Continual Graph Convolutional Network for Text Classification

Tiandeng Wu[1*], Qijiong Liu[2*], Yi Cao[1], Yao Huang[1], Xiao-Ming Wu[2†], Jiandong Ding[1†]

[1] Huawei Technologies Co., Ltd., China
[2] The Hong Kong Polytechnic University, Hong Kong
{wutiandeng1, caoyi23, huangyao11, dingjiandong2}@huawei.com,
jyonn.liu@connect.polyu.hk, xiao-ming.wu@polyu.edu.hk

（AAAI-2023)

code： https://github.com/Jyonn/ContGCN

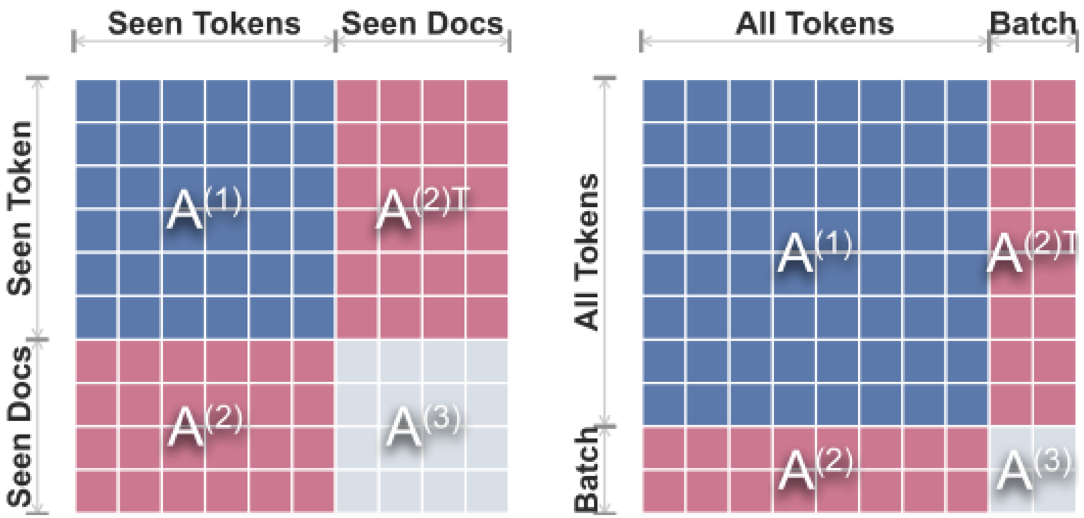
gesis
Leibniz-Institut
für Sozialwissenschaften

Reported by Zhaoze Gao

1. Introduction
2. Approach
3. Experiments

gesis
Leibniz-Institut für Sozialwissenschaften

Introduction



(a) STSD (b) ATAD

Figure 1: Comparison of the adjacency matrices. Left: seen-token-seen-document (STSD) paradigm (e.g., BertGCN). Right: proposed all-token-any-document (ATAD) paradigm.

they commonly follow a seen-token-seen-document (STSD) paradigm to construct a fixed document-token graph with all seen documents (labeled or unlabeled).

we propose a new all-token-any-document (ATAD) paradigm to dynamically construct a document-token graph.

token-token matrix $\mathbf{A}^{(1)} \in \mathbb{R}^{u' \times u'}$ $\mathbf{A}^{(1)} \in \mathbb{R}^{u \times u}$,

document-token matrix $\mathbf{A}^{(2)} \in \mathbb{R}^{m \times u'}$ $\mathbf{A}^{(2)} \in \mathbb{R}^{b \times u}$

