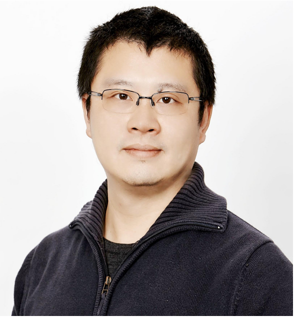
Diversity Enhanced Active Learning with Strictly Proper Scoring Rules

Wei Tan, Lan Du, Wray Buntine



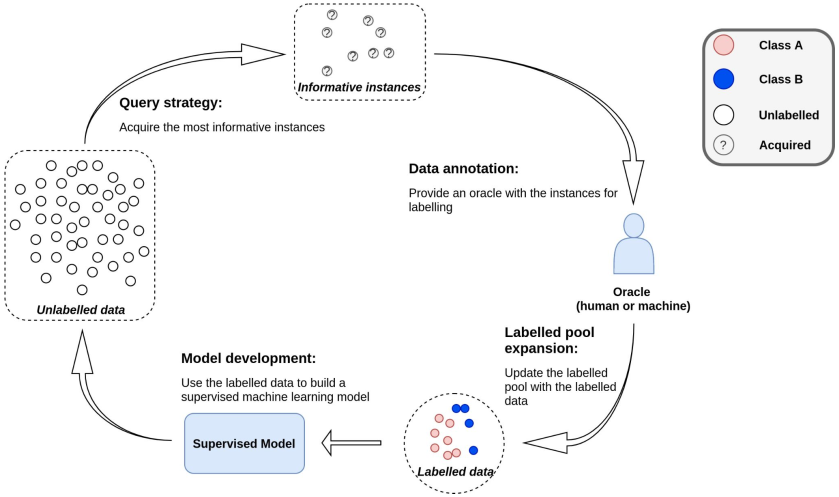
MONASH University

MONASH University

MONASH University

Active Learning

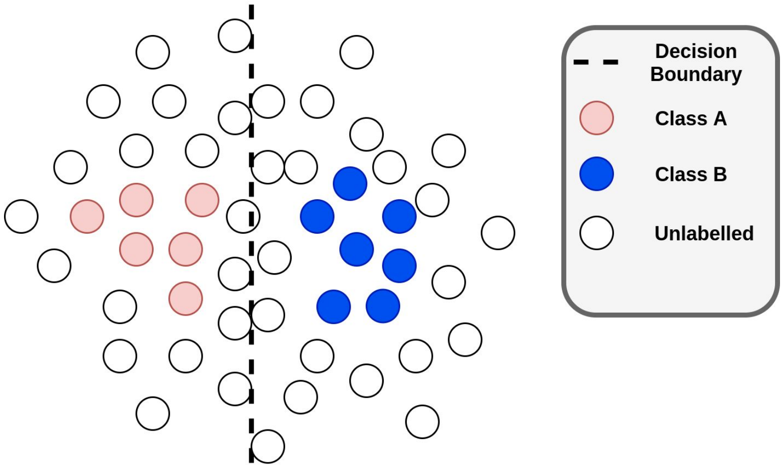
- **Purposes**:
  - recognise the most informative instances to an oracle for labelling
  - minimise the cost of labelling while preserving the model performance



Pool-based active learning

Active Learning

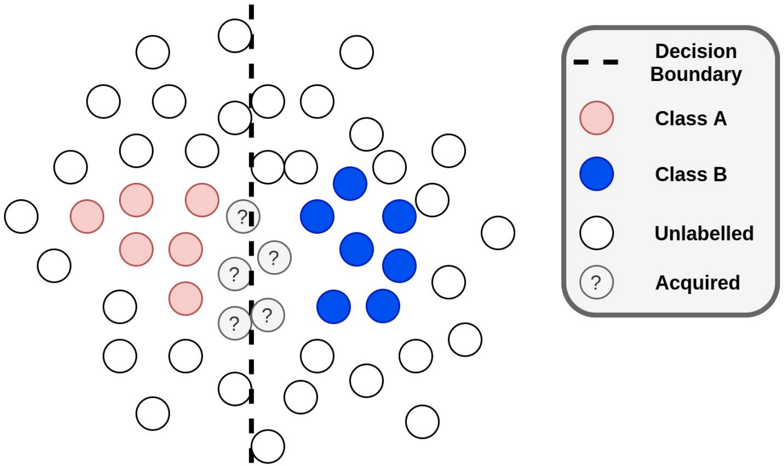
- Purposes:
    - recognise the most informative instances to an oracle for labelling
    - minimise the cost of labelling while preserving the model performance
- **Issues**:
    - ➢ what is a good uncertainty-based acquisition function



