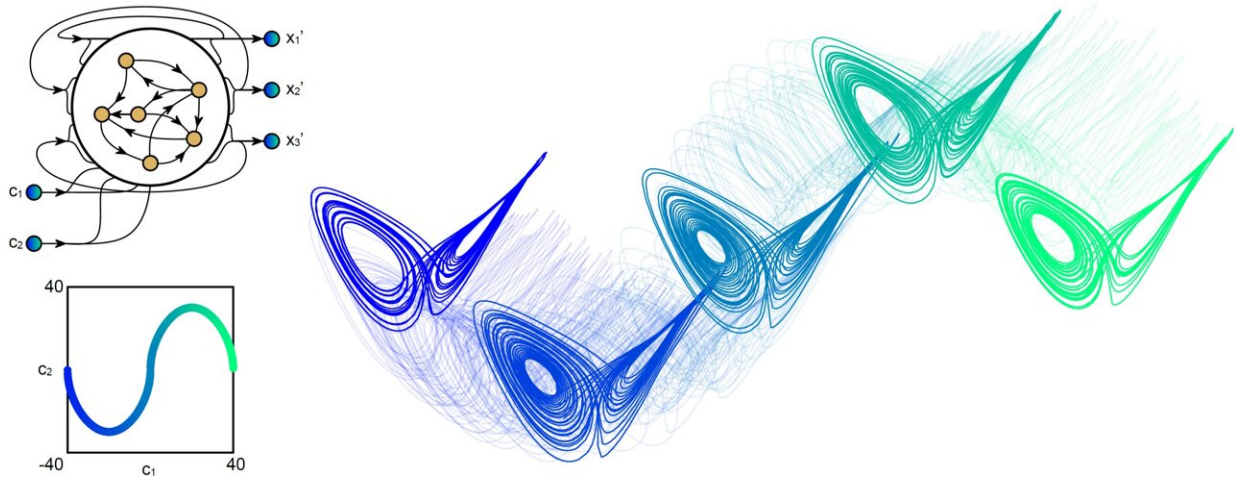
A recurrent neural network that infers the global temporal structure based on local examples

June 1 2021, by Ingrid Fadelli



After training the RNN on a few translated versions of the Lorenz attractor, the RNN stores the attractor as a memory and can translate its internal representation of the Lorenz by changing context variables. Credit: Kim et al.

Most computer systems are designed to store and manipulate information, such as documents, images, audio files and other data. While conventional computers are programmed to perform specific operations on structured data, emerging neuro-inspired systems can learn to solve tasks more adaptively, without having to be engineered to carry out a set type of operations.

Researchers at University of Pennsylvania and University of California recently trained a recurrent neural network (RNN) to adapt its representation of complex information based only on local data examples. In a paper published in *Nature Machine Intelligence*, they introduced this RNN and outlined the key learning mechanism underpinning its functioning.

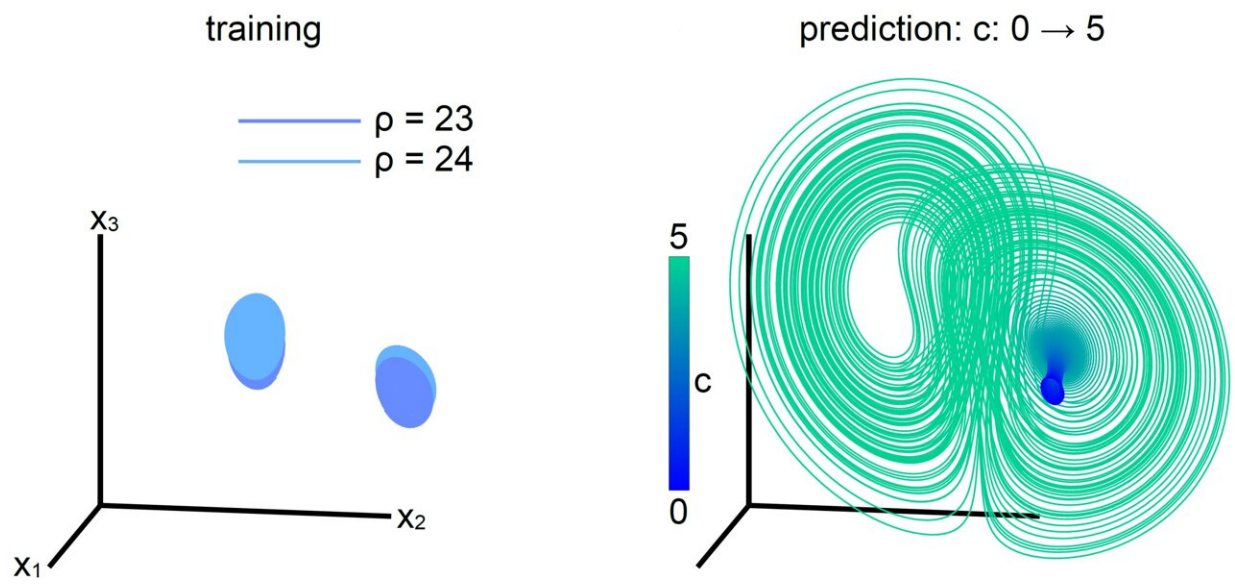
"Every day, we manipulate information about the world to make predictions," Jason Kim, one of the researchers who carried out the study, told TechXplore. "How much longer can I cook this pasta before it becomes soggy? How much later can I leave for work before rush hour? Such information representation and computation broadly fall into the category of working memory. While we can program a computer to build models of pasta texture or commute times, our primary objective was to understand how a neural network learns to build models and make predictions only by observing examples."

