Synthetic imagery sets new bar in AI training efficiency

November 20 2023, by Rachel Gordon

An MIT team studies the potential of learning visual representations using synthetic images generated by text-to-image models. They are the first to show that models trained solely with synthetic images outperform the counterparts trained with real images, in large-scale settings. Credit: Alex Shipps/MIT CSAIL via the Midjourney AI image generator

Data is the new soil, and in this fertile new ground, MIT researchers are planting more than just pixels. By using synthetic images to train machine learning models, a team of scientists recently surpassed results obtained from traditional "real-image" training methods.

At the core of the approach is a system called StableRep, which doesn't just use any synthetic images; it generates them through ultra-popular text-to-image models like Stable Diffusion. It's like creating worlds with words.

So what's in StableRep's secret sauce? A strategy called "multi-positive contrastive learning."

"We're teaching the model to learn more about high-level concepts through context and variance, not just feeding it data," says Lijie Fan, MIT Ph.D. student in electrical engineering, affiliate of the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL), lead researcher on the work currently posted to the *arXiv* preprint server.

"When multiple images, all generated from the same text, all treated as depictions of the same underlying thing, the model dives deeper into the concepts behind the images, say the object, not just their pixels."

This approach considers multiple images spawned from identical text prompts as positive pairs, providing additional information during training, not just adding more diversity but specifying to the vision system which images are alike and which are different. Remarkably, StableRep outshone the prowess of top-tier models trained on real images, such as SimCLR and CLIP, in extensive datasets.

"While StableRep helps mitigate the challenges of data acquisition in machine learning, it also ushers in a stride towards a new era of AI training techniques. The capacity to produce high-caliber, diverse

synthetic images on command could help curtail cumbersome expenses and resources," says Fan.

The process of data collection has never been straightforward. In the 1990s, researchers had to manually capture photographs to assemble datasets for objects and faces. The 2000s saw individuals scouring the internet for data. However, this raw, uncurated data often contained discrepancies when compared to real-world scenarios and reflected societal biases, presenting a distorted view of reality.

The task of cleansing datasets through human intervention is not only expensive, but also exceedingly challenging. Imagine, though, if this arduous data collection could be distilled down to something as simple as issuing a command in natural language.

A pivotal aspect of StableRep's triumph is the adjustment of the "guidance scale" in the generative model, which ensures a delicate balance between the synthetic images' diversity and fidelity. When finely tuned, synthetic images used in training these self-supervised models were found to be as effective, if not more so, than real images.

Taking it a step forward, language supervision was added to the mix, creating an enhanced variant: StableRep+. When trained with 20 million synthetic images, StableRep+ not only achieved superior accuracy but also displayed remarkable efficiency compared to CLIP models trained with a staggering 50 million real images.

Yet, the path ahead isn't without its potholes. The researchers candidly address several limitations, including the current slow pace of image generation, semantic mismatches between text prompts and the resultant images, potential amplification of biases, and complexities in image attribution, all of which are imperative to address for future advancements.

Another issue is that StableRep requires first training the generative model on large-scale real data. The team acknowledges that starting with real data remains a necessity; however, when you have a good generative model, you can repurpose it for new tasks, like training recognition models and visual representations.

The team notes that they haven't gotten around the need to start with real data; it's just that once you have a good generative model you can repurpose it for new tasks, like training recognition models and visual representations.

While StableRep offers a good solution by diminishing the dependency on vast real-image collections, it brings to the fore concerns regarding hidden biases within the uncurated data used for these text-to-image models. The choice of text prompts, integral to the image synthesis process, is not entirely free from bias, "indicating the essential role of meticulous text selection or possible human curation," says Fan.

"Using the latest text-to-image models, we've gained unprecedented control over image generation, allowing for a diverse range of visuals from a single text input. This surpasses real-world image collection in efficiency and versatility. It proves especially useful in specialized tasks, like balancing image variety in long-tail recognition, presenting a practical supplement to using real images for training," says Fan.

"Our work signifies a step forward in visual learning, towards the goal of offering cost-effective training alternatives while highlighting the need for ongoing improvements in data quality and synthesis."

"One dream of generative model learning has long been to be able to generate data useful for discriminative model training," says Google DeepMind researcher and University of Toronto professor of computer science David Fleet, who was not involved in the paper.

"While we have seen some signs of life, the dream has been elusive, especially on large-scale complex domains like high-resolution images. This paper provides compelling evidence, for the first time to my knowledge, that the dream is becoming a reality. They show that contrastive learning from massive amounts of synthetic image data can produce representations that outperform those learned from real data at scale, with the potential to improve myriad downstream vision tasks."

More information: Yonglong Tian et al, StableRep: Synthetic Images from Text-to-Image Models Make Strong Visual Representation Learners, *arXiv* (2023). DOI: 10.48550/arxiv.2306.00984

Provided by Massachusetts Institute of Technology