Q-Networks with Dynamically Loaded Biases for Personalization

Ján Magyar ✉, Peter Sinčák

DCAI, Technical University of Košice

Košice, Slovakia

✉ jan.magyar@tuke.sk

“Personalization is a process that changes the functionality, interface, information access and content, or distinctiveness of a system to increase its personal relevance to an individual or a category of individuals.”

Fan, Haiyan, and Marshall Scott Poole. "What is personalization? Perspectives on the design and implementation of personalization in information systems." *Journal of Organizational Computing and Electronic Commerce* 16, no. 3-4 (2006): 179-202.

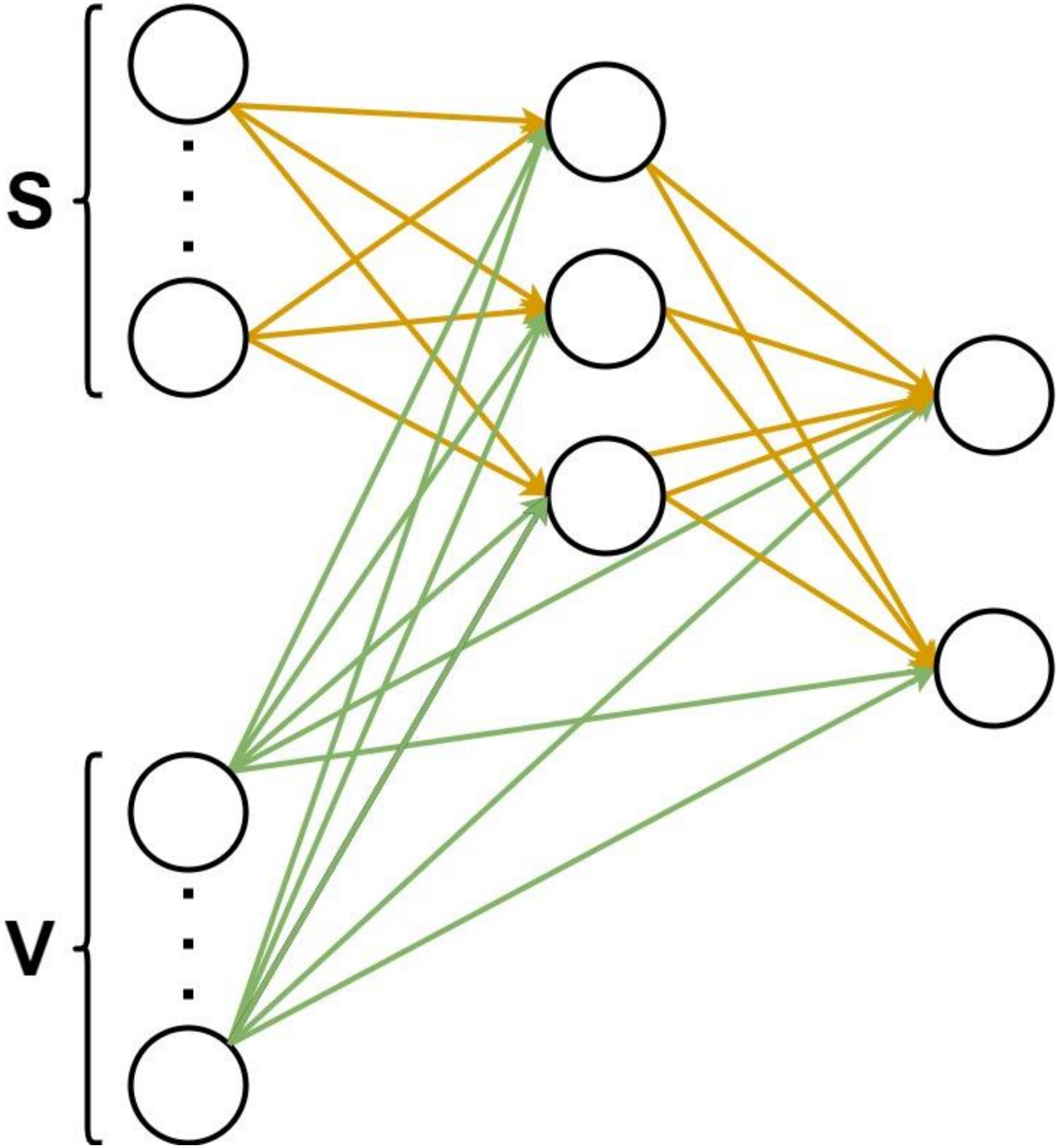
How do we personalize interactions?

