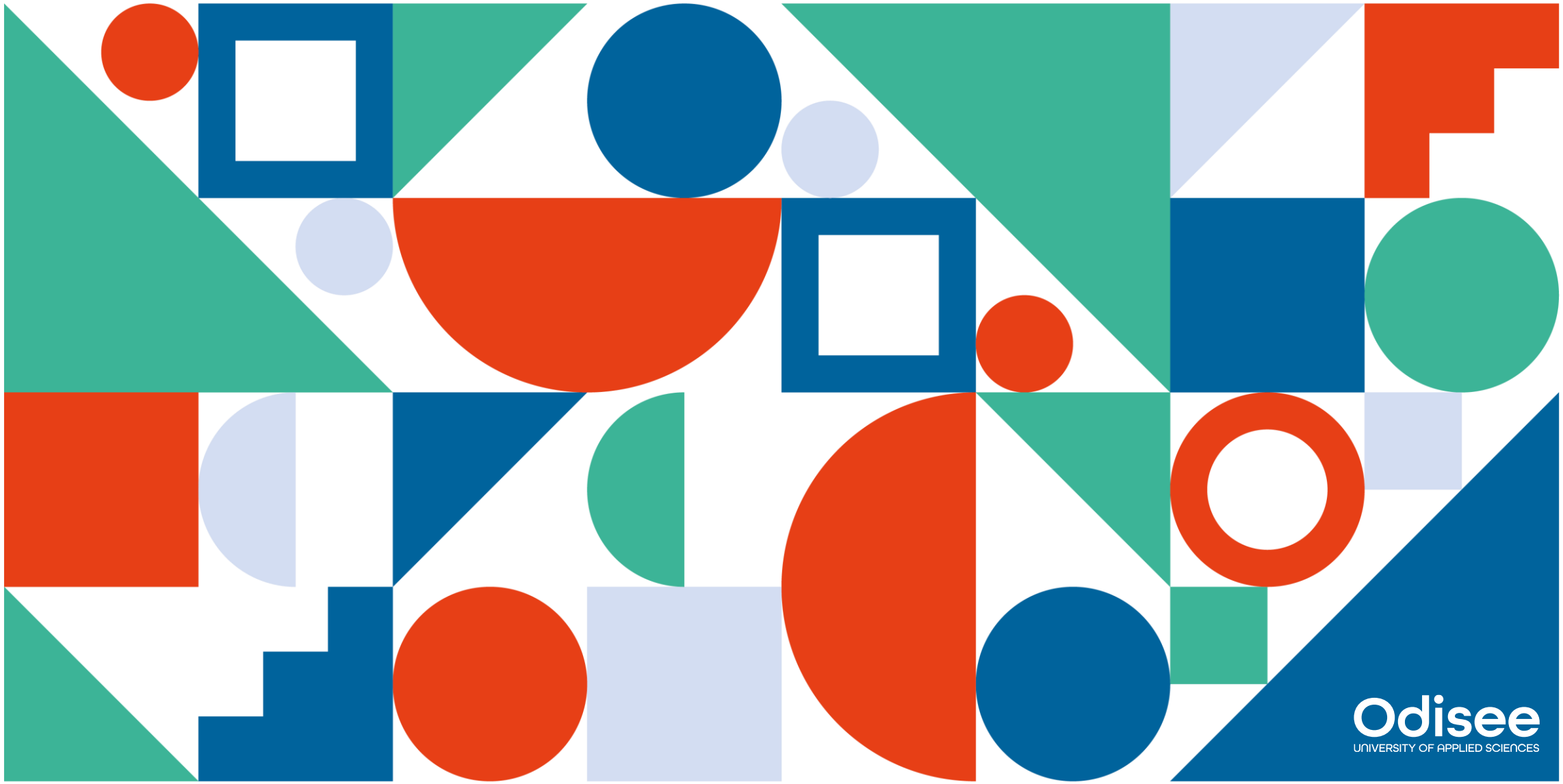

Odisee
UNIVERSITY OF APPLIED SCIENCES

Attention in Neurale Netwerken

Wat is aandacht?

5 AREAS OF ATTENTION

FOCUSED ATTENTION

When your child's attention is focused on visual (pictures) or auditory (hearing) information.

SHIFTING ATTENTION

When your child begins one task and stops to shift their focus on another task.

