I-JEPA reading notes

1. Introduction

image self-supervised learning:

① invariance-based method:

similar embeddings for ≥ 2 views of two same image

— high semantic level representation ✓

— hand-crafted data augmentation ×

— strong biases ×

② generative method:

remove/corrupt portions of the input + predict corrupted $\underbrace{\text{content}}_{= \text{pixel / token}}$

— less prior knowledge ✓

— generalize beyond modality ✓

— lower semantic level ×

— underperform ① ×

2. Methods: I-JEPA framework

<1> targets

input image $y \longrightarrow N$ non-overlapping patches

$\downarrow$

target encoder $f_{\bar{\theta}}$

$\downarrow$

patch level representation $s_y = \{s_{y_1}, \cdots, s_{y_N}\}$

$\downarrow$ random sample

$\underline{M}$ blocks : $M = 4$

aspect ratio = (0.75, 1.5 , scale = (0.15, 0.2)

the ith $\underbrace{\text{block}}_{\text{representation}}$ cover $\underbrace{\text{mask } B_i}_{0 \text{ or } 1}$: $s_y(i) = \{\underbrace{s_{y_j}}_{\text{output of target encoc}}\}_{j \in B_i}$

<2> Context

input image $\longrightarrow$ single block x + mask B_x :

unit aspect ratio + random scale $(0.85, 1.0)$

$\longrightarrow$ remove overlap with M targets

$\downarrow$

context encoder f_θ

$\downarrow$

patch level representation $s_x = \{s_{x_j}\}_{j \in B_x}$

<3> Prediction

conditioned on

input : $s_x + \underbrace{\{m_j\}_{j \in B_i}}$ a mask token for each patch we wish to predict

$\downarrow g_\phi$

output : $\hat{s}_y(i) = \{\hat{s}_{y_j}\}_{j \in B_i} = g_\phi(s_x, \{m_j\}_{j \in B_i})$

apply predictor M times : $\hat{s}_y(1), \cdots, \hat{s}_y(M)$

* mask tokens : parameterized by a shared learnable vector with an added positional embedding

* Loss : L_2 loss between $\hat{s}_{y_j}$ and s_{y_j}.

* ϕ, θ learned through gradient descent

$\bar{\theta}$ updated via exponential moving average of θ

3. Self-supervised learning architectures

objective: Energy-Based Models (EBM)

incompatible inputs $\longrightarrow$ high energy

compatible inputs $\longrightarrow$ low energy

<1> Joint-Embedding Architectures (JEA)

invariance-based pretraining

incompatible inputs $\longrightarrow$ dis-similar embeddings

compatible inputs $\longrightarrow$ similar embeddings

Challenge: <u>representation collapse</u>

the encoder produces a constant output regardless of the input

<2> Generative Architectures

reconstruction-based methods

produce compatible x, y conditioned on z — position tokens

x: copy of y with mask

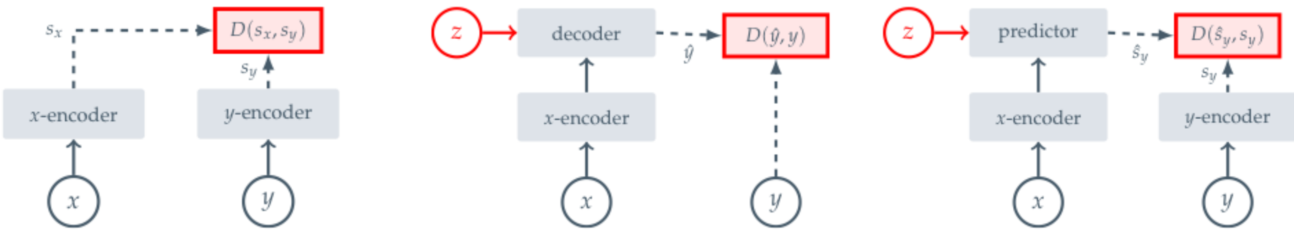
No representation collapse: the informational capacity of z is low compared to the signal y

<3> Joint-Embedding Predictive Architecture (JEPA)

predict embedding of y from a compatible signal x conditioned on z

Challenge: representation collapse

solution: asymmetric architecture between x- and y-encoders.



(a) **Joint-Embedding Architecture** (b) **Generative Architecture** (c) **Joint-Embedding Predictive Architecture**

4. Other

<1> I-JEPA有强语义性的2个原因：

① unnecessary pixel level details are potentially eliminated.

② multi-block masking strategy

<2> 接近的工作：data2vec

Context Autoencoders

GI-JEPA

noise I-JEPA