图神经网络

顾一凡　张文生

中国科学院自动化研究所

中国科学院大学人工智能学院

2020-01-02

讲课提纲

一、图神经网络发展

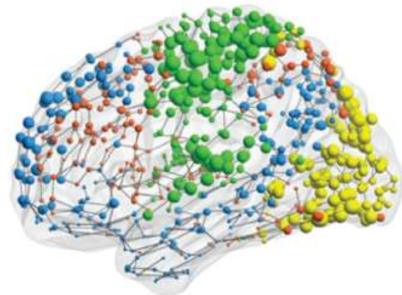
二、典型图神经网络

三、图神经网络应用

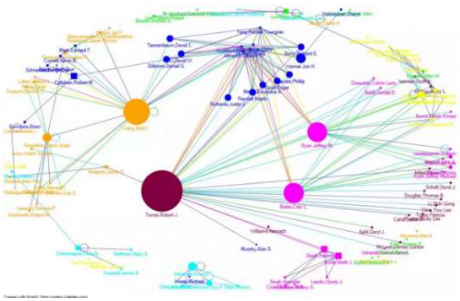
为什么要开展图神经网络研究

Social networks

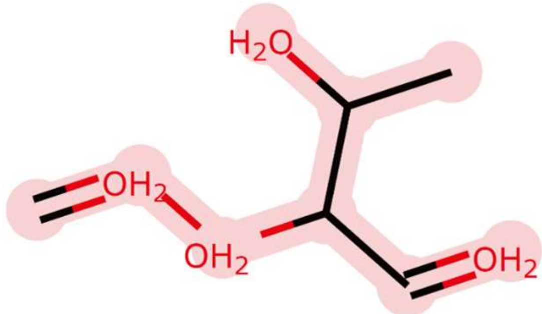
Traffic networks



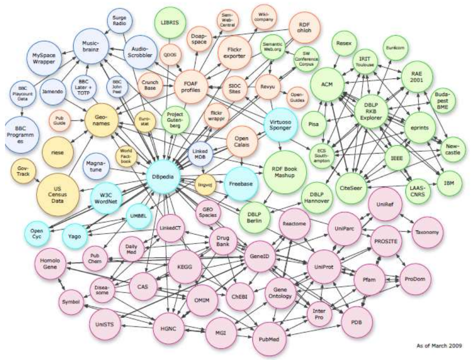
Functional networks



Citation Networks



Chemical Compound



Knowledge Graph

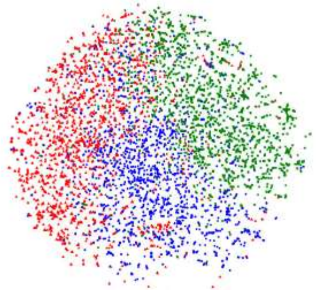
如何开展图神经网络研究

网络表示学习

G = (V, E)

模型:
DeepWalk,
Node2Ved,
LINE,SDNE…

G = (V)
Vector Space

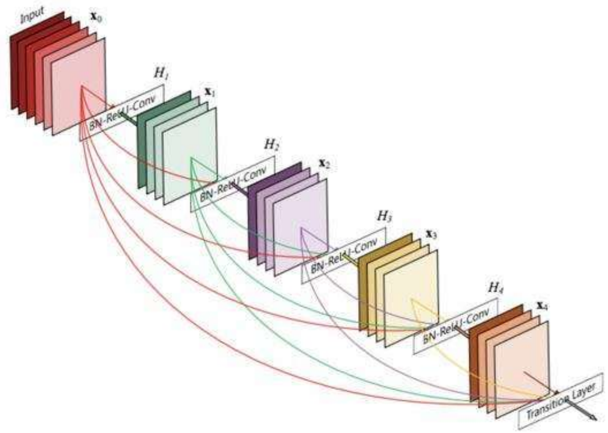


Network Inference

- 节点分类
- 连接预测
- 社区发现
- 可视化
- 相似度量
- ……

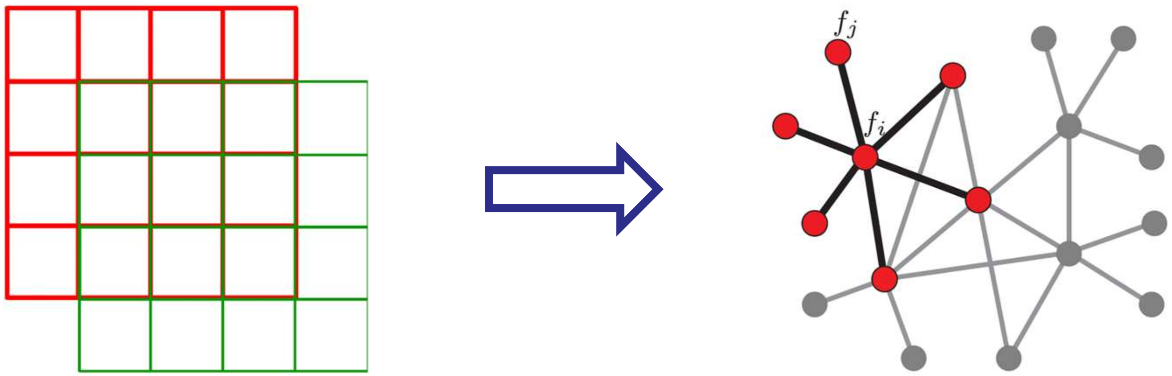
缺陷：计算复杂度高，难以在大规模图数据上进行优化。

卷积神经网络



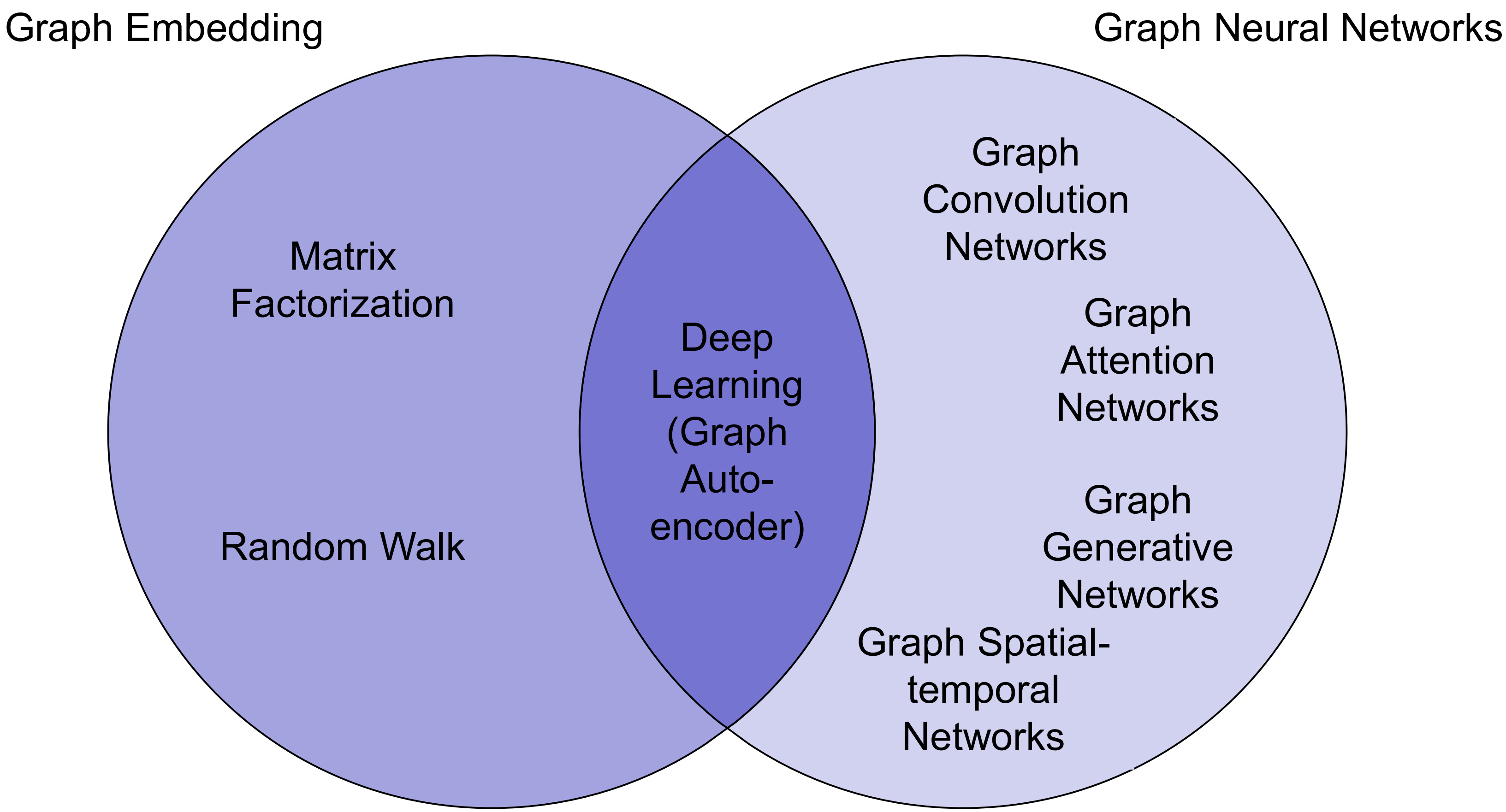
特点
- 局部连接
- 权重共享
- 多层结构

缺陷：只能处理规则的欧式数据，例如图像（2维网格）、文本（1维序列）。



中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

图神经网络VS图嵌入



Wu Z , Pan S , Chen F , et al. A Comprehensive Survey on Graph Neural Networks[J]. 2019.

