دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده مهندسی برق و کامپیوتر

تمرین شماره: ۱

هوش مصنوعی قابل اعتماد

نام و نام‌خانوادگی: مرتضی ملکی‌نژاد شوشتری

شماره دانشجویی: ۸۱۰۱۰۴۲۵۶

نیم‌سال دوم سال تحصیلی ۴۰۵-۴۰۴

فهرست مطالب

۱ تفسیرپذیری ۴
۱.۱ سوالات توضیحی ۴
۱.۱.۱ PDP ۴
۲.۱.۱ LIME ۵
۳.۱.۱ Shapely ۵
۴.۱.۱ NAM ۶
۲.۱ سوالات پیاده‌سازی ۷
۱.۲.۱ مقداردهی ۷
۲.۲.۱ بخش‌بندی تصویر ۷
۳.۲.۱ تولید نمونه‌های تغییر یافته ۹
۴.۲.۱ وزن‌دهی ۹
۵.۲.۱ آموزش مدل خطی ۹
۶.۲.۱ تصویر خرس قطبی ۱۲

۲ تفسیرپذیری ۱۴
۱.۲ توصیف‌پذیری مدل‌های هوش مصنوعی با استفاده از Grad-CAM ۱۴
۱.۱.۲ Grad-CAM ۱۴
۲.۱.۲ FinerGrad-CAM ۱۹
۳.۱.۲ مقایسه ۲۲

فهرست تصاویر

۱.۱ عکس گربه مورد بررسی ۷
۲.۱ عکس گربه که به ۱۸ بخش شکسته شده است. ۸
۳.۱ چند نمونه از تصاویر تغییریافته. ۹
۴.۱ هیستوگرام احتمالات گروه گربه ۱۰
۵.۱ هیستوگرام ضرایب مدل خطی. در اکثر موارد ضریب صفر و یا نزدیک به صفر است. . ۱۱
۶.۱ خروجی heatmap مقدار ضرایب مدل خطی. ۱۲
۷.۱ خروجی heatmap مقدار ضرایب مدل خطی. ۱۳

۱.۲ پنج نمونه از مواردی که مدل به‌درستی کلاس آن‌ها را تشخیص داده است. ۱۶
۲.۲ خروجی Grad-CAM برای نمونه‌ها. ۱۶
۳.۲ خروجی Guided Grad-CAM برای نمونه‌ها. ۱۸
۴.۲ نمونه تصویر گروه ۳ که برای d تعیین شده است. این گروه مربوط به «والیبال» است. . ۲۰
۵.۲ خروجی FinerGrad-CAM برای نمونه‌ها. ۲۱

فهرست جداول

سوال ۱

تفسیرپذیری

۱.۱ سوالات توضیحی

۱.۱.۱ PDP

۱. در Partial Dependence در حال بررسی تاثیر ویژگی‌های S روی خروجی مدل هستیم. و میانگین خروجی مدل برای مقادیر مختلف ویژگی‌های S بدست می‌آید. حال اگر ویژگی‌های S با C رابطه داشته باشند، برای مثال در حالت افراطی وقتی که یک ویژگی دوبار تکرار شده باشد و یکی در S و یکی در C باشد، در حالات زیادی ورودی مدل دو مقدار برای آن ویژگی است که در دنیای واقعی رخ نمی‌دهد و باعث می‌شود مدل مجبور به برون‌یابی [1] خروجی معنادار نباشد.

برای مثال اگر در یک اپلیکیشن درخواست سفر، برای پیش‌بینی قیمت سفر، یک ویژگی مدت زمان سفر باشد و ویژگی دیگر فاصله بین مبدا و مقصد باشد،مدل برای یافتن حالتی که مبدا و مقصد نزدیک هم هستند ولی زمان سفر خیلی زیاد است مجبور است برون‌یابی کند و احتمالا خروجی بی‌معنی بدهد که برای هوش‌مصنوعی توضیح پذیر این موضوع خوب نیست.

۲. در روش ALE مقادیر مختلف ویژگی به بازه‌های کوچکی شکسته می‌شوند و سپس برای هر کدام از این بازه‌های کوچک، مقدار بیشینه و کمینه جایگزین می‌شود (بدون تغییر سایر ویژگی‌ها) و تاثیر ویژگی

[1] Extrapolate

در بازه حساب می‌شود. سپس از مقدار کم به زیاد مقادیر جمع زده می‌شود تا تاثیر تجمعی حساب شود. سپس میانگین این مقادیر را صفر می‌کنیم و نمودار آن رسم می‌شود.

۳. خیر. با توجه به اینکه مدل PDP اصلا به هم‌بستگی [۲] اهمیت نمی‌دهد و تاثیر بود و نبود ویژگی را در نظر می‌گیرد، پس اگر مدل نسبت به یک ویژگی robust باشد، به دلیل افزونگی [۳] ممکن است با تغییر یک ویژگی خیلی تغییری رخ ندهد ولی اگر چند ویژگی باهم تغییر کنند تغییر کند.

۲.۱.۱ LIME

اگر مدل مورد بررسی یک مدل مقاوم [۴] نباشد و مرز تصمیم آن بسیار به داده‌ها نزدیک و آشفته باشد، بسته به اینکه موقع نمونه‌برداری از فضای اطراف داده، چه نمونه‌هایی انتخاب شوند، ممکن است مدل خطی دو خط کاملا متفاوت را بیابد و ضرایب گویا نباشند. اما اگر مدل یک مدل مقاوم Lipchitz با ضریب L باشد، آنگاه از آنجایی که مدل تغییرات ناگهانی ندارد، پس مرز تصمیم مقدار پایدارتری دارد و با یک نمونه‌برداری دور نمونه (با توجه به ضریب L) می‌توان دید خوبی از درون مدل یافت.

