上海海事大学
SHANGHAI MARITIME UNIVERSITY

自然语言处理

2024–2025 学年第 2 学期

信息工程学院　谢雨波

预训练语言模型

- **掩码语言模型（Masked Language Model）**
  - Transformer 编码器（Encoder-Only）
  - BERT, RoBERTa, ERNIE, ALBERT, DeBERTa, …
- **因果语言模型（Causal Language Model）**
  - Transformer 解码器（Decoder-Only）
  - GPT, PaLM, Mistral (Mixtral), Llama, DeepSeek, …
- **编码器-解码器语言模型（Encoder-Decoder Language Model）**
  - BART, T5, …

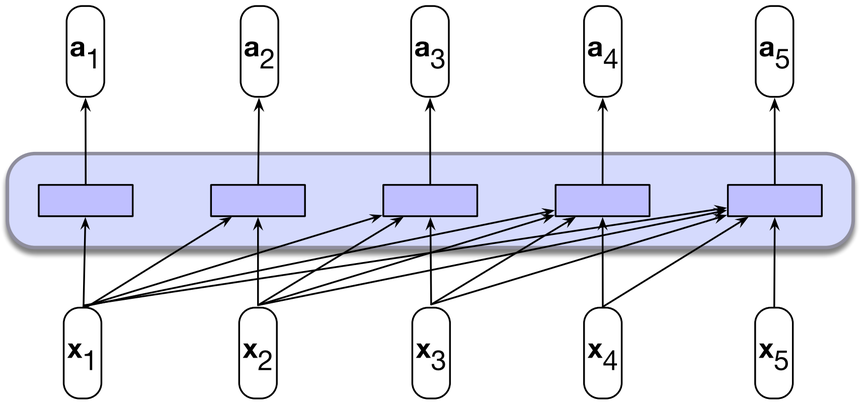
掩码语言模型

掩码语言模型

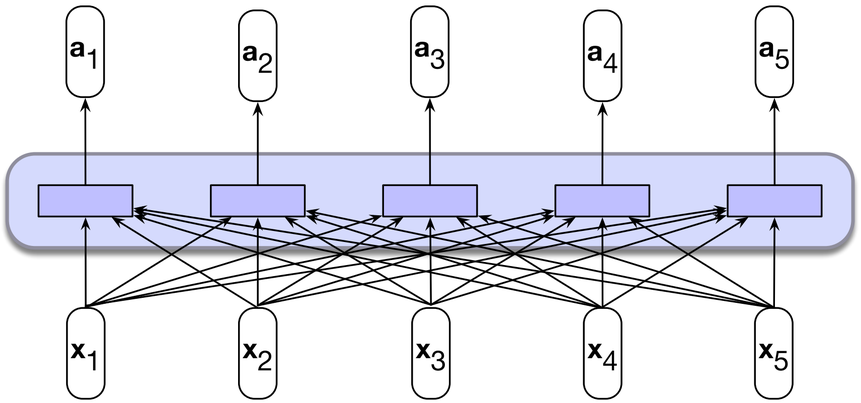
- **掩码语言模型（Masked Language Model）**
  - 预训练语言模型的一种范式
  - 通常采用双向 Transformer 编码器（Bidirectional Transformer Encoder）
- **相关概念：**
  - 微调（Fine-tuning）、迁移学习（Transfer Learning）
  - 上下文嵌入（Contextual Embedding）

双向 Transformer 编码器

- 相比较于单向的注意力机制，双向注意力：
  - 可以获取当前词之后的词的信息
  - 更适合于文本分类、序列标注



单向注意力



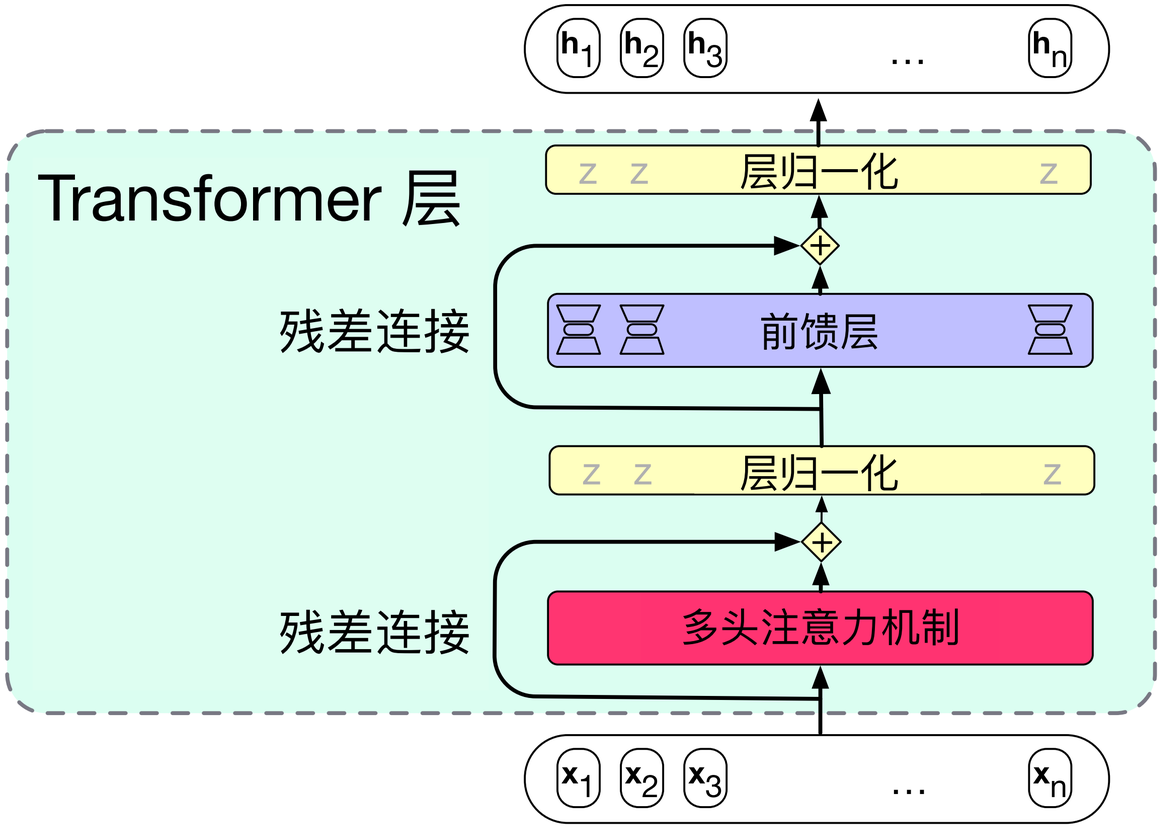
双向注意力

BERT

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2019.

