Masking Strategies in Self-Supervised Vision

Topic

Reconstruction Objectives and Representation Learning

Background

In traditional mask-based self-supervised learning, such as MAE, the standard methodology is to compute the reconstruction loss strictly on the masked regions, explicitly excluding the visible, unmasked patches from the loss function. Conversely, newer frameworks, such as TIPSv2, predict both masked and unmasked content simultaneously, which empirically enhances representation learning capabilities.

Assignment Task

Resolve the apparent contradiction between these two masking paradigms. Why does incorporating unmasked content into the predictive loss function improve feature abstraction in these newer frameworks, rather than causing the model to default to an identity mapping or trivial reconstruction? Base your explanation on gradient flow and representation density.

Submission Expectation

Prepare a rigorous, self-contained written response that defines all key assumptions, uses precise technical terminology, and supports the argument with mathematical, architectural, or conceptual reasoning where appropriate.