Unsupervised Image-to-Image Translation with Staged Transfer Learning

Abhijeet Chauhan, Cheng Duan, Fan Mo
Carleton University
Ottawa, Canada
abhijeetchauhan@cmail.carleton.ca, cduan092@uottawa.ca
fanmo@cmail.carleton.ca

Abstract

We propose a staged transfer learning approach for unsupervised image-to-image translation. Our goal is to learn a mapping from $G : A \rightarrow C$ without paired images, where A is the source domain and C is the target domain. Instead of performing image translation from source to target domain directly, to improve the performance in image-to-image translation, an intermediate domain B is used as a bridge to help domain translation from A to C. Our approach addresses the two challenges in image-to-image translation, heterogeneous domains and imbalanced training data size between the source and target image domains. Qualitative results are presented to compare the image translations between the direct translation and bridged translation with an intermediate domain.

1 Introduction

Transfer learning is a typical machine learning which focuses on using the knowledge learned from one domain on another domain. An interesting application of using the concept is doing image-to-image translation. Such problems can be concluded as trying to map an image from one domain to another domain. In fact, image-to-image translation and domain adaptation has drew attention these years and many researches have been conducted [10] [5] [3].

In image-to-image translation, paired images are very costly to obtain, while unpaired image datasets are abundant. Therefore to be able to solve the image-to-image translation problem without paired images is more practical and could be widely used in many real-world applications. In image-to-image translation, the source domain and target domain might have a very different structure and distributions and the mapping from source to target could be hard to learn. Another issue in image-to-image translation is while the source domain might have big enough dataset, the target domain might not have big enough dataset. Such imbalance of source and target domains dataset will cause image translation issue as the model and algorithm do not have sufficient look at the target domain structure given the small number of samples in the target domain.

We, therefore, propose the idea of domain bridging by introducing a bridge domain to help the domain translation. The selected bridge domain should have sufficient number of samples and have some common features and structure distribution as the target domain, to help the final translation from source to target domain.

Our contribution is the new idea of domain bridging, to help image translation from a source domain to a target domain, to address the heterogeneous issue when two domains are very different, or when there is a training data imbalance issue between the source an target domains. In our proposed method, there is only one bridging domain, but the idea can be extended to multiple bridging domains. The domain bridging can be a chain of bridges from the source to the target domain, or any directional graph structure of bridges. As this is our initial investigation of domain bridging in

transfer learning, we focus on just using single bridging domain. Multiple bridging domains will be left for future work.

2 Related Work

Generative Adversarial Networks (GANs) [1] has been widely explored in image generation and have shown very good results. On image-to-image translation, several state-of-art approaches have been investigated and proposed [9] [6] [4] [2].

Conditional GANs[9] are conditional version of the vanilla GAN[1]. In addition to noise, conditions such as labels of the images are also fed to the generator. An ideal generator will produce ideal images which meet those conditions. Similarly, discriminator takes an image with a prior condition as inputs and outputs whether the image is real or fake. In other words, generator predicts the probability of y given x where y is the image and x is the condition.

CoGAN[6] aims to learn the joint distribution of two image domains without paired images. In the paper, a constrained version of GAN was proposed, which enforce weight-sharing constraints to encourage the learning of underlying common hierarchy structure across two image domains. The assumption was made that two domains would have considerably similar data distributions and structures. By sharing weights with the generator and discriminator on the higher level of layers, a random input noise could generate images in two different domains. As the images from two domains share the same high level data distribution, it is possible to map images from domain A to domain B.

UNIT[5] extends the work in CoGAN[7] for unsupervised image-to-image translation. In [5] a shared-latent space assumption is made, which assumes a pair of corresponding images in different domains can be mapped to the same latent representation in a shared-latent space. In UNIT[5] each image domain is modeled using a VAE-GAN. The combination of adversarial training objective and weight-sharing is used to enforce a shared-latent space and generate corresponding images in two domains.

