《Heterogeneous Graph Attention Network》

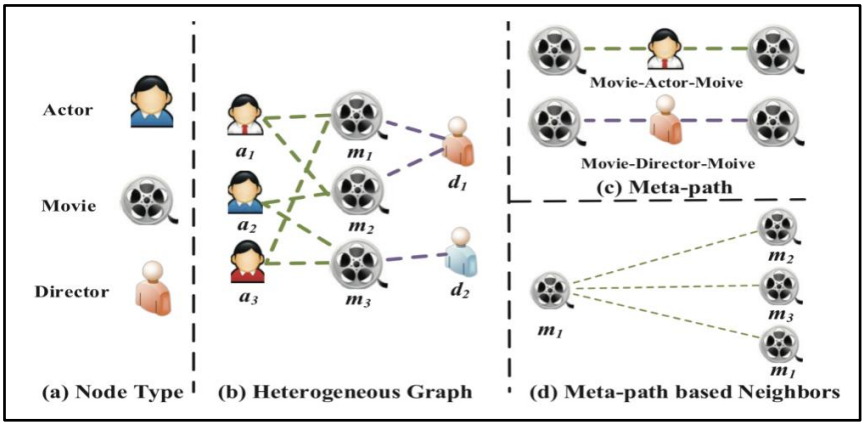
《异构图注意力网络》

摘要：

1. **本文的背景：**GCN 之后，图注意网络（GAT）开始崭露头角。作为一种新型的卷积型图神经网络，GAT 有效利用了注意力机制，并取得了更优异的结果。可是，注意力机制在异构图的图神经网络框架中并没有得到充分考虑，这便是本文的出发点，而本文的目的在于将注意力机制推广到异构图领域上。
2. **本文的贡献：**提出了一种包含**节点级（Node-level）注意力**和**语义级（Semantic-level）注意力**的层次注意力异构图神经网络（HAN），从而（第一次）将注意力机制从同构图扩展到了节点和边有不同类型的异构图上。
3. **主要创新点：**将注意力机制推广到了异质图上，通过利用节点级注意力（旨在学习节点与基于元路径的相邻节点之间的重要性）和语义级注意力（旨在学习不同元路径的重要性）在两个层面上聚合邻居特征来生成节点嵌入向量。
4. **实验结果：**在 3 个真实的异构图上的大量实验结果不仅表明了本文提出的 HAN 比现有的其他图神经网络模型有更优秀的性能，而且也证明了它对图分析具有潜在的、更好的可解释性。

异构图（Heterogeneous Graph）：

一个异构图，可记为$G=(V,E)$，由一个节点集V和一个边集E组成。而异构图的节点有多种类型，其节点类型映射函数$\phi: V \to A$；它的边也有多种类型，其边类型映射函数$\psi: E \to R$。其中，A和R表示预定义的节点类型和边类型的集合，且$|A|+|R|>2$。如下图（b）所示：



而**异质性，是异构图的内在属性**，即各种类型的节点（Node）和边（Edge）。例如，不同类型的节点具有不同的特征，它们的特征也因此可能落在不同的特征空间中。

元路径（Meta-path）：

在异构图中，两个对象可以通过不同的语义路径连接，称为**元路径（meta-path）**，如上图（c）中，Movie-Actor-Movie（MAM）和 Movie-Director-Movie (MDM) 都是不同的元路径。而不同的元路径有不同的语义，其中 MAM 表示电影的演员相同，MDM 表示电影的导演相同。于是，一个**元路径Φ**可定义为一条由$A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \dots \xrightarrow{R_l} A_{l+1}$组成的路径（也可简写为$A_1A_2 \dots A_{l+1}$）。$R = R_1 \circ R_2 \circ \dots \circ R_l$定义为节点$A_1$和$A_{l+1}$之间的复合关系，而∘表示在关系上的复合操作。

基于元路径的邻居：

给定一个元路径Φ，每个节点存在一组基于元路径的邻居，它们可以在异构图中揭示不同的结构信息和丰富的语义信息。给定一个节点i和一条元路径Φ，节点i基于元路径Φ的邻居$\mathcal{N}_i^{\Phi}$定义为通过元路径Φ和节点i相连的节点构成的集合，包括节点i自身。

