The Impact of Knowledge Distillation on the Energy Consumption and Runtime Efficiency of NLP Models

Ye Yuan, Jingzhi Zhang, Zongyao Zhang, Kaiwei Chen, Jiacheng Shi, Vincenzo Stoico, Ivano Malavolta
Vrije Universiteit Amsterdam, The Netherlands
{y.yuan3, j.zhang6, z.zhang14, k.chen2, j.shi2}@student.vu.nl, {v.stoico, i.malavolta}@vu.nl

ABSTRACT
Context. While models like BERT and GPT are powerful, they require substantial resources. Knowledge distillation can be employed as a technique to enhance their efficiency. Yet, we lack a clear understanding on their performance and energy consumption. This uncertainty is a major concern, especially in practical applications, where these models could strain resources and limit accessibility for developers with limited means. Our drive also comes from the pressing need for environmentally-friendly and sustainable applications in light of growing environmental worries. To address this, it is crucial to accurately measure their energy consumption.

Goal. This study aims to determine how Knowledge Distillation affects the energy consumption and performance of NLP models.

Method. We benchmark BERT, Distilled-BERT, GPT-2, and Distilled-GPT-2 using three different tasks from 3 different categories selected from a third-party dataset. The energy consumption, CPU utilization, memory utilization, and inference time of the considered NLP models are measured and statistically analyzed.

Results. We observed notable differences between the original and the distilled version of the measured NLP models. Distilled versions tend to consume less energy, while distilled GPT-2 uses less CPU.

Conclusion. The results of this study highlight the critical impact of model choice on performance and energy consumption metrics. Future research should consider a wider range of distilled models, diverse benchmarks, and deployment environments, as well as explore the ecological footprint of these models, particularly in the context of environmental sustainability.

ACM Reference Format:

1 INTRODUCTION
"Amidst the rapid evolution of Natural Language Processing (NLP), a pressing question arises: Can we sustain technological advancement while being environmentally conscious?" This inquiry launches our exploration into Knowledge Distillation within NLP models—a technique poised to redefine efficiency in AI.

NLP’s landscape has dramatically transformed with the advent of models like BERT and GPT-2, setting new standards for applications ranging from chatbots to content generation. Nevertheless, their extensive resource demands spotlight a critical challenge: balancing computational prowess with environmental responsibility[1]. Knowledge Distillation emerges as a compelling solution, distilling the essence of these behemoths into more compact, efficient models without sacrificing their core capabilities[2].

This study delves into the intricacies of Knowledge Distillation, driven by the urgent need to reconcile the power of NLP models with practical and sustainable usage[3]. By focusing on energy consumption, we aim to illuminate a path for developing more efficient AI models, contributing to a more responsible technological future. In an era where AI’s environmental impact is increasingly scrutinized, our research is catalyzed by the imperative to optimize the energy efficiency of state-of-the-art NLP models[4]. We investigate the dichotomy between their remarkable capabilities and substantial resource demands, aiming to bridge this gap with Knowledge Distillation. Our motivation is twofold: advancing AI’s frontiers while fostering sustainability, a balance critical for the continued growth and accessibility of NLP technologies[5].

The main goal of this paper is to empirically assess whether Knowledge Distillation can significantly enhance the energy efficiency of NLP models without compromising their performance.

We achieve the above-mentioned goal by answering the following two research questions.

RQ1. How does the Knowledge Distillation affect the energy consumption of NLP models? We explore if the process of distillation, which typically involves transferring knowledge from a larger model to a smaller one, has any significant influence on the energy footprint. To answer this RQ, we test the NLP models with the GLUE benchmark [6] and measure the energy consumed performing the SST-2, STS-B, and MNLI tasks.

RQ2. How does the Knowledge Distillation affect the performance of NLP models? This question investigates if the distilled version of the NLP models maintain, enhance, or potentially compromise their performance compared to the original version. As done for energy consumption, we observe both groups of models, i.e., the original and distilled version, executing the SST-2, STS-B, and MNLI tasks included in the GLUE benchmark.

The results of our study are multi-faceted, but also congruent. We observed a statistically significant difference in energy consumption among the considered NLP models, with distilled models tending to consume less energy compared to their non-distilled counterparts. We also observed statistically significant differences among models in terms of performance metrics (i.e., CPU usage,
memory usage, and inference time). Our analysis revealed significant variations in CPU usage across different models. The results indicated that distilled models exhibited lower CPU usage compared to their non-distilled counterparts, suggesting a more efficient utilization of processing resources. We observed that memory consumption patterns varied significantly between models, with distilled models showing higher efficiency in memory utilization than without distilled. This aspect of performance is crucial for applications where memory resources are a limiting factor. Our study also suggests there are notable differences in inference times, with distilled models achieving faster processing speeds, indicating their suitability for real-time applications.

