图神经网络导论

自监督图神经网络

周晟

浙江大学　软件学院

2022.12

课程内容

1. 图自监督学习
2. 图互信息最大化
3. 图对比学习

有/无监督学习

有监督学习：

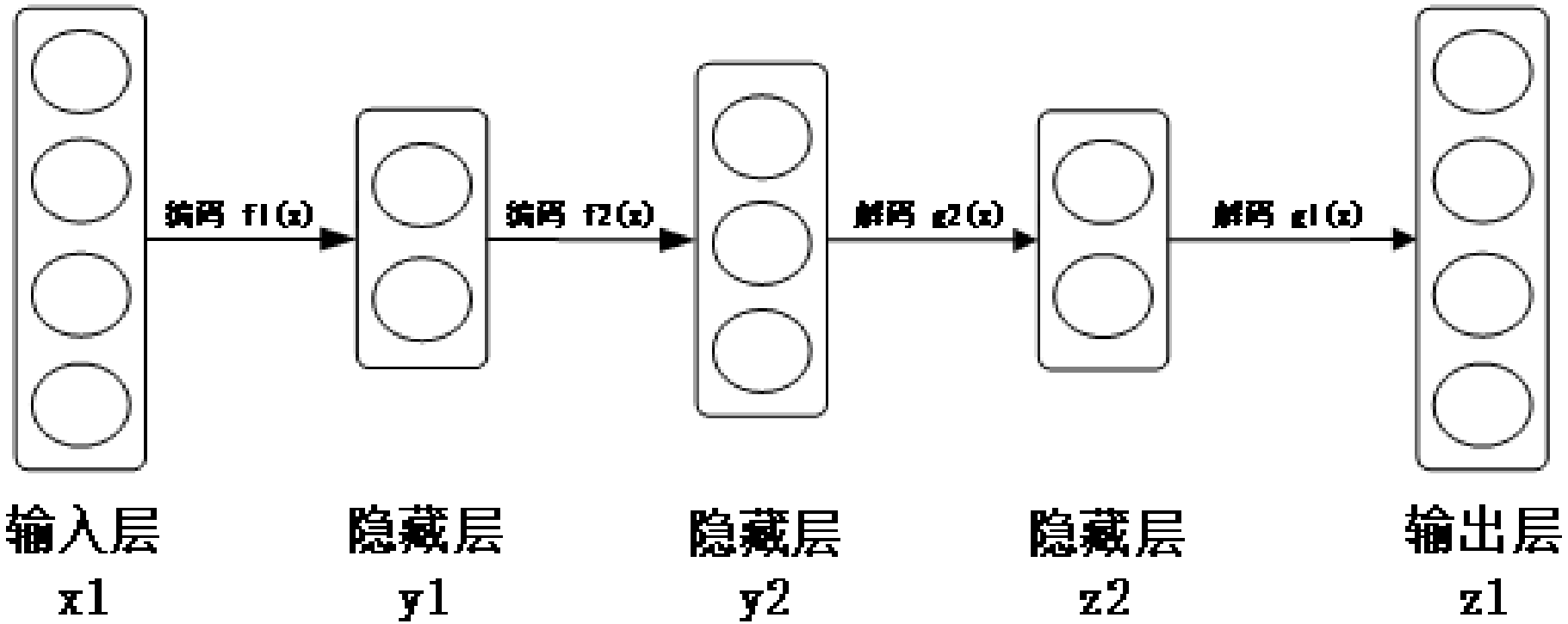
1. 效果很好
2. 需要大量人为标注
3. 模型存在上限

无监督学习：

1. 效果相对弱
2. 无标注数据容易获得
3. 类人类学习过程

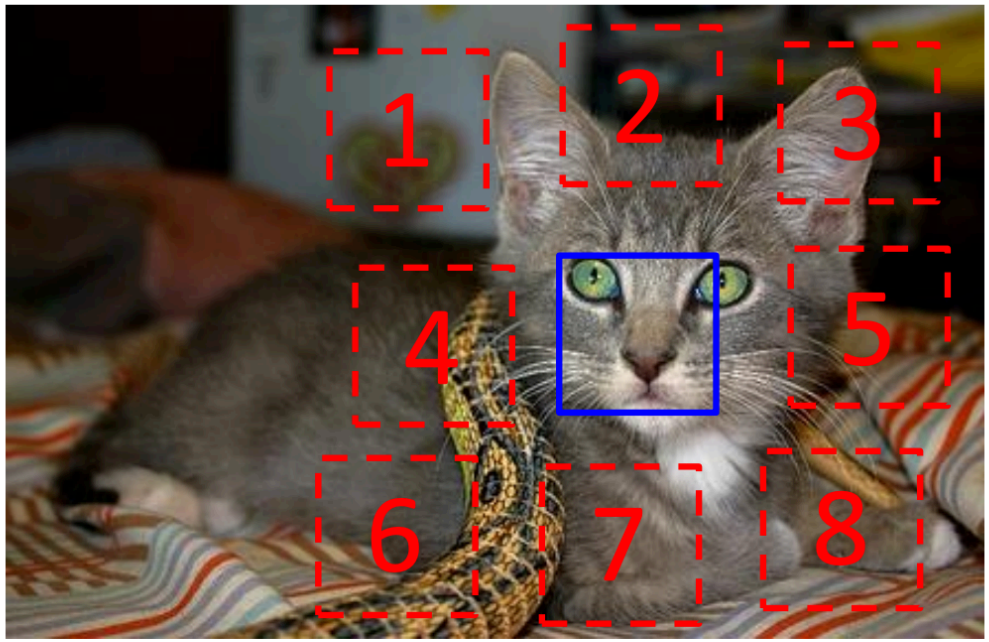
自监督学习是无监督学习的子集，其监督信号来自数据本身

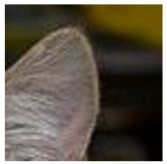
自监督学习任务



自编码器是最经典的自监督学习方法

自监督学习任务

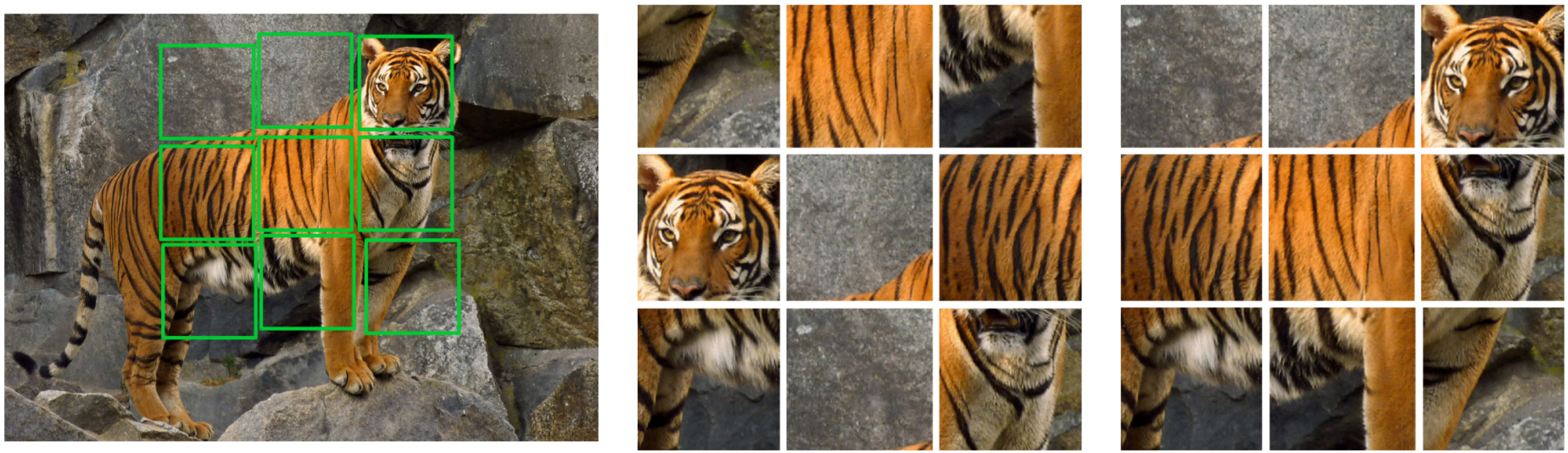


$X = ($  ,  $); Y = 3$

位置预测任务[1]

[1]Unsupervised Visual Representation Learning by Context Prediction(ICCV 2015)

自监督学习任务

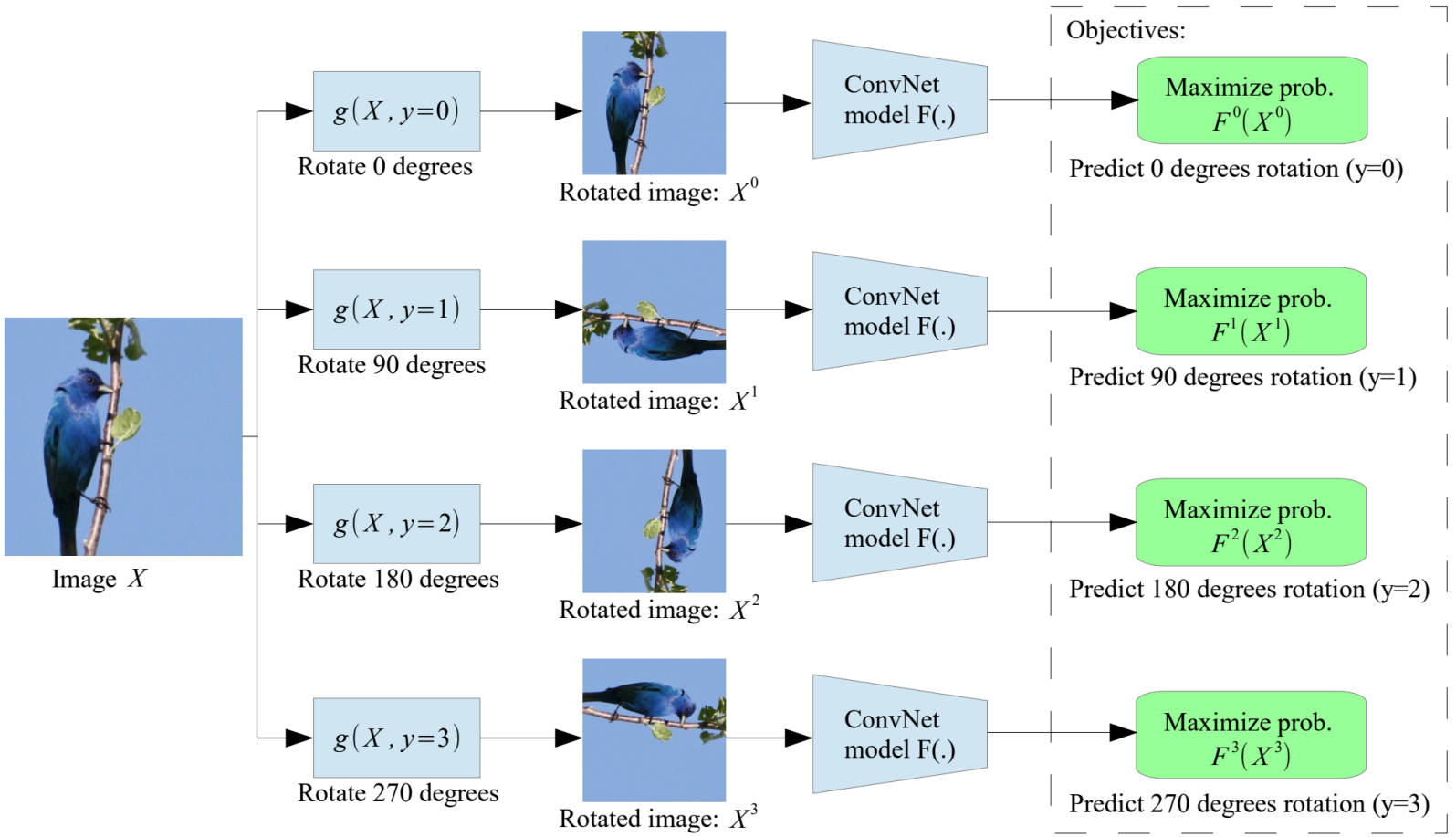


Jigsaw puzzles 任务[2]