Acquisition Example: Two classes are predicted by the decision boundary of the trained model

Active Learning

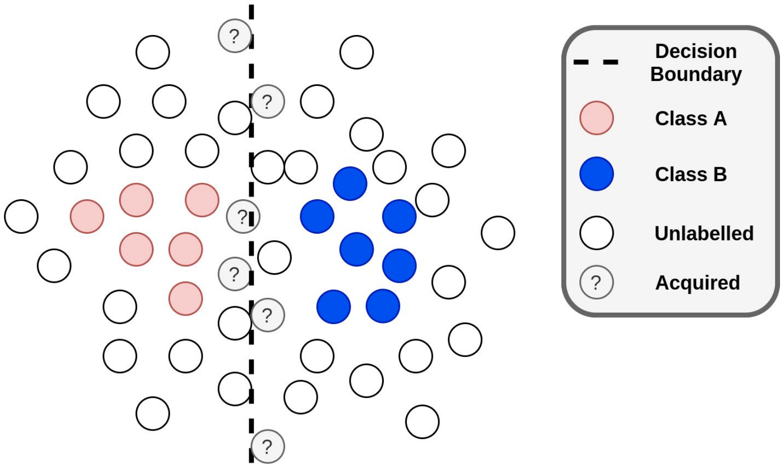
- Purposes:
  - recognise the most informative instances to an oracle for labelling
  - minimise the cost of labelling while preserving the model performance
- **Issues**:
  - ➢ what is a good uncertainty-based acquisition function



Uncertainty-based: acquiring unlabelled instances near the decision boundary

Active Learning

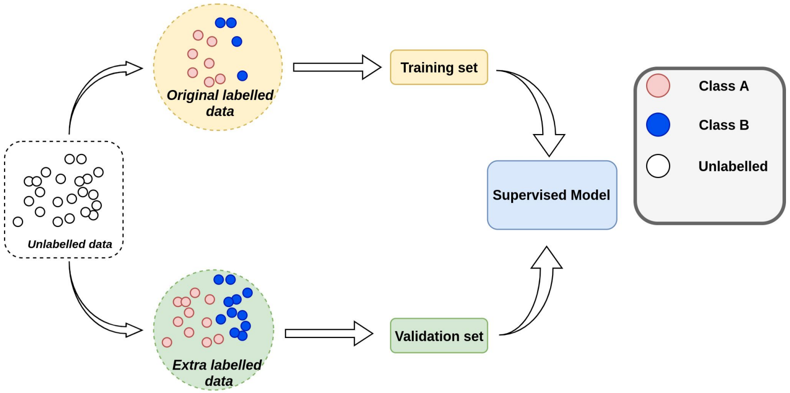
- Purposes:
  - recognise the most informative instances to an oracle for labelling
  - minimise the cost of labelling while preserving the model performance
- **Issues**:
  - what is a good uncertainty-based acquisition function
  - ➢ also, how to enhance the diversity



Uncertainty & Diversity: acquiring a diverse set of instances near the decision boundary

Active Learning

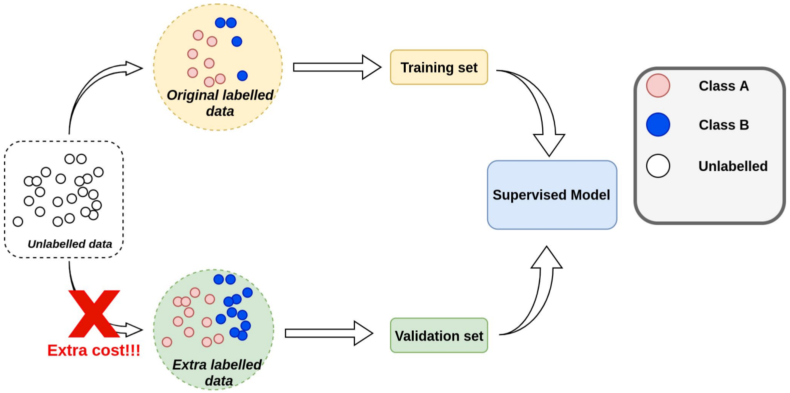
- Purposes:
  - recognise the most informative instances to an oracle for labelling
  - minimise the cost of labelling while preserving the model performance
- **Issues**:
  - what is a good uncertainty-based acquisition function
  - also, how to enhance the diversity
  - ➢ how to use a validation set required for deep learning



