# Multi-Graph Based Hierarchical Semantic Fusion For Cross-Modal Representation

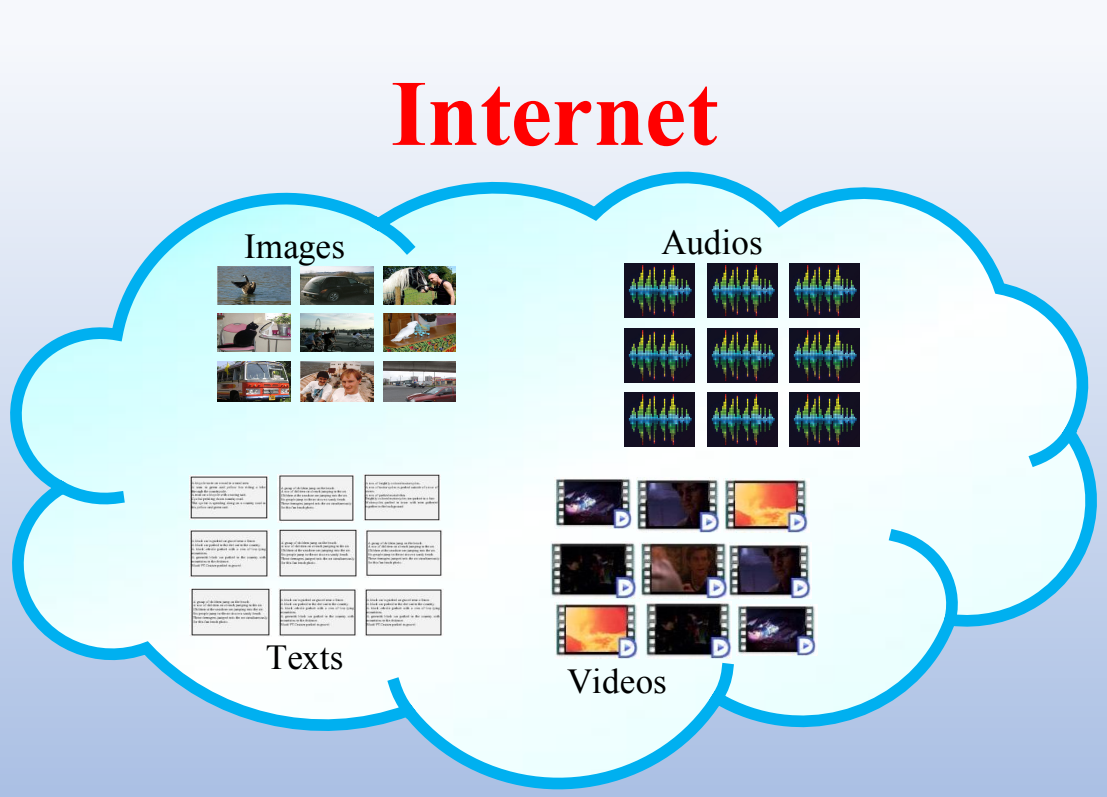
Lei Zhu, Chengyuan Zhang, Jiayu Song,
Liangchen Liu, Shichao Zhang, Yangding Li

College of Computer Science and Electronic Engineering, Hunan University
School of Computer Science and Engineering, Central South University
University of Melbourne
Hunan Normal University

# Content

- Background
- Motivation
- Contribution
- Problem Definition
- The Method
- Experiment
- Conclusion

# Background

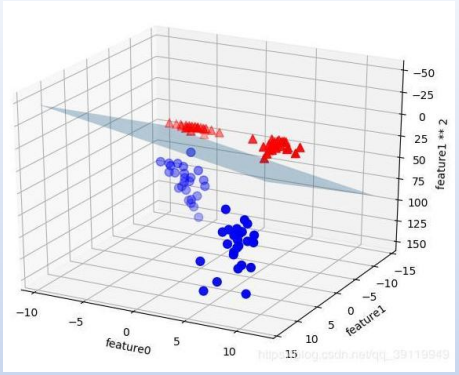




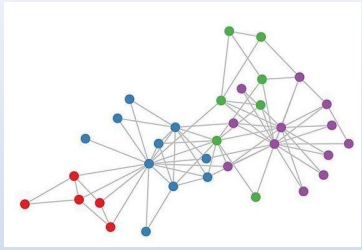
The main challenge of cross-modal retrieval task is how to efficiently realize semantic alignment and reduce the heterogeneity gap between different modalities.

