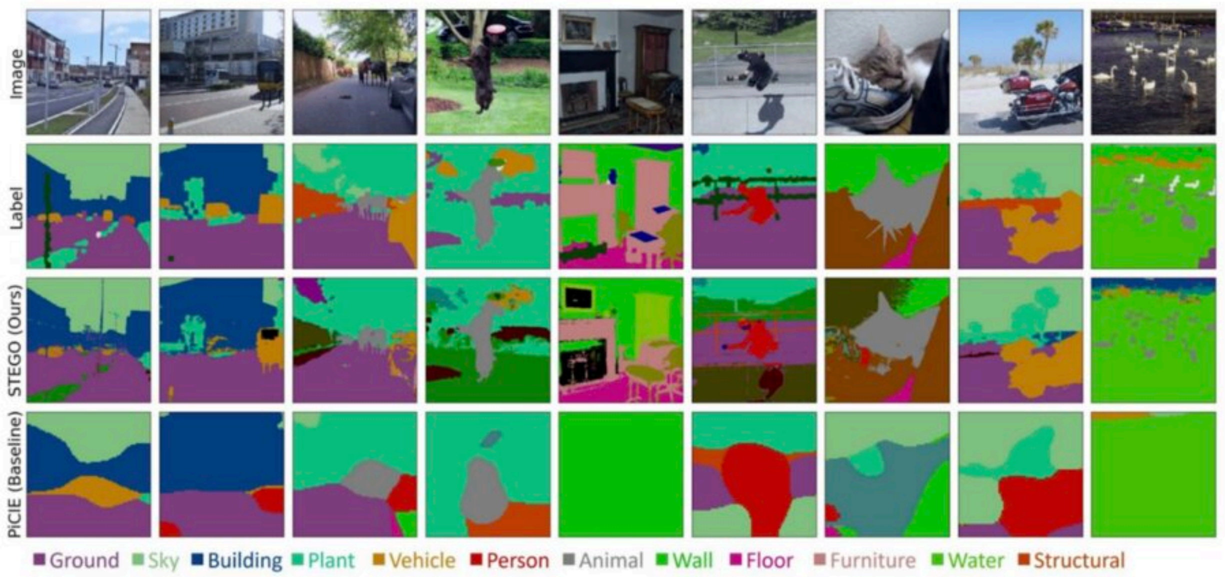
TechXplore

Scientists create algorithm to assign a label to every pixel in the world, without human supervision

April 21 2022, by Rachel Gordon



Unsupervised semantic segmentation predictions on the "CocoStuff 27" segmentation challenge. STEGO does not use labels to discover and segment consistent objects. Unlike prior algorithms, STEGO's predictions are consistent, detailed, and do not omit key objects. Credit: MIT CSAIL

Labeling data can be a chore. It's the main source of sustenance for computer-vision models; without it, they'd have a lot of difficulty identifying objects, people, and other important image characteristics.

Yet producing just an hour of tagged and labeled data can take a whopping 800 hours of human time. Our high-fidelity understanding of the world develops as machines can better perceive and interact with our surroundings. But they need more help.

Scientists from MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL), Microsoft, and Cornell University have attempted to solve this problem plaguing vision models by creating "STEGO," an algorithm that can jointly discover and segment objects without any human labels at all, down to the pixel.

STEGO learns something called "semantic segmentation"—fancy speak for the process of assigning a label to every pixel in an image. Semantic segmentation is an important skill for today's computer-vision systems because images can be cluttered with objects. Even more challenging is that these objects don't always fit into literal boxes; algorithms tend to work better for discrete "things" like people and cars as opposed to "stuff" like vegetation, sky, and mashed potatoes. A previous system might simply perceive a nuanced scene of a dog playing in the park as just a dog, but by assigning every pixel of the image a label, STEGO can break the image into its main ingredients: a dog, sky, grass, and its owner.

Assigning every single pixel of the world a label is ambitious—especially without any kind of feedback from humans. The majority of algorithms today get their knowledge from mounds of labeled data, which can take painstaking human-hours to source. Just imagine the excitement of labeling every pixel of 100,000 images. To discover these objects without a human's helpful guidance, STEGO looks for similar objects that appear throughout a dataset. It then associates these similar objects together to construct a consistent view of the world across all of the images it learns from.

