CS230

Generative Adversarial Network (GAN) Based Translation Between Medical Image Modalities

Aditi Maheshwari
Department of Material Sciences
Stanford University
aditi17@stanford.edu

Ketan Pant
Department of Electrical Engineering
Stanford University
ketan94@stanford.edu

Abstract

The project leverages capabilities of modern deep learning techniques such as GANs and its variants to address the issue regarding the scarcity of utilizable medical image data. To undertake this challenge, we evaluate the Unsupervised Image to Image Translation network, or UNIT on a brain MRI dataset from the Human Connectome project and the capability to translate between T1 and T2 modalities. Through careful inspection, we improve this method by incorporating self-attention layers, spectral normalization, and charbonnier penalty. The report describes the effectiveness of these techniques in the results obtained.

1 Introduction

In the medical community, utilizable image data is scarce because of the high costs of acquisition, rarity of patient conditions and patient confidentiality. This issue affects the generalizing capability of machine learning models trying to solve various problems like classification and segmentation in the medical domain. We aim to create utilizable translations of medical images between different modalities that would reduce the number of times medical devices would need to be used and increase the data available to doctors. This would help the machine learning community utilize disjointed data sets together. For example, if a certain MRI dataset is publicly available and if we could generate the corresponding CT scans for every image in the dataset, researchers can use both sets of images to further improve their results.

For this project, we are focusing on the translation between T1-weighted and T2-weighted brain MR images. The basic difference in the T1-weighted and T2-weighted images is that in T1-weighted, tissues with high fat content appear bright and compartments filled with water appear dark while it is the opposite in T2-weighted images. We are using a popular model used for image to image translation called UNIT proposed by Liu et al. (2), to do the translation across T1 and T2 domains as presented by Welander et al. (3). Further improvements on this baseline architecture are proposed to enhance the model's capability to reconstruct complicated, but necessary brain structures by using techniques such as self-attention and spectral normalization.

2 Related Work

There is currently active research being conducted using GANs for generating synthetic data and doing image to image translations. For our baseline, we have replicated the results presented by Welander et al. (3) using the same dataset. The code used by the authors to implement the reults of the paper is available on Github (4). The code used by us can be found at Project Code. Here, the

CS230: Deep Learning, Winter 2019, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

UNIT architecture proposed by Liu et al. (2) and CycleGAN proposed by Zhu et al. (5) are used for image translation between T1-weighted and T2-weighted brain MR images. In our project, we are only implementing the UNIT architecture. UNIT and CycleGAN were originally proposed to solve problems due to the unavailability of datasets consisting of paired images for translation. Their results show feature translation in images that include animals, faces, street views, seasons etc. In the medical imaging domain, Wolterink et al. (6) proposed a GAN consisting of two synthesis CNNs and two discriminator CNNs to translate MR images of the brain to CT images using unpaired data. Armanious et al. (7) presented a novel architecture called MedGAN which can be used for translation between PET and CT scans. Architecture of MedGAN is based on a popular model called UNet by Ronneberger et al. (8) which is widely used for segmentation of medical images.

3 Dataset and Features

We are using the dataset provided by the Human Connectome Project (1). The dataset consists of T1 and T2-weighted 3D volumes of brain MR images of 1113 patients and also includes the brain segmentations. All the images have been registered to a common template brain, such that they are in the same position and of the same size. Similar to the baseline work by Welander et al. (3), the data were split into a training set of 900 images in each domain. The remaining 213 images, in each domain, were used for testing. Images are coloured with 304x256 resolution. Fig 1 shows an example of the ground truth image in each domain.

(a) T1 Ground Truth Image (b) T2 Ground Truth Image

Figure 1: Example from dataset.

4 Methods

In the baseline implementation, the model UNsupervised Image to image Translation, or UNIT by Liu et al. (2) is being used to transform MRI images across its T1 and T2 domains. The model achieves this by learning the joint distribution between the two domains. With consideration of coupling theory in (9), learning the joint distribution from its marginal distributions is an ill-posed problem due to infinite possible sets of joint distributions. The baseline algorithm mitigates this through the use of the shared latent space assumption, which speculates that corresponding images can be mapped to the same/similar latent representation given a shared latent space. The overall UNIT framework implements this through the use of Variational Autoencoders by Larsen et al. (10) with Coupled GANS in Liu Tuzel (11) or VAE-CoGANs. In the framework, two encoders are used to map the images onto a shared latent space by tying their last layers together as well as sharing the first higher levels of its two generators to produce domain-translated images. The framework is represented in Fig 2.

