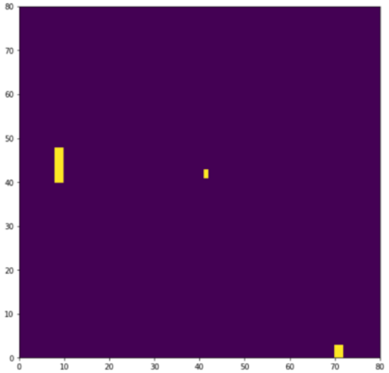
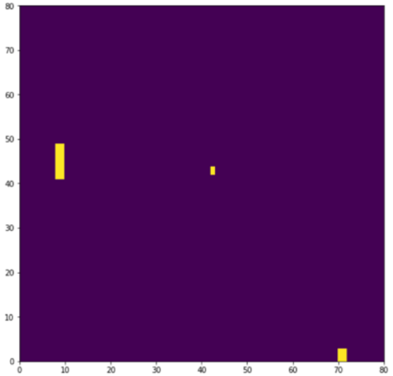
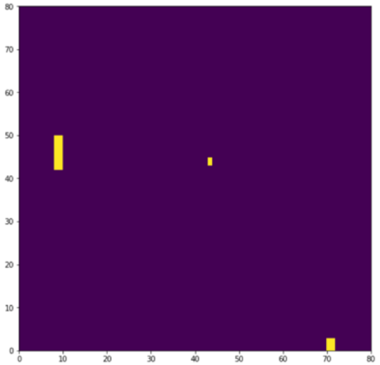
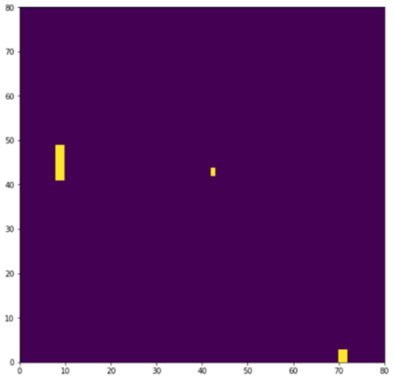
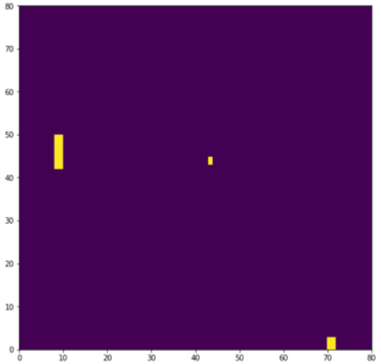
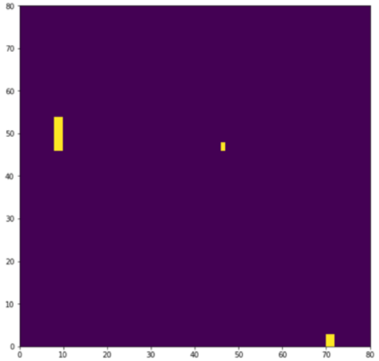
TEXAS
The University of Texas at Austin

TOPIC 4 REINFORCEMENT LEARNING

Deep Q-Learning

State	Frame 7	Frame 8	Frame 9	Button	Want to make small:
S_9				2	$(v_2(s_9) - r_9 - \delta \max\{v_0(s_{10}), v_2(s_{10}), v_3(s_{10})\})^2$

	Frame 8	Frame 9	Frame 10	Button	Want to make small
S_10				3	$(v_3(s_{10}) - r_{10} - \delta \max\{v_0(s_{11}), v_2(s_{11}), v_3(s_{11})\})^2$

Deep Q-Learning

- NN's work like regression
  - $\min \sum_t \left(predicted\ v(s_t) - true\ v(s_t)\right)^2$
- $predicted\ v(s_t)$ is like $\hat{y}$ in OLS
  - In training you just tell TF the set of s_t's
  - TF then tries to wiggle weights and biases to make predicted close to truth

Deep Q-Learning

- TF wants to minimize
  - $\left(v_0(s_t) - truth_{0,t}\right)^2 + \left(v_2(s_t) - truth_{2,t}\right)^2 + \left(v_3(s_t) - truth_{3,t}\right)^2$
- If after state t we push button 2 then we only want to minimize
  - $\left(v_2(s_t) - truth_{2,t}\right)^2$
- $truth_{3,t}$ represents future rewards if we were to have pushed button 3 at time t, but we didn't – we pushed button 2!
- Let's trick TF and minimize
  - $0 * \left(v_0(s_t) - truth_{0,t}\right)^2 + 1 * \left(v_2(s_t) - truth_{2,t}\right)^2 + 0 * \left(v_3(s_t) - truth_{3,t}\right)^2$
- To do this we have to give TF the 0,1,0
- And it doesn't matter what we tell TF for $truth_{0,t}$ or $truth_{3,t}$

Deep Q-Learning

- The problem is we don't know the true value of $v(s_t)$!
- Fake it till you make it!
  - Pretend like $r_t + \max_x \delta v(S_{t+1}, t+1)$ is the truth
- To do this, for each t, take the current weights and biases of the NN and plug s_{t+1} into the NN and see what you get out
- Tell this to TF as the 'true' y variable when it's time to update your weights and biases