GENERATIVE ADVERSARIAL NETWORKS

INF5860 — Machine Learning for Image Analysis

Ole-Johan Skrede

02.05.2018

University of Oslo

Outline

- Repetition
- Generative Adversarial Networks
- Other adversarial methods

REPETITION

Autoencoders

- An autoencoder f consist of an encoder g and an decoder h
- The encoder maps the input x to some representation z

$$g(x) = z$$

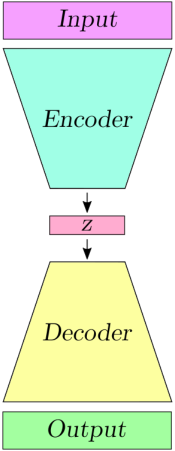
- We often call this representation z for the code or the latent vector
- The decoder maps this representation z to some output $\hat{x}$

$$g(z) = \hat{x}$$

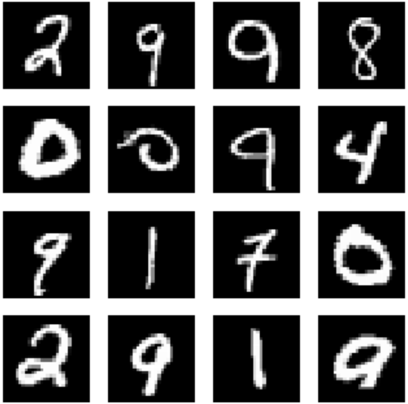
- We want to train the encoder and decoder such that

$$f(x) = h(g(x)) = \hat{x} \approx x$$

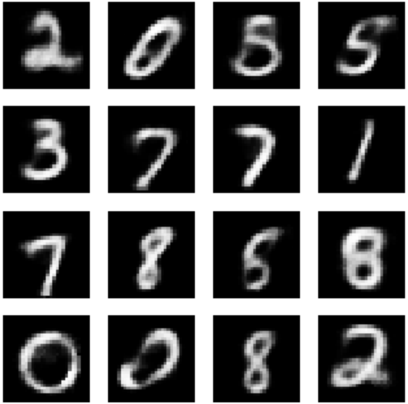
- Commonly used for compression, feature extraction and de-noising



Compression autoencoder — MNIST example

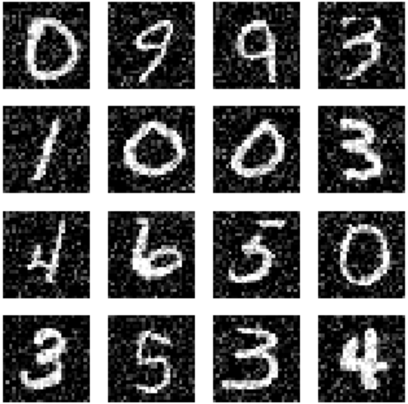


(a) Original

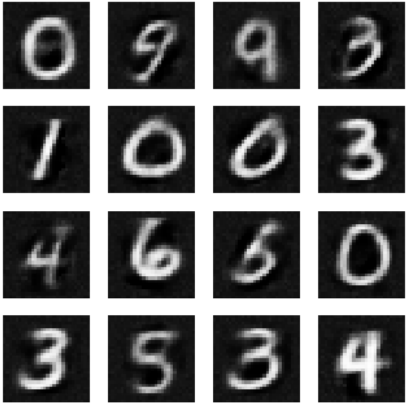


(b) Reconstructed

De-noising autoencoder — MNIST example



(a) Original

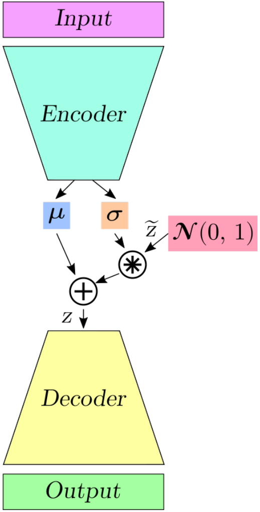


(b) Reconstructed

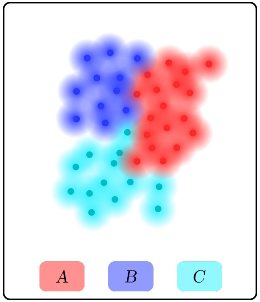
- A variational autoencoder is designed to have a continuous latent space
- This makes them ideal for random sampling and interpolation
- It achieve this by forcing the encoder g to generate Gaussian representations, $z \sim \mathcal{N}(\mu, \sigma^2)$
- More precisely, for one input, the encoder generates a mean μ and a variance σ^2
- We sample then sample a zero-mean, unit-variance Gaussian $\tilde{z} \sim \mathcal{N}(0, 1)$
- Construct the input z to the decoder from this

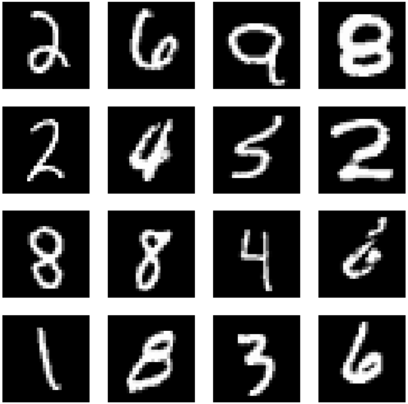
$$z = \mu + \tilde{z}\sigma^2$$

- With this, z is sampled from $q = \mathcal{N}(\mu, \sigma^2)$

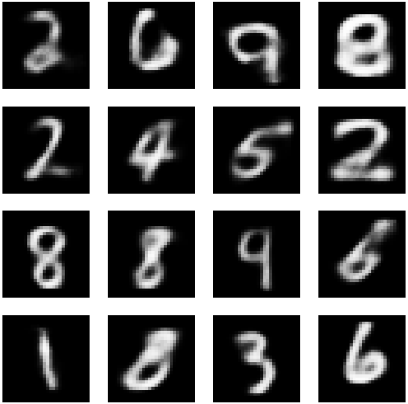


- This is a stochastic sampling
- That is, we can sample different z from the same set of (μ, σ^2)
- The intuition is that the decoder "learns" that for a given input x:
  - the point z is important for reconstruction
  - but also a neighbourhood of z
- In this way, we have smoothed the latent space, at least locally
- In the previous lecture, we learnt ways to achieve this





