CS-583: Deep Learning
BERT Model

Abdul Rafae Khan

Department of Computer Science
Stevens Institute of Technology
akhan4@stevens.edu

November 2, 2023

What is BERT?

- BERT [1] is for pre-training Transformer's [2] encoder.

- How?

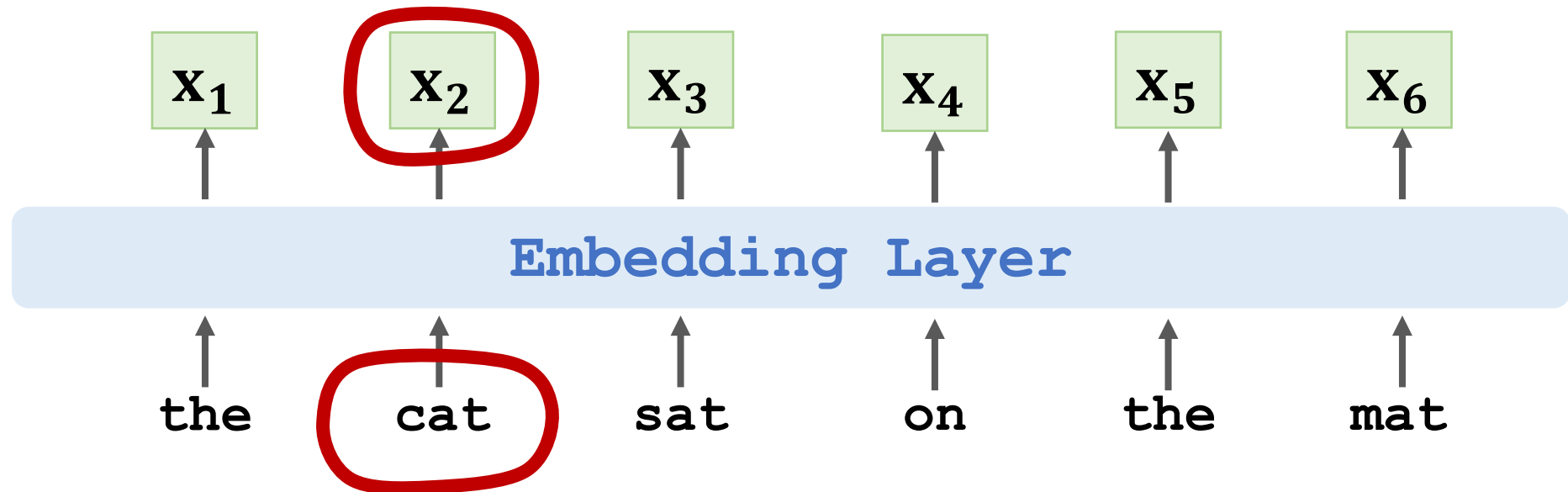- Predict masked word.

