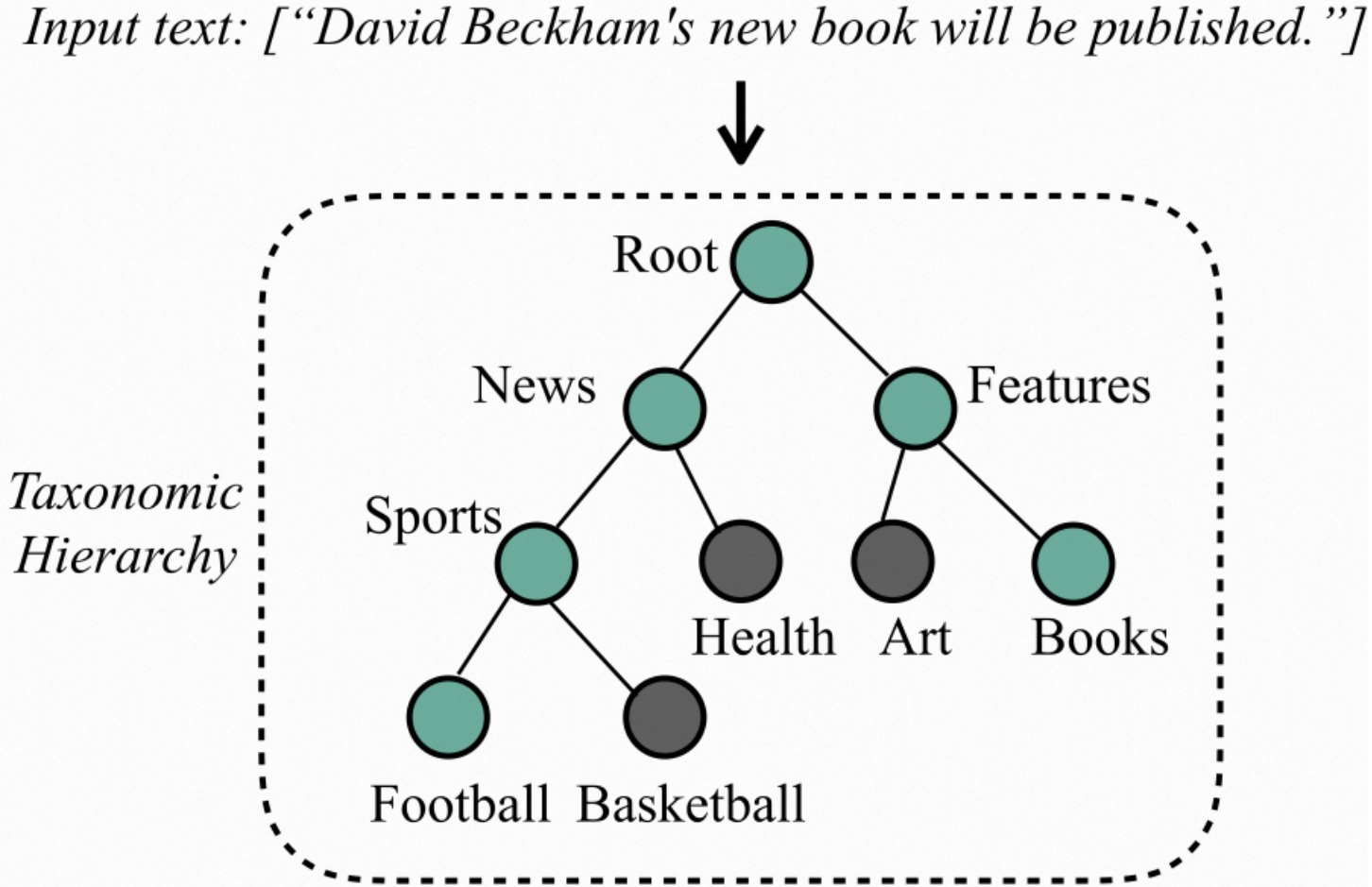
Well Begun is Half Done: An Implicitly Augmented Generative Framework with Distribution Modification for Hierarchical Text Classification

Huawen Feng, Jingsong Yan, Junlong Liu, Junhao Zheng,
School of Computer Science and Engineering
South China University of Technology (SCUT)

Outlines

01 Long-tailed Distributions in HTC

Hierarchical Text Classification (HTC) is a challenging task which aims to extract the labels in a tree structure corresponding to a given text.



