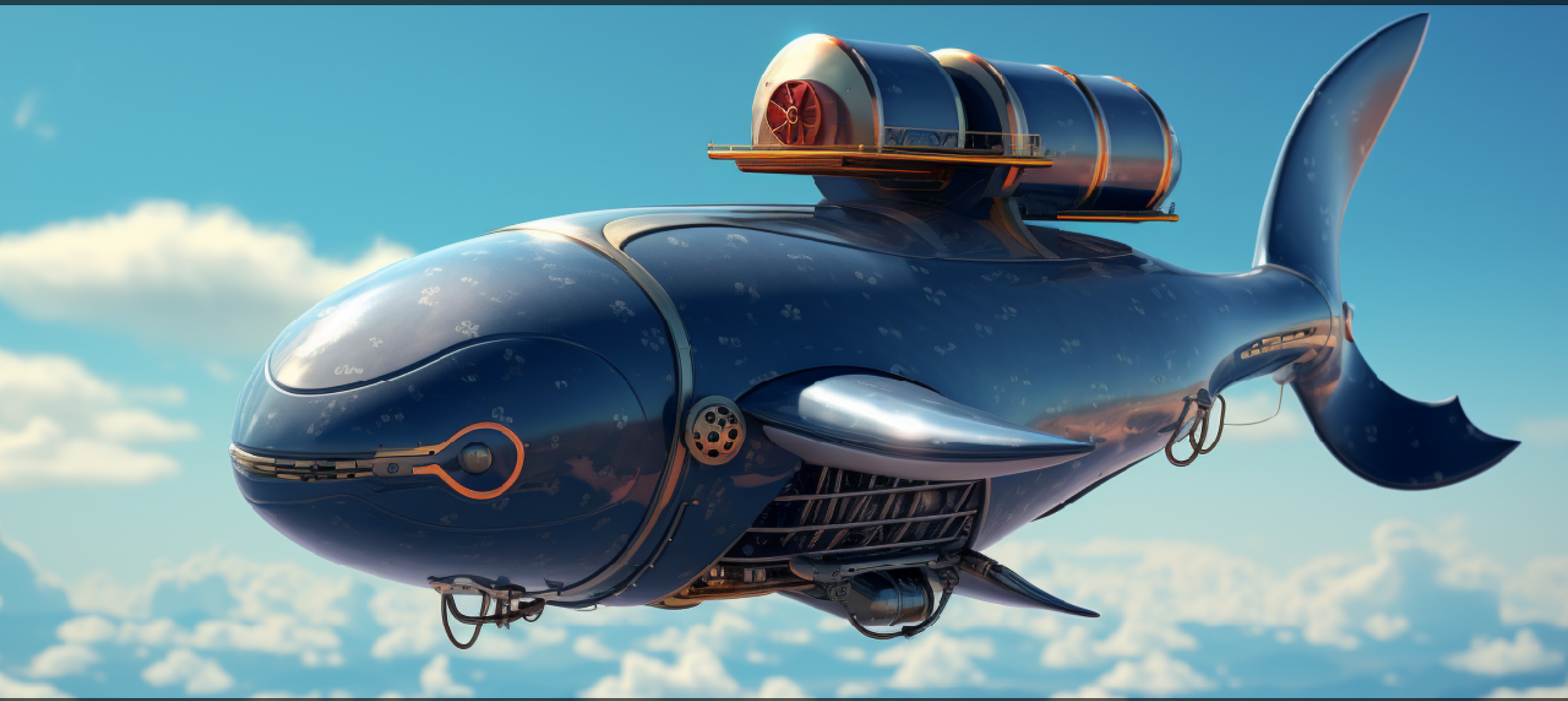
Transformerを詳しく学ぼう！



Section2

Section2の概要

講座の内容

Section1. Transformerの概要

Section2. Attentionの仕組み

Section3. Transformerにおける埋め込み

Section4. Transformerを組み立てる

今回の内容

1. Section2の概要
2. Attentionの概要
3. Scaled Dot-Product Attention
4. Multi-HeadAttention
5. 演習

教材の紹介

- **Pythonの基礎:**

  python_basic

- **Section2の教材**

  01_scaled_dot_product _attention.ipynb

  02_multi_head_attention.ipynb

  03_exercise.ipynb

https://github.com/yukinaga/learning_transformer/

Section1演習の解答例

- 03_exercise.ipynb

Attentionの概要

ChatGPT に聞いてみる

「Transformerで使われるAttentionって何ですか？」

→ ?

