Illegible Text to Readable Text: An Image-to-Image Transformation using Conditional Sliced Wasserstein Adversarial Networks

Mostafa Karimi *
Texas A&M University
mostafa_karimi@tamu.edu

Gopalkrishna Veni
Ancestry.com
gveni@ancestry.com

Yen-Yun Yu
Ancestry.com
yyu@ancestry.com

A. Wasserstein distance

Wasserstein distance is a powerful metric in the field of optimal transport and has recently drawn a lot of attention [4]. It measures the distance between two distributions. p-Wasserstein distance (WD) between two random variables X, Y is given as:

$$W_p = \inf_{\gamma \in \Gamma(\mathbb{P}_X, \mathbb{P}_Y)} \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\gamma} [d^p(\mathbf{x},\mathbf{y})]^{\frac{1}{p}}, \tag{1}$$

where $\Gamma(\mathbb{P}_X, \mathbb{P}_Y)$ denotes a set of all joint distributions $\gamma(X, Y)$ whose marginal distributions are $\mathbb{P}_X$, $\mathbb{P}_Y$. Suppose x and y are realizations or samples from random variables X and Y respectively. Let $p > 0$, then $d(\mathbf{x}, \mathbf{y})$ defines a metric for $\mathbf{x}$ and $\mathbf{y}$. For $p = 1$, 1-WD $d(\mathbf{x}, \mathbf{y})$ is named as Earth-Mover distance (EMD). Intuitively, $\gamma(X, Y)$ shows how much "mass" is going to be transported from any realization of $\mathbf{X}$ to any realization of $\mathbf{Y}$ in order to transport distribution $\mathbb{P}_X$ to the distribution $\mathbb{P}_Y$. Because the primal form of the 1-WD is generally intractable and usually the dual form is used in practice[1], a dual form of EMD is formulated through the Kantorovich-Rubinstein (KR) duality [1] and is given as:

$$W_1 = \sup_{||g||_L \leq 1} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}_X} [g(\mathbf{x})] - \mathop{\mathbb{E}}_{\mathbf{y}\sim\mathbb{P}_Y} [g(\mathbf{y})], \tag{2}$$

where the supremum is over all 1-Lipschitz functions $g(\cdot)$.

B. Wasserstein Generative adversarial networks

One challenge in applying WD on GAN is that WD is a much weaker distance compared to the JS distance, i.e., it induces a weaker topology. This fact makes a sequence of probability distributions converge in the distribution space [1], which results in bringing the model distribution closer to the real distribution. In other words, both the low dimensional support challenge in high dimensions and the gradient vanishing problem could be solved under this assumption. Due to these reasons, the Wasserstein GAN (WGAN) model has been developed based on the dual form of the EMD [1]. WGAN with generator G and discriminator D is formulated as the first term of Eq. (3). The main challenge in WGAN is to satisfy the Lipschitz continuity constraint. The original WGAN considered a weighted clipping approach that limits the capacity of the model and its performance [2]. To alleviate this problem, WGAN with gradient penalty (WGAN-GP) [2] has been developed that penalizes the norm of the discriminator's gradient with respect to a few input samples. The gradient penalty $GP = \mathop{\mathbb{E}}_{\hat{\mathbf{x}}\sim\mathbb{P}_{\hat{X}}} [(||\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}})||_2 - 1)^2]$ is added to the original WGAN loss function in Eq. (3). Therefore, WGAN-GP is formulated as:

$$\min_G \max_{||D||_L \leq 1} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}_r} [D(\mathbf{x})] - \mathop{\mathbb{E}}_{\mathbf{y}\sim\mathbb{P}_g} [D(\mathbf{y})] + \lambda GP \tag{3}$$

$\hat{\mathbf{x}}$ represents random samples following the distribution $\mathbb{P}_{\hat{X}}$, which is formed by uniformly sampling along the straight lines between pair of points sampled from $\mathbb{P}_r$ and $\mathbb{P}_g$. λ is the hyper-parameter to balance between original WGAN loss function and the gradient penalty regularization. Recently WGAN has been further improved by adding consistency term GAN (CTGAN) [5].

