فهرست مطالب

فهرست شکل‌ها

فصل 1: مقدمه

1-1 زمینه

یادگیری عمیق انقلابی در زمینه بینایی رایانه ایجاد کرده است و پیشرفت قابل‌توجهی را در کارهایی مانند طبقه‌بندی تصویر[1]، تشخیص اشیا[2] و تقسیم‌بندی معنایی[3] ممکن می‌سازد. در دهه گذشته، شبکه‌های عصبی کانولوشنال[4] (CNN) به معماری غالب تبدیل شده‌اند و به نتایج پیشرفته‌ای در معیارهای مختلف دست یافته‌اند. توانایی‌CNNها برای یادگیری خودکار ویژگی‌های سلسله مراتبی[5] از داده‌های پیکسل خام در موفقیت آن‌ها بسیار موثر بوده است. از معماری‌های اولیه مانند LeNet[1] گرفته تا مدل‌های پیچیده‌تر مانند ResNet[2] و EfficientNet[3]، CNNها به طور مداوم مرزهای آن‌چه را که در بینایی کامپیوتر ممکن است جابجا کرده‌اند.

1-2 محدودیت‌های شبکه‌های عصبی کانولوشن(CNN)

با وجود عملکرد چشمگیر، CNNها محدودیت‌های ذاتی دارند. میدان‌های دریافتی[6] محلی آن‌ها، که با اندازه هسته[7]های کانولوشنی تعیین می‌شوند، گرفتن وابستگی‌های دوربرد[8] در یک تصویر را چالش‌برانگیز می‌کنند. در حالی که تکنیک‌هایی مانند افزایش اندازه هسته یا انباشتن چندین لایه کانولوشنال می‌توانند تا حدی این مشکل را برطرف کنند، آن‌ها اغلب به قیمت افزایش پیچیدگی محاسباتی تمام می‌شوند. علاوه بر این، CNNها ذاتاً معادل ترجمه[9] هستند، به این معنی که به ترتیب فضایی ویژگی‌ها حساس هستند. در حالی که این می‌تواند در برخی موارد سودمند باشد، اما در هنگام برخورد با صحنه‌های پیچیده یا اشیاء با جهت‌گیری‌های مختلف نیز می‌تواند یک محدودیت باشد. در نهایت، طراحی معماری‌های CNN اغلب به شدت بر مهندسی دستی و انتخاب‌های معماری متکی است، که می‌تواند زمان‌بر باشد و ممکن است همیشه به نتایج مطلوب منجر نشود.

[1] Image Classification
[2] Object Detection
[3] Semantic Segmentation
[4] convolutional neural networks
[5] hierarchical features
[6] Receptive fileds
[7] kernel
[8] Long range
[9] translation-equivariant

1-3 معماری ترانسفورمر: پیشرفتی در پردازش زبان طبیعی

به موازات پیشرفت در بینایی کامپیوتر، زمینه پردازش زبان طبیعی[1] (NLP) با معرفی معماری ترانسفورمر[2] شاهد پیشرفت چشمگیری بود. بر اساس مکانیسم توجه[3]، ترانسفورمرها در گرفتن وابستگی‌های دوربرد در داده‌های متوالی، مانند متن، عالی هستند. برخلاف شبکه‌های عصبی بازگشتی[4] (RNN) که اطلاعات را به صورت متوالی پردازش می‌کنند، ترانسفورمرها می‌توانند تمام توالی‌ها را به صورت موازی پردازش کنند که منجر به بهبود قابل توجهی در سرعت و عملکرد آموزش می‌شود. مکانیسم توجه[5] به مدل اجازه می‌دهد تا اهمیت بخش‌های مختلف توالی ورودی را هنگام انجام پیش‌بینی‌ها بسنجد و آن را قادر می‌سازد تا به طور مؤثر اطلاعات متنی را بگیرد.

1-4 پل زدن بین NLP وComputer Vision

موفقیت ترانسفورمرها در NLP محققان را برانگیخت تا کاربرد آن‌ها در بینایی کامپیوتر را بررسی کنند. این منجر به توسعه ترانسفورمرهای بینایی شد که معماری ترانسفورمر را برای پردازش تصویر تطبیق می‌دهد. ترانسفورمرهای بینایی یک تصویر را به‌عنوان دنباله‌ای از وصله‌ها[6] در نظر می‌گیرند که سپس به یک رمزگذار[7] ترانسفورمر وارد می‌شوند. با استفاده از مکانیسم توجه، ترانسفورمر بینایی می‌تواند روابط سراسری بین بخش‌های مختلف تصویر را ثبت کند و بر محدودیت‌های CNN در گرفتن وابستگی‌های دوربرد غلبه کند. این رویکرد نتایج قابل توجهی را نشان داده است و عملکرد رقابتی یا حتی برتر را در مقایسه باCNNهای پیشرفته در کارهای مختلف تشخیص تصویر به دست آورده است. ترانسفورمرهای بینایی نشان‌دهنده یک تغییر پارادایم در بینایی کامپیوتر است که قدرت معماری ترانسفورمر را فراتر از قلمرو پردازش زبان طبیعی نشان می‌دهد.

1-5 ساختار و نمای کلی گزارش

این گزارش یک نمای کلی از ترانسفورمرهای بینایی ارائه می‌دهد. ما با مرور مفاهیم اساسی CNN و ترانسفورمر شروع می‌کنیم و نقاط قوت و ضعف آن‌ها را برجسته می‌کنیم. سپس به معماری ترانسفورمرهای بینایی می‌پردازیم و اجزای اصلی آن و نحوه کار آن‌ها را با هم توضیح می‌دهیم. بخش‌های بعدی انواع مختلف و پیشرفت‌های ترانسفورمرهای بینایی، کاربردهای متنوع آن‌ها، روش‌های آموزشی و مقایسه دقیق با CNNها را بررسی می‌کنند. در نهایت، مسیرهای تحقیقاتی آینده و تأثیر بالقوه ترانسفورمرهای بینایی بر آینده بینایی کامپیوتر را مورد بحث قرار

[1] Natural language proccesing
[2] Transformer
[3] Attention mechanism
[4] recurrent neural network
[5] Attention mechanism
[6] patch
[7] Encoder

می‌دهیم. هدف این گزارش ارائه یک درک کامل از ترانسفورمرهای بینایی، قابلیت‌های آن‌ها و نقش آن‌ها در شکل‌دهی چشم‌انداز یادگیری عمیق برای تجزیه و تحلیل تصویر است.

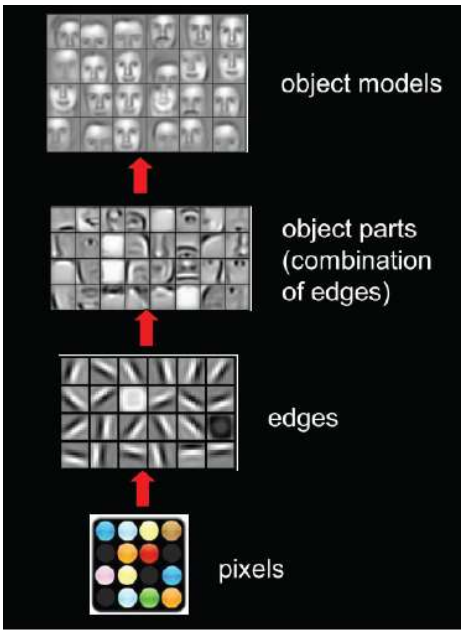
فصل 2: پیشینه و مبانی

2-1 شبکه‌های عصبی کانولوشن(CNN)

شبکه‌های عصبی کانولوشنال(CNN) برای سال‌های متمادی پیشروی بینایی کامپیوتر بوده‌اند و در طیف وسیعی از وظایف به موفقیت چشمگیری دست یافته‌اند. معماری آن‌ها به طور خاص برای بهره‌برداری از ساختار فضایی تصاویر طراحی شده است و آن‌ها را برای مشکلات مرتبط با تصویر مناسب می‌کند.

2-1-1 لایه‌های کانولوشنال

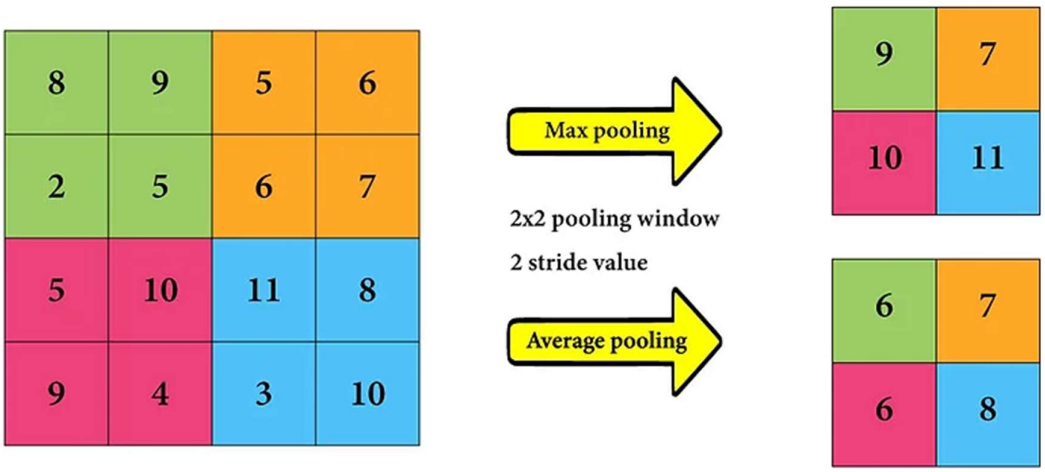
در قلب یک CNN لایه کانولوشنی نهفته است. این لایه از فیلترهای قابل یادگیری (یا هسته‌ها) برای لغزش در تصویر ورودی (یا نقشه ویژگی)، انجام ضرب عنصر و جمع‌بندی نتایج استفاده می‌کند. این عملیات پیچیدگی نامیده می‌شود. هر فیلتر الگوها یا ویژگی‌های خاصی مانند لبه‌ها، گوشه‌ها یا بافت‌ها[1] را در تصویر تشخیص می‌دهد. خروجی یک لایه کانولوشن یک نقشه ویژگی[2] است که هر مقدار نشان‌دهنده حضور و قدرت ویژگی مربوطه در یک مکان خاص در تصویر است. فیلترهای متعدد معمولاً در هر لایه کانولوشنی استفاده می‌شود که به شبکه اجازه می‌دهد مجموعه‌ای از ویژگی‌ها را بیاموزد.



شکل 2- 1 نمایش لایه‌های مختلف کانولوشن

[1] texture
[2] Feature map

لایه‌های ادغام[1] اغلب بعد از لایه‌های کانولوشن درج می‌شوند تا ابعاد فضایی نقشه‌های ویژگی را کاهش دهند و شبکه را نسبت به تغییرات کوچک در ورودی قوی‌تر کنند. عملیات ادغام متداول شامل ادغام حداکثر (که حداکثر مقدار را در یک منطقه محلی انتخاب می‌کند) و ادغام میانگین[2] (که مقدار متوسط را محاسبه می‌کند) است. ادغام به کاهش هزینه محاسباتی و حافظه مورد نیاز شبکه کمک می‌کند، در حالی که تغییرناپذیری آن را نسبت به جابه‌جایی‌ها و چرخش ورودی افزایش می‌دهد.



