Neural Collaborative Graph Machines for Table Structure Recognition

Hao Liu† Xin Li* Bing Liu Deqiang Jiang Yinsong Liu Bo Ren
Tencent YouTu Lab
{ivanhliu, fujikoli, billbliu, dqiangjiang, jasonysliu, timren}@tencent.com

Abstract

Recently, table structure recognition has achieved impressive progress with the help of deep graph models. Most of them exploit single visual cues of tabular elements or simply combine visual cues with other modalities via early fusion to reason their graph relationships. However, neither early fusion nor individually reasoning in terms of multiple modalities can be appropriate for all varieties of table structures with great diversity. Instead, different modalities are expected to collaborate with each other in different patterns for different table cases. In the community, the importance of intra-inter modality interactions for table structure reasoning is still unexplored. In this paper, we define it as heterogeneous table structure recognition (Hetero-TSR) problem. With the aim of filling this gap, we present a novel Neural Collaborative Graph Machines (NCGM) equipped with stacked collaborative blocks, which alternatively extracts intra-modality context and models inter-modality interactions in a hierarchical way. It can represent the intra-inter modality relationships of tabular elements more robustly, which significantly improves the recognition performance. We also show that the proposed NCGM can modulate collaborative pattern of different modalities conditioned on the context of intra-modality cues, which is vital for diversified table cases. Experimental results on benchmarks demonstrate our proposed NCGM achieves state-of-the-art performance and beats other contemporary methods by a large margin especially under challenging scenarios.

1. Introduction

Table structure recognition (TSR) aims to recognize the table internal structure to the machine readable data mainly presented in two formats: logical structure [18, 46] and physical structure [2, 13, 20, 22, 27, 30, 31, 34, 35, 40, 45]. More concretely, logical structure only focuses on whether two table elements belong to the same row, column or cells (i.e., logical relationships), while the physical one contains not only logical relationships but also physical coordinates of cell boxes. The recognized tabular structure is essential to many downstream applications [12, 17]. Although many previous algorithms [2, 13, 18, 20, 22, 30, 31, 34, 35, 40, 45, 46] have achieved impressive progress in the community, TSR is still a challenging task due to two factors of complicated tables. The interior factor is complex table structure where spanning cell occupies at least two columns or rows, while exterior one is table distortion incurred by capture device.

Intuitively, table elements (text segment bounding boxes or table cells) commonly have inherent relationships and natural graph structure. Therefore, recent methods [2, 30, 34] attempt to attack the problem via constructing visual cues of table elements as graphs and applying the deep graph model, such as Graph Convolutional Networks (GCN) [15] to reason their relationships. To introduce richer table information, several methods [20, 30, 34] con-
cenate the visual features with other modalities of features, such as geometry features, as a whole input to the graph model, as shown in Fig. 1 (a). Nevertheless, the relational inductive biases of different modalities would be highly discrepant, which makes naively early-fused modalities unable to deal with all table structures of great diversity. Besides, the intra-modality relationships would negatively affect each other when reasoning specific table structures. For example, the coordinates of table would dominate when recognizing a regular table, but they would become unreliable when processing distorted table cases. Instead, another alternative way is to individually model intra-modality relationships between table elements and combine them by a late-fusion strategy (Fig. 1 (b)). Unfortunately, the disentangled reasoning in terms of intra-modality interactions would introduce the curtailment of inter-modality interactions. This dilemma leads to the following question: can different modalities collaborate with each other rather than interfering under different table scenarios? We define this practical problem as heterogeneous table structure recognition (Hetero-TSR), which still lacks investigation.

In this work, we propose a novel Neural Collaborative Graph Machines (NCGM) tailored for this problem, as illustrated in Fig. 1 (c). Concretely, we adopt text segment bounding boxes as table elements in our method and extract their multi-modality feature embeddings from appearance, geometry and content dimensionality separately. To obtain the corresponding graph context and explore their interactions, we go beyond the standard attention model and propose a basic collaborative block with two successive modules, i.e., Ego Context Extractor (ECE) and Cross Context Synthesizer (CCS). Among, ECE plays a role that dynamically generates graph context for the samples of each modality while the subsequent CCS is in charge of fusing and modulating inter-modality interactive information for different table cases. We stack this elemental block multiple times. Through this way, the intra-modality context generation and inter-modality collaboration can be conducted alternatively in a hierarchical way, which enables intra-inter modality interactions to be generated constantly from the low layer to the top one. In other words, the low-level contextual information in multiple modalities and the high-level one can collaborate with each other throughout the whole network, which is similar to the human perception process [1, 26]. The yielded collaborative graph embeddings enable our method to achieve better performance compared to other TSR methods, especially under more challenging scenarios, as clearly validated by extensive experimental results. To sum up, our contributions are in the four folds:

- We investigate the importance of collaboration between different modalities in TSR and propose the Hetero-TSR problem. To our best knowledge, we are the first to research the collaborative patterns between modality interaction for predicting table structure.
- We coin a novel NCGM tailored for Hetero-TSR problem, which consists of collaborative blocks alternatively conducting intra-modality context extraction and inter-modality collaboration in a hierarchical way.
- Experimental results on public benchmarks demonstrate that our method significantly outperforms the state-of-the-arts.
- We release a synthesizing method to augment existing benchmarks to more challenging ones. Under more challenging scenarios, our method can achieve at most 11% improvement than the second best method.

2. Related Work
2.1. Table Structure Recognition

Before the flourishing of deep learning, traditional table structure recognition methods rely on pre-defined rules and hand-crafted features [9–11, 14, 41]. With the development of deep learning, table structure recognition methods have recently advanced substantially on performance, which can be classified into three categories: boundary extraction-based [13, 22, 27, 35, 40], generative model-based [18, 46], and graph-based [2, 20, 30, 34] methods.

**Boundary extraction-based methods.** To extract cell boundaries, DeepDeSRT [35] and TableNet [27] are proposed by utilizing semantic segmentation. Besides, another technique [13] exploits bi-directional GRUs to establish row and column boundaries in a context driven manner. However, these methods are struggled when identifying cells spanning multiple rows and columns. SPLERGE [40] splits the table into grid elements in which adjacent ones are merged to restore spanning cells, whereas it still suffers from boundary ambiguity problem. To tackle this issue, the hierarchical GTE [45] leverages clustering algorithm for cell structure recognition. Cycle-CenterNet [22] exploits the cycle-pairing module to simultaneously detect and group tabular cells into structured tables, which focuses on the precision of cell boundary of the wired table in the wild. In the similar spirit, LGPMA [31] applies soft pyramid mask learning mechanism on both the local and global feature maps. Nevertheless, the subsequently heuristic structure recovery pipeline cannot achieve decent performance in complex scenarios.

**Generative model-based methods.** The method [18] utilizes the encoder-decoder framework, which generates an HTML tag sequence that represents the arrangement of rows and columns as well as the type of table cells. Moreover, another generative algorithm [46], termed EDD, consists of an encoder, a structure decoder and a cell decoder.
The encoder captures visual features of input table images, while the structure decoder reconstructs table structure and helps the cell decoder to recognize cell content.

**Graph-based methods.** GraphTSR [2] employs graph attention blocks to learn the vertex and edge representations in the latent space, and classifies edges as horizontal, vertical or unrelated. The method [30] introduces DGCNN to predict the relationship between words represented by the appearance and geometry features. Also based on DGCNN, TabStruct-Net [34] proposes an end-to-end network training cell detection and structure recognition networks in a joint manner. Besides, FLAG-Net [20] leverages the modulable dense and sparse context of table elements. However, the above graph-based works are mostly designed for the different modalities and result in discretization error and insignificant performance gains in NLP community [5, 16, 21, 32, 39], but also gives birth to several pre-training methods [19, 23, 44] fusing various modalities for multimodal tasks.

**2.2. Transformer-based Multimodal Fusion**

Transformer [42] architecture not only achieves significant performance gains in NLP community [5, 16, 21, 32, 39], but also gives birth to several pre-training methods [19, 23, 44] fusing various modalities for multimodal tasks.

**Multiple embeddings fusion.** VL-BERT [38] inheriting from BERT [5] introduces additional visual feature embeddings for visual-linguistic representations. LayoutLM [44] is a document understanding pre-trained model, which jointly models the interactions between text and layout information across scanned document images. However, the above algorithms simply take early-fused multiple embeddings as inputs, which may ignore the interactions between different modalities and result in discretization error and important details missing.