如上图（d）所示，给定了元路径 Movie-Actor-Movie，电影m_1基于元路径的邻居包括m_1、m_2、m_3。类似的，给定元路径 Movie-Director-Movie，电影m_1基于元路径的邻居包括m_1、m_2。总之，这些邻居都是通过共同的演员或导演节点进行连接的。

符号语言定义：

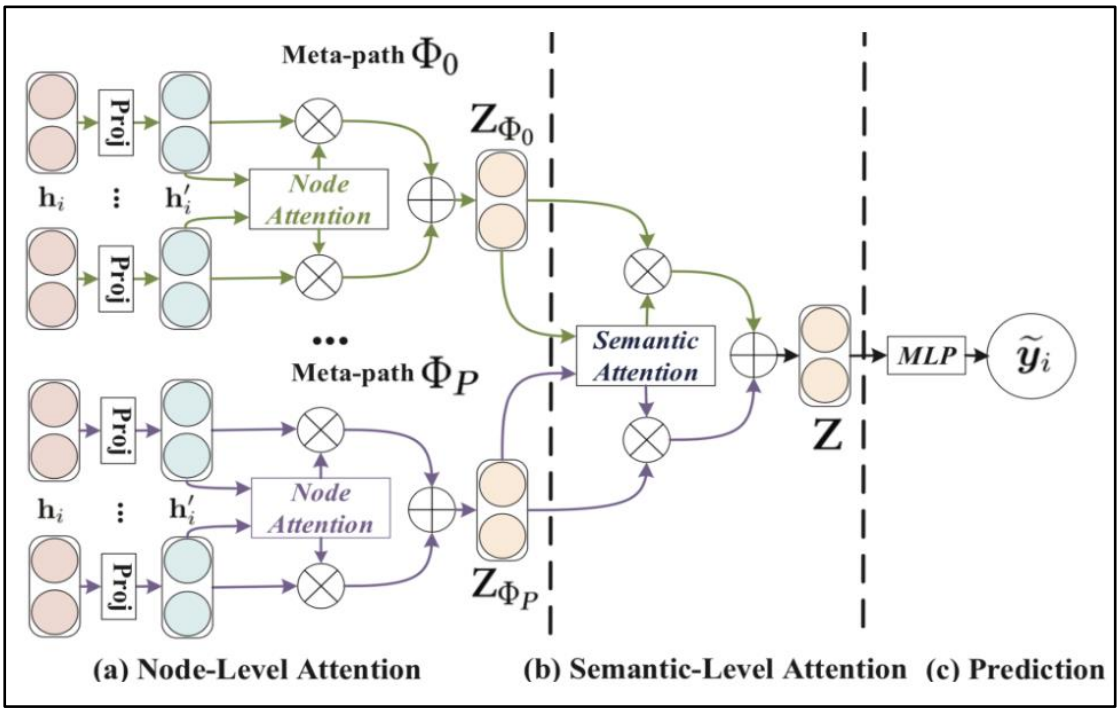
本文的符号语言表示如下：1）Φ：元路径；2）h：初始时的节点特征；3）M_{ϕ}：节点类型变换矩阵；4）h'：投影后的节点特征；5）e_{ij}^{Φ}：基于元路径Φ的节点对(i,j)的重要性；6）a_{Φ}：元路径Φ的节点级注意力向量；7）α_{ij}^{Φ}：基于元路径Φ的节点对(i,j)权重系数；8）$\mathcal{N}^{\Phi}$：基于元路径Φ的邻居集；9）Z_{Φ}：特定语义Φ的节点嵌入向量；10）q：语义级注意力向量；11）w_{Φ}：元路径Φ的重要性；12）β_{Φ}：元路径Φ的权重系数；13）Z：最终得到的节点嵌入向量。

上述内容如下表所示：

Table 1: Notations and Explanations.

