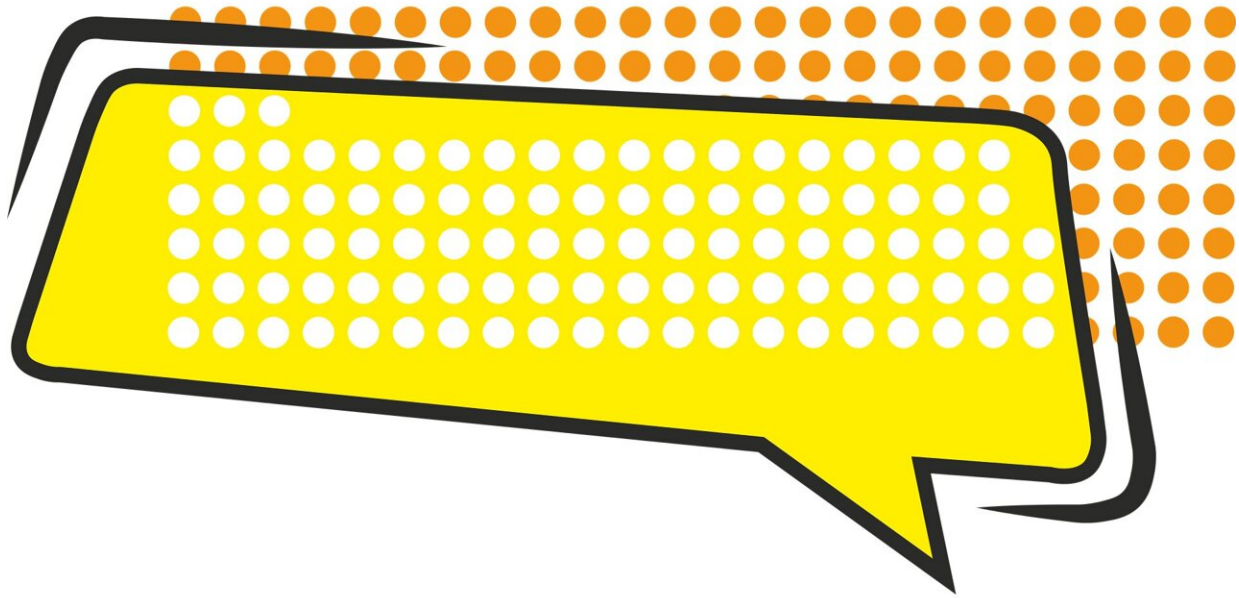
Computer scientists introduce a new method to reduce the size of multilingual language models

December 7 2023, by Jaimie Patterson



Credit: Pixabay/CC0 Public Domain

Multilingual language models, or MLMs, are machine learning models that can predict, generate, and extract text from more than one language. They're useful for cross-lingual communication, translation, and more—but tend to work best when they're only focused on a few languages.

As language models grow larger, their performance improves—as long as they only operate in a single language. Despite increasing the size of a model, the addition of more languages can undermine its performance due to "language interference," where the parameters (or variables) of a model that control its behavior in one language negatively impact its performance in another.

However, a team of Johns Hopkins computer scientists has developed a new approach to optimizing MLMs for multiple languages. Called Language-Specific Matrix Synthesis, their method reduces the number of parameters needed for a model to function in each new language.

The researchers are presenting their work this week at the 2023 Conference on Empirical Methods in Natural Language Processing in Singapore.

"Our focus was on achieving comparable performance while using fewer parameters," explains team member Haoran Xu, a doctoral candidate in the Whiting School of Engineering's Department of Computer Science who is advised by co-authors Philipp Koehn, a professor of computer science affiliated with the Center for Language and Speech Processing, and Kenton Murray, a research scientist at the Human Language Technology Center of Excellence and a member of CLSP.

As opposed to the traditional approach of designing separate dense neural networks—computing systems that loosely mimic the workings of the human brain—for each additional language in an MLM, the team opted to use low-rank matrices, which organize information by compressing data to reduce the number of parameters needed to accommodate a new language.

This allowed to the team to add new languages without needing as many parameters, avoiding what Xu calls an "explosion of parameters" at

scale.

"Imagine a classroom with 100 children, each representing a different language," explains Xu.

"Giving each child a full set of paints to express themselves—or perform tasks in their language—would require massive amounts of pigment or model parameters. Instead, if you have them share only red, yellow, and blue, the children can still create the full-color spectrum while using far less pigment and far fewer parameters. And since only one child can paint at a time, all 100 children can share that single three-color palette, drastically reducing parameter needs."

The team has proven in tests with a model capable of understanding up to 95 different languages that their method achieves superior performance in multilingual settings, all while using fewer parameters. Crucially, this allows for a significant reduction in a language model's size without compromising its performance.

Due to the reduced hardware requirements needed to deploy a smaller language model, a single, portable AI application using the Language-Specific Matrix Synthesis method may soon be capable of handling hundreds of languages instead of just a few, the team predicts.

"Our findings indicate the feasibility of deploying truly multilingual AI models in devices of all sizes," adds Xu.

The researchers say their objective is to apply their method to unwieldy MLMs and develop robust AI systems that can comprehend multiple languages while performing as effectively as they do in English.

Provided by Johns Hopkins University