Neural Collaborative Graph Machines

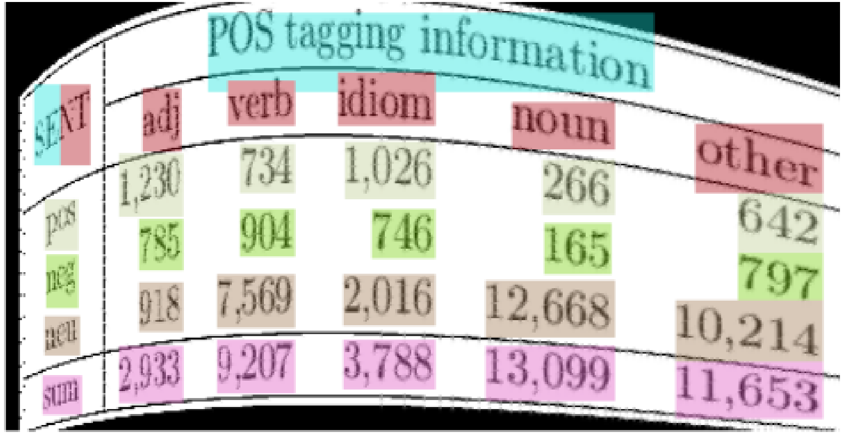
for Table Structure Recognition

CVPR 2022 Hao Liu, et al. (Tencent). Reporter-TongKe Nee, 2023/04/01

1. Background

TSR (table structure recognition) task is recognizing the physical and logical structure of tables which are usually represented by image.

The conventional digital table image recognition has been relatively mature.Recent years, some researchers have focused on more challenging table structure recognition tasks, such as **recognizing the distortion table**.



SENT	POS tagging information				
	adj	verb	idiom	noun	other
pos	1,230	734	1,026	266	642
neg	785	904	746	165	797
neu	918	7,569	2,016	12,668	10,214
sum	2,933	9,207	3,788	13,099	11,653

An example of table structure recognition[1]

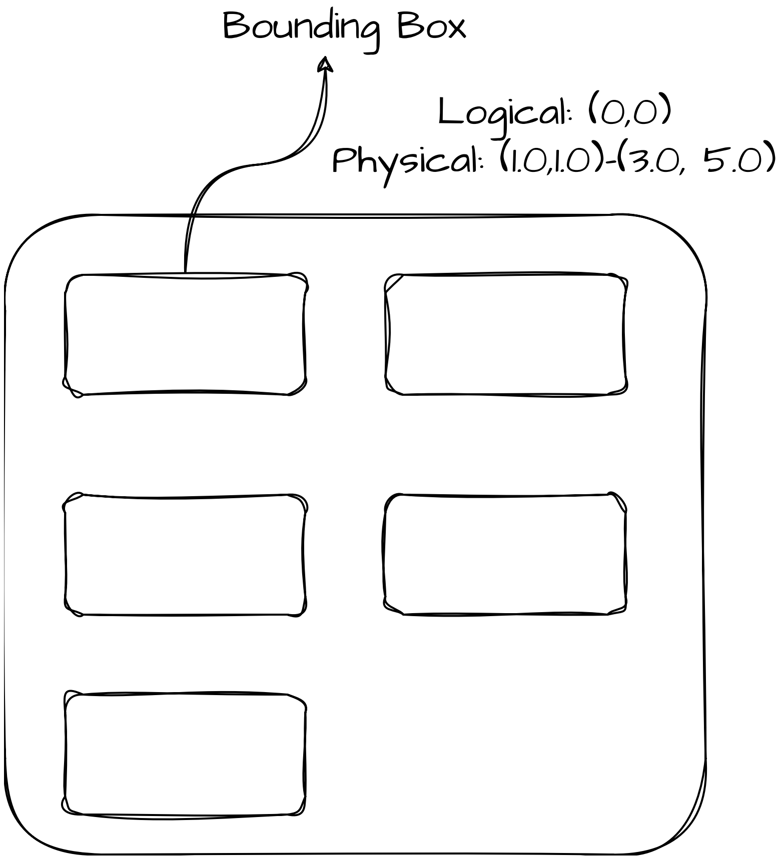
[1]Neural Collaborative Graph Machines for Table Structure Recognition

Logical Structure

Logical structure only focuses on the relationship between cells wheather they are in the same row or column or cells.

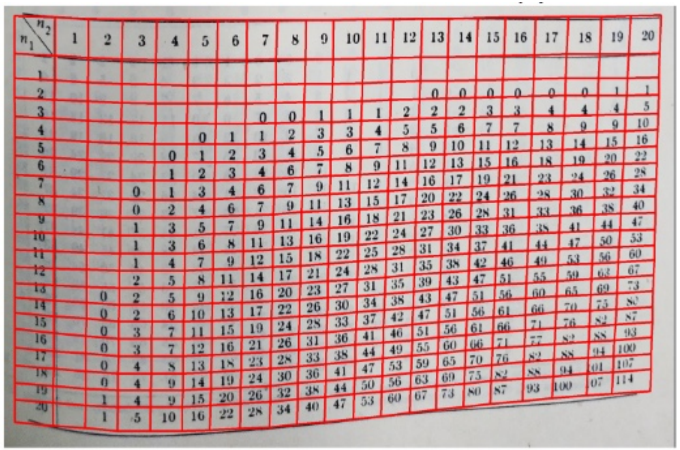
Physical Structure

Physical structure focuses on the coordinates of cells boxes.

