Advanced neural network architectures

Dr. Charles Kind

Bristol University

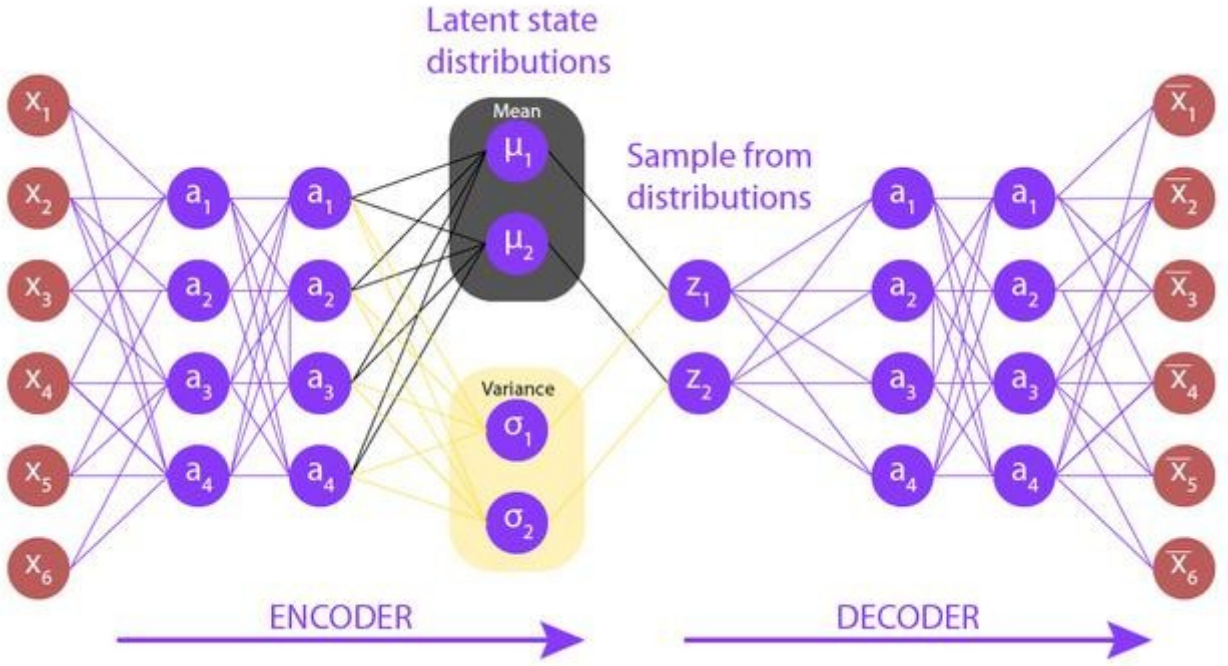
May 2023

Learning objectives

- What are variational autoencoders (VAE)?
- What is latent space?
- How do VAE provide access to latent space.
- What are generative adversarial networks (GAN) and how do they provide access to latent space.
- The differences between VAE and GAN.
- An overview of the Stable Diffusion architecture,

Variational autoencoders

- The ‘bottleneck’ is replaced with two separate vectors:
  - A vector representing the mean of your distribution
  - Another vector representing the standard deviation of your distribution
- Training changes to reflect this architecture.
- The cost function has two terms:
  - The reconstruction cost based off the expectation (in the probablistic sense as we are sampling from a distribution)
  - The KL divergence, which essentially tries to ensure that the distribution you are learning is close to a normal distribution



