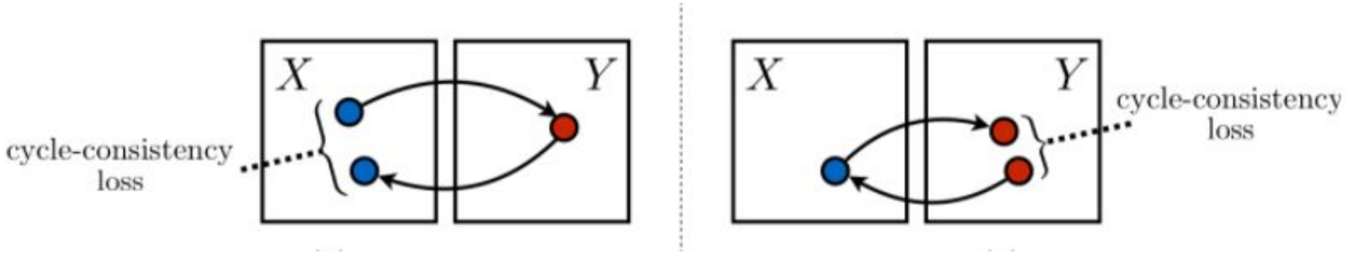
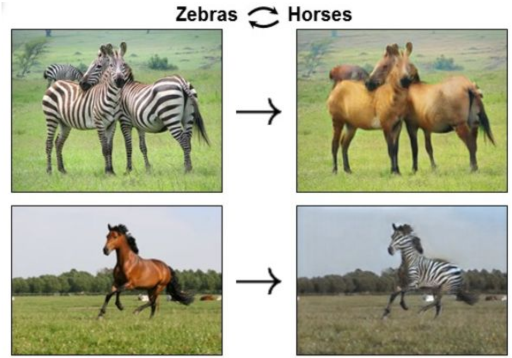
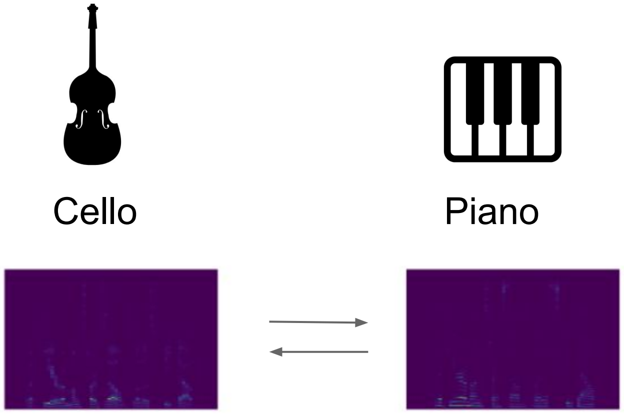
Adversarial Variational Bayes Generative Translation Network