Notation	Explanation
Φ	Meta-path
$\mathbf{h}$	Initial node feature
$\mathbf{M}_\phi$	Type-specific transformation matrix
$\mathbf{h}'$	Projected node feature
e_{ij}^{Φ}	Importance of meta-path based node pair (i,j)
$\mathbf{a}_\Phi$	Node-level attention vector for meta-path Φ
α_{ij}^{Φ}	Weight of meta-path based node pair (i,j)
$\mathcal{N}_i^{\Phi}$	Meta-path based neighbors
$\mathbf{Z}_\Phi$	Semantic-specific node embedding
$\mathbf{q}$	Semantic-level attention vector
w_Φ	Importance of meta-path Φ
β_Φ	Weight of meta-path Φ
$\mathbf{Z}$	The final embedding

HAN 模型的整体结构：



上图为本文所提出的 HAN 模型的整体框架。其中，（a）代表所有类型的节点被投影到统一的特征空间中，且基于元路径的节点对的权重可以通过节点级的注意力机制来进行学习；（b）代表联合学习每个元路径的权重，并通过语义级的注意力机制来融合不同语义的节点嵌入；而（c）表示计算 HAN 的损失函数值以及进行端到端的梯度下降优化。

总之，HAN 模型遵循一个层次注意力结构：节点级注意力→语义级注意力。首先，HAN 通过节点级注意力来学习基于元路径的节点邻居的权值，并对其进行聚合得到语义特定的节点嵌入。然后，HAN 通过语义级注意力来区分元路径的不同，从而得到特定任务的语义特定的节点嵌入的最优加权组合。

节点级注意力（Node-level Attention）：

节点级别的注意力就是将目标节点的不同邻居赋予不同的权重，表明邻居对自身的重要性程度。而在 HAN 模型中，对于每个节点，节点级注意力的目的是学习基于 meta-path 的邻居的重要性，并为它们分配不同的注意力值。具体操作过程如下：

（1）将不同类型的节点特征通过变换矩阵M_ϕ投影到统一的特征空间中：

$$h_i' = M_{\phi i} \cdot h_i$$

其中，h_i'和h_i分别是投影前后的节点i的特征向量。

（2）利用自注意力机制来学习各类节点之间的权重：

$$e_{ij}^{\Phi} = att_{node}(h_i', h_j'; \Phi) = \sigma(a_\Phi^T \cdot [h_i' || h_j'])$$

其中，att_{node}表示执行节点级注意力机制的深度神经网络，给定元路径Φ，att_{node}被所有基于元路径Φ的节点对(i,j)所共享。σ表示激活函数，a_Φ是元路径Φ的节点级注意力向量，$||$表示连接（$concat$）。此外，还需要注意e_{ij}^{Φ}是非对称的，即节点i对节点j的重要性和节点j对节点i的重要性可能有很大差异。

（3）通过 mask 操作将图的结构信息注入到模型中，即只计算节点$j \in \mathcal{N}_i^{\Phi}$的$e_{ij}^{\Phi}$，其中$\mathcal{N}_i^{\Phi}$表示节点$i$（包括自身）基于元路径$\Phi$的邻居。而在得到基于元路径的节点对之间的重要性后，我们再通过$softmax$函数将它们归一化得到权重系数α_{ij}^{Φ}：

$$\alpha_{ij}^{\Phi} = softmax(e_{ij}^{\Phi}) = \frac{\exp\left(\sigma(a_\Phi^T \cdot [h_i' || h_j'])\right)}{\sum_{k \in \mathcal{N}_i^{\Phi}} \exp\left(\sigma(a_\Phi^T \cdot [h_i' || h_k'])\right)}$$