01 Long-tailed Distributions in HTC

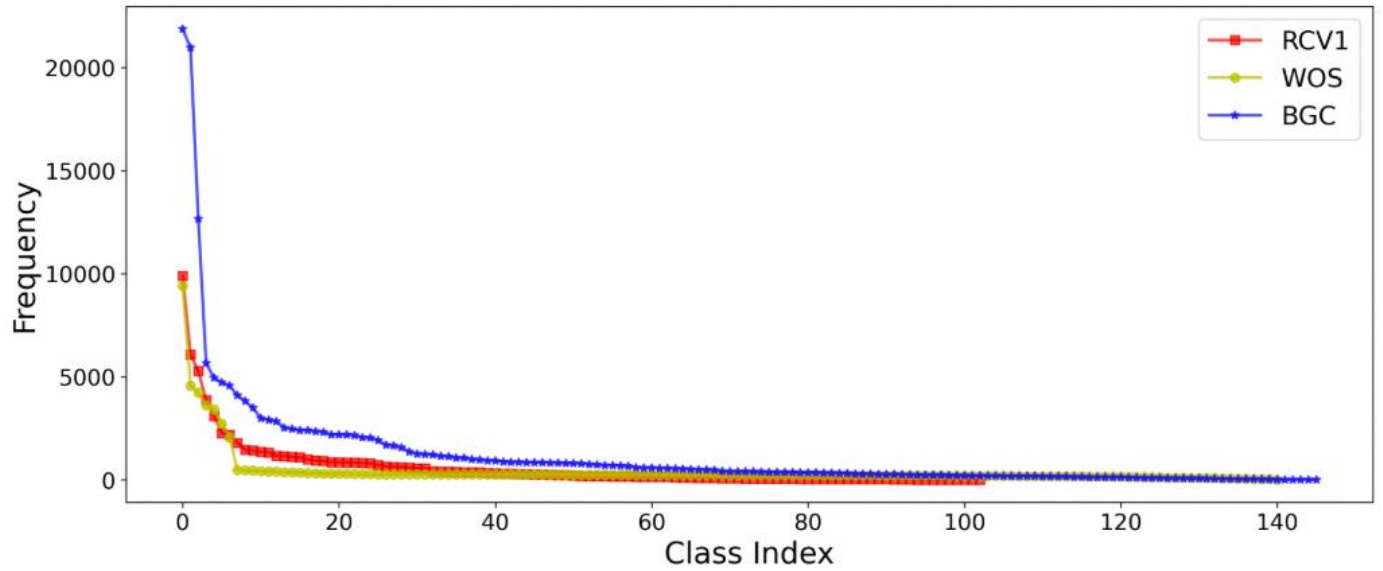


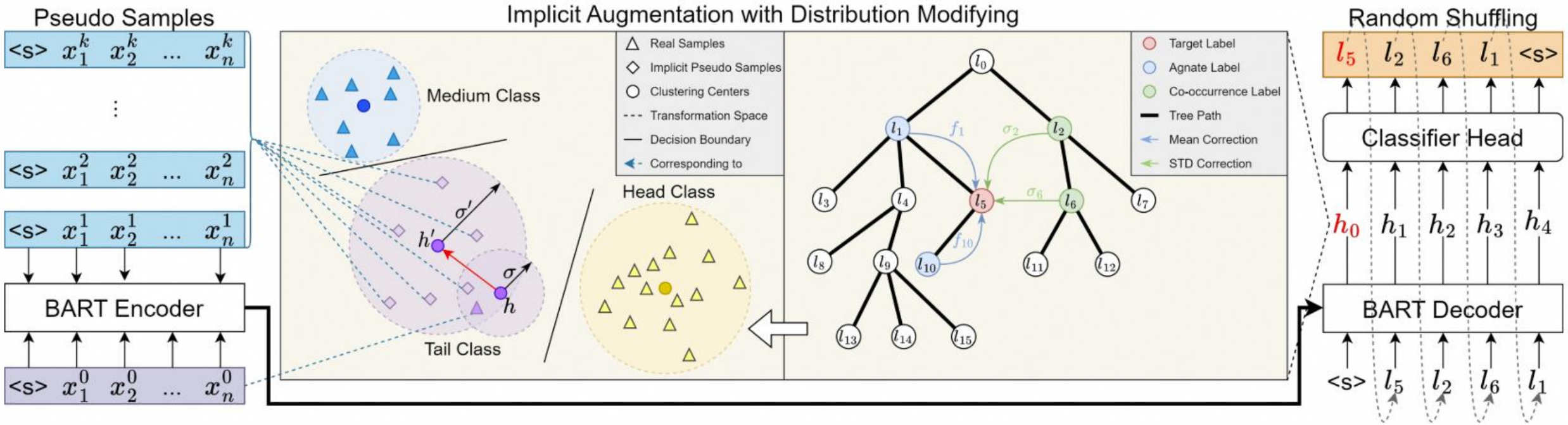
Figure 1: The diagram of the imbalance of HTC. Only a few labels (head classes) frequently appear, while others (medium and tail classes) rarely appear. The global frequency of labels shows a long-tailed distribution.

