



دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

پایان‌نامه‌ی کارشناسی
مهندسی کامپیوتر

عنوان:

تفسیرپذیری شبکه U-Net در قطعه‌بندی تصاویر پزشکی

نگارش:

زهرا فاضل

استاد راهنما:

دکتر حمیدرضا ربیعی

۱۴۰۰ بهمن

اللهُ أَكْبَرُ

چکیده

پیشرفت‌های اخیر در زمینه تصویربرداری و محاسبات منجر به افزایش استفاده از یادگیری ماشین برای تحلیل تصاویر پزشکی شده است. ظهور یادگیری عمیق سطوح بسیار بالاتری از انتزاع را برای انتخاب و گزینش ویژگی‌ها فراهم آورد. نشان داده شده است که شبکه‌های عصبی پیچشی چکیده‌هایی را که از تصاویر پزشکی چند بعدی به دست آمده‌اند، یاد می‌گیرند، ویژگی‌هایی که تعریف آن‌ها توسط انسان دشوار است. این یکی از دلایلی است که باعث می‌شود شبکه‌های عصبی پیچشی در تشخیص و قطعه‌بندی برتری داشته باشند. با این وجود، اکثر این شبکه‌ها قادر تفسیرپذیری‌اند. تفسیرپذیری در شبکه‌های یادگیری عمیق، پاسخ‌دهنده به این است که چگونه شبکه یک خروجی خاص را تولید می‌کند و این امر در کاربردهای پزشکی اهمیت بیشتری می‌یابد چرا که باعث قابل اطمینان بودن مدل برای استفاده خواهد شد. با وجود آن که در بحث تفسیرپذیری برای مسائل دسته‌بندی روش‌های زیادی توسعه داده شده است، اما پژوهش‌های زیادی در مورد تفسیرپذیری مدل‌های قطعه‌بندی صورت نگرفته است. در این پژوهه ابتدا شبکه U-Net را برای قطعه‌بندی توده درون روده‌بزرگ و لکه‌های سرطانی روی پوست آموختند. داده شده سپس با اعمال روش‌های مختلف تفسیرپذیری روی شبکه، ویژگی‌های مؤثر در تصمیم‌گیری شبکه مشخص شده و در نهایت با استفاده از معیارهای ارزیابی تفسیرپذیری نشان داده شده است که روش Gead-CAM بهترین و قابل اعتمادترین نتایج را برای تفسیرپذیری مدل U-Net ارائه می‌دهد.

کلیدواژه‌ها: شبکه U-Net، قطعه‌بندی تصاویر، تفسیرپذیری، یادگیری عمیق، بینایی ماشین

فهرست مطالب

۸	۱	مقدمه
۸	۱-۱	تعریف مسئله
۹	۲-۱	اهمیت موضوع
۱۰	۳-۱	ادبیات موضوع
۱۰	۴-۱	اهداف تحقیق
۱۱	۵-۱	ساختار پایاننامه
۱۲	۲	مفاهیم اولیه
۱۲	۱-۲	تفسیرپذیری
۱۲	۲-۲	شبکه U-Net
۱۴	۳-۲	معیارهای ارزیابی قطعه‌بندی تصاویر
۱۵	۴-۲	توابع هزینه
۱۶	۵-۲	معیارهای ارزیابی آسیب‌پذیری تفسیرپذیری
۱۷	۳	کارهای پیشین
۱۷	۱-۳	روش‌های تخصیص پیکسلی
۱۷	۱-۲-۱	روش‌های برپایه گرادیان

۲۱	روش‌های بر پایه تفاضل مرجع	۱-۳
۲۲	روش‌های بر پایه مفهوم	۲-۳
۲۲	TCAV	۱-۲-۳
۲۲	D-TCAV	۲-۲-۳
۲۳	جمع‌بندی	۳-۳
۲۴	روش‌ها و پیاده‌سازی	۴
۲۴	تنظیمات آزمایش‌ها	۱-۴
۲۴	دادگان	۲-۴
۲۵	Mجموعه دادگان KVASIR-SEG	۱-۲-۴
۲۵	HAM10000	۲-۲-۴
۲۵	پیش‌پردازش دادگان	۳-۴
۲۶	مدل	۴-۴
۲۶	تابع هزینه	۵-۴
۲۸	روش‌های تفسیرپذیری	۶-۴
۲۹	نتایج	۵
۲۹	نتایج مدل	۱-۵
۳۰	تفسیر نقشه‌ها	۲-۵
۳۱	ارزیابی آسیب‌پذیری روش‌های تفسیرپذیری	۳-۵
۳۶	جمع‌بندی و نگاه به آینده	۶

فهرست شکل‌ها

۱۳	۱-۲ معماری U-Net
۱۹	۱-۳ معماری روش شبکه عکس‌پیچشی
۲۵	۱-۴ نمونه تصاویر مجموعه‌های دادگان
۲۷	۲-۴ معماری مدل استفاده شده
۳۲	... KVASIR-SEG	۱-۵ نمونه‌ای از نقشه‌های تفسیرپذیری مدل برای مجموعه دادگان
۳۳	... HAM10000	۱-۵ نمونه‌ای از نقشه‌های تفسیرپذیری مدل برای مجموعه دادگان
۳۴	... Grad-CAM	۱-۵ نمونه‌ای از نقشه‌های Grad-CAM لایه‌ها برای مجموعه دادگان
۳۴	... HAM10000	۱-۵ نمونه‌ای از نقشه‌های Grad-CAM لایه‌ها برای مجموعه دادگان
۳۵	... KVASIR-SEG	۱-۵ نمونه‌ای از نقشه‌های DeepLift لایه‌ها برای مجموعه دادگان
۳۵	... HAM10000	۱-۵ نمونه‌ای از نقشه‌های DeepLift لایه‌ها برای مجموعه دادگان

فهرست جداول

۱-۳	بررسی روش‌های مختلف تفسیرپذیری	۲۳
۱-۵	نتایج مدل برای مجموعه دادگان KVASIR-SEG	۲۹
۲-۵	نتایج مدل برای مجموعه دادگان HAM10000	۳۰
۳-۵	ارزیابی آسیب‌پذیری روش‌های تفسیر مدل با معیار عدم صحت	۳۱
۴-۵	ارزیابی آسیب‌پذیری روش‌های تفسیر لایه با معیار عدم صحت	۳۱

فصل ۱

مقدمه

۱-۱ تعریف مسئله

تشخیص بیماری‌ها با استفاده از روش‌های یادگیری ماشین^۱ بسیار اهمیت یافته‌است، اما یک مشکل اساسی در اعتماد به خروجی مدل‌های یادگیری ماشین آن است که این مدل‌ها بر اساس ویژگی‌هایی نظری مقدار نقاط تصویر^۲ عمل می‌کنند که با مفاهیم سطح بالای قابل درک برای انسان‌ها معادل نیست. همین موضوع باعث می‌شود که در کاربردهای حیاتی مانند کاربردهای پزشکی، متخصصان آن حوزه به مدل‌های یادگیری ماشین اطمینان نداشته باشند. بنابراین لازم است طبیعت جعبه سیاه^۳ بودن مدل‌های یادگیری ماشین رمزگشایی شود تا متخصصان به آن‌ها اعتماد کنند.^[۱] برای این موضوع از روش‌های تفسیرپذیری^۴ استفاده می‌شود تا منطق و استدلال مدل قابل درک برای انسان باشد. در این پژوهه به بررسی و ارزیابی نتایج تفسیرپذیری قطعه‌بندی^۵ توده‌های^۶ سرطان روده بزرگ و لکه‌های سرطانی پوست با استفاده از مدل U-Net و تعمیم روش‌های تفسیرپذیری دسته‌بندی به قطعه‌بندی می‌پردازیم.

Machine Learning^۱

Pixel^۲

Black Box^۳

Interpretability^۴

Segmentation^۵

Polyp^۶

۱-۲ اهمیت موضوع

بسیاری از مدل‌های یادگیری ماشین مانند سیستم‌های پیشنهاد‌دهنده^۷ نیازی به توضیح چرایی خروجی خود ندارند، چرا که در محیط‌های با ریسک کم استفاده می‌شوند که بدان معناست که خروجی اشتباه، پیامدهای جدی در پی ندارد و یا بعضی از مدل‌ها به صورت گسترده مطالعه و ارزیابی شده‌اند و نیازی به توضیح چرایی تصمیم آن‌ها وجود ندارد. نیاز به تفسیرپذیری ناشی از ناکاملیت در تعریف مسئله است بدان معنا که برای بعضی مسائل، صرف گرفتن خروجی مدل کافی نیست بلکه مدل باید شرح دهد که چگونه به این پیش‌بینی رسیده‌است چرا که پیش‌بینی درست فقط بخشی از مسئله اصلی را حل می‌کند. دلایل زیر محرکه تقاضا برای تفسیرپذیری شدند:

- در مدل‌های یادگیری ماشین، خود مدل می‌تواند علاوه بر مجموعه دادگان^۸، منبع دانش باشد و تفسیرپذیری این امکان را به وجود می‌آورد تا این دانش از مدل استخراج گردد.
- برای استفاده از مدل‌های یادگیری ماشین در زندگی روزمره باید ابتدا از امنیت آن‌ها اطمینان حاصل شود. برای مثال، در یک ماشین خودران^۹ که دوچرخه‌سوارها را با استفاده از مدل یادگیری عمیق^{۱۰} تشخیص می‌دهد، باید کاملاً مطمئن بود که انتزاعی که سیستم یاد گرفته بدون هر گونه خطاست. روش‌های تفسیرپذیری می‌توانند نشان دهند که مهم‌ترین ویژگی‌های یاد گرفته‌شده آیا به صورت یکتا باعث تشخیص دوچرخه می‌شوند یا خیر.
- تفسیرپذیری ابزاری کاربردی برای خطایابی^{۱۱} و تشخیص اربی^{۱۲} در مدل است.
- تفسیرپذیر بودن مدل‌ها باعث افزایش مقبولیت آن‌ها برای کاربردهای روزمره در زندگی می‌شود.
- بررسی مواردی مانند بی‌طرفی^{۱۳}، نیرومندی^{۱۴}، پایایی^{۱۵}، علیت^{۱۶} و عدم قطعیت^{۱۷} در مدل‌های تفسیرپذیر راحت‌تر است.

