Active Learning by Feature Mixing

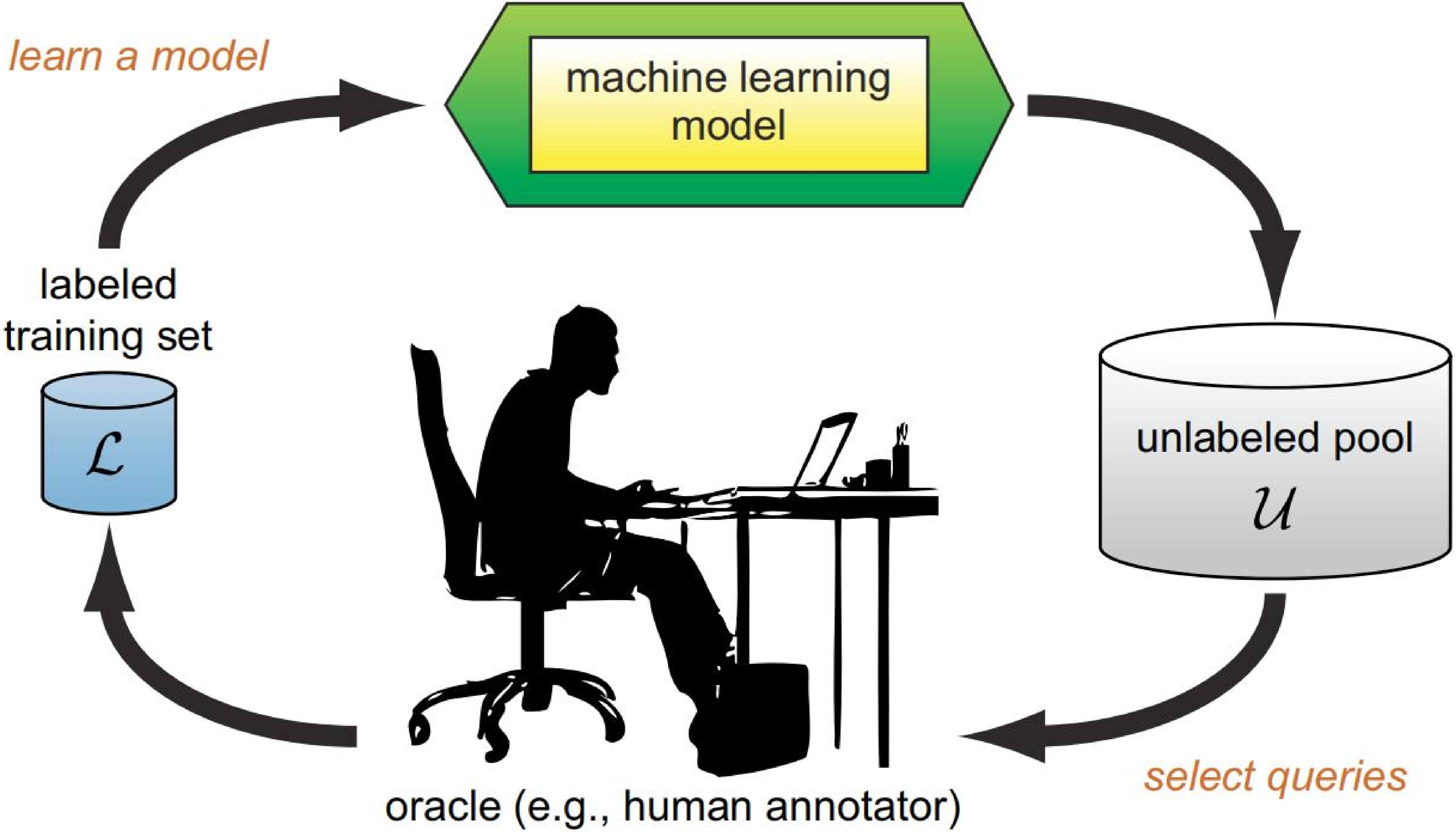
Amin Parvaneh[1] Ehsan Abbasnejad[1] Damien Teney[1,2] Reza Haffari[3],
Anton van den Hengel[1,4] Javen Qinfeng Shi[1]

[1]Australian Institute for Machine Learning, University of Adelaide
[2]Idiap Research Institute [3]Monash University [4]Amazon

{amin.parvaneh, ehsan.abbasnejad, javen.shi, anton.vandenhengel}@adelaide.edu.au
damien.teney@idiap.ch gholamreza.haffari@monash.edu