Seeing the world

Machines that can "see" are crucial for a wide array of new and emerging technologies like self-driving cars and predictive modeling for medical diagnostics. Since STEGO can learn without labels, it can detect objects in many different domains, even those that humans don't yet understand fully.

"If you're looking at oncological scans, the surface of planets, or high-resolution biological images, it's hard to know what objects to look for without expert knowledge. In emerging domains, sometimes even human experts don't know what the right objects should be," says Mark Hamilton, a Ph.D. student in electrical engineering and computer science at MIT, research affiliate of MIT CSAIL, software engineer at Microsoft, and lead author on a new paper about STEGO. "In these types of situations where you want to design a method to operate at the boundaries of science, you can't rely on humans to figure it out before machines do."

STEGO was tested on a slew of visual domains spanning general images, driving images, and high-altitude aerial photographs. In each domain, STEGO was able to identify and segment relevant objects that were closely aligned with human judgments. STEGO's most diverse benchmark was the COCO-Stuff dataset, which is made up of diverse images from all over the world, from indoor scenes to people playing sports to trees and cows. In most cases, the previous state-of-the-art system could capture a low-resolution gist of a scene, but struggled on fine-grained details: A human was a blob, a motorcycle was captured as a person, and it couldn't recognize any geese. On the same scenes, STEGO doubled the performance of previous systems and discovered concepts like animals, buildings, people, furniture, and many others.

STEGO not only doubled the performance of prior systems on the

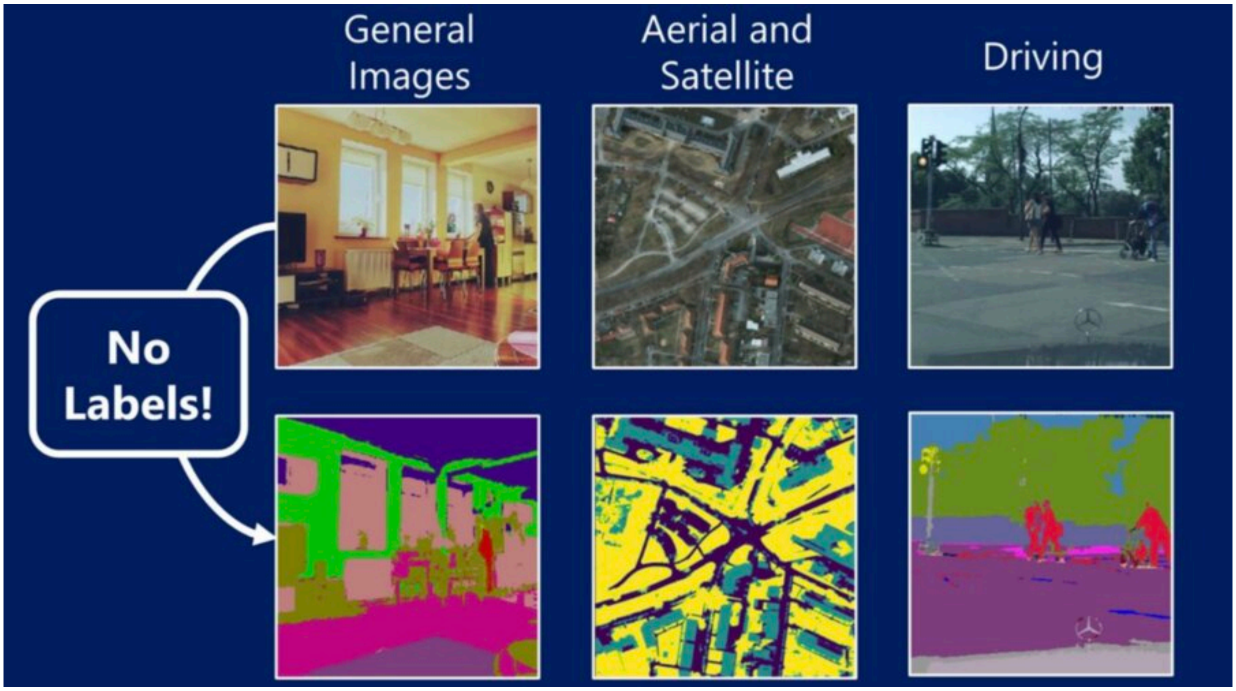
COCO-Stuff benchmark, but made similar leaps forward in other visual domains. When applied to driverless car datasets, STEGO successfully segmented out roads, people, and street signs with much higher resolution and granularity than previous systems. On images from space, the system broke down every single square foot of the surface of the Earth into roads, vegetation, and buildings.

