CIS 522: Lecture 5R

Autoencoders
02/13/20

Penn Engineering

Are you here?

Yes

No

Puppy

Feedback / Logistics

- Lecture Notes
- Puppy as option instead of kitten
- HW2 - Due tonight at midnight!
- HW3 - Computer Vision
    - Groups of 2
- Final Project Abstract Proposal (Due Tuesday 2/18)

Generative Models

Cognitive science

Imagine a kitten

Make it black

Make it play with a ball

Like this?

What is a Generative Model?

Discriminative Model: Given a set of classes C, and a sample X, what is the probability X belongs to class c?

Succinctly represented as this quantity: $P(c|X)$ (called a **class probability estimate)**

To generate a classification: $\text{argmax}_{c \in C} P(c|X)$

Examples:

1. Logistic Regression
2. Decision Trees
3. Support Vector Machines

What is a Generative Model?

Generative Model: Given a class c from a set of classes C, and a sample X, what is the probability that X was generated from that class?

Succinctly represented as this distribution: $P(X|c)$

What is a Generative Model?

Generative Model: Given a class c from a set of classes C, and a sample X, what is the probability that X was generated from that class?

Succinctly represented as this distribution: $P(X|c)$

Now to generate an X, we can simply randomly sample from this distribution!

What is a Generative Model?

Generative Model: Given a class c from a set of classes C, and a sample X, what is the probability that X was generated from that class?

Succinctly represented as this distribution: $P(X|c)$

Now to generate an X, we can simply randomly sample from this distribution!

Examples

- Naive Bayes
- **Variational Autoencoder (To be covered in this lecture)**
- Generative Adversarial Networks (To be covered in ***next*** lecture!)

Low-dimensional representations

Cute puppy… but takes ~ **half a million numbers to represent**

Artificial vs. biological representations

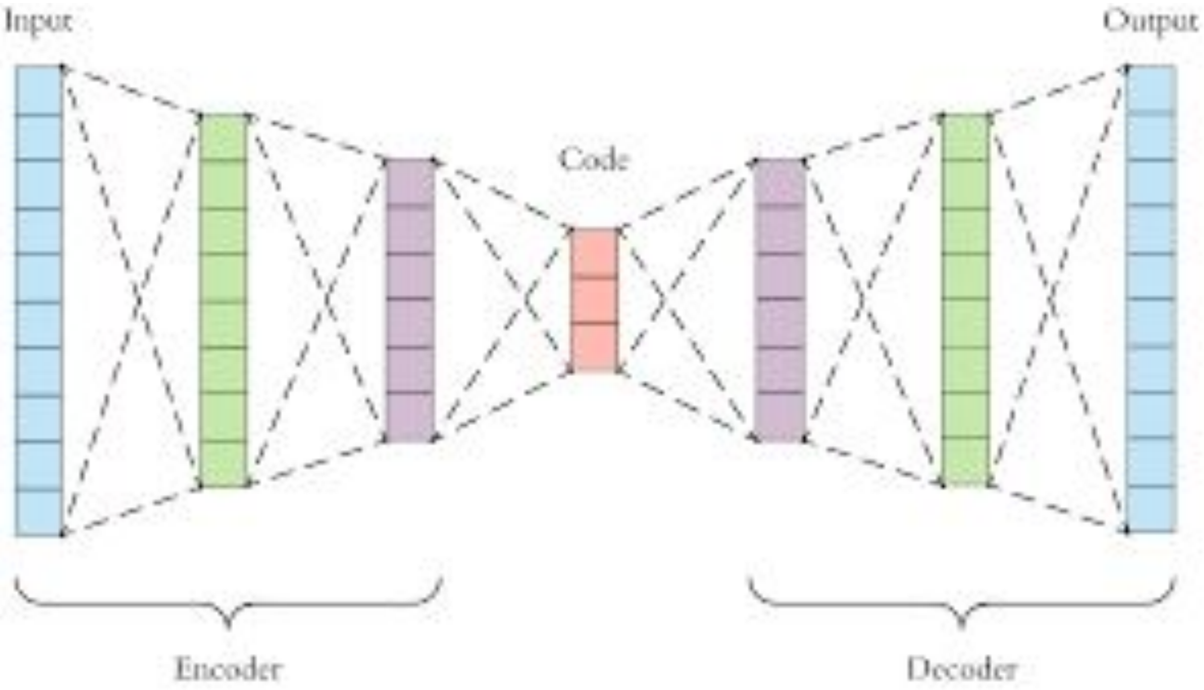
X 172k =

How can we represent a puppy with less numbers?

Autoencoders

Penn Engineering

Basic architecture



Training

- “Self-supervised” training: desired output Y is just input X
- e.g. MSE loss
- Small bottleneck layer impedes training

X

y-hat

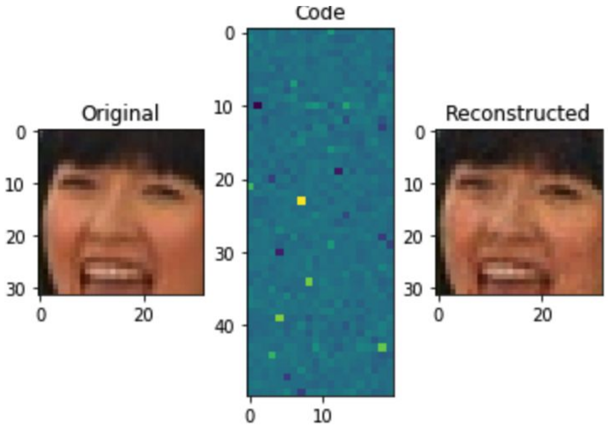
What is a linear autoencoder?

- What happens if no nonlinearity?
- Linear encoder and decoder
- If L2 on encoder and decoder, exactly PCA (Kunin et al. 2019)

Convolutional Autoencoders

- Decoder has to “undo” the convolution / pooling
- Generally involves upsampling or “deconvolutions”
- Also conv layers reducing # of features

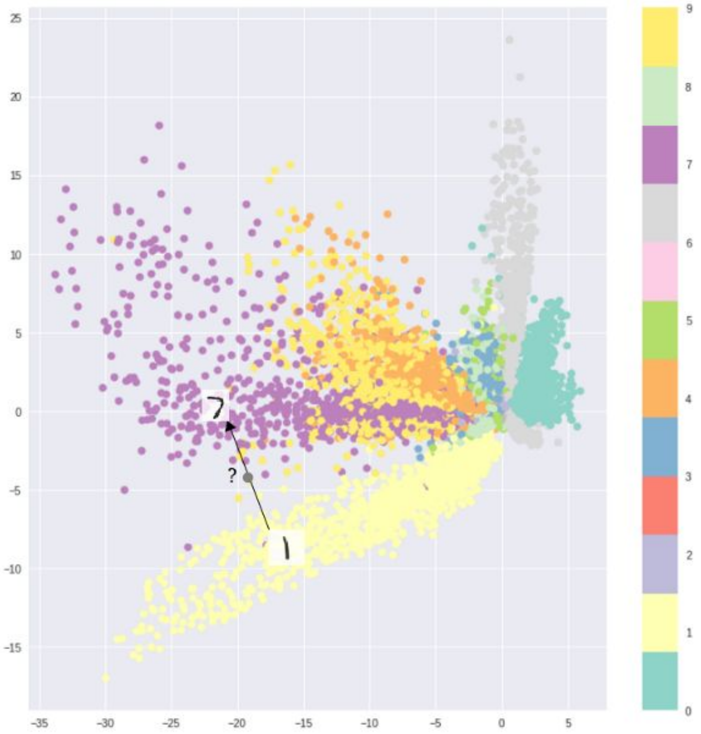
CAE:



