

ارائه روشی برای افزایش تمرکز شبکه عصبی عمیق بر سوژه اصلی مبتنی بر توضیح‌پذیری برای مدل‌های طبقه‌بندی

علی جعفری^{۱*}، شریف حسینی^۲

تاریخ دریافت: ۱۴۰۵/۰۳/۱۱

تاریخ پذیرش: ۱۴۰۵/۰۶/۲۳

چکیده

با پیشرفت استفاده از هوش مصنوعی و یادگیری عمیق در سال‌های اخیر در صنایع حساس و حیاتی مانند سامانه‌های فرماندهی و کنترل نظامی، وجود عناصری مانند قابلیت اعتماد، توانایی تفسیر نتایج و درک ساختار تصمیم‌گیری مدل‌های آموزش داده شده بیش از پیش احساس شده است. به همین منظور راهکارهایی مانند الگوریتم‌های توضیح‌پذیری مدل مانند LIME، SHAP، Grad-CAM و... ارائه شده‌اند. پس از آن در سال‌های اخیر رویکردی برای افزودن راهنمایی‌های توضیح‌پذیری در طی فرآیند آموزش مدل با نام اختصاری EGL نیز مطرح شده است. که هدف آن افزایش دقت پیش‌بینی و کسب اعتماد کاربر نسبت به نتایج مدل است. در این پژوهش راهکاری نوین با استفاده از رویکرد EGL برای افزایش تمرکز مدل‌های طبقه‌بندی بر سوژه اصلی ارائه شده است. تابع هدف پیشنهادی دارای جملاتی برای طبقه‌بندی، تمرکز بر سوژه اصلی و کاهش توجه به نقاط فاقد اهمیت در تصویر است. روش پیشنهادی بر روی شبکه ResNet^{۱۸} با مجموعه داده TinyImageNet آموزش مجدد داده شده است و نتایج آن از دو جنبه دقت پیش‌بینی و کیفیت توضیح‌پذیری ارتقا یافته است. همچنین در مقایسه با پژوهش‌های پیشین روش پیشنهادی موجب افزایش تمرکز شبکه عصبی بر روی سوژه اصلی و کاهش تمرکز مدل بر ویژگی‌های پس‌زمینه شده است.

واژگان کلیدی: شبکه‌های عصبی عمیق، توضیح‌پذیری، تفسیر پذیری، یادگیری مبتنی بر توضیح‌پذیری EGL، طبقه‌بندی تصاویر.

^۱ نویسنده مسئول، استادیار گروه علمی هوش مصنوعی، مجتمع دانشگاهی برق و رایانه، دانشگاه صنعتی مالک اشتر iustuser@mut.ac.ir

^۲ پژوهشگر دانشکده مهندسی برق و کامپیوتر، دانشگاه صنعتی مالک اشتر

ارائه روشی برای افزایش تمرکز شبکه عصبی عمیق بر سوژه اصلی مبتنی بر توضیح پذیری برای مدل‌های طبقه‌بندی

۱. کلیات

هوش مصنوعی به عنوان یکی از پیشرفته ترین و تاثیرگذارترین فناوری های نوین با ورود به عرصه های نظامی به ویژه سامانه های فرماندهی و کنترل، تحولی بنیادین را رقم زده است. سامانه های فرماندهی و کنترل مدرن با بهره گیری از توان تحلیلی و تصمیم گیری سریع هوش مصنوعی وارد مرحله ای جدید از خودکارسازی و دقت بالا شده اند. ولی با توجه به حساسیت بسیار زیاد سامانه های فرماندهی و کنترل نظامی، عدم شفافیت موجود در شبکه های عمیق مورد استفاده در هوش مصنوعی، باعث غیر قابل اعتماد شدن این فناوری در این حوزه شده است. در این راستا قابلیت توضیح پذیری در مدل های یادگیری ماشین به یکی از محورهای اصلی پژوهش های هوش مصنوعی مدرن تبدیل شده است.

معمولاً روش های توضیح پذیری به دو دسته ی روش های پس توضیحی^۱ و ذاتا توضیح پذیر^۲ تقسیم می شوند [۱]. در روش های پس توضیحی، تصمیمات مدل از طریق تحلیل اهمیت ویژگی ها پس از اتمام فرآیند آموزش تفسیر می شوند [۲]. با این حال، این روش ها عمدتاً فقط به تحلیل پیش بینی های مدل می پردازند و تأثیری در بهبود عملکرد یا فرآیند یادگیری مدل ندارند. در پاسخ به این محدودیت، رویکردی به نام یادگیری مبتنی بر توضیح پذیری^۳ (EGL) معرفی شده است که در آن توضیحات ارائه شده توسط روش های پس توضیحی مستقیماً وارد فرآیند یادگیری می شوند. این کار منجر به ساخت مدل هایی می شود که ذاتاً قابل تحلیل هستند و بر ویژگی های معنایی و قابل اعتماد تمرکز دارند. این روش ها با هم تراز کردن هدف آموزشی با استدلال های قابل درک برای انسان، به طور هم زمان قابلیت تحلیل و عملکرد مدل را در وظایف پایین دستی بهبود می دهند [۱].

با این حال، یکی از چالش های اساسی در یادگیری مبتنی بر توضیح پذیری EGL، نبود یک چارچوب ساخت یافته برای

هدایت شبکه عصبی به سمت تمرکز هدفمند بر سوژه ی اصلی تصویر و استخراج ویژگی های کلیدی، متمایز و معنادار است. این چالش به ویژه در مجموعه داده هایی که کلاس ها دارای هم پوشانی بالا در فضای ویژگی هستند، نمود بیشتری پیدا می کند. در چنین شرایطی، کنترل فرآیند یادگیری مدل بر اساس اهداف توضیح پذیری به درستی انجام نمی شود و شبکه تمایل دارد به سرنخ های زمینه ای یا نویز های مشترک میان کلاس ها اتکا کند. به عنوان نمونه، در مجموعه داده ای شامل کلاس های «قورباغه دم دار» و «قورباغه گاوی»، ویژگی های استخراج شده توسط لایه های کانولوشنی به شدت به یکدیگر نزدیک بوده و تمایز مؤثر میان این دو کلاس برای شبکه های عصبی به فرآیندی پیچیده تبدیل می شود.

رفع این مسئله مستلزم تقویت مؤلفه ی هوش^۴ از طریق بهره گیری از فراداده های حقیقت مطلق^۵ توضیح پذیر است؛ داده هایی که علاوه بر نمونه های ورودی و برچسب های خروجی، حاوی اطلاعات صریح درباره ی منطق تصمیم گیری مدل، نواحی معنادار تصویر و شواهد بصری مرتبط با هر پیش بینی هستند. این نوع فراداده ها نقش یک لایه ی ارتباطی^۶ میان انسان و مدل را ایفا می کنند و امکان انتقال دانش سطح بالا درباره ی «چرایی» و «چگونگی» تصمیم گیری را فراهم می سازند [۳]. در نتیجه، مدل نه تنها قادر به انجام وظیفه ی پیش بینی خواهد بود، بلکه می تواند مسیر استدلال خود را نیز به صورت شفاف و قابل تحلیل ارائه دهد. از سوی دیگر، در حوزه ی داده های تصویری، فقدان مجموعه داده هایی با حقیقت مطلق توضیح پذیری، بار اصلی تولید این اطلاعات را بر دوش نیروی انسانی و زیرساخت های محاسباتی^۷ قرار می دهد. این امر مستلزم افزودن حاشیه نویسی های دستی نظیر مرزبندی دقیق اشیاء، ناحیه بندی بخش های کلیدی تصویر یا تعیین اهمیت پیکسلی نواحی مختلف در تصمیم مدل است. تولید چنین داده هایی معمولاً زمان بر و پرهزینه بوده و نیازمند دانش تخصصی و

ground truth^۵
Communications^۶
Computers^۷

Post-hoc^۱
Intrinsic^۲
Explanation-Guided Learning (EGL)^۳
Intelligence^۴

بررسی دقیق هر نمونه به صورت مجزا است. در نتیجه، نبود یک چارچوب منسجم در EGL، به عنوان یک گلوگاه اساسی در توسعه مدل های شفاف، قابل اعتماد و قابل کنترل مطرح می شود.

۱-۱. مقدمه و هدف

در این پژوهش با بهره گیری از EGL، راهکاری به منظور افزایش تمرکز مدل های طبقه بندی نسبت به سوژه اصلی و استخراج ویژگی های کلیدی و قابل تمایز از تصویر ارائه شده است. به همین منظور برای تولید مجموعه داده شامل حقیقت مطلق توضیح پذیری و کاهش هزینه های زمانی و انسانی، روش ارائه شده مبتنی بر خودکارسازی تولید فراداده های توضیح پذیر ارائه شده است. در گام نخست، از مدل های از پیش آموزش دیده و قابل اعتماد در حوزه بینایی ماشین، به ویژه در وظایفی مانند بخش بندی تصویر^۱، استفاده شده است تا به صورت خودکار ماسک های دودویی مرتبط با اشیای اصلی در هر تصویر تولید شوند. این ماسک ها نشان دهنده نواحی کلیدی در تصویر هستند که بیشترین تأثیر را در تصمیم مدل دارند. سپس مجموعه داده ای ترکیبی شامل تصویر اصلی، ماسک دودویی استخراج شده، و برجسب نهایی تصویر ایجاد گردید. این ترکیب، نوعی حقیقت مطلق تقریبی برای توضیح پذیری ایجاد می کند که با دقت بالایی به واقعیت نزدیک است. در ادامه، این داده های ترکیبی به عنوان ورودی برای آموزش شبکه های عصبی با رویکرد یادگیری مبتنی بر توضیح پذیری مورد استفاده قرار گرفتند تا مدل نه تنها در وظیفه طبقه بندی عملکرد بهتری داشته باشد، بلکه قابلیت تبیین تصمیمات خود را نیز در سطح پیکسلی یا ناحیه ای ارتقا دهد.

