Human painters usually first draw some abstract stuff, then gradually add details. To mimic this process, we learn a generator that first produces high-level abstract features, then gradually generates lower level features and finally the image.

Generative Adversarial Networks (GAN):
• Two networks competing with each other.
• Discriminator $D$ tries to distinguish between real samples and samples generated by generator $G$.
• $G$ tries to “fool” $D$.
• $G$ will learn to generate samples similar to real data.

Motivation

A stack of GANs, each GAN generates lower-level features conditioned on higher-level features.

Each generator is trained with three loss terms:
• Adversarial loss: the generated features should be indistinguishable from “real” features.
  $$L_{adv} = E_{\tilde{h}_{t+1}} \sim p_{data}(\tilde{h}_{t+1}) \left[ - \log D_G(h_{t+1}, z_{t+1}) \right]$$
• Conditional loss: the generator should make use of the higher-level features it’s conditioned on:
  $$L_{cond} = E_{\tilde{h}_{t+1}} \sim p_{data}(\tilde{h}_{t+1}) \left[ f(E_G(\tilde{h}_{t}), h_{t+1}) \right]$$
• Entropy loss: encourage sample diversity by maxi-mizing a variational lower bound on the entropy
  $$L_{cond} = E_{\tilde{h}_{t+1}} \sim p_{data}(\tilde{h}_{t+1}) \left[ - \log Q_G(\tilde{v} | \tilde{h}_{t}) \right]$$

Architecture

Background

Quantitative evaluations

• Inception score on CIFAR-10:
  - Method: Score
  - Infusion training [1]: 4.62 ± 0.06
  - ALI [10] (as reported in [63]): 5.34 ± 0.05
  - EGAN-Ent-V [4]: 7.07 ± 0.10
  - LR-GAN [65]: 7.17 ± 0.07
  - Denoising feature matching [63]: 7.72 ± 0.13
  - DCGAN$^\dagger$ (with labels, as reported in [61]): 6.58
  - SteinGAN$^\dagger$ [61]: 6.35
  - Improved GAN$^\dagger$ [53] (best variant): 8.09 ± 0.07
  - AC-GAN$^\dagger$ [63]: 8.25 ± 0.07
  - DCGAN (L$^{adv}$): 6.16 ± 0.07
  - DCGAN (L$^{adv}$ + L$^{cond}$): 5.40 ± 0.16
  - DCGAN (L$^{adv}$ + L$^{cond}$)$^\dagger$: 5.40 ± 0.08
  - DCGAN (L$^{adv}$ + L$^{cond}$ + L$^{ent}$)$^\dagger$: 7.16 ± 0.10
  - SGAN-no-joint$^\dagger$: 8.37 ± 0.08
  - SGAN$^\dagger$: 8.59 ± 0.12

Real data: 11.24 ± 0.12

$^\dagger$ Trained with labels.

Human visual Turing tests on UfAH-1U: We ask AMT workers to distinguish generated images from real images. Our samples “fool” people 24.4% of the time, higher than our best DCGAN baseline (15.6%) and Improved GAN (21.3%).