Mike , Tom, & Sam

Motivation

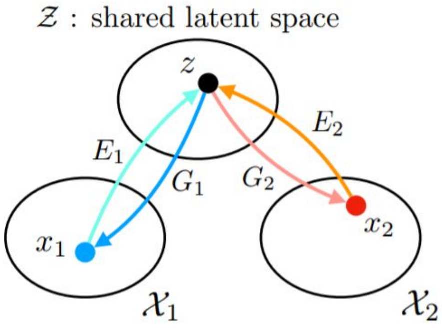


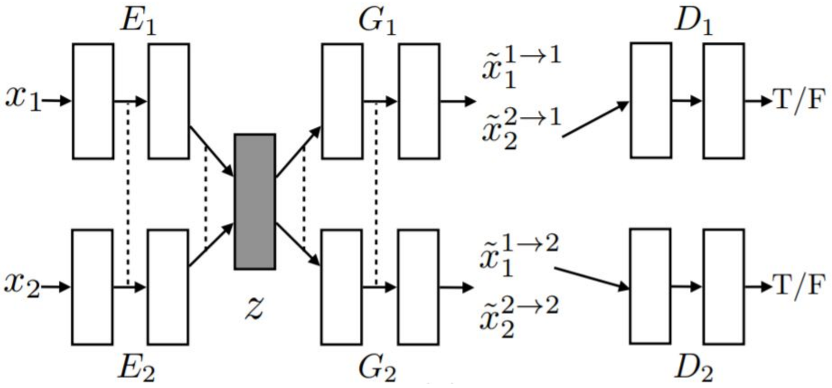




Spec2Spec

Architecture, UNIT GAN



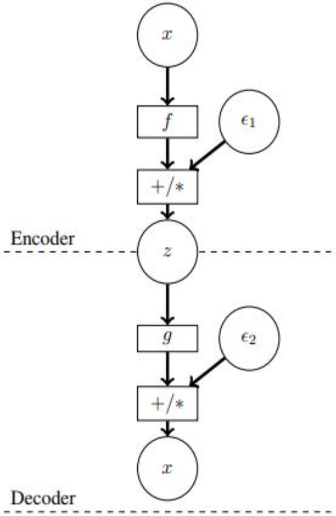


$$\mathcal{L}_{\text{GAN}_1}(E_1, G_1, D_1) = \lambda_0 \mathbb{E}_{x_1 \sim P_{\mathcal{X}_1}}[\log D_1(x_1)] + \lambda_0 \mathbb{E}_{z_2 \sim q_2(z_2|x_2)}[\log(1 - D_1(G_1(z_2)))]$$
$$\mathcal{L}_{\text{GAN}_2}(E_2, G_2, D_2) = \lambda_0 \mathbb{E}_{x_2 \sim P_{\mathcal{X}_2}}[\log D_2(x_2)] + \lambda_0 \mathbb{E}_{z_1 \sim q_1(z_1|x_1)}[\log(1 - D_2(G_2(z_1)))]$$

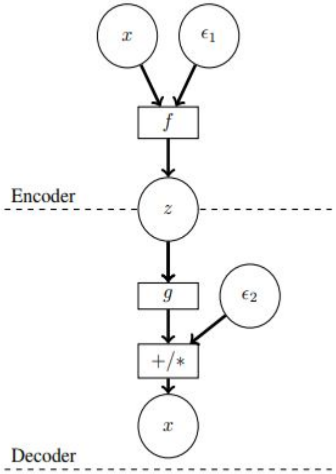
$$\mathcal{L}_{\text{VAE}_1}(E_1, G_1) = \lambda_1 \text{KL}(q_1(z_1|x_1)||p_\eta(z)) - \lambda_2 \mathbb{E}_{z_1 \sim q_1(z_1|x_1)}[\log p_{G_1}(x_1|z_1)]$$
$$\mathcal{L}_{\text{VAE}_2}(E_2, G_2) = \lambda_1 \text{KL}(q_2(z_2|x_2)||p_\eta(z)) - \lambda_2 \mathbb{E}_{z_2 \sim q_2(z_2|x_2)}[\log p_{G_2}(x_2|z_2)]$$

$$\mathcal{L}_{\text{CC}_1}(E_1, G_1, E_2, G_2) = \lambda_3 \text{KL}(q_1(z_1|x_1)||p_\eta(z)) + \lambda_3 \text{KL}(q_2(z_2|x_1^{1\to 2}))||p_\eta(z)) - \lambda_4 \mathbb{E}_{z_2 \sim q_2(z_2|x_1^{1\to 2})}[\log p_{G_1}(x_1|z_2)]$$
$$\mathcal{L}_{\text{CC}_2}(E_2, G_2, E_1, G_1) = \lambda_3 \text{KL}(q_2(z_2|x_2)||p_\eta(z)) + \lambda_3 \text{KL}(q_1(z_1|x_2^{2\to 1}))||p_\eta(z)) - \lambda_4 \mathbb{E}_{z_1 \sim q_1(z_1|x_2^{2\to 1})}[\log p_{G_2}(x_2|z_1)]$$

Architecture, Adversarial Variational Bayes



Standard VAE



AVB - VAE

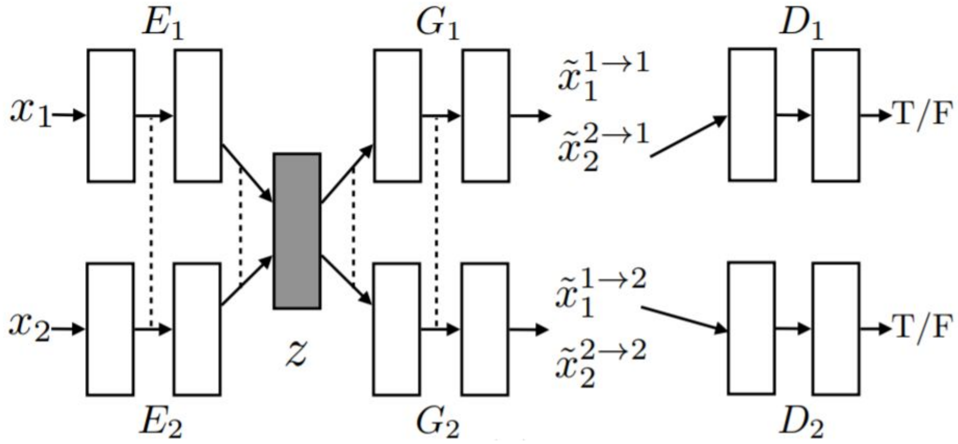
$$\max_T \mathrm{E}_{p_{\mathcal{D}}(x)} \mathrm{E}_{q_\phi(z|x)} \log \sigma(T(x,z)) + \mathrm{E}_{p_{\mathcal{D}}(x)} \mathrm{E}_{p(z)} \log\left(1 - \sigma(T(x,z))\right).$$