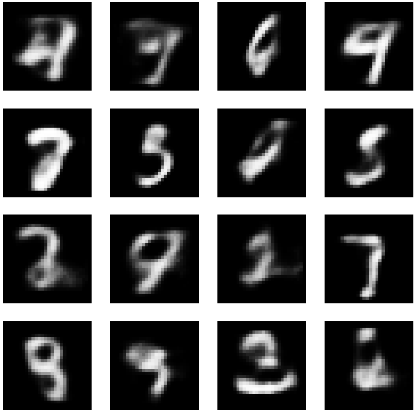
(a) Original



(b) Reconstructed

VAE example: generation of new signals

- Sample a random latent vector z from $\mathcal{N}(0, 1)$
- Decode z



VAE example: interpolation between samples

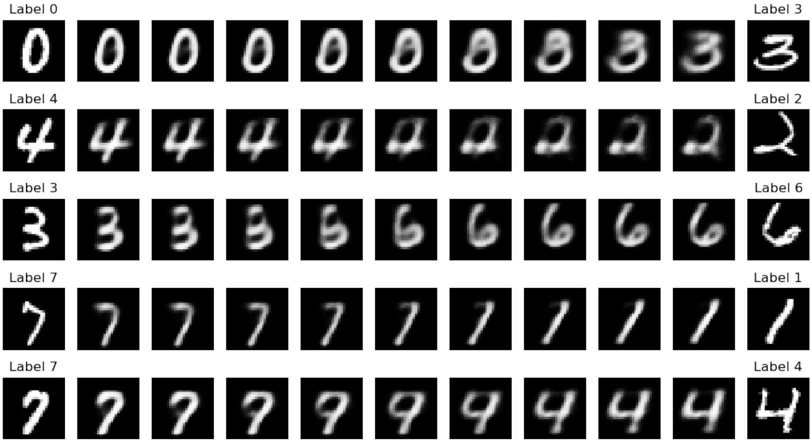
- We generate a signal c that is an interpolation between two signals a and b
- We can do this by a linear interpolation between the means

$$\mu_{c_k} = (1 - w_k)\mu_a + w_k\mu_b$$

where the different interpolation weights can be

$$w_k = \frac{k}{n+1}, \quad k = 1, \dots, n$$

VAE example: interpolation between samples



GENERATIVE MODELLING

- We have training samples from an unknown distribution p_{data}
- We want a model that can draw samples from some distribution p_{model}
- p_{model} should be an estimate of p_{data}
- A model that can sample from this p_{model} is termed a *generative model*
- For brevity, we will refer to the distributions as $p_d = p_{\mathrm{data}}$, and $p_m = p_{\mathrm{model}}$.

Overview of generative models

- Some models explicitly estimates p_m
- Some models implicitly estimates p_m by only drawing samples from it
- Some models is able to do both
- VAE explicitly approximates p_m
- GAN only samples from p_m[1]

[1] There are GAN variants that are able to do both

Different approaches to generate samples from a distribution

- In the maximum likelihood case, we often have an explicit distribution $p_\theta(x)$, and for some fixed, observed data $\{x_i\}_{i=1}^m$, we find the parameters θ^* that maximizes the likelihood

$$\theta^* = \arg\max_\theta \prod_{i=1}^m p_\theta(x_i) \tag{1}$$

- In the implicit case, we have a data distribution p_d and some generator distribution p_g
- The random variable $Z \sim p_g$ are transformed via some function to $X \sim p_m$
- This parametric function $f(x;\theta)$ can be a neural network, and the parameters θ are adjusted such that the model distribution is close to the data distribution $p_m \approx p_d$.

Motivation: why study generative modelling

- Analyse our ability to represent and manipulate high-dimensional distributions (e.g. images)
- Can be used as a tool in reinforcement learning
- Can be used in semi-supervised learning where labelled data is scarce
- Sampling of realistic examples from some high-dimensional distribution can have many applications

Application — Predicting the next frame

- A model is trained to predict the next frame in a video sequence
- There exists many possible modes (high probability events)
- A standard mean-square error model tends to predict some average of the possible futures
- A GAN model is able to select one of the possible futures, which results in a more sharp prediction

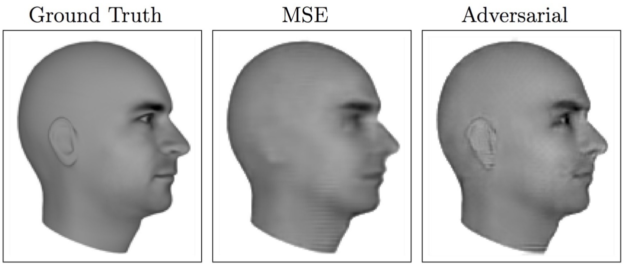


Figure 4: Source: [Goodfellow, 2016]

Application — Image super resolution

- Generating high-resolution images from low-resolution inputs
- GANs tend to produce perceptually pleasing and sharp results

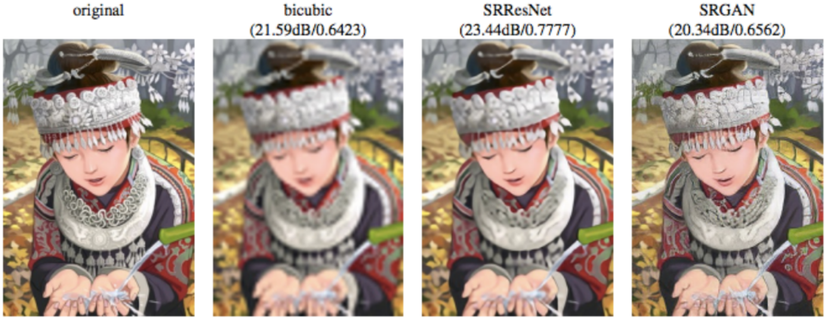


Figure 5: Source: [Goodfellow, 2016]

Figure 6: Source: [Demir and Unal, 2018]

Figure 7: Source: [Elgammal et al., 2017]

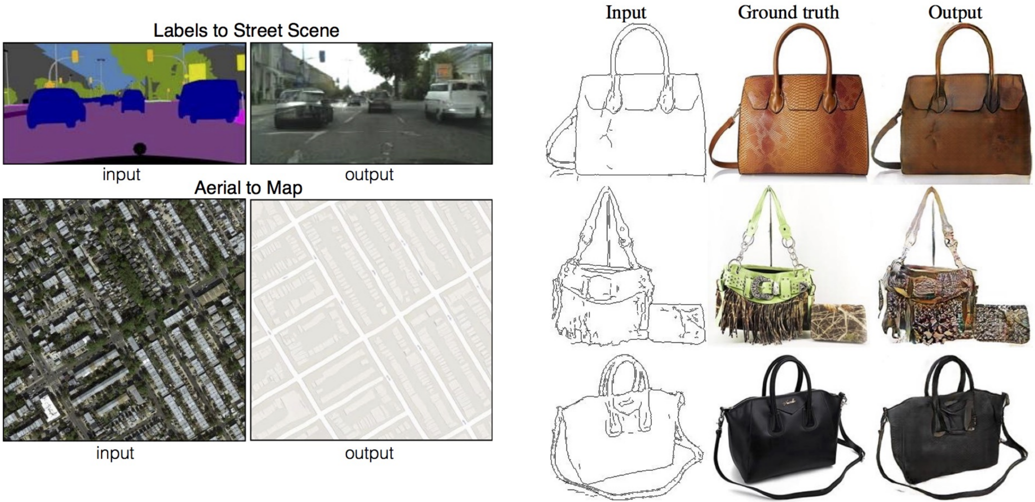


Figure 8: Source: [Goodfellow, 2016]