در این پژوهش، این سؤال مطرح شده است که آیا می توان با ترکیب قابلیت های مدل SAM^۲ در تولید خودکار ماسک های توضیح پذیری و طراحی یک تابع ضرر مبتنی بر تمرکز، مدل های طبقه بندی را به یادگیری ویژگی هایی هدایت کرد که توجه بیشتری به سوژه اصلی تصویر داشته باشند، بدون آنکه عملکرد طبقه بندی کاهش یابد؟

برای پاسخ به این سؤال، روشی مبتنی بر ترکیب SAM^۲ و یک تابع ضرر جدید (FS Loss) ارائه شده است که با استفاده از داده های توضیح پذیری تولید شده به صورت خودکار، مدل را در فرآیند آموزش به تمرکز بیشتر بر نواحی مرتبط با سوژه اصلی هدایت می کند. نتایج نشان می دهد که روش پیشنهادی ضمن افزایش تمرکز مدل های طبقه بندی بر سوژه اصلی و استخراج ویژگی های کلیدی و متمایز، وابستگی به مداخله انسانی در تولید داده های توضیح پذیری را نیز کاهش می دهد. این ویژگی، امکان تولید مجموعه داده های استاندارد، مقیاس پذیر و قابل اتکا برای ارزیابی کمی روش های توضیح پذیری را فراهم می سازد.

در مجموع، نوآوری اصلی این پژوهش تابع ضرر پیشنهادی FS Loss به منظور هدایت مدل به یادگیری متمرکزتر است. این پژوهش گامی در جهت توسعه مدل های طبقه بندی با توضیح پذیری ذاتی بیشتر، شفافیت بالاتر و قابلیت اعتماد بیشتر در کاربردهای مبتنی بر یادگیری عمیق محسوب می شود.

روش پیشنهادی و نتایج به دست آمده در این پژوهش در ۵ بخش شرح داده شده است که به ترتیب شامل موارد زیر است:

- پژوهش های پیشین: در این بخش مروری بر کارهای پژوهشی پیشین صورت گرفته است.
- روش پیشنهادی: در این بخش ایده و نوآوری اصلی روش ارائه شده برای دستیابی به هدف افزایش تمرکز مدل بر روی سوژه اصلی شرح داده شده است.
- آزمایشات: در این فصل آزمایشات انجام گرفته برای روش پیشنهادی مطرح شده اند.
- نتایج: در این بخش نتایج آزمایشات انجام شده ذکر شده است.
- بحث و نتیجه گیری: در این بخش نتیجه گیری نهایی از روش ارائه شده و تحلیل و تفسیر نتایج به دست آمده قرار داده شده است.

۲-۱. پژوهش های پیشین

روش آموزش همراه با راهنمایی های توضیح پذیری

ارائه روشی برای افزایش تمرکز شبکه عصبی عمیق بر سوژه اصلی مبتنی بر توضیح پذیری برای مدل های طبقه بندی

هدف EGL بهینه سازی همزمان عملکرد پیش بینی مدل و توضیحات متناظر آن است. به طور معمول، روش های EGL از تابع هدف زیر برای آموزش مدل f استفاده می کنند:

$$\min_f E_{(x,y)} \left[\underbrace{\mathcal{L}_{cls}(f(x), y)}_{\text{task supervision}} + \underbrace{\alpha \mathcal{L}_{exp}(g(f, \langle x, y \rangle), M)}_{\text{explanation supervision}} + \underbrace{\beta \Omega(g(f, \langle x, y \rangle))}_{\text{explanation regularization}} \right]$$

که در آن $g(f, \langle x, y \rangle)$ توضیح مدل f در ارتباط با نمونه x و برچسب y است. M نشان دهنده «توضیح درست» می باشد که معمولاً توسط حاشیه نویسی انسانی تعیین می شود [۱]. این تابع هدف شامل سه جزء کلیدی است:

- **تابع ضرر نظارت بر وظیفه^۱:** مانند آنتروپی مقاطع در وظایف طبقه بندی که مدل را به سمت پیش بینی های دقیق تر هدایت می کند.
- **تابع ضرر نظارت بر توضیح^۲:** که دانش پیشینی صریح را وارد فرآیند می کند و تضمین می نماید توضیحات مدل با آنچه یک توضیح صحیح باید باشد هم راستا باشد.
- **تابع ضرر متعادل سازی توضیح^۳:** که ویژگی های مطلوب برای توضیحات را تحمیل می کند، مانند ترویج تنگی (sparsity) و متمرکز کردن توجه بر مرتبط ترین ویژگی ها.

روش FocusEx

این روش یک چارچوب خود نظارتی EGL است که با کاهش حواس پرتی ناشی از نویز پس زمینه، مدل را بر اهداف اصلی متمرکز می کند، بدون آنکه نیاز به برچسب گذاری اضافی باشد. FocusEx به طور یکپارچه با مدل های از پیش آموزش دیده در مقیاس بزرگ^۴ ادغام می شود. این مدل ها با مشخص کردن ویژگی های کلیدی، نوعی نظارت ضعیف فراهم می کنند. نواحی برجسته شده به عنوان مناطق مورد توجه، مدل را

هدایت می کند تا بر اهداف اطلاعاتی تمرکز کند و در عین حال پس زمینه نامرتبط را با محدود سازی نقشه های توضیح به این نواحی استخراج شده، نادیده بگیرد. در این پژوهش، با مقایسه نقشه حرارتی^۵ با ناحیه هدف، تابع ضرر تقویت شده با توضیح، سازوکار توجه مدل را تنظیم می کند و باعث می شود مدل در طول آموزش به طور دقیق روی ناحیه هدف متمرکز شود و وابستگی به پس زمینه کاهش یابد. این امر قابلیت توضیح پذیری و تعمیم پذیری مدل را بهبود می بخشد. سازوکار آموزشی پیشنهادی، به طور مؤثر مدل را در جهت تمرکز بهتر بر نواحی حیاتی و در عین حال بهینه سازی دقت پیش بینی هدایت می کند و در نهایت توضیحات قابل اعتمادتر تولید می نماید [۴]. توابع ضرر به کار گرفته شده در این پژوهش به این صورت هستند:

- **تابع ضرر طبقه بندی:** برای این منظور از تابع Cross Entropy استفاده شده است.
- **تابع ضرر توضیح پذیری:** برای این منظور از تفاوت نقشه حرارتی تولید شده با ماسک تصویر با استفاده از Cross Entropy استفاده شده است.
- **تابع ضرر متعادل سازی پراکندگی (Sparsity) و داده افزایی:** برای متعادل سازی پراکندگی از تابع L_1 در نقشه حرارتی استفاده شده است. و برای جلوگیری از بیش برآزش از تابع ضرر (mixup) با راهبرد تولید نمونه های درون یابی شده^۶ بین جفت ورودی ها استفاده شده است.

با جلوگیری از بیش برآزش به الگوهای کاذب، FocusEx در آزمایش ها عملکرد بهتری نسبت به روش های پیشرفته ی موجود از نظر دقت پیش بینی و تعمیم پذیری در چندین مجموعه داده ی معیار نشان می دهد [۴]. اما از نظر کیفیت توضیحات ارائه شده و نقشه های حرارتی تولید شده، روش FocusEx در افزایش تمرکز شبکه عصبی با مشکل مواجه است و تصاویر نقشه حرارتی های تولید شده با این روش تمرکز مناسبی نسبت به تمام

^۴ Pretrained Models
^۵ heatmap
^۶ interpolated samples

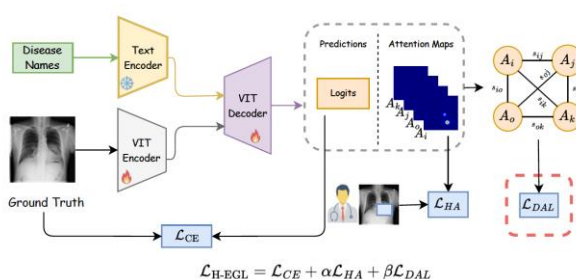
^۱ Task supervision loss
^۲ Explanation supervision loss
^۳ Explanation regularization loss

- ماژول نظارت‌شده‌ی هم‌ترازی انسان-هوش مصنوعی^۱، که تمرکز آن بر تطبیق نقشه‌های توجه با نشانه‌گذاری‌های متخصصان (مانند جعبه‌های محصورکننده)^۲ است [۵].

تابع ضرر کلی آموزش شامل سه بخش اصلی است:

- آنتروپی متقاطع ($\mathcal{L}_{CE}$) برای انجام طبقه‌بندی
- هم‌ترازی انسان-هوش مصنوعی ($\mathcal{L}_{HA}$) برای هم‌راستا کردن نقشه‌های توجه با نشانه‌های انسانی
- توجه متمایزکننده ($\mathcal{L}_{DAL}$) برای ایجاد جداسازی میان نقشه‌های توجه مربوط به کلاس‌های مختلف.

شکل ۱ مفهوم تمایز توجه^۳ را در یک سناریوی چهاربخشی نشان می‌دهد. این رویکرد برای هر تعداد کلاس قابل تعمیم است.



شکل ۱. چارچوب یادگیری هدایت‌شده با تبیین ترکیبی روش H-EGL [۵].

۳-۱. معیارهای ارزیابی توضیح‌پذیری

در رویکرد استفاده از راهنمایی‌های توضیح‌پذیری در آموزش مدل (EGL)، برای ارزیابی مدل‌های آموزش داده شده علاوه بر معیارهای پیش‌بینی مانند مساحت زیر منحنی (ACU) و دقت (Accuracy)، معیارهای زیر نیز مورد بررسی قرار می‌گیرند.

نسبت مساحت (AR)^۴

این معیار همپوشانی بین نواحی با فعال‌سازی بالا در نقشه‌های Grad-CAM و ماسک شیء را اندازه‌گیری می‌کند [۴] که به صورت رسمی به شکل زیر تعریف می‌شود:

attention distinctiveness^۳

Area Ratio^۴

سوژه اصلی نداشته است. عدم تمرکز مدل به تمام ویژگی‌های سوژه اصلی موجب کاهش دقت مدل در مجموعه داده‌هایی می‌شود که دارای فضای ویژگی مشابهی هستند. با افزایش تمرکز شبکه عصبی بر سوژه اصلی مدل توانایی درک بهتر جزئیات سوژه و عملکرد بهتری در انجام وظیفه اصلی طبقه‌بندی دارد.

