Exploring the Efficacy of Sinkhorn Loss in CycleGAN Frameworks for Image-to-Image Translation

G082 (s2596768, s2575976)

Abstract

This report explores the integration of Optimal Transport (OT) theory into the CycleGAN framework for unpaired image-to-image translation. We define a novel framework for training CycleGANs which incorporates the Sinkhorn loss into the training objective. The Sinkhorn loss serves as a metric for how closely translated images resemble the target domain. We analyze the efficacy of our new approach by testing it on a variety of datasets and tasks. We compare these results to results obtained using an implementation of the traditional CycleGAN architecture without the Sinkhorn loss. Finally, we propose further research directions into new methods of incorporating OT theory into CycleGAN model architectures.

1. Introduction

Generative modeling is a machine learning task of growing importance. As we apply machine learning to increasingly sophisticated problems, we often aim to learn functions mapping inputs to output domains that are far more complex than simple class labels (Li et al., 2017).

However, training generative models presents a significant challenge. Simpler learning problems such as classification have a clear notion of “right” and “wrong,” a formulation for which approaches based on minimizing the corresponding loss functions has been tremendously successful. In contrast, generative model training is far more complex because it is often unclear how “good” a sample from the model is (Li et al., 2017).

Generative Adversarial Networks (GANs) have been proposed to address this issue (Goodfellow et al., 2014). In a nutshell, the idea behind GANs is to learn both the generative model and the loss function at the same time. The resulting training dynamics can be thought of as a game between two models: the generator and the discriminator discriminator. The goal of the generator is to produce realistic samples that fool the discriminator, while the discriminator is trained to distinguish between the true training data and samples from the generator. GANs have shown promising results on a variety of tasks, and there is now a large body of work that explores the power of this framework (Goodfellow, 2017).

While powerful, GANs are known to be notoriously hard to train. In order for a GAN to train properly, the generator and the discriminator must learn at a comparable rate, with neither model outpacing the other. The standard strategy for stabilizing training is to carefully design the model, either by adapting the architecture (Radford et al., 2016), selecting an easy-to-optimize objective function (Yuan et al., 2018), or meticulous hyperparameter tuning.

One task to which GANs have been applied is the task of image-to-image translation. In image-to-image translation, input images from a source domain are translated to a target domain in a way that retains their fundamental characteristics. Traditional GANs struggled with this challenge, often exhibiting a phenomenon known as mode collapse wherein the generator learns to return the same sample regardless of its input, rather than learning to map input samples to unique outputs in the target domain (Metz et al., 2017). The CycleGAN model introduced the idea of cycle-consistency loss to remedy this issue, requiring the existence of a reverse mapping from the generated sample back to the original image.

As such, the CycleGAN framework uses two generators G and F along with two discriminators D_X and D_Y, one for each domain. G learns the mapping from $X \rightarrow Y$ while F learns the inverse mapping from $Y \rightarrow X$ (Gupta et al., 2022). The cycle-consistency loss can thus be calculated as the difference between an image x and the same image mapped to the target domain and back $G(F(x)$ (Yuan et al., 2018).

The addition of another generator and another discriminator makes for even more unstable training dynamics than traditional GANs, with a total of four networks that must be learned simultaneously. If any of the models learn quicker than the others, they will fail to provide meaningful gradients to the other networks and the model may not converge. A common scenario is known as discriminator saturation, whereby the discriminator quickly learns to distinguish between real and fake images and fails to provide meaningful gradients to the generator (Saad et al., 2023). Moreover, despite the introduction of the cycle-consistency loss, CycleGANs are also prone to mode collapse as well. Papers following the original paper have at times struggled to recreate the results, and changes as small as the learning rate scheduler can produce drastic changes in results (Wang, 2023).

A mathematical formalization of the domain translation problem is provided by Optimal Transport (OT) theory. By framing the problem of image-to-image translation as optimally transporting one probability distribution to another, OT theory represents an alternative method of specifying a metric over probability distributions and can be used as an objective for training generative models (Salimans et al., 2018). The Sinkhorn algorithm, also known as Sinkhorn normalization (Cuturi, 2013) is one numerical method used to efficiently compute the optimal transport (OT) matrix in optimal transport problems.

In this paper, we propose the integration of the Sinkhorn loss into the existing CycleGAN framework to address the persistent instability issues that have plagued traditional training methods. The core problem we're tackling is the lack of a robust mechanism to ensure a one-to-one correspondence between samples from different domains during training. This instability often results in mode collapse or poor convergence, severely limiting the effectiveness of CycleGANs in image-to-image translation. By embracing OT theory, we aim to introduce a sophisticated way to model the complex relationships between domain samples. Sinkhorn loss, an OT metric which can be quickly computed on modern GPUs, facilitates the calculation of an optimal transport plan, effectively guiding the alignment process between domains.

