融合模态间相互影响的多模态异构网络表示学习与节点分类

Representation Learning with Mutual Influence of Modalities for Node Classification in Multi-Modal Heterogeneous Networks

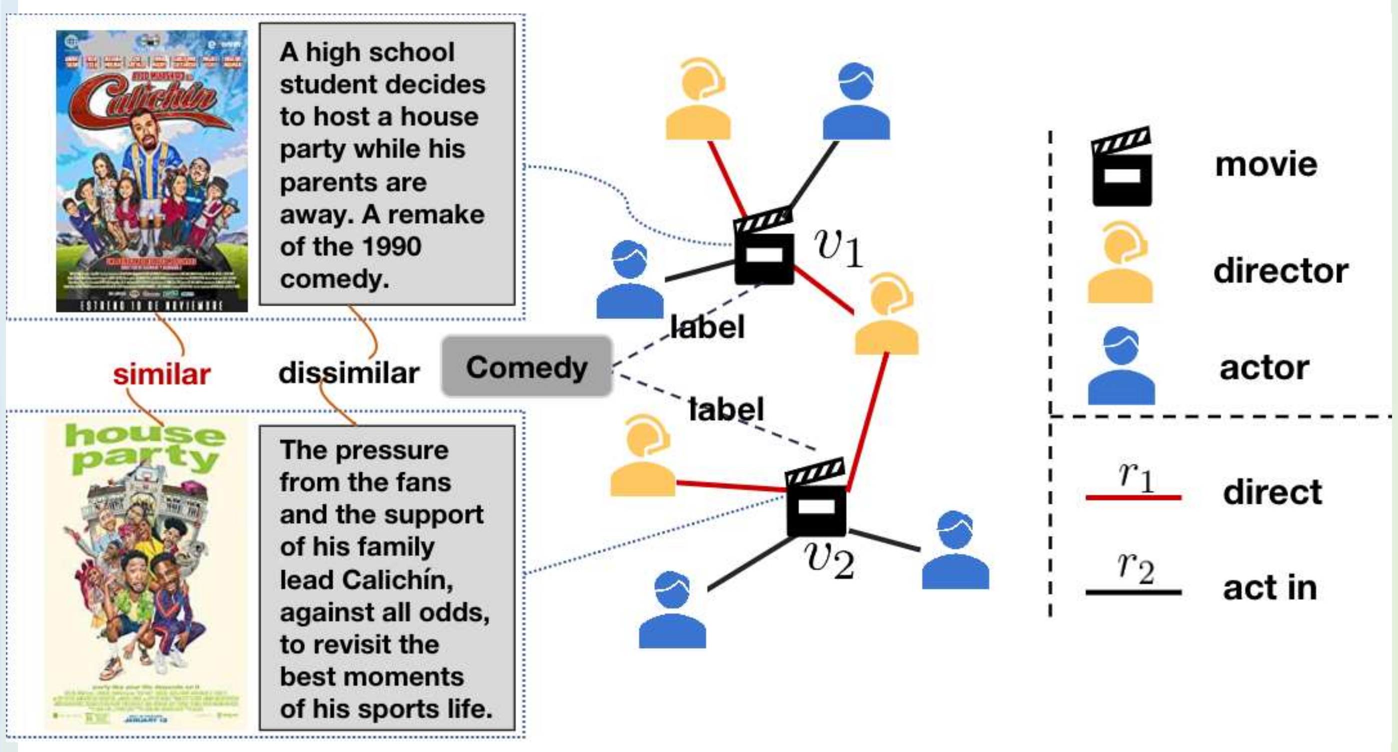
李佳璠，朱嘉奇*，常亮，李依霖，李妙妙，汪洋，杨翊，王宏安

To appear in *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2025)*, Main Track