CVPR 2022

Active Learning



Method

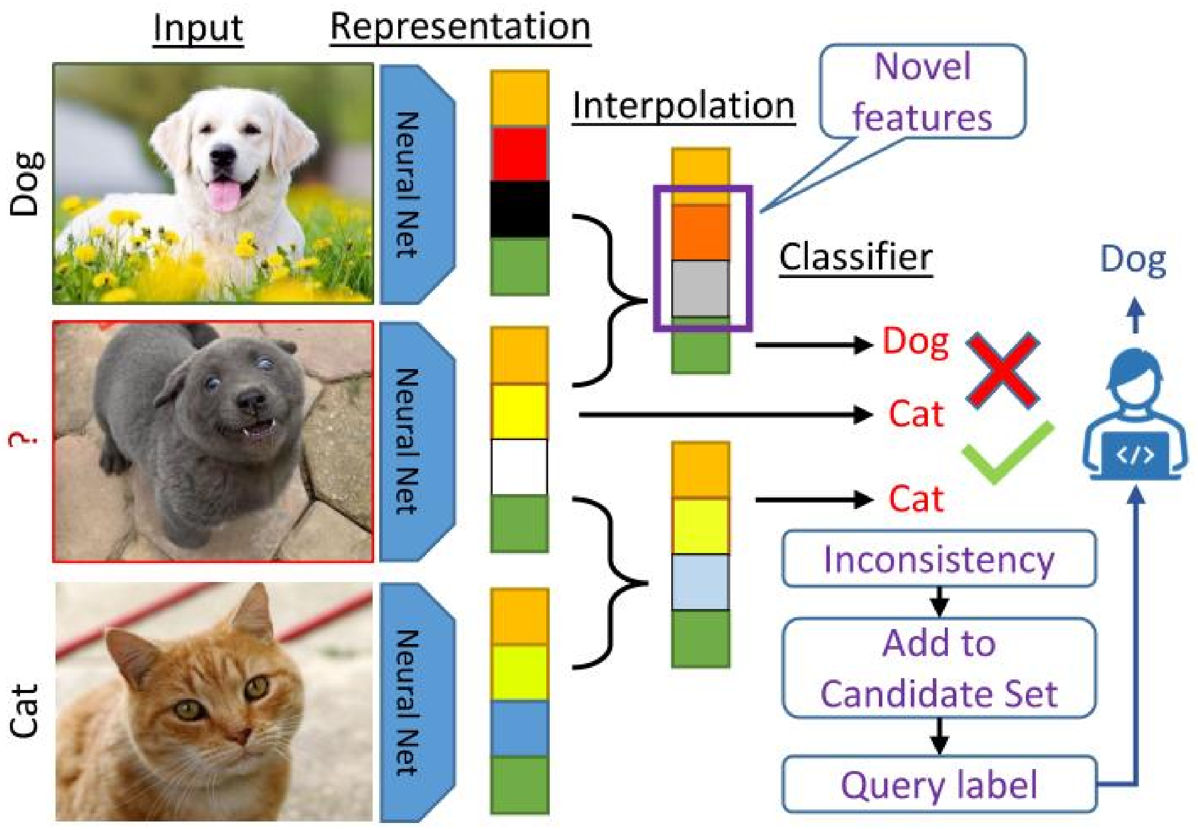


Figure 1. We propose to form linear combinations (*i.e.* interpolations or mixing) of the features of an unlabelled instance (middle image) and of labelled ones (top and bottom images). The interpolated features are passed through the current classifier. We show that inconsistencies in the predicted labels indicate that the unlabelled instance may have novel features to learn from.

Method

南京航空航天大学
Nanjing University of Aeronautics and Astronautics

$$\boldsymbol{z} = f_e(\boldsymbol{x}; \boldsymbol{\theta}_e) \qquad p(y \mid \boldsymbol{z}; \boldsymbol{\theta}) = \operatorname{softmax}(f_c(\boldsymbol{z}; \boldsymbol{\theta}_c)).$$

$$\tilde{\boldsymbol{z}}_{\boldsymbol{\alpha}} = \boldsymbol{\alpha} \boldsymbol{z}^{\star} + (1 - \boldsymbol{\alpha}) \boldsymbol{z}^{u}$$

$$\begin{aligned} \tilde{\boldsymbol{z}}_{\boldsymbol{\alpha}} - \boldsymbol{z}^{u} &= (\boldsymbol{\alpha} \boldsymbol{z}^{\star} + (1 - \boldsymbol{\alpha}) \boldsymbol{z}^{u}) - \boldsymbol{z}^{u} \\ &= \boldsymbol{\alpha} \boldsymbol{z}^{\star} + \boldsymbol{z}^{u} - \boldsymbol{\alpha} \boldsymbol{z}^{u} - \boldsymbol{z}^{u} \\ &= \boldsymbol{\alpha} \boldsymbol{z}^{\star} - \boldsymbol{\alpha} \boldsymbol{z}^{u} \\ &= \boldsymbol{\alpha}(\boldsymbol{z}^{\star} - \boldsymbol{z}^{u}) . \end{aligned}$$

