
Can ChatGPT write an academic paper? Review of “A Day in the Life of ChatGPT”

Adrien Foucart, PhD¹

2023-01-07

Introduction

Mashrin Srivastava tried an interesting ChatGPT experiment: a “paper completely written by ChatGPT” (originally published on LinkedIn, now available as a preprint on ResearchGate: <https://doi.org/10.13140/RG.2.2.34470.60480>).

He appended to the generated paper the whole sequence of prompts and responses that was used to generate it. The original prompt was for an article to submit to a workshop on *Sparsity in Neural Networks* at ICLR 2023.

A Day in the Life of ChatGPT as a researcher: Sustainable and Efficient Machine Learning - A Review of Sparsity Techniques and Future Research Directions

Title of the generated article (“A Day in the Life of ChatGPT as a researcher:”) was added by Mashrin, the rest comes from the bot

I don’t think it makes a lot of sense to review the resulting paper like I would do for a regular paper (I can’t pretend that I don’t know it’s a generated text), but I think the exercise is really interesting, so I want to take a deep dive into it.

A few disclaimers before I start:

- To be honest about my own biases, I have been quite open about the fact that I don’t think ChatGPT is revolutionary or quite as useful as the hype around it make it look, and that I’m very skeptical of the actual value of Large Language Models as anything

¹ Contact: Adrien.Foucart@ulb.be. While the author is affiliated with the LISA laboratory of the Université Libre de Bruxelles, this work was done independently from the research group and does not aim to follow usual academic standards. It was originally written as a blog post on the author’s personal research blog: <https://research.adfoucart.be/review-chatgpt-paper>.

else than fun text generators. But I'm sometimes wrong about things (allegedly), so who knows?

- I'm not an expert in sparsity. I'm enough of an expert in machine learning to know why it's an interesting topic in neural networks, and to know the basics of how to obtain it (e.g. regularization techniques), but I can't fully judge the scientific quality of a novel paper on the topic. So in terms of accuracy I'll focus on what I can easily check.

Okay, let's get started !

1. Looking at the prompts sequence

I started from the Appendix, so I could follow along the generation of the paper to better get a sense of where the ideas came from and how much "prompting" it actually needed. It's really great from Mashrin to include all of it, as it makes the analysis way more complete. Thanks!

Generating an outline

The first prompt asked for a paper idea for the workshop, giving to ChatGPT the workshop description.

Here, ChatGPT mostly takes point by point all the proposed topics from the workshop description, and arranges them into sections. This is fairly typical from what I've seen of ChatGPT: it tends to really want to include everything you mention in its answers.

Workshop details is: Deep networks with billions of parameters trained on large datasets have achieved unprecedented success in various applications, ranging from medical diagnostics to urban planning and autonomous driving, to name a few. However, training large models is contingent on exceptionally large and expensive computational resources. Such infrastructures consume substantial energy, produce a massive amount of carbon footprint, and often soon become obsolete and turn into e-waste. While there has been a persistent effort to improve the performance of machine learning models, their sustainability is often neglected. This realization has motivated the community to look closer at the sustainability and efficiency of machine learning, by identifying the most relevant model parameters or model structures. In this workshop, we examine the community's progress toward these goals and aim to identify areas that call for additional research efforts. In particular, by bringing researchers with diverse backgrounds, we will focus on the limitations of existing methods for model compression and discuss the tradeoffs among model size and performance. The following is a non-exhaustive list of questions we aim to address through our invited talks, panels, and accepted papers:

Where do we stand in evaluating and incorporating sustainability in machine learning? We make our models larger every day. Is this the right way to learn better?

Do we need better sparse training algorithms or better hardware support for the existing sparse training algorithms?

Hardware seems to be behind in supporting sparse training. What are the challenges of hardware design for sparse and efficient training? Are GPUs the answer or do we need new designs?

Our current theory can only analyze small neural networks. Can compression help us provide performance and reliability guarantees for learning?

What are the tradeoffs between sustainability, efficiency, and performance? Are these constraints competing against each other? If so, how can we find a balance?

Among different compression techniques, quantization has found more applications in industry. What is the current experience and challenges in deployment?