شکل 2- 2 تفاوت میان ادغام حداکثر و ادغام میانگین

2-1-2 توابع فعال سازی

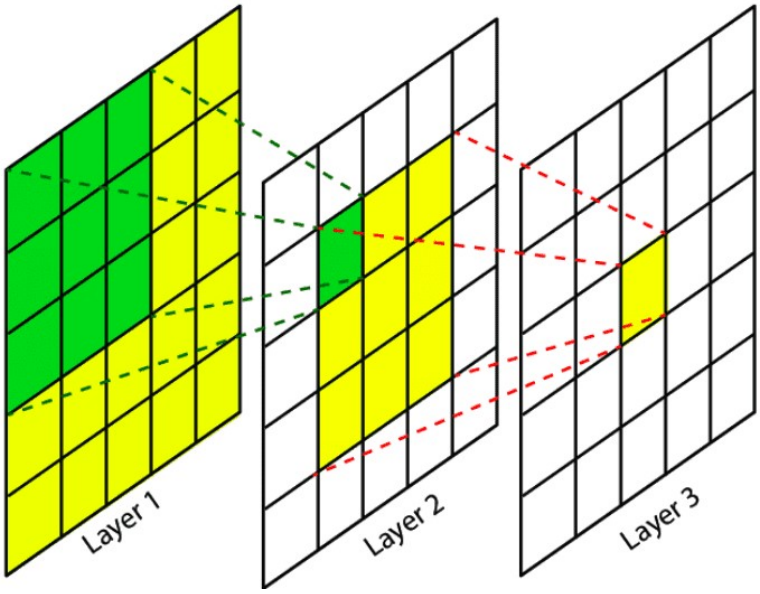
توابع فعال‌سازی[3] غیرخطی بودن[4] را به شبکه وارد می‌کند که برای یادگیری الگوهای پیچیده ضروری است. توابع فعال‌سازی رایج مورد استفاده در CNNها عبارتند از ReLU (واحد خطی اصلاح شده[5])، سیگموئید و tanh. و انواع آن (Leaky ReLU، Parametric ReLU) اغلب به دلیل عملکرد بهتر و آموزش سریع‌تر در مقایسه با سیگموئید و tanh ترجیح داده می‌شوند. تابع فعال‌سازی به صورت عنصری به خروجی لایه‌های کانولوشن یا ادغام اعمال می‌شود.

[1] Polling Layers
[2] Average Polling
[3] Activation functions
[4] Non linearity
[5] Rectified Linear Unit

اسم	ReLU (Rectified Linear Unit)	Sigmoid	Tanh (Hyperbolic Tangent)	Leaky ReLU	Parametric ReLU (PReLU)
فرمول	f(x) = max(0, x)	f(x) = 1 / (1 + exp(-x))	f(x) = (exp(x) - exp(-x)) / (exp(x) + exp(-x))	α is a small) 0 => αx if x ,0 < f(x) = x if x positive constant, e.g., 0.01)	αx if x ,0 < f(x) = x if x α is a learnable) 0 => parameter)
مزایا	از نظر محاسباتی کارآمد در بسیاری از موارد سریعتر از سیگموئید/tanh همگرا می‌شود مشکل ناپدید شدن گرادیان را برای ورودی‌های مثبت کاهش می‌دهد.	یک مقدار احتمال مانند بین 0 و 1 خروجی می‌دهد	خروجی صفر-مرکز (-1 تا 1) در برخی موارد می‌تواند سریعتر از سیگموید همگرا شود.	مشکل "ReLU Dying" را با اجازه دادن یک گرادیان کوچک و غیر صفر برای ورودی‌های منفی برطرف می‌کند. در برخی موارد می‌تواند عملکرد را در مقایسه با ReLU بهبود بخشد.	انعطاف پذیرتر از Leaky ReLU، همانطور که α در طول آموزش آموخته می‌شود. می‌تواند عملکرد را بیشتر بهبود بخشد.
معایب	مشکل "ReLU Dying": نورون‌ها می‌توانند غیرفعال شوند اگر ورودی آن‌ها به طور مداوم منفی باشد. نه در مرکز صفر.	مشکل ناپدید شدن گرادیان: گرادیان‌ها برای مقادیر ورودی شدید بسیار کوچک می‌شوند و مانع یادگیری می‌شوند. * صفر محور نیستند (خروجی‌ها همیشه مثبت هستند). از نظر محاسباتی گران تر از ReLU است.	مشکل ناپدید شدن گرادیان (اگرچه شدت کمتری نسبت به سیگموئید دارد). از نظر محاسباتی گرانتر از ReLU است.	هایپر پارامتر α باید تنظیم شود. در همه موارد به طور مداوم بهتر از ReLU نیست.	پیچیدگی مدل را افزایش می‌دهد (پارامتر قابل یادگیری).
محل استفاده	اکثر وظایف یادگیری عمیق، به ویژه آن‌هایی که مجموعه داده‌های بزرگی دارند. نقطه شروع خوب	مشکلات طبقه بندی باینری که در آن خروجی احتمال مورد نیاز است. در یادگیری عمیق مدرن برای سایر وظایف کمتر رایج است.	به صورت تاریخی استفاده شده است. در یادگیری عمیق مدرن در مقایسه با انواع ReLU کمتر رایج است، اما همچنان می‌تواند در سناریوهای خاص مانند شبکه‌های عصبی مکرر (RNN) مفید باشد.	جایگزین خوبی برای ReLU اگر "Dying ReLU" نگران کننده باشد.	جایگزین دیگری برای ReLU، به خصوص زمانی که تنظیم دقیق عملکرد مدل بسیار مهم است.

2-1-3 محدودیت در گرفتن وابستگی‌های دور از هم

CNNها علیرغم نقاط قوت خود دارای محدودیت‌هایی هستند، به ویژه در ثبت وابستگی‌های دوربرد در یک تصویر. میدان‌گیرنده[1] یک نورون در CNN با اندازه هسته‌های کانولوشن و تعداد لایه‌های کانولوشن محدود می‌شود. در حالی که انباشتن چندین لایه کانولوشن می‌تواند میدان دریافت را افزایش دهد، همچنان ثبت روابط بین بخش‌های دور از تصویر می‌تواند چالش برانگیز باشد. به عنوان مثال، درک زمینه یک شی در یک صحنه پیچیده اغلب مستلزم در نظر گرفتن روابط بین اشیاء مختلف است که ممکن است از هم دور باشند. CNNها تلاش می‌کنند تا چنین زمینه‌های سراسری را به طور مؤثر به تصویر بکشند. این محدودیت انگیزه کاوش در معماری‌های جایگزین مانند ترانسفورمرها را فراهم می‌کند که برای مدیریت وابستگی‌های دوربرد مجهزتر هستند. علاوه بر این، اگرچه تکنیک‌هایی مانند پیچش‌های متسع‌شده[2] یا اندازه‌های هسته بزرگ‌تر ممکن است کمک کنند، اما با افزایش هزینه‌های محاسباتی همراه هستند.



شکل 2- 3 میدان دریافتی در لایه‌های مختلف cnn

2-2 مدل ترانسفورمر (از NLP)

مدل ترانسفورمر که در اصل برای پردازش زبان طبیعی[3] (NLP) طراحی شده بود، انقلابی در مدل‌سازی توالی ایجاد کرد و اخیراً نفوذ قابل توجهی به بینایی کامپیوتری داشته است. قدرت اصلی آن در مکانیسم توجه نهفته است که به آن اجازه می‌دهد وابستگی‌های دوربرد را به طور موثر جذب کند.

[1] Receptive field
[2] dialated
[3] natural language processing

2-2-1 مکانیسم توجه (توجه به خود، توجه چند سر)

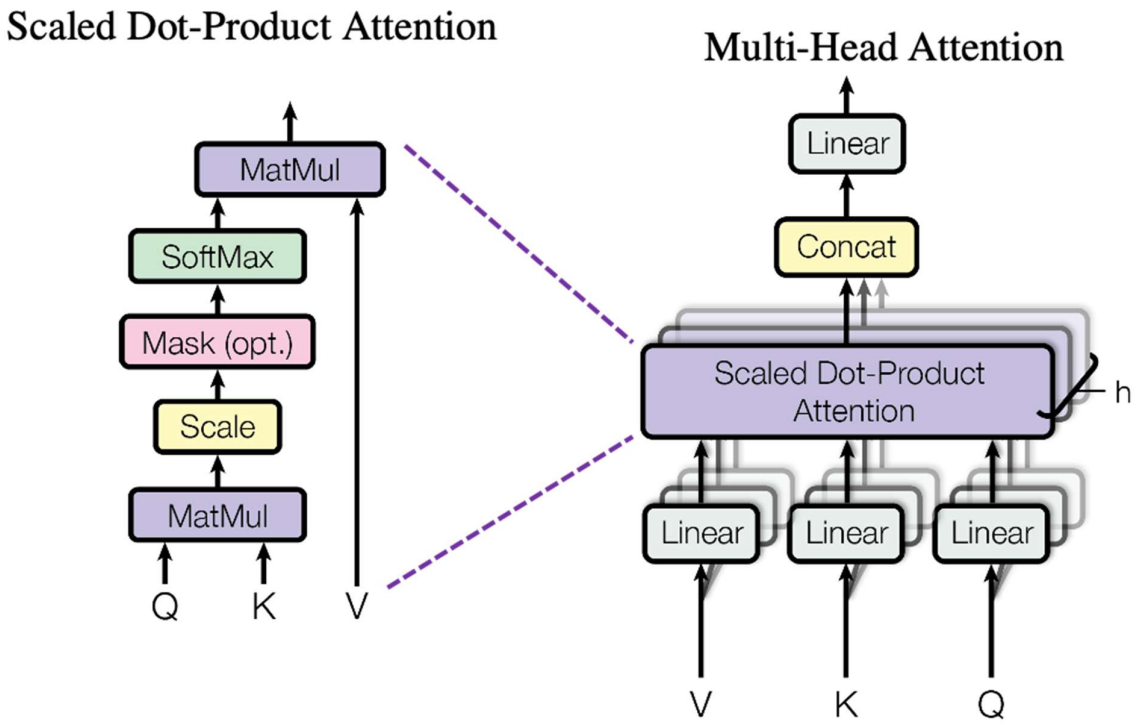
مکانیسم توجه قلب ترانسفورمر است. این به مدل اجازه می‌دهد تا اهمیت بخش‌های مختلف توالی ورودی را هنگام پردازش آن اندازه‌گیری کند. در توجه به خود، مکانیسم توجه موقعیت‌های مختلف یک دنباله واحد را برای محاسبه یک نمایش مرتبط می‌کند. با توجه به دنباله ای از بردارهای ورودی (به عنوان مثال، جاسازی کلمات در NLP، یا وصله‌های تصویر در آن‌هاترانسفورمرهای بینایی)، توجه به خود سه ماتریس را محاسبه می‌کند کوئری (Q)، کلید (K) و مقدار (V). این ماتریس‌ها از بردارهای ورودی از طریق تبدیل‌های خطی به دست می‌آیند.

وزن‌های توجه با گرفتن حاصل ضرب نقطه‌ای ماتریس‌های Query و Key، مقیاس‌گذاری با جذر ابعاد بردارهای کلید (برای جلوگیری از ناپدید شدن گرادیان)، و سپس اعمال تابع softmax محاسبه می‌شوند. این منجر به ماتریسی از وزن‌های توجه می‌شود، که در آن هر عنصر اهمیت یک موقعیت خاص در دنباله را هنگام توجه به موقعیت دیگری نشان می‌دهد.

سپس ماتریس ارزش در وزن توجه ضرب می‌شود تا خروجی نهایی تولید شود. این مجموع وزنی بردارهای مقدار با در نظر گرفتن روابط بین موقعیت‌های مختلف، دنباله ورودی را نشان می‌دهد.

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

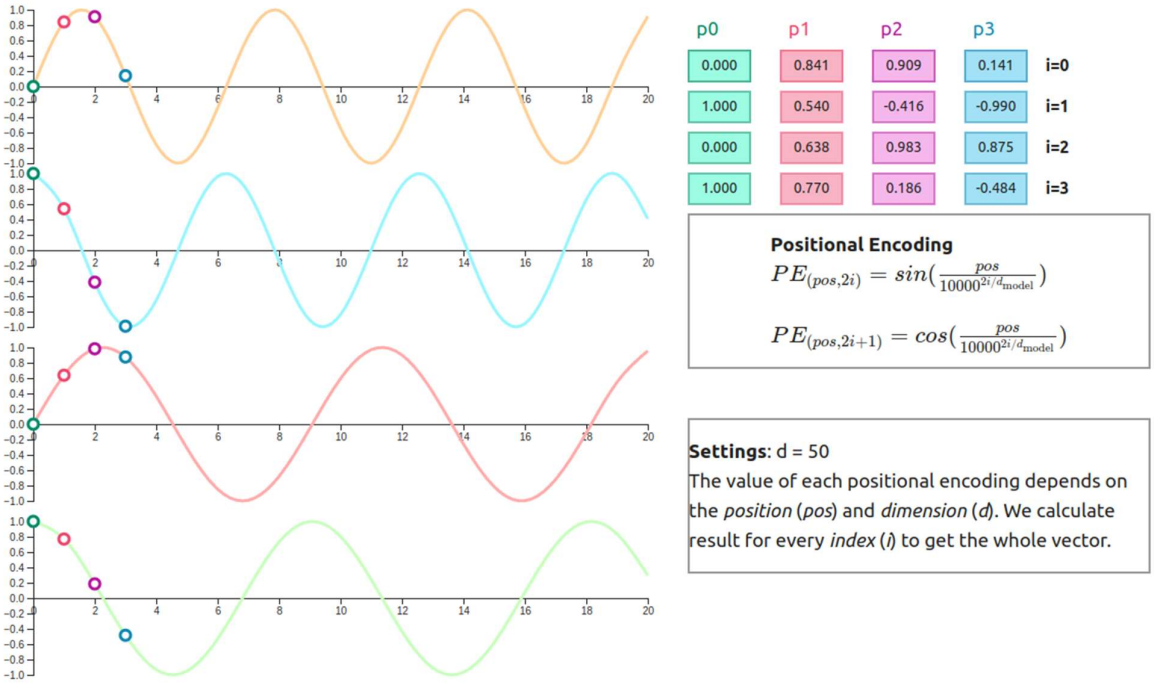
توجه چند سر این مفهوم را با استفاده از مجموعه‌های متعدد ماتریس‌های Q، K و V گسترش می‌دهد. هر «سر» الگوهای توجه متفاوتی را یاد می‌گیرد و به مدل اجازه می‌دهد تا طیف وسیع‌تری از روابط را در توالی ورودی ثبت کند. سپس خروجی‌های سرهای متعدد به هم متصل شده و به صورت خطی تبدیل می‌شوند تا خروجی نهایی تولید شود.



