RNN Overview - 3

Transformer and BERT

2021. 4. 21

Dajin Han

Index

- Transformer
  - Self Attention
  - Multi-Head Attention
- Bert
- References

Transformer

Attention is all you need

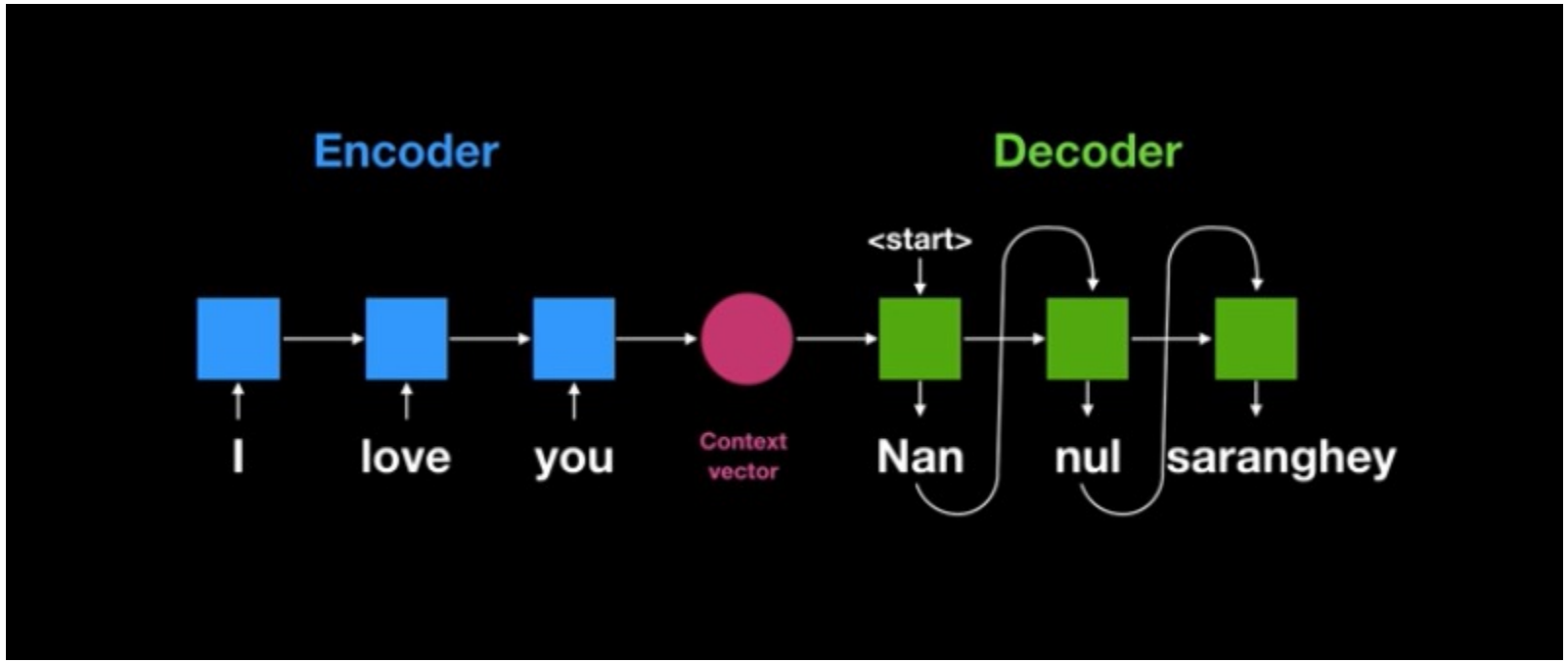
Transformer

- Encoder-Decoder Model
- Reduce RNN
- Parallelism

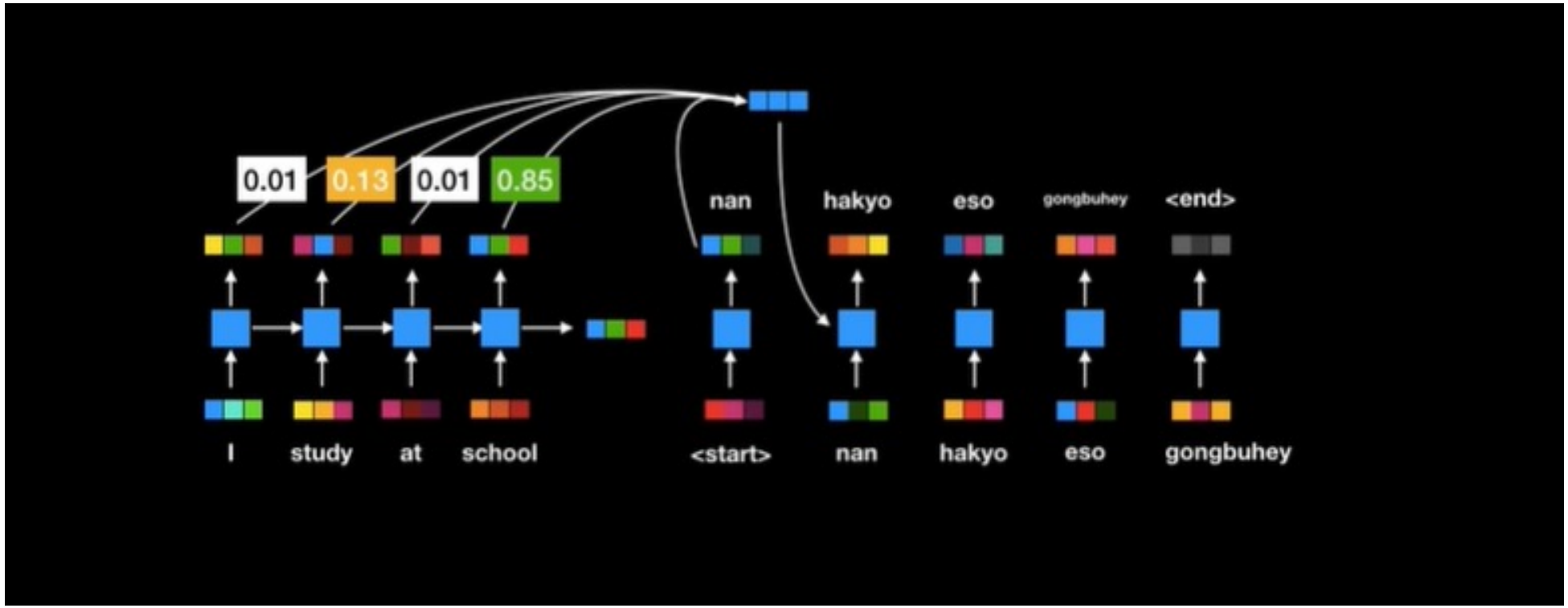
Comparison

- RNN
  - N RNN cells
  - **time step of input sequence**
  - Input is **n_th word(vector)**
- Transformer
  - N(=6) Encoder layer
  - Input is **sequence(matrix!)**