۳.۱.۱ Shapely

۱. در محاسبه مقدار shapely value خروجی‌های مختلف مدل را به ازای بود و نبود ویژگی مد نظر حساب می‌کنیم و میانگین همه، مقدار shapely value را نشان می‌دهند. اگر به فرمول داده‌شده نگاه کنیم و مقدار $f(x^i)$ را درنظر بگیریم، با توجه به ویژگی‌ها در فرمول مدل خطی، روی هم تاثیر نمی‌گذارند، می‌توان از تاثیر سایر ویژگی‌ها صرفنظر کرد و فقط دید عبارت مربوط به ویژگی j ام چه تغییری می‌کند که با استفاده از فرمول داده شده داریم:

$$f(x^i) - E(f(x)) = \beta_j x_j^i - \beta_j E(X_j) = \beta_j (x_j^i - E(x_j))$$

۲. اگر فرض کنیم در یک مدل ویژگی‌ها از هم مستقل هستند، بجای بررسی 2^n حالت مختلف، می‌توان صرفا با بررسی دو حالت به تاثیر ویژگی مدنظر رسید و مقدار Shapely value آن را بدست آورد.

درواقع اگر ویژگی‌ها مستقل باشند، دیگر نیاز نیست احتمال شرطی‌ها حساب شود و اگر برای دو حالت بود و نبود ویژگی در حالتی که سایر ویژگی‌ها همه باشند تاثیر حساب شود، چون استقلال داریم می‌توان گفت:

[2] Correlation
[3] Redundancy
[4] Robust

$P(A|B) = P(A)$ پس یعنی احتمالات شرطی نیز برابر با همین مقدار می‌شود و نیازی به حساب آن‌ها و سپس میانگین گرفتن نیست.

NAM ۴.۱.۱

این معماری پیشنهاد می‌کند که هر ویژگی بصورت مجزا به یک شبکه عصبی داده شود و بعد خروجی این شبکه‌ها باهم (همراه با bias) جمع زده شوند و خروجی را بدهند. به اینصورت می‌توان مدلی قدرتمندتر از برای مثال logistic regression داشت زیرا در logistic regression هر ویژگی تنها یک ضریب می‌گیرد؛ ولی در NAM هر ویژگی از یک شبکه عصبی عمیق رد می‌شود و شبکه می‌تواند ویژگی‌های پیچیده‌تری بیابد.

همانطور که در مقاله ذکر شده است، NAM زیرمجموعه یک خانواده بزرگ‌تر از مدل‌ها به اسم Generalized Additive Model (GAM) است. یکی از فرض‌های اصلی در این مدل‌ها، فرض مستقل بودن ویژگی‌ها است. پس اگر ویژگی‌ها مستقل نباشند و برهم‌کنش این ویژگی‌ها خروجی را تعیین کند، NAM نمی‌تواند دقت خوبی بدهد.

همچنین هرچقدر تعداد ویژگی‌ها بیشتر شود، NAM تفسیر پیچیده‌تری از ویژگی می‌دهد که می‌تواند باعث مشکل شود. یک راه‌حلی که در مقاله ارائه می‌شود، استفاده از ویژگی‌های پیچیده‌تر است (برای مثال بجای عکس، لبه‌های آن داده شود یا از یک شبکه پیچشی گذر کند و سپس ورودی داده شود.) که در اینصورت دقت مدل بهبود می‌یابد و نیاز به پیچیدگی خود مدل NAM کمتر می‌شود ولی به شرط اینکه ویژگی‌های ورودی آن خود تفسیرپذیر باشند.

در NAM چون فضای توزیع توصیف می‌شود (و نه فقط فضای نمونه‌ای) همچنان می‌توان دید در مواردی مدل مجبور به برون‌یابی [۵] می‌شود. ولی وزن این نواحی در نمودارها با شدت رنگ نشان داده شده‌اند و می‌توان در نمودار دید که کجا داده کمتر بوده است پس این مسئله نسبت به PDP که اصلا توجهی به توزیع داده ندارد بهبود دارد.

