

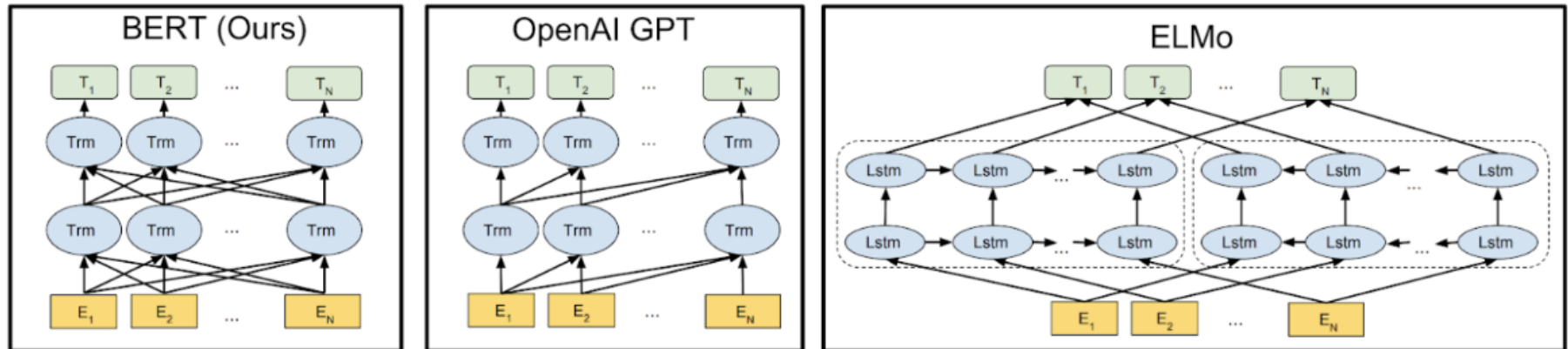
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Devlin et al.

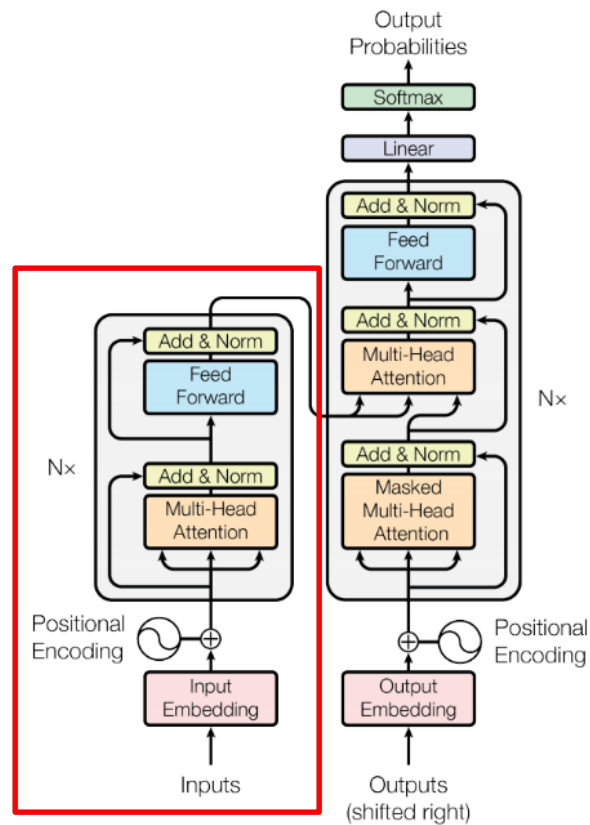
Jungsoo Park

Data Mining & Information Systems Lab.
Department of Computer Science and Engineering,
College of Informatics, Korea University

