NTT DATA
Trusted Global Innovator

Light-weight Model Technologies to Reduce CO_2 Emissions from AI

NTT DATA Corporation
Toyosu Center Building, 3-3, Toyosu 3-chome, Koto-ku, Tokyo
Phone: +81-3-5546-8202
https://www.nttdata.com/global/en/

Contents

CHAPTER.1

Introduction

In recent years, while Artificial Intelligence (AI) has improved in accuracy through technological improvements, CO_2 emissions have also rose due to the increased amount of electricity required for its development and use.

Since the advent of BERT, which revolutionized natural language processing, large-scale AI models have dramatically increased accuracy in areas such as natural language processing, image processing, and speech processing, and now AI is being applied to various business transformations. However, large-scale AI models demand substantial electric power in both the algorithm training and application phases.

Power consumption caused by AI is expected to increase more and more in the future, and it is imperative that efforts be made to reduce energy consumption.

This whitepaper discusses global trends towards reducing AI power consumption and explores software approaches to achieve this objective.

CHAPTER.2

AI and Power Consumption

2.1 The impact of AI on the global environment

Applied Materials CEO Gary Dickerson estimates that with the current rate of AI adoption and no innovation in hardware and software, the share of global power consumption in data centers will increase from 2% in 2019 to 10% in 2025. [1]

According to the Center for Low Carbon Society Strategy, data center power consumption is expected to increase from 14 TWh to 90 TWh in Japan and from 190 TWh to 3,000 TWh globally from 2018 to 2030. [2] Thus, the energy demand attributable to AI is also expected to rise sharply in the future, and efforts are required in order to reduce its power consumption.

2.2 Breakdown of power consumption by AI

There are two major phases in developing and leveraging AI models using deep learning: training and inference.

Training phase

It is the process by which the AI model developer loads a large amount of training data into the model to teach it to perform the desired AI task (e.g., machine translation, image classification).

The training phase can be roughly divided into two phases: “pre-training” for developing general models, and “fine-tuning” for optimizing the model for individual tasks. Here is an example of the development phases in natural language processing for a large-scale language model.

Pre-training develops generic, pre-trained models using large datasets.
It involves training a huge number of parameters, such as Google's proposed BERT model with 340 million parameters, or OpenAI's proposed GPT-3 with 175 billion parameters, which also required massive amounts of data. Thus, the power consumption required for training is enormous, and it is said that GPT-3 required 1,287 MWh of electricity for training, which is equivalent to 552 tons of CO_2 emissions. However, the development of pre-trained models is limited to a few companies and research institutes, and general AI developers can utilize these pre-trained models.

Fine-tuning uses relatively small amounts of data to optimize the pre-trained models for individual tasks. After setting and preparing data for specific tasks such as question answering, document summarization or classification, the model's accuracy is improved by training on thousands to millions of data samples of the task to be solved. It consumes less power than pre-training, but it must be done for each individual AI development project, which means that it must be performed more frequently.

1) https://observer.com/2019/08/artificial-intelligence-bitcoin-cloud-computing-climate-change/
2) Impact of Progress of Information Society on Energy Consumption (Vol.1): Current Status and Future Prospects for Power Consumption of IT Equipment
https://dl.ndl.go.jp/view/prepareDownload?itemId=info%3Andljp%2Fpid%2F11546567&contentNo=1
3) David Patterson et al., Carbon Emissions and Large Neural Network Training

Inference phase

It is the process by which a user inputs new data into a trained model to obtain its prediction results as output.

Comparison of training and inference phases

In general, the training phase consumes much more power per attempt than the inference phase. As the amount of data and the number of parameters increases, the power consumption of the model grows as well. This holds true for both the training and inference phases. However, the training phase needs be completed only once for a fixed amount of data, whereas the inference phase is executed by the model for every new input data, which can be tens of thousands to hundreds of billions of times or more, depending on the application.

So, during the lifecycle of an AI model, which consumes more power, the training phase or the inference phase?