شکل 2- 4 ساختار مکانسیم توجه چندسر

2-2-2 رمزگذاری‌های موقعیتی

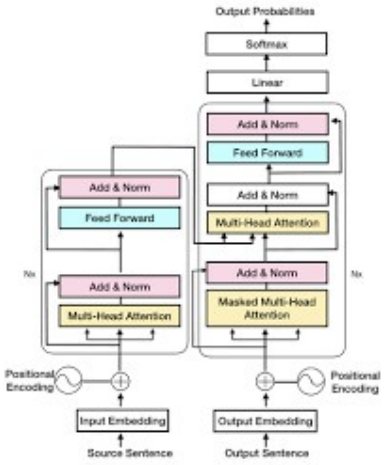
از آن‌جایی که ترانسفورمر ذاتاً اطلاعات متوالی را پردازش نمی‌کند (برخلاف RNNها)، به روشی برای رمزگذاری موقعیت هر عنصر در دنباله ورودی نیاز دارد. برای ارائه این اطلاعات، رمزگذاری‌های موقعیتی به جاسازی‌های ورودی اضافه می‌شوند. این کدگذاری‌ها معمولا بردارهای ثابتی هستند که به جاسازی‌های ورودی اضافه می‌شوند. آن‌ها را می‌توان آموخت یا از قبل تعریف کرد (به عنوان مثال، توابع سینوسی). رمزگذاری‌های موقعیتی به مدل اجازه می‌دهد تا بین عناصر در موقعیت‌های مختلف دنباله تمایز قائل شود.



شکل 2- 5 *Positional Encoding*

2-2-3 ساختار رمزگذار-رمزگشا

معماری اصلی ترانسفورمر برای کارهای ترتیب به دنباله (مانند ترجمه ماشینی) شامل رمزگذار[1] و رمزگشا[2] است. رمزگذار توالی ورودی را پردازش می‌کند، در حالی که رمزگشا توالی خروجی را تولید می‌کند. هر دو رمزگذار و رمزگشا از لایه‌های متعدد شبکه‌های خودتوجه و پیشخور تشکیل شده‌اند. برای طبقه‌بندی تصویر با آن‌هاترانسفورمرهای بینایی ، اغلب فقط بخش رمزگذار ترانسفورمر استفاده می‌شود.

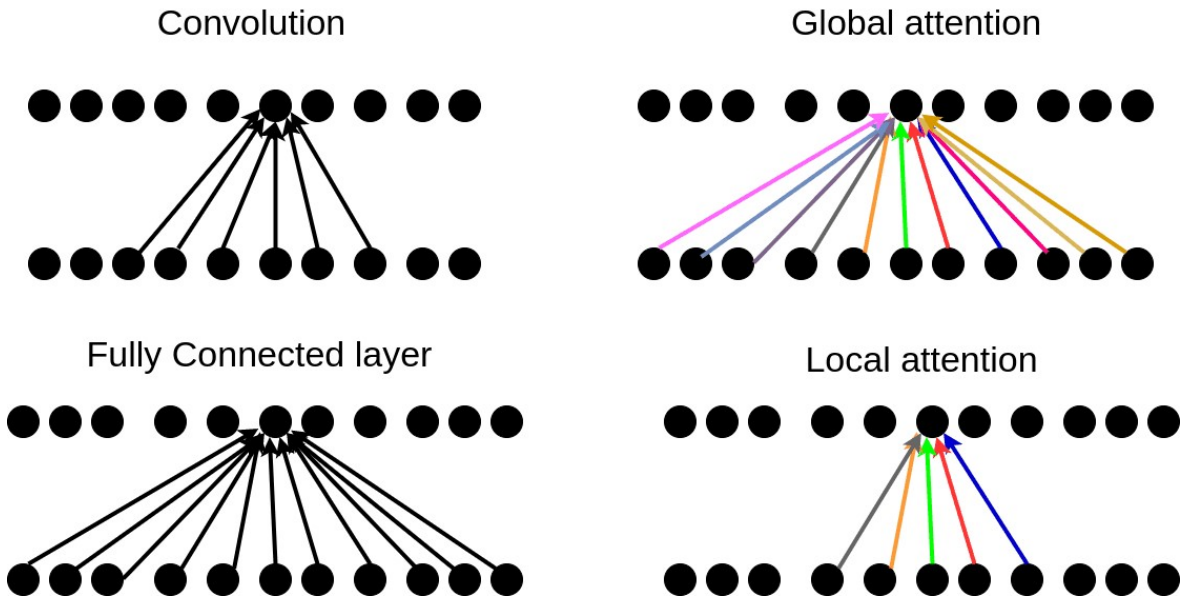


شکل 2- 6 ساختار رمزگذار-رمزگشا

Encoder[1]
Decoder[2]

2-2-4 مزایای مدیریت وابستگی‌های دوربرد

مزیت کلیدی ترانسفورمر در توانایی آن برای گرفتن وابستگی‌های دوربرد نهفته است. مکانیسم توجه به مدل اجازه می‌دهد تا بدون توجه به فاصله بین عناصر، به طور مستقیم به اطلاعات از هر بخشی از توالی ورودی دسترسی داشته باشد و آن را وزن کند. این در تضاد با RNNها است که اطلاعات را به صورت متوالی پردازش می‌کنند و به دلیل ناپدیدشدن[1] یا انفجار[2] گرادیان‌ها با توالی‌های طولانی مبارزه می‌کنند. قابلیت‌های پردازش موازی ترانسفورمر نیز آموزش آن را در مقایسه با RNN بسیار کارآمدتر می‌کند. این توانایی برای مدل‌سازی مؤثر روابط دوربرد، ترانسفورمر را به ابزاری قدرتمند برای کارهای مختلف مدل‌سازی توالی تبدیل کرده است، و اکنون با موفقیت برای بینایی رایانه سازگار شده است.



شکل 2- 7 تفاوت مکانسیم توجه در گرفتن اطلاعات و وابسگی‌های دوربرد

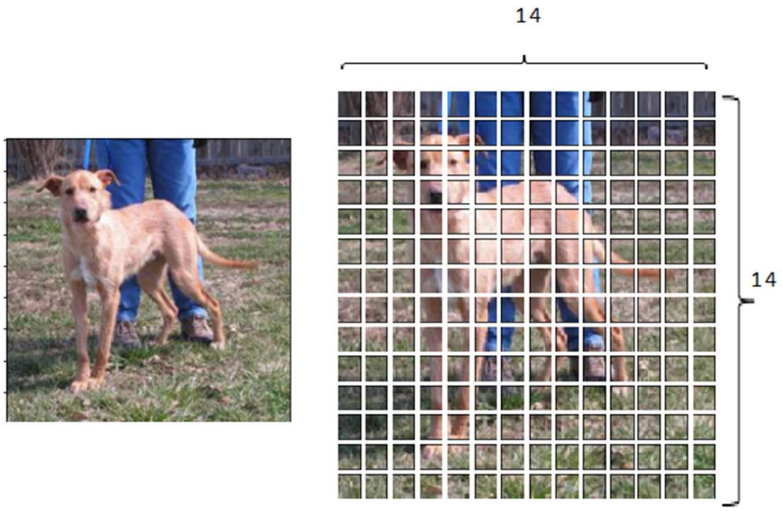
[1] Vanishing
[2] explosion

فصل 3: معماری Vision Transformer (ViT)

ترانسفورمر بینایی(ViT) [4] معماری Transformer را از NLP با وظیفه تشخیص تصویر تطبیق می‌دهد. این یک تصویر را با تقسیم آن به دنباله‌ای از وصله‌ها[1] پردازش می‌کند و سپس این وصله‌ها را به یک رمزگذار Transformer تغذیه می‌کند.

3-1 Patch Embedding: تقسیم تصاویر به واحدهای قابل مدیریت

اولین قدم در معماری ViT، تقسیم تصویر ورودی به شبکه‌ای از وصله‌های با اندازه ثابت است. به عنوان مثال، یک تصویر 224*224 را می‌توان به تکه‌های 16*16 تقسیم کرد که در نتیجه 14*14=196 وصله ایجاد می‌شود. سپس هر پچ به صورت یک بردار مسطح[2] می‌شود. این بردارهای وصله[3] مسطح شده به عنوان دنباله ورودی به رمزگذار ترانسفورمر عمل می‌کنند. اندازه پچ یک هایپرپارامتر مهم است که بر عملکرد ViT تأثیر می‌گذارد. اندازه‌های کوچک‌تر وصله‌ها جزئیات دقیق‌تری را ثبت می‌کنند، اما طول دنباله را افزایش می‌دهند، در حالی که اندازه‌های بزرگ‌تر وصله ویژگی‌های درشت‌تر را ثبت می‌کنند اما طول دنباله را کاهش می‌دهند.



شکل 3- 1 تقسیم‌کردن تصویر به وصله‌های یکسان

3-2 طرح ریزی خطی و رمزگذاری موقعیتی

[1] patch
[2] flatten
[3] patch

پس از مسطح‌کردن وصله‌ها، هر بردار وصله به صورت خطی در یک فضای تعبیه شده با ابعاد پایین‌تر نمایش داده می‌شود. این طرح خطی به عنوان مرحله استخراج ویژگی اولیه عمل می‌کند. مشابه ترانسفورمر در NLP، رمزگذاری‌های موقعیتی به تعبیه‌های وصله اضافه می‌شوند تا اطلاعاتی درباره موقعیت مکانی[1] هر وصله در تصویر اصلی ارائه کنند. این رمزگذاری‌های موقعیتی را می‌توان آموخت یا ثابت در نظر گرفت(به عنوان مثال، سینوسی). ترکیبی از تعبیه‌های وصله و کدگذاری‌های موقعیتی، دنباله ورودی رمزگذار ترانسفورمر را تشکیل می‌دهد.

3-3 بلوک‌های رمزگذار ترانسفورمر

هسته معماری ViT رمزگذار Transformer است. این شامل چندین لایه یکسان است که روی هم قرار گرفته‌اند. هر لایه شامل دو لایه فرعی اصلی است:

- **Multi-Head Self-Attention**: این لایه همان مکانیزم خود توجهی چند سر است که قبلا برای Transformer توضیح داده شد. این به مدل اجازه می‌دهد تا روابط بین وصله‌های مختلف در تصویر را ثبت کند. هر وصله به همه وصله‌های دیگر (و خودش) برای یادگیری زمینه سراسری توجه می‌کند.
- **شبکه پیش‌خور**: این لایه از دو لایه کاملاً متصل با یک تابع فعال‌سازی غیرخطی (معمولا GELU) در بین آن تشکیل شده است. خروجی لایه توجه چند سر را به صورت نقطه‌ای پردازش می‌کند.

این دو لایه فرعی با اتصالات باقیمانده و نرمال‌سازی لایه دنبال می‌شوند. اتصالات باقیمانده به آموزش شبکه‌های عمیق‌تر با کاهش مشکل ناپدیدشدن گرادیان کمک می‌کند. نرمال‌سازی لایه‌ها با عادی‌سازی فعالیت‌های هر لایه به تثبیت تمرین کمک می‌کند.