Apart from conditional GANs and CoGAN, one of the impressive method "CycleGAN" was developed by Zhu et al. [10]. The goal is to learn the image mapping from an input domain to the target domain where paired training data are not available [10]. More specifically, as shown in 1 they proposed cycleGAN which could learn a mapping function $G : X- > Y$ such that the distribution of images from G(X) is indistinguishable from the distribution [5]. In addition, they coupled the function G with an inverse mapping function $F : Y- > X$. Also, they introduce a cycle consistency loss to push $F(G(X)) = X$ and vice versa [5].

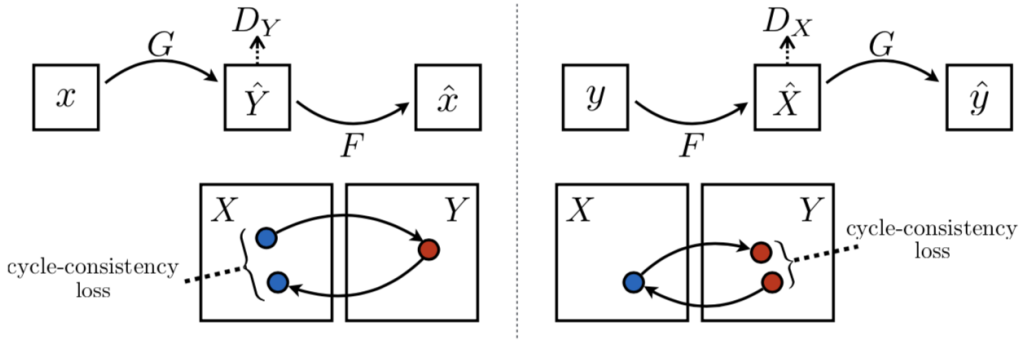

Figure 1: Image at left hand side shows the training process of mapping from domain x to $\hat{Y}$ using mapping function G and mapping from y to $\hat{x}$ using mapping function F. Besides, specifically define cycle-consistency loss. The image at right hand side shows similar process.

Qualitative results are presented in [10] on several tasks where paired training data does not exist, including collection style transfer, object transguration, season transfer, photo enhancement, etc. Quantitative comparisons against several prior methods demonstrate the superiority of our approach.

The idea of cycleGAN was inspired by Pix2Pix. With the benet of image-to-image translation, they made another step forward by taking the idea of cycle consistency into their loss calculation.

For the G net, the author suggests a network with high performance on neural style transfer and super-resolution. For the D net, a 70 * 70 Patch-GANs with the high performance was mentioned for classifying whether 70 * 70 overlapping image patches are real or fake [7].

In order to achieve the objectives, the model firstly tries to optimize the adversarial loss of two mapping functions:

$$\begin{aligned}\mathcal{L}_{GAN}(G, D_Y, X, Y) &= E_{y\sim P_{data}(y)}[logD_Y(y)]+ \\ &\quad E_{x\sim P_{data}(x)}[log(1 - D_Y(G(x)))] \\ \mathcal{L}_{GAN}(F, D_X, X, Y) &= E_{x\sim P_{data}(x)}[logD_X(x)]+ \\ &\quad E_{y\sim P_{data}(y)}[log(1 - D_X(F(y)))]\end{aligned}$$

Besides, the cycle consistency loss includes both forward and backward cycle consistency loss:

$$\begin{aligned}\mathcal{L}_{cyc}(G, F) =& E_{x\sim P_{data}(x)}[\|F(G(x)) - x\|_1]+ \\ & E_{y\sim P_{data}(y)}[\|G(F(y)) - y\|_1]\end{aligned}$$

are optimised using L1. Overall, the full loss function to be optimised is:

$$\begin{aligned}\mathcal{L}(G, F, D_X, D_Y) &= \mathcal{L}_{GAN}(G, D_Y, X, Y) \\ &= \mathcal{L}_{GAN}(F, D_X, Y, X) \\ &= \lambda\mathcal{L}_{cyc}(G, F)\end{aligned}$$

Besides, in Harry et al. implementation version of cycleGAN, they improved the original algorithm in terms of convergence speed by adding skip connection between input and output in the G net.

