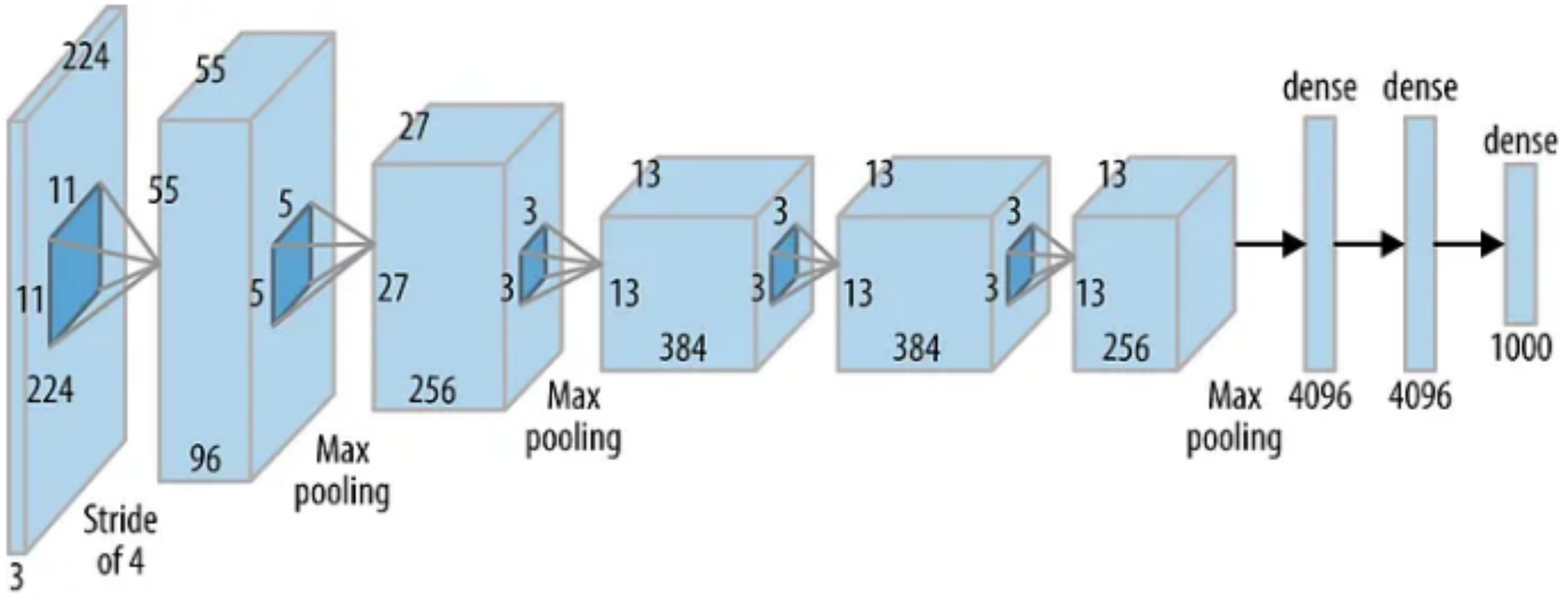
Overview of Deep Learning

From a model perspective

Mingtian Tan, Sept 2024

AlexNet 2012

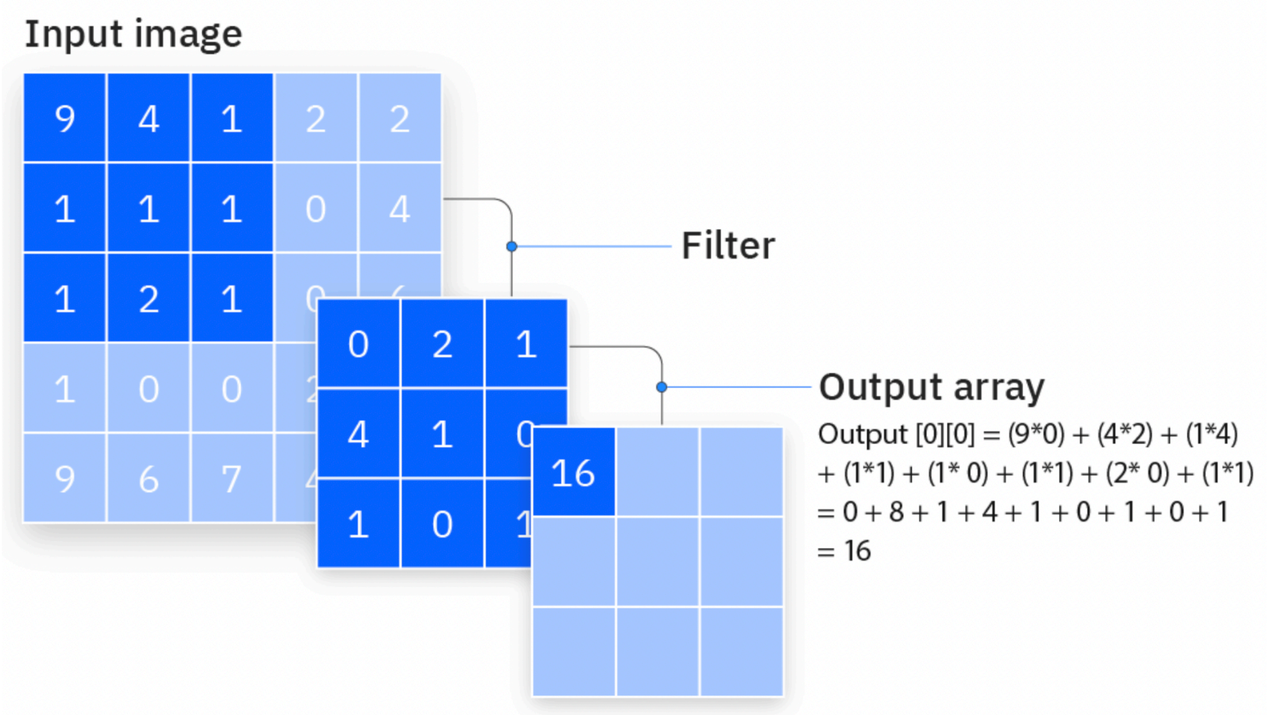
Image Source



AlexNet, introduced in 2012, was a groundbreaking model that showed how powerful deep learning could be, especially for image classification tasks. One of the key reasons it was so successful is because it used Nvidia GPUs to train the model. Hinton and his team realized that training such a deep neural network on a regular CPU would take far too long, but with GPUs, which are much faster at handling large amounts of data in parallel. This was a big turning point, as it showed the importance of using GPUs for deep learning and really pushed the field forward

ImageNet Classification with Deep Convolutional Neural Networks by Krizhevsky, Sutskever, and Hinton.

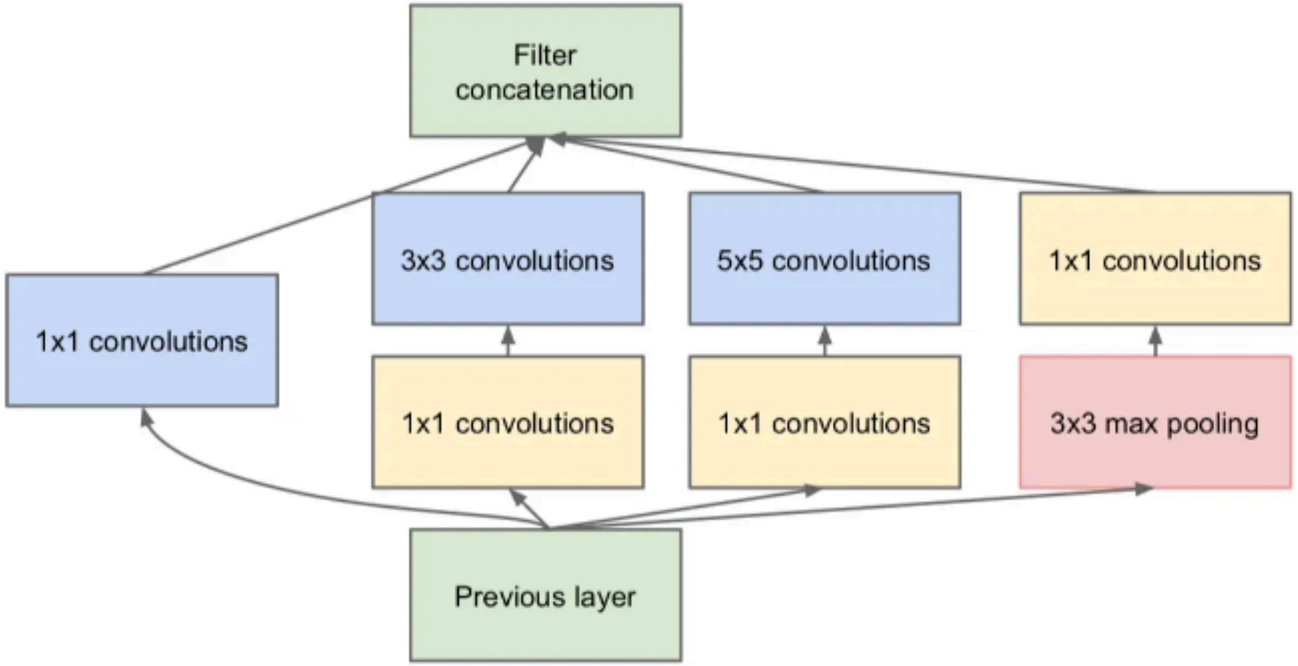
Image Source



A Convolutional Neural Network, or CNN, is a type of deep learning model that's great at handling visual data. It automatically learns patterns from images, making it really useful for things like recognizing objects or detecting them in pictures.

