20 图神经网络

Mila 研究院唐建：基于图神经网络的关系推理

整理：智源社区 熊宇轩

在 NeurIPS 2019 上，图灵奖得主、深度学习「三驾马车」之一的 Yoshua Bengio 指出深度学习需要进行从感知到认知的革命，指引研究者们尝试通过将 System 1 的快速感知系统与 System 2 的深度推理系统相结合实现更加强大的人工智能系统。

在本届智源大会上，来自 Yoshua Bengio 领导的 Mila 研究院的知名华人学者唐建为大家带来了名为「将 System 1 和 System 2 结合，用于关系推理」(Towards Integrating System I and System II for Relational Reasoning) 的精彩报告，以其研究小组近期在半监督节点分类、知识图谱推理等任务上的研究进展为例，详细介绍了将感知系统与认知系统相结合的方法。下面是演讲的主要内容:

一、System 1 VS System 2 推理

首先，我们来回顾一下 System 1 和 System 2 推理的定义。

事实上，现在大多数的深度学习系统所做的工作都属于 System 1 推理，即「感知系统」。识别图像中的物体就是一种感知任务。在感知任务中，人通常是相对无意识的，这是一个快速思考的过程。但是生活中有很多任务是非常复杂的，仅仅凭借感知系统无法很好的解决这些问题。例如，对于 VQA 任务而言，给定一幅图片，我们需要基于该图的信息回答一些问题。在上图中，右侧给出了一个视觉问答系统 (VQA) 任务的示例——咖啡机右侧碗中的红色水果是什么？。要回答这样的复杂问题需要确定图像中不同物体之间的关系，从而进行进一步的关系推理。这就涉及到 System 2 推理 (认知系统)。

认知系统是相对复杂的，它涉及到逻辑推理、知识工程、规划方法等技术，这是一个较慢的「有意识」的过程。

在本文中，我们将重点介绍如何将 System 1 (感知系统) 和 System 2 (认知系统) 用于关系推理。

目前，关系推理和预测指的往往是在图数据结构和关系数据上进行预测和推理。下面我们将介绍几个典型的关系预测和推理任务:

- 节点分类 (Node classification)：给定一些节点的标签 (如上图中的红色、蓝色节点)，预测相关节点的标签。
- 知识图谱上的推理 (Reasoning on knowledge graphs)：基于已有事实推测未知事实。例如，已知 Bill Gates 是微软的联合创始人，Paul Allen 也是微软的联合创始人，从而推测以上两人是否具有朋友关系。
- 视觉关系推理 (Visual relational reasoning)：对于 VQA 任务中的非图数据结构 (图像和文本数据集)，通过关系推理和预测回答复杂问题。
- 多跳问答系统 (multi-hop Question Answering)：综合多个事实回答复杂问题，需要理解不同实体之间的关系，在关系图谱上进行推理。

在机器学习领域中，针对于关系推理于预测任务，有两套不同的学习框架，它们分别与 System 1 和 System 2

推理相对应。

其中，System 1 推理系统通过图表示学习技术（如图神经网络）将深度学习用于图数据结构，这类技术包括：

- 节点表示方法：DeepWalk、LINE、N ode2Vec 等
- 知识图谱嵌入表示方法：TransE、TransR、RotatE 等
- 图神经网络（GNN）

然而，System 2 推理则对应于较为传统的统计关系学习，它指的是将概率图模型与知识、逻辑相结合的一些列方法，例如：

- 马尔科夫网络
- 条件随机场（CRF）
- 马尔科夫逻辑网络

其中，马尔科夫网络是将马尔科夫模型与一阶逻辑相结合的产物。下面，我们将通过几个关系预测和推理任务的例子，说明将 System 1 与 System 2 相结合。

二、示例 1：半监督节点分类

节点分类是一类非常标准的简单任务。在图 G 中，节点的集合为 V。V 是有标签节点集合 V_L 和无标签节点集合 V_U 的并集。X_v 是所有节点的特征的集合。在这里，节点分类任务指的是： 给定一些带标签的节点 V_L，预测其余无标签节点 V_U 的标签。

在统计关系学习领域中，条件随机场是一种解决节点分类问题的标准方法。CRF 定义了所有节点标签（y_v）的一个联合分布。我们通过能量函数定义如下的条件概率：

$$p(\mathbf{y}_V|\mathbf{x}_V) = \frac{1}{Z(\mathbf{x}_V)} \prod_{(i,j)\in E} \psi_{i,j}(\mathbf{y}_i, \mathbf{y}_j, \mathbf{x}_V)$$

其中，ϕ _(i,j) 是定义在每条边上的势函数，y_i、y_j 分别代表节点 i、j 的标签，x_v 代表图中所有节点的特征。

传统的统计学习的优势在于，它定义了所有节点标签的联合分布，可以建模不同节点之间的依赖关系。而其缺点在于：

- 统计关系学习需要定义势函数，因此引入了一些人为的干预。
- 势函数通常是较为线性的，其表征能力往往有限。
- 推理过程非常复杂，难以在线性、树状结构之外的图结构上进行推理。

基于图神经网络（GNN）的方法如今十分流行，其本质思想是为每个节点学习出一种性能优异的表征，并基于这种节点表征进一步进行标签预测任务。此时，我们通过神经信息传递（neural message passing）方式学习有效的节点表征。起初，每个节点的表征为其初始的节点特征。我们不断通过各种一层或多层图卷积网络实现信息聚合，将节点邻居与该节点的信息进行融合，从而更新节点的表征。基于这种节点表征，我们可以进一步独立地预测每个节点的标签。

图神经网络的优势在于，我们可以学习到较为有效的节点表征。并且由于我们可以使用多层非线性图卷积网络学习节点表征，这种表征模型的表达能力是很强的。然而，图神经网络的缺点在于，当我们基于已经学习到的每个节点的表征进行预测时，每个节点是独立的，并没有建模节点之间的依赖关系。

在 ICML 2019 上，唐建老师课题组与 Bengio 合作，提出了「图马尔科夫神经网络」。在这篇论文中，他们旨在将传统的统计关系学习方法与图神经网络相结合，学习到性能较好的节点表征，并对节点标签之间的依赖关系进行建模，从而发挥两者的长处。

类似于 CRF，我们首先对所有标签 y_v 的联合分布 $\mathbf{p}_{\boldsymbol{\phi}}(\mathbf{y}_V|\mathbf{x}_V)$ 进行建模，其中 x_v 是这些节点的特征集合。此时，我们的目标是通过最大化观测数据的对数似然 $\log \mathbf{p}_{\boldsymbol{\phi}}(\mathbf{y}_L|\mathbf{x}_V)$ 的下界。而在实际学习模型参数 ϕ 的过程中，由于我们只观测到了部分的节点标签 y_L，没有标签的节点此时都是隐变量，我们无法直接针对观测数据的对数似然进行优化。在这里，我们采用变分方法，转而最大化边缘似然函数的下界（证据下界）ELOB:

$$\log p_{\phi}(\mathbf{y}_L|\mathbf{x}_V) \geq \mathbb{E}_{q_{\theta}(\mathbf{y}_U|\mathbf{x}_V)}[\log p_{\phi}(\mathbf{y}_L, \mathbf{y}_U|\mathbf{x}_V) - \log q_{\theta}(\mathbf{y}_U|\mathbf{x}_V)]$$

其中 $q_{\theta}(\mathbf{y}_U|\mathbf{x}_V)$ 是一个用于近似无标签数据的后验分布的变分分布。

具体而言，我们基于「变分期望最大化」（Variational-EM）算法来进行优化。在 E 步中，我们固定待学习的网络 p_ϕ，更新变分分布 $\mathbf{q}_{\boldsymbol{\theta}}(\mathbf{y}_U|\mathbf{x}_V)$ 从而近似真实数据的后验分布 $\mathbf{p}_{\boldsymbol{\phi}}(\mathbf{y}_U|\mathbf{y}_L, \mathbf{x}_V)$。其中 y_U 代表无标签节点的实际标签，y_L 代表有标签节点的标签，x_v 为节点的特征。在 M 步中，我们固定 q_θ 更新 p_ϕ 从而最大化前面提到的 ELOB。在 ELOB 中，$\log q_{\theta}(\mathbf{y}_U|\mathbf{x}_V)$ 与 ϕ 无关，因此最大化 ELOB 相当于最大化 $\ell(\phi) = \mathbb{E}_{q_{\theta}(\mathbf{y}_U|\mathbf{x}_V)}[\log p_{\phi}(\mathbf{y}_L, \mathbf{y}_U|\mathbf{x}_V)]$ 。

然而，在以往我们处理 CRF 时，直接优化这个联合似然函数也是十分困难的，因为势函数 $p(\mathbf{y}_V|\mathbf{x}_V) = \frac{1}{Z(\mathbf{x}_V)} \prod_{(i,j)\in E} \psi_{i,j}(\mathbf{y}_i, \mathbf{y}_j, \mathbf{x}_V)$ 和其中的配分函数实际上是非常难以计算的。因此我们转而优化下面的伪似然函数:

$$\ell_{PL}(\phi) \triangleq \mathbb{E}_{q_\theta(\mathbf{y}_U|\mathbf{x}_V)}[\sum_{n\in V} \log p_\phi(\mathbf{y}_n|\mathbf{y}_{V\setminus n}, \mathbf{x}_V)]$$
$$= \mathbb{E}_{q_\theta(\mathbf{y}_U|\mathbf{x}_V)}[\sum_{n\in V} \log p_\phi(\mathbf{y}_n|\mathbf{y}_{\mathrm{NB(n)}}, \mathbf{x}_V)]$$

在计算伪似然的过程中，我们将原始的联合似然分解为多个边缘似然之和。对于每个节点来说，我们假设其邻居节点标签已知，我们利用其邻居节点的标签信息来预测该节点的标签。

那么我们如何通过图神经网络来定义其中的一些分布呢?

在推断的过程中，我们需要使用变分分布 q_θ 来近似真实情况下的后验概率 p_ϕ。在这里，我们用到了平均场方法。我们假设所有的隐变量（节点的标签未知）都是独立的，我们可以通过下面的方法将 q_θ 的联合分布分解为多个边缘分布的乘积:

$$q_\theta(\mathbf{y}_U|\mathbf{x}_V) = \prod_{n\in U} q_\theta(\mathbf{y}_n|\mathbf{x}_V)$$

其中，y_n 为图上所有节点的标签，x_v 为所有节点的特征。此时，我们可以通过图神经网络学习图上所有节点的特征表征，对变分分布进行参数化:

$$q_\theta(\mathbf{y}_n|\mathbf{x}_V) = \mathrm{Cat}(\mathbf{y}_n|\mathrm{softmax}(W_\theta \mathbf{h}_{\theta,n}))$$

Object representations learned by GNN

其中，h_(θ，n) 为节点 n 的表征。我们基于该表征来预测节点的标签，本质上是使用图神经网络来做推理。

在 M 步中，我们的目标是最大化上图中的伪似然。我们可以使用另一个图神经网络对 $\log p_\phi(\mathbf{y}_n|\mathbf{y}_{\mathrm{NB(n)}}, \mathbf{x}_V)$ 进行建模，即给定某节点所有邻居节点已知的标签以及节点特征时，预测当前节点的标签。

上述的两个图神经网络分别对应与推理 (inference) 与学习 (learning) 过程，它们通过 E 步和 M 步相互协作。其中，q_θ 对应于推理网络，旨在学习节点的表征，我们可以基于这种表征来预测每个节点的标签，该网络相当于 System 1；p_ϕ 对应于学习网络，旨在以邻居的标签为条件，对节点标签之间的依赖关系进行建模，该网络相当于 System 2。这两个网络的学习过程是相互促进的，我们可以基于推理网络为每个无标签节点预测出一个标签； 基于学习网络进行标签的传递，更新无标签节点的预测标签，作为反馈 / 伪标签提供给推理网络 (System 1)，提升其训练效果。

Category	Algorithm	Cora	Citeseer	Pubmed
SSL	LP	74.2	56.3	71.6
SRL	PRM	77.0	63.4	68.3
	RMN	71.3	68.0	70.7
	MLN	74.6	68.0	75.3
GNN	Planetoid *	75.7	64.7	77.2
	GCN *	81.5	70.3	79.0
	GAT *	83.0	72.5	79.0
GMNN	W/o Attr. in p_ϕ	83.4	**73.1**	81.4
	With Attr. in p_ϕ	**83.7**	72.9	**81.8**

图 1：与传统节点分类算法的性能对比。

我们在一些标准的节点分类任务上将 GMNN 与传统的统计关系学习算法（马尔科夫逻辑网络、关系马尔科夫网络等）和标准的图神经网络（图卷积网络、图注意力网络等）进行了对比实验。实验表明，由于 GMNN 结合了统计关系学习和图神经网络两者的优势，因此它在各项指标上都取得了最佳的性能。

三、示例 2：知识图谱上的推理

知识图谱可以被表征为一些三元组的集合，每个三元组都代表一些事实。知识图谱往往是不完全的，我们往往需要基于已有事实预测未知事实，这是一种标准的知识图谱推理任务。

在专家系统中，我们可以基于一些硬编码的逻辑规则进行推理。然而，逻辑规则往往并不是完全对等的，我们需要在实践中考虑逻辑规则而的不确定性。

在传统的统计关系学习领域中，马尔科夫逻辑网络完美地将概率图模型与一阶逻辑进行了结合，从而对逻辑规则的不确定性进行建模。例如，上图中蓝色的文字部分给出了三条逻辑规则，我们通过马尔科夫逻辑网络来学习每条逻辑规则的权重。在右侧的马尔科夫网络中，每个节点是一个三元组伯努利随机变量，不同的事实通过逻辑规则连接在一起。这种马尔科夫逻辑网络定义了事实的联合分布:

$$p(\mathbf{v}_O, \mathbf{v}_H) = \frac{1}{Z} \exp\left(\sum_{l \in L} w_l \sum_{g \in G_l} \mathbb{1}\{g \text{ is true}\}\right) = \frac{1}{Z} \exp\left(\sum_{l \in L} w_l n_l(\mathbf{v}_O, \mathbf{v}_H)\right)$$

