TÜBİTAK

ITU NLP NATURAL LANGUAGE PROCESSING RESEARCH GROUP

EURO^2

Lect. Tuğba Pamay Arslan

ITUNLP Research Group

AI & Data Engineering, İstanbul Technical University

ncc@ulakbim.gov.tr

The Power of LLMs: Transformer (Part 2)

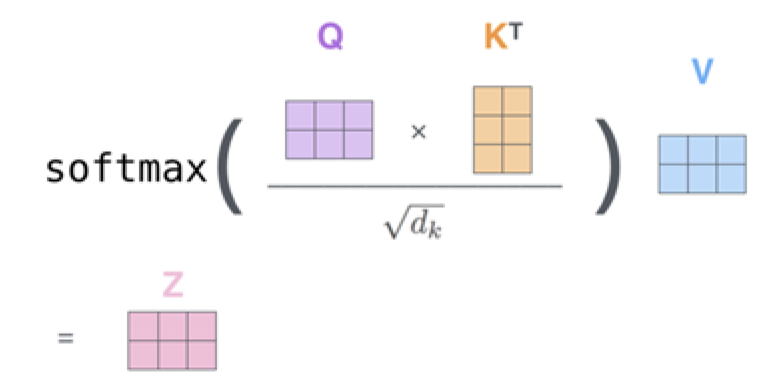
Self-Attention

➔ *Attend to* different parts of the **input sequence** when making predictions.

➔ “Self” ~ “ attention to “the same sequence which is currently being encoded.”

➔ 3 Input Tokens:

- Query (Q) / Key (K) / Value (V)

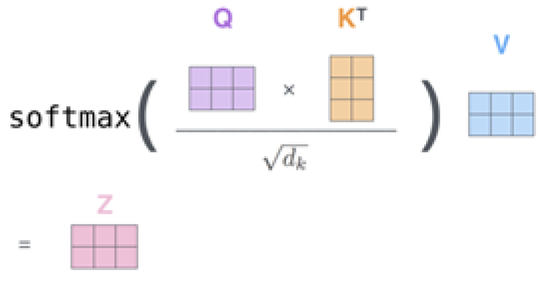


The self-attention calculation in matrix form

➔ How?

- Calculate attention scores by comparing each Query with every Key. These scores show how much each token should "attend" to others.
- Use attention scores to create a weighted sum of the Value vectors, producing a new representation for each token that considers relevant context.

Self-Attention



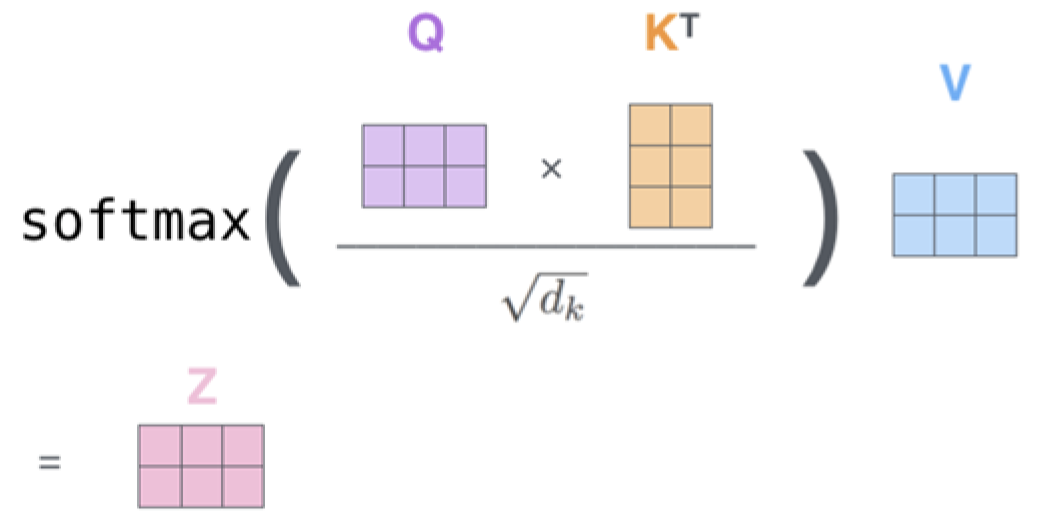
The self-attention calculation in matrix form

- Query (Q):
  - Representation of the element of interest that you want to obtain information about.
- Key (K):
  - a projection of the input data
  - Used to compute how relevant each element in the input sequence is to the Q.
- Value (V):
  - also a projection of the input data
  - Once the attention scores are calculated between Q and K, these scores are applied to the V to produce the weighted output representation.

Self-Attention - Formula

- **QK^T:** The dot product of each query with all keys, resulting in a matrix of attention scores.
- $\sqrt{d_k}$: A scaling factor where dk is the dimensionality of the keys (and queries).

  This prevents the dot products from becoming too large, stabilizing gradients during training.
- Main Calculation Steps:
  - Compute the similarity between each query and every key.
  - Normalize the similarity scores using softmax.
  - Use the normalized scores to calculate a weighted sum of the value vectors.



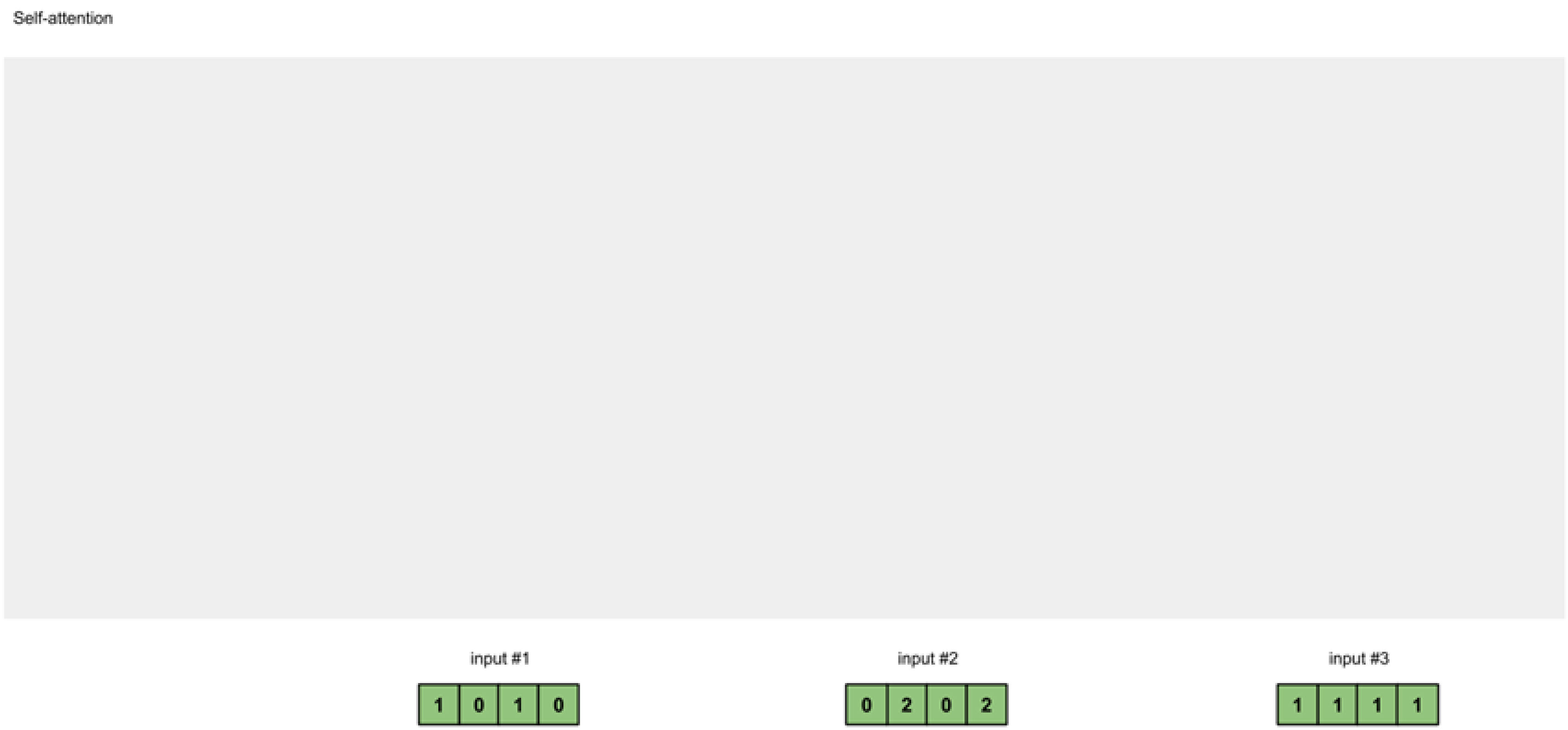
The self-attention calculation in matrix form

