بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

یادگیری عمیق

نیم‌سال دوم ۰۴-۰۳

مدرس: مهدیه سلیمانی

تمرین پنجم | ددلاین تمرین : ۲۹ خرداد

- برای ارسال هر تمرین تا ساعت ۲۳:۵۹ روز ددلاین فرصت دارید.
- در هر کدام از سوالات، اگر از منابع خارجی استفاده کرده‌اید باید آن را ذکر کنید. در صورت همفکری با افراد دیگر هم باید نام ایشان را در سوال مورد نظر ذکر نمایید.
- پاسخ تمرین باید ماحصل دانسته‌های خود شما باشد. در صورت رعایت این موضوع، استفاده از ابزارهای هوش مصنوعی با ذکر نحوه و مصداق استفاده بلامانع است.
- پاسخ ارسالی واضح و خوانا باشد. در غیر این صورت ممکن است منجر به از دست دادن نمره شود.
- پاسخ ارسالی باید توسط خود شما نوشته شده باشد. به اسکرین‌شات از منابع یا پاسخ افراد دیگر نمره‌ای تعلق نمی‌گیرد.
- در صورتی که بخشی از سوال‌ها را جای دیگری آپلود کرده و لینک آن را قرار داده باشید، حتما باید تاریخ آپلود مشخص و قابل اتکا باشد.
- محل بارگذاری سوالات نظری و عملی در هر تمرین مجزا خواهد بود. به منظور بارگذاری بایستی تمارین تئوری در یک فایل pdf با نام HW5_[First-Name]_[Last-Name]_[Student-Id].pdf و تمارین عملی نیز در یک فایل مجزای زیپ با نام HW5_[First-Name]_[Last-Name]_[Student-Id].zip بارگذاری شوند.
- در صورت وجود هرگونه ابهام یا مشکل، در کوئرای درس آن مشکل را بیان کنید و از پیغام دادن مستقیم به دستیاران آموزشی خودداری کنید.

بخش نظری (۱۰۰ نمره)

پرسش ۱. DetailCLIP (۵۰ نمره)

معماری DetailCLIP که در شکل ۱ نمایش داده شده، برای وظایفی که نیازمند قطعه‌بندی دقیق تصویر هستند، طراحی شده است. برخلاف مدل‌های سنتی که ممکن است جزئیات ریز را نادیده بگیرند، این مدل با استفاده از سه تکنیک، دقت را حفظ کرده و در عین حال از نظر محاسباتی کم‌هزینه باقی می‌ماند.

۱. (Patch-Level Self-Distillation)

در این روش، بخش‌های کوچکتر تصویر (دانش‌آموزان) از بخش‌های بزرگتر (معلمان) یاد می‌گیرند. این رویکرد به حفظ جزئیات کمک می‌کند که در غیر این صورت ممکن است از بین بروند. با تمرکز بر این تفاوت‌های ظریف، مدل از تقسیم‌بندی‌های اشتباهی که مدل‌های دیگر می‌کنند دور می‌کند.