در پژوهش [۵]، چارچوبی با نام یادگیری هدایت‌شده با تبیین ترکیبی (H-EGL) پیشنهاد شده است که محدودیت‌های خودنظارتی و هدایت‌شده توسط انسان را ترکیب می‌کند تا هم‌ترازی مکانیزم توجه را بهبود داده و قابلیت تعمیم مدل را افزایش دهد. بخش خودنظارتی H-EGL از توجه متمایزکننده‌ی کلاسی بدون تکیه بر پیش‌فرض‌های محدودکننده بهره می‌برد، که موجب افزایش پایداری و انعطاف‌پذیری می‌شود [۵]. این روش بر روی طبقه‌بندی تصاویر رادیوگرافی قفسه سینه با استفاده از Vision Transformer (ViT) ارزیابی شده است. نتایج نشان می‌دهد که H-EGL نسبت به دو روش پیشرفته‌ی موجود در حوزه‌ی یادگیری هدایت‌شده با تبیین (EGL) عملکرد بهتری دارد و دقت طبقه‌بندی و قابلیت تعمیم بالاتری ارائه می‌دهد. علاوه بر این، نقشه‌های توجه تولیدشده توسط این روش هم‌ترازی بهتری با تخصص انسانی نشان می‌دهند [۵].

H-EGL یک رویکرد ترکیبی از نوع یادگیری هدایت‌شده با تبیین (EGL) است که برای ترنسفورمرهای بینایی (ViT) طراحی شده است. این مدل ورودی‌های تصویری (مانند تصاویر رادیوگرافی قفسه سینه) و اطلاعات متنی (مانند برچسب‌های بیماری نظیر آتلکتازی یا کاردیومگالی) را از طریق یک معماری رمزگذار-رمزگشا مبتنی بر ViT با هم ترکیب می‌کند. در این ساختار، رمزگذار متنی ثابت (فریز شده) بوده و اجزای ViT قابل آموزش هستند. مدل در نهایت پیش‌بینی‌های کلاسی و نقشه‌های توجه را تولید می‌کند.

چارچوب H-EGL شامل دو مؤلفه‌ی کلیدی است:

- ماژول خودنظارتی یادگیری توجه متمایزکننده (DAL) که به تولید الگوهای توجه متمایز برای هر کلاس کمک می‌کند

Human-AI Alignment^۱

Bounding box^۲

ارائه روشی برای افزایش تمرکز شبکه عصبی عمیق بر سوژه اصلی مبتنی بر توضیح پذیری برای مدل های طبقه بندی

$$AR(G, M) = \frac{\sum_{i,j} (I(G_{ij}(x) > \tau) \cdot M(i, j))}{\sum_{i,j} I(G_{ij}(x) > \tau)}$$

که در آن τ آستانه ای فعال سازی (به طور پیش فرض ۰.۵) است و $M(i, j)$ برچسب حقیقت مطلق (ground truth) ماسک هدف می باشد.

کامل بودن توضیح (Comp)

کیفیت توضیح را با اندازه گیری اختلاف احتمال پیش بینی بین ورودی اصلی (x) و ورودی ای که بخش های پشتیبان آن حذف شده اند ($X_i \setminus g(f, \langle X_i, Y_i \rangle)$)، می سنجد [۶]، [۷]، [۸]، [۹]، [۱۰]:

$$\begin{aligned} \text{Comprehensiveness} \\ &= f(X_i)_{Y_i} \\ &- f(X_i \setminus g(f, \langle X_i, Y_i \rangle))_{Y_i} \end{aligned}$$

که در آن X_i تصویر ورودی، Y_i برچسب (کلاس) تصویر ورودی، $f(X_i)_{Y_i}$ پیش بینی مدل برای تصویر X_i و برچسب Y_i ، $g(f, \langle X_i, Y_i \rangle)$ نقشه حرارتی تولید شده برای تصویر X_i ، برچسب Y_i با استفاده از مدل $f(\cdot)$ و عملگر $\setminus$ به معنای حذف بخش های پشتیبان یا استدلال های توضیحی $g(f, \langle X_i, Y_i \rangle)$ از ورودی X_i است.

مقادیر بیشتر معیار Comp نشان می دهد که ویژگی های استخراج شده بهتر می توانند پیش بینی مدل را پشتیبانی کنند.

کفایت توضیح (Suff)

این معیار بررسی می کند که آیا ویژگی های استخراج شده به قدر کافی برای پیش بینی دقیق مدل کافی هستند یا خیر. این معیار به این صورت تعریف می شود [۷]، [۹]، [۱۱]، [۱۲]:

$$Suff = f_y(x) - f_y(g(f, \langle x, y \rangle))$$

مقادیر کمتر Suff نشان می دهد که ویژگی های استخراج شده بهتر می توانند پیش بینی مدل را پشتیبانی کنند.

۴-۱. روش تحقیق یا اصول و تئوری مقاله

فرآیند آموزش و ارزیابی مدل به صورت ساختار یافته در قالب یک چرخه یادگیری خودکار پیاده سازی شده است. در

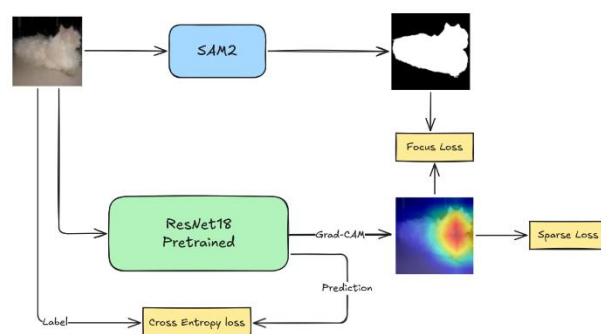
ابتدا، مدل ResNet^{18} با وزن های از پیش آموزش دیده روی مجموعه داده ImageNet بارگذاری می شود. سپس با استفاده از مدل SAM^{23} برای تصاویر موجود در مجموعه داده ماسک های دودویی تولید می شود. در فرآیند تولید ماسک های دودویی برای تعیین نقطه دستور (Prompt)، به این صورت عمل می شود، از یک شبکه ResNet^{50} از پیش آموزش داده شده برای تولید نقشه حرارتی Grad-CAM برای تصاویر استفاده می شود. سپس نقطه ای که نقشه حرارتی Grad-CAM تصویر در آن بیشینه است پیدا می شود و آن نقطه به عنوان دستور به SAM^2 ارسال می شود تا فرآیند بخش بندی دودویی نسبت به آن نقطه صورت گیرد. در مرحله بعد تابع ضرر FS Loss با اهداف زیر تعریف می شوند:

- طبقه بندی درست تصاویر
- افزایش تمرکز شبکه عصبی بر نواحی معنادار و سوژه اصلی تصویر
- کاهش نقاط حداکثری و پراکندگی در نقشه حرارتی

این تابع با بهره گیری از نقشه های حرارتی حاصل از Grad-CAM، تراکم پاسخ های مدل را در نواحی خاصی از تصویر تنظیم کرده و از توجه مدل به بخش های غیر مرتبط جلوگیری می کند. به منظور بهینه سازی پارامترهای مدل، از بهینه ساز Adam با نرخ یادگیری 10^{-4} و جمله متعادل ساز وزن L_2 برابر با (10^{-4}) استفاده شده است. نمای کلی شبکه طراحی شده در شکل ۲ نشان داده شده است.

فرآیند آموزش در طی ۱۵ دوره انجام می شود. پس از اتمام هر دوره آموزشی، مدل در حالت ارزیابی قرار گرفته و روی داده های اعتبارسنجی آزمایش می شود. در این مرحله، مدل پیش بینی های خود را بدون به روز رسانی وزن ها انجام داده و مقادیر زیان و دقت محاسبه می شوند تا عملکرد آن روی داده های مشاهده نشده بررسی گردد. بدین ترتیب، روند کاهش زیان و افزایش دقت در مجموعه های آموزش و اعتبارسنجی در طول دوره های مختلف قابل ردیابی بوده و معیارهایی مانند

Train/Test Accuracy و Train/Test Loss به عنوان شاخص‌های عملکردی مدل ذخیره می‌شوند.



$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \gamma_{sparse} \mathcal{L}_{sparse} + \beta_{focus} \mathcal{L}_{focus}$$

شکل ۲. معماری شبکه عصبی پیشنهادی به همراه تابع ضرر طراحی شده.

در مجموع، این ساختار امکان آموزش یک مدل بینایی مبتنی بر ResNet18 را با افزودن راهنمایی‌های توضیح‌پذیری از طریق مکانیزم‌های گرادین محور را فراهم می‌کند. ترکیب استفاده از تابع ضرر طبقه‌بندی و راهنمایی‌های نقشه حرارتی Grad-CAM موجب می‌شود که، مدل در کنار دقت بالای طبقه‌بندی، آموزش ببیند تا به قسمت‌های بامعنا تر تصویر و سوژه اصلی توجه کند و توجه به سایر قسمت‌های تصویر شامل پس‌زمینه کاهش یابد.

تابع ضرر FS Loss

برای بهره‌گیری از راهنمایی‌های توضیح‌پذیری در حین آموزش مدل، تابع ضرر ترکیبی که هدف آن، یادگیری ویژگی‌های مناسب افزایش دقت طبقه‌بندی، افزایش تمرکز بر سوژه اصلی و کاهش پراکندگی در نقشه‌های حرارتی است طراحی شده است. تابع ضرر FS Loss از بخش‌های زیر طراحی شده و شامل سه مؤلفه‌ی اصلی است:

- ضرر طبقه‌بندی: از تابع Cross Entropy استفاده شده است و علامت اختصاری آن $(\mathcal{L}_{CE})$ است.
- ضرر تمرکز: با علامت اختصاری $(\mathcal{L}_{Focus})$ است و زیان مرتبط با کیفیت تمرکز نقشه حرارتی بر سوژه اصلی تصویر را محاسبه می‌کند.

- ضرر پراکندگی: با علامت اختصاری $(\mathcal{L}_{sparsity})$ است و مدل را مجبور می‌کند تا پراکندگی نقشه حرارتی‌های تولید شده را تا حد ممکن کم کند و نقشه حرارتی تنگ از تصویر ورودی ایجاد کند.

ایده‌ی اصلی این تابع آن است که مدل علاوه بر یادگیری صحیح در سطح برجسب، در سطح ویژگی‌ها نیز رفتار قابل توضیح و متمرکز بر نواحی مرتبط با شیء هدف داشته باشد. این ساختار به مدل اجازه می‌دهد تا همزمان با حفظ عملکرد پیش‌بینی، ویژگی‌های درونی خود را از نظر بصری بهبود دهد و از تولید نقشه‌های غیرمتمرکز یا پر نویز جلوگیری کند.