02 Challenges

- Due to long-tailed distribution, the head classes are always allocated with large weights, which distorts the label space.

- Current methods for long-tailed distribution are inapplicable to multi-label classification because they have to handle each class separately.

03 Implicit Augmentation



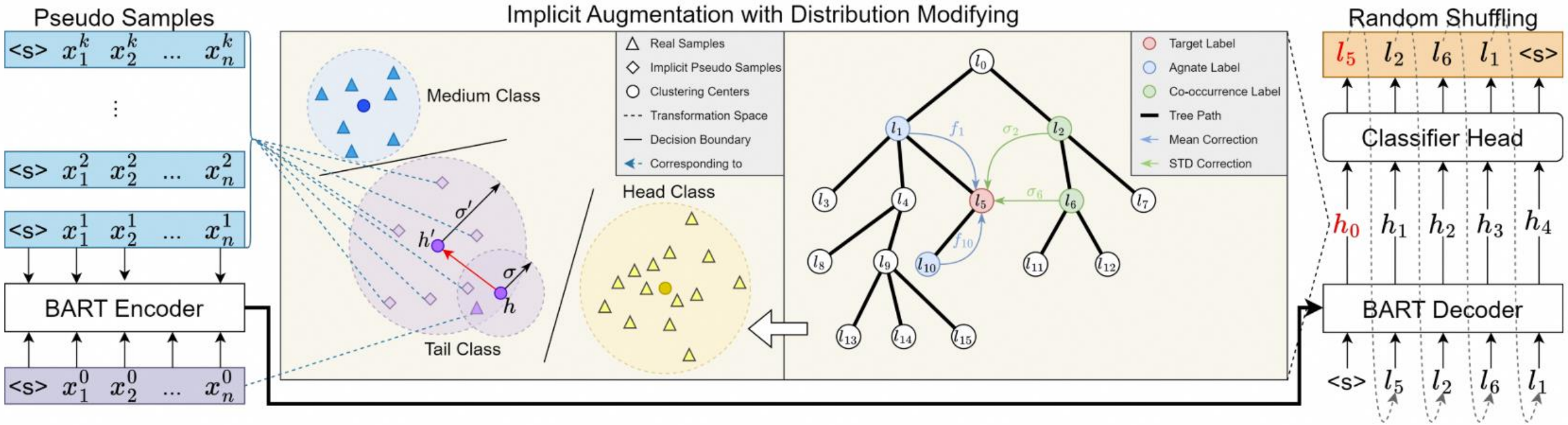
$$\widetilde{h}^p_{(q,t)} \sim \mathcal{N}(h_{(q,t)}, \sigma_k)$$

$$k = Y(q,t)$$

$$\boldsymbol{h_k} = \frac{1}{|S_k|} \sum_{(q,t)\in S_k} h_{(q,t)}$$

$$\sigma_k(a,b) = \frac{1}{|S_k|-1} \sum_{(q,t)\in S_k} (h_{(q,t)}(a) - \boldsymbol{h_k}(a)) (h_{(q,t)}(b) - \boldsymbol{h_k}(b))$$

