Infosys®
Navigate your next

VIEWPOINT

TINY AI FOR A SUSTAINABLE DIGITAL FUTURE

Artificial intelligence (AI) generates a massive carbon footprint and it will grow to unsustainable levels as more IoT devices enter the ecosystem. Now, businesses are looking to tiny AI — in which small models are processed on edge devices — as a way to reduce the environmental impact.

Infosys® | Knowledge Institute

The public is well aware of the effects that deforestation, burning fossil fuels, and other industrial activities have on climate change. But far fewer realize the environmental impact of the technology that powers typing assistance and predictive search features on their smart devices. Artificial intelligence (AI) already has a larger carbon footprint than the airline industry, even though a commercial aircraft can burn about 4 liters of jet fuel every second.[1,2]

Companies are finding more ways that AI can advance their businesses; but at the same time, these firms are under increasing pressure to rein in their carbon footprint. Globally, 21% of the largest 2,000 public companies have already committed to net-zero emissions and more can be expected to follow suit.[3] Fortunately, recent developments have created a middle ground where companies can lower emissions without sacrificing one of their most promising technologies.

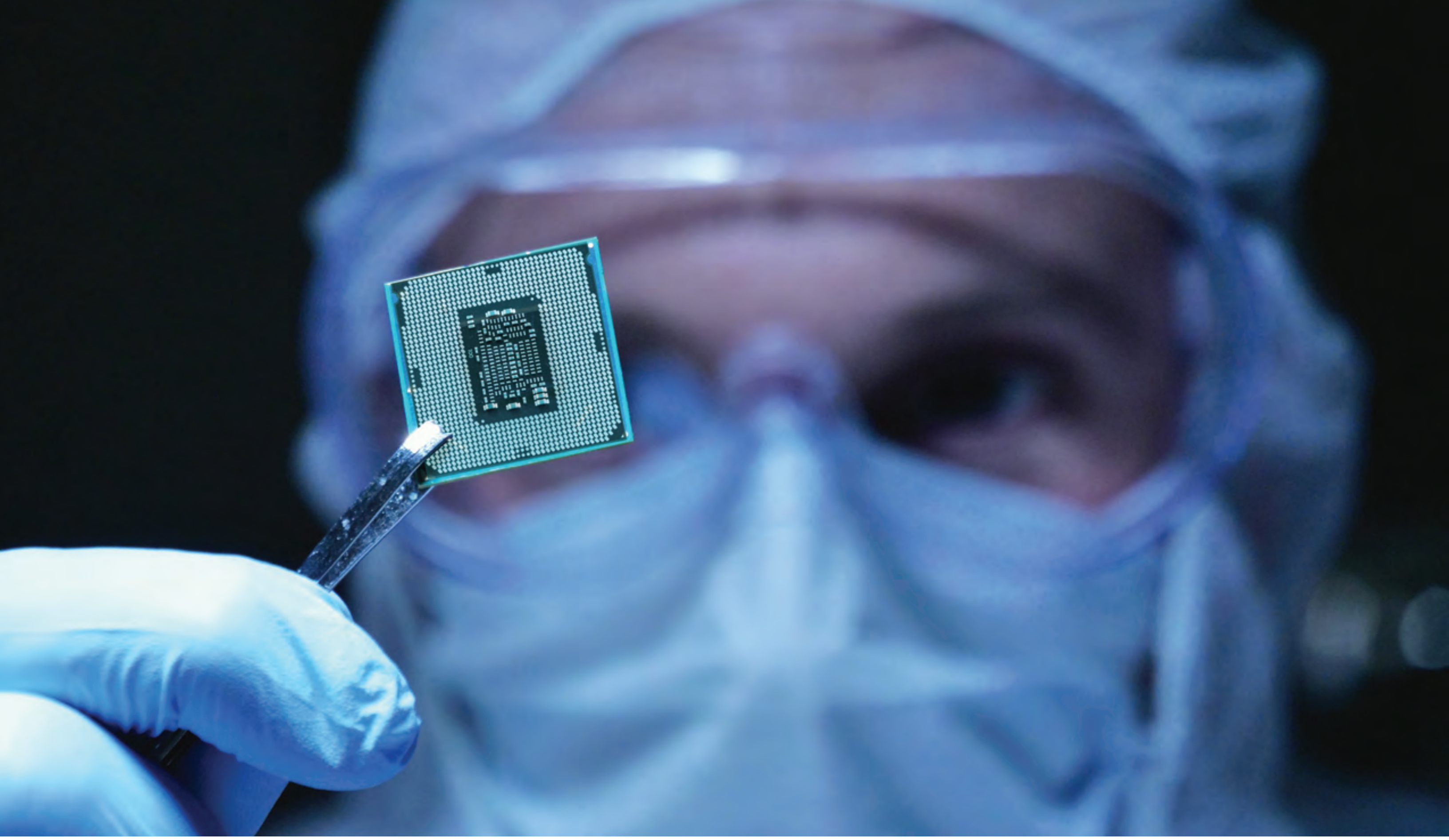
This new concept, tiny AI, helps reduce costs and energy requirements by shrinking the models and conducting AI inference and training on edge networks rather than on cloud servers. It was popularized after a 2017 Google white paper highlighted advances in federated learning, which enables edge devices to train their own models. Tiny AI can also make systems more private, secure, responsive, and contextually intelligent — even when there is no internet connectivity.[4]