Source: https://stackabuse.com/autoencoders-for-image-reconstruction-in-python-and-keras/

Issue with Convolutional Autoencoders

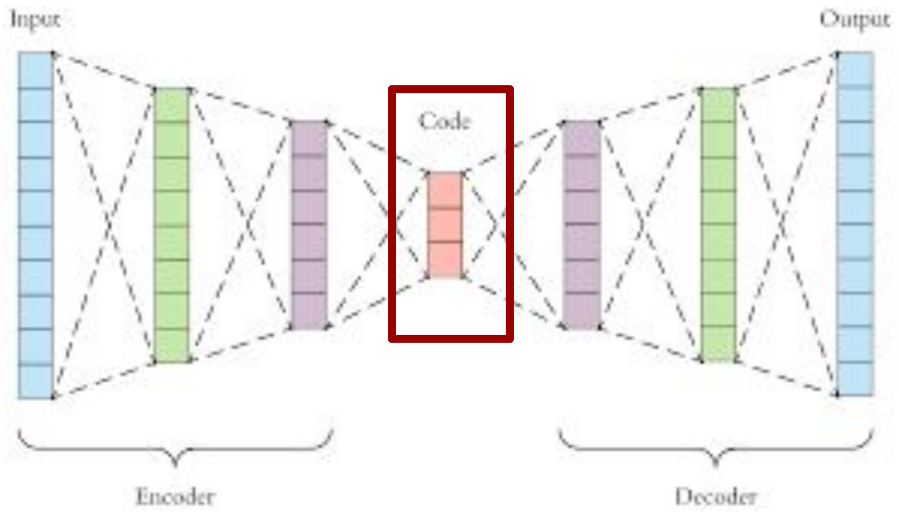
- Space forms clusters
- Clusters are far apart
- What if we want to interpolate between classes?



Source: https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf

Constraining the Features

- Currently no constraints on the features.
- What if we constrained the features to the standard normal distribution?

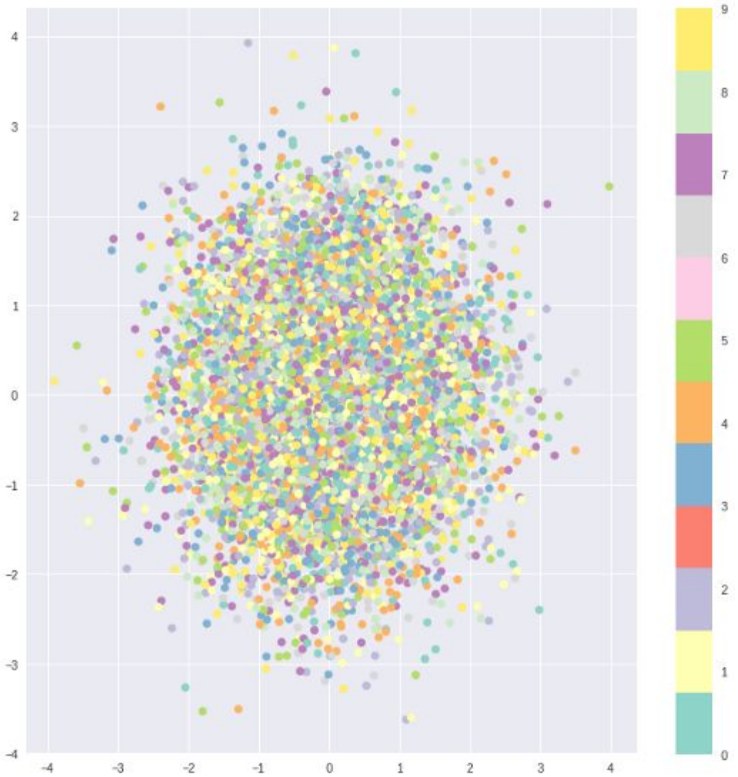


Source: https://towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-understanding-kl-divergence-2b382ca2b2a8

