Active Learning using Feature Mixing and CSVAL

Harsh Choudhary
Dept. of Electrical Engineering
IIT Bombay
Mumbai, India
200070023@iitb.ac.in

Abstract—One of the reasons for the recent progress of deep learning is the increasing size of training data which comes with higher labeling costs. Thus, to limit the labeling cost, we prioritize the important data to label using Active Learning (AL). It is the machine learning technique that enables machines to ask thoughtful questions and reduce labeling efforts by selecting the most important samples. In this report, we discuss a solution to the cold start problem of the existing algorithm and will study Feature-Mixing which is very helpful in high dimensional data with lesser samples.

I. INTRODUCTION

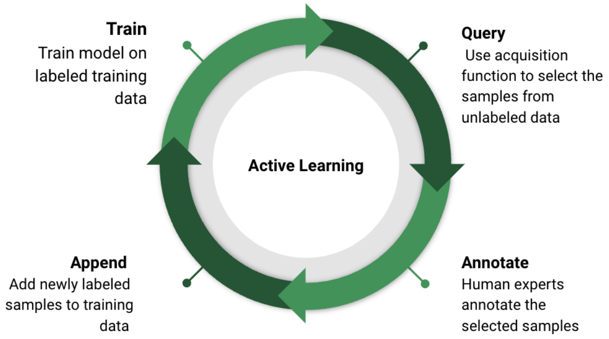
Train
Train model on labeled training data
Query
Use acquisition function to select the samples from unlabeled data
Active Learning
Append
Add newly labeled samples to training data
Annotate
Human experts annotate the selected samples

Fig. 1. Visual Guide to Active Learning [3]

AL promises to improve annotation efficiency by iteratively selecting the most important data points to annotate, the important data points are those on which the model is most uncertain. Random selection is considered as a baseline to AL because the randomly sampled query is independent and identically distributed to the entire data distribution. However, during the first few iterations, random selection outperforms most of the existing AL algorithm, we identify it as the cold start problem. [1] shows that it is caused by class imbalance and outliers in initial queries, results show that it can be solved by ensuring label diversity by generating pseudo labels and selecting hard-to-contrast data.

Current AL strategies struggle when applied to deep neural networks when they have high-dimensional data with smaller no. of training samples. To address this issue, [2] introduces Active Learning by Feature Mixing (ALFA-Mix), this method involves the mixing of features. It constructs interpolations between the representations of labeled and unlabelled data and then, analyzes the predicted labels. The paper shows that the inconsistencies in these predictions help in discovering features that the model is unable to recognize in the unlabelled samples. The discovered features are used to select the most valuable samples from the pool of unlabelled data by recognizing unlabelled samples with unique features. This method outperforms all recent AL methods on 12 benchmarks in 30 different settings.

II. METHODOLOGY

A. The Cold Start Problem

AL tends to select data that is biased to specific classes i.e. some classes are simply not selected for training, mostly because of an unbalanced datasets in real life. To address this issue label diversity of selected samples is considered an important criterion to determine its importance. To ensure label diversity, we generate the pseudo-labels using K-means clustering and select an equal number of data from each cluster in the initial query. In case, the number of classes is unknown, it is recommended to classify into a higher number of classes to increase performances on the datasets.

After applying K-means, the next step is to select data points from each cluster. Those data points which are harder to contrast with other data points in the cluster can better represent a cluster distribution and we consider them as typical data and non-outliers.

$$P_{i,j} = \frac{e^{sim(Z_i,Z_j)}/t}{\Sigma_{n=1}^{2N} I(n\neq i) \times e^{sim(Z_i,Z_n)}/t} \quad (1)$$

< u, v > gives cosine similarity between u and v,

z_{2n}, z_{2n-1} denote the projection head output of a positive pair for the input x_n in a batch, $I(n)$ is an indicator function, it equals to 1 for n ≠ i otherwise it's 0.

This Eq. 1. tells about the probability of selecting a data point j given a query point i

$$P_{\theta(e)}(y^{*}_{n}|x_{n}) = \frac{1}{2}\left[P_{2n-1,2n} + P_{2n,2n-1}\right] \qquad (2)$$

This Eq. 2. tells about the probability of assigning a pseudo-label to a data point n, given its contrastive feature x, and $\theta(e)$ denotes the parameters at the end of epoch e

$$\mu_m = \frac{1}{E}\sum_{e=1}^{E} P_{\theta(e)}(y^{*}_{n}|x_{n}) \qquad (3)$$

Here, μ_m tells about the confidence of the predicted label Data points with higher confidence are easy to contrast data whereas those with lower confidence are hard to contrast data and are thus selected.

