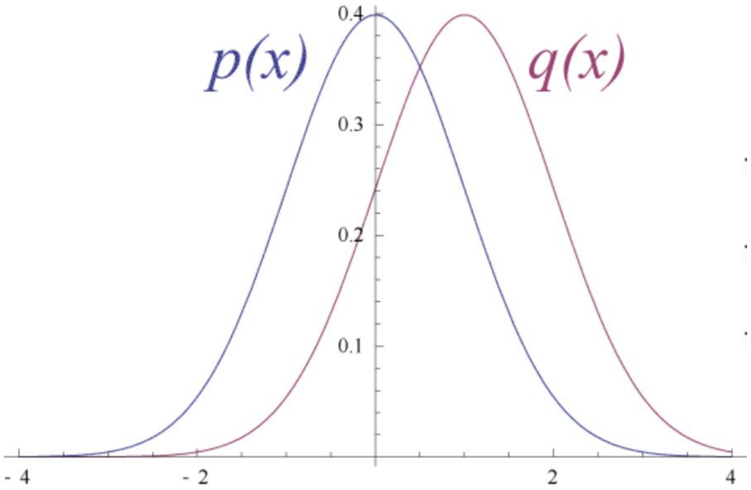
Wasserstein GAN

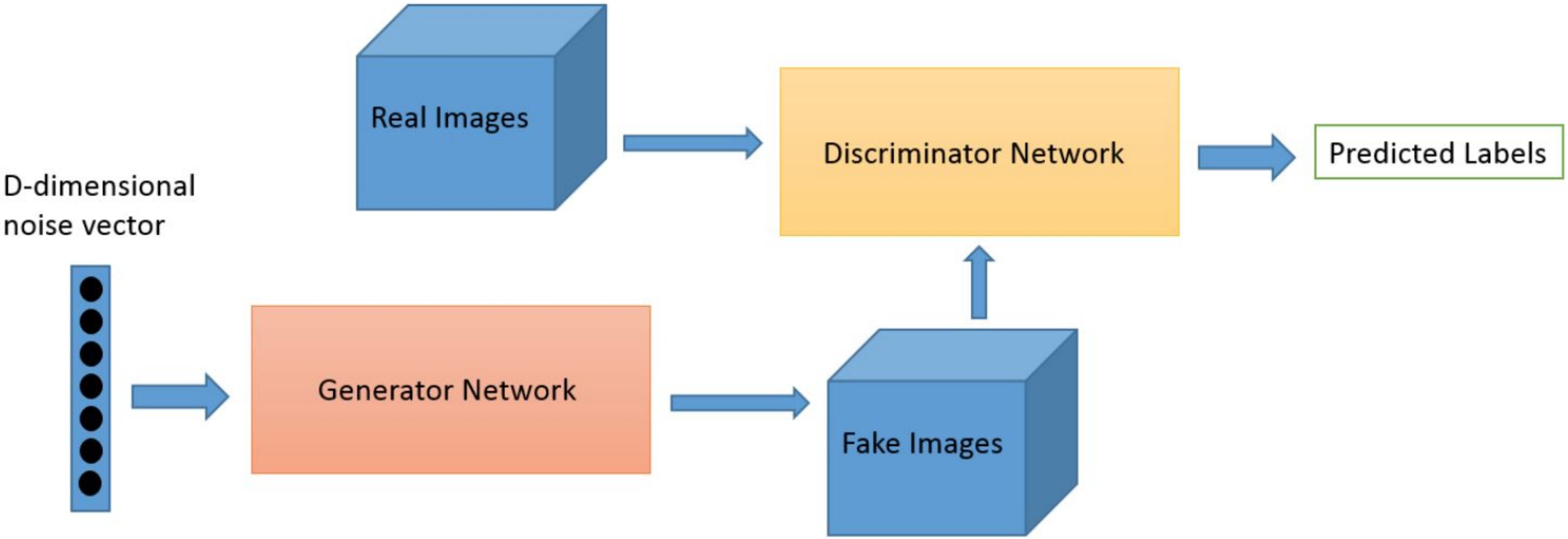
Generativni modeli



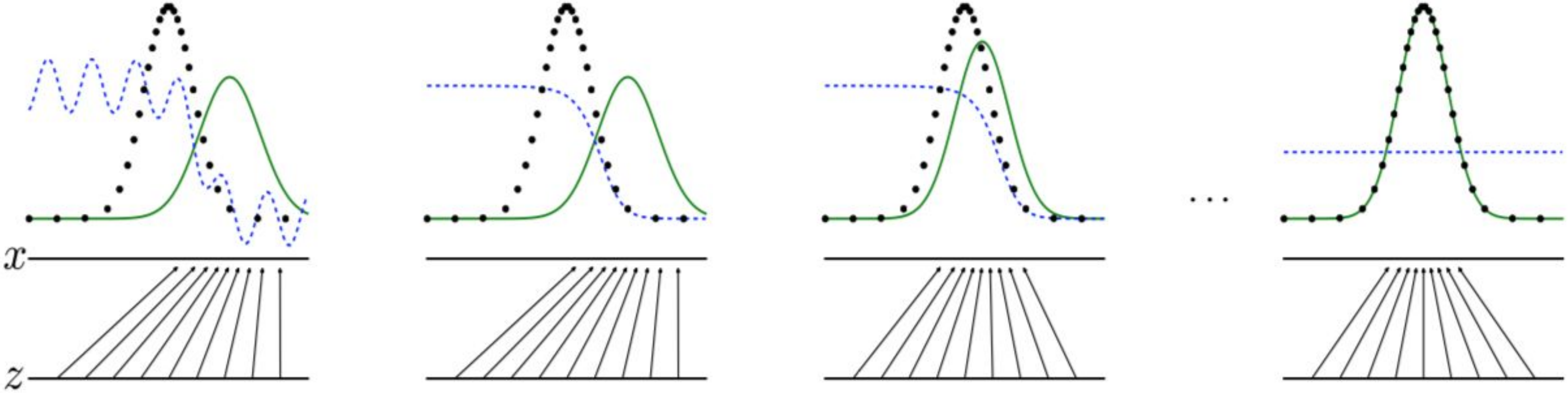
$$\max_{\theta \in \mathbb{R}^d} \frac{1}{m} \sum_{i=1}^{m} \log P_\theta(x^{(i)}) \longrightarrow KL(P||Q) = \int_x P(x) \log \frac{P(x)}{Q(x)}\, dx$$

- Predviđamo distribuciju verovatnoća nad uzoračkim prostorom.
- Centralno mesto u pristupu zauzima nekakva koncepcija distance.

