Deep Learning

Lecture 6: Adversarial models

Chris G. Willcocks

Durham University

Lecture overview

1 Generative adversarial networks

- definition
- properties
- mode collapse
- Lipschitz continuity
- spectral normalisation
- conditional GANs
- information maximizing GANs
- adversarial autoencoders

2 Popular applications

- unpaired translation
- super resolution
- adversarial anomaly detection

3 Adversarial examples

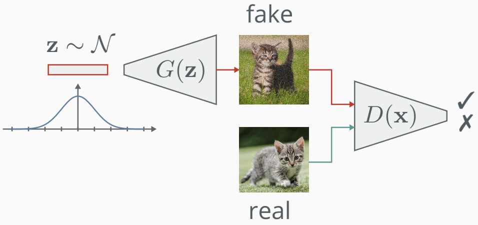
- attacks
- defences

Definition: generative adversarial networks

A generative adversarial network (GAN) is a non-coorporative zero-sum game where two networks compete against each other [1].

One network $G(\mathbf{z})$ generates new samples, whereas D estimates the probability the sample was from the training data rather than G:

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

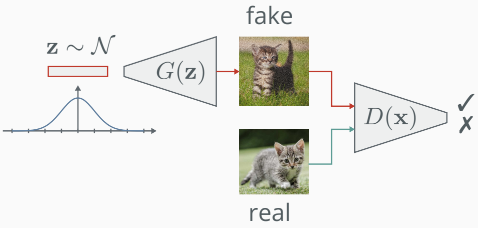


GAN **properties**

GANs benefit from differentiable data augmentation [2] for both reals and fakes, but are otherwise notoriously difficult to train:

- Non-convergence
- Diminishing gradient
- Difficult to balance
- Mode collapse (next slide)

Link to Colab example



Generative adversarial networks mode collapse

Definition: mode collapse

This is where the generator rotates through a small subset of outputs, and the discriminator is unable to get out of the trap. Mode collapse is arguably the main limitation of GANs.

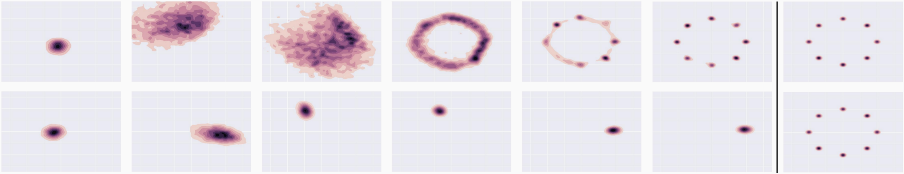


Figure from [3]. The final column shows the target data distribution and the bottom row shows a GAN rotating through the modes.

Definition: Lipschitz function

A function f is Lipschitz continous if it is bounded by how fast it can change. Specifically if there exists a positive real constant k where:

$$|f(x) - f(y)| \leq k|x - y|,$$

for all y sufficiently near x. For example, any function with a bounded first derivative is a Lipschitz function.

Wasserstein GANs [4, 5] were the first to reduce mode collapse in GANs by lowering the Lipschitz constant for the descriminator function.

Distance functions are 1-Lipschitz

Definition: spectral normalisation

The matrix (spectral) norm defines how much a matrix can stretch a vector $\mathbf{x}$:

$$\|\boldsymbol{A}\| = \max_{\mathbf{x}\neq 0} \frac{\|\boldsymbol{A}\mathbf{x}\|}{\|\mathbf{x}\|}$$

Spectral norm [6] normalises the weights for each layer using the spectral norm $\sigma(\boldsymbol{W})$ such that the Lipschitz constant for every layer and the whole network is 1:

$$\hat{\boldsymbol{W}}_{\text{SN}} = \boldsymbol{W}/\sigma(\boldsymbol{W})$$
$$\sigma(\hat{\boldsymbol{W}}_{\text{SN}}(\boldsymbol{W})) = 1$$
$$\|f\|_{\text{Lip}} = 1$$