Model Training Setup: most researchers uses the extra labelled data for the validation set in active learning

Active Learning

- Purposes:
  - recognise the most informative instances to an oracle for labelling
  - minimise the cost of labelling while preserving the model performance
- **Issues**:
  - what is a good uncertainty-based acquisition function
  - also, how to enhance the diversity
  - ➢ how to use a validation set required for deep learning



Model Training Setup: most researchers uses the extra labelled data for the validation set in active learning

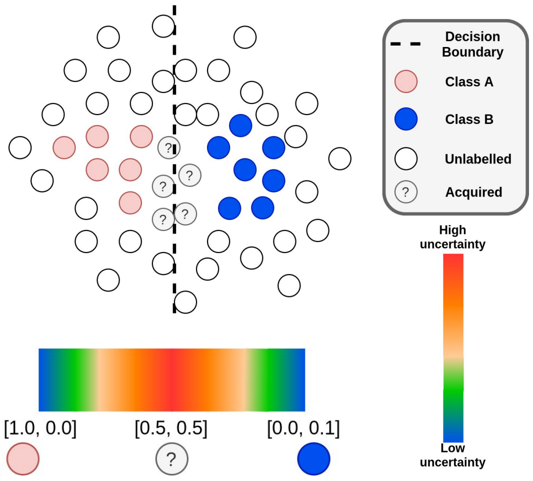
Background

- Uncertainty-based approaches:
  - ➢ Maximum Entropy (Holub et al., 2008)

Acquisition Function:

measure of uncertainty

$$x_{ME} = \arg\max_x \Big(-\sum_i p(\hat{y}_i|x;\theta) \log p(\hat{y}_i|x;\theta)\Big)$$



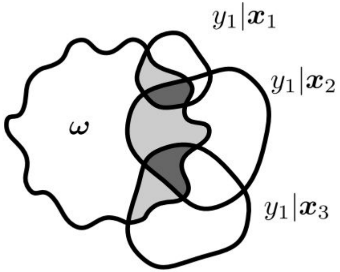
Measure of Uncertainty Diagram: Maximum Entropy acquires unlabelled instances with the high uncertainty. Maximum entropy fails when selecting instances in batches because the instances contain similar information.

Background

- Uncertainty-based approaches:
  - Maximum Entropy (Holub et al., 2008)
  - ➢ Bayesian Active Learning by Disagreement (Houlsby et al.,2011)

Acquisition Function:

$$x_{BALD} = \arg\max_{x} \overbrace{\Big(\underbrace{H(y|x,L)}_{\text{the entropy of Bayes optimal prediction}} - \mathbb{E}_{p(\theta|L)}[\underbrace{H(y|x,\theta,L)}_{\text{the expected entropy of the model prediction over the posterior of the model parameter}}]\Big)}^{\text{mutual information}}$$



$$\sum_i \mathbb{I}(y_i\,;\boldsymbol{\omega} \mid \boldsymbol{x}_i, \mathcal{D}_{\mathbf{train}}) = \sum_i \mu^*(y_i \cap \boldsymbol{\omega})$$

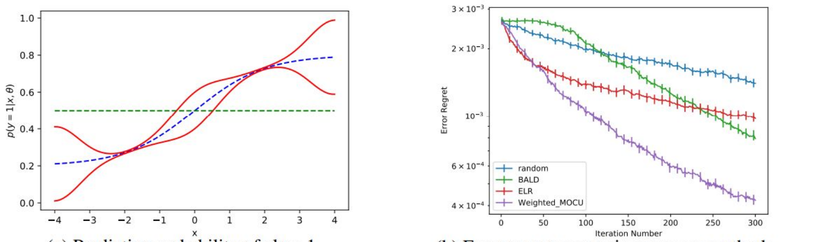
Mutual Information I-diagram: BALD acquires unlabelled instances with the high mutual information. Areas in grey contribute to the BALD score. BALD fails because the areas in dark grey are double-counted (Kirsch et al., 2019)

Background

- Uncertainty-based approaches:
  - Maximum Entropy (Holub et al., 2008)
  - Bayesian Active Learning by Disagreement (Houlsby et al.,2011)
  - ➢ Expected Error/Loss Reduction (Roy and McCallum, 2001; Zhao et al., 2021)

