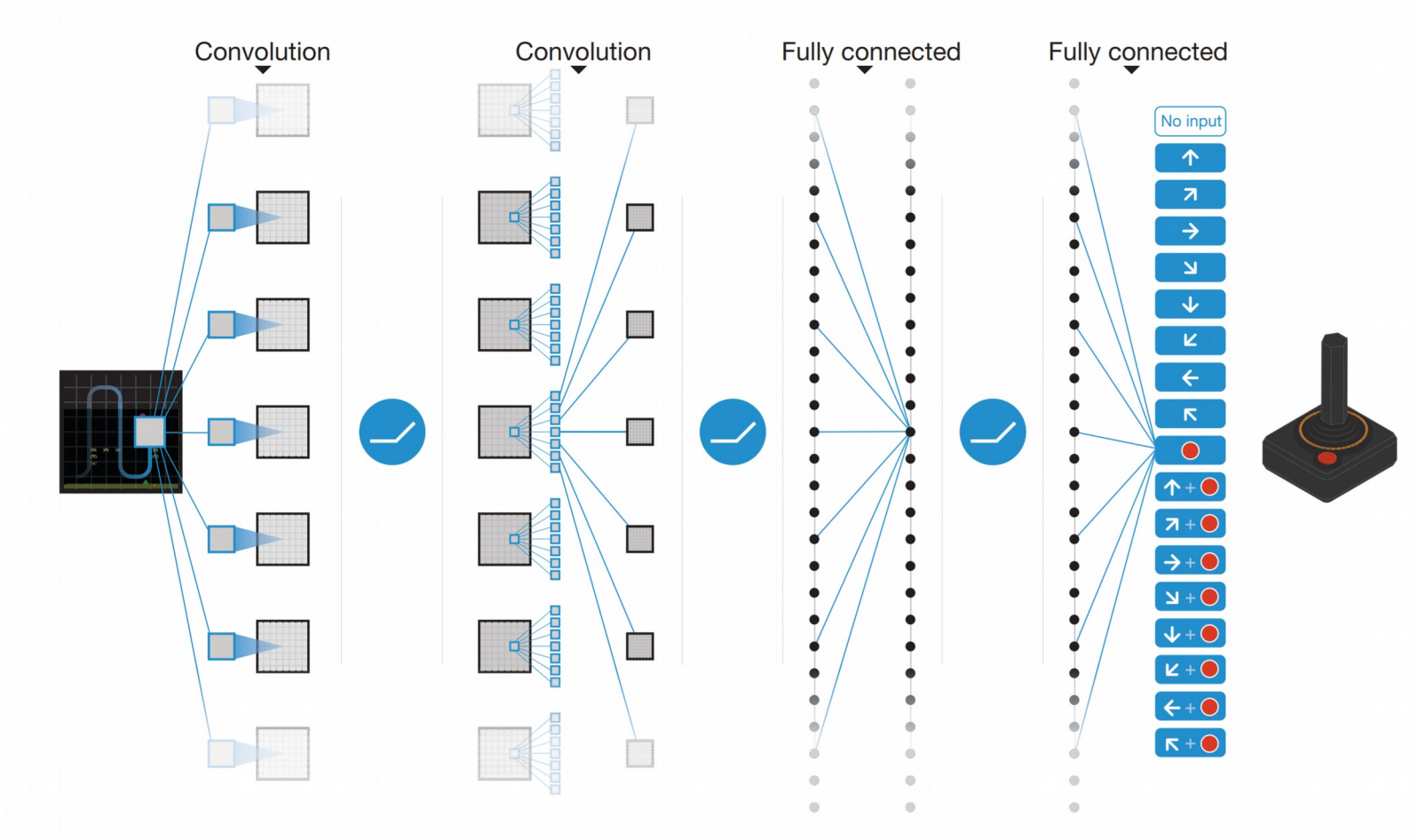
Deep Reinforcement Learning with a Natural Language Action Space

Authors: Ji He, Jianshu Chen, Xiaodong He, Jianfeng Gao, Lihong Li, Li Deng and Mari Ostendorf

