Helping computer vision and language models understand what they see

September 13 2023, by Adam Zewe

MIT researchers created a new annotated synthetic dataset of images that depict a wide range of scenarios, which can be used to help machine-learning models understand the concepts in a scene. Pictured is a scene from the synthetic dataset, and the detailed text description says, "This scene contains a box, and one human. They are in a castle ruin with old stones. The box is to the left of the human. The box is in front of the human. The human rotate jump. The human is male. The human wears a black t-shirt and dark blue jeans."

Powerful machine-learning algorithms known as vision and language models, which learn to match text with images, have shown remarkable results when asked to generate captions or summarize videos.

While these models excel at identifying objects, they often struggle to understand concepts, like object attributes or the arrangement of items in a scene. For instance, a vision and language model might recognize the cup and table in an image, but fail to grasp that the cup is sitting on the table.

Researchers from MIT, the MIT-IBM Watson AI Lab, and elsewhere have demonstrated a new technique that utilizes computer-generated data to help vision and language models overcome this shortcoming.

The researchers created a synthetic dataset of images that depict a wide range of scenarios, object arrangements, and human actions, coupled with detailed text descriptions. They used this annotated dataset to "fix" vision and language models so they can learn concepts more effectively. Their technique ensures these models can still make accurate predictions when they see real images.

When they tested models on concept understanding, the researchers found that their technique boosted accuracy by up to 10%. This could improve systems that automatically caption videos or enhance models that provide natural language answers to questions about images, with applications in fields like e-commerce or health care.

"With this work, we are going beyond nouns in the sense that we are going beyond just the names of objects to more of the semantic concept of an object and everything around it. Our idea was that, when a machine-learning model sees objects in many different arrangements, it will have a better idea of how arrangement matters in a scene," says Khaled Shehada, a graduate student in the Department of Electrical

Engineering and Computer Science and co-author of a paper on this technique.

Shehada wrote the paper with lead author Paola Cascante-Bonilla, a computer science graduate student at Rice University; Aude Oliva, director of strategic industry engagement at the MIT Schwarzman College of Computing, MIT director of the MIT-IBM Watson AI Lab, and a senior research scientist in the Computer Science and Artificial Intelligence Laboratory (CSAIL); senior author Leonid Karlinsky, a research staff member in the MIT-IBM Watson AI Lab; and others at MIT, the MIT-IBM Watson AI Lab, Georgia Tech, Rice University, École des Ponts, Weizmann Institute of Science, and IBM Research. The paper will be presented at the International Conference on Computer Vision held in Paris October 2–6.

Focusing on objects

Vision and language models typically learn to identify objects in a scene, and can end up ignoring object attributes, such as color and size, or positional relationships, such as which object is on top of another object.

This is due to the method with which these models are often trained, known as contrastive learning. This training method involves forcing a model to predict the correspondence between images and text. When comparing natural images, the objects in each scene tend to cause the most striking differences. (Perhaps one image shows a horse in a field while the second shows a sailboat on the water.)

"Every image could be uniquely defined by the objects in the image. So, when you do contrastive learning, just focusing on the nouns and objects would solve the problem. Why would the model do anything differently?" says Karlinsky.

The researchers sought to mitigate this problem by using synthetic data to fine-tune a vision and language model. The fine-tuning process involves tweaking a model that has already been trained to improve its performance on a specific task.

They used a computer to automatically create synthetic videos with diverse 3D environments and objects, such as furniture and luggage, and added human avatars that interacted with the objects.

Using individual frames of these videos, they generated nearly 800,000 photorealistic images, and then paired each with a detailed caption. The researchers developed a methodology for annotating every aspect of the image to capture object attributes, positional relationships, and human-object interactions clearly and consistently in dense captions.

Because the researchers created the images, they could control the appearance and position of objects, as well as the gender, clothing, poses, and actions of the human avatars.

"Synthetic data allows a lot of diversity. With real images, you might not have a lot of elephants in a room, but with synthetic data, you could actually have a pink elephant in a room with a human, if you want," Cascante-Bonilla says.

Synthetic data have other advantages, too. They are cheaper to generate than real data, yet the images are highly photorealistic. They also preserve privacy because no real humans are shown in the images. And, because data are produced automatically by a computer, they can be generated quickly in massive quantities.

By using different camera viewpoints, or slightly changing the positions or attributes of objects, the researchers created a dataset with a far wider variety of scenarios than one would find in a natural dataset.

Fine-tune, but don't forget

However, when one fine-tunes a model with synthetic data, there is a risk that model might "forget" what it learned when it was originally trained with real data.

The researchers employed a few techniques to prevent this problem, such as adjusting the synthetic data so colors, lighting, and shadows more closely match those found in natural images. They also made adjustments to the model's inner-workings after fine-tuning to further reduce any forgetfulness.

Their synthetic dataset and fine-tuning strategy improved the ability of popular vision and language models to accurately recognize concepts by up to 10%. At the same time, the models did not forget what they had already learned.

Now that they have shown how synthetic data can be used to solve this problem, the researchers want to identify ways to improve the visual quality and diversity of these data, as well as the underlying physics that makes synthetic scenes look realistic. In addition, they plan to test the limits of scalability, and investigate whether model improvement starts to plateau with larger and more diverse synthetic datasets.

More information: Going Beyond Nouns With Vision & Language Models Using Synthetic Data.
olivalab.mit.edu/Papers/going_beyond_nouns.pdf