Unsupervised Learning of Text and Image Joint Embedding Space: A Preliminary Study

Tuan Lai [*]

lai123@purdue.edu

Learning inter-domain relationships from unpaired data is an important research problem, with potential applications in many areas such as computer vision or natural language processing. Many unsupervised methods have been proposed for relating two domains without the need of matching samples. The success of such methods is extraordinary but mostly limited to two visual domains or to two languages. There has been little prior work on investigating how to effectively relate texts and images without paired data. In this work, we conduct a preliminary study to investigate the feasibility of linking between texts and images without requiring paired training data. We propose a method for trying to learn a joint space for texts and images using only single modal datasets (i.e., without paired data). Images and texts with similar semantics are expected to have similar representations in the joint space. We test the proposed method on the Pascal Sentences dataset and the Flickr8k dataset. The results show that it is not trivial to learn a joint space for texts and images in a completely unsupervised setting. In addition, our method is shown to be more useful than random baselines in semi-supervised settings where there exists at least a small paired validation set. Code is available online on GitHub: `https://github.com/laituan245/DAULTIS`.

1 Introduction

Recent years have witnessed the rise of interest in many tasks at the intersection of computer vision and natural language processing, including semantic image retrieval [Johnson et al., 2015],

[*] Work done when taking the subject CS590-AMS (AI Meets Sustainability) at Purdue

image captioning [Karpathy and Fei-Fei, 2015, Vinyals et al., 2017], visual question answering [Zhou et al., 2015, Antol et al., 2015], and text-to-image synthesis [Reed et al., 2016]. A core problem for many of the tasks is how to measure the semantic similarity between visual data (e.g., an input image or region) and text data (a sentence or phrase). A common solution is to learn a joint latent space to relate visual data and text data. Most existing algorithms for projecting images and texts into a shared latent space assume access to *paired* training data [Hotelling, 1936, Andrew et al., 2013, Wang et al., 2017]. In other words, the assumption is that there exists a training set of image-text pairs such as $\{(\mathbf{i}_1, \mathbf{t}_1), (\mathbf{i}_2, \mathbf{t}_2), ..., (\mathbf{i}_n, \mathbf{t}_n)\}$ where $\mathbf{i}_j$ is an image and $\mathbf{t}_j$ is a text related to the image $\mathbf{i}_j$ (e.g., a caption describing the content of the image). However, in practice, the paired data assumption is often violated: there may be only a small subset of data available with pairing annotations, or even no pair data at all.

Building a cross-modal dataset can be labor intensive. On the other hand, single modal datasets can often be collected easily. Millions of text documents can be automatically collected from Wikipedia. Millions of images can be automatically crawled from the Internet. Motivated by these observations, our ultimate goal is to develop an effective method for learning a joint space for texts and images using only single modal datasets (i.e., without paired data). Images and texts with similar semantics are expected to have similar representations in the joint space. As we will discuss in Section 2, if we succeed in developing an effective unsupervised algorithm for projecting texts and images into a joint space, the potential applications are vast.

Many unsupervised methods have been proposed for relating two domains without observing matching samples. The success of such methods is surely extraordinary but mostly limited to two visual domains [Kim et al., 2017, Liu et al., 2017, Huang et al., 2018] or to two languages [Artetxe et al., 2017, Lample et al., 2018]. There has been little prior work on investigating how to project images and texts into a joint space without paired data [Hoshen and Wolf, 2018]. It is not clear if the task is even possible because texts and images are very different in nature. As people often say, "a picture is worth a thousand words", a picture may convey information more effectively than a text description in many situations. On the other hand, it is not trivial to directly represent many abstract words by images. For example, how does an image representing the word "theory" look like? We have decided the first step is to carefully analyze whether it is possible to learn a joint latent space for texts and images without any paired data. If possible, under which conditions, the task will be more challenging? Having a solid understanding of the task will be useful during the algorithm design phase in future work. The main contributions of our work may be summarized below.