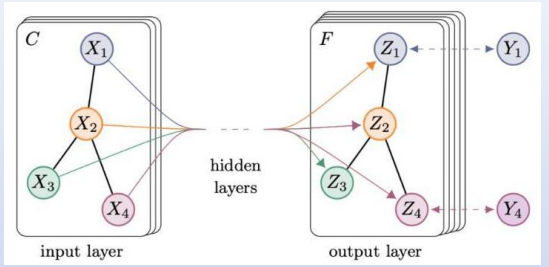
# Related Work

## Shallow Models

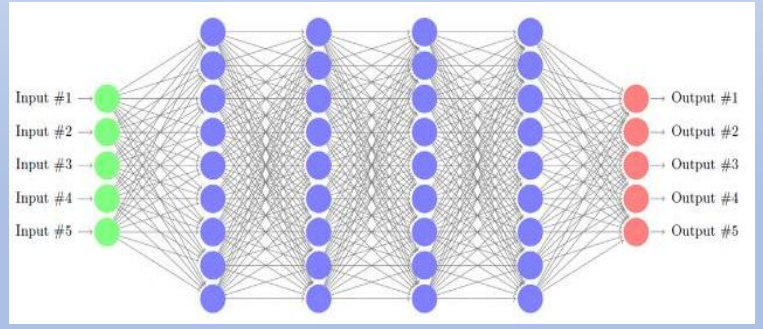



## Graph Models

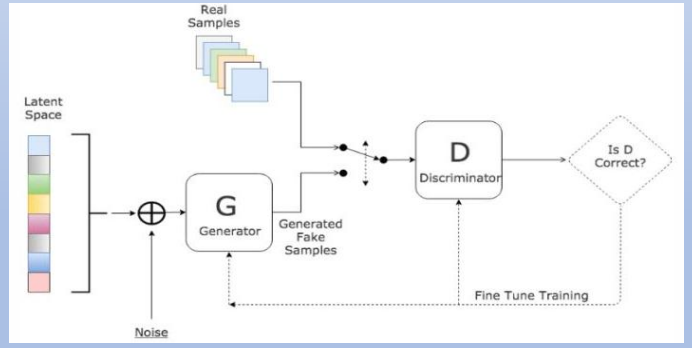





## Deep Models

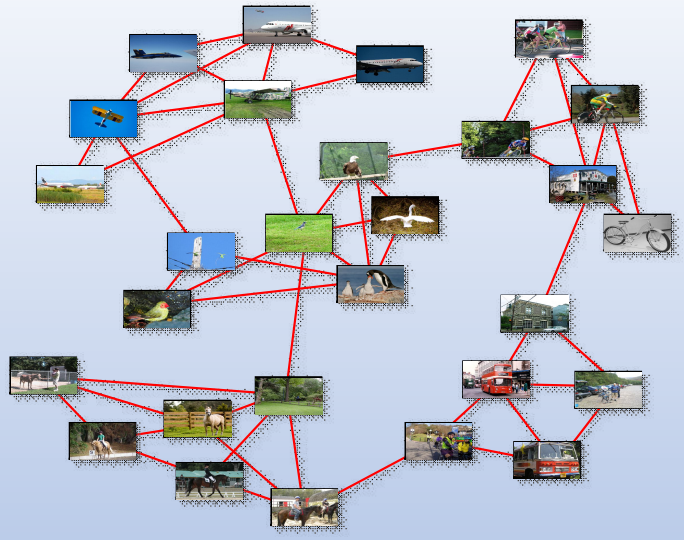



## Adversarial Models

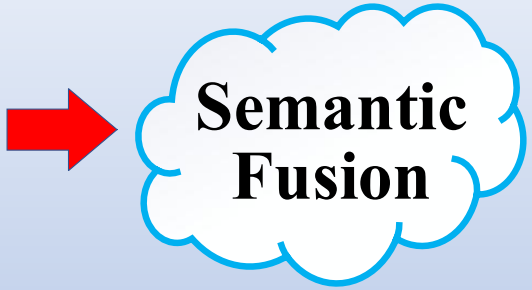


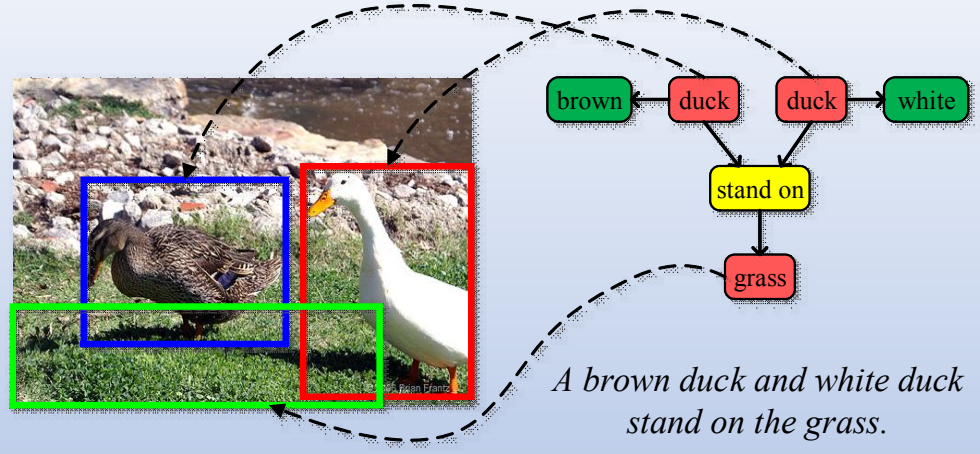

Although the existing researches have strikingly enhanced the performance of cross-modal retrieval, the semantic gap and heterogeneous gap have not been narrowed efficiently.

# Motivation



(1) Coarse-grained Semantic Knowledge




(2) Fine-grained Semantic Knowledge

The latest approaches focused on capturing high-level semantics, but ignored the hierarchical multi-grained knowledge discovery and semantic fusion. this paper investigates how to find multi-grained semantic knowledge from multi-modal objects, and realize the semantic fusion efficaciously to support cross-modal representation learning.

