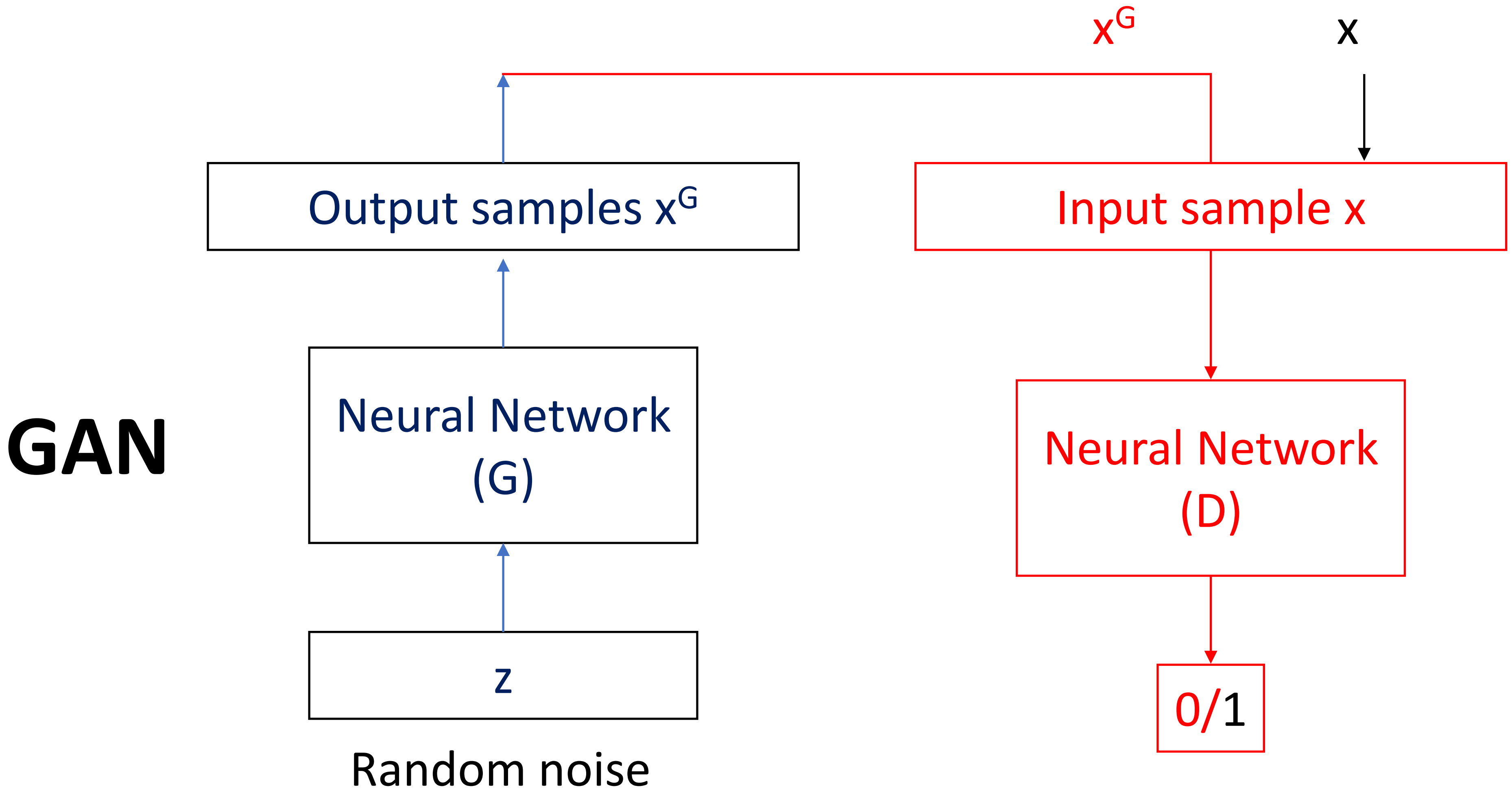
LECTURE 8:

ATTENTION AND TRANSFORMER

University of Washington, Seattle

Fall 2024

Previously in EEP 596...



OUTLINE

Part 1: Transformer motivation

- Limitation of RNNs with sequence data
- Seq2seq and attention
- Attention is all you need

Part 2: Self-attention layer

- Overview
- Key, Query and Value retrieval process
- Multi-headed attention

Part 3: Transformer architecture

- Encoder
- Decoder
- Transformer vs RNN

Part 4: Transformer applications

- NLP
- Computer vision
- Multi-modal
- Signal processing

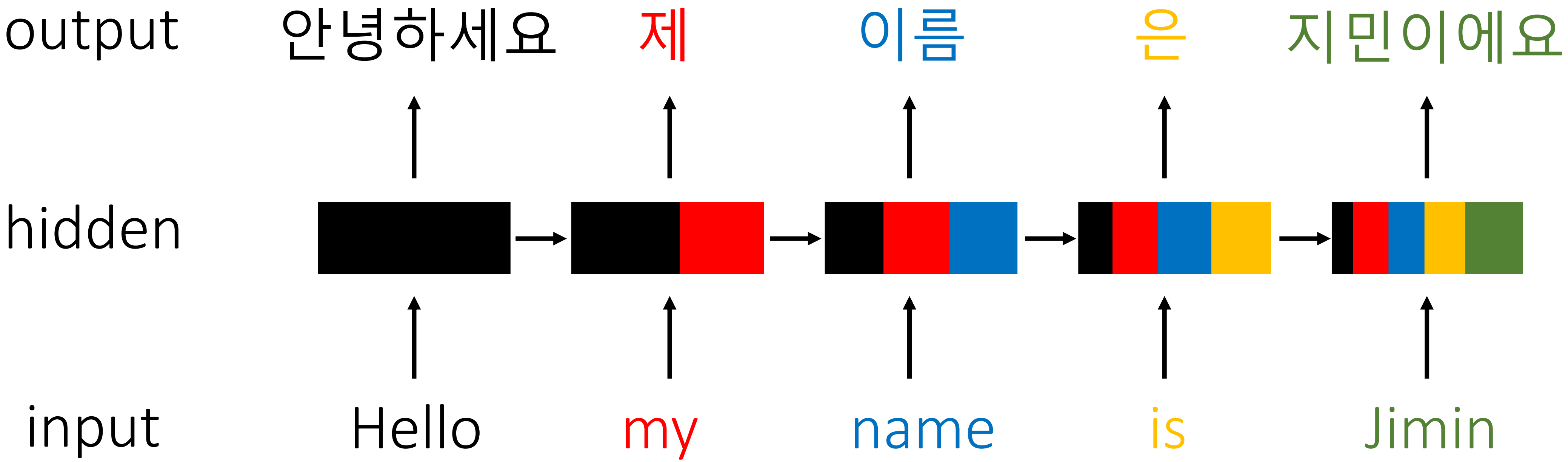
Transformer Motivation

Limitations of RNNs with sequence data

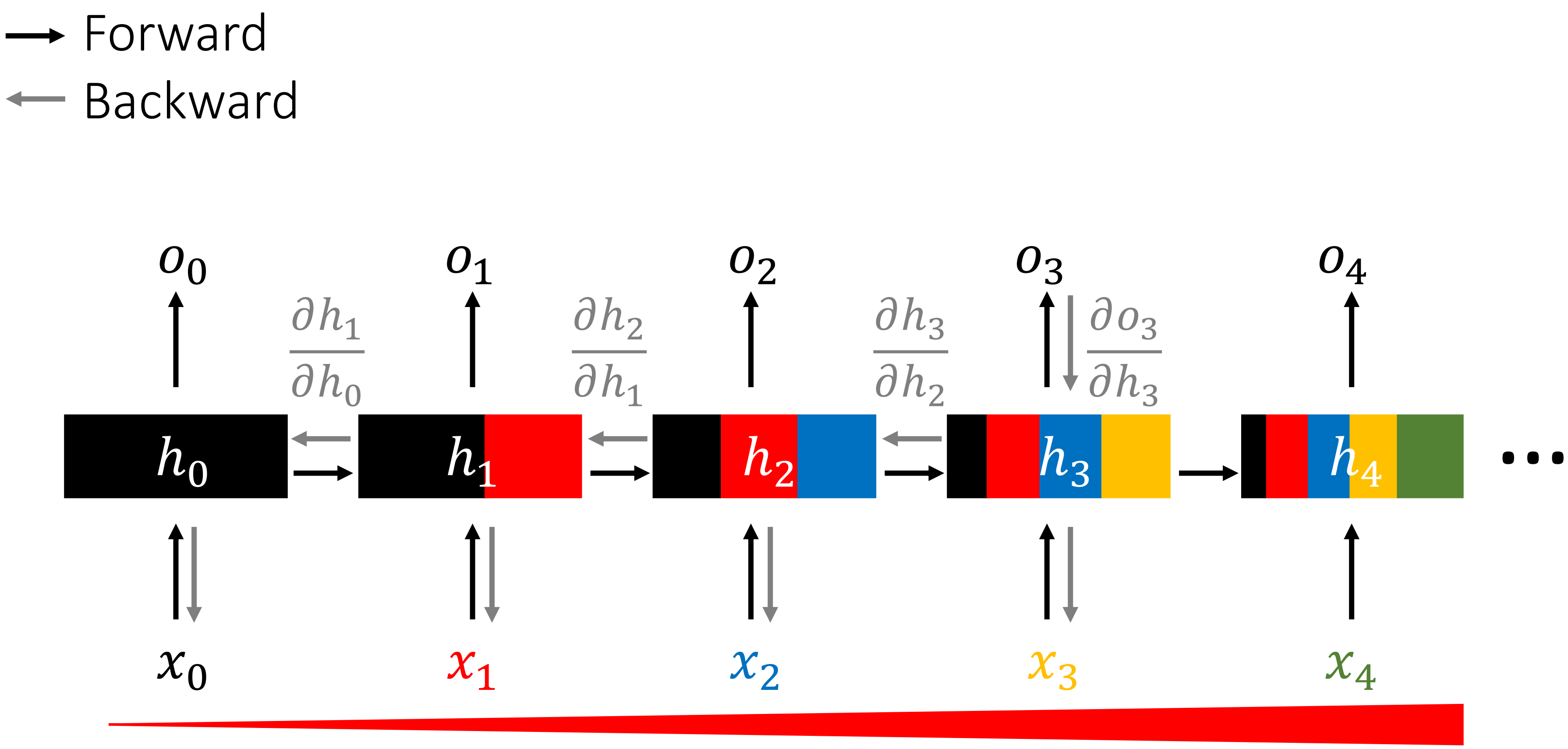
Seq2Seq and attention

Attention is all you need

Limitations of RNNs



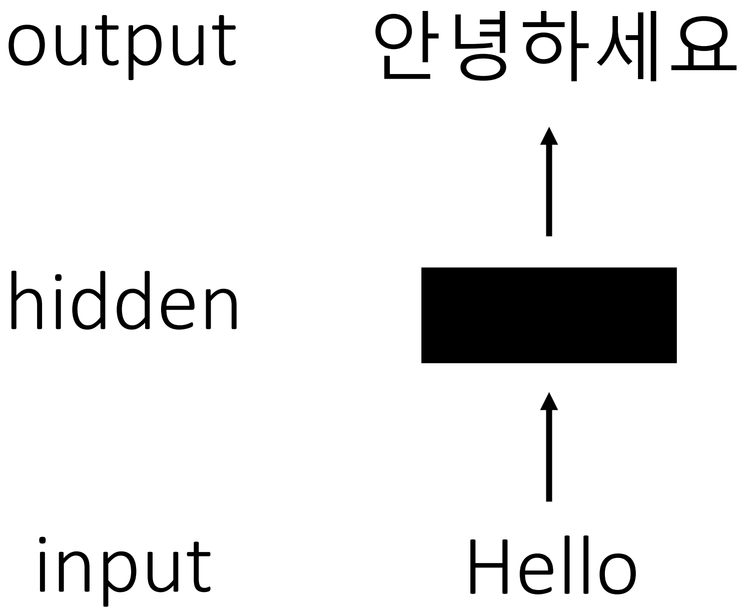
Vanishing and Exploding Gradients



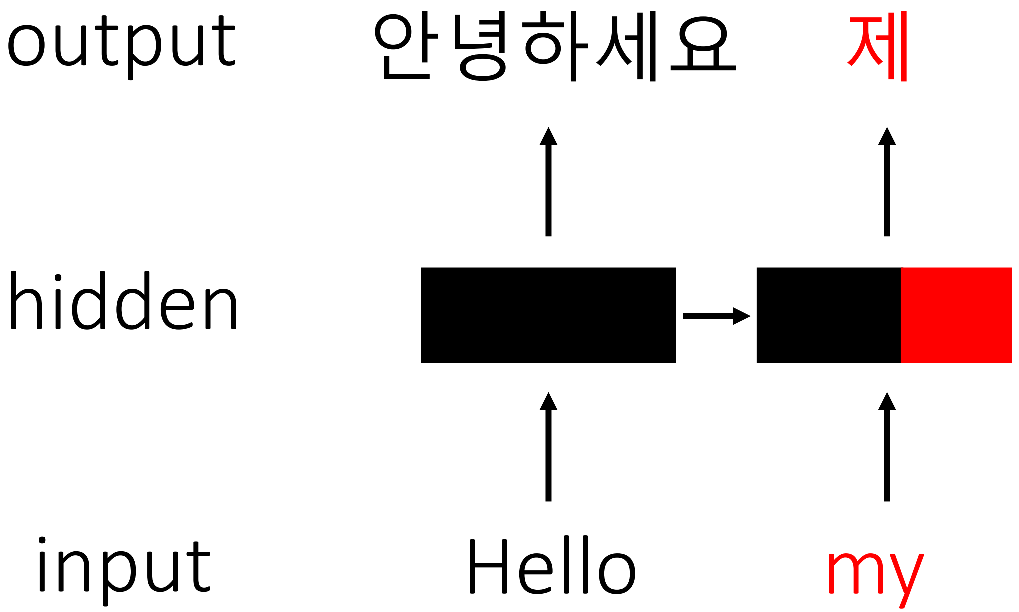
Longer input sequence →

higher risk of Vanishing/Exploding Gradients!

