CS446 HW5

Junsheng Huang

April 2024

1 GAN

1.1

since G is fixed, we need to max

$$
\begin{aligned}
f(D) &= \int_x p_r(x)log(D(x))dx + \int_z p_z(z)log(1 - D(g(z)))dz \\
&= \int_x p_r(x)log(D(x))dx + p_g(x)log(1 - D(x))dx
\end{aligned}
$$

For any $(a, b) \in R^2/\{0, 0\}$,$f(y) = alog(y) + blog(1 - y)$, $\frac{df}{dy} = \frac{a}{y} - \frac{b}{1-y} = 0$, which means $y = \frac{a}{a+b}$. Since that, $D^* = \frac{p_r(x)}{p_r(x)+p_g(x)}$

1.2

with D^*, the optimal problem become:

$$
\begin{aligned}
&min_G E_{x\sim p_r(x)}[log(\frac{p_r(x)}{p_r(x) + p_g(x)})] + E_{x\sim p_g(x)}[log(\frac{p_g(x)}{p_r(x) + p_g(x)})] \\
&= D_{KL}(p_r(x)|\frac{p_r(x) + p_g(x)}{2}) + D_{KL}(p_g(x)|\frac{p_r(x) + p_g(x)}{2}) - 2log(2) \\
&= 2D_{JS}(p_r(x), p_g(x)) - log(4)
\end{aligned}
$$

So optimizing Eq.1 is the same as minimizing the Jensen-Shannon (JS) divergence.

1.3

When D perfectly classifies generated samples from real data, we can say:$x \sim p_r(x), D(x) = 1$ and $x \sim p_g(x), D(x) = 0$. Thus, for the generator, the JS divergence become constant (since $p_r(x)$ and $p_g(x)$ is separated) and thus the gradient vanishes.

2 Diffusion Model

2.1

$$
\begin{aligned}
logp_\theta(x_0) &= log\int p_\theta(x_0...x_T)dx_1dx_2...dx_T \\
&= log\int p_\theta(x_{0:T})dx_{1:T} \\
&= log\int \frac{p_\theta(x_{0:T})q(x_{1:T}|x_0)}{q(x_{1:T}|x_0)}dx_{1:T} \\
&= logE_{q(x_{1:T}|x_0)}[\frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}] \\
&\geq E_{q(x_{1:T}|x_0)}[log\frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}] \\
&= E_{q(x_{1:T}|x_0)}[log\frac{p(x_T)\prod_{t=1}^T p_\theta(x_{t-1}|x_k)}{\prod_{t=1}^T q(x_t|x_{t-1})}]
\end{aligned}
$$

so ELBO is

$$
E_{q(x_{1:T}|x_0)}[log\frac{p(x_T)\prod_{t=1}^T p_\theta(x_{t-1}|x_k)}{\prod_{t=1}^T q(x_t|x_{t-1})}]
$$

2.2

No. The diffusion model only has estimation for $p(x_{t-1}|x_t)$ and $p(x_t|x_{t-1})$, it does not directly estimate $p_\theta(x_0)$. $p_\theta(x_0) = p_\theta(x_T)\prod_{t=0}^{T-1} p_\theta(x_t|x_{t+1})$, have to multiple the encoder to estimate the density.

2.3

$$
x_t = \sqrt{1-\beta_t}x_{t-1} + \beta_t\epsilon_t
$$

consider $\alpha_t = 1 - \beta_t$:

$$
\begin{aligned}
x_t &= \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\epsilon_t \\
&= \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + (\sqrt{\alpha_t(1-\alpha_{t-1})}\epsilon_{t-1} + \sqrt{1-\alpha_t}\epsilon_t) \\
&= \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\bar{\epsilon} \\
&= ... \\
&= \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\bar{\epsilon}
\end{aligned}
$$

which $\bar{\alpha}_t = \prod_{i=1}^t \alpha_t = \prod_{i=1}^t (1-\beta_t)$ so we have:

$$
q(x_t|x_0) = \sqrt{\prod_{i=1}^t (1-\beta_i)x_0} + N(0, (1 - \prod_{i=1}^t (1-\beta_i))I)
$$