- **Bidirectional Encoder Representations from Transformers (BERT)**
- 子词词表大小 30,000 (WordPiece)

$BERT_{BASE}$	**$BERT_{LARGE}$**
隐藏状态大小 768	隐藏状态大小 1024
12 层 Transformer 层	24 层 Transformer 层
12 头注意力	16 头注意力
参数量 110M	参数量 340M



训练 BERT

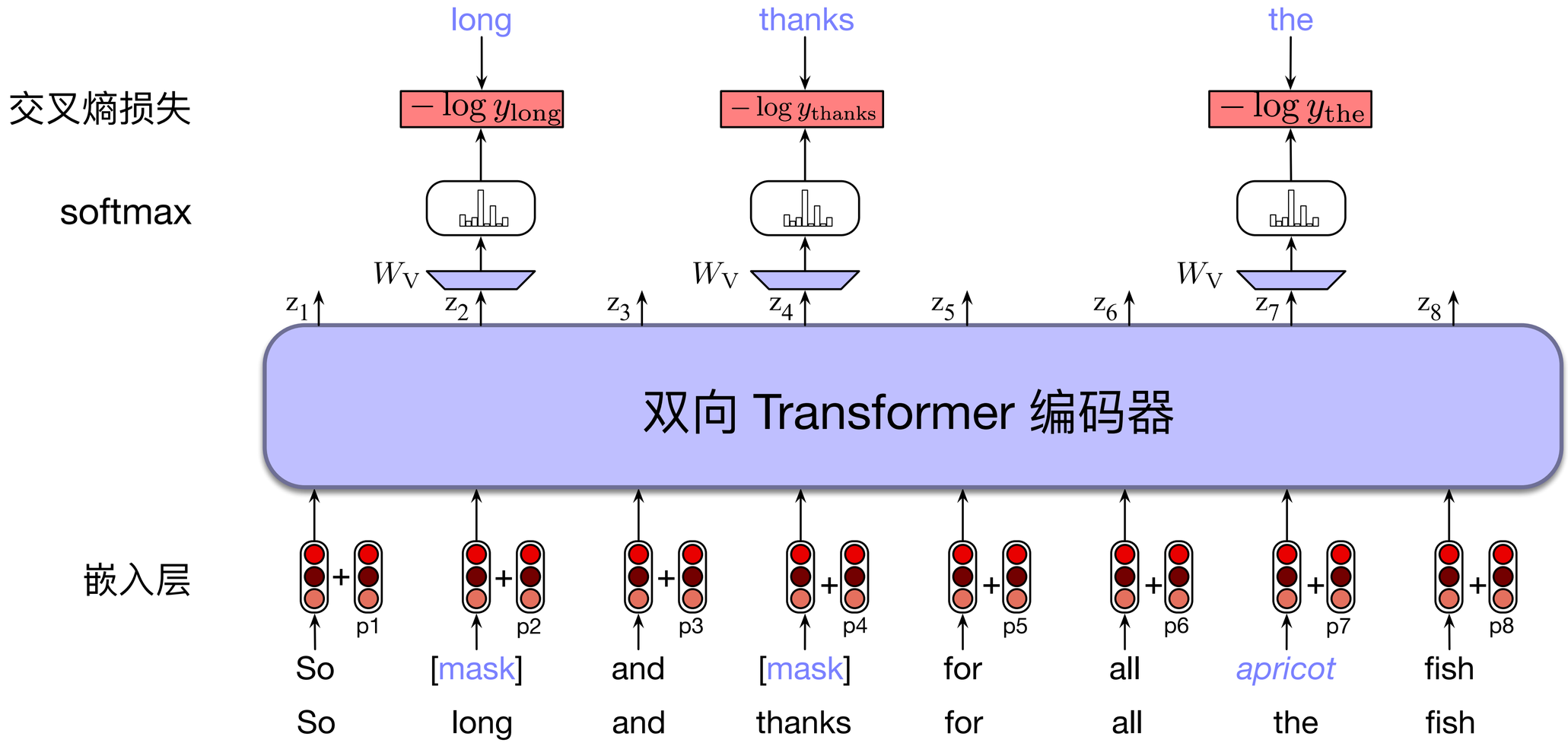
- **任务 1：掩码语言模型（Masked LM）**
  - “完形填空”（Cloze Task）
  - 输入序列的一部分是缺失的，模型需要预测缺失的部分

Please hand in ? homework .

