UNIVERSITÀ DEGLI STUDI FIRENZE

HR EXCELLENCE IN RESEARCH

Kevin Wang, Ishaan Javali, Michał Bortkiewicz, Tomasz Trzcinski

1000 Layer Networks for Self-Supervised RL

Scaling Depth Can Enable New Goal Reaching Capabilities

Nathan Acciai

Presentazione Paper Deeo Learning Applications

Introduction

- The goal of classic Reinforcement Learning is for the agent to learn how to maximize cumulative reward

Reward

$$\max_{\pi} \mathbb{E} \left[\sum_{t} \gamma^{t} r_{t} \right]$$

Discount Factor

- The problem resides in the learning signal:
  - Sparse Reward
  - Strong addiction to exploration
  - Noisy value estimation
- This leads to making learning unstable by increasing the capacity of the model
- Below we examine the three importan points of the proposed method

Goal-Condition RL

- The problem is riformulated as goal achievement
- Goal-conditioned MDP is defined as: $M_g = (S, A, p_0, p, p_g, r_g, \gamma)$
- The policy is conditioned by the goal as: $\pi(a|s, g)$
- The goal $g \in G$ are linked together with a mapping function: $f : S \to G$
- The reward now measures the probability of reaching the goal

$$r_g(s_t, a_t) \triangleq (1 - \gamma) p(s_{t+1} = g | s_t, a_t)$$

- Then the Q-function has a probabilistic interpretation:

$$\text{To maximize} \longleftarrow \quad Q_g^{\pi}(s, a) \triangleq p_{\gamma}^{\pi}(g | s, a)$$

Contrastive Reinforcement Learning

- The Contrastive RL is a Actor-Critic goal-conditioned method.
- Actor learns the goal-conditioned-policy $\pi_\theta(a|s,g) = \arg\min_a \|\phi(s,a) - \psi(g)\|$
- Critic measures a embedding distance:

$$f_{\psi,\phi}(s,a,g) \triangleq \|\underbrace{\phi(s,a)}_{\text{State-action embedding}} - \underbrace{\psi(g)}_{\text{goal embedding}}\|_2$$

- Critic is trained by infoNCE (Information Noise-Contrastive Estimation):

$$\mathcal{L} = -\mathbb{E}_{\mathcal{B}}\left[\log \frac{e^{-f(s_i,a_i,\overbrace{g_i}^{\text{Correct goal}})}}{\sum_j^K e^{-f(s_i,a_i,\underbrace{g_j}_{\text{Wrong goal}})}}\right]$$

- Policy maximise:

$$\max_\theta \mathbb{E}[-f_{\psi,\phi}(s,a,g)]$$

Residual Connection

- Residual connections allow the network to learn changes to the representation rather than completely new transformations.

$$\text{residual block}$$
$$h_{i+1} = h_i + F_i(h_i)$$

Input block

Learned transformation

- Main advantages:
    - Preserves useful information from previous layers
    - Improves gradient propagation
    - Makes it possible to train very deep networks

Combined with goal-conditioned and contrastive RL, it allows you to scale up to hundreds or thousands of layers without destabilizing your training.

Experiments and Limitations

- Experiments:
  - Deep networks (up to 1000 layers) on goal-conditioned tasks
  - Comparison with classic RL and shallow networks
  - Use of contrastive RL and residual connections
- Results:
  - Significant improvements in performance and stability
  - Generalization to previously unseen distant goals
  - Emergence of complex behaviors and implicit planning
- Limitations:
  - Requires large amounts of data and batches for contrastive learning
  - Learned distance is a quasi-metric: not perfect.
  - Very deep implementations can be compute-intensive.

Thank your for Attention!

Nathan Acciai

Deep Learning Applications