Hybrid Reinforcement Learning with Expert State Sequences

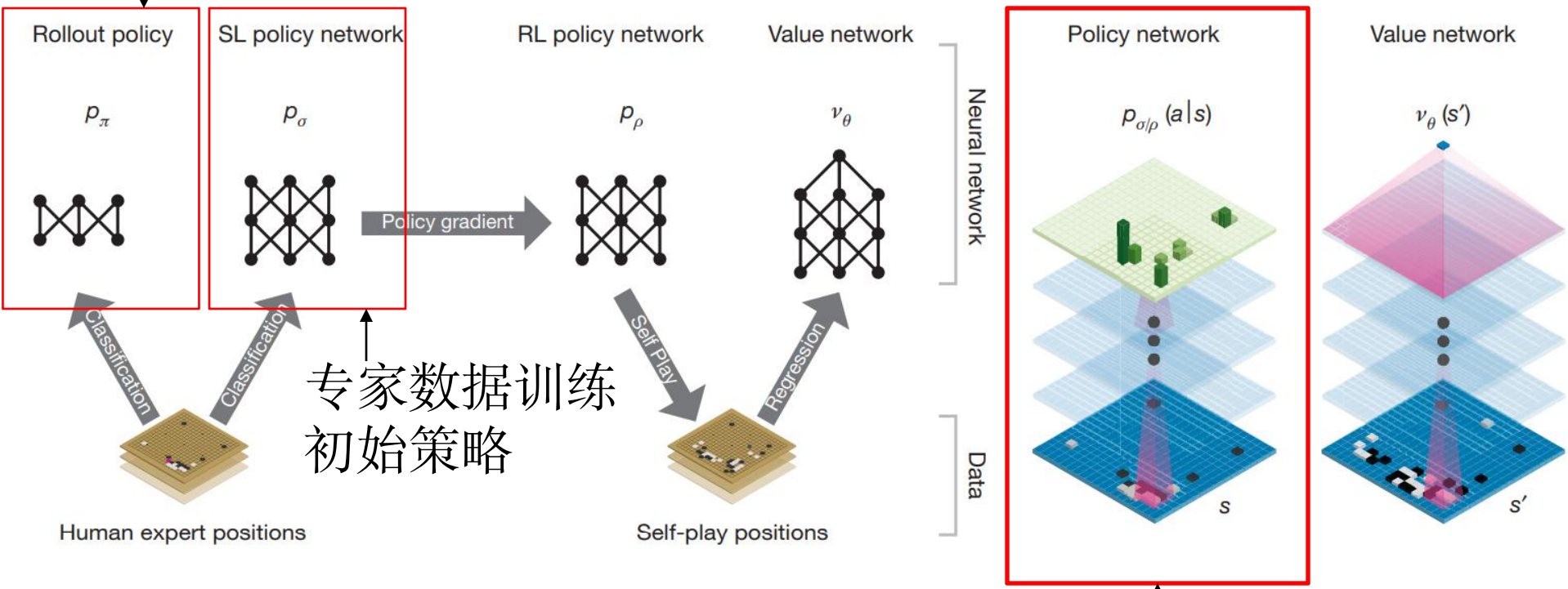
Xiaoxiao Guo, Shiyu Chang, Mo Yu, Gerald Tesauro, Murray Campbell