Inception (GoogLeNet) 2014

Image Source

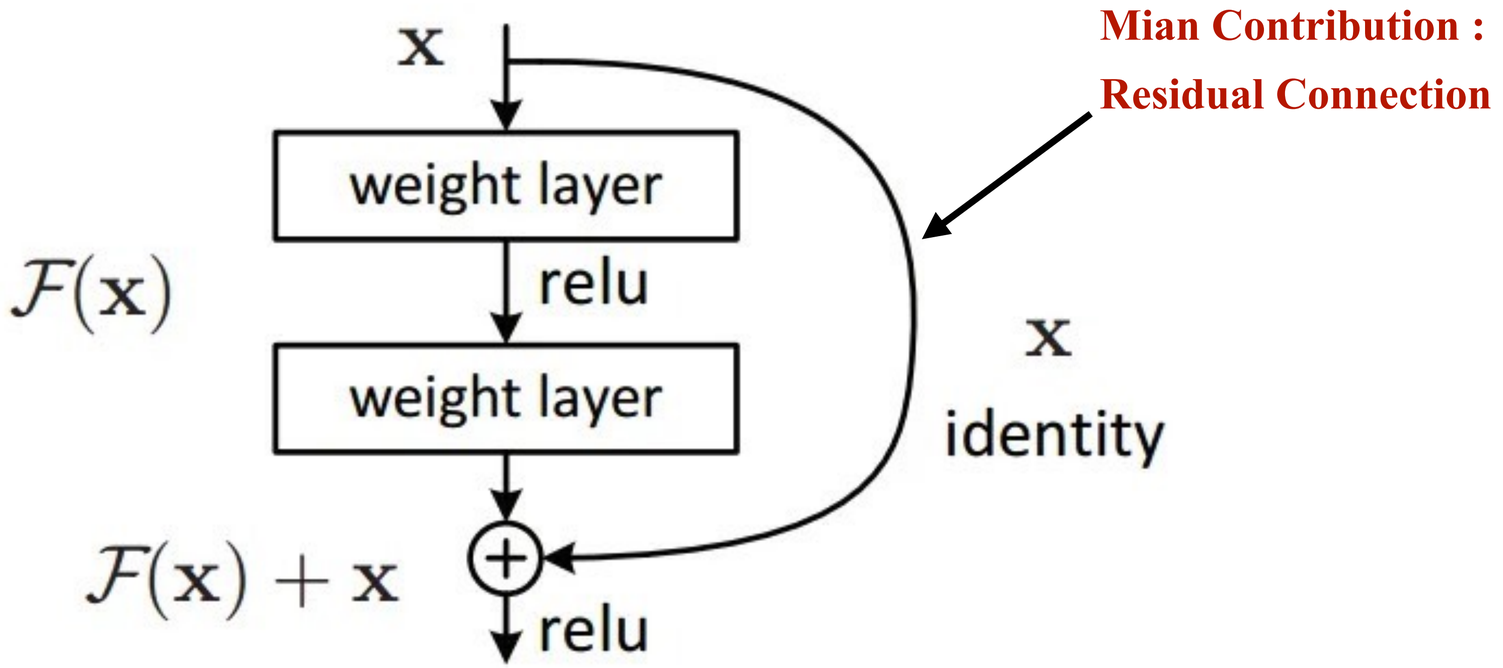


Inception was introduced by Google researchers in 2014 and brought a big breakthrough with its 'Inception module.' This idea let the model look at information at *different scales all at once* by using different filter sizes.. And it can save computational costs.

Going Deeper with Convolutions by Szegedy et al.
Rethinking the Inception Architecture for Computer Vision by Szegedy et al.

ResNet 2015

Image Source

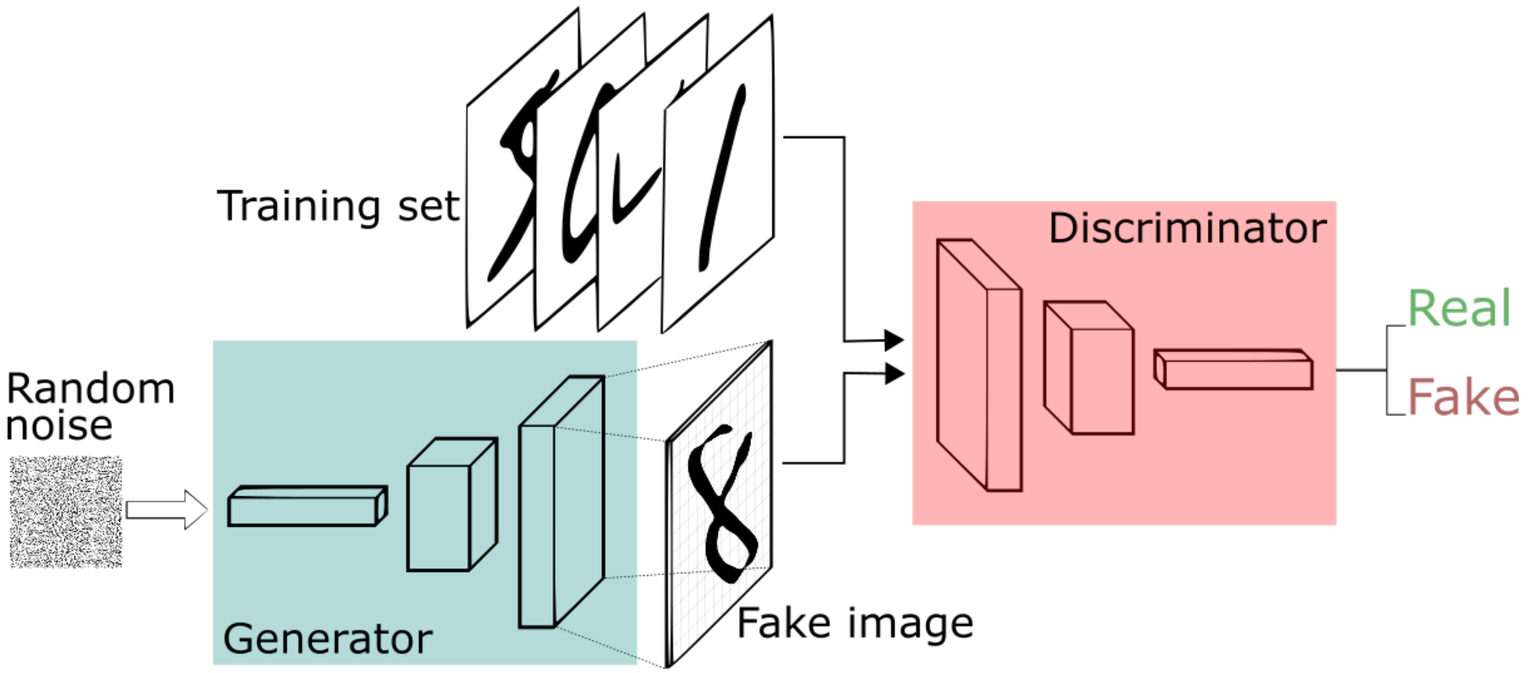


ResNet, created by Kaiming He and his team, solved the problem of training very deep networks by using something called residual connections. These connections make it easier for the network to learn, even with hundreds or thousands of layers. Thanks to this, ResNet achieved top performance in tasks like image classification.

Deep Residual Learning for Image Recognition by Kaiming He et al.
Identity Mappings in Deep Residual Networks by Kaiming He et al.
ResNeXt: Aggregated Residual Transformations for Deep Neural Networks by Saining Xie et al.

Generative Adversarial Networks (GANs) 2014

Image Source



GANs, introduced by Ian Goodfellow and his team, changed the game for generative models. They use two networks—a generator and a discriminator—that compete with each other. This back-and-forth results in highly realistic fake data, which has been used for things like creating images, art, and even expanding datasets.

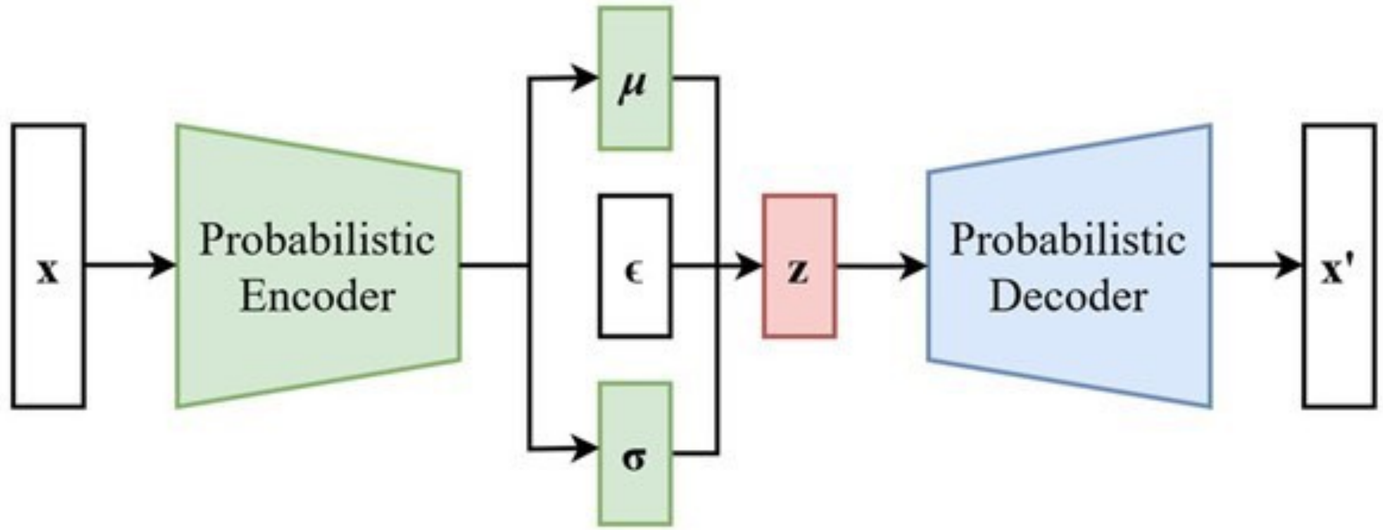
Generative Adversarial Nets by Ian Goodfellow et al.
Conditional Generative Adversarial Nets by Mehdi Mirza and Simon Osindero.

