EDITOR HSIAO-YING LIN
IEEE Member; hsiaoying.lin@gmail.com

Colors of Artificial Intelligence

Hsiao-Ying Lin, Huawei International

On the color spectrum, energy-hungry artificial intelligence (AI) is red, and energy-efficient AI is green. How will the colors of AI be represented in the future?

On 22 and 23 April, at the two-day Leaders' Summit on Climate 2021, global heads of state announced ambitious targets to manage the climate crisis. It was announced that, by 2030, the United States would cut its greenhouse gas emissions by 50%. The European Union targeted a 55% reduction in net greenhouse gases by 2030. China promised to strengthen its controls on greenhouse gas emissions beyond carbon dioxide and decelerate coal consumption. Conversely, according to a 2019 estimation, in the field of artificial intelligence (AI), the carbon footprint of training a state-of-the-art language model equates to that of five U.S. cars' entire lifetimes,[1] as depicted in Figure 1. Moreover, the carbon footprint of AI models is still increasing at a fast pace. Energy consumption is a key factor. Power requirements of modern AI are growing at a faster rate than Moore's law indicates. Because an AI model's accuracy tends not to increase at the same rate as the amount of energy that is invested, state-of-the-art AI models become energy hungry; they are termed *red AI models*. By contrast, the development of green AI is being promoted to emphasize an energy-efficient AI research agenda.[2]

Moore's law states that the number of transistors in a dense integrated circuit doubles every two years. On the other hand, the amount of compute resources used in developing the largest and most advanced AI models doubled every 3.4 months between 2012 and 2018, according to statistics from OpenAI.[3] Increases in the number of transistors cannot maintain pace with the requirements for training the latest AI models. As illustrated in Figure 2, from 2012 to 2018, a 300,000-fold increase in computation requirements was observed, whereas Moore's law indicated only an eightfold expansion. Modern AI models are energy hungry, and the emphasis on such designs seems

Digital Object Identifier 10.1109/MC.2021.3102359
Date of current version: 22 October 2021

to continue. Since breakthroughs in AI technology deliver invaluable contributions to society, making AI greener is an emerging and vital topic in science and technology.

ENERGY GUZZLERS

The main driver of energy-hungry approaches is the desire to continue advancing AI achievements. Several key factors encourage power-heavy AI implementations, including model sizes, training data set sizes, model hyperparameters (for example, model architecture), and algorithm hyperparameters (for example, the number of epochs and optimizers).

Model size is measured by the number of trainable parameters and determines the cost of processing one datum for training and inference. State-of-the-art AI models grow exponentially to achieve advances. Considering language models as an example (as shown in Table 1), Embeddings From Language Models (ELMo) has 94 million parameters, whereas Generative Pretrained Transformer 3 (GPT-3) has 175 billion. The size of training data sets is another factor; the use of a large training data set to optimize model accuracy is a common approach. Again, as illustrated in Table 1,

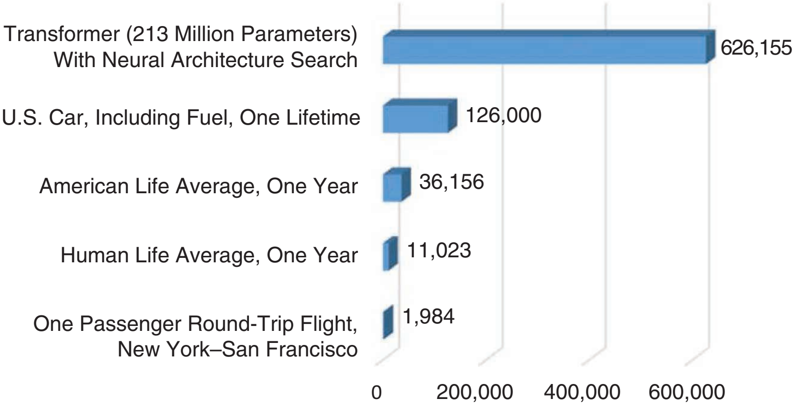

FIGURE 1. Carbon dioxide emissions, in pounds, based on data from Strubell et al.[1]