Pseudocode: 1-Lipschitz discriminator

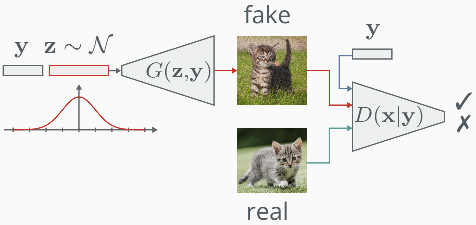
```
class Discriminator(nn.Module):
  def __init__(self, f=64):
    super().__init__()
    self.discriminate = nn.Sequential(
      spectral_norm(Conv2d(1, f, 3, 1, 1)),
      nn.LeakyReLU(0.1, inplace=True),
      nn.MaxPool2d(kernel_size=(2,2)),
      spectral_norm(Conv2d(f, f*2, 3, 1, 1)),
      nn.LeakyReLU(0.1, inplace=True),
      nn.MaxPool2d(kernel_size=(2,2)),
      spectral_norm(Conv2d(f*2, f*4, 3, 1, 1)),
      nn.LeakyReLU(0.1, inplace=True),
      nn.MaxPool2d(kernel_size=(2,2)),
      spectral_norm(Conv2d(f*4, f*8, 3, 1, 1)),
      nn.LeakyReLU(0.1, inplace=True),
      nn.MaxPool2d(kernel_size=(2,2)),
      spectral_norm(Conv2d(f*8, 1, 3, 1, 1)),
      nn.Sigmoid()
    )
```

Definition: conditional GAN

GANs can be conditioned with labels $\mathbf{y}$ if available [7] by feeding the label information into both the generator and the discriminator:

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

Link to Colab example



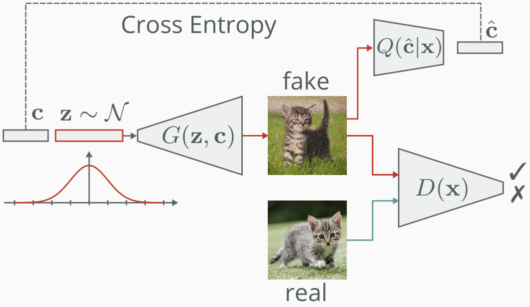
Definition: information maximizing GANs

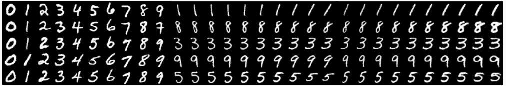
GANs can be trained to learn disentangled latent representations in a completely unsupervised manner. InfoGAN [8] popularised this by maximizing mutual information between the observation and a subset of the latents:

$$\min_{G,Q} \max_{D} V_{\text{InfoGAN}}(D, G, Q) = V(D, G) - \lambda L_I(G, Q)$$

where $L_I(G, Q)$ is a variational lower bound of the mutual information.

Link to Colab example

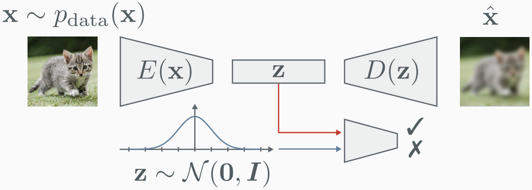




Definition: adversarial autoencoders

Adversarial autoencoders [9] are generative models that permit sampling.

In addition to the reconstruction loss, such $\|\mathbf{x} - \hat{\mathbf{x}}\|^2$, they use adversarial training to match the aggregated posterior of the hidden code vector $\mathbf{z}$ of the autoencoder with an arbitrary prior distribution, such as $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.

