Reducing Carbon Footprint of AI Models Without Compromising Performance

Austin Deng
Vandegrift High School
Austin, United States
email: austin12709@gmail.com

Xingzhi Huang
St. Anne's-Belfield School
Charlottesville, United States
email: zzandkun@gmail.com

Michael Lu
James Bowie High School
Austin, United States
email: michaelluaustin@gmail.com

***Abstract*—The widespread adoption of Artificial Intelligence (AI) models, such as ChatGPT, has resulted in a significant increase in energy consumption and carbon emissions associated with their training and inference. However, research on sustainable AI is still nascent. This paper aims to explore efficient approaches that can reduce the carbon footprint of AI models without compromising performance. Through an extensive analysis of four distinct categories of AI models (text to text summary, image classification, text to image, and image to text) across various sizes, our findings challenge the prevailing notion that larger AI models consistently outperform smaller ones. In specific AI tasks, we observe that small models can achieve comparable performance while significantly reducing carbon emissions. Moreover, we propose a carbon-aware solution that strategically directs computationally intensive AI tasks to regions with low carbon intensity, which can effectively reduce the environmental impact without compromising model quality. Our experimental results demonstrate a significant carbon savings while maintaining the desired performance levels.**

***Index Terms*—AI; Energy Efficiency; Carbon Emission.**

I. Introduction

The emergence of ChatGPT and GPT-4 has led to a remarkable surge in the popularity of AI. Within just two months, ChatGPT alone has attracted over 100 million users, showcasing the immense potential of AI models to revolutionize our daily lives and work. However, this surge in popularity has also resulted in a significant increase in energy consumption and carbon emissions attributed to AI. Unfortunately, the environmental consequences of AI models have not received sufficient attention, and efforts to mitigate their carbon footprint are still in the nascent stages of research.

One way to reduce the carbon emissions of AI is to use smaller and less energy-intensive models when possible. However, since it is commonly assumed that larger AI models consistently outperform smaller ones, smaller models are less preferred in practice.

In this study, we conduct a comprehensive analysis on 11 AI models spanning four different domains: text-to-text summary, image classification, text-to-image generation, and image-to-text generation. The analyzed text to text summary models include "t5-one-line-summary" and "t5-base-finetuned-summarize-news" [1]. For image classification models, we examine "Google-vit-base-patch16-224" [2], "Google-vit-base-patch16-384" [2], "Microsoft-cvt-13" [3], and "Microsoft-resnet-50" [4]. Regarding text-to-image generation models, we investigate "stable-diffusion-v1-4", "stable-diffsion-v1-5, and "stabilityai/stable-diffusion-2-1" [5]. We also analyze two image to text generation models ("trocr-base-printed" and "trocr-large-printed" [6]). Our experiments have demonstrated that resource efficient models can achieve comparable or superior performance in tasks such as image classification and image-to-text generation. Meanwhile, we observe that less power-hungry models can lead to a remarkable reduction in energy consumption, potentially up to 69%.

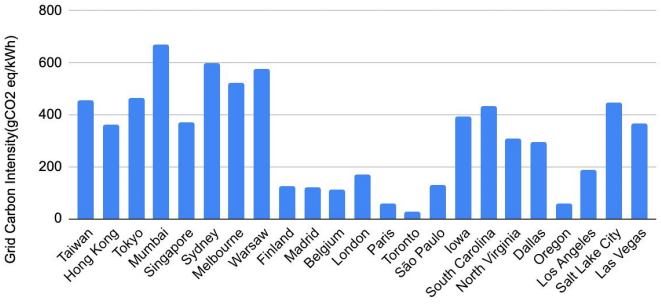

Fig. 1. Graph of carbon intensities for different locations

However, the same principle does not apply to text-to-text summarization and text-to-image generation models. In these cases, larger models (e.g., stable-diffusion-2-1) tend to generate higher quality images than smaller models such as stable-diffusion-v1-4. To balance the need for maintaining model quality while minimizing carbon emissions, we propose a carbon-aware solution. This solution strategically directs computationally intensive AI tasks to regions with lower carbon intensity in electricity production, thereby mitigating environmental impact without compromising the quality of the models. Figure 1 [7] illustrates the substantial variations in carbon intensity among different regions. For instance, the carbon intensity of Mumbai is approximately 23 times higher (670 gCO2eq) than that in Toronto (29 gCO2eq). This implies that deploying identical AI models in Toronto instead of Mumbai could potentially reduce carbon emissions by 95.67%.