掩码语言模型

- **掩码语言模型（Masked Language Model, MLM）**
- 在 BERT 中，随机从一个训练序列中选择 Token（15%），然后进行以下操作中的一种：
  - 将其替换为特殊词 [MASK] 80%
  - 将其替换为词表中的一个其他词（根据概率随机采样） 10%
  - 保持不变 10%

掩码语言模型


交叉熵损失
softmax
嵌入层
long
thanks
the
$-\log y_{\text{long}}$
$-\log y_{\text{thanks}}$
$-\log y_{\text{the}}$
W_V
W_V
W_V
z_1
z_2
z_3
z_4
z_5
z_6
z_7
z_8
双向 Transformer 编码器
p1
p2
p3
p4
p5
p6
p7
p8
So
[mask]
and
[mask]
for
all
apricot
fish
So
long
and
thanks
for
all
the
fish

掩码语言模型

- 对于某一输入词 x_i（替换前），z_i 为 BERT 的输出

$$L_{\mathrm{MLM}} = -\log P(x_i \mid z_i)$$

- 令 M 为选中的 Token 集合

$$L_{\mathrm{MLM}} = -\frac{1}{|M|}\sum_{i \in M} \log P(x_i \mid z_i)$$

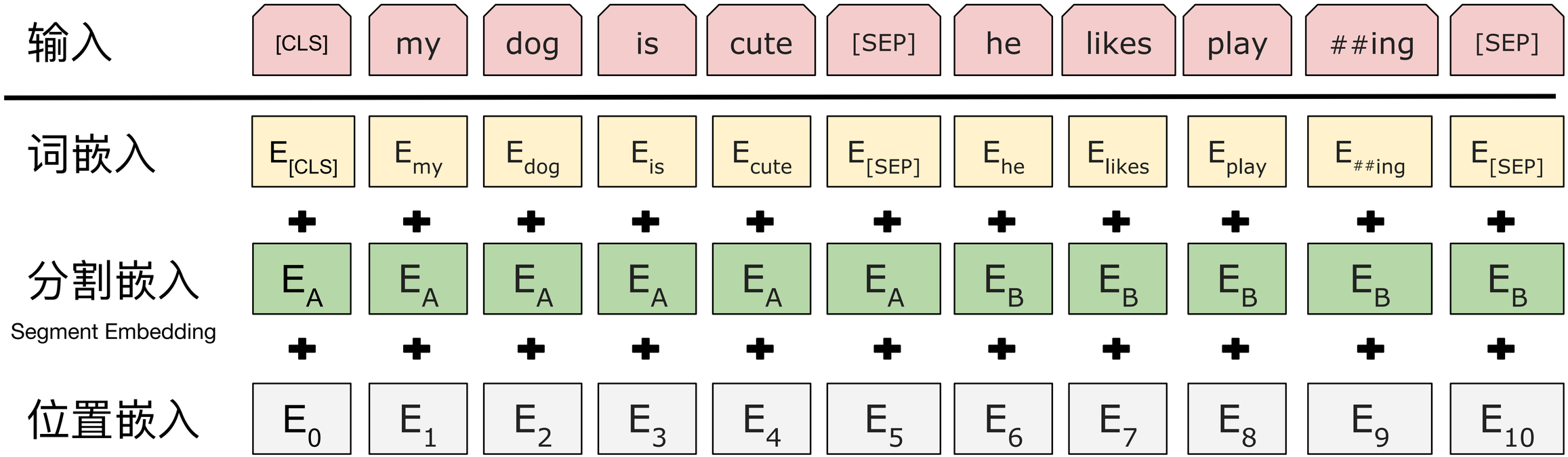
训练 BERT

- **任务 2：下一语句预测（Next Sentence Prediction, NSP）**
  - 有些 NLP 任务需要确定一对语句之间的关系
  - **释义识别（Paraphrase Detection）**：两个语句是否有相似的意思
  - **文本蕴含（Textual Entailment）**：给定前提文本，推断假说文本与其的关系，一般有蕴含关系（Entailment）和矛盾关系（Contradiction）
  - **语篇连贯性（Discourse Coherence）**：两个语句是否来自同一语篇的相邻句子

下一语句预测

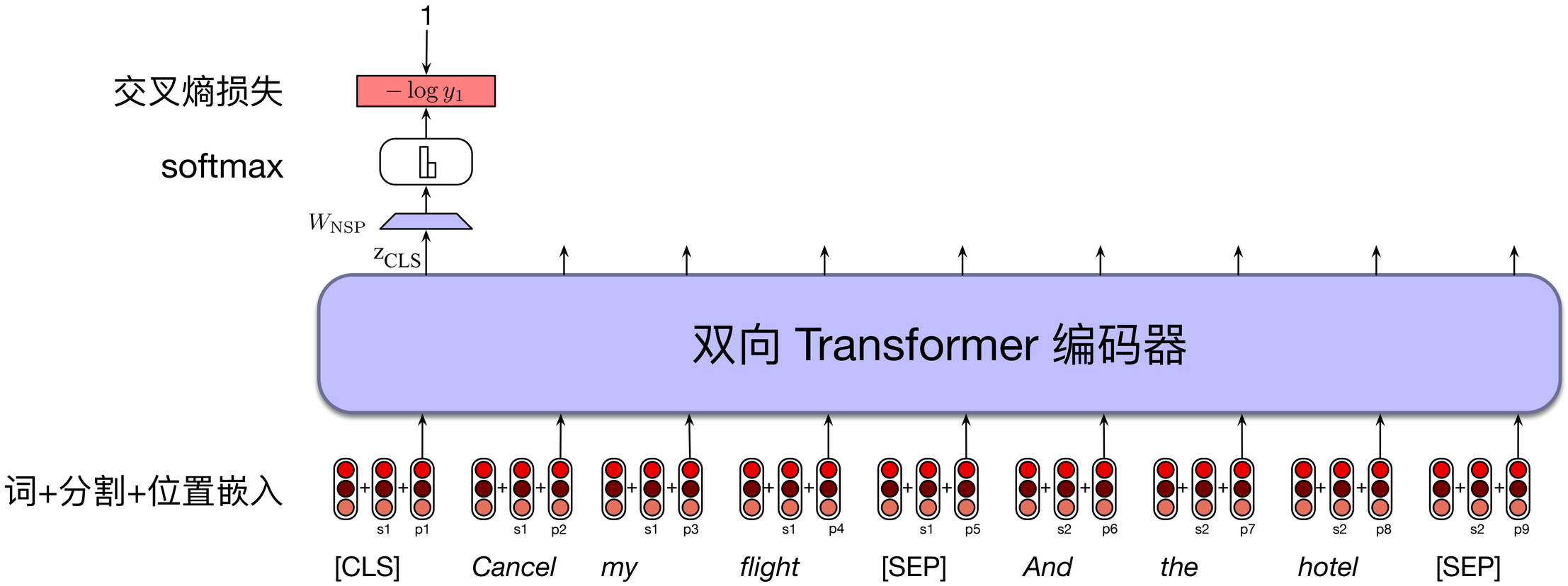
- 在 BERT 中，训练样本为**语句对（Sentence Pair）**
  - 目标为预测输入的语句对是否来自于训练语料库中的相邻语句，或是无关的语句对
  - 50% 为正向样本，50% 为负向样本（随机挑选的两个句子）
  - 两个新的 Token：`[CLS]` 和 `[SEP]`

