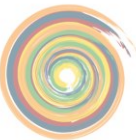# Generative AI

Vasant Jain

# Generative Adversarial Networks (GANs)

HISTORICALLY

- GANs is the first example of Generative AI, but came along with shortcomings as:

- GAN model Bottlenecks

  - lack of diversity in Image Generation, Mode Collapse, problem in learning Multimodal distribution, High Training time, Not easy to train

Generative AI

- **Thinking beyond what is already present in data**, leading to the concept of Generative Modeling.

- Overarching theme
  - **The belief is that the clearest framing of general intelligence is a system that can do anything a human can do in front of a computer.**
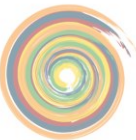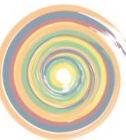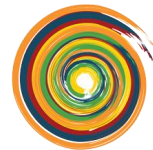
Model Evolution to Transformers

1. **Feed-forward networks**

2. **RNNs**: simplest networks that can deal with sequential data

3. **LSTMs**: special class of RNNs that can have a longer short-term memory compared to vanilla RNNs.

4. **TRANSFORMERS**: The NN architecture that made ChatGTP and other LLMs possible.
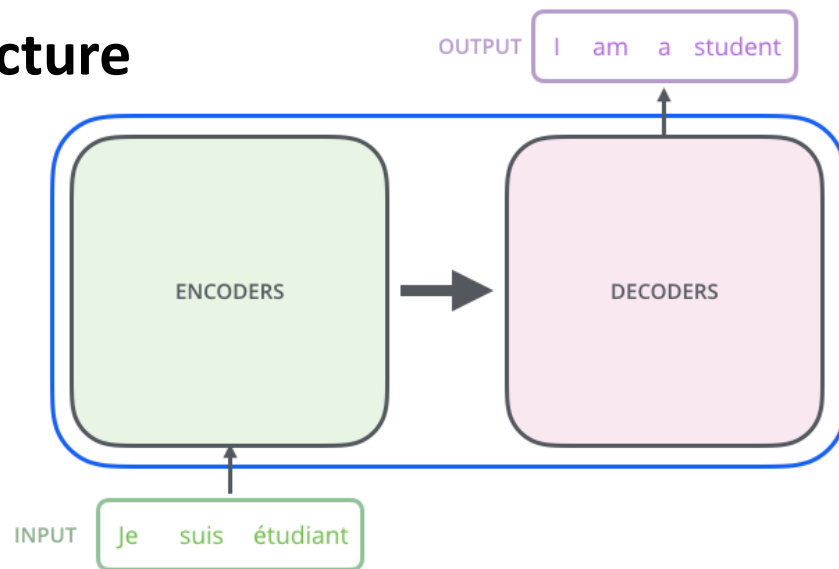
SOURCE: Jay Alammar

Transformer Architecture

Looking under the hood

High Level Look

Simple Transformer Architecture

SOURCE: Jay Alammar

Transformer architecture (stacked model)