**Co-attentional fusion.** To better utilize visual-linguistic representations, ViLBERT [23] processes both visual and textual inputs in separate streams that interact through co-attentional transformer layers. Moreover, SelfDoc [19] establishes the contextualization over a block of content via cross-modal learning to manipulate visual features and textual features. Nevertheless, these previous co-attention based methods can only handle two modalities. By comparison, our proposed NCGM focuses on modality collaboration rather than simple fusion. Further, NCGM can not only process the interaction among more than two individual modalities, but also alternatively conduct intra-modality context extraction and inter-modality collaboration, which exploits more useful information provided by different modalities.

3. Methodology

**3.1. Overall Architecture**

The overview of the proposed Neural Collaborative Graph Machines (NCGM) is shown in Fig. 2. It mainly consists of collaborative blocks, which have two successive Multi-head Attention-based [42] modules, i.e., Ego Context Extractor (ECE) and the Cross Context Synthesizer (CCS). First, three modalities of feature embeddings ($F^\sim \in \{F^G, F^A, F^C\}$) in terms of table elements are extracted, i.e., geometry, appearance and content embeddings. In each collaborative block, the extracted feature embeddings are built as context graphs which are separately applied by the ECE to shape “intra-modality stream”. Afterwards, the CCS selectively fuses individual contextual information from different modalities as inter-modality interactions maintained in “inter-modality stream”. Note, we set $M^\sim_{(0)} = F^\sim$ as the initial input of CCS. The block is stacked $L$ layers to implement the intra-inter modality collaboration in a hierarchical way. To predict the final table structure, the output collaborative graph embeddings from the $l$-th layer of inter-modality stream are sampled as pairs for cells, rows and columns classification.

**3.2. Feature Extraction**

In this component, a set of multi-modality features in terms of table elements are extracted from table image, including geometry embeddings $F^G \in \mathbb{R}^{N \times d}$, appearance embeddings $F^A \in \mathbb{R}^{N \times d}$ and content embeddings $F^C \in \mathbb{R}^{N \times d}$. $N$ denotes the number of text segment bounding boxes. A more detailed description is given in supplementary material.
3.3. Collaborative Block

Ego Context Extractor. Now we elaborate on how to extract contextual interactions within each modality of table elements with the help of the Ego Context Extractor (ECE). Specifically, each extracted modality of features input to the ECE is constructed as individual directed graph \( G^\sim = \{ V, E \} \in \{ G^G, G^A, G^C \} \). In each decoupled modality of graph, corresponding embedding of each text segment bounding box is regarded as node \( X = \{ x_1, x_2, ..., x_N \} \subseteq V \) which is connected to each other by edges \( E \subseteq V \times V \). In the similar spirit with works [30, 34], we adopt the following asymmetric edge function \( h_{\theta_i}(x_i, x_j) = x_i \| (x_i - x_j) \) to combine graph edge features to each node, which can be denoted as \( H^G_{\theta_i} \in \mathbb{R}^{(N - (N-1)/2) \times d} \). In the constructed graphs, each node can be either an anchor or one of context of others. In previous works using DGCNN [30, 34], only local context of each node is selected by \( k \)-Nearest Neighbors algorithm (KNN) to be aggregated into node feature. However, the local context is not versatile for representing relationships of all modalities. Besides, the DGCNN-based methods apply CNN to perform local context aggregation. For graph representation, CNN with strong inductive bias (e.g., local behavior) may not be the optimal choice. To tackle the above problems, our proposed ECE instead aggregates information of fully-connected graph for all three modalities via Multi-head Attention (MHA) [42] module, which has been verified that it makes few assumptions about inputs and can learn to combine local behavior and global information based on input content [3].

More concretely, \( l \)-th ECE takes intra-modality features \( C^\sim_{(l,1)} \) as queries \( Q \) and the graph edge combined features \( H^G_{(l,0)} \) as keys \( K \) and values \( V \) as illustrated in Fig. 3(a). Note, for the first layer, we input \( F^\sim \) as \( C^\sim_{(0)} \). However, the main limitation of using MHA is that the amount of input \( K \) and \( V \) can be very large (\( N \cdot (N - 1)/2 \) in our case), which is infeasible to be trained. Given \( Q \in \mathbb{R}^{N \times d_q}, K \in \mathbb{R}^{M \times d_k}, V \in \mathbb{R}^{M \times d_v} \), and \( M = N \cdot (N - 1)/2 \), the time complexity of the attention operation is \( O(NM) \) and the output is in \( N \times d_v \) dimensionalities, of which the number is only relevant to that of \( Q \). Therefore, we can extend the MHA to a more memory-efficient Compressed MHA (CMHA) by introducing memory compression module which is utilized to reduce image pixel numbers in [43], as depicted in Fig. 3(b). In detail, the compression operation can be implemented as:

\[
MC(H) = \text{Norm}(\text{Reshape}(x, \epsilon)W^h),
\]

where \( \text{Reshape}(H, \epsilon) \) denotes the operation of reshaping input \( x \in \mathbb{R}^{M \times d} \) to \( \bar{x} \in \mathbb{R}^{M \times d/\epsilon} \), and \( \epsilon \in [0, 1] \) is the compression ratio. Through this way, the complexity can be quadratically reduced from \( O(NM) \) to \( O(N \epsilon M) \). In default, we set \( \epsilon = N/M \), where \( N \) is the number of queries \( Q \). And \( \text{Norm}(\cdot) \) is the layer normalization. Additionally, we also equip the CMHA with residual connections in our method to make the query information flow unimpeded, which can be defined as:

\[
\begin{align*}
Y &= \text{Add} & \text{Norm}(\text{FFN}(\bar{P})) & \text{FFN}(\text{Add} & \text{Norm}(Q, P)), \\
\bar{P} &= \text{Add} & \text{Norm}(Q, P), \\
P &= \text{MHA}(Q, MC(K), MC(V)),
\end{align*}
\]

where “\( \text{FFN}(\cdot) \)” is the feed-forward layer and “\( \text{Add} & \text{Norm}(\cdot) \)” denotes element-wise addition and layer normalization, which is similar to the work [42]. Conclusively, the contextual graph information is baked into graph node as \( C^\sim \in \{ C^G, C^A, C^C \} \) within each modality through the CMHA in our ECE module.

Cross Context Synthesizer. Once heterogeneous context graph embeddings are obtained, our goals are to fuse them together in a collaborative way and to learn the collaborative patterns between different modalities. Also based on the CMHA, we design the Cross Context Synthesizer (CCS), as is shown in Fig. 3(c). In detail, the CCS has three parallel CMHA modules, and each of them takes one modality as queries while the other two are jointly regarded as keys and values. Take the first branch in Fig. 3(c) for example, the CMHA takes “content” modality of context graph embeddings as \( Q \), and the respective outputs of ECE for “geometry” and “appearance” are input as \( K \) and \( V \). In Fig. 3(c), “(\( U \))” denotes the union of two modality sets. For the similar purpose in ECE process, we also follow the similar rule to compress the number of “memory” to \( N \) which equals to that of \( Q \). Essentially, the query modality explores helpful information from another two modalities.

3.4. Table Structure Prediction

At the \( l \)-th layer of collaborative block, the outputs of CCS are to further fused as collaborative graph embeddings, which are denoted as \( E = \{ e_1, e_2, ..., e_N \} \in \mathbb{R}^{N \times d_e} \). Based on the embeddings \( E \), our method constructs the \( i \)-th and \( j \)-th samples as pairs and concatenate them along channel axis as vectors \( U = \{ u_{i,1}, u_{i,2}, ..., u_{i,j}, ..., u_{N,N} \} \in \mathbb{R}^{N^2 \times 2d_e} \). Then three groups of FC layers are separately applied for predicting binary-class relations of \( U \), i.e., whether
the pair of $i$-th and $j$-th sample is belong to the same row, column or cell, as illustrated in Fig. 2. Each FC group consists of three FC layers with 256 dimensions and a 2-dimension FC with softmax layer.

3.5. Training Strategy

We train our proposed NCGM in an end-to-end way. The whole loss function is defined as $L = L_{cell} + L_{col} + L_{row}$, where $L_{cell}$, $L_{col}$ or $L_{row}$ represents cell, column and row relationship losses. For each of them, we adopt the multi-task loss $L = \lambda_1 L_{class} + \lambda_2 L_{con}$ to satisfy both the contrastive objective and to predict belonging classes of the output embedding pairs. $L_{con}$ and $L_{class}$ are contrastive loss and binary classification loss functions respectively. A more detailed description is given in supplementary material.