Connecting the pixels

STEGO—which stands for "Self-supervised Transformer with Energy-based Graph Optimization"—builds on top of the DINO algorithm, which learned about the world through 14 million images from the ImageNet database. STEGO refines the DINO backbone through a learning process that mimics our own way of stitching together pieces of the world to make meaning.

For example, you might consider two images of dogs walking in the park. Even though they're different dogs, with different owners, in different parks, STEGO can tell (without humans) how each scene's objects relate to each other. The authors even probe STEGO's mind to see how each little, brown, furry thing in the images are similar, and likewise with other shared objects like grass and people. By connecting objects across images, STEGO builds a consistent view of the word.

"The idea is that these types of algorithms can find consistent groupings in a largely automated fashion so we don't have to do that ourselves," says Hamilton. "It might have taken years to understand complex visual datasets like biological imagery, but if we can avoid spending 1,000 hours combing through data and labeling it, we can find and discover new information that we might have missed. We hope this will help us understand the visual word in a more empirically grounded way."



With the STEGP algorithm, research scientists attempted to solve a massive labeling problem plaguing vision models. STEGO can jointly discover and segment objects without any human labels at all, down to the pixel. Credit: MIT CSAIL.

Looking ahead

Despite its improvements, STEGO still faces certain challenges. One is that labels can be arbitrary. For example, the labels of the COCO-Stuff dataset distinguish between "food-things" like bananas and chicken wings, and "food-stuff" like grits and pasta. STEGO doesn't see much of a distinction there. In other cases, STEGO was confused by odd images—like one of a banana sitting on a phone receiver—where the receiver was labeled "foodstuff," instead of "raw material."

For upcoming work, they're planning to explore giving STEGO a bit more flexibility than just labeling pixels into a fixed number of classes as things in the real world can sometimes be multiple things at the same time (like "food," "plant" and "fruit"). The authors hope this will give the algorithm room for uncertainty, trade-offs, and more abstract thinking.

"In making a general tool for understanding potentially complicated datasets, we hope that this type of an algorithm can automate the scientific process of object discovery from images. There's a lot of different domains where human labeling would be prohibitively expensive, or humans simply don't even know the specific structure, like in certain biological and astrophysical domains. We hope that future work enables application to a very broad scope of datasets. Since you don't need any human labels, we can now start to apply ML tools more broadly," says Hamilton.

"STEGO is simple, elegant, and very effective. I consider unsupervised segmentation to be a benchmark for progress in image understanding, and a very difficult problem. The research community has made terrific progress in unsupervised image understanding with the adoption of transformer architectures," says Andrea Vedaldi, professor of computer vision and machine learning and a co-lead of the Visual Geometry Group at the engineering science department of the University of Oxford. "This research provides perhaps the most direct and effective demonstration of this progress on unsupervised segmentation."

Hamilton wrote the paper alongside MIT CSAIL Ph.D. student Zhoutong Zhang, Assistant Professor Bharath Hariharan of Cornell University, Associate Professor Noah Snavely of Cornell Tech, and MIT professor William T. Freeman. They will present the paper at the 2022 International Conference on Learning Representations (ICLR).

More information: Mark Hamilton et al, Unsupervised semantic segmentation by Distilling Feature Correspondences (2022) is available as a PDF at marhamilresearch4.blob.core.wi … blic/stego_paper.pdf

Provided by Massachusetts Institute of Technology

Citation: Scientists create algorithm to assign a label to every pixel in the world, without human supervision (2022, April 21) retrieved 2 May 2024 from https://techxplore.com/news/2022-04-scientists-algorithm-assign-pixel-world.html