The main contributions of this work are (i) an empirical assessment of the impact of Knowledge Distillation on the performance and energy consumption of BERT and GPT-2 NLP models, (ii) a discussion of the obtained results, and (iii) the replication package of the study [7].

2 EXPERIMENT PLANNING AND EXECUTION

2.1 Subjects Selection

In this study, we compare GPT-2 and BERT, with their corresponding distilled version. GPT-2 has shown exemplary performance across a range of NLP tasks. However, its computational demands are high due to its extensive parameter count. BERT is another transformer-based model that has set new performance benchmarks on several NLP tasks. Similar to GPT-2, BERT has a high computational demand due to its large model size. Distilled GPT-2 and distilled BERT are both lighter versions created through Knowledge Distillation.

The benchmark tasks chosen for this study are derived from the General Language Understanding Evaluation (GLUE) dataset[6], specifically the tasks of SST-2, STS-B, and MNLI. GLUE is a collection of resources for training, evaluating, and analyzing natural language understanding systems. The dataset encompasses various tasks that evaluate the capacity of models to understand the nuances and subtleties of human language. The rationale for its selections is that GLUE amalgamates multiple tasks that span a wide range of NLP challenges, also, GLUE tasks derive from a mix of sources, ranging from news articles to fiction, thereby presenting a more holistic challenge for models. SST-2 is a single-sentence categorization task, that contains human annotations of sentences from movie reviews and their sentiment. This task is given the sentiment of a sentence, and the categories are divided into two types of positive sentiment and negative sentiment. STS-B is a benchmark test for evaluating the capabilities of natural language processing models, focusing on the performance of the model in measuring the semantic similarity of two sentences. MNLI requires the model to determine the relationship between one sentence (premise) and another (assumption). It uses texts from a wide range of genres and topics providing a more complex and diverse testing environment for the model to assess its ability to generalize across different contexts. In each of the assigned tasks – SST-2, STS-B, and MNLI – we perform three rounds of execution to ensure consistency and to account for any potential differences in individual runs.

2.2 Study Design

The independent variable of this study is about the application of Knowledge Distillation to NLP models; this variable has two treatments, i.e., whether knowledge distillation is applied or not. Differently, the dependent variables of this study are:

- **Energy Consumption (J):** The energy consumption of performing a run in the experiment is measured in Joules.
- **CPU Utilization (%):** The percentage of time the CPU is actively executing instructions, as opposed to being idle, is calculated as Execution Time.
- **Memory Utilization (%):** The percentage of a computer system’s memory that is used by the NLP model during its operation.
- **Inference Time (s):** The time elapsed in seconds from the moment a model receives an input until it produces a result.

CPU Utilization offers an indication of the computational resources required to process tasks, high CPU usage can result in system overheating, overloads, or performance degradation. Memory Utilization offers a direct insight into the size and complexity of a model. Distilled models, being typically smaller, should consequently occupy less memory. However, this does not directly infer a performance advantage and needs to be considered in conjunction with the other two metrics. Inference Time directly relates to the speed at which a model processes an input and produces an output. We expect to observe significant differences in processing time for the distilled model. Throughout the experiment, we collect and analyze data based on these metrics to address the specified questions and provide insights for developers regarding the energy efficiency of NLP model choices.

Based on the aforementioned variables, this study is designed as a one factor – multiple treatments experiment.

2.3 Experiment Execution

The experiment is configured and orchestrated via Experiment Runner [8], a tool for the automatic orchestration of measurement-based experiments.

2.3.1 Setup. The experiment is executed on a Linux server equipped with a 1TB hard drive, 32GB of RAM, and a 3.4GHz x 8 core CPU. Once connected via SSH, Experiment Runner executes each individual benchmark on the server. This involves initiating the necessary commands through the terminal interface to trigger the execution of the NLP Models on the designated benchmarks. During these runs, we closely monitor energy consumption and performance. Each run encompasses testing an NLP Model on three benchmarks, performed ten times consecutively.

