Patching

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.

masking ← قرار خود به جای transformer ها ابزار بود

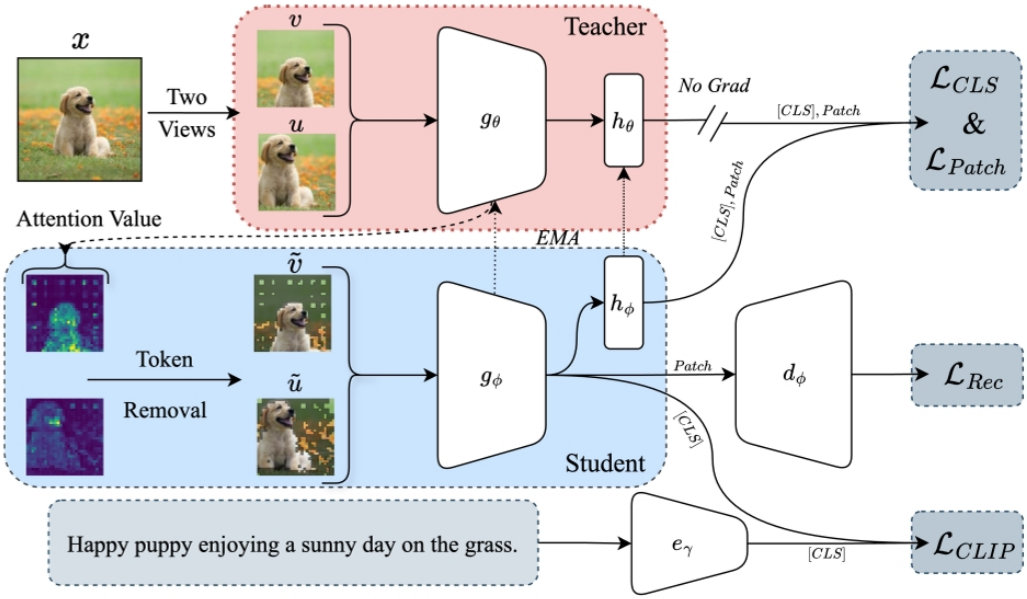
۲. حذف توکن (Attention-Based Token Removal)

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؟

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

آ) در حالتی که یادگیری در سطح Patch و Self-Distillation انجام می‌دهیم، مدل CLIP علاوه بر اینکه توکن [CLS] مدل دو classification را یاد می‌گیرد، با ماژول Patch نیز سعی می‌کند تا teacher و student مجبور به یادگیری ارتباط‌های بین هر چند عاد در سطح تصویر هست، با این کار اطلاعاتی که می‌توانست در تسیم‌های قبلی فراموش شود، حالا با این روش یادگیری می‌شود.

ب) مدل‌های Classification سنتی برای تشخیص کلاس کلی تصویر بهینه می‌شوند و بنابراین ممکن است جزئیات محلی غیرضروری برای Classification را حذف کنند. این موضوع آنها را برای وظایف Dense Prediction مانند Segmentation ضعیف می‌کند. Patch-Level Self-Distillation با نظارت مستقیم روی Patch Representationها این مشکل را کاهش می‌دهد.

۲. درباره‌ی فیلترسازی (attention-based filtering):

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

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

الف) در طول آموزش، student با یادگیری مقادیر مختلف attention بر روی پچ های مختلف، اولویت بندی اهمیت هر پچ تصویر و برای classification یاد می‌گیرد، طبیعتاً برخی قسمت های از عکس برای classification تصویر باید از اهمیت بیشتری برخوردار باشند. بنابراین به این شکل اولویت بندی می‌شود

ب) زمانی که با استفاده از attention بتوانیم اولویت بندی اهمیت به پچ های مختلف را انجام بدهیم، پس منبع پچ های کم اهمیت تر کار منطقی تری صورت می‌گیرد بنابراین با انجام این میزان محاسبات به شدت کاهش یافته و هم دقت کاهش چشمگیری پیدا نکند.

