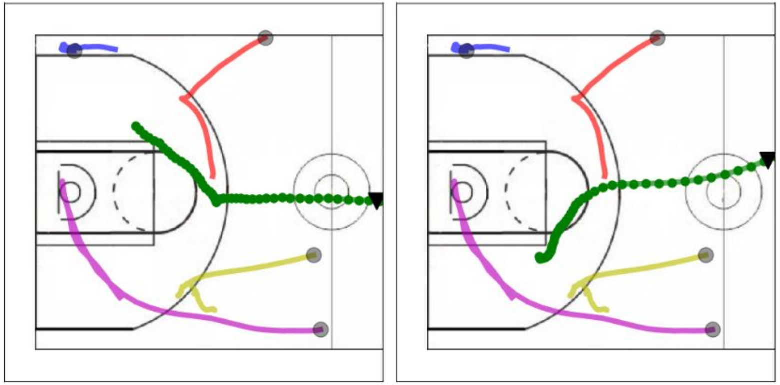
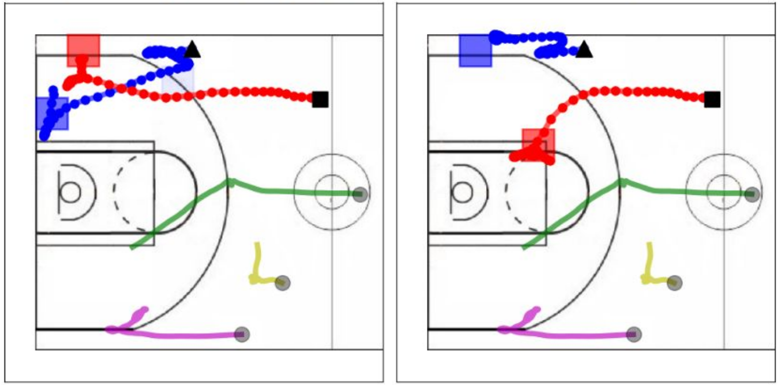
Generative Multi-Agent Behavioral Cloning

Задача

1) Рассматриваем задачу behavior cloning
2) Хотим научить агентов кооперироваться
3) Хотим, чтобы игроки были мультимодальными



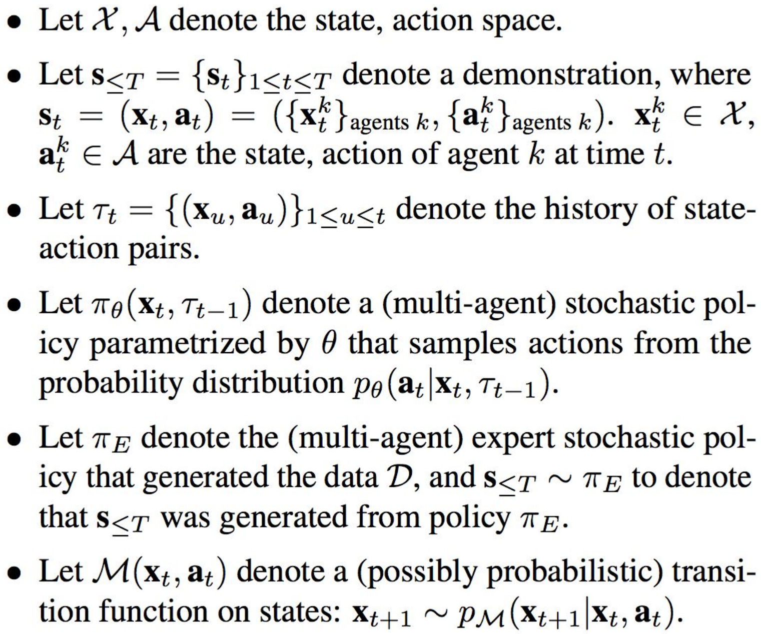
(a) Players have multi-modal behavior. For instance, in the above examples the green player (▼) moves to either the top or bottom.