Recommender Systems ^۷
Dataset ^۸
Self-driving Car ^۹
Deep Learning ^{۱۰}
Debug ^{۱۱}
Bias ^{۱۲}
Fairness ^{۱۳}
Robustness ^{۱۴}
Reliability ^{۱۵}
Causality ^{۱۶}
Uncertainty ^{۱۷}

۱-۳ ادبیات موضوع

روش‌های تفسیرپذیری بر دو نوع‌اند:

- روش‌های مبتنی بر ویژگی این روش‌ها به هر ویژگی بر اساس اهمیت آن، امتیاز اختصاص می‌دهند.
- روش‌های مبتنی بر داده این روش‌ها به هر داده ورودی بر اساس اهمیت آن در یادگیری، امتیاز اختصاص می‌دهند.^[۳]

برای مدل‌های دسته‌بندی^{۱۸} که یک خروجی نهایی از بین C دسته مختلف داریم، روش‌های تفسیرپذیری مختلفی ارائه شده که در کتابخانه‌هایی نظری Torchray، CNN Visualization و Captum پیاده‌سازی شده‌اند، اما این موضوع برای مدل‌های قطعه‌بندی همچنان یک چالش بزرگ است. یک راه برای این چالش نگاه کردن به مسئله قطعه‌بندی مانند یک مسئله دسته‌بندی است. یک مسئله قطعه‌بندی را که قطعه‌های ماسک C دسته مختلف دارند، می‌توان به صورت یک مسئله دسته‌بندی برای هر نقطه تصویر ماسک در نظر گرفت که هر نقطه تصویر باید در یکی از C دسته مختلف قرار گیرد. بدیهتاً با توجه به تعداد نقطه تصویرهای یک ماسک، اعمال روش‌های تفسیرپذیری بر روی همه آن‌ها نیازمند منابع سخت‌افزای و زمان زیاد است. برای رفع این مشکل برای هر قطعه نقاط تصویر مرزی را تعریف می‌کنیم. نقاط تصویر مرزی برای یک قطعه، نقاط تصویری هستند که حداقل یکی از چهار نقطه تصویر مجاور آن‌ها در یک دسته دیگر قرار می‌گیرد. این نقاط معین‌کننده شکل قطعه در ماسک هستند، بنابراین آن‌چه در بحث تفسیرپذیری مدل‌های قطعه‌بندی برای ما مهم است آن است که مدل چگونه در مورد این نقاط تصویر مرزی تصمیم گرفته است و لذا اعمال روش‌های تفسیرپذیری را می‌توان به این نقاط محدود کرد.

۱-۴ اهداف تحقیق

در این پایان‌نامه تلاش می‌شود مدل U-Net برای قطعه‌بندی توده‌های سرطان روده بزرگ و لکه‌های سرطانی پوست آموزش داده شده و سپس با اعمال روش‌های تفسیرپذیری بر روی ماسک‌های خروجی، ویژگی‌های مؤثر در تصمیم‌گیری شبکه شناسایی شده و میزان آسیب‌پذیر بودن تفسیرهای ارائه شده مشخص شود.

۱-۵ ساختار پایاننامه

این پایاننامه شامل شش فصل است. فصل دوم دربرگیرنده مفاهیم اولیه مرتبط با پایاننامه است. در فصل سوم، روش‌های ارائه‌شده برای تفسیرپذیری و تحقیقات در زمینه تفسیرپذیری برای مدل‌های قطعه‌بندی به تفصیل بیان می‌شوند. در فصل چهارم به توضیح مدل، مجموعه دادگان و شیوه اعمال روش‌های تفسیرپذیری پرداخته می‌شود. در فصل پنجم نتایج به دست آمده در این پایاننامه ارائه می‌گردد. در فصل ششم به جمع‌بندی و پیشنهادهایی برای کارهای آتی پرداخته خواهد شد.

فصل ۲

مفاهیم اولیه

۱-۲ تفسیرپذیری

تفسیرپذیری معیاری برای آن است که تا چه میزان انسان می‌تواند علت تصمیم مدل را درک کند. هر چقدر میزان تفسیرپذیری یک مدل یادگیری ماشین بیشتر باشد، درک چرا ای تصمیم‌های آن و پیش‌بینی خروجی‌های آن برای انسان ممکن‌تر است. در واقع تفسیرپذیری مدل‌های یادگیری ماشین به این موضوع می‌پردازد که رفتار مدل را بفهمد و بررسی کند که مدل بر اساس چه ویژگی‌هایی از ورودی، یاد می‌گیرد و آیا این ویژگی‌ها، بازتاب ارزش‌های ما در تشخیص هست یا خیر. تفسیرپذیری به این سوال پاسخ می‌دهد که «چرا یک شبکه عصبی یک خروجی خاص را می‌هد؟» [۱، ۴، ۲].

۲-۲ شبکه U-Net

شبکه عصبی پیچشی^۱ U-Net، یک ساختار کدگزار^۲ – کدگشا^۳ دارد که با اتصالات ردشونده^۴ به یکدیگر پل زده‌اند. این شبکه در اصل برای قطعه‌بندی تصاویر پزشکی تعریف شده است و در قطعه‌بندی تصاویر سایر زمینه‌ها نیز پرکاربرد و موفق بوده است [۵].

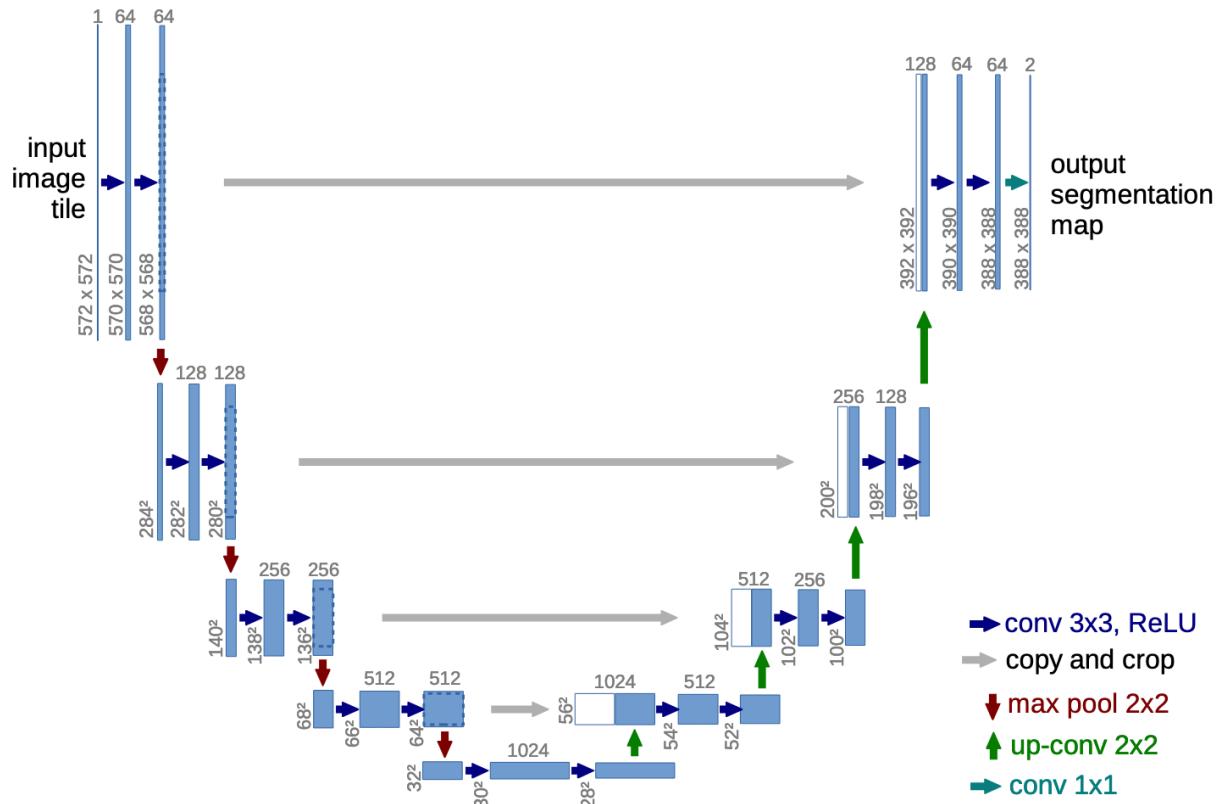
Convolutional Neural Network^۱

Encoder^۲

Decoder^۳

Skip connections^۴

این شبکه تنها از لایه‌های پیچشی^۵ و لایه‌های ادغام^۶ تشکیل شده است و به علت عدم استفاده از لایه‌های کاملاً همبند، توانایی پردازش تصاویر با ابعاد گوناگون را دارد. تصویری از معماری U-Net در شکل ۲-۱ نشان داده شده است.



شکل ۲-۱: معماری U-Net – مستطیل‌های آبی نشان‌دهندهٔ نقشه‌های ویژگی است که تعداد کانال‌های آن در بالایش نوشته شده است. عملکرد هر فاش در گوشه پایین در قسمت توضیحات نوشته شده است. مستطیل‌های سفید که در کنار مستطیل‌های آبی دیده می‌شوند اتصالات ردشونده بین کدگذار و کدگشا است [۵].

۳-۲ معیارهای ارزیابی قطعه‌بندی تصاویر

در این قسمت معیارهای ارزیابی استفاده شده در گزارش نتایج این پایان‌نامه توضیح داده‌می‌شود.

تعريف ۲-۱ (معیار Dice) معیار $Dice$, که تحت عنوان امتیاز $F1^{\text{V}}$ نیز شناخته می‌شود، یک روش برای تعیین درصد همپوشانی بین ماسک هدف و ماسک پیش‌بینی شده است. این معیار نسبت بین دو برابر اشتراک تعداد نقاط تصویر ماسک هدف و پیش‌بینی شده و مجموع تعداد نقاط دو ماسک را اندازه‌گیری می‌کند. به عبارتی برای محاسبه $Dice$ بین دو ماسک هدف (T) و پیش‌بینی (P) داریم:

$$Dice(T, P) = \frac{2|T \cap P|}{|T| + |P|} \quad (1-2)$$

که در رابطه‌ی (۱-۲) عملگر $| \cdot |$ نشان‌دهنده‌ی تعداد نقاط تصویر در ناحیه است.

در صورتی که یک مدل بتواند ماسک هدف را به طور دقیق پیش‌بینی کند، امتیاز $Dice$ دقیقاً برابر با یک می‌شود. اما در عمل این امر امکان‌پذیر نیست، بنابراین از معیار $MDice$ استفاده می‌کنیم که از مقادیر $Dice$ برای پیش‌بینی‌ها میانگین می‌گیرد.

تعريف ۲-۲ (معیار IoU) معیار IoU^{A} , که تحت عنوان شاخص جاکارد^۹ نیز شناخته می‌شود، روش دیگری برای تعیین درصد همپوشانی بین ماسک هدف و ماسک پیش‌بینی شده است. این معیار نسبت بین اشتراک تعداد نقاط تصویر ماسک هدف و پیش‌بینی شده و اجتماع بین تمامی نقاط ماسک‌های آن‌ها را اندازه‌گیری می‌کند. به عبارتی برای محاسبه IoU بین دو ماسک هدف (T) و پیش‌بینی (P) داریم:

$$IoU(T, P) = \frac{|T \cap P|}{|T \cup P|} \quad (2-2)$$

که در رابطه‌ی (۲-۲) عملگر $| \cdot |$ نشان‌دهنده‌ی تعداد نقاط تصویر در ناحیه است.