2.4

$$
\begin{aligned}
q(x_{t-1}|x_t, x_0) &= q(x_t|x_{t-1}, x_0)\frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} \\
&\propto exp(-\frac{1}{2}(\frac{(x_t - \sqrt{\alpha_t}x_{t-1})^2}{\beta_t} + \frac{(x_{t-1} - \sqrt{\bar{\alpha_{t-1}}x_0})^2}{1 - \bar{\alpha_{t-1}}}) - \frac{(x_t - \sqrt{\bar{\alpha_t}}x_0)^2}{1 - \bar{\alpha_t}})
\end{aligned}
$$

sorting the stuffs inside exp by x_{t-1} order, we have:

$$
\begin{aligned}
\mu_\theta(x_t, x_0) &= \frac{\sqrt{\alpha_t}(1 - \bar{\alpha_{t-1}})}{1 - \bar{\alpha_t}}x_t + \frac{\sqrt{\bar{\alpha_{t-1}}}\beta_t}{1 - \bar{\alpha_t}}x_0 \\
&= \frac{\sqrt{1 - \beta_t}(1 - \prod_{i=1}^{t-1}(1 - \beta_i))}{1 - \prod_{i=1}^{t-1}(1 - \beta_i)}x_t + \frac{\sqrt{\prod_{i=1}^{t-1}(1 - \beta_i)}\beta_t}{1 - \prod_{i=1}^{t-1}(1 - \beta_i)}x_0
\end{aligned}
$$

2.5

we know:

$$s_\theta(x, \delta) = \nabla_x log P_\theta(x, \delta)$$

so:

$$
\begin{aligned}
s_\theta(x, \delta|x_{known}) &= \nabla_x log P_\theta(x, \delta|x_{known}) \\
&= \nabla_x log(\frac{P(x_{known|x})P_\theta(x, \delta)}{P(x_{known})}) \\
&= \nabla_x log P(x_{known}|x) + \nabla_x log P(P_\theta(x, \delta)) - \nabla_x log P(x_{known}) \\
&= \nabla_x(- \|(x - x_{known}) \odot M\|_2^2) + s_\theta(x, \delta) \\
&= -2M \odot (x - x_{known}) + s_\theta(x, \delta)
\end{aligned}
$$

3 Unsupervised learning/ contrastive learning

3.1

True.

3.2

False, usually CV model (MAE) will have a higher mask-out rate comparing with the nlp model (BERT).

3.3

True.

3.4

False. The CLIP can be directed used to classify images on labelled image classification dataset without finetuning.

