Fine-tuning and Masked Language Models

Phạm Quang Nhật Minh

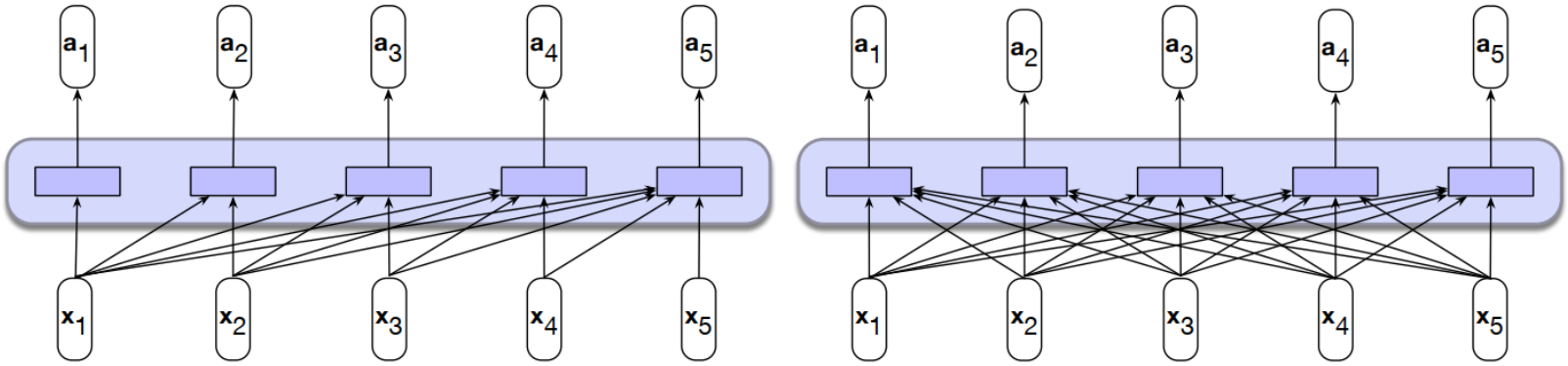
minhpham0902@gmail.com

Outline

- Research context
- Bidirectional Transformers Encoder (BERT)
- Training Bidirectional Encoders
- Fine-tuning language models

Limitation of left-to-right Transformers

- Left-to-right architecture restricts the power of pre-trained representations
- In some tasks, such as named entity recognition, we want to use information of both left and right contexts



a) A causal self-attention layer

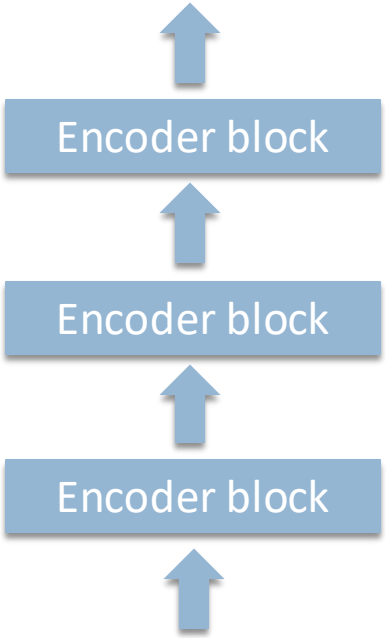
b) A bidirectional self-attention layer

Outline

- Research context
- Bidirectional Transformers Encoder (BERT)
- Training Bidirectional Encoders
- Fine-tuning language models

The architecture of bidirectional models

- BERT's model architecture is a multi-layer bidirectional Transformer encoder
- Transformer encoder uses the same self-attention mechanism as causal models, but does not mask the future