Through the use of Sinkhorn loss, our Sinkhorn CycleGAN aims to preserve the essential characteristics of each domain, resulting in more accurate translations and avoiding issues like mode collapse and discriminator saturation. By incorporating optimal transport theory and Sinkhorn loss into CycleGANs, we aim to increase the robustness and quality of the learned translation. We hope to provide a solid framework that enhances the applicability and reliability of CycleGANs to various real-world tasks.

2. Data set and Task

2.1. Task

Unpaired image-to-image translation is a challenging task in computer vision which requires learning a function which maps input images from a source domain X to a target domain Y. This task can be formulated mathematically as follows: given a source image $x \in X$ and a target domain Y, the task is to learn a mapping $G : X \to Y$ such that the distribution of images in $G(X)$ is indistinguishable from that in Y. Importantly, the mapping $G : X \to Y$ must also preserve the essential content of the source image x. The idea of "preserving the essential content" of the source image can be formalized by requiring that a reverse mapping $F : Y \to X$ be possible as well. This ensures that the translated image retains the core attributes of the original image. Unpaired image-to-image translation is a complex task because it requires the model to be able to distinguish between what characteristics are specific to the image and which ones are specific to each domain. It must then effectively translate between domains while maintaining the underlying structure of the images.

2.2. Data Description

In order to test the effectiveness of our proposed approach, we evaluate our CycleGAN model on three different datasets. The chosen datasets cover a diverse range of style domains, each presenting unique challenges and opportunities for image-to-image translation.

The Horse2Zebra dataset was originally assembled by Zhu et. al. as a test for their CycleGAN implementation. It contains images of 1,334 images of horses and 1,067 images of zebras, requiring the model to learn the distinctive coloration, texture, and overall appearance of the animals. Correct translation requires accurate segmentation of the animal and correct translation of texture between the two animals, all while leaving the background unchanged (Zhu et al., 2017)

The "Photo to Monet" dataset contains collections of digital photographs alongside a set of 1,074 landscape paintings by Claude Monet. The dataset requires transforming photographs into artwork reminiscent of Claude Monet's iconic style, emphasizing vibrant colors and fluid brushstrokes (Zhu et al., 2017).

Finally, the "10 Big Cats of the Wild - Image Classification" dataset contains close up photos of ten species of large felines. We chose this dataset because it constitutes a similar task to the Horse2Zebra dataset, but it has not been studied extensively. We chose to focus on two species of felines, tigers and pumas, because of the contrast between the distinctive patterning and coloration of the tiger and the plain brown fur of the puma. The smaller size of the tiger and puma training sets (just 237 images) also posed a good challenge for the ability of our model to work with limited data (big, 2023).

For the Horse2Zebra and Photo to Monet datasets, images were resized to 128x128x3 for computational efficiency. For the 10 Big Cats of the Wild - Image Classification dataset, images could be kept at size 224x224x3 because of the smaller training set size. By incorporating these varied datasets, we test the versatility and adaptability of the Sinkhorn CycleGAN in facilitating image-to-image translation across a range of artistic styles and subject matters.

3. Methodology

3.1. CycleGAN

The CycleGAN architecture for unpaired image-to-image translation consists of two generator networks, $G : X \to Y$ and $F : Y \to X$, and two discriminator networks, D_X and D_Y. The discriminator networks learn to distinguish images from the source domain from images translated between domains by the generators. In contrast, the generators learn to produce images that more closely remember the target domain, with the objective of minimizing their adversarial loss. The adversarial loss $\mathcal{L}_{\text{GAN}}$ is calculated as a function of the sum of the number of generated images the discriminator correctly identifies as fake and the number of real images the generator correctly identifies as real.

$$\begin{aligned}\mathcal{L}_{\text{GAN}}(G_X, D_Y, X, Y) = {} & \mathbb{E}_{y\sim p_{\text{data}}(y)}[\log D_Y(y)] \\ & +\mathbb{E}_{x\sim p_{\text{data}}(x)}[\log(1 - D_Y(G_X(x)))]\end{aligned} \tag{1}$$

In order to enforce the bi-directionality of the mapping, another term known as the cycle-consistency loss is introduced. The $\mathcal{L}$cycle term quantifies the difference between the original image x and the same image mapped to the target domain and back $F(G(x))$. This loss function ensures that only the style is altered in the domain translation while the content remains preserved.

$$\begin{aligned}\mathcal{L}\text{cycle}(G, F) = {} & \mathbb{E}x \sim p_{\text{data}}(x)[|F(G(x)) - x|1] \\ & +\mathbb{E}y \sim p_{\text{data}}(y)[|G(F(y)) - y|_1].\end{aligned} \tag{2}$$

Finally, a term known as the identity loss is introduced to ensure that images that are already in the target domain are mapped to themselves, as no transformation is required:

$$\mathcal{L}_{\text{identity}}(G, F) = \mathbb{E}_{y\sim p_{\text{data}}(y)}[||G(y)-y||_1]+\mathbb{E}_{x\sim p_{\text{data}}(x)}[||F(x)-x||_1], \tag{3}$$