Generative Adversarial Networks



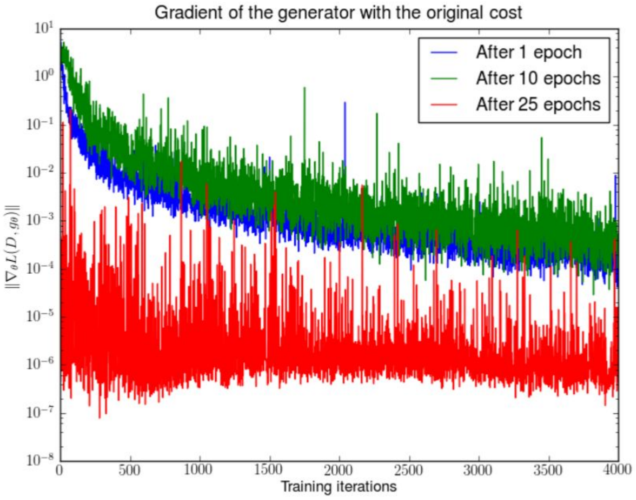
Generative Adversarial Networks

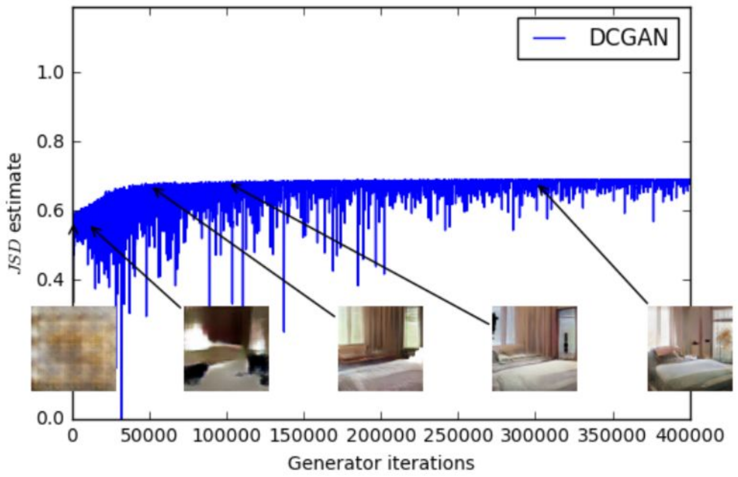


$$\min_G \max_D V(D, G) = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

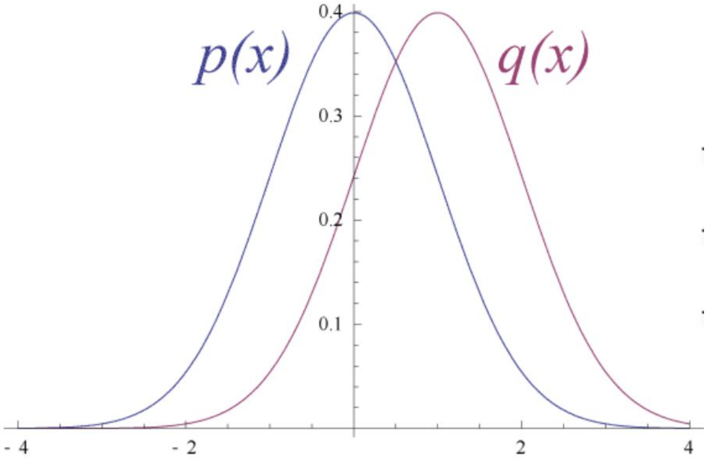
Ključni problemi

- Update postaje gori što je diskriminator bolji.
- Treniranje GAN modela je kranje nestabilno.





Disjunktni nosači gustina



Šta se dešava ako su nosači disjunktni?

Poklapanje mnogostrukosti

Neka su M i P podmnogostrukosti F i neka $x \in M \cap P$.
Kažemo da se M i P transverzalno seku u x ako $T_xM + T_xP = T_xF$

Ako se postoji $x \in M \cap P$ takvo da se u njemu M i P ne seku transverzalno,
kažemo da se ove mnogostrukosti savršeno poklapaju. Označimo ovu relaciju sa $M \sim P$.

Ukoliko su η i η' nezavisne slučajne veličine, i $\hat{M} = M + \eta$ i $\hat{P} = P + \eta'$ onda:

$$P(\hat{M} \sim \hat{P}) = 0$$

Distribucije čiji se nosači ne poklapaju

Theorem 2.2. *Let $\mathbb{P}_r$ and $\mathbb{P}_g$ be two distributions that have support contained in two closed manifolds $\mathcal{M}$ and $\mathcal{P}$ that don't perfectly align and don't have full dimension. We further assume that $\mathbb{P}_r$ and $\mathbb{P}_g$ are continuous in their respective manifolds, meaning that if there is a set A with measure 0 in $\mathcal{M}$, then $\mathbb{P}_r(A) = 0$ (and analogously for $\mathbb{P}_g$). Then, there exists an optimal discriminator $D^* : \mathcal{X} \to [0, 1]$ that has accuracy 1 and for almost any x in $\mathcal{M}$ or $\mathcal{P}$, D^* is smooth in a neighbourhood of x and $\nabla_x D^*(x) = 0$.*

Theorem 2.3. *Let $\mathbb{P}_r$ and $\mathbb{P}_g$ be two distributions whose support lies in two manifolds $\mathcal{M}$ and $\mathcal{P}$ that don't have full dimension and don't perfectly align. We further assume that $\mathbb{P}_r$ and $\mathbb{P}_g$ are continuous in their respective manifolds. Then,*

$$JSD(\mathbb{P}_r \| \mathbb{P}_g) = \log 2$$
$$KL(\mathbb{P}_r \| \mathbb{P}_g) = +\infty$$
$$KL(\mathbb{P}_g \| \mathbb{P}_r) = +\infty$$