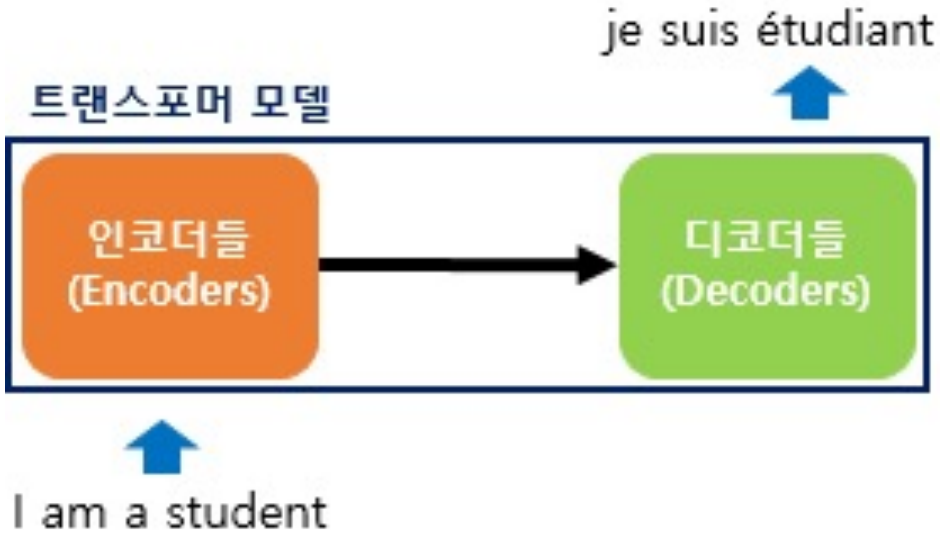
Encoder-Decoder

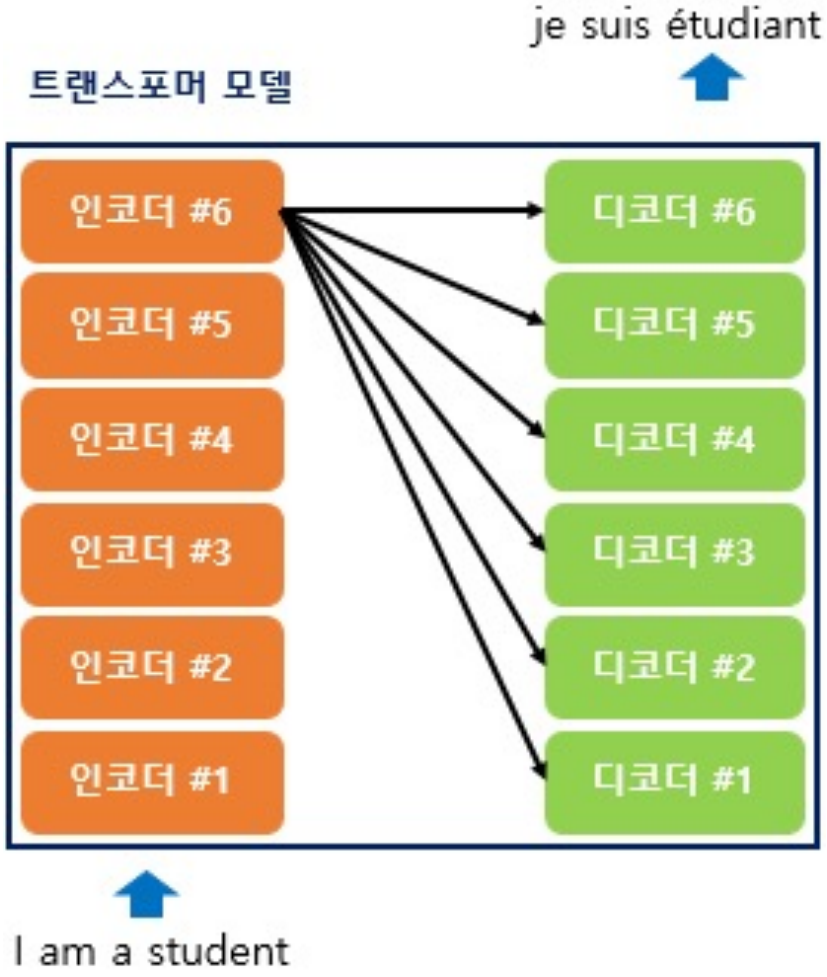


Encoder-Decoder with Attention

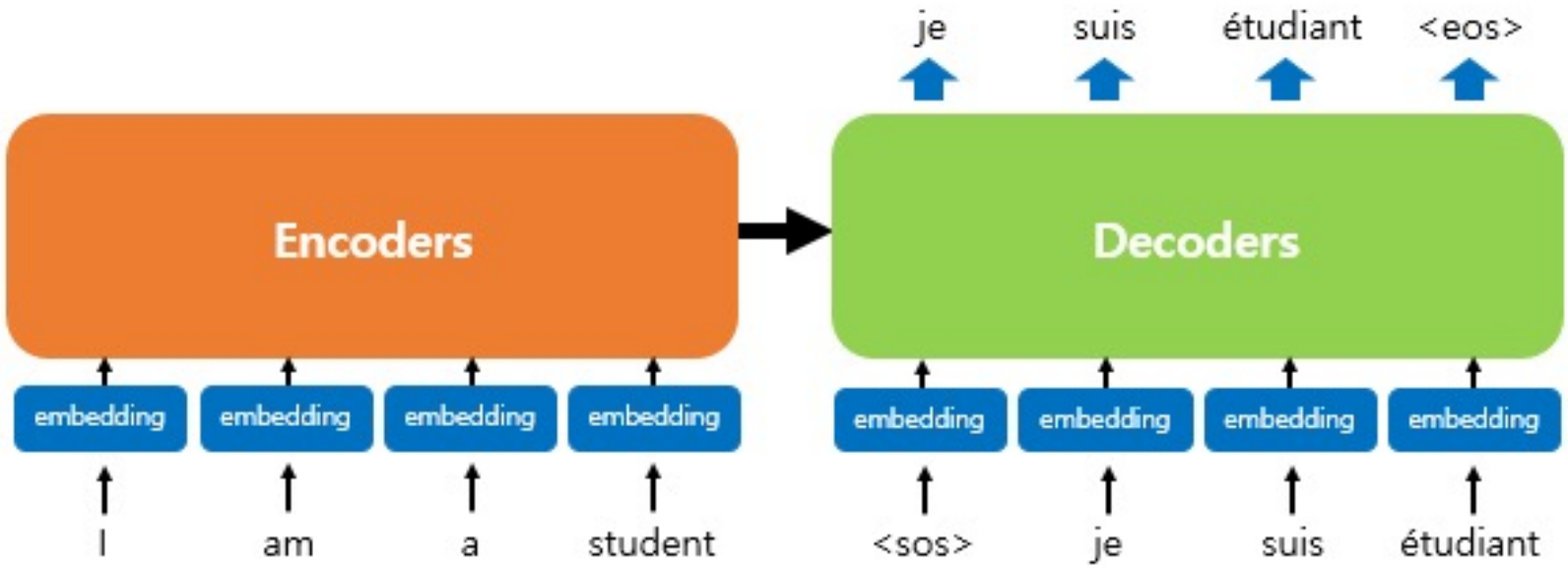


Architecture

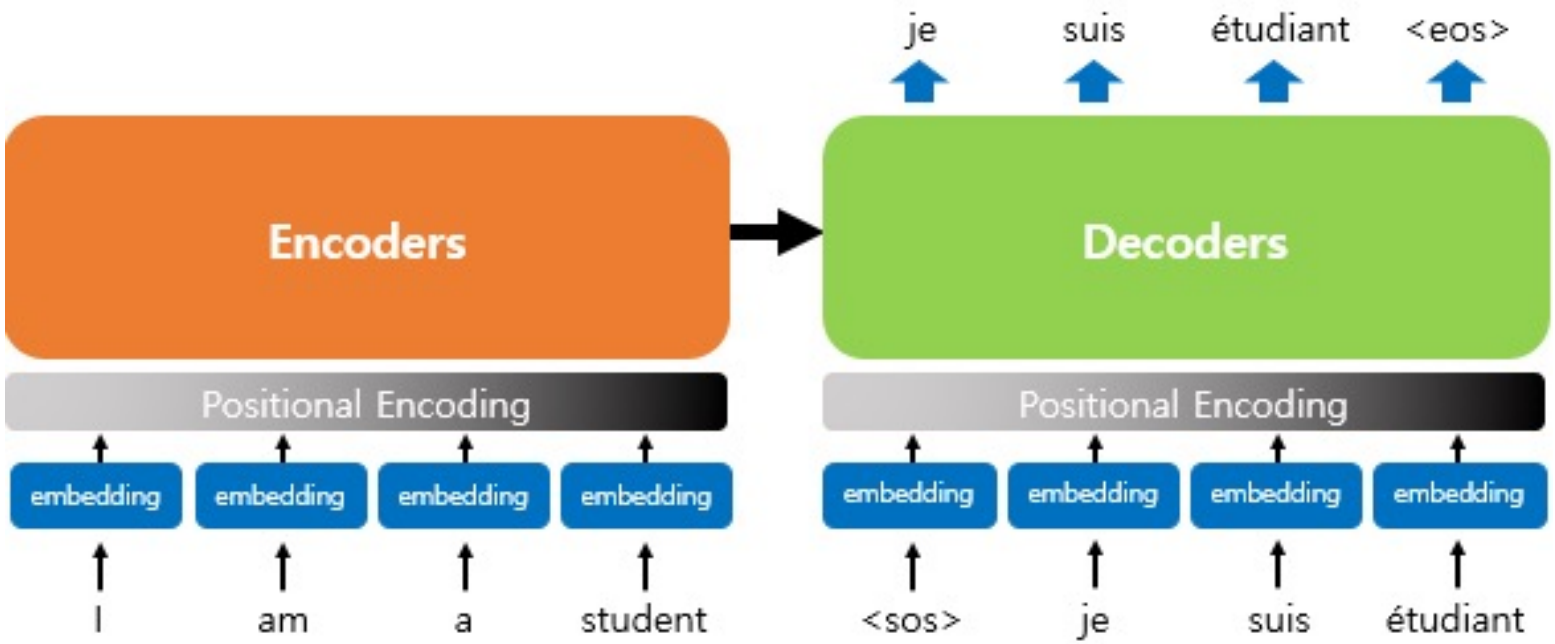




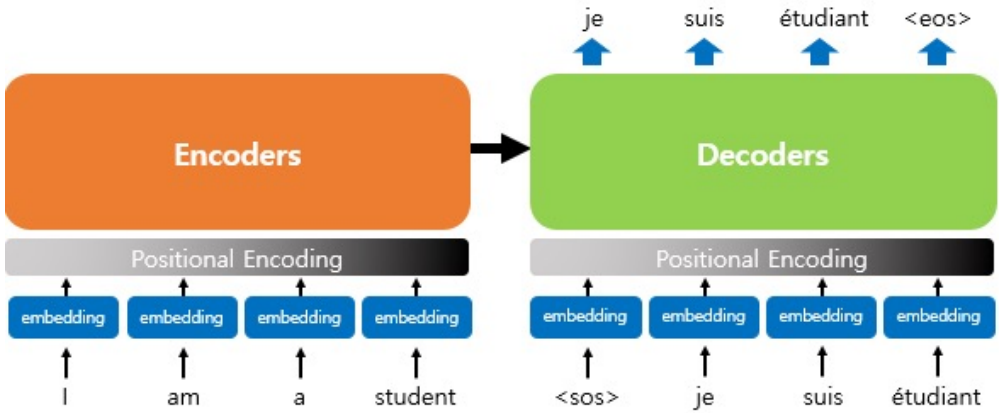
Architecture



