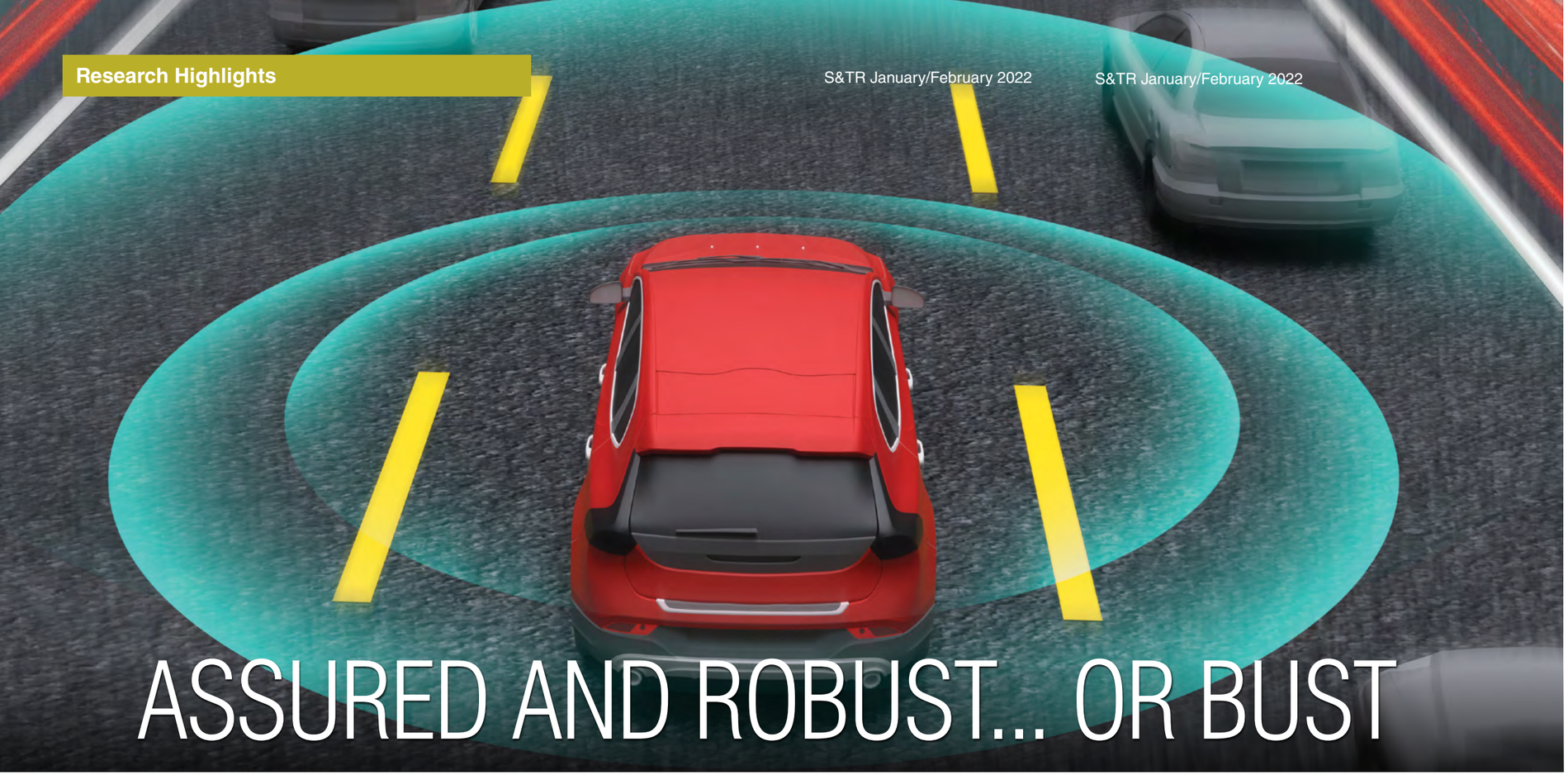
S&TR January/February 2022



ASSURED AND ROBUST... OR BUST

AN autonomous vehicle drives down a busy streeet when a strong breeze blows dust into the camera capturing the scene ahead and transmitting information to a machine-learning algorithm that detects potential safety hazards. In this hypothetical scenario, one misidentified object, such as an undetected person crossing the street, could mean the difference between life and death. Similarly, incorrect machine-learning predictions can incur significant costs in terms of experimental resources and lost opportunities in a range of scientific applications.