下一语句预测


输入
[CLS] my dog is cute [SEP] he likes play ##ing [SEP]
词嵌入
E[CLS] Emy Edog Eis Ecute E[SEP] Ehe Elikes Eplay E##ing E[SEP]
分割嵌入
Segment Embedding
EA EA EA EA EA EA EB EB EB EB EB
位置嵌入
E0 E1 E2 E3 E4 E5 E6 E7 E8 E9 E10

下一语句预测

- 基于 [CLS] 的输出，进行二元分类预测



BERT 训练

- **训练集**：
  - BookCorpus + Wikipedia（33 亿单词，包含英语维基百科与书籍语料库）
  - BookCorpus 因版权原因，不再被使用
- 输入序列长度上限：512
- 训练 Epoch 数目：~40

多语言模型训练

- 如何建立一个多语言的词表（使用 BPE 等子词分词算法）？
- 从训练语料库中随机采样句子，进行分词，得到词表
  - 结果将更偏向于常见语言的罕见词
  - 而不是罕见语言的常见词
- 因此，需要对语种的概率分布进行调整

多语言模型训练

- 将训练语料库基于语种分为 N 个子语料库
- 对于语种 i，语料库中有 n_i 个句子
- 从语种 i 中挑选句子的概率为

$$q_i = \frac{p_i^\alpha}{\sum_{j=1}^{N} p_j^\alpha} \qquad p_i = \frac{n_i}{\sum_{j=1}^{N} n_j}$$

- 0 和 1 之间的 α 值将增加低概率事件的概率值，此处一般取 $\alpha = 0.3$

多语言模型训练

- 为什么训练一个多语言模型?
  - 不需要为每一种语言训练一个单独的模型
  - 低资源语言（Low-Resource Language）的模型效果可以通过语料库中相似语言的训练样本来进行改善
- **Curse of Multilinguality**：当训练的语言数目增长至非常大时，每一种语言的模型效果将降低 (Conneau et al., 2020)
- 由于各语言资源的不平衡，高资源语言（如英语）的某些语法结构将呈现在低资源语言上 (Papadimitriou et al., 2023)

上下文嵌入

上下文嵌入

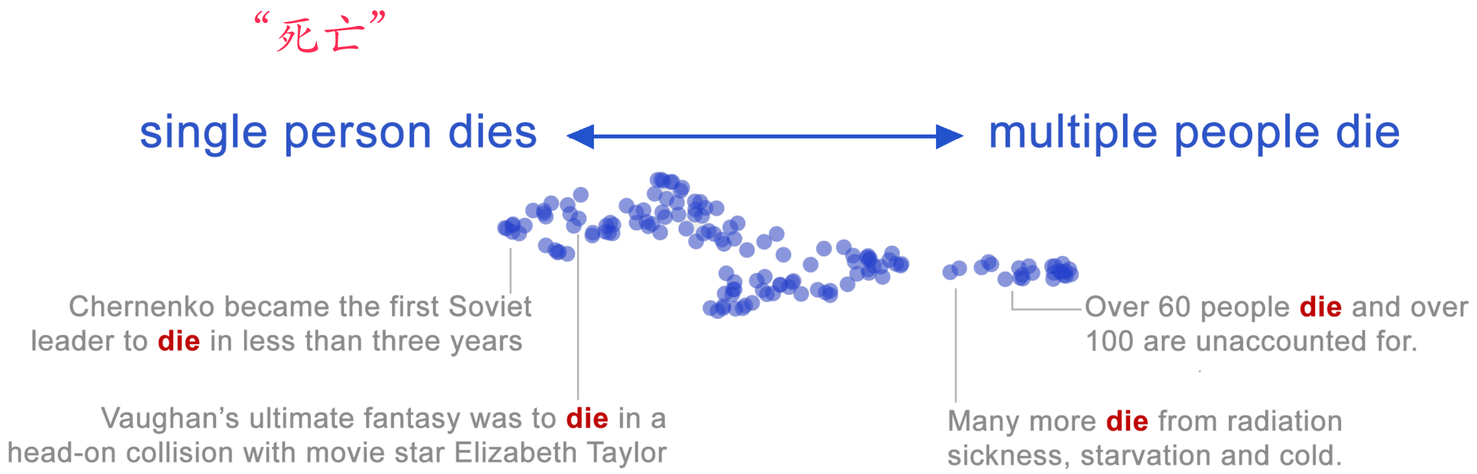
- BERT 模型为输入序列的每一个词 x_i 输出一个向量表示 z_i
- 此向量表示是输入词的**上下文嵌入（Contextual Embedding）**表示
  - 在 $x_1, \cdots, x_n$ 的上下文中，词 x_i 的语义表示
  - 也可以将模型后 4 层的输出 z_i 进行平均
- 是一种动态的词嵌入，相较于 word2vec 的静态词嵌入