> Tiny AI reduces costs and energy requirements by shrinking models and conducting inference on the edge

AI is power hungry

Although AI has broad uses, a majority of its development has focused on natural language processing (NLP) and computer vision. Those models are huge, expensive to train, and have a massive carbon footprint because of the energy required to operate the hardware. For instance, the GPT-3 model, which factors in 175 billion parameters and is the largest neural network, consumes 3 GWh of electricity and costs around $12 million to train.[5,6] This much electricity can power about 280 average American homes for a year.[7] Even the energy demands for a single training session of the smaller transformer NLP model can generate the equivalent of 284 metric tons of carbon dioxide — five times that of an average American car over its lifetime.[8] During the research and development stage, AI models are trained several times before they can be deployed, so this impact is often multiplied.

AI has now reached the point where most organizations are either exploring its applications or implementing it at scale across



multiple business units.[9] The sprawling use of AI now ranges from content recommendations on personal devices to helping robots clear farms of weeds while leaving crops untouched. Immense amounts of money and energy are being pumped into AI at a staggering rate. According to an OECD report, the global annual investment by venture capitalist firms in AI startups rose from $3 billion in 2012 to $75 billion in 2020.[10]

Industry analysts estimate that 25 billion to 30 billion internet of things (IoT) devices will be in use by 2025 (compared with about 12 billion in 2020).[11,12] These will generate even more data that will feed AI models. For instance, an autonomous vehicle can generate 4 terabytes of data in 90 minutes of driving, according to an estimate by Intel.[13]

Companies, technologies, and even governments expect AI to solve many of the world's critical problems, but it can't do so at the cost of escalating emissions.

Shrink the models for better energy efficiency

Tiny AI was developed so that smart devices would not always need to send data to the cloud for processing. By shrinking machine learning (ML) models, it is possible to retrain the models locally without compromising accuracy. At the edge, power and processing requirements are significantly reduced. This does not entirely eliminate the need to train models on the cloud, but it can reduce the load on servers by decentralizing it through smartphones and IoT devices.

> TinyBERT is much faster and more energy efficient than its "teacher," the BERT natural language model

This different approach, however, requires programmers to alter how they design AI models and select data for their use. The three main changes required to convert to tiny AI are:

1. **Edge data selection —** Diversity sampling, or valuing each dataset on its utility, helps weed out less useful data. This reduces energy requirements and can even improve model accuracy by lowering the noise level.
2. **Edge AI model —** Appropriate matchmaking between the model architecture and the edge hardware can reduce the size of the model. Studies show that smaller model architectures, such as MobileNet (20 MB), perform as well as models 25 times their size.[14]
3. **Model compression —** Reducing the number of parameters without significant loss of accuracy can be achieved through three main techniques.
   - **Knowledge distillation —** The "teacher" model trains the "student" model using only the key layers and weights.

- **Network pruning —** AI architects can remove parameters that contribute the least to model accuracy.
- **Quantization —** Model weights can be stored in quantized units of 8 bits or less. This makes the model four times smaller but only 2% less accurate.

These steps allow organizations to reduce the size and complexity of their power-hungry AI models so they can be deployed on the edge. There, learning and response can occur in real time and with low latency. The models are also personalized for each edge device environment, and critical personal data never leaves the device.

