Importance Weighted Transfer of Samples in Reinforcement Learning

ICML 2018

Introduction

The goal of transfer in Reinforcement Learning (RL) is to **speed-up** RL algorithms by reusing knowledge obtained from a set of previously learned tasks.

The intuition is that the experience made by learning *source tasks* might be useful for solving a related, but different , *target task*.

Preliminaries

$$M = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$$

$$\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$$

$$\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$$

$\gamma \in [0, 1)$ is the discount factor

$$R(s, a) = \mathbb{E}\left[\mathcal{R}(\cdot|s, a)\right]$$

$$r_t \sim \mathcal{R}(\cdot|s_t, a_t)$$

$$s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$$

Preliminaries

$$Q^{\pi}(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid M, \pi, s_0 = s, a_0 = a\right]$$

$$Q_{\max} := \frac{r_{\max}}{1-\gamma}$$

$$L^* : \mathcal{B}(\mathcal{S} \times \mathcal{A}, Q_{\max}) \to \mathcal{B}(\mathcal{S} \times \mathcal{A}, Q_{\max})$$

$$(L^*Q)(s,a) := R(s,a) + \gamma \int_{\mathcal{S}} \mathcal{P}(ds'|s,a) \max_{a'} Q(s',a')$$

$$\mathcal{D}_N = \{\langle s_i, a_i, s'_i, r_i \rangle\}_{i=1}^{N}$$

$$(\widehat{L}^*Q)(s_i,a_i) := r_i + \gamma \max_{a'} Q(s'_i,a')$$

$$\{M_j = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}_j, \mathcal{R}_j \rangle, j = 0, \ldots, m\}$$

Importance Weights for Transfer

$$Q(s,a) = r(s,a) + \gamma \max(Q(s',a'))$$

$$Q(s,a) = wr'(s,a) + \gamma \max(Q(s',a'))$$

Importance Weights for Transfer

$$w(X) = P(X)/Q(X)$$

$$\mathbb{P}((X,Y)|M_j) = \mathbb{P}(Y|X,M_j)\mu(X)$$

$$w(X_i^{(j)}, Y_i^{(j)}) = \frac{\mathbb{P}(Y_i^{(j)}|X_i^{(j)}, M_0)}{\mathbb{P}(Y_i^{(j)}|X_i^{(j)}, M_j)}$$

First, the distribution $P(Y|X, M_j)$ is, even in the case where the MDPs are known, very hard to characterize.

Second, consider a simple case where we have a source MDP with the same transition dynamics as the target, but with entirely different reward. Then, the importance weights defined above are likely to be very close to zero for any source sample, thus making transfer useless.

Importance Weights for Transfer

$$\widehat{R} = \arg\inf_{h \in \mathcal{H}} \frac{1}{Z_r} \sum_{j=0}^{m} \sum_{i=0}^{N_j} w_{r,i}^{(j)} \left| h(X_i^{(j)}) - r_i^{(j)} \right|^2$$

$$X_i = (s_i, a_i) \;\; \mathcal{H} \subset \mathcal{B}(\mathcal{S} \times \mathcal{A}, Q_{max})$$

$$w_{r,i}^{(j)} = \frac{\mathcal{R}_0(r_i^{(j)} | X_i^{(j)})}{\mathcal{R}_j(r_i^{(j)} | X_i^{(j)})}.$$

$$Q_{k+1} = \arg\inf_{h \in \mathcal{H}} \frac{1}{Z_p} \sum_{j=0}^{m} \sum_{i=0}^{N_j} w_{p,i}^{(j)} \left| h(X_i^{(j)}) - \widetilde{Y}_i^{(j)} \right|^2$$

$$\widetilde{Y}_i^{(j)} = \widetilde{L}^* Q_k(X_i^{(j)}) := \widehat{R}(s_i^{(j)}, a_i^{(j)}) + \gamma \max_{a'} Q_k(s'^{(j)}_i, a') \text{ and } Q_0 = \widehat{R}.$$

$$w_{p,i}^{(j)} = \frac{\mathcal{P}_0(s'^{(j)}_i | X_i^{(j)})}{\mathcal{P}_j(s'^{(j)}_i | X_i^{(j)})}$$

Importance Weights for Transfer

Algorithm 1 Importance Weighted Fitted Q-Iteration

Input: The number of iterations K, a dataset $\widetilde{\mathcal{D}}^{+} = \bigcup_{j=0}^{m} \bigcup_{i=0}^{N_j} \left\{ s_i^{(j)}, a_i^{(j)}, s'^{(j)}_i, r_i^{(j)}, \widetilde{w}_{r,i}^{(j)}, \widetilde{w}_{p,i}^{(j)} \right\}$, a hypothesis space $\mathcal{H}$

Output: Greedy policy π_K

$\widehat{R} \leftarrow \arg\inf_{h \in \mathcal{H}} \frac{1}{Z_r} \sum_{j,i} \widetilde{w}_{r,i}^{(j)} \left| h(s_i^{(j)}, a_i^{(j)}) - r_i^{(j)} \right|^2$

$Q_0 \leftarrow \widehat{R}$

for $k = 0, \ldots, K-1$ **do**

$Y_i^{(j)} \leftarrow \widetilde{L}^* Q_k(s_i^{(j)}, a_i^{(j)}), \quad \forall i, j$

$Q_{k+1} \leftarrow \arg\inf_{h \in \mathcal{H}} \frac{1}{Z_p} \sum_{j,i} \widetilde{w}_{i,p}^{(j)} \left| h(s_i^{(j)}, a_i^{(j)}) - Y_i^{(j)} \right|^2$

end for

$\pi_K(s) \leftarrow \arg\max_{a \in \mathcal{A}} \{Q_K(s,a)\}, \quad \forall s \in \mathcal{S}$

Estimation of Importance Weights

Gaussian Processes:

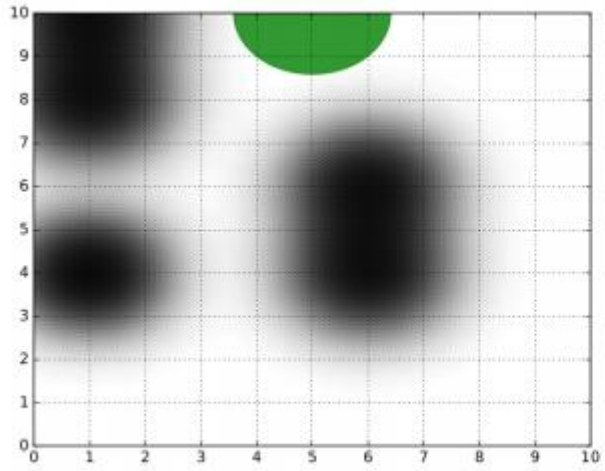
$$\bar{r}(s,a) \sim \mathcal{N}(\mu_{GP_j}(s,a), \sigma^2_{GP_j}(s,a))$$

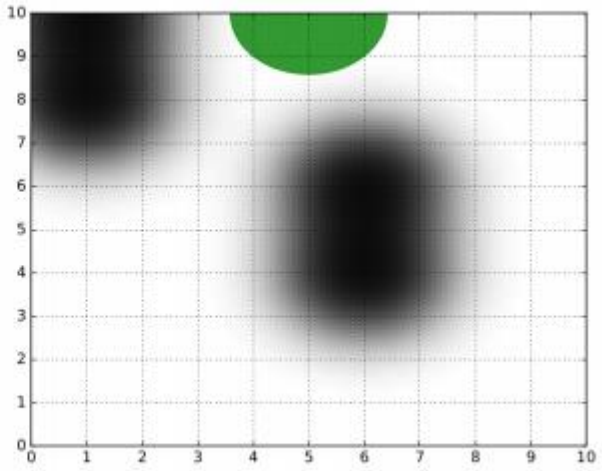
$$\mathcal{R}_j(\cdot|s,a) \quad = \quad \mathcal{N}\left(\mu_r^{(j)}(s,a), \sigma_j^2(s,a)\right)$$

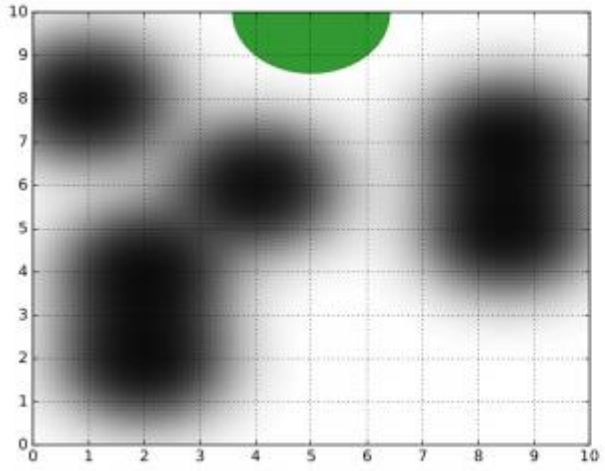
$$\mathbb{E}_{\mathcal{G}}[w] = C \, \frac{\mathcal{N}\left(r \middle| \mu_{GP_0}(s,a), \sigma_0^2(s,a) + \sigma^2_{GP_0}(s,a)\right)}{\mathcal{N}\left(r \middle| \mu_{GP_j}(s,a), \sigma_j^2(s,a) - \sigma^2_{GP_j}(s,a)\right)}$$

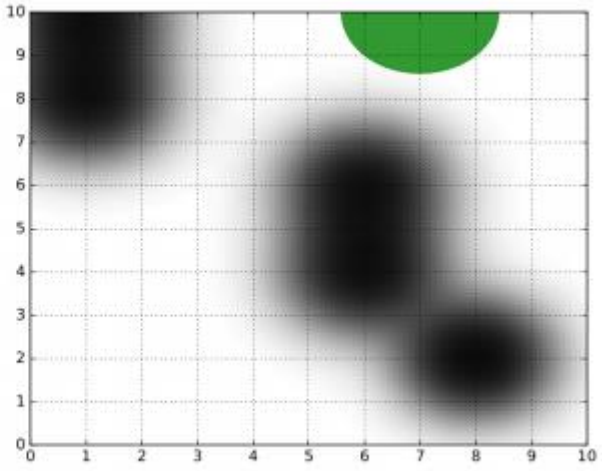
$$\mathbb{E}_{\mathcal{G}}[w] = \prod_{d=1}^{D} C_d \, \frac{\mathcal{N}\left(s'_d \middle| \mu_{GP_{0,d}}(s,a), \delta^2_{0,d}(s,a) + \sigma^2_{GP_{0,d}}(s,a)\right)}{\mathcal{N}\left(s'_d \middle| \mu_{GP_{j,d}}(s,a), \delta^2_{j,d}(s,a) - \sigma^2_{GP_{j,d}}(s,a)\right)}$$