- We propose a method for trying to learn a joint latent space to relate texts and images without any paired data.
- We test the proposed method on the Pascal Sentences dataset and the Flickr8k dataset. The experimental results show that it is not trivial to learn a joint space for texts and images without paired data. In addition, our method is shown to be more useful than random

baselines in semi-supervised settings where there exists at least a small paired validation set.

2 Applications

An effective algorithm for learning a common latent space to relate texts and images with only unpaired data will have many potential applications. We briefly discuss a few possible use-cases in this section.

Unsupervised Image Classification Suppose we have a pool of unlabelled images and a pool of category classes likely to be represented in the images. There are no explicit pairings between images and categories. If there exists an effective method for learning a joint latent space for texts and images without cross-modal data, it is possible to build an unsupervised image classifier. More specifically, we treat a category class as a sequence of one or more words. We learn a joint space for the images and the texts representing the categories. During inference, given a specific image, we simply select the category whose representation in the joint space is most similar to the representation of the image in the joint space. This task can be seen as an extreme version of zero-shot learning [Lampert et al., 2009, Tsai et al., 2017], where there is *no* auxiliary set of image + category pairs.

Unsupervised Image Captioning The task of describing an image with natural language has been a long-standing problem in computer vision. State-of-the-art methods for image captioning typically require supervised training data consisting of captions with paired image data [Vinyals et al., 2017]. Suppose it is possible to effectively project images and texts into a joint latent space without explicitly aligned data, then Figure 1 depicts an unsupervised method for image captioning. First, a decoder is trained to decode the representation of a sentence in the latent space back to the original sentence (e.g., the decoder can be a conditional language model). The decoder can be trained by a large set of unlabeled sentences. After training the decoder, given an image, we simply input the representation of the image in the joint space to the decoder to get a caption for the image.

Unsupervised Text-to-Image Synthesis The text-to-image synthesis is the reverse problem of image captioning: given a text description, the task is to generate an image which matches the description. If there is a method for relating texts and images without cross-modal data, it is possible to build an unsupervised text-to-image synthesis system. The approach is almost similar to the approach for building an unsupervised image captioning system. First, a decoder is trained to decode the representation of an image in the latent space back to the original image. The decoder can be trained by a large set of unlabeled images. After training the decoder, given a text description, we simply input the representation of the text in the joint space to the decoder to get an image for the text description.

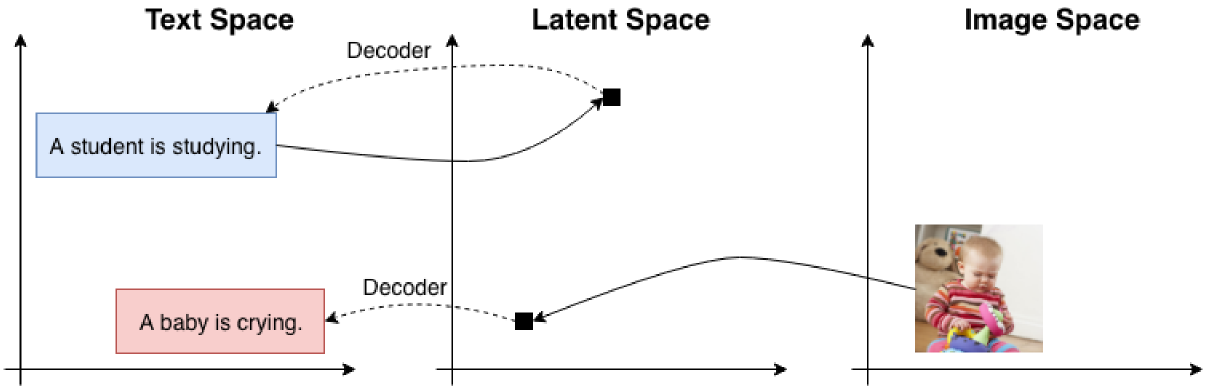


Figure 1: An unsupervised method for image captioning.

