משרד החינוך
המינהל הפדגוגי
האגף למחוננים ולמצטיינים

מכון הנרייטה סאלד
המכון הארצי למחקר במדעי ההתנהגות

תכנית מנחים–עמיתים

למידת התנהגות על ידי תחזית של מצבים עתידיים רצויים

חניך: שון גאלנצאן, תיכון קציר, רחובות
מנחה–עמית: פרופ׳ ליאור וולף, אוניברסיטת תל אביב

מטרת המחקר

מטרת המחקר היא לפתח שיטה חדשה ל-Reinforcement Learning, אשר מסוגלת ללמוד לשחק במשחקי מחשב ברמה דומה לאנשים או טובה יותר, אך בכמות קטנה יותר של משחקים וניסיונות בהשוואה לשיטות קיימות, וזאת על ידי שילוב של שיטות מסוג Unsupervised Learning ו-Supervised Learning.

מבוא

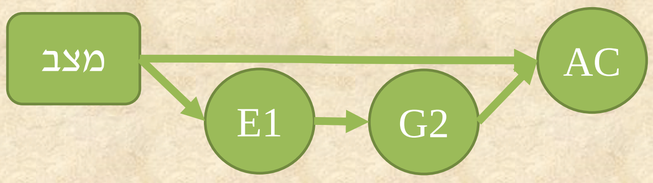
למידת מכונה מתחלקת לשלושה נושאים עיקריים: למידה מונחית (supervised learning), למידה בלתי מונחית (unsupervised learning) ולמידה מחיזוקים (reinforcement learning).

למידה מונחית – למידה המתבצעת בעזרת גישה למאגר נתונים הכולל אוסף קלטים ותוצאות רצויות, ומשתמשים בו על מנת לאמן מודל שיביא תוצאות הדומות, ככל האפשר, לתוצאות הרצויות עבור הקלטים שבמאגר.

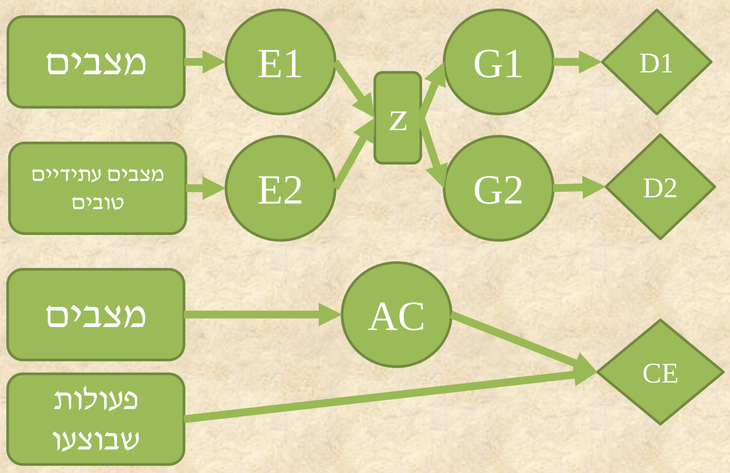
למידה בלי מונחית – למידה המתבצעת רק באמצעות מאגר של קלטים, מבלי שנתונות תוצאות רצויות. לעיבוד מידע מסוג זה יש כמה דרכים הכוללות חלוקה לקבוצות (clustering) או למידת ההתפלגות של המידע באמצעות מודל שיוכל לייצר קלטים חדשים הדומים לקלטים שבמאגר.

למידה מחיזוקים – למידה שמתבצעת באמצעות אינטראקציה עם סביבה מסוימת, כמו משחק מחשב לדוגמה, שבה המודל צריך ללמוד כיצד להתנהג כדי לקבל ניקוד גבוה ככל הניתן.