Presented by: Victor Ge

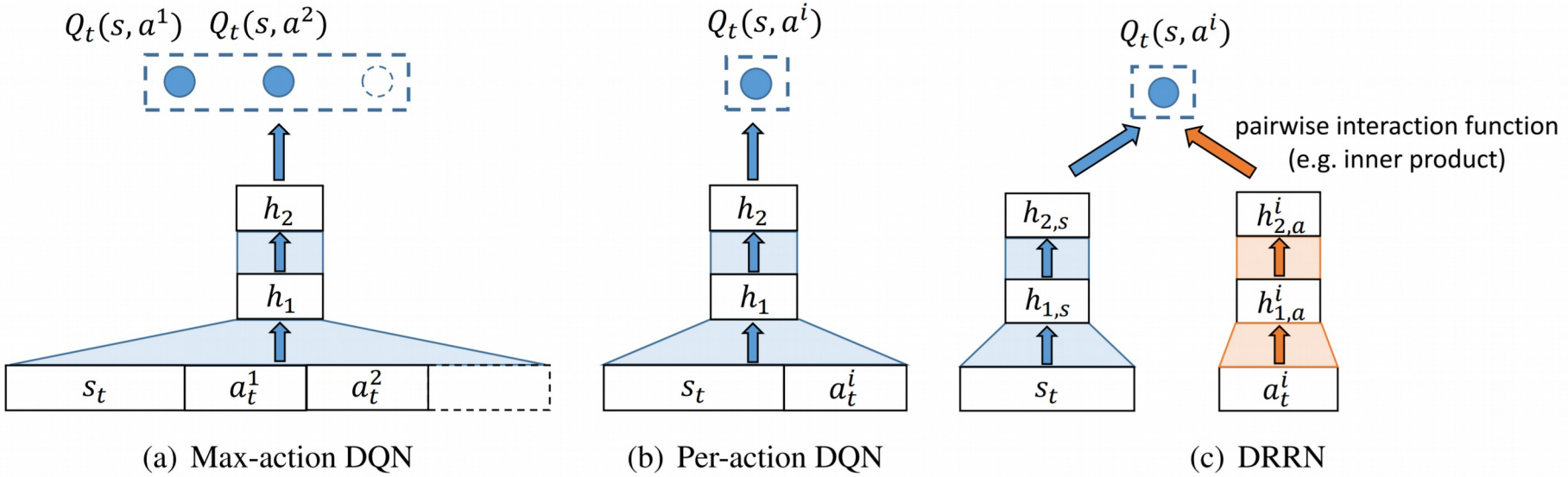
Background


Convolution
Convolution
Fully connected
Fully connected
No input

Motivation

- How to do credit assignment when the action space is discrete and potentially unbounded.
  - I.e. human-computer dialog systems, tutoring systems, and text-based games.

Q-learning architectures



Deep Reinforcement Relevance Network (DRRN)

- Factorize DQN into state representation and action representation.
- Interaction function – can be inner product, bilinear operation, nonlinear function, etc.
  - In experiments, inner product and bilinear operation give similar results.
  - Using nonlinear function (i.e. DNN) degrades performance.

Details

- Bag of words text embedding
- 1-2 hidden layers
- Experience replay buffer
- Softmax action selection:

$$\pi(a_t = a_t^i | s_t) = \frac{\exp(\alpha \cdot Q(s_t, a_t^i))}{\sum_{j=1}^{|\mathcal{A}_t|} \exp(\alpha \cdot Q(s_t, a_t^j))},$$

Experiments – text-based games

(a) Parser-based (b) Choiced-based (c) Hypertext-based

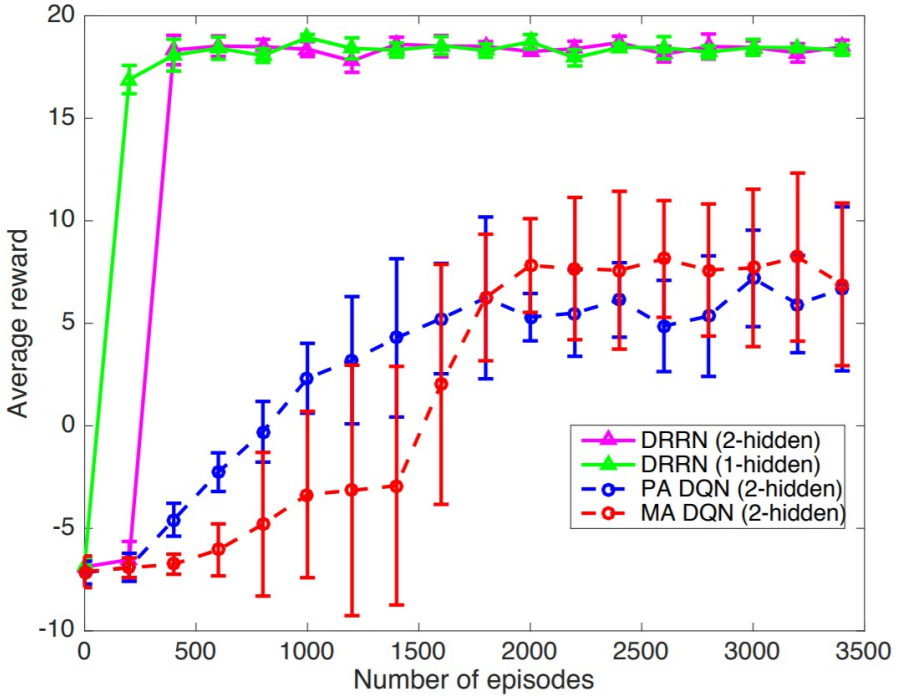
- Parser-based games can be reduced to choice-based games if there is a finite number of phrases that the parser accepts.