IBM Research AI

Presented by Quan He

Alpha-Go

专家数据训练快速规划



专家数据训练
初始策略

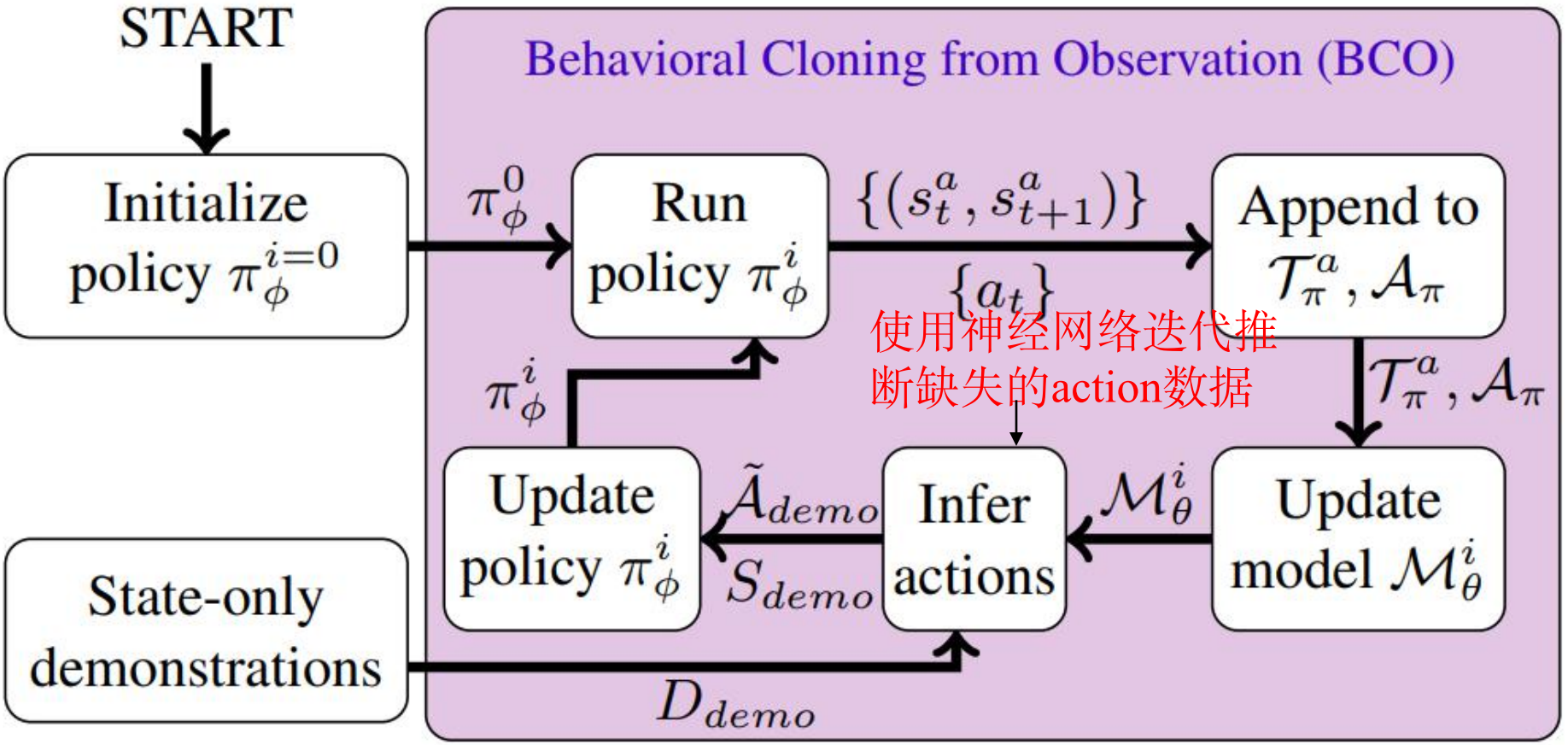
可以被视为扩展版的
Behavior-Cloning

