Machine Learning CS 460 - Project

ELSA - EXPLORATIVE LATENT SEARCH ALGORITHM

by Aaditya Vicram Saraf, A Rameswar Patro
Under the guidance of Dr. Subhankar Mishra, SCPS NISER

Email id : aadityavicram.saraf@niser.ac.in ; arameswar.patro@niser.ac.in

Introduction

Curating and labelling large datasets for training ML models has often been a major concern. There have been large manually prepared datasets for this, but it is limited in its purpose. The ELSA algorithm aims to reduce manual labelling to as low as possible using small available datasets (seeds) and discover all positive samples in a dataset with a massive pool of negative samples.

ELSA incorporates VICReg for converting a given image dataset into meaningful vector embedding where similar samples are assigned nearby vectors while dissimilar samples are farther apart. Then based on a given seed, a nearest neighbour component searches for samples lying in the same cluster as a given particular sample is discovered. For discovering multiple clusters (ideally all), a random sampler finds suitable points where further NN search can be done. The labelled points are then used to train an MLP head which will assign confidence values to our embed vectors, which would help random sampler to choose ”suitable points” in subsequent iterations.

How it works?

Intuition behind ELSA

ELSA learns to label a large unlabelled dataset containing very less positive samples hidden in a pool of negative samples. We aim to train the model to achieve a high performance with help of a smaller given labelled dataset of positive samples. These positive samples (called seeds) are used to train MLP in its initial stages so that it learns to identify how a positive sample looks.

Next, when given an unlabelled dataset, it is passed through a feature extractor (In our model VICReg [1] has performed best), and sent to MLP head to start the model training. Initially some confidence value is assigned to the points in our embedding of dataset, and a random search is performed. The randomly selected points are given to oracle (human) for labelling and then the positive samples are used for a nearest neighbour search. The discovered points are sent to oracle for labelling and the cycle for positive samples continue. Meanwhile both the positive and negatively labelled samples are sent to the MLP head for training based on which it assigns suitable confidence values to the embed vectors of our dataset.

