ML-blement session:

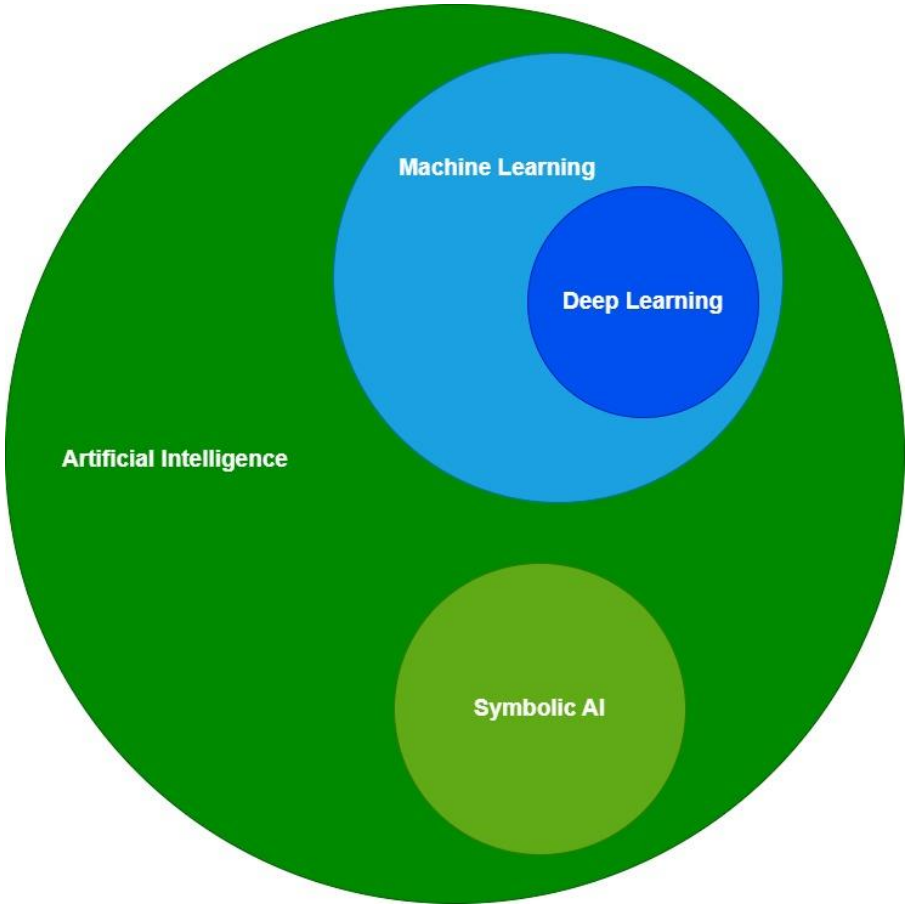
Generative AI

Documented explanations can be found here:
https://github.com/abenslimaneakawahid/generative-AI/blob/main/GenAI_simplified.pdf

Abdelwahid Benslimane
wahid.benslimane@gmail.com

What is generative AI?

- Branch of deep learning
- Produce complex data with a human quality (text, images, music etc.)



Abdelwahid Benslimane
wahid.benlimane@gmail.com

Encoder-decoder architecture/Seq2seq model:

Encoder: takes a sequence as input

Decoder: generates a new sequence

Use cases: translation, text analysis, answering questions etc., in NLP

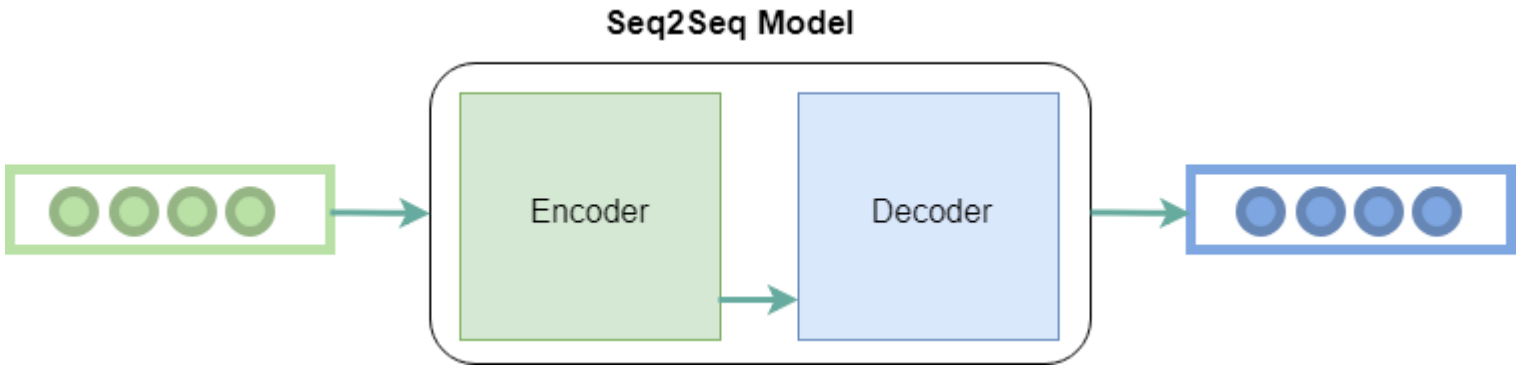


Image taken from https://www.geeksforgeeks.org/seq2seq-model-in-machine-learning/