# Contributions

I. A novel end-to-end cross-modal representation method, named **MG-HSF**.

II. A scene graph based model to learn hierarchical semantic representation to capture the fine-grained knowledge; An intra-model semantic graph based representation model is developed to capture coarse-grained knowledge.

III. A semantic fusion model integrated with intra-modal and inter-modal adversarial learning to reduce the cross-modal heterogeneity.

# Problem Definition

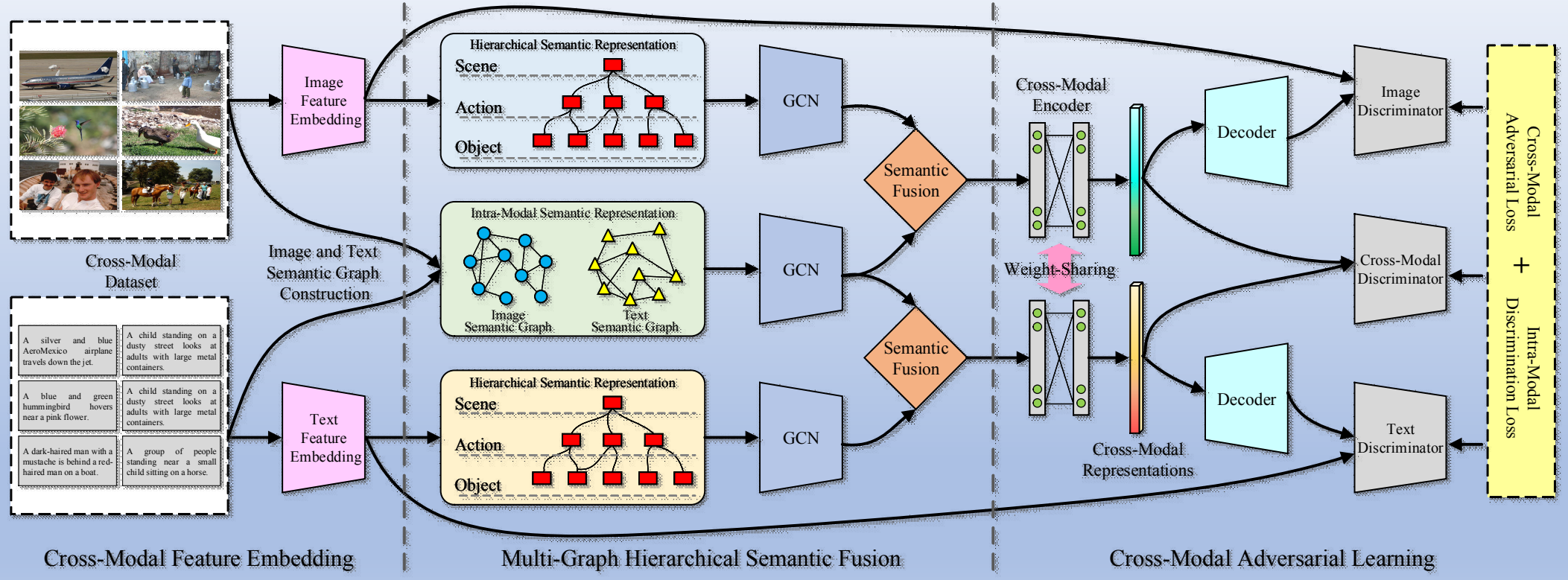
- Let $\mathcal{D}=\{\langle I_i,T_i,L_i\rangle\}_{i=1}^{N}$ be a multimedia dataset containing $N$ triplets, where $I_i$ and $T_i$ are the $i$-th image and text, $L_i$ is the label of them. Take image query$Q_I \in \mathcal{D}$ as an example, an image-to-text (Img2Txt) retrieval can be formulated as follows:

$$\mathcal{R}=\{T|Sim\left(\Phi_I(Q_I;\boldsymbol{\theta}_I),\Phi_T(T;\boldsymbol{\theta}_T)\right)\geq Sim\left(\Phi_I(Q_I;\boldsymbol{\theta}_I),\Phi_T(T';\boldsymbol{\theta}_T)\right),T\in\mathcal{D},T'\in\mathcal{D}\setminus\mathcal{R}\},$$

where $\mathcal{R}$ is the result set, $\boldsymbol{\theta}_I$ and $\boldsymbol{\theta}_T$ are the parameter vectors. $Sim\left(\right)$is the similarity function To realize the projections $\Phi_I(\cdot;\boldsymbol{\theta}_I)$ and $\Phi_T(\cdot;\boldsymbol{\theta}_T)$, we propose a novel cross-modal representation model, named MG-HSF, which is an integration of multi-graph knowledge fusion and adversarial learning technique.

# The Method

- **The Framework**

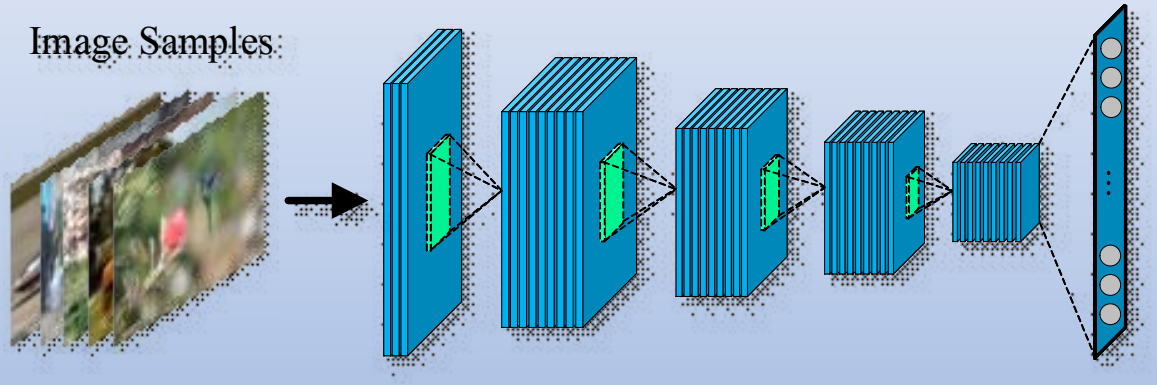



This model consists of three components: (1) cross-modal feature embedding; (2) multi-graph hierarchical semantic fusion; (3) cross-modal adversarial learning.

