MMD GAN: Towards Deeper Understanding of Moment Matching Network

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, Barnabás Póczos

---

Coby Penso

What we'll see

© Coby Penso

Introduction

Generative moment matching network (GMMN) is a deep generative model that differs from GAN by replacing the discriminator with a two-sample test based on **kernel maximum discrepency (MMD)**
In this paper we'll see an improvement of model expressiveness of GMMN and its computational efficiency by **adversarial kernel learning** techniques.
Combining key ideas from GMMN and GAN, hence called **MMD-GAN**.

© Coby Penso

Preliminary

GAN

Given $x_{i\,i=1}^{n}$, $x_i \in X$ and $x_i \sim P_X$.
Two ways to sample from P_X Estimate the density of P_X
Use Generative Adversarial Network to train a generator g_θ, to transform $z \sim P_Z$ into $g_\theta(z) \sim P_\theta$ such that $P_{theta} \approx P_X$. To measure similarity a discriminator trained to distinguish x_i and $g_\theta(z_j)$

Two-Sample Test

Distinguishing two distributions by finite samples is known as Two-Sample Test in statistics. One way is done via a kernel maximum mean discrepancy (MMD)

© Coby Penso

MMD

Given two distibutions P and Q, and a kernel k, the square of MMD distance is defined as

$$M_k(P,Q) = ||\mu_P - \mu_Q||^2_{\mathcal{H}} = \mathbb{E}_P[k(x,x')] - 2\mathbb{E}_{P,Q}[k(x,y)] + \mathbb{E}_Q[k(y,y')]$$

Theorem: Given a kernel k, if k is a characteristic kernel, then $M_k(P,Q) = 0$ iff $P = Q$

GMMN

One example of characteristic kernel is Gaussian kernel $k(x,x') = \exp ||x - x'||^2$. Based of Theorem 1, training g_θ by

$$\min_\theta M_k(P_X, P_\theta)$$

with a fixed Gaussian kernel k rather than training an additional discriminator f as GAN.

© Coby Penso

Bar-Ilan University

MMD is a distance (difference) between feature means.

Kernel

Let X be non-empty set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if there exist a Hilbert space $\mathbb{H}$ and a feature map $\phi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, x' \in \mathcal{X}$,

$$k(x, x') := \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$$

$$MMD^2(P, Q) = ||\mu_P - \mu_Q||^2_{\mathcal{H}}$$

$$MMD^2(P, Q) = \langle \mu_P - \mu_Q, \mu_P - \mu_Q \rangle = \langle \mu_P, \mu_P \rangle - 2 \langle \mu_P, \mu_Q \rangle + \langle \mu_Q, \mu_Q \rangle$$

$$MMD^2(P, Q) = \mathbb{E}_P[k(X, X)] - 2\mathbb{E}_{P,Q}[k(X, Y)] + \mathbb{E}_Q[k(Y, Y)]$$

© Coby Penso

MMD with Kernal Learning

In practice, Given $X = x_1, \ldots, x_n \sim P$ and $Y = y_1, \ldots, y_n \sim Q$ then

$$M_k'(X,Y) = \frac{1}{\binom{n}{2}} \sum_{i=i'} k(x_i, x_i') - \frac{2}{\binom{n}{2}} \sum_{i=i'} k(x_i, y_i) + \frac{1}{\binom{n}{2}} \sum_{i=i'} k(y_i, y_i')$$

<u>Note:</u> due to sampling variance $M'(X,Y)$ may not be zero even when $P = Q$.

Conduct hypothesis test with null hypothesis $H_o : P = Q$.
For a given allowable probability of false rejection α, reject H_o.

- $P \neq Q$ if $M'(X,Y) > c_\alpha$
- Otherwise Q pass the test and indistinguishable from P under the test.