Together, these terms are combined to form the loss function to be optimized, where λ_c and λ_{id} are weighting parameters.

$$\begin{aligned}\mathcal{L}_{\text{CycleGAN}} = {} & \mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) \\ & + \mathcal{L}_{\text{GAN}}(F, D_X, Y, X) \\ & + \lambda_c \mathcal{L}_{\text{cycle}}(G, F) \\ & + \lambda_{id} L_{\text{identity}}(G, F)\end{aligned} \tag{4}$$

Generally, F, G, D_X, D_Y are represented using neural networks. The discriminators are trained to optimize number of images they correctly classify as real or fake, formalized as follows:

$$\begin{aligned}\mathcal{L}_{\text{Discriminator}}(D_Y, G, X, Y) = {} & - \mathbb{E}_{y\sim p_{\text{data}}(y)}[\log D_Y(y)] \\ & - \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log(1 - D_Y(G(x)))]\end{aligned} \tag{5}$$

Notably, none of the loss terms in the CycleGAN framework constitute a direct evaluation of whether images outputted by the generators fit into the target domain. This is enforced by the discriminators, whose task it is to determine whether the outputs of the generators do indeed belong to the target domain.

3.2. Optimal Transport and the Sinkhorn Loss

A more formal approach to quantifying the difference between the generated and target images is offered by Optimal Transport (OT) theory (Villani, 2008). Given two probability distributions μ and ν on spaces X and Y, respectively, and a cost function $c : X \times Y \to \mathbb{R}$ that quantifies the expense of transporting a unit of mass from X to Y, the OT problem seeks a transport plan γ that minimizes the total cost of moving μ to match ν as closely as possible:

$$\min_{\gamma\in\Pi(\mu,\nu)} \int_{X\times Y} c(x, y) d\gamma(x, y), \tag{6}$$

Solving this problem provides a measure of the difference between the two probability distributions. However, it is generally extremely computationally intensive because it involves comparing the distance between all pairs of points in the two distributions.

The Sinkhorn loss provides a mathematically grounded metric for calculating the OT distance between the generated and target image distributions (Genevay et al., 2017). An approximation of the more complex Wasserstein distance, the Sinkhorn loss utilizes a concept known as "entropic regularization" which makes it significantly more efficient to calculate and, importantly, differentiable (Chizat et al., 2020). The Sinkhorn loss is said to "regularize" the Wasserstein distance, smoothing the transport map and encouraging the generation of a mapping that is differentiable. Recently, a new algorithm for computing the Sinkhorn loss has further sped up its implementation (Qiu et al., 2022), making it an opportune time to experiment with its implementation in algorithms using stochastic gradient descent.

$$\mathcal{L}_{\text{Sinkhorn}}(\mu, \nu) = \text{Sinkhorn}(\mu, \nu; \epsilon, \lambda), \tag{7}$$

3.3. Our Formulation

The relative computational efficiency and differentiability of the Sinkhorn loss make it an ideal choice to incorporate into neural network training. In our CycleGAN framework, we add the Sinkhorn loss as an additional term in the cost function to optimize. With this addition, we hope to enforce the fidelity of the translated images to the target domain.

In our case, μ and ν correspond to batches of real and generated images $\mu \in Y$ and $\nu = G(y)$. As we iterate over batches of images, the Sinkhorn loss between the distribution of source images will be compared to the distribution of generated images using the Sinkhorn loss. This loss will be integrated into the CycleGAN loss function, incorporating some of the principles of optimal transport directly into the learning of our model. The new loss function for our Sinkhorn CycleGAN thus becomes:

$$\begin{aligned}\mathcal{L}_{\text{Sinkhorn-CycleGAN}} = {} & \mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) \\ & + \mathcal{L}_{\text{GAN}}(F, D_X, Y, X) \\ & + \lambda_c \mathcal{L}_{\text{cycle}}(G, F) \\ & + \lambda_{id} \mathcal{L}_{\text{identity}}(G, F) \\ & + \lambda_{sink} \mathcal{L}_{\text{sinkhorn}}(G, F, X, Y)\end{aligned} \tag{8}$$

where $\lambda_{\text{Sinkhorn}}$ denotes the weighting parameter that controls the influence of the Sinkhorn loss, $\mathcal{L}_{\text{Sinkhorn}}$, on the overall training objective. The Sinkhorn loss becomes increasingly complex to calculate as the size of the batches of images increases, so we elect to process images in batches of 8 to balance sample diversity and computational efficiency. Our training process is outlined in Algorithm 1.

