Creative part: approach
CSAI: Contemporary AI Topics

December 5, 2019

Robbe Louage & Elias Storme

We want to demonstrate the performance of both approaches to safety. This requires a level playing field, so results can be compared. This leveling of the playing field is a recurring problem in AI research. To solve this problem, shared datasets have emerged and been adopted by researchers. One well known example of this is ImageNet, a database of images to do image procession on.

1 DeepMind Gridworlds

For safety in reinforcement learning, one such "dataset" being proposed is Gridworlds by Google DeepMind [?]. There are a total of 8 gridworlds. Each gridworld represents a different environments that can be used to test one aspect of AI safety. The environments include the following safety checks for a learning agent:

- Safe interruptibility
- Avoiding side effects
- Absent supervisor
- Reward gaming
- Self-modification
- Distributional shift
- Robustness to adversaries
- Safe exploration

References