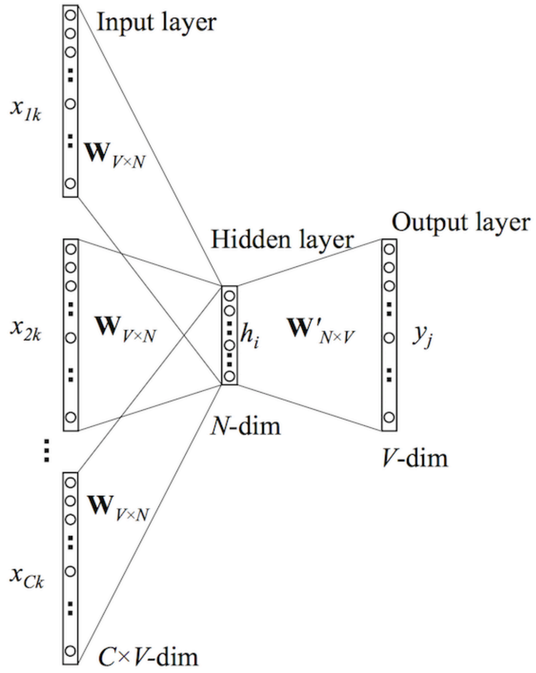
07-11–2024

Assignments/paper presentation

- Assignment 4 & 5
  - Assignment-4: VAE/GAN
  - Can we have paper presentation for assignment-5 ?
    - Based on the project group ?
    - 14 papers
    - 20 (15 + 5) mins/paper

Contextual embeddings

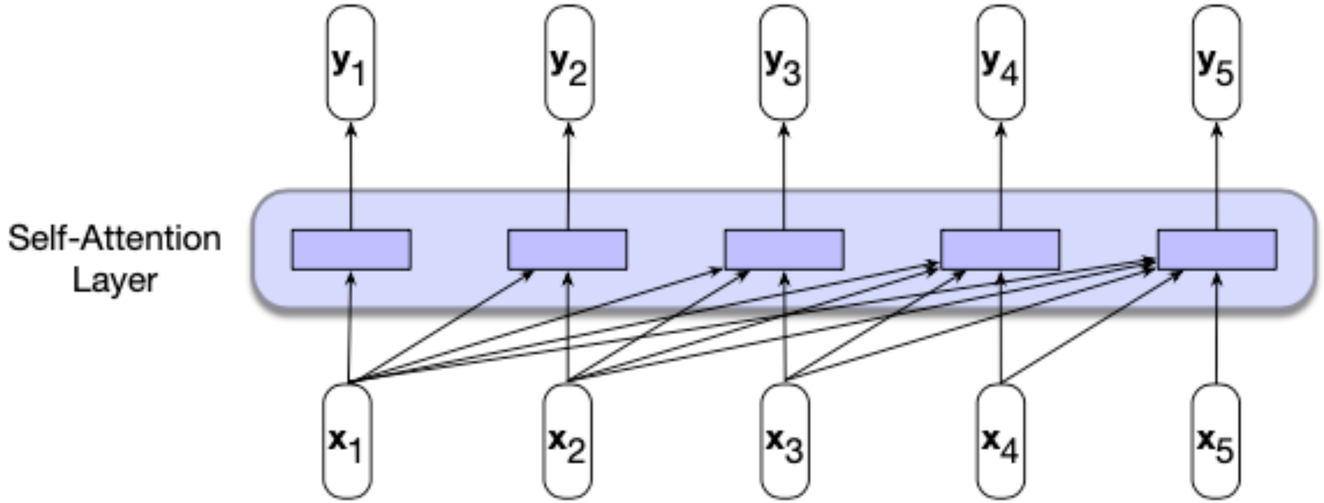
- Word2vec
  - A fixed representation for each unique word
  - But what about if a particular words is used in different context ?
    - Example?
      - a mouse controlling a computer system…
      - a quiet animal like mouse…
    - Can we represent word based on the uses in the context?
    - Multiple representation



Pretrained language model: transfer learning

- Pretraining: “the process of learning some sort of representation of meaning of words/ sentences by processing very large amount of text”
  - Big transformer like language model
- Fine-tuning: train a pertained model for a particular (NLP) task
- Transfer learning: an instance of a pretrained and fine-tuning process
  - Transfer knowledge from one task/domain to solve a new task