Abdelwahid Benslimane
wahid.benlimane@gmail.com

Other architectures exist:

Generative adversarial networks (GAN) to (mostly) generate images

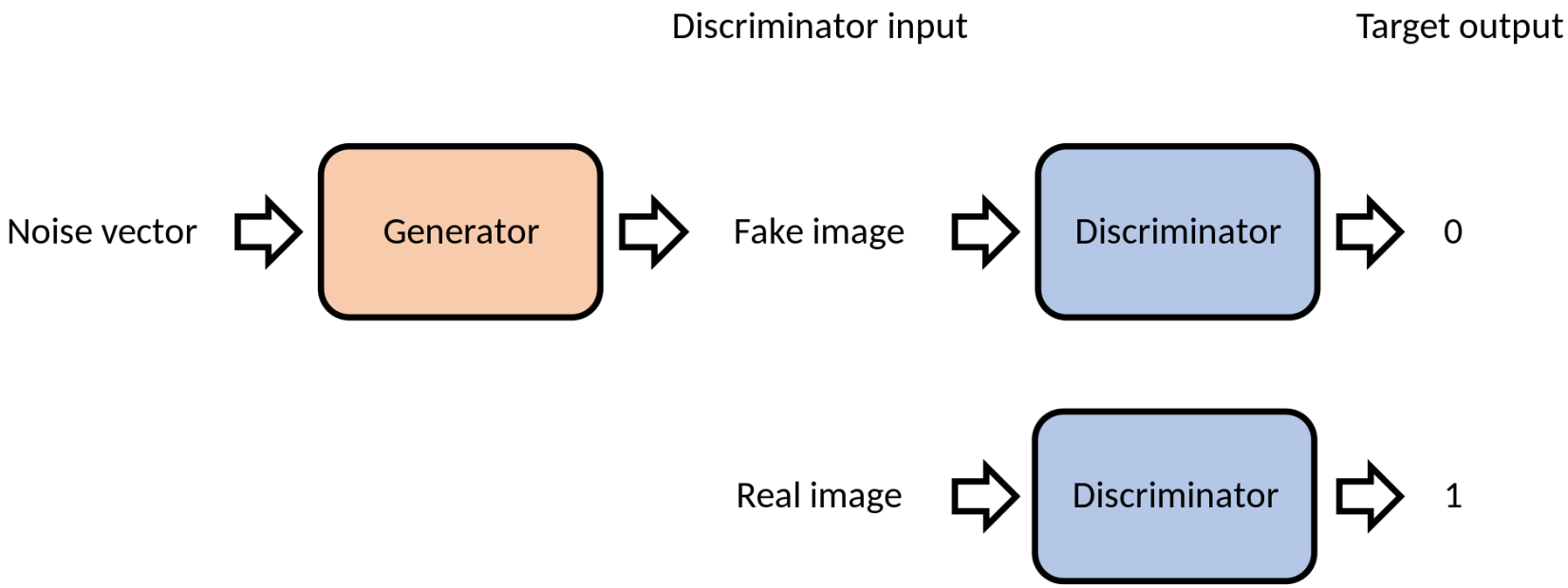


Image taken from https://en.wikipedia.org/wiki/Generative_adversarial_network

Abdelwahid Benslimane
wahid.benlimane@gmail.com

Same high-level design but different implementation:

Variational AutoEncoder (VAE), a kind of non-linear extension of PCA

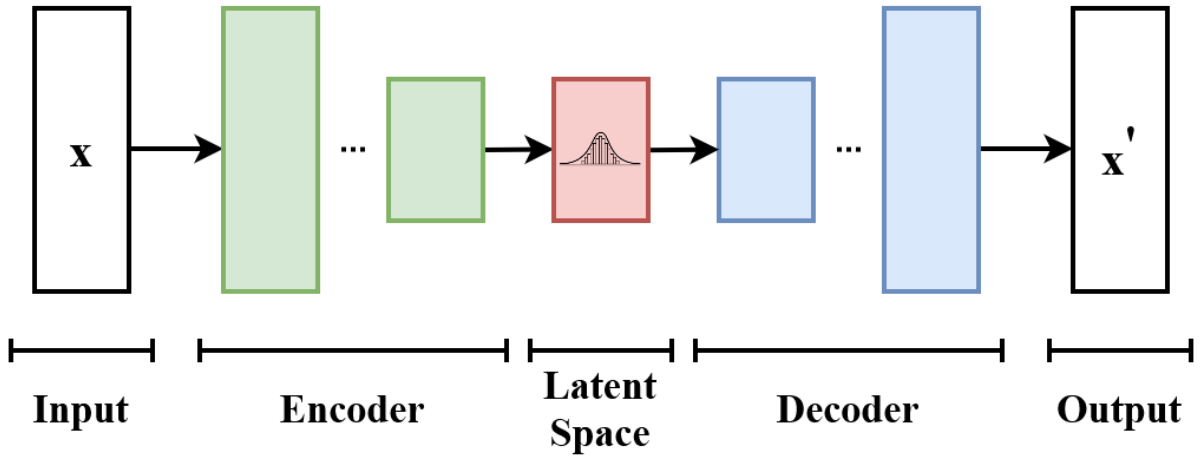


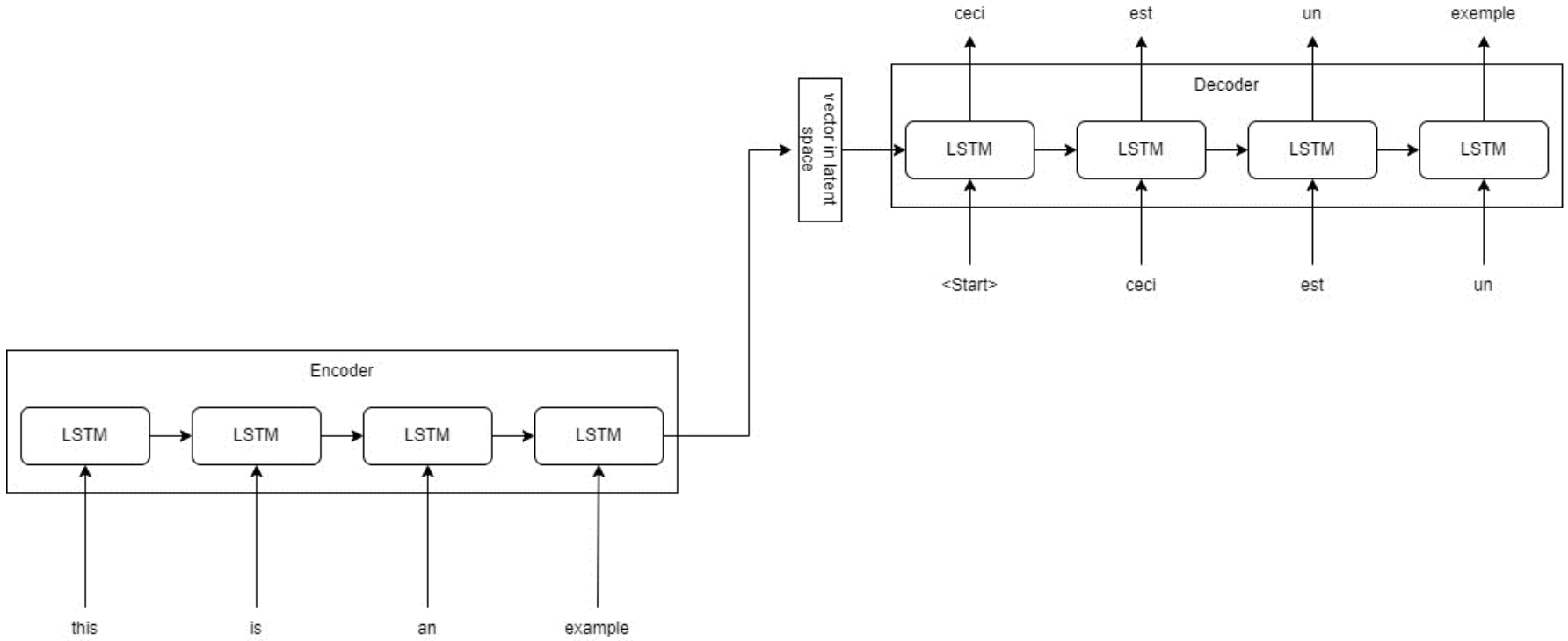
Image taken from https://en.wikipedia.org/wiki/Variational_autoencoder

Use cases: new data generation, anomaly detection, dimensionality reduction etc.

For an advanced explanation: https://github.com/abenslimaneakawahid/ML-and-maths-theory/blob/main/variational_autoencoder.pdf