زیان طبقه‌بندی^۱

مؤلفه‌ی اول تابع ضرر، همان زیان استاندارد Cross-Entropy است که برای یادگیری درست برجسب‌ها به کار می‌رود:

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^N \log p(y_i | x_i)$$

که $(p(y_i | x_i))$ خروجی شبکه برای کلاس صحیح است. این بخش مسئول اصلی یادگیری وظیفه‌ی طبقه‌بندی است.

زیان تمرکز^۲

برای اجبار مدل به یادگیری شباهت بین نقشه حرارتی تولیدشده و ماسک هدف، از معیار میانگین مربعات خطا (MSE) استفاده شده است:

$$\mathcal{L}_{focus} = \sum_{j=1}^W \sum_{i=1}^H \frac{1}{HW} (H_{pred}(i, j) - H_{mask}(i, j))^2$$

در این رابطه، H_{mask} ماسک یا نقشه هدف است و H_{pred} نقشه حرارتی تولیدی مدل است. این زیان، شبکه را وادار می‌کند تا در ناحیه‌های مشخص‌شده توسط ماسک تمرکز کند و ویژگی‌های محدود شده به ماسک هر تصویر را در نقشه حرارتی انتخاب کند.

زیان پراکندگی^۳

Sparsity Loss^۳

Classification Loss^۱

Focus Loss^۲

ارائه روشی برای افزایش تمرکز شبکه عصبی عمیق بر سوژه اصلی مبتنی بر توضیح پذیری برای مدل های طبقه بندی

برای جلوگیری از تولید نقشه های حرارتی با نواحی فعال گسترده و پرنویز، یک جمله پراکندگی بر پایه ی نرم (L_1) تعریف شده است [۴]:

$$\mathcal{L}_{sparse} = \sum_{j=1}^W \sum_{i=1}^H \frac{1}{HW} |H_{pred}(i, j)|$$

این مؤلفه باعث می شود مدل فقط در نواحی واقعاً مهم فعال شود و سایر نواحی را خاموش نگه دارد. در نتیجه، نقشه حرارتی نهایی، فشرده تر و از نظر بصری قابل تفسیرتر خواهد بود. در مجموع، تابع ضرر پیشنهادی را می توان به شکل زیر نوشت:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \gamma_{sparse} \mathcal{L}_{sparse} + \beta_{focus} \mathcal{L}_{focus}$$

این تابع ضرر سه هدف کلیدی را دنبال می کند:

- حفظ دقت مدل در سطح پیش بینی با استفاده از قرار دادن Cross-Entropy در تابع ضرر.
- افزایش هم راستایی بین نواحی تمرکز مدل و نواحی واقعی شیء با قرار دادن Focus Loss.
- بهبود فشرده گی و تفسیر پذیری نقشه های حرارتی با استفاده از Sparsity Loss.

به این ترتیب، مدل نه تنها عملکرد طبقه بندی مناسبی خواهد داشت، بلکه از نظر تبیین رفتار نیز قابل اعتمادتر و هم راستا با دانش انسانی خواهد بود.

فرآیند تولید ماسک دودویی

در چارچوب کار حاضر، از ماژول SAM^۲ برای تولید ماسک دودویی استفاده شده تا جداسازی دقیق پس زمینه و پیش زمینه روی تصاویر فراهم گردد. مدل SAM^۲ در این پژوهش به عنوان یک مدل بنیادین برای تقسیم بندی قابل پرسش^۱ به کار گرفته شده است. طبق توصیف رسمی SAM^۲، این مدل گامی به سوی حل مسئله تقسیم بندی تصویر و ویدئو با قابلیت پاسخ به نشانگرهای مختلف است. در این نسخه، تصاویر همانند یک ویدئوی تک فریمی در نظر گرفته شده اند و از یک معماری مبتنی بر ترنسفورمر با حافظه جریان پذیر برای پردازش در زمان واقعی

بهره گیری شده است. علاوه بر ساختار مدل، یک موتور داده سازی با تعامل در حلقه مدل^۲ توسعه داده شد که با کمک تعامل انسانی مجموعه داده های ویدیویی-تقسیمی را گردآوری و بهبود می بخشد. نتیجه این فرآیند، آموزش SAM^۲ بر روی داده های جمع آوری شده و تقویت عملکرد آن در حوزه های بصری متنوع بوده است [۲].

روند پردازش به صورت گام به گام به صورت زیر طراحی شده است:

- ابتدا برای هر تصویر یک نقشه حرارتی تولید شده که نشان دهنده اهمیت یا فعال سازی مقاطع مختلف تصویر برای معیار مورد نظر است.
- سپس مکان نقطه ای که مقدار نقشه گرمایی در آن حداکثر بود (نقطه ماکزیمم) با استفاده از مدل از پیش آموزش داده شده ResNet^{۵۰} با وزن های ImageNet استخراج شده.
- این مختصات نقطه به عنوان ورودی دستور^۳ در اختیار مدل SAM^۲ قرار داده شده تا بر اساس آن ناحیه مربوط به شیء مورد نظر از زمینه تفکیک شود.
- پس از تولید ماسک دودویی توسط مدل SAM^۲ تمام تصاویر تولید شده توسط ناظر انسانی بررسی می شوند و در صورت وجود خطا در فرایند تولید ماسک از مجموعه داده نهایی کنار گذاشته می شوند.

با استفاده از نقطه تعیین شده، SAM^۲ یک ماسک باینری برای تصویر تولید نموده که به وضوح نواحی پیش زمینه را از پس زمینه جدا می کند. تولید این ماسک منجر به تعریف توابع ضرری برای آموزش مدل می شود که تنها بر روی بخش های مرتبط تصویر فرآیند یا محاسباتی را انجام می دهند و اثر نویز و اطلاعات نامربوط را کاهش می دهند. همچنین استفاده از نقطه ماکزیمم نقشه حرارتی به عنوان نقطه راهنما، فرآیند تولید ماسک را هدفمند و کم هزینه می کند، زیرا فرآیند تولید ماسک از تصاویر را از یک فرآیند کاملاً دستی تبدیل به یک فرآیند کاملاً خودکار می کند که به جای ارائه چندین نشانگر یا جستجوی گسترده

^۳ prompt

^۱ promptable segmentation

^۲ model-in-the-loop data engine

- برای هر کلاس، ۵۰۰ تصویر برای آموزش، ۵۰ تصویر برای اعتبارسنجی و ۵۰ تصویر برای آزمون در دسترس است.

به منظور کاهش هزینه‌ی محاسباتی و بررسی دقیق‌تر نتایج آزمایشات، تنها ۱۰ کلاس از میان ۲۰۰ کلاس انتخاب شده است. در انتخاب کلاس‌ها سعی شده است موضوع کلاس‌ها متمایز از یکدیگر باشند تا بتوان بررسی دقیق‌تری بر روی نتایج مدل‌های آموزش داده شده با این مجموعه داده انجام داد. تصاویر انتخاب شده، طی فرآیند پیش‌پردازش به اندازه‌ی ۲۲۴×۲۲۴ پیکسل تغییر اندازه داده شده‌اند.

کلاس‌های انتخاب شده در جدول ۱ آمده‌اند.

جدول ۱. برجسب‌های انتخاب شده از مجموعه داده

TinyImageNet [۱۱].

نام فارسی کلاس	عنوان انگلیسی
صفحه کلید کامپیوتر	computer_keyboard
میز	desk
ماهیتابه	frying_pan
باجه تلفن عمومی	pay-phone
عینک آفتابی	sunglasses
پرتقال	orange
پیتزا	pizza
اسپرسو	espresso
گره ایرانی	Persian_cat
زنبر عسل	bee

همچنین برای مقایسه روش پیشنهادی با سایر پژوهش‌های پیشین، روش پیشنهادی بر روی مجموعه داده VLCS نیز پیاده‌سازی شده است و سپس نتایج به دست آمده با نتایج گزارش شده در روش FocusEx مقایسه شده‌اند.

۲-۲. آزمایشات انجام شده

در این پژوهش برای آزمایش و ارزیابی روش پیشنهادی، شبکه ResNet۱۸ (همراه با وزن‌های از پیش آموزش داده شده روی مجموعه داده ImageNet) به عنوان مدل پایه انتخاب شده و لایه تماماً متصل ۱۰۰۰ کلاس آن با یک لایه تماماً متصل ۱۰ کلاس جایگزین شده است. با توجه به نوع تابع ضرر تعریف شده و نحوه استفاده از آن در حین آموزش برای عملکرد مناسب

توسط عامل انسانی، تنها یک نقطه کلیدی که اتفاقاً مدل توجه بیشتری به آن داشته است را انتخاب کرده تا مدل بتواند مرزهای شیء را بازشناسی کند. نمونه مرزبندی تصاویر در شکل ۳ نشان داده شده است.



شکل ۳. نمونه مرزبندی تصاویر مجموعه داده با استفاده از مدل

SAM^۲

ادغام نقشه حرارتی و SAM^۲ مزایای عملی متعددی دارد. اولاً، تمرکز پردازش بر بخش‌های معنادار تصویر افزایش یافته و در نتیجه تحلیل‌های آماری و بصری قابل اطمینان‌تر می‌شوند. ثانیاً، روش پیشنهادی به صورت عمومی‌تری قابل تعمیم است، زیرا به کارگیری یک نقطه ماکزیمم از نقشه حرارتی می‌تواند برای انواع مختلف تصاویر و سناریوها به عنوان یک نشانه ساده اما مؤثر عمل نماید.

در نهایت، بهره‌گیری از SAM^۲ که با داده‌های بزرگ و متنوع آموزش دیده است، سبب می‌شود تا تفکیک مرزهای پیش‌زمینه با دقت و پایداری بیشتری انجام شود و نیاز به تنظیمات دستی یا طراحی ماسک‌های تخصصی برای هر تصویر تا حد زیادی کاهش یابد.