OUTPUT | I am a student |

ENCODER → DECODER
ENCODER → DECODER
ENCODER → DECODER
ENCODER → DECODER
ENCODER → DECODER
ENCODER → DECODER

INPUT | Je suis étudiant |
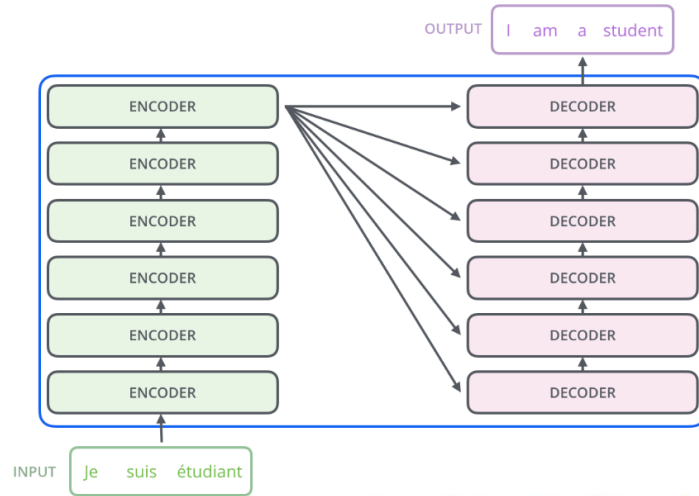
Neural Machine Translation
SEQUENCE TO SEQUENCE MODEL WITH ATTENTION
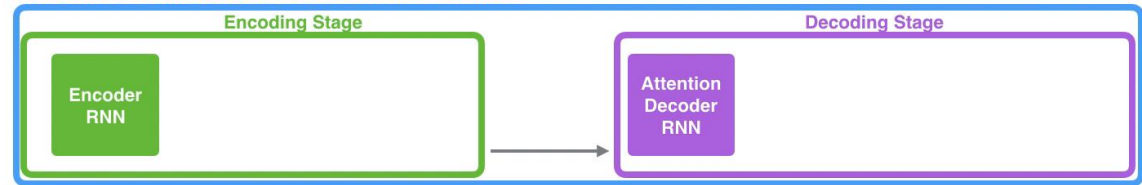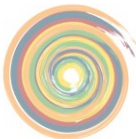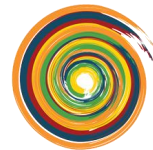
Encoding Stage | Decoding Stage

Encoder RNN → Attention Decoder RNN

Je   suis   étudiant

SOURCE: Jay Alammar

ATTENTION

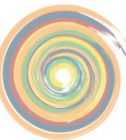|	I	Am	a	Student
je	1	0	0	0
suis	0	1	1	0
étudiant	0	0	0	1

Encoder hidden state

Je — hidden state #1

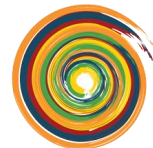suis — hidden state #2

étudiant — hidden state #3

SOURCE: Jay Alammar

Concept Building (2/3)
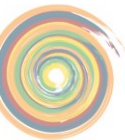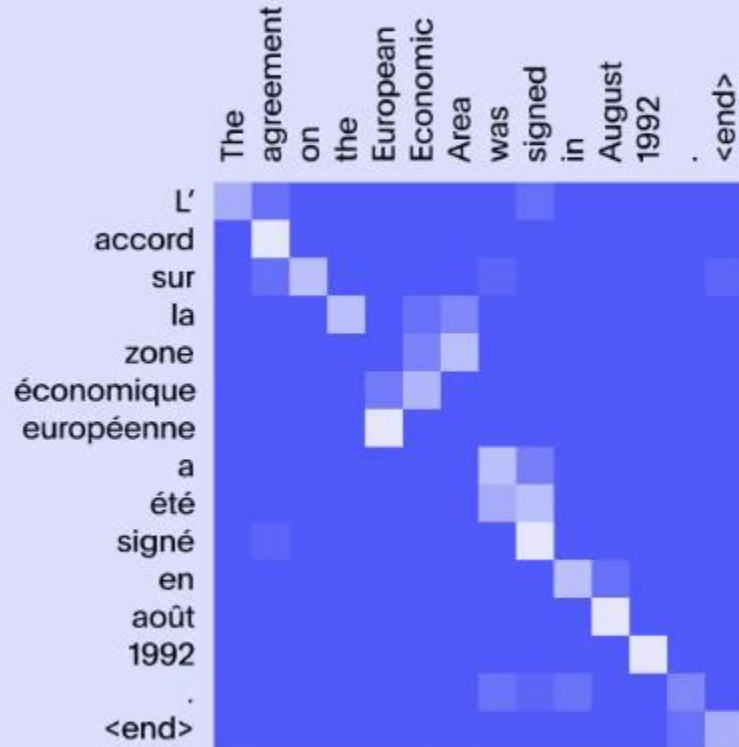
What is transformer architecture?