KL Divergence Loss

$$D_{KL}(p||q) = \sum_{i=1}^{N} p(x_i) log(\frac{p(x_i)}{q(x_i)})$$

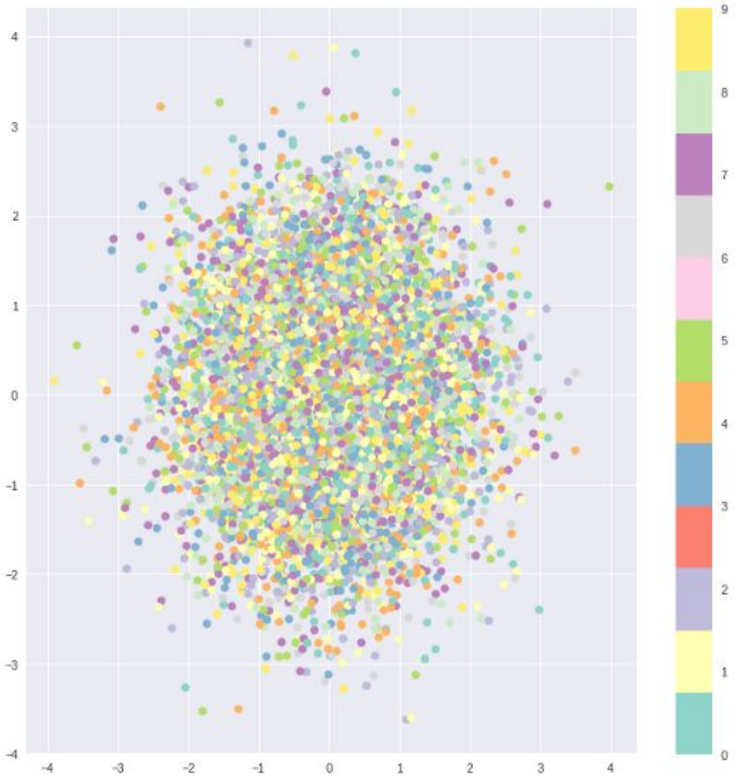
Only using the KL Divergence Loss



Source: https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf

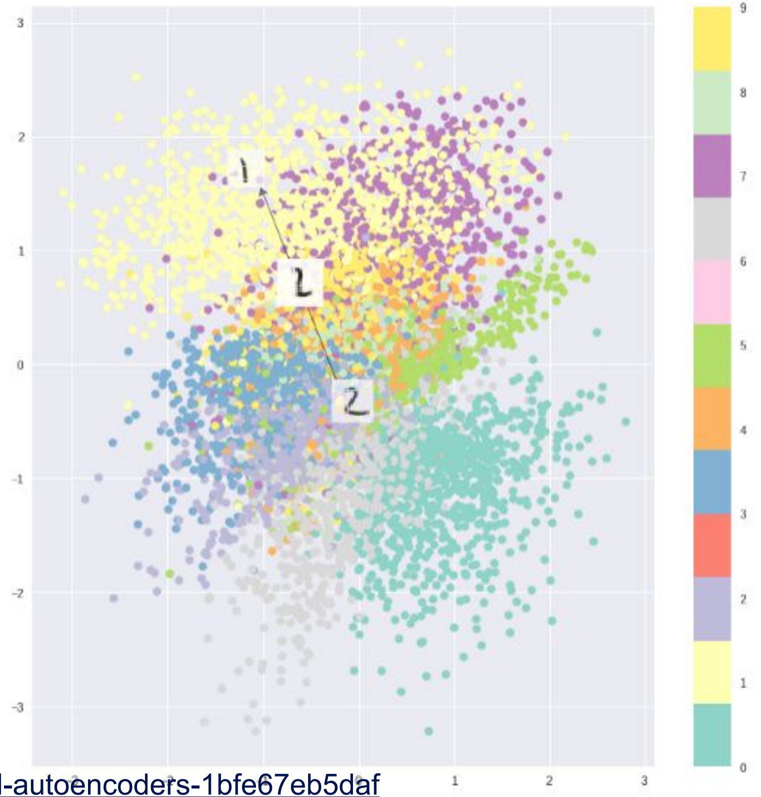
Only using the KL Divergence Loss

… clearly can’t only use a KL Divergence Loss



Source: https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf

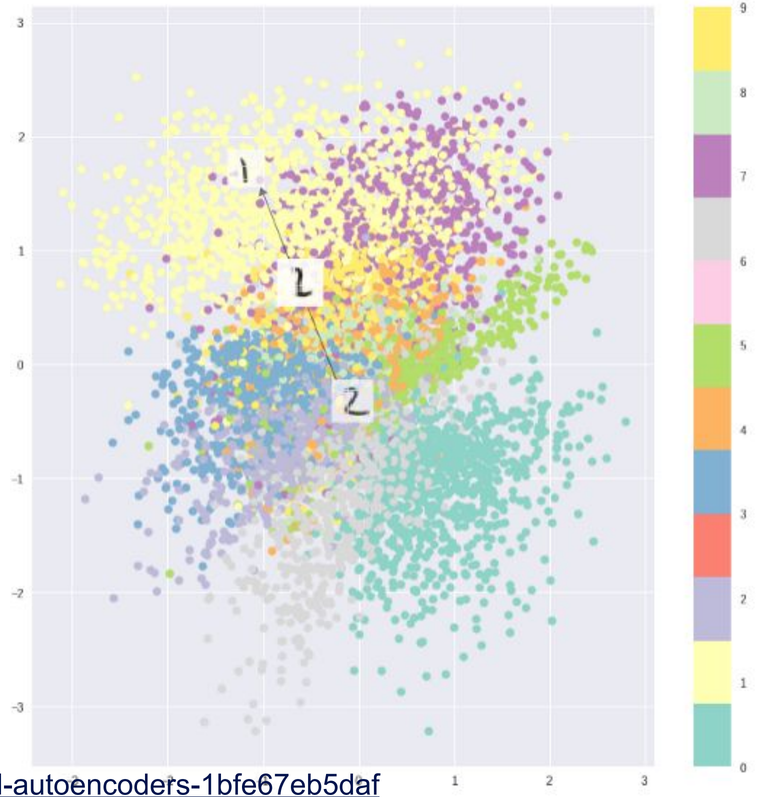
KL Divergence + Reconstruction Loss?



Source: https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf

KL Divergence + Reconstruction Loss?

Much better!



Source: https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf

KL Divergence + Reconstruction Loss?

$$\text{argmin}_{q,r}\left(KL(q(z|x)\,||\,p(z)) + \mathbb{E}_{z\sim q(z|x)}MSE(x, r(z))\right)$$

Regularization

Reconstruction loss

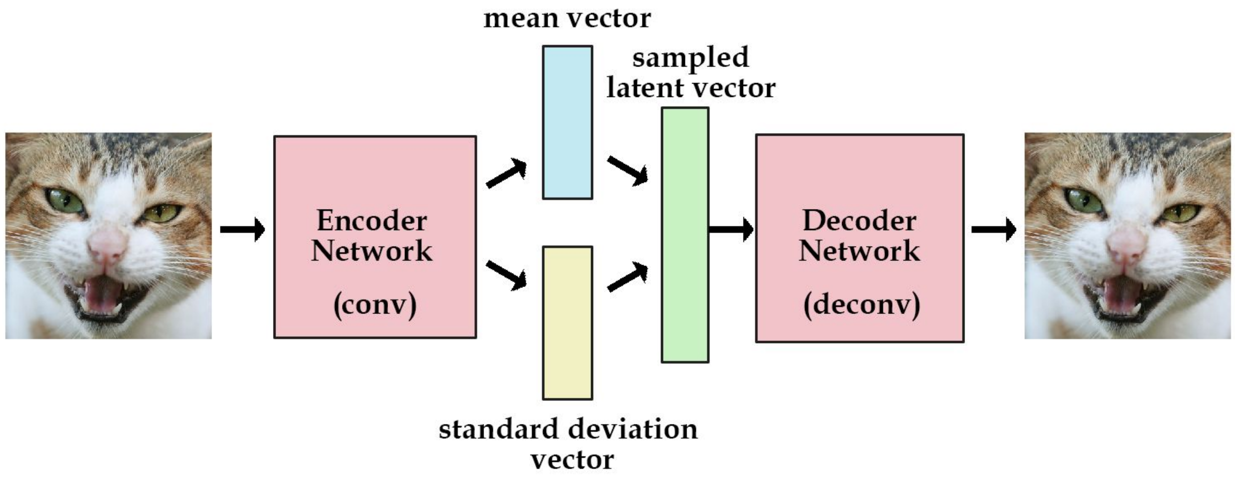
Variational autoencoders (VAEs) - motivation

- Want the low-dim representation z to have independent components
- Each component should contribute a similar amount
- z according to unit Gaussian

Example:

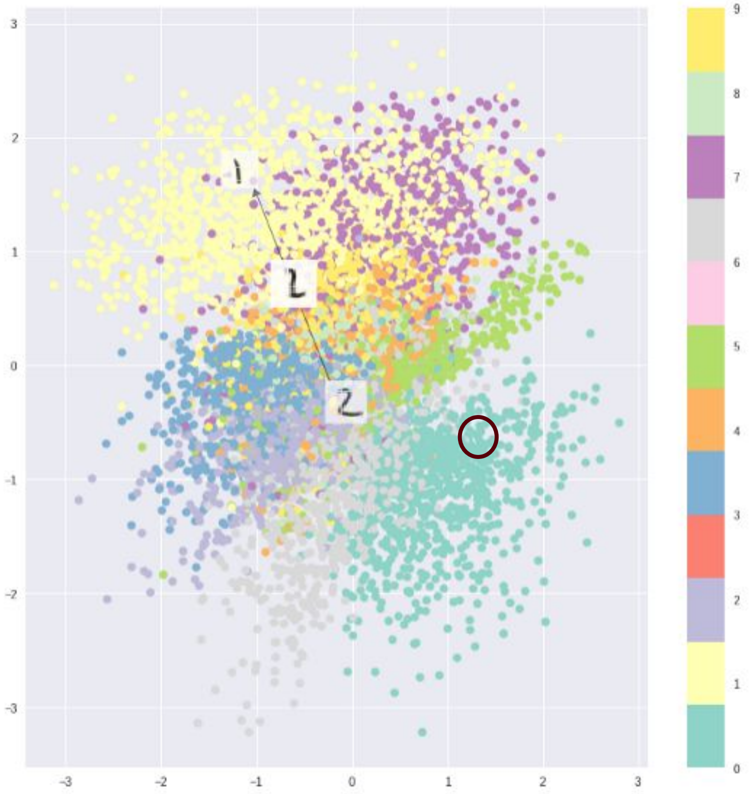
- Low-dim representation of faces - eye color, hair length, etc.
- Each feature should be independent
- Scale so each feature has same variance