Self-Attention

- Use matrices W^Q, W^K, and W^V to project input x_i into query, key, value vectors

$$q_i = W^Q x_i \qquad k_i = W^K x_i \qquad v_i = W^V x_i$$

- Output vector y_i is the weighted sum of all the input value vectors

$$y_i = \sum_{j=1}^{n} \alpha_{ij} v_j$$

- Attention weights

$$\alpha_{ij} = \frac{\exp(\text{score}_{ij})}{\sum_{k=1}^{n} \exp(\text{score}_{ik})}$$

$$\text{score}_{ij} = \frac{q_i \cdot k_j}{\sqrt{d_k}}$$

Transformer Block

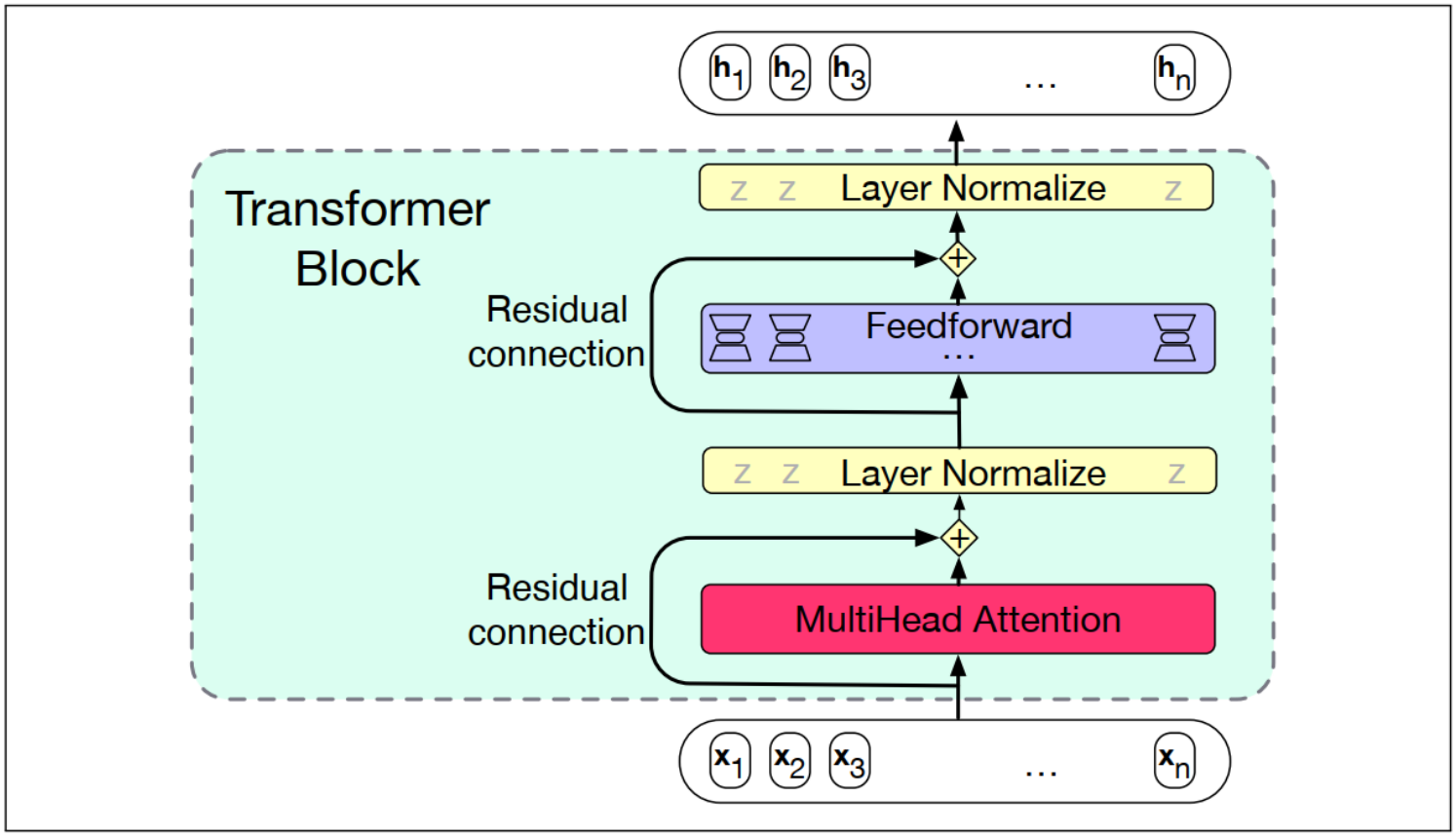


Figure 11.3 A transformer block showing all the layers.

