Hi BERT !

Chang Shen
Department of Biostatistics
Yale School of Public Health

Bidirectional Encoder Representations from Transformers

- Representations
  - Traditional word embedding
  - Contextualized embedding
  - ELMo
- Transformers
  - Bert
  - Seq2Seq
  - Self Attention Mechanism
- Applications - Transfer Learning

Make computer understand the meaning of the words

Representations

Representations

1 of N encoding

Informatics = [1, 0, 0, 0]
Computer Science = [0, 1, 0, 0]
Python = [0, 0, 1, 0]
Hogwarts = [0, 0, 0, 1]

- Sparse
- High Dimension
- Can’t express word relationship

Representations

1 of N encoding

Informatics = [1, 0, 0, 0]
Computer Science = [0, 1, 0, 0]
Python = [0, 0, 1, 0]
Hogwarts = [0, 0, 0, 1]

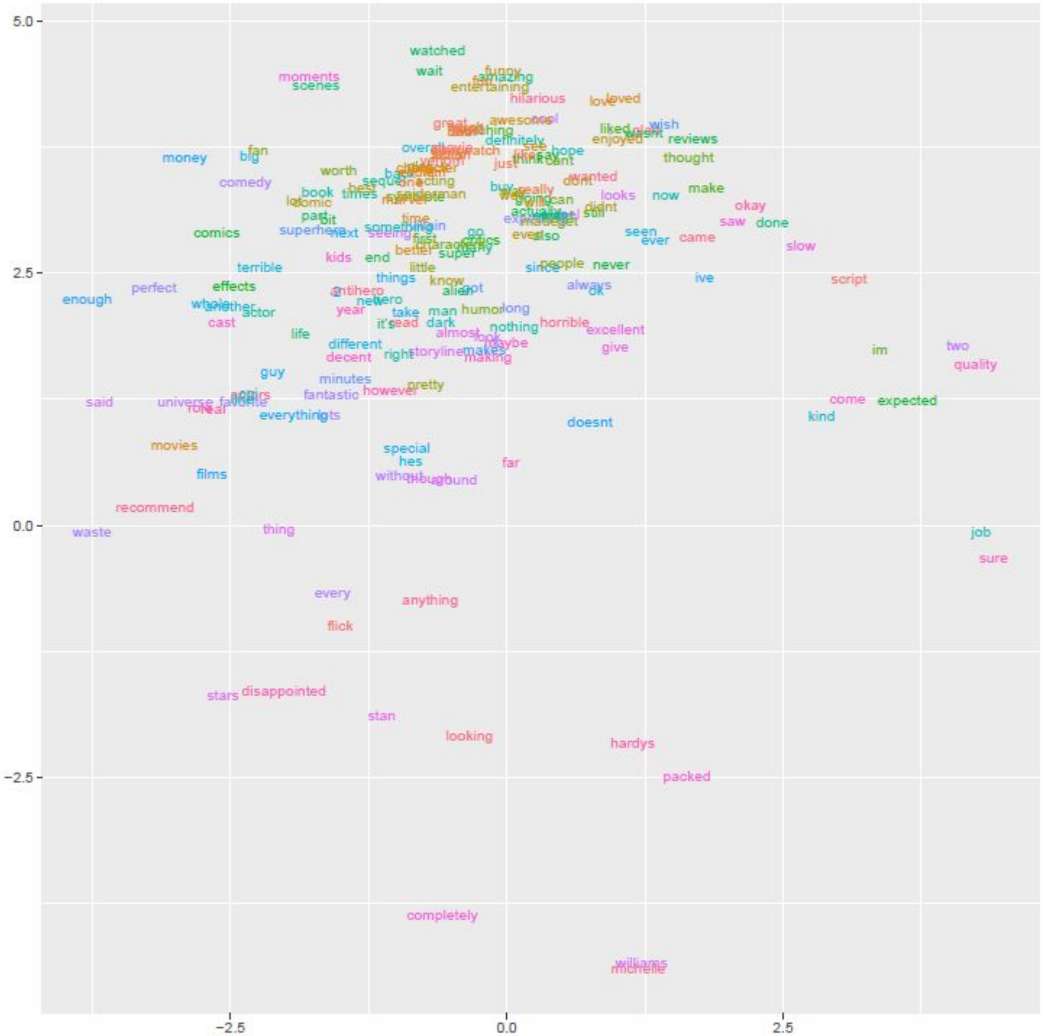
- Sparse
- High Dimension
- Can express word relationship