其中 v_O 为观测到的事实，v_H 为待预测（True/False）的事实，w_l 为第 l 条规则的权重，n_l(v_O,v_H) 代表符合实际情况的逻辑规则 l 的数量，该联合分布也可以表示为一个能量函数。

马尔科夫逻辑网络的优势在于：它可以通过逻辑规则利用领域知识，并对逻辑规则的不确定性建模。该网络的缺点在于：图结构较为复杂。推断较为困难。由于逻辑规则并不能完全覆盖所有的事实，因此召回率较低。

在图表示学习中，我们也可以进行逻辑推理。常见的方法为知识图谱表示方法，旨在学习每个实体、关系的嵌入，从而预测缺失的事实（例如，TransE、TransR、RotatE）。知识图谱嵌入同样也定义了所有事实的一个联合分布，但此时我们认为所有的事实都是独立的。我们将基于实体和关系的嵌入预测每个事实的真假（True/

False)。其中 x_h 代表头实体的嵌入，x_r 代表关系的嵌入，x_t 代表尾实体的嵌入。我们可以基于已有的知识图谱嵌入技术定义一个距离函数来定义事实为真 / 为假的概率。在优化过程中，我们将所有观测到的事实 V_O 当做正样本，将未观测到的事实 V_H 作为负样本。

使用知识图谱嵌入的优点在于：优化时可以使用 SGD 和反向传播技术，优化效率高。其劣势在于：难以利用领域知识（逻辑规则）。

四、示例 3：用于推理的概率逻辑神经网络

在 NeurIPS 2019 上，唐建老师研究组提出了「用于推理的概率逻辑神经网络」，希望将传统的基于符号逻辑规则的方法与图表示学习的方法相结合，从而同时通过逻辑规则利用领域知识，并学习较好的节点及关系表示，进行质效皆优的推断。

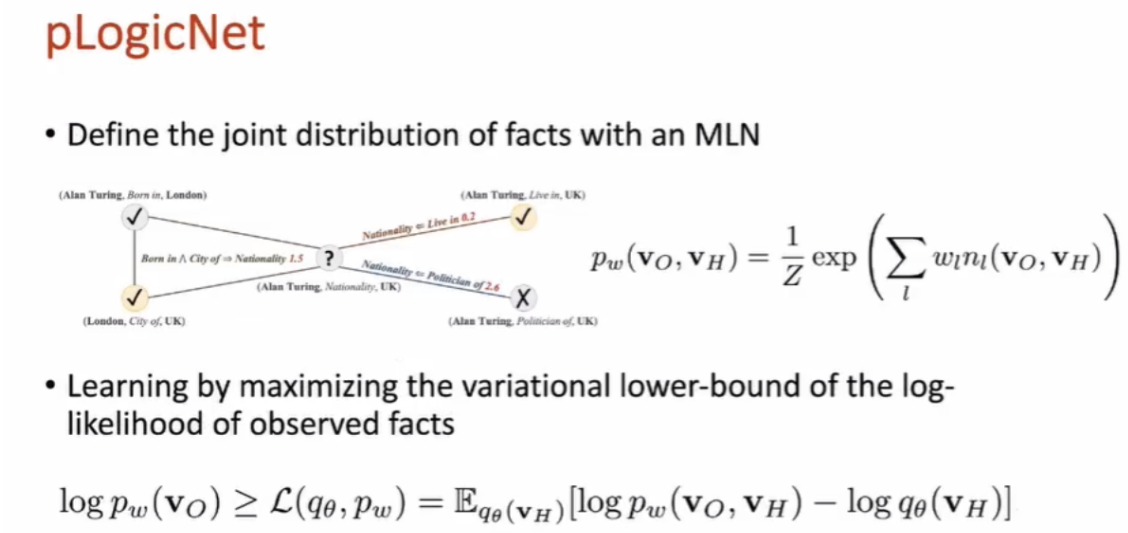


图 2：pLogicNet 的网络定义与学习方法。

在这里，我们通过马尔科夫逻辑网络对事实的联合分布进行建模，同样也利用变分 EM 算法进行优化，通过最大化观测事实对数似然的变分下界来进行学习。其中，$\log q_\theta(\mathbf{v}_H)$ 是隐变量后验分布的变分分布。

$$q_\theta(\mathbf{v}_H) = \prod_{(h,r,t)\in H} q_\theta(\mathbf{v}_{(h,r,t)}) = \prod_{(h,r,t)\in H} \mathrm{Ber}(\mathbf{v}_{(h,r,t)} | f(\mathbf{x}_h, \mathbf{x}_r, \mathbf{x}_t)),$$

图 3：pLogicNet 的推断过程

在 E 步中，我们通过平均场变分推断来近似真实的后验分布。此时，我们假设所有的事实都是独立的，因此可以将这个联合分布分解为每个事实的边缘分布的乘积。我们使用知识图谱嵌入技术，利用实体和关系的嵌入预测每个事实的真假。

$$\ell_{PL}(w) \triangleq \mathbb{E}_{q_\theta(\mathbf{v}_H)}[\sum_{h,r,t} \log p_w(\mathbf{v}_{(h,r,t)}|\mathbf{v}_{O\cup H\setminus(h,r,t)})] = \mathbb{E}_{q_\theta(\mathbf{v}_H)}[\sum_{h,r,t} \log p_w(\mathbf{v}_{(h,r,t)}|\mathbf{v}_{\mathrm{MB}(h,r,t)})].$$

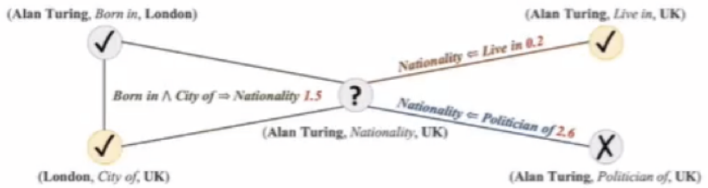


图 4：pLogicNet 的学习过程。

在学习过程中，我们同样选择优化伪似然函数，其优化过程与上文所述的马尔科夫逻辑网络相同。在 M 步中，我们旨在学习每个逻辑规则的权重。

因此，在 pLogicNet 中，我们也是通过 EM 算法框架来优化马尔科夫逻辑网络与知识图谱嵌入。在这个框架中，知识图谱嵌入对应于 System 1，其计算过程是较快的，我们可以基于实体和关系的嵌入快速预测事实的真假。当我们拥有了事实的预测结果后，可以将其提供给马尔科夫逻辑网络。当我们想预测某一个未观测过的事实的真假时，首先可以基于知识图谱嵌入技术预测事实的真假，然后将 System 1 的预测结果输入给马尔科夫逻辑网络进行进一步的推断和修正。之后，我们将马尔科夫逻辑网络的预测结果反过来又作为伪标签输入给知识图谱嵌入系统，使其学习到更好的嵌入，以此循环往复，直到收敛。

Category	Algorithm	FB15k					WN18				
		MR	MRR	H@1	H@3	H@10	MR	MRR	H@1	H@3	H@10
KGE	TransE [3]	40	0.730	64.5	79.3	86.4	272	0.772	70.1	80.8	92.0
	DistMult [17]	42	0.798	-	-	89.3	655	0.797	-	-	94.6
	HolE [26]	-	0.524	40.2	61.3	73.9	-	0.938	93.0	94.5	94.9
	ComplEx [41]	-	0.692	59.9	75.9	84.0	-	0.941	93.6	94.5	94.7
	ConvE [8]	51	0.657	55.8	72.3	83.1	374	0.943	93.5	94.6	95.6
Rule-based	BLP [7]	415	0.242	15.1	26.9	42.4	736	0.643	53.7	71.7	83.0
	MLN [32]	352	0.321	21.0	37.0	55.0	717	0.657	55.4	73.1	83.9
Hybrid	RUGE [15]	-	0.768	70.3	81.5	86.5	-	-	-	-	-
	NNE-AER [9]	-	0.803	76.1	83.1	87.4	-	0.943	**94.0**	94.5	94.8
Ours	pLogicNet	**33**	0.792	71.4	85.7	90.1	255	0.832	71.6	94.4	95.7
	pLogicNet*	**33**	**0.844**	**81.2**	**86.2**	**90.2**	**254**	**0.945**	93.9	**94.7**	**95.8**

图 5：通过算法框架来进行优化

我们在知识图谱推理的一些对比基准上进行了实验。在实验中，我们使用了合成规则 (composition rules)、逆向规则 (inverse rule)、对称规则 (symmetric rule)、子关系规则 (subrelation rule) 这四种逻辑规则。我们将 pLogicNet 与知识图谱嵌入、基于符号逻辑规则的方法进行了对比，由于我们的方法结合了前两者的优势，因此获得了更好的性能

五、ICML 2020 Workshop

在 ICML 2020 上，唐建老师团队联合 Yoshua Bengio 举办了名为「Bridge Between Perception and Reasoning: Graph Neural Networks & Beyond」的 workshop，具体时间为 2020 年 7 月 18 日。感兴趣的读者可以通过链接 (https://logicalreasoninggnn.github.io/) 在线观看。

结语

总而言之，System 1 是相对无意识的、运算较快的系统，对应于快速思维过程；System 2 更多地涉及到逻辑推理、规划等复杂任务的推理，因而是较慢的思考过程。System 1 和 System 2 可以相互促进，相互学习。我们可以基于 System 1 做出快速反应，为 System 2 提供较为初始的预测。System 2 再基于这些初始的预测结果与逻辑规则、领域知识进行更加复杂的思考，将这些推理的结果作为反馈、伪标签作为额外的监督信号提供给 System 1，从而训练出更好的 System 1。

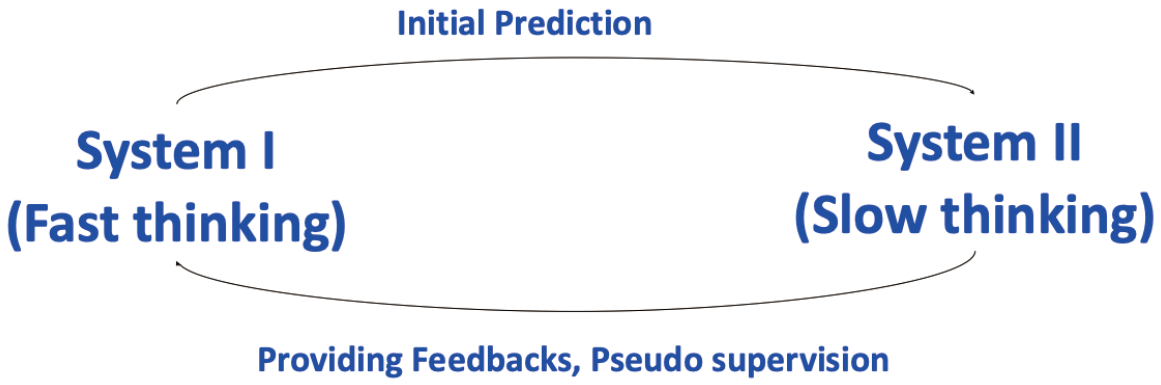


图 6：System 1 与 System 2 的融合方法

在本文中，我们主要以关系推理为例，介绍如何将这两套系统进行融合。首先，在节点表示学习任务重，我们使用图神经网络学习到较好的节点表征，再利用另一个图神经网络学习节点标签之间的依赖关系，以上两个图神经网络可以相互学习。此外，在知识图谱上的关系推理任务中，我们介绍了如何将知识图谱嵌入与传统的基于符号逻辑规则的推理相结合。

中科院研究员沈华伟：图神经网络的表达能力

整理：AI 科技评论

在第二届北京智源大会“图神经网络”专题论坛上，中科院计算所沈华伟研究员做了《图神经网络的表达能力》的报告。

在报告中，沈华伟老师提到：**这几年，虽然图神经网络在其他领域大量应用，但“ 内核”仍然停滞不前**，目前设计新图神经网络（GNN）的两种常用方式都在面临理论上的瓶颈。

沈华伟老师还对近几年图神经网络表达能力的相关研究进行了梳理，他说：“GNN 出现的早期，大家对它表达能力的认识是基于其在半监督学习，尤其是节点分类任务上的优秀表现，一些应用向的研究也只是对图神经网络表达能力经验上的证明”。

基于这个认知，在介绍完图神经网络的基本知识之后，沈华伟老师对图神经网络的表达能力给予了理论上的介绍。

以下是演讲全文。

图神经网络过去几年炙手可热，也取得了一系列的突破，但是这两年发展进入了相对停滞的状态。

当前更多的研究员是把图神经网络当做一个工具，也即把图神经网络泛化到其他领域进行应用方向的研究。例如早期图神经网络在节点分类、链路预测以及图分类上取得了一些进展之后，很快就用在了其他领域，包括推荐领域、自然语言处理领域等等。

其实，图神经网络“ 内核”仍然停滞不前。为什么呢？因为在设计新 GNN 的时候通常有两种方式：一是经验性的直觉，二是试错。而这两种方式都会面临理论上的瓶颈。
另外，如果只靠试错，那么 GNN 和深度学习就会同样存在是否“炼金术”的质疑。那么 GNN 带来的提升究竟来自哪？ 2019 年时，ICLR 发表了题为《How powerful are graph neural networks》的文章，掀起了对 GNN 表达能力的讨论。

一、GNN 表达能力的经验性证明

我们今天报告的主题也是讨论 GNN 的表达能力，看看它到底有什么特色以及限制。

在 GNN 出现的早期，大家的认识是基于其在半监督学习，尤其是节点分类任务上优秀的性能表现，其中以 GCN（图卷积网络 ）为代表。

随后，就有研究员将 GNN 强大的表达能力用在了推理、认知等领域，更有研究员用在了量子化学领域。

例如一个水分子式 H2O，它能告诉我们水分子里面有两个氢，一个氧，此表达方式和自然语言处理里面的 TF–IDF 是一样的，能够看出“词”出现的频次，但没有包含结构信息。

