Computer Vision

DINOv2: State-of-the-art computer vision models with self-supervised learning

April 17, 2023

DINOv2 is able to take a video and generate a higher-quality segmentation than the original DINO method. DINOv2 allows remarkable properties to emerge, such as a robust understanding of object parts, and robust semantic and low-level understanding of images.

- Meta AI has built DINOv2, a new method for training high-performance computer vision models.
- DINOv2 delivers strong performance and does not require fine-tuning. This makes it suitable for use as a backbone for many different computer vision tasks.
- Because it uses self-supervision, DINOv2 can learn from any collection of images. It can also learn features, such as depth estimation, that the current standard approach cannot.
- We are open-sourcing our model and sharing an interactive demo.

Today, we are open-sourcing DINOv2, the first method for training computer vision models that uses self-supervised learning to achieve results that match or surpass the standard approach used in the field.

Self-supervised learning — the same method that's used to create cutting-edge large language models for text applications — is a powerful, flexible way to train AI models because it does not require large amounts of labeled data. Like with other self-supervised systems, models using the DINOv2 method can be trained on any collection of images, without needing any associated metadata. Think of it as

Unlike many recent reconstruction-based self-supervised learning methods, our model requires no fine-tuning. DINOv2 provides high-performance features that can be directly used as inputs for simple linear classifiers. This flexibility means DINOv2 can be used to create multipurpose backbones for many different computer vision tasks. Our measurements show very strong prediction capabilities on tasks such as classification, segmentation, and image retrieval. Surprisingly, on depth estimation, our features significantly outperform specialized state-of-the-art pipelines evaluated both in-domain and out-of-domain. We believe that this strong out-of-domain performance is due to the combination of self-supervised feature learning and the use of lightweight task-specific modules, such as linear classifiers. Finally, because we don't resort to fine-tuning, the backbone remains general and the same features can be used simultaneously on many different tasks.

Self-supervised computer vision models like DINOv2 will be useful in a wide variety of applications. Meta collaborated with the World Resources Institute to use AI to map forests, tree by tree, across areas the size of continents. Our self-supervised model was trained on data from forests in North America, but evaluations confirm that it generalizes well and delivers accurate maps in other locations around the world.

DINOv2 complements our other recent computer vision research, including Segment Anything. Segment Anything is a promptable segmentation system focused on zero-shot generalization to diverse set of segmentation tasks. DINOv2 combines with simple linear classifiers to achieve strong results across multiple tasks beyond the segmentation sub-field, creating horizontal impact.

Overcoming the limitations of image-text pretraining

In recent years, a different technique, known as image-text pretraining, has been the standard approach for many computer vision tasks. But because the method relies on handwritten captions to learn the semantic content of an image, it ignores important information that typically isn't explicitly mentioned in those text descriptions. For instance, a caption of a picture of a chair in a vast purple room might read "single oak chair." Yet, the caption misses important information about the background, such as where the chair is spatially located in the purple room. Because of that, we believe caption-based features lack a proper understanding of

learning, we avoid this problem by not relying on text descriptions. This, in turn, coupled with strong execution, allows DINOv2 to provide state-of-the-art results for monocular depth estimation. For context, monocular depth estimation is a task where the goal is to predict which objects are in the foreground and which are in the background.

In general, the need for human annotations of images is a bottleneck because it limits how much data you can use to train a model. In specialized application domains, images are hard or even impossible to label. Training machine learning models on labeled cellular imaging, for instance, is challenging, as there are a limited number of experts who can annotate the cells, and certainly not at the scale required. Self-supervised training on microscopic cellular imagery, however, opens up the way for foundational cell imagery models and, consequently, biological discovery, as it becomes possible to compare known treatments with new ones, for example. The same story holds for the estimation of animal density and abundance, allowing the identification of sources of biodiversity decline and the effectiveness of conservation efforts. Both examples are based on the original open source DINO algorithm, and we hope DINOv2 can improve such lines of work. DINOv2's training stability and scalability will fuel further advances in applicative domains. One application already underway is our forest-mapping collaboration with the World Resources Institute noted above.