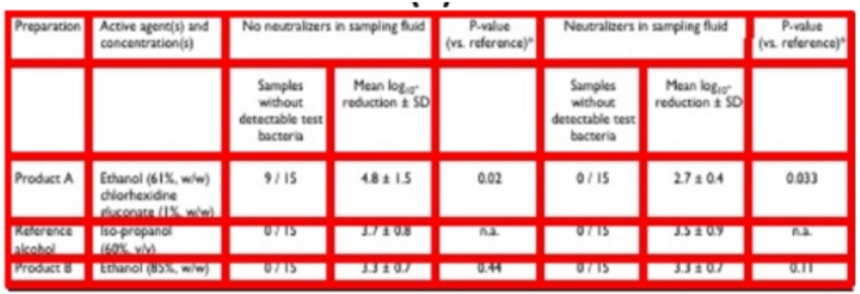


An example of physical structure and logical structure

Distorted Table and Regular Table



n_1 \ n_2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2													0	0	0	0	0	0	1	1
3							0	0	1	1	1	2	2	2	3	3	4	4	4	5
4					0	1	1	2	3	3	4	5	5	6	7	7	8	9	9	10
5				0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
6				1	2	3	4	6	7	8	9	11	12	13	15	16	18	19	20	22
7			0	1	3	4	6	7	9	11	12	14	16	17	19	21	23	24	26	28
8			0	2	4	6	7	9	11	13	15	17	20	22	24	26	28	30	32	34
9			1	3	5	7	9	11	14	16	18	21	23	26	28	31	33	36	38	40
10			1	3	6	8	11	13	16	19	22	24	27	30	33	36	38	41	44	47
11			1	4	7	9	12	15	18	22	25	28	31	34	37	41	44	47	50	53
12			2	5	8	11	14	17	21	24	28	31	35	38	42	46	49	53	56	60
13		0	2	5	9	12	16	20	23	27	31	35	39	43	47	51	55	59	63	67
14		0	2	6	10	13	17	22	26	30	34	38	43	47	51	56	60	65	69	73
15		0	3	7	11	15	19	24	28	33	37	42	47	51	56	61	66	70	75	80
16		0	3	7	12	16	21	26	31	36	41	46	51	56	61	66	71	76	82	87
17		0	4	8	13	18	23	28	33	38	44	49	55	60	66	71	77	82	88	93
18		0	4	9	14	19	24	30	36	41	47	53	59	65	70	76	82	88	94	100
19		1	4	9	15	20	26	32	38	44	50	56	63	69	75	82	88	94	101	107
20		1	5	10	16	22	28	34	40	47	53	60	67	73	80	87	93	100	107	114



Preparation	Active agent(s) and concentration(s)	No neutralizers in sampling fluid		P-value (vs. reference)*	Neutralizers in sampling fluid		P-value (vs. reference)*
		Samples without detectable test bacteria	Mean log_{10}-reduction ± SD		Samples without detectable test bacteria	Mean log_{10}-reduction ± SD	
Product A	Ethanol (61%, w/w) chlorhexidine gluconate (1%, w/w)	9 / 15	4.8 ± 1.5	0.02	0 / 15	2.7 ± 0.4	0.033
Reference alcohol	Iso-propanol (60%, v/v)	0 / 15	3.7 ± 0.8	n.a.	0 / 15	3.5 ± 0.9	n.a.
Product B	Ethanol (85%, w/w)	0 / 15	3.3 ± 0.7	0.44	0 / 15	3.3 ± 0.7	0.11

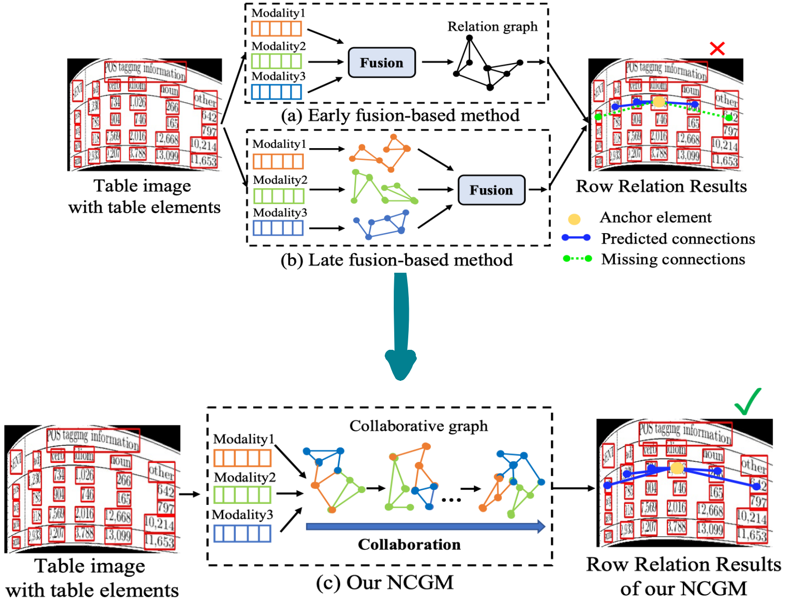
Distorted vs Regular Table[1]

[1]TSRFormer: Table Structure Recognition with Transformers

Comparison

Traditional multi-modal fusion methods lack of attention to modal collaboration, so they do not perform well on distorted tables.

NCGM utilize **graph** to construct different modality relationship, and achieves good performance on the **distorted tables**.



Early (Late) fusion-based method with NCGM.