GENERATIVE ADVERSARIAL NETWORKS

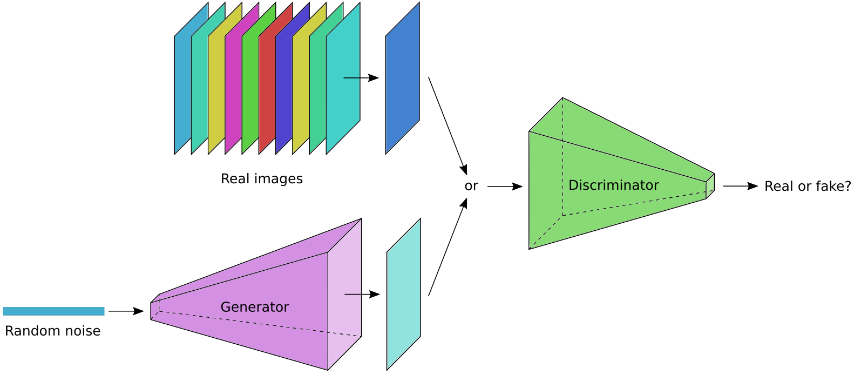
- General introduction
- Cost functions
- Challenges
- Tips and tricks

Figure 9: Source: https://deephunt.in/the-gan-zoo-79597dc8c347

- Introduced by Ian Goodfellow et al. in 2014 [Goodfellow et al., 2014]
- General idea from game theory
- Analogy
  - Counterfeiter creating fake money
  - Police trying to distinguish fake money from real money
  - The better the counterfeiter gets, the better the police gets
  - The better the police gets, the better the counterfeiter gets
- Yann LeCun dubbed adversarial training the most interesting idea in ML the last 10 years

Components

- A *generator* function that tries to create real-looking examples
- A *discriminator* function that tries to distinguish real from fake examples
- Functions are updated in a feedback loop, making each better at its task



Components

- The discriminator is a function

$$D : x \mapsto D(x; \theta_D)$$

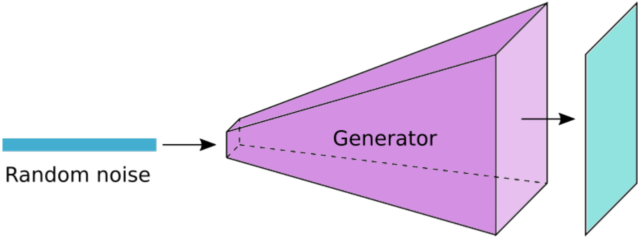
  mapping input x to $D(x; \theta_D)$ with parameters θ_D
- The generator is a function

$$G : z \mapsto G(z; \theta_G)$$

  mapping input z to $G(z; \theta_G)$ with parameters θ_G
- The discriminator has an associated loss $J_D(\theta_D, \theta_G)$, depending on both θ_D and θ_G, but can only control θ_D
- The generator has an associated loss $J_G(\theta_D, \theta_G)$, depending on both θ_D and θ_G, but can only control θ_G
- The optimal solution (θ_D^*, θ_G^*) is a *Nash equilibrium* where
  - θ_D^* is a local minimum of J_D w.r.t. θ_D
  - θ_G^* is a local minimum of J_G w.r.t. θ_G

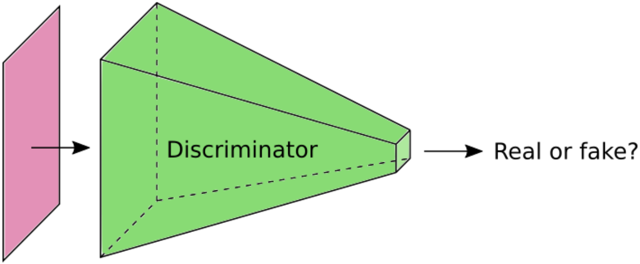
The generator

- The generator is a differentiable function
- The input z is a random vector sampled from some simple prior distribution p_g
- The output $x = G(z)$ is then sampled from p_m
- The most common form of G is some kind of generative neural network
- If we have GAN trained on data from p_d, we can use the generator to sample from p_m
- $p_m \approx p_d$
- With this, samples from the generator will look like the training data



The discriminator

- The discriminator is a standard classification network
- Trained to differentiate between real and fake (generated) images
- Outputs a single number in $[0, 1]$
  - $D(x) = 0 \rightarrow D$ believes x is fake
  - $D(x) = 1 \rightarrow D$ believes x is real



The training process

- At each update step, one mini-batch x of real images, and one mini-batch z of latent vectors are drawn
- z is fed through G, producing $G(z)$
- $D(x)$ is compared with $D(G(z))$
- θ_G is updated using gradients from J_G
- θ_D is updated using gradients from J_D
- The discriminator and generator are updated in tandem using some regular optimization routine (SGD, Adam, etc.)
- Some flexibility with regards to updating one more often than the other

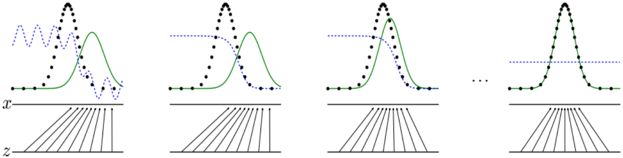


Figure 10: Source: [Goodfellow et al., 2014]

The training process

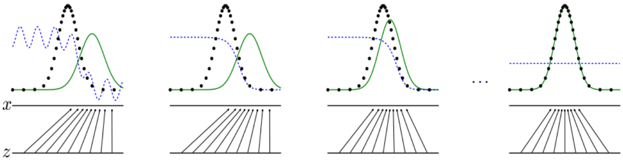


Figure 11: Source: [Goodfellow et al., 2014]

- Black arrows illustrate the mapping $z \mapsto G(z)$
- Black probability density is the data distribution p_d
- Blue probability density is the discriminator distribution
- Green probability density is the generative distribution p_m
- The generative distribution distinguishes between real and generated data
- From (a) to (d): The generative distribution (green) is guided towards high probable areas of the discriminative distribution (blue)
- The process terminates when the discriminative distribution becomes constant (is no longer able to distinguish real from fake)

The discriminator cost function

- The generator, G, and discriminator, D, are two distinct networks with distinct cost functions
- The cost functions are optimized separately
- The discriminator cost function is given by

$$\begin{aligned} J_D(\theta_D, \theta_G) &= -E_{x\sim p_d}[\log D(x;\theta_D)] - E_{z\sim p_g}[\log(1 - D(G(z;\theta_G);\theta_D))] \\ &= -E_{x\sim p_d}[\log D(x;\theta_D)] - E_{x\sim p_m}[\log(1 - D(x;\theta_D))] \end{aligned}$$

- With discrete samples, over one mini-batch $\{x_i\}$ and $\{z_i\}$, this becomes

$$J_D(\theta_D, \theta_G) = -\frac{1}{m}\sum_{i=1}^{m}[\log(D(x_i;\theta_D)) + \log(1 - D(G(z_i;\theta_G);\theta_D))]$$