Bidirectional Transformer encoders



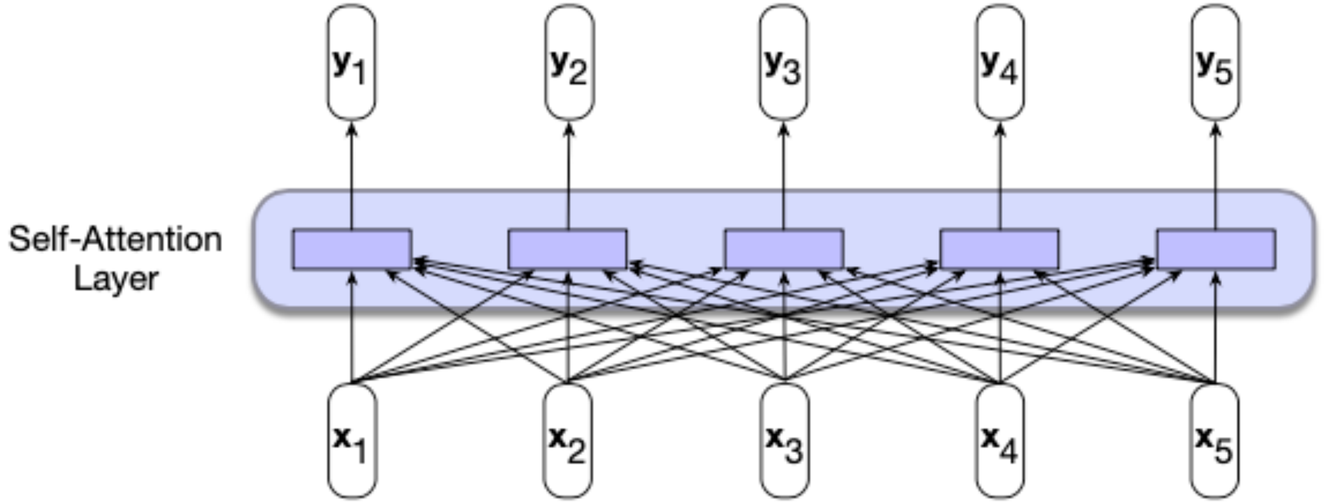


Image: Jurafsky & Martin “Speech and Language Processing, 3rd ed., 2023

Bidirectional self-Attention

- $X = [x_1; x_2; \cdots; x_N] \in \mathbf{R}^{N \times d}$, $W^Q \in \mathbf{R}^{d \times d}$,
- $W^K \in \mathbf{R}^{d \times d}$, $W^V \in \mathbf{R}^{d \times d}$
- $Y = \left[(XW^Q)(XW^K)^T\right](XW^V) \in \mathbf{R}^{N \times d}$
- $Y = softmax\left(\frac{QK^T}{d}\right)V$

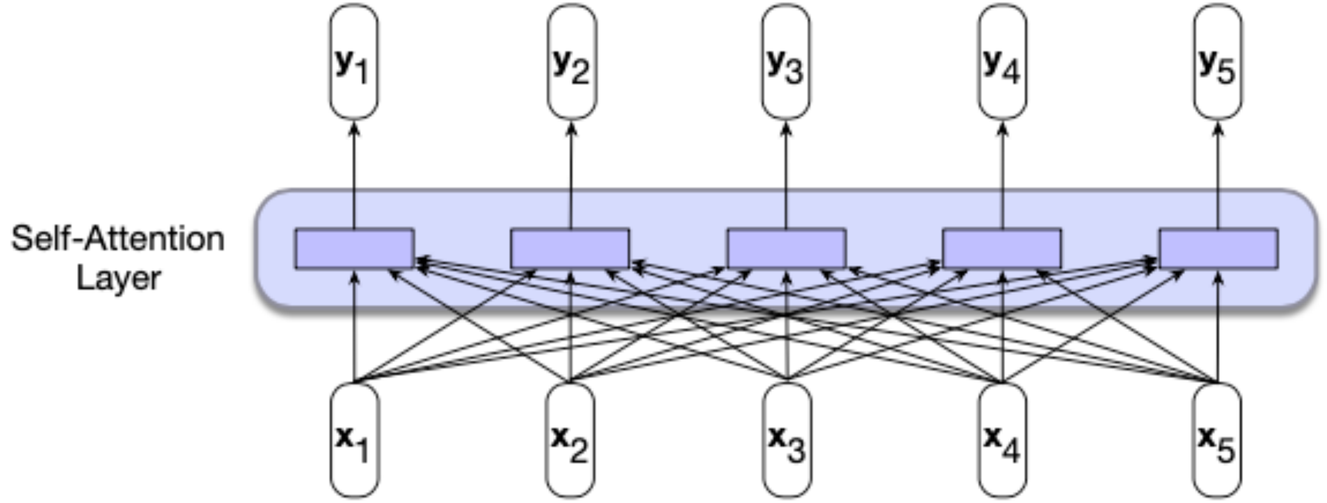


Image: Jurafsky & Martin "Speech and Language Processing, 3rd ed., 2023

N	q1•k1	−∞	−∞	−∞	−∞
	q2•k1	q2•k2	−∞	−∞	−∞
	q3•k1	q3•k2	q3•k3	−∞	−∞
	q4•k1	q4•k2	q4•k3	q4•k4	−∞
	q5•k1	q5•k2	q5•k3	q5•k4	q5•k5

N

N	q1•k1	q1•k2	q1•k3	q1•k4	q1•k5
	q2•k1	q2•k2	q2•k3	q2•k4	q2•k5
	q3•k1	q3•k2	q3•k3	q3•k4	q3•k5
	q4•k1	q4•k2	q4•k3	q4•k4	q4•k5
	q5•k1	q5•k2	q5•k3	q5•k4	q5•k5

N

Transformer block

- Layer normalisation (LN):
  - $LN_1[x + self - attention(x)]$
  - $LN_2[x + FFNN(x)]$
  - $FFNN(x) = max[0, XW_1 + b_1]W_2 + b_2$
  - $LN(\hat{x}) = \gamma\hat{x} + \beta$
    - $\hat{x} = \frac{x - \mu}{\sigma}, \mu = \frac{1}{d}\sum_{i=1}^{d} x_i; x \in \mathbf{R}^d$
    - $\sigma = \sqrt{\frac{1}{d}\sum_{i=1}^{d}(x_i - \mu)^2}$