Experiments



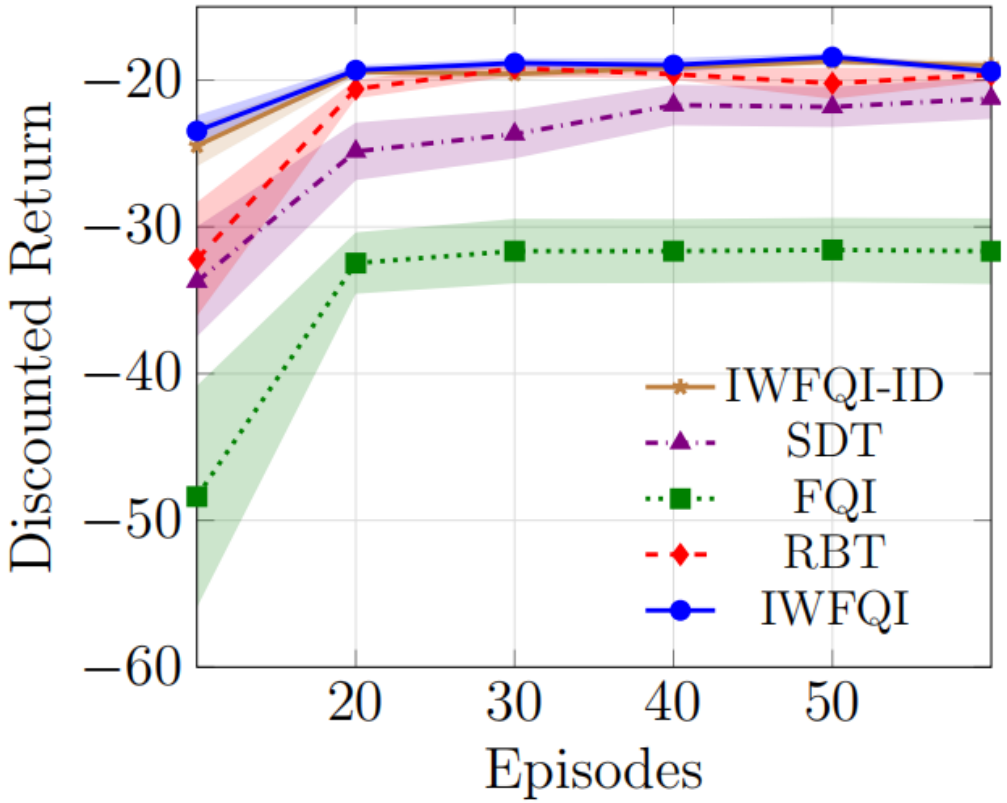


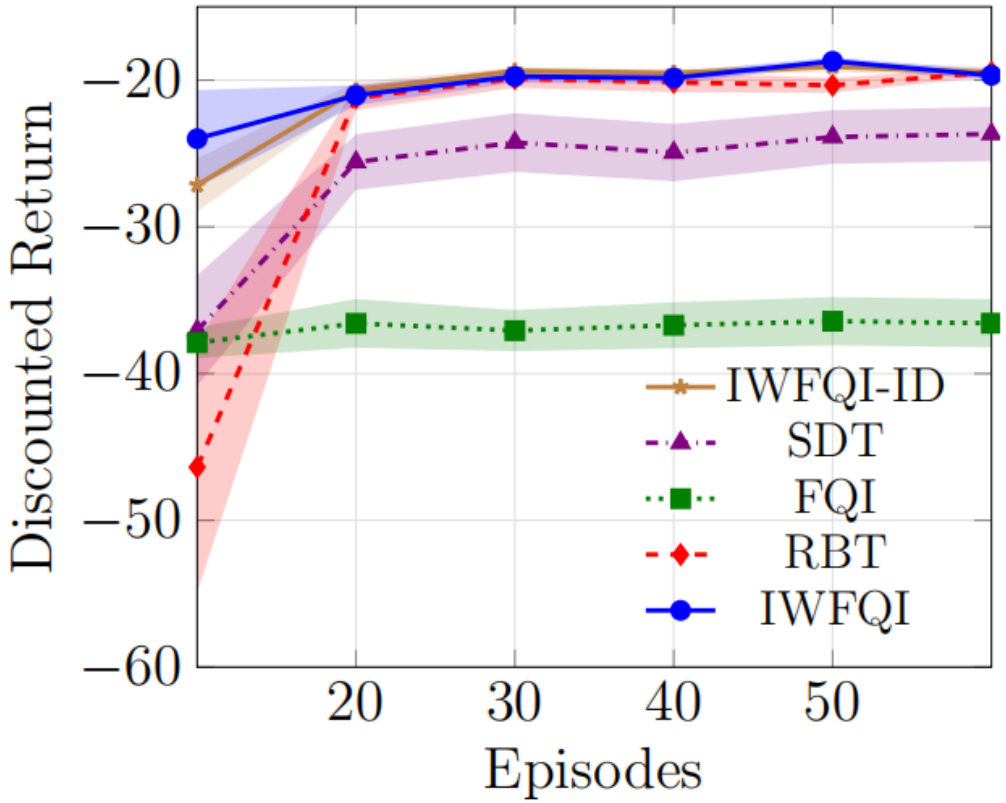




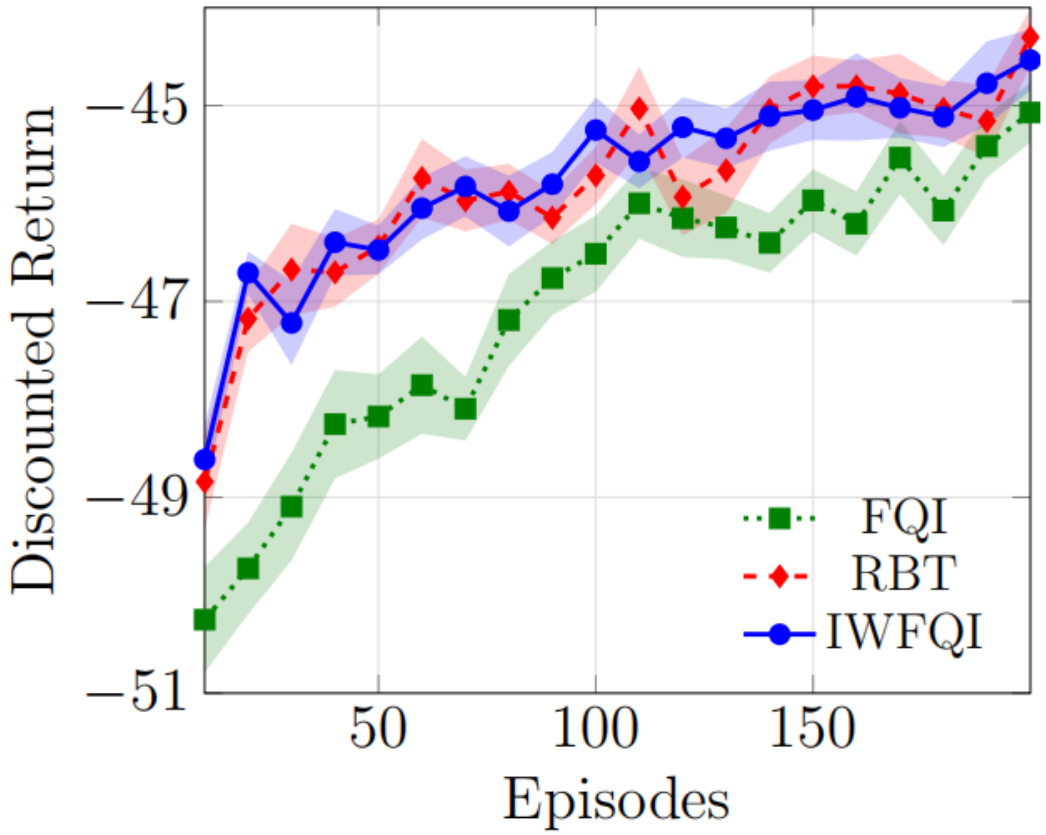
$$R(s,a) = -1 - 100 \sum_{u \in \mathcal{U}} W_u(s).$$