In addition, in [4], Kim et al. transited from supervised nature to the unsupervised nature of the problem. They used samples from 2 dierent domains and discovered the relations between them. It showed a great advantage of detecting multiple representations of the same image which can be used to increase the accuracy of downstream applications.

3 Method

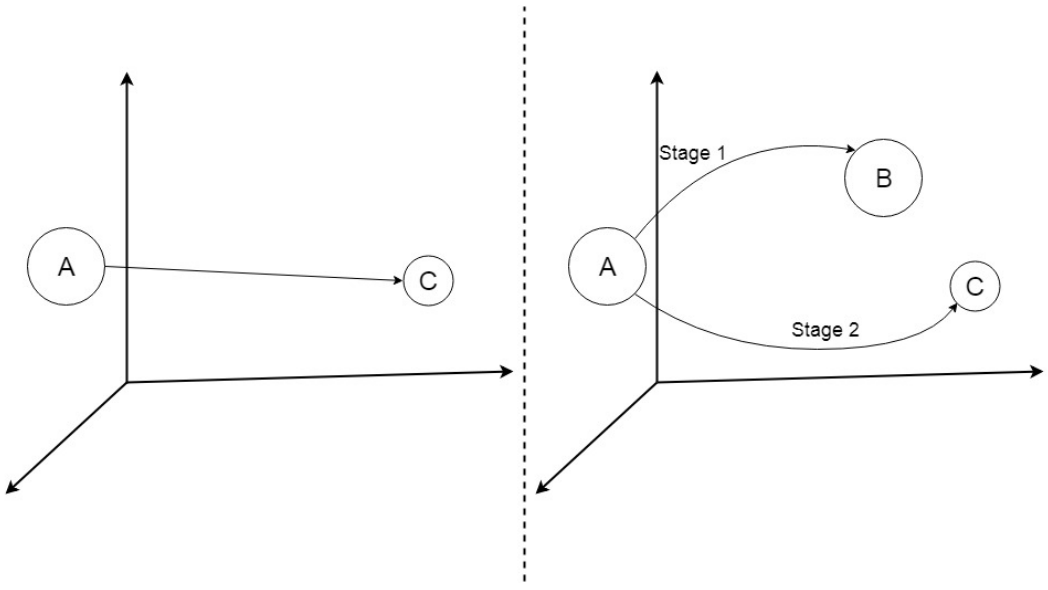

Figure 2: Our approach divides the goal of translating the image in domain A to domain C in two stages. The dataset in domains A, B and C, where volume shows the number of samples available and distance represents the similarity of the dataset. Stage 1: Learn to map from A to B which is feasible due to sufficient data. Stage 2: Learn to map from A to C.

3.1 Formulation

As shown in Figure 2, the target domain C has a small training dataset, and it is imbalanced comparing to domain A. Such training data set imbalance can cause a serious problem in classification [8]. The effect of such imbalanced training data set on unsupervised image-to-image translation has not been well examined.

Our hypothesis is in image-to-image translation, if the training data in the target domain C is much smaller than what source domain A has, the translation result will not be good, because the structure and distribution of target domain are not well presented to the model and algorithm. To address such issue, we propose to introduce a new intermediate domain B, which has similar size of training data as domain A has and also has a similar structure as the target domain C.

Our assumption is the image-to-image translation is a mapping from one latent space where the source domain resides to another latent space where target domain resides. When the latent space of the target domain is not fully represented by the target domain due to the limited training data, by using a bridging domain B, which has similar latent space as target domain C, and its latent space is fully represented by the sufficient training data in the bridging domain, the model can learn the mapping from the source domain to the target domain.