Alpha-Go

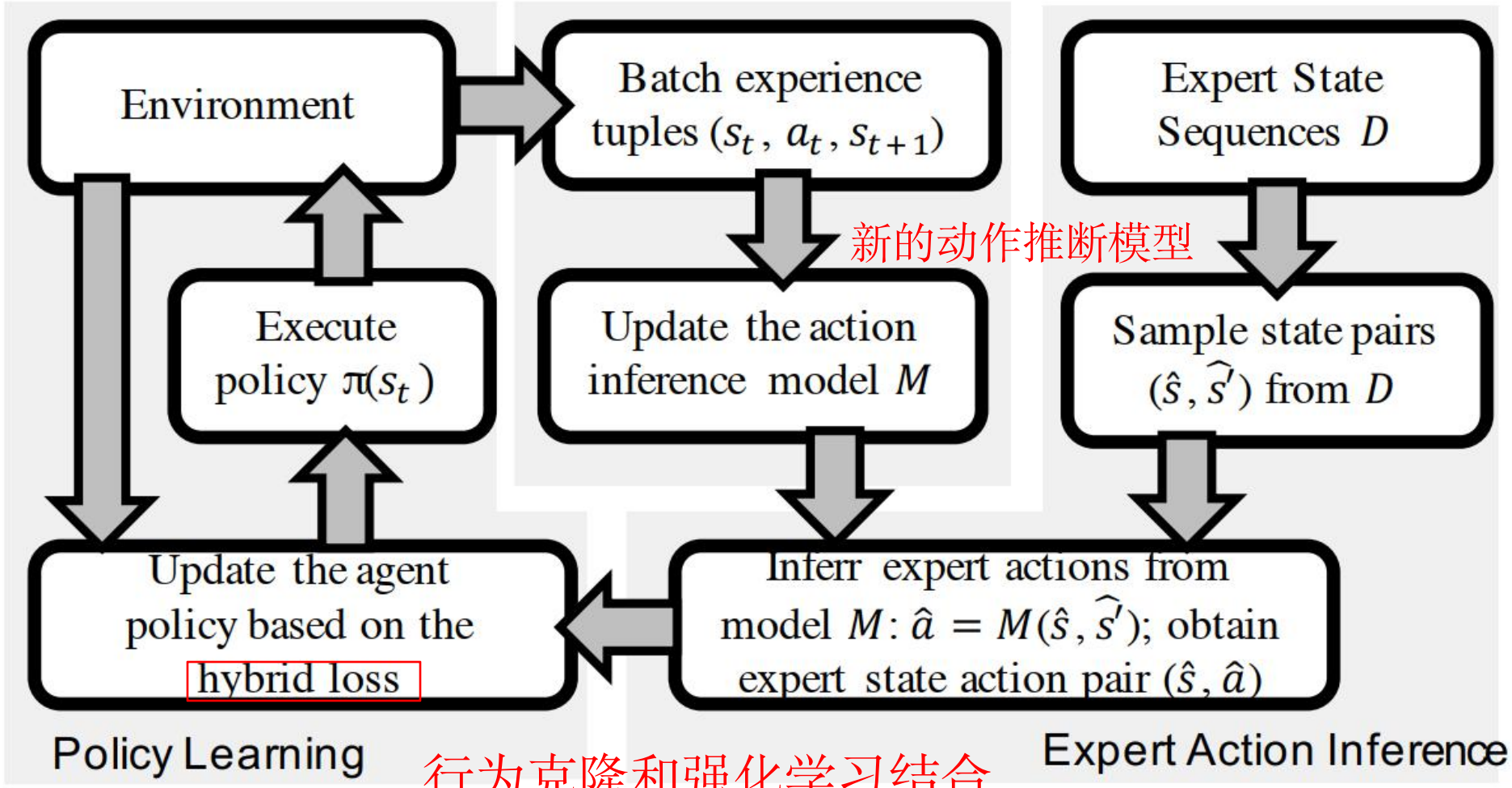
在Alpha-Go中，专家样本为棋谱（连续的state轨迹），可直接推出action（下了哪步棋），从而直接计算$P_\sigma(s, a)$进行Behavior Cloning

