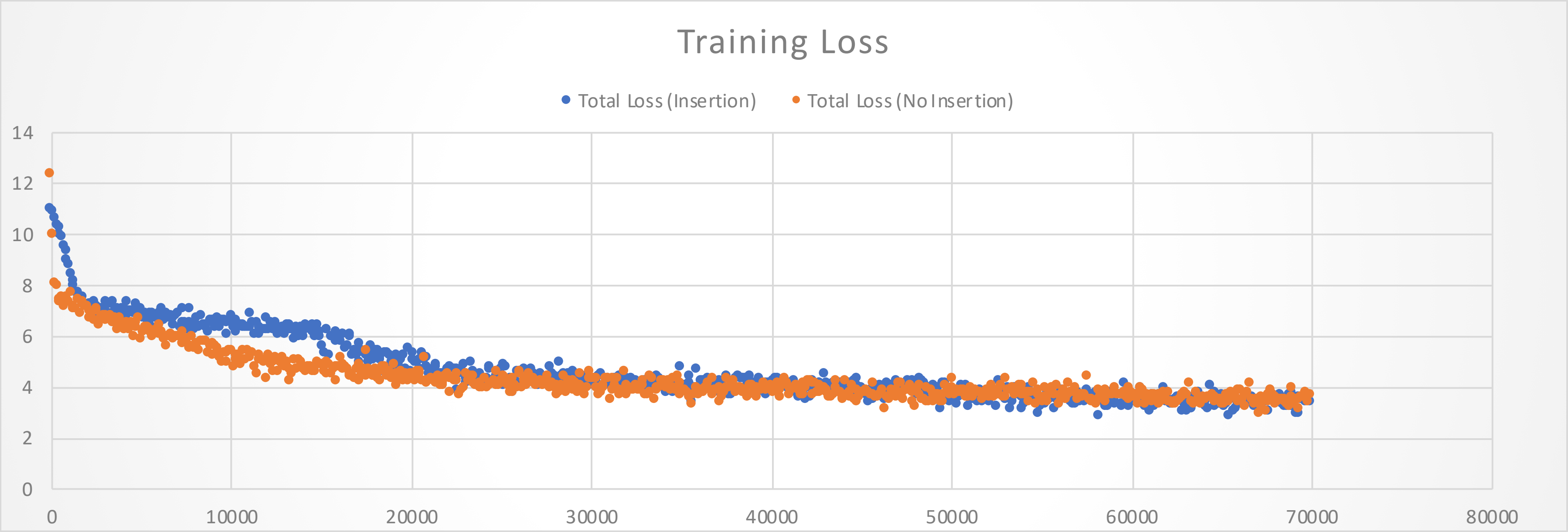
Impact of Dummy Token Insertion on BERT Pretraining

Description	Parameters/Environment		Results
The intent of this experiment is to determine whether inserting dummy tokens in the sentences of the training data, used to pretrain BERT language model; improves the overall performance of the pretraining. The rationale is to make the context recognition more robust which can lead to enhanced performance in downstream tasks.The dummy tokens are selected randomly from the prior two sentences. Single characters and fragments are excluded in the selection. Total number of intrusions in two modes are roughly equivalent. Note: in order to reduce the training time, the size of the model is reduced markedly. Multitask learning has been employed to accommodate the insertion loss implementation.	***Environment*** ***Dataset*** ***Word Embedding*** ***Learning Rate*** ***Training Steps*** ***Evaluation Steps*** ***Warm Up Steps*** ***Num. of Inserted Tokens*** ***Masked Token Probability***	Python 3.6 / Tensorflow 1.11 Project Gutenberg Byte Pair Encoding 5.00E-04 70000 20000 2000 5 15% without insertion / 10% with insertion	The insertion of dummy tokens in the training examples does not impact the long term total loss during latter stages of the training. The most notable observation of this experiment is that insertion of dummy tokens significantly accelerates the drop of the total loss during the early phase of the training. It turns out that the model is extremely effective in discerning between a valid token vs. an inserted token which explains why the impact of insertion loss dissipates quickly. Next step is to examine the impact of the insertion on the performance of downstream tasks such as reading comprehension (SQUAD).


Training Loss
● Total Loss (Insertion)
● Total Loss (No Insertion)