Natural Language Processing's Imagenet



Transformer Model



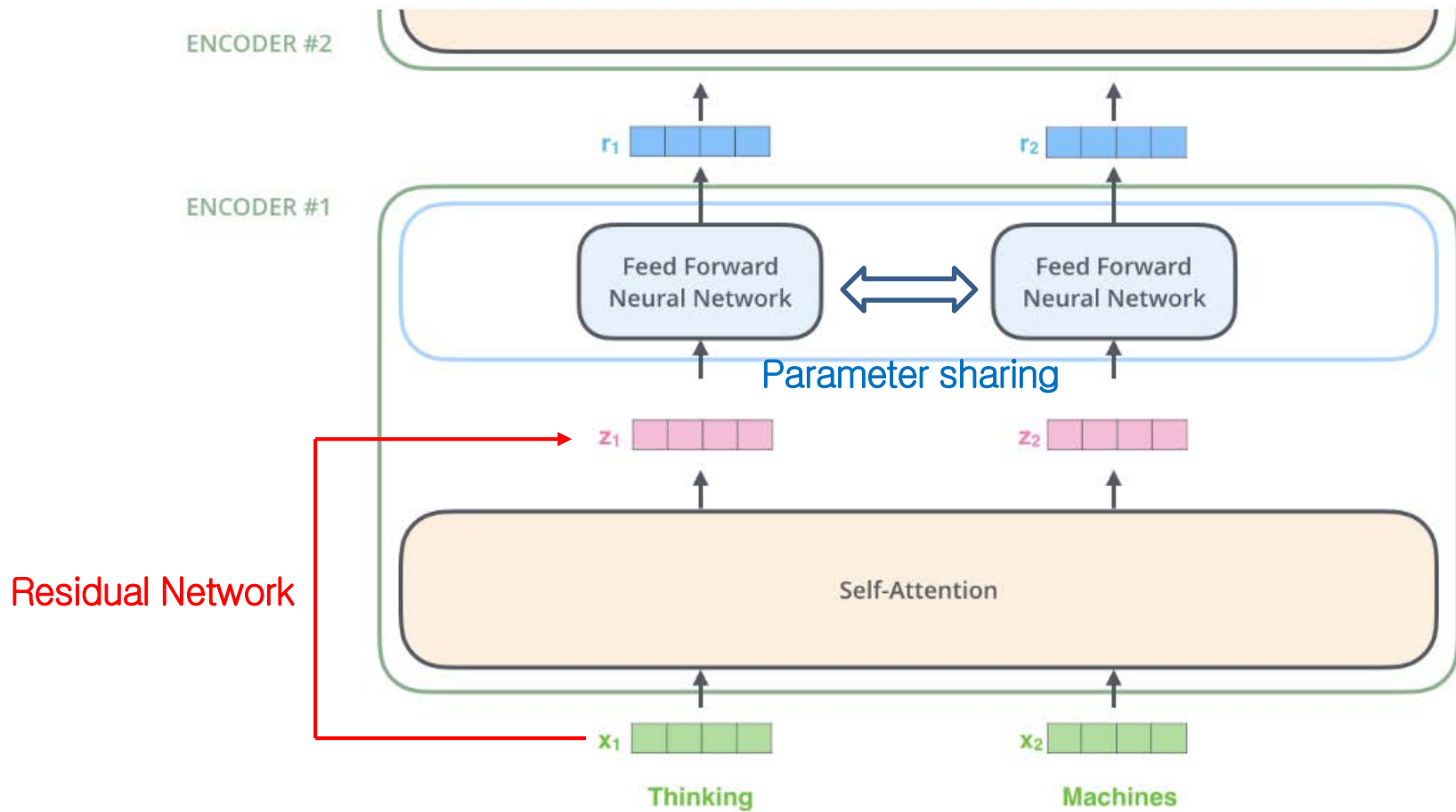


Figure from <http://jalammar.github.io/illustrated-transformer/>

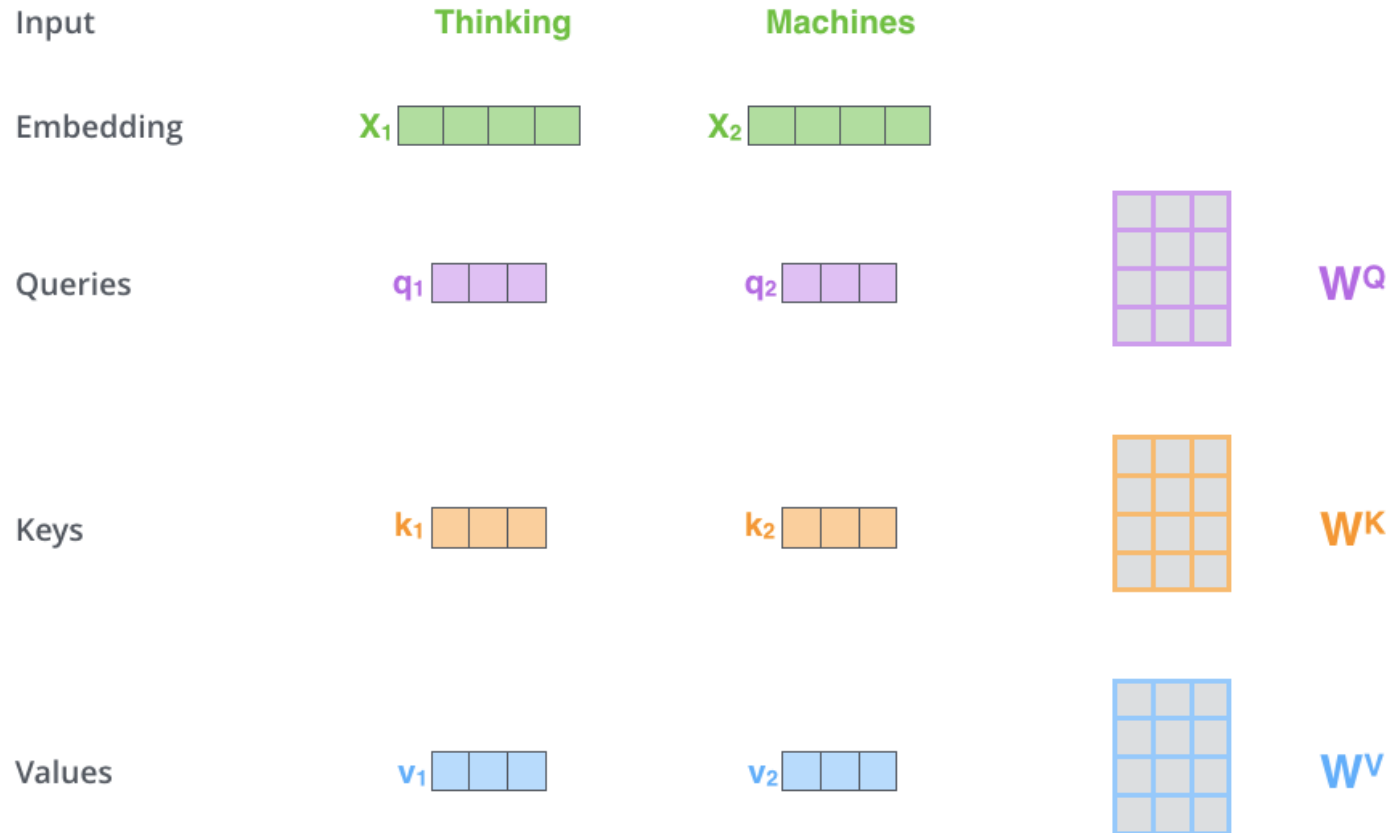
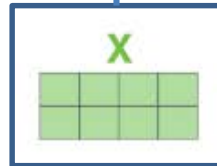
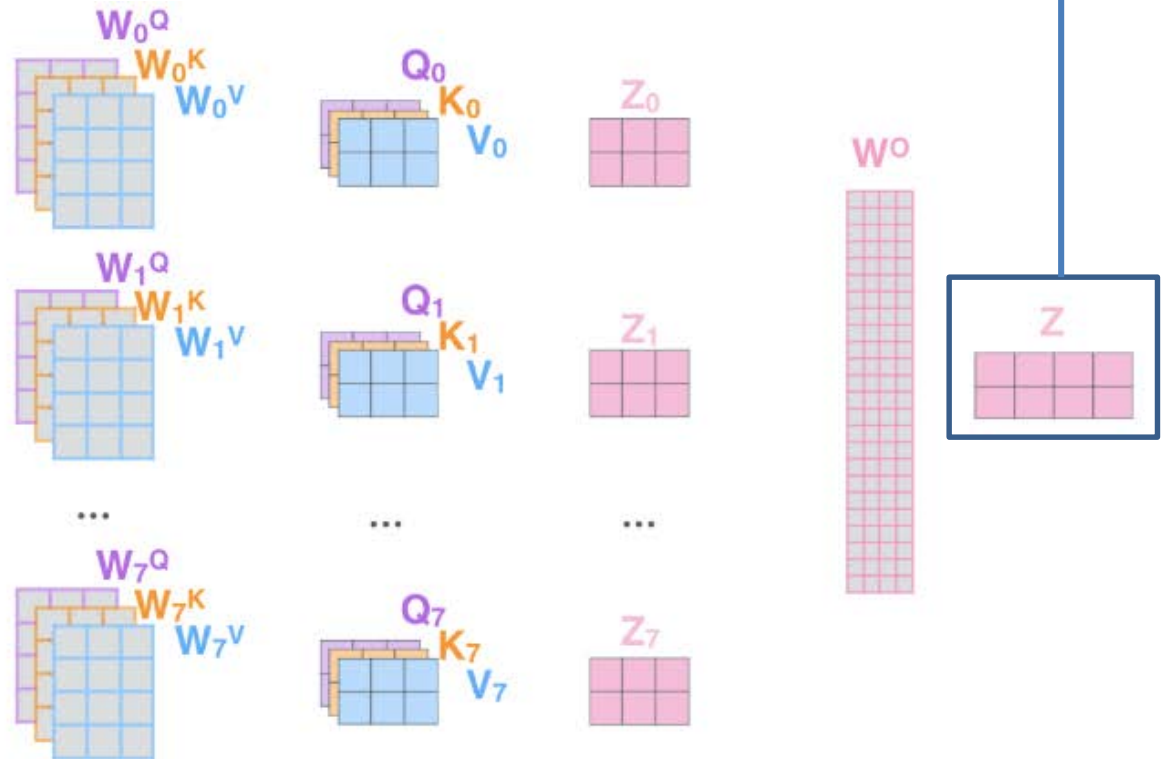


