دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

پایان‌نامه کارشناسی ارشد

تشخیص ناهنجاری با استفاده از شبکه‌های مولد تقابلی

دانشجو

زهرا دهقانیان

اساتید راهنما

دکتر محمد رحمتی

دکتر مریم امیرمزلقانی

بهار ۱۴۰۱


بسم الله الرحمن الرحیم

به نام خدا

تاریخ:

تعهدنامه اصالت اثر

صفحه فرم ارزیابی و تصویب پایان نامه- فرم تأیید اعضاء کمیته دفاع

در این صفحه (هر سه مقطع تحصیلی) باید فرم ارزیابی یا تایید و تصویب پایان‌نامه/رساله موسوم به فرم کمیته دفاع برای ارشد و دکترا و فرم تصویب برای کارشناسی، موجود در پرونده آموزشی را قرار دهند.

اینجانب زهرا دهقانیان متعهد می‌شوم که مطالب مندرج در این پایان نامه حاصل کار پژوهشی اینجانب تحت نظارت و راهنمایی اساتید دانشگاه صنعتی امیرکبیر بوده و به دستاوردهای دیگران که در این پژوهش از آنها استفاده شده است مطابق مقررات و روال متعارف ارجاع و در فهرست منابع و مآخذ ذکر گردیده است. این پایان نامه قبلاً برای احراز هیچ مدرک هم‌سطح یا بالاتر ارائه نگردیده است.

در صورت اثبات تخلف در هر زمان، مدرک تحصیلی صادر شده توسط دانشگاه از درجه اعتبار ساقط بوده و دانشگاه حق پیگیری قانونی خواهد داشت.

در صفحه تعهدنامه اصالت اثر، در قسمت بالا سمت چپ، تاریخ دفاع خود را جایگزین تاریخ نوشته شده کنید.

همچنین در صفحه تعهدنامه اصالت اثر، در خط اول، نام و نام خانوادگی خود را به صورت کامل با نام و نام خانوادگی نمونه، جایگزین کنید. در انتهای متن تعهد، در قسمت امضا نیز باید نام و نام خانوادگی کامل خود را وارد نماید.

چکیده

یکی از مهم‌ترین فعالیت‌های حوزه تحلیل داده تشخیص ناهنجاری است که در طیف وسیعی از کاربردها همچون تشخیص جعل، کاربردهای پزشکی و سیستم‌های امنیتی به کار گرفته می‌شود. علی‌رغم وجود روش‌های آماری و مبتنی بر یادگیری ماشین، طراحی مدل‌های موثر در تشخیص ناهنجاری در فضای داده پیچیده با ابعاد بالا همچنان به عنوان یک چالش اساسی باقی مانده است. شبکه‌های مولد تقابلی[1] قادرند تا بر چالش مورد نظر فائق آمده و توزیع داده‌های دنیای واقعی که دارای پیچیدگی و ابعاد بالا هستند را مدل کنند و همین امر سبب می‌شود تا عملکرد امیدوارکننده‌ای در زمینه تشخیص ناهنجاری از خود نشان دهند. در این کار سه شبکه تقابلیCALAD [2]، RALAD[3] و RCALAD[4] با هدف تشخیص ناهنجاری ارائه شده است. اساس کار هر سه مدل پیشنهادی بازسازی داده ورودی با استفاده از شبکه مولد و در ادامه محاسبه میزان اختلاف داده اصلی و بازسازی آن به منظور شناسایی نمونه‌های ناهنجار است. در مدل CALAD با تعریف متغیر جدید $\hat{\boldsymbol{z}}_{\hat{\boldsymbol{x}}}$ و افزودن تمایزگر ابتکاری $\boldsymbol{D}_{\boldsymbol{xxzz}}$ چرخه پایداری کامل میان هر دو فضای ورودی و فضای نهان برقرار می‌شود. لازمه شناسایی موثر نمونه‌های ناهنجار بازسازی ضعیف آن‌ها است. هدف از چارچوب پیشنهادی RALAD متمایل کردن تمامی بازسازی‌ها به سمت توزیع داده هنجار است. رویه مورد نظر در این مدل سبب بازسازی ضعیف نمونه‌ها ناهنجار و در نتیجه ایجاد فاصله مناسب میان داده ورودی و بازسازی متناسب با آن می‌شود. در نهایت از ترکیب هر دو ایده مدل جامع RCALAD با هدف حل مسئله تشخیص ناهنجاری در دنیای واقعی ارائه شده است. علاوه بر معماری پیشنهادی، دو امتیاز ناهنجاری جدید نیز در این کار معرفی شده است که در مقایسه با امتیازهای ناهنجاری قبلی قدرت تفکیک‌پذیری بیشتری فراهم می‌آورد. نتایج تجربی بیانگر برتری مدل‌های پیشنهادی در مقایسه با سایر مدل‌های مطرح در زمینه تشخیص ناهنجاری بوده است.

واژگان کلیدی:

تشخیص ناهنجاری، یادگیری ماشین، شبکه مولد تقابلی، چرخه پایداری کامل، امتیاز ناهنجاری.

[1] Generative Adversarially Networks

[2] Complete Adversarially Learned Anomaly Detection

[3] Regularized Adversarially Learned Anomaly Detection

[4] Regularized Complete Adversarially Learned Anomaly Detection

فهرست مطالب

صفحه

فهرست شکل‌ها

صفحه

فهرست جدول‌ها

صفحه

فهرست الگوریتم‌ها

صفحه

فهرست علائم

x	داده ورودی	π	آنتروپی شرطی		
G	شبکه مولد	CCC	چرخه پایداری کامل		
D	شبکه تمایزگر	$\hat{x}$	بازسازی x		
p_g	توزیع شبکه مولد	$\hat{z}_{\hat{x}}$	نگاشت معکوس $\hat{x}$		
p_{data}	توزیع داده ورودی	N	توزیع نرمال		
z_x	نگاشت معکوس x	$\hat{z}$	نگاشت معکوس x_z		
E	کدگذار	$\sigma(x)$	توزیع کمکی		
q	توزیع کدگذار	A	امتیاز ناهنجاری		
U	توزیع یکنواخت	z	نمونه ازتوزیع گاوسی		
x_z	خروجی مولد با ورودی توزیع گاوسی				

فصل اول:
مقدمه

هنگام تجزیه و تحلیل دادگان موجود در دنیای واقعی، شناسایی نمونه‌های غیرمشابه با سایر نمونه‌ها امری ضروری به نظر می‌رسد. چنین نمونه‌هایی با عنوان ناهنجاری شناخته می‌شوند و از عملیات شناسایی چنین نمونه‌هایی با عنوان مسئله تشخیص ناهنجاری یاد می‌شود. این مسئله یک بخش حائز اهمیت از زمینه تحقیقاتی داده‌کاوی است چرا که شامل کشف الگوهای جذاب و نادر در داده‌هاست[1]. این مسئله به طور گسترده در آمار و یادگیری ماشین مورد مطالعه قرار گرفته است و با مترادف‌هایی مانند تشخیص داده پرت[۱]، شناسایی نوآوری[۲]، تشخیص انحراف[۳] و استخراج استثنا[۴] نیز یاد می‌شود. تعریف رسمی و مورد قبول این مسئله به صورت زیر است:

"یک ناهنجاری مشاهده‌ای است که به میزانی از سایر مشاهدات منحرف می‌شود که ظن‌هایی را برای این که توسط مکانیسم متفاوتی تولید شده باشد، ایجاد می‌کند[2]."

ناهنجاری‌ها جزو پارامترهای مهم هر مجموعه داده‌ای در نظر گرفته می‌شوند و در دامنه وسیعی از کاربردها تاثیرگذار هستند. به عنوان مثال، الگوی غیر معمول ترافیک در یک شبکه کامپیوتری میتواند به معنای هک شدن رایانه و انتقال داده‌ها به مقصدهای غیرمجاز باشد. رفتار غیر عادی در معاملاتی که توسط کارت‌های اعتباری انجام می‌شوند می‌تواند نشانگر فعالیت‌های اقتصادی با هدف کلاهبرداری باشد[3]، و یا یک ناهنجاری در تصویر MRI ممکن است وجود تومور بدخیم را نشان دهد[4]. تشخیص ناهنجاری به طور گسترده در زمینه‌های کاربردی گوناگونی مانند: پزشکی، بهداشت عمومی، تشخیص کلاهبرداری، سنجش از دور[۵]، تشخیص نفوذ، پردازش تصویر، آسیب‌های صنعتی، شبکه‌های حسگر[5]، رفتار روبات‌ها و داده‌های نجومی به‌کار گرفته شده است[3].

داده‌های تمامی این مسائل یا از نوع سری زمانی هستند و یا فارغ از زمان می‌باشند. رویکرد مورد استفاده برای حل مسائل مربوط به این دو جنس داده کاملا متفاوت از یکدیگر هستند. تحقیقات بسیار گسترده‌ای

[1] Outlier detection

[2] Novelty detection

[3] Deviation detection

[4] Exception mining

[5] Remote sensing

در حوزه داده‌های سری زمانی صورت گرفته است. به عنوان مثال در مدل رگرسیون[۶] نمونه‌ای که به مقدار زیادی از مدل تعیین شده منحرف شود به عنوان داده ناهنجار شناخته می‌شود[۶]. در روش دیگر با استفاده از مدل ARIMA[۷] مقدار آینده را پیش‌بینی می‌کنند و با محاسبه میزان اختلاف داده پیش‌بینی شده و مقدار واقعی داده به شناسایی نمونه ناهنجار می‌پردازند[۷]. علاوه بر این با ظهور CNN[۸] و RNN[۹] در زمینه پیش‌بینی سری زمانی و ثابت شدن کارایی آن‌ها، توجه‌ها به سمت استفاده از این ساختارها در زمینه تشخیص ناهنجاری جلب شد، نحوه استفاده از این نوع از شبکه‌های عصبی همانند مدل رگرسیون و ARIMA است[۷]. در روش‌های آماری موجود برای مسئله تشخیص ناهنجاری فاصله داده تا توزیع یا مدل طراحی شده محاسبه می‌شود و در صورتی که از یک حد آستانه بیشتر باشد به عنوان ناهنجاری شناخته می‌شود. در دسته دیگر از روش‌های آماری هر نمونه به عنوان یک نقطه در فضای nبعدی در نظر گرفته می‌شود و حول تمامی نمونه‌ها یک فضای محدب محاسبه می‌کند و فرض می‌کند نمونه‌های ناهنجار در لبه این فضا قرار می‌گیرد[۸]. در روش‌های مبتنی بر فاصله با استفاده از یک معیار فاصله میزان تفاوت نمونه تا سایر داده‌ها سنجیده می‌شود و در صورتی که از یک حد آستانه بیشتر باشد به عنوان ناهنجاری شناخته می‌شود. روش دیگر محاسبه چگالی یک نمونه و همسایه‌های آن است تا بدین وسیله یه معیار جدید برای شناسایی نمونه‌های ناهنجار ایجاد شود. این معیار LOF[۱۰] نام دارد و هرچه این معیار بالاتر باشد، احتمال ناهنجار بودن آن نمونه بیشتر است[۹]. به طور خاص داده‌های مورد بررسی در این تحقیق همگی از جنس فارغ از زمان هستند که در فصل دوم به بررسی دسته‌بندی‌های موجود برای حل آن‌ها خواهیم پرداخت.

با هدف تشخیص نمونه‌های ناهنجار موجود در دادگان دنیای واقعی، تاکنون روش‌های متنوعی مورد استفاده قرار گرفته‌اند. به طور کلی تشخیص ناهنجاری بر دو اصل استوار است، شناسایی و مدل کردن رفتار داده هنجار و میزان انحراف از رفتار داده هنجار[10]. به طور کلی می‌توان روش‌های تشخیص

[6] Regression

[7] Autoregressive integrated moving average

[8] Convolutional Neural Networks

[9] Recurrent Neural Networks

[10] Local Outlier Factor

ناهنجاری را به دوسته روش‌های آماری و روش‌های مبتنی بر یادگیری ماشین اشاره کرد. روش‌های آماری اگرچه در برخی از موارد کارایی مناسبی دارند اما عملکرد صحیح و مناسب آن‌ها در گرو صحت پیش‌فرض‌های استفاده شده در همین روش‌هاست و در صورتی که پیش‌فرض‌های اولیه در مورد توزیع داده اشتباه باشد نتایج نهایی ناامید کننده خواهد بود[11]. مزیت اصلی روش‌های مبتنی بر یادگیری ماشین استفاده از تجربه‌های گذشته به منظور انجام پیش‌بینی‌های صحیح در آینده است. این روش‌ها تنها با مشاهده نمونه‌های گذشته به طراحی مدل می‌پردازند و پیش‌فرض خاصی نسبت به توزیع داده ندارند. این دسته از الگوریتم‌ها نیازمند تعداد مناسبی از داده‌ها هستند تا بتوانند مدل پیشنهادی خود را آموزش دهند. از جمله مهم‌ترین مدل‌های موجود در این دسته می‌توان به شبکه‌های عصبی اشاره کرد که سابقه طولانی در زمینه تشخیص ناهنجاری دارد. به عنوان مثال شبکه‌های عصبی کدگذار[11] و خودکدگذار[12] مدلی برای بازسازی داده‌های عادی آموزشی آموزش داده می‌شود و نمونه‌های با خطای بازسازی بالا به عنوان نمونه ناهنجار در نظر گرفته می‌شوند[12].

در سال ۲۰۱۷ از شبکه‌های عصبی مولد تقابلی برای تشخیص ناهنجاری در زمینه تصاویر پزشکی (تصاویر شبکیه) و در مقایسه با سایر روش‌ها به موفقیت قابل توجهی دست یافت[13]. نتایج درخشان شبکه‌های مولد تقابلی در عرصه پردازش تصویر و استخراج ویژگی سبب محبوبیت آن در زمینه‌های کاربردی مختلف شده است. به طور خاص این دسته از شبکه‌ها به عنوان یک چهارچوب قدرتمند برای مدل‌سازی مجموعه داده‌های پیچیده با ابعاد بالا شناخته می‌شوند. یک از اصلی‌ترین چالش‌های موجود در استفاده از شبکه‌های مولد تقابلی مقابله با پیچیدگی‌های استنتاج است[14]. در سال‌های اخیر تلاش‌های گسترده‌ای انجام شده است تا با استفاده از شبکه‌های عصبی خود کدگذار در ساختار شبکه‌های عصبی مولد تقابلی از پیچیدگی‌های استنتاج کاسته شده و بر چالش‌های موجود غلبه کنند. علیرغم این تلاش‌ها همچنان ضعف‌هایی در روند یادگیری بلوک‌های موجود در ساختار شبکه‌های مولد تقابلی موجود است و از تمامی ظرفیت موجود به منظور دریافت اطلاعات و آموزش هر چه بهتر مدل استفاده نمی‌شود.

[11] Encoder

[12] Autoencoder

۱-۱- ساختار گزارش

در فصل بعدی ابتدا به دسته‌بندی روش‌های تشخیص ناهنجاری از دیدگاه‌های مختلف می‌پردازیم، معیارهای ارزیابی مدل‌های تشخیص ناهنجاری را معرفی می‌کنیم و در قسمت انتهایی فصل روش‌های تشخیص ناهنجاری مبتنی بر شبکه‌های مولد تقابلی را بررسی و مرور می‌کنیم. در این قسمت سعی می‌شود تا ضمن دسته‌بندی روش‌های تشخیص ناهنجاری، مروری گذرا بر روش‌های به نسبت قدیمی‌تر نیز انجام شود. در ادامه به طور دقیق‌تر زنجیره‌ای از کارها مورد بحث قرار خواهد گرفت که در طول این زنجیره نقاط ضعف و کمبودهای مدل پیشنهادی بر طرف می‌شود. فصل سوم به معرفی مدل پیشنهادی و روش نوین تشخیص ناهنجاری اختصاص خواهد داشت. در این فصل الگوریتم پیشنهادی، که بر اساس حل یک مسئله بهینه‌سازی با در نظر گرفتن توزیع توام[۱۳] پارامترهای موجود در ساختار شبکه عصبی مولد تقابلی طراحی شده است، به شناسایی نمونه‌های ناهنجار در فضای نهفته[۱۴] و فضای ورودی می‌پردازد. در فصل چهارم عملکرد مدل روی دادگان‌های مختلف آزمایش می‌شود. در این فصل با دو نوع مختلف از دادگان روبرو خواهیم بود، دادگانی جدولی[۱۵] که شامل دادگان KDD و ARRHYTHMIA و دادگان تصاویر که شامل CIFAR-10 و SVHN[۱۶] است. در فصل پنجم و نهایی این گزارش پیشنهادهای موثر به منظور اصلاح و بهبود احتمالی روند آموزش شبکه مولد تقابلی ارائه می‌شود و در گام آخر جمع‌بندی مطالب ارائه شده و نتیجه‌گیری نهایی صورت خواهد پذیرفت.

[13] Joint distribution

[14] Latent space

[15] Tabular

[16] Street view house number

فصل دوم:
مروری بر کارهای پیشین

در این فصل ابتدا به دسته‌بندی روش‌های مختلف تشخیص ناهنجاری و مرور روش‌های شاخص هر دسته پرداخته خواهد شد. در گام بعدی معیارهای ارزیابی مدل‌های تشخیص ناهنجاری معرفی می‌شوند. در ادامه بر روی کارهایی که تاکنون در زمینه تشخیص ناهنجاری با استفاده از شبکه‌های مولد تقابلی انجام گرفته‌اند مروری خواهیم داشت. در گام اول این قسمت، مقاله پایه با عنوان شبکه‌های مولد تقابلی مورد بررسی قرار خواهد گرفت در ادامه، کار تشخیص ناهنجاری بدون نظارت با شبکه‌های عصبی تقابلی به منظور راهنمایی عملیات اکتشاف نشانگر[1] به اختصار AnoGan شرح داده خواهد شد و پس از آن f-AnoGan[2] که در ادامه کار قبلی و توسط همان نویسندگان انجام شده است بررسی خواهد شد. با توجه به ضعف‌های موجود در ساختار f-AnoGan مقاله مکمل این مدل با نام استنتاج یادگرفته شده به روش تقابلی[3] به اختصار ALI مرور خواهد شد. . در گام بعدی مدل EGBAD[4] که جزو اولین کارها در زمینه تشخیص ناهنجاری که با الهام از مدل ALI خلق شده است مرور می‌شود. در مرحله بعدی با توجه تضمین نشدن شرط سازگاری حلقه[5] در ALI مقاله آلیس[6] مورد اشاره قرار می‌گیرد و در انتها تشخیص ناهنجاری یادگرفته شده به روش تقابلی[7] که در ادامه کارهای پیشین و همچنین مقاله پایه در این پروژه است به طور دقیق بررسی خواهد شد.

۲-۱- طبقه‌بندی روش‌های تشخیص ناهنجاری از دیدگاه در دسترس بودن برچسب داده

بیشتر روش‌های تشخیص ناهنجاری در مرحله آموزش خود نیاز دارند تا برچسب نمونه‌های مختلف در دسترس آن‌ها باشد تا بتوانند در خصوص طبیعی یا ناهنجار بودن یک نمونه در مرحله آزمایش

[1] Unsupervised anomaly detection with generative adversarial networks to guid marker

[2] Fast unsupervised anomaly detection with generative adversarial networks

[3] Adversarially Learned Inference

[4] Efficient GAN-Based Anomaly Detection

[5] Cycle consistency

[6] ALICE

[7] Adversarially Learned Anomaly Detection

تصمیم‌گیری کنند[15]. فرایند تهیه و دستیابی به داده‌های دارای برچسب دقیق شامل طیف گسترده‌ای از عملیات‌های بسیار هزینه‌بر و دشوار است، از این‌رو تکنیک‌های تشخیص ناهنجاری را بر اساس میزان در دسترس بودن برچسب‌ها می‌توان به سه دسته: تشخیص ناهنجاری با نظارت، تشخیص ناهنجاری نیمه نظارتی و تشخیص ناهنجاری بدون نظارت تقسیم کرد[16].

۲-۱-۱- تشخیص ناهنجاری با نظارت

در این دسته از روش‌ها هر دو الگوری رفتاری غیرطبیعی و طبیعی مدل می‌شوند. در این مدل‌ها به داده‌های غیر طبیعی برچسب ناهنجاری و به داده‌های طبیعی برچسب عادی می‌زنند. در این رویکرد، برخی از مدل‌ها نمونه‌های ورودی را با نمونه‌های غیرعادی مقایسه می‌کنند و برخی دیگر نمونه‌ها را با نمونه‌های برچسب عادی مقایسه می‌کنند تا بر اساس آن در مورد ماهیت نمونه ورودی تصمیم‌گیری کنند[17].

۲-۱-۲- تشخیص ناهنجاری نیمه‌نظارتی

در تشخیص ناهنجاری نیمه‌نظارتی تنها الگوی رفتار طبیعی داده مدل می‌شود و به بیان دیگر تنها به برچسب‌های کلاس عادی نیاز داریم. از نظر کمی این دسته از روش‌ها کاربرد بیشتری نسبت به روش‌های تشخیص ناهنجاری نظارتی دارند[18].

۲-۱-۳- تشخیص ناهنجاری بدون نظارت

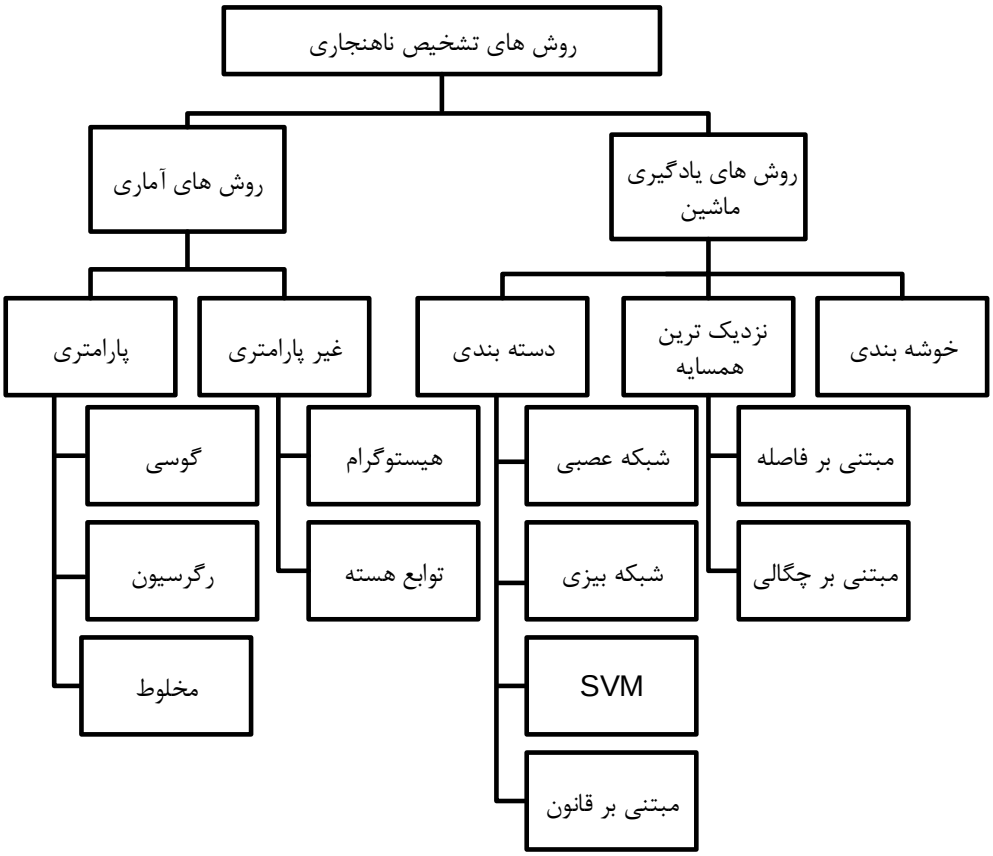
اساس کار این دسته از روش‌ها همانند روش‌های خوشه‌بندی[8] است و مدل کلاس داده‌های ناهنجاری را به صورت خودکار از سایر کلاس‌ها تمیز می‌دهد[19]. این روش خوشه‌ای از داده‌ها با رفتار نزدیک به هم پیدا می‌کند و بدین ترتیب عملیات شناسایی ناهنجاری صورت می‌گیرد. اینگونه از مدل‌ها در بسیاری از تشخیص‌ها دچار مشکل می‌شوند چراکه ممکن است نمونه‌های ناهنجار خود باعث ایجاد خوشه‌هایی با

[8] Clustering

الگوی مشابه داده‌های عادی شوند، به همین دلیل تکنیک‌های بدون نظارت در تولید نتایج دقیق کارآمد نیستند و اغلب دارای نرخ مثبت کاذب[9] هستند[20].

۲-۲- طبقه‌بندی روش‌های تشخیص ناهنجاری از نظر رویکرد حل مسئله

روش‌های تشخیص ناهنجاری به طور کلی به دو دسته روش‌های آماری و روش‌های مبتنی بر یادگیری ماشین تقسیم می‌شوند. روش‌های آماری خود شامل دو دسته پارامتری و غیرپارامتری هستند و روش‌های یادگیری ماشین شامل خوشه‌بندی، نزدیک‌ترین همسایه و دسته‌بندی است. در شکل ۲-۱ دسته‌بندی این روش‌ها به طور دقیق‌تر به تصویر کشیده شده است. در ادامه هر یک از این روش‌ها بررسی و مرور خواهد شد.



شکل ۲-۱: دسته بندی روش‌های تشخیص ناهنجاری.

[9] False positive rate

۲-۲-۱- روش‌های آماری

تشخیص ناهنجاری با روش‌های آماری به ترتیب بر اساس آمارگان‌های آماری مانند میانگین و انحراف از معیار، توزیع داده‌ها و توابع احتمال (برای ساختن نمایه‌های رفتاری) انجام می‌شود[21]. در اینگونه از روش‌ها بر اساس آزمون‌های آماری هر نوع انحراف از رفتار عادی داده‌ها تشخیص داده می‌شود و داده مورد نظر به عنوان ناهنجاری در نظر گرفته می‌شود. به منظور توسعه مدل‌های اماری در تشخیص ناهنجاری از دو نوع تکنیک پارامتری و غیرپارامتری استفاده می‌شود[22]. تکنیک‌های غیرپارامتری از اطلاعات زمینه‌ای داده‌ها استفاده نمی‌کنند، به بیان دیگر اطلاعی از توزیع داده ورودی ندارند در حالی که روش‌های پارامتری با استفاده از همین اطلاعات مدل را طراحی می‌کنند.

۲-۲-۱-۱- روش‌های پارامتری

در روش‌های آماری فرض می‌شود داده‌های واقعی بر اساس پارامترهای مشخص از یک توزیع یا تابع خاص تولید می‌شوند، این دسته از روش‌ها خود به سه دسته کلی مدل رگرسیونی، مدل گاوسی و مدل مخلوط تقسیم می‌شوند.

در مدل رگرسیونی داده‌ها بر یک مدل رگرسیونی منطبق می‌شوند و باقی‌مانده[10] هر داده که بر مدل منطبق نیست به عنوان معیار جهت تشخیص ناهنجاری به کار برده می‌شود.

در مدل گاوسی، فرض بر این است که داده‌ها به توزیع گاوسی تعلق دارند و پارامترهای مدل با استفاده از استفاده از تخمین بیشینه درست‌نمایی[11] تعیین می‌شوند. در این مدل‌ها از آزمون‌هایی نظیر آزمون کای-دو[12] جهت شناسایی نمونه ناهنجار استفاده می‌شود[23].

مدل‌های مخلوط خود ترکیبی از سایر مدل‌های پارامتری هستند. چنین مدل‌هایی در برخی از کاربردها عملکرد بسیار موفقی از خود نشان داده‌اند. به عنوان مثال با به کارگیری یک مدل مخلوط از روش‌های

[10] Residual

[11] Maximum likelihood estimation

[12] Chi-square

پارامتری برای تشخیص ناهنجاری‌های شبکه، توانسته‌اند در طی زمان بسیار کوتاهی تمام ناهنجاری‌های موجود در شبکه که توسط سناریوهای مختلف ایجاد شده بودند را شناسایی کنند[11].

۲-۲-۱-۲- روش‌های غیرپارامتری

در این روش از نمونه‌های عادی برای تولید مدل استفاده می‌شود و انحراف نمونه از مدل به عنوان امتیاز ناهنجاری در نظر گرفته می‌شود. این روش را می‌توان به دو دسته مدل‌های مبتنی بر هیستوگرام و مدل‌های مبتنی بر هسته تقسیم کرد.

در مدل‌های مبتنی بر هیستوگرام، هیستوگرام بر اساس تقریب از داده‌های عادی تولید می‌شود و برای اگر نمونه ورودی در محدوده‌های خاصی از هیستوگرام قرار گیرد به عنوان ناهنجاری شناخته می‌شود[24].

روش مدل‌سازی مبتنی بر هسته[۱۳] یک تابع تشابه بر اساس نمونه‌های موجود از داده استنباط می‌شود[25]. در اینگونه از مدل‌ها در دسترس بودن نمونه‌های کافی به منظور بازنمایی کامل رفتار مجموعه داده ضروریست چراکه در غیر این صورت دقت مدل کاهش می‌یابد.

۲-۲-۲- روش‌های یادگیری ماشین

روش‌های مبتنی بر یادگیری ماشین بر اساس تجربه حاصل از مشاهده نمونه‌های قدیمی و به کار گیری آن، ظرفیت تمایز میان رفتارهای غیرطبیعی و طبیعی داده تا حد مناسبی بهبود می‌بخشند[26]. این طبقه از روش‌ها خود به سه گروه دسته‌بندی، نزدیک‌ترین همسایه و خوشه‌بندی تقسیم می‌شوند[27] ، در ادامه به بررسی هر یک از این دسته‌ها می‌پردازیم.

[13] kernel

۲-۲-۲-۱- دسته‌بندی

هدف اصلی از روش‌های مبتنی بر دسته‌بندی، اختصاص هر نمونه از داده به یکی از کلاس‌های از پیش تعیین شده بر اساس ویژگی‌های آن نمونه است. از مزیت‌های این دسته از روش‌ها می‌توان به توانایی بالای آن‌ها در تمایز میان کلاس‌های مختلف داده در زمان آزمایش اشاره کرد. از روش‌های متداول که در این دسته جای دارند می‌توان به شبکه‌های بیزی، ماشین بردار پشتیبان[۱۴] ، برخی روش‌های مبتنی بر قانون و شبکه‌های عصبی اشاره کرد.

شبکه‌های بیزی در واقع مدل‌های گرافیکی هستند که اتصالات میان نمونه‌های مختلف را بر اساس محاسبه احتمال پیشین[۱۵] یک نمونه از داده به همراه دسته‌ای از پیش‌شروط مورد بررسی و ترجمه قرار می‌دهند. اساس کار این دسته از روش‌ها استفاده از یادگیری با نظارت است.

ماشین‌های بردار پشتیبان از دسته الگوریتم‌های یادگیری با نظارت هستند که در صورت استفاده از هسته نمونه‌ها را فضای با ابعاد بالاتر می‌برند و در فضای جدید نمونه‌ها را به دو کلاس تقسیم می‌کنند. استفاده از هسته زمانی توجیه پذیر است که نمونه‌ها در فضای با ابعاد پایین جداپذیر نباشند. این مدل به دلیل استفاده از یک مرز خطی به منظور جداسازی نمونه‌های غیرطبیعی و عادی به عنوان دسته‌بندی خطی شناخته می‌شود[28].

روش‌های مبتنی بر قانون بر اساس یک سری از قواعد رفتار و عملکرد نمونه‌های عادی را می‌آموزد، بنابراین اگر یک نمونه نتولند از این مجموعه قوانین پیروی کند به عنوان نمونه ناهنجار شناخته خواهد شد. از مطرح‌ترین روش‌هایی که در این دسته می‌گنجند می‌توان به درخت تصمیم[۱۶] اشاره کرد[29].

شبکه‌های عصبی رفتار سیستم عصبی انسان را تقلید می‌کنند و شامل مجموعه‌ای از فرایندهای بهم پیوسته هستند که به طور همزمان روی داده عمل می‌کنند. در این دسته از روش‌ها از نمونه‌های عادی برای آموزش شبکه عصبی استفاده می‌شود. از نظر دسترسی به برچسب نمونه‌ها شبکه‌های عصبی را می‌توان مشترک بین دسته‌های یادگیری با نظارت و بدون نظارت در نظر گرفت[30]. یکی از انواع این

[14] Support vector machine

[15] Posterior probability

[16] Decision tree

شبکه‌ها که در سال‌های اخیر به موفقیت چشم‌گیری دست یافته است، شبکه‌های مولد تقابلی[17] هستند. این دسته از شبکه‌ها با موفقیت بر روی داده‌های دنیای واقعی که دارای ابعاد بالا هستند اعمال شده‌اند و چهارچوب یادگیری خصمانه آن‌ها عملکرد مناسبی از خود بر جای گذاشته است. قابلیت این شبکه‌ها نشان‌دهنده ظرفیت آن‌ها برای استفاده در مسئله تشخیص ناهنجاری می‌باشد و به همین جهت به کارگیری شبکه‌های مولد تقابلی در حوزه تشخیص ناهنجاری اخیرا مورد توجه و کاوش قرار گرفته است[31]. در تشخیص ناهنجاری به کمک شبکه‌های مولد تقابلی با استفاده از فرایند آموزش تقابلی رفتار عادی داده مدل می‌شود سپس با اندازه‌گیری امتیاز ناهنجاری روی نمونه‌های مختلف عمل شناسایی نمونه ناهنجار صورت می‌پذیرد. شبکه‌های مولد تقابلی با کمک آموزش و نمونه‌گیری از مدل‌های مولد به نتایجی بسیار مناسبی در مقایسه با دیگر روش‌ها دست می‌یابند همچنین این مدل‌ها امکان آموزش داده‌های از دست رفته به کمک الگوریتم‌های یادگیری تقویتی[18] را می‌دهد. بیان جزئیات بیشتر در خصوص این دسته از شبکه‌ها را به به بخش‌هایی که در آینده خواهیم داشت موکول می‌کنیم.

۲-۲-۲-۲- نزدیک‌ترین همسایه

روش نزدیک‌ترین همسایه مبتنی بر سنجش فاصله یا تراکم میان داده‌هاست به بیان دیگر امتیاز ناهنجاری مقدار همین فاصله است و بسته به مسئله و میزان در دسترس بودن برچسب‌ها این روش می‌تواند به عنوان روش یادگیری بدون نظارت و یا با نظارت به کار گرفته شود[11].

۲-۲-۲-۳- خوشه‌بندی

روش مبتنی بر خوشه‌بندی از دسته روش‌های یادگیری بدون نظارت است که برای شناسایی مجموعه نمونه‌های شبیه به هم به کار برده می‌شود. ناهنجاری‌ها ممکن است تشکیل یک خوشه کوچک بدهند یا در هیچ خوشه‌ای جای نگیرند. این روش در مقایسه با روش‌های مبتنی بر فاصله از پیچیدگی محاسباتی کمتری برخوردار است و در عین حال نقطه ضعف این روش عملکرد نامناسب روی دادگان کوچک است

[17] Generative Adversarial Networks (GAN)

[18] Reinforcement learninig

چرا که مدل بینش مناسبی نسبت به داده ندارد و به عنوان مثال برای قسمتی از فضا که برای آن نمونه آموزشی نداریم همواره برچسب ناهنجاری در نظر می‌گیرد در صورتی که ممکن است در حضور تعداد داده کافی برچسب آن نمونه خاص برچسب عادی باشد[11].

۲-۳- دسته‌بندی بر اساس نحوه تشخیص ناهنجاری

در این قسمت بر اساس نحوه تشخیص ناهنجاری روش‌های موجود را به دسته بر اساس فاصله، دسته‌بندی تک‌کلاسی و بر اساس بازسازی[19] تقسیم می‌کنیم[32].

۲-۳-۱- بر اساس فاصله

از روش‌های مبتنی بر فاصله می‌توان به عنوان یکی از کلاس‌های اصلی روش‌های تشخیص ناهنجاری یاد کرد. در این روش‌ها با استفاده از محاسبه فاصله یک نمونه خاص با نزدیک‌ترین همسایه‌ها و یا نزدیک‌ترین کلاستر، نمونه ناهنجار شناسایی می‌شود. بدیهی‌ست به کار بردن چنین روش‌هایی نیازمند طراحی یا انتخاب معیار فاصله مناسب است[32].

۲-۳-۲- دسته‌بندی تک‌کلاسی

در رویکرد دسته‌بندی تک‌کلاسی تنها نمونه‌های عادی به یک دسته‌بند نظیر SVM آموزش داده می‌شوند[33]، در واقع این مدل‌ها یک مرز تصمیم حول نمونه‌های عادی یاد می‌گیرند، در صورتی که نمونه ورودی داخل این مرز قرار گیرد به عنوان نمونه عادی و در غیر اینصورت به عنوان نمونه ناهنجار شناخته می‌شود.