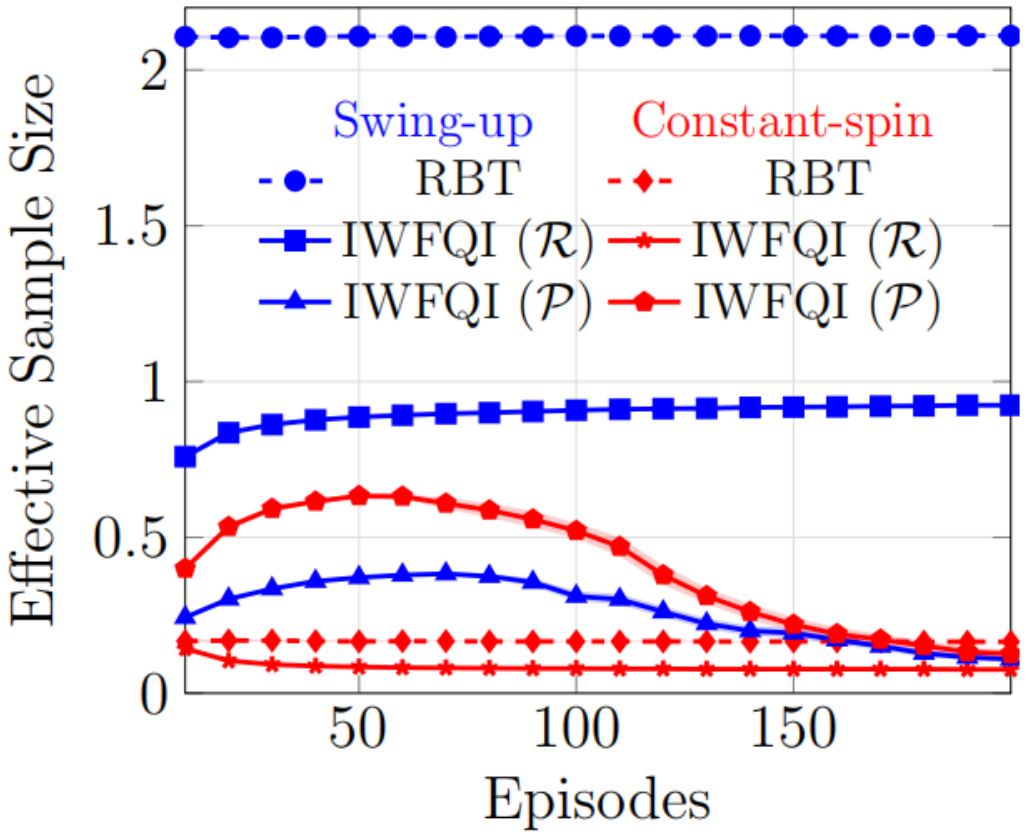
Experiments



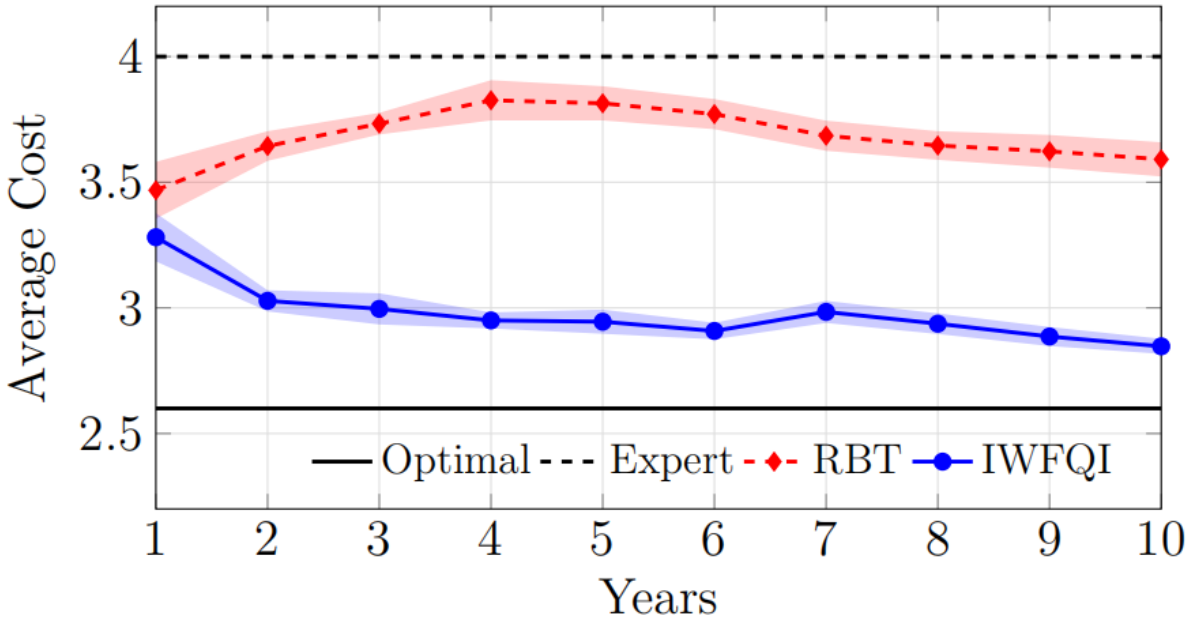


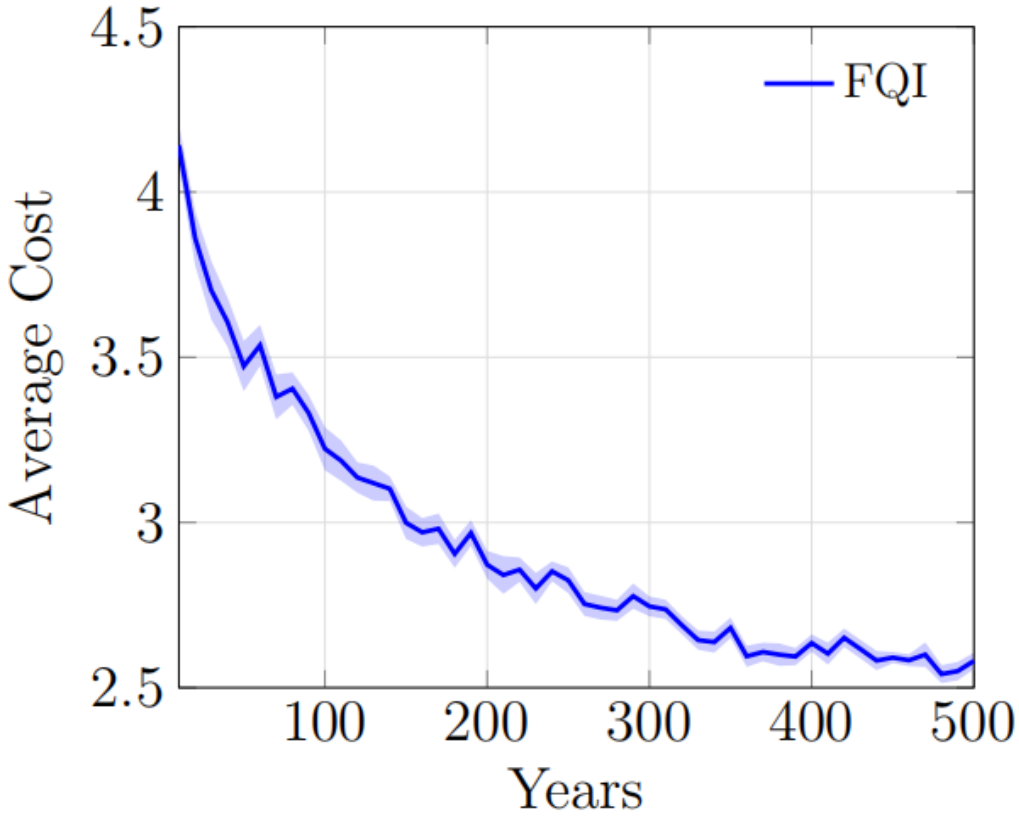