03 Implicit Augmentation



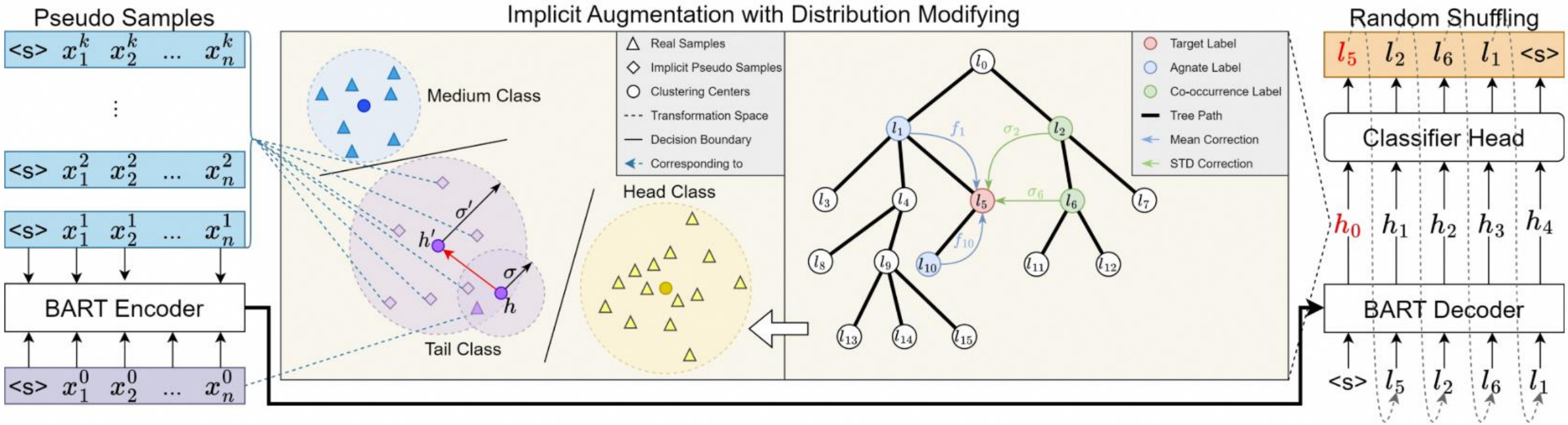
$$\widetilde{h}^p_{(q,t)} \sim \mathcal{N}(h_{(q,t)}, \sigma_k)$$

$$k = Y(q,t)$$

$$\boldsymbol{h_k} = \frac{1}{|S_k|} \sum_{(q,t)\in S_k} h_{(q,t)}$$

$$\sigma_k(a,b) = \frac{1}{|S_k|-1} \sum_{(q,t)\in S_k} (h_{(q,t)}(a) - \boldsymbol{h_k}(a)) (h_{(q,t)}(b) - \boldsymbol{h_k}(b))$$