Many machine-learning systems employ complex mathematical concepts and algorithms to create artificial neural networks (ANNs) that mimic the biological neural network patterns of the brain, identifying patterns in data and then processing that information to make decisions and predictions. The algorithms are designed to identify critical differences among objects, but the rationale used by machines to make their predictions is opaque, as the systems teach themselves how to process the data within the networks. Even researchers designing and training the networks do not always understand how the models make decisions.

The consequences of a machine-learning error that presents irrelevant advertisements to a group of social media users may seem relatively minor. However, this opacity, combined with the fact that machine-learning systems are nascent and imperfect, makes trusting their accuracy difficult in mission-critical situations, such as recognizing life-or-death risks to military personnel or advancing materials science for Livermore's stockpile stewardship mission, inertial confinement fusion experiments, radiation detectors, and advanced sensors. (See *S&TR,* July/August 2017, pp. 16–19.) While opacity remains a challenge, Livermore's machine-learning experts aim to provide assurances on performance and enable trust in machine-learning technology through innovative validation and verification techniques. (See *S&TR*, March 2019, pp. 4–11.)

Robust at the Boundaries

The algorithms used in machine learning have achieved excellent performance on specific tasks such as image classification—distinguishing cars from trucks or cats from dogs—when training data is plentiful. Researchers have found that ANNs perform extremely well on clean, well-curated data that can be controlled; but when tested against real-world scenarios, where data is often messy, incomplete, or unpredictable, the ANNs fare much worse. Recent work has shown the algorithms remain extremely brittle, failing catastrophically in the presence of variations such as small rotations of an image.

"The lack of trust and transparency in current machine-learning algorithms prevents more widespread use in important applications such as robotic autonomous systems, in which artificial intelligence algorithms enable networks to gather information and make decisions in dynamic environments without human intervention," says Ryan Goldhahn, a computational engineer at Livermore. "To be truly useful, machine-learning algorithms must be robust, reliable, and able to run on small, low-power hardware that can perform dependably in the field."

Robust and assured machine learning implies that a machine-learning model makes a correct prediction under different perturbations (changes to the data) and noise (corruptions to the data). Robustness protects against both natural shifts in the real environment and adversarial shifts, such as a defaced traffic sign or data poisoning (intentionally altered data, imperceptible to humans, that is designed to fool the machine-learning model). Achieving robustness is a notable challenge.

Preparing for the Unexpected: Red AI

A team of Livermore researchers is developing "foolproof" or "assured" artificial intelligence (AI) to solve this critical concern. The team strives to identify the types of real-world perturbations (natural and adversarial) that might impact high-regret applications, in which consequences to human life are severe if the model makes a mistake. In a task once thought impossible due to the high number of inputs and complexity of nonlinear, deep neural networks, the team trains machine-learning models to provide guarantees—making them provably robust—by testing each input point in a computationally efficient manner. "Imagine being in a self-driving car that is prone to accidents when the weather changes," says Livermore computer scientist Bhavya Kailkhura. "With provable robustness guarantees, machine learning–AI can transfer control back to the human operator or make risk-averse decisions such as reducing speed when operating conditions such as heavy rain cross the boundary of a parameter."

To avoid catastrophic failures in mission-critical spaces, Livermore approaches improvements to robustness from two angles. Kailkhura calls the first angle the "Red AI" team, which deliberately fools machine-learning systems to anticipate where they will fail and then learn how to account for these failures in future, real-world scenarios. The Red AI team has explored training data poisoning, which degrades the robustness guarantees of state-of-the-art (SOTA) classifiers (decision-making tools used by AI systems) already used in machine-learning systems. To make the poisoning attacks used to train and test robust systems harder to detect, the Livermore team applies changes to the training data that are imperceptible to current AI systems, finding they can significantly reduce the robustness of SOTA classifiers and render existing defense approaches practically useless.

To better understand how the perturbations and noise work in real-world applications, Livermore has developed a system that deliberately corrupts image classifiers or detectors. In one approach, synthetic data is leveraged. "Tools from the video game industry allow us to render photo-realistic scenes and control the environment completely, altering background, objects present, the angle of the sun, and other features," says Goldhahn. The software, created to generate different images for training and testing, leverages 3D simulations to produce accurate, unique, and interesting data difficult to gather, curate, and label in the real world.