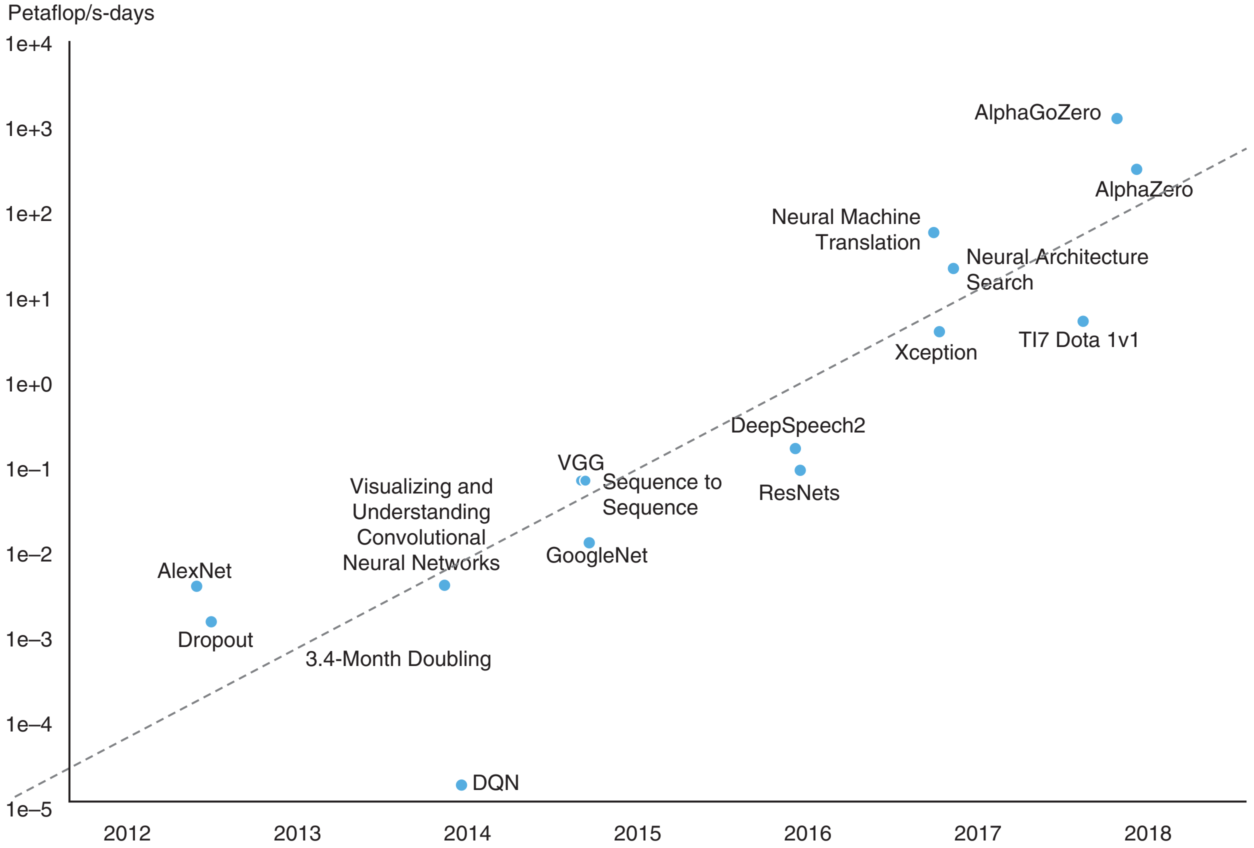

FIGURE 2. Computation demand is measured by the required number of addition and multiplication operations. The unit "petaflop/s-day" is the number of computing days, based on 10^{15} operations/s.[3] DQN: deep-Q network; VGG: Visual Geometry Group; ResNet: residual neural network.

the ELMo training data set has 5.5 billion words, whereas that of GPT-3 has hundreds of billions.

In addition to model and training data set sizes, model and algorithm hyperparameters are key factors in energy consumption. Designing a neural network architecture is critical to the final model performance, but the process can be energy intensive. A proper activation function helps in the production of a model with optimal accuracy; the search process repeats a considerable number of experiment runs on various types of AI models.[8]

Although only AI model training was considered in the preceding discussions, model inference may also considerably contribute to the energy consumption involved in AI applications. The required energy for one-time model inference is limited. However, the overall energy consumption can be significant because the number of model queries can be markedly high and increase continually. Chatbot services are an example. Assume that a user conducts a conversation with a chatbot when he or she encounters some impediment during online shopping. According to 2021 statistics,[9] approximately 2.14 billion users had completed an online purchase. Assuming that 1% of them initiate a chatbot conversation per day, at least 10 million model inference runs would be necessary, where the actual cost also depends on the conversation length. Moreover, the chatbot market is expected to expand at a compound annual growth rate of 24.9% from 2021 to 2028.[10]

TOWARD GREENER AI

Beyond environment protection, three incentives motivate green AI development. The first is diminishing returns in some mainstream models.[2] As researchers seek more accurate models by any means, a diminishing returns scenario is always taking place, whereby the same increases in model and training data set sizes generate fewer advances in accuracy. In other words, using a more extensive model no longer guarantees greater accuracy. The second impetus is a desire for inclusive AI development. When computation becomes an entry barrier, an ordinary researcher has difficulty operating AI research programs. The third motivation relates to business considerations. Edge AI products have restricted energy efficiency requirements for deployment, whereas cloud AI services can be developed on the back end of a cloud infrastructure on a limited budget.