讲课提纲

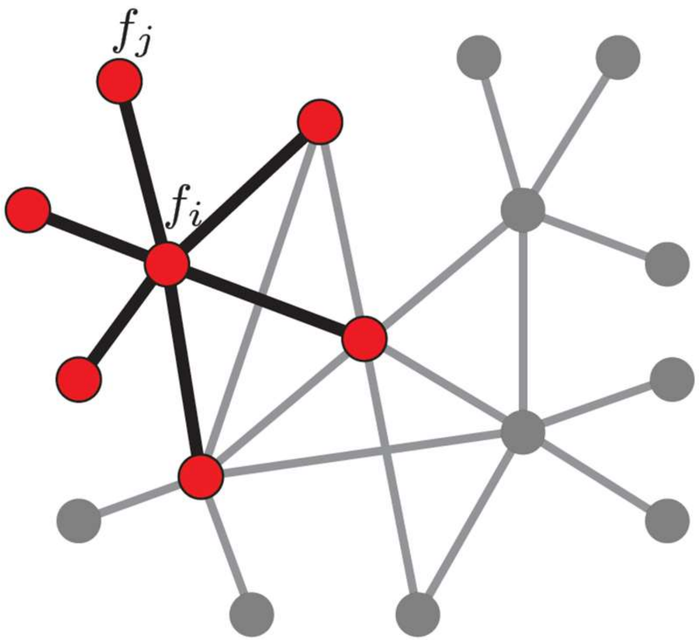
一、图神经网络发展

二、典型图神经网络

三、图神经网络应用

1、图卷积网络

Graph theory concepts



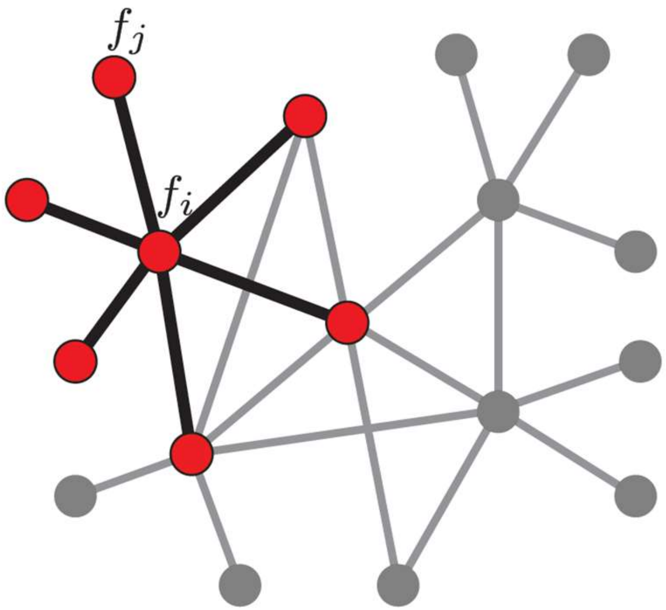
Weighted undirected graph $\mathcal{G}$

Edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ Vertices $\mathcal{V} = \{1, \dots, n\}$

Edges weights $w_{ij} \geq 0$ for $(i, j) \in \mathcal{E}$

Functions over the vertices
$L^2(\mathcal{V}) = \{f: \mathcal{V} \to \mathbb{R}\}$ represented as vectors
$\mathbf{f} = (f_1, \dots, f_n)$

Hilbert space with inner product
$< f, g >_{L^2(\mathcal{V})} = \sum_{i \in \mathcal{V}} f_i g_i = \mathbf{f}^{\mathrm{T}} \mathbf{g}$



Unnormalized Laplacian Δ:

$$(\Delta f)_i = \sum_{j:(i,j)\in\mathcal{E}} w_{ij}\left(f_i - f_j\right)$$

$$= f_i \sum_{j:(i,j)\in\mathcal{E}} w_{ij} - \sum_{j:(i,j)\in\mathcal{E}} w_{ij} f_j$$

difference between f and its local average

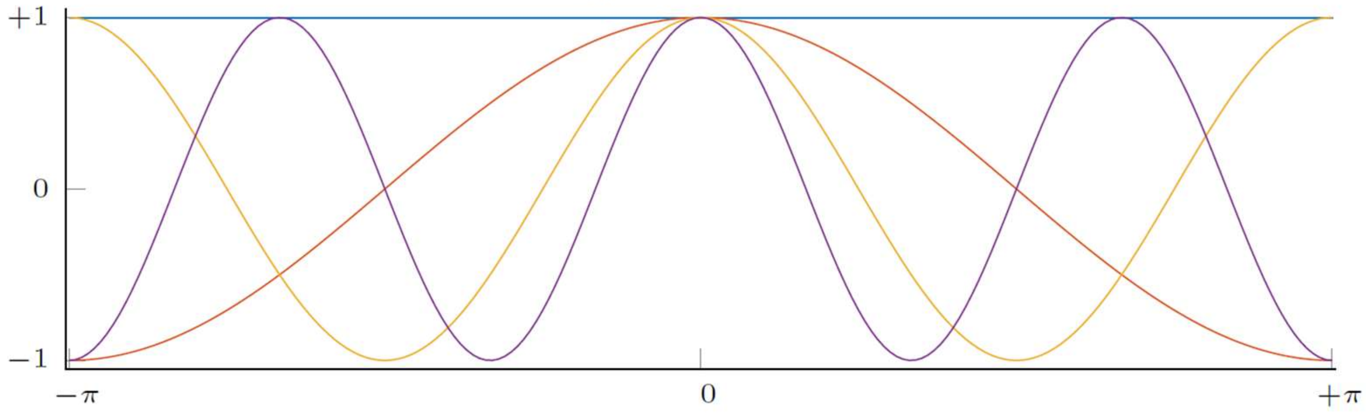
Represented as a positive semi-definite $n \times n$ matrix

$$\Delta = \mathbf{D} - \mathbf{W} \qquad \mathbf{W} = \left(w_{ij}\right) \qquad \mathbf{D} = \mathrm{diag}\left(\textstyle\sum_{j\neq i} w_{ij}\right)$$

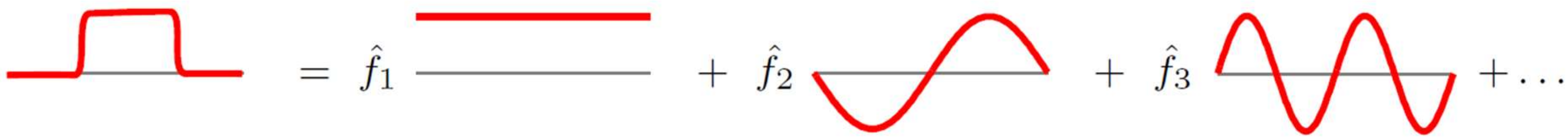
Fourier analysis on Euclidean spaces

A function $f\colon [-\pi,\pi] \to \mathbb{R}$ can be written as a Fourier series

$$f(x) = \sum_{k \geq 0} \underbrace{\langle f, e^{ikx} \rangle_{L^2([-\pi,\pi])}}_{\hat{f}_k \text{ Fourier coefficient}} e^{ikx}$$



Eigenfunctions of 1D Euclidean Laplacian



Spectral convolution

Convolution theorem: Fourier transform diagonalizes the convolution operator $\Longrightarrow$ convolution can be computed in the Fourier domain as

$$\widehat{(f * g)} = \hat{f} \cdot \hat{g}$$

Spectral convolution of $\mathrm{f}, \mathrm{g} \in: L^2(\mathcal{V})$ can be defined by analogy

$$\mathbf{f} \star \mathbf{g} = \underbrace{\sum_{k \geq 1} \underbrace{\langle \mathbf{f}, \phi_k \rangle_{L^2(\mathcal{V})} \langle \mathbf{g}, \phi_k \rangle_{L^2(\mathcal{V})}}_{\text{product in the Fourier domain}} \phi_k}_{\text{inverse Fourier transform}}$$

In matrix-vector notation

$$\begin{aligned} f * g &= \Phi(\Phi^T g) \circ (\Phi^T f) \\ &= \Phi \underline{\mathrm{diag}(\hat{g}_1, \dots, \hat{g}_n)} \Phi^T f \end{aligned}$$

Spectral filter coefficients

ChebNet

Represent spectral transfer function as a polynomial of order K

$$\tau_\theta(\Lambda) = \sum_{k=0}^{K-1} \theta_k \Lambda^k$$

- O(1) parameters per layer (K parameters)

$$y = g_\theta(L)f = \sum_{k=0}^{K-1} \theta_k T_k(\tilde{L})f \qquad \theta \in \mathbb{R}^K \qquad T_k(\tilde{L}) \in \mathbb{R}^{n\times n}$$

$$\tilde{L} = \frac{2L}{\lambda_{max}} - I_n \qquad \bar{f}_k = T_k(\tilde{L})f \in \mathbb{R}^n \qquad \bar{f}_k = 2\tilde{L}\bar{f}_{k-1} - \bar{f}_{k-2}, \bar{f}_0 = f, \bar{f}_1 = \tilde{L}f$$

$$y = g_\theta(L)f = [\bar{f}_0, \dots, \bar{f}_{k-1}]\theta$$

- Filters have guaranteed K-hops support
- No explicit computation of forward and inverse Fourier transforms $\Longrightarrow O(K|\mathcal{E}|)$

