دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مهندسی کامپییوتر

سمینار کارشناسی ارشد

در رشته‌ی مهندسی کامپیوتر گرایش هوش مصنوعی و رباتیک

عنوان سمینار

تشخیص ناهنجاری با استفاده از شبکه‌های مولد متخاصم

Anomaly Detection with Generative Adversarial Network

دانشجو

زهرا دهقانیان

استاد درس سمینار

دکتر رضاصفابخش

استاد مشاور

دکتر محمد رحمتی

دکتر مریم امیرمزلقانی

تابستان سال ۱۳۹۹

چکیده

در سال های اخیر که حجم اطلاعات با سرعت سرسام آوری در حال رشد است، توجه ویژه ای به پردازش و تحلیل این داده‌ها صورت گرفته است. یکی از مهم‌ترین فعالیت‌های تحلیل داده، تشخیص ناهنجاری می‌باشد. روش‌های تشخیص ناهنجاری در طیف وسیعی از کاربردها، هم‌چون تعاملات بانکی، کاربردهای پزشکی و سیستم-های امنیتی بکار گرفته می‌شوند. برای تشخیص ناهنجاری روش‌های مختلفی شامل مبتنی بر آمار و مبتنی بر یادگیری ماشین پیشنهاد شده‌است. از آن‌جا که نتایج خوبی از شبکه‌های عصبی و به طور خاص شبکه‌های مولد متخاصمی در حوزه‌های مختلف حاصل شده، رویکردهای اخیر به استفاده از این روش در این حوزه می‌پردازد. در این گزارش ابتدا به تعریف تشخیص ناهنجاری، طبقه‌بندی و روش‌های تشخیص ناهنجاری و جایگاه شبکه‌های مولد متخاصم می‌پردازد، سپس به تعریف شبکه مولد متخاصم، معرفی انواع و سیر تکاملی این شبکه‌ها در کاربرد تشخیص ناهنجاری پرداخته شده‌است.

کلمات کلیدی: شبکه عصبی، شبکه مولد متخاصم، تشخیص ناهنجاری، یادگیری ماشین

فهرست مطالب

صفحه

فهرست الگوریتم‌ها صفحه

فهرست اشکال

صفحه

فصل اول : مقدمه

تشخیص ناهنجاری یک کار مهم برای تجزیه و تحلیل داده‌هاست که داده‌های غیرعادی یا غیرطبیعی را از یک مجموعه داده تشخیص می‌دهد. این یک بخش حائز اهمیت از تحقیقات در زمینه داده‌کاوی است، زیرا شامل کشف الگوهای جذاب و نادر در داده‌ها است. این امر به طور گسترده در آمار و یادگیری ماشین مورد مطالعه قرار گرفته است و مترادف‌هایی چون تشخیص داده‌های پرت، شناسایی نوآوری، تشخیص انحراف و استخراج استثناء دارد. اگرچه محققان ناهنجاری را به روش های مختلف و بر اساس دامنه کاربرد آن تعریف می‌کنند، اما تعریف عام و مورد قبول آن، تعریف هاوکینز است [۱]:"یک ناهنجاری مشاهده‌ای است که به میزانی از سایر مشاهدات منحرف می‌شود که ظن‌هایی را برای این که توسط مکانیسم متفاوتی تولید شده، ایجاد می‌کند."

ناهنجاری‌ها جزو پارامترهای مهم مجموعه داده‌ها در نظر گرفته می‌شوند و می‌توانند اقدامات حیاتی را در طیف وسیعی از دامنه‌های کاربردی انجام دهند. به عنوان مثال، الگوی غیر معمول ترافیک در یک شبکه می‌تواند به معنای هک شدن رایانه و انتقال داده‌ها به مقصدهای غیرمجاز باشد. رفتار غیر عادی در معاملات کارت اعتباری می‌تواند فعالیت های کلاهبرداری را نشان دهد، یک ناهنجاری در تصویر MRI ممکن است وجود تومور بدخیم را نشان دهد. تشخیص ناهنجاری به طور گسترده‌ای در حوزه‌های بی‌شماری از کاربردها مانند: پزشکی، بهداشت عمومی، تشخیص کلاهبرداری، تشخیص نفوذ، پردازش تصویر، آسیب‌های صنعتی، شبکه‌های حسگر، رفتار روبات‌ها و داده‌های نجومی بکار گرفته شده است.

برای تشخیص ناهنجاری‌ها، تاکنون روش‌های گوناگونی مورد استفاده قرار گرفته است. روش‌های موجود بر مبنای تعریف قانون با این که دقت نسبتا قابل قبولی روی ناهنجاری‌هایی که تاکنون شناسایی شده دارند، اما به زمان زیادی در مرحله اجرا نیاز دارند و هم‌چنین در مواجهه با ناهنجاری‌های جدید و حمله‌های ناشناخته عملکرد ضعیفی دارند و به همین دلیل، روش‌های مبتنی بر یادگیری ماشین بر روش‌های دیگر برتری می‌یابند. در بین این الگوریتم‌های یادگیری ماشین، با توجه به نادر بودن ناهنجاری‌ها ، بیش‌تر الگوریتم‌ها نمی‌توانند دقت لازم را کسب کنند و روش‌هایی که نیاز به داده آموزش کم‌تری دارند، کارآمد هستند [۲].

در سال‌های اخیر، کارهای‌جدید مبتنی بر شبکه‌های عصبی عمیق انجام شده است. شبکه‌های عصبی دارای سابقه طولانی در استفاده برای تشخیص ناهنجاری هستند. رویکردهای مبتنی بر خودرمزگذار و خود رمزگذارهای خودکار از این دسته هستند. در این دسته ابتدا یک مدل برای بازسازی داده‌های عادی آموزش داده می‌شود و سپس ناهنجاری‌ها را به عنوان نمونه‌هایی با خطاهای بازسازی بالا شناسایی می‌کنند. مدل‌های مبتنی بر انرژی و مدل های مخلوط گوسی[1] با رمزگذاری عمیق خودکار نیز به طور خاص به منظور تشخیص ناهنجاری مورد کاوش قرار

[1] Gaussian Mixture Model

گرفته‌اند. چنین روش‌هایی توزیع داده را با استفاده از رمزگذار خودکار یا مدل های مشابه مدل می‌کنند و ناهنجاری را بر اساس یک معیار ناهنجاری آماری بر طبق انرژی یا مخلوط‌های گاوسی استخراج می‌کنند [۳].

