教育部产学合作协同育人项目资助
2022年北京交通大学《深度学习》课程

第十一讲 深度学习前沿与局限

主讲教师：丛润民

目录

目录

注意力机制

➢ 背景

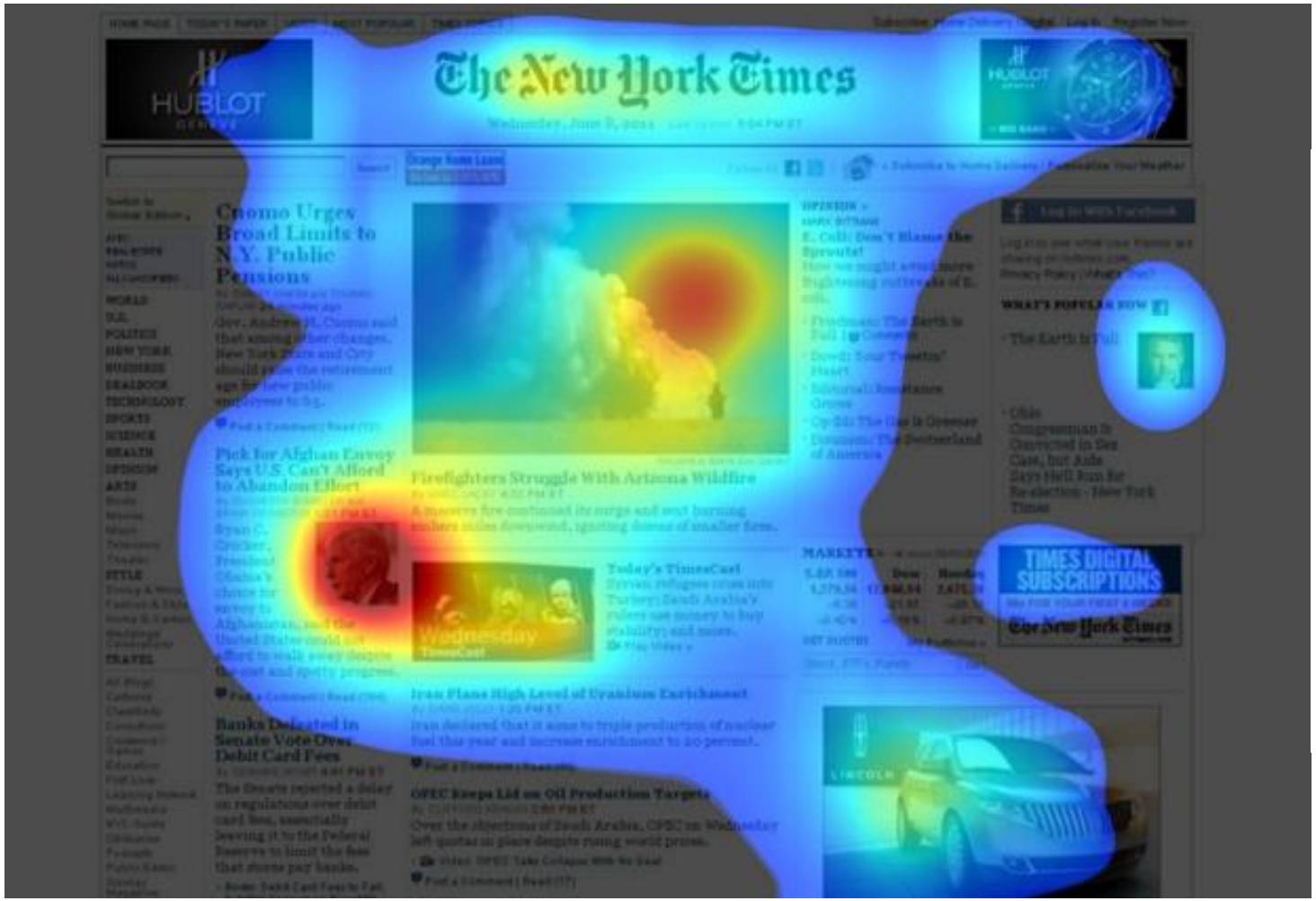
- 前馈、循环网络受到优化算法、计算能力的限制；
- 处理大量输入信息或复杂计算流程时，计算能力成为瓶颈；
- 已有简化神经网络结构的策略：局部连接、权重共享及汇聚操作（pooling）；
- 用循环神经网络解决机器阅读理解；

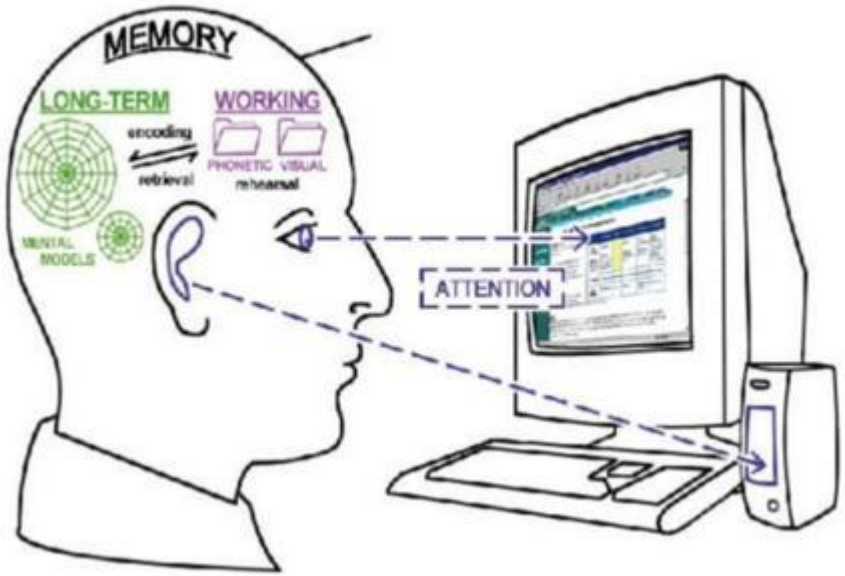
注意力机制

➢ 大脑中的注意力

- 人脑每个时刻接收的外界输入信息非常多，包括来源于视觉、听觉、触觉各种各样的信息。
- 就视觉来说，眼睛每秒钟都会发送千万比特的信息给视觉神经系统。
- 人脑通过注意力来解决信息超载问题（记忆、阅读、思考）。
- 自上而下的聚焦式，选择性注意力
- 自下而上的汇聚式，基于显著性的注意力

注意力举例


The New York Times
HUBLOT
Cuomo Urges Broad Limits to N.Y. Public Pensions
Firefighters Struggle With Arizona Wildfire
Pick for Afghan Envoy Says U.S. Can't Afford to Abandon Effort
Banks Defeated in Senate Vote Over Debit Card Fees
Iran Plans High Level of Uranium Enrichment
OPEC Keeps Lid on Oil Production Targets
TIMES DIGITAL SUBSCRIPTIONS


MEMORY
LONG-TERM
WORKING
encoding
retrieval
PHONETIC
VISUAL
rehearsal
MENTAL MODELS
ATTENTION

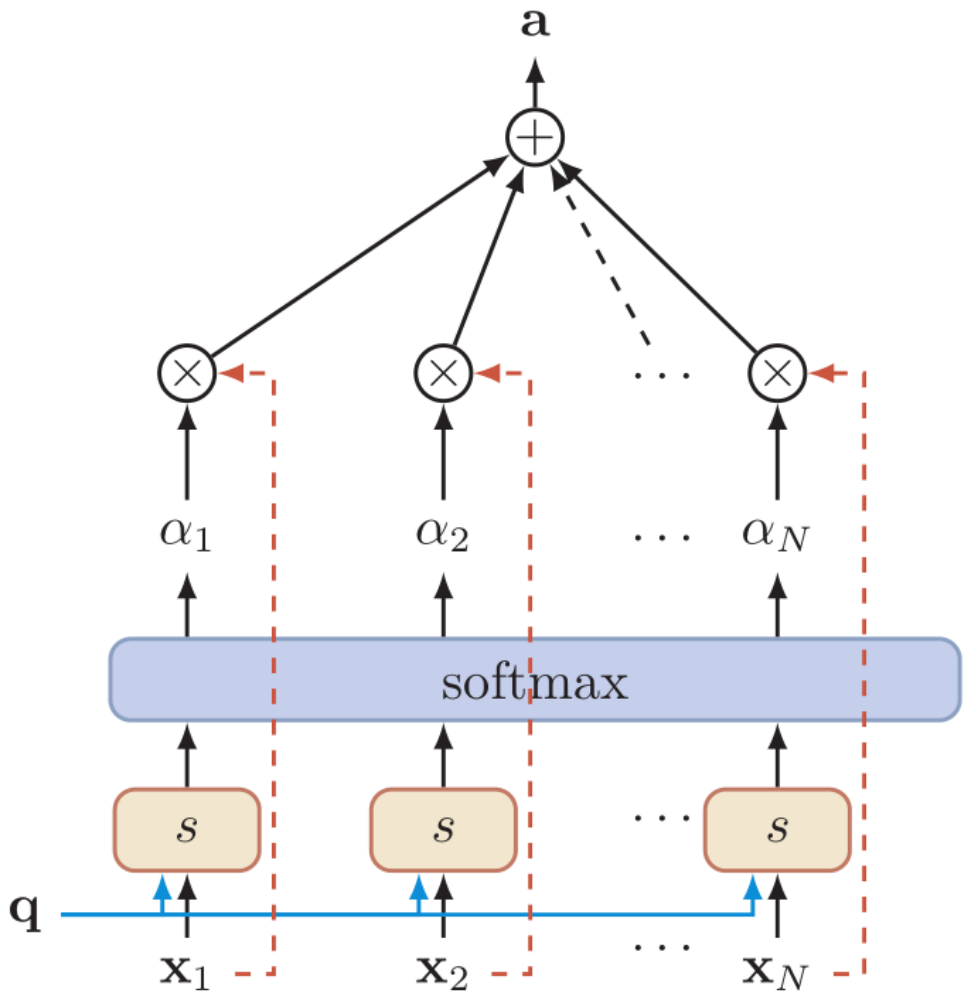
注意力机制

➢ 注意力分布

■给定查询q和输入信息$x_{1:N}$，则

$$\begin{aligned}\alpha_i &= p(z=i|\mathbf{x}_{1:N},\mathbf{q}) \\ &= \text{softmax}\left(s(\mathbf{x}_i,\mathbf{q})\right) \\ &= \frac{\exp\left(s(\mathbf{x}_i,\mathbf{q})\right)}{\sum_{j=1}^{N}\exp\left(s(\mathbf{x}_j,\mathbf{q})\right)},\end{aligned}$$

$s(\boldsymbol{x},\boldsymbol{q})$为注意力打分函数



➢ 注意力打分函数

加性模型

$$s(\boldsymbol{x}, \boldsymbol{q}) = \boldsymbol{v}^{\mathrm{T}} \tanh(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{U}\boldsymbol{q}),$$

点积模型

$$s(\boldsymbol{x}, \boldsymbol{q}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{q},$$

缩放点积模型

$$s(\boldsymbol{x}, \boldsymbol{q}) = \frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{q}}{\sqrt{D}},$$

双线性模型

$$s(\boldsymbol{x}, \boldsymbol{q}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{q},$$

注意力机制

➢ 硬注意力

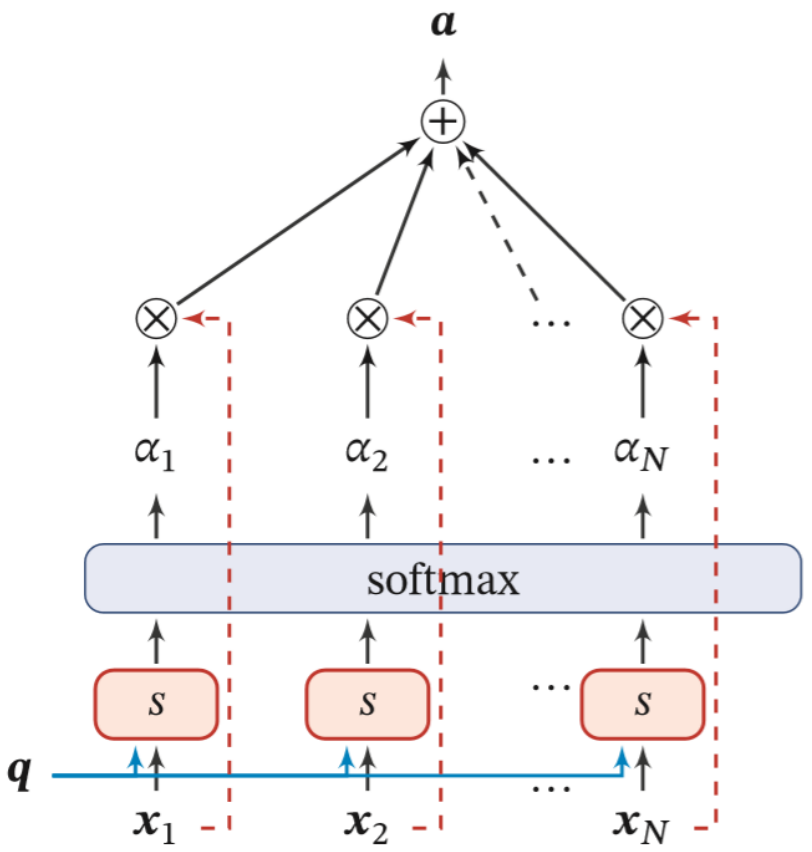
$$\text{att}(\boldsymbol{X}, \boldsymbol{q}) = \boldsymbol{x}_{\hat{n}},$$

$$\hat{n} = \arg\max_{n=1}^{N} \alpha_n.$$

➢ 软注意力

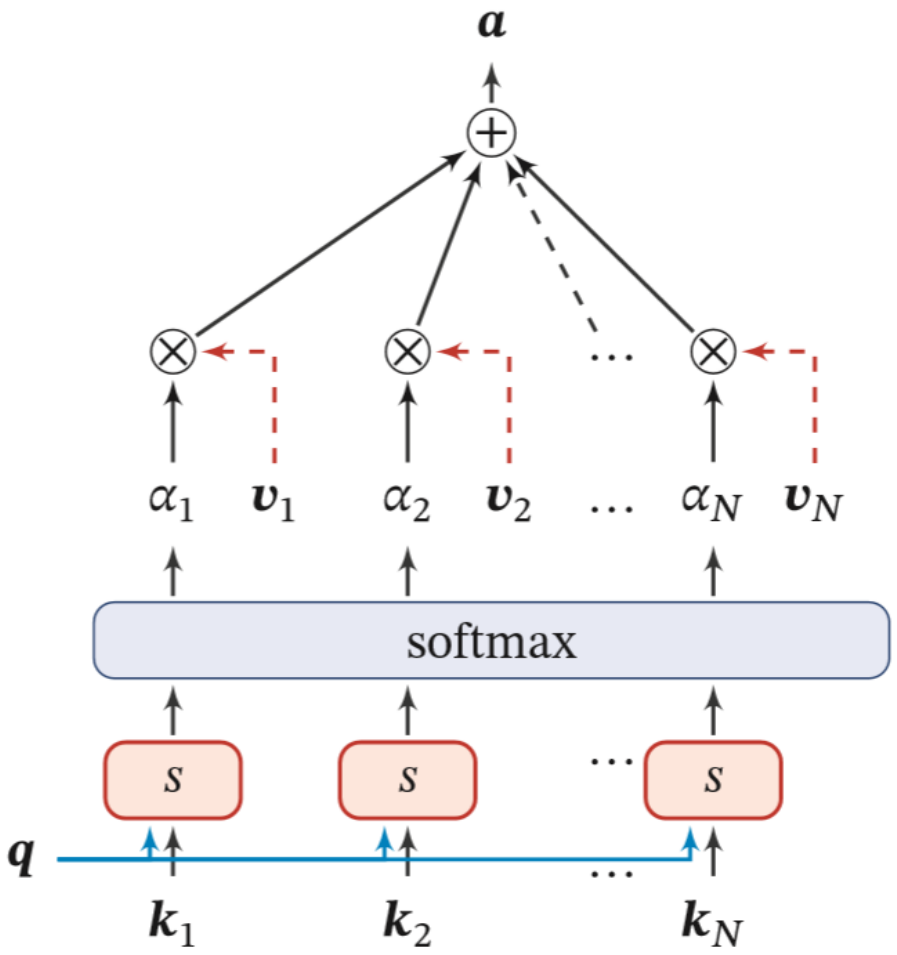
$$\text{att}(\boldsymbol{X}, \boldsymbol{q}) = \sum_{n=1}^{N} \alpha_n \boldsymbol{x}_n,$$

$$= \mathbb{E}_{z \sim p(z|\boldsymbol{X}, \boldsymbol{q})}[\boldsymbol{x}_z].$$



➢ 键值对注意力

$$\begin{aligned}\text{att}\Big((\boldsymbol{K},\boldsymbol{V}),\boldsymbol{q}\Big) &= \sum_{n=1}^{N}\alpha_n\boldsymbol{v}_n,\\ &= \sum_{n=1}^{N}\frac{\exp\Big(s(\boldsymbol{k}_n,\boldsymbol{q})\Big)}{\sum_j \exp\Big(s(\boldsymbol{k}_j,\boldsymbol{q})\Big)}\boldsymbol{v}_n,\end{aligned}$$