- Binary classification with sigmoid cross entropy where
  - Real images are given label 0
  - Generated (fake) images are given label 1

The generator cost function

- Could in principle use the negative discriminator cost

$$J_G(\theta_D, \theta_G) = -J_D(\theta_D, \theta_G)$$

- The generator is not dependent on p_d, so the loss becomes

$$J_G(\theta_D, \theta_G) = \frac{1}{m} \sum_{i=1}^{m} \log(1 - D(G(z_i; \theta_G); \theta_D))$$

- The generator is trained to minimize the probability that the discriminator classifies its generated examples as fake
- Could then summarize the entire training process as a zero-sum game

$$(\theta_D^*, \theta_G^*) = \arg\min_{\theta_G} \max_{\theta_D} V(\theta_D, \theta_G)$$

  with the *value function* $V(D, G) = -J_D(\theta_D, \theta_G)$
- Rephrasing of the discriminator cost: Find a discriminator that maximizes the probability of assigning the correct label to real and fake examples

Theoretical insights

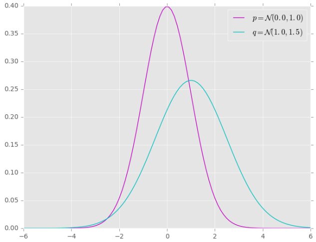
- This is generator objective formulation has some problems that we will come back to later
- It is a view that has convenient theoretical properties
- Before we return to a more useful generator loss, we are going to analyse this result
- Outline:
    - KL-divergence vs. JS-divergence
    - A closer look at the discriminator cost function
    - Consequences

- We are comparing the distributions p_X and q_X over some discrete random variable X
- The Kullback-Leibler (KL) divergence is given by

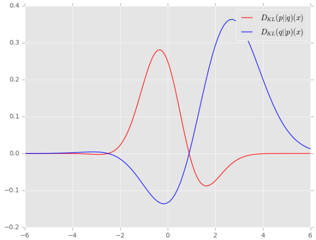
$$D_{KL}(p_X||q_X) = \sum_x p_X(x) \log \frac{p_X(x)}{q_X(x)}$$

- This is an asymmetric distance metric, meaning that, *in general*

$$p_X \neq q_X \rightarrow D_{KL}(p_X||q_X) \neq D_{KL}(q_X||p_X)$$



(a) Two unequal distributions as a function of x



(b) KL-divergence kernel as a function of x

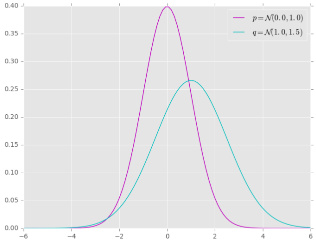
- Let p_X and q_X be as above, and let their mixture be

$$g_X = \frac{1}{2}(p_X + q_X)$$

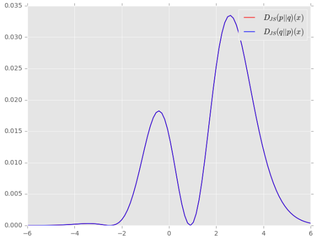
- The Jensen-Shannon (JS) divergence is then given by

$$D_{JS}(p_X||q_X) = \frac{1}{2}D_{KL}(p_X||g_X) + \frac{1}{2}D_{KL}(q_X||g_X)$$

- This is a symmetrized and smoothed version of the KL divergence



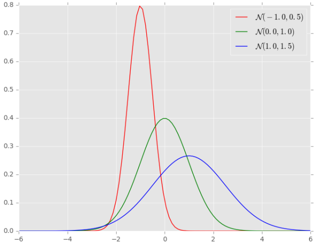
(a) Two unequal distributions as a function of x



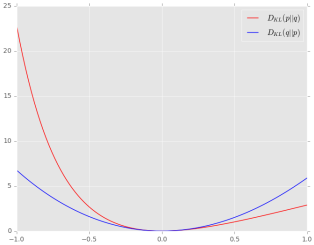
(b) JS-divergence kernel as a function of x

Comparison of distributions

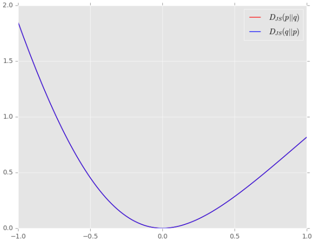
- In these figure, the KL-divergences and JS-divergences are computed for a range of distribution comparisons
- The reference distribution is $p = \mathcal{N}(0.0, 1.0)$
- The comparison distributions is $q = \mathcal{N}(\mu, \sigma^2)$ with, simultaneously
  - μ ranging from -1.0 to 1.0
  - σ^2 ranging from 0.5 to 1.5



(a) Range of normal distributions



(b) KL-divergence of range



(c) JS-divergence of range

Optimal discriminator loss

- What value of $D(x)$ is maximizing the value function?

$$\begin{aligned} V(D,G) &= -J_D(\theta_D,\theta_G) \\ &= \int_x p_d(x)\log(D(x,\theta_D)) + p_m(x)\log(1-D(x,\theta_D))\,\mathrm{d}x \\ &= \int_x \tilde{V}(D,G)(x)\,\mathrm{d}x \end{aligned}$$

where $\tilde{V}(D,G)(x)$ is the integrand.

- From variational calculus, we have that (the functional derivative is)

$$\begin{aligned} \frac{\mathrm{d}V(D,G)}{\mathrm{d}D(x)} &= \frac{\mathrm{d}\tilde{V}(D,G)}{\mathrm{d}D(x)} \\ &= \left[p_d(x)\frac{1}{\ln 10}\frac{1}{D(x)} - p_m(x)\frac{1}{\ln 10}\frac{1}{1-D(x)}\right] \\ &= \frac{1}{\ln 10}\left[\frac{p_d(x)}{D(x)} - \frac{p_m(x)}{1-D(x)}\right] \end{aligned}$$

Optimal discriminator loss

- Equating the derivative with zero yields the optimal discriminator value

$$
\begin{aligned}
0 &= \frac{\mathrm{d}V(D,G)}{\mathrm{d}D(x)} \\
&= \frac{1}{\ln 10}\left[\frac{p_d(x)}{D^*(x)} - \frac{p_m(x)}{1 - D^*(x)}\right] \\
D^*(x) &= \frac{p_d(x)}{p_d(x) + p_m(x)}
\end{aligned}
$$

- Moreover, when the generator is working optimally $p_m = p_d$, and therefore

$$D^*(x) = \frac{1}{2}$$

Optimal discriminator loss

- Inserting the optimal generator $G^*(x)$, and discriminator $D^*(x) = \frac{1}{2}$, back into the value function, we get

$$\begin{aligned} V(D^*, G^*) &= \int_x p_d(x) \log \frac{1}{2} + p_m(x) \log \frac{1}{2} \, \mathrm{d}x \\ &= \log \frac{1}{2} \left[\int_x p_d(x) + p_m(x) \, \mathrm{d}x \right] \\ &= 2 \log \frac{1}{2} \\ &= -2 \log 2 \end{aligned}$$

