1/15 4:40:03 ***

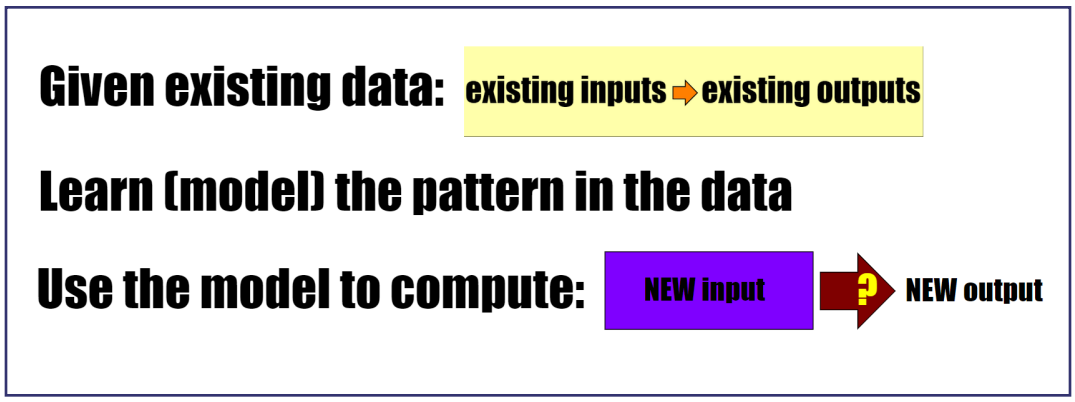
← →

genAI

[generative AI]

Good old data mining ['GODaM' :)]...

So far, we have studied many data mining algorithms (including NNs), which all learn (model) patterns (between outputs and inputs) in existing data, and use that to classify/calculate new outputs based on new inputs.



So... what's this 'genAI' thing?

'Generative AI' - a revolution

Generative AI, very loosely speaking, 'runs a neural network BACKWARDS'!

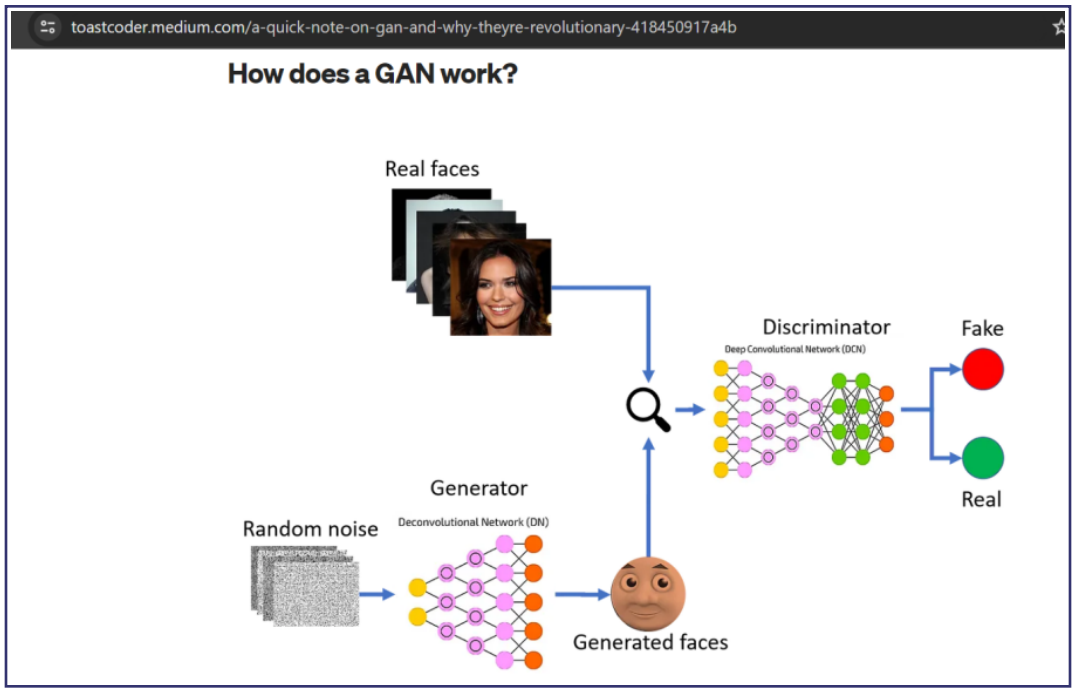
Rather than learn to classify new data using existing data, why not GENERATE new data instead?

