



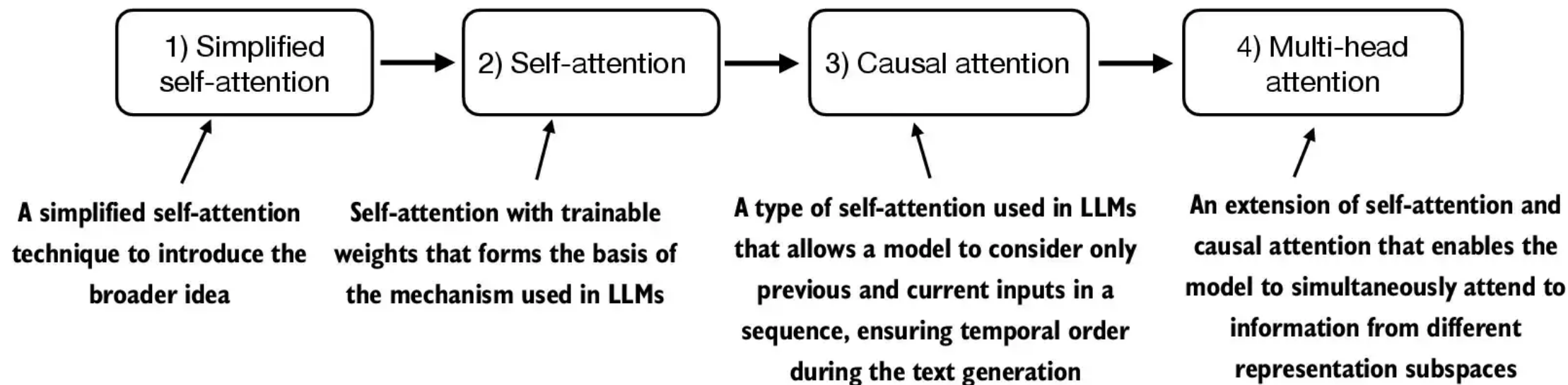
Self-Attention

Tópicos em Ciência de Dados

Pontifícia Universidade Católica de Campinas

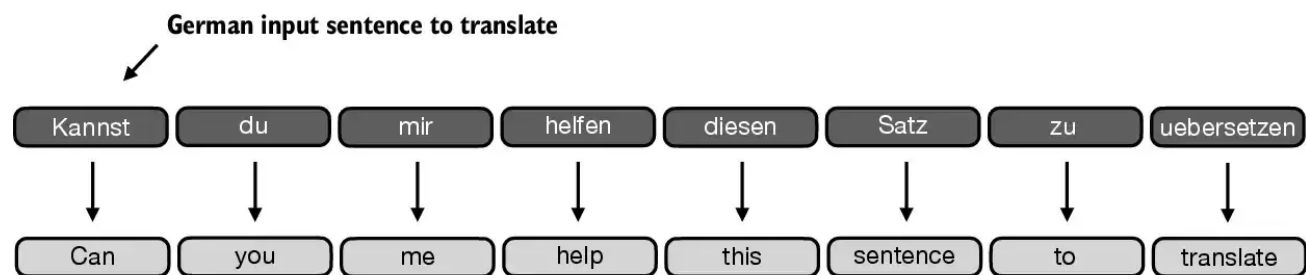
Prof. Dr. Denis M. L. Martins

Diferentes Mecanismos de Atenção

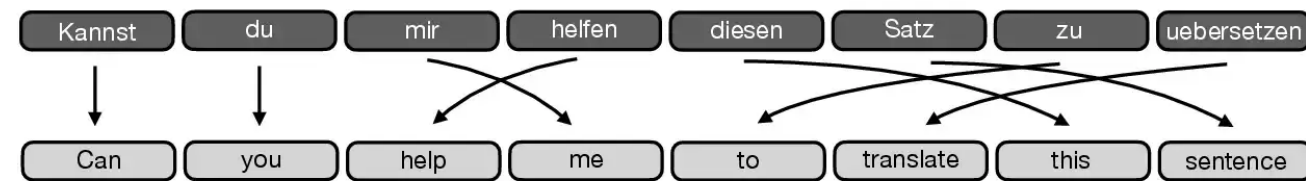


© 2024 Sebastian Raschka

