Self-Aware SGD: Reliable Incremental Adaptation Framework For Clinical AI Models

Additional Experiments

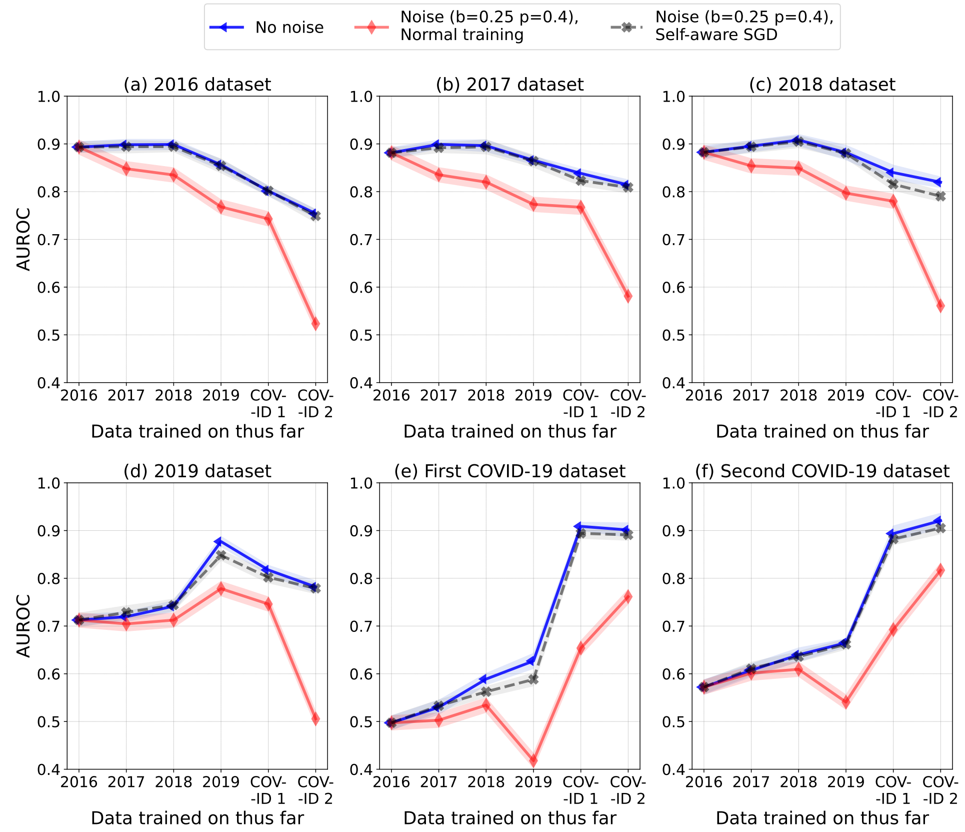

Fig. 1: Performance comparison between the normal DNN training and *self-aware SGD* under lesser label noise conditions ($b = 0.25$ and $p = 0.4$) on the test examples from (a) 2016, (b) 2017, (c) 2018, (d) 2019, (e) first and (f) second COVID-19 datasets. The baseline is normal DNN training under no label noise. Each x-axis point illustrates the datasets that have been used to update the model until that point.

I. Low-noise Experiments

We repeat the first experiment of the manuscript at two label noise configurations: ***1).*** batch probability of 0.25 and sample probability of 0.4 ($b = 0.25,\ p = 0.4$) ***2.)*** batch probability of 0.25 and sample probability of 0.6 ($b = 0.25,\ p = 0.6$). The parameter setting used in the manuscript is also used for these experiments.

The results of these experiments are listed in Fig. 1 and 2, respectively. Following inference can be drawn from the analysis of these figures:

- In both cases, the proposed self-aware SGD shows a significant improvement over standard incremental training in noisy conditions. This improvement by self-aware SGD is statistically significant at $p < 0.005$ across all test sets (or experimental configurations) as per the paired t-test. This highlights that self-aware SGD can handle low as well as high levels of label corruption.
- As desired, the performance of self-aware SGD is similar to baseline i.e. the standard incremental training in a no-noise setup.