https://chat.openai.com/

Attentionとは？

- **Attentionとは？**

  → 文章中のどの単語に注目すればいいかを表すスコア

  → Query、Key、Valueの3つのベクトルで計算される

- **Query**

  → Inputのうち「検索をかけたいもの」

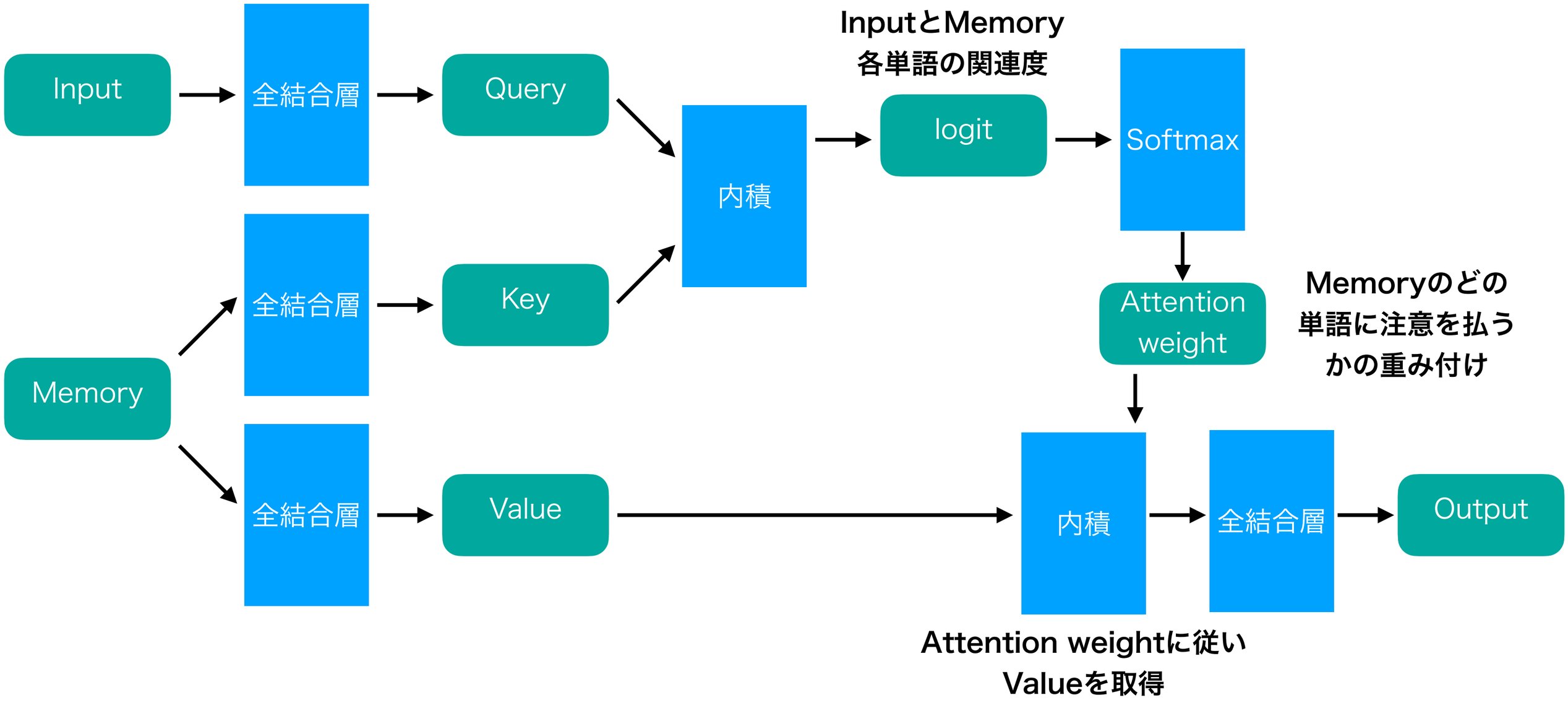
- **Key**

  → 検索対象とQueryの近さを測る

- **Value**

  → Keyに基づき、適切なValueを出力する