document-document identity matrix $\mathbf{A}^{(3)} \in \mathbb{R}^{m \times m}$

$\mathbf{A}^{(3)} \in \mathbb{R}^{b \times b}$

Approach

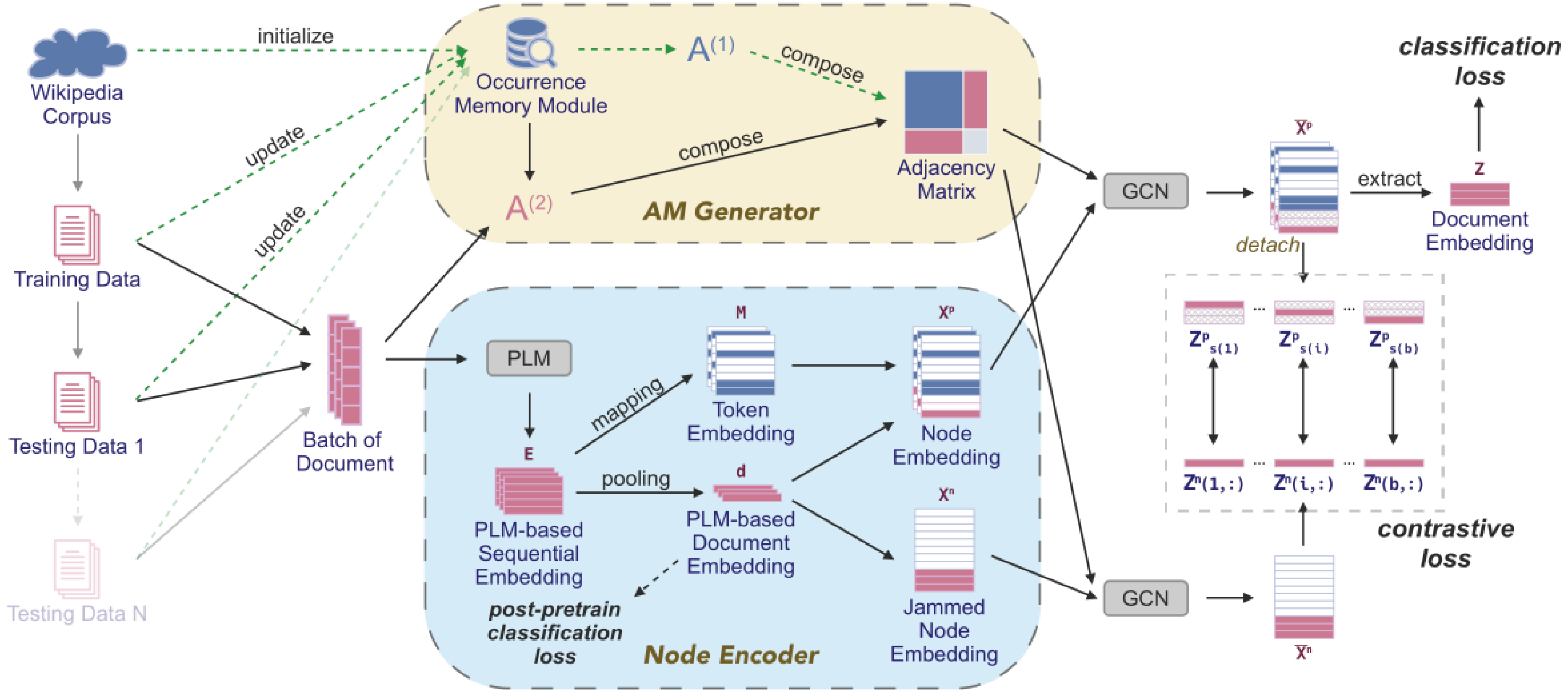
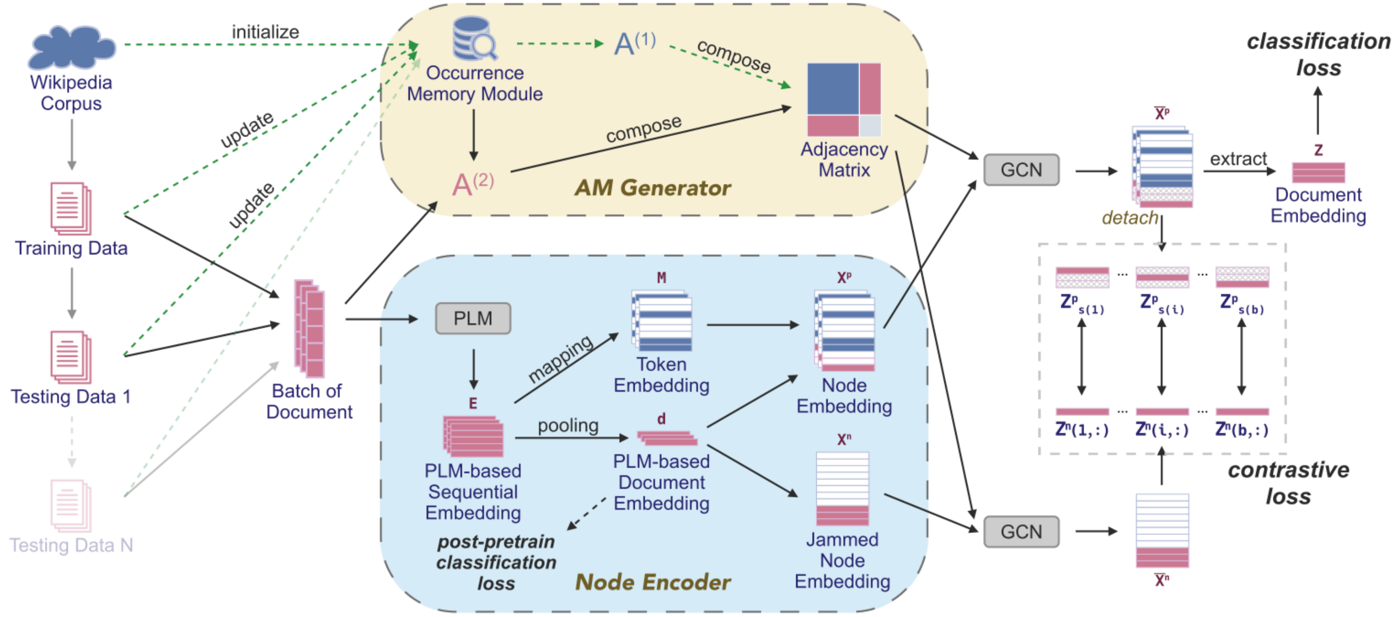