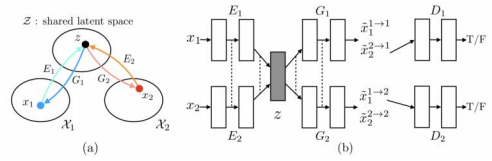


Figure 2: (a) The shared latent space assumption. (b) The proposed UNIT framework.

Liu et al. assume a pair of corresponding images (x_1, x_2) in two different domains χ_1 and χ_2 can be mapped to a same latent code z in a shared-latent space Z. E_1 and E_2 are two encoding functions, mapping images to latent codes. G_1 and G_2 are two generation functions, mapping latent codes to images as shown in Fig 1(a). E_1, E_2, G_1 and G_2 are represented using CNNs and implement the shared-latent space assumption using a weight sharing constraint where the connection weights of the last few layers in E_1 and E_2 are tied (illustrated using dashed lines) and the connection weights of the first few layers in G_1 and G_2 are tied. Here, $x_1^{1\to1}$ and $x_2^{2\to2}$ are self-reconstructed images, and $x_1^{1\to2}$ and $x_2^{2\to1}$ are domain-translated images. D_1 and D_2 are adversarial discriminators for the respective domains, in charge of evaluating whether the translated images are realistic.

After careful review, the architecture yielded good replication of larger brain structures such as the gyri, or the bumps on the edges of the brain and the sulci or grooves/fissures between tissues. However, in the lower center part of the brain, these structures were more detailed and the images generated started fused these components together leading to several inaccuracies in the translations.

In addressing this issues, the techniques from Zhang et al. (14) were investigated and applied to examine the effects it had on the UNIT architecture. The augmentations selected were the varying self-attention layers and Spectral Normalization. The self-attention layers as seen in Fig 3 is implemented with 1x1 convolutions for 3 different outputs. These 2 outputs are multiplied together and extract the attention maps via softmax to determine $\beta_{i,j}$, a mask interpreting the impact of location *i* in rendering location *j*. The third result is used as feedback to reflect the features currently used. This is represented with the equation $y_i = (\gamma)o_i + x_i$. γ was initialized to 0 so the model can first examine the local spatial information before refining the image. Spectral Normalization by Miyato et al. (16) is performed on the weights of each layer of the encoder, generator and discriminator of the model. It is a simple enough technique that does not require extra hyperparameter tuning. This method has been proven to use less discriminator updates per generator update and hence considerably reduced computational cost during training.

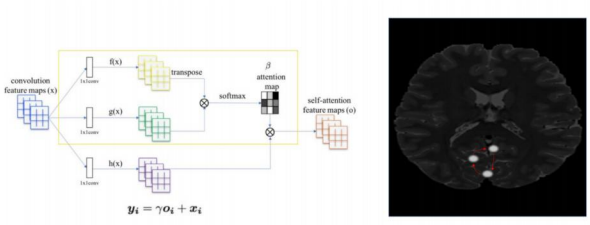


Figure 3: (a) Self-attention network layer with equation, (b) self-attention example

An additional modification conducted was the use of the charbonnier penalty function ($\rho(x) = \sqrt{x^2 + \epsilon^2}$) in order to regulate between L1 and L2 regularization as L2 regularization has been known to heavily penalize outliers.

5 Experiments and Results

We implemented different models as mentioned in Section 4 i.e., Self Attention, Spectral Normalization and different Loss Functions. We use ADAM with lr = 0.0001, β_1 = 0.5, β_2 = 0.999, binary cross entropy loss for the Generator and Discriminator, MAE loss for the reconstruction stream and VAE loss (MSE) for the encoder and cycle consistency constraint. Network architecture details are given in Table 3 of (2). We analyzed the results for each model qualitatively and quantitatively. The modified UNIT architecture inspired by (14) is shown in Fig 4. Fig 5 shows the translation and reconstruction of the T1 and T2 images for a single sample for the baseline model and our model that performed the best: Self-Attention UNIT with attention implemented in the encoder part of the model, with Spectral Normalization in the Encoder, Generator and Discriminator, and Charbonnier Loss Function. Figure 6(a) compares the loss of 3 different models - the baseline UNIT, UNIT with Charbonnier Loss and UNIT with Attention, for every epoch. Figure 6(b) compares the loss between the Self Attention UNIT and the Spectral Normalized UNIT models.