Examples generated by GANs or from our real world

Can you tell which ones are fake and which are real?

Variational Autoencoders (VAEs) 2014

Image Source



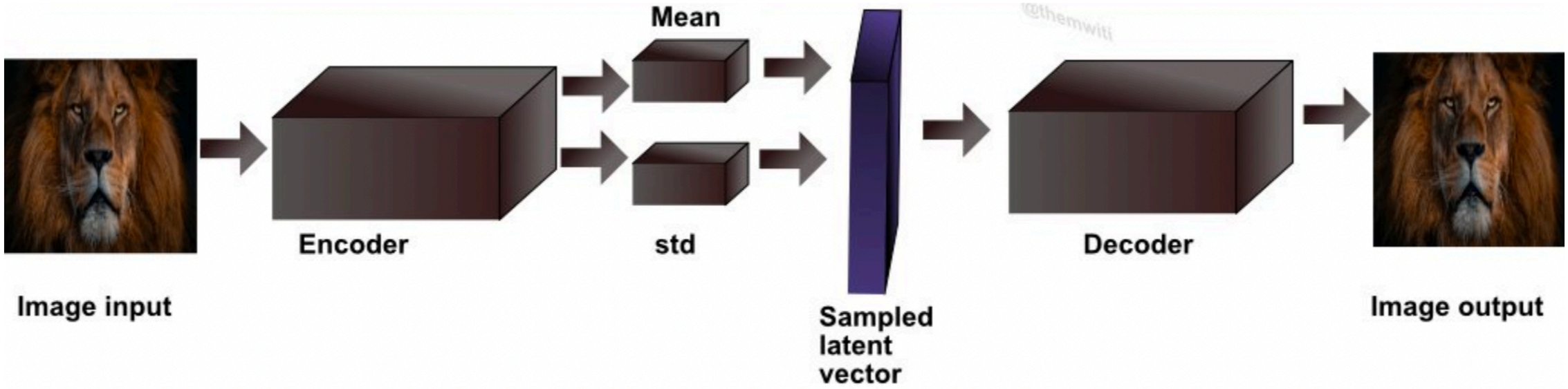
VAEs, created by Kingma and Welling, are a type of model that learns to compress data into a kind of 'hidden space' and then recreate it. Unlike regular autoencoders, VAEs can generate new data by sampling from this hidden space, which makes them great for tasks like creating images or spotting anomalies.

The biggest feature of VAEs is that the generated images are highly diverse.

Auto-Encoding Variational Bayes by Kingma and Welling.
Stochastic Backpropagation and Approximate Inference in Deep Generative Models by Rezende et al.

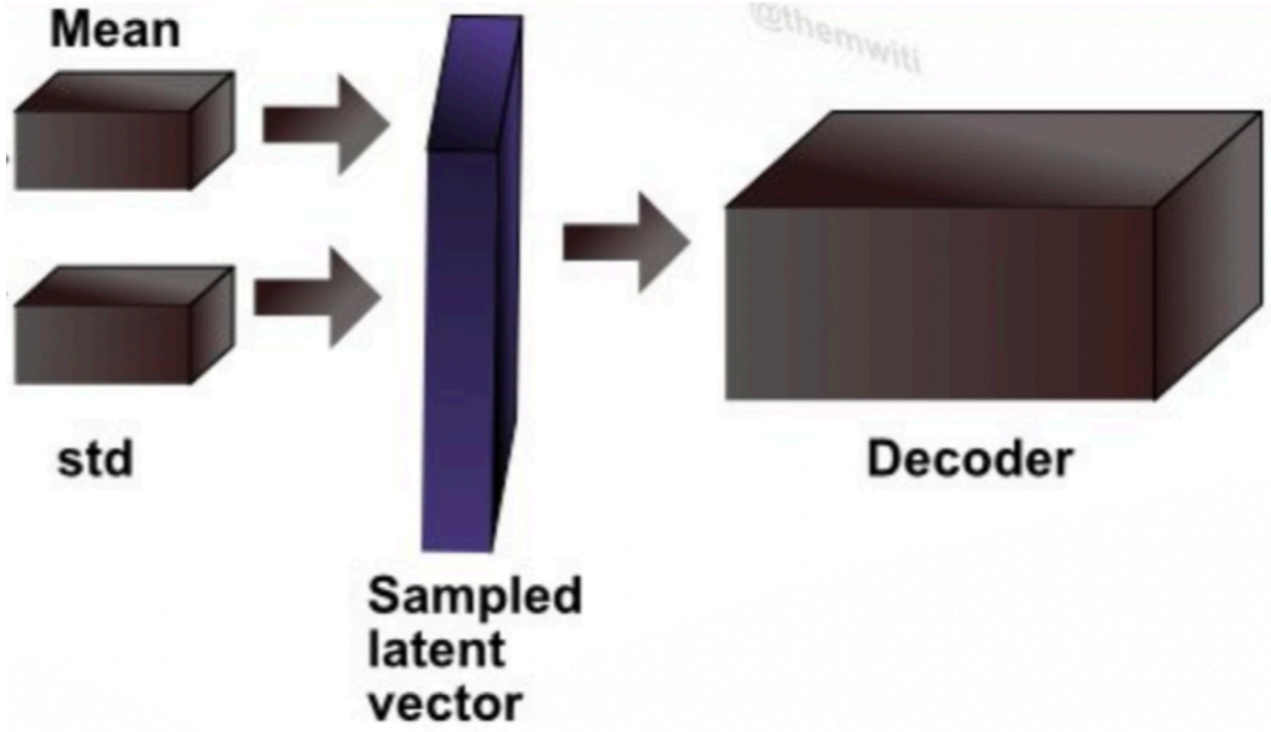
Example-VAE-Training

Image Source



Example-VAE-Generation

Image Source



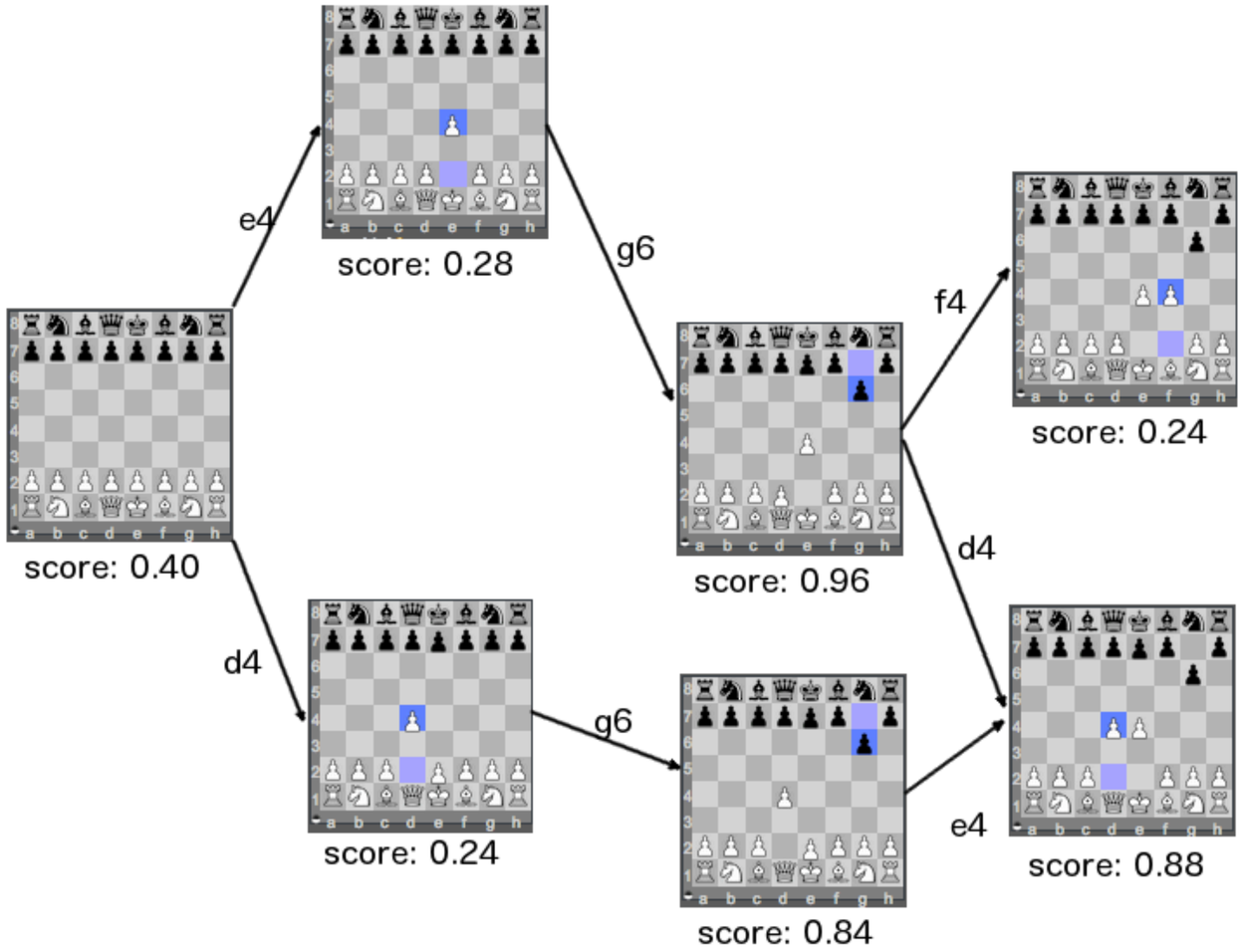
Generated Lions

...