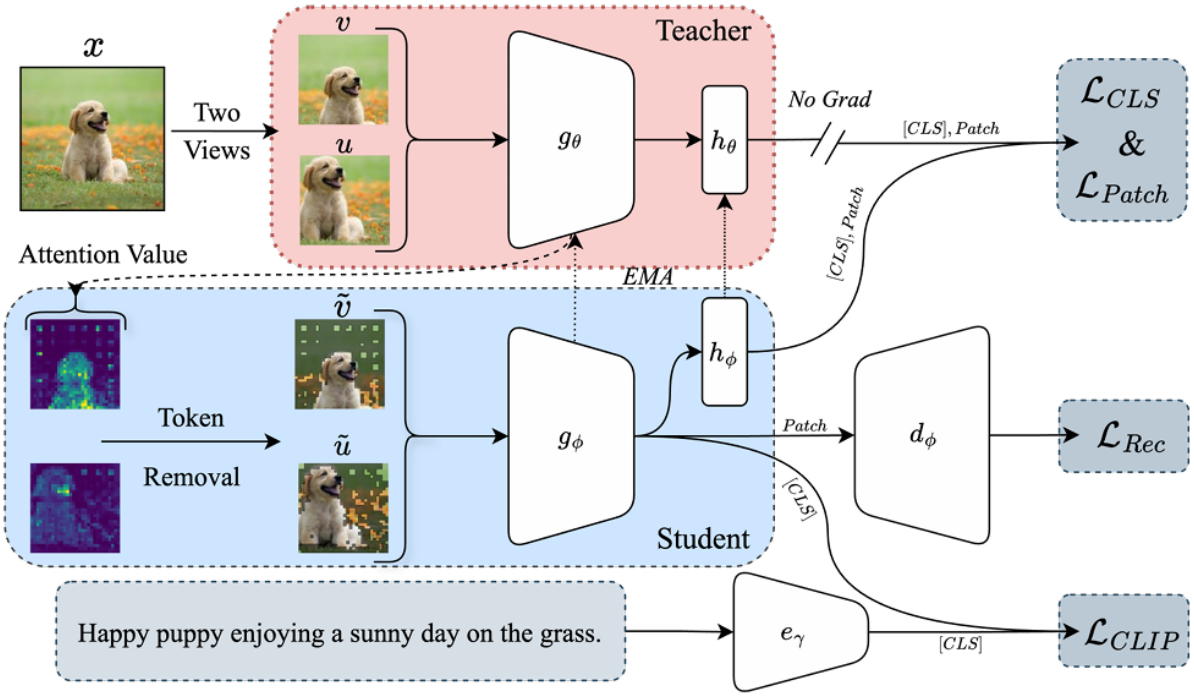
۲. حذف توکن (Attention-Based Token Removal)

این تکنیک مانند فیلتری برای داده‌ها عمل می‌کند. مدل تحلیل می‌کند که کدام بخش‌های تصویر اهمیت بیشتری دارد، برای مثال درک اهمیت شی داخل تصویر در برابر بکگراند تصور. این نه تنها سرعت تحلیل را افزایش می‌دهد بلکه با حذف نویز از مناطق غیرمهم، دقت را بهبود می‌بخشد.

۳. بازسازی در سطح پیکسل (Pixel-Level Reconstruction)

این روش برای افزایش وضوح تصویر به کار می‌رود. حتی هنگام کار با ورودی‌های با وضوح پایین، مدل می‌تواند خط‌های تیز و دقیقی را بازسازی کند. این ویژگی به‌ویژه برای اشکال پیچیده مانند خز حیوانات یا شاخ و برگ درختان ارزشمند است، جایی که لبه‌های دقیق برای تقسیم‌بندی دقیق ضروری هستند.

این روش‌ها با همدیگر همکاری می‌کنند تا مدل به دقت بالاتری برسد: Self-Distillation جزئیات ریز را حفظ می‌کند، حذف توکن پردازش را بهینه می‌کند، و بازسازی وضوح خروجی نهایی را دقیق‌تر می‌کند.



شکل ۱: با استفاده از معماری teacher-student، مدل teacher دو نمای مختلف از تصویر ورودی را پردازش کرده و مقادیر توجه (attention) را تولید می‌کند تا حذف توکن‌ها در مدل student را هدایت کند. سپس مدل student تصویر را با یک decoder vision بازسازی می‌کند، در حالی‌که هم‌زمان سه تابع هزینه شامل تابع طبقه‌بندی ($\mathcal{L}_{\text{CLS}}$)، تابع patch ($\mathcal{L}_{\text{Patch}}$)، و تابع بازسازی ($\mathcal{L}_{\text{Rec}}$) را بهینه می‌کند. همچنین تابع هزینه CLIP ($\mathcal{L}_{\text{CLIP}}$) به alignment میان انکودرهای تصویر و متن کمک می‌کند.

