DINO v2

A Preprint

The idea is to make task-agnostic pretrained representations. Authors also build a NLP-inspired automatic pipeline to filter and rebalance datasets from an extensive collection of uncurated images(data similarities are used instead of metadata).

Whole data pipeline is described the next way:

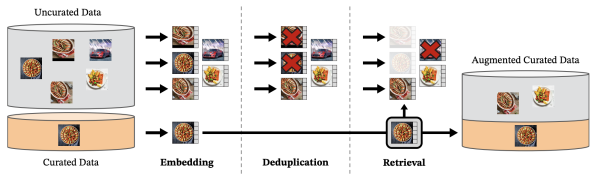

Figure 3: **Overview of our data processing pipeline.** Images from curated and uncurated data sources are first mapped to embeddings. Uncurated images are then deduplicated before being matched to curated images. The resulting combination augments the initial dataset through a self-supervised retrieval system.

The whole idea is to use some sort of curated data as data with good quality and use some uncurated data where we have a lot of noise of just no filter. Model basically chooses samples herself. Overall amount of images is 1.2B!!! For features a combination of DINO and iBOT losses with the centering of SwAV. It is mentioned that they randomly mask some tokens that are passed to student, but no to the teacher. After that IBOT head is applied to the visible teacher tokens(by visible they mean tokens corresponding to the ones masked in the student)

$$\mathcal{L}_{iBOT} = -\sum p_{ti} \log p_{si}$$

where i are patch indices for masked tokens. The whole idea is for teacher to predict for ones student do not see and for student is to predict context(masked ones) based on the part that is visible. This way is simulates NLP methods.