MMD with Kernal Learning

If kernel k cannot result high MMD when $P \neq Q$, then M' has more chance to be smaller than c_α, unlikely to reject null hypothesis and those Q is not distinguishable from P.
Therefore, instead of using pre-specified kernel k as GMMN, we consider training g_θ via

$$\min_{\theta} \max_{k \in K} M_k(P_X, P_\theta)$$

which takes different possible characteristic kernels $k \in K$ into account.
Could views as replacing the fixed kernel k with the adversarially learned kernel $\underset{k \in K}{argmax}\, M_k(P_X, P_\theta)$ to have a stronger signal where $P \neq P_\theta$

However, it is difficult to optimize over all characteristic kernels.

Theorem

Given f an injective function and k is characteristic, then the resulted kernel $\hat{k} = k \circ f$, where $\hat{k}(x, x') = k(f(x), f(x'))$ is still characteristic.

Given family of injective functions parametrized by $\{\phi, f_\phi\}$, then the objective can changed to

$$\min_{\theta} \max_{\phi} M_{k \circ f_\phi}(P_X, P_\theta)$$

In this paper - combining Gaussian kernels with injective functions

$$\hat{k}(x, x') = exp(-||f_\phi(x) - f_\phi(x')||^2)$$

Example:
$\{f_\phi | f_\phi(x) = \phi x, \phi > 0\}$
Equivalent to the kernel bandwidht tuning (length scale tuning).

© Coby Penso

Properties of MMD with Kernel Learning

Bar-Ilan University

Assumption

$g : Z \times \mathbb{R}^m \to \mathcal{X}$ is locally Lipschits. Given f_ϕ and a probability distribution $\mathbb{P}_z$ over $\mathcal{Z}$, if there are local Lipschits constants $L(\theta, z)$ for $f_\phi \circ g$, which is independent of ϕ, such that $\mathbb{E}_{z \sim \mathbb{P}_z}[L(\theta, z)] < +\infty$

Theorem 3

The generator function g_θ parametrized by θ is under the above assumption. Let $\mathbb{P}_\mathcal{X}$ be a fixed distribution over $\mathcal{X}$ and Z be a random variable over the space $\mathcal{Z}$. We denote $\mathbb{P}_\theta$ the distribution of $g_\theta(Z)$, then $\max_\phi M_{f_\phi}(\mathbb{P}_\mathcal{X}, \mathbb{P}_\theta)$ is continuous everywhere and differentiable almost everywhere in θ

If g_θ parametrized by feed-forward neural network, it satisfies the above assumption and can be trained via gradient decesnt as well as propagation, since the objective is continuous and differentiable by Theorem 3.

Theorem 4

(weak* topology) Let $\{\mathbb{P}_n\}$ be a sequence of distributions. Considering $n \to \infty$, under mild assumption, $\max_\phi M_{f_\phi}(\mathbb{P}_\mathcal{X}, \mathbb{P}_n) \to 0 \iff \mathbb{P}_n \underset{D}{\to} \mathbb{P}_\mathcal{X}$, where $\underset{D}{\to}$ means converging in distribution

Theorem 4 shows that $\max_\phi M_{f_\phi}(\mathbb{P}_\mathcal{X}, \mathbb{P}_n)$ is a sensible cost function to the distance between $\mathbb{P}_\mathcal{X}$ and $\mathbb{P}_n$.

MMD GAN

To approximate $\min_\theta \max_\phi M_{f_\phi}(P(X), P(g_\theta(Z)))$, we use NNs to parametrize - g_θ and f_ϕ. The following has to hold:

- g_ϕ is locally Lipschits, where commonly used Feed Forward NN satisfy this constraint.
- $\nabla_\theta(\max_\phi f_\phi \circ g_\theta)$ has to be bounded - done by clipping ϕ or gradient penalty.
- f_ϕ injective function - Non trivial (tackle that next slide)

© Coby Penso

Injective function requirement

Theorem: For an injective function there exists an function f^{-1} such that $f^{-1}(f(x)) = x \ \ \forall x \in X$ and $f^{-1}(f(g(z))) = g(z) \ \ \forall z \in Z$, which can be approx by an autoencoder.
Denote $\phi = (\phi_e, \phi_d)$ to be the parameter of discriminator networks, with f_{ϕ_e} and train the corresponding decoder $f_{\phi_d} \approx f^{-1}$ to regularize f.

$$\min_{\theta} \max_{\phi} M_{f_{\phi_e}}(P(X), P(g_\theta(Z))) - \lambda E_{y \in X \cup g(Z)} ||y - f_{\phi_d}(f_{\phi_e}(y))||^2$$

Note: Ignore the autoencoder objective when train .
Note: Empirical study suggests autoencoder objective is not necessary to lead to successful GAN training.