- To be clear: this is the value of the discriminator loss when using the discriminator that minimizes the loss, and the generator that samples from the (approximate) data distribution

Optimal discriminator loss and relation to Jensen Shannon divergence

- If we analyze the JS divergence

$$
\begin{aligned}
D_{JS}(p_X||q_X) &= \frac{1}{2} D_{KL}(p_X||\frac{1}{2}(p_X + q_X)) + \frac{1}{2} D_{KL}(q_X||\frac{1}{2}(p_X + q_X)) \\
&= \frac{1}{2}\left(\int_x p_d(x)\log(2\frac{p_d}{p_d+p_m})\,\mathrm{d}x + \int_x p_m(x)\log(2\frac{p_m}{p_d+p_m})\,\mathrm{d}x\right) \\
&= \frac{1}{2}\left(\log 2 + \int_x p_d(x)\log\frac{p_d}{p_d+p_m}\,\mathrm{d}x + \log 2 + \int_x p_m(x)\log\frac{p_m}{p_d+p_m}\,\mathrm{d}x\right) \\
&= \frac{1}{2}\left(2\log 2 + \int_x p_d(x)\log D^*(x)\,\mathrm{d}x + \int_x p_m(x)\log(1 - D^*(x))\,\mathrm{d}x\right) \\
&= \frac{1}{2}\left(2\log 2 + V(D^*, G)\right)
\end{aligned}
$$

Optimal discriminator loss and relation to Jensen Shannon divergence

- From the result on the previous slide, we get an expression for the discriminator loss with an optimal discriminator

$$V(D^*, G) = 2D_{JS}(p_d||p_m) - 2\log 2$$

- From this we see that minimizing the value function given an optimal discriminator is equivalent to optimizing the JS-divergence
- Also note that the optimal generator gives $p_d = p_m$, and therefore $V(D^*, G^*)$

The generator cost function

- In the discriminator, we are minimizing the cross-entropy between the target distribution and the generated distribution
- It has strong gradients when the classifier is wrong
- The gradients saturates when the classifier is right, but this is not as important
- For the generator objective, we have found a candidate with convenient theoretical interpretations
- Unfit in practice: When the discriminator successfully rejects generated examples with high confidence, the gradients of the generator loss vanishes
- We must find a generator loss that does not saturate at unwanted places

The generator cost function

- For the generator cost, we propose the following

$$
\begin{aligned}
J_G(\theta_D, \theta_G) &= -E_{z\sim p_g} \log D(G(z;\theta_G);\theta_D) \\
&= -\frac{1}{m}\sum_{i=1}^{m} \log D(G(z_i;\theta_G);\theta_D)
\end{aligned}
$$

- With this, the generator maximizes the log-probability of the discriminator being mistaken (assigning label 1 to the generated examples)
- Contrast this with the previous minimax game where we the generator minimizes the log-probability of the discriminator being correct (assigning label 0 to the generated examples)
- Both the generator and the discriminator now have strong gradients when they are "losing the game"

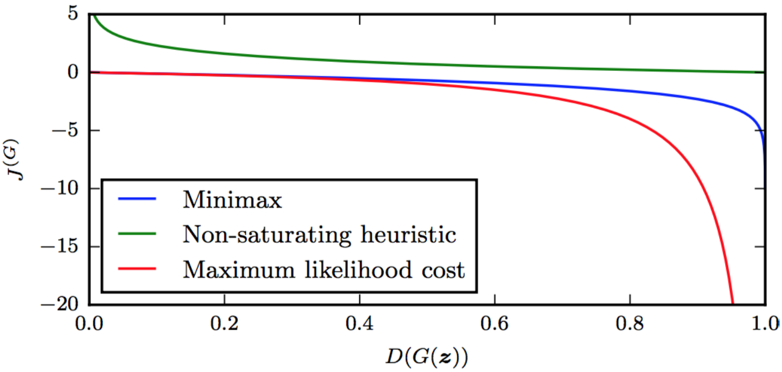


Figure 15: Graph of the gradient cost w.r.t. discriminator classification. Source: [Goodfellow, 2016]

Cost functions overview

- Minimizing the discriminative cost

$$J_D(\theta_D, \theta_G) = -\frac{1}{m}\sum_{i=1}^{m}[\log(D(x_i;\theta_D)) + \log(1 - D(G(z_i;\theta_G);\theta$$

"pushes" $D(x)$ to 1 (real class) and $D(G(z))$ to 0 (fake class)

- Minimizing the generative cost

$$J_G(\theta_D, \theta_G) = -\frac{1}{m}\sum_{i=1}^{m}\log(D(G(z_i;\theta_G)\theta_D))$$

"pushes" $D(G(z))$ to 1 (real class)

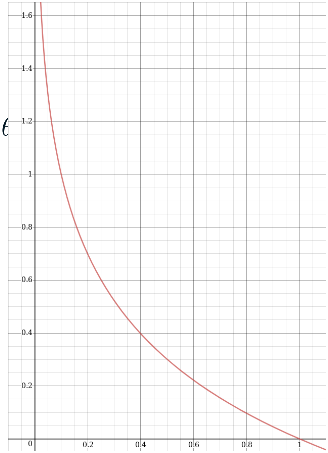


Figure 16: Graph of $f(x) = -\log x$

GAN advantages

- Tend to produce sharper examples than other generative models
- The reason was thought to be the relationship to the JS-divergence
- This view is not supported now
- It is not entirely clear why GANs tend to produce sharper images

GAN CHALLENGES

- Convergence
- Performance evaluation
- Discrete output

Non-convergence of GAN

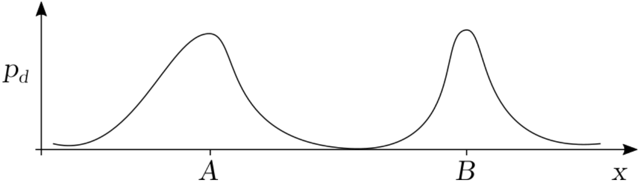
- Achieving convergence is in general difficult
- The solutions tends to oscillate
- This is connected to that one try to achieve an equilibrium in stead of a plain optimization
- The major problem is connected to what is called *mode collapse*

Mode-collapse

- A peak in the probability density is called a mode
- Real-world data tends to be *multi-model*
- This means that similar examples are clustered in separate locations
- The data probability distribution will have peaks (modes) at these locations
- Mode-collapse is the phenomenon where the generator tends to generate very similar examples
- These similar examples originates from roughly the same location in the model distribution
- This location has high probability in the data distribution.

Mode-collapse — example

- Suppose we have a dataset with two modes, around A and B
- You want the GAN to generate examples estimating the training data, from both A and B
- Mode collapse can be described as follows
  1. The generator produces examples close to A which "fools" the discriminator
  2. The discriminator classifies x from B as real with high probability, x from A are classified $50/50$ as real or fake
  3. The generator is then driven to produce examples from B
  4. The discriminator counters, and classifies examples x from A as real and examples from B real or fake with 50% probability
  5. This cycle then repeats from 1.