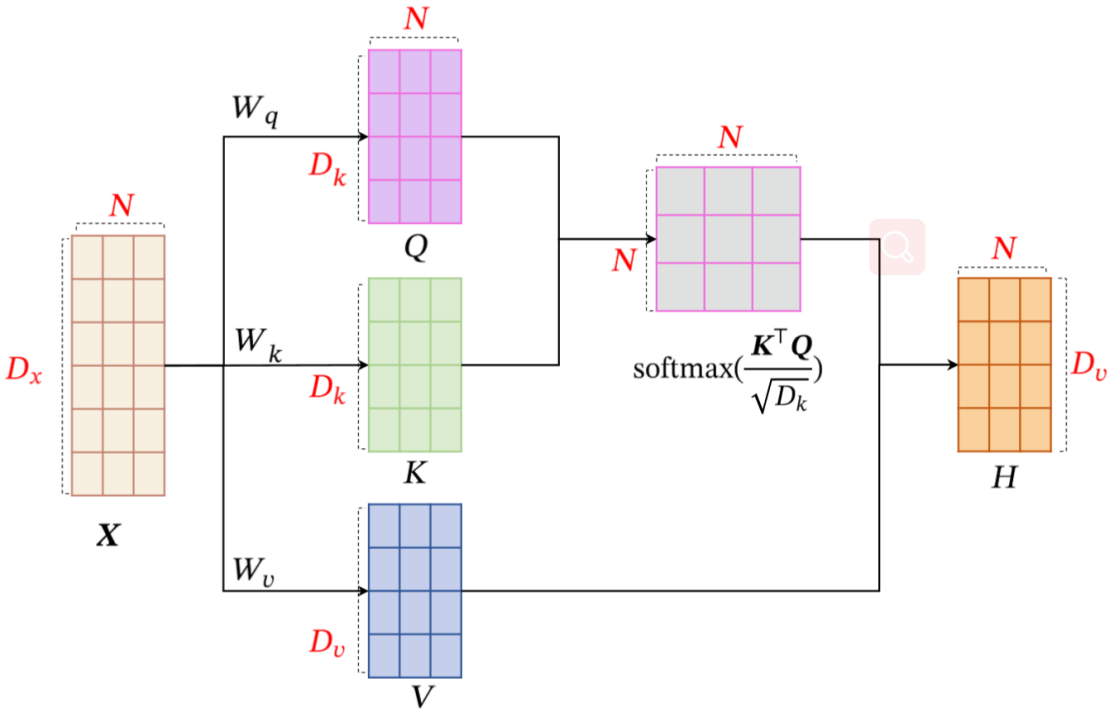
➢ 自注意力：查询-键-值（Query-Key-Value，QKV）

■线性映射

$$\boldsymbol{Q} = \boldsymbol{W}_q\boldsymbol{X} \in \mathbb{R}^{D_k \times N},$$

$$\boldsymbol{K} = \boldsymbol{W}_k\boldsymbol{X} \in \mathbb{R}^{D_k \times N},$$

$$\boldsymbol{V} = \boldsymbol{W}_v\boldsymbol{X} \in \mathbb{R}^{D_v \times N},$$

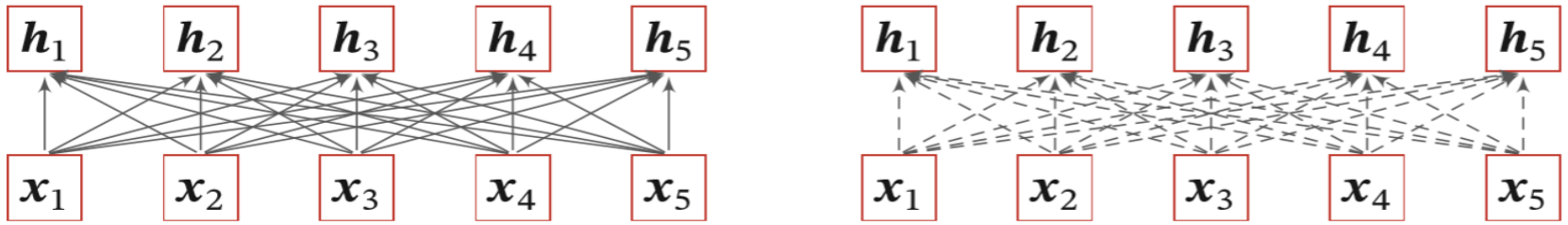


■输出向量

$$\boldsymbol{h}_n = \text{att}\Big((\boldsymbol{K}, \boldsymbol{V}), \boldsymbol{q}_n\Big) = \sum_{j=1}^{N} \alpha_{nj}\boldsymbol{v}_j = \sum_{j=1}^{N} \text{softmax}\big(s(\boldsymbol{k}_j, \boldsymbol{q}_n)\big)\boldsymbol{v}_j,$$

➢ 自注意力

- 自注意力模型的权重是动态生成，因此可以处理变长的信息序列



(a) 全连接模型　　(b) 自注意力模型

- 自注意力模型可以作为神经网络中的一层来使用，既可以用来替换卷积层和循环层，也可以和它们一起交替使用。

Vaswani A, Shazeer N, Parmar N, et al., 2017. Attention is all you need[C]//Advances in NIPS. 6000-6011.

注意力机制

注意力机制变种

1、多头注意力 Multi-Head Attention

2、硬注意力 Hard Attention

3、结构化注意力 Structure Attention

4、指针网络 Pointer Network

5、双向注意力 Bi-Directional Attention

6、键值对注意力 Key-Value Attention

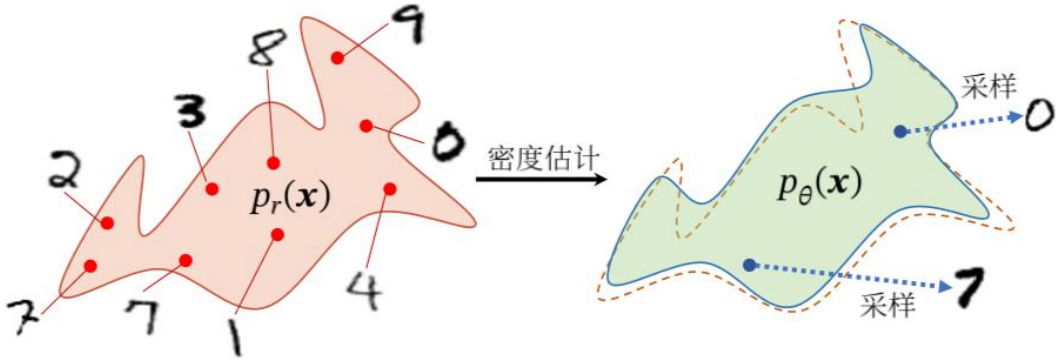
7、自注意力 Self/Intra Attention

…

目录

➢ 背景--概率生成模型

- 生成模型指一系列用于随机生成可观测数据的模型。
- 生成模型包含：概率密度估计、样本生成（即采样）



➢ 概率密度估计

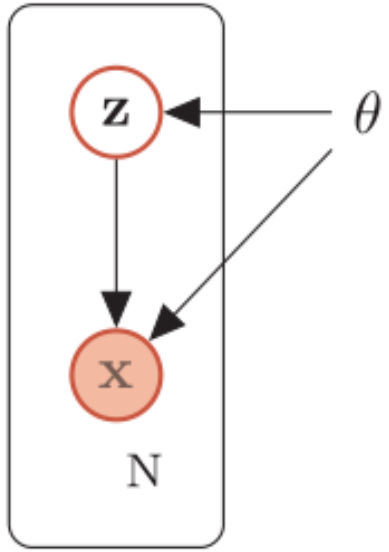
■对于由独立同分布概率密度函数$p_r(x)$产生的一组数据$D = \{x^{(n)}\}_{n=1}^{N}$，

密度估计是根据数据集D来估计密度函数$p_r(x)$。

➢ 样本生成

■生成数据x的过程分为两步：

- 根据隐变量的先验分布$p(z;\theta)$进行采样，得到样本z；
- 根据条件分布$p(x|z;\theta)$进行采样，得到x。



深度生成模型

- 深度生成模型就是利用神经网络来建模条件分布$p(x|z;\theta)$或直接生成符合分布的样本。
  - 变分自编码器（Variational Autoencoder，VAE） 自学

    [Kingma and Welling, 2013, Rezende et al., 2014]。
  - 对抗生成式网络（Generative Adversarial Network，GAN）

    [Goodfellow et al., 2014]

EM算法回顾

- 最大期望算法（Expectation-maximization algorithm）是在概率模型中寻找参数最大似然估计或者最大后验估计的算法，其中概率模型依赖于无法观测的隐性变量。
- 最大期望算法经过两个步骤交替进行计算：第一步是计算期望（E），利用对隐藏变量的现有估计值，计算其最大似然估计值；第二步是最大化（M），最大化在E步上求得的最大似然值来计算参数的值。M步上找到的参数估计值被用于下一个E步计算中，这个过程不断交替进行。

EM算法回顾

给定一个样本x，通过引入变分密度函数$q(\mathbf{z};\phi)$，其对数边际似然$\log p(\mathbf{x}|\theta)$可分解为：

$$\log p(\mathbf{x}|\theta) = \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x},\mathbf{z}|\theta)}{q(\mathbf{z})} - \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{z}|\mathbf{x},\theta))}{q(\mathbf{z})}$$

$$= ELBO(q,\mathbf{x}|\theta) + D_{\mathrm{KL}}(q(\mathbf{z})\|p(\mathbf{z}|\mathbf{x},\theta)),$$

M step　　E step

➢ 网络结构

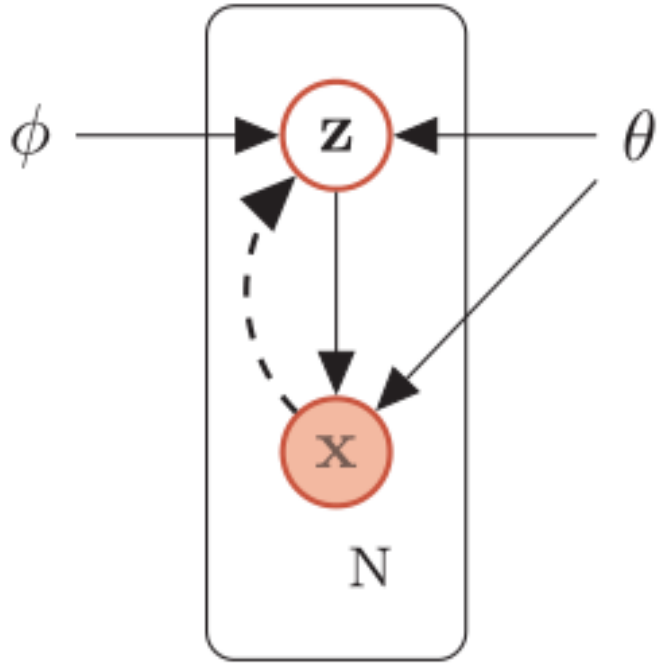
- 推断网络：

  寻找后验分布$p(z|x;\theta)$的变分近似$q(z|x;\phi)$；

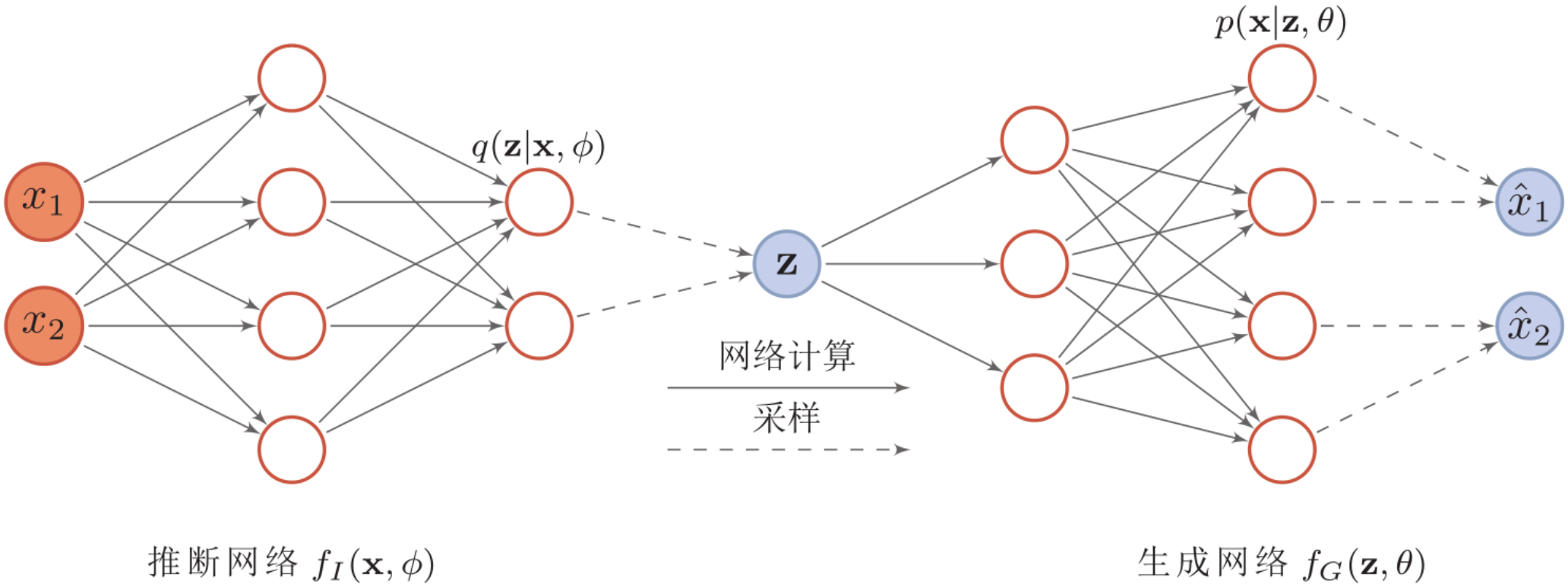
  变分推断：用简单分布q近似复杂分布p。

- 生成网络

  已知$q(z|x;\phi)$情况下，估计更好的生成$p(x|z;\theta)$。



变分自编码器


$p(\mathbf{x}|\mathbf{z},\theta)$
$q(\mathbf{z}|\mathbf{x},\phi)$
x_1
x_2
$\mathbf{z}$
$\hat{x}_1$
$\hat{x}_2$
网络计算
采样
推断网络 $f_I(\mathbf{x},\phi)$
生成网络 $f_G(\mathbf{z},\theta)$

变分自编码器

➢ 目标函数

$$\max_{\theta,\phi} ELBO(q, \mathbf{x}|\theta, \phi) = \max_{\theta,\phi} \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\phi)} \left[\log \frac{p(\mathbf{x}|\mathbf{z}, \theta)p(\mathbf{z}|\theta)}{q(\mathbf{z}|\phi)} \right]$$

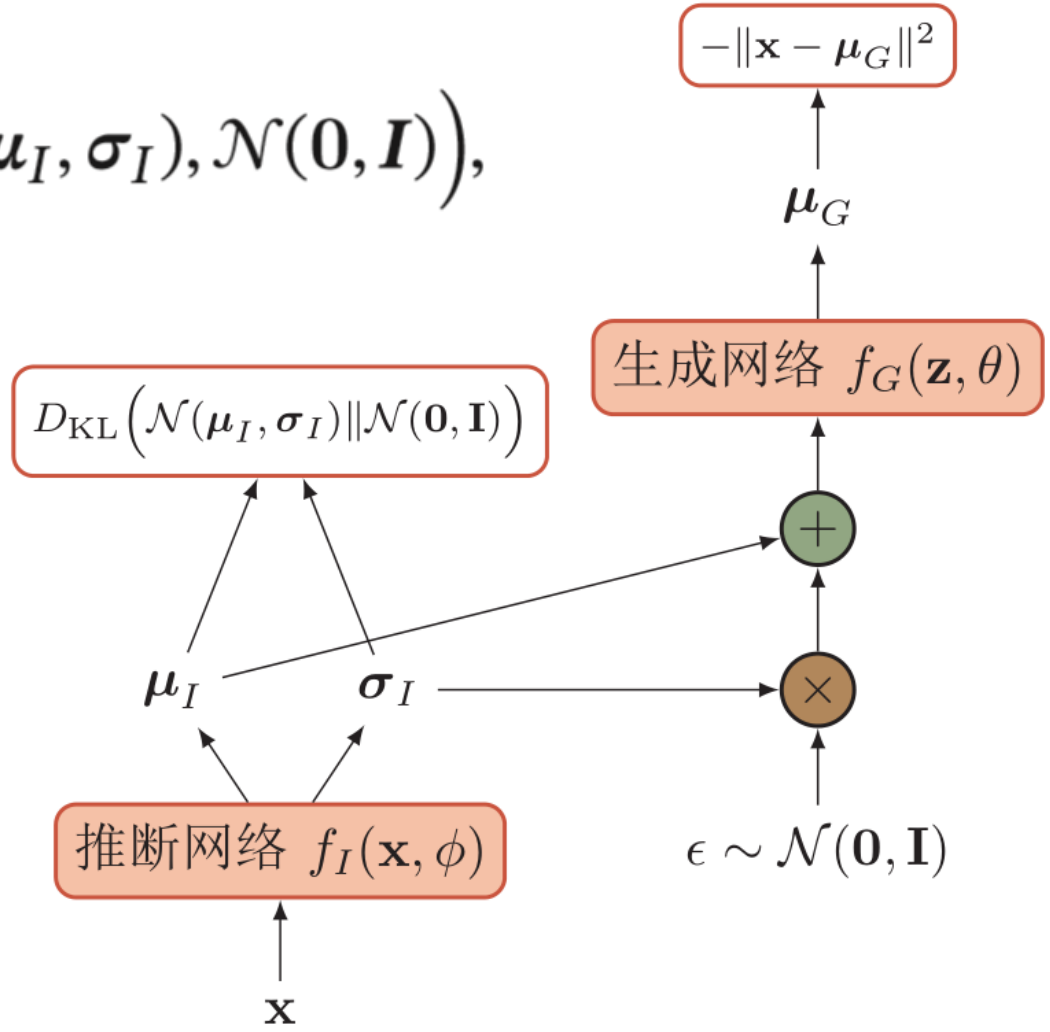
$$\max_{\theta,\phi} \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x},\phi)} \Big[\log p(\mathbf{x}|\mathbf{z}, \theta) \Big] - D_{\mathrm{KL}} \Big(q(\mathbf{z}|\mathbf{x}, \phi) || p(\mathbf{z}|\theta) \Big)$$

