# Weighted Sampling For Masked Language Modeling

Linhan Zhang[1] Qian Chen[2] Wen Wang[2] Chong Deng[2]
Yuxin Jiang[1] Kongzhang Hao[1] Wei Wang[3] Xin Cao[3]

[1]University of New South Wales, School of Computer and Engineering
[2]Speech Lab of DAMO Academy, Alibaba Group, China
[3]Hong Kong University of Science and Technology (Guangzhou), China

# Masked Language Model

達摩院
ALIBABA DAMO ACADEMY

Given a sentence $S = \{t\_1, t\_2, \ldots, t\_n\}$

- Standard Masking Strategy
  - Randomly chooses 15% of tokens to mask
  - 10% of the time replaced by a random token from corpus
  - 10% of the time remains unchanged
  - 80% of the time replaced by a special token [MASK]
- Objective
  - The language model must learn to predict the masked tokens with bidirectional context
- Use Cases
  - Helps understand the contextual relationships between words
  - Can be used for various natural language processing tasks such as text classification, question answering, and named entity recognition

# Motivation - Addressing the Frequency Bias Issue 達摩院

ALIBABA DAMO ACADEMY

- Frequency Bias in the Standard Masking Strategy
  - High-frequency tokens are masked frequently
  - More informative tokens with lower frequencies are masked much less frequently during pre-training
  - This greatly harms the efficiency of pre-training

**Input Text:**

In March 768 , he began his journey again and got as far as Hunan province , where he died in ( now Changsha ) in November or December 770 , in his 58th year . He was survived by his wife and two sons , who remained in the area for some years at least . His last known descendant is a grandson who requested a grave inscription for the poet from Yuan Zhen in 813 .

**Fig. 1.** An example from WikiText. Randomly selected tokens are in blue while Frequency Weighted Sampled tokens are in pink.

# Proposed Method

**Weighted Sampling:** masking tokens based on **(1) token frequency** or **(2) training loss**

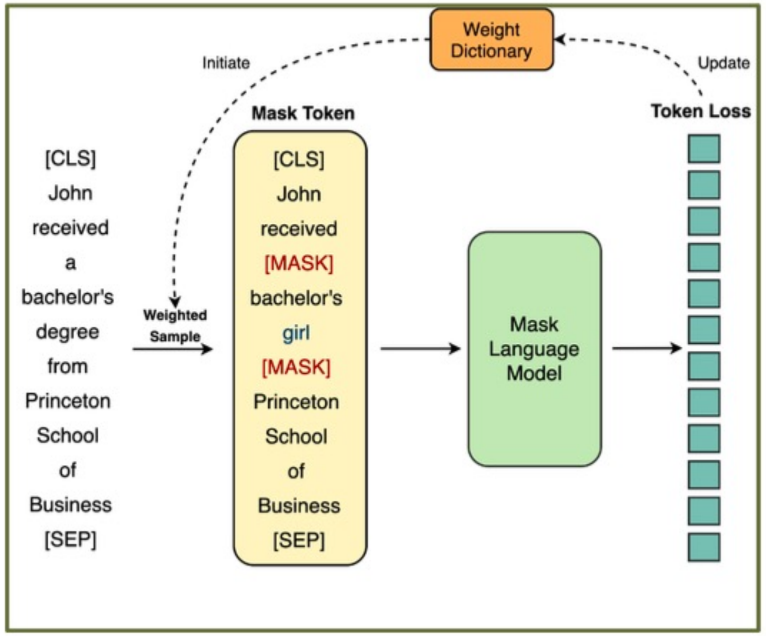


**Fig. 2.** Illustration of the proposed **Dynamic Weighted Sampling** for mask language modeling (MLM). The sampling weight of choosing a token to mask is computed based on the prediction loss of this token by the current PLM. We store the sampling weights of each token in the weight dictionary.

# Weighted Sampling Strategy

- Method 1: Frequency Weighted Sampling
  - Step 1: Remove the influences of extremely rare tokens

$$\mathrm{freq}^*(w) = \begin{cases} \mathrm{freq}(w) & \text{, if } \mathrm{freq}(w) > \theta \\ \theta & \text{, otherwise.} \end{cases} \quad (1)$$

  - Step 2: Compute Sample Weight $wt(w)$ for $w$

$$\mathrm{wt}(w) = (\mathrm{freq}^*(w))^{-\alpha} \quad (2)$$

  - Step 3: Compute Sample Probability $p(t\_i)$ for token $t\_i$ in sentence $S$

$$p(t_i) = \frac{wt(t_i)}{\sum_{j=1}^{n} wt(t_j)} \quad (3)$$

# Weighted Sampling Strategy

達摩院 ALIBABA DAMO ACADEMY

- Method 2: Dynamic Weighted Sampling
  - Step 1: Initialize Sampling Weight
    - $wt(t\_i) = 1$ for each token $t\_i \in T$ in the weight dictionary
    - $T$ denotes all tokens in the pre-training dataset
  - Step 2: Compute Total Cross-Entropy Loss for token $t\_i$

$$L_{t_i} = -logP(t_i \mid x, \theta) \quad (4)$$

  - Step 3: Compute Sampling Weight $wt(t\_i)$
    - Compute sampling weight for each token based on its prediction loss by the current pre-trained language model
    - Store these sampling weights in the weight dictionary

$$wt(t_i) = exp(\frac{L_{t_i}}{\tau}) \quad (5)$$

  - Step 4: Compute Sampling Probability $p(t\_i)$
    - Normalize $wt(t\_i)$ to obtain the sampling probability for each token $t\_i$

