



دانشگاه صنعتی امیرکبیر  
(پلی تکنیک تهران)  
دانشکده مهندسی کامپیوتر

گزارش درس بینایی ماشین

معرفی ویژن ترنسفورمر و نسخه‌های بهبود یافته آن

نگارش  
فردین آیار

استاد درس  
دکتر رضا صفابخش

دی ۱۴۰۰



## چکیده

بینایی ماشین و پردازش زبان‌های طبیعی به عنوان دو حوزه مهم از یادگیری ماشین سال‌ها در مسیرهای متفاوتی قرار داشتند؛ اگرچه همواره در حال تبادل ایده‌ها و پیشرفت‌ها بوده‌اند. در پردازش زبان‌های طبیعی ترنسفورمر به عنوان شبکه پیشفرض در سال‌های اخیر این حوزه را با پیشرفت‌های چشمگیری همراه کرده‌است. در طرف مقابل شبکه‌های کانولوشنی، سال‌های طولانی‌تری به عنوان استاندارد بدون رقیب در بینایی ماشین حضور داشته‌اند. تحت تاثیر جریان رایج در حوزه پردازش زبان‌های طبیعی، در سال‌های اخیر تلاش‌های بسیاری برای استفاده از ساختار ترنسفورمر در حوزه بینایی ماشین صورت گرفت. در حالی که تلاش‌های اولیه به دنبال ترکیب لایه‌های کانولوشنی با مکانیزم خود-توجه یا استفاده از آن‌ها در ترنسفورمر بودند، شبکه ویژن ترنسفورمر، با حذف کامل لایه‌های کانولوشنی، جریان بسیار بزرگی در این حوزه آغاز کرد. علارغم موفقیت ویژن ترنسفورمر در مسئله دسته‌بندی تصاویر در دیتاست‌های بزرگ، این شبکه در دیتاست‌های متوسط، ضعیف‌تر از شبکه‌های کانولوشنی عمل می‌کرد. خوشبختانه پتانسیل بالای ویژن ترنسفورمر و نو بودن ایده آن، محققین را به پیشرفت‌های آینده آن امیدوار کرد. به همین دلیل به سرعت پیشنهادات بسیاری برای بهبود آن در مقالات مختلف چاپ شد. به طور موازی این پیشنهادات در مسائل مختلفی از بینایی ماشین مانند تشخیص اشیا، قطعه‌بندی تصاویر و... استفاده شد که نتایج مطلوب آن روز به روز بر محبوبیت ویژن ترنسفورمر افزوده است. با توجه به اهمیت روزافزون ویژن ترنسفوررها، هدف از این گزارش معرفی ویژن ترنسفورمر و برخی از نسخه‌های بهبود یافته آن است. در این گزارش ایده‌ها و نتایج هر شبکه بررسی و با شبکه‌های کانولوشنی مقایسه خواهند شد. همچنین به دلایلی که خواهیم دید تمرکز اصلی بر روی مسئله دسته‌بندی تصاویر است، اما اشارات کوتاهی به مسائل پیشرفته‌تر مانند تشخیص اشیا نیز خواهیم داشت.

| صفحه | فهرست مطالب  |
|------|--|
| أ    | چکیده.....   |
| ۱    | فصل اول مقدمه.....                                     |
| ۶    | فصل دوم معرفی ترنسفورمر.....                           |
| ۷    | ۱-۲- ساختار کلی.....                                   |
| ۸    | ۲-۲- ماژول خود-توجه.....                               |
| ۹    | ۳-۲- ماژول فیدفوروارد.....                             |
| ۱۰   | ۴-۲- بخش کدگذار.....                                   |
| ۱۰   | ۵-۲- بخش کدگشا.....                                    |
| ۱۲   | فصل سوم ویژن ترنسفورمر.....                            |
| ۱۳   | ۱-۳- ساختار کلی.....                                   |
| ۱۵   | ۲-۳- مقایسه دقت با شبکه‌های کانولوشنی.....             |
| ۱۷   | ۳-۳- بهبود ویژن ترنسفورمر.....                         |
| ۱۸   | ۱-۳-۳- بهبود فرآیند آموزش.....                         |
| ۲۰   | ۲-۳-۳- استفاده از رویکرد فشرده‌سازی دانش.....          |
| ۲۱   | ۴-۳- مقایسه هزینه محاسباتی با شبکه‌های کانولوشنی.....  |
| ۲۳   | فصل چهارم تغییرات ساختاری در ویژن ترنسفورمر.....       |
| ۲۴   | ۱-۴- بهبود فرآیند تبدیل تصویر به توالی توکن.....       |
| ۲۷   | ۱-۱-۴- نتایج.....                                      |
| ۲۷   | ۲-۴- ویژن ترنسفورمر چند مقیاسی.....                    |
| ۲۹   | ۱-۲-۴- مکانیزم خود-توجه تجمیعی.....                    |
| ۳۰   | ۲-۲-۴- ساختار شبکه.....                                |
| ۳۱   | ۳-۲-۴- نتایج.....                                      |
| ۳۱   | ۳-۴- محدود کردن مکانیزم خود-توجه: اسوین ترنسفورمر..... |
| ۳۴   | ۱-۳-۴- نتایج.....                                      |
| ۳۶   | فصل پنجم جمع‌بندی و نتیجه‌گیری.....                    |
| ۳۷   | ۱-۵- نتیجه‌گیری.....                                   |
| ۳۸   | منابع و مراجع.....                                     |

|  |    |
|--|----|
| شکل ۱- ساختار ترنسفورمر [۸].....   | ۸  |
| شکل ۲-ساختار کلی ویژن ترنسفورمر [۱۳].....  | ۱۴ |
| شکل ۳-مقایسه معماری‌های مختلف شبکه ویژن ترنسفورمر و یک شبکه کانولوشنی [۱۳].....                      | ۱۶ |
| شکل ۴-ساختار شبکه ایمپج ترنسفورمر کارا به همراه رویکرد فشرده‌سازی دانش [۱۵].....                     | ۲۰ |
| شکل ۵-مقایسه صفحات ویژگی دو مدل ویژن ترنسفورمر و یک شبکه کانولوشنی در لایه‌های مختلف [۱۶].....       | ۲۴ |
| شکل ۶-ساختار کلی شبکه T2T-ViT [۱۶].....  | ۲۵ |
| شکل ۷-فرآیند ترکیب توکن‌ها [۱۶].....   | ۲۶ |
| شکل ۸-مکانیزم خود-توجه تجمیعی [۱۷].....  | ۲۸ |
| شکل ۹-ساختار شبکه ویژن ترنسفورمر چند مقیاسی [۱۷].....  | ۲۹ |
| شکل ۱۰-مقایسه مشخصات شبکه ویژن ترنسفورمر چند مقیاسی (راست) و ویژن ترنسفورمر استاندارد (چپ) [۱۷]..... | ۳۰ |
| شکل ۱۱-ساختار سلسه مراتبی اسوین ترنسفورمر در مقایسه با ویژن ترنسفورمر استاندارد [۱۷].....            | ۳۲ |
| شکل ۱۲-یک نمونه از معماری شبکه اسوین ترنسفورمر [۱۷].....   | ۳۲ |
| شکل ۱۳-رویکرد پنجره انتقالی در اسوین ترنسفورمر [۱۷].....   | ۳۳ |

## صفحه

## فهرست جداول

|   |    |
|---|----|
| جدول ۱- معماری‌های مختلف شبکه ویزن ترنسفورمر [۱۳].....  | ۱۶ |
| جدول ۲- جزئیات و ابرپارامترهای شبکه ایمپج ترنسفورمر کارا در مقایسه با ویزن ترنسفورمر [۱۵].....            | ۱۸ |
| جدول ۳- مقایسه شبکه ایمپج ترنسفورمر کارا، ویزن ترنسفورمر و شبکه‌های کانولوشنی [۱۵].....                   | ۱۹ |
| جدول ۴- جزئیات ساختار شبکه T2T-ViT در مقایسه با ویزن ترنسفورمر استاندارد [۱۶].....                        | ۲۶ |
| جدول ۵- مقایسه نتایج شبکه T2T-ViT با ساختارهای استاندارد ویزن ترنسفورمر در دیتاست ImageNet [۱۶].....      | ۲۷ |
| جدول ۶- نتایج شبکه ویزن ترنسفورمر چند مقیاسی در مقایسه با ساختار استاندارد و شبکه‌های کانولوشنی [۱۷]..... | ۳۱ |
| جدول ۷- جزئیات معماری‌های مختلف شبکه اسوین ترنسفورمر [۱۷].....  | ۳۴ |
| جدول ۸- نتایج شبکه اسوین ترنسفورمر در مقایسه با ساختار استاندارد و شبکه‌های کانولوشنی [۱۷].....           | ۳۴ |
| جدول ۹- نتایج شبکه اسوین ترنسفورمر به عنوان استخراج‌کننده ویژگی [۱۷].....                                 | ۳۵ |

## فصل اول

### مقدمه

دو دهه پس معرفی شبکه الکسنت<sup>۱</sup> [۱] و عملکرد فوق العاده آن در مسئله دسته‌بندی تصاویر، شبکه‌های عصبی کانولوشنی<sup>۲</sup> همچنان جریان اصلی در حوزه بینایی ماشین هستند. تحقیقات بعدی در شبکه‌هایی کانولوشنی با ساختار عمیق‌تر [۲]، لایه‌های پیچیده‌تر [۳] و معماری هوشمندانه‌تر [۴] موجب شده شبکه‌های کانولوشنی به عنوان استخراج کننده ویژگی بسیار قدرتمند، در همه کارهای بینایی ماشین برقیب باشد [۵]. فراتر از این، شبکه‌های کانولوشنی در بسیاری از حوزه‌های دیگر مانند پردازش سیگنال و حتی پردازش زبان‌های طبیعی<sup>۳</sup> به صورت گسترده مورد استفاده قرار می‌گیرند.

به طور موازی در حوزه پردازش زبان، جریان حاکم در سال‌های اخیر بسیار متفاوت بوده‌است. شبکه ترنسفورمر<sup>۴</sup> [۶] با ساختار انقلابی خود و به ویژه مکانیزم خود-توجه<sup>۵</sup>، رقیب‌های خود در حوزه پردازش زبان، شبکه‌های بازگشتی و شبکه‌های کانولوشنی را کنار زده و به شبکه استاندارد در این حوزه تبدیل شده‌است. ترنسفورمر نیز به لطف روش‌های پیش‌آموز [۷] و بهبودهای حافظه، کارایی و سرعت [۸] ضمن حفظ برتری خود، همچنان در حال پیشرفت است.

برجسته‌ترین ویژگی ترنسفورمرها را می‌توان لایه‌های خود-توجه موجود در آن‌ها دانست. با بهره‌گیری از مکانیزم توجه بین داده‌های ورودی، مکانیزم خود-توجه می‌تواند روابط ضمنی بین داده‌های ورودی را درک و با توجه به آن ویژگی‌های پیچیده‌تری استخراج کند. بنابراین جای شگفتی نیست که این مکانیزم به سرعت به سایر حوزه‌های یادگیری ماشین از جمله بینایی ماشین راه یافت. چالش اصلی خود-توجه اما، هزینه بالای محاسباتی آن برای ورودی‌های حجیم می‌باشد و متأسفانه در بینایی ماشین، تقریباً همیشه داده‌های ورودی حجیم است. یک راه ساده برای حل این مشکل استفاده از خود-توجه در لایه‌های انتهایی شبکه‌های کانولوشنی است. در این لایه‌ها معمولاً تعداد بردارهای ویژگی بسیار کمتر است و در نتیجه استفاده از این مکانیزم هزینه کمتری دارد [۹]. به طور کلی استفاده از لایه‌های خود-توجه به عنوان مکمل لایه‌های کانولوشنی پیش از این در کارهای مختلفی مورد استفاده قرار گرفته‌است؛ اما استفاده از لایه‌های

---

<sup>1</sup> AlexNet

<sup>2</sup> Convolutional neural networks

<sup>3</sup> Natural language processing

<sup>4</sup> Transformers

<sup>5</sup> Self-Attention



خود-توجه به تنهایی، همانطور که گفته شد به دلیل حجم بالای ورودی تصویر، در ابتدا منطقی به نظر نمی‌رسید. برای حل این مشکل، پیشنهاد محققین محدود کردن میدان دید هر پیکسل در فرآیند محاسبه خود-توجه بود [۱۰]. این کار، مشابه آنچه در شبکه‌های کانولوشنی انجام می‌شود، اگرچه حجم محاسبات را به شدت کاهش می‌دهد؛ اما ویژگی اصلی خود-توجه، توانایی برقراری ارتباط بین ورودی(پیکسل)های دورتر را نیز محدود می‌کند. استفاده از مکانیزم خود-توجه پراکنده<sup>۶</sup> و توجه-محوری<sup>۷</sup> [۹] راه‌های پیشنهادی دیگر برای کاهش حجم محاسبات بودند که سعی در حفظ میدان دید پیکسل‌ها دارند.

در جریانی کاملاً متفاوت، برخی از کارها از ساختار کامل ترنسفورمر(نه فقط لایه خود-توجه آن) در بینایی ماشین استفاده کردند. مطرح‌ترین کار در این زمینه، شبکه پیشنهادی کاریون و همکاران [۱۱]، از شبکه ترنسفورمر به صورت هوشمندانه‌ای برای تشخیص اشیاء استفاده می‌کند. این ایده سپس در کارهای دیگر مانند قطعه‌بندی نیز استفاده شد [۱۲]. این شبکه‌ها نیز به علت حجم بالای محاسبات، از لایه‌های کانولوشنی به عنوان شبکه شالوده<sup>۸</sup> ترنسفورمر [۱۱] یا در ترکیب با آن [۱۲] استفاده می‌کنند؛ بنابراین ساختار آن‌ها با ساختار اولیه ترنسفورمر متفاوت است.

