1.

6. **What if the time pressure is greater, and we really need to land? We can model that by changing the reward. Set the reward to -0.1, -0.2, and -0.5 and run policy iteration. Add a paragraph to assignment6.pdf explaining how the policy changes for the lander.**

For the default reward of -0.04, the simulation shows that the lander reaches the safe goal at position 5,5 every time over 1000 iterations. The same perfect success rate is observed when the reward is set to -0.1 and -0.2. Conceptually, a more negative reward per move forces the lander to choose shorter, more direct paths toward its goal since each additional step incurs a higher cost. Interestingly, when the reward is changed to -0.05, the lander ends up at position 2,3, the sandstorm cell, in every simulation run. This change illustrates how even a slight tweak in the reward can alter the balance between taking a fast route and avoiding hazards.

3. **Use your favorite LLM-based tool (I recommend NotebookLM for this, but you can use whatever you want) to create a summary of this article that explains the main ideas and helps you to understand them. Include the results in your repo.**

Briefing Document: Reinforcement Learning from Human Feedback (RLHF)

Date: October 26, 2023 Source: Excerpts from "Illustrating Reinforcement Learning from Human Feedback (RLHF)" (Published December 9, 2022) Authors: Nathan Lambert, Louis Castricato, Leandro von Werra, Alex Havrilla

Executive Summary:

This document provides a detailed overview of Reinforcement Learning from Human Feedback (RLHF), a technique that has significantly improved the ability of large language models (LLMs) to generate text aligned with complex human values and preferences. The article breaks down the RLHF training process into three key stages: pretraining a language model, training a reward model using human feedback, and fine-tuning the language model with reinforcement learning (typically using Proximal Policy Optimization - PPO). While RLHF has achieved impressive results, notably in models like ChatGPT, the article highlights ongoing challenges, limitations related to the cost and quality of human feedback data, and numerous open research questions regarding optimal model architectures, RL algorithms, and data utilization. The emergence of open-source tools is also noted as a crucial development in democratizing RLHF research and application.

Main Themes and Important Ideas:

1. The Challenge of Defining "Good" Text and the Limitations of Traditional Loss Functions and Metrics:
    - The article begins by emphasizing the subjective and context-dependent nature of what constitutes "good" text. Different applications demand varying qualities (e.g., creativity for stories, truthfulness for informative text, executability for code).
    - Traditional next token prediction loss functions (like cross-entropy) used in pretraining LMs do not directly capture these nuanced human preferences.
    - Evaluation metrics like BLEU and ROUGE, which compare generated text to references using simple rules, are also limited in their ability to assess subjective qualities.
    - Quote: "However, what makes a 'good' text is inherently hard to define as it is subjective and context dependent."
    - Quote: "While being better suited than the loss function itself at measuring performance these metrics simply compare generated text to references with simple rules and are thus also limited."
2. The Core Idea of RLHF: Directly Optimizing LMs with Human Feedback:
    - RLHF addresses the limitations of traditional methods by using human feedback as a direct signal for optimization through reinforcement learning.
    - This allows models trained on general text corpora to be aligned with complex human values and preferences.
    - Quote: "That's the idea of Reinforcement Learning from Human Feedback (RLHF); use methods from reinforcement learning to directly optimize a language model with human feedback."

Quote: "RLHF has enabled language models to begin to align a model trained on a general corpus of text data to that of complex human values."