[5] extrapolate

۲.۱ سوالات پیاده‌سازی

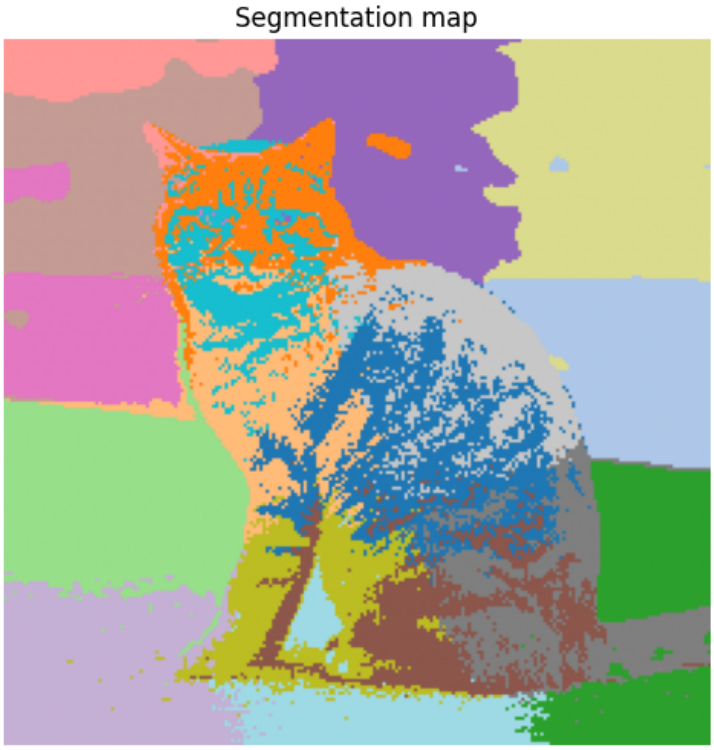
۱.۲.۱ مقداردهی

شکل ۱.۱: عکس گربه مورد بررسی

مدل با احتمال 13.68 درصد، عکس را متعلق به دسته ۲۸۲ می‌داند.

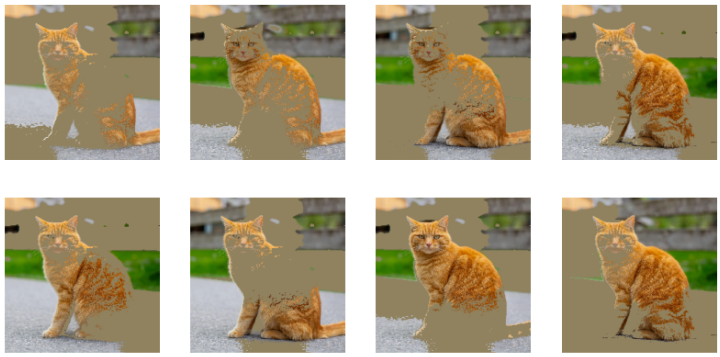
۲.۲.۱ بخش‌بندی تصویر

کد داده‌شده با استفاده از الگوریتم K means در هر مرحله، ابتدا با محاسبه فاصله بین نزدیک‌ترین مرکز دسته و هر پیکسل‌ها تعیین می‌کند هر پیکسل متعلق به چه دسته‌ای است، میانگین مختصات اعضای هردسته را می‌یابد و مرکز دسته را به آنجا تغییر می‌دهد. عملا این‌کار یک color segmentation است.



شکل ۲.۱: عکس گربه که به ۱۸ بخش شکسته شده است.

۳.۲.۱ تولید نمونه‌های تغییر یافته



شکل ۳.۱: چند نمونه از تصاویر تغییریافته.

۴.۲.۱ وزن‌دهی

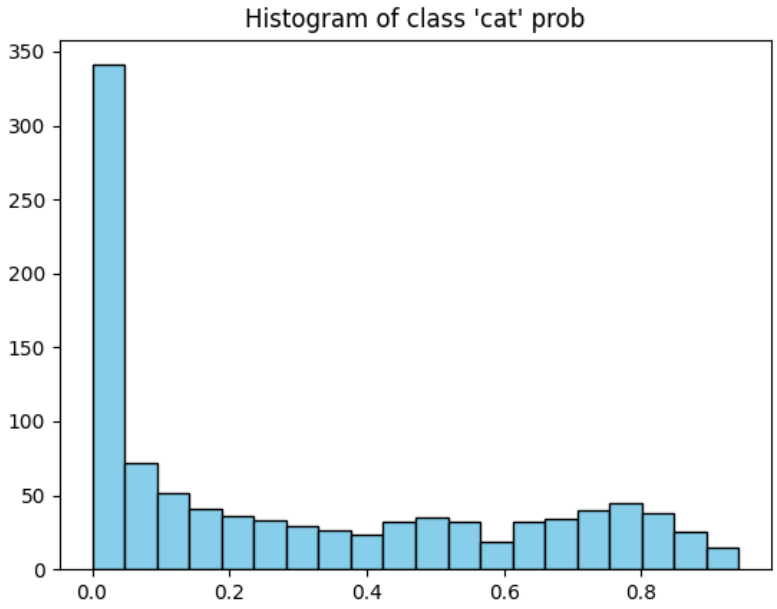
(برای فهم بهتر سوال از هوش مصنوعی کمک گرفته شد.)

این وزن‌دهی به این دلیل رخ می‌دهد که هرچقدر از تصویر اصلی دورتر می‌شویم، تصاویر وزن کم‌تری بگیرند. به این صورت هنگام آموزش مدل خطی نیز تاثیر کمتری می‌گذارند. زیرا اگر با تغییر کمی در تصویر خروجی مدل خیلی تغییر کند، پس احتمالا ویژگی‌هایی که تغییر کردند اهمیت بیشتری برای مدل داشتند.