Researchers tried this, but with unimpressive results.

In 2014, Ian Goodfellow got a much better idea than the 'SOTA' - why not pair up TWO NNs in opposing order - one a generator (eager 'student'), and the other, a discriminator (strict 'teacher')? His invention is called a 'GAN'.

GAN

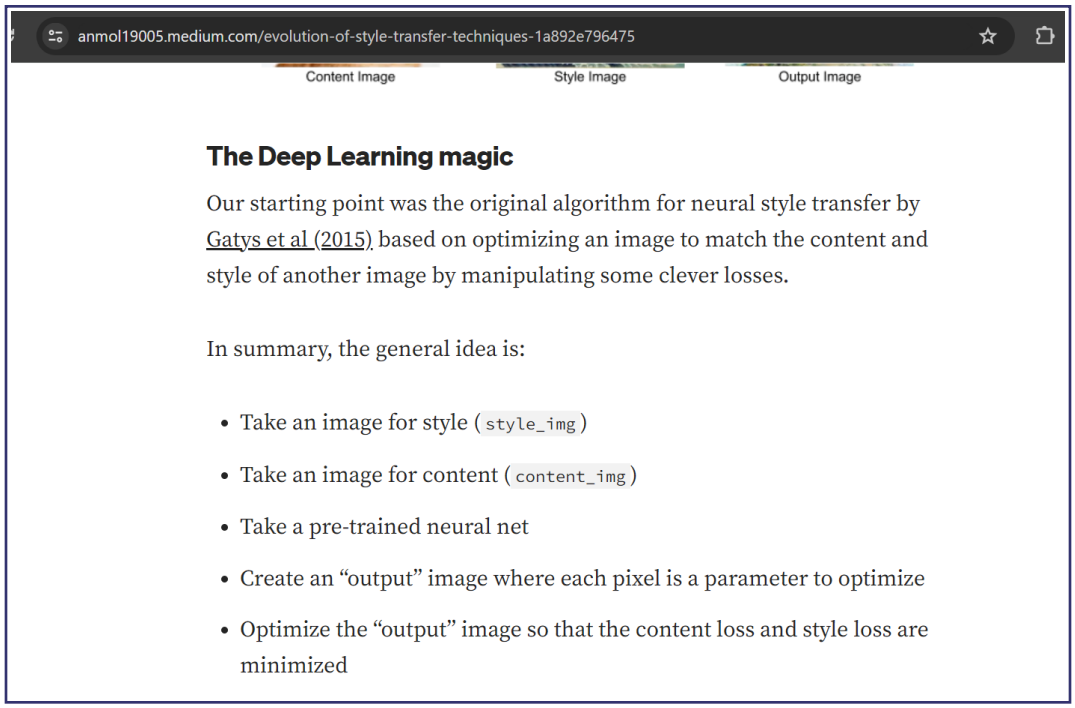
GAN stands for Generative Adversarial Network

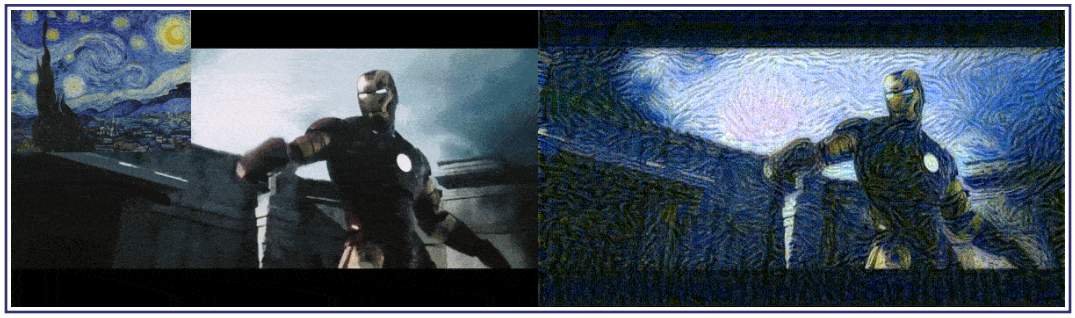


More: https://developers.google.com/machine-learning/gan/gan_structure and https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/