Figure from <http://jalammar.github.io/illustrated-transformer/>

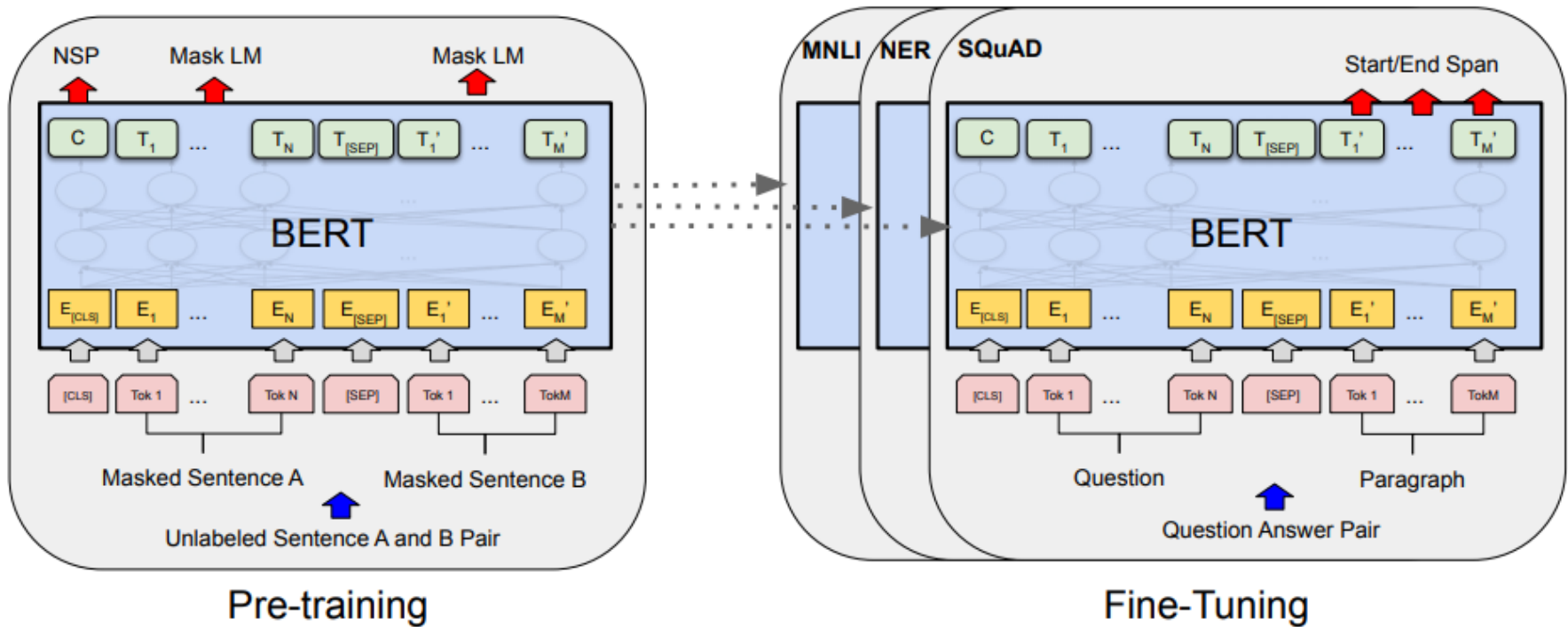
Same Dimension

Thinking
Machines

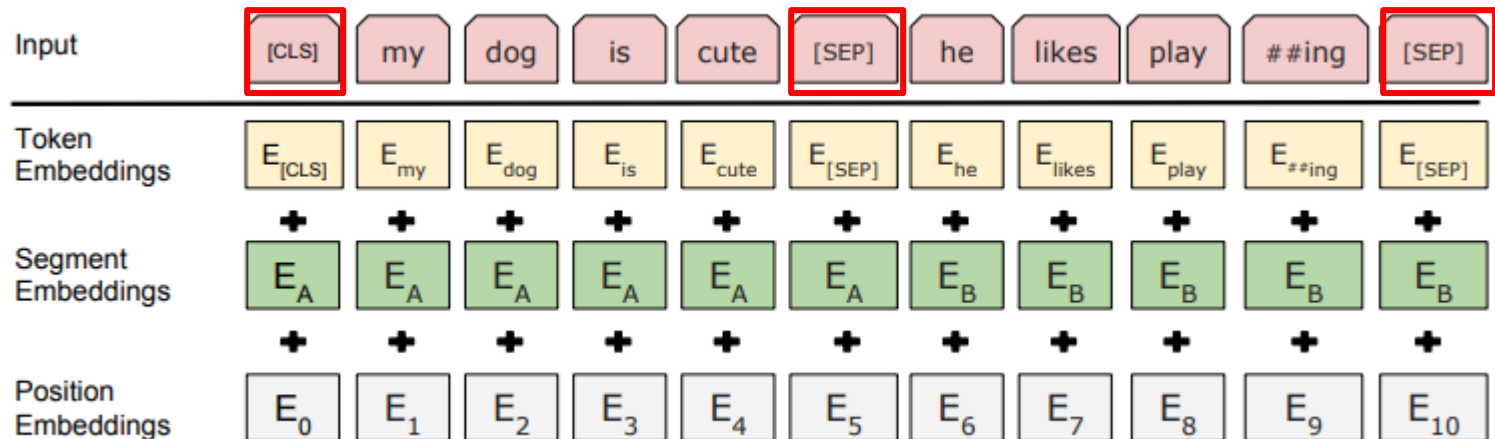
* In all encoders other than #0,
we don't need embedding.
We start directly with the output
of the encoder right below this one

Figure from <http://ialammar.github.io/illustrated-transformer/>

General Architecture



Input Representation



Token Embedding : WordPiece Embedding

Segment Embedding : Corresponding Sentence Embedding

Position Embedding : Learned Positional Embedding(up to 512 tokens)

Pre-training Task #1 : Masked Language Modeling

Masking some input tokens, and predict only those masked ones(Cloze)

- Masking 15% of WordPiece tokens in each sequence
- Rather than always replacing the token with [MASK], sometimes replace the given word with another word, and also sometimes keep the word unchanged
- This'll close the gap between pre-training phase and fine-tuning phase
- May converge slowly