[19] Reconstruction

۲-۳-۳- بر اساس بازسازی

این دسته از الگوریتم‌ها بر اساس میزان خطای بازسازی به شناسایی نمونه‌های ناهنجار می‌پردازند. PCA[20] و الگوریتم‌های مشتق از آن جزو همین دسته روش‌ها هستند. اخیرا اساس کار بیشتر کارهای پژوهشی و کاربردی در زمینه تشخیص ناهنجاری شبکه‌های عصبی هستند و به نظر می‌رسد این شبکه‌ها دارای سابقه طولانی در این زمینه هستند[32]. به عنوان مثال رویکردهای مبتنی بر خودکدگذار و خودکدگذار متغیر[21] روند بازسازی نمونه‌های عادی را فرامی‌گیرند و نمونه‌های با خطاری بازسازی زیاد را به عنوان ناهنجاری در نظر می‌گیرند. مدل‌های مبتنی بر انرژی و مدل‌های ترکیبی گاوسی با خود کدگذار عمیق[22] به طور خاص در زمینه تشخیص ناهنجاری مورد تحقیق و پژوهش قرار گرفته‌اند. چنین روش‌هایی توزیع داده را با استفاده از خودکدگذار یا روش‌های مشابه مدل می‌کنند و سپس بر اساس انرژی و یا ترکیب گاوسی‌ها یک معیار آماری تشخیص ناهنجاری پدید می‌آورند. در سال‌های اخیر از شبکه‌های مولد تقابلی به منظور تشخیص ناهنجاری استفاده شده است. در این مدل‌ها به هنگام آزمایش برای هر نمونه ورودی عمل استنتاج انجام می‌شود و با استفاده از انتشار خطای گرادیان نزولی[23] در شبکه مولد، پارامترهای فضای نهفته بازیابی می‌شوند و با استفاده از این پارامترها می‌توان به شناسایی نمونه‌های ناهنجار پرداخت.

۲-۴- معیارهای ارزیابی روش‌های تشخیص ناهنجاری

صرف نظر از رویکرد به کار گرفته شده، تشخیص ناهنجاری از مرحله یادگیری که در آن با استفاده از نمونه‌های آموزشی مدل آموزش داده می‌شود، آغاز می‌شود. پس از اتمام مرحله یادگیری، مدل دسته‌بندی نمونه‌هایی که تاکنون آن‌ها را مشاهده نکرده است را بر اساس معیارهای مورد نظر انجام می‌دهد. . نتیجه ارزیابی در میزان تشخیص ناهنجاری‌ها می‌تواند در چهار دسته مختلف گزارش شود که

[20] Principal component analysis

[21] Variational autoencoder

[22] Deep autoencoding gaussian mixture models

[23] Gradient descent

عبارتند از مثبت صحیح[۲۴] به اختصار TP، منفی صحیح[۲۵] به اختصار TN ، مثبت کاذب یا FP و منفی کاذب یا FN[۲۶]. در حوزه تشخیص ناهنجاری از معیارهای معمول و استانداردی نظیر صحت[۲۷] ، بازیابی[۲۸]، F1-score و مساحت زیر نمودار مشخصه عملکرد[۲۹] و یا به اختصار AUROC که در ادامه به بررسی جزئیات ریاضی هر یک خواهیم پرداخت.

۲-۴-۱- صحت

معیار صحت یکی از معیارهای پایه در تمامی مسائل مربوط به دسته‌بندی علی الخصوص تشخیص ناهنجاری می‌باشد. این معیار بیانگر این است که چه مقدار از داده‌هایی که به عنوان ناهنجاری تشخیص داده شده‌اند، واقعا ناهنجار هستند. بیان ریاضی این معیار مطابق معادله ۲-۱ می‌باشد.

$$\text{Precision} = \frac{TP}{TP + FP} \qquad (۱-۲)$$

۲-۴-۲- بازیابی

این معیار در کنار معیار صحت دید نسبتا خوبی از وضعیت کلی عملکرد مدل نمایان می‌کند. این معیار بیان می‌کند که چه بخشی از داده‌های ناهنجار کشف شده و مدل توانسته چه نسبتی از این دسته را بازیابی کند. معادله ۲-۲ نحوه محاسبه این معیار را نمایش می‌دهد.

$$\text{Recall} = \frac{TP}{TP + FN} \qquad (۲-۲)$$

[24] True positive

[25] True negative

[26] False negative

[27] Precision

[28] Recall

[29] Area Under Curve Receiver Operating Characteristics

F1-score -۳-۴-۲

در حالت کلی، یک مدل خوب باید هر دو معیار صحت و بازیابی بالایی داشته باشد و به هنگام مقایسه مدل‌ها، در صورتی که هر دو معیار یادشده یک مدل از دیگری بیشتر باشد، مشخصا آن مدل عملکرد بهتری داشته است. اما اگر هر دو به صورت همزمان به سمت یک مقدار مناسب همگرا نباشند، انتخاب مدل بهتر با مشکل روبرو می‌شود و نیاز به یک معیار با دید کلی‌تر می‌باشد. معیار F1-score به طور همزمان هر دو جنبه مورد سنجش را ارزیابی می‌کند و معیار جامع‌تری برای مقایسه می‌باشد. بیان ریاضی این معیار مطابق معادله ۲-۳ می‌باشد.

$$\mathrm{F1-score} = \frac{2}{\frac{1}{\mathrm{Recall}} \cdot \frac{1}{\mathrm{Precision}}} = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \qquad (۳-۲)$$

۲-۴-۴- مساحت زیر نمودار مشخصه عملکرد

این معیار به منظور تعیین میزان تولنایی مدل در تشخیص ناهنجاری به هنگام افزایش لندازه مجموعه داده تعریف شده است. این معیار با هدف ایجاد اطمینان از توانایی مدل درکنترل تغییرات سریع حجم داده ایجاد شده است و برای محاسبه آن با توجه به جنس مجموعه داده مورد آزمایش روش‌های متفاوتی ارائه شده است[11].

۲-۵- شبکه‌های مولد تقابلی و تشخیص ناهنجاری

هدف از مدل‌های یادگیری عمیق، کشف مدل‌های سلسله‌مراتبی قوی است. این مدل‌ها نشان‌دهنده توزیع احتمال انواع داده‌هایی است که در کاربردهای هوش مصنوعی نظیر تصاویر طبیعی، شکل موج صوتی حاوی گفتار به کار می‌رود. برجسته‌ترین موفقیت یادگیری عمیق در طراحی مدل‌های تمایزگر[۳۰] بوده است. این مدل‌ها قادرند تا ورودی با ابعاد بالا را دریافت کنند و کلاس هر یک از نمونه‌ها را با قدرت

[30] Discriminator models

تشـخیص خود با دقت بالا مشـخص کنند. اسـتفاده از الگوریتم‌های پس‌انتشـار[۳۱] ، حذف تصـادفی[۳۲] و واحدهای خطی تکه‌ای[۳۳] که دارای گرادیان با رفتار مناسـب هسـتند عامل موفقیت چشـم‌گیر یادگیری عمیق است.

اسـتفاده کاربردی از الگوریتم‌هایی نظیر تخمین بیشـینه درسـتنمایی و الگوریتم‌های مرتبط با آن همراه با چالش‌ها و دشواری‌های زیادی نظیر محاسبات احتمالاتی زیاد و خارج از کنترل است، وجود این چالش و همچنین سـختی‌های موجود در اسـتفاده از واحدهای خطی تکه‌ای در حوزه مدل‌های مولد سـبب شـده است تا مدل‌های مولد عمیق کمتر مورد توجه قرار گیرند. شبکه‌های مولد تقابلی بر این دست از چالش‌ها و دشواری‌ها فائق آمده و نقش مدل‌های مولد عمیق را پررنگ‌تر ساخته است.

۲-۵-۱- شبکه‌های مولد تقابلی

شـبکه‌های مولد تقابلی اولین بار در سـال ۲۰۱۴ توسـط آقای گودفلو و همکاران ابداع شـد[۲۸] ، در این شبکه‌ها یک مدل مولد در برابر یک مدل تمایزگر قرار می‌گیرد، مدل تمایزگر سعی می‌کند میان داده‌های واقعی و داده‌های تولیدی توسط شبکه مولد تمایز ایجاد کند. مدل مولد را می‌توان مانند تیمی فرض کرد که سـعی در تولید ارز جعلی با میزان شـباهت بسـیار بالا به ارز واقعی دارد و در طرف مقابل مدل تمایزگر مشـابه پلیس اسـت که سـعی در کشـف ارز تقلبی دارد. رقابت در این بازی، هر دو تیم را به سـمت بهبود روش های خود سـوق می‌دهد تا اینکه ارز تقلبی از ارز اصـلی غیرقابل تشـخیص باشـد. این چارچوب می‌تواند الگوریتم‌های آموزشـی خاصـی را برای انواع مختلف از مسـائل و مدل‌ها بهینه‌سـازی کند. بخش مولد با دریافت نویز تصـادفی، از طریق پرسـپترون چند لایه نمونه‌هایی با توزیع مشـابه داده اصـلی تولید می‌کند و مدل تمایزگر با اسـتفاده از مدل پرسـپترونی چند لایه تلاش می‌کند تا میان نمونه‌های مختلف تمایز ایجاد کند. در این نوع از تعریف شبکه می‌توان هر دو مدل را با استفاده از الگوریتم‌های پس‌انتشار و

[31] Backpropagation

[32] Dropout

[33] Piecewise linear units

حذف تصادفی ایجاد کرد و برای نمونه‌گیری از مدل مولد تنها از الگوریتم انتشار رو به جلو[۳۴] استفاده کرد و در نتیجه به کارگیری هیچ الگوریتمی نظیر استنتاج تقریبی و یا زنجیره مارکوف ضروری نیست.

در شبکه مولد تقابلی به طور همزمان دو مدل مولد و تمایزگر آموزش داده می‌شود. مدل مولد G توزیع داده را ضبط می‌کند و مدل تمایزگر D احتمال این که نمونه از داده‌های تولید شده توسط G باشد را تخمین می‌زند. تابع هدف برای شبکه مولد G به حداکثر رساندن احتمال اشتباه شبکه D است. این بستر منجر به یک بازی دو نفره مانند بازی‌های بیشینه-کمینه[۳۵] می‌شود. در فضای توابع دلخواه G و D یک راه حل منحصر به فرد وجود دارد و در این راه حل شبکه مولد G توزیع داده‌های آموزشی را یاد گرفته است و شبکه تمایزگر D احتمال را در همه جا برابر مقدار ۱/۲ نشان می‌دهد، به بیان دیگر میان داده‌های واقعی و داده تولید شده توسط شبکه مولد نمی‌تواند تمیز دهد.

مدل‌سازی چارچوب تقابلی به وسیله ایجاد یک مدل چند لایه پرسپترونی برای هر دو مدل مولد و تمایزگر انجام می‌شود. برای یادگیری توزیع مولد p روی داده x ، یک تابع نویز خالص $p(z)$ را به عنوان ورودی تعریف می‌کنیم، سپس یک نگاشت از فضای نهفته به فضای داده را به عنوان $G(\mathrm{z};\theta_g)$ نشان می‌دهیم، در اینجا G یک تابع مشتق‌پذیر است که توسط یک پرسپترون چند لایه با پارامترهای θ_g نمایش داده می‌شود. همچنین برای شبکه تمایزگر D یک پرسپترون چند لایه لایه $D\left(\mathrm{x};\theta_d\right)$ با یک خروجی اسکالر[۳۶] تعریف می‌کنیم. $D(x)$ بیانگر احتمال این است که x از داده‌های اصلی به جای توزیع p_g ناشی شده باشد. در این میان به شبکه D آموزش داده می‌شود تا احتمال تخصیص برچسب صحیح به داده‌های واقعی و نمونه‌های تولیدی از G را به حداکثر برساند. به طور همزمان به شبکه G آموزش داده می‌شود تا تابع هدف $\log\left(1-\mathrm{D}\left(\mathrm{G}(\mathrm{z})\right)\right)$ را به حداقل برساند. به عبارت دیگر، شبکه‌های D و G بازی کمینه-بیشینه دو نفره زیر با تابع $V(G,D)$ مطابق معادله ۲-۴ انجام می‌دهند.

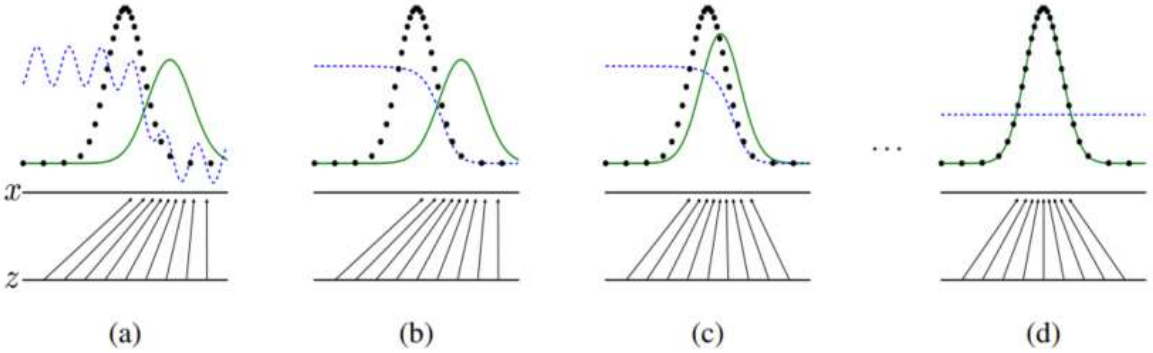
$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim q}[\log D(x)] + \mathbb{E}_{z\sim \mathrm{p}}[\log\left(1-D\left(G(z)\right)\right)] \qquad (۴-۲)$$

[34] Feed forward

[35] Minimax

[36] Scaler

با تحلیل نظری صورت گرفته بر روی شبکه‌های مولد تقابلی، نشان داده شده است که این شبکه‌ها پتانسیل کافی برای بازیابی توزیع داده‌های اصلی را در قالب شبکه مولد G دارند. در عمل معادله ۲-۱ ممکن است گرادیان کافی برای آموزش شبکه G را فراهم نکند. در اوایل یادگیری، هنگامی که شبکه مولد G ضعیف است، شبکه تمایزگر D می‌تواند با اطمینان بالا نمونه‌های غیرواقعی را شناسایی کند چرا که نمونه‌های تولیدی با نمونه‌های آموزشی کاملا متفاوت هستند. در این حالت $\log\left(1-D(G(z))\right)$ شباع می‌شود. در این حالت می‌توان به جای آموزش G برای به حداقل رساندن تابع $\log\left(1-D\left(G(z)\right)\right)$ می‌توانیم G را برای به حداکثر رساندن $D\left(G(\mathrm{z})\right)$ آموزش دهیم. این تابع در همان نقطه ثابت G و D قرار دارد اما گرادیان قوی‌تری در یادگیری فراهم می‌کند. رویکرد کلی شبکه‌های مولد تقابلی در شکل ۲-۲ نشان داده شده است.



شکل ۲- ۲: رویکرد کلی شبکه‌های مولد تقابلی[۳۱].

همانطور که در شکل ۲-۲ مشاهده می‌کنید توزیع تمایزگر (خط آبی شکسته) به‌روزرسانی می‌شود تا بتواند نمونه‌های توزیع داده‌های اصلی (خط مشکی نقطه‌چی) از داده‌های تولیدشده توسط توزیع مولد p_g (خط سبز پیوسته) تمیز دهد. خط افقی پایین بیانگر فضای نهفته است که متغیر z با توزیع یکنواخت از آن نمونه‌برداری شده است. خط افقی بالا بخشی از فضای داده واقعی x است. فلش‌های رو به بالا نشان می‌دهد که چگونه تابع G، z را به طور غیریکنواخت به x نگاشت می‌کند. به مرور زمان G در مناطق چگال‌تر منقبض می‌شود و در مناطق با چگالی کمتر باز می‌شود. در قسمت (a) شکل ۲-۲ دو بلوک تقابلی نزدیک همگرایی هستند،یعنی p به q نزدیک شده است و همچنین D دسته‌بند تا حدی دقیق می‌باشد. در قسمت (b) در حلقه داخلی الگوریتم، D آموزش می‌بیند تا بتواند نمونه‌های غیرواقعی را تشخیص دهد و به $D_G^*(x)=\frac{q(x)}{q(x)+p(x)}$ همگرا شود. در (c) پس از به روزرسانی G گرادیان ناشی از D، G را به گونه‌ای هدایت می‌کند که به سمت مناطقی مایل شود که به توزیع داده واقعی نزدیک‌تر شود. در قسمت آخر (d)

پس از انجام چند مرحله از آموزش اگر G و D ظرفیت کافی را داشته باشند به نقطه تعادلی می‌رسند که در آن $p = q$. و تمایزگر دیگر قادر نیست میان توزیع داده‌های واقعی و غیرواقعی تفاوتی قائل شود.

به منظور پیاده‌سازی این شبکه‌ها از روش عددی مبتنی بر تکرار استفاده می‌شود. تکمیل بهینه‌سازی شبکه D در حلقه داخلی مرحله آموزش همراه با چالش‌هایی نظیر هزینه محاسباتی زیاد است و همچنین روی دادگان کوچک منجر به بیش‌برازش[۳۷] است. حال به جای اینکه در هر تکرار هر دو شبکه D و G بهینه شوند، به ازای k مرحله بهینه کردن D یک مرحله G بهینه می‌شود. با این کار تا زمانی که G به اندازه کافی آهسته تغییر کند D در نزدیکی نقطه بهینه باقی خواهد ماند. شبه کد الگوریتم شبکه‌های مولد تقابلی در الگوریتم ۲-۱ آورده شده است.

k تعداد مراحل اعمال شده تمایزگر (برای کاهش هزینه محاسبات اینجا عدد یک فرض می‌شود) و n تعداد تکرار آموزش

for k steps do

  for n steps do

۱. نمونه‌برداری کوچک‌دسته‌ای[۳۸] mتایی نویز $\{\boldsymbol{z}^{(1)}, \dots, \boldsymbol{z}^{(m)}\}$ از نمونه‌های نویز$\boldsymbol{p}(\mathbf{z})$.

۲. نمونه‌برداری کوچک‌دسته‌ای mتایی نویز $\{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(m)}\}$ از داده‌های تولید توزیع$\boldsymbol{q}(\mathbf{x})$.

۳. بروزرسانی صعودی تمایزگر به وسیله گرادیان تصادفی.

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} [\log D(x^{(i)}) + \log (1 - D\left(G(z^{(i)})\right))]$$

  end for

٤. نمونه‌برداری کوچک‌دسته‌ای mتایی نویز $\{\boldsymbol{z}^{(1)}, \dots, \boldsymbol{z}^{(m)}\}$ از نمونه‌های نویز$\boldsymbol{p}(\mathbf{z})$.

٥. بروزرسانی صعودی مولد به وسیله گرادیان تصادفی.

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G(z^{(i)})\right)\right)$$

end for

الگوریتم ۲-۱: آموزش گرادیان نزولی کوچک دسته‌ای شبکه‌های مولد تقابلی.

[37] Overfitting

[38] Minibatch

۲-۵-۱-۱- تحلیل نظری شبکه مولد تقابلی

شبکه مولد G به طور ضمنی یک تابع توزیع احتمال p را به عنوان توزیع نمونه‌های $G(z)$ تعریف کرده است(توجه داشته باشید $z\sim p$). در صورتی که ظرفیت مناسب و زمان کافی برای آموزش در اختیار الگوریتم ۲-۱ قرار گیرد در نهایت این الگوریتم توزیع داده ورودی $q(x)$ را خواهد یافت. نتایج حاصل از شبکه‌های عصبی مولد تقابلی بر اساس تنظیمات غیرپارامتری به دست آمده است. در این بخش نشان می‌دهیم که در بازی کمینه-بیشینه بین دو شبکه $p = q$ یک بهینه عمومی است. برای این منظور ابتدا تمایزگر بهینه D را برای هر مولد G در می‌گیریم.

قضیه ۲-۱- برای هر تابع مولد G ثابت، تابع تمایزگر بهینه D عبارت است از:

$$D_G^*(x) = \frac{q(x)}{q(x) + p(x)} \qquad (۵-۲)$$

اثبات: معیار آموزش برای تمایزگر D، با توجه به هر مولد G، به حداکثر رساندن مقدار $V(G,D)$ می‌باشد، پس داریم:

$$\begin{aligned} V(G,D) &= \int_x^{\cdot} q(x)\log\big(D(x)\big)\,dx + \int_z^{\cdot} p(z)\log\Big(1 - D\big(g(z)\big)\Big)\,dz \\ &= \int_x^{\cdot} q(x)\log\big(D(x)\big)\,dx + p(x)\log\big(1 - D(x)\big)\,dx \end{aligned} \qquad (۶-۲)$$

از طرفی می‌دانیم برای هر $(a,b) \in \mathbb{R}^2$،تابع $y \to alog(y) + blog(1-y)$ دربازه $[۰,۱]$ در $\frac{a}{a+b}$ بیشینه است. هم‌چنین می‌دانیم تمایزگر نیاز به تعریف بیرون از مرز $Supp(p_{data}) \cup Supp(p_g)$ ندارد، پس در نتیجه $D_G^*(x)$ نقطه بهینه برای به حداکثر رساندن $V(G,D)$ می‌باشد.

می‌توان هدف از آموزش شبکه D را به حداکثر رساندن لگاریتم درست‌نمایی[۳۹] احتمال $P(Y = y|x)$ تعبیر کرد، که Y بیانگر آن است که هرجا x از توزیع q باشد y=۰ و هر جا از توزیع p باشد y=۱ است. با این تعریف بازی کمینه-بیشینه در معادله ۲-۴ را می‌توان به‌صورت زیر، بازنویسی کرد:

[39] Log-Liklihood

$$C(G) \underset{D}{=} \max V(G,D) = \mathbb{E}_{\mathrm{x}\sim\mathrm{q}}[\log D_G^*(\mathrm{x})] + \mathbb{E}_{\mathrm{z}\sim p(z)}[\log(1 - D_G^*(G(\mathrm{z})))]$$
$$+\mathbb{E}_{\mathrm{x}\sim\mathrm{q}}[\log D_G^*(\mathrm{x})] + \mathbb{E}_{\mathrm{x}\sim p}[\log(1 - D_G^*(\mathrm{x}))] \qquad (۷-۲)$$
$$= \mathbb{E}_{\mathrm{x}\sim\mathrm{q}}[\log\frac{q(x)}{q(x)+p(x)}] + \mathbb{E}_{\mathrm{x}\sim\mathrm{p}}[\log\frac{p(x)}{q(x)+p(x)}]$$

قضیه ۲-۲- کمینه سراسری $C(G)$ تنها در حالتی قابل محاسبه است که اگر و تنها اگر $q = p$ و مقدار $C(G)$ برابر با $-\log 4$ باشد.

اثبات: طبق قضیه ۲-۱، می‌دانیم هنگامی‌که $q = p$ باشد ، $D_G^* = 1/2$ می‌شود. در ادامه نشان دادیم که $C(G) = \log\frac{1}{2} + \log\frac{1}{2} = -\log 4$ حال برای اثبات این‌که بهترین مقدار $C(G)$ این مقدار است، این عبارت را از تعریف $C(G)$ کم می‌کنیم. طبق همگرایی جنسن-شانون[۴۰] داریم:

$$C(G) = -\log 4 + 2\,.\,JSD(p_{data} \,||\, p_g) \qquad (۸-۲)$$

می‌دانیم تابع همگرایی جنسن-شانون بین دو توزیع همواره نامنفی است و صفر است اگر و تنها اگر دو توزیع برابر باشند. پس ما نشان دادیم مقدار بهینه برابر با $-\log 4$ است و نقطه بهینه هنگام برابری دو توزیع رخ می‌دهد. بدین ترتیب اثبات این قضیه نیز به پایان رسید [۲۸].

۲-۵-۱-۲- مزایا و معایب

شبکه‌های مولد تقابلی نسبت به مدل‌های قبلی دارای مزایا و معایبی می‌باشد[31]. ایراد این روش آموزش این است که نمایش صریح $p(x)$ وجود ندارد و D باید در حین آموزش به خوبی با G هماهنگ شود و به طور خاص، G نباید بدون به‌روزرسانی D خیلی زیاد آموزش داده. اما از طرف دیگر، در هنگام یادگیری این مدل‌ها نیازی به استنباط نیست و می توان طیف گسترده‌ای از توابع را در مدل گنجانید. این مزیت در درجه اول محاسباتی است، مدل‌های تقابلی همچنین ممکن است برخی از مزیت‌های آماری

[40] Jensen-Shanon divergence

را از شبکه مولد به دست آورند که مستقیماً با نمونه داده‌ها به روز نمی‌شوند، و فقط با گرادیان‌هایی که از طریق تمایزگر جریان می‌یابند، بروزرسانی می‌شود.

این بدان معنی است که اجزای ورودی مستقیماً در پارامترهای مولد بکارگرفته نمی‌شوند.همچنین یکی دیگر از مزیت‌های شبکه‌های تقابلی این است که آن‌ها می‌توانند توزیع‌های بسیار تیز و لبه‌دار را نشان دهند، در حالی که روش‌های مبتنی بر زنجیره‌های مارکوف نیاز دارند که توزیع تا حدی هموار[41] باشد تا زنجیرها بتوانند میان حالت‌ها جابجا شده و مدل‌سازی را انجام دهند[31].

۲-۵-۲- مدل ANOGAN

مدل ANOGAN به منظور مدل‌سازی در زمینه پزشکی و به طور خاص مدل‌سازی وضعیت سلامت موضعی طراحی شده است[13]. این مدل از دسته الگوریتم‌های مولد و بدون‌نظارت است. با به وجود آمدن این مدل توانایی شبکه‌های مولد تقابلی در ایجاد یک مدل با قدرت بازنمایی بالا در تشریح وضعیت آناتومی ثابت شد. لازم به ذکر است مدل پیشنهادی شبکه تمایزگر و شبکه مولد را به طور همزمان آموزش می‌دهد و با استفاده از هر دو شبکه مشخص می‌کند داده ورودی از جنس داده‌های آموزشی است و یا باید به عنوان ناهنجاری دسته‌بندی شود.

به منظور تشخیص ناهنجاری مدل مورد نظر بازنمایی نمونه‌های متنوع آناتومیکی طبیعی را می‌آموزد. در این کار به جای استفاده از بهینه‌سازی تابع هزینه واحد، از تابع تعادل نش[42] میان هزینه‌ها استفاده شده است که سبب افزایش قدرت بازنمایی و رشد نرخ منفی صحیح[43] به اختصار TNR مدل تولیدی، بهبود روند نگاشت ویژگی[44] و همچنین دستیابی به دقت بالا در طبقه‌بندی داده‌های واقعی از داده‌های غیرواقعی می‌شود. در ادامه چگونگی طراحی مدل و نحوه شناسایی وضعیت و ظواهری که در داده‌های آموزش دیده نشده‌اند شرح داده خواهد شد.

[41] Smooth

[42] Nash cost

[43] True Negative Rate

[44] Feature matching

۲-۵-۲-۱- یادگیری بدون نظارت متنوع داده‌های طبیعی

M مجموعه‌ای از تصاویر پزشکی است که هر نمونه آن نمایانگر نمونه‌هایی از تصاویر آناتومی‌های سالم است و با I_m نمایش داده می‌شود که $m = 1,2, \dots, M$، که در اینجا $I_m \in \mathbb{R}^{a \times b}$ است، یعنی اندازه یک تصویر برابر $a \times b$ است. از هر تصویر I_m ، K تکه تصویر دو بعدی $x_{k,m}$، با ابعاد $c \times c$ بطور تصادفی از موقعیت‌های مختلف نمونه‌گیری می‌کنیم که منجر به داده‌های $x_{k,m} \in \mathcal{X}, k = 1,2, \dots, K$، می شود. در طول آموزش، فقط I_m را در اختیار داریم و برای یادگیری توزیع حاشیه‌ای، که نشان دهنده تنوع تصاویر آموزش است، از یک روش بدون نظارت استفاده می‌شود. برای آزمایش، داریم $\langle y_n, l_n \rangle$، که y_n تصاویر مشاهد نشده با ابعاد $c \times c$ استخراج شده از داده J است و $l_n \in \{0,1\}$ آرایه‌ای از برچسب‌های حقیقی مبتنی بر تصویر باینری با $n = 1,2, \dots, N$ است. این برچسب‌ها فقط در طول آزمایش استفاده می‌شوند، تا کارایی روش تشخیص ناهنجاری ارزیابی شود.

شبکه مولد G توزیع p را روی داده x از طریق نگاشت نمونه‌های z توسط تابع $G(\mathrm{z})$ آموزش می‌بینند؛ در واقع بردارهای تک بعدی با توزیع یکنواخت از فضای نهفته Z نمونه‌برداری می‌شوند و به فضای دو بعدی تصویر که در آن تصاویر آناتومی سالم وجود دارند نگاشت می‌شوند. در این تنظیمات، معماری شبکه مولد G معادل یک کدگذار پیچشی[۴۵] که از پشته‌های پیچشی استفاده می‌کند، در نظر گرفته می‌شود. تمایزگر D یک CNN استاندارد است که یک تصویر دو بعدی را به یک مقدار $D(.)$ نگاشت می‌کند. مقدار خروجی تمایزگر $D(.)$ بیانگر احتمال این است که ورودی تمایزگر، از فضای تصاویر واقعی یعنی فضای نمونه‌های آموزشی نمونه‌برداری شده باشد و یا توسط شبکه مولد تولید شده باشد. D و G به طور همزمان از طریق بازی کمینه-بیشینه با تابع $\mathrm{V}(D, G)$ و معادله ۲-۹ بهینه‌سازی می‌شوند[13].

$$\min_G \max_D V(D,G) = \mathbb{E}_{x \sim q(x)}[log\, D(x)] + \mathbb{E}_{z \sim p_{z(z)}}[log\, (1 - D(G(z)))] \quad (۹-۲)$$

در این بازی شبکه تمایزگر آموزش می‌بیند که احتمال اختصاص نمونه‌های واقعی را بیشینه و نمونه‌های تولیدی از p با برچسب جعلی را کمینه کند. هم‌چنین شبکه مولد G آموزش می‌بیند همزمان با حداقل

[45] Convolutional

کردن $V(G) = log\ (1 - D(G(z))$ که معادل با حداکثرکردن $V(G) = D(G(z))$ است، شبکه تمایزگر D را فریب دهد. به طور کلی در طول آموزش تقابلی، مولد در تولید تصاویر واقع بینانه و تمایزگر در شناسایی صحیح تصاویر واقعی و تولید شده بهبود می‌یابد.

۲-۵-۲-۲- نگاشت تصاویر جدید به فضای نهفته

وقتی آموزش تقابلی به پایان رسید، شبکه مولد یاد می‌گیرد که $G\ (z) = \mathrm{z} \to \mathrm{x}$ را از فضای نهفته با نمایش z به تصویر واقعی (عادی) x نگاشت کند. شبکه‌های GAN به‌طور خودکار نگاشت معکوس $= \mu(x)$ $\mathrm{x} \to \mathrm{z}$ را انجام نمی‌دهد. فضای نهفته دارای گذار خطی است، بنابراین نمونه‌گیری از دو نقطه نزدیک بهم در فضای نهفته، دو تصویر مشابه بصری نیز ایجاد می‌کند.

با فرض اینکه تصویر x را برای بررسی داریم، هدف این است که یک نقطه z را در فضای پنهان پیدا کنیم که مطابق با تصویر $G(z)$ باشد و از نظر بصری در آن نقطه $G(z)$ شبیه به تصویر x باشد و در توزیع حاشیه‌ای $\mathcal{X}$ قرار داشته باشد. میزان شباهت x و $G(z)$ بستگی به این دارد که چه تصویری از توزیع داده p_{g} برای آموزش مولد استفاده می‌شود. برای پیدا کردن بهترین z ، با نمونه‌گیری تصادفی z_1 از توزیع فضای نهفته $\mathcal{Z}$ شروع می‌کنیم و آن را به شبکه مولد آموزش دیده، برای تولید تصویر $G(z_1)$ اعمال می‌کنیم. سپس بر اساس تصویر ایجاد شده $G(z_1)$ یک تابع اتلاف تعریف می‌کنیم، که گرادیان به روزرسانی ضرایب z_1 را فراهم می‌کند و در نتیجه یک موقعیت بروز شده در فضای نهفته z_2 بدست می‌آید. به عبارتی برای پیدا کردن شبیه‌ترین تصویر $\mathrm{G}(\mathrm{z}_\Gamma)$، نقطه z در فضای نهفته $\mathcal{Z}$ در یک فرآیند تکراری از طریق $\gamma = 1,2, \dots, \Gamma$ با مراحل پس‌انتشار بهینه می‌شود.

تعریف تابع اتلاف برای نگاشت از تصاویر فضای نهفته شامل دو بخش است[34]، باقی‌مانده خطا [۴۶] و باقی‌مانده تمایز[۴۷]. باقی‌مانده اتلاف شباهت بصری بین تصویر تولید شده $G(z_\Gamma)$ و تصویر مورد بررسی x را تقویت می‌کند. باقی‌مانده تمایز، تصویر تولید شده $G(z_\Gamma)$ را در حاشیه توزیع آموزش دیده قرار می‌دهد. بنابراین، هر دو مؤلفه GAN آموزش دیده، تمایزگر D و مولد G ، برای یافتن ضرایب z از طریق پس‌انتشار مورد استفاده قرار می‌گیرند.

[46] Residual Loss

[47] Discrimination loss

باقی‌مانده خطا معیار عدم شباهت بصری بین تصویر مورد بررسی x و تصویر تولید شده $G(z_\gamma)$ در فضای تصویر اندازه‌گیری می‌کند و به‌صورت معادله ۲-۱۰ تعریف می‌شود.

$$\mathcal{L}_R(z_\gamma) = \sum |x - G(z_\gamma)| \qquad (۱۰-۲)$$

با فرض یک مولد کامل G و یک نگاشت کامل فضای نهفته، برای یک مورد بررسی ایده‌آل، تصاویر x و $G(z_\gamma)$ یکسان هستند. در این حالت باقی‌مانده خطا برابر با صفر است.
باقی‌مانده تمایز برای شبکه تمایزگر بدین ترتیب طبق معادله ۲-۱۱ تعریف می‌شود.

$$\mathcal{L}_{\hat{D}}(z_\gamma) = \sigma(D(G(z_\gamma)), \alpha) \qquad (۱۱-۲)$$

در این رابطه $G(z_\gamma)$ تصویر تولید شده توسط شبکه مولد و σ آنتروپی متقاطع سیگموئید[۴۸] است. هدف نهایی آموزش این است که خروجی شبکه تمایزگر برای تصاویر تولید برابر با ۱ باشد که به معنی این است که به ازای تصاویر تولیدی $G(z_\gamma)$ هدف نهایی آموزش $\alpha = 1$ تعریف می‌شود[13].

۲-۵-۲-۳- تشخیص ناهنجاری

در طی شناسایی ناهنجاری‌ها در داده‌ی جدید، ابتدا نمونه مورد بررسی جدید x را به عنوان یک تصویر طبیعی یا غیر عادی ارزیابی می‌کنیم. تابع اتلافی که برای نگاشت به فضای نهفته مورد استفاده قرار می‌گیرد، در هر تکرار γ بروزرسانی می‌شود و سازگاری تصاویر تولید شده $G(z_\gamma)$ با تصاویر را که در طول آموزش متخاصم مشاهده می‌شود ارزیابی می‌شود. بنابراین، این نمره ناهنجاری تناسب تصویر مورد بررسی x را با مدل تصاویر عادی بیان می‌کند، این معیار می‌تواند مستقیماً از تابع اتلاف در معادله ۲-۱۱ بدست آید. بدین ترتیب امتیاز ناهنجاری بر اساس تابع اتلاف مطابق معادله ۲-۱۲ تعریف می‌شود.

$$A(x) = (1 - \lambda) \cdot \mathcal{L}_R(z_\Gamma) + \lambda \cdot \mathcal{L}_D(z_\Gamma) \qquad (۱۲-۲)$$

که در آن به ترتیب $\mathcal{L}_R(z_\Gamma)$ مقدار باقی‌مانده خطا و $\mathcal{L}_D(z_\Gamma)$ باقی‌مانده تمایزگر است. این مدل، نمره ناهنجاری بزرگی برای تصاویر غیر عادی بدست می آورد و یک نمره ناهنجاری کوچک بدین معنی است که این تصویر بسیار مشابه تصاویر هنجاریست که قبلاً در طول آموزش دیده شده است. برای تشخیص

[48] Sigmoid cross entropy

ناهنجاری مبتنی بر تصویر، از نمره ناهنجاری $A(x)$ استفاده می‌شود. علاوه بر این در این‌جا، برای شناسایی مناطق غیر عادی در یک تصویر، از رابطه $x_R = |x - G(z_\Gamma)|$ استفاده شده است و بخش‌هایی از تصویر که x_R بزرگ‌تری دارند به عنوان بخش ناهنجار شناسایی می‌شوند[9].