变分自编码器

➢ 训练过程

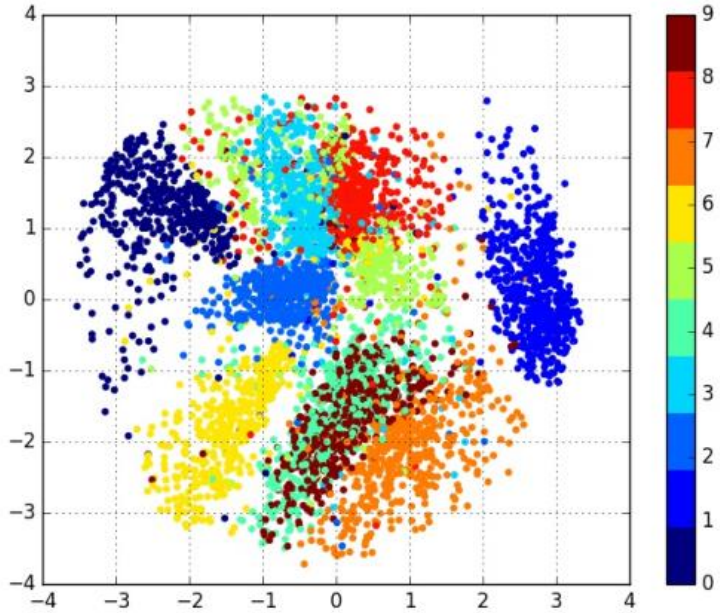
$$\mathcal{J}(\phi,\theta|\boldsymbol{x}) = -\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{\mu}_G\|^2 + \lambda\,\mathrm{KL}\Big(\mathcal{N}(\boldsymbol{\mu}_I,\boldsymbol{\sigma}_I),\mathcal{N}(\boldsymbol{0},\boldsymbol{I})\Big),$$

如果采用随机梯度方法，每次从数据集中采集一个样本$\boldsymbol{x}$和一个对应的随机变量 ϵ，并进一步假设$p(\boldsymbol{x}|\boldsymbol{z};\theta)$服从高斯分布$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_G,\lambda\boldsymbol{I})$，其中$\boldsymbol{\mu}_G=f_G(\boldsymbol{z};\theta)$是生成网络的输出，其中第一项可以近似看作输入$\boldsymbol{x}$的重构正确性，第二项可以看作正则化项，$\lambda$可以看作正则化系数 。这和自编码器在形式上非常类似，但它们的内在机理是完全不同的。

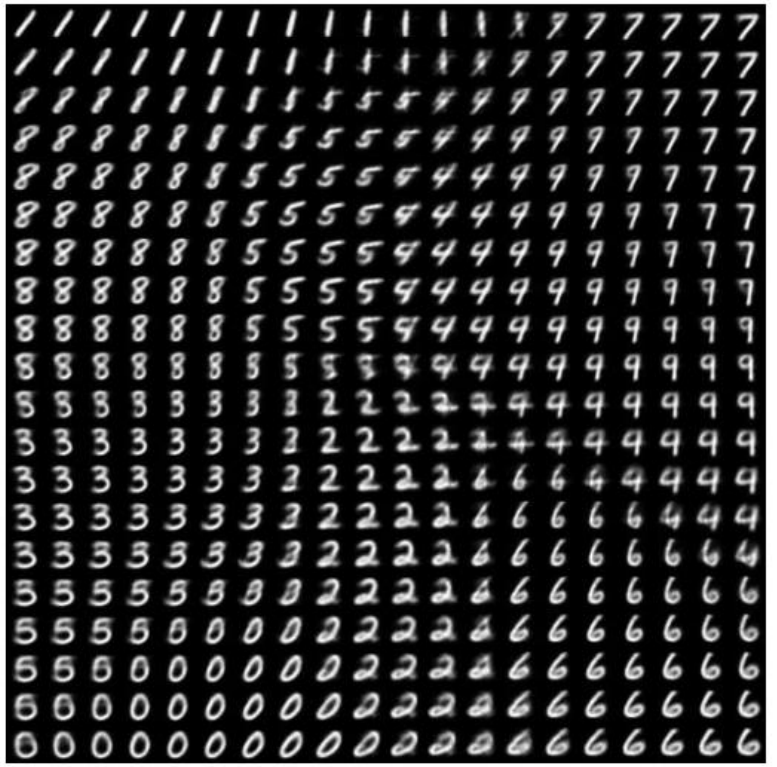


变分自编码器

➢ 学到的隐变量流形



(a) 训练集上所有样本在隐空间上的投影。

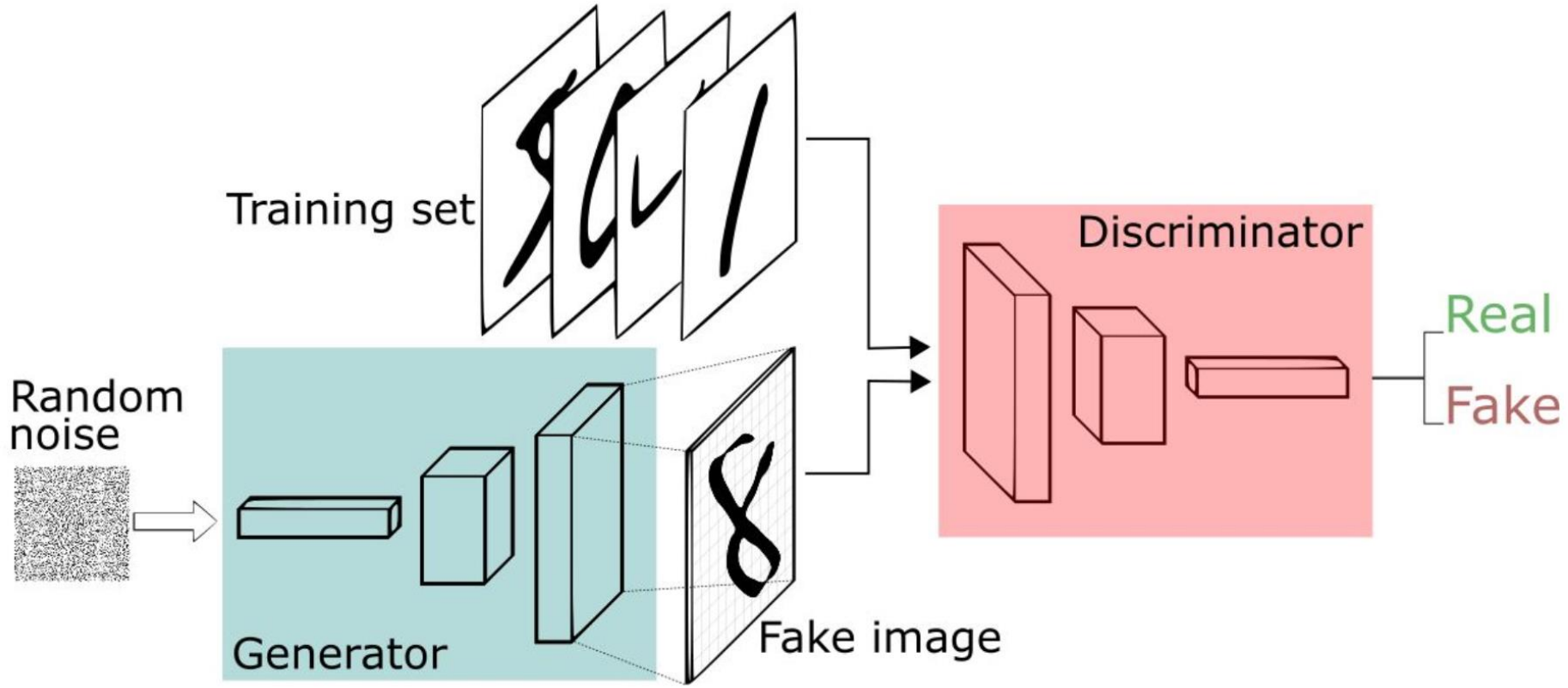


(b) 隐变量 $\mathbf{z}$ 在图像空间的投影。

生成对抗网络

- 生成对抗网络：生成网络+判别网络。
- 生成网络：输入为从潜在空间中随机采样，其输出结果需要尽量模仿训练集中的真实样本。
- 判别网络：输入为真实样本、生成网络输出的假样本，其目的尽可能分辨出真实样本与假样本。

生成对抗网络



➢ 目标函数

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1 - D(G(z)))]$$

https://zhuanlan.zhihu.com/p/33752313

生成对抗网络

1、DCGAN

2、LSGAN

3、WGAN

4、EBGAN

5、InfoGAN

6、Conditional GAN

7、Cycle GAN, Disco GAN

…

Year	Month	Abbr.	Title
2014	6	GAN	Generative Adversarial Networks
2014	11	CGAN	Conditional Generative Adversarial Nets
2015	6	LAPGAN	Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks
2015	11	CatGAN	Unsupervised and Semi-supervised Learning with Categorical Generative Adversarial Networks
2015	11	DCGAN	Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks
2015	12	VAE-GAN	Autoencoding beyond pixels using a learned similarity metric
2016	2	GRAN	Generating images with recurrent adversarial networks
2016	3	S^2GAN	Generative Image Modeling using Style and Structure Adversarial Networks
2016	4	MGAN	Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks
2016	5	BiGAN	Adversarial Feature Learning

494	2018	9	GcGAN	Geometry-Consistent Adversarial Networks for One-Sided Unsupervised Domain Mapping
495	2018	9	GraphSGAN	Semi-supervised Learning on Graphs with Generative Adversarial Nets
496	2018	9	IGMM-GAN	Coupled IGMM-GANs for deep multimodal anomaly detection in human mobility data
497	2018	9	MeRGAN	Memory Replay GANs: learning to generate images from new categories without forgetting
498	2018	9	SAM	Sample-Efficient Imitation Learning via Generative Adversarial Nets
499	2018	9	SiftingGAN	SiftingGAN: Generating and Sifting Labeled Samples to Improve the Remote Sensing Image Scene Classification Baseline in vitr
500	2018	9	SLSR	Sparse Label Smoothing for Semi-supervised Person Re-Identification
501	2018	9	Twin-GAN	Twin-GAN -- Unpaired Cross-Domain Image Translation with Weight-Sharing GANs
502	2018	9	WaveletGLCA-GAN	Global and Local Consistent Wavelet-domain Age Synthesis

https://github.com/hindupuravinash/the-gan-zoo

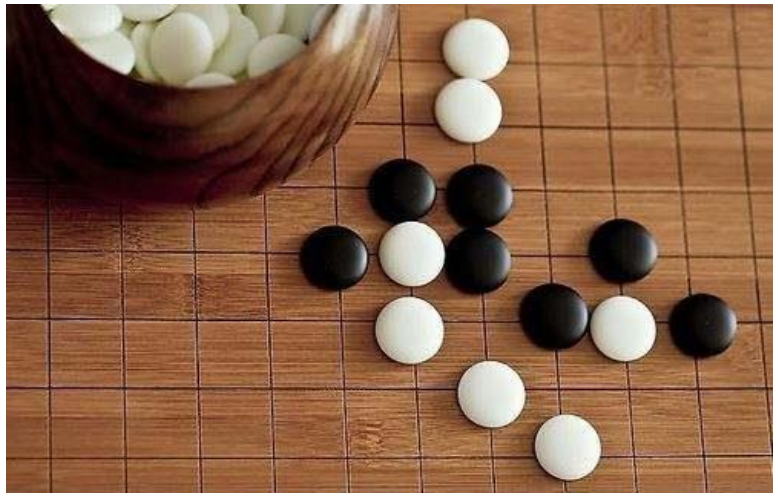
目录

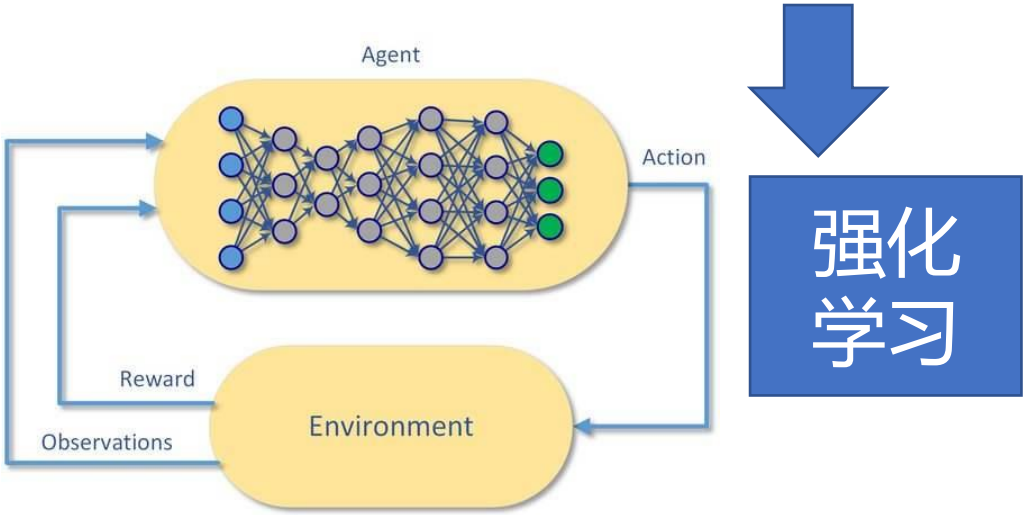
背景

- 一些应用场景中，通过人工标注给数据打标签很难；
- 通过训练监督学习模型自动下围棋，难点：
  - 对于每一种棋盘状态，即便专家也很难给出“正确”的动作；
  - 获取大量数据的成本较高；

➢ 背景

- 在下棋任务中，虽然很难知道每一步的“正确”动作，但其最后结果（赢输）却很容易判断；
- 若通过大量的模拟数据，通过最后结果（奖励）来倒推每一步棋的好坏，从而学习出“最佳”的下棋策略；

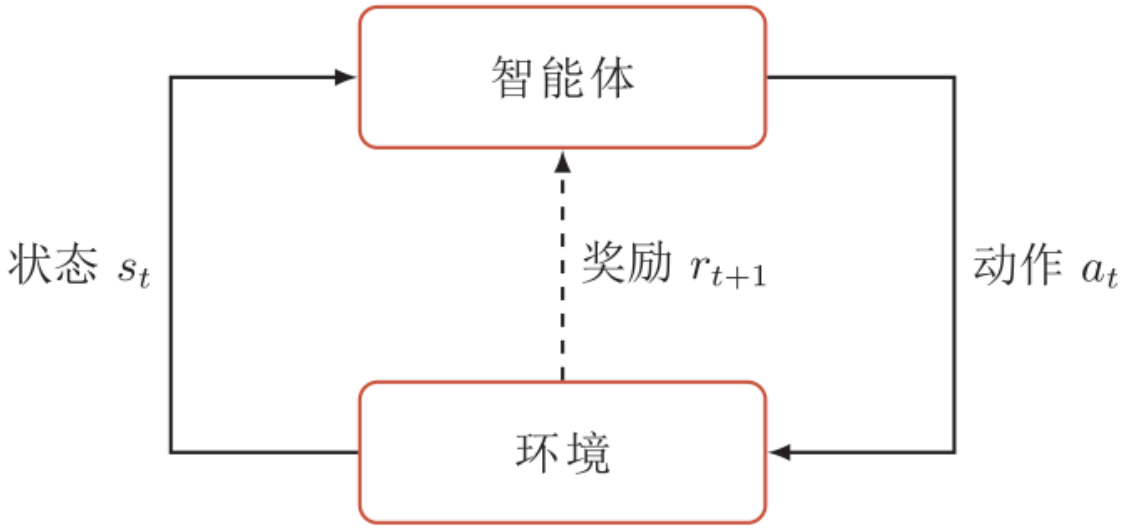




强化学习

- 强化学习问题可以描述为一个智能体从与环境的交互中不断学习以完成特定目标（比如取得最大奖励值）。
- 强化学习就是智能体不断与环境进行交互，并根据经验调整其策略来最大化其长远的所有奖励的累积值。