Problema com longas sequências de texto



The word-by-word translation results in a grammatically incorrect sentence

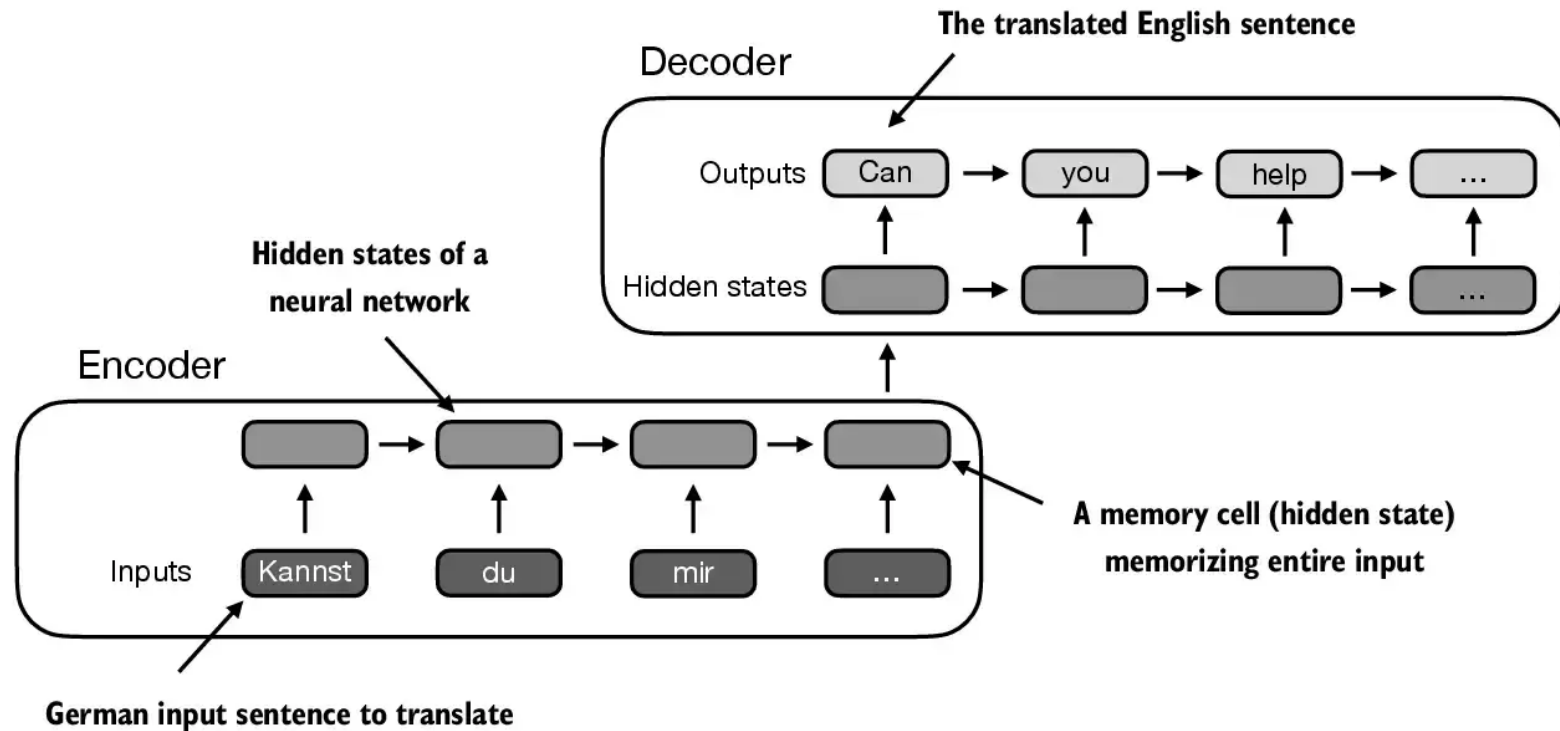


The correct translation

Certain words in the generated translation require access to words that appear earlier or later in the original sentence