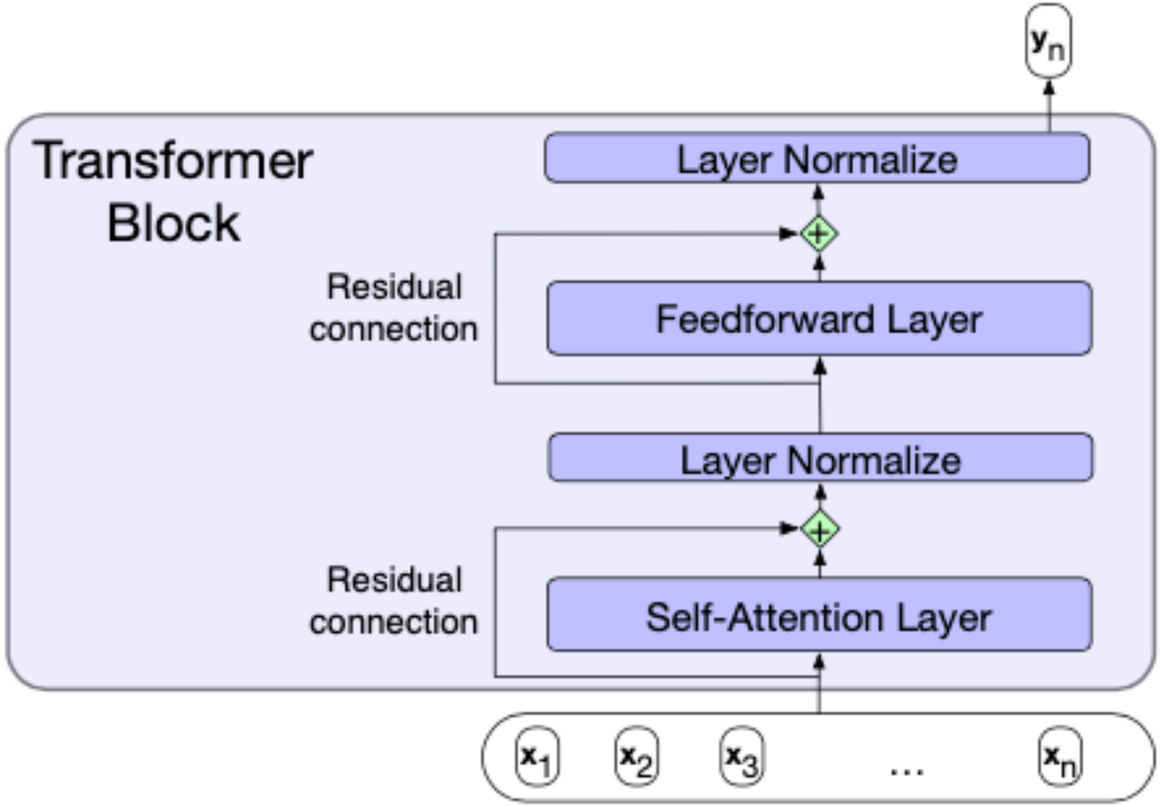


Image: Jurafsky & Martin "Speech and Language Processing, 3rd ed., 2023

Bidirectional Encoder Representations from Transformers (BERT)

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

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

Go to page 11

We argue that current techniques restrict the power of the pre-trained representations, especially for the fine-tuning approaches. The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in OpenAI GPT, the authors use a left-to-right architecture, where every token can only attend to previous tokens in the self-attention layers

BERT

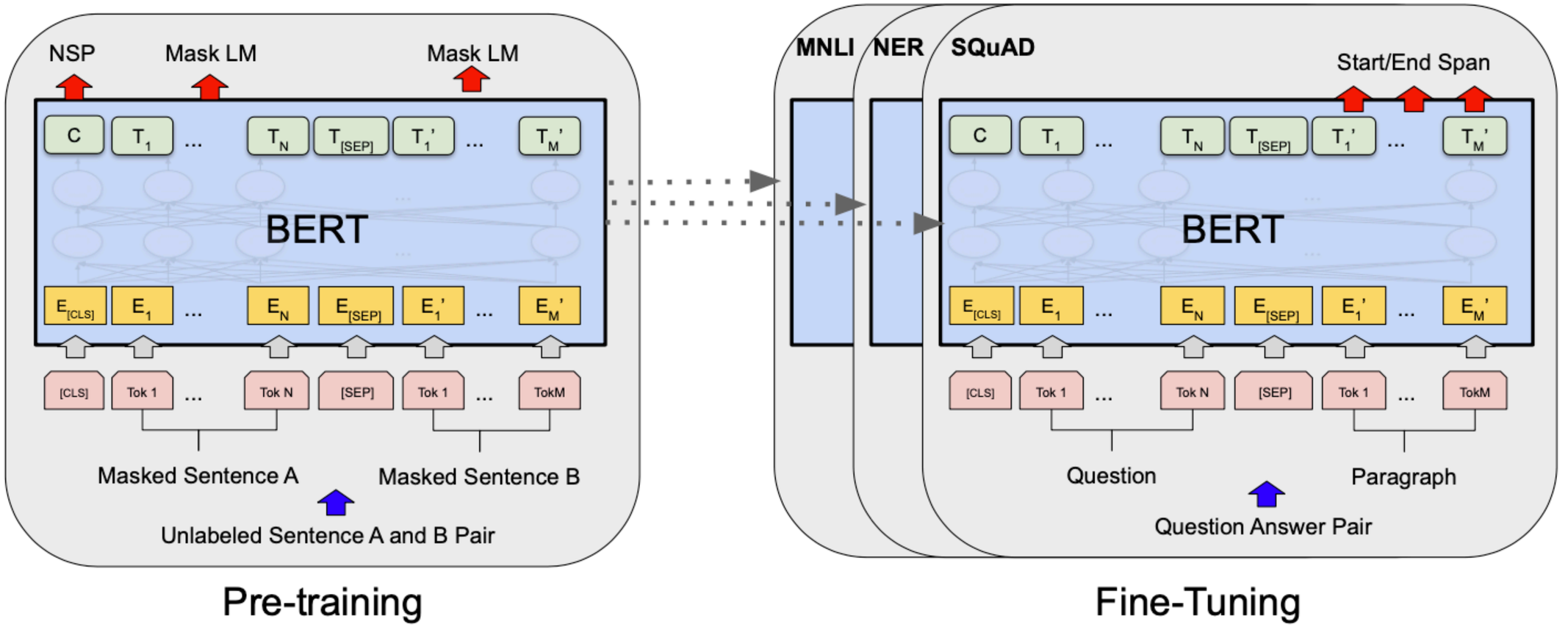


Image: Jurafsky & Martin “Speech and Language Processing, 3rd ed., 2023

BERT (cont.)

- Model in brief:
  - Subword tokenisation (WordPiece algo. by Schuster and Nakajima, ICASSP, 2012
    - 30,000 tokens
  - Size of hidden layers: 768
  - # Transformer blocks: 12
    - #multihead attention layers: 12

Transformer original:
#tokens: 25,000
Size of hidden layers: 2048
#transformer blocks: 6(E) and 6(D)
#multihead attention layers: 8

Training bidirectional transformer encoders

- Masked language model (MLM)
  - Masking words
    - Sample 15% of tokens in an input sequence for learning
      - 80% filled with a special token [MASK]
      - 10% randomly selected token from outside
      - 10% unchanged
  - Can you see what the model will learn ?
    - Contextual embedding?

