המאמר היומי של טדי מייק 01.07.2025  
DINO-WM:World Models on Pre-trained Visual Features enable Zero-shot Planning

חוזר לסקור מאמרים בראייה ממוחשבת משולבת עם למידה באמצעות חיזוקים או RL. המאמר מציע גישה לאימון של מודל עולם(world model) לשימושים בעולם הרובוטיקה. כלומר המחברים מציעים גישה המאפשרת ללמוד איך לגרום לרובוט לבצע פעולות מסוימות בהתבסס על התיאור הויזואלי של הסביבה (קרי תמונות).

המודל DINO-WM מציג גישה חדשנית ל־World Modeling (בניית מודל של העולם) על־ידי הפרדה בין למידת דינמיקה חזותית(המתוארת על ידי ייצוג לטנטי של הסביבה) לבין שחזור פיקסלים ישיר ואופטימיזציה של תגמולים תלויי-משימה. החידוש המרכזי טמון בארכיטקטורה ובשיטת האימון של המודל, המנצלות תכונות חזותיות מאומנות מראש כדי לאפשר תכנון אפס-שוט (zero-shot planning). בגדול DINO-WM מציע לאמן את מודל החיזוי של הצעד הבא במרחב הלטנטי כאשר הדקודר שמשחזר את הפיקסלים מהייצוג הלטנטי מאומן בנפרד.

מודלים מסורתיים של העולם נאבקים לרוב בעלות החישובית של חיזוי במרחב הפיקסלים, או במגבלות של מודלים לטנטיים הקשורים למטרות של שחזור תמונה. DINO-WM פותר את הבעיה על־ידי פעולה מלאה במרחב לטנטי קומפקטי המאומן מראש. הוא משתמש ייצוג פאצ'ים שמופקות מ־DINOv2, בתור מודל התצפית שלו (תמונות). זהו שינוי מהותי לעומת עבודות קודמות שבהן מודל התצפית נלמד מאפס, לרוב בתלות במשימה. שימוש במקודד של DINOv2 שאינו ניתן ללמידה מאפשר ל־DINO-WM להנות מהצגות עשירות של אובייקטים ומרחבים שנלמדו ממאגרי מידע עצומים באינטרנט. זה הופך את מודל התצפית לבלתי תלוי במשימות וסביבות.

מודל המעברים בתוך DINO-WM, הבנוי על־גבי טרנספורמר חזותי (ViT), חוזה את ייצוגים העתידיים של הפאצ'ים ולא את הפיקסלים עצמם. החיזוי מתבצע במרחב הלטנטי, תוך התניה על היסטוריה של מצבים ופעולות. רכיב טכני מרכזי כאן הוא יישום של מנגנון attention סיבתי בתוך ה־ViT. בניגוד לגישות קודמות שמבצעות חיזוי אוטורגרסיבי ברמת טוקנים, DINO-WM חוזה ברמת פריים כלומר מתייחס לכל וקטורי הפאצ'ים של תצפית כאובייקט שלם. לפי המחברים, עיצוב זה קולט טוב יותר את המבנה הגלובלי והדינמיקה הטמפורלית, מה שמוביל להכללה טמפורלית טובה יותר. החיזוי מותנה גם בפעולות הסוכן, שממופות לממד גבוה יותר באמצעות MLP ומוצמדות לכל וקטור פאצ'ים.

אחד החידושים הבולטים של DINO-WM הוא הניתוק המוחלט של הדקודר (Decoder) מחיזוי המעבר הבא. ניתן להשתמש בדקודר על מנת לשחזר תמונות ממצבים לטנטיים לצרכי פרשנות, אך אימונו מנותק לחלוטין ממודל המעברים. המשמעות היא שיכולת התכנון והדינמיקה הפנימית של המודל אינן תלויות בשחזור פיקסלים, מה שמביא ליעילות חישובית רבה יותר בזמן אימון ובזמן ריצה. הדבר שונה ממודלים שבהם חיזוי במרחב לטנטי מחובר לשחזור תמונה, מה שעלול לפגוע באוניברסליות של ייצוגים הנלמדים שנדרשות לתיאור מוצלח של פיקסלים, במקום דינמיקה רלוונטית למשימה.

בזמן מבחן, תהליך האופטימיזציה של ההתנהגות מנוסח כבעיית הגעה ליעד חזותי במרחב הלטנטי. לוס של תכנון(חיזוי מעברים) מוגדרת כשגיאת ריבועים ממוצעת בין המצב הסופי החזוי למצב היעד הלטנטי. היכולת הזו לבצע תכנון zero-shot ללא הסתמכות על הדגמת מומחה או מודלי תגמולים, נובעת ישירות מהיכולת של המודל ללמוד דינמיקה חזותית כללית רובסטית ביחס למשימה מתוך מרחב לטנטי מאומן מראש.

היכולת של DINO-WM להכליל לקונפיגורציות חדשות של סביבות – כמו מבוכי אקראיים או אובייקטים בצורות מגוונות – מדגישה עוד יותר את חידושו. ההכללה הזו נובעת מהלמידה האפקטיבית של מושגים ודינמיקות כלליים בתוך ייצוגי הפאצ'ים הלטנטיים המאומנים מראש, מה שמפחית את התלות בביצועים שבריריים הנובעים מאימון על נתונים תלויי-משימה.

לסיכום, החידוש של DINO-WM טמון בשילוב בין מקודד פאצ'ים קפוא ומאומן מראש מ־DINOv2 לצורך מודל התצפית, מודל מעברים מבוסס ViT ברמת פריים הפועל כולו במרחב הלטנטי, והניתוק המוחלט בין הדינמיקה הפנימית לשחזור הפיקסלים. הארכיטקטורה הזו מאפשרת למידה של דינמיקה חזותית כללית, רובסטית ביחס למשימות רבות, מנתוני אופליין בלבד ומובילה לתכנון zero-shot אפקטיבי והכללה חזקה בסביבות מגוונות.

https://arxiv.org/abs/2411.04983