给定打乱顺序后的 patch，预测 patch 之间的相对位置关系

[2]Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles(ECCV 2016)

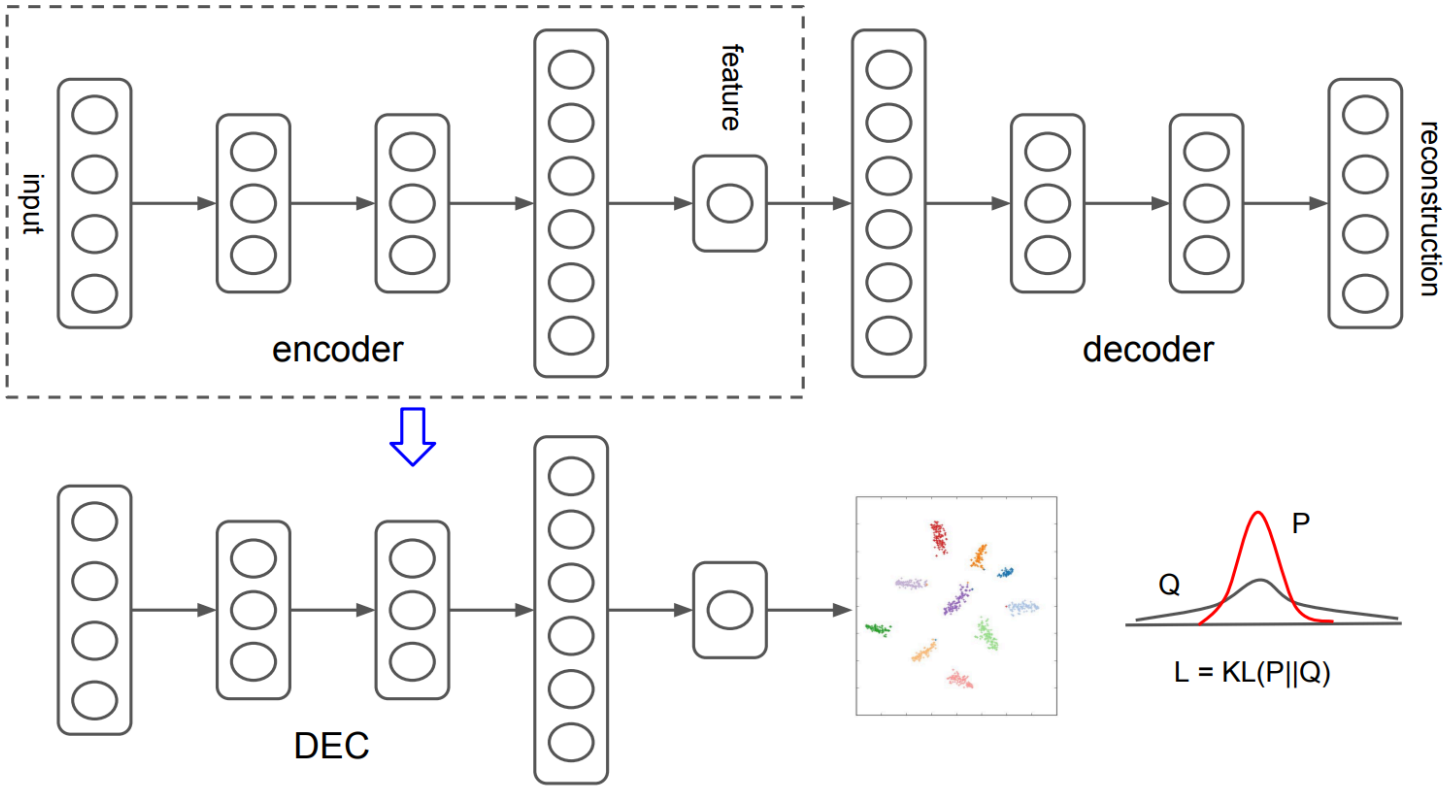
自监督学习任务



旋转预测任务[3]

[3]Unsupervised Representation Learning by Predicting Image Rotations(ICLR 2018)

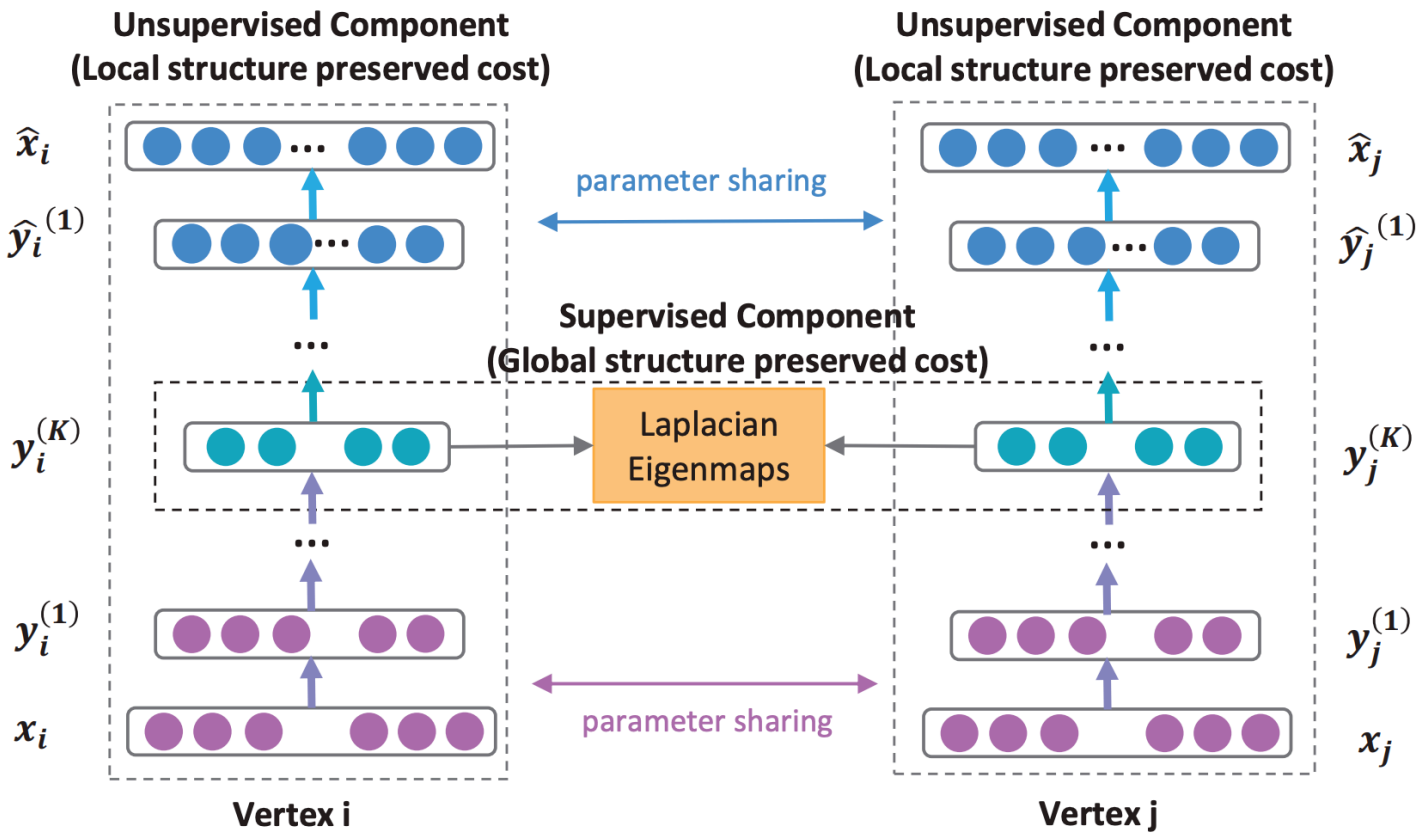
自监督学习任务



聚类预测任务[4]

[4]Unsupervised Deep Embedding for Clustering Analysis(ICML 2016)

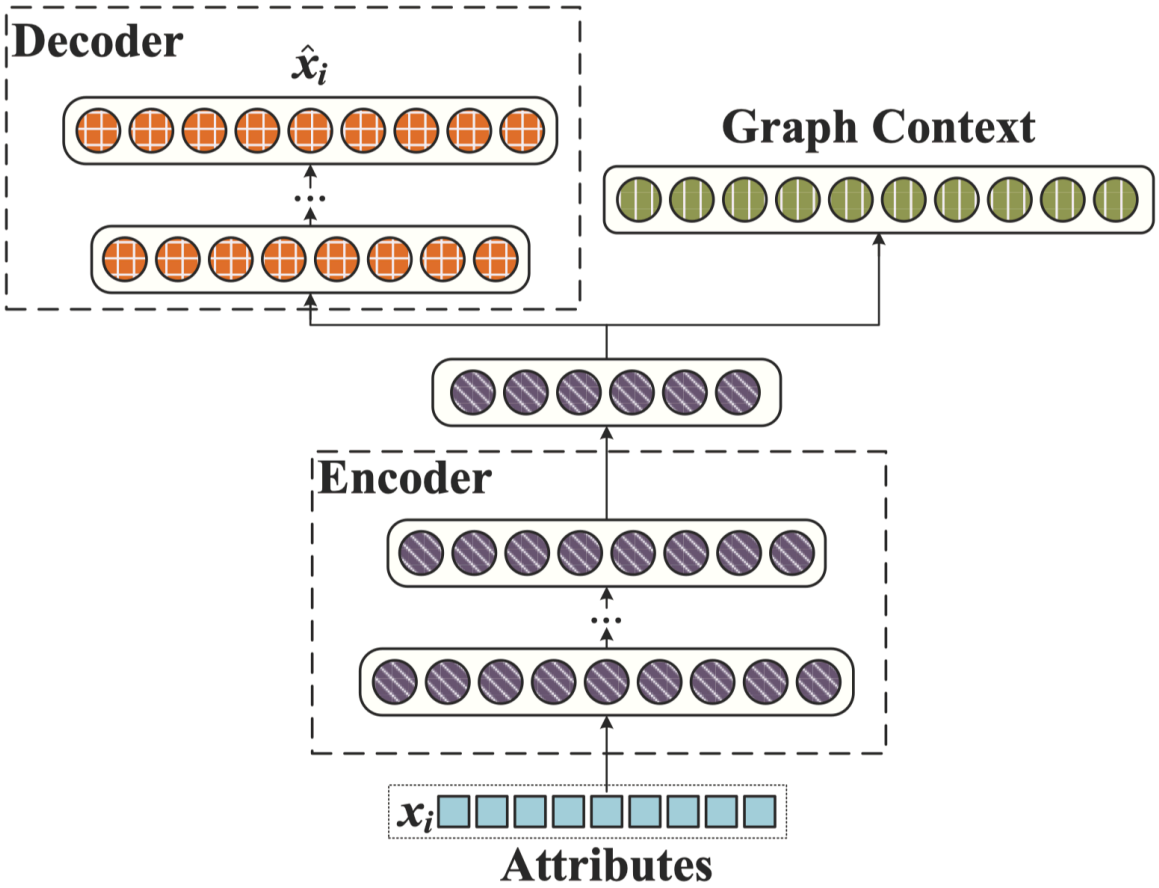
图自监督学习: 自编码器



SDNE[5]