© 2024 Sebastian Raschka

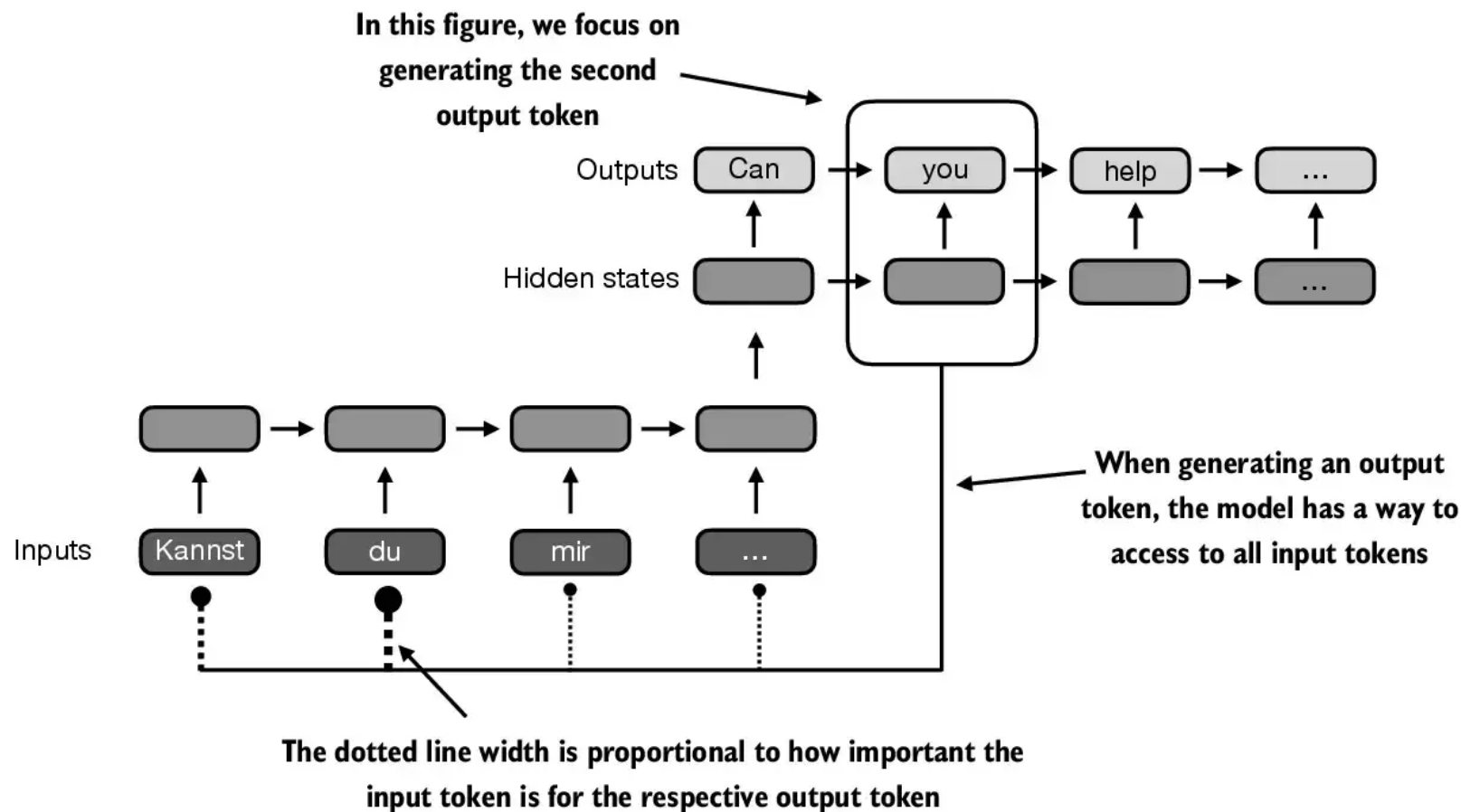
RNN perdem referência em longas sequências



