**ViT**

חלה התקדמות משמעותית בתחום הבינה המלאכותית בשנים האחרונות ועקב כך, העניין במודלים גנריים של למידה למטרות כלליות (מוגדרות לכל משימה) גובר. הבלוג דן באחד מסוגי הדגמים אלו - Transformers.

**1. Transformers של הטקסט**

הרעיון של Transformers מקורו ביישומי עיבוד שפה טבעית (NLP), לפיהם המשימה היא להבין את הטקסט ולהסיק מסקנות משמעותיות. דגמי Transformers השיגו ביצועים גבוהים והפכו לסטנדרט בתחום ה-NLP בשל הפשטות וההכללה שלהם. ביישומי עיבוד טקסט, הנתונים מורכבים מרצף גדול של מילים המגיעות ממערכת קבועה של אוצר מילים. כפי שאנו יכולים לראות שהארכיטקטורה של רשת Transformers היא גנרית מכיוון שפלט המקודד אינו בנוי עבור משימה מסוימת (למשל סיווג) אלא הוא מספק קידוד כללי שניתן להשתמש בו עבור מספר רב של יישומים על ידי הוספת ראש MLP מתאים ולכן Transformers שימושיים בלמידה של העברה ומבטיחים להשגת מטרת למידה כללית.

**2. Vision Transformers (ViT)**

המושג Vision Transformer (ViT) הוא הרחבה של המושג המקורי של Transformer. ViT הוא היישום של Transformer בתחום התמונה עם שימוש בשיטות שונות על מנת לטפל במודל הנתונים השונה (תמונות). תמונת קלט מפוצלת לקבוצות ונקראים token חזותיים. הם מוטמעים בקבוצה של וקטורים מקודדים בעלי מימד קבוע ונשלחים לרשת כקלט.

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

1. ה-Layer Norm שומרת על תהליך האימון ונותן למודל להתאים את עצמו לווריאציות בין תמונות האימון.
2. MSP היא רשת האחראית ליצירת מפות attition מה-token הוויזואליים המוטבעים. תפקידן לעזור להתמקד ברשת באזורים החשובים ביותר בתמונה כגון אובייקט(ים).
3. ה-MLP היא רשת סיווג דו-שכבתית עם GELU (Gaussian Error Linear Unit) בסוף. בלוק זה הוא הפלט של ה-Transformer. יישום של softmax על פלט זה יכול לספק תוויות סיווג.

**3. יישומים**

בשל האופי הגנרי שלהם, היישומים של ViTs מקיפים כמעט את כל ההיבטים החזותיים, כגון: סיווג תמונה, יצירת תמונה לטקסט/טקסט לתמונה, חשיבה חזותית, למידת אסוציאציות ולמידה רב-מודאלית.

a. מתמונה -> לתווית

המשימה של סיווג תמונה היא הבעיה הנפוצה ביותר בתחום. ViTs אינם מייצרים ביצועים דומים לרשת CNN במערך נתונים קטן עד בינוני (כי CNN מקודדים את המידע המקומי בתמונה בצורה יעילה מאוד), אך הם השיגו ביצועים טובים יותר על מערכי נתונים גדולים מאוד.

b. מתמונה -> למשפט

ניתן להשיג צורה מתקדמת יותר של סיווג תמונה על ידי יצירת כיתוב המתאר את תוכן התמונה במקום תווית של מילה אחת. זה הפך אפשרי עם השימוש ב-ViTs. ב - ViTsלומדים ייצוג כללי של שיטת נתונים נתונה במקום קבוצה גסה של תוויות, לכן, ניתן ליצור טקסט תיאורי לתמונה נתונה.

c. תמונה <-> קטע טקסט

יישום מתקדם יותר של ViTs הוא למידה של אסוציאציה הקיימת בין תמונה לפיסת טקסט. זהו היישום שבו נדרשת היכולת האמיתית של הכללה מכיוון שהיא דורשת ייצוג מופשט הן של הטקסט והן של התמונה. ניתן להשיג זאת על ידי אימון שני מקודדי טרנספורמציה נפרדים עבור קטע טקסט והתמונה. לאחר מכן ניתן להשוות את התמונה המקודדת כמו גם את תכונות הטקסט עבור הדמיון שלהם על ידי בניית מטריצת דמיון קוסינוס. יישום רלוונטי עבור מודל שנאי כזה מוצע ב-CLIP (הדרכה מקדימה של שפה-תמונה נגדית).