Bidirectional Encoder Representations from Transformers (BERT)

Shusen Wang

What is BERT?

18 年 提出 19年发表

- BERT [1] is for pre-training Transformer's [2] encoder.
- How?
- Predict masked word. 想法1： 输入一个句子，随机遮挡一个或者多个单词，预测被遮挡的单词
- Predict next sentence. 想法2：把两个句子放在一起，判断是否是相邻的上下文

Reference

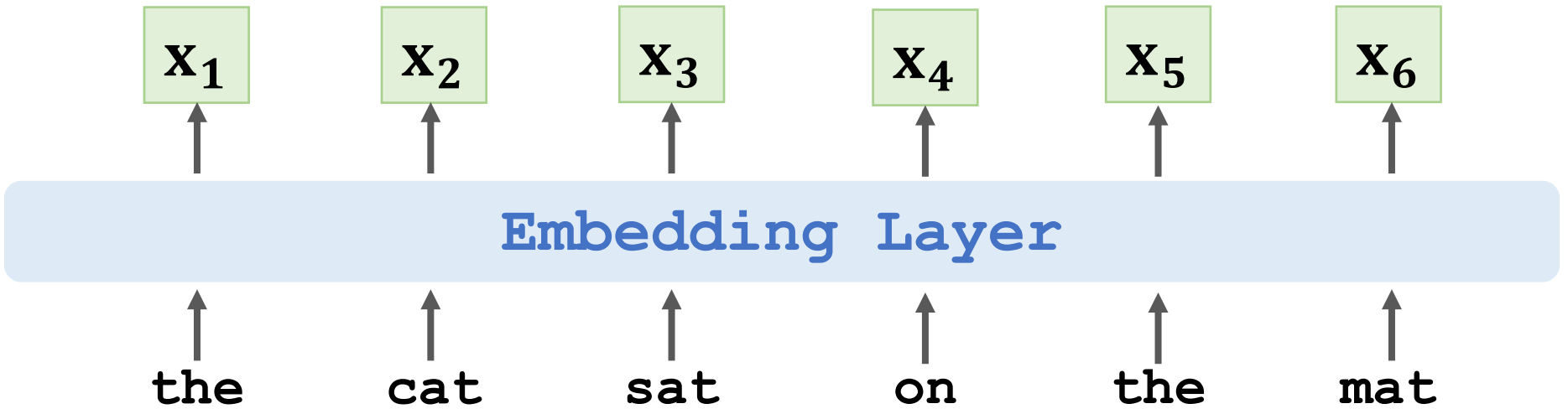
1. Devlin, Chang, Lee, and Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *ACL*, 2019.
2. Vaswani and others. Attention is all you need. In *NIPS*, 2017.