马尔可夫决策过程

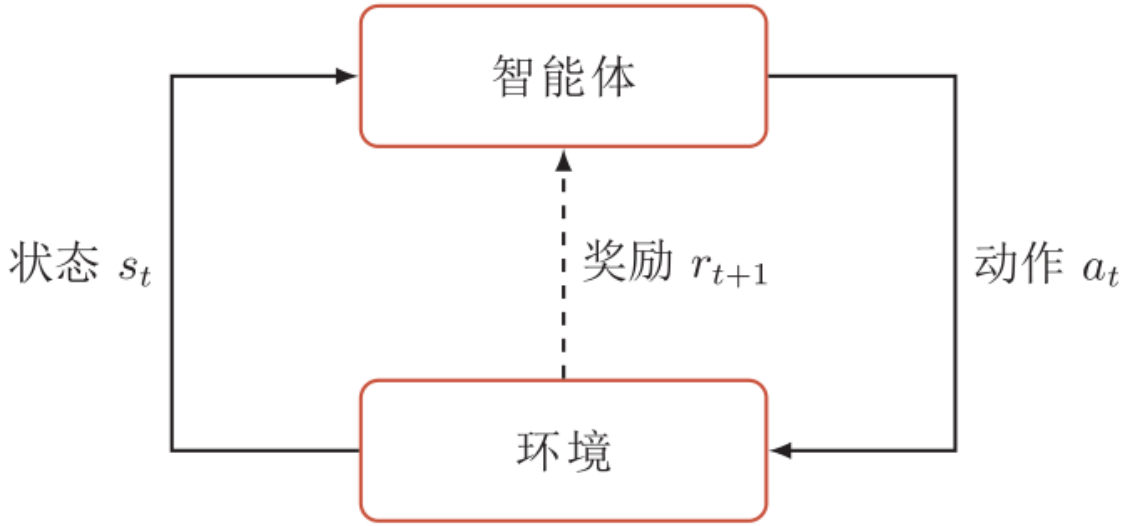


$$s_0, a_0, s_1, r_1, a_1, \cdots, s_{t-1}, r_{t-1}, a_{t-1}, s_t, r_t, \cdots,$$

马尔可夫决策过程

- **马尔可夫过程**

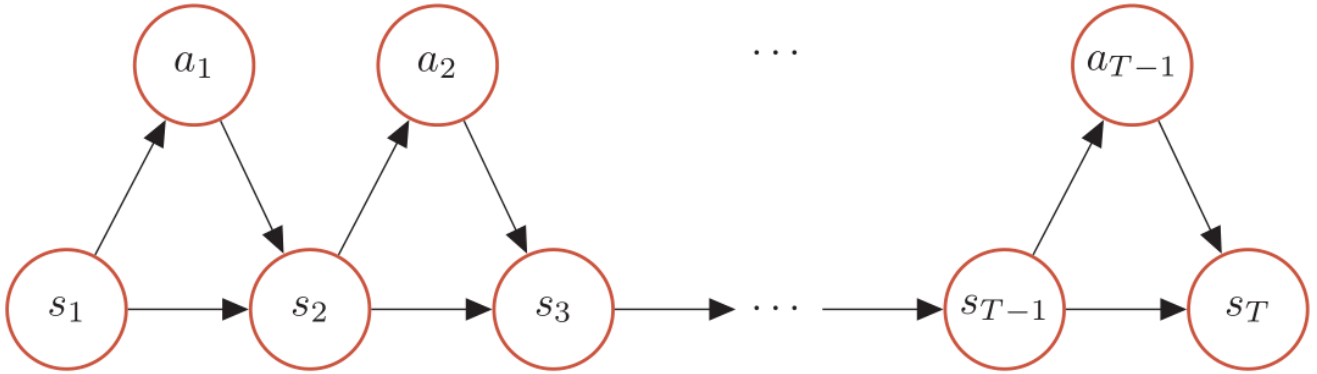
$$p(s_{t+1}|s_t, \cdots, s_0) = p(s_{t+1}|s_t),$$



- **马尔可夫决策过程**

$$p(s_{t+1}|s_t, a_t, \cdots, s_0, a_0) = p(s_{t+1}|s_t, a_t),$$

➢ 马尔可夫决策过程



- 给定策略π(a|s)，马尔可夫决策过程的一个轨迹（trajectory）

$$\tau = s_0, a_0, s_1, r_1, a_1, \cdots, s_{T-1}, a_{T-1}, s_T, r_T$$ 的概率

$$p(\tau) = p(s_0, a_0, s_1, a_1, \cdots) = p(s_0) \prod_{t=0}^{T-1} \pi(a_t|s_t) p(s_{t+1}|s_t, a_t).$$

➢ 总回报

- 给定策略$\pi(a|s)$，智能体和环境一次交互过程的轨迹τ所收到的累积奖励为总回报（return）

$$G(\tau)=\sum_{t=0}^{T-1}\gamma^{t}r_{t+1},$$

$\gamma\in$ [0,1]是折扣率。当γ接近于0时，智能体更在意短期回报；而当γ接近于1时，长期回报变得更重要。环境中有一个或多个特殊的终止状态（terminal state）。

➢ 目标函数

■强化学习的目标是学习到一个策略$\pi_\theta(a|s)$来最大化期望回报(expected return)

$$\mathcal{J}(\theta) = \mathbb{E}_{\tau\sim p_\theta(\tau)}[G(\tau)] = \mathbb{E}_{\tau\sim p_\theta(\tau)}[\sum_{t=0}^{T-1}\gamma^t r_{t+1}],$$

- θ为策略函数的参数

➢ 值函数　状态值函数

■策略π的期望回报分解：

$$\mathbb{E}_{\tau\sim p(\tau)}[G(\tau)] = \mathbb{E}_{s\sim p(s_0)}\left[\mathbb{E}_{\tau\sim p(\tau)}\left[\sum_{t=0}^{T-1}\gamma^t r_{t+1}|\tau_{s_0}=s\right]\right]$$

$$= \mathbb{E}_{s\sim p(s_0)}\left[V^{\pi}(s)\right],$$

$$V^{\pi}(s) = \mathbb{E}_{\tau\sim p(\tau)}\left[\sum_{t=0}^{T-1}\gamma^t r_{t+1}|\tau_{s_0}=s\right],$$

➢ 值函数 状态-动作值函数（Q函数）

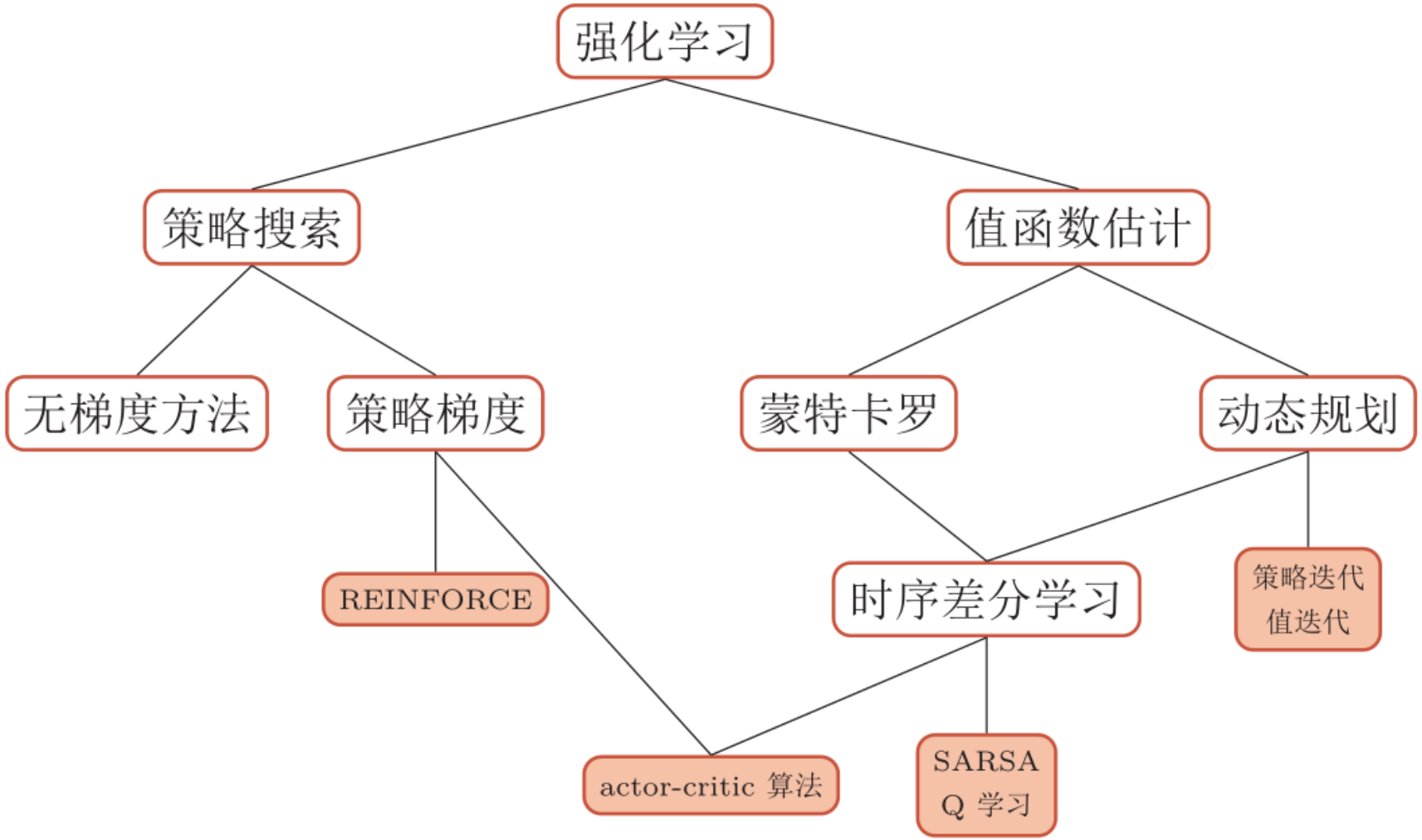
$$
\begin{aligned}
V^{\pi}(s) &= \mathbb{E}_{\tau_{0:T}\sim p(\tau)}\left[r_1 + \gamma\sum_{t=1}^{T-1}\gamma^{t-1}r_{t+1}|\tau_{s_0}=s\right] \\
&= \mathbb{E}_{a\sim\pi(a|s)}\mathbb{E}_{s'\sim p(s'|s,a)}\mathbb{E}_{\tau_{1:T}\sim p(\tau)}\left[r(s,a,s') + \gamma\sum_{t=1}^{T-1}\gamma^{t-1}r_{t+1}|\tau_{s_1}=s'\right] \\
&= \mathbb{E}_{a\sim\pi(a|s)}\mathbb{E}_{s'\sim p(s'|s,a)}\left[r(s,a,s') + \gamma\mathbb{E}_{\tau_{1:T}\sim p(\tau)}\left[\sum_{t=1}^{T-1}\gamma^{t-1}r_{t+1}|\tau_{s_1}=s'\right]\right] \\
&= \mathbb{E}_{a\sim\pi(a|s)}\mathbb{E}_{s'\sim p(s'|s,a)}\left[r(s,a,s') + \gamma V^{\pi}(s')\right].
\end{aligned}
$$

$$
Q^{\pi}(s,a) = \mathbb{E}_{s'\sim p(s'|s,a)}\left[r(s,a,s') + \gamma V^{\pi}(s')\right],
$$

深度强化学习

- 深度强化学习是将强化学习和深度学习结合在一起，用强化学习来定义问题和优化目标，用深度学习来解决状态表示、策略表示等问题。

深度强化学习之间的关系



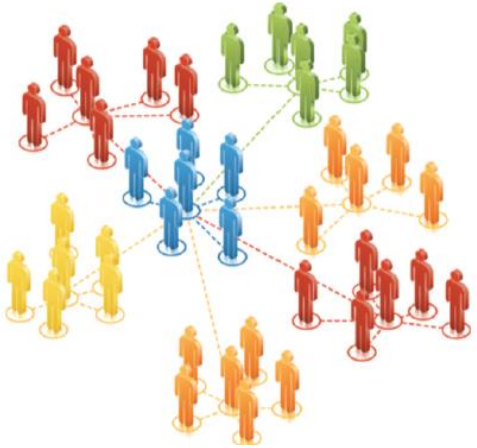
➢ 优化步骤

算法	步骤
SARSA	（1）执行策略，生成样本：s, a, r, s', a' （2）估计回报：$Q(s,a) \leftarrow Q(s,a) + \alpha\Big(r + \gamma Q(s',a') - Q(s,a)\Big)$ （3）更新策略：$\pi(s) = \arg\max_{a \in \lvert\mathcal{A}\rvert} Q(s,a)$
Q学习	（1）执行策略，生成样本：s, a, r, s' （2）估计回报：$Q(s,a) \leftarrow Q(s,a) + \alpha\Big(r + \gamma \max_{a'} Q(s',a') - Q(s,a)\Big)$ （3）更新策略：$\pi(s) = \arg\max_{a \in \lvert\mathcal{A}\rvert} Q(s,a)$
REINFORCE	（1）执行策略，生成样本：$\tau = s_0, a_0, s_1, a_1, \cdots$ （2）估计回报：$G(\tau) = \sum_{t=0}^{T-1} r_{t+1}$ （3）更新策略：$\theta \leftarrow \theta + \sum_{t=0}^{T-1}\Big(\frac{\partial}{\partial\theta}\log\pi_\theta(a_t\lvert s_t)\gamma^t G(\tau_{t:T})\Big)$
Actor-Critic	（1）执行策略，生成样本：s, a, s', r （2）估计回报：$G(s) = r + \gamma V_\phi(s')$ $\phi \leftarrow \phi + \beta\Big(G(s) - V_\phi(s)\Big)\frac{\partial}{\partial\phi}V_\phi(s)$ （3）更新策略：$\lambda \leftarrow \gamma\lambda$ $\theta \leftarrow \theta + \alpha\lambda\Big(G(s) - V_\phi(s)\Big)\frac{\partial}{\partial\theta}\log\pi_\theta(a\lvert s)$