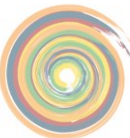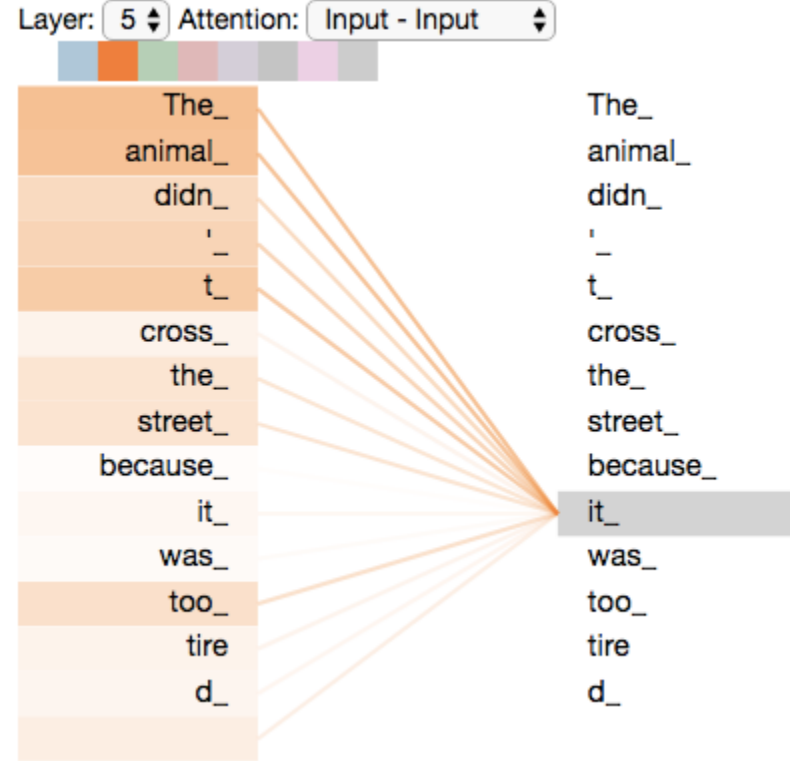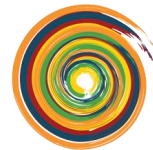# Let's Pay Attention Now

- Attention allows the model to focus on the relevant part of the input sequence as needed.

- Say the following sentence is an input sentence we want to translate:

- "The animal didn't cross the street because it was too tired"

- *Self-attention is the method the Transformer uses to bake the "understanding" of other relevant words into the one we're currently processing.*
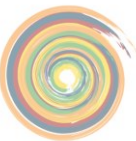
Self Attention Concept

- The first step in calculating self-attention is to create three vectors from each of the encoder's input vectors (in this case, the embedding of each word). So for each word, we create a Query vector, a Key vector, and a Value vector.

- The second step in calculating self-attention is to calculate a score.

- The third and fourth steps are to divide the scores by 8

- The fifth step is to multiply each value vector by the softmax score

- The sixth step is to sum up the weighted value vectors.

Layer: 5 | Attention: Input - Input

The_ animal_ didn_ '_ t_ cross_ the_ street_ because_ it_ was_ too_ tire d_

The_ animal_ didn_ '_ t_ cross_ the_ street_ because_ it_ was_ too_ tire d_

Uses of Transformers

Transformers

- Language
  - GPT-3 / 4
  - PaLM
  - Chinchilla
- Code
  - Codex
  - AlphaCode
- Image Generation
  - DALL-E
  - Imagen

The next era of computing will be defined by the natural language interfaces that allows us to tell our computers what we want directly, rather than doing it by hand.

Looking ahead

- Most interactions with computers will be done using natural language, not GUIs.

- **We'll tell our computer what to do, and it'll doit.**

- Today's user interfaces will soon seem as archaic as landlines phones do to smartphone users.