در صورتی که یک مدل بتواند ماسک هدف را به طور دقیق پیش‌بینی کند، امتیاز IoU دقیقاً برابر با یک می‌شود. اما در عمل این امر امکان‌پذیر نیست، بنابراین از معیار $MIoU$ استفاده می‌کنیم که از مقادیر IoU برای پیش‌بینی‌ها میانگین می‌گیرد.

F1 Score^V
Intersection over Union^A
Jaccard's Index⁹

۴-۲ توابع هزینه

در این قسمت توابع هزینه استفاده شده در این پایان نامه توضیح داده می شود.

تعريف ۲-۳ (تابع هزینه آنتروپی متقاطع دودویی^{۱۰}) در یک مسئله دسته بندی با دو دسته مختلف برای خروجی، فرض کنیم y_i مقدار واقعی خروجی و \hat{y}_i مقدار پیش‌بینی شده توسط مدل به ازای ورودی x_i باشد. در این صورت تابع هزینه آنتروپی متقاطع دودویی برای ورودی‌های x_1, \dots, x_n به صورت زیر تعریف می شود:

$$\mathcal{L}_{BCE} = -\frac{1}{n} \sum_{i=1}^n y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \quad (3-2)$$

اگر به مسئله قطعه بندی نمونه‌ای^{۱۱} مانند یک مسئله دسته بندی دودویی نقاط عکس نگاه کنیم، آنگاه می‌توان از این تابع هزینه در قطعه بندی نیز استفاده کرد.

تعريف ۲-۴ (تابع هزینه Dice) به ازای ماسک هدف T و ماسک پیش‌بینی شده P ، تابع هزینه Dice برابر است با:

$$\mathcal{L}_{Dice} = 1 - Dice(T, P) = 1 - \frac{2|T \cap P|}{|T| + |P|} = \frac{|T \cup P| - |T \cap P|}{|T| + |P|} \quad (4-2)$$

در واقع با کم کردن این تابع هزینه، تعداد نقاط اشتباه در ماسک پیش‌بینی شده کم می شود.

تعريف ۲-۵ (تابع هزینه IoU) به ازای ماسک هدف T و ماسک پیش‌بینی شده P ، تابع هزینه IoU برابر است با:

$$\mathcal{L}_{IoU} = 1 - IoU(T, P) = 1 - \frac{|T \cap P|}{|T \cup P|} = \frac{|T \cup P| - |T \cap P|}{|T \cup P|} \quad (5-2)$$

در واقع با کم کردن این تابع هزینه، اختلاف اجتماع و اشتراک نقاط دو ماسک که همان تعداد نقاط اشتباه در ماسک پیش‌بینی شده کم می شود.

۲-۵ معیارهای ارزیابی آسیب‌پذیری تفسیرپذیری

معیارهای کمی ارزیابی تفسیرپذیری عبارتند از: عدم صحت^{۱۲} و حساسیت^{۱۳}. [۶]

تعريف ۲-۶ (معیار عدم صحت) تابع جعبه سیاه f ، تابع تفسیر ϕ و ورودی x را داریم. فرض کنیم $I \in \mathbb{R}^d$ که d اندازه x است، یک متغیر تصادفی باشد که نمایانگر آشفتگی^{۱۴} های معنادار است. در این صورت عدم صحت تابع تفسیر ϕ به شکل زیر تعریف می‌شود:

$$INFD(\phi, f, x) = \mathbb{E}_I [(I^T \phi(f, x) - (f(x) - f(x - I)))^2] \quad (6-2)$$

تعريف ۲-۷ (معیار حساسیت) تابع جعبه سیاه f ، تابع تفسیر ϕ ، ورودی x و شعاع آشفتگی r را داریم. در این صورت حساسیت تابع تفسیر ϕ به شکل زیر تعریف می‌شود:

$$SENS(\phi, f, x, r) = \max_{\|y-x\| \leq r} \|\phi(f, y) - \phi(f, x)\| \quad (7-2)$$

در این پروژه به دلیل محدودیت امکانات سخت‌افزاری از معیار عدم صحت برای ارزیابی استفاده شده است.

Infidelity^{۱۲}
Sensitivity^{۱۳}
Perturbation^{۱۴}

فصل ۳

کارهای پیشین

روش‌های تفسیرپذیری مبتنی بر ویژگی به دو دسته کلی تخصیص^۱ پیکسلی و بر پایه مفهوم تقسیم می‌شوند. روش‌های تخصیص پیکسلی خود دو دسته‌اند: بر پایه گرادیان و بر پایه تفاضل مرجع.

۱-۱-۱ روشهای تخصیص پیکسلی

۱-۱-۱-۱ روشهای بر پایه گرادیان

Image-Specific Class Saliency روش

این روش از اولین روش‌های تخصیص پیکسلی است. در این روش، با محاسبه گرادیان امتیازهای لایه آخر مدل برای هر کلاس نسبت به نقاط تصویر ورودی، به نقشه‌ای همانندازه با ویژگی‌های ورودی می‌رسیم که دارای مقادیر منفی و مثبت است. مراحل الگوریتم به شرح زیر است:

- یک گذر جلو^۲ از مدل با عکس دلخواه I . انجام می‌دهیم.
- شبکه عصبی پیچشی امتیاز $(I_c)_S$ را برای کلاس c به ازای ورودی I خروجی می‌دهد.
- یکتابع غیرخطی از عکس ورودی است اما می‌توان با استفاده از بسط تیلور مرتبه اول آن

Attribution^۱
Forward Pass^۲

را به صورت یک تابع خطی تخمین زد:

$$S_c(I) \simeq w^T I + b \quad (1-3)$$

- نقشه خروجی همان w است که به صورت زیر به دست می‌آید:

$$w = \frac{\partial S_c}{\partial I} \Big|_I. \quad (2-3)$$

برای محاسبه این گرادیان از الگوریتم پس‌انتشار^۴ استفاده می‌شود. ایراد این روش آن است که به دلیل وجود توابع فعال‌سازی^۵ مانند تابع یک‌سوساز خطی واحد^۶ در گذر عقب^۷ که علامت را حذف می‌کنند، وقتی مقدار فعال‌سازی یک نورون^۸ باشد نمی‌توان مقدار پس‌انتشار را به صورت یکتا تعیین کرد. در چنین شرایطی همان مقدار^۹ را به عنوان مقدار پس‌انتشار می‌فرستد. که در نتیجه آن در ادامه نیز همین مقدار^۹ انتشار می‌یابد و دیگر مقدار فعال‌سازی نغیر نمی‌کند.^[۲، ۷]

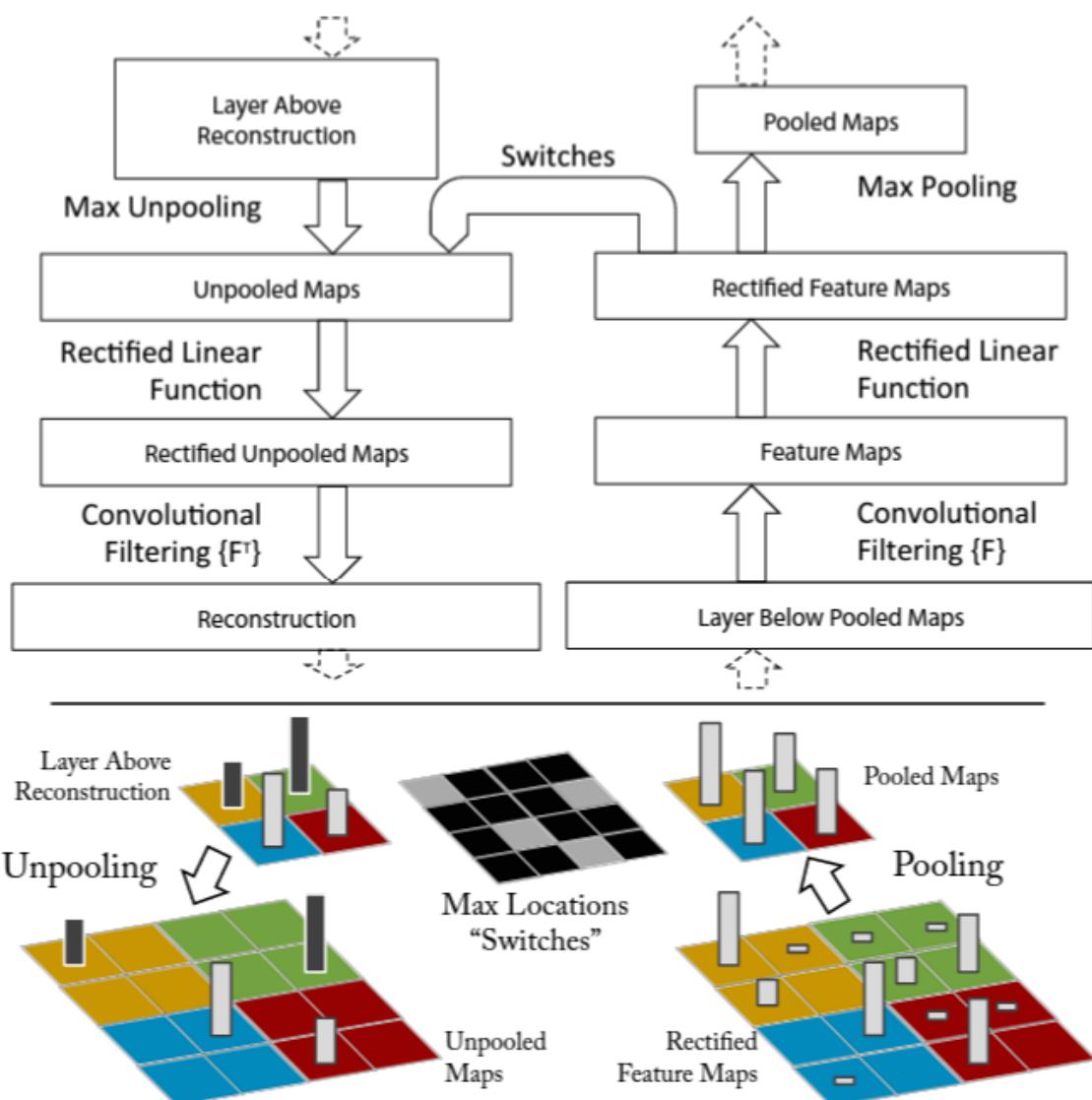
روش DeconvNet

در این روش که تعمیم روش قبل است، شبکه عصبی با معکوس کردن همه لایه‌ها، معکوس می‌شود. همان‌گونه که در ساختار این شبکه در شکل ۱-۳ می‌توان دید، برای عکس لایه ادغام از لایه عکس ادغام^{۱۰} و برای عکس لایه پیچشی از ترانهاده فیلترهای یادگرفته شده در لایه قبل استفاده می‌کند. از آنجا که خروجی لایه ادغام برگشت‌پذیر نیست، شبکه در هنگام گذر جلو در شبکه پیچشی اصلی مکان بیشینه محلی در لایه‌های ادغام را ذخیره می‌کند و در لایه‌های عکس ادغام با استفاده از این مقادیر ذخیره شده، بیشینه محلی در هر ناحیه ادغام را بازیابی می‌کند.^[۸]

روش Guided Backpropagation

این روش که ترکیبی از دو روش قبل است، با فرض آن که فقط ویژگی‌های با امتیاز مثبت برای ما اهمیت دارند، هنگام انتشار گرادیان فقط مقادیر مثبت گرادیان را نگه می‌دارد و مقادیر منفی را با^{۱۱} جایگزین می‌کند. در این روش در گذر عقب هم انتشار و هم عکس‌پیچش صورت می‌گیرد.^[۹]

Backpropagation ^۳
Activation Function ^۴
Rectifying Linear Unit (ReLU) ^۵
Backward Pass ^۶
Unpooling ^۷



شكل ۳-۱: بالا: یک لایه عکس‌پیچشی وصل شده به لایه پیچشی - پایین: عملگر عکس‌ادغام [۸].