- Amazon Web Services customers say that inference accounts for 90% of the cost of their machine learning services. [4]
- Jensen Huang, CEO of NVIDIA, estimated that in 2019, 80-90% of machine learning costs were due to inference.[5]
- Even Google, which does extensive model development in-house, reported in 2022 that the power it consumes for machine learning comes 60% from inference and 40% from training. [6]
- Facebook reports that it uses AI models via its own servers to perform 200 trillion inferences per day on smartphones worldwide. [7]

As explained above, inference consumes less power per run than training, but the number of executions is orders of magnitude larger, so the total power consumption of inference is greater than training in most cases as can be seen from Figure 1.

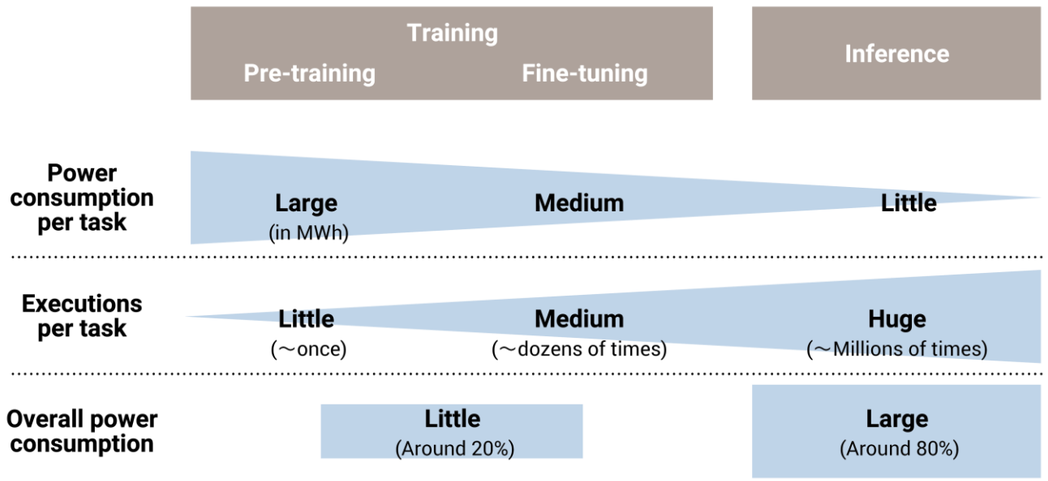

Figure 1: Relationship between power consumption and number of executions for each AI model development phase.

4) https://aws.amazon.com/jp/blogs/aws/amazon-ec2-update-inf1-instances-with-aws-inferentia-chips-for-high-performance-cost-effective-inferencing/

5) https://www.hpcwire.com/2019/03/19/aws-upgrades-its-gpu-backed-ai-inference-platform/

6) https://ai.googleblog.com/2022/02/good-news-about-carbon-footprint-of.html

7) https://engineering.fb.com/2018/05/02/ai-research/announcing-pytorch-1-0-for-both-research-and-production/

2.3 Trends in AI power consumption

Trends in computational complexity and power consumption in training

AlexNet was introduced in 2012 as a convolutional neural network (CNN) for object recognition from images. It won the Image Classification Contest (ILSVRC) at the time with an error rate of 15.3%, more than 10 points lower than the runner-up, and sparked interest into deep learning.

While AlexNet boasted groundbreaking accuracy at the time, the technological innovation around deep learning algorithms has been tremendous, and the computational complexity of the training required to achieve the same accuracy as AlexNet has been reduced to 1/44 in the 7 years from 2012 to 2019. This means that approximately every 16 months the required computational complexity is halved. [8]

On the other hand, the size and complexity of deep learning models has also exploded, as the more parameters are used for training, the higher the accuracy becomes. The computational power required to train state-of-the-art models in the 6 years from 2012 to 2018 increased 300,000-fold. This means that about every 3.4 months the required computations double, increasing at a much faster rate than the reduction in computation due to algorithmic improvements, leading to increased power consumption and CO_2 emissions. [9]

Trends in computational complexity and power consumption in inference

The computational effort required for inference is also on the rise.