Figure 2: Framework of our ContGCN model. Green dotted lines represent operations before each phase of model training or testing. Two key components, i.e., AM Generator and Node Encoder, dynamically construct the adjacency matrix and generate node embeddings, which are then fed into a GCN encoder. Finally, our ContGCN model is trained with a classification loss and an anti-interference contrastive loss.

Approach



$$\widetilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{\frac{1}{2}}, \tag{1}$$

$$\mathbf{H}^{(k)} = \sigma\left(\widetilde{\mathbf{A}}\mathbf{H}^{(k-1)}\mathbf{W}_k\right), \tag{2}$$

token vocabulary set $\mathcal{T}(u = |\mathcal{T}|)$.

a document counter $s \in \mathbb{Z}^1$

token occurrence counter $\mathbf{c} \in \mathbb{Z}^u$

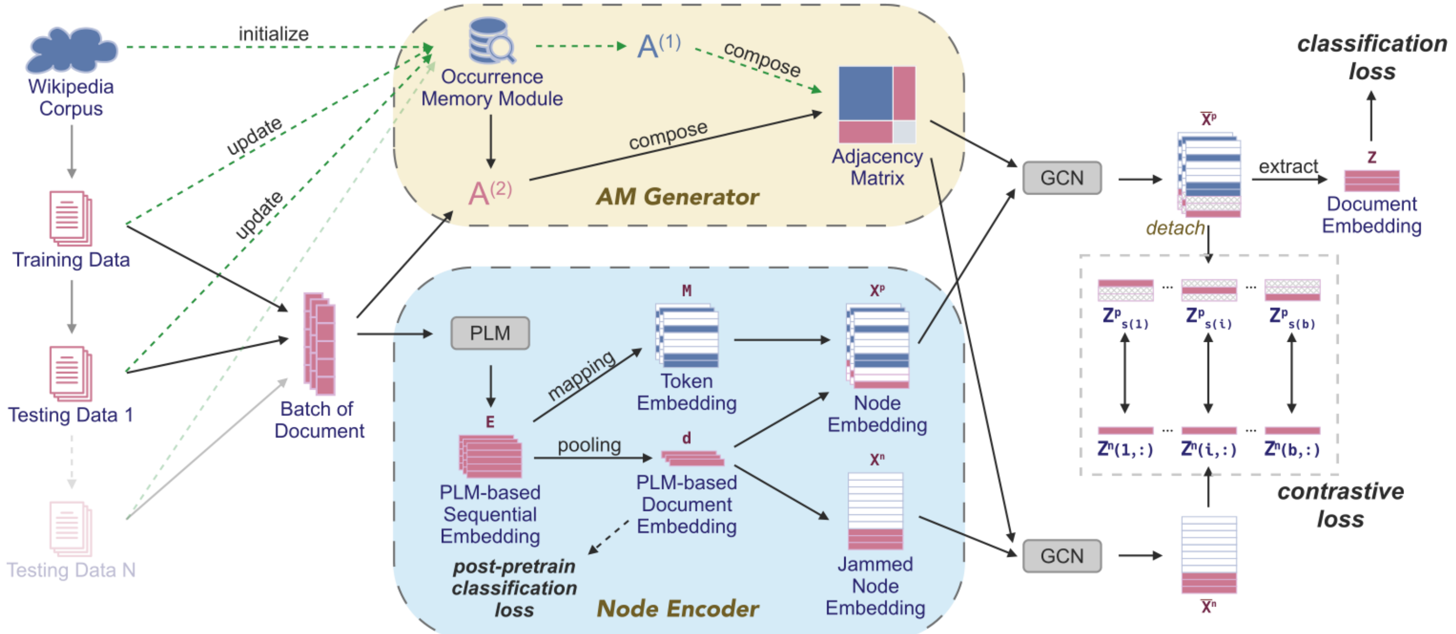
co-occurrence counter $\mathbf{C} \in \mathbb{Z}^{u\times u}$

$$\mathbf{A}^{(1)}_{i,j} = \begin{cases} 1, & \text{if } i = j, \\ \max\left(\log\left(s\frac{C_{i,j}}{c(i,:)c_j}\right), 0\right), & \text{else.} \end{cases} \tag{3}$$

$$\mathbf{A}^{(2)}_{\mathbf{s},t} = \frac{\mathrm{g}(\mathbf{s},t)}{|\mathbf{s}|}\log\frac{s}{c_t+1}, \tag{4}$$

$$\mathbf{A}^{(3)}_{i,j} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{else.} \end{cases} \tag{5}$$

Approach



$$\mathbf{A} = \begin{pmatrix} \mathbf{A}^{(1)} & \mathbf{A}^{(2)\top} \\ \mathbf{A}^{(2)} & \mathbf{A}^{(3)} \end{pmatrix}. \tag{6}$$

$$\mathbf{s} = \left(t_1^{(\mathbf{s})}, t_2^{(\mathbf{s})}, \cdots, t_{|\mathbf{s}|}^{(\mathbf{s})}\right)$$

$$\mathbf{E}_{(\mathbf{s})} = \mathtt{PLM}(\mathbf{s}) \in \mathbb{R}^{l \times d}, \tag{7}$$

$$\mathbf{X}^n = \left(\mathbf{0}, \cdots, \mathbf{0}, \mathbf{d}^{(\mathbf{s}_1)}, \cdots, \mathbf{d}^{(\mathbf{s}_b)}\right)^\top, \tag{8}$$