$$\begin{aligned} \ell\left(f_c\left(\tilde{\boldsymbol{z}}_{\boldsymbol{\alpha}}\right), y^*\right) \approx\, & \ell\left(f_c(\boldsymbol{z}^u), y^*\right) + \\ & \left(\tilde{\boldsymbol{z}}_{\boldsymbol{\alpha}} - \boldsymbol{z}^u\right)^{\mathsf{T}} . \nabla_{\boldsymbol{z}^u} \ell\left(f_c\left(\boldsymbol{z}^u\right), y^*\right) \end{aligned}$$

$$\begin{aligned} \ell\left(f_c\left(\tilde{\boldsymbol{z}}_{\boldsymbol{\alpha}}\right), y^*\right) \approx\, & \ell\left(f_c(\boldsymbol{z}^u), y^*\right) + \\ & \left(\boldsymbol{\alpha}(\boldsymbol{z}^{\star} - \boldsymbol{z}^u)\right)^{\mathsf{T}} . \nabla_{\boldsymbol{z}^u} \ell\left(f_c\left(\boldsymbol{z}^u\right), y^*\right) , \end{aligned} \tag{2}$$

$$\begin{aligned} \max_{\boldsymbol{z}^{\star} \sim \boldsymbol{Z}^{\star}} \left[\ell\left(f_c\left(\tilde{\boldsymbol{z}}_{\boldsymbol{\alpha}}\right), y^*\right)\right] - \ell\left(f_c(\boldsymbol{z}^u), y^*\right) \approx \\ \max_{\boldsymbol{z}^{\star} \sim \boldsymbol{Z}^{\star}} \left[\left(\boldsymbol{\alpha}(\boldsymbol{z}^{\star} - \boldsymbol{z}^u)\right)^{\mathsf{T}} . \nabla_{\boldsymbol{z}^u} \ell\left(f_c\left(\boldsymbol{z}^u\right), y^*\right)\right] . \end{aligned} \tag{3}$$

Method

南京航空航天大学
Nanjing University of Aeronautics and Astronautics

$$\boldsymbol{\alpha}^{*} = \underset{\|\boldsymbol{\alpha}\| \leq \epsilon}{\arg\max} \, (\boldsymbol{\alpha}(\boldsymbol{z}^{\star} - \boldsymbol{z}^{u}))^{\intercal} . \nabla_{\boldsymbol{z}^{u}} \ell(f_{c}(\boldsymbol{z}^{u}), y^{*}), \quad (4)$$

approximate the solution to this optimisation using dual norm formulation

$$\boldsymbol{\alpha}^{*} \approx \epsilon \frac{\|(\boldsymbol{z}^{\star} - \boldsymbol{z}^{u})\|_{2} \nabla_{\boldsymbol{z}^{u}} \ell(f_{c}(\boldsymbol{z}^{u}), y^{*})}{\|\nabla_{\boldsymbol{z}^{u}} \ell(f_{c}(\boldsymbol{z}^{u}), y^{*})\|_{2}} \oslash (\boldsymbol{z}^{\star} - \boldsymbol{z}^{u}), \quad (5)$$

$\oslash$ indicates element-wise division

Method

Algorithm 1: Our active learning algorithm.

Inputs: initial labelled set $\mathcal{D}^l$; unlabelled pool $\mathcal{D}^u$; labelling budget at each round B; mixing parameter ϵ;

for $i = 1$ ***to*** *max_rounds* **do**

  Train the model f using the labelled data $\mathcal{D}^l$.

  Initialise $\boldsymbol{Z}^\star$ based on the representations of $\mathcal{D}^l$.

  $\mathcal{I} = \{\}$.

  for $\boldsymbol{x}^u \in \mathcal{D}^u$ **do**

    $\boldsymbol{z}^u = f_e(\boldsymbol{x}^u)$.

    for $\boldsymbol{z}^\star \in \boldsymbol{Z}^\star$ **do**

      Calculate $\boldsymbol{\alpha}^*$ using ϵ and Eq. 5.

      $\tilde{\boldsymbol{z}} = \boldsymbol{\alpha}^* \boldsymbol{z}^\star + (1 - \boldsymbol{\alpha}^*) \boldsymbol{z}^u$.

      if $\arg\max_y (f_c^y(\boldsymbol{z}_u)) \neq \arg\max_y (f_c^y(\tilde{\boldsymbol{z}}))$

       then

        $\mathcal{I} = \mathcal{I} \cup (\boldsymbol{x}^u, \boldsymbol{z}^u)$.

