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

Aditi Maheshwari (aditi17@stanford.edu), Ketan Pant (ketan94@stanford.edu) | CS 230, Winter 2019 | Stanford University

Introduction

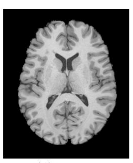
- The objective is to improve the fidelity in which we can translate medical images from one modality to another using Generative Adversarial Networks(GANs).
- Evaluate Unsupervised Image to Image Translation network(UNIT) as baseline and incorporate techniques like Self-Attention, Spectral Normalization to improve results.

Motivation

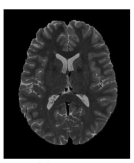
- Utilizable image data is scarce in medical community because of the high costs of acquisition, rarity of patient conditions and patient confidentiality.
- Translations of medical images between different modalities would reduce the number of times medical devices need to be used and increase the data available to doctors thereby helping the machine learning community utilize disjointed data sets together.

Dataset

- Using Dataset provided by the Human Connectome Project.
- Consists of paired T1 and T2-weighted 3D volumes of brain MR images of 1113 patients, spilt into 900 training and 213 test images in each domain.
- Resolution : 304x256 pixel colour images.



(a) T1 GT



(b) T2 GT

Figure: Example from dataset.

UNIT

- UNsupervised Image to image Translation, or UNIT by Liuet al. [1] is being used to transform MRI images across its T1 and T2 domains.
- In T1-weighted MRI, tissues with high fat content appear bright and compartments filled with water appear dark while it is the opposite in T2-weighted images.
- The model learns the joint distribution between the 2 domains by the 'shared latent space assumption'.

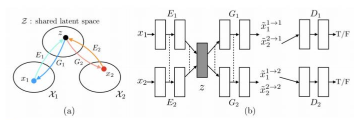


Figure: (a) The shared latent space assumption. (b) The UNIT framework.

Architecture

- Here, N=Neurons, K=Kernel size, S=Stride size.

Layer	Encoders	Shared?
1	CONV-(N64,K7,S1), LeakyReLU	No
2	CONV-(N128,K3,S2), LeakyReLU	No
3	CONV-(N256,K3,S2), LeakyReLU	No
4	RESBLK-(N512,K3,S1)	No
5	RESBLK-(N512,K3,S1)	No
6	RESBLK-(N512,K3,S1)	No
μ	RESBLK-(N512,K3,S1)	Yes
Layer	**Generators**	**Shared?**
1	RESBLK-(N512,K3,S1)	Yes
2	RESBLK-(N512,K3,S1)	No
3	RESBLK-(N512,K3,S1)	No
4	RESBLK-(N512,K3,S1)	No
5	DCONV-(N256,K3,S2), LeakyReLU	No
6	DCONV-(N128,K3,S2), LeakyReLU	No
7	DCONV-(N3,K1,S1), TanH	No
Layer	**Discriminators**	**Shared?**
1	CONV-(N64,K3,S2), LeakyReLU	No
2	CONV-(N128,K3,S2), LeakyReLU	No
3	CONV-(N256,K3,S2), LeakyReLU	No
4	CONV-(N512,K3,S2), LeakyReLU	No
5	CONV-(N1024,K3,S2), LeakyReLU	No
6	CONV-(N1,K2,S1), Sigmoid	No

Figure: Network Architecture

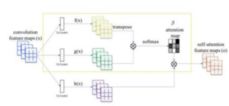
- Key points:
  - Binary Cross Entropy Loss for G and D.
  - MAE loss for the reconstruction stream.
  - VAE loss (MSE) for encoder and cycle consistency constraint.
  - ADAM with learning rate = 0.0001, $\beta_1 = 0.5$, $\beta_2 = 0.999$

Modified UNIT

- We incorporate techniques such as Self-Attention, Spectral Normalization (SN) and Charbonnier Penalty to improve fidelity of translations.
- Spectral Normalization was shown to improve stability in GAN training.
- Since L2 regularization penalizes outliers heavily, it prevents learning minute deltails, Hence we use Charbonnier penalty: $\rho(x) = \sqrt{x^2 + \epsilon^2}$

Self Attention [2]

- $f(x) = W_f x$, $g(x) = W_g x$, $s_{i,j} = f(x_i)^T g(x_j)$
- $h(x_i) = W_h x_i$, $\beta_{j,i} = {exp(s_{i,j})}/{\sum_{i=1}^{N} exp(s_{i,j})}$
- $o_i = \sum_{i=1}^{N} \beta_{j,i} h(x_i)$
- $\beta_{j,i}$ indicates extent to which model attends to i^{th} location when synthesizing j^{th} region.



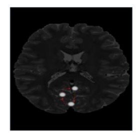


Figure: (a) Self-attention network layer, (b) Example

Qualitative Results

- With self attention in encoder, SN in encoder, gennerator and discriminator, and Charbonnier penalty, below is a qualitative comparsion of the two models.

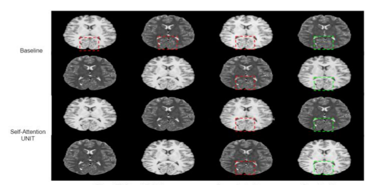


Figure: The translated and reconstructed images for the baseline model (top) and Self- Attention UNIT (bottom).

Quantitative Results

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

Figure: 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

Figure: SSIM scores for the baseline model and our best model over epochs 1, 25, 50, 75, 100.

Conclusion and Future Work

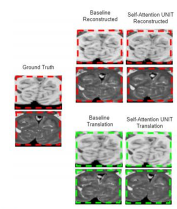


Figure: Concluding comparison.

- Even though quantitative results were similar to baseline, qualitatively, our model was better at reproducing intricate details $\implies$ search for other evaluation metrics?

References

[1] 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.

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