联系人：朱嘉奇，zhujq@ios.ac.cn，13683257241

Introduction

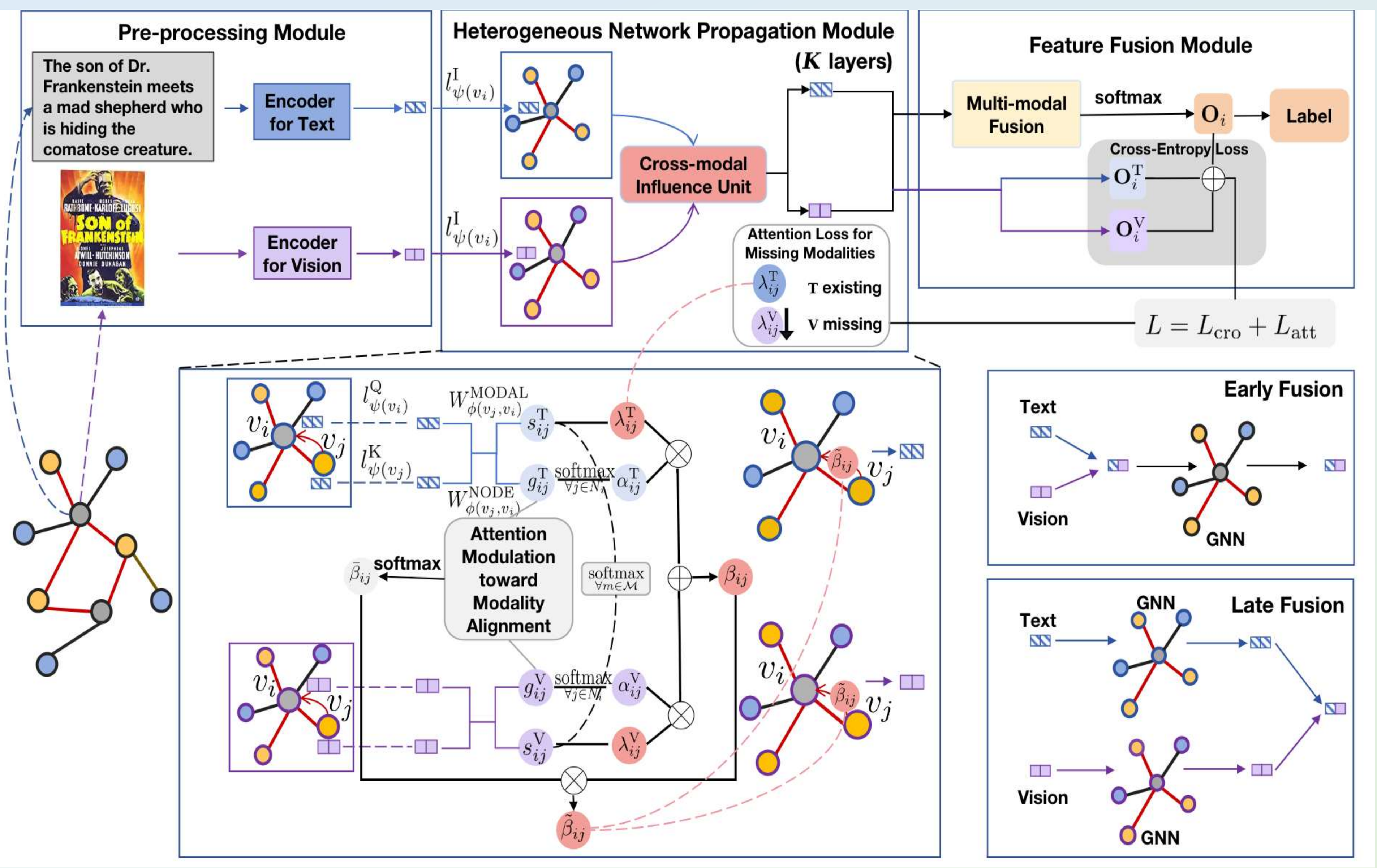
- **Task:** Node Classification in Multi-Modal Heterogeneous Networks (MMHNs)
- **Motivation:** the limitations of single-modality similarity and the necessity of multi-modal fusion in node representation learning



Core Idea

- **Core idea:** we propose a heterogeneous graph transformer with nested inter-modal attention and similarity-consistent modulation for modality alignment, in order to adaptively integrate multi-modal features during the propagation process in MMHNs.
- **Challenges:**
  - Consider the mutual influence of modalities during the information propagation process in MMHNs and learn it in an adaptive way
  - Choose the appropriate granularity to define and distinguish the cross-modal influence
  - Missing attributes for specific modalities
  - Mis-alignment among modalities