VAE Process



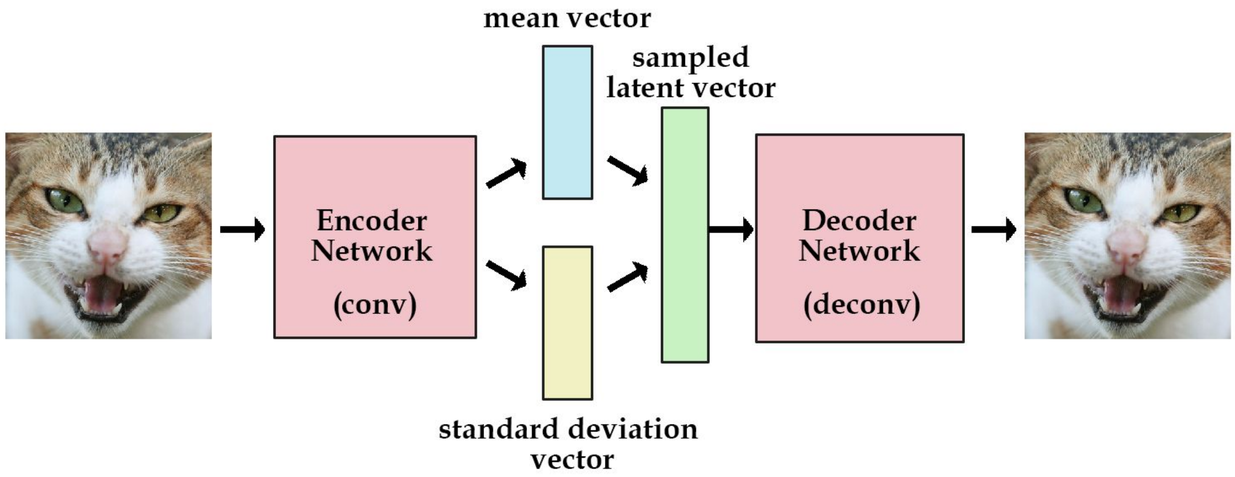
Source: http://kvfrans.com/variational-autoencoders-explained/

Why sample a latent vector?



Source: https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf

VAE Process



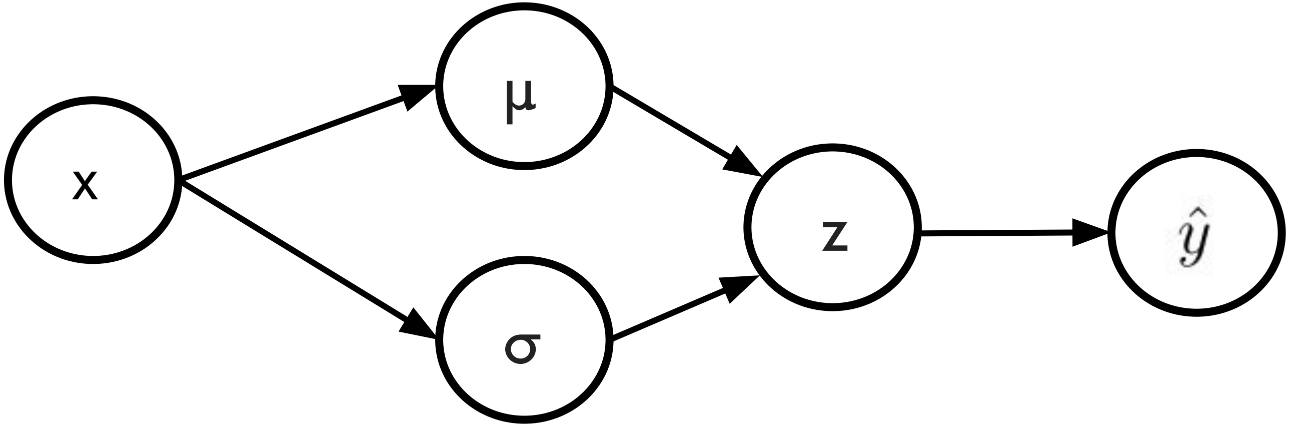
$$\operatorname{argmin}_{q,r} \left(KL(q(z|x) \,||\, p(z)) + \mathbb{E}_{z \sim q(z|x)} MSE(x, r(z)) \right)$$