SUSTAINED ATTENTION

When your child can attend and focus on a task for a continuous stretch of time.

SELECTIVE ATTENTION

When your child attends to one specific task while filtering out other distractions around them.

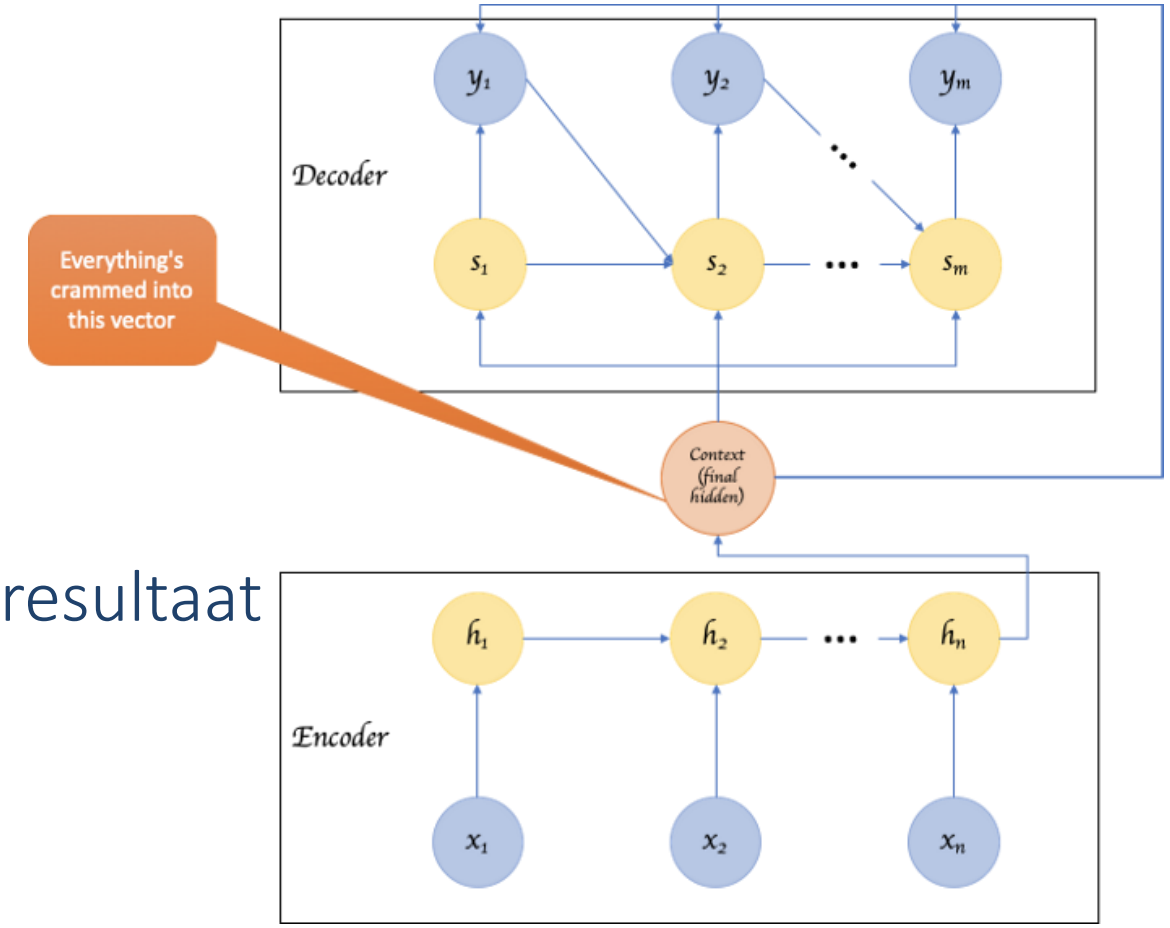
DIVIDED ATTENTION

When your child can multitask, focusing only a part of their attention on multiple items at the same time.

https://youtu.be/ubNF9QNEQLA

Waarom is Attention belangrijk bij AI?

