# Optimization of Unpaired Data Translation

## Group Members:

Jalansh Munshi - 201601042
Raksha Rank - 201601055
Dharmil Patel - 201601103
Sakshee Patel - 201601132
Akshat Kaneria - 201601242

# Introduction

- Vision, graphics, and speech problems often involve mapping between input and target data using a training set of aligned data pairs.
- But it is not necessary that paired data is available for all the tasks.
- We present an optimization problem which performs conversion between two different data distributions, in the absence of any parallel data.

# Proposed Approach

- The primary objective function of the problem consists of Adversarial Loss, which was introduced by Goodfellow et. al. [1]
- Adversarial loss tries to learn the mapping such that the generated data cannot be distinguished from the original data distribution.
- Furthermore, we also require cycle consistency loss, in order to prevent the learned mappings from contradicting each other.
- Lastly, we also include identity loss for generating near identity mappings.
- We optimize the final objective function using Stochastic Gradient Descent.

[1] Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." In *Advances in neural information processing systems*, pp. 2672-2680. 2014.
[2] Zhu, Jun-Yan, 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*, pp. 2223-2232. 2017.

# Optimization of Adversarial Loss

- Our inherent architecture consists of two generators and discriminators [2]. The generator tries to fool the discriminator by generating fake data, and discriminator tries to classify whether the generated data is fake or real.
- Let us consider the adversarial loss for only one generator (G) and discriminator (D) for now.
- The data distribution is denoted as $x \sim p_{data}(x)$ and the generated data is denoted by z.
- We get the following cost function for the problem statement – (equation 1)

$$V(D,G) = E_{x \sim p_{data}(x)}[logD(x)] + E_{x \sim p_z(z)}[\log(1 - D\big(G(z)\big)$$

- D maximizes the probability of assigning the correct label to the training examples and samples from G. Simultaneously, G tries to minimize $\log(1 - D\big(G(z)\big)$. They are basically trying to play a minimax game with value function $V(D,G)$.

$$\min_G \max_D V(D,G)$$

- For a given G, it will first find the optimal D, and then for a given optimal D, it will find the optimal value of G.

[2] Zhu, Jun-Yan, 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*, pp. 2223-2232. 2017.

# Optimization of Adversarial Loss

- Converting everything to a single variable $x$.

$$x = G(z) \rightarrow z = G^{-1}(x) \rightarrow dz = (G^{-1})'(x)dx$$

- Also, $p_g(x) = p_z\big(G^{-1}(x)\big)(G^{-1})'(x)dx$, where $P_g$ is the generated distribution.
- So, equation (1) becomes –

$$V(D,G) = \int_x (p_{data}(x)\log(D(x)) + \log(1 - D(x)))dx$$

- For a given G, optimal D comes out to be $D(x) = \dfrac{p_{data}}{p_{data} + p_g}$
- Now, we solve the optimization problem $C(G) = \min_G V(D,G)$

# Optimization of Adversarial Loss

- The final expression comes out to be –

$$C(G) = KL[p_d \parallel \frac{p_d + p_g}{2}] + KL[p_g \parallel \frac{p_d + p_g}{2}] - \log 4$$

- where KL[.] represents the KL divergence.
- KL divergence is always non-negative. Since we are minimizing, the minimum value of KL divergence will be 0. Therefore.

$$\min_G \max_D V(D, G) = -\log 4$$

- Also, since KL divergence is 0, we get $p_{data}(x) = p_g(x)$, which is exactly what we desired to achieve.

# Optimization of Cycle-consistency Loss

- Adversarial loss alone cannot guarantee that the learned mapping function can map an individual input to the desired output.
- Cycle-consistency loss minimizes the loss between source and target data distributions.
- For two data distributions, $x$ and $y$, the cycle-consistency losses are given by

$$\mathbb{E}_{x \sim p_{data}(x)}[||F(G(x)) - x||_1]$$

$$\mathbb{E}_{y \sim p_{data}(y)}[||G(F(x)) - y||_1]$$

- These two equations represent forward and backward cycle losses respectively. ||.|| represents the L1 norm.

# Optimization of Identity Loss

- This additional loss is introduced for emphasizing the mapping between the input and the output.
- Main objective is to regularize the generator so that it produces a near-identity mapping.
- The identity loss for two data distributions $x$ and $y$ is given by

$$\mathcal{L}_{identity}(G, F) = \mathbb{E}_{y \sim p_{data}(y)}[||G(y) - y||_1] + \mathbb{E}_{x \sim p_{data}(x)}[||F(x) - x||_1], \quad (17)$$

- The final objective function for the problem is as follows –

$$\mathcal{L}_{final} = \mathcal{L}_{GAN} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{id}\mathcal{L}_{id},$$

- Here, $\lambda_{cyc}$ and $\lambda_{id}$ are empirically chosen trade-off parameters.
- At the end of every iteration, the loss is calculated, and the model tunes its parameters on the basis of this loss. We use Stochastic Gradient Descent method for tuning or optimizing the parameters.

