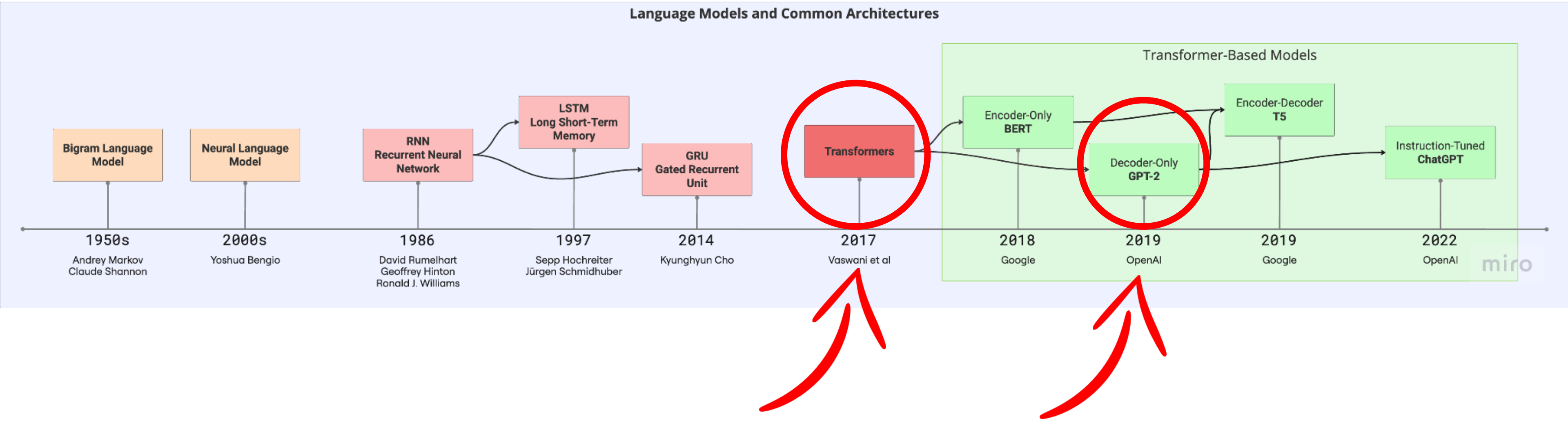
NLP4Web
Practice Session 10

Encoder-only (BERT) and Seq2seq (T5) Modeling
Decoding Strategies

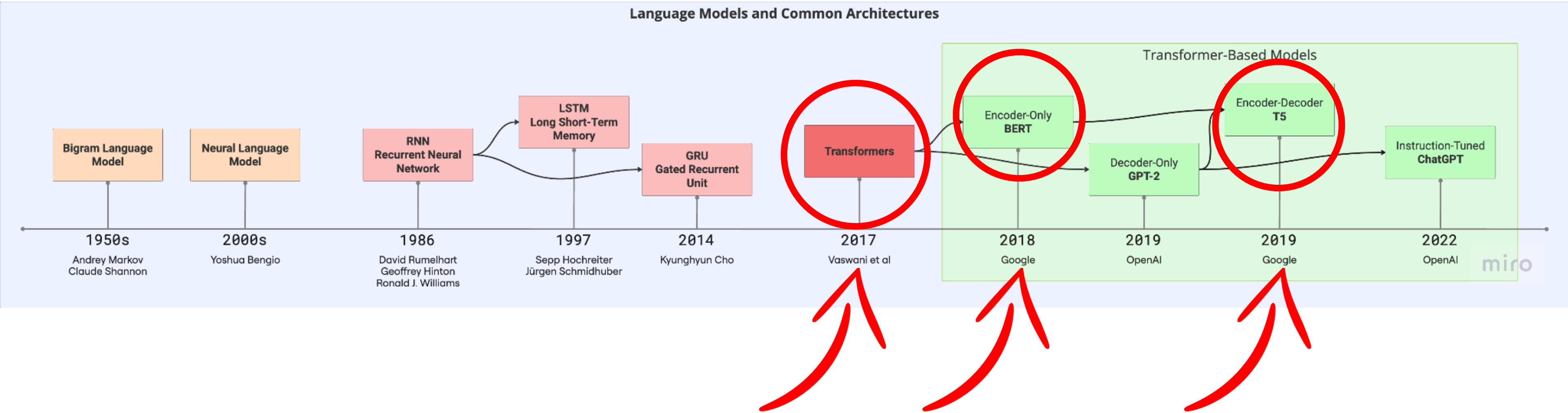
Hovhannes Tamoyan
tamohannes.com

To not get lost in space over time, let’s
Use a **mind map**

Last time we covered

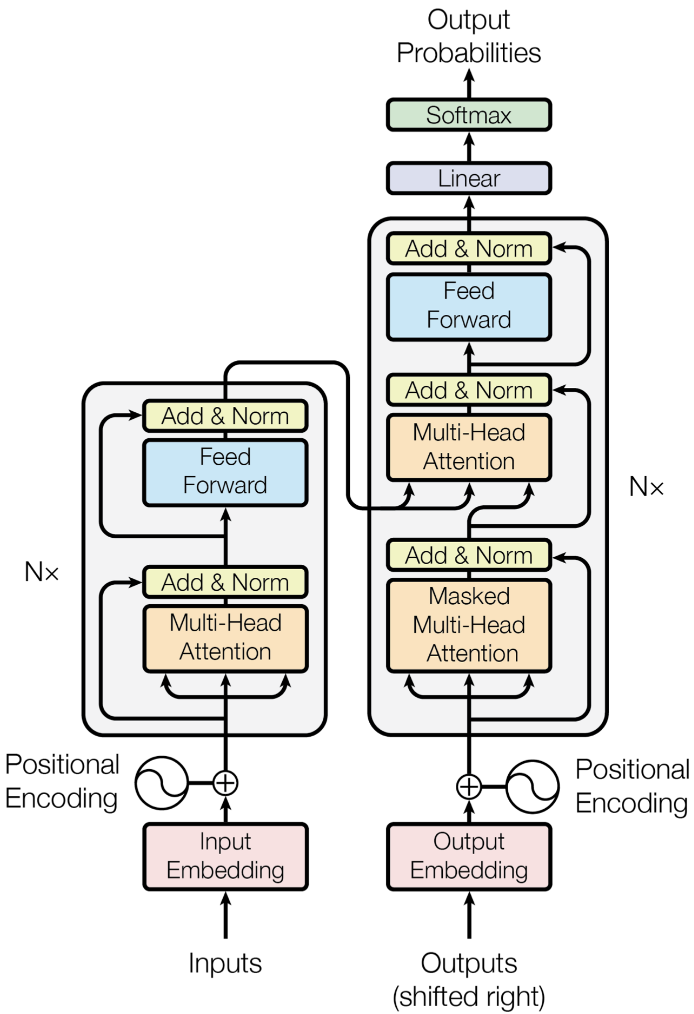


Today’s subject: Transformers
(Encoder-only and Encoder-Decoder)



Recap of Transformer architecture

- The main components
    - Embedding
    - Positional Encoding
    - Self-Attention
    - Feed Forward
    - Layer Normalization
    - Residual Connections

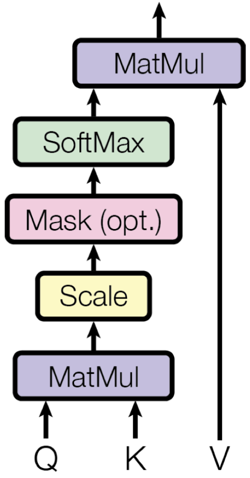


Recap of Attention mechanism

- Scaled Dot-Product attention
- where $\sqrt{d_k}$ is the dimension of the key vector k and query vector q