در سال ۲۰۱۷ از شبکه‌های مولد متخاصم برای تشخیص ناهنجاری در زمینه تصویربرداری پزشکی بر روی تصاویر شبکیه استفاده شد و به موفقیت قابل توجهی در مقایسه با سایر روش‌ها دست یافت [۴]. شبکه مولد متخاصم در سال های اخیر به دقت بسیار بالایی دست یافته و در زمینه‌های مختلف به‌کار گرفته شده‌اند و به موفقیت‌های چشم‌گیری در عرصه پردازش تصویر و استخراج ویژگی از آن‌ها دست یافته‌اند . مدل‌های مولد عمیق به عنوان یک چارچوب قدرتمند برای مدل سازی مجموعه داده‌های پیچیده چند بعدی عمل می‌کند. این مدل‌ها از نمونه‌برداری سریع بهره می‌جویند، اما اغلب به دلیل پیچیدگی‌های استنتاج، دچار چالش‌های جدی هستند [۵].

در سال‌های اخیر، تلاش‌شده تا با بکارگیری شبکه‌های خودرمزنگار در کنار شبکه‌های مولد متخاصم از این پیچیدگی‌ها کاسته و بر چالش‌های موجود غلبه کنند.

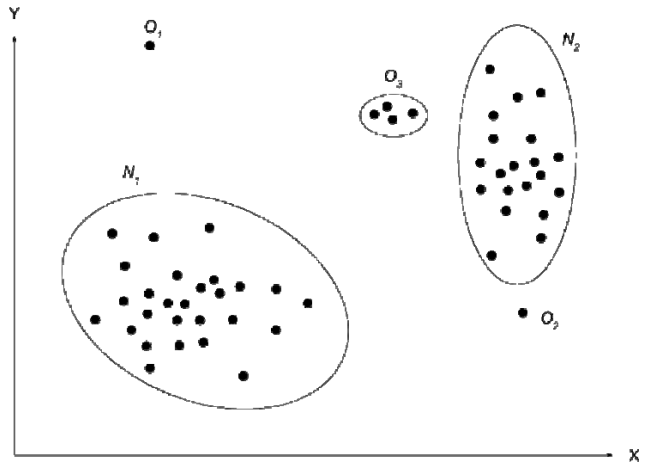
۱-۲- سازماندهی گزارش

ابتدا در فصل دوم مفاهیم پایه و روش‌های تشخیص ناهنجاری ارائه می‌گردد. در فصل سوم به معرفی شبکه مولد متخاصم خواهیم پرداخت. در فصل چهارم چندین مدل معروف شبکه‌های مولد متخاصم که در کاربرد تشخیص ناهنجاری مورد استفاده قرار گرفته است، بررسی خواهد شد. در نهایت در فصل پنجم، نتیجه‌گیری، جمع‌بندی و هم‌چنین روال آتی توضیح داده می‌شود.

فصل دوم : روش‌های تشخیص ناهنجاری

۲-۱- مقدمه

امروزه ناهنجاری‌های[1] موجود در مجموعه داده‌ها می‌تواند به دلیل خطای ناخواسته، اشکالات عمدی یا حملات سایبری اتفاق بیفتد که برای جلوگیری از بروز مشکلات اساسی، به تکنیک‌های تشخیص ناهنجاری نیاز می‌باشد. تشخیص ناهنجاری به جستجوی الگوهای غیرعادی در داده‌ها که با رفتار معمول داده ها مطابقت ندارد، اطلاق می‌شود. شکل ۲-۱ نمونه‌ای از ناهنجاری‌های دو بعدی را نشان می‌دهد؛ مناطق N_1 و N_2 داده‌های عادی را نشان می‌دهد زیرا بیشتر مشاهدات در این مناطق است، درحالی که نقاط O_2 ، O_1 و منطقه O_3 که از مناطق عادی دور هستند، ناهنجاری شناخته می‌شوند.



شکل ۲-۱: ناهنجاری دو بعدی

اصطلاح ناهنجاری و داده‌پرت[2] دو عبارتی است که غالباً در زمینه تشخیص ناهنجاری به طور متناوب به جای یکدیگر به کار می‌رود. روش‌های تشخیص ناهنجاری را می‌توان در بسیاری از حوزه‌ها استفاده کرد. به عنوان نمونه، در یک شبکه با الگوی ترافیکی غیر عادی می‌توان وجود یک وسیله مخرب یا یک گره آسیب پذیر در برابر حملات سایبری را نشان داد. ویا از ناهنجاری‌های موجود در تراکنش‌های کارت اعتباری می‌توان برای اثبات کلاهبرداری در کارت اعتباری استفاده کرد.

تحقیقات در مورد تشخیص ناهنجاری به طور گسترده از قرن نوزدهم مورد مطالعه قرار گرفته است. با گذشت زمان، تکنیک‌های مختلفی برای تشخیص ناهنجاری ایجاد شده و تعداد قابل توجهی از این تکنیک‌ها بطور خاص برای برخی از برنامه‌های کاربردی ایجاد شده است [۶].

[1] anomalies
[2] outlier

۲-۲- کاربردهای تشخیص ناهنجاری

الگوسازی ناهنجاری اساساً به دو چیز بستگی دارد. اول ساخت نمایه[1]های رفتاری برای فعالیت های عادی و دوم، تکنیک‌های مختلف برای شناسایی هر نوع انحراف از این نمایه‌ها می‌باشد. تشخیص ناهنجاری دارای کاربردهای گسترده‌ای در تجارت همانند سیستم تشخیص نفوذ، سیستم نظارت بر سلامت، سیستم تشخیص جعل در کارت اعتباری و یا سیستم تشخیص خطا در سیستم های مهم اطلاعاتی، می‌باشد. در ادامه به بررسی بیشتر این روش‌ها در بعضی از این کاربردها می‌پردازیم [۶].