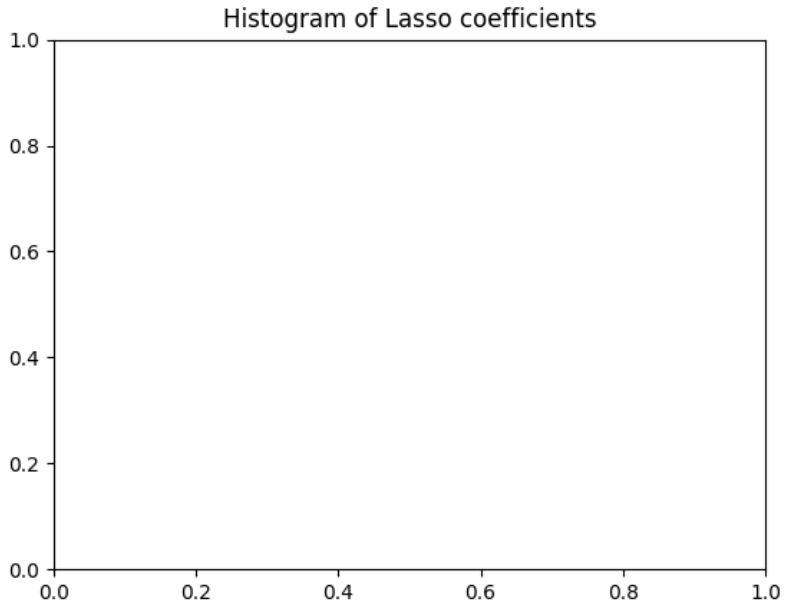
۵.۲.۱ آموزش مدل خطی

با آموزش مدل خطی، در تصویر

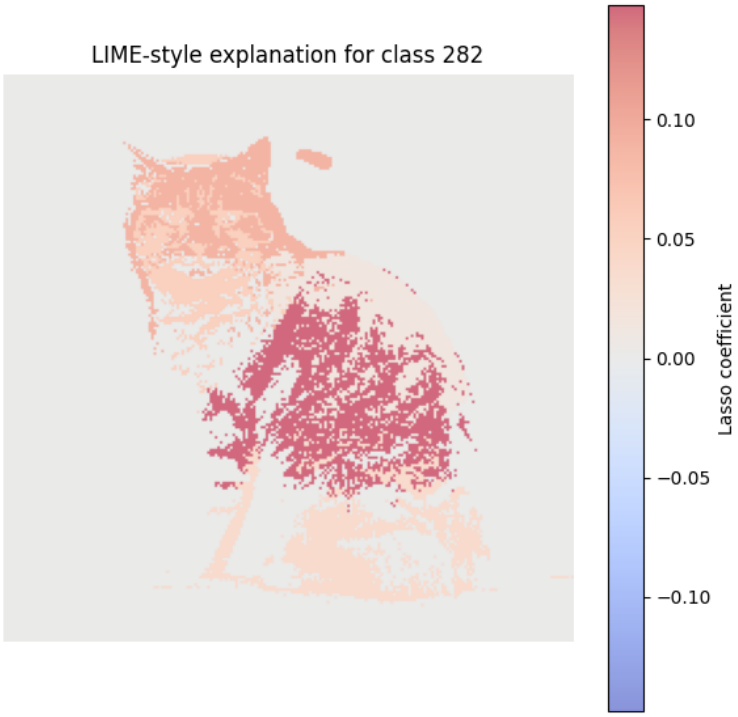
می‌توان دید که مدل به اکثر نواحی توجه کمی دارد ولی نواحی که بدن و سر گربه را تشخیص می‌دهند وزن بیشتری می‌گیرند.



شکل ۴.۱: هیستوگرام احتمالات گروه گربه



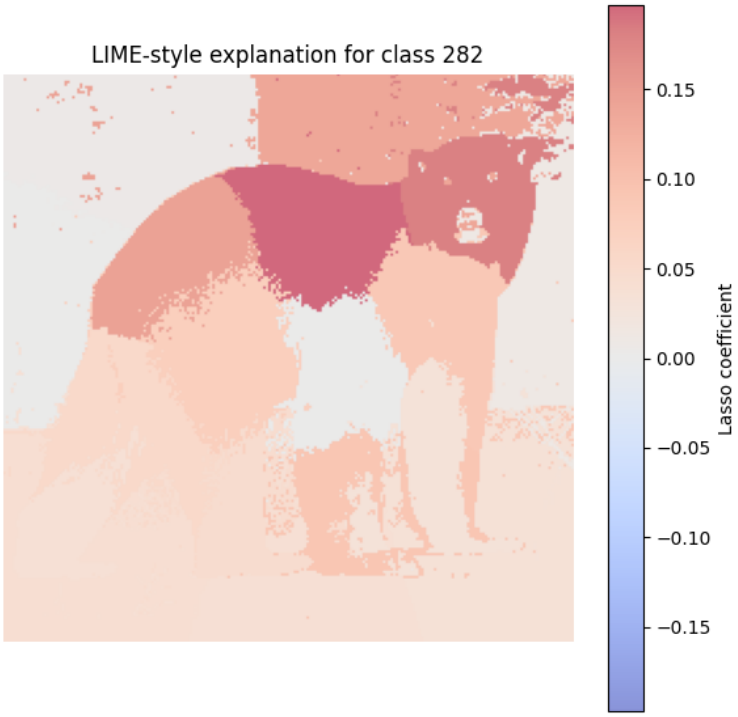
شکل ۵.۱: هیستوگرام ضرایب مدل خطی. در اکثر موارد ضریب صفر و یا نزدیک به صفر است.



شکل ۶.۱: خروجی heatmap مقدار ضرایب مدل خطی.

۶.۲.۱ تصویر خرس قطبی

در heatmap تصویر مربوط به خرس قطبی، خروجی به اندازه تصویر گربه تمیز نیست. مدل تمرکز بیشتری روی سر و بخشی از بدن خرس دارد ولی روی پس‌زمینه و سایر موارد نیز تمرکز دارد. این موضوع خوب نیست و احتمالا به سوگیری مدل اشاره دارد (برای مثال در پس‌زمینه برفی و درختان شمالی احتمال خرس بیشتر است.)