UNSUPERVISED MACHINE TRANSLATION Suppose we need to build a system for translating English to Korean. If there is an unsupervised method for relating texts and images, it is possible to build the system without the need of parallel sentences. First, we create monolingual corpora of the source and target languages. For example, in order to create a corpus for English, we can simply crawl million of English documents from Wikipedia. Second, we can crawl million of images from the Internet. Most websites use HTML and the <img> tag defines an image in an HTML page. After that, we build a text-to-image synthesis system using the English corpus and the dataset of images. Finally, we build an image captioning system using the dataset of images and the Korean corpus. During inference, given a text in English, we generate an image related to the text. After that, we generate a Korean caption for the image. We can consider the Korean caption as a translation of the original English text.

3 METHOD

Let's formally define the problem of interest. Suppose we have a set $\{x_1, x_2, ..., x_n\}$ of n image feature vectors from domain X and we have a set $\{y_1, y_2, ..., y_m\}$ of m text feature vectors from domain Y. The relationship between any pair (x_i, y_j) is not known. In other words, it is not known whether the image represented by the vector x_i and the text represented by the vector y_j are related or not. Typically the number of dimensions of x_i is different from the number of dimensions of y_j and it is not possible to directly measure the similarity between x_i and y_j using metrics such as Euclidean distance or Cosine similarity. Even if the numbers of dimensions are the same, the image feature vectors and the text feature vectors typically have very different statistical properties and follow different distributions. The main task is to learn a mapping $E_x : X \to H$ and a mapping $E_y : Y \to H$ such as $E_x(x_i)$ and $E_y(y_j)$ can be directly compared and their similarity reflects their relatedness. H can be considered as a joint latent space.

In this section, we propose a deep architecture for tackling the problem of interest. We refer to the architecture as DAULTIS (Deep Autoencoders for Unsupervised Learning of Text and Image Space). The main components of the architecture are two deep autoencoders, one for the image feature vectors and one for the text feature vectors (Figure 2). The image autoencoder consists of an image encoder $E_x : X \to H$ and an image decoder $F_x : H \to X$. The role of the encoder E_x is to map an image feature vector x_i into a representation $E_x(x_i)$ in the latent space. The role of the decoder F_x is to generate a reconstruction $F_x(E_x(x_i))$ of the original feature vector x_i. If the latent representation $E_x(x_i)$ is informative, the decoder will be more likely to accurately reconstruct the original feature vector (i.e., $F_x(E_x(x_i)) \approx x_i$). We define the reconstruction loss for the image autoencoder as:

$$L_{Rec_X} = \sum_{i=1}^{n} \left\| F_x(E_x(x_i)) - x_i \right\|^2 \tag{1}$$

Similarly, the text autoencoder consists of a text encoder $E_y : Y \to H$ and a text decoder $F_y : H \to Y$. We define the reconstruction loss for the text autoencoder as:

$$L_{Rec_Y} = \sum_{j=1}^{m} \left\| F_y(E_y(y_j)) - y_j \right\|^2 \tag{2}$$

Minimizing the reconstruction losses would encourage the latent representations to have useful features describing the original feature vectors. However, without additional constraints, it is likely that the latent representations of images and the latent representations of texts will have very different characteristics and statistical properties. Inspired by recent works on mapping images between domains without aligned image pairs by combining generative adversarial networks with "cycle-consistent" constraints [Liu et al., 2017, Hoffman et al., 2018], we introduce additional constraints to DAULTIS to try to minimize the discrepancy between texts and images.