4 Coding: GAN

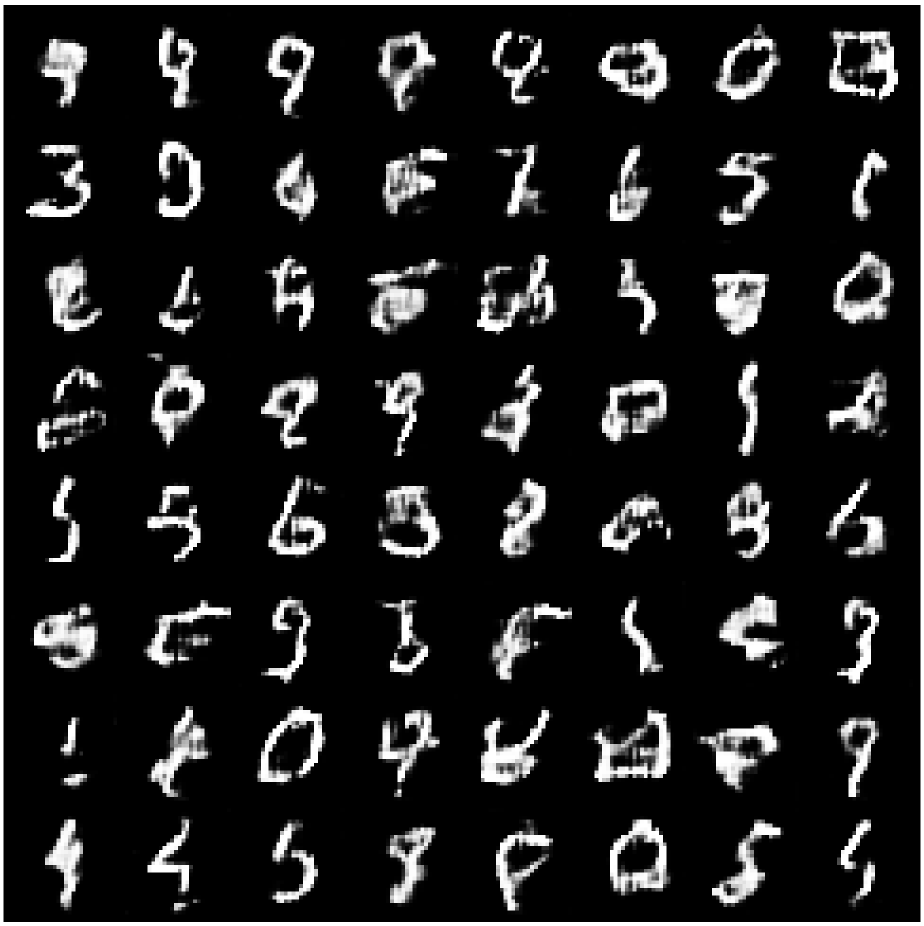

Figure 1: test 30

5 Coding: Diffusion Model

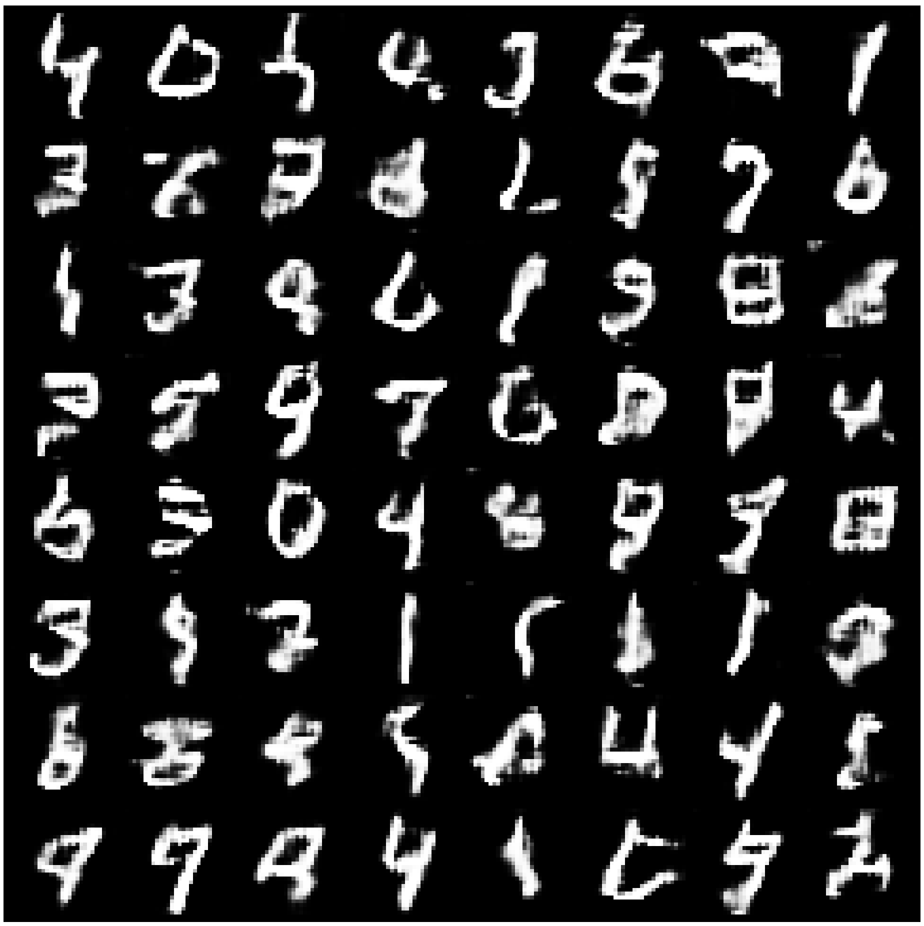

Figure 2: test 60

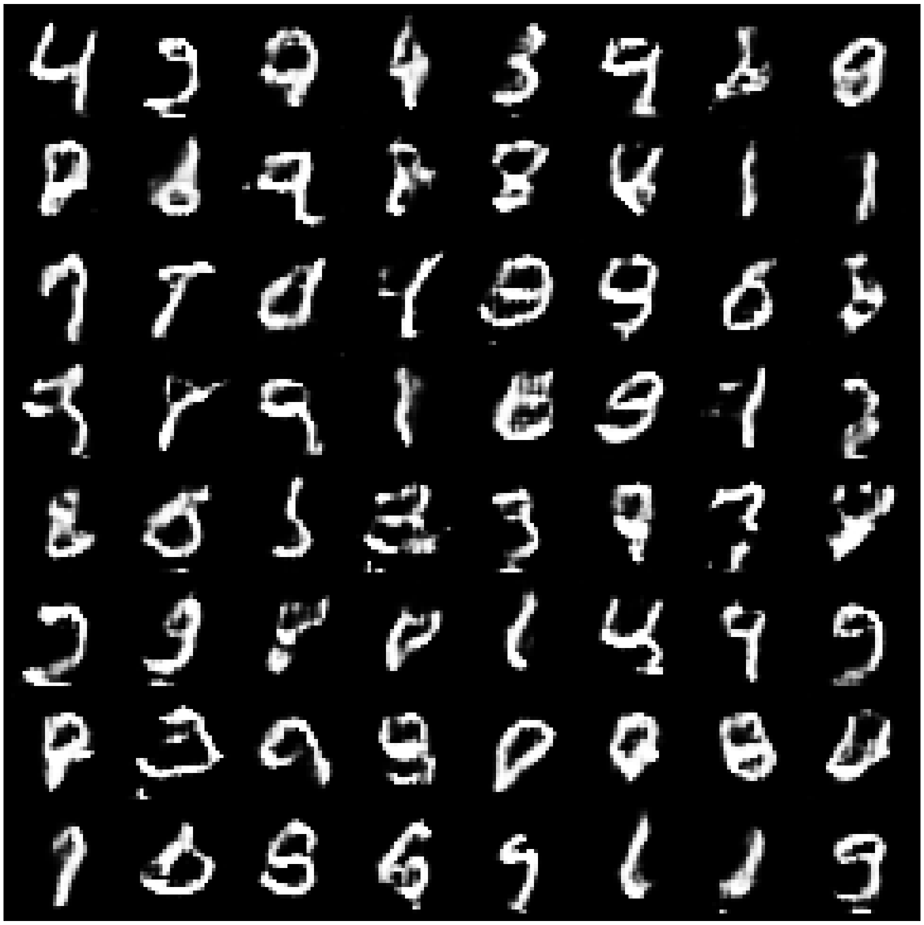

Figure 3: test 90

