

BERT Training Objective: Mathematics

Masked Language Modeling Objective

$$\mathcal{L}_{MLM} = - \sum_{i \in \text{masked}} \log P(w_i | \text{context})$$

Masked
tokens
(15%)

Cross-entropy
loss

Conditional
probability

Full sentence
context

Example:

"The [MASK] sat" $\rightarrow P(\text{cat}) = 0.73$

$$\text{Loss} = -\log(0.73) = 0.31$$