8-BIT GRADIENT APPROXIMATION FOR PARALLELISM IN DEEP LEARNING

Tim Dettmers
The Faculty of Informatics
Universi della Svizzera italiana
Via Giuseppe Buffi 13, CH-6904 Lugano, Switzerland
tim.dettmers@usi.ch

ABSTRACT

The application of deep learning to large data sets is important to create practical data products featuring language and visual understanding. Parallelization across processors and computers is often needed to make deep learning on large data sets feasible but bottlenecks in communication bandwidth make it difficult to attain good speedups through parallelism. Here we develop an algorithm, which provides improved utilization of the available bandwidth by compressing 32-bit gradients to 8-bit approximations. We show that these approximations do not decrease predictive performance for both model and data parallelism and provide a speedup of 2x relative to 32-bit parallelism. Thus 8-bit approximation provides an universally applicable algorithm which achieves state-of-the-art parallelism for convolutional networks.

1 INTRODUCTION

ICLR requires electronic submissions, processed by `http://arxiv.org`. See ICLR's website for more instructions.

If your paper is ultimately accepted, the statement `\iclrfinalcopy` should be inserted to adjust the format to the camera ready requirements.

The format for the submissions is a variant of the NIPS format. Please read carefully the instructions below, and follow them faithfully.

1.1 STYLE

Papers to be submitted to ICLR 2016 must be prepared according to the instructions presented here.

Authors are required to use the ICLR LaTeX style files obtainable at the ICLR website. Please make sure you use the current files and not previous versions. Tweaking the style files may be grounds for rejection.

1.2 RETRIEVAL OF STYLE FILES

The style files for ICLR and other conference information are available on the World Wide Web at

`http://www.iclr.cc/`

The file `iclr2016_conference.pdf` contains these instructions and illustrates the various formatting requirements your ICLR paper must satisfy. Submissions must be made using LaTeX and the style files `iclr2016_conference.sty` and `iclr2016_conference.bst` (to be used with LaTeX2e). The file `iclr2016_conference.tex` may be used as a "shell" for writing your paper. All you have to do is replace the author, title, abstract, and text of the paper with your own.

The formatting instructions contained in these style files are summarized in sections 2, 3, and 4 below.

2 General formatting instructions

The text must be confined within a rectangle 5.5 inches (33 picas) wide and 9 inches (54 picas) long. The left margin is 1.5 inch (9 picas). Use 10 point type with a vertical spacing of 11 points. Times New Roman is the preferred typeface throughout. Paragraphs are separated by 1/2 line space, with no indentation.

Paper title is 17 point, in small caps and left-aligned. All pages should start at 1 inch (6 picas) from the top of the page.

Authors' names are set in boldface, and each name is placed above its corresponding address. The lead author's name is to be listed first, and the co-authors' names are set to follow. Authors sharing the same address can be on the same line.

Please pay special attention to the instructions in section 4 regarding figures, tables, acknowledgments, and references.

3 Headings: first level

First level headings are in small caps, flush left and in point size 12. One line space before the first level heading and 1/2 line space after the first level heading.

3.1 Headings: second level

Second level headings are in small caps, flush left and in point size 10. One line space before the second level heading and 1/2 line space after the second level heading.

3.1.1 Headings: third level

Third level headings are in small caps, flush left and in point size 10. One line space before the third level heading and 1/2 line space after the third level heading.

4 Citations, figures, tables, references

These instructions apply to everyone, regardless of the formatter being used.

4.1 Citations within the text

Citations within the text should be based on the `natbib` package and include the authors' last names and year (with the "et al." construct for more than two authors). When the authors or the publication are included in the sentence, the citation should not be in parenthesis (as in "See **?** for more information."). Otherwise, the citation should be in parenthesis (as in "Deep learning shows promise to make progress towards AI (**?**).").

The corresponding references are to be listed in alphabetical order of authors, in the References section. As to the format of the references themselves, any style is acceptable as long as it is used consistently.

4.2 Footnotes

Indicate footnotes with a number[1] in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).[2]

[1] Sample of the first footnote

[2] Sample of the second footnote

Table 1: Sample table title

PART	DESCRIPTION
Dendrite	Input terminal
Axon	Output terminal
Soma	Cell body (contains cell nucleus)

4.3 FIGURES

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction; art work should not be hand-drawn. The figure number and caption always appear after the figure. Place one line space before the figure caption, and one line space after the figure. The figure caption is lower case (except for first word and proper nouns); figures are numbered consecutively.

Make sure the figure caption does not get separated from the figure. Leave sufficient space to avoid splitting the figure and figure caption.

You may use color figures. However, it is best for the figure captions and the paper body to make sense if the paper is printed either in black/white or in color.

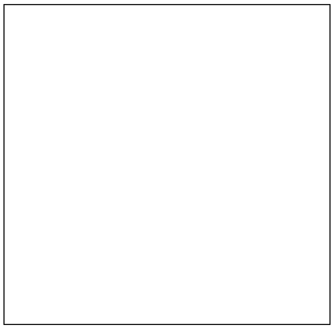

Figure 1: Sample figure caption.

4.4 TABLES

All tables must be centered, neat, clean and legible. Do not use hand-drawn tables. The table number and title always appear before the table. See Table 1.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

5 FINAL INSTRUCTIONS

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle the text should fit into, and do not change font sizes (except perhaps in the REFERENCES section; see below). Please note that pages should be numbered.

6 PREPARING POSTSCRIPT OR PDF FILES

Please prepare PostScript or PDF files with paper size "US Letter", and not, for example, "A4". The -t letter option on dvips will produce US Letter files.

Consider directly generating PDF files using `pdflatex` (especially if you are a MiKTeX user). PDF figures must be substituted for EPS figures, however.

Otherwise, please generate your PostScript and PDF files with the following commands:

```
dvips mypaper.dvi -t letter -Ppdf -G0 -o mypaper.ps
ps2pdf mypaper.ps mypaper.pdf
```

6.1 MARGINS IN LATEX

Most of the margin problems come from figures positioned by hand using `\special` or other commands. We suggest using the command `\includegraphics` from the graphicx package. Always specify the figure width as a multiple of the line width as in the example below using .eps graphics

```
   \usepackage[dvips]{graphicx} ...
   \includegraphics[width=0.8\linewidth]{myfile.eps}
```

or

```
   \usepackage[pdftex]{graphicx} ...
   \includegraphics[width=0.8\linewidth]{myfile.pdf}
```

for .pdf graphics. See section 4.4 in the graphics bundle documentation (`http://www.ctan.org/tex-archive/macros/latex/required/graphics/grfguide.ps`)

A number of width problems arise when LaTeX cannot properly hyphenate a line. Please give LaTeX hyphenation hints using the `\-` command.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Bengio, Yoshua and LeCun, Yann. Scaling learning algorithms towards AI. In *Large Scale Kernel Machines*. MIT Press, 2007.

Hinton, Geoffrey E., Osindero, Simon, and Teh, Yee Whye. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.