上下文嵌入与词义

- 单词是具有歧义的，即一个单词可能有多种**词义（Word Sense）**
- 可以根据上下文来确定一个单词的词义
  - **mouse**[1] : … a *mouse* controlling a computer system in 1968.
  - **mouse**[2] : … a quiet animal like a *mouse*
  - **bank**[1] : … a *bank* can hold the investments in a custodial account …
  - **bank**[2] : … as agriculture burgeons on the east *bank*, the river …

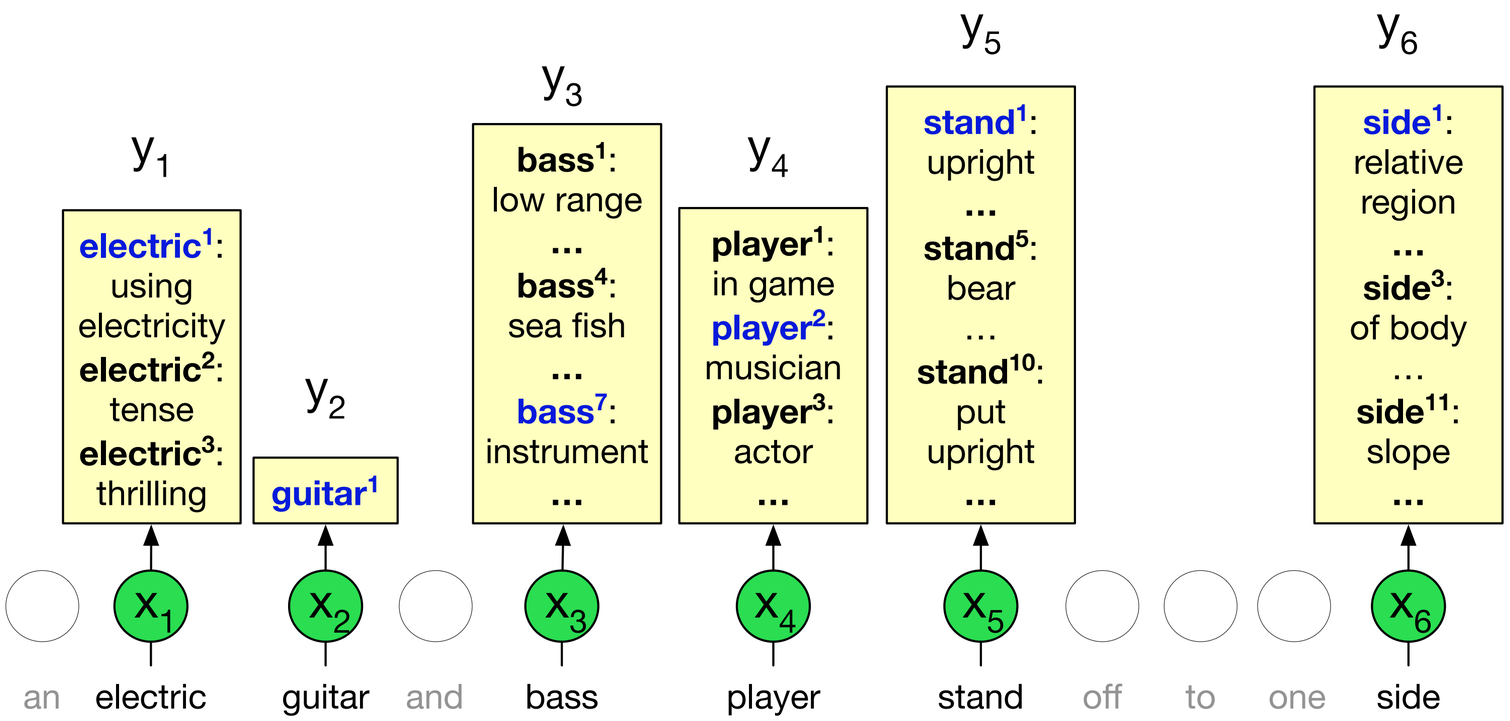
上下文嵌入与词义

Coenen, A., E. Reif, A. Yuan, B. Kim, A. Pearce, F. Viégas, and M. Wattenberg. Visualizing and measuring the geometry of BERT. NeurIPS 2019.



词义消歧

- **词义消歧（Word Sense Disambiguation, WSD）**
  - 为输入单词选择正确的词义



词义消歧

1-邻近算法（1-Nearest-Neighbor Algorithm）

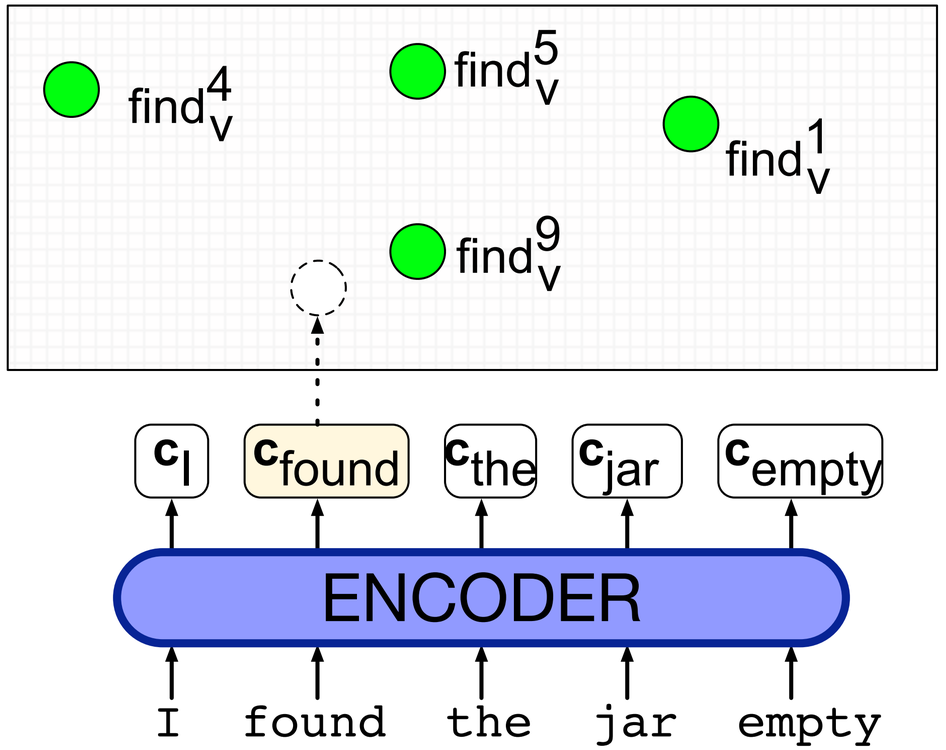
- 将词义标注数据集（如 SemCore 和 SenseEval）输入 BERT 获得每一个标注词的上下文嵌入
- 对于语料库中任一词的每一个词义 s，对符合词义 s 的 n 个词的上下文嵌入 v_i 进行平均得到词义 s 的上下文嵌入：