Attention vs. Self-Attention

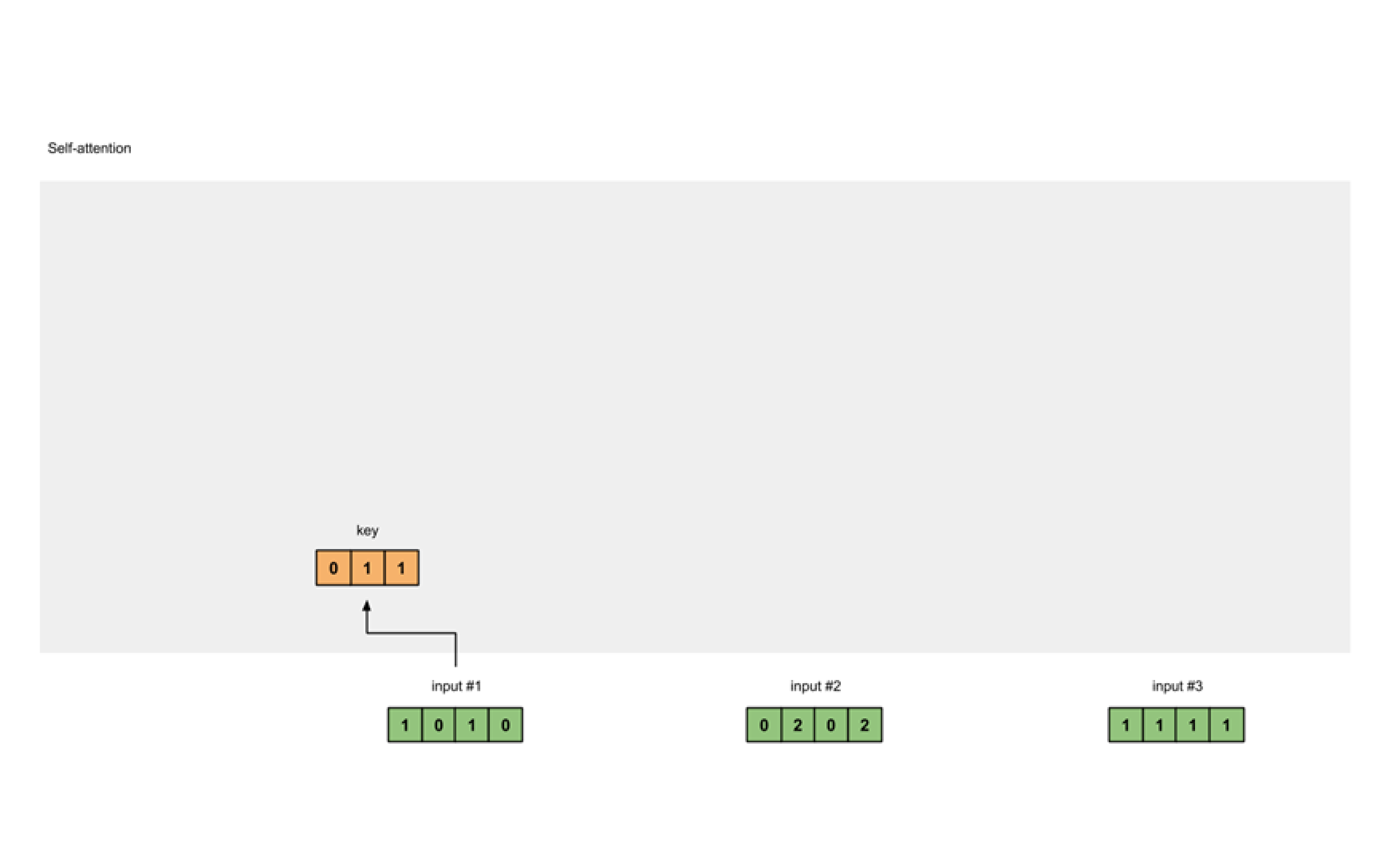
- ➔ *Focus of Attention:*
  - ◆ Self-Attention: Each token in an input attends to all other tokens within the same sequence, capturing relationships within the input itself.
  - ◆ Attention: Attend to relevant parts of the input when producing each output token.

- ➔ *Computation:*
  - ◆ Self-Attention: Computes attention weights across all positions within the input sequence. Providing a global view of the sequence.
  - ◆ Attention: Calculates attention weights from the decoder's current state to each position in the encoder's output. Improving alignment and accuracy in generation tasks between input and output.