Model



Cross-modal Influence Unit

- Inter-node attention score

$$g_{ij}^{(k),m'} = l^{\mathrm{K}}_{\psi(v_j)}(\mathbf{h}_j^{(k-1),m'}) \cdot W^{\mathrm{NODE}}_{\phi(v_j,v_i)} \cdot l^{\mathrm{Q}}_{\psi(v_i)}(\mathbf{h}_i^{(k-1),m'})$$

$$\alpha_{ij}^{(k),m'} = \operatorname*{softmax}_{\forall j\in N_i}\left(g_{ij}^{(k),m'}\right) = \frac{\exp\left(g_{ij}^{(k),m'}\right)}{\sum_{j'\in N_i}\exp\left(g_{ij'}^{(k),m'}\right)}$$

- Inter-modal attention score

$$s_{ij}^{(k),m'} = l^{\mathrm{K}}_{\psi(v_j)}(\mathbf{h}_j^{(k-1),m'}) \cdot W^{\mathrm{MODAL}}_{\phi(v_j,v_i)} \cdot l^{\mathrm{Q}}_{\psi(v_i)}(\mathbf{h}_i^{(k-1),m'})$$

$$\lambda_{ij}^{(k),m'} = \operatorname*{softmax}_{\forall m'\in\mathcal{M}}\left(s_{ij}^{(k),m'}\right) = \frac{\exp\left(s_{ij}^{(k),m'}\right)}{\sum_{m''\in\mathcal{M}}\exp\left(s_{ij}^{(k),m''}\right)}$$

- Cross-modal attention weight

$$\beta_{ij}^{(k)} = \operatorname*{softmax}_{\forall j\in N_i}\left(\sum_{m'=1}^{\mathcal{M}}\left(\lambda_{ij}^{(k),m'}\alpha_{ij}^{(k),m'}\right)\right)$$

Attention Modulation toward Modality Alignment

- Consistency-weighted attention

$$\bar\beta_{ij}^{(k)} = \operatorname*{softmax}_{\forall j\in N_i}\left(\sum_{m_1,m_2\in\mathcal{M}}|g_{ij}^{(k),m_1} - g_{ij}^{(k),m_2}|\right)$$

- Final inter-node attention

$$\tilde\beta_{ij}^{(k)} = \operatorname*{softmax}_{\forall j\in N_i}(\beta_{ij}^{(k)}\cdot\bar\beta_{ij}^{(k)})$$

Training Objective

- Attention Loss for Missing Modalities

$$L_{\text{att}} = \frac{1}{K\cdot|\mathcal{M}|}\sum_{v_i\in\mathcal{V}}\sum_{v_j\in N_i}\sum_{1\le k\le K}\sum_{m'\notin f(\psi(v_j))}\lambda_{ij}^{(k),m'}$$

- Cross-entropy loss of individual modalities and the fused one

$$L_{\text{cro}} = \frac{1}{1+|\mathcal{M}|}\Big(\sum_{v_i\in\mathcal{V}_{\mathrm{L}}^*}\mathbf{y}_i^\top\cdot\log(\mathbf{O}_i) + \sum_{m\in\mathcal{M}}\sum_{v_i\in\mathcal{V}_{\mathrm{L}}^*}\mathbf{y}_i^\top\cdot\log(\mathbf{O}_i^m)\Big)$$