چیزی که شبکه ویژن ترنسفورمر [۱۳] را از همه کارهای پیشین مجزا می‌کند، استفاده از ساختار کامل بخش کدگذار<sup>۹</sup> ترنسفورمر، تقریباً بدون هیچ تغییری است. جالب‌تر اینکه این شبکه از هیچ لایه کانولوشنی نیز در سراسر ساختار خود استفاده نمی‌کند و برای مشکل حجم بالای تصاویری رویکرد دیگری را در پیش می‌گیرد. نتایج این شبکه در مسئله دسته‌بندی تصاویر نشان داد اگرچه در دیتاست‌های کوچک و متوسط عملکرد ضعیف‌تری در مقایسه با شبکه‌های کانولوشنی دارد؛ اما در دیتاست‌های بزرگ این شبکه حتی از آنها بهتر عمل می‌کند. ویژن ترنسفورمر علاوه بر مشکلاتی که دارد، در دو سال اخیر جریان کاملاً جدیدی در بینایی ماشین شروع کرد که رقابت آن را با لایه‌های کانولوشنی به تدریج نزدیک‌تر می‌کند. کارهای جدیدتر، سعی در حل مشکلات ویژن ترنسفورمر به ویژه نیاز آن به داده‌های آموزشی زیاد، دارند و هم

---

<sup>۶</sup> Sparse

<sup>۷</sup> Axial-Attention

<sup>۸</sup> منظور از شبکه شالوده، استخراج کننده ویژگی، به عنوان ترجمه اصطلاح Backbone Network است.

<sup>۹</sup> Encoder

اکنون ایده استفاده از ویژن ترنسفورمر به عنوان شبکه شالوده جایگزین شبکه‌های کانولوشنی، بسیار مورد استقبال بوده است [۱۴].

هدف از این گزارش معرفی ویژن ترنسفورمر و نسخه‌های بهبود یافته آن به صورت مختصر می‌باشد. بدین منظور شبکه ویژن ترنسفورمر را به تفصیل شرح خواهیم داد و خواهیم دید آزمایشات داسایوفسکی و همکاران [۱۳] چگونه پتانسیل بالای ویژن ترنسفورمر را نشان می‌دهد. شبکه آن‌های که تقریباً بدون تغییر مشابه ساختار استاندارد است، تصاویر ورودی را به صورت مجموعه‌ای از توکن‌هایی می‌بیند که حاصل افکنش وصله‌های دوبعدی تصویر به فضای یک بعدی هستند. این شبکه اگرچه در مسئله دسته‌بندی تصاویر در دیتاست‌های بزرگ می‌تواند بهتر از شبکه‌های کانولوشنی عمل کند اما در دیتاست‌های کوچک و متوسط، چیزی که در مسائل عملی رایج‌تر است، برتری خاصی نسبت به آن‌ها ندارد. مدتی بعد توفرون و همکاران [۱۵] با بررسی دقیق‌تر فرآیند آموزش ویژن ترنسفورمرها، نشان دادند بخش زیادی از این ضعف را می‌توان تنها با عوض کردن پارامترهای آموزش شبکه جبران کرد. تغییر دیگر آن‌ها استفاده از رویکرد فشرده‌سازی دانش<sup>۱۰</sup> بود که می‌تواند عملکرد شبکه را بیش‌ازپیش بهبود دهد.

برای بهبود ویژن ترنسفورمر در دیتاست‌های کوچک و متوسط، محققان به سمت ایجاد تغییرات ساختاری در آن متمایل شده‌اند. جنبه مشترک این تغییرات را می‌توان به نوعی افزودن دانش پیشین به شبکه برای تطابق آن با حوزه تصویر دانست. با فرض اینکه فرآیند تبدیل تصویر به توالی توکن‌ها در ویژن ترنسفورمر استاندارد، مناسب حوزه تصویر نیست؛ یان و همکاران [۱۶] روش جدیدی برای اینکار ارائه می‌کنند که شکاف بین ویژن ترنسفورمرها و شبکه‌های کانولوشنی در دیتاست‌های متوسط را تا حد زیادی پر می‌کند. در رویکردی متفاوت، فان و همکاران [۱۷]، با ایجاد ساختاری چند مقیاسی در ویژن ترنسفورمر توانستند نتایج آن را بیش از پیش بهبود دهند. ایده آن‌ها این بود که ثابت نگه داشتن تعداد و بُعد توکن‌ها در کل شبکه در حوزه تصویر مناسب نیست، بنابراین از ساختاری سلسله مراتبی مانند شبکه‌های کانولوشنی استفاده کردند.

متاسفانه رزلوشن پایین ویژن ترنسفورمرها امکان استفاده از آن‌ها را به عنوان استخراج کننده ویژگی تصاویر در مسائل پیشرفته‌تر، مانند تشخیص اشیاء، از بین می‌برد. برای رفع این مشکل لیو و همکاران

<sup>10</sup> Knowledge distillation

[۱۸] با محدود کردن مکانیزم خود-توجه و رویکردهای مبتکرانه دیگر، شبکه‌ای معرفی کردند که نه تنها قابلیت استفاده از آن در مسائل پیشرفته‌تر وجود دارد؛ بلکه می‌تواند از شبکه‌های کانولوشنی نیز بهتر عمل کند.

ادامه این گزارش به این صورت سازمان دهی شده است: در فصل ۲ ساختار ترنسفورمر استاندارد در حوزه پردازش زبان‌های طبیعی به اختصار شرح داده می‌شود. در فصل ۳ ضمن معرفی ویژن ترنسفورمر [۱۳] و بررسی نتایج آن، تغییرات پیشنهادی توفرون و همکاران [۱۵] برای بهبود آن ارائه می‌شود. سه کار که به منظور تغییر در ساختار ویژن ترنسفورمر منتشر شده‌اند [۱۶-۱۸] به صورت مجزا در فصل ۴ معرفی می‌شوند. در نهایت فصل ۵ شامل مرور مختصری بر این گزارش و نتیجه‌گیری آن خواهد بود.

## فصل دوم

### معرفی ترنسفورمر

از آن جا که ساختار ویژن ترنسفورمر تقریباً بدون هیچ تغییری مشابه ساختار نسخه اولیه ترنسفورمر است، برای شناخت آن بهتر است ابتدا به معرفی ساختار ترنسفورمر بپردازیم. در سال‌های اخیر نسخه‌های متفاوتی از ترنسفورمر در زمینه‌های مختلف معرفی شده‌است؛ اما در این گزارش منظور از ترنسفورمر، نسخه اولیه آن است که در سال ۲۰۱۷ توسط واسوانی و همکاران [۶] برای حوزه پردازش زبان طبیعی و به طور خاص ترجمه ماشینی معرفی شد. در ادامه این فصل سعی خواهیم کرد ساختار ترنسفورمر را با جزئیات بیشتری ارائه کنیم.

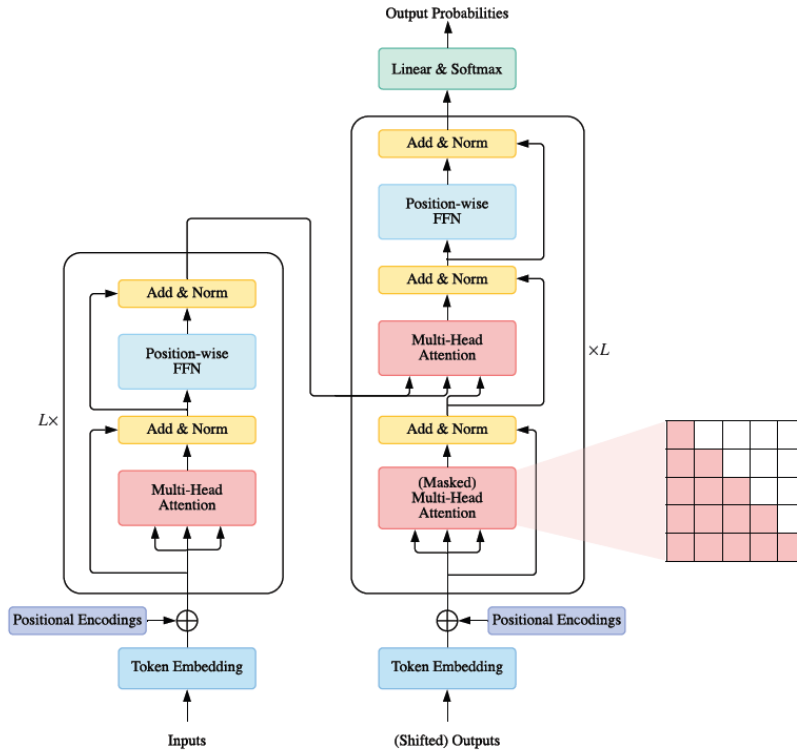
## ۲-۱- ساختار کلی

ترنسفورمر [۶] یک مدل برای تبدیل یک توالی ورودی به توالی خروجی متناظر با آن است. این شبکه شامل دو بخش کدگذار و کدگشا<sup>۱۱</sup> است که هر یک از  $L$  بلوک مشابه تشکیل شده‌اند. هر بلوک کدگذار شامل دو ماژول خود-توجه و شبکه فیدفوروارد<sup>۱۲</sup> است که برای تسهیل آموزش، از اتصال باقی‌مانده‌ای [۲] و نرمال‌سازی لایه‌ای [۱۹] به همراه هر ماژول استفاده شده‌است. در مقایسه با کدگذار، تنها تفاوت بخش کدگشا، وجود یک ماژول توجه اتصالی<sup>۱۳</sup> بین ماژول‌های خود-توجه و فیدفوروارد است. اولین ماژول خود-توجه در بخش کدگشا به گونه‌ای تنظیم شده که از توجه به کلمات بعدی توالی (که هنوز مقدار آن‌ها مشخص نشده) جلوگیری کند. این بخش در هر مرحله یک کلمه از خروجی را پیش‌بینی می‌کند. در شکل ۱ ساختار ترنسفورمر به همراه جزئیاتی که شرح داده شد ارائه شده‌است.

<sup>11</sup> Decoder

<sup>12</sup> Feed forward

<sup>13</sup> Cross-Attention



شکل ۱- ساختار ترنسفورمر [۸]

## ۲-۲- مازول خود-توجه

به ازای هر کلمه در توالی ورودی  $X \in \mathbb{R}^{N \times D}$ ، مکانیزم خود-توجه از جمع وزن دار کل مقادیر  $V$  توالی ورودی به عنوان بازنمایی جدید آن‌ها استفاده می‌کند.

$$SA(X) = AV \quad (1)$$

در رابطه فوق  $SA$  نشان دهنده عملیات خود-توجه و  $A$  که اصطلاحاً ماتریس توجه نامیده می‌شود از رابطه زیر بدست می‌آید.

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{D_h}}\right) \quad (2)$$

در روابط ۱ و ۲ ماتریس‌های  $Q \in \mathbb{R}^{N \times D_h}$ ،  $K \in \mathbb{R}^{N \times D_h}$  و  $V \in \mathbb{R}^{N \times D_v}$  به ترتیب ماتریس جستار<sup>۱۴</sup>، ماتریس کلید<sup>۱۵</sup> و ماتریس مقدار<sup>۱۶</sup> نامیده می‌شوند. این سه ماتریس به ترتیب از ضرب ماتریس توالی ورودی  $X$  در سه ماتریس  $W_Q \in \mathbb{R}^{D \times D_h}$ ،  $W_K \in \mathbb{R}^{D \times D_h}$  و  $W_V \in \mathbb{R}^{D \times D_v}$  به وجود می‌آید. سه ماتریس اخیر به عنوان پارامترهای شبکه در حین آموزش یاد گرفته خواهند شد.

آزمایشات واسوانی و همکاران [۶] نشان داد استفاده از چند مکانیزم خود-توجه به صورت همزمان به شبکه اجازه می‌دهد با افکنش توالی ورودی به زیرفضاهای مختلف، جنبه‌های مختلف شبکه را بهتر یاد بگیرد. بنابراین ماژول خود-توجه را به صورت چندشاخه‌ای<sup>۱۷</sup> و به صورت زیر تعریف کردند:

$$MSA(X) = [SA_1(X), SA_2(X), \dots, SA_h(X)]W_0 \quad (3)$$

در رابطه فوق  $MSA$  نشان‌دهنده عملیات خود-توجه چندشاخه‌ای،  $SA_i$  نماینده مکانیزم خود-توجه شاخه  $i$ -ام،  $[.]$  عمل الحاق،  $h$  تعداد شاخه‌های خود-توجه و  $W_0 \in \mathbb{R}^{hD_h \times D}$  ماتریسی افکنش برای بدست آوردن خروجی نهایی است. ماتریس  $W_0$  نیز به عنوان پارامتر شبکه، در حین آموزش یاد گرفته خواهد شد.

در ترنسفورمر  $D_v = D_h = \frac{h}{8}$  و  $D=512$  تنظیم شده است اما به صورت کلی محدودیتی در مورد این ابعاد وجود ندارد. با توجه به روابط بالا  $A \in \mathbb{R}^{N \times N}$ ،  $SA(X) \in \mathbb{R}^{N \times D_h}$  و  $MSA(X) \in \mathbb{R}^{N \times D}$ ؛ بنابراین مشکلی از نظر استفاده از اتصال باقی‌مانده‌ای وجود نخواهد داشت.

## ۲-۳- ماژول فیدفوروارد

ماژول فیدفوروارد به سادگی یک شبکه فیدفوروارد دولایه است که بعد از هر لایه یک تابع فعالسازی ریلو<sup>۱۸</sup> اعمال می‌شود. این شبکه روی هر کلمه از توالی ورودی به صورت مجزا اعمال می‌شود اما وزن‌های آن بین همه کلمات مشترک است؛ به همین دلیل می‌توان این دو لایه را مانند دو لایه کانولوشنی  $1 \times 1$

<sup>14</sup> Query

<sup>15</sup> Key

<sup>16</sup> Value

<sup>17</sup> Multi-head

<sup>18</sup> ReLU

در نظر گرفت. به دلیل استفاده از اتصال باقی مانده‌ای، ابعاد خروجی این شبکه با ورودی آن برابر است اما لایه پنهان آن در ترنسفورمر دارای بُعد 4D می‌باشد.

