تمرین ۶ یادگیری عمیق

سید محمد عرفان موسوی منزه

۴۰۱۷۲۲۱۹۹

سوال ۱ - الف

شبکه های عصبی کانولوشنی معمولا از لایه های متوالی از عملگر های کانولوشن و ادغام تشکیل شده است. لایه کانولوشنی با اعمال یک فیلتر بر روی نواحی مختلف تصویر ورودی به دنبال یافتن الگو های مختلف در نواحی مختلف تصویر است. (هر کانولوشن به دنبال یافتن یک الگو در هر ناحیه از تصویر است.)

لایه ادغام با داون سمپل کردن نقشه فعال سازی حاصل از کانولوشن، با حفظ اطلاعات مهم ابعاد را کاهش می دهد.

شبکه های عصبی کانولوشنی برای بدست آوردن الگو های محلی و روابط محلی بین الگو ها مناسب هستند. اما توانایی درک یک الگوی بزرگ گلوبال در کل تصویر را ندارند. به زبان دیگر لایه های ابتدایی این شبکه ها به استخراج الگو های ابتدایی مانند خطوط، حاشیه ها و بافت ها می پردازد و با عمیق تر شدن شبکه به ویژگی های عمیق تر می پردازد. نکته مهم این است که شبکه رابطه بین این الگو ها در سر تا سر تصویر را یاد نمی گیرد و نهایتا بتواند به طور مختصر در یک بخش کوچک از تصویر این روابط را بیاموزد.

شبکه های عصبی مبتنی بر توجه با جمع کردن توجه شبکه بر روی یک ناحیه از تصویر کار می کنند.

با این کار شبکه میتواند یاد بگیرد که متناسب با وظیفه ای که دارد به نقاط کلیدی تصویر توجه کند و اطلاعات بسیار کلیدی برای انجام آن تسک را استخراج کند. مدل های مبتنی بر توجه بر خلاف مدل های کانولوشنی علاوه بر استخراج ویژکی های محلی برای استخراج ویژگی های جهانی نیز مناسب هستند. این مدل ها ذاتا توانایی پیدا کردن رابطه بین اشیا یا الگو های درون تصویر را دارند. (بر خلاف مدل های کانولوشنی)

سوال ۱ - ب

در شبکه های مبتنی بر کانولوشن هر فیلتر کانولوشنی به استخراج یک ویژگی از تصویر می پردازد. برای مثال برای وظیفه مشخص شده که تشخیص چهره انسان در تصویر است، یک فیلتر کانولوشنی ممکن است به تشخیص یکی از اعضای صورت بپردازد. ایراد اصلی شبکه کانولوشنی در آن است که این شبکه به ارتباط بین الگو های یافته شده نمی پردازد و صرفا وجود یا عدم وجود این الگو ها در تصویر را در نظر میگیرد. به همین دلیل به احتمال زیاد تصویر داده شده را به عنوان چهره انسان اعلام می کند.

اما شبکه های مبتنی بر توجه با در نظر گرفتن هم الگو های محلی و هم رابطه بین این الگو ها این توانایی را دارند که عدم انسان بودن این تصویر را تشخیص دهند. این شبکه ها در هنگام آموزش می آموزند که برای تشخیص انسان بودن به کدام قسمت از تصاویر توجه کنند همچنین این قسمت چگونه با سایر قسمت ها در ارتباط باشد هم یک بخش دیگر از دانشی است که آن ها یاد میگیرند. به همین دلیل هرچند که اجزای صورت در این تصویر نمایان است اما این شبکه ها برخلاف شبکه های کانولوشنی صرفا به دلیل وجود اجزای صورت آن را به عنوان صورت انسان اعلام نمی کنند.

نکته مهم: شبکه های کانولوشنی نسبت به چرخش الگو ها حساس هستند به همین دلیل الگو هایی مانند لب یا دماغ در این تصویر چون با الگو های مرسوم و افقی لب و دماغ تفاوت دارند نمی توانند توسط شبکه ای که صرفا تصویر صورت انسان در حالت عادی را دیده است استفاه شوند. پس در صورتی که ما سایر اجزای صورت را نیز بچرخانیم شانس شبکه کانولوشنی برای تشخیص این تصویر به عنوان انسان را کاهش می دهیم.

