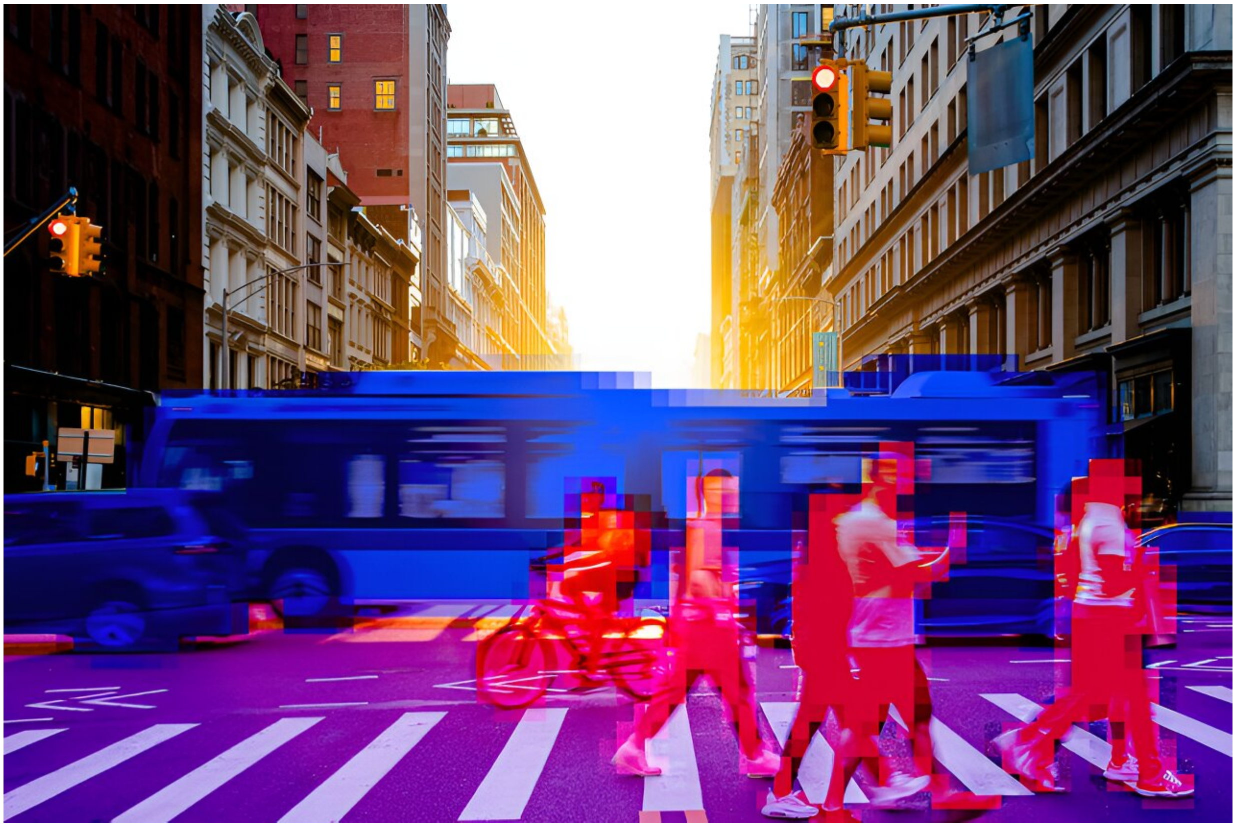
AI model speeds up high-resolution computer vision

September 12 2023, by Adam Zewe



A machine-learning model for high-resolution computer vision could enable computationally intensive vision applications, such as autonomous driving or medical image segmentation, on edge devices. Pictured is an artist's interpretation of the autonomous driving technology. Credit: Massachusetts Institute of Technology

An autonomous vehicle must rapidly and accurately recognize objects that it encounters, from an idling delivery truck parked at the corner to a cyclist whizzing toward an approaching intersection.

To do this, the vehicle might use a powerful computer vision model to categorize every pixel in a high-resolution image of this scene, so it doesn't lose sight of objects that might be obscured in a lower-quality image. But this task, known as semantic segmentation, is complex and requires a huge amount of computation when the image has high resolution.

Researchers from MIT, the MIT-IBM Watson AI Lab, and elsewhere have developed a more efficient computer vision model that vastly reduces the computational complexity of this task. Their model can perform semantic segmentation accurately in real-time on a device with limited hardware resources, such as the on-board computers that enable an autonomous vehicle to make split-second decisions.

Recent state-of-the-art semantic segmentation models directly learn the interaction between each pair of pixels in an image, so their calculations grow quadratically as image resolution increases. Because of this, while these models are accurate, they are too slow to process high-resolution images in real time on an edge device like a sensor or mobile phone.