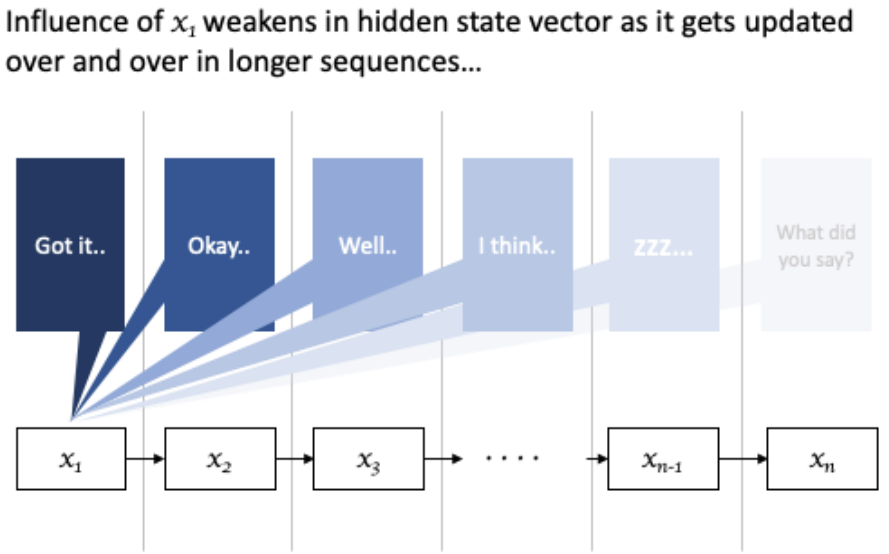
- Basic RNN
  - Standaard NN op beelden
  - Werkt maar niet zo efficient als CNN
- Volgorde van de woorden bepaalt het resultaat



https://towardsdatascience.com/an-introduction-to-attention-transformers-and-bert-part-1-da0e838c7cda

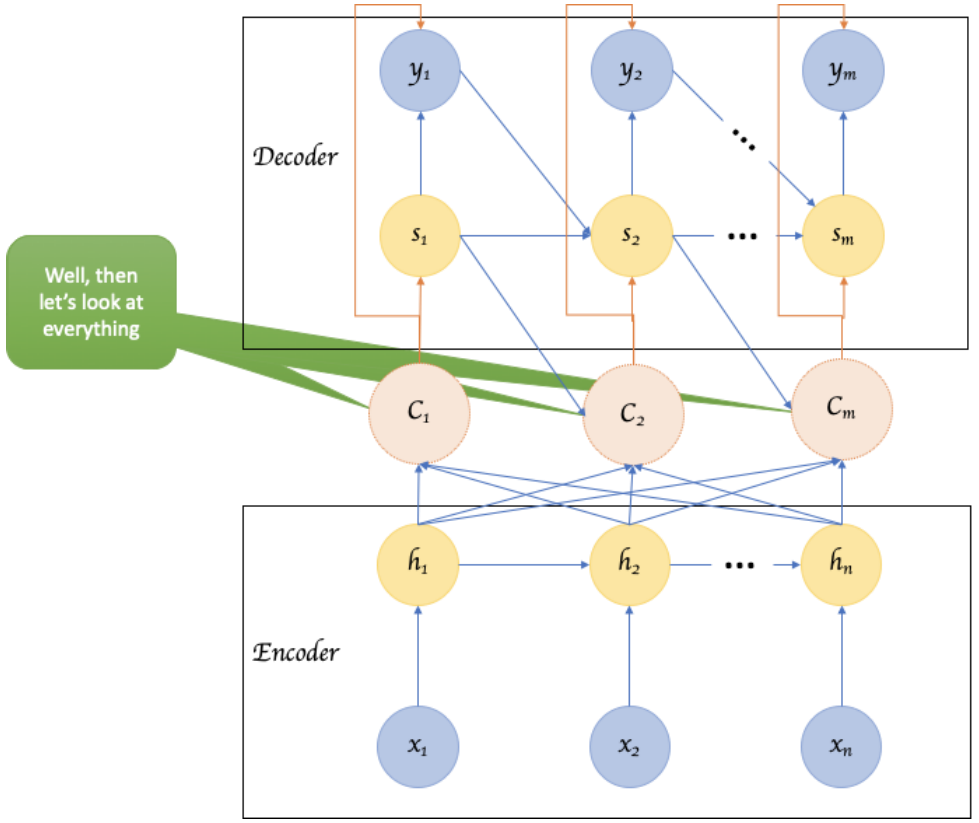
Waarom is Attention belangrijk bij AI?