Experiments





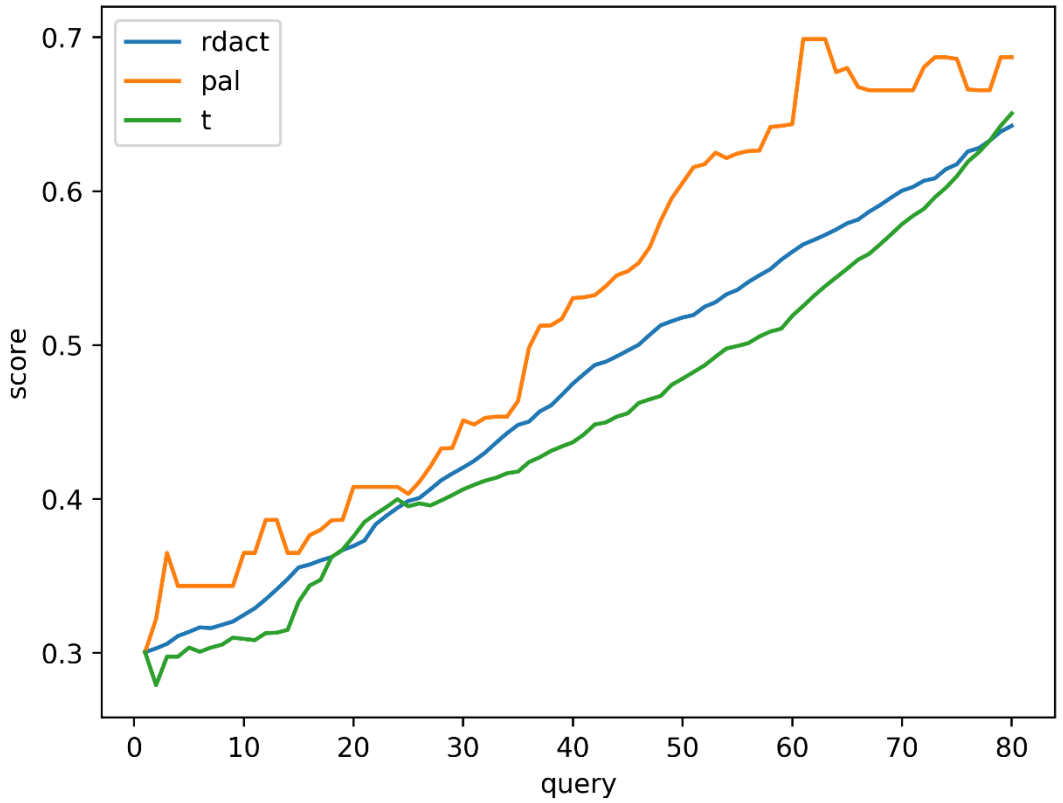
Experiments



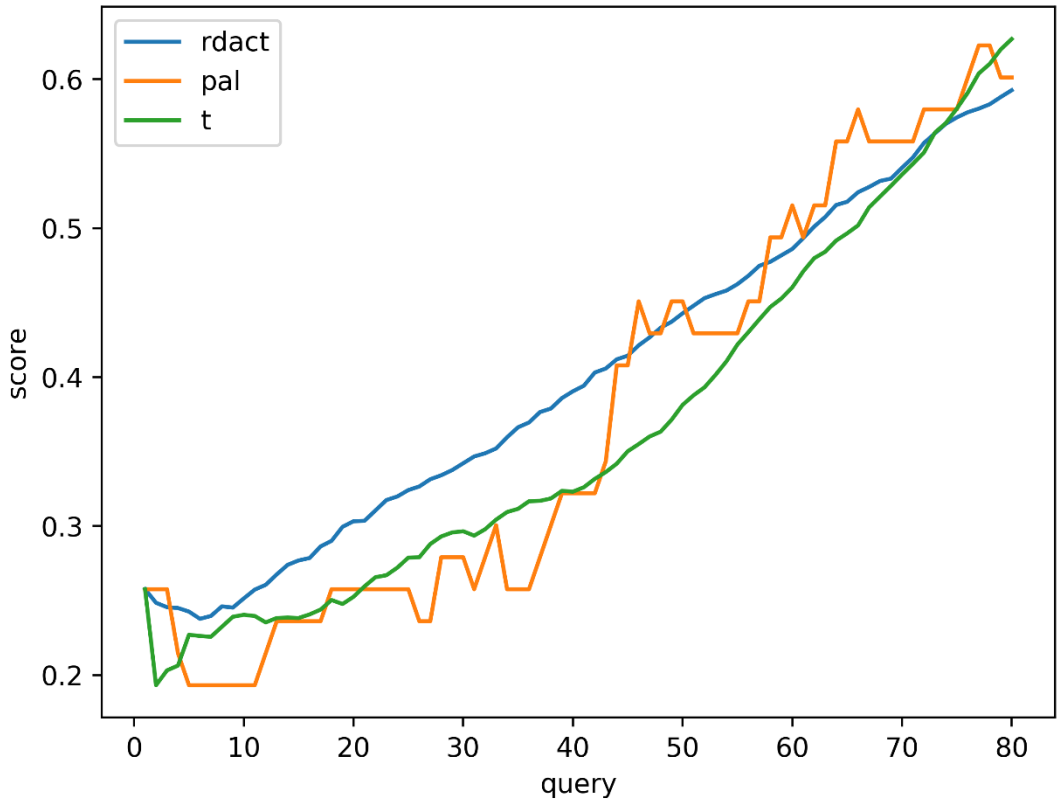