Kim, his mentor Danielle S. Bassett and the rest of their team showed that the two key mechanisms through which a neural network learns to make predictions are associations and context. For instance, if they wanted to teach their RNN to change the pitch of a song, they fed it the original song and two other versions of it, one with a slightly higher pitch and the other with a slightly lower pitch.

For each shift in pitch, the researchers 'biased' the RNN with a context variable. Subsequently, they trained it to store the original and modified songs within its working memory. This allowed the RNN to associate the pitch shifting operation with the context variable and manipulate its memory to change a song's pitch further, simply by changing the context variable.

"When one of our collaborators, Zhixin Lu, told us about an RNN that could learn to store information in working memory, we knew our

objective was in sight," Kim said. "Theoretically, the RNN evolves forward in time according to an equation. We derived the equation that quantifies how a small change in the context variable causes a small change in the RNN's trajectory and asked what conditions need to be met for the small change in the RNN's trajectory to yield the desired change in representation."



After training the RNN on a few stable trajectories of the Lorenz system, the RNN learns to correctly infer the bifurcation into the global Lorenz structure. Credit: Kim et al.

Kim and his colleagues observed that when the differences between training data examples were small (e.g., small differences/changes in pitch), their RNN associated the differences with the context variable. Notably, their study also identifies a simple mechanism through which neural networks can learn computations using their working memory.

"A great example is actually seen in a popular video of a stalking cat," Kim explained. "Here, the camera periodically moves in and out of view and the recorded cat inches closer only when the camera is out of view and stays frozen when the camera is in view. Just by observing the first few motions, we can predict the end result: a proximal cat."

While many past studies showed how neural networks manipulate their outputs, the work by Kim and his colleagues is among the first to identify a simple neural mechanism through which RNNs manipulate their memories, while retaining them even in the absence of inputs.

"Our most notable finding is that, not only do RNNs learn to continuously manipulate information in working memory, but they actually make accurate inferences about global structure when only trained on very local examples," Kim said. "It's a bit like accurately predicting the flourishing melodies of Chopin's Fantaisie Impromptu after only having heard the first few notes."

The recent paper by Kim and his colleagues introduces a quantitative model with falsifiable hypotheses of working memory that could also be relevant in the field of neuroscience. In addition, it outlines design principles that could aid the understanding of neural networks that are typically perceived as black boxes (i.e., that do not clearly explain the processes behind their predictions).

"Our findings also demonstrate that, when designed properly, neural networks have incredible power to accurately generalize outside of their training examples," Kim said. "We are now exploring many other exciting research directions. These go from studying the changes in the RNN's internal representation during learning to using context variables to switch between memories, to programming computations in RNNs without training."

More information: Jason Z. Kim et al, Teaching recurrent neural networks to infer global temporal structure from local examples, *Nature Machine Intelligence* (2021). DOI: 10.1038/s42256-021-00321-2