Canziani et al. (2017) take issue with the fact that since the advent of deep learning, the goal of AI models has been skewed towards achieving the highest accuracy regardless of the actual inference time, and point out that improvements in model accuracy have had an impact on power consumption and computational resource usage. [10]
Desislavov et al. (2021) warn about trends in the computational complexity and power consumption of inference in the fields of computer vision (CV) and natural language processing (NLP). [11]
According to this report, the computational effort for inference required to run the state-of-the-art model in CV increased approximately 3,700-fold from 1.42 GFLOPs (2012) to 5,270 GFLOPs (2021), while the computational power required by NLP inference increased approximately 13,700-fold from 54 GFLOPs (2017) to 740,000 GFLOPs (2020).[12]
Although the energy efficiency of GPUs has improved over the years, the amount of computational power required by machine learning algorithms has multiplied, resulting in a 300-fold increase in power consumption for inference in CV and 500-fold or more in NLP.

8) https://arxiv.org/abs/2005.04305

9) https://openai.com/blog/ai-and-compute/

10) https://arxiv.org/abs/1605.07678

11) https://arxiv.org/abs/2109.05472

12) GFLOPs: A commonly used metric to express the computational cost of a model. Billion times FLOPs representing the number of floating-point operations.

CHAPTER.3

Approaches to Reducing AI Power Consumption

There are three main approaches to reducing AI power consumption:

3.1 Reductions in the cloud/data centers

Reductions in the power consumption of facilities where computational resources of training and inference are actually consumed. In a breakdown of data centers' electricity costs, it is reported that 50% is for servers, 30% is for power and cooling, and 20% is for others including storage. Energy savings for these facilities will lead to lower power consumption.

- For example, the data centers' Immersion Cooling System, which is being tested at NTT DATA, has been confirmed to reduce the power consumption required for cooling by up to 97% compared to conventional data centers. [13]
- Through a visualization system of the data centers' indoor environment it is also possible to reduce cooling energy by approximately 35% by suppressing server room overcooling. [14]
- Google says it has used machine learning to optimize the power used to cool servers in its data centers, resulting in a 40% reduction in power consumption. [15]

3.2 Reductions in hardware

Reductions by optimizing the power use of individual devices and components, such as the individual servers in the cloud or in data centers, or the edge devices that perform inference and the respective GPUs and CPUs inside them. Energy savings may be achieved by selecting the most appropriate hardware based on the characteristics of the AI model and the constraints of the devices on which inference is performed. Hardware manufacturers are also working to develop devices specifically for deep learning training and inference, and the power consumption for performing the same calculations is decreasing year by year.

- According to the report by Desislavov et al. mentioned earlier, GPUs were able to process less than 7 GFLOPs/W of computation in 2010, but as of 2020, they were able to process more than 100 GFLOPs/W, and some CV and NLP specialized GPUs are now capable of processing more than 300 GFLOPs/W.
- Processors specialized for deep learning are also being developed. Preferred Networks, for example, has developed a dedicated chip optimized for "matrix operations", a hallmark of deep learning, to improve execution performance while lowering costs. In 2020 and 2021, a supercomputer equipped with this chip has won the world's No. 1 position in the Green500 power-saving performance ranking.

3.3 Reductions in software

Reduction of power consumption by AI developers' own algorithms. Examples include AI model compression, selection of file formats suitable for data analysis, and selection of energy-efficient programming languages.