Algorithm 1 Training CycleGAN with Sinkhorn Loss

```
Require: Source domain images $X$, Target domain images $Y$
Require: Initialized models $G$, $F$, $D_X$, $D_Y$
Require: Hyperparameters: learning rate, $\lambda_{cyc}$, $\lambda_{id}$, $\lambda_{sink}$
for each iteration do
    Sample batch $\{x_i\}_{i=1}^m$ from $X$
    Sample batch $\{y_j\}_{j=1}^m$ from $Y$
    Calculate batch of fake images $\{G_(x_i)\}_{i=1}^m$ from $X$ to $Y$
    Calculate batch of fake images $\{F(y_j)\}_{j=1}^m$ from $Y$ to $X$
    $L_{GAN} \leftarrow$ Compute adversarial loss
    $L_{cyc} \leftarrow$ Compute cycle consistency loss
    $L_{Sink} \leftarrow$ Compute Sinkhorn loss
    $L_{id} \leftarrow$ Compute identity loss
    $L_{Sinkhorn-CycleGAN} \leftarrow L_{GAN} + \lambda_{cyc}L_{cyc} + \lambda_{id}L_{id} + \lambda_{Sink}L_{Sink}$
    Update $G$, $F$ by descending $\nabla_G L_{Sinkhorn-CycleGAN}$
    Update $D_X$, $D_Y$ by descending $\nabla_D L_{Discriminator}$
end for=0
```

For our experiments, we mirror the architecture described by Zhu et. al. in the original CycleGAN paper (Zhu et al., 2017). Our generator network contains three convolutional layers, six residual layers, two fractionally strided convolutions, and one final convolution mapping features back to RGB output. Instance normalization and ReLU activation functions are implemented after each of the convolutional layers. Our discriminator networks implement the PatchGAN discriminator architecture, which divide the image into overlapping patches each of which is individually classified as real or fake (Isola et al., 2018).

4. Experiments

4.1. Implementation and Hyperparameters

We carried out experiments on three datasets in order to determine whether the addition of the Sinkhorn loss had a positive effect in the training of CycleGANs. Code was written in Python and implemented using PyTorch. All experiments were run using the Adam Optimiser with a learning rate of 0.0002 and $\beta = (0.5, 0.999)$. We also set $\lambda_c = 10$ and $\lambda_{id} = 5$, in alignment with Zhu et. al. Experiments were run for 100 epochs, with the exception of experiments run on the 10 Big Cats dataset, which were run for 200 epochs because of the smaller size of the dataset (236 images per class). Models were run with a batch size of 8 to balance computational efficiency and memory space. Experiments were run with $\lambda_{sink} = 0, 1e-4, 2e-4$. The values of λ_{sink} were chosen to scale the sinkhorn term to within the range of the other loss terms, so as not to overpower their effect completely. Experiments on the Photo to Monet dataset and the Big Cats dataset were run with $\lambda_{sink} = 1e-4$ as we observed greater success with this value in our baseline testing on the Horse2Zebra dataset.

4.2. Evaluation Metrics

There are few good standard metrics for evaluating generative models such as CycleGANs. Ultimately, the most important benchmark for success is subjective human judgement as the quality, fidelity, and authenticity of generated images is difficult to capture in numerical metrics. For application to specific tasks such as medical imaging, domain specific metrics are generally preferred to assess the quality of the output images (Dou et al., 2018).

To assess a model's performance on more general tasks, however, there are a few metrics that can be used. To evaluate the ability of the model to maintain the content of the original image in the new domain we use Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM). PSNR ranges between 0 and 100 and is a relatively rudimentary metric that measures the peak error percentage between the generated image and the target image. SSIM is also a metric of image reconstruction quality which aims to account for differences in luminance, contrast, and brightness as well. The output of an SSIM evaluation is a value between -1 and 1 indicative of the correlation between pixels in the input image and the output image. Higher values of PSNR and SSIM indicate that the model is more capably mapping features of the input domain to the output image. It is important to note that with CycleGAN models, unlike with image sharpening or reconstruction, it is not expected or even ideal to maximize PSNR or SSIM values, as maximal values for these metrics would indicate that no mapping whatsoever is occurring. Minimum values, on the other hand, may indicate that the model is overly fitting to the target domain.

In order to evaluate the ability of the model to output images that resemble the target domain, we will use the Frechet Inception Distance (FID) (Heusel et al., 2018). FID evaluates the similarity between two datasets of images by comparing the feature vectors extracted by a pre-trained Inception network. We will use FID to compare the features of our real test dataset and a set of generated images derived from the test set. A lower FID score is indicative of higher similarity between the images. While FID can provide a useful measure for dataset similarity, it is not without its limitations and lower FID numbers may not always correspond to images that appear to be more similar to the human eye. Metrics for all runs are displayed in Table 1.