Fig 7(a) shows the quantitative results for various models implemented using a scoring method called Structural Similarity Index (SSIM) proposed by Wang et al. (15), at the 75th epoch of training. Fig

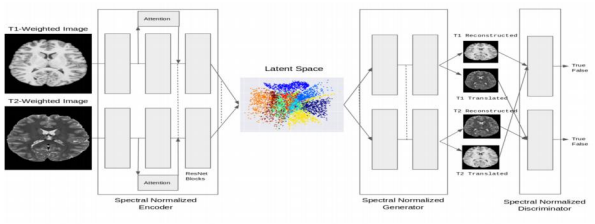


Figure 4: Modified UNIT architecture: Self-attention UNIT.

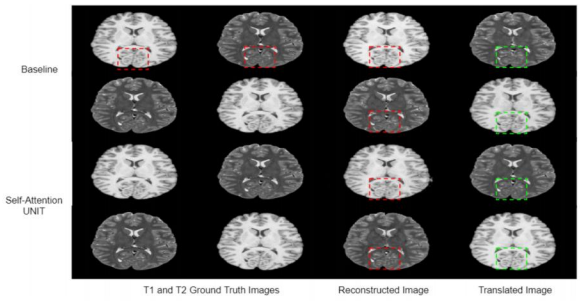
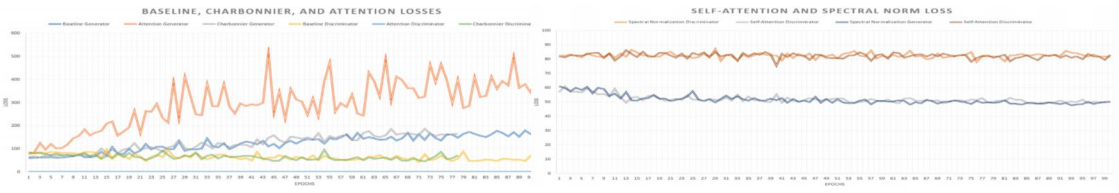


Figure 5: The translated images and reconstructed images for the baseline model (top) and Self-Attention UNIT (bottom).

7(b) is a table comparing the SSIM scores between the baseline UNIT model and our best model obtained (UNIT with Self Attention, Spectral Normalization and Charbonnier Loss) over epochs 1, 25, 50, 75, 100.



(a) Baseline, Charbonnier and Attention Loss (b) Self Attention and Spectral Normalization Loss

Figure 6: Generator and Discriminator Loss for different models.

6 Discussion

6.1 Qualitative Analysis

The main limitation that we observed in the translation performed by the baseline was that it did not reproduce the intricate details of the brain structure with good fidelity. As we can see in Fig 8, the bridge separating the left and right lobes is blurred in the translated images produced by the baseline. Our best model was qualitatively better at reproducing the intricate details. It is evident from Fig 8 that the translation performed by our best model reproduces the bridge separating the two lobes with better fidelity and the structures are clearer.

6.2 Analysis of the loss

As seen in Fig 6, the combined generator and combined discriminator loss with respect to the baseline and each respective modification has been illustrated across the different epochs. This has been done to determine when/where overfitting may between the generator and the discriminator. In the baseline, the overfitting appears to occur around 13 epochs, while charbonnier occurred around 11,

Model	SSIM	
	T1	T2
Baseline (RGB)	0.7563	0.716
Baseline (gray scale)	0.7617	0.7217
Baseline with Charbonnier loss (CL)	0.7584	0.7181
Baseline with Spectral Normalization (SN)	0.7108	0.7402
Attention in generator	0.7166	0.5743
Attention in generator with CL	0.7134	0.5256
Attention in generator with CL and SN	0.6722	0.6678
Attention in encoder with CL and SN	0.7443	0.6834

(a) SSIM scores for different models after 75th epoch

Model	Epoch No.	SSIM	
		T1	T2
Baseline	1	0.7596	0.7141
	25	0.724	0.7086
	50	0.7561	0.716
	75	0.7563	0.7145
	100	0.7567	0.716
Self-Attention UNIT (with spectral normalization, charbonnier loss)	1	0.6427	0.6245
	25	0.7396	0.6502
	50	0.7341	0.651
	75	0.7443	0.6834
	100	0.7486	0.7176

(b) SSIM scores for the baseline model and our best model over epochs 1, 25, 50, 75, 100.

Figure 7: Quantitative Results.

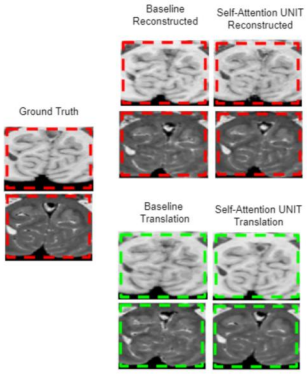


Figure 8: Qualitative Comparsion of Baseline and our best Model.

and attention occurred since the first epoch. However, after applying the self-attention layer with the spectral normalization and seeing its effect separately, we see almost no overfitting occurring as both the generator and discriminator were actively tricking each other and improving the result.