۲-۲-۱- تشخیص نفوذ

نفوذ به فعالیت‌های مخرب مانند نقض سیستم مبتنی بر رایانه اشاره می‌کند. از نظر امنیت اطلاعات، این نقض‌ها باارزش هستند. این نقض از رفتار استاندارد سیستم، سبب می‌شود تا تکنیک‌های تشخیص ناهنجاری برای حوزه سیستم‌های تشخیص نفوذ[2] از نیازهای اساسی تلقی شود. سیستم های تشخیص نفوذ را می توان به دو دسته طبقه‌بندی کرد که سیستم‌های شناسایی نفوذ مبتنی بر شبکه و سیستم‌های شناسایی نفوذ مبتنی بر میزبان هستند. سیستم‌های شناسایی نفوذ مبتنی بر شبکه به طور معمول، با حمله‌ها که به عنوان ناهنجاری در داده‌های شبکه رخ می‌دهد، برخورد می‌کنند. برای این منظور، داده‌های شبکه پی‌درپی مدل می‌شوند تا به محض وقوع الگوهای ناهنجار، آن را تشخیص دهند. سیستم‌های شناسایی نفوذ مبتنی بر میزبان، تکنیک‌های تشخیص ناهنجاری دارند که بتواند داده ها را دائما کنترل کند، زیرا روش تشخیص ناهنجاری نقطه برای سیستم تشخیص نفوذ مبتنی بر میزبان مناسب نیست.

۲-۲-۲- تشخیص جعل

اصطلاح تشخیص جعل به شناسایی فعالیت‌های غیرقانونی که در کاربردهای تجاری مانند شرکت‌های کارت اعتباری، بانک‌ها و شرکت‌های بیمه انجام می شود، اطلاق می گردد. کاربران مخرب می‌توانند به عنوان یک کاربر مشروع از خدمات ارائه شده توسط سازمان‌های تجاری بهره برداری کنند. کلاهبرداری زمانی اتفاق می‌افتد که این کاربران از شیوه‌های غیرقانونی از منابع ارائه شده، استفاده کنند. از این رو، سازمان‌های تجاری به دنبال شناسایی چنین کلاهبرداری‌هایی هستند تا ضررهای مالی را کاهش دهند.

---

[1] Profile
[2] intrusion detection systems

در حوزه شناسایی جعل کارت اعتباری، از تکنیک‌های تشخیص ناهنجاری برای شناسایی معاملات جعلی استفاده می شود. داده‌های کارت اعتباری معمولاً از سوابق کاربران مانند: میزان هزینه، شناسه کاربر و مدت زمان معامله تشکیل شده است. کلاهبرداری‌ها معمولاً به عنوان ناهنجاری‌های نقطه‌ای در سوابق معاملاتی منعکس می‌شوند؛ به این معنا که در یک بازه زمانی کوتاه، تعداد زیادی خرید یا پرداخت صورت می‌گیرد که جزو روال رفتار مالی کاربر نیست. تکنیک‌های مبتنی بر خوشه‌بندی و نمایه‌سازی معمولاً توسط شرکت‌های کارت اعتباری برای متمایز کردن داده‌ها بر اساس کاربر کارت اعتباری استفاده می‌شوند، زیرا شرکت‌های کارت اعتباری سوابق داده‌های دارای برچسب را به صورت کامل دارند.

۲-۲-۳- تشخیص ناهنجاری های پزشکی

در حوزه پزشکی، تشخیص ناهنجاری شامل سوابق بیمار نیز می‌شود. ناهنجاری در اطلاعات پرونده بیمار به دلیل خطای دستگاه، خطای هنگام ذخیره و نگهداری و یا وضعیت غیرمعمول بیمار، ممکن است اتفاق بیفتد. علاوه بر این یکی دیگر از کاربردها، تشخیص ناهنجاری در شناسایی شیوع بیماری در یک منطقه خاص می باشد. تشخیص ناهنجاری در حوزه پزشکی بسیار حیاتی است و نیاز به دقت بالایی دارد.

در این حوزه اکثر تکنیک‌ها بر شناسایی ناهنجاری های نقطه‌ای متمرکز شده اند. به‌طور کلی داده‌ها شامل وزن بیمار، سن بیمار و گروه خونی و همچنین داده‌های سری زمانی مانند الکتروکاردیوگرام[1] می‌باشد .برای تشخیص ناهنجاری‌ها در این نوع داده‌ها از تکنیک‌های تشخیص ناهنجاری جمعی استفاده شده است. با این حال، بخش چالش برانگیز هزینه‌ای است که در طبقه‌بندی یک ناهنجاری باید پرداخت شود.

۲-۳- طبقه‌بندی روش‌های تشخیص ناهنجاری

اکثر تکنیک‌های تشخیص ناهنجاری نیاز دارند تا از برچسب‌ها برای تشخیص برای این که نمونه داده طبیعی یا ناهنجاری است، استفاده کنند. تهیه و دستیابی به داده‌های دارای برچسب دقیق، شامل طیف گسترده‌ای از رفتارهای بسیار هزینه‌بر و دشوار است. تکنیک‌های تشخیص ناهنجاری بر اساس در دسترس بودن برچسب‌ها، می توانند به سه گروه طبقه بندی شوند.

۲-۳-۱- تشخیص ناهنجاری نظارت‌شده

در این روش‌ها هر دو الگوی رفتاری غیر طبیعی و عادی با استفاده از تشخیص ناهنجاری نظارت شده، مدل می شوند. در این طبقه، برای شناسایی ناهنجاری به داده های از پیش نشانه گذاری با برچسب های غیر طبیعی و

[1] Electrocardiograms

عادی نیاز است. برای شناسایی داده‌های غیر طبیعی یا عادی در مجموعه داده‌ها، از چندین مدل آموزش استفاده می‌شود. تکنیک‌های نظارت شده با دنبال کردن این رویکردها کار می‌کنند؛ مدل در حال آموزش در پایگاه‌داده‌ای که به‌عنوان داده‌های عادی طبقه‌بندی شده، برای شناسایی داده‌های غیر عادی مقایسه می‌شود و در مقابل برخی از داده‌های غیر عادی با مدل آموزش برای یافتن داده‌های غیر طبیعی مقایسه می‌شود.