Model Architecture

- In the original BERT paper, two models with different sizes were investigated
  - $BERT_{BASE}$: L=12, H=768, A=12, Total Parameters=110M
    - (L: number of layers (Transformer blocks), H is the hidden size, A: the number of self-attention heads)
  - $BERT_{LARGE}$: L=24, H=1024, A=16, Total Parameters=340M

Outline

- Research context
- Bidirectional Transformers Encoder (BERT)
- Training Bidirectional Encoders
- Fine-tuning language models

Two pre-training tasks

- Task#1: Masked Language Models
  - Instead of predicting next words as causal transformers
    - Please turn your homework __ .
  - Predict a missing item given the rest of the sentence
    - Please turn __ homework in .
- Task#2: Next Sentence Prediction
  - Details in next sections

Task#1: Masked Language Model (MLM)

- 15% of the words are masked at random
  - and the task is to predict the masked words based on its left and right context
- Not all tokens were masked in the same way (example sentence “My dog is hairy”)
  - 80% were replaced by the <MASK> token: “My dog is <MASK>”
  - 10% were replaced by a random token: “My dog is apple”
  - 10% were left intact: “My dog is hairy”

MLM Training

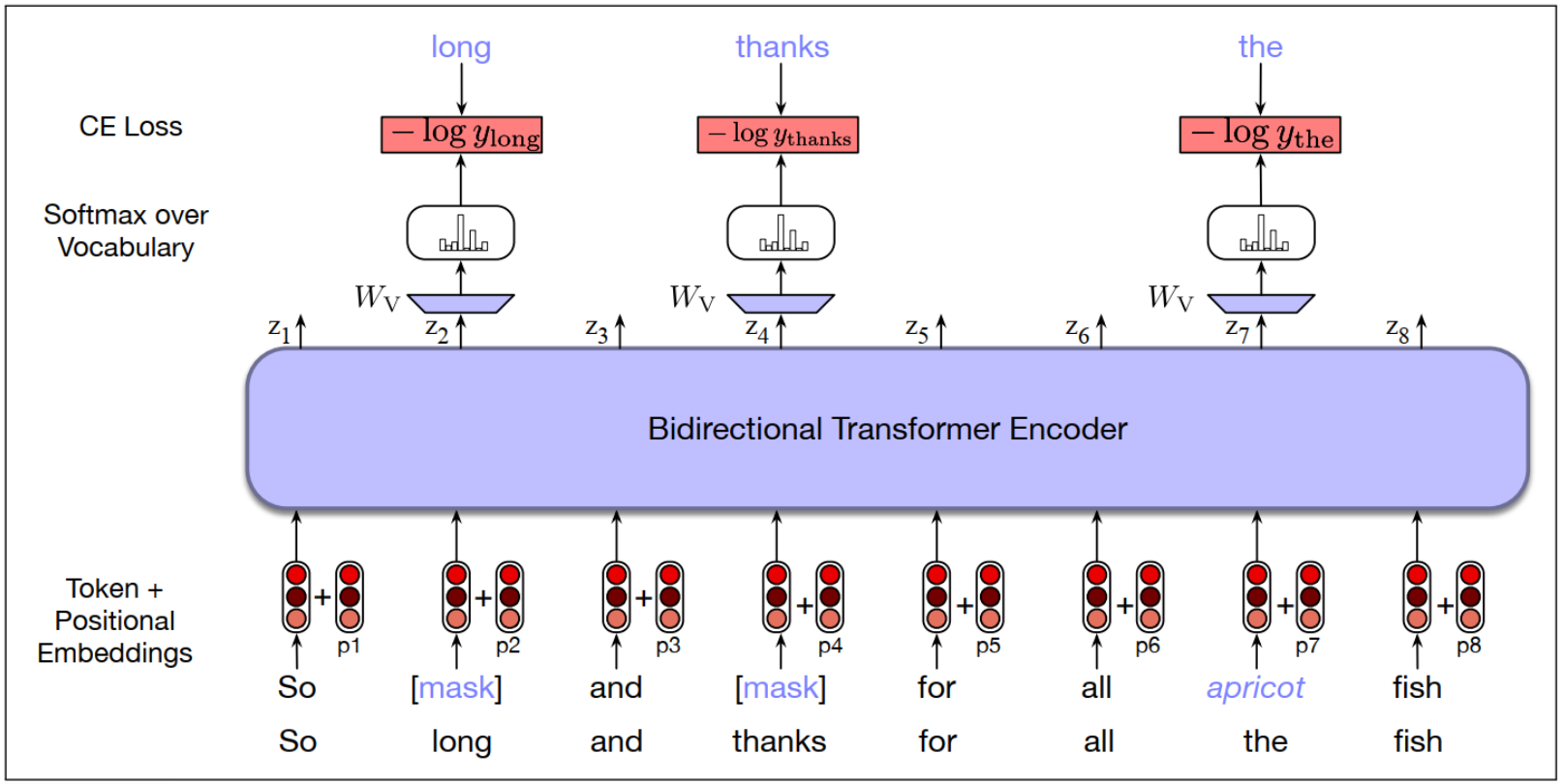


Figure 11.4 Masked language model training. In this example, three of the input tokens are selected, two of which are masked and the third is replaced with an unrelated word. The probabilities assigned by the model to these three items are used as the training loss. The other 5 words don't play a role in training loss. (In this and subsequent figures we display the input as words rather than subword tokens; the reader should keep in mind that BERT and similar models actually use subword tokens instead.)

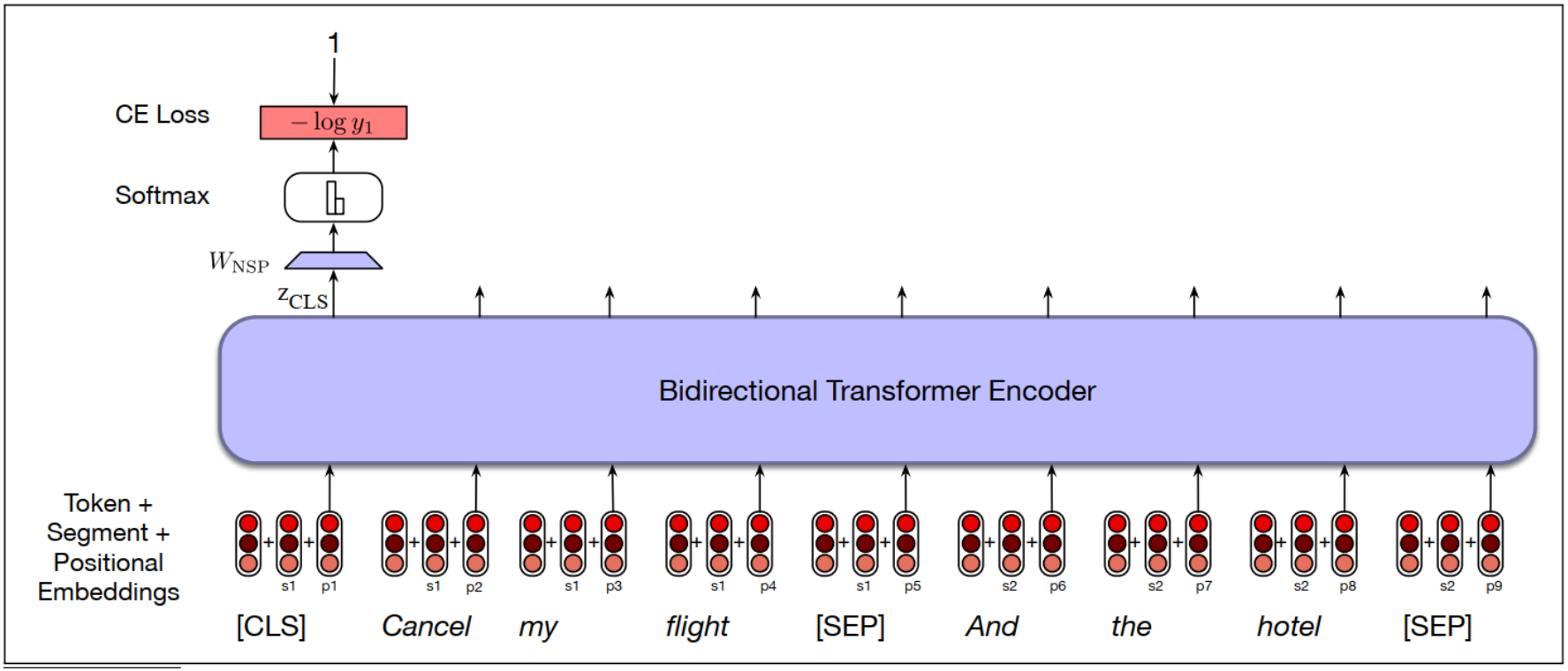
Task#2: Next Sentence Prediction (NSP)

