New method improves efficiency of vision transformer AI systems

June 1 2023, by Matt Shipman

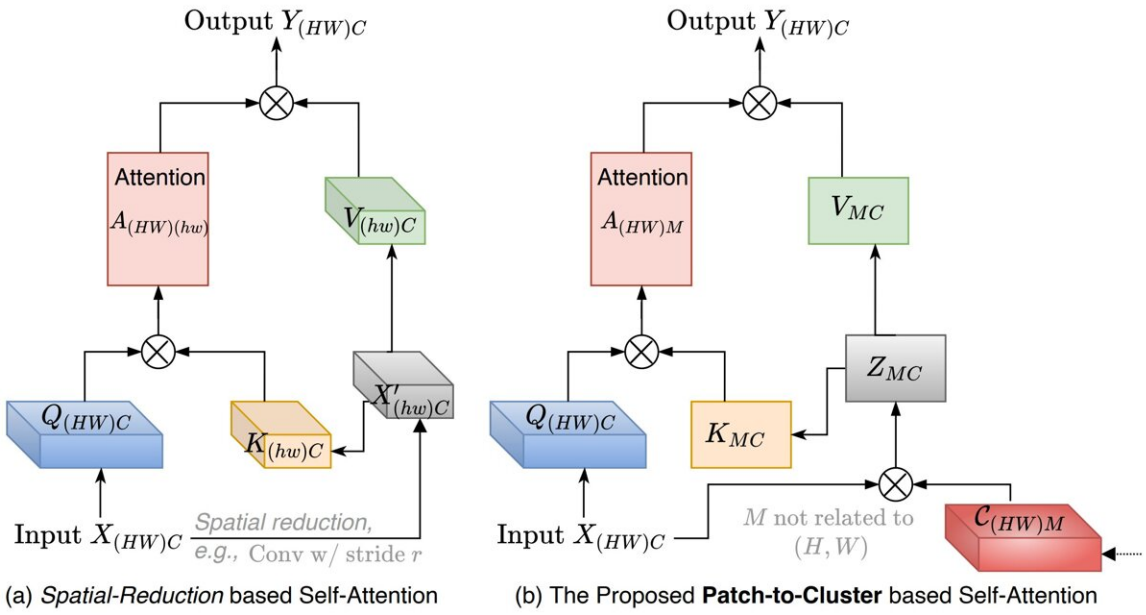

Illustration of (a) the spatial reduction based selfattention and (b) the proposed PaCa module in vision applications, where (HW) represents the number of patches in the input with H and W the height and width respectively, and M a predefined small number of clusters (e.g., M = 100). Credit: *arXiv* (2022). DOI: 10.48550/arxiv.2203.11987

Vision transformers (ViTs) are powerful artificial intelligence (AI) technologies that can identify or categorize objects in images—however, there are significant challenges related to both computing power

requirements and decision-making transparency. Researchers have now developed a new methodology that addresses both challenges, while also improving the ViT's ability to identify, classify and segment objects in images.

Transformers are among the most powerful existing AI models. For example, ChatGPT is an AI that uses transformer architecture, but the inputs used to train it are language. ViTs are transformer-based AI that are trained using visual inputs. For example, ViTs could be used to detect and categorize objects in an image, such as identifying all of the cars or all of the pedestrians in an image.

However, ViTs face two challenges.

First, transformer models are very complex. Relative to the amount of data being plugged into the AI, transformer models require a significant amount of computational power and use a large amount of memory. This is particularly problematic for ViTs, because images contain so much data.

Second, it is difficult for users to understand exactly how ViTs make decisions. For example, you might have trained a ViT to identify dogs in an image. But it's not entirely clear how the ViT is determining what is a dog and what is not. Depending on the application, understanding the ViT's decision-making process, also known as its model interpretability, can be very important.

The new ViT methodology, called "Patch-to-Cluster attention" (PaCa), addresses both challenges.

"We address the challenge related to computational and memory demands by using clustering techniques, which allow the transformer architecture to better identify and focus on objects in an image," says

Tianfu Wu, corresponding author of a paper on the work and an associate professor of electrical and computer engineering at North Carolina State University.

"Clustering is when the AI lumps sections of the image together, based on similarities it finds in the image data. This significantly reduces computational demands on the system. Before clustering, computational demands for a ViT are quadratic. For example, if the system breaks an image down into 100 smaller units, it would need to compare all 100 units to each other—which would be 10,000 complex functions."

"By clustering, we're able to make this a linear process, where each smaller unit only needs to be compared to a predetermined number of clusters. Let's say you tell the system to establish 10 clusters; that would only be 1,000 complex functions," Wu says.

"Clustering also allows us to address model interpretability, because we can look at how it created the clusters in the first place. What features did it decide were important when lumping these sections of data together? And because the AI is only creating a small number of clusters, we can look at those pretty easily."

The researchers did comprehensive testing of PaCa, comparing it to two state-of-the-art ViTs called SWin and PVT.

"We found that PaCa outperformed SWin and PVT in every way," Wu says. "PaCa was better at classifying objects in images, better at identifying objects in images, and better at segmentation—essentially outlining the boundaries of objects in images. It was also more efficient, meaning that it was able to perform those tasks more quickly than the other ViTs."

"The next step for us is to scale up PaCa by training on larger,

foundational data sets."

The paper, "PaCa-ViT: Learning Patch-to-Cluster Attention in Vision Transformers," will be presented at the IEEE/CVF Conference on Computer Vision and Pattern Recognition, being held June 18-22 in Vancouver, Canada.

It is published on the *arXiv* preprint server.

More information: Ryan Grainger et al, PaCa-ViT: Learning Patch-to-Cluster Attention in Vision Transformers, *arXiv* (2022). DOI: 10.48550/arxiv.2203.11987

Conference: cvpr2023.thecvf.com/

Provided by North Carolina State University