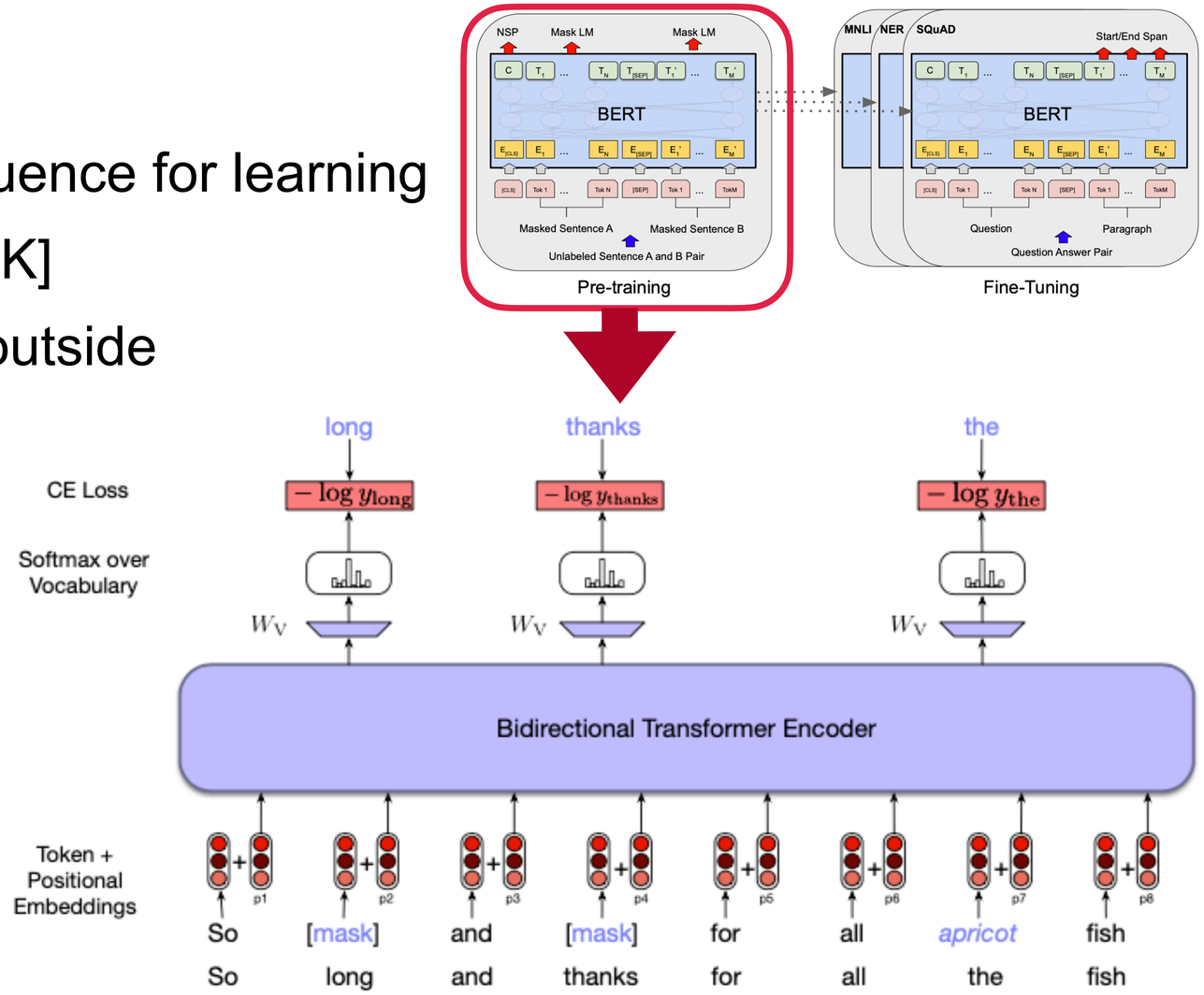


Image: Jurafsky & Martin "Speech and Language Processing, 3rd ed., 2023

Training bidirectional transformer encoders (cont.)

- Masked language model
  - Next sentence prediction (NSP)
    - Pairs of sentences and two special tokens [CLS] and [SEP]
    - Learning goal: positive vs negative pairs