[5]Structural Deep Network Embedding(KDD 2016)

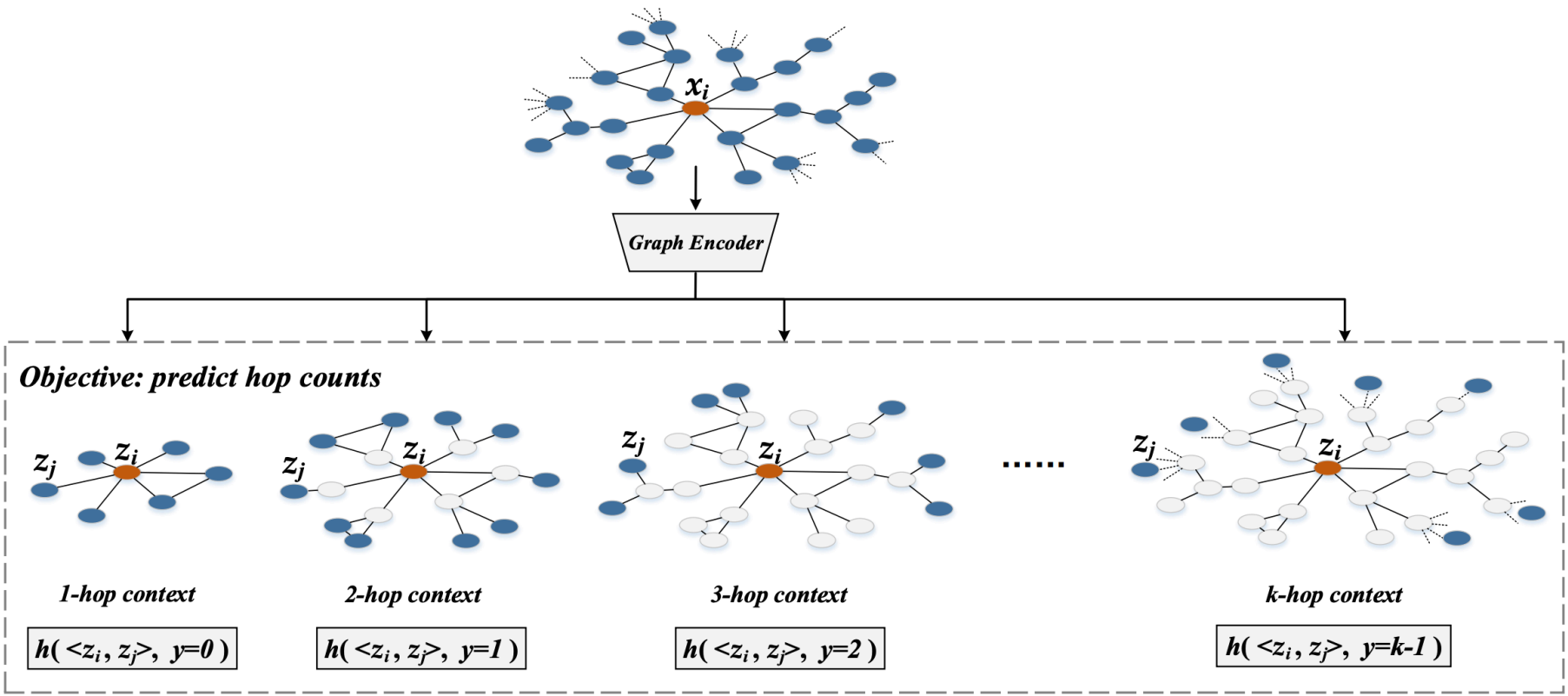
图自监督学习: 自编码器



同时重构节点属性和图结构[6]

[6]ANRL: Attributed Network Representation Learning via Deep Neural Networks(IJCAI 2018)

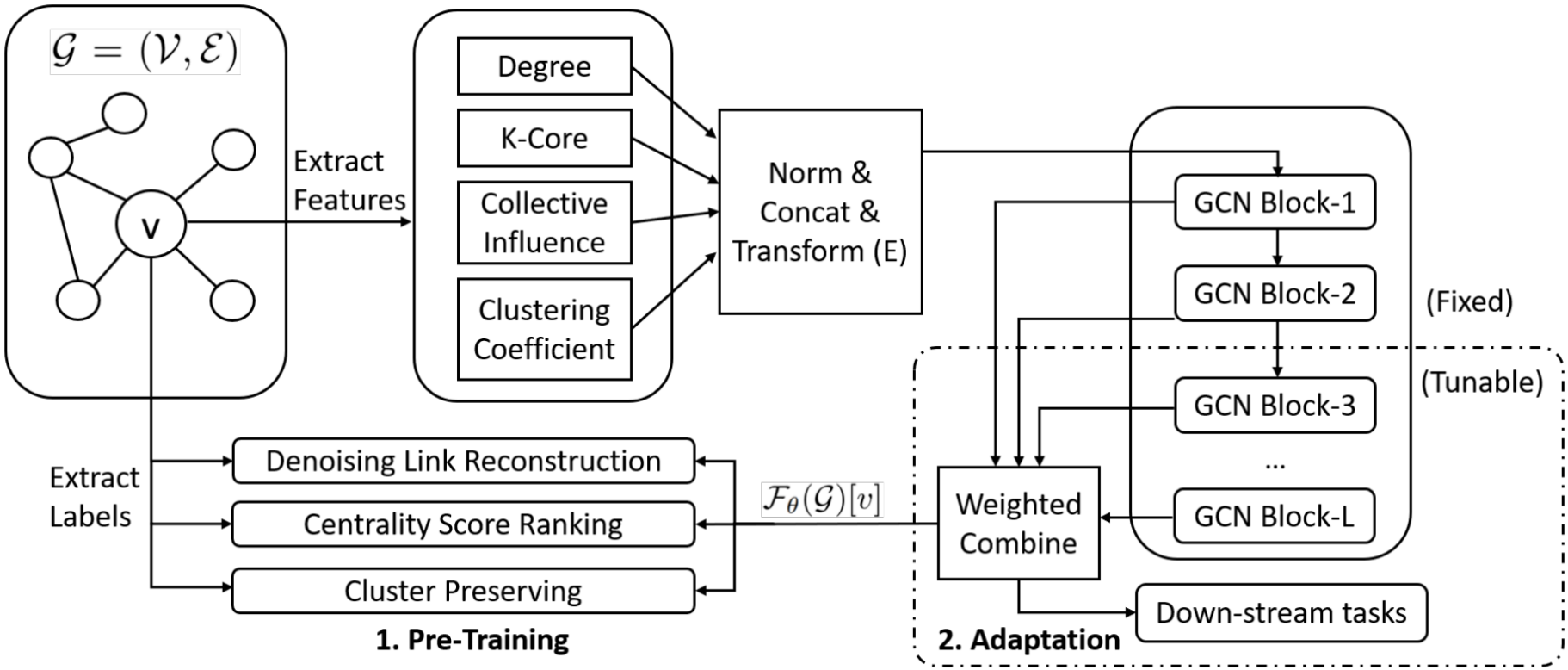
图自监督学习: 图性质预测



预测图上节点之间的最短路径[7]

[7]Self-Supervised Graph Representation Learning via Global Context Prediction(IJCAI 2020)

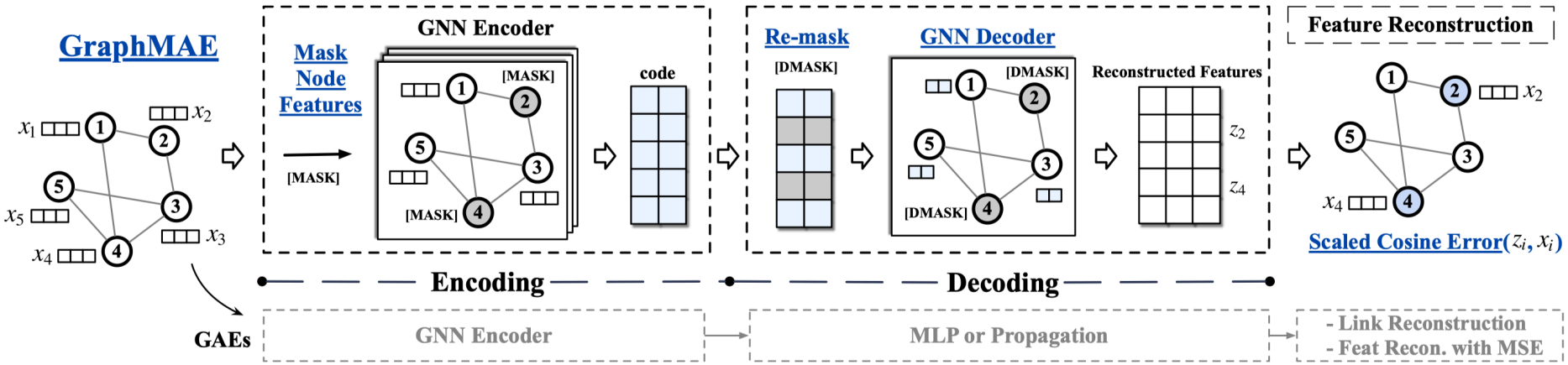
图自监督学习: 图性质预测



预测图上其他性质[8]

[8]Pre-Training Graph Neural Networks for Generic Structural Feature Extraction(Arxiv 2019)

GraphMAE



GraphMAE[9]

GraphMAE 核心设计：

1. 带 [MASK] 的节点特征重建
2. 重掩码的 GNN 解码器
3. 放缩余弦误差

[9] GraphMAE: Self-Supervised Masked Graph Autoencoders(KDD 2022)

GraphMAE

1. 带 [MASK] 的节点特征重建
   1. 随机采样部分节点
   2. 使用 [MASK] 来替换这些节点的特征
   3. 未被选中的节点保留原始特征
2. 重掩码的 GNN 解码器
   1. 使用单层 GNN 作为解码器
   2. 使用 [DMASK] 来替换采样节点的表征
   3. 解码器实现搜集邻居信息来重建原始特征
3. 放缩余弦误差

$$\mathcal{L}_{\text{SCE}} = \frac{1}{|\widetilde{\mathcal{V}}|} \sum_{v_i \in \widetilde{\mathcal{V}}} \left(1 - \frac{x_i^T z_i}{\|x_i\| \cdot \|z_i\|}\right)^{\gamma}, \gamma \geq 1$$