Generative models

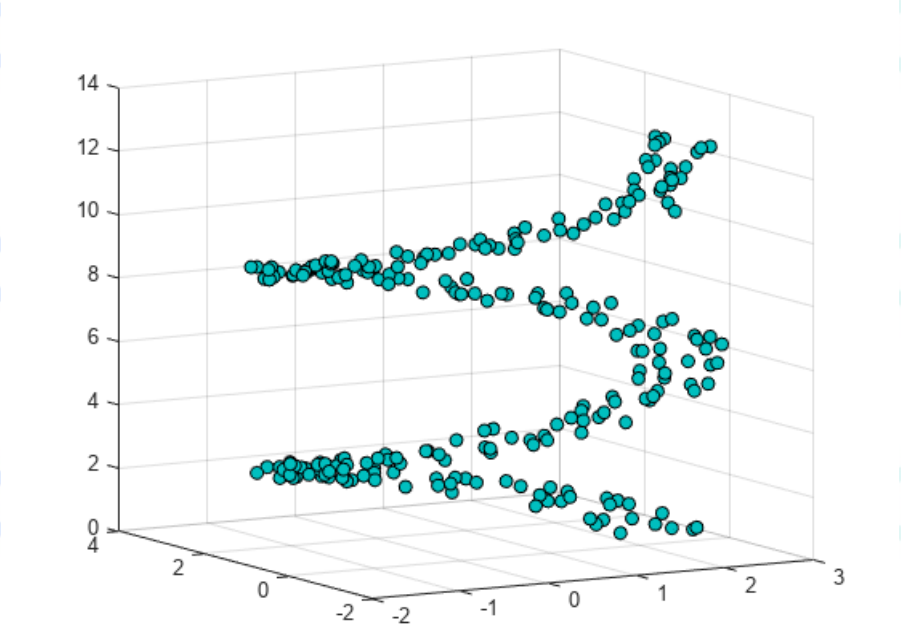
- You can think of generative neural networks as ‘backwards’ neural networks!
  - These networks generate NEW data with similar statistics to their training data

Stable diffusion prompt: a photograph of an astronaut riding a horse

- GAN: Generative adversarial networks
  - Two networks compete against each other
  - One networks gain is the others loss
  - These networks are therefore trained to fool their discriminator networks whose purpose is to spot generated data
  - ChatGPT and Stable Diffusion both use this architecture in their training

Generative adversarial networks

- What kind of problems could we have using standard DNN to generate data?
- Imagine we had trained a DNN, on data samples like the below, to output a function that minimises the error.
- This function would essentially be a helix.



- I we asked this DNN to output data that looks like the original it would not be able to.
- It would just produce spirals.
- Although there are an essentially infinite number of valid representations that you or I would consider like the helix with some noise.