شکل ۷.۱: خروجی heatmap مقدار ضرایب مدل خطی.

سوال ۲

تفسیرپذیری

۱.۲ توصیف‌پذیری مدل‌های هوش مصنوعی با استفاده از Grad-CAM

۱.۱.۲ Grad-CAM

پرسش و پاسخ

۱. زیرا Grad-CAM همان ایده و هدف CAM را دارد: تولید نقشه‌ی حرارتی[1] که نشان دهد هر ناحیه از تصویر چقدر در تصمیم‌گیری مدل مؤثر است.

CAM فقط در مدل‌هایی قابل استفاده است که لایه‌ی Global Average Pooling به همراه لایه‌ی آخر Fully Connected داشته باشند.

اما Grad-CAM با بهره‌گیری از این ایده که به جای وزن‌های لایه‌ی FC می‌توان از مشتق نمره‌ی کلاس نسبت به نقشه‌های ویژگی[2] استفاده کرد، روی تقریباً هر مدل کانولوشنی بدون نیاز به تغییر معماری قابل اعمال است.

[1] heatmap
[2] feature maps

در واقع اگر مدلی دارای معماری مورد نیاز CAM باشد، Grad-CAM به فرم CAM فرو می‌ریزد و خروجی آن‌ها بسیار مشابه خواهد بود.

۲. هدف از اعمال ReLU در انتهای Grad-CAM این است که فقط نواحی‌ای در نقشه‌ی حرارتی نمایش داده شوند که احتمال کلاس مورد نظر را **تقویت** می‌کنند.

اگر گرادیان[3] یک ناحیه منفی باشد، یعنی افزایش فعالیت آن ناحیه باعث کاهش نمره‌ی کلاس هدف می‌شود؛ بنابراین آن ناحیه در جهت تصمیم فعلی مدل عمل نمی‌کند. چنین نواحی‌ای معمولاً برای تصمیم‌گیری فعلی مفید نیستند.

بنابراین با اعمال ReLU، تأثیر مقادیر منفی حذف شده و نقشه‌ی نهایی فقط نواحی «مثبت» و مؤثر را نشان می‌دهد.

(دقت شود که گرادیان منفی لزوماً به معنای «تشخیص کلاس دیگر» نیست، بلکه یعنی آن ناحیه با تصمیم فعلی هم‌جهت نیست.)

۳. Grad-CAM نشان می‌دهد مدل برای پیش‌بینی یک کلاس به کدام نواحی تصویر توجه می‌کند.

اگر مدل به جای ویژگی‌های معنادار (مانند لباس یا ابزار کار در تشخیص «پرستار» در مقابل «دکتر»)، روی ویژگی‌های نادرست و جانبی مانند جنسیت، نژاد، پس‌زمینه یا watermark تمرکز کند، این نواحی در نقشه‌ی حرارتی پررنگ دیده می‌شوند.

بنابراین با تحلیل نقشه‌های حرارتی روی نمونه‌های مختلف (به‌خصوص نمونه‌هایی که برچسب آن‌ها با ویژگی‌های زمینه‌ای در تضاد است) می‌توان وجود سوگیری[4] را کشف کرد.

هرچند Grad-CAM به تنهایی سوگیری را کمّی نمی‌کند، اما ابزار قدرتمندی برای بازشناسی بصری آن است.

۴. Guided Grad-CAM ترکیبی از دو روش است:

- **Guided Backpropagation**: در هنگام محاسبه‌ی گرادیان، در لایه‌های ReLU علاوه بر شرط مثبت بودن ورودی، شرط مثبت بودن گرادیان برگشتی نیز اعمال می‌شود. این کار باعث می‌شود گرادیان نهایی تا سطح پیکسل‌های ورودی فقط از مسیرهایی عبور کند که هم فعال بوده‌اند و هم تأثیر مثبت داشته‌اند.

[3] gradient
[4] bias

نتیجه، یک نگاشت با جزئیات پیکسلی بالا است که بیشتر ساختارهای محلی و لبه‌های تصویر را برجسته می‌کند؛ هرچند الزاماً تفکیک‌کننده‌ی کلاس نیست.

- **Grad-CAM**: نقشه‌ی حرارتی با رزولوشن پایین که نواحی مهم را مشخص می‌کند، اما به دلیل بزرگ‌نمایی[۵] شدن، فاقد دقت پیکسلی است.

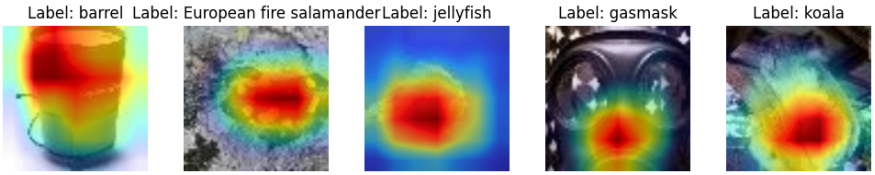
در Guided Grad-CAM، خروجی Guided Backpropagation به‌صورت نقطه‌به‌نقطه در خروجی Grad-CAM (که به همان ابعاد بزرگ‌نمایی شده است) ضرب می‌شود.