- humans are hard to model
- for large states spaces use neural networks
  - one network / user – no transfer
  - one network with user information – what information is relevant?
  - one network / group – how do we group users correctly?
- limited training data

DLBQN

- **S** for state representation
- **V** sparse vector identifying the user

- hidden units have no bias
- upper half (orange) of the network is responsible for general functionality
- lower half (green) of the network personalizes the functionality



Mathematical background

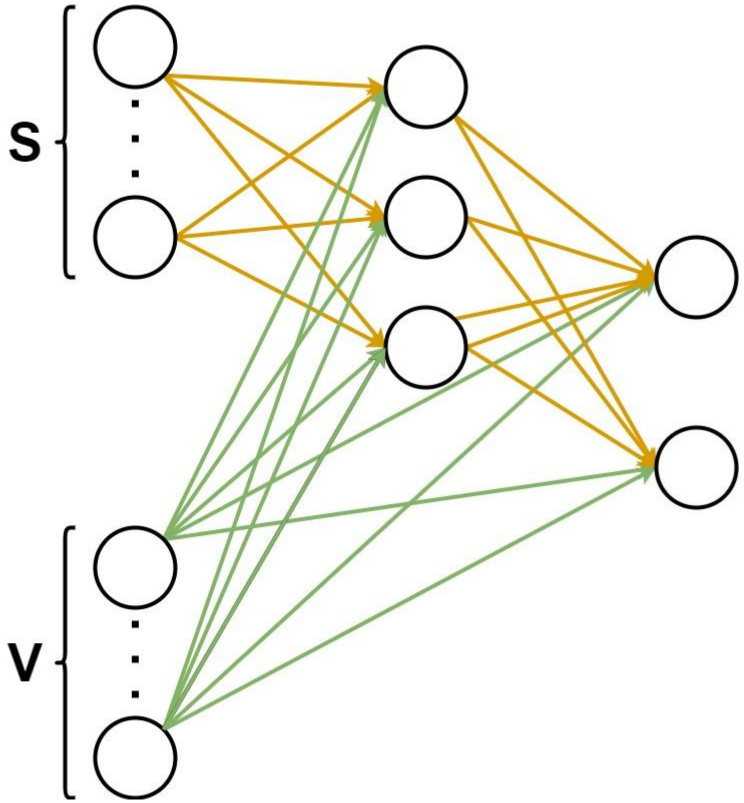
- the input function of a neuron can be calculated as:

$$Z_k = \sum_{j=1}^{m} w_{kj} x_j + \sum_{l=1}^{n} w_{kl} v_l$$

- but since **V** is a sparse vector (with a single 1 value):