۲-۳-۲- تشیص ناهنجاری نیمه نظارتی

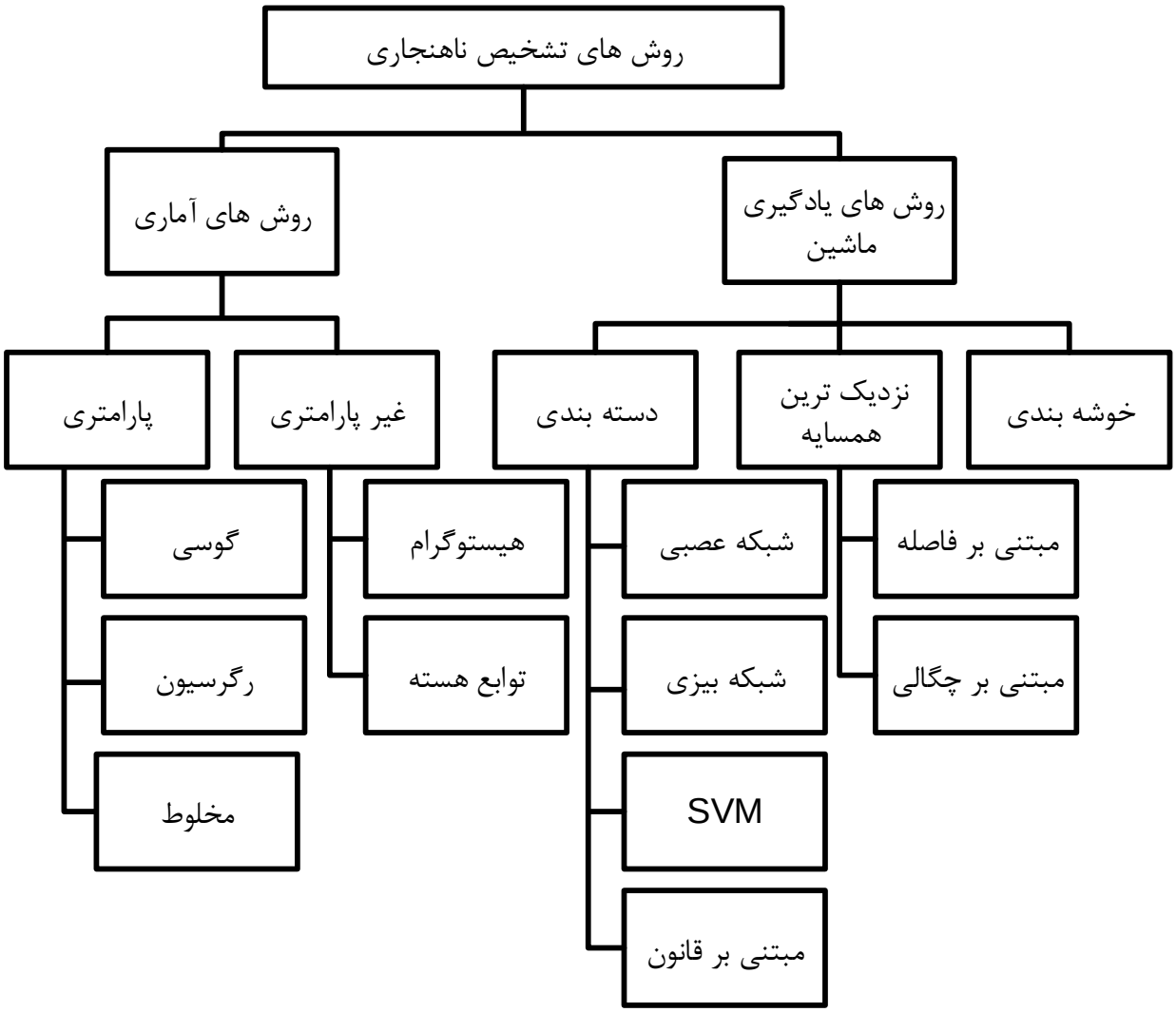
این دسته از روش‌ها تنها به داده‌های آموزش از کلاس عادی نیاز دارند و برچسب داده‌های ناهنجاری لازم نیست. این دسته از روش‌ها بیش از روش‌های نظارت شده استفاده می‌شود.

۲-۳-۳- تشخیص ناهنجاری بدون نظارت

در روش‌های بدون نظارت، مدل آموزش کلاس داده‌های ناهنجاری را به صورت خودکار از مجموعه داده‌ها تشخیص می دهد. این عملیات با استفاده از روش‌های خوشه‌بندی کار می‌کند. این روش خوشه‌ای از گره‌ها را پیدا می‌کند که در آن رفتار اعضای گروه مشابه است. با این حال، این فرض در بسیاری از تشخیص‌ها دچار مشکل می‌شود؛ زیرا بسیاری از ناهنجاری‌ها باعث ایجاد خوشه‌هایی با الگوی مشابه داده عادی می‌شوند. به همین دلیل تکنیک‌های بدون نظارت در تولید نتایج دقیق کارآمد نیستند و اغلب از نرخ مثبت کاذب بالا رنج می‌برند.

۲-۴- معرفی روش‌های تشخیص ناهنجاری

در این بخش، چندین روش که برای تشخیص ناهنجاری پیشنهاد شده و مورد استفاده قرار می‌گیرد را مرور می‌کنیم. این تقسیم‌بندی شامل روش‌های آماری و روش‌های یادگیری ماشین می‌باشد. این دسته‌بندی به طور دقیق تر در شکل ۲-۲ نشان داده شده است [۶].



شکل ۲-۲: روش‌های تشخیص ناهنجاری

۲-۴-۱- روش‌های آماری

در این روش‌ها مدل‌های ناهنجاری آماری شامل عناصر زیر می باشد: اول آمارگان‌های آماری مانند میانگین و انحراف از معیار، دوم توزیع داده‌ها و سوم توابع احتمال برای ساختن نمایه های رفتاری. آزمون‌های آماری برای تشخیص هر نوع انحراف از رفتار عادی استفاده می‌شود. در توسعه مدل‌های تشخیص ناهنجاری آماری، از تکنیک

های غیر پارامتری و پارامتری استفاده می‌شود. تفاوت این تکنیک‌ها در استفاده از اطلاعات توزیع زمینه‌ای می‌باشد؛ تکنیک‌های غیر پارامتری اطلاعات توزیع زمینه‌ای از داده ها را ندارند در حالی که تکنیک‌های پارامتری این اطلاعات را دارند.