Cost funkcije

Theorem 2.4 (Vanishing gradients on the generator). *Let $g_\theta : \mathcal{Z} \to \mathcal{X}$ be a differentiable function that induces a distribution $\mathbb{P}_g$. Let $\mathbb{P}_r$ be the real data distribution. Let D be a differentiable discriminator. If the conditions of Theorems 2.1 or 2.2 are satisfied, $\|D - D^*\| < \epsilon$, and $\mathbb{E}_{z\sim p(z)}\left[\|J_\theta g_\theta(z)\|_2^2\right] \leq M^2$,* [2] *then*

$$\|\nabla_\theta \mathbb{E}_{z\sim p(z)}[\log(1 - D(g_\theta(z)))]\|_2 < M\frac{\epsilon}{1-\epsilon}$$

$$\lim_{\|D-D^*\|\to 0} \nabla_\theta \mathbb{E}_{z\sim p(z)}[\log(1 - D(g_\theta(z)))] = 0$$

logD trik

Theorem 2.5. *Let $\mathbb{P}_r$ and $\mathbb{P}_{g_\theta}$ be two continuous distributions, with densities P_r and P_{g_θ} respectively. Let $D^* = \frac{P_r}{P_{g_{\theta_0}} + P_r}$ be the optimal discriminator, fixed for a value θ_0*[3]. *Therefore,*

$$\mathbb{E}_{z\sim p(z)}\left[-\nabla_\theta \log D^*(g_\theta(z))|_{\theta=\theta_0}\right] = \nabla_\theta \left[KL(\mathbb{P}_{g_\theta}\|\mathbb{P}_r) - 2JSD(\mathbb{P}_{g_\theta}\|\mathbb{P}_r)\right]|_{\theta=\theta_0} \quad (3)$$

Noisy generator

Jedno od rešenja ovog problema je dodavanje normalno raspodeljenog šuma generatoru i podacima. U tom slučaju loss ima stabilne gradijente i konvergira simetričnom izrazu:

$$2\nabla_\theta JSD(\mathbb{P}_{r+\epsilon}\|\mathbb{P}_{g+\epsilon})$$

Sa druge strane, intenzitet šuma koji je potreban da bi se postiglo stabilno treniranje je krajnje veliki i drastično obara kvalitet generisanih vrednosti.

Alternativna metrika/divergencija

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma} \big[\|x - y\| \big]$$

Wasserstein u odnosu na druge divergencije

Theorem 1. *Let $\mathbb{P}_r$ be a fixed distribution over $\mathcal{X}$. Let Z be a random variable (e.g Gaussian) over another space $\mathcal{Z}$. Let $g : \mathcal{Z} \times \mathbb{R}^d \to \mathcal{X}$ be a function, that will be denoted $g_\theta(z)$ with z the first coordinate and θ the second. Let $\mathbb{P}_\theta$ denote the distribution of $g_\theta(Z)$. Then,*

1. *If g is continuous in θ, so is $W(\mathbb{P}_r, \mathbb{P}_\theta)$.*
2. *If g is locally Lipschitz and satisfies regularity assumption 1, then $W(\mathbb{P}_r, \mathbb{P}_\theta)$ is continuous everywhere, and differentiable almost everywhere.*
3. *Statements 1-2 are false for the Jensen-Shannon divergence $JS(\mathbb{P}_r, \mathbb{P}_\theta)$ and all the KLs.*

Wasserstein dual i treniranje

$$W(\mathbb{P}_r, \mathbb{P}_\theta) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x \sim \mathbb{P}_r}[f(x)] - \mathbb{E}_{x \sim \mathbb{P}_\theta}[f(x)]$$

$$\downarrow$$

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

Algoritam

Algorithm 1 Minibatch stochastic gradient descent training of generative adversarial nets. The number of steps to apply to the discriminator, k, is a hyperparameter. We used $k = 1$, the least expensive option, in our experiments.

for number of training iterations **do**

for k steps **do**

- Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
- Sample minibatch of m examples $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ from data generating distribution $p_{\text{data}}(\boldsymbol{x})$.
- Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right)\right].$$

end for

- Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
- Update the generator by descending its stochastic gradient:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right).$$

end for

The gradient-based updates can use any standard gradient-based learning rule. We used momentum in our experiments.

Algorithm 1 WGAN, our proposed algorithm. All experiments in the paper used the default values $\alpha = 0.00005$, $c = 0.01$, $m = 64$, $n_{\text{critic}} = 5$.