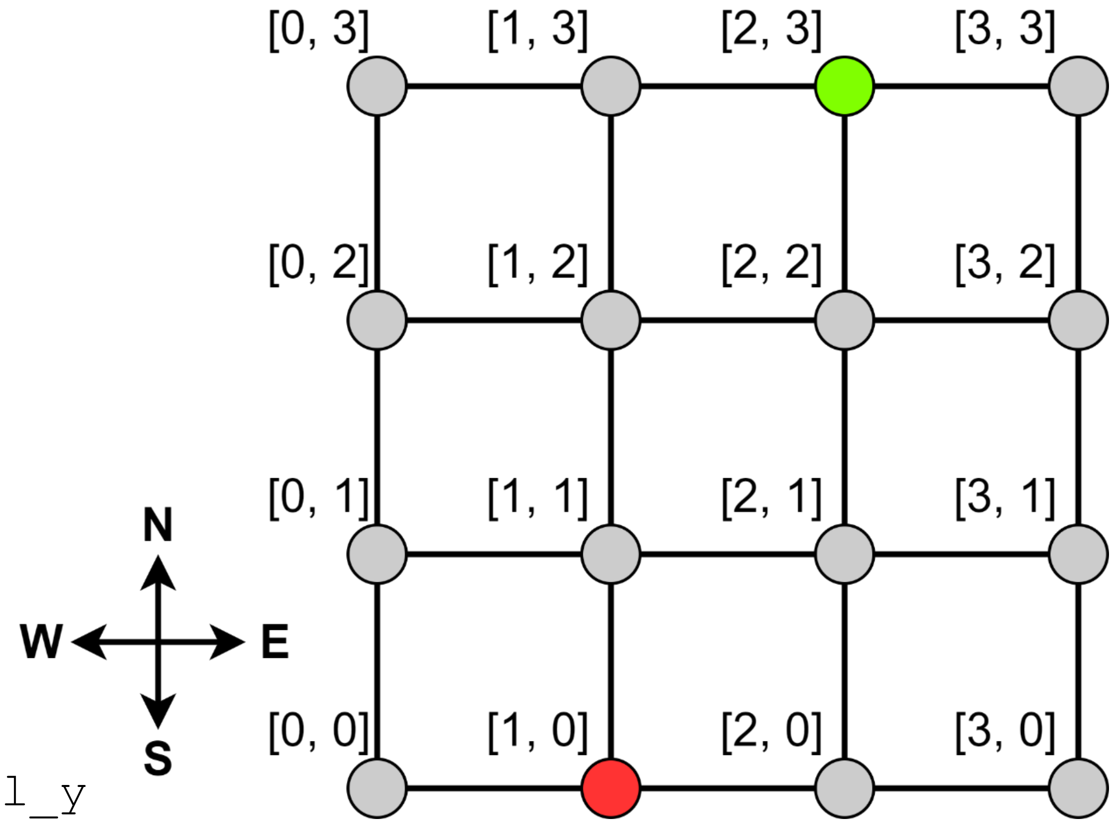
$$Z_k = \sum_{j=1}^{m} w_{kj} x_j + w_{kl}$$