互信息

互信息是一种随机变量之间非线性依赖关系的度量：

$$I(X;Z) := H(X) - H(X \mid Z)$$

$$I(X,Z) = D_{KL}\left(\mathbb{P}_{XZ} \| \mathbb{P}_X \otimes \mathbb{P}_Z\right)$$

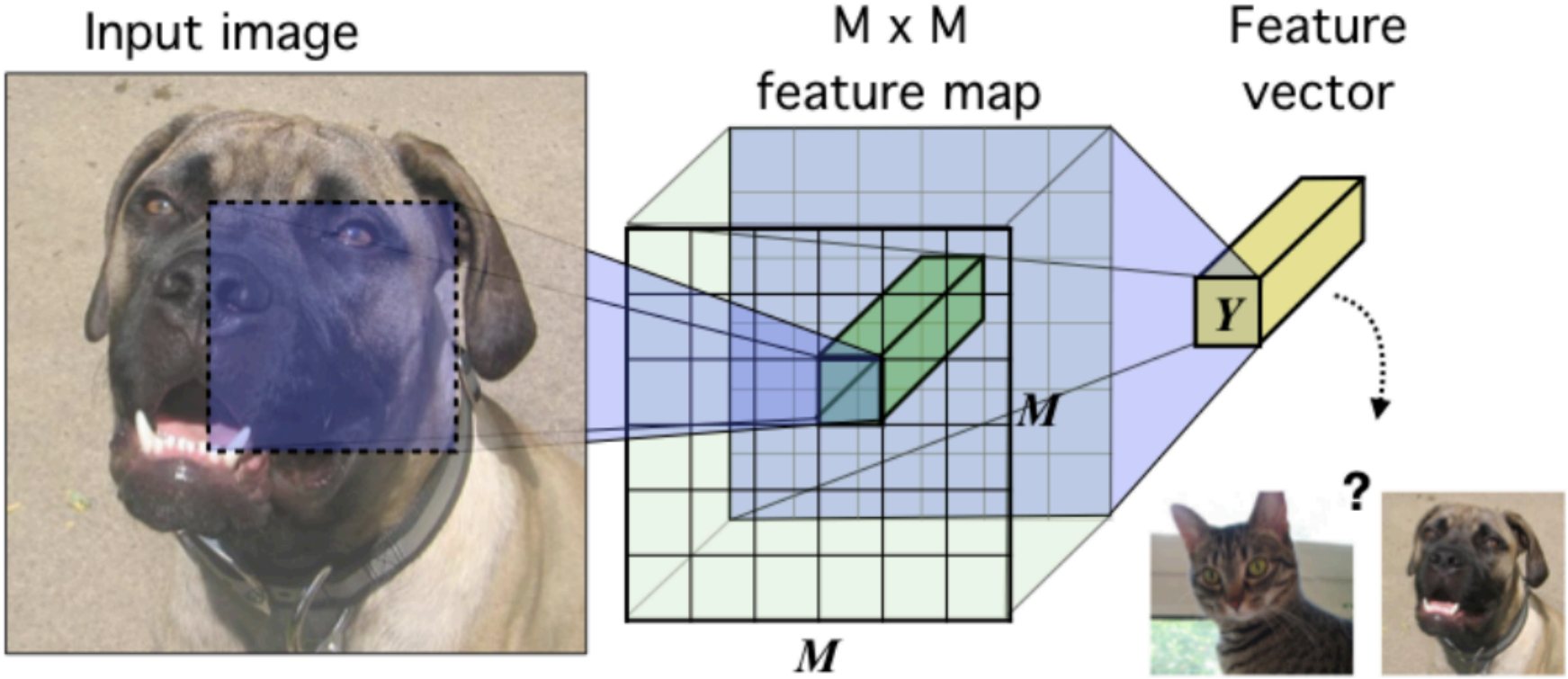
使用神经网络估计互信息[10]：

$$\widehat{I(X;Z)}_n = \sup_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_{XZ}^{(n)}}\left[T_\theta\right] - \log\left(\mathbb{E}_{\mathbb{P}_X^{(n)} \otimes \hat{\mathbb{P}}_Z^{(n)}}\left[e^{T_\theta}\right]\right)$$

互信息的神经网络估计为后续表征学习提供了思路。

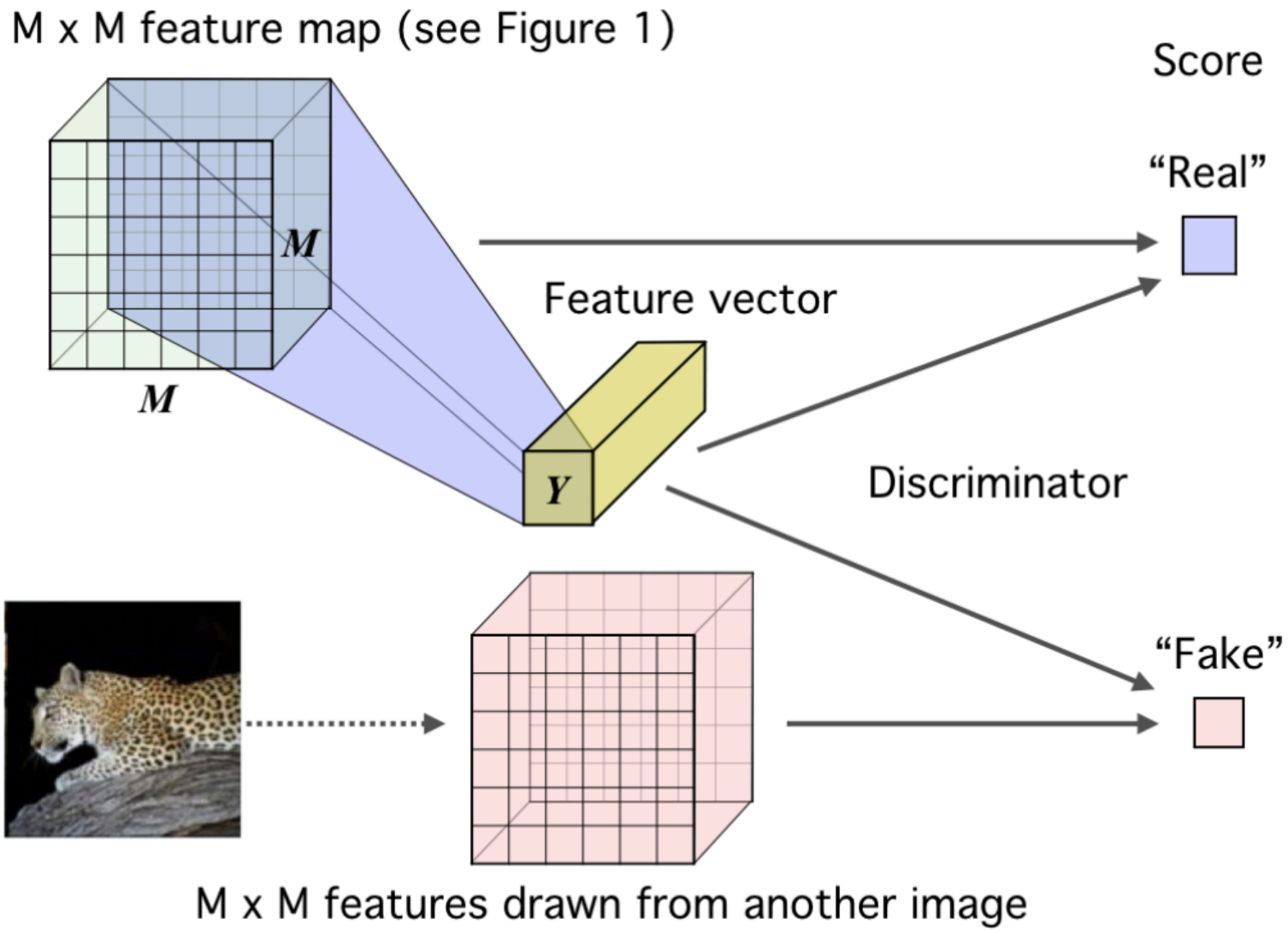
[10] Mutual Information Neural Estimation(ICML 2018)

Deep Infomax



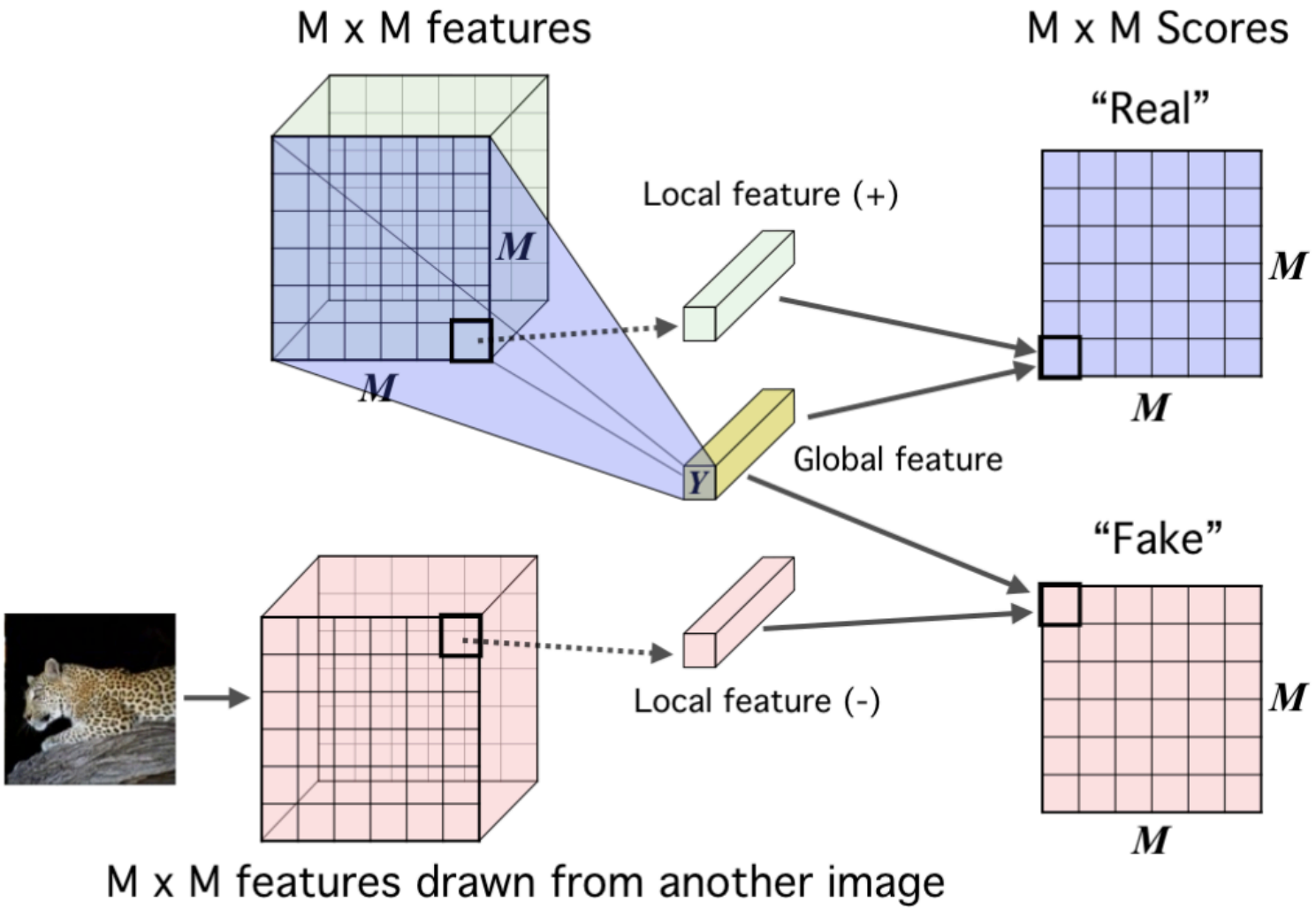
好的表征应该尽可能多地保留数据的特征

Deep Infomax



保留原始特征的表征可以和其他数据相区分

Deep Infomax



好的全局表征可以与其对应的局部表征对齐

DeepInfomax

DeepInfomax 希望通过学到和原始数据互信息最大的表征：

$$(\hat{\omega}, \hat{\psi})_G = \arg\max_{\omega,\psi} \widehat{\mathcal{I}}_\omega \left(X; E_\psi(X)\right)$$

$$(\hat{\omega}, \hat{\psi})_L = \arg\max_{\omega,\psi} \frac{1}{M^2} \sum_{i=1}^{M^2} \widehat{\mathcal{I}}_{\omega,\psi} \left(C_\psi^{(i)}(X); E_\psi(X)\right)$$