Style transfer

An early use of genAI was/is to "add style" to imagery: https://arxiv.org/abs/1508.06576



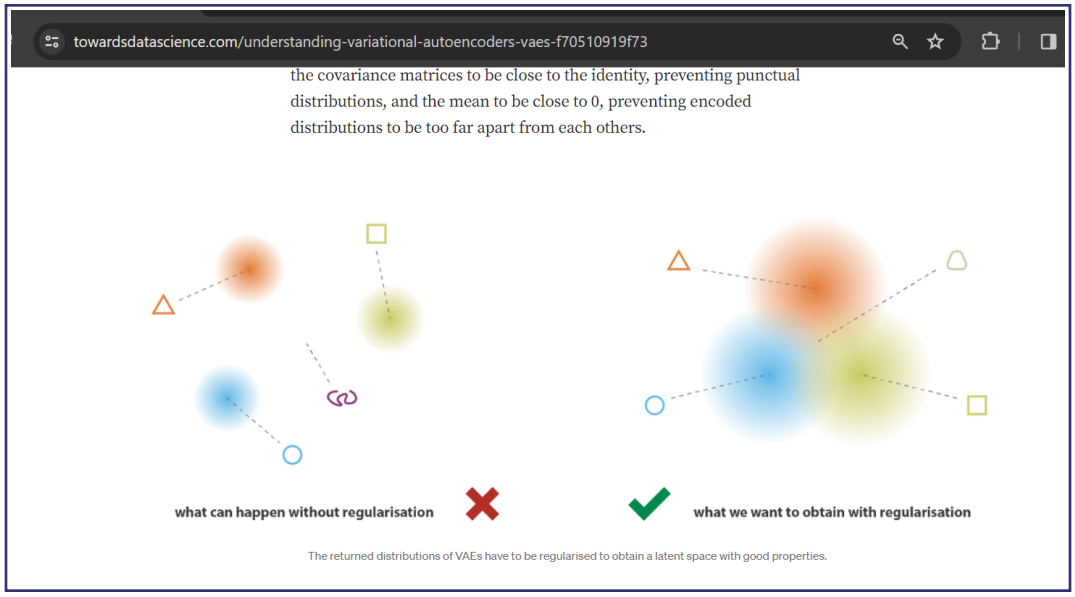


Encoders/decoders, autoencoders, VAEs

Encoder: a function (NN) that maps original input to LATENT/ENCODED space [decoder is the reverse]

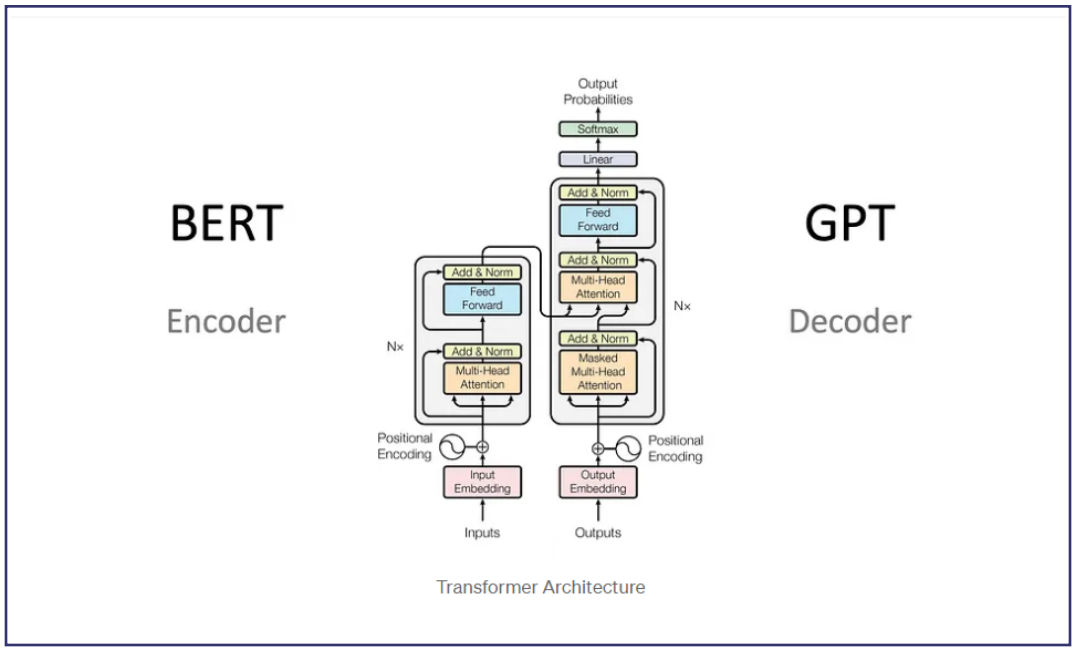
Autoencoder: encoder + decoder combination - can GENERATE NEW OUTPUT (by interpolating a random point in latent space)!

