Bidirectional Encoder Representations from Transformers (BERT)

Shusen Wang

What is BERT?

- BERT [1] is for pre-training Transformer's [2] encoder.
- How?
- Predict masked word.
- Predict next sentence.

Reference

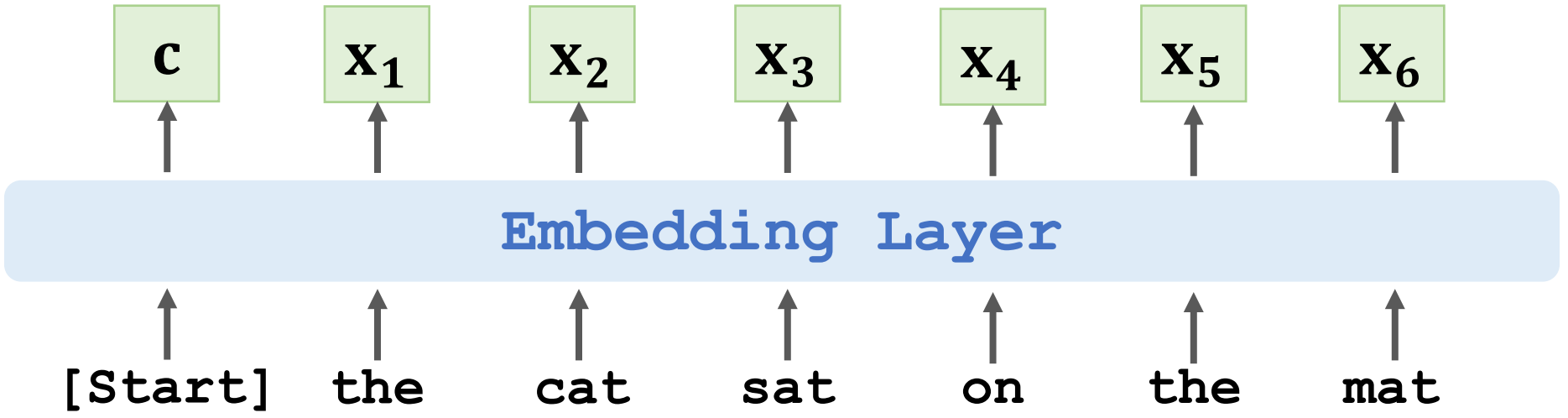
1. Devlin, Chang, Lee, and Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *ACL*, 2019.
2. Vaswani and others. Attention is all you need. In *NIPS*, 2017.