Amazon-Dslr

Multi-source

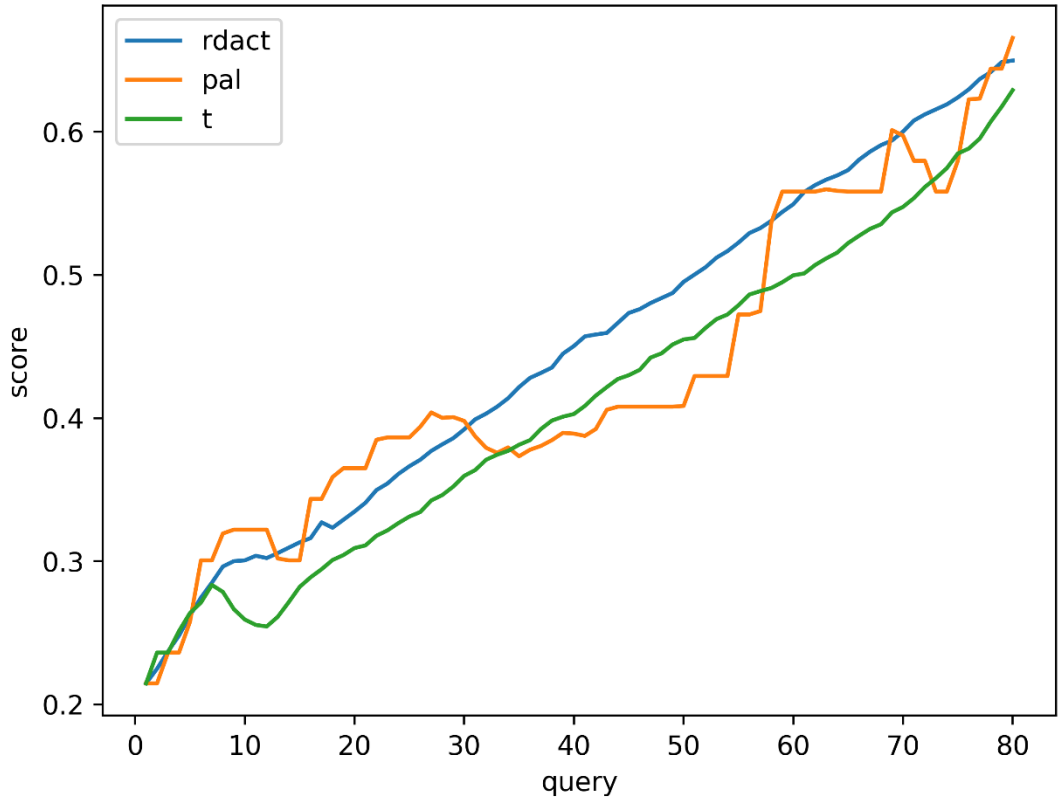


one-source

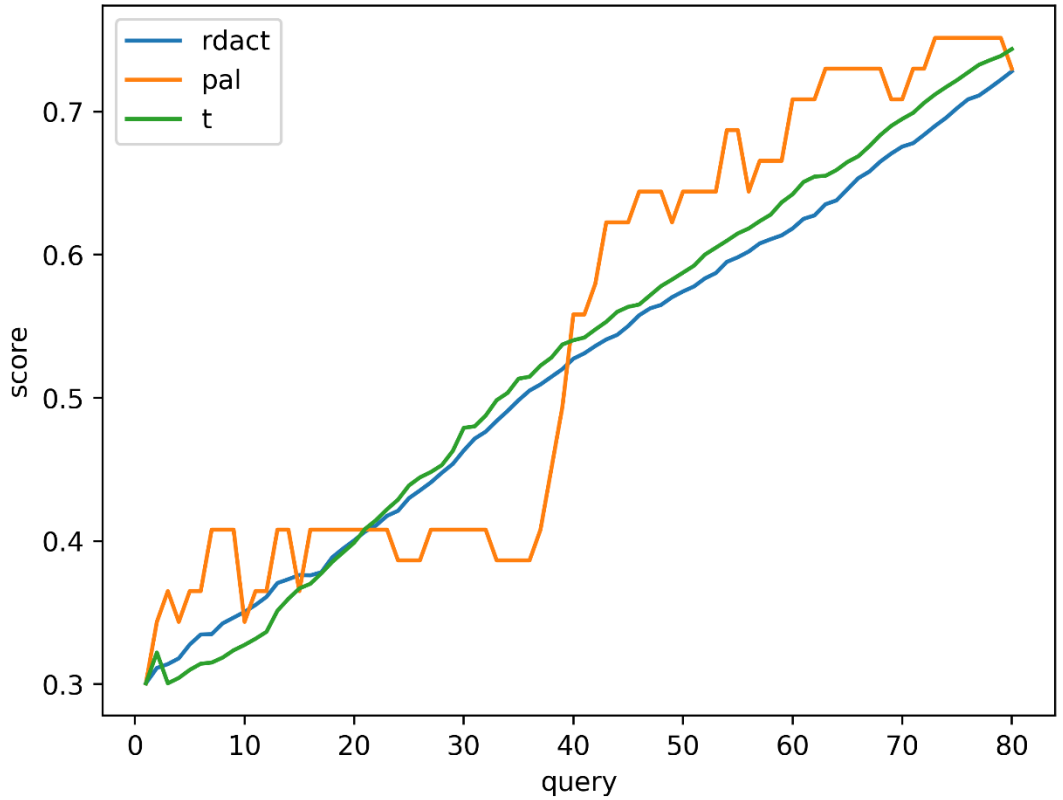


