Investigating Vulnerabilities of Deep Neural Policies Supplementary Material

Ezgi Korkmaz[1]

[1]KTH Royal Institute of Technology , Stockholm, Sweden.

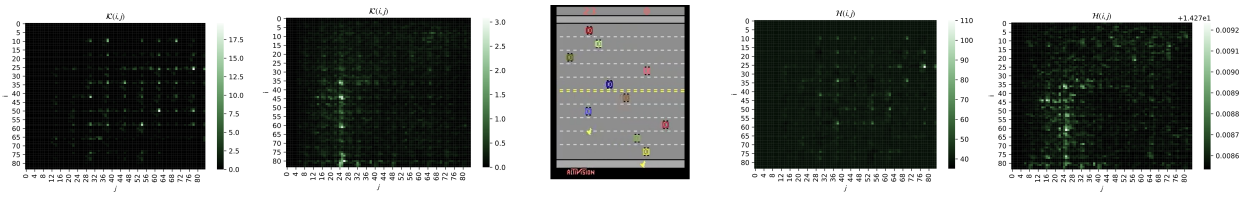
1 COMPREHENSIVE RESULTS ON FEATURE VULNERABILITY MAPPING

In this section we provide complementary results discussed in Section 6 in the main body of the paper on KMAP and HMAP feature vulnerability mapping. In more detail, Figure 1, Figure 2 and Figure 3 shows KMAP $\mathcal{K}(i,j)$ heatmaps on the left and HMAP $\mathcal{H}(i,j)$ heatmaps on the right for adversarially trained deep neural policies and vanilla trained deep neural policies with representative state in the center. The compact form of the KMAP $\mathcal{K}(i,j)$ heatmaps and HMAP $\mathcal{H}(i,j)$ heatmaps with representative states can help to visualize the sensitivity pattern analysis provided in Section 6 of the main body of the paper. For instance, the sensitivity pattern in BankHeist does not change with adversarial training. The high sensitivity region in BankHeist corresponds to the fuel gauge. The player loses when the fuel runs out. The high sensitivity region in BankHeist and corresponding representation can be seen in Figure 2.

Figure 1: KMAP $\mathcal{K}(i,j)$ and HMAP $\mathcal{H}(i,j)$ heatmaps for state-of-the-art adversarially trained deep neural policy and vanilla trained deep neural policy in RoadRunner. Left: KMAP $\mathcal{K}(i,j)$ heatmaps. Center: Representative State. Right: HMAP $\mathcal{H}(i,j)$ heatmaps.

Figure 2: KMAP $\mathcal{K}(i,j)$ and HMAP $\mathcal{H}(i,j)$ heatmaps for state-of-the-art adversarially trained deep neural policy and vanilla trained deep neural policy in BankHeist. Left: KMAP $\mathcal{K}(i,j)$ heatmaps. Center: Representative State. Right: HMAP $\mathcal{H}(i,j)$ heatmaps.

In Freeway we found that the state-of-the-art adversarial training removes the non-robust features relevant to the optimal policy. In particular, KMAP $\mathcal{K}(i,j)$ and HMAP $\mathcal{H}(i,j)$ heatmaps for vanilla trained deep neural policy in Figure 3 show a clear path where the agent crosses the street as a high sensitivity region. However, the sensitivity region for the adversarially trained deep neural policy is in a clear grid pattern completely decoupled from the region where the optimal policy is executed.



Adversarially Trained | Vanilla Trained | Representative State | Adversarially Trained | Vanilla Trained

Figure 3: KMAP $\mathcal{K}(i,j)$ and HMAP $\mathcal{H}(i,j)$ heatmaps for state-of-the-art adversarially trained deep neural policy and vanilla trained deep neural policy in Freeway. Left: KMAP $\mathcal{K}(i,j)$ heatmaps. Center: Representative State. Right: HMAP $\mathcal{H}(i,j)$ heatmaps.