DINOv2 필기

목표: Pretrained model, without finetuning & supervision

결과: ViT model with 1B parameters

Requirement: Generate visual features:

1. Image lever for classification
2. Pixel level for segmentation
3. 선행연구 및 비교:
4. Text-guide pretraining

Shortcoming:

1. limit the info that can be retained, ignore the pixel-lever data
2. Can only from raw data since text issue.
3. Self-supervise learning

Advantage: likely to pre-text task

Shortcoming: pretrained with ImageNet-1k, require curated dataset

1. Pretrain on a large quantity of curated data

Advantage: 2× faster and require 3× less memory than similar

discriminative self-supervised methods

Work on pretraining data:

similar pipeline like used in NLP, used data similarity

Rebalance concepts and avoid overfitting on a few dominant modes: clustering

Conclude: Self-supervised pretraining alone is a good candidate for learning transferable frozen features that are competitive with the best openly available weakly-supervised models.

1. 관련 작업:
2. Intra-image self-supervised training

A first family of self-supervised methods focuses on pretext tasks built from the image, i.e., extracting a signal from the image to be predicted from the rest of the image.

1. Discriminative self-supervised learning.

The second line of work, closer to ours, is using discriminative signals between images or groups of images to learn features, hard to scale to larger model sizes

1. Scaling self-supervised pretraining.

Due to the poor quality of the pretraining data, need to fine tuning the feature.

1. Automatic data curation.

inspired by text curation pipelines

1. Data 처리

Assemble curated LVD-142M dataset by retrieving, from a large pool of uncurated data, images that are close to those in several curated datasets.

Post-process the downloaded images (PCA hash deduplication, NSFW filtering, and blurring identifiable faces). This results in 1.2B unique images.

Deduplication.

Self-supervised image retrieval

1. 변별형 Self-supervised Pre-training:

discriminative self-supervised method that can be seen as a combination of DINO and iBOT losses with the centering of SwAV

1. Image-level objective

Minimize the difference between the student and teacher

Cross entropy loss, use student net and teacher net, improve student net

1. Patch-level objective
2. **Untying** head weights between both objectives.

선행연구와 반대

1. Sinkhorn-Knopp centering

최적 전송 문제

1. KoLeo regularize

鼓励批次内的样本在特征空间中均匀分布，增大样本之间的最小距离，提高模型的泛化能力

1. Adapting the resolution

Increasing image resolution is key to pixel-level downstream tasks such as segmentation or detection

training at high resolution is time and memory demanding

increase the resolution of images to 518×518 during a short period at the end of pretraining.

DINO 방법: student 네트워크와 teacher 네트워크 간의 교차 엔트로피 손실을 학습하여 student 네트워크의 이미지 **분류 성능**을 향상시킵니다.

iBOT 방법: 입력 패치를 random로 차단하여 학생들이 네트워크에서 차단된 부분의 특성을 학습하여 **robust** 성을 향상시킵니다.

1. 효율적 구현
2. Fast and memory-efficient attention

Use an embedding dimension of 1536 with 24 heads (64 dim/head), rather than 1408 with 16 heads (88 dim/head)

1. Sequence packing.

reproducing the output of a large model with a smaller model by minimizing some distance between both outputs for a set of given inputs. Since our objective function is a form of distillation from the teacher network to the student network, we

1. Efficient stochastic depth

high drop rates (d = 40% in this work),

1. Fully-Sharded Data Parallel (FSDP).

the model size is not bounded by the memory of a single GPU but by the total sum of GPU memory across compute nodes

1. Model distillation.

For smaller models, distill them from largest model, the ViT-g, instead of training them from scratch.

1. 삭마 연구 Aablation Studies

**Pretext Task:** Pretext tasks are pre-designed tasks for networks to solve, and visual features are learned by learning objective functions of pretext tasks.

frozen features: 이런 경우 데이터의 특징을 추출(피처 추출)하는 부분의 변수는 동결하고(freeze), 분류 파트에 해당되는 fully connected layer의 변수만 업데이트 할 수 있음. 이런 방법을 모델 동결(model freezing)이라고 함.