- Lengte van de sequentie bepaalt de output



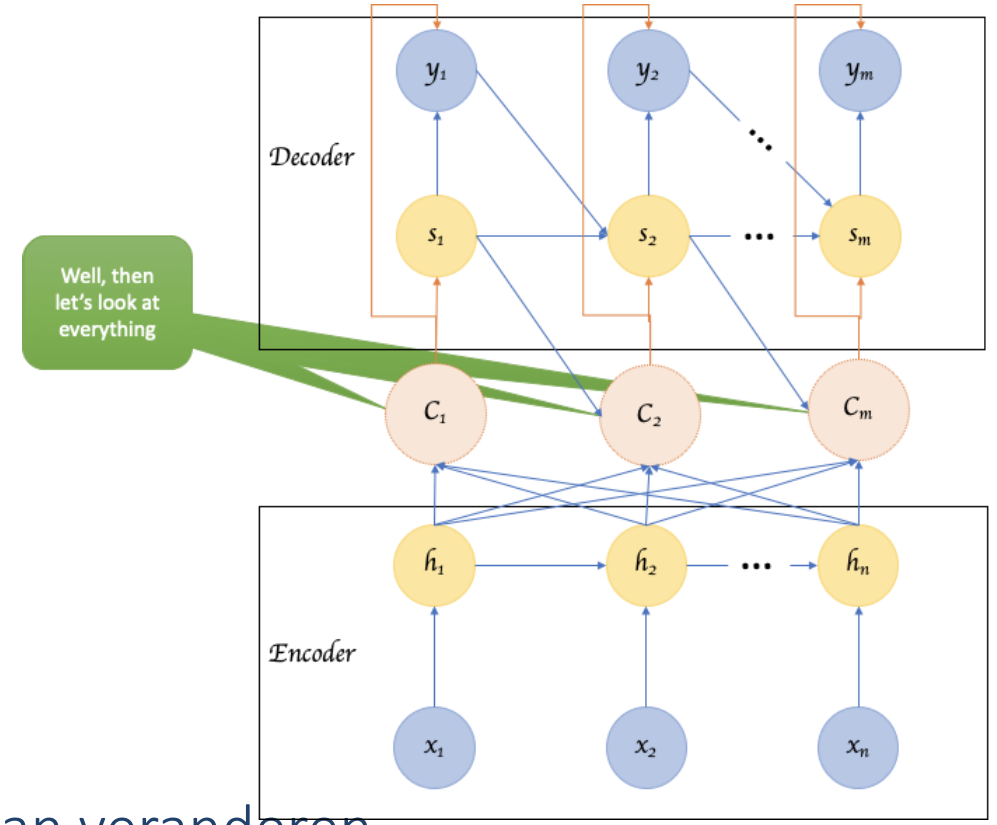
Hoe kunnen we deze problemen oplossen

- Bekijk de hidden state van elke tussenstap in de sequentie



Attention layer

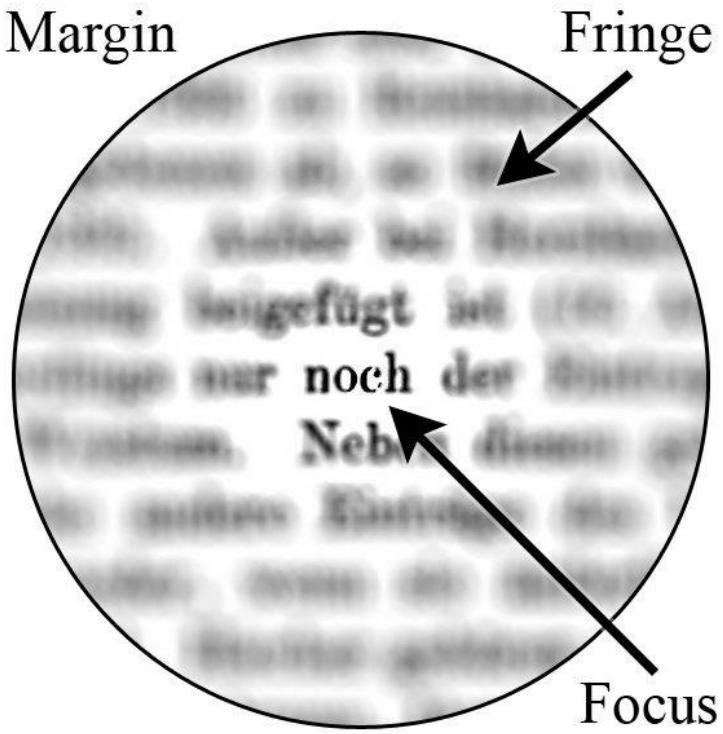
- Geen eenvoudige operatie
  - Anders geen verschil met standaard RNN
  - Maar een “Attention” operation