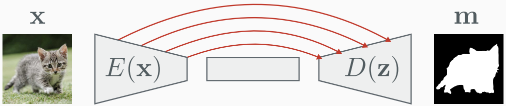


Definition: skip connections (U-Net)

Skip connections (U-Net) is a popular residual approach used for paired image translation tasks [10]. For example for images $\mathbf{x}$ and paired masks $\mathbf{m}$, where: $\mathcal{L} = \mathbb{E}_{\mathbf{x},\mathbf{m}\sim p_{\text{data}}}\left[\|U(\mathbf{x}) - \mathbf{m}\|^2\right]$

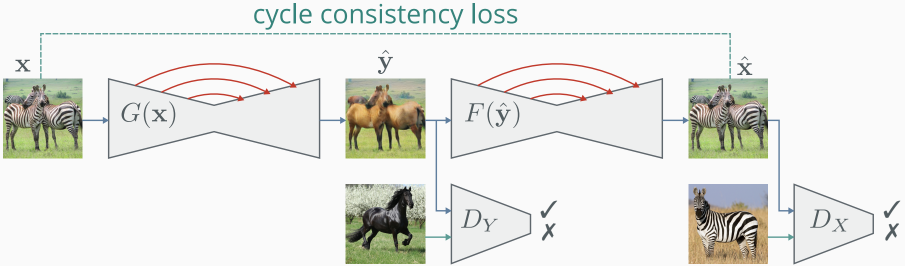
Link to Colab example

Note: U-Net is not an adversarial method, but the use of skip connections is popular in many papers, so now is a good time to introduce it.



Definition: unpaired translation (CycleGAN)

CycleGAN [11] propose an adversarial architecture that enables unpaired image translation. It has twin residual generators and two discriminators, which translate between the domains, alongside a cycle consistency loss (an L1 norm) which ensures the mapping can recover the original image.

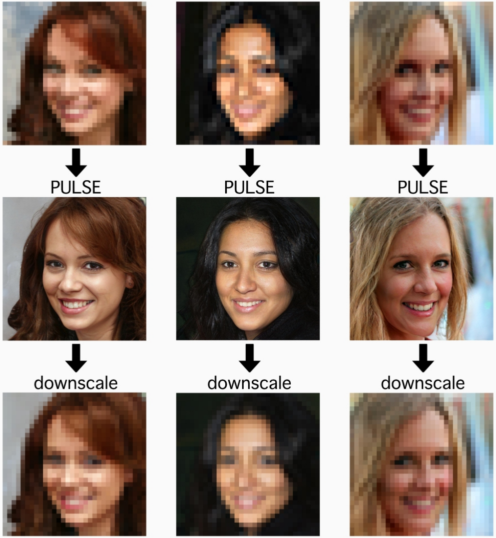


Definition: super-resolution

Adversarial models are popular in super-resolution approaches. The challenge is that a single low-resolution (LR) input can map to a distribution of high-resolution (HR) outputs.

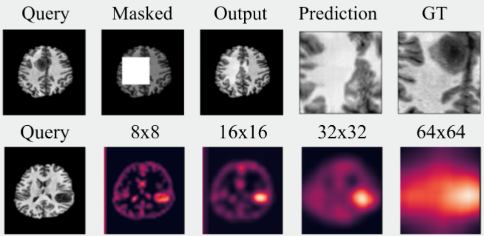
PULSE [12] investigates this by projecting points in the search of the latent space of StyleGAN (a large conditional GAN) onto a hypersphere, which ensures probable outputs in the high-dimensional latent space.

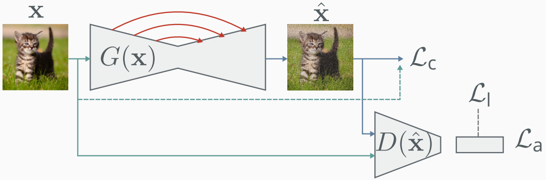
Online example



Definition: anomaly detection

Unsupervised anomaly detectors [13] learn a normal distribution over (healthy) observations. Then, when they observe something not observed in training (unhealthy/dangerous), they fail to reconstruct - detecting it as an anomaly. Region-based anomaly detectors [14] learn a distribution over inpainted (erased) regions.