Foolproof (Assured) AI: Blue AI

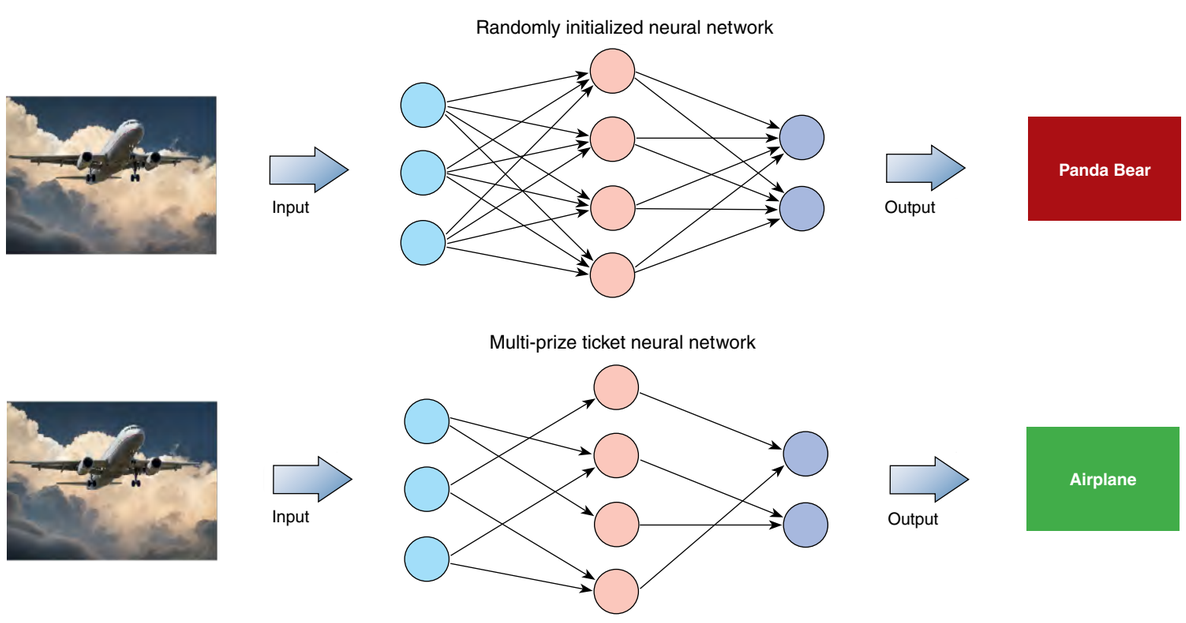
A "Blue AI" team complements the Red AI team's attacks by developing certified defenses that cannot be broken, regardless of the attacking algorithm, to provide robustness guarantees. Predictions from the certified model during testing are accompanied by a mathematical radius in which the predictions of the classifier are guaranteed to remain constant, thereby making them resilient to adversarial attacks on the system. If the input falls outside the mathematical radius under these circumstances, the predictions are still resilient.

Borrowing motivations from formal verification (computer-supported mathematical analysis ensuring software correctness), the Blue AI team developed certification schemes to provide rigorous guarantees on the correctness and robustness of pretrained machine-learning models under a range of natural and adversarial data shifts. Next, they developed training schemes to design machine-learning systems that are not only accurate but also provably correct and robust to multiple real-world perturbations such as shifting or rotating images, blurring, or altering lighting and contrast. "For the first time, the Livermore team has developed a mathematical tool that can provide decision makers with assurances about robustness to large, natural perturbations that can happen in the real world," says Kailkhura. "This tool essentially guarantees that the AI will do the right thing and make the right predictions under numerous specifications and constraints."