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

Scaled Dot-Product Attention

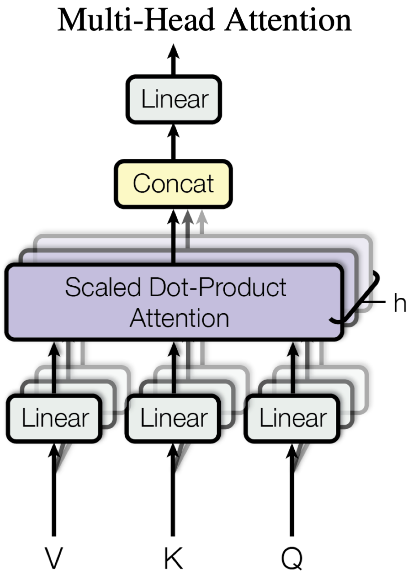


Recap of Attention mechanism

- Multi-head attention

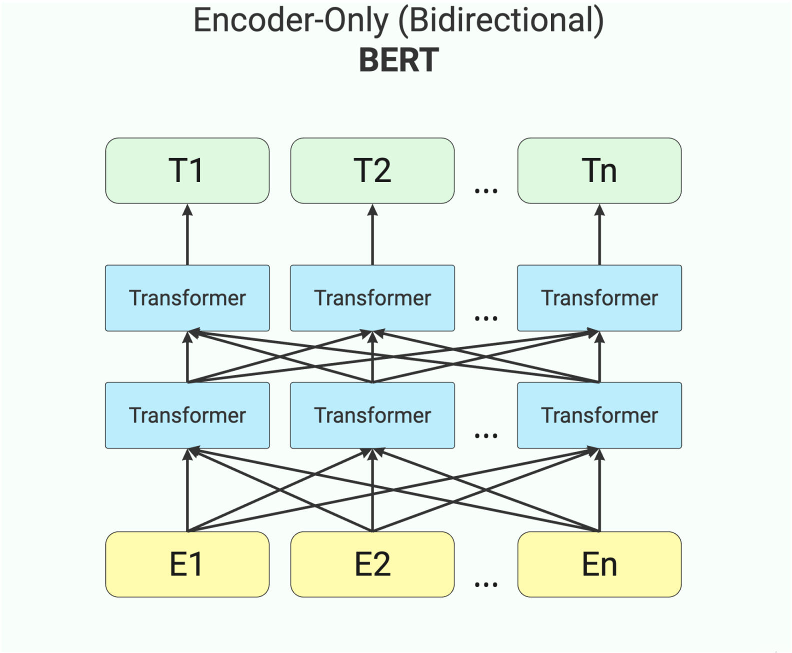
$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^O$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$



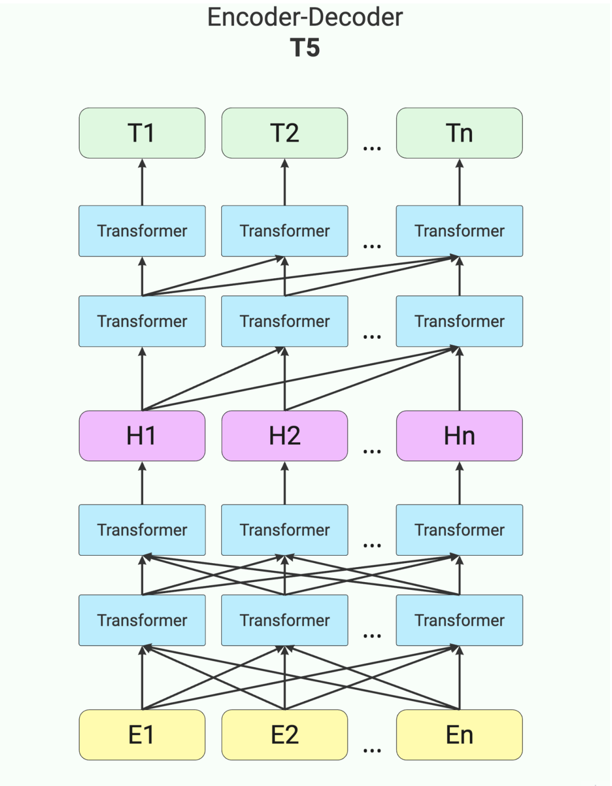
BERT: Encoder-Only, Bidirectional Architecture

- **Encoder-only**: processes text to produce representations in the latent space
- **Bidirectional**: it reads text both left-to-right and right-to-left for deeper context
- **Masked Language Modeling** (**MLM**): trains by predicting missing words in sentences
- **Self-attention**: focuses on important parts of text regardless of word position
- **Pre-trained**: fine-tuned for specific tasks with minimal extra training
- *Next Sentence Prediction: Learns relationships between consecutive sentences*



T5: Encoder-Decoder Architecture

- **Encoder-Decoder**: transforms input text into latent representations (encoder) and generates output text (decoder)
- **Text-to-Text**: converts all tasks (translation, summarization, etc.) into a text-to-text format
- **Pre-trained on Span Corruption**: Trains by masking spans of text and predicting the missing content
- **Bidirectional Encoding**: the encoder reads text in both directions for better understanding
- **Decoder Attention**: uses **self-attention** and **cross-attention** for generating accurate outputs
- **Fine-tuned for Multiple Tasks**: Adaptable to various NLP tasks with additional task-specific training



Decoding Strategies

Decoding Strategies

- Crucial for determining text quality and characteristics.
- Dictate how the model chooses the next word.
- Influence coherence, diversity, and relevancy of the output.

Most Common Decoding Strategies

- Greedy Decoding:
  - Selects the word with the highest probability at each step.
  - Fast and efficient but may lead to repetitive text.
- Beam Search:
  - Considers multiple possibilities (“beam width”) at each step.
  - Keeps track of the most probable sequences; more computationally intensive.

Most Common Decoding Strategies

- Top-k Sampling:
    - Chooses the next word from the top k most likely candidates.
    - Introduces randomness, enhancing diversity in text.
- Top-p (Nucleus) Sampling:
    - Selects words from the smallest set whose cumulative probability exceeds threshold p.
    - Balances randomness with high probability, improving coherence and variety.