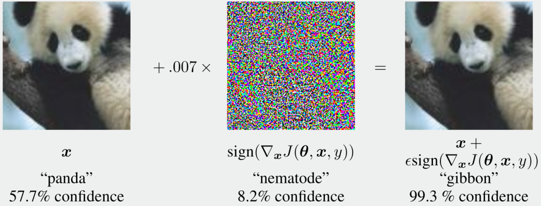
Adversarial examples attacks

Definition: adversarial examples

These are small but intentionally worst-case perturbations that fool the model to output incorrect answers with high confidence [15]. It is possible to generate examples that also fool the human visual system [16]. Cat or dog?

Example: adversarial examples

Example adding an imperceptibly small vector by the sign of the elements of the gradient of the cost function with respect to the input [15]:



Adversarial examples defences

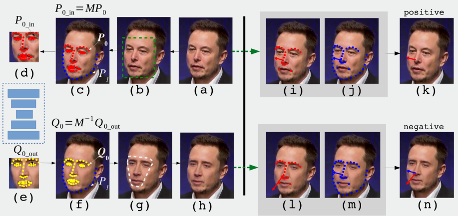
Definition: adversarial defence

There are several defence strategies that introduce the adversarial examples into training [15]. A popular approach uses U-Net to denoise and reduce the amplification of the adversarial perturbations [17].

Black-box adversarial defence is where an adversary can only monitor the outputs of the model. White-box methods are more difficult, as an adversary has access to the model allowing for specific attacks. White-box defence generally overfits to the attack used during training.

Example: adversarial defences

Question: What is the behaviour at the limit of the adversarial generative model arms-race? Who wins at convergence?



References I

[1] Ian Goodfellow et al. “Generative adversarial nets”. In: Advances in neural information processing systems. 2014, pp. 2672–2680.

[2] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. “Differentiable augmentation for data-efficient gan training”. In: arXiv preprint arXiv:2006.10738 (2020).

[3] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. “Unrolled generative adversarial networks”. In: arXiv preprint arXiv:1611.02163 (2016).

[4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. “Wasserstein GAN”. In: arXiv preprint arXiv:1701.07875 (2017).

[5] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. “Improved training of Wasserstein GANs”. In: Advances in neural information processing systems. 2017, pp. 5767–5777.

[6] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. “Spectral normalization for generative adversarial networks”. In: arXiv preprint arXiv:1802.05957 (2018).

References II

[7] Mehdi Mirza and Simon Osindero. "Conditional generative adversarial nets". In: arXiv preprint arXiv:1411.1784 (2014).

[8] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets". In: Advances in neural information processing systems. 2016, pp. 2172–2180.

[9] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. "Adversarial autoencoders". In: arXiv preprint arXiv:1511.05644 (2015).

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional networks for biomedical image segmentation". In: International Conf on Medical image comp and comp-assisted intervention. Springer. 2015, pp. 234–241.

References III

[11] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. “Unpaired image-to-image translation using cycle-consistent adversarial networks”. In: Proceedings of the IEEE international conference on computer vision. 2017, pp. 2223–2232.

[12] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. “PULSE: Self-supervised photo upsampling via latent space exploration of generative models”. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, pp. 2437–2445.

[13] Samet Akçay, Amir Atapour-Abarghouei, and Toby P Breckon. “Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection”. In: 2019 International Joint Conference on Neural Networks (IJCNN). IEEE. 2019, pp. 1–8.

[14] Bao Nguyen, Adam Feldman, Sarath Bethapudi, Andrew Jennings, and Chris G Willcocks. “Unsupervised Region-based Anomaly Detection in Brain MRI with Adversarial Image Inpainting”. In: arXiv preprint arXiv:2010.01942 (2020).

References IV

[15] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples". In: arXiv preprint arXiv:1412.6572 (2014).

[16] Gamaleldin Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. "Adversarial examples that fool both computer vision and time-limited humans". In: Advances in Neural Information Processing Systems. 2018, pp. 3910–3920.

[17] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. "Defense against adversarial attacks using high-level representation guided denoiser". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, pp. 1778–1787.

[18] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. Available online, MIT press. 2016.