How effective sparsity could be in different domains, ranging from reinforcement learning to vision and robotics?

I. Introduction

1. Brief overview of the importance of sustainability and efficiency in machine learning
2. Introduction of sparsity as a way to address these issues
3. Thesis statement outlining the main focus of the paper

II. Background

1. Definition and explanation of sparsity in the context of neural networks
2. Overview of existing sparsity-inducing techniques, such as pruning, quantization, and low-rank decomposition
3. Discussion of the tradeoffs between model size, efficiency, and performance

III. Sparse Training Algorithms

1. Overview of current sparse training algorithms, including weight decay and gradient-based pruning
2. Evaluation of the effectiveness and limitations of these algorithms
3. Suggestions for future research directions in this area

IV. Hardware Support for Sparse Training

1. Overview of current hardware architectures and their support for sparse training
2. Discussion of the challenges and limitations of hardware design for sparse and efficient training
3. Evaluation of the potential for new hardware designs, such as specialized sparse processors, to improve efficiency and sustainability

V. Compression for Large-Scale Neural Networks

1. Discussion of the current limitations of compression techniques in providing performance and reliability guarantees for large-scale neural networks
2. Suggestions for future research directions in this area, including the potential for compression to enable the analysis of larger networks

VI. Sparsity in Different Machine Learning Domains

1. Evaluation of the effectiveness of sparsity in various domains, such as reinforcement learning, vision, and natural language processing
2. Discussion of the specific challenges and tradeoffs of applying sparsity in these different domains

VII. Conclusion

1. Summary of the main points discussed in the paper
2. Reiteration of the importance of sustainability and efficiency in machine learning
3. Final thoughts on the potential for sparsity to address these issues and future research directions in this area. I hope this outline provides a helpful starting point for your research. It is important to carefully evaluate the reliability of sources and accurately document your work with citations. Good luck with your paper!

Workshop description (left) in the prompt vs proposed outline (right) by ChatGPT

The result here is that the outline proposes a wide but superficial survey paper, which would not really be fitting for a workshop. The structure reads more like the chapters of a book than the sections of a conference paper.

So: not a *great* start, but nothing particularly bad right now.

Generating the introduction

Large portions of the introduction are again directly taken from the workshop description. It then rewrites the paper outline, which itself also came from the prompt. Having the outline explained in the intro is normal, but this means that right now we still don't have any added value from ChatGPT: if we want to know what are interesting topics related to sparsity, we can read the workshop description and we'll have as much info as with the current intro.

Deep neural networks have achieved remarkable success in a wide range of applications, from medical diagnostics to autonomous driving. However, the training and deployment of these models often requires significant computational resources and energy, leading to a large carbon footprint and the potential for e-waste. In light of these sustainability and efficiency concerns, the machine learning community has begun to explore techniques for reducing the size and complexity of neural networks.

One promising approach is sparsity, which involves the explicit or implicit reduction of the number of non-zero parameters in a neural network. Sparsity has the potential to significantly reduce the computational and storage requirements of a model, as well as improve its generalization performance. In this paper, we will explore the current state of the art in sparsity-inducing techniques and their application to various machine learning domains.

The main focus of this paper will be on the limitations and potential future research directions of sparse training algorithms and hardware support for sparse training. We will also examine the tradeoffs between model size, efficiency, and performance, as well as the potential for compression to enable the analysis of larger neural networks. Finally, we will evaluate the effectiveness of sparsity in different machine learning domains, such as reinforcement learning, vision, and natural language processing.

Through this paper, we aim to provide a comprehensive overview of the current state of research on sparsity in neural networks and identify areas that call for additional efforts. By bringing together researchers from academia and industry with diverse expertise and points of view, we hope to further develop these research directions for the machine learning community and facilitate the incorporation of sustainability and efficiency constraints into machine learning pipelines.

Introduction generated by ChatGPT

Also, at this point we have zero citations, including for assertions of facts such as:

- The range of applications of DNN success (medical diagnostics, autonomous driving)... There should at least be some citation to review papers in those domains.
- The large carbon footprint and e-waste.
- The potential of a "sparsity" approach.