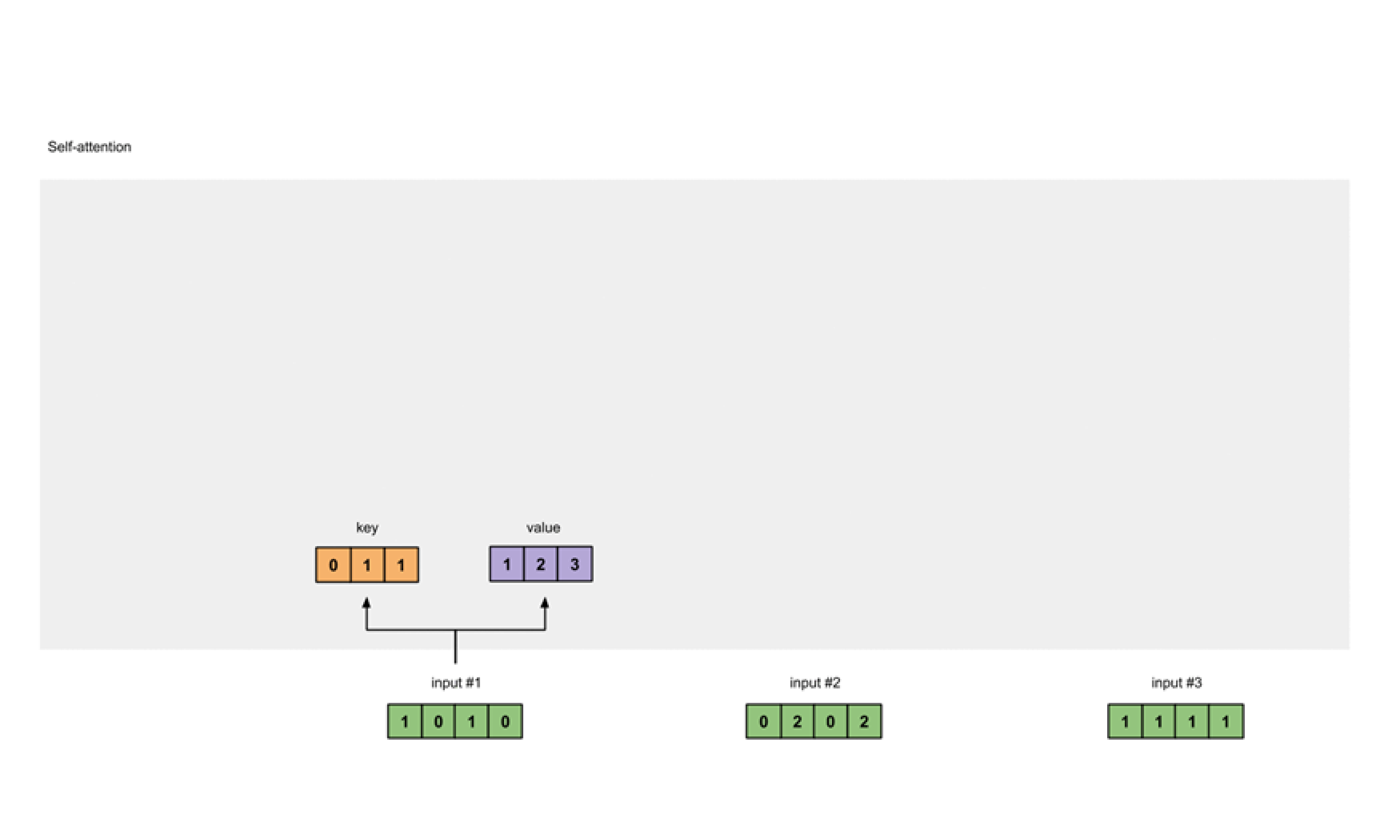
Step-by-Step / Self-Attention



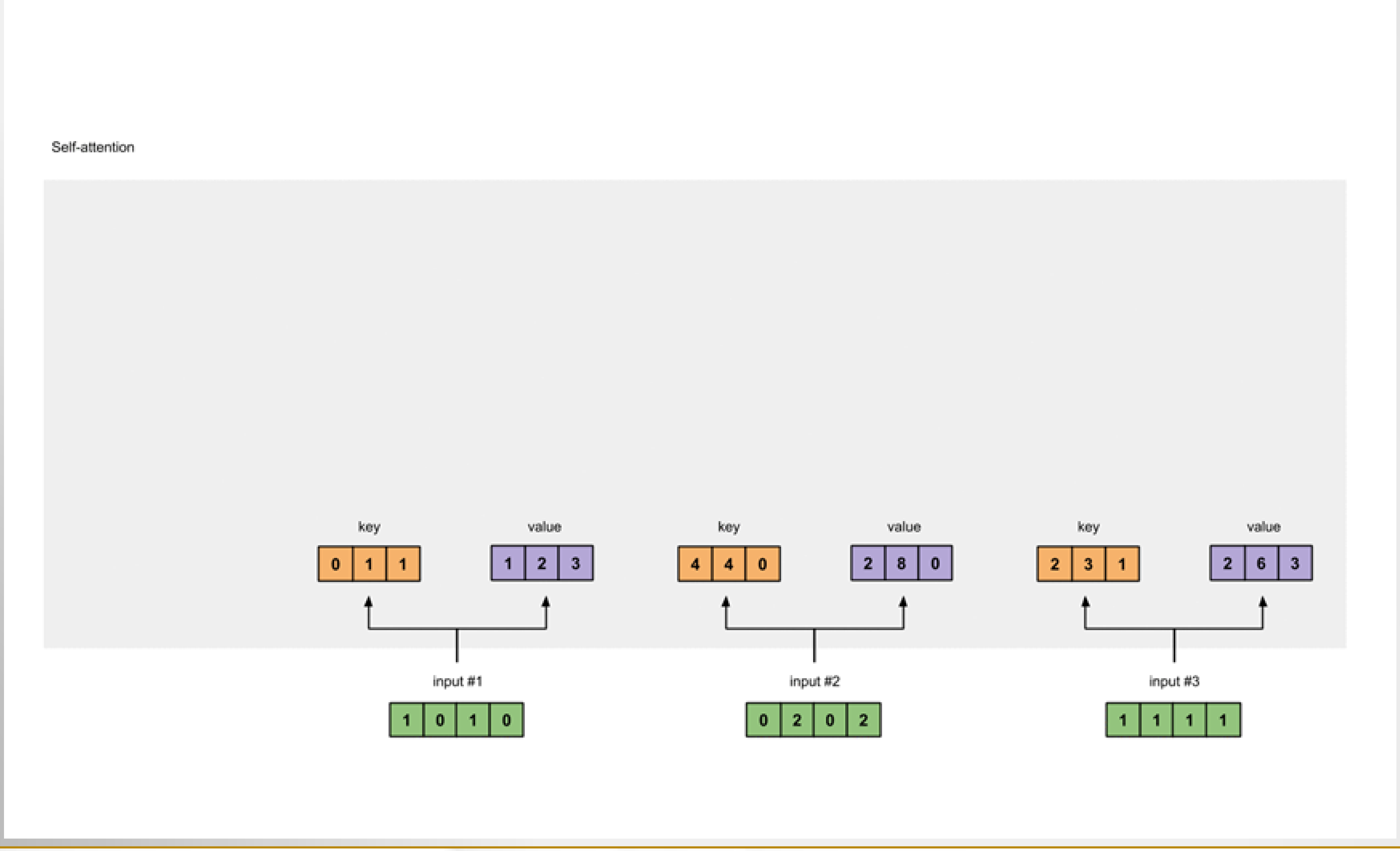
Step-by-Step / Self-Attention



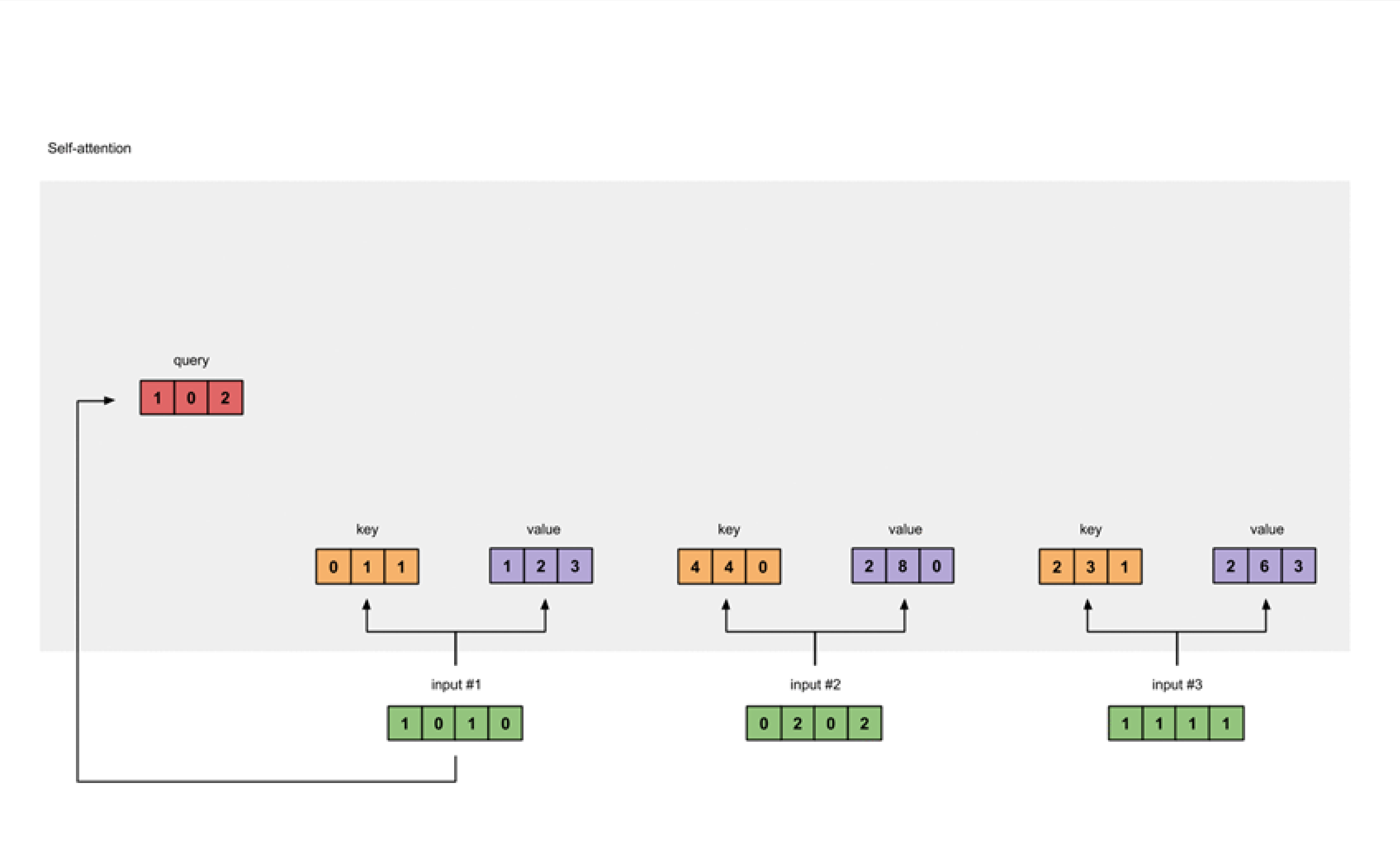
Step-by-Step / Self-Attention



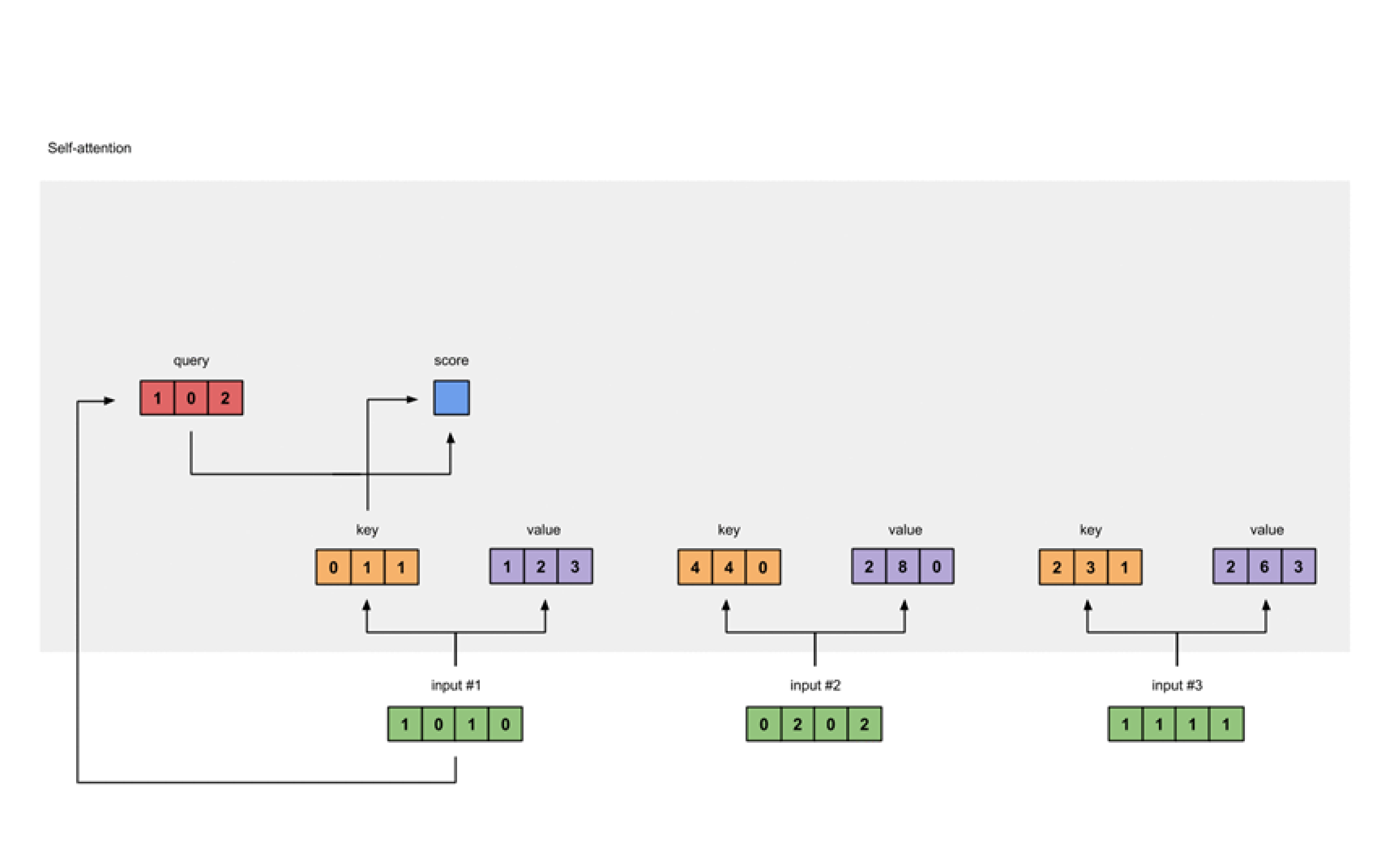
Step-by-Step / Self-Attention