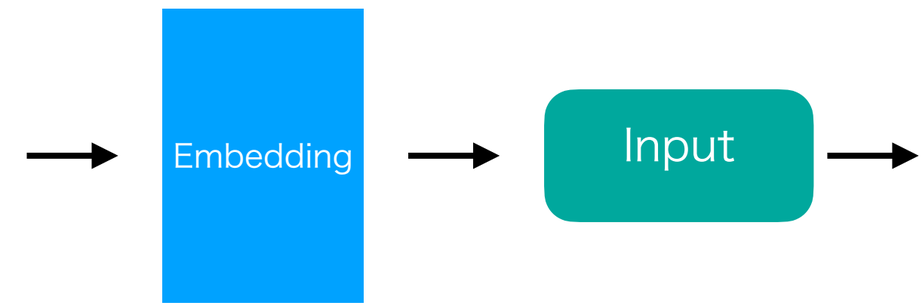
Attentionとは？



InputとMemory

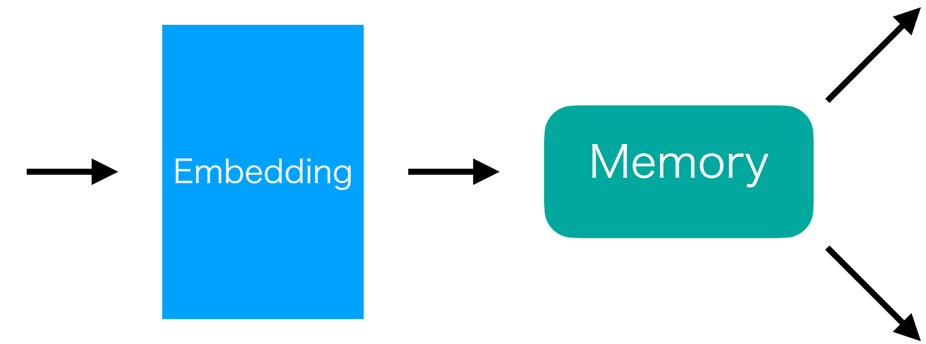
得意　な　スポーツ　は　？

（各単語はidで表される）

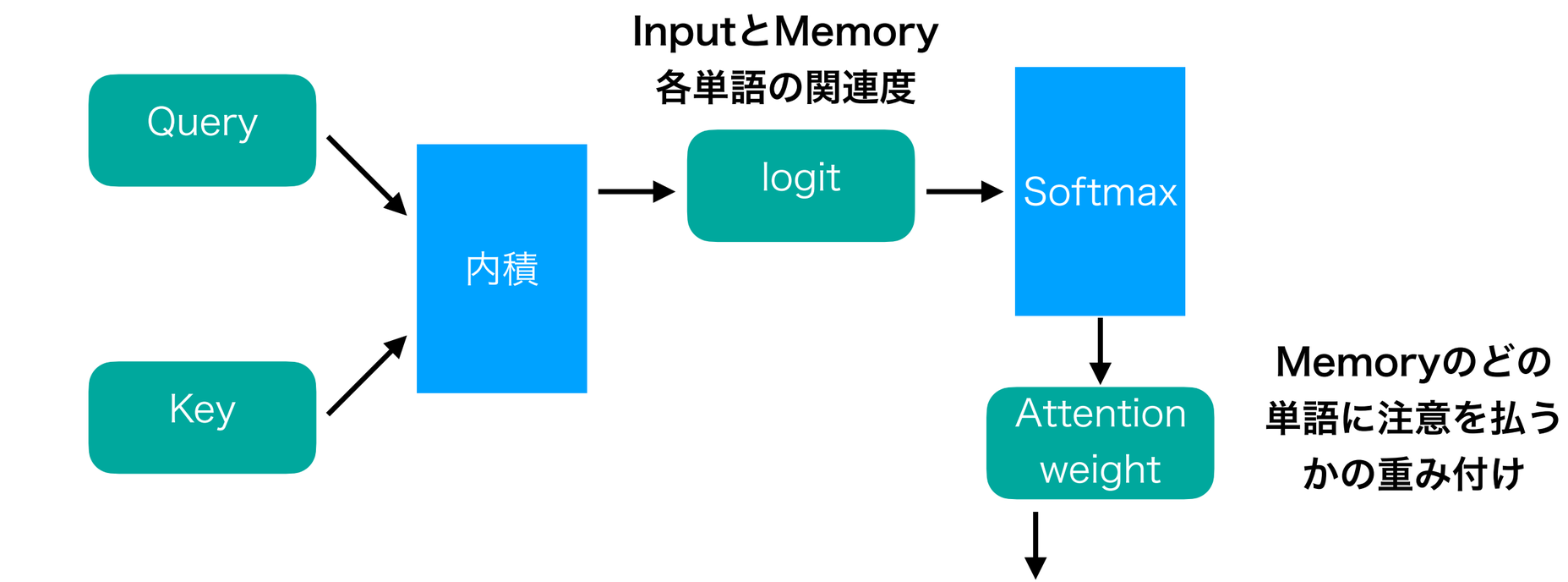


野球　が　得意

（各単語はidで表される）



Attention weightの計算

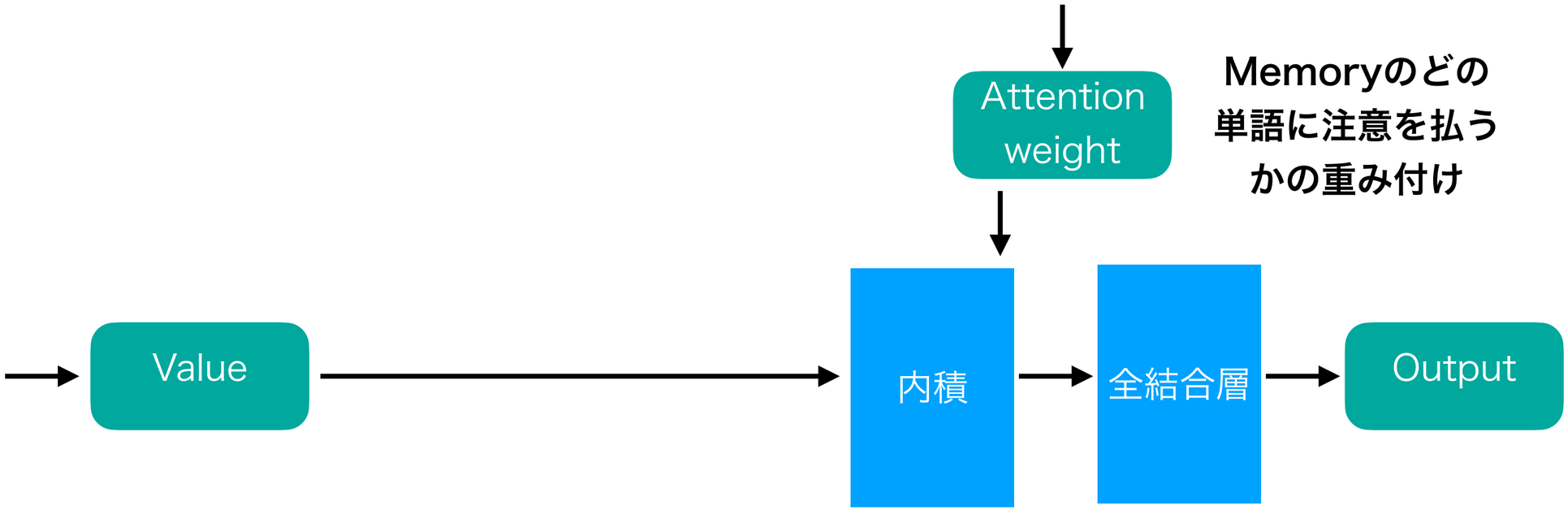


Input: 得意 な スポーツ は ？