نتیجه، نقشه‌ای با **جزئیات پیکسلی بالا** و در عین حال **تفکیک‌کننده‌ی کلاس**[۶] است.

به بیان ساده‌تر، از Grad-CAM به‌عنوان **ماسک**[۷] استفاده می‌شود که تعیین می‌کند «کدام نواحی مهم هستند» و از Guided Backpropagation برای مشخص کردن «کدام پیکسل‌ها» در آن نواحی اهمیت دارند استفاده می‌شود.

پیاده‌سازی

شکل ۱.۲: پنج نمونه از مواردی که مدل به‌درستی کلاس آن‌ها را تشخیص داده است.



شکل ۲.۲: خروجی Grad-CAM برای نمونه‌ها.

[5] upscale
[6] class-discriminative
[7] mask

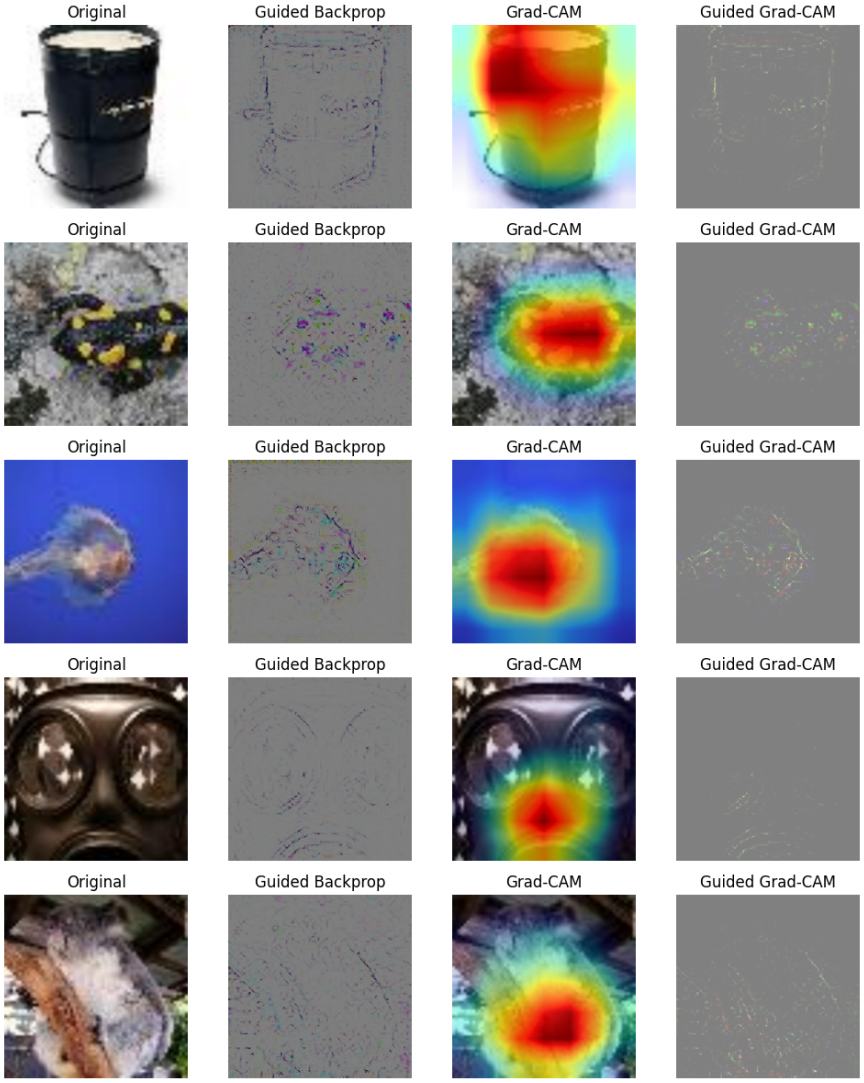
همان‌طور که در سؤال خواسته شد، Grad-CAM در نوت‌بوک q2_gradcam پیاده‌سازی شد. در تصویر ۲.۲ می‌توان دید که مدل روی چه نواحی‌ای تمرکز کرده است. برای مثال، در تصویر مربوط به سمندر، مدل روی نواحی‌ای که حیوان در آن‌ها قرار دارد تمرکز بیشتری نشان می‌دهد.

البته می‌توان دید که مدل کاملاً دقیق نیست و برای مثال در تصاویر مربوط به سطل و ماسک تنها روی بخش کوچکی از تصویر تمرکز دارد.

با پیاده‌سازی Guided Grad-CAM و Guided Backpropagation و مقایسه‌ی خروجی آن‌ها که در شکل ۳.۲ آمده است، می‌توان دید که خروجی‌ها تقریباً شبیه لبه‌ها و ساختارهای محلی تصویر هستند که مطابق انتظار است.

یک تفاوت جالب که در تصویر عروس دریایی دیده می‌شود این است که در Guided Backpropagation خروجی دارای مؤلفه‌های آبی و قرمز بیشتری است که در تصویر نهایی به صورت رنگ‌های متمایل به بنفش دیده می‌شوند؛ ولی در Guided Grad-CAM از آنجا که مدل تشخیص می‌دهد زمینه‌ی آبی نقش زیادی در نواحی مؤثر برای تصمیم‌گیری ندارد، بخش‌های سبز و قرمز بیشتر برجسته می‌شوند.