Step-by-Step / Self-Attention



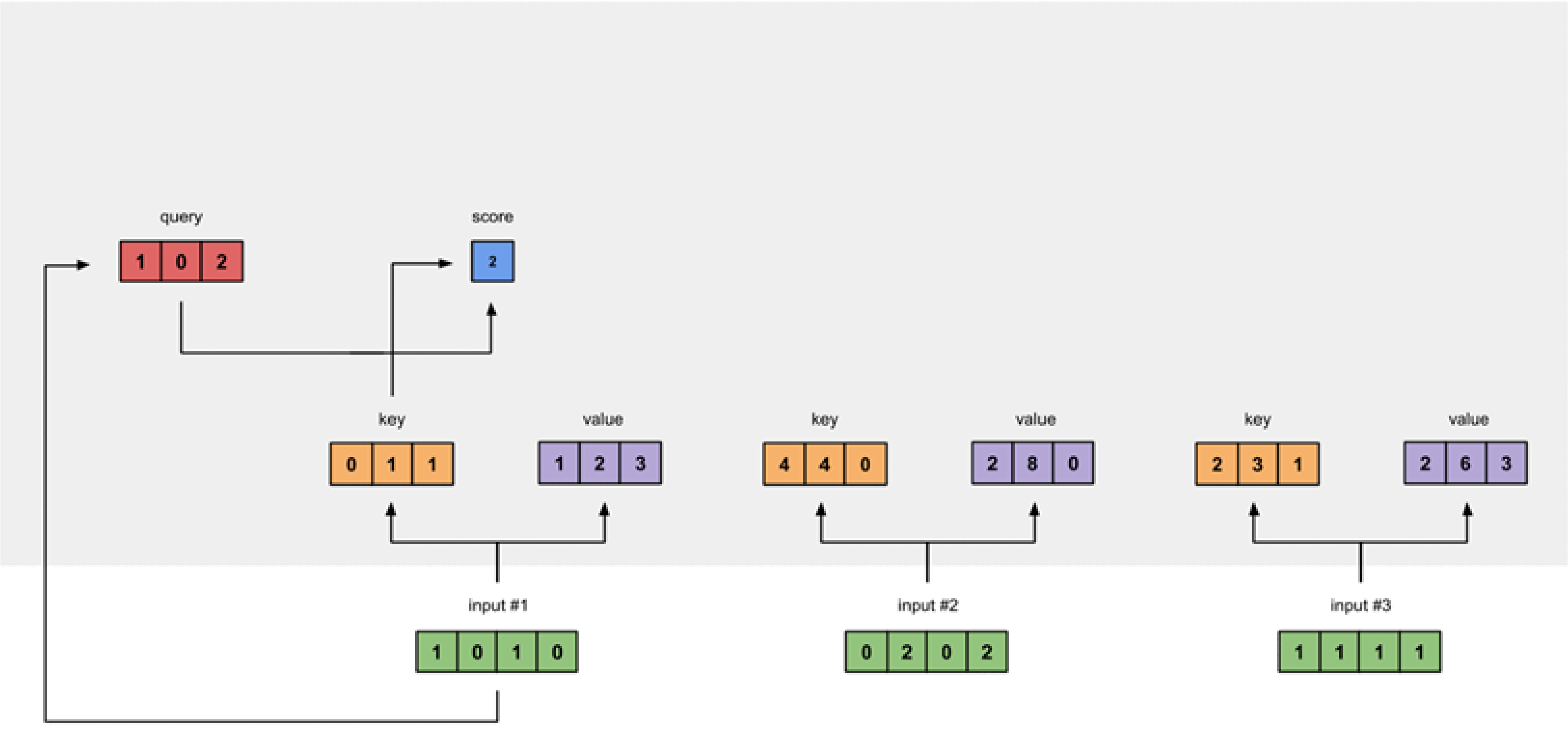
Step-by-Step



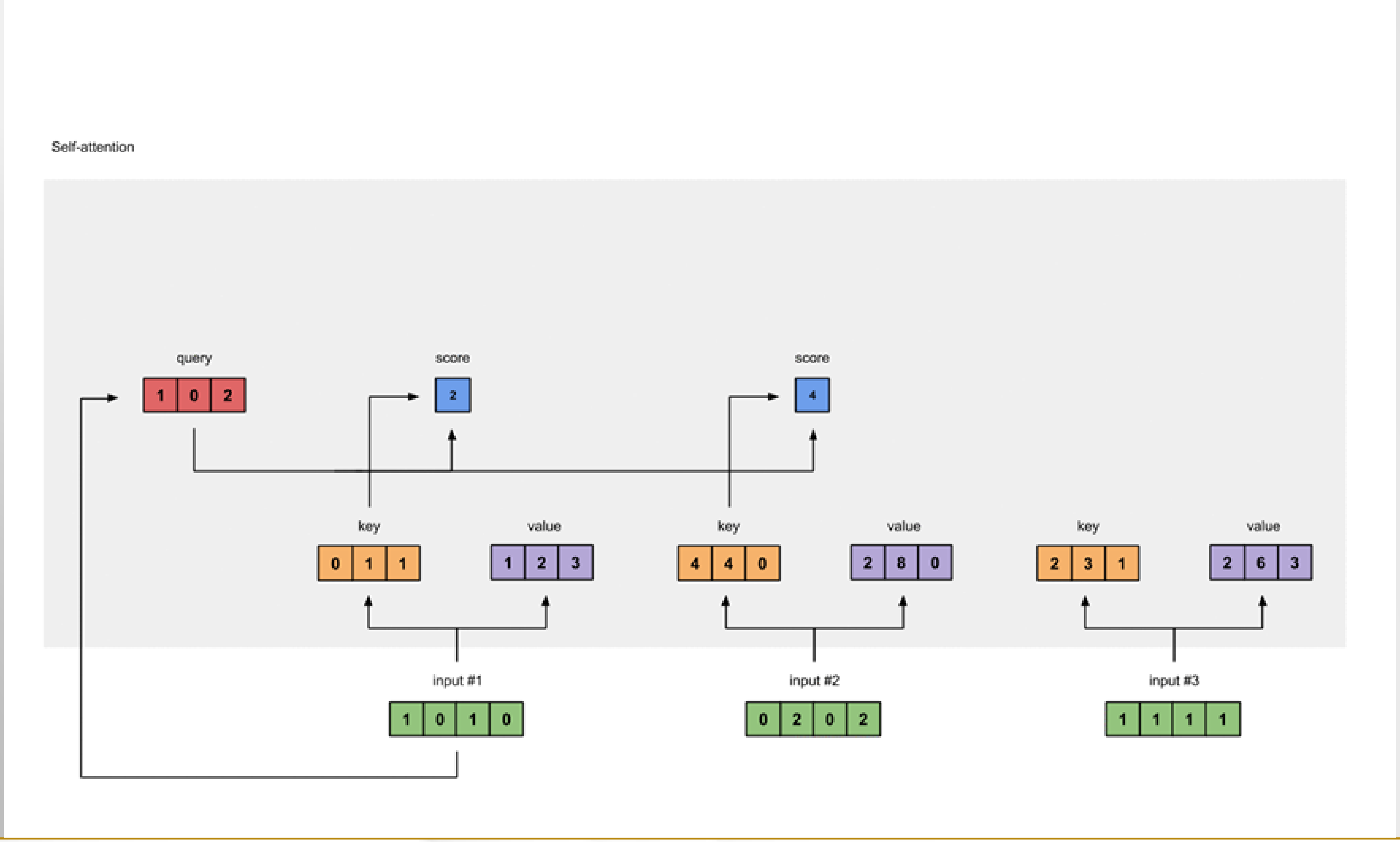
Self-attention
query
1 0 2
score
key
0 1 1
value
1 2 3
key
4 4 0
value
2 8 0
key
2 3 1
value
2 6 3
input #1
1 0 1 0
input #2
0 2 0 2
input #3
1 1 1 1