# The Method

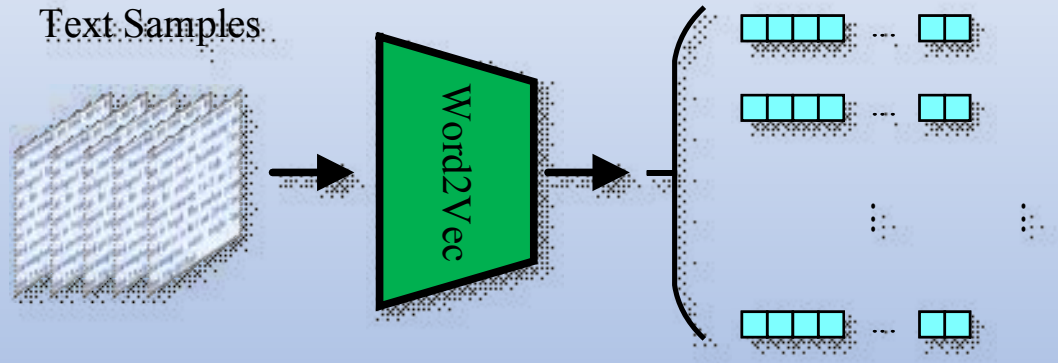
- **Cross-Model Feature Embedding**

Image Feature Embedding




For images, we employ deep CNN to learn visual features

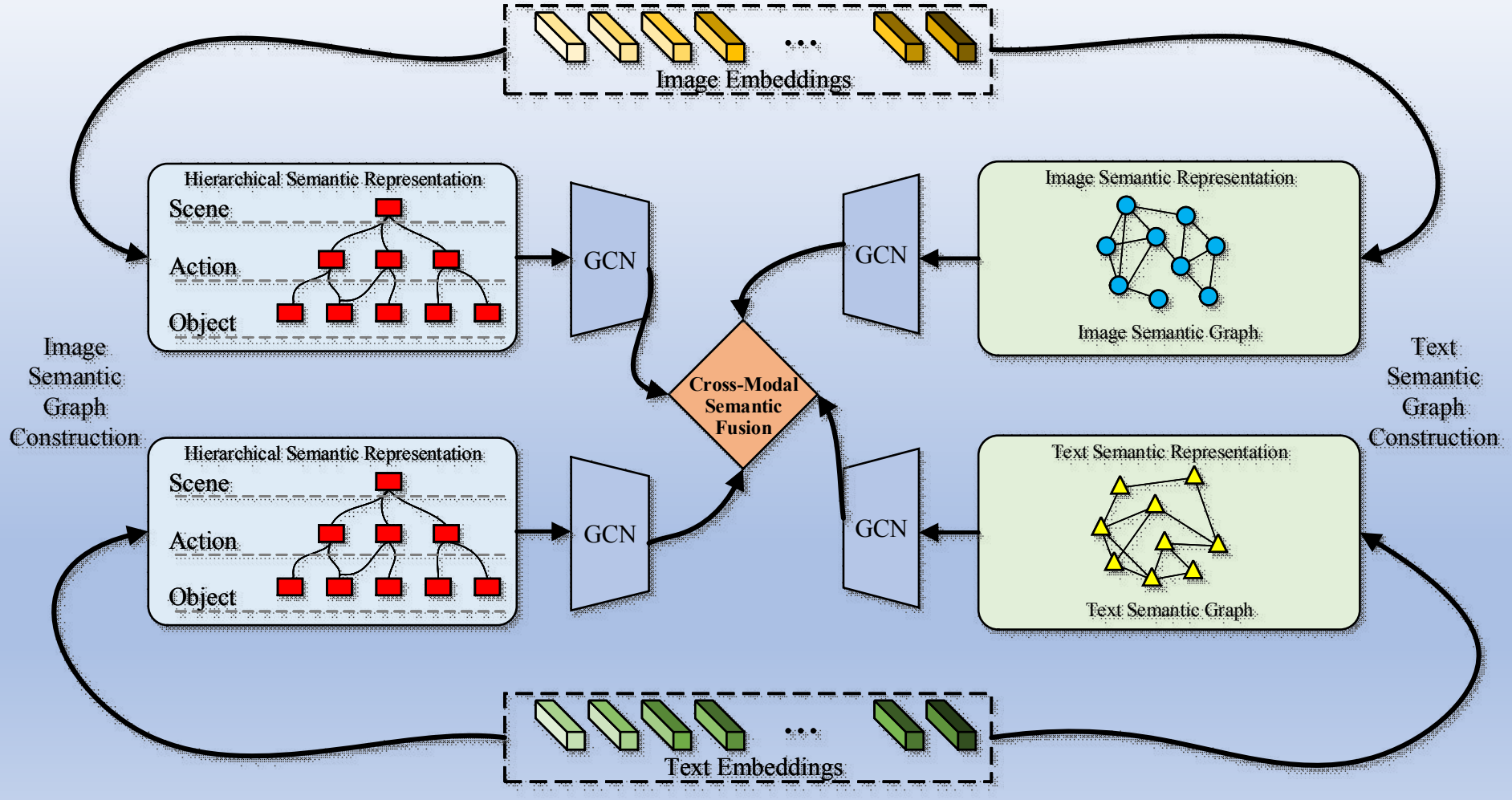
Text Feature Embedding




For texts, we embed each word by a word2vec model

# The Method

- **Multi-Graph Hierarchical Semantic Fusion**



# The Method

For image and text, we construct the hierarchical semantic graph with three levels: scene, action and object

**For text**: vertices can be learned via bi-LSTM from the word embeddings:

$$\boldsymbol{\varpi}_j = \frac{1}{2}(L_1(\boldsymbol{\omega}_j, \boldsymbol{\omega}_{j-1}; \boldsymbol{\theta}_{L1}) + L_2(\boldsymbol{\omega}_j, \boldsymbol{\omega}_{j+1}; \boldsymbol{\theta}_{L2})),$$

where $L_1(\cdot, \cdot; \boldsymbol{\theta}_{L1})$ and $L_2(\cdot, \cdot; \boldsymbol{\theta}_{L2})$ are LSTM models of two direction. a text attention mechanism is used to capture important information:

$$\boldsymbol{\varphi}^s = \sum_{j=1}^{K}\left(\frac{exp\left(\boldsymbol{A}_s \boldsymbol{\varpi}_j\right)\boldsymbol{\varpi}_j}{\sum_{k=1}^{K} exp\left(\boldsymbol{A}_s \boldsymbol{\varpi}_k\right)}\right)$$

# The Method

For image and text, we construct the hierarchical semantic graph with three levels: scene, action and object