۲-۵-۲-۴- مزایا و معایب

علیرغم نتایج قابل قبول این مدل در تشخیص ناهنجاری در تصاویر پزشکی، همانطور که مشاهده کردیم، برای بررسی تصویر x ، باید نقطه متناظر با آن در فضای نهفته پیدا شود به طوری که در آن نقطه $G(z)$ از نظر بصری بیشتر شبیه به تصویر x باشد. برای پیدا کردن بهترین z ، باید با نمونه‌گیری تصادفی z_1 از توزیع فضای نهفته $\mathcal{Z}$ شروع کنیم و آن را به شبکه مولد آموزش بدهیم و برای تولید تصویر، به $G(z_1)$ اعمال می‌کنیم. سپس بر اساس تصویر ایجاد شده $G(z_1)$ می‌بایست یک تابع اتلاف تعریف کنیم، که گرادیان به روزرسانی ضرایب z_1 را فراهم کند و در نتیجه یک موقعیت بروز شده در فضای نهفته z_2 بدست آورد. به عبارتی برای پیدا کردن شبیه‌ترین تصویر $G(z_\Gamma)$، نقطه z در فضای نهفته $\mathcal{Z}$ در یک فرآیند تکراری از طریق $\gamma = 1,2,\dots,\Gamma$ با مراحل پس‌انتشار بهینه و پیدا می‌شود. این فرایند تکراری مبتنی بر تصادف است و هزینه و پیچیدگی محاسباتی زیادی به مدل تحمیل می‌کند[35]. با توجه به وجود این مشکل می‌توان مکانیزمی طراحی کرد که مدل پیشنهادی فرایند نگاشت معکوس تصاویر در حین آموزش مدل فرابگیرد، در ادامه به بررسی همین مدل‌ها می‌پردازیم.

۲-۵-۳- مدل f-AnoGan

مدل f-AnoGan در ادامه کار قبلی و توسط همان نویسندگان در سال ۲۰۱۹ ارائه شد[35]. در مدل AnoGan از شبکه‌های عصبی عمیق کانولوشنی مولد تقابلی[۴۹] یا به اختصار DCGAN برای آموزش بدون نظارت شبکه مولد و تمایزگر استفاده شده است. در مدل قبلی برای شناسایی نقطه متناسب با تصویر ورودی بر اساس الگوریتم پس‌انتشار از یک فرایند مبتنی بر تکرار استفاده می‌شد. به هنگام شناسایی ناهنجاری در کاربردهای دنیای واقعی، این فرایند مبتنی بر تکرار از نظر پیچیدگی زمانی

[49] Deep convolutional generative adversarial network

مشـــکلات قابل توجهی ایجاد می‌کند. f-AnoGan فرایند مبتنی بر تکرار مورد نظر را با یادگیری یک نگاشت معکوس از فضای اصلی به فضای نهفته جایگزین می‌کند. علاوه بر این در ساختار مدل جدید از WGAN[۵۰] به جای DCGAN استفاده شده است.

چهارچوب ارائه شده در این کار شامل دو گام آموزشی روی تصاویر عادی است، در گام اول شبکه مولد تقابلی آموزش می‌بیند و در گام بعدی بر اساس شبکه مولد آموزش دیده شده شبکه کدگذار آموزش می‌بیند. پس از آموزش قسمت‌های مختلف مدل بر اساس استنتاج برای هر تصویر یک امتیاز ناهنجاری محاسبه می‌شود. همانند کار قبلی شبکه مولد روی تصاویر عادی آموزش داده می‌شود و بازنمایی‌های تصاویر سالم در فضای نهفته به دست می‌آید. علاوه بر این کدگذار نیز نگاشت تصاویر به فضای نهفته را آموزش می‌بیند. در ادامه در سه بخش به بررسی شبکه‌های تمایزگر و مولد، کدگذار و نحوه امتیازدهی به داده‌ها به منظور تشخیص داده‌های ناهنجار می‌پردازیم.

۲-۵-۳-۱- یادگیری بدون نظارت تصاویر طبیعی

داده‌های آموزشـــی در این مدل به شـــکل $x = x_{k,n} \in \mathcal{X}$ نمایش داده می‌شـــود. در این نوع از نمایش $k = 1,2, \dots, K$ و $n = 1,2, \dots, N$ است. به بیان دیگر تعداد N تصویر پزشکی داریم و از هر تصویر $\mathrm{I_n}$ به صورت تصادفی از مناطق مختلف تصویر تعداد K نمونه با ابعاد $s \times s$ انتخاب می‌شود. توجه شود $\mathrm{I_n} \in \mathbb{R}^{u \times v}$ و $u \gg s$ و $v \gg s$. برای ارزیابی مدل از داده تصـــویر y_m با ابعاد $s \times s$ که از میان دادگـان آزمـایش انتخـاب شـــده اســـت، حـاشـــیـه نویســـی پیکســـل مربوطـه بـه صـــورت $a_m \in \{0,1\}^{s \times s}$ و برچسب مربوطه استفاده می‌شود. این مجموعه داده آزمایش به صورت $\langle y_m, a_m \rangle$ است و به طور همزمان شامل داده عادی و ناهنجار است.

برای یادگیری تنوع موجود در تصاویر از WGAN استفاده می‌شود. این شبکه نگاشت غیرخطی از فضای نهفته $\mathbb{Z}$ به فضـای ورودی را یاد می‌گیرد. مانند مدل‌های دیگر در این مدل نیز شـبکه مولد و تمایزگر به طور همزمان بهینه می‌شـوند. در ابتدا از فضـای $\mathbb{z} \in \mathbb{R}^d$ با ابعاد d نمونه نویز انتخاب می‌کنیم. در طول فرایند آموزش شـــبکه مولد تلاش می‌کند تا توزیع شـــبکه مولد یعنی p را تا حد امکان به توزیع داده

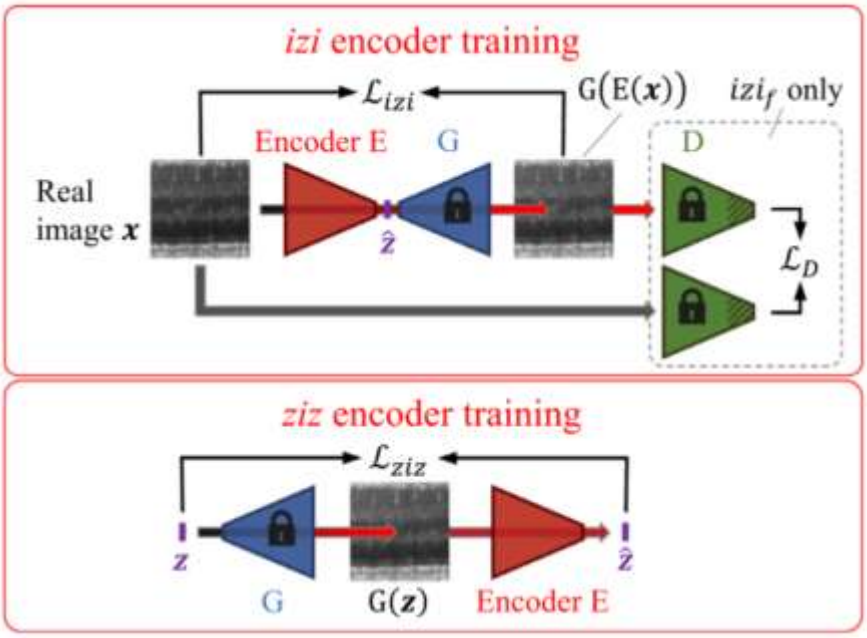
[50] Wasserstein GAN

ورودی یعنی q نزدیک کند و در نتیجه بتواند شبکه تمایزگر را به گونه‌ای فریب دهد که نتواند تشخیص دهد از توزیع داده واقعی است و یا توسط شبکه مولد ایجاد شده است. در پایان فرایند آموزش شبکه مولد توزیع داده‌های آموزشی را یاد گرفته است و شبکه تمایزگر می‌تواند تخمین بزند داده تولیدی توسط مولد تا چه اندازه به توزیع داده واقعی نزدیک است.

۲-۵-۳-۲- یادگیری نگاشت سریع از فضای تصویر به فضای نهفته

شبکه مولد GAN را می‌توان به صورت $G(z) = z \mapsto x$ نمایش داد. این نمایش به توانایی نگاشت شبکه مولد از فضای نهفته به فضای داده ورودی اشاره دارد. در روند آموزشی شبکه GAN اولیه هیچ نگاشتی از فضای داده ورودی به فضای نهفته آموزش داده نمی‌شود. در مدل پیشنهادی این مقاله نگاشت معکوس به صورت $E(x) = x \mapsto z$ نمایش داده می‌شود و این نگاشت توسط یک کدگذار آموزش دیده می‌شود. آموزش این کدگذار با دو معماری مختلف قابل پیاده‌سازی است، روش اول z-image-z و روش دوم image-z-image نام دارد. روش اول که به صورت خلاصه ziz و روش دوم به طور خلاصه izi نامیده می‌شود. در هر دو معماری از خودکدگذارهای کانولوشنی استفاده می‌شود. از کدگذار E برای نگاشت معکوس استفاده می‌شود و از شبکه مولد که در واقع یک WGAN با وزن‌های یادگرفته‌شده ثابت است، به عنوان کدگشا استفاده می‌شود. تفاوت دو روش فوق در ترتیب استفاده از کدگذار و کدگشاست.

هنگام آموزش کدگذار تنها پارامترهای کدگذار بهینه می‌شوند و پارامترهای شبکه مولد ثابت هستند. معماری‌های مختلف جهت آموزش کدگذار در شکل ۲-۳ قابل مشاهد است.



شکل ۲-۳: شمای کلی روند آموزش کدگذار[35].

درآموزش کدگذار به روش izi خطای $\mathcal{L}_{izi}$ بر اساس باقی‌مانده[51] از تفاوت تصاویر ورودی واقعی و تصویر بازسازی شده بهینه می‌شوند کد گذار آموزش می‌بینند. در حین آموزش کدگذار به روش $\mathcal{L}_{izi_f}$ به صورت توام خطای $\mathcal{L}_{izi}$ که همان خطای باقی‌مانده میان تصویر ورودی واقعی و تصویر بازسازی شده است به همراه خطای $\mathcal{L}_D$ که خطای باقی‌مانده روی ویژگی‌های شبکه تمایزگر است، بهینه می‌شود. در آموزش کدگذار به روش $\mathcal{L}_{ziz}$ خطای باقی‌مانده میان نمونه‌های تصادفی و موقعیت‌های موجود در فضای نهفته بهینه می‌شود.

الگوریتم ziz با معکوس کردن ترتیب کدگذار و کدگشا در ساختار معمول یک خودکدگذار به وجود می‌آید. در هنگام آموزش یک نمونه از فضای نهفته انتخاب می‌شود و با استفاده از شبکه مولد که وزن‌های آن ثابت نگاه داشته شده است به فضای داده واقعی نگاشت می‌شود و کدگذار تلاش می‌کند تا معکوس این نگاشت به فضای نهفته را یاد بگیرد، بنابراین در این روش به هیچ تصویر واقعی یا به بیان دیگر به هیچ نمونه‌ای از فضای داده واقعی نیاز نیست. در واقع در معماری ziz ساختار یک کدگذار از فضای نهفته به فضای نهفته است و در همین حال نگاشت مورد نیاز از فضای نهفته به فضای واقعی داده ورودی ثابت در نظر گرفته

[51] Residual

شده است. تابع هدف این آموزش به صورت خطای MSE[۵۲] روی نمونه اولیه z و مقدار بازسازی شده آن توسط کدگذار تعریف شده است. تابع هدف این معماری را در زیر مشاهده می‌کنید:

$$\mathcal{L}_{ziz}(z) = \frac{1}{d} \| z - E(G(z)) \|^2 \qquad (۱۳-۲)$$

d در معادله بیانگر ابعاد نمونه‌ها در فضای نهفته است. در این روش کدگذار بر خلاف روش izi هیچ نمونه‌ای از فضای تصاویر واقعی نمی‌بیند و این مسئله می‌تواند بر آموزش صحیح کدگذار تاثیر منفی بگذارد.

در آموزش کدگشا به روش izi از ساختار کدگذار استاندارد استفاده می‌شود، بدین صورت که در ادامه کدگذار کدگشا (شبکه مولد) قرار خواهد گرفت. در فرایند آموزش ابتدا نگاشت معکوس از فضای داده واقعی به فضای نهفته توسط کدگذار انجام می‌شود و در ادامه نگاشت از فضای نهفته به فضای داده واقعی توسط کدگشا با ضرایب ثابت صورت می‌پذیرد. ساختار این روش به صورت از فضای واقعی به فضای واقعی است. تابع هدف این روش با استفاده از خطای MSE بدین شکل پیاده‌سازی می‌شود که خطای باقی‌مانده میان تصویر واقعی و تصویر خروجی مولد کمینه می‌شود، تابع هدف مورد نظر در ادامه آمده است:

$$\mathcal{L}_{izi}(x) = \frac{1}{n} \| x - G(E(x)) \|^2 \qquad (۱۴-۲)$$

در معادله ۲-۱۴ $\|\cdot\|^2$ بیانگر جمع مربعات خطا در سطح پیکسل میان دو تصویر است. داده‌های آموزشی این روش داده‌های آموزش همان WGAN یعنی داده‌های عادی است. در این روش نمی‌توان به طور مستقیم میزان دقت کدگذار را در فضای نهفته اندازه گرفت و تنها می‌توان به صورت غیرمستقیم نگاشت مربوط به فضای نهفته را به فضای داده واقعی انتقال داده و در این فضا میزان دقت را اندازه گرفت به بیان دیگر میزان دقت به صورت تصویر-تصویر محاسبه می‌شود.

در روش دیگر به نام izi_f از تمایزگر نیز استفاده می‌شود. در روش izi تابع هدف میزان شباهت در فضای تصویر را تحمیل می‌کند. هنگام نگاشت تصاویر جدید ممکن است با نمونه‌هایی روبرو شویم که در مرحله

[52] Mean squared error

آموزش به صورت تنک از فضای نهفته متناظر آن‌ها نمونه گرفته باشیم وقتی نقطه متناظر را به فضای تصویر(فضای داده ورودی) می‌بریم با تصاویر تولیدی دیگر نمی‌توان تمایزگر را متقاعد کرد. در نتیجه این اتفاق تنها کمینه کردن تفاوت تصاویر در سطح پیکسل گاهی اوقات منجر به تولید تصاویر عادی می‌شوند که واقعی به نظر نمی‌رسند ولی هنوز خطای باقی‌ملنده کمی حتی برای نمونه‌های ناهنجار دارند و این مورد سبب می‌شود تا دیگر خطای باقی مانده (خطای بازسازی) در فضای داده ورودی دیگر به عنوان معیار مناسب تشخیص ناهنجاری در نظر گرفته نشود.

نویسندگان مقاله دریافتند که باقی‌مانده که خود معیار مورد نظر ما برای تشخیص ناهنجاری است در فضای ویژگی توسط تمایزگر انباشته می‌شود و این عبارت حتما باید در تابع هدف مربوط به آموزش کدگذار گنجانده شود. بنابر این آمارگان تصاویر ورودی و تصاویر خروجی محاسبه می‌شود تا با استفاده از آن‌ها تصاویر خروجی شبکه مولد شبیه تصاویر ورودی بشود و بدین ترتیب روش izi_f پدید آید. تابع هدف این روش به شکل زیر است:

$$\mathcal{L}_{izi_f}(x) = \frac{1}{n} \cdot \| x - G\big(E(x)\big) \|^2 + \frac{k}{\text{n}_\text{d}} \cdot \| f(x) - f\big(G\big(E(x)\big)\big) \|^2 \qquad (۱۵-۲)$$

در معادله ۲-۱۵ ویژگی‌های شبکه تمایزگر که در واقع بردار ویژگی لایه‌های میانی این شبکه است با $f(\cdot)$ و ابعاد این ویژگی با نماد n_d نمایش داده می‌شود، همچنین κ عامل وزن است. اوزان شبکه تمایزگر همان اوزانی است که در آموزش WGAN یاد گرفته شده‌اند و هنگام آموزش کدگشا ثابت در نظر گرفته شده‌اند. این مدل سبب می‌شود تا هم در فضای تصویر و هم در فضای نهفته کدگذار به جهت مناسبی حرکت کند.

۲-۵-۳-۳- شناسایی ناهنجاری

در مرحله آزمایش میزان انحراف تصویر اصلی از تصویر بازسازی شده به منظور تشخیص ناهنجاری اندازه‌گیری می‌شود. تمامی موارد مورد نیاز برای بازسازی تصویر و تشخیص ناهنجاری در هنگام آموزش WGAN و کدگذار یادگرفته می‌شود. برای محاسبه امتیاز ناهنجاری مستقیما از تعریف خطای استفاده شده در آموزش کدگذار استفاده می‌شود. امتیاز نهایی که برای تشخیص ناهنجاری در مدل f-AnoGan استفاده می‌شود به صورت زیر است.

$$\mathcal{A}(X) = \mathcal{A}_R(X) + \kappa.\mathcal{A}_D(X) \qquad (۱۶-۲)$$

در اینجا $\mathcal{A}_R(X) = \frac{1}{n}\cdot\| X - G(E(X)) \|^2$ و $\mathcal{A}_D(X) = \frac{1}{n_d}\cdot\| f(x) - f(G(E(x))) \|^2$ ، همچنین κ عامل وزن اســـت. عبارت مورد نظر برای نمونه‌های عادی دارای خطای کمی اســـت و برای نمونه‌های ناهنجار دارای مقدار بزرگی اســـت. از آنجایی که مدل تنها روی نمونه‌های عادی آموزش دیده است نمونه‌های بازسازی شده از نظر بصری شبیه تصویر ورودی هستند. توانایی بازسازی تصویر به طوری که شبیه تصویر ورودی باشد رابطه عکس دارد با درجه یا میزان تمایز ناهنجاری دارد. تصاویر عادی میزان انحراف کمی دارند در حالی‌که تصاویر ناهنجار که به بازسازی خود نگاشت می‌شوند میزان انحراف زیادی دارند. قدر مطلق خطای باقی‌مانده در سطح پیکسل به صورت زیر تعریف می‌شود.

$$\dot{\mathcal{A}}_R(X) = |X - G(E(X))| \qquad (۱۷-۲)$$

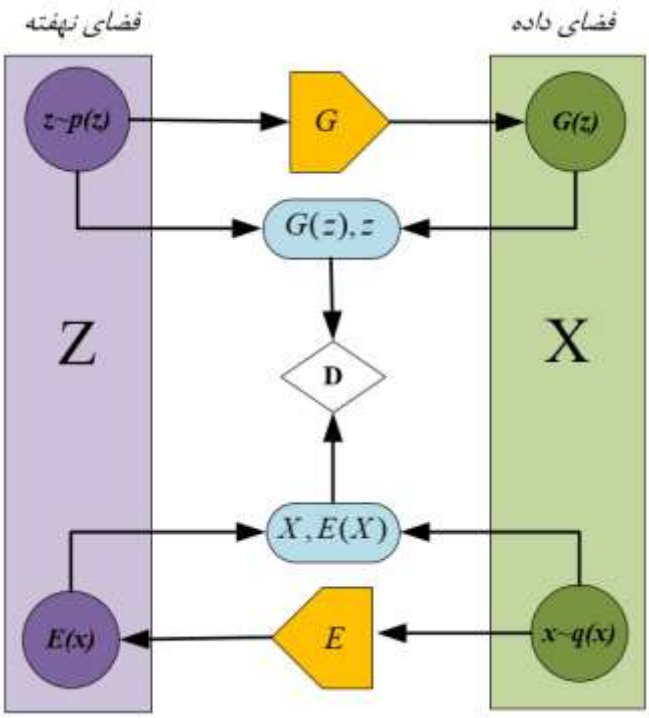
۲-۵-۳-۴- مزایا و معایب

f-AnoGan مشـــکلات مدل قبلی (AnoGan) نظیر پیچیدگی زمانی بالا در هنگام اجرا رفع کرد. همچنین از شـــبکه مولد تقابلی قوی‌تری در ســـاختار مدل خود اســـتفاده کرد و به نتایج قابل قبولی نیز دست یافت. در مدل قبلی از یک فرایند تصادفی و مبتنی بر تکرار برای نگاشت معکوس از فضای داده واقعی به فضـــای نهفته اســـتفاده می‌شـــد که از نظر زمانی هزینه گزافی را به هنگام اجرا به مدل تحمیل می‌کرد. در مدل جدید f-AnoGan با اســـتفاده از یک کدگشـــا پارامترهای نگاشـــت معکوس مورد نظر فراگرفته می‌شود[14]. علی‌رغم این موفقیت این مدل همچنان از مشکلاتی نظیر عدم استفاده توام از هر دو فضای تصویر و نهفته و نبود چرخه پایداری در حین آموزش رنج می‌برد.

۲-۵-۴- مدل ALI

این شبکه در سال ۲۰۱۷ در کنفرانس ICLR معرفی شد[14]. این شبکه‌ها با هدف یادگیری نگاشت معکوس از فضای ورودی x به فضای نهفته z تعریف شد. در این شبکه، علاوه بر شبکه مولد G که در

معماری اصلی نیز تعریف شده بود، یک کدگذار[۵۳] E نیز وجود دارد که از دامنه داده‌های ورودی x به دامنه ویژگی‌ها z می‌برد. بدین ترتیب خروجی بخش مولد یک دوتایی[۵۴] است؛ که یکی از دامنه ویژگی‌ها و دیگری از دامنه داده‌های ورودی است. این مدل به طور همزمان شبکه مولد و شبکه استنتاج را با استفاده از یک فرآیند تقابلی به کار می‌برند. شبکه مولد، نمونه‌ها را از یک فضای نهفته آماری به فضای داده‌ها نگاشت می‌کند و شبکه استنتاج نمونه‌های آموزش را از فضای داده به فضای متغیرهای نهفته نگاشت می‌کند. به این صورت یک بازی خصمانه بین دو شبکه انجام می‌شود. در این جا شبکه متمایزگر باید یاد بگیرد تا تفاوت بین جفت ورودی فضای نهفته و فضای داده را تشخیص دهد. شبکه تمایزگر D در این‌جا علاوه بر تفکیک در فضای داده، در فضای ویژگی نیز تفکیک می‌کند. به این معنا که تشخیص می‌دهد دوتایی واردشده، داده واقعی است با ویژگی تولیدشده توسط E و یا داده جعلی است که توسط G و به همراه ویژگی‌های z درست شده، است. در تصویر زیر چارچوب کلی این الگوریتم، به نمایش درآمده است:



شکل ۲-۴: معماری شبکه ALI.

[53] Encoder

[54] Tuple

دو تابع توزیع احتمال روی x و z در نظر بگیرید:

- $q(x,z) = q(x)q(z|x)$ تابع توزیع تعریف شده برای کدگذار E
- $p(x,z) = p(z)p(x|z)$ تابع توزیع تعریف شده برای کدگشا G

این دو توزیع، توابع توزیع حاشیه‌ای دارند که برای ما آشناست: توزیع حاشیه‌ای کدگذار $q(x)$ تابع توزیع داده‌های اصلی است و توزیع حاشیه‌ای کدگشا $p(z)$ معمولا به عنوان یک تابع توزیع ساده مانند تابع توزیع استاندارد $p(z) = N(0, \mathrm{I})$ در نظر بگیریم. بدین ترتیب روند تولید $p(x,z)$ و $q(\mathrm{x}, \mathrm{z})$ معکوس می‌باشد.

هدف اصلی شبکه ALI مطابقت این دو توزیع است. اگر این شرط محقق شود، ما اطمینان حاصل می‌کنیم که تمام توزیع‌های حاشیه‌ای و توزیع‌های شرطی مطابقت دارد. برای دستیابی به این توابع توزیع، یک بازی تقابلی صورت می‌گیرد. جفت (x,z) از دو توزیع $q(x,z)$ یا $p(x,z)$ در نظر گرفته می‌شود و یک شبکه تمایزگر می‌آموزد تا بین این دو خروجی، تمایز قائل شود؛ در حالی که دو شبکه کدگشا و کدگذار می‌آموزند تا این شبکه را فریب دهند. در نهایت تابع هدفی که این بازی بر اساس آن صورت می‌گیرد به صورت زیر است:

$$\min_{\mathrm{G,E}} \max_{\mathrm{D}} \mathrm{V(D,G)} = \mathbb{E}_{\mathrm{q(x,z)}}[\log \mathrm{D(x}, E(\mathrm{x}))] + \mathbb{E}_{\mathrm{p(x,z)}}[\log\,(1 - \mathrm{D}(G(z), z)] \qquad (۱۸-۲)$$

ویژگی جالب رویکردهای خصمانه این است که آن‌ها نیازی به محاسبه تابع چگالی شرطی ندارند و تنها نیاز دارند که به نحوی نمونه برداری شوند که این امکان را به وجود آورد که بتواند از پس انتشار گرادیان برای آموزش شبکه استفاده شود. در مورد شبکه ALI ، این بدان معنی است که گرادیان ها باید از شبکه تمایزگر به شبکه‌های مولد و کدگذار انتشار یابند.

به طور دقیق‌تر شبکه تمایزگر آموزش می بیند که بین نمونه‌هایی که از کدگذار $(x\ ،\ \hat{z}) \sim q\ (x\ ،\ z)$ و نمونه هایی که از کدگشا $(\hat{x}\ ،\ z) \sim p\ (x\ ،\ z)$ تولید می‌شود، تمایز بگذارد. شبکه مولد و شبکه کدگذار نیز می‌آموزند که شبکه تمایزگر را فریب دهند؛ یعنی جفت $x\ ،\ z$ تولید کنند که $p(x\ ،\ z)$ از $q(x\ ،\ z)$ غیر قابل تشخیص باشد. شاید در این‌جا سوال مطرح شود که چرا باید شبکه کدگذار تلاش کند تا شبکه تمایزگر را فریب دهد و چرا باید شبکه کدگذار تلاش کند تا در معادله ۲-۱۸ خروجی تمایزگر برای داده‌های هنجار کمینه شود.

برای پاسخ به این سوال باید به این نکته توجه کرد که در حین آموزش شبکه کدگذار، داده آموزشی برای این شبکه وجود ندارد و تنها اطلاعات موجود قابل استفاده، جفت داده‌هایی است که توسط شبکه مولد ساخته‌شده است و این شبکه بر اساس آموزش و نگاشتی که شبکه مولد استخراج کرده است، نگاشت معکوس را یاد می گیرد. در واقع در حین آموزش شبکه کدگذار مفهوم داده آموزشی عوض شده و این شبکه با تلاش به فریب دادن شبکه تمایزگر آموزش می یابد. در این‌جا اساس کار یادگیری، تولید نگاشت معکوس به فضای نهفته z برای تصاویر ورودی اصلی شبیه به جفت تولید شده توسط شبکه مولد، می‌باشد. یعنی داده آموزشی برای این شبکه تصاویر تولیدی مولد و متغیر z آن است و داده ورودی برای آزمون و تولید خروجی، تصاویر x اولیه اصلی شبکه هستند.

در الگوریتم ۲-۲ شبکه ALI توصیف شده است. اثبات می‌شود که با فرض یک تمایزگر بهینه، شبکه مولد، واگرایی جنسن-شانون[۵۵] را بین $p\ (x\ ,\ z)$ و $q\ (x, z)$ به حداقل می‌رساند.

روش استفاده شده در مدل ALI تنها راه استنتاج در شبکه‌های عصبی مولد تقابلی نیست. راه دیگر برای انجام این کار استفاده از شبکه استنتاج جلورو[۵۶] در ساختارGAN است. در مدل InfoGAN [۳۲] با کمینه کردن اطلاعات متقابل[۵۷] میان مجموعه C از فضای نهفته و X به وسیله توزیع کمکی $Q(c \mid x)$ نگاشت معکوس را یاد می‌گیرد. InfoGan نیاز دارد تا تابع احتمال پسین[۵۸] $Q(c \mid x)$ قابل تخمین و ارزیابی باشد. در مدل ALI تنها نیاز است که بتوان از شبکه استنتاج نمونه گرفت تا بدین وسیله تابع پیچیده توزیع پسین را بازنمایی کرد. عمل انجام شده در این کار مشابه این است که یک کدگذار برای بازسازی z آموزش دهیم. به عنوان مثال پیدا کردن کدگذار به طوری که $\mathbb{E}_{z\sim p(z)}\left[\|z - \mathrm{E}(G(z))\|_2^2\right] \approx 0$. نمونه که در جمله قبل بدان اشاره شد از نظر رویه‌ای شبیه به InfoGAN اما در این روش از یک شبکه مولد با ضرایب ثابت و همچنین تابع توزیع پسین گاوسی با واریانس قطری ثابت استفاده شده است.

[55] Jenesen-Shannon Divergence

[56] Feedforward

[57] Mutual information

[58] Posterior

روند آموزش را می‌توان به دو فاز تقسیم کرد. در فاز اول شبکه مولد تقابلی به صورت معمول آموزش می‌بینند. در فاز دوم کدگشا ثابت در نظر گرفته می‌شود و کدگذار به روش مدل ALI آموزش داده می‌شود. در این روش کدگذار و کدگشا در هنگام آموزش هیچ تعاملی با هم ندارند و در واقع کدگذار بر اساس هر چه کدگشا آموخته است آموزش می‌بینند. مشخص است اگر کدگذار و مولد با هم تعامل داشته باشند روند مدل‌سازی داده بهبود خواهد یافت.

رویه آموزش یادگیری خصمانه استنتاج

مقداردهی اولیه پارامترها $\boldsymbol{\theta_g, \theta_d} \leftarrow$

Repeat

$\boldsymbol{x^{(1)}, \dots, x^{(M)} \sim q(x)}$ نمونه برداری اولیه از مجموعه داده M

$\boldsymbol{z^{(1)}, \dots, z^{(M)} \sim p(z)}$

$\boldsymbol{\hat{x}^{(i)} \sim q\left(\left(z|x = x^{(i)}\right)\right), \quad i = 1, \dots, M}$ انتخاب شرطی

$\boldsymbol{\hat{x}^{(j)} \sim q\left(\left(z|x = x^{(j)}\right)\right), \quad j = 1, \dots, M}$

$\boldsymbol{\rho_{(q)}^{(i)} \leftarrow D\left(x^{(i)}, \hat{z}^{(i)}\right), \quad i = 1, \dots, M}$ محاسبه پیش‌بینی تمایزگر

$\boldsymbol{\rho_{(p)}^{(j)} \leftarrow D\left(\hat{x}^{(i)}, z^{(j)}\right), \quad j = 1, \dots, M}$

$\boldsymbol{\mathcal{L}_d \leftarrow -\frac{1}{M}\sum_{i=1}^{M} \log\left(\rho_q^{(i)}\right) - \frac{1}{M}\sum_{j=1}^{M} \log\left(1 - \rho_q^{(j)}\right)}$ محاسبه تلفات تمایزگر

$\boldsymbol{\mathcal{L}_g \leftarrow -\frac{1}{M}\sum_{i=1}^{M} \log\left(1 - \rho_q^{(i)}\right) - \frac{1}{M}\sum_{j=1}^{M} \log\left(\rho_q^{(j)}\right)}$ محاسبه تلفات مولد

$\boldsymbol{\theta_d \leftarrow \theta_d - \nabla_{\theta_d}\mathcal{L}_d}$ بروزرسانی گرادیان تمایزگرشبکه

$\boldsymbol{\theta_g \leftarrow \theta_d - \nabla_{\theta_g}\mathcal{L}_g}$ بروزرسانی گرادیان مولد شبکه

Until

الگوریتم ۲-۲: رویه آموزش یادگیری خصمانه استنتاج

۲-۵-۴-۱- مقایسه مدل‌های ALI و GAN

شبکه ALI شباهت زیادی به شبکه GAN دارد ، اما دو تفاوت اساسی با آن دارد:

۱- بخش مولد دارای دو مؤلفه است: بخش کدگذار،$E(x)$ که نمونه‌های داده x را به فضای z نگاشت می‌کند و بخش مولد $G(z)$ که نمونه‌ها را از $p(z)$ (منبع منبع نویز) به فضای ورودی نگاشت می‌کند.

۲- بخش تمایزگر به منظور تمایز بین جفت $(\hat{x} = G(z)، z)$ و $(x ، \hat{z} = E(x))$، آموزش دیده می‌شود.

۲-۵-۴-۲- رویکردهای جایگزین برای استنتاج در GAN

روش استفاده شده در مدل ALI تنها راه استنتاج در شبکه‌های عصبی مولد تقابلی نیست. راه دیگر برای انجام این کار استفاده از شبکه استنتاج جلورو[۵۹] در ساختارGAN است. در مدل InfoGAN[۶۰] [۳۲] با کمینه کردن اطلاعات متقابل[۶۱] میان مجموعه C از فضای نهفته و X به وسیله توزیع کمکی $Q(c \mid x)$ نگاشت معکوس را یاد می‌گیرد. InfoGan نیاز دارد تا تابع احتمال پسین[۶۲] $Q(c \mid x)$ قابل تخمین و ارزیابی باشد. در مدل ALI تنها نیاز است که بتوان از شبکه استنتاج نمونه گرفت تا بدین وسیله تابع پیچیده توزیع پسین را بازنمایی کرد. عمل انجام شده در این کار مشابه این است که یک کدگذار برای بازسازی z آموزش دهیم. به عنوان مثال پیدا کردن کدگذار به طوری که $\mathbb{E}_{\boldsymbol{z}\sim\boldsymbol{p}(\boldsymbol{z})}[\|z - G_z(G_x(z))\|_2^2] \approx 0$. نمونه که در جمله قبل بدان اشاره شد از نظر رویه‌ای شبیه به InfoGAN اما در این روش از یک شبکه مولد با ضرایب ثابت و همچنین تابع توزیع پسین گاوسی با واریانس قطری ثابت استفاده شده است.

روند آموزش را می‌توان به دو فاز تقسیم کرد. در فاز اول شبکه مولد تقابلی به صورت معمول آموزش می‌بینند. در فاز دوم کدگشا ثابت در نظر گرفته می‌شود و کدگذار به روش مدل ALI آموزش داده می‌شود. در این روش کدگذار و کدگشا در هنگام آموزش هیچ تعاملی با هم ندارند و در واقع کدگذار بر اساس هر چه کدگشا آموخته است آموزش می‌بینند. مشخص است اگر کدگذار و کدگشا با هم تعامل داشته باشند روند مدل‌سازی داده بهبود خواهد یافت.

۲-۵-۴-۳- مزایا و معایب

اگرچه وجود مکانیزم مناسب جهت نگاشت معکوس از فضای داده واقعی به فضای نهفته سبب موفقیت و عملکرد قابل قبول مدل ALI شده است اما در این مدل هیچ سازوکاری جهت کنترل میزان شباهت تصویر اصلی و تصویر بازسازی شده در آن تعبیه نشده است و به همین جهت در برخی موارد هیچ شباهتی میان

---

[59] Feedforward

[60] Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets

[61] Mutual information

[62] Posterior

داده اصلی و داده بازسازی شده وجود ندارد. به بیان دیگر پس از یافتن نقطه متناسب با داده مورد نظر در فضای داده واقعی در فضای نهفته، نقطه مورد نظر را به شبکه مولد می‌دهیم و خروجی حاصل را با ورودی ابتدایی مقایسه می‌کنیم. انتظار می‌رود داده ورودی و خروجی در این چرخه شباهت زیادی داشته باشند ولی همانطور که گفته شد در برخی از نمونه‌ها شباهتی میان این دو تصویر وجود ندارد.

۲-۵-۵- مدل EGBAD

همانطور که گفته شد شبکه‌های عصبی تقابلی قادرند توزیع‌های پیچیده دنیای واقعی با ابعاد بالا را مدل کنند و همین امر سبب می‌شود تا بتوان از این شبکه‌ها در زمینه تشخیص ناهنجاری نیز کرد. با توجه به تعداد کارهای انگشت شمار در این زمینه، مدل EGBAD را می‌توان از اولین کارها در تشخیص ناهنجاری با استفاده از شبکه‌های عصبی تقابلی به شمار آورد[۳۶].

مدل ارائه شده بر اساس شبکه عصبی تقابلی دوطرفه به اختصار BiGAN[۶۳] بنا نهاده شده است. وظیفه نگاشت معکوس از فضای داده ورودی به فضای نهفته نیز بر عهده کدگذار است. کدگذار، شبکه مولد و تمایزگر در اینجا به طور همزمان آموزش می‌بینند و وجود بلوک کدگذار سبب کاهش هزینه‌های محاسباتی در گام آزمایش می‌شود. بر خلاف ساختار استاندارد GAN که در آن تمایزگر تنها تصویر واقعی و تصویر تولیدی شبکه مولد را ورودی می‌گیرد، بازنمایی این تصاویر در فضای نهفته هم به عنوان ورودی به شبکه تمایزگر داده می‌شود.