Once you've trained on enough "lions," you can generate various lions. In a sense, these generated lions are “random combinations” of lions the model has seen.

AlphaGo 2016

Image Source

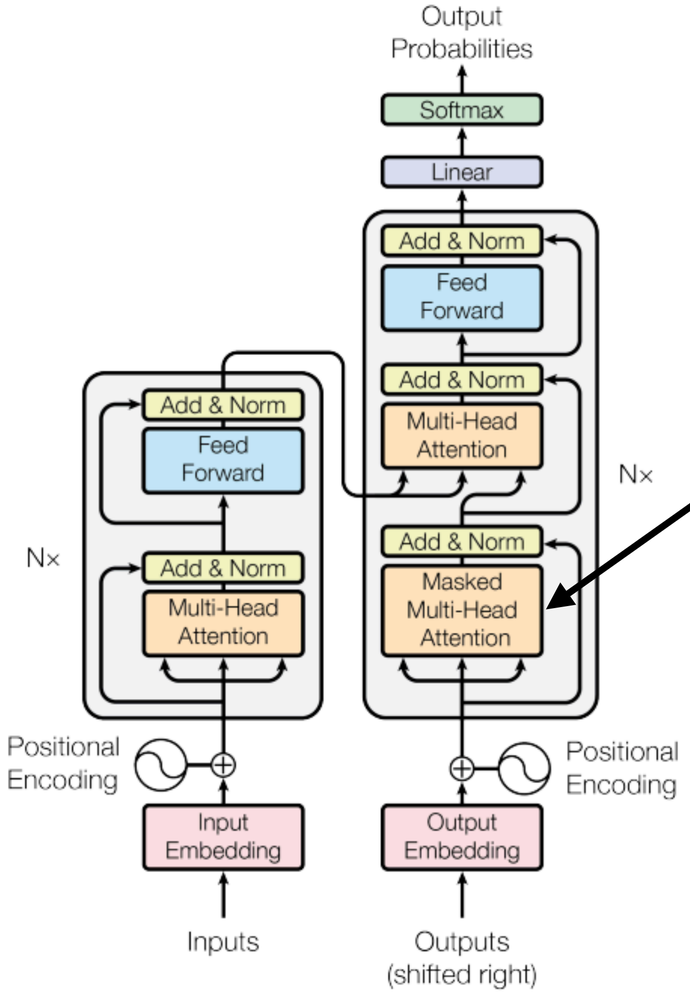


AlphaGo, created by DeepMind, was the first AI to beat the most professional Go player. It used deep neural networks along with a strategy called Monte Carlo tree search.

Mastering the game of Go with deep neural networks and tree search by Silver et al.
Mastering the game of Go without human knowledge by Silver et al.
Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm by Silver et al.

Transformers 2017

Image Source



Mian Contribution :

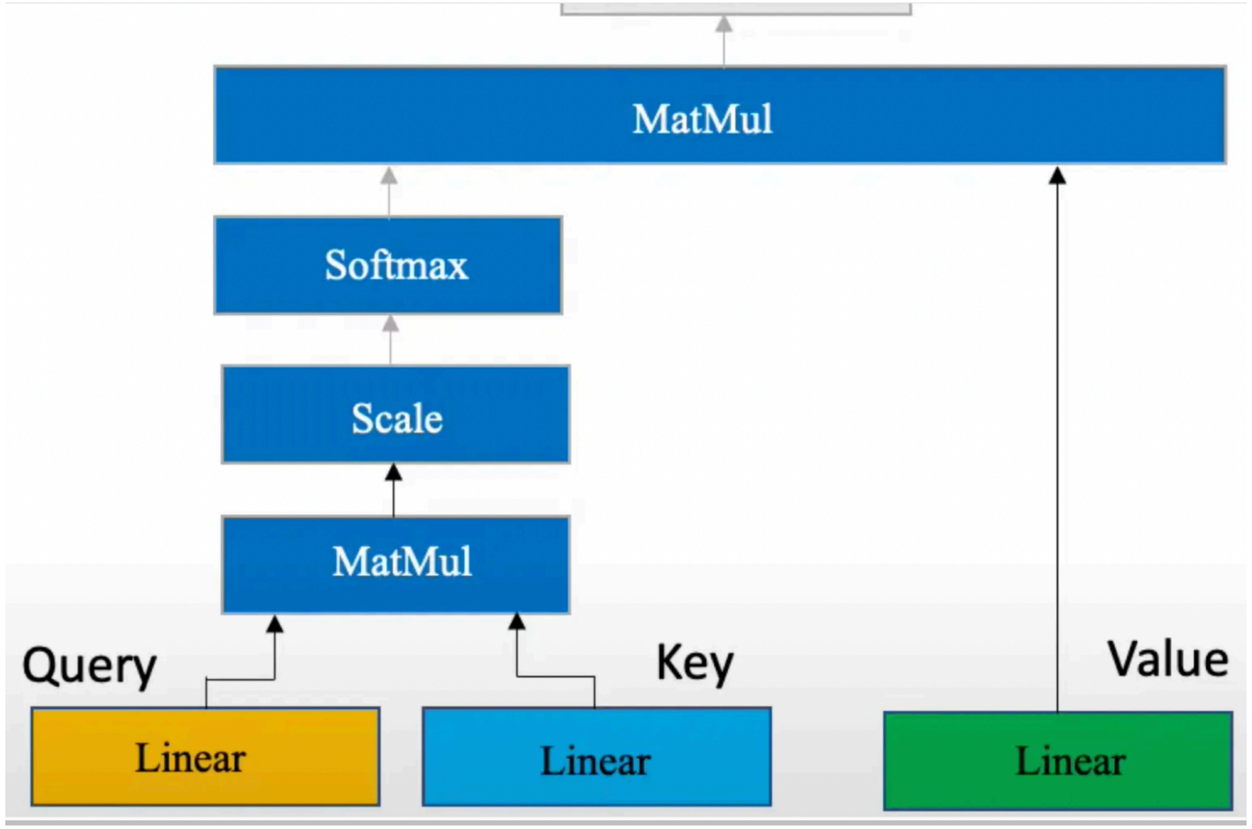
Self-Attention

The Transformer architecture, introduced by Vaswani and his team, changed natural language processing by getting rid of the need for recurrent and convolutional layers. Instead, it uses self-attention, which made it the base for powerful models like BERT and GPT.

Attention Is All You Need by Vaswani et al.
The Annotated Transformer by Harvard NLP.
Universal Transformers by Dehghani et al.

Self-Attention.

Image Source

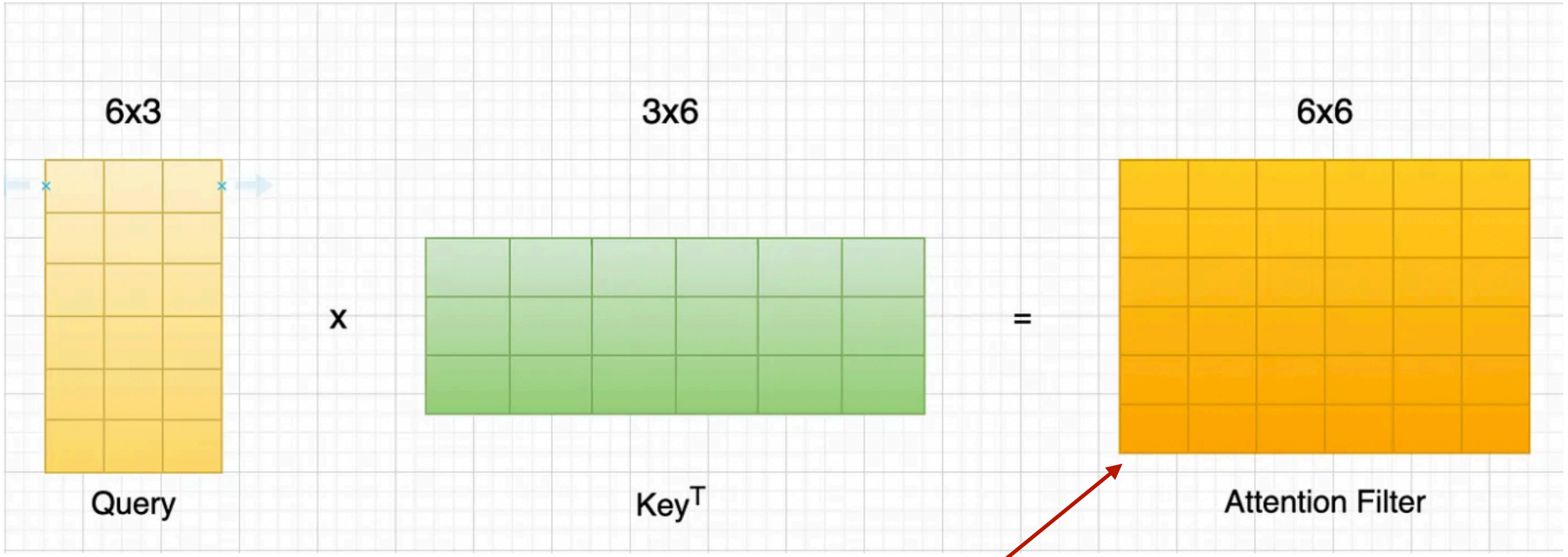


Hi	89	12	32
,	20	11	15
How	33	25	91
are	52	68	12
you	28	27	54
?	39	30	73

“Hi, How are you?” “Hi, How are you?” “Hi, How are you?”

matrix multiplication

Image Source



This matrix stores the relationships between tokens in a sentence.

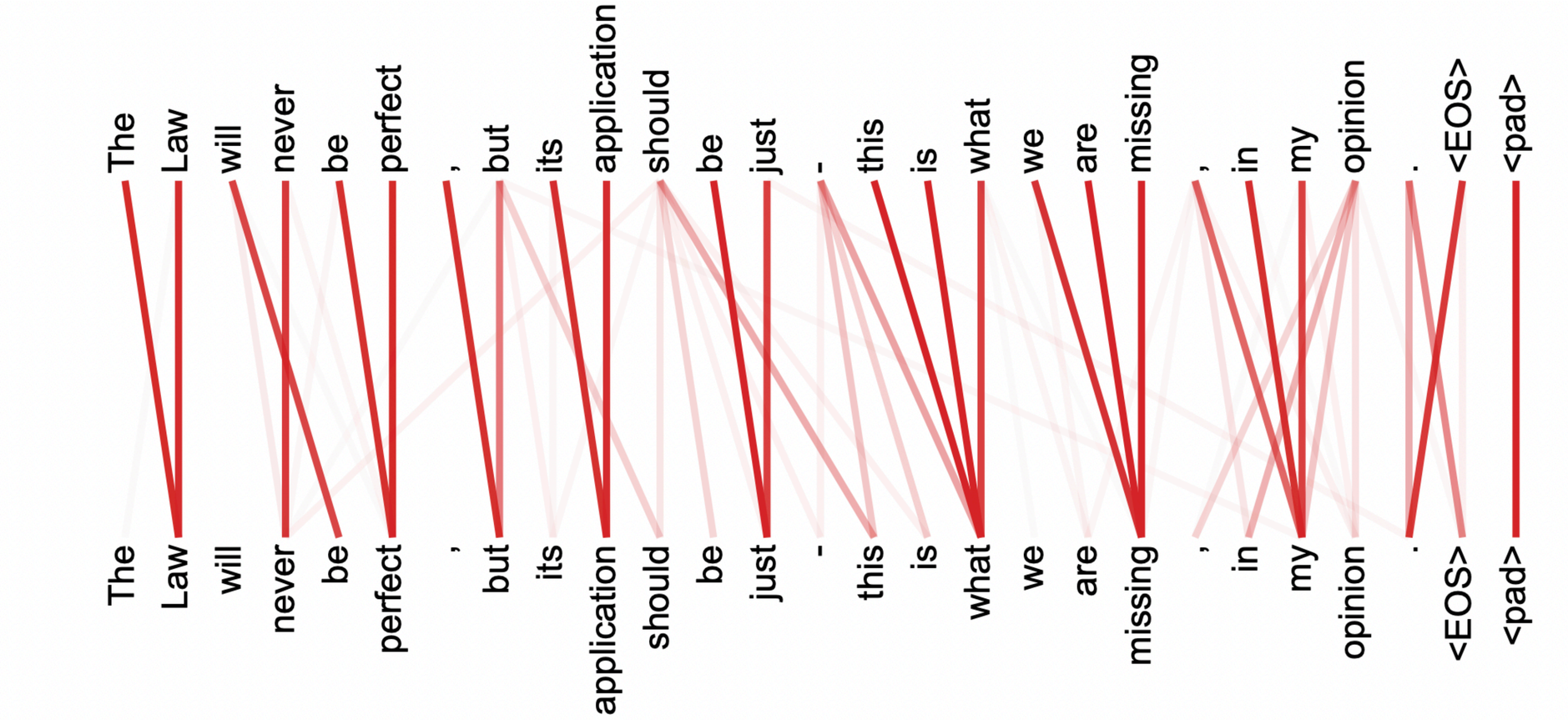


Figure 5: Many of the attention heads exhibit behaviour that seems related to the structure of the sentence. We give two such examples above, from two different heads from the encoder self-attention at layer 5 of 6. The heads clearly learned to perform different tasks.

Attention Is All You Need by Vaswani et al.

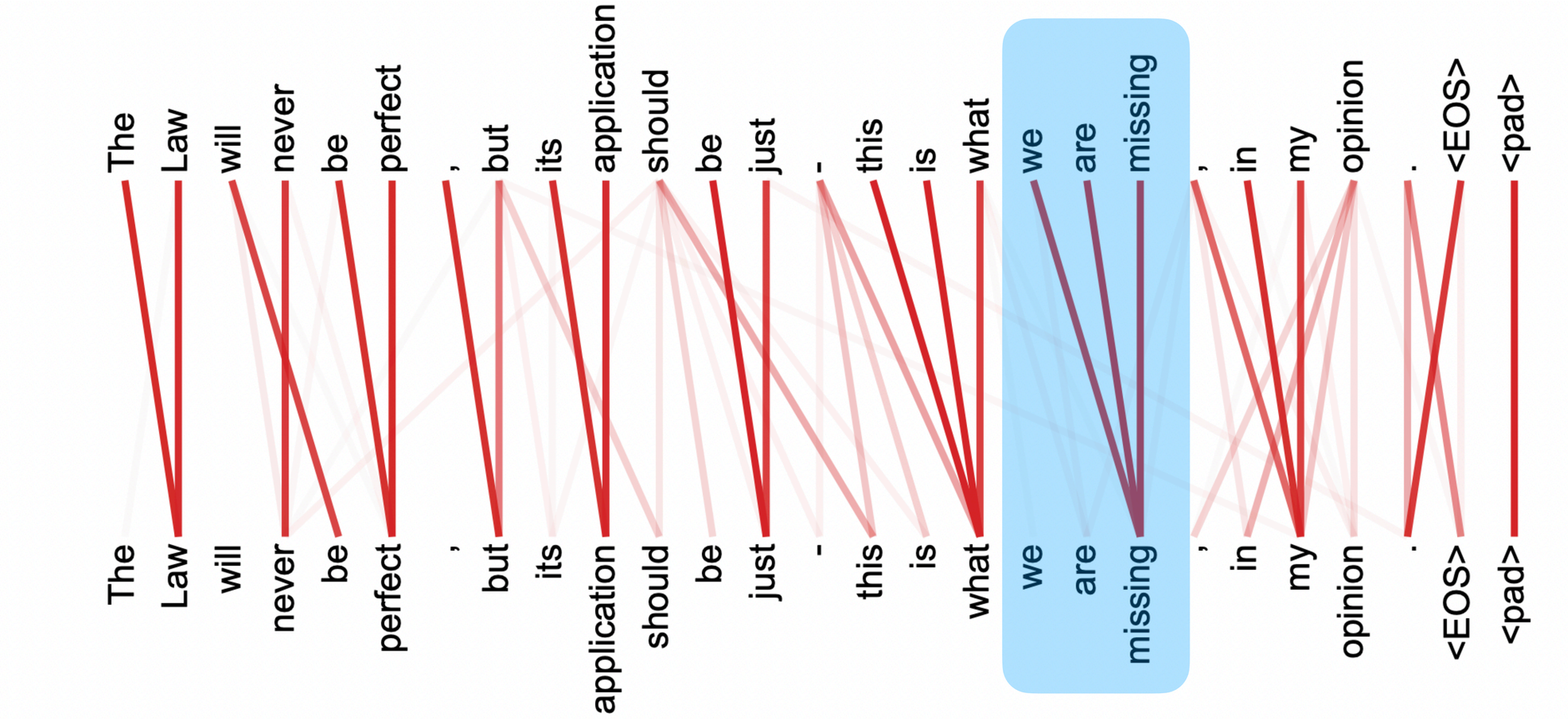
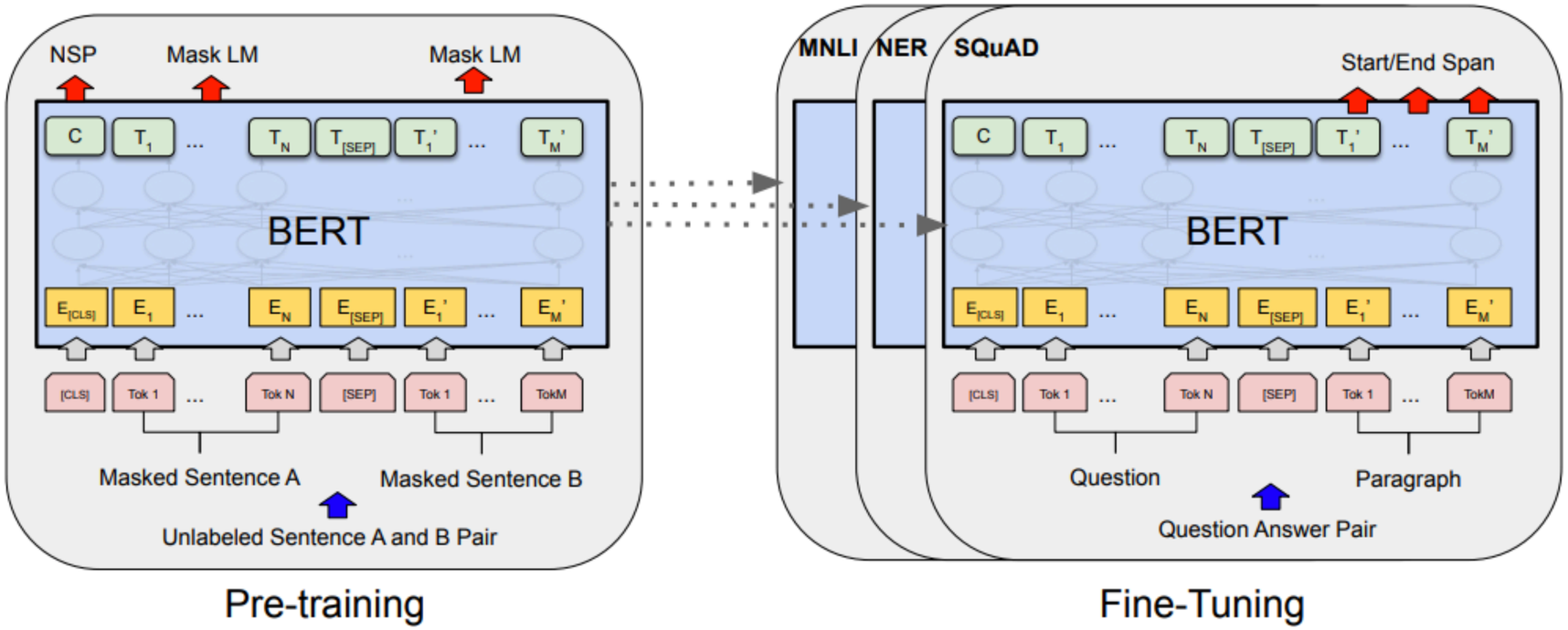


Figure 5: Many of the attention heads exhibit behaviour that seems related to the structure of the sentence. We give two such examples above, from two different heads from the encoder self-attention at layer 5 of 6. The heads clearly learned to perform different tasks.