This paper makes the following contributions:

- We quantitatively evaluate the energy consumption and carbon emission of 11 AI models.
- We reveal that using smaller models in image classification and image-to-text generation tasks can significantly reduce carbon footprint without compromising model quality.
- We propose a carbon-aware approach to mitigate the carbon emissions associated with AI tasks requiring large models, while ensuring no compromise on model quality.

 ISBN: 978-1-68558-097-1

The subsequent sections of this paper are structured as follows. Section II discusses related work, and Section III presents the detailed information about the AI models we evaluate. The methodologies for model quality evaluation, carbon emission measurement and carbon-aware model deployment are presented in Section IV. Section V presents experimental results and Section VI concludes this study.

II. Related Work

In recent years, the environmental impact of AI has garnered increasing attention, despite its status as a relatively young field. Numerous studies and reports have been published to better understand the significant environmental impact of AI training and inference.

Amodei et al. revealed that the computing demand for training AI models have increased 300,000 times in recent years [8]. The environmental impact of large AI models was highlighted by MIT Technology Review [10], stating that training a single AI model can emit 626,000 pounds of carbon dioxide, equivalent to the lifetime emissions of five average American cars [13]. However, few works have proposed ways to reduce the amount of carbon emissions generated by AI models. In [12], Schwartz et al. pointed out that traditional AI research (a.k.a. Red AI) focused on improving accuracy through the use of massive computational power while disregarding the cost and environmental impact. Red AI is not sustainable because the relationship between model performance and model complexity is understood to be logarithmic, meaning an exponentially larger model is required to gain a linear increase in performance [9]. For example, Mahajan et al. [11] reported that object detection accuracy increases linearly as the number of training examples increases exponentially. The opposite approach is Green AI, which emphasizes the importance of developing AI research that considers the computational cost and resource utilization [12]. According to Wu C. et al. [14], a deliberate and responsible approach is necessary when developing AI technologies, taking into account the environmental impact of innovations.

Although previous studies offered different approaches to enhance AI efficiency and decrease its carbon footprint through technological advancements, they generally overlooked the quantification of carbon emissions associated with distinct AI models. In contrast, our research takes a novel standpoint by quantitatively measuring the carbon emissions of 11 AI models. The results of our study demonstrate that using smaller AI models can be an effective and easily implementable strategy to reduce carbon emissions in AI systems. In situations where larger models are necessary for optimal performance, we propose an innovative carbon-aware solution, which can reduce carbon emissions by deploying AI models in regions with low carbon intensity.

III. AI Models

In this section, we describe the specific AI models we used for each of the four categories discussed earlier.

A. Text to Text Summary Models

Text-To-Text Transfer Transformer (a.k.a. T5) model was trained on up to 770 million parameters. The model takes a text as its parameter and can achieve four tasks: Translation, Corpus of Linguistic Acceptability, Semantic Textual Similarity Benchmark, and Summary. We evaluate two T5 models and focus on their summary functionality. The "t5-one-line-summary" [1] model was trained additionally on 370,000 research papers and can generate one line summary based on the abstract of the papers. It is conservatively estimated that the model has been downloaded 466,000 times a month from the Huggingface downloading data. The "t5-base-finetuned-summarize-news" [1] model was trained on news articles and the summarized versions of corresponding news articles. The model takes a small piece of news articles and can generate a brief summary of the article. The model is conservatively estimated to have 267,000 downloads a month.

B. Text to Image Generation Models

Stable Diffusion is a text-to-image generation AI model developed by Stability AI. It employs a technique where Gaussian noise is added to an image then removed in a manner that generates images corresponding to a given text or image prompt. We evaluate three different versions of Stable Diffusion models available on Huggingface: "stable-diffusion-v1-4", "stable-diffsion-v1-5, and "stabilityai/stable-diffusion-2-1" [5].

