Final Report Review: Towards More Robust Solutions in Low Precision Training

Summary of Report:

The authors address performance of deep neural networks trained with low-precision storage used during training. Building on recent ideas such as SWALP, they propose and test two means of alleviating the performance gap between low-precision and full-precision training. First, SWA restarting, where low-precision training is resumed after weight averaging (SWA) is introduced, and second a modified loss function which promotes flat local minima is described. After developing an efficient means of evaluating this modified loss function, the authors test the performance of these methods in conjunction with SWALP and find that full-precision performance can nearly be recovered on CIFAR-10 and CIFAR-100.

Review:

This paper is complete, fully developed and contains thorough background. Methods are described at a high level several times, and in detail where necessary. Helpful visualizations of the approaches are provided which significantly improve clarity. In addition, the authors seem to have a very well-developed notion of how this work fits in to the landscape of low-precision methods and previous work, which is evident in the background section. Experimental results are certainly meaningful and significant, and even without strong motivation in introductory sections are motivating in and of themselves.

The paper’s strengths lie in its detail of explanation and visualization. A somewhat esoteric topic becomes easy to conceptualize with a few key plots, and results are well-presented overall. Its weaknesses, as in the midterm report, still lie in writing style and flow, with a small number of typos, awkward word choice and grammatical errors detracting occasionally from the strength of a paragraph. However, these issues have been massively improved since the midterm report, and I would rate it as having fair quality, strong originality and significance, and strong clarity.

Since the midterm report, very impressive results have been achieved – nearly full recovery of high-precision training performance is quite a feat, and speaks to the quality of the methods themselves. As compared to the results presented in the midterm report (table 1), these new results (table 2, specifically performance on CIFAR-100) completely change the narrative of the project in my opinion. In my review of the midterm report, writing quality was my main issue with the submission. Here that has been nearly completely turned around, with very few remaining issues. Those should be polished before submission, but overall the project is significantly improved by the work that has already been done in this area. By comparison to the midterm checkpoint this report was a pleasure to read.

Suggestions:

Typos: Introduction, final paragraph: “posterior”, section 4 first paragraph: “week”, which I believe should say “look”. One or two mild grammatical issues exist, but are not critical. Again I suggest reading each section aloud, or having it read aloud to you by somebody else (perhaps somebody who has not yet read the paper).

My other suggestion is to more prominently feature your core results. I find them very impressive, but they feel buried at the end and are not discussed in much detail. Perhaps this is due to the wealth of background provided, which may not be necessary in the context of submission to a conference but is certainly necessary in this context. Either way, it seems to me you could make a bigger deal about them.

Summary:

This project posits and proves the merits of two approaches to alleviate the performance gap between low-precision and high-precision training of deep neural networks. In combination with existing methods, the authors nearly recover full performance on two standard datasets. The background and methodology are very thorough, and the writing is clear even though a small number of details still should be fixed. Overall this report is solid and tells the story of a successful project.