Our release comes at a time when the performance of joint embedding models that train features by matching data augmentations is plateauing. Specifically, the evaluation performance on ImageNet had moved by 10 percent between 2019 and 2021, and not much since then (+1 percent since 2021). The community focused more on developing alternatives, such as masked-image modeling, limiting progress in that field. In addition, the DINO class of models, among other SSL methods, was difficult to train outside of the classical scope of ImageNet, limiting their adoption for research.

Making progress from DINO to DINOv2 required overcoming several challenges: creating a large and curated training dataset, improving the training algorithm and implementation, and designing a functional distillation pipeline.

Building a large, curated, and diverse dataset to train the models

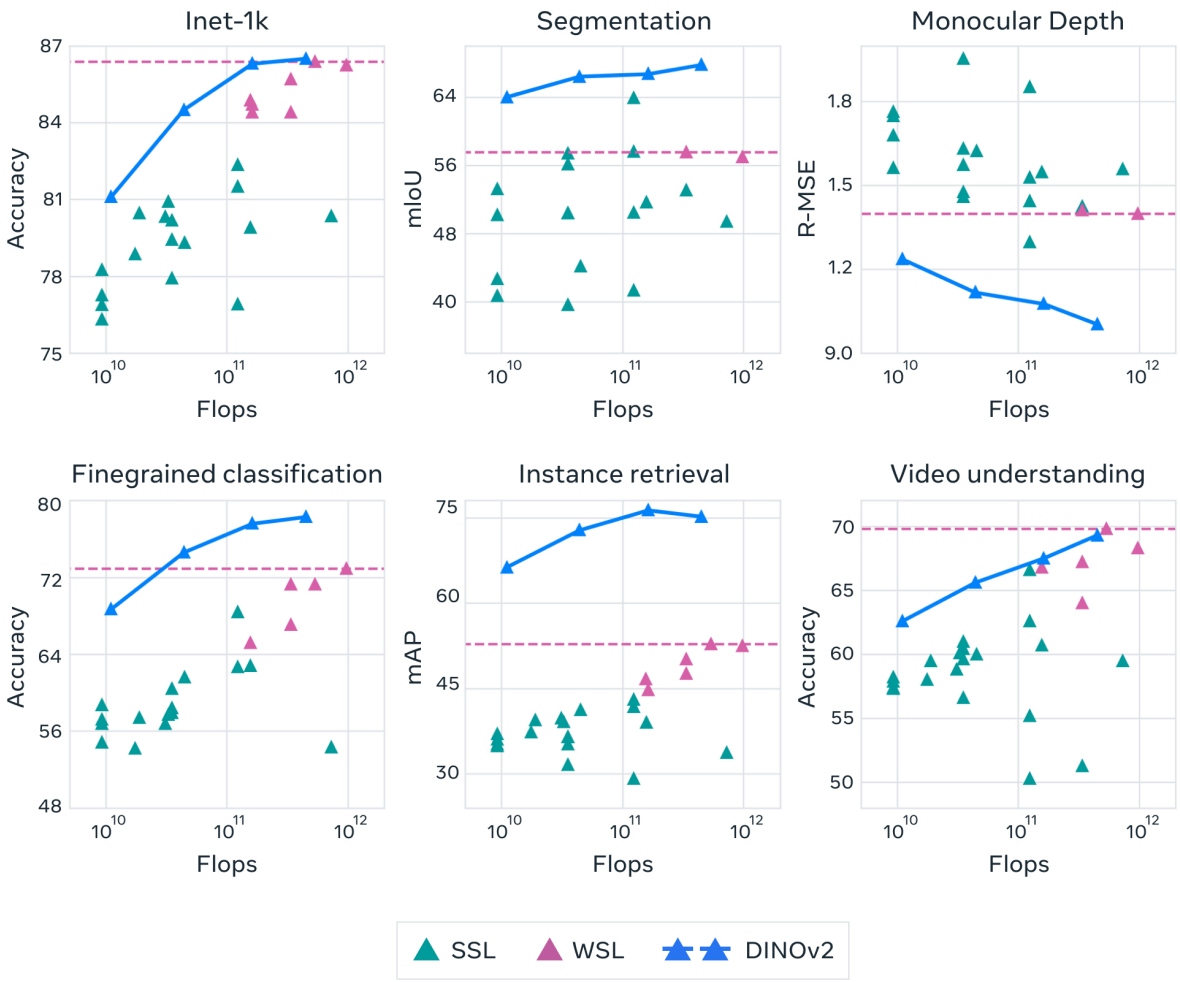
accessing more data is not always possible. With no sufficiently large curated dataset available to suit our needs, we looked into leveraging a publicly available repository of crawled web data and built a pipeline to select useful data inspired by LASER. Two key ingredients are required for building a large-scale pretraining dataset from such a source: discarding irrelevant images and balancing the dataset across concepts. Such delicate curation can't realistically be done manually, and we wanted a method that allowed capturing distributions not easily associated with metadata. This was achieved by curating a set of seed images from a collection of about 25 third-party datasets and extending it by retrieving images sufficiently close to those seed images. This approach enabled us to produce a pretraining dataset totaling 142 million images out of the 1.2 billion source images.