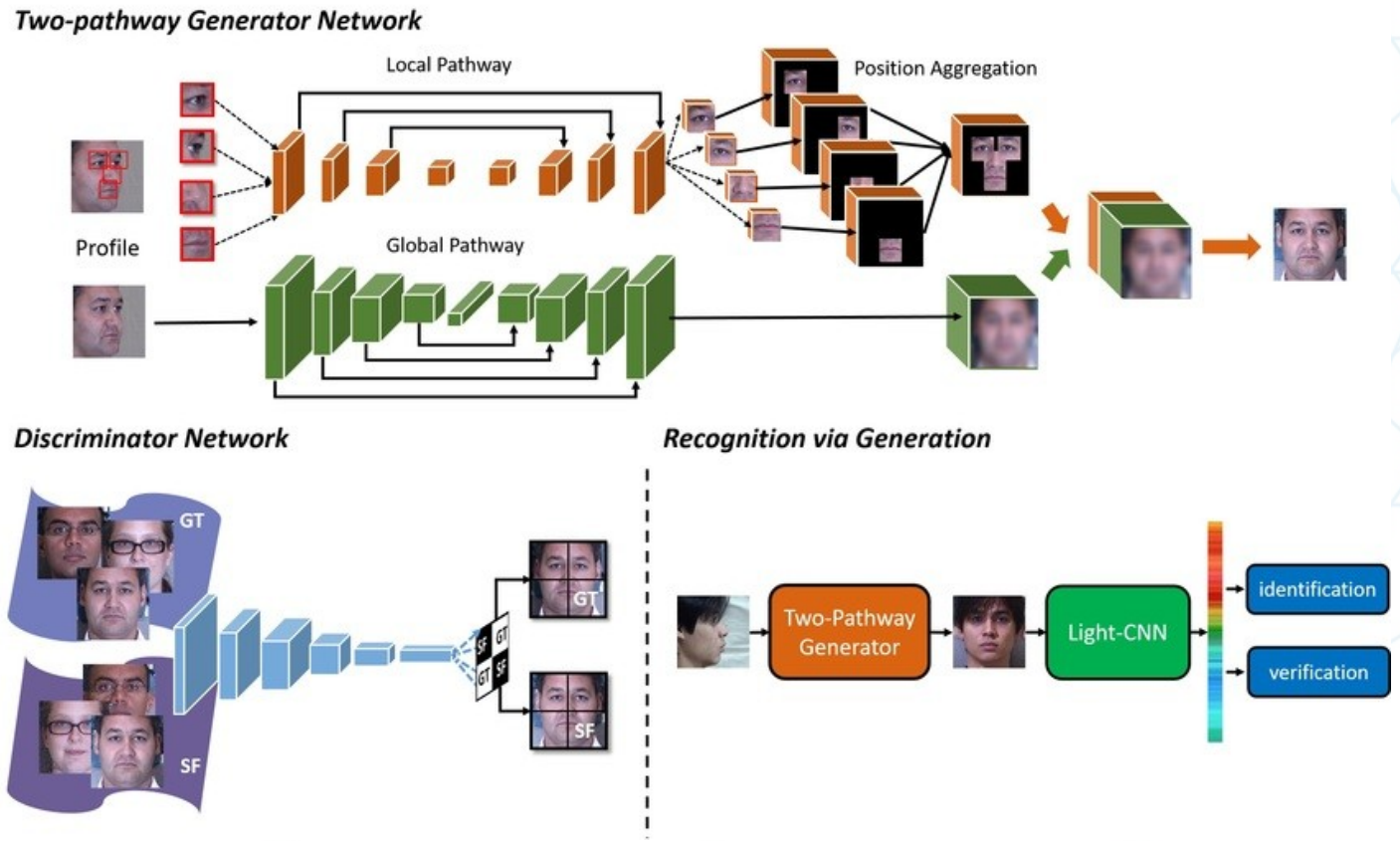
- How can we get a DNN to produce ‘realistic’ output?
- You or I would say that a series of points normally distributed about the helix, within a standard deviation, would look like the original.

Generative adversarial networks

- Left: An example from Nvidias styleGAN.
- The network learns to produce faces by competing against a discriminator.
- Below: A simple flow chart of the GAN architecture.

Generative adversarial networks

- Can we start from randomly initialised networks?
- It depends on the data. Highly complex data with deep structure, like images of faces, may require pre-training prior to the GAN stage.



Huang et. Al, 2017,Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis

GAN vs VAE

- GAN and VAE are very similar in important ways.
- A GAN learns to map from the entire dimensional space it can represent into the ‘latent’ space it is being trained on. For example images of horses and astronauts.
- A VAE is training a probabilistic function to map to the latent space of the data it is learning. The network basically convolves the known representation with a mollifier and enforces that the result is close to a standard distribution.
  - This is similar to the way a physicist would measure the temperature at various points in a space and then apply a smoothing function, via convolution, to make the data continuous for use with calculus.
- VAEs are frequently simpler to train than GANs as they can be unsupervised and do not require synchronisation with a discriminator.
- GANs are likely to recognise more complicated insights of the input and generate higher and more detailed plausible data than VAEs
- GANs are therefore generally used in more demanding tasks like image-to-image translation. VAEs are widely used in image denoising and generation.
- For example Stable Diffusion can use a VAE to clean up generated images.