Related Work

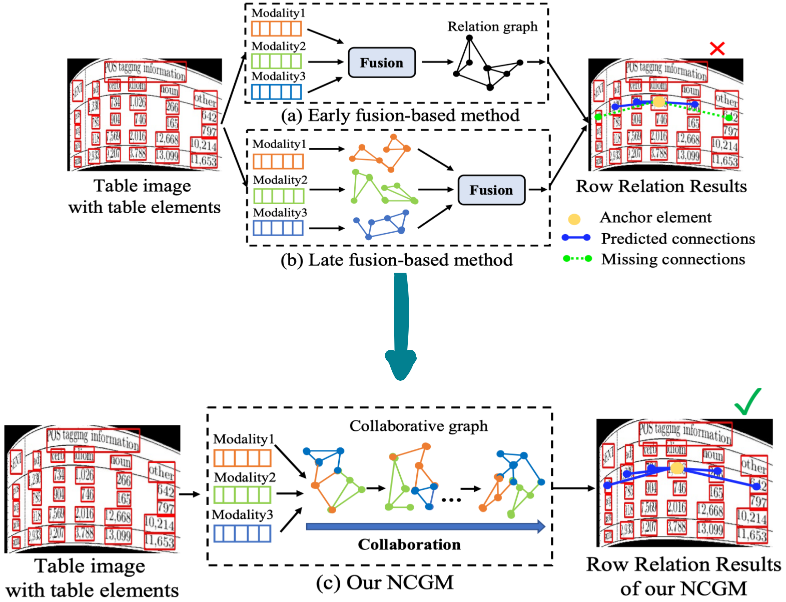
After entering the era of deep learning, there are several main ways to identify table structure:

Boundary extraction-based

Based on the boundary extraction technology, the table structure is identified by detecting the **horizontal and vertical separator** in the table.

Generative model-based

This method uses the **encoder-decoder** schema to generate HTML text to represent the table structure.



Early (Late) fusion-based method with NCGM.

Related Work

After entering the era of deep learning, there are several main ways to identify table structure:

Graph-based

This method is mainly based on the **graph** model, and describes the table structure by representing the table elements as nodes and the relationship between them as edges.

Transformer-based multimodal

Use early fused embedding as input (VLBERT, LayoutLM etc.), and the method of using two modalities to do co-attentional fusion (ViLBERT).

Motivation

In table recognition, inductive biases of different modalities may affect the performance of the modality.

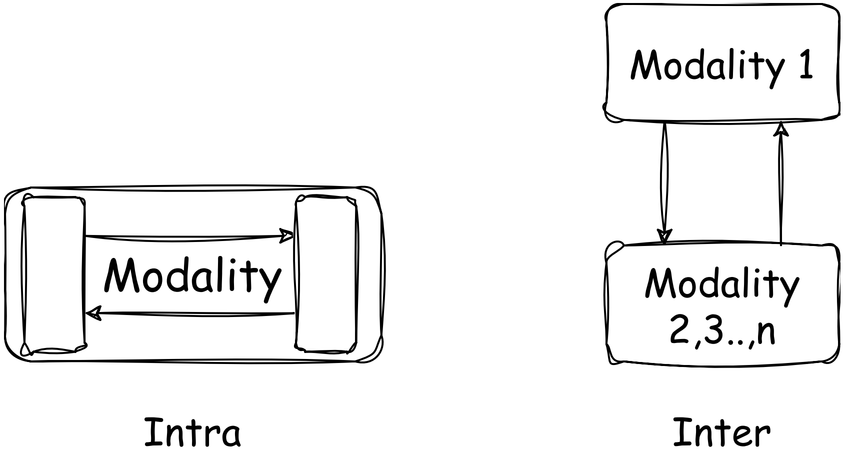
For example, when identifying regular tables, the coordinates of tables will dominate, but when dealing with distorted tables, they will become unreliable.

Hetero-TSR Problem

Can different modalities **collaborate** with each other rather than **interfering** with each other?

Tow aspects to utilize the modalities.

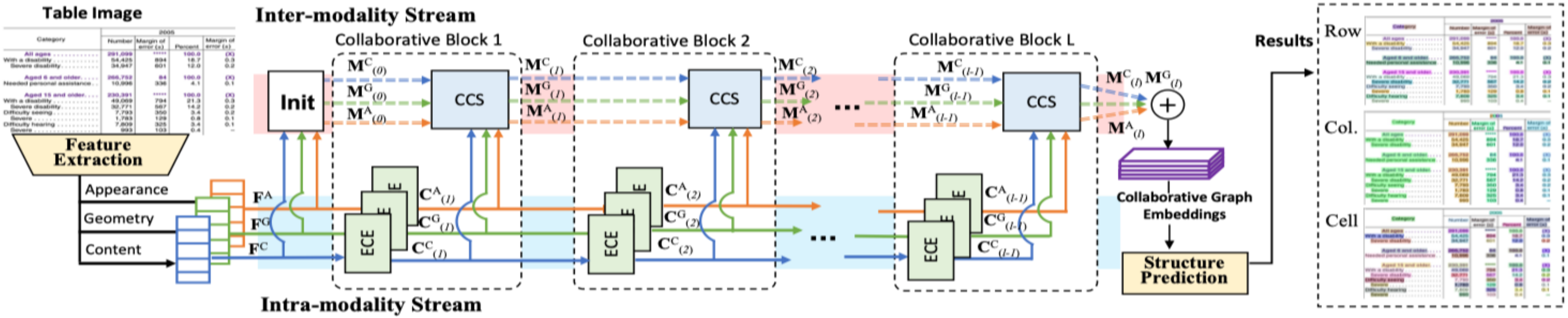
- **Intra-Modality** (Among the modality)
- **Inter-Modality** (Between the modalities)



Intra-modality and Inter-modality.

2. Model Architecture

1. Three modalities: **Geometry**, **Appearance** and **Content**
2. Collaborative block: **ECE** (Intra modality) and **CCS** (Inter modality)
3. Structure prediction: **Fully connected layers**



The architecture of NCGM.[1]

Feature Extraction

Three different modalities are used in the NCGM model, namely **Geometry**, **Appearance** and **Content**.

Geometry

$\mathbf{F}^{\mathrm{G}} \in \mathbb{R}^{N \times d}$, Geometry modality mainly uses the **spatial information** of the cell.