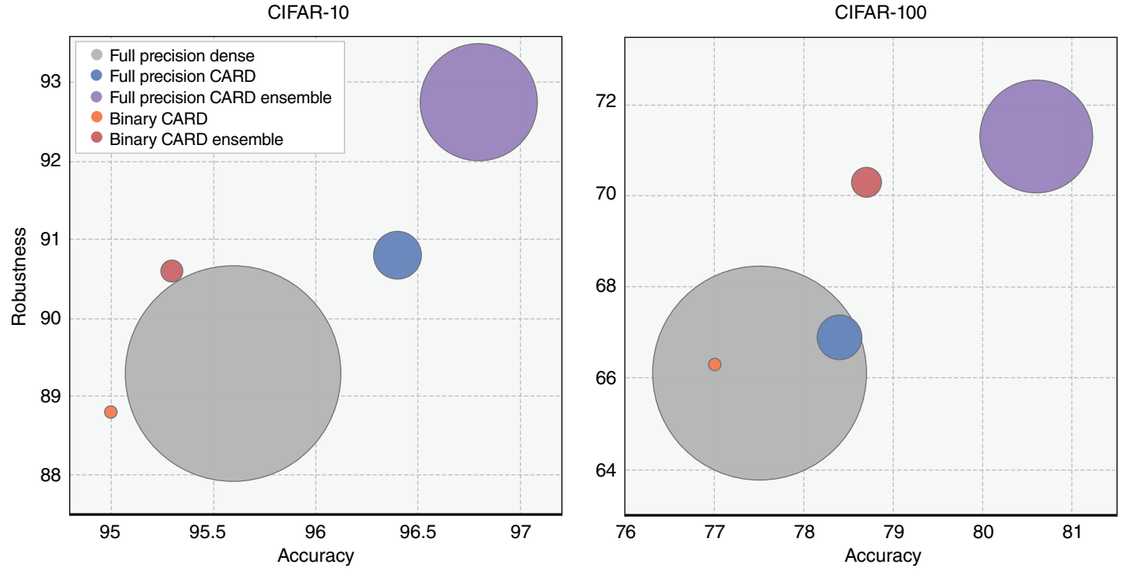
Real-World Deployment

The prevailing belief among researchers had been that only extremely large, dense neural networks can be robust, creating a challenge for running machine-learning systems on small, resource-constrained devices such as drones, smartphones, and other edge devices. The Livermore team has overcome this challenge with a multi-prize ticket approach creating models Kailkhura describes as "compact, accurate, and robust deep neural networks (CARDs)."

CARDS, sparse binary subnetworks, exist within a random-weight neural network in which weights are the model parameters that determine the influence of inputs on outputs. Despite having untrained weights, CARDS are drastically more robust compared to their more computationally expensive,



Livermore's computational researchers have trained the multi-prize ticket neural networks to accurately classify information using fewer nodes than a randomly initialized neural network to make the right decisions in high-regret situations, despite perturbations such as blur and brightness.



As illustrated for two data sets (CIFAR-10 and CIFAR-100), certain model compression strategies can produce compact, accurate, and robust deep neural networks (CARDs) to provide comparable (or better) accuracy and robustness than the dense baseline while significantly reducing the memory overhead.

traditional counterparts. "The math indicates, in principle, that you can make networks smaller while matching the accuracy and robustness of larger models," says Livermore postdoctoral researcher Brian Bartoldson. "This finding stands in sharp contrast to the current trend of making models bigger to achieve robustness."

CARDs have important implications for Livermore's mission-critical space by enhancing the accuracy and reliability of machine learning for national security applications. Opportunities for commercial and research uses have emerged as well; CARDs enable accurate and robust AI models on computing platforms that had been too small to run large neural networks and that lacked the power to process huge networks on the devices themselves. Livermore computer scientist James Diffenderfer says, "If we could equip edge devices with their own neural networks or deep-learning models to classify the data they're collecting, we would accelerate discovery of what edge devices could do." In addition to everyday commercial uses—smartphones, watches, and other hardware—CARDs can benefit research areas such as space exploration that require equipment with small, efficient AI systems and models providing accurate predictions for months or years with few interventions.

These breakthroughs offer a huge opportunity to rapidly accelerate new discoveries in a variety of scientific fields and offer the potential to be much more accessible to institutions and research teams that lack the storage capacity, power, and high-performance computing requirements that large-scale machine-learning models require. Kailkhura says, "CARDs are significantly lighter, faster, and more power-efficient while maintaining state-of-the-art performance. They can unlock a range of potential uses in which deep learning struggles due to its inefficiency and brittleness. Collaborative autonomy; natural-language-processing models; and environmental, ocean, or urban monitoring using edge computing applications with affordable sensors are just some of the applications that could be better enabled by this research."

— Sheridan Hyland

Key Words: artificial intelligence (AI); artificial neural network (ANN); compact, accurate, and robust deep neural networks (CARDs); machine learning; multi-prize ticket; neural network.

For further information contact Bhavya Kailkhura (925) 422-5810 (kailkhura1@llnl.gov).