C. Image Classification Models

Image classification models aim to accurately classify the contents of different images. We evaluate four popular image classification models published on Huggingface. Both the "google-vit-base-patch16-224" model and the "google-vit-base-patch16-384" [2] model are derived from Vision Transformer (ViT). The "google-vit-base-patch16-224" model was trained using 224x224 resolution images while the "google-vit-base-patch16-384" model was trained on 384x384 resolution images. The "microsoft-cvt-13" [3] is an image classification model that adds convolutional neural networks to the ViT architecture. This model aims to help reduce the effects of distortions on the accuracy of ViT-based models. The "microsoft-resnet-50" model is a deep convolutional neural network that does not use Transformers to classify images. Instead, it explores the concept of using residual learning to train deeper models [4].

D. Image To Text Models

The image-to-text models use both image and text Transformers to recognize the words inside an image and print them out. Li et al. [6] pioneered the development of the "Transformer-based Optical Character Recognition" (TrOCR) model. To optimize its performance, these models have been fine-tuned using the Scanned Receipts OCR and Information Extraction (SROIE) dataset. We evaluate two TrOCR models presented by Microsoft in our study. The "trocr-base-printed" model is an encoder-decoder model which uses an image and

 ISBN: 978-1-68558-097-1

text Transformer to scan an image with text in it and prints a string of the word/phrase in the image. This is the base sized model for the Microsoft trocr-image-to-text models [6]. The "trocr-large-printed" model is the large sized model for the microsoft trocr-image-to-text models [6].

IV. Methodology

In this section, we describe our methodology for evaluating the quality of different AI models, measuring the energy usage and carbon emission of each AI model, and carbon aware deployment in the cloud.

A. Model Quality Evaluation

While reducing the carbon footprint of AI models is essential, it should not come at the expense of significantly compromising the quality of these models. Therefore, we carefully evaluate the quality of each AI model and only opt for smaller models when they can match or surpass the capabilities of larger models.

Specifically, the accuracy of each image classification model is evaluated by obtaining a random subset of 100 images from the CIFAR-10 dataset, which consists a set of 32x32 images that are labeled with one of the 10 categories: "airplane", "automobile", "bird", "cat", "deer", "dog", "frog", "horse", "ship", or "truck". Please refer to Figure 2 and Table 1 for the output of different models on a sample image of an automobile from CIFAR-10.

Fig. 2. Sample image of an "automobile" from the CIFAR-10 dataset

TABLE I
SAMPLE-IMAGE CLASSIFICATION MODEL OUTPUTS

Model	*Output*
Google-vit-base-patch16-384	moving van
Google-vit-base-patch16-224	sports car
Microsoft-cvt-13	beach wagon
Microsoft-resnet-50	cassette player

The accuracy of the small and large image to text models are evaluated by whether or not the model outputs the correct word in the image. We use 20 images downloaded from the internet with words in the image and finding the percentage of correct outputs of the 20 test images.

Assessing the quality of Text to Text Summary and Image Generation models can be challenging due to subjectivity, leading to varying ratings from different individuals. When comparing the outputs of three models prompted with "a photo of a beautiful desert landscape at night" (see Figure 3), it is evident that all three Stable Diffusion models successfully generated a desert landscape. However, the "stable-diffusion-v1-5" model (small) failed to produce a nighttime image and displayed several odd streaks of yellow throughout the image. Additionally, the resolution of the image generated by the "stable-diffusion-2-1" model (large) was significantly higher (768x768) than the other two images (512x512). In general, larger models for Text to Text Summary and Image Generation tend to produce higher quality outputs compared to smaller models.

