Analyzing Masked Relational Knowledge Distillation For Self-Supervised Learning

Background

Self-Supervised Learning (SSL) has been highly successful in the field of natural language processing. BERT from [9], an autoregressive language model, learns language representations for downstream tasks through masked modeling. 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 (CV). In the past two years, we have seen the emergence of approaches such as ***Contrastive Learning (CL)***, ***Self-Distillation Learning (SDL)***, and ***Masked Image Modeling (MIM)***.

Problem

Collecting labelled data is expensive and time-consuming. The emergence of SSL in the field of CV has, to some extent, mitigated this issue. However, because the field is still in its infancy, the approaches and architectures used are not yet well understood. In this work, we aim to better understand SSL in CV by analyzing different architecture configurations. We also aim to introduce the notion of ***Relational Knowledge Distillation (RKD)***, which has been shown to perform better than regular Knowledge Distillation (KD), as a novel SDL approach.

Literature

[19] was one of the first to introduce CL, which to this day still performs fairly well, as seen in [8]. By matching an encoded augmented image with many other momentum-encoded augmented images, and through CL[16], they are able to train an encoder to create visual representations. [6] has shown that the type of image augmentation chosen heavily impacts the performance of the model, and using a non-linear transformation between representations and contrastive loss improves the quality of the representations. These methods need careful treatment of negative pairs, mandate the use of large batch sizes, and rely heavily on augmentation.

In [15], negative pairs are completely omitted. Instead, they iteratively bootstrap the weights of an encoder to serve as the weights of a target network. By doing so, they improve performance robustness to different image augmentation and lower the necessity of using large batch sizes. [5] incorporates the ***Vision Transformer (ViT)*** [10] into this approach, and presents it as SDL. They are able to create segmentation masks from images without using predefined ground truth labels, something that has been difficult to do, even with CNNs. Many variations of this model have been studied, such as [11], [22] and [32].

Very recently, [3] proposes a MIM task to pre-train ViTs, very reminiscent of the pre-training task of [9]. 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. [37] combines the ideas of masking image models from [3] and SDL from [5] to create a new current SoTA that is able to outperform certain SoTA supervised methods. Many papers have incorporated MIM, such as [18], [31] and [35]

Though KD as the proposed concept for SDL has been shown to work well, [27] demonstrated that it is outperformed by RKD in generalizing. That said, no work has been done on incorporating this concept into SSL. [24] use a multi-stage distillation framework, by freezing the teacher and re-initializing the student after every step. This is different from RKD because they still perform instance-to-instance KD. We theorize that incorporating RKD may help improve the results of current SSL frameworks.

Proposal

Our goal is not to obtain better results than the state of the art, instead we aim to explore the properties of current SSL models, mainly those that use use SDL and MIM. We propose analyzing the behaviour of several architecture configurations. We also propose to use RKD as a new form of SDL, and believe that it will help models better generalize. We will train and evaluate our models on the ImageNet Dataset, as most SSL papers have done. Evaluation will consist of taking the trained backbone architecture and treating it as a linear classifier or a KNN classifier. We will use tools such as ***PyTorch***, ***Hugging Face*** and ***VISSL***, the latter being a library meant purely for SoTA SSL research. We hope to uncover some of the mysteries surrounding SSL and present new ideas.