سؤالات:

۱. با توجه به پویایی teacher-student در یادگیری patch-level :

(آ) این رویکرد سلسله‌مراتبی چگونه به حفظ جزئیاتی از تصویر که ممکن است در غیر این صورت از بین بروند کمک می‌کند؟

در رویکرد سلسله‌مراتبی teacher-student در سطح patch که در DetailCLIP معرفی شده، مدل معلم تصویر کامل را پردازش می‌کند و بردارهای دقیق CLS و patch را تولید می‌نماید، در حالی که مدل دانش‌آموز نسخه‌ای ماسک‌شده از تصویر را دریافت کرده و تلاش می‌کند خروجی مشابهی با معلم تولید کند. با اعمال KL-divergence بین بردارهای patch و CLS معلم و دانش‌آموز و همچنین بازسازی نواحی ماسک‌شده با استفاده از یک decoder و اعمال loss بازسازی، مدل مجبور می‌شود به روابط موضعی و جزئیات ظریف تصویر توجه بیشتری نشان دهد. این فرآیند باعث می‌شود مدل دانش‌آموز قادر باشد حتی از اطلاعات ناقص نیز به شکل دقیق‌تری معنا استخراج کند و از بین رفتن جزئیات حیاتی تصویر جلوگیری شود.

(ب) مدل‌های دسته‌بندی سنتی چه محدودیت‌هایی دارند که این تکنیک به رفع آن‌ها کمک می‌کند؟

مدل‌های سنتی دسته‌بندی (مانند CLIP یا ViT در حالت پایه) معمولاً بر روی ویژگی‌های کلی و سطح بالا از تصویر تمرکز می‌کنند تا بتوانند برچسب نهایی (مثلاً "سگ" یا "ماشین") را پیش‌بینی کنند. این مدل‌ها به‌دلیل استفاده از توکن CLS و تمرکز بر کل تصویر، اغلب قادر نیستند به جزئیات موضعی (مانند شکل دقیق گوش، بافت پوست، یا مرزهای اشیاء) توجه کافی نشان دهند. این موضوع باعث می‌شود عملکرد آن‌ها در وظایفی مثل سگمنتیشن، تشخیص اشیاء کوچک، یا تطبیق جزئیاتی بین تصویر و متن کاهش یابد. تکنیک DetailCLIP با افزودن یادگیری در سطح patch، بازسازی نواحی ماسک‌شده، و استفاده از teacher-student، مدل را وادار می‌کند تا به اطلاعات محلی و جزئیات ریز تصویر توجه کند. در نتیجه، این روش به رفع محدودیت مدل‌های سنتی در درک دقیق ساختارهای تصویر و افزایش عملکرد در وظایف جزئی‌محور (fine-grained tasks) کمک می‌کند.

۲. درباره‌ی فیلترسازی (attention-based filtering):

(آ) مدل با چه معیارهایی تصمیم می‌گیرد کدام نواحی تصویر را در اولویت قرار دهد؟

در روش attention-based filtering مدل DetailCLIP تصمیم‌گیری برای حفظ یا حذف نواحی تصویر بر اساس مقدار attention بین توکن CLS و توکن‌های پچ انجام می‌شود؛ بدین صورت که در آخرین لایه ترنسفورمر، مدل یک نقشه توجه تولید می‌کند که نشان می‌دهد هر پچ چقدر در درک کلی تصویر از دید توکن [CLS] نقش دارد. سپس پچ‌هایی که کمترین attention را دریافت کرده‌اند، به‌عنوان نواحی کم‌اهمیت شناخته شده و ماسک می‌شوند. این فیلتر کردن هدفمند باعث می‌شود مدل به‌جای یادگیری از کل تصویر، روی نواحی مهم‌تر تمرکز کرده و نواحی حذف‌شده را از طریق زمینه بازسازی کند، که این موضوع موجب بهبود یادگیری جزئیات موضعی و دقت در وظایف ظریف می‌شود.

