AGE Networks and BEGAN

It Takes (Only) Two: Adversarial Generator-Encoder Networks
D. Ulyanov¹, A. Vedaldi² and V. Lempitsky¹
¹Skolkovo Institute of Science and Technology, ²University of Oxford

BEGAN: Boundary Equilibrium Generative Adversarial Networks
D. Berthelot, T. Schumm and L. Metz
Google

Discussion by Ikenna Odinaka
July 14, 2017
1. GAN Recap
2. Bi-Directional GAN Recap
3. BEGAN Networks
4. AGE Networks
5. My Concerns with BEGAN
GAN: Main Idea

- An adversarial game between:
  1. Discriminator $D$
  2. Generator $G$

- Discriminator is trained to discriminate between samples $x$ from:
  1. Real data distribution $q(x)$
  2. Generated data distribution $p(x) = \int p(z)p(x|z)dz$, implemented via a generator by sampling $z \sim p(z)$ and then sampling $x \sim p(x|z)$

- Generator learns conditional $p(x|z)$ to fool the discriminator
GAN: Formulation

The adversarial game is formalized by the value function

$$\min_G \max_D V(D, G) = \mathbb{E}_{q(x)}[\log(D(x))] + \mathbb{E}_{p(z)}[\log(1 - D(G(z)))]$$  \hspace{1cm} (1)

- Lacks an efficient mechanism to infer \(z\) given \(x\)
- Goal of extensions: Align real data distribution to generated data distribution and create a correspondence between data and latent samples
Cast the learning of both an inference machine (encoder) and a deep directed generative model (decoder) in a GAN-like adversarial framework.

- **Discriminator** is trained to discriminate between joint samples \((x, z)\) from:
  1. Encoder distribution \(q(x, z) = q(x) q(z|x)\), or
  2. Decoder distribution \(p(x, z) = p(x|z) p(z)\)

- Generator learns conditionals \(q(z|x)\) and \(p(x|z)\) to fool the discriminator.
The adversarial game is formalized by the value function

\[
\min_G \max_D V(D, G) = \mathbb{E}_{q(x)}[\log(D(x, G_z(x)))] + \mathbb{E}_{p(z)}[\log(1 - D(G_x(z), z))] \quad (2)
\]
BEGAN: Main Ideas and Contributions

- Similar to EBGAN, use auto-encoder as discriminator
- Unlike EBGAN and other GANs, match auto-encoder loss distributions
- Divergence between distributions using a lower bound of Wasserstein distance
- Proportional Control Theory to maintain equilibrium between power of discriminator and generator
- Control image diversity and visual quality trade-off using diversity ratio
- Similar to Wasserstein GAN, define an approximate measure of convergence
- Simple training with fast and stable convergence
  - No need to pretrain discriminator
  - No need to alternate training discriminator and generator. Can update parameters of generator and encoder in parallel per learning step
BEGAN: Notation

- $x$ sample from real data distribution. $N_x$ dimension of $x$
- $z \in [-1, 1]^{N_z}$ uniform random samples of dimension $N_z$
- $z_G$ samples used in generator, $z_D$ samples used in auto-encoder/discriminator
- $D : \mathbb{R}^{N_x} \leftrightarrow \mathbb{R}^{N_x}$ auto-encoder, $G : \mathbb{R}^{N_z} \rightarrow \mathbb{R}^{N_x}$ generator
- Pixel-wise auto-encoder loss given as
  \[
  \mathcal{L}(\nu) = |\nu - D(\nu)|^\eta
  \]
  where $\eta \in \{1, 2\}$, $\nu \in \mathbb{R}^{N_x}$
- $\mu_1$ auto-encoder loss distribution for real data samples $\mathcal{L}(x)$, $\mu_2$ for generated samples $\mathcal{L}(G(z))$
- $\Gamma(\mu_1, \mu_2)$ set of all couplings of $\mu_1$ and $\mu_2$
- $m_i \in \mathbb{R}^+$ mean of $\mu_i$
- $k$ control variable. $k_t \in [0, 1]$, $t$ learning step, $\lambda_k$ learning rate for $k$
BEGAN: Formulation

- Using Jensen’s inequality for convex function $f$, $\mathbb{E}[f(x)] \geq f(\mathbb{E}[x])$
- The lower bound to the Wasserstein distance between auto-encoder loss distributions is given by

$$W_1(\mu_1, \mu_2) = \inf_{\tau \in \Gamma(\mu_1, \mu_2)} \mathbb{E}(y_1, y_2) \sim \tau [|y_1 - y_2|]$$

$$\geq \inf_{\tau \in \Gamma(\mu_1, \mu_2)} [\mathbb{E}(y_1, y_2) \sim \tau [y_1 - y_2]]$$

$$= |m_1 - m_2|$$

$$= \begin{cases} m_1 - m_2, & \text{if } m_1 \geq m_2 \\ m_2 - m_1, & \text{if } m_2 > m_1 \end{cases} \quad (3)$$