在其他环境中（例如机器人运动），专家样本通常为连续的状态，但未必能立即获得action信息——需要推断专家所使用的action

State-only BC Framework

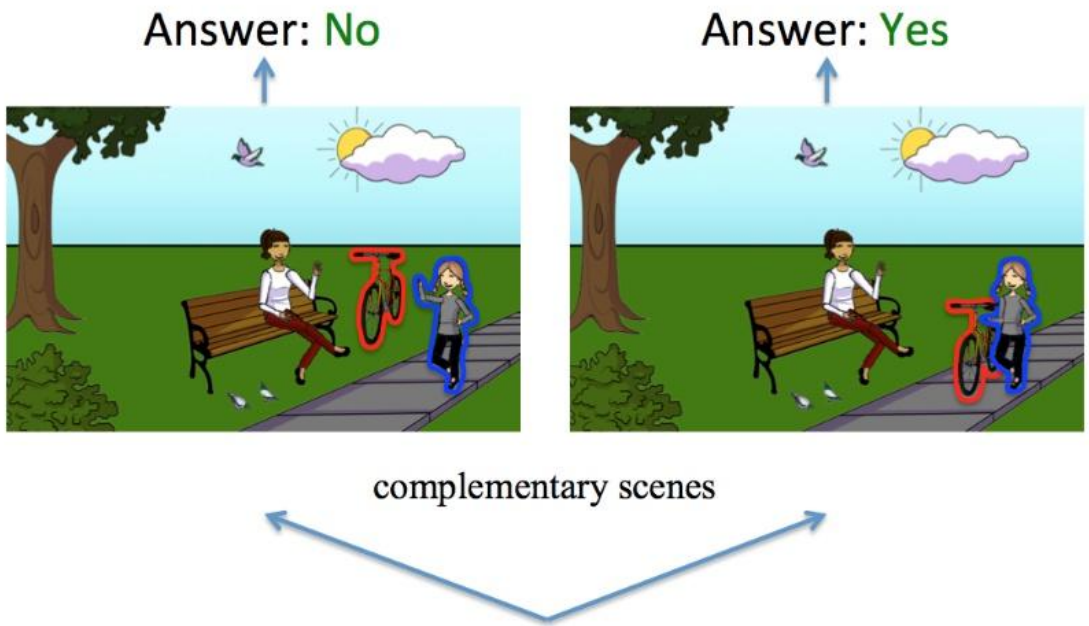


Hybrid Framework



行为克隆和强化学习结合

Visual Question Answering



已知Image,Question,求Answer： 涉及模态融合,三元组预测

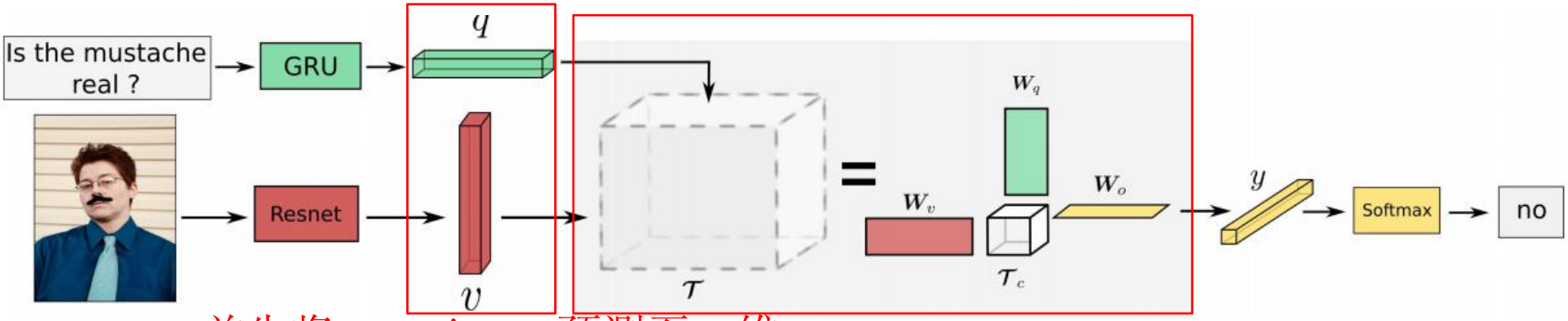
Visual Question Generation

INPUT		OUTPUT
Image	Expected answer category	Generated questions
	object	What is the person throwing? What is on top of the person's head?
	attribute	What material are those white pants made out of? What color is the frisbee?
	relationship	What is the person on the right doing with the frisbee? Did the person on the left throw or catch the frisbee?

已知Image,text,求question

MUTAN： a bilinear model

基于三维类table压缩的预测



首先将text， image预测至一维

Figure 2: MUTAN fusion scheme for global Visual QA. The prediction is modeled as a bilinear interaction between visual and linguistic features, parametrized by the tensor $\mathcal{T}$. In MUTAN, we factorise the tensor $\mathcal{T}$ using a Tucker decomposition, resulting in an architecture with three intra-modal matrices $\boldsymbol{W}_q$, $\boldsymbol{W}_v$ and $\boldsymbol{W}_o$, and a smaller tensor $\mathcal{T}_c$. The complexity of $\mathcal{T}_c$ is controlled *via* a structured sparsity constraint on the slice matrices of the tensor.

T:3-way tensor $\mathcal{T} \in R^{d_q * d_v * d_a}$

即基于三维类tabular的方法预测特定answer的条件概率

Ben-Younes, et al. MUTAN: Multimodal Tucker Fusion for Visual Question Answering, ICCV 2017

MUTAN：a bilinear model

$\times_n$:**mode-n product**（张量的模式-n积）

$$(\mathcal{T} \times_n W)_{i_1,\ldots,j,\ldots,i_k} = \sum_{i_n=1}^{d_n} \mathcal{T}_{i_1,\ldots,i_n,\ldots,i_k} W_{n,j}$$

相当于改变第n维的长度

对答案的预测$\tilde{a}$:

$$\tilde{a} = (\mathcal{T} \times_1 q) \times_2 v$$

q为GRU提取question向量， v为Resnet提取图片信息向量

例：若1(q),1(v)为one-hot 向量(在q, v处为1)， $\tilde{a}$向量为$\mathcal{T}$[q, v, :]

Ben-Younes, et al. MUTAN: Multimodal Tucker Fusion for Visual Question Answering, ICCV 2017