Abdelwahid Benslimane
wahid.benlimane@gmail.com

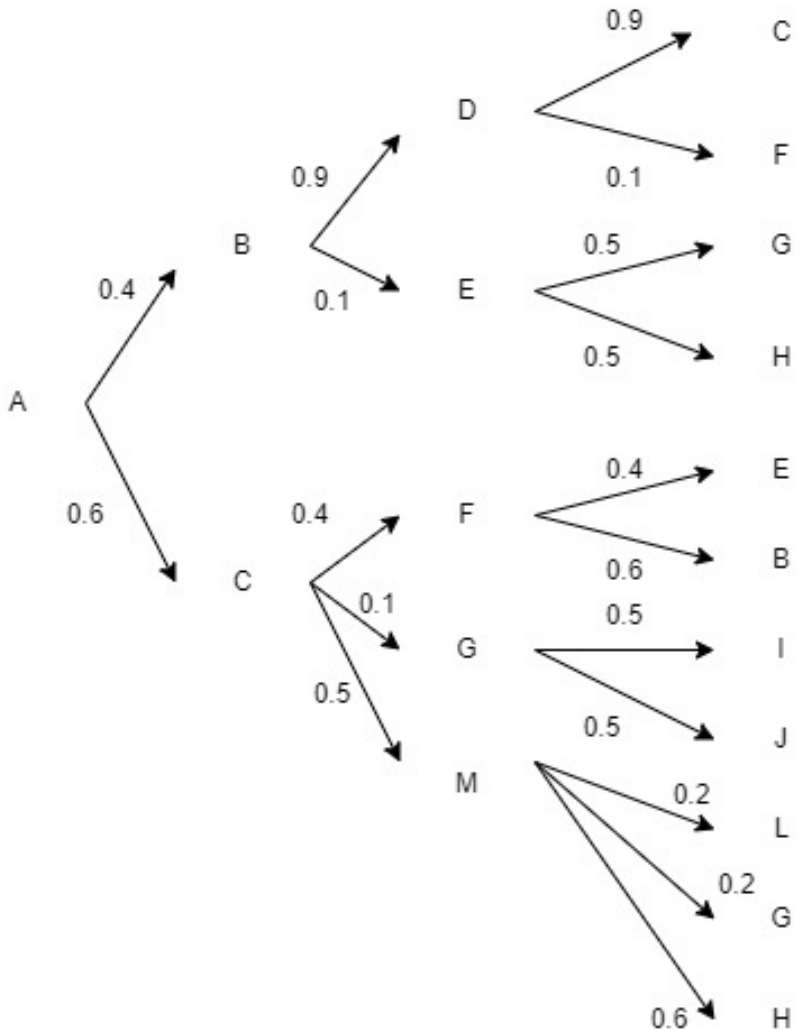
Seq2Sseq model with RNNs



Pre-processing steps: tokenization and word embedding

Abdelwahid Benslimane
wahid.benlimane@gmail.com

Beam Search vs Greedy Search

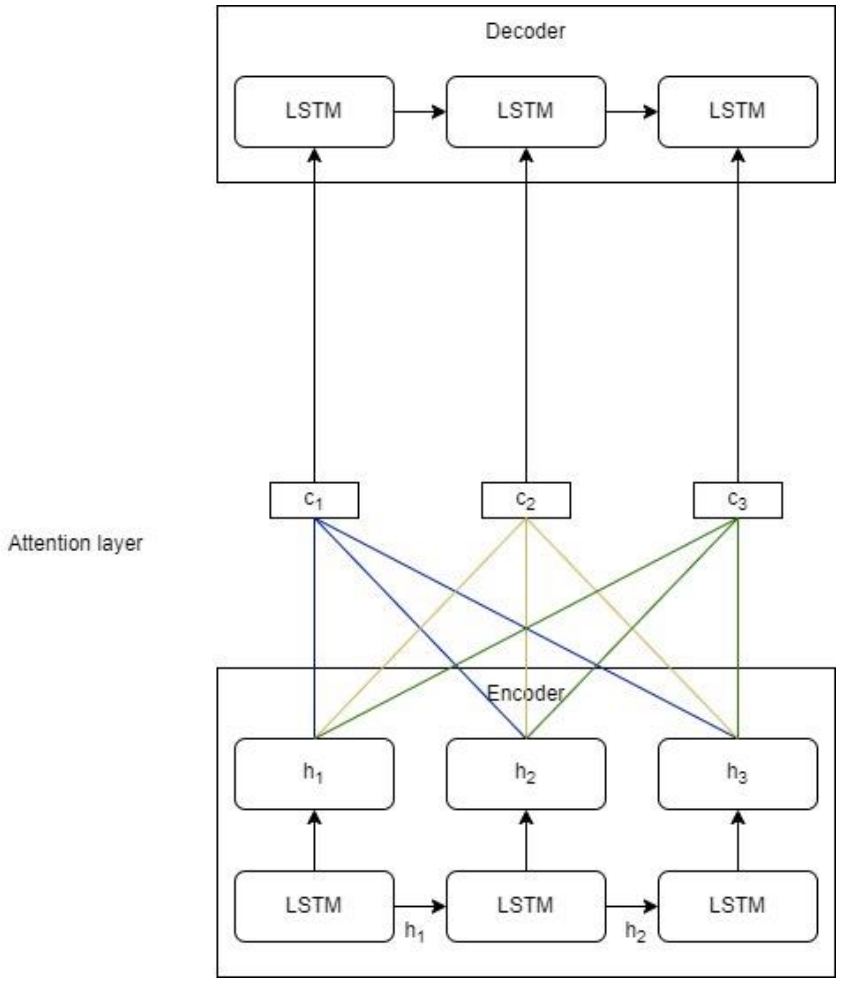