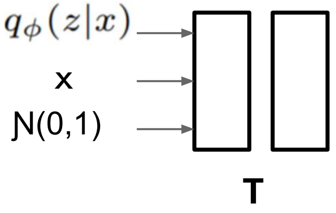
The discriminator is considered to be a nonparametric limit and a **universal function approximator to approximate the Kullback-Leibler regularization term.**

$$\max_\theta \max_\phi \mathrm{E}_{p_{\mathcal{D}}(x)} \Big[-\mathrm{KL}(q_\phi(z \mid x), p(z)) + \mathrm{E}_{q_\phi(z|x)} \log p_\theta(x \mid z) \Big].$$

$$\max_{\theta,\phi} \mathrm{E}_{p_{\mathcal{D}}(x)} \mathrm{E}_\epsilon \big(-T^*(x, z_\phi(x,\epsilon)) + \log p_\theta(x \mid z_\phi(x,\epsilon)) \big)$$

Our Architecture





Implementation:

- ResNet style generator with shared middle layers
- Discriminator with shared final layer between both domains
- 2 KL Discriminators for each posterior for our model
- Standard KL between Normal posterior for UNIT

Results: CelebA Hair Color (Original, UNIT, Ours)

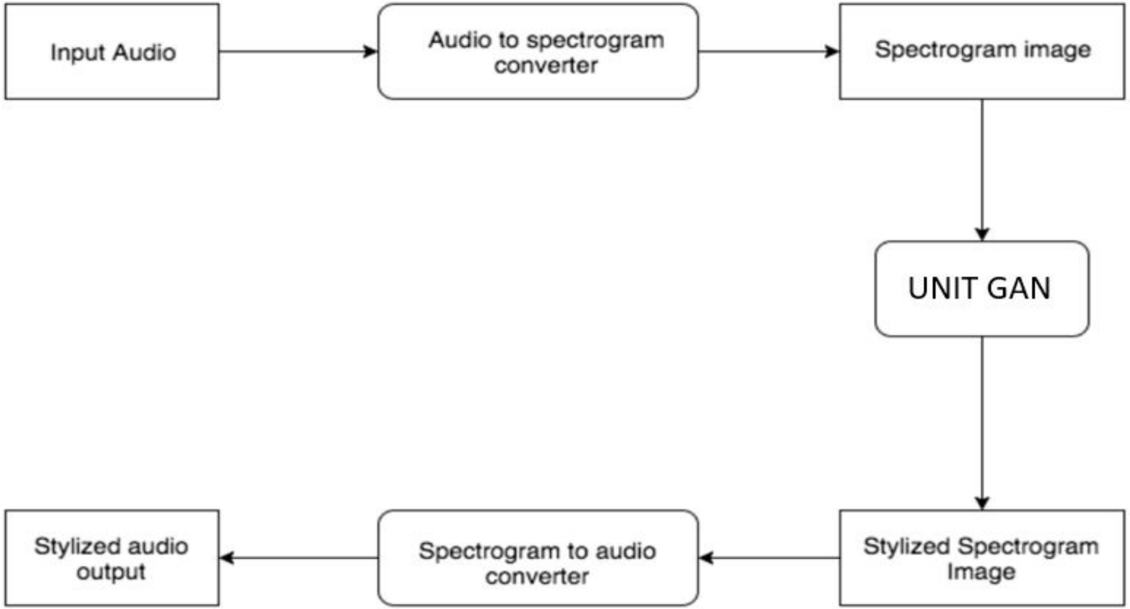
Epoch

5000

20000

50000

Audio Conversion



Results: Audio Conversion

- Results:
  - Pretty good conversion back to original domain
  - Not great conversion to alternate domain
- Issues:
  - Lack of data
  - Each sound clip not very representative of the instrument
  - GPU memory not large enough for larger spectrograms
  - Didn’t experiment with frequency only training/reconstruction

Conclusion

- Much faster convergence to good results
- Better results in the end
- Simple add-on to existing VAE style architectures
- Generative model as opposed to pix2pix