4.3. Horses to Zebras

We began by testing on the Horses2Zebras dataset. This dataset has served as a useful benchmark for style translation since it was assembled by Zhu et. al. in their original Cycle GAN paper (Zhu et al., 2017). In this task, the two domains X and Y represent images of horses and zebras. The task presents an interesting test case because it requires the model to segment the image correctly, locating the body of the animal and translating the patterning. While the authors of the original paper showed impressive results, other papers have struggled since then to recreate the re-

Table 1. Experimental Results

Experiment	Dataset	λ_{sink}	Metrics					
			PSNR		SSIM		FID	
			A → B	B → A	A → B	B → A	A	B
1	Horses → Zebras	0	16.4	15.5	0.435	0.413	188.1	158.5
2	Horses → Zebras	$1e-4$	9.25	8.54	-0.104	-0.140	242.1	265.6
3	Horses → Zebras	$2e-4$	8.78	8.51	-0.127	-0.175	232.4	240.6
4	Pumas → Tigers	0	16.2	16.7	0.399	0.512	201.7	376.4
5	Pumas → Tigers	$1e-4$	9.18	8.92	-0.146	-0.067	315.9	432.8
7	Photo → Monet	0	0.223	0.169	11.3	12.0	304.5	273.6
8	Photo → Monet	$1e-4$	0.219	0.209	12.9	12.5	248.1	195.9

sults (Wang, 2023). As such, testing the efficacy of our baseline model was vital before proceeding.

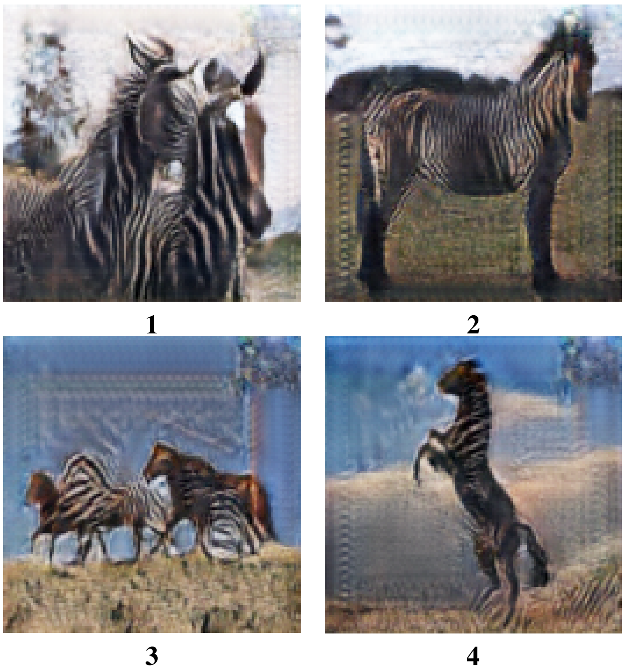

Figure 0. Horses to Zebras translations with $\lambda_{sink} = 0$

Given that the architecture was initially designed and tested on the dataset, it is unsurprising that our model fared best on the Horses to Zebras translation task. The baseline model with $\lambda_{sink} = 0$ achieved a PSNR score of 16.4 and an SSIM score of 0.435 for translating horses to zebras. These results, while unspectacular, suggest that the model is overall succeeding in the task of translating the information from the source image to the target domain.

However, there are still a variety of cases for which the model struggles. While it often performs admirably at locating the horse, segmenting it, and applying the patterning when the horses are in common positions, it struggles when it encounters horses in an unusual poses or when attempting to segment multiple overlapping horses. In these cases, it tends to add the characteristic black and white patterning of the zebra ad-hoc to the image, as shown in images 3 and 4 in Figure 0.

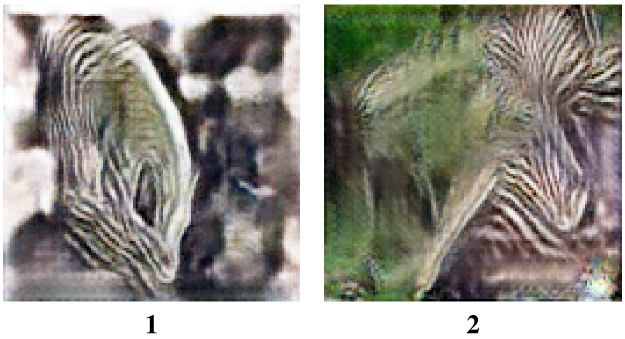

Figure 1. Example translations on the Horses to Zebras task performed by Sinkhorn CycleGAN with $\lambda_{sink} = 1e-4$

In comparison, the Sinkhorn CycleGAN implementation achieved less impressive results. The PSNR scores of 9.25 and 8.78 for the $\lambda_{sink} = 1e-4$ and $\lambda_{sink} = 2e-4$ models suggest that there is a high degree of error between the generated images and the original images. Negative SSIM scores mean that there is a negative correlation between the input image and the output images, indicating that the model is overfitting to the target domain. Looking at the example images in Figure 1, we can see that the model has focused on adding the characteristic black and white patterning of the zebra to the image but is failing to segment the zebra properly.