3-4 طبقه بندی

خروجی بلوک رمزگذار نهایی ترانسفورمر، دنباله‌ای از تعبیه‌های وصله است. برای طبقه‌بندی تصویر، یک سر طبقه‌بندی در بالای رمزگذار اضافه می‌شود. این سر طبقه‌بندی معمولاً از یک لایه کاملاً متصل تشکیل شده است که خروجی رمزگذار را به تعداد کلاس‌های مجموعه داده ترسیم می‌کند. یک نشانه طبقه‌بندی خاص (اغلب به عنوان [CLS] نشان داده می‌شود) قبل از وارد شدن به رمزگذار ترانسفورمر، به دنباله تعبیه‌های وصله اضافه می‌شود. خروجی مربوط به این نشانه [CLS] پس از ترانسفورمر به عنوان ورودی سر طبقه‌بندی استفاده می‌شود.

---

[1] Spatial Location

فصل 4: انواع ترانسفورمرهای بینایی

4-1 ترانسفورمر DeiT

DeiT (تبدیل کننده تصویر کارآمد داده[1])[5]، توسعه یافته توسط فیس‌بوک AI Research (FAIR)، به یک چالش کلیدی با ترانسفورمرهای بینایی یعنی نیازهای داده قابل توجه آن‌ها می‌پردازد. در حالی که ترانسفورمرهای بینایی عملکرد قابل توجهی از خود نشان دادند، اغلب برای دستیابی به نتایج پیشرفته نیاز به پیش‌آموزش بر روی مجموعه‌داده‌های عظیم داشتند، که باعث می‌شود محققان با منابع محاسباتی محدود یا دسترسی به چنین مجموعه‌داده‌های بزرگی کمتر در دسترس باشند. هدف DeiT بهبود کارایی داده‌های ViT است و به آن‌ها اجازه می‌دهد به طور موثر بر روی مجموعه داده‌های کوچکتر و در دسترس‌تر مانند ImageNet آموزش داده شوند.

تکنیک‌های کلیدی:

4-1-1 بهبود استراتژی آموزشی

DeiT از تکنیک‌های گسترده افزایش داده‌ها[2] در طول آموزش استفاده می‌کند. این افزایش‌ها شامل برش تصادفی، تغییر اندازه، چرخش، لرزش رنگ، و RandAugment، یک استراتژی تقویت خودکار قدرتمند است. این تقویت‌ها مدل را در معرض طیف گسترده‌تری از تبدیل‌های تصویر قرار می‌دهند و به آن کمک می‌کنند تا از داده‌های محدود بهتر تعمیم یابد.

DeiT از تکنیک‌های منظم‌سازی[3]، مانند dropout یا عمق‌تصادفی[4]، برای جلوگیری از برازش بیش از حد استفاده می‌کند، که به ویژه هنگام آموزش بر روی مجموعه‌های داده کوچک‌تر مهم است. منظم‌سازی به مدل کمک می‌کند تا ویژگی‌های قوی‌تر و قابل تعمیم بیشتری بیاموزد.

DeiT همچنین استفاده از استراتژی‌های افزایش داده Mixup و CutMix را بررسی می‌کند. Mixup با ترکیب خطی جفت تصاویر و برچسب‌های مربوط به آن‌ها، نمونه‌های آموزشی جدیدی ایجاد می‌کند. CutMix بخش‌هایی از تصاویر مختلف و برچسب‌های آن‌ها را ترکیب می‌کند. این تکنیک‌ها با ایجاد نمونه‌های آموزشی متنوع‌تر، کارایی داده‌ها را بیشتر افزایش می‌دهند.

---

[1] Data-efficient Image Transformer
[2] Data Augmentation
[3] Regularization
[4] stochastic depth

شکل 4- 1 نمایش دو نوع افزایش داده به روش *mixup* و *cutmix*

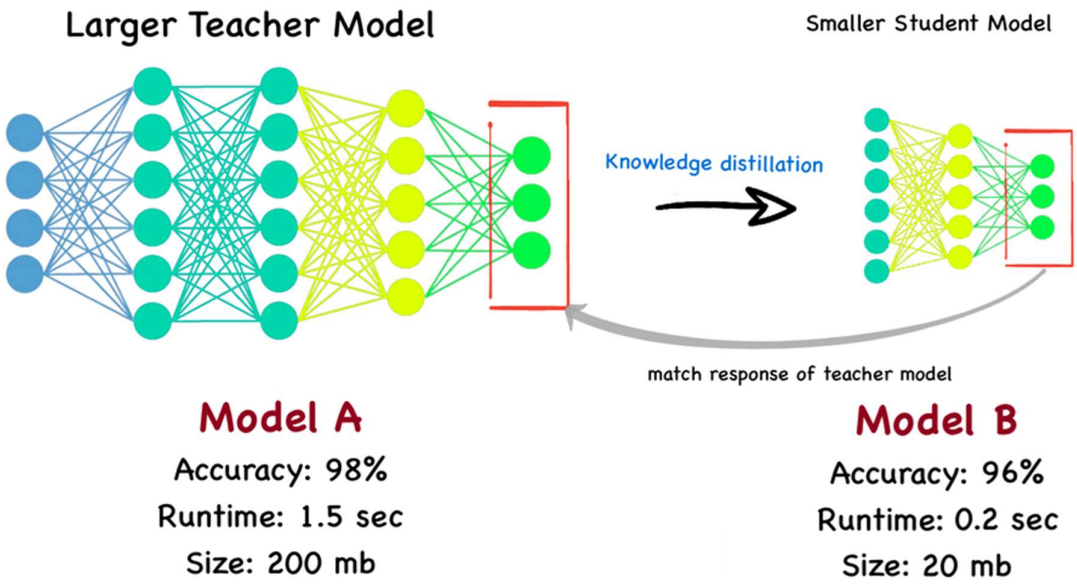
4-1-2 تقطیر[1]

یک عنصر حیاتی DeiT استفاده از تقطیر دانش است. به جای آموزش مستقیم مدل ViT، DeiT آن را برای تقلید از پیش‌بینی‌های یک مدل معلم، معمولاً یک شبکه عصبی کانولوشنال (CNN) مانند RegNet، آموزش می‌دهد. مدل معلم اهداف «نرم» (توزیع احتمال در کلاس‌ها) را به جای برچسب‌های «سخت» (یک کلاس صحیح) ارائه می‌کند. این اهداف نرم حاوی اطلاعات بیشتری در مورد روابط بین کلاس‌های مختلف است که می‌تواند برای آموزش مدل ViT دانش‌آموز مفید باشد. فرآیند تقطیر به ViT کمک می‌کند تا به طور موثرتری از داده‌های محدود یاد بگیرد.

4-1-3 آموزش با یک معلم CNN

مدل معلم در DeiT یک CNN است، اغلب یک مدل RegNet، که از قبل در ImageNet آموزش دیده است. این شبکه به عنوان یک پیشین قوی برای کار طبقه‌بندی تصویر عمل می‌کند. مدل دانش‌آموز ViT یاد می‌گیرد که رفتار این معلم را تقلید کند و به طور موثر دانش را از CNN به ViT منتقل می‌کند.

[1] Distillation



شکل 4- 2 نحوه تقطیر دانش در مدل‌ها

4-1-4 کارایی محاسباتی

در حالی که تقطیر یک سربار محاسباتی جزئی اضافه می‌کند (آموزش معلم)، فرآیند کلی آموزشی برای DeiT اغلب کارآمدتر از آموزش یک ViT استاندارد از ابتدا در مجموعه داده‌های بزرگ است. این امر DeiT را برای محققان با منابع محدود کاربردی‌تر می‌کند.

4-2 نتایج و تاثیر

DeiT به نتایج چشمگیری در ImageNet دست یافت و نشان داد که ViT را می‌توان به طور موثر با داده‌های بسیار کمتری نسبت به آن‌چه قبلا تصور می‌شد آموزش داد. عملکرد DeiT با CNNهای پیشرفته قابل مقایسه و در برخی موارد پیشی گرفته است، حتی زمانی که در ImageNet بدون آموزش قبلی روی مجموعه داده‌های بزرگ‌تر آموزش داده می‌شد. این کار به طور قابل توجهی دسترسی به ViT را گسترش داد و راه را برای تحقیقات بیشتر در آموزش ترانسفورمر بینایی کارآمد داده هموار کرد.

نوآوری کلیدی DeiT در استراتژی آموزشی موثر آن نهفته است که ترکیبی از افزایش داده‌ها، منظم‌سازی و تقطیر دانش است. با آموزش یک ViT برای تقلید از معلم CNN از قبل آموزش دیده، DeiT به طور قابل توجهی کارایی داده‌ها را بهبود می‌بخشد و آن‌ها را به ابزاری کاربردی‌تر و قدرتمندتر برای تشخیص تصویر تبدیل می‌کند. این نشان داد که بایاس القایی از CNNها می‌تواند به طور موثر از طریق تقطیر به ViT منتقل شود.

4-3 ترانسفورمر BEiT

BEiT، مخفف نمایش رمزگذار دوطرفه از ترانسفورمر تصویر[1][6]، یک روش یادگیری خود نظارت برای ترانسفورمرهای بینایی است که از موفقیت مدل‌سازی زبان ماسک‌شده[2] (MLM) در NLP، به‌ویژه با مدل‌هایی مانند BERT الهام می‌گیرد. این روش از تکنیکی به نام مدل‌سازی تصویر ماسک‌دار[3] برای آموزش ViT بدون تکیه بر برچسب‌های واضح استفاده می‌کند و آن‌ها را قادر می‌سازد تا بازنمایی‌های بصری غنی را از تصاویر خام بیاموزند.

4-3-1 مدل سازی تصویر ماسک شده

ایده اصلی پشت مدل‌سازی تصویر ماسک‌دار، آموزش مدلی برای بازسازی بخش‌های پوشانده شده از تصویر است. این شبیه به نحوه آموزش BERT برای پیش‌بینی کلمات پوشیده شده در یک جمله است. در زمینه تصاویر، درصد معینی از وصله‌های تصویر به‌طور تصادفی ماسک می‌شوند (با یک نشانه ماسک ویژه جایگزین می‌شوند)، و مدل وظیفه دارد مقادیر پیکسل اصلی یا ویژگی‌های بصری این وصله‌های ماسک‌شده را پیش‌بینی کند.

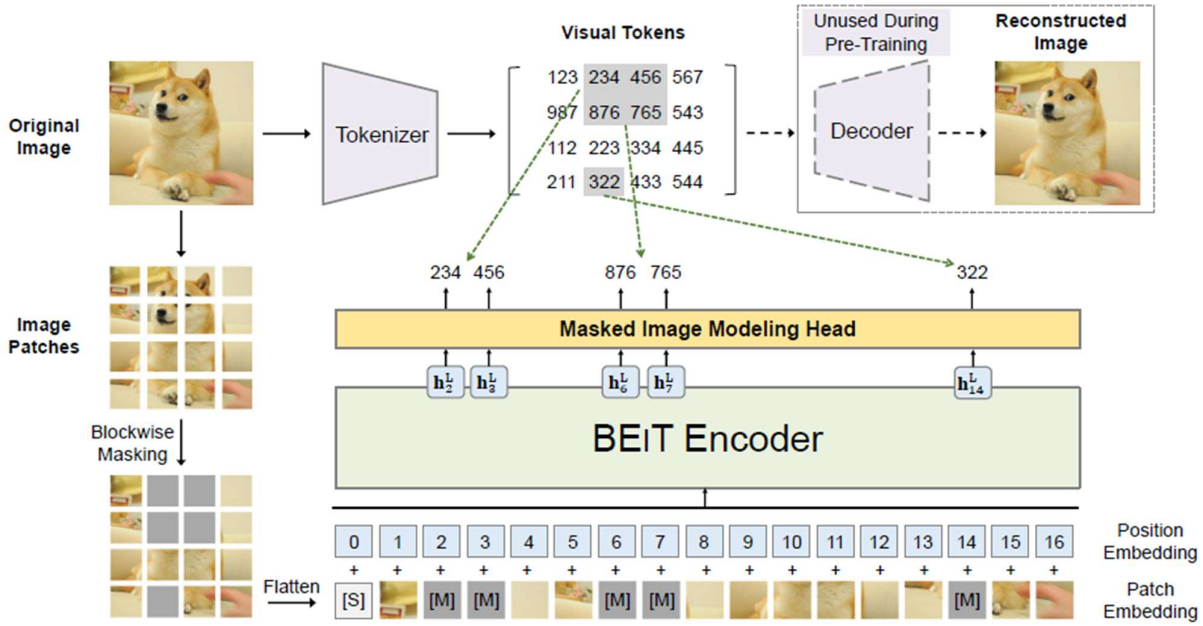
4-3-2 BEiT چگونه کار می‌کند؟

مانند ViT استاندارد، BEiT با تقسیم تصویر ورودی به وصله‌ها و نمایش خطی آن‌ها در یک فضای جاسازی شروع می‌شود. رمزگذاری‌های موقعیتی نیز اضافه شده است.