Step-by-Step / Self-Attention

Self-attention

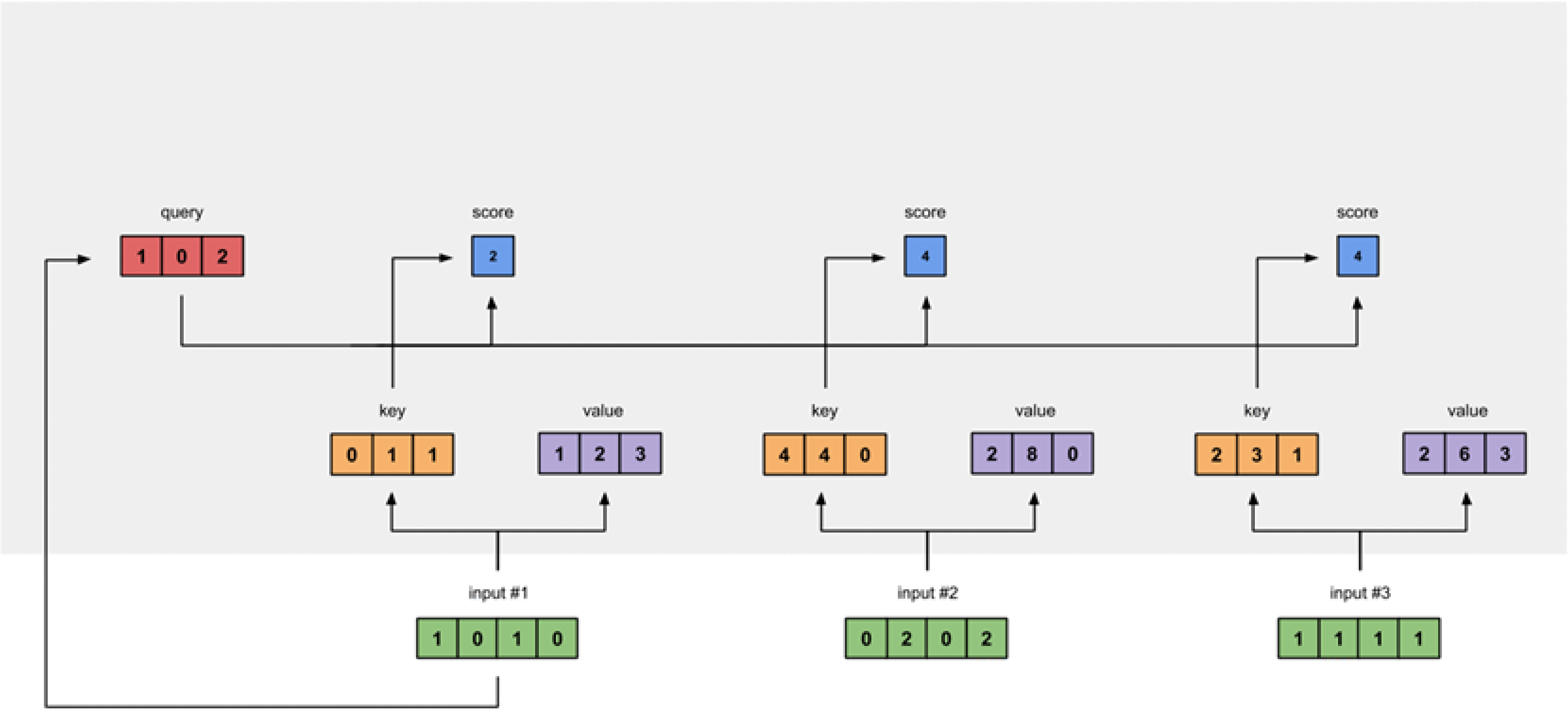


Step-by-Step / Self-Attention

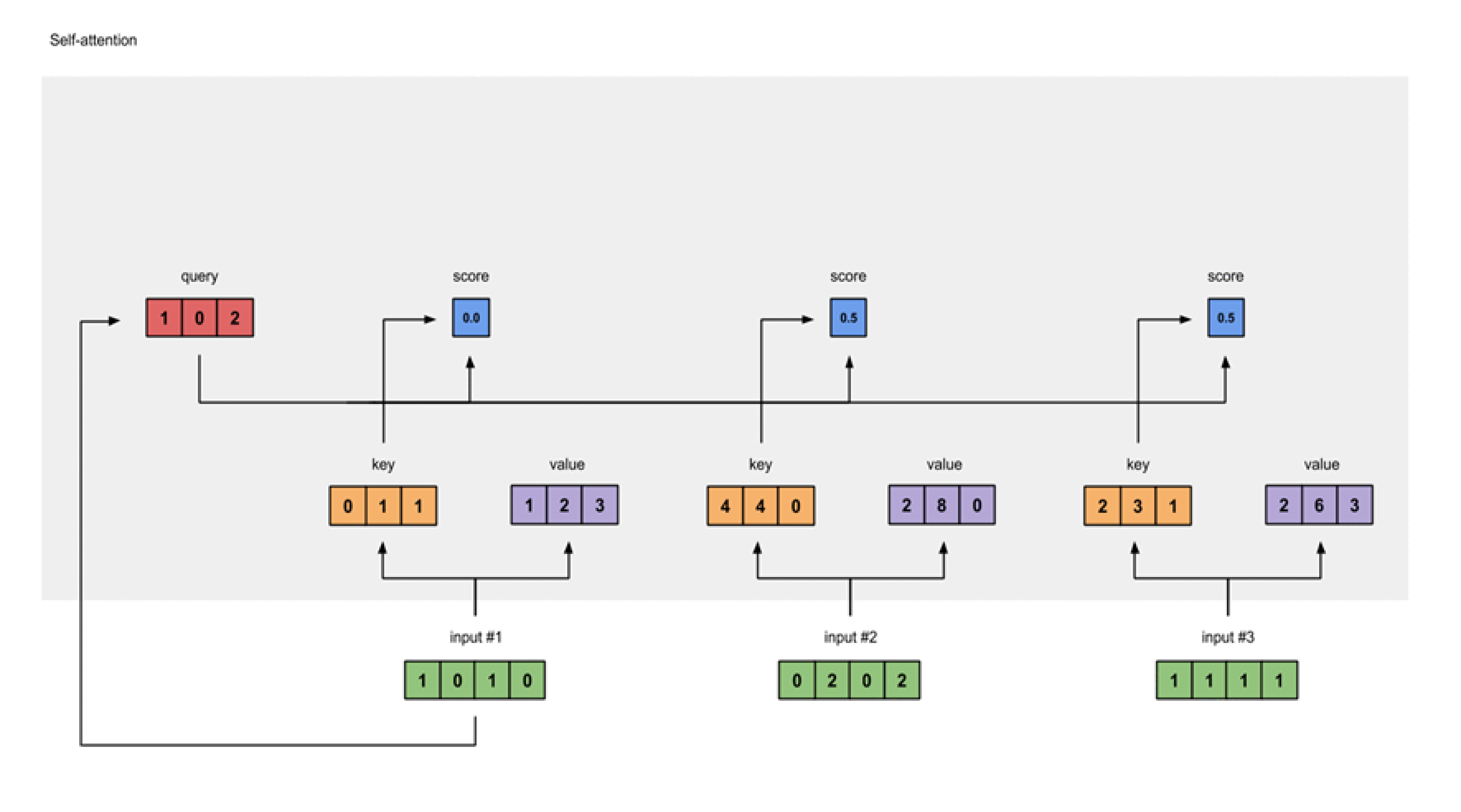


Step-by-Step / Self-Attention

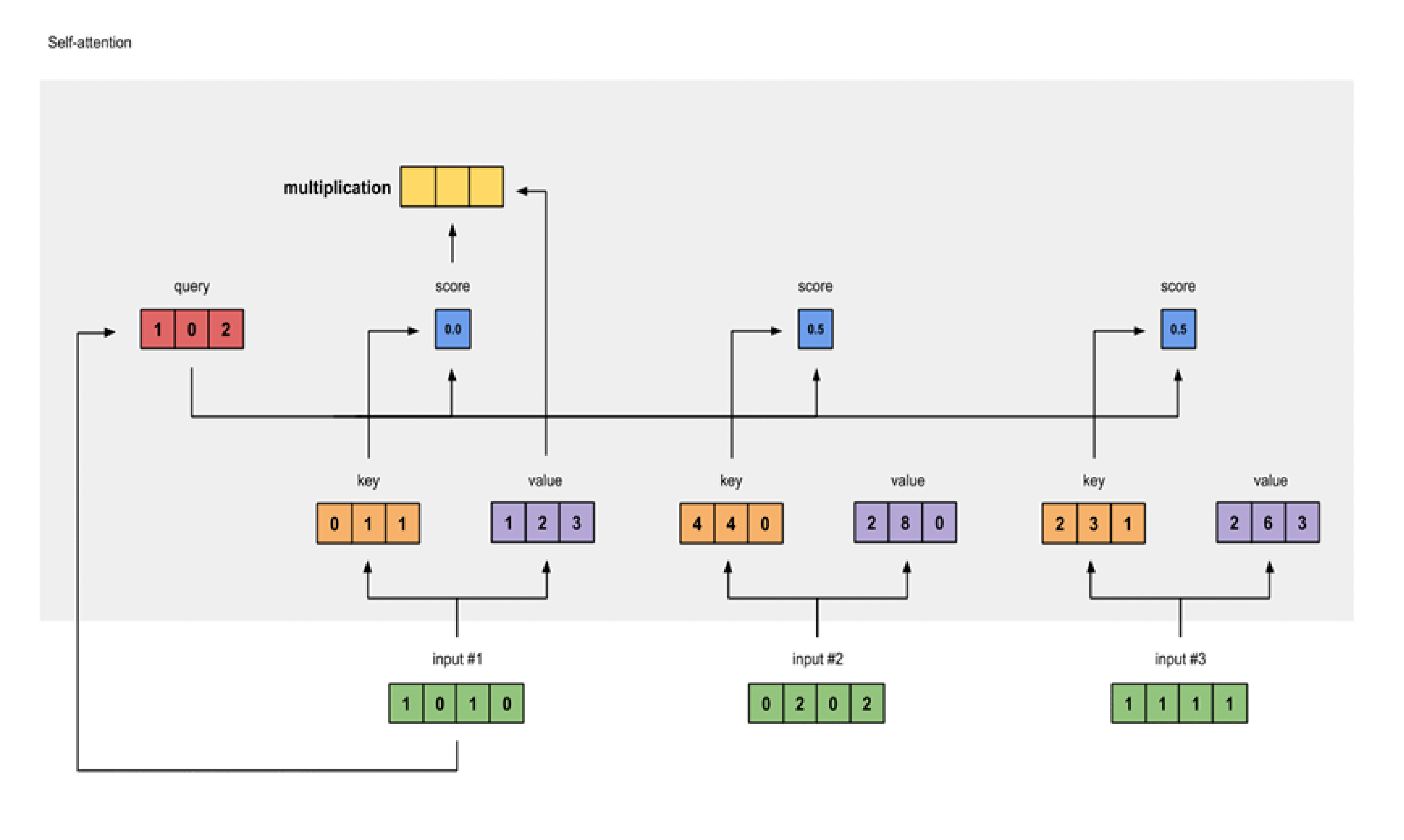
Self-attention



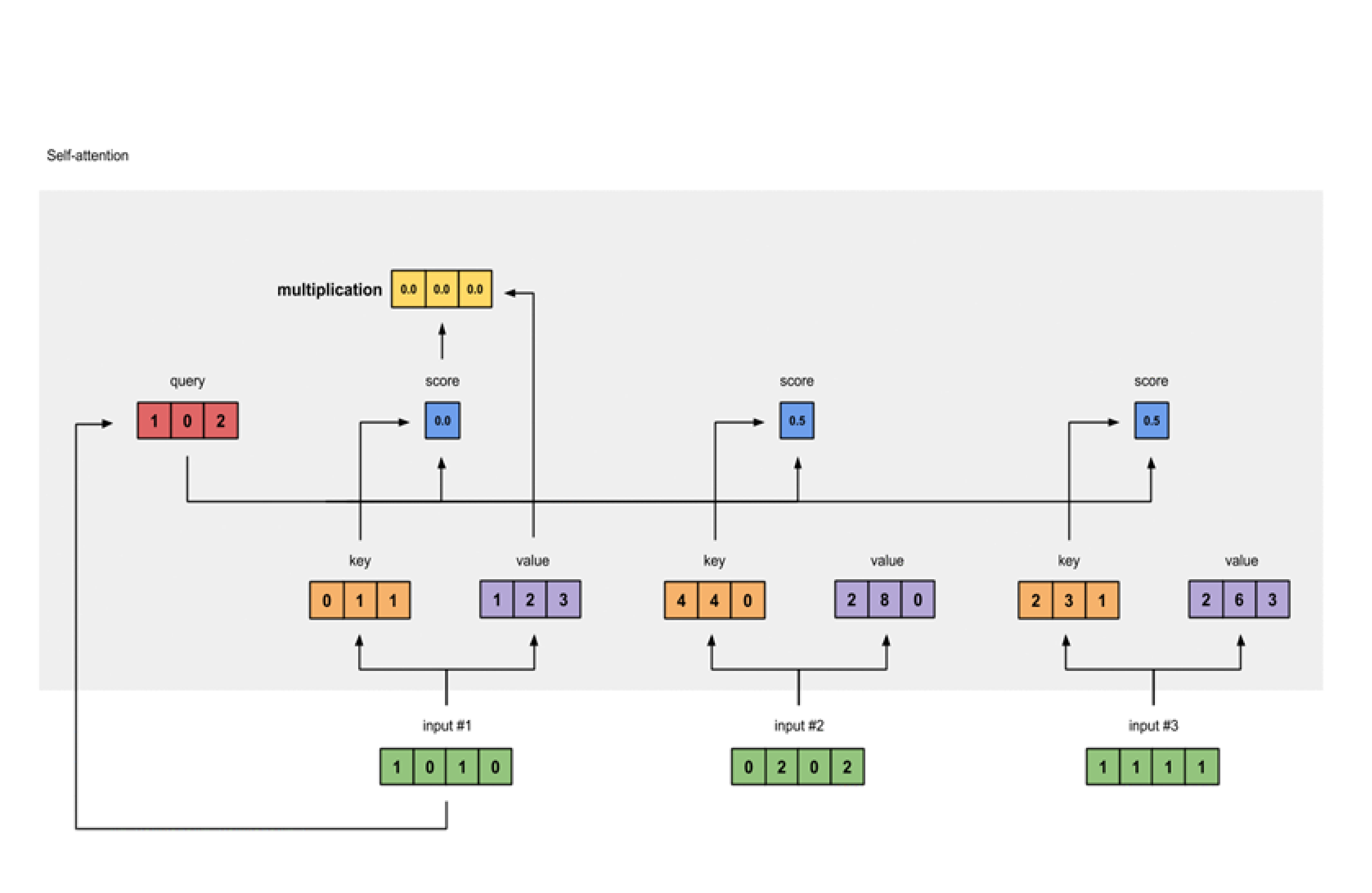
Step-by-Step / Self-Attention



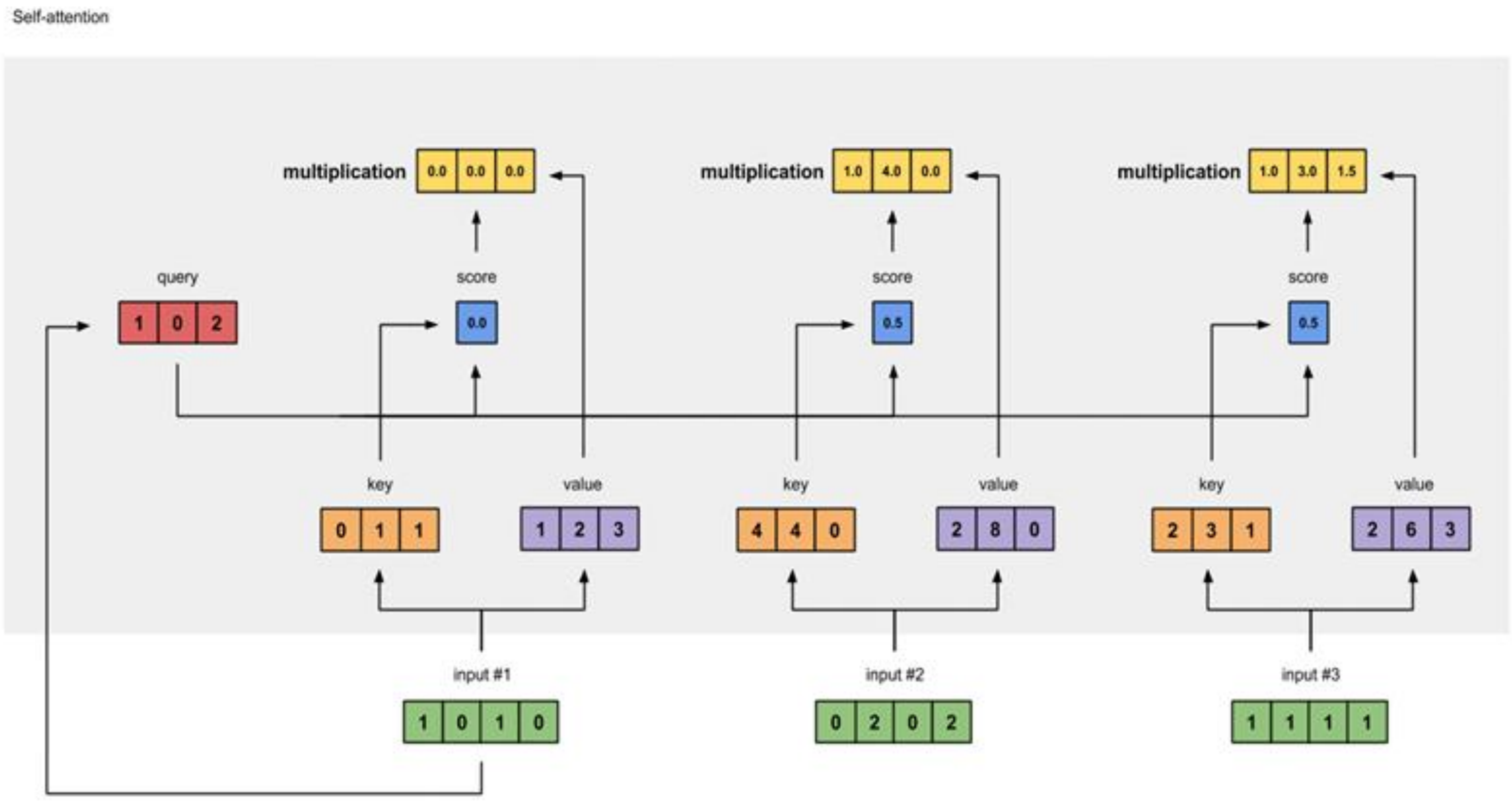
Step-by-Step / Self-Attention



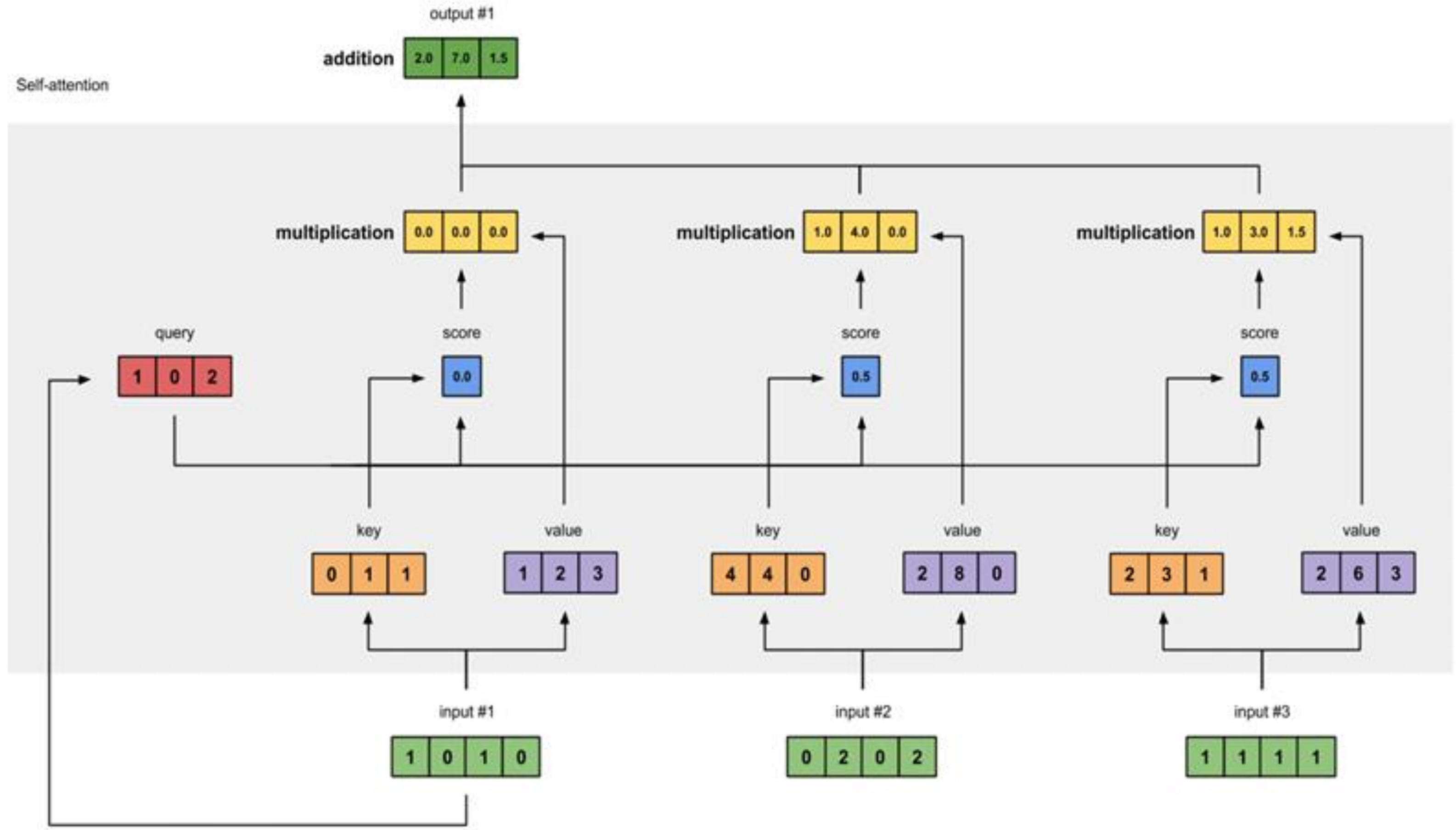
Step-by-Step / Self-Attention



Step-by-Step / Self-Attention