Regularization

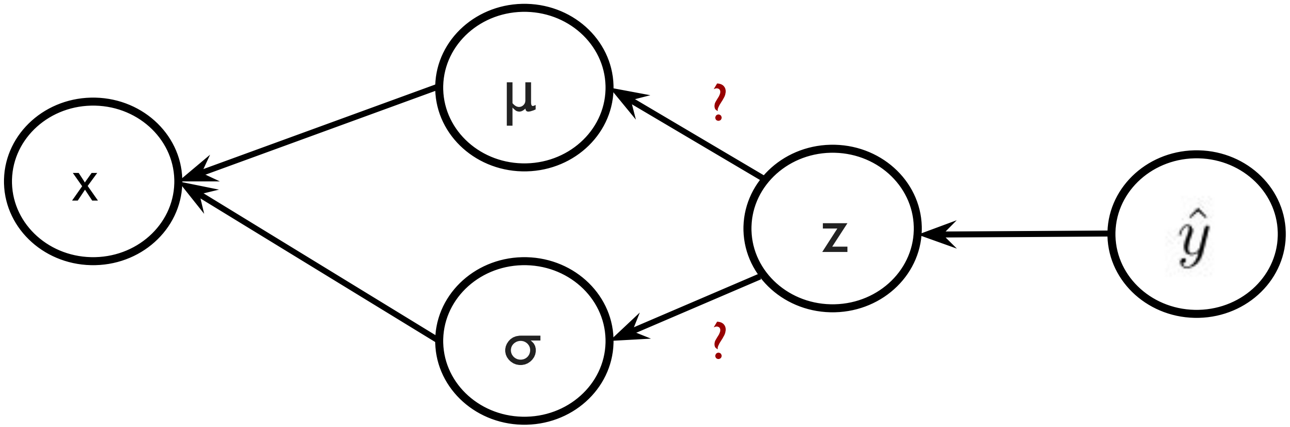
Reconstruction loss

Source: http://kvfrans.com/variational-autoencoders-explained/

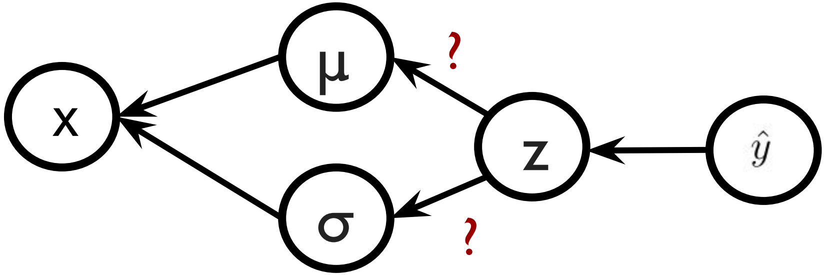
VAE Computational Graph - Forward Pass



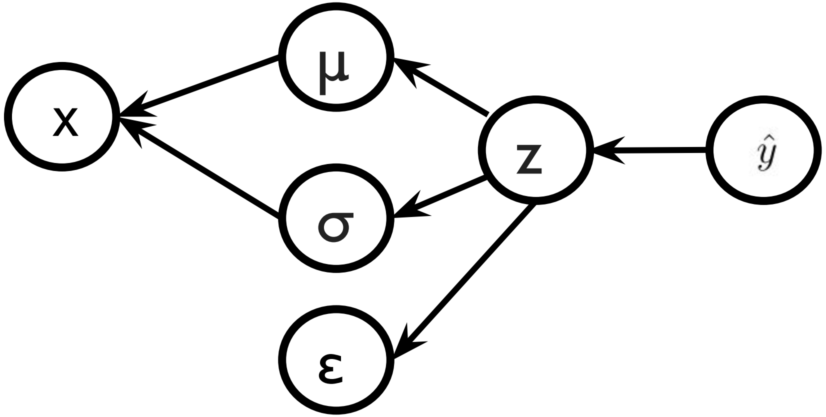
VAE Computational Graph - Backward Pass



Reparameterization Trick



$Z \sim N(\mu,\sigma)$



$\varepsilon \sim N(0,1)$

$Z = \varepsilon\sigma + \mu$

Where did this loss function come from?

$$\text{argmin}_{q,r}\left(KL(q(z|x)\,||\,p(z)) + \mathbb{E}_{z\sim q(z|x)} MSE(x, r(z))\right)$$

Regularization

Reconstruction loss

Estimating the VAE Lower Bound

VAE Objective Function: $\mathrm{argmax}_\theta p_\theta(X)$

Estimating the VAE Lower Bound

VAE Objective Function: $\text{argmax}_\theta p_\theta(X)$

Latent Representation: $p_\theta(X) = \int p_\theta(z) p_\theta(X|z) dz$ (Think HMM's)

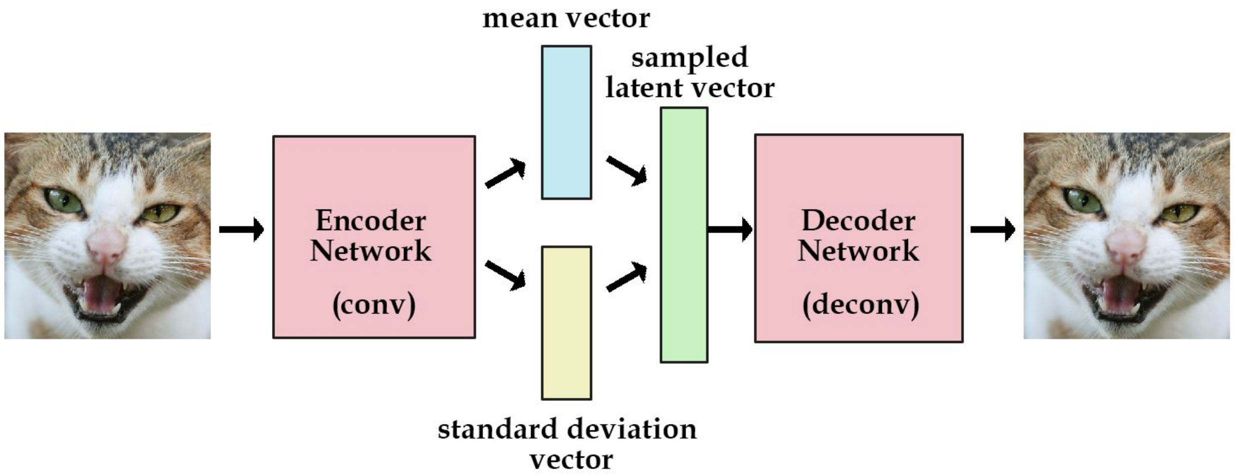
Estimating the VAE Lower Bound

VAE Objective Function: $\text{argmax}_\theta p_\theta(X)$

Latent Representation: $p_\theta(X) = \int p_\theta(z) p_\theta(X|z) dz$ (Think HMM's)

Can compute individual terms of integral, but not the integral itself!

Estimating the VAE Lower Bound



Encoder Network: $q_\phi(z|x)$ Decoder Network: $p_\theta(x|z)$

Source: http://kvfrans.com/variational-autoencoders-explained/

Estimating the VAE Lower Bound

$$
\begin{aligned}
\log p_\theta(x^{(i)}) &= \mathbf{E}_{z\sim q_\phi(z|x^{(i)})}\left[\log p_\theta(x^{(i)})\right] \qquad (p_\theta(x^{(i)}) \text{ Does not depend on } z) \\
&= \mathbf{E}_z\left[\log \frac{p_\theta(x^{(i)} \mid z)p_\theta(z)}{p_\theta(z \mid x^{(i)})}\right] \qquad (\text{Bayes' Rule})
\end{aligned}
$$

Equations from Stanford lecture linked here: https://www.youtube.com/watch?v=5WoItGTWV54

Estimating the VAE Lower Bound

$$\begin{aligned}
\log p_\theta(x^{(i)}) &= \mathbf{E}_{z\sim q_\phi(z|x^{(i)})}\left[\log p_\theta(x^{(i)})\right] \qquad (p_\theta(x^{(i)}) \text{ Does not depend on } z) \\
&= \mathbf{E}_z\left[\log \frac{p_\theta(x^{(i)} \mid z)p_\theta(z)}{p_\theta(z \mid x^{(i)})}\right] \qquad (\text{Bayes' Rule}) \\
&= \mathbf{E}_z\left[\log \frac{p_\theta(x^{(i)} \mid z)p_\theta(z)}{p_\theta(z \mid x^{(i)})}\frac{q_\phi(z \mid x^{(i)})}{q_\phi(z \mid x^{(i)})}\right] \qquad (\text{Multiply by constant}) \\
&= \mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - \mathbf{E}_z\left[\log \frac{q_\phi(z \mid x^{(i)})}{p_\theta(z)}\right] + \mathbf{E}_z\left[\log \frac{q_\phi(z \mid x^{(i)})}{p_\theta(z \mid x^{(i)})}\right] \qquad (\text{Logarithms})
\end{aligned}$$

Equations from Stanford lecture linked here: https://www.youtube.com/watch?v=5WoItGTWV54

Estimating the VAE Lower Bound

$$
\begin{aligned}
\log p_\theta(x^{(i)}) &= \mathbf{E}_{z\sim q_\phi(z|x^{(i)})}\left[\log p_\theta(x^{(i)})\right] \qquad (p_\theta(x^{(i)}) \text{ Does not depend on } z) \\
&= \mathbf{E}_z\left[\log \frac{p_\theta(x^{(i)} \mid z)p_\theta(z)}{p_\theta(z \mid x^{(i)})}\right] \qquad (\text{Bayes' Rule}) \\
&= \mathbf{E}_z\left[\log \frac{p_\theta(x^{(i)} \mid z)p_\theta(z)}{p_\theta(z \mid x^{(i)})}\frac{q_\phi(z \mid x^{(i)})}{q_\phi(z \mid x^{(i)})}\right] \qquad (\text{Multiply by constant}) \\
&= \mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - \mathbf{E}_z\left[\log \frac{q_\phi(z \mid x^{(i)})}{p_\theta(z)}\right] + \mathbf{E}_z\left[\log \frac{q_\phi(z \mid x^{(i)})}{p_\theta(z \mid x^{(i)})}\right] \qquad (\text{Logarithms}) \\
&= \mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z)) + D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z \mid x^{(i)}))
\end{aligned}
$$