Attention Is All You Need by Vaswani et al.

BERT 2018

Image Source



BERT, introduced by Google, was the first model to really popularize the Transformer architecture. Its bidirectional training on large amounts of text helped it understand language in a more detailed way. This led to top results in many NLP tasks like question answering and sentiment analysis.

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding by Devlin et al.
RoBERTa: A Robustly Optimized BERT Pretraining Approach by Liu et al.
XLNet: Generalized Autoregressive Pretraining for Language Understanding by Yang et al.

Auto-regression (Decoder-only LLM, GPT3) -2021

Image Source

Encoder-only *Decoder-only*

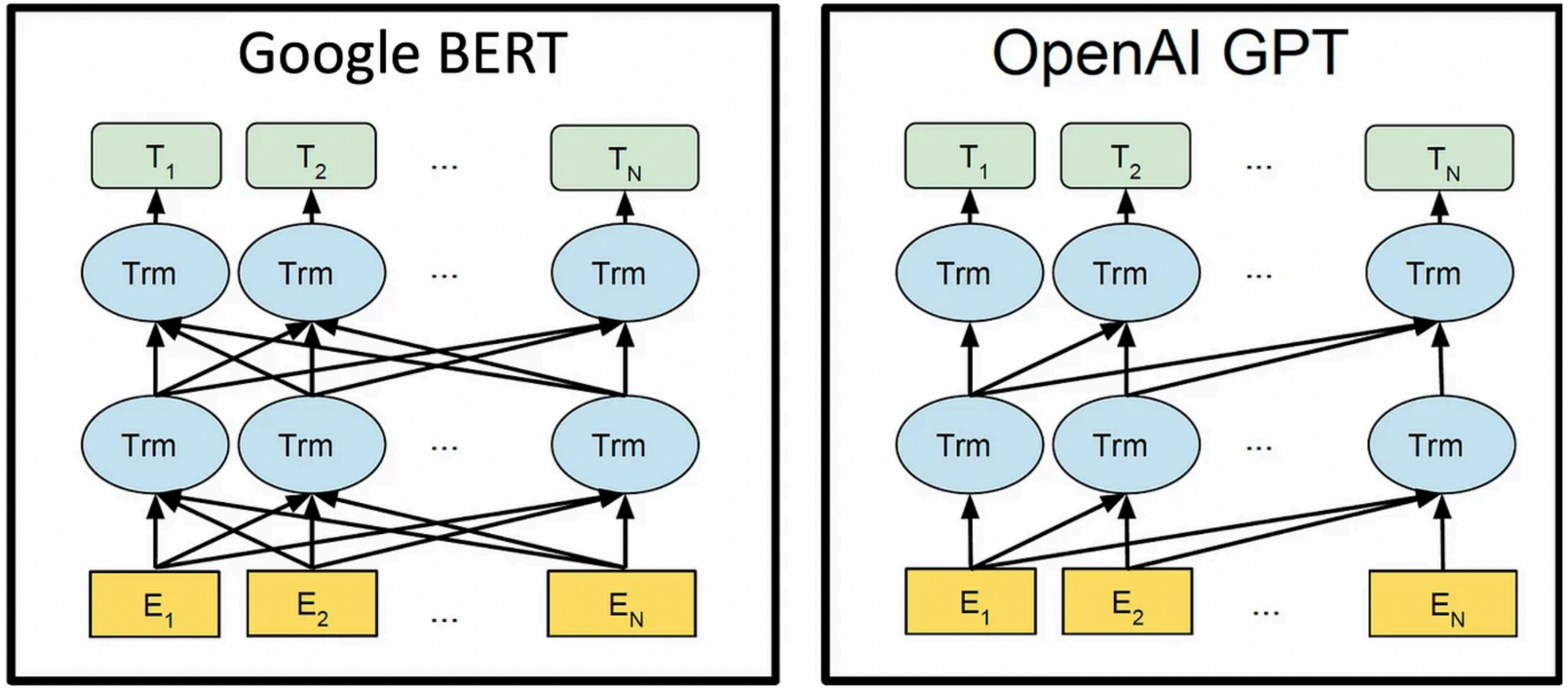
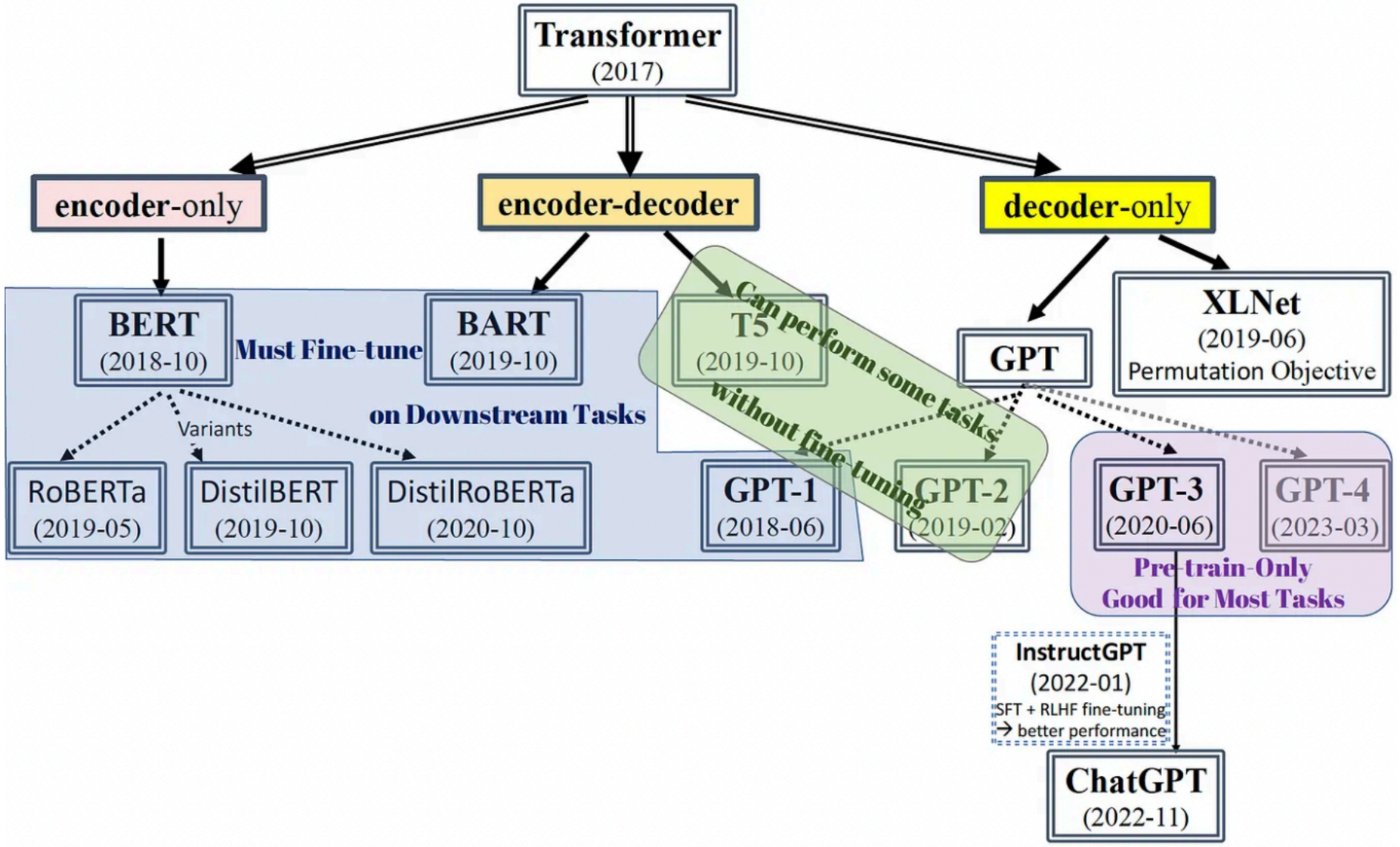


Fig. 2: BERT vs GPT — typical autoencoding (AE), encoder-only, bidirectional model vs typical autoregressive (AR), decoder-only, left-to-right model. Note that this figure is only qualitatively correct. To accurately understand transformer encoder and decoder structures, please read my blog "Step-by-Step Illustrated Explanations of Transformer". (Image Source: Devlin, et. al., 2018)

Encoder-only are great at understanding the context of a sentence by reading all the words at once. On the other hand, decoder-only models, like GPT, focus on generating text. They predict the next word in a sentence, so they're used when you want a model to create or continue a piece of text. In short, encoders are better at understanding, while decoders are better at generating

