Exercises for summer school on self-supervised learning at DTU

Introduction

Learning a meaningful representation from data without human supervision is a well established problem in machine learning. However, recent works on representation learning through self-supervision have achieved impressive performance on a wide range of different tasks involving data types such as images, time series, and text. This exercise set will be focused on the RELAX framework [1], which can be used to explain representations of data. The exercises in this set is focused on both theoretical and technical aspects of RELAX. RELAX explains representations by measuring similarities in the representation space between an input and masked out versions of itself.

Let $\mathbf{X} \in \mathbb{R}^{H \times W}$ represent an image[1] consisting of $H \times W$ pixels, and f denote a feature extractor that transforms an image into a representation $\mathbf{h} = f(\mathbf{X}) \in \mathbb{R}^D$. To mask out regions of the input, we apply a stochastic mask $\mathbf{M} \in (0, 1)^{H \times W}$, where each element M_{ij} is drawn from some distribution. The stochastic variable $\bar{\mathbf{h}} = f(\mathbf{X} \odot \mathbf{M})$, where $\odot$ denotes element-wise multiplication, is a representation of a masked version of $\mathbf{X}$. Moreover, we let $s(\mathbf{h}, \bar{\mathbf{h}})$ represent a similarity measure between the unmasked and the masked representation. Intuitively, $\mathbf{h}$ and $\bar{\mathbf{h}}$ should be similar if $\mathbf{M}$ masks *non-informative* parts of $\mathbf{X}$. Conversely, if *informative* parts are masked out, the similarity between the two representations should be low.

Problem 1

In the original form of RELAX, the masks are assumed to be continuous in (0, 1). In this problem, we will derive a related expression but where we assume binary masks (i.e. $M_{ij} \in \{0, 1\}$). Note, this derivation is closely related to how the RISE algorithm is constructed [2].

Let R_{ij} define the importance of pixel (i, j) to the representation $\mathbf{h}$. With binary masks where the probability of $M_{ij} = 1$ is p, R_{ij} can be computed as the expected similarity score, conditioned on the pixel being present:

$$R_{ij} = \mathrm{E}_{\mathbf{M}}\big[s(\mathbf{h}, \bar{\mathbf{h}}) | M_{ij} = 1\big]. \tag{1}$$

Show that Equation 1 can be rewritten into the following form:

$$R_{ij} = \frac{1}{p} \sum_{\mathbf{M}'} s(\mathbf{h}, \bar{\mathbf{h}}_{\mathbf{M}'}) M_{ij} P(\mathbf{M} = \mathbf{M}'), \tag{2}$$

[1] To enhance readability, we do not include image channels, but this can be easily included by letting the masks span the channel dimension.

where the sum is over all possible masks, and $\bar{\mathbf{h}}_{\mathbf{M}'}$ is the representation of the image masked with mask $\mathbf{M}'$.

Note: it is not feasible to sum over all masks, so instead we randomly sample masks to approximate the sum in Equation 2. After further refinment, the expression with continious versus binary masks becomes very similar.

Problem 2

Implement the main algorithm of RELAX. Use the notebook at:

- https://github.com/Wickstrom/ssl-summer-school-dtu.

Problem 3

The following inequality relates the number of masks needed to obtain an absolute estimation error of less than t with a probability of at least $1 - \delta$ (see theorem 3.1 of [1] for more details):

$$2e^{\frac{-2t^2}{\sum_{n=1}^{N} c_n^2}} \leq \delta, \tag{3}$$

where $c_n = b_n - a_n$. In the case of RELAX, $a_n = 0/N$ and $b_n = 1/N$, which gives $c_n = 1/N$.

Find the number of masks required to obtain an estimator error of less than 0.01 (i.e. $t = 0.01$) with a probability of 0.99 (i.e. $\delta = 0.99$).

The value of a_n and b_n are determined by the similarity measure used in RELAX, where $a_n = 0/N$ since the minimum of the cosine similarity is 0 (in the case of non-negative vectors), and $b_n = 1/N$ since the maximum value of the cosine similarity is 1. However, in practice, a similarity of 0 almost never occurs.

Explain why we usually do not see a similarity of 0 between the masked and unmasked representation. Use your code from problem 2 and plot a historgram of the similarity scores for the four example and 1000 masks. What is the minimum similarity you observe? Use this minimum similarity to derive a new empirical bound for the number of masks required to obtain a desired error with a desired probability. Using your empirical bound, how many masks do you now need to obtain an estimator error of less than 0.01 (i.e. $t = 0.01$) with a probability of 0.99 (i.e. $\delta = 0.99$).

Problem 4 (bonus)

Implement the one-pass version of RELAX (Equation 12 and 13 of [1]).

References

[1] Wickstrøm, K.K., Trosten, D.J., Løkse, S. et al., RELAX: Representation Learning Explainability. International Journal of Computer Vision, 2023

[2] Vitali Petsiuk and Abir Das and Kate Saenko, RISE: Randomized Input Sampling for Explanation of Black-box Models. Proceedings of the British Machine Vision Conference, 2018