(ب) این عمل (انتخاب اینکه به چه مکانی توجه شد) چه تأثیری بر کیفیت تحلیل دارد؟

انتخاب هدفمند اینکه مدل به چه نواحی‌ای از تصویر توجه کند، تاثیر مستقیمی بر کیفیت تحلیل داده دارد، زیرا به مدل اجازه می‌دهد تمرکز خود را بر بخش‌های معنادار تصویر معطوف کرده و از پردازش نواحی کم‌اهمیت یا بی‌ربط صرف‌نظر کند. این تمرکز موجب استخراج ویژگی‌های دقیق‌تر و درک بهتر روابط موضعی در تصویر می‌شود که به بهبود عملکرد مدل در وظایفی مانند طبقه‌بندی جزئی، سگمنتیشن و تشخیص اشیاء کوچک کمک می‌کند. در نتیجه، مدل نه‌تنها بازدهی بالاتری دارد، بلکه قادر است تحلیل عمیق‌تری از داده‌های تصویری ارائه دهد.

۳. در مورد فرایند بازسازی (reconstruction):

(آ) چرا توانایی افزایش وضوح ورودی‌های کم‌کیفیت در کاربردهای دنیای واقعی ارزشمند است؟

در بسیاری از کاربردهای دنیای واقعی مانند پزشکی، نظارت تصویری، خودروهای خودران، یا سیستم‌های امنیتی، تصاویر دریافتی ممکن است به دلایل مختلف مانند نویز، فشرده‌سازی، یا شرایط نوری ضعیف، کیفیت پایین یا جزئیات ناقص داشته باشند. توانایی مدل در بازسازی و افزایش وضوح این ورودی‌ها به آن امکان می‌دهد تا حتی در شرایط نامطلوب نیز اطلاعات معنادار را استخراج کند و تصمیم‌گیری دقیقی داشته باشد. این قابلیت به‌ویژه در موقعیت‌هایی که داده‌های اصلی قابل تکرار یا جایگزینی نیستند بسیار ارزشمند است، چرا که می‌تواند عملکرد مدل را پایدارتر، قابل اعتمادتر و کاربردی‌تر در محیط‌های غیرایده‌آل کند.

(ب) کدام انواع اشیاء یا صحنه‌ها بیشتر از این قابلیت بهبود بهره‌مند می‌شوند؟

اشیایی مانند چهره انسان، بافت‌های پزشکی و علائم نوشتاری به دلیل نیاز به جزئیات بالا بیشترین بهره را از افزایش وضوح می‌برند. در صحنه‌های طبیعی یا صنعتی که عناصر ریز و دقیق اهمیت دارند، بازسازی کیفیت پایین حیاتی است. این قابلیت باعث بهبود درک مدل در شرایط واقعی با داده‌های ناقص یا نویزی می‌شود.

.۴

(آ) این سه مؤلفه چگونه یکدیگر را تکمیل کرده و در مجموع یک مدل قدرتمند می‌سازند؟

مولفه‌های مدل DetailCLIP از طریق تمرکز بر اطلاعات مهم، استخراج دقیق ویژگی‌ها و بازسازی جزئیات، به‌صورت مکمل با یکدیگر همکاری می‌کنند تا درک عمیق‌تری از تصویر ایجاد کرده و عملکرد مدل را در وظایف دقیق‌محور به‌طور چشمگیری بهبود دهند. در واقع مولفه اول این مدل یاد می‌گیرد که امبدینگ‌های درستی برای توکن CLS و هر پچ از تصویر بسازد و این امبدینگ‌ها در ساخت ماتریس توجه که در بخش حذف توکن استفاده می‌شود، به کار می‌روند پس ساخت امبدینگ‌های مناسب برای هر یک از این توکن‌ها روی ساخت نقشه ویژگی مناسب تاثیر دارد. از طرفی با کمک ماتریس توجه بدست آمده از ترنسفورمر انکدر مدل معلم، قسمت‌های غیر مهم حذف و از مدل خواسته می‌شود آن‌ها را باتوجه به قسمت‌های مهم که در تصویر باقی مانده و قابل مشاهده است، بازسازی کند. پس محاسبه یک ماتریس توجه درست که از مولفه قبلی بدست می‌آید در ارضای هدف بازسازی به کمک دیکدر نقش دارد.