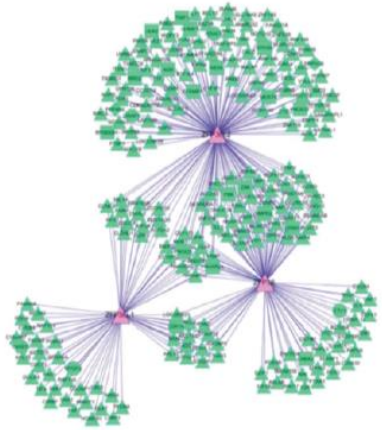
目录

图神经网络

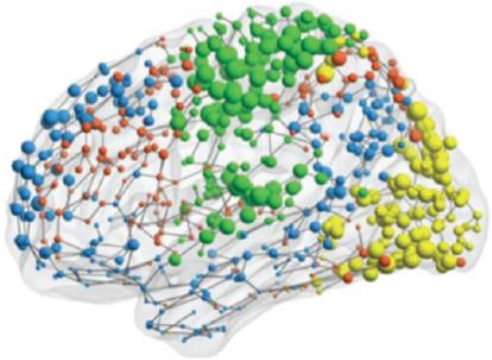
➢ 背景--图数据



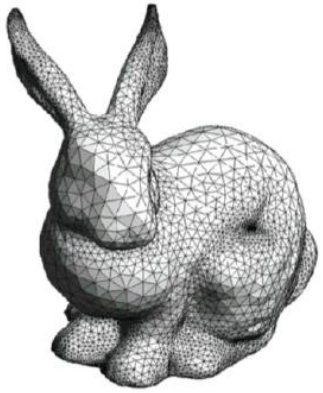
社交网络



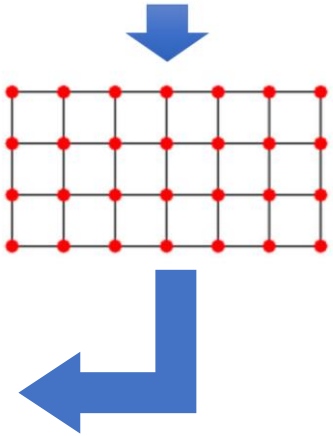
基因调控网络



大脑功能网络

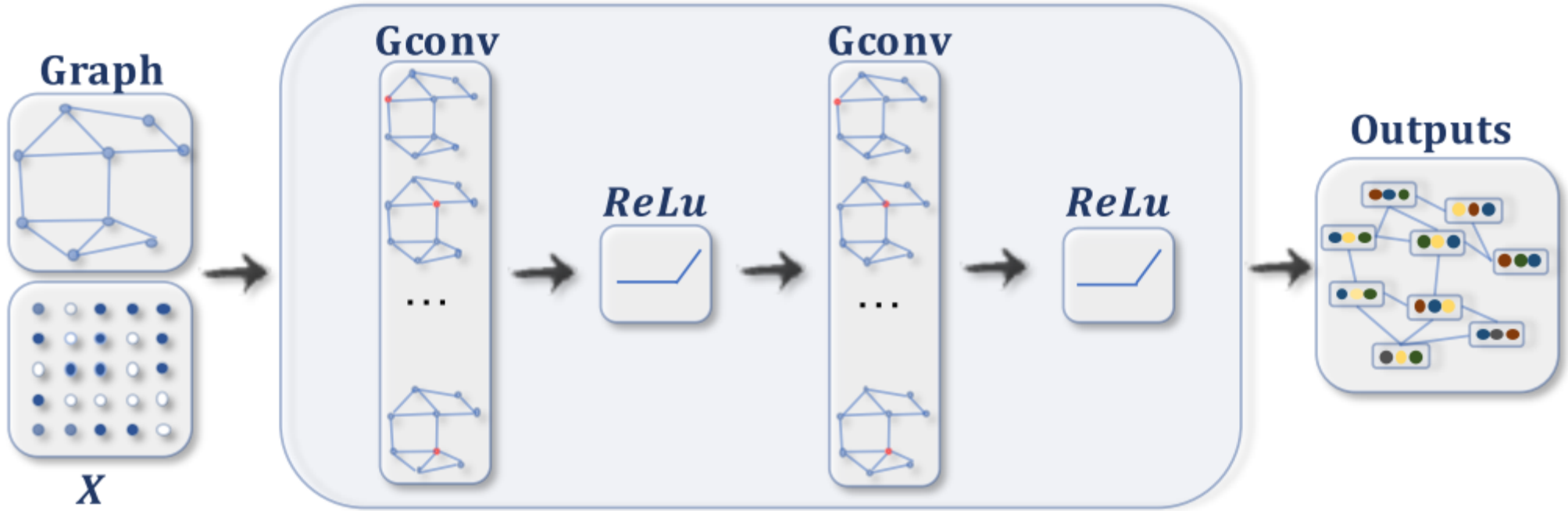


3维形状



➢ 起源

■借鉴卷积网络、循环网络和深度自动编码器等的思想，定义、设计用于处理图数据的神经网络结构。



图神经网络

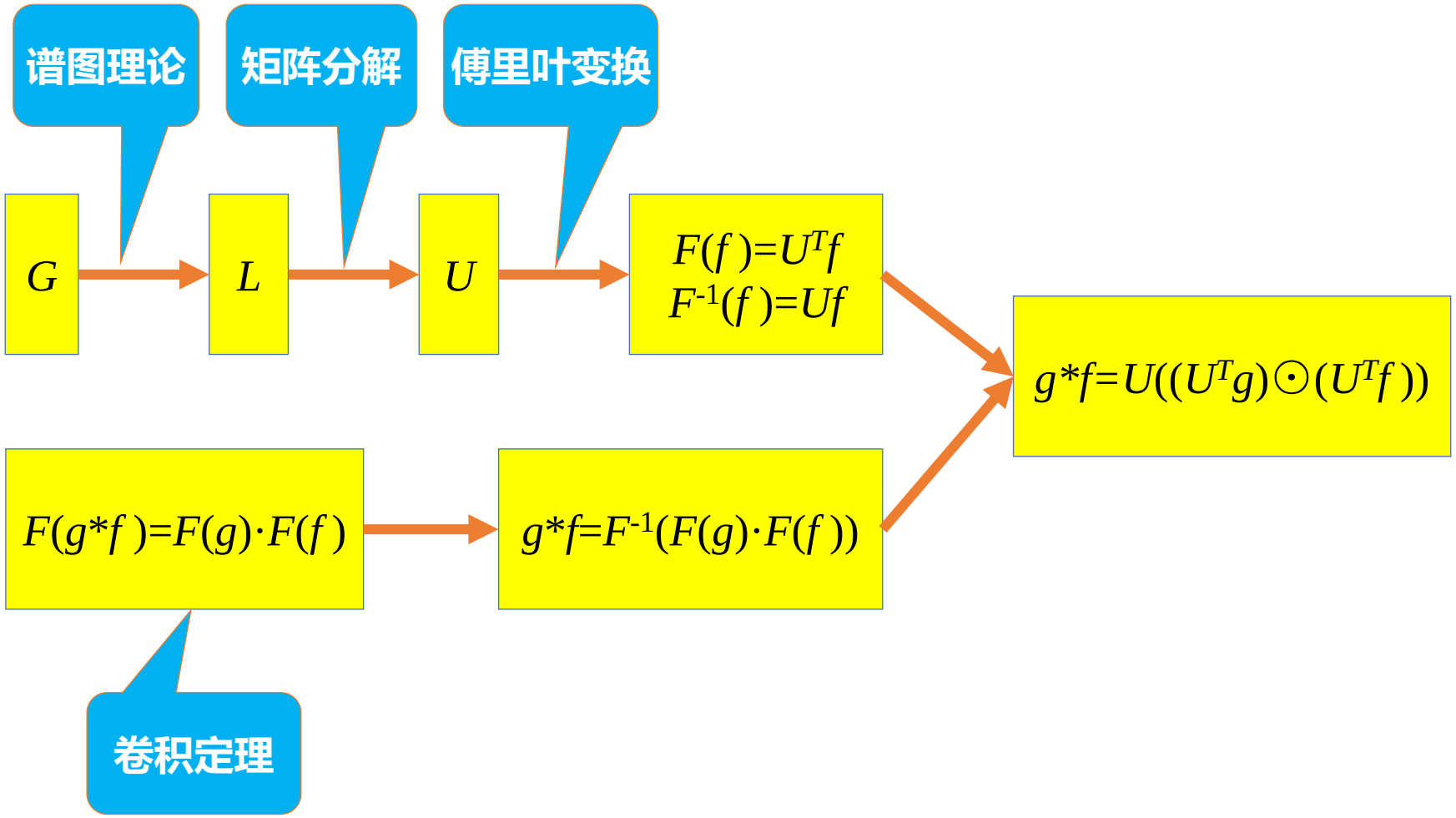
分类

- 图卷积网络（Graph Convolution Networks，GCN）
- 图注意力网络（Graph Attention Networks，GAT）
- 图自编码器（ Graph Autoencoders）
- 图生成网络（ Graph Generative Networks）
- 图时空网络（Graph Spatial-temporal Networks）

图卷积网络

- GCN方法又可以分为两大类，基于谱（spectral-based）和基于空间（spatial-based）。
- 基于谱的方法从图信号处理的角度引入滤波器来定义图卷积，其中图卷积操作被解释为从图信号中去除噪声。
- 基于空间的方法将图卷积表示为从邻域聚合特征信息，当图卷积网络的算法在节点层次运行时，图池化模块可以与图卷积层交错，将图粗化为高级子结构。

➢ 基于谱的图卷积


谱图理论
矩阵分解
傅里叶变换
G
L
U
$F(f)=U^Tf$
$F^{-1}(f)=Uf$
$g*f=U((U^Tg)\odot(U^Tf))$
$F(g*f)=F(g)\cdot F(f)$
$g*f=F^{-1}(F(g)\cdot F(f))$
卷积定理

图卷积网络

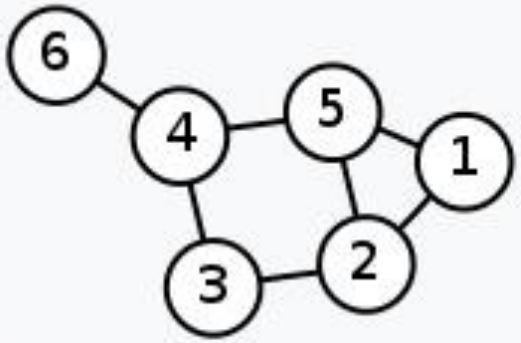
➢ 谱图理论

■借助于图的拉普拉斯矩阵的特征值和特征向量来研究图的空间性质。

➢ 拉普拉斯矩阵

■给定一个图G=(V,E), D为顶点度矩阵, A为邻接矩阵, 则拉普拉斯矩阵

$$L = D - A$$

Labeled graph	Degree matrix	Adjacency matrix	Laplacian matrix
	$\begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 2 & -1 & 0 & 0 & -1 & 0 \\ -1 & 3 & -1 & 0 & -1 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ -1 & -1 & 0 & -1 & 3 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix}$

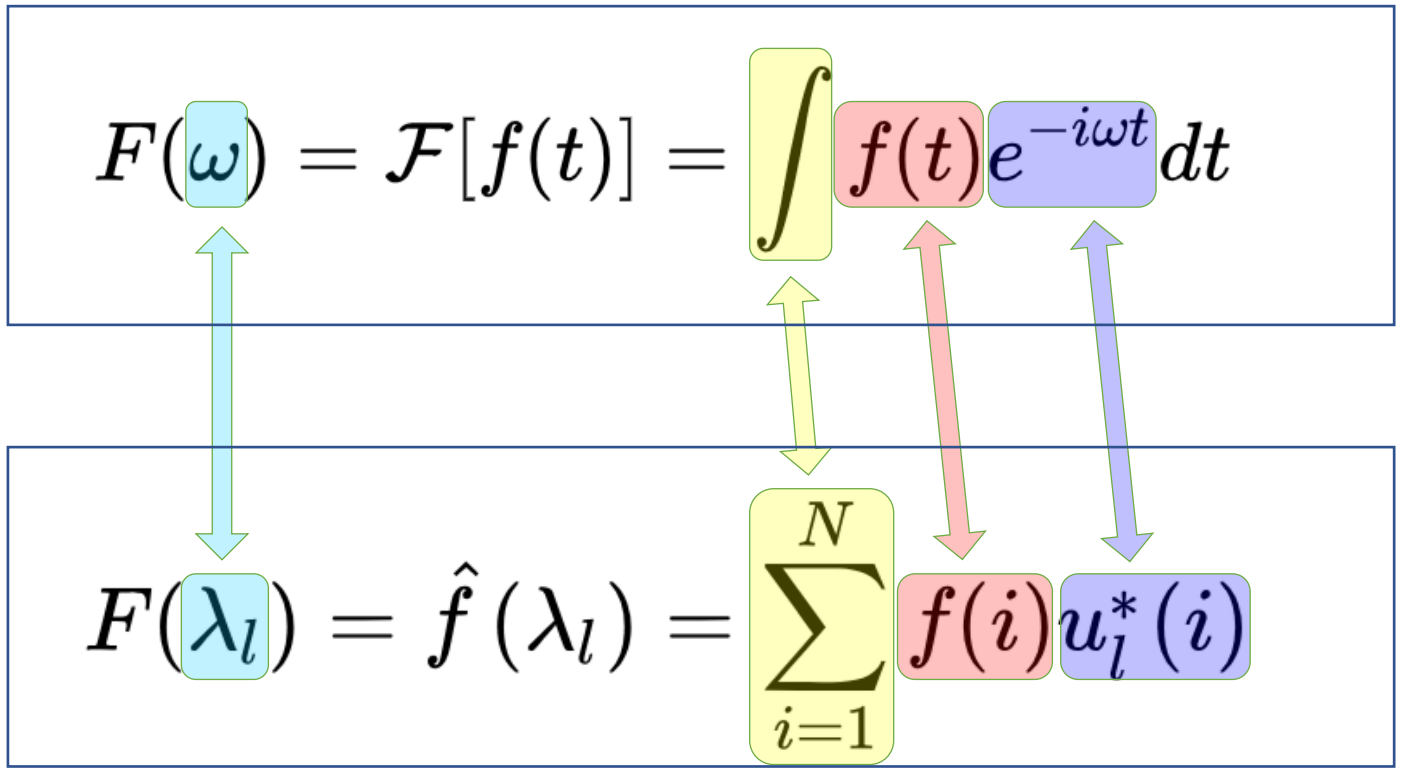
➢ 拉普拉斯矩阵分解

$$L = U \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} U^{-1}$$

U为特征向量矩阵，λ为特征值构成的对角阵。由于U是正交的，所以：

$$L = U \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} U^{T}$$

图信号的傅里叶变换



传统傅里叶变换

图信号傅里叶变换

- 图的第i个点上的signal为f(i)
- 图x表示为 $x = (f(1) \dots f(n)) \in \mathbb{R}^n$
- 互为对偶向量 u_l^* 是 u_l 的的对偶向量（线性代数的知识）
- 其中 u_l^* 是矩阵 U^T 的第 l 行，u_l 是矩阵 U 的第 l 行

➢ 图信号的傅里叶变换

$$F(\lambda_l) = \hat{f}(\lambda_l) = \sum_{i=1}^{N} f(i) u_l^*(i)$$

$$\begin{pmatrix} \hat{f}(\lambda_1) \\ \hat{f}(\lambda_2) \\ \vdots \\ \hat{f}(\lambda_N) \end{pmatrix} = \begin{pmatrix} u_1(1) & u_1(2) & \dots & u_1(N) \\ u_2(1) & u_2(2) & \dots & u_2(N) \\ \vdots & \vdots & \ddots & \vdots \\ u_N(1) & u_N(2) & \dots & u_N(N) \end{pmatrix} \begin{pmatrix} f(1) \\ f(2) \\ \vdots \\ f(N) \end{pmatrix}$$

$$\mathcal{GF}\{x\} = U^T x$$

$$\mathcal{IGF}\{x\} = Ux$$

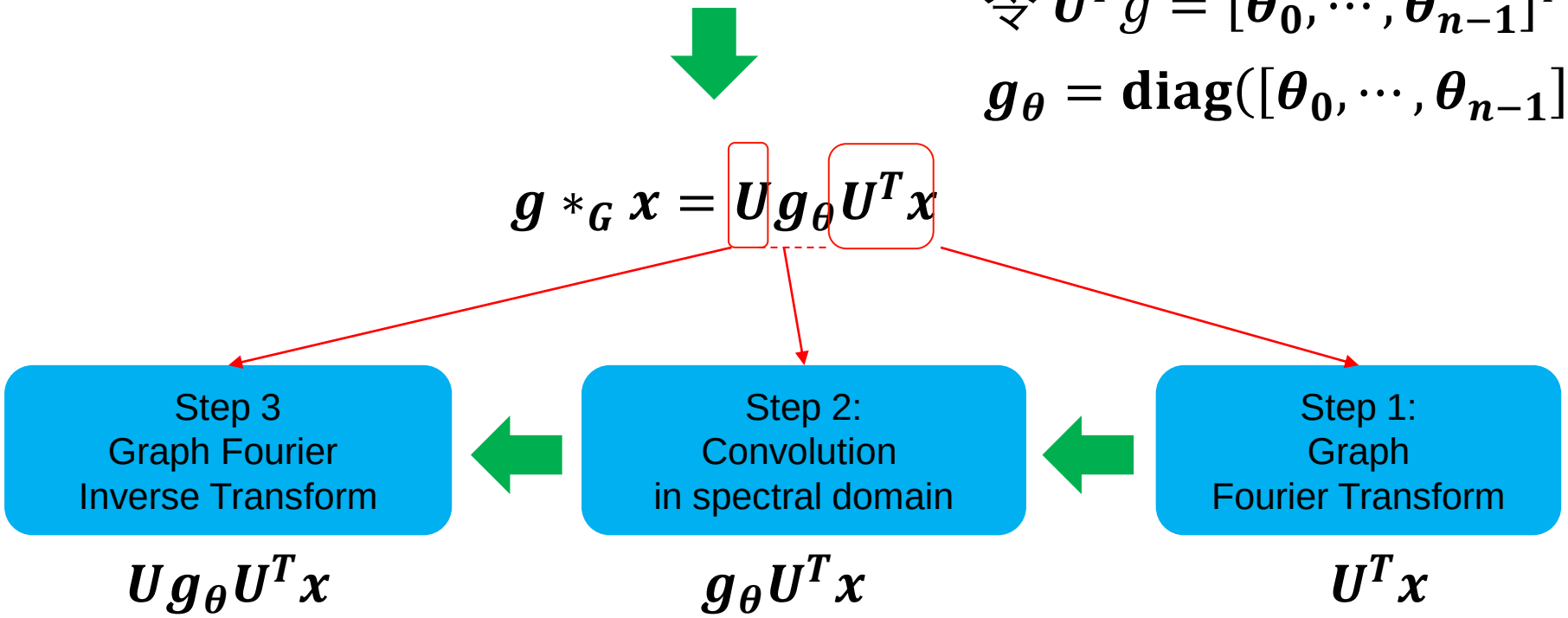
图卷积操作

$$g *_G x = U\left((U^T g)\odot(U^T x)\right)$$

令 $U^T g = [\theta_0, \cdots, \theta_{n-1}]^T$ 且

$g_\theta = \mathbf{diag}([\theta_0, \cdots, \theta_{n-1}])$, 有

$$g *_G x = U g_\theta U^T x$$

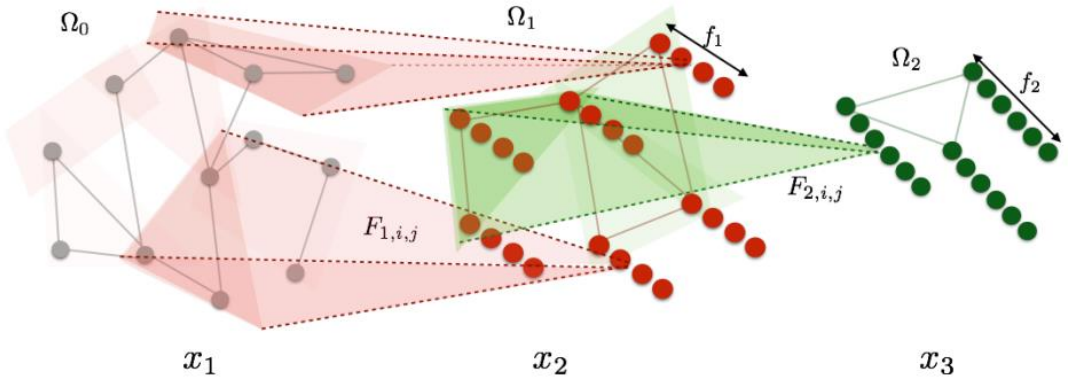


➢ SCNN：第一代图卷积模型

- 直接把卷积核的傅里叶变换的对角阵 $diag(\hat{g}(\lambda_l))$ 当作参数

$$y_{output} = \sigma\left(U\begin{pmatrix}\theta_1 & & \\ & \ddots & \\ & & \theta_n\end{pmatrix}U^T x\right)$$

- 不足
  - 计算U、Θ、U^T之间的乘积的计算代价太高，$O(n^2)$
  - 卷积核参数过多（n个参数），感受野太大，不够局部化



J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun. Spectral networks and locally connected networks on graphs. ICLR, 2014.