# Experiments - Semantic Textual Similarity

達摩院
ALIBABA DAMO ACADEMY

- **Objective:** To evaluate unsupervised sentence representation on STS tasks
- **Evaluation Metric:** Spearman's correlation coefficient between the predicted similarity and the gold standard similarity scores
- **Baselines**
  - BERT: bert-base-uncased
  - BERT-CP: continue pre-training on BERT with **random sampling** on the Wiki-Text
- **Proposed Method**
  - WSBERT_Freq: continue pre-training on BERT with **Frequency Weighted Sampling** on the Wiki-Text
  - WSBERT_Dynamic: continue pre-training on BERT with **Dynamic Weighted Sampling** on the Wiki-Text

# Experiments - Semantic Textual Similarity

| Method | STS12 | STS13 | STS14 | STS15 | STS16 | STS-B | SICK-R | Avg. |
|---|---|---|---|---|---|---|---|---|
| BERT | 39.70 | 59.38 | 49.67 | 66.03 | 66.19 | 53.87 | 62.06 | 56.70 |
| BERT-CP | 41.00 | 60.02 | 51.11 | 68.43 | 64.59 | 56.32 | 62.07 | 57.65 |
| WSBERT_Freq | 42.60 | 61.32 | 52.04 | 69.84 | 66.61 | 59.89 | 61.94 | 59.18 |
| WSBERT_Dynamic | 47.80 | 67.28 | 57.13 | 71.41 | 68.87 | 65.28 | 64.90 | 63.24 |
| BERT-Whitening | 54.28 | **78.07** | **65.44** | 64.83 | 70.16 | 71.43 | 62.23 | 66.43 |
| WSBERT-Whitening | 55.14 | **78.45** | **66.13** | 65.47 | **70.68** | **71.98** | 61.91 | 67.10 |
| BERT + Prompt† | **60.96** | 73.83 | 62.18 | **71.54** | 68.68 | 70.60 | **67.16** | **67.85** |
| WSBERT + Prompt | **63.03** | 71.66 | 63.80 | **75.32** | **76.67** | **74.79** | **65.32** | **70.08** |

- Findings
  - Weighted sampling methods, WSBERT_Freq and WSBERT_Dynamic, outperform the baselines (BERT and BERT-CP)
  - For instance, WSBERT_Dynamic outperforms BERT and BERT-CP by **6.54** and **5.59** absolute points respectively
  - WSBERT_Dynamic can be effectively combined with Whitening and Prompt to further improve performance

# Experiments - GLUE Evaluation

達摩院
ALIBABA DAMO ACADEMY

- **Purpose:** to evaluate transfer learning capability
- **Findings**
  - WSBERT achieves the best average GLUE score compared to BERT and BERT-CP, outperforming BERT by **0.52** absolute
  - BERT-CP degrades GLUE AVG by 0.35 absolute compared to BERT
  - WSBERT outperforms BERT- CP by 0.87 absolute
  - The gain of WSBERT over BERT is from continual pre-training with Dynamic Weighted Sampling, not from continual pre-training

STS and GLUE evaluations demonstrate that *Dynamic Weighted Sampling improves the transfer learning capability while enhancing sentence representations.*

| Dataset | BERT | BERT-CP | WSBERT |
|---|---|---|---|
| MNLI | $84.30_{\pm 0.26}$ | $84.26_{\pm 0.19}$ | $\mathbf{84.42}_{\pm \mathbf{0.35}}$ |
| QQP | $91.31_{\pm 0.04}$ | $90.94_{\pm 0.59}$ | $\mathbf{91.43}_{\pm \mathbf{0.05}}$ |
| QNLI | $\mathbf{91.47}_{\pm \mathbf{0.01}}$ | $91.32_{\pm 0.17}$ | $91.14_{\pm 0.17}$ |
| SST-2 | $\mathbf{92.86}_{\pm \mathbf{0.13}}$ | $92.78_{\pm 0.43}$ | $91.35_{\pm 0.47}$ |
| CoLa | $56.47_{\pm 0.65}$ | $57.44_{\pm 0.95}$ | $\mathbf{58.29}_{\pm \mathbf{0.33}}$ |
| STS-B | $89.68_{\pm 0.26}$ | $89.52_{\pm 0.37}$ | $\mathbf{89.86}_{\pm \mathbf{0.18}}$ |
| MRPC | $86.13_{\pm 1.63}$ | $85.13_{\pm 0.53}$ | $\mathbf{88.20}_{\pm \mathbf{2.39}}$ |
| RTE | $69.23_{\pm 0.4}$ | $67.25_{\pm 1.84}$ | $\mathbf{70.89}_{\pm 0.17}$ |
| **AVG** | $82.68_{\pm 0.33}$ | $82.33_{\pm 0.32}$ | $\mathbf{83.20}_{\pm 0.10}$ |

**Table 2**. GLUE Validation results from *BERT-base-uncased* (BERT-base), *BERT-base-uncased* continually pre-trained (BERT-CP), and Weighted-Sampled BERT (WSBERT). BERT-CP and WSBERT both continually train on BERT with the same training settings. WS-BERT refers to WSBERT_Dynamic. The best results for each dataset and AVG are in bold.

# Takeaway and Future work

達摩院
ALIBABA DAMO ACADEMY

- Proposed two Weighted Sampling methods to address the frequency bias issue in conventional masked language modeling
- Developed a new PLM, WSBERT, by applying Weighted Sampling to BERT
- WSBERT outperforms BERT in both sentence representation quality and transfer learning capability
- Future work includes investigating other dynamic sampling methods and exploring training objectives with a penalty for frequency bias