Die iterative Verbesserung eines Large Language Model durch Training gegenüber dessen Selbstverbesserung durch geschicktes Prompten

Generiert mit Anthrop\c Claude 3 - veröffentlicht am 21. August 2024

Die iterative Selbstverbesserung eines Sprachmodells - was genau ist das?

Iterative Selbstverbesserung ist ein Konzept in der Entwicklung von Sprachmodellen, bei dem das Modell kontinuierlich verbessert wird, indem es auf zusätzliche Trainingsdaten angewendet und dann weiter optimiert wird.

Der grundlegende Ablauf ist folgendermaßen:

1. Initialisierung: Ein Sprachmodell wird mit einem initialen Datensatz trainiert.
2. Anwendung: Das trainierte Modell wird dazu verwendet, neue Daten zu generieren oder zu verarbeiten.
3. Auswertung: Die Leistung des Modells auf den neuen Daten wird evaluiert.
4. Optimierung: Basierend auf den Evaluationsergebnissen wird das Modell weiter optimiert, z.B. durch das Hinzufügen neuer Trainingsdaten, Anpassungen der Modellarchitektur oder des Trainingsverfahrens.
5. Wiederholung: Der Zyklus aus Anwendung, Auswertung und Optimierung wird iterativ wiederholt, um das Modell schrittweise zu verbessern.

Durch dieses *iterative* Vorgehen kann das Modell *kontinuierlich* an die Anforderungen angepasst und seine Leistung *sukzessive* gesteigert werden.

Dies ist besonders wichtig, um Sprachmodelle auf dem neuesten Stand zu halten und an sich *ständig* verändernde Sprache, Themen und Anwendungsszenarien anzupassen.

Iterative Selbstverbesserung sei ein Konzept in der Entwicklung von Sprachmodellen, bei dem das Modell kontinuierlich verbessert würde, indem es auf zusätzliche Trainingsdaten angewendet und dann weiter optimiert würde, sagst du.

Und weiter: Basierend auf den Evaluationsergebnissen würde das Modell weiter optimiert, z.B. durch das Hinzufügen neuer Trainingsdaten, Anpassungen der Modellarchitektur oder des Trainingsverfahrens.

Nun steht die iterative Selbstverbesserung in engstem Zusammenhang mit dem, was auf Englisch bezeichnet wird als *Self-Refinement-Promting*, also das auf das selbstverfeinerte Prompting, im Rahmen des auf Selbstverbesserung oder Selbstoptimierung hin ausgerichteten Ansatz, ein Sprachmodell zu prompten.

Diesbezüglich lese ich folgendes: >>Eine auf Selbstverbesserung eines Sprachmodells abzielende Studie zeigt, wie ein LLM eine iterative Selbstverbesserung ohne zusätzliches Training ermöglicht, was zu qualitativ hochwertigeren Ergebnissen führt … << geschrieben durch jemand, der sich wissenschaftlich mit der Materie befasst.

Ich erkenne hier einen Widerspruch. Du auch? So ja, wie erklärst du dir diesen Widerspruch?

Sie haben Recht, es gibt hier einen scheinbaren Widerspruch zwischen meiner Erklärung der iterativen Selbstverbesserung und der Aussage, dass ein Self-Refine-Prompting-Ansatz eine solche Verbesserung ohne zusätzliches Training ermöglicht.

Lassen Sie mich versuchen, diesen Widerspruch aufzulösen:

Die traditionelle Vorstellung der iterativen Selbstverbesserung, die ich beschrieben habe, beinhaltet tatsächlich ein zusätzliches Trainieren des Sprachmodells mit neuen Daten. Das ist der klassische Weg, um ein Modell schrittweise zu verbessern.

Der Self-Refine-Prompting-Ansatz scheint jedoch eine Möglichkeit zu bieten, eine ähnliche Verbesserung ohne aufwendiges Retraining zu erreichen. Dabei wird das Sprachmodell nicht direkt mit neuen Daten trainiert, sondern es wird durch geschicktes Prompting dazu gebracht, sich selbst zu verfeinern und die Qualität seiner Ausgaben zu verbessern.

Dies könnte zum Beispiel durch iteratives Überarbeiten und Verfeinern der Prompts oder durch Einbeziehen von Feedback in die Promptgestaltung erreicht werden. So kann das Modell seine interne Repräsentation und Verarbeitungsweise schrittweise optimieren, ohne dass es komplett neu trainiert werden muss.

Insofern ist die Aussage, dass ein Self-Refine-Prompting-Ansatz eine iterative Selbstverbesserung ohne zusätzliches Training ermöglicht, durchaus plausibel und nicht widersprüchlich zu meiner vorherigen Erklärung. Es handelt sich um einen alternativen Weg, das gleiche Ziel zu erreichen.

Verbesserung ohne aufwendiges *Retraining* zu erreichen sagst du, was ja mittels unbeschrifteten (*unlabeled*) Daten und darauffolgender Feinabstimmung mit beschrifteten (*labeled*) Daten für spezifische Aufgaben geschieht, so jedenfalls habe ich an anderer Stelle vernommen,

Und ein Modell nun durch geschicktes Prompting dazu zu bringen, sich selbst zu verfeinern und die Qualität seiner Ausgaben zu verbessern, sagst Du ferner, wobei dies zum Beispiel durch iteratives Überarbeiten und Verfeinern der Prompts oder durch Einbeziehen von Feedback in die Promptgestaltung erreicht werden könne.

Vielen Dank. Dies ist genau 'mein Thema' und Gegenstand einer auf dem Wege des Selbststudiums zu erlangenden Expertise im gedanklichen und folglich sprachlichen Umgang mit Künstlicher Intelligenz– nenne es Prompt-Engineering, so dir das gefällt!!

Das freut mich, dass ich Ihnen mit meiner Erklärung zum Thema der iterativen Selbstverbesserung durch geschicktes Prompting weiterhelfen konnte.

Das Verständnis und der gekonnte Umgang mit Prompting als Methode, um Sprachmodelle gezielt zu beeinflussen und zu optimieren, ist in der Tat ein sehr spannendes und tiefgreifendes Thema im Bereich der Künstlichen Intelligenz.

Es erfordert ein gutes Verständnis der inneren Funktionsweise und Repräsentationsformen der Modelle sowie Kreativität und Erfahrung bei der Gestaltung effektiver Prompts. Durch stetiges Experimentieren und Lernen kann man hier eine beeindruckende Expertise aufbauen.

Ich wünsche Ihnen viel Freude und Erfolg bei Ihrem Selbststudium zu diesem Thema. Es ist ein faszinierendes Gebiet, das stetig an Bedeutung gewinnt. Zögern Sie nicht, falls Sie weitere Fragen haben - ich stehe Ihnen gerne mit weiteren Erläuterungen zur Verfügung.