استراتژی مورد استفاده در گام آموزش مدل مشابه شبکه ALI است که در فصل قبل به تفضیل به توضیح آن پرداختیم. همانطور که در قسمت قبل بررسی شد در این استراتژی آموزشی تاکید بر آن است که شبکه مولد و کدگذار به طور توامان آموزش داده شوند. تابع هزینه در هنگام آموزش مطابق معادله ۲-۱۹ بهینه می‌شود.

$$V(D,E,G) = \mathbb{E}_{x\sim p_X}\left[\mathbb{E}_{z\sim p_E(\cdot|x)}[\log\ D(x,z)]\right] + \mathbb{E}_{z\sim p_Z}\left[\mathbb{E}_{x\sim p_G(\cdot|z)}[1-\log\ D(x,z)]\right] \qquad (۲-۱۹)$$

[63] Bidirectional Generative Adversarial Model

در معادله ۲-۱۹ $p_X(x)$ بیانگر تابع توزیع داده ورودی است، $p_Z(z)$ بیانگر توزیع نمونه‌ها در فضای نهفته است، $p_E(z \mid x)$ و $p_G(x \mid z)$ به ترتیب بیانگر تابع توزیع کدگذار و مولد هستند.
پس از آموزش مدل نوبت به تعریف معیاری می‌رسد که به وسیله آن بتوان نمونه‌های ناهنجار را تشخیص داد. بیان ریاضی این معیار در ادامه آمده است.

$$A(x) = \alpha L_G(x) + (1-\alpha) L_D(x) \qquad (۲۰-۲)$$

در معادله ۲-۲۰ $L_G(x) = \| x - G(E(x)) \|_1$ و $L_D(x)$ را می‌توان به دو روش تعریف کرد. در روش اول تابع خطای آنتروپی[۶۴] متقابل استفاده می‌شود. ورودی این تابع در واقع خروجی تمایزگر است و خود شامل خروجی کدگذار و نمونه متناظر با آن است و به صورت $\sigma(D(x, E(x)),1)$ تعریف می‌شود. در اینجا خروجی تمایزگر میزان اطمینان آن در قبال واقعی بودن نمونه ورودی است. در روش دوم تعریف $L_D(x)$ از خطای تطبیق ویژگی[۶۵] استفاده می‌شود. تابع امتیاز این روش به صورت $L_D(x) = \|f_D(x, E(x)) - f_D(G(E(x)), E(x))\|_1$ تعریف می‌شود. f_D در واقع لاجیت‌های تمایزگر است و بیان می‌کند ویژگی‌های تصویر بازسازی شده تا چه اندازه شبیه ویژگی‌های تصویر واقعی هستند. توجه شود هر چه مقدار امتیاز محاسبه شده بالاتر باشد نمونه مورد نظر با احتمال بیشتری نمونه ناهنجار است.

۲-۵-۵-۱- مزایا و معایب

مدل مورد بحث در این قسمت ثابت کرد می‌توان با تعریف تابع امتیاز مناسب و استفاده از ساختارهای تقابلی به روز ناهنجاری را داده‌های دنیای واقعی شناسایی کرد. در این مدل هیچ نظارتی بر میزان تشابه داده ورودی و داده بازسازی وجود ندارد در حالی که انتظار می‌رود داده ورودی و داده بازسازی شده توسط شبکه مولد برای نمونه‌های هنجار یکسان باشد. در ادامه به بررسی این مشکل و روش ارائه شده برای حل آن پرداخته می‌شود.

[64] Cross entropy

[65] Feature matching loss

۲-۵-۶- مدل ALICE

در حالت استاندارد شبکه GAN تنها نگاشت یک طرفه از فضای نهفته به فضای داده بدست می‌آورد، یعنی فاقد مکانیسم معکوس (از فضای داده به فضای نهفته) است و این امر مانع می‌شود که این شبکه ها قادر به استنباط باشند. توانایی محاسبه تابع توزیع متغیر نهفته شرطی ممکن است برای تفسیر داده‌ها و برای برنامه‌های پایین دستی (به عنوان مثال، طبقه‌بندی متغیر نهفته) مهم باشد.

تلاش‌های زیادی برای یادگیری همزمان یک مدل دو طرفه کارآمد برای تولید نمونه‌هایی با کیفیت بالا برای هر دو فضای نهفته و داده صورت گرفته است. در میان این طرح‌ها، یکی از طرح‌ها که به موفقیت چشم‌گیری دست یافته است، شبکه یادگیر استنباط خصمانه ALI است[14]. همان‌طور که شرح داده شد، دراین مدل در یک چارچوب شبکه مولد متخاصم، شبکه تمایزگر می‌آموزد تا تفاوت بین دو توزیع توامان را تشخیص دهد.

همان‌طور که ذکر شد، با این که شبکه ALI یک رویکرد جالب و خلاقانه است، اما یک ایراد اساسی دارد؛ این که بازسازی‌های صورت گرفته از داده‌ها در بعضی موارد حتی به داده های اصلی شبیه هم نیستند. دلیل این امر این است که شبکه ALI تنها به دنبال مطابقت دو توزیع توامان است، اما همبستگی بین دو متغیر تصادفی شرطی در هر یک از این توابع مشخص و اعمال نمی‌شود. در نتیجه حاصل، راه‌حل‌هایی می‌شود که هدف ALI را برآورده سازند اما در بازسازی داده‌های مشاهده شده با مشکل روبرو هستند. این شبکه هم‌چنین مشکلاتی در کشف رابطه صحیح جفت‌ها در زمان تغییر دامنه دارد[۳۷].

۲-۵-۶-۱- یادگیری تقابلی با اندازه‌گیری اطلاعات

به یاد داریم که تابع هدف در شبکه ALI به صورت معادله ۲-۲۰ بود:

$$\min_{G,E}\max_{D} V(D,G) = \mathbb{E}_{q(x,z)}[\log D(x,E(x))] + \mathbb{E}_{p(x,z)}[\log(1 - D(G(z),z)] \qquad (۲۰-۲)$$

نقطه تعادل این معادله هنگامی است که $q(x,z) = p(x,z)$ باشد. ارتباط بین متغیرهای تصادفی z و x توسط ALI محدود و مقید نشده است. در نتیجه، این امکان وجود دارد که توزیع همسان $p(x,z) = q(x,z)$ برای یک کاربرد خاص نامطلوب باشد. در واقع بسیاری از کاربردها به ثبات چرخه و وجود یک نگاشت معنی‌دار دو طرفه بین دامنه ها احتیاج دارند.

جهت مقابله با مشکل توزیع‌های نامطلوب اما برابر، بر روی راه‌حل‌های شبکه ALI باید محدویتی بر روی توزیع های $q(x,z)$ و $p(x,z)$ اعمال شود. این کار با کنترل "عدم قطعیت" بین جفت متغیرهای تصادفی، یعنی z و x با استفاده از آنتروپی‌های شرطی انجام می‌شود.

۲-۵-۶-۲- آنتروپی شرطی[۶۶]

آنتروپی شرطی یک معیار نظریه اطلاعاتی است که عدم قطعیت متغیر تصادفی x را به شرط متغیر Z با کمک توزیع توامان $\pi(x\ ،\ z)$ تعیین می‌کند:

$$H_\pi(x|z) \cong -E_\pi(x,z)[log\ \pi(x|z)]$$
$$H_\pi(z|x) \cong -E_\pi(x,z)\ [log\ \pi(z|x)] \qquad (۲۱-۲)$$

عدم قطعیت متغیر x به شرط متغیر Z با $H_\pi(x \mid z)$ مرتبط است. در حقیقت، اگر $H_\pi(x \mid z) = 0$ باشد در این صورت x به طور قطعی به Z وابسته می‌باشد. به کمک کنترل میزان عدم قطعیت $q(z \mid x)$ و $p(x \mid z)$، می توان راه حل های ALI را در توزیع های توامانی که نگاشت آن ها منجر به نتایج بهتری می شود، محدود کرد. در نهایت با افزودن یک عامل تنظیم کننده آنتروپی شرطی، به تابع هدف زیر دست می یابیم:

$$V_{ALICE}(D_{xz}\ ,E\ ,G) = V(D_{xz}\ ,E\ ,G) + V_{CE}(E\ ,G) \qquad (۲۲-\ ۲)$$

$V_{CE}(E\ ,G)$ وابسته به متغیرهای تصادفی توزیع های توامان است. در حالت ایده آل، پس از شناسایی تمام نقاط تعادل تابع هدف ALI ، می‌توان با محاسبه آنتروپی شرطی آن‌ها، راه‌حل مطلوب را انتخاب کنیم. با این حال، در عمل این راه غیرقابل استفاده است، زیرا ما از قبل به نقاط تعادل دسترسی نداریم. در ادامه یک راه‌حل برای محاسبه آنتروپی شرطی ارائه می‌شود.

[66] Conditional entropy

۲-۵-۶-۳- فرایند یادگیری

در نبود تابع توزیع احتمال صریح که برای محاسبه آنتروپی شرطی مورد نیاز است، می‌توان حدود آنتروپی شرطی را با استفاده از معیار ثبات چرخه[۶۷] محدود کرد. در این‌جا برای بازسازی $\hat{x}$ به طریق زیر عمل می شود:

$$\hat{x} \sim p(\hat{x} \mid z) \,, z \sim q(z \mid x) \,, x \sim q(x) \qquad (۲۳-۲)$$

به کمک روال تولید بالا، تلاش می شود تا $\hat{x}$ با احتمال بالایی شبیه x اصلی باشد. اثبات می‌شود که به کمک این روال تولید $\hat{x}$ ها، حد بالای آنتروپی شرطی $V_{CE}(E\,,G)$ می‌باشد.

نکته حائز اهمیت این است که می‌توان عامل تنظیم آنتروپی را به تابع هدف شبکه ALI، بدون اعمال تغییرات اضافی دیگری، در روال آموزش این شبکه اضافه کرد. بدین ترتیب تابع بهینه‌سازی برای شبکه ALICE به‌صورت ۲-۲۴ خواهد بود.

$$\begin{aligned}\min_{E,G}\max_{D_{xz},D_{xx}} V_{ALICE} \\ = \mathrm{V_{ALI}} + \mathrm{E}_{x\sim q(x)}[\log \mathrm{D}_{xx}(x,x) + \log 1 \\ - \mathrm{D}_{xx}(x, \mathrm{G}(\mathrm{E}(x)))]\end{aligned} \qquad (۲۴-۲)$$

ویژگی پایداری چرخش در مقالات پیش از نیز وجود داشته است این ویژگی در این مقالات به کمک نرم درجه[۶۸] یک و دو و داده‌های واقعی مانند تصاویر محاسبه شده است. وجود تابع اتلاف بر اساس نرم درجه ۲ مبتنی بر پیکسل، سبب می‌شود که نمونه‌های خروجی این شبکه‌ها تصاویر تاری باشند. به همین علت در این شبکه از یک شبکه تمایزگر که اختلاف بین X ها و $\widehat{\mathrm{X}}$ های بازسازی‌شده را اندازه‌گیری می‌کند، استفاده شده است.

۲-۵-۶-۴- مزایا و معایب

همانطور که بررسی شد در این کار با استفاده از کدگذار تلاش شد تا شرط پایداری در شبکه‌های عصبی مولد برقرار شود به بیان دیگر اگر تصویر ورودی را به کدگذار بدهیم و نقطه متناظر در فضای نهفته را

[67] Cycle Consistency

[68] L-norm

بدست آوریم و سپس آن را به عنوان ورودی به شبکه مولد بدهیم انتظار داریم نتیجه نهایی شبیه به تصویر اولیه باشد. علی‌رغم عملکرد مناسب این مدل هنوز توزیع‌های توامی وجود دارد که از آن‌ها استفاده نشده است و همین امر سبب می‌شود تا از تمامی اطلاعات موجود استفاده نشود.

۲-۵-۷- مدل RCGAN[۶۹]

این مقاله [38] با تمرکز بر ارائه ساختاری مبتنی بر شبکه‌های عصبی تقابلی به منظور هر چه بهتر کردن تشخیص نمونه‌های ناهنجار به وسیله پوشش تمام فضای نهفته و فضای داده ورودی، در سال ۲۰۲۰ در ادامه مقاله ALICE ارائه شد. اساس این کار بر پایه تعریف تابع جریمه، بیان جدیدی از تابع هزینه و همچنین استفاده نوآورانه از تمایزگر در مسئله تشخیص ناهنجاری است. عملکرد مناسب این مدل در خلال بررسی نتایج آن قابل مشاهده است. در بخش‌های بعدی به بررسی جزئیات بیشتر این مدل می‌پردازیم.

۲-۵-۷-۱- منظم‌سازی شبکه مولد و تمایزگر

بیشتر مدل‌هایی که اخیرا به تشخیص ناهنجاری به وسیله شبکه‌های عصبی تقابلی پرداخته‌اند مبتنی بر شبکه‌های عصبی دو طرفه هستند، اگرچه این مدل‌ها قادرند تا تصاویر هنجار را با امتیاز پایین ناهنجاری بازسازی کنند اما هیچ ضمانتی در اختصاص نمره بالای ناهنجاری به نمونه‌های ناهنجار وجود ندارد.

برای رفع این محدودیت‌ها و قادر ساختن شبکه‌های عصبی تقابلی برای تشخیص نمونه‌های ناهنجار از هنجار در این کار توزیع جریمه $t(x)$ به گونه‌ای تعریف می‌شود که $x \sim t(x)$ باشد، نمونه‌های تولیدی از این توزیع باید به عنوان داده تقلبی توسط تمایزگر شناخته شود. تابع هدف مدل پیشنهادی مطابق معادله ۲-۲۵ به صورت زیر است.

$$\begin{aligned}\min_{E,G}\max_{D_{xz}} V_{\text{ano}}(D_{xz},G,E) &= \mathbb{E}_{x\sim q(x)}[\log\ D_{xz}(x,E(x))] \\ &+\mathbb{E}_{z\sim p(z)}[\log(1-D_{xz}(G(z),z))] \\ &+\mathbb{E}_{x\sim t(x)}[\log(1-D_{xz}(x,E(x)))]\end{aligned} \qquad (۲۵-۲)$$

[69] Regularized Cycle onsistent Generative Adversarial Network for anomaly detection

تابع توزیع $t(x)$ یک تابع چگالی احتمال مانند توزیع گاوسی انتخاب می‌شود. توزیع گاوسی کاربردهای زیادی در زمینه آموزش تقابلی دارد. در انتهای آموزش، شبکه‌های مولد و تمایزگر متمایل به توزیع داده‌های هنجار خواهد بود. در مدل پیشنهادی این مقاله نمونه $G(E(x))$ که در آن x نمونه ناهنجار است، نزدیک به توزیع داده هنجار خواهد بود و بنابراین بازسازی x با خود x فاصله خواهد داشت. این فاصله سبب می‌شود تا تشخیص نمونه ناهنجار آسان شود. همانطور که در شکل ۳-۳ مشخص است، نمونه‌های ناهنجار به گونه‌ای در مدل هدایت می‌شوند که بازسازی آن‌ها به سمت داده هنجار صفر متمایل شود. چنین رویکردی سبب می‌شود تا میان داده‌های ناهنجار و بازسازی آن‌ها تفاوت زیادی حاصل شود و در نتیجه تشخیص نمونه‌های ناهنجار گارانتی شود. قابل توجه است که الگوریتم ارائه شده سبب می‌شود تا نقاطی که $q(x)$ در آن‌ها کوچک است پوشش داده شود و این تنظیم مستقل از ارتباط میان $t(x)$ و توزیع نمونه‌های ناهنجار انجام می‌شود.

۲-۵-۷-۲- پایداری چرخه

به منظور ارضای شرط پایداری چرخه تمایزگر D_{xx} به معماری مورد نظر اضافه شده است. تابع هدف بخش چرخه پایداری همانند معادله ۲-۲۶ است.

$$\min_{E,G}\max_{D_{xx}} V_{\text{cycle}}(D_{xx},G,E) = \mathbb{E}_{x\sim q(x)}[\log D_{xx}(x,x)] + \mathbb{E}_{x\sim q(x)}[\log(1-D_{xx}(x,\hat{x}))] \qquad (۲۶-۲)$$

در معادله ۲-۲۶ $\hat{x} = G(E(x))$ بازسازی داده ورودی x است. تابع هدف کامل ارائه شده در این کار در نهایت مطابق معادله ۲-۲۷ به شکل زیر است.

$$\min_{E,G}\max_{D_{xz},D_{xx}} V_{\text{ano}}(D_{xz},G,E) + V_{\text{cycle}}(D_{xx},G,E) \qquad (۲۷-۲)$$

پس از آموزش مدل نوبت به محاسبه امتیاز ناهنجاری می‌رسد، در مدل پشنهادی این کار تابع امتیاز پیشنهادی مطابق معادله زیر است.

$$A(x) = 1 - D_{xx}(x, G(E(x))) \qquad (۲۸-۲)$$

امتیاز ناهنجاری $A(x)$ بیانگر میزان کیفیت بازسازی x است. الگوریتم ارائه شده در این کار مدل را مجبور به تولید خطای بزرگ برای نمونه‌های ناهنجار می‌کند در حالی که تابع هدف چرخه پایداری مدل مجبور به تولید بازسازی مناسب برای نمونه‌های هنجار می‌کند. این اختلاف امتیاز میان نمونه‌های هنجار و ناهنجار معیار مناسبی برای تشخیص نمونه‌های ناهنجار است.

۲-۵-۸- مدل ALAD

در این بخش، یک روش تشخیص ناهنجاری مبتنی بر شبکه مولد متخاصم را بررسی می‌کنیم که در زمان آزمون بسیار کارآمد است. در این روش به طور هم‌زمان یک شبکه کدگذار را در حین آموزش فرا می‌گیرد و بدین ترتیب استنتاج سریع‌تر و کارآمدتر را در زمان آزمون امکان‌پذیر می‌کند. علاوه بر این در شبکه معرفی‌شده، تکنیک‌هایی که اخیرا برای بهبود بیشتر شبکه کدگذار و تثبیت آموزش شبکه مولد متخاصم ترکیب‌شده و نشان داده شده که این تکنیک‌ها عملکرد و کارایی را در کاربرد تشخیص ناهنجاری بهبود می‌بخشند. آزمایشات روی طیف وسیعی از داده‌های جدولی و تصویری، کارایی و اثربخشی این رویکرد را در عمل نشان می‌دهد[31].

همان‌طور که پیش از این گفته شد، شبکه‌های GAN استاندارد از نمونه‌گیری کارآمد پشتیبانی می‌کنند و روش‌های مختلفی وجود دارد که می‌تواند آن‌ها را برای تشخیص ناهنجاری تطبیق دهد. به عنوان مثال، برای یک نقطه داده x، می‌توان از نمونه‌گیری استفاده کرد تا احتمال ناهنجار بودن x را تخمین زند. تخمین دقیق احتمال به تعداد زیادی نمونه نیاز دارد و در نتیجه محاسبه احتمال، بار محاسباتی سنگینی دارد.

روش دیگر معکوس کردن[۷۰] شبکه مولد برای یافتن متغیرهای نهفته z است که به معنای به حداقل رساندن خطای بازسازی با تابع هدف گرادیان نزولی تصادفی می‌باشد. این روش هم‌چنین از نظر محاسباتی بسیار پرهزینه است زیرا هر محاسبه گرادیان نیاز به یک پس انتشار از طریق شبکه مولد دارد.

به واسطه بهره‌وری محاسباتی بالا و قابلیت مدل‌سازی داده‌های ابعاد بالا، از شبکه‌های مولد متقابلی به همراه یک شبکه کدگذار E (که نمونه‌ها را از فضای داده x به فضای نهفته z نگاشت می‌کند) استفاده

[70] Invert

می‌شود. نمایش نهفته هر نمونه از فضای داده در چنین مدل‌هایی صرفاً با عبور از شبکه کدگذار انجام می‌شود. هم‌چنین این مدل پیشرفت‌های اخیر که برای بهبود شبکه کدگذار صورت گرفته مانند افزودن یک شبکه تمایزگر برای بهبود سازگاری چرخه $(G(E(x)) \approx x)$ را شامل می‌شود.

پیش از این توضیح داده شد که شبکه ALI توزیع توامان داده‌ها را به همراه یک شبکه کدگذار مدل می کند. این مدل یک شبکه تمایزگر D_{xz} دارد که z و x را به عنوان ورودی می‌گیرد و بررسی می‌کند که این جفت ورودی از کدام منبع – شبکه مولد و یا شبکه کدگذار – تولید شده است.

با این‌که به لحاظ نظری توزیع توامان شبکه مولد و شبکه کدگذار به یک نقطه میل می‌کند، اما در عمل اغلب نتیجه یکسان نیست و لزوما به یک نقطه همگرا نمی‌شوند و این پدیده سبب نقض پایداری چرخه می‌شود. نبود پایداری چرخه به این معناست که $G(E(x)) \not\approx x$ باشد. این مشکل برای روش های تشخیص ناهنجاری مبتنی بر بازسازی چالش‌های جدی ایجاد می‌کند. برای حل این مشکل، چارچوب ALICE پیشنهاد می‌کند که آنتروپی شرطی را با افزودن یک شبکه تمایزگر بین متغیر x و بازسازی آن به روش تقابلی برای سازگاری چرخه تقریب بزنیم.

۲-۵-۸-۱ تابع هزینه

برای تثبیت آموزش در مدل پایه ALICE، توزیع‌های شرطی را با اضافه‌کردن یک قید آنتروپی شرطی دیگر تنظیم می‌کنیم و سپس عملیات نرمال‌سازی طیفی را انجام می‌دهیم.

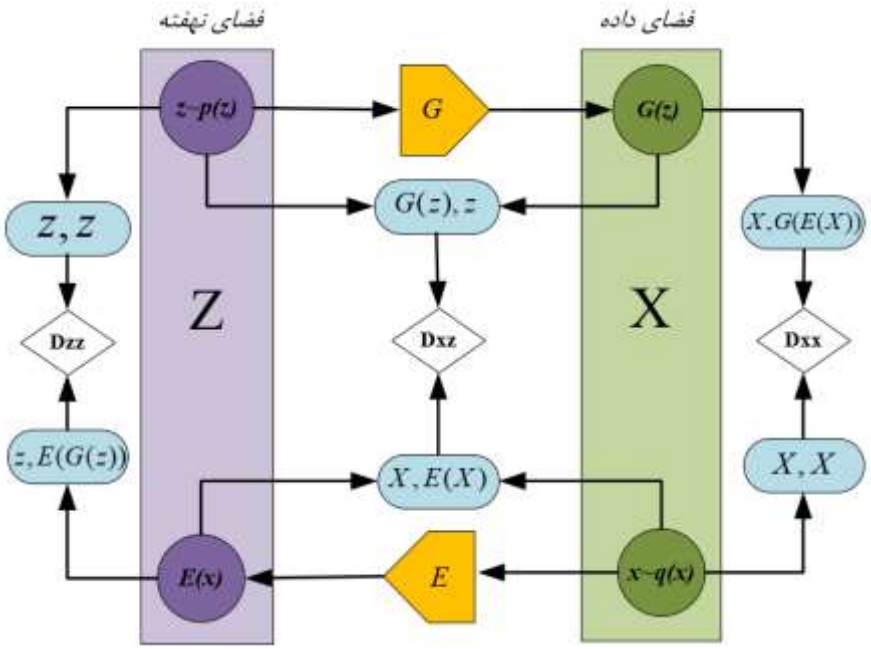
توضیح دقیق‌تر این‌که، در این‌جا فضای نهفته شرطی $\mathrm{H}^{\pi}(\mathrm{z} \mid \mathrm{x}) = -\mathrm{E}_{\pi(x,z)}[\log \pi(\mathrm{z} \mid \mathrm{x})]$ را با یک شبکه تمایزگر مخالف دیگر D_{zz} با نقطه تعادل مشترک تنظیم می‌کنیم که، مطابق افزودن معادله ۲-۲۹ به تابع هزینه در چارچوب ALICE می‌باشد.

$$\begin{aligned} \mathrm{V}(D_{zz}, \mathrm{G}, \mathrm{E}) = {} & V_{ALICE} + \mathbb{E}_{z\sim \mathrm{p(z)}}[log(D_{zz}(z,z)) \\ & + \mathbb{E}_{z\sim \mathrm{p(z)}}[log(1 - D_{zz}(z, G(E(z))))] \end{aligned} \qquad (۲۹-۲)$$

با کنار هم قرار دادن تمامی این اجزا، در نهایت تابع هزینه این شبکه مطابق معادله ۲-۳۰ خواهد بود. شبکه ALAD تلاش می‌کند تا نقطه تعادل این مسئله را آموزش ببیند.

$$min_{G,E} max_{D_{xz},D_{xx},D_{zz},} V_{ALAD}(D_{xz}, D_{xx}, D_{zz}, E, G) \quad (۳۰-۲)$$
$$= \mathbb{E}_{z\sim p(z),x\sim q(x)}[log\left(D_{xz}(x, E(x))\right)$$
$$+ log(1 - D_{xz}(G(z), z)) + \mathbb{E}_{x\sim q(x)}[log(D_{xx}(x, x))$$
$$+ log(1 - D_{xx}(x, G(E(x)))) + \mathbb{E}_{z\sim p(z)}[log(D_{zz}(z, z))$$
$$+ log(1 - D_{zz}(z, E(G(z))))]$$

در نهایت معماری کلی شبکه ALAD به صورت شکل ۲-۶ خواهد بود.



شکل ۲-۵: شمای کلی شبکه ALAD [۳۲].

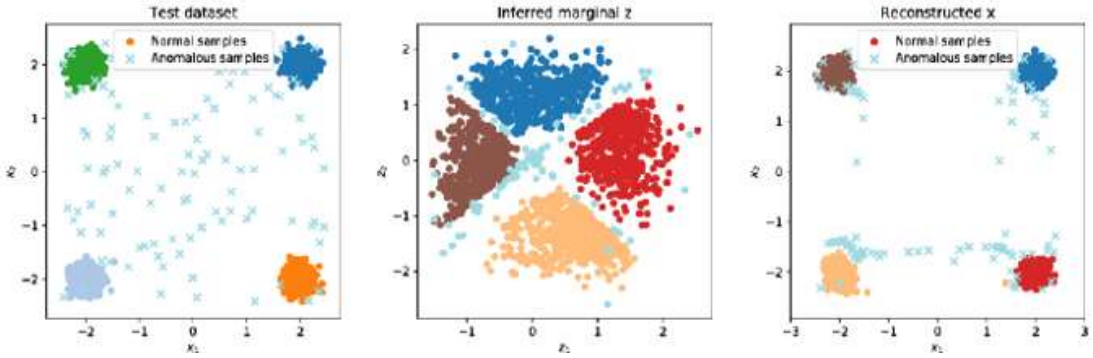
اضافه کردن مرحله نرمال‌سازی طیفی با انگیزه یادشده مقاله [39] می‌باشد. در این مقاله نشان داده‌شده که با افزودن قیود لیپسچیتز[۷۱] به تمایزگر شبکه GAN، فاز آموزش تثبیت خواهد شد. در عمل نشان داده شده که با تنظیم مجدد پارامترهای وزن، بهبود بسیار خوبی روی عملکرد شبکه خواهیم داشت. بدین صورت که بزرگ‌ترین مقادیر ویژه ماتریس وزن را در شبکه تمایزگر ثابت نگه داریم. این روش از نظر محاسباتی کارآمد است و هم‌چنین آموزش را تثبیت می‌کند. با آزمایش‌های صورت‌گرفته نشان داده شد

[71] Lipschitz Constraints

که افزودن این قیود، نه تنها برای شبکه تمایزگر، بلکه برای شبکه کدگذار نیز سودمند است. قابل توجه است که مدل ALICE شامل این مرحله نمی باشد.

۲-۵-۸-۲ تشخیص ناهنجاری

شبکه ALAD یک روش تشخیص ناهنجاری مبتنی بر بازسازی است و بدین صورت عمل می‌کند که فاصله نمونه از بازسازی را توسط شبکه GAN ارزیابی می‌کند. نمونه‌های عادی باید به طور دقیق بازسازی شوند در حالی که نمونه‌های ناهنجار احتمالاً به طور ضعیف‌تری بازسازی می‌شوند. نحوه تشخیص ناهنجاری در شکل ۲-۶ نشان شده است.



شکل۲-۶: نمونه‌ای از خروجی شبکه ALAD به همراه داده‌های ناهنجار[۳۲].

در شکل ۲-۶ ضربدرها نمونه‌های ناهنجار و دایره‌های رنگی نمونه‌های عادی هستند. همان‌طور که دیده می‌شود شبکه ALAD تا حدی توانسته برای داده‌های ناهنجار بازسازی ضعیف داشته باشد

مؤلفه کلیدی دیگر ALAD نمره ناهنجاری است که فاصله بین نمونه‌های اصلی و بازسازی آن‌ها را اندازه‌گیری می‌کند. انتخاب اولیه‌ای که به ذهن می‌رسد، فاصله اقلیدسی بین نمونه‌های اصلی و بازسازی آن‌ها در فضای داده است. اما، این معیار ممکن است معیار مطمئنی برای اندازه‌گیری تشابه نباشد. به عنوان مثال، این معیار در مورد تصاویر می‌تواند بسیار پرخطا باشد؛ زیرا تصاویر با ویژگی‌های تصویری مشابه الزاماً از نظر فاصله اقلیدسی نزدیک به یک‌دیگر نیستند. معیار تعریف‌شده در این روش از فاصله بین نمونه‌ها در فضای ویژگی‌های تمایزگر D_{xx} محاسبه می‌شود، که توسط لایه قبل از لاجیت تعریف شده است. از این ویژگی‌ها هم‌چنین به عنوان کدهای CNN یاد می‌شود. به‌طور دقیق‌تر می‌توان گفت با آموزش یک مدل

برای داده‌های عادی و محاسبه E، G، D_{xx}، D_{xz} و D_{zz} یک تابع نمره‌دهی را بر اساس خطای بازسازی نرم۱ مطابق معادله ۲-۳۱ تعریف می‌شود. در این تعریف $f(.,.)$ تابع فعال‌ساز[۷۲]های لایه قبل از لاجیت و یا همان کد CNN می‌باشد. این نوع تعریف A به ما این اطمینان را می‌دهد که نمونه به درستی کدگذاری و بازسازی‌شده و در نتیجه از توزیع داده واقعی می‌باشد.

$$A(x) = ||f_{xx}(x,x) - f_{xx}\left(x, G\left(E(x)\right)\right)||_1 \qquad (۳۱-۲)$$

با این تعریف، نمونه‌ها بامقدار بیش‌تر A با احتمال بالاتری داده ناهنجار خواهند بود. در ادامه در الگوریتم ۲-۳ روال محاسبه $A(X)$ ارائه می‌شود.

الگوریتم محاسبه امتیاز ناهنجاری مدل ALAD

ورودی $x \sim p_{X_{Test}}(x), E, G, f_{xx}$ **لایه ویژگی مربوط به تمایزگر** D_{xx}

خروجی $A(x)$

انجام روال استنتاج

1. $\tilde{z} \leftarrow E(x)$ کدگذاری نمونه
2. $\hat{x} = G(\tilde{z})$ کدگشایی نمونه
3. $f_\delta \leftarrow f_{xx}(x, \hat{x})$
4. $f_\alpha \leftarrow f_{xx}(x, x)$
5. **بازگرداندن** $||f_\delta - f_\alpha||_1$
6. **اتمام روال محاسبه نمره ناهنجاری**

الگوریتم۲-۳: شبه کد الگوریتم ALAD

معیار استفاده شده در این‌جا از ایده تطابق ویژگی‌های از دست‌رفته الهام گرفته شده است[40]. اما در این جا به جای استفاده از ویژگی‌های محاسبه شده درشبکه تمایزگر GAN استاندارد (که اختلاف را بین نمونه های تولید شده و داده‌های واقعی را محاسبه می‌کند)، از ویژگی‌های محاسبه‌شده در شبکه تمایزگر D_{xx}

[72] Activation

استفاده می‌شود. هم‌چنین در این‌جا به جای استفاده از این معیار در حین آموزش شبکه GAN، از این معیار در هنگام روال استنتاج بهره می‌جوییم.

سوالی که در این‌جا مطرح می‌شود این است که : چرا نباید از خروجی تمایزگر D_{xx} به عنوان معیار فاصله استفاده کرد. پاسخ این سوال بدین صورت است که هدف از شبکه تمایزگر D_{xx} تمایز بین یک جفت نمونه واقعی (x, x) و بازسازی آن $(x, G(E(x)))$ می‌باشد و شبکه کدگذار و شبکه مولد داده‌های واقعی و توزیع متغیر نهفته را کاملاً ضبط خواهند کرد. در این حالت D_{xx} قادر به تفکیک بین نمونه‌های واقعی و نمونه‌های بازسازی شده نخواهد بود و بدین ترتیب یک پیش بینی تصادفی را تولید می‌کند که معیار ناهنجاری مناسبی نخواهد بود.

۲-۶- جمع‌بندی

در این فصل ابتدا به دسته‌بندی روش‌های مختلف تشخیص ناهنجاری پرداختیم، در خلال همین دسته‌بندی ها برخی مدل‌های به نسبت قدیمی‌تر و سنتی بررسی شدند. در ادامه اهمیت جایگاه شبکه‌های عصبی تقابلی روشن شد. در گام بعدی معیارهای ارزیابی مدل‌های تشخیص ناهنجاری معرفی شده‌اند. در ادامه پس از بررسی مدل پایه شبکه عصبی تقابلی، زنجیره کارهایی که روی شبکه‌هایی عصبی تقابلی به منظور بهبود صورت گرفته است، بررسی شده است. در ادامه مدل AnoGan شرح داده شد و با توجه به مشکل آن در نگاشت معکوس از فضای داده ورودی به فضای نهفته، مدل f-AnoGan که در ادامه کار قبلی است، بررسی شد. با توجه به ضعف‌های موجود در ساختار f-AnoGan، مقاله مکمل این مدل یعنی شبکه ALI مرور شد. در گام بعدی مدل EGBAD که جزو اولین کارها در زمینه تشخیص ناهنجاری که با الهام از مدل ALI خلق شده است مرور شد. در مرحله بعدی با توجه تضمین نشدن شرط سازگاری حلقه در ALI مقاله ALICE مورد بررسی دقیق‌تر قرار گرفت. در ادامه این روش دو شبکه دیگر ارائه شده است، شبکه RCGAN که ایده اصلی آن متمایل سازی بیشتر توزیع شبکه مولد و کدگذار به سمت توزیع داده‌های هنجار است و شبکه ALAD با هدف تضمین بیشتر چرخه پایداری، یک تمایزگر بین متغیر z و بازسازی آن توسط شبکه، به ساختار شبکه ALICE اضافه کرد.

در فصل بعد روش پیشنهادی RCALAD که با تمرکز بر بکارگیری حداکثری جریان اطلاعات موجود در شبکه و همچنین تاکید بر بازسازی ضعیف نمونه‌های ناهنجار، به تفضیل ارائه خواهد شد.

فصل سوم:
روش پیشنهادی

در فصول قبل به بررسی روش‌های تشخیص ناهنجاری با رویکردهای مختلف پرداختیم. همان‌طور که پیش‌تر ذکر شد، با پیشرفت روزافزون زیرساخت‌های محاسباتی و افزایش توان پردازشی، علاقه و توجه محققان حوزه هوش مصنوعی به سمت شبکه‌های عصبی جلب شد. در میان این انواع شبکه‌های عصبی ارائه شده در چند سال اخیر، شبکه های مولد تقابلی به نتایج قابل توجه و درخشانی در کاربردهای مختلف دست یافته‌اند. علی‌رغم نتایج قابل دفاع این نوع شبکه‌ها در زمینه‌های پردازش تصویر[13]، پردازش گفتار و پردازش متن[41] در حوزه تشخیص ناهنجاری آن طور که شایسته است بدان توجه نشده است. در زمینه شناسایی نمونه‌های ناهنجار در دادگان دنیای واقعی به ندرت می‌توان الگوریتمی یافت که بر مبنای شبکه‌های مولد تقابلی طراحی شده باشد؛ در حقیقت بیشتر ساختارهای تقابلی برای کاربردهای دیگر طراحی شده است و صرفا همان ساختار در کاربرد تشخیص ناهنجاری مورد استفاده قرار گرفته است، با این حال همین مدل‌های عام منظوره، توانسته‌اند به نتایج نسبتا معقولی دست یابند[32].