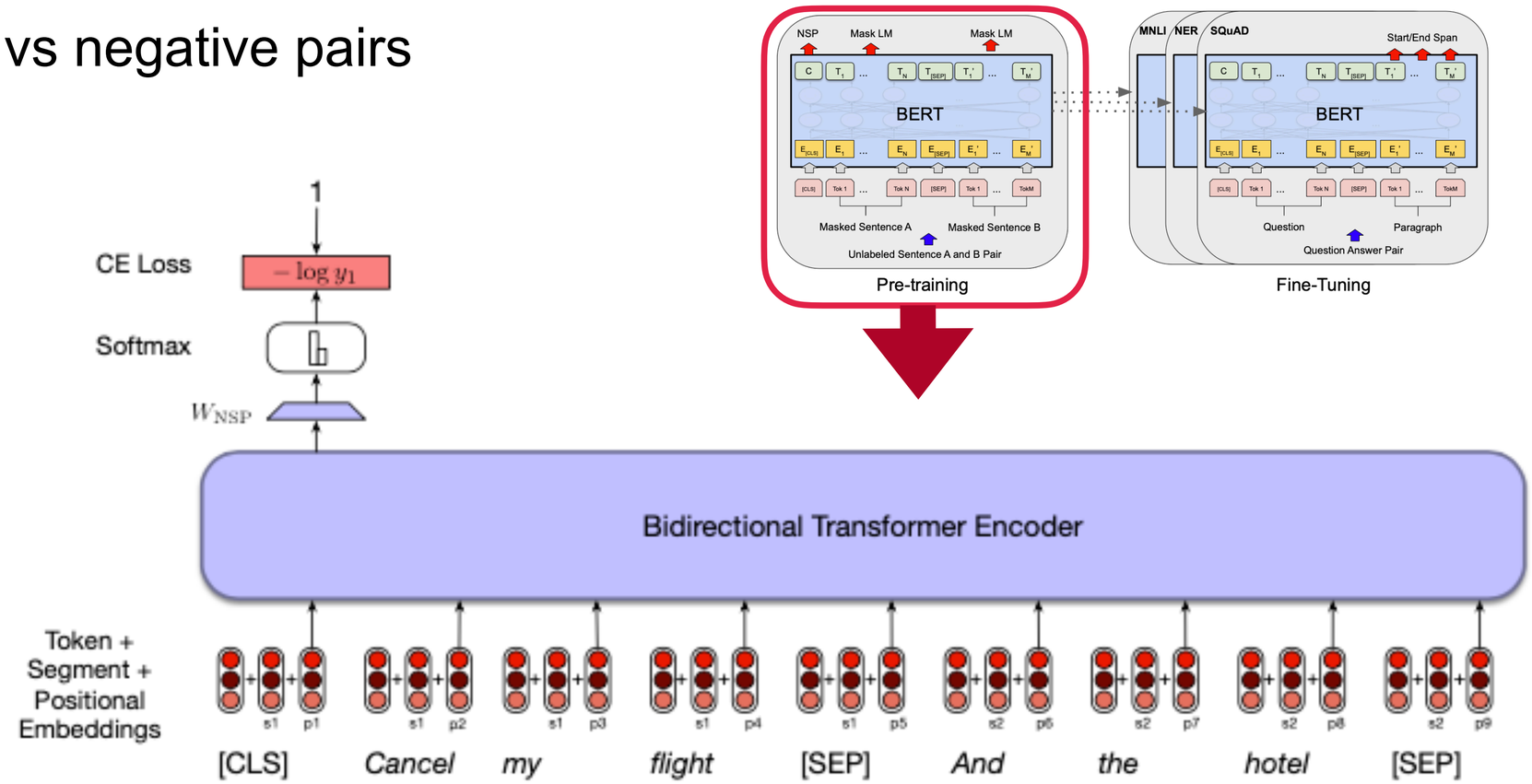
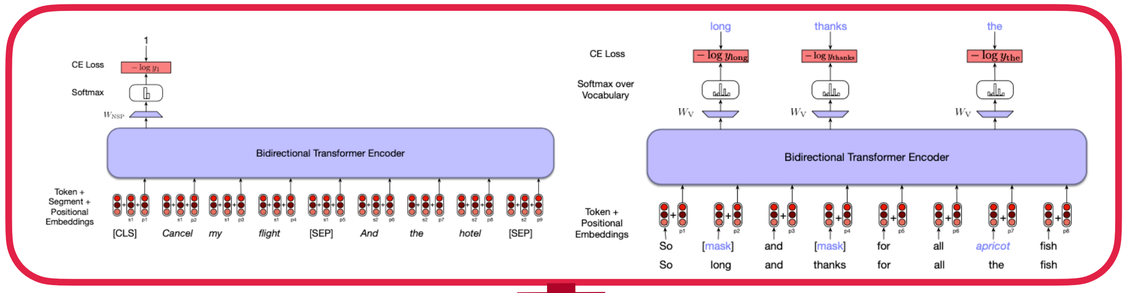


Image: Jurafsky & Martin "Speech and Language