(ب) در صورت حذف یکی از این تکنیک‌ها، چه ضعف‌هایی ممکن است در عملکرد مدل ایجاد شود؟

حذف مکانیسم حذف توکن مبتنی بر توجه باعث می‌شود مدل نتواند نواحی مهم تصویر را به‌درستی شناسایی و تمرکز کند؛ در نتیجه، بخش‌های غیرضروری پردازش می‌شوند و کیفیت ویژگی‌های استخراج‌شده کاهش می‌یابد، به‌ویژه در وظایف ریزدانه مانند بخش‌بندی. حذف self-distillation در سطح patch موجب می‌شود مدل از دانش سطح بالای معلم بی‌بهره بماند و نتواند ارتباط مؤثر میان اجزای تصویر را در سطوح مختلف یاد بگیرد، که منجر به ضعف در یادگیری امبدینگ‌های دقیق و هماهنگ می‌شود. حذف بازسازی در سطح پیکسل توانایی مدل در حفظ و بازسازی جزئیات تصویری را کاهش می‌دهد، که به‌طور مستقیم بر دقت در وظایف حساس به جزئیات (مانند تشخیص مرز اشیاء) تأثیر منفی می‌گذارد.

پرسش ۲. یادگیری خودنظارتی (۵۰ نمره)

(الف) به طور کلی در روش‌های Self Supervised Learning تلاش بر این است که شبکه برای هر تصویر یک representation خروجی بدهد، به‌گونه‌ای که مفاهیم آن تصویر را در خود در بر داشته باشد. بسیاری از روش‌ها همچون روش‌هایی که در شکل می‌بینیم، این کار را با تلاش برای نزدیک‌کردن representation دو تصویر مشابه (positive pairs) انجام می‌دهند. با این حال چرا چرا برخی روش‌ها مانند SimCLR به نمونه‌ها نامشابه (negative samples) نیاز دارند تا خروجی مطلوبی داشته باشند؟

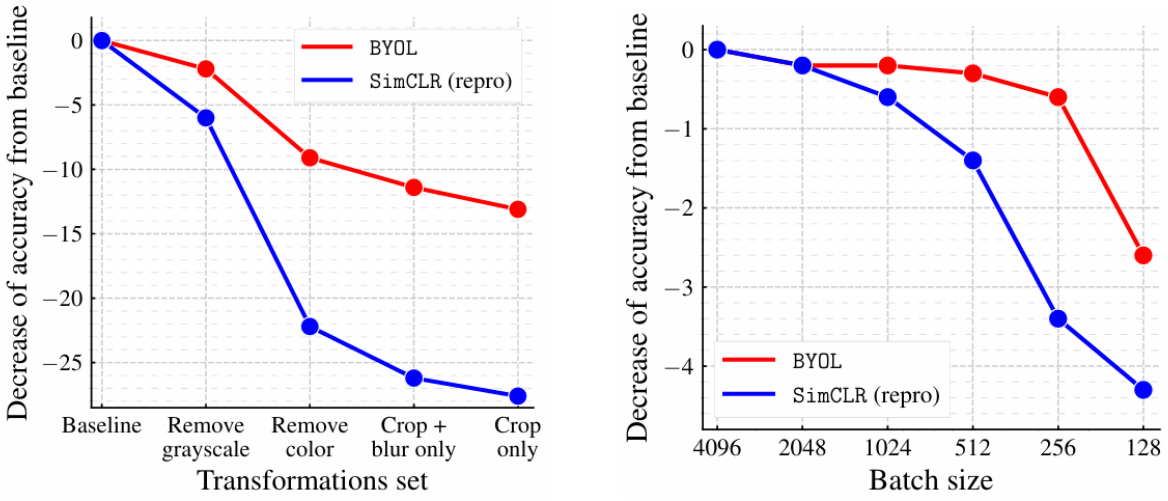
در روش‌هایی مانند SimCLR که مبتنی بر contrastive learning هستند، مدل با استفاده از جفت‌های مثبت (مانند دو view مختلف از یک تصویر) سعی می‌کند نمایش‌های مشابهی برای آن‌ها تولید کند، و در مقابل، با استفاده از نمونه‌های منفی (تصاویر متفاوت)، تلاش می‌کند نمایش‌های متمایزی ایجاد کند. دلیل نیاز به نمونه‌های منفی این است که مدل بتواند بین تصاویر مختلف تمایز قائل شود. اگر فقط نمونه‌های مثبت وجود داشته باشند، مدل ممکن است تمام تصاویر را به یک ناحیه مشترک در فضای نمایش ببرد (کاری که یک مدل با وزن صفر هم می‌تواند انجام دهد)، که در نتیجه اطلاعات مفید برای تفکیک از بین می‌رود. نمونه‌های منفی به مدل کمک می‌کنند تا هم شباهت‌ها و هم تفاوت‌ها را یاد بگیرد و در نتیجه خروجی دقیق‌تر و قابل تعمیم‌تری تولید کند. همچنین لاس این مدل لاس کانترستیو یا تقابلی است و برای این نوع لاس‌ها نیاز به هر دو نمونه مثبت و منفی داریم. مدل‌هایی که نیاز به نمونه منفی ندارند، مدل‌هایی هستند که لاسشان منفی شباهت کسینوسی است و همچنین دو شبکه انکدرشان متقارن نیستند.