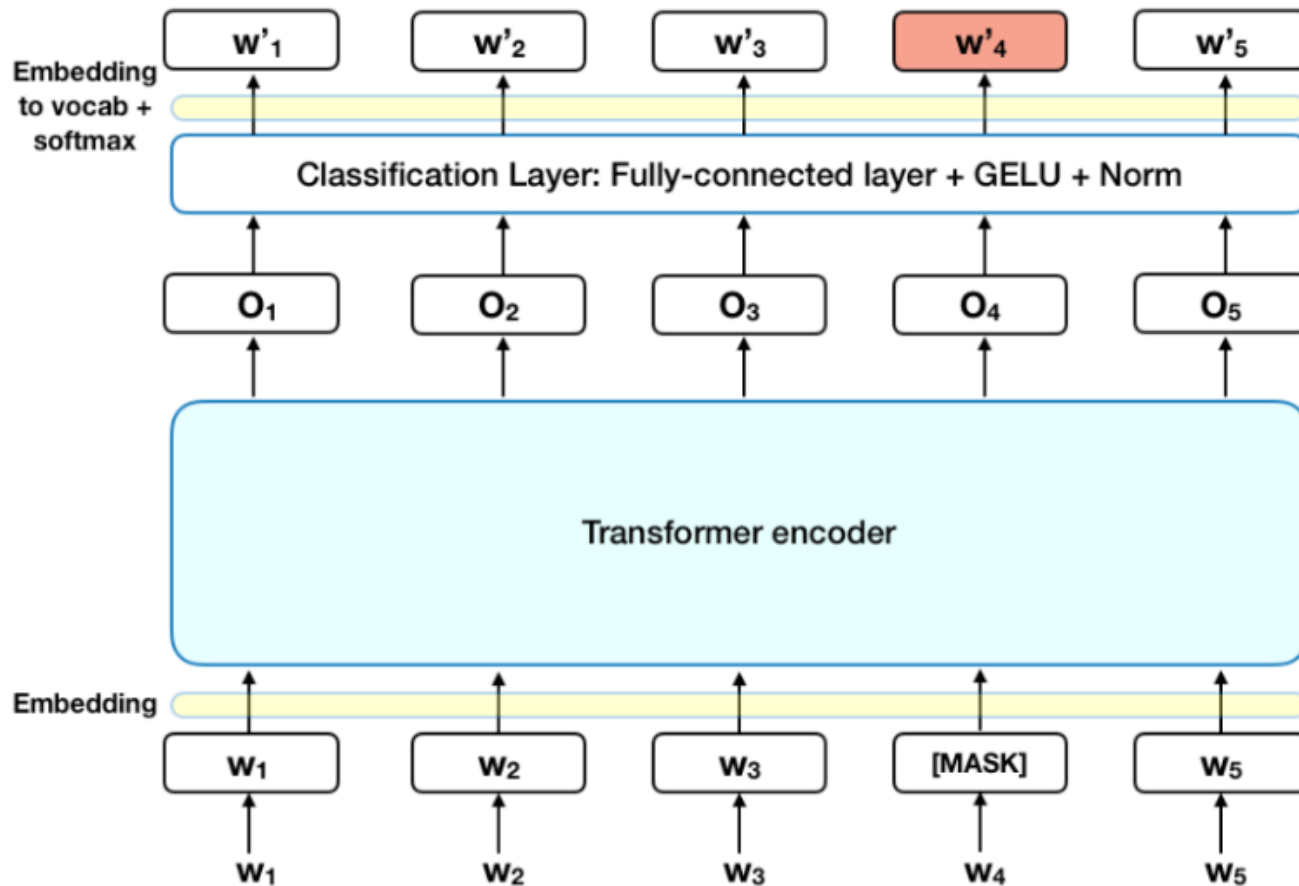


Figure from <https://www.lyrn.ai/2018/11/07/explained-bert-state-of-the-art-language-model-for-nlp/>

Pre-training Task #2 :Next Sentence Prediction

Understanding the relationship between two sentences

- Binarized next sentence prediction task
- In monolingual text, 50% of the time, next sentence is the actual sentence which follows the previous