Beam search output: ABDC

Greedy search output: ACMH

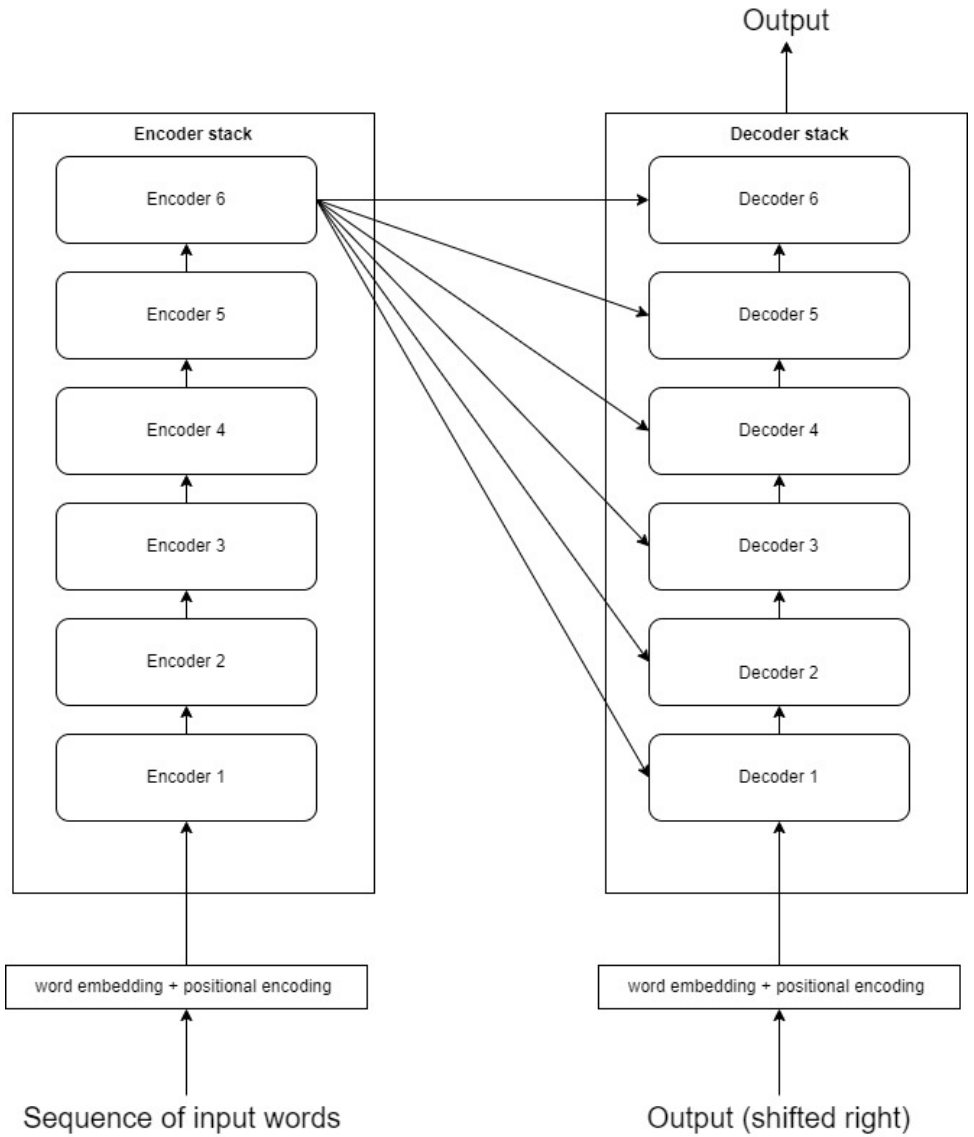
Abdelwahid Benslimane
wahid.benlimane@gmail.com

RNN + Attention



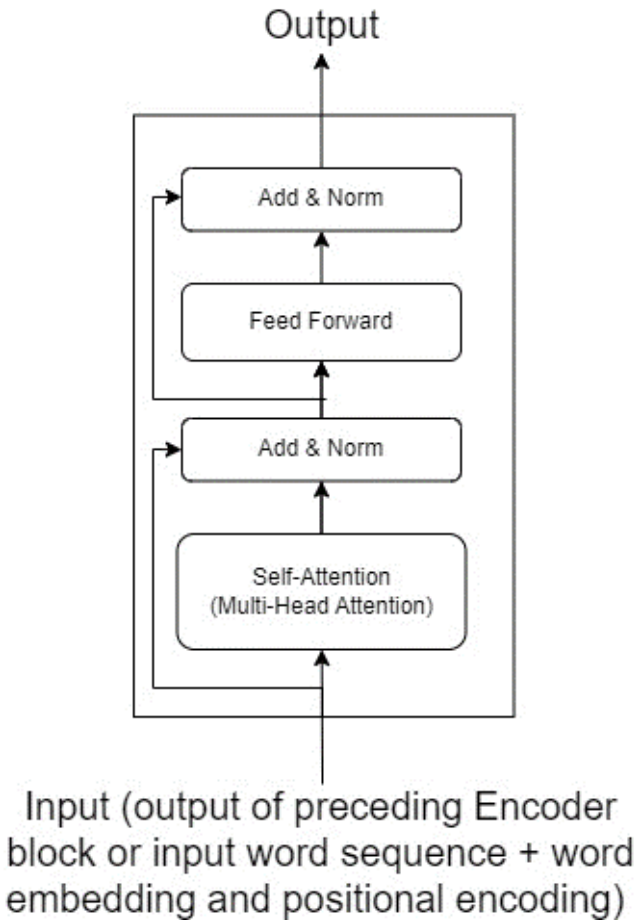
Abdelwahid Benslimane
wahid.benlimane@gmail.com