First, we want to enforce the "cycle-consistent" constraints, in the sense that if we e.g., first transform an image into a text description and then try to generate a new image which matches the text description, the newly generated image and the original image should be similar. More specifically, suppose we have an image feature vector x_i, we would want $F_x(E_y(F_y(E_x(x_i))))$ to be similar to x_i. Figure 3 illustrates the cycle consistent constraint. Similarly, given a text feature vector y_j, we would want $F_y(E_x(F_x(E_y(y_j))))$ to be similar to y_j. We explicitly define the cycle consistency losses as:

$$L_{Cycle_X} = \sum_{i=1}^{n} \left\| F_x(E_y(F_y(E_x(x_i)))) - x_i \right\|^2 \tag{3}$$

$$L_{Cycle_Y} = \sum_{j=1}^{m} \left\| F_y(E_x(F_x(E_y(y_j)))) - y_j \right\|^2 \tag{4}$$

Second, a recent popular approach for reducing the discrepancy between representations learned from two different domains is via adversarial training [Ganin and Lempitsky, 2015]. Inspired by the work, we use the approach to introduce additional constraints to DAULTIS. More specifically, we want to train a discriminator network D_H which takes a projected representation in the latent space H as input and predicts whether the representation is from the image domain or from the text domain. We define the loss function of the discriminator D_H as:

$$L_{D_H} = BCE(D_H(C_x), 0) + BCE(D_H(C_y), 1) \tag{5}$$

where BCE denotes the Binary Cross Entropy function, $C_x = \{E_x(x_1), E_x(x_2), ..., E_x(x_n)\}$, and $C_y = \{E_y(y_1), E_y(y_2), ..., E_y(y_m)\}$. If the discriminator D_H does well, the value of L_{D_H} will be small. On the other hand, we would want the image encoder E_x and the text encoder E_y to have an opposite goal, which is to minimize the following loss:

$$L_{G_H} = BCE(D_H(C_x), 1) + BCE(D_H(C_y), 0) \tag{6}$$

Minimizing L_{D_H} and minimizing L_{G_H} are two opposite goals. When the discriminator predicts many image latent representations to have label 0 and predicts many text latent representations to have label 1, L_{D_H} will be small but L_{G_H} will be large (and vice versa). The intuition is that if the image latent representations produced by E_x and the text latent representations produced by E_y have very different properties and characteristics, the discriminator will be likely to do well. In order to fool the discriminator, the encoders would need to produce latent representations that are modality invariant. By trying to optimize the discriminator to minimize L_{D_H} and to optimize the encoders to minimize L_{G_H} at the same time, we implicitly try to reduce discrepancy between latent representations from the two different domains.

Based on the same idea of adversarial learning, we define the following losses:

$$L_{D_X} = BCE(E_x(C_x), 0) + BCE(E_x(C_y), 1) \tag{7}$$

$$L_{D_Y} = BCE(E_y(C_x), 0) + BCE(E_y(C_y), 1) \tag{8}$$

$$L_{G_X} = BCE(E_x(C_x), 1) + BCE(E_x(C_y), 0) \tag{9}$$

$$L_{G_Y} = BCE(E_y(C_x), 1) + BCE(E_y(C_y), 0) \tag{10}$$

Taken together, the two deep autoencoders try to minimize the following total loss:

$$L_G = \lambda_{adv}(L_{G_X} + L_{G_Y} + L_{G_H}) + \lambda_{cycle}(L_{Cycle_X} + L_{Cycle_Y}) + \lambda_{rec}(L_{Rec_X} + L_{Rec_Y}) \tag{11}$$

where λ_{adv}, λ_{cycle}, and λ_{rec} are hyper-parameters.

On the other hand, the discriminator networks D_H, D_X, and D_Y try to minimize the following total loss:

$$L_D = L_{D_X} + L_{D_Y} + L_{D_H} \tag{12}$$

We train DAULTIS using mini-batch stochastic gradient descent (SGD) using the ADAM optimization algorithm. During training, we alternate between optimizing the deep autoencoders for minimizing L_G and optimizing the discriminators for minimizing L_D. In every experiment, training was performed for 50 epochs.