روش نقشه فعالسازی دسته

یک نقشه فعالسازی دسته ^۱ برای یک دسته خاص مشخص‌کننده ناحیه‌هایی از عکس است که در تشخیص آن دسته استفاده شده‌اند. برای ساختن نقشه فعالسازی کلاس از ادغام میانگین سراسری^۹ در شبکه‌های عصبی پیچشی استفاده می‌شود. فرض کنیم برای عکس دلخواه، $f_k(x, y)$ نمایانگر فعالسازی نورون k در لایه پیچشی آخر در محل (x, y) باشد. در این صورت برای نورون k ، مقدار ادغام میانگین w_k^c سراسری یا F_k برابر با $\Sigma_{(x,y)} f_k(x, y)$ است و بنابراین امتیاز دسته c یا S_c برابر با $\Sigma_k w_k^c F_k$ است که وزن معادل دسته c در نورون k است.

$$S_c = \Sigma_k w_k^c \Sigma_{(x,y)} f_k(x, y) = \Sigma_{(x,y)} \Sigma_k w_k^c f_k(x, y) \quad (۳-۳)$$

نقشه فعالسازی دسته یا M_c به شکل زیر تعریف می‌شود:

$$M_c(x, y) = \Sigma_k w_k^c f_k(x, y) \implies S_c = \Sigma_{(x,y)} M_c(x, y) \quad (۴-۳)$$

و بنابراین M_c نشان دهنده اهمیت فعالسازی در محل (x, y) است که منجر به دسته‌بندی تصویر در دسته c شده‌است.^[۱۰]

روش Grad-CAM

یکی از مشکلات روش نقشه فعالسازی دسته آن است که فقط برای شبکه‌های عصبی پیچشی قابل استفاده است که قبل از لایه آخر، لایه ادغام میانگین سراسری دارند. روش Grad-CAM که تعیین روش نقشه فعالسازی دسته است، این مشکل را با به کارگیری گرادیان حل می‌کند. برای به دست آوردن نقشه موضع‌یابی $L_{Grad-CAM}^c \in \mathbb{R}^{u \times v}$ ، که u عرض تصویر، v طول آن و c دسته است، ابتدا باید گرادیان امتیاز برای دسته c پیش از لایه بیشینه نرم^{۱۰} را نسبت به فعالسازی نقشه ویژگی A^k حساب کنیم. این گرادیان روی عرض و طول ادغام میانگین سراسری شده تا وزن اهمیت نورون یا α_k^c به دست آید:

$$\alpha_k^c = \Sigma_i \Sigma_j \frac{\partial y^c}{\partial A^k i, j}$$

Class Activation Map (CAM)^{۱۱}
Global Average Pooling^۹
Softmax Layer^{۱۲}.

وزن α_k^c نمایانگر خطیسازی جزئی از شبکه عمیق با شروع از A است و بنابراین میزان اهمیت نقشه ویژگی k را برای دسته c به دست می‌آورد. برای به دست آوردن نقشه موضع‌یابی، تابع ReLU روی جمع $\alpha_k^c A^k$ اعمال می‌شود تا فقط ویژگی‌های با تأثیر مثبت باقی بماند.^[۱۱]

$$L_{Grad-CAM}^c = \text{ReLU}(\sum_k \alpha_k^c A^k) \quad (5-3)$$

روش Guided Grad-CAM

این روش که ترکیبی از روش‌های Guided Backpropagation و Grad-CAM است، نقشه موضع‌یابی را با ضرب درایه‌ای نقشه‌های Guided Backpropagation و Grad-CAM به دست می‌آورد.^[۱۱]

۲-۱-۳ روش‌های بر پایه تفاضل مرجع

روش DeepLift

روش DeepLift که بر اساس تفاضل مرجع عمل می‌کند، بر روش‌های بر پایه گرادیان برتری دارد، چرا که در این روش سیگنال‌های اهمیت حتی در صورت صفر بودن گرادیان نیز انتشار می‌یابند و ناپیوستگی گرادیان در آن تأثیری ندارد.

فرض کنیم t نمایانگر نورون دلخواه در لایه آخر و x_1, \dots, x_n نمایانگر نورون‌هایی در لایه‌های میانی باشند که برای محاسبه t لازم و کافی‌اند. همچنین فرض کنیم t' نشان‌دهنده فعال‌سازی مرجع t باشد. در این صورت تفاضل از مرجع را $\delta t = t - t'$ تعریف می‌کنیم. روش DeepLift امتیاز‌های مشارکت $C_{\delta x_i \delta t}$ را به δx_i اختصاص می‌دهد به صورتی که:

$$\sum_{i=1}^n C_{\delta x_i \delta t} = \delta t$$

تعریف مرجع باید بر اساس دامنه مسئله صورت گیرد.^[۱۲]

۲-۳ روش‌های بر پایه مفهوم

۱-۲-۳ روش TCAV

در روش TCAV^{۱۱} به دو مجموعه دادگان نیاز داریم که یکی برای آموزش و دیگری برای تعریف مفاهیم است. ابتدا بردارهای فعال‌سازی مفهوم^{۱۲} تعریف می‌شوند که نمایش عددی از یک مفهوم در فضای فعال‌ساز شبکه‌اند. برای مثال در مسئله دسته‌بندی حیوانات، یک مفهوم برای تشخیص گورخر راهراه بودن است. برای به دست آوردن بردارهای فعال‌سازی مفهوم شبکه را با مجموعه دادگان مربوط به مفاهیم آموزش می‌دهیم تا بتواند مشخص کند چه تصویری در آن مجموعه دارای یک مفهوم خاص است. در مثال قبل خروجی شبکه برای تصاویر ورودی دو حالت دارد: راهراه هست و راهراه نیست. در نهایت برای ورودی x حساسیت مفهومی^{۱۳} در لایه l به شکل زیر تعریف می‌شود:

$$S_{C,k,l}(x) = \nabla h_{l,k}(f_l(x)).v_l^c$$

که l نمایانگر لایه شبکه، f_l نگارنده x به بردار فعال‌سازی لایه l ، $h_{l,k}$ نگارنده بردار فعال‌سازی به خروجی کلاس k و v_l^c بردار فعال‌سازی مفهوم برای مفهوم C در لایه l است.

حساسیت مفهومی کلی برابر است با:

$$TCAV_{C,k,l} = \frac{|x \in X_k : S_{C,k,l}(x) > 0|}{|X_k|} \quad (6-3)$$

که $|X_k|$ زیرمجموعه‌ای از داده‌هاست که در دسته k قرار می‌گیرند. در مثال گفته شده پیش از این، اگر $TCAV_{\text{striped}, \text{zebra}, l} = 0/8$ باشد، نتیجه می‌شود که ۸۰ درصد خروجی‌های پیش‌بینی شده به عنوان گورخر به صورت مثبت تحت تأثیر مفهوم راهراه بودن، بوده‌اند.^[۴]

۲-۲-۳ روش D-TCAV

روش D-TCAV^{۱۴} که تعمیم بر روش قبلی است، به جای تعریف مفاهیم، آن‌ها را از روی تصاویر به دست می‌آورد. برای این کار از الگوریتم SLIC^{۱۵} استفاده می‌کند و با استفاده از آن نقاط تصویر را بر

Testing with Concept Activation Vectors^{۱۱}

Concept Activation Vector^{۱۲}

Conceptual Sensitivity^{۱۳}

Discovering and Testing with Concept Activation Vectors^{۱۴}

Simple Linear Iterative Clustering^{۱۵}

اساس تشابه مقدار به ۵ دسته تقسیم می‌کند. به هر دسته یک ابرنقطه^{۱۶} گفته می‌شود. سپس هر کدام از این ابرنقطه‌ها را به عنوان یک مفهوم در نظر گرفته و روش TCAV را اجرا می‌کند.^[۱۳]

۳-۳ جمع‌بندی

به طور کلی تا به امروز روش‌های زیادی برای تفسیرپذیری ارائه شده است که به توضیح تعدادی از آن‌ها در بالا پرداخته شد. اکثریت این روش‌ها بر پایه گرادیان هستند و تمامی آن‌ها برای مسئله دسته‌بندی تعریف شده‌اند. اما می‌توان با تعمیم این روش‌ها به مسئله قطعه‌بندی، از آن‌ها برای تفسیرپذیری خروجی شبکه‌های قطعه‌بندی نیز استفاده کرد. باید توجه کرد که هر کدام از این روش‌ها محدودیت‌های خاص خود را دارند که در جدول ۳-۱ نشان داده شده است.

جدول ۳-۱: بررسی روش‌های مختلف تفسیرپذیری

روش	مسئله	مراجع
روش‌های بر پایه گرادیان	عدم نشان دادن تأثیر ویژگی‌های با گرادیان منفی عدم نشان دادن چگونگی ارتباط نواحی مشخص شده در نقشه با خروجی مدل	[۸]، [۲]، [۷] [۱۰]، [۹] [۱۴] و [۱۱]
روش‌های بر پایه تفاضل مرجع	نیاز به متخصص حوزه استفاده برای تعیین تصویر مرجع عدم نشان دادن چگونگی ارتباط نواحی مشخص شده در نقشه با خروجی مدل	[۱۴] و [۱۲]
روش‌های بر پایه مفهوم	نیاز به محاسبات اضافه‌تر و آموزش جداگانه مدل برای یافتن مفاهیم و نقشه فعال‌سازی آن‌ها	[۱۴]، [۱۳] و [۱۴]

فصل ۴

روش‌ها و پیاده‌سازی

۱-۴ تنظیمات آزمایش‌ها

کلیه آزمایش‌ها با استفاده از کتابخانه‌های متن‌باز پایتورچ^۱، کپتم^۲ [۱۵] و SMP[۱۶]^۳ و روی سایت کنگل^۴ با حافظه اصلی ۱۳ گیگابایت و با بهره‌گیری از پردازنده گرافیکی Nvidia Telsa P100 حافظه ۱۶ گیگابایت انجام شده‌اند. تعداد تصاویر هر گروه^۵ برای مجموعه دادگان توده روده بزرگ ۸ و برای مجموعه دادگان سرطان پوست ۶۴ در نظر گرفته شده‌است. تعداد دوره‌ها^۶ برای آموزش مدل روی مجموعه دادگان اول ۵۰۰ دوره و برای آموزش مدل روی مجموعه دادگان دوم ۱۰۰ دوره است. بهینه‌ساز استفاده شده Adam و نرخ یادگیری برای آموزش روی هر دو مجموعه دادگان ۰/۰۰۵ بوده است.