➢ ChebNet & GCN：第二代图卷积模型

■直接把卷积把 $diag(\hat{g}(\lambda_l))$ 设计为关于特征值的多项式形式 $\sum_{j=0}^{K} \alpha_j \lambda_l^j$

$$y_{output} = \sigma\left(U\begin{pmatrix}\sum_{j=0}^{K}\alpha_j\lambda_1^j & & \\ & \ddots & \\ & & \sum_{j=0}^{K}\alpha_j\lambda_n^j\end{pmatrix}U^T x\right)$$

■优势

- 参数个数由n变为k，缩小了感受野为节点的k跳邻居
- 计算复杂度下降

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

图神经网络

应用

Area	Application	Algorithm	Deep Learning Model	References
Text	Text classification	GCN	Graph Convolutional Network	[1], [21], [38] [2], [20], [36]
		GAT	Graph Attention Network	[54]
		DGCNN	Graph Convolutional Network	[79]
		Text GCN	Graph Convolutional Network	[80]
		Sentence LSTM	Graph LSTM	[48]
	Sequence Labeling (POS, NER)	Sentence LSTM	Graph LSTM	[48]
	Sentiment classification	Tree LSTM	Graph LSTM	[46]
	Semantic role labeling	Syntactic GCN	Graph Convolutional Network	[81]
	Neural machine translation	Syntactic GCN	Graph Convolutional Network	[82], [83]
		GGNN	Gated Graph Neural Network	[31]
	Relation extraction	Tree LSTM	Graph LSTM	[84]
		Graph LSTM	Graph LSTM	[34], [85]
		GCN	Graph Convolutional Network	[86]
	Event extraction	Syntactic GCN	Graph Convolutional Network	[87], [88]
	AMR to text generation	Sentence LSTM	Graph LSTM	[89]
		GGNN	Gated Graph Neural Network	[31]
	Multi-hop reading comprehension	Sentence LSTM	Graph LSTM	[90]
	Relational reasoning	RN	MLP	[69]
		Recurrent RN	Recurrent Neural Network	[91]
		IN	Graph Neural Network	[4]
Image	Social Relationship Understanding	GRM	Gated Graph Neural Network	[92]
	Image classification	GCN	Graph Convolutional Network	[93], [94]
		GGNN	Gated Graph Neural Network	[95]
		ADGPM	Graph Convolutional Network	[29]
		GSNN	Gated Graph Neural Network	[96]
	Visual Question Answering	GGNN	Gated Graph Neural Network	[92], [97], [98]
	Object Detection	RN	Graph Attention Network	[99], [100]
	Interaction Detection	GPNN	Graph Neural Network	[101]
		Structural-RNN	Graph Neural Network	[102]
	Region Classification	GCNN	Graph CNN	[103]
	Semantic Segmentation	Graph LSTM	Graph LSTM	[49], [104]
		GGNN	Gated Graph Neural Network	[105]
		DGCNN	Graph CNN	[106]
		3DGNN	Graph Neural Network	[107]
Science	Physics Systems	IN	Graph Neural Network	[4]
		VIN	Graph Neural Network	[64]
		GN	Graph Networks	[3]
	Molecular Fingerprints	NGF	Graph Convolutional Network	[33]
		GCN	Graph Convolutional Network	[72]
	Protein Interface Prediction	GCN	Graph Convolutional Network	[5]
	Side Effects Prediction	Decagon	Graph Convolutional Network	[108]
	Disease Classification	PPIN	Graph Convolutional Network	[109]
Knowledge Graph	KB Completion	IN	Graph Neural Network	[4]
	KG Alignment	GCN	Graph Convolutional Network	[110]
Combinatorial Optimization		structure2vec	Graph Convolutional Network	[7]
		GNN	Graph Neural Network	[111]
		GCN	Graph Convolutional Network	[112]
		AM	Graph Attention Network	[113]
Graph Generation		NetGAN	Long short-term memory	[114]
		GraphRNN	Rucurrent Neural Network	[111]
		Regularizing VAE	Variational Autoencoder	[115]
		GCPN	Graph Convolutional Network	[116]
		MolGAN	Relational-GCN	[117]

文本

文本分类　关系抽取
阅读理解　机器翻译
....

图像

图像分类　视觉问答
目标检测　语义分割
....

科学

分子指纹解析　疾病分类
药物副作用预测
....

知识图谱

联合优化

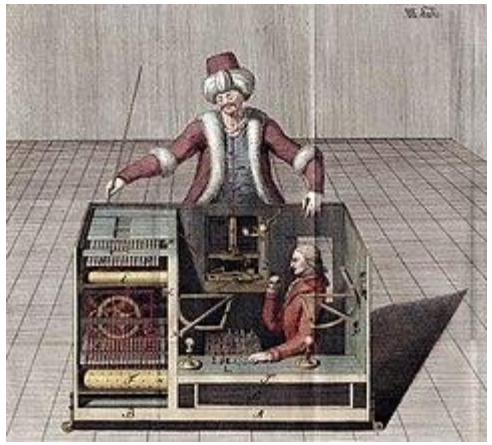
图生成

Zhou J , Cui G , Zhang Z , et al. Graph Neural Networks: A Review of Methods and Applications[J]. 2018.

目录

“人工”智能

1769



TheTurk

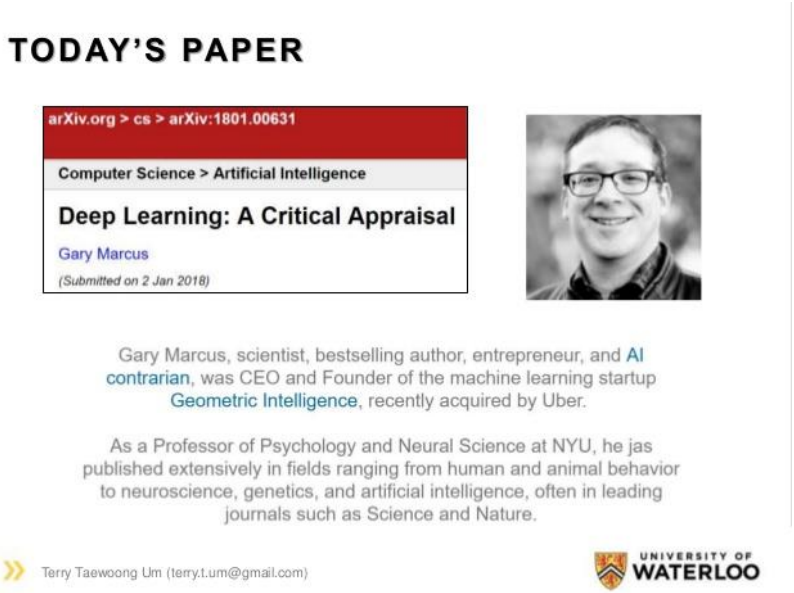
匈牙利作家兼发明家Wolfgang von Kempelen建造了机器人TheTurk，能够跟国际象棋高手对弈，但最终谜底揭开，机器人之所以会下棋是因为箱子里藏着一个人。

2019

KiwiBot

美国校园外卖机器人由在哥伦比亚的操作员远程人工操纵：依靠 GPS 信号和摄像头获得的视频进行监控和操作。

深度学习的“不能”



Deep learning thus far

3.1. is data hungry

3.2. is shallow & has limited capacity for transfer

3.3. has no natural way to deal with hierarchical structure

3.4. has struggled with open-ended inference

3.5. is not sufficiently transparent

3.6. has not been well integrated with prior knowledge

3.7. cannot inherently distinguish causation from correlation

3.8. presumes a largely stable world

3.9. its answer often cannot be fully trusted

3.10. is difficult to engineer with

Gary Marcus. 2018. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631* (2018).

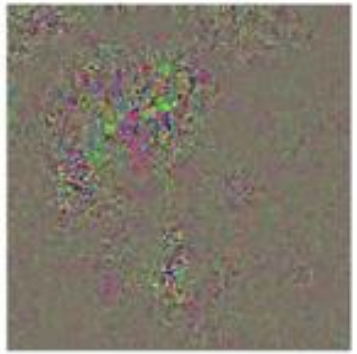
深度学习的“不能”（1）

- 算法输出不稳定，容易被“攻击”

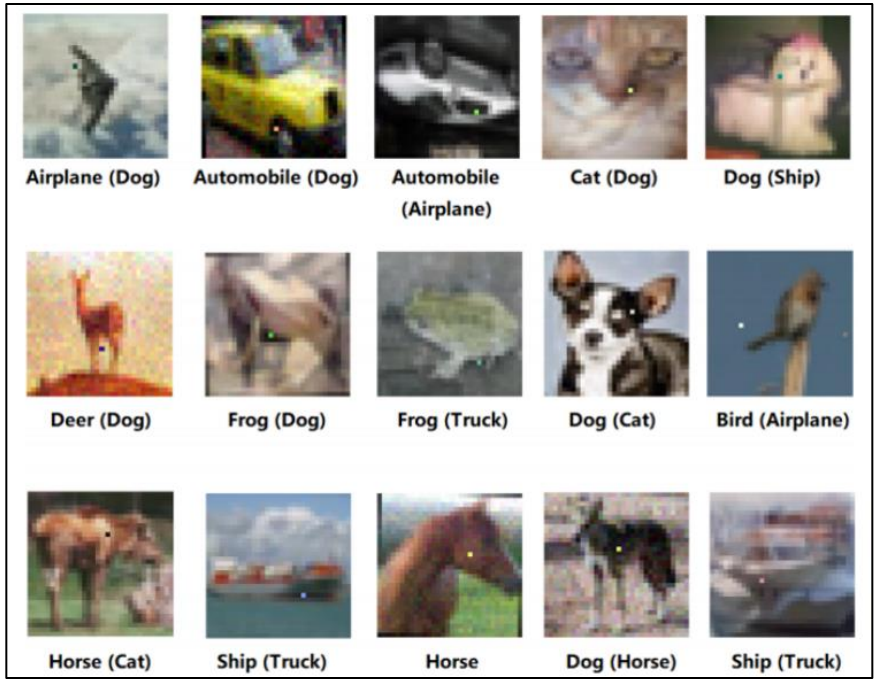
“African elephant”

“koala”

difference



10x difference



one pixel attack

深度学习的“不能” （1）

- 算法输出不稳定，容易被“攻击”

原始图片
（分类结果:Bulldog）

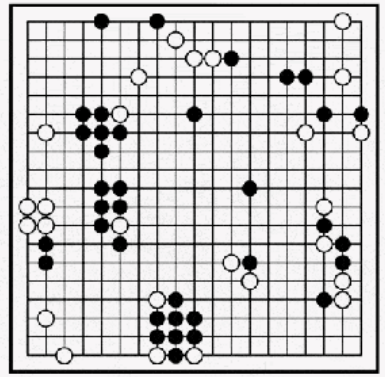
攻击图片1
（系统崩溃，DoS）

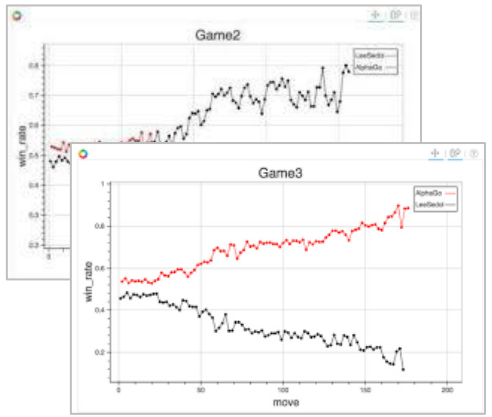
攻击图片2
（分类错误）

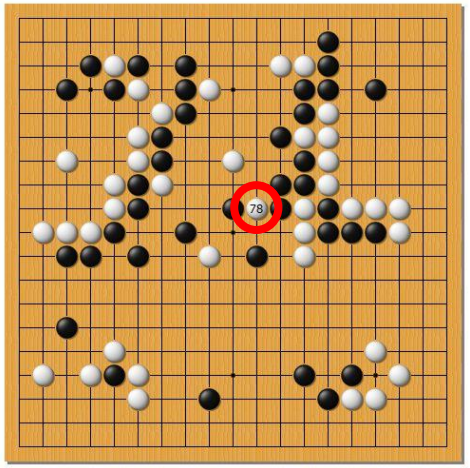
攻击图片3
（系统越权）

深度学习的“不能” (2)

- 模型复杂度高，难以纠错和调试

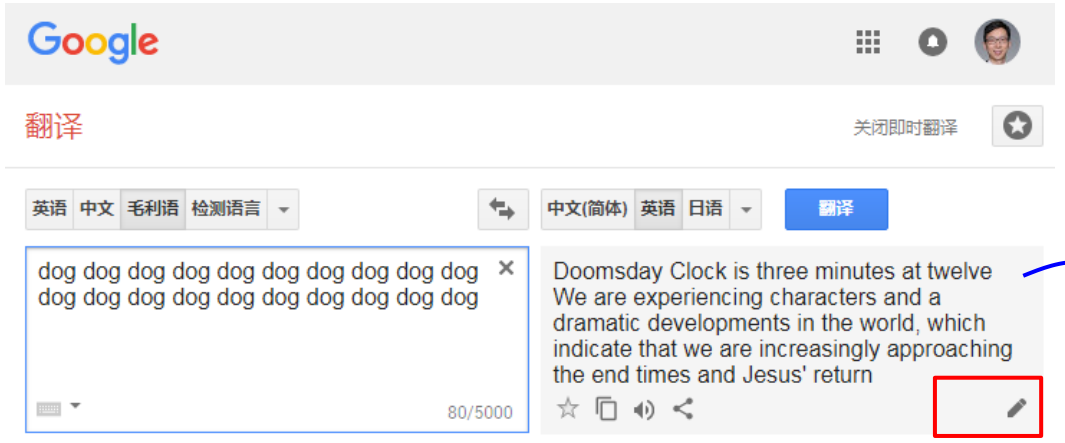






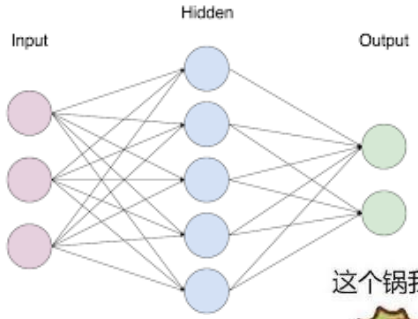
深度学习的“不能” (2)

- 模型复杂度高，难以纠错和调试

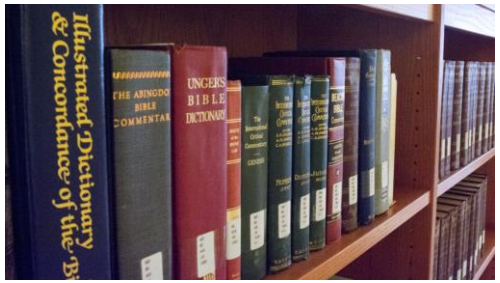


“世界末日时钟还差3分钟到12点。我们正在经历世界上的人物和戏剧性的发展，这表明我们越来越接近末日和耶稣的回归。”

恶作剧?