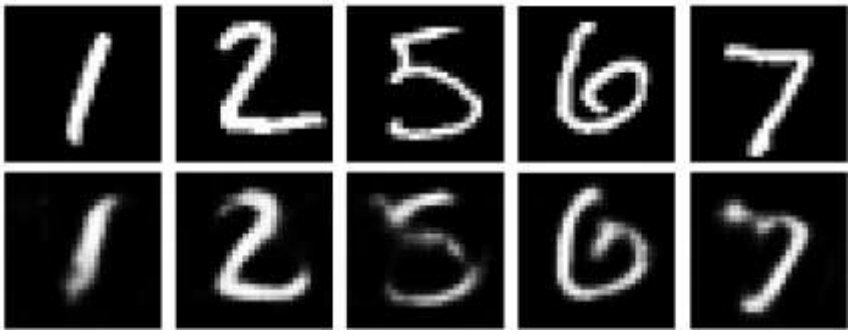
همان‌طور که پیش از این گفته شد، پس از معرفی شبکه GAN اولیه، مقاله AnoGAN از این ساختار برای کاربرد تشخیص ناهنجاری استفاده کرد. مشکل اصلی این روش، پیچیدگی زمانی بالای روش پیشنهادی به دلیل استفاده از یک ساختار مبتنی بر تکرار[1] برای یافتن نگاشت معکوس از فضای داده به فضای نهان بود. در سال بعد شبکه ALI دقیقا با همین هدف، یعنی دستیابی به نگاشت از فضای x به فضای z ارائه شد. در این روش یک شبکه کدگذار همزمان با شبکه مولد آموزش داده می‌شود. این شبکه توانست تا حدی مشکل نگاشت معکوس را برطرف کند. در این میان مشکل ساده اما جدی چرخه پایداری ایجاد شد. برای حل این مشکل، شبکه ALICE پیشنهاد افزودن یک تمایزگر بین تصاویر اولیه و بازسازی آن‌ها را داد. شبکه ALAD با افزودن یک تمایزگر دیگر بین نگاشت‌های تصاویر در فضای نهان و z های بازسازی شده آن‌ها، تلاش کرد تا مفهوم چرخه پایداری را در فضای نهان نیز برقرار کند. اشکال این روش در فرض استقلال بین دو چرخه پایداری فضای داده ورودی و چرخه پایداری فضای نهان بود. تمایزگر افزوده شده در مدل ALAD در واقع در یک چرخه مستقل از چرخه پایداری فضای داده ورودی قرار دارد در حالی که این دو چرخه کاملا به یکدیگر وابسته می‌باشند. این مشکل در بخش ۳-۳-۱ تحت عنوان مشکل CCC[2] با جزئیات بیشتر شرح داده خواهد شد. در واقع در این مدل از همه اطلاعات در دسترس

[1] Iterative

[2] Complete Cycle Consistency

برای بهبود روند آموزش و تقویت ساختار تقابلی استفاده نشده است، همانطور که می‌دانیم روند تبدیل داده با نگاشت از فضای داده ورودی به فضای نهفته آغاز می‌شود و در ادامه به فضای داده ورودی بازگردانده می‌شود تا با استفاده از میزان تفاوت داده ورودی و داده بازسازی شده امتیاز ناهنجاری محاسبه شود. این ایده در مدل CALAD[3] ارائه شد و در بخش نتایج عملی نشان داده شد که استفاده از اطلاعات فضای نهفته و فضای داده ورودی به صورت توام در یک چرخه کامل، سبب بهبود عملکرد و پایداری شبکه‌های عصبی تخاصمی می‌شود. قابل توجه است که تاکنون در هیچ یک از کارهای قبلی، از اطلاعات موجود در یک چرخه کامل برای آموزش شبکه استفاده نشده است.

علاوه بر این، در مدل پایه ALAD فرض شده است که با آموزش انجام گرفته بر روی داده هنجار، بازسازی نمونه‌های ناهنجار لزوما بازسازی ضعیفی خواهد بود در صورتی که هیچ قیدی برای تاکید بر این مهم و متمایل کردن مدل به سمت تولید بازسازی ضعیف در نظر گرفته نشده است. علی‌رغم نتایج قابل قبولی که مدل‌های پیشین ارائه داده‌اند، اما در تمامی آن‌ها با سهل انگاری فرض شده که اگر یک شبکه بر روی داده‌های هنجار آموزش ببیند، لزوما برای داده‌های هنجار بازسازی خوب و برای ناهنجارها بازسازی ضعیف دارد. اما این فرض لزوما برقرار نیست و ممکن است شبکه برای تصاویر ناهنجار هم بازسازی نزدیک به تصویر ورودی داشته باشد، همانند تصویر ۳-۱ که در آن کلاس هنجار کلاس صفر می‌باشد و بقیه کلاس‌ها، کلاس ناهنجاری به حساب می آیند و همان‌طور که مشاهده می‌کنید، مدل در بازسازی نامناسب نمونه‌های ناهنجار ضعیف عمل کرده و بازسازی بسیار نزدیک و شبیه به کلاس ورودی داشته است.



شکل ۳-۱: بازسازی نامطلوب نمونه ناهنجار.

[3] Complete Adversarially Learned Anomaly Detection

قابل توجه است که این بازسازی نزدیک تصویر ورودی اولیه، عملا فرض ابتدایی برای روش‌های تشخیص ناهنجاری مبتنی بر بازسازی را نقض کرده و روند تفکیک داده‌ها را مختل خواهد کرد و بدین ترتیب دیگر با این روش داده هنجار و ناهنجار از یکدیگر قابل شناسایی نخواهند بود.

در این بخش همچنین مدل RALAD[۴] که مبتنی بر شبکه‌های مولد تقابلی با هدف مقیدسازی مدل برای داشتن بازسازی ضعیف برای داده‌های ناهنجار معرفی می‌شود. در نهایت به کمک ترکیب هر دو ایده و با هدف معرفی یک چارچوب قوی و جامع برای تمامی کاربردهای تشخیص ناهنجاری، مدل RCALAD[۵] معرفی شده است. در طراحی این روش تمرکز اصلی بر روی ارائه مدلی است که بتواند در کاربردهای دنیای واقعی نظیر داده‌های پزشکی مورد استفاده قرار بگیرد. مانند بسیاری از الگوریتم‌های مبتنی بر یادگیری، در این جا دو مرحله اصلی آموزش و آزمایش وجود دارد. در قسمت آموزش همانند دیگر چارچوب‌های تقابلی به نوبت بخش مولد و بخش تمایزگر را آموزش می‌دهیم تا هر دو بخش در عین تناسب به نوبت به‌روزرسانی شود. همچنین یک مرحله پیش‌پردازش شامل نرمال‌سازی تصاویر به منظور افزایش دقت مدل نهایی انجام شده است. در مرحله آموزش نمونه‌های ناهنجار را از تصاویر آموزشی حذف می‌کنیم و فقط از آن‌ها در مرحله آزمایش برای ارزیابی مدل پیشنهادی استفاده می‌کنیم. به بیان دیگر مدل تنها توزیع داده‌های عادی را می‌آموزد؛ روند آموزش مدل باید به حدی قدرتمند است که بتواند به خوبی فضای داده‌های ناهنجار را از داده‌های عادی تفکیک کند.

در ادامه و در قسمت تشخیص ناهنجاری، یک ورودی بدون برچسب به ساختار شبکه وارد می‌شود و به کمک اختلاف بازسازی ارائه شده توسط شبکه برای آن ورودی در هر دو فضای داده و فضای نهان، امتیاز ناهنجاری برای هر ورودی محاسبه می‌شود. با توجه به این نکته که شبکه بر روی دادگان نرمال آموزش داده می‌شود انتظار می‌رود که برای داده‌های ناهنجار بازسازی ضعیف‌تری داشته باشد و بدین ترتیب این اختلاف بیشتر شود و در نهایت این داده‌ها امتیاز ناهنجار بیشتری بگیرند. بدین ترتیب با انتخاب یک حدآستانه[۶] و یا انتخاب یک نسبت معین از میان داده‌های با بیشترین امتیاز، داده‌هایی که ناهنجار هستند شناسایی می‌شوند.

[4] Regularized Adversarially Learned Anomaly Detection

[5] Regularized Complete Adversarially Learned Anomaly Detection

[6] Threshold

در ادامه این بخش به بررسی دقیق‌تر مشکلات پیشنهادی و راه حل ارائه شده و جزئیات هر قسمت از مدل پیشنهادیCALAD، RALAD و RCALAD خواهیم پرداخت. لازم به ذکر است نتایج آزمایش‌ها بر روی هر دو نوع داده تصویر و جدولی بیانگر کارایی و اثربخشی روش‌های پیشنهادی و نمایانگر سازگاری نتایج تئوری و عملی بدست‌آمده برای این مسائل می‌باشد.

۳-۱- مدل CALAD[۷]

این بخش به معرفی اولین مدل پیشنهادی اختصاص داده شده است. روش مورد بحث به منظور تشخیص ناهنجاری به دسته روش‌های مبتنی بر بازسازی تعلق دارد و از نظر دسترسی برچسب داده‌ها همانند سایر مدل‌های مبتنی بر شبکه‌های عصبی تقابلی بررسی شده، جزو دسته الگوریتم‌های بدون نظارت به حساب می‌آید. در این روش نیز شبکه مولد موجود در ساختار GAN نگاشت از فضای نهان به فضای داده ورودی را فرا می‌گیرد، به بیان دیگر در این قسمت با استفاده از شبکه عصبی تقابلی، توزیع داده هنجار ورودی مدل می‌شود. برای یادگیری نگاشت معکوس از فضای داده ورودی به فضای نهان همانند مدل‌های پیشین از یک شبکه کدگذار استفاده شده است. در ادامه مشابه با مدل ALAD از دو تمایزگر D_{xx} و D_{zz} برای تضمین چرخه پایداری در هر دو فضای ورودی و نهان استفاده شده است.

نوآوری به کار گرفته شده در این مدل، استفاده از متغیر جدید $\hat{z}_{\hat{x}}$ و افزودن تمایزگر D_{xxzz} به ساختار ارائه شده برای تضمین شرط چرخه پایداری کامل در هر دو فضای نهفته و فضای ورودی می‌باشد. در واقع متغیر $\hat{z}_{\hat{x}}$ بازسازی شبکه از نگاشت تصویر اولیه در فضای نهان است. هدف از تعریف این متغیر جدید دستیابی به یک حلقه کامل نگاشت‌های متوالی در فضای داده ورودی و فضای نهان به صورت وابسته است. قابل توجه است که تمایزگر مورد استفاده در کارهای قبلی، تنها به توزیع‌های مستقل داده‌ها در هر دو فضای نهفته و داده ورودی توجه می‌کردند و عملا برخی از جریان اطلاعات موجود در شبکه که متعلق به توزیع توامان متغیرهای فضای داده ورودی و نگاشت متناسب آن در فضای نهفته است، بلا استفاده باقی می‌ماند.

[7] Complete Adversarialy Learned Anomaly Detection

شایان ذکر است که این روال و اثر بکارگیری این نوع تمایزگر توامان یک مرتبه پیش از این ثابت شده است؛ در واقع برای آموزش به هنگام افزودن شبکه کدگذار به ساختار اولیه GAN در هنگام معرفی مدل ALI، دو راه پیش‌رو بود. یک راه افزودن یک شبکه تمایزگر در کنار تمایزگر اولیه موجود، برای تمیز بین متغیرهای فضای پنهان بود و یک راه، تقویت شبکه تمایزگر اولیه و آموزش این شبکه به نحویست که توزیع توامان در هردو فضای نهان و واقعی را فرا بگیرند و بتوانند داده‌های آموزشی را از داده‌های تولیدشده توسط این مدل تشخیص دهد.

در مدل ارائه شده، جریان اطلاعات شامل یک فرایند دو مرحله‌ایست، به این ترتیب که ابتدا از روی داده اولیه x در فضای واقعی یک نگاشت توسط کدگذار به فضای نهان $E(x)$ انجام می‌شود و سپس از روی همین داده یک نگاشت معکوس به عنوان بازسازی به فضای داده اولیه توسط شبکه مولد $G(E(x))$ انجام می‌شود. سپس بار دیگر همین بازسازی به کدگذار فرستاده شده و در واقع بازسازی متغیر در فضای نهان $E(G(E(x)))$ بدست می‌آید. در مدل‌های قبلی متغیرها در هر کدام از فضاها با بازسازی آن‌ها به صورت جداگانه مورد بررسی قرار می‌گرفتند و مستقل از هم تمیز داده می‌شدند و بدین ترتیب زنجیره ارتباطات موجود بین این دو فضا نادیده گرفته می‌شد.

بر مبنای استفاده از همین اطلاعات از دست رفته و سابقه بکارگیری توزیع‌های توامان در این زمینه، در این جا مدل CALAD پیشنهاد شده‌است. در این ساختار در کنار چهارچوب اولیه شبکه‌های مولد تقابلی، یک شبکه تمایزگر توامان افزوده شده، تا با به‌کارگیری بیشترین اطلاعات موجود، مدل به جهت بهتری هدایت شده و سازگاری چرخه‌ها به صورت وابسته به هم بررسی شود؛ یعنی برای آموزش مدل از اطلاعات هر دوفضا به صورت توامان استفاده شود و در نهایت مدل به وزن‌های بهتر و دقت بالاتری دست یابد.

۳-۱-۱- معماری شبکه

در این بخش به معرفی تک تک اجزای مدل ارائه‌شده خواهیم پرداخت. در این ساختار همانند کارهای پیشین با هدف کاهش پیچیدگی زمانی، یک کدگذار توام با شبکه مولد در ساختار کلی شبکه عصبی تقابلی آموزش داده می‌شود. نگاشت معکوس از فضای داده ورودی به فضای نهفته به سادگی با تعبیه کدگذار E در معماری پیشنهادی به‌دست می‌آید. این کدگذار تنها از فضای داده ورودی نمونه می‌گیرد و

به طور تقریبی بازنمایی متناسب با آن را در فضای نهفته را تولید می‌کند. در اینجا برای آموزش هم‌زمان هر دو شبکه مولد و کدگذار از یک شبکه تمایزگر توامان با نام D_{xz} استفاده شده است. این تمایزگر بررسی می‌کند که جفت متغیر ورودی متعلق به توزیع داده ورودی x و نقطه متناظر با آن در فضای نهفته $E(\mathrm{x})$ است و یا توسط شبکه مولد $G(z)$ و نمونه‌گیری از فضای نهفته z تولید شده است.

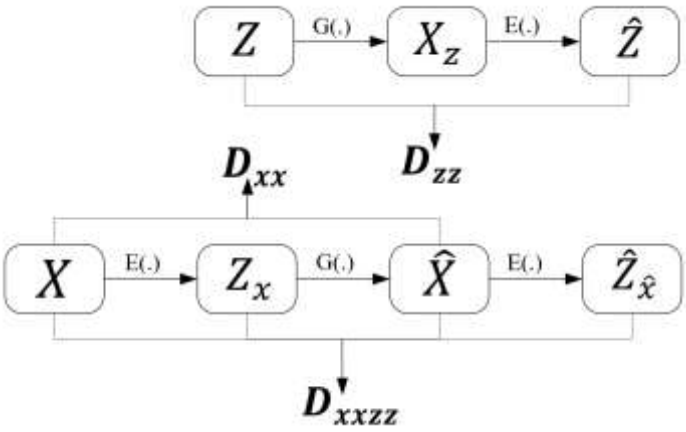
به منظور ارضای شرط پایداری حلقه در فضای داده ورودی از تمایزگر D_{xx} افزوده شده است، این تمایزگر به صورت توام نمونه داده ورودی x و نمونه بازسازی شده متناظر آن را $G(E(\mathrm{x})) = \hat{x}$ به عنوان ورودی دریافت می‌کند. همچنین برای تقویت شرط پایداری حلقه در فضای نهفته، تمایزگر D_{zz} به این مجموعه اضافه شده است. این تمایزگر شرط پایداری حلقه را در خلال روند تولید نمونه بازسازی شده چک می‌کند. ورودی این تمایزگر نمونه ورودی شبکه مولد از فضای نهفته z و نمونه بازسازی متناظر با آن در فضای نهان است.

با اضافه شدن تمایزگر D_{xxzz} به ساختار موجود، تلاش شده است تا از تمامی اطلاعات موجود در یک چرخه کامل به صورت توام استفاده شود یعنی در کنار بررسی هر دو متغیر و بازسازی آن‌ها در همان فضا، توزیع توامان چهارتایی آن‌ها در روند تشخیص نمونه ناهنجار به کار گرفته شود. این شبکه وظیفه تمییز بین نمونه‌های چهارتایی (x, x, z_x, z_x) و $(x, G(E(x)), z_x, E(G(z_x)))$ را دارد به بیان دیگر تلاش می‌کند تا x و بازسازی ارائه شده توسط شبکه و همینطور نگاشت تصویر ورودی در فضای نهان z_x و بازسازی خروجی شبکه مولد توسط کدگذار $E(G(z_x))$ تا حد امکان به یکدیگر نزدیک کند. هدف از تعبیه این تمایزگر در این ساختار حل مشکل چرخه پایداری کامل[8] یا به اختصار CCC می‌باشد. تعریف دقیق مسئله CCC در ادامه بررسی می‌شود.

بیان ریاضی مسئله CCC بدین صورت می‌باشد که به ازای هر متغیر x از فضای ورودی شبکه ابتدا کدگذار نگاشت معکوس به فضای نهفته را تخمین می‌زند که معادل $E(x) = z_x$ می‌باشد. در مرحله بعد بازنمایی بدست آمده را به شبکه مولد وارد می‌کنیم تا بازسازی شبکه از متغیر ورودی $G(z_x) = G\big(E(x)\big) = \hat{x}$ تولید کند. سپس همین بازسازی را بار دیگر به شبکه کدگذار می‌دهیم تا بازسازی در فضای نهفته را نیز محاسبه شود یعنی $E(\hat{x}) = E\big(G(z_x)\big) = \hat{z}_{\hat{x}}$. در این جریان،

[8] Compete Cycle Consistency

انتظار منطقی از هر شبکه مبتنی بر بازسازی این است که دو متغیر $\hat{x}$ و x و همچنین دو متغیر $\hat{z}_{\hat{x}}$ و z_x تا حد امکان کمترین اختلاف را داشته باشند. در مدل ALAD شباهت میان داده ورودی و بازسازی آن و همچنین شباهت z و بازسازی آن مستقل از هم و در دو چرخه جداگانه بررسی می‌شد و فرض شده بود که مستقل از هم هستند در حالی‌که می‌دانیم این دو چرخه کاملا به یکدیگر وابسته بوده و فرض استقلال این دو مسئله غلط است. در اینجا سعی شده است با بررسی توام متغیرهای موجود در چرخه CCC در تمایزگر جدید D_{xxzz} عدم استقلال میان متغیرها مدل شود و جریان اطلاعات موجود در این زنجیره برای بهبود آموزش شبکه و تشخیص هر چه بهتر داده‌های ناهنجار به کار گرفته شود. تفاوت میان ورودی تمایزگر D_{xxzz} و ورودی تمایزگر D_{zz} که در مدل ALAD استفاده شده است در شکل ۳-۲ قابل مشاهده است.

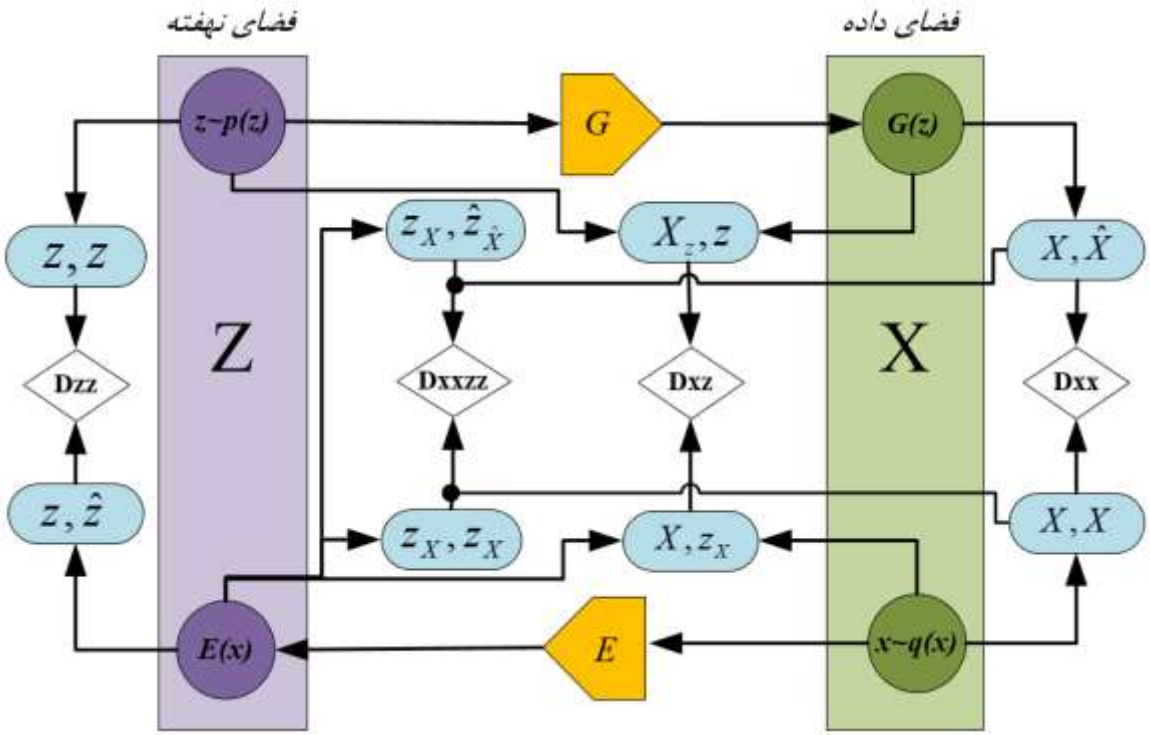


شکل۳-۲: نمایش جریان اطلاعات در مدل CALAD.

همانطور که در شکل ۳-۲ قابل مشاهده است، چرخه مورد نظر در این مسئله شامل سه گام متوالی است، در کارهای قبلی چرخه کامل CCC وجود نداشت و از اطلاعات به طور کامل در شبکه استفاده نمی‌شد. در حقیقت، مدل‌های پیشین ارائه شده به بررسی مستقل این دو جفت متغیر در فضای جداگانه می‌پرداخت و تمایزگر دید کاملی از جریان اطلاعات و وضعیت داده در هر دو فضا داده ورودی و نهفته به طور همزمان نداشت. قابل توجه است که این متغیر $\hat{z}_{\hat{x}}$ پیش از این محاسبه نمی‌شده و این چرخه در مدل‌های قبلی تعریف نمی‌شده است.

وجود تمایزگر D_{xxzz} سبب می‌شود تا ویژگی‌های جدید و عمیق‌تری (به نسبت تمایزگرهای تک گامی D_{xx} و D_{zz}) استخراج شود. این شبکه با در اختیار گرفتن خروجی کل چرخه داده، دید جامعی از

وضعیت تمام قسمت‌های شبکه دارد و با استفاده از تمام این خروجی‌ها به طور همزمان، به ویژگی‌های ترکیبی قوی‌تری برای تمیز نمونه‌های ناهنجار از نمونه داده‌های هنجار دست می‌یابد. جزئیات معماری مدل CALAD در شکل ۳-۳ نمایش داده شده است.



شکل ۳-۳: معماری CALAD.

نامگذاری‌های به‌کارگرفته شده در شکل ۳-۳ دقیقا مطابق با توضیحات ابتدای همین بخش می‌باشد. در بخش ۳-۱-۲ به توضیح دقیق‌تر روال آموزش، بررسی جزئیات بلوک‌های موجود و توضیح کامل تابع هدف این مدل خواهیم پرداخت.

۳-۱-۲- تابع هدف

همانند سایر مدل‌های پیشین در روند آموزش مدل پیشنهادی از روال‌های آموزش تقابلی استفاده می‌شود، بدیت ترتیب که به صورت متوالی بخش مولد و کدگذار و پس از آن تمایزگرها آموزش داده می‌شوند، یعنی به ترتیب با ثابت نگه داشتن وضعیت بخش مولد و کدگذار، پارامترهای شبکه‌های تمایزگر را به‌روزرسانی می‌کنیم. سپس آموزش این بخش‌ها را متوقف کرده و با توجه به خروجی بهبود یافته

آن‌ها، پارامترهای بخش مولد و کدگذار به روزرسانی می‌شود و این روال بارها و بارها تکرار شده تا مدل به کیفیت خروجی مطلوب دست یابد. هدف از آموزش تمامی مدل‌های مبتنی بر بازسازی از جمله مدل CALAD تولید بازسازی مناسب برای داده‌های هنجار و بازسازی ضعیف برای نمونه داده‌های ناهنجار است.

بیان ریاضی تابع هدف مدل پیشنهادی حاصل از جمع دو بخش کلی V_{ano} و V_{CCC} مطابق معادله ۳-۱۴ می‌باشد.

$$\min_{E,G} \max_{D_{xz},D_{xx},D_{zz},D_{xxzz}} V_{\text{ano}}(D_{xz}, G, E) + V_{\text{CCC}}(D_{xxzz}, D_{xx}, D_{zz}, G, E) \qquad (۱۴-۳)$$

در ادامه به بررسی هدف و جزئیات ریاضی هر یک از این دو بخش خواهیم پرداخت.

اولین بخش از این تابع هدف V_{ano} می‌باشد. در حالت کلی بخش اول تابع هدف این مسئله یعنی V_{ano} مطابق معادله ۳-۱۵ تعریف می‌شود.

$$\min_{E,G}\max_{D_{xz}} V_{\text{ano}}(D_{xz}, G, E) = \mathbb{E}_{x\sim q(x)}\left[\log D_{xz}\left(x, E(x)\right)\right] + \mathbb{E}_{z\sim p(z)}\left[\log\left(1 - D_{xz}(G(z), z)\right)\right] \qquad (۱۵-۳)$$

در این معادله، داده‌های هنجار مورد استفاده در مرحله آموزش با تابع توزیع احتمال $q(x)$ تعریف می‌شود و $p(z)$ به عنوان تابع توزیع ورودی برای شبکه مولد در نظر گرفته شده است.

بخش دوم تابع هدف مورد استفاده در تابع هزینه این مدل V_{CCC} است که تلاش می‌کند تا شباهت میان داده تولیدی و بازسازی آن توسط شبکه مولد برای داده‌های هنجار را تضمین کند. در واقع این تابع هدف شرط پایداری چرخه را هم به صورت تک مرحله‌ای و هم به صورت دو مرحله ای یعنی چرخه کامل ارضا می‌کند. این بخش مطابق معادله ۳-۱۶ فرمول‌بندی می‌شود.

$$\min_{E,G}\max_{D_{xx},D_{zz},D_{xxzz}} V_{ccc}(D_{xxzz}, D_{xx}, D_{zz}, E, G)$$
$$= \mathbb{E}_{z\sim \mathrm{p(z)}}[\log D_{zz}(z,z)] + \mathbb{E}_{z\sim \mathrm{p(z)}}\left[1 - \log D_{zz}\left(z, E(G(z))\right)\right]$$
$$+ \mathbb{E}_{x\sim q(x)}[\log D_{xx}(x,x)] + \mathbb{E}_{x\sim \mathrm{q(x)}}\left[1 - \log D_{xx}\left(x, G(E(x))\right)\right] \qquad (۱۶-۳)$$
$$+ \mathbb{E}_{x\sim \mathrm{q(x)}}\left[\log D_{xxzz}(x, x, E(x), E(x))\right]$$
$$+ \mathbb{E}_{x\sim \mathrm{q(x)}}\left[1 - \log D_{xxzz}\left(x, G(E(x)), E(x), E\left(G(E(x))\right)\right)\right]$$

در معادله ۳-۱۶ که متعلق به بخش چرخه پایداری تابع هدف می‌باشد از جمع چهار جزء تشکیل شده است، جزء دوم مربوط به تمایزگر D_{xx} می‌باشد. این تمایزگر تلاش می‌کند که شرط چرخه پایداری را به صورت تک گامی با شروع از فضای داده در این مدل ایجاد کند. جزء دوم مربوط به تمایزگر D_{zz} است که با هدف ایجاد چرخه پایداری تک گامی با شروع از فضای نهفته به مدل اضافه شده است و در نهایت دو عبارت آخر مربوط به تمایزگر جامع D_{xxzz} است. این تمایزگر در دو گام جریان اطلاعات را دنبال می‌کند و خروجی تمامی مراحل را بررسی می‌کند و چرخه پایدار را برای همه مراحل نگاشت داده در مدل ایجاد می‌نماید. همان‌طور که در بخش قبل نیز اشاره شد، هدف از تعبیه این تمایزگر در این ساختار، حل مشکل چرخه پایداری کامل یا CCC می‌باشد.

شاید پس از بررسی دقیق V_{CCC} این سوال پیش بیاید که آیا با وجود D_{xxzz} همچنان به دو تمایزگر دیگر نیز احتیاجی هست؟ در پاسخ به این سوال باید گفت که بله احتیاج هست، زیرا همانطور که در بخش ۳-۱-۱ و در شکل ۳-۲ نشان داده شد، این تمایزگر ورودی متفاوتی از تمایزگرهای قبلی دارد و از متغیر z که با هدف آموزش شبکه مولد G و یادگیری نگاشت معکوس از فضای نهان به فضای ورودی است، استفاده نمی‌کند و اگر دو تمایزگر یادشده حذف شوند، عملا تضمین چرخه پایداری برای متغیر z که اساس بخش V_{ano} می‌باشد نیز حذف شده که این اتفاق مطلوب نیست. در بخش ۴-۵-۴ به بررسی تاثیر هر یک از اجزا به تفکیک پرداخته شده است و نشان داده شده که بهترین نتیجه در حضور هر سه تمایزگر حاصل می‌شود.

۳-۲- مدل RALAD[۹]

بخش فعلی به معرفی مدل پیشنهادی دوم اختصاص داده شده است. این مدل بر پایه مدل ALAD بنا نهاده شده است. در مدل‌های مبتنی بر بازسازی فرض بر این است که اگر آموزش و بازسازی برای داده‌های هنجار به خوبی انجام بگیرد، بازسازی داده‌های ناهنجار لزوما ضعیف و متفاوت از داده اولیه ورودی خواهد بود. در صورتی که این فرض سهل‌انگارانه است و هیچ استلزام یا قید کنترلی برای این مشکل در هیچ یک از مدل‌های پیشین ارائه نشده است. در مدل ارائه شده با افزودن نمونه‌گیری از کل فضا، این استلزام برای بازسازی در فضای هنجار ایجاد و مدل را به سمت فضای بازسازی داده‌های هنجار متمایل[۱۰] کردیم. در بخش‌های آتی به بررسی جزئیات دقیق‌تر این مدل می‌پردازیم.

۳-۲-۱- معماری شبکه

روند آموزشی مدل ALAD با همه مزیت‌هایی که نسبت به مدل‌های پیشین دارد، اما از یک مشکل اساسی چشم‌پوشی کرده است؛ مشکل استلزام بازسازی ضعیف. تعریف دقیق این مشکل بدین ترتیب می‌باشد که در تمامی روال‌های آموزش مدل‌های تشخیص ناهنجاری، بازسازی دقیق نمونه‌های هنجار با کمترین خطا به مدل آموزش داده می‌شود و در مرحله آزمایش، نمونه‌های ناهنجار و هنجار به مدل داده می‌شود و همواره فرض می‌شود که برای نمونه‌های هنجار میزان اختلاف تصویر ورودی با تصویر بازسازی شده کم و برای نمونه‌های ناهنجار این اختلاف زیاد خواهد بود.

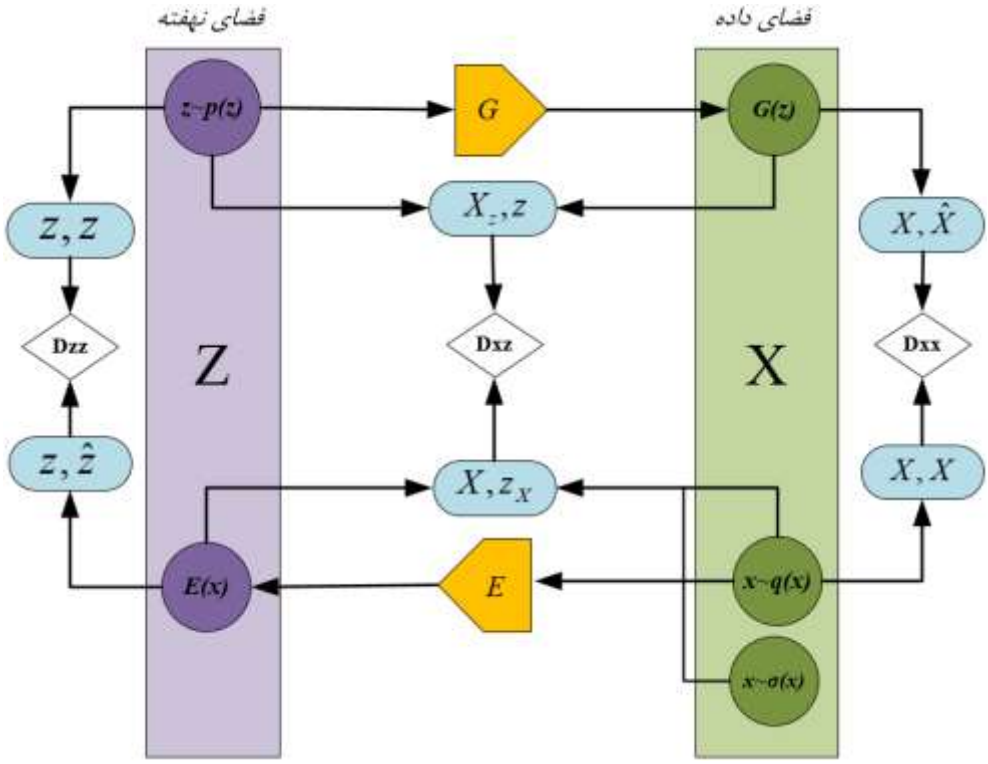
در برخی موارد پیش‌فرض فوق صحیح نیست و نمونه بازسازی شده داده ناهنجار، میزان اختلاف کمی با نمونه ورودی دارد و به همین سبب تشخیص آن به عنوان نمونه ناهنجار دشوار خواهد بود. در واقع در مدل‌های ارائه شده پیشین هیچ استلزامی برای بازسازی ضعیف نمونه ناهنجار وجود ندارد. علت وقوع این امر نگاشت تنک از فضای داده ورودی به فضای نهفته است. در حالت عادی آموزش فضای ورودی تنها به قسمت کوچکی از فضای نهفته نگاشت می‌شود و در نتیجه نمونه‌گیری از فضای نهفته به منظور نگاشت دوباره به فضای ورودی تنک خواهد بود. در زمان مواجهه با نمونه‌های هنجار این امر مشکلی ایجاد نخواهد

[9] Regularized Complete Adversarialy Learned Anomaly Detection

[10] Bias

کرد چرا که فضای متناسب Z برای نمونه‌های هنجار به خوبی مدل شده است ولی در نمونه‌های ناهنجار با توجه به اینکه مدل تا به حال چنین داده‌های را ندیده است ممکن است نمونه را به نقطه‌ای ناشناخته از فضای نهفته نگاشت کند و در نتیجه بازسازی نمونه ناهنجار نیز ممکن است به نقطه‌ای نامناسب در فضای ورودی نگاشت شود. نگاشت به‌دست آمده از این فرایند، هیچ ضمانتی برای ایجاد بازسازی ضعیف از نمونه ناهنجار ارائه نمی‌دهد.

بلوک $\sigma(x)$ به منظور پوشش حداکثری فضای نهفته به ساختار مدل ALAD اضافه می‌شود. هدف از تعبیه این بلوک تولید نمونه‌های جدید در فضای داده ورودی و سپس نگاشت آن به فضای نهفته است. انتظار می‌رود در این روند فضای نهفته به شکل مناسب‌تری نسبت به کارهای قبلی پوشش داده شود. نتایج عملی نمایانگر صحت تئوری ارائه شده در این قسمت است. در نهایت شمای کلی مدل پیشنهادی RALAD در شکل ۳-۴ قابل مشاهده است.



شکل ۳-۴: معماری RALAD.

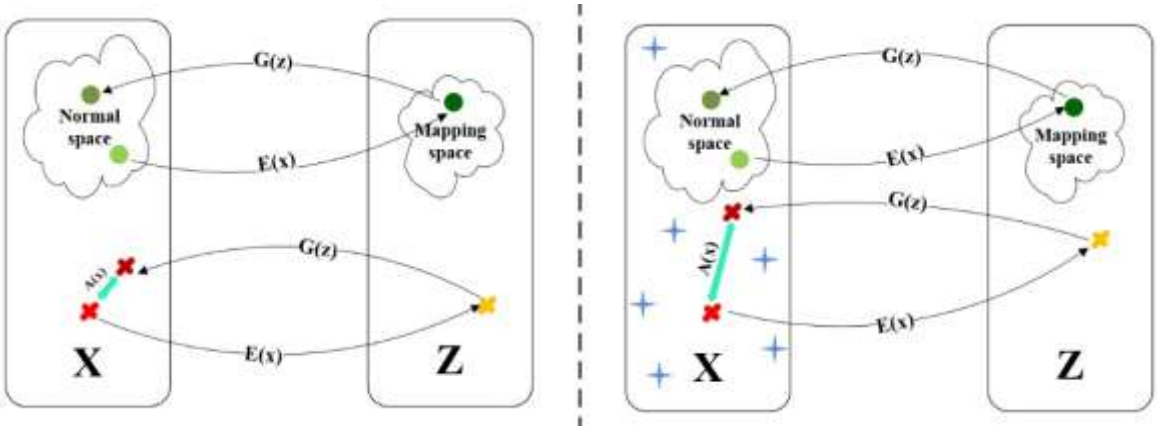
نامگذاری‌های به‌کارگرفته شده در شکل۳-۴ مطابق با توضیحات بخش ۳-۱-۱ می‌باشد. در بخش ۳-۲-۲ به توضیح دقیق‌تر روال آموزش و بررسی جزئیات بلوک‌های موجود در مدل خواهیم پرداخت.