## ۲-۴- بخش کدگذار

پس از معرفی ماژول‌های اصلی ترنسفورمر، در این قسمت بخش کدگذار را که شامل ۶ بلاک کدگذار مشابه است، مجدداً و با جزئیات بیشتر ارائه می‌کنیم. به طور رسمی ورودی  $X \in \mathbb{R}^{N \times D}$  که شامل N کلمه پس از اعمال امبدینگ<sup>۱۹</sup> است، پس از ورود به بخش کدگذار ترنسفورمر، فرآیند زیر را طی می‌کند:

$$Z_0 = X + E_{pos} \quad (4)$$

$$R_l = LN(MSA(Z_{l-1}) + Z_{l-1}) \quad l = 1, 2, \dots, L \quad (5)$$

$$Z_l = LN(FF(R_l) + R_l) \quad l = 1, 2, \dots, L \quad (6)$$

در رابطه بالا L تعداد بلاک‌های کدگذار ( $L=6$ )، LN نرمال‌سازی لایه‌ای [۱۹] و FF نماینده ماژول فیدفوروارد است. از آنجا که ساختار ترنسفورمر درکی از مکان کلمات در توالی ورودی ندارد،  $E_{pos}$  که نشان‌دهنده ماتریس امبدینگ مکانی<sup>۲۰</sup> است در ابتدای بخش کدگذار به ماتریس ورودی اضافه می‌شود. امبدینگ مکانی می‌تواند به صورت ثابت باشد یا در حین آموزش شبکه یادگرفته شد. در هر دو صورت این ماتریس شبکه را از محل کلمات در جمله آگاه می‌کند و از این طریق عملکرد شبکه را بسیار بهبود می‌دهد. در نهایت خروجی کدگذار ریال  $Z_L$ ، به تمام بلاک‌های بخش کدگشا ارسال می‌شود.

## ۲-۵- بخش کدگشا

همانطور که در فصل آینده خواهیم دید، شبکه ویژن ترنسفورمر تنها از بخش کدگذار ترنسفورمر به عنوان استخراج‌کننده ویژگی از تصاویر استفاده خواهد کرد؛ بنابراین در این گزارش از ارائه جزئیات این بخش صرف‌نظر می‌کنیم. به هر حال برای تکمیل بحث ترنسفورمر و حفظ جامعیت گزارش، در این بخش به صورت کلی ساختار بخش کدگذار شرح داده می‌شود.

<sup>19</sup> Embedding

<sup>20</sup> Position embeddings



بخش کدگشا مانند بخش کدگذار از ۶ بلاک کدگشا تشکیل شده است و در هر مرحله یک کلمه از خروجی را پیش‌بینی می‌کند. کلمه خروجی هر مرحله سپس برای ادامه پیش‌بینی به عنوان ورودی به بخش کدگشا داده می‌شود؛ بنابراین برای حفظ خاصیت خودبرگشتی<sup>۲۱</sup> این بخش، ماژول‌های خود توجه آن به نحوی تغییر یافته که از توجه به کلماتی آینده جلوگیری شود. این کار به سادگی با قراردادن مقدار منفی بی‌نهایت در درایه‌های مربوطه در ماتریس توجه انجام می‌شود.

تفاوت دیگر بخش کدگشا، وجود ماژول توجه اتصالی در هر بلاک برای ارتباط بخش‌های کدگذار و کدگشا است. ساختار این ماژول کاملاً مشابه ماژول خود-توجه است با این تفاوت که ماتریس کلید و ماتریس مقدار آن به جای اینکه از روی ورودی خود ماژول بدست آید، از خروجی کدگذار بدست می‌آید. سایر جزئیات این بخش مشابه بخش کدگذار است. (شکل ۱)

---

<sup>21</sup> Auto-regressive

## فصل سوم

### ویژن ترنسفورمر

همانطور که در مقدمه این گزارش نیز تاکید شد، شبکه ای که تحت عنوان ویژن ترنسفورمر [۱۳] شناخته می‌شود، اولین شبکه‌ای نیست که از مکانیزم خود-توجه و یا ساختار ترنسفورمر در حوزه بینایی ماشین استفاده می‌کند؛ اما چیزی که این شبکه را از سایر کارها متمایز می‌کند، ایده جسورانه استفاده از ساختار اصلی ترنسفورمر بدون تغییر و کنار گذاشتن لایه‌های کانوشولنی به طور کامل است. ایده‌ای که در ابتدا ممکن است چندان مناسب به نظر نرسد، اما همانطور که در ادامه این گزارش خواهیم دید، بسیار موفق بوده‌است به طوری که در دو سال اخیر جریان بزرگی را در حوزه بینایی ماشین شروع کرده‌است. در این فصل علاوه بر معرفی شبکه اولیه ویژن ترنسفورمر، خواهیم دید توفرون و همکاران [۱۵] چگونه با حداقل تغییرات ممکن، عملکرد آن را در دیتاست‌های متوسط بهبود دادند و پتانسیل ویژن ترنسفورمر را برای بهبودهای آینده ثابت کردند.

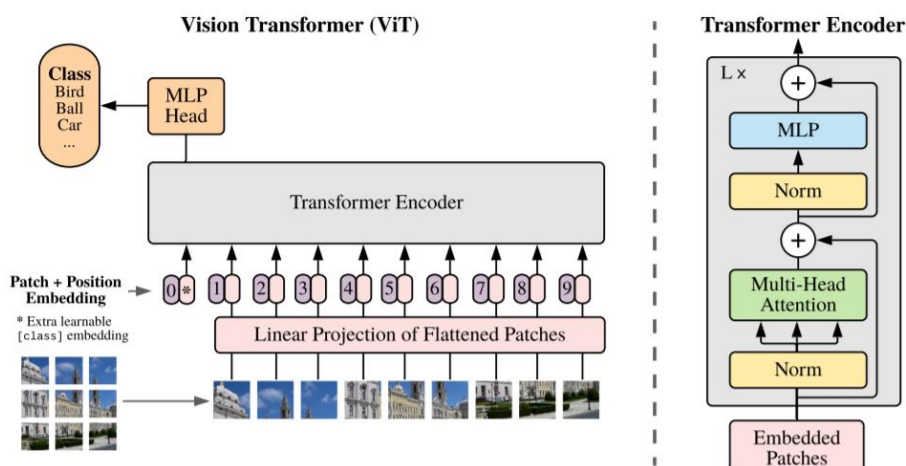
### ۳-۱- ساختار کلی

شاید کوتاه‌ترین توصیف برای ویژن ترنسفورمر را بتوان عنوان مقاله دسایوفسکی و همکاران [۱۳] دانست: ((یک تصویر معادل کلمات  $16 \times 16$  است))<sup>۱</sup>. ویژن ترنسفورمر یک نسخه از بخش کدگذار نسخه اصلی ترنسفورمر [۶] است که حداقل تغییرات ممکن را نسبت به آن دارد (شکل ۱). ورودی نسخه اصلی ترنسفورمر، یک توالی از امبدینگ کلمات (توکن‌ها)<sup>۲</sup> است. برای استفاده از آن در تصاویر، از وصله‌های  $16 \times 16$  تصویر استفاده می‌کنیم. این وصله‌ها دقیقاً معادل کلمات در ترنسفورمر هستند. به طور رسمی یک تصویر  $X \in \mathbb{R}^{H \times W \times C}$  به توالی از وصله‌ها  $X_p \in \mathbb{R}^{N \times (p^2 \cdot C)}$  تغییر شکل داده می‌شود که  $(H, W)$  رزولوشن تصویر،  $C$  تعداد کانال‌های تصویر،  $P$  سایز وصله‌ها و  $N = \frac{HW}{p^2}$  تعداد کل توکن‌های توالی ورودی است. سایز توکن‌ها در تمام لایه‌های ترنسفورمر مقدار ثابت  $D$  است؛ بنابراین وصله‌ها با اعمال یک تبدیل خطی به این سایز تبدیل می‌شوند ( $E$  در رابطه ۷). مشابه ترنسفورمر این توکن‌ها سپس با

<sup>۱</sup> An image is worth  $16 \times 16$  words

<sup>۲</sup> Token

<sup>۳</sup> Path



شکل ۲- ساختار کلی ویژن ترنسفورمر [۱۳]

امبدینگ مکانی یادگرفتنی<sup>۱</sup> جمع خواهد شد ( $E_{pos}$  در رابطه ۷). آزمایشات نشان داد که استفاده از امبدینگ مکانی دو بعدی برتری محسوسی نسبت به نوع یک بعدی آن ندارد؛ بنابراین امبدینگ مکانی استفاده شده از نوع یک بعدی است.

مشابه آنچه در پیش آموزش ترنسفورمر در حوزه پردازش زبان رایج است [۷]، یک توکن ثابت امبدینگ کلاس<sup>۲</sup> به ابتدای توالی توکن‌ها اضافه می‌شود. این توکن در طول لایه‌های ترنسفورمر از طریق ماژول‌های خود-توجه با سایر توکن‌ها تبادل اطلاعات می‌کند و در نهایت خروجی متناظر با آن به عنوان بازنمایی فشرده‌ای از تصویر برای دسته‌بندی استفاده می‌شود ( $Z_L^0$  در رابطه ۱۰). در نهایت کل فرآیند یک تصویر در ویژن ترنسفورمر را می‌توان در روابط زیر خلاصه کرد:

$$Z_0 = [x_{class}, x_p^1 E, x_p^2 E, \dots, x_p^N E] + E_{pos} \quad E \in \mathbb{R}^{(P^2 \cdot C) \times D}, E_{pos} \in \mathbb{R}^{(N+1) \times D} \quad (7)$$

$$R_l = MSA(LN(Z_{l-1}) + Z_{l-1}) \quad l = 1, 2, \dots, L \quad (8)$$

$$Z_l = FF(LN(R_l) + R_l) \quad l = 1, 2, \dots, L \quad (9)$$

$$y = LN(Z_L^0) \quad (10)$$

<sup>۱</sup> Learnable

<sup>۲</sup> Class embedding

در روابط بالا FF، MSA و LN مشابه روابط ۴ تا ۶ هستند و سایر علائم پیش از این در همین بخش توضیح داده شدند. همانطور که مشاهده می‌کنید بر خلاف نسخه اصلی ترنسفورمر، در اینجا نرمال‌سازی لایه‌ای در هر بلاک کدگذار پیش از ورود به ماژول‌ها اعمال شده‌اند.

به جای استفاده از تصویر به جای ورودی ترنسفورمر، مشابه آنچه در کارهای قبلی رایج بود می‌توان از نقشه‌های ویژگی بدست آمده از شبکه‌های کانولوشنی را به عنوان ورودی ترنسفورمر استفاده کرد. در این صورت با توجه به سائز نقشه‌ها می‌توان اندازه وصله‌ها را تعیین کرد. در حالت خاص این وصله‌ها می‌توانند  $1 \times 1$  باشند که در این صورت از هر بردار در نقشه ویژگی به عنوان یک کلمه استفاده خواهد شد. داسایوفسکی و همکاران [۱۳] مدل ترکیبی کانولوشن و ترنسفورمر را مدل هیبریدی<sup>۱</sup> نامیدند و در آزمایشات خود آن را با مدل پایه ویژن ترنسفورمر مقایسه کردند.

### ۳-۲- مقایسه دقت با شبکه‌های کانولوشنی

شبکه‌های کانولوشنی نسبت به شبکه ویژن ترنسفورمر از بایاس القایی بیشتری برخوردار هستند. در شبکه‌های کانولوشنی محلی بودن<sup>۲</sup>، ساختار همسایگی دو بعدی و مقاومت در برابر انتقال<sup>۳</sup> به صورت صریح در هر لایه لحاظ شده‌است [۲۰]. در طرف مقابل، تنها ماژول فیدفوروارد در ویژن ترنسفورمر تا حدی محلی است و ساختار همسایگی دو بعدی تنها در زمان تبدیل تصویر به وصله‌ها صریحاً در نظر گرفته شده‌است. علاوه بر این امبدینگ مکانی یادگرفتنی است و در ابتدای آموزش هیچ درکی از همسایگی توکن‌ها ندارد. مجموع این عوامل موجب می‌شود انتظار داشته باشیم شبکه‌های کانولوشنی در حوزه تصویر بهتر از دیگر ساختارهای شبکه‌های عصبی، از جمله ویژن ترنسفورمر عمل کند.

<sup>۱</sup> Hybrid

<sup>۲</sup> Locality

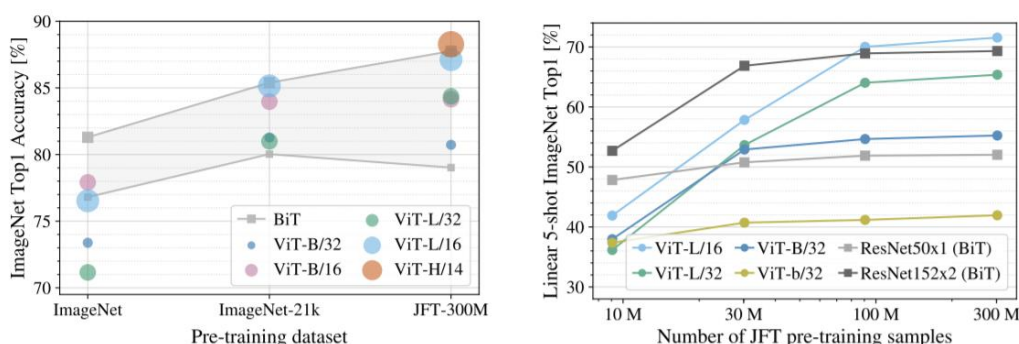
<sup>۳</sup> Translation Equivariance

جدول ۱- معماری‌های مختلف شبکه ویژن ترنسفورمر [۱۳]

| Model     | Layers | Hidden size $D$ | MLP size | Heads | Params |
|-----------|--------|-----------------|----------|-------|--------|
| ViT-Base  | 12     | 768             | 3072     | 12    | 86M    |
| ViT-Large | 24     | 1024            | 4096     | 16    | 307M   |
| ViT-Huge  | 32     | 1280            | 5120     | 16    | 632M   |

ویژن ترنسفورمر از ابتدا به طور خاص برای مسئله دسته‌بندی تصاویر معرفی و آزمایش شد. داسایوفسکی و همکاران [۱۳] معماری‌های مختلف از شبکه ویژن ترنسفورمر را با یک نسخه بهبود یافته از شبکه رزنت (BiT) [۲۱] مقایسه کردند. این معماری‌ها هر یک از نظر تعداد لایه‌ها، سائز وصله و بُعد لایه پنهان با هم متفاوت هستند (جدول ۱). به طور مثال منظور از ViT-L/16 شبکه ویژن ترنسفورمر Large با سائز وصله  $16 \times 16$  است.