- Discriminator attempts to maximize equation 3.
- Maximizing second lower bound $W_1(\mu_1, \mu_2) \geq m_2 - m_1$ implies $m_1 \to 0$ and $m_2 \to \infty$
- Since discriminator is an auto-encoder, naturally select this lower bound
GAN objective based on loss distributions is

\[
\begin{align*}
\mathcal{L}_D &= \mathcal{L}(x; \theta_D) - \mathcal{L}(G(z_D; \theta_G); \theta_D) \quad \text{for } \theta_D \\
\mathcal{L}_G &= -\mathcal{L}_D|_{z_D \rightarrow z_G} = \mathcal{L}(G(z_G; \theta_G); \theta_D) \quad \text{for } \theta_G
\end{align*}
\]  

(4)

Similar objective to WGAN, but discriminator not required to be K-Lipschitz and distributional divergence/similarity on loss not data space
BEGAN: BEGAN Objective

- Equilibrium between generator and discriminator when
  \[ \mathbb{E}[\mathcal{L}(G(z))] = \mathbb{E}[\mathcal{L}(x)] \]
- Relax equilibrium \[ \mathbb{E}[\mathcal{L}(G(z))] = \gamma \mathbb{E}[\mathcal{L}(x)] \] using diversity ratio \( \gamma \in [0, 1] \)
- BEGAN objective using **Proportional Control Theory** is (ignoring the network parameters)

\[
\begin{align*}
\mathcal{L}_D &= \mathcal{L}(x) - k_t \mathcal{L}(G(z_D)) \\
\mathcal{L}_G &= \mathcal{L}(G(z_G)) \\
k_{t+1} &= k_t + \lambda_k (\gamma \mathcal{L}(x) - \mathcal{L}(G(z_G)))
\end{align*}
\]  
for \( \theta_D \)  
for \( \theta_G \)  
for each training step \( t \)  

(5)

- Convergence measure is given as

\[
\mathcal{M}_{global} = \mathcal{L}(x) + |\gamma \mathcal{L}(x) - \mathcal{L}(G(z_G))|
\]

(6)
**Figure**: Random samples comparison. (a) CelebA (b) private 360K celebrity faces

**Figure**: Random samples varying diversity ratio
Figure: Latent space interpolation of real images
BEGAN: Experiments - Diagnostics

Figure: Convergence and image quality

(a) Starved generator (\(z = 16\) and \(h = 128\))  
(b) Starved discriminator (\(z = 128\) and \(h = 16\))

Figure: Generator-Discriminator Imbalance
AGE: Main Ideas and Contributions

- Uses only two networks: generator and encoder
- Generator maps samples from prior distribution (in latent space) to data space
- Encoder maps real data and generated data, inducing two latent distributions
- AGE learning considers the divergence of the two latent distributions to the original prior distribution
- Divergence function must be non-negative and zero iff the distributions are identical
- No need for a separate discriminator network as in other GAN extensions like BiGAN or BEGAN
- AGE is adversarial because encoder tries to increase the distance between the divergences, while generator tries to decrease it
AGE: Notation

- data space $\mathcal{X}$, latent space $\mathcal{Z}$
- real data distribution $X$, prior distribution $Z$
- encoder $e_\psi(x)$, generator $g_\theta(z)$
- $f(Y)$ denotes the distribution of the random variable $f(y), y \sim Y$
- $\Delta(X \| Y)$ is the divergence between two distributions $X$ and $Y$
- $\psi$ encoder parameters, $\theta$ generator parameters
- $\bar{\psi}$ current value of encoder parameters, $\bar{\theta}$ current value of generator parameters
- $\lambda$ and $\mu$ user-defined coefficients
Theorem (optimal encoder and generator)

- A pair \((g^*, e^*)\) forms a saddle point of the game
  \[
  \max_e \min_g V_1(g, e) = \Delta(e(g(Z)) \| e(X)) \text{ if and only if the generator } g^*
  \]
  matches the data distribution, i.e. \(g^*(Z) = X\).

- Let \(Y\) be any fixed distribution in the latent space. A pair \((g^*, e^*)\) forms a
  saddle point of the game
  \[
  \max_e \min_g V_2(g, e) = \Delta(e(g(Z)) \| Y) - \Delta(e(X) \| Y) \text{ if and only if the}
  \]
  generator \(g^*\) matches the data distribution, i.e. \(g^*(Z) = X\).