- 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.
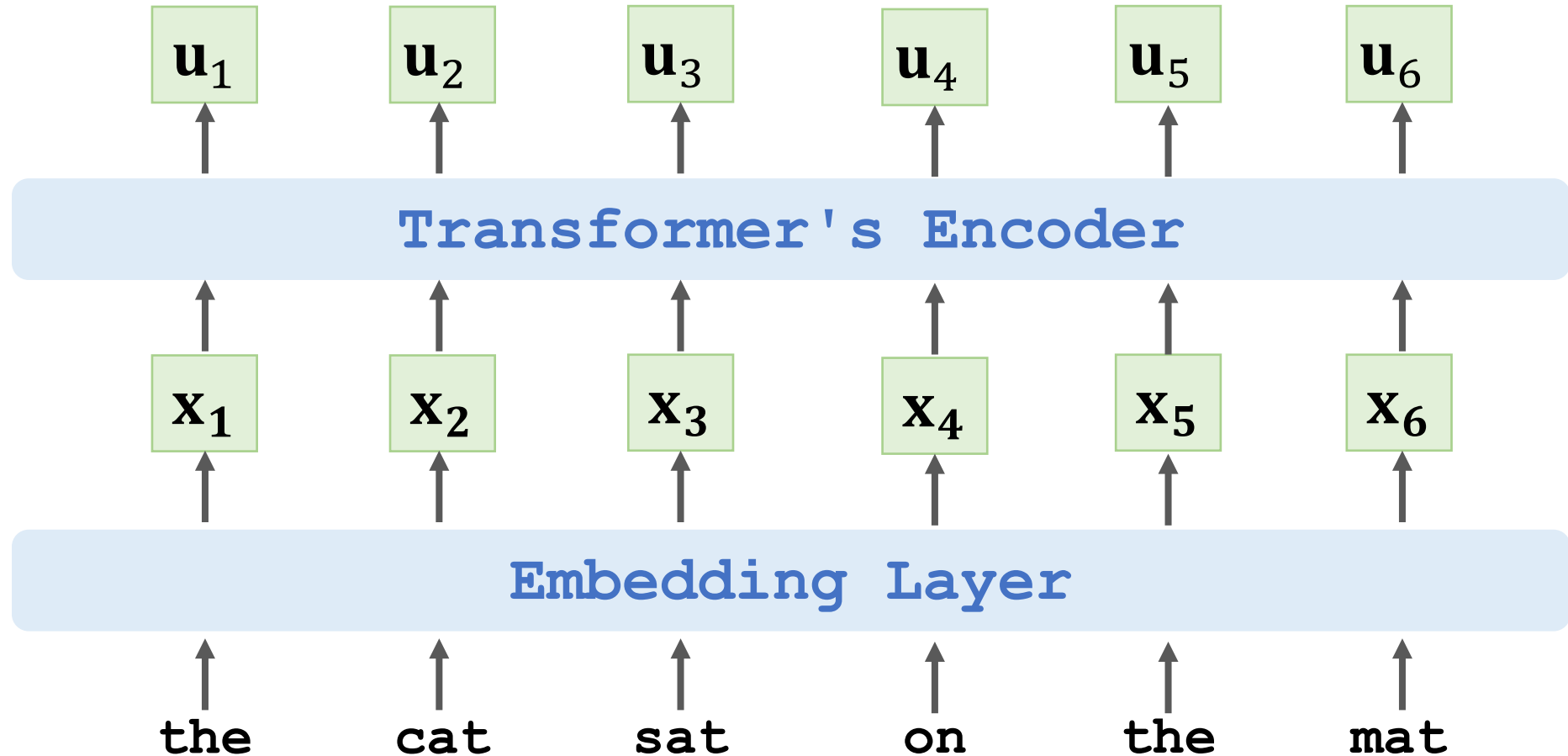
Task 1: Predict Masked Words