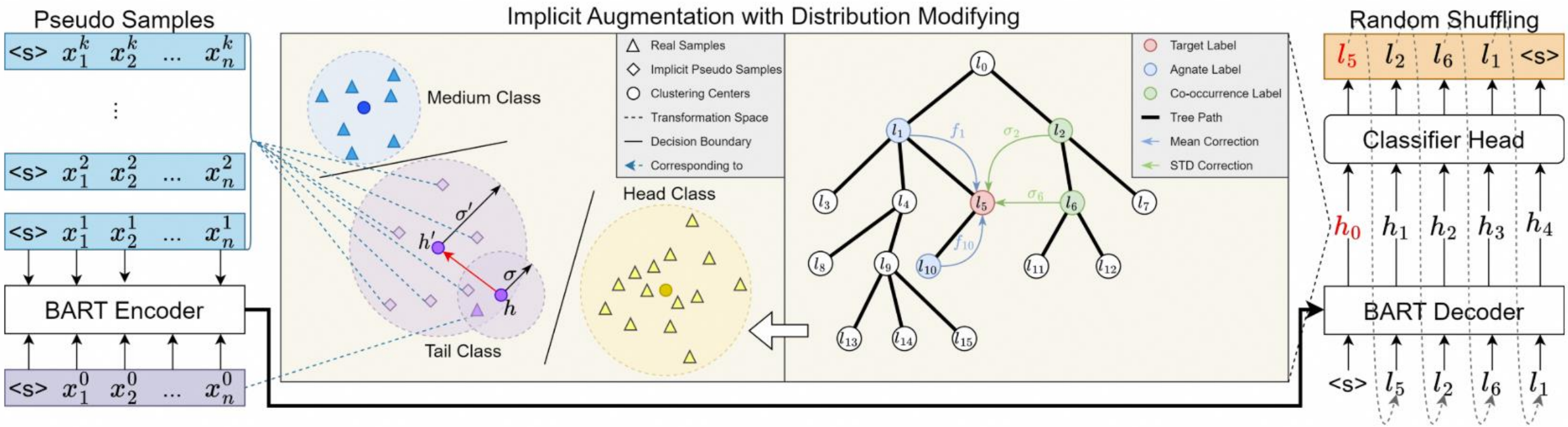
03 Implicit Augmentation



$$\triangle h_{(q,t)} = \sum_{i \in S_k^{agn}} w_{ik} \boldsymbol{h_i}$$

$$w_{ik} = Norm(\frac{1}{\log |r_i - r_k| + 1})$$

03 Implicit Augmentation



$$\triangle\sigma_k = \sum_{i \in S_k^{co}} w_{ik}\sigma_k$$

$$L_\sigma = \sum_{k=1}^{N_c} \frac{1}{|S_k^{co}|} \sum_{i \in S_k^{co}} (w_{ik} - \log X_{ik})^2$$

$$\widetilde{h}^p_{(q,t)} \sim \mathcal{N}(h_{(q,t)} + \alpha \triangle h_{(q,t)}, \beta(\sigma_k + \triangle\sigma_k))$$

03 Implicit Augmentation

(1)
$$L_{aug} = -\frac{1}{P}\frac{1}{Q}\frac{1}{T}\sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{t=1}^{T}\log g^{p}_{(q,t)}$$
$$= -\frac{1}{P}\frac{1}{Q}\frac{1}{T}\sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{t=1}^{T}\log\frac{e^{w_k\widetilde{h}^{p}_{(q,t)}}}{\sum_{j=1}^{N_C}e^{w_j\widetilde{h}^{p}_{(q,t)}}}$$

(2)
$$L_{\substack{aug\\P\to\infty}} = \frac{1}{Q}\frac{1}{T}\sum_{q=1}^{Q}\sum_{t=1}^{T}G$$
$$G = \mathbb{E}_{\widetilde{h}^{p}_{(q,t)}}[\log\sum_{j=1}^{N_C}e^{U}]$$
$$U = \psi^{j}_{k}\widetilde{h}^{p}_{(q,t)}, \quad \psi^{j}_{k} = w_j - w_k$$

(3)
$$G \le \log\mathbb{E}_{\widetilde{h}^{p}_{(q,t)}}[\sum_{j=1}^{N_C}e^{U}] = \log\sum_{j=1}^{N_C}\mathbb{E}_{\widetilde{h}^{p}_{(q,t)}}[e^{U}]$$

(4)
$$U \sim \mathcal{N}(\psi^{j}_{k}(h_{(q,t)} + \alpha\triangle h_{(q,t)}), \Sigma_{jk})$$
$$\Sigma_{jk} = \beta\psi^{j}_{k}(\sigma_k + \triangle\sigma_k){\psi^{j}_{k}}^{T}$$

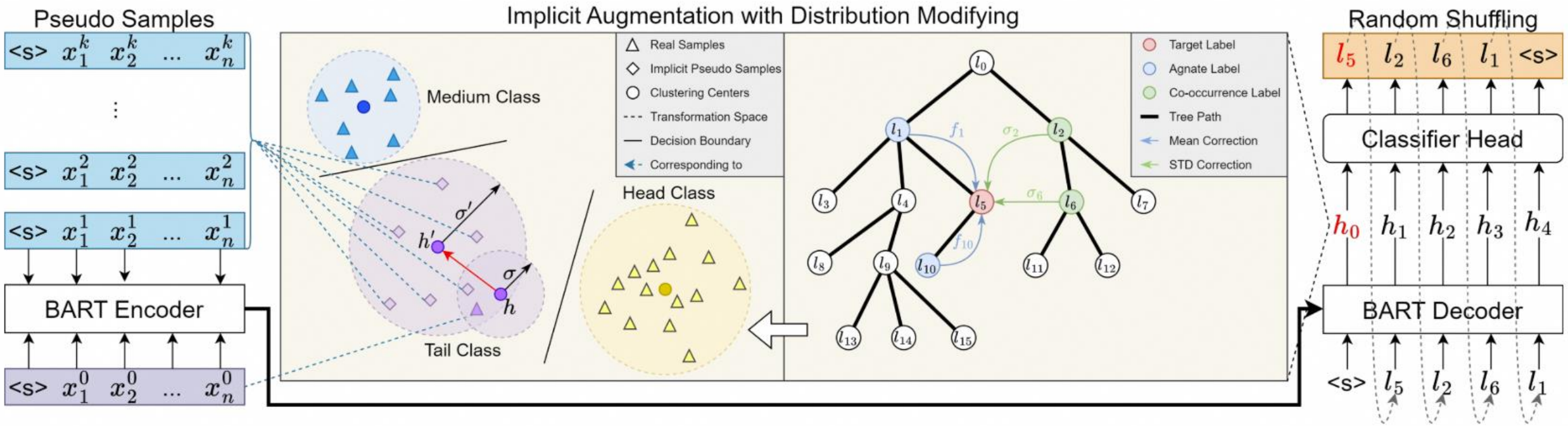
(5)
$$\mathrm{E}\left[e^{tX}\right] = e^{t\mu+\frac{1}{2}\sigma^2t^2}, \quad X \sim \mathcal{N}\left(\mu,\sigma^2\right)$$