Generating the "background" section

The first problem here – caught by Mashrin – is that there is still zero citations, which becomes very problematic in what is typically a citation-heavy part of a paper.

The provided definitions are then very surface-level, and arguably wrong: for instance, “weight decay” is not by itself a *pruning* technique. It’s a regularization technique which can be used in conjunction with a pruning method which removes connections with weights that are close to zero.

In general, the level of the explanations at this point could be acceptable for a student assignment (assuming it’s correct, I don’t see many obvious errors but there could be), not quite for an actual research paper.

Mashrin then asks ChatGPT to provide citations and to add more details “for a research paper”. We’re moving away a bit from the “completely generated by ChatGPT” scenario, but that’s a bit nitpicky so let’s keep going.

The response doesn’t have more details, but did add citations and a reference list, so let’s go through that.

Sparsity refers to the explicit or implicit reduction of the number of non-zero parameters in a neural network (LeCun et al., 1990). Sparsity can be introduced in various ways, including through the use of pruning, quantization, and low-rank decomposition (Han et al., 2015).

Pruning involves the removal of individual parameters or connections in a neural network based on their importance or relevance to the model’s performance (Hassibi et al., 1993). Pruning can be performed either manually or through the use of automated algorithms. One common automated pruning method is weight decay, which adds a regularization term to the objective function that encourages the weights to be small (Krogh & Hertz, 1992). Another method is gradient-based pruning, which removes the weights with the smallest gradients during training (Han et al., 2015).

Start of the background section generated by ChatGPT

The first reference is “Sparsity refers to the explicit or implicit reduction of the number of non-zero parameters in a neural network (LeCun et al., 1990)”. LeCun et al. is not in the reference list (which I think just got truncated due to limits on the output size). Much later in the process, Mashrin asks ChatGPT to provide a BibTeX citation for it, to which we get a reply mentioning a paper called “Optimal Brain Damage” by LeCun, Denker, Solla, Howard and Jackel. This is not exactly correct, as that paper is from 1989 and with only LeCun, Denker, and Solla. While the paper does talk about pruning and reducing the size of a network, it does not provide such a definition of sparsity.

So let’s check some other references:

- *Sparsity can be introduced in various ways, including through the use of pruning, quantization, and low-rank decomposition (Han et al., 2015)*: the reference is correct, but the paper really only deal with pruning (although it does mention quantization and low-rank approximation in its related works).
- *Pruning involves the removal of individual parameters or connections in a neural network based on their importance or relevance to the model’s performance (Hassibi et al., 1993). Pruning can be performed either manually or through the use of automated algorithms.*: the reference provided has the wrong year (it’s 1992), but it exists. The paper talks about pruning, but it doesn’t really state such a definition, so it’s not ideal. Also, “manual” pruning of a deep neural network doesn’t sound like a very realistic prospect...

- *One common automated pruning method is weight decay, which adds a regularization term to the objective function that encourages the weights to be small (Krogh & Hertz, 1992):* The reference got truncated as well, but it's in the final paper, so let's use the info from there. The paper does explain weight decay, but does not state that it is a pruning method (it's explored as a way to combat noise).

Let's stop here for a moment: the citations are generally "on topic", but they are clearly not the *source* of the information.

Credit where it's due, however: ChatGPT shows itself here to be a decent source of papers to look at to actually get to know the topic... at least from a fairly general perspective (up to now at least this is all very superficial).

Generating the "sparse training algorithms" section

Still looking at the citations, here our first is to Krogh and Hertz, 1992, but in the references the title corresponds to another paper from the same authors (from 1991). Again, the citation is "on topic", but does not explicitly provide the assertion cited.

Weight decay methods, such as L1 and L2 regularization, add a penalty term to the objective function that encourages the weights to be small (Krogh & Hertz, 1992). The L1 regularization term, also known as the Lasso, adds a penalty proportional to the absolute value of the weights, while the L2 regularization term, also known as the Ridge, adds a penalty proportional to the square of the

weights. The use of weight decay can significantly reduce the number of non-zero parameters in a neural network and improve its generalization performance, but may also degrade its fitting ability (van der Maaten et al., 2010).