Revisit Transformer's Encoder
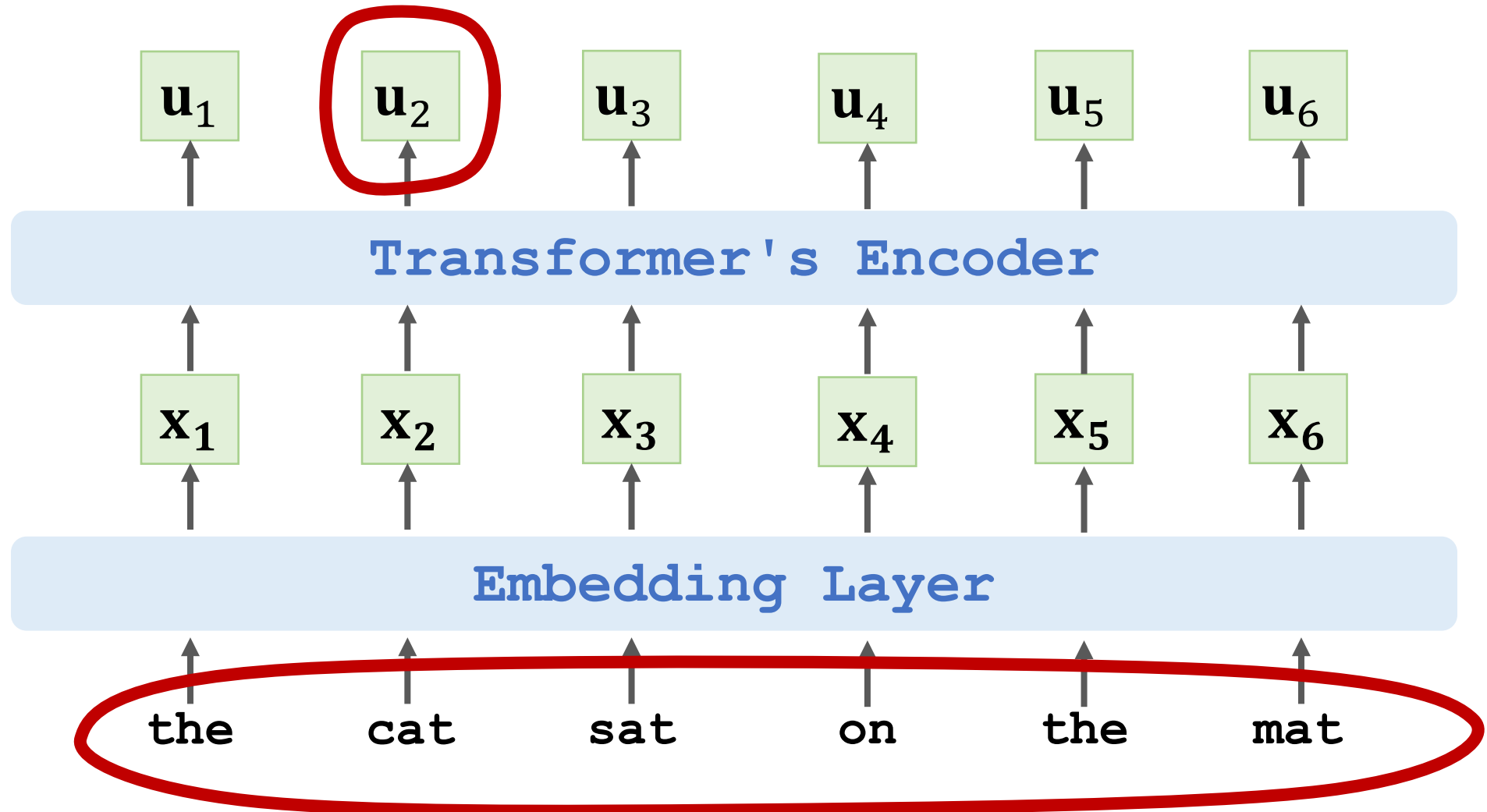
the   cat   sat   on   the   mat