۳. در مورد فرایند بازسازی (reconstruction):

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

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

الف) توی دنیای واقعی، برخلاف شرایط آزمایشگاهی و تئوری، داده ها اغلب کیفیت بالا و اون شرایط استاندارد رو ندارند، بنابراین اگر مدل بتواند با تکنیک های بازسازی نسبت به این موارد مقاوم شود، در واقع Representation اصلی تصویر را متوجه میشود، از لحاظ شهودی هم انسان تا حد زیادی از تصاویر کم کیفیت را می تواند به خوبی طبقه بندی کند، چرا که در واقع اون نمایش زیر لایه ای اشیاء رو متوجه شده است.

ب) اشیا و صحنه‌هایی که بیشترین سود را می‌برند شامل موارد زیر هستند:

اشیای کوچک مثل سوزن

اشیای دارای اجزای باریک مثل مو

اشیای دارای مرز پیچیده مثل سیم

صحنه‌های شلوغ مثل شهر

تصاویر دوربین مداربسته یا کم‌وضوح

تصاویر پزشکی یا ماهواره‌ای که جزئیات کوچک مهم‌اند

چون در همه این موارد، از دست رفتن چند جزئیات کوچک می‌تواند نتیجه نهایی را به‌شدت خراب کند.

.۴

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

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

الف) این سه مؤلفه مکمل یکدیگرند، زیرا Attention-Based Token Removal هزینه محاسباتی را کاهش می‌دهد، ولی ممکن است مقداری اطلاعات را حذف کند. Patch-Level Self-Distillation کمک می‌کند Student با وجود این حذف‌ها، Representationهای محلی و معنایی مشابه Teacher را حفظ کند. در نهایت Pixel-Level Reconstruction مدل را مجبور می‌کند جزئیات بسیار دقیق تصویری و مرزها را نیز در Featureهای خود نگه دارد.

ب) گر Self-Distillation حذف شود، Representationهای Patch ضعیف‌تر شده و اطلاعات Local Semantic کمتر حفظ می‌شوند.

اگر Attention-Based Token Removal حذف شود، هزینه محاسباتی و مصرف حافظه افزایش می‌یابد.

اگر Pixel-Level Reconstruction حذف شود، جزئیات ریز، مرزها و ساختارهای باریک ضعیف‌تر حفظ می‌شوند و کیفیت Dense Prediction کاهش پیدا می‌کند.

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؟

روش‌هایی مثل SimCLR به Negative Samples نیاز دارند، چون فقط نزدیک کردن Representationهای Positive Pair می‌تواند به یک راه‌حل بد منجر شود که در آن شبکه برای تمام تصاویر Representation یکسان تولید کند. این پدیده Representation Collapse نام دارد. Negative Samples مدل را مجبور می‌کنند علاوه بر اینکه برای تصاویر مربوط به هم شباهت ($\mathrm{Sim}(x_i, x_j)$) بین آنهارا بالا ببرد، باید بتواند میان تصاویر غیر مرتبط فاصله بیاندازد، بنابراین دچار Collapse نمیشود.

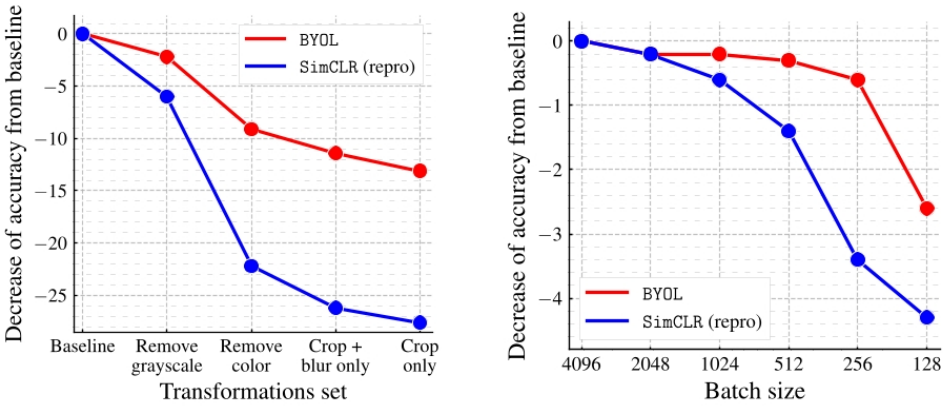
(ب) تحقیق کنید و بگویید دلیل اینکه چنین مشکلی در روش‌هایی چون BYOL رخ نمی‌دهد چیست.

