Vision Transformer Pruning

Bui, Vy Chen, Chiu-Chun Helms, Derek Mueller, Sebastian
Oregon State University
{buivy, chenchiu, helmsd, FILLIN}@oregonstate.edu

Abstract

The ABSTRACT is to be in fully justified italicized text, at the top of the left-hand column, below the author and affiliation information. Use the word "Abstract" as the title, in 12-point Times, boldface type, centered relative to the column, initially capitalized. The abstract is to be in 10-point, single-spaced type.

1. Introduction

Give a general introduction on vision transformers, discuss start in NLP and transfer to vision tasks, who first implemented ViT, brief discussion on CNN's and how they are the main architecture for vision tasks, discuss what we will do in the paper (implementing ViT and pruning methods), can discuss results if we want to add a brief statement on how much the model was reduced by.

2. Related Work

Vision Transformers. *Discuss background on transformers, who introduced ViT, where it started and where it is now (in terms of size/performance), compare to CNN models, etc.*

Pruning Methods. *Discuss general idea of pruning methods, can discuss transformer pruning in NLP, discuss multiple pruning methods (not only the ones we implement), what are the advantages or disadvantages, etc.*

3. Methodology

3.1. Implementation of DeiT Baseline

Discuss short background of DeiT model, who implemented/create the model architecture, that it is pre-trained on ImageNet-1k, discuss baseline results (accuracy, flops, and params), etc.

3.2. Pruning Method Implementation

Discuss the pruning method being used (reference paper and discuss what parts of ViT are being pruned - which layers), what we expect the reduction to be, short discussion on accuracy/FLOPs trade-off, any limitations/known issues with pruning.

3.3. Evaluating Model Performance

Discuss how we are evaluating the model, what data set are we using, maybe add brief background on ImageNet-1k, define FLOPs in terms of our project, define what parameters are being measured (i.e. layers), etc.

4. Results

5. Conclusion

References

[1] FirstName Alpher. Frobnication. *IEEE TPAMI*, 12(1):234–778, 2002. 2

[2] FirstName Alpher and FirstName Fotheringham-Smythe. Frobnication revisited. *Journal of Foo*, 13(1):234–778, 2003. 2

[3] FirstName Alpher, FirstName Fotheringham-Smythe, and FirstName Gamow. Can a machine frobnicate? *Journal of Foo*, 14(1):234–778, 2004. 2

[4] FirstName Alpher and FirstName Gamow. Can a computer frobnicate? In *CVPR*, pages 234–778, 2005. 2

6. Final Report Guidelines

This document serves as a style guide for the AI535 Final Report and describes the expected content.

6.1. Content

Reports should be similar to real papers – having a title, abstract, introduction, discussion of background material and related work, detailed technical approach, experimental results, and a conclusion. The typical section layout is below along with questions typically covered by each in a full paper. Depending on your project (and its progress), you may not have full answers to each.

1) **Abstract.** A short 5-6 sentence overview. What is the problem? Why should we care? How do you address it? What are your general results? Think of this as an "ad" for your work that helps readers decide if it is relevant.

2) **Introduction.** What is being studied? Why is it important? What problems keep it from being solved already? How does your approach resolve these? What experiments suggest your approach was effective? There are dedicated sections later for some of these questions, just provide a summary and motivation here.

3) **Related Work.** How has prior work addressed this problem? How is your approach different / similar to the prior work? Organize, contrast, and compare with others – don't just list prior work, that adds little value!

4) **Methodology.** What did you do? And why? Be specific about algorithmic details. Be clear about what ideas are novel vs. what you are using from others. Try to organize this section appropriately – if your algorithm has multiple stages, separate them out as subsections.

5) **Results.** What is the experimental setting used to evaluate your approach? Why is that setting appropriate? How were the experiments run? What are the results? Are there reasonable baselines or prior work to compare to? If so, how does your approach compare?

6) **Conclusion.** What was learned from this work? What does it suggest are useful things to do next? What are limitations of the approach or experiments?

Your report should also end with a bibliography of works cited. In this template, that is handled automatically with BibTex. For example, I might cite papers listed in `egbib.bib` like [1–3] as a group or individually [4]. See this tutorial for more information on BibTex.

Method	Frobnability
Theirs	Frumpy
Yours	Frobbly
Ours	Makes one's heart Frob

Table 1. Results. Ours is better.

6.2. Formatting

This document is the report template and is based off of the CVPR 2022 author template. Modifying text size, line spacing, or margins is prohibited.

Length. Reports (including references and all figures) must be at least four pages and can be at most 8 pages.

Figures and Tables Figures and tables must include captions and be centered.

When placing figures in LaTeX, it's almost always best to use `\includegraphics`, and to specify the figure width as a multiple of the line width as in the example below

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

Further, aligning figures at the top of the page is preferred

```
\begin{figure}[t]
```

For tables, no vertical lines should be present. Please us the `booktabs` package's `toprule`, `midrule`, and `bottomrule` commands for horizontal lines as in the example in Table 1.

Mathematics. All equations appearing in the report should be numbered. For example, using the `equation` or `eqnarray` environments will handle this automatically:

$$E = m \cdot c^2 \tag{1}$$

$$\begin{aligned} a^2 + b^2 &= c^2 && (2)\\ &= ln(e^{c^2}) && (3) \end{aligned}$$

For further style tips regarding equations in prose, see `www.pamitc.org/documents/mermin.pdf`.

Language. Reports are expected to be in English with minimal spelling and grammar errors. Point will be deducted if these errors make understanding the report difficult.

6.3. Submission

Papers must be submitted as PDFs to Canvas – no other format will be accepted. Only one report per group needs to be submitted.

References

[1] FirstName Alpher. Frobnication. *IEEE TPAMI*, 12(1):234–778, 2002. 2

[2] FirstName Alpher and FirstName Fotheringham-Smythe. Frobnication revisited. *Journal of Foo*, 13(1):234–778, 2003. 2

[3] FirstName Alpher, FirstName Fotheringham-Smythe, and FirstName Gamow. Can a machine frobnicate? *Journal of Foo*, 14(1):234–778, 2004. 2

[4] FirstName Alpher and FirstName Gamow. Can a computer frobnicate? In *CVPR*, pages 234–778, 2005. 2