۳-۲-۲- تابع هدف

تابع هدف و روال آموزش مدل RALAD همانند مدل ALAD است و تنها تفاوت آن وجود توزیع $\sigma(x)$ است. در حالت کلی مدل‌های مولد تخاصمی تلاش می‌کنند تا نزدیک‌ترین بازسازی برای تمامی داده‌ها فارغ از هنجار یا ناهنجار بودن آن‌ها ایجاد شود اما شبکه مورد نیاز برای تشخیص ناهنجاری باید برای داده‌های هنجار بازسازی نزدیک و برای ناهنجارها بازسازی متفاوت از داده ورودی تولید کند و هدف از آموزش تمامی مدل‌های مبتنی بر بازسازی از جمله مدل RALAD تولید بازسازی مناسب برای داده‌های هنجار و بازسازی ضعیف برای نمونه داده‌های ناهنجار است. ایده مدل RALAD برای ایجاد این فاصله بازسازی تمامی داده‌ها در زیرفضای توزیع داده‌های هنجار است.

ممکن است این سوال مطرح شود که در عمل در برخی موارد این رویه نگاشت، سبب نزدیک‌تر شدن فاصله داده ناهنجار ورودی و بازسازی آن شود و این نوع از آموزش خلاف هدف مطلوب عمل کند. در پاسخ به این مسئله باید ذکر کرد که امکان این اتفاق در پاره‌ای از موارد وجود دارد، اما در مقایسه با مدل پایه ALAD که در آن هیچ اطلاعی از وضعیت نگاشت داده‌های ناهنجار وجود نداشت حال با اطمینان بالاتری می‌دانیم که یک فاصله حداقلی میان داده ورودی با بازسازی آن وجود دارد و بنابر نتایج به دست آمده در فصل چهارم ثابت شده است همین فاصله سبب تخصیص امتیاز ناهنجاری مناسب به نمونه‌های ناهنجار می‌شود.

در اینجا برای تاکید بر ایجاد بازسازی ضعیف برای داده ناهنجار، تلاش شده است تا تمامی فضای داده ورودی را به کمک نویز $\sigma(x)$ پوشش دهیم و مولد شبکه را به سمت بازسازی هر بیشتر نزدیک به توزیع داده هنجار متمایل کنیم. در شکل ۳-۴ چگونگی عملکرد این قسمت از مدل پیشنهادی و نحوه تاثیر آن در فرایند آموزش به تصویر کشیده شده است.



شکل ۳-۵: تاثیر حضور توزیع σ(x) در روند آموزش مدل.

در شکل ۳-۵ x بیانگر فضای داده ورودی و z بیانگر فضای داده ورودی است. نمونه‌ها توسط مولد G از فضای داده ورودی به فضای نهفته نگاشت می‌شوند و وظیفه انجام نگاشت معکوس بر عهده کدگذار E است. دایره‌های سبز رنگ نماد نمونه داده‌های هنجار و ضربدرهای قکد نماد نمونه‌های ناهنجار هستند. بعلاوه آبی رنگ نشانگر نمونه‌های تولید شده توسط توزیع $\sigma(x)$ هستند که در تنها مرحله آموزش مورد استفاده قرار گرفته‌اند. فلش فیروزه‌ای مقدار امتیاز ناهنجاری را نشان می‌دهد . همانطور که در شکل ۳-۵ مشاهده می‌شود در صورت عدم حضور $\sigma(x)$ (در سمت چپ شکل) در روند آموزش، امتیاز ناهنجاری برای نمونه‌های غیرعادی کمتر از زمانی است که از این توزیع استفاده شده است، در واقع در تصویر سمت راست، توزیع $\sigma(x)$ مدل را به سمت بازسازی همه نمونه‌ها اعم از ناهنجار و هنجار به سمت توزیع داده‌های هنجار متمایل کرده است.

بیان ریاضی تابع هدف مدل پیشنهادی RALAD همانند مدل پیشنهادی قبلی حاصل از جمع دو بخش کلی V_{ano} و V_{CC} و مطابق معادله ۳-۱۷ می‌باشد.

$$\min_{E,G} \max_{D_{xz},D_{xx},D_{zz}} V_{\text{ano}}(D_{xz},G,E) + V_{CC}(D_{xx},D_{zz},G,E) \qquad (۱۷-۳)$$

تنها تفاوت موجود در این مدل در بخش V_{ano} نمایان می‌شود. با هدف حل مسئله استلزام بازسازی و توانمندسازی شبکه پیشنهادی برای تمایز هر چه بهتر بین داده‌های هنجار و ناهنجار از توزیع جریمه $\sigma(x)$ استفاده شده است. ابعاد داده‌های خروجی این توزیع برابر با ابعاد داده ورودی x می‌باشد و در حالتی که هیچ اطلاعات اضافه ای از داده نداشته باشیم، با توجه به این که داده‌ها در مرحله پیش‌پردازش

نرمال شده اند، از توزیع گاوسی نرمال برای پوشش کلی‌تر فضا استفاده می‌کنیم. البته اگر اطلاعات اضافه ای به مدل از توزیع داده‌های ناهنجار داده شود، می‌توان با بکارگیری به جای توزیع $\sigma(\mathrm{x})$ اعمال کرد. در حالت کلی بخش اول تابع هدف این مسئله V_{ano} طبق معادله ۳-۱۸ تعریف می‌شود.

$$\begin{aligned}\min_{E,G}\max_{D_{xz}} V_{\text{ano}}(D_{xz},G,E) &= \mathbb{E}_{x\sim q(x)}\left[\log D_{xz}\left(x,E(x)\right)\right]\\ &+\mathbb{E}_{z\sim p_g(z)}\left[\log\left(1-D_{xz}(G(z),z)\right)\right]\\ &+\mathbb{E}_{x\sim\sigma(\mathrm{x})}\left[\log\left(1-D_{xz}\left(x,E(x)\right)\right)\right]\end{aligned} \qquad (۱۸\text{-}۳)$$

در این معادله، داده‌های هنجار مورد استفاده در مرحله آموزش با تابع توزیع احتمال $q(x)$ تعریف می‌شود و $p_g(\boldsymbol{z})$ به عنوان تابع توزیع ورودی برای شبکه مولد در نظر گرفته شده است. در این معادله از تابع توزیع جریمه $\sigma(\mathrm{x})$ برای پوشش بهتر فضای داده ورودی و حل مشکل تنک بودن فضای نهفته استفاده می‌شود. در این معادله با داده‌های خروجی از توزیع $\sigma(\mathrm{x})$ و بازسازی آن‌ها به عنوان نمونه‌های تقابلی برخورد می‌شود و بدین ترتیب با افزودن عبارت سوم یعنی $\mathbb{E}_{\boldsymbol{x}\sim\sigma(\boldsymbol{x})}\left[\log\left(1-D_{\boldsymbol{xz}}\left(x,E(x)\right)\right)\right]$ به تابع هزینه کلی، تلاش می‌شود تا تمامی فضای ورودی به زیرفضا متعلق به توزیع داده‌های هنجار در فضای نهفته نگاشت شود و با این روش بازسازی‌های ارائه شده توسط شبکه به سمت توزیع داده هنجار ورودی متمایل شود. توضیح بیشتر این که ابتدا داده تولیدی از توزیع $\sigma(x)$ به شبکه کدگذار وارد می‌شود، در طی فرایند آموزش کدگذار می‌آموزد تمام فضای داده ورودی را به فضای متناسب توزیع داده هنجار در فضای نهفته نگاشت کند، در نتیجه این روند، توزیع داده ورودی شبکه مولد همواره از فضای متناسب با توزیع داده هنجار خواهد بود و خروجی حاصل نیز متعلق به توزیع متناسب با آن در فضای داده ورودی می‌شود. از آن جایی که شبکه مولد نگاشت از فضای نهفته متعلق به داده هنجار به فضای داده ورودی متناسب را یاد گرفته است، تمامی فضای ورودی فارغ از هنجار یا ناهنجار بودن به فضای داده هنجار نگاشت می‌شود. پس به ازای داده‌های ناهنجار اختلاف بازسازی ارائه شده توسط شبکه بیشتر از داده‌های هنجار خواهد بود.

برای اثبات نظری تمایل شبکه‌های مولد و تمایزگر به سمت داده‌های هنجار ابتدا تمایزگر و مولد بهینه را آموزش می‌دهیم. توزیع توام کدگذار روی داده‌های هنجار به صورت $q(x,z) = q(x)e(z \mid x)$ و توزیع توام کدگذار روی داده‌های ناشی از توزیع جریمه به صورت $\sigma(x,z) = \sigma(x)e(z \mid x)$ و

$p(x,z) = p(z)p(x \mid z)$ اســـت. نقطه بهینه برای تمایزگر D_{zz} که از معادله ۳-۱۸ بدســـت می‌آید مطابق معادله ۳-۱۹ می‌باشد.

$$D_{xz}^{*} = \frac{q(x,z)}{q(x,z) + \sigma(x,z) + p(x,z)} = \frac{q(x,z)}{\left(1 + \frac{\sigma(x)}{q(x)}\right) q(x,z) + p(x,z)} \qquad (۱۹ - ۳)$$

در معادله ۳-۱۹ هر دو توزیع داده ورودی و توزیع داده جریمه در نظر گرفته است. از معادله فوق می‌توان نتیجه گرفت بر خلاف GAN اســتاندارد که تنها روی نمونه‌های عادی آموزش می‌بینند تمایزگر مدل ارائه شـده در این کار احتمال بیشـتری به نمونه‌های عادی اختصـاص می‌دهد و با توجه به این که به نمونه‌های ناهنجار با $q(x)$ کوچکتر اختصـاص می‌یابد، خروجی تمایزگر بهینه برای داده‌های ناهنجار کم می‌شـود. نتایج عملی ارائه شده در فصل چهارم گزاره‌های ارائه شده در این قسمت را پشتیبانی می‌کنند.

بخش دوم تابع هدف مورد اســـتفاده در تابع هزینـه این مدل V_{CC} اســـت که دقیقا همانند V_{CC} مدل ALAD است. تعریف این بخش در معادله ۳-۲۰ آمده است.

$$\min_{E,G} \max_{D_{xx}, D_{zz}} V_{cc}(D_{xx}, D_{zz}, E, G) \qquad (۲۰ - ۳)$$
$$= \mathbb{E}_{z\sim \mathrm{p(z)}}[\log D_{zz}(z,z)] + \mathbb{E}_{z\sim \mathrm{p(z)}}\left[1 - \log D_{zz}\left(z, E\left(G(z)\right)\right)\right]$$
$$+ \mathbb{E}_{x\sim q(x)}[\log D_{xx}(x,x)] + \mathbb{E}_{x\sim \mathrm{q(x)}}\left[1 - \log D_{xx}\left(x, G\left(E(x)\right)\right)\right]$$

۳-۳- مدل RCALAD[11]

در این بخش به بررسـی مدل اصـلی و کلی پیشـنهادی این تحقیق که از ترکیب دو شـبکه CALAD و RALAD بدسـت می‌آید، می‌پردازیم. در این شـبکه به هر دو مسـئله چرخه پایداری کامل و اسـتلزام بازسـازی ضـعیف به طور همزمان پرداخته شـده اسـت و تلاش شـده تا یک چارچوب جامع، کاربردی و سازگار برای تمامی مسائل تشخیص ناهنجاری ارائه شود.

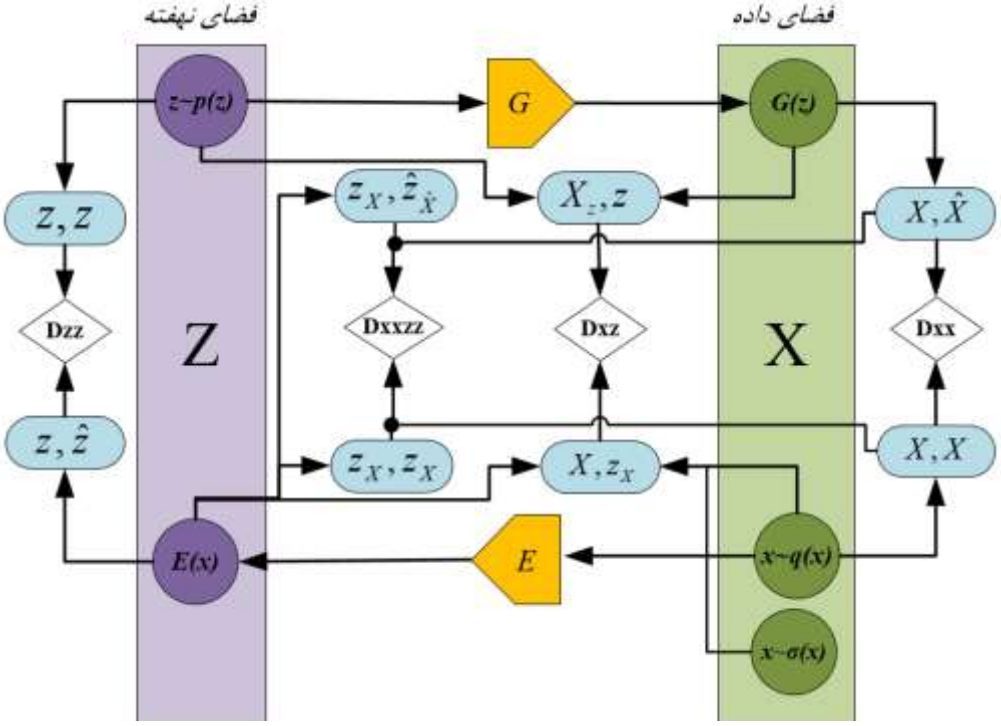
[11] Regularized Complete Adversarialy Learned Anomaly Detection

نتایج آزمایش‌ها بر روی هر دو نوع داده تصـــویر و جدولی بیانگر کارایی و اثربخشـــی روش پیشـــنهادی RCALAD می‌باشد و نمایانگر سازگاری نتایج تئوری و عملی بدست‌آمده برای این مسئله است.

۳-۳-۱- معماری شبکه

همان‌طور که پیش از این ذکر شد، ایده بکارگرفته شده در این پروژه برای حل مشکل چرخه پایداری کامل، در این‌جا متغیر جدید $\hat{z}_{\hat{x}}$ تعریف و تمایزگر D_{xxzz} به ساختار کلی افزوده گردیده است که سبب تامین شرط چرخه پایداری کامل میان هر دو فضای ورودی و فضای نهان می‌شود. به جزئیات دقیق تعریف این مسئله، تعریف کامل این متغیر و نحوه آموزش تمایزگر D_{xxzz} در بخش ۳-۱ به تفضیل پرداخته شده است. همچنین با هدف متمایل سازی بازسازی ارائه شده توسط شبکه به سمت توزیع داده هنجار، از توزیع $\sigma(x)$ برای نمونه گیری از کل فضا استفاده شده است و با استفاده از این توزیع تلاش شده تا که کل فضای ورودی به شبکه کدگذار نشان داده شود و کدگذار ملزم به ایجاد یک نگاشت نزدیک‌تر به فضای داده هنجار بشود. توضیحات کامل‌تر این موضوع نیز در بخش ۳-۲ شرح داده شده است.

با تجمیع این دو ایده در یک چارچوب در نهایت مدل RCALAD بدســت خواهد آمد. شــمای کلی مدل پیشنهادی و جزئیات ورودی هر یک از اجزای RCALAD در شکل ۳-۶ قابل مشاهده است.



شکل ۳-۶: معماری RCALAD.

نمادگذاری بکارگرفته شده در شکل ۳-۶ در زیربخش ۳-۱-۱ به طور کامل شرح داده شده است. در بخش بعد به توضیح دقیق‌تر روال آموزش و بررسی تابع هدف این شبکه خواهیم پرداخت.

۳-۳-۲- تابع هدف

تابع هدف مدل پیشنهادی RCALAD همانند مدل‌های پیشنهادی قبلی حاصل از جمع دو بخش کلی V_{ano} و V_{CCC} و مطابق معادله ۳-۲۰ می‌باشد.

$$\min_{E,G} \max_{D_{xz},D_{xx},D_{zz},D_{xxzz}} V_{\text{ano}}(D_{xz}, G, E) + V_{CCC}(D_{xxzz}, D_{xx}, D_{zz}, G, E) \qquad (۱۷-۳)$$

در بخش V_{ano} با هدف حل مسئله استلزام بازسازی و توانمندسازی شبکه پیشنهادی برای تمایز هر چه بهتر بین داده‌های هنجار و ناهنجار از توزیع جریمه $\sigma(x)$ همانند مدل RALAD استفاده شده است. بخش اول تابع هدف این مسئله یعنی V_{ano} طبق معادله ۳-۲۱ تعریف می‌شود.

$$\begin{aligned}\min_{E,G}\max_{D_{xz}} V_{\text{ano}}(D_{xz}, G, E) &= \mathbb{E}_{x\sim q(x)}\left[\log D_{xz}\left(x, E(x)\right)\right] \\ &+\mathbb{E}_{z\sim p(z)}\left[\log\left(1 - D_{xz}(G(z), z)\right)\right] \\ &+\mathbb{E}_{x\sim\sigma(x)}\left[\log\left(1 - D_{xz}\left(x, E(x)\right)\right)\right]\end{aligned} \qquad (۲۱-۳)$$

در این معادله، داده‌های هنجار مورد استفاده در مرحله آموزش با تابع توزیع احتمال $q(x)$ تعریف می‌شود و $p(z)$ به عنوان تابع توزیع ورودی برای شبکه مولد در نظر گرفته شده است. در این معادله از تابع توزیع جریمه $\sigma(x)$ برای پوشش بهتر فضای داده ورودی و حل مشکل تنک بودن فضای نهفته استفاده می‌شود.

بخش دوم تابع هدف مورد استفاده در تابع هزینه این مدل V_{CCC} است که مطابق با V_{CCC} معرفی شده در مدل CALAD است. در این بخش یک متغیر و تمایزگر جدید برای تضمین چرخه پایداری وابسته هر دو فضا معرفی شده است. تعریف این بخش از تابع هزینه مطابق معادله ۳-۲۲ می‌باشد.

$$\min_{E,G}\max_{D_{xx},D_{zz},D_{xxzz}} V_{ccc}(D_{xxzz}, D_{xx}, D_{zz}, E, G)$$
$$= \mathbb{E}_{z\sim \mathrm{p(z)}}[\log D_{zz}(z,z)] + \mathbb{E}_{z\sim \mathrm{p(z)}}\left[1 - \log D_{zz}\left(z, E(G(z))\right)\right]$$
$$+ \mathbb{E}_{x\sim q(x)}[\log D_{xx}(x,x)] + \mathbb{E}_{x\sim \mathrm{q(x)}}\left[1 - \log D_{xx}\left(x, G(E(x))\right)\right] \qquad (۲۲-۳)$$
$$+ \mathbb{E}_{,x\sim \mathrm{q(x)}}\left[\log D_{xxzz}(x, x, E(x), E(x))\right]$$
$$+ \mathbb{E}_{x\sim \mathrm{q(x)}}\left[1 - \log D_{xxzz}\left(x, G(E(x)), E(x), E\left(G(E(x))\right)\right)\right]$$

۳-۴- تشخیص ناهنجاری

همانطور که گفته شد مدل پیشنهادی در این پروژه به منظور تشخیص ناهنجاری بر اساس بازسازی داده ورودی است. داده هنجار به صورت دقیق و شبیه به داده ورودی بازسازی می‌شود در حالی که بازسازی نمونه ناهنجار ضعیف خواهد بود. پس از بازسازی نمونه‌ها، عنصر کلیدی در تشخیص ناهنجاری تعریف امتیاز ناهنجاری با هدف محاسبه فاصله میان نمونه ورودی و خروجی بازسازی شده، توسط شبکه است. اولین انتخاب در زمینه محاسبه میزان فاصله این دو داده فاصله اقلیدسی است ولی در فضای داده این معیار ممکن است به اندازه کافی قابل اتکا نباشد[32]. به عنوان مثال در زمینه تشخیص ناهنجاری در تصاویر ممکن است علی‌رغم ویژگی‌های بصری مشابه، فاصله اقلیدسی زیادی داشته باشند.

در این کار به جای محاسبه فاصله میان نمونه‌ها در فضای داده ورودی، از فضای ویژگی موجود در تمایزگر D_{xxzz} استفاده می‌شود. به این منظور خروجی توابع فعالیت[۱۲] لایه یکی مانده به آخر به عنوان ویژگی استفاده می‌شوند. امتیاز ناهنجاری مورد استفاده به صورت زیر و با استفاده از خطای بازسازی نرم یک و مطابق معادله ۳-۲۳ تعریف می‌شود.

$$A_{fm}(x) = \|f_{xxzz}(x, x, E(x), E(x)) - f_{xxzz}(x, G(E(x)), E(x), E(G(E(x))))\|_1 \qquad (۲۳-۳)$$

در معادله ۳-۲۳ $f(\cdot)$ بیانگرتابع فعالیت لایه یکی ملنده به آخر در ساختار تمایزگر D_{xxzz} است. $A_{fm}(x)$ میزان اطمینان تمایزگر از کیفیت روند کدگذاری و بازسازی داده توسط مولد است که اگر

[12] Activation function

خوب انجام شـده باشـد در واقع نمونه متعلق به توزیع داده هنجار اسـت که مدل روی آن آموزش دیده است. بنابر مطالب گفته شده هر چه مقدار این معیار بیشتر باشد، اختلاف بازسازی‌ها بیشتر و احتمال ناهنجاری بودن آن ورودی بیشتر است. عملکرد مناسب این معیار در مقایسه با سایر معیارهای تشخیص ناهنجاری در فصل آینده نمایش داده شده است.

معیار پیشنهادی در این قسمت بر اساس روش تطبیق ویژگی[۱۳] و یا به اختصار fm، پایه‌گذاری شده است [40] ، در ساختار GAN استاندارد برای تطبیق ویژگی از خروجی شبکه مولد استفاده شده است و در مقاله ALAD [32] از خروجی تمایزگر D_{xx} برای محاسبه این معیار استفاده شده است.

در مدل ALAD تنها از خروجی تمایزگر D_{xx} برای شناسایی نمونه‌های ناهنجار استفاده شد و بنابر نتایج به‌دست آمده برای مدل RCALAD این تمایزگر فاقد بخشی از اطلاعات مفید موجود در این مدل است. به منظور بهره‌گیری از تمامی اطلاعات موجود در این مدل برای تشخیص ناهنجاری در این‌جا یک معیار جدید با نام A_{all} نیز تعریف شده است. این امتیاز از جمع خروجی هر سه تمایزگر D_{xx}، D_{zz} و D_{xxzz} تشکیل شده است. تمام تمایزگرهای موجود در مدل پیشنهادی تنها روی نمونه‌های هنجار آموزش دیده‌اند و بازسازی برای تمامی فضای داده ورودی به سمت فضای داده هنجار متمایل شده است، پس انتظار می‌رود تصویر بازسازی شده نمونه ناهنجار و همچنین بازنمایی آن در فضای نهفته که توسط کدگذار تولید می‌شود، ضعیف باشد و تمایزگرهای موجود در مدل این ورودی‌های ناهنجار را شناسایی کنند. بیان ریاضی این معیار در معادله ۳-۲۴ آورده شده است.

$$A_{all}(x) = D_{xxzz}(x, \hat{x}, z_x, \hat{z}_{\hat{x}}) + D_{xx}(x, \hat{x}) + D_{zz}(z_x, \hat{z}_{\hat{x}}) \qquad (۲۴-۳)$$

حال مسئله قابل بررسی این موضوع می‌باشد که آیا معیار A_{all} حاوی اطلاعات کافی برای تشخیص داده‌های هنجار از ناهنجار می‌باشد یا خیر. پاسخ به این سوال در حالت کلی بله می‌باشد زیرا این تمایزگرها در طی فرایند آموزش یاد می‌گیرند که به اختلاف میان دوتایی (x, x) و $(x, \hat{x})$ و همچنین دوتایی (z_x, z_x) و $(z_x, \hat{z}_{\hat{x}})$ توجه کنند یعنی هر چه $\hat{x}$ از x و یا $\hat{z}_{\hat{x}}$ از z_x فاصله بگیرد، تشخیص آن برای تمایزگرها ساده‌تر می‌شود. حال با افزودن توزیع $\sigma(x)$ به این مجموعه و تلاش برای متمایل کردن

[13] Feature Matching

تمامی بازسازی به سمت توزیع داده‌های هنجار، اختلاف بازسازی برای داده‌های ناهنجار زیادتر هم شده و عملاً خروجی تمایزگرها یک معیار قلبل اتکا برای تشـــخیص ناهنجاری می‌شـــود. در نهایت الگوریتم محاسبه امتیازهای ناهنجاری پیشنهادی مطابق الگوریتم ۳-۲ قابل مشاهده است.

Algorithm 2 Regularized Comprehensive Adverarially Learned Anomaly Detection

Input $x \sim p_{X_{\text{Test}}}(x), E, G, D_{xx}, D_{zz}, D_{xxzz}, f_{xxzz}$ where f_{xxzz} is the feature layer of , D_{xxzz}

Output $A_{all}(x)$, $A_{fm}(x)$, **where A is the anomaly score**

1: **procedure** INFERENCE

2: $z_x \leftarrow E(x)$ Encode samples, Construct latent Embedding

3: $\hat{x} \leftarrow G(z_x)$ Reconstruct samples

4: $\hat{z}_{\hat{x}} \leftarrow E(\hat{x})$ Reconstruct latent Embedding

5: $A_{fm}(x) \leftarrow \|f_{xxzz}(x, x, z_x, z_x) - f_{xx}(x, \hat{x}, z_x, \hat{z}_{\hat{x}})\|_1$

6: $A_{all}(x) \leftarrow D_{xxzz}(x, \hat{x}, z_x, \hat{z}_{\hat{x}}) + D_{xx}(x, \hat{x}) + D_{zz}(z_x, \hat{z}_{\hat{x}})$

7: return $A_{all}(x)$, $A_{fm}(x)$

8: **end procedure**

الگوریتم ۳-۲: روند محاسبه‌ی امتیاز ناهنجاری

در فصل آینده این دو معیار با سایر معیارهای تشخیص ناهنجاری مقایسه شده و در طی آزمایشات مختلف کارایی این امتیازها نشان داده می‌شود.

۳-۵- جمع‌بندی

در این فصل ابتدا به مرور کوتاهی از مدل‌های پایه و ایرادات هر یک از آن‌ها پرداخته شد. سپس معماری و تابع هدف ســه مدل پیشــنهادی این تحقیق با جزییات دقیق‌تر مورد بررســی قرار گرفت. در ابتدا مدل CALAD که با هدف تامین چرخه پایداری کامل، متغیر $\hat{z}_{\hat{x}}$ تعریف شـــد و از آن برای آموزش شـــبکه تمایزگر توامان D_{xxzz} اسـتفاده شـد. بدین ترتیب از تمامی جریان اطلاعات موجود در شـبکه به طور توام

استفاده شد و روند آموزش و چرخه پایداری به بیشینه دقت در بین سایر مدل‌های مبتنی بر شبکه‌های مولد تقابلی رسید.

در گام بعدی به منظور هدایت کردن مدل به سمت تولید بازسازی ضعیف برای نمونه‌های ناهنجار و متمایل سازی تمامی بازسازی‌های ارائه شده توسط شبکه به سمت توزیع هنجار، با الهام از مدل RCGAN از یک توزیع نویز در فضای داده ورودی با نام $\sigma(x)$ استفاده شد. بدین ترتیب که تلاش شد، با نمونه‌های کدشده توسط شبکه کدگذار همانند نمونه‌های خصمانه تولید شده توسط شبکه مولد برخورد شود و بدین ترتیب نگاشت تمامی فضای ورودی به سمت توزیع داده هنجار متمایل شود.

در گام نهایی دو امتیاز ناهنجاری جدید با نام‌های A_{all} و A_{fm} معرفی شدند، همانطور که مشاهده شد معیار اول مبتنی بر خروجی تمایزگرهای موجود بر شبکه تعریف شد بود و اساس کار معیار دوم بر استفاده از تطبیق ویژگی در لایه‌های تمایزگر D_{xxzz} بنا نهاده شده بود. در فصل آینده کارایی مدل پیشنهادی روی دادگان مختلف بررسی خواهد شد.

فصل چهارم:
آزمایش‌ها و نتایج

به منظور بررسی و ارزیابی کارایی الگوریتم پیشنهادی در تشخیص ناهنجاری در این فصل نتایج عملکرد آن طی چندین روال آزمایشی ارائه می‌شود. در ابتدا به معرفی مجموعه داده‌های آزمایشی و تنظیمات متناسب با هر یک از آن‌ها برای مدل پیشنهادی می‌پردازیم. سپس دیگر الگوریتم‌های تشخیص ناهنجاری مطرح پایه را معرفی کوتاهی خواهیم کرد، در گام بعدی به بررسی نتایج بر روی مجموعه داده‌های جدولی[1] و تصویری ارائه شده می‌پردازیم و سپس نتایج به‌دست آمده بررسی، مقایسه و تحلیل می‌شود. درنهایت به بحث بیشتر و موشکافی نحوه عملکرد هر یک از امتیازهای معرفی شده در بخش ۳-۳-۳ در مجموعه داده‌های مختلف اشاره می‌شود.

۴-۱- دادگان و پیش‌پردازش

برای سنجش عملکرد مدل پیشنهادی و بررسی کارایی آن ار جنبه‌های مختلف از مجموعه داده‌هایی با ویژگی متفاوت استفاده می‌شود. در این قسمت مشخصات و ویژگی‌های این مجموعه داده‌ها معرفی خواهد شد.

۴-۱-۱- مجموعه داده KDDCup99

مجموعه دادگان KDDCUP یک دادگان جدولی در ارتباط با نفوذ به شبکه‌های کامپیوتری است. این دادگان شامل ۴۹۴۰۲۱ نمونه با ۳۴ ویژگی اسمی[2] و هفت ویژگی پیوسته[3] است. در مرحله پیش‌پردازش ویژگی‌های اسمی به روش بازنمایی one-hot کدگذاری می‌شوند و نمونه‌های نهایی هر کدام ۱۲۱ بعد خواهند داشت. داده‌های با برچسب غیرنفوذ با توجه به اینکه حدود بیست درصد مجموعه داده را شامل می‌شود و در اقلیت است به عنوان ناهنجاری در نظر گرفته می‌شود. در مرحله آزمون بیست درصد از داده‌ها با بیشترین امتیاز ناهنجاری به عنوان داده ناهنجار در نظر گرفته می‌شود. معیارهای ارزیابی برای

[1] Tabular

[2] Categorical

[3] Continuous

سنجش این مجموعه داده شامل صحت، بازیابی و F1-score است. توضیحات این معیارها در بخش ۲-۴ به تفصیل مورد بررسی قرار گرفته است.

۴-۱-۲- مجموعه داده Arrhytthmia

دادگان جدولی آریتمی قلبی شامل ۴۵۲ نمونه با ۲۷۴ ویژگی است و هر داده می‌تولند به ۱۶ کلاس مختلف دسته‌بندی شود. کوچکترین کلاس‌ها به ترتیب شامل ۳،۴،۵،۷،۸،۹،۱۴ و ۱۵ نمونه هستند و جمعا ۱۵ درصد از نمونه‌های این دادگان را شامل می‌شوند و همین نمونه‌ها در واقع ناهنجاری هستند. سایر کلاس‌ها به عنوان داده هنجار در نظر گرفته می‌شوند. در اینجا نیز در مرحله آزمایش ۱۵ درصد از دادگان با بیشترین امتیاز به عنوان داده‌های ناهنجار در نظر گرفته می‌شوند. معیارهای ارزیابی برای سنجش این مجموعه داده شامل صحت، بازیابی و F1-score است.

۴-۱-۳- مجموعه داده Thyroid

این دادگان مربوط به بیماری تیرویید و جدولی است و شامل ۳۷۷۲ نمونه در سه کلاس و شامل شش ویژگی پیوسته است. کلاس با برچسب hyperfunction که شامل ۲.۵ درصد از مجموعه داده است به عنوان داده ناهنجار دسته‌بندی می‌شود، بنابراین در مرحله آزمون ۲.۵ درصد از نمونه‌ها با امتیاز ناهنجاری بالا به عنوان داده ناهنجار تشخیص داده می‌شود. در این مجموعه داده ۵۰٪ از نمونه‌های موجود به صورت تصادفی به عنوان داده آموزشی انتخاب شده است. توجه شود که نمونه‌های ناهنجار در تمامی مراحل آموزش مدل از داده‌های آموزشی حذف می‌شود. معیارهای ارزیابی برای سنجش این مجموعه داده همانند دیگر مجموعه داده‌های جدولی شامل صحت، بازیابی و F1-score است.

۴-۱-۴- مجموعه داده Musk

دادگان Musk دادگانی جدولی مربوط به دسته‌بندی شش کلاسی روی مشک مولکولی[4] شامل ۳۰۶۲ نمونه با ۱۶۶ ویژگی است. کلاس‌های موجود در دسته به نام ۲۱۳ و ۲۱۱ به عنوان داده ناهنجار در نظر

[4] Musk molecular

گرفته می‌شوند و شامل ۳.۲ درصد دادگان است. در مرحله پیش‌پردازش دو ستون اسمی این مجموعه داده که شامل اسامی مولکول‌ها و ساختار آن‌ها حذف شده است. معیارهای ارزیابی برای سنجش این مجموعه داده مانند بخش‌های گذشته شامل صحت، بازیابی و F1-score است.

۴-۱-۵- مجموعه داده CIFAR-10

این دادگان شامل ۶۰۰۰۰ تصویر ۳۲*۳۲ در ده کلاس است. برای آزمایش مدل پیشنهادی روی این دادگان هر بار یک کلاس به عنوان داده هنجار در نظر گرفته می‌شود و نه کلاس دیگر به عنوان داده ناهنجار در نظر گرفته می‌شود. معیار مقایسه مدل پیشنهادی با سایر مدل‌ها در این مجموعه داده AUROC[5] است. ۸۰ درصد داده به عنوان داده آموزشی در نظر گرفته می‌شود و ۲۰ درصد باقی به عنوان داده آزمون و ۲۵ درصد نمونه‌های آزمون به عنوان داده اعتبارسنجی در نظر گرفته می‌شود. توجه شود نمونه‌های ناهنجار از داده آموزشی و اعتبارسنجی حذف می‌شود.

۴-۱-۶- مجموعه داده SVHN

دادگان SVHN مربوط به دسته‌بندی اعداد بین صفر تا نه روی پلاک خانه‌هاست. این مجموعه داده شامل ۹۹۲۸۹ تصویر ۳۲*۳۲ است. رویکرد آموزش، اعتبار سنجی و آزمون در این دادگان همانند دادگان CIFAR10 است.

۴-۲- تنظیمات مدل

در این قسمت به معرفی و بررسی جزئیات به کارگیری و تنظیمات معماری شبکه عصبی مدل پیشنهادی بر روی مجموعه داده‌های مختلف می‌پردازیم. تمامی نتایج گزارش شده در پایان‌نامه با استفاده از چارچوب Tensorflow1 و به زبان پایتون تولید شده است. توان پردازشی مورد نیاز در این کار با استفاده از NVIDIA A100 تامین شده است. تنظیمات آزمون[6] همانند مدل پایه ALAD می‌باشد.

[5] Area Under Receiver Operating Curve

[6] Experimental Setup

جزئیات ساختاری و معماری مدل پیشنهادی در بخش ۳-۲ به تفضیل بررسی شده است. فراپارامتر[۷]ها برای بهینه ساز Adam از کتابخانه Tensorflow1 همانند [32] معادل $\alpha = 10^{-5}$ و $\beta_1 = 0.5$ می‌باشد. اندازه هر دسته[۸] در تمامی آزمایش‌ها ۳۲ در نظر گرفته شده است.

۴-۳- مدل‌های پایه

در این قسمت چارچوب پیشنهادی با تعداد زیادی از روش‌های تشخیص ناهنجاری مقایسه می‌شود، بخش اعظمی از مدل‌های مورد مقایسه در فصل۲ مورد بررسی قرار گرفت. در ادامه به طور مختصر مدل‌هایی که پیش از بررسی نشده، شرح داده می‌شود.

۴-۳-۱- روش OC-SVM[۹]

این روش یک مرز حول نمونه‌های هنجار یاد می‌گیرد و نمونه‌هایی که خارج از این مرز قرارگیرند به عنوان ناهنجاری در نظر گرفته می‌شوند. این مرز توسط روش ماشین بردار پشتیبان ایجاد می‌شود و هسته[۱۰] مورد استفاده در این روش RBF[۱۱] است. پارامتر ν برابر نسبت تعداد داده‌های ناهنجار به کل داده‌هاست و γ برابر 1/m قرار داده می‌شود، m برابر تعداد ویژگی‌های دادگان است[۳۳], [۴۲].