Fig. 3. Images generated by CompVis/stable-diffusion-v1-4, runwayml/stable-diffusion-v1-5, and stabilityai/stable-diffusion-2-1 from left to right

B. Energy Usage and Carbon Emission Measurement

As AI models continue to gain widespread use across various sectors and industries, their environmental impact has grown significantly. Therefore, it is important to quantitatively measure both the energy used and the carbon emissions produced by training and deploying AI models. We leverage CodeCarbon [15] as a valuable tool for monitoring carbon emissions associated with various AI models. CodeCarbon provides a user-friendly API that facilitates the tracking of energy consumption and carbon emissions of different AI models. CodeCarbon enables us to monitor the power usage of underlying hardware components, such as GPUs and CPUs, at regular time intervals. In our study, we express carbon emissions in kilograms of CO2-equivalent per kilowatt-hour, and the power consumption is measured using the default sampling rate of 15 seconds.

C. Carbon Aware AI Deployment

When large models provide superior performance than smaller models, we propose to reduce their carbon footprint by deploying large AI models in regions with low carbon intensity. The carbon intensity of the consumed electricity is determined by taking into account the emissions from various energy sources used for electricity generation, encompassing both fossil fuels and renewables. The carbon intensity considers fossil fuels such as coal, petroleum, and natural gas, each linked to specific carbon intensities, signifying the amount of carbon dioxide released per kilowatt-hour of electricity produced. On the other hand, renewable or low-carbon fuels like solar power, hydroelectricity, biomass, and geothermal are also factored in.

Table II presents the Grid Carbon Intensity data provided by Google in grams of CO2 equivalent per kilowatt-hour (gCO2eq/kWh) for various cloud regions/locations [7]. The carbon intensity values indicate the amount of carbon dioxide equivalent emissions produced per unit of electricity consumed in each region. Lower carbon intensity values suggest that the electricity generation in those regions is more environmentally

friendly and emits fewer greenhouse gases. Looking at the data, we can observe that Toronto, Paris, Finland, Madrid, Oregon, London, and Belgium have relatively low carbon intensities. These regions appear to have a strong focus on renewable energy sources or nuclear power, resulting in significantly reduced carbon emissions. On the other hand, Mumbai, Sydney, Melbourne, Salt Lake City, South Carolina, and Warsaw have relatively high carbon intensities. These regions might be relying heavily on fossil fuels for electricity generation, leading to higher emissions.

TABLE II
GRID CARBON INTENSITY FOR DIFFERENT REGIONS

Cloud Region/Location	*Grid Carbon Intensity(gCO2eq/kWh)*
Taiwan	456
Hong Kong	360
Tokyo	464
Mumbai	670
Singapore	372
Sydney	598
Melbourne	521
Warsaw	576
Finland	127
Madrid	121
Belgium	110
London	172
Paris	59
Toronto	29
São Paulo	129
Iowa	394
South Carolina	434
North Virginia	309
Dallas	296
Oregon	60
Los Angeles	190
Salt Lake City	448
Las Vegas	365

In our experiments, we measure the energy consumption in kilowatt-hours (kWh) and the run time in seconds for each model. Then, using the carbon intensity data, we estimate the carbon emissions in various regions. The variation in carbon emissions across regions is multiplied by the actual usage of the AI models, allowing us to assess the potential CO2 savings.

V. EXPERIMENTAL RESULTS

In this section, we will present experimental results and the cloud platform that we use to deploy these AI models. Specifically, Subsection A discusses the cloud deployment details. Subsections B and C present the carbon reduction results of replacing larger models with small models. Subsections D and E illustrate the penitential carbon savings when deploying AI models on low carbon regions.

A. Cloud Deployment

In our experiments, all AI models are assessed using the A10 GPU within the Lambda Cloud [16], which provides us with instant access to cloud GPUs at highly competitive prices. The Lambda Cloud follows a pay-by-the-second billing model, ensuring that users are charged only for the actual time their instances are utilized. Furthermore, Lambda cloud pre-installs popular machine learning frameworks like TensorFlow, PyTorch, CUDA, and cuDNN, enabling us to promptly deploy models without any installation hassles. Additionally, the Lambda cloud supports deployment in multiple regions, each with varying carbon intensity levels. This feature allowed us to examine the efficacy of our proposed carbon-aware AI deployment approach.

