Dear Reviewers,

We would like to thank you for your careful assessment of our paper. In order to improve the paper we intend to make the following changes:

1. One of the goals of Section 3.1 in the paper is to determine whether *customary confidence intervals* become poor indicators when one overuses the testing set. This is why we prefer using the simple and popular Wald approach over more precise approaches such as binomial tails or empirical Bernstein bounds. We intend to make this clear. The second goal of section 3.1 is to show how pairing improves the confidence intervals so much that Bonferroni-style corrections have far less impact. We intend to re-word this argument to emphasize the assumptions we are making and discuss how the argument resists when the assumptions are weakened.

2. We also intend to add additional experimental results. In particular we would like to have a modern CNN (ReLU instead of sigmoids, many more feature maps) and a system that relies on manually crafted features. However it is clear that whatever list of systems we test will remain very small in comparison to the countless results published in the literature. This is why we are looking forward to seeing more results using the QMNIST testing set.

3. We believe that the 2.3 MNIST trivia section is important and needs to be documented. In fact we received substantial positive feedback about this section and inquiries for even more information.

Regards,

Authors