DSC190: Machine Learning with Few Labels

Reinforcement Learning

Zhiting Hu

Lecture 23, November 22, 2024

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Outline

Reinforcement learning

Presentations

- **Brandon Chiou:** Scaling Rectified Flow Transformers for High-Resolution Image Synthesis
- **Samuel Zhang:** What Matters in Transformers? Not All Attention is Needed
- **Andrew Yin: ??**
- **Gloria Kao:** ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate
- **Yi Zhang**: Fast Inference from Transformers via Speculative Decoding
- **Bill Wang**: Can AI Be as Creative as Humans?
- **Arul Mathur**: The Geometry of Concepts: Sparse Autoencoder Feature Structures

Intuition of Policy Gradient

Gradient: $$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p(\tau;\theta)} \left[r(\tau) \nabla_\theta \log p(\tau;\theta) \right]$$

Interpretation:

- If r(τ) is high, push up the probabilities of the actions seen
- If r(τ) is low, push down the probabilities of the actions seen

Intuition of Policy Gradient

Gradient: $\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p(\tau;\theta)} \left[r(\tau) \nabla_\theta \log p(\tau;\theta) \right]$

Interpretation:

- If r(τ) is high, push up the probabilities of the actions seen
- If r(τ) is low, push down the probabilities of the actions seen

Might seem simplistic to say that if a trajectory is good then all its actions were good. But in expectation, it averages out!

Intuition of Policy Gradient

Gradient: $\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p(\tau;\theta)} \left[r(\tau) \nabla_\theta \log p(\tau;\theta) \right]$

Interpretation:

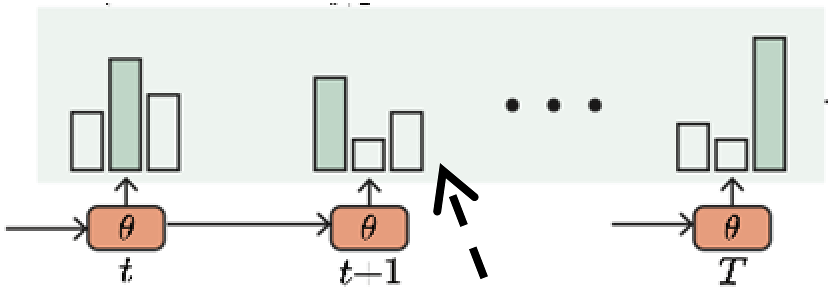
- If r(τ) is high, push up the probabilities of the actions seen
- If r(τ) is low, push down the probabilities of the actions seen

Might seem simplistic to say that if a trajectory is good then all its actions were good. But in expectation, it averages out!

However, this also suffers from high variance because **credit assignment** is really hard.

RL for LLMs

RL for Text Generation: Formulation



- (Autoregressive) text generation model:

Sentence $\boldsymbol{y} = (y_0, \ldots, y_T)$

$\pi_\theta(y_t \mid \boldsymbol{y}_{<t}) = \mathrm{softmax}(\ f_\theta(y_t | \boldsymbol{y}_{<t})\ )$

logits

In RL terms:

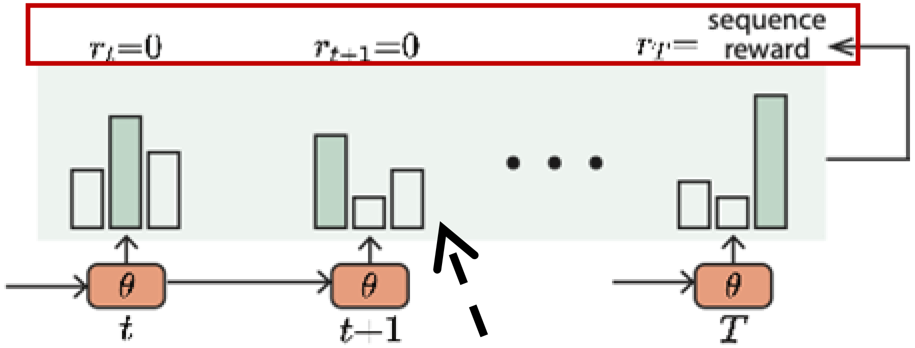
trajectory, τ

action, a_t

state, $\boldsymbol{s}_t$

policy $\pi_\theta(a_t \mid \boldsymbol{s}_t\ )$

RL for Text Generation: Formulation



- (Autoregressive) text generation model:

Sentence $\boldsymbol{y} = (y_0, \dots, y_T)$ $\qquad \pi_\theta(y_t \mid \boldsymbol{y}_{<t}) = \text{softmax}(\ f_\theta(y_t | \boldsymbol{y}_{<t})\ )$

logits

In RL terms: trajectory, τ action, a_t state, $\boldsymbol{s}_t$ policy $\pi_\theta(a_t \mid \boldsymbol{s}_t\ )$

- Reward $r_t = r(\boldsymbol{s}_t, a_t)$
  - Often **sparse**: $r_t = 0$ for $t < T$
- The general RL objective: maximize cumulative reward

$$J(\pi) = \mathbb{E}_{\tau \sim \pi}\left[\sum_{t=0}^{T} \gamma^t r_t\right]$$

- Q-function: expected *future* reward of taking action a_t in state $\boldsymbol{s}_t$

$$Q^\pi(\boldsymbol{s}_t, a_t) = \mathbb{E}_\pi\left[\ \sum_{t'=t}^{T} \gamma^{t'}\ r_{t'} \mid\ \boldsymbol{s}_t, a_t\ \right]$$

Presentations

Questions?