Explaining Reinforcement Learning and GRPO with Manim

Declan Young

April 25, 2025

Abstract

New AI models, such as Deepseek, typically use models on the cutting edge of technology and thus may not have easily digestible resources explaining how the model itself works. This project aimed to bridge that gap by creating an educational video using the open-source animation library Manim to make it more accessible for viewers without prior knowledge through visual storytelling and step-by-step breakdowns. The video provides a high-level introduction to reinforcement learning and the GRPO (Group Relative Policy Optimization) algorithm used in Deepseek, and appeared to succeed to a moderate extent in achieving this goal.

Introduction

As the field of artificial intelligence continues to progress, new AI chatbots – each with differing functions and technologies – are constantly being created and improved. However, as a large majority of these models are on the cutting edge of AI, and thus use new models and findings, resources detailing how the model works are limited or difficult to understand digestible – especially for those who lack the required prior knowledge or technical experience. One such example is relatively recent creation of Deepseek. More specifically, that of its V3 and R1 models.

Deepseek R1 is a Chinese AI large language model (LLM) created with reinforcement learning strategies. At the core of its success, is the Group Relative Policy Optimization (GRPO) algorithm that builds upon the existing Proximal Policy optimization (PPO) [8]. As the name implies, GRPO is an algorithm that aims to optimize the policy in reinforcement learning.

This project aims to create a video explaining how the reinforcement learning algorithm behind Deepseek's success GRPO works through the use of the open-sourced Manim library [3]. This video should accurately and effectively explain the concept to a person without any prior knowledge about the subject. This will ideally give the viewer a thorough understanding of not only how Deepseek works mathematically, but also how reinforcement learning works to a high level as well.

The code for all the Manim animations created for the video can be found on the project's GitHub

Related Work

There are a couple of videos that explain how the equation behind Deepseek, group relative policy optimization (GRPO), works. While these videos do seem to use Manim to explain how GRPO works, they hone in specifically on how the GRPO equation works and can be difficult to follow for someone less versed in the subject as a whole. A notable example is [5], which explains the math behind GRPO quite well, but is unfortunately hard to follow for someone unfamiliar with the subject. Another notable video is [6], which explains the topic very well and in simpler terms, but it lacks some nicer Manim animations. Furthermore, there are also articles such as [4] that explain how the math behind Deepseek works as well. Although it explains the math behind Deepseek pretty well, it is still an article, and thus may be less digestible for the average person.

Approach (or Methodology)

To address this issue, a video, edited with Final Cut Pro, explaining the math behind Deepseek was created. This video first gives a brief introduction to Deepseek, followed up by an introduction to reinforcement learning, and finally concludes with the GRPO equation. The reason why an introduction to reinforcement learning was given, was to hopefully give the viewer the perquisite knowledge required to understand why GRPO is used in the first place. Additionally, as this video is catered towards viewers without much prior knowledge on Deepseek and/or how machine learning models work, the depth of the video was kept at a relatively high level. Instead, it attempted to focus on how the GRPO equation works and related it back to Deepseek. To better facilitate learning and understanding by allowing viewers to visualize how the equation works, animations created through the use of Manim were used. Manim is an open-sourced python library used to create mathematical animations created by the Grant Sanderson, the creator of the YouTube channel 3Blue1Brown, to use for his mathematical-based YouTube videos. For this project, the community version of Manim was used since it can be thought of as technically more stable and well documented due to the volume of people maintaining the library as opposed to just its creator. To briefly outline the structure of the video for more clarity, it is:

1. Introduction to Deepseek (What is it?)
2. Introduction to Reinforcement Learning
3. Introduction and explanation of Group Relative Policy Optimization (GRPO)

4. GRPO Example
5. Conclusion

For the introduction section, due to the lack of math related concepts to explain, images small clips, and text relating to what was being explained at the time were used to accompany the explanation and provide engagement and/or reinforcement of concepts. Some examples were the use of text highlighting the reported cost of Deepseek (6 million), and a clip showcasing Deepseek R1's "thinking".