        Break

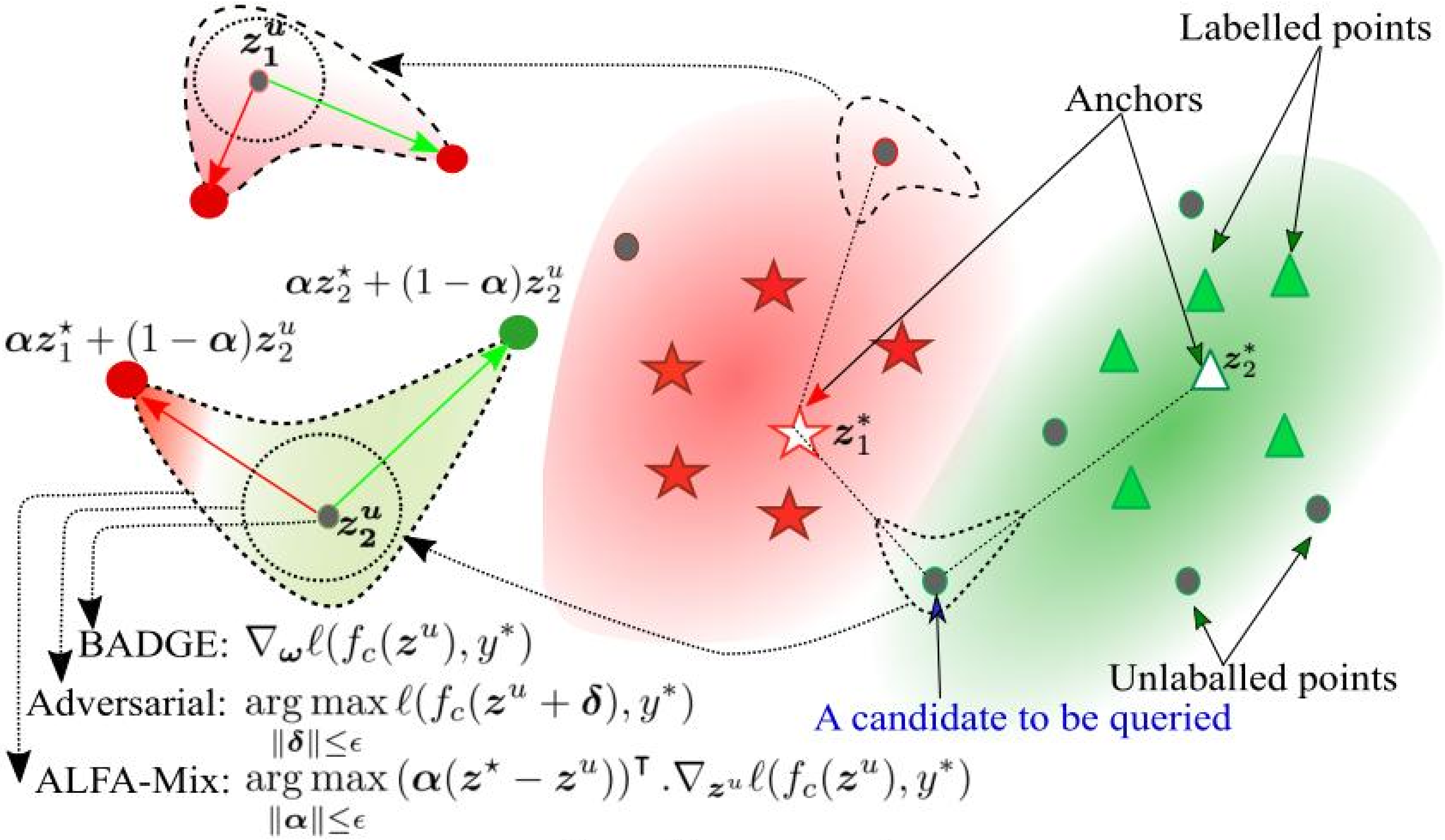
  Cluster the samples in $\mathcal{I}$ into B clusters.

  Select samples at the centre of each cluster ($\mathcal{C}$).

  $Y^{new} = \text{Query}(\mathcal{C})$.