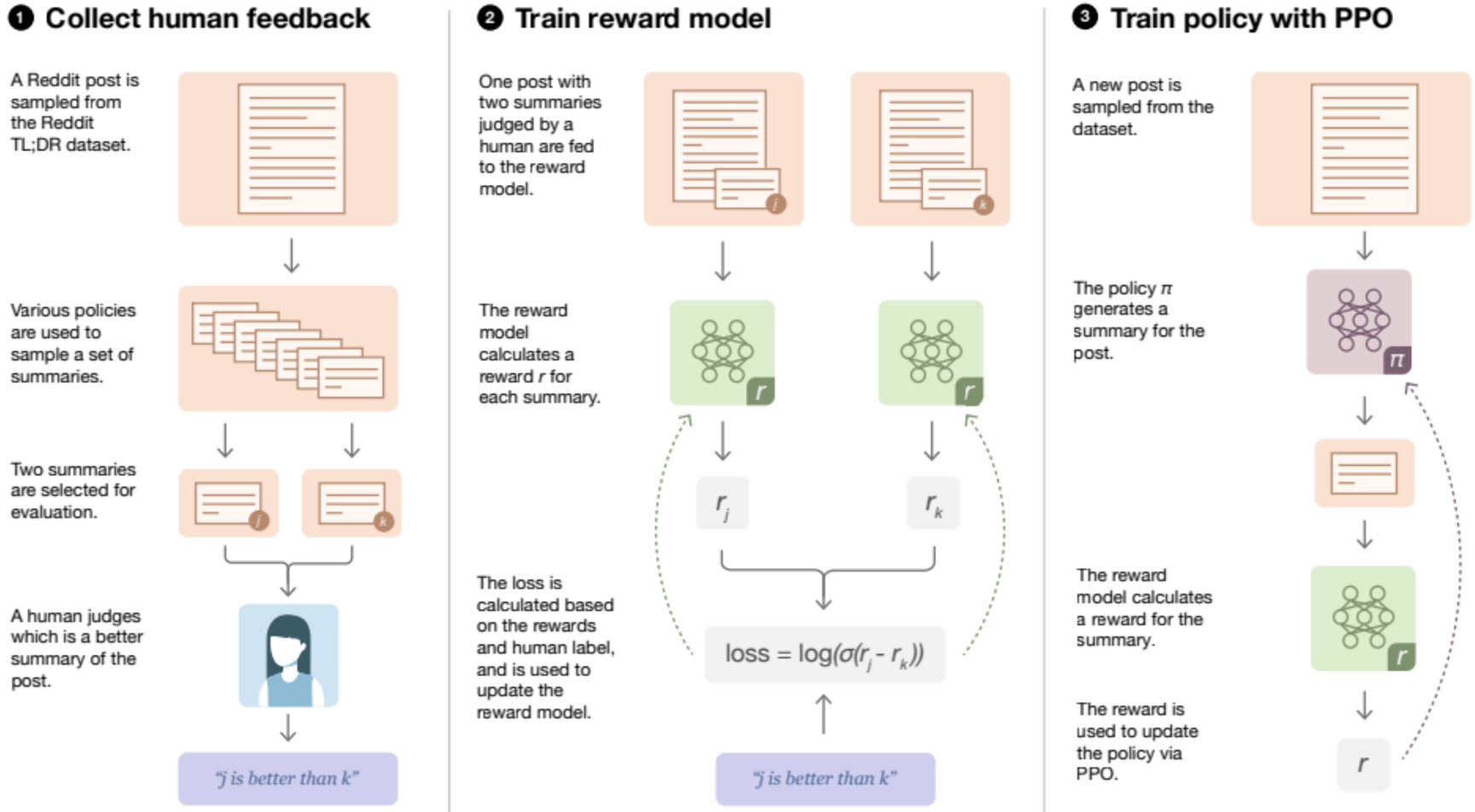
Summary of various language models.

Image Source



Reinforcement Learning from Human Feedback (RLHF) 2022 - GPT3.5

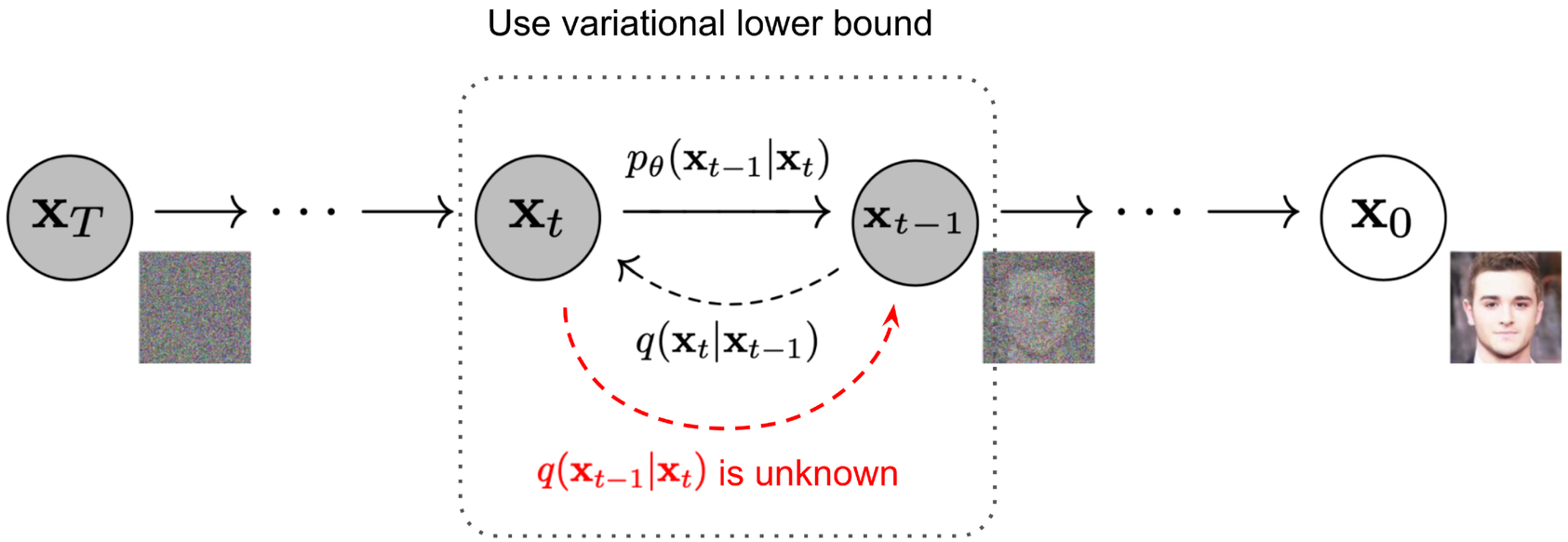
Image Source



RLHF combines reinforcement learning with human feedback to make AI models better aligned with our values and preferences. It's been key in fine-tuning large language models, especially in conversational AI, ensuring they give not only accurate but also helpful, safe, and ethical responses.

Learning to Summarize with Human Feedback by Stiennon et al.
Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback by Bai et al.

Image Source



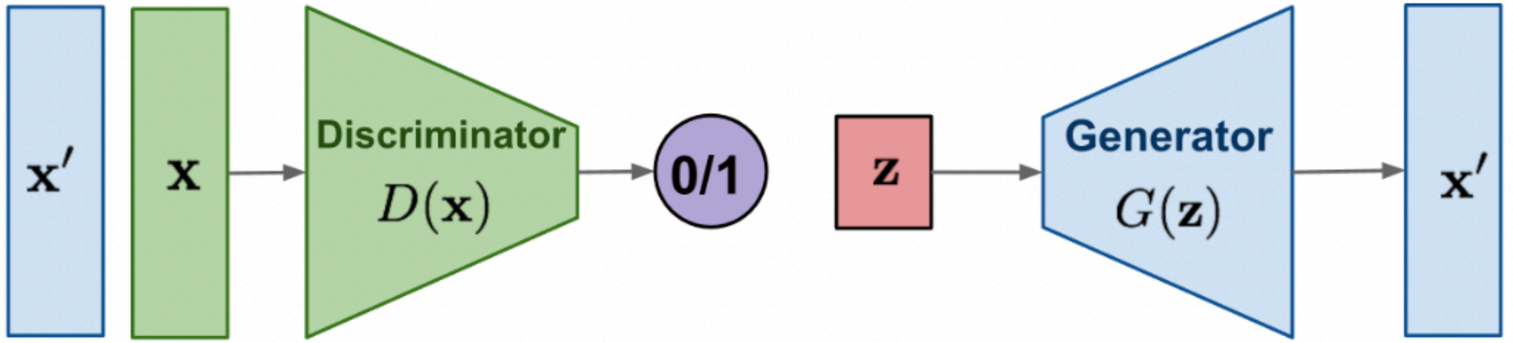
Diffusion models are a powerful type of generative model that create high-quality images by gradually removing noise from a signal. They've become popular for image generation, often producing more diverse and higher-quality results than GANs.

Denoising Diffusion Probabilistic Models by Jonathan Ho et al.
Improved Denoising Diffusion Probabilistic Models by Nichol and Dhariwal.