- Attention operation
  - Unique vector voor elke tijdstap
  - Gewichten van de verschillende hidden states kan veranderen
  - Er bestaan een aantal verschillende varianten
- Kunnen ook in CNN gebruikt worden
  - Focussen op delen van de figuur

Attention layers in Tensorflow

- tf.keras.layers.AdditiveAttention
- tf.keras.layers.MultiHeadAttention



Nadelen

- Attention layers zijn bedoeld om de state beter te kunnen analyseren
  - Het gebruik van een attention layer kan snel je model heel groot maken
  - Parallellisatie is moeilijk omdat tijdstappen sequentieel zijn
    - Reeds aanwezig bij standaard RNN netwerken

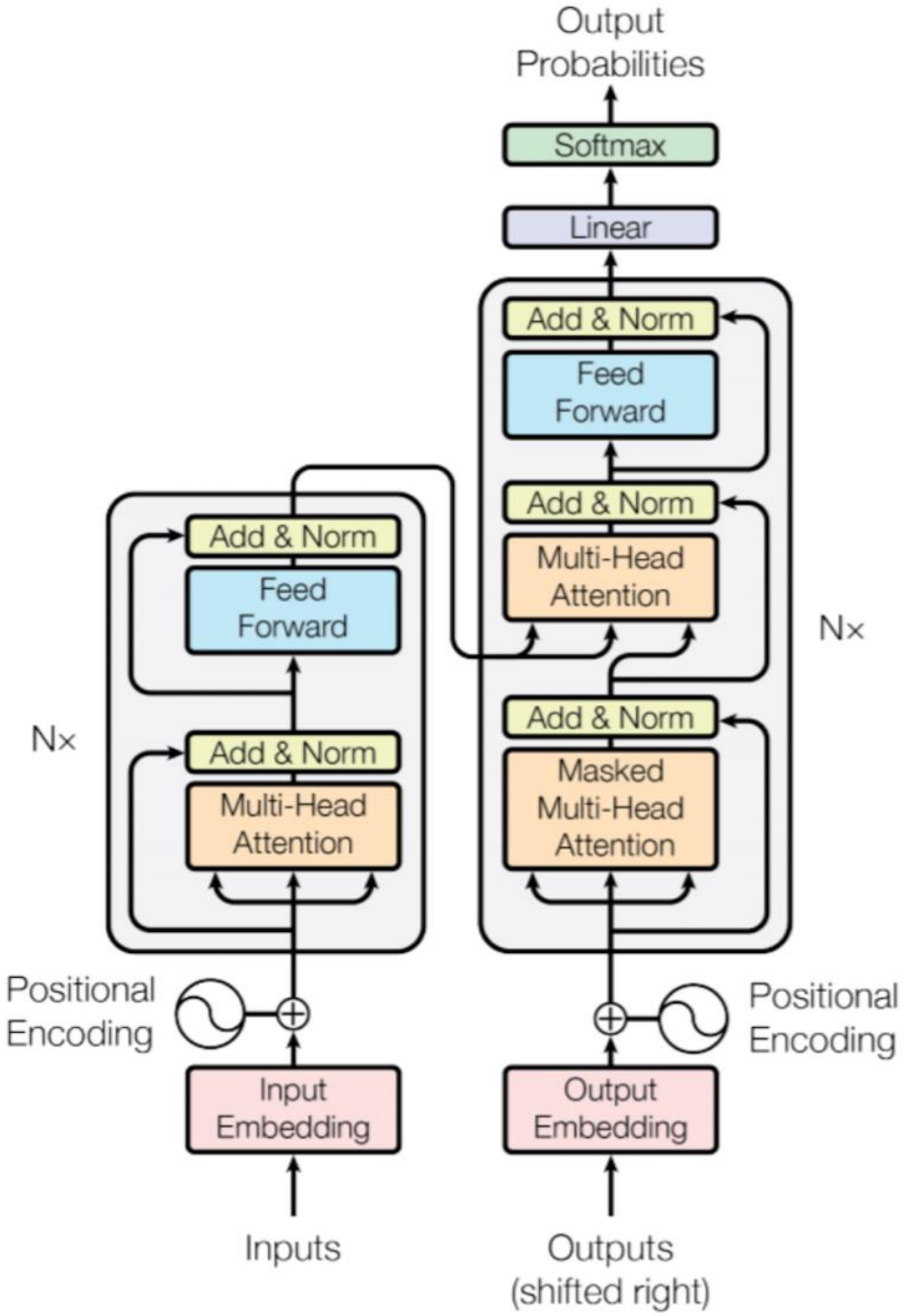
Transformers

Transformers