۲. یافته‌ها

۲-۱. مجموعه داده انتخاب شده

در این پژوهش از مجموعه داده Tiny ImageNet برای انجام آزمایش‌ها استفاده شده است. این مجموعه داده نسخه‌ای فشرده و کوچک شده از ImageNet است و به صورت گسترده در ارزیابی مدل‌های یادگیری عمیق مورد استفاده قرار می‌گیرد [۱۴]. مشخصات این مجموعه داده به صورت زیر است:

- شامل ۱۰۰,۰۰۰ تصویر در ۲۰۰ کلاس مختلف (هر کلاس دارای ۵۰۰ نمونه تصویری) است.
- تمامی تصاویر به اندازه‌ی ۶۴×۶۴ پیکسل تغییر مقیاس یافته‌اند.

ارائه روشی برای افزایش تمرکز شبکه عصبی عمیق بر سوژه اصلی مبتنی بر توضیح پذیری برای مدل های طبقه بندی

و قابل توجه، بهتر است از زیان های $\mathcal{L}_{focus}$ و $\mathcal{L}_{sparse}$ در زمانی استفاده شود که مدل در حال آموزش ویژگی های اصلی مسئله بوده و فضای عمومی مسئله را قبلاً آموزش دیده باشد. بدین ترتیب محل استفاده صحیح توابع ضرر پیشنهاد شده در این حالت ها می باشد:

- استفاده از $\mathcal{L}_{focus}$ و $\mathcal{L}_{sparse}$ پس از چند دوره آموزش با استفاده از تابع ضرر طبقه بندی.

- استفاده از $\mathcal{L}_{focus}$ و $\mathcal{L}_{sparse}$ برای بهینه سازی مدل از پیش آموزش داده شده بر روی مسئله جدید.

آزمایشات صورت گرفته در این پژوهش طبق حالت دوم صورت گرفته است. شبکه ResNet^{18} با فرایندهای β_{focus} و γ_{sparse} متفاوت آموزش داده شده و نتایج آن در دو فاز طبقه بندی و توضیح پذیری مقایسه شده اند. همچنین برای مقایسه بهتر تاثیر روش پیشنهادی، به صورت جداگانه شبکه ResNet^{18} تنها با تابع ضرر $\mathcal{L}_{CE}$ نیز آموزش داده شده است تا مقایسه صحیحی بین مدل های آموزش داده شده با راهنمایی توضیح پذیری و مدل های آموزش داده شده با روش های معمول صورت گیرد.

۳. نتایج

۳-۱. مقایسه مدل های توضیح پذیر و غیر توضیح پذیر

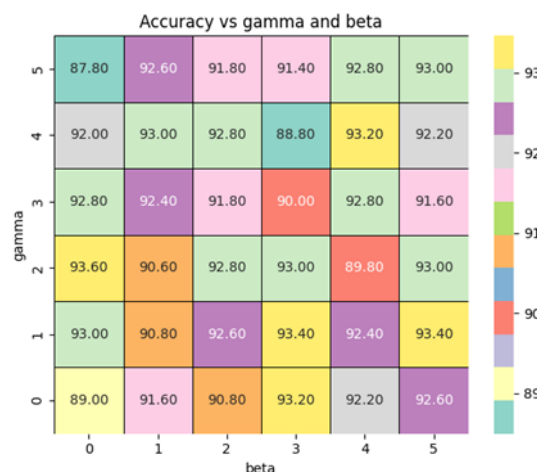
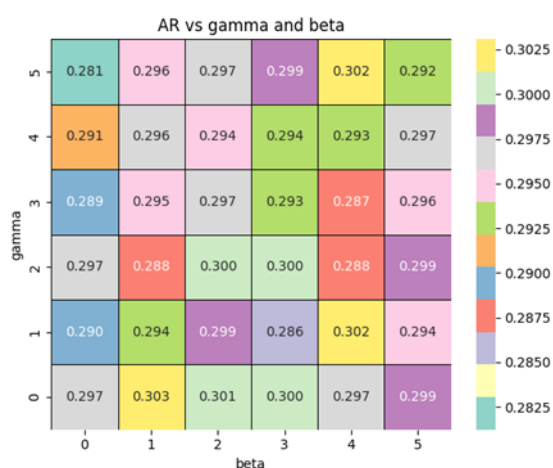
در این بخش ابتدا مدل های آموزش داده شده همراه با فرایندهای آن بررسی می شوند و سپس به بررسی نتایج آن از جمله نمودارهای فرآیند آموزش مدل و مقایسه دقت هر مدل پرداخته می شود.

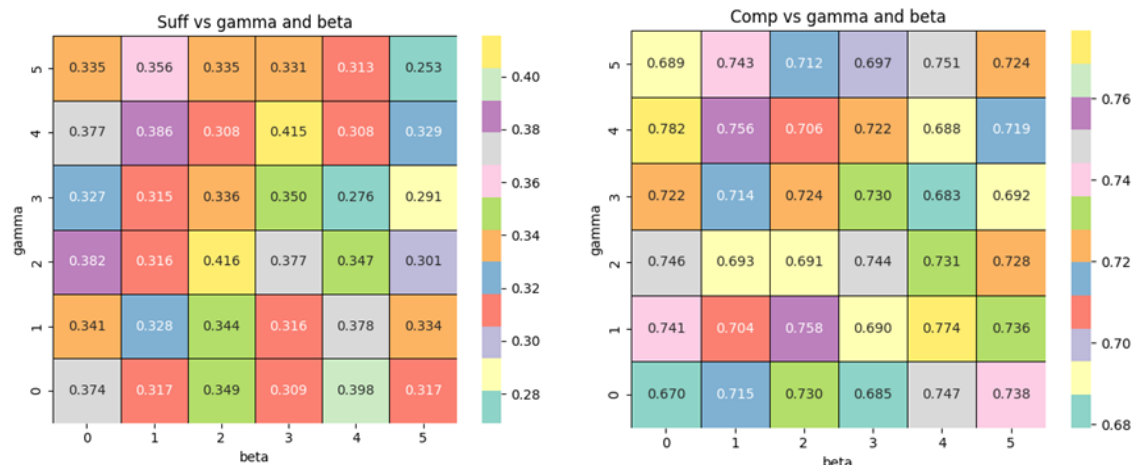
در این پژوهش شبکه عصبی ResNet^{18} با فرایندهای ثابت ذکر شده در جدول ۲ آموزش داده شده است:

جدول ۲. فرایندهای تنظیم شده در آزمایشات صورت گرفته.

نوع فرایندها	مقدار
Weight decay	۰.۰۰۰۱
Learning rate	۰.۰۰۰۱
Batch size	۱۲۸
Grad-CAM target layer	آخرین لایه کانولوشن در مدل ResNet^{18}

اما سایر فرایندها مانند β_{focus} و γ_{sparse} با مقادیر متفاوت آزمایش شده اند. در ادامه نتایج آموزش مدل های در شکل ۴ قرار داده شده است.





شکل ۴. نتایج معیارهای ارزیابی Suff, AR, Accuracy و Comp روش ارائه شده.

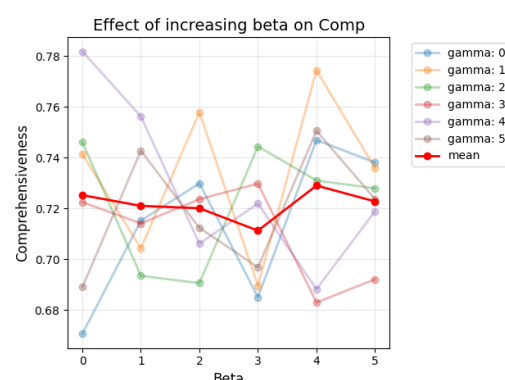
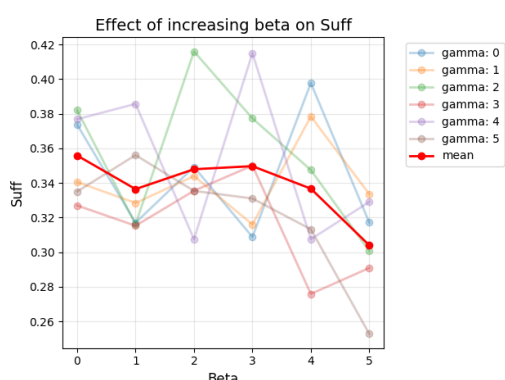
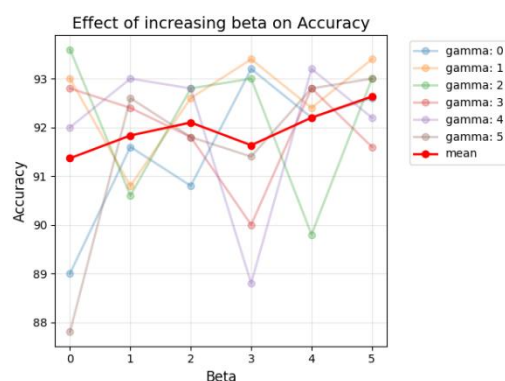
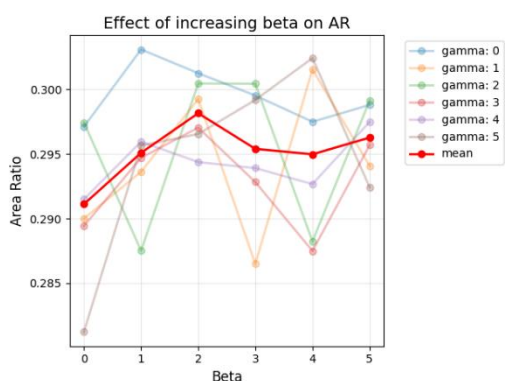
دقت مدل‌های آموزش داده شده می‌توان افزایش دقت پیش‌بینی برای مدل را برداشت کرد.

یکی از موارد کلیدی که به درک بهتر میزان تاثیر تابع ضرر طراحی شده کمک می‌کند، مقایسه نتایج معیارهای ارزیابی توضیح‌پذیری است. در شکل ۴ نتایج ارزیابی مدل‌های آموزش داده شده با معیارهای مطرح شده «در بخش معیارهای ارزیابی توضیح‌پذیری» نشان داده شده است. مدل‌های آموزش داده شده با روش ارائه شده در معیار AR نسبت به مدل پایه در بیشترین حالت حدود ۱٪ بهبود داشته‌اند. این میزان بهبود برای معیار Comp حدود ۷٪ و برای معیار Suff حدود ۱۰٪ است. با توجه به موارد بررسی شده، مدل‌های آموزش داده شده با روش پیشنهادی همزمان با برتری در دقت پیش‌بینی نسبت به مدل پایه، در ارزیابی‌های کمی توضیح‌پذیری نیز برتری قابل توجهی دارند.