总之，节点对(i,j)的权重系数α_{ij}^{Φ}取决于节点i,j各自的特征。并且，权重系数α_{ij}^{Φ}也是非对称的，即它们对彼此的贡献是不同的。

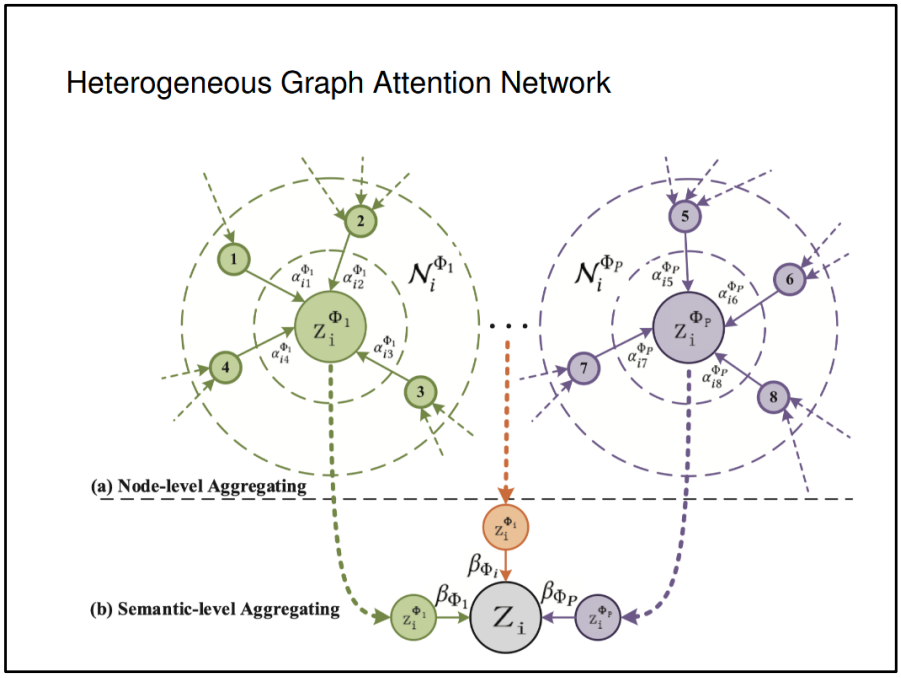
（4）节点i的基于元路径Φ的嵌入向量z_i^{Φ}可以通过聚合邻居的投影特征乘以对应的权重系数来得到，具体公式如下：

$$z_i^{\Phi} = \sigma\left(\sum_{j\in\mathcal{N}_i^{\Phi}} \alpha_{ij}^{\Phi} \cdot h_j'\right)$$

（5）由于异构图具有无标度特性，因此图数据的方差会相当大。而为了解决上述难题，我们将节点级注意力扩展为多头注意力，从而使训练过程更加稳定。即重复进行K次上述的节点级注意力操作，并将各头的结果进行拼接：

$$z_i^{\Phi} = ||_{k=1}^{K}\sigma\left(\sum_{j\in\mathcal{N}_i^{\Phi}} \alpha_{ij}^{\Phi} \cdot h_j'\right)$$

此外，为了更好地理解节点级注意力的聚合过程，在下图（a）中作者也给出了更加简要的解释。



如图（a）所示，每个节点都由其邻居聚合而成。而由于注意力权重α_{ij}^{Φ}是针对某一元路径Φ生成的，因此它也是语义特定的，即能够捕获某一种特定的语义信息。所以，给定元路径集合$\{\Phi_1, \Phi_2, \ldots, \Phi_P\}$，并将节点特征输入到节点级注意力网络后，我们就可以得到P组特定语义的节点嵌入，记为$\{Z_{\Phi_1}, Z_{\Phi_2}, \ldots, Z_{\Phi_P}\}$。

语义级注意力（Semantic-level Attention）：

通常来说，异构图会涉及到不同的有意义和复杂的语义信息，而这些信息一般由元路径来反映。对于某个具体的任务，不同元路径表达的语义不同，因此对目标任务的贡献度也会不同。给定元路径集合$\{\Phi_1, \Phi_2, \dots, \Phi_P\}$以及通过节点级注意力学习到的不同语义下的节点表示$\{Z_{\Phi_1}, Z_{\Phi_2}, \dots, Z_{\Phi_P}\}$，进一步，我们可以利用语义级注意力来学习不同语义的重要性并融合多个语义下的节点表示。因此，语义级注意力的形式化描述如下：

$$(\beta_{\Phi_1}, \beta_{\Phi_2}, \dots, \beta_{\Phi_P}) = att_{sem}(Z_{\Phi_1}, Z_{\Phi_2}, \dots, Z_{\Phi_P})$$

