Addis Ababa University

Self supervised learning in computer vision

Prepared by Tsion Meride
June, 2024

Introduction

Contrastive self-supervised learning has emerged as a powerful paradigm in deep learning, particularly for computer vision tasks. Unlike supervised learning, which requires labeled data, self-supervised learning leverages the inherent structure and patterns within unlabeled data to learn rich and generalizable representations.

Traditional supervised learning approaches rely heavily on the amount of annotated training data available. Even though there is a plethora of data available, the lack of annotations has pushed researchers to find alternative approaches that can leverage them.

This is where self-supervised methods play a vital role in fueling the progress of deep learning without the need for expensive annotations and learn feature representations where data provide supervision.

Pretext Tasks

Pretext tasks are self-supervised tasks that act as an important strategy to learn representations of the data using pseudolabels. These pseudolabels are generated automatically based on the attributes found in the data. The learned model from the pretext task can be used for any downstream tasks such as classification, segmentation, detection, etc. in computer vision.

pretext tasks can be designed for any kind of data such as image, video, speech, signals, and so on. For a pretext task in contrastive learning, the original image acts as an anchor, its augmented(transformed) version acts as a positive sample, and the rest of the images in the batch or in the training data act as negative samples.

Most of the commonly used pretext tasks are divided into four main categories.

Color Transformation

Color transformation involves basic adjustments of color levels in an image such as blurring, color distortions, converting to grayscale, During this pretext task, the network learns to recognize similar images invariant to their colors.

. *Geometric Transformation*

A geometric transformation is a spatial transformation where the geometry of the image is modified without altering its actual pixel information. The transformations include scaling, random cropping, flipping (horizontally, vertically). the original image is **considered as the global view and the transformed version is considered as the local view**.

Context-Based

Jigsaw Puzzle

Traditionally, solving jigsaw puzzles has been a prominent task in learning features from an image in an unsupervised way. It involves identifying the correct position of the scrambled patches in an image by training an encoder

2. Frame Order Based

This approach applies to data that extends through time. An ideal application would be in the case of sensor data or a sequence of image frames (video). A video contains a sequence of semantically related frames. This implies that frames that are nearby with respect to time are closely related and the ones that are far away are less likely to be related.

1. Contrastive Losses: Contrastive learning objectives, such as InfoNCE and SimCLR, encourage the model to learn representations that maximize the similarity between positive (similar) pairs of data samples and minimize the similarity between negative (dissimilar) pairs. This contrastive approach has been shown to be highly effective in learning robust and transferable features.
2. Data Augmentation: Contrastive learning models often rely on extensive data augmentation techniques, such as random cropping, color jittering, and rotation, to create positive and negative pairs of samples. These augmentations help the model learn invariances to various transformations, leading to better generalization.

1. Momentum Encoders: Some contrastive learning approaches, such as MoCo and SimSiam, utilize momentum-based encoders to maintain a slowly-evolving representation of the positive samples, which can help stabilize the training process and improve the quality of the learned representations.
2. Unsupervised Pre-training: Contrastive self-supervised models are often pre-trained on large-scale unlabeled datasets, such as ImageNet, and then fine-tuned on specific downstream tasks. This pre-training approach has been shown to significantly improve performance compared to training from scratch, especially when labeled data is scarce.

Applications and Impact

Contrastive self-supervised learning has had a profound impact on the field of computer vision, leading to state-of-the-art results across a wide range of tasks, including:

- Image classification: Contrastive models like SimCLR and BYOL have achieved impressive performance on standard benchmarks, often matching or even surpassing their supervised counterparts.
- Object detection and segmentation: Contrastive pre-training has been shown to improve the performance of downstream detection and segmentation models, leveraging the learned visual representations.

- Medical image analysis: Contrastive learning has been successfully applied to medical imaging tasks, such as disease diagnosis and organ segmentation, where labeled data is often scarce.
- Robustness and out-of-distribution generalization: Contrastive models have demonstrated improved robustness to distribution shifts and enhanced out-of-distribution generalization capabilities.

Future works and Challenges

- Scalability and efficiency: Developing more efficient contrastive learning algorithms and architectures to handle large-scale datasets and enable real-world deployment.
- Interpretability and explainability: Understanding the inner workings of contrastive models and the reasons for their superior performance.
- Domain-specific adaptations: Exploring how contrastive learning can be tailored to specific domains, such as medical imaging or autonomous driving, to maximize its impact.
- Incorporating structured knowledge: Investigating ways to integrate contrastive learning with other forms of prior knowledge, such as graph-based representations, to further enhance performance.

Conclusion

Contrastive self-supervised learning has emerged as a transformative approach in computer vision, pushing the boundaries of what is possible with deep learning. By leveraging the inherent structure in unlabeled data, contrastive models have demonstrated impressive performance across a wide range of computer vision tasks, paving the way for more efficient and robust visual understanding systems.