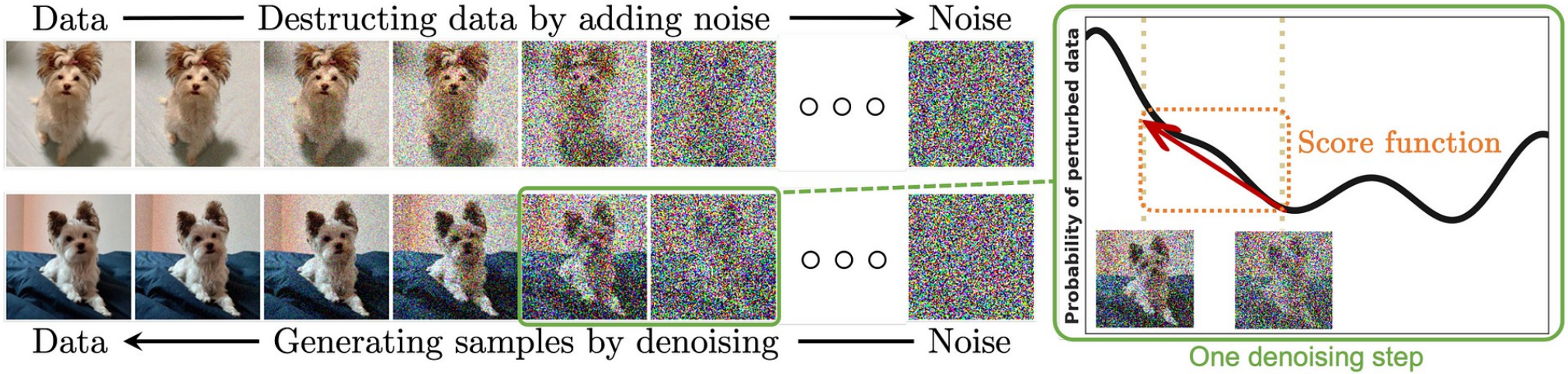
GAN arithmetic!

- It turns out, due to the structure of latent space, that we can use basic linear algebra to manipulate our output.
- Imagine we have trained our GAN on images of peoples faces. A very common thing to do!
- After training we can input random values from our latent space and recover novel images of faces.
- If we average the vectors of a large number of images of women, we would have an novel image of an 'average' woman. Call that vector W.
- If we average the vectors of a large number of images of men, we would have an novel image of an 'average' man. Call that vector M.
- We then average the vectors of women wearing hats. Call that vector WH.
- Then we could do the following:
  - Move from W to M in small steps and generate a video of a transition of a female character to a male character.
  - WH – W + M = an image of a man wearing a hat

- This is an amazing and powerful result. It also works on the latent space of word relationships from a transformer. Famously:
  king – man + woman = queen and London – England + Japan = Tokyo

So what about Stable Diffusion?

- It’s all in the diffusion!
- We want to generate novel images from trained data. We want in some sense to add noise and recover a new meaningful, high quality image.
- GANs suffer from a lack of ability to generate truly new, surprising images that also conform to the viewers expectations.
- What we do is add noise in steps and then train the network to denoise.



- The denoising is done in small steps. The network ‘guesses’ at the final denoised image and then most of the noise is added back and the process is done again (and again ...).

Stable Diffusion

- To generate meaningful images from text we also need a transformer neural network trained to label images.
- The transformer is trained against the convolutional image classifier so that labels are related to features of images.
- This network is then used to guide the denoising process.
- So called ‘Classifier free guidance’ is also used.
- This is where the guided denoised image step is compared to a random denoised image and the differences amplified.
- This has the effect of forcing the denoising to concentrate on the encoded features from the transformer.

