Introduction to Generative AI

2023.11.30

Administrative Questions (incl. test outlook) / Theory Recap / … / HW

01

Administrative Questions

Test Characteristics

- Date & Time: **07.12.2023 | 13:20 – 13:50 (30min)** | Lecture afterwards
- Online on TUWEL (both in presence as well as at home, but only at this exact time)
- Multiple Choice:
  - 100% for a question only if all selected answers are correct
  - 0% awarded for any incorrect or incomplete responses
  - no negative points for wrong answers
- Closed Booked

Test Content

- Birds Eye View of the Architecture of a Large Language Model
- Overview of the purpose of the different layers (till MHA)
- Overview of how the layers transform their input (till MHA)
- High-Level Overview of the training steps from nothing to RLHF
- Basics of RL and PPO

Exemplary Question

- What conceptual analogy describes the dot product in $\text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$ the best?
    - A) The dot product produces the attention score and describes how much tokens should pay attention to each other in a given sequence; one can call it the affinity for each word with each other word in a given sequence.
    - B) The dot product measures the syntactic similarity between tokens, focusing on grammatical and structural parallels rather than contextual relevance.
    - C) The dot product represents the literal position of each token in the sequence, assigning a unique spatial value to each word that determines its order in the sequence.
    - D) The dot product is used to directly modify the inherent meanings of a given token based on their interactions within the sequence.

Homework Grading

- Homework 1 is graded
- By the end of this week, it will be visible in TUWEL.

02

Recap RL & PPO

Where is J in my homework?

The Roles in PPO

- An *agent* interacts with an *environment* by taking *actions*.
- The agent uses a *policy* (a strategy/rule system) to decide which action to take based on the current *state* of the environment.
- A judge then rewards/penalises the agent for its actions.
- It is our goal to find a policy that *maximizes* the cumulative reward J

Objective Function: Where we left off

$$J_{CLIP+VS+S}(\theta) := \hat{\mathbb{E}}_t \left[J_{CLIP}(\theta) - c_1 J_{VF}(\theta) + c_2 S[\pi\theta](s_t) \right]$$

Objective Function: Where we started

$$J_{PG}(\theta) = \hat{\mathbb{E}}_t[\log \pi_\theta(a_t|s_t)\hat{A}_t]$$

So What?

A step-by-step guide for the objective function

Objective Function: Where we started

$$J_{PG}(\theta) = \hat{\mathbb{E}}_t[\log \pi_\theta(a_t|s_t)\hat{A}_t]$$

$$J_{PG}(\theta) = \hat{\mathbb{E}}_t[\log \pi_\theta(a_t|s_t)\hat{A}_t]$$

- What does the objective function return, what does it mean, and where is it in our homework?
  - A scalar value [-∞, ∞]
  - Indicator of how well a policy performs on average across a range of states and actions
  - Code: *loss = -torch.log(agent_output[word_index]) * advantage (Go to Notebook)*
- What is θ, what does it mean, and where is it in our homework?
  - A set of variables that define our policy
  - i.e. the parameters of our LLM
  - Code: Under the hood of our Agent class (Go to Notebook)

$$J_{PG}(\theta) = \hat{\mathbb{E}}_t[\log \pi_\theta(a_t|s_t)\hat{A}_t]$$

- What is E, what does it mean and where is it in our homework?
  - A weighted average of all t values
  - Code: *Not existing*
- What is π, what does it mean, and where is it in our homework?
  - A policy that selects an action *a* given the state *s*
  - The forward function (i.e. the brain) of our LLM
  - Code: *Under the hood of our Agent class (Go to Notebook)*

$$J_{PG}(\theta) = \hat{\mathbb{E}}_t[\log \pi_\theta(a_t|s_t)\hat{A}_t]$$

- What is A, what does it mean and where is it in our homework?
  - The *Advantage Function* returns a scalar value [-∞, ∞]
  - Measures how much better / worse a given action a state s is compared to the average action in that state under the current policy.
  - Code: *(We cheated) advantage = critic_score - baseline.get()* (Go to Notebook)

Objective Function: Where we started | Recap

$$J_{PG}(\theta) = \hat{\mathbb{E}}_t[\log \pi_\theta(a_t|s_t)\hat{A}_t]$$

But what about the test though?

Do I have to calculate the objective value on the board in front of the whole class?

But what about the test though?

- Understand the different components of a PPO system (agent, policy, environment, …)
- Understand the simplified objective function (just discussed) and its components (i.e. what is on the slides)
- Discussed last time: Understand the potential issues of the simplified objective function and what is done to counteract them. (This will be the 1,2 hard questions in the test)
    - Clipping to avoid overaggressive updates
    - Penalty term with Kullback-Leibler to penalise overaggressive updates

Exemplary Question

- What is the role of π in PPO, and how is it represented in RL of an LLM?
    - A) π is the learning rate in PPO that determines how quickly an LLM adapts to new states. In RL, this is represented by the rate at which the loss function of the LLM decreases.
    - B) π is the policy that decides on a specific action the agent should take at a given state s. In the context of RL for LLMs, it is represented by the LLM itself or, more precisely, the forward function.
    - C) π refers to the discount factor that quantifies future rewards' importance. In the RL framework for LLMs, it is represented by the activation function in the neural network layers.
    - D) π is the reward function that assigns a value to each action taken by the agent. For LLMs in RL, this reward function is encapsulated in the backpropagation algorithm.

What about the slides / the code?

- These slides will go online till EOD.
- The example solution for homework 3 will go online till EO tomorrow.

Questions / Concerns / Fears / Whishes?