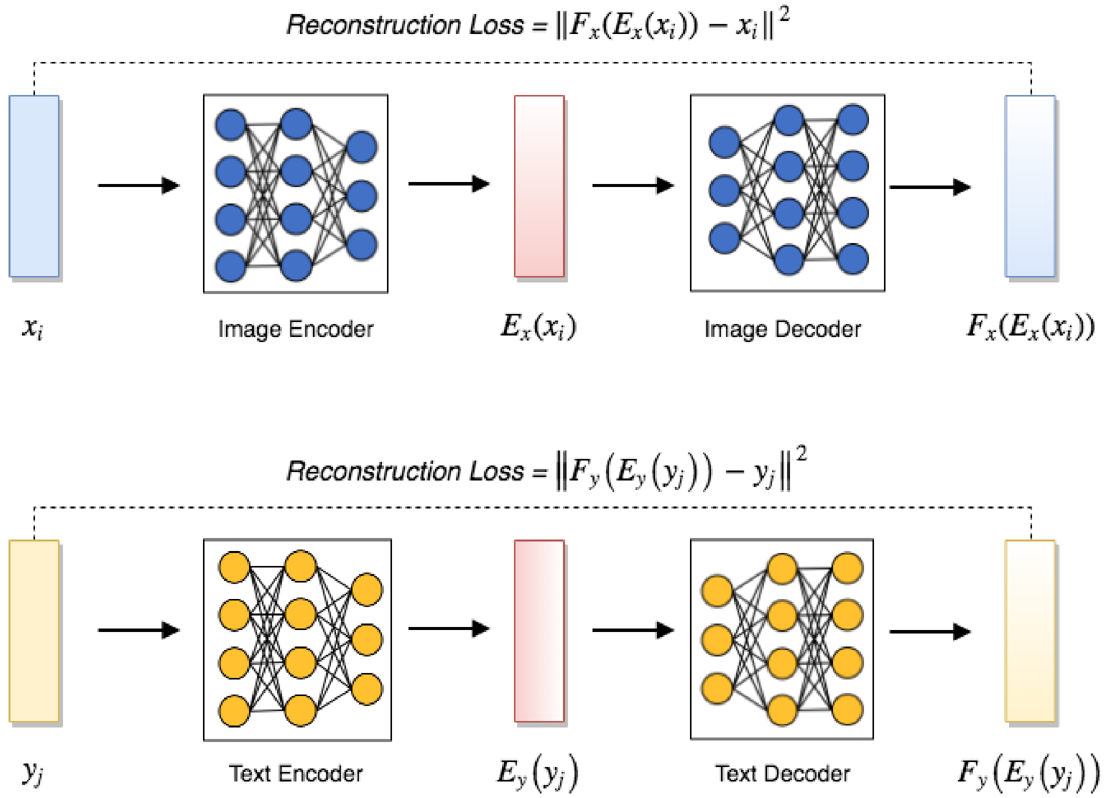


Figure 2: The main components of `DAULTIS` are two deep autoencoders, one for the image feature vectors and one for the text feature vectors.

4 Experiments

Our experiments employ two datasets:

- **Flickr8k dataset** [Hodosh et al., 2013] includes 8,000 images obtained from the Flickr website. Each image was annotated by 5 sentences. We split the dataset into three subsets, 6,000 images for training, 1,000 images for validation, and 1,000 images for testing. We encode the images using ResNet152 [He et al., 2016] and encode the sentences using pre-trained GloVe word embeddings [Pennington et al., 2014]. More specifically, the text associated with an image was taken as the average of the word embeddings of all words in the five sentences describing it. The dimension of an image feature vector is 2,048. The dimension of a text feature vector is 300.