Positional Encoding

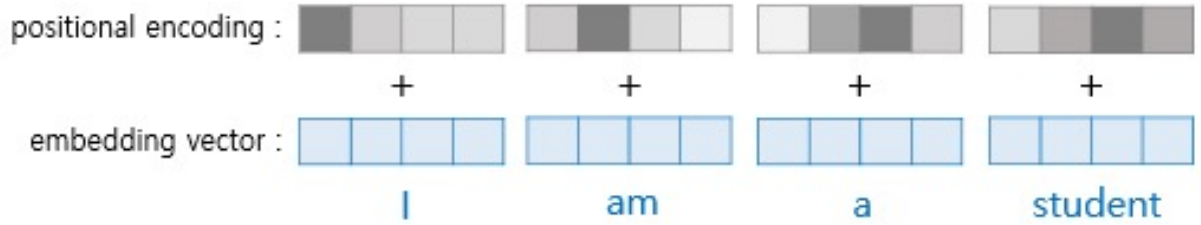


Positional Encoding

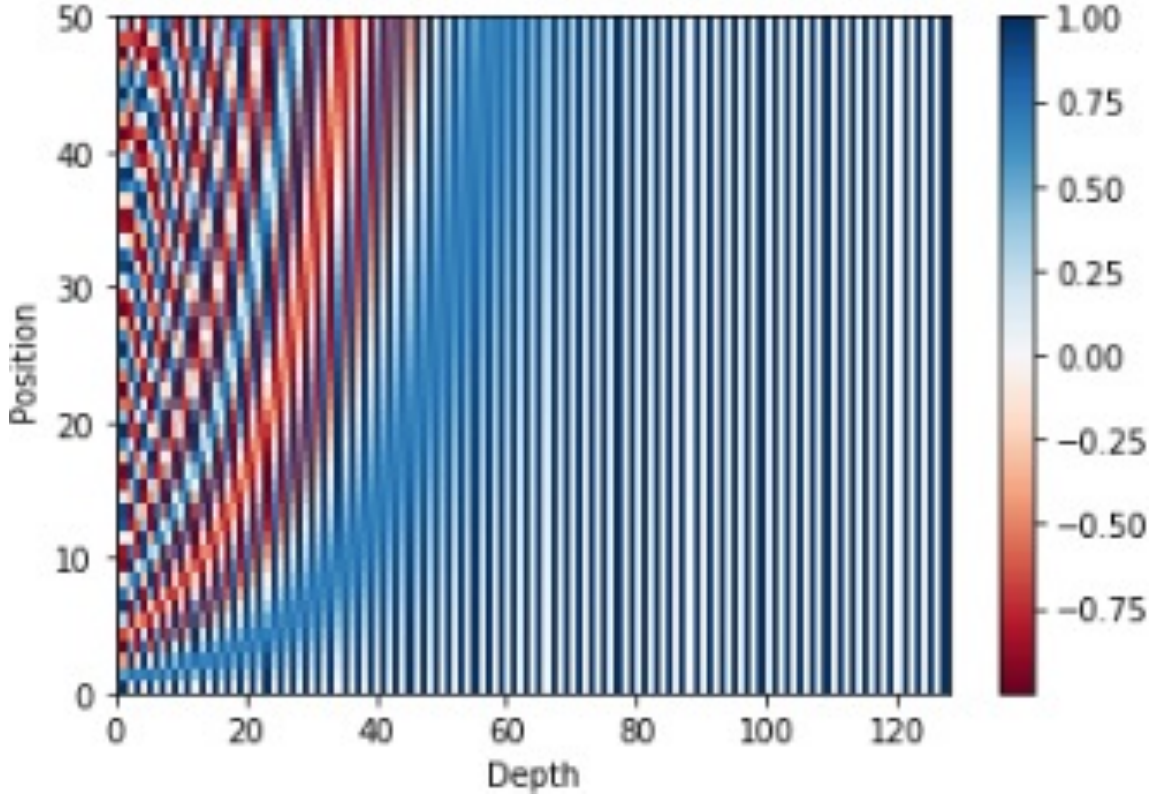


$$PE_{(pos,\,2i)} = sin(pos/10000^{2i/d_{model}})$$

$$PE_{(pos,\,2i+1)} = cos(pos/10000^{2i/d_{model}})$$



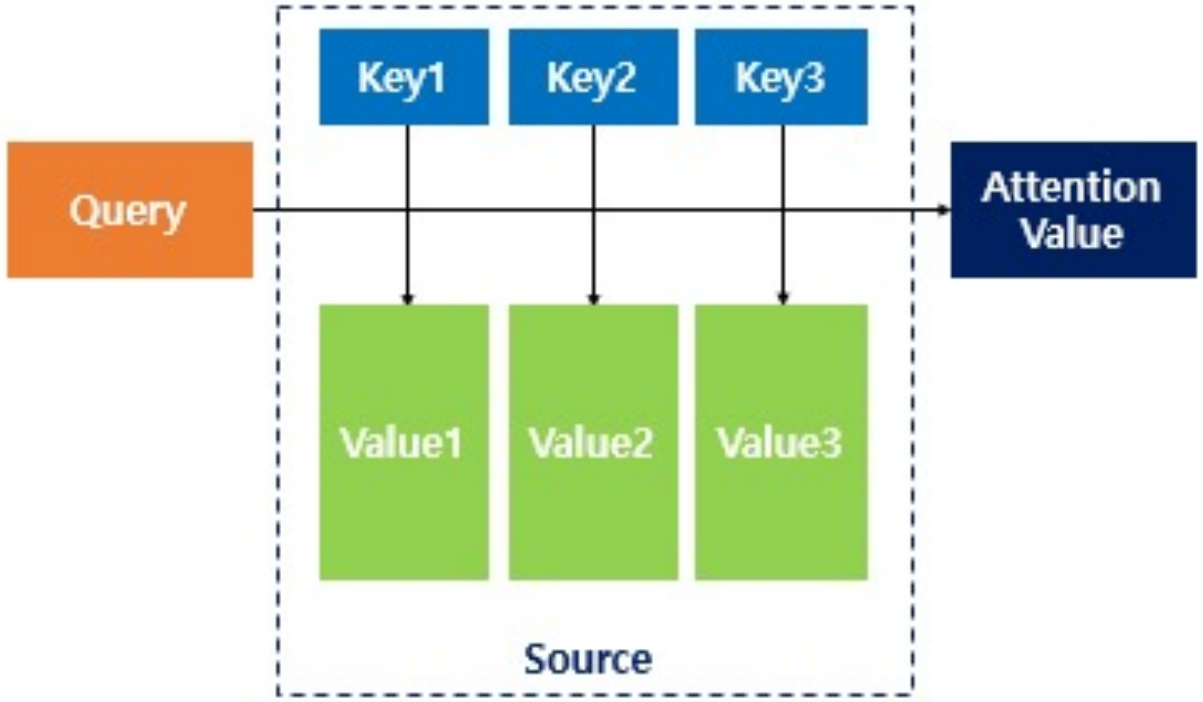
Positional Encoding