Appearance

$\mathbf{F}^{\mathrm{A}} \in \mathbb{R}^{N \times d}$, Appearance modality mainly uses cell **visual information**, background color, pixel information and so on.

N denotes the number of text segment bounding boxes.

Feature Extraction

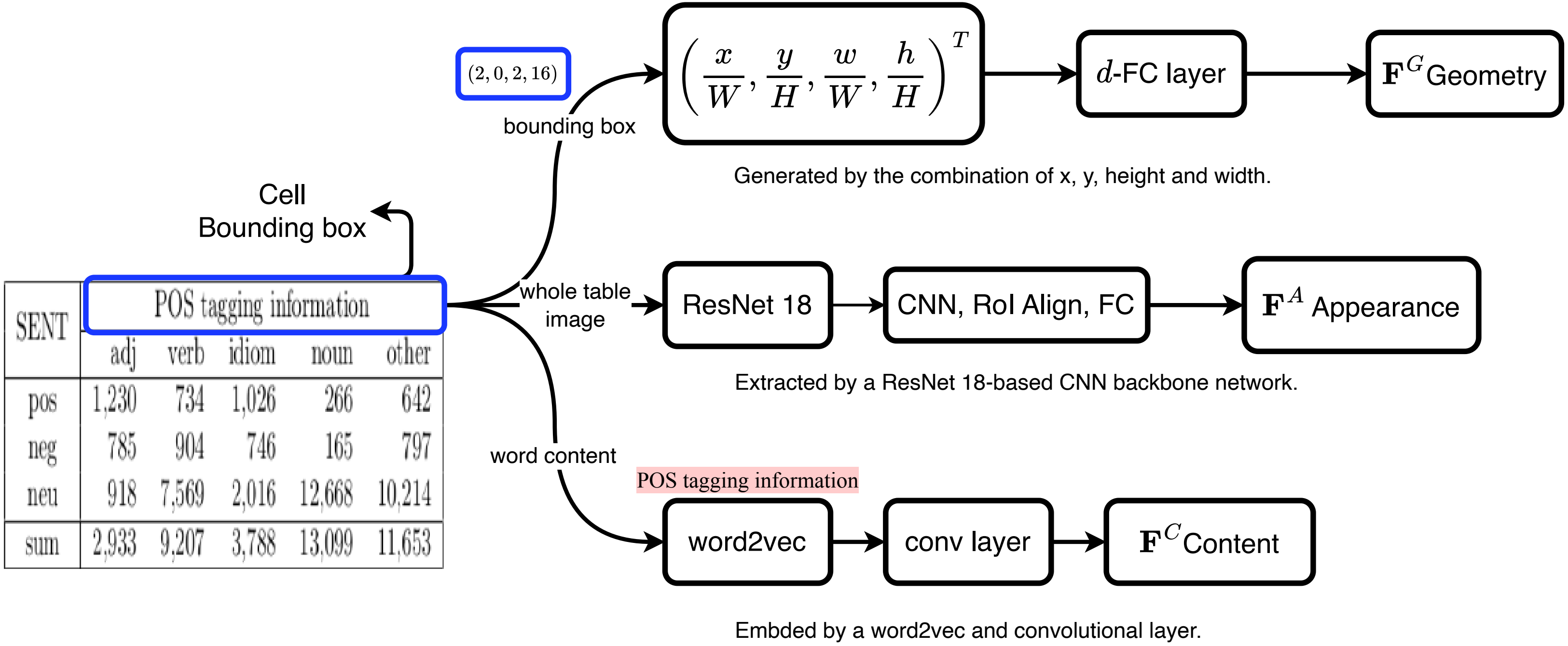
Three different modalities are used in the NCGM model, namely **Geometry**, **Appearance** and **Content**.

Content

$\mathbf{F}^{\mathrm{C}} \in \mathbb{R}^{N \times d}$, The Content modality mainly uses the **text content** of the cell.

N denotes the number of text segment bounding boxes.

Feature Extraction



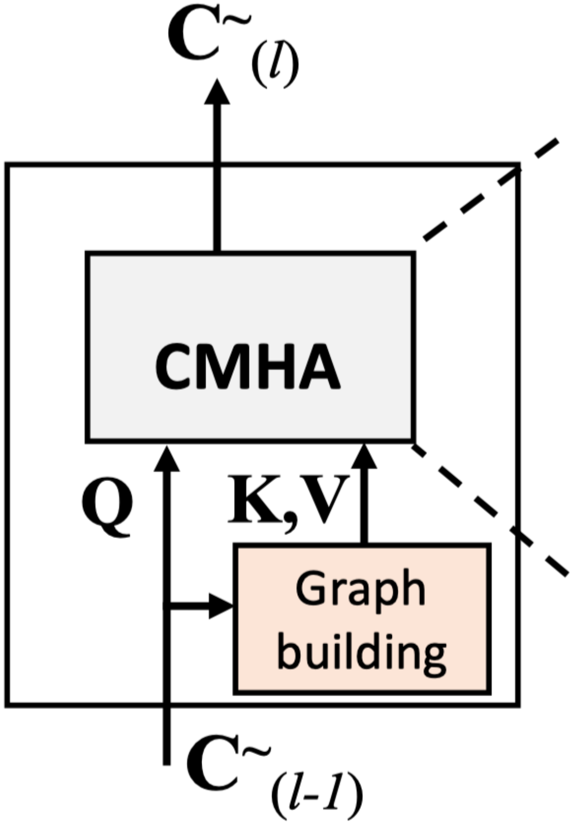
SENT	POS tagging information				
	adj	verb	idiom	noun	other
pos	1,230	734	1,026	266	642
neg	785	904	746	165	797
neu	918	7,569	2,016	12,668	10,214
sum	2,933	9,207	3,788	13,099	11,653

Three modalities

ECE Module

Ego Context Extractor module (ECE) is used to extract the context information of intra modality. The feature of modality that input ECE will be constructed as a graph.