- We will never search through forums on "how to do X in Salesforce or Unity or Figma" -- **the model will do that work, allowing us to focus on the higher-order task at hand.**
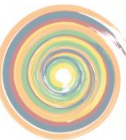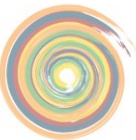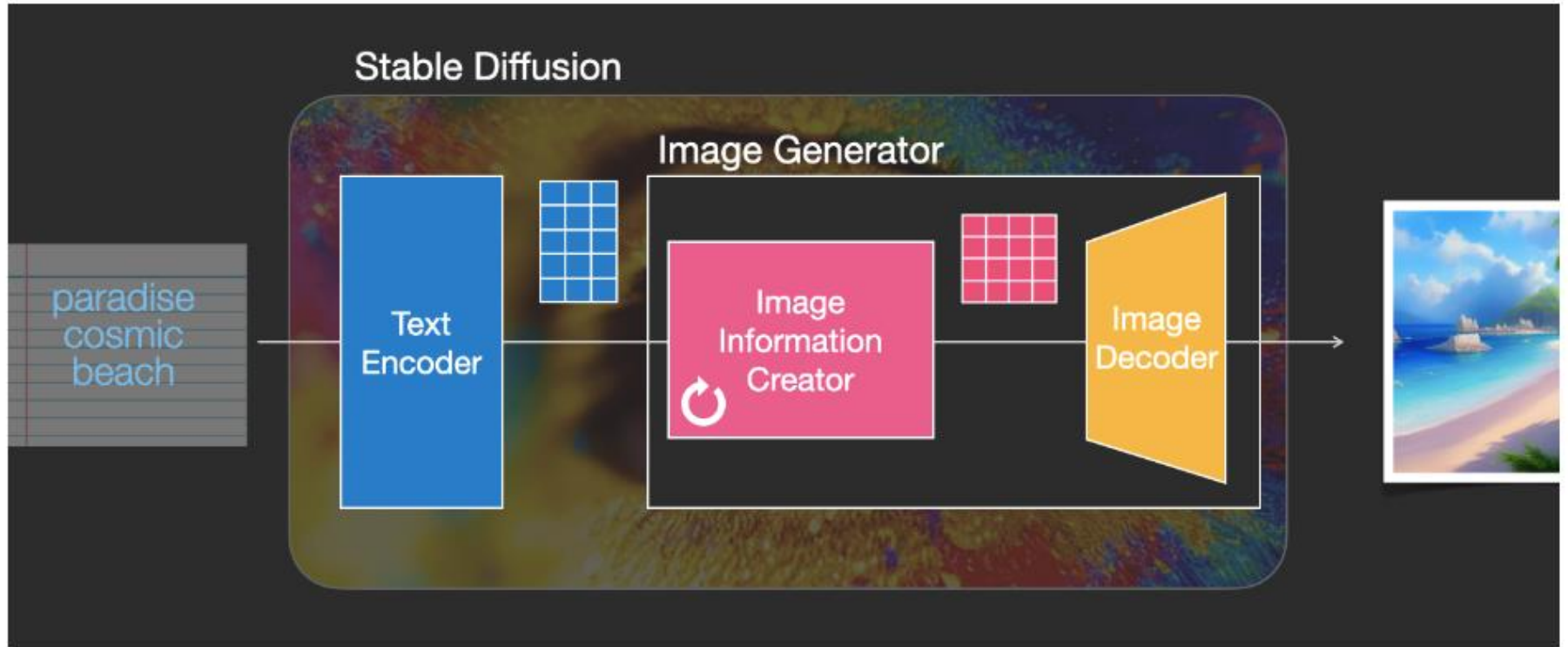
Stable Diffusion: Architecture and Examples