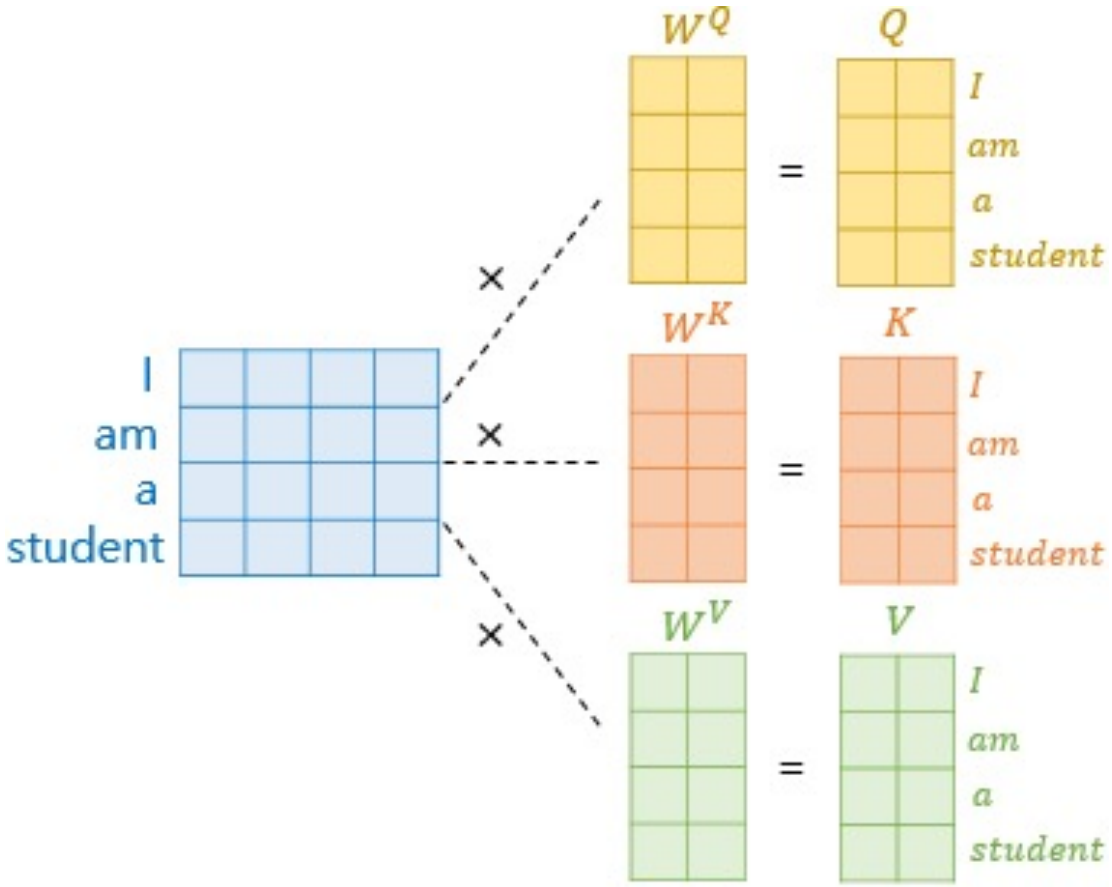
Encoders

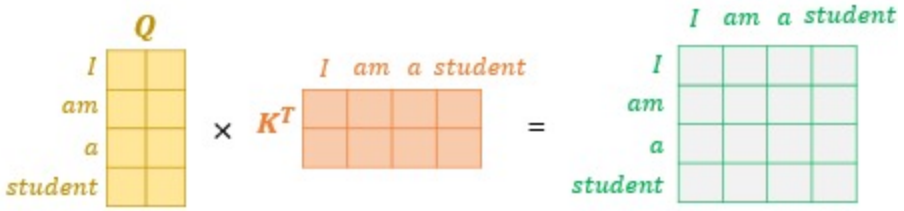
Encoder Layers

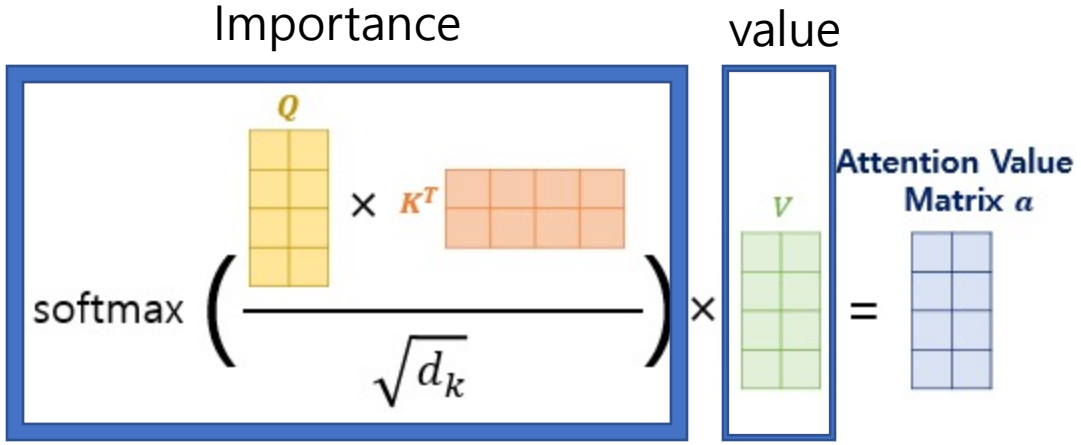
Self Attention



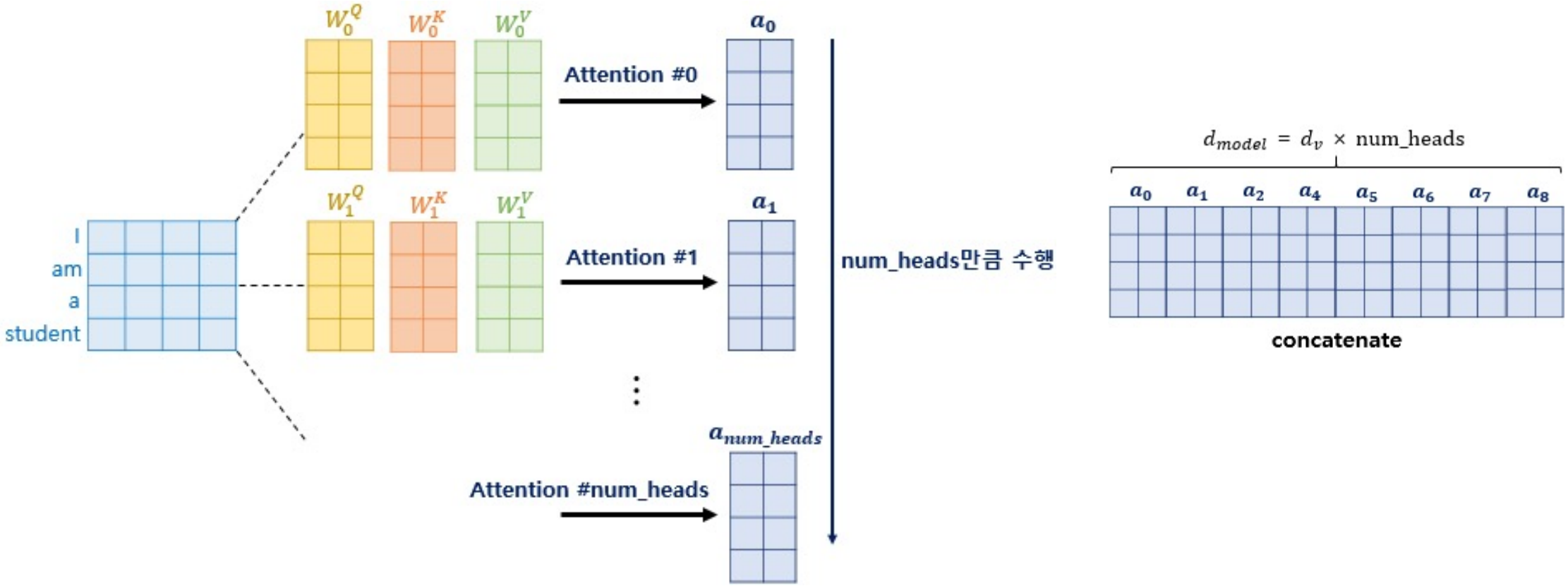
Self Attention



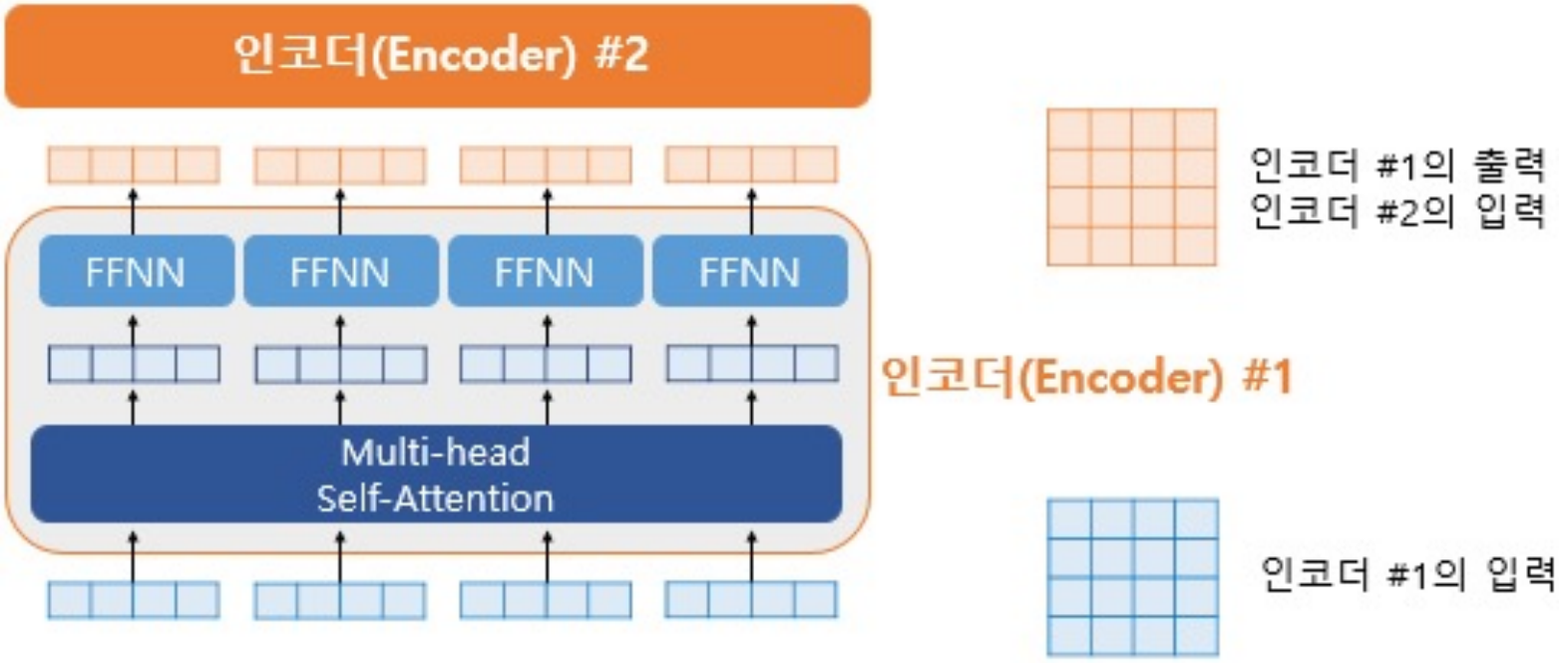




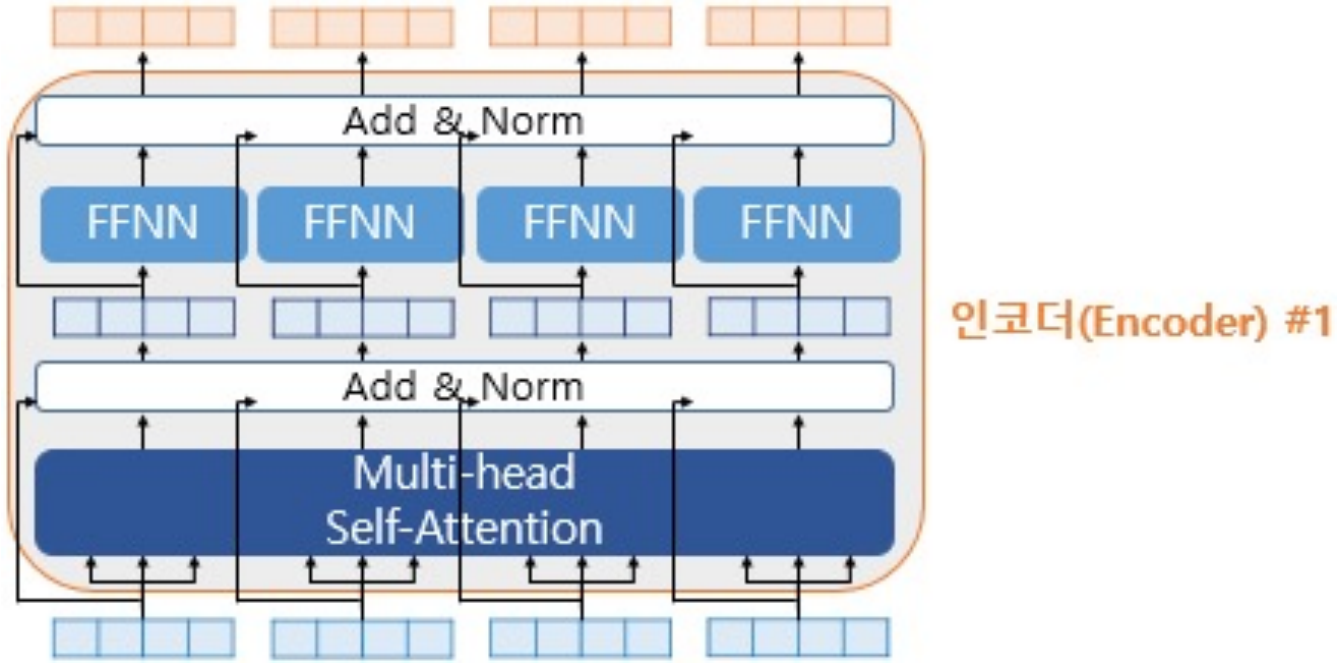
