# Review 113: [Short] UnIVAL: Unified Model for Image, Video, Audio and Language Tasks, 01.08.23

**Paper: https://arxiv.org/abs/2307.16184v2**

https://huggingface.co/papers/2307.16184

מכירים את המודלים המולטימודליים כמו ImageBind של מטה והמודל הטרי Med-PaLM שיכולים לעבוד עם סוגים שונים של דאטה? מודלים אלו מכילים עשרות מיליארדים של פרמטרים והם גם מאומנים על דאטהסטים עצומים. האם ניתן לאמן מודל מולטימודלי יחסית קטן שלא דורש כמויות עצומות של דאטה לאימון?

היום ב- #shorthebrewpapereviews סוקרים מאמר שמחבריו טוענים שהם הצליחו לאמן מודלי מולטימודלי קטן יחסי היודע לשפה טבעית, תמונות, וידאו ואודיו. איך הם עשו זאת? הם אימנו (pretraining) את המודל שלהם קודם כל על משימות פשוטות יותר (כמו מידול שפה שזה יותר פשוט ממשימות מולטימודליות) ואז המשיכו לאמן אותו על משימות המערבות טקסט ותמונות.

לאחר מכן הם הלכו ונתנו למודל משימות קשות יותר דרך הוספה של מודליטי נוסף לדאטה מהשלב הקודם. הם קראו לשיטה הזו למידה curriculum מולטימודלי. המחברים טוענים שככה המודל לומד לייצג כל מודליטי במרחב השיכון המשותף. כלומר מתחילים מאמנים כמה אפוקים על משימה פשוטה וכל כמה אפוקים מוספים סוג דאטה נוסף ו״מסבכים״ את המשימה. דרך אגב שמודל כזה מצריך טוקניזציה שתתרגם סוגי דאטה שונים לאותו מרחב. התוצאות של המודל מציג תוצאות לא רעות ביחס למודלים גדולים הרבה יותר ממנו.