- **Pascal Sentences dataset** [Rashtchian et al., 2010] has 1,000 images from 20 categories. Each image has 5 description sentences. We split the dataset into two subsets, 800 images for training and 200 images for testing. The feature extraction for images and texts is the same as for the Flickr8k dataset. This dataset is available online at `http://vision.cs.uiuc.edu/pascal-sentences/`.

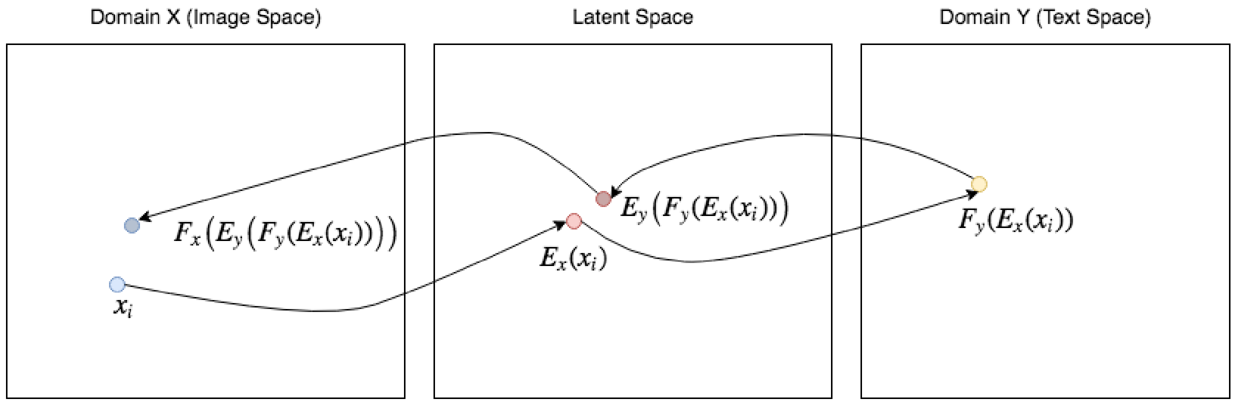


Figure 3: An illustration of a cycle consistent constraint. Suppose we have an image feature vector x_i, we would want $F_x(E_y(F_y(E_x(x_i))))$ to be similar to x_i.

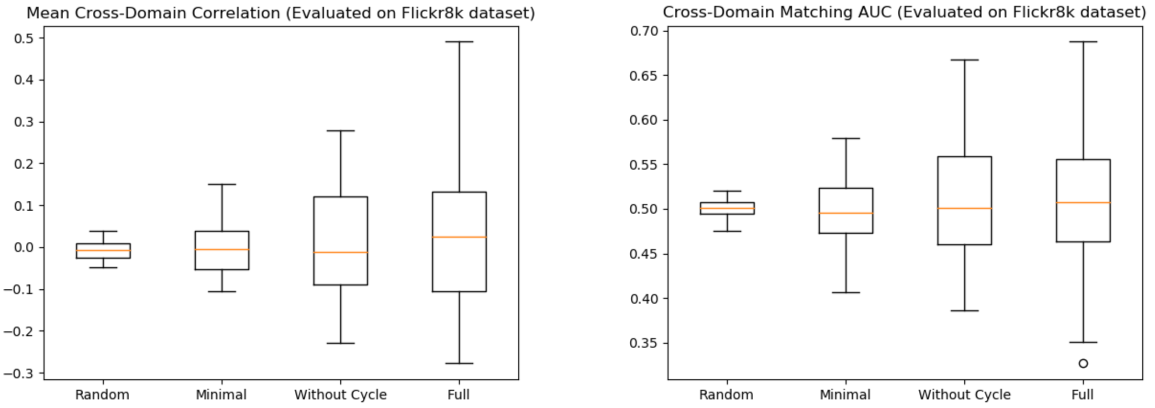
Two metrics are used for evaluation: (1) Correlation - average of the 1D dimension by dimension correlations of the latent representations of matching text-image pairs, and (2) Area Under Curve (AUC) - We compute the similarity, in the latent space, between pairs of positive and negative matches (each pair has one sample from the text domain and one from the image domain) and report the area under the ROC curve.