Predict Masked Word

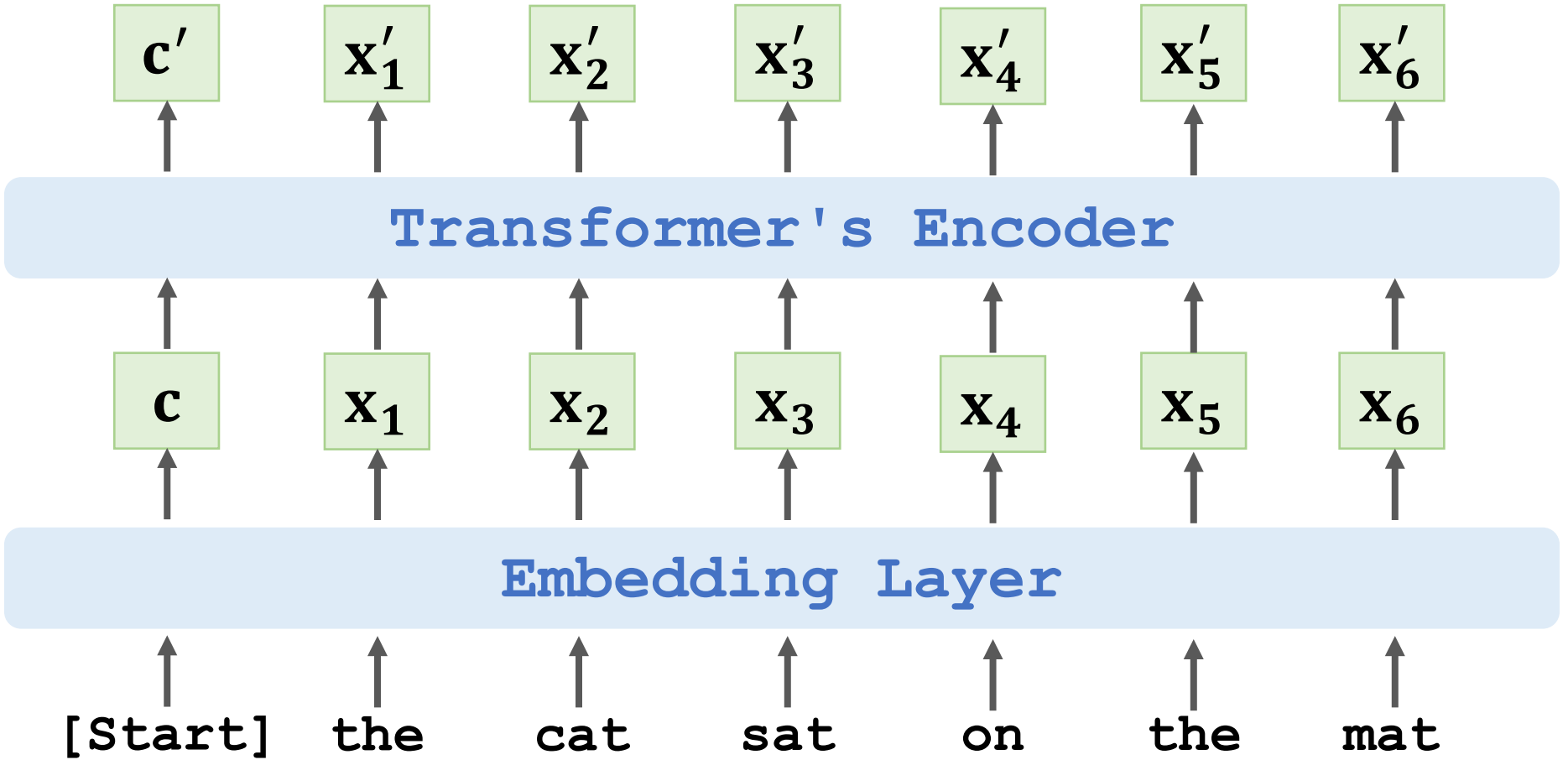
Revisit Transformer's Encoder

[Start] the cat sat on the mat

Revisit Transformer's Encoder



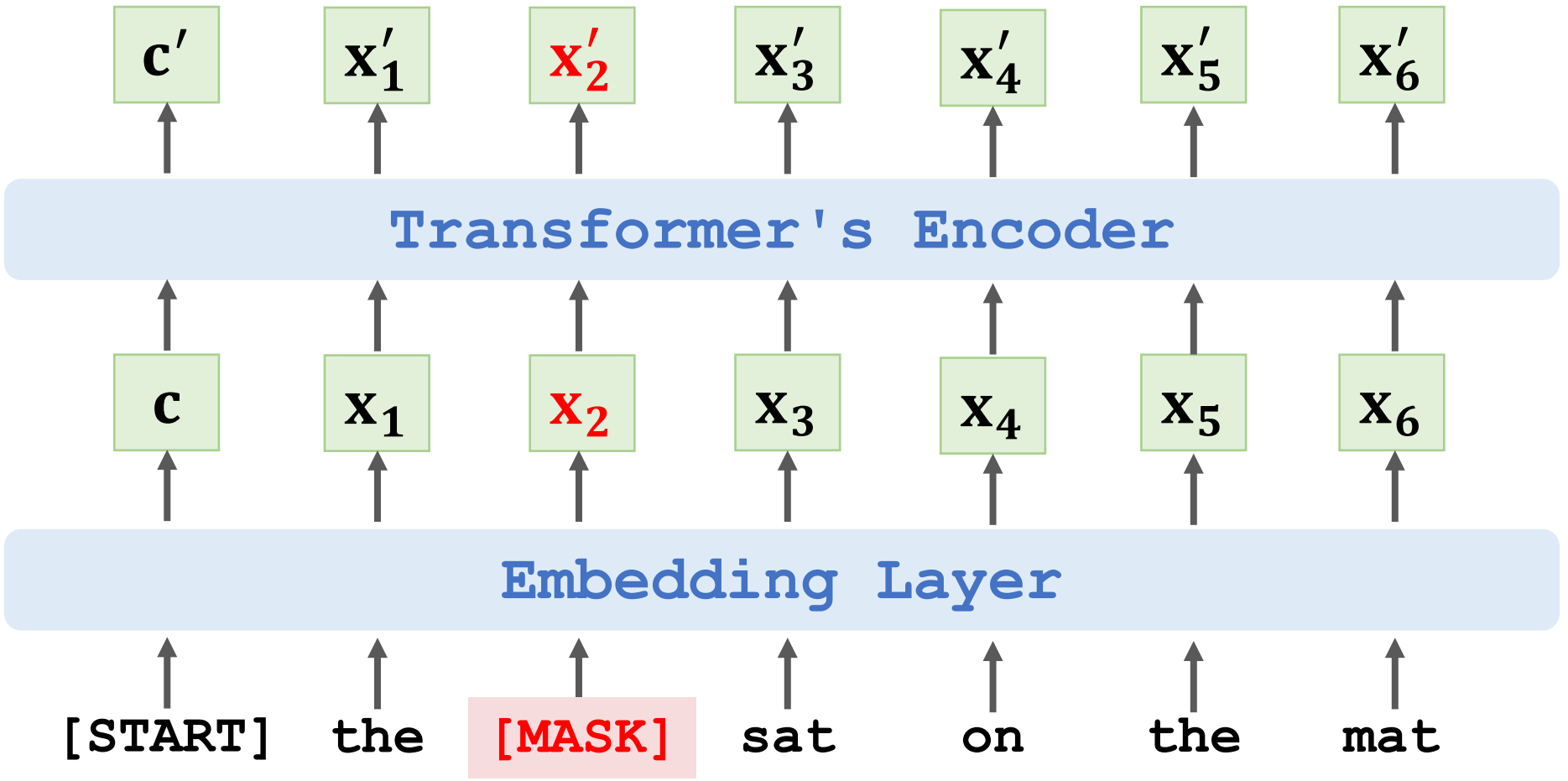
Revisit Transformer's Encoder

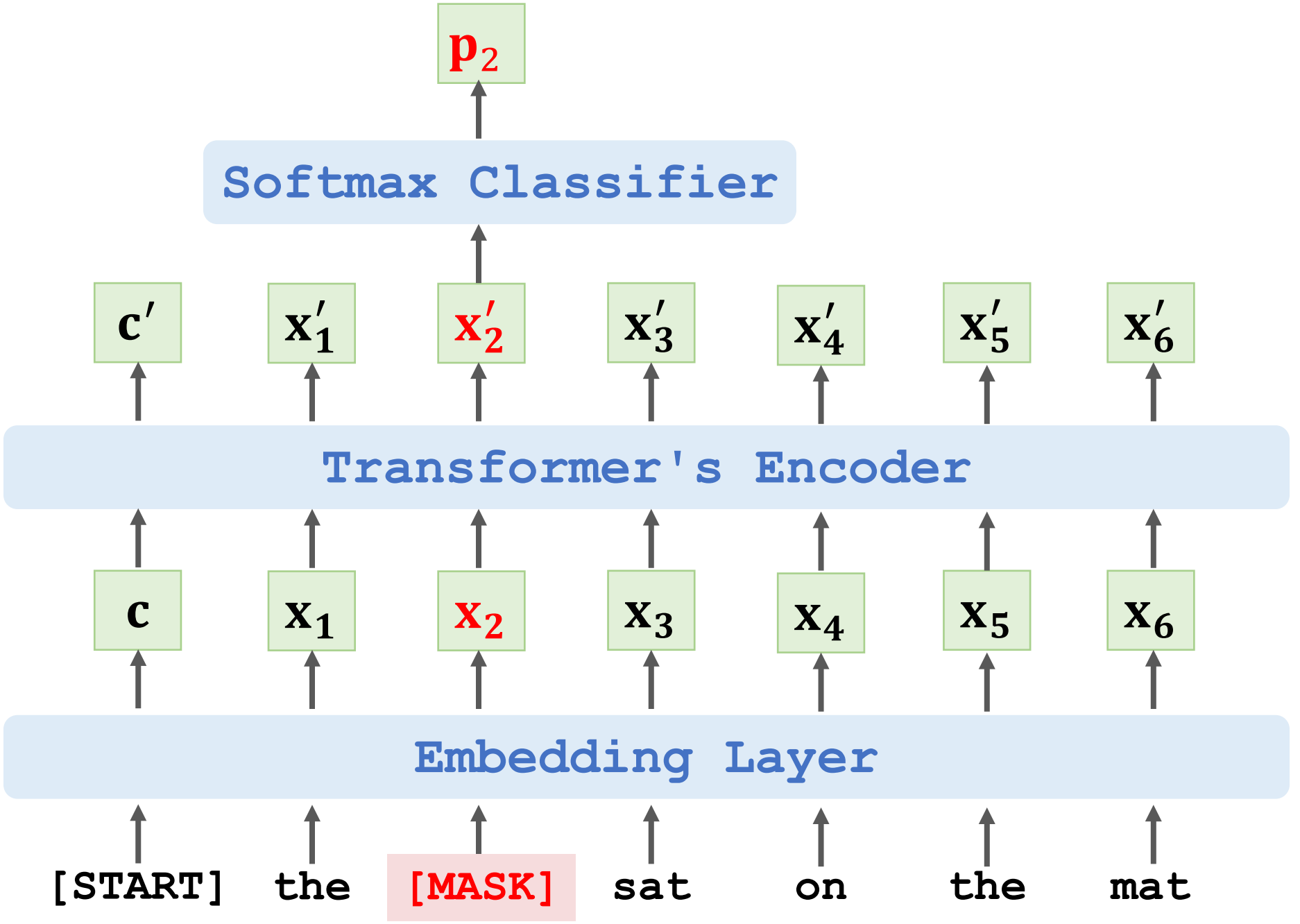


Randomly mask a word

- "The _____ sat on the mat"
- What is the masked word?

Randomly mask a word




p_2
Softmax Classifier
c'
x'_1
x'_2
x'_3
x'_4
x'_5
x'_6
Transformer's Encoder
c
x_1
x_2
x_3
x_4
x_5
x_6
Embedding Layer
[START]
the
[MASK]
sat
on
the
mat

Predict the masked word

- $\mathbf{e}_2$: one-hot encode of the masked word "cat".
- $\mathbf{p}_2$: output probability distribution at the masked position.
- $\text{Loss} = \text{CrossEntropy}(\mathbf{e}_2, \mathbf{p}_2)$.
- Performing one gradient descent to update the layers' parameters.

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"

- Is this the next sentence?

  "panda is native to south central china"

Input Representation

- **Input:**

"[START] calculus is a branch of math [SEP] it was developed by newton and leibniz".

