How does the Actor-Critic-LSTM work?

Policy-based RL has a different workflow compared to Deep-Q Learning. In a OpenAI game episode, image frames from OpenAI are stacked and fed into the Policy’s Convolutional Block. The output of the Convolutional Block is vectorized and fed into the LSTM Cell as x. The output hx of the LSTM Cell is fed simultaneously into 2 fully-connected Linear Layers – the Action and the Critic heads. The output of the Action head is fed into a softmax classifier (modified by temperature to control explore/exploit) to generate the probability distribution that is then used to generate the Action. The output of the Critic head is a scalar which is the estimate of the value function of the state V(s). This will later be used to calculate the Advantage Function = Q(s,a) – V(s).

This forward pass described above is repeated thousands of game steps until the episode is done. When the episode is done, a backprop is performed once utilized all the accumulated gradients to generate the gradient update to the policy parameters, which are the weights of the Convolutional Block and the LSTM cell.