Baselines:

- **Random**. The parameters of E_x and E_y are randomly initialized. There is no training.
- **Minimal Constraint**. λ_{cycle} is set to be 0, λ_{rec} is set to be 0, λ_{adv} is set to be 1 and we discard the discriminator networks D_X and D_Y (and discard L_{D_X}, L_{D_Y}, L_{G_X}, and L_{G_Y}). There is only the discriminator network D_H for adversarial learning.
- **Without Cycle Constraints**. λ_{cycle} is set to be 0, λ_{rec} is set to be 0.001, λ_{adv} is set to be 1.
- **Full Constraints**. λ_{cycle} is set to be 0.001, λ_{rec} is set to be 0.001, λ_{adv} is set to be 1.

4.1 Results

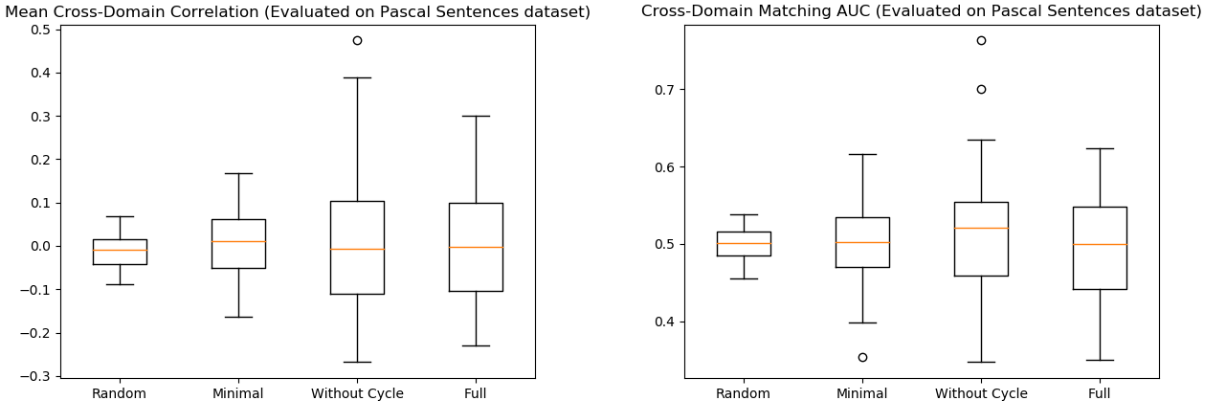
Running the proposed method multiple times results in a number of runs with bad performance and a number of runs with decent performance. In semi-supervised settings where there exists at least a small paired validation set for choosing the best performing runs, our method is more useful than random baselines.



(a) Mean Cross-Domain Correlation

(b) Cross-Domain Matching AUC

Figure 4: Flickr8k Dataset



(a) Mean Cross-Domain Correlation

(b) Cross-Domain Matching AUC

Figure 5: Pascal Sentences Dataset

Table 1: Flickr8k Dataset

Method	Mean Cross-Domain Correlation	Cross-Domain Matching AUC
Random	-0.01 ± 0.02 (**0.04**)	0.50 ± 0.01 (**0.52**)
Minimal Constraint	-0.00 ± 0.06 (**0.15**)	0.50 ± 0.04 (**0.58**)
Without Cycle Constraints	0.01 ± 0.14 (**0.28**)	0.51 ± 0.07 (**0.67**)
Full Constraints	0.03 ± 0.17 (**0.49**)	0.51 ± 0.08 (**0.69**)