# Gradient Descent

- It is used in Machine Learning to minimize or optimize the cost function, so that our model can make accurate predictions in accordance with our data. [3]
- The weights of our network are initialized randomly. Thereafter, we calculate the gradient, and modify the weights of the network.
- The degree at which the modification should occur, is controlled by the learning rate $\alpha$.
- We update the weights at the end of every iteration, and hence it is known as Stochastic Gradient Descent method.

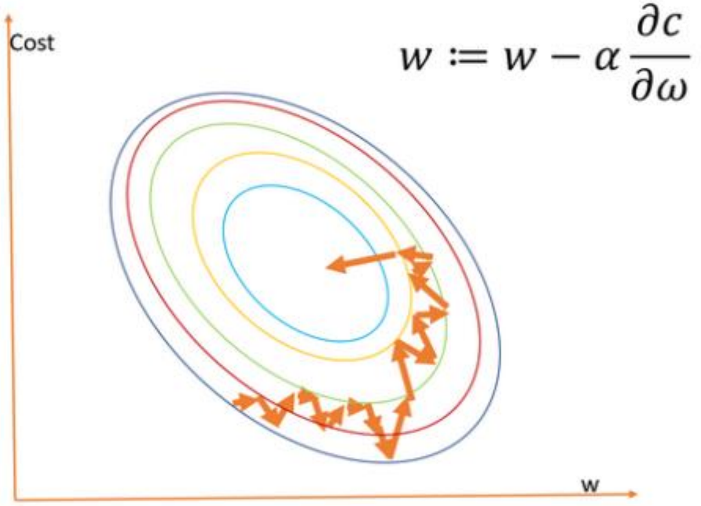


**Fig. 1.** Depiction of a sample run of the Gradient Descent algorithm.

[3] Renu Khandelwal, “Machine learning gradient descent,” in Medium - https://medium.com/datadriveninvestor/gradient-descent-5a13f385d403, Last Accessed: 14 November 2019 . Medium, 2018

# Results and Analysis

- We performed the experiment of Non-Audible Murmur (NAM) to Whisper conversion, to improve the intelligibility of NAM speech.
- We analyzed the effectiveness of our model by comparing the scatter plots of the original whispered speech and the whispered speech converted by our model.
- Our model's effectiveness can be clearly observed from the figure.

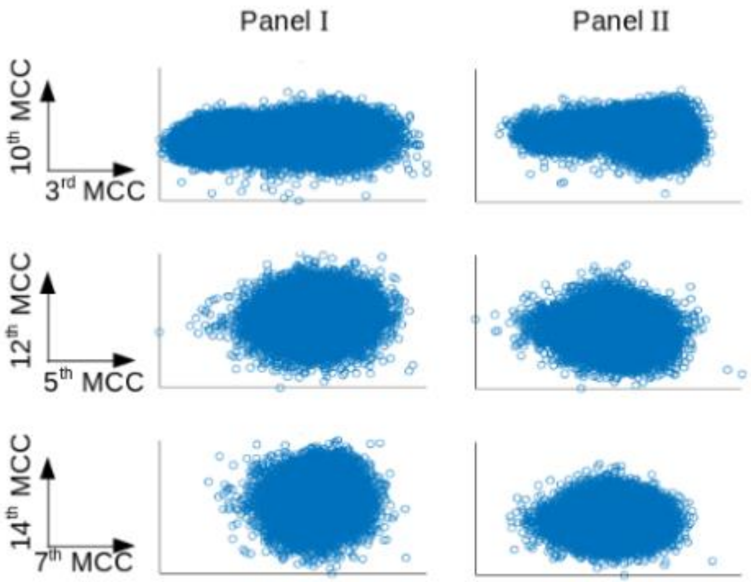


**Fig. 2.** Scatter plot analysis which shows the effectiveness of our approach.

# References

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

[2] Zhu, Jun-Yan, 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*, pp. 2223-2232. 2017.

[3] Renu Khandelwal, "Machine learning gradient descent," in Medium - https://medium.com/datadriveninvestor/gradient-descent-5a13f385d403, Last Accessed: 14 November 2019 . Medium, 2018

Thank you