Revisit Transformer's Encoder
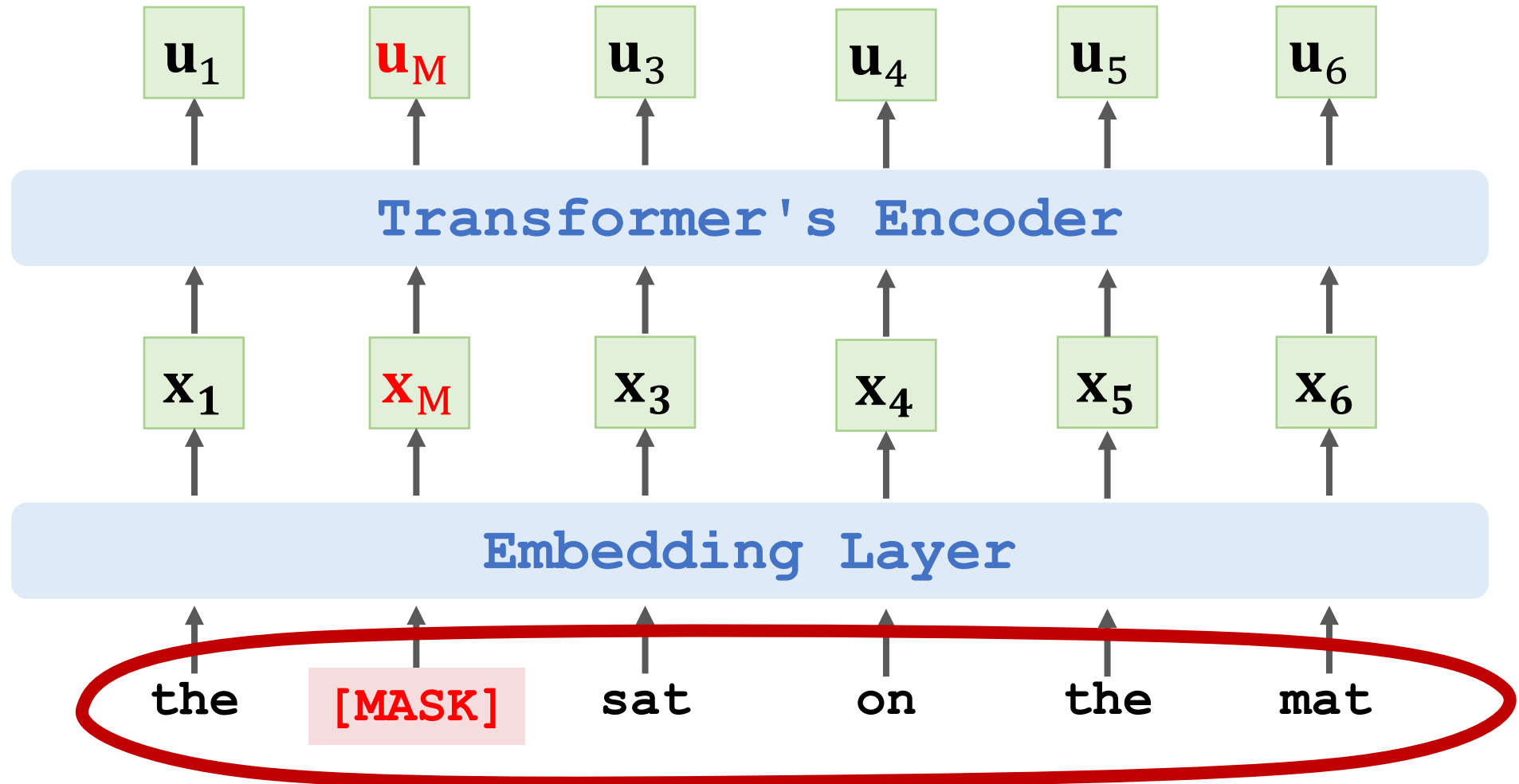
Revisit Transformer's Encoder