© 2024 Sebastian Raschka

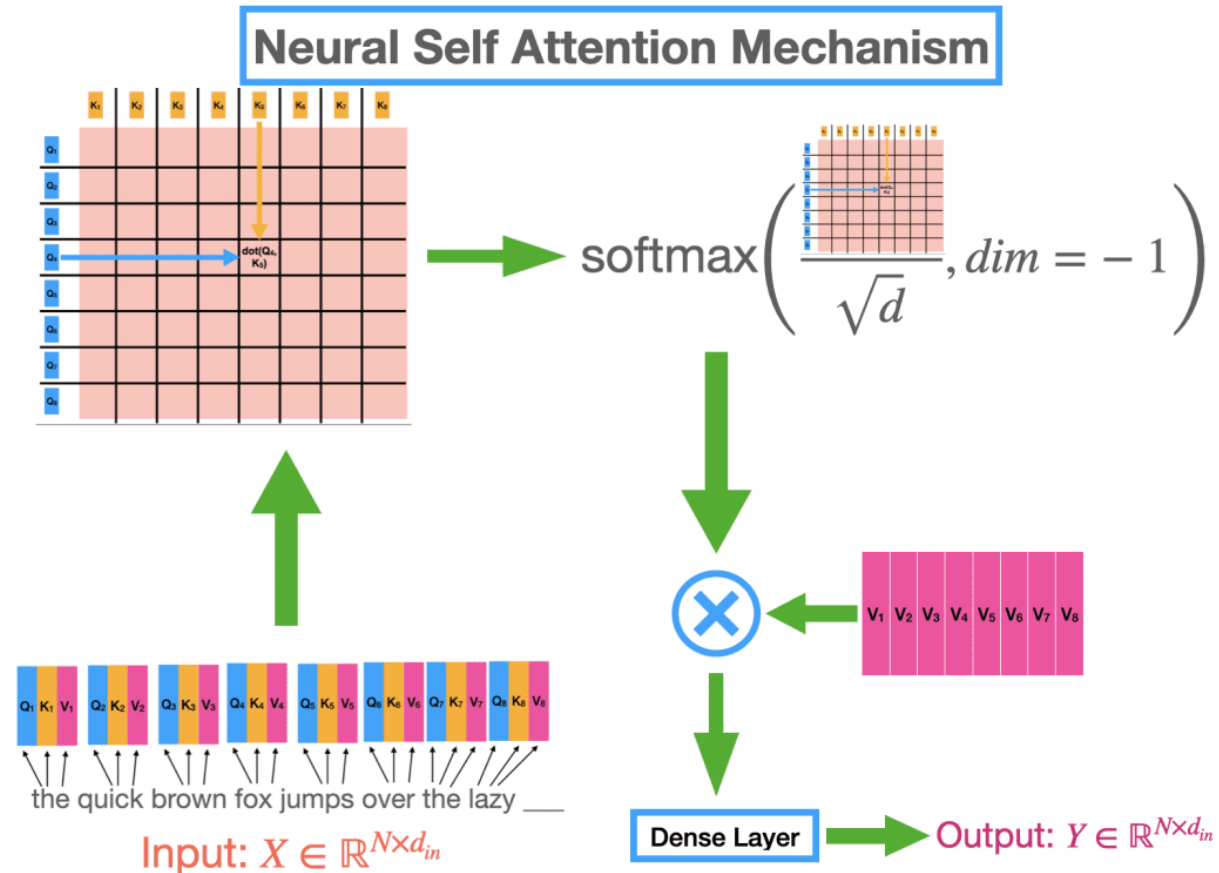
RNNs/GRUs apresentam dificuldade de capturar dependências longas.
Custo computacional linear no comprimento da sequência.