$$v_s = \frac{1}{n}\sum_i v_i \qquad \forall v_i \in \text{tokens}(s)$$

词义消歧

- 对于测试词 t，计算它的上下文嵌入和每一个词义嵌入的相似度，选择相似度最高的词义作为返回结果

$$\text{sense}(t) = \underset{s \in \text{senses}(t)}{\text{argmax}} \cos\left(t, v_s\right)$$



微调

微调

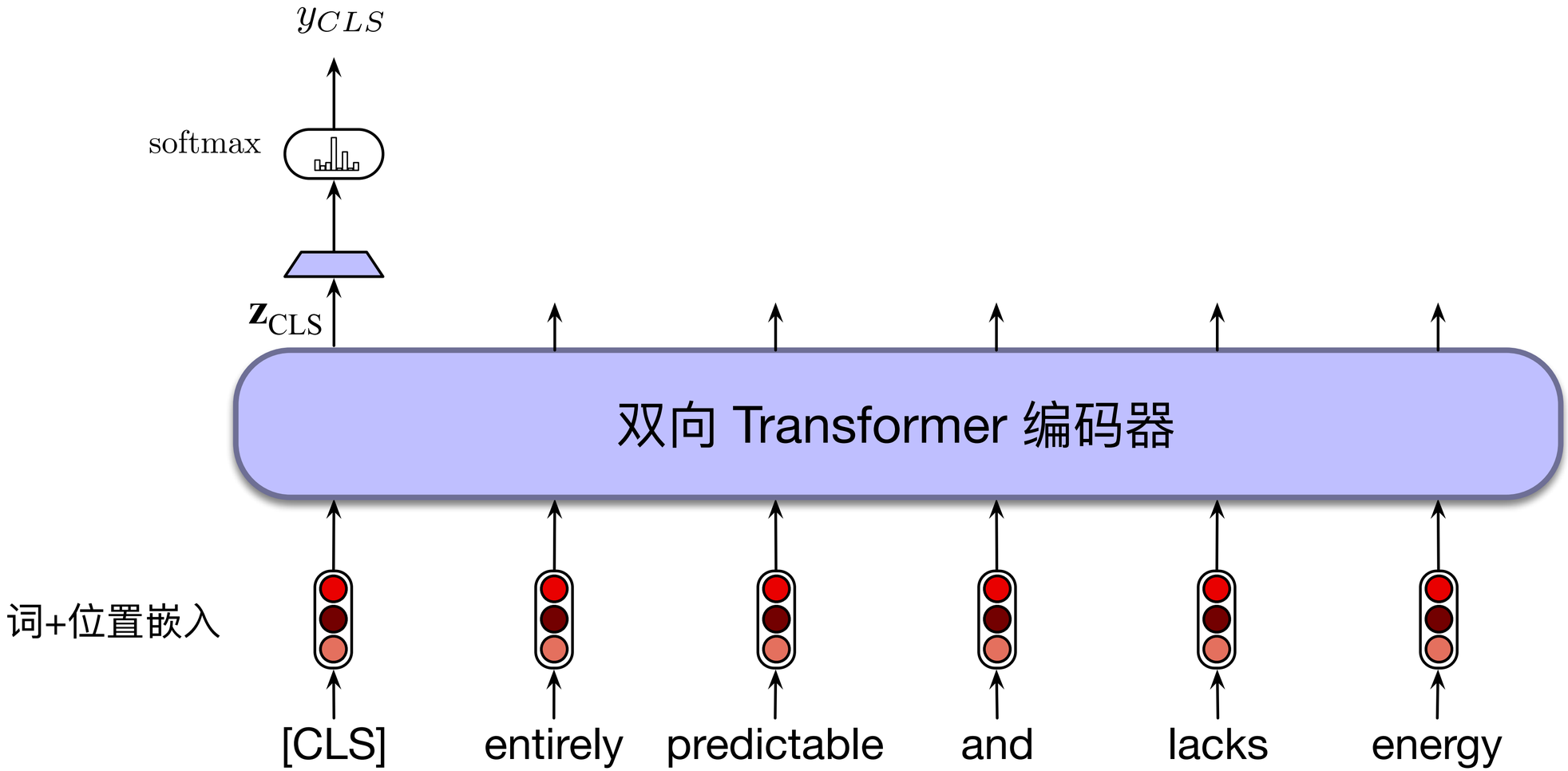
- 将预训练语言模型用于特定的下游任务
- **微调（Fine-tuning）**
  - 在预训练模型的基础上增加少量与特定任务相关的参数
  - 使用特定任务的标注数据对这些参数进行训练
  - 通常会保持原预训练模型的参数不变，或仅做少量调整（例如，仅调整原预训练模型的最后几层参数）

序列分类

- 在对输入序列进行分类时，通常将输入序列表示为一个向量
  - 例如，RNN 中的最后一个隐藏状态
- 在 BERT 中，可以使用 [CLS] 的输出向量 z_{CLS} 表示整个输入序列
  - 也称为**句子嵌入（Sentence Embedding）**
- 将其输入一个**分类头（Classifier Head）**进行分类：

$$y = \text{softmax}(W_C z_{\text{CLS}})$$

序列分类


y_{CLS}
softmax
$\mathbf{z}_{\text{CLS}}$
双向 Transformer 编码器
词+位置嵌入
[CLS]
entirely
predictable
and
lacks
energy

成对序列分类

- **成对序列分类（Pairwise Sequence Classification）**
  - 输入由 A 和 B 两个句子组成
  - 释义识别（Paraphrase Detection）：A 和 B 是否是对方的转述?
  - 文本蕴含（Textual Entailment）：A 是否在逻辑上蕴含 B?
  - 语篇连贯性（Discourse Coherence）：B 接在 A 后面是否连贯?