DeepInfomax 提供了两种度量互信息的方式：

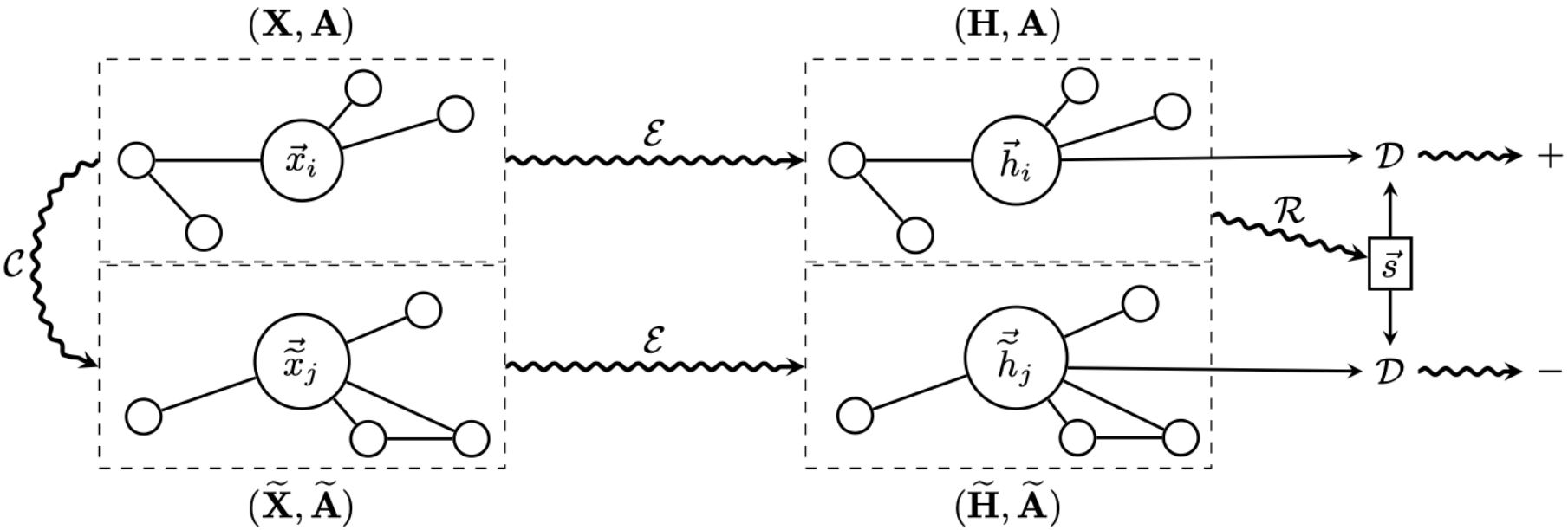
1. JSD

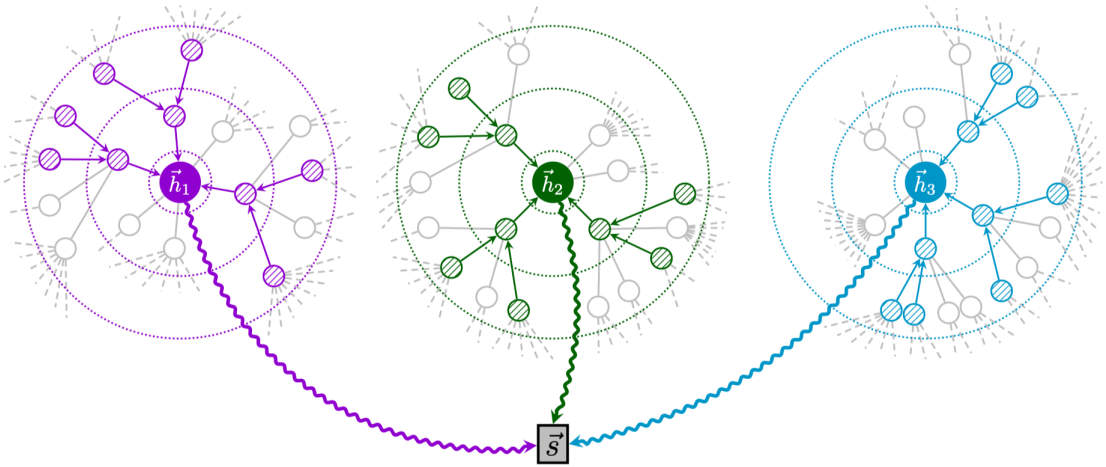
$$\mathbb{E}_{\mathbb{P}} \left[-\mathrm{sp}\left(-T_{\psi,\omega}\left(x, E_\psi(x)\right)\right)\right] - \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}} \left[\mathrm{sp}\left(T_{\psi,\omega}\left(x', E_\psi(x)\right)\right)\right]$$

2. InfoNCE

$$\mathbb{E}_{\mathbb{P}} \left[T_{\psi,\omega}\left(x, E_\psi(x)\right) - \mathbb{E}_{\tilde{\mathbb{P}}} \left[\log \sum_{x'} e^{T_{\psi,\omega}\left(x', E_\psi(x)\right)}\right]\right]$$

Deep Graph Infomax





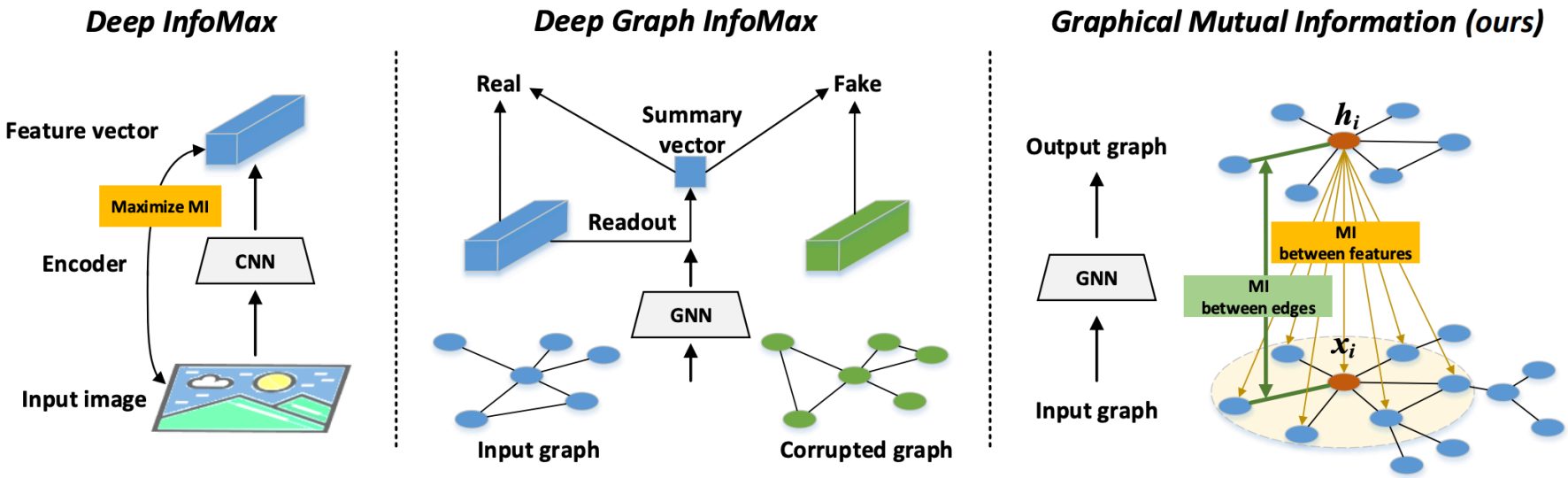
Deep Graph Infomax

DGI 的核心流程：

1. 使用扰动函数生成负样本
2. 对于原图中每个节点周围的子图使用图神经网络获得表征
3. 对于负样本图中每个节点周围的子图使用图神经网络获得表征
4. 将所有子图的表征融合后得到全局的表征
5. 利用互信息最大化优化图神经网络

$$\frac{1}{N+M}\left(\sum_{i=1}^{N}\mathbb{E}_{(\mathbf{X},\mathbf{A})}\left[\log\mathcal{D}\left(\vec{h}_i,\vec{s}\right)\right]+\sum_{j=1}^{M}\mathbb{E}_{(\tilde{\mathbf{X}},\tilde{\mathbf{A}})}\left[\log\left(1-\mathcal{D}\left(\vec{h}_j,\vec{s}\right)\right)\right]\right.$$

Graph Representation Learning via Graphical Mutual Information Maximization(GMI)



DIM,DGI,GMI 对比[11]

[11] Graph Representation Learning via Graphical Mutual Information Maximization(WWW 2020)

GMI

属性互信息

$$I\left(\boldsymbol{h}_i; \boldsymbol{X}_i\right) = \int_{\mathcal{H}} \int_{\mathcal{X}} p\left(\boldsymbol{h}_i, \boldsymbol{X}_i\right) \log \frac{p\left(\boldsymbol{h}_i, \boldsymbol{X}_i\right)}{p\left(\boldsymbol{h}_i\right) p\left(\boldsymbol{X}_i\right)} d\boldsymbol{h}_i d\boldsymbol{X}_i$$

$$I\left(\boldsymbol{h}_i; \boldsymbol{x}_j\right) = -sp\left(-\mathcal{D}_w\left(\boldsymbol{h}_i, \boldsymbol{x}_j\right)\right) - \mathbb{E}_{\tilde{\mathbb{P}}}\left[sp\left(\mathcal{D}_w\left(\boldsymbol{h}_i, \boldsymbol{x'}_j\right)\right)\right]$$

结构互信息：

$$I\left(\boldsymbol{h}_i; \mathcal{G}_i\right) := \sum_{j}^{i_n} w_{ij} I\left(\boldsymbol{h}_i; \boldsymbol{x}_j\right) + I\left(w_{ij}; \boldsymbol{a}_{ij}\right),$$

$$\text{with } w_{ij} = \sigma\left(\boldsymbol{h}_i^{\mathrm{T}} \boldsymbol{h}_j\right),$$

$$I\left(w_{ij}; \boldsymbol{a}_{ij}\right) = \boldsymbol{a}_{ij} \log w_{ij} + \left(1 - \boldsymbol{a}_{ij}\right) \log\left(1 - w_{ij}\right)$$