Caltech10-Dslr

Multi-source

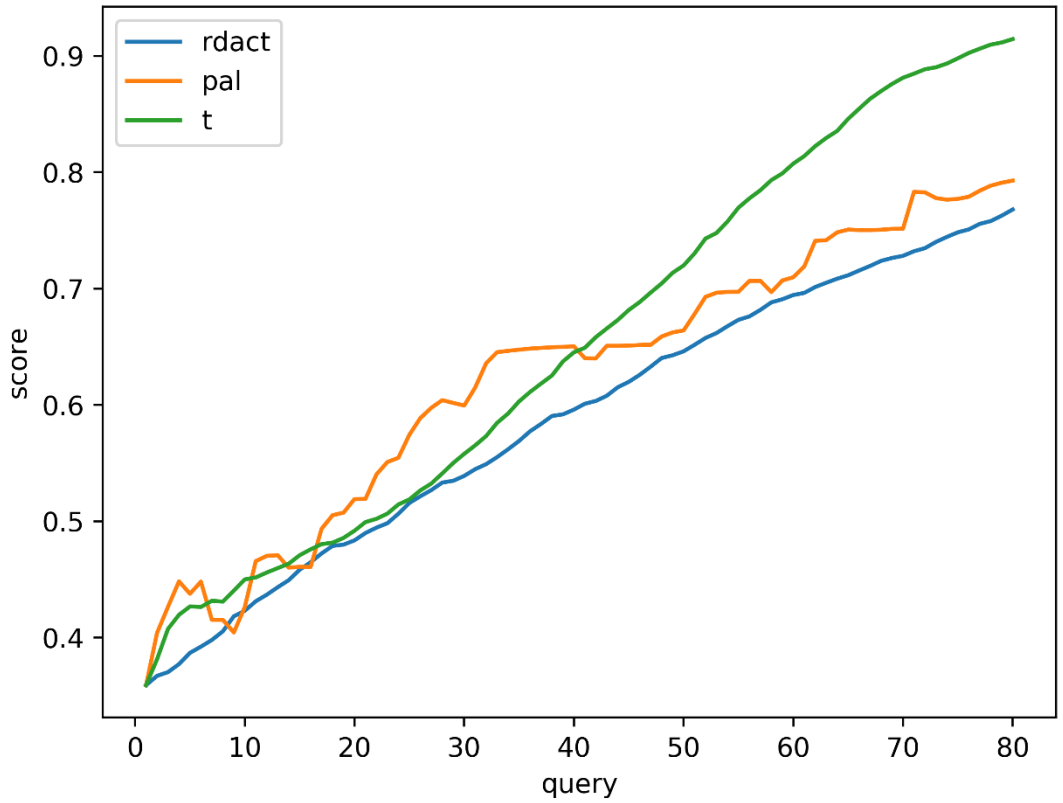


one-source

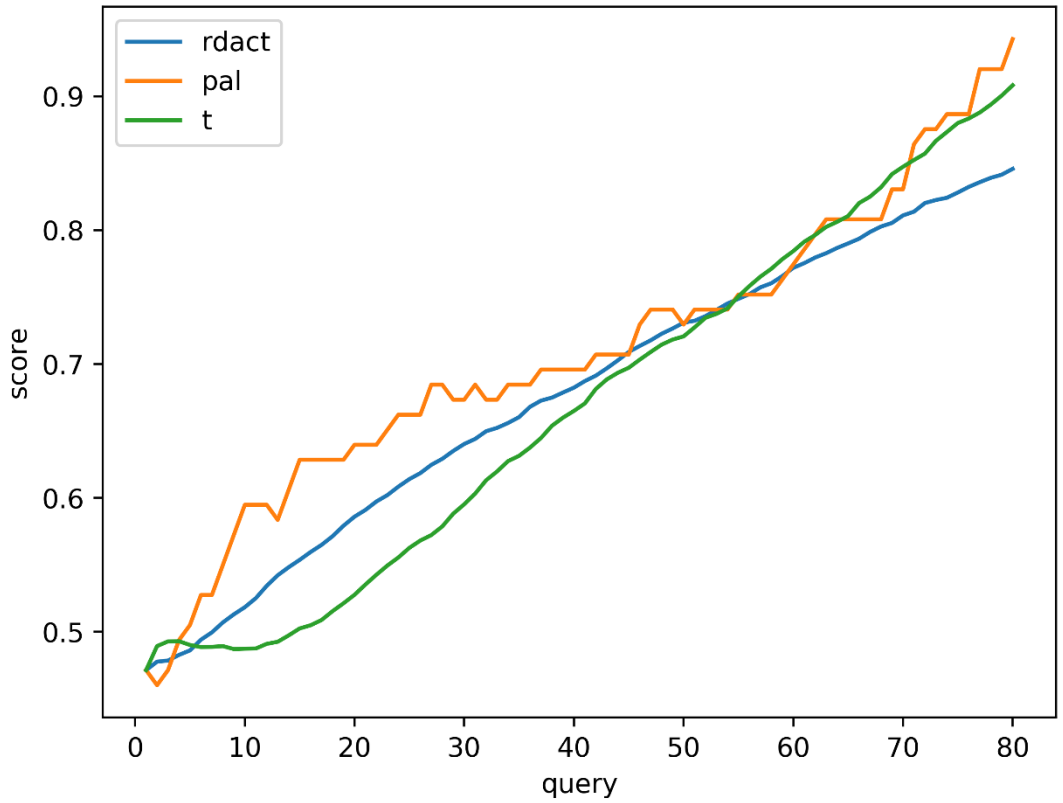