Acquisition Function:

$$\Delta Q(x|L) = \overbrace{\underbrace{Q(L)}_{\substack{\text{the expected loss difference}\\ \text{between the OBC and the}\\ \text{optimal classifier}}} - \mathcal{E}_{\Pr(y|L,x)}[\underbrace{Q(L \cup \{(x,y)\})}_{\substack{\text{the expected loss difference}\\ \text{between the OBC and the}\\ \text{optimal classifier given new data}}}]}^{\substack{\text{the expected reduction of the}\\ \text{OBC error given new data}}}$$



(a) Predictive probability of class 1 (b) Error regret comparison among methods

Figure 1: (a) Predictive probability of class 1 under uncertainty: the red lines indicate the upper and lower bounds of the predictive probability; the blue dash line is the mean of the predictive probability; the green dash line indicates that the probability is equal to 0.5. (b) Active learning performance comparison.

Uncertainty and Error Analysis: Provides an example of binary classification with one feature where both BALD and ELR methods fail (Zhao et al., 2021)

Bayesian Estimate of Mean Proper Scores

- Apply the Expected Error Reduction framework

$$\Delta Q(x|L) = Q(L) - \mathcal{E}_{\Pr(y|L,x)}[Q(L \cup \{(x,y)\})], \qquad (1)$$

- Define a better expected loss function with strictly proper scoring rules

$$\begin{aligned} Q_S(L) &= \mathcal{E}_{\Pr(x)\Pr(\theta|L)}[\mathcal{E}_{\Pr(y|\theta,x)}[S(\Pr(\cdot \mid \theta, x), y) - S(\Pr(\cdot \mid L, x), y)]] & (3) \\ &= \mathcal{E}_{\Pr(x)\Pr(\theta|L)}[B(\Pr(\cdot \mid L, x), \Pr(\cdot \mid \theta, x))] & (4) \\ &= \mathcal{E}_{\Pr(x)}[\mathcal{E}_{\Pr(\theta|L)}[G(\Pr(\cdot \mid \theta, x))] - G(\Pr(\cdot \mid L, x))] & (5) \end{aligned}$$

(3) Arbitrary strictly proper scoring rule

(4) Bregman divergence

(5) Arbitrary strictly convex function

- The acquisition function in a general form can be defined as

$$\Delta Q_S(x|L) = \mathcal{E}_{\Pr(x')}[\mathcal{E}_{\Pr(y|L,x)}[G(\Pr(\cdot \mid L, (x,y), x'))] - G(\Pr(\cdot \mid L, x'))] \qquad (6)$$

Bayesian Estimate of Mean Proper Scores

- Algo 2 shows the Non-batch approach

$$\Delta Q_S(x|L) \quad = \quad \mathcal{E}_{\Pr(x')}[\mathcal{E}_{\Pr(y|L,x)}[G(\Pr(\cdot \mid L,(x,y),x'))] - G(\Pr(\cdot \mid L,x'))] \qquad (6)$$

- Apply two strictly convex functions
  - CoreMse (Squared error scoring rule as Brier score)

$$G_{MSE}(q(\cdot)) = \sum_y q(y)^2 - 1$$

  - CoreLog (logarithmic scoring rule)

$$G_{log}(q(\cdot)) = -I(q(\cdot))$$

Algorithm 2 Estimate of $\text{argmax}_{x \in U} \, \Delta Q(x|L)$

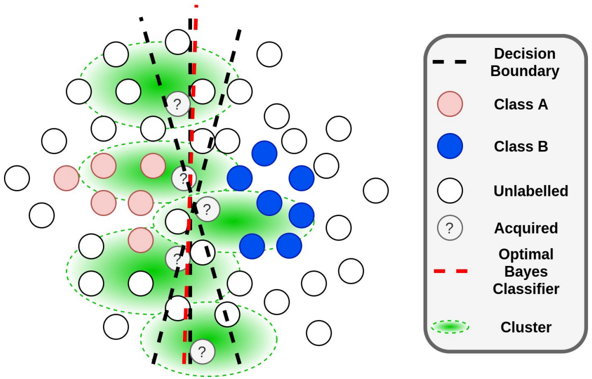
Require: unlabelled pool U, estimation pool X

1: **for** $x \in U$ **do**
2: $\quad Q_x = 0$
3: $\quad$ **for** $x' \in X$ **do**
4: $\quad\quad Q_x \mathrel{+}= \Delta Q(x|L, x')$
5: **return** $\text{argmax}_{x \in U} \, Q_x$

