מה הם Vision Transformers וכיצד הם חשובים ללמידה למטרות כלליות?

בעולם הלמידה אנו ניצבים בפני בעיה, ישנם המון מודלים שמתמודדים עם בעיות למידה שונות ומגוונות אך הבעיה שהמודלים מתמודדים עם בעיות ספציפיות.

מודלים אלו דורשים בנייה ושינויים ייחודיים כדי להתמודד עם בעיה מעט שונה ממה שהם פותרים בדרך כלל.

לכן, מודלים של למידה למטרות כלליות מעניינים אותנו שכן לא נצטרך לשנות אותם דרסטית לפני שניגש לפתרון של בעיה שונה במעט.

הקונספט של Transformers נולד עבור עיבוד שפה טבעית שם המשימה היא להבין טקסט ולהסיק לגביו מסקנות(למשל השלמת משפט, מציאת מילה חסרה ועוד).

וכעת נדבר על Vision Transformers:
המושג Vision Transformer הוא הרחבה של המושג המקורי של Transformer,

ViT בעצם משתמשת בשיטות שונות של Token ושל הטמעה (Embedded) כדי להתמודד עם שינויים בתמונה.

איך זה בעצם עובד?

תמונת קלט מפוצלת לקבוצה של תיקוני תמונה הנקראים אסימונים חזותיים(Visual Tokens) אסימונים אלו מוטמעים בקבוצה של וקטורים מקודדים בעלי מימד קבוע, המיקום של שינוי בתמונה מוטבע יחד עם הווקטור המקודד ומוזן לרשת שזהה לזו האחראית על עיבוד קלט הטקסט.

ישנם מספר בלוקים במקודד ViT וכל בלוק מורכב משלושה מרכיבי עיבוד עיקריים:

1. נורמה שכבתית.
2. MSP (Multi-head attention network)
3. MLP (Multi-layer perceptrons)

נורמה שכבתית שומרת על תהליך האימון ונותן למודל להתאים את עצמו לווריאציות בין תמונות האימון.

MSP היא רשת האחראית ליצירת מפות קשב מהאסימונים הוויזואליים המוטבעים. מפות אלו עוזרות להתמקד ברשת באזורים החשובים ביותר בתמונה כגון אובייקטים.

הרעיון של מפות קשב זהה לזה שנמצא בספרות הראייה הממוחשבת המסורתית (למשל מפות בולטות ואלפא-מטינג).

ה-MLP היא רשת סיווג דו-שכבתית עם GELU (יחידת חישוב ליניארית עם שגיאה גאוסית).

בלוק MLP הסופי, המכונה גם ראש MLP משמש כפלט של ה-Transformer.

יישום של softmax על פלט זה יכול לסווג תוויות (כלומר אם היישום הוא סיווג תמונה).

היישומים של ViT בעצם מקיפים את כל היבטי הראייה החל מסיווג תמונה דרך יצירת תמונה לטקסט או יצירת טקסט עבור תמונה, חשיבה חזותית ולמידה אסוציאטיבית.