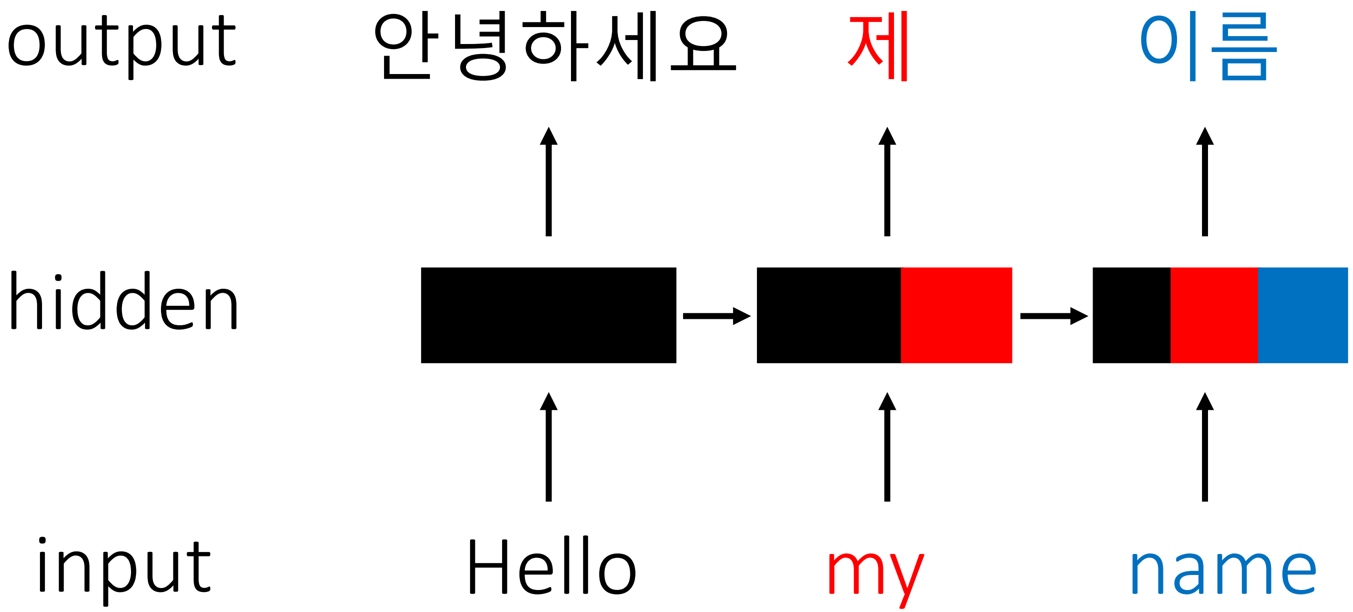
RNN Architecture



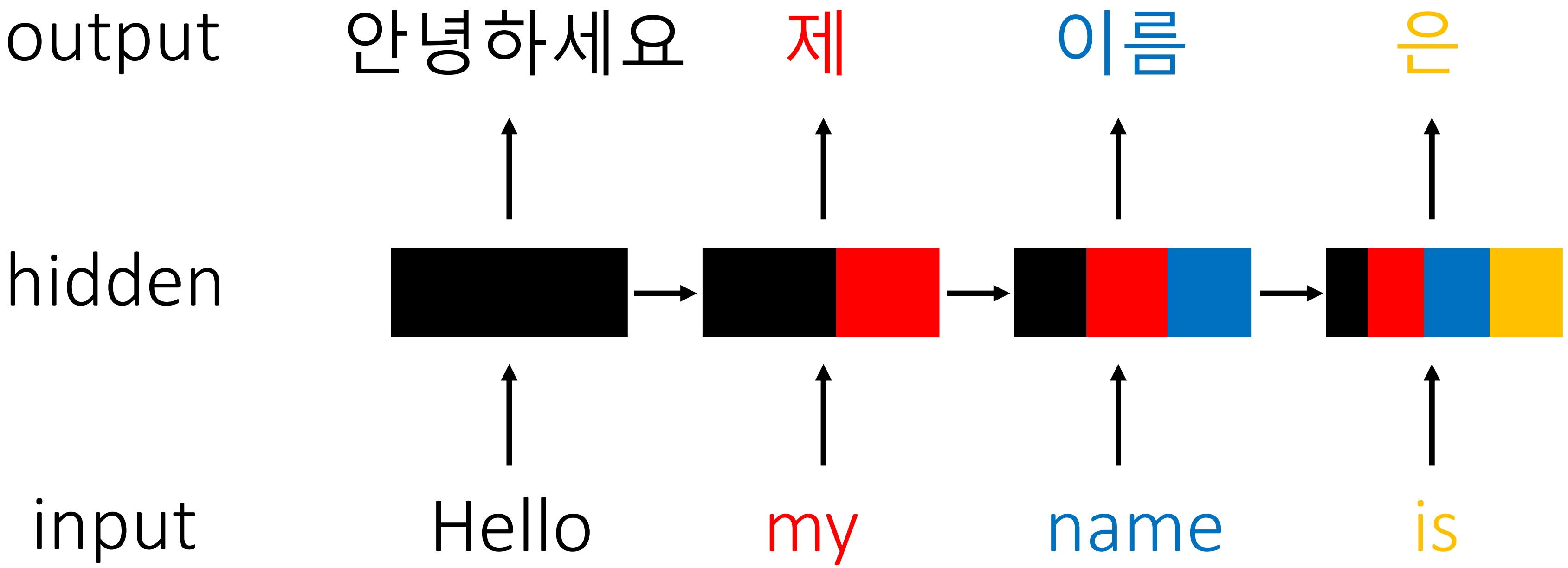
RNN Architecture



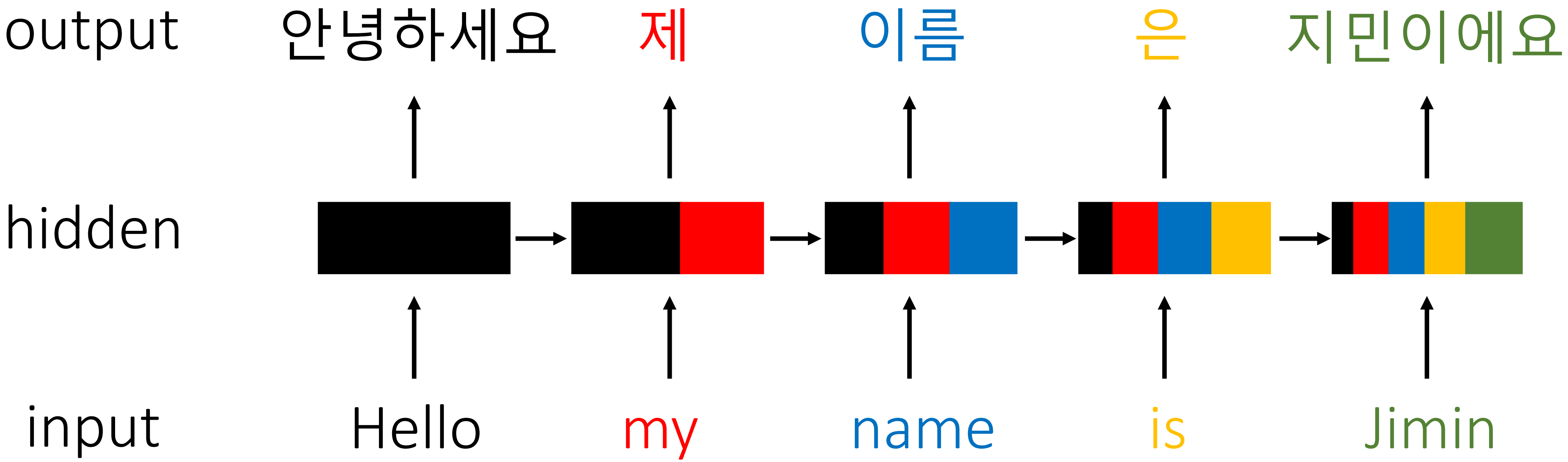
RNN Architecture



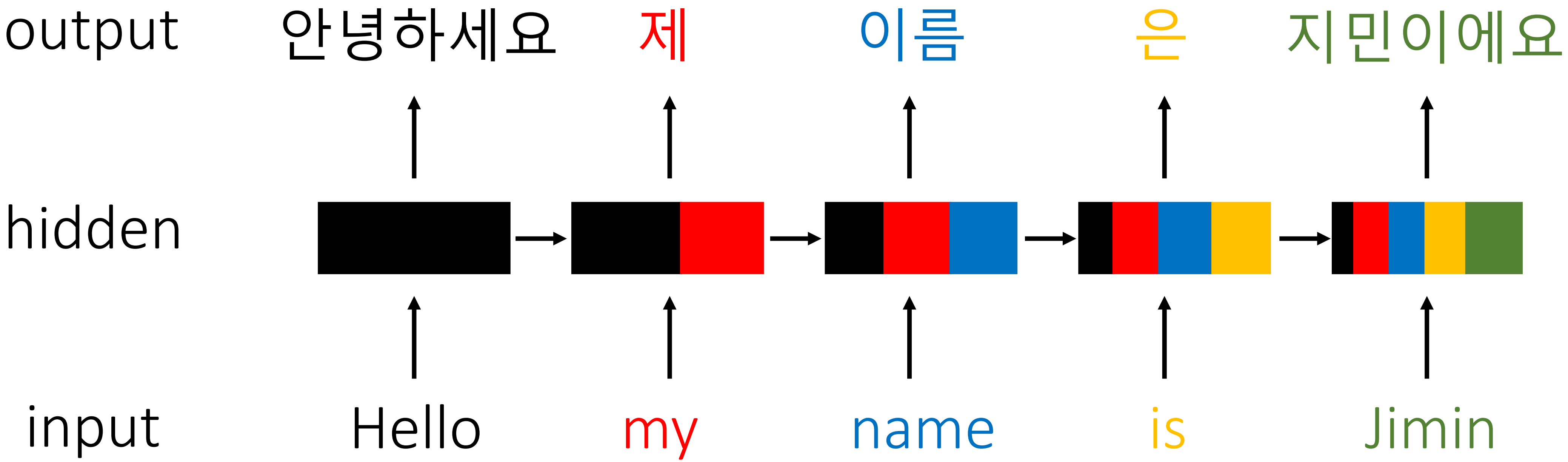
RNN Architecture



RNN Architecture

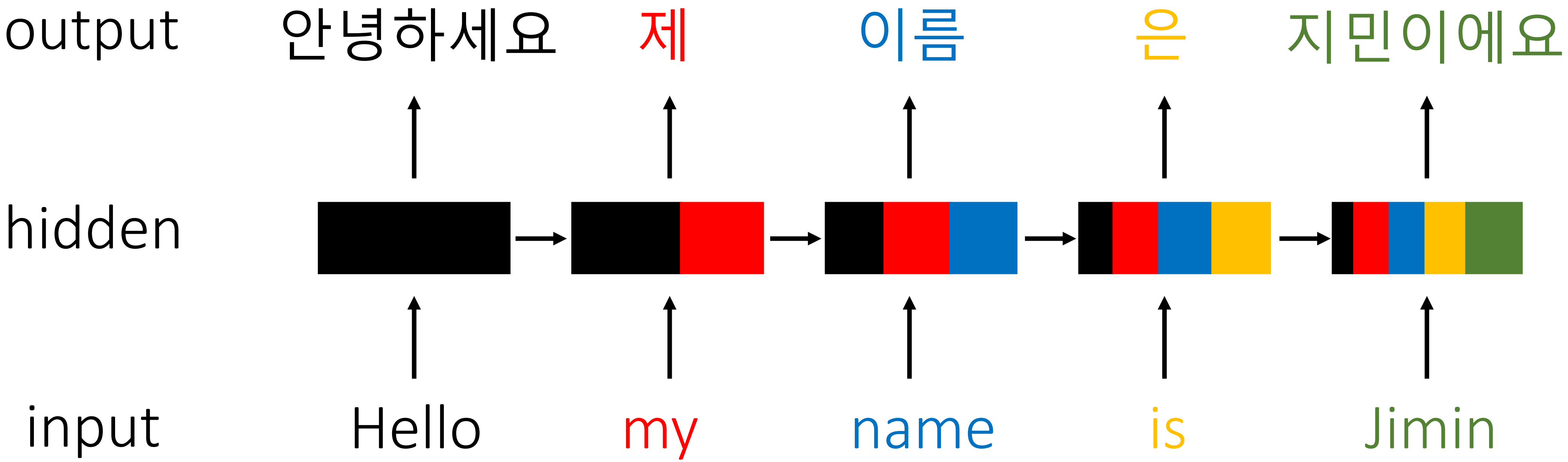


RNN Architecture



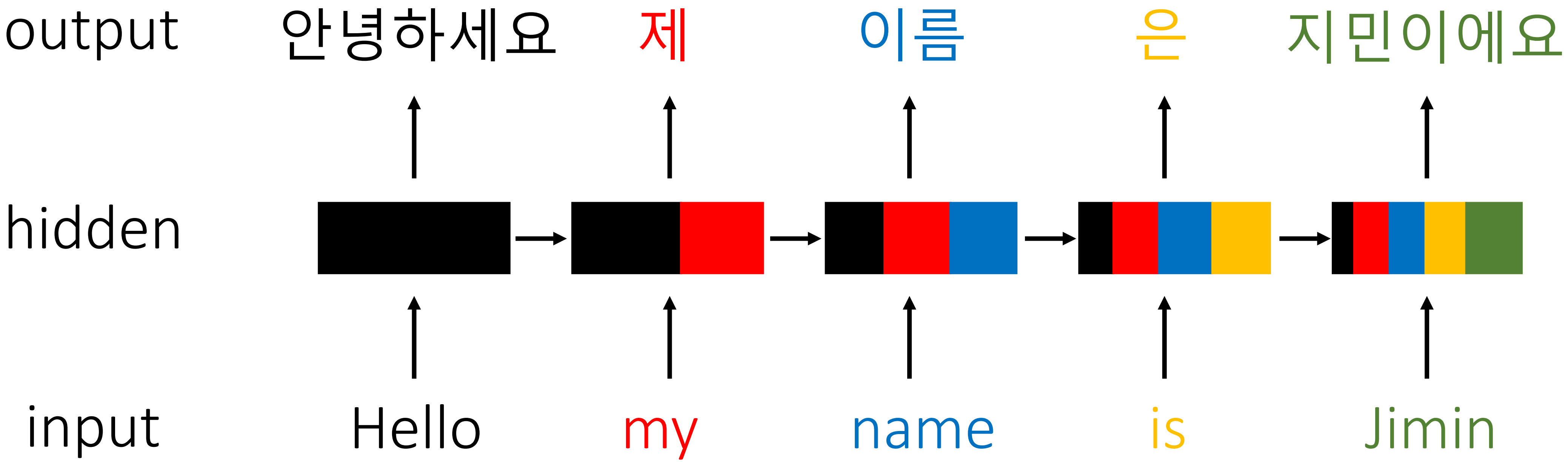
Each input (token) is fed sequentially →
No parallelization