- **تکنیک‌های پارامتری**

در این تکنیک‌ها، فرض بر این است که داده‌های عادی با استفاده از پارامترها و امتیازها از نمونه داده‌ها تولید می‌شوند. تکنیک‌های پارامتری را می‌توان به سه دسته کلی تقسیم کرد: مدل رگرسیونی، مدل گاوسی و مدل مخلوط. در مدل رگرسیونی، داده‌ها بر یک مدل رگرسیونی منطبق می‌شوند و باقی‌مانده[1] هر داده که بر مدل منطبق نیست، اندازه‌گیری می‌شود و این معیار به‌عنوان امتیاز ناهنجاری نمونه، به حساب می‌آید.

در مدل گاوسی، فرض بر این است که داده‌ها به توزیع گاوسی تعلق دارند. پارامترهای مدل با استفاده از تخمین حداکثر احتمال نمونه‌های داده، تعیین شده است. در این مدل‌ها از آزمون‌هایی مانند آزمون chi-square می‌توان برای تصمیم‌گیری در مورد اینکه یک نمونه داده ناهنجاری است یا خیر، استفاده کرد.

در مدل‌های مخلوط، از مدل ترکیبی مدل‌های پارامتری استفاده می‌شود. استفاده از این روش‌ها در بعضی از کاربردها، بسیار موفق عمل کرده است. مثلاً با بکارگیری یک مدل مخلوط از تکنیک‌های پارامتری برای تشخیص ناهنجاری‌های شبکه، توانستند در طی چند ثانیه یا کمتر، تمام ناهنجاری های شبکه را در تمام سناریوها تشخیص دهند [۶].

- **تکنیک‌های غیرپارامتری**

در تکنیک‌های غیر پارامتری، از نمونه‌های داده عادی برای تولید یک مدل استفاده می‌شود و انحراف داده نمونه از مدل، امتیاز ناهنجاری نامیده می‌شود. در مدل‌های مبتنی بر هیستوگرام، هیستوگرام بر اساس تقریب از داده‌های عادی تولید می‌شود. برای تصمیم‌گیری این که آیا نمونه مشخص غیر طبیعی است یا نه، نمونه رسم می‌شود و در صورتی که در محدوده‌هایی از هیستوگرام قرار گیرد، نمونه ناهنجاری به حساب می‌آید. روش مدل‌سازی مبتنی بر هسته[2] با هدف استنباط یک تابع تشابه بر اساس داده‌های ارائه شده، امکان ساخت مدل را بر اساس نمونه‌های داده دارد. قابل توجه است که اگر نمونه‌های داده شده، رفتار مجموعه داده را به طور کامل به تصویر نکشد، مدل دقت کافی نخواهد داشت.

[1] Residual
[2] Kernel

۲-۴-۲- روش‌های یادگیری ماشین

مهم‌ترین مزیت روش‌های مبتنی بر یادگیری ماشین، توانایی آن در بهبود ظرفیت برای تمایز بین رفتارهای غیر طبیعی از رفتارهای عادی بر اساس تجربه و بکارگیری آن‌ها برای شناسایی نمونه‌های جدید است. دسته‌بندی تکنیک‌های مبتنی بر یادگیری ماشین شامل طبقه‌بندی، نزدیک‌ترین همسایه و خوشه‌بندی است.

- **دسته‌بندی**

هدف اصلی از تکنیک‌های مبتنی بر دسته‌بندی، اختصاص هر نمونه داده به یکی از کلاس‌های از پیش تعیین شده با توجه به ویژگی‌های آن است. از مزایای استفاده از تکنیک‌ها می‌توان به توانایی آن‌ها در تمایز بین کلاس‌های مختلف از طریق الگوریتم‌های قدرتمند و راندمان بالا در مرحله آزمایش اشاره کرد؛ زیرا هر نمونه از داده‌های آزمون باید با مدل پیش‌پردازش شده مقایسه شود. با این حال، نتیجه این تکنیک‌ها به برچسب‌های دقیق و نماینده‌هایی که برای کلاس‌های مختلف تعیین می‌شود، متکی است. روش‌های متداول کاربردی این تکنیک عبارتند از:

- **شبکه‌های بیزی**

یک مدل گرافیکی است که احتمال اتصالات را در بین نمونه‌های مورد بررسی، ترجمه می‌کند. این کار بر اساس یادگیری نظارت شده است. این عمل با محاسبه احتمال پیشین[1] یک نمونه داده به همراه یک سری پیش‌شرط[2] عمل می‌کند.

- **ماشین بردار پشتیبان (SVM)**

یک الگوریتم یادگیری نظارت شده است که نمونه‌های داده‌های آموزشی را به یک صفحه چند بعدی انتقال می‌دهد. و در فضای چند بعدی، نمونه‌های داده را به دو گروه جداکننده تقسیم می‌کند. SVM فقط با داده‌های عادی آموزش داده می‌شود، از ماشین‌های بردار پشتیبان یه عنوان یک طبقه‌بندی کننده خطی یاد می‌شود، به دلیل اینکه از یک مرزبندی خطی برای جداسازی داده‌ها به حالت عادی و غیر عادی استفاده می‌کند.

- **مبتنی بر قانون**

در این مدل یک سری قواعدی می‌آموزد که عملکرد نمونه‌های داده عادی را یاد می‌گیرد. بنابراین، اگر قوانین نتوانند یک نمونه داده را دنبال کنند، آنگاه نمونه داده غیر عادی تلقی می‌شود. تکنیک درخت تصمیم‌گیری از مجموعه تکنیک‌های مبتنی بر قانون است که برای مطالعه قوانین از طریق نمونه‌های داده‌های آموزش استفاده می‌شود

[1] Posterior
[2] Precondition

• شبکه های عصبی

یک شبکه عصبی سیستم عصبی انسان را تقلید می‌کند و از مجموعه‌ای از فرایندهای بهم پیوسته تشکیل شده است که به طور همزمان با داده‌های محلی عمل می‌کنند. از نمونه‌های داده‌های عادی برای آموزش یک شبکه عصبی استفاده می شود. شبکه های عصبی هم در یادگیری نظارت شده و هم بدون نظارت کار می‌کنند. یکی از انواع این شبکه که در سال‌های اخیر به موفقیت‌های چشم‌گیری دست یافته است شبکه‌های مولد متخاصم می‌باشد. شبکه های مولد متخاصم[1] و چارچوب آموزش متخاصمی با موفقیت برای مدل های پیچیده و با ابعاد بالا از داده‌های دنیای واقعی استفاده شده‌اند. قابلیت این شبکه نشان دهنده ظرفیت آن‌ها برای کاربرد تشخیص ناهنجاری می‌باشد، اگرچه بکارگیری آن‌ها اخیراً مورد کاوش قرار گرفته است [۷].