در تصویر سمت چپ شکل ۳ دقت معماری‌های مختلف شبکه ویژن ترنسفورمر (نقاط رنگی) و شبکه BiT (ناحیه خاکستری) به ازای دیتاست‌های پیش‌آموزش<sup>۱</sup> مختلف مقایسه شده‌اند. مطابق شکل شبکه ویژن ترنسفورمر در حالتی که تنها روی دیتاست ImageNet آموزش دیده بسیار ضعیف‌تر از شبکه کانولوشنی عمل می‌کند. این نتیجه در نگاه اول اگرچه ناامیده‌کننده است؛ اما قابل پیش‌بینی بود. بایاس القایی در شبکه‌های کانولوشنی آن‌ها را به طور خاص در حوزه تصویر بسیار قدرتمند می‌کند. خوشبختانه ادامه آزمایشات نشان می‌دهد در صورت آموزش روی دیتاست بزرگ‌تر ImageNet-21k، شکاف بین ویژن ترنسفورمر و BiT از بین می‌رود اما هنوز عملکرد ویژن ترنسفورمر اندکی ضعیف‌تر است. روند صعودی افزایش دقت شبکه ویژن ترنسفورمر بر روی دیتاست بسیار بزرگ JFT-300M ادامه دارد و



شکل ۳- مقایسه معماری‌های مختلف شبکه ویژن ترنسفورمر و یک شبکه کانولوشنی [۱۳]

<sup>۱</sup> Pre-training

مطابق شکل در این حالت معماری Huge ویژن ترنسفورمر از معماری شبکه کانولوشنی متناظر آن (از نظر تعداد پارامتر)، پیشی می‌گیرد. تصویر سمت راست شکل ۳ نتایج مشابهی را این بار به ازای اندازه‌های مختلف دیتاست پیش‌آموزش JFT-300 نشان می‌دهد.

به طور خلاصه دیدیم که یادگیری مدل‌های ویژن ترنسفورمر با افزایش اندازه دیتاست پیش‌آموزش، به نسبت بیشتری افزایش می‌یابد و این نتیجه می‌تواند بسیار امیدوارکننده باشد. مقایسه بسیار مهم دیگری که باید انجام شود، مقایسه تعداد پارامترها و هزینه محاسباتی ویژن ترنسفورمر و شبکه‌های کانولوشنی است. پیش از انجام این مقایسه، در بخش بعد دو بهبود مهم روی شبکه ویژن ترنسفورمر که توسط توفرون و همکاران [۱۵] ارائه شد، شرح داده می‌شود.

### ۳-۳- بهبود ویژن ترنسفورمر

شبکه ویژن ترنسفورمر در دیتاست‌های بسیار بزرگ توانست از شبکه‌های کانولوشنی پیشی بگیرد. متأسفانه در کاربردهای عملی وجود این حجم از تصویر (حدود ۳۰۰ میلیون در دیتاست JFT) بسیار نادر و در صورت وجود آموزش آن بسیار مشکل است. با این وجود آیا ادامه مسیر ویژن ترنسفورمر منطقی است و چرا محققین نتایج ویژن ترنسفورمر را ((امیدوارکننده)) می‌دانند؟

در طی دو دهه حضور بدون رقیب شبکه‌های کانولوشنی در حوزه بینایی ماشین، تحقیقات بسیار زیادی برای بهینه‌سازی آن‌ها هم از نظر معماری [۴] و هم از نظر پارامترها و شیوه آموزش صورت گرفته است؛ به ویژه اینکه آزمایشات مربوطه معمولاً شامل ارزیابی در دیتاست‌های رایج بوده که تا حدی شبکه‌های کانولوشنی را به سمت بیش‌برازش سوق می‌دهد [۱۵]. بنابراین مقایسه سابقه بیست ساله شبکه‌های کانولوشنی با سابقه دو ساله ویژن ترنسفورمر این امید را می‌دهد که بخش زیادی از ضعف ویژن ترنسفورمر به دلیل همین سابقه کمتر تحقیقات آن است. این فرضیه زمانی که توفرون و همکاران [۱۵] با صرفاً تغییر روند آموزش شبکه ویژن ترنسفورمر توانستند عملکرد آن را به طرز شگفت‌انگیزی افزایش دهند، بیش از پیش پررنگ شد.

## ۳-۱-۳- بهبود فرآیند آموزش

مهم‌ترین تغییری که توفرون و همکاران [۱۵] در فرآیند آموزش ویژن ترنسفورمر دادند افزایش استفاده از روش‌های افزونگی داده<sup>۱</sup> است. آزمایشات ثابت کرد که تقریباً همه روش‌های افزونگی داده تاثیر بسیار خوبی در عملکرد ویژن ترنسفورمر دارد. علاوه بر این، برخی ابرپارامترهای این شبکه پس از آزمایشات متعدد تعیین شده که بعضاً با مقادیر ویژن ترنسفورمر استاندارد بسیار متفاوت هستند (جدول ۲). آن‌ها نام شبکه پیشنهادی خود را ایمیج ترنسفورمر کارا<sup>۳</sup> (DeiT) نامیدند. مجدداً تاکید می‌گردد که نسخه پایه ایمیج ترنسفورمر کارا از نظر ساختاری هیچ تفاوتی با ویژن ترنسفورمر ندارند و این تفاوت نام فقط برای متمایز کردن آن‌ها است.

نتیجه آزمایش روی دیتاست ImageNet در جدول ۳ نشان می‌دهد شبکه ایمیج ترنسفورمر کارا با همین تغییرات توانست بدون پیش آموزش در دیتاست‌های بزرگتر، بسیار بهتر از شبکه ویژن ترنسفورمر عمل کند و شکاف بین این شبکه و شبکه‌های کانولوشنی را تا حد زیادی پر کند. لازم به ذکر است که

جدول ۲- جزئیات و ابرپارامترهای آموزش شبکه ایمیج ترنسفورمر کارا در مقایسه با ویژن ترنسفورمر [۱۵]

| Methods                    | ViT-B [15]   | DeiT-B                                       |
|----------------------------|--------------|--|
| Epochs                     | 300          | 300  |
| Batch size                 | 4096         | 1024   |
| Optimizer                  | AdamW        | AdamW  |
| learning rate              | 0.003        | $0.0005 \times \frac{\text{batchsize}}{512}$ |
| Learning rate decay        | cosine       | cosine                                       |
| Weight decay               | 0.3          | 0.05   |
| Warmup epochs              | 3.4          | 5  |
| Label smoothing $\epsilon$ | $\times$     | 0.1  |
| Dropout                    | 0.1          | $\times$                                     |
| Stoch. Depth               | $\times$     | 0.1  |
| Repeated Aug               | $\times$     | $\checkmark$                                 |
| Gradient Clip.             | $\checkmark$ | $\times$                                     |
| Rand Augment               | $\times$     | 9/0.5  |
| Mixup prob.                | $\times$     | 0.8  |
| Cutmix prob.               | $\times$     | 1.0  |
| Erasing prob.              | $\times$     | 0.25   |

<sup>1</sup> Data-Augmentation

<sup>2</sup> Hyper-Parameter

<sup>3</sup> Data-efficient image transformers



جدول ۳-مقایسه شبکه ایمج ترنسفورمر کارا، ویژن ترنسفورمر و شبکه‌های کانولوشنی [۱۵]

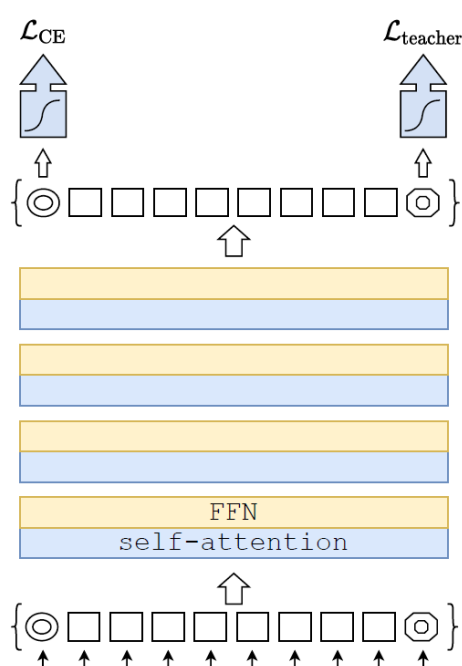
| Network                   | #param. | image throughput<br>size (image/s) |        | ImNet<br>top-1 | Real<br>top-1 | V2<br>top-1 |
|---------------------------|---------|------------------------------------|--------|----------------|---------------|-------------|
| Convnets                  |         |                                    |        |                |               |             |
| ResNet-18 [21]            | 12M     | 224 <sup>2</sup>                   | 4458.4 | 69.8           | 77.3          | 57.1        |
| ResNet-50 [21]            | 25M     | 224 <sup>2</sup>                   | 1226.1 | 76.2           | 82.5          | 63.3        |
| ResNet-101 [21]           | 45M     | 224 <sup>2</sup>                   | 753.6  | 77.4           | 83.7          | 65.7        |
| ResNet-152 [21]           | 60M     | 224 <sup>2</sup>                   | 526.4  | 78.3           | 84.1          | 67.0        |
| RegNetY-4GF [40]★         | 21M     | 224 <sup>2</sup>                   | 1156.7 | 80.0           | 86.4          | 69.4        |
| RegNetY-8GF [40]★         | 39M     | 224 <sup>2</sup>                   | 591.6  | 81.7           | 87.4          | 70.8        |
| RegNetY-16GF [40]★        | 84M     | 224 <sup>2</sup>                   | 334.7  | 82.9           | 88.1          | 72.4        |
| EfficientNet-B0 [48]      | 5M      | 224 <sup>2</sup>                   | 2694.3 | 77.1           | 83.5          | 64.3        |
| EfficientNet-B1 [48]      | 8M      | 240 <sup>2</sup>                   | 1662.5 | 79.1           | 84.9          | 66.9        |
| EfficientNet-B2 [48]      | 9M      | 260 <sup>2</sup>                   | 1255.7 | 80.1           | 85.9          | 68.8        |
| EfficientNet-B3 [48]      | 12M     | 300 <sup>2</sup>                   | 732.1  | 81.6           | 86.8          | 70.6        |
| EfficientNet-B4 [48]      | 19M     | 380 <sup>2</sup>                   | 349.4  | 82.9           | 88.0          | 72.3        |
| EfficientNet-B5 [48]      | 30M     | 456 <sup>2</sup>                   | 169.1  | 83.6           | 88.3          | 73.6        |
| EfficientNet-B6 [48]      | 43M     | 528 <sup>2</sup>                   | 96.9   | 84.0           | 88.8          | 73.9        |
| EfficientNet-B7 [48]      | 66M     | 600 <sup>2</sup>                   | 55.1   | 84.3           | -             | -           |
| EfficientNet-B5 RA [12]   | 30M     | 456 <sup>2</sup>                   | 96.9   | 83.7           | -             | -           |
| EfficientNet-B7 RA [12]   | 66M     | 600 <sup>2</sup>                   | 55.1   | 84.7           | -             | -           |
| KDforAA-B8                | 87M     | 800 <sup>2</sup>                   | 25.2   | 85.8           | -             | -           |
| Transformers              |         |                                    |        |                |               |             |
| ViT-B/16 [15]             | 86M     | 384 <sup>2</sup>                   | 85.9   | 77.9           | 83.6          | -           |
| ViT-L/16 [15]             | 307M    | 384 <sup>2</sup>                   | 27.3   | 76.5           | 82.2          | -           |
| DeiT-Ti                   | 5M      | 224 <sup>2</sup>                   | 2536.5 | 72.2           | 80.1          | 60.4        |
| DeiT-S                    | 22M     | 224 <sup>2</sup>                   | 940.4  | 79.8           | 85.7          | 68.5        |
| DeiT-B                    | 86M     | 224 <sup>2</sup>                   | 292.3  | 81.8           | 86.7          | 71.5        |
| DeiT-B↑384                | 86M     | 384 <sup>2</sup>                   | 85.9   | 83.1           | 87.7          | 72.4        |
| DeiT-Ti🍄                  | 6M      | 224 <sup>2</sup>                   | 2529.5 | 74.5           | 82.1          | 62.9        |
| DeiT-S🍄                   | 22M     | 224 <sup>2</sup>                   | 936.2  | 81.2           | 86.8          | 70.0        |
| DeiT-B🍄                   | 87M     | 224 <sup>2</sup>                   | 290.9  | 83.4           | 88.3          | 73.2        |
| DeiT-Ti🍄 / 1000 epochs    | 6M      | 224 <sup>2</sup>                   | 2529.5 | 76.6           | 83.9          | 65.4        |
| DeiT-S🍄 / 1000 epochs     | 22M     | 224 <sup>2</sup>                   | 936.2  | 82.6           | 87.8          | 71.7        |
| DeiT-B🍄 / 1000 epochs     | 87M     | 224 <sup>2</sup>                   | 290.9  | 84.2           | 88.7          | 73.9        |
| DeiT-B🍄↑384               | 87M     | 384 <sup>2</sup>                   | 85.8   | 84.5           | 89.0          | 74.8        |
| DeiT-B🍄↑384 / 1000 epochs | 87M     | 384 <sup>2</sup>                   | 85.8   | 85.2           | 89.3          | 75.2        |

منظور از DeiT $\uparrow$ 384، تنظیم دقیق شبکه DeiT روی رزولوشن بالاتر (۳۸۴) است. پیش از این ثابت شده که این کار، یعنی آموزش شبکه در رزولوشن پایین و سپس تنظیم دقیق آن در رزولوشن بالا می‌تواند سرعت آموزش شبکه و دقت آن را بهبود ببخشد [۲۲]. به طور کلی مشاهده می‌شود تغییرات در فرآیند آموزش به تنهایی می‌تواند ویژن ترنسفورمر را بهبود دهد (اگرچه همچنان نسبت به شبکه‌های کانولوشنی

در دیتاست‌های متوسط، ضعیف‌تر هستند). این نتیجه، پتانسیل بالای ویژن ترنسفورمر برای بهبودهای آینده را نشان می‌دهد.