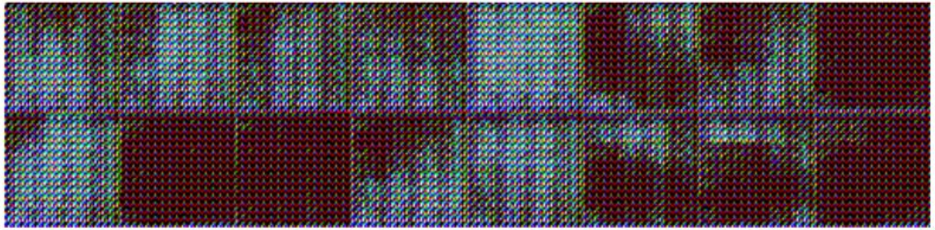
Require: : α, the learning rate. c, the clipping parameter. m, the batch size. n_{critic}, the number of iterations of the critic per generator iteration.

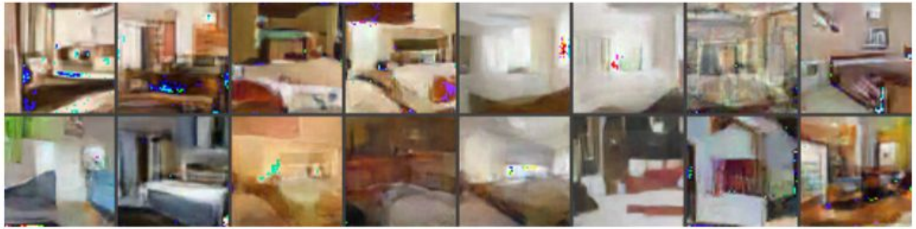
Require: : w_0, initial critic parameters. θ_0, initial generator's parameters.

while θ has not converged **do**
  for $t = 0, \ldots, n_{\text{critic}}$ **do**
    Sample $\{x^{(i)}\}_{i=1}^{m} \sim \mathbb{P}_r$ a batch from the real data.
    Sample $\{z^{(i)}\}_{i=1}^{m} \sim p(z)$ a batch of prior samples.
    $g_w \leftarrow \nabla_w \left[\frac{1}{m}\sum_{i=1}^{m} f_w(x^{(i)}) - \frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))\right]$
    $w \leftarrow w + \alpha \cdot \text{RMSProp}(w, g_w)$
    $w \leftarrow \text{clip}(w, -c, c)$
  end for
  Sample $\{z^{(i)}\}_{i=1}^{m} \sim p(z)$ a batch of prior samples.
  $g_\theta \leftarrow -\nabla_\theta \frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))$
  $\theta \leftarrow \theta - \alpha \cdot \text{RMSProp}(\theta, g_\theta)$
end while

