**AI model speeds up high-resolution computer vision**

The system could improve image quality in video streaming or help autonomous vehicles identify road hazards in real-time.

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.

**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.”

**Модель искусственного интеллекта ускоряет компьютерное зрение высокого разрешения**

Система может улучшить качество изображения при потоковом видео или помочь автономным транспортным средствам идентифицировать дорожные опасности в режиме реального времени.

Автономное транспортное средство должно быстро и точно распознавать объекты, с которыми оно сталкивается: от стоящего на холостом ходу грузовика, припаркованного на углу, до велосипедиста, мчащегося к приближающемуся перекрестку.

Для этого автомобиль может использовать мощную модель компьютерного зрения, чтобы классифицировать каждый пиксель изображения этой сцены с высоким разрешением, чтобы не упускать из виду объекты, которые могут быть скрыты на изображении более низкого качества. Но эта задача, известная как семантическая сегментация, сложна и требует огромного количества вычислений, когда изображение имеет высокое разрешение.

Исследователи из Массачусетского технологического института, Лаборатории искусственного интеллекта MIT-IBM Watson и других организаций разработали более эффективную модель компьютерного зрения, которая значительно снижает вычислительную сложность этой задачи. Их модель может точно выполнять семантическую сегментацию в режиме реального времени на устройстве с ограниченными аппаратными ресурсами, например, на бортовых компьютерах, которые позволяют автономному транспортному средству принимать решения за доли секунды.

Последние современные модели семантической сегментации напрямую изучают взаимодействие между каждой парой пикселей изображения, поэтому их расчеты растут квадратично по мере увеличения разрешения изображения. По этой причине, хотя эти модели и точны, они слишком медленны для обработки изображений высокого разрешения в реальном времени на периферийном устройстве, таком как датчик или мобильный телефон.

Исследователи Массачусетского технологического института разработали новый строительный блок для моделей семантической сегментации, который обеспечивает те же возможности, что и эти современные модели, но с линейной вычислительной сложностью и эффективными аппаратными операциями.

Результатом стала новая серия моделей для компьютерного зрения высокого разрешения, которая при развертывании на мобильном устройстве работает до девяти раз быстрее, чем предыдущие модели. Важно отметить, что эта новая серия моделей продемонстрировала такую ​​же или лучшую точность, чем эти альтернативы.

Этот метод можно использовать не только для помощи автономным транспортным средствам в принятии решений в режиме реального времени, но и для повышения эффективности других задач компьютерного зрения с высоким разрешением, таких как сегментация медицинских изображений.

«Хотя исследователи уже довольно давно используют традиционные преобразователи зрения, и они дают потрясающие результаты, мы хотим, чтобы люди также обращали внимание на аспект эффективности этих моделей. Наша работа показывает, что можно значительно сократить объем вычислений, чтобы сегментация изображений в реальном времени могла происходить локально на устройстве», — говорит Сун Хан, доцент кафедры электротехники и информатики (EECS), член Лаборатории искусственного интеллекта MIT-IBM Watson и старший автор статьи, описывающей новую модель.

В работе над статьей к нему присоединяется ведущий автор Хан Цай, аспирант EECS; Джуньянь Ли, студентка Чжэцзянского университета; Муян Ху, студентка Университета Цинхуа; и Чуанг Ган, главный научный сотрудник лаборатории MIT-IBM Watson AI Lab. Исследование будет представлено на Международной конференции по компьютерному зрению.

**Упрощенное решение**

Классификация каждого пикселя изображения с высоким разрешением, которое может содержать миллионы пикселей, является сложной задачей для модели машинного обучения. В последнее время эффективно используется новый мощный тип модели, известный как преобразователь зрения.

Трансформеры изначально были разработаны для обработки естественного языка. В этом контексте они кодируют каждое слово в предложении как токен, а затем генерируют карту внимания, которая фиксирует отношения каждого токена со всеми другими токенами. Эта карта внимания помогает модели понимать контекст, когда она делает прогнозы.