Processing, 3rd ed., 2023

BERT: Pre-training



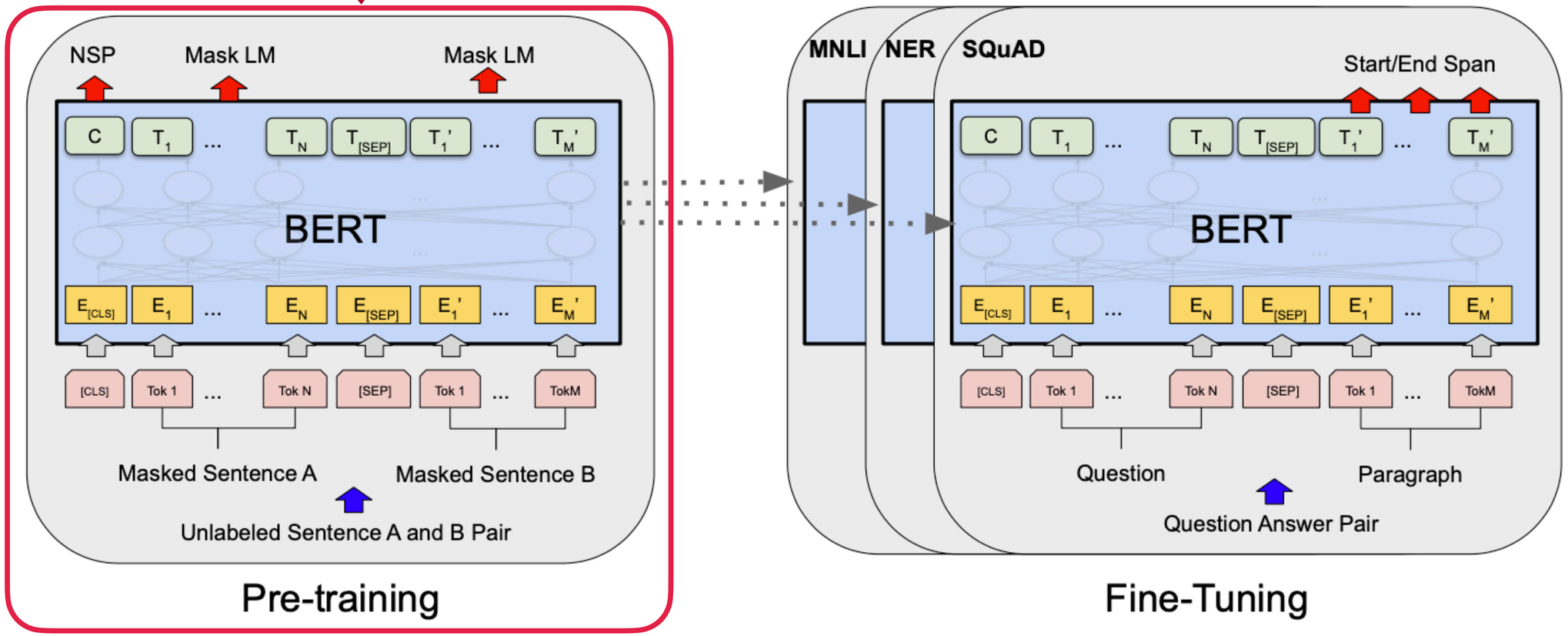
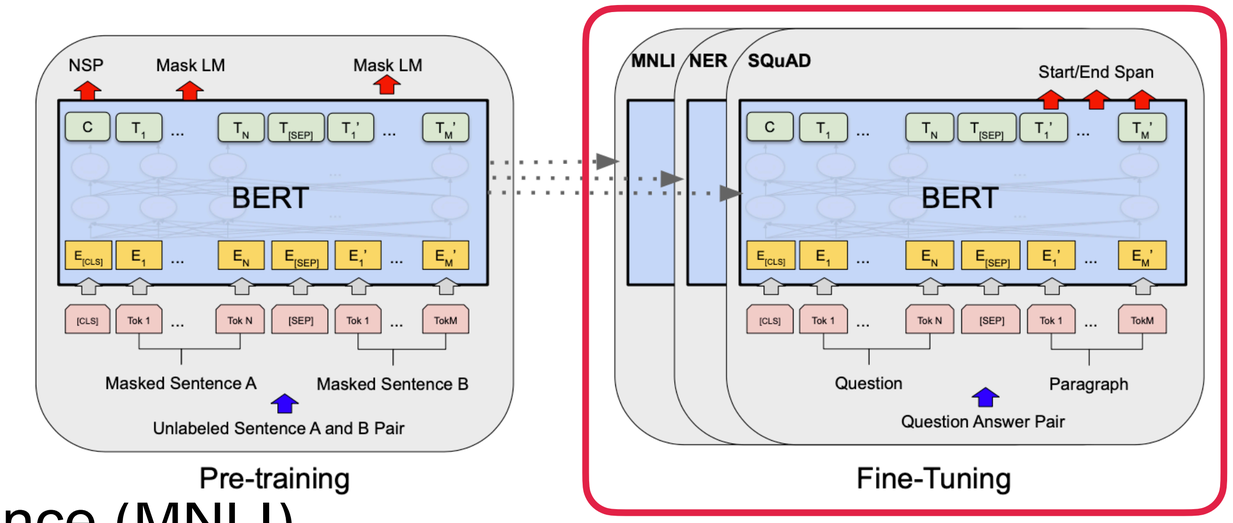