- **AI model compression:** Reducing the computational complexity of a model can shorten inference time, reduce memory usage, and reduce power consumption during inference. Details are described in the next CHAPTER 3. Approaches to Reducing AI Power Consumption chapter.
- **Programming languages:** Interpreted languages such as Python are relatively slow in processing speed as they execute programs while interpreting code sequentially. Choosing a pre-compiled language such as C or C++ can improve processing speed and potentially reduce power consumption.
- **File format:** Model training requires reading and pre-processing large amounts of data, so efficient data processing and storage can reduce power consumption. For example, Parquet is a column-oriented file format designed for data storage and retrieval that can process data more efficiently than CSV files.

13) https://www.nttdata.com/jp/ja/news/release/2022/060601/
14) https://www.nttdata.com/jp/ja/news/release/2022/072901/
15) https://www.deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-by-40

CHAPTER.4

AI and Power Consumption Reduction by Model Compression

There are three typical methods of model compression to reduce power consumption during inference:

(1) Pruning
(2) Distillation
(3) Quantization

Each of these model compression methods has a tradeoff between power reduction and model prediction accuracy. Compressing a model means that, although there is a reduction in inference time and power consumption, the ability of the model to understand words and grammar rules decreases and the accuracy of the model becomes worse.

NTT DATA is developing an approach to optimize algorithms and parameters by combining these methods, while keeping accuracy as high as possible.

4.1 Pruning

In general, pruning a model equates to deleting the parameters with the smallest importance (usually the nodes or weights with the value closest to 0) from a deep learning model.

There are two approaches to model compression: Matrix Pruning, which removes rows and columns of a weight matrix, and Layer Removal, which removes the layers themselves from the model (Figure 2).

Matrix Pruning

When considering the reduction of power consumption during inference, it is effective to reduce the overall computations of the model by deleting rows and columns from the weight matrices.

Layer Removal

For large language models that consist of many layers, it is possible to remove some layers from the model. A moderate amount of layer removal has been shown to have low impact on the final accuracy. [16]

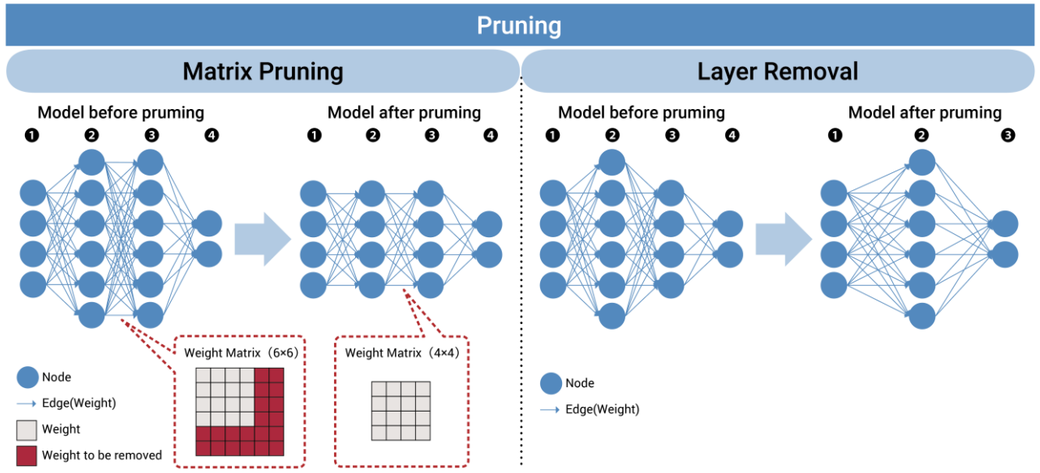

Figure 2: Image of Pruning

16) https://arxiv.org/pdf/2004.03844.pdf

4.2 Distillation

Distillation is one of the most common ways to compress a model. During training, a small “student” model tries to mimic the output of a larger finetuned model, the “teacher”, so that it learns to closely match the same output of the teacher model. This approach generally improves the accuracy of the student model, but it requires training with large amounts of data to achieve sufficient improvements in accuracy. Some examples of well-known applications include DistillBERT, which is widely used as a small-scale language model, and TinyBERT and MobileBERT, which are small enough to be used on edge and mobile devices.