(6)
$$\log\sum_{j=1}^{N_C}\mathbb{E}_{\widetilde{h}^{p}_{(q,t)}}[e^{U}] = \log\sum_{j=1}^{N_C}e^{V}$$

(7)
$$V = \psi^{j}_{k}(h_{(q,t)} + \alpha\triangle h_{(q,t)}) + \tfrac{1}{2}\beta\psi^{j}_{k}(\sigma_k + \triangle\sigma_k){\psi^{j}_{k}}^{T}$$

(8)
$$\overline{L_{aug}} = \frac{1}{Q}\frac{1}{T}\sum_{q=1}^{Q}\sum_{t=1}^{T}\log\sum_{j=1}^{N_C}e^{V} = -\frac{1}{Q}\frac{1}{T}\sum_{q=1}^{Q}\sum_{t=1}^{T}$$
$$\log(\frac{e^{w_k(h_{(q,t)}+\alpha\triangle h_{(q,t)})}}{\sum_{j=1}^{N_C}e^{w_j(h_{(q,t)}+\alpha\triangle h_{(q,t)})+\frac{1}{2}\beta\psi^{j}_{k}(\sigma_k+\triangle\sigma_k){\psi^{j}_{k}}^{T}}})$$

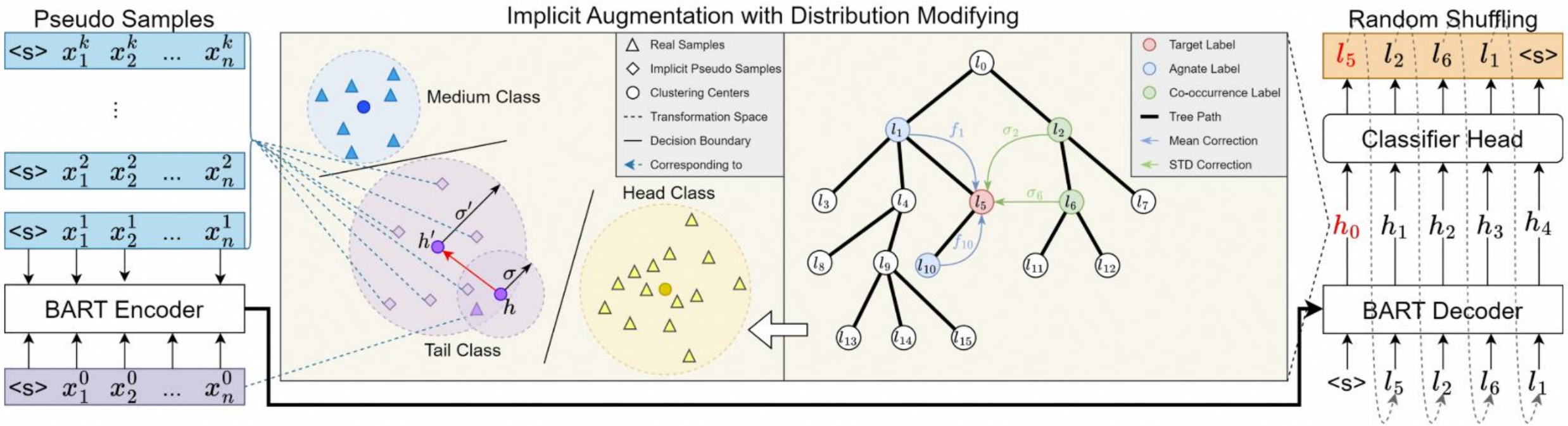
03 Implicit Augmentation + Random Shuffling