نتایج ارزیابی مدل‌های آموزش داده با معیارهای مطرح شده نشان داده شده است. مدل‌های آموزش داده شده با روش ارائه شده در معیار نسبت مساحت، نسبت به مدل پایه در بیشترین حالت حدود ۱٪ بهبود داشته‌اند. این میزان بهبود برای معیار کامل بودن توضیحات، حدود ۷٪ و برای معیار کفایت توضیحات، حدود ۱۰٪ است.

در شکل ۴ تمامی مدل‌ها در ۱۵ دوره آموزش داده شده و به منظور مقایسه بهتر سایر فرآیندها ثابت شده‌اند. در نمودارها، محور x نشان دهنده مقادیر β_{focus} و محور y مقادیر γ_{sparse} آزمایش شده هستند. هر خانه از نمودارها بیانگر مقدار نهایی معیار ارزیابی مدل آموزش داده شده است. به عنوان مثال خانه پایین گوشه چپ، نشانگر مدل آموزش داده شده با فرآیندهای $\gamma_{sparse} = 0.0$ و $\beta_{focus} = 0.0$ که یعنی تنها تابع ضرر موجود در این آزمایش خاص تابع ضرر $\mathcal{L}_{CE}$ بوده است. اگر آزمایش با فرآیندهای $\gamma_{sparse} = 0.0$ و $\beta_{focus} = 0.0$ را به عنوان مدل پایه در نظر بگیریم، با توجه به نتایج گزارش شده در تمامی معیارهای ارزیابی استفاده از تابع ضرر مبتنی بر توضیح‌پذیری از مدل پایه نتایج بهتری را کسب کرده است.

با بررسی دقیق‌تر می‌توان دریافت که، راهنمایی‌های توضیح‌پذیری به طور موثری در بهبود نتایج شبکه عصبی نقش داشته‌اند. در آزمایشاتی که صورت گرفته دقت نهایی مدل پایه بر روی مجموعه داده تست حدود ۸۹٪ است. اما این محدوده برای مدل‌هایی که از راهنمایی‌های توضیح‌پذیری استفاده می‌کنند در حدود $92\% \pm 1\%$ است. که این مقدار در بیشترین حالت خود به ۹۳٫۸٪ نیز می‌رسد. در همین قسمت با مقایسه نتیجه



شکل ۵. تاثیر تابع ضرر تمرکز و پارامترهای β_{focus} و γ_{sparse} بر دقت مدل و معیارهای کمی توضیح پذیری.

بر اساس روند نمایان شده توسط خط میانگین، می توان استنباط کرد که افزایش وزن مؤلفه تمرکز در تابع زیان (از طریق افزایش مقدار β_{focus}) به صورت کلی منجر به بهبود دقت پیش بینی شده است. این مشاهده نشان می دهد که هدایت شبکه به سمت تمرکز مکانی بیشتر بر نواحی معنادار تصویر، نه تنها باعث بهبود ساختار توضیحی نقشه های حرارتی می شود، بلکه از منظر عملکرد طبقه بندی نیز سودمند است. از سوی دیگر، پارامتر γ_{sparse} که نقش تنظیم شدت قید پراکندگی را بر عهده دارد، ابزاری برای پالایش دقیق تر ساختار نقشه های حرارتی محسوب می شود. تنظیم مناسب این پارامتر می تواند از ایجاد فعال سازی های گسترده و غیرهدفمند جلوگیری کرده و تمرکز ویژگی ها را به نواحی کلیدی محدود سازد. نتایج نشان می دهد که انتخاب مقدار بهینه برای γ_{sparse} علاوه بر ارتقای معیارهای

نتایج گزارش شده به صورت نمودار نشان داده شده اند. در این نمودارها، رفتار مدل در مواجهه با تغییرات شدت هر یک از مؤلفه های زیان به صورت نظام مند بررسی شده است تا حساسیت عملکرد مدل نسبت به این ابرپارامترها مشخص گردد. به طور مشخص، در نمودار مربوط به تأثیر بر دقت پیش بینی، پنج منحنی مجزا برای مقادیر مختلف γ_{sparse} ترسیم شده اند که به صورت خطوط کم رنگ نمایش داده شده اند. هر نقطه روی این منحنی ها بیانگر دقت نهایی مدل پس از آموزش با یک ترکیب مشخص از β_{focus} و γ_{sparse} است. این نمایش چندمنحنی امکان تحلیل هم زمان اثر متقابل دو پارامتر را فراهم می سازد. علاوه بر این، یک خط قرمز رنگ به عنوان روند میانگین تغییرات بر روی نمودار افزوده شده است تا تصویر کلان تری از رفتار سیستم ارائه دهد و وابستگی عمومی عملکرد مدل به تغییرات β_{focus} را آشکار سازد.

توضیح‌پذیری (مانند تمرکز و تمایزپذیری نقشه‌ها)، می‌تواند موجب بهبود یا تثبیت دقت پیش‌بینی نیز شود.

در مجموع، تحلیل نمودارهای شکل ۵ نشان می‌دهد که ترکیب هم‌افزای دو مؤلفه تمرکز و پراکندگی در قالب ضرایب γ_{sparse}

۲-۳. تحلیل نقشه حرارتی‌های مدل آموزش داده شده
پس از بررسی و ارزیابی نتایج طبقه‌بندی مدل پایه و مدل آموزش داده شده با روش پیشنهادی، برای بررسی کیفی نتایج توضیح‌پذیری نقشه حرارتی‌های تولید شده توسط هر مدل نیز بررسی و ارزیابی می‌شوند. در صورتی که نقشه حرارتی‌های تولید شده از روش پیشنهادی بهتر و معنادارتر از روش پایه باشند می‌توان با اطمینان بیشتری به نتایج مدل اعتماد پیدا کرد. در جدول ۳ نقشه حرارتی‌های تولید شده توسط مدل پایه و مدل‌های برتر آموزش داده شده با روش پیشنهادی مقایسه شده‌اند. برای نمایش تاثیر فرآیند بازآموزش، نقشه حرارتی تولید شده با مدل پایه نیز قرار داده شده است تا تاثیر کلی بازآموزش بر روی شبکه نیز نمایش داده شود و بتوان درک بهتری از تاثیر روش ارائه شده داشت.

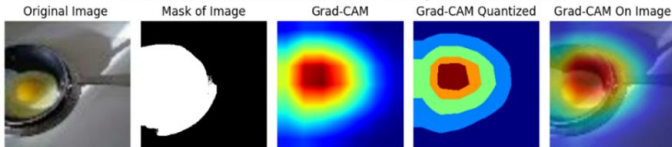
و β_{focus} امکان تنظیم توازن میان دقت پیش‌بینی و کیفیت توضیحات را فراهم می‌کند. این ویژگی، چارچوب پیشنهادی را به یک سازوکار قابل کنترل برای طراحی مدل‌هایی با کارایی بالا و توضیح‌پذیری تقویت‌شده تبدیل می‌نماید.

برای درک بهتر نقشه‌های حرارتی در جدول ۳، این نقشه در سه تصویر به نمایش درآمده است که توضیحات هر یک به این صورت است:

- تصویر با عنوان Grad-CAM، همان نقشه حرارتی تولید شده با روش Grad-CAM است.
- تصویر با عنوان Grad-CAM Quantized نقشه حرارتی Grad-CAM است با این تفاوت که مقادیر در چند بازه به صورت زیر کوانتیده شده‌اند:
 - مقدار پیکسل در بازه [۰, ۰.۲۵] مقدار ۰ دارد
 - مقدار پیکسل در بازه [۰.۲۵, ۰.۵] مقدار ۰.۲۵ دارد
 - مقدار پیکسل در بازه [۰.۵, ۰.۷۵] مقدار ۰.۵ دارد
 - مقدار پیکسل در بازه [۰.۷۵, ۰.۹] مقدار ۰.۷۵ دارد
 - مقدار پیکسل در بازه [۰.۹, ۱.۰] مقدار ۱.۰ دارد
- تصویر با عنوان Grad-CAM On Image نقشه حرارتی تولید شده بر روی تصویر ورودی به مدل است.

جدول ۳. مقایسه نقشه حرارتی‌های تولید شده از مدل‌های برتر با مدل پایه و مدل از پیش آموزش داده شده.

تابع ضرر	β_{focus}	γ_{sparse}	Test Accuracy	نقشه حرارتی تولید شده پس از آموزش مدل
$\mathcal{L}_{CE}$	۰,۰	۰,۰	۸۹,۰٪	For Model: Trained model: ResNet18 With CE Loss Original Image Mask of Image Grad-CAM Grad-CAM Quantized Grad-CAM On Image
$\mathcal{L}_{CE} + \gamma_{sparse}\mathcal{L}_{sparse} + \beta_{focus}\mathcal{L}_{focus}$	۵,۰	۰,۰	۹۲,۸٪	For Model: Trained model: ResNet18 With CE + (5.0)Focus Loss Original Image Mask of Image Grad-CAM Grad-CAM Quantized Grad-CAM On Image
$\mathcal{L}_{CE}$	۰,۰	۰,۰	۸۹,۰٪	For Model: Trained model: ResNet18 With CE Loss Original Image Mask of Image Grad-CAM Grad-CAM Quantized Grad-CAM On Image

$\mathcal{L}_{CE} + \gamma_{sparse}\mathcal{L}_{sparse} + \beta_{focus}\mathcal{L}_{focus}$	۱,۰	۳,۰	۹۳,۸٪	For Model: Trained model: ResNet18 With CE + (3.0)Sparse + (1.0)Focus Loss 
---	-----	-----	-------	--

موجب تمرکز بیشتر بر روی سوژه اصلی تصویر شده و مدل را مجبور می کند تا ویژگی هایی را انتخاب کند که در محدوده مورد نظر از تصویر هستند.

- همچنین روش پیشنهاد شده بر روی تصاویری که نقشه حرارتی مناسبی با مدل پایه ندارند نیز عملکرد بهتری دارد، به این صورت که، به دلیل وجود ماسک در روش پیشنهادی موجب تمرکز بیشتر مدل آموزش داده شده به نقاط مهم تصویر می شود و باعث تولید نقشه حرارتی های بهتری از نظر تمرکز بیشتر بر سوژه اصلی و پراکندگی کمتر نقاط پر اهمیت می شود. تفاوت این نتایج در نقشه حرارتی های شکل ۶ نشان داده شده است.

