JEPA: Joint-Embedding Predictive Architecture

1. Idea

Self-supervised learning often trains models to reconstruct missing data. For example, given the left half of an image x_{left}, a generative model may learn

$$p(x_{\text{right}} \mid x_{\text{left}})$$

to reconstruct the missing pixels. However, pixel prediction requires modeling many irrelevant details (texture, lighting, noise) even though many completions are plausible.

JEPA (Joint-Embedding Predictive Architecture) instead predicts a *representation* of the missing information rather than the raw signal.

Given two views of the same input:

- context view x_c
- target view x_t

the model learns to predict the representation of x_t from x_c.

Key idea: learn predictable structure rather than reconstruct sensory detail.

2. Architecture

JEPA operates in an embedding space.

$$z_c = f_\theta(x_c), \qquad z_t = f_\xi(x_t)$$

where

- f_θ : online encoder
- f_ξ : target encoder

The target embedding is predicted from the context:

$$\hat{z}_t = g_\theta(z_c)$$

where g_θ is a predictor network.

The target encoder parameters are updated using an exponential moving average (EMA):

$$\xi \leftarrow \tau\xi + (1-\tau)\theta$$

and a stop-gradient is applied to z_t.

3. Training Objective

Training minimizes a regression loss between predicted and target embeddings:

$$\mathcal{L}(\theta) = \mathbb{E}_{x_c, x_t} \left[\| g_\theta(f_\theta(x_c)) - f_\xi(x_t) \|_2^2 \right].$$

Often embeddings are normalized:

$$\tilde{z} = \frac{z}{\|z\|}$$

leading to a cosine-style loss:

$$\mathcal{L} = \mathbb{E} \left[\| \tilde{\hat{z}}_t - \tilde{z}_t \|_2^2 \right].$$

4. Preventing Representation Collapse

A trivial solution is a constant embedding:

$$f_\theta(x) = \text{constant}.$$

JEPA avoids collapse through:

- stop-gradient on the target branch
- a slowly updated EMA target encoder
- an asymmetric predictor g_θ

This stabilization mechanism is similar to approaches such as BYOL.

5. Non-Generative Nature

JEPA does not model a probability distribution over the data.

There is no learning of

$$p(x), \qquad p(x|z), \qquad p(x_t|x_c).$$

In particular:

- no likelihood objective
- no decoder to pixel space
- no sampling procedure

The model simply learns a representation space in which predictable relationships between views can be inferred.

6. Conceptual Interpretation

Suppose the data decomposes as

$$x = (s, \epsilon)$$

where

- s represents semantic structure
- ϵ represents unpredictable detail

Pixel-level generative models must model both components.
JEPA instead aims to learn a representation

$$z \approx s$$

that captures predictable structure while ignoring unpredictable details.

7. Summary

JEPA is a self-supervised representation learning framework that:

- predicts latent representations instead of reconstructing pixels,
- trains via regression in embedding space,
- uses a momentum target encoder and stop-gradient for stability,
- focuses on predictable semantic structure rather than full data generation.

The learned representation can later be used for downstream tasks or combined with generative models operating in latent space.