Multi-Head Attention



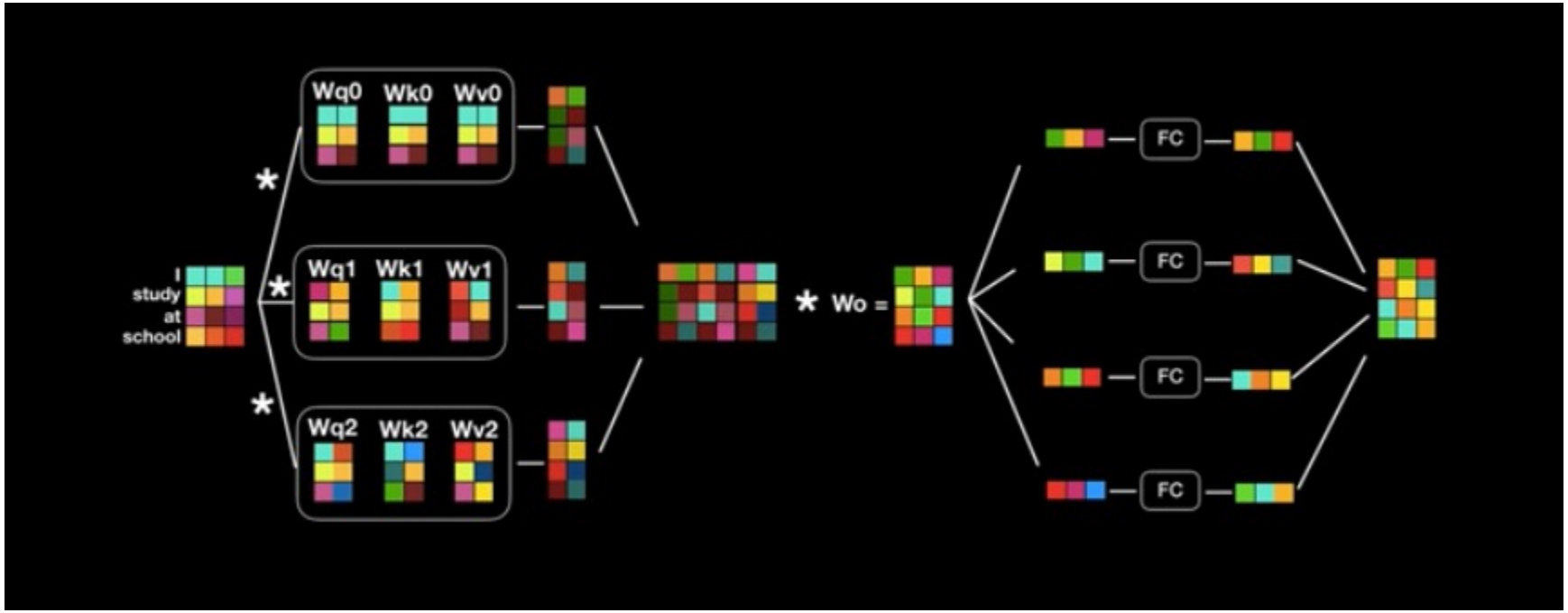
Position wise FFNN



Residual Connection and Normalization



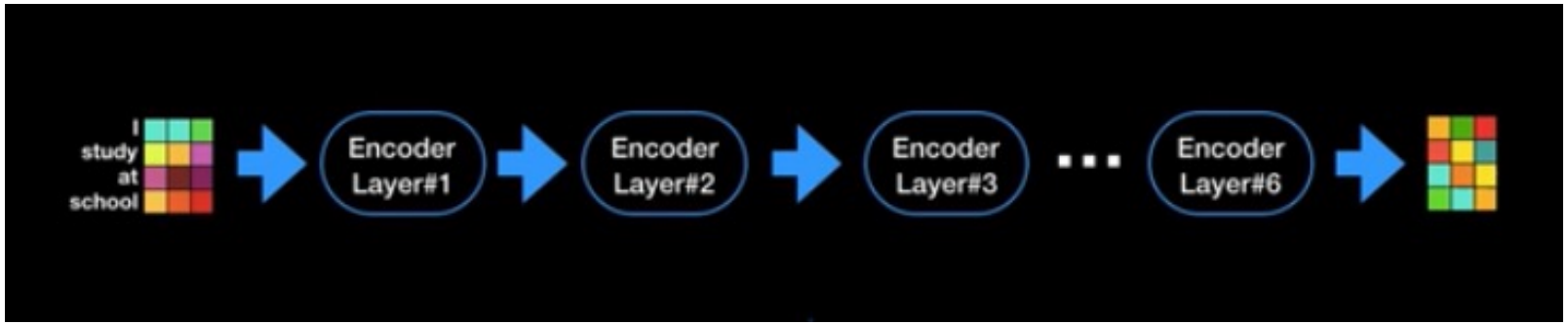
Architecture of Single Encoder

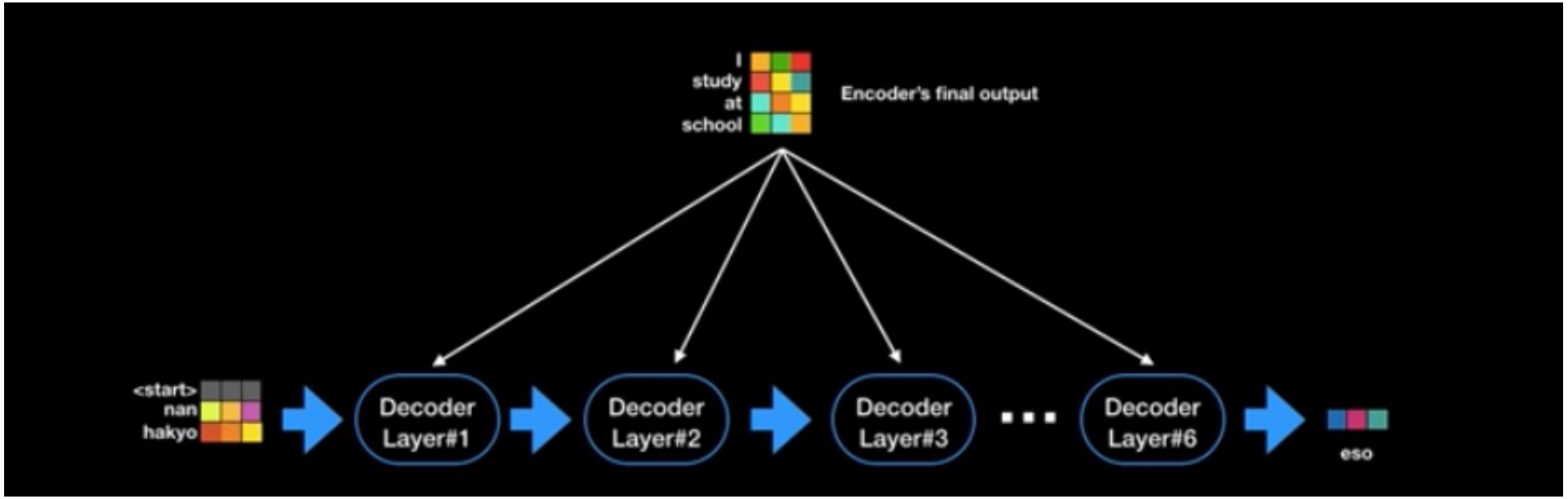


Decoders

Masked Multi-Head Attention