۴-۳-۲- روش IF[۱۲]

این روش از دسته روش‌های کلاسیک یادگیری ماشین است و به جای مدل کردن توزیع داده هنجار داده ناهنجار را از سایر نمونه‌ها جدا می‌کند. ابتدا در این روش ابتدا تعدادی ویژگی انتخاب می‌شود و یک

[7] Hyper Parameter

[8] Batch

[9] One Class Support Vector Machine

[10] Kernel

[11] Radial Basis Function

[12] Isolated Forest

مقدار تصـادفی برای هر ویژگی انتخاب می‌شـود تا بتوان داده‌ها را جدا کرد. در ادامه میانگین فاصـله هر نمونه تا ریشه به عنوان امتیاز ناهنجاری در نظر گرفته می‌شود[۴۳].

۴-۳-۳- روش DSEBM[13]

اساس کار این روش بر استفاده از انرژی لایه‌هایی که در خودگذارهای حذف نویز به کار برده می‌شوند بنا نهاده شده است. از خطای بازسازی و همچنین خود انرژی به عنوان امتیاز ناهنجاری در این مدل استفاده شـده اسـت. DSEBM-r بیانگر تشـخیص ناهنجاری با خطای بازسـازی و DSEBM-e بیانگر تشـخیص ناهنجاری با امتیاز انرژی است[۲۶].

۴-۳-۴- روش DAGMM[14]

این مدل بر اسـاس خودکدگذارهای مورد اسـتفاده در تشـخیص ناهنجاری طراحی شـده اسـت. در گام اول مدل یک خودکدگذار را برای تولید فضـــای نهفته معقول و بازســـازی ویژگی‌ها آموزش می‌دهد. شـــبکه تخمین‌گر دیگری نیز در این مدل آموزش داده می‌شـــود که به عنوان خروجی پارامترهای GMM که فضـــای نهفته با ابعاد کوچک را مدل‌ســـازی می‌کند، تولید می‌کند. در مرحله آزمون میزان درســـتنمایی بازنمایی مدل در فضای نهفته و ویژگی‌های بازسازی شده توسط GMM محاسبه می‌شود و این مقدار به عنوان امتیاز ناهنجاری در نظر گرفته می‌شود[۴۲].

۴-۳-۵- روش DCAE[15]

مدل DCAE یک مدل کلاسیک خودکدگذار است که در آن کدگذار و کدگشا دارای ساختار کانولوشنی هستند. امتیاز ناهنجاری در این مدل نرم l_2 خطای بازسازی است[۴۴].

---

[13] Deep Structures Energy Based Models

[14] Deep Autoencoding Gaussian Mixture Model

[15] Deep Convolutional Autoencoder

۴-۳-۶- روش DSVDD[16]

در این روش یک شـــبکه عصـــبی که حجم ابر کره محیط بر داده هنجار را کمینه می‌کند آموزش داده می‌شود. امتیاز ناهنجاری در این مدل فاصله اقلیدسی میان مرکز این ابرکره تا داده ورودی است[۴۵].

۴-۴- نتایج

در این بخش به مقایسه مدل پیشنهادی با مدل‌های پایه بخش ۴-۳ روی مجموعه داده‌هایی‌که در بخش ۴-۱ معرفی شد، می‌پردازیم. در ادامه مقایسه‌ها را در دو بخش دادگان جدولی و تصویری ارائه می‌کنیم.

۴-۴-۱- دادگان جدولی

مدل پیشـنهادی بر روی چهار مجموعه داده جدولی شـامل KDDCup99، Arrhythmia، Thyroid و MUSK آزمایش شـــده اســـت. دادگان Thyroid و MUSK به دلیل نرخ پایین ناهنجاریشـــان انتخاب شده‌اند تا مقاومت[17] مدل در شرایط تفاوت فاحش در میزان نسبت داده‌های ناهنجار به هنجار نیز بررسی شـود. در جدول ۴-۱ نتایج حاصـل از آزمودن مدل‌های پایه و مدل پیشـنهادی را با سـه معیار اسـتاندارد صـحت، بازیابی و F1-score ارزیابی می‌شـوند. نتایج زیر مقادیر متوسـط به ازای ۳۰ اجرا برای هر یک از مدل‌ها می‌باشد.

[16] Deep Support Vector Data Description

[17] Robustness

جدول ۴-۱: نتایج خروجی مدل پیشنهادی در مقایسه با مدل‌های پایه بر روی مجموعه داده‌های جدولی.

Model	KDDCUP			Arrhythmia			Thyroid			Musk		
	Prec.	Recall	F_1	Prec.	Recall	F_1	Prec.	Recall	F_1	Prec.	Recall	F_1
IF	92.16	93.73	92.94	51.47	54.69	53.03	**70.13**	**71.43**	**70.27**	47.96	47.72	47.51
OC-SVM	74.57	85.23	79.54	53.97	40.82	45.18	36.39	42.39	38.87	–	–	–
DSEBMr	85.12	64.72	73.28	15.15	15.13	15.10	4.04	4.03	4.03	–	–	–
DSEBMe	86.19	64.46	73.99	46.67	45.65	46.01	13.19	13.19	13.19	–	–	–
AnoGAN	87.86	82.97	88.65	41.18	43.75	42.42	44.12	46.87	45.45	3.06	3.10	3.10
DAGMM	92.97	94.22	93.69	49.09	50.78	49.83	47.66	48.34	47.82	–	–	–
ALAD	94.27	**95.77**	95.01	50.00	53.13	51.52	22.92	21.57	22.22	58.16	59.03	58.37
DSVDD	89.81	94.97	92.13	35.32	34.35	34.79	22.22	23.61	23.29	–	–	–
RCALAD	**95.36**	95.62	**95.49**	**58.82**	**62.50**	**60.60**	53.76	51.53	52.62	**62.96**	**63.33**	**63.14**
error bar	0.28	0.29	0.28	6.6	6.8	5.8	4.3	2.7	2.8	5.06	2.53	2.62

نتایج جدول ۴-۱ بیانگر این مسئله است که مدل RCALAD در سه مورد از چهار مجموعه داده جدولی آزمایشی با اختلاف بهتر عمل کرده است و تنها در مجموعه داده Thyroid از مدل IF دقت کمتری به‌دست آورده است. با نگاه دقیق‌تر در این مجموعه داده می‌توان از علت این پدیده آگاه شد، ویژگی‌های این مجموعه داده شامل نتایج آزمایشگاهی برای بررسی بیماری تیرویید می‌باشد و شامل ۳۵ ویژگی است. اما ثابت شده است که وجود/عدم وجود این بیماری تنها تحت تاثیر دو ویژگی T3 و T4 است و به همین دلیل مدل IF که مبتنی بر انتخاب ویژگی مهم‌تر و با ارزش‌تر است، در این مسئله بهتر عمل می‌کند. یک ایده کلی برای بهبود نتایج مدل پیشنهادی روی این دادگان می‌تواند به کارگیری مدل IF در مرحله پیش‌پردازش و با هدف استخراج ویژگی‌های قوی‌تر باشد و پس از این مرحله از این ویژگی‌ها برای آموزش مدل RCALAD استفاده شود.

۴-۴-۱- دادگان تصویری

در این بخش عملکرد مدل پیشنهادی روی مجموعه داده تصویری شامل SVHN و CIFAR10 آزمایش شده است. نتایج آزمایش به تفکیک برای هر کلاس از داده‌ها مطابق جدول ۴-۲ و ۴-۳ گزارش می‌شود. این نتایج حاصل از میانگین ۳ مرتبه اجرای هر یک از مدل‌ها می باشد. همچنین میانگین معیار AUROC برای هر مجموعه داده در جدول ۴-۴ آورده شده است.

جدول ۴-۲: نتایج خروجی مدل پیشنهادی در مقایسه با مدل‌های پایه برروی مجموعه داده CIFAR10.

Normal	DCAE	DSEBM	DAGMM	IF	AnoGAN	ALAD	RCALAD
Airplane	59.1 ± 5.1	41.4 ± 2.3	56.0 ± 6.9	60.1 ± 0.7	67.1 ± 2.5	64.7 ± 2.6	$\mathbf{72.8 \pm 0.8}$
auto.	$\mathbf{57.4 \pm 2.9}$	57.1 ± 2.0	56.0 ± 6.9	50.8 ± 0.6	54.7 ± 3.4	45.7 ± 0.8	50.2 ± 0.3
Bird	48.9 ± 2.4	61.9 ± 0.1	53.8 ± 4.0	49.2 ± 0.4	52.9 ± 3.0	67.0 ± 0.7	$\mathbf{72.6 \pm 0.2}$
Cat	58.4 ± 1.2	50.1 ± 0.4	51.2 ± 0.8	55.1 ± 0.4	54.5 ± 1.9	59.2 ± 1.1	$\mathbf{64.2 \pm 0.9}$
Deer	54.0 ± 1.3	73.2 ± 0.2	52.2 ± 7.3	49.8 ± 0.4	65.1 ± 3.2	72.7 ± 0.6	$\mathbf{74.9 \pm 0.5}$
Dog	$\mathbf{62.2 \pm 1.8}$	60.5 ± 0.3	49.3 ± 3.6	58.5 ± 0.4	60.3 ± 2.6	52.8 ± 1.2	60.1 ± 1.1
Frog	51.2 ± 5.2	68.4 ± 0.3	64.9 ± 1.7	42.9 ± 0.6	58.5 ± 1.4	69.5 ± 1.1	$\mathbf{75.3 \pm 0.4}$
Horse	58.6 ± 2.9	53.3 ± 0.7	55.3 ± 0.8	55.1 ± 0.7	$\mathbf{62.5 \pm 0.8}$	44.8 ± 0.4	56.6 ± 0.2
Ship	$\mathbf{76.8 \pm 1.4}$	73.9 ± 0.3	51.9 ± 2.4	74.2 ± 0.6	75.8 ± 4.1	73.4 ± 0.4	$\mathbf{77.5 \pm 0.3}$
Truck	$\mathbf{67.3 \pm 3.0}$	63.6 ± 3.1	54.2 ± 5.8	58.9 ± 0.7	66.5 ± 2.8	43.2 ± 1.3	52.6 ± 0.6
Mean	59.4	60.3	54.4	55.5	61.8	59.3	**65.7**

همان‌طور که در جدول ۴-۲ مشاهده می‌شود، مدل پیشنهادی RCALAD می‌تواند در نگاه کلی (۷ کلاس از ۱۰ کلاس داده را) دیگر مدل‌های پایه از جمله ALAD و RCGAN را در مجموعه داده CIFAR10 مغلوب کند و حتی در کلاس‌هایی که به بهترین نتیجه دست نیافته است، به نتایج قابل قبولی دست یافته است. در ادامه عملکرد مدل پیشنهادی بر روی مجموعه داده SVHN مورد ارزیابی قرار گرفته است.

همانطور که از جدول ۴-۳ مشخص است اگرچه مدل پیشنهادی RCALAD در بیشتر کلاس‌ها عملکرد بهتری نسبت به سایر مدل‌های پایه دارد برای برخی از کلاس‌ها قادر به تشخیص نمونه‌های ناهنجار نیست. مقایسه وضعیت کلی هر یک از این مدل‌ها بر روی هر دو مجموعه‌داده در یک نگاه در سطر آخر جدول ۴-۳ انجام شده است. همان‌طور که می‌بینیم عملکرد کلی مدل پیشنهادی RCALAD از دیگر مدل‌ها بهتر است.

جدول ۴-۳: نتایج خروجی مدل پیشنهادی در مقایسه با مدل‌های پایه بر روی مجموعه داده SVHN.

Normal	OCSVM	DSEBMr	DSEBMe	IF	ANOGAN	ALAD	RCALAD
0	52.0 ± 1.6	56.1 ± 0.2	53.4 ± 1.8	53.0 ± 0.6	57.3 ± 0.4	58.7 ± 0.9	$\mathbf{60.4 \pm 0.1}$
1	48.6 ± 5.3	52.3 ± 0.9	52.1 ± 0.3	51.2 ± 0.9	57.0 ± 0.8	$\mathbf{62.8 \pm 1.7}$	59.2 ± 0.3
2	49.7 ± 7.7	51.9 ± 0.8	51.8 ± 0.4	52.3 ± 0.1	53.1 ± 0.4	$\mathbf{55.2 \pm 2.3}$	54.9 ± 0.1
3	50.9 ± 1.4	51.8 ± 0.4	51.7 ± 0.5	52.2 ± 0.3	52.6 ± 0.4	53.8 ± 3.3	$\mathbf{55.8 \pm 1.9}$
4	48.4 ± 5.2	52.5 ± 0.1	52.4 ± 0.2	49.1 ± 0.6	53.9 ± 0.5	58.0 ± 0.1	$\mathbf{58.5 \pm 0.2}$
5	51.1 ± 2.6	52.4 ± 2.3	52.3 ± 2.6	52.4 ± 0.8	52.8 ± 0.1	56.1 ± 0.9	$\mathbf{56.2 \pm 0.4}$
6	50.1 ± 3.9	52.1 ± 1.8	52.2 ± 1.8	51.8 ± 0.2	53.2 ± 0.0	57.4 ± 0.6	$\mathbf{59.4 \pm 0.5}$
7	49.6 ± 1.3	53.4 ± 0.9	55.3 ± 1.1	52.0 ± 0.4	55.0 ± 0.0	$\mathbf{58.8 \pm 0.3}$	58.0 ± 0.4
8	45.0 ± 4.2	51.9 ± 0.3	52.5 ± 0.6	52.3 ± 0.8	52.2 ± 0.7	$\mathbf{55.2 \pm 0.4}$	$\mathbf{56.1 \pm 0.5}$
9	52.5 ± 3.9	55.8 ± 1.7	52.7 ± 1.4	53.7 ± 0.6	53.1 ± 0.1	57.3 ± 0.6	$\mathbf{58.3 \pm 0.2}$
Mean	50.2	52.9	52.4	51.6	54.0	57.3	**57.7**

عملکرد ضعیف مدل پیشنهادی روی عدد سه می‌تواند به دلیل شباهت ظاهری میان عدد سه به عدد پنج و دو در زبان انگلیسی باشد. برای مشاهده نحوه عملکرد مدل روی کلاس عدد سه در شکل ۴-۴ را مشاهده کنید.

شکل ۴-۱: عملکرد مدل RCALAD روی کلاس عدد سه.

در شکل ۴-۱ ردیف اول داده هنجار، ردیف دوم بازسازی داده هنجار، ردیف سوم داده ناهنجار و ردیف چهارم بازسازی داده ناهنجار است.

۴-۵- بحث

در این بخش به بررسی کارایی و تاثیر جز به جز هر یک از عناصر موجود در مدل می‌پردازیم. سپس توابع توزیع جریمه مختلف و تاثیر انتخاب هر کدام بحث قرار می‌گیرد و نشان داده می‌شود که نتایج بدست آمده وابستگی چندانی به یک تابع توزیع جریمه خاص ندارند. در نهایت دو امتیاز ناهنجاری ارائه شده مقایسه و حوزه عملکرد متناسب با هر یک بررسی می‌شود.

۴-۵-۱- مطالعه فرسایشی[18]

در این قسمت تاثیر جزء به جزء قسمت‌های مختلف مدل را بر دقت نهایی به دست آمده بررسی می‌کنیم. آزمایش‌ها در این‌جا در شرایط حضور و عدم حضور هر جزء تکرار می‌شوند و نتایج حاصل از آن‌ها با هم مقایسه می‌شود تا میزان تاثیر هر قسمت به طور جداگانه مشخص شود. نماد D_{xxzz} نشان‌دهنده افزودن همین تمایزگر به مدل پایه ALAD می‌باشد. و همچنین نماد $\sigma(x)$ به معنای اضافه شدن توزیع کمکی برای پوشش فضای X می‌باشد.

جدول ۴-۴: تاثیر بخش‌های مختلف پیشنهادی در بهبود نتایج دادگان جدولی.

Model	Precision	Recall	F1 score
KDD99			
Baseline (ALAD)	0.942 ± 0.008	$\mathbf{0.957 \pm 0.006}$	0.950 ± 0.007
Baseline + D_{xxzz} (CALAD)	$\mathbf{0.959 \pm 0.004}$	0.957 ± 0.007	$\mathbf{0.958 \pm 0.005}$
Baseline + $\sigma(x)$ (RALAD)	0.943 ± 0.005	0.955 ± 0.004	0.949 ± 0.004
Baseline + D_{xxzz} + $\sigma(x)$ (RCALAD)	0.953 ± 0.007	0.956 ± 0.005	0.954 ± 0.006
Arrhythmia			
Baseline (ALAD)	0.500 ± 0.049	0.531 ± 0.047	0.515 ± 0.048
Baseline + D_{xxzz} (CALAD)	0.574 ± 0.021	0.605 ± 0.022	0.575 ± 0.021
Baseline + $\sigma(x)$ (RALAD)	0.546 ± 0.035	0.565 ± 0.039	0.555 ± 0.037
Baseline + D_{xxzz} + $\sigma(x)$ (RCALAD)	$\mathbf{0.588 \pm 0.42}$	$\mathbf{0.625 \pm 0.41}$	$\mathbf{0.606 \pm 0.41}$
Thyroid			
Baseline (ALAD)	0.229 ± 0.067	0.215 ± 0.067	0.222 ± 0.067
Baseline + D_{xxzz} (CALAD)	0.529 ± 0.071	$\mathbf{0.518 \pm 0.075}$	0.523 ± 0.073

[18] Ablation studies

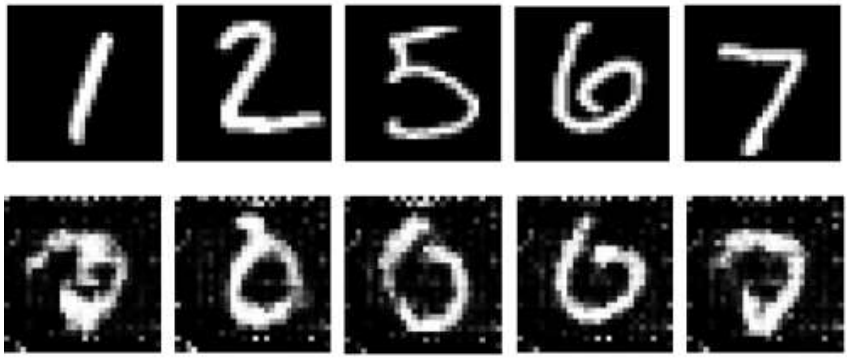
Baseline + $\sigma(x)$ (RALAD)	0.431 ± 0.039	0.457 ± 0.043	0.443 ± 0.041
Baseline + D_{xxzz} + $\sigma(x)$ (RCALAD)	$\mathbf{0.537 \pm 0.054}$	0.515 ± 0.057	$\mathbf{0.526 \pm 0.055}$
Musk			
Baseline (ALAD)	0.500 ± 0.068	0.531 ± 0.070	0.515 ± 0.069
Baseline + D_{xxzz} (CALAD)	0.574 ± 0.026	0.605 ± 0.027	0.575 ± 0.026
Baseline + $\sigma(x)$ (RALAD)	0.546 ± 0.051	0.565 ± 0.051	0.555 ± 0.051
Baseline + D_{xxzz} + $\sigma(x)$ (RCALAD)	$\mathbf{0.629 \pm 0.011}$	$\mathbf{0.633 \pm 0.016}$	$\mathbf{0.631 \pm 0.013}$

نتایج جدول ۴-۴ نشان می‌دهد اضافه کردن تمایزگر D_{xxzz} به طور قابل ملاحظه‌ای کارایی مدل را روی دادگان جدولی افزایش داده است. همچنین این تمایزگر نتایج روی مجموعه داده‌های تصویری CIFAR-10 و SVHN را نیز به شکل مناسبی بهبود داده است. همان‌طور که در جدول ۴-۵ مشاهده می‌شود، نتایج مجموعه داده جدولی درخشان‌تر از مجموعه داده‌های تصویری می‌باشد. شایان ذکر است که اگرچه بهبود حاصل‌شده در دادگان تصویری کمتر می‌باشد، اما در این جنس مجموعه داده‌ها حتی بهبودهای جزئی‌تر از این هم باارزش بوده و قابل توجه می‌باشد.

جدول۴-۵ : تاثیر بخش‌های مختلف پیشنهادی در بهبود نتایج دادگان تصویری.

Model	AUROC
SVHN	
Baseline (ALAD)	0.573 ± 0.016
Baseline + D_{xxzz} (CALAD)	0.576 ± 0.014
Baseline + $\sigma(x)$ (RALAD)	0.568 ± 0.018
Baseline + D_{xxzz} + $\sigma(x)$ (RCALAD)	$\mathbf{0.577 \pm 0.019}$
CIFAR-10	
Baseline (ALAD)	0.593 ± 0.017
Baseline + D_{xxzz} (CALAD)	0.634 ± 0.018
Baseline + $\sigma(x)$ (RALAD)	0.642 ± 0.012
Baseline + D_{xxzz} + $\sigma(x)$ (RCALAD)	$\mathbf{0.657 \pm 0.016}$

برای مشاهده هر چه بهتر تاثیر توزیع $\sigma(x)$ بر بازسازی نمونه‌های ناهنجار شکل ۴-۲ آورده شده است. همانطور که در شکل مشخص است این توزیع مدل را متمایل به تولید کلاس هنجار صفر به ازای همه ورودی‌های ناهنجار کرده است.



شکل ۴-۲: تاثیر توزیع $\sigma(x)$ بر بازسازی نمونه‌های ناهنجار.

۴-۵-۲- انتخاب تابع توزیع جریمه

در این زیربخش به آزمایش تابع توزیع‌های مختلف $\sigma(x)$ و میزان تاثیر آن‌ها بر نتایج نهایی مدل می‌پردازیم. توزیع‌های مورد استفاده در این‌جا دو توزیع گاوسی به صورت $N(0,I)$ ، $N(0,2I)$ و یک توزیع یکنواخت به صورت $U(-1,+1)$ است. با استفاده از هر کدام از توابع توزیع یاد شده در مدل ارائه شده همانطور که در جدول ۴-۶ مشخص است، بهبودهای پایداری حاصل شده است. تابع توزیع $\sigma(x)$ به منظور تقلید و یا تخمین توزیع داده ناهنجار طراحی نشده است و می‌تواند مستقل از توزیع داده ناهنجار حتی در شزایطی که متفاوت با آن است تاثیر مورد نظر خود را بر مدل بگذارد.

جدول ۴-۶: تاثیر $\sigma(x)$ های مختلف بر عملکرد مدل RCALAD.

$t(x)$	KDDCUP			Arrhythmia		
	Prec.	Recall	F_1	Prec.	Recall	F_1
$\mathcal{N}(\mathbf{0},\boldsymbol{I})$	**0.629**	**0.633**	**0.631**	**0.588**	0.625	0.606
$\mathcal{N}(\mathbf{0},2\boldsymbol{I})$	0.626	**0.633**	0.629	0.580	0.629	0.603
$\mathcal{U}(-\mathbf{1},\mathbf{1})$	0.608	0.604	0.606	0.584	**0.633**	**0.607**

۴-۵-۳- ارزیابی کارایی امتیازهای ناهنجاری

در این قسمت امتیازهای ناهنجاری پیشنهادی در این پروژه با دیگر معیارهای مبتنی بر بازسازی مقایسه می‌شود. خروجی خام تمایزگرها را در این‌جا لاجیت[19] می‌نامیم و غرض از ویژگی[20]، ویژگی‌های تولیدی تمایزگر در لایه‌های پنهان است. x نمونه ورودی، z_x نگاشت این نمونه در فضای نهان، $\hat{x} = G(E(x))$ بازسازی تولیدی همان نمونه و $\hat{z}_{\hat{x}}$ نگاشت تصویر بازسازی شده در فضای نهان به وسیله مدل RCALAD است. امتیازهای مورد استفاده در این بخش به شرح زیر است.

$$
\begin{aligned}
&L_1 : A(x) = \|x - \hat{x}\|_1 \\
&L_2 : A(x) = \|x - \hat{x}\|_2 \\
&\text{Logits} : A(x) = \log\left(D_{xx}(x, \hat{x})\right) \\
&\text{Features} : A(x) = \|f_{xx}(x, x) - f_{xx}(x, \hat{x})\|_1 \\
&\text{FM} : A_{fm}(x) = \|f_{xxzz}(x, x, z_x, z_x) - f_{xxzz}(x, \hat{x}, z_x, \hat{z}_{\hat{x}})\|_1 \\
&\text{ALL} : A_{ALL}(x) = D_{xxzz}(x, \hat{x}, z_x, \hat{z}_{\hat{x}}) + D_{zz}(z_x, \hat{z}_{\hat{x}}) + D_{xx}(x, \hat{x})
\end{aligned}
\qquad (۱-۴)
$$

در ادامه ارزیابی نتایج هر یک از امتیازها روی دادگان جدولی مطابق جدول ۴-۷ قابل مشاهده است.

جدول ۴-۷: مقایسه عملکرد امتیازهای ناهنجاری پیشنهادی با سایر امتیازها روی دادگان جدولی.

Model	Precision	Recall	F1 score
KDD99			
L_1	0.9081 ± 0.0638	0.9108 ± 0.0638	0.9094 ± 0.0638
L_2	0.9011 ± 0.0155	0.9004 ± 0.0157	0.9007 ± 0.0156
Logits	0.9169 ± 0.0162	0.9168 ± 0.0164	0.9168 ± 0.0163
Features	0.9127 ± 0.0029	0.9177 ± 0.0039	0.9151 ± 0.0034
Features_xxzz	$\mathbf{0.9327 \pm 0.0017}$	$\mathbf{0.9377 \pm 0.0017}$	$\mathbf{0.9301 \pm 0.0017}$
Logits_all	0.9231 ± 0.0018	0.9207 ± 0.0018	0.9218 ± 0.0018
Arrhythmia			
L_1	0.3529 ± 0.0148	0.3750 ± 0.0164	0.3636 ± 0.0256
L_2	0.3529 ± 0.0107	0.3750 ± 0.0108	0.3636 ± 0.0107

[19] Logit

[20] Feature

Logits	0.5588 ± 0.0334	0.5937 ± 0.0386	0.5757 ± 0.0359
Features	0.2325 ± 0.0029	0.2500 ± 0.0029	0.2424 ± 0.0029
Features_xxzz	0.4411 ± 0.0013	0.4687 ± 0.0013	0.4545 ± 0.0013
Logits_all	$\mathbf{0.6176 \pm 0.0208}$	$\mathbf{0.6562 \pm 0.0221}$	$\mathbf{0.6363 \pm 0.0214}$
Thyroid			
L_1	0.4981 ± 0.0028	0.4908 ± 0.0024	0.4994 ± 0.0024
L_2	0.5011 ± 0.0330	0.5004 ± 0.0318	0.5007 ± 0.0324
Logits	0.4969 ± 0.0142	0.4968 ± 0.0144	0.4968 ± 0.0143
Features	0.5127 ± 0.0119	0.5177 ± 0.0119	0.5151 ± 0.0119
Features_xxzz	0.5227 ± 0.0083	0.5123 ± 0.0083	0.5174 ± 0.0083
Logits_all	$\mathbf{53.76 \pm 0.0029}$	$\mathbf{51.53 \pm 0.0029}$	$\mathbf{52.62 \pm 0.0029}$
Musk			
L_1	0.5979 ± 0.0103	0.5931 ± 0.0109	0.5954 ± 0.0106
L_2	0.6008 ± 0.0021	0.6018 ± 0.0028	0.6013 ± 0.0024
Logits	0.5868 ± 0.0124	0.5897 ± 0.0127	0.5882 ± 0.0125
Features	0.5824 ± 0.0011	0.5883 ± 0.0019	0.5883 ± 0.0015
Features_xxzz	0.6111 ± 0.0481	0.6187 ± 0.0468	0.6148 ± 0.0474
Logits_all	$\mathbf{62.96 \pm 0.0013}$	$\mathbf{63.33 \pm 0.0013}$	$\mathbf{63.14 \pm 0.0013}$

همانطور که در جدول ۴-۷ مشاهده می‌شود روی دادگان جدولی خروجی تمایزگر D_{xxzz} دارای بهترین نتایج به نسبت سایر امتیازهای ناهنجاری است. با توجه به اینکه تعداد ویژگی‌ها روی دادگان جدولی به نسبت دادگان تصویری کمتر است تمایزگر D_{xxzz} قادر به تشخیص مناسب نمونه‌های ناهنجار است. نتایج روی دادگان تصویری در جدول ۴-۸ قابل مشاهده است.

جدول ۴-۸: مقایسه عملکرد امتیازهای ناهنجاری پیشنهادی با سایر امتیازها روی دادگان تصویری.

Anomaly Score	AUROC
SVHN	
L_1	0.5778 ± 0.0141
L_2	0.5636 ± 0.0251
Logits	0.5369 ± 0.0785
Features	0.5763 ± 0.0367
Logits_all	0.5768 ± 0.0251
Features_xxzz	0.5778 ± 0.0161
CIFAR-10	
L_1	63.41 ± 0.0321
L_2	63.27 ± 0.0782
Logits	62.97 ± 0.0643
Features	63.12 ± 0.0368
Logits_all	64.77 ± 0.0227
Features_xxzz	65.73 ± 0.0194

همانطور که در جدول ۴-۸ مشخص است عملکرد امتیاز مبتنی بر ویژگی‌ها روی دادگان تصویری بسیار مناسب است، این می‌تواند به این دلیل باشد بردار ویژگی‌ها برای هر عکس نسبت به بردار ویژگی‌های موجود در دادگان جدولی بزرگتر است و استفاده از پارامترهای بیشتر به منظور تشخیص ناهنجاری سبب بهبود عملکرد امتیاز مورد نظر شده است.

۴-۵-۴- ارزیابی کفایت تمایزگر D_{xxzz}

در این بخش به بررسی سوالی که در بخش ۳-۱-۲ مطرح شده می‌پردازیم، توضیح دقیق مسئله این بخش بدین ترتیب است: آیا با وجود D_{xxzz} همچنان به دو تمایزگر دیگر نیز احتیاجی هست؟ در واقع

D_{xxzz} برای تامین شـــرط پایداری چرخه کامل کافی نیســـت؟ آیا دیگر تمایزگرها اطلاعاتی در مدل استخراج کرده و یا اضافه بوده و می‌توان حذف شوند؟

برای پاسخ‌گویی به این سوالات مطابق جدول ۴-۹ به ترتیب هر یک از تمایزگرها را از مدل کنار گذاشته و عملکرد مدل را گزارش می‌کنیم.

جدول ۴-۹: ارزیابی عملکرد مدل در حضور/عدم حضور هر یک از اجزا.

Model	D_{zz}	D_{xx}	D_{xxzz}	Prec.	Recall	F_1
KDD99						
ALAD	√	√	×	0.942 ± 0.008	$\mathbf{0.957 \pm 0.006}$	0.950 ± 0.007
ALI + D_{xxzz}	×	×	√	0.938 ± 0.007	0.951 ± 0.010	0.944 ± 0.009
ALI + D_{zz} + D_{xxzz}	√	×	√	0.946 ± 0.005	0.955 ± 0.004	0.950 ± 0.004
ALICE + D_{xxzz}	×	√	√	0.941 ± 0.005	0.954 ± 0.008	0.947 ± 0.006
CALAD	√	√	√	$\mathbf{0.959 \pm 0.004}$	0.957 ± 0.007	$\mathbf{0.958 \pm 0.005}$
RCALAD	√	√	√	0.953 ± 0.007	0.956 ± 0.005	0.954 ± 0.006
Arrhythmia						
ALAD	√	√	×	0.500 ± 0.049	0.531 ± 0.047	0.515 ± 0.048
ALI + D_{xxzz}	×	×	√	0.522 ± 0.054	0.529 ± 0.049	0.525 ± 0.052
ALI + D_{zz} + D_{xxzz}	√	×	√	0.571 ± 0.033	0.582 ± 0.028	0.576 ± 0.031
ALICE + D_{xxzz}	×	√	√	0.543 ± 0.052	0.561 ± 0.044	0.551 ± 0.048
CALAD	√	√	√	0.574 ± 0.021	0.605 ± 0.022	0.575 ± 0.021
RCALAD	√	√	√	$\mathbf{0.588 \pm 0.42}$	$\mathbf{0.625 \pm 0.41}$	$\mathbf{0.606 \pm 0.41}$

مطابق نتایج تئوری، افزودن تمایزگر D_{xxzz}به چارچوب کلی و در کنار دیگر تمایزگرها بالاترین کارایی را داشـته اسـت. پس از آن حذف D_{xx} ضـربه کمتری به مدل می‌زند زیرا بخشـی از اطلاعاتی که اسـتخراج می‌کند، توسط تمایزگر D_{xxzz} پوشش داده می‌شود. همان‌طور که نتایج نشـان می‌دهد، نتیجه این بخش این اسـت که تمایزگر D_{xxzz} به تنهایی کافی نیسـت و این سـه تمایزگر در کنار هم بیشـترین کارایی را دارند و تمایزگر D_{xxzz} به تنهایی نمی‌تواند تمامی جنبه‌ها را دیده و اطلاعات مورد نیاز را استخراج کند.

۴-۶- جمع‌بندی

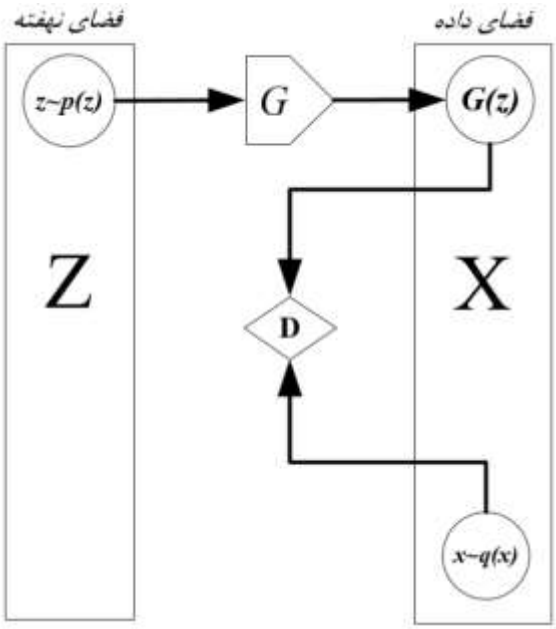
در این فصل ابتدا مجموعه داده‌های استفاده شده در این پروژه معرفی شد. همچنین تنظیمات مورد نیاز برای هر مجموعه داده در تقسیم نمونه‌ها به مجموعه آموزشی و آزمون و چگونگی انتخاب تعداد نمونه‌های ناهنجار در هر دادگان مورد بررسی قرار گرفت. در گام بعدی مدل‌های پایه‌ای که به منظور مقایسه با مدل پیشنهادی به کار برده شده‌اند به طور مختصر بررسی شدند. در قدم بعدی مدل پیشنهادی با سایر مدل‌ها مقایسه و برتری کلی آن ثابت شد سپس میزان مشارکت هر قسمت از مدل پیشنهادی بر بهبود نهایی بررسی شد. در مرحله بعدی توابع مختلف به عنوان توزیع $\sigma(x)$ امتحان شدند و ثابت شد بهبود حاصل از این توزیع مستقل از توزیع داده‌های ناهنجار است. سپس امتیازهای ناهنجاری پیشنهادی در مقایسه با سایر امتیازهای ناهنجاری مبتنی بر بازسازی آزموده شدند و کارایی آن‌ها تایید شد. در گام آخر به این سوال پاسخ داده شد که با وجود تمایزگر D_{xxzz} به دیگر تمایزگرها نیاز است یا خیر، که نتایج آزمایش‌ها نشان داد که بهترین نتیجه در حضور هر سه تمایزگر بدست می‌آید.

فصل پنجم:
جمع‌بندی، نتیجه‌گیری و کارهای آتی

در بخش پایانی به بر مرور و جمع‌بندی هر آنچه در این پروژه گفته شـــد می‌پردازیم و خط مش کلی کارهای آتی را مشخص می‌کنیم.

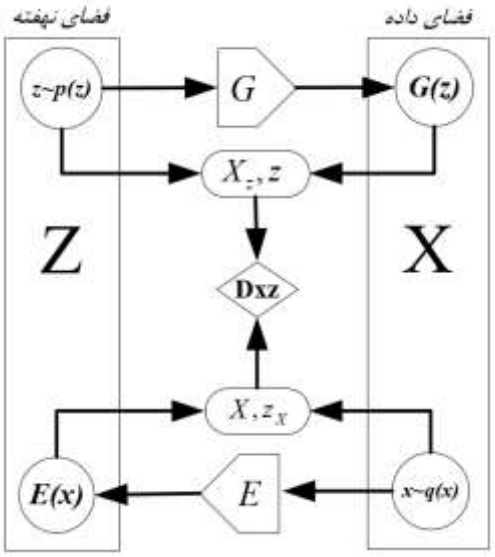
۵-۱- جمع‌بندی و نتیجه‌گیری

در بخش اول به اهمیت مسئله تشخیص ناهنجاری و کاربردهای آن در دنیای واقعی پرداختیم و سپس در همین فصل جایگاه شبکه‌های مولد تقابلی در زمینه تشخیص ناهنجاری مشخص شد. در فصل دوم به دسته‌بندی روش‌های که تا کنون در تشخیص ناهنجاری به کار برده شده‌اند از دیدگاه‌های مختلف و همچنین معرفی معیارهای پرکاربرد در ارزیابی مدل‌های تشـــخیص ناهنجاری پرداختیم. در خلال بررســـی‌ها لزوم وجود روش‌هایی برای مدل‌ســـازی داده‌هایی پیچیده و با ابعاد بالا احســـاس و در نتیجه توجه‌ها به سمت شبکه‌های مولد تقابلی که قادر به انجام این مهم هستند معطوف شد. در ادامه همین فصل به بررسی تعریف و اصول این نوع از شبکه‌ها پرداخته شد و روند تکاملی و محلی‌سازی آن‌ها با هدف تشـــخیص ناهنجاری مورد اشـــاره قرار گرفت. همانطور که مشـــاهده شـــد شـــبکه مولد تقابلی که معماری آن در شکل ۵-۱ مشخص است در سال ۲۰۱۴ معرفی شد[۳۱].