Transformer’s high-level design



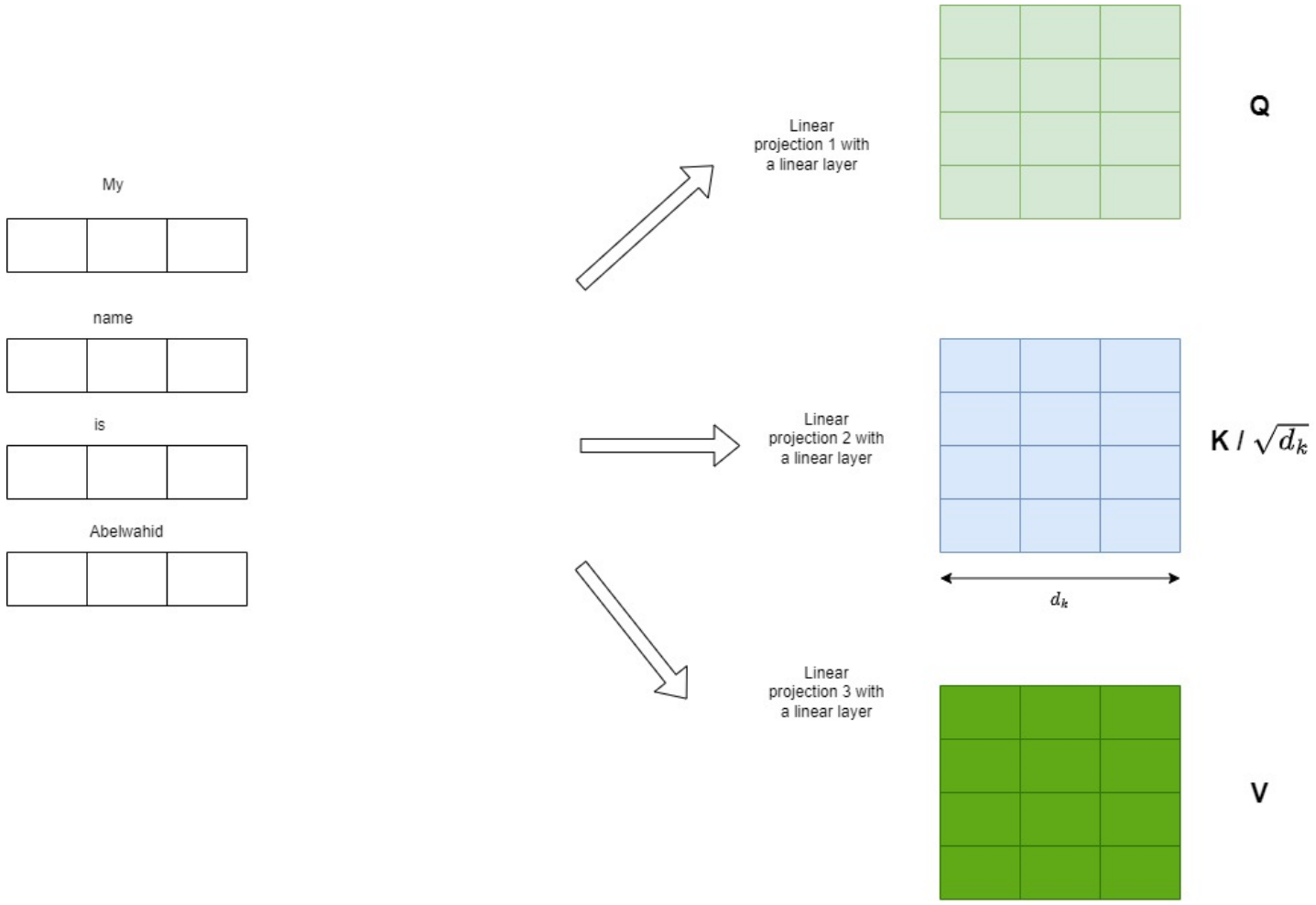
Abdelwahid Benslimane
wahid.benlimane@gmail.com

Encoder block



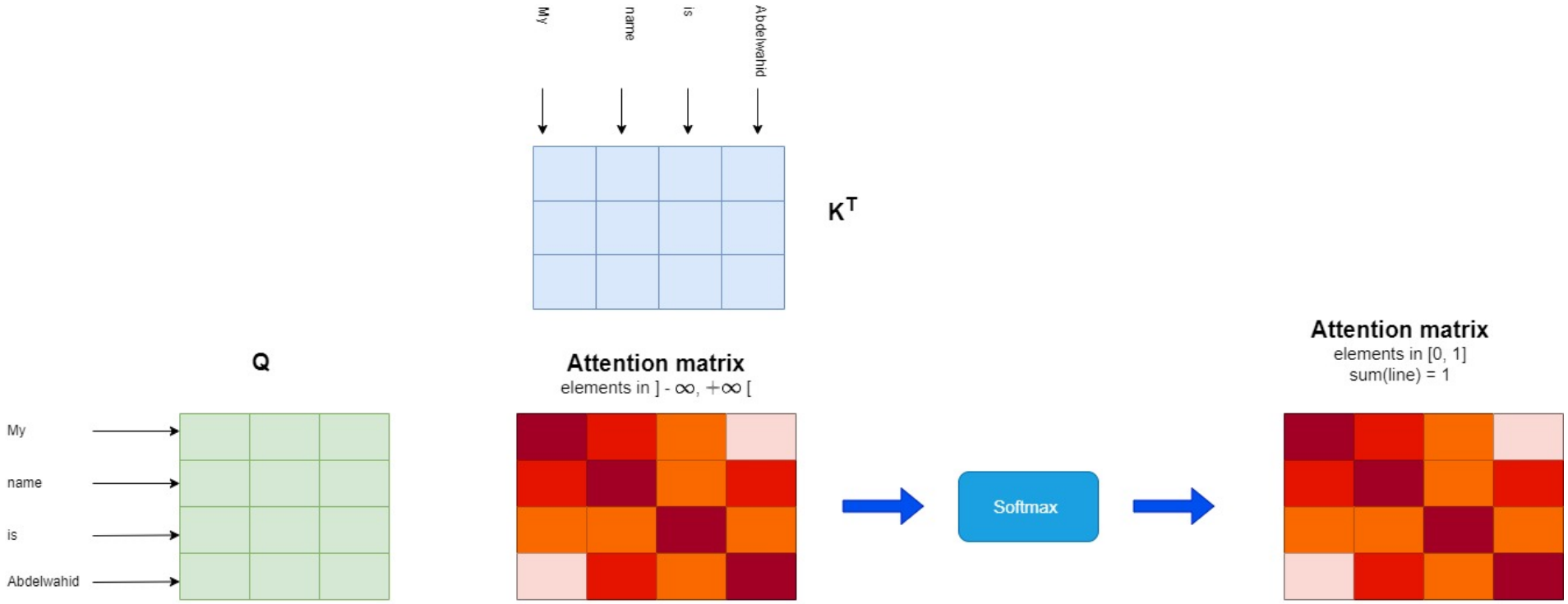
Abdelwahid Benslimane
wahid.benlimane@gmail.com