### ۳-۳-۲- استفاده از رویکرد فشرده‌سازی دانش

برای افزایش بیشتر کارایی ویژن ترنسفورمر، توفرون و همکاران [۱۵] از رویکردی به نام فشرده‌سازی دانش [۲۳] استفاده کردند. فشرده‌سازی دانش در شبکه‌های عصبی به معنی انتقال یک مدل شاگرد<sup>۱</sup> از دانش یک مدل مربی<sup>۲</sup> است. این انتقال می‌تواند به صورت ملایم<sup>۳</sup>، از طریق ارائه خروجی سافت‌مکس مدل مربی به مدل شاگرد، یا به صورت سخت<sup>۴</sup>، تنها از طریق خروجی نهایی مدل مربی صورت بگیرد. این کار به ویژه در هنگام استفاده از روش افزونگی داده می‌تواند بسیار مفید باشد. حالتی را فرض کنید که یک عکس دارای برچسب گربه شامل یک گربه کوچک در گوشه تصویر باشد. اعمال افزونگی داده در



شکل ۴- ساختار شبکه ایمپج ترنسفورمر کارا به همراه رویکرد فشرده‌سازی دانش [۱۵]

<sup>1</sup> Student


<sup>2</sup> Teacher

<sup>3</sup> Soft

<sup>4</sup> Hard

این حالت ممکن گربه را از تصویر پاک کند که در این حالت برچسب گربه دیگر معتبر نخواهد بود. این تغییر برچسب به صورت خودکار از طریق مدل مربی به شاگرد منتقل می‌شود.

راه پیشنهادی توفرون و همکاران [۱۵] برای استفاده از رویکرد فشرده سازی دانش، استفاده از یک توکن فشرده سازی<sup>۱</sup> مشابه توکن کلاس امبدینگ است (شکل ۴). آزمایشات آن‌ها نشان داد استفاده از شبکه‌های کانولوشنی به عنوان شبکه مربی، نتیجه بهتری دارد. جالب اینکه در این حالت شبکه ویژن ترنسفورمر شاگرد از شبکه مربی خود پیشی می‌گیرد. از نظر آن‌ها این اتفاق به نحوی معادل انتقال بایاس القایی شبکه کانولوشنی به شبکه ترنسفورمر است. نکته جالب دیگر اینکه آزمایشات نشان داد که استفاده از روش سخت برای فشرده سازی دانش در شبکه‌های ویژن ترنسفورمر نتیجه بهتری دارد. این کار دقیقاً مشابه داشتن دو خروجی و دو برچسب مبنای<sup>۲</sup> است که تابع هزینه نهایی از ترکیب تابع هزینه آنتروپی متقاطع<sup>۳</sup> آن‌ها بدست می‌آید.

در جدول ۳ شبکه‌های ایمج ترنسفورمر کارا که در آن‌ها از رویکرد فشرده سازی دانش استفاده شده با علامت  مشخص شده‌اند. همانطور که مشاهده می‌کنید این شبکه‌ها بدون پیش آموزش در دیتاست‌های بزرگ‌تر، حتی از شبکه‌های کانولوشنی قدرتمند مانند ایفیشینتنت<sup>۴</sup> [۴] نیز بهتر عمل می‌کنند.

### ۳-۴- مقایسه هزینه محاسباتی با شبکه‌های کانولوشنی

شبکه ویژن ترنسفورمر با پیش آموزش بسیار زیاد [۱۳] یا با روش فشرده سازی دانش [۱۵] می‌توانند از نظر دقت بهتر از شبکه‌های کانولوشنی عمل کنند؛ اما در مسائل واقعی معیار بسیار مهم دیگر هزینه محاسباتی مدل‌ها است. از این نظر اطلاعات موجود در جدول ۳، برای مقایسه شبکه‌های کانولوشنی و ویژن ترنسفورمرها مناسب به نظر می‌رسد. مطابق این اطلاعات به ازای دقت‌های نسبتاً برابر، تعداد

<sup>1</sup> Distillation token

<sup>2</sup> Ground truth

<sup>3</sup> Cross entropy

<sup>4</sup> EfficientNet

پارامترها و نرخ گذر<sup>۱</sup> این شبکه‌ها در یک محدوده قرار دارد. بنابراین استفاده از شبکه‌های ویژن ترنسفورمر، حداقل برای مسئله دسته‌بندی تصاویر، کاملاً به صرفه است. متأسفانه ضعف بسیار بزرگ شبکه‌های ویژن ترنسفورمر که در این جدول مشاهده می‌شود، رزلوشن بسیار کوچک ورودی شبکه‌های ویژن ترنسفورمر می‌باشد. در مقایسه با شبکه‌های کانولوشنی متناظر (از نظر تعداد پارامتر و دقت)، شبکه‌های ویژن ترنسفورمر تا حدود ۵۰ درصد رزلوشن ورودی کوچکتری دارند. این موضوع باعث می‌شود در بسیاری از کاربردهای پیشرفته‌تر مانند تشخیص اشیا یا قطعه‌بندی تصاویر که نیاز به رزلوشن‌های بالاتری دارند، استفاده از ویژن ترنسفورمرها ممکن نباشد. برای حل این مشکل به نظر می‌رسد تغییرات ساختاری مهمی باید در ویژن ترنسفورمرها اعمال شود که در فصل بعد به بعضی از آن‌ها می‌پردازیم.

---

<sup>1</sup> Throughput

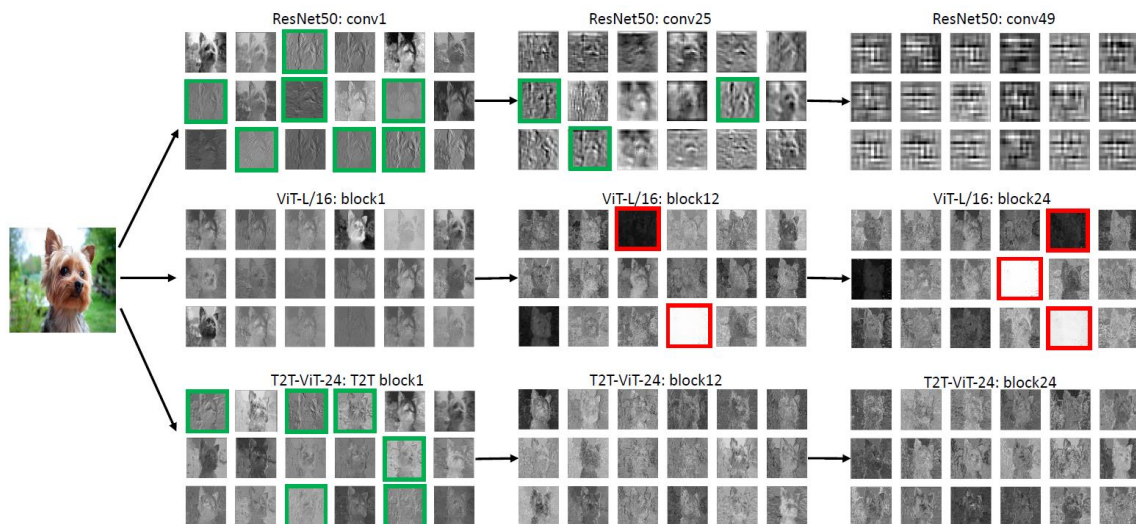
## فصل چهارم

### تغییرات ساختاری در ویژن ترنسفورمر

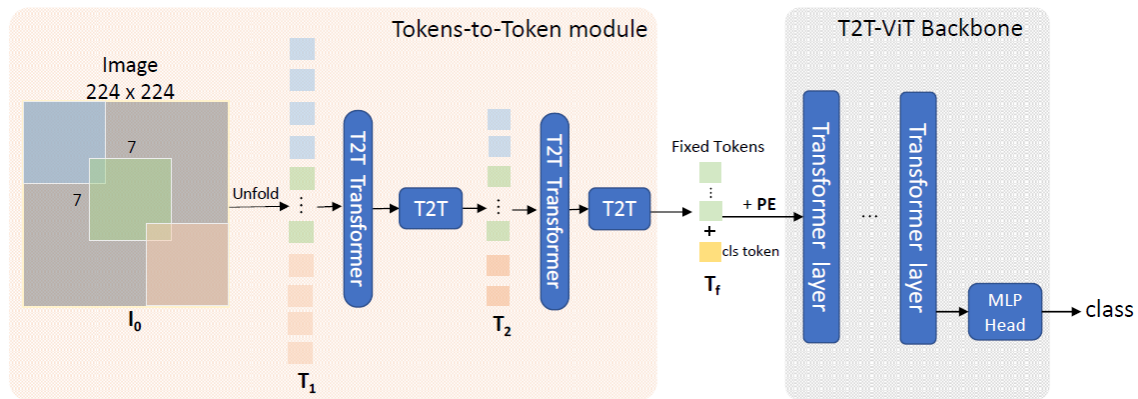
جریان جدیدی که ویژن ترنسفورمر در حوزه بینایی ماشین آغاز کرد موجب شد تا در سالهای اخیر محققان زیادی به بررسی آن علاقه‌مند شوند. علاوه بر تلاش‌هایی که برای بهبود آموزش ویژن ترنسفورمر صورت گرفت [۱۵]، بخش زیادی از تلاش‌ها صرف تغییر ساختار ویژن ترنسفورمر برای بهبود عملکرد و یا تطبیق آن با نیازهای حوزه بینایی ماشین بوده‌است. به بیان دقیق‌تر، هر کدام از تلاش به نحوی با افزودن بایاس القایی به ویژن ترنسفورمر آن را با ورودی تصویر سازگارتر می‌کنند. در این فصل ضمن بررسی سه کار برتر در این زمینه، نتایج آن را با شبکه‌های کانولوشنی و شبکه‌های ویژن ترنسفورمر استاندارد مقایسه می‌کنیم. نقطه مشترک همه کارهایی که در این فصل معرفی می‌شوند، ایجاد تغییرات بسیار زیاد در معماری ویژن ترنسفورمر است؛ بنابراین بر خلاف ایمپج ترنسفورمر کارا که در فصل قبل معرفی شد، برای این سه کار یک فصل مجزا در نظر گرفته شده است.

#### ۴-۱- بهبود فرآیند تبدیل تصویر به توالی توکن

عملکرد ضعیف ویژن ترنسفورمرها در دیتاست‌های متوسط، یان و همکاران [۱۶] را به سمت دو فرضیه برای توجیه این ضعف سوق داد. اولین فرض آن‌ها این بود که احتمالاً فرآیند تبدیل تصویر به توالی توکن، با روش ساده‌ای که در ویژن ترنسفورمر دیدیم، قابلیت درک ساختارهای دو بعدی مانند لبه‌ها و خط‌ها را از



شکل ۵-مقایسه صفحات ویژگی دو مدل ویژن ترنسفورمر و یک شبکه کانولوشنی در لایه‌های مختلف [۱۶]

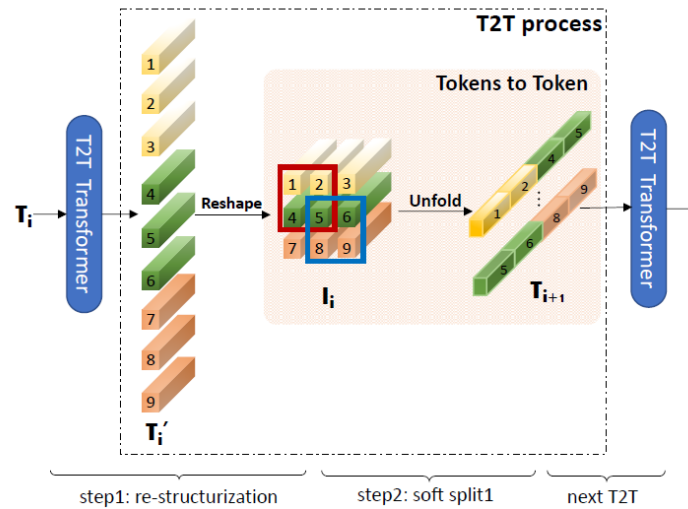


شکل ۶- ساختار کلی شبکه T2T-ViT [۱۶]

مدل سلب می‌کند. دومین فرض این بود که ساختار شبکه ویژن ترنسفورمر به اندازه شبکه‌های کانولوشنی برای کار با تصاویر مناسب نیست. از نظر آن‌ها مقایسه صفحات ویژگی شبکه ویژن ترنسفورمر و شبکه رزنت در لایه‌های مختلف تا حدی این فرضیه‌ها را اثبات می‌کند. با توجه به دو تصویر بالای شکل ۵، لایه‌های مختلف در شبکه‌های کانولوشنی به تدریج از لایه‌های اول تا آخر جزئیات ریز تا درشت تصویر را استخراج می‌کنند. در حالی که در ویژن ترنسفورمر تقریباً همه لایه‌ها ساختار کلی تصویر را مدل می‌کنند و جزئیاتی مانند لبه در این ویژگی‌ها دیده نمی‌شود. علاوه بر این، تعداد زیادی از لایه‌های ویژگی، مقادیر تقریباً ثابتی دارند (مربع‌های قرمز در شکل ۵).