RNN Architecture



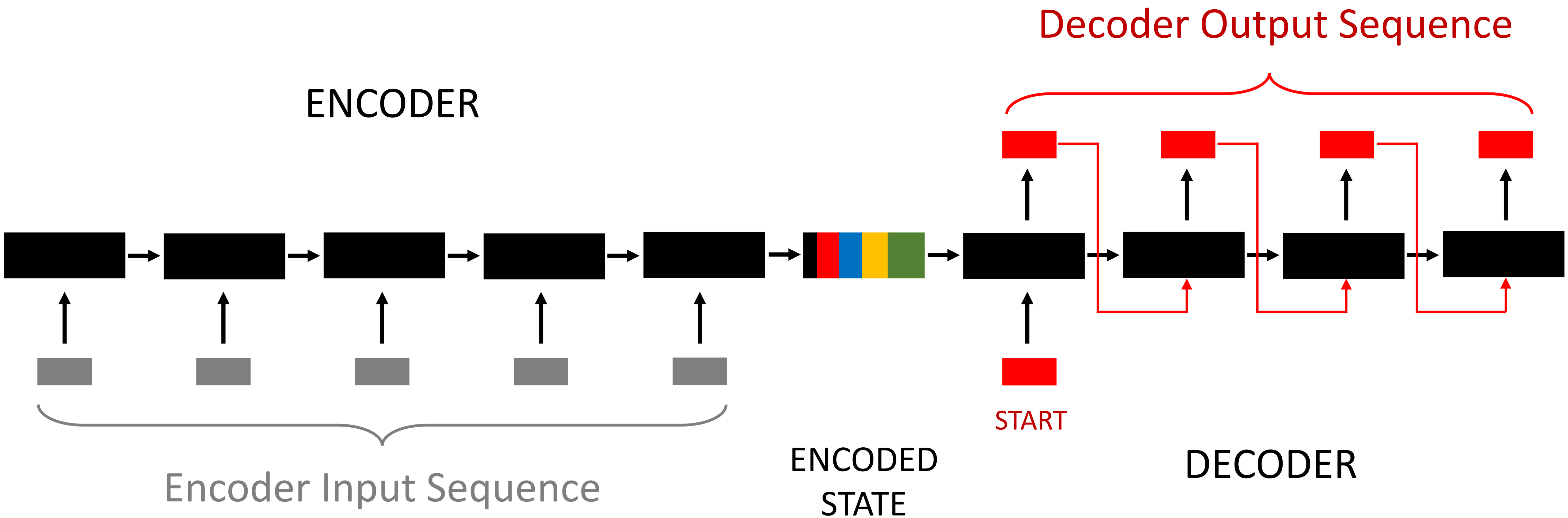
Difficult to store long-term context when sequence is long

RNN Architecture



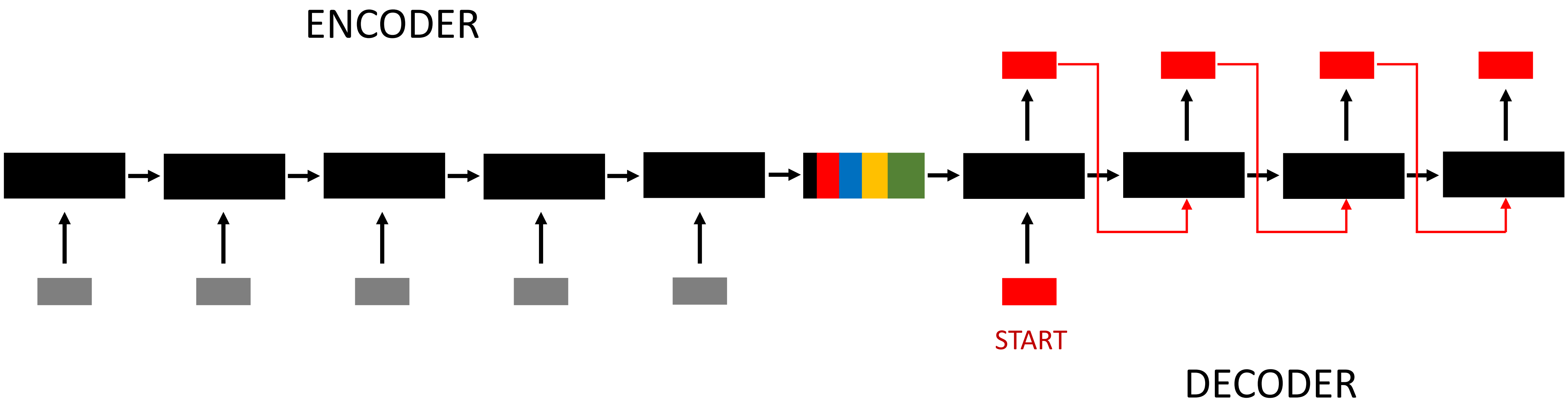
If using time-synced many-to-many →
$len(input\ seq) == len(output\ seq)$

Seq2Seq



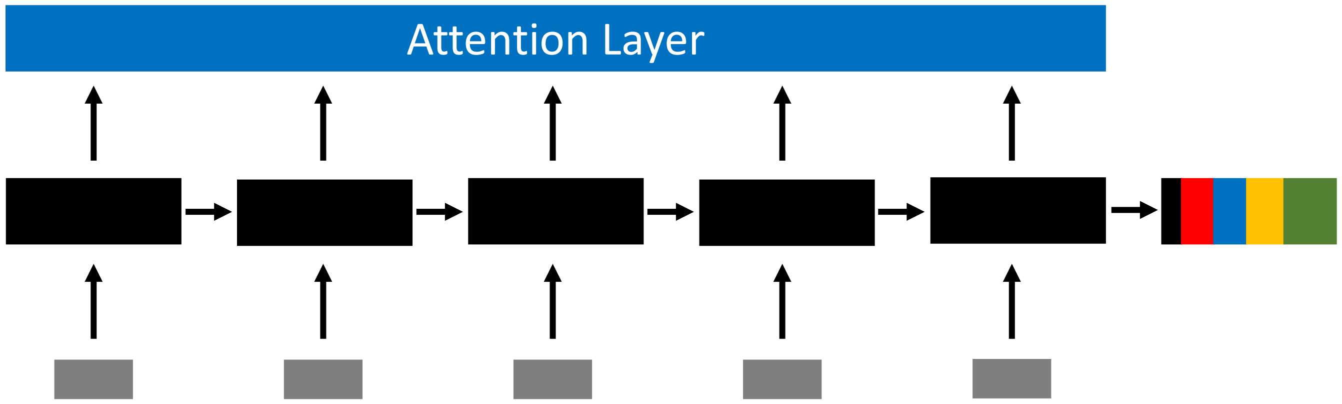
(+) Can be trained to translate input sequence to output sequence with two different lengths

Seq2Seq

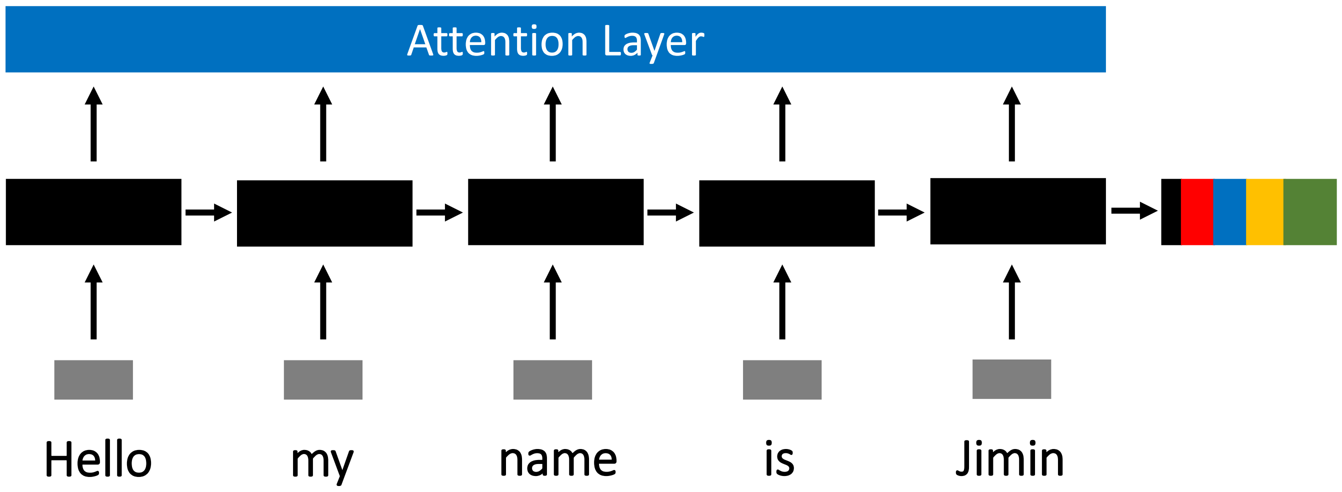


(-) Suffers from identical limitations as RNNs → Can't process long context, Hard to parallelize