Task 1: Predict Masked Words

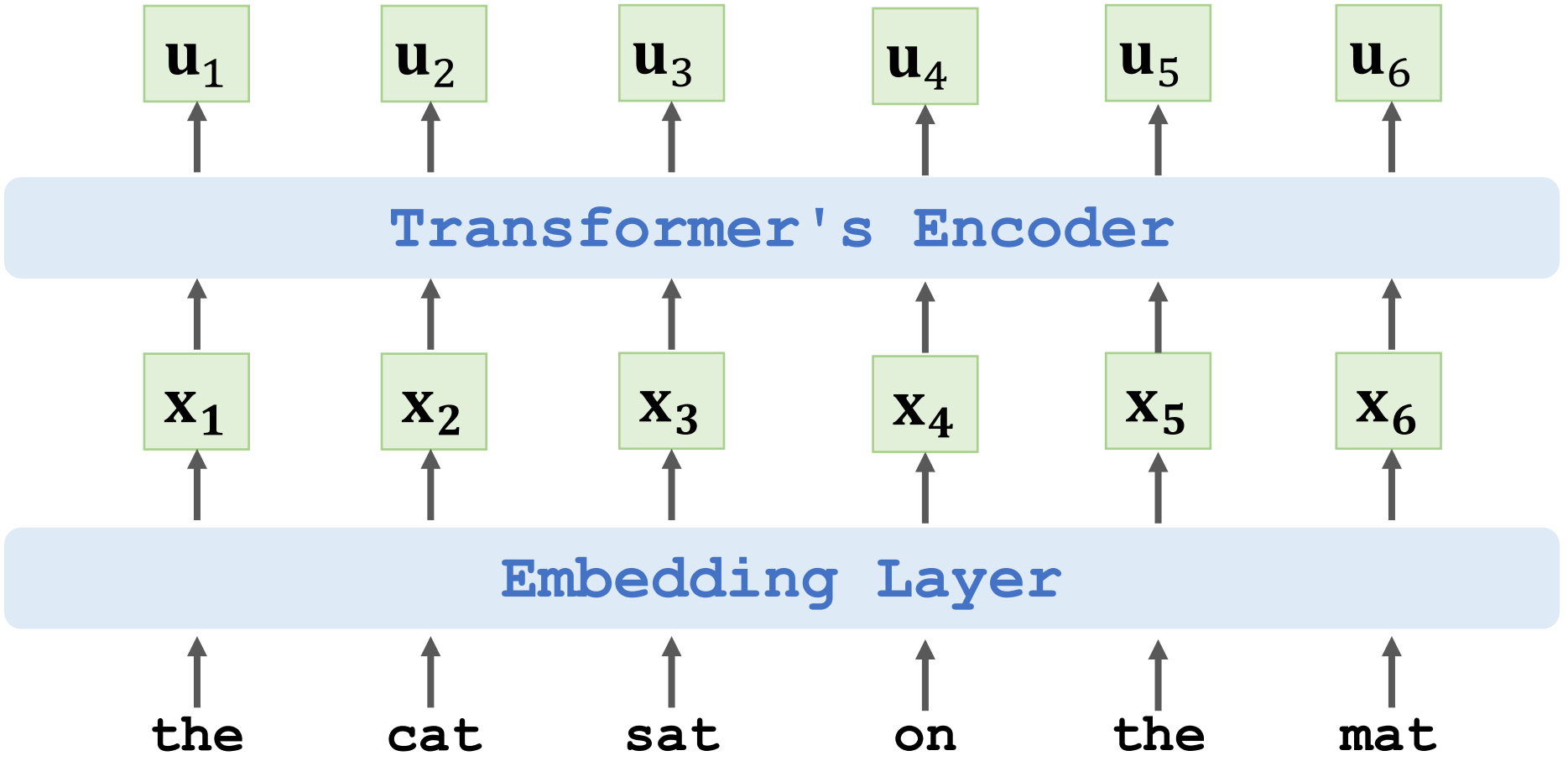
Revisit Transformer's Encoder

the cat sat on the mat

Revisit Transformer's Encoder



Revisit Transformer's Encoder

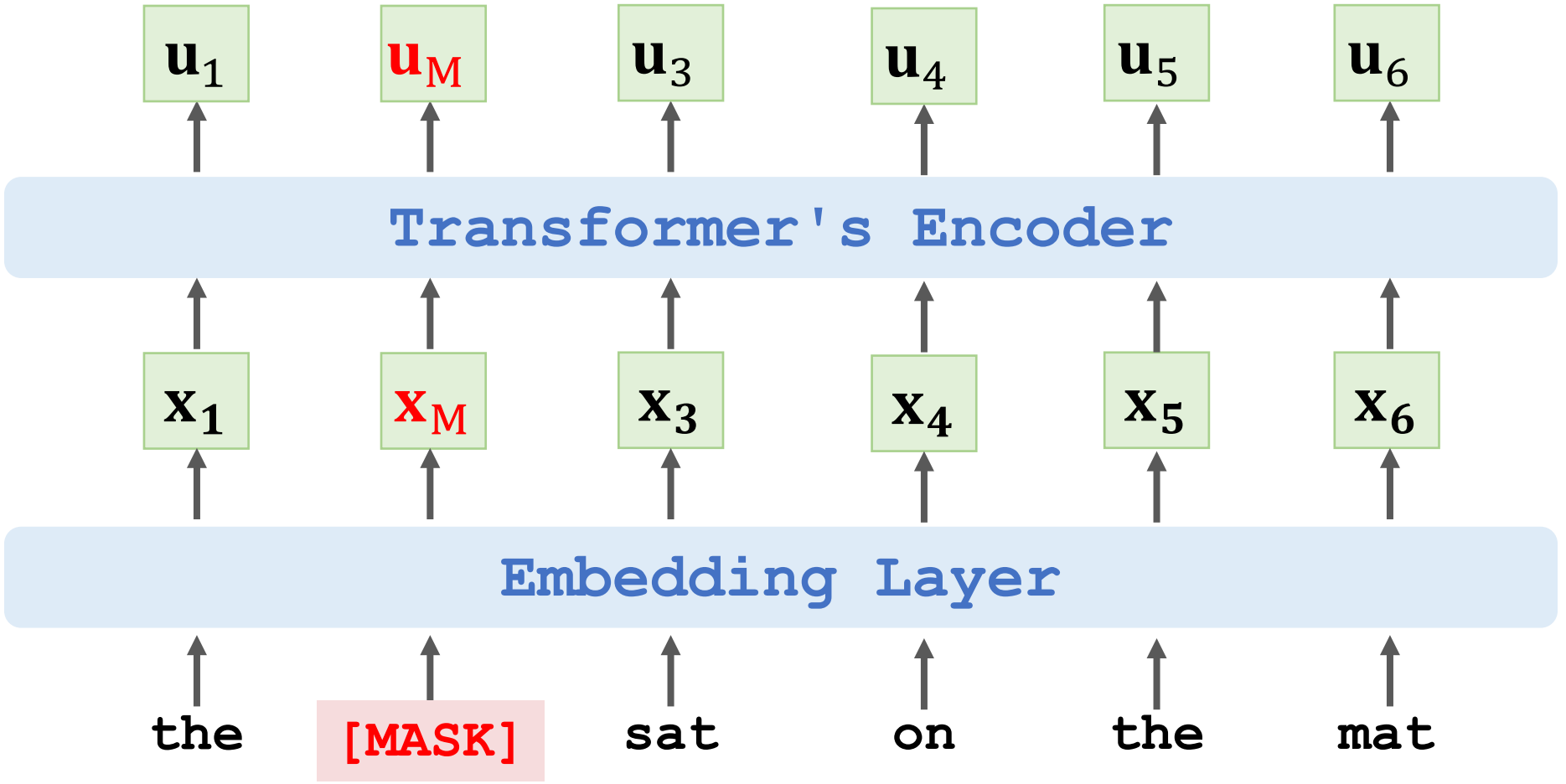


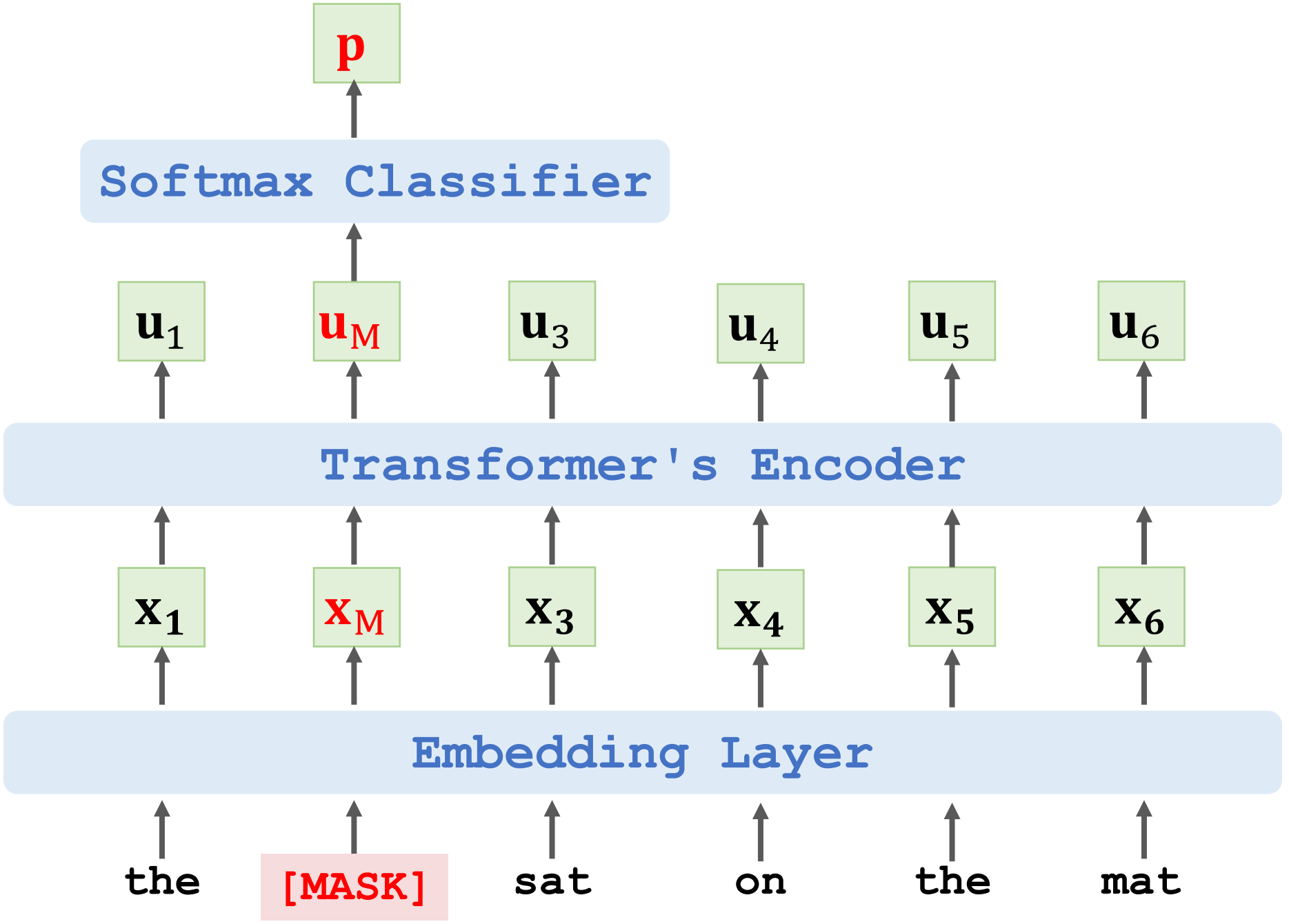
Randomly mask a word

- "The _____ sat on the mat"

- What is the masked word?

Randomly mask a word




p
Softmax Classifier
u_1
u_M
u_3
u_4
u_5
u_6
Transformer's Encoder
x_1
x_M
x_3
x_4
x_5
x_6
Embedding Layer
the
[MASK]
sat
on
the
mat

Predict the masked word

- $\mathbf{e}$: one-hot vector of the masked word “cat”.
- $\mathbf{p}$: output probability distribution at the masked position.
- $\text{Loss} = \text{CrossEntropy}(\mathbf{e}, \mathbf{p})$.
- Performing one gradient descent to update the model parameters.

e: groud truth 标签

在掩蔽位置的输出概率分布

思考：这个任务

无需人工标注

有文本就可以

Task 2: Predict the Next Sentence

Predict the next sentence

- Given the sentence:

  "calculus is a branch of math".

- Is this the next sentence?

  "it was developed by newton and leibniz"

Predict the next sentence

- Given the sentence:

  "calculus is a branch of math".

- Is this the next sentence?

  "it was developed by newton and leibniz"

- Is this the next sentence?

  "panda is native to south central china"

Input Representation

- **Input:**

[CLS] “calculus is a branch of math”

[SEP] “it was developed by newton and leibniz”

- **[CLS]** is a token for classification.
- **[SEP]** is for separating sentences.

Input Representation

- **Input:**

  [CLS] “calculus is a branch of math”

  [SEP] “it was developed by newton and leibniz”