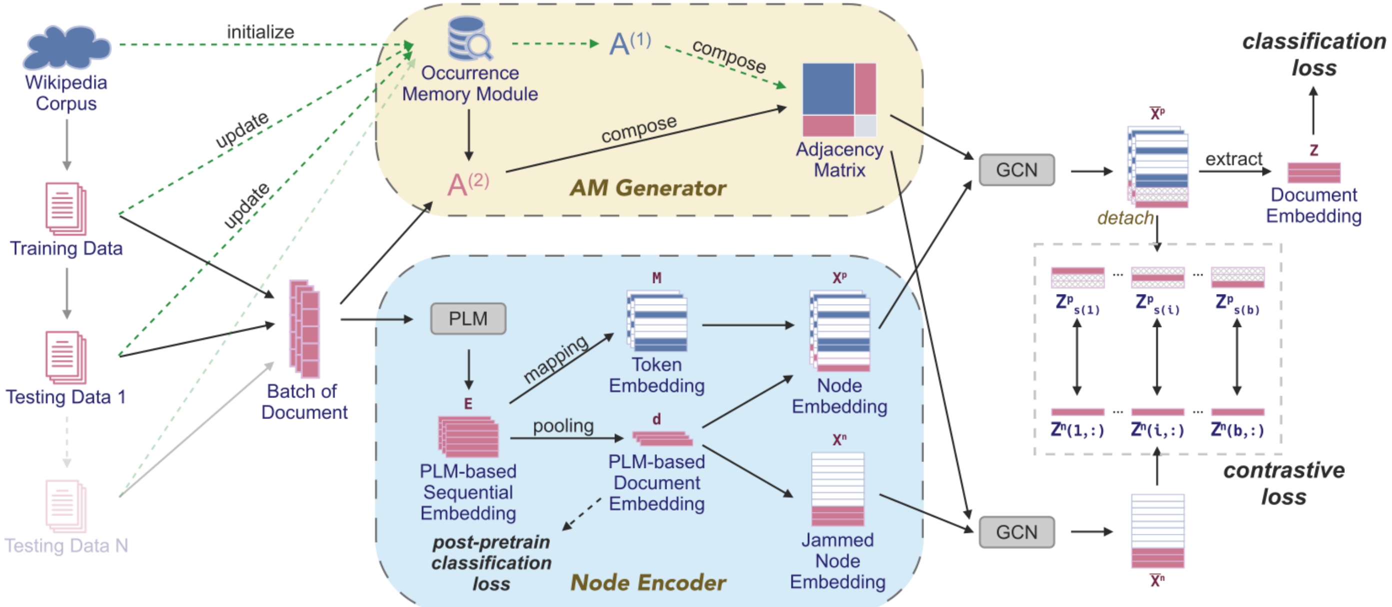
$$\mathbf{X}^n \in \mathbb{R}^{(u+b) \times d};$$

$$\mathbf{X}^p_{(\mathbf{s}_j)} = \left(\mathbf{M}_{(\mathbf{s}_j)}, \mathbf{0}, \cdots, \mathbf{0}, \mathbf{d}^{(\mathbf{s}_j)}, \mathbf{0}, \cdots, \mathbf{0}\right)^\top, \tag{9}$$

$$\mathbf{X}^p_{(\mathbf{s}_j)} \in \mathbb{R}^{(u+b) \times d} \qquad \mathbf{M}_{(\mathbf{s}_j)} \in \mathbb{R}^{u \times d}$$

$$\mathbf{M}_{(\mathbf{s}_j)}(i,:) = \begin{cases} \mathbf{E}_{(\mathbf{s}_j)}(k,:), & \text{if token } i \text{ of the vocabulary is the } k\text{-th token in } \mathbf{s}_j, \\ \mathbf{0}, & \text{otherwise.} \end{cases} \tag{10}$$