BYOL بدون Negative Samples هم دچار Representation Collapse نمی‌شود، چون معماری و نحوه آموزش آن نامتقارن است. سه مؤلفه ی اصلی این موضوع عبارتند از:

- Predictor in Online branch
- Stop-Gradient on Target Branch
- EMA / Momentum Update of Target Network

Target Network مستقیماً با Gradient به‌روزرسانی نمی‌شود و تنها به‌آرامی از Online Network پیروی می‌کند. بنابراین Online Network مجبور است Representation یک هدف نسبتاً پایدار را پیش‌بینی کند و این دینامیک آموزشی از رسیدن ساده سیستم به Representation ثابت و بی‌معنا جلوگیری می‌کند.

(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).



transformations removing of Impact (ب) size batch of Impact (آ)

شکل ۲: BYOL: Bootstrap Your Own Latent

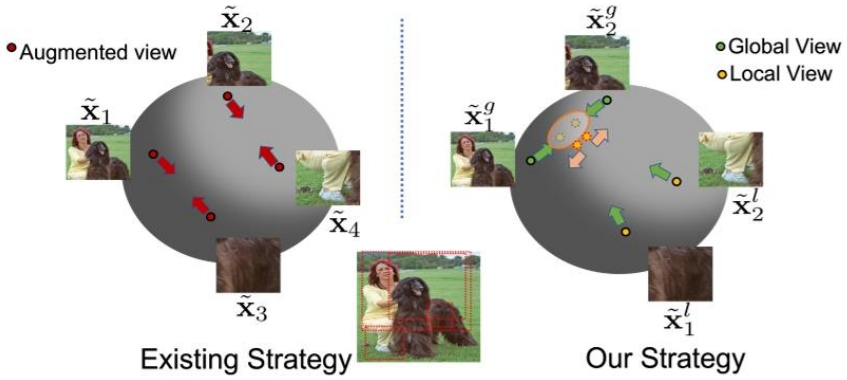
به نظر شما دلیل این موضوع چیست؟

BYOL نسبت به Batch Size مقاوم‌تر از SimCLR است، زیرا SimCLR از نمونه‌های دیگر Batch به‌عنوان Negative Samples استفاده می‌کند؛ بنابراین با کاهش Batch Size تعداد Negativeها کاهش یافته و کیفیت Contrastive Learning افت می‌کند. BYOL Negative Sample ندارد و در نتیجه وابستگی بسیار کمتری به Batch Size دارد.

همچنین BYOL نسبت به حذف بعضی Transformationها مقاوم‌تر است، زیرا یادگیری آن تنها بر Contrast بین نمونه‌ها متکی نیست و از ساختار Online–Target، Predictor، Stop-Gradient و EMA برای تولید یک Target پایدار استفاده می‌کند. بنابراین تغییر Augmentationها کمتر از SimCLR به فرایند یادگیری آن آسیب می‌زند.

بااین‌حال Augmentation همچنان برای BYOL مهم است؛ نمودار نشان می‌دهد حذف شدید آنها عملکرد BYOL را هم کاهش می‌دهد.

(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؟

Global Cropهای یک تصویر باید Representationهای نزدیک داشته باشند، چون هرکدام بخش بزرگی از همان صحنه را دیده‌اند و Semantic کلی مشابهی دارند. **"باید Global Views نزدیک باشند به همدیگر"**.

اما Local Cropها ممکن است مربوط به اجزای کاملاً متفاوت تصویر باشند. اگر همه Local Cropها نیز مجبور شوند به یک Representation مشترک نزدیک شوند، اطلاعات Part-Level و Fine-Grained از بین می‌رود. از طرفی برای Local Views شرط اصلی این هستش که مفاهیم مشترک به همدیگر نزدیک شوند، این کار Representationهایی می‌سازد که هم مفهوم کلی تصویر را می‌فهمند و هم تفاوت بین اجزای مختلف آن را حفظ می‌کنند.