而化学分子式另一种表达，画出结构图的模式相当于把分子结构给理清了，于是，研究员开始思考，这“结构图”是否比原来“频次”方式有更好的表达能力，如果有了表达能力，那能否用包含结构、包含节点的方式对分子式的化学特性进行预测。如果能，表达能力自然得到了证明。

GNN 方法对比传统密度泛函的分析方法，在算力方面有大幅度的节省。毕竟，密度泛函的分析用模拟的方式进行计算，通常需要高性能计算机进行操作。在现实情况中，GNN 确实做到了化学精度以内的预测，用的就是 message passing GNN 的方式。

所以，如果能用 GNN 直接拟合一个模型，从而对任意一个新的分子进行性能预测，自然就体现出来了 GNN 的表达能力，但这只是经验性表达能力的认识。

另外，在《自然》子刊上，也有一些用 GNN 建模预测的文章发表，例如预测玻璃的动力学特点、预测地震等等。这些案例也都是对 GNN 表达能力提供了经验性的认知，我们接下来想从理论的角度讨论 GNN 的表达能力。

二、GNN 基本知识

介绍一下基本的知识，GNN 的输入是一张图，图中有节点集合，参数包括 V、E、W、X 等。GNN 早期典型的两类任务是节点分类和图分类，在节点分类任务中，GNN 的目的是训练得出网络当中节点的表达，然后根据节点的表达进行下游的任务；在图分类任务中，GNN 的目的是要训练得到图的表达，有了图表达之后再对图做分类。在这两个典型的任务中，目前 80%、90% 的工作都倾向于节点分类，只有少数是图分类。

关于 GNN 的标准框架，目前用得比较多的是 Aggregate+Combine，此框架比较灵活，图分类任务和节点分类任务都适用，其策略方式是通过迭代，用邻居的表示迭代更新自己的表示。策略一共分两步，第一步是聚合邻居信息；第二步是把邻居信息和自己上一轮的信息进行耦合。

下面举几个这种框架的例子，第一个是 2017 年提出的 GraphSAGE，其操作是把邻居的表达进行变换之后，取里面最大的一个赋给自己，然后再把学到的表达和自己上一轮的表达做整合，随后得到新的表达。值得一提的是，GraphSAGE 用了 Max–pooling 的方式，此方式限制了他的表达能力，是导致表达能力丧失关键的一步。

GCN 的表达方式直接，将邻居进行特征变换，然后按照矩阵规划进行传递，它的特点是把 AGGREGATE 和 COMBINE 两个操作进行了整合。值得注意的是，GCN 采用了 Mean–pooling 的方式，此方式也限制了它的表达能力。另外，GCN 的改进版是 GAT，其采用的方式是 weighted mean pooling。

三、图神经网络的表达能力如何

前面是关于图神经网络基本介绍，现在回到今天的主题：图神经网络的表达能力。我们先讨论 2019 年发表在 ICLR 上的《How powerful are graph neural networks》。

首先明确什么是表达能力，所谓表达能力一般有两个方面，**第一个方面是表示的空间大小**，例如，一个 N 位的

二进制能表达 N 的二次方个数。这种表达能力旨在表示多少不同的东西，不同的结果。

第二个方面是关于近似能力。例如设计一个神经网络能够近似什么样的函数。值得一提的是，在 1989 年的时候就有了证明： 神经网络的层次只要足够深，就可以逼近任意连续函数。这个“万能近似定理”也解释了为什么深度学习从来不担心表达能力的原因。

但是 GNN 提出之后，深度学习表达能力的话题又被提起，2017 年有研究员发现深度学习的表达能力和深度神经网络的层次是指数关系，假如网络有 D 层，那么表达能力与“某个数”的 D 次方成正比，大家感兴趣可以看相关的论文。

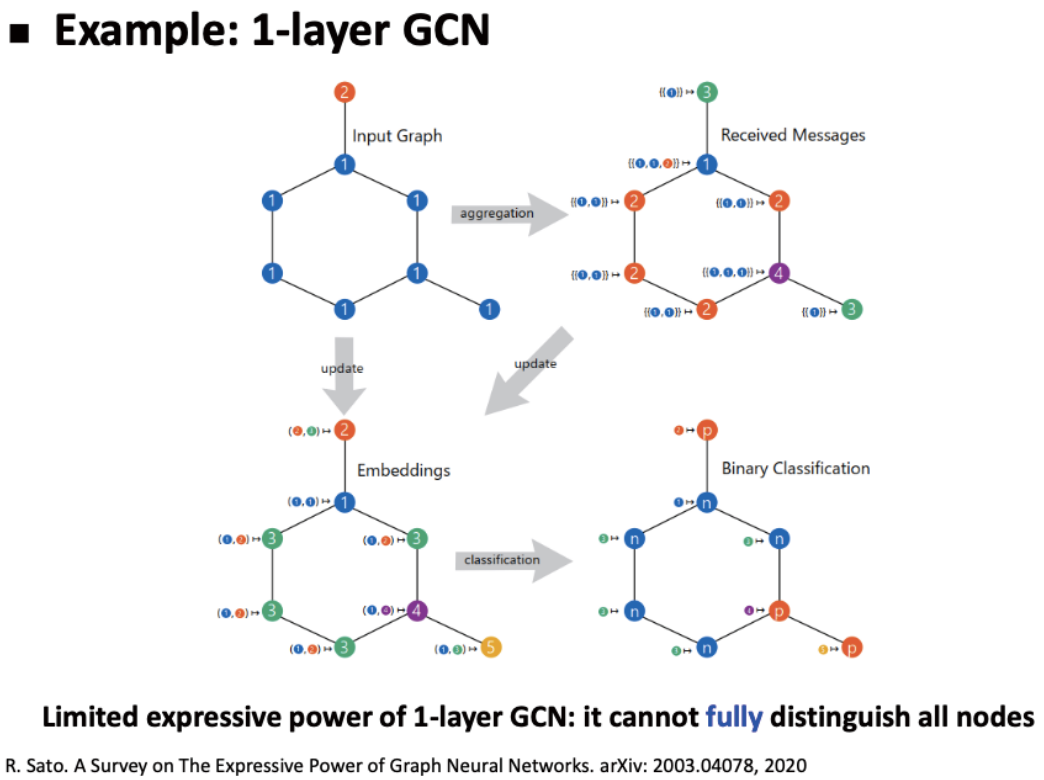


图 1： 1–layer GCN

GNN 引进之后，对于表达能力有什么样的启示呢？如上述左图所示，如果不看结构，节点的标号标 1 还是标 2 是区分不开。如果想区分这个不同的“标号 1”，需要观察标号的邻居，可以通过邻居信息进行区分。GCN 可以把邻居信息进行聚合，提升区分节点的能力。如上图左下所示，在一层 GCN 操作完成之后，已经可以区分一些标号，但左下图四个“标 3”的点还是区分不出来。

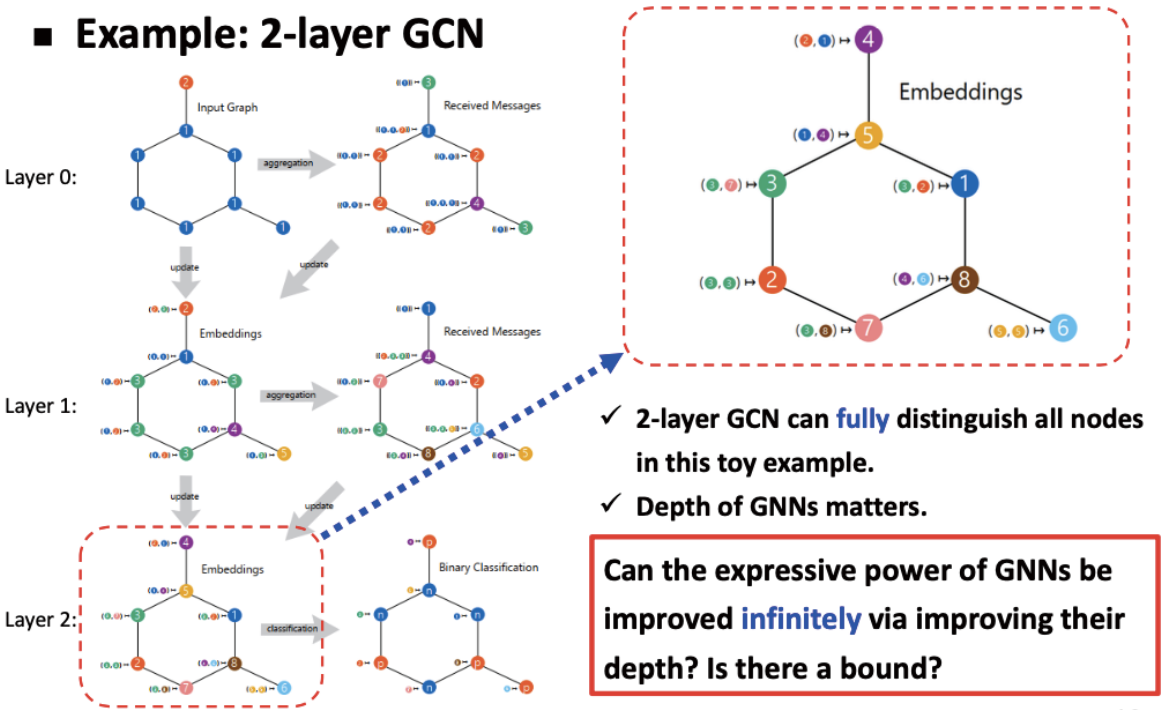


图 2：2–layer GCN

所以，一层 GCN 区分能力并不够，那能否通过加深层次解决表达能力呢？两层 GNN 之后，如上图所示，变成了八个点，并可以完全区分开。

所以，如果用两层 GCN 对上述节点进行分类，无论 Label 标记成何种方式，GCN 的表达能力都能满足分类要求。

上面是两层 GCN 完全可以区分的例子。回到刚才的问题，**把 GNN 加深就一定能把表达能力做上去吗？**也就是说，我们能不能通过深度实现无穷大的表达能力？ 2019 年那篇 ICLR 文章回答：不可以！

上面是节点的角度，下面从图的角度进行讨论，也即如果把不同的图做成一个表示，GNN 表示图的方面表达能力如何。这里有两个关键因素，**一个是节点的特征，一个是图的结构，**节点的特征刚才已经讲过了，如果把节点做深度神经网络，已经可以帮我们解决掉表达能力问题。

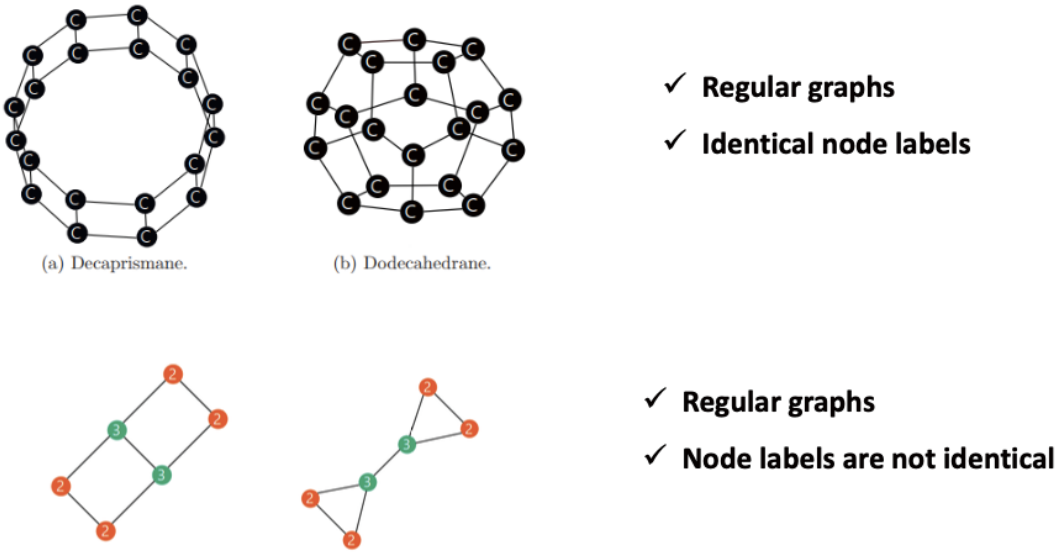


图 3：通过 GCN 区分有机化学分子式

所以，另外一个表达能力的限制就来自于图的结构。

下面看两个简单的例子，左上角的图都是 24 个碳原子，有两个有机化学的分子式： 左边的结构是两层，一层 12 个碳，24 个碳分成平行的两层，都和自己的邻居相连。右边图是 24 个碳结构变成一个正球体，每个面都由五个碳构成。这两个结构，人一眼就能看出它俩的不同，但是 GCN 无法区分，即使把层次加深到无穷多层也区分不了它俩。

即使简化成左下角两个更加简单的图，GCN 也区分不出来。所以，这给出的启示是： 通过提升 GCN 的深度来提升图的表达能力，是无法做到的。

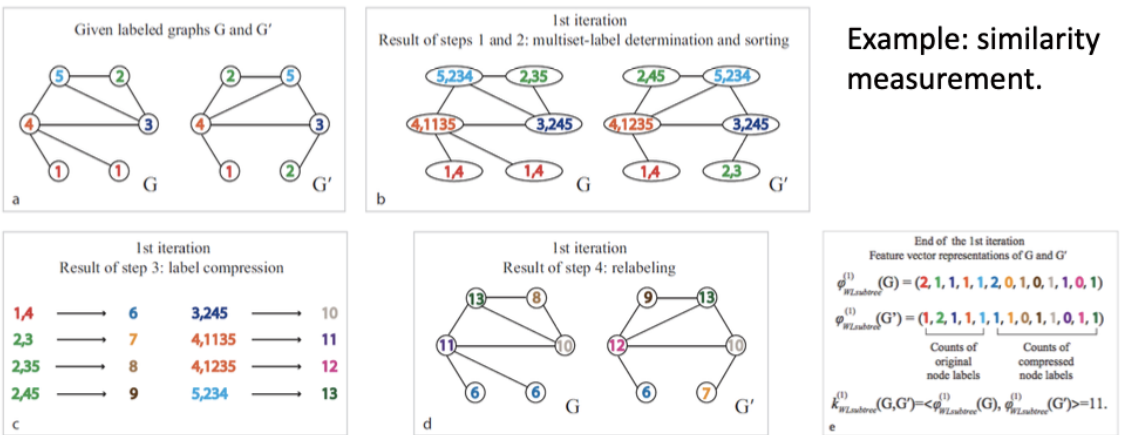


图 4：Example: similarity measurement

那么它的表达能力受限在哪儿？既然是图的结构相关，那我们就想到可以采用 WL-test，对两个图是否同构进行测试。

WL test 的主要是通过聚合节点邻居的 label，然后通过 hash 函数得到节点新的 label，不断重复知道每个节点的 label 稳定不变。稳定后，统计两张图的 label 的分布，如果分布相同，则一般认为两张图时同构的。