4. Experiments

4.1. Datasets and Evaluation Protocol

Datasets. We perform extensive experiments on various benchmark datasets. Among, ICDAR-2013 [8], ICDAR-2019 [6], WTW [22], UNLV [36], SciTSR [2] and SciTSR-COMP [2] are employed for physical structure recognition, while TableBank [18] and PubTabNet [46] are adopted for evaluating logical structure recognition performance. It should be noted that there is no training set in ICDAR-2013 and UNLV datasets, so we extend the two datasets to the partial versions, which is similar to TabStruct-Net [34]. A more detailed description about public benchmarks is given in supplementary material.

To further investigate the capacity of our proposed method under more challenging scenarios, we expand “SciTSR-COMP” dataset to “SciTSR-COMP-A” by applying two kinds of distortion algorithms. A more detailed description is given in supplementary material.

Evaluation settings. Several existing works are only applicable to table images alone, while others utilize additional information including text segment/cell bounding boxes or text contents. To compare in a unified protocol, we follow two different experimental setups in [34]: (a) Setup-A where only table image is taken as input without additional information and (b) Setup-B where table image along with additional features such as cell/text segment bounding boxes and text contents. For a fair comparison, we also incorporate the result boxes of detection in FLAG-Net [20] and the OCR results of Tesseract [37] as inputs in Setup-A.

Evaluation protocol. We employ precision, recall and F1-score [7] as protocol to evaluate the performance of our model for recognizing table physical structure including vertical and horizontal relations. For the recognition of table logical structure, BLEU score [28] used in [18] and Tree-Edit-Distance-based Similarity (TEDS) proposed in [46] are exploited.

4.2. Implementation Details

The framework is built on Pytorch [29]. We scale the input table images to a fixed size $512 \times 512$ to introduce scale invariance. In default, the layer number of collaborative blocks is set to 3 and the hidden size $d$ is set to 64. Further, we set $h = 8, d_{m} = 64, d_{k} = d_{v} = 8$ for both Ego Context Extractor (ECE) and Cross Context Synthesizer (CCS) of each collaborative block. During training, the learning rate is initialized as $1e^{-4}$ and divided by 10 when the loss stops decreasing. For the training loss, we empirically set all weight parameters $\lambda_1 = \lambda_2 = 1$. For all experiments, the models are pre-trained on SciTSR for 10 epochs, and then fine-tuned on different benchmarks for 50 epochs, which is conducted on the platform with one Nvidia Tesla V100 GPU and 32 GB memory.

4.3. Comparison with State-of-the-arts

Results of physical structure recognition. As is shown in Tab. 1, our NCGM outperforms most of previous methods on different datasets for physical structure recognition. Compared with the strong baseline FLAG-Net [20], NCGM increases average F1-score on all datasets by round 2% under both Setup-A settings and Setup-B settings. When processing table images with complex distortions (“SciTSR-COMP-A”), it is worth mentioning that our NCGM can achieve about 11% and 12% higher F1-scores under Setup-A and Setup-B than the second-best FLAG-Net [20] without using distorted images as training data. If taking distorted data as training set, the performance of NCGM still can surpass it round 7% and 9% under both settings respectively. We also visualize row and column physical relationships of distorted table in Fig. 4. Note, the different color blocks in it merely visualize the belonging relationship rather than dividing the entire box. Taking right column of Fig. 4 for example, “POS tagging information” is one whole text segment bounding box. In logical, one can observe that “POS tagging information” box spans across five columns of word bounding boxes below it in column dimension. Therefore, the five columns attribute their respective colors to the “POS tagging information” box. By comparison, our method correctly recognizes both relationships while the FLAG-Net performs unsatisfactorily under distorted table scenes.

Results of logical structure recognition. In order to evaluate our model on logical structure recognition task benchmarks, i.e., TableBank and PubTabNet, we perform lightweight post-processing (see supplementary material) on the NCGM’s output results of row/column relationships to convert them to the HTML representation. Tab. 2 presents that our method achieves significant improvement compared with other methods for logical structure recognition task.
Computational complexity. A more detailed description is given in supplementary material.