One prominent example of this new approach is TinyBERT. Researchers from Huawei and Huazhong University of Science and Technology developed this model from Google's original BERT natural language model. BERT is a pretrained model that suggests contextually relevant words based on the words that have already been typed into the user interface. However, the electricity required for this innovative model is estimated to produce greenhouse gas emissions equivalent to an average passenger vehicle being driven about 2,600 kilometers.[15,16] TinyBERT is 7.5 times smaller and consists of only 28% of the original BERT's parameters, making it much more energy efficient than its parent model. At the same time, it is 97% as accurate as its "teacher" is while also being 9.4 times faster at generating predictive sentences.[17]

Leverage external advances for tiny AI

While the concept of tiny AI offers cost and sustainability benefits, it may struggle to achieve mass adoption solely based on smaller AI models and edge processing capabilities. The ecosystem around it must evolve as well to facilitate practical applications whose benefits aren't offset by other environmental concerns. Federated learning, batteryless IoT sensors, and decentralized network coverage are three such external advances that will further expand adoption of this technology and enhance its sustainability.

Federated learning

Introduced by Google in 2017, federated learning is an ML approach that decentralizes the training of data. It requires less power consumption and offers lower latency and improved privacy. With federated learning, a connected device trains its own model using its own data and shares just a summary output of the model with the cloud. This reduces the need for continuous interaction between the device and cloud and enables fewer iterations. Only the most relevant updates — rather than all data — are exchanged with the cloud. Federated learning also ensures immediate implementation of local learning on edge devices, since the lag between cloud learning and device learning is eliminated. Gboard, the Google keyboard on Android devices, uses this concept to improve text recommendations by learning from what suggestions users click or ignore.[18]

Batteryless IoT

Batteryless IoT can eliminate some of the solid waste problems that could accompany an exponential increase in IoT devices. The replacement of batteries in billions of new sensors would be an environmental and logistical nightmare. With 25 billion to 30 billion IoT devices in circulation by 2025 and an average battery life of 3 years, the world would eventually discard about 23 million to 27 million batteries each day.[19,20] However, advancements in ultra-low power, always-on circuits that operate on energy derived from their immediate environment could solve this battery problem. Companies like Everactive, e-peas, and Nowi Energy are developing new energy harvesting solutions that could further enhance tiny AI's sustainability benefits. While still in the early stages of development, batteryless sensors are likely to propel IoT toward greater adoption.

> Proof-of-coverage based blockchain networks can help increase the adoption of tiny AI

Decentralization of the network coverage

Proof-of-coverage based blockchain networks can also help increase adoption of tiny AI. Unlike expensive data plans from traditional network providers, these blockchain networks crowdsource network coverage at fraction of the price. Individuals who participate set up long-range wide-area networks from their homes and earn cryptocurrency. These networks operate on an unlicensed low-power, long-range sub-GHz spectrum. This has the potential to connect billions of IoT devices in an energy-efficient manner instead of requiring a significant number of 5G antennas and fiber optic cables. Apple has already created one such crowdsourced network. Its proprietary network for the AirTags product range — coin-shaped tracking tags that can be attached to items — work using Bluetooth and low-power ultra-wideband technology. AirTags ping off iPhones or other Apple devices nearby, which send the location of the AirTag to Apple servers and allow the owner to track it using the Find My app.