برای حل این دو مشکل، شبکه پیشنهادی آن‌ها T2T-ViT<sup>۱</sup>، دارای ساختاری کاملاً جدید به صورت دو مرحله‌ای است. در حقیقت این شبکه به جز ساختار استاندارد ویژن ترنسفورمر، یک مرحله ابتدایی نیز دارد. تصویر ورودی در این ساختار جدید به وصله‌های هم‌پوشان با اندازه کوچکتر  $7 \times 7$  تقسیم و مشابه شبکه استاندارد به توکن‌های یک بعدی تبدیل می‌شوند. سپس در طول مرحله اول شبکه (بخش سمت چپ در شکل ۶)، توکن‌ها به تدریج (در طی دو مرحله)، پس از گذشت از هر لایه ترنسفورمر T2T به صورت دو بعدی بازسازی شده و مجدداً به صورت هم‌پوشان ولی با سایز وصله بزرگتر با هم ترکیب می‌شوند (شکل ۷). هر لایه ترنسفورمر T2T به سادگی ترکیبی از ماژول‌های خود-توجه و فیدفوروارد به همراه نرمال‌سازی لایه‌ای هستند. در انتهای این مرحله از شبکه، تعداد توکن‌ها به تعداد ثابتی کاهش می‌یابد. این فرآیند تدریجی و توکن‌های هم‌پوشان باعث می‌شود شبکه از ساختار دو بعدی تصاویر درک بهتری داشته باشد و

<sup>1</sup> Tokens-to-Token Vision Transformers



شکل ۷-فرآیند ترکیب توکن‌ها [۱۶]

توکن‌های مجاور اطلاعات محلی هم را بهتر درک کنند [۱۶]. با توجه به اینکه اندازه وصله‌ها در این مرحله کوچک‌تر و در نتیجه تعداد توکن‌ها بیشتر است. برای محدود نگه داشتن حجم محاسبات و حافظه مصرفی، بعد توکن‌ها مقدار بسیار کوچک‌تر انتخاب شده است (۳۲ یا ۶۴). همچنین برای کاهش حجم حافظه می‌توان به جای ساختار عادی ترنسفورمر از ساختارهای کارآمدتر (از نظر حافظه) مانند پرفورمر<sup>۱</sup> [۲۴] استفاده کرد. پیش از ورود به مرحله دوم (بخش سمت راست در شکل ۶) توکن‌های بدست آمده از مرحله اول با امبدینگ مکانی جمع شده و توکن کلاس به آن‌ها افزوده می‌شود.

جدول ۴-جزئیات ساختار شبکه T2T-ViT در مقایسه با ویژن ترنسفورمر استاندارد [۱۶]

| Models                        | Tokens-to-Token module |       |            |          | T2T-ViT backbone |            |          | Model size |          |
|-------------------------------|------------------------|-------|------------|----------|------------------|------------|----------|------------|----------|
|                               | T2T transformer        | Depth | Hidden dim | MLP size | Depth            | Hidden dim | MLP size | Params (M) | MACs (G) |
| ViT-S/16 [14]                 | -                      | -     | -          | -        | 8                | 786        | 2358     | 48.6       | 10.1     |
| ViT-B/16 [14]                 | -                      | -     | -          | -        | 12               | 786        | 3072     | 86.8       | 17.6     |
| ViT-L/16 [14]                 | -                      | -     | -          | -        | 24               | 1024       | 4096     | 304.3      | 63.6     |
| T2T-ViT-14                    | Performer              | 2     | 64         | 64       | 14               | 384        | 1152     | 21.5       | 5.2      |
| T2T-ViT-19                    | Performer              | 2     | 64         | 64       | 19               | 448        | 1344     | 39.2       | 8.9      |
| T2T-ViT-24                    | Performer              | 2     | 64         | 64       | 24               | 512        | 1536     | 64.1       | 14.1     |
| <b>T2T-ViT<sub>t</sub>-14</b> | Transformer            | 2     | 64         | 64       | 14               | 384        | 1152     | 21.5       | 6.1      |
| T2T-ViT-7                     | Performer              | 2     | 64         | 64       | 8                | 256        | 512      | 4.2        | 1.2      |
| T2T-ViT-12                    | Performer              | 2     | 64         | 64       | 12               | 256        | 512      | 6.8        | 2.2      |

<sup>1</sup> Performer



جدول ۵-مقایسه نتایج شبکه T2T-ViT با ساختارهای استاندارد ویژن ترنسفورمر در دیتاست ImageNet [۱۶]

| Models                                    | Top1-Acc (%) | Params (M)  | MACs (G)    |
|---|--------------|-------------|-------------|
| ViT-S/16 [14]                             | 78.1         | 48.6        | 10.1        |
| DeiT-small [38]                           | 79.9         | 22.1        | 4.6         |
| DeiT-small-Distilled [38]                 | 81.2         | 22.1        | 4.7         |
| <b>T2T-ViT-14</b>                         | <b>81.5</b>  | 21.5        | 5.2         |
| <b>T2T-ViT-14<math>\uparrow</math>384</b> | <b>83.3</b>  | 21.5        | 17.1        |
| ViT-B/16 [14]                             | 79.8         | 86.4        | 17.6        |
| ViT-L/16 [14]                             | 81.1         | 304.3       | 63.6        |
| <b>T2T-ViT-24</b>                         | <b>82.3</b>  | <b>64.1</b> | <b>14.1</b> |

مرحله دوم در این شبکه مشابه ساختار استاندارد ویژن ترنسفورمر، از چند بلاک کدگذار تشکیل شده است. تنها تفاوت تعداد لایه‌ها و بُعد توکن‌ها در طول شبکه است. یان و همکاران [۱۶] با الهام از ساختارهای رایج در شبکه‌های کانولوشنی، به این نتیجه رسیدند که کاهش بُعد توکن‌ها ضمن افزایش تعداد بلاک‌های کدگذار، عملکرد شبکه را بهبود می‌دهد.

#### ۴-۱-۱- نتایج

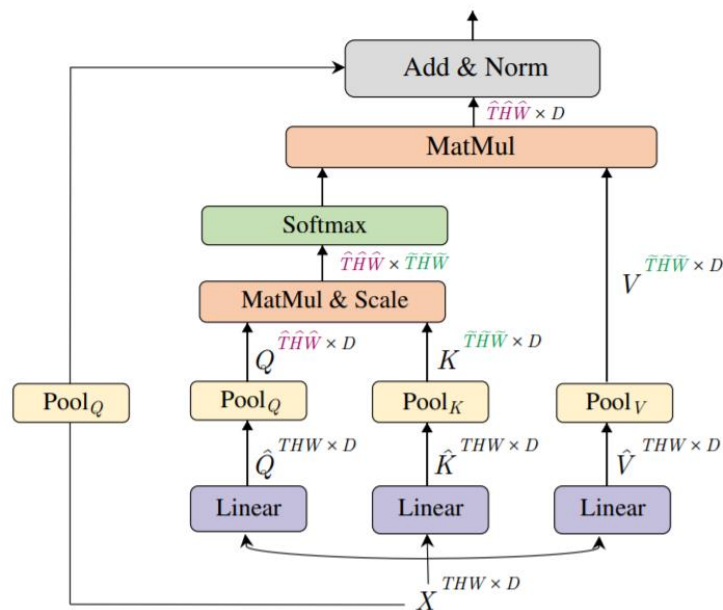
آموزش و ارزیابی شبکه در دیتاست ImageNet نشان می‌دهد که به ازای تعداد پارامترها و هزینه‌ی برابر، شبکه T2T-ViT از ساختارهای استاندارد بهتر عمل می‌کند (جدول ۵). جالب‌تر آنکه به ازای دقت‌های تقریباً برابر، شبکه T2T-ViT از نظر تعداد پارامتر و هزینه چندین برابر بهتر است. اگرچه همچنان بهترین نتایج این شبکه نسبت به شبکه افیشینت نت (جدول ۳) اندکی ضعیف‌تر است.

#### ۴-۲- ویژن ترنسفورمر چند مقیاسی

ویژن ترنسفورمر استاندارد تعداد و بُعد توکن‌های ورودی را در سرتاسر ساختار خود به سمت خروجی، حفظ می‌کند. به بیان دیگر رزلوشن مکانی و عمق بردار ویژگی در همه بلاک‌ها یکسان است. از طرفی معمولاً ویژگی‌های استخراج شده در بلاک‌های اول سطح پایین و در بلاک‌های بالا ویژگی‌ها معمولاً معنایی‌تر هستند؛ در نتیجه مشابه شبکه‌های کانولوشنی تعداد بیشتر ویژگی‌ها در لایه‌های انتهایی ممکن است

مفیدتر باشد. علاوه بر این تنها خروجی توکن کلاس برای دسته‌بندی استفاده می‌شود، و خروجی سایر توکن‌ها هیچ نقشی در دسته‌بندی ندارند. با این اوصاف آیا ثابت نگه داشتن تعداد توکن‌ها و عمق بردار آن‌ها منطقی است؟

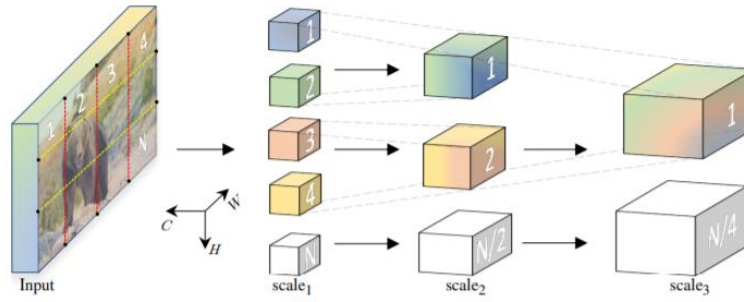
ایده اصلی فان و همکاران [۱۷] برای ارائه شبکه ویژن ترنسفورمر چندمقیاسی<sup>۱</sup> از همین سوال نشأت می‌گیرد. به نظر آن‌ها یکی از عوامل موفقیت شبکه‌های کانولوشنی ساختار چند مقیاسی آن‌هاست. در شبکه‌های کانولوشنی رایج، رزلوشن مکانی در طول شبکه کاهش و تعداد کانال‌های ویژگی به تدریج افزایش می‌یابد. این کار ضمن کاهش حجم محاسبات به شبکه امکان می‌دهد ویژگی‌های معنایی بیشتری استخراج کند. عنصر کلیدی در شبکه‌های کانولوشنی که امکان تغییر رزلوشن مکانی را فراهم می‌کند، لایه تجمیع<sup>۲</sup> است. بنابراین باید سعی شود تا سازوکار مشابهی برای ویژن ترنسفورمر ایجاد شود.



شکل ۸- مکانیزم خود-توجه تجمیعی [۱۷]

<sup>۱</sup> Multiscale Vision Transformers

<sup>۲</sup> Pooling



شکل ۹- ساختار شبکه ویژن ترنسفورمر چند مقیاسی [۱۷]

#### ۴-۲-۱- مکانیزم خود-توجه تجمیعی<sup>۱</sup>

در حالت کلی برای کاهش رزلوشن مکانی-زمانی (به ترتیب  $H$ ،  $W$  و  $T$  برای ارتفاع، عرض و تعداد فریم) توکن‌های ورودی، فان و همکاران [۱۷] از مکانیزمی شبیه لایه تجمیع در شبکه‌های کانولوشنی استفاده کردند. اگر ماتریس‌های جستار، کلید و مقدار در ماژول خود توجه را به ترتیب با  $\hat{Q}$ ،  $\hat{K}$  و  $\hat{V}$  نشان دهیم، ماتریس‌های متناظر کاهش یافته با اعمال تابع تجمیع  $P(\cdot; \theta)$  بدست می‌آیند. با فرض  $\theta = (k, s, p)$  این تابع یک عملیات تجمیع با هسته  $k$  و ابعاد  $k_T \times k_H \times k_W$ ، گام  $s$  متناظر در هر بعد  $S_T \times S_H \times S_W$  و گسترش<sup>۲</sup>  $p$  متناظر در هر بعد  $p_T \times p_H \times p_W$  روی تانسور ورودی اعمال می‌کند که تعداد توکن‌های ورودی  $N = T \times H \times W$  را به  $\tilde{N}$  کاهش می‌دهد.

$$\tilde{N} = \left\lfloor \frac{L + 2p - k}{s} \right\rfloor + 1 \quad (11)$$

با توجه به اینکه تابع تجمیع در سه بعد اعمال می‌شود، تانسور حاصل مجدداً به ماتریس  $\mathbb{R}^{\tilde{N} \times D}$  تبدیل می‌شود. پس از این ماتریس‌های کاهش یافته  $Q$ ،  $K$  و  $V$ ، ادامه فرآیند اعمال توجه را مشلبه قبل طی می‌کنند (شکل ۸). تعریف مکانیزم خود-توجه تجمیعی را می‌توان به حالت چندشاخه‌ای نیز بسط داد.

با توجه به معادلات خود-توجه، تنها تجمیع ماتریس  $Q$  می‌تواند باعث کاهش تعداد توکن‌ها شود. با این حال برای کاهش حجم محاسبات، عملیات تجمیع روی دو ماتریس  $K$  و  $V$  نیز اعمال می‌شود. با تعریف دقیق چگونگی کاهش تعداد توکن‌ها، حال می‌توانیم ساختار ویژن ترنسفورمر چند مقیاسی را تعریف کنیم.

<sup>۱</sup> Pooling self-attention

<sup>۲</sup> Padding

## ۴-۲-۲- ساختار شبکه

مطابق شکل ۹ شبکه ویژن ترنسفورمر چندمقیاسی از چند مرحله تشکیل شده است. ویدیو ورودی به مکعب‌هایی<sup>۱</sup> تقسیم شده و به عنوان توکن‌ها وارد مرحله اول می‌شوند. هر مرحله شامل تعداد مشخصی بلاک کدگذار مشابه شبکه ویژن ترنسفورمر استاندارد است. در این ساختار بین هر دو مرحله دو تغییر رزولوشن اتفاق می‌افتد: (۱) تغییر رزولوشن مکانی - زمانی (کاهش تعداد توکن‌ها) (۲) تغییر رزولوشن بردار ویژگی (افزایش بُعد بردار ویژگی هر توکن).

تغییر رزولوشن اول از طریق مکانیزم خود-توجه جمعی همانطور که در بخش قبل توضیح داده شد، اتفاق می‌افتد. در اولین بلاک هر مرحله، ماتریس  $Q$  جمع شده و بنابراین تعداد توکن‌ها در ابتدای مرحله کاهش می‌یابد. در سایر بلاک‌های مرحله، ماتریس‌های  $K$  و  $V$  نیز به تدریج جمع شده تا حجم محاسبات کاهش یابد.