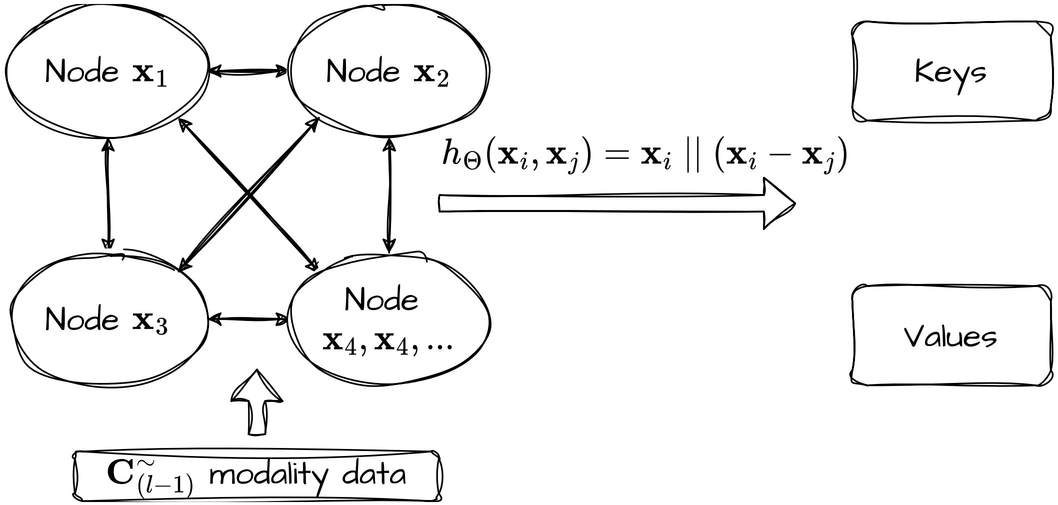
This input $\mathbf{C}^{\sim}_{(l-1)}$ is copied into two copies, one as the **Query** input to CMHA and the other into the Graph building module to model the intra-modality context which is taken as Keys $\mathbf{K}$ and values $\mathbf{V}$.



ECE Module Architecture

ECE Module

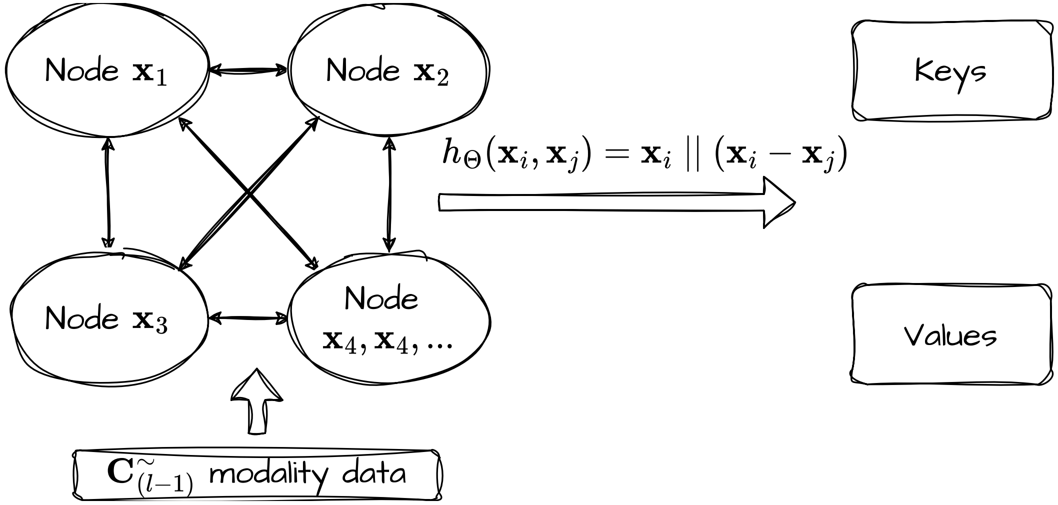
$\mathbf{G}^{\sim} = \{\mathcal{V}, \mathcal{E}\} \in \left\{\mathbf{G}^{\mathrm{G}}, \mathbf{G}^{\mathrm{A}}, \mathbf{G}^{\mathrm{C}}\right\}$. The nodes in $\mathcal{V}$ is the node (cell) feature set, and $\mathcal{E}$ is the full connected graph edge set.



Graph Building Module

ECE Module

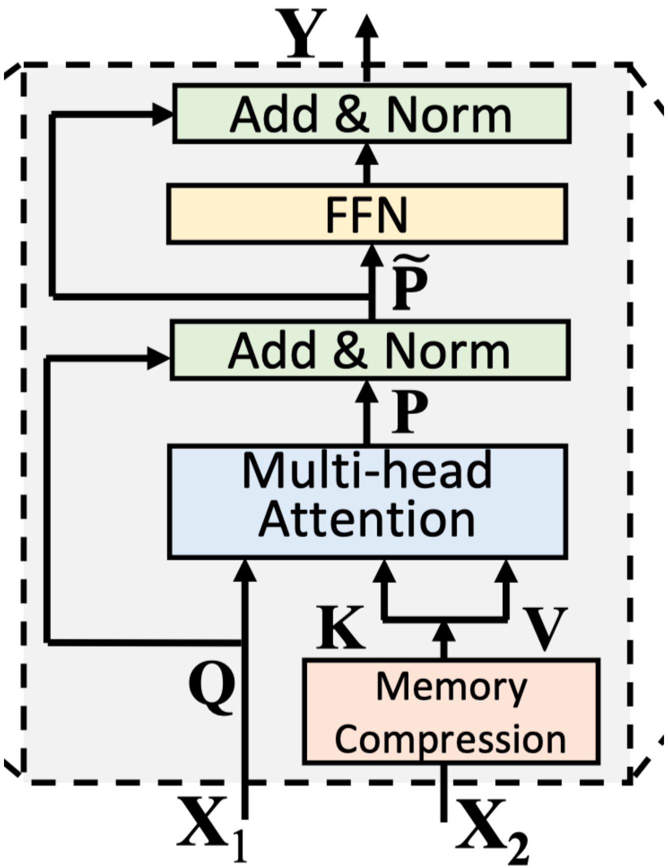
Authors adopt the following asymmetric edge function to combine graph edge features to each node.