0.6 0.02 0.38

Memory: 野球 が 得意

Valueとの内積



Input: 得意 な スポーツ は ？

Memory: 野球 が 得意

0.6 0.02 0.38

内積 = Value(野球)×0.6

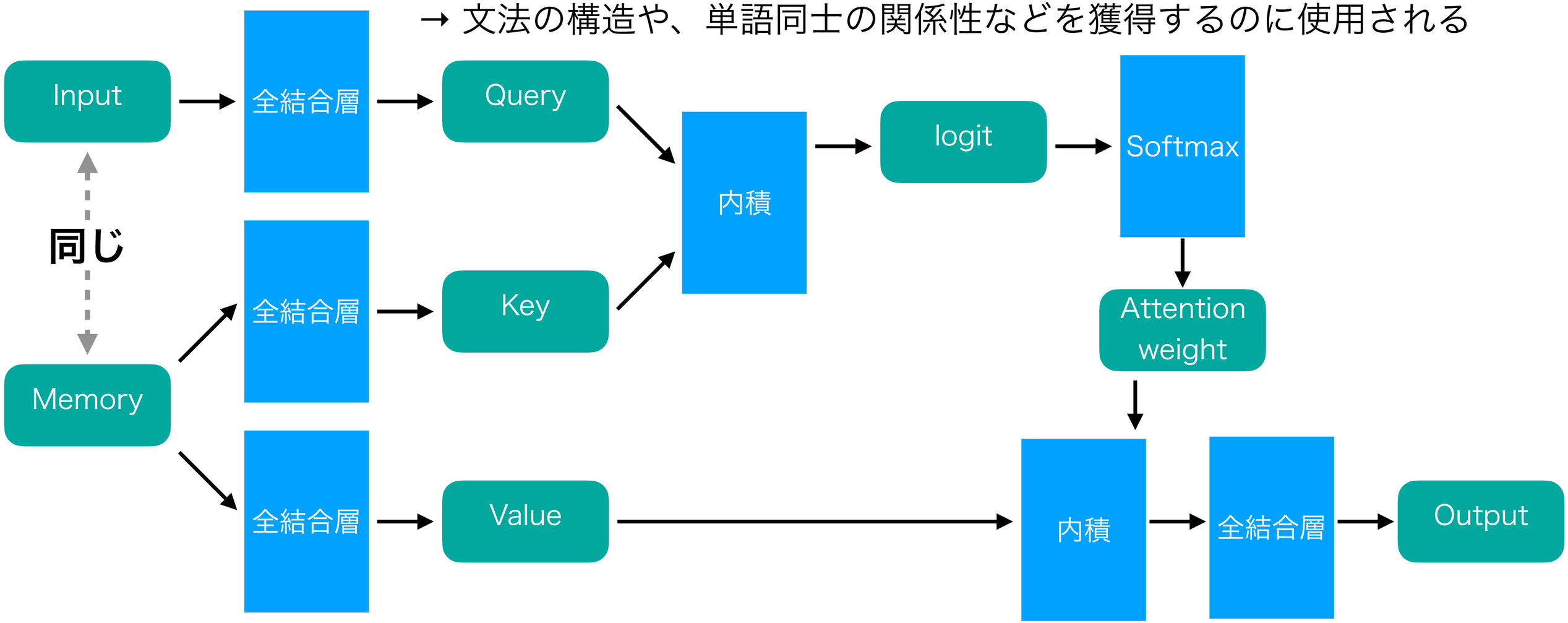
+Value(が)×0.02

+Value(得意)×0.38

Self-Attention

- **Self-Attention**
  → InputとMemoryが同一のAttention

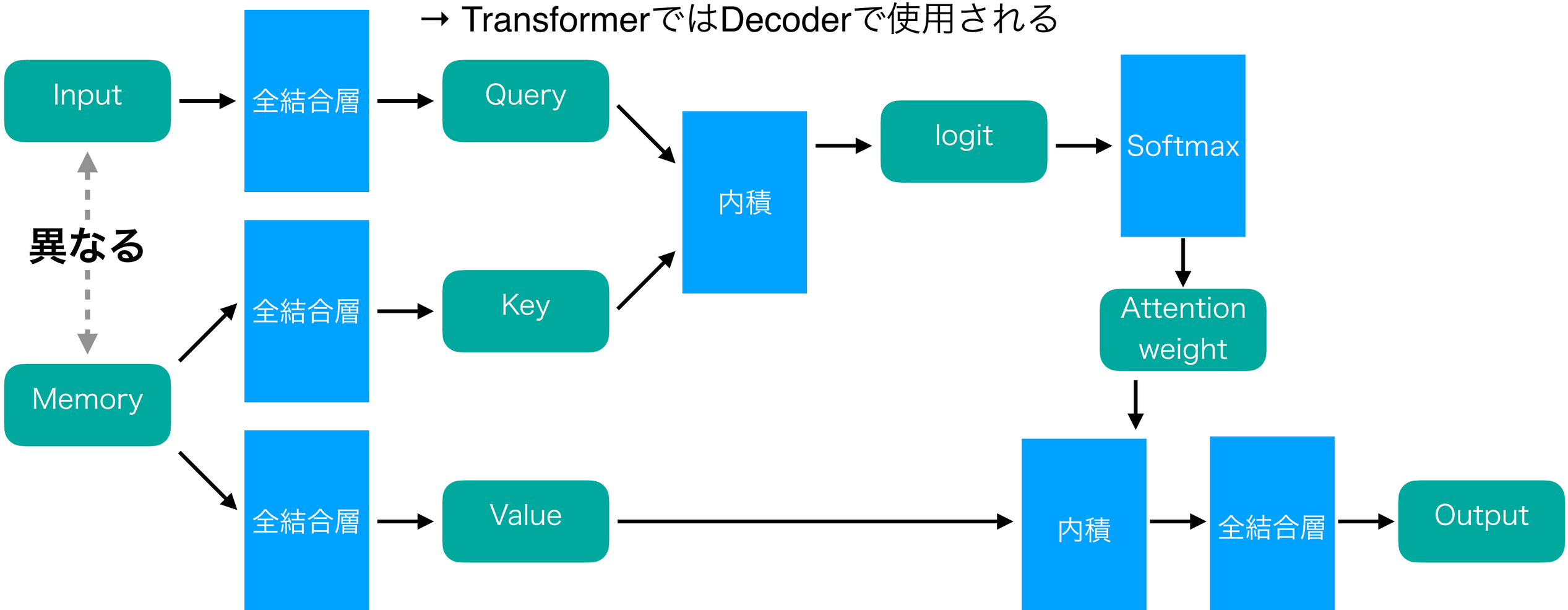