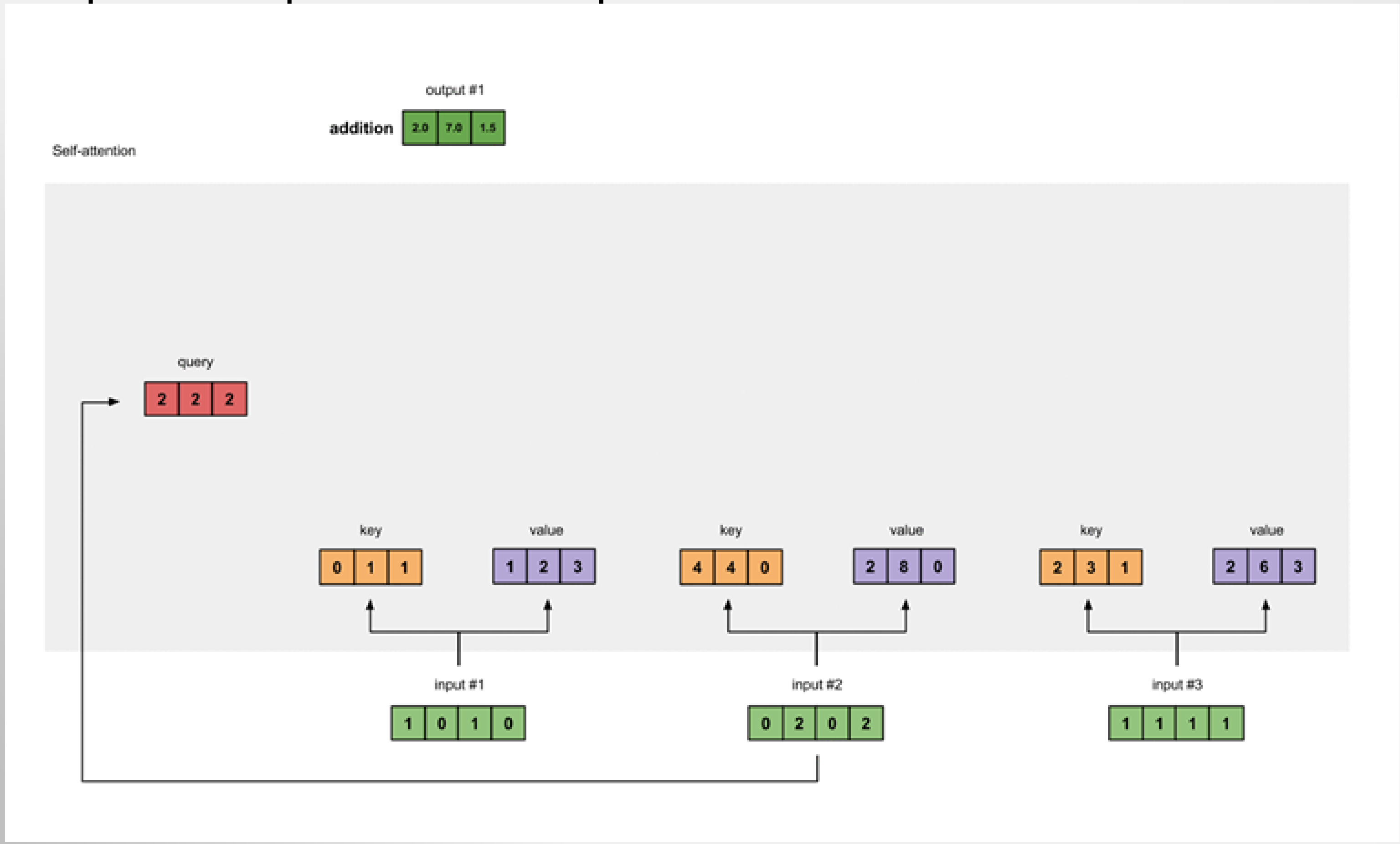


Step-by-Step / Self-Attention



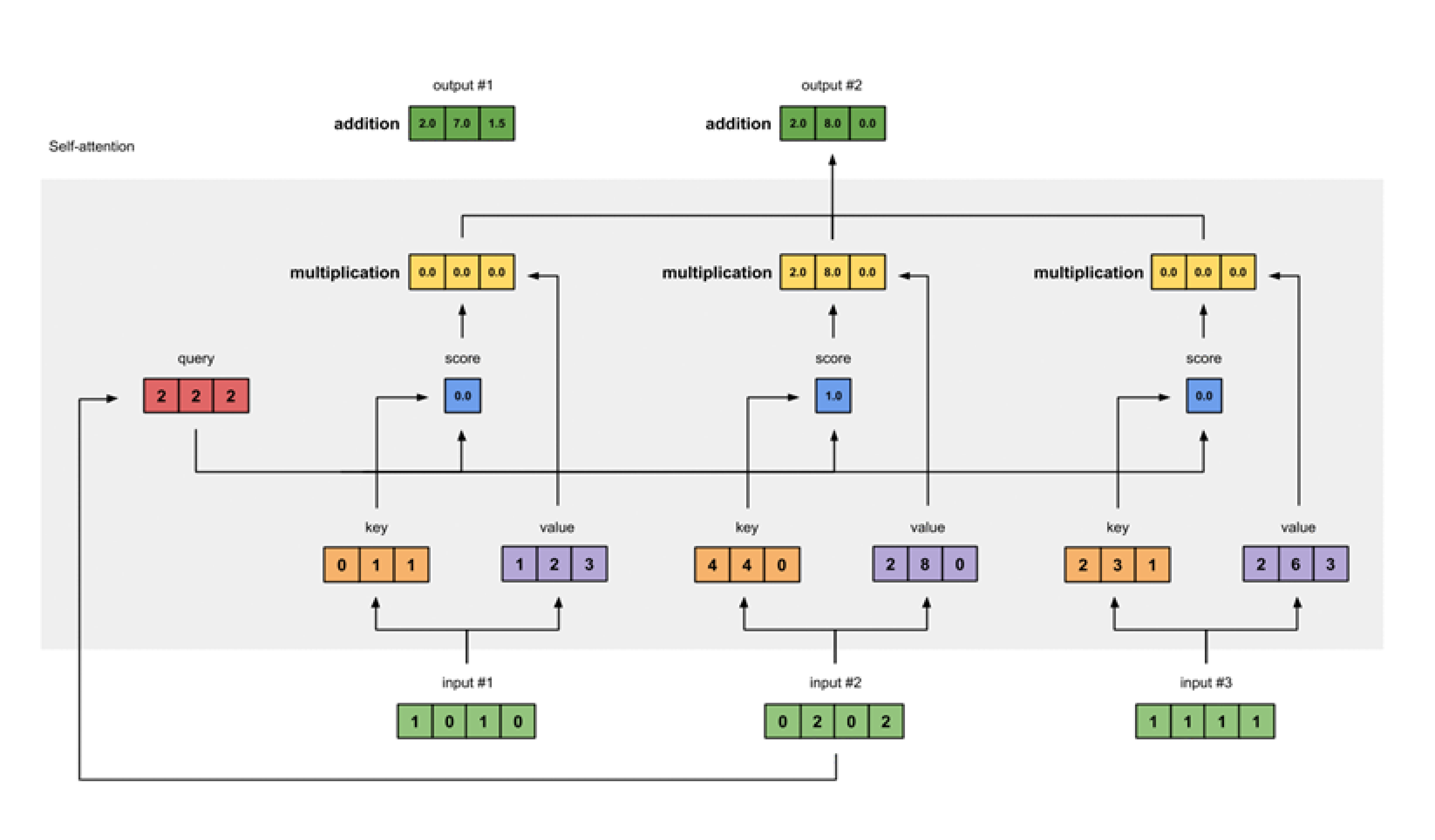
Step-by-Step / Self-Attention

Repeat the process for Input #2



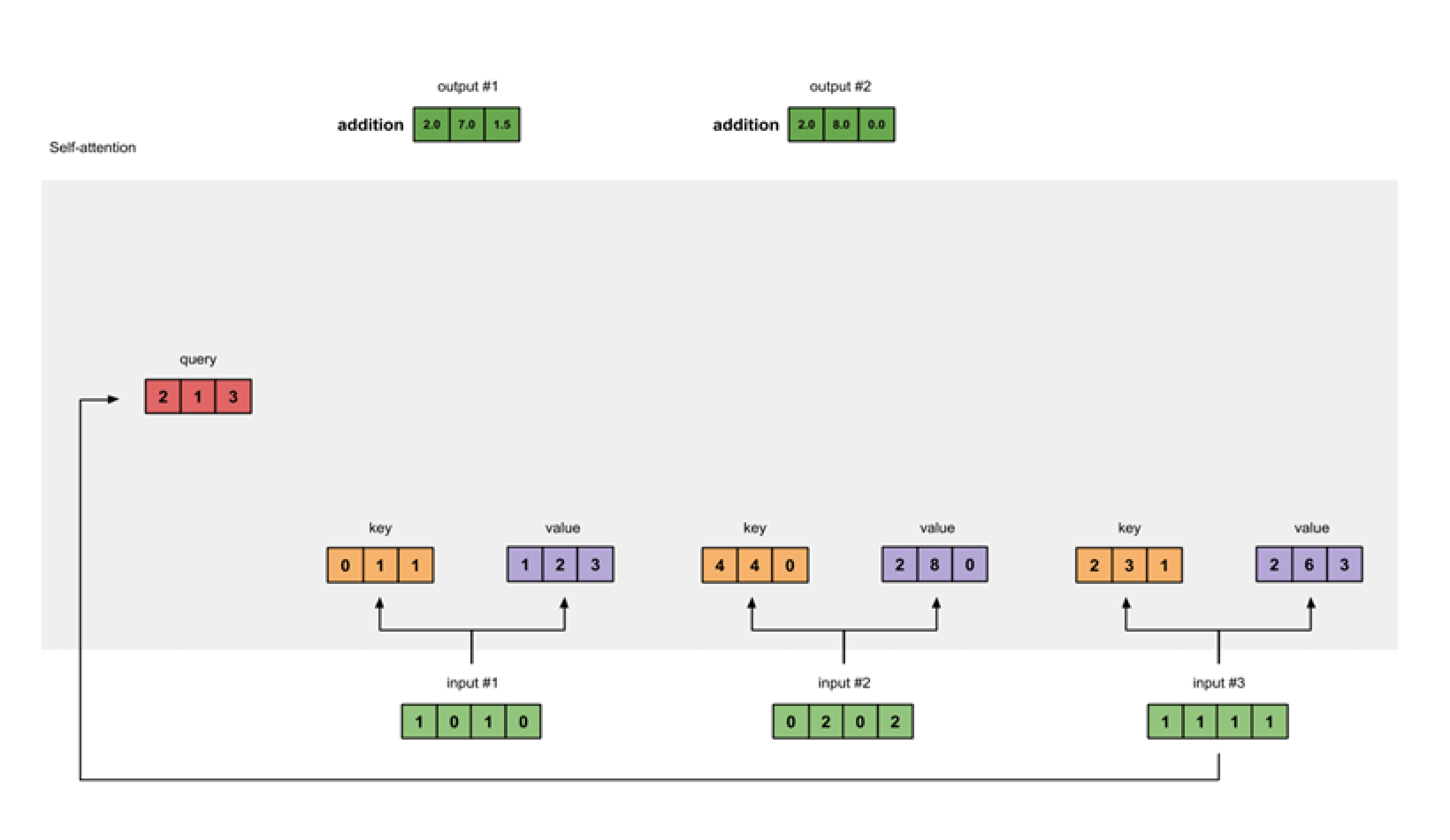
Step-by-Step / Self-Attention

Repeat the process for Input #2



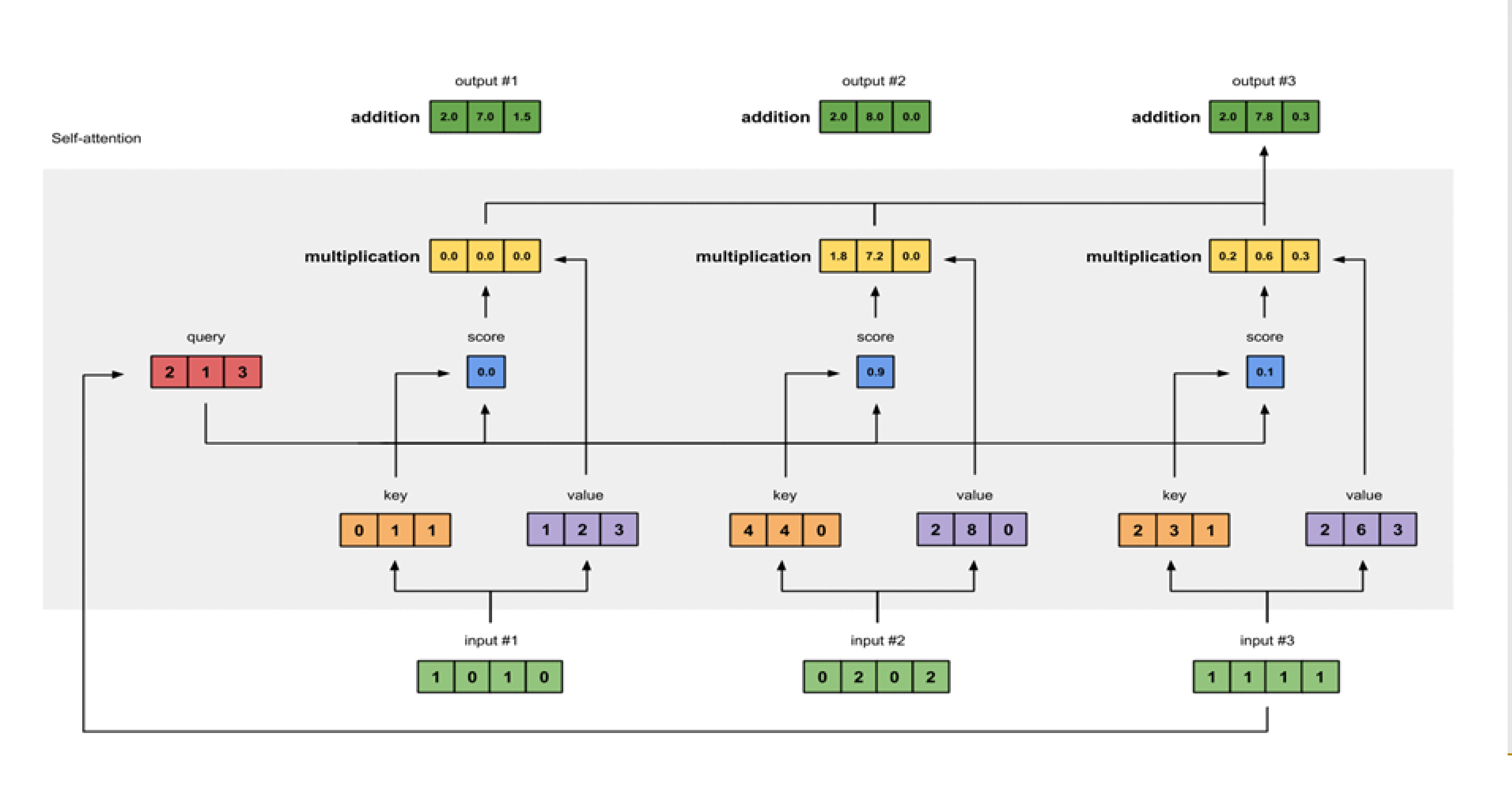
Step-by-Step / Self-Attention

Repeat the process for Input #3

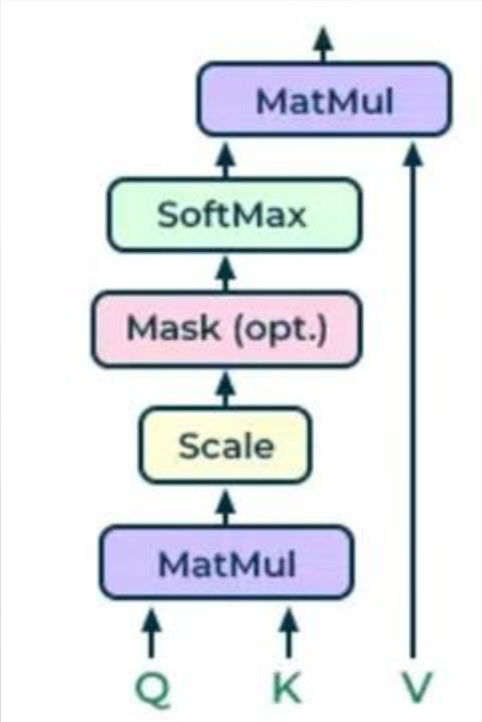


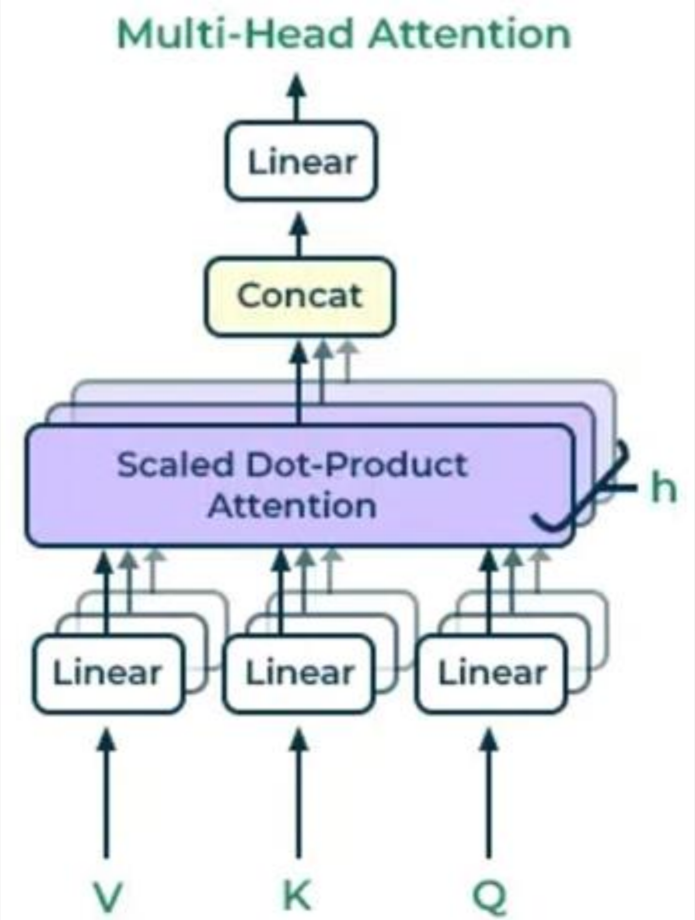
Step-by-Step / Self-Attention

Repeat the process for Input #3