Gradient-based pruning methods remove the weights with the smallest gradients during training (Han et al., 2015). These methods typically involve the use of a pruning threshold, below which the weights are set to zero and the corresponding connections are removed. Gradient-based pruning can effectively induce sparsity in a neural network and improve its efficiency, but may also lead to a decrease in performance (van der Maaten et al., 2010).

Start of the sparse training algorithms section generated by ChatGPT

We also here see repetitions from things already written in the previous chapter. This section barely brings any new details or new pieces of information to the table. The algorithms provided are:

- *Weight decays methods:* this is indeed a standard way of inducing sparsity. The "Lasso / Ridge" nicknames are only used in regression models, and it doesn't apply to the "regularization term" (see e.g. [here](#) for more), so that part is not correct. Last sentence is correct, but the citation is incorrect. Van der Maaten and Hinton is from 2008, not 2010, and the title is "Visualizing data using t-SNE", not "Getting the most out of a neural network by minimizing the amount of parameter" (which doesn't exist as far as I can tell). Also, it has nothing to do with weight decay.
- *Gradient-based pruning methods:* the explanation is mostly incorrect (the "standard approach" would rather be to sort the gradients and remove the smallest ones rather than using a hard threshold), and neither of the cited papers (Han et al., van der Maaten et al.) talk about it.

Then it does some improv on the limitations of those two “methods” (with again van der Maaten cited for no reason).

Mashrin prompts it again for “more sparsity algorithms”. We get:

- *Structured pruning*, pointing again to Han et al. Han et al. does not mention anything called structured pruning. It does not mention pruning an entire layer either. In fact, the definition of structured pruning here is just incorrect.
- *Sparse initialization*. The provided source is not about sparse initialization, but about ReLU activation functions.
- *Column sampling*. There is no “column sampling” algorithm as far as I can find. The reference does present a method to reduce parameters, but not by the described method.
- *Binary weights*. This method does not reduce the amount of parameters or induce sparsity. Also, it’s badly described: the weights are only restricted *during the forward and backward passes* of the training, but they are still stored as usual. It would also not necessarily require “specialized hardware”.
- *Low-precision weights*. That’s just quantization again, which was previously mentioned in the paper.

Mashrin asks for more again, and we get “*Dynamic sparsity*” (the text doesn’t really correspond to what the paper describes); “*Structural sparsity*”, which cites a paper by Gao et al. that doesn’t exist, and provides an incorrect definition; and “*Functional sparsity*”, which cites an article with the wrong authors and year (and which doesn’t correspond to the explanation, which is kind of nonsensical as “constraining the activation to follow a specific function” doesn’t mean anything: that’s just what activation functions do in general).

So, to summarize: rehashing things that superficially develop what was in the prompt, with “algorithms” that are not particularly well explained, are often badly attributed, are badly named so that it’s hard to find more about them, or just don’t exist.

Moving on... to “novel” ideas

I’m not going to be as detailed for the rest, as it would get very repetitive very quickly, but in the next section on Hardware the first citation is already wrongly attributed, and we keep the same pattern: superficial explanations that “sound right” but are often imprecise, incomplete or just wrong, with citations that don’t match (when they exist).

In the “compression” section, we see the same techniques that were already presented before (pruning, quantization), so no new information.

At some point, Mashrin asks for a novel idea for future research in the area of compression for large-scale neural networks. This moves again away from the “all-written-by-ChatGPT” concept, but it’s interesting. Lack of novelty is often seen as one of the main limitations of LLMs. So what can ChatGPT come up with?

20. ChatGPT: Certainly, here is a novel idea for future research in the area of compression for large-scale neural networks:

One potential direction for future research is the development of compression algorithms that adaptively adjust the level of sparsity in the network based on the specific characteristics of the input data. For example, the sparsity level could be increased in regions of the network where the data is more predictable or well-behaved, and decreased in regions where the data is more complex or noisy. This approach could potentially improve the efficiency and sustainability of the network while maintaining its performance on a wide range of tasks.

