# Review 160: [Short] Vision Transformers Need Registers https://arxiv.org/abs/2309.16588

**Paper: https://arxiv.org/abs/2309.16588v2**

ממש אהבתי את המאמר זה אבל לא בגלל שהוא הציע איזה רעיון מהפכני (הרעיון די נחמד אבל לא איזו פצצה). הסיבה לכך היא שהמאמר הזה הדגיש שוב את העובדה כמה מעט אנחנו מבינים מודלי ענק בעל ביליוני פרמטרים

היום ב- #shorthebrewpapereviews אנו סוקרים מאמר שבמילים ממש פשוטות מצא שייצוג הדאטה שהטרנספורמרים הוויזואליים(כמו DINOv2) מפיקים מכילים דאטה מיותר שלא תורם לביצועי המודל יותר מדי. אזכיר שהמטרה העיקרית של מודלי הענק האלו היא לבנות ייצוג של דאטה המכיל את הפיצ'רים המהותיים ביותר שלו.

כלומר הטרנספורמרים הויזאליים לא מצליחים לקודד את המידע בצורה המיטבית יש חלקים מיותרים בייצוג הזה. איך המחברים בכלל הגיעו לזה? הם שמו לב שיש פאצ'ים בתמונה שנורמה של ייצוגם (מהשכבה האחרונה) היא גדולה באופן אנומלי יחסית לייצוגי הפאצ'ים האחרים.

המחברים גם שמו לב שייצוגים של פאצ'ים חריגים אלו מאוד דומים לייצוגי הפאצ'ים הסמוכים (מבחינת מרחק קוסיין). בנוסף יכולת של ייצוגי פאצ'ים אנומליים אלו להציג את מיקום הפאץ' בתמונה היא משמעותית יותר נמוכה מהפאצ'ם הרגילים (אימנו מודל לזיהוי המיקום). הם עשו עוד בדיקות נוספות ששיכנעו אותם שייצוג הפאצ'ים האלו לא משפר את איכות המודל.

אז מה הם עשו? משהו די אלגנטי (זה לא רעיון חדש כי כבר עשו זאת לפני כמה שנים במאמר על מודלי שפה). אז הם הוסיפו כמה טוקנים (אחרי טוקן cls) שמטרתם היא להכיל מידע לא רלוונטי. ייצוגי טוקנים אלה פשוט נזרקים ולא משמשים לא לאימון ולא לאינפרנס. וזה אכן משפר את ביצועי המודל בכמה משימות.