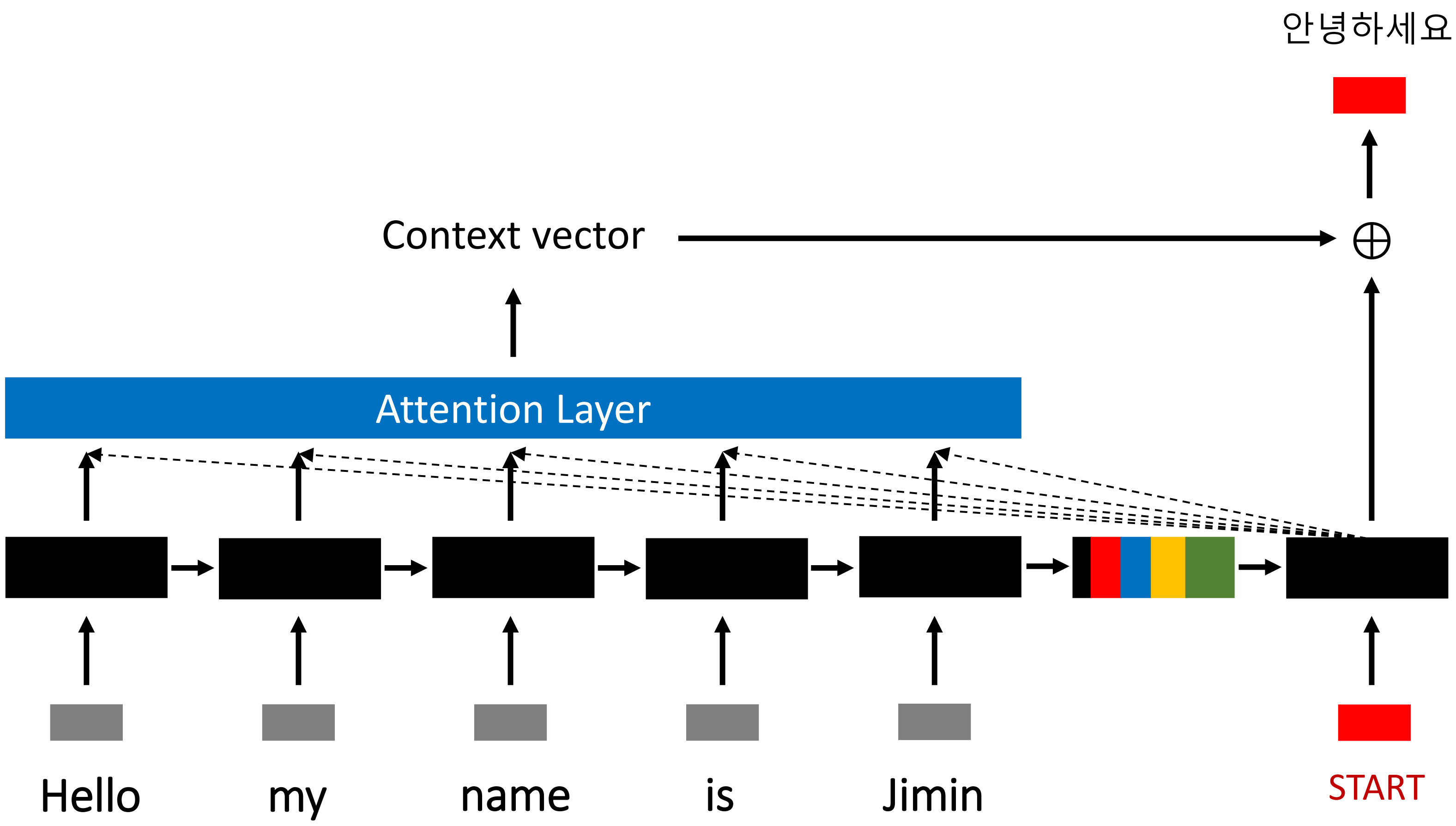
Seq2Seq with Attention



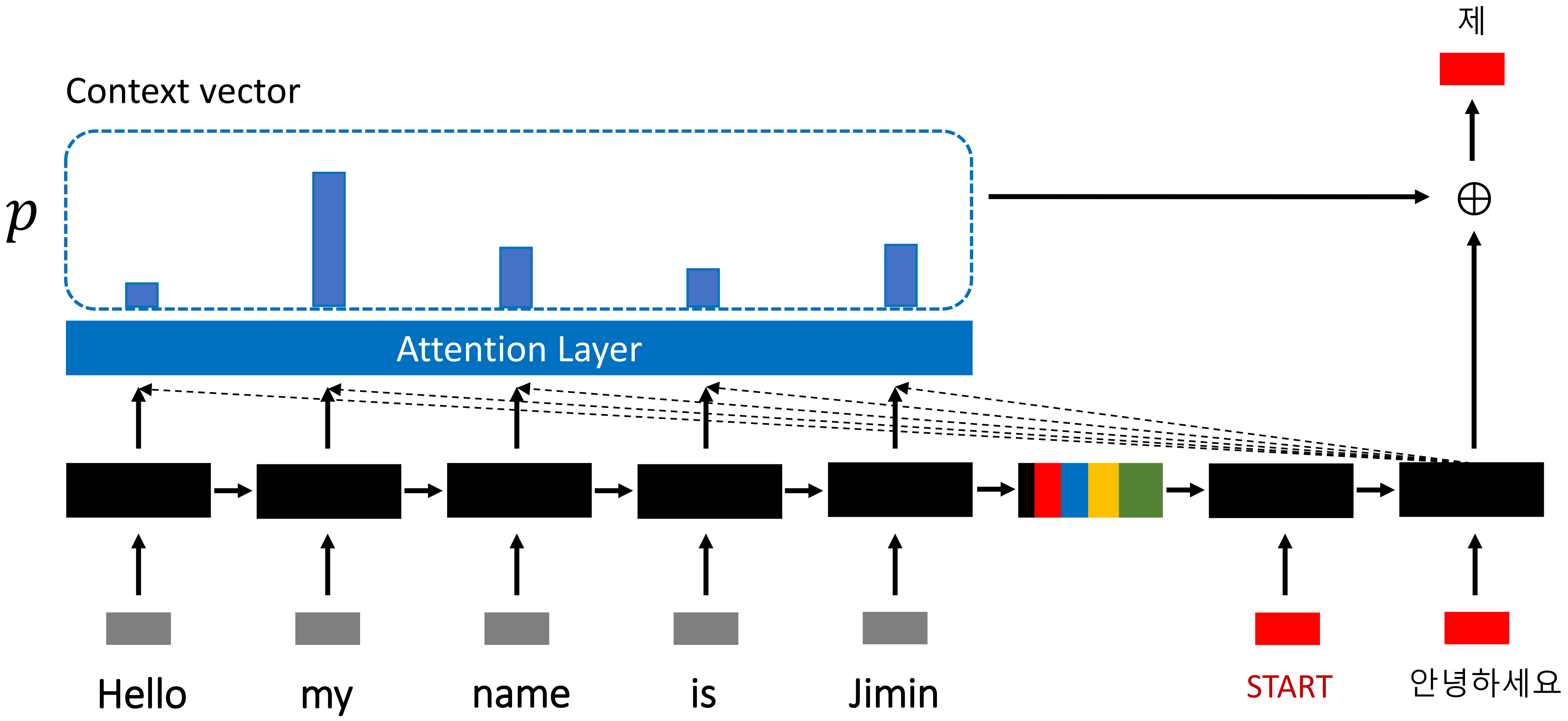
Seq2Seq with Attention



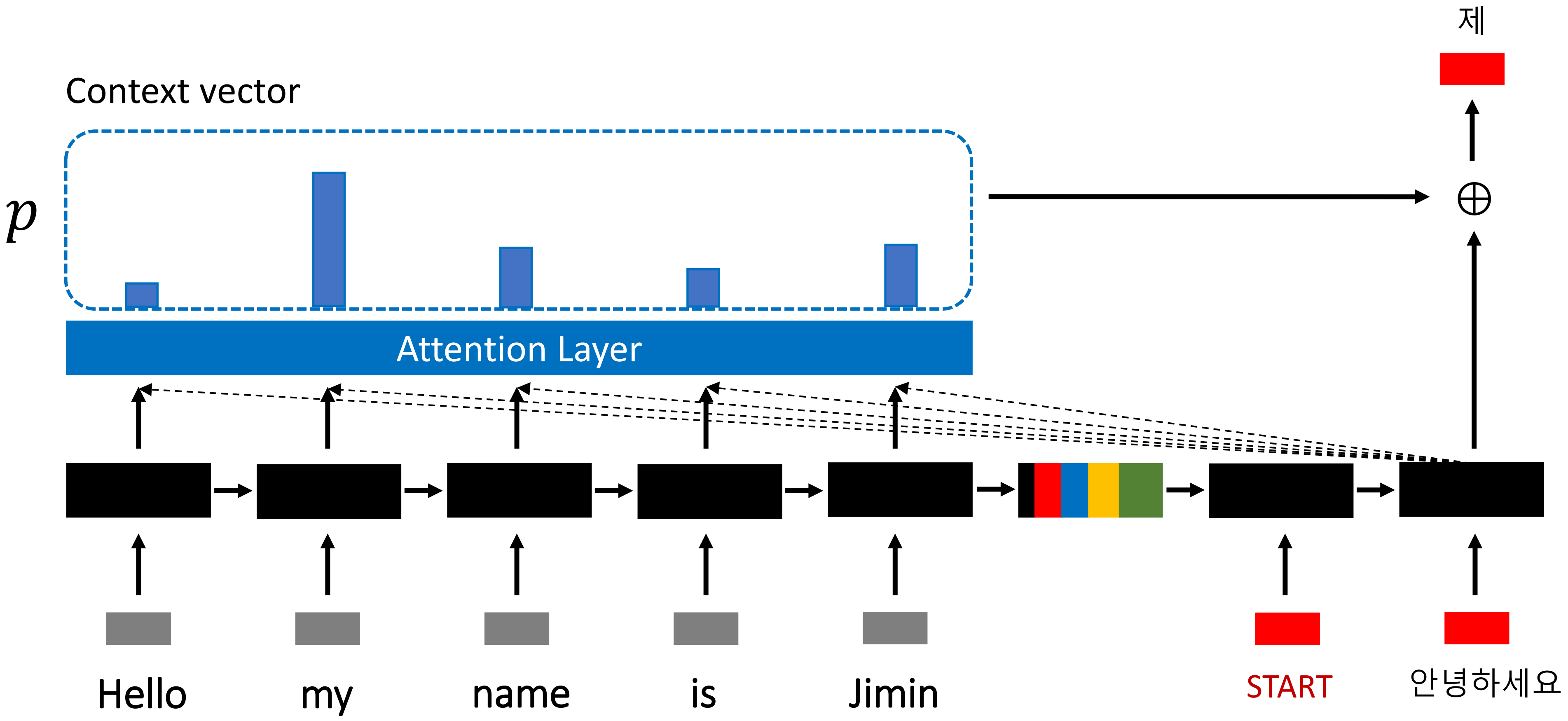
Seq2Seq with Attention



Seq2Seq with Attention

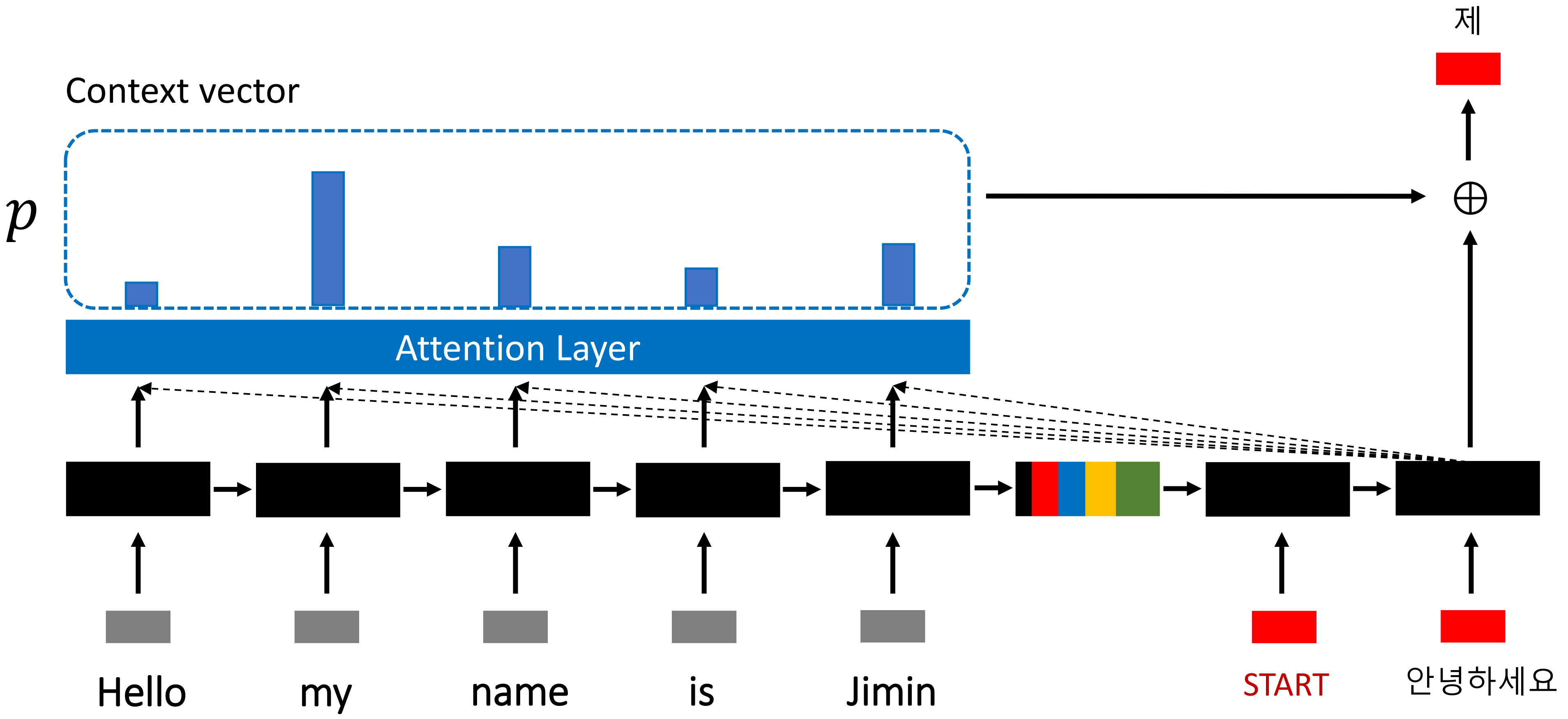


Seq2Seq with Attention



(+) Addresses long context issue

