# Review 116: [Short] Multimodal Neurons in Pretrained Text-Only Transformers, 05.08.23 https://huggingface.co/papers/2308.01544

**Paper: https://arxiv.org/abs/2308.01544v2**

תופעה מעניינת מתרחשת כאשר ממפים תמונות וטקטסט לאותו מרחב וקטורי.מחברים אנקודר לתמונות עם מודל שפה מוקפא, מחברים ביניהם עם שכבת לינארית אחת בלבד ואז מאמנים את האנקודר את השכבה הלינארית לחזות כותרת של תמונה. מסתבר כי בתוך מודל שפה יש נוירונים שנדלקים חזק עבור פאץ של תמונה ועבור תיאורה הטקסטואלי.  
  
היום ב-#shorthebrewpapereviews סוקרים מאמר המראה שלמרות שמודלי השפה לא מאומן למשימה זו עדיין יש בו נויורים ה״מסמנים״ בו זמנית קונספטים (אובייקטים) בשני העולמות: הויזואלי והטקסטואלי. נגיד אם יש בתמונה כלב אז יהיה נוירון בשכבה מסוימת של מודל שפה שמי שהדליק אותו הכי חזק הם הפאצ'ים שבהם מופיע הכלב ובאותו הזמן הם נדלקים חזק גם עבור המילים שמסמנים כלב (dog, doggy וכדומה). המחברים קוראים לנוירונים אלו נוירונים מולטימודליים. עכשיו בואו נבין מה זה אומר להדליק נוירון כתלות במשהו (נגיד פאץ או מילה). המאמר מגדיר ציון ייחוס (attribution score) של נוירון עבור מילה(טוקן) נתונה c באופן הבא.   
  
נניח שמודל חוזה מילה c עם ערך הלוגית הגבוה ביותר בין כל הטוקנים. אז ציון הייחוס מוגדר בתור מכפלה של ערך הנוירון עצמו ובין הנגזרת של ערך הלוגיט ביחס לנוירון הזה. כלומר נוירון שערכו גבוה וגם ערך הלוגיט הוא תלוי (נגזרת גבוהה) בנוירון הזה מקבל ציון ייחוס גבוה (= נדלק חזק). ציון יחס של נוירון בפאץ מסוים מוגדר בצורה דומה אך הנגזרת הפעם מחושבת לנוירון זה ביחס לייצוג של פאץ.