- **Target:** true

Input Representation

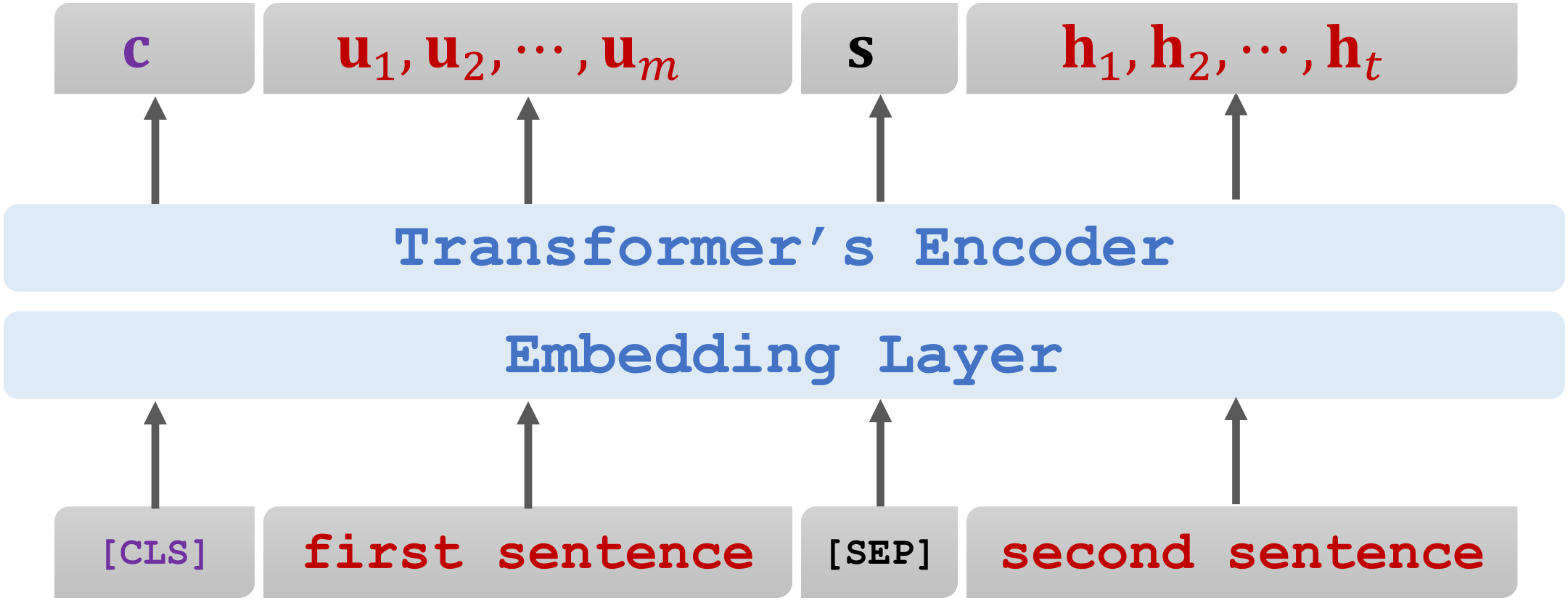