(ب) تحقیق کنید و بگویید دلیل اینکه چنین مشکلی در روش‌هایی چون BYOL رخ نمی‌دهد چیست.

در روش BYOL برخلاف روش‌های مبتنی بر contrastive learning، مدل بدون استفاده از نمونه‌های منفی آموزش می‌بیند. این روش از دو شبکه استفاده می‌کند: یک شبکه اصلی (online) و یک شبکه هدف (target) که با

میانگین‌گیری نمایی از وزن‌های شبکه اصلی به‌روزرسانی می‌شود. هدف مدل، پیش‌بینی نمایش‌های شبکه هدف بر اساس ورودی‌های تغییریافته است. عامل کلیدی که از فروپاشی نمایشی جلوگیری می‌کند، وجود شبکه هدف و به‌روزرسانی کند آن است؛ زیرا این ساختار نوعی ثبات در هدف یادگیری ایجاد می‌کند که مانع می‌شود همه نمایش‌ها به یک مقدار ثابت همگرا شوند. در واقع، مدل نمی‌تواند فقط با تولید خروجی ثابت، خطا را کاهش دهد، چرا که خروجی انکدر هدف همواره متغیر است. به همین دلیل، روش BYOL بدون نیاز به نمونه‌های منفی نیز از یادگیری معنای تصویر پشتیبانی می‌کند و collapse رخ نمی‌دهد. دلیل دیگری که جلوی این مشکل را می‌گیرد این است که معماری انکدر اصلی و هدف متقارن نیست و در مسیر انکدر اصلی، یک هد پردیکتور اضافه‌تر وجود دارد تا خروجی امبدینگ انکدر اصلی را به خروجی انکدر هدف نزدیک کند. این عدم تقارن مانع از ایجاد مشکل گفته شده در معماری BYOL می‌شود.

(ج) در مقایسه‌ی BYOL می‌بینیم که این روش در برابر انتخاب برخی هایپرپارامترها همچون batch size و یا انتخاب transformationهایی که روی تصاویر اعمال می‌شوند مقاوم‌تر است (شکل ۲).