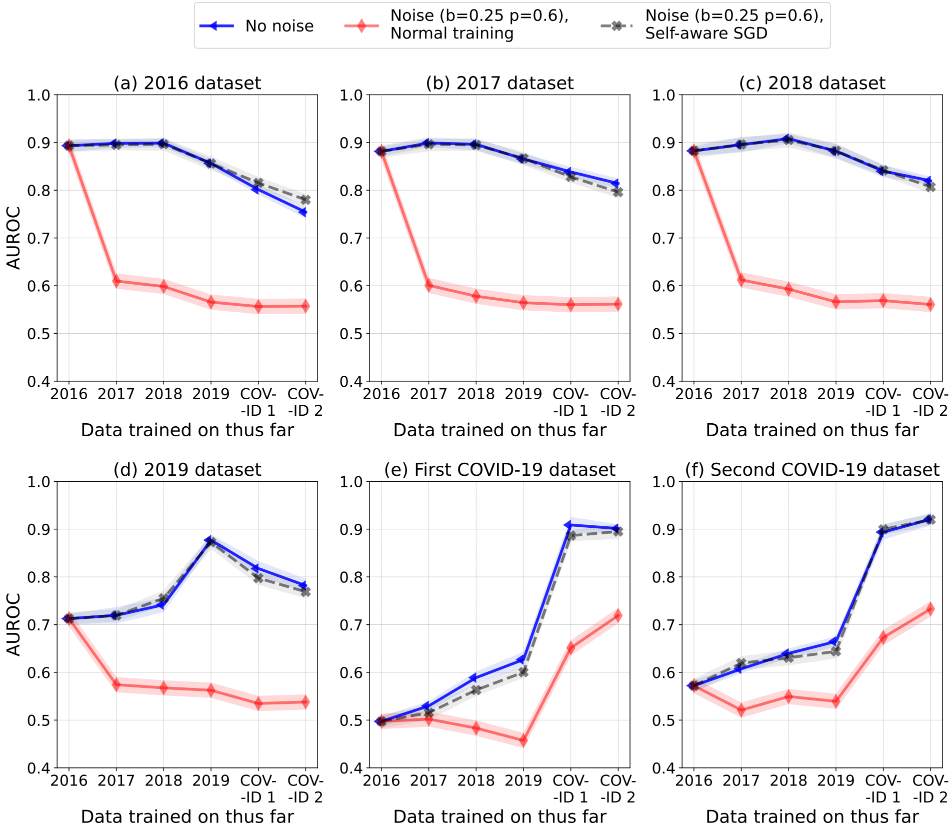

Fig. 2: Performance comparison between the normal DNN training and *self-aware SGD* under lesser label noise conditions ($b = 0.25$ and $p = 0.6$) on the test examples from (a) 2016, (b) 2017, (c) 2018, (d) 2019, (e) first and (f) second COVID-19 datasets. The baseline is normal DNN training under no label noise. Each x-axis point illustrates the datasets that have been used to update the model until that point.

- As expected, the standard training under high label noise ($p = 0.6$, Fig. 2) exhibits a greater performance drop during incremental training than the low label noise setup ($p = 0.4$, Fig. 1).

II. No noise experiment

In real-life scenario, we can always encounter scenarios where all incremental batches are pure/no label noise (best-case scenario), or all are corrupted (worst-case scenario). In both cases, we won't be able to train the bandit model. Bandit model is a binary classifier and requires gradient features as well as deviation in AUROC from both noisy and clean batches.

One simple solution to overcome this issue is to maintain a handful of batches with the correct labels. This small set (referred to as meta set) of correctly labeled batches has been used in many noise-robust deep learning frameworks such as meta-weight nets. We can artificially corrupt these batches to obtain their noisy versions. Then, these batches (both corrupted and no-noise batches) are mixed with incremental batches for training the bandit models. Hence, we make sure that there is always a representation from both classes during bandit training.

This information is already provided in Section V-C (Implementation details).