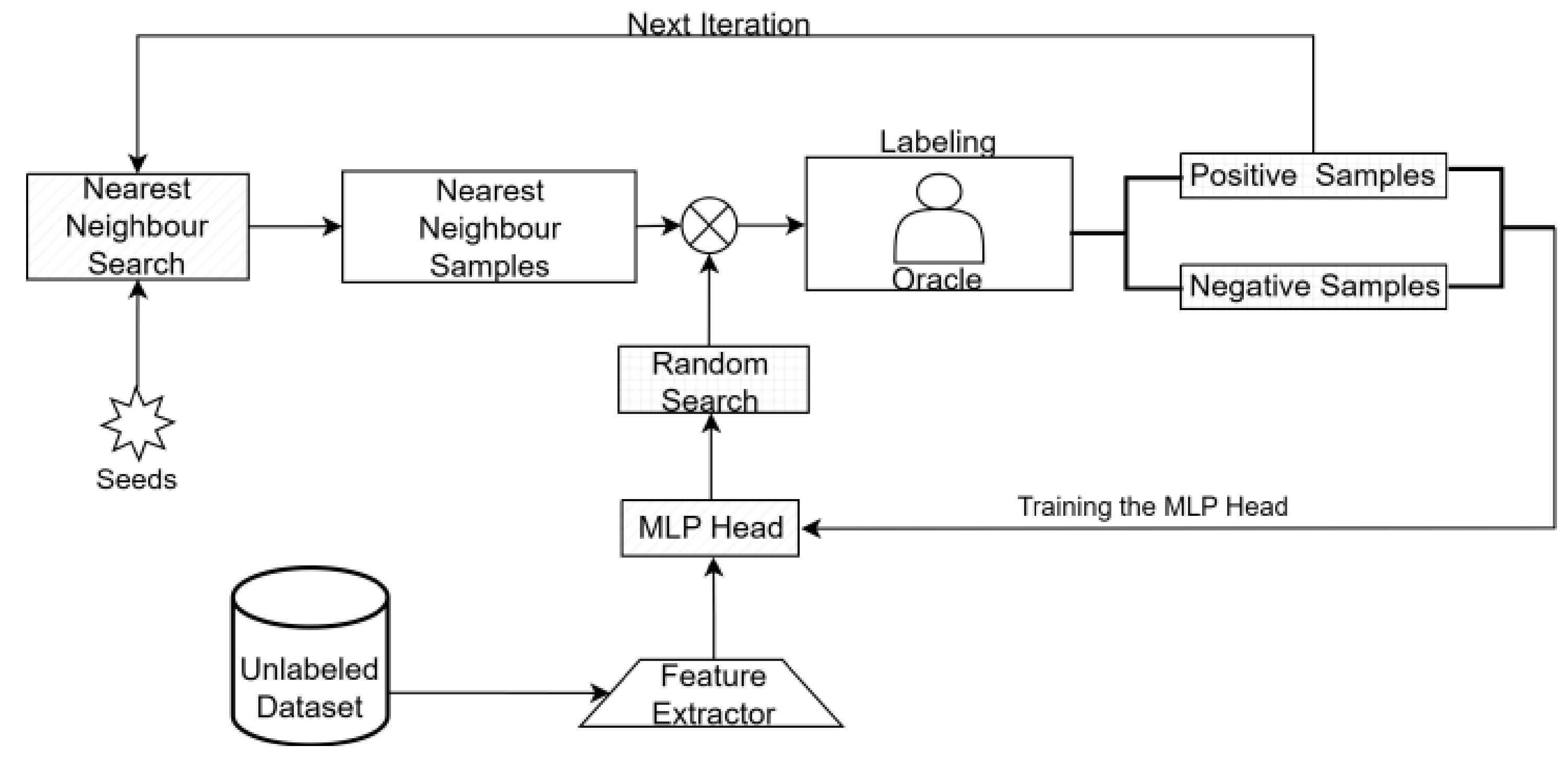

Figure 4: Architecture of ELSA

(Image from the original ELSA paper. Refer to [2])

Working of the model

ELSA introduces 3 major components which execute the complete task at hand. These are back-boned by VICReg algorithm which provides a suitable vector representation for our dataset points. Individually, these components are :

- **MLP head :** This is a neural network, which basically works like a function that assigns confidence values between +1 and -1 to the images in our complete dataset. The points labelled positive have +1 and the points discovered negative are assigned -1. This is based on the positive and negative samples which come from either seeds or after labelling through ELSA.
- **Nearest neighbour search component (kNN) :** This is a kNN algorithm which discovers the nearest neighbours of a particular vector based on MSE distance in the embedding space. It sorts the vector arguments in an increasing value of MSE with the root vector and picks the k closest neighbours.
- **Random Search Component (RandS) :** The random search component used in ELSA is NOT completely random. Based on how the MLP is trained, RandS chooses high confidence vectors which are far from each other. They are given to the oracle for labelling, and positive samples so obtained are used for an NN search.

Our contribution

Several modifications and improvements have been done by us in regards to ELSA. Apart from clarifying and clearing the notations and minor mistakes with the ELSA paper, on understanding the working of ELSA, we had two primary goals. We were mostly successful in both of our major aspirations, which have been described below :

- **Devised an Error for labelling efficiency :** The labelling efficiency L_e of ELSA is given by
$$L_e = \frac{\|\text{Points labelled positive } (L^+)\|}{\|\text{Total points labelled } (L)\|}$$
We discovered that if ε is the percentage labelling error made by the oracle, then the new labelling efficiency can be computed by
$$\frac{L^+ - L_e \varepsilon * r * t}{L + (1 - L_e)\varepsilon * r * t}$$
which simplifies to
$$L_e \left[\frac{(1-\varepsilon)}{1 + (1 - L_e) * \varepsilon} \right].$$
- **Proposed other RandS components :** The old RandS components was complicated and had the issue that it might give close-by points, hence not very efficient in discovering new positive clusters. We propose some new RandS algorithms, which are explained in brief below :

1. **GreedyS :** We will select high confidence points say those whose confidence range in (0.85,0.95), and to maximize the distance between the chosen points from each other and already discovered positives, we choose points one by one for an NN iteration.
2. **FarS :** Farthest Sampler aims to maximize the function given below and works similar to GreedyS but chooses multiple points at once and it only maximizes mutual distance of selected points in the embed space. It chooses points $(x_1, x_2, \ldots, x_n)$ such that :
$$g(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} \|x_i - x_j\| + \sum_{i=1}^{n} f_c(x_i)$$
3. **GridS :** Grid Sampler entails the idea to find one initial point in the whole dataset and find the diameter (say d) of the whole dataset from this sample in our embed space. Then we consider our embed space restricted to a hypersphere of radius ’d’ and divide it into smaller sections by using grids. Then in each small grid block so obtained, we can choose high confidence points.

References

[1] A. Bardes, J. Ponce, and Y. LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

[2] A. Nath, D. Choudhury, and S. Mishra. Explorative latent self-supervised active search algorithm (ELSA), 2024.