Experiments – text-based games

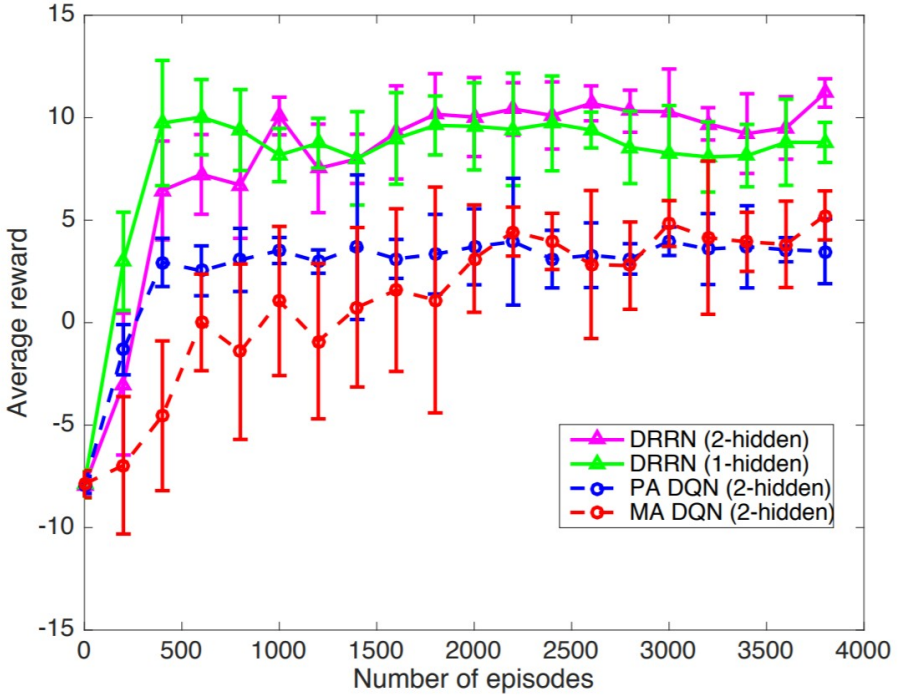
Game	Saving John	Machine of Death
Text game type	Choice	Choice & Hypertext
Vocab size	1762	2258
Action vocab size	171	419
Avg. words/description	76.67	67.80
State transitions	Deterministic	Stochastic
# of states (underlying)	≥ 70	≥ 200

Table 1: Statistics for the games "Saving John" and and "Machine of Death".

Experiments – text-based games



(a) Game 1: "Saving John"



