Variational learning across domains with triplet information

Rita Kuznetsova[1,2], **Oleg Bakhteev**[1,2] **and Alexandr Ogaltsov**[2,3]
[1]Moscow Institute of Physics and Technology
[2]Antiplagiat Company
[3]National Research University Higher School of Economics
{rita.kuznetsova, bakhteev}@phystech.edu, avogaltsov@edu.hse.ru

Abstract

The work investigates deep generative models, which allow us to use training data from one domain to build a model for another domain. We propose the Variational Bi-domain Triplet Autoencoder (VBTA) that learns a joint distribution of objects from different domains. We extend the VBTAs objective function by the relative constraints or triplets that sampled from the shared latent space across domains. In other words, we combine the *deep generative models* with a *metric learning* ideas in order to improve the final objective with the triplets information. The performance of the VBTA model is demonstrated on different tasks: image-to-image translation, bi-directional image generation and cross-lingual document classification.

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) Kingma and Welling [2013] 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 Karaletsos et al. [2015], Kingma et al. [2014], Suzuki et al. [2016] 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 Liu et al. [2017] 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) we define the joint probability of the proposed model. Our domains $\mathbf{x}$ and $\mathbf{y}$ have similar structures and dimensions, and we suppose approximate posterior distributions will be represented in form of $q_{\phi_{\mathbf{x}}(\mathbf{z}_x|\mathbf{x})}(\mathbf{z}_x|\mathbf{x})$ and $q_{\phi_{\mathbf{y}}(\mathbf{z}_y|\mathbf{y})}(\mathbf{z}_y|\mathbf{y})$. 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 Karaletsos et al. [2015] 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. 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

Third workshop on Bayesian Deep Learning (NeurIPS 2018), Montréal, Canada.

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 $f : \mathbf{x} \rightarrow \mathbf{y}$ and inverse mapping $g : \mathbf{y} \rightarrow \mathbf{x}$, 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 Zhu et al. [2017] 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. We proved this in (6.3), where we do not use parallel corpora for our algorithm.

In Section (6) we describe the results on several different datasets and different tasks. The first dataset is MNIST LeCun et al. [1998], the second dataset is CelebA Liu et al. [2015], the third is RCV1/RCV2 corpora Lewis et al. [2004]. 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 across domains with learning triplet information. We propose negative sampling method that samples from the shared latent space purely unsupervised during training.
- We demonstrate the performance of the proposed model on different tasks such as bi-directional image generation, image-to-image translation, cross-lingual document classification.

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. Kingma and Welling [2013] introduced Variational Autoencoder, where it is assumed that the data is generated using some latent continuous random variable $\mathbf{z}$. In paper Kingma et al. [2014] extended the approach for semi-supervised settings. Chung et al. [2015] presented a Recurrent Latent Variable Model for Sequential Data. Kulkarni et al. [2015] presented Deep Convolution Inverse Graphics Network and Goodfellow et al. [2014] 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 Wang et al. [2016] for multi-view representation learning. Suzuki et al. [2016] 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. Vedantam et al. [2017] presented an extension of joint VAE for multimodal setting and introduced the TELBO objective. However, Suzuki et al. [2016] and Vedantam et al. [2017] 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 Bellet et al. [2013], 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. Karaletsos et al. [2015] proposed the OPBN model with the VAEs objective extension with triplets. Norouzi et al. [2012] sampled the triplets that are close to each other by Hamming distance. Wieting et al. [2015] sampled triplets from the training batches using combination of some strategies. The triplet loss for face recognition has been introduced by the paper Schroff et al. [2015]. They describe a new approach for training face embeddings using online triplet mining with different strategies.

Distributed representation learning Mikolov et al. [2013] demonstrated the potential of distributed representations for crosslingual case. In works Su et al. [2015], Zhang et al. [2016] bilingual autoencoder was demonstrated. Recent works by Wei and Deng [2017], Su et al. [2018b] 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 Isola et al. [2016] using Cycle-Consistent Adversarial Networks Zhu et al. [2017], DualGANs Yi et al. [2017], Coupled GANs Liu et al. [2017], Triangle GANs Gan et al. [2017].