(b) Players are coordinated. **Left**: The red player (■) moves to the top-left corner of the court. **Right**: The blue player (▲) moves to the top-left corner so the red player goes elsewhere. Knowing each other's macro-goals (boxes) is crucial for team coordination.

Постановка задачи

Введем обозначения



- Let $\mathcal{X}, \mathcal{A}$ denote the state, action space.
- Let $\mathbf{s}_{\leq T} = \{\mathbf{s}_t\}_{1 \leq t \leq T}$ denote a demonstration, where $\mathbf{s}_t = (\mathbf{x}_t, \mathbf{a}_t) = (\{\mathbf{x}_t^k\}_{\text{agents } k}, \{\mathbf{a}_t^k\}_{\text{agents } k})$. $\mathbf{x}_t^k \in \mathcal{X}$, $\mathbf{a}_t^k \in \mathcal{A}$ are the state, action of agent k at time t.
- Let $\tau_t = \{(\mathbf{x}_u, \mathbf{a}_u)\}_{1 \leq u \leq t}$ denote the history of state-action pairs.
- Let $\pi_\theta(\mathbf{x}_t, \tau_{t-1})$ denote a (multi-agent) stochastic policy parametrized by θ that samples actions from the probability distribution $p_\theta(\mathbf{a}_t | \mathbf{x}_t, \tau_{t-1})$.
- Let π_E denote the (multi-agent) expert stochastic policy that generated the data $\mathcal{D}$, and $\mathbf{s}_{\leq T} \sim \pi_E$ to denote that $\mathbf{s}_{\leq T}$ was generated from policy π_E.
- Let $\mathcal{M}(\mathbf{x}_t, \mathbf{a}_t)$ denote a (possibly probabilistic) transition function on states: $\mathbf{x}_{t+1} \sim p_{\mathcal{M}}(\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{a}_t)$.

Постановка задачи

Что оптимизируем?

$$\theta^* = \operatorname{argmin}_\theta \mathbb{E}_{\mathbf{s}_{\leq T} \sim \pi_E} \left[\sum_{t=1}^{T} \ell\big(\mathbf{a}_t, \pi_\theta(\mathbf{x}_t, \tau_{t-1})\big) \right]$$

$$\approx \operatorname{argmin}_\theta \sum_{\mathcal{D}} \sum_{t=1}^{T} \ell\big(\mathbf{a}_t, \pi_\theta(\mathbf{x}_t, \tau_{t-1})\big) \qquad (9)$$

Упростим себе жизнь

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{a}_t. \qquad (10)$$

Что оптимизируем дубль 2?

$$\theta^* = \operatorname{argmin}_\theta \sum_{\mathcal{D}} \sum_{t=1}^{T} \ell\big(\mathbf{x}_t, \pi_\theta(\tau_{t-1})\big) \qquad (11)$$

Напишем лосс в виде правдоподобия

$$\ell\big(\mathbf{x}_t, \pi_\theta(\tau_{t-1})\big) = -\log p_\theta\big(\mathbf{x}_t | \tau_{t-1}\big), \qquad (13)$$

$$\theta^* = \operatorname{argmax}_\theta \sum_{\mathcal{D}} \sum_{t=1}^{T} \log p_\theta\big(\mathbf{x}_t | \tau_{t-1}\big). \qquad (14)$$

Перепишем все еще раз с учетом того, что у нас много агентов

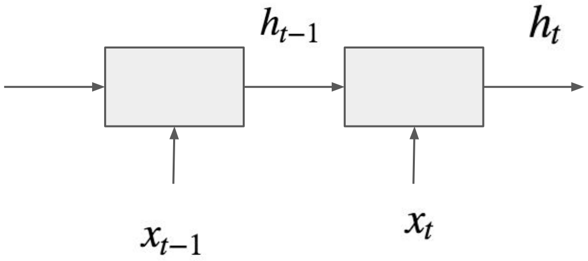
$$\theta^* = \operatorname{argmin}_\theta \sum_{\mathcal{D}} \sum_{t=1}^{T} \ell\big(\mathbf{x}_t, \pi_\theta(\tau_{t-1})\big)$$

$$= \operatorname{argmin}_\theta \sum_{\mathcal{D}} \sum_{t=1}^{T} \sum_{k=1}^{K} \ell\big(\mathbf{x}_t^k, \pi_\theta^k(\tau_{t-1})\big)$$

$$= \operatorname{argmax}_\theta \sum_{\mathcal{D}} \sum_{t=1}^{T} \sum_{k=1}^{K} \log p_\theta^k(\mathbf{x}_t^k | \mathbf{x}_{<t}) \qquad (16)$$

