Summary of report – Towards More Robust Solutions in Low Precision Training:

This report proposes several modified training schemes for deep neural networks, all leveraging the use of low-precision (8-bit) computation at some point in the training process, and the mitigation of error introduced in the naïve implementation thereof. Weight-averaged restarts after low-precision stochastic gradient descent, and a loss regularizer for low-precision computation are proposed. A strategy for efficiently evaluating the regularizer is developed, and experiments are conducted to evaluate the two proposed approaches. Mixed results are reported, with the regularized loss function and restarts providing marginal improvement over baseline methods.

Review:

DISCLAIMER: I am far from an expert on neural networks, so some amount of background that is understood by those in the field may be lost on me. From my perspective, however, there are many important details missing from this report.

With respect to structural completeness, the paper is acceptable. It has a motivating introduction, a brief description of related work, and a somewhat detailed description of the methods employed. It is missing some amount of background information, especially for readers not intimately familiar with deep neural networks and the standard training procedures involved. Experiments are done on the CIFAR-100 benchmark, but network architecture choices are not discussed anywhere that I can see, certainly an oversight. The report contains the realization of the stated midterm goals in the proposal. However, the tone of the paper diverges from the intent stated in the proposal (which is to demonstrate the failings of precision-switching and explore two solutions), and to some degree fails to emphasize that precision-switching experiments are expected to underperform without a mitigating strategy. A paragraph in the spirit of the first paragraph of section 3 of the proposal would go a long way to setting the stage for the rest of the paper.

This paper has a number of major issues with respect to clarity of presentation. Frequent grammatical and spelling errors distract significantly from its message, and the choice of words is often awkward and unclear. Several statements in the methods section are repeated verbatim from subsection to subsection, further halting the flow of the paper, and a number of imprecise word choices are made (e.g. referring to “good” and “bad” optima, while colloquially clear, is never explicitly described and feels sloppy). Oversights such as introducing Stochastic Weight Averaging twice in the introduction, once with capitalized first letters and the second time all lowercase in italics further indicate a need for thorough proofreading and editing.

In spite of those issues, the concepts discussed in the paper appear novel, and work towards fast evaluation of the proposed regularizer seems well founded. As for the significance of the work, it seems that currently performance improvement is marginal, but shows some promise, especially future work aimed at combining both proposed modifications to training.

One major oversight in the content of the paper is with respect to computational efficiency. Despite the motivating fact that 8-bit computations improve efficiency, all studies to date are focused on post-training accuracy and none deal with the overall training time of the low-precision methods as compared to their final test set accuracy. In a fully developed paper I would expect to see at least some mention of these results, if not a complete study.

Suggestions:

1. Edit and proofread. This report is not suitably polished for submission as an assignment, let alone to a conference. There’s very little to say here except for that spelling and grammatical errors have completely overshadowed the work being presented. Perhaps a fresh pair of eyes would help – I often find that if I miss an error the first time I’ll miss it the second time, and so I usually have at least one other person take a look at important submissions for me.
2. Include more details about the specific experiments that were run. Nowhere is the network architecture mentioned, and nowhere are training times reported. These details seem vital to evaluating the quality of results – how can one attempt to reproduce and extend them if information like that is missing?
3. Formalize choice of language, and read the paper out loud as a narrative (one suggestion made by Prof. Wilson in his guide to writing reports). This should clarify what I mean when I talk about issues with the flow of the paper. Consider a full rewrite, perhaps after diagramming each section of the paper and distilling major ideas for those sections down to a few bullet points. Take plenty of time with the write-up, and revisit each section a few times.
4. In figure 3(c), there are two trajectories labeled “Full-precision”. The first is the same as the other two plots in the figure, but it is difficult to determine what the second refers to. Additionally, the “simple alternative method” described when referring to figure 3(c) seems to be the red curve, meaning the green curve is completely without context. Please make this figure clearer, hopefully by expanding the caption to be more descriptive.

Technical Questions:

1. In section 3.3.1, there are two claims that feel unsupported, though they may certainly be true. Specifically the claim that $\alpha=0$ is a known regularization technique should probably have a citation attached to it, and the claim that the $\mathcal{L}(w)$ term centers the ultimate solution in the local minimum is not immediately obvious and should be explained in some detail. In short, my technical question is “why?” with respect to that second statement.
2. In the last paragraph of section 3.3.2, it is claimed that noise reduction of the modified loss function approximation can be achieved by increasing the samples tested for each $X\_j$, and that this can be done nearly in parallel by duplicating objects in the minibatch in question and using the local reparameterization trick, but that the number of samples is restricted by GPU memory. Here I have two questions, because this seems like an important ingredient in making the modified loss function viable. First, how many samples per $X\_j$ are viable, and how many were used? Second, has there been any study (mathematical or numerical) of this variance with respect to the number of samples – and if not, how can we judge whether this approximation is reasonable in another way (perhaps empirically)?

Review Summary:

In my opinion this report is based on several good ideas and has some strong math behind the technical portion, which is to say the approximation of the modified loss function. However, glaring issues with the writing itself significantly detract from the paper’s ability to get its point across and justify the methods introduced. A number of missing details such as network architecture and efficiency studies, as well as a few poor word choices further this issue. Overall this report is quite far from finished, and significant portions need to be revisited and re-evaluated, if not rewritten.