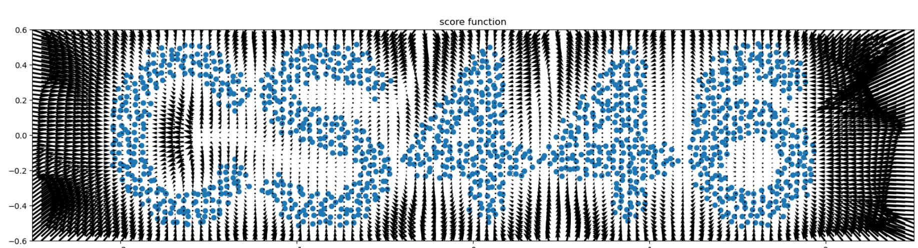

Figure 4: score

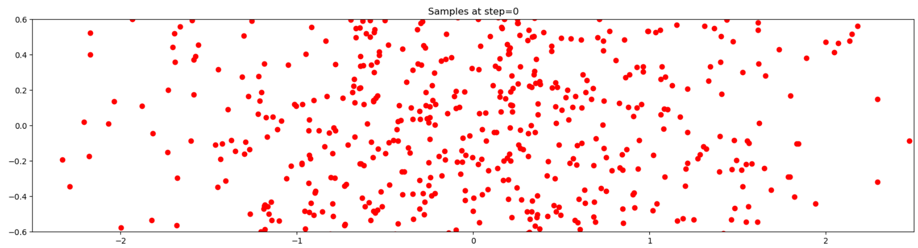

Figure 5: step 0

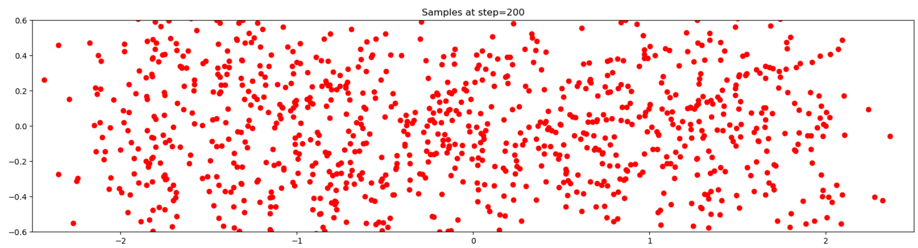

Figure 6: step 200

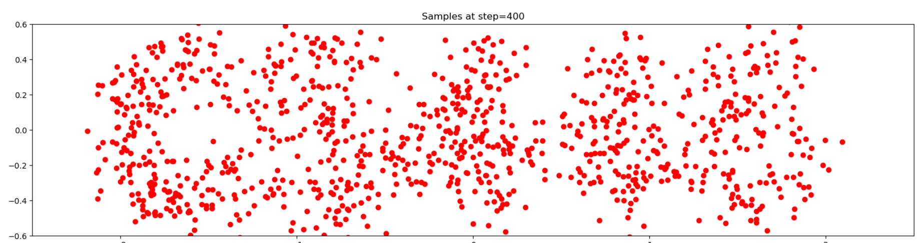

Figure 7: step 400