所以，WL test 递归聚合邻居的方式和 GNN 类似，都是一个递归地来更新自己的特征，只不过更新的方式，WL test 做了一个单射函数，但是 GNN 用的聚合函数，无论是 Max、Mean 还是 Sum，大部分情况下都不是单射的。也就是说 GNN 非单射的聚合方式，把原本不一样的东西聚合后变得一样了，这让 GNN 丧失了表达能力。

更直白一些，WL Test 是一个单射函数，GNN 不是单射函数，于是 WL Test 为 GNN 的表达能力提供了一个理论上的上界。(注：这里 GNN 说的是通过邻居聚合，如果别的聚合方式，性能可能超过 WL test 的)

为什么当前流行的 GNN，例如 GCN、GraphSAGE 为什么不是单射呢？原因有两个，**一个是每层做特征变换做得不够；另一个是，这些图神经网络用的聚合函数天然具有单射属性。**

所以，在论文《How Powerful are Graph Neural Networks》中，作者提出来了新的图模型 GIN(Graph Isomorphism Network)，其中 I 表示图同构，关键思想是构建了一个单射函数。有了单射函数，GIN 也达到了和 W L test 类似的表达能力，达到了图神经网络表达能力的上界。

后来作者对 GIN 模型的表达能力进行了验证，具体是用数据的拟合能力进行测评，验证思想是：**如果表达能力足够强，那么数据集上的每个数据都能精确拟合。**验证结果确实符合作者的预期，通过构造一个单设的聚合函数，能够实现和 WL Test 一样的表达能力。

那么，泛化能力如何呢？也即在 Test Loss 表现如何呢？这里有一个直观上的事实是，如果 Training Loss 做得不好，那么 Test Loss 表现也不会好，毕竟 Train 是 Test 的基准，另外，如果训练集上表现非常好，而在测试集上表现非常差，那么就出现过拟合现象，没有泛化能力了。

提出 GIN 的作者也在论文中证实了，GIN 在表达能力上很强，但是在其他任务上，泛化能力还不如表达能力差的模型，如上图 GIN 在几个数据集上的表现。

所以，这给图神经网络带来的启示是：高的表达能力，并不总意味着对下游任务友好，但是低表达能力总是不好的。综上，我们还是希望构造出一个强表达能力的图神经网络，这是当前学界研究员的共识。

总结一下 ICLR 2019 那篇论文的工作：首先 GNN 在理论上有一个表达能力的上界，这个上界是 WL Test；为什么是 WL Test？因为它的聚合函数是单射；同时这篇论文又提出了 GIN 这一有着单射聚合函数的图神经网络，并对其表达能力进行了验证。

四、待讨论问题：图神经网络的前沿

那篇文章在 2019 发表之后，引起了很大的关注，其实后来有很多人进行了讨论，我把这些问题抛在这里，大

家讨论一下。

第一个问题，高的表达能力，到底是不是必要的，我们有没有必要构造出这么高表达能力的图神经网络？我们能否做出一个通用的极强表达能力 GNN，然后再也不用考虑表达能力这个问题了？

我们现在并没有得到这个问题的答案。对于节点分类，基本上可以提供 universal approximitor，对于图分类无法做到，不仅做不到，而且有些场景下表现还特别差。

那么，对于特殊的任务，我们有没有必要构造出高表达能力的东西呢？前面提到，如果表达能力很差，泛化能力肯定不好，表达能力好的，泛化能力也未必好。这在一定程度上也解释了为什么 GNN 和 GraphSAGE 聚合函数不是单射，表达能力有限的情况下，为什么还能在一些任务上表现非常棒。

在一些场景下，GNN 的大部分表达能力其实用不上。我们真正需要什么呢？**我们需要的是它可以把相似的对象，例如相似的节点和图映射成相近的表达。**

那么问题又来了，用什么衡量是否相似？所以就有很多度量两个图是否相似的工作。另外，判断一个复杂对象，能不能分解成简单对象进行表达这也是个值得探讨的问题。

第二个问题，关于结构。其实我们都希望 GNN 学到结构，大家研究 GNN 这几年，也都明白了 GNN 在结构上无能为力，只是用结构进行了约束，做了平滑。

举例而言，什么是一个好的图表达？假设一个分子结构图里面有一个苯环，能不能把这个分子式分成苯，还是说分子式中有很多苯环的情况下，才能分成苯。

这个问题的本质，其实在回答我们做的是信息抽取还是相似性度量。如果想做信息抽取，也就是想识别出分子式里面有没有苯结构，现在的 GNN 做不到这一点，或者必须再设计一些别的方式才能达到。

所以，这两年大家一直在思考，GNN 研究的是模式识别，还是说只是图相似性的度量方式。

第三个问题，能不能构造一个更强大的 GNN 呢？也即表达能力更强大的 GNN？关于表达能力，一阶 WL Test 已经在理论限制突破能力。这两年大家更多的研究方式是把常用的 1 阶 WL Test 拓展成 K 阶，所以就有了 KGN 的方式。在这样 K 阶的 WL Test 方式下，表达方式已经突破 1 阶的能力，但是成本也比较大，因为需要处理的对象增加了很多。

这种方式给大家起了抛转引玉的作用，给提升 GNN 表达能力提供了一种思路。但是这种把所有可能都列出来的方式并不是我们所需要的，我们想要的是一个 layer-by-layer 的网络，也即如果网络每一层非常简单，层次的堆叠是逐渐提升的，然后获得一个更强大的表达能力。

所以，layer-by-layer 网络也是未来几年大家应该去探索的一个问题。所以现在我把这个问题抛出来了，你能设计一个这样 layer-by-layer 的网络，从而获得一个比 GNN 更强大的表达能力吗？

微软高级研究员熊辰炎：图神经网络在信息检索等领域的应用

整理：智源社区 熊宇轩

对于许多信息检索和知识图谱研究者来说，究竟应该使用抽象的结构化信息进行表示学习还是使用海量的文本信息始终是一个富有争议的话题。在本届智源大会上，来自微软研究院的高级研究员熊辰炎博士带来了题为“利用半结构化知识的表示学习与信息检索”的主题报告，结合其近年来在 ICLR、ACL、WebConf 上发表的相关工作，介绍了如何从半结构化知识的视角同时利用符号知识与纯文本信息，从而提升表征性能与效率。

尽管本次演讲的标题中没有「图神经网络」等字眼，但其内容都围绕图神经网络展开。本次演讲将侧重于实际的问题、知识、工业界常用任务中的半结构化数据，探讨如何利用图神经网络对半结构化数据进行表示学习，以及如何使用较为统一的框架解决实际中的问题。

演讲正文:

本次演讲的内容主要分为两部分，首先，我们将从统一的「半结构化」的视角讨论知识图谱以及各种信息检索任务（例如，问答系统、事实验证、假新闻检测、信息搜索）；接着，我将介绍我们近期提出的一种 Transformer 模型，它能够整合各种不同任务的信息，并学习其表征，从而完成这些任务。

一、知识和信息检索任务的「半结构化」视角

图 1：符号知识 V.s. 自由的纯文本信息

在知识工程与自然语言表征学习领域，往往有两种对信息建模的视角。首先，对于许多任务来说，我们拥有的是结构化的数据（即符号化的知识），知识图谱就是其中一种形式。例如，图中每个节点都是一个命名实体，实体之间的边代表关系。另一方面，在有的任务中，信息则存在于原始的纯文本之中。

这两种建模信息的方式各有千秋。一方面，结构化数据十分干净而精确，数据十分规整，我们可以在这种该结构化数据上进行各种推理，或者基于它们开发一些可执行的程序（如 SQL 查询，或图数据库的查找或搜索）。然而，构建结构化数据的成本是很高的，并且现实世界中的一些信息也很难被表示成这种规整的结构。

而对于纯文本信息来说，可以使用的语料的数据量往往非常大，我们可以利用各种自然语言处理（NLP）技术处理这些文本（例如，信息提取、文本表征）。然而，纯文本信息往往不够精准，存在各种噪声，其结构也不太明显，这不利于我们进行后续的操作。

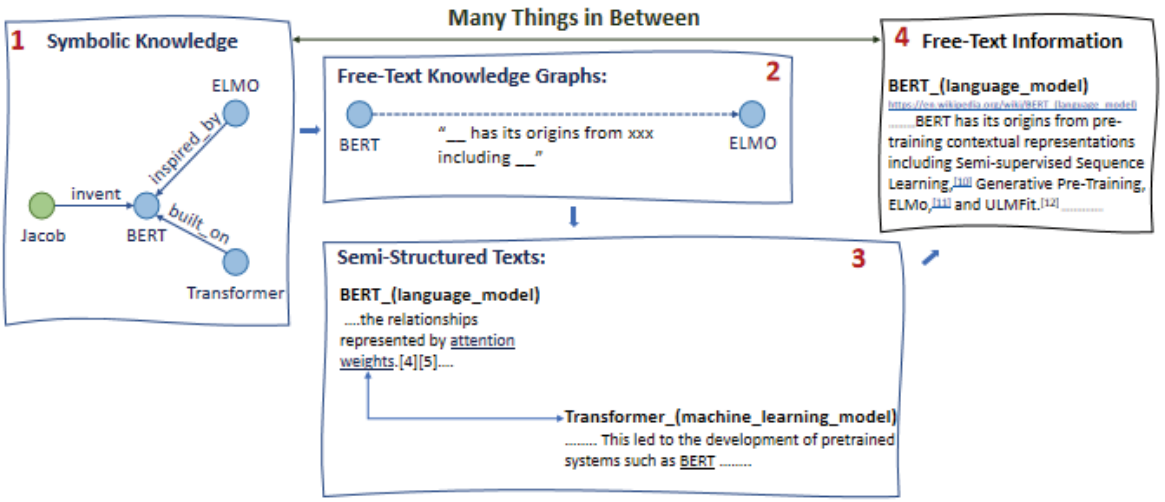


图 2：半结构化信息

我们认为，许多信息处于结构化数据与纯文本数据之间，我们将其称为「半结构化」数据。
举例而言，对于知识图谱来说，我们可以对知识图谱进行一定的松弛（例如，用某些纯文本作为边）。而对于文本信息来说，文档之间可能存在各种各样的关系，我们可以将这些文档作为图谱中的各个节点，用各种关系边将它们相连。如上图所示，1 → 4 展示了我们如何一步步将结构化符号信息松弛为纯文本信息的层次化过程。

二、Free-Text Knowledge Graph

如今，研究人员开发了各种各样包含大量优质信息的知识图谱（例如，Freebase、WikiPedia）。然而，真实场景下有很多并不完全针对这些知识图谱设计的实际问题，我们应该如何将这些知识图谱应用于这类问题呢（例如，将知识图谱引入网页搜索引擎，从而提升文档排序的性能）？

实际上，要实现这一目标仍然存在许多挑战，其中最主要的一个挑战是： 由于进行信息提取和知识图谱扩容是相对困难的，因此其召回率往往较低。例如，某系统中仅仅 1% 的流量数据可以被知识图谱中的边覆盖，无论我们在 1% 的流量上执行某任务的效果有多好，其对整体系统的影响仍然十分有限。

尽管如此，我们发现知识图谱对于命名实体（通常为节点）的覆盖率仍然是很高的。2015 年，我在一个实际的网页语料库上使用各种实体链接，从而观察每个查询或文档中包含多少个实体。实验结果表明，在大约只有 2–3 个词的查询中，平均覆盖了高达 1.5 个实体； 而在大约有 500 个单词的文档中，平均覆盖了

约 252 个实体。几乎所有的查询、文档都包含至少一个实体。当我在 Allen 人工智能研究院实习时，我也对其 SemanticScholar 的查询上使用了实体链接，我惊讶地发现有高达 70% 的学术搜索查询在 Freebase 或 WikiPedia 这种非学术的通用知识图谱上有相对应的实体。

此外，从新的文本中抽取出实体也并非十分困难。例如，新冠疫情爆发后，学者们发表了大约 10 万篇相关的医学论文，我们直接将基于 BERT 的关键词标注器（并未在医学领域上进行预训练）应用于跨领域、零样本的关键词抽取。实验结果表明，该关键词标注器确实能够抽取出重要的关键词。

Google's Knowledge Vault [KDD 2014] Dong et al. Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion

Name	# Entity types	# Entity instances	# Relation types	# Confident facts (relation instances)
Knowledge Vault (KV)	1100	45M	4469	271M
DeepDive [32]	4	2.7M	34	7M[a]
NELL [8]	271	5.19M	306	0.435M[b]
PROSPERA [30]	11	N/A	14	0.1M
YAGO2 [19]	350,000	9.8M	100	4M[c]
Freebase [4]	1,500	40M	35,000	637M[d]
Knowledge Graph (KG)	1,500	570M	35,000	18,000M[e]

Table 1: Comparison of knowledge bases. KV, DeepDive, NELL, and PROSPERA rely solely on extraction, Freebase and KG rely on human curation and structured sources, and YAGO2 uses both strategies. Confident facts means with a probability of being true at or above 0.9.