- Belangrijke paper voor dit type: https://arxiv.org/abs/1508.04025
- RNN architectuur gebaseerd op Attention layers
- Maakt gebruik van een encoder-decoder architectuur
- Belangrijke verschillen met standaard RNN architecturen:
  - Input sequentie kan in parallel berekend worden
    - Hierdoor kan een GPU gebruikt worden
  - Maakt gebruik van een Multi-head Attention Layer
    - Hierdoor wordt het vanishing gradient problem aangepakt
- Resultaten: https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

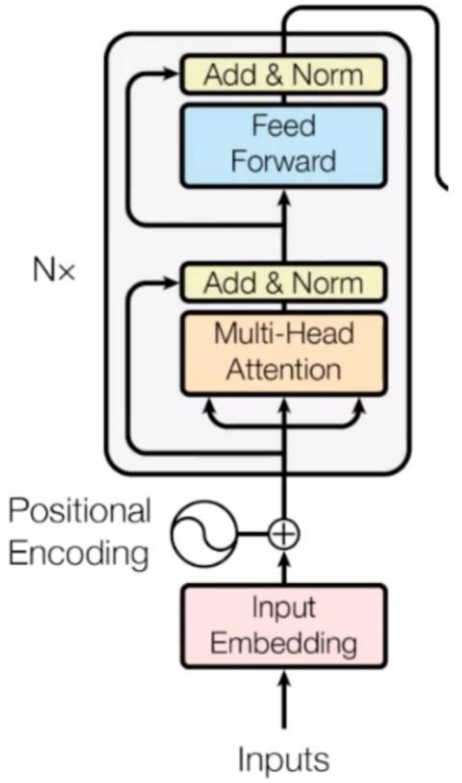
Transformers

- Voorbeeld: Engels – Nederlands vertaling
  - Engelse zinnen in inputs
    - Hele zinnen tegelijkertijd
  - Output is Nederlands
    - Woord per woord opgebouwd