Resolvendo o problema através de atenção



© 2024 Sebastian Raschka

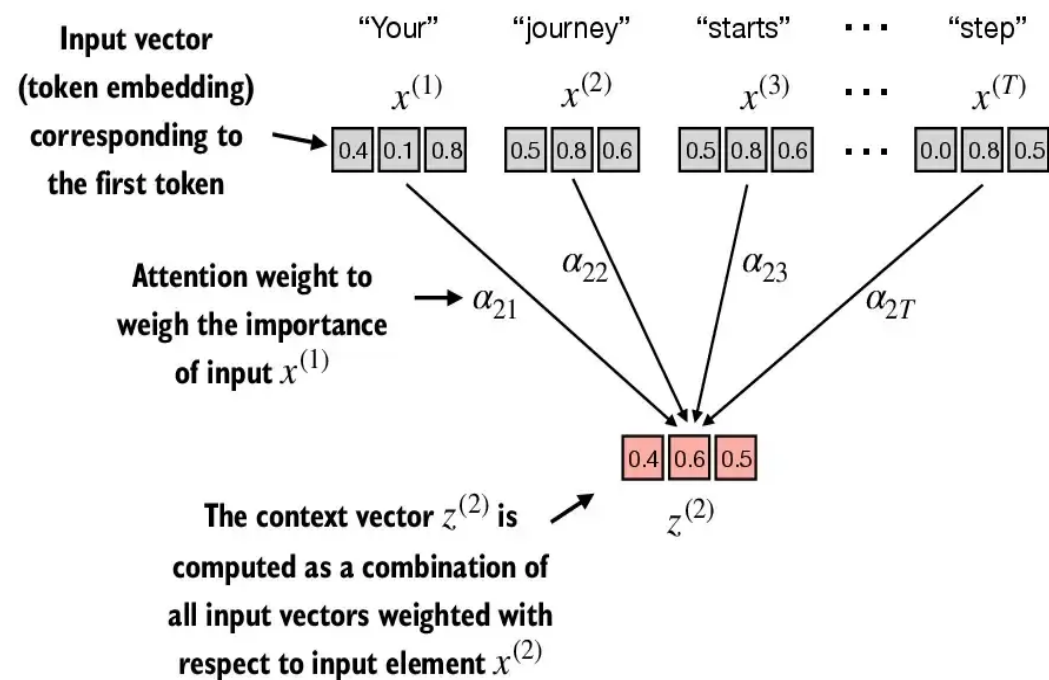
Self-Attention: Visão Geral



Visão geral do mecanismo de atenção em Transformers. Fonte: [Jaiyam Sharma @LearnOpenCV](#).

