# Review 96: [Short] Multiscale Vision Transformers (MViT): A hierarchical architecture for representing image and video information (Meta)

**Paper: https://arxiv.org/abs/2104.11227v1**

בגדול MViT זה טרנספורמר היררכי לבניית ייצוג של תמונה ווידאו. בדרך כלל מודלים המבוססים על ארכיטקטורת הטרנספורמרים משתמשת ברזולוציה קבועה ובאותו מימד של ייצוג הטוקנים כדי לשערך את הקשרים בין טוקנים (כגון פאצ'ים של תמונה או פריימים של וידאו) שונים בתמונה. לעומת זאת MViT מורכב כמה שכבות של טרנספורמרים שכל אחד מהם פועל ברזולוציה שונה ומימד שונה של ייצוג טוקנים. בדומה לרשתות קונבולוציה ככל שמתעמקים יותר לתוך השכבות של MViT מספר טוקנים (רזולוציה) יורד כאשר מימד הייצוג שלהם עולה. כמובן שרשת הטרנספורמרים ההיררכית הזו לוקחת בחשבון גם את המימד של הזמן בנושא למימד המרחבי (של התמונה).

המאמר טוען שניתן לאמן את MViT מאפס והוא מצטיין במשימות זיהוי בוידאו וגם בלוקליזציה של פעולות בוידאו.

פרוייקט: project link

מאמר: https://arxiv.org/abs/2104.11227