Mode-collapse — example

- Some rationale can be that when the generator should be the solution to

$$\min_{\theta_G} \max_{\theta_D} V(D, G)$$

there seems to be difficult to guarantee that it is not the solution to

$$\max_{\theta_D} \min_{\theta_G} V(D, G)$$

which would explain the mode collapse
- Partial mode collapse is more common than complete mode collapse
- Generated images then tend to have the same colors, or some of the same features
- See the figure below for another example

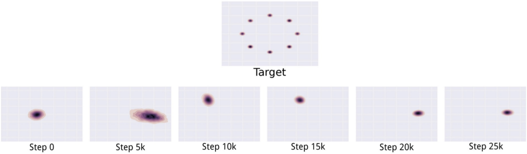


Figure 18: Mode collapse on data of a mixture of gaussians. Source: [Metz et al., 2016]

Performance evaluation

- Results from generative models can be hard to quantify and evaluate
- Often, in terms of images, perceptual similarity is important
- Other generative models may have an explicit objective function
- GANs lack this, which makes it even harder
- [Salimans et al., 2016] discusses this:
- Human evaluation using Amazon Mechanical Turk
  - Subjective
  - Work intensive
  - Overly pessimistic when thought
- Automatic evaluation using a classifier to produce conditional distributions $p(y|x)$
  - Examples with a clear class should have $p(y|x)$ with low entropy
  - A generative model should produce varied results: $\int p(y|x = g(z))\,\mathrm{d}z$ should have high entropy

GAN tricks and advice

- We will present some useful tricks for GAN
- Some are related to preventing the mode collapse problem
- See [Salimans et al., 2016] and [Goodfellow, 2016] for a more thorough discussion

Minibatch discrimination

- The discriminator in a standard GAN compares single examples
- The idea is to aid this comparison with information from the whole mini-batch of real and generated examples
- The rationale is that the discriminator can detect if one example is unusually similar to other generated examples
- This technique is shown to work quite well

Feature matching

- This is related to the minibatch discrimination
- Also attempts addressing the mode-collapse problem by increasing diversity
- Extends (or replaces) the discriminator loss with a comparison of intermediate features from both the real and generated data
- In stead of explicitly discriminating on the output, we also discriminate on hidden layers

Train with labels

- If you have a labeled training set, use the labels
- If you have K classes, add the fake data as class $K + 1$
- The discriminator now tries to classify examples as one of $K + 1$ classes
- This improves the perceptual quality of generated examples
- This technique can be used in semi-supervised learning

One-sided label smoothing

- Neural network classifiers tend to classify with too high confidence
- We can encourage the discriminator to produce more soft predictions
- Set the true label for the real samples to be 0.9 in stead of 1
- This penalizes models producing too large logits on real samples
- Important to not smooth the generated sample label

Batch norm

- Batch normalization in GAN is, in general, very useful
- Batch normalization is not ideal for small batch sizes as the mean and variance varies too much between batches
- This is problematic for GANs as these fluctuations can dominate over the latent variable z in the generator (see figure below)
- Reference batch norm and virtual batch norm can aid this [Goodfellow, 2016]

Figure 19: GAN on ImageNet. Source: [Goodfellow, 2016]

Decent looking examples

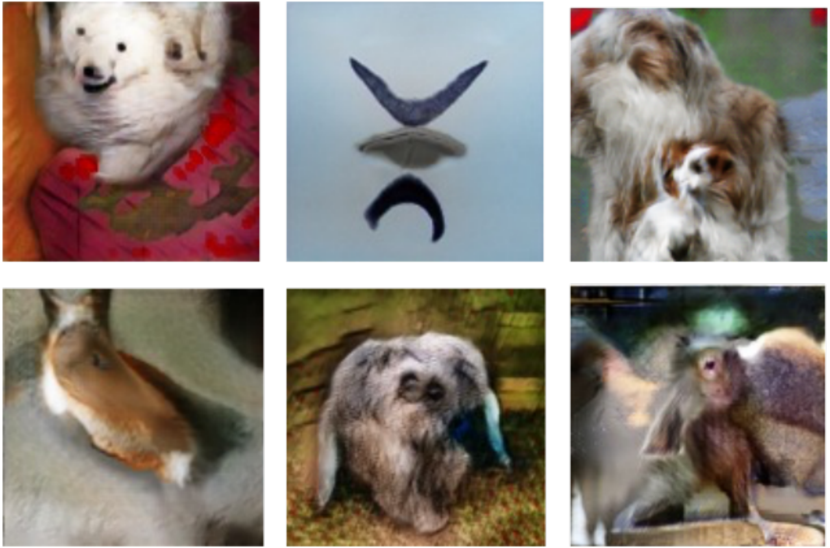


Figure 20: GAN on ImageNet. Source: [Goodfellow, 2016]

Problems with counting

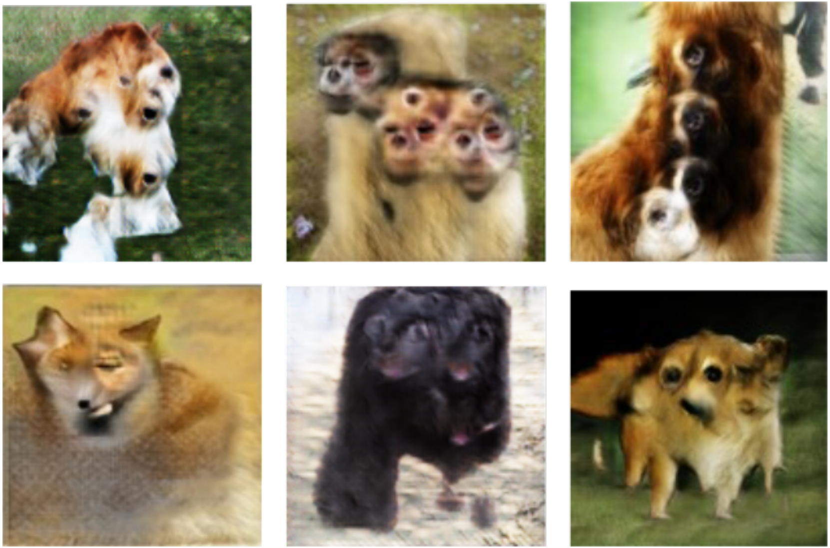


Figure 21: GAN on ImageNet. Source: [Goodfellow, 2016]

Problems with 3D (one of these are real)

Figure 22: GAN on ImageNet. Source: [Goodfellow, 2016]