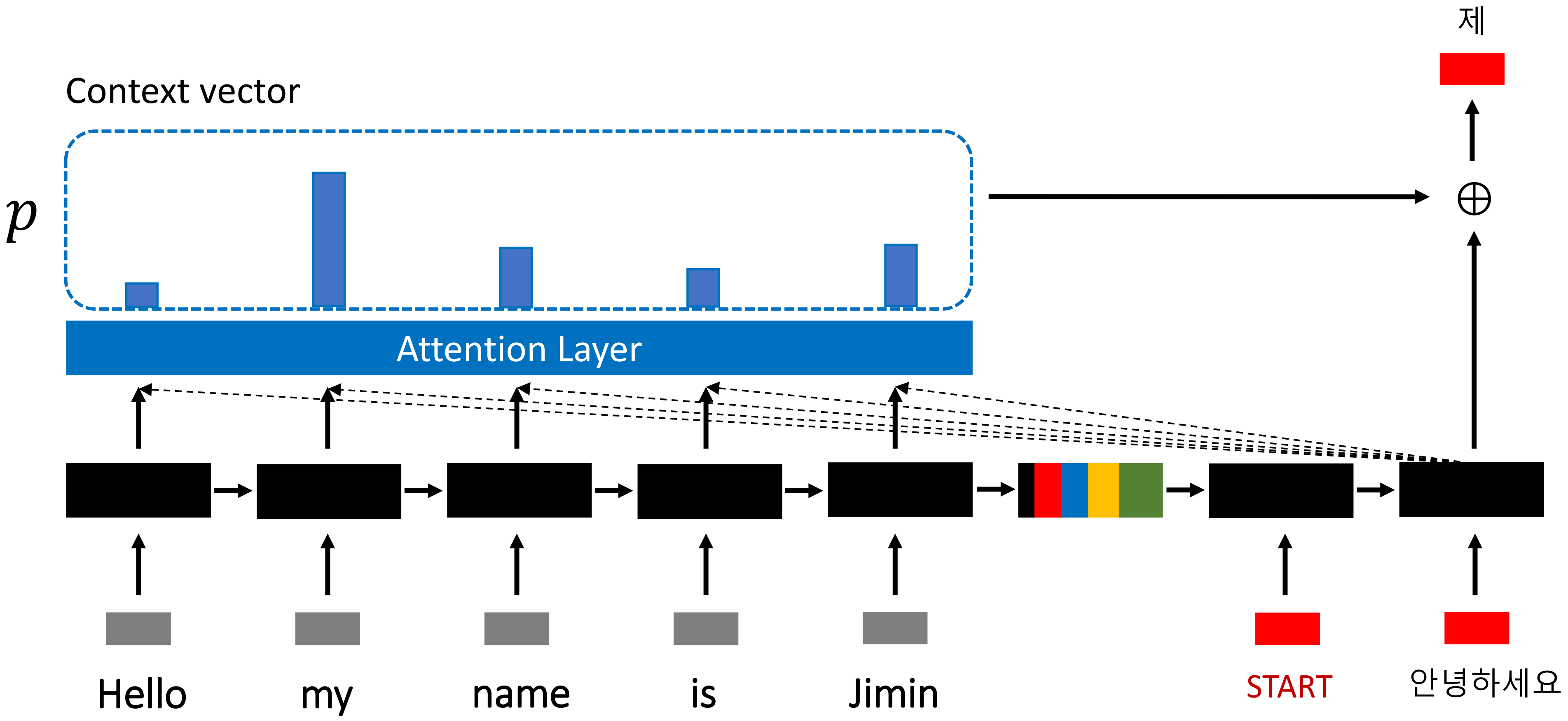
Seq2Seq with Attention



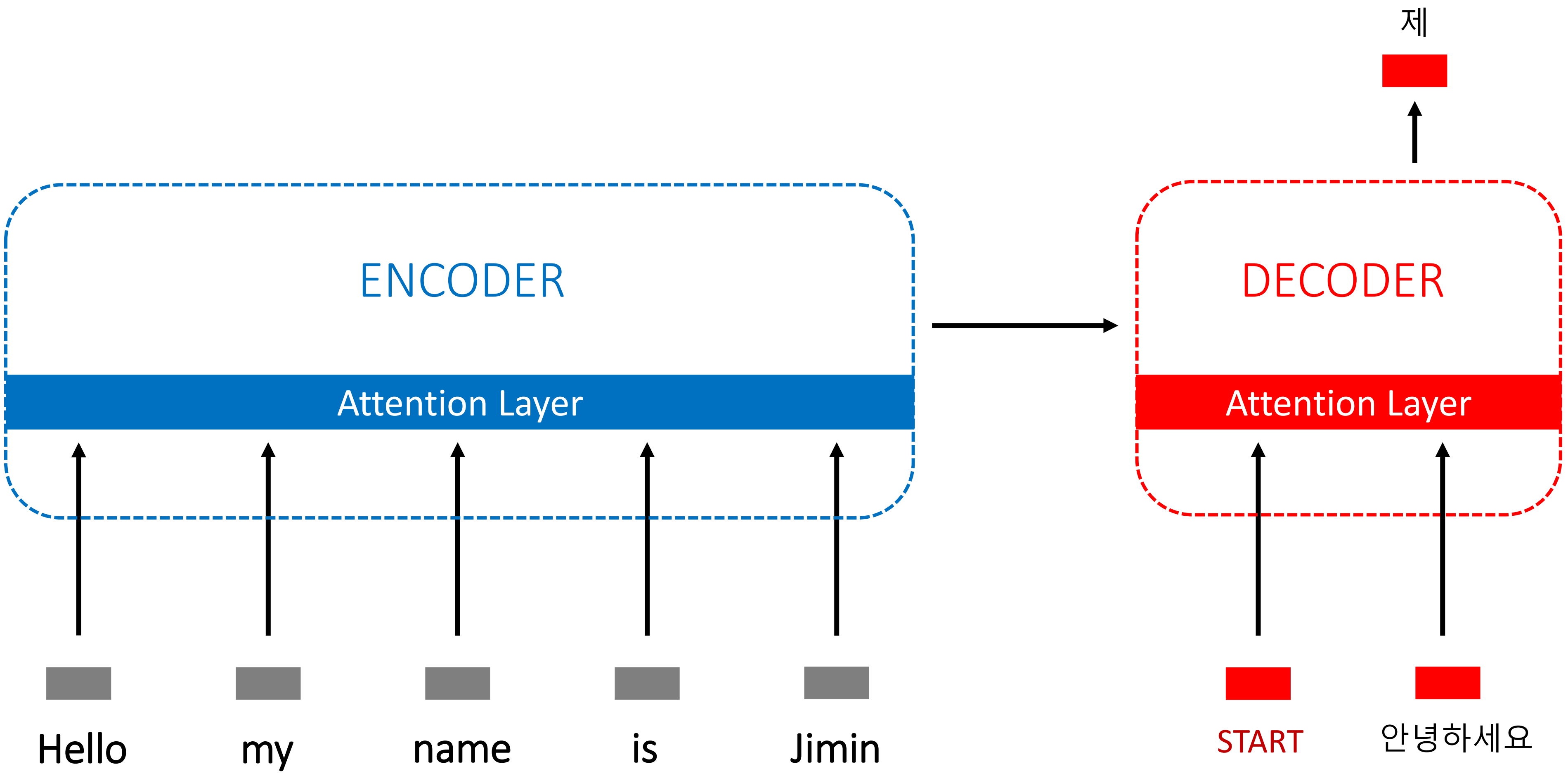
(+) Addresses long context issue

(-) Difficult to parallelize

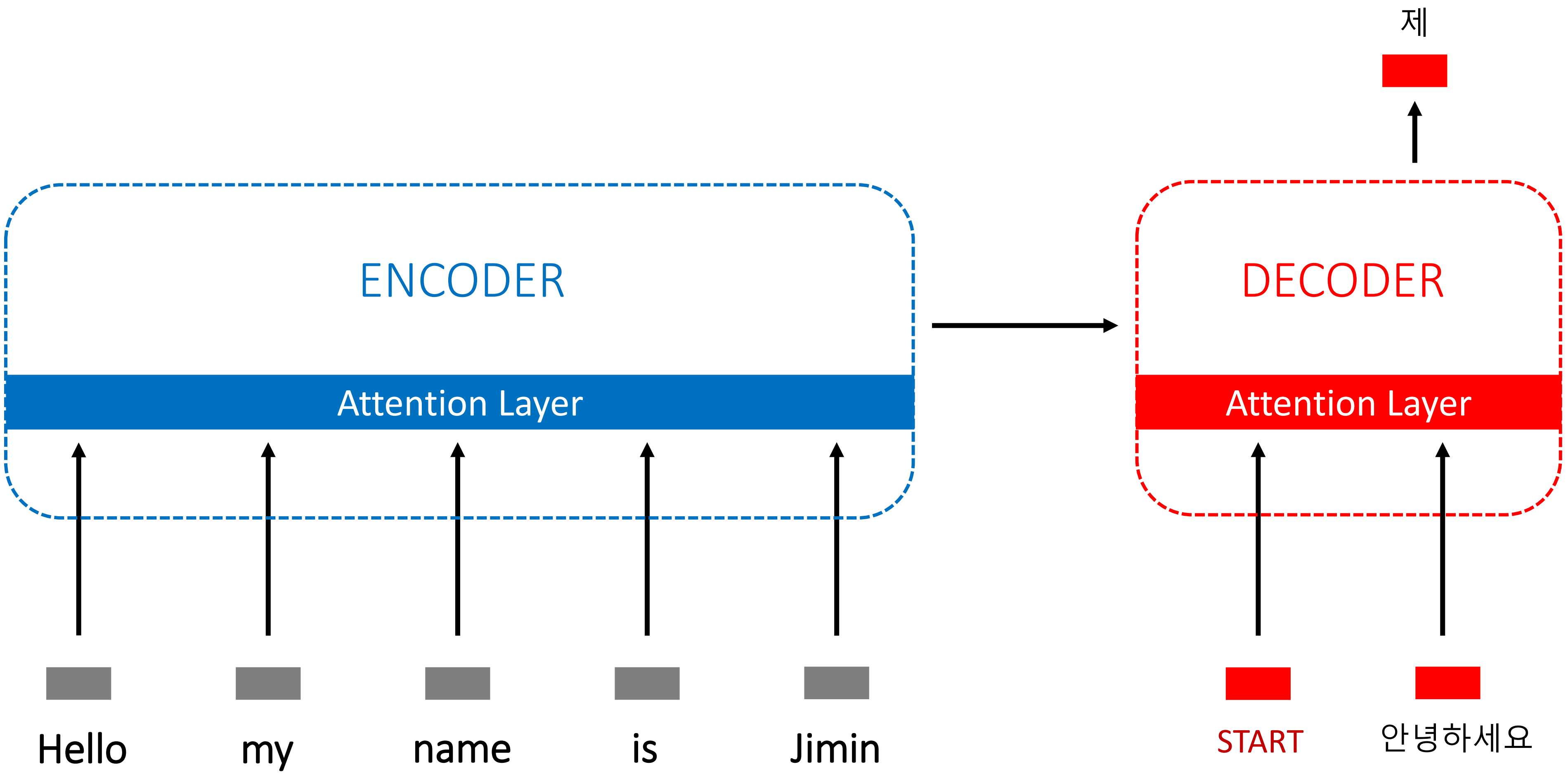
Attention is all you need (2017)



Attention is all you need (2017)



Attention is all you need (2017)