Table 2: Pascal Sentences Dataset

Method	Mean Cross-Domain Correlation	Cross-Domain Matching AUC
Random	-0.01 ± 0.04 (**0.07**)	0.50 ± 0.02 (**0.54**)
Minimal Constraint	0.00 ± 0.08 (**0.17**)	0.50 ± 0.05 (**0.62**)
Without Cycle Constraints	0.02 ± 0.16 (**0.47**)	0.51 ± 0.08 (**0.76**)
Full Constraints	0.00 ± 0.14 (**0.30**)	0.50 ± 0.07 (**0.62**)

5 Conclusions and Future Work

In this work, we conduct a preliminary study to investigate the feasibility of relating between texts and images without the need of matching samples. We propose a method for trying to learn a common latent space for texts and images without using cross-modal data. We test the proposed method on the Pascal Sentences dataset and the Flickr8k dataset. The results show that it is not trivial to learn a joint space for texts and images in a completely unsupervised setting. Running the proposed method multiple times results in a number of runs with bad performance and a number of runs with decent performance. In semi-supervised settings where there exists at least a small paired validation set for choosing the best performing runs, our method is more useful than random baselines. In the future, we would like to experiment with variational autoencoders (VAEs).

6 Acknowledgement

The author wants to thank Professor Yexiang Xue and Dr. Quan Tran of Adobe Research for their guidance on this project.

References

[Andrew et al., 2013] Andrew, G., Arora, R., Bilmes, J. A., and Livescu, K. (2013). Deep canonical correlation analysis. In *ICML*.

[Antol et al., 2015] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). Vqa: Visual question answering. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433.

[Artetxe et al., 2017] Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2017). Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

[Ganin and Lempitsky, 2015] Ganin, Y. and Lempitsky, V. S. (2015). Unsupervised domain adaptation by backpropagation. In *ICML*.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

[Hodosh et al., 2013] Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics (extended abstract). *J. Artif. Intell. Res.*, 47:853–899.

[Hoffman et al., 2018] Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A. A., and Darrell, T. (2018). Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*.

[Hoshen and Wolf, 2018] Hoshen, Y. and Wolf, L. (2018). Unsupervised correlation analysis. *CoRR*, abs/1804.00347.

[Hotelling, 1936] Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

[Huang et al., 2018] Huang, X., Liu, M.-Y., Belongie, S. J., and Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In *ECCV*.

[Johnson et al., 2015] Johnson, J., Krishna, R., Stark, M., Li, L.-J., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L. (2015). Image retrieval using scene graphs. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3668–3678.

[Karpathy and Fei-Fei, 2015] Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137.

[Kim et al., 2017] Kim, T., Cha, M., Kim, H., Lee, J. K., and Kim, J. (2017). Learning to discover cross-domain relations with generative adversarial networks. In *ICML*.

[Lampert et al., 2009] Lampert, C. H., Nickisch, H., and Harmeling, S. (2009). Learning to detect unseen object classes by between-class attribute transfer. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958.

[Lample et al., 2018] Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. (2018). Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

[Liu et al., 2017] Liu, M.-Y., Breuel, T., and Kautz, J. (2017). Unsupervised image-to-image translation networks. In *NIPS*.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP*.

[Rashtchian et al., 2010] Rashtchian, C., Young, P., Hodosh, M., and Hockenmaier, J. (2010). Collecting image annotations using amazon's mechanical turk. In *Mturk@HLT-NAACL*.

[Reed et al., 2016] Reed, S. E., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. (2016). Generative adversarial text to image synthesis. In *ICML*.

[Tsai et al., 2017] Tsai, Y.-H., Huang, L.-K., and Salakhutdinov, R. (2017). Learning robust visual-semantic embeddings. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3591–3600.

[Vinyals et al., 2017] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2017). Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:652–663.

[Wang et al., 2017] Wang, B., Yang, Y., Xu, X., Hanjalic, A., and Shen, H. T. (2017). Adversarial cross-modal retrieval. In *ACM Multimedia*.

[Zhou et al., 2015] Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., and Fergus, R. (2015). Simple baseline for visual question answering. *CoRR*, abs/1512.02167.