MUTAN：a bilinear model

tucker分解(高阶奇异值分解):

$$\mathcal{T} = ((\mathcal{T}_c \times_1 W_q) \times_2 W_v) \times_3 W_o$$

则

$$\vec{a} = ((\mathcal{T}_c \times_1 q^T W_q) \times_2 v^T W_v) \times_3 W_o$$

令$q^T W_q = q$, $v^T W_v = v$, $a^T W_o = a$,有

$$\vec{a} = (\mathcal{T}_c \times_1 \vec{q}) \times_2 \vec{v}$$

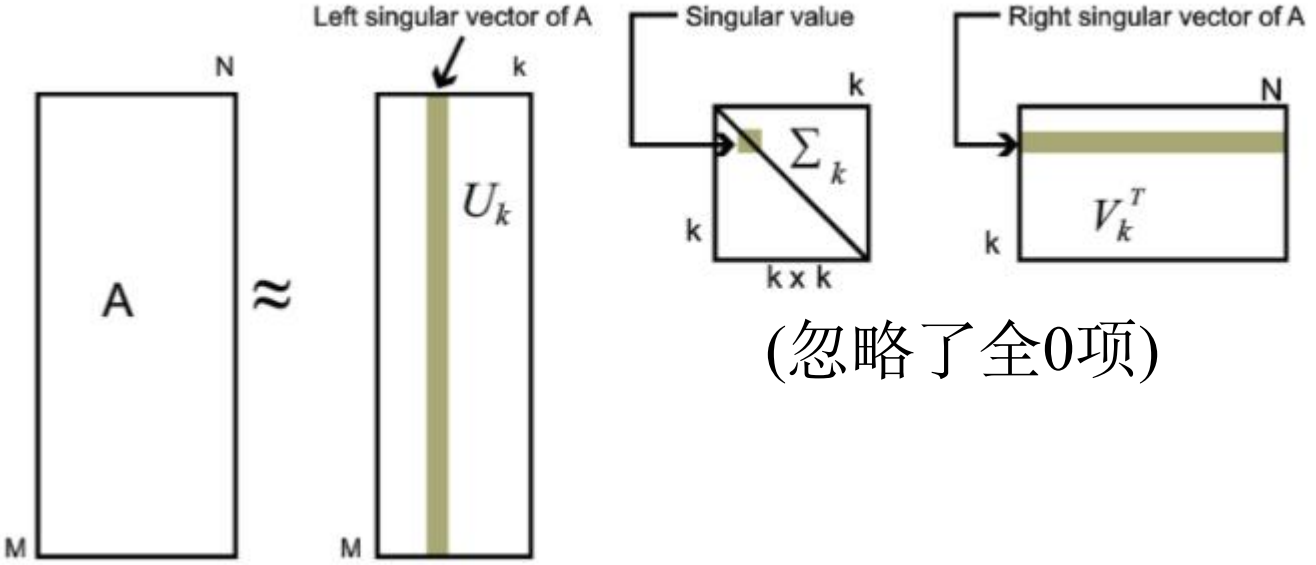
$$\vec{a}[k] = \vec{q}^T \mathcal{T}_c[:,:,k]\ \vec{v} \qquad (1)$$

Ben-Younes, et al. MUTAN: Multimodal Tucker Fusion for Visual Question Answering, ICCV 2017

MUTAN：a bilinear model

对$T_c[:,:,k]$降维(奇异值分解)：

若$T_c[:,:,k]$为m*n矩阵，则$T_c[:,:,k]=U\sum V^T$,其中U为m*m矩阵，V为n*n矩阵，$\sum$为m*n对角矩阵，Rank与$T_c[:,:,k]$相同



(忽略了全0项)

Ben-Younes, et al. MUTAN: Multimodal Tucker Fusion for Visual Question Answering, ICCV 2017