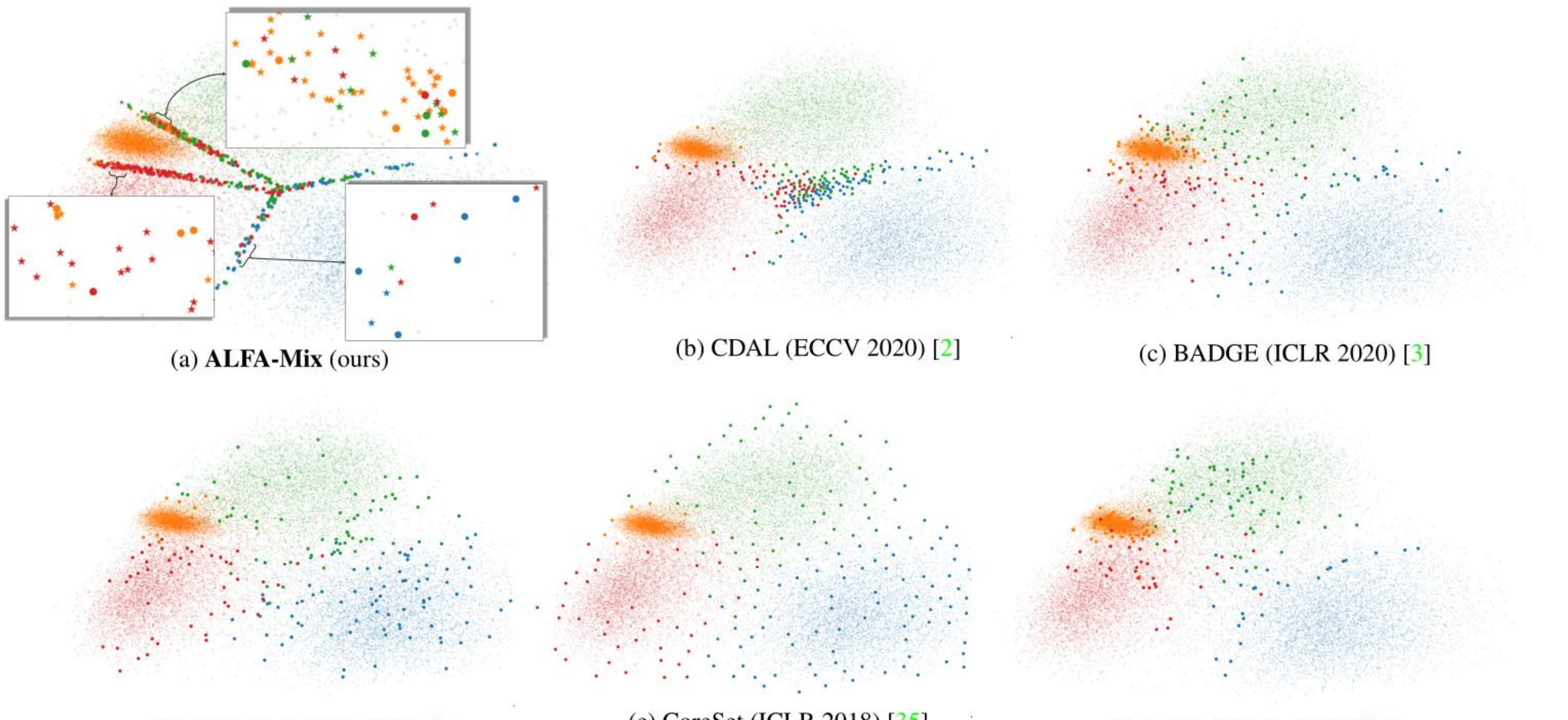
  $\mathcal{D}^l = \mathcal{D}^l \cup (\mathcal{C}, Y^{new})$, $\mathcal{D}^u = \mathcal{D}^u \setminus \mathcal{C}$.

Method



(a) Sampling strategies.

Experiments



(a) **ALFA-Mix** (ours) (b) CDAL (ECCV 2020) [2] (c) BADGE (ICLR 2020) [3]

(d) GCNAL (CVPR 2021) [6] (e) CoreSet (ICLR 2018) [35] (f) BALD (ICML 2017) [16]

Figure 2. Visualization of sample selection behaviours of various AL methods in the latent space (see the Appendix for additional methods). The larger dots represent the selected samples to label; smaller dots represent unlabelled ones. Our approach finds a candidate set (demonstrated by stars in 2a) of unlabelled instances with inconsistencies in their label prediction when interpolated with labelled representations. It selects a diverse set of samples lying close to the all four borders for the labelling (with three zoom-in windows). The demonstration problem is that of identifying 4 classes from MNIST (illustrated above by 4 colours) using a MLP. An initial training set of 200 randomly selected points and their labels was provided, with each method given a budget of 200 additional labels. The features are projected to two-dimensions for visualization.