Equations from Stanford lecture linked here: https://www.youtube.com/watch?v=5WoItGTWV54

Estimating the VAE Lower Bound

$$\mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z)) + D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z \mid x^{(i)}))$$

$$\mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right]$$

Log likelihood from our decoder network!

$$- D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z))$$

KL Divergence constraining z to STD Gaussian

$$D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z \mid x^{(i)}))$$

Intractable term

Equations from Stanford lecture linked here: https://www.youtube.com/watch?v=5WoItGTWV54

Estimating the VAE Lower Bound

$$\mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z)) + D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z \mid x^{(i)}))$$

$$\mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right]$$ Log likelihood from our encoder network!

$$- D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z))$$ KL Divergence constraining z to STD Gaussian

$$\geq \mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z))$$ (Since KL Div always non-negative)

Equations from Stanford lecture linked here: https://www.youtube.com/watch?v=5WoItGTWV54

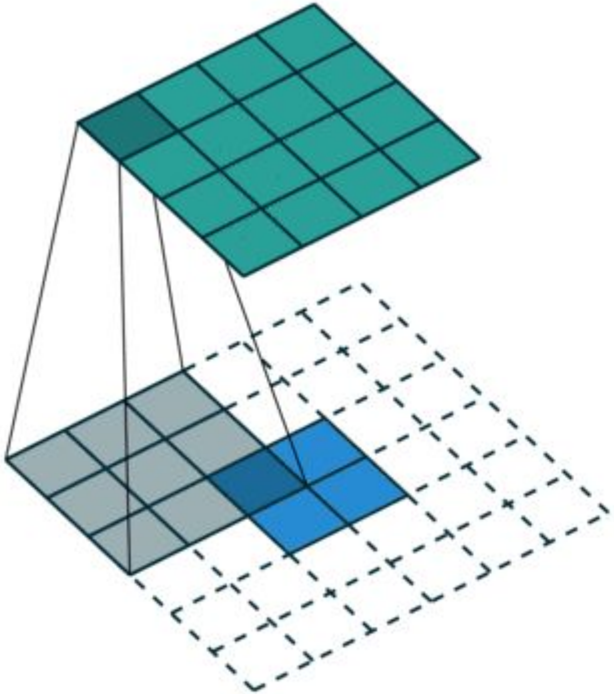
Estimating the VAE Lower Bound

$$\mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z))$$

$$\operatorname{argmin}_{q,r}\left(KL(q(z|x)\,||\,p(z)) + \mathbb{E}_{z\sim q(z|x)} MSE(x, r(z))\right)$$

Equations from Stanford lecture linked here: https://www.youtube.com/watch?v=5WoItGTWV54

Transposed Convolution (“Deconvolution”)



Source: https://github.com/vdumoulin/conv_arithmetic

What works best to increase resolution?

- Upsampling an image
- Upsampling + Convolution
- Unpooling
- Transposed Convolution
- None of the above

Issues with VAEs

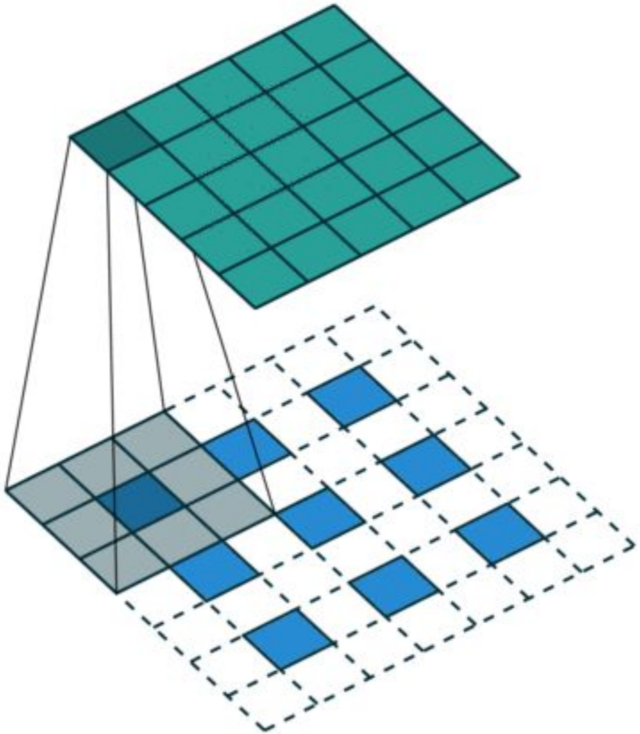
Input

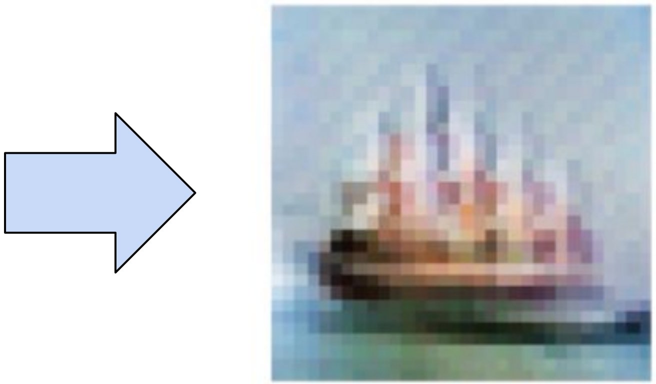
VAE reconstruction

VAE/GAN reconstruction

Current fix is to use a VAE + GAN (you'll learn about GANs in the next lecture)

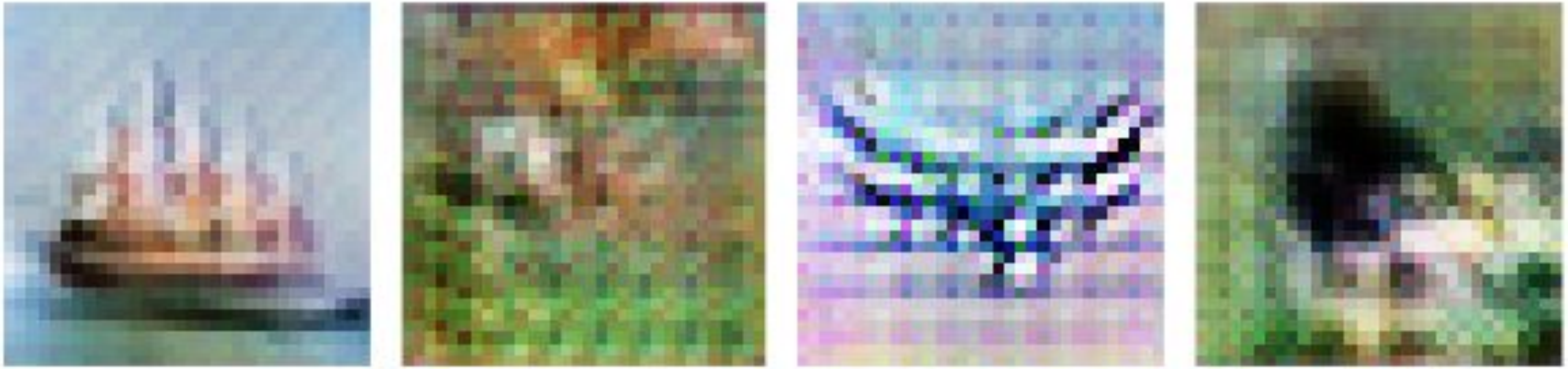
Transposed Convolution ("Deconvolution")