MUTAN：a bilinear model

对$\mathcal{T}_c[:,:,k]$取前R个特征值，有$\mathcal{T}_c[:,:,k]$为R个列向量与R个行向量的外积之和：

$$\mathcal{T}_c[:,:,k] = \sum_{r=1}^{R} M_r[:,k] \otimes N_r[:,k]^T$$

代入前(1)式得双线性模型MUTAN：

$$a = \sum_{r=1}^{R} (q^T M_r) * (v^T N_r)$$

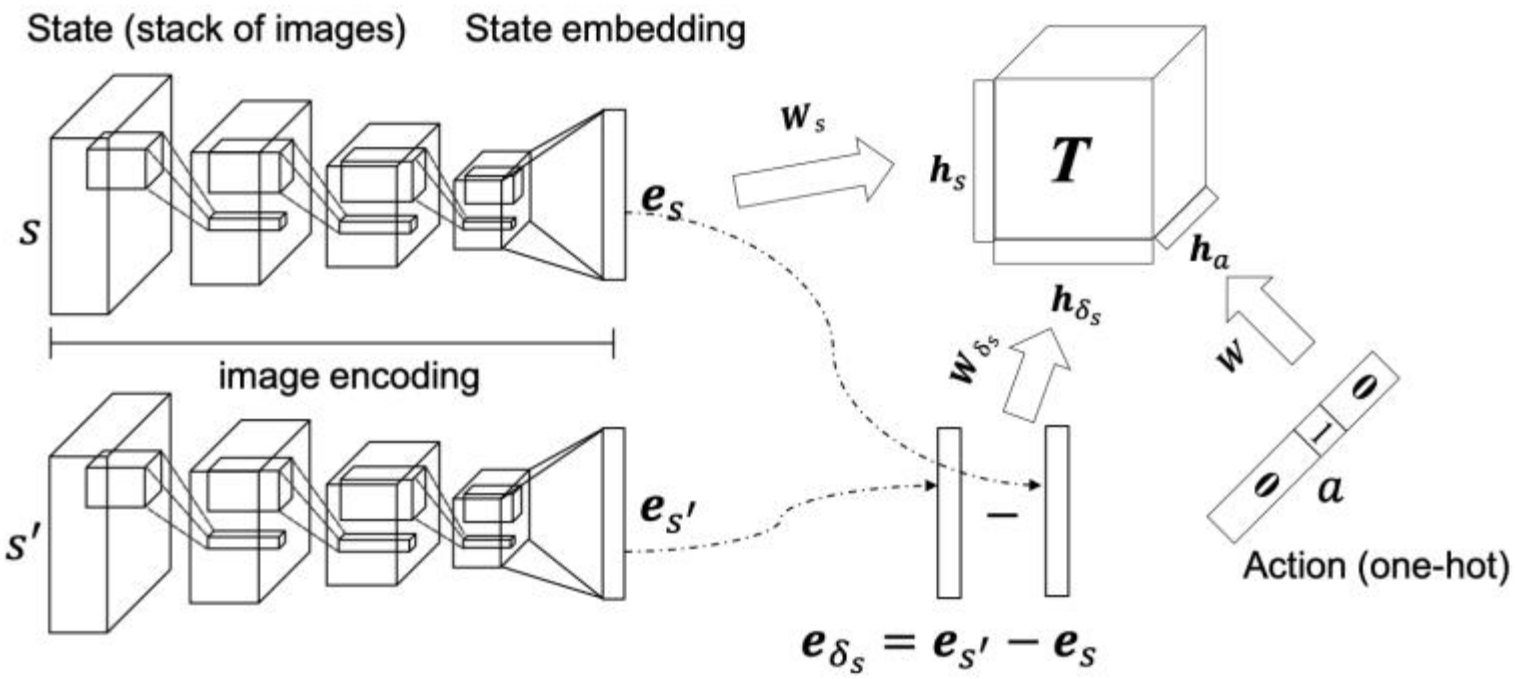
其中*为逐元素点乘

一种理解：$a_r[k] = (q \text{ similar to } M_r[:,k])$ AND $(v \text{ similar to } N_r[:,k])$

$$a[k] = a_1[k] \text{ OR } \dots \text{ OR } a_R[k]$$

Ben-Younes, et al. MUTAN: Multimodal Tucker Fusion for Visual Question Answering. ICCV 2017

tensor-based inference model



用δ_s代替s′后，使用类MUTAN模型进行预测

tensor-based inference model

$$h_{\delta_s} = \mathcal{T} \times_1 h_s \times_2 h_a$$
$$h_a = \mathcal{T} \times_1 h_s \times_3 h_{\delta_s}$$

使用双线性模型：

$$h_a = \sum_{r=1}^{R} (h_s M_r) * (h_{\delta_s} N_r)$$

假设在相同的state下，embedding所得h_a与h_{δ_s}有相似性，引入对偶性简化（借鉴了VQA与VQG对偶的思路）：

$$\mathcal{T}[:,:,k] = \mathcal{T}[:,k,:], \quad d_{\delta_s} = d_a = d$$
$$h_{\delta_s} = \sum_{r=1}^{R} (h_s M_r) * (h_a N_r)$$

tensor-based inference model

动作推断： $\mathrm{P}^{i}(.|s,s') = \mathrm{SoftMax}(\mathrm{W}_a \mathrm{h}_a + b_a)$

状态推断： $\mathrm{P}^{i}(.|s,a) = \mathrm{SoftMax}((\mathrm{W}_{\delta_s} \mathrm{h}_{\delta_s} + \mathrm{W}_s \mathrm{h}_s) + b_{s'})$