one, whereas for the 50% of the time, the random sentence from corpus is given.

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]

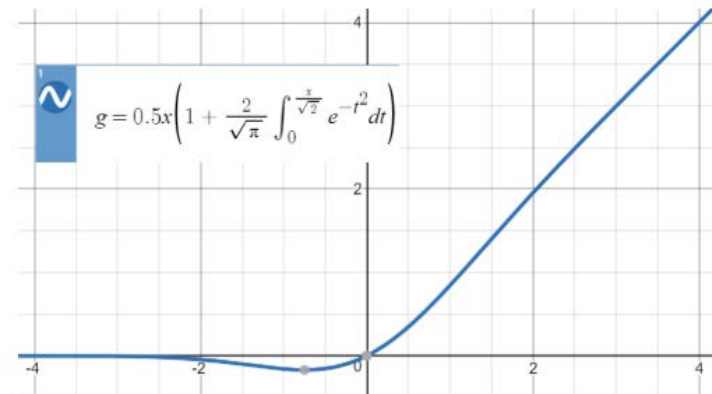
penguin [MASK] are flight ##less birds [SEP]

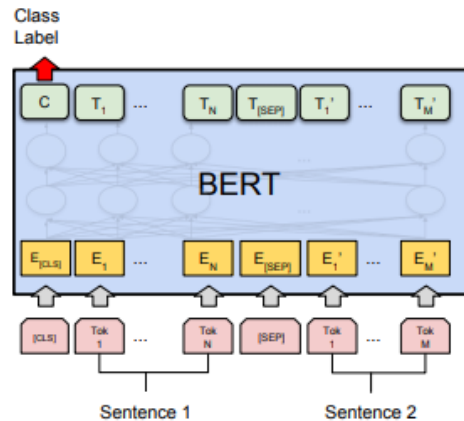
Label = NotNext

Pre-training Procedure

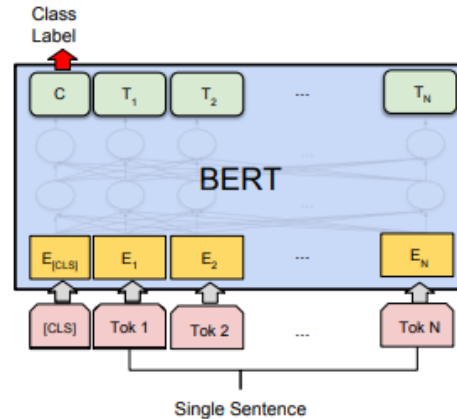
Understanding the relationship between two sentences

