An LM-Powered Deep-Q Dive into Complex Reasoning

Matt Wise
mattwise@stanford.edu

Introduction

Our objective is to improve performance on complex reasoning tasks by pairing a Language Model with a Deep-Q Network – abbreviated LMDQN. LMDQN can augment human activity by autonomously completing complex tasks that typically require human research and reasoning.

We focus on "multi-hop" questions, which involve considering multiple facts or relations to derive an answer. LMDQN produces leading results for yes/no questions and shows promising opportunity for additional performance gains.

Data

Training/validation/test: 1000/250/250 entries from the HotpotQA dataset[2], including this sample entry:

Question: "What stock exchange lists a competitor of Ladbrokes?"
Answer: "London Stock Exchange"
Difficulty: Medium *(options are Easy, Medium, Hard)*
Type: Bridge *(options are Bridge, Comparison)*

Language Model (LM) Architecture

A standard GPT 3.5 setup feeds outputs into DQN.

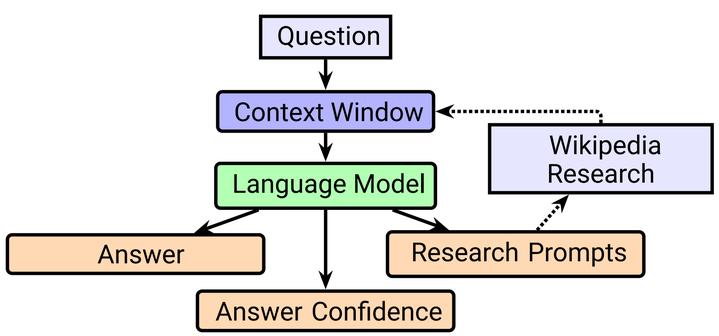

Figure 1. LM generates outputs that are fed to the DQN

Deep-Q Network (DQN) Architecture

Our DQN incorporates LM outputs as part of the state and decides the next action.

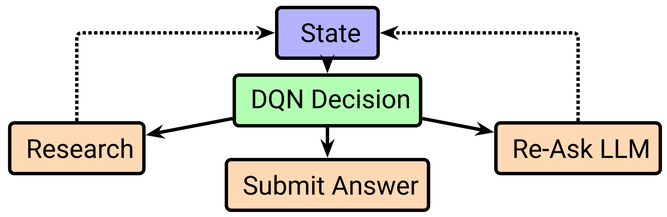

Figure 2. DQN decides next action

DQN applies MSE to optimize the Bellman equation[1]:

$$J(\theta) = \mathbb{E}_{s,a,s' \sim \mathcal{D}_{\text{Replay}}} \left\{ \left(r + \gamma \max_{a'} Q_\theta(s', a') - Q_\theta(s, a) \right)^2 \right\}$$

Results

We primarily mirror the HotpotQA paper metrics for easy comparison:

Model	Exact Match (%)	F1 (%)
LMDQN	30.80	**86.36**
Beam Retrieval[3]	**67.46**	80.52
HotpotQA[2]	23.95	32.89

Table 1. Test results vs. benchmarks at hotpotqa.github.io

Training rewards increase but have more room to improve:

Figure 3. Average Reward by Episode

Discussion Points

Strong F1 Score Performance

On yes/no questions, LMDQN surpasses current leaders (F1 Score), underscoring its promising potential.

Expanding max depth to improve EM Score

Resource constraints forced us to limit the model's max depth, which prevented deep research and especially limited performance on harder bridge questions.

Potential with deeper training

Training rewards improved but show room for continued optimization. With deeper training (larger sample, more episodes, slower epsilon decay), we believe the model's performance can increase significantly.

Next Steps

With additional resources, the most promising areas for improvement would be to explore **increased max action depth**, **deeper training**, and **expanded state definitions**.

References

[1] Shengbo Eben Li.
Reinforcement Learning for Sequential Decision and Optimal Control.
Springer Singapore, 2023.

[2] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning.
Hotpotqa: A dataset for diverse, explainable multi-hop question answering, 2018.

[3] Jiahao Zhang, Haiyang Zhang, Dongmei Zhang, Yong Liu, and Shen Huang.
Beam retrieval: General end-to-end retrieval for multi-hop question answering, 2023.