On the Implicit Bias of Predicting in Latent Space in Self-Supervised Learning

Authors

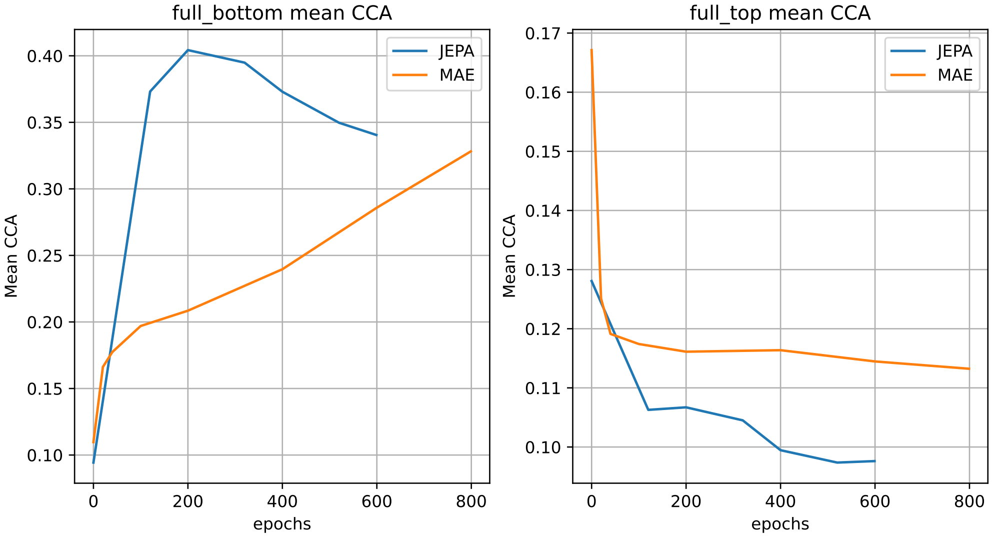

Figure 1: **Feature similarity dynamics for I-JEPA vs MAE on Imagenet dataset**. Our results in the paper predict that JEPA will tend to focus on the lower subspace of data variance where most of the perceptual features reside as claimed by Balestriero and Lecun [20]. To test this prediction in a realistic setting, we use the empirical setup described in [20] to study the differences in the feature learning dynamics of JEPA and MAE. We first describe the empirical setup and then make our observations.

Setup: We train a vision transformer (ViT) encoder using I-JEPA [3] and MAE [8] on ImageNet-1K data resized to 64 × 64 pixels. We use MAE-style masking to ensure parity with the masking function and copy all other hyperparameters described in I-JEPA [3] and MAE [8]. We build additional datasets by removing certain principal components of the full data subspace as described by Balestriero and Lecun [20] — **bottom** refers to the dataset that preservers bottom 25% of explained variance while **top** denotes the dataset with top 75% of explained variance. We use the original (**full**) images as well the filtered images described above and extract representations using encoders trained with I-JEPA and MAE. We then compare the representation between **full-bottom** and **full-top** via canonical correlation analysis (CCA) for several checkpoints that are gathered during training.

Observe that I-JEPA's features obtained from the full and bottom datasets show higher similarity compared to MAE throughout training while the trend is reversed for full and top images. This implies that JEPA learns features from the bottom portion of the PCA space **faster** than MAE. Balestrieo and Lecun [20] suggest that the bottom portion of the spectrum contain features useful for discrimination tasks which is what JEPA focuses on during optimization.