Attention without RNN is sufficient
Can utilize parallelization with GPUs

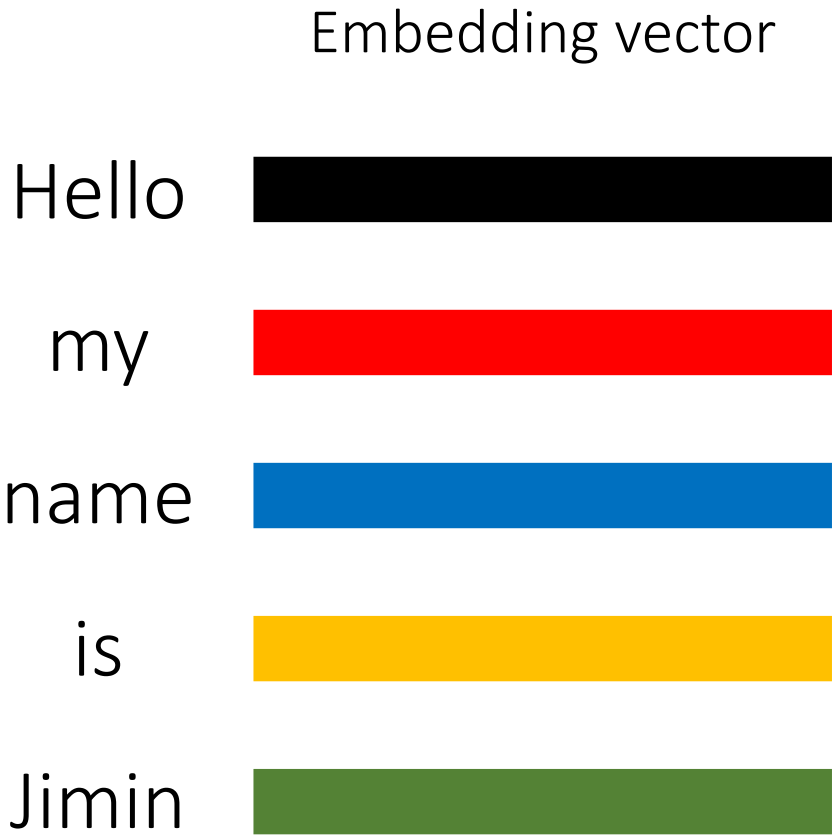
Self-attention layer

Overview

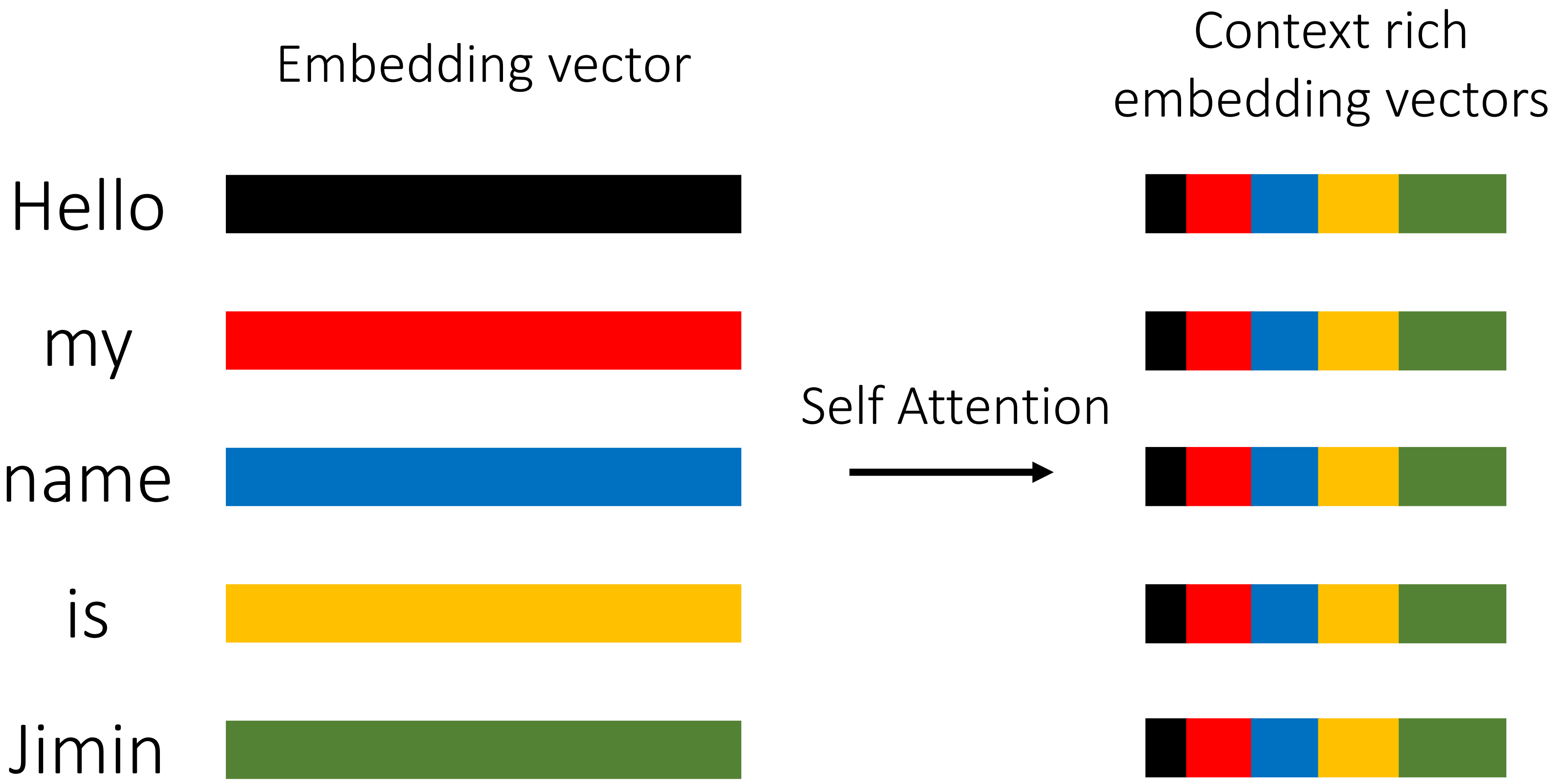
Key, Query, Value retrieval process

Multi-headed attention

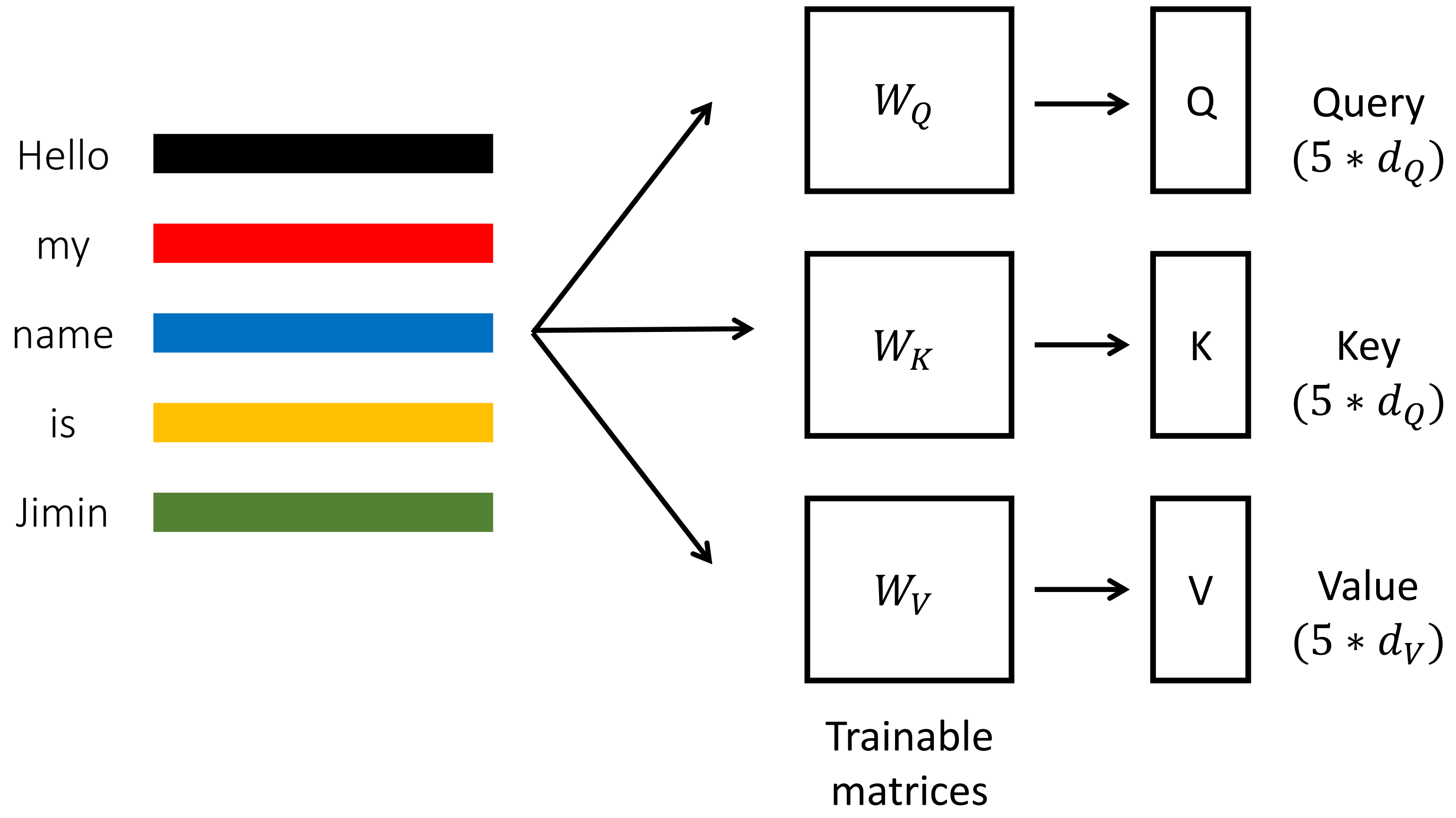
Overview of self-attention layer



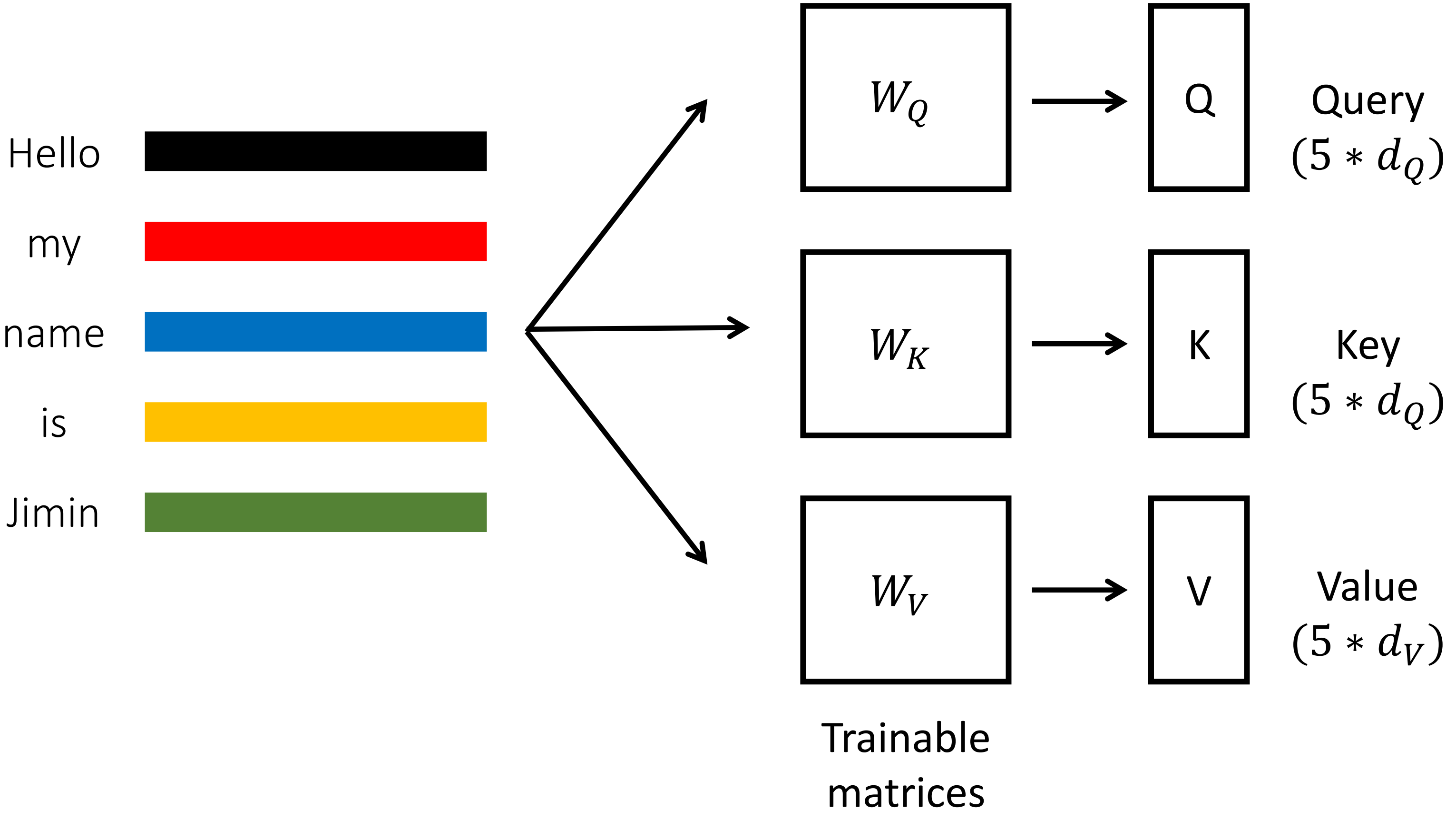
Overview of self-attention layer



Key, Query, Value retrieval

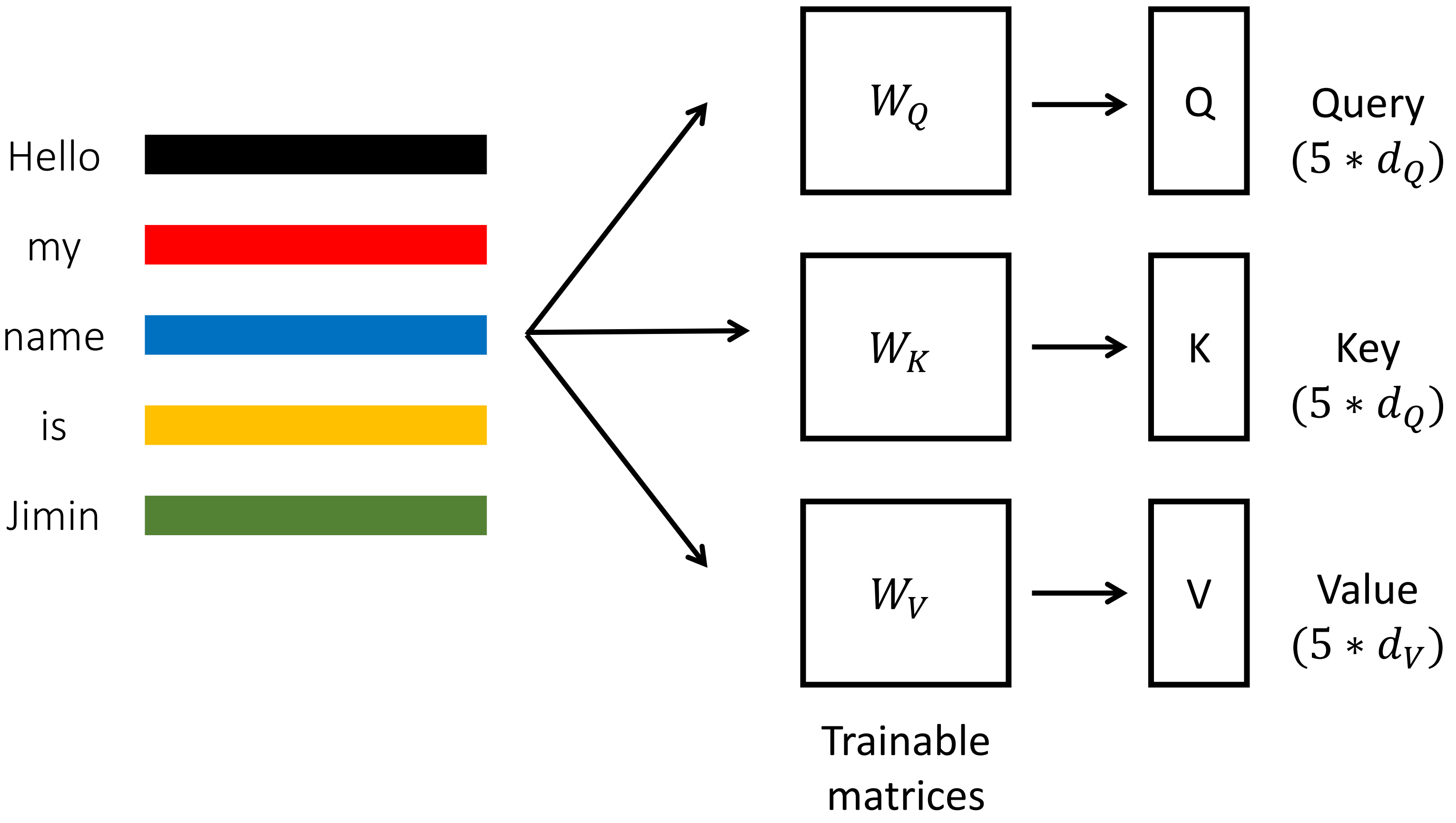


Key, Query, Value retrieval



≈ looking for a specific context (e.g., is prior word a subject?)

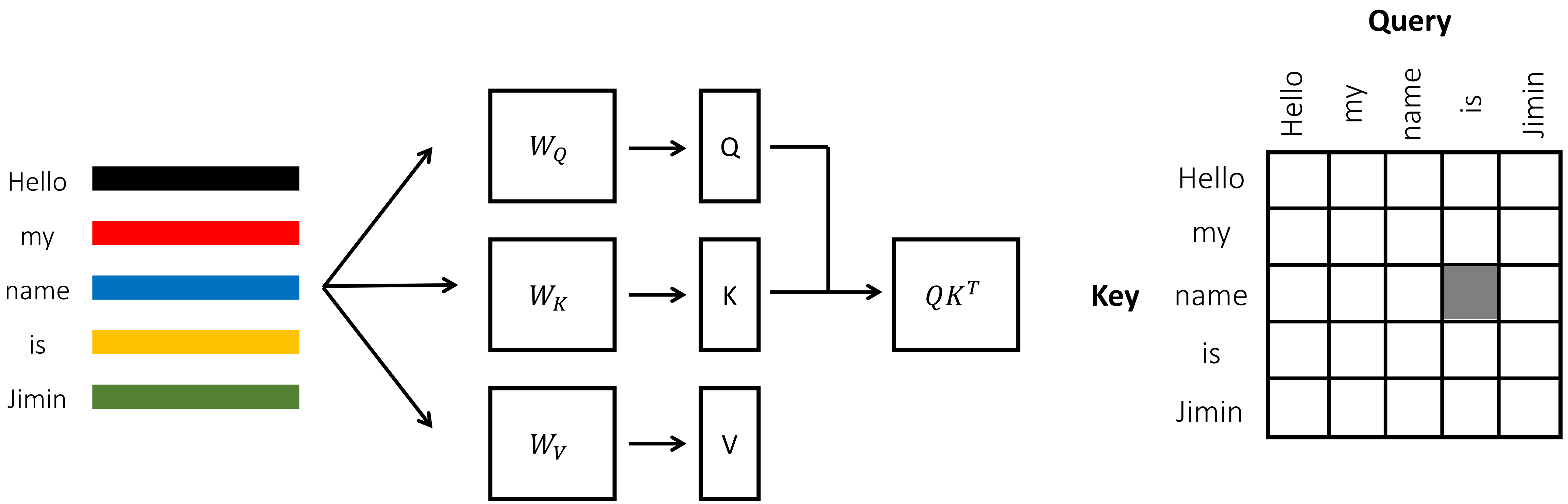
Key, Query, Value retrieval



≈ looking for a specific context
(e.g., is prior word a subject?)

≈ answering the query
(e.g., yes, I am the subject)