Word Embedding

Informatics = [0.5, 0.4, 0, 1, 0.9]
Computer Science= [0.5, 1, 0, 0.9, 0.8]
Python = [0.3, 0.99, 0, 0.1, 0.8]
Hogwarts = [0, 0, 1, 0, 0]

- Dense
- Lower-dimension
- Learn from data

Representations



Word Embedding

Informatics = [0.5, 0.4, 0, 1, 0.9]
Computer Science= [0.5, 1, 0, 0.9, 0.8]
Python = [0.3, 0.99, 0, 0.1, 0.8]
Hogwarts = [0, 0, 1, 0, 0]

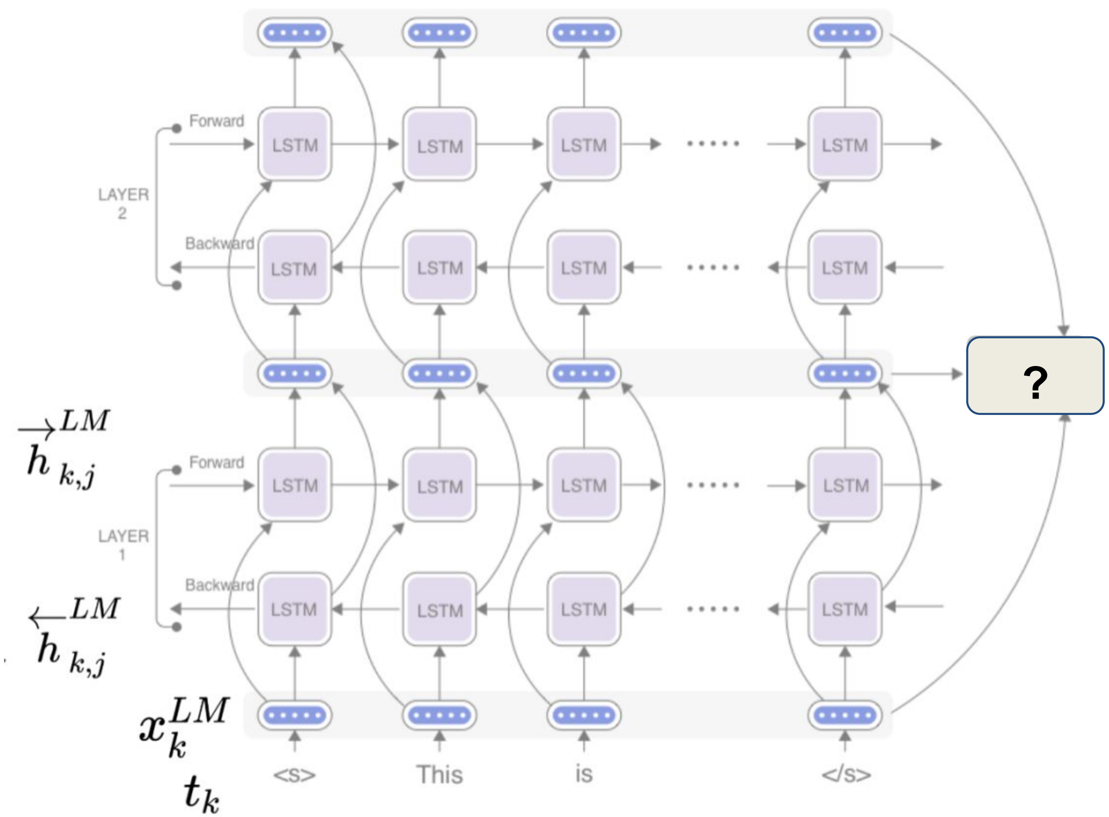
- Dense
- Lower-dimension
- Learn from data

polysemy?