دومین تغییر رزولوشن، افزایش بُعد بردار ویژگی، به سادگی با افزایش تعداد نرون‌های خروجی در ماژول فید فوروارد در آخرین بلاک قبل صورت می‌گیرد. این دو تغییر رزولوشن امکان استفاده از اتصال باقی مانده‌ای را از بین می‌برد. برای حل این مشکل در بلاک‌های متناظر با هر تغییر، به ترتیب از جمع ماتریس ورودی ماژول خود-توجه یا اعمال تبدیل خطی روی ورودی ماژول فیدفوروارد استفاده می‌شود. برای جمع‌بندی

| stage              | operators   | output sizes                       |
|--------------------|---|------------------------------------|
| data               | stride $8 \times 1 \times 1$                                    | $8 \times 224 \times 224$          |
| patch <sub>1</sub> | $1 \times 16 \times 16$ , 768<br>stride $1 \times 16 \times 16$ | $768 \times 8 \times 14 \times 14$ |
| scale <sub>2</sub> | MHA(768)<br>MLP(3072) $\times 12$                               | $768 \times 8 \times 14 \times 14$ |

| stage              | operators  | output sizes                       |
|--------------------|--|------------------------------------|
| data               | stride $4 \times 1 \times 1$                               | $16 \times 224 \times 224$         |
| cube <sub>1</sub>  | $3 \times 7 \times 7$ , 96<br>stride $2 \times 4 \times 4$ | $96 \times 8 \times 56 \times 56$  |
| scale <sub>2</sub> | MHPA(96)<br>MLP(384) $\times 1$                            | $96 \times 8 \times 56 \times 56$  |
| scale <sub>3</sub> | MHPA(192)<br>MLP(768) $\times 2$                           | $192 \times 8 \times 28 \times 28$ |
| scale <sub>4</sub> | MHPA(384)<br>MLP(1536) $\times 11$                         | $384 \times 8 \times 14 \times 14$ |
| scale <sub>5</sub> | MHPA(768)<br>MLP(3072) $\times 2$                          | $768 \times 8 \times 7 \times 7$   |

شکل ۱۰- مقایسه مشخصات شبکه ویژن ترنسفورمر چند مقیاسی (راست) و ویژن ترنسفورمر استاندارد (چپ) [۱۷]

<sup>۱</sup> این مکعب‌ها معادل وصله‌ها در ساختار ویژن ترنسفورمر استاندارد هستند. استفاده از واژه مکعب برای تأکید روی سه بعدی بودن ورودی است. (رزولوشن مکانی و زمان)

جدول ۶- نتایج شبکه ویژن ترنسفورمر چند مقیاسی در مقایسه با ساختار استاندارد و شبکه‌های کانولوشنی [۱۷]

| model                                      | Acc         | FLOPs (G) | Param (M) |
|--|-------------|-----------|-----------|
| RegNetZ-4GF [24]                           | 83.1        | 4.0       | 28.1      |
| RegNetZ-16GF [24]                          | 84.1        | 15.9      | 95.3      |
| EfficientNet-B7 [93]                       | 84.3        | 37.0      | 66.0      |
| DeiT-S [95]                                | 79.8        | 4.6       | 22.1      |
| DeiT-B [95]                                | 81.8        | 17.6      | 86.6      |
| DeiT-B $\uparrow 384^2$ [95]               | 83.1        | 55.5      | 87.0      |
| MViT-B-16, max-pool                        | 82.5        | 7.8       | 37.0      |
| MViT-B-24, max-pool                        | 83.1        | 10.9      | 53.5      |
| MViT-B-24-wide-320 <sup>2</sup> , max-pool | 84.3        | 32.7      | 72.9      |
| MViT-B-16                                  | 83.0        | 7.8       | 37.0      |
| MViT-B-24-wide-320 <sup>2</sup>            | <b>84.8</b> | 32.7      | 72.9      |

در شکل ۱۰ مشخصات یک معماری از ویژن ترنسفورمر چندمقیاسی (MViT-B) با ویژن ترنسفورمر استاندارد مقایسه شده است.

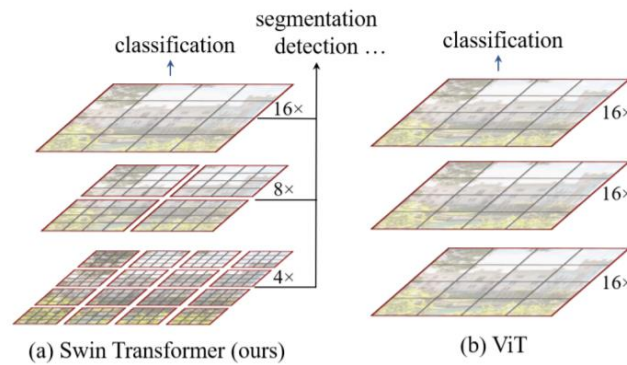
#### ۴-۲-۳- نتایج

فان و همکاران [۱۷]، شبکه پیشنهادی خود را در چند دیتاست شناسایی ویدیو ارزیابی کردند؛ اما در این گزارش، صرفاً نتایج مربوط به دسته‌بندی تصاویر در دیتاست ImageNet بررسی می‌شود. مطابق جدول ۶ ویژن ترنسفورمر چند مقیاسی هم از نظر تعداد پارامتر و حجم محاسبات و هم از نظر دقت از ساختار استاندارد ویژن ترنسفورمر عملکرد بهتری داشته است. همچنین برای اولین بار از بین شبکه‌هایی که تاکنون معرفی شد، توانسته از نظر دقت عملکردی مشابه و یا اندکی بهتر از شبکه ایفیشنت نت [۴] داشته باشد.

#### ۴-۳- محدود کردن مکانیزم خود-توجه: اسوین ترنسفورمر

مسئله دسته‌بندی تصاویر، صرف‌نظر از اهمیت ذاتی خود، دروازه ورود به مسائل پیشرفته‌تر در بینایی ماشین، مانند تشخیص اشیا و قطعه‌بندی تصاویر است. از آنجا که همواره از شبکه‌های دسته‌بندی به عنوان شبکه شالوده (استخراج‌کننده ویژگی) در مسائل پیشرفته‌تر استفاده می‌شود، در طی دو دهه اخیر هربار پیشرفت مسئله دسته‌بندی تصاویر، به سرعت موجب پیشرفت در سایر مسائل نیز شده است.

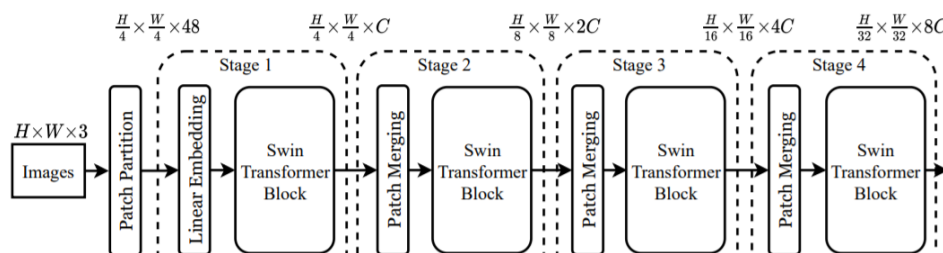
تمام کارهایی که تاکنون در زمینه ویژن ترنسفورمر معرفی شد، صرفاً تلاش‌هایی برای افزایش دقت و سرعت آن در مسئله دسته‌بندی تصاویر (یا ویدیو) بوده‌اند. مشکلی که در این میان کمتر مورد توجه قرار گرفته این است که ویژن ترنسفورمر در رزلوشن فعلی خود، به هیچ عنوان نمی‌تواند گزینه مناسبی برای



شکل ۱۱- ساختار سلسه مراتبی اسوین ترنسفورمر در مقایسه با ویژن ترنسفورمر استاندارد [۱۸]

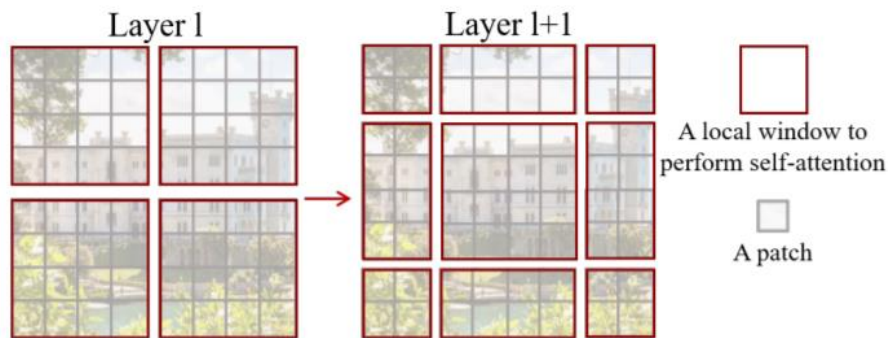
استخراج ویژگی در مسائل پیشرفته‌تر باشد. متأسفانه ساده‌ترین راه حل ممکن، یعنی کاهش سایز وصله‌ها و افزایش تعداد توکن‌ها هزینه محاسباتی را به شدت بالا می‌برد. همین مورد درباره افزایش رزولوشن تصاویر نیز صادق است؛ زیرا با ثابت نگه‌داشتن سایز وصله‌ها، هزینه محاسباتی ترنسفورمر با توان دوم رزولوشن ورودی متناسب است [۱۸].

راه حل پیشنهادی لیو و همکاران (شبکه اسوین ترنسفورمر<sup>۱</sup>) [۱۸] برای این مشکل، محدود کردن میدان اعمال مکانیزم خود-توجه و کاهش تدریجی رزولوشن تصویر است. مطابق تصویر سمت چپ شکل ۱۱، تصویر ورودی به پنجره‌هایی با اندازه ثابت تقسیم می‌شود (مربع‌های قرمز). هر پنجره شامل تعداد ثابتی توکن است که مکانیزم خود-توجه تنها در بین این توکن‌ها اعمال می‌شود. این کار اجازه می‌دهد توکن‌هایی با سایز بسیار کوچکتر از معمول ( $4 \times 4$  به جای  $16 \times 16$ ) داشته باشیم، بدون اینکه حجم محاسباتی بسیار سنگین شود. در لایه‌های بعدی توکن‌های مجاور با هم ترکیب شده تا مشابه شبکه‌های کانولوشنی، رزولوشن‌های پایین‌تر با ویژگی‌های معنایی‌تر بدست آید. این ساختار سلسله مراتبی به اسوین ترنسفورمر



شکل ۱۲- یک نمونه از معماری شبکه اسوین ترنسفورمر [۱۸]

<sup>۱</sup> Swin Transformer



شکل ۱۳- رویکرد پنجره انتقالی در اسوین ترنسفورمر [۱۸]

امکان می‌دهد به عنوان یک شبکه شالوده عمومی، جایگزین شبکه‌های هرمی رایج مانند شبکه هرم ویژگی<sup>۱</sup> [۲۵] شود. همانطور که در شکل ۱۲ مشاهده می‌کنید ساختار شبکه اسوین ترنسفورمر از چند مرحله ساخته شده است که هر مرحله شامل چند بلاک انکودر با ماژول خود-توجه تغییر یافته است. به جز بلاک اول که وصله‌های تصویر را به توکن‌هایی با بُعد ثابت تبدیل می‌کند، در ابتدای سایر بلاک‌های یک ماژول برای ترکیب توکن‌های همسایه و در نتیجه کاهش رزلوشن تصویر وجود دارد. همچنین بُعد توکن‌ها در هر بلاک نسبت به بلاک قبل دو برابر می‌شود (به جز بلاک ۲).

محدود کردن میدان اعمال مکانیزم خود-توجه، اگرچه هزینه محاسباتی را کاهش می‌دهد؛ اما مزیت اصلی ترنسفورمر، برقرای ارتباط بین توکن‌های دورتر را نیز از بین می‌برد. به طور دقیق‌تر پنجره‌های غیر همپوشان در شکل ۱۱ باعث می‌شود توکن‌هایی که در یک پنجره نیستند هیچ ارتباطی با هم نداشته باشند. این مورد به خصوص در توکن‌هایی که در پنجره‌های غیر همسایه هستند بحرانی‌تر است؛ زیرا ارتباط آن‌ها ممکن است فقط در آخرین مرحله اسوین ترنسفورمر صورت گیرد (بالاترین لایه شکل ۱۱). لیو و همکاران [۱۸] برای حل این مشکل از رویکرد ابتکاری پنجره انتقالی<sup>۲</sup> استفاده کردند (شکل ۱۳). در این رویکرد شیوه پنجره‌بندی توکن‌ها در دو بلاک متوالی از ترنسفورمر دائماً بین دو حالت شکل ۱۳ جابه‌جا می‌شود. ایده این کار این است که توکن‌های که در یک پنجره نیستند به صورت مستقیم یا با یک واسطه

<sup>۱</sup> Feature pyramid network

<sup>۲</sup> Shifted window



جدول ۷- جزئیات معماری‌های مختلف شبکه اسوین ترنسفورمر [۱۸]

|         | downsp. rate<br>(output size) | Swin-T                                | Swin-S                                 | Swin-B                                 | Swin-L                                 |
|---------|-------------------------------|---------------------------------------|--|--|--|
| stage 1 | 4×<br>(56×56)                 | concat 4×4, 96-d, LN                  | concat 4×4, 96-d, LN                   | concat 4×4, 128-d, LN                  | concat 4×4, 192-d, LN                  |
|         |                               | win. sz. 7×7,<br>dim 96, head 3 × 2   | win. sz. 7×7,<br>dim 96, head 3 × 2    | win. sz. 7×7,<br>dim 128, head 4 × 2   | win. sz. 7×7,<br>dim 192, head 6 × 2   |
| stage 2 | 8×<br>(28×28)                 | concat 2×2, 192-d, LN                 | concat 2×2, 192-d, LN                  | concat 2×2, 256-d, LN                  | concat 2×2, 384-d, LN                  |
|         |                               | win. sz. 7×7,<br>dim 192, head 6 × 2  | win. sz. 7×7,<br>dim 192, head 6 × 2   | win. sz. 7×7,<br>dim 256, head 8 × 2   | win. sz. 7×7,<br>dim 384, head 12 × 2  |
| stage 3 | 16×<br>(14×14)                | concat 2×2, 384-d, LN                 | concat 2×2, 384-d, LN                  | concat 2×2, 512-d, LN                  | concat 2×2, 768-d, LN                  |
|         |                               | win. sz. 7×7,<br>dim 384, head 12 × 6 | win. sz. 7×7,<br>dim 384, head 12 × 18 | win. sz. 7×7,<br>dim 512, head 16 × 18 | win. sz. 7×7,<br>dim 768, head 24 × 18 |
| stage 4 | 32×<br>(7×7)                  | concat 2×2, 768-d, LN                 | concat 2×2, 768-d, LN                  | concat 2×2, 1024-d, LN                 | concat 2×2, 1536-d, LN                 |
|         |                               | win. sz. 7×7,<br>dim 768, head 24 × 2 | win. sz. 7×7,<br>dim 768, head 24 × 2  | win. sz. 7×7,<br>dim 1024, head 32 × 2 | win. sz. 7×7,<br>dim 1536, head 48 × 2 |

Table 7. Detailed architecture specifications.