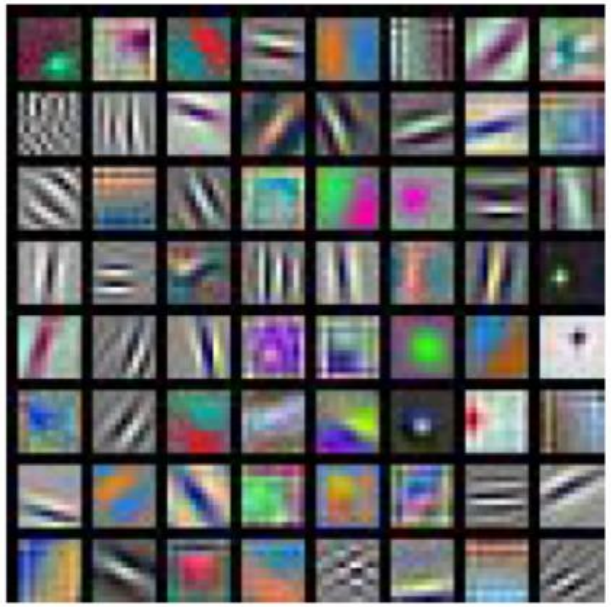
Deep Dream



《Bible》

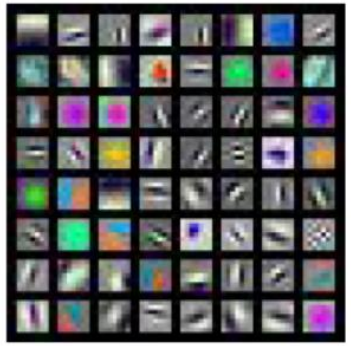
深度学习的“不能”（3）

- 模型层级复合程度高，参数不透明

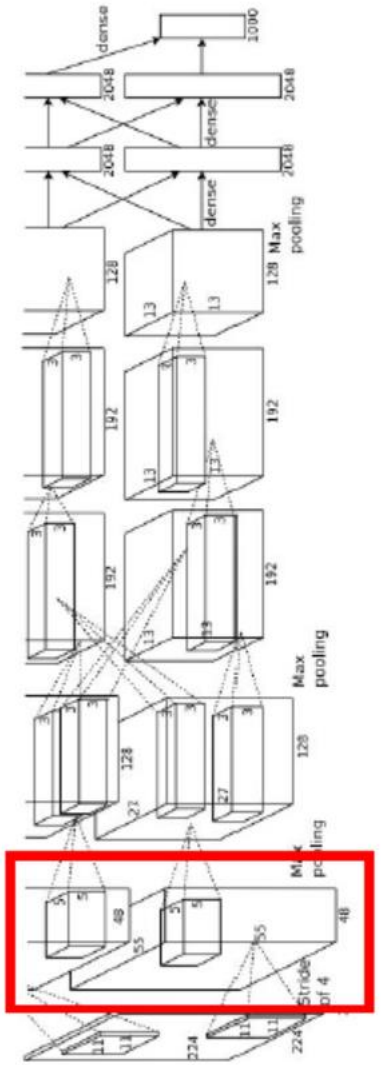


AlexNet:
64 x 3 x 11 x 11

ResNet-101:
64 x 3 x 7 x 7



DenseNet-121:
64 x 3 x 7 x 7

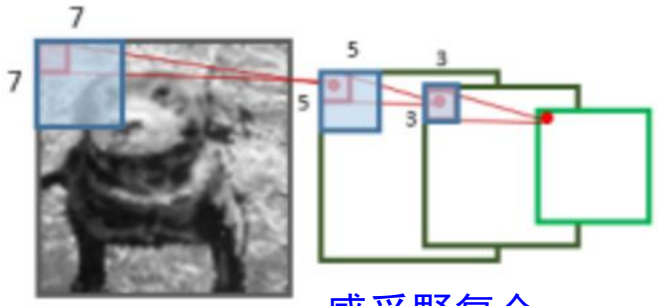


深度学习的“不能”（3）

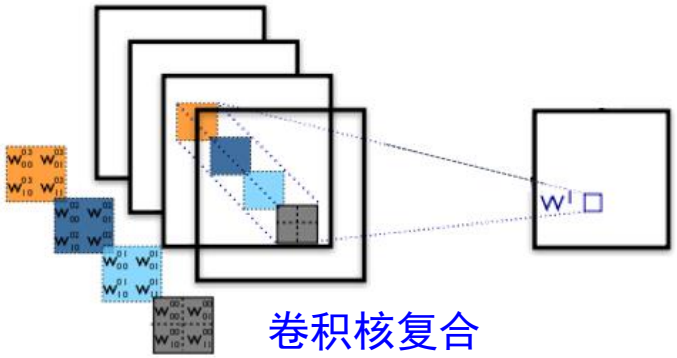
- 模型层级复合程度高，参数不透明

第一层：16*3*3*3

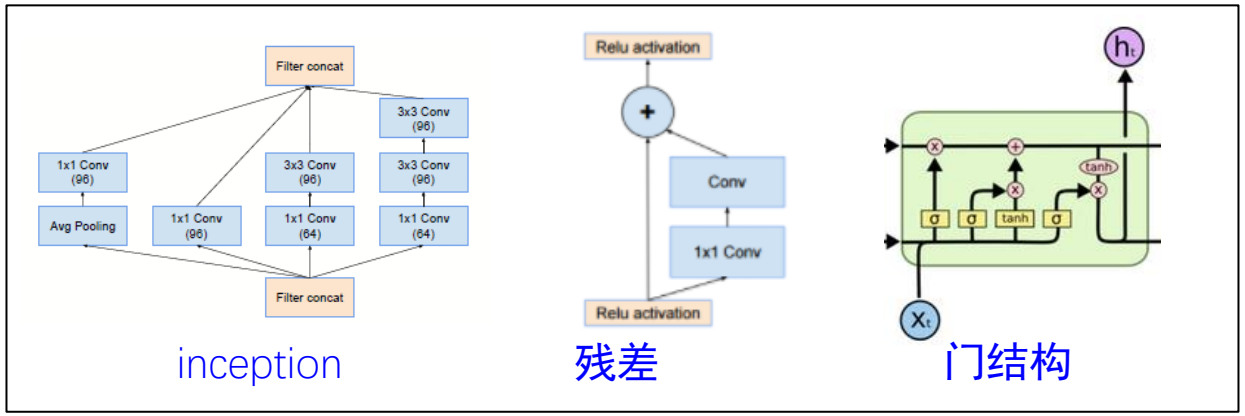
第二层：N*16*3*3



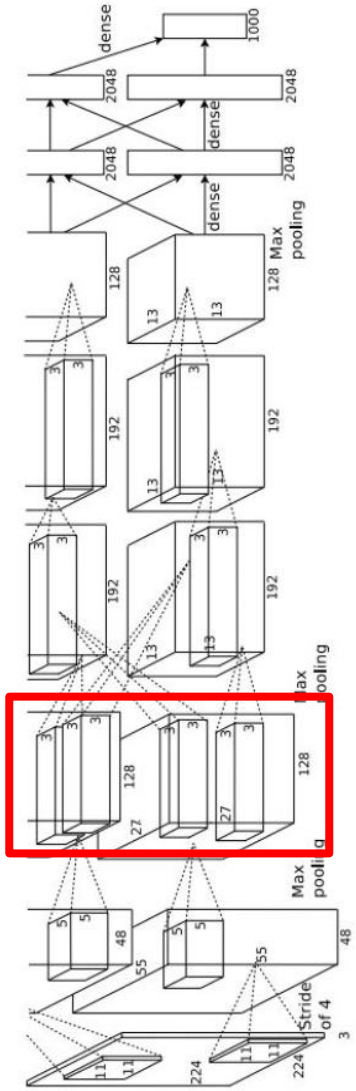
感受野复合



卷积核复合



inception　　残差　　门结构



深度学习的“不能” (3)

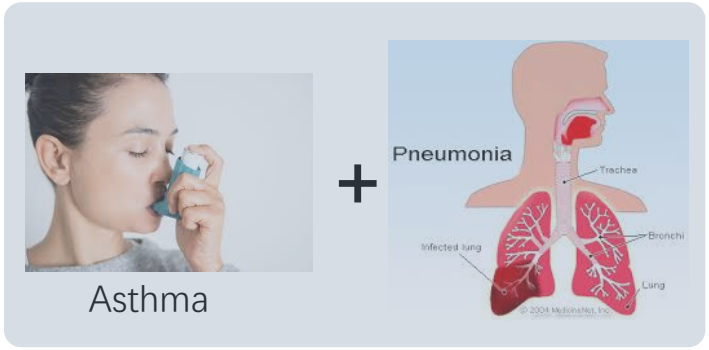
- 基于相关关系，学习得到的是不可靠关联模式

Patient ID		Has Asthma		Risk of Death
84	...	Yes	...	5%
85	...	Yes	...	6%
86	...	No	...	12%
87	...	No	...	15%
...	...	...	...	...

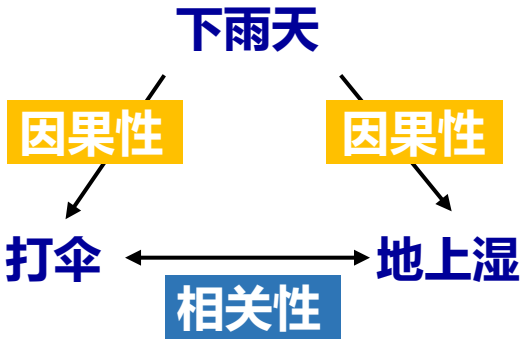
Feature Importance (Higher risk of death): Low High

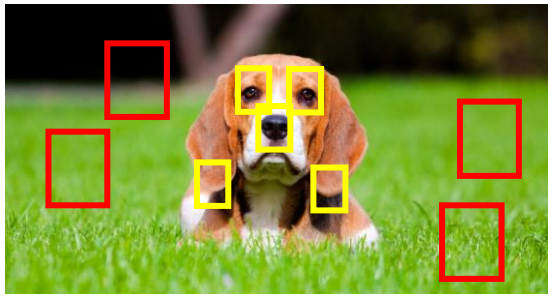
Feature Importance (Lower risk of death): Low High

With Context: Patients with asthma have a lower risk of death from pneumonia because they receive more intensive care.



⇨ 哮喘对治疗肺炎有效





Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission.

深度学习的“不能” (4)

■端到端训练方式对数据依赖性强，模型增量性差

$$\text{Test loss} - \text{training loss} \le \sqrt{\frac{N}{m}}$$

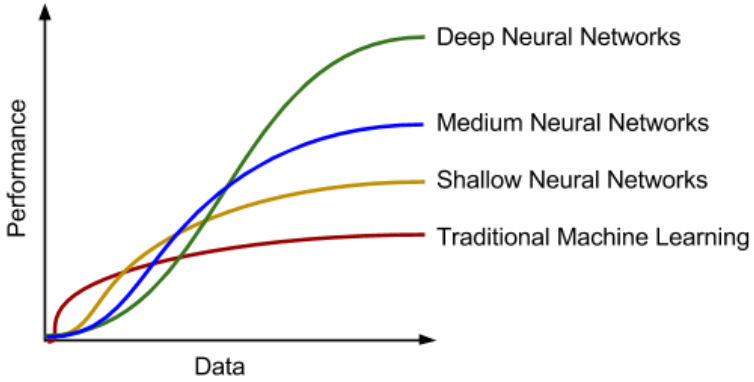
泛化误差　　训练误差

m = #训练样本

N = effective capacity (模型有效容量)

↑ 上界

#参数 / VC维 / Rademacher复杂度



当样本数据量小的时候，深度学习无法体现强大拟合能力

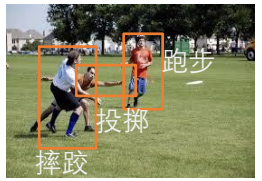

马


草地


农舍


山

语义标注

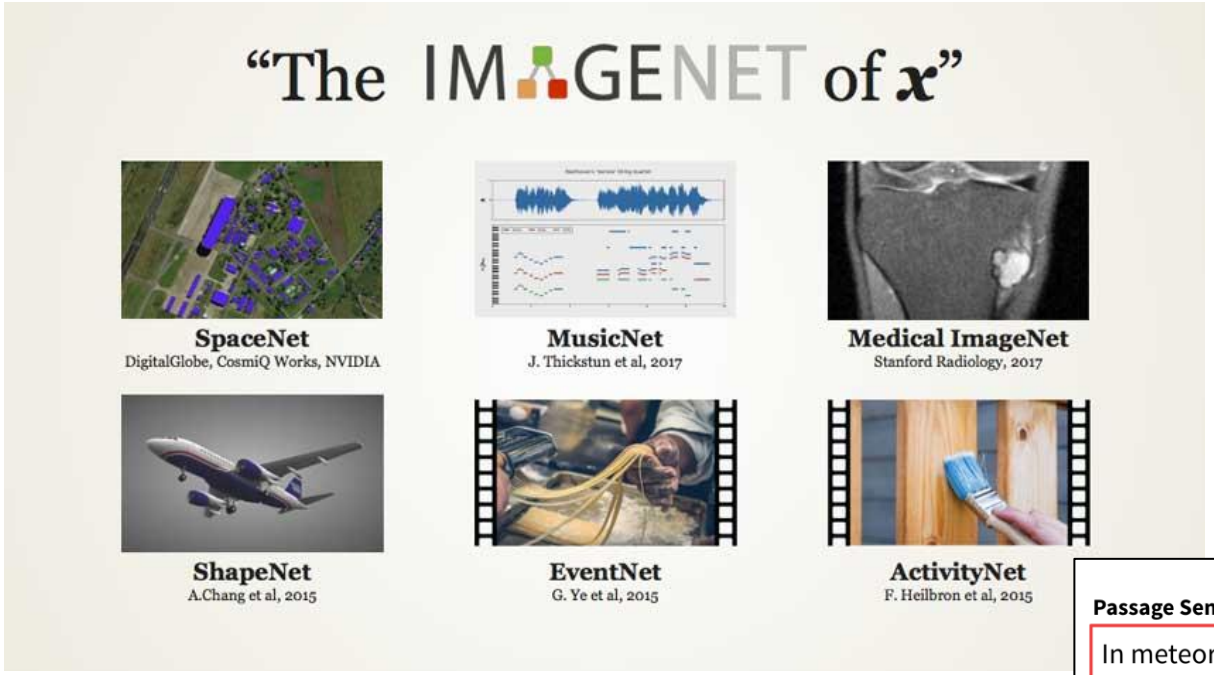


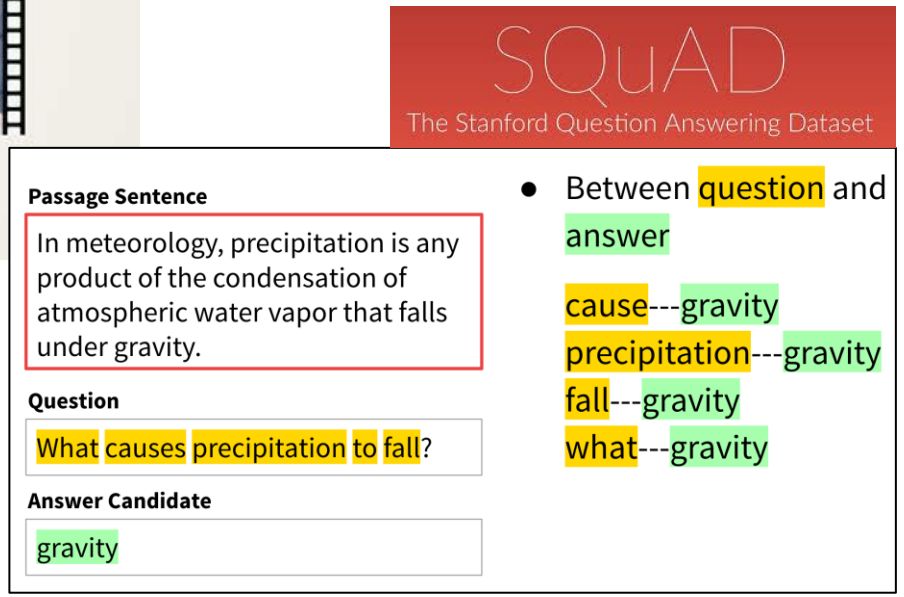
关系检测

图像描述？

深度学习的“不能”（5）

- 专注直观感知类问题，对开放性推理问题无能为力



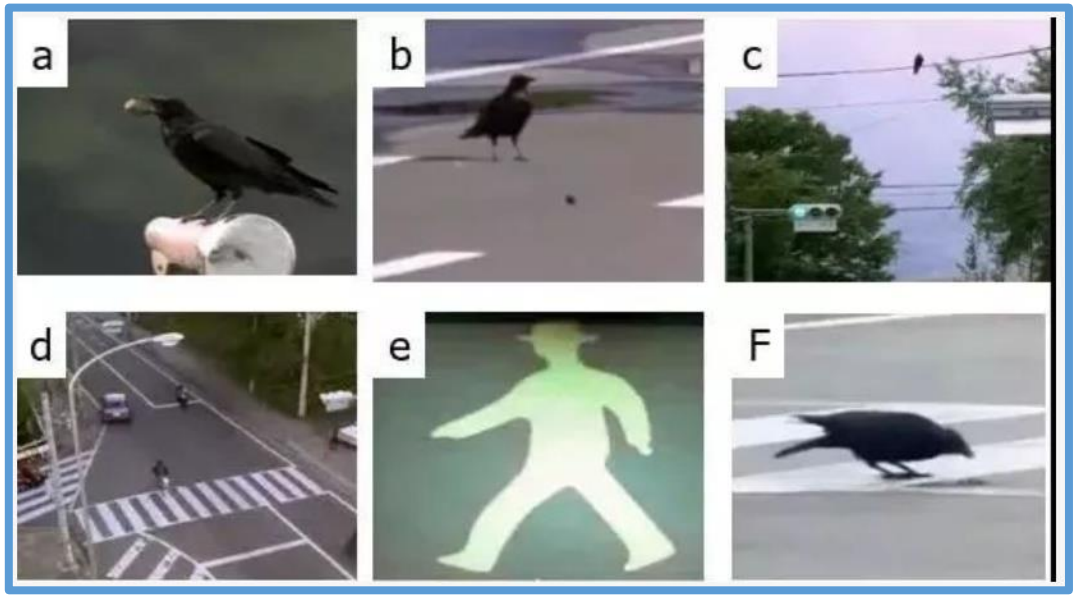


深度学习的“不能” （5）

- 专注直观感知类问题，对开放性推理问题无能为力

“鹦鹉”智能

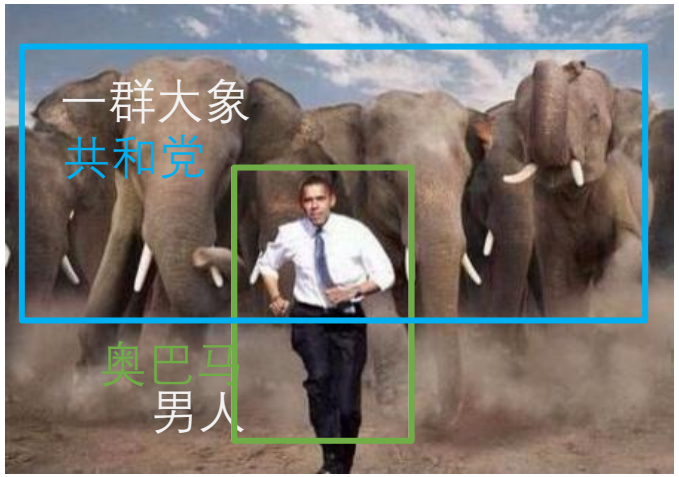
“乌鸦”智能

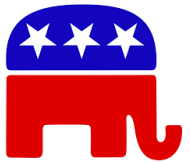


本页改编自朱松纯老师的报告 “人工智能的现状、任务、构架与统一”