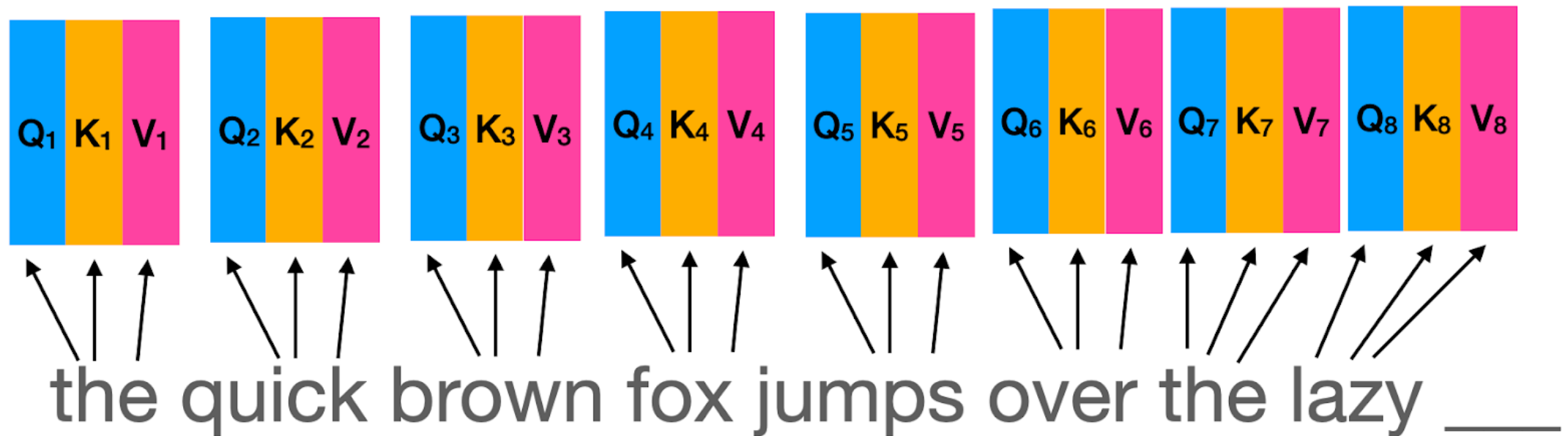
Self-Attention

- Keys (K), queries (Q) e values (V) são vetores fundamentais em mecanismos de atenção. Eles são obtidas através de projeções lineares, onde os embeddings dos tokens são multiplicados por matrizes de projeção (pesos) W^K , W^Q e W^V que funcionam como parâmetros da rede neural e, portanto, são aprendidas durante o treinamento.
- K : representam o token de origem.
- Q : representam os tokens de destino.
- V : representam a semântica e contexto dos tokens.



© 2024 Sebastian Raschka

Self-Attention



Primeiro passo no mecanismo de atenção. Fonte: [Jaiyam Sharma @LearnOpenCV](#).

Self-Attention

Mapeia uma query e pares de key-value em uma saída.

$$A = \text{Softmax} \left(\frac{Q K^\top}{\sqrt{d_k}} \right) V$$

onde:

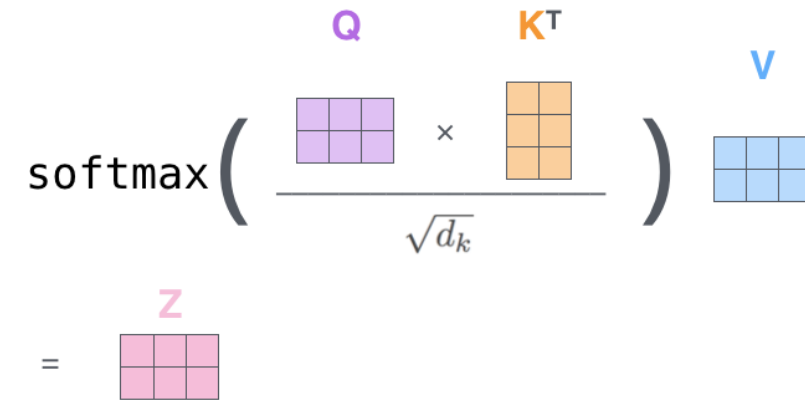
- $X \in \mathbb{R}^{n \times d_{\text{model}}}$ é a matriz de embeddings da sequência (n = comprimento).
- $W_Q, W_K, W_V \in \mathbb{R}^{d_{\text{model}} \times d_k}$ são as matrizes de projeção treináveis.

$$Q = X W_Q, \quad K = X W_K, \quad V = X W_V$$

- d_k é a dimensionalidade dos vetores **query** e **key** (normalmente d_{model}/h quando há h cabeças).
- Softmax opera ao longo da dimensão das posições de sequência, produzindo pesos de atenção.