- Motivation
  - Many downstream tasks are based on understanding the relationship between two text sentences
    - Question Answering (QA) and Natural Language Inference (NLI)
  - Language modeling does not directly capture that relationship
- The task is pre-training binarized next sentence prediction task

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

NSP Training



Pre-training procedure

- Training data: BooksCorpus (800M words) + English Wikipedia (2,500M words)
- To generate each training input sequences: sample two spans of text (A and B) from the corpus
  - The combined length is ≤ 512 tokens
  - 50% B is the actual next sentence that follows A and 50% of the time it is a random sentence from the corpus
- The training loss is the sum of the mean masked LM likelihood and the mean next sentence prediction likelihood

Outline

- Research context
- Bidirectional Transformers Encoder (BERT)
- Training Bidirectional Encoders
- Fine-tuning language models

Transfer Learning

- Fine-tuning is an instance of transfer learning

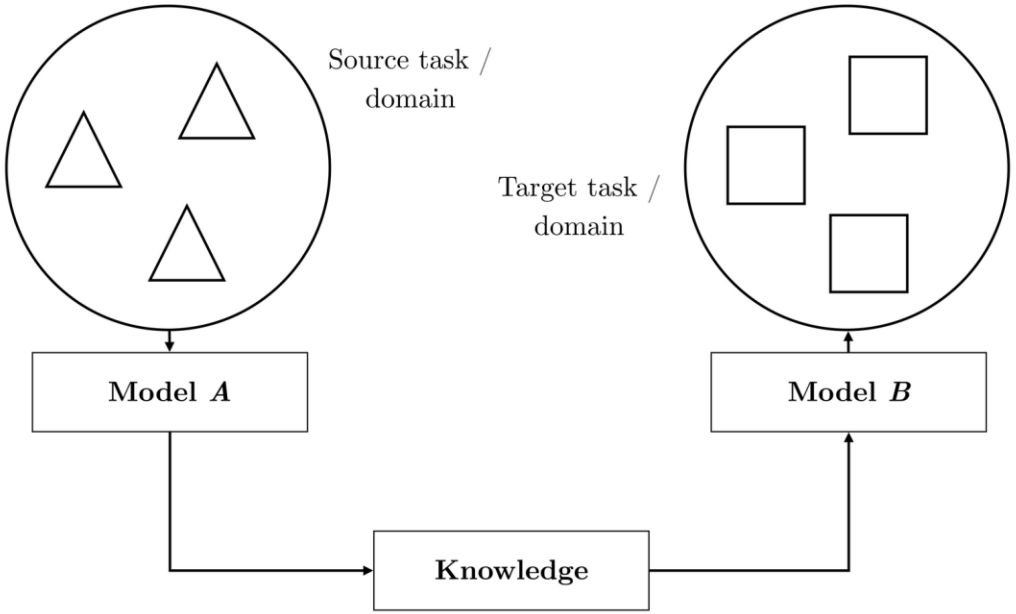


Image source: https://www.ruder.io/state-of-transfer-learning-in-nlp/

Sequence Classification

- Obtain the representation of the input sequence by using the final hidden state (hidden state at the position of the special token [CLS]) $C \in R^H$
- Just add a classification layer and use softmax to calculate label probabilities. Parameters $W \in R^{K \times H}$

$$P = \text{softmax}(CW^T)$$

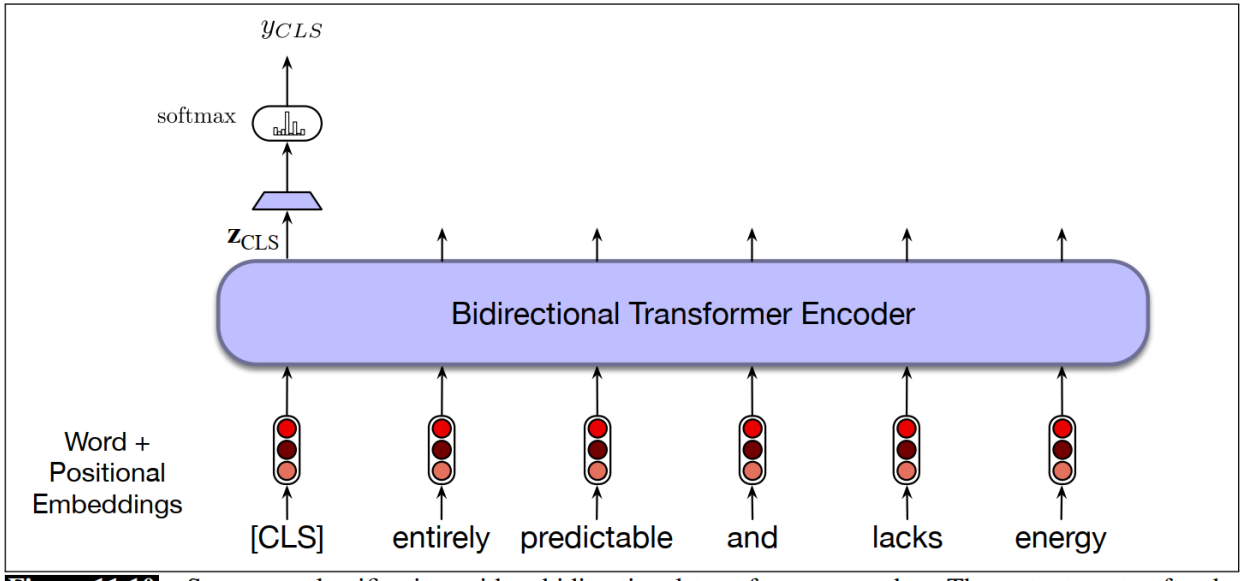


Figure 11.10 Sequence classification with a bidirectional transformer encoder. The output vector for the [CLS] token serves as input to a simple classifier.