یک زیرمجموعه تصادفی از این وصله‌ها ماسک شده است. وصله‌های ماسک‌شده با یک جاسازی توکن قابل یادگیری ویژه [MASK] جایگزین می‌شوند. نسبت پوشش یک هایپرپارامتر بسیار مهم است که معمولاً از 15٪ تا 40٪ متغیر است.

---

[1] Bidirectional Encoder representation from Image Transformers
[2] masked language modeling
[3] masked image modeling



شکل 4- 3 ساختار *biet*

دنباله‌ای از تعبیه‌های وصله، از جمله نشانه‌های [MASK]، به یک رمزگذار ترانسفورمر وارد می‌شود. رمزگذار کل توالی را پردازش می‌کند و به مدل اجازه می‌دهد تا روابط بین هر دو وصله قابل مشاهده و ماسک شده را بیاموزد.

خروجی رمزگذار ترانسفورمر مربوط به وصله‌های پوشانده شده سپس از طریق یک رمزگشا (اغلب یک لایه خطی ساده) عبور می‌کند تا مقادیر پیکسل اصلی یا ویژگی‌های بصری وصله‌های ماسک شده را پیش‌بینی کند. BEiT توکن‌های بصری یا نشانه‌های گسسته وصله‌های ماسک‌شده را پیش‌بینی می‌کند. این نشانه‌های بصری با کمّی‌کردن مقادیر پیکسل تکه‌های اصلی با استفاده از رویکرد کوانتیزه‌برداری[1] (VQ) به دست می‌آیند. این مقادیر پیکسل‌های پیوسته را به مجموعه‌ای مجزا از نشانه‌های بصری یا نمایش‌های گسسته تبدیل می‌کند.

این مدل برای به حداقل رساندن خطا بازسازی آموزش داده شده است، که تفاوت بین نشانه‌های بصری پیش‌بینی‌شده و نشانه‌های بصری واقعی وصله‌های ماسک‌شده را اندازه‌گیری می‌کند. معمولاً از خطای آنتروپی متقاطع استفاده می‌شود.

4-3-3 مزایای کلیدی BEiT

BEiT یک روش خود نظارت است، به این معنی که برای آموزش به برچسب‌های دستی نیاز ندارد. این به آن اجازه می‌دهد تا از مقادیر عظیمی از داده‌های تصویر بدون برچسب استفاده‌کند که به راحتی در دسترس است.

[1] vector quantized

با آموزش بازسازی وصله‌های تصویر ماسک دار، BEiT بازنمایی‌های بصری غنی و معنی دار را می‌آموزد. این نمایش‌ها ساختار زیربنایی و معنایی تصاویر را به تصویر می‌کشند. نمایش‌هایی که توسط BEiT آموخته می‌شود را می‌توان برای کارهای مختلف پایین‌دستی، مانند طبقه‌بندی تصویر، تشخیص اشیا، و تقسیم‌بندی معنایی، به خوبی تنظیم کرد. مدل‌های اولیه BEiT اغلب در مقایسه با مدل‌هایی که از ابتدا آموزش دیده‌اند، عملکرد بهتر و همگرایی سریع‌تری دارند.

4-3-4 اتصال به مدل سازی زبان ماسک شده (MLM)

رویکرد مدل‌سازی تصویر ماسک‌دار BEiT مستقیماً از تکنیک مدل‌سازی زبان پوشانده شده در مدل‌های NLP مانند BERT الهام گرفته شده است. هر دو روش، مدل‌ها را برای پیش‌بینی بخش‌های پوشانده شده از داده‌های ورودی آموزش می‌دهند و مدل را مجبور می‌کنند تا بازنمایی‌های متنی را بیاموزد. تفاوت اصلی این است که BEiT بر روی تصاویر عمل می‌کند، در حالی که BERT بر روی متن عمل می‌کند.

4-3-5 نتایج و تاثیر

BEiT از مدل‌سازی تصویر پوشانده برای آموزش ترانسفورمرهای بینایی به شیوه‌ای تحت نظارت خود استفاده می‌کند. BEiT با یادگیری بازسازی وصله‌های تصویر ماسک‌دار، بازنمایی‌های بصری قدرتمندی را می‌آموزد که می‌توانند به طور مؤثر به وظایف پایین دستی منتقل شوند و پتانسیل یادگیری خود نظارتی برای بینایی رایانه را نشان می‌دهند. استفاده از نشانه‌های بصری گسسته جنبه کلیدی موفقیت BEiT است.

4-4 ترانسفورمر PiT

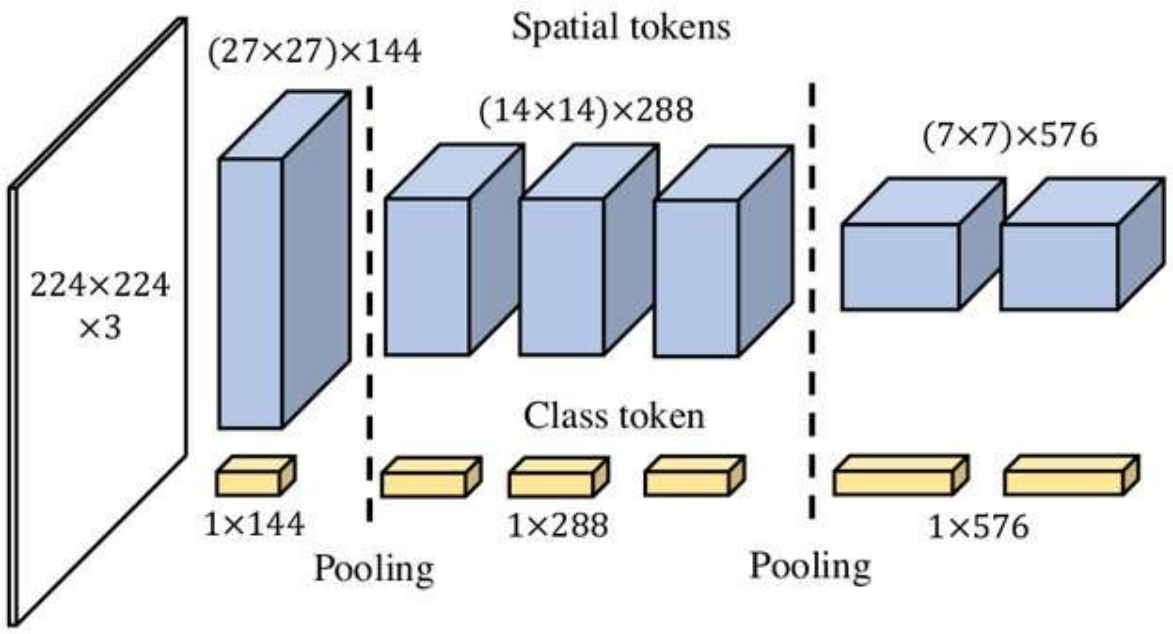
یکی از چالش‌های اصلی ترانسفورمرهای بینایی هزینه محاسباتی آن‌ها است، به‌ویژه در هنگام برخورد با تصاویر با وضوح بالا. مکانیسم توجه به خود، در حالی که قدرتمند است، دارای پیچیدگی درجه دوم نسبت به تعداد وصله‌ها است، که از نظر محاسباتی برای تصاویر بزرگ یا اندازه‌های وصله کوچک گران است. چندین تلاش تحقیقاتی بر بهبود کارایی ViT متمرکز شده‌اند و PiT (ترانسفورمر بینایی مبتنی بر ترکیب[1])[7] یکی از این رویکردها است.

4-4-1 ایده اصلی PiT

PiT با معرفی یک مکانیسم ادغام جدید در لایه‌های توجه به خود، تنگنای کارایی را برطرف می‌کند. به جای محاسبه توجه بین تمام جفت‌های وصله، PiT توجه را بین مجموعه‌ای کاهش‌یافته از نمایش‌های تلفیقی از وصله‌ها انجام می‌دهد. این به طور قابل توجهی هزینه محاسباتی را کاهش می‌دهد در حالی که هنوز وابستگی‌های طولانی برد مهم را در بر می‌گیرد.

PiT، مانند سایر ViTها، با تقسیم تصویر ورودی به وصله‌ها و نمایش خطی آن‌ها در یک فضای تعبیه[2] شروع می‌شود. رمزگذاری‌های موقعیتی نیز برای ارائه اطلاعات در مورد موقعیت مکانی هر پچ اضافه شده است. نوآوری کلیدی PiT در معرفی یک عملیات ادغام قبل از مکانیسم توجه نهفته است. یک لایه ادغام اطلاعات را از وصله‌های همسایه جمع‌آوری می‌کند و مجموعه کوچک‌تری از نمایش‌های تلفیقی را ایجاد می‌کند. این ادغام را می‌توان با استفاده از تکنیک‌های مختلفی از جمله جمع‌آوری متوسط یا حداکثر جمع آوری انجام داد. اندازه پنجره ادغام یک فراپارامتر مهم است که میزان اطلاعات جمع‌آوری شده را کنترل می‌کند.

[1] Pooling-based Vision Transformer
[2] Embedding



شکل 4- 4 ساختار شبکه *PiT*

سپس مکانیسم توجه به خود به جای تعبیه‌های وصله اصلی، روی این نمایش‌های ترکیبی اعمال می‌شود. از آن‌جایی که تعداد نمایش‌های تلفیقی بسیار کمتر از تعداد وصله‌های اصلی است، هزینه محاسباتی مکانیسم توجه به طور قابل توجهی کاهش می‌یابد. پس از عملیات توجه بر روی نمایش‌های تلفیقی، ویژگی‌ها به وضوح وصله اصلی نمونه‌برداری می‌شوند. این را می‌توان با استفاده از تکنیک‌هایی مانند درون‌یابی دو خطی یا کانولوشن‌های انتقالی انجام داد. PiT از چندین لایه از این عملیات ادغام، توجه و نمونه‌برداری استفاده می‌کند که با سایر اجزای استاندارد ترانسفورمر مانند شبکه‌های پیش‌خور، اتصالات باقی‌مانده و عادی‌سازی لایه‌ها ترکیب شده‌اند.

4-4-2 مزایای کلیدی PiT

PiT چندین مزیت کلیدی را ارائه می‌دهد. با توجه به نمایش‌های تلفیقی، کارایی را به‌طور قابل‌توجهی بهبود می‌بخشد، هزینه‌های محاسباتی را به‌ویژه برای تصاویر با وضوح بالا کاهش می‌دهد. این کارایی بهبود یافته، PiT را برای تصاویر و مجموعه داده‌های بزرگتر مقیاس پذیرتر می‌کند. علیرغم کاهش هزینه محاسباتی، PiT عملکرد رقابتی را حفظ می‌کند و اغلب از ترانسفورمرهای بینایی استاندارد در وظایف مختلف تشخیص تصویر پیشی می‌گیرد.

4-4-3 تفاوت PiT با ترانسفورمرهای بینایی استاندارد

تفاوت اصلی بین PiT و ترانسفورمرهای بینایی استاندارد، معرفی مکانیسم ادغام قبل از عملیات توجه است. این به PiT اجازه می‌دهد تا توجه را بین مجموعه کوچک‌تری از ویژگی‌ها محاسبه کند که منجر به بهبود کارایی می‌شود. استاندارد ViT توجه را بین تمام جفت‌های وصله محاسبه می‌کند که می‌تواند از نظر محاسباتی گران باشد.