where $v_l = 1$

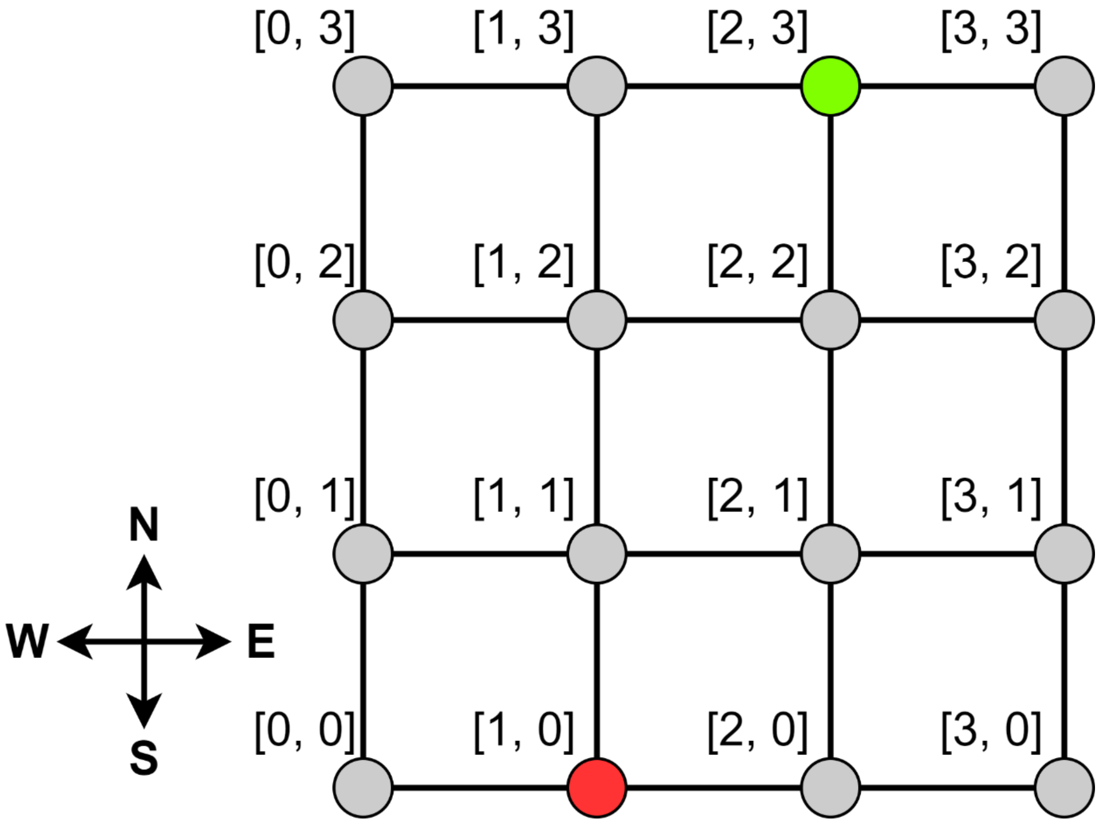


Gridworld

- 2D world with 4 actions
- two state representations:
  - `[agt_x, agt_y, goal_x, goal_y]`
  - `[agt_x, agt_y]`
- reward:
  - 1 if `agt_x == goal_x and agt_y == goal_y`
  - -1 otherwise
- training on one world and 10 worlds with randomly generated goals



Push the Box



- variation on gridworld
- the agent must cooperate with a simulated human
- two types of simulated humans
  - vertical-first
  - horizontal-first
- update state only if robot and human actions are identical

Methodology

- train for 100 epochs per environment
- precision of best policies (how often would the agent select a valid action?)
- rate of convergence (average precision over first *n* iterations)
- stability of the policy (average precision over last *n* iterations)

Precision of best policies

agent	state	worlds	max. precision		
			max	count	mean
DQN	2	1	100	100	100
		10	81.3	1	70.48
	4	1	100	100	100
		10	100	99	99.99
DLBQN	2	1	100	100	100
		10	100	67	99.32
	4	1	100	100	100
		10	100	77	99.53

gridworld

agent	state	worlds	max. precision		
			max	count	mean
DQN	2	1	100	100	100
		10	51.25	1	48.75
	4	1	100	100	100
		10	100	8	91.05
DLBQN	2	1	100	100	100
		10	100	96	99.9
	4	1	100	100	100
		10	100	98	99.98

Push the Box

Rate of convergence

agent	state	worlds	mean of first *n* iterations		
			0-5	5-10	10-30
DQN	2	1	58.97	83.88	96.99
		10	42.79	52.23	52.04
	4	1	60.61	83.85	97.49
		10	45.93	70.03	87.02
DLBQN	2	1	61.2	82.97	96.68
		10	40.59	50.2	69.12
	4	1	59.25	82.56	96.63
		10	39.3	49.64	70.91

gridworld

agent	state	worlds	mean of first *n* iterations		
			0-5	5-10	10-30
DQN	2	1	52.2	85.95	98.21
		10	33.4	33.53	34.14
	4	1	53.38	89.43	98.37
		10	37.43	89.97	75.41
DLBQN	2	1	46.55	84.08	97.81
		10	28.68	41.91	61.38
	4	1	48.03	87.08	98.21
		10	28.71	42.99	63.18

Push the Box

Stability of policy

agent	state	worlds	mean of last *n* iterations		
			0-5	5-10	10-30
DQN	2	1	99.99	99.98	99.95
		10	42.43	41.12	41.67
	4	1	99.99	99.93	99.9
		10	98.54	98.45	98.44
DLBQN	2	1	99.99	99.97	99.98
		10	96.43	96.3	96.42
	4	1	99.97	99.93	99.9
		10	96.1	96.1	96.26

gridworld

agent	state	worlds	mean of last *n* iterations		
			0-5	5-10	10-30
DQN	2	1	99.98	100	99.99
		10	24.81	24.78	25.41
	4	1	100	99.98	99.99
		10	83.06	83.18	82.98
DLBQN	2	1	100	99.98	100
		10	99.73	99.76	99.72
	4	1	100	100	99.98
		10	99.77	99.77	99.71

Push the Box

Conclusion

- a single DQN is not enough to personalize, we would need to train one network/environment
- DLBQN provides personalized policies for different environments, even for two environments with the same goal position
- for the same topology (weights *w* and biases *b*) training on *n* worlds requires
  - $n \cdot (|w| + |b|)$ parameters for *n* DQN networks
  - $|w| + n \cdot |b|$ parameters for a single DLBQN network
- the convergence rate of DLBQN is slightly lower, which must be addressed in the future