A METHODS FOR REPRESENTATIONS

We evaluate different representations method in Transformer-base models, including CMLM and BERT base (using the model on official Tensorflow Hub). The experiments are conducted on SentEval. Results in Table 8 show that MEAN representation exhibit better performance than CLS and MAX representations.

Model	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	SICK-E	SICK-R	Avg.
CMLM MAX	82.8	88.9	96.2	89.2	87.81	89.8	72.1	82.1	83.7	85.8
CMLM MEAN	83.6	89.9	96.2	89.3	88.5	91.0	69.7	82.3	83.4	**86.0**
CMLM CLS	79.1	84.3	94.2	86.9	84.9	82.6	68.4	79.3	81.7	82.4
BERT base MAX	79.6	85.5	94.6	87.3	83.0	90.0	65.6	75.5	78.1	82.1
BERT base MEAN	81.6	87.4	95.2	87.8	85.8	90.6	71.1	79.3	80.5	**84.3**
BERT base CLS	79.9	83.9	93.8	85.4	86.1	81.0	69.5	62.5	48.8	76.8

Table 8: Performance of sentence representations model with different representations method (MAX, MEAN and CLS).

B EXPERIMENTS WITH DIFFERENT MASKING RATIOS

We test with different masking ratios in CMLM training data. Specifically, We tried masking 40, 60, 80and 100 tokens of 256 tokens in the CMLM data. Performance of obtained models on SentEval are presented in Table 9.

Mask Tokens	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	SICK-E	SICK-R	Avg.
40	81.8	89.3	95.3	87.8	87.0	90.2	68.5	77.5	77.6	83.9
60	**83.7**	89.5	95.8	88.9	88.0	90.3	68.7	79.5	82.8	85.4
80	83.6	**89.9**	**96.2**	**89.3**	**88.5**	**91.0**	69.7	**82.3**	**83.4**	**86.0**
100	83.2	89.5	95.5	88.7	88.0	90.8	**70.0**	81.5	82.7	85.6

Table 9: Performance with different masking ratios in data (X-out-of-256) of CMLM base on SentEval.

C TRAINING AND IMPLEMENTATION DETAILS

Projection P in the CMLM modeling. Let h denote the dimension of the input sentence vector (e.g. h = 768 in BERT base; h = 1024 in BERT large). Let $FC(h_1, h_1, n)$ denote a fully connected layer with input dimension h_1, output dimension h_2 and nonlinearity function n. The three layers are $FC(h, 2 \times h, \text{ReLU})$, $FC(2 \times h, 2 \times h, \text{ReLU})$, $FC(2 \times h, h, \text{None})$.