BFS: 1=>2=>5=>6

DFS: 1=>5=>2=>6

03 Implicit Augmentation + Random Shuffling



BFS: 1=>2=>5=>6

DFS: 1=>5=>2=>6

Random: 5=>2=>6=>1

04 Experimental Results

IMplicitly **A**ugmented **G**enerativ**E** framework with distribution modification (**IMAGE**)

Model	WOS		RCV1-V2		BGC	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
TextRNN	77.94	69.65	-	-	-	-
TextCNN	82.00	76.18	76.60	43.00	-	-
TextRCNN	83.55	76.99	81.57	59.25	-	-
HiAGM	85.82	80.28	83.96	63.35	77.22	57.91
BERT+HiAGM	86.04	80.19	85.58	67.93	-	-
HTCInfoMax	85.58	80.05	83.51	62.71	-	-
BERT+HTCInfoMax	86.30	79.97	85.53	67.09	-	-
HiMatch	86.20	80.53	84.73	64.11	76.57	58.34
BERT+HiMatch	86.70	81.06	86.33	68.66	78.89	63.19
HGCLR	87.11	81.20	86.49	68.31	-	-
HPT	87.16	81.93	87.26	69.53	-	-
SGM-T5	85.83	80.79	84.39	65.09	77.84	60.91
Seq2Tree	87.20	82.50	87.20	70.01	79.72	63.96
PAAM-HiA-T5	90.36	81.64	87.22	70.02	-	-
IMAGE	87.76	82.38	87.69	70.77	81.16	67.37

Table 2: Experimental results of on three benchmarks. The best result is in red, the second is in green, and the third is in blue. The table is divided into three parts, and from top to bottom they are: discriminative methods, generative methods, and our method.

Model	Size
BERT+GCN	148M
BERT+Graphormer	156M
T5	220M
BART	140M

Table 3: The size of different backbones.

