm that our method can generate images and translate them between two domains. Figure 4 shows face images from datasets and their translation into different domains.

Table 2: Marginal log-likelihood for MNIST as log$p$(x) and MNIST-transpose datasets as log$p$(y). The JVMAE results are taken from [19]. For VAE results we tested standart VAE (1).

Model	$< logp(\mathbf{x})$	$< logp(\mathbf{y})$
VAE	-81.13	-81.01
JVMAE	-85.35	-85.44
Proposed	**−80.92**	**−80.91**

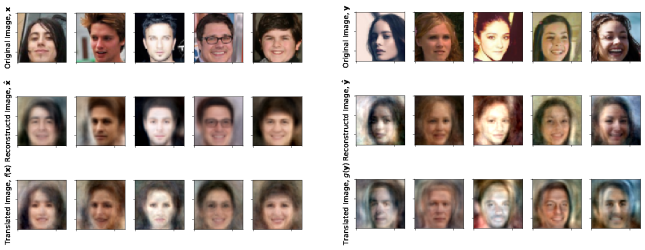

Figure 4: Results of image-to-image translation for CelebA dataset. The first row corresponds to the original images that were considered as similar because of high amount of matching attrbutes. The second row shows the reconstruction of the images. The third row illustrates the image translation from domain x into domain y and from y into x.

Figure 5 shows faces generated from Gaussian distribution. We found that our algorithm works well enought and can reproduce similar faces for both domains from one sample in latent space.

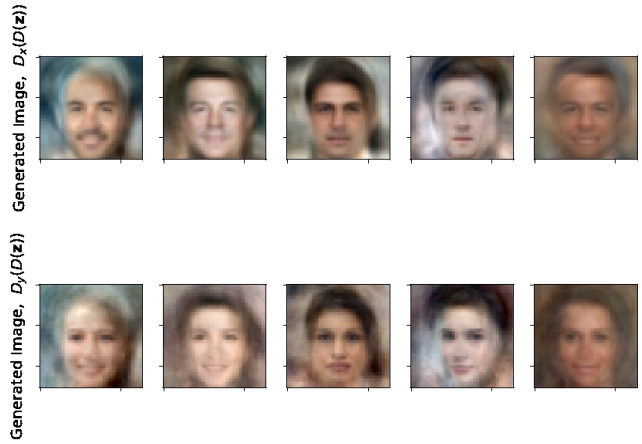

Figure 5: Results of image generation from the common shared space. Each column corresponds to the faces generated from one sample of z. The latent variable z was sampled from Gaussian distribution: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

7 CONCLUSION AND FUTURE WORK

In this paper we proposed the Variational Bi-domain Triplet Autoencoder (VBTA) that learns a joint distribution of objects from different domains. We consider that domains have similar structure (texts, images). The proposed model built the joint probability of domains, that are conditioned independently on a latent variable. We extended the VBTAs objective function by the relative constraints or triplets that sampled from the shared latent space across domains. We did that extension to deal with the task when there is no possibility to obtain labeled data.

We demonstrated the performance of the VBTA model on different tasks: bi-directional image generation, image-to-image translation, even on unpaired data. We also provided the qualitative analysis. We showed that VBTA model is comparable and outperforms some of the existing generative models.

In future work we would like to provide experiments on text data and extend the model on more than one domain.

REFERENCES

[1] Aurélien Bellet, Amaury Habrard, and Marc Sebban. 2013. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709* (2013).

[2] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. 2015. A Recurrent Latent Variable Model for Sequential Data. In *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.). 2980–2988.

[3] Zhe Gan, Liqun Chen, Weiyao Wang, Yunchen Pu, Yizhe Zhang, Hao Liu, Chunyuan Li, and Lawrence Carin. 2017. Triangle Generative Adversarial Networks. *CoRR* abs/1709.06548 (2017). arXiv:1709.06548 http://arxiv.org/abs/1709.06548

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). 2672–2680.

[5] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2016. Image-to-Image Translation with Conditional Adversarial Networks. *CoRR* abs/1611.07004 (2016). arXiv:1611.07004 http://arxiv.org/abs/1611.07004

[6] Theofanis Karaletsos, Serge Belongie, and Gunnar Rätsch. 2015. Bayesian representation learning with oracle constraints. *arXiv preprint arXiv:1506.05011* (2015).

[7] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[8] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised Learning with Deep Generative Models. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). 3581–3589.

[9] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[10] Tejas D Kulkarni, William F. Whitney, Pushmeet Kohli, and Josh Tenenbaum. 2015. Deep Convolutional Inverse Graphics Network. In *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.). 2539–2547.

[11] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[12] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised Image-to-Image Translation Networks. *CoRR* abs/1703.00848 (2017). arXiv:1703.00848 http://arxiv.org/abs/1703.00848

[13] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*. 3730–3738.

[14] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168* (2013).

[15] Mohammad Norouzi, David J Fleet, and Ruslan R Salakhutdinov. 2012. Hamming Distance Metric Learning. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.). 1061–1069.

[16] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[17] Jinsong Su, Shan Wu, Biao Zhang, Changxing Wu, Yue Qin, and Deyi Xiong. 2018. A neural generative autoencoder for bilingual word embeddings. *Information*

Sciences 424 (2018), 287–300.

[18] Jinsong Su, Deyi Xiong, Biao Zhang, Yang Liu, Junfeng Yao, and Min Zhang. 2015. Bilingual correspondence recursive autoencoder for statistical machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 1248–1258.

[19] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. 2016. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891* (2016).

[20] Ramakrishna Vedantam, Ian Fischer, Jonathan Huang, and Kevin Murphy. 2017. Generative models of visually grounded imagination. *arXiv preprint arXiv:1705.10762* (2017).

[21] Weiran Wang, Honglak Lee, and Karen Livescu. 2016. Deep Variational Canonical Correlation Analysis. *CoRR* abs/1610.03454 (2016). arXiv:1610.03454 http://arxiv.org/abs/1610.03454

[22] Liangchen Wei and Zhi-Hong Deng. 2017. A Variational Autoencoding Approach for Inducing Cross-lingual Word Embeddings. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17)*. 4165–4171.

[23] John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198* (2015).

[24] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. 2017. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. *CoRR* abs/1704.02510 (2017). arXiv:1704.02510 http://arxiv.org/abs/1704.02510

[25] Biao Zhang, Deyi Xiong, and Jinsong Su. 2016. BattRAE: Bidimensional Attention-Based Recursive Autoencoders for Learning Bilingual Phrase Embeddings. *CoRR* abs/1605.07874 (2016). arXiv:1605.07874 http://arxiv.org/abs/1605.07874

[26] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *CoRR* abs/1703.10593 (2017). arXiv:1703.10593 http://arxiv.org/abs/1703.10593