با توجه به آزمایشات انجام شده، روش پیشنهادی عملکرد مطلوبی در افزایش توضیح پذیری مدل ResNet۱۸ داشته است. این عملکرد در جنبه های افزایش دقت پیش بینی مدل، بهبود نقشه های حرارتی و تمرکز بیشتر مدل بر روی سوژه اصلی مشهود است.

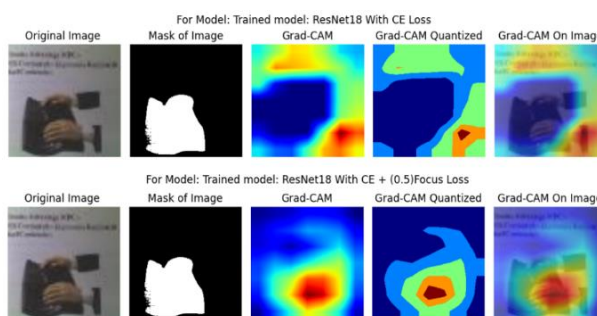
۳-۳. مقایسه با پژوهش های پیشین

به منظور ارزیابی دقیق تر عملکرد روش پیشنهادی و بررسی میزان اثربخشی آن در مقایسه با سایر پژوهش های مرتبط، آزمایش های تکمیلی بر روی چند مجموعه داده ای استاندارد مورد استفاده در تحقیقات پیشین انجام شد. پژوهش هایی که برای مقایسه انتخاب شدند، عمدتاً از مدل های متداول بینایی کامپیوتر مانند ResNet، VGG یا DenseNet همراه با روش های توضیح پذیری پس پردازشی نظیر Grad-CAM و LIME استفاده کرده بودند. در این پژوهش، همان مجموعه داده هایی که در آن مطالعات به کار رفته بودند، شامل VLCS مورد استفاده قرار گرفتند و مدل پیشنهادی با توابع زیان ($\mathcal{L}_{focus}$) و ($\mathcal{L}_{sparse}$) بر روی آنها بازآموزی و ارزیابی شد.

طبق نقشه های حرارتی جدول ۳، مدل های توضیح پذیر آموزش داده شده نسبت به مدل پایه نقشه های حرارتی با پوشش بیشتری از سوژه اصلی تولید کرده اند و هدف اصلی روش ارائه شده که افزایش تمرکز مدل بر روی سوژه اصلی و کاهش تاثیر سایر ویژگی های کم اهمیت ورودی را نشان می دهند. برای تحلیل نقشه های حرارتی می توان موارد زیر را به عنوان سنجه های ارزیابی مدنظر قرار داد:

- تمرکز نقشه حرارتی تولید شده بر سوژه اصلی و میزان پوشش سوژه اصلی با سطح های تاثیر بالاتر (مطلوب است بخش بیشتری از سوژه اصلی تصویر با نقاط دارای مقدار بیشتر در نقشه حرارتی تکمیل شوند)
- میزان پوشش پس زمینه تصویر با سطح های تاثیر پایین تر (به این صورت که مطلوب است بخش بیشتری از پس زمینه تصویر با نقاط دارای مقدار کمتر در نقشه حرارتی تکمیل شوند)

برای بررسی دقیق تر نقشه های حرارتی تولید شده دو نمونه دیگر از نقشه حرارتی های تولید شده توسط مدل پایه و روش پیشنهادی در شکل ۶ نشان داده شده است.



شکل ۶. تفاوت نقشه حرارتی تولید شده توسط مدل پایه و روش پیشنهادی.

طبق شکل ۶ و نتایج بررسی شده در جدول ۳، می توان تاثیر آموزش مدل با رویکرد پیشنهادی را به صورت زیر تحلیل کرد:

- روش پیشنهادی بر روی نمونه تصاویری که نقشه حرارتی نسبتاً مناسبی با مدل پایه دارند، به این صورت است که،

۴۹,۶۸	۵,۹۱	۵۰,۷۵	۹۷,۸۸	GroupDRO
۳۷,۹۵	۷,۵۹	۵۸,۳۶	۹۹,۰۵	Mixup
۴۷,۳۶	۰,۷۳	۹۸,۹۹	۹۹,۳۵	SGT
۲۷,۳۳	۱۰,۰۹	۹۹,۴۱	۹۹,۳۱	SGDrop
۲۹,۱۹	۳,۲۵	۹۸,۵۵	۹۹,۲۴	DRE
۲۸,۲۲	۲۸,۵۳	۹۹,۸۳	۹۹,۹۸	روش FocusEx
۲,۱۷	۶۰,۰۴	۲۸,۷۴	۹۷,۱	روش پیشنهادی

به‌ویژه در معیارهای کیفی مربوط به تمرکز مدل بر نواحی معنادار تصویر و کاهش پراکندگی نقشه‌های حرارتی.

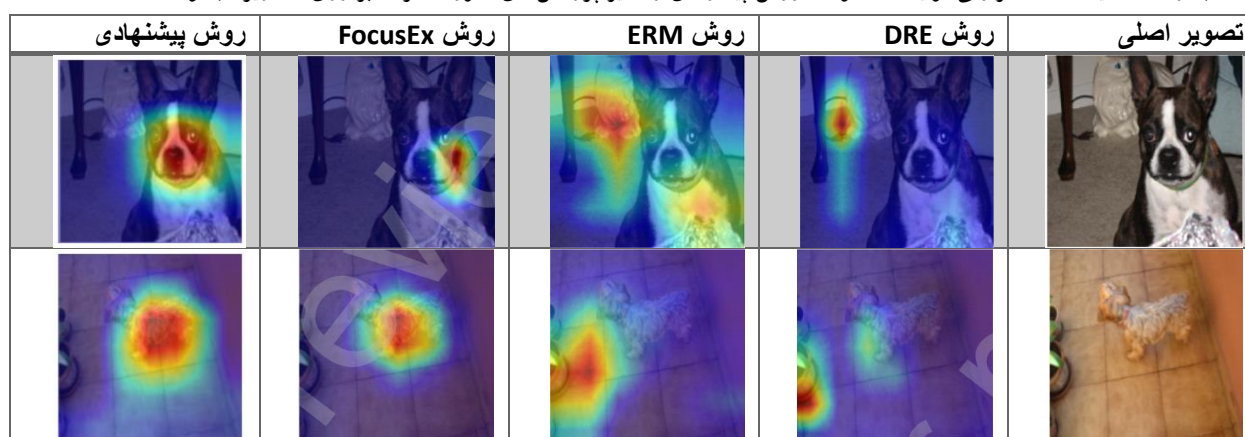
علاوه بر نتایج کمی، در معیارهای کیفی مربوط به تمرکز مدل بر نواحی معنادار تصویر و کاهش پراکندگی نقشه‌های حرارتی نیز روش پیشنهادی بهبود بسیار خوبی داشته است. تحلیل کیفی نقشه‌های حرارتی تولیدشده در جدول ۵ نشان می‌دهد که مدل پیشنهادی قادر است نواحی مرتبط‌تری از تصویر را شناسایی و بر بخش بیشتری از سوژه اصلی تمرکز کند. در مقایسه با روش‌های پیشین، نقشه‌های حرارتی حاصل از روش پیشنهادی تمرکز بالاتری و پوشش بیشتری از سوژه اصلی را داشتند. این موضوع بیانگر آن است که استفاده از روش پیشنهاد شده موجب هدایت مؤثرتر فرآیند یادگیری شبکه به سمت ویژگی‌های کلیدی و مؤثر تصویر شده است.

نتایج کمی به‌دست‌آمده از مقایسه با پژوهش‌های موجود در جدول ۴ آورده شده است. همان‌طور که از این جدول مشخص است، روش پیشنهادی در تعدادی از معیارهای ارزیابی از جمله **Comp** و **Suff** نسبت به روش‌های موجود عملکرد بهتری داشته است. به عنوان مثال، بر روی مجموعه داده‌ی **VLCS**، معیار **Comp** که در واقع بیانگر میزان تاثیر ویژگی‌های انتخاب شده بر پیش‌بینی مدل است، در مدل آموزش داده شده با روش پیشنهادی ۶۰.۰۴ گزارش شد که نسبت به بهترین روش پیشین با مقدار ۲۸.۵۳، بهبودی معادل ۳۲ واحد را نشان می‌دهد. با توجه به مفهوم معیار **Comp** این بهبود در معیار ارزیابی **Comp** بیانگر تاثیر افزایش تمرکز و استفاده از رویکرد اصلی روش پیشنهادی بوده است که در نتایج کمی نیز به طور واضحی قابل مشاهده و ارزیابی است. همچنین این بهبود در شاخص **Suff** نیز قابل مشاهده است.

جدول ۴. مقایسه نتایج روش پیشنهادی با سایر پژوهش‌های انجام شده.

Caltech ۱۰۱				شبکه آموزش داده شده
Suff	Comp	AR	Accuracy	
۳۸,۶۹	۰,۴۸	۹۸,۳۴	۹۸,۶۸	ERM
۳۰,۳۱	۲,۶۵	۵۶,۱۲	۹۵,۳۵	IRM

جدول ۵. مقایسه نقشه حرارتی تولید شده توسط روش پیشنهادی و سایر پژوهش‌های صورت گرفته بر روی تصاویر مجموعه داده VLCS.



توضیح‌پذیر با طراحی تابع زیان هدف‌دار می‌تواند گامی مؤثر در جهت توسعه مدل‌هایی باشد که نه تنها دقیق‌تر، بلکه قابل اعتمادتر و قابل تفسیرتر هستند.