4.3 Quantization

Quantization differs from the previous two methods in that it approximates the calculations by shortening the number of bits of the model’s weights, rather than changing the structure of the model. This means that the model’s accuracy does not change substantially, but the memory and computations required for the model are greatly reduced. However, since these calculations are not supported by most GPUs, they are mostly applied to inference on edge devices and mobile devices that primarily use CPUs.

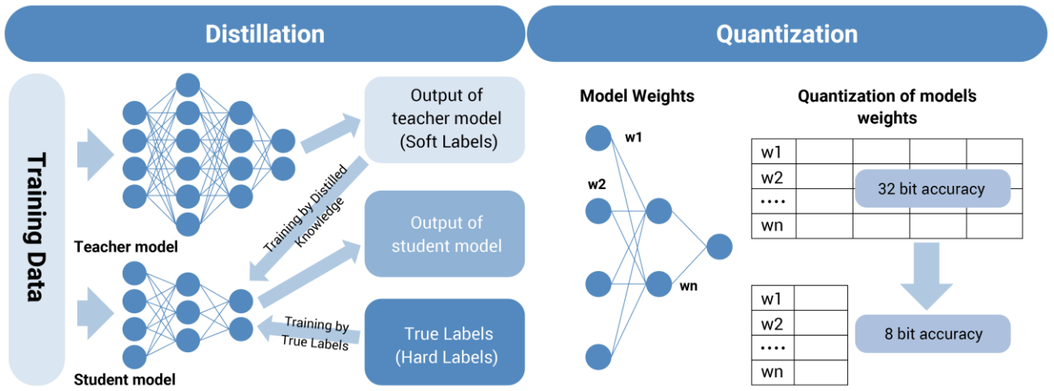

Figure 3: Image of Distillation and Quantization

4.4 Model compression techniques combination

Each approach has its advantages and disadvantages, but when properly combined, they have the potential to create sustainable, green AI models that can meet a variety of requirements.

The speed of technological evolution in AI in recent years has been rapid, and since BERT, pre-trained large-scale AI models have been appearing one after another, so we believe that the technology to reduce their power consumption must be versatile enough to respond to various AI models in a timely manner, rather than specializing in a particular model.

At NTT DATA, we have been verifying the optimal combination of methods in terms of accuracy, power consumption (CO_2 emissions), and availability on GPUs, for any language model, rather than limiting it to a specific one, with the aim of establishing a general-purpose model compression methodology that can reduce power consumption at inference time. In the next chapter, the method under verification by NTT DATA is explained in detail.

CHAPTER.5

Verification of Model Compression Techniques

5.1 Conditions of verification

While the three methods of model compression described in the previous section are common, the optimization and the combination of these methods is left to the know-how of each AI developer. NTT DATA is aiming to realize a combination of optimal lightweighting methods by leveraging its expertise in natural language processing.

The advantage of our method is that it can build lightweight models without any adjustments for any model, not only BERT, but also T5 and the GPT-3 structured OPT model among others.

In order to evaluate this method, we compare the accuracy and amount of CO_2 emissions required to perform the inference.

For the verification of our approach, to adjust the degree of model compression, combinations of Layer Removal, Matrix Pruning, and Distillation were applied under various conditions.

- Specifically, for each language model, the model obtained by "fine tuning" without any lightweighting was used as the standard, and three types of compression were applied, in descending order of model size: "2/3 Layer," "1/3 Layer," and "1/3 Layer + Matrix Pruning".
- In addition, Distillation was applied to all lightweight models to improve accuracy.

Through this validation, Layer Removal and Matrix Pruning were found to be optimal when Layer Removal was applied first. Applying Layer Removal first allows for model compression with minimal accuracy degradation. Then, Matrix Pruning is applied while performing fine-tuning, which compresses the weight matrices while non-pruned weights are trained to relearn the lost information.