B. Image Classification Models

We evaluate four image classification models, including “Google-vit-base-patch16-224” [2], “Google-vit-base-patch16-384” [2], “Microsoft-cvt-13” [3], and “Microsoft-resnet-50” [4]. The accuracy and energy consumption of all four models are presented in Table III, from which we can observe that two Google-vit models perform significantly better than the other two image classification models. However, they also consume significantly more energy. Surprisingly, even though the ”Google-vit-base-patch16-224” model uses 69% less electricity than the larger ”Google-vit-base-patch16-384” model, it achieves a higher accuracy level. This finding suggests that utilizing the ”Google-vit-base-patch16-224” model not only leads to better model quality but also generates less than one-third of the carbon emissions produced when using the ”Google-vit-base-patch16-384” model.

TABLE III
IMAGE CLASSIFICATION MODEL ACCURACY AND ENERGY USAGE

Model	*Accuracy*	*Energy Consumption (kWh)*
Google-vit-base-patch16-384	61%	0.1229
Google-vit-base-patch16-224	68%	0.0381
Microsoft-cvt-13	49%	0.0122
Microsoft-resnet-50	39%	0.0062

C. Image to Text Models

We analyze two image to text generation models, namely “trocr-base-printed” and “trocr-large-printed”. Both models exhibit nearly identical accuracy levels, with 96.59% for the large model and 96.37% for the base model [6]. However, our experiments reveal a significant disparity in energy consumption and processing time between the two models. Specifically, the base model outperform the large model in terms of energy efficiency and processing speed. For instance, when converting the image shown in Figure 4 to text, the large model takes 17.6 seconds and consumes 0.000585 kWh of energy, whereas the same task is accomplished by the base model in only 14.4 seconds, consuming just 0.000403 kWh of energy. Similarly, Figure 5’s image conversion is 18% faster and result in 22% energy savings when using the base model with no compromise on output quality.

These findings challenge the prevailing notion that larger models are inherently superior to smaller ones. In cases where both models achieve comparable accuracy, opting for the smaller model proves to be more efficient. Larger models not only demand more energy for running and training but

 ISBN: 978-1-68558-097-1

also take longer to process data. As such, the superiority of larger models is not universal, and in certain scenarios, smaller models can perform equally well, offering the additional advantages of reduced energy consumption and processing time.

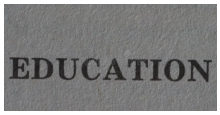

Fig. 4. Image with the text "Education" used to test the Text-to-Text Models

Fig. 5. Image with the text "Advantage" used to test the Text-to-Text Models

D. Carbon-Aware Deployment of Stable Diffusion Models

We evaluate three popular Stable Diffusion models including "stable-diffusion-v1-4", "stable-diffsion-v1-5, and "stabilityai/stable-diffusion-2-1" [5]. As previously discussed in Section III and illustrated in Figure 6, images generated by the large "stabilityai/stable-diffusion-2-1" model have much higher quality than the other two smaller models. Nevertheless, the "stabilityai/stable-diffusion-2-1" model consumes 4 times more energy than the other two smaller models, as illustrated in Table IV.

Fig. 6. An image of flying pigs generated by different Stable Diffusion models

TABLE IV
STABLE DIFFUSION MODEL ENERGY CONSUMPTION

Model	*Energy Consumption (kWh)*
stabilityai/stable-diffusion-2-1	0.01438
runwayml/stable-diffusion-v1-5	0.00321
CompVis/stable-diffusion-v1-4	0.00323

We leverage a carbon-aware solution to balance the need for maintaining model quality while minimizing carbon emissions. This solution uses the high quality "stabilityai/stable-diffusion-2-1" model but strategically deploys it to regions with lower carbon intensity, thereby mitigating environmental impact without compromising the quality of models.

This approach works because we now can easily deploy AI models at different regions using cloud computing. For example, deploying the "stabilityai/stable-diffusion-2-1" model at the "us-west4" region (based in Salt Lake City) results in a carbon intensity of approximately 448. In contrast, the "us-west1" region (based in Oregon) has a much lower carbon intensity of 60. The notable difference in carbon intensity is attributed to Oregon's utilization of renewable electricity sources, such as Hydro Power and Wind Power [17], while "us-west3", where Salt Lake City is located, relies significantly on fossil fuels like natural gas and coal to generate electricity [18].