**For image**: three matrices are adopted to encode scene vertex, action vertex and object vertex representations:

$$\boldsymbol{h}_i^s = \boldsymbol{M}_s C_i \qquad \boldsymbol{h}_i^a = \boldsymbol{M}_a \boldsymbol{C}_i \qquad \boldsymbol{h}_i^o = \boldsymbol{M}_o \boldsymbol{C}_i$$

Each element of $\boldsymbol{h}_i^a$ or $\boldsymbol{h}_i^o$ corresponds to an action vertex or an object vertex.

We propose to utilize the text semantic representation to train this model.

# The Method

For the intra-modal semantic graph construction, $k$NN algorithm is used on image set and text set to produce two graphs: $\mathcal{G}_I$ and $\mathcal{G}_T$,

$$\mathcal{G}_I = \langle \{\boldsymbol{C}_i\}, \mathcal{E}_I \rangle \qquad \mathcal{G}_T = \langle \{\boldsymbol{W}_j\}, \mathcal{E}_T \rangle$$

where $\mathcal{E}_I = \{E_{I:ij}\}$ and $\mathcal{E}_T = \{E_{T:ij}\}$ denote the edge sets that represent the connection of nearest neighbors for image or text. The weight of each edges is calculated by Euclidean distance:

$$E_{I:ij} = EucDst(\boldsymbol{C}_i, \boldsymbol{C}_j)$$

$$E_{T:ij} = EucDst(\boldsymbol{W}_i, \boldsymbol{W}_j)$$

where $\boldsymbol{C}_i \in \mathcal{N}_{knn}(\boldsymbol{C}_j)$ $\boldsymbol{C}_j \in \mathcal{N}_{knn}(\boldsymbol{C}_i)$ $\boldsymbol{W}_i \in \mathcal{N}_{knn}(\boldsymbol{W}_j)$ $\boldsymbol{W}_j \in \mathcal{N}_{knn}(\boldsymbol{W}_i)$