推断模型学习目标：

$$\mathcal{L}^{\text{dual-model}} = \mathbb{E}_{(s,a,s')}\Big[-\log P^f(s'|s,a) - \log P^i(a|s,s') + \|\mathbf{h}_a - \hat{\mathbf{h}}_a\|_1 + \|\mathbf{h}_{\delta s} - \hat{\mathbf{h}}_{\delta s}\|_1 \Big]$$

推断结果：

$$\mathcal{M}(\hat{s},\hat{s}') = \arg\max_a P^i(a|\hat{s},\hat{s}')$$

Hybrid BC and RL

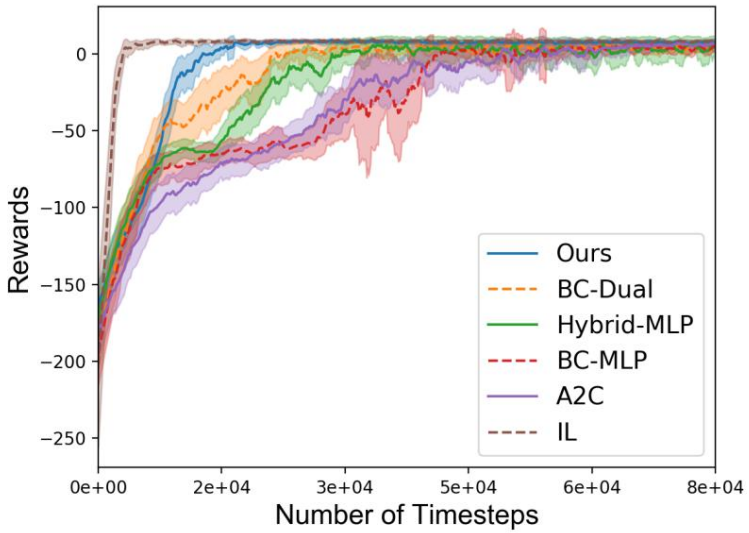
与RL（A2C）结合的策略训练目标:

$$\mathcal{U}^{\text{hybrid}}(\theta) = \mathbb{E}_{s,a}\Big[A(s)\log\pi(a|s;\theta) + \alpha\mathcal{H}(\pi(.|s))\Big] + \mathbb{E}_{(\hat{s},\hat{s}')\sim\rho(\mathcal{D})}\Big[\log\pi(\mathcal{M}(\hat{s},\hat{s}')|\hat{s};\theta)\Big]$$

其中A(s)为优势值函数

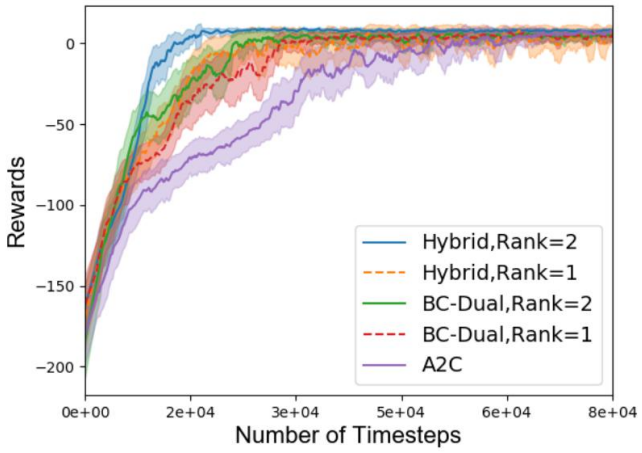
$$A(s_t) = \sum_{n=1}^{N}\gamma^{n-1}r_{t+n} + \gamma^N V(s_{t+N}) - V(s_t)$$

Experiments on Taxi Domain

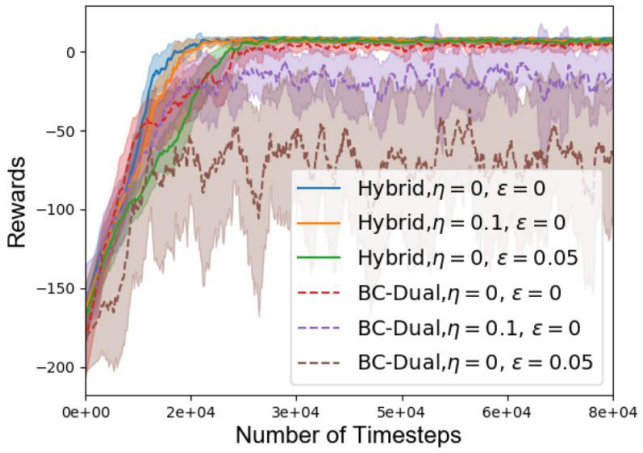


BC-Dual:仅做模仿学习，使用本文tensor-based推断模型
Hybrid-MLP：结合BC与RL，使用神经网络推断模型
BC-MLP：仅做模仿学习，使用神经网络推断模型
A2C：纯强化学习
IL：有action信息的模仿学习（相当于上界）