Theorem (encoder-generator reciprocity)

- Let the two distributions \(W\) and \(Q\) be aligned by the mapping \(f\) (i.e. \(f(W) = Q\)) and let
  \[
  \mathbb{E}_{w \sim W} ||w - h(f(w))||^2_2 = 0. \text{ Then, for } w \sim W \text{ and}
  \]
  \(q \sim Q\), we have \(w = h(f(w))\) and \(q = f(h(q))\) almost certainly, i.e. the
  mappings \(f\) and \(h\) invert each other almost everywhere on the supports of
  \(W\) and \(Q\). Furthermore, \(Q\) is aligned with \(W\) by \(h\), i.e. \(h(Q) = W\).
AGE: Formulation

- The adversarial game is formalized by the value function

$$\max_e \min_g V_2(g, e) = \Delta(e(g(Z))\|Y) - \Delta(e(X)\|Y)$$

where $Y = Z$

- For encoder-decoder reciprocity, define reconstruction losses in latent and data spaces

$$L_X(g_\theta, e_\psi) = \mathbb{E}_{x \sim X} \|x - g_\theta(e_\psi(x))\|_1$$

(8)

$$L_Z(g_\theta, e_\psi) = \mathbb{E}_{z \sim Z} \|z - e_\psi(g_\theta(z))\|_2^2$$

(9)

- Following theorems above, the generator and encoder parameters are optimized following

$$\hat{\theta} = \arg\min_\theta \left( V_2(g_\theta, e_\psi) + \lambda L_Z(g_\theta, e_\psi) \right)$$

(10)

$$\hat{\psi} = \arg\max_\psi \left( V_2(g_{\tilde{\theta}}, e_\psi) - \mu L_X(g_{\tilde{\theta}}, e_\psi) \right)$$

(11)
Inception score on CIFAR-10: AGE 5.90, ALI 5.34, BEGAN 5.62
Faster convergence than ALI, faster computation per epoch than other GAN extensions (2 networks instead of 3)

Figure: Tiny ImageNet test dataset (top), SVHN test dataset (bottom). For (c) and (d), odd columns show real examples and even columns show their reconstructions
Figure: CelebA test dataset
AGE: Experiments - Conditional

- Images in *Lab* color space
- $L$ is input to both encoder and generator
- Colorization: Use different color embeddings applied to image $L$ channel
- Color transfer: Transfer color from one image to another’s $L$ channel

Figure: AGE (odd rows), pix2pix (bottom rows)
Contradiction Between Relaxed Equilibrium and Lower Bound

- Authors chose second lower bound for Wasserstein distance $W_1(\mu_1, \mu_2) \geq m_2 - m_1$ because discriminator maximization of the lower bound implies $\mathbb{E}[\mathcal{L}(x)] = m_1 \to 0$ (desirable for an auto-encoder for real data samples) and $\mathbb{E}[\mathcal{L}(G(z))] = m_2 \to \infty$

- Authors used $\gamma = \frac{\mathbb{E}[\mathcal{L}(G(z))]}{\mathbb{E}[\mathcal{L}(x)]}$, $\gamma \in [0, 1]$ to ensure relaxed equilibrium between generator and discriminator.

- This implies $m_2 = \gamma m_1$ and $m_2 \leq m_1$. This contradicts the choice of the second lower bound for the Wasserstein distance, where $m_2 \geq m_1$

- Authors found that diversity and quality are dependent contradicting the work of Poole et al. [Improved techniques for training GANs]
In the passage following equation 5 the authors mentioned that for the control variable update

\[ k_{t+1} = k_t + \lambda_k (\gamma \mathcal{L}(x) - \mathcal{L}(G(z_G))) \]

- \( k_t \in [0, 1], \ k_0 = 0, \ \lambda_k = 10^{-3}, \) and \( \gamma \in [0, 1] \)
- With this setup, \( k_t \) is not guaranteed to remain in \([0, 1]\) unless there is a missing projection onto \([0, 1]\) in the text
- With the projection onto \([0, 1]\), \( \gamma = 0 \) and \( k_0 = 0 \) is not a valid pair, since \( \mathcal{L}(G(z)) \geq 0 \) and \( k_t = 0, \ \forall t \)
- More generally, with the projection onto \([0, 1]\), if \( k_0 = 0 \), then \( \gamma \mathcal{L}(x) > \mathcal{L}(G(z_G)) \). Otherwise, \( k_t = 0, \ \forall t \)