<table>
<thead>
<tr>
<th>Method</th>
<th>Train Dataset</th>
<th>Setup-A</th>
<th>Setup-B</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>DGCNN [30]</td>
<td>Sci. + ICI3-P</td>
<td>-</td>
<td>-</td>
<td>98.6</td>
<td>99.0</td>
<td>98.8</td>
<td>99.1</td>
<td>99.3</td>
<td>99.2</td>
</tr>
<tr>
<td>TabStr. [34]</td>
<td>Sci. + ICI3-P</td>
<td>93.0</td>
<td>90.8</td>
<td>91.9</td>
<td>99.1</td>
<td>99.3</td>
<td>99.9</td>
<td>99.3</td>
<td>99.2</td>
</tr>
<tr>
<td>GTE [45]</td>
<td>Pub. + ICI3-P</td>
<td>94.4</td>
<td>92.7</td>
<td>93.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LGPMA [31]</td>
<td>Sci. + ICI3-P</td>
<td>96.7</td>
<td>99.1</td>
<td>97.9</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C-CTRNet [22]</td>
<td>WTW + ICI19</td>
<td>95.5</td>
<td>88.3</td>
<td>91.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FLAG-Net [20]</td>
<td>Sci. + ICI3-P</td>
<td>97.9</td>
<td>99.3</td>
<td>98.6</td>
<td>99.2</td>
<td>99.5</td>
<td>99.9</td>
<td>99.3</td>
<td>99.2</td>
</tr>
<tr>
<td>NCGM</td>
<td>Sci. + ICI3-P</td>
<td>98.4</td>
<td>99.3</td>
<td>98.8</td>
<td>99.3</td>
<td>99.9</td>
<td>99.6</td>
<td>99.6</td>
<td>99.6</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Method</th>
<th>Train Dataset</th>
<th>Setup-A</th>
<th>Setup-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>DGCNN [30]</td>
<td>Sci. + ICI19</td>
<td>80.3</td>
<td>77.8</td>
</tr>
<tr>
<td>TabStr. [34]</td>
<td>Sci. + ICI19</td>
<td>82.2</td>
<td>78.7</td>
</tr>
<tr>
<td>C-CTRNet [22]</td>
<td>WTW</td>
<td>-</td>
<td>80.8</td>
</tr>
<tr>
<td>FLAG-Net [20]</td>
<td>Sci. + ICI19</td>
<td>85.2</td>
<td>83.8</td>
</tr>
<tr>
<td>NCGM</td>
<td>Sci. + ICI19</td>
<td>84.6</td>
<td>86.1</td>
</tr>
</tbody>
</table>

ICDAR-2019

<table>
<thead>
<tr>
<th>Method</th>
<th>Train Dataset</th>
<th>Setup-A</th>
<th>Setup-B</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>DGCNN [30]</td>
<td>Sci. + ICI19</td>
<td>80.3</td>
<td>77.8</td>
<td>90.9</td>
<td>89.0</td>
<td>90.9</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TabStr. [34]</td>
<td>Sci. + ICI19</td>
<td>82.2</td>
<td>78.7</td>
<td>80.4</td>
<td>97.5</td>
<td>95.8</td>
<td>96.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C-CTRNet [22]</td>
<td>WTW</td>
<td>-</td>
<td>80.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FLAG-Net [20]</td>
<td>Sci. + ICI19</td>
<td>85.2</td>
<td>83.8</td>
<td>84.5</td>
<td>96.1</td>
<td>96.3</td>
<td>96.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NCGM</td>
<td>Sci. + ICI19</td>
<td>84.6</td>
<td>86.1</td>
<td>85.3</td>
<td>98.9</td>
<td>98.8</td>
<td>98.8</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

WTW

<table>
<thead>
<tr>
<th>Method</th>
<th>Train Dataset</th>
<th>Setup-A</th>
<th>Setup-B</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>DGCNN [30]</td>
<td>Sci. + UNLV-P</td>
<td>80.3</td>
<td>77.8</td>
<td>90.9</td>
<td>89.0</td>
<td>90.9</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TabStr. [34]</td>
<td>Sci. + UNLV-P</td>
<td>82.2</td>
<td>78.7</td>
<td>80.4</td>
<td>97.5</td>
<td>95.8</td>
<td>96.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C-CTRNet [22]</td>
<td>WTW</td>
<td>-</td>
<td>80.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FLAG-Net [20]</td>
<td>Sci. + UNLV-P</td>
<td>85.2</td>
<td>83.8</td>
<td>84.5</td>
<td>96.1</td>
<td>96.3</td>
<td>96.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NCGM</td>
<td>Sci. + UNLV-P</td>
<td>84.6</td>
<td>86.1</td>
<td>85.3</td>
<td>98.9</td>
<td>98.8</td>
<td>98.8</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

UNLV-P


<table>
<thead>
<tr>
<th>Method</th>
<th>Train Dataset</th>
<th>Setup-A</th>
<th>Setup-B</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image-to-Text [18]</td>
<td>TableBank</td>
<td>73.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TabStruct-Net [34]</td>
<td>SciTSR</td>
<td>91.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLAG-Net [20]</td>
<td>SciTSR</td>
<td>93.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NCGM</td>
<td>SciTSR</td>
<td>94.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Comparison results of logical structure recognition on TableBank and PubTabNet datasets.

