Augmenting Low-Resource Text Classification with Graph-Grounded Pre-training and Prompting

Zhihao Wen and Yuan Fang

School of Computing and Information Systems

Singapore Managent University

SMU

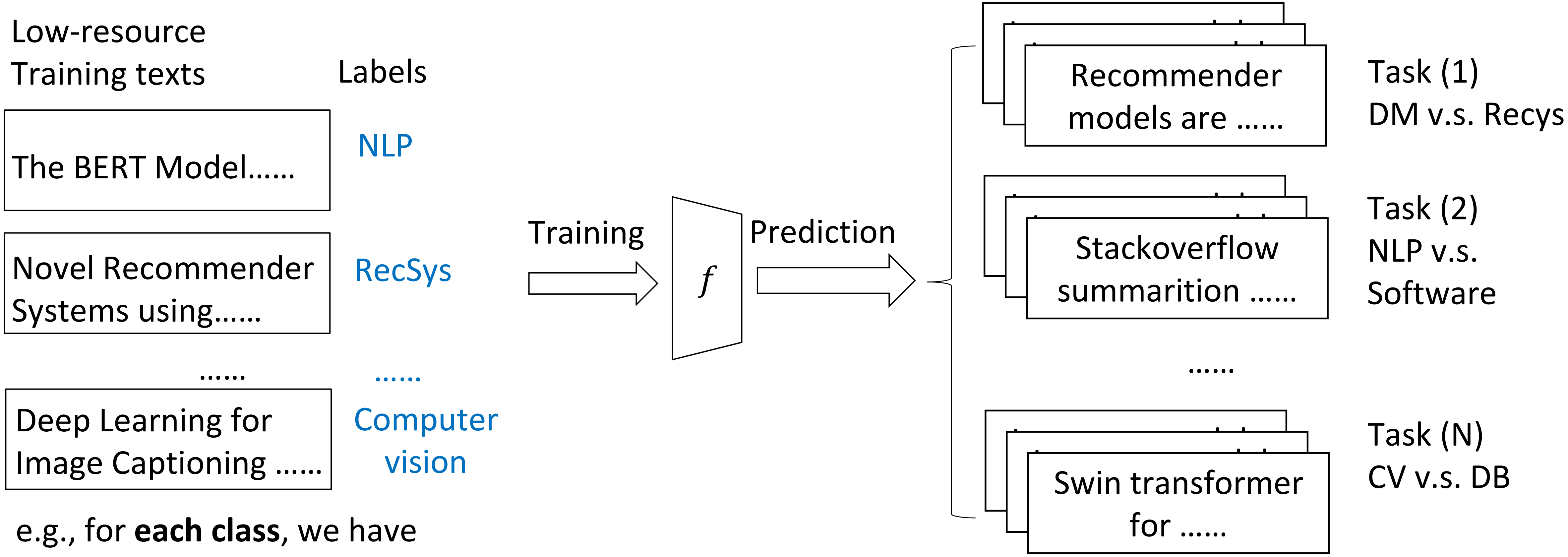
SINGAPORE MANAGEMENT UNIVERSITY

School of Computing and Information Systems

Outline

- ❑ **Introduction**
- ❑ Methodology
- ❑ Experiment
- ❑ Conclusion & Future work

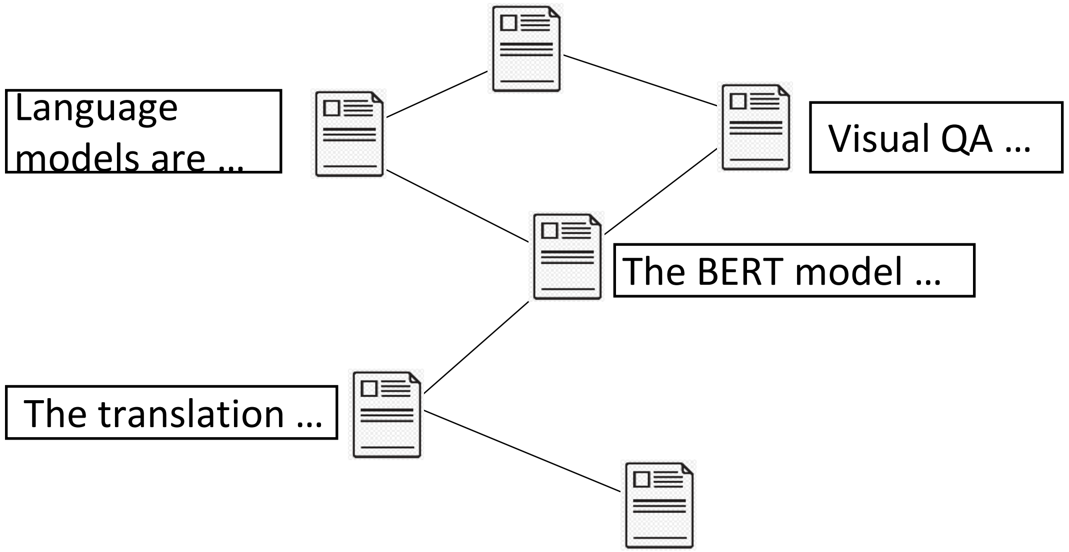
Low-resource multi-task text classification



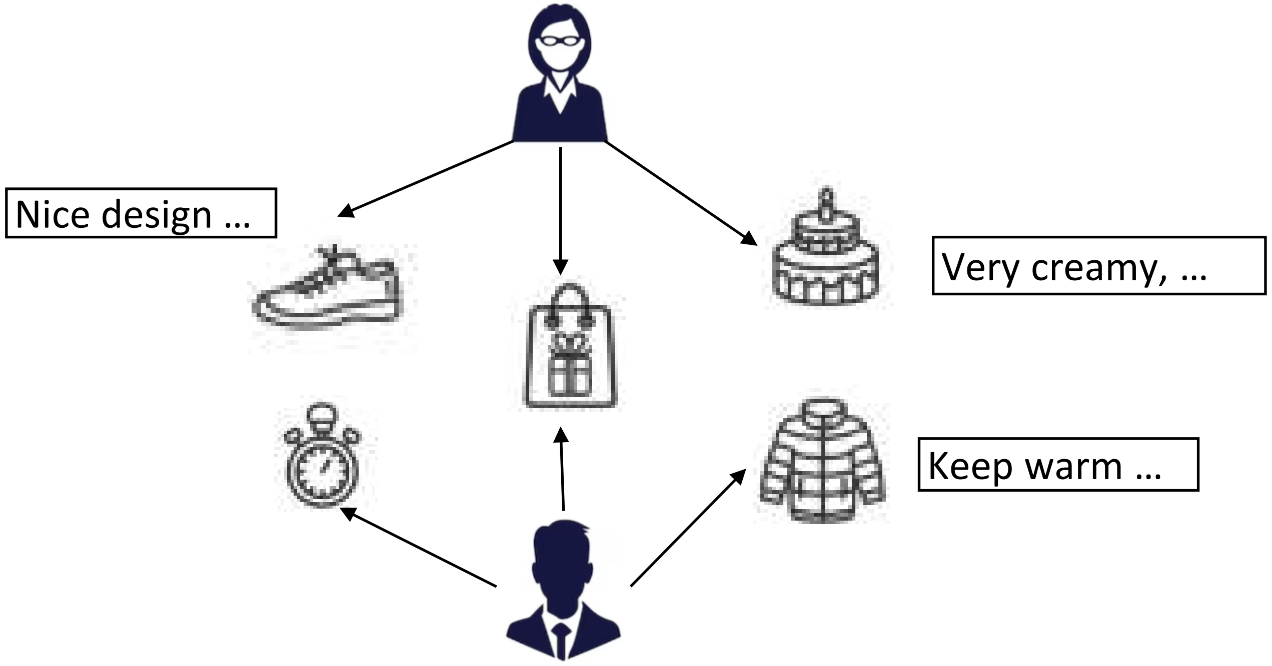
e.g., for **each class**, we have only **one labeled** training samples