Self-Attention

1. Cada palavra cria uma **query** e recebe **keys** e **values** das demais palavras.
2. O produto escalar $Q \cdot K$ mede a similaridade entre a pergunta de A e as chaves dos outros termos.
3. **Softmax** transforma esses números em pesos (probabilidades).
4. Os **values** são então somados ponderadamente, produzindo uma representação que leva em conta todas as palavras relevantes.


$$\text{softmax}\left(\frac{Q \times K^T}{\sqrt{d_k}}\right) \times V = Z$$

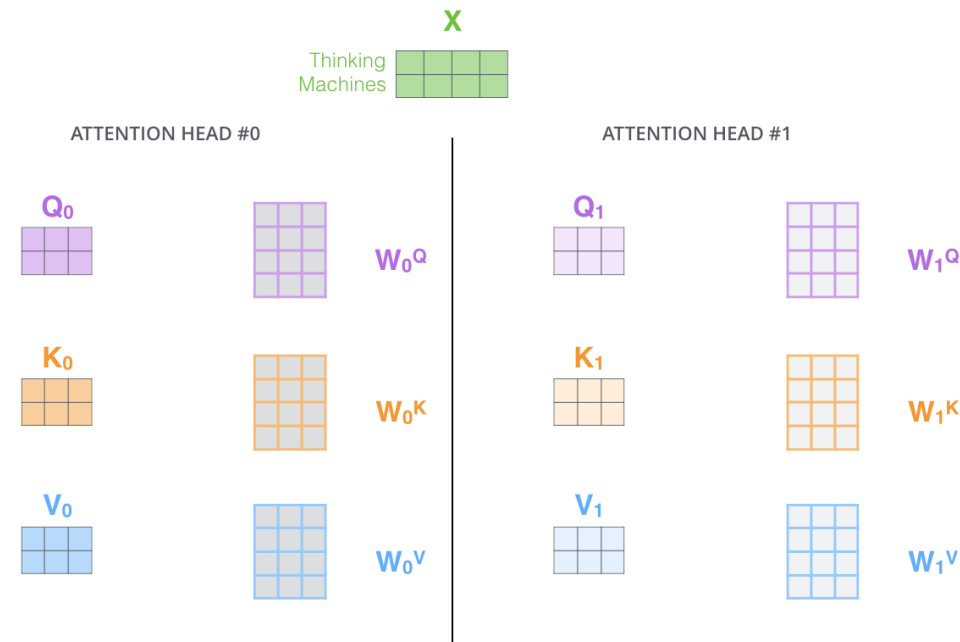
Fonte: [Illustrated Transformer](#)

Self-Attention

Elemento	Intuição	Como aparece na prática?
Queries (Q)	Pergunta : Cada palavra da frase está “fazendo uma pergunta” sobre quais outras palavras ela quer saber.	Vetor que representa a própria palavra, gerado por multiplicação do embedding pela matriz W_Q .
Keys (K)	Chaves de um armário : As demais palavras têm “chaves” que podem ser comparadas com as perguntas. Se uma chave for semelhante à pergunta, ela “abre” a porta para a informação relevante.	Vetor gerado pela mesma palavra, mas usando W_K .
Values (V)	Conteúdo guardado nas portas : Quando a porta abre, o que vem dentro é a informação que a palavra quer transmitir à pergunta.	Vetor resultante da multiplicação do embedding por W_V .

Resumo e Próximos Passos

- **Objetivo:** Permitir que cada token acesse e combine informação de **todas** as posições da sequência simultaneamente.
- **Queries (Q):** Vetores “perguntas” gerados a partir do próprio token.
- **Keys (K):** Vetores “chaves” que representam o conteúdo de cada token na mesma sequência.
- **Values (V):** Vetores contendo a informação real que será combinada.
- $A = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$
- **Próximos Passos:** Compreender **Multi-Head Attention**: Repete o mecanismo em (h) sub-espacos diferentes e concatena os resultados, permitindo capturar múltiplas relações simultaneamente.



Multi-Head Attention. Fonte: [Illustrated Transformer](#)