4.4. Ablation Study

In this subsection, we perform several analytic experiments under Setup-B settings on SciTSR-COMP benchmark to investigate the contributions of intra-modality and inter-modality interactions in our proposed NCGM.

Effect of intra-modality interactions. For intra-modality interactions, Tab. 3 compares the effectiveness of various extractors, including DGCNN [30] and Transformer [42], with ECE in our method. “Mixed” means all modality features are early-fused by concatenation as the input and “Individual” denotes each modality is input into
context extractor separately. Tab. 3 shows ECE can achieve the best performance when taking either mixed features or individual features as input while “Transformer” performs the worst. For “DGCNN”, it only aggregates information from top K similar nodes of each node instead of all ones. Compared with “DGCNN”, although “Transformer” can deploy the global information of nodes, it ignores the directed edge effects between nodes. Encouragingly, our CMHA-based ECE can not only consider the directed relationships between nodes, but also extract the context information from all nodes. Additionally, we can also observe that individual features can yield better results than the mixed ones, which proves that decoupling the individual modality from each other is indeed a more preferable way to solve the Hetero-TSR problem.

<table>
<thead>
<tr>
<th>Fusion Method</th>
<th>Input</th>
<th>Intra.</th>
<th>Inter.</th>
<th>Setup-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early Fusion</td>
<td>✓ ✔</td>
<td>✓ ✔</td>
<td>✓ ✔</td>
<td>96.3 97.4 96.8</td>
</tr>
<tr>
<td>Late Fusion</td>
<td>✗ ✔</td>
<td>✔ ❌</td>
<td>✗ ✔</td>
<td>95.1 95.6 95.3</td>
</tr>
<tr>
<td>NCGM</td>
<td>✗ ✔</td>
<td>✔ ❌</td>
<td>✗ ✔</td>
<td>97.8 98.3 98.0</td>
</tr>
<tr>
<td></td>
<td>✗ ✔</td>
<td>✔ ❌</td>
<td>✗ ✔</td>
<td>96.9 98.2 97.5</td>
</tr>
<tr>
<td></td>
<td>✗ ✔</td>
<td>✔ ❌</td>
<td>✗ ✔</td>
<td>94.9 96.1 95.5</td>
</tr>
<tr>
<td></td>
<td>✗ ✔</td>
<td>✔ ❌</td>
<td>✗ ✔</td>
<td>98.4 98.2 98.3</td>
</tr>
<tr>
<td></td>
<td>✗ ✔</td>
<td>✔ ❌</td>
<td>✗ ✔</td>
<td>98.8 99.3 99.0</td>
</tr>
</tbody>
</table>

Table 3. Ablation studies of NCGM on SciTSR-COMP dataset. “Intra.” and “Inter.” stand for intra-modality interactions and inter-modality interactions respectively. “Mix.” and “Ind.” are short for “Mixed” and “Individual”. “DG.” and “Tr.” denote “DGCNN” and “Transformer”. “Con.” represents “Concatenation”.

Effect of inter-modality interactions. We compare the proposed CCS with the “Concatenation” operation of multi-modal features in Tab. 3. It can be observed that CCS improves the accuracy of predicting adjacency relationship compared with directly late-fused multiple model features via concatenation. This confirms the benefits of CCS that enables one modality to positively collaborate with the others, and can capture the complex implicit modality relationships. Moreover, it also proves that the CCS module combined with ECE can further boost the performance.