2.3.2 Measurement. For each run of the experiment, an NLP model is tested on the three considered benchmarks. Since the goal of our study is not the performance of the model on a specific benchmark therefore we do not consider it necessary to collect data from the three benchmark tests separately. In order to take into account the variability of the collected variables, each run is executed 10 times consecutively. We introduce a 5-minute cool-off period after each run to take into account the problem of CPU performance degradation due to temperature increase when the NLP model performs inference tasks.
To gather the values of our dependent variables, we integrate Experiment Runner with PowerJoular and Linux’s ps command. PowerJoular [9] is a monitoring tool that can be used to monitor the CPU usage and energy consumption of a particular process. The ps command, which is one of the most commonly used commands in Linux, can capture the CPU usage as well as memory usage of processes. Inference time is collected via Python’s time module.

3 RESULTS

3.1 Descriptive Statistics

Before our statistical analysis, we carefully check the correctness of the collected measures, e.g., in case the energy profiler produced values not matching the technical specifications of the machine. We identified two problematic runs, which involved re-running those runs to guarantee that each model had precisely 10 valid runs.

We initiate our analysis by exploring the data via basic descriptive statistics and visualization. Table 1 presents the average values of all our measured metrics across all models. The analysis reveals intriguing differences between the models. For both GPT-2 and BERT, the distilled versions exhibited lower CPU and memory usage, reduced energy consumption, and faster inference times.

![Figure 1: CPU usage, memory usage, energy consumption, and total inference time across the four language models.](image)

<table>
<thead>
<tr>
<th>Model</th>
<th>Average CPU usage (%)</th>
<th>Average memory usage (%)</th>
<th>Total inference time (s)</th>
<th>Total energy consumption (J)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT</td>
<td>36.7</td>
<td>5.68</td>
<td>886.22</td>
<td>26892.170</td>
</tr>
<tr>
<td>Distilled-BERT</td>
<td>26.4</td>
<td>5.55</td>
<td>724.77</td>
<td>16234.187</td>
</tr>
<tr>
<td>GPT-2</td>
<td>46.9</td>
<td>3.95</td>
<td>797.90</td>
<td>31228.713</td>
</tr>
<tr>
<td>Distilled-GPT-2</td>
<td>48.3</td>
<td>3.17</td>
<td>557.61</td>
<td>22464.590</td>
</tr>
</tbody>
</table>

Table 1: Overview of collected measures.

Specifically, by referring to Table 1 and the box plots in Figure 1 we make the following observations:

- **Energy Consumption**: the distilled models, i.e., Distilled-BERT and Distilled-GPT-2, consume 16,234.187J and 22,464.590J, respectively. Distilled-BERT exhibits a remarkable 43.96% reduction in energy consumption compared to its non-distilled counterpart, demonstrating its higher energy efficiency. Distilled-GPT-2 exhibits a notable 28.1% decrease in energy consumption compared to its non-distilled counterpart, GPT-2.

- **Inference Time**: distilled models tend to have shorter inference time. Distilled-BERT takes an average of 674.770ms, which is faster than BERT’s 891.072ms. Similarly, Distilled-GPT-2 took 557.614ms, a noticeable reduction from GPT-2’s 798.988ms.

- **CPU Utilization**: Distilled-BERT shows a sensible reduction in CPU utilization (0.264%) when compared to BERT (0.367%). In the case of GPT models, Distilled-GPT-2 (0.483%) exhibits a sensible lower CPU utilization than GPT-2 (0.496%).

- **Memory Utilization**: for BERT models, the difference in memory utilization is marginal, with Distilled-BERT using 5.546% and BERT consuming 5.686%. However, for GPT models, Distilled-GPT-2 tends to use considerably less memory (3.167%) compared to GPT-2 (3.951%).

In Figure 1 we can also observe that the data displays an asymmetrical pattern in some cases, with a few outliers present. These deviations are indicative of a distribution that deviates from perfect normality. This aspect will be the subject of formal investigation in the subsequent phases of our analysis. Additionally, the boxplots reinforce our initial assumption that distilled models consume less energy and exhibit enhanced performance. Testing for normality is a crucial step before hypothesis testing. Depending on whether the collected data is assumed to be normally distributed or not, either paired t-tests or Wilcoxon Signed Rank tests are employed. Figure 2 presents density plots for each metric, considering different models. Within each subplot, one metric is plotted for the four models. These plots illustrate that most data adheres to a bell curve, indicating a tendency towards normal distribution. To delve deeper into normality assumptions, Q-Q plots were generated (see our replication package [7]). While most figures closely follow the diagonal line (highlighted in red), it is important to note that some exhibit slight deviations. Due to our relatively modest sample size, we could expect that the Q-Q plots would display more dispersion and variability. In such instances, discerning a clear linear pattern can be challenging, and deviations from linearity might occur due to sampling variability. Consequently, in the subsequent analysis, we employ quantile-quantile (Q-Q) plots and the Shapiro-Wilk test to rigorously assess if our data adheres to a normal distribution.

