DINO & SAM

Self-supervised learning and Transformers

Un-supervised vs Self-supervised learning

Aspect	Unsupervised Learning	Self-Supervised Learning
Labels	No labels at all	Still no human labels, but generates *pseudo-labels* from data
Goal	Discover structure in data	Learn representations useful for downstream tasks
Examples	Clustering, PCA, autoencoders	Contrastive learning, masked autoencoders, DINO, SimCLR
Tasks	Find patterns, groups, density	Solve a *pretext task* designed by us (e.g., match image views)

Self-supervised learning with ViTs

1. A self-supervised system is a type of machine learning system that learns from unlabeled data by generating its own labels from the data itself.

DINO

1. Our model can discover and segment objects in an image or a video with absolutely no supervision
2. Segmenting objects helps facilitate tasks ranging from swapping out the background of a video chat to teaching robots that navigate through a cluttered environment.
3. It is considered one of the hardest challenges in computer vision because it requires that AI truly understand what is in an image.

DINO

1. This is traditionally done with supervised learning and requires large volumes of annotated examples.
2. But our work with DINO shows highly accurate segmentation may actually be solvable with nothing more than self-supervised learning and a suitable architecture.
3. By using self-supervised learning with Transformers, DINO opens the door to building machines that understand images and video much more deeply.

DINO

1. Training ViT with our DINO algorithm, we observe that our model automatically learns an interpretable representation and separates the main object from the background clutter.
2. It learns to segment objects without any human-generated annotation
3. When visualizing the local attention maps in the network, we see that they correspond to coherent semantic regions in the image.

DINO

1. DINO works by interpreting self-supervision as a special case of self-distillation, where no labels are used at all.
2. Indeed, we train a student network by simply matching the output of a teacher network over different views of the same image.
3. We identified two components from previous self-supervised approaches that are particularly important for strong performance on ViT,
    a. the momentum teacher and
    b. multicrop training

Step-by-step: DINO + k-NN for Classification

1. Pretrain a model using DINO

- Train a ViT or ResNet using DINO on **unlabeled** data.
- The model learns to extract **rich embeddings** from images.
- No labels are involved during this phase.

2. Build an embedding database (train set)

- For every image in your labeled training dataset (even a small one like CIFAR-10):
  - Pass it through the pretrained DINO model.
  - Extract the **embedding** (typically from the CLS token for ViT or penultimate layer for CNNs).
  - Store the embedding along with its **label**.

3. Classify a new image using k-NN

- For a new (test) image:
  - Get its embedding using the same DINO model.
  - Compute its **cosine similarity** (or Euclidean distance) to all embeddings in the database.
  - Pick the **top-k most similar** embeddings.
  - Predict the label using **majority vote** (or weighted vote based on similarity).

Image retrieval works in the similar fashion

SAM

1. SAM: A generalized approach to segmentation
2. Previously two approaches
    a. interactive segmentation, allowed for segmenting any class of object but required a person to guide the method by iteratively refining a mask.
    b. automatic segmentation, allowed for segmentation of specific object categories defined ahead of time (e.g., cats or chairs) but required substantial amounts of manually annotated objects to train
3. With SAM, practitioners will no longer need to collect their own segmentation data and fine-tune a model for their use case.

References

1. https://ai.meta.com/blog/dino-paws-computer-vision-with-self-supervised-transformers-and-10x-more-efficient-training/
2. https://arxiv.org/abs/2104.14294
3. https://ai.meta.com/blog/segment-anything-foundation-model-image-segmentation/
4. https://arxiv.org/pdf/2304.02643