Masked Knowledge Distillation – Analyzing Self-Supervised ViT Frameworks

Author: Cristopher McIntyre Garcia
SN: 300025114
Email: cmcin019@uottawa.ca

Topics

Self-Supervised Learning, Computer Vision, Transformers, Multi-teacher Self-Distillation, Masked Image Modelling, Momentum/Moving-Averaged Encoders, Masked Autoencoders

Proposal

Self-Supervised Learning (SSL) has been highly successful in the field of natural language processing. As an example, [37], an autoregressive language model, learns language representations through a masked modelling task; by randomly masking tokens within text and then recovering the information, this model has quickly become SoTA. Researchers have proposed many different approaches for introducing the notion of SSL to the field of Computer Vision. [34] was one of the first to introduce contrastive learning and momentum encoders, which to this day still perform fairly well, as seen in [17]. By matching two encoded augmented views of an image to one another, through ***Contrastive Loss*** [36], they are able to train an encoder to create visual representations of images. More specifically, a key encoder learns to encode image representations by matching encoded keys to encoded queries of a momentum encoder dictionary, the latter being driven by momentum update or ***Exponential Moving Average (EMA)***. [32] has shown that the type of image augmentation chosen heavily impacts the performance of the model. Also demonstrated was that using a non-linear transformation between representations and contrastive loss improve the quality of the representations, and also using large batches sizes and longer training. As one may think, these methods need careful treatment of negative pairs, and rely heavily on augmentation. In [33], negative pairs are completely omitted, and instead iteratively bootstrapping the outputs of the encoder to serve as targets for an enhanced representation is opted. By doing so, they improve robustness against different augmentation processes and lower the necessity of using large batch sizes. The way the target network is trained is through an EMA, and a projection head is used. [31] introduces the ***Vision Transformer (ViT)*** to this approach, and presents it as a ***Self-Distillation*** method. They are able to create segmentation masks for images without using predefined ground truth labels – something that has been difficult to do, even with Convolution Neural Networks. Very recently, [15] proposes a ***Masked Image Modelling (MIM)*** task to pre-train ViTs, very reminiscent of the pre-training task of [37]. In this method, random portions of an image are masked and fed into a backbone ViT. The pre-training task aims to reconstruct the visual information that was masked from the original image. [14] combines the ideas of masking image models from [15] and self-distillation from [31], to creates a new current SoTA that is able to outperform certain SoTA supervised methods. Up until now, the models presented have used an architecture with two or fewer networks. [22] has shown the potential of using ***Multiple Teacher Networks in Knowledge Distillation***. Yet no papers have discussed this possibility in SSL. [1] uses a similar idea, however instead of using multiple teachers, they use multi-stage distillation by freezing the teacher network and re-initializing the student network. We theorize that incorporating the idea of a multi-teacher multi-level framework will aid the model better generalize and speed up training. Our goal is not to obtain better results than the state of the art, instead we aim to explore the properties of current self-supervised models. To do this, we will analyze several ***Model Configuration*** that use different SSL techniques as discussed above. Notably, we are looking into comparing these methods, and introducing the idea of multiple teachers in self-distillation. Because we are limited in our computational power, we will be using smaller datasets, and comparing the results found between our configurations only. We will use a subset of the ***ImageNet Dataset***, because it is widely available, and used in every SSL paper above. We will perform evaluation at different stages of training by taking the backbone architecture and using it as a ***Linear Classifier***. The research questions that we wish to answer are as follows; What is MIM's impact on representation in a self-distillation framework? Will introducing multiple teacher networks aid in the pre-training of a backbone vision transformer? How does momentum and contrastive learning hold up against these other approaches in short term training? We will use tools such as ***PyTorch*** and ***VISSL*** that are ML libraries, the latter being purely for SoTA SSL research. Training will be done on my computer on a NVIDIA 3070 ti. We hope to uncover some of the mysteries surrounding SSL, and provide clarity as to why these approaches seem to work.