**Highlights**—The distilled versions of the measured NLP models tend to exhibit better energy efficiency, faster inference times, and in some cases, reduced CPU utilization and memory usage. The collected data suggests that Distilled-BERT is especially more energy-efficient compared to BERT, while Distilled-GPT-2 offers significant improvements in inference time and memory usage over its original counterpart, GPT-2.
3.2 Hypothesis Testing
In our study, the process of hypothesis testing is conducted in two primary stages. Initially, we apply an Analysis of Variance (ANOVA) to discern if there are any significant differences in our metrics across the various models. ANOVA is a statistical method used to test differences between two or more means, and it is particularly useful when comparing multiple groups simultaneously [10]. This approach provides a holistic view and tests differences across multiple groups. Furthermore, to ensure the validity of our ANOVA results, we conduct Mauchly’s Test for Sphericity. This test is crucial as it checks the assumption of sphericity in repeated measures ANOVA, which is essential for the accuracy of the F-ratio in the ANOVA test. Sphericity refers to the condition where the variances of the differences between all combinations of related groups are equal [11]. Following the ANOVA, we apply a t-test for pairwise comparisons to understand the differences between the specific models. The t-test is a statistical test that allows us to compare the means of two groups and determine if they are significantly different from each other. This step is particularly important to pinpoint where the significant differences lie between individual models [12].

3.2.1 CPU Utilization (RQ1). Our analysis shows a profound impact of the model type on average CPU utilization, with an $F(3, 27) = 4358.705$ and $p < .001$. The observed generalized eta squared (ges) was 0.9969644, indicating a substantial effect size, thus suggesting that the model type significantly influences the CPU utilization. This is particularly noteworthy because it suggests that different NLP models have markedly different computational demands. Sphericity, an important assumption in our ANOVA test, was not met ($W = 0.1398375$, $p = 0.01017544$). This prompted us to apply the Greenhouse-Geisser correction ($\epsilon = 0.69218$), which confirmed the results remained statistically significant at $p < .001$.

3.2.2 Memory Usage (RQ1). The analysis shows a statistically significant difference when analyzing memory usage, $F(3, 27) = 2666.646$, $p < .001$. The substantial effect size (ges = 0.9956913) demonstrates that the type of NLP model in use can have a significant impact on the memory resources consumed. The test indicated that the assumption of sphericity was violated ($W = 0.1418206$, $p = 0.0163451$). Even after using the Greenhouse-Geisser correction ($\epsilon = 0.4912353$), the findings remained robust and statistically significant, $p < .001$.

3.2.3 Total Inference Time (RQ1). The results pointed towards a significant difference among the models in terms of total inference time, $F(3, 27) = 1335.418$, $p < .001$. The robust effect size (ges = 0.9899065) indicates that model choice can substantially affect the time taken for inferences, an essential factor in real-time applications. The sphericity assumption is satisfactorily met ($W = 0.284992$, $p = 0.08669588$).

3.2.4 Energy Consumption (RQ2). The analysis shows a statistically significant difference when analyzing the energy consumed by the models, with $F(3, 27) = 64777.89$, $p < .001$. The extremely large effect size (ges = 0.9998128) underscores the importance of selecting the right NLP model, especially in scenarios where energy efficiency is paramount. Our data met the assumption of sphericity, as confirmed by $W = 0.3804188$, $p = 0.1915562$.

Highlights – The distilled models exhibit significant differences in performance metrics when compared to their original counterparts. The extent and direction of these differences vary depending on the specific metric being considered. Our results indicate that while GPT2 has a slightly higher CPU utilization on average, distilledGPT2 consumes more power. This suggests that while distillation might optimize certain aspects of a model, it could lead to higher power consumption in certain scenarios. This highlights the importance of carefully considering trade-offs when employing model distillation techniques.

