Distilling Knowledge from Self-Supervised Teacher by Embedding Graph Alignment

Yuchen Ma¹, Yanbei Chen¹, Zeynep Akata¹,²,³
¹University of Tübingen, ²MPI for Informatics, ³MPI for Intelligent Systems
*Equal contribution. The work was done when Yuchen Ma was enrolled in the MSc program at the University of Heidelberg.

Introduction

Task Introduction

Knowledge Distillation

Goal: Learn visual representation by knowledge distillation

Motivation

Contribution:

➢ Propose a new knowledge distillation method to transfer the instance-wise structural knowledge.
➢ Establish a comprehensive benchmark on three image classification datasets
➢ Demonstrate the superiority of our model under a variety of evaluation setups.

Model Overview

Graph Construction

Node: Feed the extracted features to individual node embedding layers.
Edge:based on the Pearson’s correlation coefficient (PPC)
\[ r_{ij} = \frac{\sum_{k} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k} (x_{ik} - \bar{x}_i)^2 \sum_{k} (x_{jk} - \bar{x}_j)^2}} \]
Edge matrix: encode the correlation between every pair of images among the same batch

Embedding Graph Alignment

Teacher Graph

Student Graph

➢ Edge matching loss
\[ L_{edge} = \| E_E - E_T \| \]
➢ Node matching loss
\[ L_{node} = \| N_E - N_T \| \]
➢ Distillation loss
\[ L = L_{edge} + \lambda L_{node} \]
➢ Training loss
\[ L = L_{node} + \lambda L_{edge} \]

Experiment

Evaluation on different network architectures

Training loss

Visualization of self-supervised embeddings

Analyzing the learning dynamics

Reference