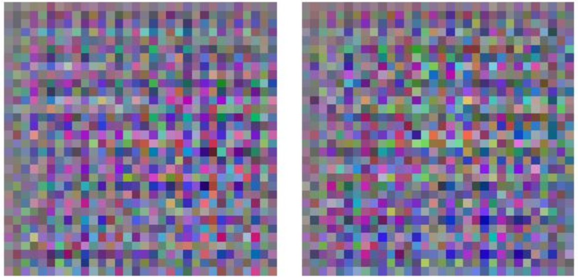
Source: https://github.com/vdumoulin/conv_arithmetic

Checkerboarding Artifact

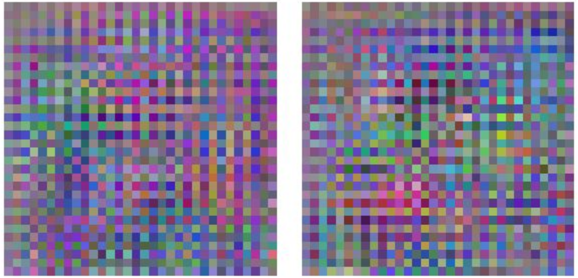


Source: https://distill.pub/2016/deconv-checkerboard/

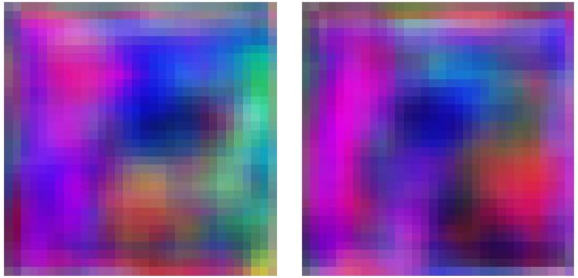
Resize Convolutions



Deconvolution in last two layers.
Artifacts prior to any training.



Deconvolution only in last layer.
Artifacts prior to any training.



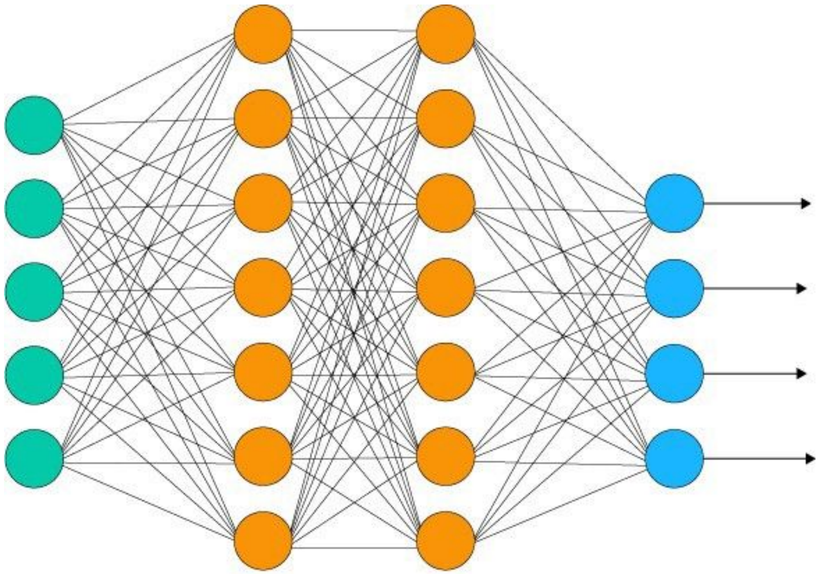
All layers use resize-convolution.
No artifacts before or after training.

Source: https://distill.pub/2016/deconv-checkerboard/

Better Weight Initialization

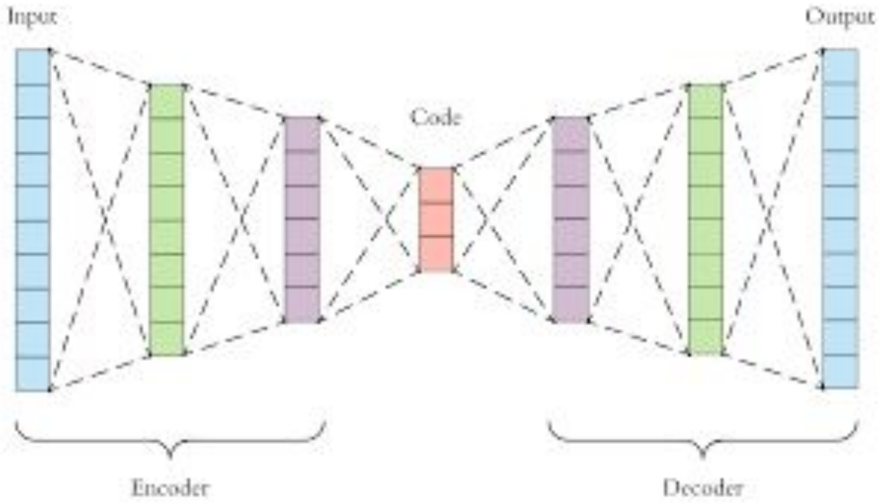
- Train unsupervised with unlabeled inputs
- Throw away decoder network
- Start training with labeled data!

“Explainable” AI



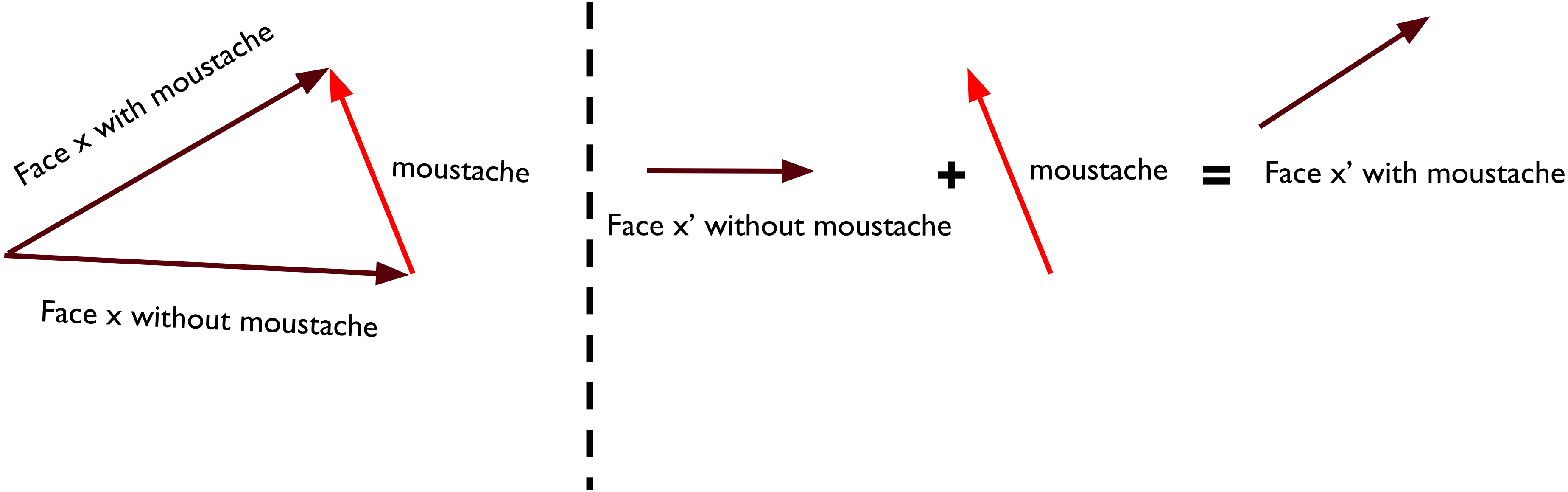
What do these weights mean?
What if we do a mis-prediction?