Representations

Contextualized Word Embedding

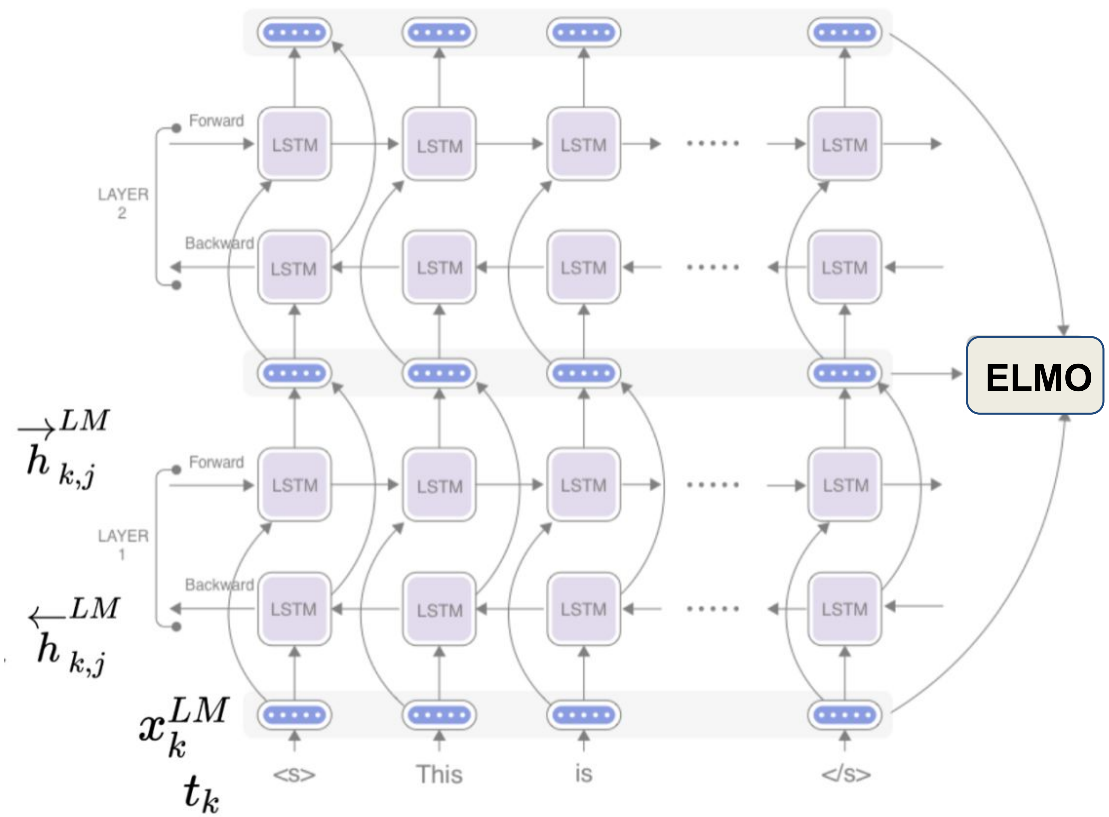
- Each word token has it’s own embedding



