Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, Nicolas Ballas

Meta AI (FAIR). McGill University. Mila, Quebec AI Institute. New York University

Conference on Computer Vision and Pattern Recognition. June 2023

Summary

The paper discusses self-supervised learning in computer vision and compares two common approaches: invariance-based methods and generative methods. Invariance-based methods train an encoder to produce similar embeddings for different views of the same image, using hand-crafted data augmentations. These methods produce high-level semantic representations but introduce biases that may not generalize well to different tasks or data types. Generative methods, inspired by cognitive learning theories, focus on predicting corrupted or missing parts of the input. They are more versatile across modalities but tend to yield lower-level semantic representations. Fine-tuning is often required to maximize their utility.

The paper introduces a new approach called I-JEPA (Joint-Embedding Predictive Architecture for Images) to improve the semantic level of self-supervised representations without additional image transformations. I-JEPA predicts missing information in an abstract representation space, using a target-encoder network. It differs from generative methods by targeting abstract prediction goals, potentially reducing pixel-level details and encouraging the learning of more semantic features. The article also highlights the importance of predicting sufficiently large target blocks using a multi-block masking strategy.

The empirical evaluation shows that I-JEPA achieves strong off-the-shelf representations without hand-crafted view augmentations, outperforming pixel-reconstruction methods on various tasks. It competes well with view-invariant pretraining approaches on semantic tasks and performs better on low-level vision tasks like object counting and depth prediction. I-JEPA is scalable and efficient, requiring less GPU hours for pretraining compared to other methods, making it a promising approach for self-supervised learning in computer vision.

Why Is This Paper Interesting?

Advancement in Self-Supervised Learning: Self-supervised learning is a prominent topic in computer vision, offering a solution to reduce reliance on extensive labeled datasets. This paper introduces a novel approach (I-JEPA) that improves the semantic level of self-supervised representations without additional image transformations. It presents a valuable alternative to existing methods.

Performance Across Various Tasks: I-JEPA not only competes effectively with view-invariant pretraining approaches on semantic tasks but also surpasses them in low-level vision tasks such as object counting and depth prediction. This versatility renders it applicable to a broader spectrum of computer vision tasks, enabling AI engineers to harness its potential across a wide range of projects.

Scalability and Efficiency: The paper underscores the scalability and efficiency of I-JEPA, which are crucial considerations for real-world applications. AI engineers frequently contend with resource limitations, and a method that demands fewer computational resources can offer a significant advantage.