Algo 2 represents a single instance acquisition based on the maximum expected loss reduction given the new data

Bayesian Estimate of Mean Proper Scores

- Batch mode algorithm
  - Generate the vector representation based on the expected loss reduction given new data
  - Apply the K-means clustering to find a diverse batch of samples with the most uncertainty information



Algorithm 3 Finding a diverse batch

Require: unlabelled pool U, batch size B
Require: estimation pool X, top fraction T

$\forall_{x \in U} Q_x = 0$
for $x \in U, x' \in X$ **do**
  $Q_x \mathrel{+}= vec_{x,x'} = \Delta Q(x|L, x')$
$V \leftarrow topk(Q, T * |U|)$
$batch = \emptyset$
$centroids$ = k-Means centers $(vec_{x \in V}, B)$
for $c \in centroids$ **do**
  $batch \cup= \{\text{argmin}_{x \in V} ||c - vec_x||\}$
return $batch$

Algo 3 represents the batch instances acquisition based the representation of the expected loss reduction

Experiment Setup

- Dataset and Model

TABLE 3.2: Datasets and the used language model

Dataset	Unlabelled/Test sizes	#Classes	Lang. Model	Initial labelled size
AG NEWS	120,000 / 7,600	4	DistilBERT	26
PUBMED 20K RCT	15,000 / 2,500	5	DistilBERT	26
IMDB	25,000 / 25,000	2	DistilBERT	26
SST-5	8544 / 2210	5	DistilBERT	26

- Dynamic validation:
  - generate a new train/validation pair for each element of the ensemble

- Baseline
  - Random
  - Maximum Entropy (Holub et al., 2008)
  - BALD (Houlsby et al.,2011)
  - MOCU/ELR (Zhao et al., 2021)
  - WMOCU (Zhao et al., 2021)
  - BADGE (Ash et al., 2020)
  - ALPS (Yan et al., 2020)

Experiment Result

- Model Performance for non-batch (batch size 1)
    - Learning curve
    - Pairwise comparison matrix

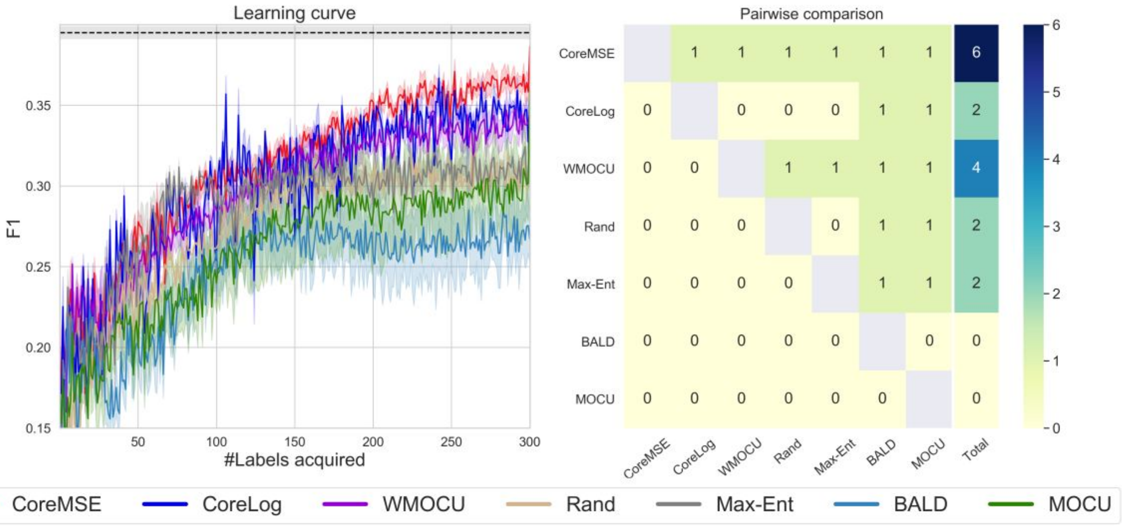


Figure 1: Performance on SST-5 dataset. The left half illustrates the learning curve, while the right half illustrates the matrix of paired comparisons.