Revisit Transformer's Encoder

$\mathbf{u}_1$ $\mathbf{u}_2$ $\mathbf{u}_3$ $\mathbf{u}_4$ $\mathbf{u}_5$ $\mathbf{u}_6$

Transformer's Encoder

$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$ $\mathbf{x}_4$ $\mathbf{x}_5$ $\mathbf{x}_6$

Embedding Layer

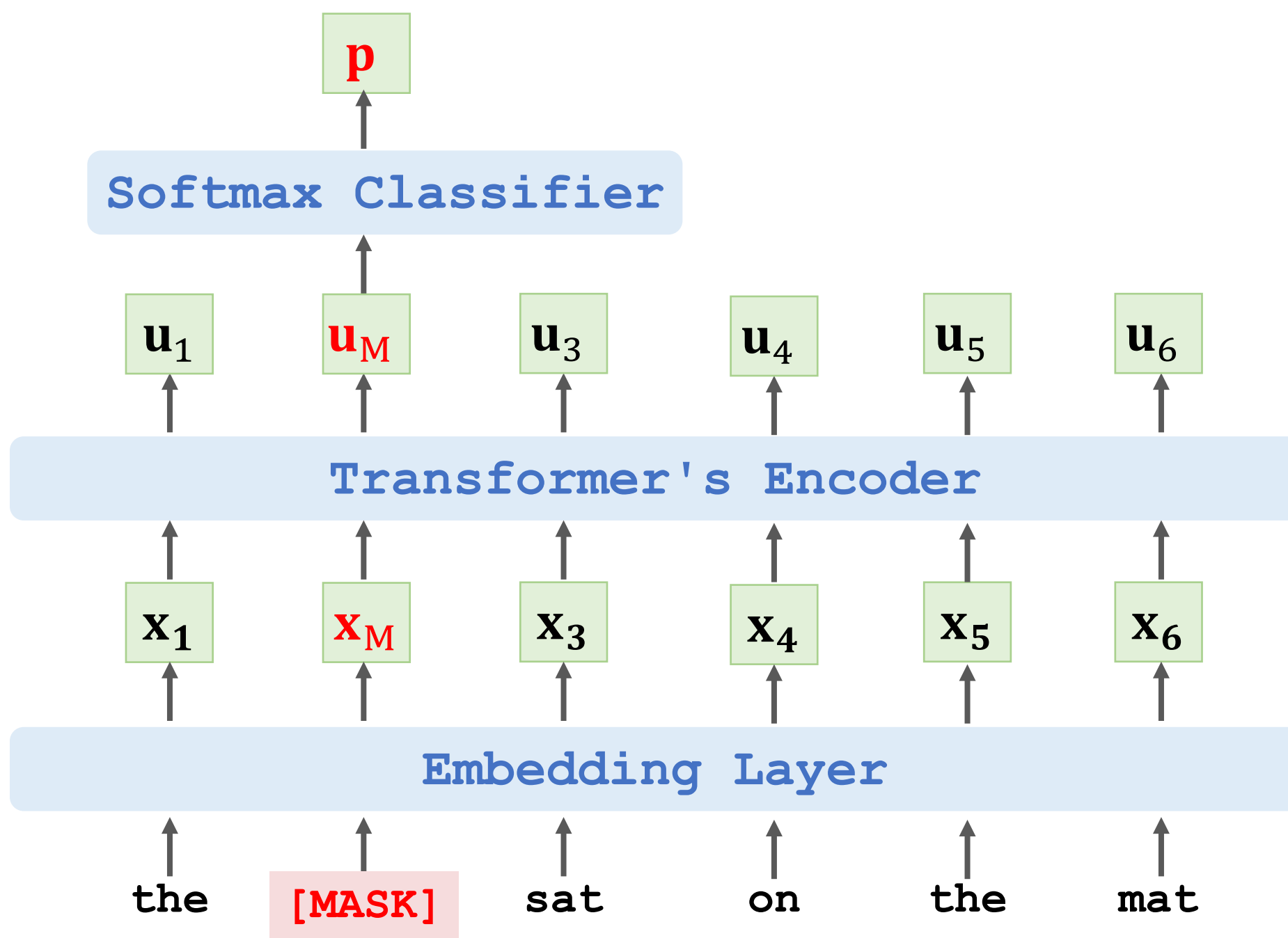the cat sat on the mat

Randomly mask a word

- "The _____ sat on the mat"

- What is the masked word?

Randomly mask a word

Predict the masked word

- **e**: one-hot vector of the masked word "cat".

- **p**: output probability distribution.

- Loss = CrossEntropy($\mathbf{e}, \mathbf{p}$).

- Performing one gradient descent to update the model parameters.

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"

