



UNIVERSITÀ
DEGLI STUDI
FIRENZE



HR EXCELLENCE IN RESEARCH

Kevin Wang, Ishaan Javali, Michał Bortkiewicz, Tomasz Trzcinski

1000 Layer Networks for Self-Supervised RL

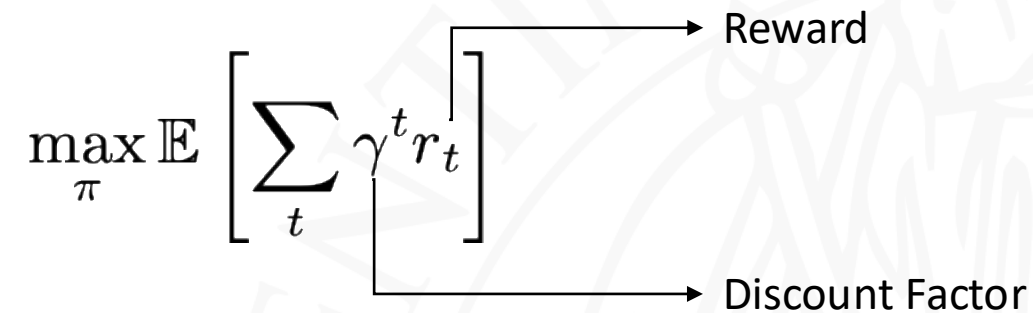
Scaling Depth Can Enable New Goal Reaching Capabilities

Nathan Acciai

Presentazione Paper Deep Learning Applications

Introduction

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

$$\max_{\pi} \mathbb{E} \left[\sum_t \gamma^t r_t \right]$$


The diagram shows the equation $\max_{\pi} \mathbb{E} \left[\sum_t \gamma^t r_t \right]$ with two arrows pointing from the terms inside the brackets to labels on the right. One arrow points from r_t to the label 'Reward', and another arrow points from γ to the label '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 important points of the proposed method

Goal-Condition RL

- The problem is reformulated 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**: $f : S \rightarrow G$
- 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:

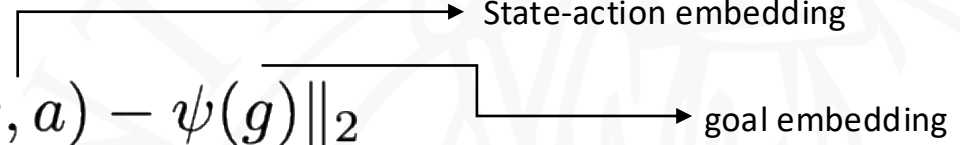
To maximize $\longleftarrow Q_g^\pi(s, a) \triangleq p_\gamma^\pi(g|s, a)$

Contrastive Reinforcement Learning

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

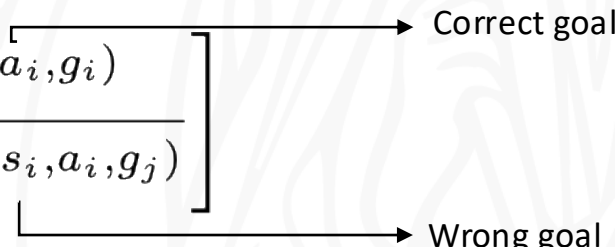
- Critic measures a **embedding distance**:

$$f_{\psi, \phi}(s, a, g) \triangleq \|\phi(s, a) - \psi(g)\|_2$$



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

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

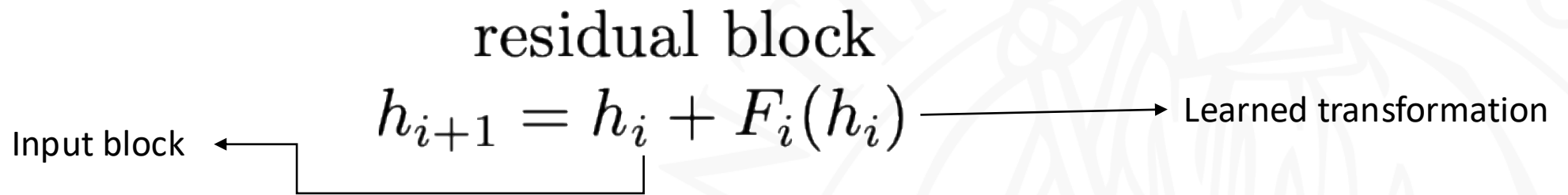


- Policy maximise**:

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

Residual Connection

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



- **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 you for Attention!

Nathan Acciai

Deep Learning Applications