图 3：覆盖关系更为困难

然而，覆盖关系（知识图谱中的边）相较于覆盖实体来说要更为困难，对边的召回率往往要低一些。Google 研究院在 Knowledge Vault 上的实验表明，其使用的本体非常多，其关系边的数目也很多，但是其边的类型相较于 Freebase 并没有改变（仍然为 35,000 种边），这是因为定义边的关系类型是十分困难的。

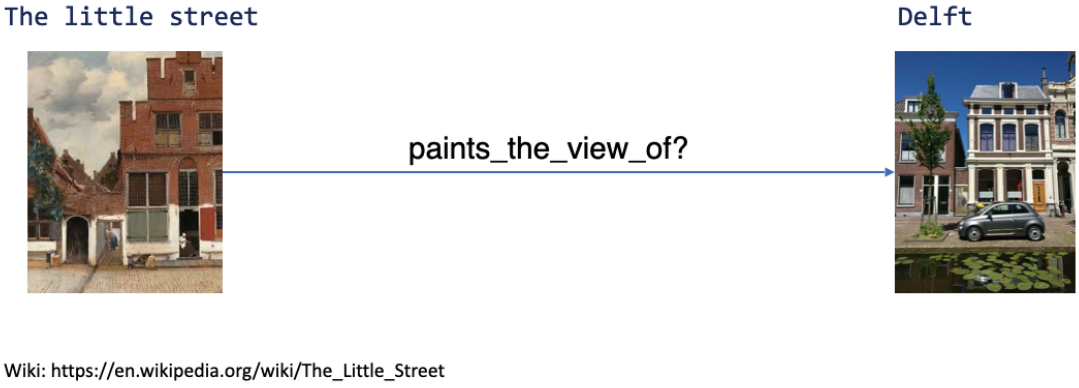


图 4：关系符合长尾分布

在许多实际问题中，实体之间的关系实际上满足具有严重长尾效应的分布。如上图所示，在实际的视觉问答系统数据集中，左侧的实体「The little street」是一幅画，画的内容是右侧荷兰小城 Delft 的街景，这两个实体在维基百科中都可以很容易找到。但是，此时我们希望识别这两个实体之间的关系（paints the view of），并用资源描述框架（RDF）将其记录下来。

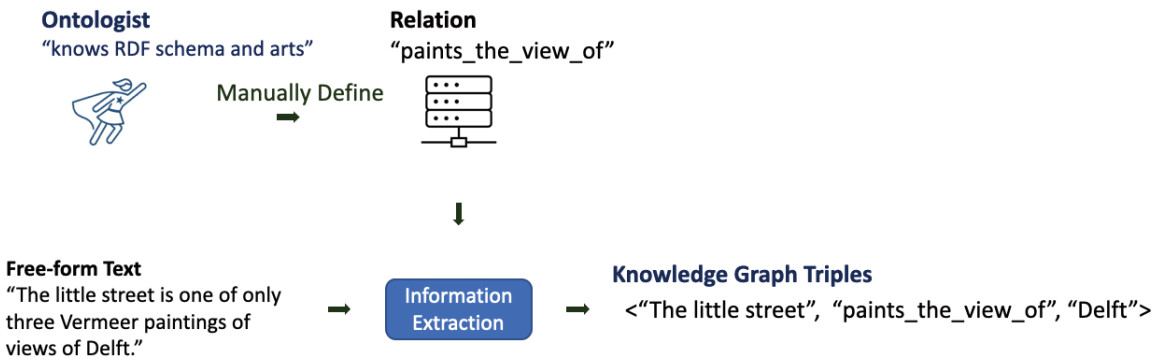


图 5：关系构建的处理流程

为了构建一个可以覆盖该关系的知识图谱，我们需要即熟悉 RDF 的模式，又了解艺术的本体学家（Ontologist）来定义该关系，而这种人才往往很稀缺。接着，我们需要通过信息提取技术从纯文本中抽取出相应的（实体、关系、实体）三元组，而能够顺利通过上面提到的整个处理流程的关系实际上并不多。

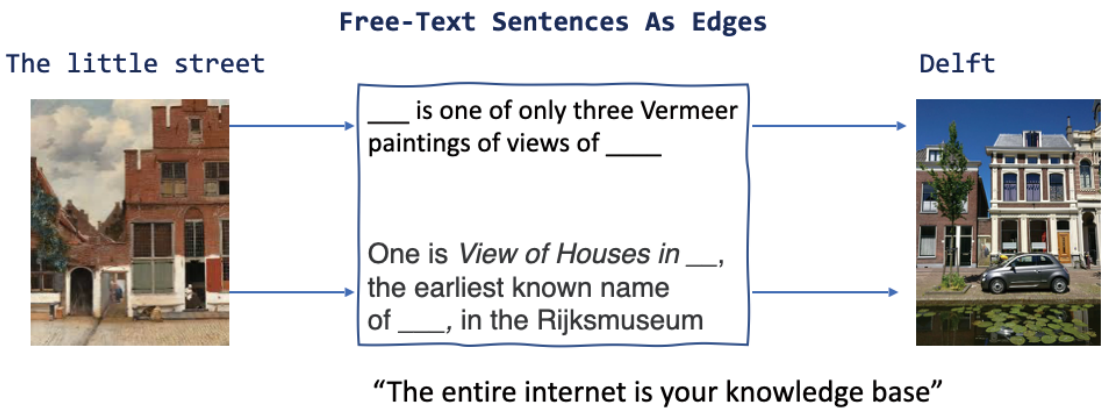
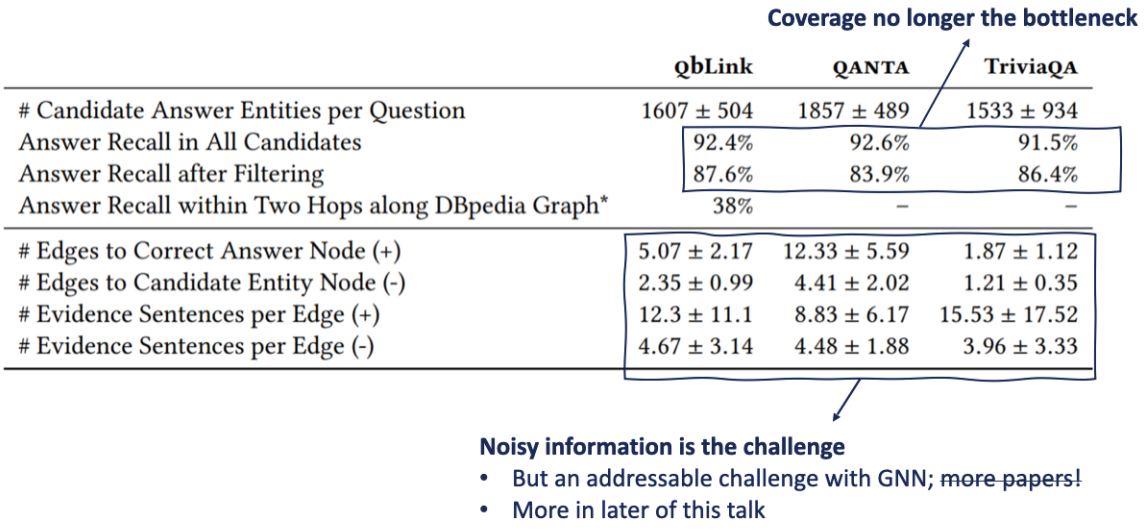


图 6：Free-Text 知识图谱

实际上，这里的纯文本本身就存在于维基百科或其它语料库中，上述复杂的处理流程费时费力，不一定能够满足一些具有时效性的任务的需求。因此，我们考虑跳过上述流程中的一些步骤，不再要求边是规整的关系，转为使用各种半结构化的边（句子）。近年来，许多知识图谱领域的研究者们认为「整个互联网就是我们的知识库」，这种半结构化的思想在近期发表的一些论文中也有体现。因此，要构建这种 Free-Text 知识图谱，我们首先找出已有的实体，然后在网页语料库中使用实体链接，将所有包含该实体的句子抽取出来，基于这些句子和实体连成一张图，从而提升关系的召回率。由此得到的图是覆盖率非常高的密集图。



	qbLink	QANTA	TriviaQA
# Candidate Answer Entities per Question	1607 ± 504	1857 ± 489	1533 ± 934
Answer Recall in All Candidates	92.4%	92.6%	91.5%
Answer Recall after Filtering	87.6%	83.9%	86.4%
Answer Recall within Two Hops along DBpedia Graph*	38%	–	–
# Edges to Correct Answer Node (+)	5.07 ± 2.17	12.33 ± 5.59	1.87 ± 1.12
# Edges to Candidate Entity Node (-)	2.35 ± 0.99	4.41 ± 2.02	1.21 ± 0.35
# Evidence Sentences per Edge (+)	12.3 ± 11.1	8.83 ± 6.17	15.53 ± 17.52
# Evidence Sentences per Edge (-)	4.67 ± 3.14	4.48 ± 1.88	3.96 ± 3.33

图 7：在问答系统数据集上的 Free-text 知识图谱的覆盖率

针对 qbLink 数据集，我们在 DBpedia 知识图谱上从核心实体开始，沿着其边游走，我们发现在 2 步之内有 38% 的概率找出对应的答案，我们可以将 38% 看做仅仅使用 DBpedia 这种基于纯结构化数据的知识图谱时所能达到的覆盖率上限。而如果我们将句子作为图谱中的边，那么有 80–90% 的答案可以被找到，此时召回率明显提升。

然而，此时图的规模可能会非常大，噪声非常多，与每个实体、答案相连的边数会显著增长，而其中可能存在大量无效的路径。此外，每条边可能对应于很多的句子，而这些句子可能良莠不齐。因此，我们虽然突破率覆盖率的瓶颈，但是引入了更多的噪声。**幸运的是，我们可以使用图神经网络很好地处理这些信息，具体处理方式将在后半段演讲中介绍。**如何设计模型从这种噪声较大的图中提取信息是一个有待研究的方向。

如图 2 中第 2 步与第 3 步所示，这种半结构化的形式化定义的适用范围非常广泛。例如，在事实验证和假新闻检测任务中，我们需要根据语料库（如维基百科）中的文档和段落（证据）验证声明（claim）的真实性，给出「真」（support）、「假」（Refute）、「信息不足无法验证」（Not Enough INFO）其中的一种结果。该任务可能需要综合考虑多篇文档的信息，因此对于该任务而言，我们可以将声明与纯文本证据作为 Free-Text，将语料库中的证据之间的推理链条、共现、超链接、指代关系等作为结构关系，从而构建半结构化数据。

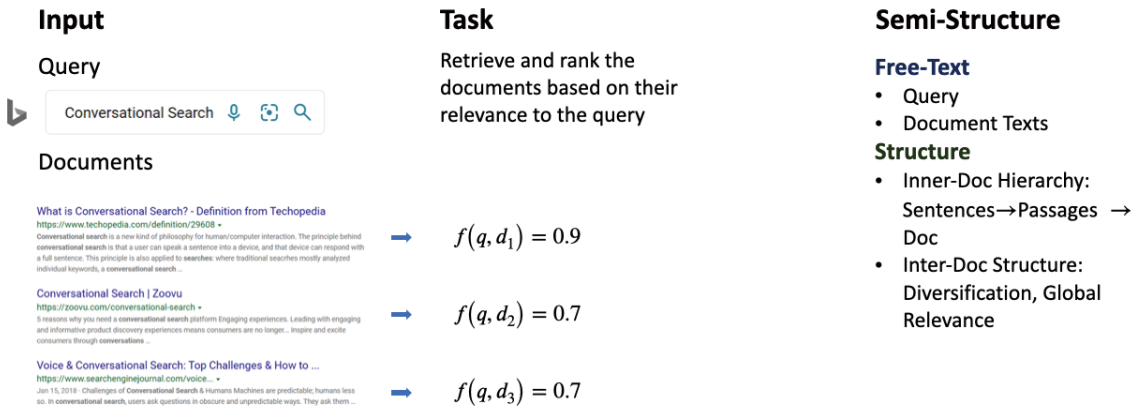


图 8：文档排序问题

在搜索引擎的文档排序问题中，当我们输入一条查询时，搜索引擎会检索出若干文档，我们需要在语料库中找到这些文档并对它们进行排序，输出一个排序得分。此时，在构建半结构化数据的过程中，我们可以将查询与文档作为 Free-Text，而将文档内部的层次化语法结构、文档之间的关系（相似、全局相关性等）作为结构关系。

在多跳问答系统中，我们需要利用多种树状结构、链式结构的信息作为证据。在构建半结构化数据时，我们使用不同文本之间隐含的推理关系、超链接、相关性作为结构关系。

三、Transformer-XH

为了统一地解决面向半结构化数据的任务（QA、文档排序、事实验证等），我们可以使用 Transformer 类模型（如图注意力网络 GAT、BERT 等）对这种半结构化数据进行建模。

在多头注意力机制中，对于节点、单词等表征，我们通过注意力机制来更新它们，得到更复杂、能够表达全局语义的信息表征形式。对于每一个注意力层而言，面对从下一层输入的表征，用查询权值矩阵、键权值矩阵、值权值矩阵与该输入相乘，再将这种投影作为查询向量、键向量、值向量。当网络对某个单词进行表征时，会使用该单词的查询向量 q_i 与每个单词的键向量 k_j 相乘得，再对其乘积进行 softmax 操作得到加权和为 1 的注意力得分，用该得分将所有单词的值 v_j 结合在一起，用最后计算的结果更新该单词的表征。

这种自注意力机制是非常普适的。例如在 GAT 中，我们会沿着已有的图中的边，构建注意力路径，对节点之间的关系进行建模；对于 BERT 而言，我们可以将其视作一个全连接的词袋图。