To demonstrate the impact of carbon aware deployment in the cloud on carbon emissions, we conduct experiments with Stable Diffusion and estimated the carbon emissions for both "us-west3" and "us-west1." Although the energy consumption of individual requests might not seem substantial, considering Stable Diffusion models serve 10 million daily users worldwide, the accumulated energy and carbon savings become noteworthy. Assuming each user generates one image, we are looking at 10 million images being produced daily. Our experimental results indicate that this process would consume approximately 14,380 kWh of electricity per day. If the "stabilityai/stable-diffusion-2-1" model were to generate all its images using "us-west3," it would produce 6.4 million gCO2eq of greenhouse gases. On the other hand, if deployed on "us-west1," it would emit only about 860 thousand gCO2eq. By making environmentally conscious decisions about where the model is deployed, we manage to reduce carbon emissions by an impressive 86.6%.

E. Carbon-Aware Deployment for Summary Models

In this experiment, we evaluate two Text-To-Text Transfer ransformer summary models: "t5-one-line-summary" and "t5-base-finetuned-summarize-news" [1]. It is evident that the large "t5-base-finetuned-summarize-news" model generally provides higher quality summary than the small "t5-one-line-summary" model. Similarly, we can reduce its carbon footprint by deploying the large model in regions with low carbon intensity.

It is worth noting that the carbon-aware deployment approach can also benefit small AI models. Table V presents the carbon emissions of both small and large models when deployed in different regions. By choosing Oregon over Dallas as the deployment location for the "t5-one-line-summary" model, a single request can save 0.084 grams of CO_2, amounting to an 80% reduction. With an estimated usage of 466,000 times a month, the carbon aware deployment approach could save at least 39,144 grams of CO_2. Similarly, the "t5-base-finetuned-summarize-news" model could save a total of 106,533 grams of CO_2.

Since carbon intensity varies throughout the day due to electricity demand, computational tasks could switch between different cloud locations to optimize carbon savings.

 ISBN: 978-1-68558-097-1

TABLE V
ESTIMATED CO2 EMISSION IN DIFFERENT CARBON INTENSITY REGION

Region	***"one-line-summary"***	***"finetuned-summarize-news"***
Taiwan	0.162	0.771
Hong Kong	0.128	0.608
Tokyo	0.165	0.784
Mumbai	0.238	1.13
Singapore	0.132	0.629
Sydney	0.213	1.01
Melbourne	0.185	0.881
Warsaw	0.205	0.973
Finland	0.045	0.215
Madrid	0.043	0.204
Belgium	0.039	0.186
London	0.061	0.291
Paris	0.021	0.010
Toronto	0.010	0.049
São Paulo	0.046	0.218
Iowa	0.140	0.666
South Carolina	0.154	0.733
North Virginia	0.110	0.522
Dallas	0.105	0.500
Oregon	0.021	0.101
Los Angeles	0.068	0.321
Salt Lake City	0.159	0.757
Las Vegas	0.130	0.617

VI. CONCLUSIONS AND FUTURE WORK

The growing adoption of AI models has led to a notable rise in energy consumption and carbon emissions associated with their training and inference processes. Despite this concern, research on sustainable AI is still in its early stages. This study aims to investigate efficient approaches that can diminish the carbon footprint of AI models without sacrificing their performance.

We propose two methods to mitigate CO2 emissions while utilizing AI models. Firstly, by employing more energy-efficient models where feasible, we showcase instances where smaller AI models, consuming less energy, could deliver comparable or even superior performance to larger, more energy-intensive counterparts. Secondly, we advocate for a carbon-aware deployment of AI models. The geographical location where AI models are executed significantly influences their carbon intensity, as the carbon emissions generated during electricity production depend on the carbon intensity of the local energy grid. Adopting low carbon-intensity cloud services for running AI models can substantially reduce the carbon footprint of AI applications. This approach is applicable to both AI training and inference, thereby reducing the carbon emissions associated with electricity consumption. By implementing these carbon-reduction strategies, we can harness the power of AI for societal benefits while ensuring AI's carbon emissions remain sustainable. Our findings indicate that the utilization of smaller models can potentially reduce energy usage by up to 69% in specific scenarios, and the second method aligns AI's carbon footprint with the carbon intensity of the least carbon-intense cloud computing server.