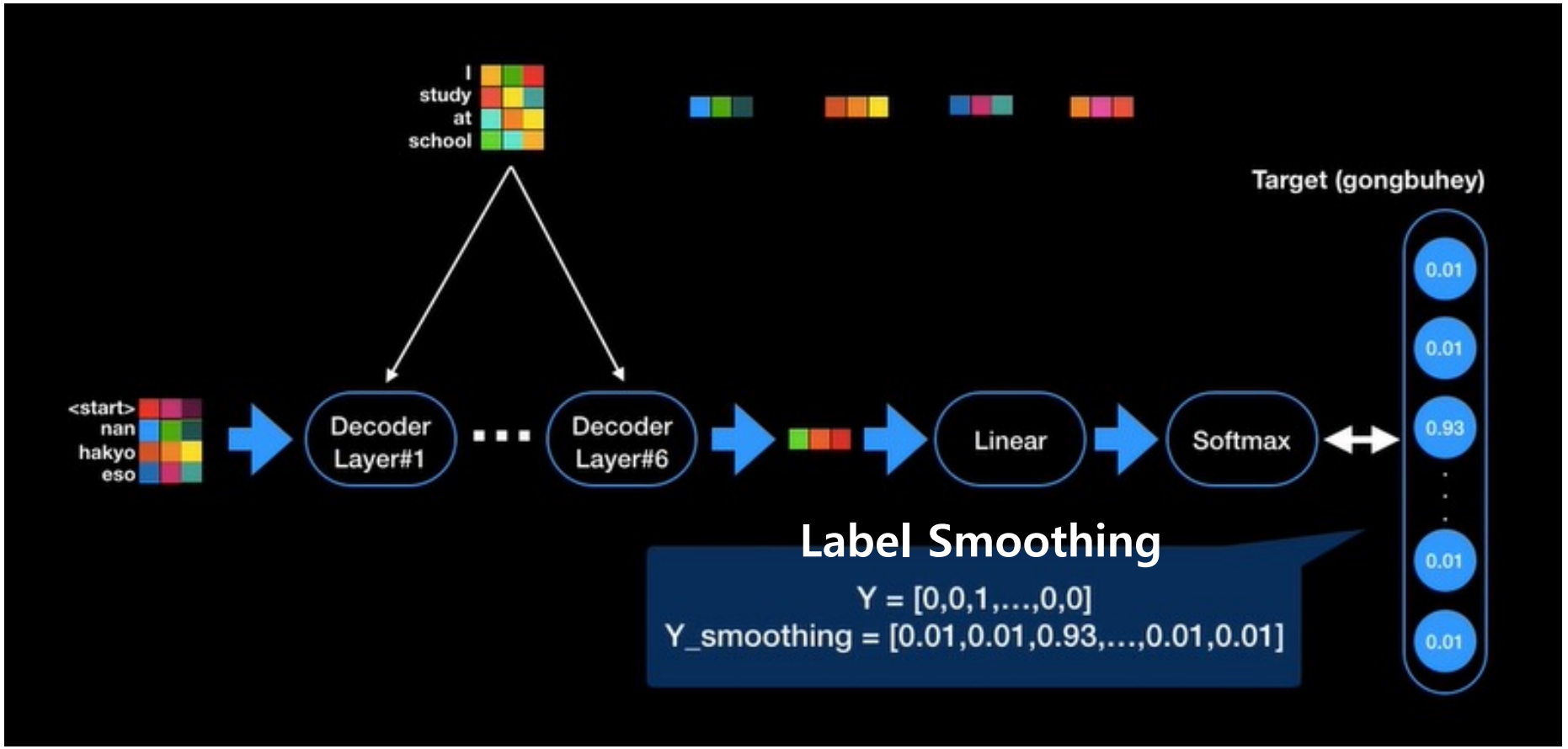
Architecture of Decoders



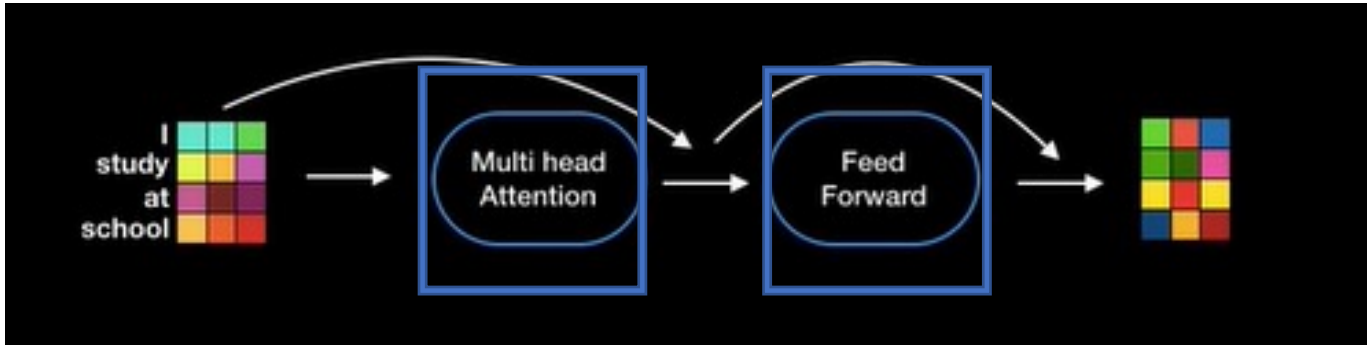


x res_seq_len

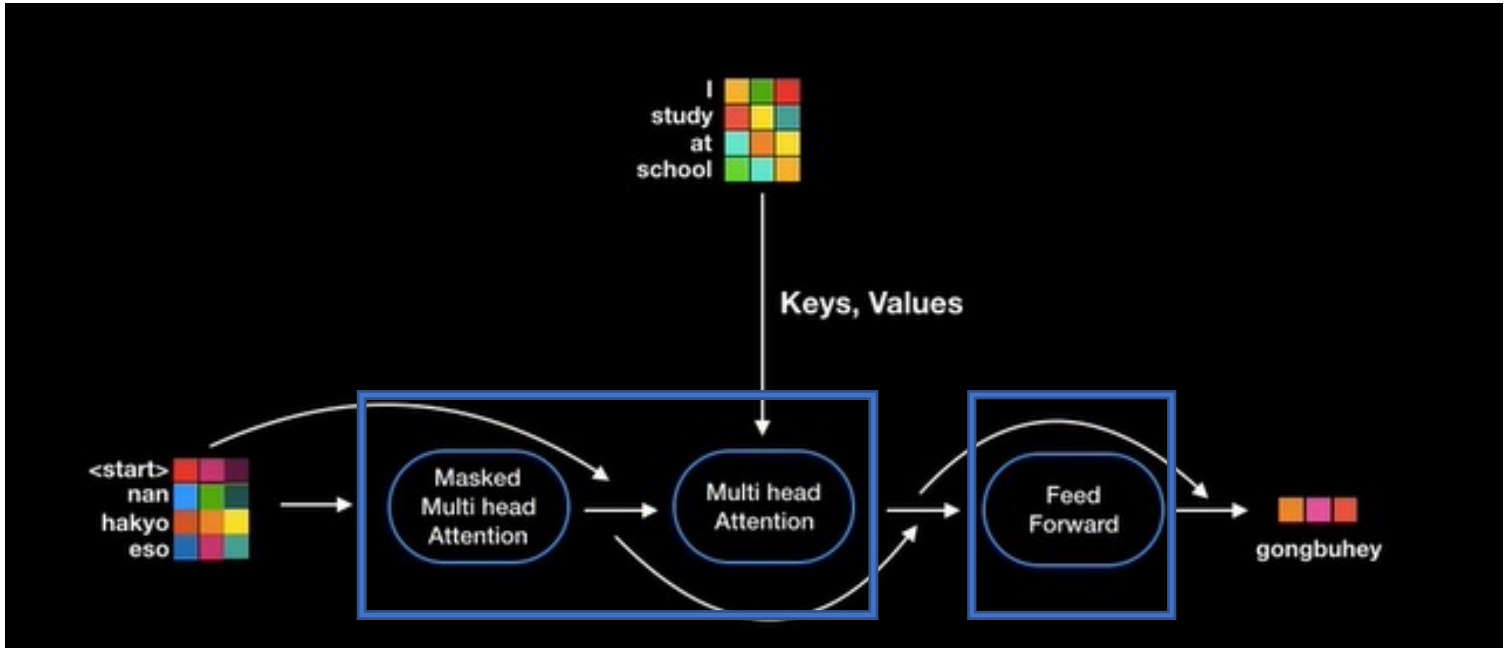
Architecture of Decoders



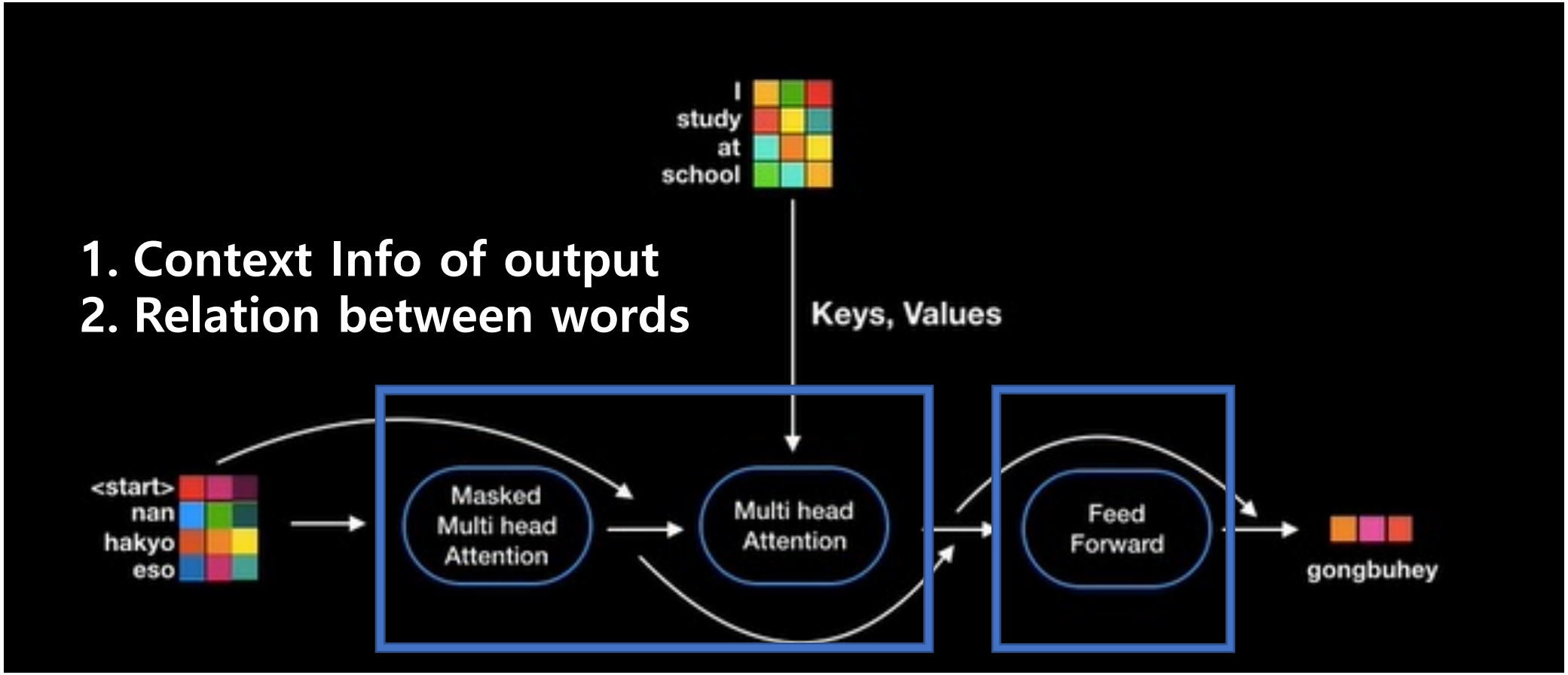
Architecture of Single Decoder



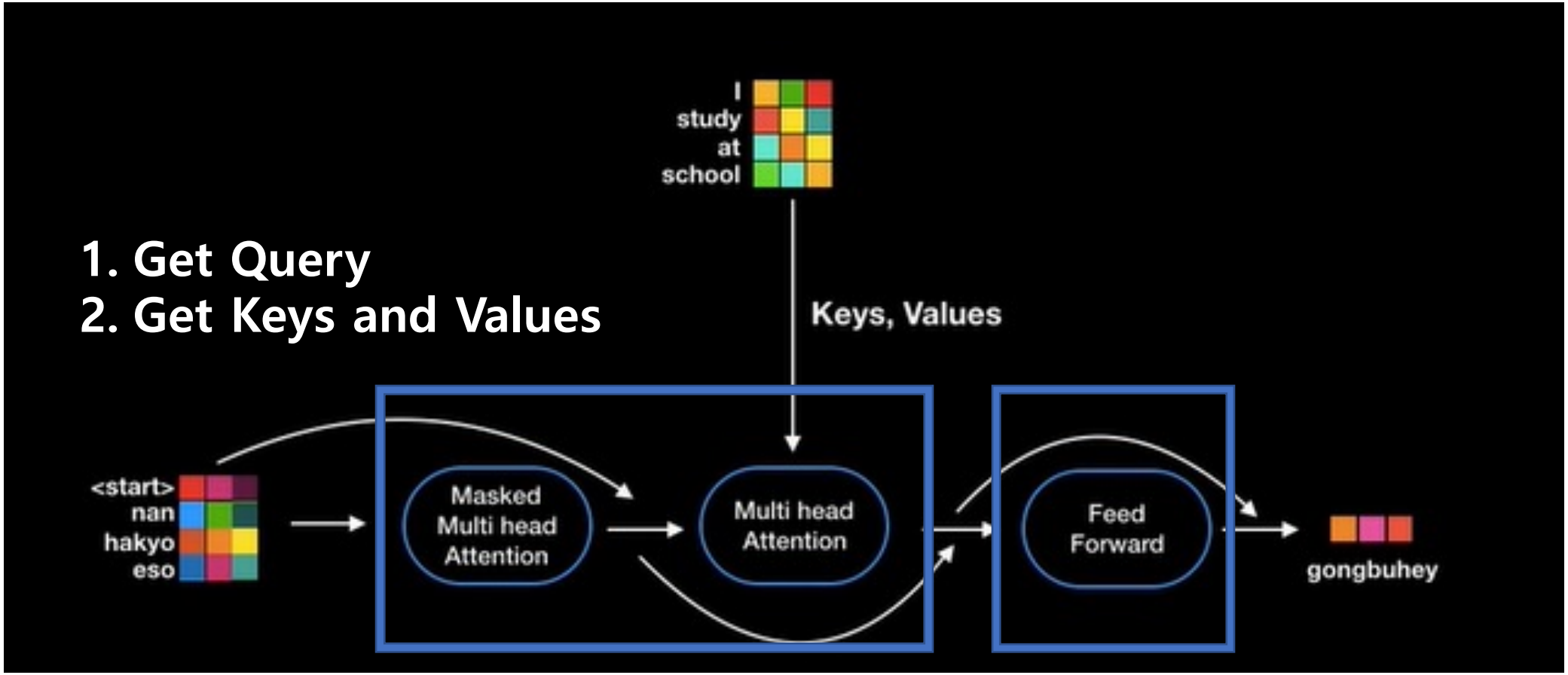
1. Get Features