Experiments

Overall Results

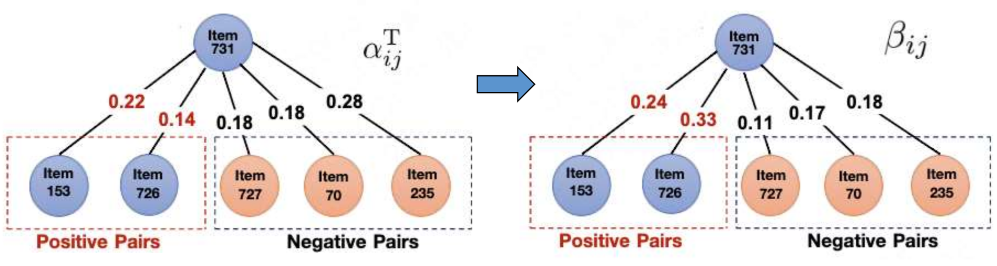
Datasets		DOUBAN		IMDB		AMAZON		AMAZON-1		AMAZON-2	
Metrics		Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
HAN	early	0.8707	0.8666	0.7267	0.7262	0.8594	0.8015	0.8532	0.6866	0.8542	0.6927
	late	0.8737	0.8699	0.7300	0.7286	0.8337	0.7737	0.8250	0.6157	0.8208	0.5936
SHGP	early	0.8319	0.8288	0.5488	0.5447	0.7483	0.6344	0.5989	0.3311	0.5678	0.3038
	late	0.8224	0.8256	0.5320	0.5180	0.7748	0.6920	0.5911	0.3205	0.5844	0.3255
SeHGNN	early	0.8667	0.8652	*0.7496*	*0.7478*	0.8726	0.8289	0.8554	0.7323	0.8522	0.7561
	late	0.8677	0.8624	0.7453	0.7438	0.8550	0.8122	0.8571	0.7638	0.8554	0.7660
HERO	early	0.8533	0.8493	0.6517	0.6102	0.8295	0.7699	0.8023	0.6755	0.8058	0.6546
	late	0.8283	0.8252	0.6936	0.6848	0.8207	0.7547	0.8136	0.6862	0.8012	0.6723
HGT	early	0.8508	0.8483	0.7407	0.7381	*0.8773*	*0.8302*	0.8883	0.7682	*0.8882*	*0.7807*
	late	0.8654	0.8629	0.7419	0.7407	0.8703	0.8212	*0.8931*	*0.7990*	0.8871	0.7799
HetGNN (early)		0.8366	0.8332	0.5068	0.4906	0.8328	0.7636	0.7012	0.5187	0.7129	0.4977
MHGAT (late)	max	0.8629	0.8574	0.7364	0.7249	0.8638	0.8084	0.8011	0.6729	0.8003	0.6486
	sum	0.8545	0.8468	0.7220	0.7127	0.7963	0.6734	0.7975	0.5929	0.7873	0.5433
IDKG		0.8462	0.8451	0.7410	0.7387	0.8752	0.8276	0.8504	0.5164	0.8604	0.5268
XGEA		*0.8765*	*0.8728*	0.7126	0.7047	0.8596	0.8001	0.8847	0.7226	0.8872	0.7301
HGNN-IMA		**0.8778**	**0.8758**	**0.7578**	**0.7560**	**0.8870**	**0.8427**	**0.8946**	**0.8233**	**0.8905**	**0.8182**

Ablation Study

Variants	DOUBAN	IMDB	AMAZON	AMAZON-1	AMAZON-2
-cross	*0.8661*	0.7344	0.8173	0.8125	0.8067
-adapt	0.8594	0.7360	*0.8397*	0.8044	0.7696
+inf	0.8614	0.7319	0.8332	0.8101	0.7798
-nei	0.8599	0.7256	0.8241	0.8082	0.7921
-align	0.8562	*0.7487*	0.8389	0.8056	0.8095
-L_{att}	0.8467	0.7436	0.8334	\	\
-L_{ind}	0.8602	0.7469	0.8392	*0.8218*	**0.8270**
Ours	**0.8758**	**0.7560**	**0.8427**	**0.8233**	*0.8182*

- Removing or changing the Cross-modal Influence Unit
- Removing attention modulation for modality alignment
- Removing some part of loss functions

Case study



Inter-node attention
✓ positive pairs ↑
✓ negative pairs ↓

Conclusion

- This paper delves into the intricate problem of node representation learning within multi-modal heterogeneous networks, characterized with complicated interactions of modalities and node/edge types.
- The innovative inter-modal attention acting on the modal-specific inter-node attention is proposed to enable adaptive modal fusion, based on the heterogeneous graph transformer framework.
- Another two critical factors in multi-modal data, modality alignment and modality missing, are also integrated into the model in a straightforward way to achieve significant improvements on node classification.

Applications

- HGNN-IMA learns superior node representations by effectively integrating multi-modal interactions, which significantly improves performance in node classification, link prediction, and recommendation tasks for platforms such as Amazon and Douban.
- The model effectively handles real-world networks with diverse modalities including images, numerical data, and audios, demonstrating strong adaptability to complex multi-modal scenarios.