# The Method

## • Cross-Modal Adversarial Learning

A cross-modal encoder that is implemented by fully connected networks with weight-sharing is used to generate modality invariant representation: $\bar{\boldsymbol{\xi}}_i = Enc_I(\boldsymbol{\xi}_i; \boldsymbol{\theta}_{Ei})$ and $\bar{\boldsymbol{\zeta}}_i = Enc_T(\boldsymbol{\zeta}_i; \boldsymbol{\theta}_{Et})$. Besides, two decoders are integrated, and then re-construct the feature embeddings of image and text: $\hat{\boldsymbol{C}}_i$ and $\hat{\boldsymbol{W}}_i$. To realize intra-modal and inter-modal adversarial learning, three discriminators: $D_I(\cdot,\cdot;\boldsymbol{\theta}_{Di})$, $D_T(\cdot,\cdot;\boldsymbol{\theta}_{Dt})$ and cross-modal discriminator $D_{CI}(\cdot;\boldsymbol{\theta}_{Dci})$ and $D_{CT}(\cdot;\boldsymbol{\theta}_{Dct})$. The loss functions:

$$\mathcal{L}_{g1} = \mathbb{E}_{\boldsymbol{C}}[logD_I(\boldsymbol{C})] + \mathbb{E}_{\boldsymbol{C}}[log(1 - D_I(\hat{\boldsymbol{C}}))],$$

$$\mathcal{L}_{g2} = \mathbb{E}_{\boldsymbol{W}}[logD_I(\boldsymbol{W})] + \mathbb{E}_{\boldsymbol{W}}[log(1 - D_I(\hat{\boldsymbol{W}}))],$$

$$\mathcal{L}_{g3} = \mathbb{E}_{\boldsymbol{C},\boldsymbol{W}}[logD_{CI}(\bar{\boldsymbol{\xi}}) - logD_{CI}(\bar{\boldsymbol{\zeta}}) + logD_{CT}(\bar{\boldsymbol{\xi}}) - logD_{CT}(\bar{\boldsymbol{\zeta}})].$$

# Experiments

- **Data sets**
  - I. Wikipedia
  - II. NUS-WIDE
  - III. NUS-WIDE
- **Competitors**

  Shallow: CCA and SM

  Deep: Deep-SM and DSCMR

  Adversarial: ACMR and MHTN

Wikipedia

NUS-WIDE

Pascal Sentence

# Experiments

- The results (mAP) on Wikipedia

**Table 1**. The results (mAP@50 in %) with 6 competitors on Wikipedia dataset.

| Method | Img2Txt | Txt2Img | Average |
|---|---|---|---|
| CCA | 21.01 | 17.84 | 19.43 |
| SM | 23.34 | 28.51 | 25.93 |
| Deep-SM | 39.90 | 35.43 | 37.67 |
| DSCMR | 52.13 | 47.82 | 49.97 |
| ACMR | 47.74 | 43.42 | 45.58 |
| MHTN | 51.41 | 44.45 | 47.93 |
| MG-HSF | **52.85** | **53.21** | **53.03** |

# Experiments

- The results (mAP) on NUS-WIDE

**Table 2.** The results (mAP@50 in %) with 6 competitors on NUS-WIDE dataset.

| Method | Img2Txt | Txt2Img | Average |
|---|---|---|---|
| CCA | 38.17 | 36.80 | 37.49 |
| SM | 39.16 | 42.37 | 40.77 |
| Deep-SM | 57.80 | 62.55 | 60.18 |
| DSCMR | 61.10 | 61.54 | 61.32 |
| ACMR | 58.41 | 57.85 | 58.13 |
| MHTN | 52.03 | 53.41 | 52.72 |
| MG-HSF | **62.06** | **64.88** | **63.47** |