Sequence Classification (2)

- All of the parameters of BERT and W are fine-tuned jointly
- Most model hyperparameters are the same as in pre-training
  - except the batch size, learning rate, and number of training epochs

Sequence Labeling

- Token tagging task (e.g., Named Entity Recognition)
  - Feed the final hidden representation z_i for each token i into a classification layer for the tagset (NER label set)

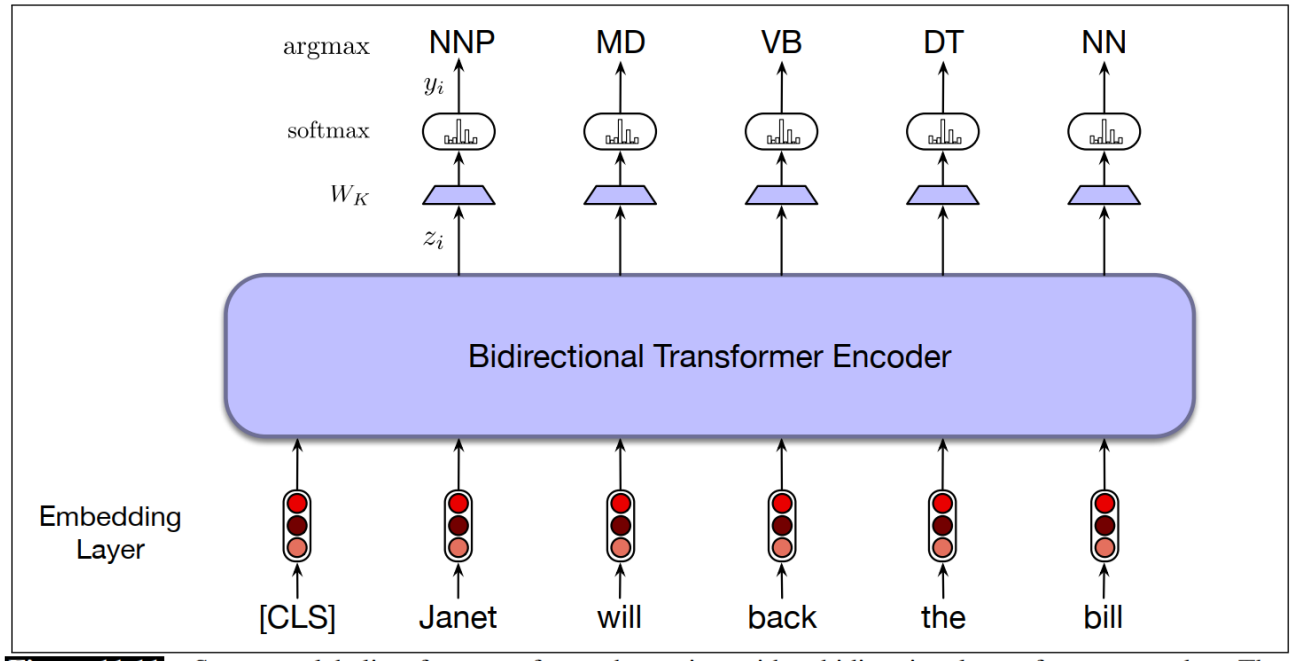


Figure 11.11 Sequence labeling for part-of-speech tagging with a bidirectional transformer encoder. The output vector for each input token is passed to a simple k-way classifier.

Sequence Labeling: Decoding

Mt.	Sanita	is	in	Sunshine	Canyon	.
B-LOC	I-LOC	O	O	B-LOC	I-LOC	O

Mt	.	San	##itas	is	in	Sunshine	Canyon	.
B-LOC	X	I-LOC	X	O	O	B-LOC	I-LOC	O

To make the task compatible with WordPiece tokenization

- Predict the tag for the first sub-token of a word
- No prediction is made for X

Pair-wise Sequence Classification (1)

- Paraphrase Detection
- Natural Language Inference. E.g,
  - Neutral
    - a: Jon walked back to the town to the smithy.
    - b: Jon traveled back to his hometown.
  - Contradicts
    - a: Tourist Information offices can be very helpful.
    - b: Tourist Information offices are never of any help.
  - Entails
    - a: I'm confused.
    - b: Not all of it is very clear to me.