The difference between the FID scores for the two models is not as pronounced. While the $\lambda_{sink} = 0$ model achieves translations that are on average slightly more faithful to the features of the output domain, none of the models are producing images that the Inception model believes to be overly close to the target domain.

Both models perform better at the task of mapping horses to zebras, likely owing to the black and white patterning of the zebra making it difficult for the model to perform the segmentation of the animal.

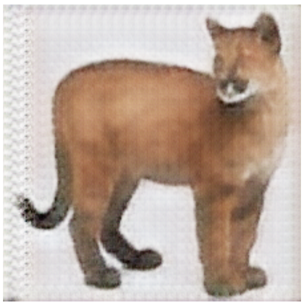

Figure 2. Failed Puma → Tiger translation outputted by the model with $\lambda_{sink} = 0$

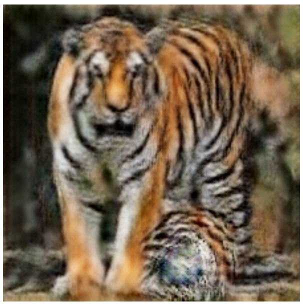

Figure 3. Failed Tiger → Puma translation outputted by the model with $\lambda_{sink} = 0$

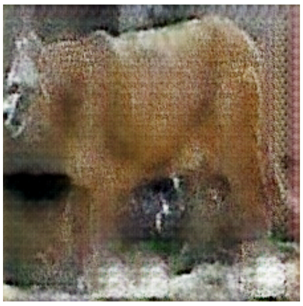

Figure 4. Tiger translated with regular cycle GAN

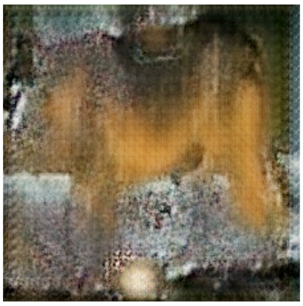

Figure 5. Tiger translated with sinkhorn Cycle GAN

4.4. Pumas to Tigers

We further tested our model on the Big Cats dataset. The PSNR and SSIM scores achieved by the $\lambda_{sink} = 0$ model on the Pumas to Tigers translation task rival that achieved on the Horses to Zebras task, indicating that the model is creating images that retain the characteristics of the input domain. The scores of SSIM = 0.512 and PSNR = 16.7 achieved for the reverse translation from Tigers to Pumas are the highest marks for any of the models tested.

The FID scores of 201.7 and 376.4, however, indicate that the output images are not characteristic of the target domain. Upon inspecting the images produced by the model in Figure 2 and Figure 3, it is clear that the model is not producing meaningful translations at all, instead outputting images that are basically unchanged from their inputs. This occurs because of the reliance of the model on cycle-consistency loss and identity loss. Although these terms are vital for ensuring the feasibility of reverse translations, they do not ensure that the model accurately translates the image into the target domain. In this case, fast convergence of the cycle and identity loss caused the model to overemphasize the importance of retaining the input domain features while failing to learn the correct translation to the target domain.

The $\lambda_{sink} = 1e - 4$ implementation achieved significantly worse numerical metrics, again getting negative SSIM scores for the transformations in both directions. However, visual comparison of images produced by the models shows that the addition of the Sinkhorn loss term did encourage the model to attempt to add the distinctive orange and black coloring of the tiger to the body of the animal. While the high FID scores emphasize that the translation does not maintain many other characteristics of the animal, the Sinkhorn loss term does appear to cause the model to more faithfully emulate the coloration and patterning of the target domain.

4.5. Photo to Monet

In the Photo to Monet task, the implementation with the Sinkhorn loss actually achieved better metrics than the no Sinkhorn control. The $\lambda_{sink} = 1e - 4$ implementation achieved similar PSNR and SSIM scores but significantly higher FID scores. Investigation of the images, however, reveals that this is a case where numerical metrics may be misleading.

Although the images generated with $\lambda_{sink} = 1e - 4$ do display coloration and patterning characteristic of Monet's work, they do not retain the fundamental characteristics of the image and tend to blur it significantly. In this case, the Sinkhorn loss term appears to have resulted in images that blend together many of the characteristics into one "mask" that is overlayed over all images regardless of the input, as seen in Figure 6. This likely occurs because the Sinkhorn loss calculation is happening for batches of data, meaning that the generators are encouraged to output images that exhibit characteristics of the target domain averaged over this batch.