- **Random**: a simple baseline that randomly selects B samples from the unlabelled pool at each round.
- **Entropy** [40]: A conventional AL approach that picks unlabelled instances with highest entropy.
- **BALD** [16]: An uncertainty model relying on Bayesian approaches that selects a set of samples with the highest mutual information between label predictions and posterior of the model approximated using dropout (Figure 2f).
- **Coreset** [35]: An approach based on the core-set technique that chooses a batch of diverse representative samples of the whole unlabelled set (Figure. 2e).
- **Adversarial Deep Fool** [13]: An uncertainty method that utilises deep fool attacks to select a batch of unlabelled samples whose predictions flip with small perturbations.
- **GCNAL** [6]: A model-based approach that learns a graph convolutional network to measures the relation between labelled and unlabelled instances (Figure. 2d)[4].
- **BADGE** [3]: A hybrid approach that queries the centroids obtained from the clustering of the gradient embeddings (Figure. 2c).
- **CDAL** [2]: A hybrid approach that exploits the contextual information in the predicted probabilities to choose samples with varied contexts (Figure. 2b)

Experiments

experiments in 30 different settings on multiple datasets

dataset——MNIST, EMNIST, CIFAR10, CIFAR100, MiniImageNet, DomianNet-Real, Two subsets of DomainNet-Real for image datasets, Two more non-visual datasets from the OpenML5 repository.

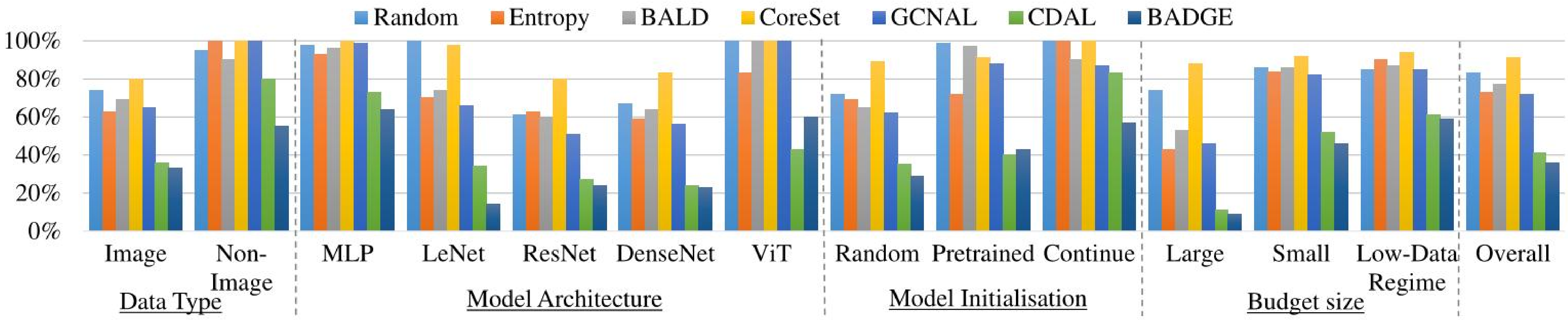


Figure 3. A summary of the performance of our proposed AL method (ALFA-Mix) compared with state-of-the-art across all the 30 settings considered. Each bar represents the percentage of AL rounds in which our approach outperforms others (lower indicates stronger baseline). It is worth noting that our approach (ALFA-Mix) under-performs others in close to zero cases.