سوال ۳

$$4 \cdot 128 \cdot 128 \cdot 128 \Rightarrow 512\ patches\ with\ 16 \cdot 16 \cdot 16\ sizes$$

$$4 \cdot 16 \cdot 16 \cdot 16 = 16{,}384$$

$$We\ have\ a\ linear\ 16{,}384\ x\ 512$$

To convert this 512 to 768, we need this positional embeddings:

$$512\ x\ 768$$

سوال ۴ - الف

ترنسفورمر Swin برای حل چالش استفاده از ترنسفورمر های بینایی در وظایف تشخیص تصویر با وضوح بالا ارایه شده است. ترنسفورمر های بینایی سنتی در هنگام برخورد با تصاویر بزرگ به دلیل نیازهای محاسباتی و حافظه از مشکلات مقیاس پذیری رنج می برند. اما این مقاله یک معماری ترنسفورمر بینایی سلسله مراتبی را پیشنهاد می‌کند که از پنجره‌های جابجا شونده برای کاهش هزینه‌های محاسباتی و در عین حال حفظ عملکرد در تصاویر با وضوح بالا استفاده می‌کند.
Swin Transformer با تقسیم تصویر ورودی به تکه های غیر همپوشانی و پردازش آنها به صورت سلسله مراتبی، پردازش کارآمد و مقیاس پذیر تصاویر بزرگ را امکان پذیر می کند.

سوال ۴ - ب

ابتدا به معنی این دو اصطلاح بپردازیم:

MSA (multi-head self attention)

W-SMA (window based multi-head self attention)

در MSA ما یک توکن را با تمام توکن های دیگر ورودی مقایسه می کنیم. این رویکرد که به اصطلاح گلوبال است، پیچیدگی زمانی درجه دو دارد. از آنجایی که در وظایف مربوط به بینایی با تصاویر سر و کار داریم و تصاویر حتی آنهایی که کوچک هستند از تعداد زیادی پیکسل و ناحیه تشکیل شده اند باعث می شود که MSA بهینگی مناسبی برای آنها نداشته باشد.

اما W-SMA یک نوع مکانیزم توجه مبتنی بر پنجره است که به طور خاص برای پردازش بهینه تصاویر طراحی شده است. W-SMA تصویر ورودی را به پنجره‌هایی تقسیم می‌کند، محاسبات توجه را در هر پنجره انجام می‌دهد. این تکنیک W-SMA را برای هندل کردن تصاویر با وضوح بالا در معماری های ترنسفورور بینایی مناسب می کند.

سوال ۴ - ج

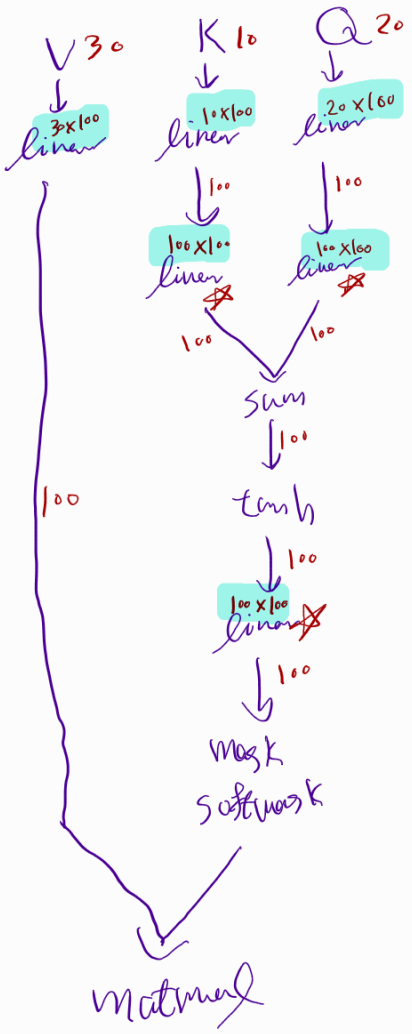
پنجره‌های جابجا شده نقش مهمی در استراتژی پردازش سلسله مراتبی ترنسفورمر Swin دارند. این استراتژی جابجایی تضمین می‌کند که هر پنجره اطلاعات را از پنجره‌های همسایه هم ترکیب می‌کند و یکپارچه‌سازی اطلاعات گلوبال را تسهیل می‌کند و حساسیت مکانی را حفظ می‌کند. به طور نمونه اگر یک اطلاعات مهم بین دو پنجره قرار گرفته باشد و ما توجه را به سمت آن ها نبریم می تواند در عملکرد مدل تاثیر منفی بگذارد. به عبارت دیگر توجه نکردن به بعضی از پیکسل ها در حالت بدن جا به جایی ریسک عدم توجه به اطلاعات و الگو های کلیدی را در بر دارد.

توجه افزودنی

$Q_{\in R^q}$ → linear $h \times q$ → h

$K_{\in R^k}$ → linear $h \times k$ → h

h + h → tanh → h → linear $V \times h$ → V



$$3 \times q \begin{cases} 30 \times 100 + 10 \times 100 + 20 \times 100 \\ + \\ 100 \times 100 + 100 \times 100 \\ + \\ 100 \times 100 \end{cases}$$

توجه افزودنی 3تا

$$+$$

$$300 \times 50$$

$$51,000$$

فرضیات: توجه افزودنی سایزش (همان متدار) ۱۰۰ حفظ می‌کند. ← تقاط ⭐ دار