Representations

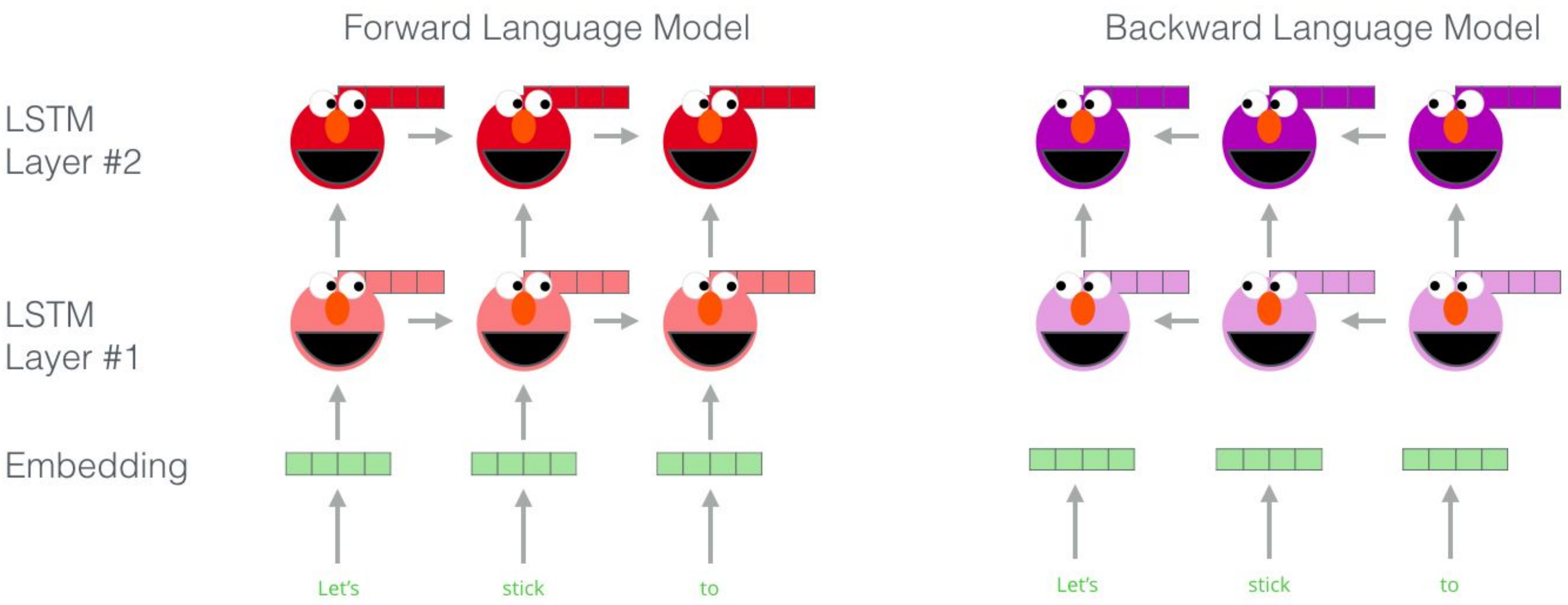
Contextualized Word Embedding

- Each word token has it’s own embedding



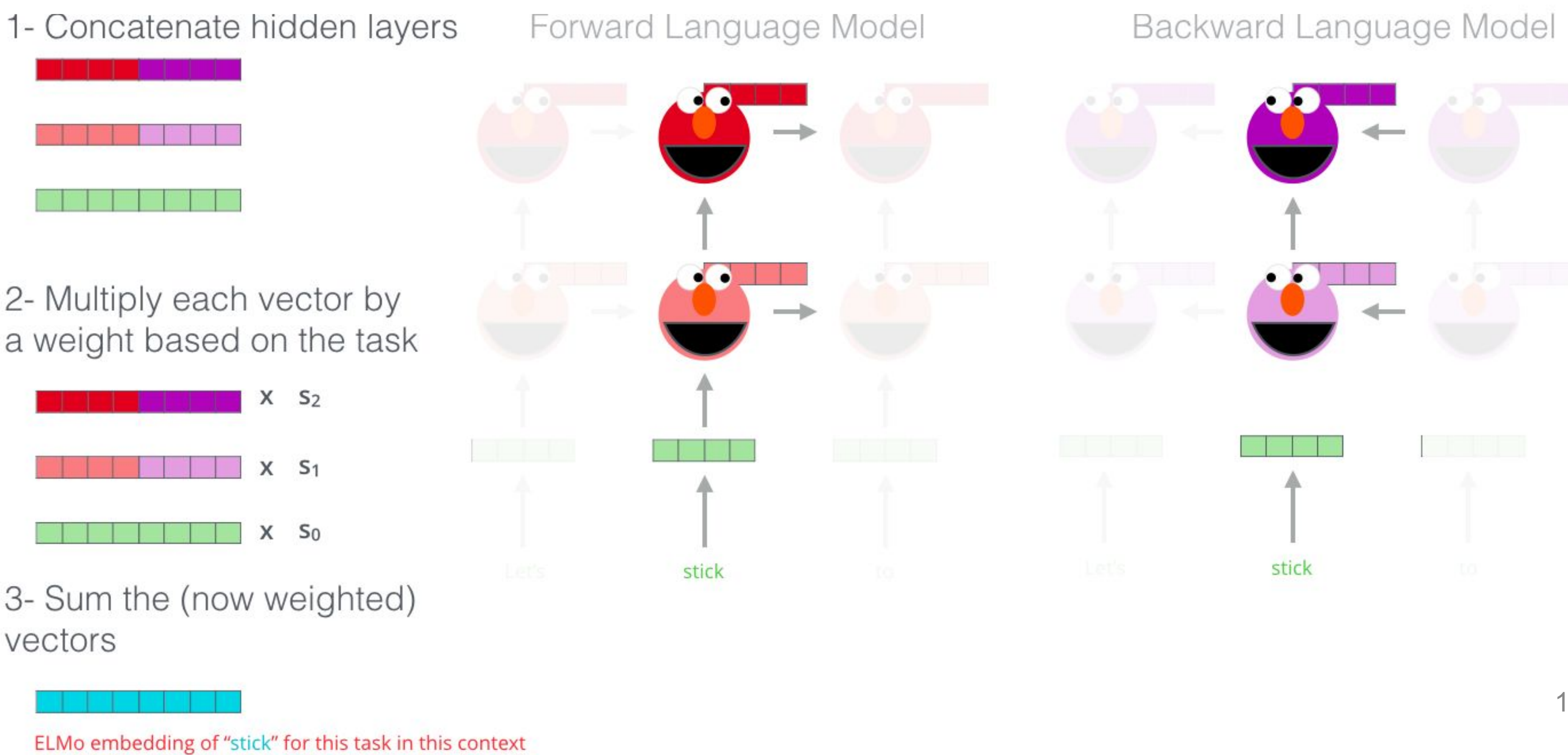
Representations & Bidirections

Embedding from **L**angage **Mo**del(ELMo)