  → 文法の構造や、単語同士の関係性などを獲得するのに使用される



SourceTarget-Attention

- **Self-Attention**
  → InputとMemoryが異なるAttention

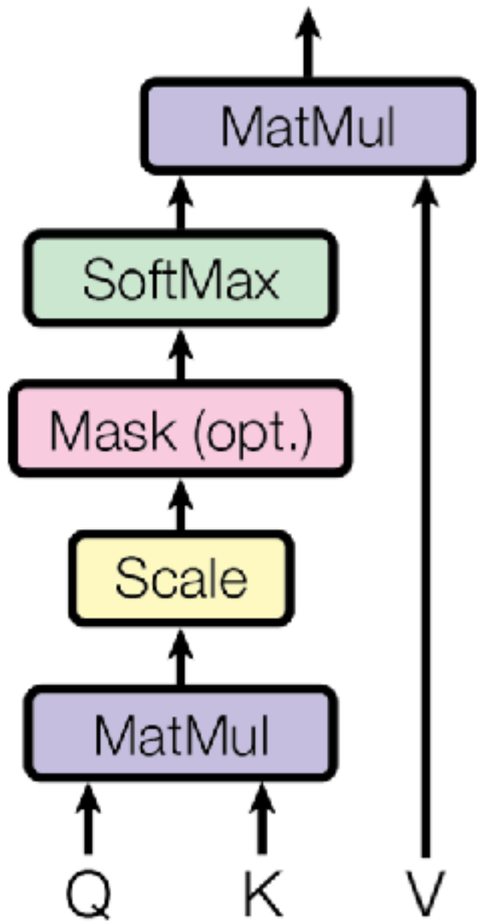
  → TransformerではDecoderで使用される



Scaled Dot-Product Attention

- **Scaled Dot-Product Attention**