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 $\mathbf{h}_x$, $\mathbf{h}_y$ according to equations: $\mathbf{z} = E(\mathbf{h}_{x_i}) = E\left(E_x(\mathbf{x}_i)\right)$, $\mathbf{z} = E(\mathbf{h}_{y_j}) = E\left(E_y(\mathbf{y}_j)\right)$, where $\mathbf{h}_{x_i}$ and $\mathbf{h}_{y_j}$ are produced from $\mathbf{x}_i$ and $\mathbf{y}_j$ accordingly: $\mathbf{h}_{x_i} = E_x(\mathbf{x}_i)$, $\mathbf{h}_{y_j} = E_y(\mathbf{y}_j)$. We define a shared intermediate variable $\mathbf{h}$, which is used to obtain corresponding domain variables $\hat{\mathbf{x}}_i$, $\hat{\mathbf{y}}_j$ from $\mathbf{y}_j$, $\mathbf{x}_i$ through $\mathbf{z}$: $\mathbf{h} = D(\mathbf{z}) = D\left(E(E_x(\mathbf{x}_i))\right) = D\left(E(E_y(\mathbf{y}_j))\right)$.

$\hat{\mathbf{y}}_j = D_y(\mathbf{z}) = D_y\left(D(E(E_x(\mathbf{x}_i)))\right) = f(\mathbf{x}_i) \approx \mathbf{y}_j, \hat{\mathbf{x}}_i = D_x(\mathbf{z}) = D_x\left(D(E(E_y(\mathbf{x}_i)))\right) = g(\mathbf{y}_j) \approx \mathbf{x}_i.$

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 $\mathbf{z}_k$ form triplet.

$$\begin{tikzcd}
 & & \mathbf{x}_i \arrow[r] \arrow[dl, "E_x"'] \arrow[dd, "E \circ E_x"] & \hat{\mathbf{x}}_i \\
\mathbf{z}_i & \mathbf{h}_{x_i} \arrow[dr, "E"] & & \\
\mathbf{z}_j & & \mathbf{z} \arrow[ull] \arrow[ll] \arrow[dll] \arrow[r, "D"'] \arrow[uur, "D_x \circ D"'] \arrow[ddr, "D_y \circ D"] & \mathbf{h} \arrow[uu, "D_x"'] \arrow[dd, "D_y"] \\
\mathbf{z}_k & \mathbf{h}_{y_j} \arrow[ur, "E"'] & & \\
 & & \mathbf{y}_j \arrow[r] \arrow[ul, "E_y"] \arrow[uu, "E \circ E_y"'] & \hat{\mathbf{y}}_j
\end{tikzcd}$$

Figure 1: VBTA generative process

4 Variational Bi-domain Triplet Autoencoder

The marginal likelihood defined by this model is:

$$p(\mathbf{x}, \mathbf{y}, \mathbf{t}) = \int_{\mathbf{z}} p_{\theta_{\mathbf{x}}}(\mathbf{x}|\mathbf{z}) p_{\theta_{\mathbf{y}}}(\mathbf{y}|\mathbf{z}) p(\mathbf{t}_{i,j,k}|\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k) p(\mathbf{z}) d\mathbf{z} \tag{1}$$

We can assume the following generative process:

- generate $\mathbf{z}$ from prior distribution $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$,

- value $\mathbf{x}$ is generated from some conditional distribution $p_{\theta_{\mathbf{x}}}(\mathbf{x}|\mathbf{z})$,
- value $\mathbf{y}$ is generated from some conditional distribution $p_{\theta_{\mathbf{y}}}(\mathbf{y}|\mathbf{z})$.

The lower bound of the log-likelihood:

$$
\begin{aligned}
\mathcal{L}_{VBTA} &= \mathbb{E}_{q_{\phi_x}(\mathbf{z}_x|\mathbf{x})} \log \frac{p_{\theta_x}(\mathbf{x},\mathbf{y},\mathbf{t},\mathbf{z}_x)}{q_{\phi_x}(\mathbf{z}_x|\mathbf{x})} + \mathbb{E}_{q_{\phi_y}(\mathbf{z}_y|\mathbf{y})} \log \frac{p_{\theta_y}(\mathbf{x},\mathbf{y},\mathbf{t},\mathbf{z}_y)}{q_{\phi_y}(\mathbf{z}_y|\mathbf{y})} = \\
&= \underbrace{-\Big[KL\big(q_{\phi_{\mathbf{x}}(\mathbf{z}_x|\mathbf{x})}(\mathbf{z}_x|\mathbf{x}) \parallel p_{\theta_{\mathbf{x}}}(\mathbf{z}_x)\big) + KL\big(q_{\phi_{\mathbf{y}}(\mathbf{z}_y|\mathbf{y})}(\mathbf{z}_y|\mathbf{y}) \parallel p_{\theta_{\mathbf{y}}}(\mathbf{z}_y)\big)\Big]}_{\text{Penalty}} + \\
&\quad + \underbrace{\Big[\mathbb{E}_{q_{\phi_{\mathbf{x}}}(\mathbf{z}_x|\mathbf{x})}\big[\log p_{\theta_{\mathbf{x}}}(\mathbf{x}|\mathbf{z}_x)\big] + \mathbb{E}_{q_{\phi_{\mathbf{y}}}(\mathbf{z}_y|\mathbf{y})}\big[\log p_{\theta_{\mathbf{y}}}(\mathbf{y}|\mathbf{z}_y)\big]\Big]}_{\text{Reconstruction}} + \\
&\quad + \underbrace{\Big[\mathbb{E}_{q_{\phi_{\mathbf{x}}}(\mathbf{z}_x|\mathbf{x})}\big[\log p_{\theta_{\mathbf{x}}}(\mathbf{y}|\mathbf{z}_x)\big] + \mathbb{E}_{q_{\phi_{\mathbf{y}}}(\mathbf{z}_y|\mathbf{y})}\big[\log p_{\theta_{\mathbf{y}}}(\mathbf{x}|\mathbf{z}_y)\big]\Big]}_{\text{Cycle-consistency}} + \\
&\quad + \underbrace{\mathbb{E}_{q_{\phi_{\mathbf{x}}}(\mathbf{z}_x|\mathbf{x})}\big[\log p(\mathbf{t}|\mathbf{z}_x)\big] + \mathbb{E}_{q_{\phi_{\mathbf{y}}}(\mathbf{z}_y|\mathbf{x})}\big[\log p(\mathbf{t}|\mathbf{z}_y)\big]}_{\text{Triplet likelihood}} \qquad (2)
\end{aligned}
$$

Both $q_{\phi_{\mathbf{x}}(\mathbf{z}_x|\mathbf{x})}(\mathbf{z}_x|\mathbf{x})$ and $q_{\phi_{\mathbf{y}}(\mathbf{z}_y|\mathbf{y})}(\mathbf{z}_y|\mathbf{y})$ are encoders, $p_{\theta_{\mathbf{x}}}(\mathbf{x}|\mathbf{z}_x)$ and $p_{\theta_{\mathbf{y}}}(\mathbf{y}|\mathbf{z}_y)$ are decoders, modeled by the deep neural networks. Similar to Liu et al. [2017] our decoders and encoders use the common functions E and D, see (3). We apply the Stochastic Gradient Variational Bayes (SGVB) and optimize the variational parameters $\theta_{\mathbf{x}}$, $\theta_{\mathbf{y}}$, $\phi_{\mathbf{x}}$ and $\phi_{\mathbf{y}}$.

5 Learning Triplets

Based on the metric learning approach and similar to Karaletsos et al. [2015] we extend our model by relative constraints or triplets: $\mathcal{T} = \{(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k) : d(\mathbf{z}_i, \mathbf{z}_j) < d(\mathbf{z}_i, \mathbf{z}_k)\}$, but in our case we sampled triplets across domains $\mathbf{X}$ and $\mathbf{Y}$. We define the conditional triplet likelihood in the following form:

$$
p(t_{i,j,k} = True|i,j,k) = \int_{\mathbf{z}} p(\mathbf{t}_{i,j,k}|\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k) p(\mathbf{z}_i) p(\mathbf{z}_j) p(\mathbf{z}_k) d\mathbf{z}_i d\mathbf{z}_j d\mathbf{z}_k, \qquad (3)
$$

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

$$
p(t_{i,j,k}|i,j,k) = \frac{e^{-d(\mathbf{z}_i,\mathbf{z}_j)}}{e^{-d(\mathbf{z}_i,\mathbf{z}_j)} + e^{-d(\mathbf{z}_i,\mathbf{z}_k)}} \qquad (4)
$$

Triplets — three objects from shared latent space $\mathbf{z}$. $\mathbf{z}_i$, $\mathbf{z}_j$ — shared latent representation of objects from $\mathbf{X}$ and $\mathbf{Y}$ domains. The third object $\mathbf{z}_k$ is sampled from domain $\mathbf{y}$ with the minimal distance function to the corresponding objects from domain $\mathbf{x}$ (and vice versa):

$$
\mathbf{z}_k = \operatorname*{arg\,min}_{\mathbf{z}_{i'} \in \mathcal{S}_b \setminus (\mathbf{z}_i, \mathbf{z}_j)} d(\mathbf{z}_i, \mathbf{z}_{i'}), \qquad (5)
$$

where $\mathcal{S}_b \in \mathcal{S}$ — current mini-batch, $\mathbf{z}_i$ and $\mathbf{z}_j$ — the paired objects from different domains. As d we use approximate form of JS-divergence, like Karaletsos et al. [2015]. In other words, we want to choose an example $\mathbf{z}_k$ that is similar to $\mathbf{z}_i$ according to the current model parameters.