- Document Level Corpus was used for training(Book Corpus, Wikipedia)
- Sentences are generally longer than the usual sentences that we know
- Each sequence(combined of two sentences)'s length is smaller than 512
- GELU activation was used

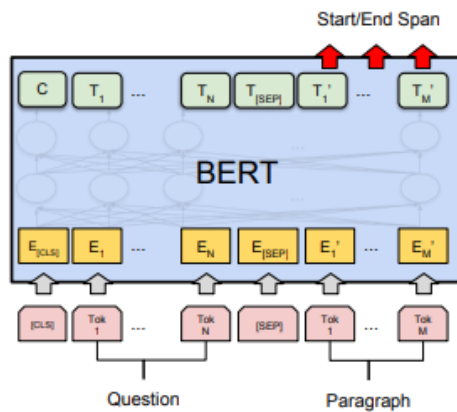




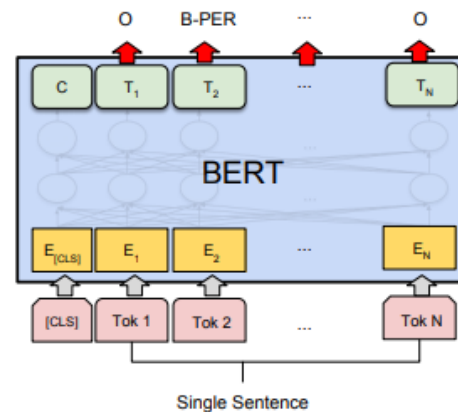
(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG



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



(c) Question Answering Tasks:
SQuAD v1.1



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

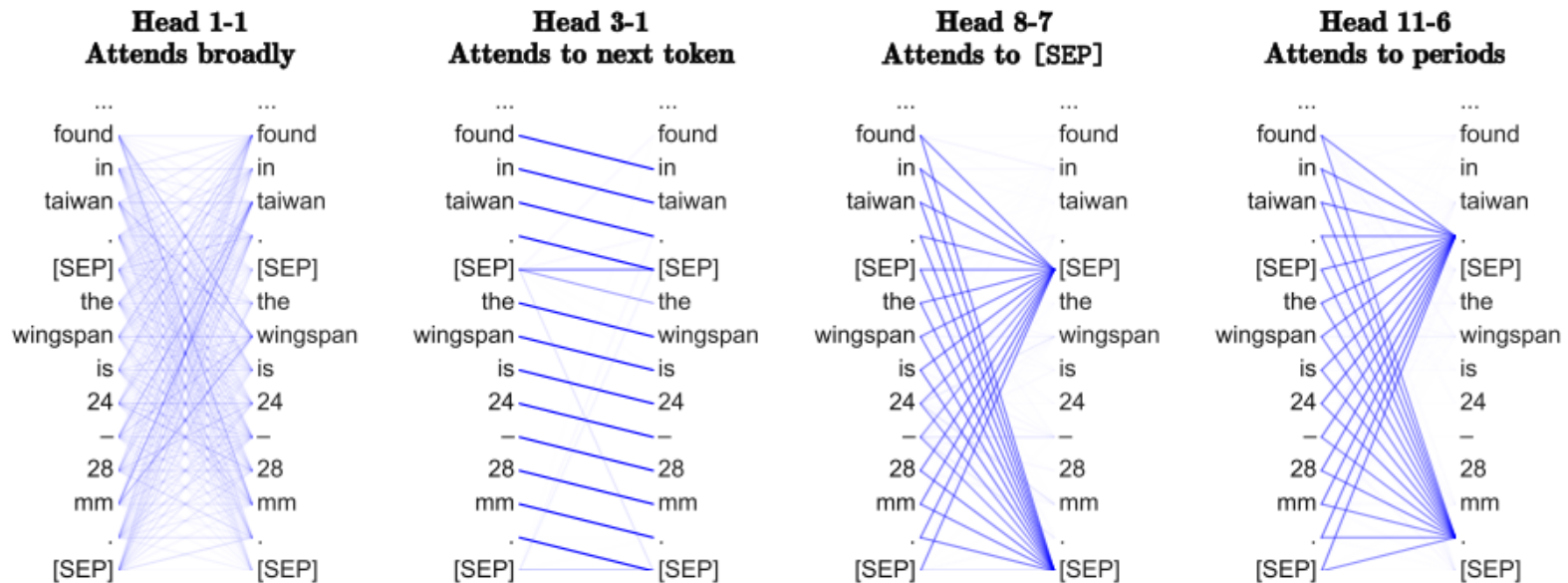
What Does BERT Look At?