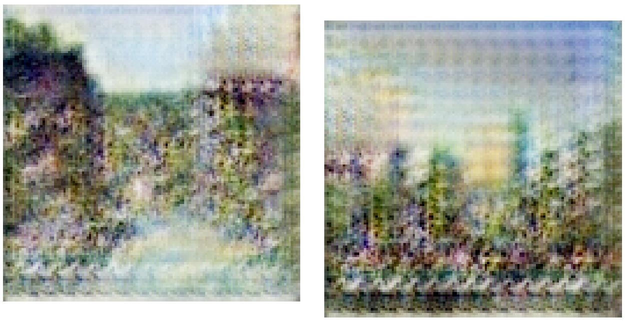

Figure 6. Monets generated with $\lambda_{sink} = 1e - 4$

4.6. Discussion

Across the datasets, we found that the addition of the Sinkhorn loss term had mixed results. While the models ran with the Sinkhorn loss term successfully maintained the necessary color gradients, they struggled to accurately map the content between different domains. One potential reason for this discrepancy stems from the pixel-to-pixel style transfer approach for calculating the Sinkhorn loss we implemented in our model.

In our setup, the Sinkhorn loss directly compares pixel values of the generated images and the target domain, causing the models to seek to minimize the difference between pixel values at similar locations in the image. This approach, while well-intentioned, appears to be oversimplistic for the problem at hand. The Sinkhorn loss encouraged the network to focus on matching pixel values rather than aligning significant features and distributions in the data. Preserving coloration and pixel location is important for image-to-image translation, but an emphasis on these two alone may lead to loss of higher level features of images that are more important for human perception.

We believe that rethinking the utility of Sinkhorn as a technique to regularize how probabilities are mapped could lead to better translation results. Applying the Sinkhorn loss to feature representations rather than raw pixel values may be one path towards more meaningful comparisons between image batches and result in better model convergence.

These insights underscore the complexity of achieving nuanced style transfer through CycleGANs and highlight avenues for future research, such as changing the methodology for calculating the Sinkhorn loss term, refining the training regimen or modifying the network architecture to more faithfully capture and replicate the textural attributes of the target domain, thereby enhancing the overall fidelity of the translations.

5. Related work

The combination of Optimal Transport methods with the CycleGAN framework has been gaining traction. Wasserstein GANs modify the traditional GAN framework by replacing the discriminator architecture used in GAN frameworks with a "critic" which provides feedback on the optimal transport distance required to transport the generated samples to the real samples (Arjovsky et al., 2017).

Wasserstein CycleGANs extend the Wasserstein GAN framework to image-to-image translation (Hu et al., 2018). They have been proposed to solve issues with CycleGAN training such as discriminator saturation, as the OT distance will always provide meaningful gradients regardless of the learning rate of the discriminator. This approach contrasts with our Sinkhorn CycleGAN which continues to implement discriminators as before but adds the Sinkhorn loss as another term in the generator loss function. The Wasserstein CycleGAN can be slower to train than traditional CycleGANs, so research into optimizing the calculation of the Wasserstein distance in the training loop using approximations has been conducted (Gulrajani et al., 2017) (Deshpande et al., 2018) (Wu et al., 2019).

Other approaches to integrating optimal transport theory into CycleGANs have aimed to ensure the one-to-one nature of the CycleGAN's learned mapping by using optimal transport (Sim et al., 2020). Recent advances in Optimal Transport theory have even seen the the optimal transport plan used directly as a generative model itself, as proposed by Korotin et. al. (Korotin et al., 2023). This approach yielded visually impressive results, although it was only tested on datasets of size greater than 50,000 images, which significantly limits its potential for generalization.

One approach to integrating optimal transport theory with CycleGANs that we believe merits further investigation is to apply the Sinkhorn loss to feature representations of images rather than the full images. Feature representations of images can be used to capture higher level characteristics of the images, which could provide a more useful basis for comparison between images and their target domains. This change could potentially mitigate some of the issues our Sinkhorn CycleGAN encountered and allow the model to more accurately recreate images in the target domain.

This approach would require the use of a neural network pre-trained on a large dataset of images, such as the ImageNet dataset, as a feature extractor. The Sinkhorn loss could then be calculated between the feature representations of the images, rather than the raw pixel data. Methods to apply optimal transport to feature representations have been proposed before (Li et al., 2021). To our knowledge, however, no one has proposed the use of feature level comparison to guide optimal transport mappings for domain translation. Introducing feature representations would introduce another layer of abstraction to an already complex process, however, potentially complicating training dynamics and requiring extensive hyperparameter tuning.

6. Conclusions

Our experiments into the efficacy of our Sinkhorn CycleGAN framework yielded mixed results. While the integration of the Sinkhorn loss displayed potential in enforcing color and stylistic attributes of the target domain, it limited the learned model's ability to retain higher level features of the input image. For tasks where both models struggled such as the Pumas to Tigers translation task, models run with the Sinkhorn loss displayed a better ability to implement the coloration and style of the output domain in some cases. However, this improvement of often came at the expense of retaining the underlying structure of the input image, a significant limitation which was exemplified by its poor performance on the Horses to Zebras task.

These findings highlight the difficulty of the image-to-image translation task and the delicate balance that must be struck to solve it properly. The direct pixel-level calculation of the Sinkhorn loss over batches of images skews the model towards a mapping which emphasizes the color and patterning of the target domain over retaining the features of the input image.