MMD GAN - Algorithm

Algorithm 1: MMD GAN, our proposed algorithm.

input : α the learning rate, c the clipping parameter, B the batch size, n_c the number of iterations of discriminator per generator update.

initialize generator parameter θ and discriminator parameter ϕ;

while *θ has not converged* **do**

  for $t = 1, \ldots, n_c$ **do**

    Sample a minibatches $\{x_i\}_{i=1}^{B} \sim \mathbb{P}(\mathcal{X})$ and $\{z_j\}_{j=1}^{B} \sim \mathbb{P}(\mathcal{Z})$

    $g_\phi \leftarrow \nabla_\phi M_{f_{\phi_e}}(\mathbb{P}(\mathcal{X}), \mathbb{P}(g_\theta(\mathcal{Z}))) - \lambda \mathbb{E}_{y \in \mathcal{X} \cup g(\mathcal{Z})} \|y - f_{\phi_d}(f_{\phi_e}(y))\|^2$

    $\phi \leftarrow \phi + \alpha \cdot \text{RMSProp}(\phi, g_\phi)$

    $\phi \leftarrow \text{clip}(\phi, -c, c)$

  Sample a minibatches $\{x_i\}_{i=1}^{B} \sim \mathbb{P}(\mathcal{X})$ and $\{z_j\}_{j=1}^{B} \sim \mathbb{P}(\mathcal{Z})$

  $g_\theta \leftarrow \nabla_\theta M_{f_{\phi_e}}(\mathbb{P}(\mathcal{X}), \mathbb{P}(g_\theta(\mathcal{Z})))$

  $\theta \leftarrow \theta - \alpha \cdot \text{RMSProp}(\theta, g_\theta)$

Encoding Perspective of MMD GAN

Another way besides from using a kernel selection to explain MMD GAN is viewing f_{ϕ_e} as a feature transformation function, and the kernel two-sample test is performed on this transformed feature space (i.e the code space of the autoencoder).

Feasible Set Reduction

Theorem: For any f_ϕ there exists f'_ϕ such that $M_{f_\phi}(P_r, P_\theta) = M_{f'_\phi}(P_r, P_\theta)$ and $E_x[f_\phi(x)] \geq E_z[f_{\phi'}(g_\theta(z))]$

With this theorem, we can reduce the feasible set of ϕ during the optimization by solving

$$\min_\theta \max_\phi M_{f_\phi}(P_r, P_\theta) \quad s.t \;\; E[f_\phi(x)] \geq E[f_\phi(g_\theta(z))]$$

which the optimal solution is still equivalent to solving (2)

In practice, the following objective

$$\min_\theta \max_\phi M_{f_\phi}(P_r, P_\theta) + \lambda \min(E[f_\phi(x)] - E[f_\phi(g_\theta(z))], 0)$$

Which penalizes the objective when the constraint is violated.

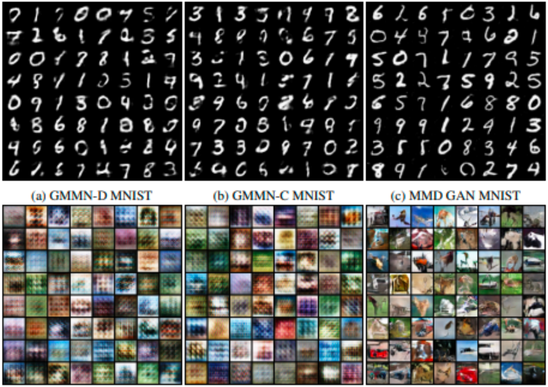
<u>Note:</u> reducing the feasible set makes the training faster and stabler.