- **Input:**

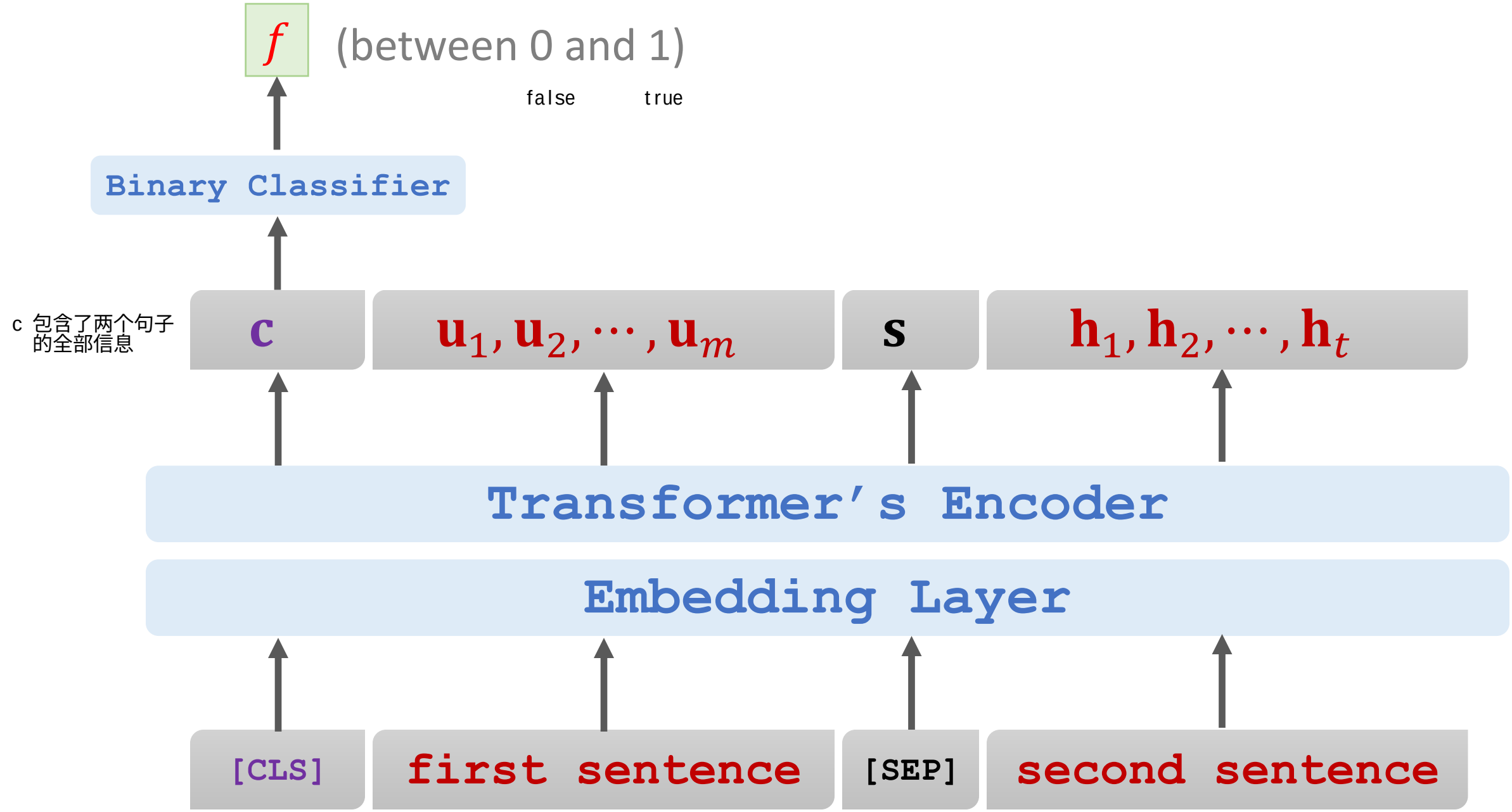
  [CLS] “calculus is a branch of math”