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
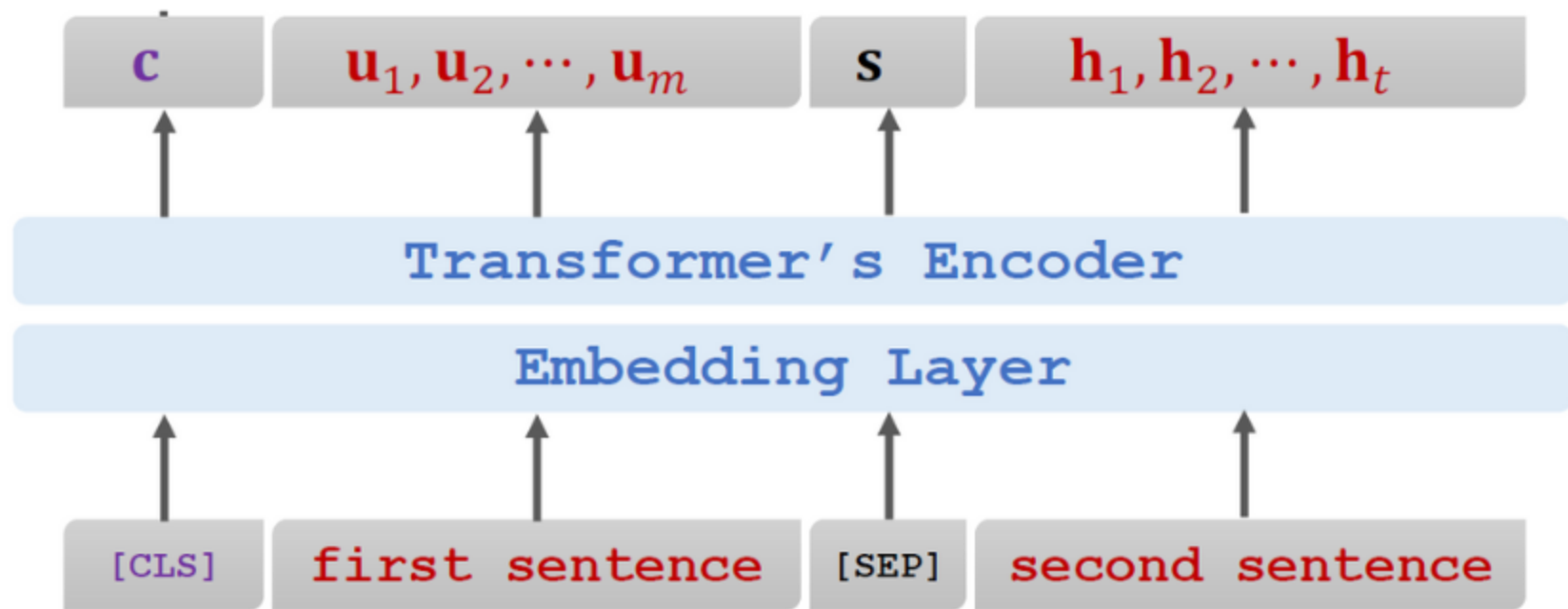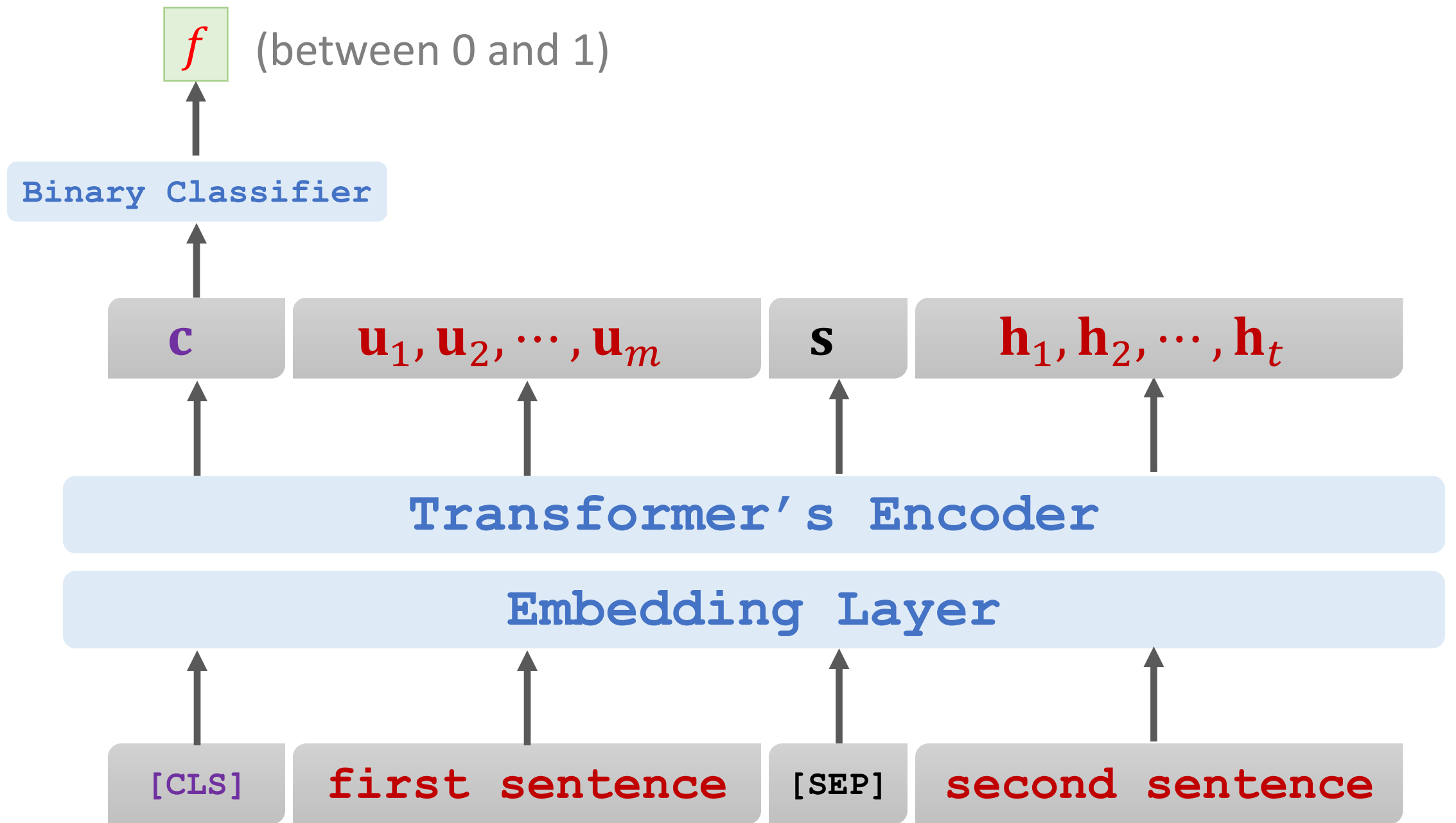
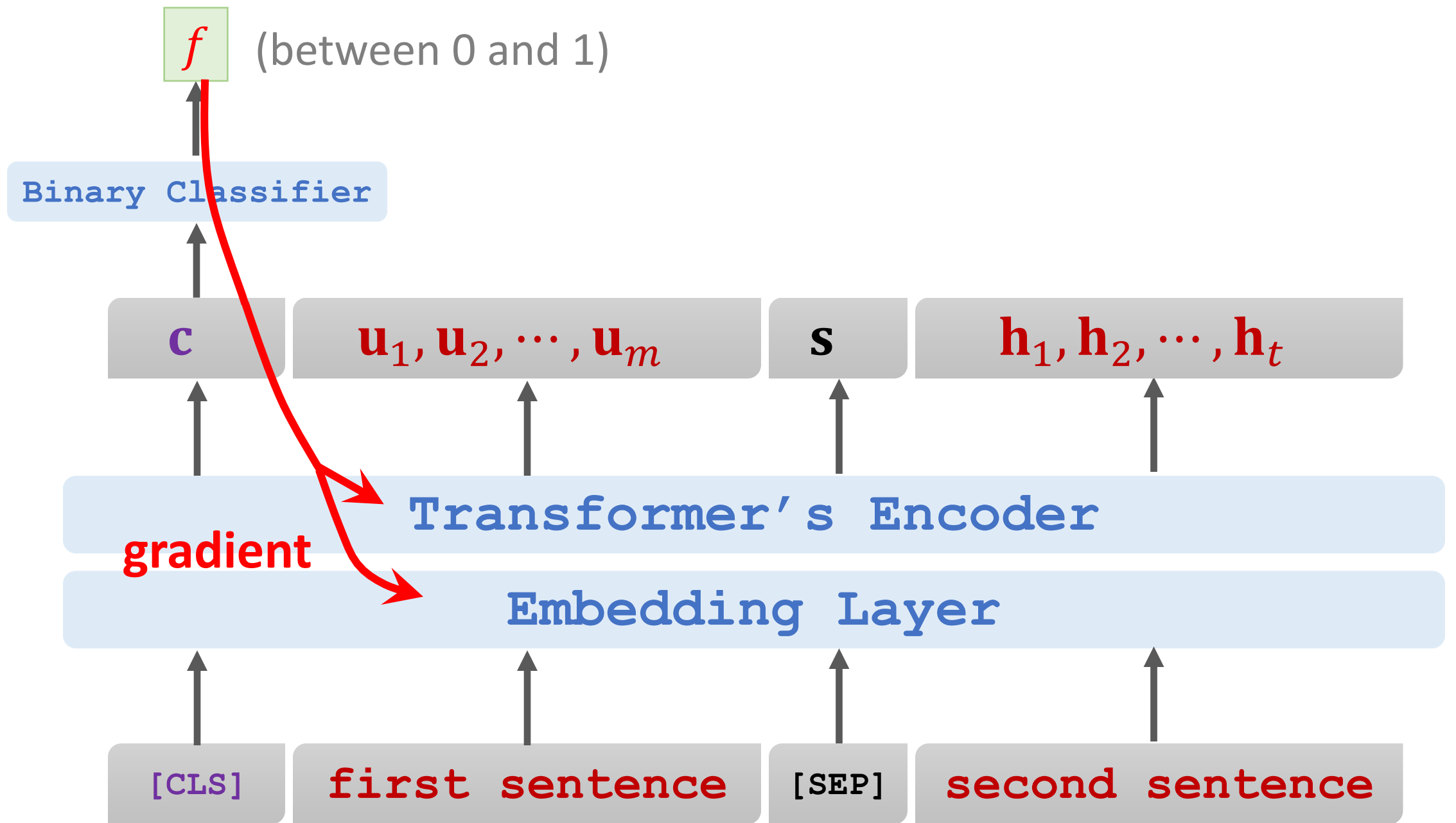# Predict the next sentence

[CLS] **first sentence** [SEP] **second sentence**

Predict the next sentence

Why is the classification task helpful?

- The task helps improve the embedding layer.

- The task helps improve the self-attention layers.

Combining the two methods

Input Representation

- **Input:**

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

Input Representation

- **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., deciding if the next sentence is real.)

- Loss 2 and Loss 3 are for multi-class classification (i.e., predicting the masked words.)

- Objective function is the sum of the three loss functions.

- Update model parameters by performing one gradient descent.

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 15% 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.)

The End