Self-Attention algorithm 1/3



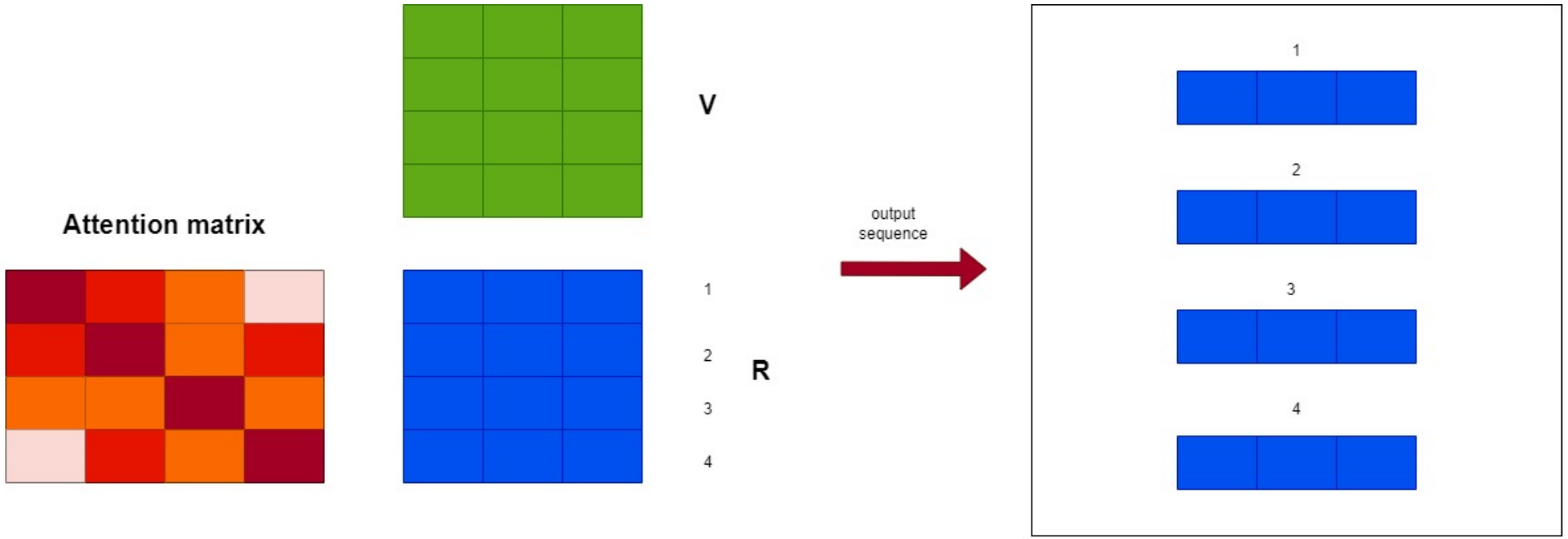
Abdelwahid Benslimane
wahid.benlimane@gmail.com

