Supplementary Material for "Deep Reinforcement Learning with a Natural Language Action Space"

A Percentage of Choice-based and Hypertext-based Text Games

As shown in Table 1.[1]

Year	2010	2011	2012	2013	2014
Percentage	7.69%	7.89%	25.00%	55.56%	61.90%

Table 1: Percentage of choice-based and hypertext-based text games since 2010, in archive of interactive fictions

B Back Propagation Formula for Learning DRRN

Let $h_{l,s}$ and $h_{l,a}$ denote the l-th hidden layer for state and action side neural networks, respectively. For state side, $W_{l,s}$ and $b_{l,s}$ denote the linear transformation weight matrix and bias vector between the $(l-1)$-th and l-th hidden layers. For actions side, $W_{l,a}$ and $b_{l,a}$ denote the linear transformation weight matrix and bias vector between the $(l-1)$-th and l-th hidden layers. The DRRN has L hidden layers on each side.

Forward:

$$h_{1,s} = f(W_{1,s}s_t + b_{1,s}) \tag{1}$$

$$h^i_{1,a} = f(W_{1,a}a^i_t + b_{1,a}), \quad i = 1, 2, 3, ..., |\mathcal{A}_t| \tag{2}$$

$$h_{l,s} = f(W_{l-1,s}h_{l-1,s} + b_{l-1,s}), \quad l = 2, 3, ..., L \tag{3}$$

$$h^i_{l,a} = f(W_{l-1,a}h^i_{l-1,a} + b_{l-1,a}), \quad i = 1, 2, 3, ..., |\mathcal{A}_t|, l = 2, 3, ..., L \tag{4}$$

$$Q(s_t, a^i_t) = h^T_{L,s}h^i_{L,a} \tag{5}$$

where $f(\cdot)$ is the nonlinear activation function at the hidden layers, which is chosen as $\tanh(x) = (1 - \exp(-2x))/(1 + \exp(-2x))$, and $\mathcal{A}_t$ denotes the set of all actions at time t.

Backward:

Note we only back propagate for actions that are actually taken. More formally, let a_t be action the DRRN takes at time t, and denote $\Delta = [Q(s_t, a_t) -$

[1] Statistics are obtained from http://www.ifarchive.org

Reward	Endings (partially shown)
-20	Suspicion fills my heart and I scream. Is she trying to kill me? I don't trust her one bit...
-10	Submerged under water once more, I lose all focus...
0	Even now, she's there for me. And I have done nothing for her...
10	Honest to God, I don't know what I see in her. Looking around, the situation's not so bad...
20	Suddenly I can see the sky... I focus on the most important thing - that I'm happy to be alive.

Table 2: Final rewards defined for the text game "Saving John"

$(r_t + \gamma \max_a Q(s_{t+1}, a))]^2/2$. Denote $\delta_{l,s} = \delta b_{l,s} = \partial Q/\partial b_{l,s}$, $\delta_{l,a} = \delta b_{l,a} = \partial Q/\partial b_{l,a}$, and we have (by following chain rules):

$$\delta Q = \frac{\partial \Delta}{\partial Q} = Q(s_t, a_t) - (r_t + \gamma \max_a Q(s_{t+1}, a)) \tag{6}$$

$$\begin{cases} \delta_{L,s} = \delta Q \cdot h_{L,a} \odot (1 - h_{L,s}) \odot (1 + h_{L,s}) \\ \delta_{l-1,s} = W_{l,s}^T \delta_{l,s} \odot (1 - h_{l-1,s}) \odot (1 + h_{l-1,s}), \quad l = 2, 3, ..., L \end{cases} \tag{7}$$

$$\begin{cases} \delta_{L,a} = \delta Q \cdot h_{L,s} \odot (1 - h_{L,a}) \odot (1 + h_{L,a}) \\ \delta_{l-1,a} = W_{l,a}^T \delta_{l,a} \odot (1 - h_{l-1,a}) \odot (1 + h_{l-1,a}), \quad l = 2, 3, ..., L \end{cases} \tag{8}$$

$$\begin{cases} \delta W_{1,s} = \partial Q/\partial W_{1,s} = \delta_{1,s} \cdot s_t^T \\ \delta W_{l,s} = \partial Q/\partial W_{l,s} = \delta_{l,s} \cdot h_{l-1,s}^T, \quad l = 2, 3, ..., L \end{cases} \tag{9}$$

$$\begin{cases} \delta W_{1,a} = \partial Q/\partial W_{1,a} = \delta_{1,a} \cdot a_t^T \\ \delta W_{l,a} = \partial Q/\partial W_{l,a} = \delta_{l,a} \cdot h_{l-1,a}^T, \quad l = 2, 3, ..., L \end{cases} \tag{10}$$

where $\odot$ denotes element-wise Hadamard product.

C Final Rewards in the Two Text Games

As shown in Table 2 and Table 3.

D Game 2 Learning curve with shared state and action embedding

As shown in Figure 1. For the first 1000 episodes, parameter tying gives faster convergence, but learning curve also has high variance and unstable.