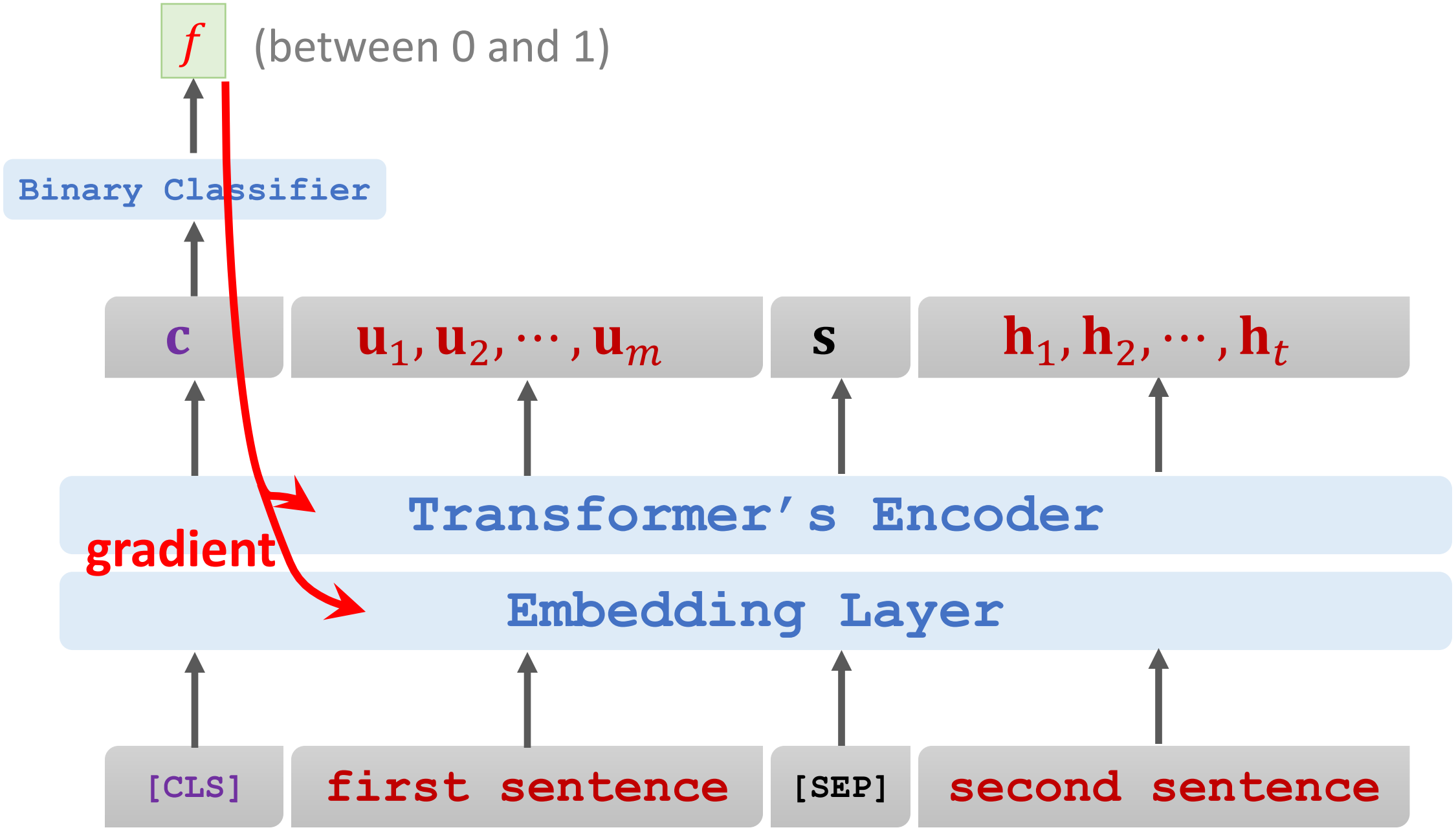
  [SEP] “panda is native to south central china”

- **Target:** false

Predict the next sentence




f
(between 0 and 1)
false
true
Binary Classifier
c 包含了两个句子的全部信息
$\mathbf{c}$
$\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_m$
$\mathbf{s}$
$\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_t$
Transformer's Encoder
Embedding Layer
[CLS]
first sentence
[SEP]
second sentence


f
(between 0 and 1)
Binary Classifier
$\mathbf{c}$
$\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_m$
$\mathbf{s}$
$\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_t$
Transformer's Encoder
gradient
Embedding Layer
[CLS]
first sentence
[SEP]
second sentence

Combining the two methods

Input Representation

输入表示

- **Input:**

"[CLS] calculus is a **[MASK]** of math [SEP] it **[MASK]** developed by newton and leibniz".

- **Targets:** true, "branch", "was".

Input Representation

- **Input:**

"[CLS] calculus is a branch of math [SEP] panda is native to **[MASK]** central china".

- **Targets:** false, "south".

Training

- Loss 1 is for binary classification (i.e., predicting the next sentence.) 损失1用于二元分类（即，预测下一个句子）。
- Loss 2 and Loss 3 are for multi-class classification (i.e., predicting the masked words.) 损失2和损失3用于多类分类（即，预测掩码词）。
- Objective function is the sum of the three loss functions. 目标函数是三个损失函数的和
- Update model parameters by performing one gradient descent.

通过执行一次梯度下降来更新模型参数。

Data

- BERT does not need manually labeled data. (Nice! Manual labeling is expensive.)
- Use large-scale data, e.g., English Wikipedia (2.5 billion words.)
- Randomly mask words (with some tricks.)
- 50% of the next sentences are real. (The other 50% are fake.)

Cost of Computation

- BERT Base
  - 110M parameters.
  - 16 TPUs, 4 days of training (without hyper-parameter tuning.)
- BERT Large
  - 235M parameters.
  - 64 TPUs, 4 days of training (without hyper-parameter tuning.)

Thank you!