Pair-wise Sequence Classification (2)

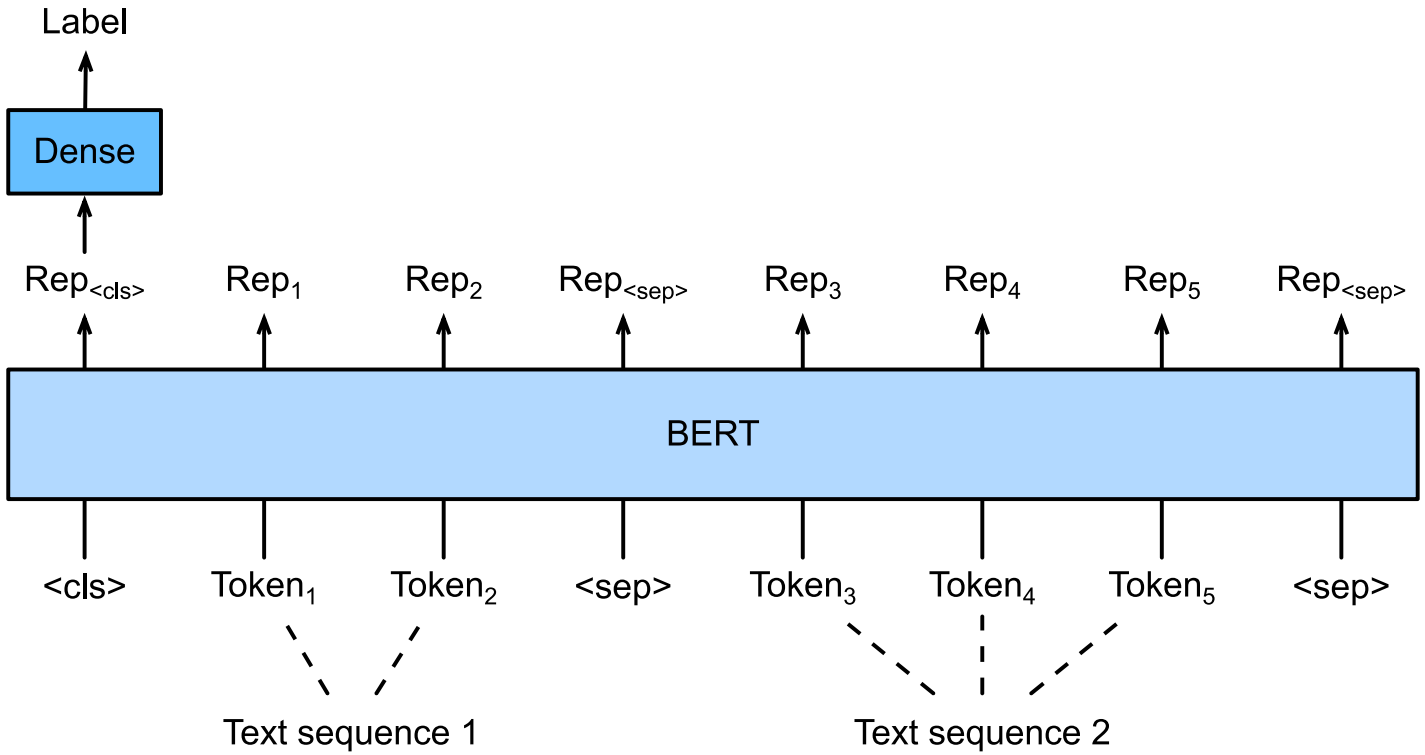


Image source: https://tinyurl.com/mr2xb9e3

Implementation of Transformers

- https://github.com/huggingface/transformers
- https://huggingface.co/learn/nlp-course/chapter1