Reward	Endings (partially shown)
-20	You spend your last few moments on Earth lying there, shot through the heart, by the image of Jon Bon Jovi.
-20	you hear Bon Jovi say as the world fades around you.
-20	As the screams you hear around you slowly fade and your vision begins to blur, you look at the words which ended your life.
-10	You may be locked away for some time.
-10	Eventually you're escorted into the back of a police car as Rachel looks on in horror.
-10	Fate can wait.
-10	Sadly, you're so distracted with looking up the number that you don't notice the large truck speeding down the street.
-10	All these hiccups lead to one grand disaster.
10	Stay the hell away from me! She blurts as she disappears into the crowd emerging from the bar.
20	You can't help but smile.
20	Hope you have a good life.
20	Congratulations!
20	Rachel waves goodbye as you begin the long drive home. After a few minutes, you turn the radio on to break the silence.
30	After all, it's your life. It's now or never. You ain't gonna live forever. You just want to live while you're alive.

Table 3: Final rewards for the text game "Machine of Death." Scores are assigned according to whether the character survives, how the friendship develops, and whether he overcomes his fear.

E Examples of State-Action Pairs in the Two Text Games

As shown in Table 4 and Table 5.

F Examples of State-Action Pairs that do not exist in the feasible set

As shown in Table 6.

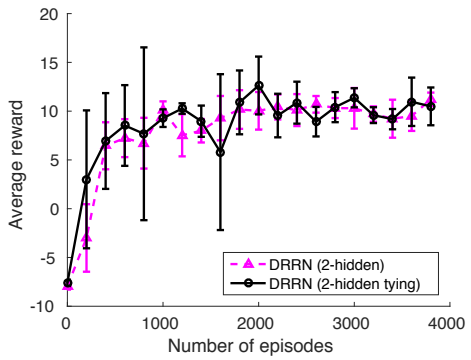

Figure 1: Learning curves of shared state-action embedding vs. proposed DRRN in Game 2

State	Actions (with Q values)
A wet strand of hair hinders my vision and I'm back in the water. Sharp pain pierces my lungs. How much longer do I have? 30 seconds? Less? I need to focus. A hand comes into view once more.	I still don't know what to do. (-8.981) Reach for it. (18.005)
"Me:" Hello Sent: today "Cherie:" Hey. Can I call you? Sent: today	Reply "I'll call you" (14.569) No (-9.498)
"You don't hold any power over me. Not anymore." Lucretia raises one eyebrow. The bar is quiet. "I really wish I did my hair today." She twirls a strand. "I'm sorry," "Save it." //Yellow Submarine plays softly in the background.// "I really hate her." "Cherie? It's not her fault." "You'll be sorry," "Please stop screaming."	I laugh and she throws a glass of water in my face. (16.214) I look away and she sips her glass quietly. (-7.986)
My dad left before I could remember. My mom worked all the time but she had to take care of her father, my grandpa. The routine was that she had an hour between her morning shift and afternoon shift, where she'd make food for me to bring to pops. He lived three blocks away, in a house with red steps leading up to the metal front door. Inside, the stained yellow wallpaper and rotten oranges reeked of mold. I'd walk by myself to my grandfather's and back. It was lonely sometimes, being a kid and all, but it was nothing I couldn't deal with. It's not like he abused me, I mean it hurt but why wouldn't I fight back? I met Adam on one of these walks. He made me feel stronger, like I can face anything.	Repress this memory (-8.102) Why didn't I fight back? (10.601) Face Cherie (14.583)

Table 4: Q values (in parentheses) for state-action pair from "Saving John", using trained DRRN. High Q-value actions are more cooperative actions thus more likely leading to better endings

State	Actions (with Q values)
Peak hour ended an hour or so ago, alleviating the feeling of being a tinned sardine that?s commonly associated with shopping malls, though there are still quite a few people busily bumbling about. To your left is a fast food restaurant. To the right is a UFO catcher, and a poster is hanging on the wall beside it. Behind you is the one of the mall's exits. In front of you stands the Machine. You're carrying 4 dollars in change.	fast food restaurant (1.094) the Machine (3.708) mall's exits (0.900) UFO catcher (2.646) poster (1.062)
You lift the warm mug to your lips and take a small sip of hot tea.	Ask what he was looking for. (3.709) Ask about the blood stains. (7.488) Drink tea. (5.526) Wait. (6.557)
As you move forward, the people surrounding you suddenly look up with terror in their faces, and flee the street.	Ignore the alarm of others and continue moving forward. (-21.464) Look up. (16.593)
Are you happy? Is this what you want to do? If you didn't avoid that sign, would you be satisfied with how your life had turned out? Sure, you're good at your job and it pays well, but is that all you want from work? If not, maybe it's time for a change.	Screw it. I'm going to find a new life right now. It's not going to be easy, but it's what I want. (23.205) Maybe one day. But I'm satisfied right now, and I have bills to pay. Keep on going. (One minute) (14.491)
You slam your entire weight against the man, making him stumble backwards and drop the chair to the ground as a group of patrons race to restrain him. You feel someone grab your arm, and look over to see that it?s Rachel. Let's get out of here, she says while motioning towards the exit. You charge out of the bar and leap back into your car, adrenaline still pumping through your veins. As you slam the door, the glove box pops open and reveals your gun.	Grab it and hide it in your jacket before Rachel can see it. (21.885) Leave it. (1.915)

Table 5: Q values (in parentheses) for state-action pair from "Machine of Death", using trained DRRN

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

Table 6: Q values (in parentheses) for state-action pair from "Machine of Death", using trained DRRN, with made-up actions that were not in the feasible set