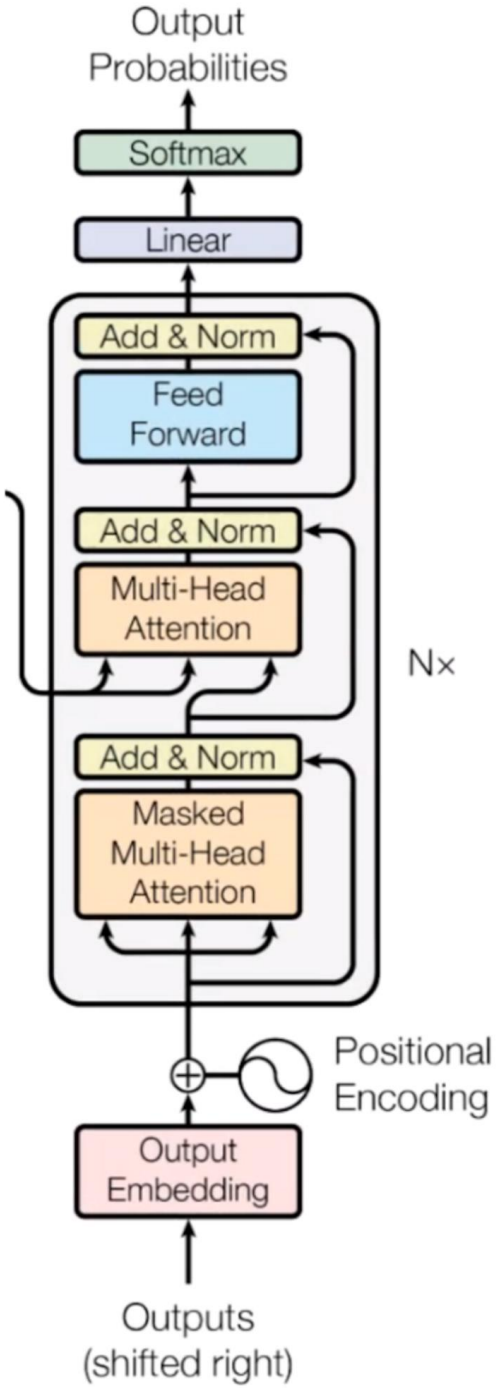
Transformer – Encoder block

- Embedding:
  - Word → Embedding → Positional Embedding → Final Vector, framed as Context
- Multi-head Attention
  - Wordt ook self-attention genoemd
  - Contextueel verband tussen woorden in een tekst
- Standaard feedforward neural network



Transformer – Decoder block

- Begin is hetzelfde
  - Embedding en positional encoding
- Attention part om zinvolle teksten op te bouwen
  - Masked: Gebruik enkel het verleden, niet de toekomst
- Combineer outputs encoder layer en de masked multi-head attention layer
- Feed forward gedeelte
  - One-hot encoding van alle woorden in de woordenboek



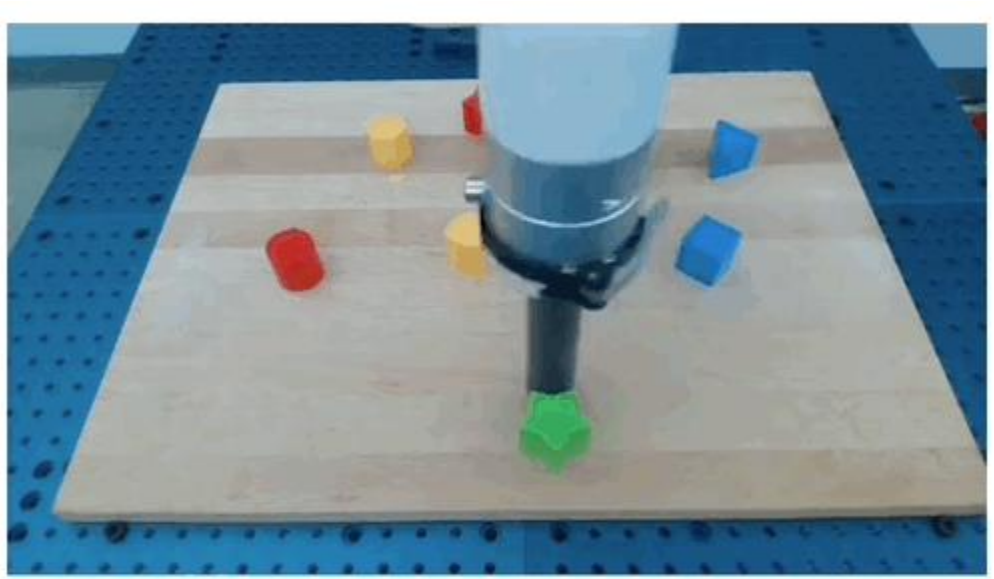
BERT

- Bidirectional Encoder Representations from Transformers (BERT)
  - Ontwikkeld door Google in 2018
  - Gebruikt in Google Search
  - Baseline in NLP
- 2 varianten
  - Base: 12 encoders met 12 bidirectional self attention heads
  - Large: 24 encoders met 16 bidirectional self attention heads

GPT

- GPT-3 laatste versie
  - Record van grootste neural network met 175 miljard parameters
- Voordelen
  - Heel krachtig, goed begrip van wat taal is
- Nadelen
  - Heel rekenintensief: Evenveel CO2 uitstoot als naar de maan rijden en terug
  - Biased (religieus, ras-gebaseerd, …): kan verstrekt worden door het model
  - Potentieel voor fake-news te maken


BERT



- https://ai.googleblog.com/2022/12/talking-to-robots-in-real-time.html?m=1