C. Sliced Wasserstein distance (SWD)

WD is generally intractable for multi-dimensional probability distribution [3]. However, there is a closed-form solution (i.e., WD is tractable) if the distribution is in the low-dimensional space (in this paper, we use an one dimensional space). Let F_X and F_Y be the cumulative distribution function (CDF) for probability distributions $\mathbb{P}_X$ and $\mathbb{P}_Y$ respectively. The WD between these two distributions is uniquely defined as $F_Y^{-1}(F_X(x))$.The primal p-WD between them can be re-defined as:

$$W_p = \left(\int_0^1 d^p(F_X^{-1}(z), F_Y^{-1}(z)) dz \right)^{\frac{1}{p}} \tag{4}$$

The change of variable $z := F_x(x)$ is used to derive the equation. For empirical distributions, Eq. (4) is calculated

*Work done during internship with the Data Science team at Ancestry.com

by sorting two distributions and then calculating the average distance $d^p(\cdot,\cdot)$ between two sorted samples which requires $O(M)$ at best and $O(M \log M)$ at worst, where M is number of samples for each distribution[3].

Sliced Wasserstein distance (SWD) utilizes this property by factorizing high-dimensional probabilities to multiple marginal distributions [6] with standard Radon transform, denoted by $\mathcal{R}$. Given any distribution $P(\cdot)$, the Radon transform of $P(\cdot)$ is defined as:

$$\mathcal{R}P(t,\theta) = \int_{\mathbb{R}^d} \mathbb{P}(\mathbf{x})\delta(t - \langle\theta,\mathbf{x}\rangle)d\mathbf{x}, \tag{5}$$

where $\delta(\cdot)$ is the one-dimensional Dirac delta function and $\langle\cdot,\cdot\rangle$ is the Euclidean inner-product. The hyper-parameters in Radon transform include a level set parameter $t \in \mathbb{R}$ and a normal vector $\theta \in \mathbb{S}^{d-1}$ (θ is a unit vector, and $\mathbb{S}^{d-1}$ is the unit hyper-sphere in d-dimensional space). Radon transform $\mathcal{R}$ maps a function to the infinite set of its integrals over hyperplanes $\langle\theta,\mathbf{x}\rangle$ of $\mathbb{R}^d$. For a fixed θ, the integrals over all hyperplanes define a continuous function $\mathcal{R}P(.,\theta) : \mathcal{R} \to \mathcal{R}$ which is a slice or projection of P. The p-WD in Eq. (4) can be rewritten as the sliced p-WD for a pair of distributions $\mathbb{P}_X$ and $\mathbb{P}_Y$:

$$SW_p = \left(\int_{\theta\in\mathbb{S}^{d-1}} W_p(\mathcal{R}P_X(.,\theta),\mathcal{R}P_Y(.,\theta))d\theta\right)^{\frac{1}{p}} \tag{6}$$

The dual of Eq. (6) can be derived based on KR duality:

$$SW_p = \left(\int_{\theta\in\mathbb{S}^{d-1}} \sup_{||g||_L\le 1} \mathop{\mathbb{E}}_{\mathbf{x}_\theta}[g(\mathbf{x}_\theta)] - \mathop{\mathbb{E}}_{\mathbf{y}_\theta}[g(\mathbf{y}_\theta)]d\theta\right)^{\frac{1}{p}} \tag{7}$$

where x_θ and y_θ are sampled from $\mathcal{R}P_X(.,\theta)$ and $\mathcal{R}P_Y(.,\theta)$ respectively. SWD is not only a valid distance which satisfies positive- definiteness, symmetry and the triangle inequality [6], but also equivalent to WD based on Lemma 0.1.

