1611.01779 - Learning to Act by Predicting the Future

`RL` `DL` `AI`

- **Author : Yunqiu Xu**

---

1. Introduction

- https://zhuanlan.zhihu.com/p/23454387
    - 给定当前图像，当前的各游戏数据（血量，子弹数和分数）及提高这些数据的迫切程度的权值（Goal
    - 对每个动作输出一个提高值f（比如说做这个动作之后，血量提高了多少，或者又杀死了几个敌人)
    - 然后用最高的提高值来选下一步动作
    - 这个实际上是Q值网络的变种
    - 他们生成了各种类型的地图做了训练，效果比DQN及A3C都要好些
    - 因为迫切程度的权值是一个输入，所以这个模型具有在线改变目标的能力，比如说可以先让它去加血，加完了再去杀敌
- Highlights:
    - High-dim sensory stream + low-dim measurement stream $\rightarrow$ train a sensorimotor control model by interacting with the environment
    - Supervised learning without extraneous supervision
    - Learn without a fixed goal at training time
    - Pursue dynamically changing goals at test time
    - The Track 2 (full deathmatch with unknown maps) champion of ViZDoom AI Competition 2016
- Challenges of RL:
    - Sensorimotor control from raw sensory input in complex and dynamic three-dimensional environments, learned directly from experience

- The acquisition of general skills that can be flexibly deployed to accomplish a multitude of dynamically specified goals
- The paper:
    - Propose an approach to sensorimotor control $\rightarrow$ assist progress towards overcoming these challenges
    - Use monilithic state and a scalar reward to replace reward-based formalisation
    - High/multi-dim sensory stream: more appropriate for an organism that is learning to function in an immersive environment
    - Low-dim measurement stream: provides rich and temporally dense supervision $\rightarrow$ stabilize and accelerate training.
    - Given present sensory input, measurements, and goal, the agent can be trained to predict the effect of different actions on future measurements
    - Reduces sensorimotor control to supervised learning,
      $\rightarrow$ can learn from raw experience and without extraneous data

2. Model

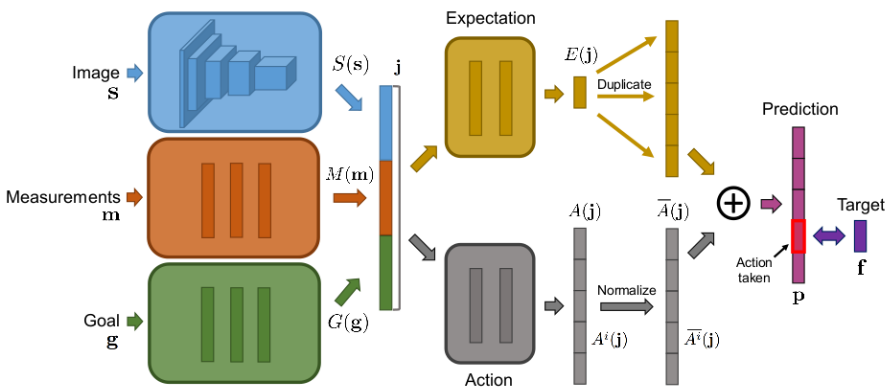

Figure 1: Network structure. The image $\mathbf{s}$, measurements $\mathbf{m}$, and goal $\mathbf{g}$ are first processed separately by three input modules. The outputs of these modules are concatenated into a joint representation $\mathbf{j}$. This joint representation is processed by two parallel streams that predict the expected measurements $E(\mathbf{j})$ and the normalized action-conditional differences $\{\overline{A^i}(\mathbf{j})\}$, which are then combined to produce the final prediction for each action.

3. Result

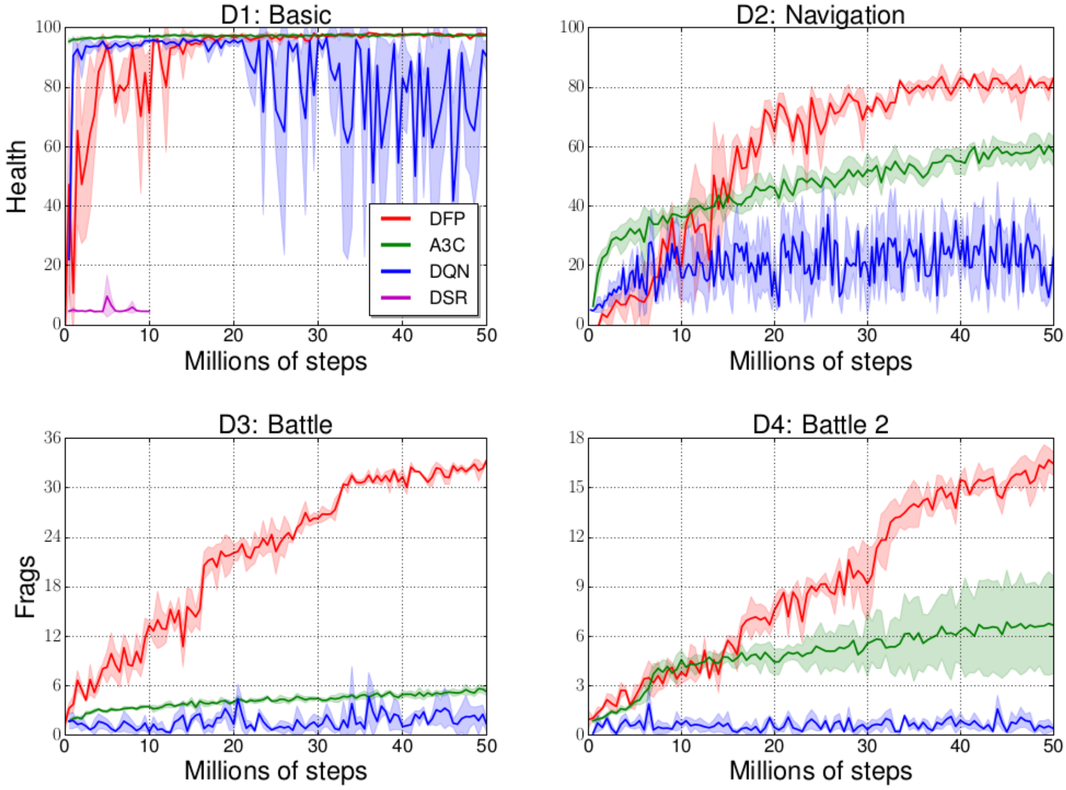

Figure 3: Performance of different approaches during training. DQN, A3C, and DFP achieve similar performance in the Basic scenario. DFP outperforms the prior approaches in the other three scenarios, with a multiplicative gap in performance in the most complex ones (D3 and D4).

		Train				
		D3	D4	D3-tx	D4-tx	D4-tx-L
Test	D3	33.6	17.8	29.8	20.9	22.0
	D4	1.6	17.1	5.4	10.8	12.4
	D3-tx	3.9	8.1	22.6	15.6	19.4
	D4-tx	1.7	5.1	6.2	10.2	12.7

Table 2: Generalization across environments.