4-4-4 نتایج و تاثیر

PiT نمونه‌ای از معماری ترانسفورمرهای بینایی کارآمدتر است. PiT با معرفی مکانیزم ادغام قبل از عملیات خودتوجهی، هزینه محاسباتی را کاهش می‌دهد و ترانسفورمرهای بینایی را برای تصاویر و مجموعه داده‌های بزرگ‌تر مقیاس پذیرتر می‌کند. این نشان می‌دهد که بهبود کارایی را می‌توان بدون قربانی کردن عملکرد به دست آورد و راه را برای کاربردهای عملی‌تر ترانسفورمرهای بینایی هموار کرد.

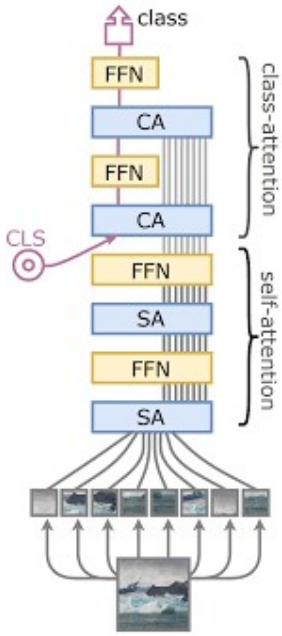
4-5 ترانسفورمرCaiT

CaiT یا کلاس توجه در ترانسفورمرهای تصویر[1]، رویکرد دیگری است که با هدف بهبود کارایی و عملکرد ترانسفورمرهای بینایی، به ویژه برای کارهای طبقه‌بندی تصویر، انجام می‌شود. بر اصلاح مکانیسم توجه و معرفی یک ماژول تخصصی "توجه به کلاس" تمرکز دارد.

4-5-1 ایده‌های کلیدی در CaiT

CaiT[8] یک لایه جدید "توجه به کلاس" را معرفی می‌کند. در ترانسفورمرهای بینایی استاندارد، نشانه طبقه‌بندی ([CLS]) به همه پچ توکن‌های دیگر توجه می‌کند و زمینه سراسری را به تصویر می‌کشد. CaiT این را یک قدم فراتر می‌برد. لایه توجه به کلاس به طور خاص بر اصلاح نمایش نشانه [CLS] با توجه به خروجی آخرین بلوک ترانسفورمر متمرکز است. این به مدل اجازه می‌دهد تا به صراحت بر یادگیری یک نمایش بهتر برای کار طبقه‌بندی تمرکز کند. به جای اینکه توکن [CLS] با همه وصله‌ها در کل شبکه تعامل داشته باشد، در درجه اول با اطلاعات جمع‌آوری شده از همه وصله‌ها در مرحله نهایی تعامل دارد.

[1] Class-attention in Image Transformers



شکل 4- 5 ساختار شبکه *CaiT*

CaiT همچنین توکن کلاس را به هر بلوک ترانسفورمر تزریق می‌کند. این بدان معناست که نشانه کلاس نه تنها در ابتدای شبکه بلکه در هر لایه بعدی نیز وجود دارد. این توکن کلاس را قادر می‌سازد تا به تدریج نمایش خود را هنگام عبور از شبکه اصلاح کند و اطلاعات را از سطوح مختلف انتزاع ویژگی جمع‌آوری کند.

در حالی که CaiT لایه توجه کلاس را معرفی می‌کند، برخی از نسخه‌های CaiT مکانیسم استاندارد خودتوجهی چند سر را ساده می‌کنند. به عنوان مثال، ممکن است از سر توجه کمتر یا ابعاد کلید/پرس و جو/مقدار کوچکتر استفاده کنند. این بیشتر به بهبود کارایی مدل کمک می‌کند.

CaiT همچنین استفاده از یک میکسر توکن مبتنی بر کانولوشن را به جای مکانیسم استاندارد خود توجهی در برخی از لایه‌ها بررسی می‌کند. این می‌تواند از نظر محاسباتی، به خصوص در لایه‌های اولیه شبکه، کارآمدتر باشد.

4-5-2 چگونه CaiT کارایی و عملکرد را بهبود می‌بخشد

CaiT کارایی و عملکرد را با تمرکز صریح بر اصلاح نمایش نشانه کلاس از طریق لایه توجه کلاس، کاهش افزونگی محاسباتی با ساده کردن مکانیسم توجه استاندارد و استفاده از کانولوشن‌های عمیق، و امکان اصلاح لایه‌ای نشانه کلاس با تزریق آن به هر لایه، بهبود می‌بخشد.

4-5-3 تفاوت‌های اصلی با ترانسفورمرهای بینایی استاندارد

تفاوت اصلی بین CaiT و ترانسفورمرهای بینایی استاندارد، معرفی لایه توجه کلاس و تزریق نشانه کلاس به صورت لایه است. CaiT همچنین به بررسی ساده‌سازی مکانیسم توجه استاندارد و استفاده از پیچش‌های عمیق برای افزایش کارایی می‌پردازد.

4-5-4 نتایج و تاثیر

CaiT نمونه دیگری از معماری ترانسفورمر بینایی کارآمد است که بر بهبود عملکرد در وظایف طبقه‌بندی تصاویر تمرکز دارد. CaiT با معرفی یک لایه توجه به کلاس و تزریق نشانه کلاس به صورت لایه، نمایش توکن کلاس را اصلاح می‌کند و به دقت بهبود یافته دست می‌یابد. ساده‌سازی مکانیسم توجه استاندارد و استفاده از پیچش‌های عمقی به افزایش کارایی کمک می‌کند. CaiT نشان می‌دهد که طراحی دقیق مکانیسم توجه و استفاده از نشانه کلاس می‌تواند منجر به بهبود عملکرد و کارایی در ترانسفورمرهای بینایی شود.

4-6 ترانسفورمر Swin

ترانسفورمرSwin[9] که توسط تحقیقات مایکروسافت توسعه یافته است، به یک محدودیت کلیدی ترانسفورمرهای بینایی استاندارد می‌پردازد پیچیدگی محاسباتی آن‌ها با توجه به وضوح تصویر. ترانسفورمرهای بینایی استاندارد، توجه به خود را در سراسر همه وصله‌ها محاسبه می‌کند، که برای تصاویر با وضوح بالا بسیار گران می‌شود. ترانسفورمر Swin مکانیزم توجه مبتنی بر پنجره سلسله مراتبی[1] را معرفی می‌کند که به طور قابل توجهی کارایی را بهبود می‌بخشد و امکان مدیریت بهتر اندازه‌های مختلف تصویر را فراهم می‌کند.

4-6-1 ویژگی‌ها

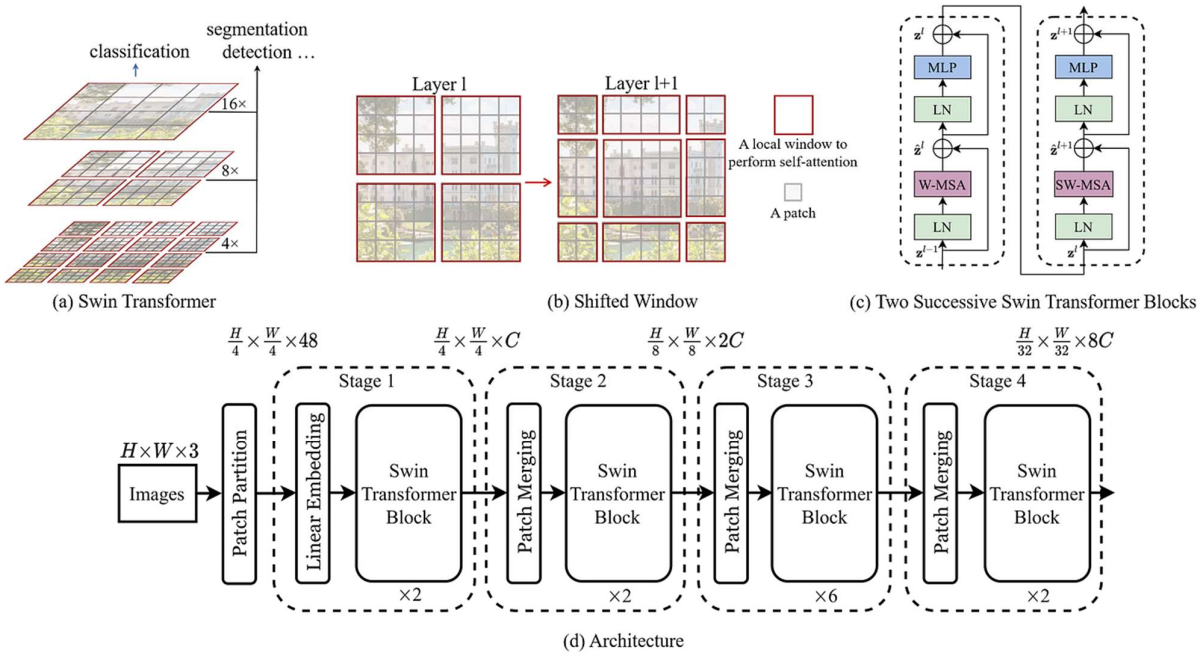
ترانسفورمر Swin از مکانیسم سراسری خود توجهی ترانسفورمرهای بینایی استاندارد خارج می‌شود. در عوض، توجه به خود را در پنجره‌های محلی محاسبه می‌کند. تصویر به پنجره‌های غیرهمپوشانی تقسیم می‌شود و توجه به خود فقط به تکه‌های داخل هر پنجره اعمال می‌شود. این رویکرد به طور چشمگیری پیچیدگی محاسباتی را از درجه دوم در اندازه تصویر به خطی در تعداد پنجره‌ها کاهش می‌دهد.

در حالی که توجه مبتنی بر پنجره کارآمدتر است، اما می‌تواند تعامل بین وصله‌های واقع در پنجره‌های مختلف را محدود کند. برای کاهش این موضوع، ترانسفورمر Swin پنجره‌های جابجاشده[2] را در لایه‌های متناوب معرفی می‌کند. پنجره‌های این لایه‌ها به مقدار مشخصی جابجا می‌شوند و مناطق همپوشانی با پنجره‌های لایه قبلی

[1] hierarchical window-based attention mechanism
[2] shifted windows

ایجاد می‌کنند. این تغییر استراتژیک اجازه می‌دهد تا اطلاعات بین پنجره‌های مختلف جریان یابد و مدل را قادر می‌سازد تا ضمن حفظ کارایی، زمینه سراسری را به تصویر بکشد.

ترانسفورمر Swin ساختاری سلسله مراتبی دارد که یادآور شبکه‌های عصبی کانولوشنی است. با پنجره‌های کوچک در لایه‌های اولیه شروع می‌شود و به تدریج اندازه پنجره را در لایه‌های بعدی افزایش می‌دهد. این طراحی سلسله مراتبی مدل را قادر می‌سازد تا ویژگی‌های محلی و سراسری را به طور موثر یاد بگیرد. ساختار سلسله مراتبی با ادغام وصله‌های همسایه در لایه‌های متوالی پیاده سازی می‌شود. این فرآیند ادغام تعداد وصله‌ها را کاهش می‌دهد و در نتیجه میدان دریافت موثر شبکه را افزایش می‌دهد.



شکل 4- 6 ایده و ساختار *swin*

بلوک ترانسفورمر Swin یک واحد با دقت ساخته شده است که از چندین لایه تشکیل شده است. در هر لایه، ماژول‌های خودتوجهی چند سر مبتنی بر پنجره و ماژول‌های خودتوجهی چند سر مبتنی بر پنجره تغییر یافته را خواهید یافت. این ماژول‌ها با لایه‌های نرمال‌سازی و لایه‌های MLP تکمیل می‌شوند که به اثربخشی کلی بلوک کمک می‌کنند.

4-6-2 مزایای ترانسفورمر Swin

مکانیسم‌های توجه پنجره مبتنی بر پنجره و تغییر مکان به طور قابل توجهی هزینه محاسباتی را کاهش می‌دهد و ترانسفورمر Swin را برای تصاویر با وضوح بالا و مجموعه داده‌های بزرگ‌تر مقیاس پذیرتر می‌کند. علیرغم بهره‌وری بهبودیافته، ترانسفورمر Swin به عملکردی پیشرفته در کارهای بینایی مختلف، از جمله طبقه بندی تصویر،

تشخیص اشیا، و تقسیم بندی معنایی دست می‌یابد. اندازه پنجره و مقدار جابجایی را می‌توان برای کنترل مبادله بین کارایی و عملکرد تنظیم کرد.