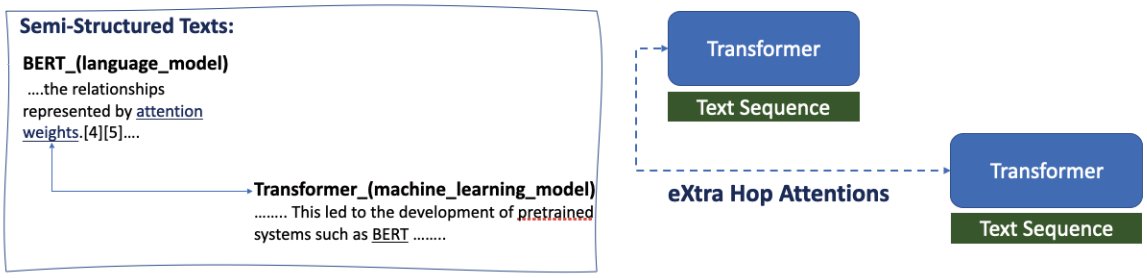


图 9：Transformer–XH：eXtra–Hop Attention

在我们于 ICLR 2020 上发表的论文「Transformer–XH: Modeling Semi–structured Information with eXtra–Hop Attentions」中，我们用 Transformer 类的模型（如 BERT）生成文本块的表征嵌入，再用 eXtra–Hop 注意力将这些文本的嵌入联系起来。这相当于一个双层结构，在每一个文本内部，我们使用全连接的 GAT 对每个词之间的关系建模，而在不同的文本序列之间，我们也会利用 eXtra–Hop 注意力建立与 GAT 相类似的注意力路径，这是一种 Transformer（BERT）+GNN 的模式。

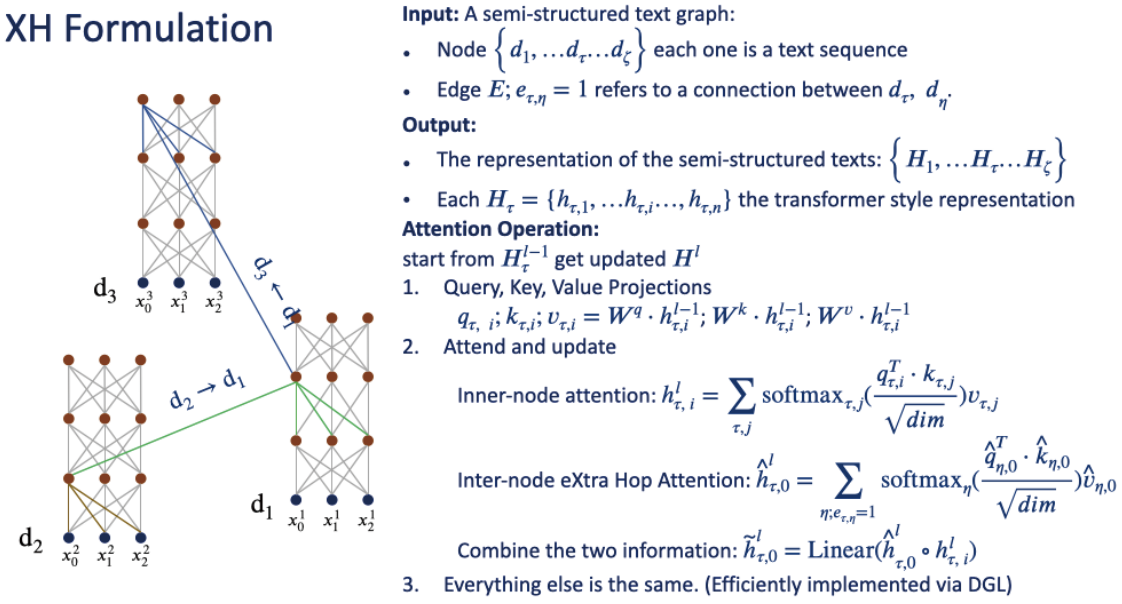


图 10：Transformer–XH 的形式化定义

对于文本序列节点 {d_1,...,d_τ,...,d_ξ}，我们会用一些邻接矩阵记录将这些文本序列连接起来的边，该模型旨在综合考虑各节点、各文本序列之间的信息，输出每个节点上的表征。该模型与普通 Transformer 的区别在于，我们用 τ 表示某一个文本序列的编号，每个文本序列独立使用一套查询、键、值投影矩阵。此外，我们还将每一个文本序列第一个位置上的特殊标记 [CLS] 作为注意力中心（attention hub），将其与其它文本的注意力中心通过 eXtra hop 注意力相连。通过以上机制，我们可以同时考虑本文序列内部各词例（token）之间的关系表征，以及其它文本序列的 [CLS] 的信息表征，对这两部分表征进行连接（concatenate）后再对进行线性投影。

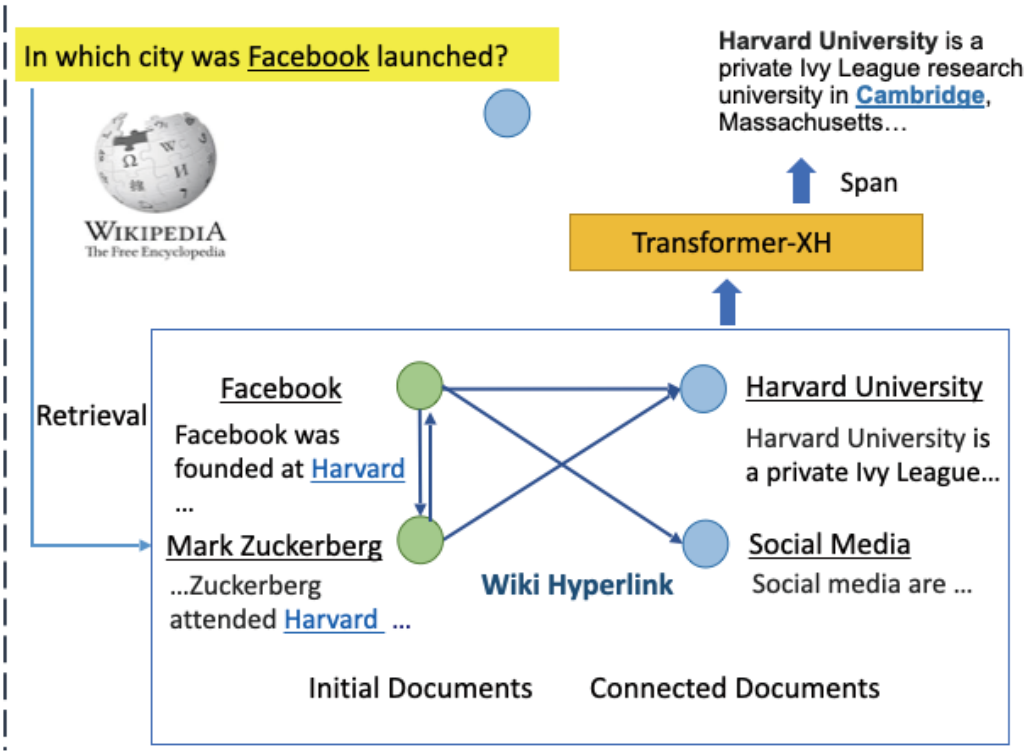


图 11：使用多文档信息提取答案区间（answer span）

在具体的多跳 QA 任务中，针对上图黄色方框中的查询，我们检索出了若干文档，沿着维基百科的超链接扩展出相应的文档节点。接着，我们整合所有文档的信息，通过 Transformer-XH 得到表征，将其表征用于提取答案区间。具体而言，我们会在 Transformer-XH 的最后一层后面接上一个阅读理解（MRC）任务常用的预测层，用于预测答案区间的开始于结束位置。

在进行事实验证时，针对一个声明（claim），我们检索出与其相关的若干文档，通过与提取答案区间任务中相似的方式生成文档表征。针对这一任务，我们预测每个文档起始位置节点 [CLS] 的标签，作为节点级别的分类概率。接着，我们学习出各 [CLS] 节点的权重，对它们进行加权平均，从而完成事实验证任务。

因此，对于面向半结构数据的任务而言，底层的表征模型都是类似的，都需要将词、关系输入到基于注意力的表征模型中，而我们重点关注的是上层针对特定任务的层如何设计。

Hotpot QA 是由 CMU、MILA 等单位发布的针对多跳问答系统开发的数据集，旨在向亚马逊的众包「土耳其机器人」提供两个段落的文本，要求他们写出一些需要利用这两个才能回答的问题，而最后确实有大约一半的问题需要用到多个段落的信息。

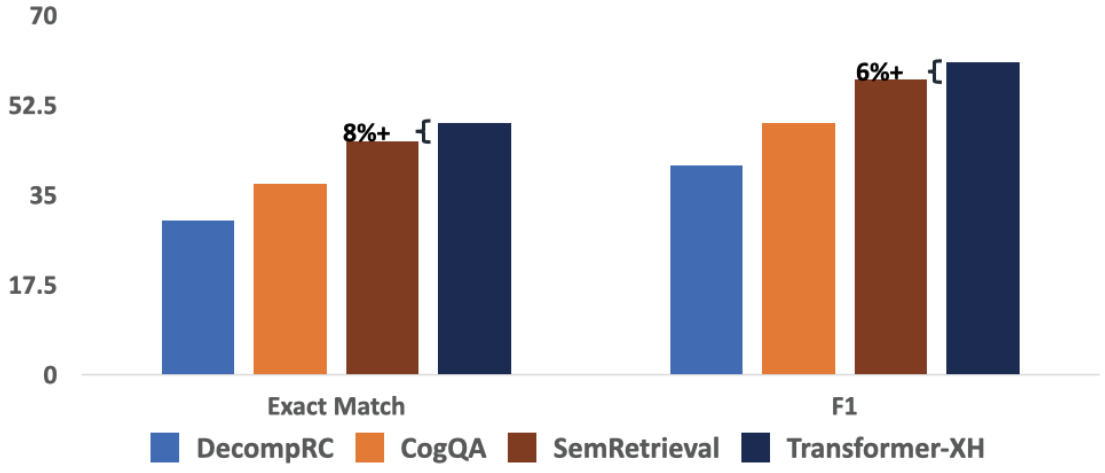


图 12：在 Hotpot QA 数据集上的实验结果

相较于使用纯粹的基于 GNN（将每个文档的嵌入输入到 GNN 中）或 Free-text（进行精准化的文本检索，将检索结果连接在一起，输入到 BERT、RoBerta 等模型中进行阅读理解）的模型，我们提出的 Transformer-XH 是一种更为简单的模型，它将原始的半结构化数据作为输入，从而得到目标表征，Transformer-XH 最终的性能效果往往更优。

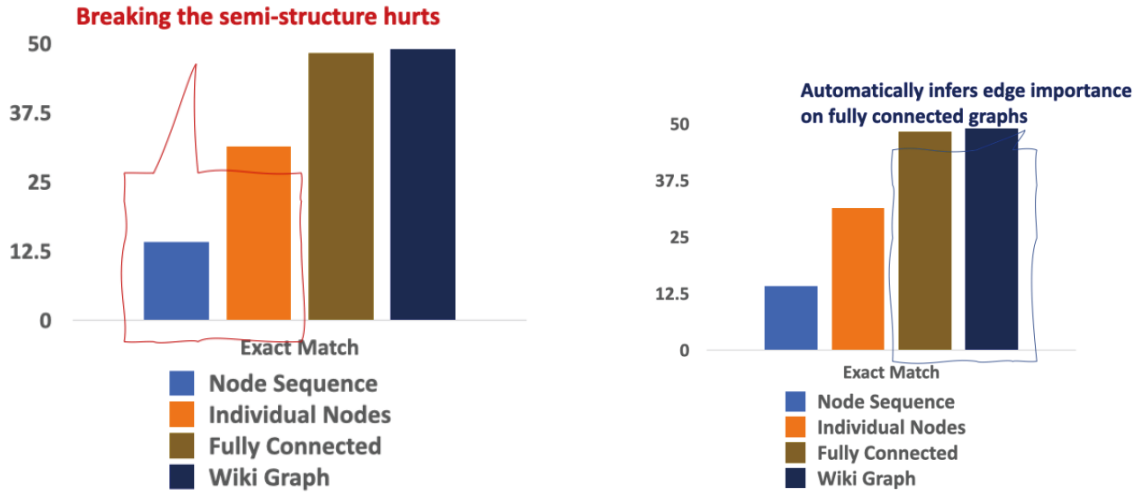


图 13：结构对模型性能的影响；使用不同文档图结构时的性能对比

如上图所示，相较于我们提出的模型（深蓝色），当我们使用独立的节点（橙色）时，模型性能几乎减半。当我们使用错误的结构（浅蓝色，如链式关系）时，性能相较于使用独立节点的情况甚至更差。这说明，正确的结构信息缺失对于提升模型性能很有帮助。

此外，对于 Hotpot QA 数据集而言，由于其保证在写问题时的两个真实段落之间必定存在维基百科中的链接，

因此维基百科中的信息对其性能的影响十分巨大。但是，当使用我们的 Transformer-XH 时，我们甚至可以不使用维基百科的信息，即使使用一个全联通的图进行学习就可以得到与使用维基百科时相近的结果。因此，我们的方法可以自动推断出边的重要性。

在图神经网络中，每一层中的节点可以聚合其邻居节点的信息（一跳），而网络的层数决定了每个节点可以聚合信息的范围。实验结果表明，当我们使用三层 Transformer 的注意力机制时，模型性能最优。这是因为 Hotpot QA 大多数是 2 跳的，当网络层数继续上升时，往往难以进行学习。