تفاوت دیگر که در خروجی‌های مقاله نیز مشاهده می‌شود این است که Guided Backpropagation لبه‌ها را بهتر مشخص می‌کند، ولی Guided Grad-CAM فقط ساختارهایی را برجسته می‌کند که در انتخاب کلاس نیز مؤثر هستند.



شکل ۳.۲: خروجی Guided Grad-CAM برای نمونه‌ها.

FinerGrad-CAM ۲.۱.۲

پرسش و پاسخ

۱. زیرا Grad-CAM نواحی‌ای را نشان می‌دهد که باعث تقویت کلاس شده‌اند، اما این نواحی لزوماً متمایزکننده[8] نیستند.

مثالی که در مقاله مطرح می‌شود مربوط به دسته‌بندی پرندگان است؛ ممکن است دو کلاس مختلف هر دو به بخش آبی بدن پرنده وزن بالایی بدهند. در نتیجه، این ویژگی اگرچه برای هر دو کلاس مهم است، اما الزاماً عامل تمایز میان آن‌ها نیست.

۲. ادعای مطرح‌شده در مقاله این است که مغز هنگام تشخیص یک نمونه در میان گروه‌های مختلف، به دنبال ویژگی‌های متمایزکننده می‌گردد.

FinerGrad-CAM نیز با نشان دادن نواحی‌ای که احتمال کلاس هدف را بدون تقویت سایر کلاس‌های مشابه افزایش می‌دهند، تلاش می‌کند ویژگی‌های یکتای متمایزکننده را مشخص کند.

به جای استفاده‌ی مستقیم از گرادیان کلاس هدف، از گرادیان‌های کنتراستی استفاده می‌شود تا ویژگی‌هایی که بین کلاس هدف و کلاس‌های مشابه مشترک هستند وزن کمتری بگیرند.

۳. هرچه مقدار γ بیشتر باشد، FinerGrad-CAM نواحی کوچک‌تر و متمایزکننده‌تری را نمایش می‌دهد؛ یعنی نواحی‌ای که فقط احتمال کلاس هدف را افزایش می‌دهند.

اما در مقابل، بخشی از اطلاعات کلی مورد استفاده‌ی مدل برای انتخاب کلاس از دست می‌رود.

برای مثال، در کاربردهای زیست‌شناسی اگر هدف یافتن مهم‌ترین نواحی مؤثر در پیش‌بینی کلاس باشد، مقدار گاما باید کوچک انتخاب شود. ولی اگر هدف یافتن ویژگی‌هایی باشد که باعث ترجیح یک کلاس نسبت به کلاس‌های مشابه شده‌اند، استفاده از مقادیر بزرگ‌تر گاما مناسب‌تر است.

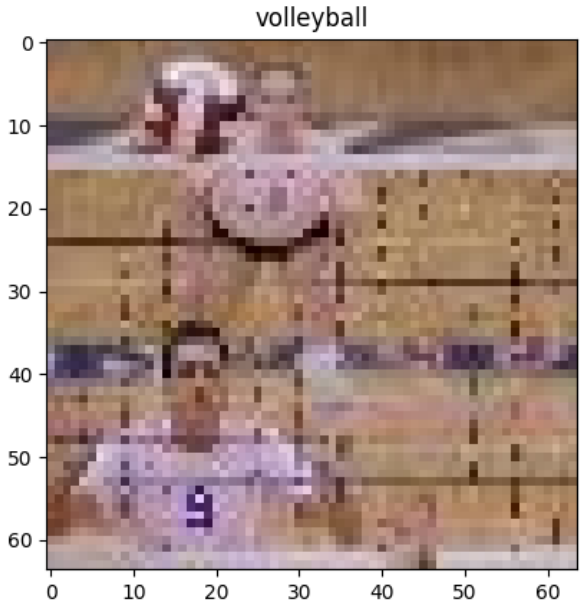
۴. در مثالی که مقاله مطرح می‌کند، در یک مجموعه‌داده‌ی تصویر–توضیح بررسی می‌شود که آیا مدل به بخش‌هایی از تصویر که در توضیح ذکر شده‌اند توجه می‌کند یا خیر.

این موضوع باعث می‌شود فارغ از دقت نهایی مدل، اعتمادپذیری تصمیم‌های آن افزایش یابد و رفتار مدل پایدارتر باشد.