To implement this idea, one could design an optimization algorithm that jointly optimizes the network weights and the sparsity pattern of the network based on the input data. The sparsity pattern could be encoded as a binary mask applied to the weights, which is updated during training based on the gradient information. The optimization algorithm could also incorporate additional constraints or inductive biases, such as structural or functional sparsity, to further improve the efficiency and performance of the resulting network.

A novel idea on compression proposed by ChatGPT

Well, we see a reformulation of ideas that were already presented before (“adaptively adjust the level of sparsity in the network based on the specific characteristics of the input data”, which nearly matches what *Dynamic Sparsity* was presented as). The proposed method also doesn’t really make sense. “Jointly optimizes the network weights and the sparsity pattern of the network based on the input data”? ... yes? The “sparsity pattern” is directly linked to the “network weights” (connections that can be pruned are connections with weights close to zero), and the training is obviously based on the training data, so what does that even mean? The rest of the explanation similarly makes no sense (but it “reads” nicely!)

ChatGPT is prompted for more novel ideas.

We get... really rewording of the same, or of previously explained concepts, or so vague as to be completely unusable.

Anyway, I think we get the point. The conclusion is also mostly empty of actual content. It’s basically all “sparsity would be more efficient, so it’s nice, but there are tradeoffs, and it’s difficult”.

Finally, Mashrin asks for an abstract and bibtex citations of the references. The references, as we’ve seen, are mostly accurate, but sometimes made up, which is exactly what you want from an academic paper!

2. Taking a step back

So all of these results from the prompts were then compiled into a full paper. At least one citation has been changed in the full paper from the prompts results (LeCun, 1990 has become Liu, 2015, I haven’t gone through all), but otherwise it’s just some reformatting.

What can we make of all of this? Here are my main thoughts:

ChatGPT didn't understand the assignment

That's really important to note again, I think, because it shows a key problem with LLMs – they don't have, nor understand, intent. Well, they don't *understand* anything, but let's move on from that.

The original prompt was to “suggest a paper for the below conference workshop on Sparsity in Neural Networks”. This is not the kind of venue where you try to write a global review of the whole domain discussed in the workshop. Such workshop are typically for really novel – often work-in-progress – ideas that can move the field forward in one or a few of the specific topics ([just look at the [list of paper from the 2022 edition](#)]). Everything that is in this ChatGPT paper would (hopefully) be considered basic, common knowledge for attendees of such a workshop.

I mean, the parts that are correct.

It's really just a (excellent) word generator

Some people seem to get really upset when they are reminded that LLMs are “stochastic parrots”, but this paper is a great example of how true that is.

It's really just a word generator.

The structure of the “proposed paper” is taken directly from the prompt. For all the “detailed” section, it's a lot of superficial, sometimes nonsensical paragraphs that are “on topic” but don't bring anything new to the table.

It doesn't even really work as a “quick review”, which would be fine for personal usage if not for publication, because too much of it is just empty of content or wrong. The citations are sometimes relevant, but most of it is way too old to be really good starting points to get the state-of-the-art.

Can ChatGPT produce valuable science?

From what I can see here: no. I can't say *exactly* how I would feel seeing this as a reviewer *without* knowing that it was a ChatGPT-generated text, or without going through the prompt, but it would certainly be a hard reject from me.

Besides the fact that it's not appropriate for the workshop, there are too many obvious red flags. The superficiality, the repetitions, the citations that don't match the text... These are all things that I notice when I review papers or grade student work, and there are so many of them here that there is no chance that I would let it pass.

I'm not enough of a specialist in questions of sparsity techniques and efficiency in neural networks, so I can't really judge how wrong it is... but that's even more of a red flag: even as a non-specialist, it's very obviously wrong in many places.

Can ChatGPT be useful as a study pal?

If it's a topic you don't actually care about and you just want to have a superficial level of knowledge to get the bare minimum to pass a class, sure.

Otherwise I would avoid.

3. Conclusions

I was skeptical before reading the paper, but I tried to keep an open mind (believe it or not!).

I must say I'm absolutely impressed by the quality of the text. It reads like a scientific paper. I totally understand why many people see this as amazing.

But in the end... Stochastic parrot remains the best description there is. As soon as you go beyond the superficial level of "reading the text", and you try to parse its meaning and verify what it says, it breaks down into noise.