# Review 68: PATCHES ARE ALL YOU NEED?

**Paper: https://arxiv.org/abs/2201.09792v1**

אתם אולי מכירים את חולשתי למאמרים מכילים ביטוי ״all you need". מרגיש מחובתי לסקור קצרות כל מאמר כזה שעיני נתקלת בו. אציין שלא הייתי סוקר את המאמר הזה אחרת

הפעם במוקד נמצא לא אחר אלא פאץ'. לאור ההצלחה של הטרנספורמרים גם בדומיין הויזואלי (Visual Transformer ודומיו). אזכיר שהקלט לטרנספורמר ויזואלי הינו פאצ'ים של תמונות. כלומר מחלקים תמונה לפאצ'ים, מייצגים כל אחד באמצעות וקטור (בדרך כלל פשוט משטחים את מטריצת הפיקסלים בפאץ' ומכפילים במטריצה).

אז המאמר שואל את השאלה הבאה: האם ההצלחה של הטרנספורמרים הויזואליים נובעת מעוצמתו של ארכיטקטורת הטרנספורמר או שזה פשוט עניין של בנייה נכונה של קלט לרשת, כלומר הפאצ'ים. בשביל לבדוק את העניין הזה המחברים בנו רשת ממש ממש פשוטה (אפילו יותר פשוטה מ- MLP-Mixer המפורסם) כשהקלט שלה הוא הייצוגים של פאצ'ים. הרשת בנויה בצורה מאוד פשוטה מבלוקים residual שכל אחד מהם מכיל קונבולוציה depthwise ו- pointwise אחד אחרי השני. בסוף יש שכבת pooling ו-FC. וזהה זה.

עם ארכיטקטורה מאוד פשוטה זו הם הגיעו לביצועים ברי השוואה עם קונפיגורציה מסוימת של ViT (הם גם השוו ביצועים עם איזה סוג של טרנספורמר, הנקרא DeIT). הם גם עשו השוואה על עם מודלים פשוטים כמו ResNet152 ו- ResMLP.

אבל

הבדיקה נעשתה על שתי משימות בלבד (סיווג על ImageNet ו- CIFAR10). הם לא ביצועי שום אימון self-supervised והוכיחו שהייצוגים שהארכיטקטורה שלהם, שקיבלה שם ConvMixer, יודעת להפיק, חזקים יותר. אלא רק אימון על דאטהסטים מתויגים (תקנו אותי אם פספסתי משהו).

בקיצור מאמר די מאכזב, השם מאוד באזזי אך לצערי לא הולם את התוכן…

אשמח לדעת מה דעתכם….

למאמר: https://openreview.net/forum?id=TVHS5Y4dNvM