Head


MatMul
SoftMax
Mask (opt.)
Scale
MatMul
Q
K
V


Multi-Head Attention
Linear
Concat
Scaled Dot-Product Attention
h
Linear
Linear
Linear
V
K
Q

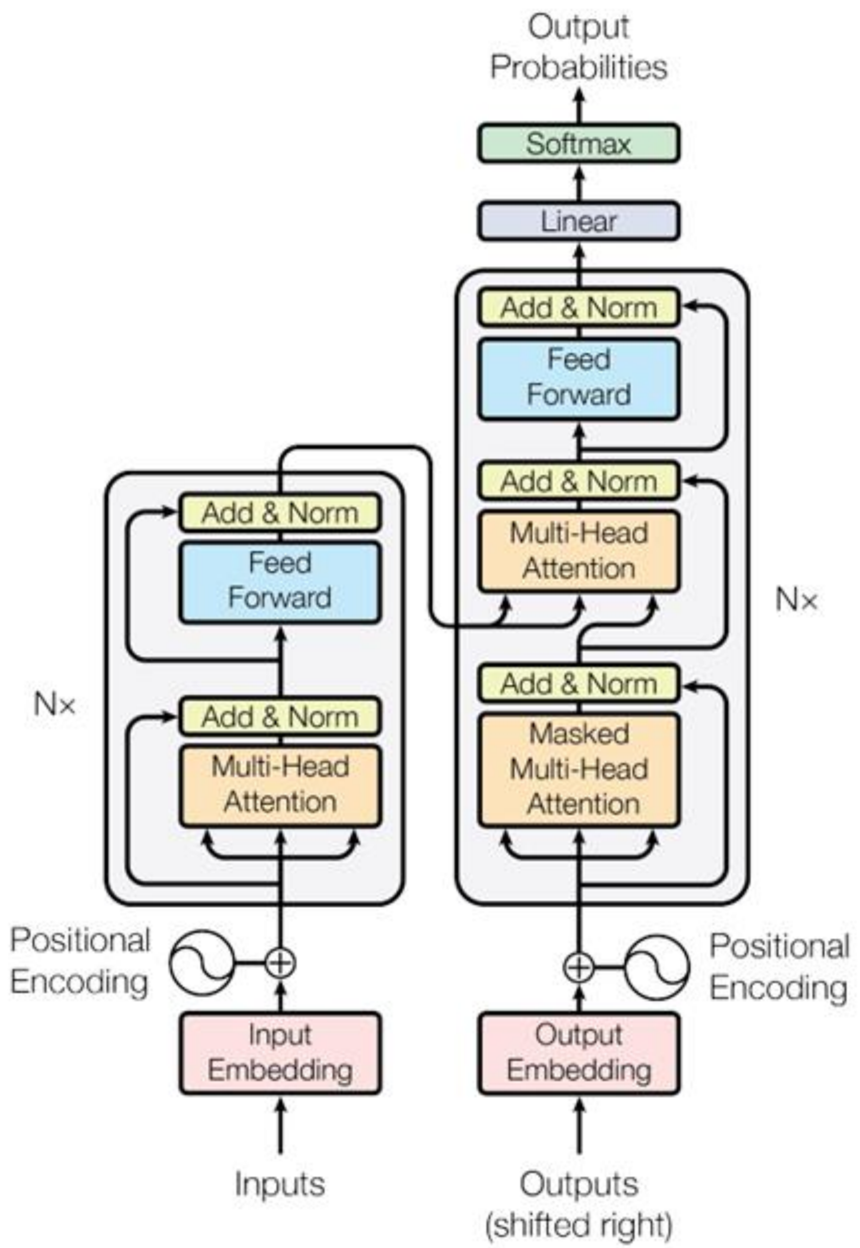
Transformer

➔ "Self-Attention" mechanism is the key component of Transformers

Extending to Transformers:

- Inputs to the self-attention module:
  - Embedding module
  - Positional encoding

- Modules between self-attention modules:
  - Linear transformations
  - LayerNorm



Referenc

[1] Daniel Jurafsky and James H. Martin. 2024. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition. Online manuscript released August 20, 2024. https://web.stanford.edu/~jurafsky/slp3.

[2] https://towardsdatascience.com/illustrated-self-attention-2d627e33b20a

[3] https://user.phil.hhu.de/~cwurm/wp-content/uploads/2020/01/7181-attention-is-all-you-need.pdf

[4] https://www.geeksforgeeks.org/self-attention-in-nlp/

[5] https://www.youtube.com/watch?v=dqoEU9Ac3ek - MIT 6.S191: Recurrent Neural Networks, Transformers, and Attention