Approach



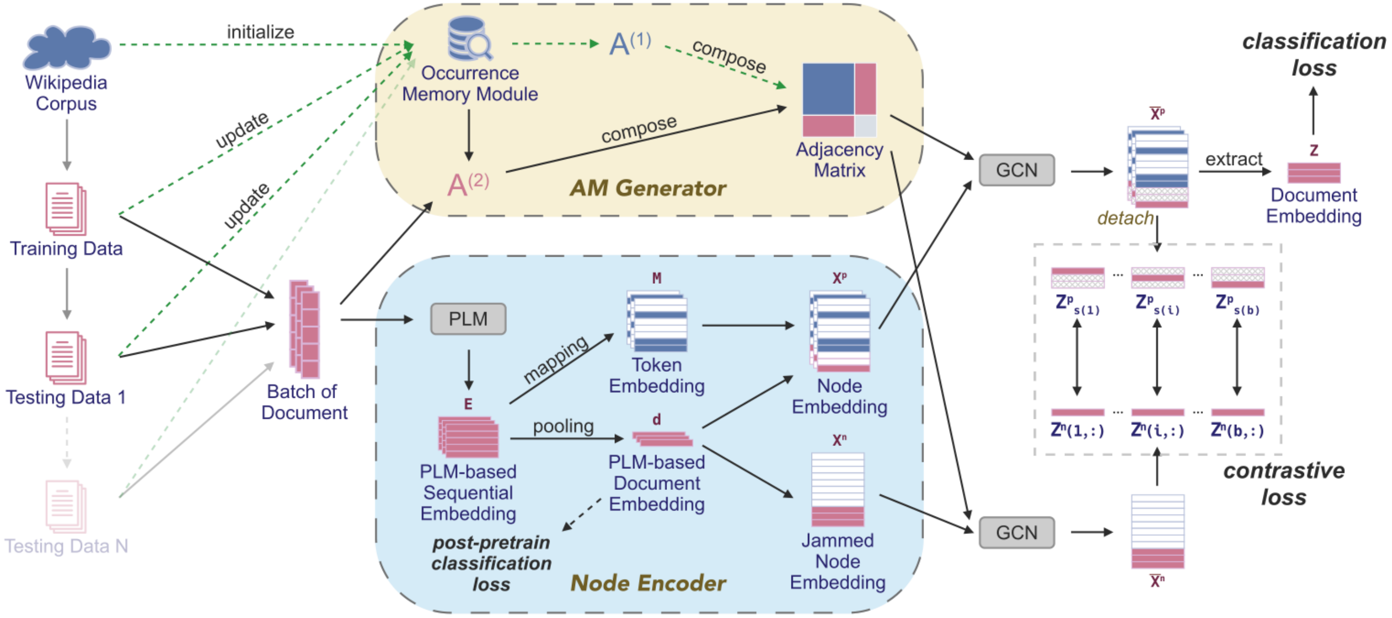
$$\mathbf{Z}(j,:) = \bar{\mathbf{X}}^{p}_{(\mathbf{s}_j)}(j+u,:). \tag{11}$$

$$\mathbf{Z}^{p}_{(\mathbf{s}_j)}(i,:) = \bar{\mathbf{X}}^{p}_{(\mathbf{s}_j)}(i+u,:) \text{ and} \tag{12}$$

$$\mathbf{Z}^{n}(i,:) = \bar{\mathbf{X}}^{n}(i+u,:), \tag{13}$$

$$\mathcal{L}_{\text{cls}} = -\frac{1}{b}\sum_{j=1}^{b} \log\left(f\left(\mathbf{Z}(j,:)\right)_{l_j}\right), \tag{14}$$

Approach



$$\mathcal{L}_{\text{aic}} = -\frac{1}{b}\sum_{j=1}^{b} \log\left(\mathbf{y}_{(\mathbf{s}_j)}(j)\right), \text{ where} \tag{15}$$

$$\mathbf{y}_{(\mathbf{s}_j)} = \mathtt{softmax}\left(\mathbf{Z}^{p}_{(\mathbf{s}_j)} \left(\mathbf{Z}^{n}(j,:)\right)^{\top}\right) \in \mathbb{R}^{b}. \tag{16}$$

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda\mathcal{L}_{\text{aic}}. \tag{17}$$

Experiments

Dataset	20NG	R8	R52	Ohsumed	MR
# Docs	18,846	7,674	9,100	7,400	10,662
# Training	11,314	5,485	6,532	3,357	7,108
# Test	7,532	2,189	2,568	4,043	3,554
# Classes	20	8	52	23	2
Avg. Length	221	66	70	136	20

Table 1: Dataset statistics.