Many tasks and each task is a different **text classification** task

Text data are grounded on network structures



Citation graph for online articles



E-commerce item review graph

- Text data are frequently grounded on **network structures**
- Graph structures expose valuable **relationships**
- **GNNs** are designed to learn from graph structures

Challenges and present work

Q1: How do we capture **fine-grained textual** semantics, while leveraging **graph structure** information jointly?

⇒

We propose a **graph-grounded contrastive pre-training**, to maximize the **alignment** between **text** and **graph** representations based on three types of graph interaction.

Q2: How do we **augment** low-resource multi-task text classification given a jointly pre-trained **graph-text** model?

⇒

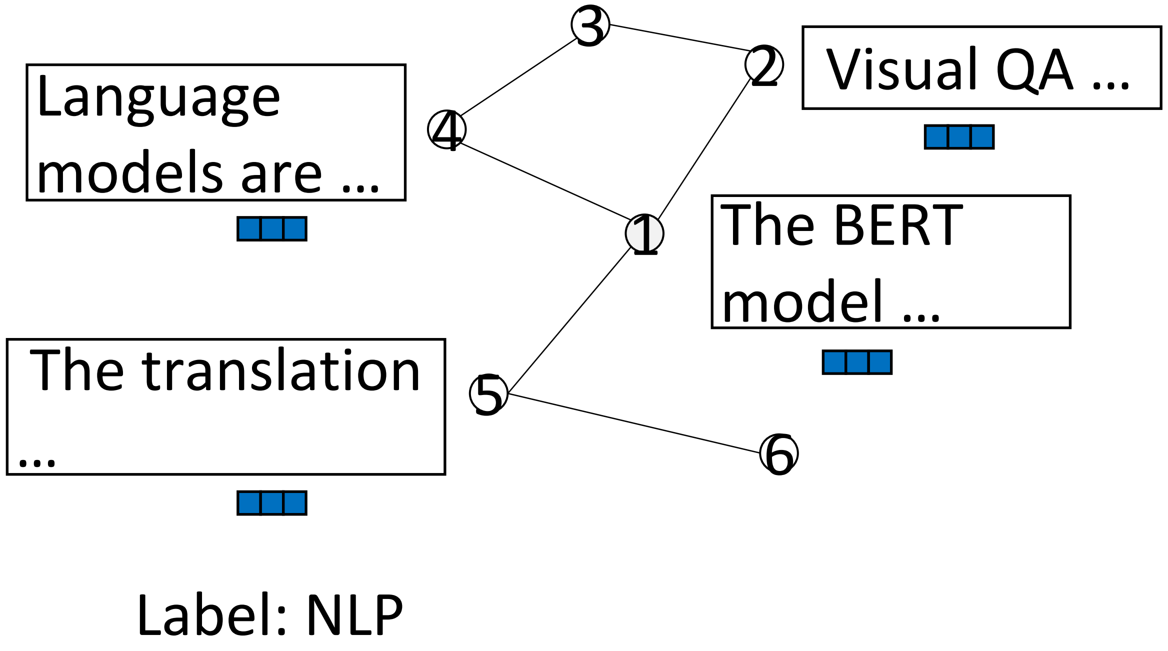
We propose a novel approach of **"prompting"** a jointly pre-trained **graph-text** model instead of fine-tuning it.

Outline

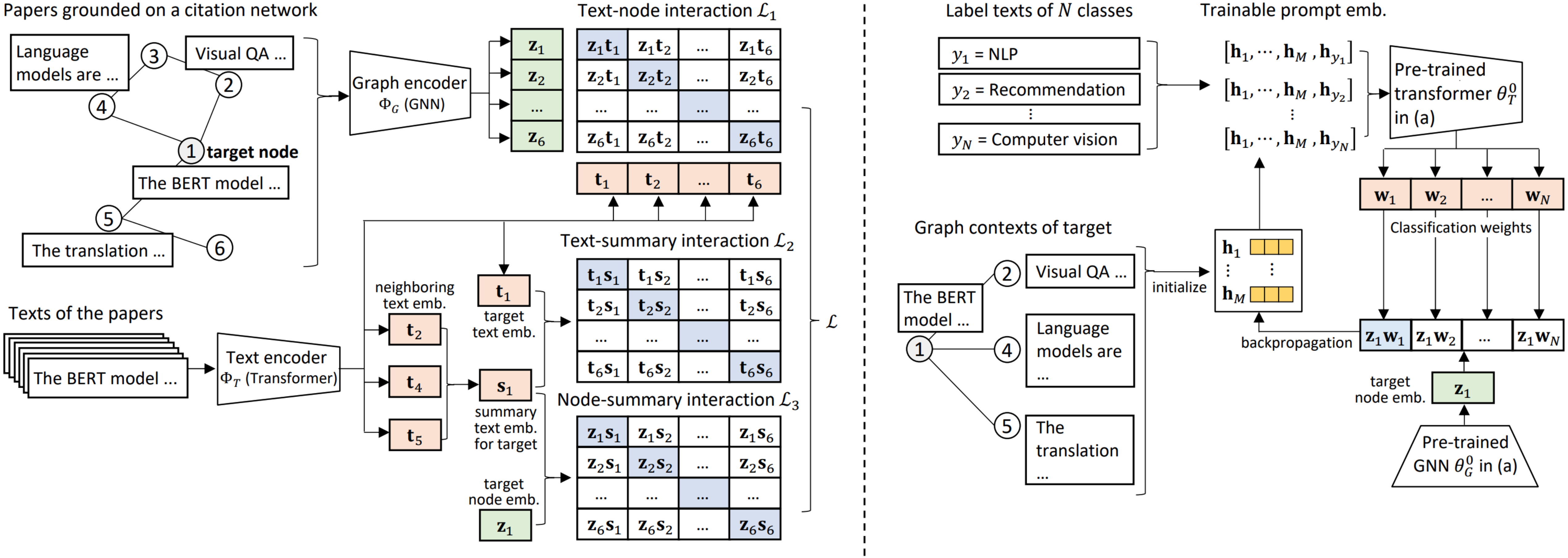
- ❑ Introduction
- ❑ **Methodology**
- ❑ Experiment
- ❑ Conclusion & Future work