2. NN Production



Architecture of Single Decoder



Architecture of Single Decoder



BERT

Improve accuracy of seq2seq model

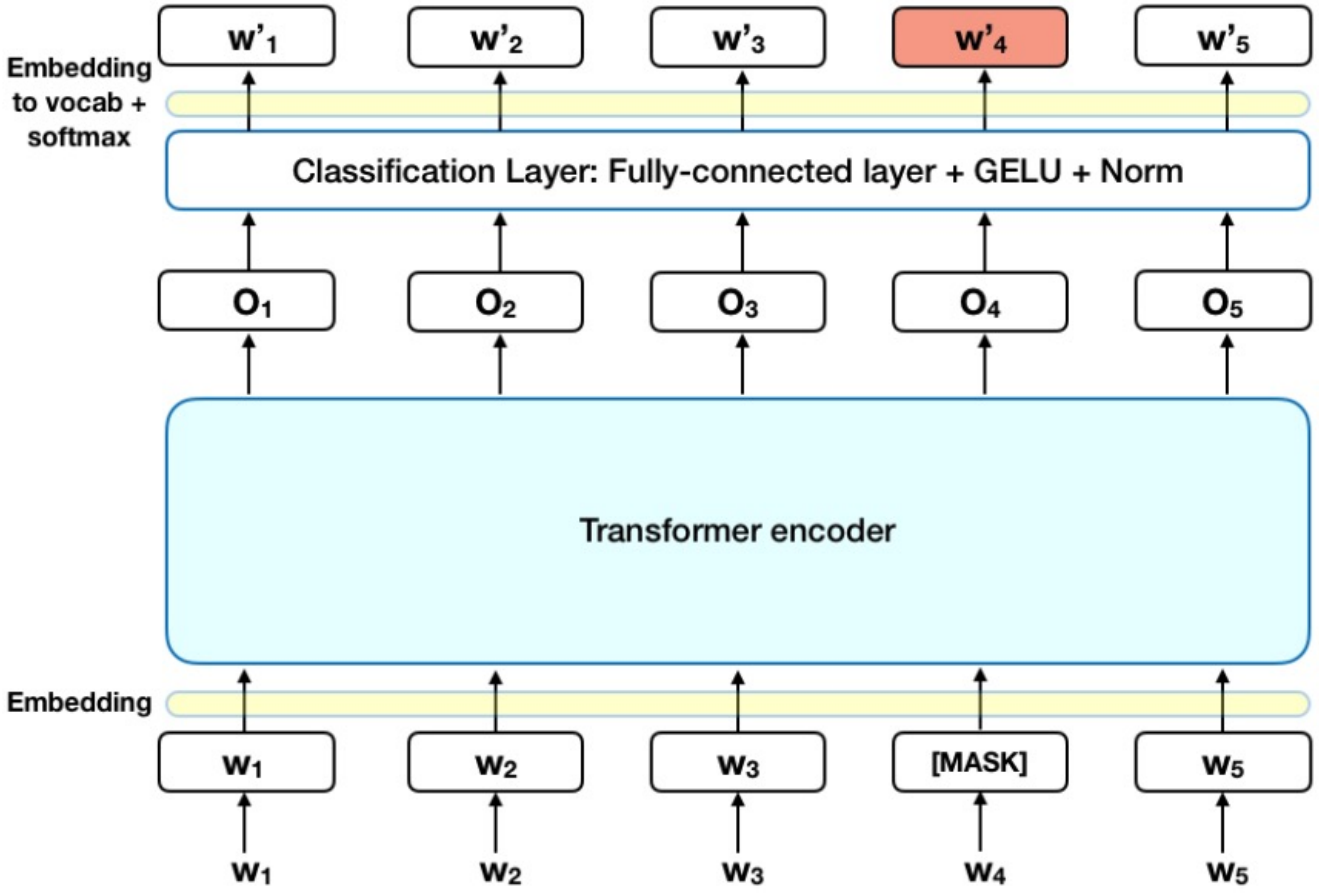
Keypoints

- Encoder part of Transformer
- **Tranfer learning**
- **Bidirectional**
- Fine-tuning without additional network