4 DISCUSSION
4.1 Energy Consumption (RQ1)
Our primary research question (RQ1) focuses on the energy consumption of distilled models compared to standard NLP models. The results of our hypothesis testing demonstrate a significant difference between the two, with distilled models consuming less energy. This finding is crucial in scenarios where energy efficiency is a priority. However, it is important to note that the choice of model, benchmark tasks, and the computing environment can influence these results. Further research is needed to assess the broader applicability of these findings.

4.2 Performance (RQ2)
The second research question (RQ2) concerns the performance of distilled models and standard ones. Both Distilled BERT and Distilled GPT-2 have significantly lower average inference time compared to their original counterparts, BERT and GPT-2. This indicates that Knowledge Distillation indeed has a positive impact
on model performance, making them more competitive for real-time applications. For memory usage, all models in the experiment performed well. The analysis reveals that the choice of the NLP model significantly impacts memory resources consumed. Distilled-GPT-2, in particular, consumes considerably less memory than the non-distilled GPT-2, while the memory difference between Distilled-BERT and BERT is marginal. In terms of CPU usage, both distilled versions of models have a slightly lower number than their counterparts. Additionally, different models influence CPU usage. To summarize, the distillation model performs better compared to the standard model, however, different experimental environments may have an impact on the results.

4.3 Practical insights for developers

Our results offer practical insights for developers in the NLP field. Distilled models are advantageous in scenarios where energy efficiency is paramount, such as in mobile applications or devices with limited processing power. However, for tasks that demand high accuracy or complex language processing, the original, non-distilled models might be more suitable. This decision-making is critical in balancing the trade-offs between efficiency and computational requirements. Our study highlights the critical impact of model selection in the realm of NLP, particularly in terms of energy efficiency and performance. The findings from this research not only contribute to the technical understanding of distilled NLP models but also offer a perspective on their practical applications and environmental implications. As the field of AI continues to evolve, these insights will be invaluable in guiding the development of sustainable and efficient AI technologies.

4.4 Other remarks

It is important to note that the choice of benchmarks from the GLUE dataset played a pivotal role in our analysis. These benchmarks, encompassing diverse tasks, provided a comprehensive ground for testing. However, it is crucial to consider that certain tasks in these benchmarks, which demand intensive memory or computational power, might inherently favor the distilled models due to their optimized architecture. The extent to which our results can be generalized may depend on the nature of the tasks and the specific requirements of different applications. Finally, we acknowledge certain limitations in our study, such as the specific scope of benchmarks and the range of models examined. Future research should expand to include a broader spectrum of NLP tasks, encompassing real-world applications, and should also consider the long-term environmental impacts of these models. Such studies would contribute significantly to the field of sustainable AI development, providing deeper insights into the ecological footprint of NLP technologies.

5 CONCLUSION

Our comparative analysis brought to light the pronounced differences between distilled and non-distilled NLP models across various metrics: energy consumption, inference time, CPU utilization, and memory usage. Distilled models, notably Distilled-BERT and Distilled-GPT-2, generally showcased superior energy efficiency and swifter inference. Particularly, Distilled-BERT emerged as notably more energy-efficient compared to its original counterpart, BERT. Conversely, Distilled-GPT-2 demonstrated marked improvements over GPT-2 in both inference time and memory usage. The results from our hypothesis testing further corroborated these observations, emphasizing the paramount influence of the choice of NLP model on both performance and energy consumption metrics.

With the insights garnered from this study, there lies a plethora of possibilities for further investigation. An immediate and apparent extension would be to compare a wider array of distilled models, expanding the horizons to incorporate other cutting-edge architectures and their distilled versions. A pivotal extension would be to assess the performance of models across diverse benchmarks, ascertaining if the observed benefits of distillation remain consistent across various challenges and tasks. Teams could also explore the performance trade-offs inherent in model distillation techniques in different deployment environments, especially on edge devices where computational constraints are accentuated. Furthermore, with multi-modal data becoming increasingly ubiquitous, a study into whether the advantages of distillation transcend data types would be invaluable. Lastly, in our era that is progressively emphasizing environmental sustainability, which is a deep dive into the ecological footprint of these models might reveal if distilled models truly serve as a more eco-friendly alternative in large-scale deployments.

ACKNOWLEDGEMENTS

This project is partially supported by the LETS- GO Project, promoted by the Netherlands Enterprise Agency (Rijksdienst voor Ondernemend Nederland); the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 871342 “uDEVOPS”; the European COST programme under action CAI19135 (CERCIRAS).

REFERENCES