其中，$(\beta_{\Phi_1}, \beta_{\Phi_2}, \dots, \beta_{\Phi_P})$是各个元路径的注意力权重，$att_{sem}$表示用于语义级注意力的神经网络。简单来说，**语义级自注意力机制**就是通过利用单层神经网络和节点级嵌入向量来学习各个语义（元路径）的重要性并通过$softmax$函数来进行归一化。具体操作过程如下：

（1）为了学到每一个 meta-path 的重要性，首先使用一个线性变换+$tanh$激活（文中使用一层MLP）来转换节点级嵌入向量。接着，转换后的嵌入向量与一个语义级别的注意力向量q相乘。最后，平均所有特定语义节点嵌入的重要性。因此，每个元路径的重要性w_{Φ_p}表示如下：

$$w_{\Phi_p} = \frac{1}{|V|}\sum_{i \in V} q^T \cdot \tanh\left(W \cdot z_i^{\Phi_p} + b\right)$$

其中，W为权重矩阵，b为偏置向量，q为语义级注意力向量。

（2）在获得每个元路径的重要性后，我们再通过$softmax$函数对其进行归一化得到权重系数：

$$\beta_{\Phi_p} = \frac{\exp\left(w_{\Phi_p}\right)}{\sum_{p=1}^{P} \exp\left(w_{\Phi_p}\right)}$$

（3）利用学习到的权重作为系数，我们就可以将这些语义特定的嵌入进行加权融合，得到最终的节点嵌入表示Z：

$$Z = \sum_{p=1}^{P} \beta_{\Phi_p} \cdot Z_{\Phi_p}$$

而为了更好地理解语义层面的聚合过程，上图（b）中也给出了更加简要的解

释，即最终的嵌入表示由所有特定语义的嵌入聚合而成。然后，为了将最终的嵌入应用到具体的任务中，我们还需要设计不同的损失函数。而对于半监督节点分类任务，我们可以最小化所有标记节点在真实节点和预测节点之间的交叉熵，具体公式如下：

$$L = -\sum_{l \in y_L} Y^l \ln\left(C \cdot Z^l\right)$$

其中，C是分类器的参数，y_L是有标签节点的索引集合，Y^l和Z^l是有标签节点的标签向量和嵌入向量。最后，我们再通过反向传播来不断优化上述所有步骤中的模型参数。

HAN 模型的伪代码流程：

Algorithm 1: The overall process of HAN.

Input : The heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$,
The node feature $\{\mathbf{h}_i, \forall i \in \mathcal{V}\}$,
The meta-path set $\{\Phi_0, \Phi_1, \ldots, \Phi_P\}$.
The number of attention head K,

Output : The final embedding $\mathbf{Z}$,
The node-level attention weight α ,
The semantic-level attention weight β .

for $\Phi_i \in \{\Phi_0, \Phi_1, \ldots, \Phi_P\}$ **do**
  for $k = 1 \ldots K$ **do**
    Type-specific transformation $\mathbf{h}'_i \leftarrow \mathbf{M}_{\phi_i} \cdot \mathbf{h}_i$;
    for $i \in \mathcal{V}$ **do**
      Find the meta-path based neighbors $\mathcal{N}_i^{\Phi}$;
      for $j \in \mathcal{N}_i^{\Phi}$ **do**
        Calculate the weight coefficient α_{ij}^{Φ} ;
      end
      Calculate the semantic-specific node embedding $\mathbf{z}_i^{\Phi} \leftarrow \sigma\left(\sum_{j \in \mathcal{N}_i^{\Phi}} \alpha_{ij}^{\Phi} \cdot \mathbf{h}'_j\right)$;
    end
    Concatenate the learned embeddings from all attention head $\mathbf{z}_i^{\Phi} \leftarrow \Big\Vert_{k=1}^{K} \sigma\left(\sum_{j \in \mathcal{N}_i^{\Phi}} \alpha_{ij}^{\Phi} \cdot \mathbf{h}'_j\right)$;
  end
  Calculate the weight of meta-path β_{Φ_i} ;
  Fuse the semantic-specific embedding $\mathbf{Z} \leftarrow \sum_{i=1}^{P} \beta_{\Phi_i} \cdot \mathbf{Z}_{\Phi_i}$;
end
Calculate Cross-Entropy $L = -\sum_{l \in y_L} \mathrm{Y}_l \ln(\mathrm{C} \cdot \mathrm{Z}_l)$;
Back propagation and update parameters in HAN;
return Z, α, β.