نتایج و تاثیر

ترانسفورمر Swin یک معماری ترانسفورمر بینایی بسیار کارآمد و موثر است که به محدودیت‌های محاسباتی ترانسفورمرهای بینایی استاندارد می‌پردازد. ترانسفورمر Swin با معرفی توجه پنجره‌ای و تغییر یافته به پنجره، همراه با ساختار سلسله مراتبی، به عملکرد پیشرفته‌تری در وظایف مختلف بینایی دست می‌یابد در حالی که از نظر محاسباتی بسیار کارآمدتر است. این یک معماری بسیار تأثیرگذار در زمینه ترانسفورمر بینایی شده است.

4-7 سایر معماری‌های قابل توجه

فراتر از معماری‌هایی که قبلاً مورد بحث قرار گرفت، حوزه ترانسفورمرهای بینایی با رویکردهای جدید و نوآورانه در حال توسعه دائماً در حال تکامل است. در اینجا چند نمونه آورده شده است:

- **ViLT**[10]**:** (Vision-and-Language Transformer) معماری ترانسفورمر را برای مدیریت اطلاعات بصری و زبانی گسترش می‌دهد. این برای کارهایی طراحی شده است که نیاز به درک رابطه بین تصاویر و متن دارند، مانند پاسخگویی بصری به سؤال[1] (VQA) یا نوشتن شرح تصویر[2]. ViL معمولاً از رمزگذارهای ترانسفورمر جداگانه برای تصویر و متن استفاده می‌کند و سپس نمایش‌های این رمزگذارها را برای انجام وظیفه مورد نظر ترکیب می‌کند.
- **MAE**[11]**:** MAE (Masked Autoencoders) یک روش یادگیری خود نظارت است که از استراتژی پوششی مشابه BEiT استفاده می‌کند، اما به جای تکیه بر نشانه‌های بصری، بر بازسازی پیکسل‌های پوشانده شده به طور مستقیم تمرکز دارد. MAE به طور تصادفی نسبت بالایی از وصله‌های تصویر (به عنوان مثال، 75٪) را ماسک می‌کند و مدل را برای بازسازی پیکسل‌های از دست رفته آموزش می‌دهد. این رویکرد برای یادگیری بازنمایی‌های بصری قدرتمند بسیار موثر است. یکی از جنبه‌های کلیدی MAE طراحی رمزگشای آن است که فقط روی تکه‌های ماسک شده عمل می‌کند و فرآیند بازسازی را کارآمدتر می‌کند.
- **DINO**[12]: (تقطیر با برچسب‌های NO) DINO یکی دیگر از روش‌های یادگیری خود نظارت است که قدرت ترانسفورمرها را با تقطیر دانش ترکیب می‌کند، اما بدون تکیه بر برچسب‌های صریح. این یک مدل ترانسفورمر دانش آموز را آموزش می‌دهد تا با خروجی یک مدل ترانسفورمر معلم مطابقت داشته باشد، جایی

[1] Visual Question Answering
[2] Image Captioning

که هر دو مدل به طور تصادفی مقداردهی اولیه می‌شوند. از طریق یک فرآیند آموزشی با دقت طراحی شده، DINO بازنمایی‌های بصری عالی را می‌آموزد که می‌تواند برای کارهای مختلف پایین دستی استفاده شود.

- **CrossViT**[13]: برای کنترل وضوح ورودی‌های متعدد طراحی شده است. از شاخه‌های ترانسفورمر جداگانه برای وضوح‌های مختلف استفاده می‌کند و سپس اطلاعات این شاخه‌ها را ترکیب می‌کند. این به ویژه برای کارهایی مفید است که اطلاعات چند مقیاسی مهم است، مانند تشخیص اشیا.

این‌ها فقط یک نمونه کوچک از بسیاری از معماری‌های ترانسفورمر بینایی است که توسعه یافته‌اند. تحقیقات در این زمینه بسیار فعال است و معماری‌های جدید مدام پیشنهاد می‌شود. تمرکز مداوم بر بهبود کارایی، عملکرد و بررسی برنامه‌های جدید ترانسفورمرهای بینایی است.

فصل 5: کاربردهای ترانسفورمرهای بینایی

ترانسفورمرهای بینایی عملکرد قابل توجهی را در طیف گسترده‌ای از وظایف بینایی کامپیوتری نشان داده‌اند و اغلب به نتایج پیشرفته‌ای دست می‌یابند. توانایی آن‌ها در گرفتن زمینه سراسری و وابستگی‌های دوربرد آن‌ها را برای کاربردهای مختلف مناسب می‌کند.

5-1 طبقه بندی تصویر (ImageNet و مجموعه داده‌های دیگر)

طبقه‌بندی تصویر، وظیفه اختصاص یک برچسب به یک تصویر، یکی از اولین زمینه‌هایی بود که ترانسفورمرهای بینایی پتانسیل خود را نشان دادند. ترانسفورمرهای بینایی ، از جمله تغییراتی مانند DeiT، Swin Transformer و CaiT، در مجموعه داده‌هایی مانند ImageNet به نتایج عالی دست یافته‌اند، که اغلب از عملکرد CNNهای سنتی پیشی می‌گیرند. آن‌ها همچنین با موفقیت در سایر مجموعه داده‌های طبقه‌بندی تصویر اعمال شده‌اند که تطبیق‌پذیری آن‌ها را نشان می‌دهد.

5-2 تشخیص شی

تشخیص شی شامل شناسایی و محلی‌سازی[1] اشیاء درون یک تصویر است. ViT در چارچوب‌های تشخیص اشیا ادغام شده‌اند و اغلب جایگزین ستون فقرات CNN مورد استفاده برای استخراج ویژگی می‌شوند. معماری‌هایی مانند Swin Transformer عملکرد چشمگیری در معیارهای تشخیص اشیا مانند COCO نشان داده‌اند و توانایی خود را در گرفتن اطلاعات متنی مورد نیاز برای شناسایی دقیق و محلی‌سازی اشیا نشان می‌دهند.

5-3 تقسیم بندی معنایی

تقسیم‌بندی معنایی وظیفه اختصاص یک برچسب به هر پیکسل در یک تصویر است که به طور موثر تصویر را به مناطق مربوط به کلاس‌های شی مختلف تقسیم می‌کند. ViTها همچنین برای تقسیم‌بندی معنایی استفاده شده‌اند و نتایج رقابتی در مجموعه داده‌هایی مانند ADE20K به دست می‌آید. توانایی آن‌ها در گرفتن وابستگی‌های دوربرد برای درک زمینه هر پیکسل و تقسیم‌بندی دقیق صحنه‌های پیچیده بسیار مهم است.

---

[1] localizing

5-4 تولید تصویر

در حالی که ترانسفورمرهای بینایی به اندازه سایر برنامه‌ها مورد بررسی قرار نگرفته‌اند، اما در حال نفوذ به تولید تصویر هستند. با ترکیب ترانسفورمرها با شبکه‌های متخاصم مولد (GAN) یا سایر مدل‌های مولد، محققان در حال بررسی پتانسیل ترانسفورمرهای بینایی برای تولید تصاویر با کیفیت بالا هستند.

5-5 تجزیه و تحلیل ویدئو

توانایی ترانسفورمرها برای مدیریت داده‌های متوالی آن‌ها را برای کارهای تجزیه و تحلیل ویدیو مناسب می‌کند. ترانسفورمرهای بینایی را می‌توان برای طبقه‌بندی ویدیو، تشخیص اقدام و سایر کارهای مرتبط با ویدیو با در نظر گرفتن فریم‌های ویدیو به عنوان دنباله‌ای از وصله‌های تصویر اعمال کرد. مکانیسم توجه زمانی در Transformers به مدل اجازه می‌دهد تا اطلاعات حرکت و وابستگی‌های بلندمدت را در ویدیوها ثبت کند.

5-6 سایر برنامه‌های در حال ظهور

تطبیق پذیری ترانسفورمرهای بینایی منجر به کاوش آن‌ها در طیف گسترده ای از کاربردهای دیگر شده است، از جمله:

- **تجزیه و تحلیل تصویر پزشکی:** ترانسفورمرهای بینایی برای کارهایی مانند تشخیص بیماری، تقسیم‌بندی تصویر، و تشخیص ناهنجاری در تصاویر پزشکی استفاده می‌شوند. توانایی آن‌ها در گرفتن زمینه سراسری به ویژه برای تجزیه و تحلیل اسکن‌های پزشکی مهم است.
- **سنجش از دور:** ترانسفورمر بینایی برای تجزیه و تحلیل تصاویر ماهواره‌ای و هوایی برای کارهایی مانند طبقه‌بندی پوشش زمین، برنامه ریزی شهری، و نظارت بر بلایا استفاده می‌شوند.
- **رباتیک**: ترانسفورمر بینایی در رباتیک برای کارهایی مانند تشخیص اشیا، درک صحنه، و ناوبری استفاده می‌شوند.
- **وظایف چندوجهی:** ترانسفورمر بینایی با سایر روش‌ها، مانند متن یا صدا، ترکیب می‌شوند تا به وظایف چندوجهی مانند پاسخ‌گویی بصری به سؤال یا شرح تصویر بپردازند.
- **بینایی سطح پایین:** در حالی که در ابتدا بر روی وظایف بینایی سطح بالا متمرکز بودند، ViTها برای مشکلات بینایی سطح پایین مانند وضوح تصویر فوق العاده و حذف نویز تصویر نیز مورد بررسی قرار می‌گیرند.

کاربردهای ترانسفورمرهای بینایی به سرعت در حال گسترش است زیرا محققان به کشف قابلیت‌های آن‌ها و توسعه معماری‌های جدید ادامه می‌دهند. توانایی آن‌ها در گرفتن زمینه سراسری و وابستگی‌های دوربرد آن‌ها را به ابزاری قدرتمند برای طیف گسترده‌ای از وظایف بینایی رایانه تبدیل می‌کند.

فصل 6: آموزش و بهینه سازی ترانسفورمرهای بینایی

آموزش ترانسفورمرهای بینایی به طور موثر چندین جنبه حیاتی را شامل می‌شود، از انتخاب مجموعه داده‌ها و استراتژی‌های قبل از آموزش گرفته تا تنظیم هایپرپارامتر و منابع محاسباتی. هر یک از این عناصر نقش بسزایی در دستیابی به عملکرد مطلوب دارند.

6-1 مجموعه داده‌ها و استراتژی‌های قبل از آموزش

ترانسفورمرهای بینایی، به‌ویژه مدل‌های قبلی، اغلب از مجموعه داده‌های مقیاس بزرگ برای پیش‌آموزش بهره می‌برند. مجموعه داده‌هایی مانند ImageNet، JFT-300M، و حتی مجموعه داده‌های بزرگ‌تر و تخصصی‌تر، اغلب برای پیش آموزش ViTها استفاده می‌شوند. هدف از پیش‌آموزش، یادگیری بازنمایی‌های بصری کلی است که می‌توان آن‌ها را برای کارهای پایین دستی خاص تنظیم کرد. انتخاب مجموعه داده به کار در دست و در دسترس بودن داده‌ها بستگی دارد. اخیراً، استراتژی‌های پیش‌آموزشی تحت نظارت خود، مانند مدل‌سازی تصویر ماسک‌دار (همانطور که در BEiT و MAE استفاده می‌شود)، به طور فزاینده‌ای محبوب شده‌اند. این روش‌ها به ViT اجازه می‌دهند تا نمایش‌های قدرتمندی را از داده‌های بدون برچسب بیاموزند و اتکا به مجموعه داده‌های برچسب‌دار عظیم را کاهش دهند. زمانی که داده‌های برچسب‌گذاری‌شده کمیاب یا پرهزینه است، پیش‌آموزش خود نظارتی می‌تواند مفید باشد.

6-2 تنظیم هایپر پارامتر

تنظیم هایپر پارامتر یک مرحله حیاتی در آموزش ViTها است. چندین فراپارامتر کلیدی باید به دقت تنظیم شوند، از جمله نرخ یادگیری، اندازه دسته، اندازه وصله، تعداد سرهای توجه، عمق شبکه و پارامترهای تنظیم (به عنوان مثال، نرخ دراپ‌اوت، کاهش وزن). مقادیر بهینه برای این ابرپارامترها بسته به معماری خاص، مجموعه داده و وظیفه می‌تواند متفاوت باشد. تکنیک‌هایی مانند جستجوی شبکه‌ای، جستجوی تصادفی و بهینه‌سازی بیزی را می‌توان برای کشف فضای هایپرپارامتر و یافتن بهترین ترکیب استفاده کرد. زمان‌بندی نرخ یادگیری، نیز برای بهینه‌سازی فرآیند تمرین مهم است. تنظیم دقیق هایپر پارامتر می‌تواند به طور قابل توجهی بر عملکرد نهایی مدل ViT آموزش دیده تأثیر بگذارد.

