

Selective Masking

Feb 10 2021

Most content is borrowed from <https://virtual.2020.emnlp.org/>

Train No Evil: Selective Masking for Task-Guided Pre-Training (Gu et al., EMNLP'20)

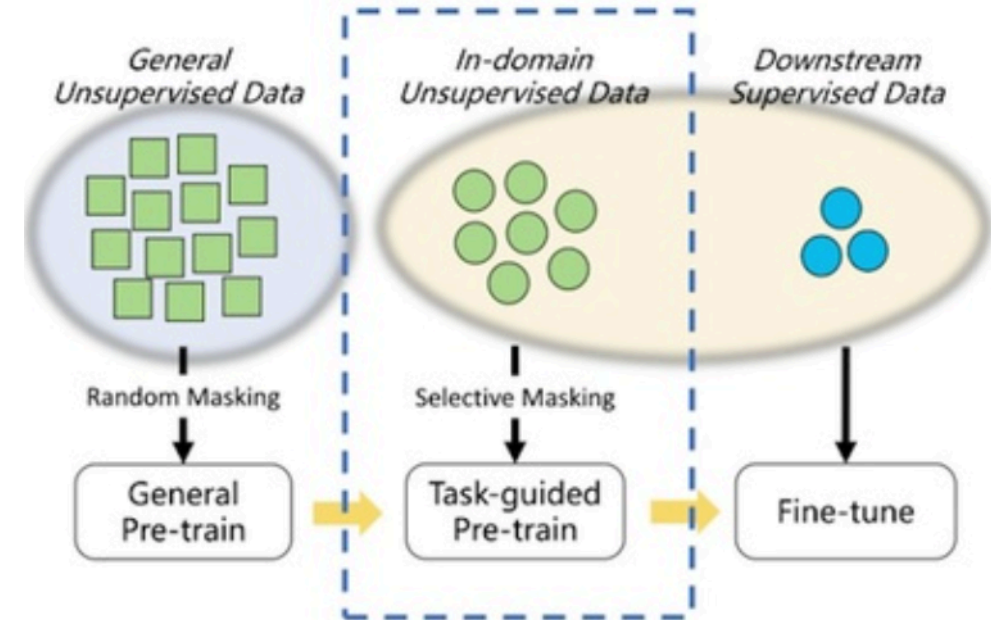
- Standard pipeline
 - Pre-training on large-scale unsupervised general domain data
 - Fine-tuning on small-scale supervised downstream data
- Mid-scale in-domain unsupervised data

Wikipedia & Bookcorpus	~1000M words
Unlabeled Yelp data	~10M words
Restaurant review classification	~10K words

- Task specific patterns
 - *Sentiment Classification*: Constantly **touching**, an exploration of the **creative** act.
 - *Relation extraction*: Bob Dylan **wrote** Blowin' in the Wind in 1962.

Train No Evil: Selective Masking for Task-Guided Pre-Training (Gu et al., EMNLP'20)

- Random masking strategy in MLM
 - k% of randomly picked input tokens for prediction (k=15, for BERT)
 - Replace i-th token
 - With [MASK] 80% of the time
 - Random token 10% of the time
 - Unchanged i-th token 10% of the time
 - Aimless and inefficient for capturing domain specific and task specific data
- Task-guided pre-train
 - Selectively mask important words on in-domain unsupervised data for specific tasks
 - Pre-train model on in-domain data to learn domain-specific and task-specific patterns



Train No Evil: Selective Masking for Task-Guided Pre-Training (Gu et al., EMNLP'20)

We propose a simple method to find important tokens of D_{Task} . Given the n -token input sequence $s = (w_1, w_2, \dots, w_n)$, we use an auxiliary sequence buffer s' to help evaluating these tokens one by one. At time step 0, s' is initialized to empty. Then, we sequentially add each token w_i to s' and calculate the task-specific score of w_i , which is denoted by $S(w_i)$. If the score is lower than a threshold δ , we regard w_i as an important token. Note that we will remove previous important tokens from s' to make sure the score is not influenced by previous important tokens.

Assume the buffer at the time step $i - 1$ is s'_{i-1} . We define the token w_i 's score as the difference of classification confidences between the original input sequence s and the buffer after adding w_i , which is denoted by $s'_{i-1}w_i$:

$$S(w_i) = P(y_t | s) - P(y_t | s'_{i-1}w_i), \quad (1)$$

where y_t is the target classification label of the input s and $P(y_t | *)$ is the classification confidence computed by a PLM fine-tuned on the task. Note that the PLM used here is the model with GenePT introduced in Section 2.1, not a fully pre-trained PLM. In experiments, we set $\delta = 0.05$. The important token criterion $S(w_i) < \delta$ means after adding w_i , the fine-tuned PLM can correctly classify the incomplete sequence buffer with a close confidence to the complete sequence.

The food tastes good here . \rightarrow  $\rightarrow P(\text{positive} | \text{The food tastes good here .})$

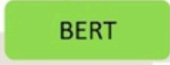
$$S(\text{good}) = P(\text{positive} | \text{The food tastes good here .}) - P(\text{positive} | \text{The food tastes good})$$

$S(\text{good}) < \delta \rightarrow$ Important, mask "good" in the sentence

• In-Domain Mask

- Train a model to learn the distribution of important tokens.
- Use the model to annotate important words on unsupervised in-domain data.

The food tastes **good** here .
The movie is **impressive**.
I am **disappointed** in this laptop. \rightarrow  \rightarrow Predict whether a token is important.

The shirt looks good in this store \rightarrow  \rightarrow The shirt looks **good** in this store

Train No Evil: Selective Masking for Task-Guided Pre-Training (Gu et al., EMNLP'20)

		MR	Sem14-Rest
w/o Task-guided pre-training		87.37	88.60
Amazon	Random	88.35	90.40
	Selective	89.51**	91.56**
Yelp	Random	87.20	90.70
	Selective	88.15**	91.87*

Table 1: Test accuracies of models trained with different methods (without task-guided pre-training or task-guided pre-training with different masking strategies) after full general pre-training (1M steps). * and ** indicate statistically significant ($p < .05$ and $p < .001$).