实验及结果：

论文在三个数据集上做了大量充分的实验（包括节点分类，节点聚类，可视化）来验证 HAN 模型的有效性。同时，为了验证节点级和语义级注意力机制的作用，作者还分别去除了节点级或语义级注意力机制来进行相关的消融实验。

数据集：

Table 2: Statistics of the datasets.

Dataset	Relations(A-B)	Number of A	Number of B	Number of A-B	Feature	Training	Validation	Test	Meta-paths
DBLP	Paper-Author	14328	4057	19645	334	800	400	2857	*APA*
	Paper-Conf	14328	20	14328					*APCPA*
	Paper-Term	14327	8789	88420					*APTPA*
IMDB	Movie-Actor	4780	5841	14340	1232	300	300	2687	*MAM*
	Movie-Director	4780	2269	4780					*MDM*
ACM	Paper-Author	3025	5835	9744	1830	600	300	2125	*PAP*
	Paper-Subject	3025	56	3025					*PSP*

实验中所用的三个数据集的统计信息如上表所示。

实验设置：

对于所提出的 HAN，实验中对它进行随机初始化参数，采用 Adam 优化器进行优化，并将学习率设置为 0.005，正则化参数设置为 0.001，语义级注意力向量 q 的维度设置为 128，注意力头 K 的数量设置为 8，注意力参数的丢弃率设置为 0.6。此外，如果验证损失在连续 100 个 epoch 内都没有明显减少，就停止训练。而对于 GCN 和 GAT，实验中使用验证集来优化它们的参数。

对于半监督图神经网络，包括 GCN，GAT 和 HAN，实验时拆分了完全相同的训练集，验证集和测试集，以确保公平性。对于基于随机游走的方法，包括 DeepWalk，ESim，metapath2vec 和 HERec，实验中将窗口大小统一设置为 5，游走长度设置为 100，每个节点的漫游数设置为 40，负样本的数量设置为 5。并且，为了公平比较，实验中将上述所有算法的最终嵌入维度都设置为 64。

节点分类任务实验结果：

使用$k=5$的KNN分类器来执行节点分类。

Table 3: Qantitative results (%) on the node classification task.

Datasets	Metrics	Training	DeepWalk	ESim	metapath2vec	HERec	GCN	GAT	HAN_{nd}	HAN_{sem}	HAN
ACM	Macro-F1	20%	77.25	77.32	65.09	66.17	86.81	86.23	88.15	89.04	**89.40**
		40%	80.47	80.12	69.93	70.89	87.68	87.04	88.41	89.41	**89.79**
		60%	82.55	82.44	71.47	72.38	88.10	87.56	87.91	**90.00**	89.51
		80%	84.17	83.00	73.81	73.92	88.29	87.33	88.48	90.17	**90.63**
	Micro-F1	20%	76.92	76.89	65.00	66.03	86.77	86.01	87.99	88.85	**89.22**
		40%	79.99	79.70	69.75	70.73	87.64	86.79	88.31	89.27	**89.64**
		60%	82.11	82.02	71.29	72.24	88.12	87.40	87.68	**89.85**	89.33
		80%	83.88	82.89	73.69	73.84	88.35	87.11	88.26	89.95	**90.54**
DBLP	Macro-F1	20%	77.43	91.64	90.16	91.68	90.79	90.97	91.17	92.03	**92.24**
		40%	81.02	92.04	90.82	92.16	91.48	91.20	91.46	92.08	**92.40**
		60%	83.67	92.44	91.32	92.80	91.89	90.80	91.78	92.38	**92.80**
		80%	84.81	92.53	91.89	92.34	92.38	91.73	91.80	92.53	**93.08**
	Micro-F1	20%	79.37	92.73	91.53	92.69	91.71	91.96	92.05	92.99	**93.11**
		40%	82.73	93.07	92.03	93.18	92.31	92.16	92.38	93.00	**93.30**
		60%	85.27	93.39	92.48	93.70	92.62	91.84	92.69	93.31	**93.70**
		80%	86.26	93.44	92.80	93.27	93.09	92.55	92.69	93.29	**93.99**
IMDB	Macro-F1	20%	40.72	32.10	41.16	41.65	45.73	49.44	49.78	**50.87**	50.00
		40%	45.19	31.94	44.22	43.86	48.01	50.64	52.11	50.85	**52.71**
		60%	48.13	31.68	45.11	46.27	49.15	51.90	51.73	52.09	**54.24**
		80%	50.35	32.06	45.15	47.64	51.81	52.99	52.66	51.60	**54.38**
	Micro-F1	20%	46.38	35.28	45.65	45.81	49.78	55.28	54.17	55.01	**55.73**
		40%	49.99	35.47	48.24	47.59	51.71	55.91	56.39	55.15	**57.97**
		60%	52.21	35.64	49.09	49.88	52.29	56.44	56.09	56.66	**58.32**
		80%	54.33	35.59	48.81	50.99	54.61	56.97	56.38	56.49	**58.51**

