**Бочковой И.П М211**

**Learning to grow machine-learning models**

**New LiGO technique accelerates training of large machine-learning models, reducing the monetary and environmental cost of developing AI applications.**

It’s no secret that OpenAI’s ChatGPT has some incredible capabilities — for instance, the chatbot can write poetry that resembles Shakespearean sonnets or debug code for a computer program. These abilities are made possible by the massive machine-learning model that ChatGPT is built upon. Researchers have found that when these types of models become large enough, extraordinary capabilities emerge.But bigger models also require more time and money to train. The training process involves showing hundreds of billions of examples to a model. Gathering so much data is an involved process in itself. Then come the monetary and environmental costs of running many powerful computers for days or weeks to train a model that may have billions of parameters. “It’s been estimated that training models at the scale of what ChatGPT is hypothesized to run on could take millions of dollars, just for a single training run. Can we improve the efficiency of these training methods, so we can still get good models in less time and for less money? We propose to do this by leveraging smaller language models that have previously been trained,” says Yoon Kim, an assistant professor in MIT’s Department of Electrical Engineering and Computer Science and a member of the Computer Science and Artificial Intelligence Laboratory (CSAIL).Rather than discarding a previous version of a model, Kim and his collaborators use it as the building blocks for a new model. Using machine learning, their method learns to “grow” a larger model from a smaller model in a way that encodes knowledge the smaller model has already gained. This enables faster training of the larger model.Their technique saves about 50 percent of the computational cost required to train a large model, compared to methods that train a new model from scratch. Plus, the models trained using the MIT method performed as well as, or better than, models trained with other techniques that also use smaller models to enable faster training of larger models.Reducing the time it takes to train huge models could help researchers make advancements faster with less expense, while also reducing the carbon emissions generated during the training process. It could also enable smaller research groups to work with these massive models, potentially opening the door to many new advances.“As we look to democratize these types of technologies, making training faster and less expensive will become more important,” says Kim, senior author of a [paper](https://openreview.net/pdf?id=cDYRS5iZ16f) on this technique.Kim and his graduate student Lucas Torroba Hennigen wrote the paper with lead author Peihao Wang, a graduate student at the University of Texas at Austin, as well as others at the MIT-IBM Watson AI Lab and Columbia University. The research will be presented at the International Conference on Learning Representations.

**The bigger the better**

Large language models like GPT-3, which is at the core of ChatGPT, are built using a neural network architecture called a transformer. A neural network, loosely based on the human brain, is composed of layers of interconnected nodes, or “neurons.” Each neuron contains parameters, which are variables learned during the training process that the neuron uses to process data.Transformer architectures are unique because, as these types of neural network models get bigger, they achieve much better results. “This has led to an arms race of companies trying to train larger and larger transformers on larger and larger datasets. More so than other architectures, it seems that transformer networks get much better with scaling. We’re just not exactly sure why this is the case,” Kim says.These models often have hundreds of millions or billions of learnable parameters. Training all these parameters from scratch is expensive, so researchers seek to accelerate the process. One effective technique is known as model growth. Using the model growth method, researchers can increase the size of a transformer by copying neurons, or even entire layers of a previous version of the network, then stacking them on top. They can make a network wider by adding new neurons to a layer or make it deeper by adding additional layers of neurons. In contrast to previous approaches for model growth, parameters associated with the new neurons in the expanded transformer are not just copies of the smaller network’s parameters, Kim explains. Rather, they are learned combinations of the parameters of the smaller model.