[1]Kronecker product operator

CMHA Module

Compressed Multi-Head Attention (CMHA) module which has been verified that it makes few assumptions about inputs and can learn to combine local behavior and global behavior on input stream.

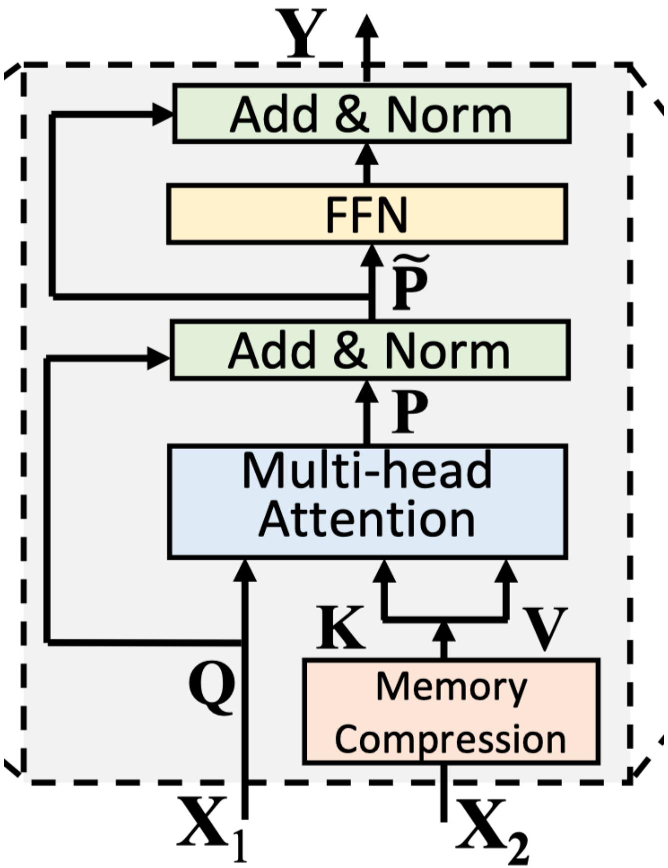


Compressed Multi-head Attention Module

CMHA Module

In order to reduce the amount of computation caused by too many dimensions, we introduce a memory compression module. In addition, we also introduce residual links to make the query information flow unimpeded.

$$\mathbf{Y} = Add\&\,\mathrm{Norm}(FFN(\widetilde{\mathbf{P}}), \widetilde{\mathbf{P}})$$
$$\widetilde{\mathbf{P}} = Add\&\,\mathrm{Norm}(\mathbf{Q}, \mathbf{P}),$$
$$\mathbf{P} = MHA(\mathbf{Q}, MC(\mathbf{K}), MC(\mathbf{V})),$$

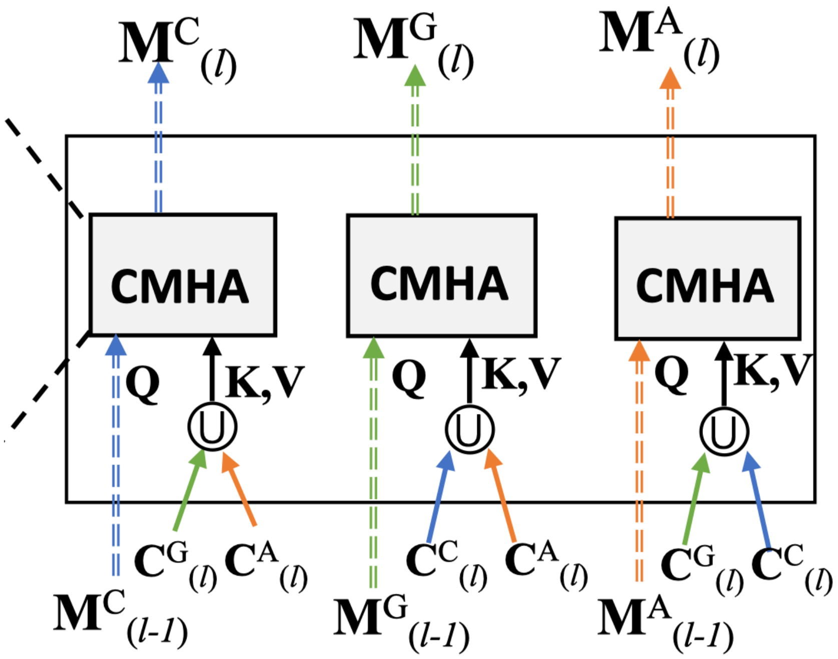


Compressed Multi-head Attention Module

CCS Module

The Cross Context Synthesis module (CCS) fuses heterogeneous context graph embeddings collaboratively and learns collaborative patterns between different modalities.

Specifically, it takes three modality inputs as queries, and the union of the other two modalities as keys and values inputs to enable query modality to fully learn information from the other two modalities.