从互信息到对比学习

等价性推导 [Oord et al., 2018]：

$$
\begin{aligned}
\mathcal{L}_{\mathrm{N}}^{\mathrm{opt}} &= -\mathop{\mathbb{E}}_{X} \log \left[\frac{\frac{p(x_{t+k}|c_t)}{p(x_{t+k})}}{\frac{p(x_{t+k}|c_t)}{p(x_{t+k})} + \sum_{x_j \in X_{\mathrm{neg}}} \frac{p(x_j|c_t)}{p(x_j)}} \right] \\
&= \mathop{\mathbb{E}}_{X} \log \left[1 + \frac{p\left(x_{t+k}\right)}{p\left(x_{t+k} \mid c_t\right)} \sum_{x_j \in X_{\mathrm{neg}}} \frac{p\left(x_j \mid c_t\right)}{p\left(x_j\right)} \right] \\
&\approx \mathop{\mathbb{E}}_{X} \log \left[1 + \frac{p\left(x_{t+k}\right)}{p\left(x_{t+k} \mid c_t\right)} (N-1) \mathop{\mathbb{E}}_{x_j} \frac{p\left(x_j \mid c_t\right)}{p\left(x_j\right)} \right] \\
&= \mathop{\mathbb{E}}_{X} \log \left[1 + \frac{p\left(x_{t+k}\right)}{p\left(x_{t+k} \mid c_t\right)} (N-1) \right] \\
&\geq \mathop{\mathbb{E}}_{X} \log \left[\frac{p\left(x_{t+k}\right)}{p\left(x_{t+k} \mid c_t\right)} N \right] = -I\left(x_{t+k}, c_t\right) + \log(N),
\end{aligned}
$$

对比学习

对比学习

对比学习（Contrastive Learning）的核心思想是在特征空间中将样本分别与正例样本和负例样本进行对比，以此来学习样本的特征表示。

对比学习的核心目标：

$$score(f(x), f(x^+)) >> score(f(x), f(x^-))$$

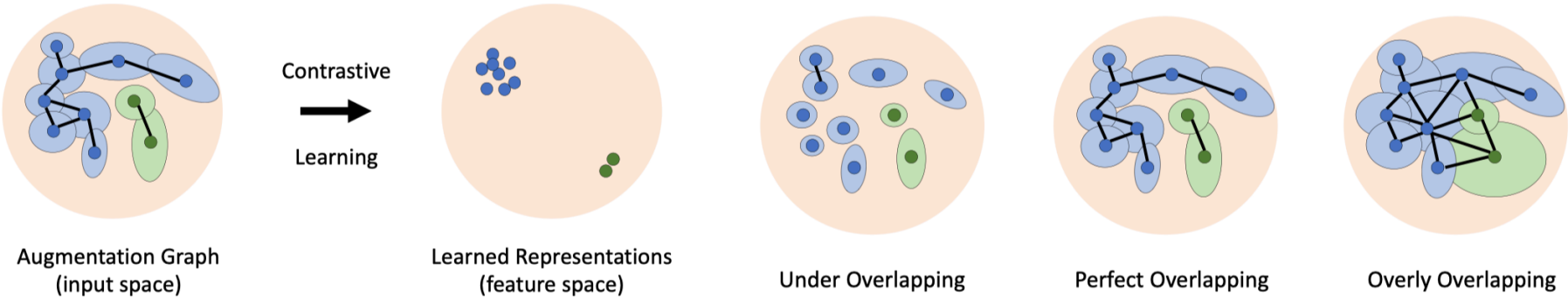
其中，x^+ 为与 x 相似的样本，称为正样本，两者组成正对。x^- 为与 x 不相似的样本，称为负样本，两者组成负对。

如何构造正负样本?

数据增强

数据增强

以图像数据为例，数据增强（Data Augmentation）是在原始图像上进行一系列图像变换，使原始图像变为语义等价但表现形式不同的图像。



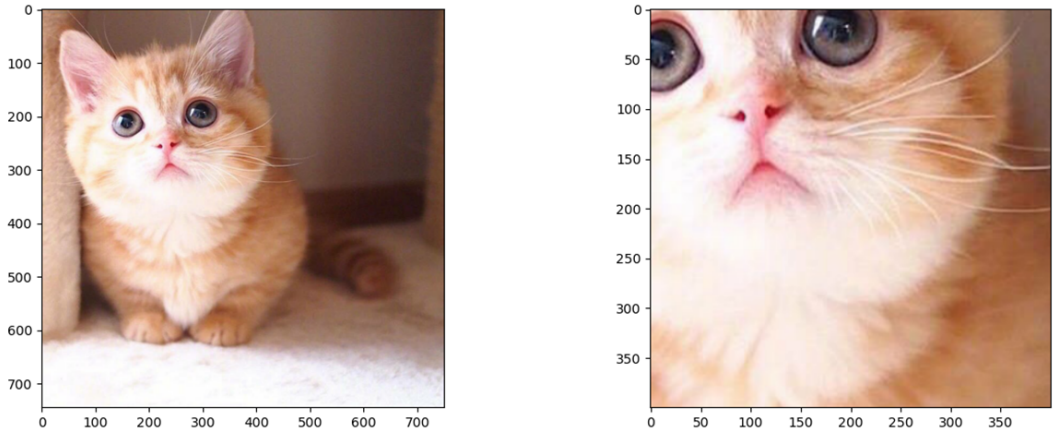
(a) Contrastive learning with an augmentation graph satisfying intra-class connectivity.

(b) Augmentation graph under increasing augmentation strengthes (left to right).

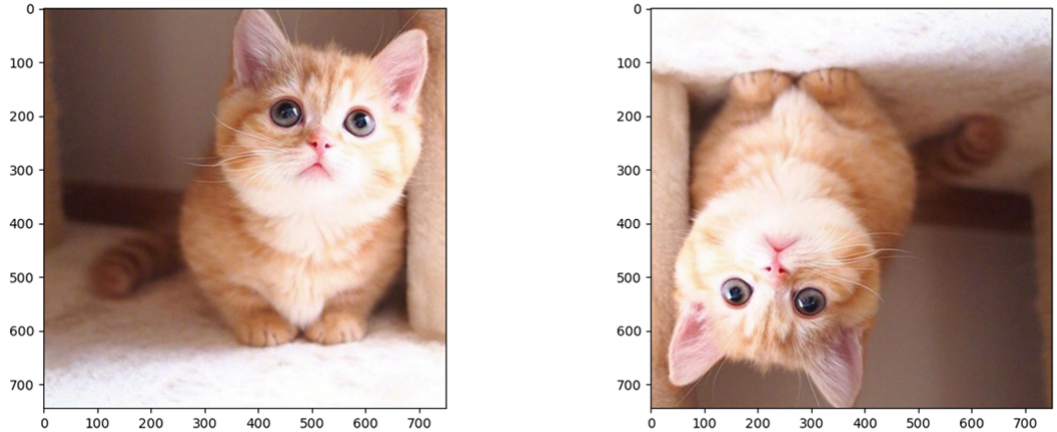
数据增强的作用[12]

[12] Chaos is a Ladder: A New Theoretical Understanding of Contrastive Learning via Augmentation Overlap(ICLR 2022)