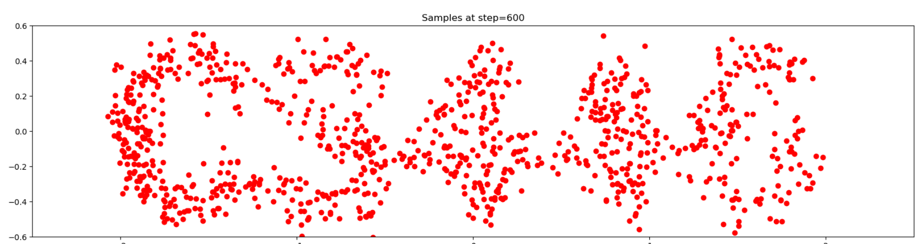

Figure 8: step 600

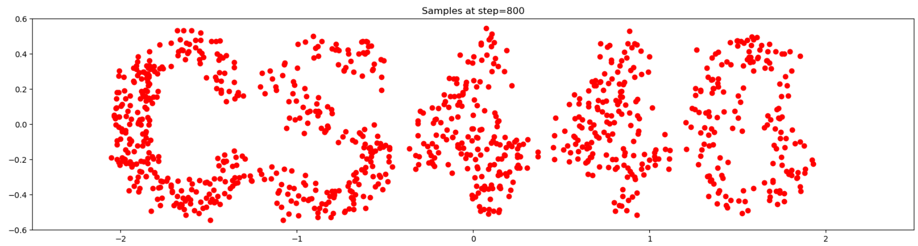

Figure 9: step 800

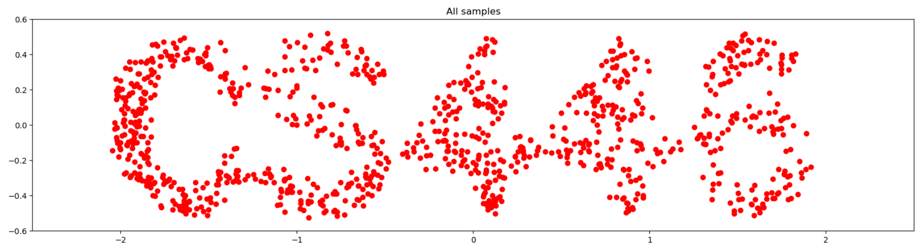

Figure 10: final

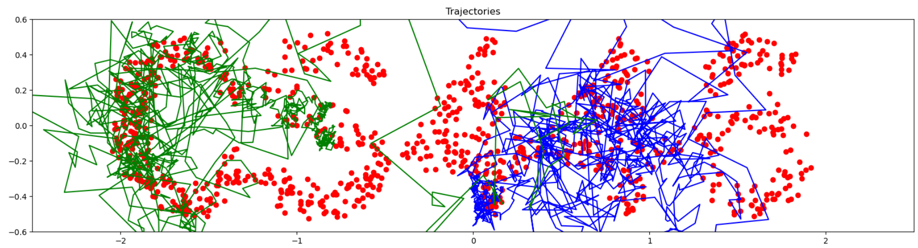

Figure 11: trajectories