Experiments Settings

- MNIST (50K), CIFAR10 (50K), CelebA (160K), LSUN (3M)
- DCGAN architecture based
- Kernel - Mixture of K RBF kernels $k(x, x') = \sum_{q=1}^{K} k_{\sigma_q}(x, x')$ where k_{σ_q} is a Gaussian kernel with bandwidth σ_q. Fixed $K = 5 - 1, 2, 4, 8, 16$ (left the f_ϕ learn under these σ_q
- RMSProp with LR 0.00005 (such has WGAN)
- Ensure boundedness of model parameters of the discriminator by clipping weights point-wish to $[-0.01, 0.01]$
- batch size set to 64

© Coby Penso

Qualitative Analysis - GMMN



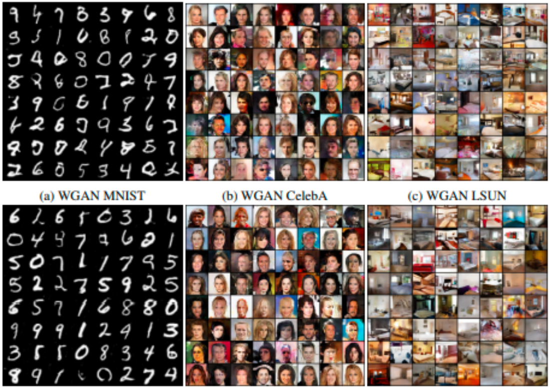
(a) GMMN-D MNIST (b) GMMN-C MNIST (c) MMD GAN MNIST

(d) GMMN-D CIFAR-10 (e) GMMN-C CIFAR-10 (f) MMD GAN CIFAR-10

Short reminder - WGAN: Change GAN critiria to:

$$\max_{w \in \mathcal{W}} \mathbb{E}_{x \sim \mathbb{P}_r}[f_w(x)] - \mathbb{E}_{z \sim p(z)}[f_w(g_\theta(z))]$$

Such that the discriminator is from a family of functions $\{f_w\}_{w \in \mathcal{W}}$ that are all K-Lipschits for some K.



(a) WGAN MNIST (b) WGAN CelebA (c) WGAN LSUN

(d) MMD GAN MNIST (e) MMD GAN CelebA (f) MMD GAN LSUN

Short reminder - Inception Score: Higher is Better - Measures diversity and quality of samples.

$$IS = exp\left(\mathbb{E}_{p_G}[KL(p(y|x)||p_\theta(y)]\right)$$

Method	Scores ± std.
Real data	11.95 ± .20
DFM [36]	7.72
ALI [37]	5.34
Improved GANs [28]	4.36
MMD GAN	**6.17** ± .07
WGAN	5.88 ± .07
GMMN-C	3.94 ± .04
GMMN-D	3.47 ± .03

Table 1: Inception scores

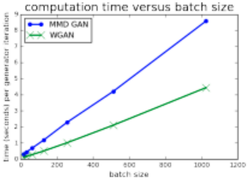
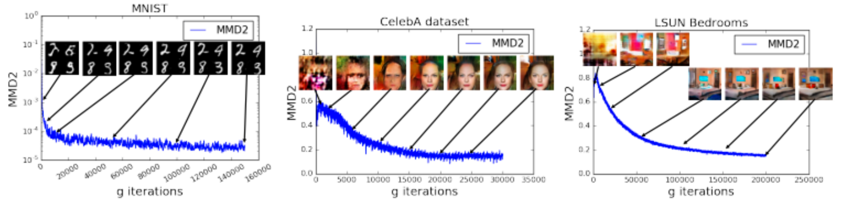


Figure 3: Computation time

Stability of MMD GAN

Moving average of MMD loss through training.



(a) MNIST (b) CelebA (c) LSUN Bedrooms

Figure 4: Training curves and generative samples at different stages of training. We can see a clear correlation between lower distance and better sample quality.

Questions?