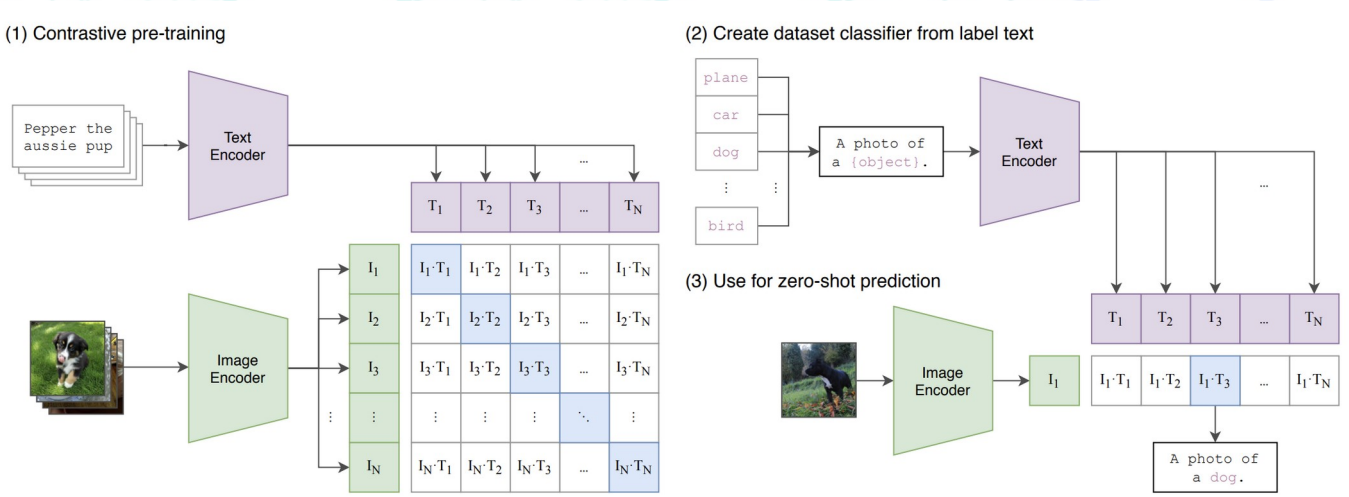
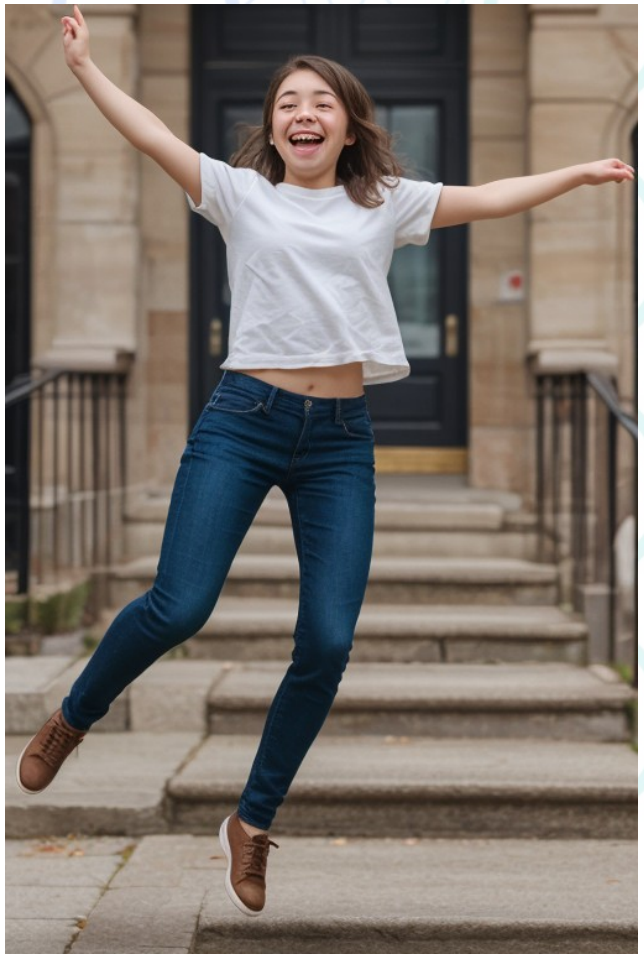


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset’s classes.

Stable Diffusion

- Then we can generate whatever we like!
- RIGHT: Prompt: “dr. charlie, a 50 year old physicist, wearing a crown, sitting on a throne, ((in a lecture hall))”



LEFT: Prompt: “A 20 year old student leaping for joy after passing comsm0094”

Thank you all for attending my lectures. I have really enjoyed comsm0094 in 2023.

:)