**به نام خدا**

\*\*Vision Transformers (ViT)\*\* یک مدل یادگیری عمیق است که از معماری \*\*ترانسفورمر\*\* (Transformer) برای وظایف پردازش تصویر استفاده می‌کند. این مدل ابتدا در مقاله‌ای توسط گوگل در سال 2020 با عنوان \*"An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale"\* معرفی شد. برخلاف شبکه‌های کانولوشنی (CNN) که به طور سنتی برای پردازش تصویر استفاده می‌شوند، ViT از مکانیزم توجه (**Attention**) ترانسفورمرها بهره می‌برد که ابتدا برای پردازش زبان طبیعی (NLP) توسعه یافته بود.

**### نحوه کار Vision Transformers:**

1. \*\*تقسیم تصویر به پچ‌ها (Patches):\*\*

تصویر ورودی به قطعات کوچک‌تر (مثلاً 16\*16 پیکسل) تقسیم می‌شود. هر پچ به یک بردار تبدیل شده و به عنوان یک توکن (مشابه کلمات در NLP) در نظر گرفته می‌شود.

2. \*\*تبدیل به Embedding:\*\*

هر پچ به یک بردار ویژگی (Embedding) تبدیل می‌شود. یک توکن ویژه به نام \*\*[CLS]\*\* (مشابه ترانسفورمرهای NLP) ممکن است اضافه شود تا اطلاعات کلی تصویر را جمع‌آوری کند.

3. \*\*اضافه کردن اطلاعات موقعیتی (Positional Encoding):\*\*

برای حفظ اطلاعات مکانی پچ‌ها در تصویر، کدگذاری موقعیتی به Embeddingها اضافه می‌شود، زیرا ترانسفورمرها به طور ذاتی ترتیب توکن‌ها را درک نمی‌کنند.

4. \*\*پردازش با ترانسفورمر:\*\*

پچ‌های تبدیل‌شده به ترانسفورمر وارد می‌شوند، که از لایه‌های \*\*توجه چندسر (Multi-Head Attention)\*\* و \*\*شبکه‌های پیش‌خور (Feed-Forward Networks)\*\* تشکیل شده است. این لایه‌ها روابط بین پچ‌ها را مدل‌سازی می‌کنند.

5. \*\*خروجی:\*\*

خروجی ترانسفورمر (معمولاً بردار [CLS] یا میانگین خروجی پچ‌ها) برای وظایفی مثل طبقه‌بندی تصویر استفاده می‌شود.

**### مزایا:**

- \*\*مقیاس‌پذیری:\*\* ViT در مجموعه داده‌های بزرگ (مانند ImageNet یا JFT-300M) عملکرد بسیار خوبی دارد.

- \*\*انعطاف‌پذیری:\*\* به دلیل استفاده از توجه، می‌تواند روابط پیچیده و غیرمحلی بین بخش‌های تصویر را مدل کند.

- \*\*یکپارچگی با NLP:\*\* معماری مشابه ترانسفورمرها امکان استفاده از تکنیک‌های مشترک بین تصویر و متن را فراهم می‌کند.

**### معایب:**

- \*\*نیاز به داده زیاد:\*\* ViT برای عملکرد خوب نیاز به پیش‌آموزش روی مجموعه داده‌های بزرگ دارد، برخلاف CNNها که با داده کمتر هم کار می‌کنند.

- \*\*مصرف محاسباتی بالا:\*\* پردازش پچ‌ها و مکانیزم توجه می‌تواند پرهزینه باشد.

- \*\*عملکرد ضعیف‌تر در مقیاس کوچک:\*\* در مجموعه داده‌های کوچک‌تر، معمولاً از CNNها عقب می‌ماند.

**### کاربردها:**

- \*\*طبقه‌بندی تصویر:\*\* مانند شناسایی اشیا در تصاویر.

- \*\*تشخیص اشیا (Object Detection):\*\* با ترکیب با مدل‌های دیگر.

- \*\*تقسیم‌بندی تصویر (Segmentation):\*\* با معماری‌های خاص.

- \*\*کاربردهای چندوجهی:\*\* ترکیب تصویر و متن در مدل‌هایی مثل CLIP.

**### تفاوت با CNN:**

- CNNها از فیلترهای محلی (کانولوشنی) استفاده می‌کنند، در حالی که ViT از توجه جهانی برای مدل‌سازی روابط بین پچ‌ها بهره می‌برد.

- ViT نیازی به پیش‌فرض‌های قوی در مورد ساختار تصویر (مانند محلی بودن) ندارد.

Vision Transformers به دلیل عملکرد قوی و انعطاف‌پذیری، در سال‌های اخیر به یکی از معماری‌های کلیدی در بینایی کامپیوتری تبدیل شده‌اند و در مدل‌های پیشرفته‌تر مانند Swin Transformer یا DeiT توسعه یافته‌اند.

اگر سوال خاصی در مورد ViT دارید یا نیاز به توضیحات بیشتر دارید، بفرمایید!