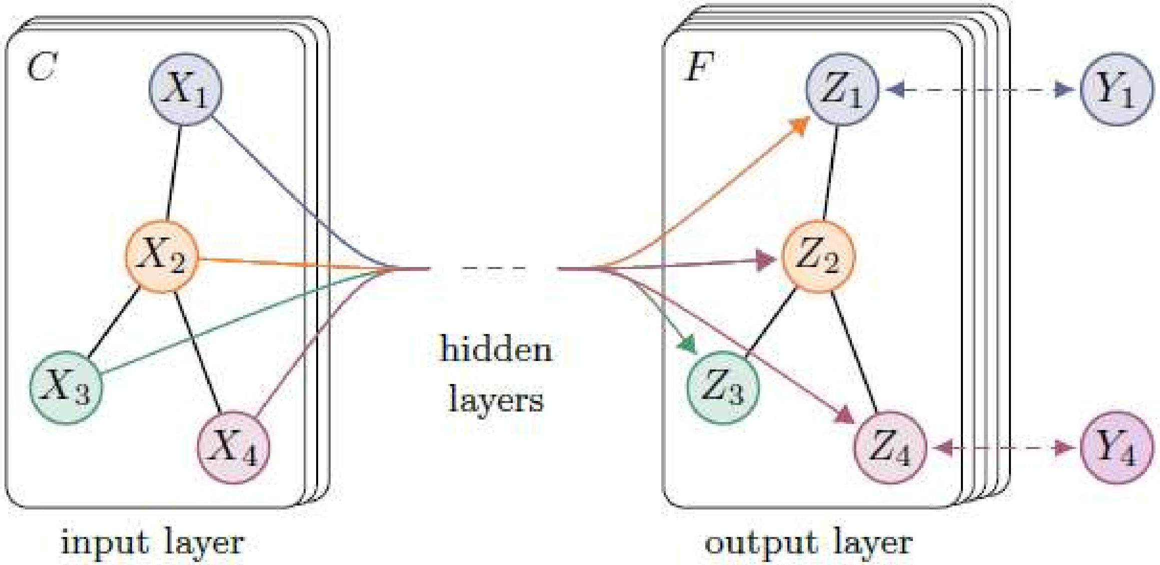
Defferrard M, Bresson X, Vandergheynst P. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. *NIPS 2016*.

Graph Convolution networks (GCN)

Linear approximation: $\lambda_{max} \approx 2, K = 1$

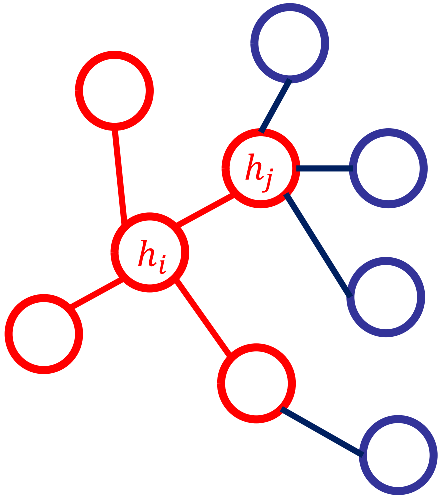
$g * f \approx \theta_0 f + \theta_1 (L - I_n) f = \theta_0 f - \theta_1 D^{-1/2} A D^{-1/2} f$ 2 params per layer

$\theta_0 = -\theta_1 = \theta \quad g * f \approx \theta\left(I_n + D^{-1/2} A D^{-1/2}\right) f$ 1 param per layer



Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[C]. *ICLR 2017*

2、图注意力网络



Node features $\mathbf{h} = \{\vec{h}_1, \vec{h}_2, \dots, \vec{h}_N\}, \vec{h}_i \in \mathbb{R}^F$

Linear transformation $\mathbf{h}' = \mathbf{W}\mathbf{h} = \{\vec{h}_1, \vec{h}_2, \dots, \vec{h}_N\}, \vec{h}_i \in \mathbb{R}^{F'}$

Attention coefficients $e_{ij} = a(\mathbf{W}\vec{h}_i, \mathbf{W}\vec{h}_j)$

Normalized coefficients $\alpha_{ij} = \mathrm{softmax}_j(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k\in\mathcal{N}_i}\exp(e_{ik})}$

$$\alpha_{ij} = \frac{\exp\left(\mathrm{LeakyReLU}\left(\vec{a}^T\left[\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_j\right]\right)\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(\mathrm{LeakyReLU}\left(\vec{a}^T\left[\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_k\right]\right)\right)}$$

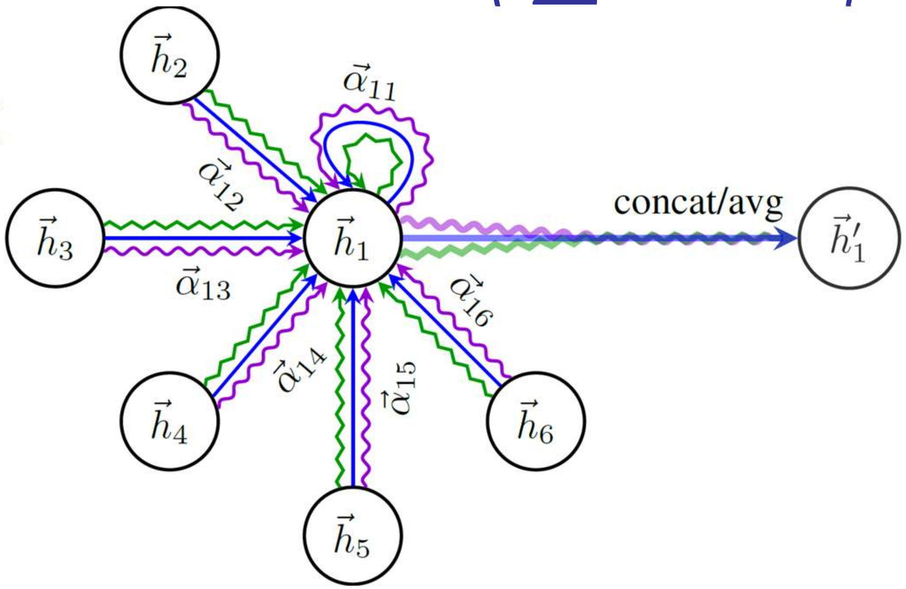
Velickovic P, Cucurull G, Casanova A, et al. Graph attention networks. *ICLR 2018*.

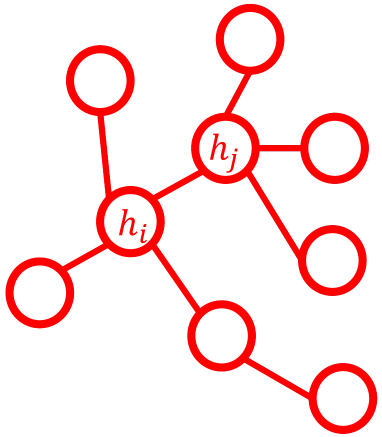
Attention layer output $\overrightarrow{h'}_i = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W} \vec{h}_j\right)$

Multi-head attention[1]

$$\overrightarrow{h'}_i = ||_{k=1}^{K} \sigma\left(\sum \alpha_{ij}^{k} \mathbf{W}^k \vec{h}_j\right)$$

Final (prediction



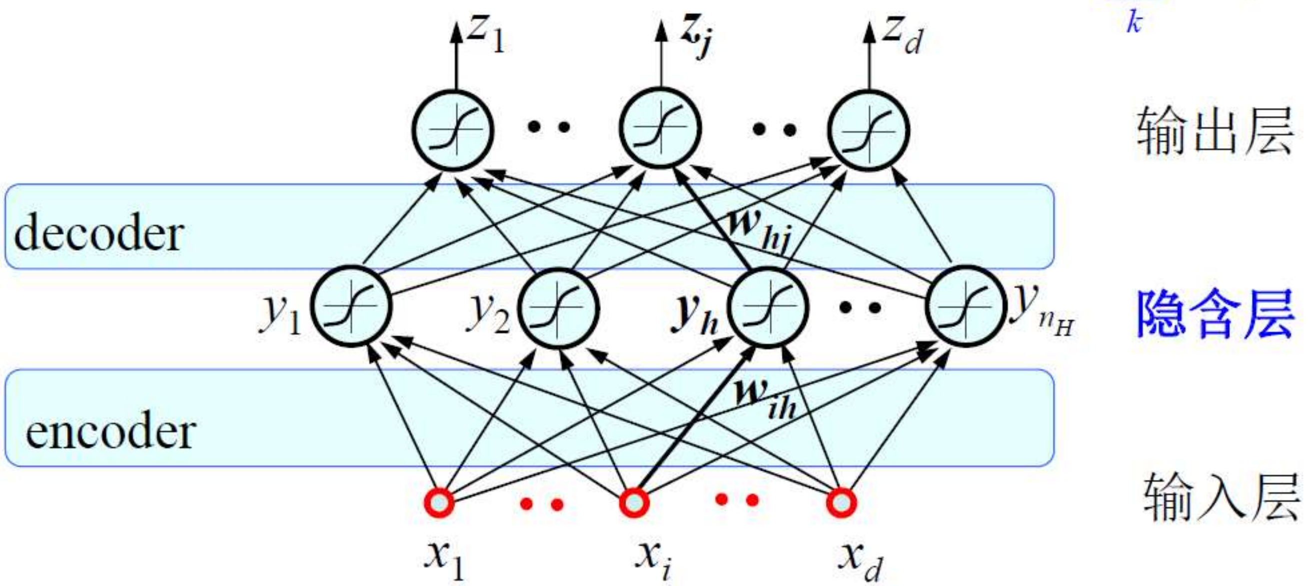


[1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *NIPS 2017*.