Preliminary: Graph-grounded text corpus

- Consider a set of documents $\mathcal{D}$, which is grounded on a graph $\mathcal{G}$ such that each **document** d_i is a **node** v_i in the graph
- Documents are linked via **edges**
- Each node v_i is also associated with a feature vector X_i
- Each document/node has a class label



Overall framework of our proposed G2P2



(a) Graph-grounded contrastive pre-training

(b) Graph-grounded prompt tuning (few-shot classification)

Overall framework of G2P2. (a) During pre-training, it jointly trains a text and a graph encoder through three contrastive strategies. (b) During testing, it performs prompt-assisted zero- or few-shot classification

Preliminary: prompt learning

- Prompt learning in NLP: the process of **formulating** effective **prompts or instructions** to guide **pre-trained language models** to generate **desired outputs**.

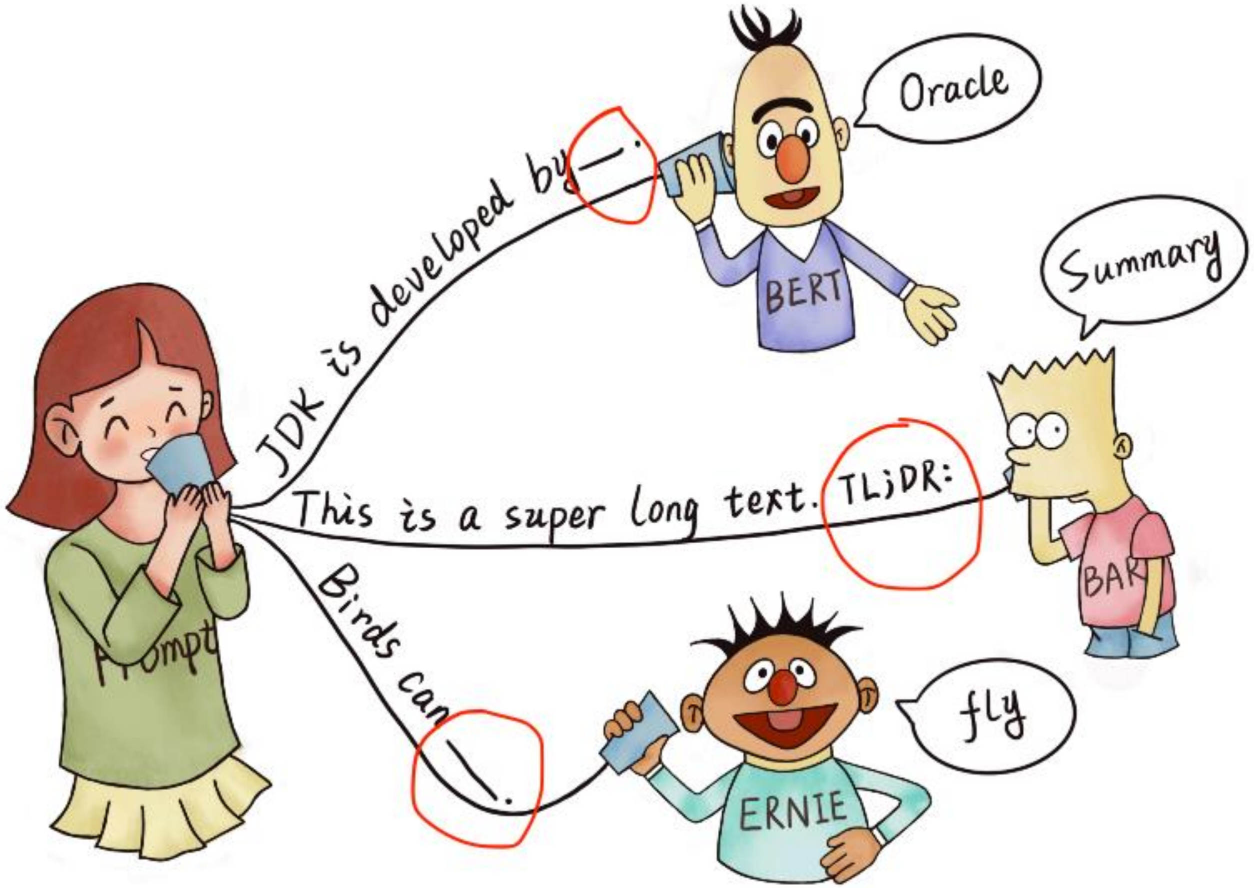
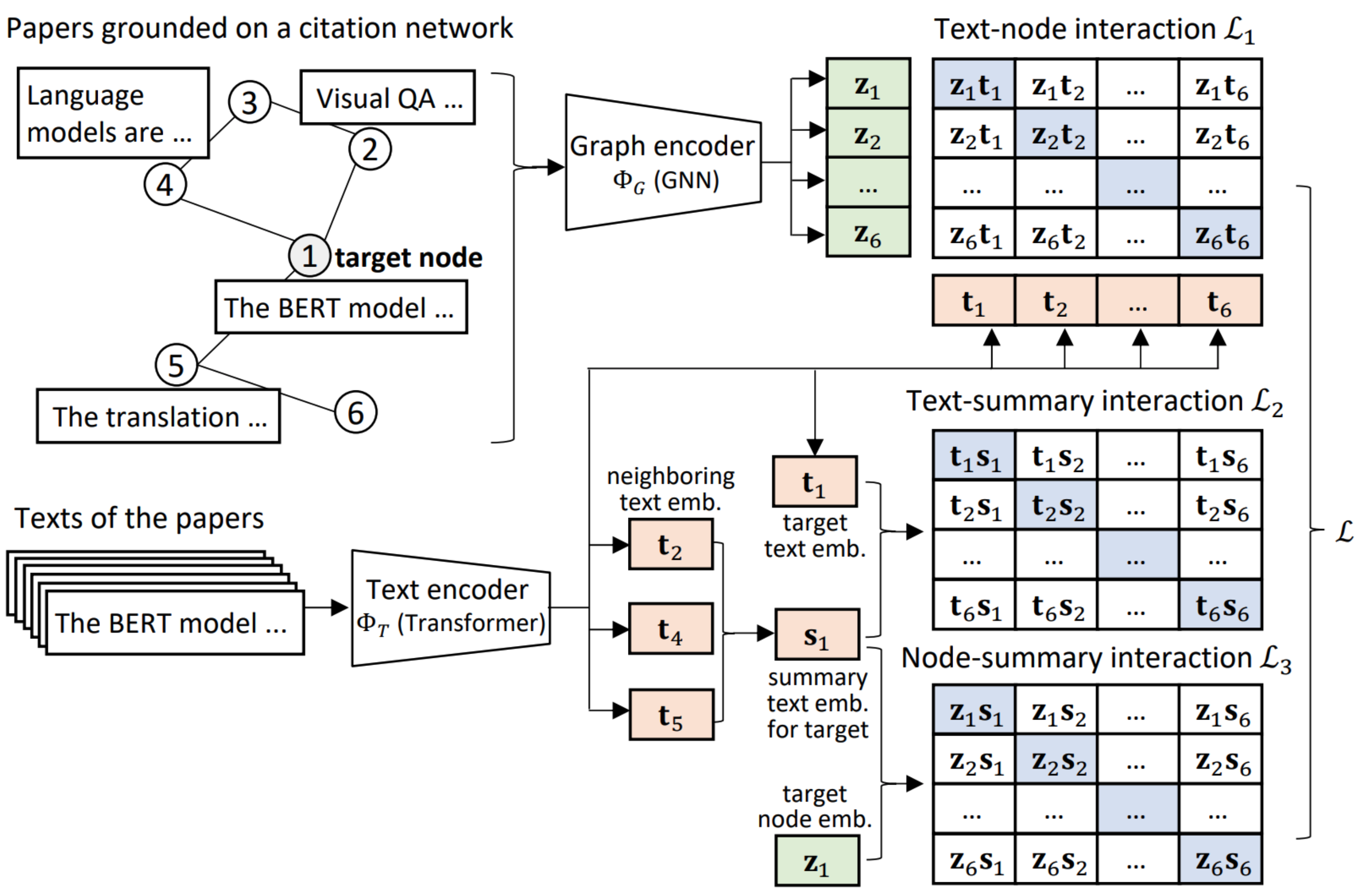


