MLM pre-training objective

Transformer Encoder Block L_c

Transformer Encoder Block 2

Transformer Encoder Block 1

Segment Embedding

h_1 h_2 h_3 h_4 h_5 $\Longleftrightarrow$ h_1 h_2 h_3 h_4 h_5

Concat

Transformer Encoder Block L

Transformer Encoder Block 1

Transformer Encoder Block L

Transformer Encoder Block 1

Semantic Embedding $\Longleftrightarrow$ Semantic Embedding

Shared

X_1 X_2 X_3 X_4 X_5 Y_1 Y_2 Y_3 Y_4 Y_5

The masked operation list of X

The masked operation list of Y

Computationally similar architecture pair (X, Y)

(X, Y)

(X, X)

(Y, X)

(Y, Y)

Cross-attention mask for variable length input pairs

(Augmented) adjacency matrix for masked self-attention