Models	20NG	R8	R52	Ohsumed	MR
TextGCN	86.3	97.1	93.6	68.4	76.7
TensorGCN	87.7	98.0	95.0	70.1	77.9
BERT	85.3	97.8	96.4	70.5	85.7
RoBERTa	83.8	97.8	96.2	70.7	89.4
XLNet	85.1	98.0	<u>96.6</u>	70.7	87.2
TG-Transformer	-	98.1	95.2	70.4	-
BertGCN	89.3	98.1	<u>96.6</u>	<u>72.8</u>	86.0
RoBERTaGCN	<u>89.5</u>	<u>98.2</u>	96.1	<u>72.8</u>	<u>89.7</u>
$\text{ContGCN}_{\text{BERT}}$	89.4	98.3	96.9	73.1	86.4
$\text{ContGCN}_{\text{XLNet}}$	89.7	98.5	**97.0**	73.1	88.7
$\text{ContGCN}_{\text{RoBERTa}}$	**90.1**	**98.6**	96.6	**73.4**	**91.3**

Table 2: Comparison of ContGCN with state-of-the-art models in offline evaluation. The best results are in boldface, and the second best results are underlined.

Experiments

Models	20NG	R8	Ohsumed
$\text{ContGCN}_{\texttt{RoBERTa}}$	90.1	98.6	73.4
w/o Wikipedia Init	89.9	98.2	73.1
w/o OMM Updating	89.6	98.3	73.0
w/o Contrastive Loss	89.7	98.5	73.2
$\text{ContGCN}_{\texttt{XLNet}}$	89.7	98.5	73.1
w/o Wikipedia Init	89.8	98.3	72.8
w/o OMM Updating	89.4	98.2	72.7
w/o Contrastive Loss	89.5	98.2	73.0

Table 3: Influence of Wikipedia initialization, OMM updating, and the anti-interference contrastive task.

Experiments

Variants	1/6	2/6	3/6	4/6	5/6	6/6
ContGCN*	86.4	87.3	88.1	88.6	89.0	89.6
ContGCN	86.3	87.1	87.8	88.2	88.7	89.1
ContGCN$^{\alpha}$	86.1	86.9	87.5	87.9	88.3	88.7
ContGCN$^{\beta}$	86.0	86.2	86.4	86.6	86.9	87.1

Table 4: Comparisons of variants of ContGCN$_{\texttt{RoBERTa}}$ in the online learning scenario on the 20NG dataset. ContGCN* is retrained from scratch in each session with all previously seen data. ContGCN$^{\alpha}$ is updated without the contrastive loss. ContGCN$^{\beta}$ is updated without LUM.