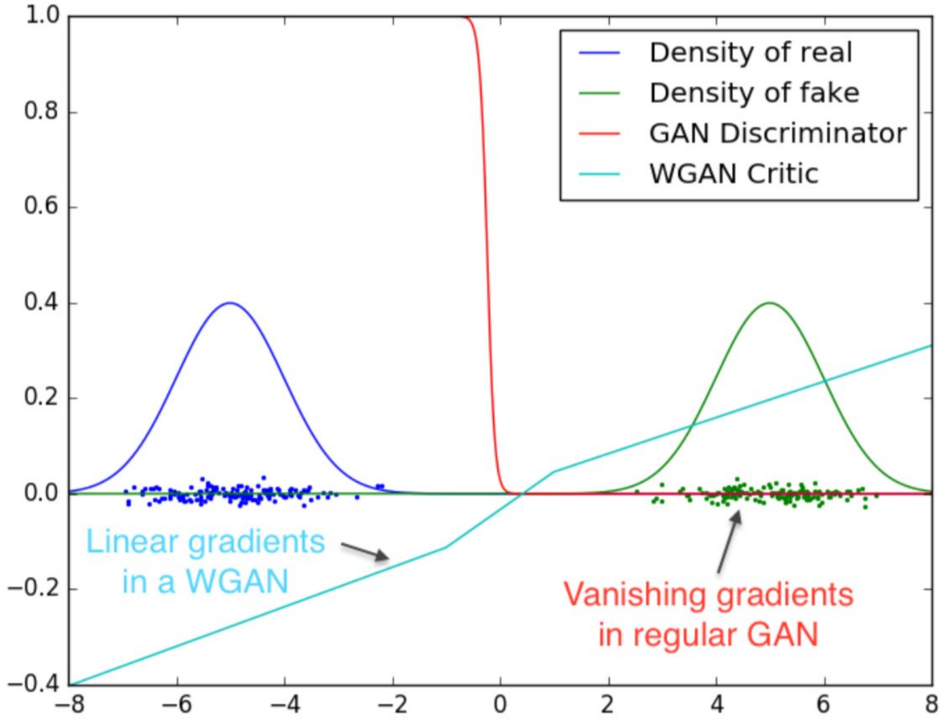
Praktični rezultati

- Batchnorm nema nikakvog efekta na proces treniranja
- Treniranje postaje praktično neosetljivo na arhitekturu generatora





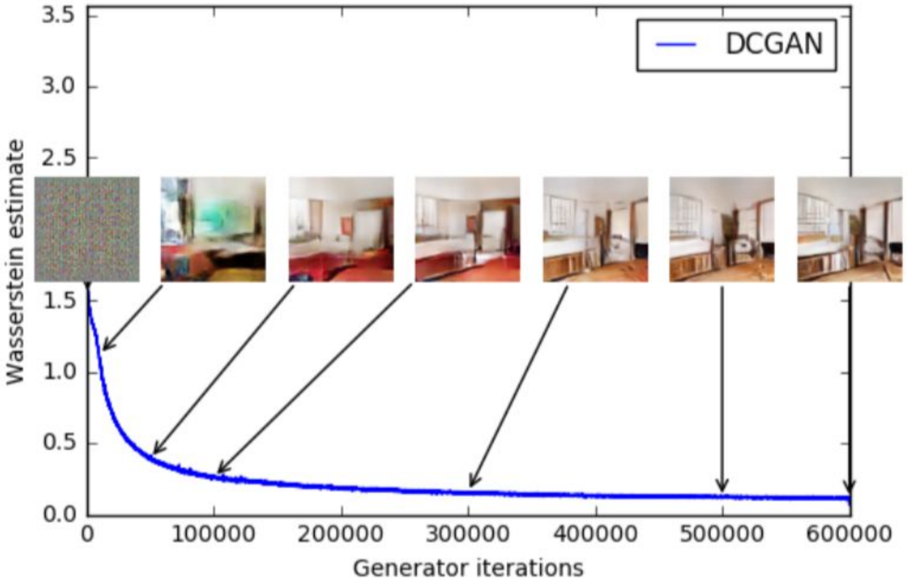
Praktični rezultati



Diskriminator(kritičar) ima skoro konstantan gradijent

Praktični rezultati

Loss funkcija se može interpretirati na razuman način



Problemi

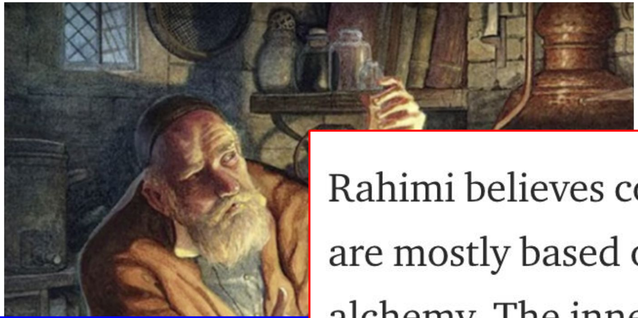
- "Weight clipping" kod diskriminatora.
- Spora konvergencija na visokodimenzionim podacima.

Teorija i praksa

Medium

Sign

LeCun vs Rahimi: Has Machine Learning Become Alchemy?



Yann LeCun
December 6 at 8:57am ·

My take on Ali Rahimi's "Test of Time" award talk at NIPS.

Ali gave an entertaining and well-delivered talk. But I fundamentally disagree with the message.
The main message was, in essence, that the current practice in machine learning is akin to "alchemy" (his word).
It's insulting, yes. But never mind that: It's wrong!

Rahimi believes contemporary machine learning models' successes—which are mostly based on empirical methods—are plagued with the same issues as alchemy. The inner mechanisms of machine learning models are so complex and opaque that researchers often don't understand why a machine learning model can output a particular response from a set of data inputs, aka the black box problem. Ranimi believes the lack of theoretical understanding or