Tiny AI and Californian wildfires

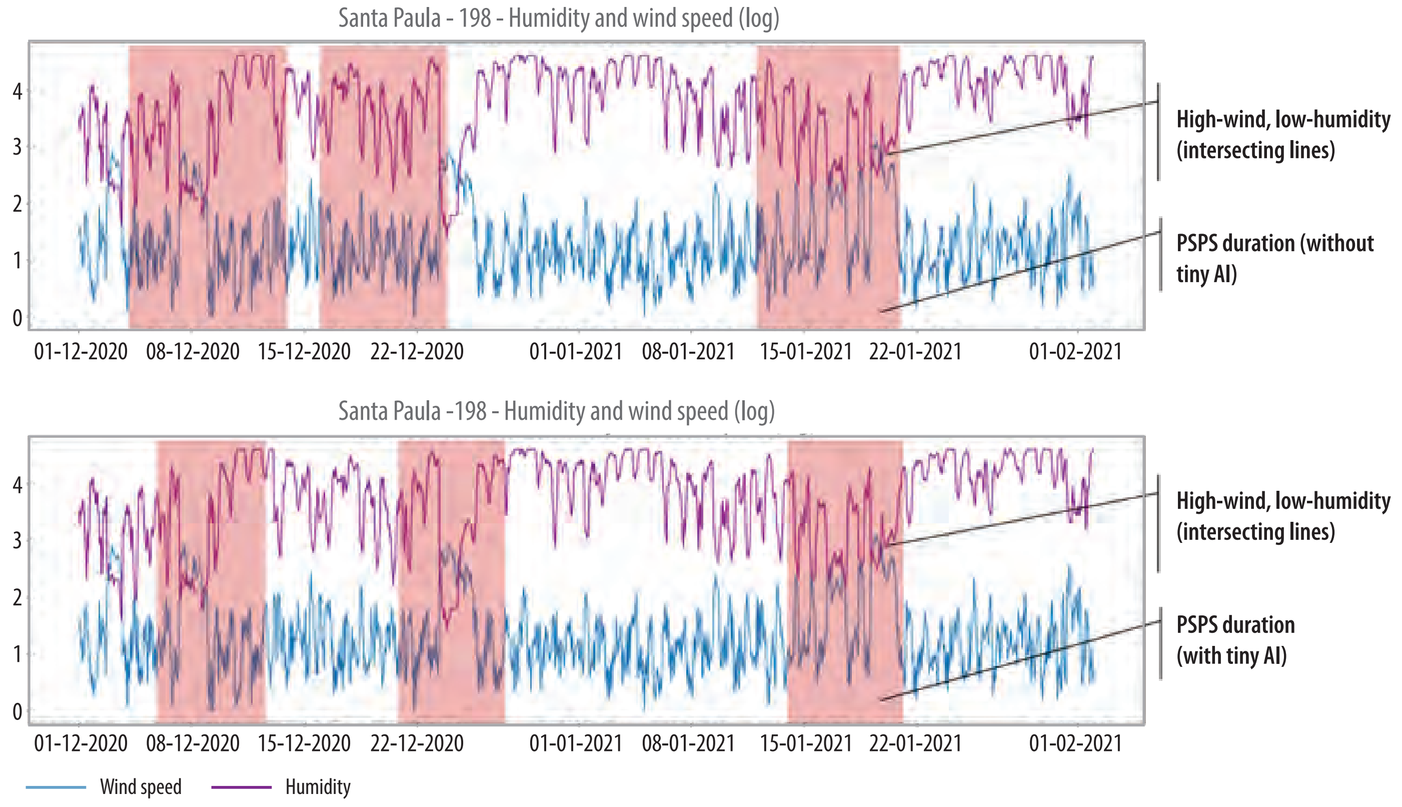
Tiny AI is a new technology — so new that its full environmental benefits have not yet been calculated. But organizations are finding more ways to use it, including cases where sensors monitor environmental conditions. In one real-world example, a California electricity provider uses tiny AI to prevent wildfires. There, utilities conduct public safety power shutoffs (PSPS) in certain areas when weather conditions are ripe for wildfires, which can be caused by electric lines falling on dry leaves or trees falling on power lines. However, determining when to cut off power and when to turn it back on is a difficult task, as it's always a trade-off between public convenience and safety. To solve this problem, the utility collects local weather data from 1-square-mile areas. The model uses sensors to collect data on temperature, wind, humidity, and other factors. Since data from each unit of area is collected and modeled separately, it helps better predict localized threat windows. At the same time, this local data is also fed into a central model that gets trained and updated on a wider regional basis, which helps predict local disruptions likely to be caused by broader weather variations. This enabled the electricity provider to make precise predictions for each localized region and therefore, avoid excess downtime in addition to allowing up to 48 hours of advance notice prior to a PSPS incident.

In California, this technique significantly improved the accuracy of these windows (Figure 1). The pink blocks here indicate shutoff periods, which are imposed when wind speed is high, and humidity is low (marked by intersecting lines). The top half shows PSPS windows predicted by the utility using previous technology, while the bottom half shows windows predicted by the tiny AI model for the same time period in Santa Paula, California.[21] The tiny AI model approach was not only able to shrink shutoff periods but also capture high-probability incidents more accurately. For instance, in the incident around Dec. 22, 2020, power would have been restored in the middle of a high-wind, low-humidity period using the earlier prediction method, while the tiny AI model accurately captured that incident safely between a shutoff.

Future-proof the AI ecosystem

When designing AI, it is imperative that environmental and social costs don't outweigh both the perceived and real benefits of the technology.