Figure from [1]

[1] Liu, Pengfei, et al. "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing." *ACM Computing Surveys 2023.*

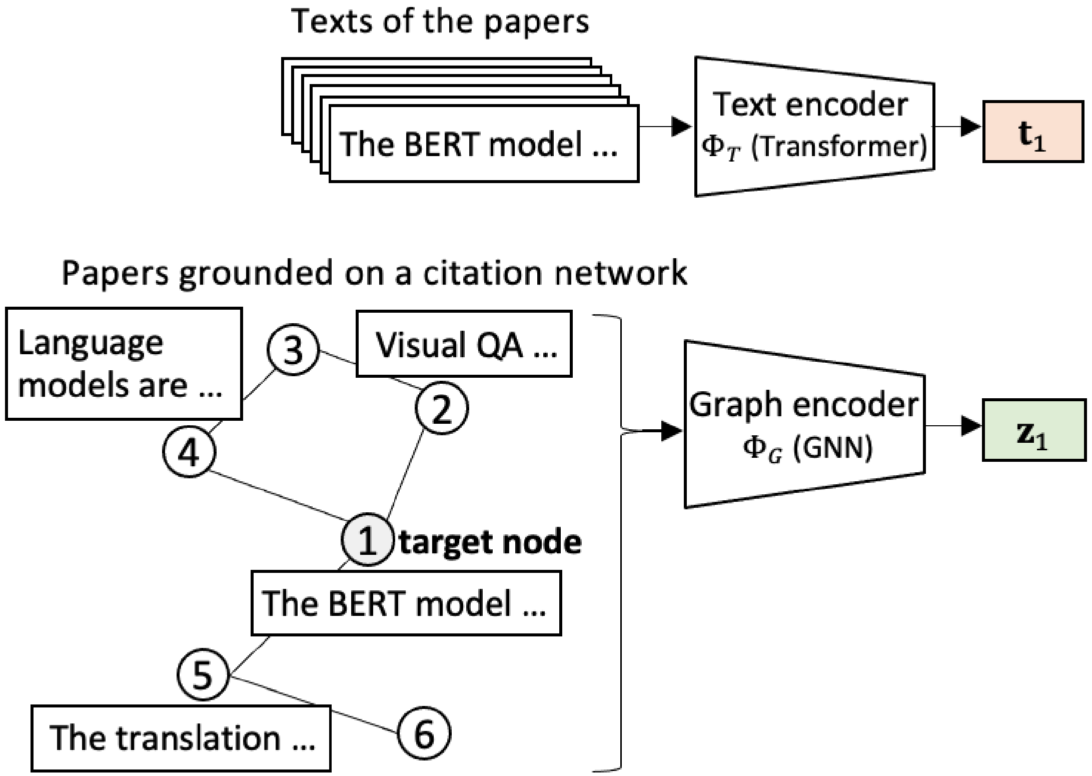
Our proposed graph-grounded contrastive pre-training



(a) Graph-grounded contrastive pre-training

- Learn a dual-modal embedding space jointly training a **text encoder** and **graph encoder** through **3 contrastive strategies**.