CCS Module Architecture

Structure Prediction

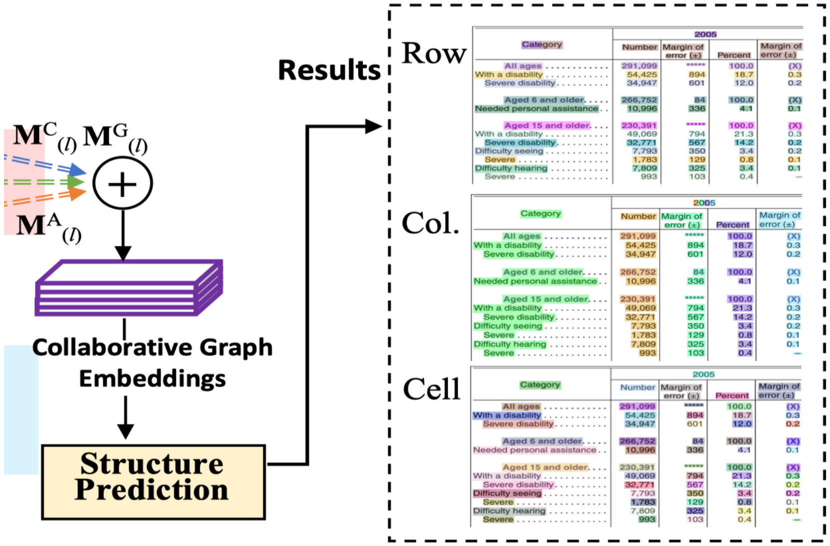
Based on the output embeddings

$$\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N\} \in \mathbb{R}^{N \times d_e}$$

a set of node pairs is constructed, where each element is a vector formed by two node vectors concatenate, and then predicted by the full connection layer.

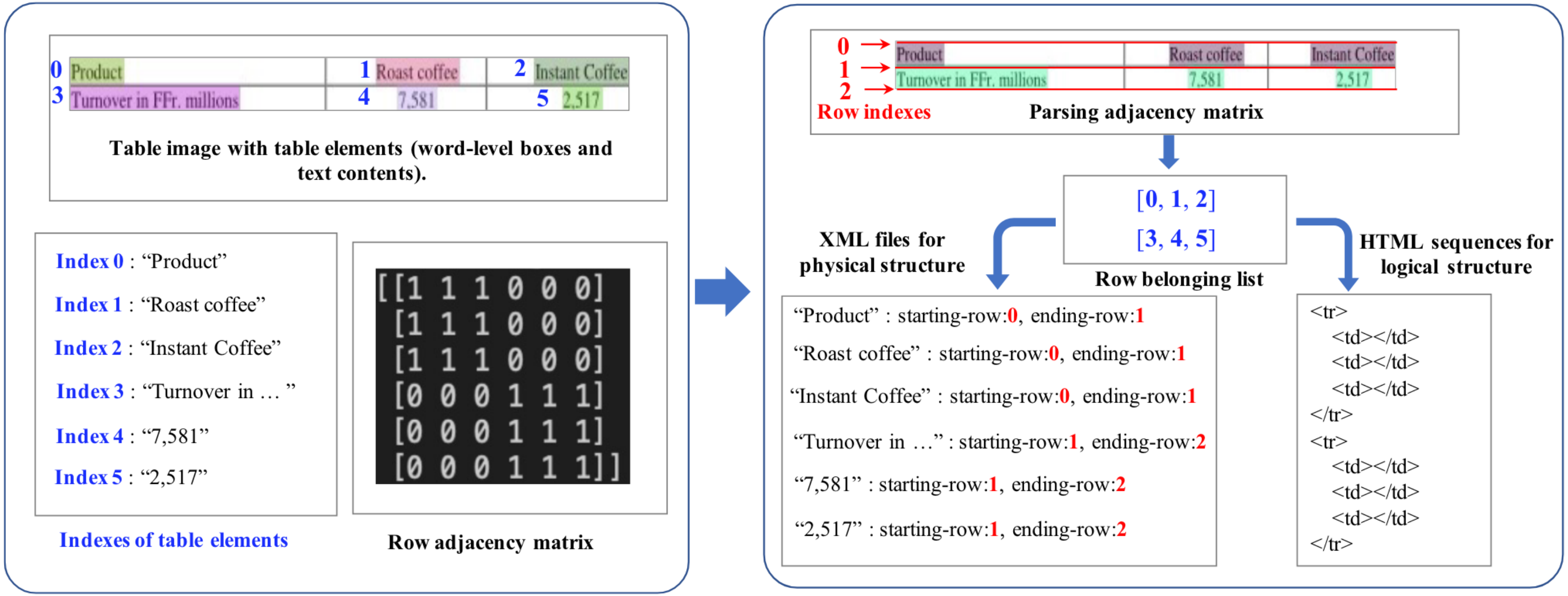
$$\mathbf{U} = \{\mathbf{u}_{1,1}, \mathbf{u}_{1,2}, \ldots, \mathbf{u}_{i,j}, \ldots, \mathbf{u}_{N,N}\} \in \mathbb{R}^{N^2 \times 2d_e}$$

Is to predict whether two nodes (cell) are in the same row, the same column, and to restore the table structure.



Predicting the structure of the table.

Structure Prediction



Post processing. Convert adjacency matrix containing relationships to spanning information.