Experiment Result

- Model Performance for batch mode (batch size 50)
    - Learning curve
    - Pairwise comparison matrix

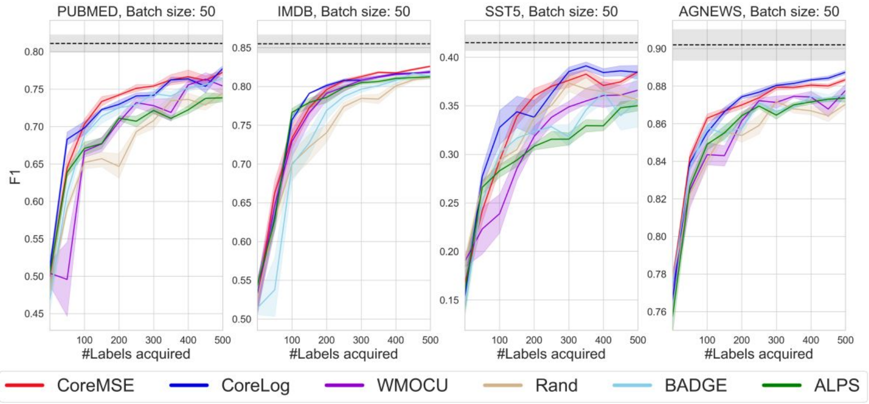
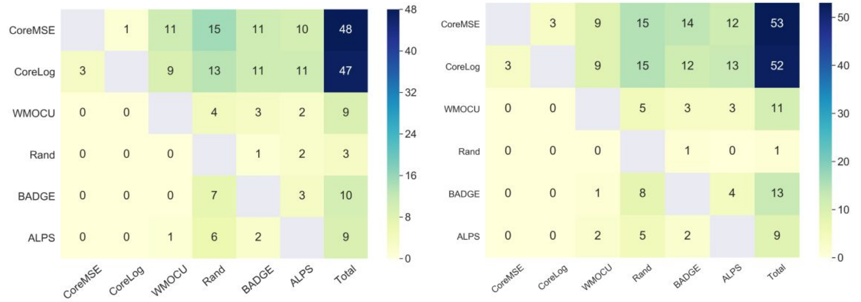


Figure 2: Learning curves of batch size 50 for PUBMED, IMDB, SST-5 and AG NEWS.



(a) F1-based pairwise comparison (b) Accuracy-based pairwise comparison

Figure 3: Pairwise comparison matrices of batch active learning strategies.

Experiment Result

- Model performance comparison by different validation setup
  - Dynamic validation set
  - Constant validation set
  - No validation set

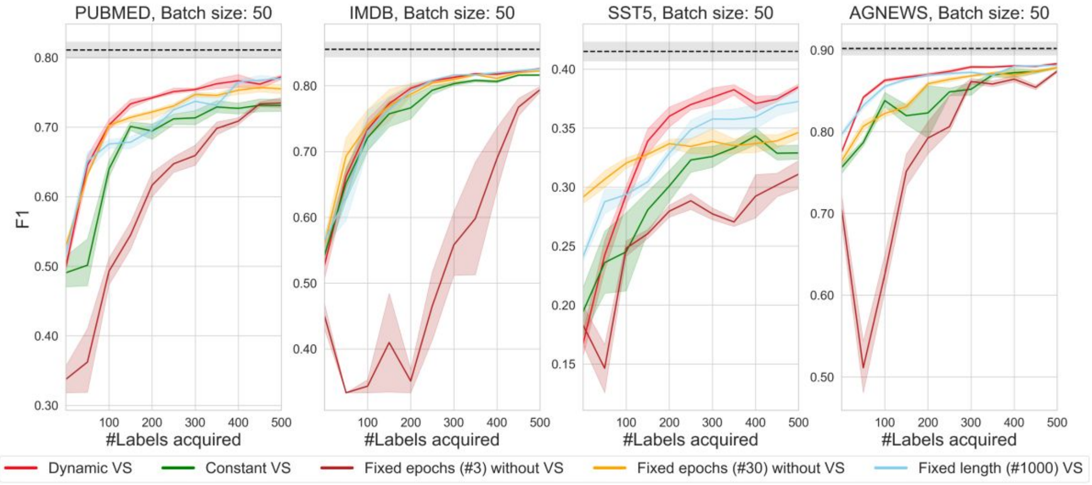


Figure 6: Learning curves of the model training with a dynamic validation set, constant validation set, fixed # epochs without validation set, fixed length # labels validation set for CoreMSE

Future work

- Utilise the computation cost via Monte Carlo dropout
- Extend our algorithm on the other tasks such as image classification, NER etc

Thank you !!!