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



- **Mask**

  → 入力した単語が「先読み」されるのを防

  ぐために、特定の key に対して

  Attention weight を0にする

Attention Is All You Need, Ashish, V. et al. (2017) より引用

Multi-Head Attention

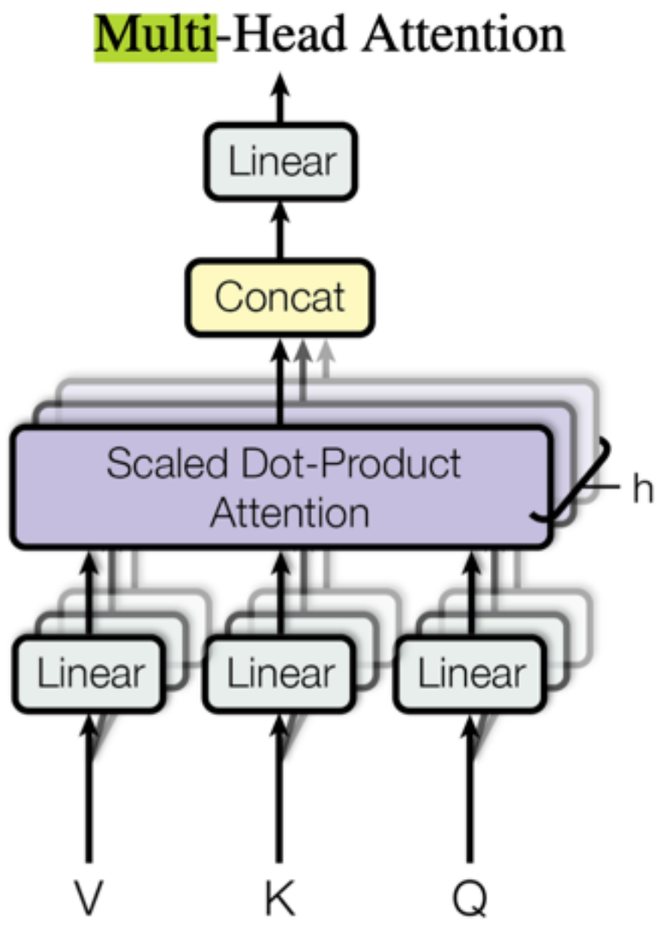
- **Multi-Head Attention**
  → Attentionを並行に並べる