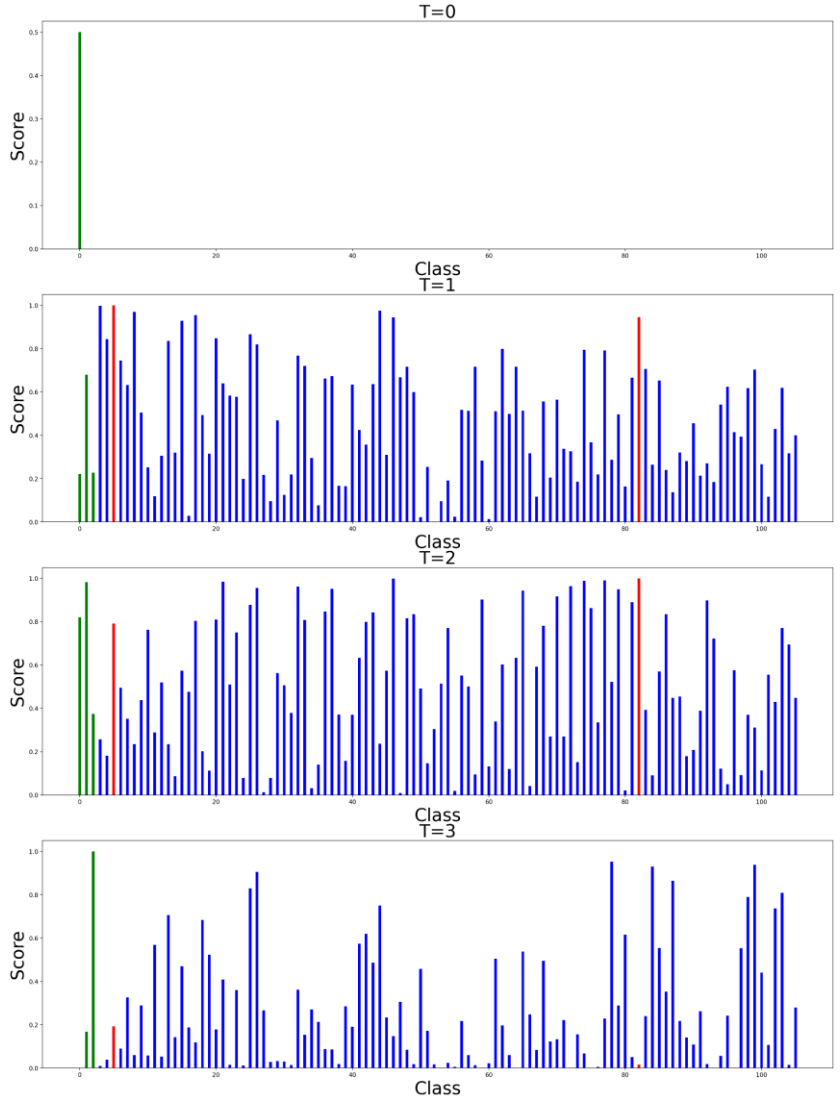
04 Experimental Results

Model	Micro-F1	Macro-F1
IMAGE	87.69	70.77
-w/o Distribution Modification	87.43	69.26
-w/o Implicit Augmentation	87.06	69.22
-w/o Random Shuffling	86.67	69.35
BART (Backbone)	86.51	68.15

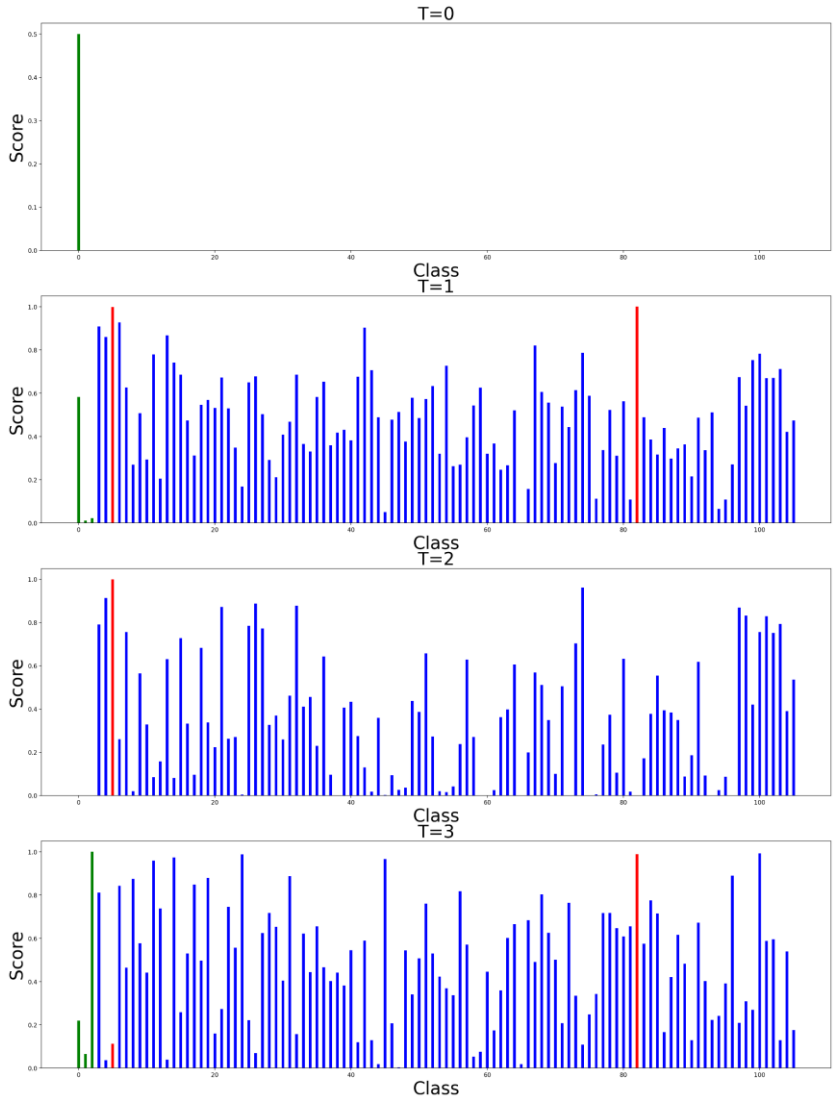
Table 4: The results of the ablation study on the benchmark dataset - RCV1-V2.

04 Experimental Results

BFS:



Ours:



The output of BART (the left graph) compared with IMAGE (the right graph). Class 0, 1 and 2 (green bars) corresponds to <s>, <pad>, and </s>, which are ignorable in the global predictions.

Thanks for your listening.

Huawen Feng