Algorithmic and technical improvements

With more training data, larger models perform better than smaller ones, but their training poses two major challenges. First, increasing the model size makes the training more challenging because of potential instability. In DINOv2, we included additional regularization methods inspired by the similarity search and classification literature, making the training algorithm much more stable. Second, in order to remain tractable, larger models require more efficient implementations. The DINOv2 training code integrates the latest mixed-precision and distributed training implementations proposed in the cutting-edge PyTorch 2 (fully sharded data parallel), an efficient implementation of the stochastic depth technique, as well as the latest compute algorithm implementations of xFormers (in particular, variable-length memory-efficient attention). This allows faster and more efficient iteration cycles. Overall, with equivalent hardware, our code runs around twice as fast with only a third of the memory usage, allowing scaling in data, model size, and hardware.

Strong, lightweight models with distillation

Running inference for larger models requires more powerful hardware, potentially limiting many practical use cases. To circumvent this problem, researchers typically resort to *model distillation*, to compress the knowledge of a large model into a smaller one. Our training algorithm is based on self-distillation, making it straightforward to compress our large models into smaller ones. This procedure allows us to compress our highest-performance architecture into significantly smaller ones at only a minimal cost in accuracy, for a dramatically decreased



The DINOv2 family of models drastically improves over the previous state of the art in self-supervised learning (SSL), and reaches performance comparable with weakly-supervised features (WSL).

Releasing a family of high-performance pretrained models

We release DINOv2 pretrained models to the community with a matching stable, accurate, and scaled implementation: We share pretraining code and recipe for ViT-L/16 (300 M params) and ViT-g/14 (1.1 B params) architectures, as well as checkpoints for a range of pretrained models from the larger ViT-g/14 down to

as CLIP and OpenCLIP on a wide array of tasks, some of which are illustrated in our demo. Don't hesitate to play with it! Our features can be used out of the box for nearest neighbor classification or paired with linear classification, yielding strong performance. DINOv2 allows skipping the model adaptation phase (fine-tuning) — our linear evaluation performance is close to their fine-tuned counterpart (within 2 percent on ImageNet-1k) .

Our features can be used out of the box for nearest neighbor classification or paired with linear classification, yielding strong performance. DINOv2 allows skipping the model adaptation phase (fine-tuning) — our linear evaluation performance is close to their fine-tuned counterpart (within 2 percent on ImageNet-1k) .

Going forward, the team plans to integrate this model, which can function as a building block, in a larger, more complex AI system that could interact with large language models. A visual backbone providing rich information on images will allow complex AI systems to reason on images in a deeper way than describing them with a single text sentence. Models trained with text supervision are ultimately limited by the image captions. With DINOv2, there is no such built-in limitation.

Written By

The DINOv2 team

Research Areas

Computer Vision

Read the paper →

Get the code

Explore the demo →

Research

Meta AI

Latest news

Foundational models