3、图自编码器

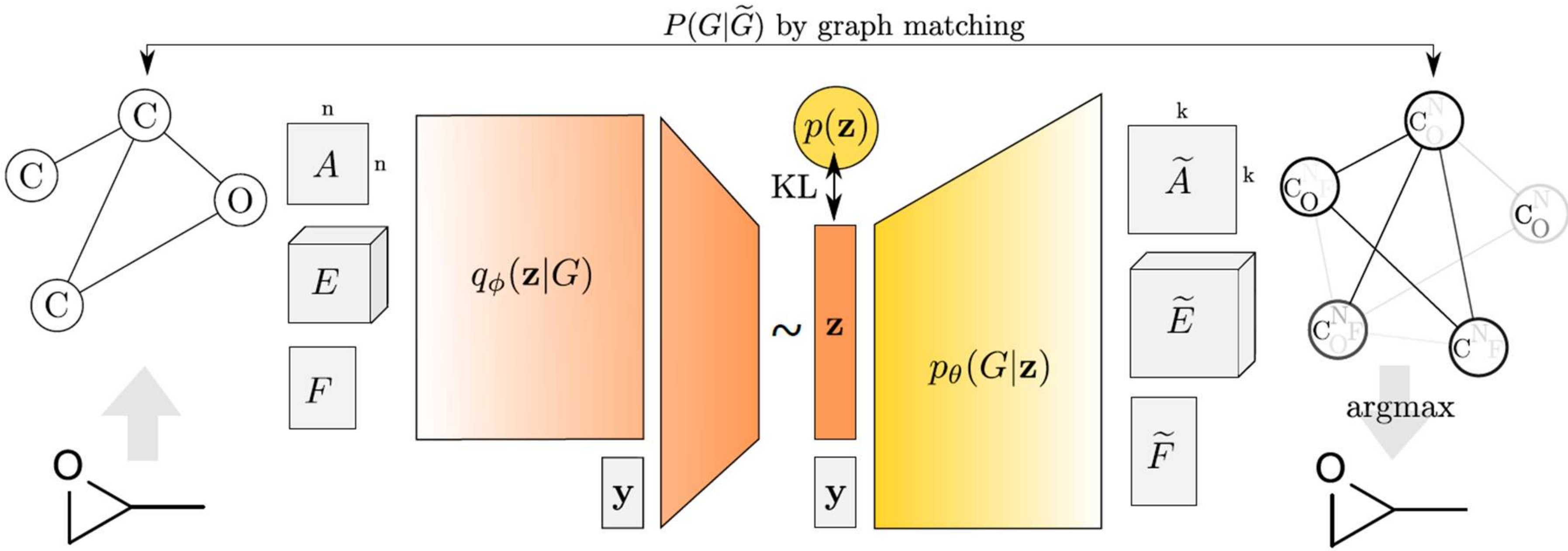
自编码器

$$\min \sum_{k} \| \mathbf{x}_k - \mathbf{z}_k \|^2$$



自动编码器是一种尽可能重构输入信号的神经网络

图编码器



G (A, E, F): Input graph. A: Adjacency matrix; E: Edge attribute tensor; F: Node attribute matrix

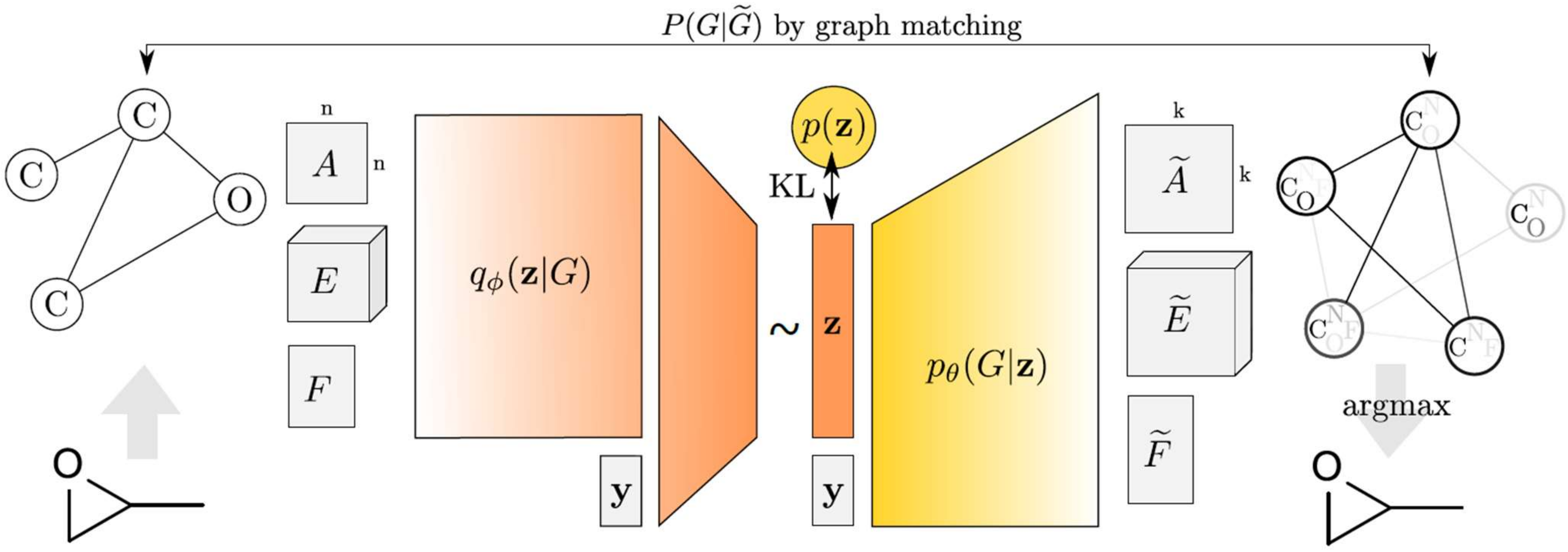
q_{ϕ}: Stochastic graph encoder

p_{θ}: Graph decoder

$\widetilde{G}$ ($\widetilde{A}$, $\widetilde{E}$, $\widetilde{F}$): Output graph

Simonovsky M, Komodakis N. Graphvae: Towards generation of small graphs using variational autoencoders[C]//International Conference on Artificial Neural Networks. Springer, Cham, 2018: 412-422.

图编码器



$$\log p(A'|\mathbf{z}) =$$

$$= 1/k \sum_a A'_{a,a} \log \widetilde{A}_{a,a} + (1 - A'_{a,a}) \log(1 - \widetilde{A}_{a,a}) +$$

$$+ 1/k(k-1) \sum_{a \neq b} A'_{a,b} \log \widetilde{A}_{a,b} + (1 - A'_{a,b}) \log(1 - \widetilde{A}_{a,b})$$

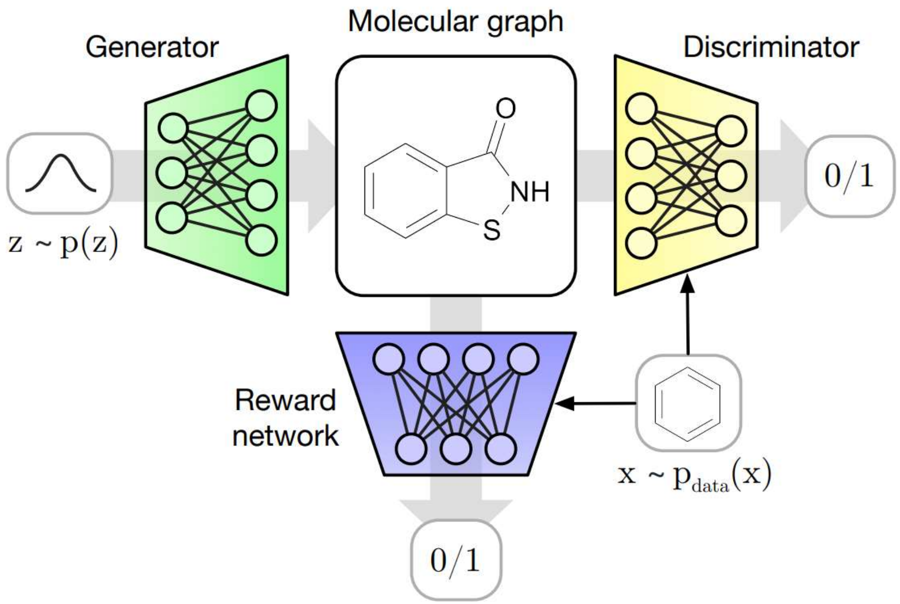
$$\log p(F|\mathbf{z}) = 1/n \sum_i \log F^T_{i,\cdot} \widetilde{F}'_{i,\cdot}$$

$$\log p(E|\mathbf{z}) = 1/(||A||_1 - n) \sum_{i \neq j} \log E^T_{i,j,\cdot} \widetilde{E}'_{i,j,\cdot}, \qquad (2)$$

$$-\log p(G|\mathbf{z}) = -\lambda_A \log p(A'|\mathbf{z}) - \lambda_F \log p(F|\mathbf{z}) - \lambda_E \log p(E|\mathbf{z}) \qquad (3)$$