تشخیص ناهنجاری با استفاده از شبکه مولد متخاصم بدین صورت است که ابتدا به کمک فرایند آموزش تخاصمی مدل‌سازی رفتار عادی صورت می‌گیرد، سپس تشخیص ناهنجاری‌ها به کمک اندازه‌گیری نمره ناهنجاری روی داده‌ها انجام می شود. در تعریف اصلی چارچوب مولد متخاصم، شبکه مولد یک نگاشت از فضای نهفته[2] به فضای داده می‌آموزد و شبکه تمایزگر سعی بر تفکیک بین نمونه‌های واقعی و نمونه‌های تولید شده توسط شبکه مولد را دارد. در معماری‌های ارائه شده برای کاربرد تشخیص ناهنجاری، عموما یک شبکه رمزگذار به چارچوب اصلی اضافه شده تا نگاشت معکوس از فضای داده به فضای نهفته را بیاموزد.

بکارگیری این شبکه ها چندین مزیت دارد. اولین مزیت این است که شبکه مولد متخاصم با کمک آموزش و نمونه‌گیری از مدل‌های مولد، نتایج بسیار خوبی در مقایسه با دیگر روش‌ها در زمان آزمون دارند. علاوه بر این، امکان آموزش داده‌های از دست رفته به کمک الگوریتم‌های یادگیری تقویت شده در مدل شبکه مولد متخاصم وجود دارد. مزیت دیگر این که این شبکه‌ها می‌تواند با همکاری الگوریتم‌های یادگیری ماشین برای تولید خروجی های چندحالته بکار گرفته شود. در ادامه در فصل سوم به بررسی تعاریف پایه شبکه شبکه مولد متخاصم و در فصل چهارم به بررسی جزییات بکارگیری این شبکه ها در کاربرد تشخیص ناهنجاری خواهیم پرداخت.

[1] Generative Adversarial Nets
[2] latent space

- **نزدیک‌ترین همسایه**

این روش از توابع مبتنی بر فاصله یا تراکم برای سنجش فاصله بین نمونه داده تا نزدیک‌ترین همسایه خود استفاده می‌کند. امتیاز ناهنجاری هر داده، همین فاصله است. بسته به برچسب داده‌ها، این روش می‌تواند در یادگیری بدون نظارت و بانظارت بکار گرفته شود.

- **خوشه‌بندی**

تکنیک‌های مبتنی بر خوشه‌بندی از روش‌های یادگیری بدون نظارت است که برای شناسایی مجموعه نمونه‌های شبیه به‌هم بکار برده می‌شود. ناهنجاری ها ممکن است در یک خوشه کوچک مدل شود و یا در هیچ خوشه ای قرار نگیرند. نکته مثبت تکنیک‌های مبتنی بر خوشه بندی این است که سریع‌تر از روش‌های مبتنی بر فاصله است زیرا پیچیدگی محاسباتی کمتری دارد. با این حال نقطه ضعف این روش‌ها، این است که در مجموعه داده‌های کوچک بینش دقیقی ندارد و هم‌چنین برای بخش‌هایی از فضا که نمونه‌ای از آن در داده آموزش وجو ندارد، همواره برچسب داده ناهنجار اختصاص می‌دهد؛ در صورتی که ممکن است با داشتن مجموعه داده بزرگ‌تر این داده برچسب عادی بگیرد.

۲-۵- معیارهای ارزیابی روش‌های تشخیص ناهنجاری

صرف نظر از رویکرد بکار گرفته شده، تشخیص ناهنجاری با مرحله یادگیری که در آن با مجموعه داده‌ی آموزش، مدل آموزش داده می‌شود، آغاز می‌شود. پس از اتمام مرحله یادگیری، مدل آماده طبقه‌بندی نمونه‌های مشاهد نشده و معیارهای موردنظر برای محاسبه عملکرد می‌باشد. به منظور ارزیابی یک تکنیک تشخیص ناهنجاری، معیارهای استانداردی از قبیل نرخ تشخیص، دقت، کارایی و مقیاس‌پذیری اهمیت زیادی دارد.

۲-۵-۱- نرخ تشخیص[1]

نرخ تشخیص یک معیار بسیار مقبول برای سنجش روش تشخیص ناهنجاری است. نتیجه ارزیابی در میزان تشخیص ناهنجاری‌ها می‌تواند در چهار دسته گزارش شود که عبارتند از: مثبت صحیح[2]، منفی صحیح[3]، مثبت کاذب[4] و

[1] Detection Rate
[2] True Positive
[3] True Negative
[4] False Positive

منفی کاذب[1]. معیار معمول مورد استفاده در میزان تشخیص، مثبت صحیح و منفی صحیح است که در آن نسبت مواردی که توسط آشکارساز طبقه‌بندی می‌شود، تعریف می‌شود. مثبت کاذب معیار مهم دیگری است که به مواردی که به اشتباه به عنوان ناهنجار طبقه‌بندی می شوند، اشاره دارد. نرخ منفی کاذب، نسبت موارد غیر طبیعی است که به طور نادرست به عنوان داده عادی طبقه‌بندی شده‌اند.

۲-۵-۲- دقت[2]

این معیار به تمام موارد صحیح طبقه‌بندی شده عادی یا غیر عادی به نسبت کل داده‌ها اشاره دارد معیار صحت[3] نیز یک معیار کاربردی است که نسبت داده‌های واقعا ناهنجار به تمام موارد طبقه بندی شده به عنوان ناهنجاری می‌باشد.

۲-۵-۳- کارایی[4]

از ویژگی‌های عملکرد گیرنده[5] برای محاسبه این معیار استفاده می‌شود. این مشخصه با ترسیم نرخ مثبت صحیح در برابر نرخ مثبت کاذب در مقادیر آستانه متغیر ایجاد می‌شود و در آن آستانه به عنوان نقطه برش در تعیین عادی و یا غیرعادی بودن یک نمونه عمل می‌کند.