An Analysis of BERT's Attention

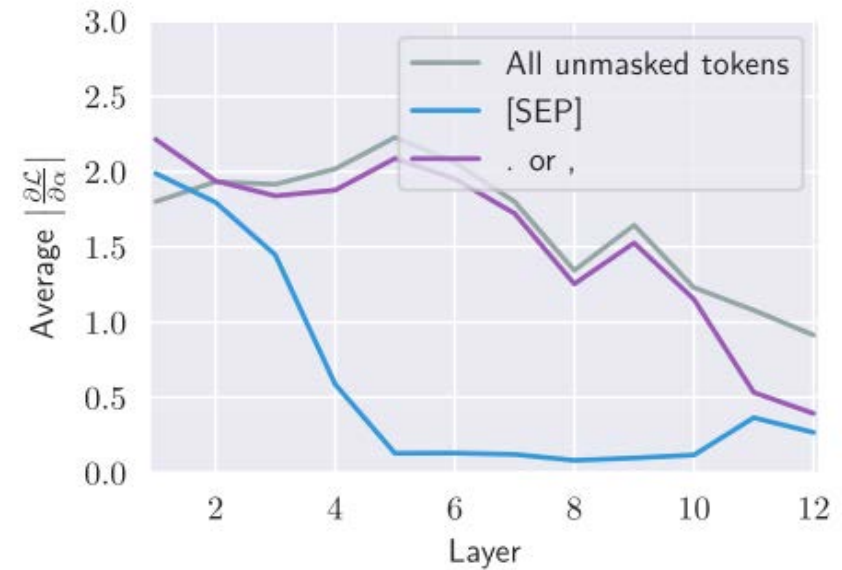
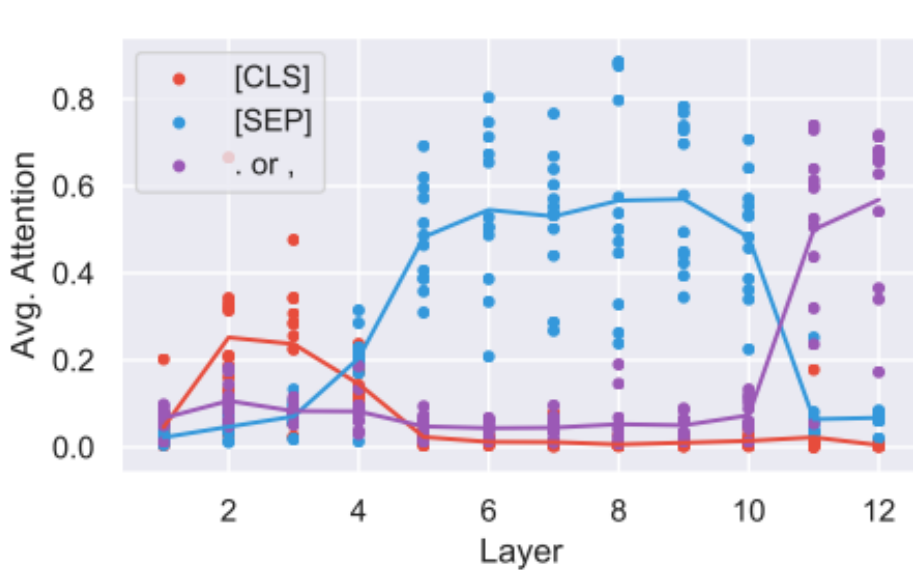
Kevin Clark et al.

Jungsoo Park

Data Mining & Information Systems Lab.
Department of Computer Science and Engineering,
College of Informatics, Korea University



- In earlier layers, attention heads tend to put their attention to previous, or next token of given token's position.
- In the higher layers, entropies of attention distribution gets smaller.
- Substantial amount of BERT's attention focuses on a few tokens ([SEP], [CLS], “,”, “.”)



- Rather than being an artifact of stochastic training, high attention would have been caused by systematic reason. (Because, special tokens don't get masked away)
- Another possible explanation is that [SEP] is used to aggregate segment level information, however doubtful. (Right Figure) => Sort of “no-op”