4、图生成网络

MolGAN



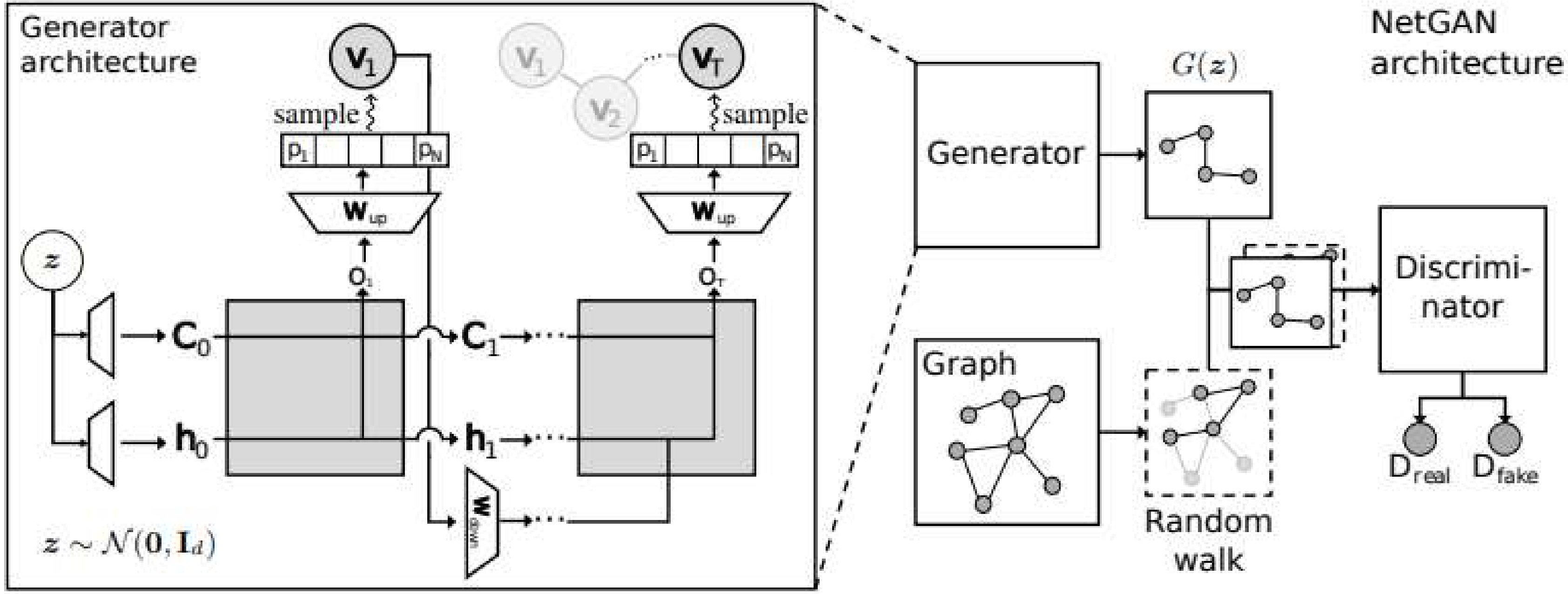
Generator: generating molecules from a prior distribution

Discriminator: distinguishing the generated samples and real samples

Reward network:
Learns to assign a reward to each molecule to match a score provided by an external software. Invalid molecules always receive zero rewards.

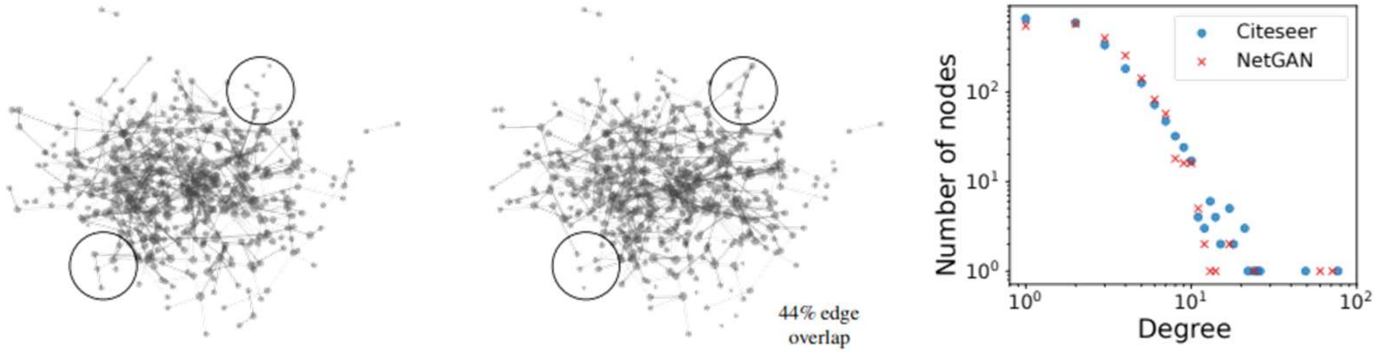
De Cao N , Kipf T . MolGAN: An implicit generative model for small molecular graphs. 2018.

NetGAN



(a) Generator architecture

(b) NetGAN architecture



(a) Original graph

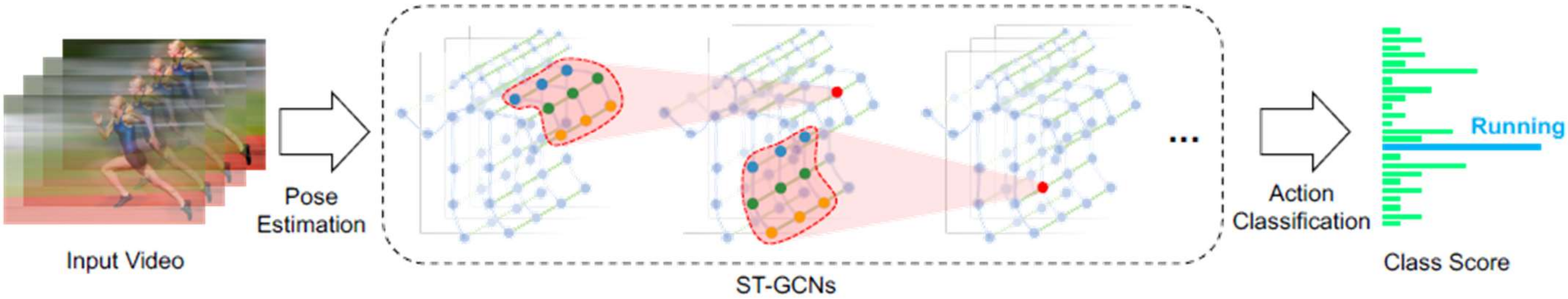
(b) Graph generated by NetGAN

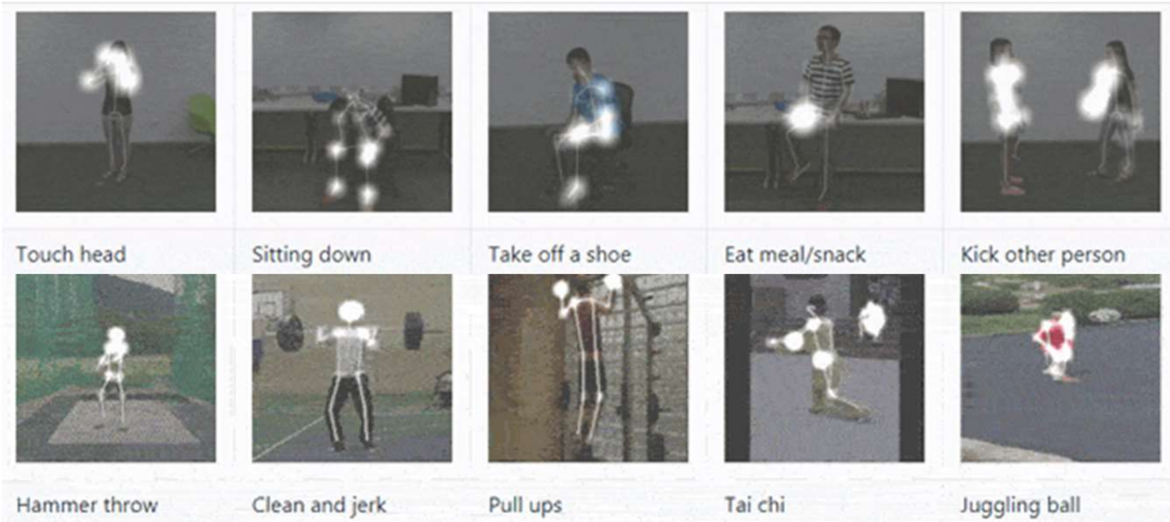
(c) Degree distribution comparison

Bojchevski A, Shchur O, Zügner D, et al. Netgan: Generating graphs via random walks. ICML 2018.

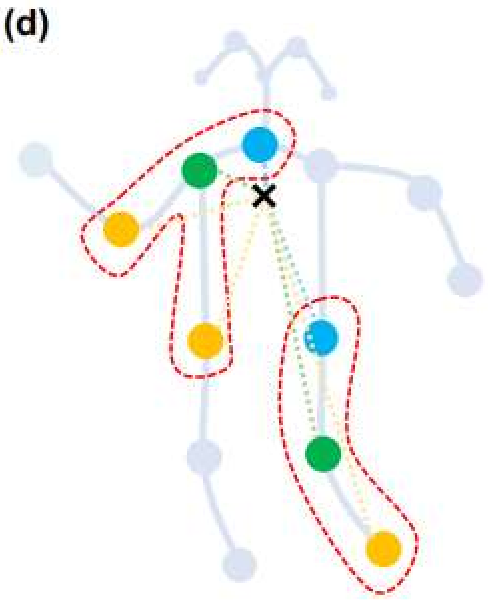
5、时空图网络

ST-GCN





Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. Sijie Yan, Yuanjun Xiong and Dahua Lin, AAAI 2018.