6 Experiment and Results

We presented the results on an image-to-image translation task: MNIST LeCun et al. [1998] and CelebA Liu et al. [2015]. We presented results on cross-lingual text classification task on RCV1/RCV2 corpora Lewis et al. [2004].

6.1 Image-to-image translation for MNIST dataset

We 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. 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 the linear mappings are used. The shared latent space dimension was set to 64. Training set consist 50,000 objects and the test set consist 10,000.

Similar to Gan et al. [2017] we used the classifier that trained on MNIST images as a ground-truth evaluator. For the classification evaluation we set $p_{\theta_\mathbf{x}}(\mathbf{x}|\mathbf{z}_x)$ and $p_{\theta_\mathbf{y}}(\mathbf{y}|\mathbf{z}_y)$ to be Gaussian distribution. For all the transposed images we encoded them via the model encoder $E \circ E_\mathbf{y}$ and decoded via decoder $D_\mathbf{x} \circ D$. Then we classified it. The results of the classification are shown in Table 1, where n is the number of objects used for triplets sampling and cycle-consistency.

Table 1: Classification accuracy (%) on the MNIST-transpose dataset. The DiscoGAN, Triple GAN and Δ-GAN results are taken from Gan et al. [2017]

Model	$n = 0$	$n = 10$	$n = 100$	$n = 1000$	All
DiscoGAN	-	-	-	-	15.00 ± 0.20
Triple GAN	-	-	63.79 ± 0.85	84.93 ± 1.63	86.70 ± 1.52
Δ-GAN	-	-	83.20 ± 1.88	88.98 ± 1.50	93.34 ± 1.46
VBTA	18.89 ± 3.59	86.57 ± 6.338	$\mathbf{90.44 \pm 0.3}$	$\mathbf{90 \pm 0.26}$	$\mathbf{95 \pm 0.06}$

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.

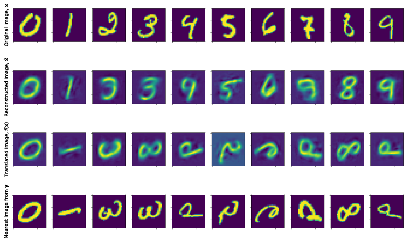

Figure 2: Intermediate results of training model for 10 epochs. As we can see, the digit “2” is purely reconstructed and similar to “3”. Therefore the corresponding negative sample from domain $\mathbf{y}$ is chosen to distinguish them.

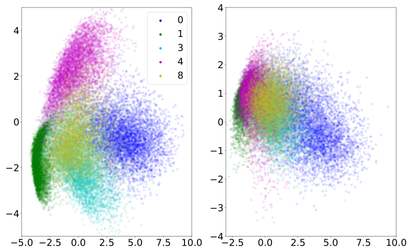

Figure 3: PCA projection of the dataset $\mathbf{y}$ (left) and the translation of $\mathbf{Y}$, i.e. $g(\mathbf{y})$ (right). In both cases the PCA model was optimized only using the dataset $\mathbf{y}$.

We evaluated the marginal log-likelihood of our model on binarized versions of MNIST and MNIST-transpose. The results are listed in Table 2. For the comparison to JMVAE model we set $p_{\theta_\mathbf{x}}(\mathbf{x}|\mathbf{z})$ and $p_{\theta_\mathbf{y}}(\mathbf{y}|\mathbf{z})$ to be Bernoulli. We set model of JMVAE to the same configuration.

Table 2: Marginal log-likelihood for MNIST as log$p(\mathbf{x})$ and MNIST-transpose datasets as log$p(\mathbf{y})$.

Model	$< logp(\mathbf{x})$	$< logp(\mathbf{y})$
VAE Kingma and Welling [2013]	-81.13	-81.01
JVMAE Suzuki et al. [2016]	-85.35	-85.44
VBTA	$\mathbf{-80.92}$	$\mathbf{-80.91}$

6.2 Qualitative results for CelebA dataset

CelebA consists of 202,599 face images with 40 binary attributes. In this section we considered this dataset as a union of two domains: faces of men $\mathbf{X}$ and faces of women $\mathbf{Y}$. Similar to Suzuki et al. [2016] we cropped and normalized the images and resized them to 64x64. 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.