Other validation conditions are as follows:

- BERT, GPT-2, T5 and OPT were used as target language models.
- As datasets, IMDb and GLUE, which are widely used benchmark datasets in natural language processing were used to verify the accuracy. [17] [18]
- The power consumption required for inference was converted into the relative CO_2 emissions for evaluation. Relative accuracy (%accuracy) and relative emissions (%emissions) are compared with the accuracy and CO2 emissions of the finetuned model, which is set to 1.
- All inference was performed on a NVIDIA A40 GPU.

When using a CPU device for inference, quantization can also be applied. Although the following results do not include quantization as they were obtained using GPUs, the results of a separate verification confirmed that the quantized models can reduce both inference time and CO_2 emissions by half while maintaining a relative accuracy of 99% compared to non-quantized models.

However, since using only a CPU requires longer inference times and releases more emissions compared to GPUs in the first place, we suggest quantization to be applied to edge and mobile devices that do not have GPUs.

In the next section, we show the results of the method NTT DATA is currently validating, which include the accuracy, as well as the CO_2 emissions required for such inference.

17) IMDb: A widely used data set for verification of natural language processing, where movie review comments are binarized as positive or negative. It contains 25,000 data for both training and testing.

18) GLUE: A data set for natural language processing, consisting of 9 tasks that are a relatively high degree of difficulty compared to IMDb. The number of data varies by task.

5.2 Verification Results

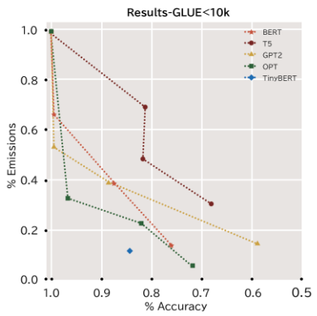

Figure 4: Results for the GLUE dataset (Data size less than 10,000)

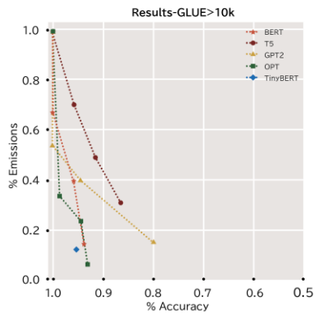

Figure 5: Results for the GLUE dataset (Data size greater than 10,000)

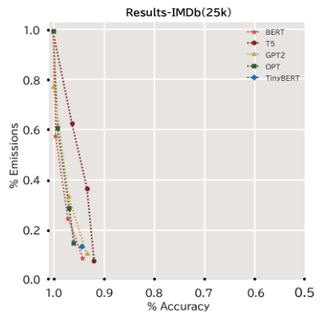

Figure 6: Results for IMDb dataset (Data size 25,000)

The figures above are the results of the validation of four different models: BERT, T5, GPT-2, and OPT. Since the number of data samples for GLUE varies by task, the results were divided depending on the size of the dataset, and we group together tasks with less than 10,000 data samples (<10K) and tasks with more than 10,000 data samples (>10K).

Figure 4 shows the results for GLUE (<10K), Figure 5 for GLUE (>10K), and Figure 6 for IMDb. The results of TinyBERT, a state-of-the-art pre-trained distilled model, are also shown for comparison with the BERT verification results.

As an overall trend, it was confirmed that CO_2 emissions can be reduced for both the IMDb and GLUE tasks, while maintaining relative accuracy as high as possible for all language models.

Task-specific Trends

- For GLUE (<10K) we can maintain 95% or greater accuracy when applying a low amount of layer removal. However, as the model weight reduction progresses, the information lost by pruning cannot be effectively relearned, resulting in a high accuracy reduction.
- For GLUE (>10K), accuracy degradation is suppressed as the longer training of pruned models allows them to learn the lost information.
- IMDb, with its large 25,000 dataset size, showed better performance for all language models. In the most lightweight case, total emissions were reduced to about 10% while maintaining a relative accuracy higher than 90% for all models.