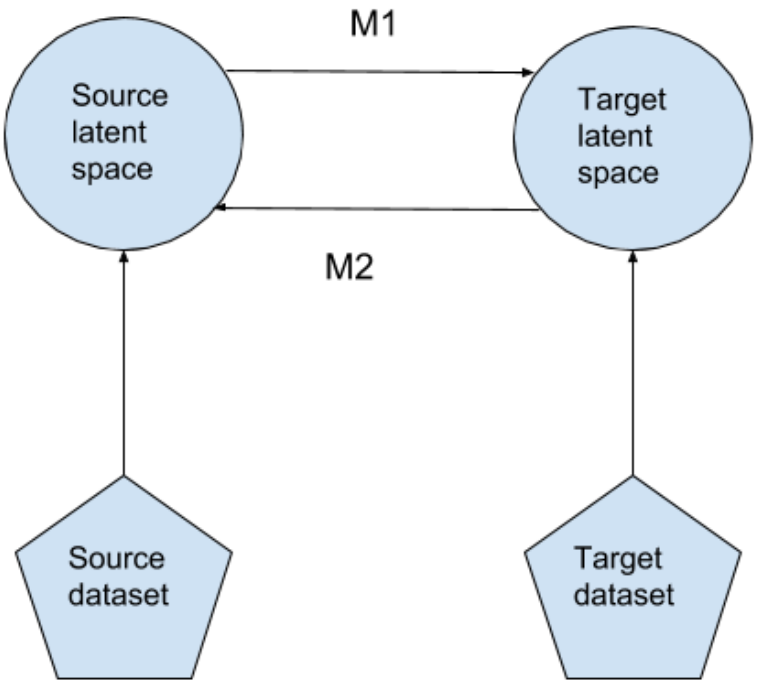

Figure 3: Latent space mapping

As shown in Figure 3, the source dataset is represented in the source latent space, and the target data set is represented in the target latent space. If the training data set in the target domain is small, the target latent space will not be well constructed by the model, which will cause image translation with lower quality. The same issue will happen when the source data set has small training data comparing to the target domain due to the symmetric nature of translation task.

To address the issue when the target domain does not have enough training data, our method is to add an intermediate domain B, which has enough training data and have some common structure and distribution with the target domain C. It is an open question on how to choose bridge domain, and how to measure the similarity between two domains. In our experiment we choose the bridge

domain which has similar visual structure as the target domain. The two breeds of dogs picked are similar.

3.2 Implementation

Network Architecture

We implement the system based on CycleGAN [10]. The reason we choose CycleGAN as our base is in CycleGAN, there is no additional assumption beyond cycle consistency. Therefore there is no assumption between the latent space of the source domain and the target domain, whether they are homogeneous or heterogeneous, therefore the source and target domains do not need to be similar.

Another reason to choose GAN based generative model for image translation is GAN has been widely experimented and proved to be able to generate images with high visual qualities [1].

We didn't choose CoGAN and UNIT based architecture because, in CoGAN and UNIT model, there is an assumption of shared latent space between the source and target domain. Such an assumption is unnecessary in image-to-image translation. In image-to-image translation, the only assumption we need is cycle consistency which is inherent in any translation tasks, including image-to-image translation, language-to-language translation.

Another reason we didn't choose CoGAN/UNIT as our base design is in UNIT, there is Variational Auto-Encoder (VAE) introduced [4]. The VAE serves the purpose of auto-encoder of input images to latent space, then the output of VAE is fed into the generator for image generation and translation. In the cycleGAN design, the generator consists of convolution layers, followed by some layers of residual blocks, then de-convolution layers. The convolution layers already serve the purpose of encoding input images to latent space, and the deconvolution layers generate translated images from latent space. Therefore there is no need for a separate VAE as auto-encoder, which will cause the increased cost of computations and increased training time.

Training Details

Stage1: CycleGAN architecture is used to learn the mapping from domain A to domain B where both domains have sufficient data to make learning feasible. The weights of the generator are saved to be used in the next stage.

Stage2: To learn the mapping from A to C the weights from the pretrained generator is used. This lead to quicker convergence and better results in the case when domain A and C are very different in structure and there is data imbalance problem.

4 Experiments

The code for this project adopts the existing cycleGAN code introduced by the author Liu et al. [5]. We used an tensorflow version of it. We transferred entire project into a notebook and made several modifications based on our objectives.