To evaluate this strategy, we perform the first experiment of the manuscript in a *no label noise* setup. During each incremental training bout, we set aside a measly 1% of the available batches (i.e. 25) for creating meta set. We corrupted the labels of this meta set using different sample probabilities ($p = 0.2, 0.4, 0.6, 0.8$). Both corrupted and no-noise versions of meta set are appended with the incremental setup. The rest of the process remains same. This

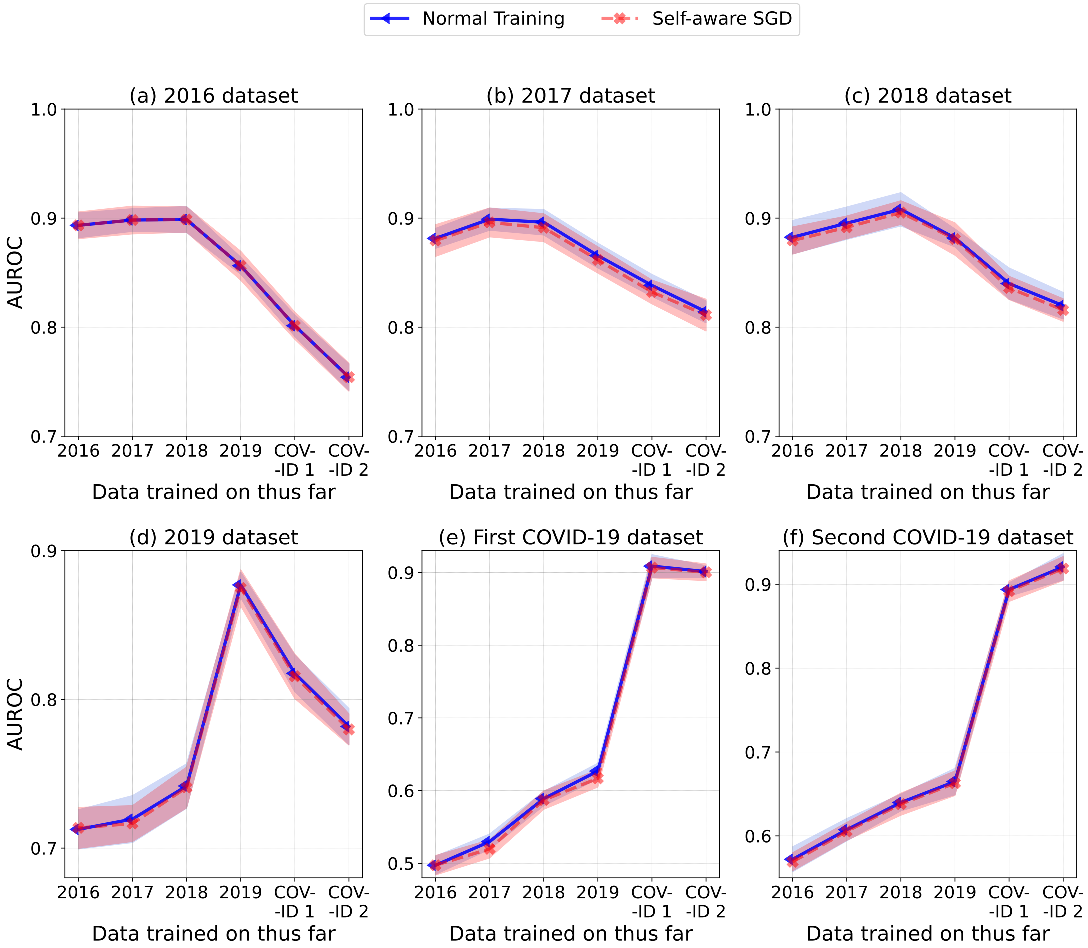

Fig. 3: Performance comparison between the normal incremental training and *self-aware SGD* under no label noise conditions on the test examples from (a) 2016, (b) 2017, (c) 2018, (d) 2019, (e) first and (f) second COVID-19 datasets. Each x-axis point illustrates the datasets that have been used to update the model until that point.

meta set is also used as a validation set for obtaining the best classification threshold over the bandit predictions (instead of 0.5 in Line 16, Algorithm 2 of the manuscript).

Fig. 3 illustrates the results of this experiment. The analysis of this figure highlights that across all test sets, the performance of self-aware SGD is comparable to the standard incremental training (standard DNN training). Also, the paired t-test further confirms that there is no statistical difference between the performance of self-aware SGD and the standard training ($p > 0.005$ across all test sets). Hence, self-aware SGD can be used in both noisy and non-noisy conditions.