۲-۴ دادگان

در این آزمایش‌ها از دو مجموعه دادگان KVASIR-SEG و HAM10000 استفاده شده‌است. در زمان اجرا هرگز از داده‌های آزمون در فرایند یادگیری مدل استفاده نشده و از آن‌ها فقط برای گزارش نتایج استفاده شده است.

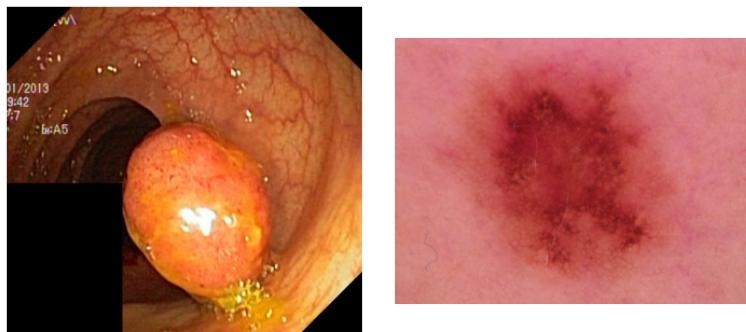
Pytorch^۱
Captum^۲
Kaggle^۳
Batch^۴
Epoch^۵

۱-۲-۴ مجموعه دادگان KVASIR-SEG

این مجموعه دادگان برای قطعه‌بندی توده‌های روده بزرگ ارائه شده است که شامل ۱۰۰۰ تصویر کلونوسکوپی در ابعاد مختلف می‌باشد. ۸۰۰ تصویر برای آموزش، ۱۰۰ تصویر برای اعتبارسنجی^۶ و ۱۰۰ تصویر برای آزمون به کار گرفته شده است. نمونه‌ای از این مجموعه دادگان در چپ شکل ۱-۴ قابل مشاهده است.
[۱۷].

۲-۲-۴ مجموعه دادگان HAM10000

این مجموعه دادگان برای قطعه‌بندی و دسته‌بندی لکه‌های ناشی از سرطان پوست در چالش ISIC 2018 ارائه شده است که شامل ۱۰۰۰۰ تصویر درماتوسکوپیک در ابعاد 450×600 می‌باشد. ۸۰۰ تصویر برای آموزش، ۱۰۰۰ تصویر برای اعتبارسنجی و ۱۰۰۰ تصویر برای آزمون به کار گرفته شده است. نمونه‌ای از این مجموعه دادگان در راست شکل ۱-۴ قابل مشاهده است.
[۱۸، ۱۹].



شکل ۱-۱: چپ: تصویری از مجموعه دادگان KVASIR-SEG – راست: تصویری از مجموعه دادگان KVASIR-SEG

۳-۴ پیش‌پردازش دادگان

تصاویر مجموعه دادگان KVASIR-SEG به ابعاد 400×400 و تصاویر مجموعه دادگان HAM10000 به ابعاد 256×256 در آمدند. ماسک‌ها و تصاویر در داده‌شاختار تنسور به یکدیگر الحاق شدند.

Validation^۶

برای آگمنتیشن^۷ از توابع وارون افقی^۸، وارون عمودی^۹، چرخش، محو کاوی^{۱۰}، قرینه کردن^{۱۱}، تغییر بزرگ‌نمایی و انتقال کتابخانه Albummentation استفاده شده است و این آگمنتیشن‌ها به صورت تصادفی با احتمال ۰/۵ روی داده آموزش اعمال می‌شوند.

۴-۴ مدل

مدل استفاده شده در این پروژه، یک مدل پیش‌آموزش داده شده U-Net^{۱۲} [۲۰] است که ساختار آن مطابق شکل ۲-۴ است. تفاوت این مدل با مدل U-Net استندارد در آن است که تعداد کانال‌های لایه‌های پیچشی نصف است، در لایه‌های پیچشی از Batch Norm استفاده شده و ابعاد عکس و ماسک خروجی یکی است.

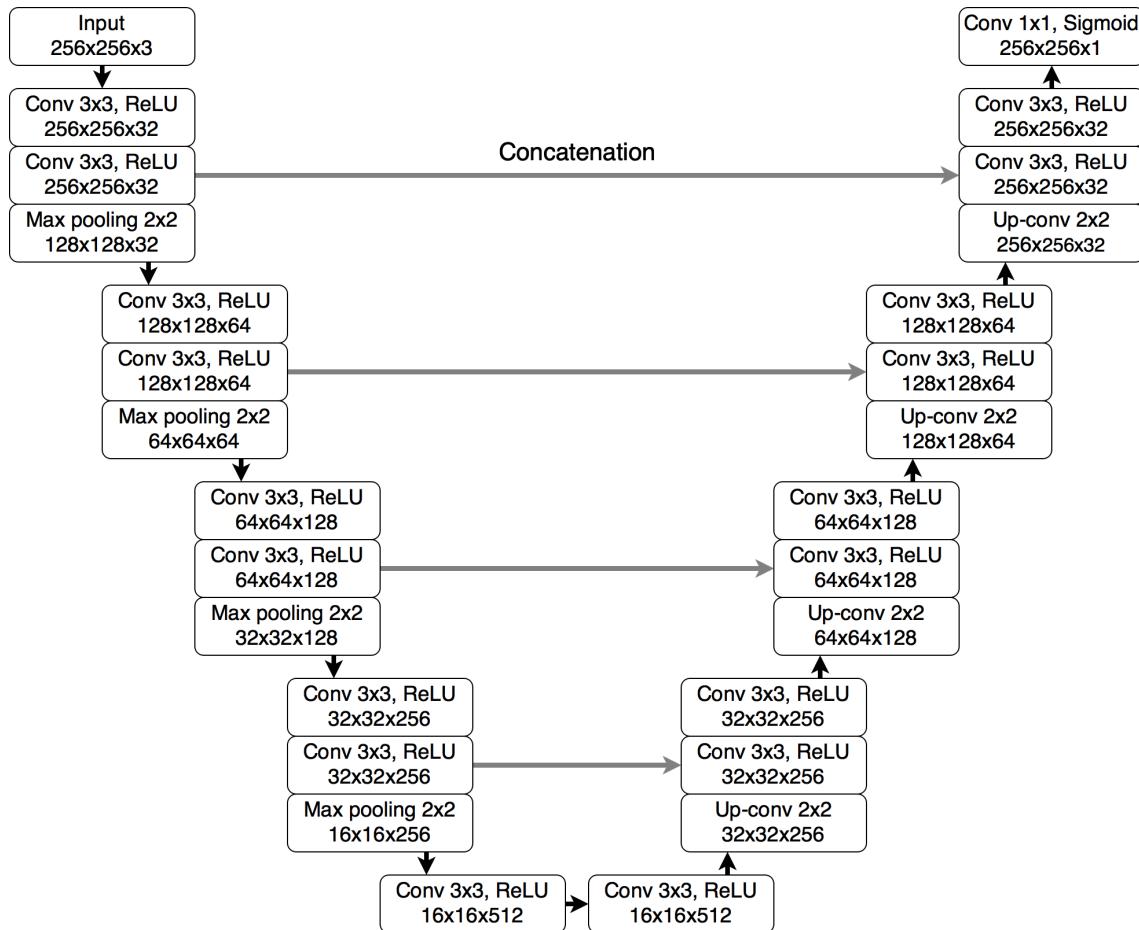
۴-۵ تابع هزینه

برای آموزش مدل از تابع هزینه زیر استفاده شده است که ترکیب خطی تابع هزینه آنتروپی متقطع دودویی، Dice و IoU است:

$$\begin{aligned} \mathcal{L}_{total} &= \mathcal{L}_{BCE} + \mathcal{L}_{IoU} + \mathcal{L}_{Dice} = \\ &\frac{1}{N} \sum_{k=1}^N \left(-\frac{1}{w \times h} \sum_{i=1}^w \sum_{j=1}^h T_{i,j}^k \log P_{i,j}^k + (1 - T_{i,j}^k) \log (1 - P_{i,j}^k) + \right. \\ &\quad \left. (|T^k \cup P^k| - |T^k \cap P^k|) \left(\frac{1}{|T^k \cup P^k|} + \frac{1}{|T^k| + |P^k|} \right) \right) \end{aligned} \quad (1-4)$$

که N تعداد داده‌ها، w و h به ترتیب عرض و طول تصویر، T^k ماسک واقعی ورودی k ، P^k ماسک پیش‌بینی شده برای ورودی k و (i, j) نشان‌دهنده نقطه تصویر در سطر i و ستون j است.

Augmentation ^۷
Horizontal Flip ^۸
Vertical Flip ^۹
Gaussian Blur ^{۱۰}
Transpose ^{۱۱}
Pretrained ^{۱۲}



شکل ۴-۲: معماری مدل پیش‌آموزش‌داده شده U-Net که در آزمایش‌ها استفاده شده و تفاوت آن با مدل ارائه شده در [۵] در کمتر بودن تعداد کانال‌های لایه‌های پیچشی وجود لایه سیگموید^{۱۴} در آخر آن است [۲۰].

۶-۴ روش‌های تفسیرپذیری

برای تعمیم روش‌های تفسیرپذیر دسته‌بندی که پیشتر گفته شد، در قطعه‌بندی استفاده کنیم، مسئله قطعه‌بندی را به صورت مسئله دسته‌بندی نقاط تصویر در نظر می‌گیریم که هر نقطه دو حالت دارد: جزو قطعه باشد و یا جزو پس‌زمینه باشد. آنچه تعیین‌کننده شکل قطعه است، نقاط مرزی آن است که در این جا منظور از نقاط مرزی، نقاطی از قطعه هستند که حداقل یکی از چهار همسایه آن در دسته پس‌زمینه قرار گرفته‌اند. بنابراین آنچه در تفسیرپذیری مدل‌ها قطعه‌بندی مهم است آن است که مدل این مرز را چگونه تشخیص داده است.