如上表所示，可以发现本文所提出的 HAN 模型在所有数据集上都实现了最佳性能。而消融实验的结果表明，在异构图分析中捕捉节点和元路径的重要性都非常重要。

节点聚类任务实验结果：

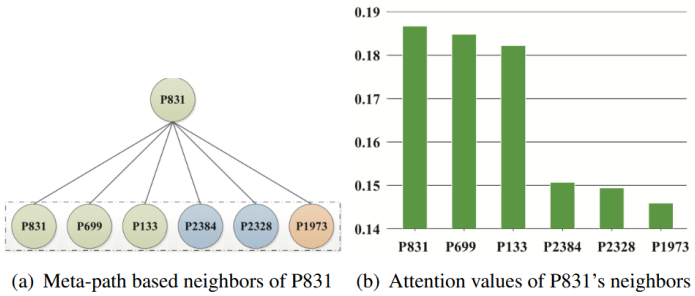
使用 k-means 聚类算法进行聚类。

Table 4: Qantitative results (%) on the node clustering task.

Datasets	Metrics	DeepWalk	ESim	metapath2vec	HERec	GCN	GAT	HAN_{nd}	HAN_{sem}	HAN
ACM	NMI	41.61	39.14	21.22	40.70	51.40	57.29	60.99	61.05	**61.56**
	ARI	35.10	34.32	21.00	37.13	53.01	60.43	61.48	59.45	**64.39**
DBLP	NMI	76.53	66.32	74.30	76.73	75.01	71.50	75.30	77.31	**79.12**
	ARI	81.35	68.31	78.50	80.98	80.49	77.26	81.46	83.46	**84.76**
IMDB	NMI	1.45	0.55	1.20	1.20	5.45	8.45	9.16	10.31	**10.87**
	ARI	2.15	0.10	1.70	1.65	4.40	7.46	7.98	9.51	**10.01**

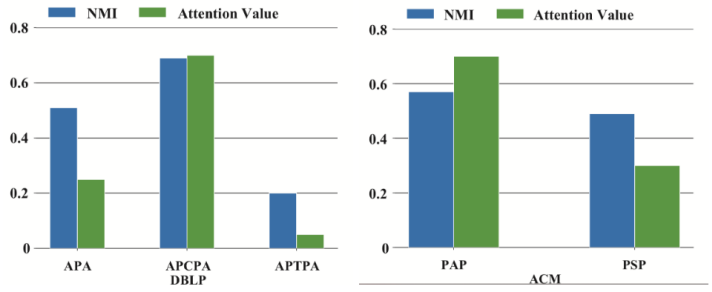
如上表所示，可以发现 HAN 的性能始终优于所有基线。同时，在去除节点级和语义级注意力后，模型的效果有不同程度的降低。这也进一步验证了节点级注意力和语义级注意力的有效性。

层次注意力机制分析：



(a) Meta-path based neighbors of P831 (b) Attention values of P831's neighbors

Figure 4: Meta-path based neighbors of node P831 and corresponding attention values (Different colors mean different classes, e.g., *green* means Data Mining, *blue* means Database, *orange* means Wireless Communication).