The MIT researchers designed a new building block for semantic segmentation models that achieves the same abilities as these state-of-the-art models, but with only linear computational complexity and hardware-efficient operations.

The result is a new model series for high-resolution computer vision that performs up to nine times faster than prior models when deployed on a mobile device. Importantly, this new model series exhibited the same or better accuracy than these alternatives.

Not only could this technique be used to help autonomous vehicles make decisions in real-time, it could also improve the efficiency of other high-resolution computer vision tasks, such as medical image segmentation.

"While researchers have been using traditional vision transformers for quite a long time, and they give amazing results, we want people to also pay attention to the efficiency aspect of these models. Our work shows that it is possible to drastically reduce the computation so this real-time image segmentation can happen locally on a device," says Song Han, an associate professor in the Department of Electrical Engineering and Computer Science (EECS), a member of the MIT-IBM Watson AI Lab, and senior author of the paper describing the new model.

He is joined on the paper by lead author Han Cai, an EECS graduate student; Junyan Li, an undergraduate at Zhejiang University; Muyan Hu, an undergraduate student at Tsinghua University; and Chuang Gan, a principal research staff member at the MIT-IBM Watson AI Lab. The research will be presented at the International Conference on Computer Vision held in Paris, October 2–6. It is available on the *arXiv* preprint server.

A simplified solution

Categorizing every pixel in a high-resolution image that may have millions of pixels is a difficult task for a machine-learning model. A powerful new type of model, known as a vision transformer, has recently been used effectively.

Transformers were originally developed for natural language processing. In that context, they encode each word in a sentence as a token and then generate an attention map, which captures each token's relationships with all other tokens. This attention map helps the model understand context when it makes predictions.

Using the same concept, a vision transformer chops an image into patches of pixels and encodes each small patch into a token before generating an attention map. In generating this attention map, the model uses a similarity function that directly learns the interaction between each pair of pixels. In this way, the model develops what is known as a global receptive field, which means it can access all the relevant parts of the image.

Since a high-resolution image may contain millions of pixels, chunked into thousands of patches, the attention map quickly becomes enormous. Because of this, the amount of computation grows quadratically as the resolution of the image increases.

In their new model series, called EfficientViT, the MIT researchers used a simpler mechanism to build the attention map—replacing the nonlinear similarity function with a linear similarity function. As such, they can rearrange the order of operations to reduce total calculations without changing functionality and losing the global receptive field. With their model, the amount of computation needed for a prediction grows linearly as the image resolution grows.

"But there is no free lunch. The linear attention only captures global context about the image, losing local information, which makes the accuracy worse," Han says.

To compensate for that accuracy loss, the researchers included two extra components in their model, each of which adds only a small amount of computation.

One of those elements helps the model capture local feature interactions, mitigating the linear function's weakness in local information extraction. The second, a module that enables multiscale learning, helps the model recognize both large and small objects.

"The most critical part here is that we need to carefully balance the performance and the efficiency," Cai says.

They designed EfficientViT with a hardware-friendly architecture, so it could be easier to run on different types of devices, such as virtual reality headsets or the edge computers on autonomous vehicles. Their model could also be applied to other computer vision tasks, like image classification.

Streamlining semantic segmentation

When they tested their model on datasets used for semantic segmentation, they found that it performed up to nine times faster on a Nvidia graphics processing unit (GPU) than other popular vision transformer models, with the same or better accuracy.

"Now, we can get the best of both worlds and reduce the computing to make it fast enough that we can run it on mobile and cloud devices," Han says.

Building off these results, the researchers want to apply this technique to speed up generative machine-learning models, such as those used to generate new images. They also want to continue scaling up EfficientViT for other vision tasks.

"Efficient transformer models, pioneered by Professor Song Han's team, now form the backbone of cutting-edge techniques in diverse computer vision tasks, including detection and segmentation," says Lu Tian, senior director of AI algorithms at AMD, Inc., who was not involved with this paper. "Their research not only showcases the efficiency and capability of transformers, but also reveals their immense potential for real-world applications, such as enhancing image quality in video games."

"Model compression and light-weight model design are crucial research topics toward efficient AI computing, especially in the context of large foundation models. Professor Song Han's group has shown remarkable progress compressing and accelerating modern deep learning models, particularly vision transformers," adds Jay Jackson, global vice president of artificial intelligence and machine learning at Oracle, who was not involved with this research. "Oracle Cloud Infrastructure has been supporting his team to advance this line of impactful research toward efficient and green AI."

More information: Han Cai et al, EfficientViT: Lightweight Multi-Scale Attention for On-Device Semantic Segmentation, *arXiv* (2022). DOI: 10.48550/arxiv.2205.14756

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology

Citation: AI model speeds up high-resolution computer vision (2023, September 12) retrieved 8 May 2024 from https://techxplore.com/news/2023-09-ai-high-resolution-vision.html