  → それぞれのAttentionはHeadと呼ばれる

  → 「Attention Is All You Need」では

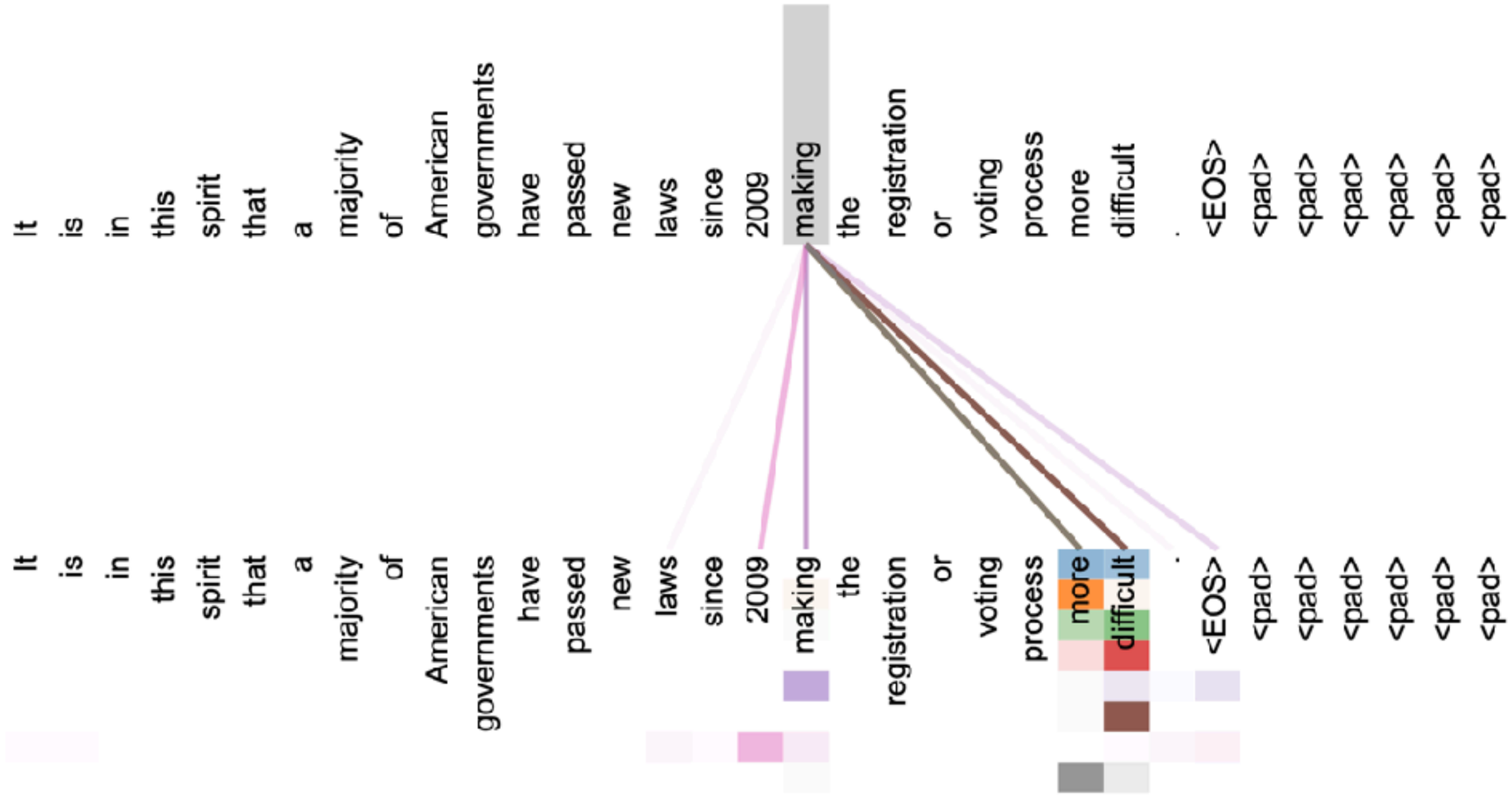
  Multi-Head化による性能の向上が述べられている

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, ..., \mathrm{head}_\mathrm{h})W^O$$
$$\text{where } \mathrm{head}_\mathrm{i} = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$



