[Active Learning by Feature Mixing]

<Problem Definition>

- labeled data $D^l = \{(x_i, y_i)\}_{i=0}^{M}$
- unlabeled data $D^u \xrightarrow{\text{extract}} B$ instances for query

Learner : $f = f_c \odot f_e$ parameterised by $\theta = \{\theta_e, \theta_c\}$

$f_e : \mathcal{X} \to \mathbb{R}^D$

↳ backbone which encodes the input to a D-dimensional representation in a latent space

$z = f_e(x; \theta_e)$ ← encoded input to the latent space

$f_c : \mathbb{R}^D \to \mathbb{R}^K$

↳ classifier (ex. multi-layer perceptron (MLP) that maps the instances from their representations to their corresponding logits which can be converted to class likelihoods by

$p(y|z;\theta) = \text{Softmax}(f_c(z,\theta))$ ← class likelihoods of instances mapped from their representations.

✓ Optimise parameters end-to-end by minimizing cross-entropy loss over the labelled set: $\mathbb{E}_{(x,y)\sim D^l}[\ell(f_c \odot f_e(x;\theta), y)]$

ℓ: categorical cross entropy

✓ The prediction of the label (i.e. pseudo-label) for an unseen instance:

$$y_z^* = \underset{y}{\text{argmax}}\ f_c^y(z;\theta_c)$$

(y^*)

$z = f_e(x, \theta_e)$

f_c^y : logit output for class y

✓ The logit of the predicted label: $f_c^*(z) := f_c^{y_z^*}(z)$

✓ The set of representations of the unlabelled data: $Z^u = \{f_e(x), \forall x \in D^u\}$

✓ The set of representations of the labelled data: $Z^l = \{f_e(x), \forall x \in D^l\}$

z^*: the average representation of labelled samples per class
(= anchor)

Z^*: The set of anchors for all classes.
(= representatives of the labelled instances)

⟨Feature Mixing⟩

Intuition: The model's incorrect prediction is mainly due to novel "features" in the input that are not recognisable.

→ We use interpolation as a way to explore novel features in the vicinity of each unlabeled point.

Interpolation between the representations of the unlabelled and labelled instances, z^u and z^*, respectively:

$\tilde{z}_\alpha$ z^u z^*

$\tilde{z}_\alpha = \alpha z^* + (1-\alpha) z^u \qquad \alpha \in [0,1)^D$: interpolation ratio

$z \sim p(z \mid z^u, Z^*, \alpha) \equiv \alpha z^* + (1-\alpha) z^u, \; z^* \sim Z^*$ ← can be seen as sampling a new instance without explicitly modelling the joint probability of the labelled and unlabelled instances.

Consider how model's prediction changes:

{ ① The change in pseudo-label (i.e. y^*) for the unlabelled instance
② The loss incurred with the interpolation

The model's loss for predicting the pseudo-label (y^*) of an unlabelled instance at its interpolation ($\tilde{z}_\alpha$) with a labeled one:

$$\ell(f_c(\tilde{z}_\alpha), y^*) \approx \ell(f_c(z^u), y^*) + (\alpha(z^* - z^u))^T \cdot \nabla_{z^u} \ell(f_c(z^u), y^*)$$

↑ for sufficiently small $\|\alpha\| \leq \epsilon$ is almost exact

Consequently, for the full labelled set,
by choosing the max loss from both sides we have:

$$\max_{z^* \sim Z^*} [\ell(f_c(\tilde{z}_\alpha), y^*)] - \ell(f_c(z^u), y^*)$$
$$\approx \max_{z^* \sim Z^*} [(\alpha(z^* - z^u))^T \cdot \nabla_{z^u} \ell(f_c(z^u), y^*)]$$

When performing interpolation, the change in the loss is proportionate to two terms:

① The difference of features of z^* and z^u proportionate to their interpolation α.
↳ determines which features are novel and how their value could be different between the labelled and unlabelled instance.

② The gradient of the loss with respect to the unlabelled instance.
↳ determines the sensitivity of the model to those features.

∴ If the features of the labelled and unlabelled instances are completely different but the model is reasonably consistent, there is ultimately no change in the loss, and hence those features are not considered novel to the model.

α : · input specific
· determines the features to be selected

⟨ Optimising the Interpolation Parameter α ⟩

The objective for choosing α

$$\alpha^* = \underset{\|\alpha\| \leq \epsilon}{\operatorname{argmax}} \left(\alpha (z^* - z^u)\right)^T \nabla_{z^u} \ell(f_c(z^u), y^*)$$

ϵ : hyperparameter governing the magnitude of the mixing

chooses the hardest case of α for each unlabeled instance and anchor

Approximate solution to this optimization from using dual norm formulation :

$$\alpha^* \approx \epsilon \frac{\|(z^* - z^u)\|_2 \nabla_{z^u} \ell(f_c(z^u), y^*)}{\|\nabla_{z^u} \ell(f_c(z^u), y^*)\|_2} \oslash (z^* - z^u)$$

$\oslash$: element-wise division

⟨ Candidate Selection ⟩

It is reasonable to choose instances to be queried <u>whose loss substantially change with interpolation</u>

↳ instances for which the model's prediction change and has novel features.
= placed close to the decision boundary in the latent space.

Candidate set : $I = \{ z^u \in Z^u \mid \exists z^* \in Z^*, f_c^*(\tilde{z}_\alpha) \neq y^*_{z^u} \}$

- The size of selected set I could potentially be larger than the budget B.
- Seek diverse samples since most instances in I could be chosen from the same region. (i.e. they might share the same novel features.)

K-Means → Cluster the instances in I into B groups based on their feature similarities and further choose the closest samples to the centre of each cluster to be labelled by oracle.

<Algorithm>

D^l : Initial labelled set

D^u : Unlabelled pool

B : labelling budget at each round

ϵ : mixing parameter

for $i=1$ to max_rounds do :

Train the model f using the labelled data D^l

Initialize Z^* on the representations of D^l

$I = \{\}$

for $x^u \in D^u$ do

$z^u = f_e(x^u)$

for $z^* \in Z^*$ do # Z^* (z^*) : The set of anchors for all classes (representatives of labelled instances)

Calculate α^* using ϵ

$\tilde{z} = \alpha^* z^* + (1-\alpha^*) z^u$ # interpolation

if $\underset{y}{\text{argmax}}(f_c^y(z_u)) \neq \underset{y}{\text{argmax}}(f_c^y(\tilde{z}))$

then

$I = I \cup (x^u, z^u)$

Break

클래스가 무엇이 있고 모든 클래스에 대해 한 unlabelled instance로 interpolation을 하라?

Cluster the samples in I into B clusters.

Select samples at the centre of each cluster (C)

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

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