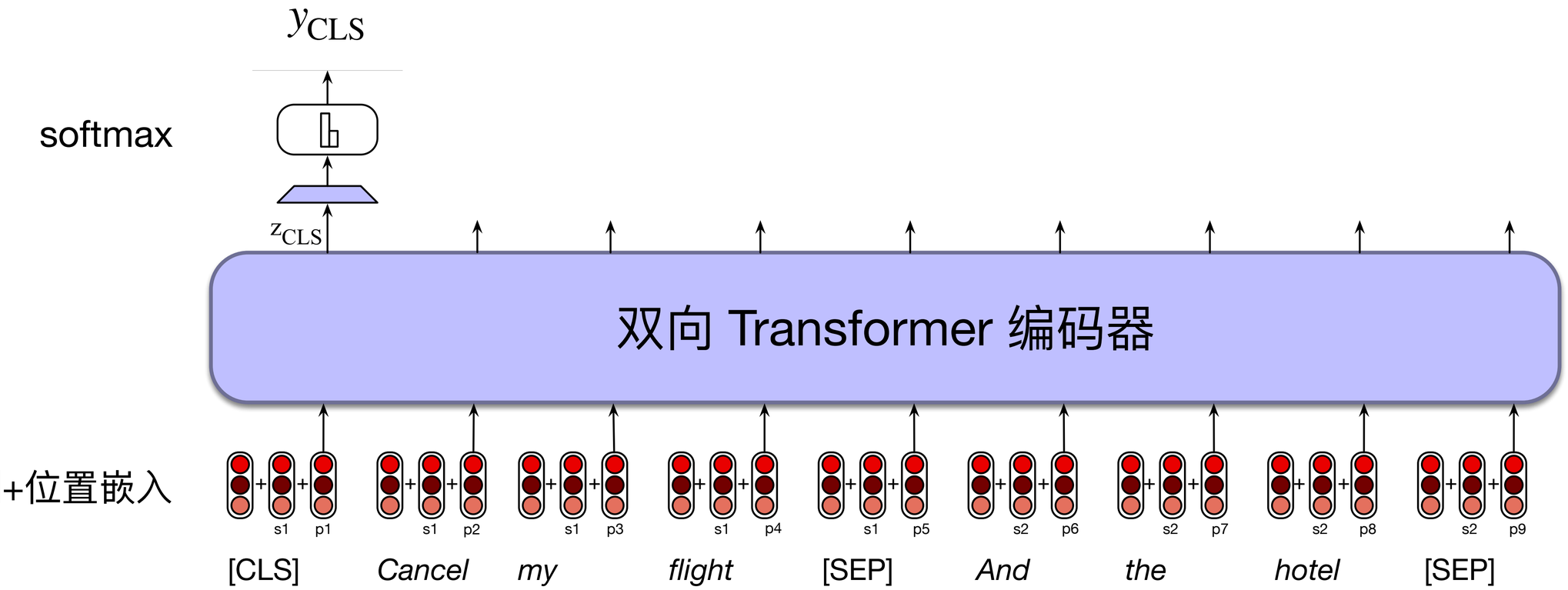
成对序列分类

- 例如，文本蕴含任务又称**自然语言推断（Natural Language Inference, NLI）**
  - Multi-Genre Natural Language Inference (MultiNLI) 数据集
  - 句子对：前提（Premise）+ 假设（Hypothesis）
  - 有 3 个标签：*entails*, *contradicts*, *neutral*

- `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.