4.5. Further Analysis on Collaborative Block

What does ECE learn from the intra-modality? As suggested by recent works [24, 25, 33] on interpreting attention mechanism, separate attention heads may learn to look for various relationships between inputs and introducing more sparsity and diversity for attention may improve performance and interpretability. To explore the intra-modality interactions learned by ECE in collaborative block, we in Fig. 6 visualize the multi-head attention maps from last blocks of ECE. For comparison, we also visualize the multi-head self-attention maps from the last blocks of “Transformer-Mixed” [42] and KNN (K = 5) selection heatmaps of all layers in DGCNN [30], where a lighter color indicates a closer relationship. The KNN results of DGCNN show that the feature aggregation of one node only pays attention to the top K similar features of other nodes instead of all the nodes, and relies on the choice of K. The attention maps of Transformer-Mixed present equilibrium status, which lacks sparsity and diversity. Comparatively, our “ECE-Mixed” taking mixed features presents more diversified attention maps in eight heads, which indicates ECE can more effectively capture context information. Moreover, the attention maps generated by “ECE-Individual” show different intriguing focus patterns for different features. Specifically, ECE prefers to extract interactions for appearance and geometry features in global scope while content features bring more local focus patterns.

![Figure 5. Diversities of attention maps for different modalities with or without CCS. Solid lines (~ w/ CCS) represent the diversity distributions of attention in CCS when one modality features are regarded as queries and others as keys/values. Dashed lines (~ w/o CCS) present diversity of attention weights in ECE for each modality.](image)

How do different modalities collaborate with each other? To investigate the working pattern of CCS, we adopt Jensen-Shannon Divergence [4] (see supplementary materials) to measure the average diversity of attention map in CCS when the model also takes input table image shown in Fig. 6. As shown in Fig. 5, solid lines (~ w/ CCS) represent the diversity distributions when one modality features are regarded as queries and others as keys/values. After removing CCS, diversity of attention weights in ECE for each modality is also presented by dashed lines (~ w/o CCS). For those with CCS, the higher value indicates the query modality is in a closer collaboration with the other modalities. Particularly, appearance modality has the strongest collaborative relationship with others while geometric one requires the least collaboration. By comparison, the diversities of attention weights in ECE also follow a similar trend, but with lower values on average.
Figure 6. Visualization of the heat-maps generated by DGCNN and multi-head attention maps from the Transformer and ECE. Y-axis (red) and X-axis (blue) are “probes” and “candidates” respectively. For ECE, probes are graph node features and candidates are edge combined features. For Transformer and DGCNN, probes and candidates are both non-graph features. The heat-maps of DGCNN show a local hard selection way in terms of context. And Transformer yields attention maps lacking sparsity and diversity. In contrast, ECE-Mixed presents more diversified attention maps and ECE-Individual extracts interactions in global or local pattern conditioned on different features. Best viewed in color.

Figure 7. The relationship between block number of NCGM and F1-score on SciTSR-COMP dataset.

The more collaborative blocks, the better performance?
To further explore the effect of the collaborative block number on the NCGM performance, we conduct a set of experiments setting block numbers from 1 to 9, respectively. It can be seen from Fig. 7 that it is a trade-off problem. Small block number can render faster convergence to the model. As the number increases, the performance keeps improving until block number increases to 5, but the convergence speed of the network keeps slowing down. In particular, we observe that the F1-score decreases sharply when NCGM with more than 7 blocks is trained over round 50 epochs, which indicates more blocks are easier to cause model training collapse problem. Based on the above observation, we set it to 3 as default number.

5. Conclusion and Limitation

We present a novel graph-based method for heterogeneous table structure recognition through learning intra-inter modality collaboration. Extensive experiments on public benchmarks demonstrate its superiority over state-of-the-art methods, especially under challenge scenarios. However, there exist two limitations could be improved in future. The first one is the inevitable problem of computational complexity increase brought by introducing multiple modalities and decoupled processing. The second one lies in the fact that NCGM with deeper blocks is easier to suffer from the training collapse problem. We may introduce more inductive bias into the attention model to tackle it.
References


[26] Bruno A Olshausen, Charles H Anderson, and David C Van Essen. A neurobiological model of visual attention and


[34] Sachin Raja, Ajoy Mondal, and CV Jawahar. Table structure recognition using top-down and bottom-up cues. In *European Conference on Computer Vision*, pages 70–86. Springer, 2020. 1, 2, 3, 4, 5, 6


[41] Scott Tupaj, Zhongwen Shi, C Hwa Chang, and Hassan Alam. Extracting tabular information from text files. *EECS Department, Tufts University, Medford, USA*, 1996. 2


