# Review 144: Transformers as Support Vector Machines, 12.09.23 https://arxiv.org/abs/2308.16898.pdf

**Paper: https://arxiv.org/abs/1610.09152v2**

המאמר הזה שונה מרוב המאמרים על למידה עמוקה. הוא לא מנסה לשפר איזה מודל או להכות ביצועי SOTA במשימה זו או אחרת. הוא מנסה להסביר סודות של ארכיטקטורת הטרנספורמרים שעליה מבוססים רוב המודלים החזקים של ימינו.   
  
אז היום ב-#shorthebrewpapereviews אנו סוקרים מאמר שמנסה לחקור מה למעשה מאופטם בטרנספורמרים כאשר אנחנו מאמנים אותו דרך מזעור של פונקציית ההפסד (נגיד על הטוקן הבא). המחברים מצאו בהקשר המדובר בעיית אופטימיזציה של הטרנספורמרים שקולה למודל מסוג Support Vector Machine או SVM בדרך מאוד מעניינת. קודם כל מה SVM? זו טכניקה שהומצאה בשנות ה-80 ובמהותה היא מנסה למצוא ישר (או עקומה מורכבת יותר כאשר מדובר בקרנל SVM) המרחיק כמה   
  
שיותר בין הנקודות השייכות לקטגוריות שונות. אז מה משותף בין אופטימיזציה של הטרנספורמרים לבין SVM שלכאורה נראים דברים לא קשורים לחלוטין. נניח שיש לנו בעיית SVM (דואלית למי שמכיר אבל זה פרט פחות חשוב). מטרתה למצוא מטריצות Q ו- K בעלות נורמה גרעינית (סכום של ערכים עצמיים) מינימלית של W = QK עם ראנק קטן מ-m (מספר השורות ב-W) שמפרידות את הטוקן האופטימלי משאר הטוקנים (הם חקרו את חיזוי הטוקן הראשון).   
  
ההפרדה גם מתבצעת עם המכפלה של K ו- Q. אז מתברר שהפתרון של בעיה זו (מטריצות Q ו- K) הן פתרון של בעיית מזעור של בלוק טרנספורמר אחד (כולל MLP בסוף). כאן בעיה SVM לא קמורה. עוד משפט מעניין שהם הוכיחו שאם עושים רפרמטריזציה של בעיית אופטימיזציה של הטרנספורמים עם W = QK ומאפטמים את שכבת הטרנספורמר שלנו במונחי W זה שקול בפתרון בעיית SVM כאשר ממזערים את נורמת פרובניוס של W (ששוה לשורש מ- trace של מטריצה מוכפלת בשחלופה). בקיצור מאמר סופר מעניין ודי כבד מתמטית אבל עדיין שווה קריאה.