常用的图像变换

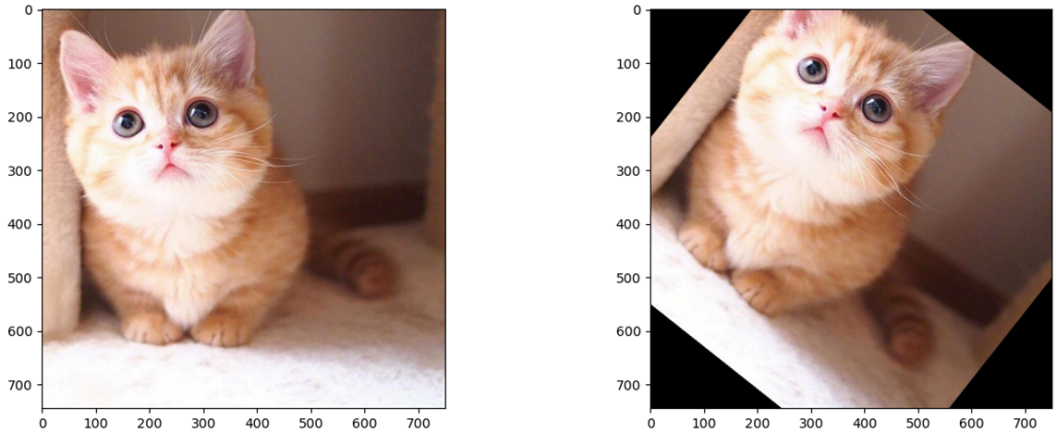


随机裁剪

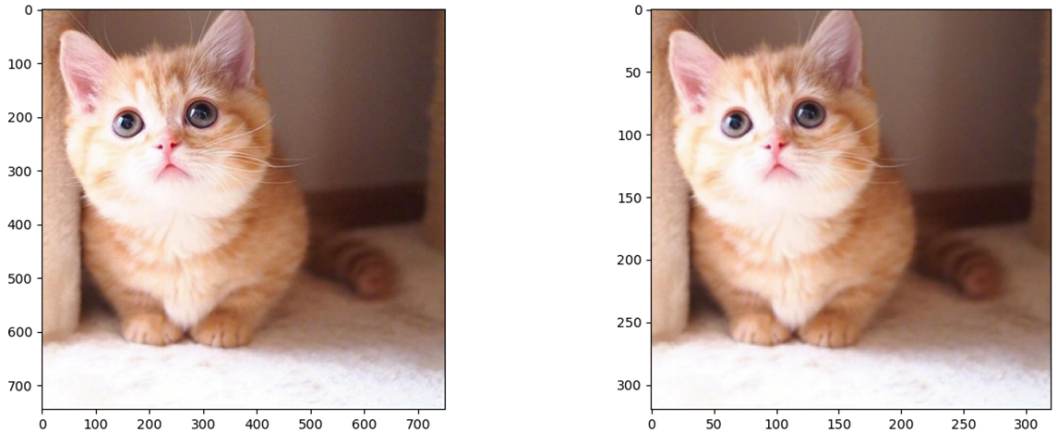


水平和竖直翻转

常用的图像变换

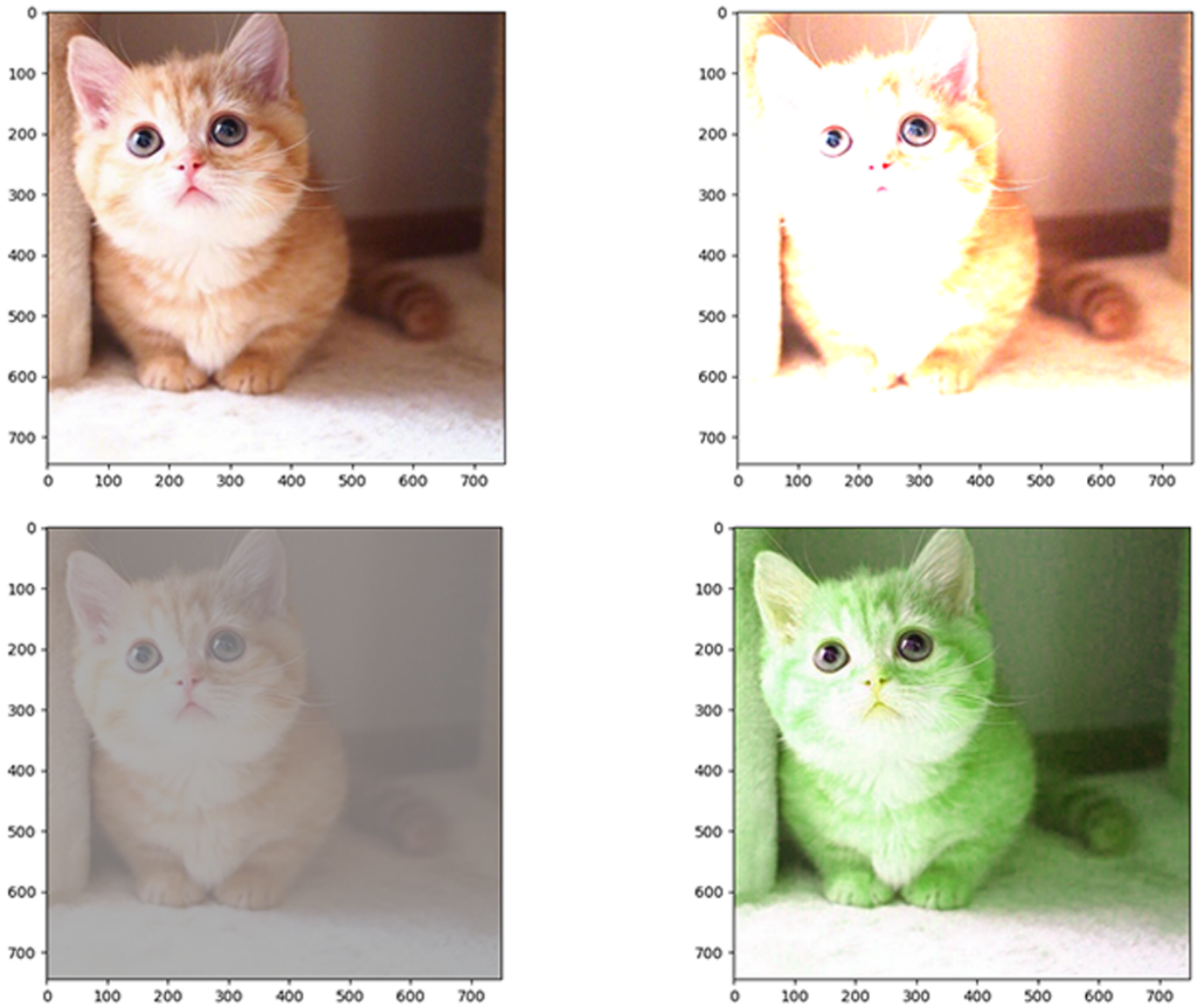


随机旋转



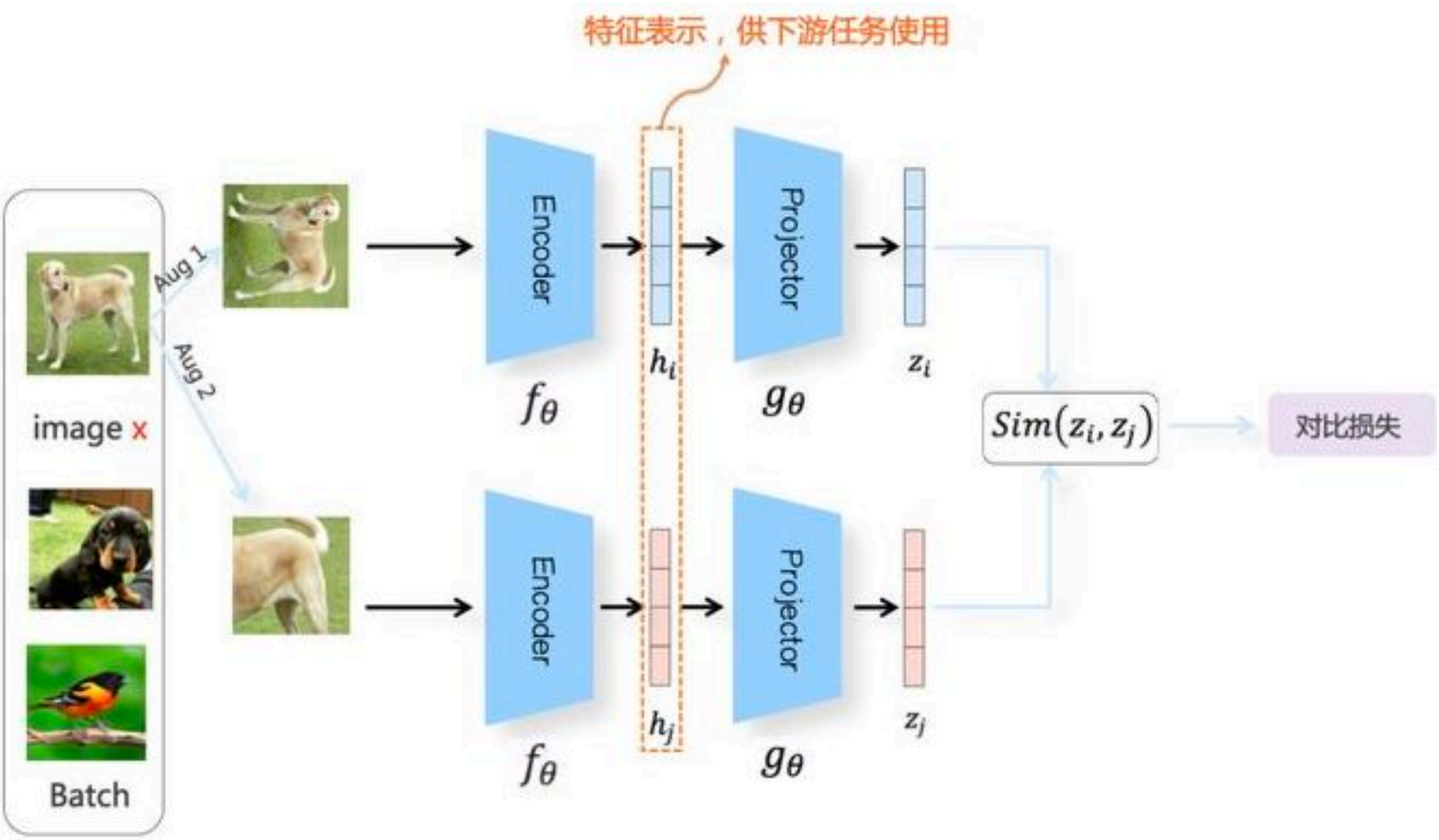
图像缩放

常用的图像变换



亮度、对比度和颜色的随机变换

SimCLR


特征表示，供下游任务使用
image x
Aug 1
Aug 2
Batch
Encoder
Encoder
Projector
Projector
h_i
h_j
z_i
z_j
f_θ
f_θ
g_θ
g_θ
$Sim(z_i, z_j)$
对比损失

SimCLR

样本对的相似度度量:

$$S(z_i, z_j) = \frac{z_i^T z_j}{\|z_i\|_2 \|z_j\|_2}$$

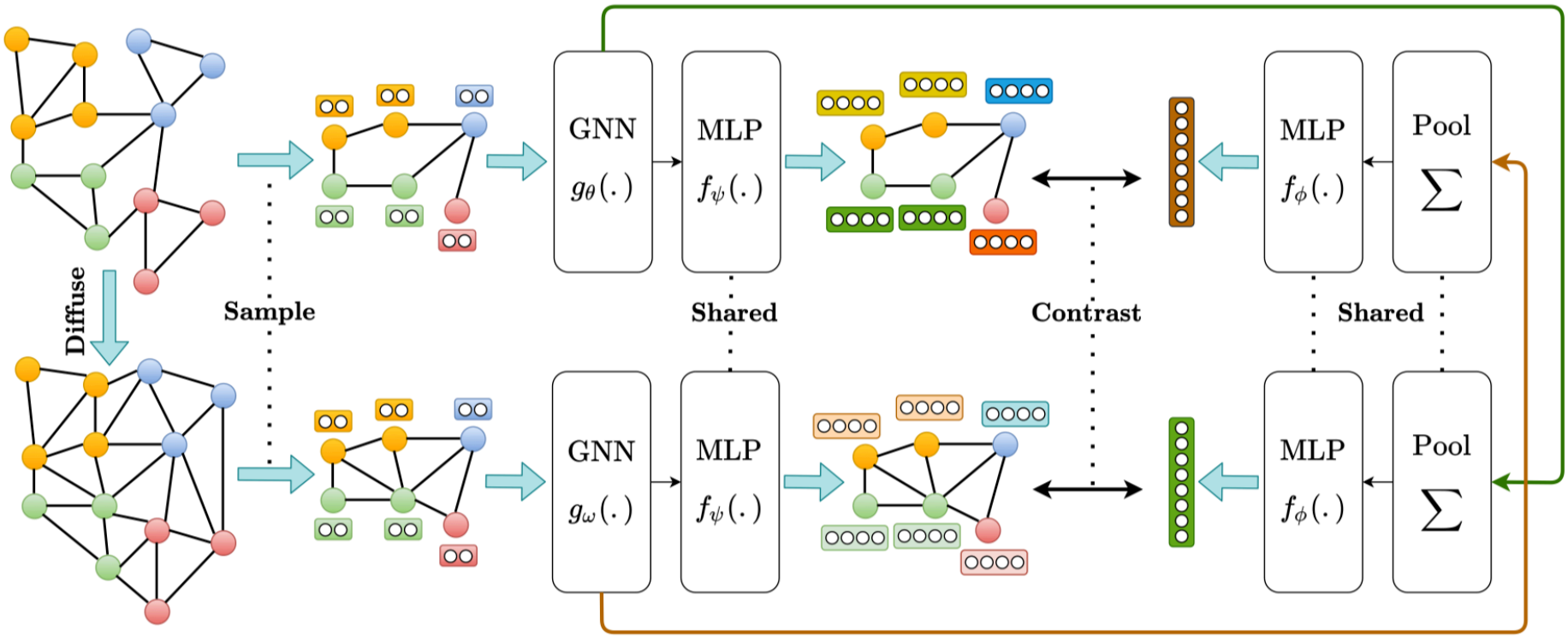
SimCLR 的优化目标：InfoNCE Loss

$$L_i = -\log \frac{\exp(S(z_i, z_i^+)/\tau)}{\sum_{j=1}^{N} \exp(S(z_i, z_j)/\tau)}$$

其中 τ 为温度参数，用于控制整体分布的均匀程度。

分子部分鼓励正样本对间相似度越大越好，分母部分鼓励负样本对之间相似度越小越好。

Contrastive Multi-View Representation Learning on Graphs(MVGRL)



节点子图交叉对比学习

MVGRL 的核心模块

1. Graph Augmentation

$$\mathbf{S} = \sum_{k=0}^{\infty} \Theta_k \mathbf{T}^k \in \mathbb{R}^{n \times n}$$

$$\mathbf{S}^{\text{heat}} = \exp\left(t\mathbf{A}\mathbf{D}^{-1} - t\right)$$

$$\mathbf{S}^{\text{PPR}} = \alpha\left(\mathbf{I}_n - (1-\alpha)\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}\right)^{-1}$$