Several experiments were taken in order to justify our statements. In original cycleGAN paper [5], the authors trained the model using ImageNet. However, due to the limitation of computing resources, we choose several smaller datasets for training and testing. Apart from that, all experiments were performed using Google's Colaboratory platform which is a online notebook platform with free GPU accessing.

4.1 Datasets

In order to implement our assumption, we chose three species of dogs, Maltese, Chow and Dhole. As the obtained images are in different sizes, we re-sized them all into 256 by 256 with 3 channels (as shown in figure 4).

(a) Maltese dog (b) Chow dog (c) Dhole dog

Figure 4: Examples of three species of dogs where Maltese dog is used as domain A, Chow dog is used as domain B and Dhole dog is used as domain C

In order speed up the training speed, we limited the number of images for training and testing. For training, there are 196 images of Chow dogs, 150 images of Dhole dogs and 252 images of Maltese dogs. For testing, there are 30 images for each species respectively.

4.2 Experiments Design

The experiments were divided in two parts.

Firstly, another model was trained to directly transfer dogs from domain A to domain C. During this training session, Maltese dogs and Dhole dogs will be used.

Secondly, a model was trained using Maltese dogs and Chow dogs which is trying to translate the dogs from domain A to domain B. After that, the trained model will be used to continually trained using Maltese dogs and Dhole dogs. This is to transfer Maltese dogs (domain A) into Dhole dogs (domain C). During this process, we adopted a feature of tensorflow which could save weights of the trained model and restore them back during new training session.

Each training session was designed to train 100 epochs while each epoch contains approximately 150 iterations (slightly differences between training session due to the different of the number of training images).

4.3 Results

Based on experiments design, several experiments were performed so as to achieve our objectives.

In most cases, standards such as prediction accuracy, mean square error or evaluation plots like ROC curve are usually been used to evaluate the quality of a trained model. However, these evaluation methods require corresponding standards or labels such that the calculated loss or label can be mathematically compared based on L1, mean square error, etc. While in this case, our model focuses on unpaired image-to-image translation task. Therefore, traditional evaluation method cannot be applied. During this research, we find most of the papers also do not have an uniformed mature measurement solution. Comparing pictures is the most adopted approach.

First Phase

For our first phase of experiments, a model was firstly trained to translate Maltese dogs (domain A) to Dhole dogs (domain C). Followed by training, the trained model was used to transfer the images in testing sets.

As shown in figure 5, three example sets of testing results are shown, where the first column are input images from domain A, the second column are corresponding translated fake images in domain C, and the third column are cycle-backed images by translating the generated domain C image back to domain A. While the figure 6 is a control group which containing results of translating images from domain C to domain A.

Figure 5: Three examples of testing results

Figure 6: Three examples of testing results

Figure 7: Results of A $\rightarrow C$ at 11 epoch before A $\rightarrow B$ training.

Figure 8: Results of A $\rightarrow C$ at 11 epoch after A $\rightarrow B$ training.

By looking at the images in Figure 5-6, it can be seen that the models are trying to translating images in terms of colour only. A white maltese is transferred to a brown dog as a dhole is brown. Besides, a brown dhole is transferred to a while dog as maltese is white. However, it is pretty obvious that both translation tasks performed not well. The generated fake images are not clear and the background of the images seems being grey-scaled to some extent. On one hand, such results meet our expectation as there is not enough domain C images for training. On the other hand, the results could be slightly better if more computing resources can be accessed even with limited number of domain C images.

Second Phase

For our second phase of experiments, a model was firstly trained to translate Maltese dogs (domain A) to Chow dogs (domain B). Followed by training, the pretrained model was tuned/trained to translate Maltese dogs (domain A) to Dhole dogs (domain C). Followed by that, the trained model was used to transfer the images in testing sets.