Model-specific Trends

- Looking at the results of the BERT model in more detail, for GLUE (< 10K) tasks with small datasets, the accuracy decreases as pruning increases. TinyBERT also shows some accuracy degradation, but the effect remains comparatively little. On the other hand, in the case of IMDb and GLUE (>10K) with larger datasets, our model compression approach can produce a good relative accuracy, keeping it on the same levels as TinyBERT. The overall trend is similar for OPT, whose structure as a language model is similar to that of BERT.
- T5 shows the highest accuracy degradation among the four language models overall. T5 is a sequence-to-sequence generative model, which requires a higher level of language understanding to generate sentences compared to the other models. As a result, the negative impact on the accuracy is more significant, especially in conditions with few data samples such as GLUE (<10K), even when few layers are deleted.
- For GPT-2, the overall trend is similar to BERT and OPT, but the rightmost results of the graphs, obtained by applying Matrix Pruning, show a different tendency. In fact, GPT-2 has a rare architecture where weight matrices are appended to one another, meaning that the order of rows and columns is extremely important. Applying Matrix Pruning breaks the learned order and significantly reduces the accuracy of the model.

The above results show that depending on the model and its architecture, different pruning options should be utilized. As an example:

- If there is little available data, high accuracy can be maintained by limiting layer removal to a reduction of only few layers, except for generative models such as T5.
- For models with specific architectures such as GPT-2, it is suggested to not utilize Matrix Pruning, but perform only Layer Removal.

If model weight reduction is implemented correctly, model compression can create AI models that can perform inference at a fraction of the cost and computational complexity of the original models, with as high as a tenfold reduction in emissions.

An additional benefit of this approach is that also model size and inference time are reduced with a similar trend as CO_2 emissions, as shown in Table 1. While Table 1 shows the results of BERT, similar results were obtained for all the other language models under consideration.

As such, we can create smaller models that meet a variety of constraints such as accuracy, CO_2 emissions, model size, and inference time, regardless of the type of language model, by combining different model compression techniques.

Model	Model Type	% Accuracy	% Size	% Time
BERT	Finetuned	1	1	1
BERT	8 layers (2/3)	99.7%	74.2%	86.9%
BERT	4 layers (1/3)	91.9%	48.2%	61.4%
BERT	4 layers (1/3) + Matrix pruning	85.6%	21.5%	26.9%

Table 1: Relative Accuracy, Size and Time results for BERT

CHAPTER.6

Future Outlook

As described above, the balance between accuracy and CO_2 emissions varies depending on the type of model and dataset used. However, a uniform approach can be applied to validate the potential reduction of emissions while preserving most of the model's accuracy. By utilizing this technique during AI development, it is possible to optimize the search for models that satisfy customer demands, such as specific accuracy and CO_2 emissions targets.

Further research is underway on AI model lightweighting methods. In the future, new techniques are expected to be developed that can reduce CO_2 emissions with less accuracy degradation and without spending time training. For example:

- Practical application of GPUs that support quantization
- Application of sparse matrices to deep learning leading to faster matrix computations

We will continue to work on the reduction of CO_2 emissions and optimization of accuracy through appropriate combinations of model lightweighting methods, including technologies that will emerge in the future. Moreover, in addition to language models, we will apply and verify the effectiveness of these methods to image, sound, and multimodal models.

As described in this white paper, there are different approaches to CO_2 reduction from a software perspective. NTT DATA is a member of the Green Software Foundation (GSF), a non-profit organization that organizes and standardizes these methods, in order to strengthen its efforts towards decarbonization. By developing environmentally friendly AI, we will continue to promote innovation and the realization of a sustainable society through AI.

Authors Information

Yuji Nomura
D&I Technology Department, System Engineering HQ, Technology and Innovation General HQ

Atsushi Nakano
D&I Technology Department, System Engineering HQ, Technology and Innovation General HQ

Paolo Tirotta
D&I Technology Department, System Engineering HQ, Technology and Innovation General HQ

March 2023