Experiments

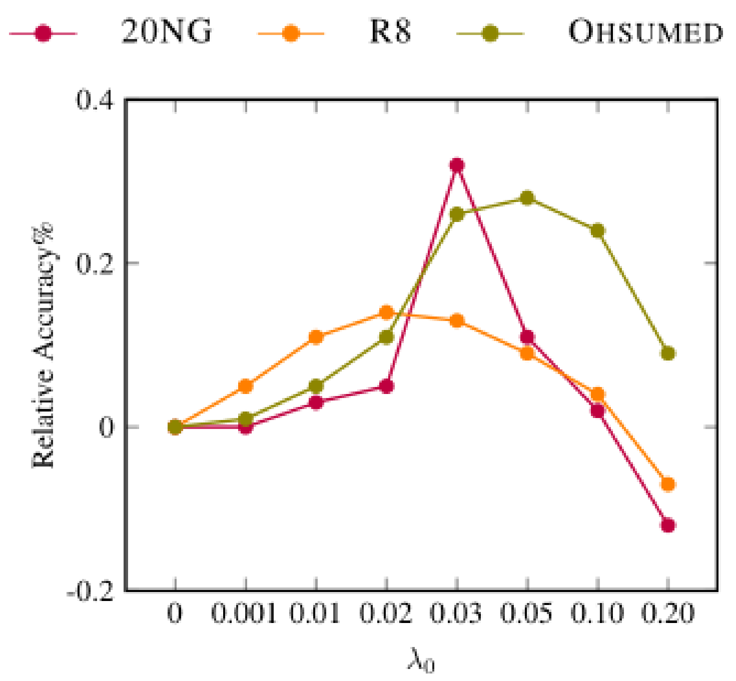
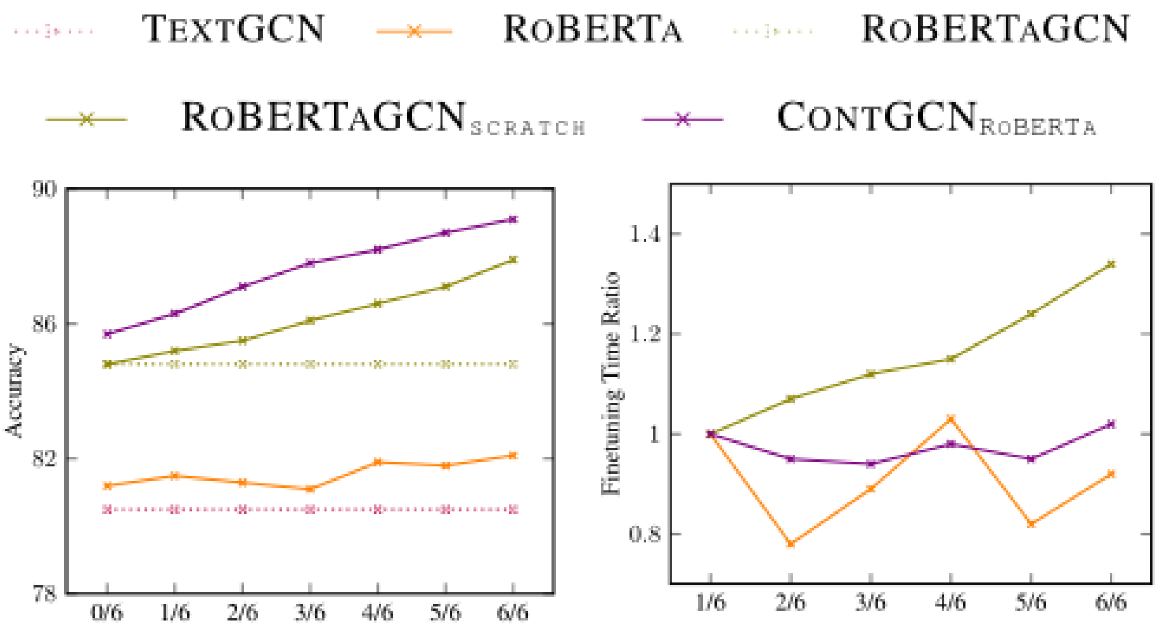


Figure 3: Influence of the parameter λ that weights the anti-interference contrastive loss. *Relative accuracy (%)* means the difference between the accuracy achieved with $\lambda = \lambda_0$ and that achieved with $\lambda = 0$.

Experiments



(a) Reasoning Quality

(b) Finetuning Efficiency

Figure 4: Comparison between our ContGCN model and baselines in an online learning scenario. We divide the 20NG dataset into training, testing, and updating sets by the ratio of 2:2:6. We trained each model with the training set to learn an initial version. Then, we divided the updating set into six equal parts and gradually fed each part to the model for fine-tuning. The *finetuning time ratio* in (b) is calculated by the finetuning time of the current session over that of the first session. For each training or updating session, we used 10% of the training set as the validation set.

Experiments

Models	0th	1st	2nd	3rd
RoBERTaGCN	91.7	N/A	N/A	N/A
RoBERTa	87.6	86.8	85.2	83.5
$\text{ContGCN}^{\beta}_{\texttt{RoBERTa}}$	**92.8**	90.3	89.9	88.2
$\text{ContGCN}_{\texttt{RoBERTa}}$	**92.8**	**92.5**	**92.0**	**90.9**

Table 5: Comparison of our ContGCN model with RoBERTa in an industrial online learning scenario. All models are first trained offline (in the 0th month) with a labeled dataset. After deployed, $\text{ContGCN}_{\texttt{RoBERTa}}$ performs online learning with LUM. $\text{ContGCN}^{\beta}_{\texttt{RoBERTa}}$ is a static network with parameters fixed after the initial training.

Thank you！