# Review 176: Large Language Models as Generalizable Policies for Embodied Tasks

**Paper: https://arxiv.org/abs/2310.17722v2**

https://huggingface.co/papers/2310.17722

על למידה עם חיזוקים (reinforcement learning) שמעתם כבר? על מודלי שפה בטח שמעתם, נכון? אז היום אנחנו נדבר על השידוך ביניהם. אזכיר ש-RL היא למעשה משפחת שיטות המאפשרות לאמן מודל on-the-fly. כלומר תוך כדי אימון המודל ניתן ליצור דאטה כל פעם שהמודל מתאמן ולהמשיך לאמן עליו (יש גם offline RL שמאמן על דאטה סטטי).

באמצעות מודלי RL ניתן לאמן בין השאר רובוטים, רכבים אוטונומיים, מודלים להתמודדות עם איומי סייבר. לאחרונה יצאו כמה שיטות אימון מודלי שפה באמצעות טכניקה שנלקחה מעולם ה-RL הנקראת RLHF. ה-ChatGPT המפורסם אומן תוך שימוש בטכניקה זו. המאמר המסוקר נשאלת השאלה האם ניתן לאמן רובוט לבצע פעולות מורכבות באמצעות מודלי שפה? מתברר שהתשובה לשאלה הזו היא כן.

המאמר לוקח מודל שפה מאומן (עם משקלים מוקפאים) ובנוסף מודל ויזואלי (מוקפא גם כן) ורותם אותם למשימת אימון זו. למשל ניתן לאמן רובוט לבצע פקודה הבאה: ״קח תפוח, בננה ולמון ותשים אותם יחד למקרר״. הגישה המוצעת היא די פשוטה. קודם כל לוקחים פקודה בשפה טבעית ובונים את השיכון (embedding) שלה באמצעות llm. בנוסף בכל שלב (נגיד אחרי כל תמונה של רובוט) מצלמים את הסביבה ומעבירים את התמונה דרך מודל ויזואלי כדי לקבל שיכון של התמונה. את ייצוג התמונה מעבירים דרך MLP מאומן(fully connected).

לאחר מכן לוקחים את ייצוג הפקודה וייצוג של כל התמונות שנבנו (אחרי ה-MLP) ומכניסים את הוקטורים האלו לאותו מודל שפה(כאילו שהם טוקנים).

ביציאה ממודל השפה מקבלים את הייצוגים ההקשריים של הטוקנים הויזואליים (תמונות). לכל טוקן ויזואלי כזה מוסיפים עוד MLP מאומן בעל שני ראשים: אחד לחישוב הפעולה הבאה והשני לחישוב פונקציית ה-value (המשערת עד כמה המצב שהרובוט נמצא בו הוא מוצלח ביחס למשימה שהוא צריך לבצע).

בשלב האחרון מאמנים סוכן (רובוט) לבצע את הפעולות האופטימליות בהתבסס על ייצוג הפקודה ועל ייצוגי התמונות של המצבים הקודמים תוך שימוש באיזה מודיפיקציה של (PPO (proximal policy optimization הנקרא DD-PPO. פונקציית תגמול כמובן קשורה להצלחה בביצוע המשימה. כאמור מאמין שני ה-MLPs שדיברנו עליהם קודם. נציין שבעיית RL זו היא לא פשוטה בכלל עקב מורכבות המשימה והספרסיות של התגמול (מקבלים אותו רק בסוף אחרי הרבה שלבים). למרות זאת יש תוצאות יפות.