Graph-grounded contrastive pre-training



Dual-encoders

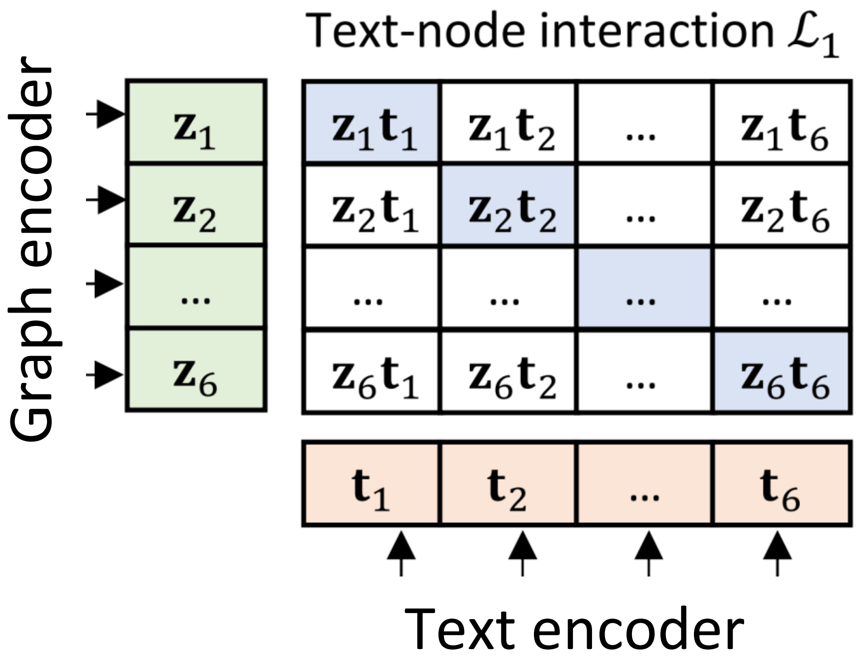
1. Text-encoder: a transformer

$$\mathbf{t}_i = \Phi_T(d_i; \theta_T)$$

2. Graph-encoder: a GCN

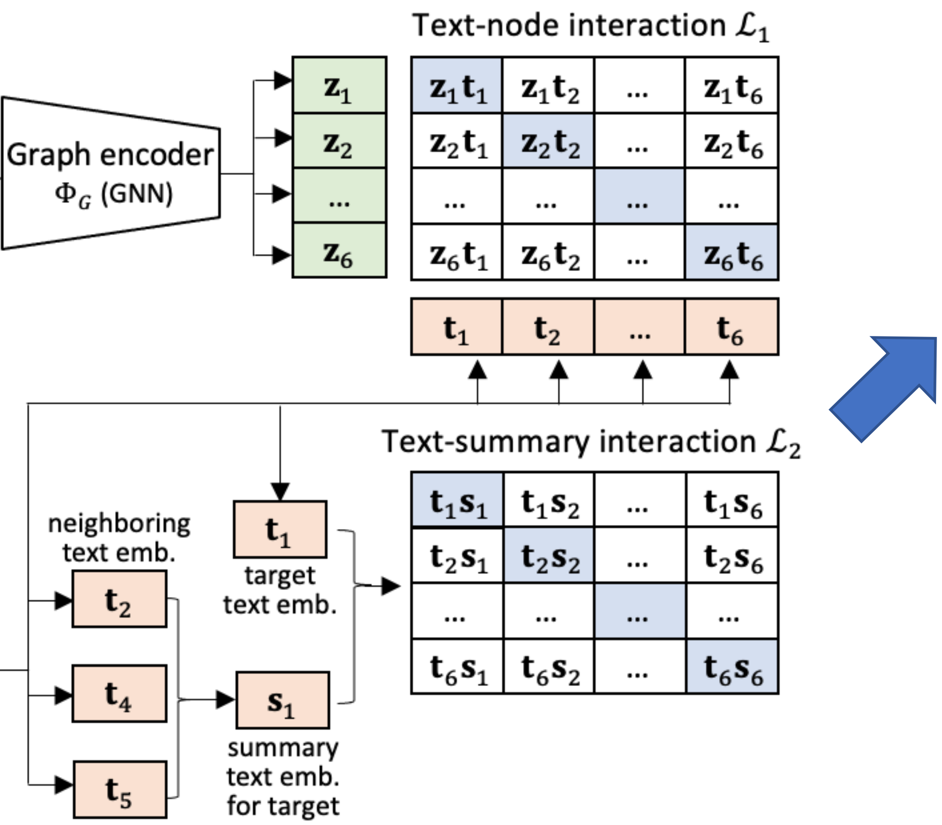
$$\mathbf{z}_i = \Phi_Z(v_i; \theta_G)$$

Text-node interaction



- Graph-grounded texts naturally implies a **bijection** between **nodes** and **texts**
- Predict the **text** of a document **matches** which **node** in the graph.
- Given **n documents** and the corresponding **n nodes**, there are **n^2** possible document node pairs
- Only n pairs with **i = j** are true matching
- The **remaining** n^2−n pairs are **false matching**
- **Maximize** the cosine similarity of n **matching** pairs, while **minimizing** that of the n^2 − n **unmatching** pairs