As for the remaining sections, multiple animations were created to through the use of Manim to aid with the understanding of the respective concept worked. There were 2 prominent types of Manim animations created for both the introduction to reinforcement learning and GRPO sections. The first of which was the decision to use multiple GridWorld animations for the reinforcement learning section, while the second were animations highlighting and transforming the GRPO equation. Firstly, GridWorld animations were used since they provided a relatively simple way to code a beginner-friendly reinforcment learning example in Manim. GridWorld is a simple game in which the player, placed in a grid environment (in this case a 5x5 one), has to move to the top right corner of the grid whilst avoiding traps and potentially collecting coins in the process. Initially, the idea to use an analogy of a baby to explain reinforcement learning was going to be used, however, as there is not really a method to animate and explain that through manim, it was rejected in favour of GridWorld. Next, most of the animations for the GRPO equation were transformations of the GRPO equation to showcase value substitutions as there were not too many good animations to showcase how the equation worked. Certainty, some animations showing the clipping portions and probability distributions were created, but they were rather few and far between. Hence, it was decided to use detailed substitutions to showcase the inner-workings of GRPO.

The full script can be found at this google docs link

As for the video, it can be found at this YouTube Video

Datasets

As this project is a Pedagogy/Explainer Project, a dataset will not be created or used.

Evaluation Results

The primary goal of this project was to explain the reinforcement learning algorithm behind Deepseek, GRPO, to audiences without a technical background. Although formal user testing was not extensively conducted, a Google form with the following questions was sent to a couple of people for feedback:

1. Do you have prior experience with any of the presented subjects (i.e machine learning, reinforcement learning, GRPO, etc.)?
2. How clearly did the video explain the topic of reinforcement learning?
3. How well do you understand the GRPO algorithm after watching the video?
4. Did the animations help your understanding?
5. How was the pacing and structure of the explanation?
6. Was the content accessible to someone without prior technical knowledge?
7. Were examples or analogies used effectively to simplify concepts?
8. Any other comments?

Not too many people filled it out. However, from the handful of responses that were collected as of the publishing of this report, it seems that the video does a moderately well job of accomplishing the initial goal of creating an educational video using the open-source animation library Manim, through visual storytelling and step-by-step breakdowns, that provides a high-level introduction to reinforcement learning and the GRPO (Group Relative Policy Optimization) algorithm used in Deepseek for viewers without prior knowledge. This can be attributed to the fact that most of the responses to the questionaire were 3s or above.

The form can be found at this link

Conclusion

The project aimed to create an educational video for a non-technical audience that explained the math behind Deepseek – specifically that of the GRPO algorithm – through the use of animated visuals made with Manim. In addition to GRPO, the video also briefly introduced both Deepseek-R1 itself and reinforcement learning. Although not extensively evaluated, early feedback suggests that the video succeeded to a moderate extent. To summarize, through the process of creating the video, I not only learned how to use manim to a moderate extent, but also learned how GRPO itself worked as well. Manim certainty is can be a useful tool in helping understanding of emerging AI technologies similar to that of standard math concepts.

References

[1] DeepSeek-AI. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. DeepSeek-AI, 2025.

[2] AI with Alex. *DeepSeek R1 Explained to your grandma.* YouTube 2025. https://www.youtube.com/watch?v=kv8frWeKoeo

[3] The Manim Community Developers Manim - Mathematical Animations Github, 2025

[4] Daniel Warfield DeepSeek-R1 — Intuitively and Exhaustively ExplainedThe Manim Substack, 2025 https://iaee.substack.com/p/deepseek-r1-intuitively-and-exhaustively

[5] AGI Lambda. *Group Relative Policy Optimization(GRPO) Visualized.* YouTube, 2025. https://www.youtube.com/watch?v=EX8-ucKOBbA

[6] Dr Mihai Nika. *How DeepSeek learns: GRPO explained with Triangle Creatures.* YouTube, 2025. https://www.youtube.com/watch?v=wXEvvg4YJ9I&t=982s

[7] Patel et al. *DeepSeek Debates: Chinese Leadership On Cost, True Training Cost, Closed Model Margin Impacts* SemiAnalysis, 2025. https://semianalysis.com/2025/01/31/deepseek-debates/

[8] Shao et al. *DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models* arXiv, 2024. https://arxiv.org/abs/2402.03300

[9] Kalimanie. *The Shocking Truth About DeepSeek's AI Training Costs!* Medium, 2025. https://medium.com/@kalimanie58/the-shocking-truth-about-deepseeks-ai-training-costs-92f2f8abac99

[10] Gabriele De Luca and Michal Aibin. *What Is a Policy in Reinforcement Learning?* Baeldung, 2025. https://www.baeldung.com/cs/ml-policy-reinforcement-learning

[11] Ben Thompson. *DeepSeek FAQ* Stratechery, 2025. https://stratechery.com/2025/deepseek-faq/