۴. بحث و نتیجه گیری

در مجموع، مقایسه با پژوهش‌های پیشین نشان می‌دهد که روش پیشنهادی علاوه بر بهبود معیارهای کمی توضیح‌پذیری، از نظر جنبه‌های کیفی نیز خروجی‌های معنادارتری تولید کرده است. این نتایج بیانگر آن است که ترکیب رویکردهای یادگیری

ارائه روشی برای افزایش تمرکز شبکه عصبی عمیق بر سوژه اصلی مبتنی بر توضیح پذیری برای مدل های طبقه بندی

در این پژوهش، یک رویکرد نوین در زمینه یادگیری مبتنی بر توضیح پذیری (EGL)، برای بهبود توضیح پذیری شبکه های عصبی عمیق از طریق طراحی توابع زیان ($\mathcal{L}_{focus}$) و ($\mathcal{L}_{sparse}$) پیشنهاد و ارزیابی شد. روش پیشنهادی بر روی شبکه ی ResNet^{۱۸} با وزن های از پیش آموزش دیده بر روی مجموعه داده ImageNet پیاده سازی شده است. اهداف اصلی این روش شامل موارد زیر است:

- هدایت مدل به سمت تمرکز بر نواحی معنادار تصویر و افزایش تمرکز مدل بر روی سوژه اصلی
- کاهش وابستگی مدل به ویژگی های غیر ضروری یا پراکنده
- آزمایش ها برای ارزیابی عملکرد در دو فاز انجام گرفتند تا تأثیر دقیق راهنمایی های توضیح پذیری بر کیفیت یادگیری مدل به صورت جامع بررسی شود. ارزیابی عملکرد طبقه بندی و تحلیل توضیح پذیری مدل.
- نتایج به دست آمده نشان دادند که استفاده از ترکیب زیان های ($\mathcal{L}_{focus}$) و ($\mathcal{L}_{sparse}$) در فاز بازآموزی مدل از پیش آموزش دیده، موجب بهبود قابل توجهی در عملکرد کلی شبکه می شود. این افزایش دقت بیانگر آن است که افزودن راهنمایی های توضیح پذیری با روش ارائه پیشنهادی نه تنها به درک بهتر ویژگی های مهم توسط مدل کمک کرده اند، بلکه منجر به بهبود قابلیت تعمیم و عملکرد کلی شبکه نیز شده اند.

در بخش تحلیل نقشه های حرارتی حاصل از مدل های مختلف، تفاوت معناداری میان مدل پایه و مدل های آموزش داده شده با روش پیشنهادی مشاهده شد. نقشه های حرارتی تولید شده توسط مدل های توضیح پذیر، تمرکز بسیار بیشتری بر سوژه اصلی تصویر داشتند و نواحی پس زمینه با شدت پایین تری فعال شدند. این موضوع نشان می دهد که مدل توانسته است نواحی غیر مؤثر تصویر را بهتر نادیده گرفته و تصمیم گیری خود را بر اساس بخش های مرتبط تری از داده انجام دهد. علاوه بر این، در نمونه هایی که مدل پایه نتوانسته بود به درستی بر سوژه تمرکز کند، روش پیشنهادی توانست با هدایت مدل از طریق ماسک های تولیدی، تمرکز مکانی دقیق تر و منسجم تری ایجاد نماید.

همچنین، مقایسه با پژوهش های پیشین نشان می دهد که روش پیشنهادی نه تنها موجب بهبود معیارهای کمی مرتبط با توضیح پذیری شده است، بلکه از نظر جنبه های کیفی نیز، نقشه های حرارتی منسجم تر و معنادارتری تولید کرده است. این نتایج تأیید می کنند که ترکیب رویکردهای یادگیری توضیح پذیر با طراحی هدف مند توابع زیان می تواند به توسعه ی مدل هایی منجر شود که علاوه بر دقت بیشتر، قابلیت تفسیر و اعتماد پذیری بالاتری نیز دارند.

۴-۱. نتیجه گیری

به طور کلی، یافته های این پژوهش تأیید می کنند که استفاده از توابع زیان مبتنی بر توضیح پذیری در قالب زیان های ($\mathcal{L}_{focus}$) و ($\mathcal{L}_{sparse}$) می تواند به عنوان ابزاری مؤثر برای بهبود همزمان دقت و قابلیت تبیین در مدل های بینایی عمیق به کار گرفته شود. روش پیشنهادی عملکرد قابل قبولی در افزایش تمرکز نقشه های حرارتی تولید شده داشته و نه تنها منجر به تولید نقشه های حرارتی معنادارتر شده است، بلکه رفتار تصمیم گیری منسجم تر و قابل اعتمادتر در مدل را نیز تقویت کرده است. در نتیجه، می توان گفت رویکرد پیشنهادی گامی مؤثر در جهت ترکیب آموزش هدف محور با تبیین پذیری مدل ها بوده و می تواند به عنوان مسیری عمومی برای طراحی مدل های توضیح پذیرتر در کاربردهای مبتنی بر داده های تصویری مورد استفاده قرار گیرد.

نتایج این پژوهش نشان می دهد که ادغام قیود توضیح پذیری در فرآیند آموزش، نقش مهمی در کنترل رفتار مدل ایفا می کند؛ به گونه ای که شبکه عصبی به صورت هدفمند به نواحی معنادار تصویر هدایت شده و از اتکا به سرنخ های نامرتبط یا نویزی پرهیز می نماید. همچنین، نقشه های حرارتی تولید شده به عنوان یک لایه ی ارتباطی میان سامانه ی هوشمند و کاربر انسانی عمل کرده و امکان تحلیل، پایش و اعتماد سازی در تصمیمات مدل را فراهم می سازند. از سوی دیگر، استفاده از توابع زیان پیشنهادی باعث ارتقای سطح هوش سیستم از طریق هم راستاسازی فرآیند استنتاج با منطق قابل تبیین شده و این امر، استقرار چنین مدل هایی را در سامانه های مبتنی بر محاسبات و تصمیم یارهای عملیاتی حساس، واقع بینانه تر می سازد. بنابراین، رویکرد ارائه شده می تواند به عنوان گامی مؤثر در جهت توسعه ی

تحقیقاتی موارد زیر از جمله بخش‌هایی هستند که موجب تاثیر بسزایی افزایش توضیح‌پذیری می‌شوند.

- آزمون و بررسی توابع ریاضیاتی دیگر برای هم‌راستایی ماسک و نقشه حرارتی

- تولید ماسک تصاویر با استفاده از روشی با اطمینان بالاتر و بدون نیاز به دستور

گسترش بحث یادگیری مبتنی بر توضیح‌پذیری برای سایر کاربردهای یادگیری عمیق شامل متن، صوت و سیگنال.

[۷] Mitsuhashi, Masahiro, et al. (۲۰۱۹), "Embedding human knowledge into deep neural network via attention map." *arXiv preprint arXiv:1905.03540*.

[۸] Ismail, Aya Abdelsalam, Hector Corrada Bravo, and Soheil Feizi. (۲۰۲۱), "Improving deep learning interpretability by saliency guided training." *Advances in Neural Information Processing Systems* ۳۴: ۲۶۷۲۶-۲۶۷۳۹.

[۹] DeYoung, Jay, et al. (۲۰۲۰), "ERASER: A benchmark to evaluate rationalized NLP models." *Proceedings of the 58th annual meeting of the association for computational linguistics*.

[۱۰] Chrysostomou, George, and Nikolaos Aletras. (۲۰۲۱), "Enjoy the salience: Towards better transformer-based faithful explanations with word salience." *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

[۱۱] Atanasova, Pepa, et al. (۲۰۲۲), "Diagnostics-guided explanation generation." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. ۳۶. No. ۱۰.

[۱۲] Situ, Xuelin, et al. (۲۰۲۱), "Learning to explain: Generating stable explanations fast." *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

[۱۳] Ravi, Nikhila, et al. (۲۰۲۰), "Sam ۲: Segment anything in images and videos." *International Conference on Learning Representations*. Vol. ۲۰۲۰

[۱۴] Le, Yann, and Xuan Yang. (۲۰۱۰), "Tiny imagenet visual recognition challenge." *CS 231N* ۷, ۷: ۳.

سامانه‌های بینایی کامپیوتری مبتنی بر شبکه‌های عصبی عمیق با قابلیت تصمیم‌گیری شفاف در نظر گرفته شده و این شبکه‌های عصبی برای استفاده در سامانه‌های فرماندهی و کنترل نظامی قابل کنترل و قابل اعتماد تلقی شود.

در پایان، می‌توان نتیجه گرفت که رویکرد پیشنهادی گامی مؤثر در جهت ترکیب آموزش هدف‌محور با تبیین‌پذیری مدل‌ها بوده و می‌تواند به عنوان یک مسیر عمومی برای طراحی مدل‌های توضیح‌پذیرتر در کاربردهای مبتنی بر داده‌های تصویری مورد استفاده قرار گیرد. برای تبیین ادامه این مسیر

۵. مراجع

[۱] Y. Gao, S. Gu, J. Jiang, S. R. Hong, D. Yu, and L. Zhao (Apr. ۲۰۲۴), "Going Beyond XAI: A Systematic Survey for Explanation-Guided Learning", *ACM Comput Surv*, Vol ۵۶, No ۷, PP ۱۸۸:۱-۱۸۸:۳۹, , doi: ۱۰.۱۱۴۵/۳۶۴۴۰۷۳.

[۲] I. E. Nielsen, D. Dera, G. Rasool, R. P. Ramachandran, and N. C. Bouaynaya (July ۲۰۲۲), "Robust Explainability: A tutorial on gradient-based attribution methods for deep neural networks," *IEEE Signal Process. Mag.*, Vol ۳۹, No ۴, PP ۷۳-۸۴, doi: ۱۰.۱۱۰۹/۰۰۰۰.۲۰۲۲.۳۱۴۲۷۱۹.

[۳] A. Hedström et al. (Apr. ۲۸, ۲۰۲۳), "Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond", *arXiv:arXiv:۲۲۰۲.۰۶۸۶۱*. doi: ۱۰.۴۸۵۵۰/arXiv.۲۲۰۲.۰۶۸۶۱.

[۴] Chen, Yanhui, Chao Chen, Xujia Li, Dongsheng Hong, and Xiangwen Liao. (Oct. ۲۴, ۲۰۲۰), "Focusex: Generalizable Explanation Guided Learning Through Noise Reduction and Fine-Grained Localization." *Available at SSRN 5403119*.

[۵] Shu, Shelley Zixin, et al. (۲۰۲۰), "Hybrid Explanation-Guided Learning for Transformer-Based Chest X-Ray Diagnosis." *International Workshop on Interpretability of Machine Intelligence in Medical Image Computing*. Cham: Springer Nature Switzerland.

[۶] Zeng, Guohang, et al. (۲۰۲۱), "Generating deep networks explanations with robust attribution alignment." *Asian Conference on Machine Learning*. PMLR.