We thank the reviewers for their comments. We will carefully modify the paper according to the suggestions.

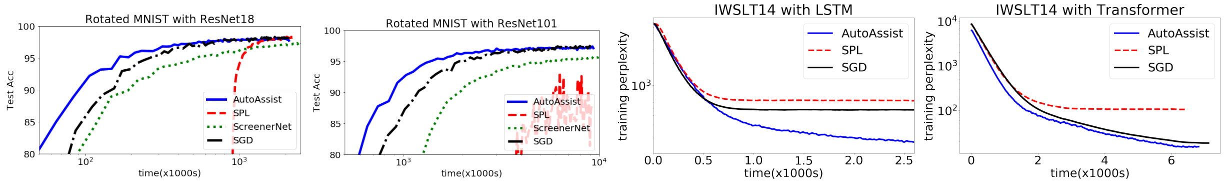

Figure 1: Comparison of different learning schemes on RotMNIST classification and IWSLT translation tasks.

Comparison with the vanilla SGD baseline. In the current manuscript, we only focused on the comparison on curriculum-based approaches (SPL [18] and ScreenerNet [16]) which demonstrates superiority over plain SGD and uses the final accuracy of the SGD as a sanity check for the quality of models trained with AutoAssist (e.g.g, BLEU score results in Appendix D). We thank and agree with the reviewers for the constructive suggestion that demonstration of superior performance of AutoAssist over the vanilla SGD without a curriculum would make the paper stronger. We will add this SGD baseline to the all experiments in the revised version. Due to the limited time of the rebuttal period, only partial results are included in Figure 1. Among the three curriculum-based models and the non-curriculum SGD baseline, we can see that AutoAssist shows the best performance.

To reviewer 1:

– We experimented with learning rates from 1^{-4} to 5^{-3} and pick the one with the best performance for all three models. For the NMT tasks, we used the same parameter settings from previous papers, as described in section 5.2. We've done sanity check with baseline SGD that the setting can reach the similar BLEU score as reported in the original paper.

– The linear assistant model we used is already one of the simplest ML models for the given format, and we still observe better performance.

– The aggressiveness is decided by Assistant through the safeguard variable γ, which is discussed in section 4.1 and plotted in Figure 2(R).

– It is true that a more complex Assistant model might yield a better shrinking accuracy. However, in order to minimize the time overhead, it is better to choose an Assistant model such that it requires less batch training time on CPU than the training time of Boss on GPU so that Assistant training can be hidden behind GPU training.

– We've experimented with batch size varying from 16 to 128 for image classification tasks and number of tokens varying from 1000 to 5000 for NMT tasks. Assistant model shows similar performance over different batch sizes.

– As described in previous paragraph, even with zero loss there is still a safeguarding base possibility that the instance will be incorporated into a training batch. In the extreme case that all training loss go to zero, the Assistant will gradually reduce to uniform sampling.

To reviewer 2:

– The reason why we didn't use raw ImageNet is that there is no available implementation for ScreenerNet on ImageNet dataset. However, we will provide results on raw ImageNet dataset and large Transformer model in the revised version.

– More ablation study will be included in our revision. The comparison of the portion of batch generation time over the entire training time is showed in Figure 2(L). With the parallel framework, the batch generation can be done in parallel to the GPU training and thus result in similar percentage of time used for batch generation (3% to 4%) in the plain SGD.

– Since the Assistant is learning from an evolving Boss model, it will not converge to a stationary point as most linear models do. However, we do observe the Assistant usually reaches a relatively stable state with high accuracy ($\sim 90\%$), during the first several epochs of the Boss (please see the blue carve in Figure 2(R)).

– We will include the listed related works into discussion.

To reviewer 3:

– In most cases while the data can be preprocessed and loaded into the memory, the CPU has idle periods. However, it is true that in the case that CPU needs to load every training batch from disk (such as raw ImageNet), the CPU utilization could be high. We will make this more clear in the revised version.

– As long as the CPU training is faster than the GPU training, the overhead could be hide behind GPU time, thus a shallow CNN is plausible.

– We will include raw ImageNet in the revised version.

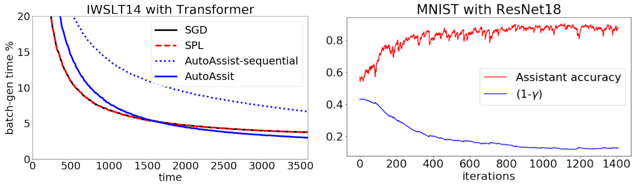

Figure 2: (L) Percentage of time used for batch generation during training. (R) Assistant predictive accuracy and safeguarding rate $(1-\gamma)$.