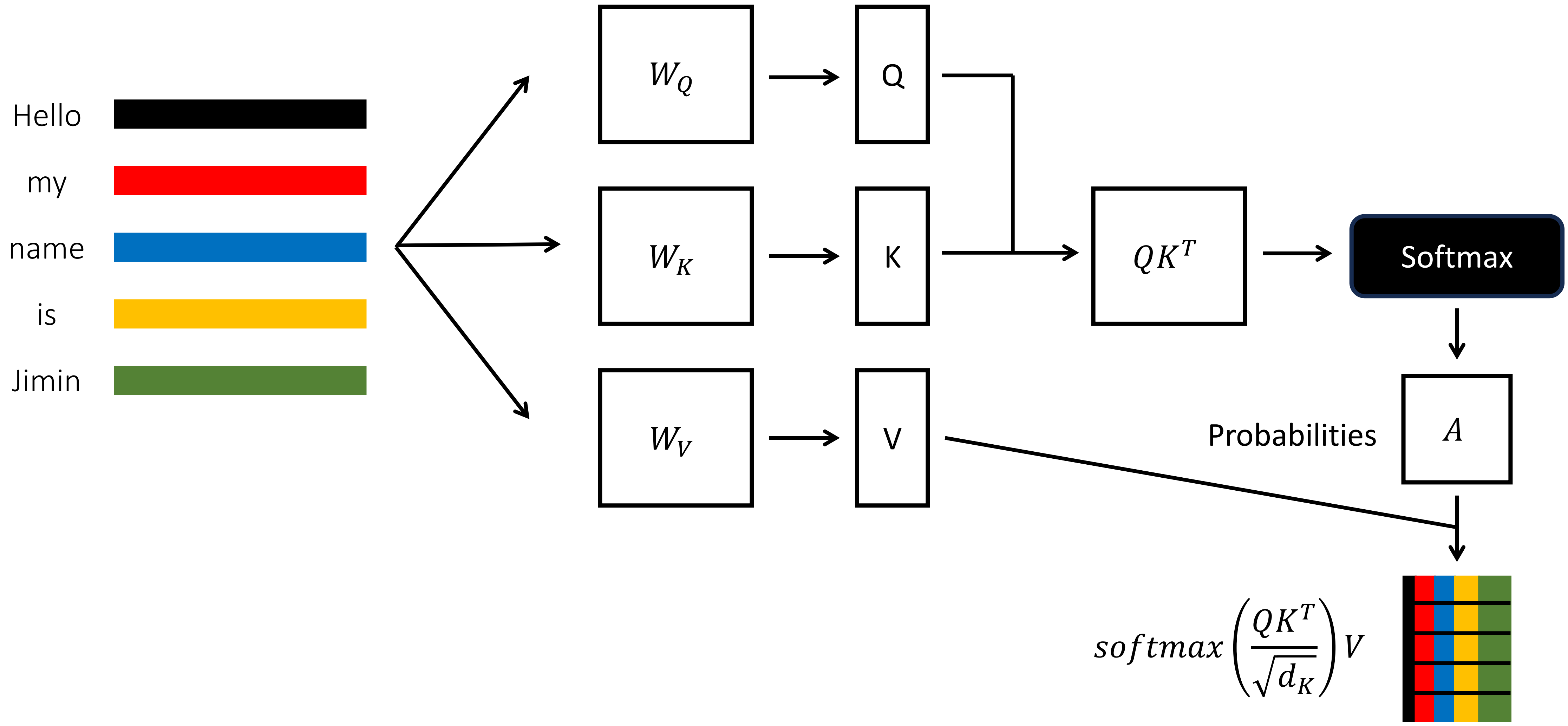
Key, Query, Value retrieval



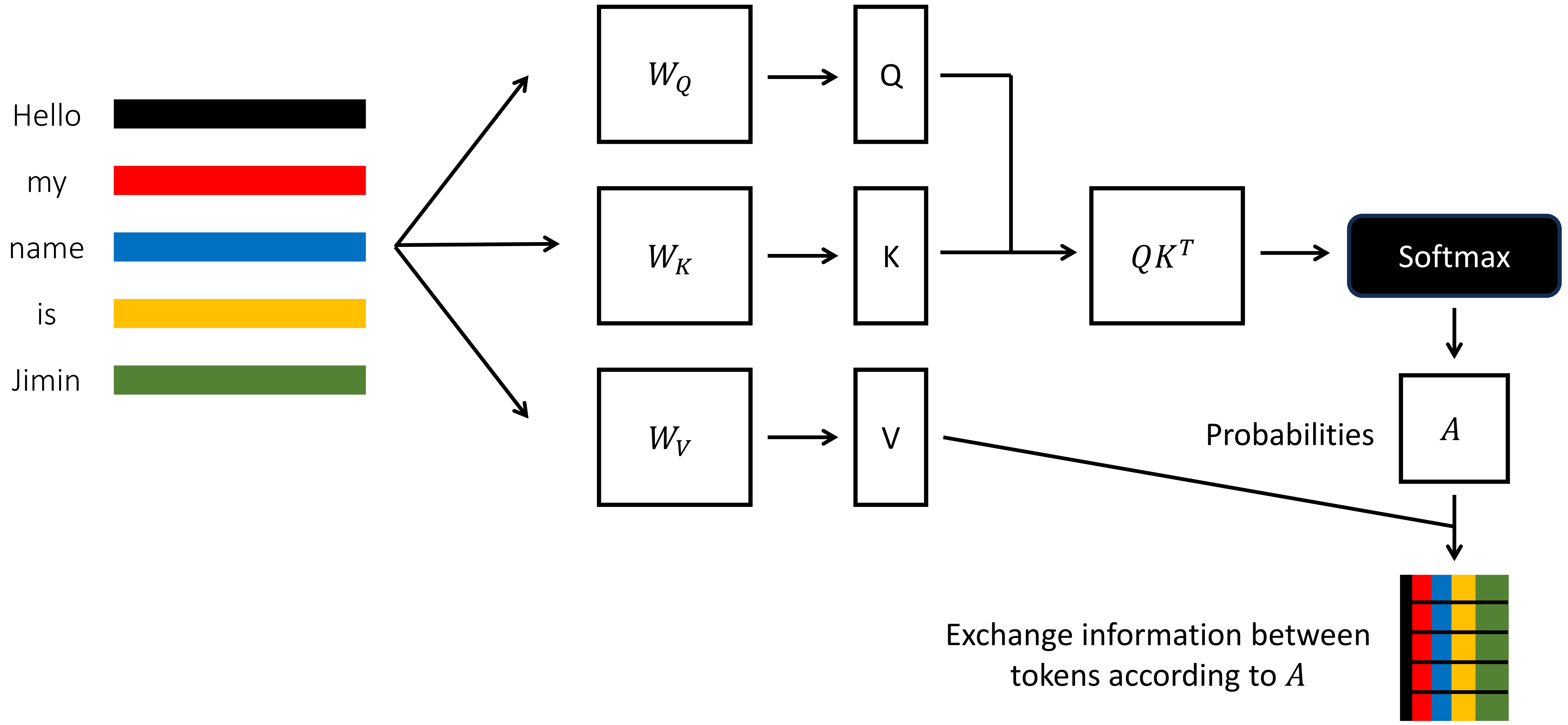
Query $\approx$ is prior word a subject?

Key $\approx$ yes, I am the subject

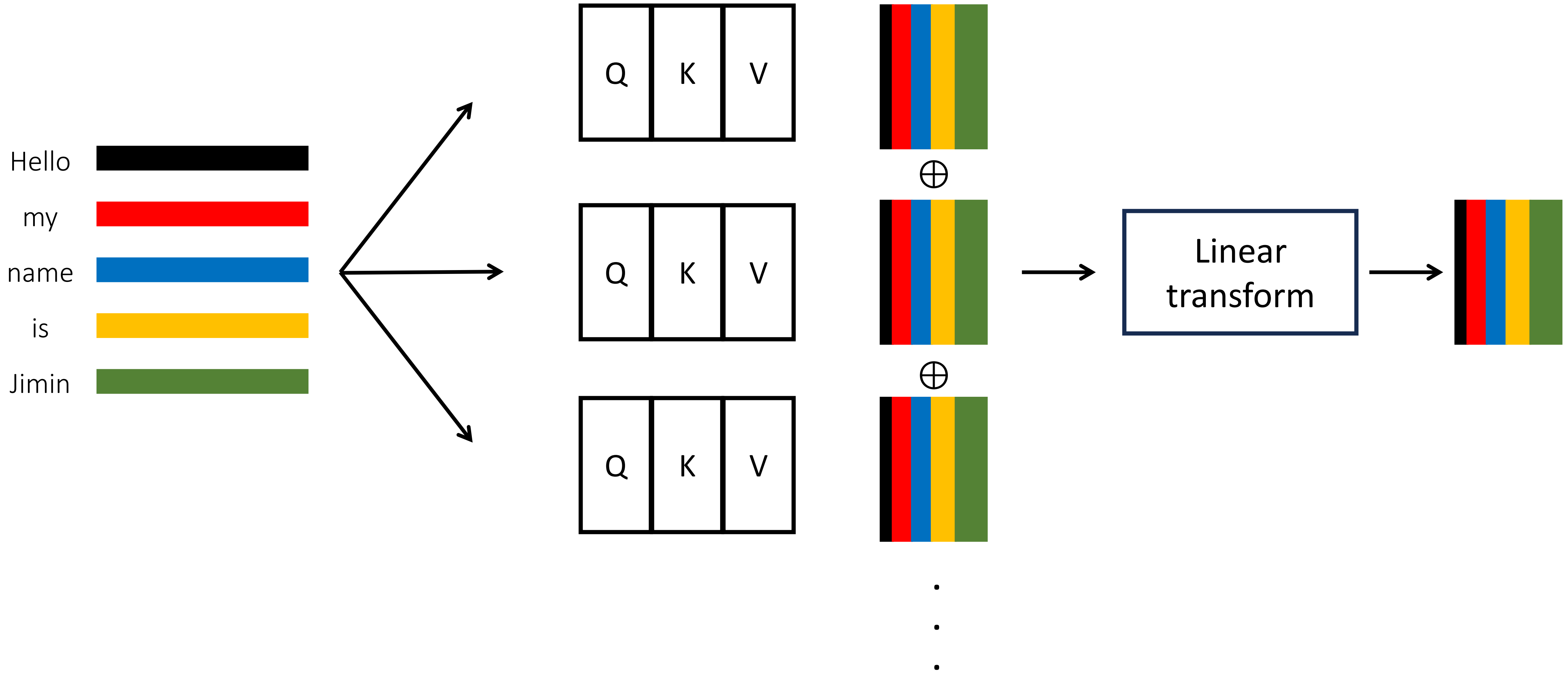
Key, Query, Value retrieval



Key, Query, Value retrieval



Multi-headed attention



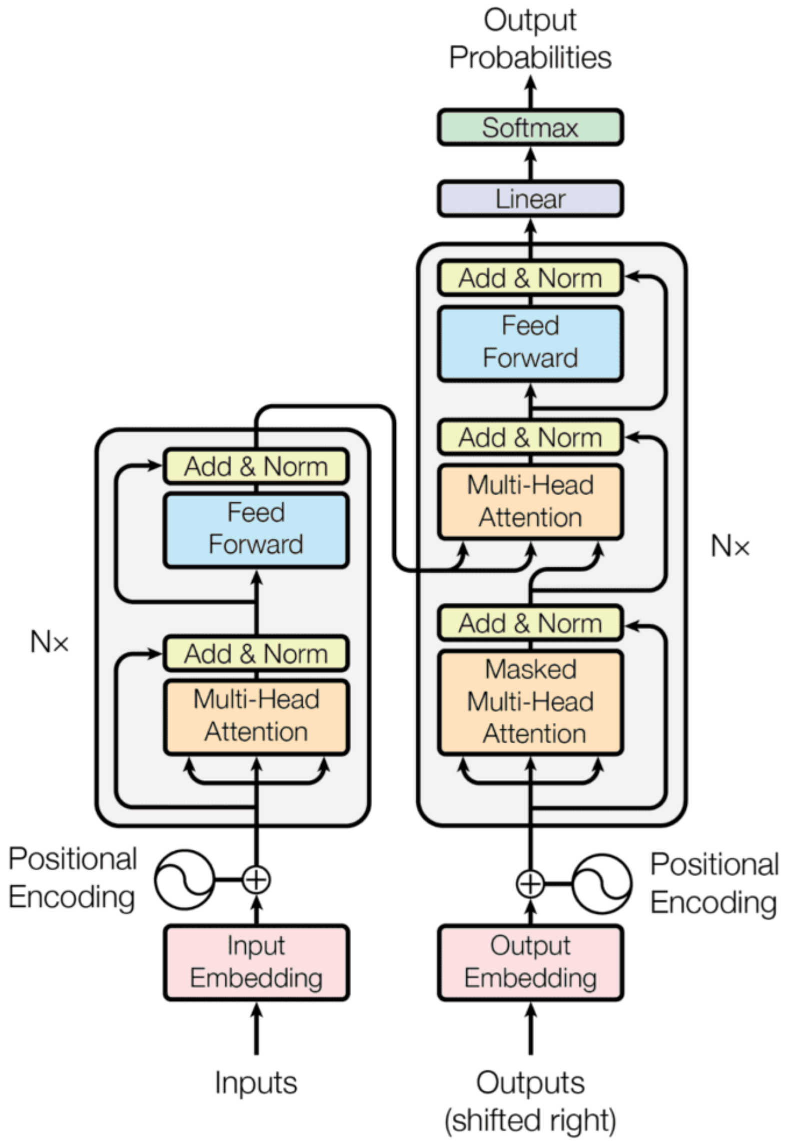
Transformer Architecture

Encoder

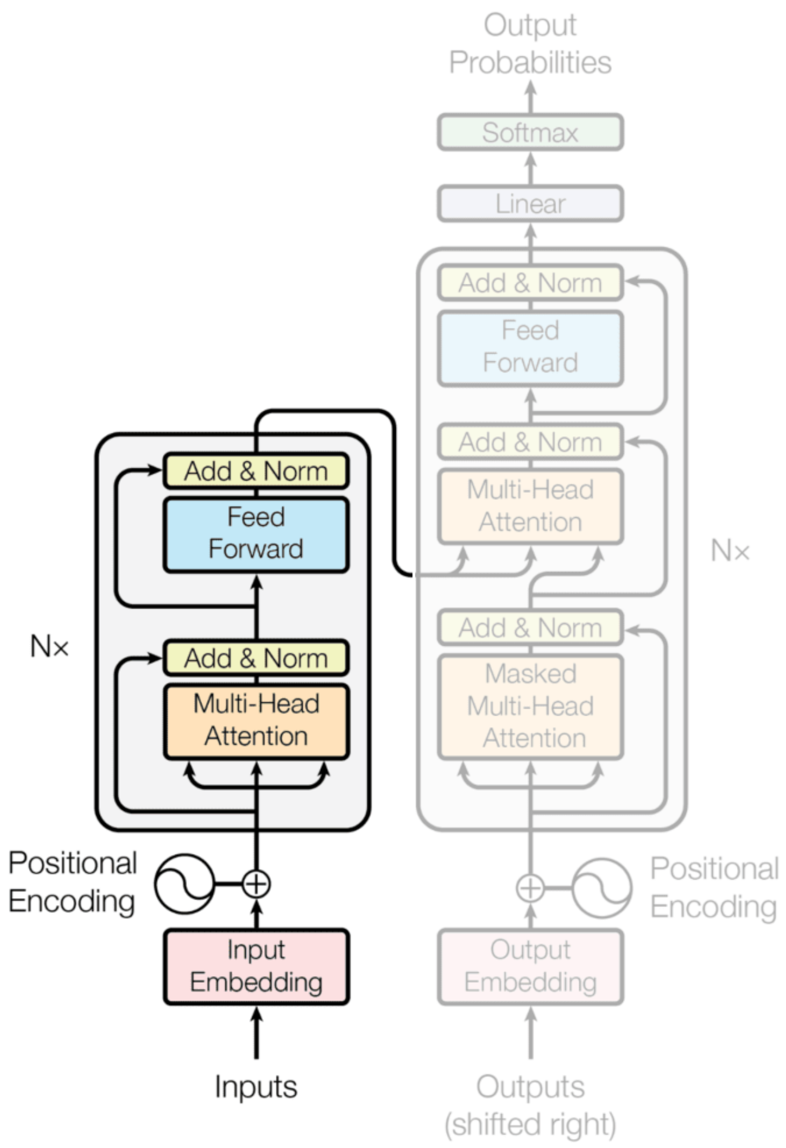
Decoder

Transformer vs RNN

Transformer Architecture



Transformer Architecture



Encoder layer with

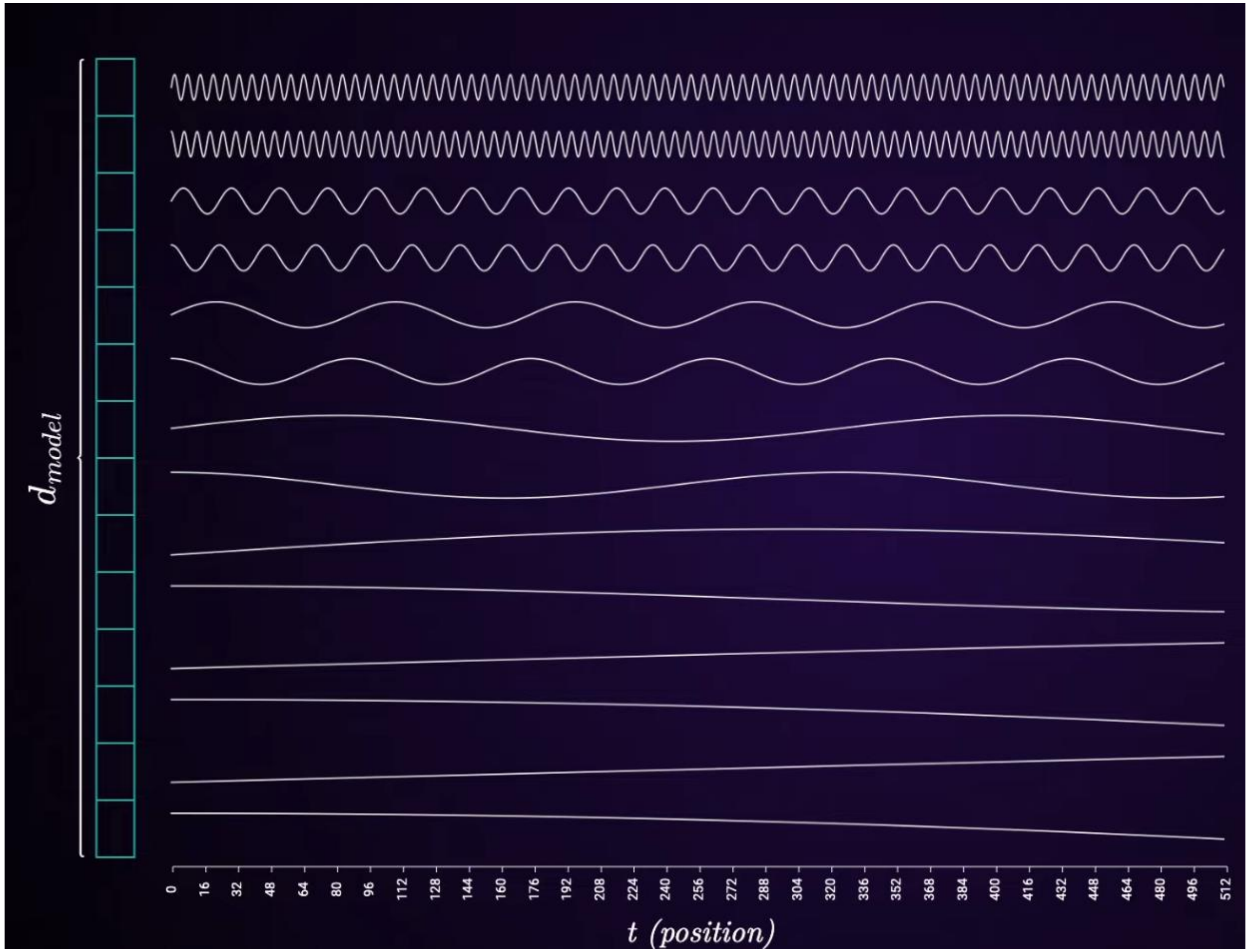
- Input embedding with positional encoding
- multi-headed self attention
- Residual connections, Layer norm & dropout