Attention Is All You Need, Ashish, V. et al. (2017) より引用

Attentionの可視化



異なる色は異なるAttentionのHeadを表す

Attention Is All You Need, Ashish, V. et al. (2017) より引用

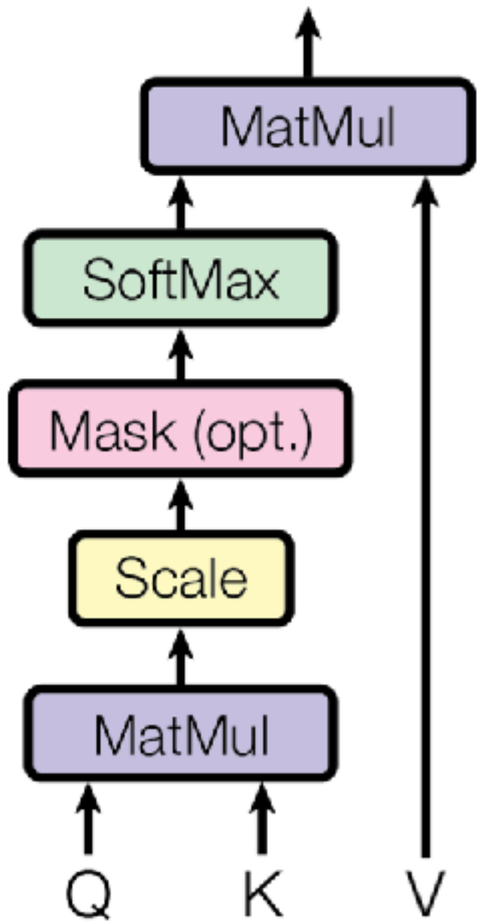
休憩…

Scaled Dot-Product Attention

Scaled Dot-Product Attention

- **Scaled Dot-Product Attention**

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



- **Mask**

  → 入力した単語が「先読み」されるのを防

  ぐために、特定の key に対して

  Attention weight を0にする

Attention Is All You Need, Ashish, V. et al. (2017) より引用

Scaled Dot-Product Attention

- 01_scaled_dot_product _attention.ipynb

Multi-HeadAttention

Multi-Head Attention

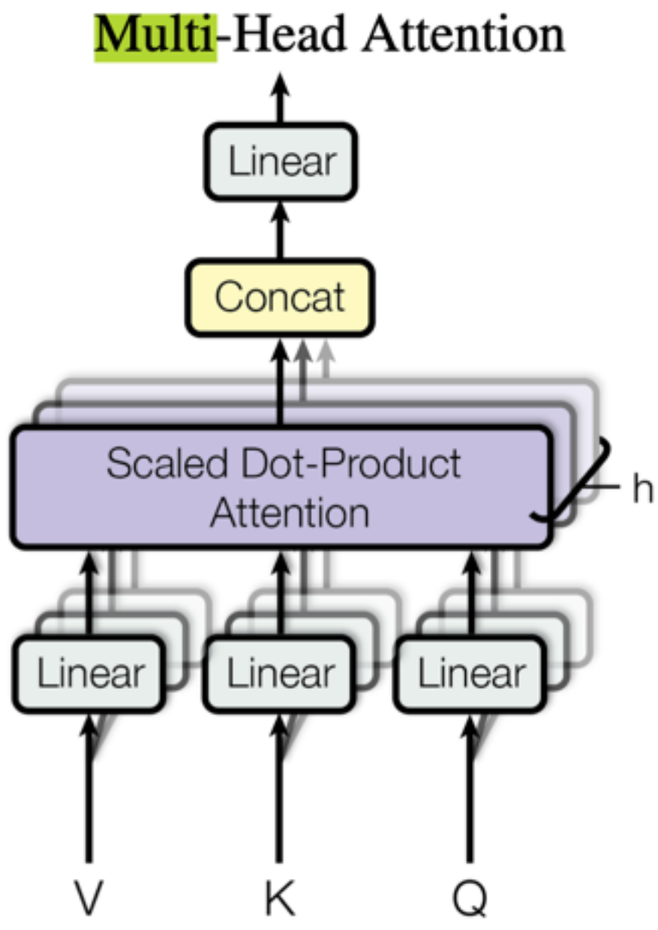
- **Multi-Head Attention**
  → Attentionを並行に並べる

  → それぞれのAttentionはHeadと呼ばれる

  → 「Attention Is All You Need」では

  Multi-Head化による性能の向上が述べられている

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, ..., \mathrm{head}_\mathrm{h})W^O$$
$$\text{where } \mathrm{head}_\mathrm{i} = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$



Attention Is All You Need, Ashish, V. et al. (2017) より引用

Multi-HeadAttention

- 02_multi_head_attention.ipynb

演習

演習

- 03_exercise.ipynb

次回の内容

Section1. Transformerの概要

Section2. Attentionの仕組み

Section3. Transformerにおける埋め込み

Section4. Transformerを組み立てる