Experiments on Taxi Domain



(b)

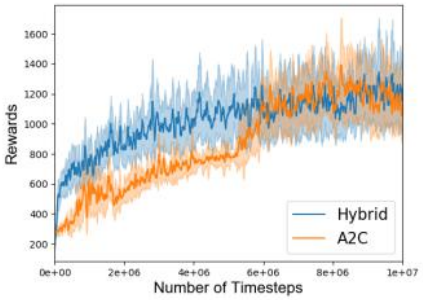


(c)

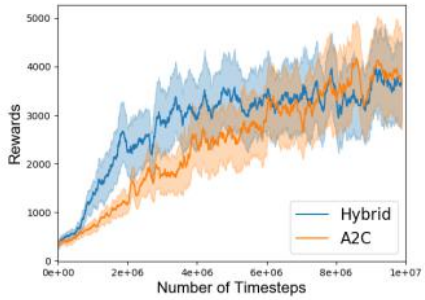
左图：推断模型rank参数的影响

右图：引入RL对模型robust能力的影响（ η 表示专家样本在state上的缺失率， ε 表示专家样本的错误率）

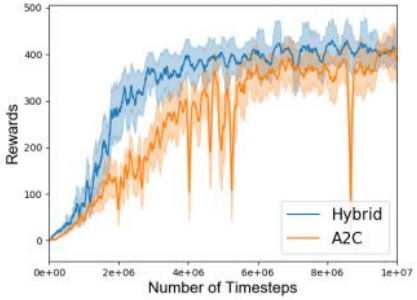
Experiments on Atari



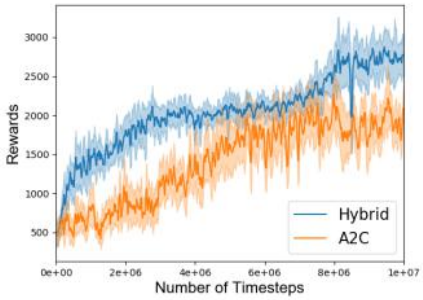
(a) Alien



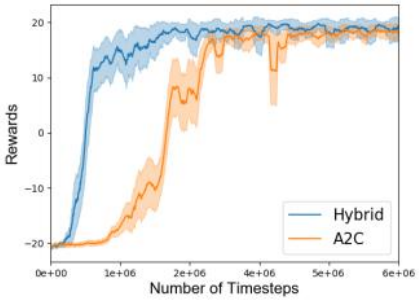
(b) BeamRider



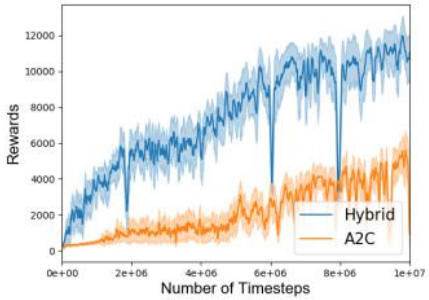
(c) Breakout



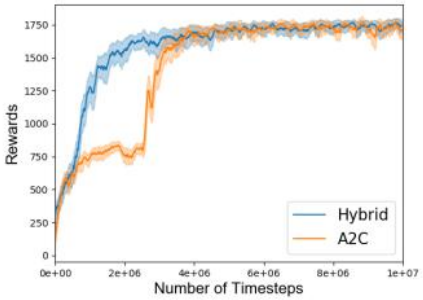
(d) MsPacman



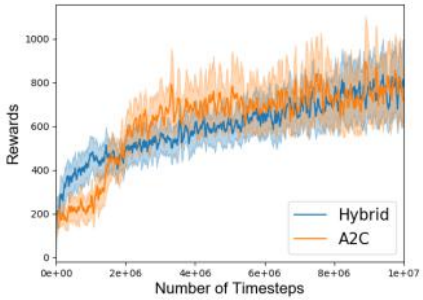
(e) Pong



(f) Qbert



(g) Seaquest



(h) SpaceInvaders


LAMDA
Learning And Mining from DatA
http://lamda.nju.edu.cn
南京大學
NANJING UNIVERSITY
1902

Thanks!