IMPROVING SELF-SUPERVISED PRE-TRAINING *via* A FULLY-EXPLORED MASKED LANGUAGE MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Masked language model (MLM) framework has been widely adopted for self-supervised language pre-training. Given a training sequence, MLM randomly masks some tokens for prediction. In this paper, we observe that this random behavior will result in a undesirable high gradient variance. Thus, we first theoretically quantify the gradient variance during training via correlating the gradient covariance with the hamming distance between two different masked sequences. To address the high gradient variance introduced by the standard random masking, we propose a *fully-explored* masking strategy to make sure each token have equal opportunity to be masked and predicted using the remaining context. For an efficient implementation, we assign a segment id for each token and divide the entire sequence into multiple discrete segments, so we can mask all tokens in a segment and use the remaining tokens from other segments as context for their predictions. We prove, from a theoretical perspective, that the gradients derived from this new masking schema have a smaller variance and can lead to more efficient pre-training. We conduct extensive experiments on both continual pre-training and general pre-training from scratch. Empirical results continue to confirm that this new masking strategy can consistently outperform standard random masking. Detailed efficiency analysis and ablation studies further validate the advantages of our *fully-explored* masking strategy for masked language model.