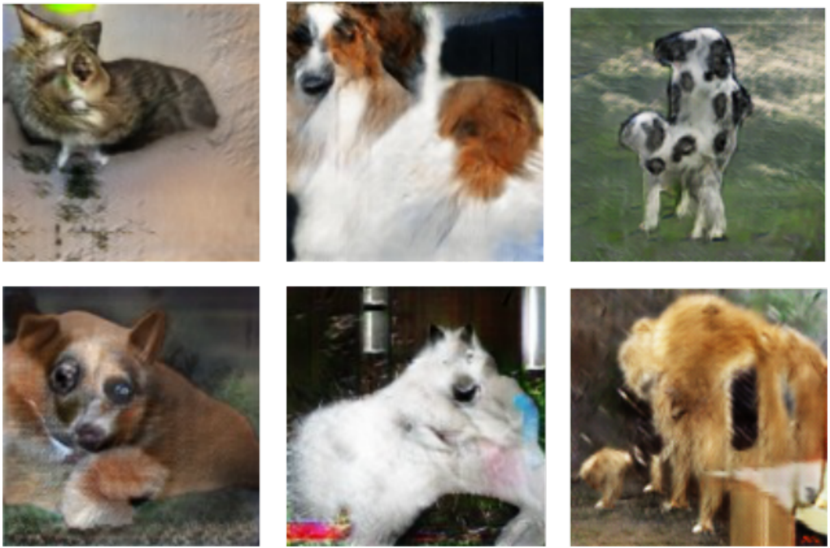


Figure 23: GAN on ImageNet. Source: [Goodfellow, 2016]

NOTABLE GAN VARIANTS

DCGAN

- GANs have gained a lot of interest
- For an impression of the amount of models, take a look at this post: `https://deephunt.in/the-gan-zoo-79597dc8c347`
- We are only going to look briefly at two architectures:
  - DCGAN
  - WGAN
- Both have been selected because of their generality and popularity

DCGAN

- *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks* [Radford et al., 2016]
- Wants to learn good intermediate image representations from unlabeled data
- VAEs and standard GANs produces generates blurry images
- GANs are difficult to train and can generate non-sensical results
- DCGAN enables the coupling of CNNs with GANs

DCGAN — INTRODUCED CHANGES

- Uses techniques from the (then) resent lessons learned
- Replaces deterministic spatial pooling operations (such as maxpool) with learned spatial up- and down-sampling
    - strided convolutions for the discriminator
    - fractionally strided convolutions (transposed convolutions) for the generator
- Elimination of dense layers on top of the convolutional layers at the end of the networks
- Use batch-normalization between layers in both the generator and discriminator
- Use ReLU activation in the generator (except in the output layer, which uses tanh)
- Use leaky ReLU activation in the discriminator

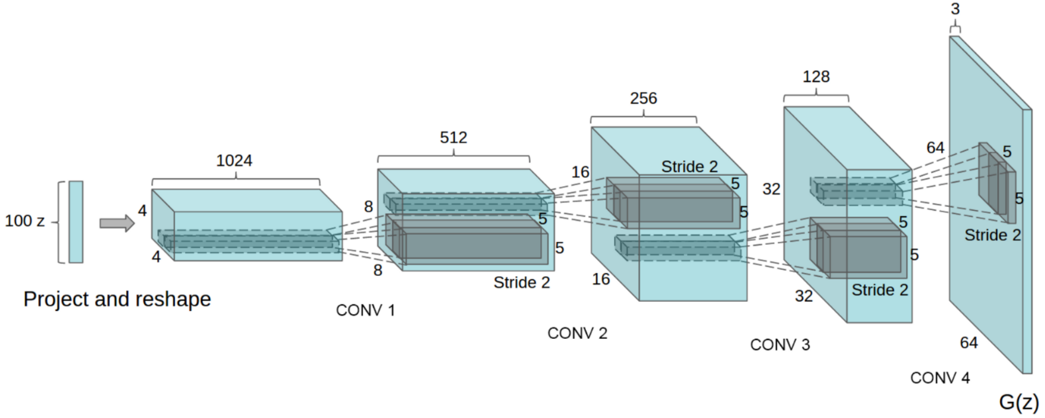


Figure 24: DCGAN generator architecture. Source: [Radford et al., 2016]

DCGAN — TRAINING DETAILS

- Image values are scaled to $[-1, 1]$
- Adam optimizer with momentum 0.5
- Learning rate of 2×10^{-4}
- Mini-batch size of 128
- Initialize weights from a zero-mean normal distribution with standard deviation 0.02

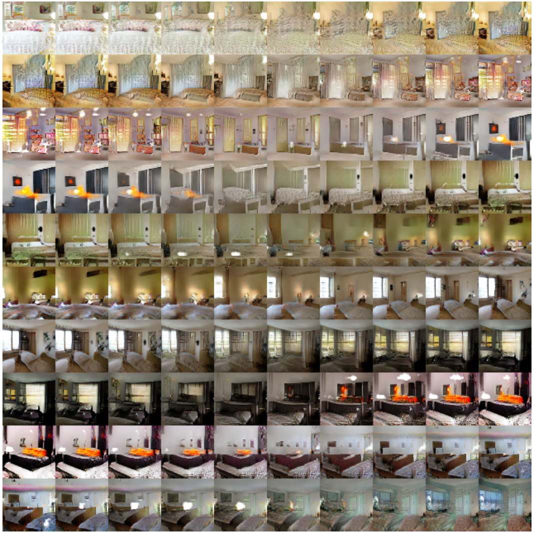


Figure 25: Bedroom interpolation

DCGAN — Results

Figure 26: Faces looking left to faces looking right

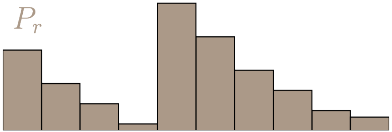
WGAN — Wasserstein Generative Adversarial Network

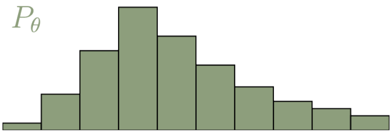
- Introduced in 2017, [Arjovsky et al., 2017]
- Claims to solve, or reduce many of the problems with training GANs
- Is based on the Wasserstein distribution similarity metric
- Has become quite popular with > 500 citations and > 1700 github stars in little over a year
- It is quite technical, so we will only look at the wasserstein distance

Wasserstein distance

- Also known as *Earth Mover Distance*
- Intuitively easy to grasp
- Quite complicated to derive, compute, and fully understand
- We will only concern ourself with the intuition
- For more details, I refer to `https://vincentherrmann.github.io/blog/wasserstein/`,
- The figures for this section are from the above resource

- It measures the smallest amount of "work" that needs to be done in order to transform one distribution to the other.
- Let our distributions be P_r and P_θ

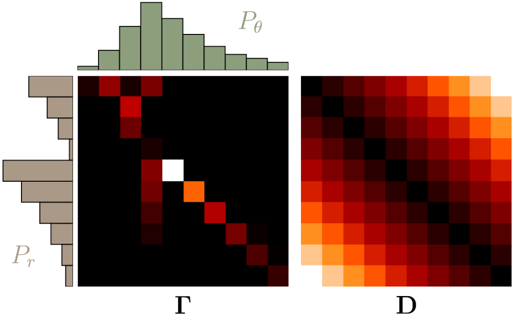


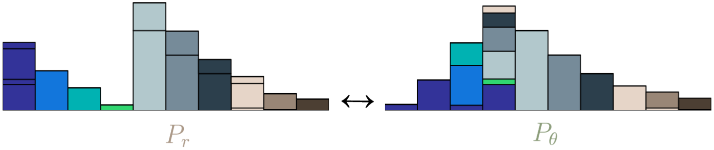


- Let $\gamma(x, y)$ be the difference between $P_r(x)$ and $P_\theta(y)$
- The Wasserstein distance is then

$$W(P_r, P_\theta) = \inf_{\gamma \in \Gamma} \sum_{x,y} ||x - y|| \gamma(x, y)$$