در بلاک‌های بعدی به هم متصل شوند و از این طریق همه توکن‌ها بتوانند با هم ارتباط برقرار کنند. نتایج آزمایشات نشان می‌دهد این کار می‌تواند تا حد زیادی کاهش دقت ناشی از مکانیزم خود-توجه محدود شده را جبران کند [۱۸]. در نهایت جزئیات معماری‌های مختلف شبکه اسوین ترنسفورمر در جدول ۷ ارائه شده‌است.

#### ۴-۳-۱- نتایج

نتایج ارزیابی شبکه اسوین ترنسفورمر در دیتاست ImageNet (جدول ۸) نشان می‌دهد این شبکه نه تنها از شبکه‌های ویژن ترنسفورمر استاندارد بهتر عمل می‌کند، بلکه از نظر سرعت و دقت نیز در حد شبکه‌های کانولوشنی یا از آن‌ها بهتر است. ویژگی اصلی این شبکه اما، ساختار سلسله مراتبی آن به عنوان شالوده استخراج ویژگی برای مسائل پیشرفته‌تر حوزه بینایی ماشین است. به همین منظور یک نمونه از نتایج این

جدول ۸- نتایج شبکه اسوین ترنسفورمر در مقایسه با ساختار استاندارد و شبکه‌های کانولوشنی [۱۸]

|                  | size             | params | flops (image / s) | top-1 acc. |      |
|------------------|------------------|--------|-------------------|------------|------|
| RegNetY-4G [48]  | 224 <sup>2</sup> | 21M    | 4.0G              | 1156.7     | 80.0 |
| RegNetY-8G [48]  | 224 <sup>2</sup> | 39M    | 8.0G              | 591.6      | 81.7 |
| RegNetY-16G [48] | 224 <sup>2</sup> | 84M    | 16.0G             | 334.7      | 82.9 |
| EffNet-B3 [58]   | 300 <sup>2</sup> | 12M    | 1.8G              | 732.1      | 81.6 |
| EffNet-B4 [58]   | 380 <sup>2</sup> | 19M    | 4.2G              | 349.4      | 82.9 |
| EffNet-B5 [58]   | 456 <sup>2</sup> | 30M    | 9.9G              | 169.1      | 83.6 |
| EffNet-B6 [58]   | 528 <sup>2</sup> | 43M    | 19.0G             | 96.9       | 84.0 |
| EffNet-B7 [58]   | 600 <sup>2</sup> | 66M    | 37.0G             | 55.1       | 84.3 |
| ViT-B/16 [20]    | 384 <sup>2</sup> | 86M    | 55.4G             | 85.9       | 77.9 |
| ViT-L/16 [20]    | 384 <sup>2</sup> | 307M   | 190.7G            | 27.3       | 76.5 |
| DeiT-S [63]      | 224 <sup>2</sup> | 22M    | 4.6G              | 940.4      | 79.8 |
| DeiT-B [63]      | 224 <sup>2</sup> | 86M    | 17.5G             | 292.3      | 81.8 |
| DeiT-B [63]      | 384 <sup>2</sup> | 86M    | 55.4G             | 85.9       | 83.1 |
| Swin-T           | 224 <sup>2</sup> | 29M    | 4.5G              | 755.2      | 81.3 |
| Swin-S           | 224 <sup>2</sup> | 50M    | 8.7G              | 436.9      | 83.0 |
| Swin-B           | 224 <sup>2</sup> | 88M    | 15.4G             | 278.1      | 83.5 |
| Swin-B           | 384 <sup>2</sup> | 88M    | 47.0G             | 84.7       | 84.5 |



شبکه در مسئله شناسایی اشیا و مسئله قطعه‌بندی اشیا بررسی می‌شود. برای مقایسه از شبکه ماسک آر-سی‌ان‌ان آبشاری<sup>۱</sup> [۲۶] با شبکه‌های شالوده مختلف استفاده شده‌است. مطابق جدول ۹، استفاده از استفاده اسوین ترنسفورمر به جای شبکه‌های کانولوشنی، بدون تغییر محسوس در هزینه و تعداد پارامتر، عملکرد شبکه ماسک-آرسی‌ان‌ان آبشاری را تا حد زیادی بهبود داده‌است. این نتایج نشان می‌دهد ویژن ترنسفورمر برای مسائل پیشرفته بینایی ماشین نیز می‌تواند با شبکه‌های کانولوشنی رقابت کنند.

جدول ۹- نتایج شبکه اسوین ترنسفورمر به عنوان استخراج‌کننده ویژگی [۱۷]

| (b) Various backbones w. Cascade Mask R-CNN |                   |                                 |                                 |                    |                                  |                                  |       |       |      |
|---|-------------------|---------------------------------|---------------------------------|--------------------|----------------------------------|----------------------------------|-------|-------|------|
|   | AP <sup>box</sup> | AP <sub>50</sub> <sup>box</sup> | AP <sub>75</sub> <sup>box</sup> | AP <sup>mask</sup> | AP <sub>50</sub> <sup>mask</sup> | AP <sub>75</sub> <sup>mask</sup> | param | FLOPs | FPS  |
| DeiT-S <sup>†</sup>                         | 48.0              | 67.2                            | 51.7                            | 41.4               | 64.2                             | 44.3                             | 80M   | 889G  | 10.4 |
| R50   | 46.3              | 64.3                            | 50.5                            | 40.1               | 61.7                             | 43.4                             | 82M   | 739G  | 18.0 |
| Swin-T                                      | <b>50.5</b>       | <b>69.3</b>                     | <b>54.9</b>                     | <b>43.7</b>        | <b>66.6</b>                      | <b>47.1</b>                      | 86M   | 745G  | 15.3 |
| X101-32                                     | 48.1              | 66.5                            | 52.4                            | 41.6               | 63.9                             | 45.2                             | 101M  | 819G  | 12.8 |
| Swin-S                                      | <b>51.8</b>       | <b>70.4</b>                     | <b>56.3</b>                     | <b>44.7</b>        | <b>67.9</b>                      | <b>48.5</b>                      | 107M  | 838G  | 12.0 |
| X101-64                                     | 48.3              | 66.4                            | 52.3                            | 41.7               | 64.0                             | 45.1                             | 140M  | 972G  | 10.4 |
| Swin-B                                      | <b>51.9</b>       | <b>70.9</b>                     | <b>56.5</b>                     | <b>45.0</b>        | <b>68.4</b>                      | <b>48.7</b>                      | 145M  | 982G  | 11.6 |

<sup>1</sup> Cascade Mask R-CNN

## فصل پنجم

### جمع‌بندی و نتیجه‌گیری

در این گزارش پس از معرفی ویژن ترنسفورمر، دیدیم که این شبکه با بایاس القایی بسیار کمتر در مقایسه با شبکه‌های کانولوشنی، می‌تواند در مسئله دسته‌بندی تصاویر در دیتاست‌های بسیار بزرگ نتایج بسیار خوبی کسب کند [۱۳]. این نتایج در کنار سابقه بسیار کوتاه ویژن ترنسفورمر محققین را به پیشرفت‌های بیشتر این شبکه امیدوار کرده است [۱۵]. در مسائل دنیای واقعی معمولاً تامین دیتاست‌های بزرگ و آموزش شبکه روی آن‌ها پرهزینه و مشکل است؛ به همین جهت کارهای بعدی سعی در تغییر ساختار ویژن ترنسفورمر به منظور افزودن بایاس القایی به آن، برای انطباق بیشتر با حوزه تصویر داشتند. به طور خاص در این گزارش دیدیم که بهبود فرآیند تبدیل تصویر به توالی توکن [۱۶] و تغییر تدریجی در رزلوشن مکانی و بُعد توکن‌ها [۱۷]، می‌تواند این شبکه را در دیتاست‌های کوچک در مقایسه با شبکه‌های کانولوشنی رقابتی کند. به هر حال کسب دقت یا سرعت روی مسئله دسته‌بندی تصاویر، به تنهایی برای رقابت با شبکه‌های کانولوشنی کافی نیست. متأسفانه ساختار استاندارد ویژن ترنسفورمر به دلیل محدودیت در رزلوشن تصاویر و تعداد توکن‌ها نمی‌تواند به عنوان شبکه شالوده در مسائل پیشرفته‌تر، مانند قطعه‌بندی تصاویر مورد استفاده قرار بگیرد. برای حل این مشکل دیدیم که اسوین ترنسفورمر [۱۸] با ایجاد محدودیت روی میدان اعمال مکانیزم خود-توجه، نه تنها در مسئله دسته‌بندی تصاویر بسیار بهتر عمل می‌کند بلکه در مسائل پیشرفته‌تر می‌تواند از شبکه‌های کانولوشنی پیشی بگیرد.

## ۵-۱- نتیجه‌گیری

رقابت بین شبکه‌های کانولوشنی و ویژن ترنسفورمرها با معرفی ایفیشینتنت ۲ [۲۷] و اسوین ترنسفورمر ۲ [۲۸] همچنان ادامه دارد. بنابراین این سوال که ویژن ترنسفورمر بهتر است یا کانولوشن همچنان بی پاسخ می‌ماند. اما آیا پاسخ به این سوال اهمیتی دارد؟ آیا در نهایت واقعا مجبور به انتخاب بین یکی از این دو هستیم؟ کارهای اخیر به این سوال جواب منفی داده‌اند. ترکیب شبکه‌های کانولوشنی با ترنسفورمرها [۲۹-۳۱] در یک سال اخیر به شدت مورد توجه قرار گرفته‌است. در حقیقت حتی موفق‌ترین کارهای قبلی ویژن ترنسفورمر، مانند اسوین ترنسفورمر [۱۸]، با وجود اینکه از لایه‌های کانولوشنی استفاده نمی‌کنند، موفقیت خود را مدیون الهامات زیاد از شبکه‌های کانولوشنی هستند. بنابراین به نظر می‌رسد ترکیب بایاس القایی شبکه‌های کانولوشنی با قدرت مکانیزم خود-توجه در ترنسفورمر، جریان جدیدی در بینایی ماشین خواهد بود.

## منابع و مراجع

- [1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," *Advances in neural information processing systems*, vol. 25, pp. 1097-1105, 2012.
- [2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.
- [3] J. Dai *et al.*, "Deformable Convolutional Networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764-773.
- [4] M. Tan and Q. Le, "Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks," in *International Conference on Machine Learning*, 2019: PMLR, pp. 6105-6114.
- [5] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
- [6] A. Vaswani *et al.*, "Attention Is All You Need," in *Advances in neural information processing systems*, 2017, pp. 5998-6008.
- [7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [8] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A Survey of Transformers," *arXiv preprint arXiv:2106.04554*, 2021.
- [9] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-Deeplab: Stand-Alone Axial-Attention for Panoptic Segmentation," in *European Conference on Computer Vision*, 2020: Springer, pp. 108-126.
- [10] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-Alone Self-Attention in Vision Models," *arXiv preprint arXiv:1906.05909*, 2019.
- [11] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," in *European Conference on Computer Vision*, 2020: Springer, pp. 213-229.
- [12] H. Wang, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "Max-Deeplab: End-to-End Panoptic Segmentation with Mask Transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5463-5474.

- [13] A. Dosovitskiy *et al.*, "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations*, 2020.
- [14] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in Vision: A Survey," *arXiv preprint arXiv:2101.01169*, 2021.
- [15] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training Data-Efficient Image Transformers & Distillation through Attention," in *International Conference on Machine Learning*, 2021: PMLR, pp. 10347-10357.
- [16] L. Yuan *et al.*, "Tokens-to-Token Vit: Training Vision Transformers from Scratch on Imagenet," in *International Conference on Computer Vision*, 2021.
- [17] H. Fan *et al.*, "Multiscale Vision Transformers," in *International Conference on Computer Vision*, 2021.
- [18] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in *International Conference on Computer Vision (ICCV)*, 2021.
- [19] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer Normalization," *arXiv preprint arXiv:1607.06450*, 2016.
- [20] P. W. Battaglia *et al.*, "Relational Inductive Biases, Deep Learning, and Graph Networks," *arXiv preprint arXiv:1806.01261*, 2018.
- [21] A. Kolesnikov *et al.*, "Big Transfer (Bit): General Visual Representation Learning," in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, 2020: Springer, pp. 491-507.
- [22] H. Touvron, A. Vedaldi, M. Douze, and H. Jégou, "Fixing the Train-Test Resolution Discrepancy," *arXiv preprint arXiv:1906.06423*, 2019.
- [23] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," *arXiv preprint arXiv:1503.02531*, 2015.
- [24] K. Choromanski *et al.*, "Rethinking Attention with Performers," *arXiv preprint arXiv:2009.14794*, 2020.
- [25] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117-2125.
- [26] Z. Cai and N. Vasconcelos, "Cascade R-Cnn: Delving into High Quality Object Detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154-6162.
- [27] M. Tan and Q. V. Le, "Efficientnetv2: Smaller Models and Faster Training," *arXiv preprint arXiv:2104.00298*, 2021.
- [28] Z. Liu *et al.*, "Swin Transformer V2: Scaling up Capacity and Resolution," *arXiv preprint arXiv:2111.09883*, 2021.
- [29] J. Guo *et al.*, "Cmt: Convolutional Neural Networks Meet Vision Transformers," *arXiv preprint arXiv:2107.06263*, 2021.

- [30] S. d'Ascoli, H. Touvron, M. Leavitt, A. Morcos, G. Biroli, and L. Sagun, "Convit: Improving Vision Transformers with Soft Convolutional Inductive Biases," *arXiv preprint arXiv:2103.10697*, 2021.
- [31] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "Coatnet: Marrying Convolution and Attention for All Data Sizes," *arXiv preprint arXiv:2106.04803*, 2021.