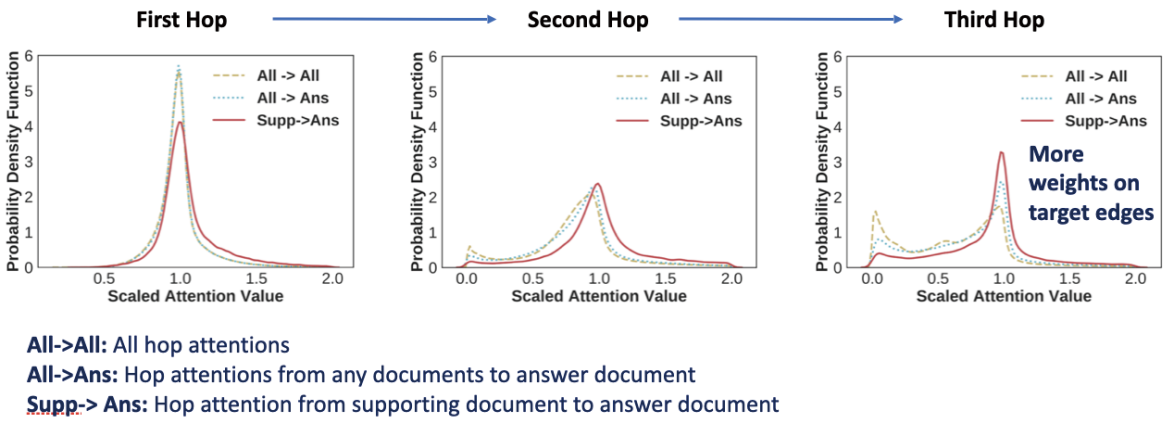


图 14：全联通状态下不同节点之间的注意力权重

在这里，我们考虑三种节点之间的连边：（1）任意两个文档之间的边（2）所有节点与实际包含答案的节点相连的边（3）包含关键信息的节点到包含答案的节点之间的边。如上图所示，当使用一跳信息时，注意力权值的概率密度分布近似为均值为 1 的较为集中的「瘦高」高斯分布；当使用两跳信息时，概率密度分布变得较为「矮胖」，模型可以将边的重要性区分开来。由于我们希望 Transformer-XH 能够从文章中提取出有助于确定答案区间的信息，所以我们最关注红色的实线。而在使用三跳信息时，红色实线的值相较于另外两种边确实越来越大，说明 Transformer-XH 确实能够学习到推理的关系。

四、结语

综上所述，对于知识图谱、语义信息的处理，以及一些现实中的信息检索任务都可以看做是一种针对半结构化数据的问题。「如何结合结构化符号知识与纯文本信息」这一问题仍然有很大的研究空间。

我们可以用 Transformer、BERT 等模型对半结构化数据进行建模和表征（在文本节点内部运用 BERT 的思想，在文本节点之间运用 GAT 的思想），针对具体的任务设计后续的网络层。

中科大教授何向南：图神经网络在推荐系统的前沿研究

整理：智源社区 高洛生

推荐系统中用户和物品之间的交互、物品知识图谱、用户社交网络等数据可以天然地表示成图。受益于图神经网络在图数据上进行表示学习的优势，近年来图神经网络推动了推荐系统技术的发展。近期，在第二届智源大会的“图神经网络专题论坛”中，何向南教授针对这一进展做了主题为“图神经网络在推荐系统的前沿研究”的报告。

何向南教授首先介绍了推荐系统的背景，以及将图神经网络和推荐系统结合在一起的初衷； 随后，他从建模的角度介绍了他们团队在图神经网络方面所做的一些工作，并就如何从学习的角度对推荐场景下的图卷积网络进行更好优化进行了阐述。

何向南，中国科学技术大学教授、博士生导师，国家青年千人计划学者。主要研究方向： 推荐系统、数据挖掘。在 CCF A 类会议和期刊发表论文 80 余篇，谷歌学术引用 5300 余次，包括国际顶级会议 SIGIR、WWW、KDD 和顶级期刊 TKDE、TOIS、TNNLS 等； 长期担任这些会议和期刊的审稿人，CCIS 2019 的程序委员会主席，AI Open 期刊编委。研究成果曾获 SIGIR 2016 最佳论文提名奖、WWW 2018 最佳论文提名奖等，主持国家基金委面上项目 1 项，重点项目 1 项。

一、推荐系统与图神经网络的渊源

如今人类正处于一个信息大爆炸的时代，每天都有海量的信息等待人们挖掘，从中找到有价值或感兴趣的信息。例如，微博每天有超过五亿条博文，Flickr 每天有三亿张图片，快手每天上传超过两千万个短视频。这就导致，最近几年在信息检索和挖掘领域，推荐的热度已经超过搜索； 在各类学术会议中，推荐方向的文章数量已经是搜索方向的两倍以上。

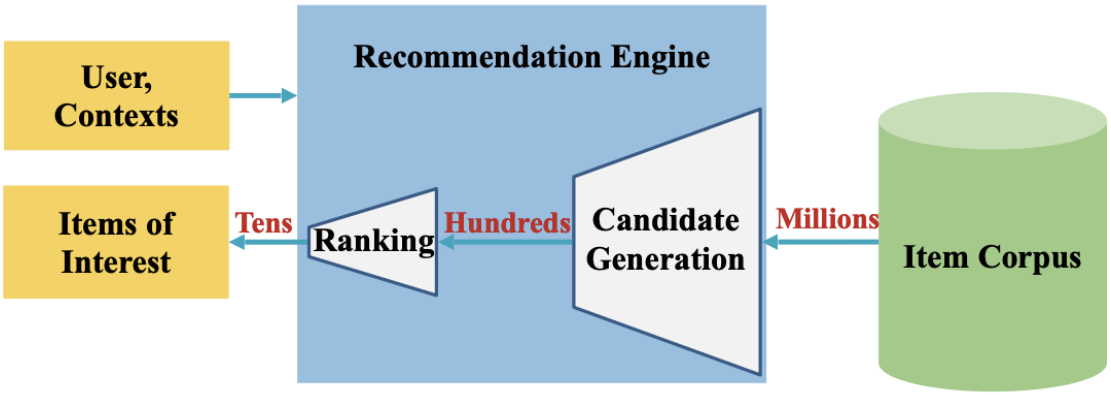


图 1： 推荐系统示意图

推荐系统最基本的技术是协同过滤技术。协同过滤技术指，若给定一个用户与物品的历史交互，则希望可以预测在未来该用户与商品产生交互的可能性。其基本假设是基于相似的用户可能会对相似物品产生相似的喜好。

早期的协同过滤方法主要是通过手算相似度，然后对相似度进行融合。最近十年，推荐的主流方法则是在隐空间中学习用户的表示，然后在隐空间衡量相似度。

早期工作在对用户或者物品进行表示学习时，使用的是用户 ID 或者物品 ID，建模单个用户和物品之间的交互。然而，研究人员逐渐发现，用一个用户的历史数据作为该用户的特征或将物品与历史用户产生的交互作为该物品的特征，然后进行表示学习，也会得到不错的效果。这相当于在图数据上，早期的工作是考虑一个单独的节点进行表示学习，现在则是将其邻居节点也考虑进来。

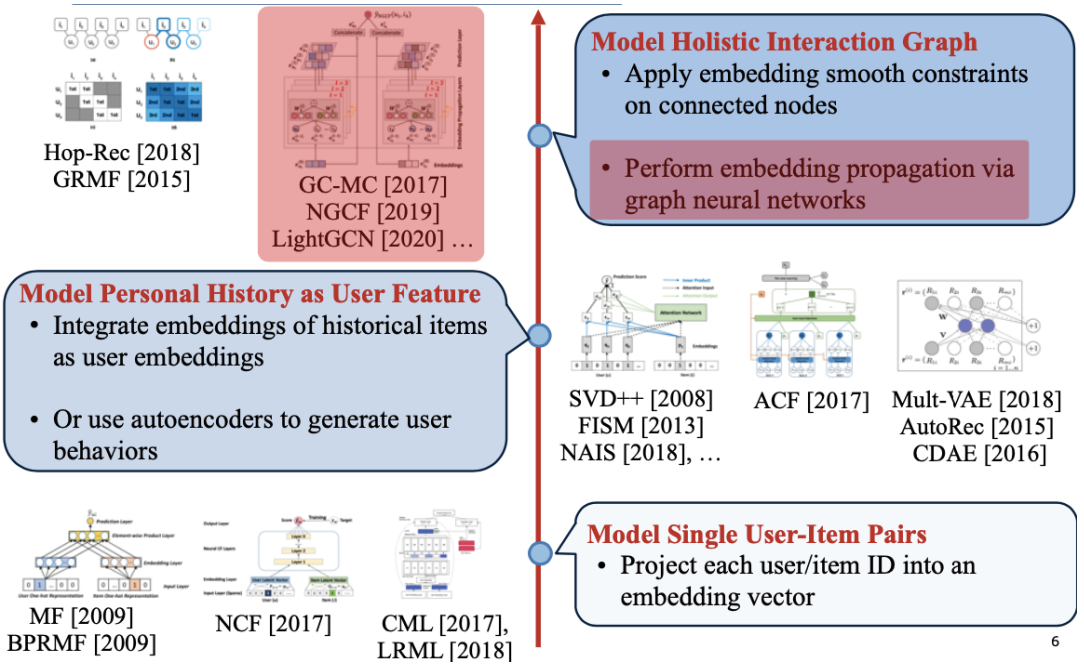


图 2：协同过滤技术的演化

二、从建模的角度看图神经网络在推荐场景下的协同过滤

2.1 神经网络协同过滤技术 NGCF

NGCF 的主要想法是将用户与物品之间的交互建模成二分图，并在该二分图上通过高阶连接性建模协同过滤的信号。高阶连接性的意思是指对于节点 u1，任何可以到达的其他节点，其路径长度需要大于 1。以往大部分的协同过滤方法都没有考虑高阶连接性，仅通过用户 ID 和历史记录来表示用户，这导致并没有显示的表明用户与物品之间的高阶连接性。NGCF 的工作则可以在模型层面显示的建模用户和物品的高阶连接性，从而学习到更好的用户和物品的表示。

具体而言，NGCF 直接借鉴了 GCN 的经典设计。首先，定义用户和物品之间的消息，两个节点之间的相似度越高，它们之间扩散的消息就越多； 其次，考虑两个待扩散节点之间的相似度。定义好单个节点的消息后，再将其邻节点和自身的消息累加，从而得到下一层的节点表示。通过多层的表示便可以获得高阶连接性的建模。当获得每一层都有扩散之后，将各个层拼接起来便可以捕获更全面的信号，从而方便预测和推荐。通过这种方式，可以在扩散的过程中将高阶的协同过滤信号建模出来。

When:

- Organize historical interactions as a user-item bipartite graph
- Capture CF signal via high-order connectivity
  - Definition: the paths that reach u_1 from any node with the path length l larger than 1.

Most CF methods (e.g., MF, FISM, NAIS, AutoRec) fail to model high-order connectivity explicitly.
- Embedding function only considers descriptive features (e.g., ID, attributes)
- User-item interactions are not considered

NGCF's contribution: modeling CF with high-order connectivity via GNN.

- We can revisit CF via high-order connectivity

图 3：卷积网络协同过滤（NGCF）

图 4 展示的是一个十分有意思的可视化结果。(a) 图是将用户和物品投影到二维平面的可视化结果，每一个五角星代表一个用户，每一个圆圈代表一个物品，如果两者是同一个颜色，则表示该用户与该物品产生过交互，对应图数据的一阶邻居关系。(b) 图展示的是三层的 NGCF 用户和物品的可视化结果。不难看出，右边的结果明显比左边的结果更能清楚的看出聚类结果，尤其是每一种颜色都形成了一个边界相对清晰的聚类。由可视化结果可知，通过 NGCF 做高阶连接的建模可以得到更高质量的表示学习结果。

- Powerful Representation Ability

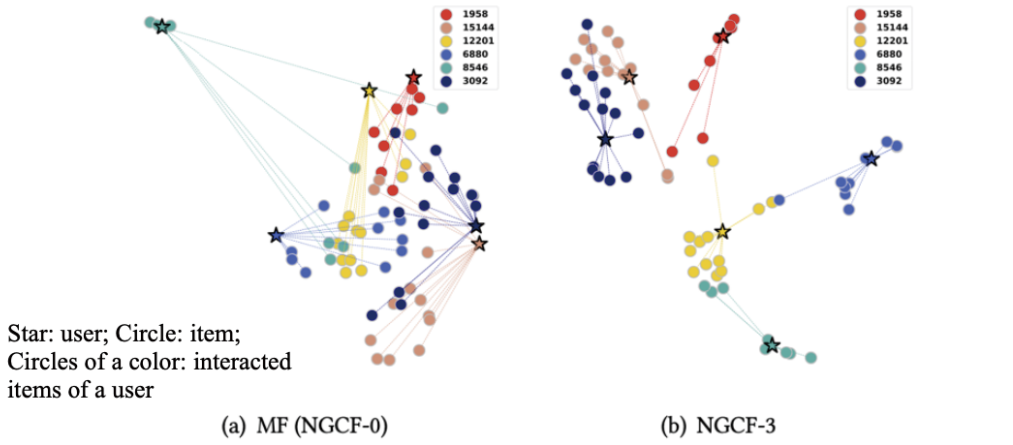


图 4：NGCF 可视化结果对比

2.2 轻型图卷积网络 LightGCN

然而，NGCF 在很多数据集上很难进行训练。何向南等人通过对比传统 GCNs 和 NGCF（对比结果如图 5 所示）发现，在推荐系统场景下，NGCF 的诸多设计是冗余的； 此外，NGCF 之所以难训练，以及最后的效果不尽如

人意，主要原因是加入了很多非必要的神经网络的操作符。

- Designs of NGCF are rather heavy and burdensome
  - Many operations are directly inherited from GCN without justification.