After the $A \rightarrow B$ (Maltese to Chow) training, the experiment results of $A \rightarrow C$ (Maltese to Dhole) to presented in Figure 7 - 12. From the results shown in Figure 8, 10, and 12, where the bridging domain is used for training first, we can see the visual effect is much better comparing to the case where no bridging domain is used. From the training using bridging domain B (Chow), the main character of Chow is learned by the model and applied to the translate image. But from the results

Figure 9: Results of A $\rightarrow C$ at 30 epoch before A $\rightarrow B$ training.

Figure 10: Results of A $\rightarrow C$ at 30 epoch after A $\rightarrow B$ training.

Figure 11: Results of A $\rightarrow C$ at 95 epoch before A $\rightarrow B$ training.

Figure 12: Results of A $\rightarrow C$ at 95 epoch after A $\rightarrow B$ training.

we can see that during training the model only learned the color mapping and how to generate image back from translated image.

From the experiment results we also see that in the staged training, the Maltese-to -Dhole has slightly better results than Dhole-to-Maltese, especially on on generated back images. This could be caused by the fact that in the first stage training, model generators have learned how to generate images from domain B (Chow) back to domain A (Maltese) well, but because C (Dhole) is new to the model, to generate from C back to A, is more difficult. But from the results, we can see there is no big differences on this even when C is a new domain to the pretrained model.

From the results, we can see that cycle-consistency is a very strong factor in the system. Even though the translation result is not good, the images generated back to the source domain has much better visual quality. We suspect that in our experiment, the cycle-consistency loss is given too high emphasis. More future works required to further experiment with more emphasis on adversarial loss instead of cycle-consist loss.

Comparing to CycleGAN paper [10], their experiments also only have good results on domains with very small difference, e.g., horse to zebra. The authors in the paper mentioned that they have little success to translate images with geometric changes. We have not seen any published good results on images from two domains with significant geometric changes. All results we have seen are based on one attribute, for example, color, or face with/without glasses. One possible reason is the GAN used in the implementation might have mode collapse on one attribute during training. If this is the case, WGAN might help to learn multiple attributes mapping. Replace GAN with WGAN is left for future work.

5 Conclusion

Our proposed staged translation model performed well when compared with direct translation and is the first model to introduce the idea of staged translation. The results from the experiments substantiate our hypothesis that when the data available for target domain is not enough, the structure and distribution of the target domain are not well presented to the model and algorithm. In such cases, a new intermediate domain B, which has sufficient size of training data and also has a similar structure as the target domain C can be used as a bridge domain. Both models (our and original CycleGAN paper) wasn't able to capture the structural changes of the dogs. They were able to capture only color and texture changes. Handling these type of structural changes is a significant future work.

Experiment different weights between cycle-consistency loss and adversarial loss could be a way to improve the translation results for images with geometric changes; Replacing GAN with WGAN could be another way.

The results of the model highly dependent on the selection of the intermediate domain (B). Selecting a domain B which is not similar to domain C in terms of the color, texture and other features will lead to poor results. In our experiments, we chose our domains very carefully. Selection of the appropriate domains for the image to image translation task is a very important future work.

References

[1] Jean Pouget-Abadie Mehdi Mirza Bing Xu David Warde-Farley Sherjil Ozair Aaron Courville Goodfellow, Ian and Yoshua Bengio. Generative adversarial nets. In *In Advances in neural information processing systems*, pages 2672–2680, 2014.

[2] Jun-Yan Zhu Tinghui Zhou Isola, Phillip and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *arXiv preprint*, 2017.

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[4] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Auto-encoding variational bayes. In *arXiv preprint*, 2013.

[5] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017.

[6] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in neural information processing systems*, pages 469–477, 2016.

[7] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *NIPS*, 2016.

[8] pages=427–436-year=2008 Mazurowski, Maciej A., Piotr A. Habas, Jacek M. Zurada, Joseph Y. Lo, Jay A. Baker, and Georgia D. Tourassi, booktitle=Neural networks 21, no. 2-3. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance.

[9] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. In *arXiv preprint arXiv:1411.1784*, 2014.

[10] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2242–2251. IEEE, 2017.