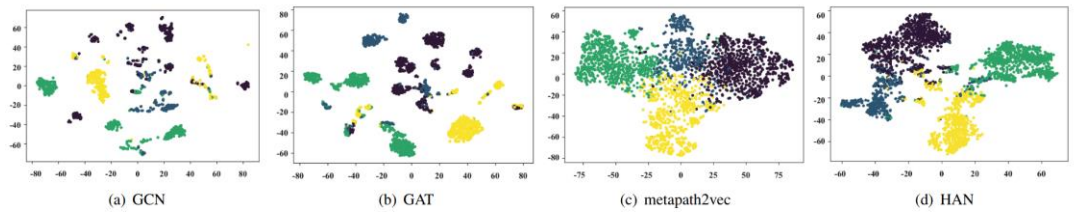


(a) NMI values on DBLP (b) NMI values on ACM

Figure 5: Performance of single meta-path and corresponding attention value.

节点级和语义级注意力机制分析分别如上图 4 和图 5 所示，可以看到在节点级注意力层面上，主要赋予了同类型的邻居更高的权重。而在语义级注意力层面上，对于较为重要的元路径，即该条元路径在聚类任务上会具有较大的 NMI 值，HAN 会赋予它们相应更大的权重。因此，可以表明 HAN 能够自动地选取较为重要的节点邻居及元路径来进行邻居聚合。

可视化分析：

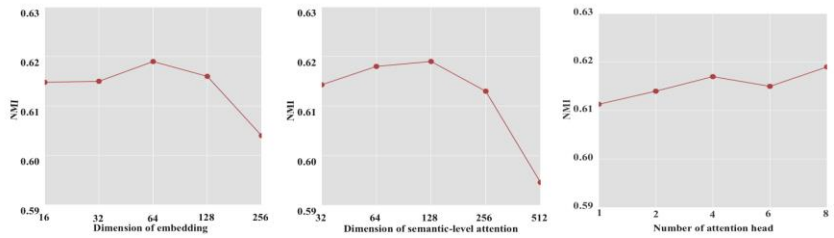


(a) GCN (b) GAT (c) metapath2vec (d) HAN

Figure 6: Visualization embedding on DBLP. Each point indicates one author and its color indicates the research area.

如上图所示，可以看出 HAN 的可视化效果最好。在多条元路径的引导下，HAN 学习到的嵌入具有较高的类内相似性，并且能以明显的边界将不同类别的数据分开。

参数敏感性分析：



(a) Dimension of the final embedding Z (b) Dimension of the semantic-level attention vector q (c) Number of attention head K.

Figure 7: Parameter sensitivity of HAN w.r.t. Dimension of the final embedding Z, Dimension of the semantic-level attention vector q and Number of attention head K.

如图所示，可以看到聚类指标 NMI 随着参数Z（最终嵌入维度）的增加先增大后减小；对于参数q（语义级注意力向量）的增加也是先增大而后减小，说明参数Z和q都需要一个合适的大小才能使 HAN 的模型效果最好。而对于多头参数K，K 越大，通常会提高 HAN 的性能，但性能仅略有提高。此外，实验还发现多头注意力可以使整个训练过程变得更加稳定。

总结：

本文是第一篇基于注意力机制的异构图神经网络研究，并且提出了一种新的基于注意力机制的异构图神经网络 Heterogeneous Graph Attention Network（HAN），它可以有效地应用于异构图分析。而大量的实验结果也表明 HAN 模型与现有模型相比更具优越性，且相对于 meta-path 节点对的数目具有线性复杂度，可以应用于大规模异构图。此外，在本文的最后，通过分析这种分层的注意力机制，还证明了 HAN 对异构图分析具有潜在地良好的可解释能力。

对本文的感悟：

本文将注意力机制从同构图成功推广到了异构图，也再次有力证明了注意力机制在深度学习领域上的有效性，不仅是对计算机视觉领域，对图结构分析领域也充满着无限的可能性。