Amazon-Webcam

Multi-source

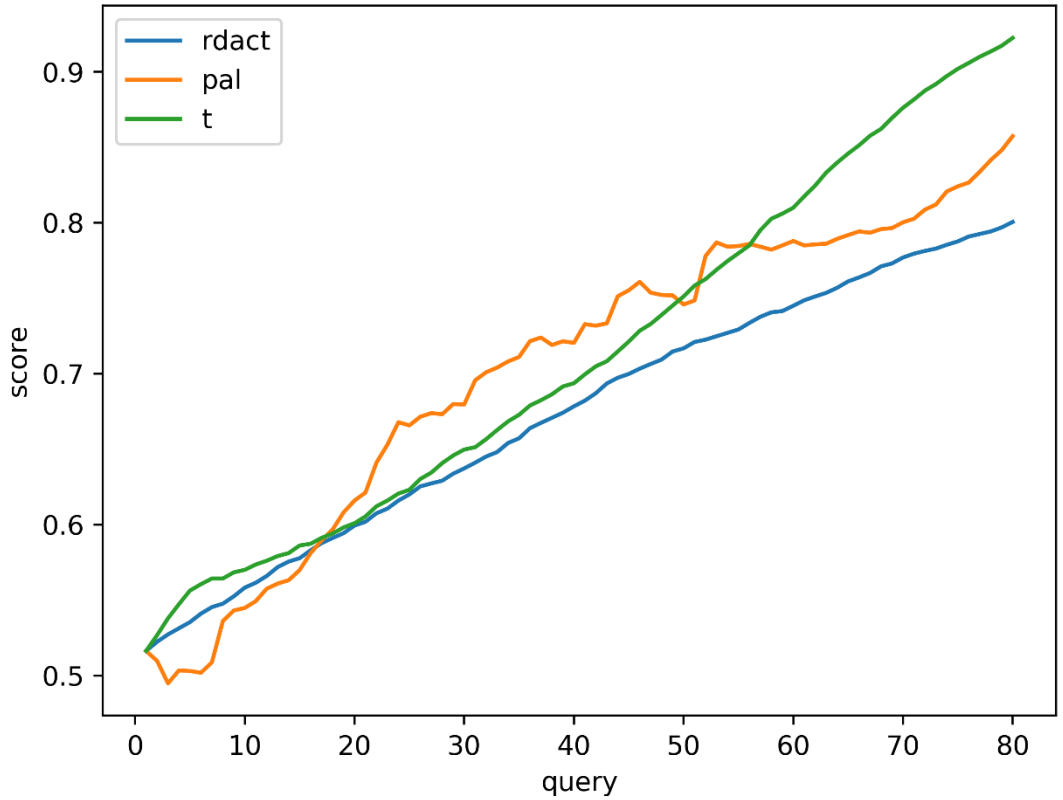


one-source



Caltech10-Webcam

Multi-source



one-source