- Here Γ contains all "valid" γ (details are not important here)
- inf means *infimum* and can be thought of as the greatest lower bound





WGAN — MOTIVATING EXAMPLE

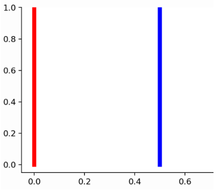


Figure 28: Two point distributions

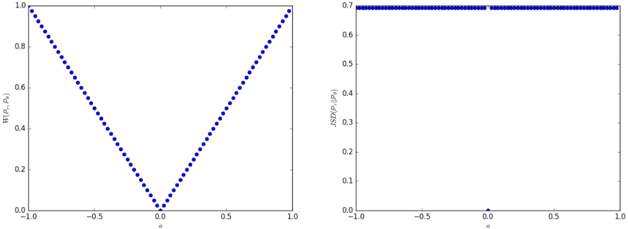


Figure 29: Wasserstein distance (left) and JS-divergence (right) when the above two distributions come closer, overlap, and then move away from each other again. Source: [Arjovsky et al., 2017]

ADVERSARIAL DOMAIN ADAPTATION

Introduction

- Generalization of results on training to the real world is crucial for a useful method
- This can be hard enough when training and test comes from the same distribution
- Even worse when the train and test data comes from different distributions, known as *domain shift* or *dataset bias*
- Domain adaptation methods addresses this problem
- Adversarial domain adaptation methods uses principles from GANs
- In essence, they try to train models that are invariant to the dataset domain by trying to fool a discriminator that tries to classify domains

Case study: Adversarial discriminative domain adaption

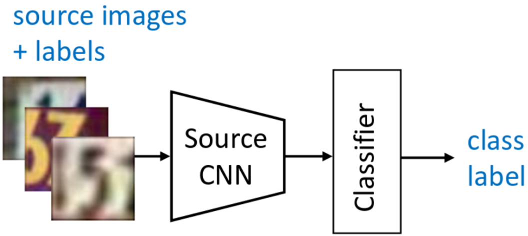
- There are several approaches to this problem
- Notable works are e.g.
    - Gradient reversal [Ganin et al., 2016]
    - Domain confusion [Tzeng et al., 2015]
    - CoGAN [Liu and Tuzel, 2016]
- The adversarial discriminative domain adaption (ADDA) is illustrated because of its simplicity and performance

ADDA: OVERVIEW

- We have labeled data (x_s, y_s) for the source domain
- The target domain data, y_t is unlabeled
- We are going to learn a source mapping $M_s : x_s \mapsto y_s$
- We are also going to learn a target mapping $M_t : x_t \mapsto y_t$
- The target mapping should be invariant to the domain difference between the source and the target
- We are going to use a discriminator with an associated loss to learn this domain invariance

ADDA: Example mappings

- For the source mapping M_s, we can use a standard classification network with cross-entropy loss
- The target mapping M_t is equal to M_s, except for the classifier part, but with separate and independent parameters
- The parameters of M_t are initialized with the parameters of a trained M_s
- M_s is fixed when M_t is trained



ADDA: Discriminator

- The discriminator D should classify outputs of these networks as either originating from the source or the target domain
- This is a similar situation as with regular GANs, and the loss is

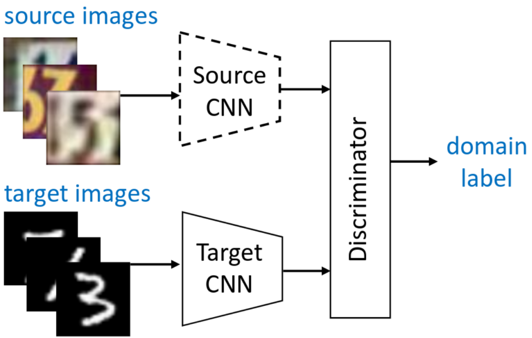
$$J_D(M_s, M_t) = -E_s\left[\log D(M_s(x_s))\right] - E_t\left[\log(1 - D(M_t(x_t)))\right]$$

- The discriminator wants to maximize the probability that it predicts the correct domain
- The target mapping should produce examples that maximizes the probability of being classified as coming from the source
- We therefore chose the generator loss from GANs

$$J_M(M_s, M_t) = -E_t\left[\log D(M_t(x_t))\right]$$

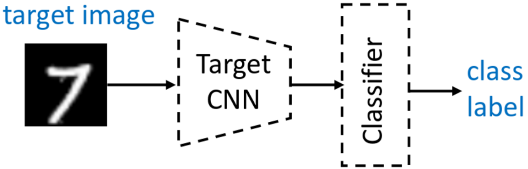
- E_d is the expectation over examples in $d \in \{\text{source}, \text{target}\}$

ADDA: Discriminator



ADDA: Testing

- We now have a classifier that can classify examples from features
- We also have a base mapping M_t that should generate domain-invariant features
- We reuse those parts in the testing



Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. *arXiv preprint*, 2017. URL https://arxiv.org/pdf/1701.07875.pdf.

Ugur Demir and Gozde Unal. Patch-Based Image Inpainting with Generative Adversarial Networks. *arXiv preprint*, 2018. URL http://arxiv.org/abs/1803.07422.

Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. CAN: Creative Adversarial Networks, Generating "Art" by Learning About Styles and Deviating from Style Norms. *arXiv preprint*, (Iccc):1–22, 2017. doi: 10.1089/cyber.2017.29084.csi. URL http://arxiv.org/abs/1706.07068.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Francois Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 2016.

Ian Goodfellow. NIPS 2016 Tutorial: Generative Adversarial Networks. *arXiv preprint*, 2016. ISSN 0253-0465. doi: 10.1001/jamainternmed.2016.8245. URL http://arxiv.org/abs/1701.00160.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *arXiv preprint*, pages 1–9, 2014. ISSN 10495258. doi: 10.1001/jamainternmed.2016.8245. URL http://arxiv.org/abs/1406.2661.

Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. *CoRR*, 2016.

Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled Generative Adversarial Networks. *arXiv preprint*, 2016. URL https://arxiv.org/pdf/1611.02163.pdf.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *ICLR 2016*, pages 1–16, 2016. URL http://arxiv.org/abs/1511.06434.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs. *arXiv preprint*, 2016. doi: arXiv:1504.01391. URL https://arxiv.org/pdf/1606.03498.pdf.

Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. *International Conference in Computer Vision (ICCV)*, 2015.

Questions?