B. Feature Mixing

[2] proposes way to select unlabelled instances with unique features by constructing interpolations between representations of labeled and unlabelled instances and optimizing the mixing ratios. The optimal mixing ratio is chosen by seeking the most complex case of mixing ratios for each unlabelled instance and anchor.

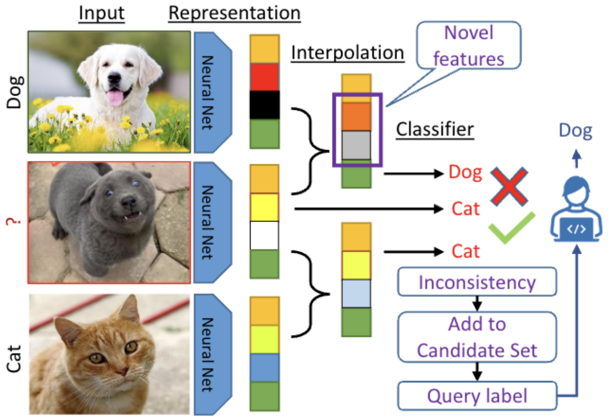

Fig. 2. Representation of extraction function f_e, classifier function f_c, identification of novel features and labeling of important samples [2]

During each iteration, this active learner has access to a small set of labeled data. The learner also has access to a set of unlabelled data from which B instances are chosen to be labeled. The learner is a deep neural network $f = f_e \odot f_c$ parameterized by $\theta = \{\theta_e, \theta_c\}$. Here $f_e : X \to R_D$ is the backbone which converts the input into a D-dimensional representation in a latent space (or extracts features from images), given as $z = f_e(x;\theta_e)$. Further, f_c is a classifier that is used to predict the class labels of the input data given as $p(y\,|\,z;\theta) = softmax(f_c(z_i;\theta_c))$ Here, intuition is that the model's incorrect prediction is mainly caused by features which are not recognizable in the input. Thus, to reduce their effect, we use a convex combination (i.e. interpolation) of the features in the vicinity of each unlabelled point given as

$$z^{\alpha} = \alpha z^{*} + (1-\alpha)z^{u} \qquad (4)$$

Equation 4, z^{*} is the average representation of the labeled samples per class and is called as the anchor, z^{u} is the output of f_e when unlabeled data is given as input to it, and α is the interpolation parameter lying in the set $[0,1)^{D}$. When we put the convex combination of point in the loss term and using Taylor series expansion w.r.t. z^{u}, we get

$$\begin{aligned} &max\left[L(f_c(z^{*}), y^{*})L(f_c(z^{u}), y^{*})\right] \\ &= max[(\alpha(z^{*} - z^{u}))^{T}.\nabla_{z^{u}} l(f_c(z^{u}), y*)]. \qquad (5) \end{aligned}$$

From Eq. 5. we can observe that the change in the loss term is proportionate to two terms: (i) α times the difference of features of z* and z^{u}, (ii) the gradient of the loss w.r.t the unlabelled inputs in the latent space.
(i) determine the features in unlabelled data which are new and different from those in labeled data and (ii) determine the sensitivity of the model to those features. In case the features of the labeled and unlabelled samples are entirely different but the model is almost the same, there is no change in the loss, and hence those features are not considered important to the model and those samples are not selected.s
The optimum value of α is α* for which r.h.s of Eq. 5. is maximized. The approximate solution for this optimization solution as given in [2] as:

$$\alpha^{*} \approx \frac{\epsilon \times \|(z^{*}-z^{u})\|_{2}\nabla_{z_u}(f_c(z^{u}), y*)}{\|\nabla_{z_u}(f_c(z^{u}), y*)\|_{2}} \oslash (z^{*} - z^{u}) \qquad (6)$$

To identify the points with unique features, the point is selected in set I if the predicted label for an interpolated representation of the unlabelled instance is inconsistent with the true label of the unlabeled instance.

The dimension of *I* can be larger than the required number of selections, so we take k-means and select the required number of samples.