3. The Three Core Steps of RLHF Training:

Pretraining a Language Model (LM):RLHF starts with a conventionally pretrained language model (e.g., a version of GPT, Transformer models).
This initial model provides a strong base for further fine-tuning.
The article notes that the optimal choice of the pretrained model is an open research question.
Quote: "As a starting point RLHF use a language model that has already been pretrained with the classical pretraining objectives..."
Gathering Data and Training a Reward Model (RM):This step involves collecting human preferences on generated text to train a reward model.
The reward model's goal is to take a text sequence and output a scalar reward representing human preference.
Instead of directly scoring text, humans typically rank multiple generated outputs for the same prompt, which provides a more robust and calibrated dataset.
Techniques like pairwise comparisons and Elo systems are used to generate rankings.
The size of the reward model relative to the language model can vary.
Quote: "Generating a reward model (RM, also referred to as a preference model) calibrated with human preferences is where the relatively new research in RLHF begins."
Quote: "Instead, rankings are used to compare the outputs of multiple models and create a much better regularized dataset."
Fine-tuning the LM with Reinforcement Learning:The pretrained LM is fine-tuned using a reinforcement learning algorithm, most commonly Proximal Policy Optimization (PPO).
The reward function for the RL process combines the output of the reward model (preferability score) with a penalty term, typically based on the Kullback-Leibler (KL) divergence.
The KL divergence penalty prevents the fine-tuned model from deviating too drastically from the initial pretrained model, ensuring coherent outputs.
The policy in this RL setup is the language model, the action space is the vocabulary tokens, and the observation space is the input prompt.
Due to the size of LLMs, often only some parameters are updated during fine-tuning for computational efficiency.
Quote: "What multiple organizations seem to have gotten to work is fine-tuning some or all of the parameters of a copy of the initial LM with a policy-gradient RL algorithm, Proximal Policy Optimization (PPO)."
Quote: "The reward function is a combination of the preference model and a constraint on policy shift."

4. Iterative RLHF and Online Data Collection:

The article mentions the possibility of iteratively updating both the reward model and the policy.
This often involves collecting new human feedback on the latest versions of the model, particularly relevant for dialogue agents.
Anthropic's "Iterated Online RLHF" is cited as an example of this approach.
This introduces complex dynamics as both the policy and the reward signal evolve.

5. Open-Source Tools for RLHF:

The article highlights the emergence of open-source libraries in PyTorch that facilitate RLHF research and development.

Key repositories mentioned include Transformers Reinforcement Learning (TRL), TRLX (focused on larger models), and Reinforcement Learning for Language models (RL4LMs) (offering a wider variety of RL algorithms).
The availability of these tools democratizes access to RLHF techniques.
Quote: "Today, there are already a few active repositories for RLHF in PyTorch that grew out of this. The primary repositories are Transformers Reinforcement Learning (TRL), TRLX which originated as a fork of TRL, and Reinforcement Learning for Language models (RL4LMs)."
The availability of datasets like Anthropic's on the Hugging Face Hub is also noted.

6. Limitations and Future Directions of RLHF:

Despite its success, RLHF models still have limitations, including the potential to generate harmful or factually incorrect text.
Gathering high-quality human preference data is expensive, involving either costly human-generated text or the labeling of model outputs.
Disagreements among human annotators can introduce noise and variance into the training data.
Numerous unexplored design options exist, particularly in improving the RL optimizer (exploring algorithms beyond PPO, offline RL techniques like Implicit Language Q-Learning (ILQL)).
The trade-offs in the RL process, such as exploration-exploitation balance, are not yet well-documented for RLHF.
The formulation of the reward function is expected to continue evolving.
Quote: "While these techniques are extremely promising and impactful and have caught the attention of the biggest research labs in AI, there are still clear limitations."
Quote: "RLHF performance is only as good as the quality of its human annotations..."

7. Key Papers in the Field:

The article provides a chronological list of significant papers in RLHF, categorized into earlier work on preference-based reinforcement learning and more recent papers focusing on language model alignment.
This list serves as a valuable resource for further exploration of the field.
Examples include: TAMER (2008), Deep Reinforcement Learning from Human Preferences (2017), Fine-Tuning Language Models from Human Preferences (2019), InstructGPT (2022), ChatGPT (2022), and Llama 2 (2023).

Conclusion:

RLHF represents a significant advancement in aligning language models with human preferences and values. By directly incorporating human feedback into the training process through a multi-stage approach involving a reward model and reinforcement learning, LLMs have achieved unprecedented levels of helpfulness, honesty, and harmlessness in their generated text. However, the field is still relatively young, with substantial challenges and open research questions remaining, particularly concerning the cost and quality of human feedback data and the exploration of alternative RL optimization techniques. The growing availability of open-source tools and datasets is crucial for fostering further innovation and democratizing this powerful technology.

convert_to_textConvert to source