Representations & Bidirections

Embedding from **L**angage **Mo**del(ELMo)

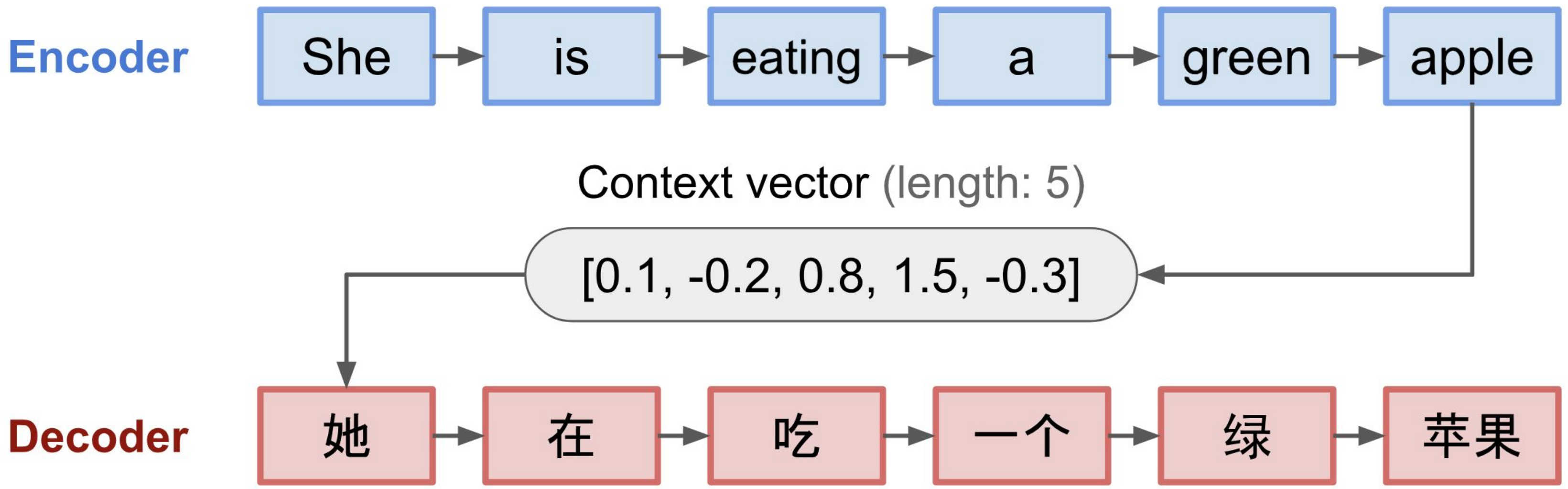


A step from BERT

Transformer

Seq2Seq Model

- An **encoder** processes the input sequence and compresses the information into a context vector of a *fixed length*. This representation is expected to be a good summary of the meaning of the *whole* source sequence.
- A **decoder** is initialized with the context vector to emit the transformed output. The early work only used the last state of the encoder network as the decoder initial state.

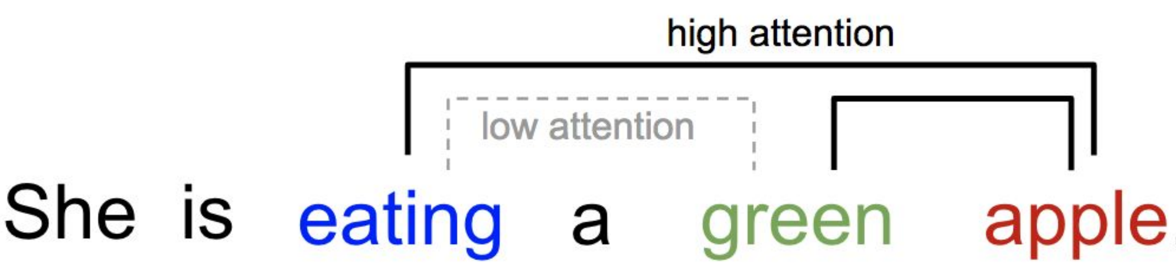


Transformers

Attention Mechanism

Attention is all you need

Attention is, to some extent, motivated by how we pay visual attention to different regions of an image or correlate words in one sentence