در پروژه از روش‌های DeepLift و Guided Grad-CAM، Guided Backpropagation برای به دست آوردن نقشه‌های تفسیرپذیری مدل و از Grad-CAM و DeepLift برای به دست آوردن نقشه‌های تفسیرپذیری لایه‌ها استفاده شده است. ابتدا نقاطه‌ها مرزی ماسک پیش‌بینی شده به دست آورده شده، سپس نقشه تفسیرپذیری به ازای هر نقطه را به دست می‌آوریم. هر کدام از این نقشه‌ها به ابعاد تصویر اصلی است و در هر کدام هر نقطه یک امتیاز دارد. برای به دست آوردن نقشه کلی، میانگین نقشه‌های نقاط مرزی را محاسبه می‌کنیم.

برای بررسی میزان آسیب‌پذیری نتایج تفسیرپذیری از معیار عدم صحت استفاده می‌کنیم. مشابه به روش به دست آوردن نقشه‌ها، برای به دست آوردن عدم صحت، آن را به ازای هر نقطه مرزی حساب کرده و میانگین آن را به عنوان میزان آسیب‌پذیری گزارش می‌کنیم. همچنین از تابع نویز گاوی با انحراف معیار ۰/۰۰ به عنوان متغیر تصادفی ایجاد کننده آشفتگی استفاده شده است.

فصل ۵

نتایج

۱-۵ نتایج مدل

در جدول‌های ۱-۵ و ۲-۵ به بررسی نتایج مدل بر روی مجموعه دادگان آزمون و مقایسه آن با سایر مدل‌ها می‌پردازیم. امتیازهایی که گزارش شده‌اند امتیازهای MIoU و MDice است که همان‌گونه که پیشتر در فصل مفاهیم اولیه گفته شد، میانگین امتیازهای IoU و Dice تصاویر به دست می‌آید.

جدول ۵-۱: مقایسه نتایج مدل با معیارهای MDice و MIoU روی مجموعه دادگان KVASIR-SEG با سایر مدل‌ها

MIoU	MDice	مدل
-	۰/۸۱۸۰	استاندارد U-Net
۰/۷۲۳۹	۰/۸۲۰۶	ColonSegNet
-	۰/۸۸۰۳	FANet
۰/۸۹۱۴	۰/۹۲۱۷	MSRF-Net
۰/۷۵۹۸	۰/۸۴۱۲	خروجی مدل این پروژه

جدول ۵-۲: مقایسه نتایج مدل با معیارهای MDice و MIoU روی مجموعه دادگان HAM10000 با سایر مدل‌ها

MIoU	MDice	مدل
-	۰/۹۲۵۳	Polar Res-U-Net++
-	۰/۸۹۶۲	DoubleU-Net
۰/۸۴۳	۰/۹۱۲	BAT
-	۰/۸۸۱۳	MSRF-Net
۰/۸۹۷۶	۰/۹۴۰۸	خروجی مدل این پروژه

۲-۵ تفسیر نقشه‌ها

خروجی‌های حاصل برای مدل و لایه‌ها به ازای مجموعه دادگان KVASIR-SEG و HAM10000 به ترتیب در شکل‌های ۱-۵، ۲-۵، ۳-۵، ۴-۵ و ۵-۶ آمده است. با توجه به شکل ۱-۵ می‌توان فهمید نقاط مرزی توده اثر مثبت در یادگیری مدل و نقاط داخلی توده اثر منفی دارند. دلیل این موضوع می‌تواند شباهت سطح توده و سطح روده باشد و بنابراین عملاً مدل برای تشخیص توده، در حال یادگیری تشخیص محل‌های دارای برآمدگی در تصویر است.

از شکل ۲-۵ می‌توان فهمید که تنها نقاط مرزی در تصمیم‌گیری مدل مؤثرند که یعنی مدل بر اساس تفاوت رنگ دو قسمت پوست، لکه سرطانی را تشخیص می‌دهد. همچنین اگر توجه کنیم، می‌بینیم که این سمت داخل این خط مرزی قرمز و خارج آن سبز است که بدان معناست که در جایی که تشابه رنگ بین لکه و سطح پوست افزایش پیدا کرده، در تصمیم مدل تأثیر منفی داشته است.

با توجه به شکل‌های ۴-۵ و ۳-۵ می‌توان دریافت که در لایه‌های کدگذار هر چه عمق بیشتر می‌شود، توجه مدل به نقاط مرزی بیشتر شده و این نقاط امتیاز بالاتری گرفته‌اند. در لایه Bottleneck ناحیه اطراف قطعه به شدت تأثیر منفی روی یادگیری مدل دارند. در لایه‌های کدگشا نقاط قطعه تأثیر مثبت دارند و نقاط اطراف آن تأثیر منفی دارند. این تأثیرهای مثبت و منفی تا لایه کدگشای Decoder2 به ترتیب روند صعودی و نزولی دارند و پس از این لایه این ترتیب عکس می‌شود. بنابراین می‌توان استدلال کرد که در قسمت کدگشای شبکه، لایه دوم آن یعنی Decoder2 بیشترین تأثیر را در یادگیری مدل دارد.

از مقایسه شکل‌های ۴-۵ و ۳-۵ با شکل‌های ۵-۴ و ۵-۶ می‌توان فهمید که روش DeepLift

الزاماً تأثير همه ویژگی‌ها را نشان نمی‌دهد. همان‌گونه که در مجموعه دادگان اول ویژگی‌های منفی را به کلی نشان نداده است و در مجموعه دادگان دوم نیز ویژگی‌های منفی محدود به ویژگی‌های نزدیک مرز بین لکه و سطح پوست است.

۳-۵ ارزیابی آسیب‌پذیری روش‌های تفسیرپذیری

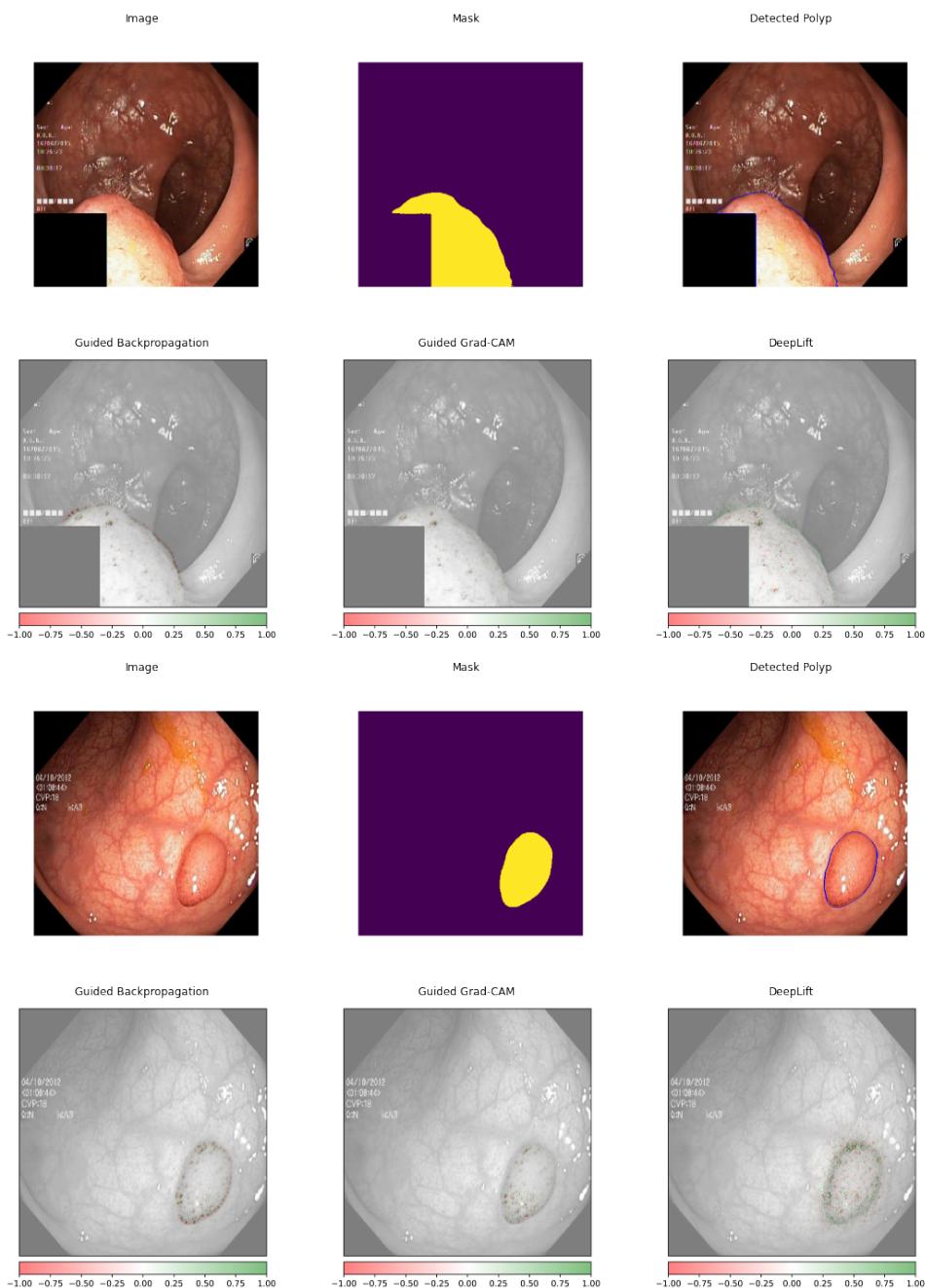
همان‌گونه که پیشتر در فصل مفاهیم اولیه گفته شد، برای ارزیابی آسیب‌پذیری از معیار عدم صحت استفاده می‌کنیم. نتایج این ارزیابی برای مدل و لایه‌ها به ترتیب در جدول‌های ۳-۵ و ۴-۵ آمده است. با توجه به این نتایج می‌توان فهمید که روش‌های Grad-CAM و Guided Grad-CAM بهترین روش‌ها از نظر آسیب‌پذیر نبودن نقشه تفسیرپذیری در برابر آشتفتگی‌اند و نتایج به دست آمده در مورد مدل و لایه‌ها از این دو روش در قسمت قبل قابل اعتمادتر است.