- the root node itself
- centripetal group: the neighboring nodes that are closer to the gravity center of the skeleton than the root node
- otherwise the centrifugal group

$$l_{ti}(v_{tj}) = \begin{cases} 0 & \text{if } r_j = r_i \\ 1 & \text{if } r_j < r_i \\ 2 & \text{if } r_j > r_i \end{cases}$$

Spatial configuration partitioning

$$\mathbf{f}_{out} = \mathbf{\Lambda}^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{f}_{in}\mathbf{W} \quad \Longrightarrow \quad \mathbf{f}_{out} = \sum_j \mathbf{\Lambda}_j^{-\frac{1}{2}}\mathbf{A}_j\mathbf{\Lambda}_j^{-\frac{1}{2}}\mathbf{f}_{in}\mathbf{W}_j$$

讲课提纲

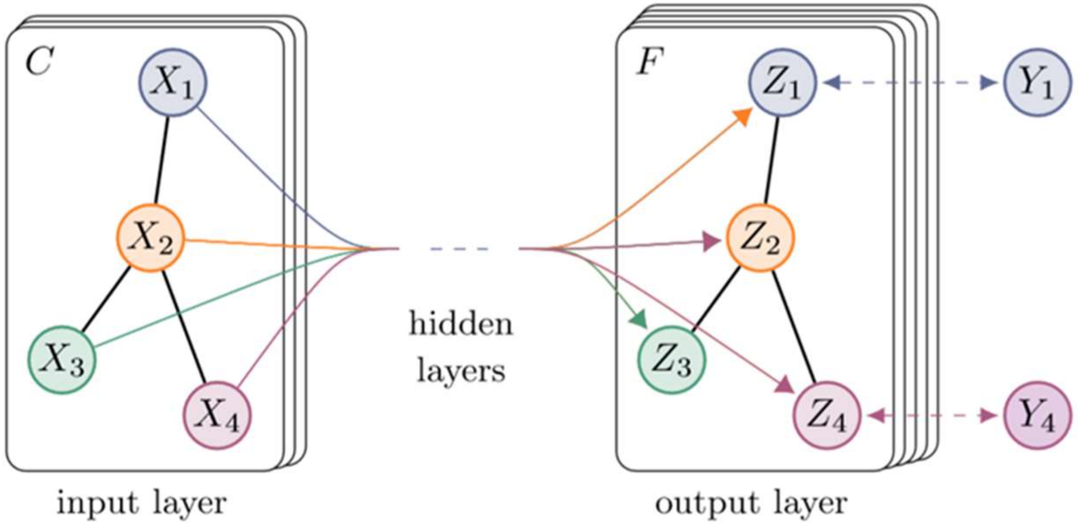
一、图神经网络发展

二、典型图神经网络

三、图神经网络应用

1、引文网络节点半监督分类

节点分类



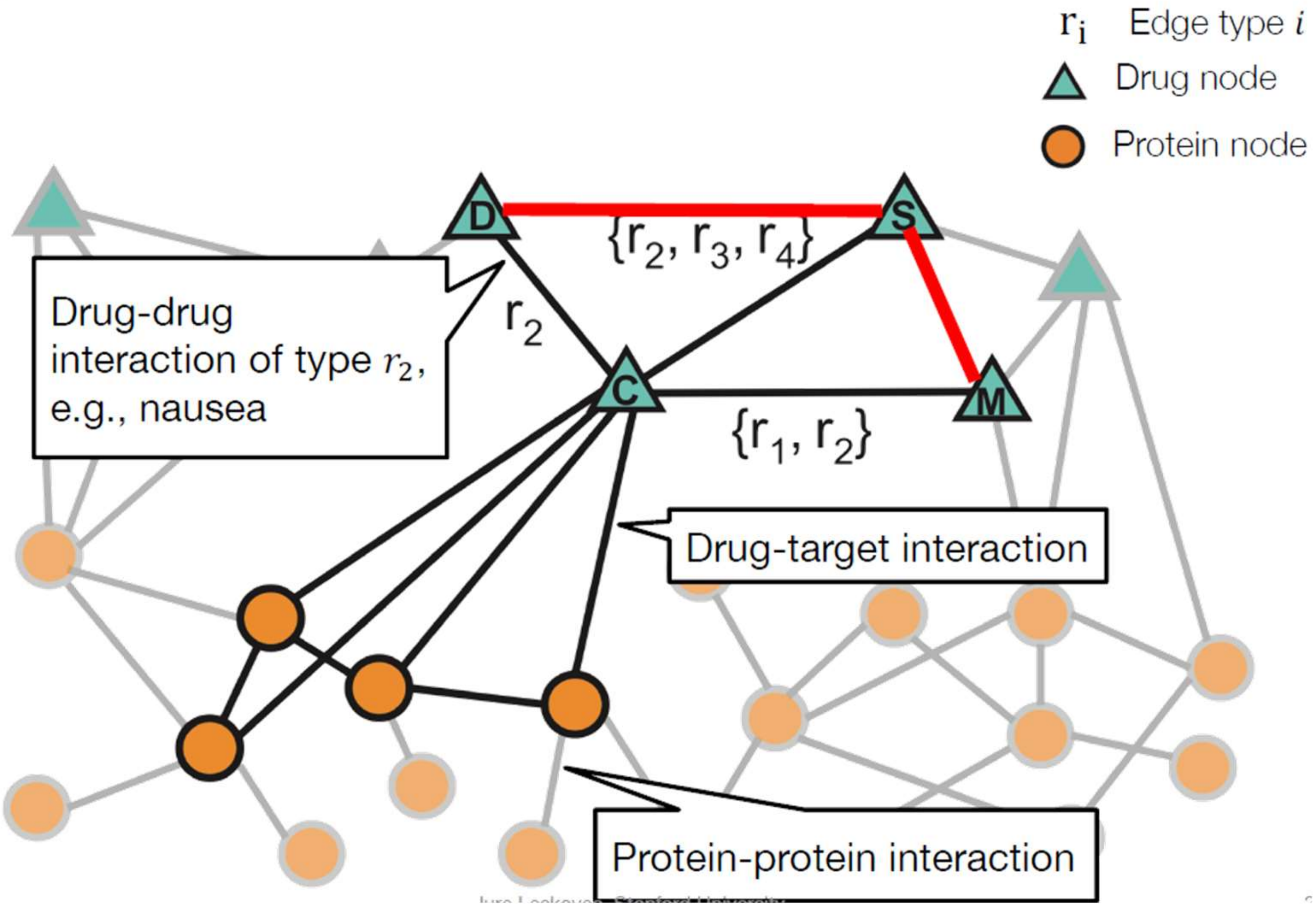
给定图结构和部分节点的标签，通常标注率较小(Citeseer 0.036, Cora 0.052, Pubmed 0.003)

目标：对未给出标签的节点进行分类

$$\mathcal{L} = -\sum_{i \in L} \sum_{j=1}^{c} Y_{ij} \ln Z_{ij}$$

2、多药物副作用预测

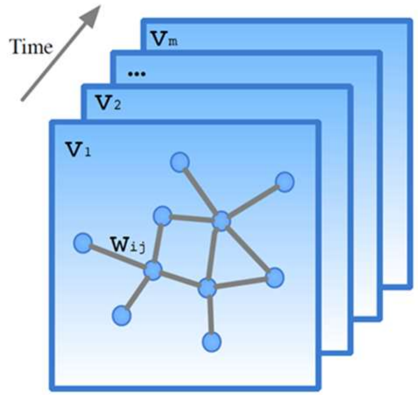
链接预测



Marinka Z , Monica A , Jure L . Modeling polypharmacy side effects with graph convolutional networks[J]. Bioinformatics, 2018, 34(13):i457-i466.

3、交通预测

交通预测



$G = (\mathcal{V}, \varepsilon, W)$

The purpose of traffic prediction task is to use previously observed road speed records to forecast the future status in a certain period of a specified region. Historical traf-

在第t时刻，在图$G = (\mathcal{V}, \varepsilon, W)$中，$\mathcal{V}$是一个有限的顶点集合，对应于交通网络中n个监测站的观测值; ε是一组边，表示站点之间的连通性; W表示G的加权邻接矩阵

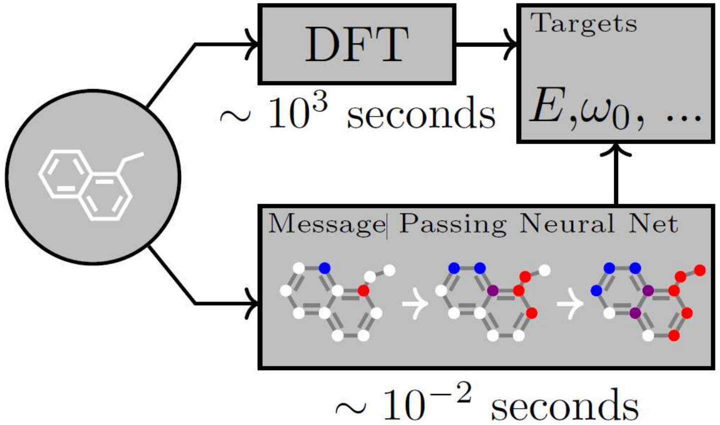
Yu B, Yin H, Zhu Z. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting[C]. IJCAI 2018.