**Learning to grow**

Kim and his collaborators use machine learning to learn a linear mapping of the parameters of the smaller model. This linear map is a mathematical operation that transforms a set of input values, in this case the smaller model’s parameters, to a set of output values, in this case the parameters of the larger model.Their method, which they call a learned Linear Growth Operator (LiGO), learns to expand the width and depth of larger network from the parameters of a smaller network in a data-driven way. But the smaller model may actually be quite large — perhaps it has a hundred million parameters — and researchers might want to make a model with a billion parameters. So the LiGO technique breaks the linear map into smaller pieces that a machine-learning algorithm can handle. LiGO also expands width and depth simultaneously, which makes it more efficient than other methods. A user can tune how wide and deep they want the larger model to be when they input the smaller model and its parameters, Kim explains. When they compared their technique to the process of training a new model from scratch, as well as to model-growth methods, it was faster than all the baselines. Their method saves about 50 percent of the computational costs required to train both vision and language models, while often improving performance. The researchers also found they could use LiGO to accelerate transformer training even when they didn’t have access to a smaller, pretrained model. “I was surprised by how much better all the methods, including ours, did compared to the random initialization, train-from-scratch baselines.” Kim says. In the future, Kim and his collaborators are looking forward to applying LiGO to even larger models. The work was funded, in part, by the MIT-IBM Watson AI Lab, Amazon, the IBM Research AI Hardware Center, Center for Computational Innovation at Rensselaer Polytechnic Institute, and the U.S. Army Research Office.

**Перевод**

Не секрет, что ChatGPT от OpenAI обладает невероятными возможностями — например, чат-бот может писать стихи, напоминающие шекспировские сонеты, или отлаживать код для компьютерной программы. Эти возможности стали возможными благодаря массивной модели машинного обучения, на которой построен ChatGPT. Исследователи обнаружили, что, когда эти типы моделей становятся достаточно большими, появляются экстраординарные способности. Но более крупные модели также требуют больше времени и денег для обучения. Процесс обучения включает показ модели сотен миллиардов примеров. Сбор такого количества данных сам по себе является сложным процессом. Затем следуют денежные и экологические затраты, связанные с запуском множества мощных компьютеров в течение нескольких дней или недель для обучения модели, которая может иметь миллиарды параметров. «Было подсчитано, что модели обучения в масштабе того, на чем предполагается работать ChatGPT, могут потребовать миллионы долларов только для одного тренировочного прогона. Можем ли мы повысить эффективность этих методов обучения, чтобы получить хорошие модели за меньшее время и за меньшие деньги? Мы предлагаем сделать это, используя более мелкие языковые модели, которые были обучены ранее», — говорит Юн Ким, доцент кафедры электротехники и компьютерных наук Массачусетского технологического института и член Лаборатории компьютерных наук и искусственного интеллекта (CSAIL). отбрасывая предыдущую версию модели, Ким и его сотрудники используют ее в качестве строительных блоков для новой модели. Используя машинное обучение, их метод учит «выращивать» большую модель из меньшей модели таким образом, чтобы кодировать знания, которые меньшая модель уже получила. Это позволяет быстрее обучать более крупную модель. Их метод экономит около 50 % вычислительных затрат, необходимых для обучения большой модели, по сравнению с методами обучения новой модели с нуля. Кроме того, модели, обученные с использованием метода Массачусетского технологического института, работали так же или даже лучше, чем модели, обученные с помощью других методов, которые также используют модели меньшего размера для более быстрого обучения более крупных моделей. Сокращение времени, необходимого для обучения больших моделей, может помочь исследователям добиться прогресса. быстрее с меньшими затратами, а также сокращая выбросы углекислого газа, образующиеся во время тренировочного процесса. Это также может позволить небольшим исследовательским группам работать с этими массивными моделями, потенциально открывая двери для многих новых достижений. «По мере того, как мы стремимся демократизировать эти типы технологий, более важным станет ускорение и удешевление обучения», — говорит Ким, старший специалист. автор статьи об этой технике. Ким и его аспирант Лукас Торроба Хенниген написали статью вместе с ведущим автором Пейхао Ваном, аспирантом Техасского университета в Остине, а также другими сотрудниками Лаборатории искусственного интеллекта Watson MIT-IBM и Колумбийского университета. Университет. Исследование будет представлено на Международной конференции по обучающим представлениям. Большие языковые модели, такие как GPT-3, лежащие в основе ChatGPT, построены с использованием архитектуры нейронной сети, называемой преобразователем. Нейронная сеть, основанная на человеческом мозге, состоит из слоев взаимосвязанных узлов или «нейронов». Каждый нейрон содержит параметры, которые представляют собой переменные, полученные в процессе обучения и используемые нейроном для обработки данных. Архитектуры преобразователей уникальны, потому что по мере того, как эти типы моделей нейронных сетей становятся больше, они достигают гораздо лучших результатов. «Это привело к гонке вооружений компаний, пытающихся обучать все более и более крупные преобразователи на все более и более крупных наборах данных. Больше, чем другие архитектуры, кажется, что сети-трансформеры становятся намного лучше с масштабированием. Мы просто не совсем уверены, почему это так», — говорит Ким. Эти модели часто имеют сотни миллионов или миллиарды обучаемых параметров. Обучение всех этих параметров с нуля стоит дорого, поэтому исследователи стремятся ускорить процесс. Один из эффективных методов известен как рост модели. Используя метод роста модели, исследователи могут увеличить размер преобразователя, скопировав нейроны или даже целые слои предыдущей версии сети, а затем наложив их сверху. Они могут расширять сеть, добавляя новые нейроны в слой, или углублять ее, добавляя дополнительные слои нейронов. Ким объясняет, что в отличие от предыдущих подходов к росту модели параметры, связанные с новыми нейронами в расширенном преобразователе, являются не просто копиями параметров меньшей сети. Скорее, это заученные комбинации параметров меньшей модели.