6-3 منابع محاسباتی و مقیاس پذیری

آموزش ترانسفورمرهای بینایی بزرگ به منابع محاسباتی قابل توجهی نیاز دارد. این مدل‌ها اغلب دارای میلیون‌ها یا حتی میلیاردها پارامتر هستند که به پردازنده‌های گرافیکی قدرتمند و مقدار زیادی حافظه نیاز دارند. آموزش توزیع شده اغلب برای سرعت بخشیدن به فرآیند آموزش با توزیع حجم کار در چندین GPU استفاده می‌شود. مقیاس‌پذیری یک نگرانی کلیدی است، به ویژه هنگامی که با تصاویر با وضوح بالا یا مجموعه داده‌های بزرگ سروکار داریم. معماری‌های کارآمد، مانند Swin Transformer و PiT، با کاهش پیچیدگی محاسباتی مکانیسم توجه، این چالش را برطرف می‌کنند. پلتفرم‌های رایانش ابری دسترسی به منابع محاسباتی لازم برای آموزش مدل‌های بزرگ ViT را فراهم می‌کنند.

6-4 چالش‌ها و بهترین شیوه‌ها

آموزش ترانسفورمرهای بینایی چالش‌های متعددی را به همراه دارد. یکی از چالش‌ها هزینه محاسباتی بالا است که می‌تواند آموزش مدل‌های بزرگ را وقت گیر و گران کند. چالش دیگر، پتانسیل بیش برازش است، به ویژه هنگامی که در مجموعه داده‌های کوچک‌تر آموزش می‌بینید. تکنیک‌های تقویت و منظم‌سازی داده‌ها برای کاهش بیش برازش بسیار مهم هستند. آموزش با ثبات نیز می‌تواند چالش برانگیز باشد و تنظیم دقیق هایپرپارامترها ضروری است. برخی از بهترین شیوه‌ها برای آموزش ViTها عبارتند از استفاده از تکنیک‌های تقویت داده‌های مناسب، به‌کارگیری روش‌های منظم‌سازی، استفاده از معماری‌های کارآمد در صورت لزوم، استفاده از مدل‌های پیش‌آموزش (به‌ویژه با self-supervised)، و تنظیم دقیق هایپرپارامترها. نظارت بر فرآیند آموزش و تجزیه و تحلیل عملکرد مدل در مجموعه‌های اعتبارسنجی نیز برای شناسایی و پرداختن به مسائل بالقوه ضروری است.

فصل 7: جهت گیری‌های آینده و روندهای تحقیقاتی

حوزه ترانسفورمرهای بینایی به سرعت در حال تکامل است و جهت‌های تحقیقاتی متعددی به طور فعال دنبال می‌شود. این روندها نوید افزایش بیشتر قابلیت‌های ترانسفورمرهای بینایی و گسترش کاربردهای آن‌ها را می‌دهند.

7-1 بهبود کارایی و مقیاس پذیری

علیرغم پیشرفت‌های اخیر، بهبود کارایی و مقیاس‌پذیری ترانسفورمرهای بینایی همچنان یک حوزه تحقیقاتی حیاتی است. پیچیدگی درجه دوم مکانیزم توجه به خود با توجه به تعداد وصله‌ها چالشی برای تصاویر با وضوح بالا و مجموعه داده‌های بزرگ ایجاد می‌کند. محققان به طور فعال در حال بررسی مکانیسم‌های توجه جدید، مانند توجه خطی و توجه sparse هستند تا هزینه محاسباتی را کاهش دهند. معماری‌های کارآمد، مانند Swinترانسفورمر و PiT، پتانسیل توجه مبتنی بر پنجره و pooling base را نشان داده‌اند. تحقیقات بیشتر در این راستا برای استقرار ترانسفورمرهای بینایی در محیط‌های محدود به منابع و فعال‌کردن کاربرد آن‌ها در مجموعه داده‌های بزرگ‌تر و پیچیده‌تر ضروری خواهد بود.

7-2 کاوش در معماری‌های جدید و مکانیسم‌های توجه

فراتر از بهبود کارایی، محققان در حال بررسی معماری‌های کاملاً جدید و مکانیسم‌های توجه برای ترانسفورمرهای بینایی هستند. این شامل بررسی روش‌های مختلف ترکیب CNN و ترانسفورمر، کاوش در اشکال جدید توجه است که می‌تواند انواع مختلفی از روابط را در تصاویر جلب کند و توسعه معماری‌های سلسله مراتبی که می‌توانند ویژگی‌ها را در مقیاس‌های مختلف یاد بگیرند. هدف این کاوش‌ها این است که مرزهای آنچه را که با ترانسفورمرهای بینایی امکان‌پذیر است پیش ببرد و قابلیت‌های جدید را باز کند.

7-3 ترانسفورمرهای دید چندوجهی

ادغام ترانسفورمرهای بینایی با سایر روش‌ها، مانند متن، صدا یا اطلاعات عمقی، یک جهت تحقیقاتی امیدوارکننده است. ViTهای چندوجهی می‌توانند با ترکیب اطلاعات از منابع متعدد، بازنمایی‌های غنی‌تری را بیاموزند. این به ویژه برای کارهایی که نیاز به درک رابطه بین روش‌های مختلف دارند، مانند پاسخ‌گویی بصری به سؤال، شرح تصویر، یا روباتیک مرتبط است.

7-4 یادگیری خود نظارتی با آن‌هاترانسفورمرهای بینایی

یادگیری خود نظارتی به عنوان یک تکنیک قدرتمند برای آموزش ترانسفورمرهای بینایی بدون تکیه بر برچسب‌های صریح ظاهر شده است. روش‌هایی مانند مدل‌سازی تصویر ماسک‌دار (BEiT، MAE) موفقیت چشمگیری در یادگیری بازنمایی‌های بصری غنی از تصاویر بدون برچسب نشان داده‌اند. تحقیقات بیشتر در زمینه یادگیری خود نظارتی برای بازگشایی پتانسیل کامل ViTها و قادر ساختن آن‌ها به یادگیری از مقادیر گسترده داده‌های بدون برچسب در دسترس بسیار مهم خواهد بود. توسعه روش‌های یادگیری خود نظارتی جدید و بهبودیافته برای ترانسفورمرهای بینایی یک حوزه فعال تحقیقاتی است.

7-5 قابلیت توضیح و تفسیر آن‌هاترانسفورمرهای بینایی

همانطور که ViTها پیچیدتر و قدرتمندتر می‌شوند، درک نحوه تصمیم‌گیری آن‌ها اهمیت فزاینده‌ای پیدا می‌کند. هدف تحقیق در مورد توضیح‌پذیری و تفسیرپذیری توسعه تکنیک‌هایی برای تجسم و درک الگوهای توجه آموخته شده توسط ترانسفورمرهای بینایی است. این می‌تواند بینش‌هایی را در مورد اینکه کدام بخش از تصویر برای پیش‌بینی‌های مدل مهم است و به شناسایی سوگیری‌ها یا محدودیت‌های بالقوه کمک می‌کند. بهبود قابلیت توضیح ترانسفورمرهای بینایی برای ایجاد اعتماد در این مدل‌ها و استقرار مسئولانه آن‌ها در برنامه‌های کاربردی دنیای واقعی بسیار مهم است.

فصل 8: نتیجه‌گیری

ترانسفورمرهای بینایی به عنوان یک معماری قدرتمند و همه کاره در زمینه بینایی کامپیوتر ظاهر شده است که عملکرد قابل توجهی را در طیف گسترده‌ای از وظایف نشان می‌دهد. با تطبیق معماری ترانسفورمر از پردازش زبان طبیعی، ViTها نشان داده‌اند که مکانیسم‌های مبتنی بر توجه می‌توانند به طور موثر وابستگی‌های دوربرد و زمینه سراسری را در تصاویر ثبت کنند و بر برخی از محدودیت‌های شبکه‌های عصبی کانولوشنال سنتی غلبه کنند. از طبقه‌بندی تصویر و تشخیص اشیا گرفته تا تقسیم‌بندی معنایی و تجزیه و تحلیل ویدئو، آن‌هاترانسفورمرهای بینایی به نتایج پیشرفته‌ای دست یافته‌اند و امکانات جدیدی را برای درک تصویر باز کرده‌اند.

توسعه معماری‌های مختلف ترانسفورمرهای بینایی، از جمله DeiT، BEiT، PiT، Swinترانسفورمر و غیره، چالش‌های کلیدی مانند کارایی داده، هزینه محاسباتی و مدیریت تصاویر با وضوح بالا را برطرف کرده است. این پیشرفت‌ها ترانسفورمرهای بینایی را برای برنامه‌های کاربردی دنیای واقعی در دسترس‌تر و کاربردی‌تر کرده است. تحقیقات در حال انجام در یادگیری خود نظارتی، مکانیسم‌های توجه جدید و ادغام چندوجهی نویدبخش افزایش بیشتر قابلیت‌های ViT و گسترش کاربردهای آن‌ها به مشکلات پیچیده‌تر و چالش‌برانگیزتر است.

در حالی که چالش‌هایی مانند بهبود کارایی و مقیاس‌پذیری، افزایش قابلیت توضیح و پرداختن به استحکام دشمنان باقی مانده است، پیشرفت‌های حاصل در ترانسفورمرهای بینایی قابل توجه بوده است. آن‌ها نشان‌دهنده یک تغییر پارادایم قابل توجه در بینایی کامپیوتر هستند که قدرت معماری ترانسفورمر را فراتر از قلمرو پردازش زبان طبیعی نشان می‌دهد. با ادامه تحقیقات و ظهور نوآوری‌های جدید، ترانسفورمرهای بینایی نقش مهمی در شکل دادن به آینده بینایی رایانه ایفا می‌کند و امکان پیشرفت‌های جدید در درک تصویر و هوش مصنوعی را فراهم می‌کند. آن‌ها نه تنها به عملکرد قابل توجهی دست یافته‌اند، بلکه درک عمیق‌تری از نحوه استفاده از مکانیسم‌های توجه برای داده‌های بصری ایجاد کرده‌اند و راه را برای تحقیق و توسعه آینده در این زمینه هموار می‌کنند.

فصل 9: منابع

[1] “Lecun98.pdf.” Accessed: Feb. 08, 2025. [Online]. Available: http://vision.stanford.edu/cs598_spring07/papers/Lecun98.pdf
[2] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.
[3] M. Tan and Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” 2019, doi: 10.48550/ARXIV.1905.11946.
[4] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” 2020, *arXiv*. doi: 10.48550/ARXIV.2010.11929.
[5] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” 2020, *arXiv*. doi: 10.48550/ARXIV.2012.12877.
[6] H. Bao, L. Dong, S. Piao, and F. Wei, “BEiT: BERT Pre-Training of Image Transformers,” 2021, *arXiv*. doi: 10.48550/ARXIV.2106.08254.
[7] B. Heo, S. Yun, D. Han, S. Chun, J. Choe, and S. J. Oh, “Rethinking Spatial Dimensions of Vision Transformers,” 2021, *arXiv*. doi: 10.48550/ARXIV.2103.16302.
[8] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, “Going deeper with Image Transformers,” Apr. 07, 2021, *arXiv*: arXiv:2103.17239. doi: 10.48550/arXiv.2103.17239.
[9] Z. Liu *et al.*, “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada: IEEE, Oct. 2021, pp. 9992–10002. doi: 10.1109/ICCV48922.2021.00986.
[10] W. Kim, B. Son, and I. Kim, “ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision,” 2021, *arXiv*. doi: 10.48550/ARXIV.2102.03334.
[11] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked Autoencoders Are Scalable Vision Learners,” 2021, *arXiv*. doi: 10.48550/ARXIV.2111.06377.
[12] M. Caron *et al.*, “Emerging Properties in Self-Supervised Vision Transformers,” 2021, *arXiv*. doi: 10.48550/ARXIV.2104.14294.
[13] C.-F. Chen, Q. Fan, and R. Panda, “CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification,” 2021, *arXiv*. doi: 10.48550/ARXIV.2103.14899.