These limitations lead us to propose future research into applying the Sinkhorn loss to feature-level representations of images. This approach could lead to the model learning a more nuanced mapping that is more faithful to higher-level characteristics of the target domain. Further exploration into the best implementation of this approach, along with the correct balance between the Sinkhorn loss term and the other loss terms, could be a promising route for future research.

References

10 Big Cats of the Wild - Image Classification

Dataset. https://www.kaggle.com/datasets/gerry/10-big-cats-of-the-wild-image-classification, 2023. Accessed: 10/03/23.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 214–223. PMLR, July 2017. URL https://proceedings.mlr.press/v70/arjovsky17a.html. ISSN: 2640-3498.

Lénaïc Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster Wasserstein Distance Estimation with the Sinkhorn Divergence. In *Advances in Neural Information Processing Systems*, volume 33, pages 2257–2269. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/17f98ddf040204eda0af36a108cbdea4-Abstract.html.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances, 2013.

Ishan Deshpande, Ziyu Zhang, and Alexander G. Schwing. Generative Modeling Using the Sliced Wasserstein Distance. pages 3483–3491, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Deshpande_Generative_Modeling_Using_CVPR_2018_paper.html.

Qi Dou, Cheng Ouyang, Cheng Chen, Hao Chen, and Pheng-Ann Heng. Unsupervised Cross-Modality Domain Adaptation of ConvNets for Biomedical Image Segmentations with Adversarial Loss, June 2018. URL http://arxiv.org/abs/1804.10916. arXiv:1804.10916 [cs].

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning Generative Models with Sinkhorn Divergences, October 2017. URL http://arxiv.org/abs/1706.00292. arXiv:1706.00292 [stat].

Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks, 2017.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved Training of Wasserstein GANs, December 2017. URL http://arxiv.org/abs/1704.00028. arXiv:1704.00028 [cs, stat].

Rajeev Kumar Gupta, Santosh Bharti, Nilesk Kunhare, Yatendra Sahu, and Nikhlesh Pathik. Brain tumor detection and classification using cycle generative adversarial networks. *INTERDISCIPLINARY SCIENCES-COMPUTATIONAL LIFE SCIENCES*, 14(2):485–502, JUN 2022. doi: 10.1007/s12539-022-00502-6.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, January 2018. URL http://arxiv.org/abs/1706.08500. arXiv:1706.08500 [cs, stat].

Weining Hu, Meng Li, and Xiaomeng Ju. Improved cyclegan for image-to-image translation. *arXiv preprint arXiv:insert$_a$rxiv$_n$umber$_h$ere*, 2018.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-Image Translation with Conditional Adversarial Networks, November 2018. URL http://arxiv.org/abs/1611.07004. arXiv:1611.07004 [cs] version: 3.

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural Optimal Transport, March 2023. URL http://arxiv.org/abs/2201.12220. arXiv:2201.12220 [cs].

Jerry Li, Aleksander Madry, John Peebles, and Ludwig Schmidt. Towards understanding the dynamics of generative adversarial networks. *arXiv preprint arXiv:1706.09884*, 1, 2017.

Xuhong Li, Yves Grandvalet, Rémi Flamary, Nicolas Courty, and Dejing Dou. Representation Transfer by Optimal Transport, February 2021. URL http://arxiv.org/abs/2007.06737. arXiv:2007.06737 [cs, stat].

Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled Generative Adversarial Networks, May 2017. URL http://arxiv.org/abs/1611.02163. arXiv:1611.02163 [cs, stat].

Yixuan Qiu, Haoyun Yin, and Xiao Wang. Efficient, Stable, and Analytic Differentiation of the Sinkhorn Loss. September 2022. URL https://openreview.net/forum?id=uATOkwOZaI.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016.

Muhammad Muneeb Saad, Ruairi O'Reilly, and Mubashir Husain Rehmani. A Survey on Training Challenges in Generative Adversarial Networks for Biomedical Image Analysis, August 2023. URL http://arxiv.org/abs/2201.07646. arXiv:2201.07646 [cs].

Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving gans using optimal transport, 2018.

Byeongsu Sim, Gyutaek Oh, Jeongsol Kim, Chanyong Jung, and Jong Chul Ye. Optimal Transport driven CycleGAN for Unsupervised Learning in Inverse Problems, August 2020. URL http://arxiv.org/abs/1909.12116. arXiv:1909.12116 [cs, eess, stat].

Cédric Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer, 2008.

Junqing Wang. The study of object transformation based on cyclegan. In *Fifth International Conference on Computer Information Science and Artificial Intelligence (CISAI 2022)*, volume 12566, pages 82–87. SPIE, 2023.

Jiqing Wu, Zhiwu Huang, Dinesh Acharya, Wen Li, Janine Thoma, Danda Pani Paudel, and Luc Van Gool. Sliced wasserstein generative models, 2019.

Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. June 2018.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.