Модели, которые авторы используют

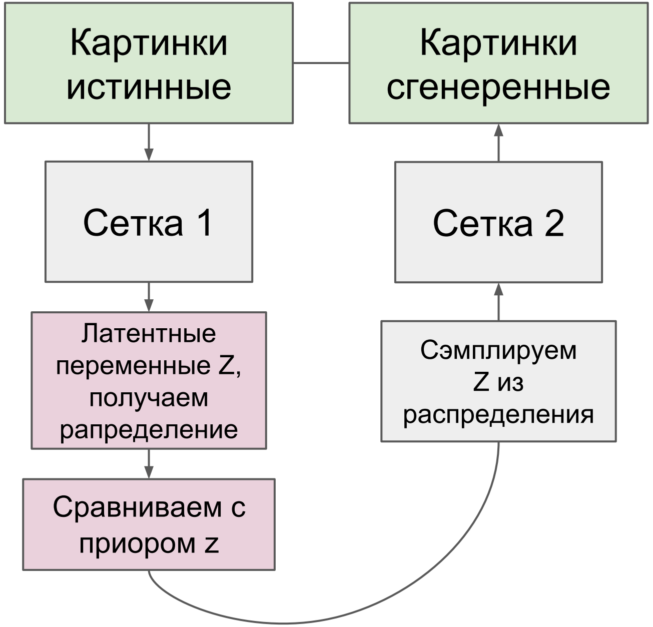
RNN



$$p_\theta(\mathbf{x}_t|\mathbf{x}_{<t}) = \varphi(\mathbf{h}_{t-1}), \quad (1)$$

$$\mathbf{h}_t = f(\mathbf{x}_t, \mathbf{h}_{t-1}), \quad (2)$$

VAE


Картинки истинные

Картинки сгенеренные

Сетка 1

Сетка 2

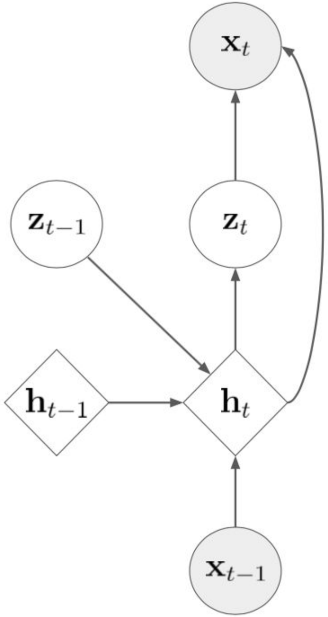
Латентные переменные Z, получаем рапределение

Сэмплируем Z из распределения

Сравниваем с приором z

$$\boxed{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]} + \boxed{D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x})||p_\theta(\mathbf{z}))} \quad (3)$$

VRNN



(a) VRNN

$$p_\theta(\mathbf{z}_t|\mathbf{x}_{<t}, \mathbf{z}_{<t}) = \varphi_{\text{prior}}(\mathbf{h}_{t-1}) \qquad \text{(prior)} \qquad (4)$$

$$q_\phi(\mathbf{z}_t|\mathbf{x}_{\leq T}, \mathbf{z}_{<t}) = \varphi_{\text{enc}}(\mathbf{x}_t, \mathbf{h}_{t-1}) \qquad \text{(inference)} \qquad (5)$$

$$p_\theta(\mathbf{x}_t|\mathbf{z}_{\leq t}, \mathbf{x}_{<t}) = \varphi_{\text{dec}}(\mathbf{z}_t, \mathbf{h}_{t-1}) \qquad \text{(generation)} \qquad (6)$$

$$\mathbf{h}_t; = f(\mathbf{x}_t, \mathbf{z}_t, \mathbf{h}_{t-1}). \qquad \text{(recurrence)} \qquad (7)$$