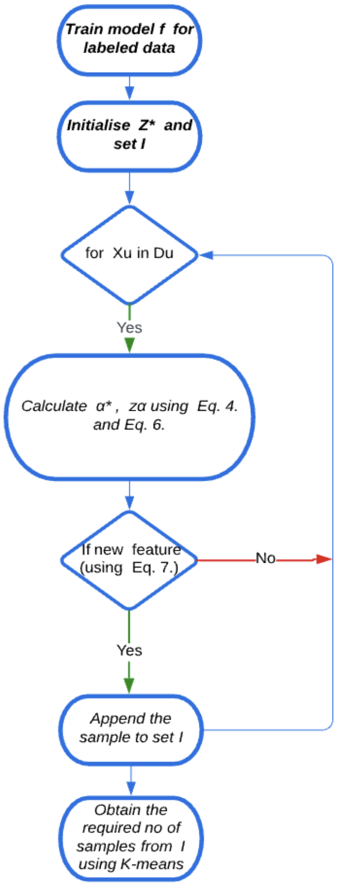
Train model f for labeled data
Initialise Z* and set I
for Xu in Du
Yes
Calculate α* , zα using Eq. 4. and Eq. 6.
If new feature (using Eq. 7.)
No
Yes
Append the sample to set I
Obtain the required no of samples from I using K-means

Fig. 3. Block diagram for AL using Feature mixing [2]

III. CODE

For feature mixing, run the colab file and select the dataset from MNIST, EMNIST, SVHN, CIFAR10, CIFAR100, MiniImageNet, openml_6, openml_155 or load custom datasets and add dataloader for it by making changes in main.py and datasets.py.

For Cold Start, run the file and select the dataset from MNIST, EMNIST, SVHN, CIFAR10, CIFAR100, MiniImageNet, openml_6, openml_155 or load custom datasets and add dataloader for it.

We tried to integrate these code with DISTIL but due to highly different dependencies, task is out of scope of this project.

IV. RESULTS

Contrastive learning solution of Feature-Mixing is evaluated on MiniImageNet datasets by [1], and the results show that the initial query selected by the proposed method outperforms existing active querying strategies and random selection by a large margin. The performance was compared with various algorithms including BALD, Coreset, Consistency, VAAL, and random selection

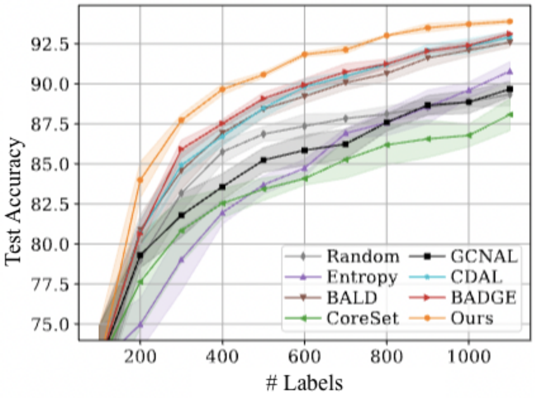

Fig. 4. Test accuracy of Feature mixing vs other algorithms on MiniImageNet

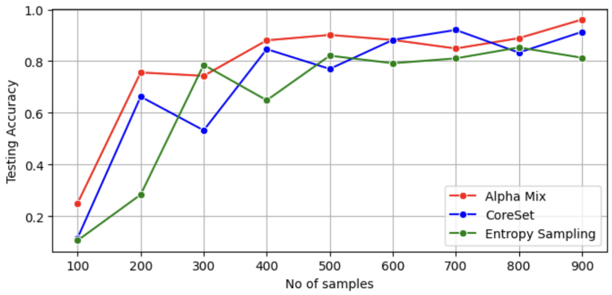

Fig. 5. We have compare AlphaMixSampling, CoreSet and EntropySampling on Mnist Dataset

V. CONCLUSIONS

In this project, we studied the active learning in different situations. The cold start problem which is caused by imbalanced training data and outlier selection which can be solved by ensuring balanced sampling and selecting hard-to-contrast data, and Feature mixing for high dimensionality in low data regime. Code have some dependencies which can not be satisfied together, thus I am working on better implementation which can be added to Distil.

ACKNOWLEDGMENTS

I would like to thank Prof Ganesh Ramakrishnan for guiding me throughout the course, TA Durga Sivasubramanian, for guiding me through DISTIL.

REFERENCES

[1] Liangyu Chen, Yutong Bai, Siyu Huang, Yongyi Lu, Bihan Wen, Alan L. Yuille, Zongwei Zhou. Making Your First Choice: To Address Cold Start Problem in Vision Active Learning (2022). https://arxiv.org/abs/2210.02442

[2] Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Reza Haffari, Anton van den Hengel, Javen Qinfeng Shi. Active Learning by Feature Mixing (2022). https://arxiv.org/abs/2203.07034

[3] https://www.v7labs.com/blog/active-learning-guide