۲-۵-۴- مقیاس‌پذیری[6]

معیار مقیاس‌پذیری توانایی تشخیص ناهنجاری را برای مقیاس‌بندی و مقابله موثر با افزایش اندازه مجموعه داده‌ها تعریف می کند. هدف این معیار اطمینان از این است که روش بکارگرفته‌شده می‌تواند تغییرات سریع حجم داده‌های بزرگ را کنترل کند. برای محاسبه این معیار با توجه به جنس مجموعه داده مورد آزمایش روش‌های متفاوتی ارائه شده است [۶].

۲-۶- جمع‌بندی

در این فصل ابتدا به تعریف ناهنجاری پرداخته و با کاربردهای آن در زمینه های مختلف آشنا شدیم. سپس به گروه‌بندی روش‌های تشخیص ناهتجاری در سه گروه با نظارت، نیمه‌نظارتی و بدون نظارت پرداختیم و در ادامه انواع روش‌های تشخیص ناهنجاری را در دو زیر گروه مبتنی برآمار و مبتنی بر یادگیری ماشین بررسی کردیم و با

[1] False Negative
[2] Accuracy
[3] Precision
[4] Performance
[5] Receiver Operating Characteristics
[6] Scalability

جایگاه مدل شبکه مولد متخاصم آشنا شدیم. در انتهای فصل نیز به بررسی چهار معیار موثر اندازه‌گیری تشخیص ناهنجاری پرداختیم.

فصل سوم : شبکه‌های مولد متخاصم

۳-۱- مقدمه

هدف از مدل‌های یادگیری عمیق، کشف مدل‌های سلسله مراتبی قوی است. این مدل‌ها نشان‌دهنده توزیع احتمال انواع داده‌هایی است که در کاربردهای هوش مصنوعی مانند تصاویر طبیعی، شکل موج صوتی حاوی گفتار و نمادها بکار می‌رود. برجسته‌ترین موفقیت یادگیری عمیق در مدل‌های تمایزگر[۱] بوده است؛ این مدل‌ها دارای قدرت دریافت ورودی با ابعاد بالا و تشخیص بادقت زیاد از برچسب کلاس‌ها می‌باشد. این موفقیت‌های چشمگیر در درجه اول مبتنی بر الگوریتم‌های پس‌انتشار[۲] و حذف تصادفی[۳] بوده است و با استفاده از واحدهای خطی که دارای گرادیان مناسبی هستند، بهبود یافته است [۷].

در شبکه‌های متخاصم، مدل مولد در برابر یک مدل تمایزگر قرار می‌گیرد: مدل تمایزگر می‌آموزد که مشخص کند نمونه از توزیع مدل است و یا از توزیع داده است. این مدل مولد را می‌توان مثل تیمی از جعل فرض کرد که سعی در تولید ارز جعلی و استفاده از آن بدون شناسایی دارد، و در طرف مقابل مدل تمایزگر مشابه پلیس است که سعی در کشف ارز تقلبی دارد. رقابت در این بازی، هر دو تیم را به سمت بهبود روش های خود سوق می‌دهد تا این‌که ارز تقلبی از ارز اصلی غیرقابل تشخیص باشد.

این چارچوب می‌تواند الگوریتم‌های آموزشی خاصی را برای انواع مختلف مسئله‌ها، مدل و بهینه‌سازی کند. شبکه مولد متخاصم که در این بخش قصد معرفی آن را داریم، بخش مدل مولد با دریافت نویز تصادفی و از طریق پرسپترون چند لایه، نمونه داده تولید می‌کند و مدل تمایزگر نیز از یک پرسپترون چند لایه تشکیل شده است. از این مورد خاص می‌توان به‌عنوان شبکه‌های دشمن استفاده کرد. در این نوع تعریف شبکه می‌توان هر دو مدل را با استفاده از الگوریتم‌های پس‌انتشار و حذف تصادفی ایجاد کرد و برای نمونه‌گیری از مدل مولد تنها از الگوریتم انتشار رو به جلو استفاده کرد و در نتیجه بکارگیری هیچ الگوریتمی برای استنباط تقریبی و یا زنجیره مارکوف ضروری نیست.

۳-۲- شبکه مولد متخاصم

در شبکه مولد متخاصم به طور همزمان دو مدل آموزش داده می‌شود؛ یک مدل مولد G که توزیع داده را ضبط می‌کند، و یک مدل تمایزگر D که 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 مینیماکس[۴] می‌شود. در فضای توابع دلخواه G و D، یک راه حل منحصر به فرد وجود دارد؛ این‌که شبکه

[1] Discriminative Models
[2] Backpropagation
[3] Dropout
[4] Minmax

مولد G توزیع داده‌های آموزشی را بازیابی کند و شبکه تمایزگر D احتمال را در همه جا برابر و مقدار ۱/۲ نشان دهد. با توجه به این‌که شبکه‌های G و D توسط پرسپترون های چند لایه تعریف می‌شود، کل سیستم را می توان با پس‌انتشار آموزش داد و در طول آموزش یا تولید نمونه‌ها، نیازی به زنجیره مارکوف یا شبکه‌های استنتاج نیست.

مدل‌سازی چارچوب متخاصم با اعمال مدل چند لایه پرسپترون برای هر دو مدل مولد و تمایزگر است. برای یادگیری توزیع مولد p_g روی داده x ، یک تابع نویز خالص $p_z(z)$ را به‌عنوان ورودی تعریف می‌کنیم، سپس یک نگاشت به فضای داده را به عنوان $G(z; \theta_g)$ 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 $D(x; \theta_d)$ ، با یک خروجی اسکالر تعریف می‌کنیم. $D(x)$ بیانگر احتمال این‌است که x از داده‌های اصلی به جای توزیع p_g ارائه شده باشد. به شبکه D آموزش داده می‌شود تا احتمال تخصیص برچسب صحیح را برای هر دو داده‌های آموزش و نمونه‌های تولیدی از G به حداکثر برساند. به طور هم‌زمان به شبکه G آموزش داده می‌شود تا تابع هدف $\log(1 - D(G(z)))$ را به حداقل برساند. به عبارت دیگر، شبکه‌های D و G بازی مینیماکس دو نفره زیر را با تابع $V(G, D)$ مطابق معادله ۳-۱ انجام می‌دهند:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim pdata(x)}[log\, D(x)] + \mathbb{E}_{z \sim p_{z(z)}}[log(1 - D(G(z)))]$$ معادله ۳-۱