Добавим макроцели

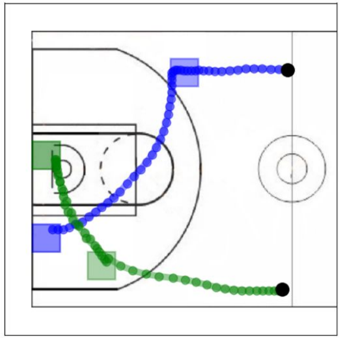


Figure 2. Showing macro-goals (boxes) for two players.

$$p(\mathbf{g}_t|\mathbf{g}_{<t}) = \varphi_g(\mathbf{h}_{g,t-1}, \mathbf{x}_{t-1}), \qquad (20)$$

$$\mathbf{h}_{g,t} = f_g(\mathbf{g}_t, \mathbf{h}_{g,t-1}). \qquad (21)$$

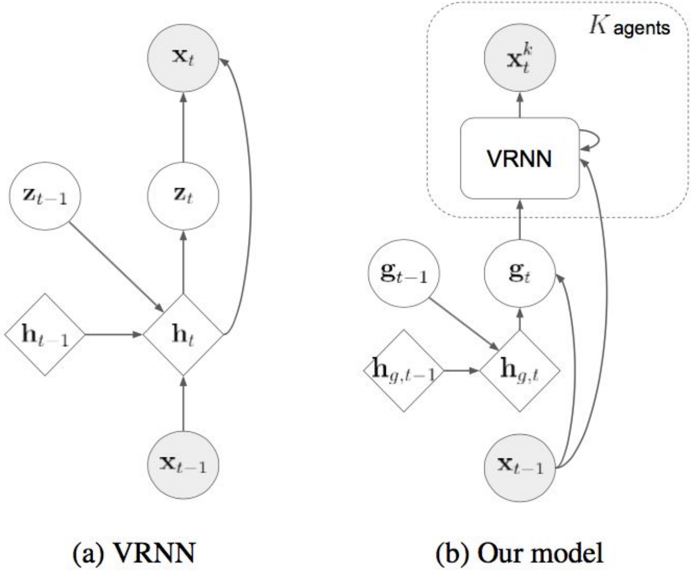


Figure 3. Computation graph of VRNN (Chung et al., 2015) and our model. Circles are stochastic variables whereas diamonds are deterministic states. Macro-goal $\mathbf{g}_t$ is shared among all agents.

Распределение макроцелей

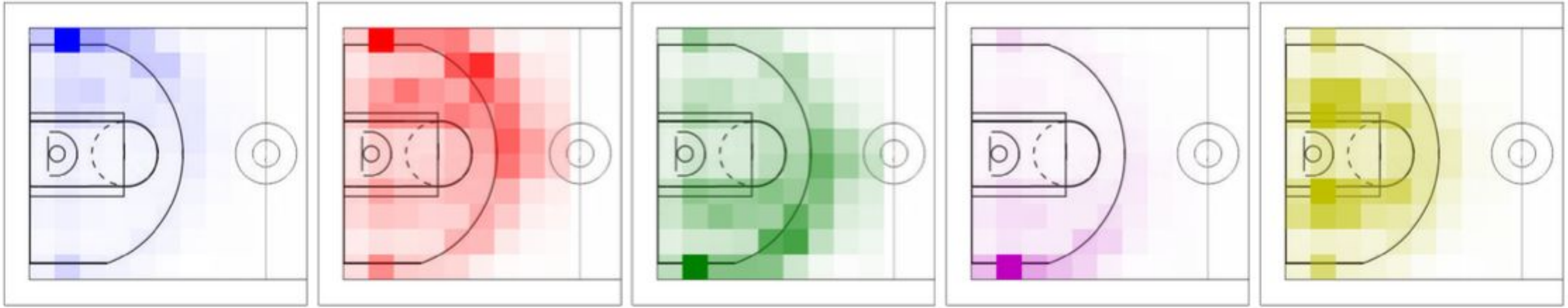
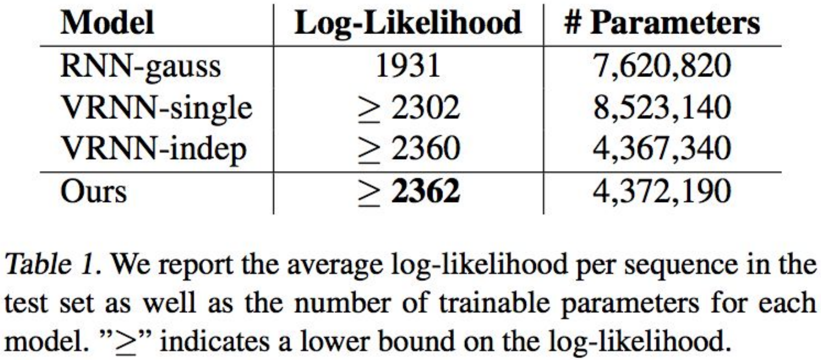