Despite our paper's valuable contributions, we acknowledge certain limitations. The accuracy and energy usage measurements of AI models may not be entirely precise due to a relatively small sample size. To enhance our research, conducting in-depth experiments with a larger dataset could more accurately determine the models' accuracies. Moreover, future work should focus on quantifying the quality of both image generation models and summary models, enabling a more comprehensive comparison of their accuracies and energy usage. Another limitation pertains to the carbon intensity data, which is based on average values and may not account for real-time fluctuations. Carbon emissions during electricity production can vary due to several factors. To minimize carbon emissions more effectively, AI models might need to dynamically adapt to different cloud regions. Therefore, further investigation is necessary to further improve this approach.

Large computing organizations might see a more pronounced reduction in CO2 emission since they have more resources to optimize cloud usage and a wider range of AI models to choose from, but these solutions can be effective even on a small scale.

In conclusion, our study emphasizes the urgency of addressing the environmental impact of AI and presents viable strategies for reducing carbon emissions of employing AI models without compromising model quality. Embracing energy-efficient models and adopting carbon-aware deployment practices will contribute to a more sustainable and environmentally friendly integration of AI technology into our society.

REFERENCES

[1] C. Raffel, et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", pp. 11-41, 2021. https://www.jmlr.org/papers/volume21/20-074/20-074.pdf. Accessed Aug 10, 2023.

[2] A. Dosovitskiy et al., "An Image Worth 16x16 Words: Transformers for Image Recognition at Scale", pp. 3-7, 2021. https://arxiv.org/pdf/2010.11929.pdf. Accessed Aug 10, 2023.

[3] H. P. Wu, et al., "CvT: Introducing Convolutions to Vision Transformers", pp. 4-6, 2021. https://arxiv.org/pdf/2103.15808.pdf. Accessed Aug 10, 2023.

[4] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Deep residual learning for image recognition", in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684-10695, June 2022.

[6] M. H. Li et al., "TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models", https://arxiv.org/abs/2109.10282v5, 2021. Accessed Aug 10, 2023

[7] https://cloud.google.com/sustainability/region-carbon. Accessed Aug 10, 2023.

[8] D. Amodei and D. Hernandez, "AI and compute", https://openai.com/research/ai-and-compute, 2018. Accessed Aug 10, 2023.

[9] J. Huang et al., "Speed/accuracy trade-offs for modern convolutional object detectors", published in Proceedings of CVPR, 2017.

[10] K. Hao, "Training a single AI model can emit as much carbon as five cars in their lifetimes", published in MIT Technology Review, pp. 8-14, Jun. 2019.

[11] D. Mahajan et al., "Exploring the limits of weakly supervised pretraining", https://arxiv.org/abs/1805.00932, pp. 7-13, 2018. Accessed Aug 10, 2023.

[12] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI", pp. 5-9, Dec. 2020.

 ISBN: 978-1-68558-097-1

[13] E. Strubell, A. Ganesh, and A. McCallum, “Energy and Policy Considerations for Deep Learning in NLP”, https://arxiv.org/abs/1906.02243, pp. 1-5, 2019. Accessed Aug 10, 2023.
[14] C. J. Wuet al., “Sustainable AI: Environmental Implications, Challenges and Opportunities”, https://arxiv.org/abs/2111.00364, pp. 1-8, 2021. Accessed Aug 10, 2023.
[15] Codecarbon https://codecarbon.io/. Accessed Aug 10, 2023.
[16] Lambda Cloud https://lambdalabs.com/service/gpu-cloud. Accessed Aug 10, 2023.
[17] Oregon Department of Energy, “2020 Biennial Energy Report”, 2020.
[18] U.S. Energy Information Administration, “Utah State Profile and Energy Estimates”, 2023

 ISBN: 978-1-68558-097-1