	GNNs	NGCF
Original task	Node classification	Collaborative filtering
Input data	Rich node features • Attributes, text, image data	Only node ID • One-hot encoding
Feature transformation	Distill useful information	Generate ID embeddings
Neighborhood aggregation	Pass messages from neighbors to the egos	Pass messages from neighbors to the egos
Nonlinear activation	Enhance representation ability	Negatively increases the difficulty for model training

图 5：传统 GCNs 和 NGCF 对比结果

基于此考虑，何向南团队提出了 LightGCN 方法。该方法的核心是轻型的图卷积。轻型图卷积是指，在每次进行图卷积时，只考虑一个节点的邻居节点，然后将这些邻居节点的上一层表示信息加权在一起。与之前相对冗余的 NGCF 相比，LightGCN 相当于删除了 NGCF 里面的激活函数，同时也将 GCN 中常见的自环操作丢掉了。NGCF 获得每一层表示以后会将结果拼接起来，然而这会导致向量长度变长，最终模型需要更长的时间进行向量乘法运算，LightGCN 则不会。因此，与 NGCF 相比，LightGCN 具备无功能转换、无非线性激活、无自连接等优点。

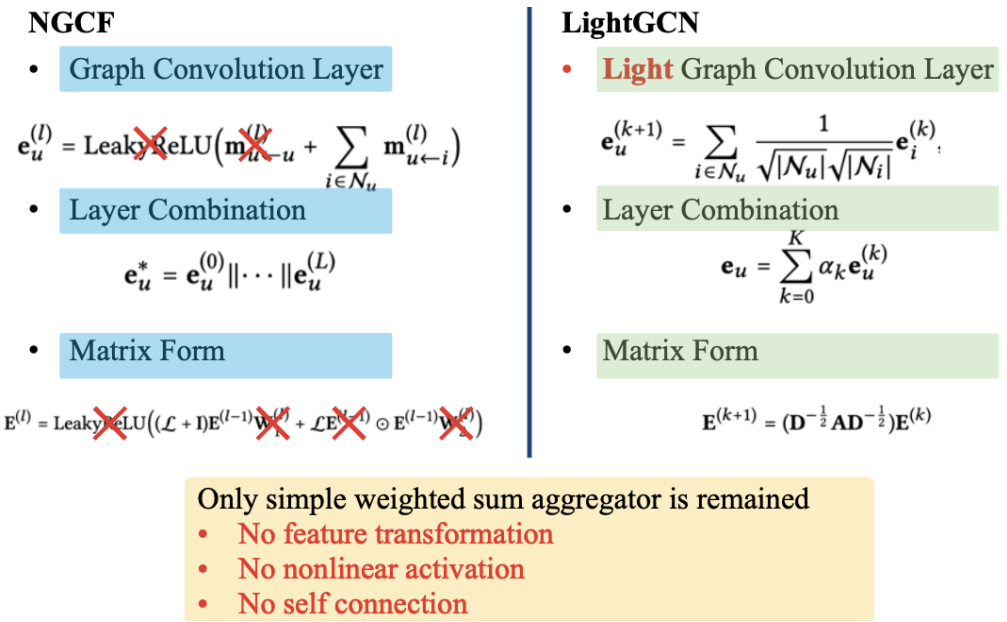


图 6：NGCF 与 LightGCN 对比

2.3 解耦图协同过滤 GDCF

通过研究发现，用户在选择物品时会抱有不同的意图，这会激发出不同的用户行为。因此，在用户表示学习时，如何将不同维度下交互背后隐含的用户意图刻画出来是关键所在。

基于此考虑，何向南团队提出了解耦图协同过滤方法 GDCF。具体的做法是:

首先，提出图解耦层（Graph Disentangling Layer）的概念。假设具有潜在意图，并重组特定于 K 的交互图（K 表示共有 K 种意图），对每一种意图建立用户 / 物品图，这样就得到了 K 个图，每个图都是针对某个意图。

然后，在该图上进行动态邻居聚合。这样做的目的是自适应地测量用户项连接的特定意图强度。
最后，使用“距离相关”来规范特定于意图的表示独立建模。他们希望通过正则化表示学习，同时考虑用户两两意图之间的关联关系，尽量让两两意图之间的关联性为零，从而最大化表达的信息量。

DGCF 模型如图 7 所示。

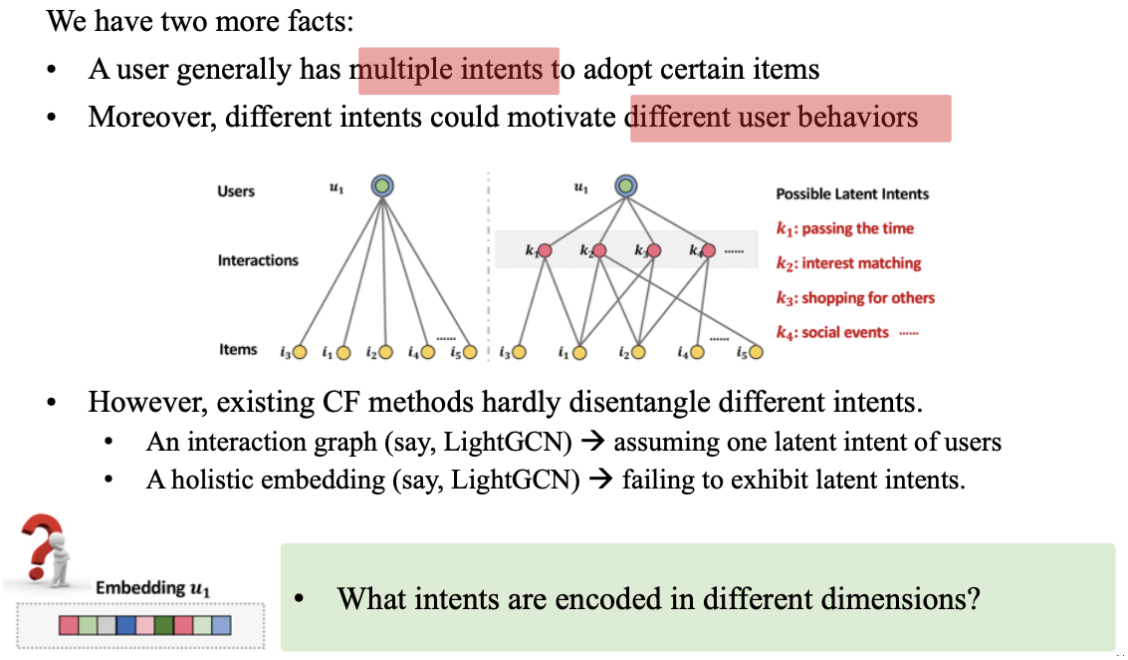


图 7：解耦图协同过滤 GDCF

通过对比实验（如图 8 所示）可知，在相同的数据集上，GDCF 与 NGCF 相比可以学习到更好的表示，在表示效果上提升了 15% 左右； 与 Light GCN 相比，DGCF 与 Light GCN 处于相同的水平，没有显著性的差异，但是将两者投射到隐空间，GDCF 会有更好的解释性。

- In recommendation accuracy, DGCF betters NGCF → **better representations**

	Gowalla		Yelp2018*		Amazon-Book	
	recall	ndcg	recall	ndcg	recall	ndcg
MF	0.1291	0.1109	0.0433	0.0354	0.0250	0.0196
GC-MC	0.1395	0.1204	0.0462	0.0379	0.0288	0.0224
NGCF	0.1569	0.1327	0.0579	0.0477	0.0337	0.0266
DisenGCN	0.1356	0.1174	0.0558	0.0454	0.0329	0.0254
MacridVAE	0.1618	0.1202	0.0612	0.0495	0.0383	0.0295
DGCF-1	**0.1794***	**0.1521***	**0.0640***	**0.0522***	**0.0399***	**0.0308***
%improv.	10.88%	14.62%	4.58%	5.45%	4.17%	4.41%
p-value	6.63e-8	3.10e-7	1.75e-8	4.45e-9	8.26e-5	7.15e-5

- DGCF vs. LightGCN: same performance level, but **better interpretability of representations**.

图 8：GDCF 实验结果对比

以上三项工作侧重于 GNN 建模的层面。早期的 NGCF 为后续研究提供了基础，Light GCN 又对 NGCF 做了大量的简化，有更好的效果，可以让表示学习的效果具有更强的解释性。

三、从学习的角度看图卷积网络在推荐场景下的优化

上文主要是从建模的角度看图神经网络在推荐场景下的协同过滤技术。在这一节，将着重介绍何向南团队在推荐场景中通过图卷积网络所做的工作。

3.1 Fast Loss

通过实验分析发现，在计算 BPR Loss 时，使用现有的主流深度学习框架 TensorFlow 在 GPU 上的计算速度远不及使用 C++ 在单个 CPU 上的计算速度。

	C++ (CPU)	TensorFlow (GPU)
time/epoch	1.1s	55s

图 9：在 CPU 上使用 C++ (i9 9000kf) 和 GPU 上使用 TensorFlow (2080Ti) 运行有 amazon-book 数据集的 BPRMF

为了解决该问题，何向南教授团队提出了 Fast Loss 方法。Fast Loss 计算过程如图 10 所示：

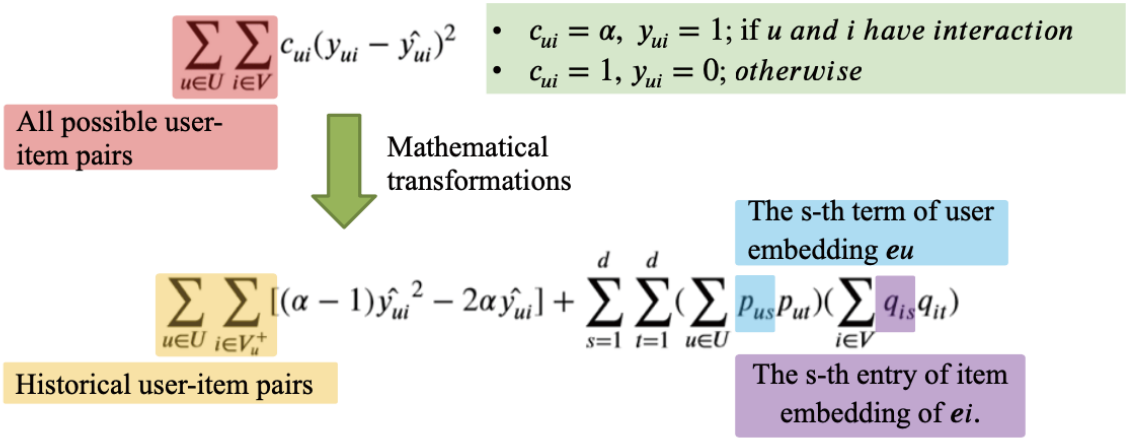


图 10：Fast Loss 计算过程

该方法时间复杂度为 O(|R|d+|N|d2)，并且与观察到的相互作用数呈线性关系。在训练模型时，无论用户是否与物品有历史交互，都将整个矩阵纳入训练。如果存在历史交互，则记为正样本，并给正样本比较高的权重α，再将其向 1 回归。对于负样本，给其一个比较低的权重并让其向 0 回归。

如果直接优化损失函数（Loss），其复杂度是惊人的，即 M × N。在实际应用中，百万用户和百万物品相乘后的结果是难以置信的。但是通过对损失函数进行数学上的等价变换，可以得到其变型形式。通过数学上非常聪明的等价变化，可以将原始损失函数的复杂度，从 M × N 降低到了 R，即降低到观察到正样本的个数，使得总体的复杂度可以被认为是和观察到的样本呈线性关系。

经过快速损失优化的 LightGCN 可以达到与 BPR 损失相当的性能。当然，哪种损失更好取决于数据特征，但是 Fast Loss 在长尾用户 / 项目上似乎更好。

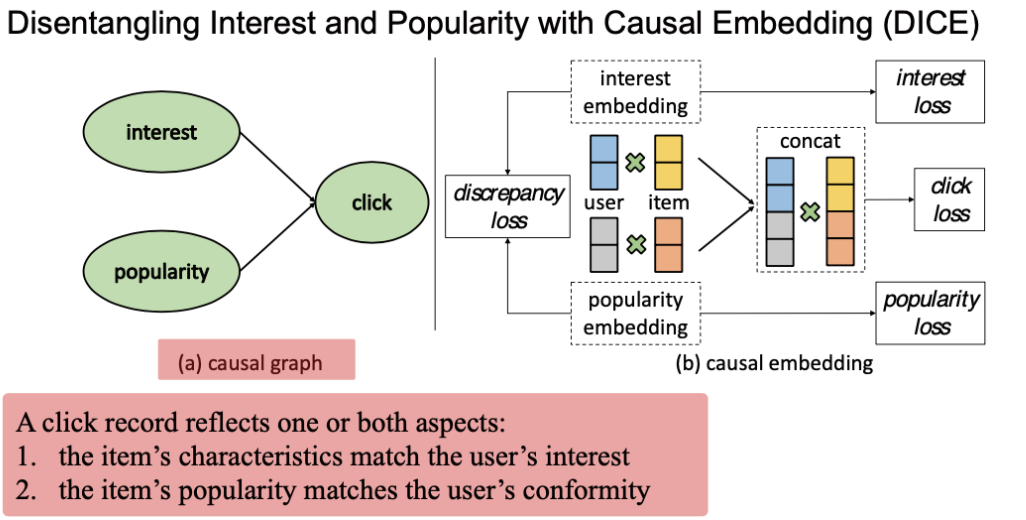
3.2 解耦用户兴趣和物品受欢迎程度 DICE

在推荐系统中，记录的数据往往表现出各种偏差，例如人气偏差。这是由于模型在历史数据中进行训练，模型训练完成后将推荐结果展示给用户，用户在推荐的历史上产生新的数据，随后又将新的数据训练模型，如此循环往复形成闭环，这个闭环导致越是流行的物品就变得越流行，曝光度增高，用户点击就越多。这种偏差很大程度上阻碍了模型发现用户真正的兴趣。

现有的方法一般归为两类：IPS 和 CausE。这两类方法在没有合理假设的情况下，仅根据点击数据来区分兴趣和受欢迎程度，是非常难的。其次，利用它们来学习推荐的解纠缠表示比较困难，因为根本没有用户兴趣的 ground-truth。此外，利用它们进行推荐，在合并或区分 cause 时，需要进行细致的设计。

基于此，何向南团队提出了 DICE 方法。该方法首先建立一个因果图，然后何向南教授认为一个点击行为背后

要么是该用户的属性和用户的兴趣相匹配；要么就是该物品的流行度和该用户当前的消费态度相匹配。基于此，分别基于兴趣和流行度分别计算概率并叠加在一起即可。该方法优化的目标是希望获得解耦的表示，以至于每次嵌入仅捕获一个原因。



Assumption 2.1 *A click record in biased recommendation scenario results from two independent causes, which are the user's interest and the item's popularity.*

$$P_{\text{click}} = P_{\text{interest}} + P_{\text{popularity}}, \tag{1}$$

图 10：解耦用户兴趣和物品受欢迎程度 DICE 方法

在演讲的最后，何向南教授还介绍了其团队未来将在模型的有效性、推荐效果的鲁棒性和对话式推荐系统等方面继续探索，力争取得更好的成果。