Experiments

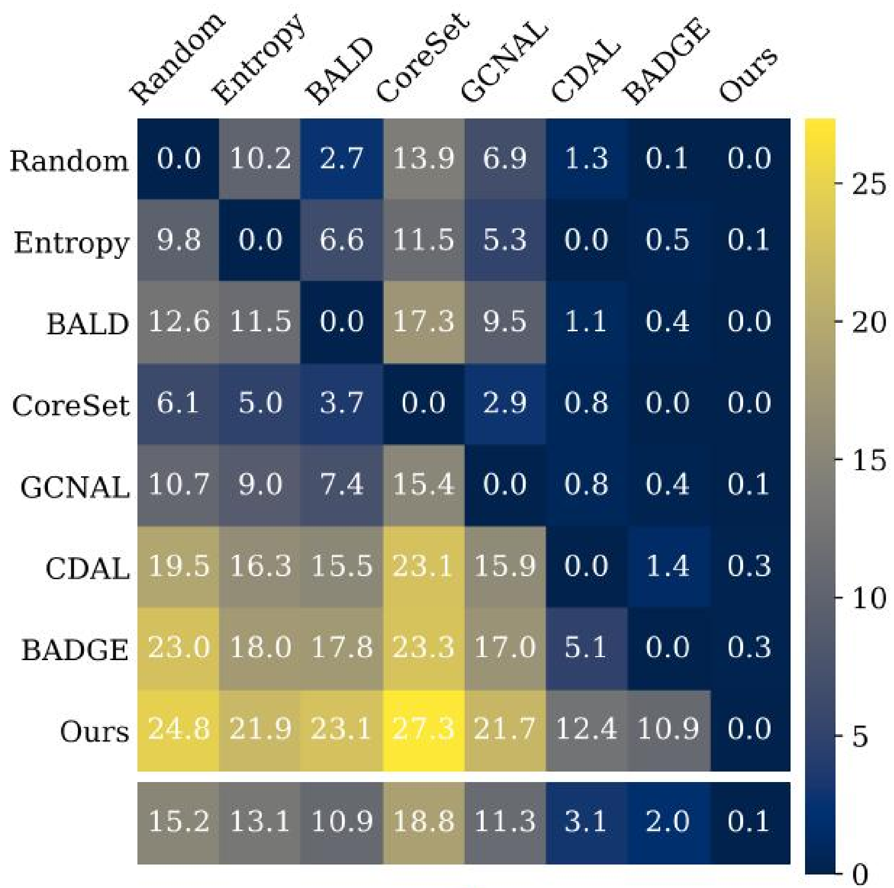


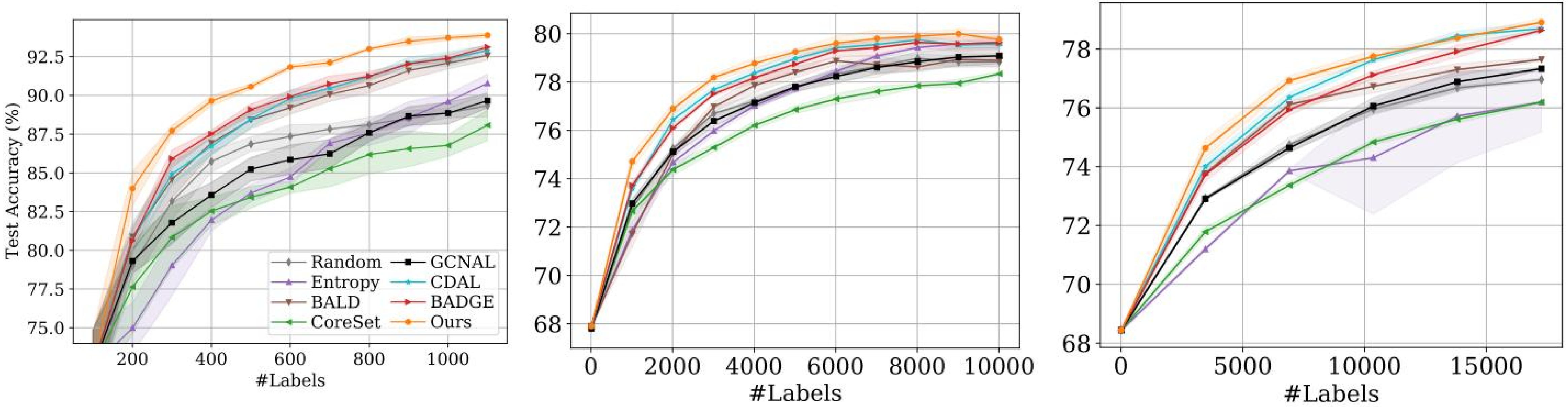
Figure 4. Pairwise comparison [3] of different approaches. Lower values shown at each column reveal the better performances of that AL method across all the experiments. Maximum value of each cell is 30, which represents the number of experimental settings.

HMDB: a large video database for human motion recognition

	AL Rounds			
Method	**204***	**408**	**765**	**1530**
MViT (initial accuracy with 102 instances: 50.9±1.2)				
Random	56.7±1.4	64.1±1.2	72.0±1.1	75.3±0.4
Entropy [40]	55.5±0.6	65.5±0.3	70.2±2.0	76.5±0.7
BALD [16]	56.7±0.4	65.5±0.6	72.4±1.3	76.6±1.8
CoreSet [35]	59.3±1.3	65.8±1.2	72.8±1.6	78.5±0.7
GCNAL [6]	54.9±1.4	63.3±2.2	70.8±1.4	77.0±1.3
CDAL [2]	60.9±0.1	67.2±0.4	74.6±0.2	78.4±0.5
BADGE [3]	60.6±1.3	67.3±0.2	73.2±1.1	**78.7**±0.2
Ours	**62.5**±0.6	**69.4**±0.7	**75.1**±0.3	78.3±0.1

Table 1. Top-1 test accuracy of various AL methods on HMDB [23].
* Values on top of each column reveal the size of the labelled set at the end of each round.