Image: Jurafsky & Martin "Speech and Language Processing, 3rd ed., 2023

BERT: Fine-tuning

- Tasks
  - Sequence labelling
    - NER, POS
  - Sequence classification
    - Sentiment
  - Multi-Genre Natural Language Inference (MNLI)
    - Given two sentences S1 & S2
    - task is to predict S2 is an entailment, contradiction or neutral w.r.t S1
  - Question answering (SQuAD- Stanford Question Answering) Dataset)
  - ….



BERT: Fine-tuning (cont.)

- Sequence labelling:
  - Parts-of-speech tagging

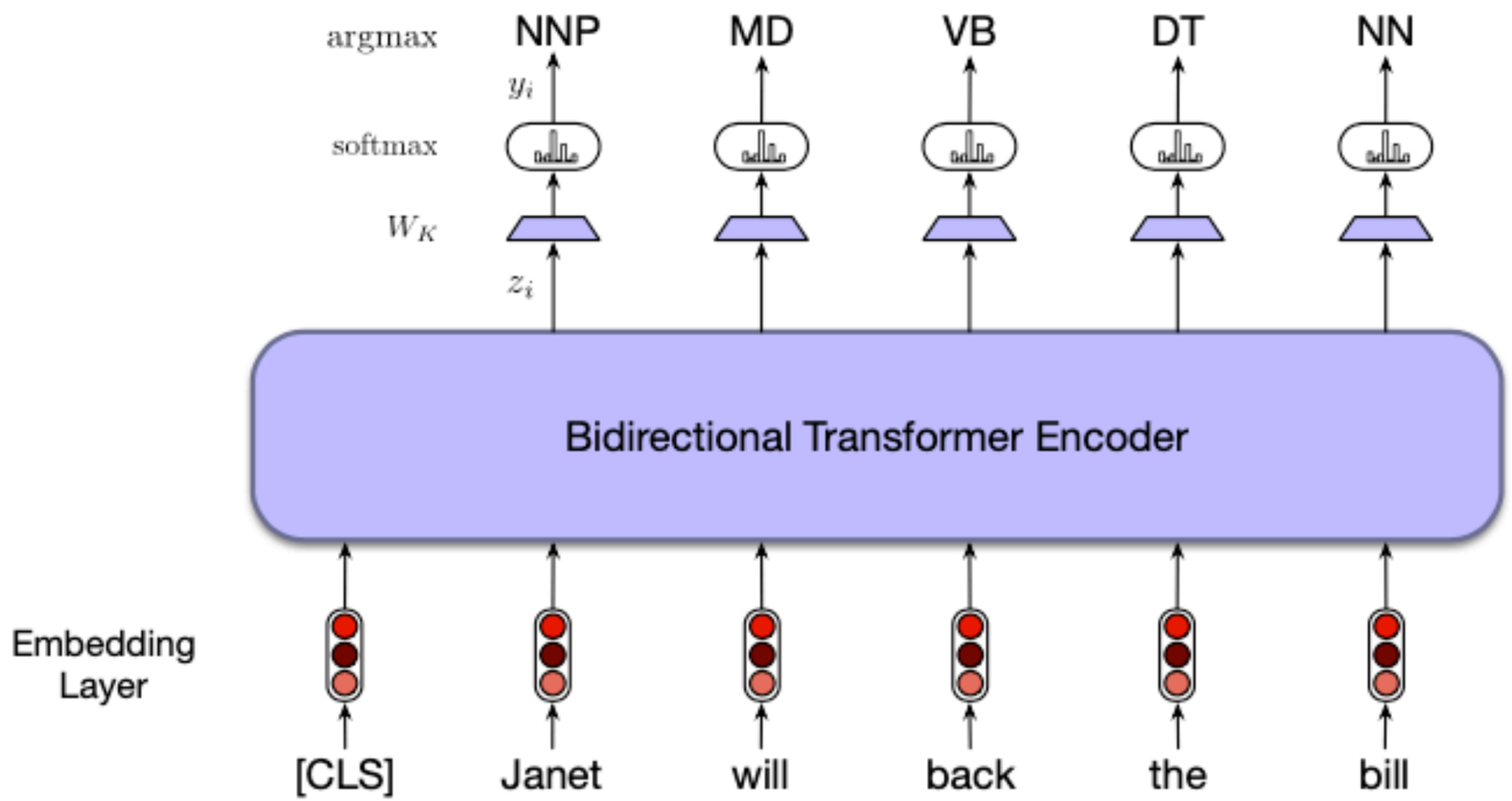


Image: Jurafsky & Martin "Speech and Language Processing, 3rd ed., 2023

BERT: Fine-tuning (cont.)

- Sequence classification
  - Sentiment classification
    - Positive
    - Negative
    - Neutral

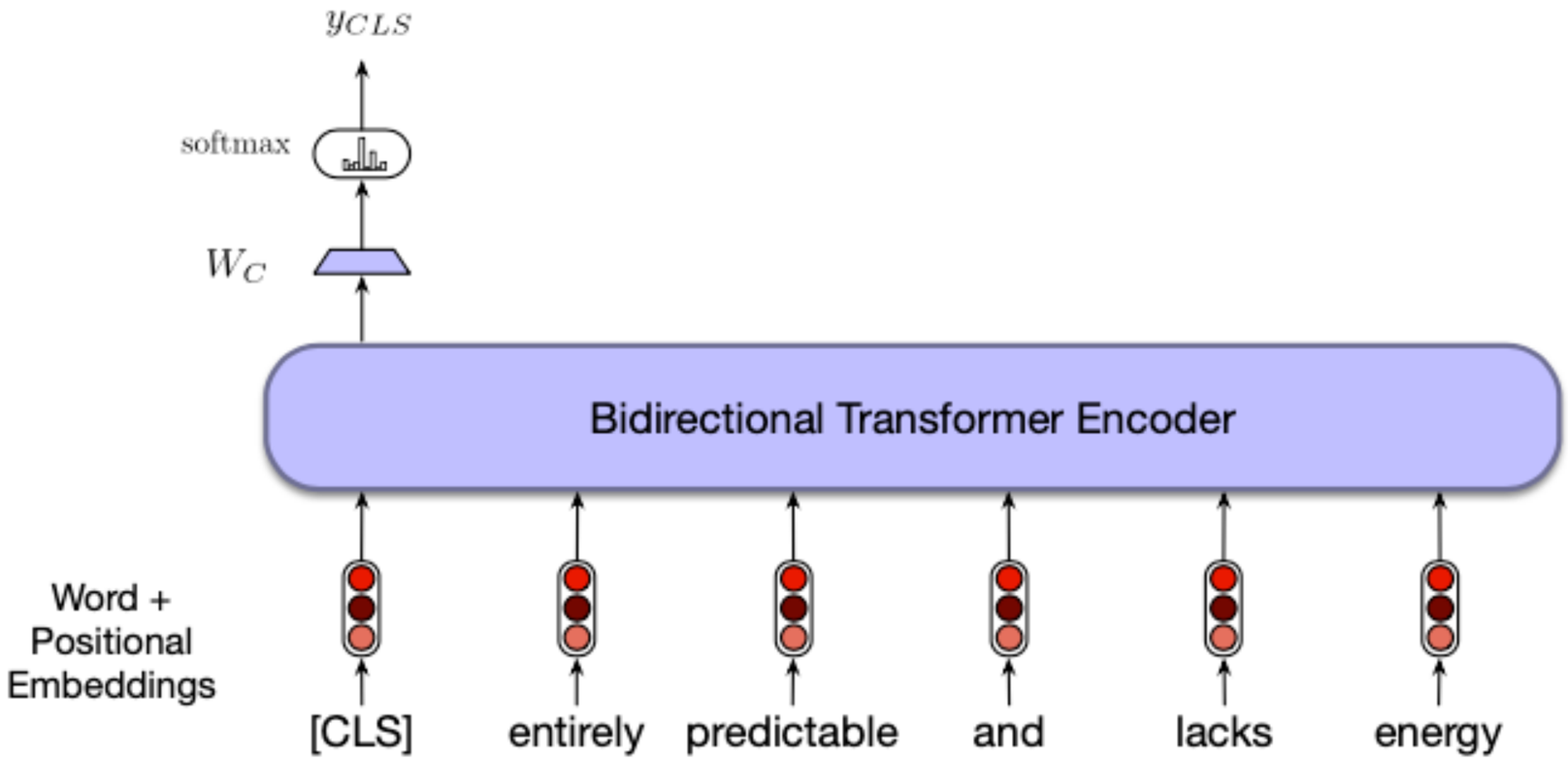


Image: Jurafsky & Martin "Speech and Language Processing, 3rd ed., 2023

Different variants of Transformers

- Span based language model (SpanBERT, Joshi et al., TACL 2020)
  - Masking span
    - Sample a span from input sequence for learning
    - All the token of that span substituted by:
      - 80% filled with a special token [MASK]
      - 10% randomly selected token from outside
      - 10% unchanged

