💡

Self-Attentional Credit Assignment for Transfer in Reinforcement Learning

论文试图解决什么问题？

ca

这是否是一个新的问题？

todo

这篇文章要验证一个什么科学假设？

todo

有哪些相关研究？如何归类？谁是这一课题在领域内值得关注的研究员？

todo

论文中提到的解决方案之关键是什么？

核心：SECRET weighs the contribution of observation-action pairs to future reward

- 奖励的预测：

1. We create a sequence-to- sequence (seq2seq) model (Sutskever et al., 2014) that takes as input the sequence of observation-action pairs and has to reconstruct the corresponding sequence of environment rewards
2. the reward prediction model does not share representations with the agent(模型单独训练，使用经验回放池数据进行线下训练)
3. We equip our seq2seq model with an attention mechanism.
4. the seq2seq model looks into the past to find predictive signal in order to reconstruct the reward. so observation-action pairs it attends to should be those which reduce its uncertainty about the future, in other words those that explain future reward and should be credited (未来的状态-动作对会attend到过去的状态-动作对，从而发现过去哪些状态-动作对的相关效应强弱)

- 利用reward shaping进行信用分配：

什么是reward shaping？

For a given MDP M = (S, A, γ, R, P), we define a new MDP M′ = (S,A,γ,R′,P) where R′ = R + F is the shaped reward and F the shaping.

Since Secret weighs the contribution of observation-action pairs to future reward, we use it to derive a shaped reward that corresponds to the sum of future reward reachable from the underlying state, weighted by the attention calculated by the model。（使用学到的注意力权重构造一个reward shaping的潜在函数F）

定义：

$$R^{\leftarrow}_{\tau}(s,a) = \sum_{t=1}^{T} \mathbb{I}\{s_t = s, a_t = a\} \sum_{i=t}^{T} \alpha_{t\leftarrow i} r(s_i, a_i),$$

潜在函数为：

$$\hat{\phi}(s) = \frac{1}{|D|}\sum_{\tau\in D}\sum_{t=1}^{T}\mathbb{I}\{s_t^{(\tau)} = s\}R_\tau^{\leftarrow}(s_{t-1}^{(\tau)}, a_{t-1}^{(\tau)}).$$

- 模型架构：

We use a Transformer decoder with a single self-attention layer (Lin et al., 2017) and a single attention head.

The model input is a sequence of observation-action couples (ot, at)t=0,...,T .

Each observation goes through a series of convolutional layers (for visual inputs) followed by a series of feed-forward layers.

attention weights themselves can be viewed as a form of credit assignment, and will be used as such in what follows.

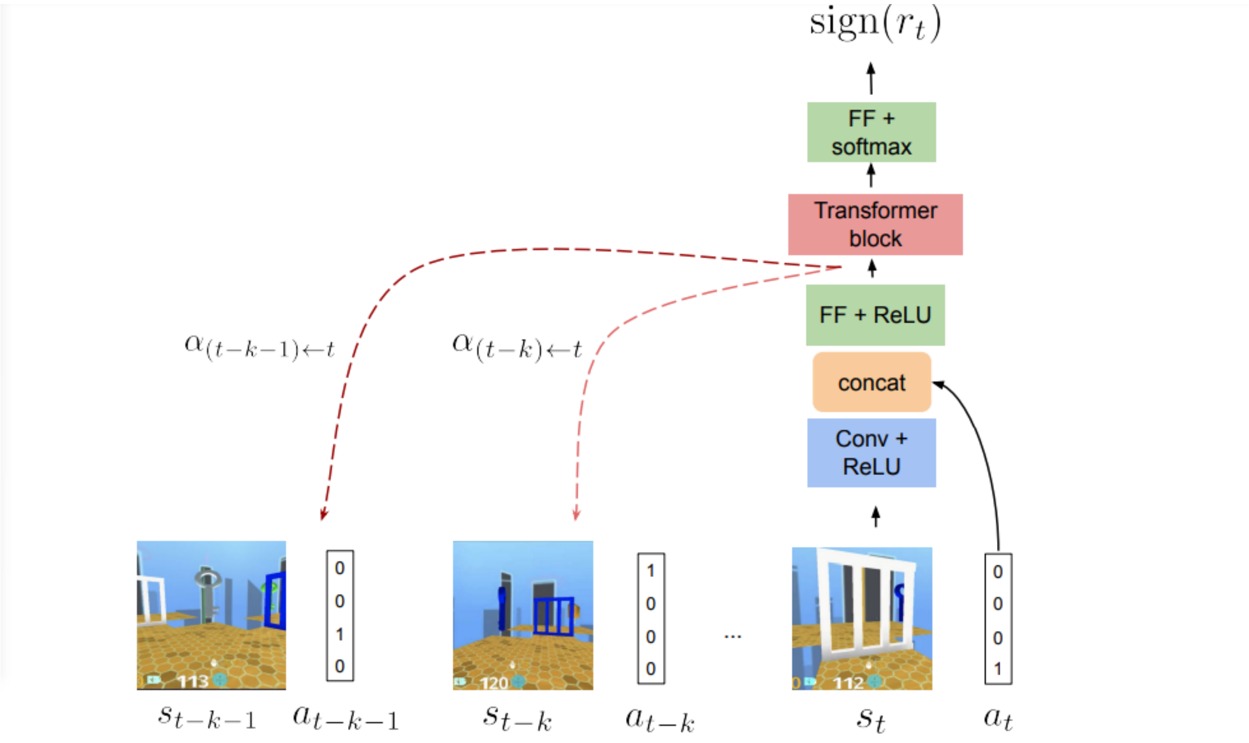

Figure 7: The architecture used for SECRET. $\alpha_{\cdot\leftarrow t}$ is the vector containing the attention weights of the model for its prediction at step t.

论文中的实验是如何设计的？

todo

用于定量评估的数据集是什么？代码有没有开源？

todo

论文中的实验及结果有没有很好地支持需要验证的科学假设？

todo

这篇论文到底有什么贡献？

todo

下一步呢？有什么工作可以继续深入？

todo