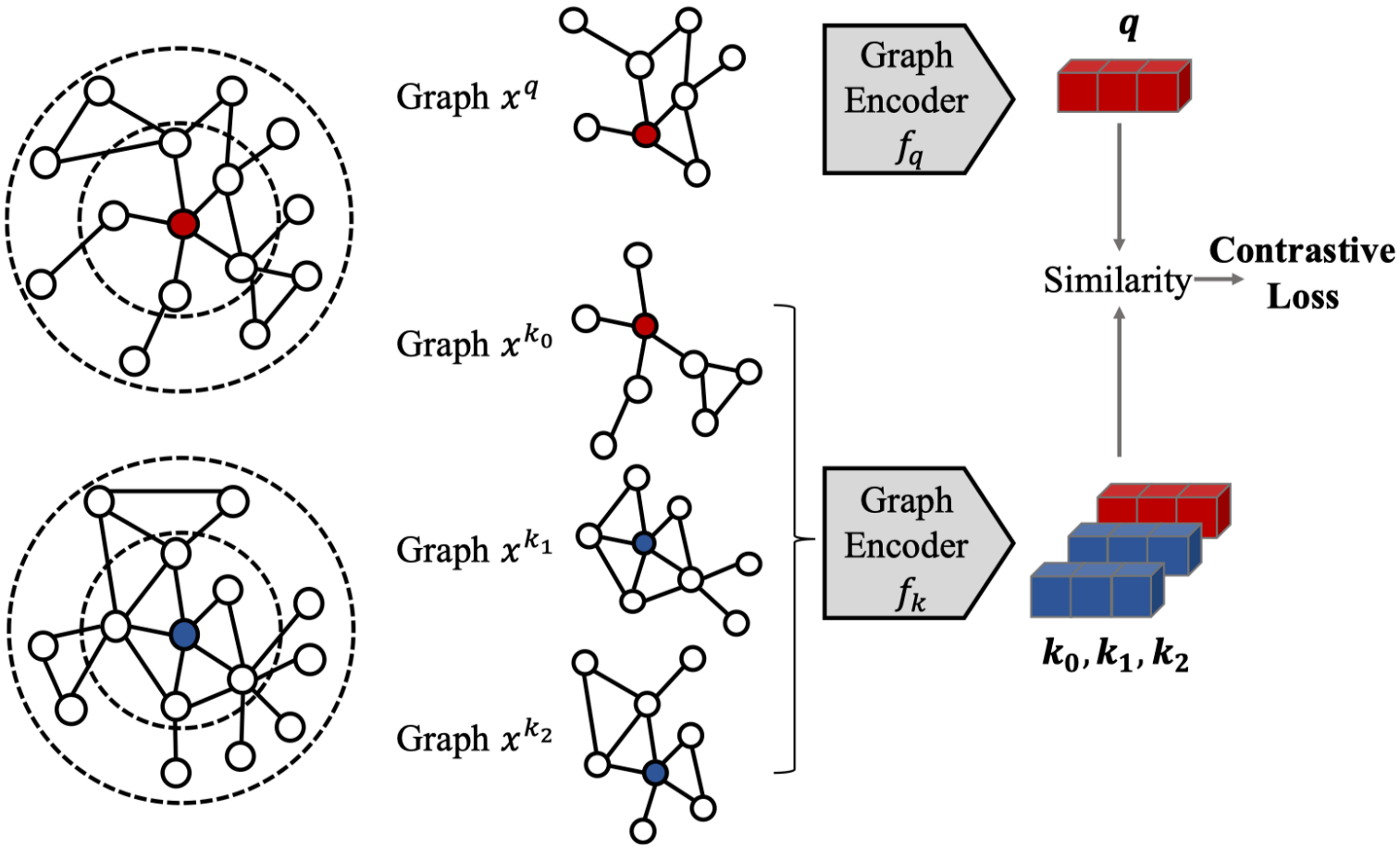
2. JK-Net

$$\vec{h}_g = \sigma\left(\|_{l=1}^{L}\left[\sum_{i=1}^{n}\vec{h}_i^{(l)}\right]\mathbf{W}\right) \in \mathbb{R}^{h_d}$$

3. View Contrastive (Alignment)

$$\max_{\theta,\omega,\phi,\psi} \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}\left[\frac{1}{|g|}\sum_{i=1}^{|g|}\left[\text{MI}\left(\vec{h}_i^{\alpha},\vec{h}_g^{\beta}\right)+\text{MI}\left(\vec{h}_i^{\beta},\vec{h}_g^{\alpha}\right)\right]\right]$$

GCC: Graph Contrastive Coding for Graph Neural Network Pre-Training(KDD 2020)



子图间对比学习

GCC

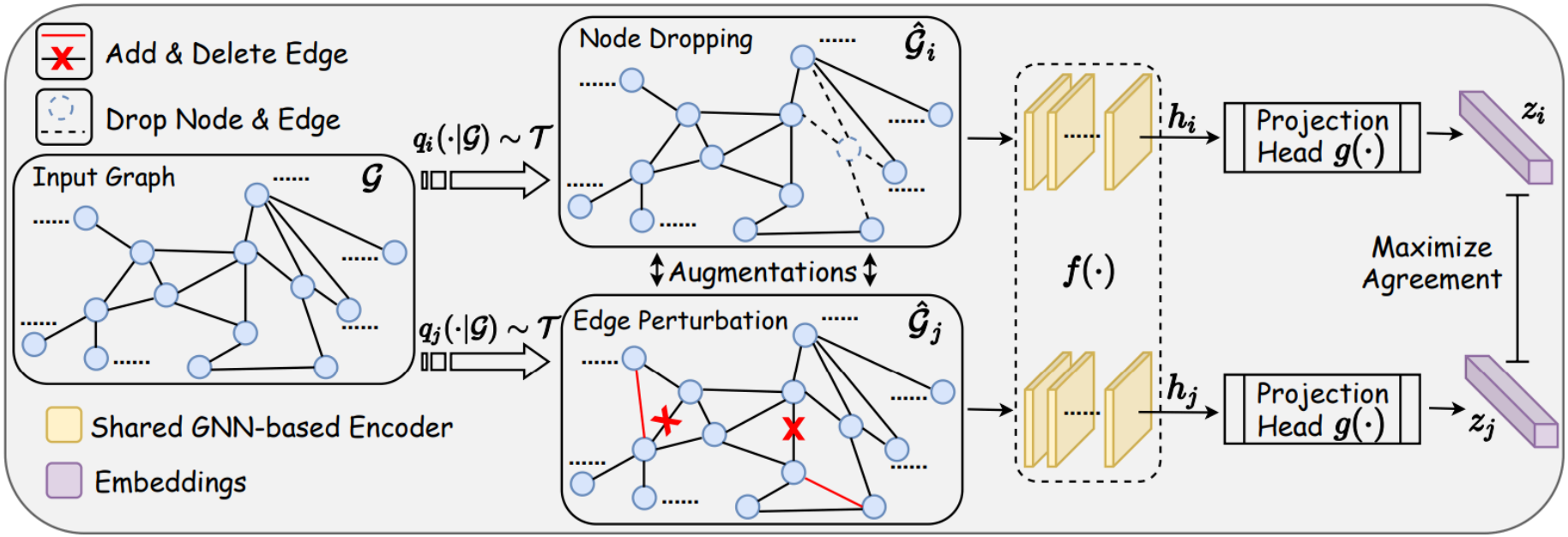
GCC 的核心模块：

1. r-ego subgraph sampling

$$S_v = \{u : d(u, v) \leq r\}$$

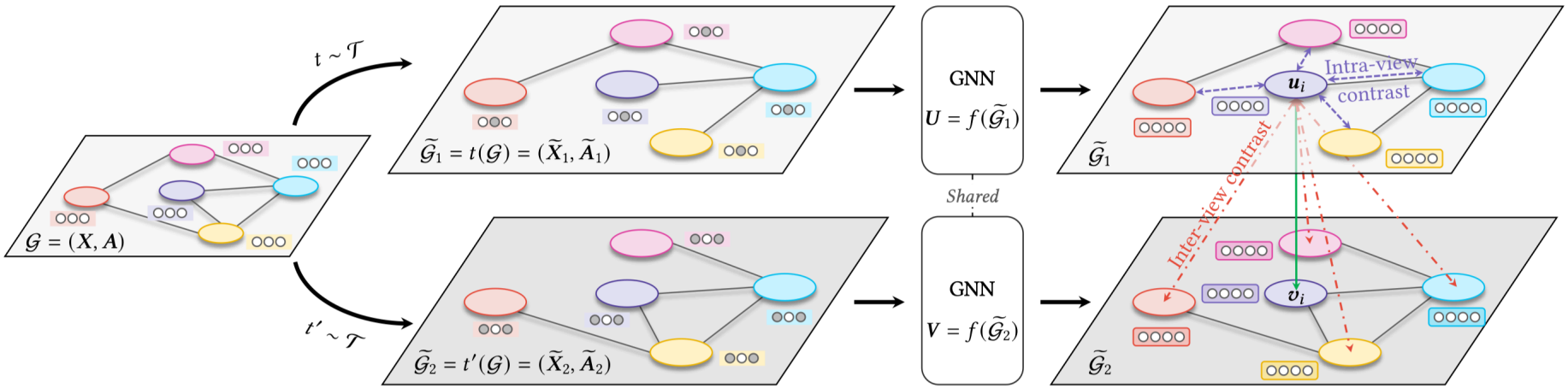
2. Positive/Negative sample generation
   1. Random walks with restart (RWR)
   2. Subgprah induction
   3. Anonymization
3. Generalized position encoding
   1. Eigenvector of normalized graph Laplacian
   2. Node degree Encoding
   3. Binary indicator of the ego vertex

GraphCL



Data augmentation	Type	Underlying Prior
Node dropping	Nodes, edges	Vertex missing does not alter semantics.
Edge perturbation	Edges	Semantic robustness against connectivity variations.
Attribute masking	Nodes	Semantic robustness against losing partial attributes per node.
Subgraph	Nodes, edges	Local structure can hint the full semantics.

Graph Contrastive Learning with Adaptive Augmentation(GCA)



$$\log \frac{e^{\theta(\boldsymbol{u}_i,\boldsymbol{v}_i)/\tau}}{\underbrace{e^{\theta(\boldsymbol{u}_i,\boldsymbol{v}_i)/\tau}}_{\text{positive pair}} + \underbrace{\sum_{k\neq i} e^{\theta(\boldsymbol{u}_i,\boldsymbol{v}_k)/\tau}}_{\text{inter-view negative pairs}} + \underbrace{\sum_{k\neq i} e^{\theta(\boldsymbol{u}_i,\boldsymbol{u}_k)/\tau}}_{\text{intra-view negative pairs}}},$$

总结与展望

图对比学习的主要挑战：

1. 图数据增强 VS 噪声
2. 去数据增强
3. 去负样本
4. 对比对象选择
5. 图数据的全局特性保留

References I

Oord, A. v. d., Li, Y., and Vinyals, O. (2018).
Representation learning with contrastive predictive coding.
arXiv preprint arXiv:1807.03748.