Учимся расти

Ким и его сотрудники используют машинное обучение, чтобы изучить линейное отображение параметров меньшей модели. Эта линейная карта представляет собой математическую операцию, которая преобразует набор входных значений, в данном случае параметры меньшей модели, в набор выходных значений, в данном случае параметры большей модели. Их метод, который они называют обученным линейным ростом Оператор (LiGO) учится расширять ширину и глубину более крупной сети за счет параметров меньшей сети на основе данных. Но меньшая модель на самом деле может быть довольно большой — возможно, она имеет сто миллионов параметров — и исследователи могут захотеть создать модель с миллиардом параметров. Таким образом, метод LiGO разбивает линейную карту на более мелкие части, с которыми может справиться алгоритм машинного обучения. LiGO также одновременно увеличивает ширину и глубину, что делает его более эффективным, чем другие методы. Ким объясняет, что пользователь может настроить, насколько широкой и глубокой должна быть большая модель, когда он вводит меньшую модель и ее параметры. Когда они сравнили свою методику с процессом обучения новой модели с нуля, а также с методами роста модели, она оказалась быстрее всех базовых. Их метод экономит около 50 процентов вычислительных затрат, необходимых для обучения как зрительных, так и языковых моделей, часто повышая производительность. Исследователи также обнаружили, что они могут использовать LiGO для ускорения обучения трансформеров, даже если у них не было доступа к предварительно обученной модели меньшего размера. «Я был удивлен тем, насколько лучше работают все методы, включая наши, по сравнению со случайной инициализацией и обучением с нуля». — говорит Ким. В будущем Ким и его сотрудники надеются применить LiGO к еще более крупным моделям. Работа частично финансировалась Лабораторией искусственного интеллекта MIT-IBM Watson, Amazon, Центром аппаратного обеспечения искусственного интеллекта IBM, Центром вычислительных инноваций Политехнического института Ренсселера и Исследовательским бюро армии США.

**Summary**

The LiGO technique, developed by researchers at MIT and the University of Texas at Austin, aims to accelerate the training of large machine-learning models by using smaller language models as building blocks for a larger model. This saves around 50% of the computational cost required to train a large model, while still producing models that perform as well as, or better than, those trained with other techniques. The researchers hope their approach will democratize access to these technologies and reduce carbon emissions generated during the training process. Large language models like GPT-3 are built using a neural network architecture called a transformer, which achieves much better results as the model gets bigger.