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 $p_{\theta_{\mathbf{x}}}(\mathbf{x}|\mathbf{z})$ and $p_{\theta_{\mathbf{y}}}(\mathbf{y}|\mathbf{z})$ as a Gaussian distribution. Figure 4 shows the face images from datasets and their translation into different domains.

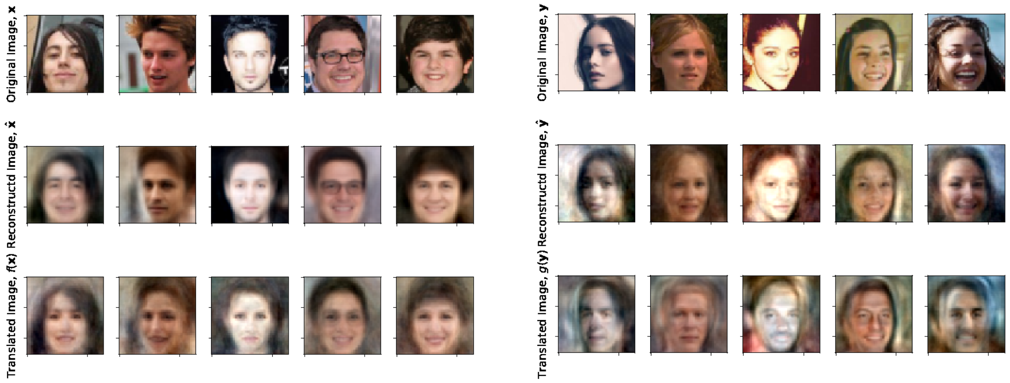
Original Image, $\mathbf{x}$
Reconstructd Image, $\hat{\mathbf{x}}$
Translated Image, $f(\mathbf{x})$
Original Image, $\mathbf{y}$
Reconstructd Image, $\hat{\mathbf{y}}$
Translated Image, $g(\mathbf{y})$

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 $\mathbf{X}$ into domain $\mathbf{Y}$ and from $\mathbf{Y}$ into $\mathbf{X}$.

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

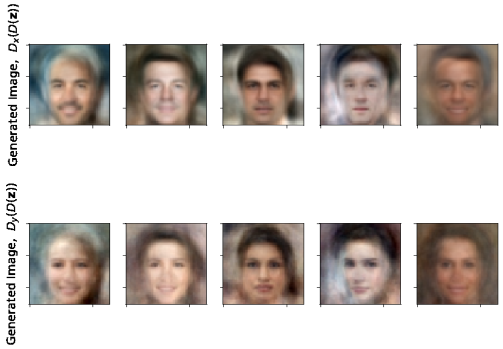
Generated Image, $D_x(D(\mathbf{z}))$
Generated Image, $D_y(D(\mathbf{z}))$

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

6.3 Cross Lingual Document Classification

We use experimental setup similar to introduced in Klementiev et al. [2012]. Given a classifier trained on documents in language *A* ($\mathbf{X}$ domain), one should use that classifier to predict labels of documents in language *B* ($\mathbf{Y}$ domain).

Previous work Chandar et al. [2014], Wei and Deng [2017] and Gouws et al. [2015] used Europarl v7 parallel corpus Koehn [2005] to pretrain embeddings and then utilize it to classify subset of RCV1/RCV2 corpora Lewis et al. [2004]. To handle this task we need to construct meaningful bilingual text representations. For train and test we used RCV1/RCV2 corpora, where documents are

assigned to one of four predefined topics: CCAT (Corporate/Industrial), ECAT (Economics), GCAT (Government/Social), MCAT (Markets). In contrast to previous work, we *do not* use parallel data at all. We artificially paired documents according to their topics. We select 15000 documents from both English and German for classification experiments. The other part of RCV corpora was used to construct training triplets. Algorithm was trained for approximately 300K iterations with batch size equals to 50. We use Moses Koehn et al. [2007] preprocessing tools to lowercase and tokenize texts. Bag-of-words was used as an initial document representation. We keep 30000 top-frequency words for each language as a vocabulary.
For classification experiment, 10000 documents in English was used to train classifier and test it on 5000 documents in German and vice versa. We train logistic regression using low-dimensional representation obtained by our algorithm as features. Classification results are in Table 3.