Text-summary interaction

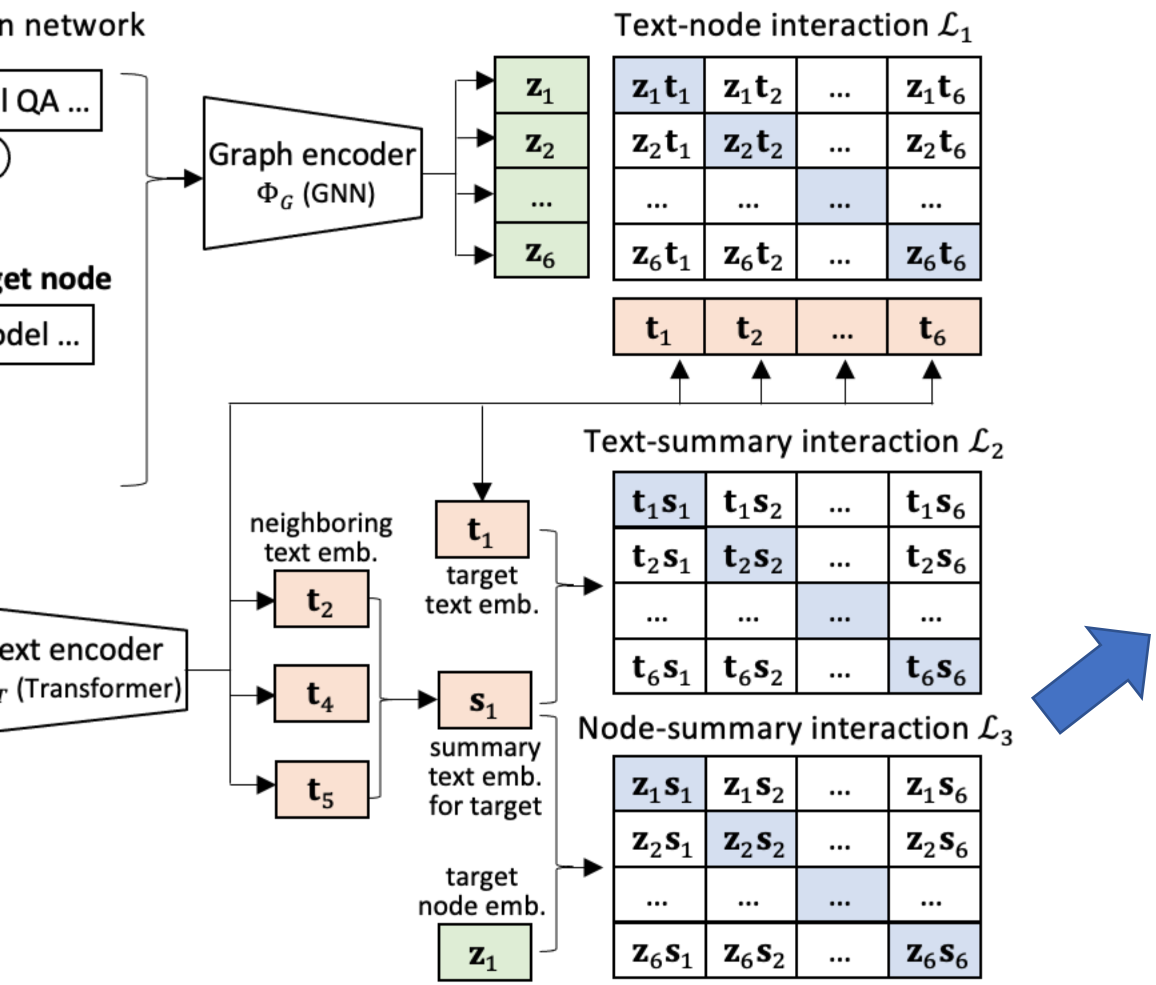


- Each document has **a set of neighboring documents** defined by graph topology
- The neighboring documents are a **summary** of the target document
- Employ a simple **mean** pooling to generate the summary embedding

$$\mathbf{s}_i = \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \mathbf{t}_j$$

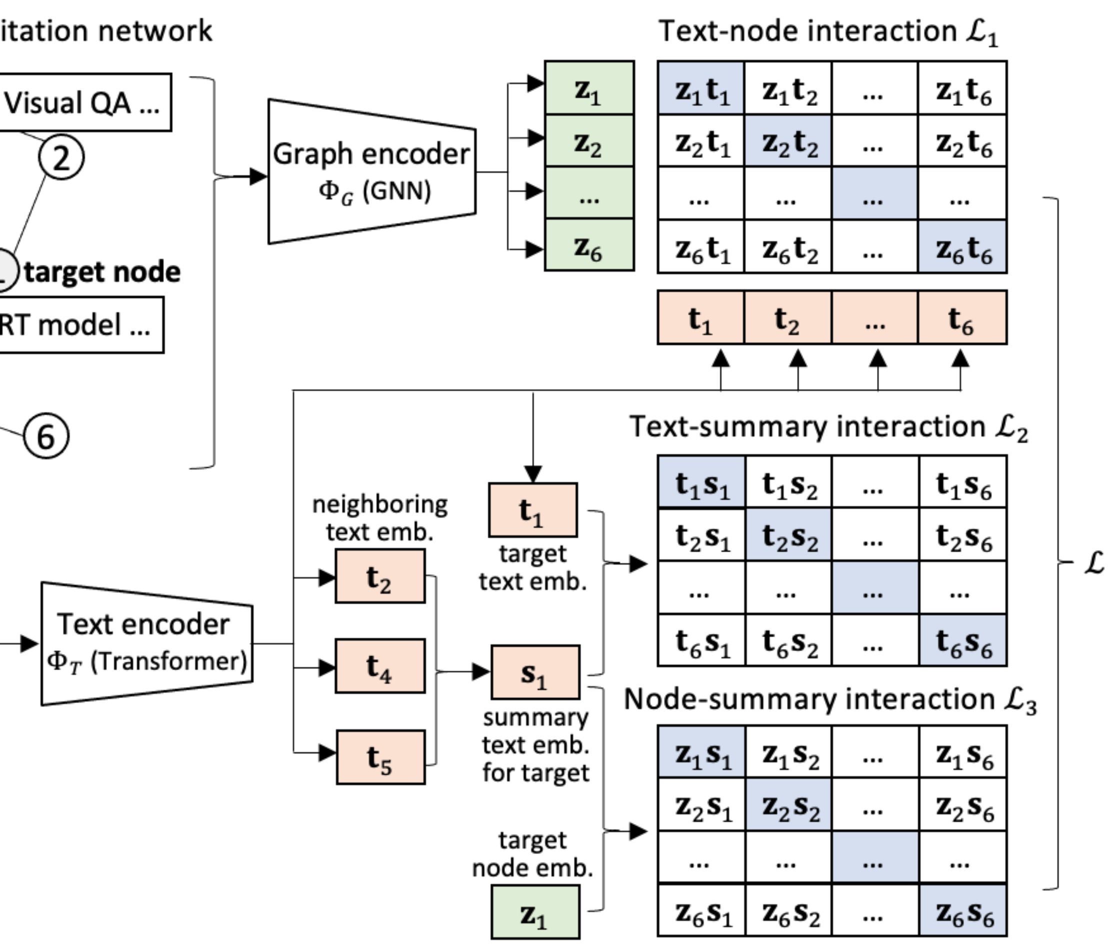
- Align the **text** embedding and its corresponding **summary** text embedding

Node-summary interaction



- Neighborhood based summary $\mathbf{s}_i$ for document d_i also serves as a semantic description of **node** $\boldsymbol{v_i}$.
- Align the **node** embedding $\mathbf{z}_i$ and its neighborhood-based **summary** text embedding $\mathbf{s}_i$.

Overall pre-training objective



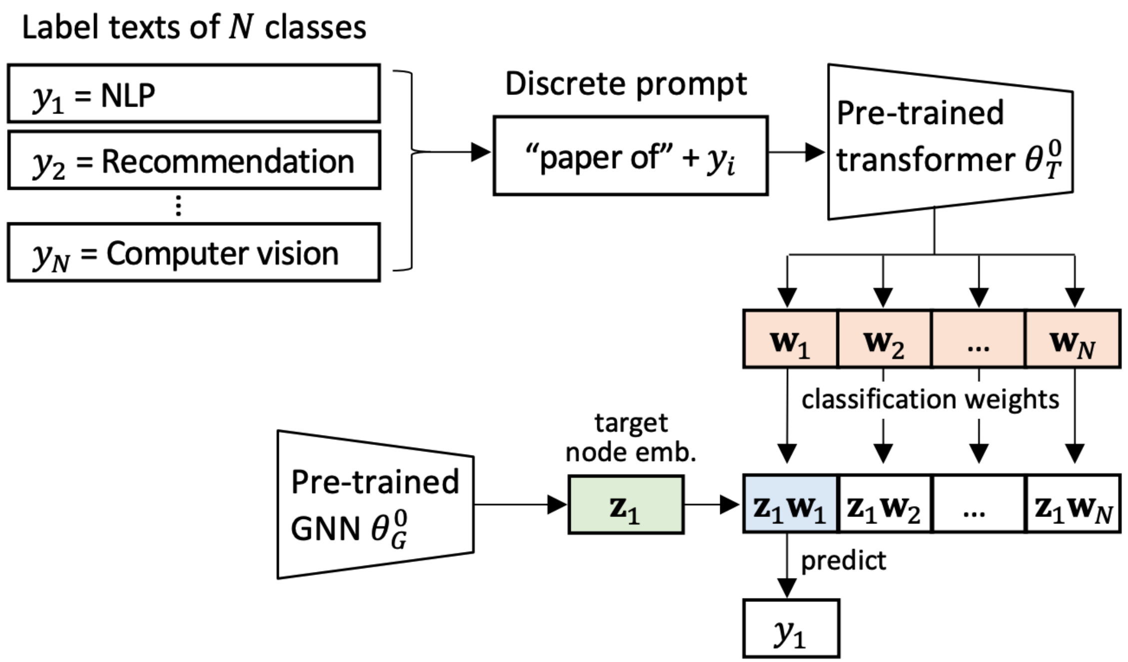
- Integrate the three contrastive losses based on the **text-node**, **text-summary** and **node-summary** interactions