The first step toward greener AI is having an appropriate quantitative tool. With that, the energy consumption of an AI model can be described and later compared with that of alternate models and with a greener version of the present one. The Allen Institute for AI proposed using the number of floating-point operations as a metric and encouraged researchers to measure and document it in their work. Another machine learning emissions calculator[11] was proposed based on hardware types, cloud providers, and geographical regions.[12] OpenAI tracks the energy efficiency of state-of-the-art AI models in terms of vision and translation.[13] On the basis of the efficiency benchmark—which quantifies the energy consumption associated with the computational gain yielded by algorithmic progress—image classification now has 44 times fewer computational requirements for training a neural network to the level of AlexNet than it did in 2012.

To advance the green AI movement, energy-efficient approaches in edge AI may inspire broader innovation. In edge AI research, common strategies for improving model inference energy efficiency include model compression and special hardware designs. Model compression aims to convert an original model to a functionally equivalent yet smaller one that consumes less energy during inference. Typical techniques include knowledge distillation, quantization, and pruning. Knowledge distillation transfers information from an original model to a target one through a training process applied to the latter. As the amount of data used in the training process grows, the target model functions increasingly similarly to the original. Quantization is employed to represent typically 64-bit floating-point model parameters as smaller ones, such as 32-bit, 16-bit, and even 1-bit integers. Pruning removes

> In the field of artificial intelligence, the carbon footprint of training a state-of-the-art language model equates to that of five U.S. cars' entire lifetimes.

TABLE 1. Characteristics of selected AI language models.

Model	Model size (parameters)	Training data set size (words)	Year
ELMo[4]	94 million	5.5 billion	2018
BERT (large)[5]	350 million	3 billion	2018
GPT-2[6]	1.5 billion	40 billion	2019
GPT-3[7]	175 billion	Hundreds of billions	2020

ELMo: Embeddings From Language Models; BERT: Bidirectional Encoder Representations From Transformers; GPT: Generative Pretrained Transformer.

less-critical neurons from a network to obtain a smaller model with acceptable accuracy loss. Tiny machine learning is an example of applying model compression to reduce the size sufficiently for a model to run on edge devices.[14]

Special hardware design is another common approach for enhancing energy efficiency in edge AI in particular and the AI field in general. Hardware acceleration for AI operations, such as those using tensor processing units and neural network processing units, has long been in progress. Special hardware design for edge AI also aims to deliver a range of AI functions from the cloud to edge platforms.[15] In early 2021, IBM announced the results of its work on the world's first energy-efficient AI chip with low-precision training and inference, constructed with 7-nm technology.[16] The chip integrates power management with improved model performance and power use. Codesign approaches integrating special AI models and hardware to achieve energy efficiency are also noteworthy. One example is the combination of spiking neural networks and neuromorphic accelerators.[17]

Reducing the amount of required training data is another principle of green AI. It remains an active research area, where comparatively more is learned with less data, as exemplified by zero-, one-, and few-shot learning for classification tasks. Zero-shot classification means a model is trained on some classes and then predicts inference data for a new class, which the model has never been exposed to. One-shot learning evaluates the possibility that two pieces of input data are in the same class, where one input datum serves as a class reference. Few-shot learning, or low-shot learning, classifies inputs among classes even when a training data set contains only a small amount of data in certain classes. Zero-, one-, and few-shot learning require considerably less information to learn new classes than other conventional learning techniques do.

DISCLAIMER

This column includes the views of the author. The opinions expressed here are her own.

During the development of AI models, various experiments are conducted, such as those for tuning hyperparameters and model parameters. A high-reuse scheme can assist in obtaining better energy efficiency from the entire process. For instance, using a pretrained model is a common method of faster AI model development. It is also a more energy-efficient technique. Devising reproducible and portable experiments that can be conducted on various machine learning frameworks is another means of sharing research experiences. Approaches for reusing and sharing useful intermediate data and learned experiences from the development of AI models should be further explored.