Variational AE: the encoder produces a distribution rather than a single point [and the decoder uses a sampled point from the distribution].



Transformers

'Attention is all you need': https://arxiv.org/abs/1706.03762



Revolution: non-fixed size and parallelizable self-attention mechanism (ie. computing word affinities).

More:

- https://jalammar.github.io/illustrated-transformer/
- https://medium.com/@yulemoon/detailed-explanations-of-transformer-step-by-step-dc32d90b3a98
- https://www.youtube.com/watch?v=zjkBMFhNj_g

The decoder takes a prompt (new point in latent space), INTERPOLATES over inputs (ALL English!!), generates output.

Transformers to... ChatGPT!

Here is a great explanation of ChatGPT: https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/

The core pre-trained LLM needs to be augmented with HFRL [a form of fine-tuning], to produce acceptable responses.

GPT extensions

- multimodal (eg text, images)
- plugin API
- GPT Store
- LLM apps, eg. https://blog.llamaindex.ai/introducing-llama-packs-e14f453b913a [and https://llamahub.ai/]

LLM extensions

- larger context [eg. RMT: https://arxiv.org/abs/2304.11062, and https://hazyresearch.stanford.edu/blog/2023-03-07-hyena and https://bdtechtalks.com/2023/11/27/streamingllm/amp/]
- infinite memory! [https://medium.com/@jordan_gibbs/how-to-create-your-own-gpt-voice-assistant-with-infinite-chat-memory-in-python-d8b8e93f6b21]
- architecture alterations (eg Rethinking Attention: https://arxiv.org/abs/2311.10642), alternate position encodings [including NO: https://arxiv.org/pdf/2203.16634.pdf, context-aware, rotary encoding...]
- open source!
- quantization [of weights to 16/8...bits, to compress model size]
- SLMs! [latest: https://mistral.ai/, https://starling.cs.berkeley.edu/, Orca 2...] (almost all at https://huggingface.co/models)

Prompt extensions/ LLM frameworks

- LangChain
- LlamaIndex
- Haystack
- https://cassio.org/
- ...
- CoT [chain of thoughts], ToT [tree of thoughts]...
- LangChain Expression Language [LCEL]
- agents!

Response extensions

- fine-tuning, eg. https://magazine.sebastianraschka.com/p/practical-tips-for-finetuning-llms and https://www.databricks.com/blog/efficient-fine-tuning-lora-guide-llms
- **RAG**!! Two kinds (eg. https://ai.plainenglish.io/beyond-tables-and-vectors-knowledge-graphs-for-ai-reasoning-46f0f8721894), more kinds... [eg. https://artificialcorner.com/ive-created-a-custom-gpt-that-scrapes-data-from-websites-9086aff58105]
- vector DBs https://medium.com/@zilliz_learn/what-is-a-real-vector-database-b391b0468279 and https://tiledb.com/blog/why-tiledb-as-a-vector-database
- LMSQL! https://towardsdatascience.com/lmql-sql-for-language-models-d7486d88c541

Custom GPTs

As a result of the extensions listed above, there is bound to be numerous, narrow-purpose GPTs, eg. https://chat.openai.com/g/g-kCfSC3b10-analystgpt - it's in a way, back to 'expert systems' AI from the mid-80s and early 90s :)

Here is an architecture useful for building at-scale RAG apps: https://www.pinecone.io/learn/aws-reference-architecture/

genAI: other (non-plaintext) content

- CODE!!!
- images
- video
- music
- 3D CG
- ...

Issues?

https://www.theguardian.com/books/2023/nov/15/hallucinate-cambridge-dictionary-word-of-the-year

GenAI Against Humanity: https://arxiv.org/pdf/2310.00737.pdf