成对序列分类

- 模型结构和 NSP 相同



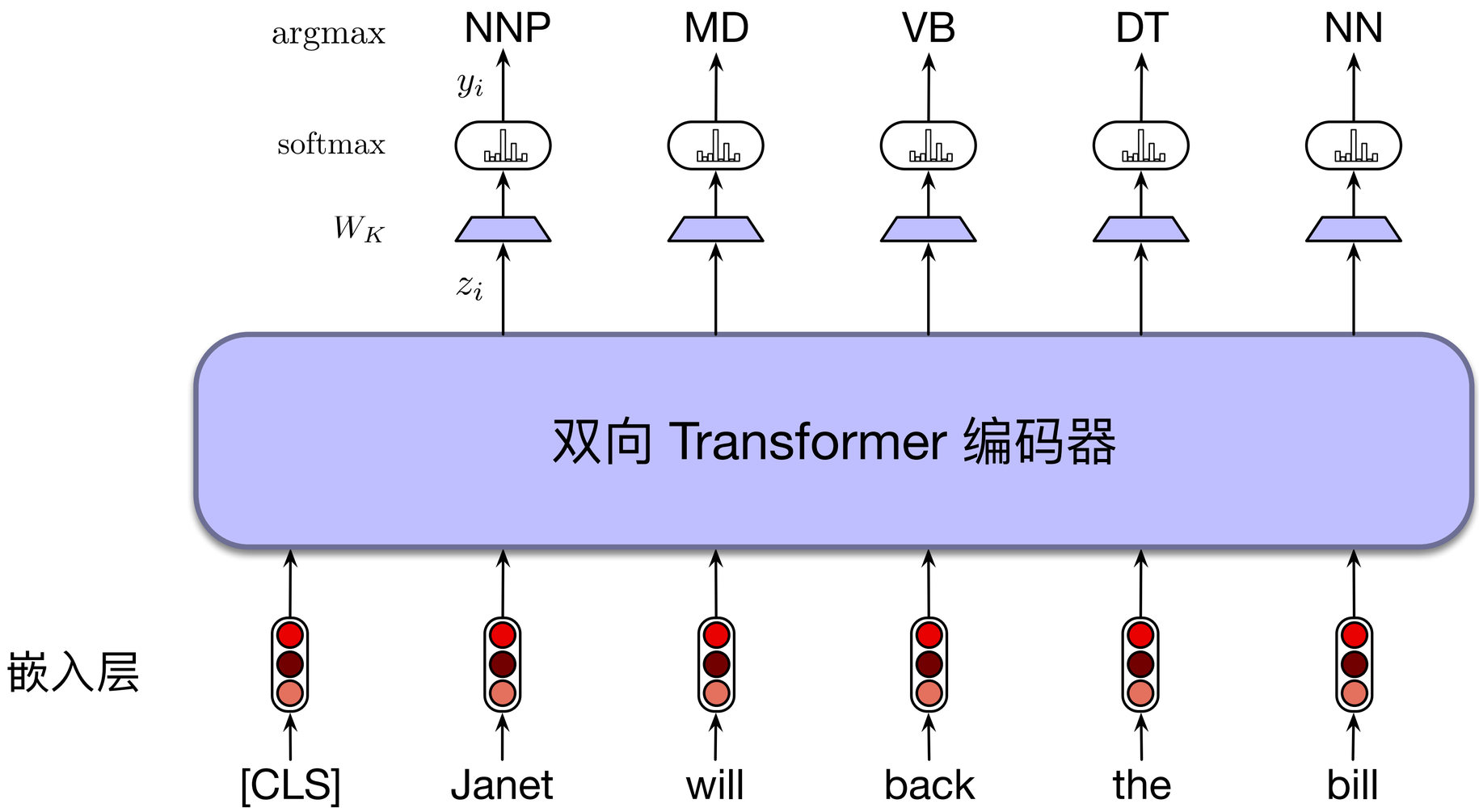
序列标注

- **序列标注（Sequence Labeling）**
  - 词性标注（Part-Of-Speech Tagging）
  - 命名实体识别 (Named Entity Recognition, NER)
- 每一个输入词的输出向量都用于预测标签

$$y_i = \text{softmax}\left(W_K z_i\right)$$

$$t_i = \underset{k}{\text{argmax}}\left(y_i\right)$$

序列标注


argmax
NNP
MD
VB
DT
NN
y_i
softmax
W_K
z_i
双向 Transformer 编码器
嵌入层
[CLS]
Janet
will
back
the
bill

BERT 相关模型

RoBERTa

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019.

- **RoBERTa (Robustly optimized BERT approach)**
- **训练数据更多**：BookCorpus, CC-News, OpenWebText, Stories, 共计 160GB
- **动态掩码**：每次将序列输入模型时，进行掩码操作，而不是在数据预处理时进行掩码操作

Masking	SQuAD 2.0	MNLI-m	SST-2
reference	76.3	84.3	92.8
Our reimplementation:			
static	78.3	84.3	92.5
dynamic	78.7	84.0	92.9

RoBERTa

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019.

- **去除 NSP 任务**：在去除 NSP 损失后，发现模型在下游任务上的效果稍稍变好
  - 可能的原因：NSP 任务较为简单：**Topic Prediction** + Coherence Prediction；原 BERT 论文在去除 NSP 后，依旧保留了 Segment-Pair 的输入格式

Model	SQuAD 1.1/2.0	MNLI-m	SST-2	RACE
Our reimplementation (with NSP loss):				
SEGMENT-PAIR	90.4/78.7	84.0	92.9	64.2
SENTENCE-PAIR	88.7/76.2	82.9	92.1	63.0
Our reimplementation (without NSP loss):				
FULL-SENTENCES	90.4/79.1	84.7	92.5	64.8
DOC-SENTENCES	90.6/79.7	84.7	92.7	65.6
$\text{BERT}_{\text{BASE}}$	88.5/76.3	84.3	92.8	64.3
$\text{XLNet}_{\text{BASE}}$ (K = 7)	–/81.3	85.8	92.7	66.1
$\text{XLNet}_{\text{BASE}}$ (K = 6)	–/81.0	85.6	93.4	66.7

RoBERTa

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019.

- 使用 BPE 算法进行分词，词表大小为 50K
  - 分词单位为字节（Byte），而不是 Unicode 字符
  - 可以将词表控制在一个合适的大小，同时泛化能力更强
- 增大 Batch Size

bsz	steps	lr	ppl	MNLI-m	SST-2
256	1M	1e-4	3.99	84.7	92.7
2K	125K	7e-4	**3.68**	**85.2**	**92.9**
8K	31K	1e-3	3.77	84.6	92.8

ALBERT

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. ICLR 2020.

- ALBERT: A Lite BERT，BERT 的轻量化版本
- 嵌入矩阵分解：不直接将输入的独热向量映射到模型的隐藏状态大小 H，而是先映射到较低维度的空间（大小为 E），然后再映射到 H

$$O(V \times H) \rightarrow O(V \times E + E \times H) \qquad H \gg E$$

- 将所有的 Transformer 层的参数共享
- 去除 NSP，加入句子顺序预测（Sentence-Order Prediction, SOP）
- **ALBERT-large: 60M** (BERT-large: 340M)

DistilBERT

Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. NeurIPS 2019 Workshop.

- 对 BERT 模型进行压缩
- **知识蒸馏（Knowledge Distillation）**
  - Teacher：原 BERT 模型
  - Student：基本结构和 BERT 相同，层数减少 1/2（去掉 Segment Embedding）
- 参数量减少 40%，速度提升 60%，性能达到原 BERT 的 97%