جدول ۳-۵: ارزیابی آسیب‌پذیری روش‌های تفسیر مدل با معیار عدم صحت

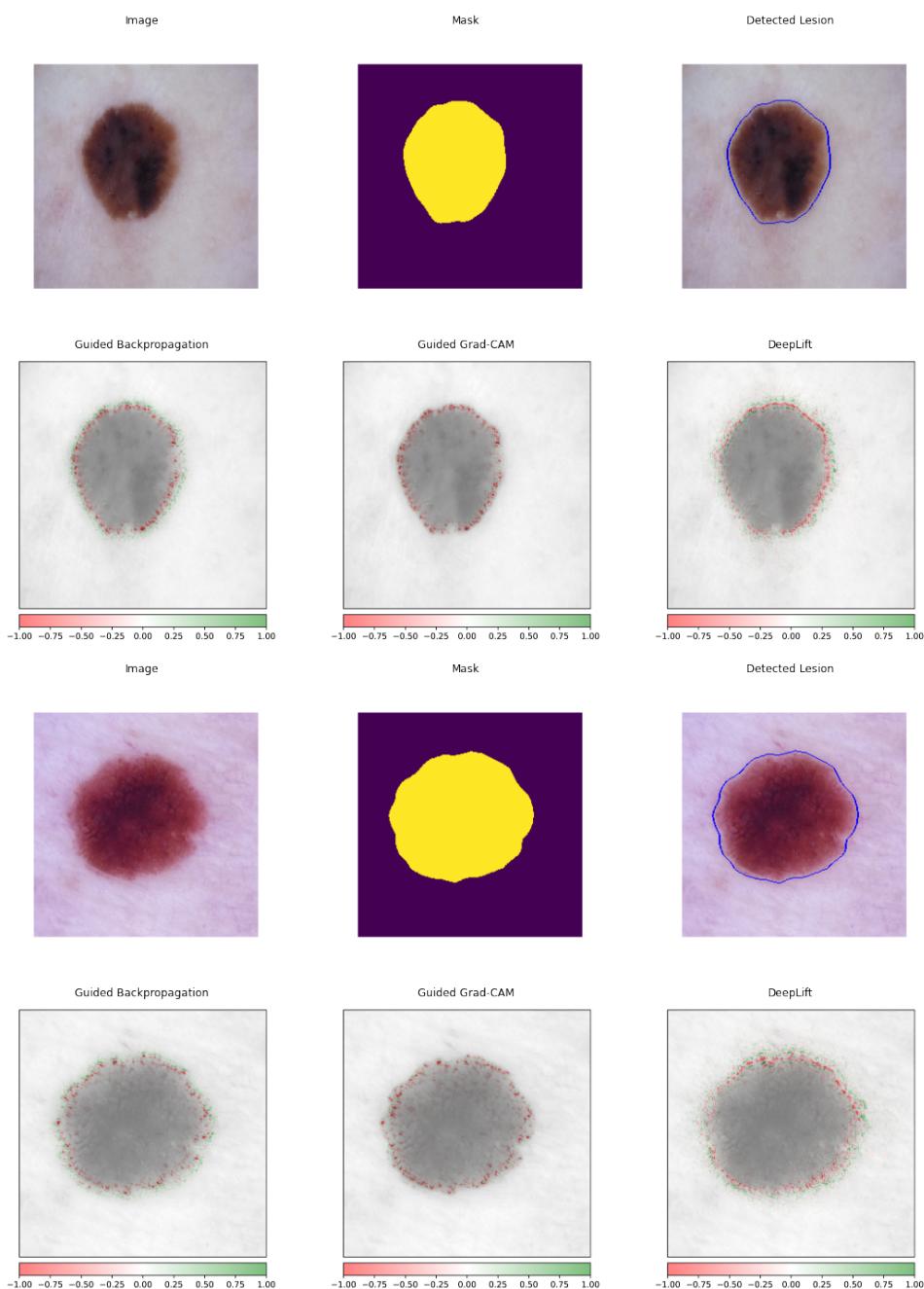
عدم صحت		روش تفسیرپذیری
HAM10000	KVASIR-SEG	
۰/۰۱۲۱	۰/۲۴۱۵	Guided Backpropagation
۰/۰۰۶۸	۰/۰۲۶۵	Guided Grad-CAM
۰/۰۱۹۱	۰/۰۹۰۷	DeepLift

جدول ۴-۵: ارزیابی آسیب‌پذیری روش‌های تفسیر لایه با معیار عدم صحت

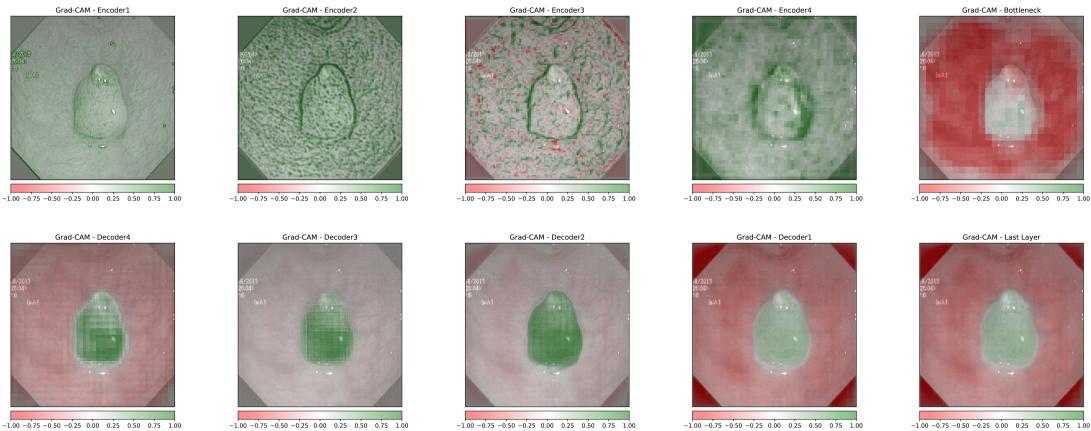
عدم صحت		روش تفسیرپذیری
HAM10000	KVASIR-SEG	
۰/۲۵۰۷	۰/۳۶۰۶	Grad-CAM
۰/۲۷۰۴	۲/۴۵۷۶	DeepLift



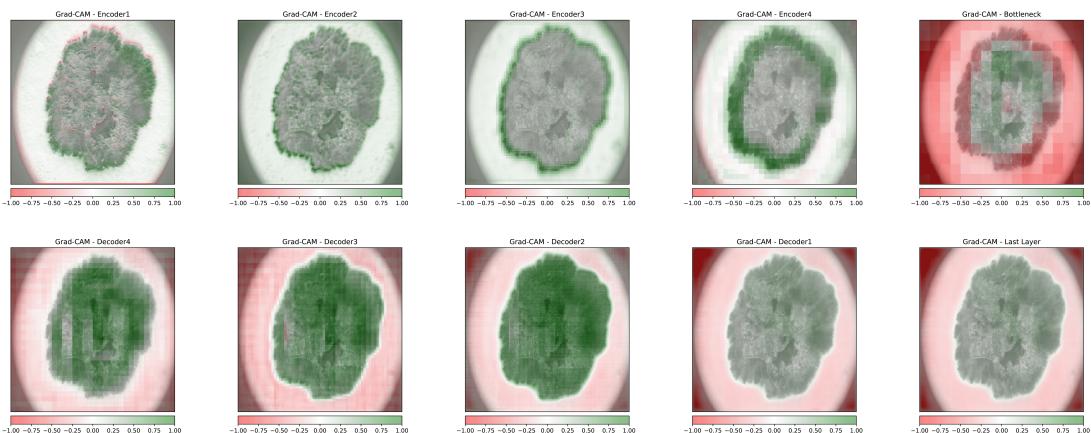
شکل ۵-۱: در سطرهای اول و سوم به ترتیب از چپ به راست، داده ورودی، ماسک پیش‌بینی شده و توده تشخیص داده شده آمده است. در سطرهای دوم و چهارم به ترتیب از چپ به راست، خروجی روشهای DeepLift Guided Grad-CAM، Guided Backpropagation برای ماسک پیش‌بینی شده در سطرهای اول و سوم آمده است.



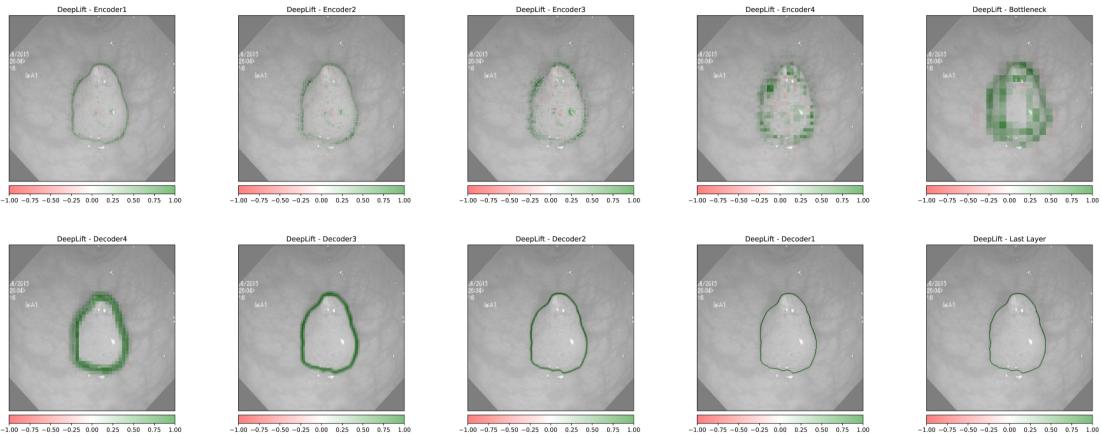
شکل ۵-۲: در سطرهای اول و سوم به ترتیب از چپ به راست، داده ورودی، ماسک پیش‌بینی شده و توده تشخیص داده شده آمده است. در سطرهای دوم، و چهارم به ترتیب از چپ به راست، خروجی روش‌های DeepLift Guided Grad-CAM، Guided Backpropagation برای ماسک پیش‌بینی شده در سطرهای اول و سوم آمده است.