7 Conclusion/Future Work

We demonstrated improved fidelity in translating medical images from one modality to another using GANs. The biggest challenge was obtaining a quantitative evaluation metric that would reflect the results obtained qualitatively by the different models. We considered different evaluation scores like the Inception Score, Average Pixel Accuracy and Structural Similarity Index (SSIM). Among these, we decided to choose SSIM as we wanted to focus on how the images generated by the different models compared with the ground truth images with respect to the minute details and structures present in the brain. The SSIM is calculated between the ground truth image and the translated image generated by the model. The table in Figure 7 shows the different SSIM scores we obtained for the different models on the test dataset. The final scores mentioned in the table are the scores averaged over the images in the test set. According to the scores in table in Figure 7, our modified model after training for 100 epochs performs similar to the baseline model. However, as mentioned in the section describing the results qualitatively, we think our model actually performs better than the baseline with regards to the structural details. In this regard, in future, we would like to utilize some other metric to better quantify the results as it is difficult to optimize something which we are not able to quantify properly. Some other possible ways to obtain a score are by segmenting the smaller structures in the brain and then just comparing the segmented images. Another possible way is to use descriptors like SURF or SIFT to obtain features in the images and then compare them. Another analysis as shown in the table in the Figure 7(b) shows that our model performs better with training over the epochs and the baseline model did not improve that much after training over the epochs.

8 Contributions

Aditi Maheshwari: Initial literature review to select the project, setting up dependenicies to help replicate baseline code and intergrating charbonnier penalty in final work.

Ketan Pant: Initial literature review to select the project, literature review to add extensions to the baseline i.e Self-Attention, Spectral Normalization (SN) etc. and integrate self attention and SN in final work.

Putting everything together to a final model and comparing various models was a combined effort.

References

[1] Human Connectome Project. Connectome - homepage. URL https://www.humanconnectome.org/.

[2] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. CoRR, abs/1703.00848, 2017. URL http://arxiv.org/abs/1703.00848.

[3] Per Welander, Simon Karlsson, and Anders Eklund. Generative adversarial networks for image-to-image translation on multi-contrast MR images - A comparison of cyclegan and unit. arXiv preprint arXiv:1806.07777, 2018.

[4] Simontomaskarlsson. simontomaskarlsson/gan-mri, Oct 2018. URL https://github.com/simontomaskarlsson/GAN-MRI.

[5] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image trans- lation using cycle-consistent adversarial networks. CoRR, abs/1703.10593, 2017. URL http://arxiv.org/abs/1703.10593.

[6] Jelmer M. Wolterink, Anna M. Dinkla, Mark H. F. Savenije, Peter R. Seevinck, Cornelis A. T. van den Berg, and Ivana Isgum. Deep MR to CT synthesis using unpaired data. CoRR, abs/1708.01155, 2017. URL http://arxiv.org/abs/1708.01155.

[7] Karim Armanious, Chenming Yang, Marc Fischer, Thomas Ku stner, Konstantin Nikolaou, Ser- gios Gatidis, and Bin Yang. Medgan: Medical image translation using gans. arXiv preprint arXiv:1806.06397, 2018.

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomed- ical image segmentation. CoRR, abs/1505.04597, 2015. URL http://arxiv.org/abs/ 1505.04597.

[9] T. Lindvall. Lectures on the Coupling Method. Dover Books on Mathematics Series. Dover Publi- cations, Incorporated, 2002. ISBN 9780486421452. URL https://books.google.com/ books?id=GUwyU1ypd1wC.

[10] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. CoRR, abs/1512.09300, 2015. URL http://arxiv. org/abs/1512.09300.

[11] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. CoRR, abs/1606.07536, 2016. URL http://arxiv.org/abs/1606.07536.

[12] Shane Barratt and Rishi Sharma. A note on the inception score. arXiv preprint arXiv:1801.01973, 2018.

[13] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. arXiv preprint arXiv:1711.10337, 2017.

[14] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. arXiv preprint arXiv:1805.08318, 2018.

[15] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4):600–612, April 2004. ISSN 1057-7149. doi: 10.1109/TIP.2003.819861.

[16] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. CoRR, abs/1802.05957, 2018. URL http://arxiv.org/ abs/1802.05957.