Table 3: Text classification accuracy

Model	$en \to de$	$de \to en$
Majority Baseline	46.8	46.8
MT Baseline	68.1	67.4
Klementiev et al. [2012]	77.6	71.1
Gouws et al. [2015]	86.5	75.0
Chandar et al. [2014]	91.8	74.2
Wei and Deng [2017]	92.7	80.4
Su et al. [2018a]	91.3	77.8
This work	**94.3**	**82.8**

7 Conclusion

In this paper we proposed the Variational Bi-domain Triplet Autoencoder (VBTA) that learns a joint distribution of objects from different domains with the help of the learning triplets that sampled from the shared latent space across domains. We demonstrated the performance of the VBTA model on different tasks: image-to-image translation, bi-directional image generation and cross-lingual document classification.

Acknowledgments. This work was supported by RFBR project No.18-07-01441.

References

A. Bellet, A. Habrard, and M. Sebban. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*, 2013.

A. P. S. Chandar, S. Lauly, H. Larochelle, M. M. Khapra, B. Ravindran, V. Raykar, and A. Saha. An autoencoder approach to learning bilingual word representations. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 1853–1861, Cambridge, MA, USA, 2014. MIT Press.

J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio. A recurrent latent variable model for sequential data. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2980–2988. 2015.

Z. Gan, L. Chen, W. Wang, Y. Pu, Y. Zhang, H. Liu, C. Li, and L. Carin. Triangle generative adversarial networks. *CoRR*, abs/1709.06548, 2017. URL `http://arxiv.org/abs/1709.06548`.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. 2014.

S. Gouws, Y. Bengio, and G. Corrado. Bilbowa: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 748–756. JMLR.org, 2015.

P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016. URL `http://arxiv.org/abs/1611.07004`.

T. Karaletsos, S. Belongie, and G. Rätsch. Bayesian representation learning with oracle constraints. *arXiv preprint arXiv:1506.05011*, 2015.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling. Semi-supervised learning with deep generative models. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3581–3589. 2014.

A. Klementiev, I. Titov, and B. Bhattarai. Inducing crosslingual distributed representations of words. In *COLING*, 2012.

P. Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand, 2005. AAMT, AAMT.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum. Deep convolutional inverse graphics network. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2539–2547. 2015.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361–397, Dec. 2004.

M. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. *CoRR*, abs/1703.00848, 2017. URL `http://arxiv.org/abs/1703.00848`.

Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

T. Mikolov, Q. V. Le, and I. Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

M. Norouzi, D. J. Fleet, and R. R. Salakhutdinov. Hamming distance metric learning. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1061–1069. 2012.

F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

J. Su, D. Xiong, B. Zhang, Y. Liu, J. Yao, and M. Zhang. Bilingual correspondence recursive autoencoder for statistical machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1248–1258, 2015.

J. Su, S. Wu, B. Zhang, C. Wu, Y. Qin, and D. Xiong. A neural generative autoencoder for bilingual word embeddings. *Inf. Sci.*, 424(C):287–300, Jan. 2018a. ISSN 0020-0255.

J. Su, S. Wu, B. Zhang, C. Wu, Y. Qin, and D. Xiong. A neural generative autoencoder for bilingual word embeddings. *Information Sciences*, 424:287–300, 2018b.

M. Suzuki, K. Nakayama, and Y. Matsuo. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*, 2016.

R. Vedantam, I. Fischer, J. Huang, and K. Murphy. Generative models of visually grounded imagination. *arXiv preprint arXiv:1705.10762*, 2017.

W. Wang, H. Lee, and K. Livescu. Deep variational canonical correlation analysis. *CoRR*, abs/1610.03454, 2016. URL `http://arxiv.org/abs/1610.03454`.

L. Wei and Z.-H. Deng. A variational autoencoding approach for inducing cross-lingual word embeddings. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 4165–4171, 2017. ISBN 978-0-9992411-0-3.

J. Wieting, M. Bansal, K. Gimpel, and K. Livescu. Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*, 2015.

Z. Yi, H. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. *CoRR*, abs/1704.02510, 2017. URL `http://arxiv.org/abs/1704.02510`.

B. Zhang, D. Xiong, and J. Su. Battrae: Bidimensional attention-based recursive autoencoders for learning bilingual phrase embeddings. *CoRR*, abs/1605.07874, 2016. URL `http://arxiv.org/abs/1605.07874`.

J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017. URL `http://arxiv.org/abs/1703.10593`.