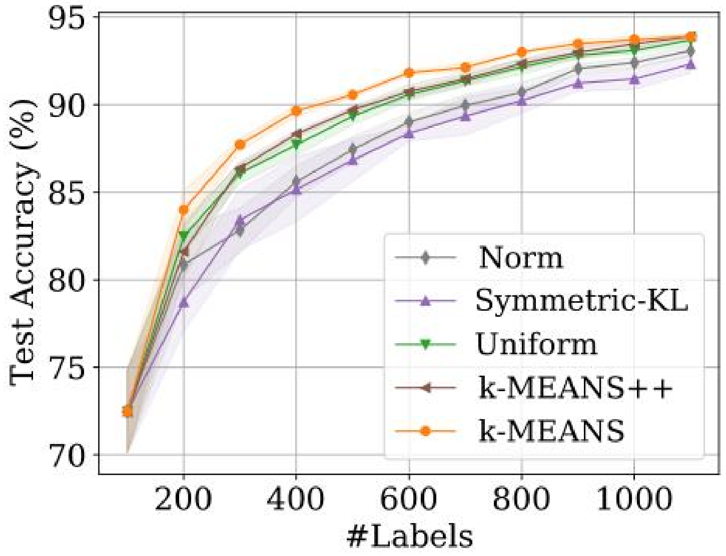
Experiments



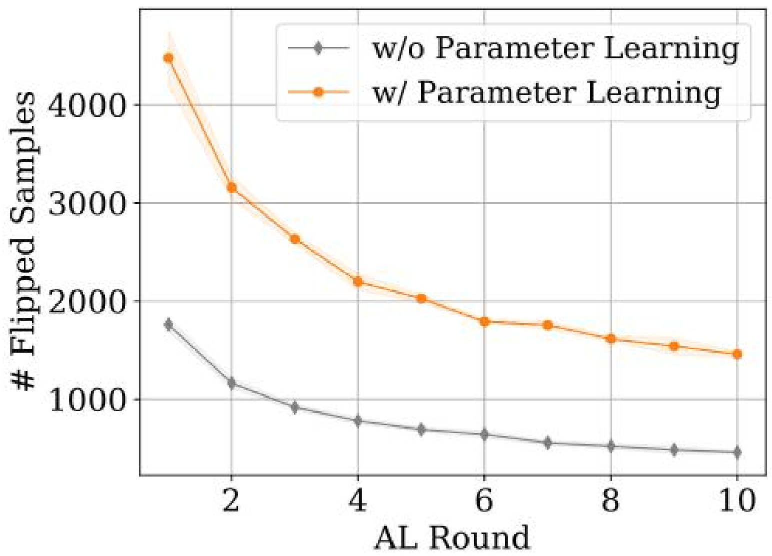
(a) MNIST (MLP) (b) MiniImageNet (ViT-Small) (c) DomainNet-Real (ViT-Base)

Figure 5. Test accuracy plots across some of the employed settings. Each experiment has been repeated 5 times.

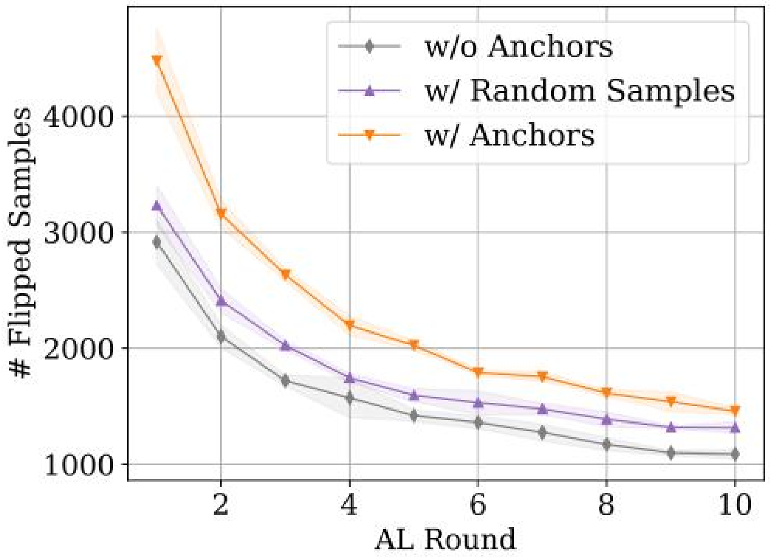
Ablation Study



(a) Diversity impact of the sample selection from the candidate set ($\mathcal{I}$).
*k-MEANS is our proposed full model.



(b) Number of unlabelled samples whose predictions flip with and without learning the interpolation parameter α.



(c) The impact of anchors on identifying samples whose labels flip during the interpolation.

Figure 6. Ablations of our AL approach. Experiments are conducted on MNIST datasets using an MLP model and a small AL budget.

Thanks