Figure 6. Distribution of weak macro-goal labels extracted for each player from the training data. Color intensity corresponds to frequency of macro-goal label. Players are ordered by their relative positions on the court, which can be seen from the macro-goals.

Эксперименты

Бэйзлайны

1) RNN-gauss
2) VRNN-single
3) VRNN-indep

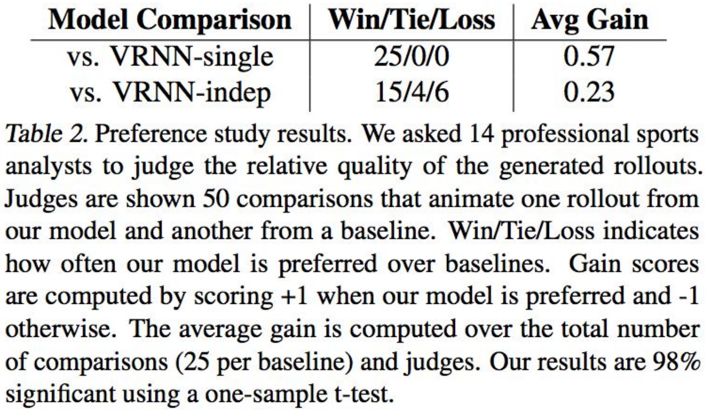


Model	Log-Likelihood	# Parameters
RNN-gauss	1931	7,620,820
VRNN-single	$\geq$ 2302	8,523,140
VRNN-indep	$\geq$ 2360	4,367,340
Ours	$\geq$ **2362**	4,372,190

Table 1. We report the average log-likelihood per sequence in the test set as well as the number of trainable parameters for each model. ”$\geq$” indicates a lower bound on the log-likelihood.

Кто лучше - мы или бэйзлайны?

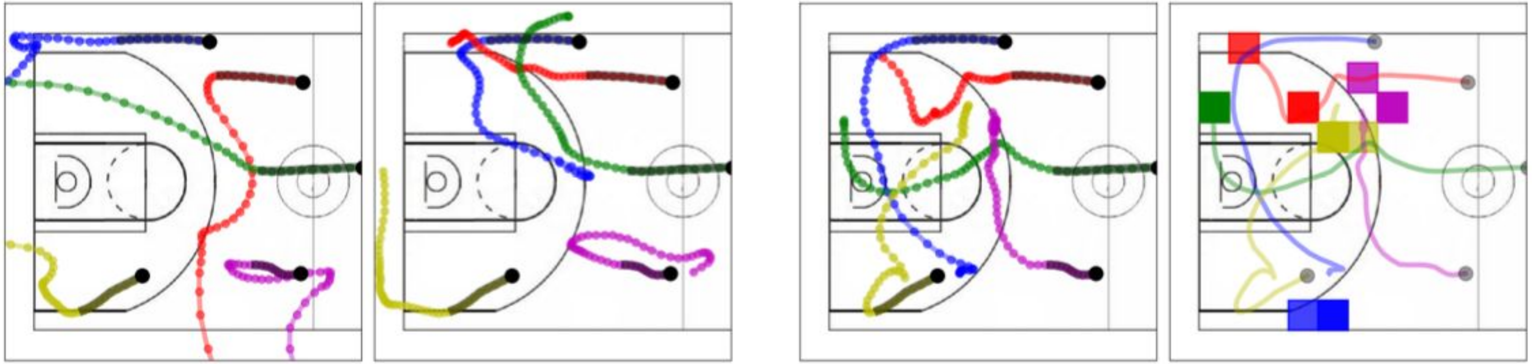
Посадим смотреть профессиональных судей творчество алгоритмов