(ب) transformations removing of Impact

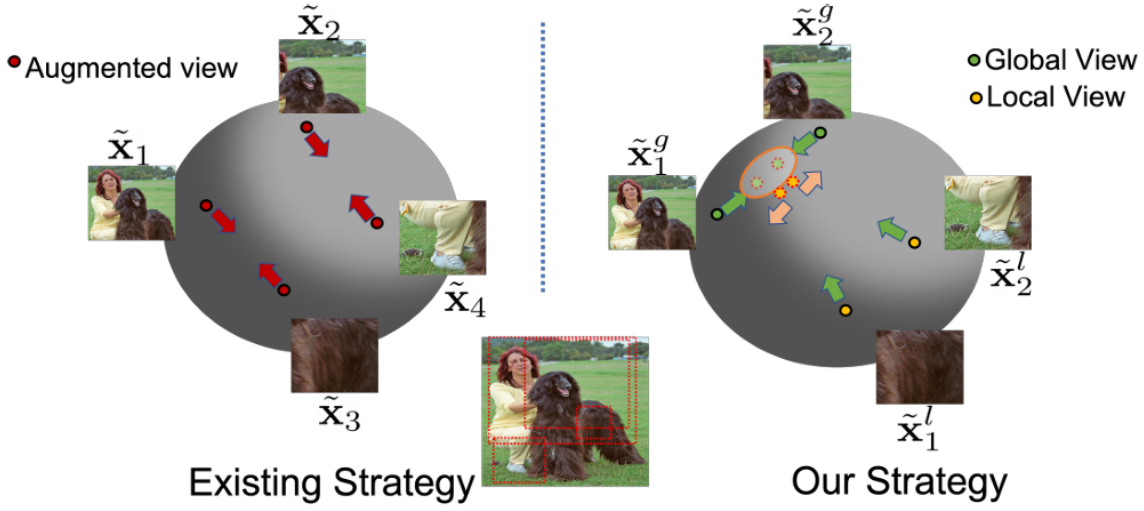
(آ) size batch of Impact

شکل ۲: BYOL: Bootstrap Your Own Latent

به نظر شما دلیل این موضوع چیست؟

روش BYOL نسبت به SimCLR در برابر تغییرات اعمال‌شده روی تصویر و اندازه کوچک batch مقاوم‌تر است، زیرا برخلاف SimCLR که برای یادگیری مؤثر به تعداد زیادی نمونه منفی نیاز دارد و عملکرد آن به batch بزرگ وابسته است، BYOL بدون استفاده از نمونه‌های منفی کار می‌کند و از یک شبکه هدف با به‌روزرسانی آهسته برای ایجاد ثبات در فرآیند یادگیری بهره می‌برد. این ویژگی باعث می‌شود که BYOL حتی در شرایطی که transformation های مختلف (مانند crop یا blur) روی تصاویر اعمال می‌شود، بتواند نمایش‌های پایدار و معنادار یاد بگیرد و دچار فروپاشی نمایش یا افت عملکرد نشود.

(د) در برخی روش‌ها سعی می‌شود بازنمایی برش‌های بزرگ یک تصویر (global crops) به یکدیگر نزدیک و برش‌های کوچک local crops در عین حال که به بازنمایی برش‌های بزرگ نزدیک باشند، از یکدیگر دور شوند (شکل ۳).



شکل ۳: Representations. Global and Local Your Leverage

دلیل این موضوع را چه می‌دانید؟

دلیل این کار آن است که global crops نمایانگر محتوای کلی تصویر هستند و باید نمایش‌های آن‌ها مشابه باشند تا مدل بتواند مفاهیم اصلی تصویر را یاد بگیرد. در مقابل، local crops معمولاً بخش‌های کوچکی از تصویر را شامل می‌شوند که ممکن است حاوی جزئیات متفاوت یا حتی غیرمرتبط باشند. اگر crops lrlocal بیش از حد به یکدیگر نزدیک شوند، مدل ممکن است تفاوت‌های ظریف میان نواحی مختلف تصویر را نادیده بگیرد. بنابراین، نزدیک کردن امبدینگ local crops به global crops کمک می‌کند تا ارتباط معنایی آن‌ها با تصویر کلی حفظ شود، اما دور نگه‌داشتن local crops از یکدیگر باعث می‌شود مدل بتواند تفاوت‌های محلی و جزئیات تصویر را بهتر یاد بگیرد، که این برای درک دقیق‌تر ساختار تصویر و یادگیری نمایش‌های غنی‌تر بسیار مهم است.