Lemma 0.1 *Following inequality holds for SWD and WD where α_1 and α_2 are constants and n is the dimension of sample vectors from X and Y [6]:*

$$SW_p(\mathbb{P}_X,\mathbb{P}_Y)^p \le \alpha_1 W_p(\mathbb{P}_X,\mathbb{P}_Y)^p \le \alpha_2 SW_p(\mathbb{P}_X,\mathbb{P}_Y)^{\frac{1}{n+1}}$$

D. Sliced Wasserstein Generative adversarial networks (SWGAN)

Recently Sliced Wasserstein Generative adversarial network (SWGAN) [6] has been proposed by utilizing the dual form of WGAN and approximating SWD in generative models. The discriminator is composed of an encoding network E and M dual SWD blocks $\{S_m\}_{m=1}^M$, that is, $D := \{S_m \circ E\}_{m=1}^M = [S_1 \circ E, \cdots, S_M \circ E]^T$. Where The operation $S_i \circ E = S_i(E(\cdot))$. The encoder $E : \mathcal{R}^{b\times n} \to \mathcal{R}^{b\times r}$ maps a batch of data $X \in \mathcal{R}^{b\times n}$ to the latent space of $X^{\text{embd}} \in \mathcal{R}^{b\times r}$ where b is the batch size, n is the data dimension and r is the latent dimension. The first part of each dual SWD block will operate on the orthogonalization operation $X^{\text{orth}} = X^{\text{embd}}\Theta$ with $\Theta \in \mathcal{R}^{r\times r}$ to make sure that the encoded matrix is orthogonal. The second part of each dual SWD block will perform an element-wise non-linear neural network function $T_i(\mathbf{x}_i^{\text{orth}}) = u_i\text{LeakyReLU}(w_i\mathbf{x}_i^{\text{orth}} + b_i)$ to approximate one-dimensional optimal g function [6] in Eq. (7) for all $i = 1, ..., r$ where $u_i, w_i,$ and b_i are scalar parameters. Eventually, the model can be approximated by integrating over $\mathbb{S}^{n-1}$ and summing the output mean value of the dual SWD blocks.

The Lipschitz constraint can be easily applied over one-dimensional functions followed by the gradient penalty on each dimension of T_i's. The projection matrices should remain orthogonal throughout the training process. Accordingly, a manifold-valued update rule has been developed based on the Stiefel manifolds [6]. SWGAN's final objective function is as follows:

$$\min_G \max_D \int_{\theta\in\mathbb{S}^{n-1}} \Big(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}_r}[D(\mathbf{x})] - \mathop{\mathbb{E}}_{\mathbf{y}\sim\mathbb{P}_g}[D(\mathbf{y})]\, d\theta\Big) + \lambda_1 \mathop{\mathbb{E}}_{\hat{\mathbf{x}}\sim\mathbb{P}_{\hat{X}}}[||\nabla_{\hat{\mathbf{x}}}D(\hat{\mathbf{x}})||_2^2] + \lambda_2 \mathop{\mathbb{E}}_{\hat{\mathbf{y}}\sim\mathbb{P}_{\hat{y}}}[(||\nabla_{\hat{\mathbf{y}}}T(\hat{\mathbf{y}}) - \mathbf{1}||_2^2] \tag{8}$$

where θ represents trainable parameters embedded in D, $\mathbf{1}$ is a vector with all entries equal to 1, λ_1 and λ_2 are the hyper-parameters for balancing the gradient penalty terms and dual SWD.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223, 2017. 1

[2] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017. 1

[3] Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo K Rohde. Generalized sliced wasserstein distances. *arXiv preprint arXiv:1902.00434*, 2019. 1, 2

[4] Soheil Kolouri, Phillip E Pope, Charles E Martin, and Gustavo K Rohde. Sliced-wasserstein autoencoder: an embarrassingly simple generative model. *arXiv preprint arXiv:1804.01947*, 2018. 1

[5] Xiang Wei, Boqing Gong, Zixia Liu, Wei Lu, and Liqiang Wang. Improving the improved training of wasserstein gans: A consistency term and its dual effect. *arXiv preprint arXiv:1803.01541*, 2018. 1

[6] Jiqing Wu, Zhiwu Huang, Dinesh Acharya, Wen Li, Janine Thoma, Danda Pani Paudel, and Luc Van Gool. Sliced wasser-

stein generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3713–3722, 2019. 2