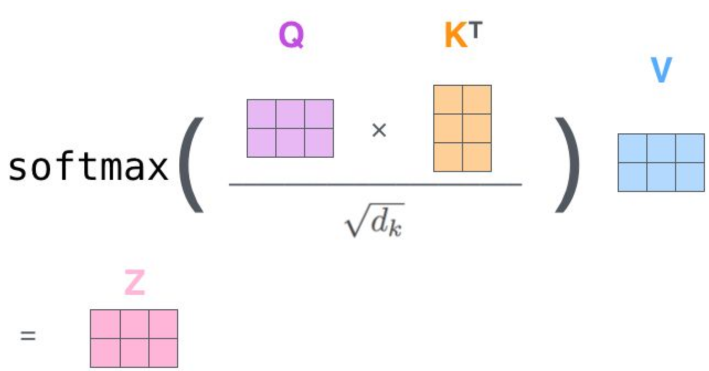


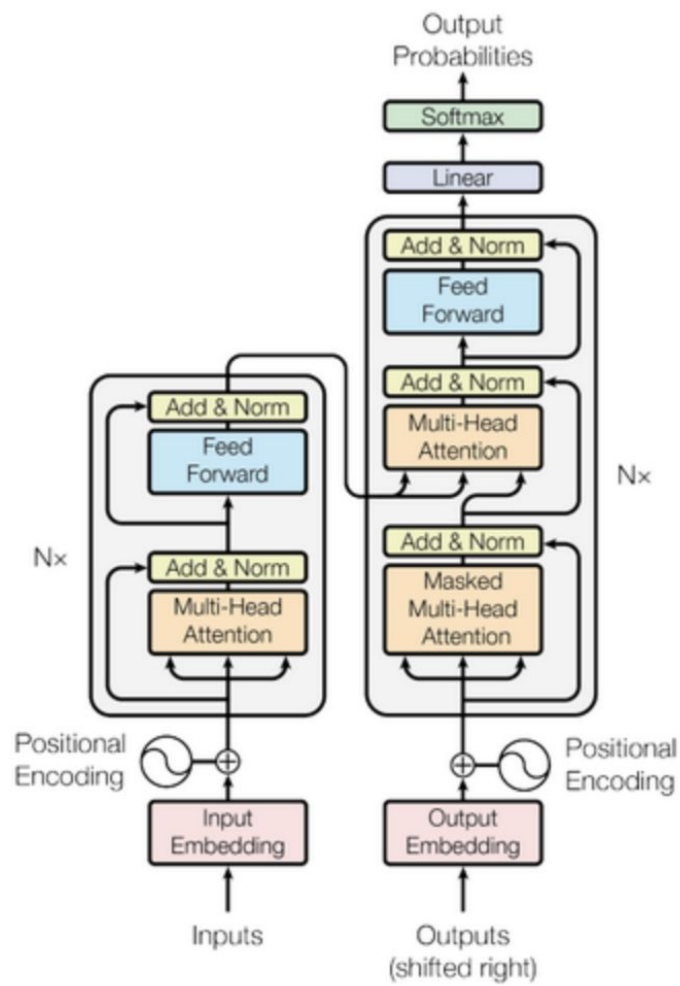
Transformers

Key The word to be matched

Query Match other words

Value attention weight
Information need to be matched





BERT

1. **Masked language model**
   Where some words are hidden (15% of words are masked) and the model is trained to predict the missing words

2. **Next sentence prediction**
   Where the model is trained to identify whether sentence B follows (is related to) sentence A

Why Bert is so popular

Applications

Applications

Downstream tasks

- Machine Translation
- Sentiment Analysis
- Text summarization
- Recommended system
- Inference
- ……

XL-BERT, RoBERTa
ALBERT, ERNIE,
DistillBERT,
Multilingual-BERT
…...

Applications

Medical related BERT?

- Readmission prediction- **ClincalBERT**
- Patient matching - **DeepEnroll**
- Pretrain on Medical paper- **BlueBERT/BioBERT**
- ……..

Conclusions and discussions