Используя ту же концепцию, преобразователь зрения разбивает изображение на участки пикселей и кодирует каждый небольшой участок в токен перед созданием карты внимания. При создании этой карты внимания модель использует функцию сходства, которая напрямую изучает взаимодействие между каждой парой пикселей. Таким образом, модель создает так называемое глобальное рецептивное поле, что означает, что она может получить доступ ко всем соответствующим частям изображения.

Поскольку изображение с высоким разрешением может содержать миллионы пикселей, разбитых на тысячи участков, карта внимания быстро становится огромной. Из-за этого объем вычислений растет квадратично по мере увеличения разрешения изображения.

В своей новой серии моделей под названием EfficientViT исследователи из Массачусетского технологического института использовали более простой механизм построения карты внимания — замену нелинейной функции сходства на линейную функцию сходства. Таким образом, они могут изменить порядок операций, чтобы сократить общий объем вычислений, не меняя при этом функциональность и не теряя глобального рецептивного поля. В их модели объем вычислений, необходимый для прогнозирования, растет линейно по мере роста разрешения изображения.

«Но бесплатных обедов не бывает. Линейное внимание фиксирует только глобальный контекст изображения, теряя локальную информацию, что ухудшает точность», — говорит Хан.

Чтобы компенсировать эту потерю точности, исследователи включили в свою модель два дополнительных компонента, каждый из которых требует лишь небольшого объема вычислений.

Один из этих элементов помогает модели фиксировать взаимодействие локальных объектов, смягчая слабость линейной функции при извлечении локальной информации. Второй модуль, обеспечивающий многомасштабное обучение, помогает модели распознавать как большие, так и маленькие объекты.

«Самое важное здесь заключается в том, что нам необходимо тщательно сбалансировать производительность и эффективность», — говорит Цай.

Они разработали EfficientViT с дружественной к аппаратному обеспечению архитектурой, поэтому его будет проще запускать на различных типах устройств, таких как гарнитуры виртуальной реальности или периферийные компьютеры на автономных транспортных средствах. Их модель также может быть применена к другим задачам компьютерного зрения, таким как классификация изображений.

**Оптимизация семантической сегментации**

Когда они протестировали свою модель на наборах данных, используемых для семантической сегментации, они обнаружили, что на графическом процессоре (GPU) Nvidia она работает до девяти раз быстрее, чем другие популярные модели преобразователей зрения, с такой же или лучшей точностью.

«Теперь мы можем получить лучшее из обоих миров и сократить объем вычислений, чтобы сделать их достаточно быстрыми, чтобы мы могли запускать их на мобильных и облачных устройствах», — говорит Хан.

Основываясь на этих результатах, исследователи хотят применить эту технику для ускорения генеративных моделей машинного обучения, например тех, которые используются для создания новых изображений. Они также хотят продолжить масштабирование EfficientViT для решения других задач машинного зрения.

«Эффективные модели-трансформеры, впервые разработанные командой профессора Сун Хана, теперь составляют основу передовых технологий в различных задачах компьютерного зрения, включая обнаружение и сегментацию», — говорит Лу Тянь, старший директор по алгоритмам искусственного интеллекта в AMD, Inc., который был не связан с этой статьей. «Их исследования не только демонстрируют эффективность и возможности преобразователей, но также раскрывают их огромный потенциал для реальных приложений, таких как повышение качества изображения в видеоиграх».

«Сжатие моделей и проектирование облегченных моделей являются важнейшими темами исследований для эффективных вычислений ИИ, особенно в контексте крупных базовых моделей. Группа профессора Сон Хана продемонстрировала замечательный прогресс в сжатии и ускорении современных моделей глубокого обучения, в частности преобразователей зрения», — добавляет Джей Джексон, глобальный вице-президент по искусственному интеллекту и машинному обучению в Oracle, который не участвовал в этом исследовании. «Oracle Cloud Infrastructure поддерживает его команду в продвижении этого направления эффективных исследований в направлении эффективного и экологичного искусственного интеллекта».