Another green approach is to use AI for energy-efficient purposes, such as Google employing AI to optimize power usage in data centers and reduce a cooling system's energy consumption by 40%.[18] AI can substantially lower energy consumption in buildings,[19] which are the source of approximately 40% of all U.S. power demand.[20]

Although modern AI has made many groundbreaking achievements in computer vision, natural language processing, and decision making, the human brain is still a more efficient source of intelligence. While red AI technology pursuing more cutting-edge technologies with considerable energy investment is expected to stimulate AI advances that help people overcome burdens and flourish, green AI development should be boosted for environment protection and inclusiveness. It is going to be a colorful world.

REFERENCES

1. E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," 2019, arXiv:1906.02243.
2. R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," *Commun. ACM*, vol. 63, no. 12, pp. 54–63, Dec. 2020. doi: 10.1145/3381831.
3. D. Amodei and D. Hernandez, "AI and compute," OpenAI. https://openai.com/blog/ai-and-compute/ (accessed May 16, 2021).
4. M. E. Peters et al., "Deep contextualized word representations," 2018, arXiv:1802.05365.
5. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, arXiv:1810.04805.
6. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI Blog, 2019. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (accessed May 16, 2021).
7. T. B. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst. 33 (NeurIPS 2020)*, Dec. 2020. [Online]. Available: https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html
8. P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," in *Proc. 6th Int. Conf. Learn. Representations (ICLR) Workshop Track*, 2018. [Online]. Available: https://openreview.net/forum?id=Hkuq2EkPf
9. D. Coppola, "E-commerce worldwide - Statistics & facts," Statista. https://www.statista.com/topics/871/online-shopping/ (accessed May 15, 2021).
10. "Chatbot market size worth $2,485.7 million by 2028," Grand View Research, Apr. 2021. https://www.grandviewresearch.com/press-release/global-chatbot-market (accessed May 21, 2021).

11. “Machine learning emissions calculator,” GitHub. https://mlco2.github.io/impact/#compute (accessed May 16, 2021).

12. A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres, “Quantifying the carbon emissions of machine learning,” 2019, arXiv:1910.09700.

13. D. Hernandez and T. Brown, “AI and efficiency,” OpenAI. https://openai.com/blog/ai-and-efficiency/ (accessed May 16, 2021).

14. “TinyML brings AI to smallest arm devices,” Arm Blueprint. https://www.arm.com/blogs/blueprint/tinyml (accessed May 16, 2021).

15. J. Lee, S. Kang, J. Lee, D. Shin, D. Han, and H.-J. Yoo, “The hardware and algorithm co-design for energy-efficient DNN processor on edge/mobile devices,” *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 67, no. 10, pp. 3458–3470, Oct. 2020. doi: 10.1109/TCSI.2020.3021397.

16. A. Agrawal et al., “9.1 A 7nm 4-core AI chip with 25.6TFLOPS hybrid FP8 training, 102.4TOPS INT4 inference and workload-aware throttling,” in *Proc. 2021 IEEE Int. Solid- State Circuits Conf. (ISSCC)*, pp. 144–146. doi: 10.1109/ISSCC42613.2021.9365791.

17. S. R. Kulkarni, D. V. Kadetotad, S. Yin, J.-S. Seo, and B. Rajendran, “Neuromorphic hardware accelerator for SNN inference based on STT-RAM crossbar arrays,” in *Proc. 2019 26th IEEE Int. Conf. Electron., Circuits Syst. (ICECS)*, pp. 438–441. doi: 10.1109/ICECS46596.2019.8964886.

18. W. Knight, “Google just gave control over data center cooling to an AI,” MIT Technology Review, 2018. https://www.technologyreview.com/2018/08/17/140987/google-just-gave-control-over-data-center-cooling-to-an-ai/ (accessed May 16, 2021).

19. B. Yan, F. Hao, and X. Meng, “When artificial intelligence meets building energy efficiency, a review focusing on zero energy building,” *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 2193–2220, 2021. doi: 10.1007/s10462-020-09902-w.

20. “The issue,” Alliance to Save Energy. https://www.ase.org/categories/buildings (accessed May 16, 2021).

HSIAO-YING LIN is a senior researcher at Huawei International, Singapore. She is a Member of IEEE. Contact her at hsiaoying.lin@gmail.com.

Digital Object Identifier 10.1109/MC.2021.3115381