Pre-training

- Masked Language Model(MLM)
  - Deep Bidirectional
- Next Sentence Prediction
  - NLP task optimization

Masked Language Model



- 85% normal token
- 15% **masked token**
  - 80% [MASK] token
  - 10% random token
  - 10% default token

Next Sentence Prediction

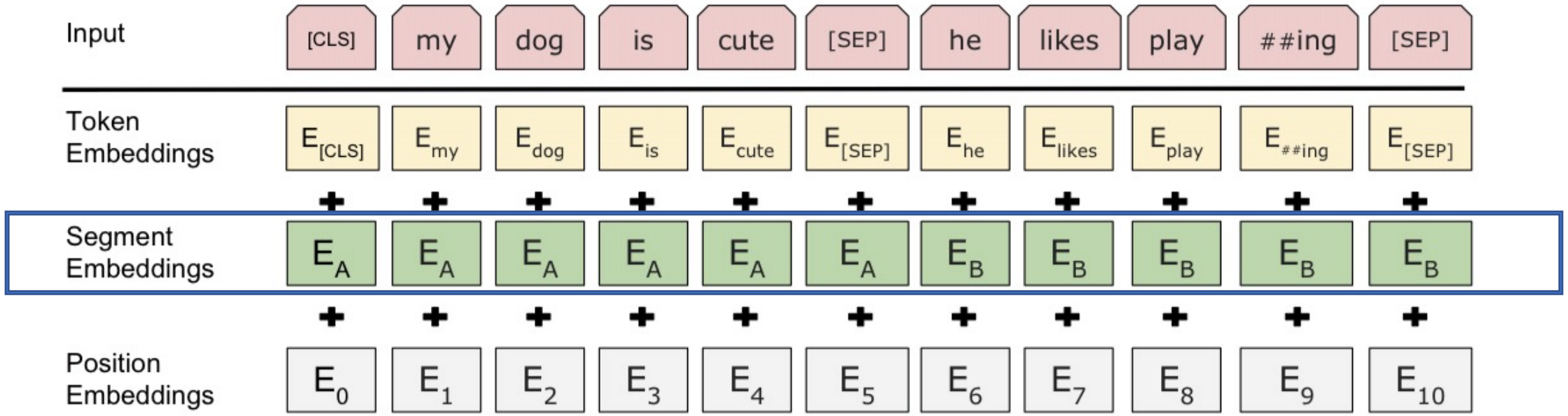
- QA, NLI
- Relation between successive sentences
- Binarized next sentence prediction task

- 50:50

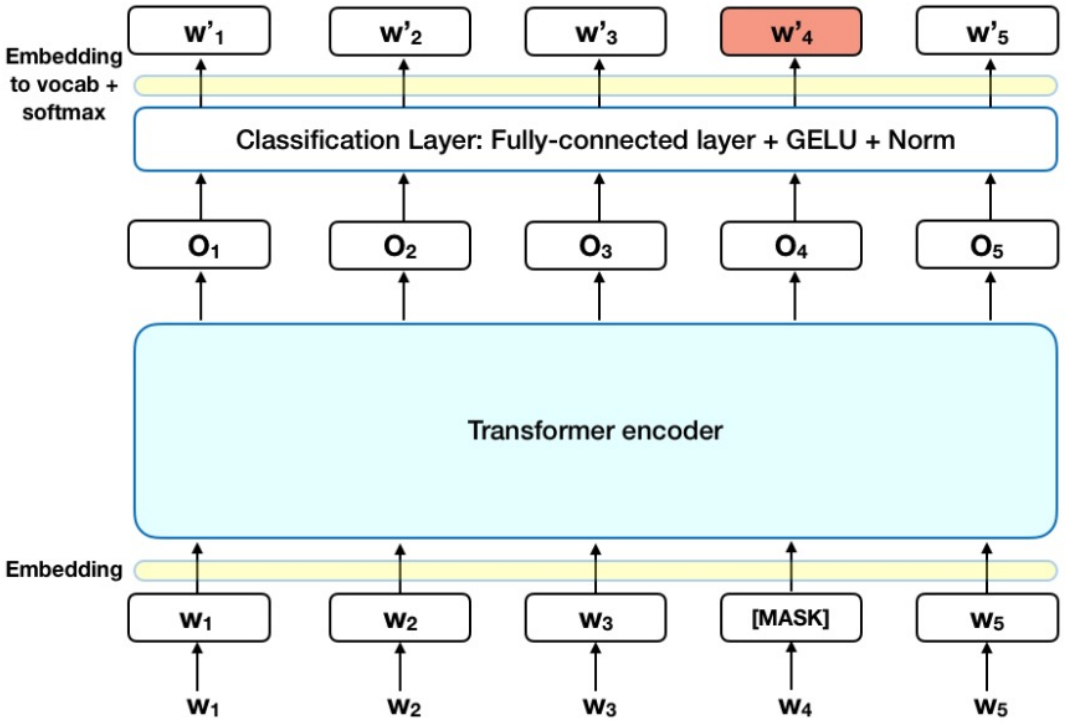
Input = `[CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP]` LABEL = `IsNext`

Input = `[CLS] the man [MASK] to the store [SEP] penguin [MASK] are flight ##less birds [SEP]` Label = `NotNext`

Input Representation

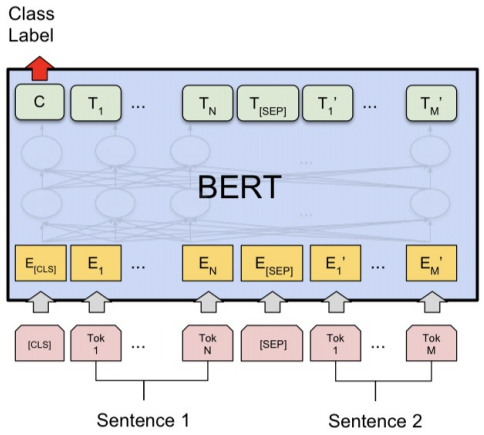


Fine-tuning

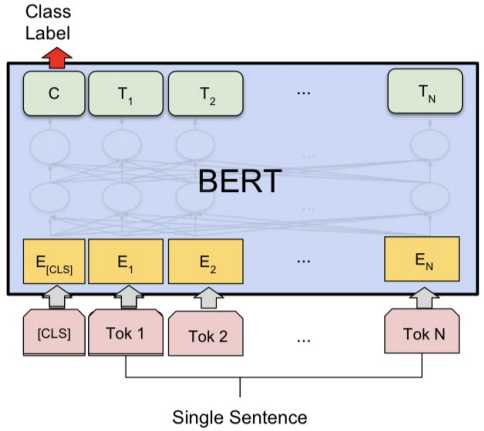


- Depends on task
- Sequence-level classification task
  - Edit or Add Classification layer
- Others

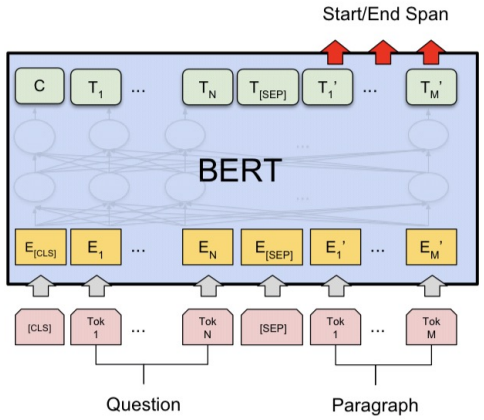
Fine-tuning



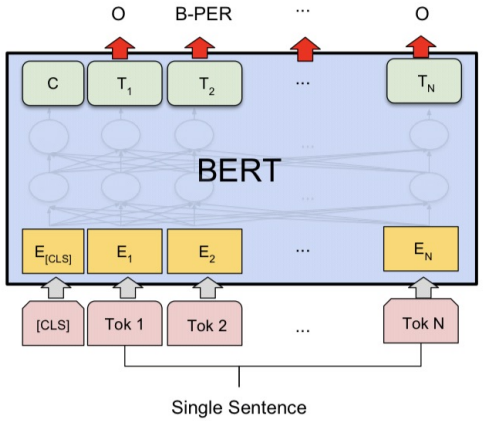
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Experiments

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

References

References

- Ashish, V., Noam, S., Niki, P., Jakob, U., Llion, J., Aidan, N., … Lukasz, K. (2017). Attention Is All You Need
- Jakob, D., Ming-Wei, C., Kenton, L., Kristina, T. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

- https://wikidocs.net/31379
- https://pozalabs.github.io/transformer/
- https://www.youtube.com/watch?v=mxGCEWOxfe8&t=826s
- https://mino-park7.github.io/nlp/2018/12/12/bert-%EB%85%BC%EB%AC%B8%EC%A0%95%EB%A6%AC/?fbclid=IwAR3S-8iLWEVG6FGUVxoYdwQyA-zG0GpOUzVEsFBd0ARFg4eFXqCyGLznu7w