3. Result

Evaluation setting: Different methods utilize different information. Some methods use the cell/text bounding box, while others do not. Therefore, they design two different steps:

- **Step A**: Only with table image
- **Step B**: Along with the cell/text segment bounding box and text content

Method	Train Dataset	Setup-A P	Setup-A R	Setup-A F1	Setup-B P	Setup-B R	Setup-B F1
ICDAR-2013-P							
DGCNN [34]	Sci. + IC13-P	-	-	-	98.6	99.0	98.8
TabStr. [38]	Sci. + IC13-P	93.0	90.8	91.9	99.1	99.3	99.2
GTE [49]	Pub. + IC13-P	94.4	92.7	93.5	-	-	-
LGPMA [35]	Sci. + IC13-P	96.7	99.1	97.9	-	-	-
C-CTRNet [26]	WTW + IC19	95.5	88.3	91.7	-	-	-
FLAG-Net [24]	Sci. + IC13-P	97.9	**99.3**	98.6	99.2	99.5	99.3
NCGM	Sci. + IC13-P	**98.4**	**99.3**	**98.8**	**99.3**	**99.9**	**99.6**
ICDAR-2019							
DGCNN [34]	Sci. + IC19	80.3	77.8	79.0	-	-	-
TabStr. [38]	Sci. + IC19	82.2	78.7	80.4	97.5	95.8	96.6
C-CTRNet [26]	WTW	-	-	80.8	-	-	-
FLAG-Net [24]	Sci. + IC19	**85.2**	83.8	84.5	96.1	96.3	96.2
NCGM	Sci. + IC19	84.6	**86.1**	**85.3**	**98.9**	**98.8**	**98.8**

ICDAR 2013 Partial and ICDAR 2019

P: Precision, R: Recall, F1: F1 score. Metric: IoU(Coincidence ratio between the predicted area and the target area)

3. Result

SciTSR

DGCNN [34]	Sci.	-	-	-	97.0	98.1	97.6
TabStr. [38]	Sci.	92.7	91.3	92.0	98.9	99.3	99.1
LGPMA [35]	Sci.	98.2	99.3	98.8	-	-	-
FLAG-Net [24]	Sci.	**99.7**	99.3	99.5	**99.8**	99.5	99.6
NCGM	Sci.	**99.7**	**99.6**	**99.6**	99.7	**99.8**	**99.7**

SciTSR-COMP

DGCNN [34]	Sci.	-	-	-	96.3	97.4	96.9
TabStr. [38]	Sci.	90.9	88.2	89.5	98.1	98.7	98.4
LGPMA [35]	Sci.	97.3	98.7	98.0	-	-	-
FLAG-Net [24]	Sci.	98.4	98.6	98.5	98.6	99.0	98.8
NCGM	Sci.	**98.7**	**98.9**	**98.8**	**98.8**	**99.3**	**99.0**

SciTSR and SciTSR-COMP

P: Precision, R: Recall, F1: F1 score. Metric: IoU(Coincidence ratio between the predicted area and the target area)

Result

$$\mathrm{TEDS}\left(T_a, T_b\right) = 1 - \frac{\mathrm{EditDist}\left(T_a, T_b\right)}{\max\left(\left|T_a\right|, \left|T_b\right|\right)}$$

where $EditDist$ denotes tree-edit distance, and $|T|$ is the number of nodes in T. The table recognition performance of a method on a set of test samples is defined as the mean of the TEDS score between the recognition result and ground truth of each sample.[1]

TableBank

Method	Train Dataset	Setup-A BLEU
Image-to-Text [22]	TableBank	73.8
TabStruct-Net [38]	SciTSR	91.6
FLAG-Net [24]	SciTSR	93.9
NCGM	SciTSR	**94.6**

PubTabNet

Method	Train Dataset	Setup-A TEDS
EDD [50]	PubTabNet	88.3
TabStruct-Net [38]	SciTSR	90.1
GTE [49]	PubTabNet	93.0
LGPMA [35]	PubTabNet	94.6
FLAG-Net [24]	SciTSR	95.1
NCGM	SciTSR	**95.4**

Logical structure recognition

[1] Image-based table recognition: data, model, and evaluation

Ablation Study of Modality Fusion

Fusion Method	Input		Intra.			Inter.		Setup-B		
	Mix.	Ind.	DG.	Tr.	ECE	Con.	CCS	P	R	F1
Early Fusion	✓	✗	✓	✗	✗	✗	✗	96.3	97.4	96.8
	✓	✗	✗	✓	✗	✗	✗	95.1	95.6	95.3
	✓	✗	✗	✗	✓	✗	✗	97.8	98.3	98.0
Late Fusion	✗	✓	✓	✗	✗	✓	✗	96.9	98.2	97.5
	✗	✓	✗	✓	✗	✓	✗	94.9	96.1	95.5
	✗	✓	✗	✗	✓	✓	✗	98.4	98.2	98.3
NCGM	✗	✓	✗	✗	✓	✗	✓	**98.8**	**99.3**	**99.0**

Modality Fusion Abalation Study

P: Precision, R: Recall, F1: F1 score. Metric: IoU(Coincidence ratio between the predicted area and the target area)

Ablation Study of Multi-Modality

Input Modality			Setup-B		
A	G	C	P	R	F1
✓	✗	✗	89.8	47.9	62.5
✗	✓	✗	97.9	97.7	97.8
✗	✗	✓	70.5	39.0	50.2
✓	✓	✗	98.6	98.3	98.4
✗	✓	✓	98.0	95.0	96.5
✓	✗	✓	87.6	89.3	88.4
✓	✓	✓	**98.8**	**99.3**	**99.0**

Modality Fusion Abalation Study

P: Precision, R: Recall, F1: F1 score. Metric: IoU(Coincidence ratio between the predicted area and the target area)

Thinking about modalities collaboration

- What does ECE learn from the intra-modality?

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[1].

- How do different modalities collaborate with each other?

Multi-head Attention.

[1] Sparse and constrained attention for neural machine translation.

Conclusion

Authors proposed a novel graph-based method for heterogeneous table structure recognition through **intra-modality and inter-modality collaboration**.

Tests on various open data sets show the **effectiveness** of the method, and the importance of multi-modal cooperation for table structure identification.

Limitations include increased computational complexity and potential training collapse with deeper blocks. Future work can address these through refining the attention model.

Thank You!

Any questions?

2023/04/01