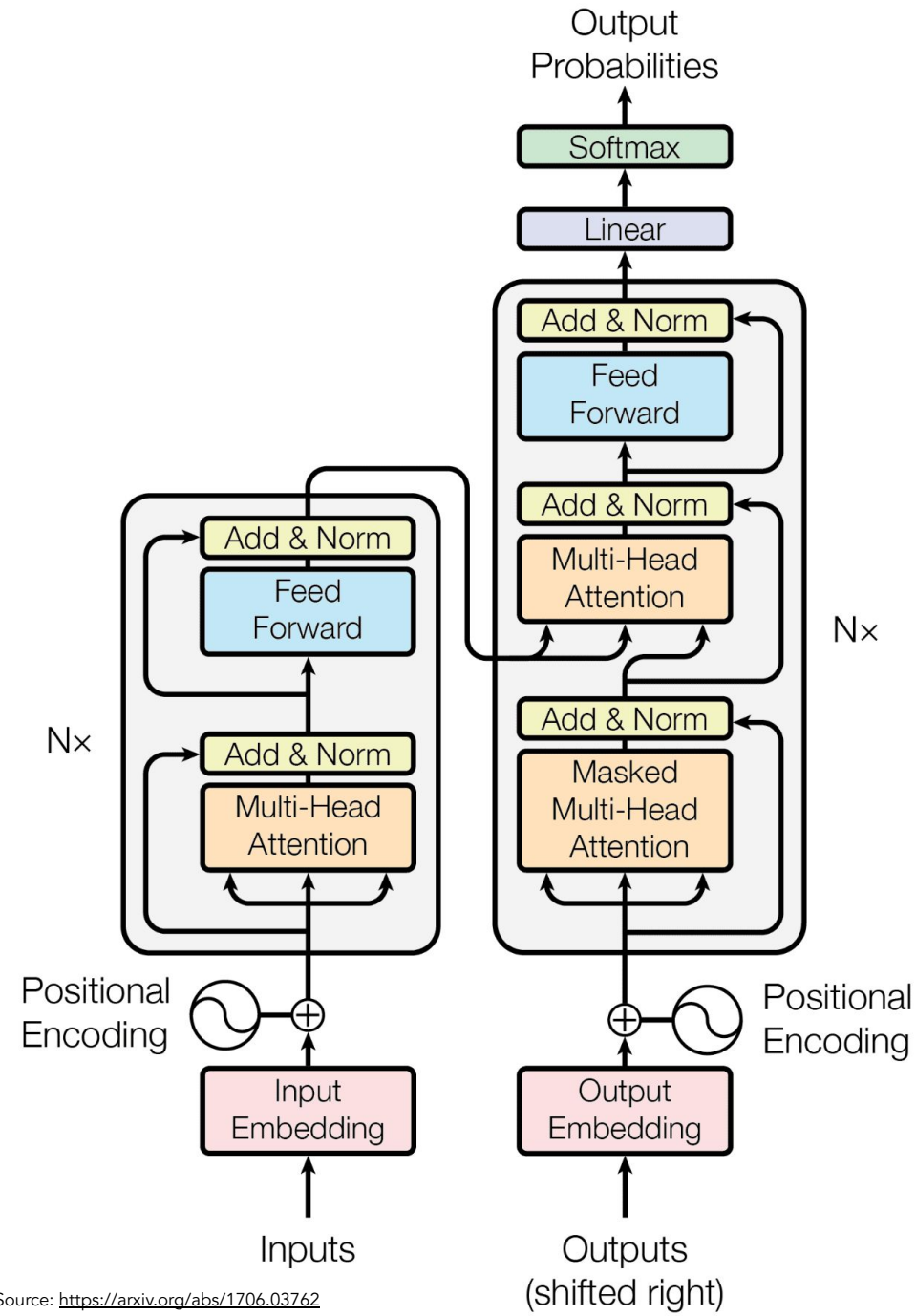
Transformer

- Attention is all you need (2017)
- How information flows through?
- State-of-the-art attention-based model
  - Many variants: BERT (encoder, for classification), GPT (decoder, for generation), T5 (generation)
- **Multi-head Attention**: Contextual embeddings
- Limitations
  - Fixed size input: usually between 256 and 128k tokens
- Can get very large
  - Largest LLMs have trillions parameters