Self-Attention algorithm 2/3



Abdelwahid Benslimane
wahid.benlimane@gmail.com

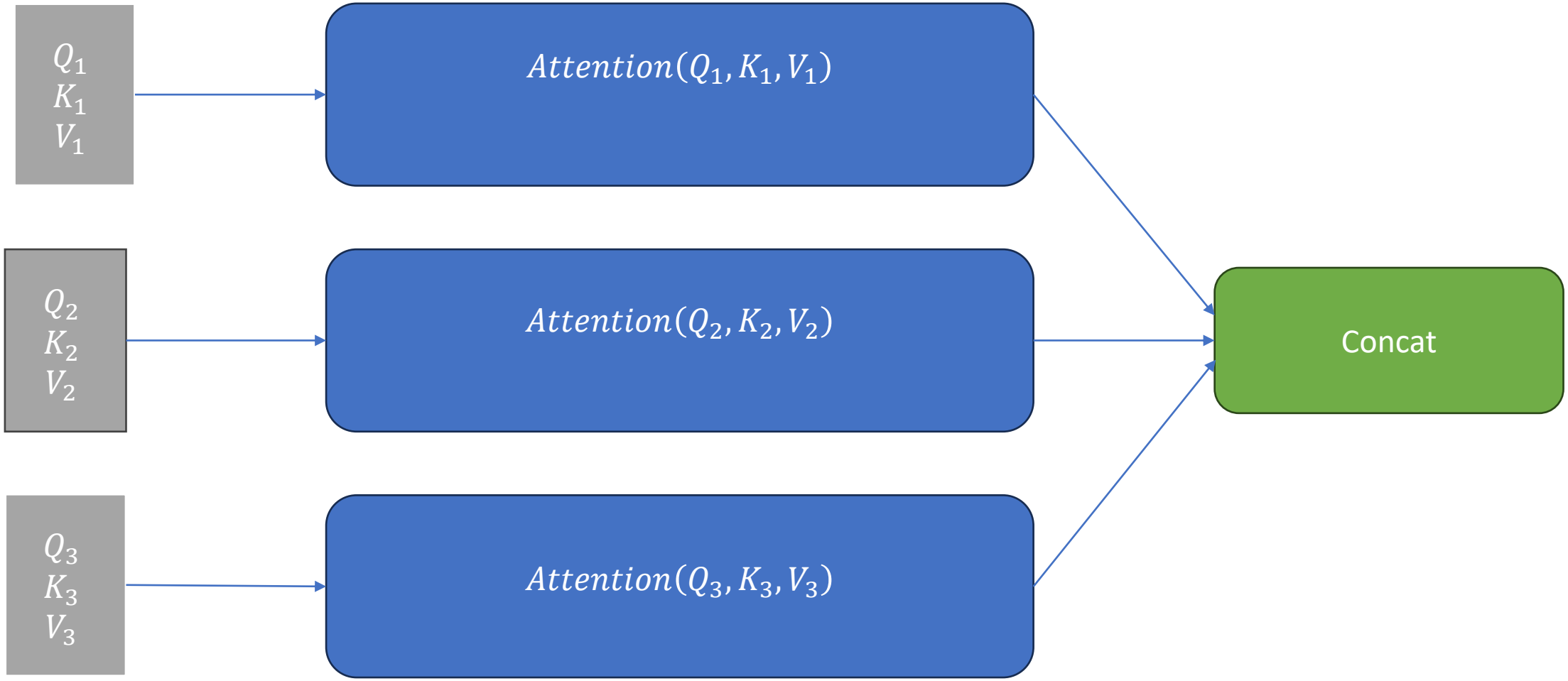
Self-Attention algorithm 3/3



$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

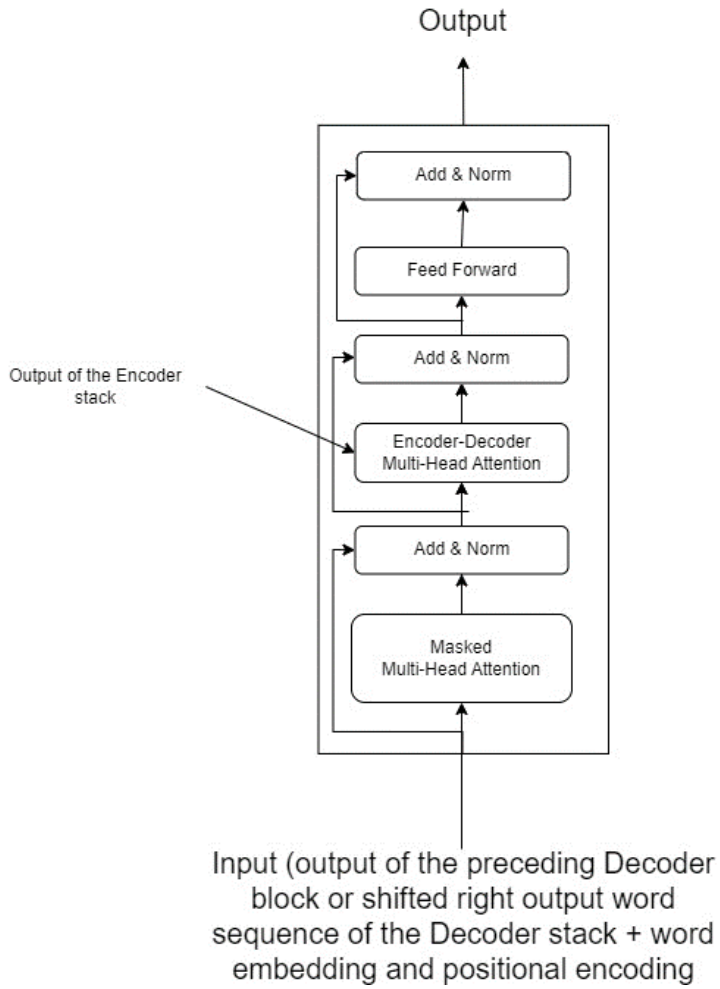
Abdelwahid Benslimane
wahid.benlimane@gmail.com

Multihead Attention



Abdelwahid Benslimane
wahid.benlimane@gmail.com

Decoder block



Abdelwahid Benslimane
wahid.benlimane@gmail.com

Advantages and drawbacks of Transformers:

- Very large architectures (hundreds of millions of parameters) -> gigantic amount of data is required for the training
- Highly parallelizable -> GPU power can be leveraged
- A lot of pretrained models are available through APIs (Hugging face, PyTorch etc.)
- Generally outperforms other models (but not always: 2009.05451.pdf (arxiv.org))

Scope:

- NLP: sentiment analysis, translation, text summarization, chatbot etc.
- Computer vision: image recognition with Transformers as an alternative to convolutional neural networks (CNNs)
- Scientific research: prediction of the 3D structure of proteins (https://daleonai.com/how-alphafold-works)

Abdelwahid Benslimane
wahid.benlimane@gmail.com

Popular Transformer-based models

BERT (Google):

- Encoder stack only
- Different models for different (NLP) tasks
- Can be fine-tuned on custom data
- Free and open source

GPT (OpenAI):

- Decoder stack only
- Same model for all NLP tasks
- Can’t be fine-tuned (except GPT-3.5/GPT-3.5 Turbo, fine-tuning available soon for GPT-4)
- Basic version of GPT-based tools for free, APIs and pro versions are paid

Abdelwahid Benslimane
wahid.benlimane@gmail.com