深度学习的“不能” （5）

- 专注直观感知类问题，对开放性推理问题无能为力





大象

共和党

本页部分例子来自芮勇博士的报告 “计算机视觉从感知到认知的长征”

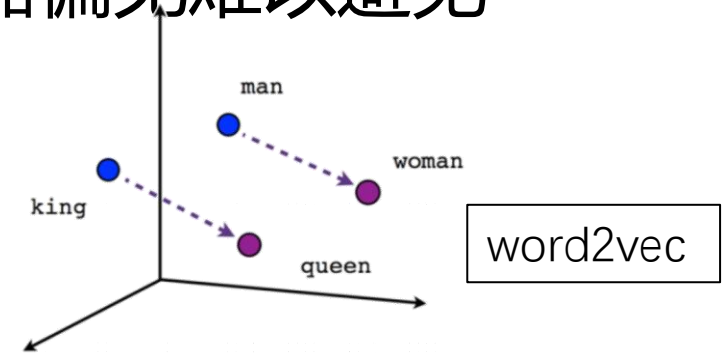
深度学习的“不能”（6）

- 人类知识无法有效引入进行监督，机器偏见难以避免

Man:Woman as King:Queen

Father:Doctor as Mother: Nurse

Man:Computer_Programmer as Woman: Homemaker



word2vec

Gender stereotype *she-he* analogies

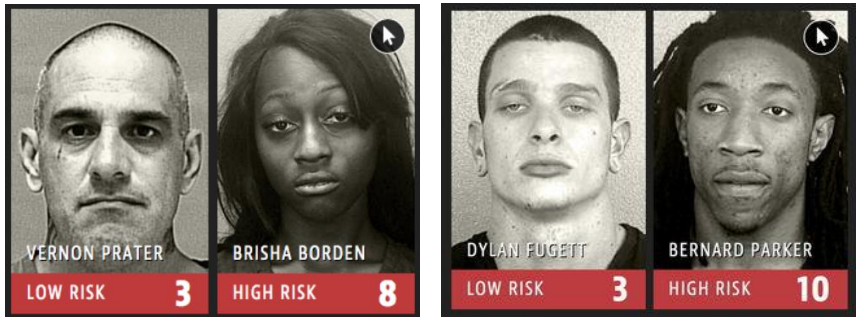
sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	lovely-brilliant

深度学习的“不能”（6）

人类知识无法有效引入进行监督，机器偏见难以避免

- 微软开发了Tay聊天机器人，模仿年轻网民的语言模式。
- 试用24小时后，被引入歧途，成为偏激的种族主义者，甚至发出了“希特勒无罪”的消息。

美国法院用以评估犯罪风险的算法COMPAS，被证明对黑人造成了系统性歧视。



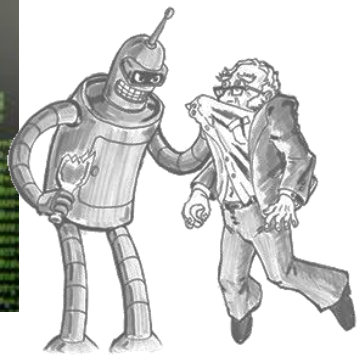
算法依赖于大数据，但数据不是中立的：从真实社会中抽取，必然带有社会固有的不平等、排斥性和歧视。

深度学习的“不能” (6)

- 人类知识无法有效引入进行监督，机器偏见难以避免

阿西莫夫机器人三定律

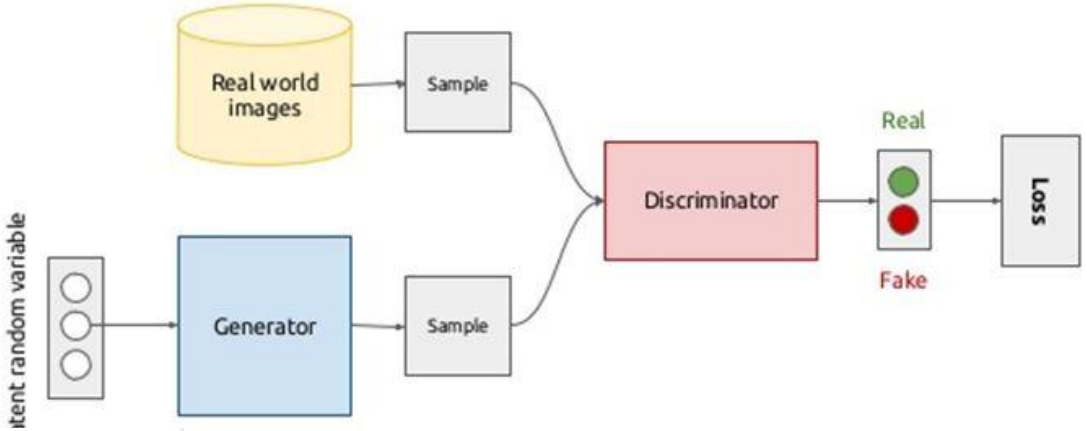
- 机器人不得伤害人类，或坐视人类受到伤害；
- 除非违背第一法则，否则机器人必须服从人类命令；
- 除非违背第一或第二法则，否则机器人必须保护自己

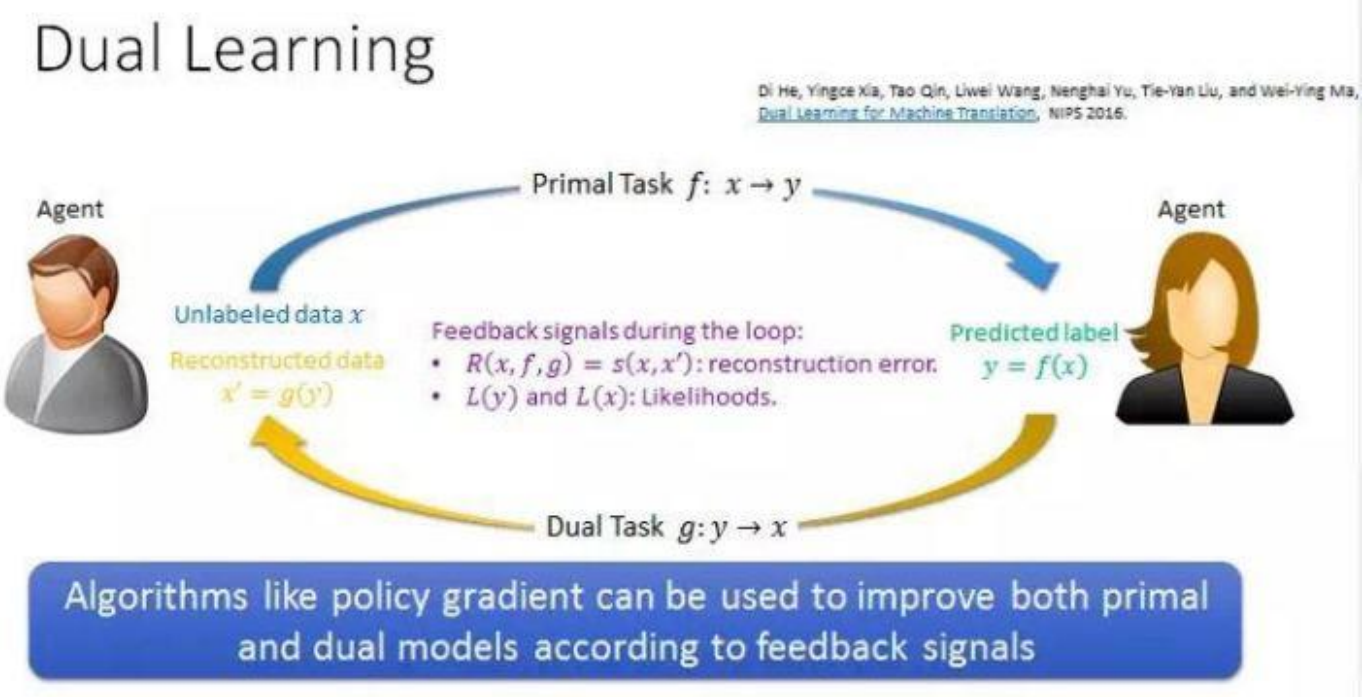


目录

深度学习趋势

- 挑战1：标注数据代价昂贵
- 趋势1：从无标注的数据里学习





深度学习趋势

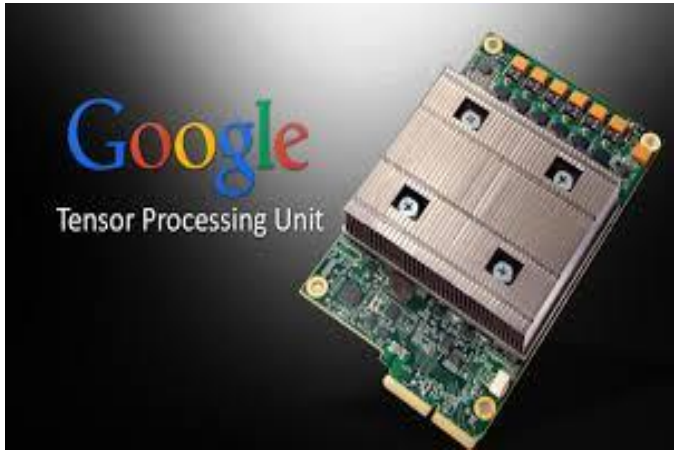
- 挑战2：大模型不方便在移动设备上使用
- 前沿2：降低模型大小

剪枝　　权值共享　　量化（二进制网络）

Network	Original Size	Compressed Size	Compression Ratio	Original Accuracy	Compressed Accuracy
LeNet-300	1070KB	27KB	**40x**	98.36%	98.42%
LeNet-5	1720KB	44KB	**39x**	99.20%	99.26%
AlexNet	240MB	6.9MB	**35x**	80.27%	80.30%
VGGNet	550MB	11.3MB	**49x**	88.68%	89.09%
GoogleNet	28MB	2.8MB	**10x**	88.90%	88.92%
SqueezeNet	4.8MB	0.47MB	**10x**	80.32%	80.35%

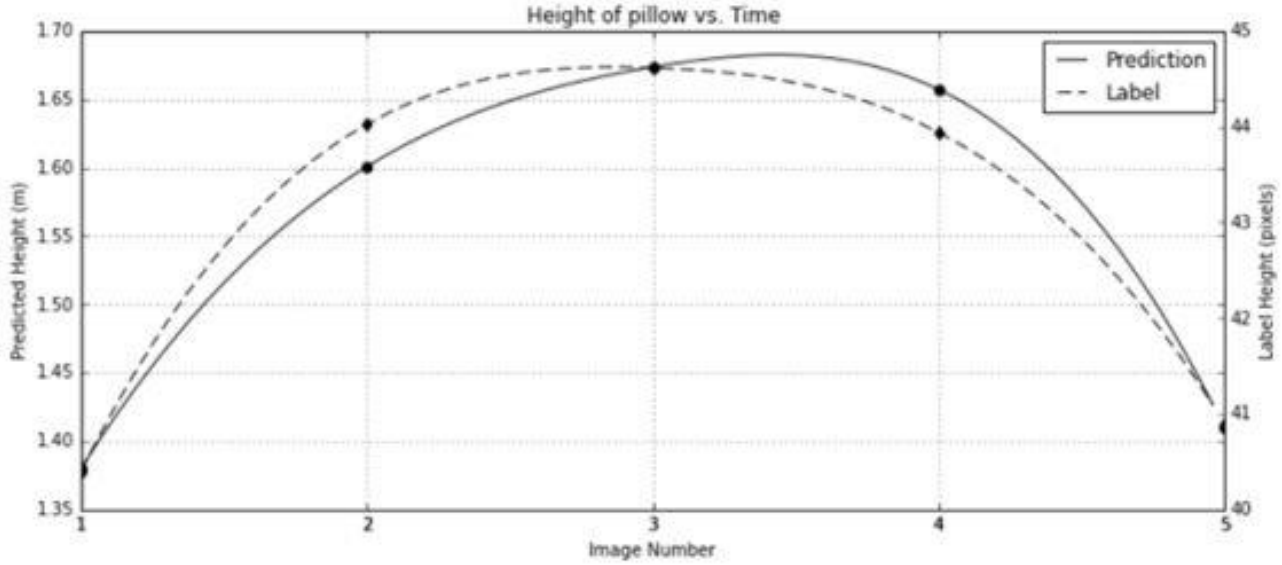
深度学习趋势

- 挑战3：大计算需要昂贵的物质、时间成本
- 趋势3：全新的硬件设计、算法设计、系统设计



深度学习趋势

- 挑战4：如何像人一样从小样本进行有效学习？
- 趋势4：数据+知识，深度学习与符号主义(知识图谱、逻辑推理)相结合



无监督信息，利用物理知识（抛物线）跟踪
(AAAI 2017最佳论文)

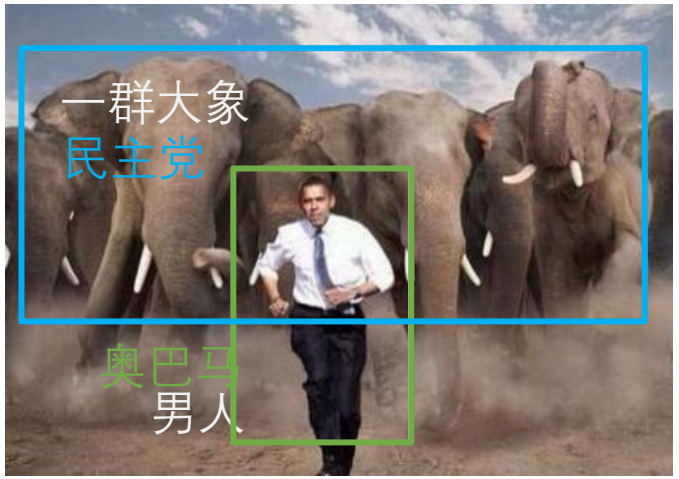
深度学习趋势

- 挑战5：如何从感知认知性的任务扩展到决策性任务？
- 趋势5：

隐含义理解

大象

民主党



博弈机器学习

参考

参考内容

[1] 复旦大学，邱锡鹏，神经网络与深度学习：https://nndl.github.io/；
[2] 北京交通大学，万怀宇，“复杂网络基础-图卷积神经网络概述”；
[3] 北京交通大学，桑基韬，“深度学习的局限与趋势”；
[4] 西安电子科技大学，高新波，“深度学习前世今生”；

部分参考文献

[1] Bruna J, Zaremba W, Szlam A, et al. Spectral Networks and Locally Connected Networks on Graphs. International Conference on Learning Representations, 2014.
[2] Defferrard M, Bresson X, Vandergheynst P, et al. Convolutional neural networks on graphs with fast localized spectral filtering[J]. Neural Information Processing Systems, 2016: 3844-3852.
[3] Kipf T N, Welling M. Semi-Supervised Classification with Graph Convolutional Networks. International Conference on Learning Representations, 2017.
[4] Seo Y, Defferrard M, Vandergheynst P, et al. Structured Sequence Modeling with Graph Convolutional Recurrent Networks. arXiv: Machine Learning, 2017.
[5] Yu B, Yin H, Zhu Z, et al. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. International Joint Conference on Artificial Intelligence, 2018: 3634-3640.
[6] Hamilton W L, Ying Z, Leskovec J, et al. Inductive Representation Learning on Large Graphs. Neural Information Processing Systems, 2017: 1024-1034.
[7] Yan S, Xiong Y, Lin D, et al. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. National Conference on Artificial Intelligence, 2018.
[8] Gao H, Wang Z, Ji S, et al. Large-Scale Learnable Graph Convolutional Networks. Knowledge Discovery and Data Mining, 2018: 1416-1424.
[9] Ying R, He R, Chen K, et al. Graph Convolutional Neural Networks for Web-Scale Recommender Systems. Knowledge Discovery and Data Mining, 2018: 974-983.
[10] Velickovic P, Cucurull G, Casanova A, et al. Graph Attention Networks. International conference on learning representations, 2018.
[11] Li Q, Han Z, Wu X, et al. Deeper Insights into Graph Convolutional Networks for Semi-Supervised Learning. National Conference on Artificial Intelligence, 2018.