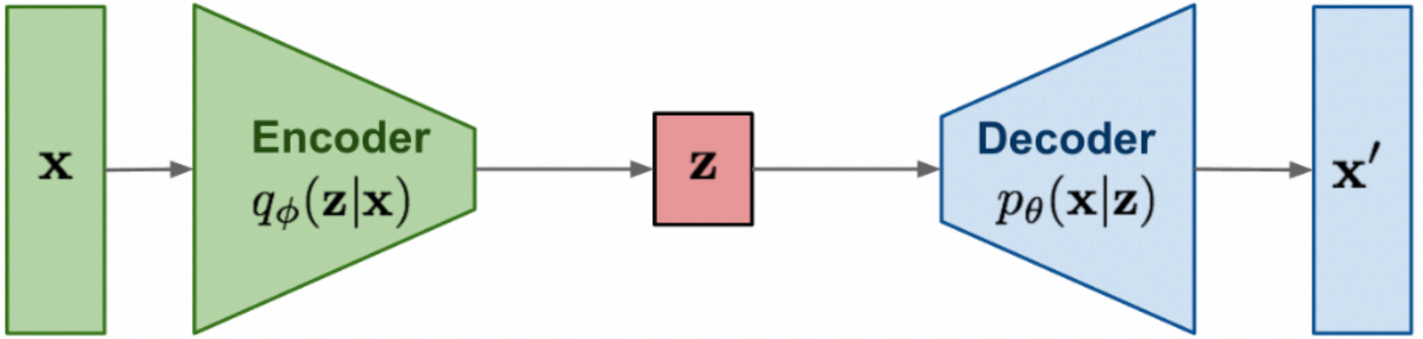
Comparison-GANs-VAE-Diffusion

Image Source

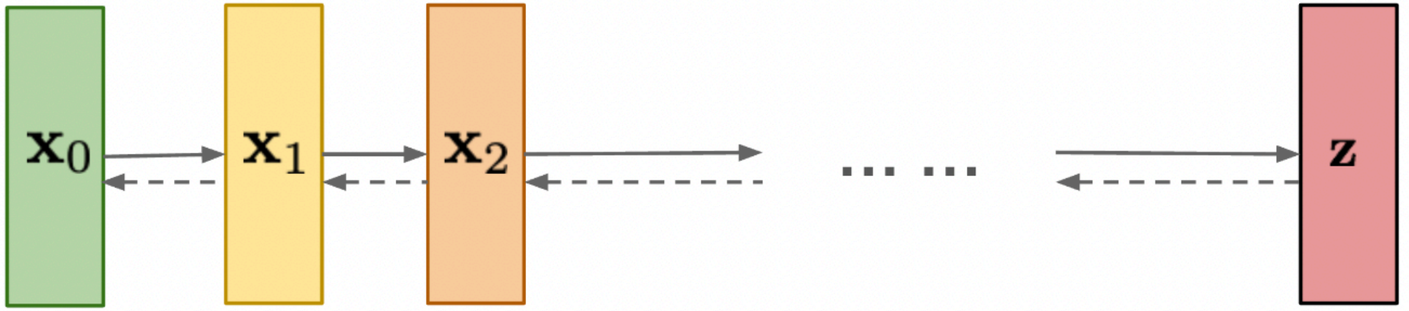
GAN: Adversarial training



VAE: maximize variational lower bound

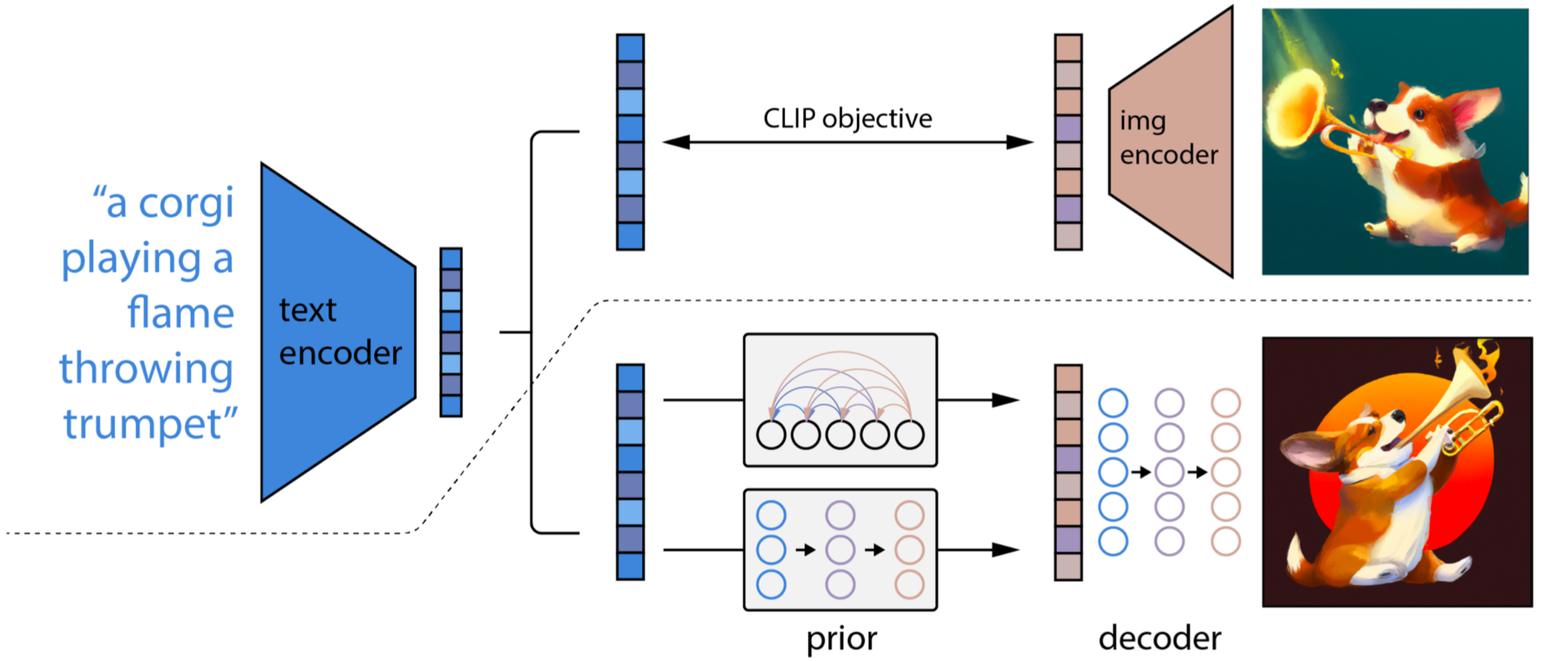


Diffusion models: Gradually add Gaussian noise and then reverse



Text-to-Image Product: DALL-E and CLIP (2021)

Image Source

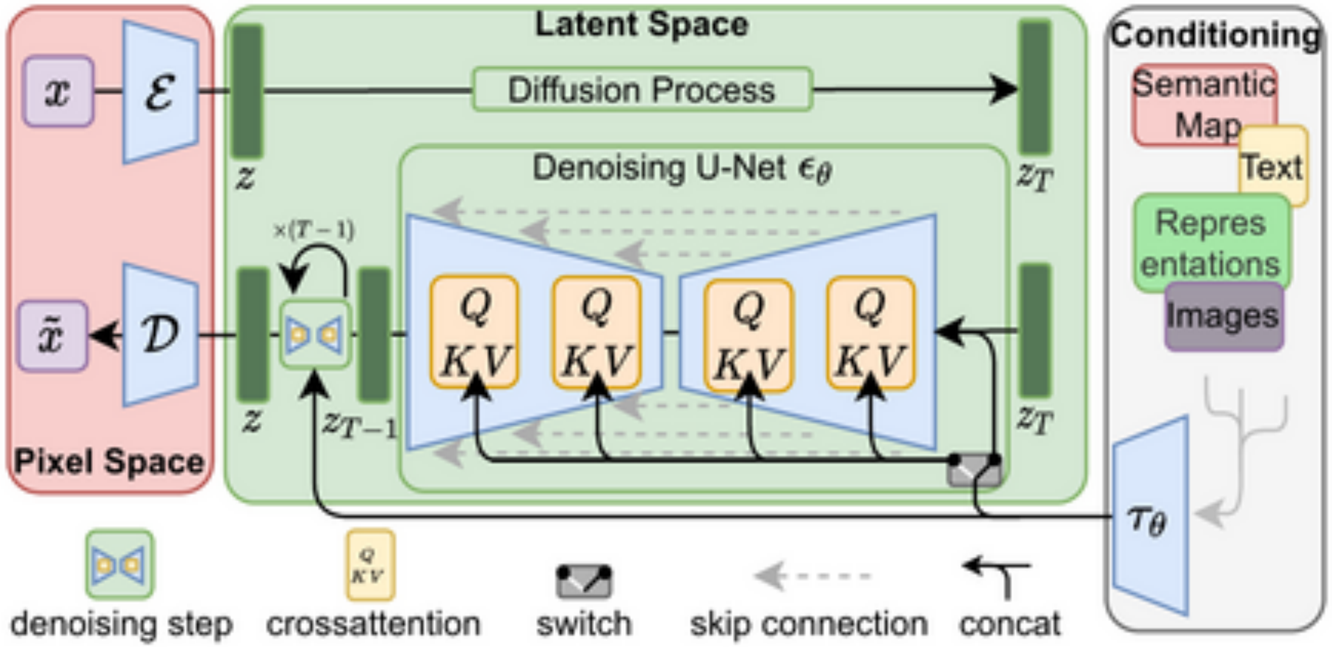


When we talk about image generation, a key topic is 'Text-to-Image.' One well-known example is DALL-E, developed by OpenAI, which creates unique images based on text descriptions. It uses a transformer model to generate high-quality images that match the text. CLIP, introduced with DALL-E, helps the model better understand the connection between images and text, making the results even more accurate.

Learning Transferable Visual Models From Natural Language Supervision (CLIP) by Radford et al.
Zero-Shot Text-to-Image Generation (DALL-E) by Ramesh et al.

Text-to-Image: Stable Diffusion (2022)

Image Source



The open-source project Stable Diffusion gives us a great chance to understand how text-to-image models work. It's a latent diffusion model that generates high-quality images efficiently. Unlike earlier models, it operates in a lower-dimensional space, which means it uses less computing power.

High-Resolution Image Synthesis with Latent Diffusion Models by Rombach et al.

Thanks