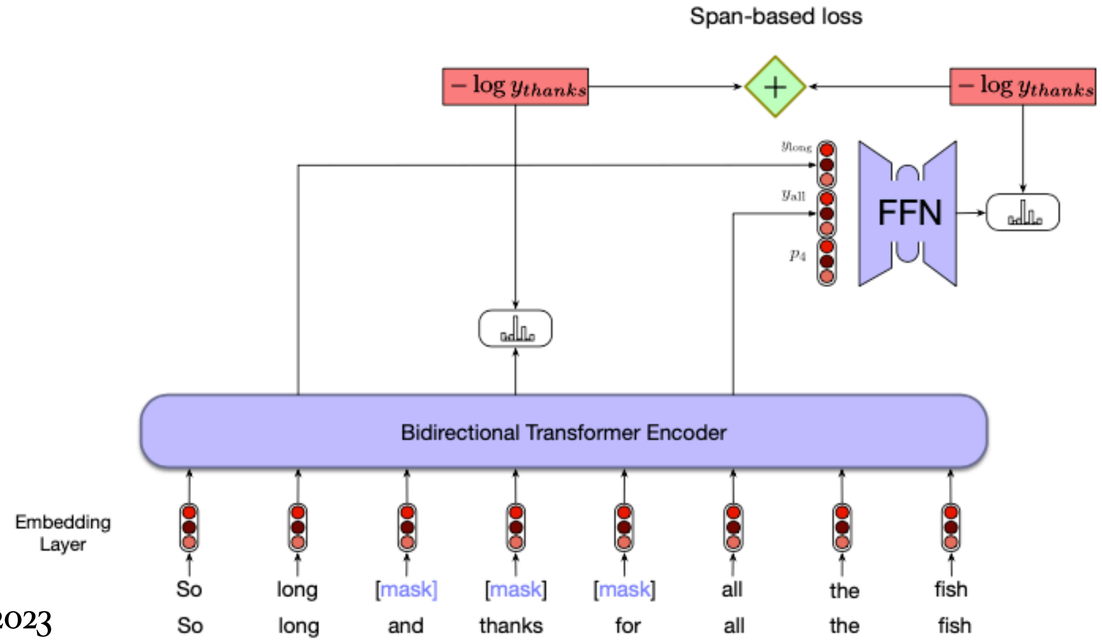


Image: Jurafsky & Martin "Speech and Language Processing, 3rd ed., 2023

Different variants of Transformers (cont.)

- #parameters

✹ GPT-3 (175 B), OpenAI

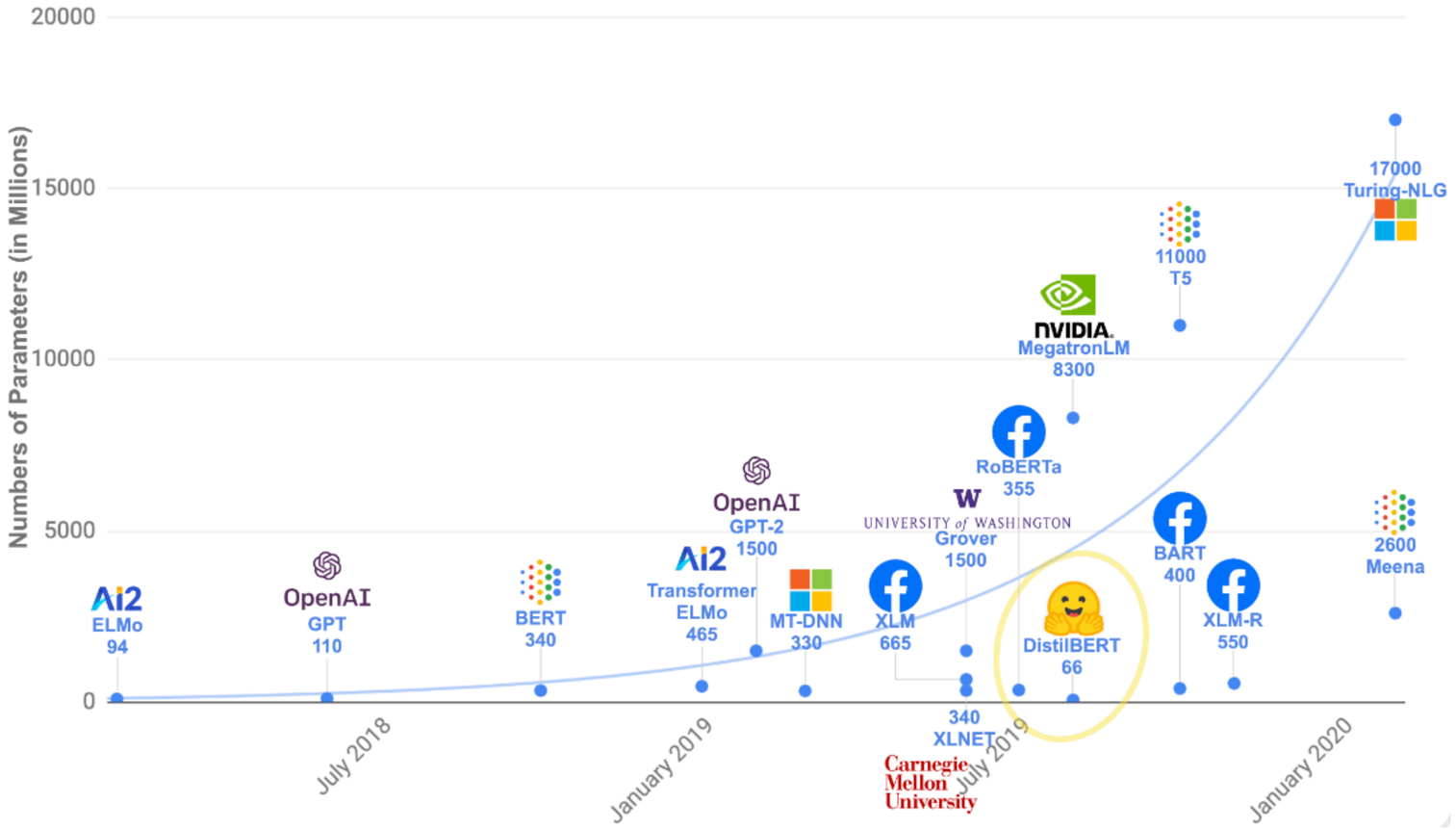


Image: https://huggingface.co/course/chapter1/4