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

Train task directly

Target task

New network

Original training

Knowledge distillation

Teacher (expert of target task)

Mimic the teacher (trained network)

Loss

Student (new network)

Loss

Goal: Learn visual representation by knowledge distillation

Model Overview

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.

Motivation

Embedding Graph Alignment

Modeling the instance-instance correlations

Transferring the graph structural knowledge

Use self-supervised knowledge

➢ Construct the teacher graph and the student graph

➢ Align the teacher graph and the student graph

➢ Jointly optimize an edge matching constraint and a node matching constraint.

Experiment

Graph Construction

Node:

Feed the extracted features to individual node embedding layers.

Edge:

Based on the Pearson’s correlation coefficient (PPC)

\[ r_{ij} = \frac{\sum_{k=1}^{N} (x_{ik} - \bar{x}_i) (x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{N} (x_{ik} - \bar{x}_i)^2} \sqrt{\sum_{k=1}^{N} (x_{jk} - \bar{x}_j)^2}} \]

Edge matrix: encode the correlation between every pair of images among the same batch

Embedding Graph Alignment

➢ Edge matching loss

\[ L_{edge} = \| E_T - E_S \| \]

➢ Node matching loss

\[ L_{node} = \| \nabla X_T - \nabla X_S \| \]

➢ Distillation loss

\[ L_{distill} = L_{node} + \lambda L_{edge} \]

➢ Training loss

\[ L = L_{node} + \lambda L_{edge} \]

Graph Embedding

Node embedding: Feed the extracted features to individual node embedding layers.

Edge embedding: Based on the Pearson’s correlation coefficient (PPC)

\[ r_{ij} = \frac{\sum_{k=1}^{N} (x_{ik} - \bar{x}_i) (x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{N} (x_{ik} - \bar{x}_i)^2} \sqrt{\sum_{k=1}^{N} (x_{jk} - \bar{x}_j)^2}} \]

Graph alignment

[1] Yanbei Chen, Yongqin Xian, Akiko Koepke, Yong Shan, and Zeynep Akata. Distilling audio-visual knowledge by compositional contrastive learning. In CVPR, 2021.
[2] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.
[3] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In CVPR, 2019.
[4] Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In ECCV, 2018.
[5] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In ICLR, 2015.
[6] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In ICLR, 2019.