- Obtain a pre-trained model $\boldsymbol{\theta^0}$ consisting of the parameters of the **dual encoders**

$$\theta^0 = \arg\min_{\theta_T, \theta_G} \mathcal{L}_1 + \lambda(\mathcal{L}_2 + \mathcal{L}_3)$$

Hyperparameter

Prompt-assisted text classification



- Discrete prompt for zero-shot classification
- Predict the class whose **label text** embedding has the **highest similarity** to the **node** embedding
- **Classification weights** can be generated by the **text encoder** based on the **class label texts**

$$\mathbf{w}_y = \phi_T(\text{“}\texttt{prompt [CLASS]}\text{”}; \theta_T^0)$$

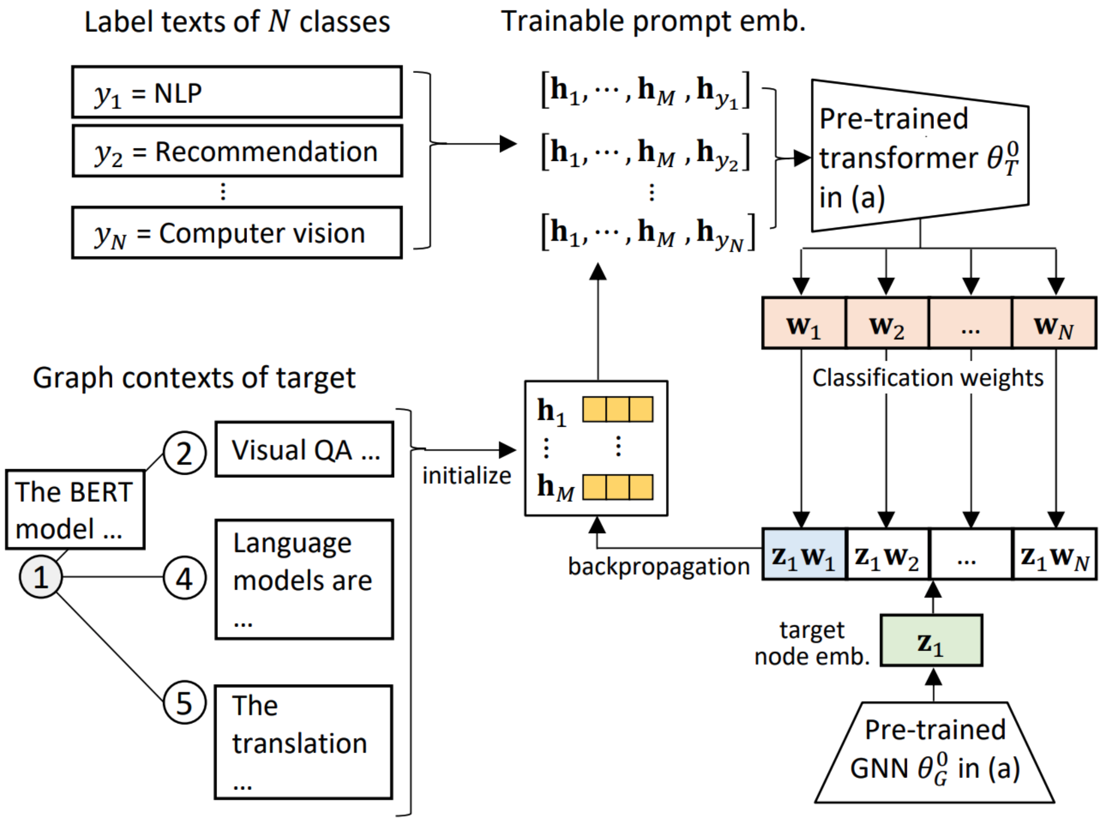
e.g., "A paper of " label text, e.g., "NLP"

- Class distribution is predicted as

$$p(y \mid \mathbf{z}_i) = \frac{\exp(\langle \mathbf{z}_i, \mathbf{w}_y \rangle)}{\sum_{y=1}^{N} \exp(\langle \mathbf{z}_i, \mathbf{w}_y \rangle)}$$

cosine similarity

Graph-grounded prompt tuning



- Discrete prompts are difficult to optimize.
- Resort to **prompt tuning**, substituting discrete prompts with **learnable continuous** vectors, while keeping the parameters of PLM **frozen**
- Instead of a sequence of **discrete tokens**, we use a sequence of **continuous embeddings**

$$\mathbf{w}_y = \phi_T([\mathbf{h}_1, \cdots, \mathbf{h}_M, \mathbf{h}_{\text{CLASS}}]; \theta_T^0)$$

- We initialize the prompt embeddings with **graph contexts**.
- A node v_i and its neighbor set $\{v_j | j \in \mathcal{N}_i\}$ are collectively called the *graph contexts* of v_i .

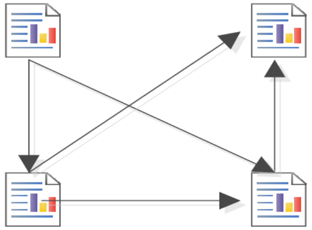
Outline

- ❑ Introduction
- ❑ Methodology
- ❑ **Experiment**
- ❑ Conclusion & Future work

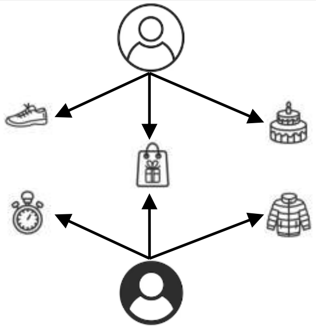
Datasets

Table 1: Statistics of datasets.