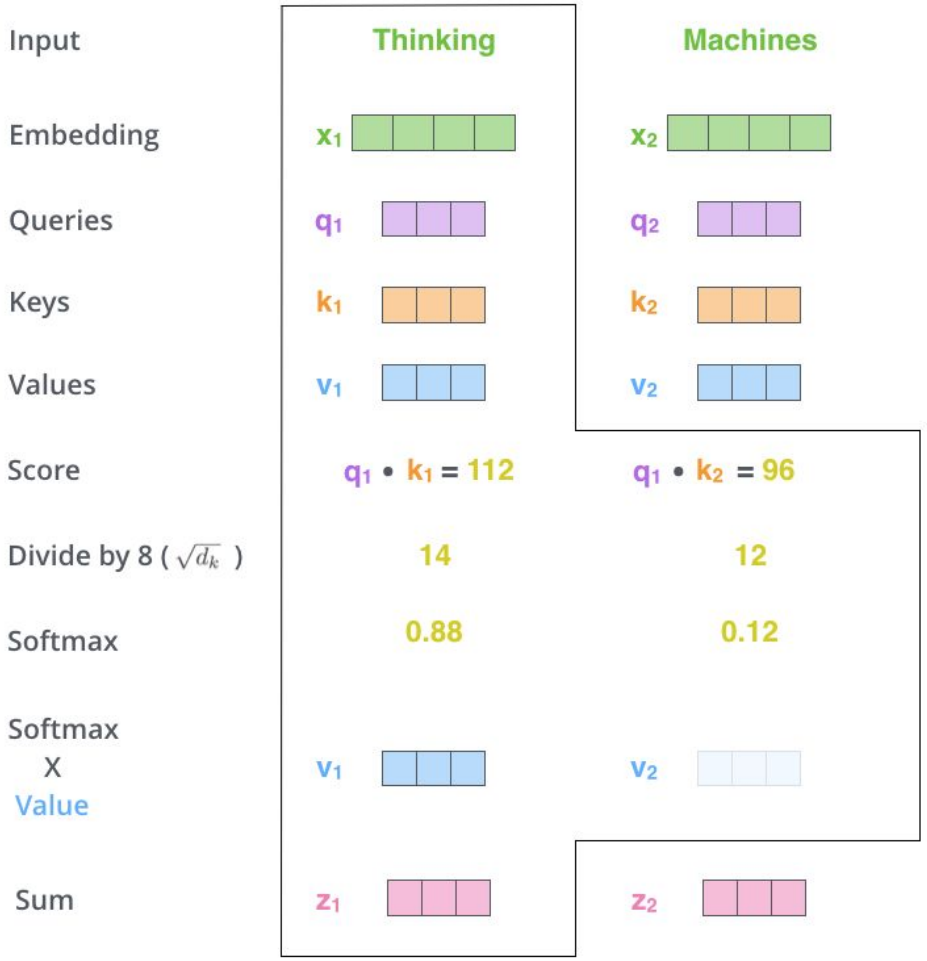


Source: https://arxiv.org/abs/1706.03762

Attention Mechanism

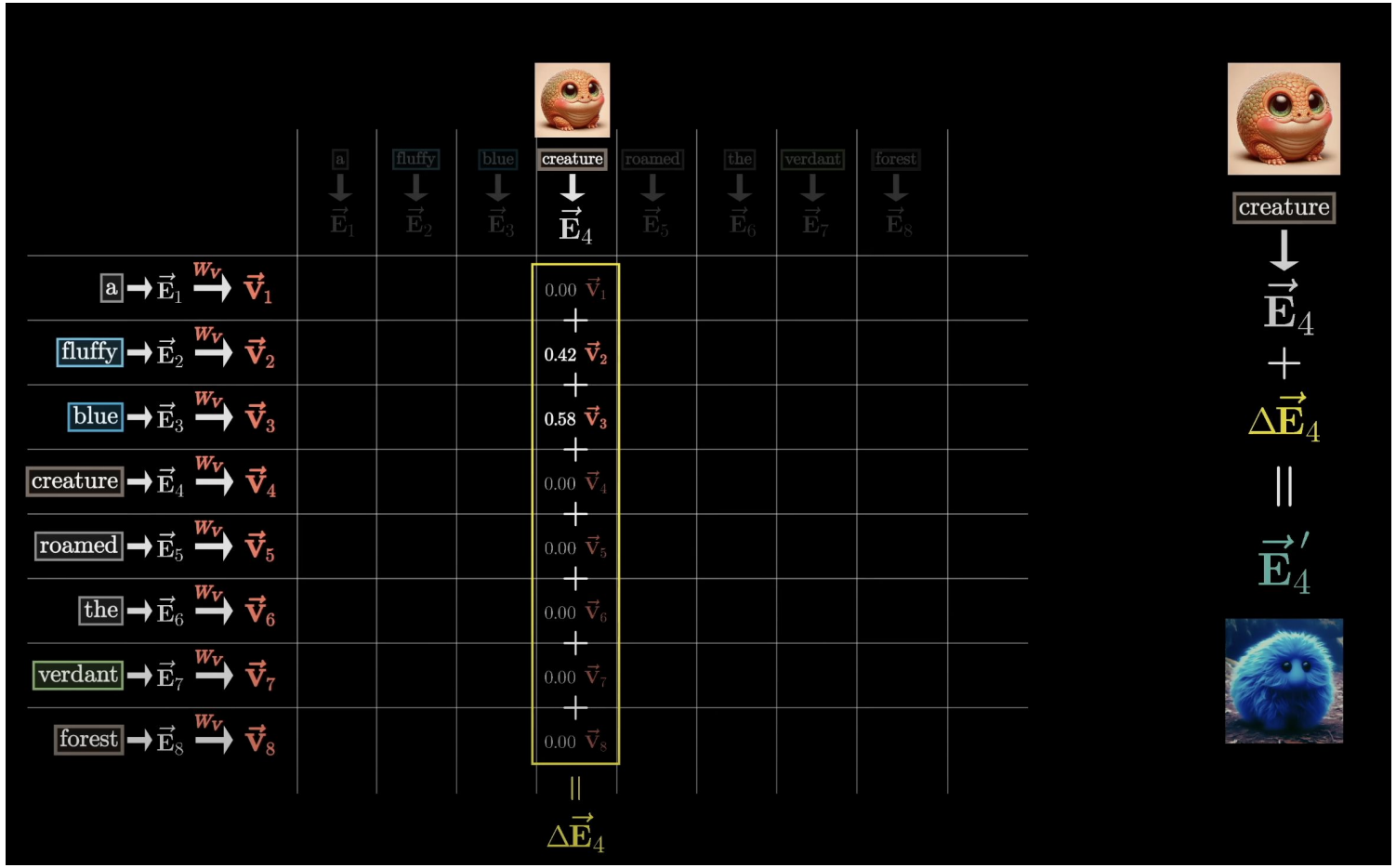
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

- Attention mechanism allows the model to **focus on specific parts of the input sequence**.
- Model dynamically weights the importance of different parts of the input.
- How it works?
  - **Input Representation**: Each word in the input sequence is represented as a vector (embedding).
  - **Scoring**: The attention mechanism computes a score that represents the relevance of each input word to the current output word being generated.
  - **Context Vector**: A weighted sum of input vectors is calculated using these weights, resulting in a context vector that emphasizes relevant parts of the input.



Source: https://jalammar.github.io/illustrated-transformer/

Contextual Embedding



Source: https://www.youtube.com/watch?v=eMlx5fFNoYc

Tokenization (building Vocabulary)

- Create tokens (e.g. sentencepiece, byte-pair encoding)
- 100 tokens ~= 75 words
- https://platform.openai.com/tokenizer
- Vocabularies are usually very large, approx. 50k-200k in size
- Special tokens
  - **[CLS]:** Used for classification tasks in some models.
  - **[SEP]:** Used to separate segments or sentences in some models.
  - **[MASK]:** Used for masking tokens during pre-training or for applying a masked language model objective.