The ability to create striking visuals from text descriptions has a magical quality to it and points clearly to a shift in how humans create art.
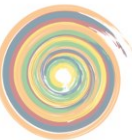
The release of Stable Diffusion is a clear milestone in this development because it made a high-performance model available to the masses.
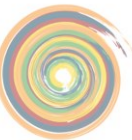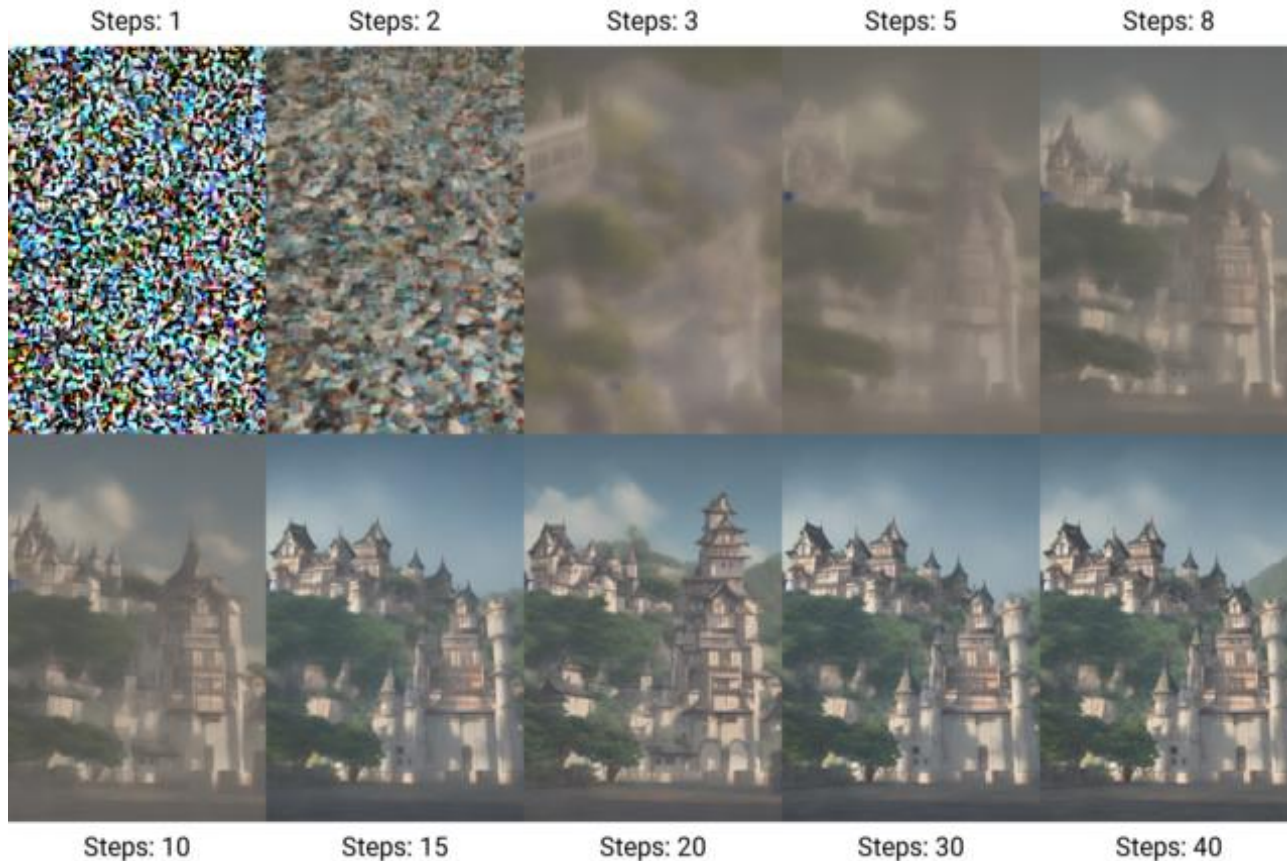
Stable Diffusion Architecture

- Special class of RNNs that can have a longer short-term memory compared to vanilla RNNs
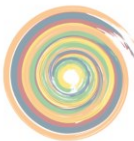
Image appearing from noise

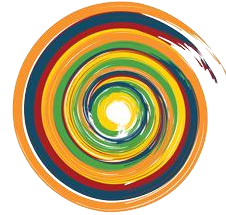# An example of denoising process used by Stable Diffusion

| Steps: 1 | Steps: 2 | Steps: 3 | Steps: 5 | Steps: 8 |
| Steps: 10 | Steps: 15 | Steps: 20 | Steps: 30 | Steps: 40 |

- Used a lot in Anomaly Detection, based on Teacher-Student Learning Model.

Myelin Foundry

Everything AI, Anything Creative

Thank You

Concept Building (3/3)

- Interestingly, if the words are changed slightly the attention matrix also changes.

  - For example, in the sentence,

  - "The lamp could not be packed into the suitcase because *it* was too large" the word "it" refers to the lamp.

  - But in the sentence,

  - "The lamp could not be packed into the suitcase because *it* was too small." The word "it" refers to the suitcase.

- So, the only way to obtain these attention matrices would be to learn from data.