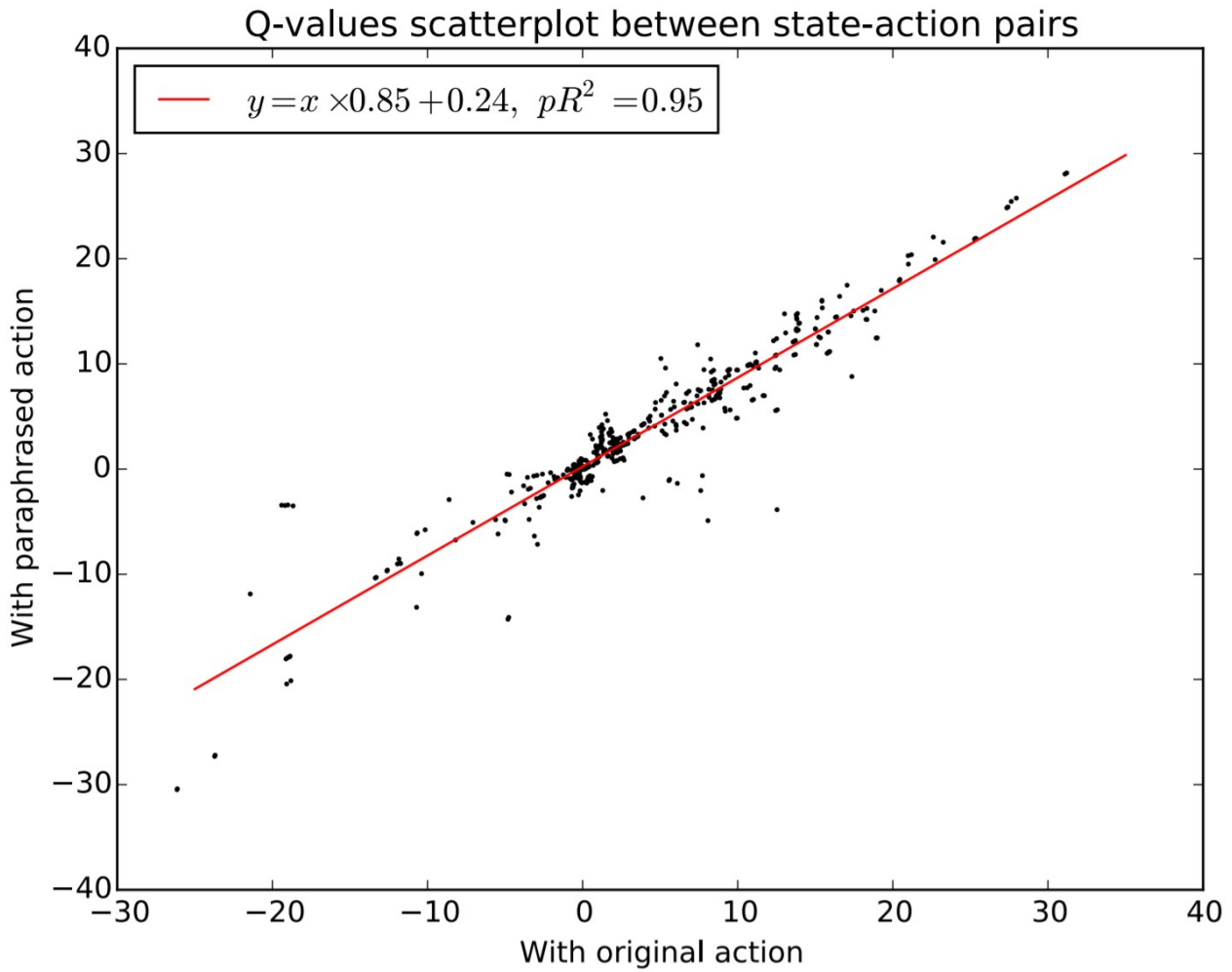
(b) Game 2: "Machine of Death"

- Human baselines:
  - "Saving John": -5.5
  - "Machine of Death": 16.0

Experiments – paraphrased actions

- Question: Is DRRN memorizing the right action?
  - State space is small (<1000)
- Replace 81.4% of action descriptions with human paraphrased descriptions.
  - Standard 4-gram BLEU score between paraphrased and original actions is 0.325
  - DRRN gets 10.5 average reward on paraphrased game vs 11.2 for original "Machine of Death" game

Experiments – paraphrased actions



Experiments – paraphrased actions

	Text (with predicted Q-values)
State	As you move forward, the people surrounding you suddenly look up with terror in their faces, and flee the street.
Actions in the original game	Ignore the alarm of others and continue moving forward. (-21.5) Look up. (16.6)
Paraphrased actions (not original)	Disregard the caution of others and keep pushing ahead. (-11.9) Turn up and look. (17.5)
Positive actions (not original)	Stay there. (2.8) Stay calmly. (2.0)
Negative actions (not original)	Screw it. I'm going carefully. (-17.4) Yell at everyone. (-13.5)
Irrelevant actions (not original)	Insert a coin. (-1.4) Throw a coin to the ground. (-3.6)

Table 4: Predicted Q-value examples