شکل ۵-۱: معماری اولیه شبکه مولد تقابلی.

مطابق با آنچه در فصل دو بیان شد با هدف تشخیص ناهنجاری علاوه شبکه مولد که وظیفه نگاشت از فضای نهفته به فضایی داده ورودی را بر عهده دارد نیازمند فرایند استنتاجی برای نگاشت معکوس از فضای داده ورودی به فضای نهفته هستیم. مدل AnoGan با استفاده از یک فرایند مبتنی بر تکرار نقطه متناظر با داده ورودی را در فضای نهفته محاسبه می‌کرد. مشکل مدل مورد بحث پیچیدگی محاسباتی بالا و همچنین تا حد زیادی تصادفی بودن فرایند آن است. مدل‌های ALI و f-anogan و EGBAD با استفاده از یک کدگذار نگاشت معکوس به فضای نهفته آموخته می‌شود. معماری مدل ALI در شکل ۵-۲ قابل مشاهده است[۱۴], [۳۵], [۳۶].

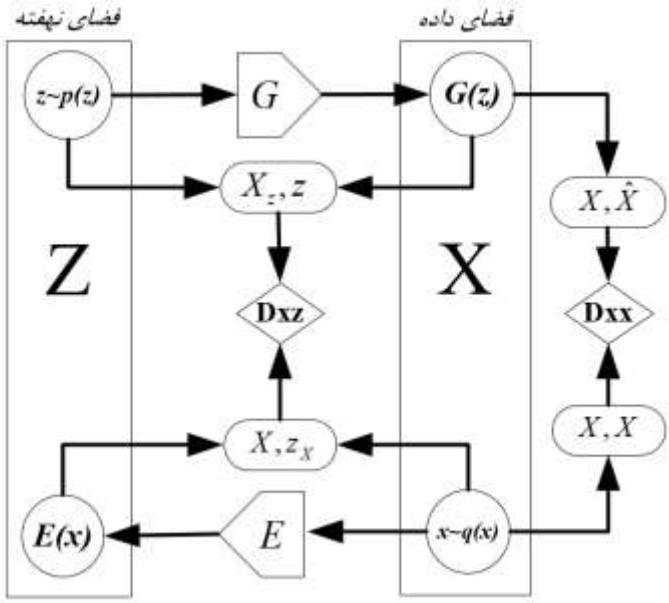


شکل ۵-۲: معماری مدل ALI.

این کارها اگرچه در جزئیات متفاوتند اما چالش اصلی آن‌ها یکسان است. در ادامه با توجه به اینکه هیچ ساختاری برای کنترل مشابهت میان تصویر ورودی و تصویر بازسازی شده توسط شبکه مولد تا به حال وجود نداشته است مدل ALICE به رفع این نقصان پرداخته است. مدل ALICE با اضافه کردن یک تمایزگر که داده ورودی و بازسازی آن را به عنوان ورودی دریافت می‌کند مشکل مورد نظر را که شرط پایداری حلقه نام داشت برطرف نمود. معماری ALICE د شکل ۵-۳ آورده شده است[۳۷].

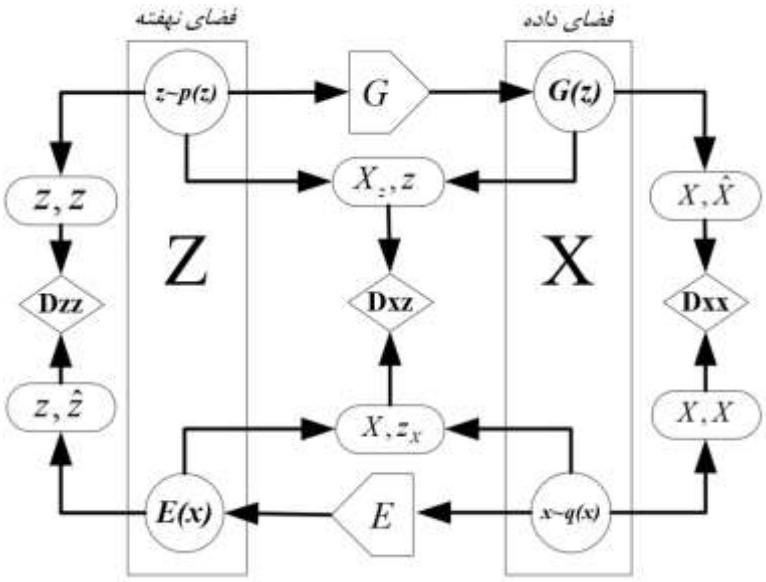
در فصل سه ابتدا دو مدل پایه مورد استفاده در این شبکه به طور دقیق مورد بررسی قرار گرفت. مدل ALAD بر پایه مدل قبلی بنا نهاده شده است و در معماری خود برای تضمین بیشتر پایداری حلقه تمایزگر دیگری برای فضای نهفته اضافه کرده است و همچنین ورودی‌های تمازگرهای موجود در شبکه

را به صورت توام در نظر گرفته است. معماری پیشنهادی ALAD سبب افزایش بازدهی در زمان آزمایش شده است و همچنین روند آزمایش را تثبیت کرده است. معماری این شبکه نیز در ۵-۴ نمایش داده شده است[۳۲].



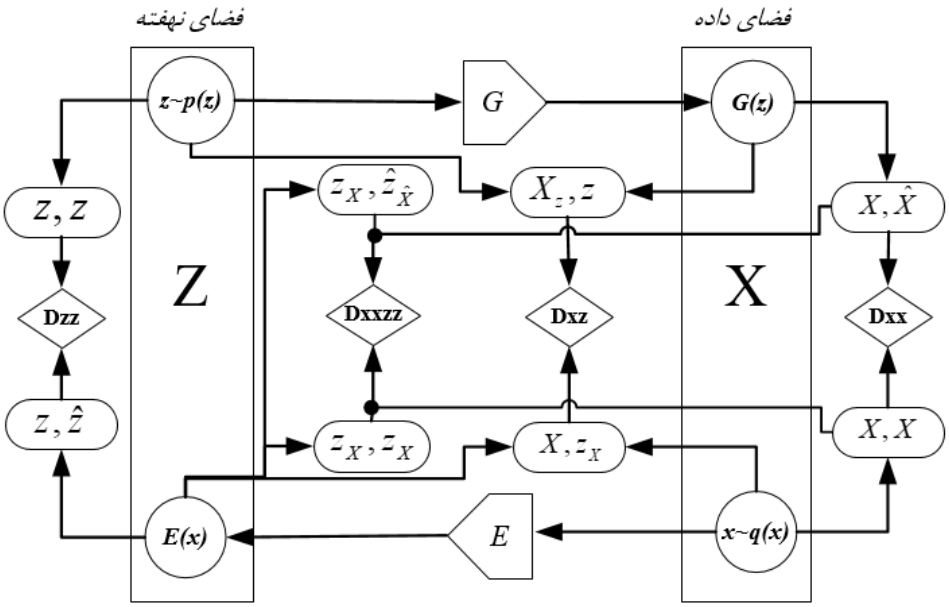
شکل۵-۳: معماری شبکه ALICE.

مدل پایه دیگر که در این فصل بررسی شده است، مدل RCGAN است[۳۸]. این شبکه با هدف ضمانت بازسازی ضعیف برای نمونه‌های ناهنجار تابع توزیع $t(x)$ را به ساختار تقابلی ALICE اضافه کرد تا مدل را به گونه‌ای متمایل به سمت بازسازی تمامی نمونه‌های ورودی به فضای داده هنجار در فضای داده ورودی کند. با انجام این کار فاصله میان داده ناهنجار و بازسازی آن زیاد خواهد شد و در نتیجه شناسایی نمونه ناهنجار ساده‌تر خواهد بود.



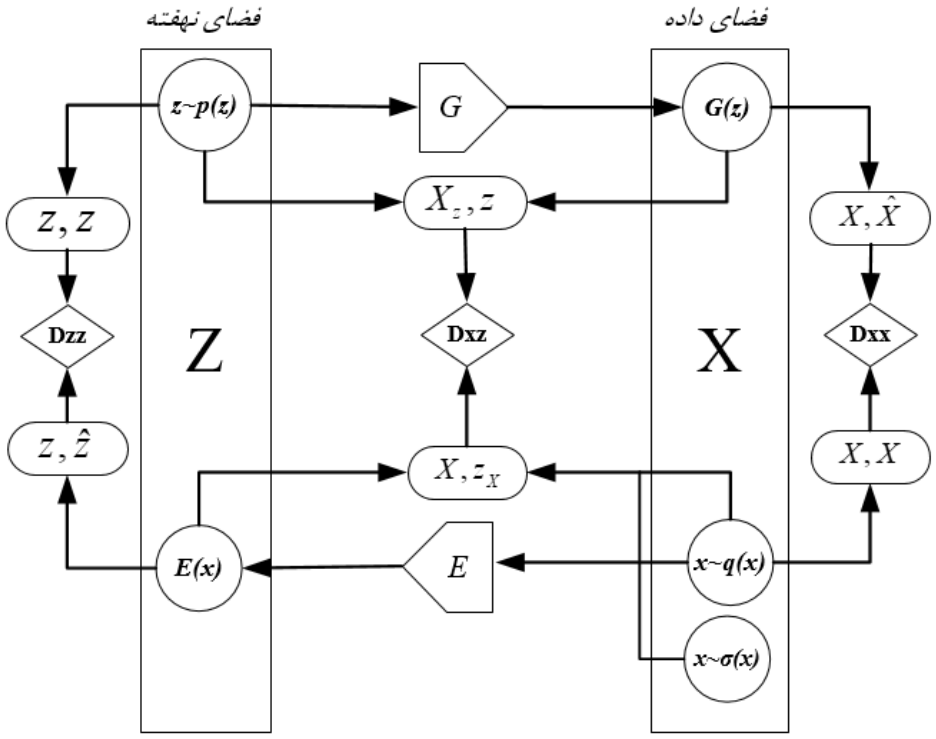
شکل ۵-۴: معماری شبکه ALAD.

پس از این قسمت‌ها نوبت به معرفی مدل‌های پیشنهادی این پروژه می‌رسد. همانطور که در فصل سه گفته شد به منظور تقویت قدرت تشخیص تمایزگر اطلاعات حاصل از روند دگردیسی داده ورودی در تمامی مراحل چرخه، باید توسط تمایزگر قابل دسترس باشد. چرخه مورد نظر در این مسئله شامل سه گام متوالی است، در کارهای قبلی از خروجی‌های چرخه اطلاعات به طور کامل در شبکه استفاده نمی‌شد. به منظور پوشش این نقص تمایزگر D_{xxzz} به ساختار تقابلی قبلی اضافه شد. نتیجه افزودن این تمایزگر مدل CALAD شد. معماری این مدل در شکل ۵-۵ نمایش داده شده است.



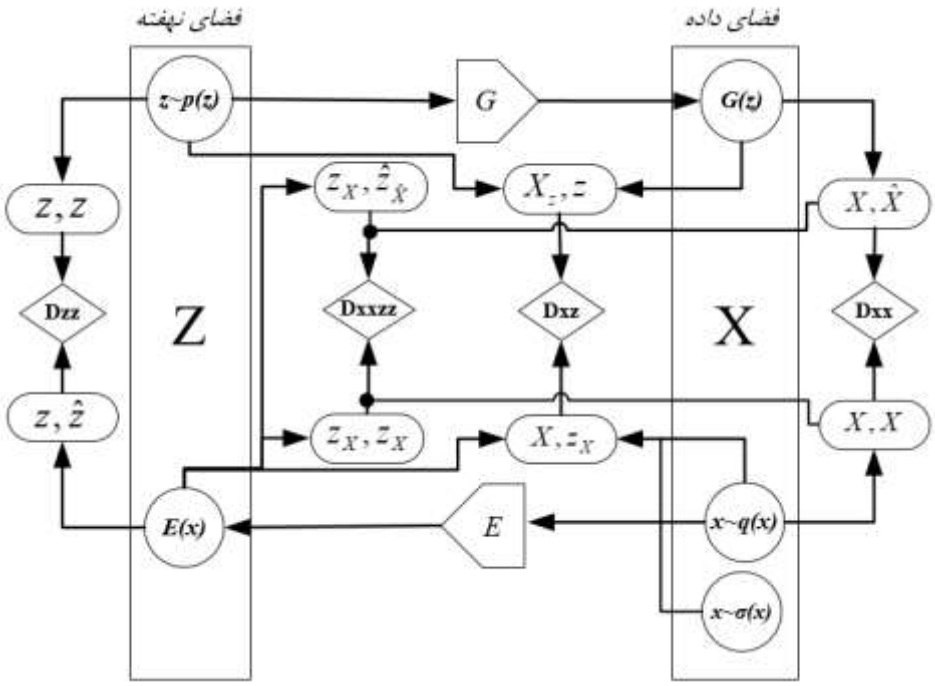
شکل ۵-۵: معماری شبکه CALAD.

دیگر مدل معرفی شده در این تحقیق با استفاده از توزیع $\sigma(x)$ برای ضمانت بازسازی ضعیف داده ناهنجار استفاده می‌کند. ساختار مدل RALAD در شکل ۵-۶ نشان داده شده است.



شکل ۵-۶: معماری شبکه RALAD.

در نهایت مدل جامع RCALAD که از تجمیع هر دو ایده بکارگیری متغیر $\hat{z}_{\hat{x}}$ و همچنین توزیع $\sigma(x)$ بدست می‌آید و همانطور که در بخش چهارم نشان داده شد، بیشترین کارایی د میان مدل‌ها را بدست می‌آورد. معماری این مدل در شکل ۵-۷ نمایش داده شده است.



شکل ۵-۷: معماری شبکه RCALAD.

توابع بهینه‌سازی هر یک از شبکه‌های مورد بحث در جدول ۵-۱ آمده است. همانطور که مشخص است توابع بهینه‌سازی این شبکه‌ها کاملا در امتداد هم و در راستای رفع نقاط ضعف کارهای قبلی هستند.

جدول ۵-۱ : روند تکامل توابع بهینه‌سازی شبکه‌های مولد تقابلی.

نام شبکه	تابع بهینه‌سازی
GAN	$\min_G \max_D V_{GAN}(D,G) = \mathbb{E}_{x\sim q(x)}[log\ D(x)] + \mathbb{E}_{z\sim \mathrm{p(z)}}[log\ (1 - D(G(z)))]$
ALI	$\min_G \max_D \mathrm{V}_{ALI}(\mathrm{D},\mathrm{G}) = \mathbb{E}_{x\sim q(x)}[\log \mathrm{D}(\mathrm{x}, G_z(\mathrm{x}))] + \mathbb{E}_{z\sim \mathrm{p(z)}}[\log (1 - \mathrm{D}(G_z(z), z)]$
ALICE	$\min_{E,G} \max_{D_{xz},D_{xx}} V_{ALICE} = \mathrm{V_{ALI}} + \mathbb{E}_{x\sim q(x)}[\log\ \mathrm{D}_{xx}(\mathrm{x},\mathrm{x}) + \log\ 1 - \mathrm{D}_{xx}(\mathrm{x}, \mathrm{G}(\mathrm{E}(\mathrm{x})))]$
ALAD	$min_{G,E} max_{D_{xz},D_{xx},D_{zz}} V_{ALAD}(D_{xz}, D_{xx}, D_{zz}, E, G) = V_{ALICE} + \mathbb{E}_{z\sim \mathrm{p(z)}}[log(D_{zz}(z,z)) + log(1 - D_{zz}(z, E(G(z))))]$
CALAD	$min_{G,E} max_{D_{xxzz},D_{xz},D_{xx},D_{zz}} V_{CALAD}(D_{xxzz}, D_{xz}, D_{xx}, D_{zz}, E, G) = V_{ALAD} + \mathbb{E}_{x\sim \mathrm{q(x)}}[\log D_{xxzz}(x,x,E(x),E(x))] + \mathbb{E}_{x\sim \mathrm{q(x)}}\left[1 - \log D_{xxzz}\left(x, G(E(x)), E(x), E\left(G(E(x))\right)\right)\right]$
RALAD	$min_{G,E} max_{,D_{xz},D_{xx},D_{zz}} V_{RALAD}(D_{xz}, D_{xx}, D_{zz}, E, G) = V_{ALAD} + \mathbb{E}_{\boldsymbol{x}\sim\sigma(x)}\left[\log\left(1 - D_{\boldsymbol{xz}}(\boldsymbol{x}, E(\boldsymbol{x}))\right)\right]$
RCALAD	$min_{G,E} max_{D_{xxzz},D_{xz},D_{xx},D_{zz}} V_{RCALAD}(D_{xxzz}, D_{xz}, D_{xx}, D_{zz}, E, G) = V_{ALAD} + \mathbb{E}_{\boldsymbol{x}\sim\sigma(x)}\left[\log\left(1 - D_{\boldsymbol{xz}}(\boldsymbol{x}, E(\boldsymbol{x}))\right)\right] + \mathbb{E}_{x\sim \mathrm{q(x)}}[\log D_{xxzz}(x,x,E(x),E(x))] + \mathbb{E}_{x\sim \mathrm{q(x)}}\left[1 - \log D_{xxzz}\left(x, G(E(x)), E(x), E\left(G(E(x))\right)\right)\right]$

نتایج عملی که در فصل چهارم مشاهده کردیم بیانگر کارایی مدل RCALAD در زمینه تشخیص ناهنجاری است. از مجموع مطالب گفته شده تا اینجا می‌توان چنین برداشت کرد استفاده توام از اطلاعات موجود در ساختارهای تقابلی سبب بهبود عملکرد آن‌ها می‌شود. علاوه بر این وجود توزیع $\sigma(x)$ که

مستقل از توزیع داده ناهنجار است سبب می‌شود که مدل به سمت تولید بازسازی‌ها در فضای داده هنجار سوق داده شود.

۵-۲-کارهای آتی

در کارهای آتی تلاش خواهد شد با روش‌های معرفی شده در مقاله [40] آموزش شبکه به شکل هر چه بهتر صورت پذیرد. در این کار با استفاده از تطبیق ویژگی به جای آن‌که شبکه مولد روی تمایزگر آموزش بیش از حد ببیند، تلاش می‌شود تا آمارگان توزیع داده ورودی نیز به شبکه مولد آموزش داده شود. علاوه بر این با استفاده از روش تمایز کوچک دسته‌ای[1] مولد را مجبور به تولید خروجی‌های متفاوت خواهیم کرد تا کار شبکه تمایزگر سخت‌تر و روند آموزش بهبود یابد. میانگین‌گیری تاریخی[2] از دور باطل حول یک نقطه بهینه جلوگیری می‌کند و انتظار می‌رود در صورت استفاده از این روش به بهینه محلی مناسب‌تری دست یابیم. همچنین روش نرمال‌سازی مجازی دسته[3] سبب می‌شود تا نمونه‌های موجود در یک دسته مستقل از هم شوند و در نتیجه روند بهینه‌سازی شبکه عصبی بهبود یابد. کارایی این روش روی DCGAN ثابت شده است. انتظار می‌رود با پیاده‌سازی روش‌های نام برده شده در این بخش، روند تشخیص ناهنجاری به حالت بهینه نزدیک‌تر شود.

روش دیگر که برای بهبود روش آموزش در مدل RCALAD پیشنهاد می‌شود، یادگیری ضریب اهمیت برای هر یک از تمایزگرها با توجه به جنس مسئله است. در واقع با توجه به نتایجی که در بخش ۴-۵-۳ بدست آمد و دیده شد که میزان اهمیت تمایزگرها در مسائل از جنس مختلف (داده جدولی و تصویری) متفاوت است، بنظر می‌رسد که با یادگیری ضریب اهمیت برای هر یک از تمایزگرها در تابع هزینه و استفاده از آن‌ها برای محاسبه امتیاز ناهنجاری، می‌توان به دقت‌های بالاتری دست یافت. دیگر ایده به کار گرفته شده در حوزه تشخیص ناهنجاری در سال‌های اخیر، استفاده از داده‌های کمکی یعنی نمونه‌های ناهنجار شناخته شده (هرچند تعداد آن‌ها بسیار کم باشد) با استفاده از روش‌های یادگیری روی داده‌های

[1] Minibatch discrimination

[2] Historical averaging

[3] Virtual batch normalization

نامتوزان[۴] می‌باشـــد. در [۴۶] دو تابع هزینه جدید با هدف یادگیری نامتوازن در کاربرد تشـــخیص ناهنجاری در شـــبکه‌های مولد تقابلی با نام‌های Patch loss و Anomaly adversarial loss معرفی شده که هر دوی این تابع‌ها قابلیت به کارگیری در چارچوب معرفی شده RCALAD را دارند.

[4] Imbalenced

منابع و مراجع

[1] X. Shu, L. Cheng, and S. J. Stolfo, “Anomaly Detection as a Service”.

[2] D. M. Hawkins, *Identification of Outliers*. Netherlands: Springer, 1980. doi: 10.1007/978-94-015-3994-4.

[3] C. Jiang, J. Song, G. Liu, L. Zheng, and W. Luan, “Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism,” *IEEE Internet of Things Journal*, vol. 5, no. 5, pp. 3637–3647, 2018, doi: 10.1109/JIOT.2018.2816007.

[4] X. Dai and M. Bikdash, “Distance-based outliers method for detecting disease outbreaks using social media,” *Conference Proceedings - IEEE SOUTHEASTCON*, vol. 2016-July, 2016, doi: 10.1109/SECON.2016.7506752.

[5] S. A. Haque, M. Rahman, and S. M. Aziz, “Sensor anomaly detection in wireless sensor networks for healthcare,” *Sensors (Switzerland)*, vol. 15, no. 4, pp. 8764–8786, 2015, doi: 10.3390/s150408764.

[6] H. S. Wu, “A survey of research on anomaly detection for time series,” *2016 13th International Computer Conference on Wavelet Active Media Technology and Information Processing, ICCWAMTIP 2017*, no. 1, pp. 426–431, 2017, doi: 10.1109/ICCWAMTIP.2016.8079887.

[7] K. Choi, J. Yi, C. Park, and S. Yoon, “Deep Learning for Anomaly Detection in Time-Series Data: Review, Analysis, and Guidelines,” *IEEE Access*, vol. 9, pp. 120043–120065, 2021, doi: 10.1109/ACCESS.2021.3107975.

[8] I. Ruts and P. J. Rousseeuw, “Computing depth contours of bivariate point clouds,” *Computational Statistics and Data Analysis*, vol. 23, no. 1, pp. 153–168, 1996, doi: 10.1016/S0167-9473(96)00027-8.

[9] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos, “LOCI: Fast outlier detection using the local correlation integral,” *Proceedings - International Conference on Data Engineering*, pp. 315–326, 2003, doi: 10.1109/ICDE.2003.1260802.

[10] F. Sönmez, M. Zontul, O. Kaynar, and H. Tutar, “Anomaly Detection Using Data Mining Methods in IT Systems: A Decision Support Application,”

Sakarya University Journal of Science, vol. 22, no. 4, pp. 1–1, 2018, doi: 10.16984/saufenbilder.365931.

[11] G. Muruti, F. A. Rahim, and Z. A. Bin Ibrahim, "A survey on anomalies detection techniques and measurement methods," *2018 IEEE Conference on Application, Information and Network Security, AINS 2018*, no. 1, pp. 81–86, 2019, doi: 10.1109/IISA.2018.8631436.

[12] M. Ahmed, A. Naser Mahmood, and J. Hu, "A survey of network anomaly detection techniques," *Journal of Network and Computer Applications*, vol. 60, pp. 19–31, 2016, doi: 10.1016/j.jnca.2015.11.016.

[13] T. Schlegl, P. Seeb, S. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery," *International Confrence on Information Processing in Medical Imaging*, vol. 2, pp. 146–157, 2017, doi: 10.1007/978-3-319-59050-9.

[14] V. Dumoulin *et al.*, "Adversarially learned inference," *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, pp. 1–18, 2017.

[15] H. Issa and M. A. Vasarhelyi, "Application of Anomaly Detection Techniques to Identify Fraudulent Refunds," *SSRN Electronic Journal*, 2012, doi: 10.2139/ssrn.1910468.

[16] R. Kaur and S. Singh, "A survey of data mining and social network analysis based anomaly detection techniques," *Egyptian Informatics Journal*, vol. 17, no. 2, pp. 199–216, 2016, doi: 10.1016/j.eij.2015.11.004.

[17] N. Görnitz, M. Kloft, K. Rieck, and U. Brefeld, "Toward supervised anomaly detection," *Journal of Artificial Intelligence Research*, vol. 46, pp. 235–262, 2013, doi: 10.1613/jair.3623.

[18] R. N. Reza Hassanzadeh, "A SemiSupervised GraphBased Algorithm for Detecting Outliers in OnlineSocialNetworks," pp. 577–582, 2013.

[19] D. Kwon, H. Kim, J. Kim, S. C. Suh, I. Kim, and K. J. Kim, "A survey of deep learning-based network anomaly detection," *Cluster Computing*, vol. 22, pp. 949–961, 2019, doi: 10.1007/s10586-017-1117-8.

[20] Z. Zhao, C. K. Mohan, and K. G. Mehrotra, “Adaptive sampling and learning for unsupervised outlier detection,” *Proceedings of the 29th International Florida Artificial Intelligence Research Society Conference, FLAIRS 2016*, pp. 460–465, 2016.

[21] D. Digitalcommons@uri and Y. Chae, “Representing Statistical Network-Based Anomaly Detection by Representing Statistical Network-Based Anomaly Detection by Using Trust Using Trust,” 2017.

[22] M. A. Rassam, A. Zainal, and M. A. Maarof, “Advancements of data anomaly detection research in Wireless Sensor Networks: A survey and open issues,” *Sensors (Switzerland)*, vol. 13, no. 8, pp. 10087–10122, 2013, doi: 10.3390/s130810087.

[23] G. Thatte, U. Mitra, and J. Heidemann, “Parametric methods for anomaly detection in aggregate traffic,” *IEEE/ACM Transactions on Networking*, vol. 19, no. 2, pp. 512–525, 2011, doi: 10.1109/TNET.2010.2070845.

[24] J. Wu, W. Zeng, and F. Yan, “Hierarchical Temporal Memory method for time-series-based anomaly detection,” *Neurocomputing*, vol. 273, pp. 535–546, 2018, doi: 10.1016/j.neucom.2017.08.026.

[25] S. Zou, Y. Liang, H. V. Poor, and X. Shi, “Unsupervised nonparametric anomaly detection: A kernel method,” *2014 52nd Annual Allerton Conference on Communication, Control, and Computing, Allerton 2014*, pp. 836–841, 2014, doi: 10.1109/ALLERTON.2014.7028541.

[26] D. B. Araya, K. Grolinger, H. F. ElYamany, M. A. M. Capretz, and G. Bitsuamlak, “An ensemble learning framework for anomaly detection in building energy consumption,” *Energy and Buildings*, vol. 144, pp. 191–206, 2017, doi: 10.1016/j.enbuild.2017.02.058.

[27] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, “Network anomaly detection: Methods, systems and tools,” *IEEE Communications Surveys and Tutorials*, vol. 16, no. 1, pp. 303–336, 2014, doi: 10.1109/SURV.2013.052213.00046.

[28] M. S. Mohd Pozi, M. N. Sulaiman, N. Mustapha, and T. Perumal, “Improving Anomalous Rare Attack Detection Rate for Intrusion Detection System Using Support Vector Machine and Genetic Programming,” *Neural Processing Letters*, vol. 44, no. 2, pp. 279–290, 2016, doi: 10.1007/s11063-015-9457-y.

[29] R. Ul Islam, M. S. Hossain, and K. Andersson, “A novel anomaly detection algorithm for sensor data under uncertainty,” *Soft Computing*, vol. 22, no. 5, pp. 1623–1639, 2018, doi: 10.1007/s00500-016-2425-2.

[30] A. H. Moghaddam, M. H. Moghaddam, and M. Esfandyari, “Stock market index prediction using artificial neural network,” *Journal of Economics, Finance and Administrative Science*, vol. 21, no. 41, pp. 89–93, 2016, doi: 10.1016/j.jefas.2016.07.002.

[31] I. Goodfellow *et al.*, “Generative Adversarial Nets,” in *Advances in Neural Information Processing Systems*, Oct. 2014, pp. 2672–2680. doi: 10.1109/ICCVW.2019.00369.

[32] H. Zenati, M. Romain, C. S. Foo, B. Lecouat, and V. Chandrasekhar, “Adversarially Learned Anomaly Detection,” *Proceedings - IEEE International Conference on Data Mining, ICDM*, vol. 2018-Novem, pp. 727–736, 2018, doi: 10.1109/ICDM.2018.00088.

[33] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Piatt, “Support vector method for novelty detection,” *Advances in Neural Information Processing Systems*, pp. 582–588, 2000.

[34] R. A. Yeh, C. Chen, T. Yian Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, “Semantic image inpainting with deep generative models,” *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 6882–6890, 2017, doi: 10.1109/CVPR.2017.728.

[35] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, “f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks,” *Medical Image Analysis*, vol. 54, pp. 30–44, 2019, doi: 10.1016/j.media.2019.01.010.

[36] H. Zenati, C. S. Foo, B. Lecouat, G. Manek, and V. R. Chandrasekhar, “Efficient GAN-Based Anomaly Detection,” 2018, [Online]. Available: http://arxiv.org/abs/1802.06222

[37] C. Li *et al.*, “ALICE : Towards Understanding Adversarial Learning for Joint Distribution Matching arXiv : 1709 . 01215v2 [stat . ML] 5 Nov 2017,” no. Nips, pp. 1–22, 2017.

[38] Z. Yang, I. S. Bozchalooi, and E. Darve, “Regularized Cycle Consistent Generative Adversarial Network for Anomaly Detection”.

[39] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, “Spectral normalization for generative adversarial networks,” 2018.

[40] T. Salimans, I. Goodfellow, V. Cheung, A. Radford, and X. Chen, “Improved Techniques for Training GANs,” pp. 1–10.

[41] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, “Generative Adversarial Networks: An Overview,” *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018, doi: 10.1109/MSP.2017.2765202.

[42] S. Zhai, Y. Cheng, W. Lu, and Z. Zhang, “Deep structured energy based models for anomaly detection,” *33rd International Conference on Machine Learning, ICML 2016*, vol. 3, pp. 1742–1751, 2016.

[43] F. Tony Liu, K. Ming Ting, and Z.-H. Zhou, “Isolation Forest ICDM08,” *Icdm*, 2008, [Online]. Available: https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf%0Ahttps://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf?q=isolation-forest

[44] A. Makhzani and B. Frey, “Winner-take-all autoencoders,” *Advances in Neural Information Processing Systems*, vol. 2015-Janua, pp. 2791–2799, 2015.

[45] L. Ruff *et al.*, “Deep one-class classification,” *35th International Conference on Machine Learning, ICML 2018*, vol. 10, pp. 6981–6996, 2018.

[46] J. Kim, K. Jeong, H. Choi, and K. Seo, “GAN-Based Anomaly Detection In Imbalance Problems*.” Lecture Notes in Computer Science*, 12540 LNCS, 128–145. https://doi.org/10.1007/978-3-030-65414-6_118, 2020.

فهرست واژگان انگلیسی به فارسی

A

Ablation studies.....مطالعه فرسایشی

Activation function.....تابع فعال‌ساز

Area Under Curve Receiver Operating Characteristics.....مساحت زیر نمودار مشخصه عملکرد

Autoencoder.....خودکدگذار

B

Backpropagation.....پس انتشار

Batch.....دسته

Bias.....متمایل

Bidirectional.....دوطرفه

C

Categorical.....اسمی

Clustering.....خوشه‌بندی

Complete Cycle Consistency.....چرخه پایداری کامل

Conditional entropy.....آنتروپی شرطی

Continuous.....پیوسته

Cycle consistency.....چرخه پایداری

D

Decision tree.....درخت تصمیم

Deviation detection.....تشخیص انحراف

Discrimination loss.....خطای تمایزگر

Discriminator models.....مدل‌های تمایزگر

Dropout.....حذف تصادفی

E

Encoder.....کدگذار

Exception mining.....استخراج استثنا

Experimental Setup.....تنظیمات آزمون

F

False negative.....منفی کاذب

False positive rate.....نرخ مثبت کاذب

Feature matching.....تطبیق ویژگی

Feature matching loss.....خطای تطبیق ویژگی

Feed forward.....انتشار رو به جلو

G

Generative Adversarially Networks.....شبکه مولد تقابلی

Gradient descent.....گرادیان نزولی

H

Historical averaging.....میان‌گیری تاریخی

Hyper Parameter.....فراپارامتر

I

Invert.....معکوس کردن

Iterative.....مبتنی بر تکرار

J

Joint distribution.....توزیع توام

K

kernel.....هسته

L

Latent space.....فضای نهفته

Log-Liklihood.....لگاریتم درست‌نمایی

M

Maximum likelihood estimation.....بیشینه تخمین درست‌نمایی

Mean squared error.....میانگین مجموع مربعات خطا

Minibatch.....کوچک دسته‌ای

Minibatch discrimination.....تمایز کوچک دسته‌ای

Minimax.....بیشینه-کمینه

Mutual information.....اطلاعات متقابل

N

Novelty detection.....شناسایی نوآوری

O

Outlier detection.....تشخیص داده پرت

Overfitting.....بیش‌برازش

P

Piecewise linear units.....واحدهای خطی تکه‌ای

probability Posterior.....احتمال پسین

Precision.....صحت

Principal component analysis.....تحلیل مولفه اصلی

R

Recall.....بازیابی

Reconstruction.....بازسازی

Reinforcement learninig.....یادگیری تقویتی

Remote sensing.....سنجش از راه دور

Residual.....باقی‌مانده

Residual Loss.....خطای باقی‌مانده

Robustness.....مقاومت

S

Scaler.....اسکالر

Sigmoid cross entropy.....آنتروپی متقاطع سیگموئید

Smooth.....هموار

Support vector machine.....ماشین بردار پشتیبان

T

Tabular.....جدولی

Threshold.....حد آستانه

True negative.....منفی صحیح

True Negative Rate.....نرخ منفی صحیح

True positive.....مثبت صحیح

Tuple.....دوتایی

V

Variational autoencoder.....خودکدگذار متغیر

Virtual batch normalization.....نرمال کردن مجازی دسته

Abstract

Anomaly detection is a significant and hence well studied problem in field of data analysis which is used in a wide range of applications such as fraud detection, medical application and cyber security systems. Despite the existence of statistical and machine learning-based methods, designing effective models for anomaly detection in complex high-dimensional data space remains a major challenge. As generative adversarial networks are able to handle this challenge and model the complex high-dimensional distribution of real-world data. as a result it can operate promisingly in field of anomaly detection. In this work we propose CALAD[1], RALAD[2] and RCALAD[3] models to detect anomalies. Our reconstruction based method reconstruct the input data through generative network and compute reconstruction error to find anomalous example. In the CALAD model defining new variable $\hat{\boldsymbol{z}}_{\hat{\boldsymbol{x}}}$ and using an innovative discriminator $\boldsymbol{D}_{\boldsymbol{xxzz}}$, complete cycle consistency between input space and hidden space is established. Poor reconstruction for anomalous data is a prerequisites in reconstruction based models. RALAD aims to bias the model towards normal data distribution. This bias leads to poor reconstruction of anomalous data and as a result the distance between anomalous input data and its reconstruction will increase.With combining these two ideas, comprehensive RCALAD model is proposed. In addition, two new anomaly score are proposed which provide high resolution power in contrast to other anomaly scores. Finally, experimental results demonstrate the effectiveness of our approach by showing the results if outperforming the current state of the art approaches in terms of the average area under the ROC[4] and F1-score.

Key Words: anomaly detection, machine learning, generative adversarial networks, reconstruction error, anomaly score

[1] Complete Adversarially Learned Anomaly Detection

[2] Regularized Adversarially Learned Anomaly Detection

[3] Regularized Complete Adversarially Learned Anomaly Detection

[4] Reciver Operating Characteristic

Amirkabir University of Technology
(Tehran Polytechnic)

Department of Computer Engineering and Information Technology

Master Thesis

Anomaly Detection with Generative Adversarial Network

By
Zahra Dehghanian

supervisor

Dr. Mohammad Rahmati
Dr. Maryam Amirmazlaghani

May 2022