השיטה שלנו



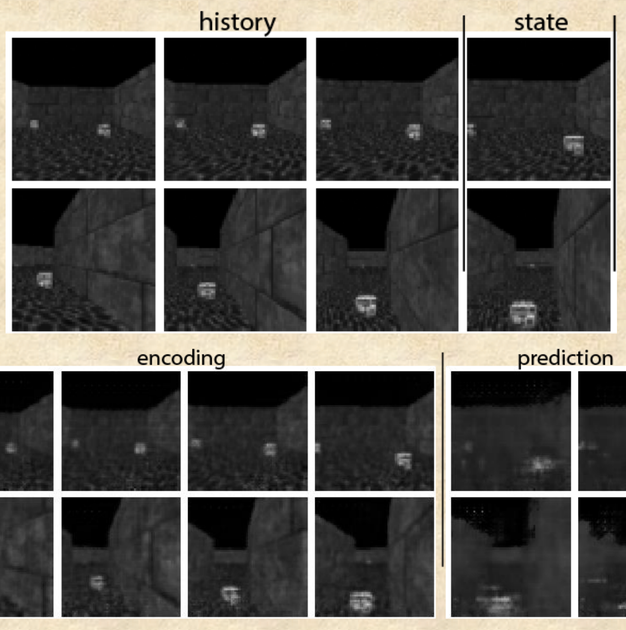
ארכיטקטורת המודל בזמן משחק



ארכיטקטורת המודל בזמן למידה

המודל שלנו מורכב משבע רשתות שונות של נוירונים – E1, E2, G1, G2, D1, D2 ו-AC – כשלכל רשת יש תפקידים אחרים: רשתות E1 ו-E2 מקבלות את המצב הנוכחי של המשחק או מצבים עתידיים טובים של המשחק, בהתאמה, וצריכות לתרגם אותם לשפה משותפת של הבינה המלאכותית; רשתות G1 ו-G2 מקבלות את הקידודים השונים של המצבים בשפה המשותפת, וצריכות להפוך אותם חזרה לתמונות המתארות את המצב הנוכחי או את המצבים הבאים; רשתות D1 ו-D2 מקבלות את התמונות שנוצרו על ידי G1 ו-G2, ובוחנות את האיכות שלהן, ולפי זה אפשר יהיה לאמן את רשתות E1, E2, G1 ו-G2; רשת AC מקבלת את המצב הנוכחי ואת התחזית של המודל למצבים הבאים ומצביעה על הפעולות שהמודל צריך לבצע.

תוצאות



דוגמאות לתחזיות הנוצרות על ידי E1 ו-G2 בזמן משחק

במודל שלנו, וגם במודלים אחרים, ללמידה מחיזוקים יש הרבה מאוד פרמטרים המשפיעים על תהליך הלמידה, ולכן, בדרך כלל, נעזרים ב״מחשב על״ המסוגל להריץ במקביל הרבה מאוד ניסויים עם פרמטרים שונים, על מנת למצוא את הפרמטרים הטובים ביותר ללמידה עבור כל סביבה.

הבדיקות שעשינו הראו שמספיק שהמודל יחזה רק שני מצבים קדימה. תחזית של מצבים נוספים לא תורמת ואף גורעת מהביצועים של המודל. באיור למעלה ניתן לראות דוגמאות של אחת מההרצות שהצליחה, כאשר היא חוזה רק שני מצבים קדימה.

ניקוד המשחקים עבור D2_navigation	לאחר 200,000 מצבי משחק	לאחר 400,000 מצבי משחק	לאחר 600,000 מצבי משחק
השיטה שלנו	513.77	553.46	612.41
DQN	416.67	438.12	452.72

מסקנות

עבור הסביבה D2_navigation הראינו שהמודל שלנו מצליח ללמוד יותר מהר מאשר DQN והוא מגיע לניקוד טוב יותר עבור הכמות הקטנה יחסית של משחקים ששוחקו.

נדרשת עבודה נוספת והרצת חיפושי פרמטרים נוספים על מנת לקבל תוצאות טובות עבור שאר הסביבות.

דוגמאות לסביבות שבהן בדקנו את המודל