"Explainable" AI

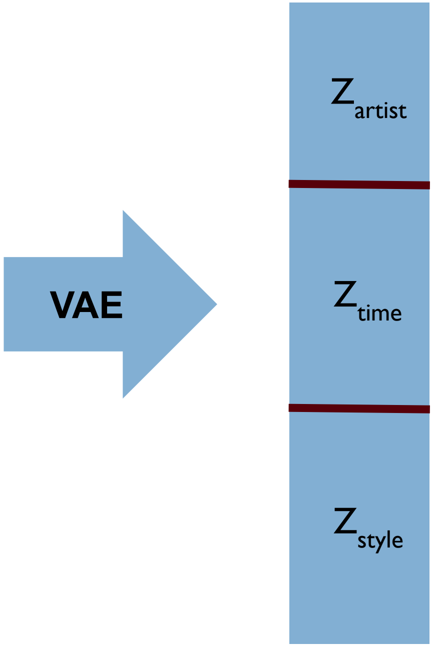


Vary each of the features in the code vector and see what the output looks like!

VAE Vector Arithmetic



Style Disentanglement



Mona lisa painted by Dali in 2019?

Cubist "Starry Night" in 1500?

Style Disentanglement

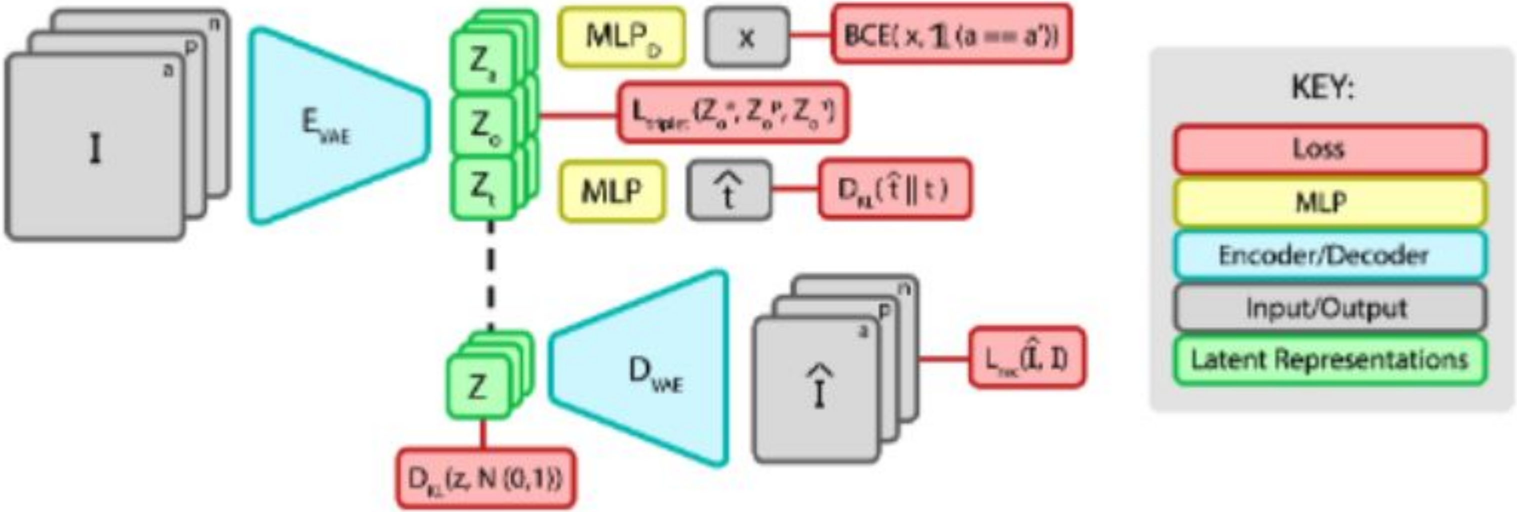
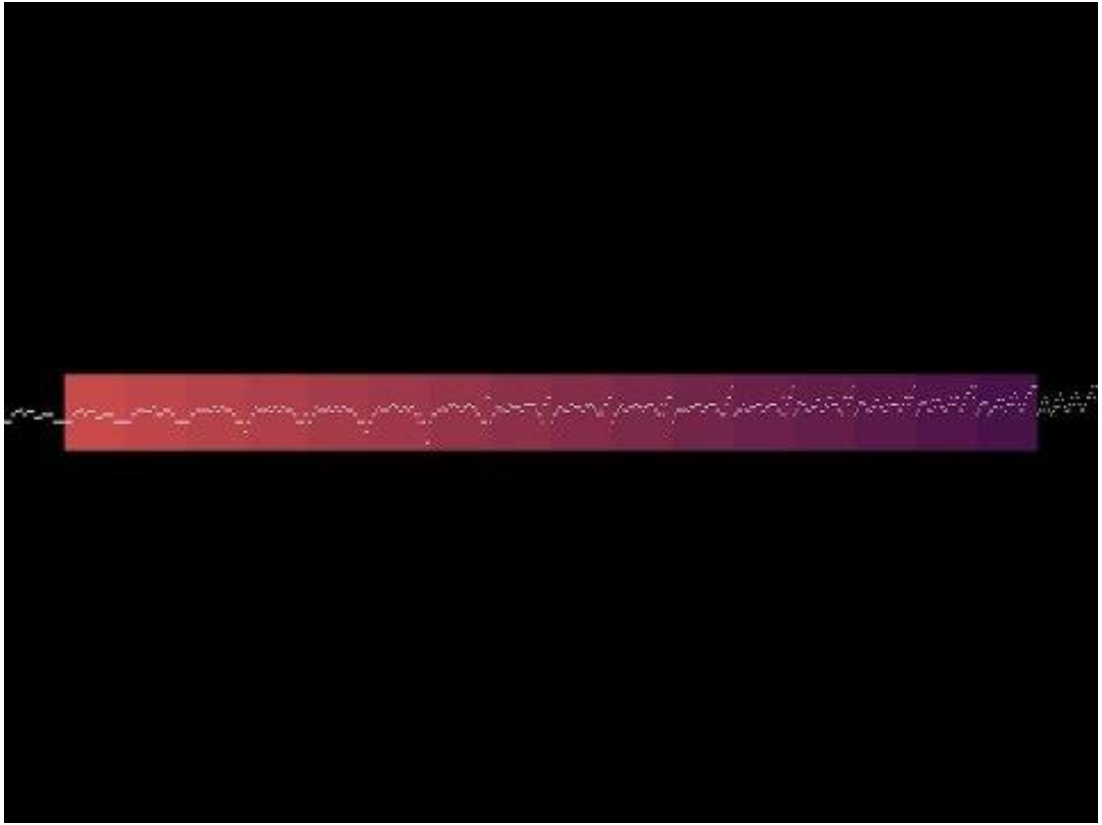


Figure 1. VAE architecture.

Interpolation of Latent Vectors - Music VAE



How could lecture 9 have been better?