Dataset	Cora	Art	Industrial	M.I.
# Documents	25,120	1,615,902	1,260,053	905,453
# Links	182,280	4,898,218	3,101,670	2,692,734
# Avg. doc length	141.26	54.23	52.15	84.66
# Avg. node deg	7.26	3.03	2.46	2.97
# Classes	70	3,347	2,462	1,191



Cora is a collection of research papers



Art, Industrial and Music Instruments (M.I.) are 3 Amazon review datasets

Performance comparison with baselines

		Cora		Art		Industrial		M.I.	
		Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
End-to-end GNN	GCN	41.15±2.41	34.50±2.23	22.47±1.78	15.45±1.14	21.08±0.45	15.23±0.29	22.54±0.82	16.26±0.72
	SAGE_{sup}	41.42±2.90	35.14±2.14	22.60±0.56	16.01±0.28	20.74±0.91	15.31±0.37	22.14±0.80	16.69±0.62
	TextGCN	59.78±1.88	55.85±1.50	43.47±1.02	32.20±1.30	53.60±0.70	45.97±0.49	46.26±0.91	38.75±0.78
Pre-trained GNN	GPT-GNN	76.72±2.02	72.23±1.17	65.15±1.37	52.79±0.83	62.13±0.65	54.47±0.67	67.97±2.49	59.89±2.51
	DGI	78.42±1.39	74.58±1.24	65.41±0.86	53.57±0.75	52.29±0.66	45.26±0.51	68.06±0.73	60.64±0.61
	$\text{SAGE}_{\text{self}}$	77.59±1.71	73.47±1.53	76.13±0.94	65.25±0.31	71.87±0.61	65.09±0.47	77.70±0.48	70.87±0.59
Pre-trained Transformers	BERT	37.86±5.31	32.78±5.01	46.39±1.05	37.07± 0.68	54.00±0.20	47.57±0.50	50.14±0.68	42.96±1.02
	BERT*	27.22±1.22	23.34±1.11	45.31±0.96	36.28±0.71	49.60±0.27	43.36±0.27	40.19±0.74	33.69±0.72
	RoBERTa	62.10±2.77	57.21±2.51	72.95±1.75	62.25±1.33	76.35±0.65	70.49±0.59	70.67±0.87	63.50±1.11
	RoBERTa*	67.42±4.35	62.72±3.02	74.47±1.00	63.35±1.09	77.08±1.02	71.44±0.87	74.61±1.08	67.78±0.95
Prompt tuning	P-Tuning v2	71.00±2.03	66.76±1.95	76.86±0.59	66.89±1.14	79.65±0.38	74.33±0.37	72.08±0.51	65.44±0.63
	G2P2-p	79.16±1.23	74.99±1.35	79.59±0.31	68.26±0.43	80.86±0.40	74.44±0.29	81.26±0.36	74.82±0.45
	G2P2	**80.08***±1.33	**75.91***±1.39	**81.03***±0.43	**69.86***±0.67	**82.46***±0.29	**76.36***±0.25	**82.77***±0.32	**76.48***±0.52
	(improv.)	(+2.12%)	(+1.78%)	(+5.43%)	(+4.44%)	(+3.53%)	(+2.7%)	(+6.53%)	(+7.92%)

- G2P2 outperforms the best baseline by around 3–7%, showing the advantage of our contrastive pre-training and graph grounded prompt tuning

Outline

- ❑ Introduction
- ❑ Methodology
- ❑ Experiment
- ❑ **Conclusion & Future work**

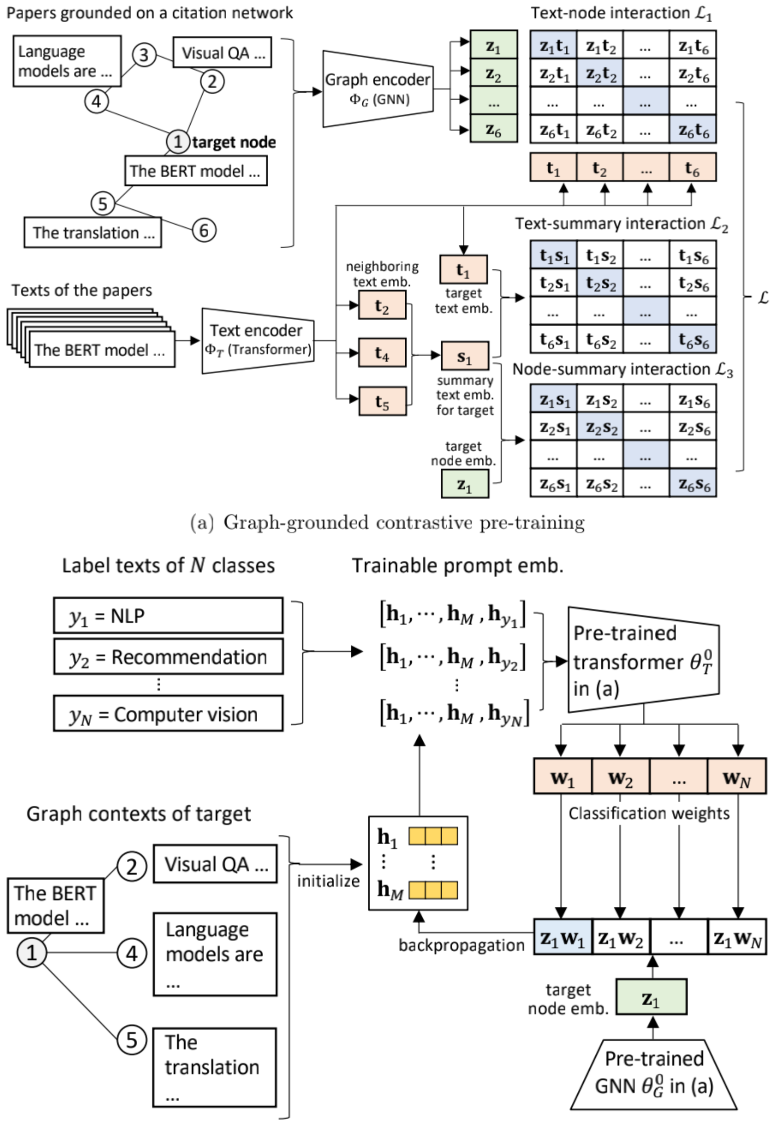
Conclusion

Key contributions

- Addressed the problem of **low-resource multi-task text classification**;
- Proposed G2P2, consisting of **three graph interaction-based** contrastive strategies in pre-training, and a **prompting** mechanism for the jointly pre-trained graph-text model in downstream classification.

Limitations

- The need of a **graph** to complement the texts
- Cannot do prompt tuning for zero-shot



(a) Graph-grounded contrastive pre-training

(b) Graph-grounded prompt tuning (few-shot classification)

THANK YOU FOR YOUR ATTENTION

Paper, code, data…
www.yfang.site

SMU
SINGAPORE MANAGEMENT UNIVERSITY

School of
Computing and
Information Systems