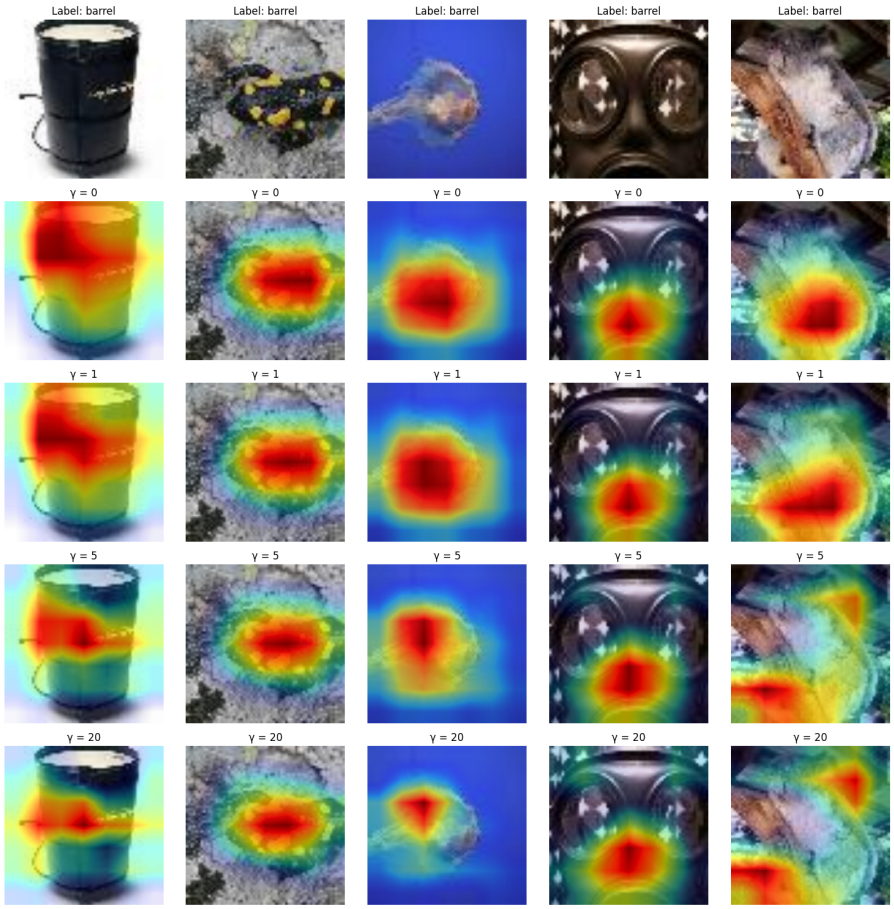
[8] Discriminative

اما اگر مدل بدون توجه به توضیحات به پاسخ صحیح برسد، احتمال بیش‌برازش روی یک ویژگی خاص افزایش می‌یابد و سامانه احتمالاً از پایداری و مقاومت کمتری برخوردار خواهد بود.

پیاده‌سازی



شکل ۴.۲: نمونه تصویر گروه ۳ که برای d تعیین شده است. این گروه مربوط به «والیبال» است.



شکل ۵.۲: خروجی FinerGrad-CAM برای نمونه‌ها.

همان‌طور که در تصویر ۵.۲ مشاهده می‌شود، با افزایش گاما، نواحی مشخص‌شده کوچک‌تر می‌شوند و ویژگی‌های شاخص‌تر تصویر باقی می‌مانند.

البته این مجموعه‌داده و این شیوه‌ی آزمایش، داده‌ی مناسبی برای نمایش کامل توانایی FinerGrad-CAM نیست؛ زیرا کلاس‌ها شباهت زیادی به یکدیگر ندارند و برای مشاهده‌ی تغییرات محسوس، باید از مقادیر بسیار بزرگ گاما استفاده شود.

(برای مثال در اینجا از مقادیر ۱، ۵ و ۱۰۰ استفاده شد، در حالی که مقدار متعارف گاما معمولاً بین صفر و یک است.)

در تصویر ۵.۲ خروجی مربوط به $\gamma = 0$ نیز رسم شده است که عملاً همان خروجی Grad-CAM است.

نکته‌ی قابل مشاهده‌ی دیگر در تصویر مربوط به کوالا است. می‌توان دید که با افزایش زیاد گاما، ناحیه‌ی قبلی دیگر برجسته نمی‌شود و مدل روی سایر نواحی متمایزکننده‌ای که در خروجی اولیه چندان مشخص نبودند نیز تمرکز می‌کند. این موضوع ایده‌ی اصلی FinerGrad-CAM را به‌خوبی نشان می‌دهد.

۳.۱.۲ مقایسه

در تصویر ۵.۲ خروجی سطر مربوط به $\gamma = 0$ برابر خروجی Grad-CAM و خروجی سطر مربوط به $\gamma = 1$ برابر با خروجی استاندارد FinerGrad-CAM است.

(سایر مقادیر گاما خارج از بازه‌ی متعارف هستند و صرفاً برای ایجاد تغییرات قابل مشاهده انتخاب شده‌اند.)

می‌توان دید که در FinerGrad-CAM در بعضی تصاویر، مانند تصویر سطل، ناحیه‌ی کوچک‌تری مشخص شده است. این موضوع نشان می‌دهد که بخش‌هایی از ناحیه‌ی اولیه احتمالاً چندان متمایزکننده نبوده‌اند و احتمال چند کلاس مختلف را هم‌زمان افزایش می‌داده‌اند.

در مقابل، در برخی تصاویر مانند کوالا یا ماسک، ناحیه‌ی برجسته‌شده بزرگ‌تر شده است. وجود این ناحیه‌ی جدید که در خروجی قبلی مشخص نبود، نشان می‌دهد بخشی از ناحیه‌ی اولیه میان چند کلاس مشترک بوده است. بنابراین با استفاده از گرادیان‌های کنتراستی و حذف فعال‌سازی‌های مشترک بین کلاس‌ها، می‌توان نواحی واقعاً متمایزکننده را بهتر مشخص کرد.

همچنین در تصویر سمندر (و تا حدی عروس دریایی) مشاهده می‌شود که افزایش گاما تأثیر زیادی بر خروجی ندارد، مگر در مقادیر بسیار بزرگ. این موضوع نشان می‌دهد که مدل از ابتدا روی ویژگی‌های نسبتاً یکتای سمندر تمرکز کرده است و ویژگی مشترک قابل‌توجهی با کلاس مرجع انتخاب‌شده وجود ندارد.