# Experiments

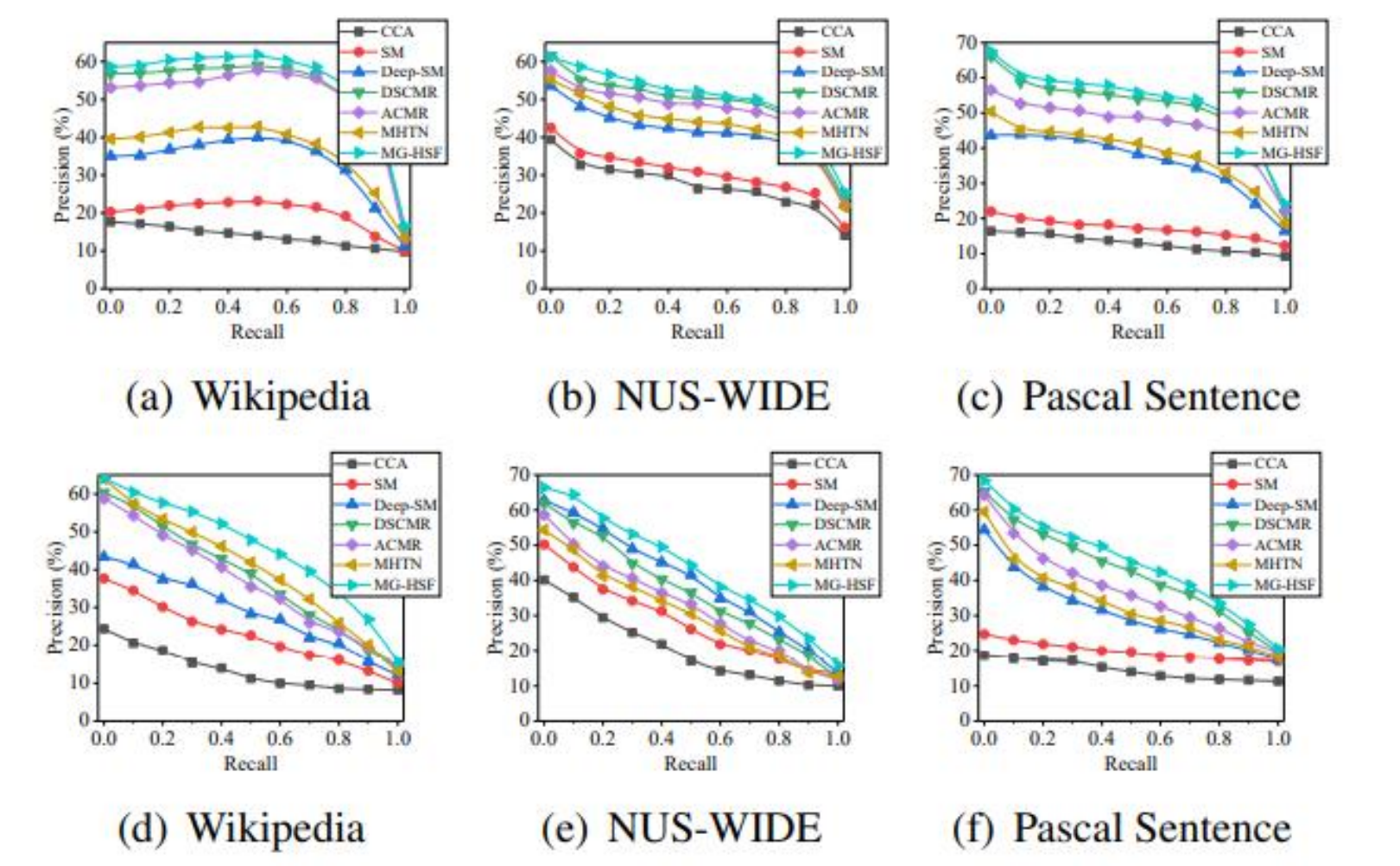
- The results (mAP) on Pascal Sentence

**Table 3**. The results (mAP@50 in %) with 6 competitors on Pascal Sentence dataset.

| Method | Img2Txt | Txt2Img | Average |
|---|---|---|---|
| CCA | 11.21 | 12.06 | 11.64 |
| SM | 18.74 | 21.12 | 20.14 |
| Deep-SM | 44.63 | 48.05 | 46.34 |
| DSCMR | 69.10 | 71.08 | 70.09 |
| ACMR | 60.48 | 59.75 | 60.12 |
| MHTN | 49.61 | 50.04 | 49.83 |
| MG-HSF | **69.62** | **71.55** | **70.59** |

# Experiments

- The PR-Curves on Wikipedia, NUS-WIDE and Pascal Sentence



# Conclusion

- This paper proposes a novel cross-modal representation method, called Multi-Graph based Hierarchical Semantic Fusion (MG-HSF).
- A multi-graph based hierarchical semantic fusion method is developed to learn fine-grained and coarse-grained knowledge.
- Intra-modal and inter-modal adversarial learning networks are utilized to reduce the cross-modal heterogeneity.

# Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62072166, 61836016, 61672177, 61702560).

# Thank You !