Model Comparison	Win/Tie/Loss	Avg Gain
vs. VRNN-single	25/0/0	0.57
vs. VRNN-indep	15/4/6	0.23

Table 2. Preference study results. We asked 14 professional sports analysts to judge the relative quality of the generated rollouts. Judges are shown 50 comparisons that animate one rollout from our model and another from a baseline. Win/Tie/Loss indicates how often our model is preferred over baselines. Gain scores are computed by scoring +1 when our model is preferred and -1 otherwise. The average gain is computed over the total number of comparisons (25 per baseline) and judges. Our results are 98% significant using a one-sample t-test.

Что с кооперацией?



(a) Baseline rollouts of representative quality. **Left**: VRNN-single. **Right**: VRNN-indep. Common problems in baseline rollouts include players moving out of bounds or in the wrong direction. Players do not appear to behave cohesively as a team.

(b) **Left**: Rollout from our model. All players remain in bounds. **Right**: Corresponding macro-goals for left rollout. Macro-goal generation is stable and suggests that the team is creating more space for the blue player (perhaps setting up an isolation play).

Что с мультимодальностью?

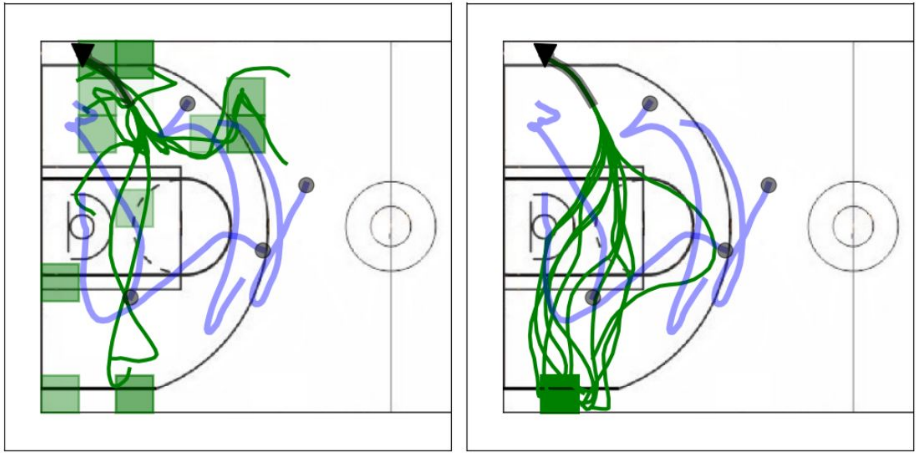


Figure 5. 10 rollouts of the green player (▼) overlayed on top of each other. A burn-in period of 20 timesteps is applied. Blue trajectories (●) are ground truth and black symbols indicate starting positions. **Left**: The model generates macro-goals. **Right**: We ground the macro-goals at the bottom-left. In both cases, we observe a multi-modal generating distribution of trajectories.

Что в итоге?

1) Сделали крутую модель для обучения с подкреплением для нескольких агентов сразу
2) Сделали предсказания мультимодальными
3) Обеспечили кооперацию
4) Превзошли другие методы
5) Успех

Спасибо за внимание!

Subscribe @tldr_arxiv