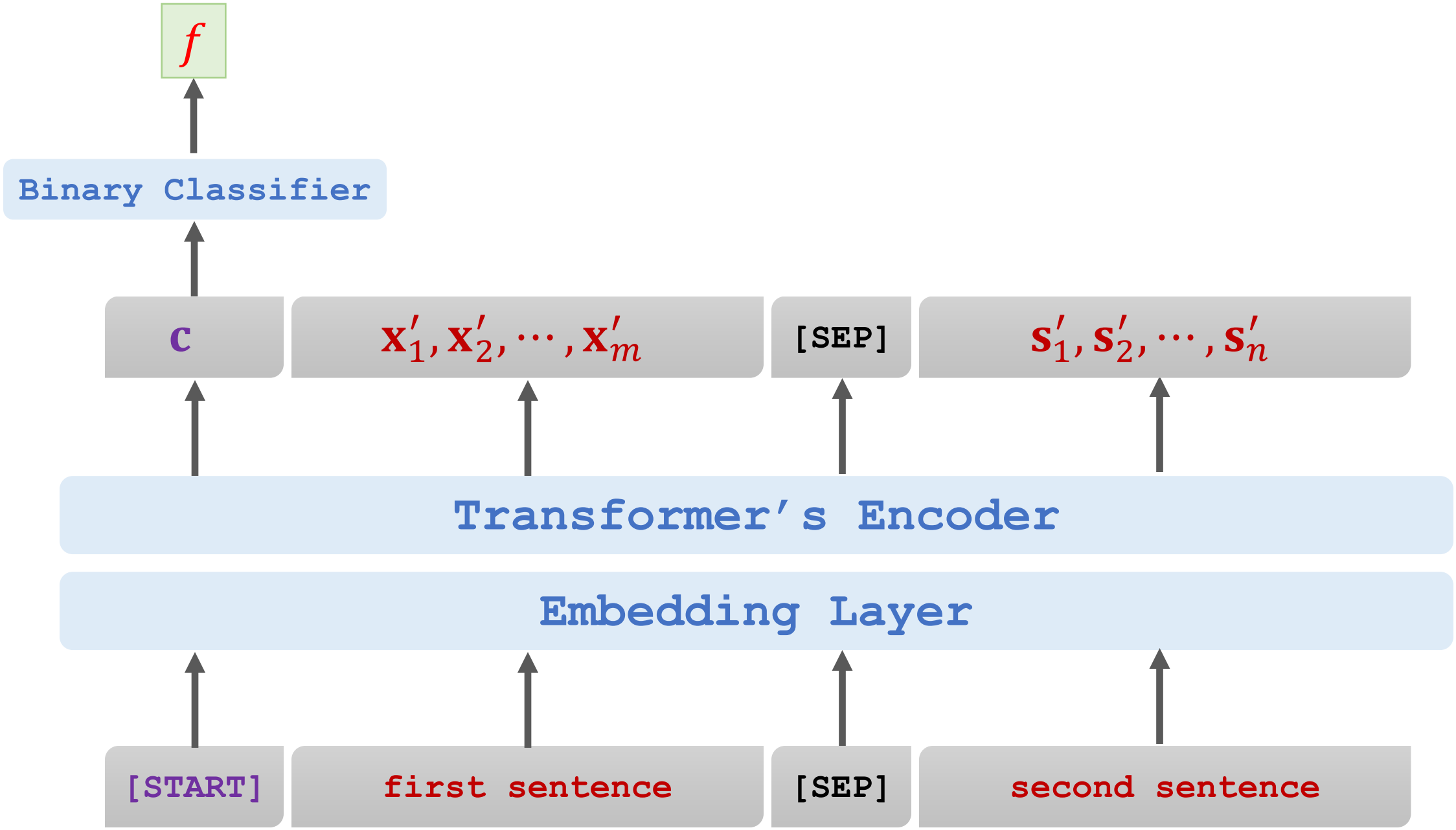
- **Target:** true

Input Representation

- **Input:**

  "[START] calculus is a branch of math [SEP] panda is native to south central china".

- **Target:** false


f
Binary Classifier
$\mathbf{c}$
$\mathbf{x}'_1, \mathbf{x}'_2, \cdots, \mathbf{x}'_m$
[SEP]
$\mathbf{s}'_1, \mathbf{s}'_2, \cdots, \mathbf{s}'_n$
Transformer's Encoder
Embedding Layer
[START]
first sentence
[SEP]
second sentence

Combining the two methods

Input Representation

- **Input:**

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

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

[START]	masked first sentence	[SEP]	masked second sentence

Input Representation

- **Input:**

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

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

[START]	masked first sentence	[SEP]	masked second sentence

Training

- Loss 1 is from the binary classification (for the next sentence.)
- Loss 2 is from the prediction of the masked word (in the first sentence.)
- Loss 3 is from the prediction of the masked word (in the second sentence.)
- Minimize the sum of the three losses.

Data

- Use large-scale data, e.g., English Wikipedia (2.5 billion words.)
- 50% of the next sentences are real. (The other 50% are fake.)
- Randomly mask words (with some tricks.)

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.)