Positional Encoding

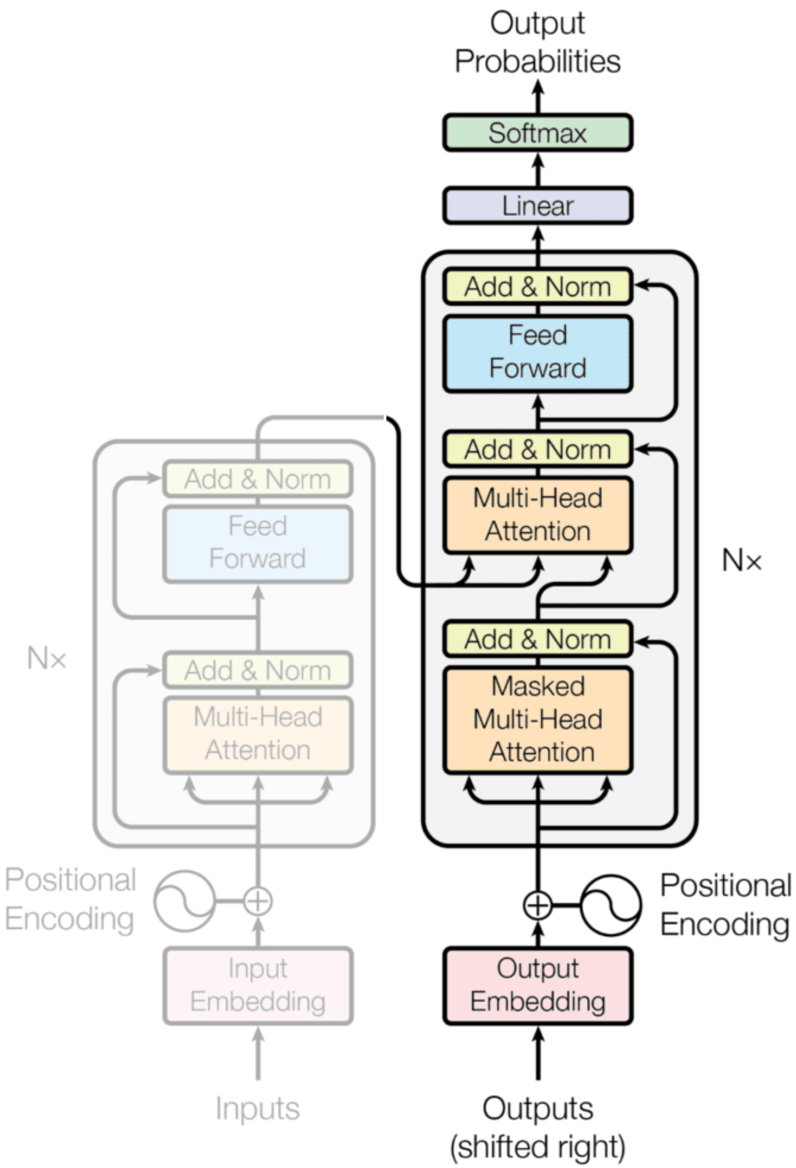
Hello my name is Jimin

$$k\ is\ even: \sin\left(\frac{t}{10000^{\frac{k}{d_{embedding}}}}\right)$$

$$k\ is\ odd: \cos\left(\frac{t}{10000^{\frac{k}{d_{embedding}}}}\right)$$



Encoder



Decoder layer with

- Masked multi-headed self attention
- Multiheaded cross attention
  - Inputs → Key, Query
  - Outputs → Value

Transformer vs RNN

	Transformers	RNNs
Sequential	No	No
Parallel computation	Yes	No
Long-term dependencies	Yes	Kind of
Scalability	Yes	Problematic
Fine tuning	Yes	Difficult

Transformer Applications

NLP

Computer Vision

Multi-modal

Audio and Speech

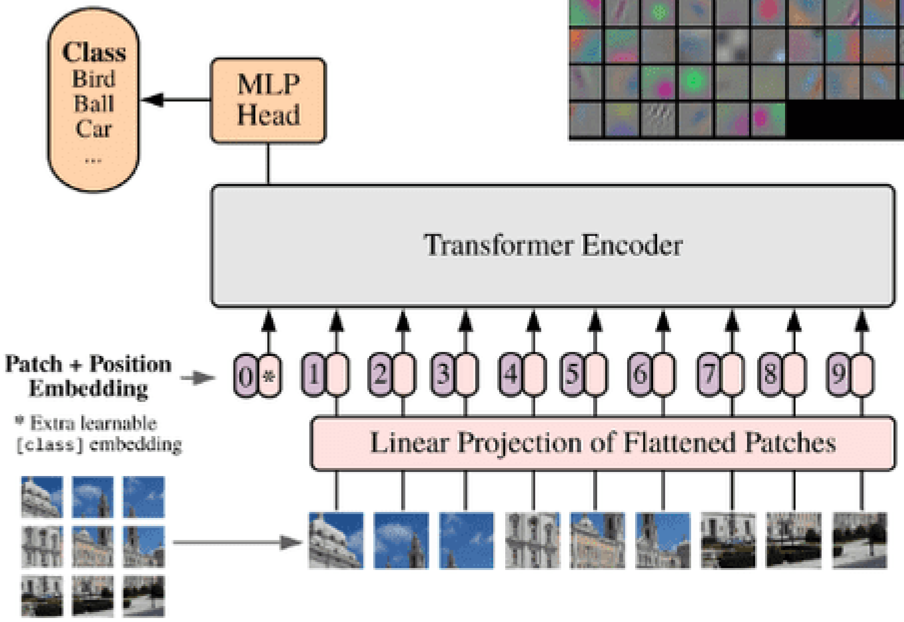
Signal processing

NLP

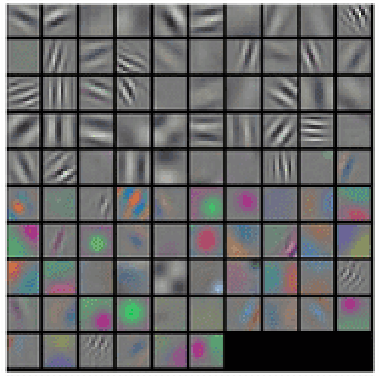
LLaMA

ChatGPT

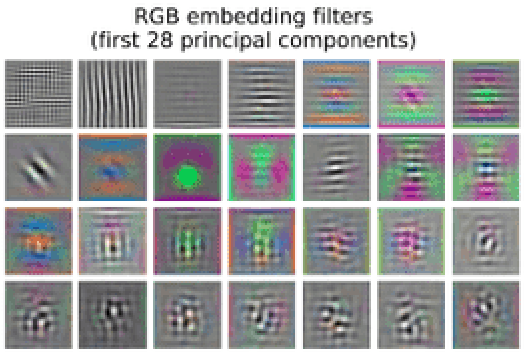
Computer Vision



Alexnet 1st conv filters



ViT 1st linear embedding filters



Multi-modal

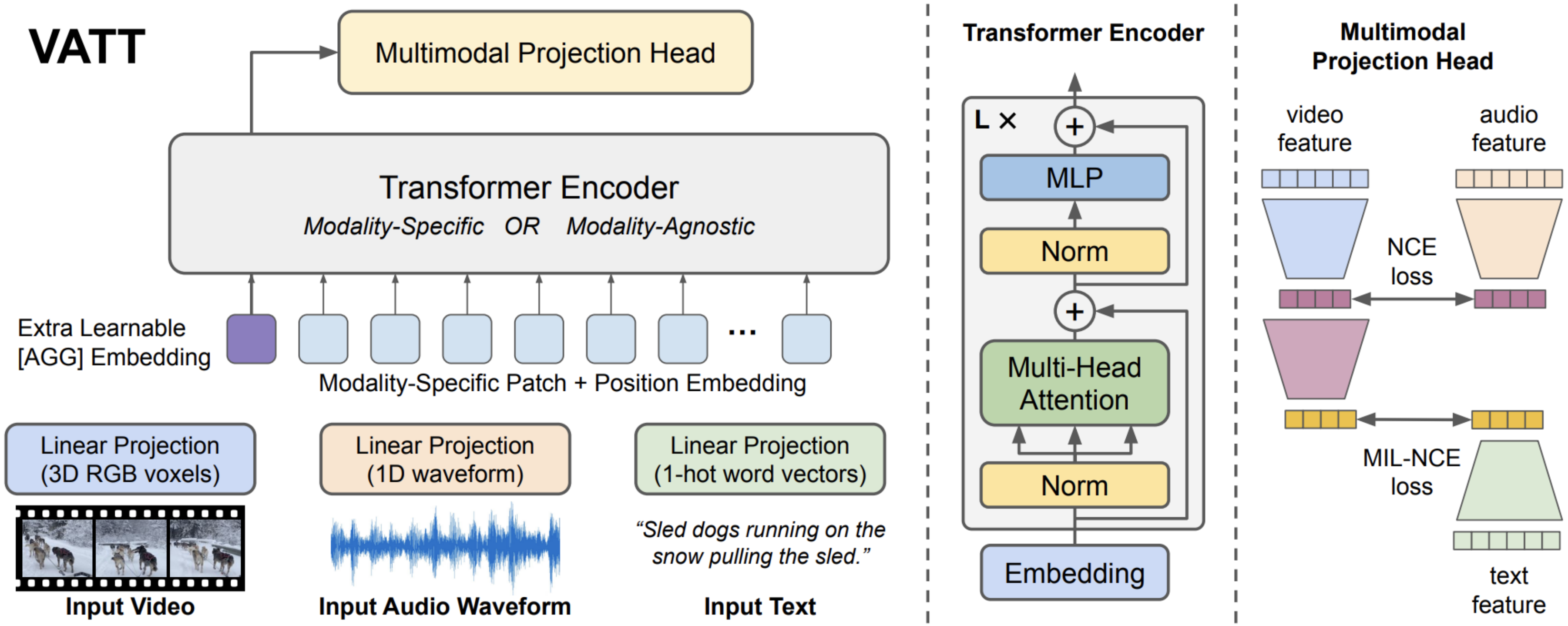
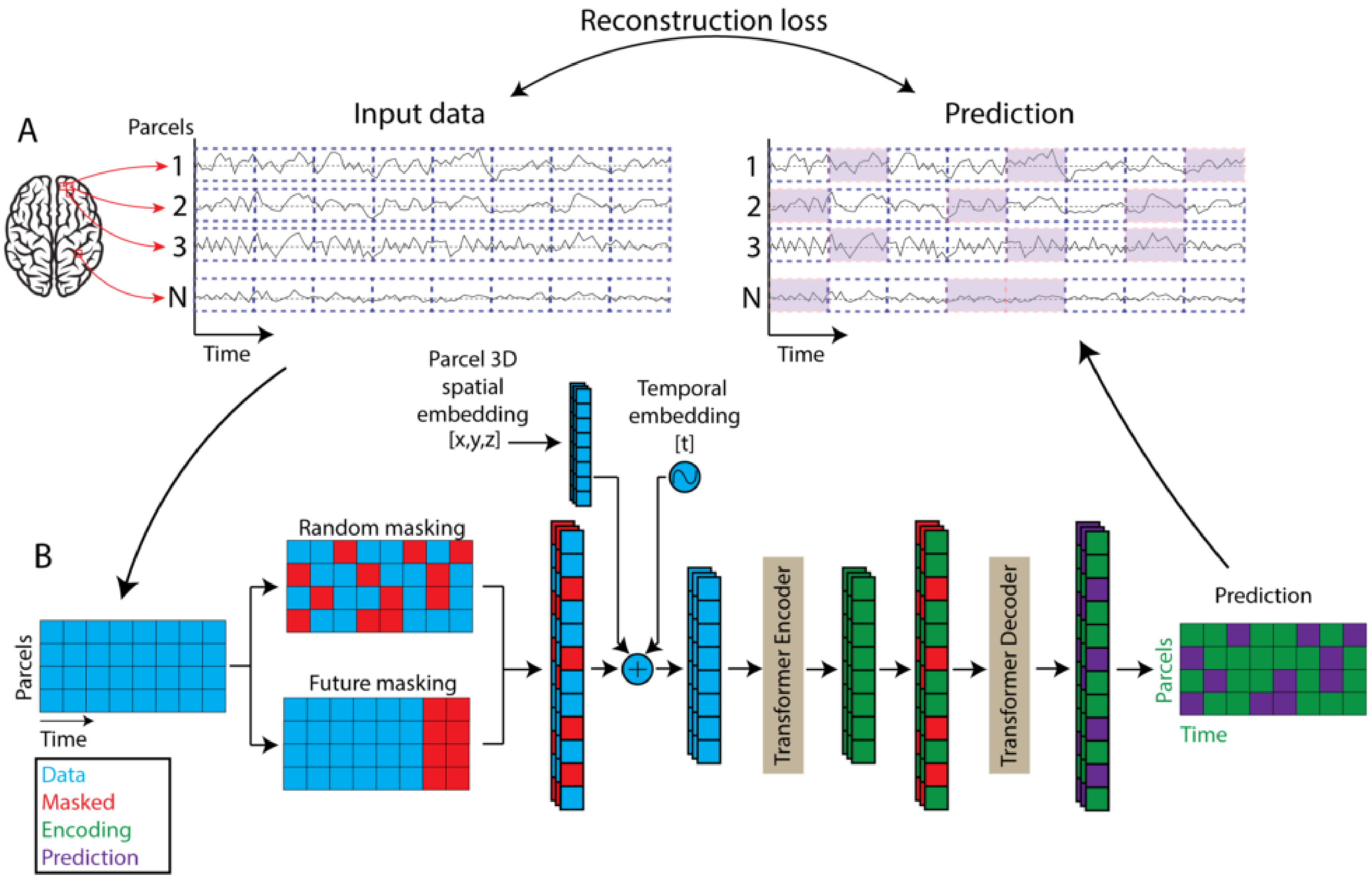


Figure 1. **Overview of the VATT architecture and the self-supervised, multimodal learning strategy**. VATT linearly projects each modality into a feature vector and feeds it into a Transformer encoder. We define a semantically hierarchical common space to account for the granularity of different modalities and employ the noise contrastive estimation to train the model.

Signal processing



Next episode in EEP 596…