4、分子特性预测

图分类

- 目标：预测分子的化学/物理特性，如原子化能 (atomization energy)、基本振动 (fundamental vibrations)、电子能隙 (electron-energy gap)等
- 传统方法：量子力学模拟方法DFT (Density Functional Theory)

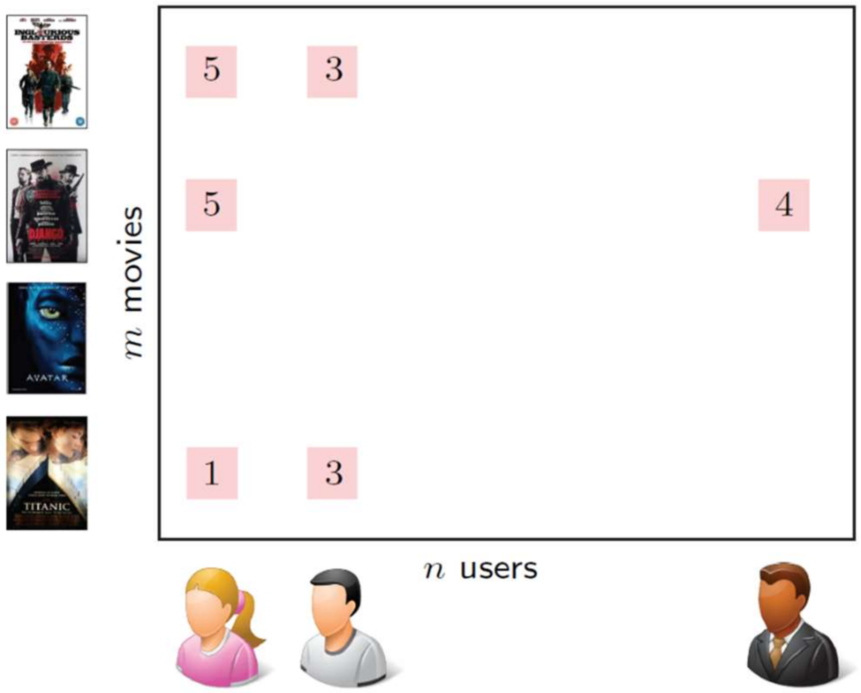
  计算复杂度高 ($\mathcal{O}(N_e^3)$), 难以扩展到大分子结构！
- 图神经网络方法：Message Passing Neural Network

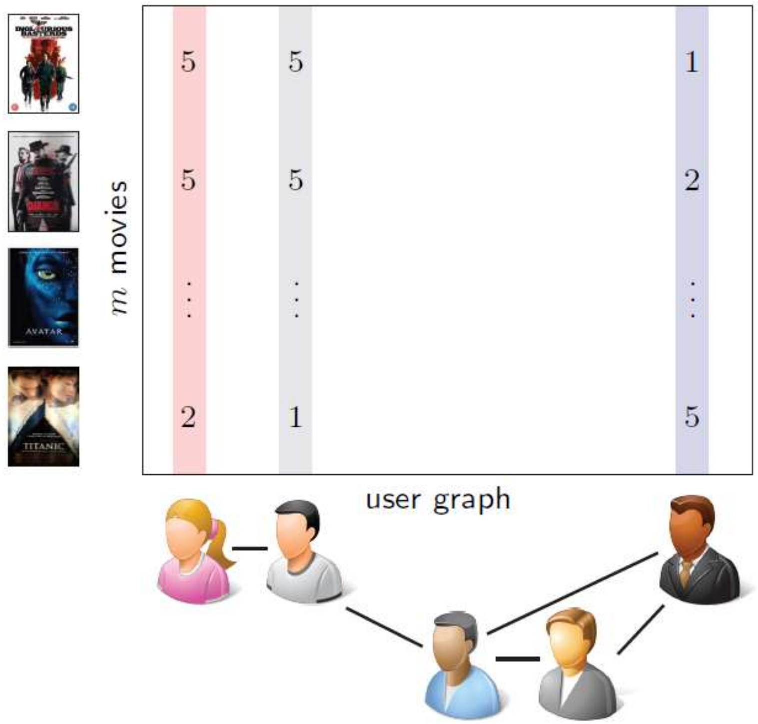


在QM9数据集上MPNN的推理时间比传统DFT方法快300k倍

Gilmer J, Schoenholz S S, Riley P F, et al. Neural message passing for quantum chemistry[C]//Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017: 1263-1272.

5、推荐系统



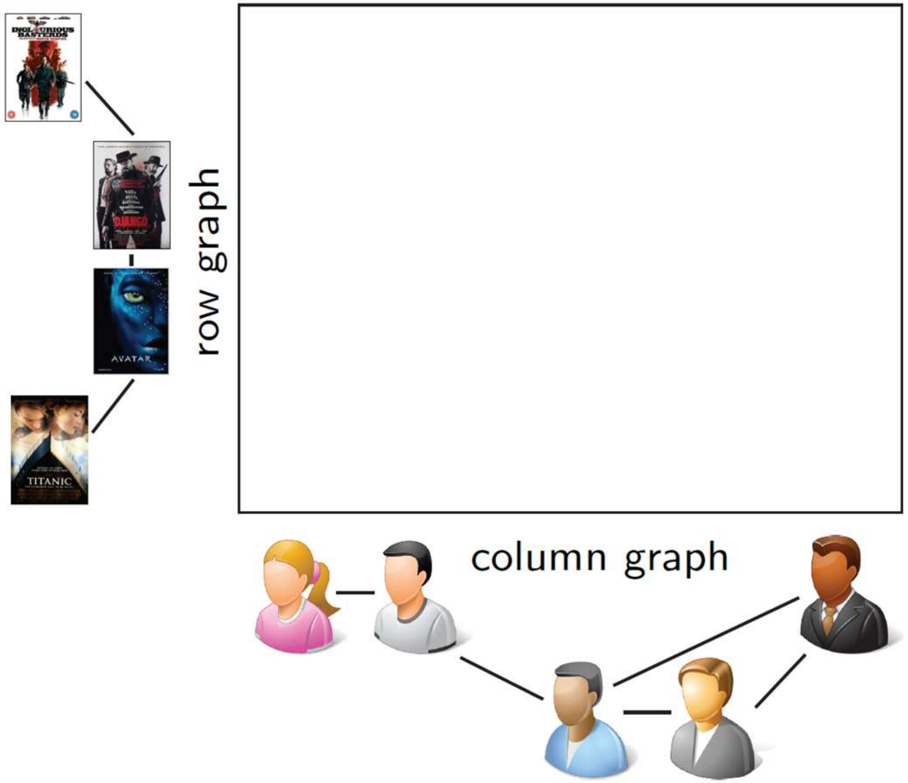


$$\min_{\mathbf{X}\in\mathbb{R}^{m\times n}} \|\mathbf{X}\|_* + \mu\|\Omega\circ(\mathbf{X}-\mathbf{A})\|_F^2$$

Candès. 2008

$$\min_{\mathbf{X}\in\mathbb{R}^{m\times n}} \mu\|\Omega\circ(\mathbf{X}-\mathbf{A})\|_F^2 + \underbrace{\mu_c \mathrm{tr}\big(\mathbf{X}\boldsymbol{\Delta}_c\mathbf{X}^T\big)}_{\|\mathbf{X}\|_{\mathcal{G}_c}^2}$$

Kalofolias et al. 2014



Multi-graph Fourier transform

$$\hat{\mathbf{X}} = \boldsymbol{\Phi}_{\mathrm{r}}^{\mathrm{T}} \mathbf{X} \boldsymbol{\Phi}_{\mathrm{c}}$$

Multi-graph spectral convolution

$$\mathbf{X} \star \mathbf{G} = \boldsymbol{\Phi}_{\mathrm{r}} \left(\left(\boldsymbol{\Phi}_{\mathrm{r}}^{\mathrm{T}} \mathbf{X} \boldsymbol{\Phi}_{\mathrm{c}} \right) \circ \left(\boldsymbol{\Phi}_{\mathrm{r}}^{\mathrm{T}} \mathbf{G} \boldsymbol{\Phi}_{\mathrm{c}} \right) \right) \boldsymbol{\Phi}_{\mathrm{c}}^{\mathrm{T}}$$

$$\mathbf{X} \star \mathbf{G} = \boldsymbol{\Phi}_{\mathrm{r}} \left(\hat{\mathbf{X}} \circ \hat{\mathbf{G}} \right) \boldsymbol{\Phi}_{\mathrm{c}}^{\mathrm{T}}$$

Multi-graph bi-variate polynomial filter

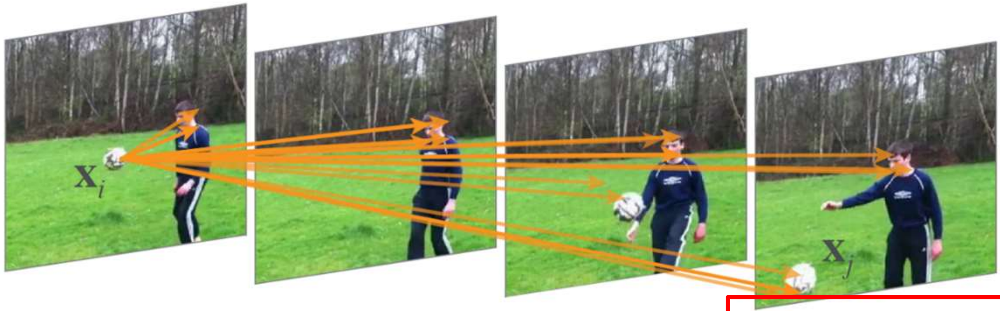
$$\mathbf{Y} = \tau_{\boldsymbol{\Theta}}(\mathbf{X}) = \sum_{j,j'=1}^{r} \theta_{jj'} \boldsymbol{\Delta}_{\mathrm{r}}^{j} \mathbf{X} \boldsymbol{\Delta}_{\mathrm{c}}^{j'}$$

$\boldsymbol{\Theta} = \left(\boldsymbol{\theta}_{jj'} \right)$是$r \times r$的滤波器参数矩阵

Monti F, Bronstein M, Bresson X. Geometric matrix completion with recurrent multi-graph neural networks[C]//Advances in Neural Information Processing Systems. 2017: 3697-3707.