Figure 1: Tiny AI predicts shorter, more precise shutoff periods for Californian electricity provider


Santa Paula - 198 - Humidity and wind speed (log)

High-wind, low-humidity (intersecting lines)

PSPS duration (without tiny AI)

Santa Paula -198 - Humidity and wind speed (log)

High-wind, low-humidity (intersecting lines)

PSPS duration (with tiny AI)

Source: Infosys

A future-proof AI ecosystem must evolve in tandem with climate change goals set by world leaders. Today, employees, customers, regulators, and investors expect environmental, social, and governance factors to play a key role in all business activities.

> Visionary executives will embed tiny AI on the edge to save money and energy, and improve data privacy and security

In this climate, organizations are moving away from centralized and proprietary networks and toward ones that are decentralized and democratized. The move from centralized cloud servers toward edge computing is following closely. Visionary executives must now start embedding sustainable practices in their use of AI — not just to save money and energy but also to improve data privacy and security. This will enhance their credibility with a diverse pool of stakeholders. Applying tiny AI while leveraging federated learning, batteryless IoT devices, and the imminent decentralization of network coverage, organizations stand to defend themselves from economic, regulatory, and societal threats.

References

1. Deploying artificial intelligence at the edge: Key takeaways from SEMI CTO Forum, Dr. Pushkar P. Apte and Tom Salmon, Sept. 13, 2021, SEMI.
2. How much fuel does an international plane use for a trip?, May 19, 2021, HowStuffWorks.
3. Taking stock: A global assessment of net zero targets, Richard Black, Kate Cullen, Byron Fay, et al., March 2021, The Energy & Climate Intelligence Unit, Oxford Net Zero.
4. Tiny AI for the enterprise edge, Rajeshwari Ganesan, Rafee Tarafdar and, Harry Keir Hughes, Aug. 2021, Infosys Knowledge Institute.
5. Tiny machine learning: The next AI revolution, Matthew Stewart, Oct. 2, 2020, Towards Data Science.
6. OpenAI's massive GPT-3 model is impressive, but size isn't everything, Kyle Wiggers, June 1, 2020, VentureBeat.
7. How much electricity does an American home use?, Oct. 7, 2021, U.S. Energy Information Administration.
8. Training a single AI model can emit as much carbon as five cars in their lifetimes, Karen Hao, June 6, 2019, MIT Technology Review.
9. Digital Radar survey, 2021, Infosys Knowledge Institute.
10. Venture Capital Investments in Artificial Intelligence, Sept. 2021, OECD.
11. IoT connections forecast: The rise of enterprise, Dec. 16, 2019, GSMA.
12. State of the IoT 2020: 12 billion IoT connections, surpassing non-IoT for the first time, Knud Lasse Lueth, Nov. 19, 2020, IoT Analytics.
13. Intel editorial: For self-driving cars, there's big meaning behind one big number: 4 terabytes, Kathy Winter, April 14, 2017, Business Wire.
14. Deep learning has a size problem, Jameson Toole, Nov. 5, 2019, Heartbeat.
15. OpenAI's massive GPT-3 model is impressive, but size isn't everything.
16. Greenhouse Gas Equivalencies Calculator, March 2021, U.S. Environmental Protection Agency.
17. TinyBERT: Distilling BERT for natural language understanding, Xiaoqi Jiao, Yichun Yin, Lifeng Shang, et al., Oct. 16, 2020, Huawei Technologies,Huazhong University of Science and Technology.
18. Federated learning: Collaborative machine learning without centralized training data, Brendan McMahan and Daniel Ramage, April 6, 2017, Google.
19. Everactive's Evernet Protocol, Everactive.
20. Overcoming the battery obstacle to ubiquitous sensing — finally, Everactive.
21. Tomi Tyrrell, PhD contributed to building the model for wildfire prediction using Auto regressive RNNs from 200+ weather station data.

36.5
91.0
16-06

Authors

Rajeshwari Ganesan

Associate Vice President and Senior Director, Infosys

Jitesh Gera

Infosys Knowledge Institute

Harry Keir Hughes

Infosys Knowledge Institute

Contributor

Tomi Tyrrell

Specialist Programmer, Strategic Technology Group, Infosys

About Infosys Knowledge Institute

The Infosys Knowledge Institute helps industry leaders develop a deeper understanding of business and technology trends through compelling thought leadership. Our researchers and subject matter experts provide a fact base that aids decision making on critical business and technology issues.
To view our research, visit Infosys Knowledge Institute at infosys.com/IKI or email us at iki@infosys.com.

For more information, contact askus@infosys.com

Infosys®
Navigate your next