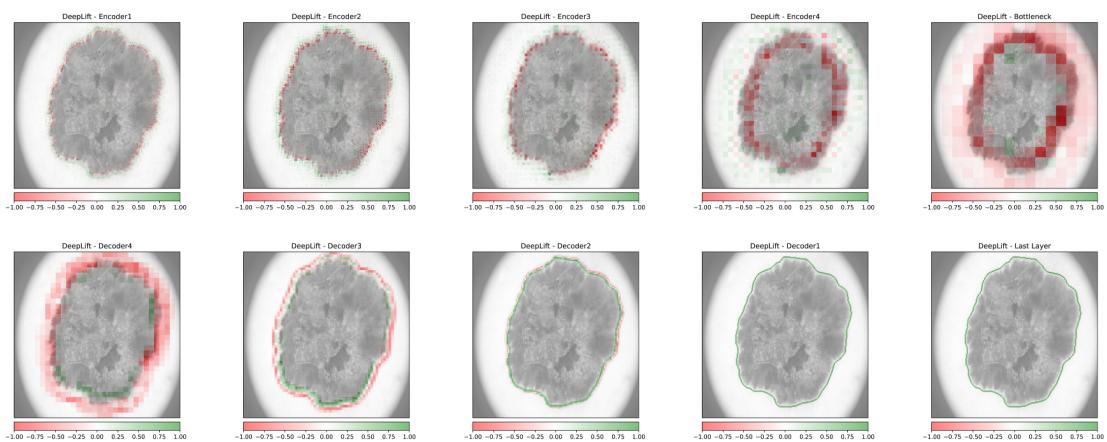
شکل ۳-۵: نمونه‌ای از نقشه‌های Grad-CAM لایه‌ها برای مجموعه دادگان KVASIR-SEG



شکل ۴: نمونه‌ای از نقشه‌های Grad-CAM لایه‌ها برای مجموعه دادگان HAM10000



شکل ۵-۵: نمونه‌ای از نقشه‌های DeepLift لایه‌ها برای مجموعه دادگان KVASIR-SEG



شکل ۵-۶: نمونه‌ای از نقشه‌های DeepLift لایه‌ها برای مجموعه دادگان HAM10000

فصل ۶

جمع‌بندی و نگاه به آینده

در طی مراحل این پایان‌نامه، شبکه U-Net را برای قطعه‌بندی توده‌های روده بزرگ و لکه‌های سرطانی پوست آموزش دادیم، با تعمیم روش‌های تفسیرپذیری مدل‌های دسته‌بندی، نقشه‌های تفسیرپذیری برای مدل و لایه‌ها آن به دست آوردیم، سپس معیارهای تصمیم‌گیری مدل را در مورد هر مجموعه دادگان با استفاده از آن‌ها یافتیم و رفتار مدل را حد خوبی فهمیدیم. در راستای این کار نقشه‌های حاصل از Grad-CAM بیشترین اطلاعات را فراهم کردند. در نهایت بررسی کردیم که این نقشه‌ها چقدر در مقابل آشفتگی‌های معنادار آسیب‌پذیرند و نشان دادیم خروجی‌های Grad-CAM و Guided Grad-CAM بهترین مقاومت را دارند و در نتیجه تفسیرهای حاصل از آن‌ها قابل اعتمادتر است.

تعمیم روش‌های تفسیرپذیری مدل‌های دسته‌بندی به صورت نقطه تصویری به مدل‌های قطعه‌بندی، نیازمند محاسبات زیاد برای محاسبه نقشه‌ها و ارزیابی آن‌هاست و نیاز به زمان و منابع سخت‌افزاری زیاد دارد. بنابراین نیاز است که برای معادل کردن مسئله قطعه‌بندی به مسئله دسته‌بندی از تعداد کمتری نفطه تصویر استفاده شود که یعنی به معادل‌سازی سطح بالاتری نیاز داریم و یا آنکه باید روش‌های تفسیرپذیری برای مدل‌های قطعه‌بندی ارائه شود که محاسبات نقشه‌ها مبتنی بر محاسبات نقشه‌ها برای نقاط تصویر نباشد.

اگر چه در راستای تفسیرپذیری مدل‌های قطعه‌بندی به صورت جعبه سفید^۱ مانند مدل‌هایی [۲۱] SAUnet ارائه شده‌است، اما در در حالتی که به مدل به صورت جعبه سیاه نگاه شود، تاکنون روشنی مخصوص قطعه‌بندی ارائه نشده و همه تلاش‌ها برای تعمیم روش‌های تفسیرپذیری دسته‌بندی به قطعه‌بندی با

White Box^۱

فصل ۶. جمع‌بندی و نگاه به آینده

معادل‌سازی‌های متفاوت مسئله قطعه‌بندی به مسئله دسته‌بندی بوده است. ارائه روش تفسیرپذیری خاص مسئله قطعه‌بندی می‌تواند گامی بزرگ در مسئله تفسیرپذیری شبکه‌های عصبی باشد.

همان‌گونه که پیشتر گفته شد، بحث تفسیرپذیری در کاربردهای پزشکی از اهمیت بزرگی برخوردار است اما تاکنون توجه به مسئله آسیب‌پذیری نقشه‌های حاصل نشده است. در حالی که تغییرات بسیار کوچک در تصویر ورودی می‌توانند تغییرات زیادی در نقشه‌ها خروجی ایجاد کنند. تمرکز بر این موضوع می‌تواند مقبولیت استفاده از مدل‌های یادگیری ماشین را در کاربردهای حساس مانند کاربردهای پزشکی بالا بیرد.

روش‌های تفسیرپذیری بر پایه مفهوم و روش‌های مبتنی بر داده پتانسیل استفاده در کاربردهای پزشکی را دارند و می‌توانند تفسیرهای خوبی را در اختیار بگذارند اما این روش‌ها تاکنون در مورد مدل‌های قطعه‌بندی به کار نرفته‌اند.^[۱۴] در آینده می‌توان از این ایده‌ها استفاده کرد.

مراجع

- [1] S. Chatterjee, A. Das, C. Mandal, B. Mukhopadhyay, M. B. Vipinraj Bhandari, A. Shukla, O. Speck, and A. Nürnberg. Interpretability techniques for deep learning based segmentation models. 2021.
- [2] C. Molnar. *Interpretable Machine Learning*. 2019.
- [3] A. Ghorbani, A. Abid, and J. Zou. Interpretation of neural networks is fragile. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:3681–3688, 2019.
- [4] B. Kim, M. Wattenberg, J. Gilmer, C. J. Cai, J. Wexler, F. B. Viégas, and R. Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *ICML*, 2018.
- [5] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [6] C.-K. Yeh, C.-Y. Hsieh, A. S. Suggala, D. I. Inouye, and P. Ravikumar. On the (in)fidelity and sensitivity of explanations. In *NeurIPS*, 2019.
- [7] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2014.
- [8] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014.
- [9] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller. Striving for simplicity: The all convolutional net. *CoRR*, abs/1412.6806, 2015.

-
- [10] B. Zhou, A. Khosla, Á. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016.
 - [11] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128:336–359, 2019.
 - [12] A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. In *ICML*, 2017.
 - [13] A. Janik, J. D. Dodd, G. Ifrim, K. Sankaran, and K. M. Curran. Interpretability of a deep learning model in the application of cardiac mri segmentation with an acdc challenge dataset. In *Medical Imaging*, 2021.
 - [14] Z. Salahuddin, H. C. Woodruff, A. Chatterjee, and P. Lambin. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in Biology and Medicine*, 140:105111, 2022.
 - [15] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson. Cap-tum: A unified and generic model interpretability library for pytorch, 2020.
 - [16] P. Yakubovskiy. Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch, 2020.
 - [17] D. Jha, P. H. Smedsrud, M. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen. Kvasir-seg: A segmented polyp dataset. *ArXiv*, abs/1911.07069, 2020.
 - [18] N. C. F. Codella, V. M. Rotemberg, P. Tschandl, M. E. Celebi, S. W. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. A. Marchetti, H. Kittler, and A. C. Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *ArXiv*, abs/1902.03368, 2019.
 - [19] P. Tschandl, C. Rosendahl, and H. Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5, 2018.

- [20] M. Buda, A. Saha, and M. A. Mazurowski. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Computers in Biology and Medicine*, 109, 2019.
- [21] J. Sun, F. Darbeha, M. Zaidi, and B. Wang. Saunet: Shape attentive u-net for interpretable medical image segmentation. *ArXiv*, abs/2001.07645, 2020.

واژه‌نامه

الف

ت

تابع فعال‌سازی	Activation Function
تابع هزینه	Loss Function.....
tribution	Attribution
فسیرپذیری	Interpretability
نسور	Tensor
پوده	Polyp.....

ج

جعبه سیاه	Black Box
جعبه سفید	White Box.....

ح

حساسیت مفهومی	Conceptual Sensitivity.....
---------------------	-----------------------------

خ

خطایابی	Debug.....
---------------	------------

الف

آسیب‌پذیر	Fragile
آشتفتگی	Perturbation
آگمنشن	Augmentation.....
آنتروپی متقاطع دودویی ..	Binary Cross Entropy ..
ابر نقطه تصویر	Super-Pixel
اتصالات ردشونده	Skip Connections
ادغام میانگین سراسری ..	Global Average Pooling ..
اریبی ..	Bias
اعتبارسنجی	Validation.....

ب

بردار فعال‌سازی مفهوم	Concept Activation Vector
بی‌طرفی	Fairness

پ

پایایی	Reliability
پسانشار	Backpropagation
پیش‌آموزش‌داده شده	Pretrained

ک	Encoder کدگذار	دسته دسته
	Decoder کدگشا	Class دسته‌بندی
گ		Classification دسته‌بندی
		Epoch دوره
Forward Pass گذر جلو		
Backward Pass گذر عقب		س
Batch گروه		Recommender System سیستم پیشنهاددهنده
ل		
Layer لایه		شبکه عصبی Neural Network
Pooling Layer لایه ادغام		شبکه عصبی پیچشی Convolutional Neural Network
Softmax Layer لایه بیشینه‌نرم		شبکه عکس‌پیچشی Deconvolutional Network
Convolutional Layer لایه پیچشی		
Fully Connected Layer لایه کاملاً همبند		
م		ع
Mask ماسک		عدم صحت Infidelity
Self-driving Car ماشین خودران		عدم قطعیت Uncertainty
Dataset مجموعه دادگان		عکس‌پیچش Deconvolutional
Gaussian Blur محو گاوسی		عکس ادغام Unpooling
		علیت Causality
ن		
Class Activation Map نقشه فعال‌سازی دسته		قطعه‌بندی Segmentation
Feature Map نقشه ویژگی		قطعه‌بندی نمونه‌ای Instance Segmentation
Pixel نقطه تصویر		

ی	Gaussian Noise	نویز گاوسی
	Robustness.....	نیرومندی
Machine Learning	یادگیری ماشین	
Deep Learning	یادگیری عمیق	و
Rectifying Linear Unit...	یکسوساز خطی واحد ...	وارون افقی
		Horizontal Flip.....
		وارون عمودی
		Vertical Flip.....

Abstract

Recent advances in imaging and computation have led to a drastic rise in the use of machine learning for medical imaging. The advent of deep learning allows for much higher levels of abstraction for feature selection and discretization. Convolutional neural networks have been shown to learn abstractions obtained from multidimensional medical images and learning features hard to define by humans. This is one of the reasons why convolutional neural networks excel at object recognition and segmentation. However, most of these networks lack interpretability. Interpretability of deep learning models answers the question as to why a neural network model provides a particular output. In medical image analysis, this is an essential matter in order to build confidence while using. For classification models, there are multiple interpretability techniques; however, there is not much research done on the interpretability of segmentation models. In this project, we trained the U-Net model to segment polyps in colonoscopy images and skin cancer lesions. We applied various interpretability techniques to this model and detected important features that the network's decision is based on. Then, using interpretability evaluation metrics, we showed that Grad-CAM has the most reliable outputs between the applied methods.

Keywords: U-Net Model, Segmentation, Interpretability, Deep Learning, Computer Vision



Sharif University of Technology
Department of Computer Engineering

B.Sc. Thesis

Interpretability of U-Net Model in the Segmentation of Medical Images

By:

Zahra Fazel

Supervisor:

Dr. Hamid Reza Rabiee

February 2022