6、视频分类

Non-local Neural Networks



Gaussian $f(\mathbf{x}_i, \mathbf{x}_j) = e^{\mathbf{x}_i^T \mathbf{x}_j}$
Embedded Gaussian $f(\mathbf{x}_i, \mathbf{x}_j) = e^{\theta(\mathbf{x}_i)^T \phi(\mathbf{x}_j)}$
Dot product $f(\mathbf{x}_i, \mathbf{x}_j) = \theta(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$
Concatenation $f(\mathbf{x}_i, \mathbf{x}_j) = \mathrm{ReLU}(\mathbf{w}_f^T [\theta(\mathbf{x}_i), \phi(\mathbf{x}_j)])$

Non-local Operation $\boldsymbol{y_i} = \dfrac{1}{\mathcal{C}(\mathbf{x})} \sum_{\forall j} \boxed{f(\mathbf{x}_i, \mathbf{x}_j)} g(\mathbf{x}_j)$

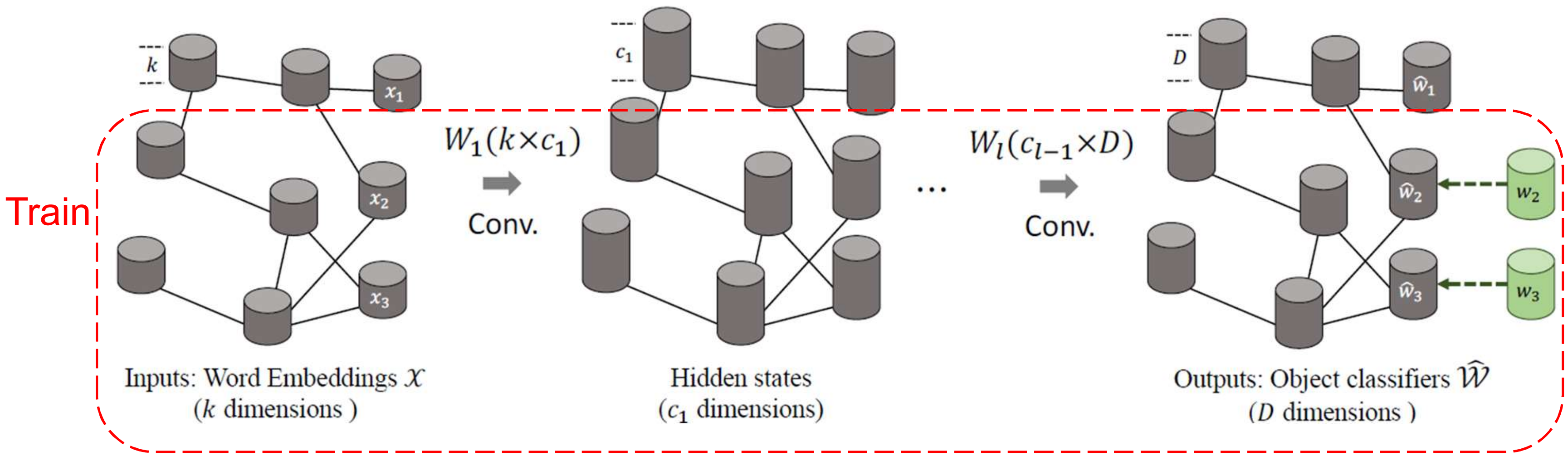
建立视频中某一位置与其他所有位置 (时间、空间) 之间的关联，构成“图模型”，捕捉数据间的长距离依赖关系

Non-local Neural Networks可以视为不同”self-attention”方法的统一形式

Wang X, Girshick R, Gupta A, et al. Non-local neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7794-7803.

7、Zero-shot图像分类

Zero-shot图像分类任务：对训练集中没有出现过的图像进行分类



输入训练类别文本的
word embeddings

采用GCN进行训练, 图结构为知识图谱
(NELL / WordNet, 简化为无向图)

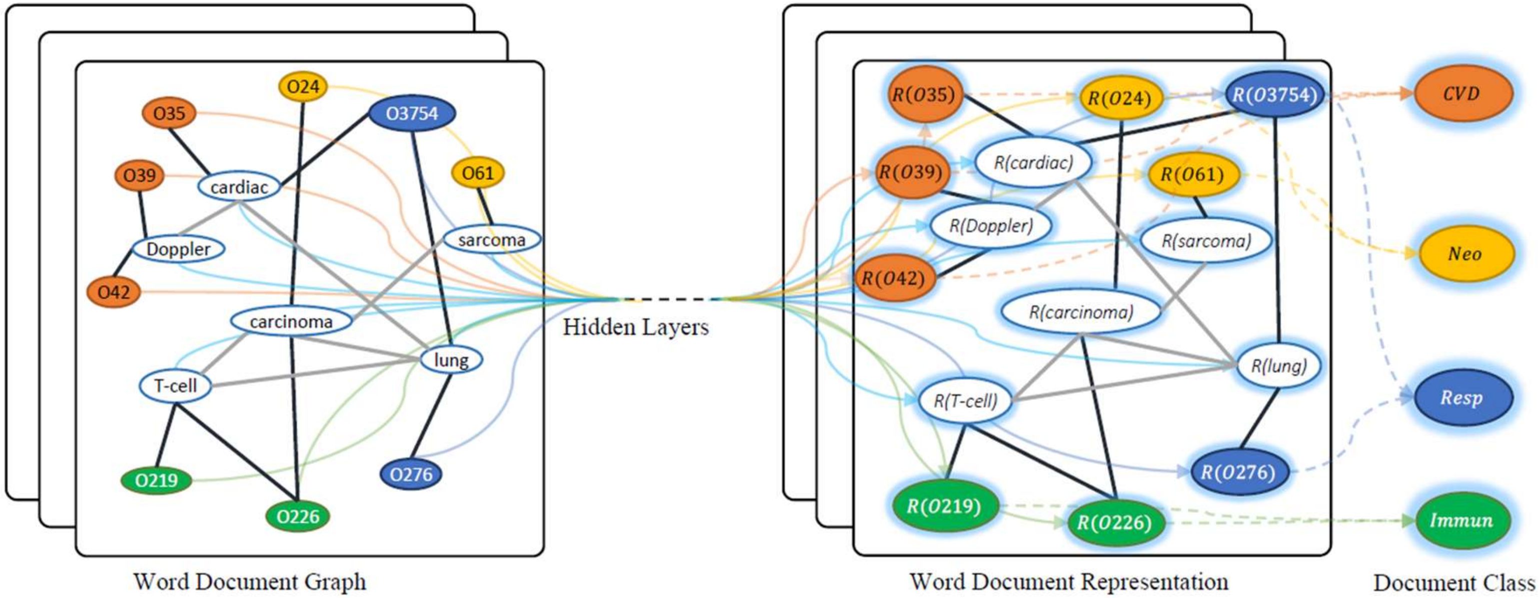
输出分类器$\widehat{w}$,拟合对训练图样本
采用CNN训练得到的分类器w

Loss: $\frac{1}{m}\sum_{i=1}^{m} L_{mse}(\widehat{w}_i, w_i)$

Wang X, Ye Y, Gupta A. Zero-shot recognition via semantic embeddings and knowledge graphs[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6857-6866.

中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

8、文本分类



$$A_{ij} = \begin{cases} \mathrm{PMI}(i,j) & i \text{ and } j \text{ are words and } \mathrm{PMI}(i,j) > 0 \\ \mathrm{TFIDF}_{ij} & i \text{ is a document and } j \text{ is a word} \\ 1 & i = j \\ 0 & \text{otherwise} \end{cases}$$

$$\mathrm{PMI}(i,j) = \log \frac{p(i,j)}{p(i)p(j)}$$

$$p(i,j) = \frac{\#W(i,j)}{\#W}$$

$$p(i) = \frac{\#W(i)}{\#W}$$

Yao L, Mao C, Luo Y. Graph convolutional networks for text classification. AAAI 2019.

Github代码链接

GCN

https://github.com/tkipf/gcn

GAT

https://github.com/PetarV-/GAT

GraphSAGE

https://github.com/williamleif/GraphSAGE

Text-GCN

https://github.com/yao8839836/text_gcn

Graph Nets Library

https://github.com/deepmind/graph_nets

Deep Graph Library (DGL)

https://github.com/dmlc/dgl

PyTorch Geometric

https://github.com/rusty1s/pytorch_geometric

感谢大家!

欢迎批评指正