۳با تحلیل نظری صورت گرفته بر روی شبکه‌های متخاصم، نشان داده شده که این شبکه‌ها پتانسیل کافی برای بازیابی توزیع داده‌های اصلی را در قالب شبکه مولد G دارند. رویکرد کلی در شکل ۳-۱ نشان داده شده است. در پیاده‌سازی این شبکه‌ها، از یک روش تکراری و عددی استفاده می‌شود. بهینه سازی D برای تکمیل حلقه درونی آموزش، محاسبات سنگینی دارد و در مجموعه داده‌های محدود، منجر به بیش‌برازش[1] می‌شود. در مدل مولد متخاصم، k مرحله بهینه سازی شبکه D و یک مرحله بهینه سازی شبکه G، بطور متناوب انجام می‌گیرد. این نوع بهینه‌سازی سبب می‌شود تا شبکه تمایزگر در حوالی جواب بهینه باقی بماند و شبکه مولد بتواند به آهستگی کافی، مدل داده را بیاموزد. روال آموزش این شبکه‌ها در الگوریتم ۳-۱ ارائه شده است.

در عمل، معادله ۳-۱ممکن است گرادیان کافی برای آموزش شبکه G فراهم نکند. در اوایل یادگیری، هنگامی که شبکه مولد G ضعیف است، شبکه تمایزگر D می تواند نمونه‌های تولیدشده را با اطمینان بالا رد کند زیرا آن‌ها با داده‌های آموزش کاملاً متفاوت هستند. در این حالت، $\log(1 - D(G(z)))$اشباع می‌شود. در این حالت، میتوان به جای آموزش شبکه G برای به حداقل رساندن تابع $\log(1 - D(G(z)))$ می توانیم G را برای به حداکثر رساندن

[1] Overfitting

$D(G(z))$ آموزش دهیم. این تابع هدف، در همان نقطه ثابت G و D قرار دارد اما گرادیان قوی‌تری در یادگیری فراهم می‌کند.

---

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**

- نمونه‌برداری کوچک‌دسته‌ای[1] mتایی نویز $\{z^{(1)}, \dots, z^{(m)}\}$ از نمونه‌های نویز $p_g(z)$.
- نمونه‌برداری کوچک‌دسته‌ای mتایی نویز $\{x^{(1)}, \dots, x^{(m)}\}$ از داده‌های تولید توزیع $p_{data}(x)$.
- بروزرسانی صعودی تمایزگر به وسیله گرادیان تصادفی.

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

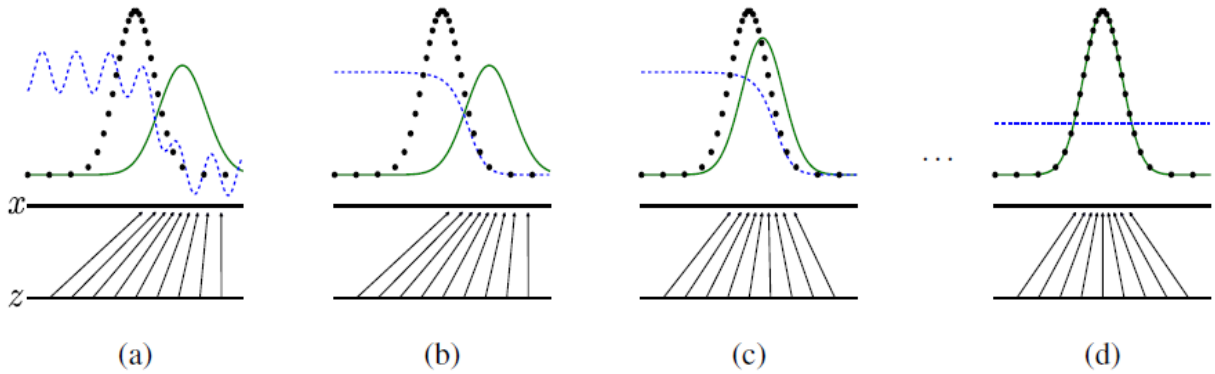
  end for

- نمونه‌برداری کوچک‌دسته‌ای mتایی نویز $\{z^{(1)}, \dots, z^{(m)}\}$ از نمونه‌های نویز $p_g(z)$.
- بروزرسانی صعودی مولد به وسیله گرادیان تصادفی.

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

end for

الگوریتم ۳-۱: آموزش گرادیان نزولی کوچک دسته‌ای[2] شبکه‌های مواد متخاصم.



شکل ۳-۱: شبکه‌های مولد متخاصم [۷]

در توضیح بیشتر شکل ۳-۱ می‌توان گفت که در ابتدا با به روز کردن توزیع تمایزگر (خط آبی شکسته) آموزش داده می‌شود تا بدین نمونه‌های توزیع داده‌های اصلی (خط مشکی نقطه‌چی) از داده‌های تولیدشده توسط توزیع مولد p_g (خط سبز پیوسته)، قابل تمایز باشد. خط افقی پایین دامنه، نشان‌دهنده z است که به طور یکنواخت

[1] Minibatch

نمونه گرفته شده است. خط افقی بالا بخشی از دامنه x است. فلش های رو به بالا نشان می دهد که چگونه $x = G(z)$ توزیع غیر یکنواخت p_g را بر روی نمونه‌ها نگاشت می‌کند. و به مرور زمان G به سمت مناطقی که p_g چگالی بالایی دارد، متمایل می‌شود. در مرحله اول شکل ۳-۱، ابتدا شبکه تمایزگر خود را اصلاح می‌کند. در مرحله بعد شبکه مولد، با توجه به تغییرات شبکه تمایزگر، بهبود می‌یابد و این بهبود اگر G و D از ظرفیت کافی برخوردار باشند، به نقطه‌ای می‌رسند که نمی‌توانند بهبود بیشتری داشته باشند، زیرا توزیع تابع مولد بر توزیع داده اصلی منطبق شده است. $p_g = p_{data}$. در چنین حالتی شبکه تمایزگر قادر به تفکیک بین دو توزیع نیست و $D(x) = 1/2$ می‌باشد.

۳-۳- تحلیل نظری شبکه مولد متخاصم

شبکه مولد G بطور ضمنی یک تابع توزیع احتمال p_g را به عنوان توزیع نمونه‌های G(z) تعریف کرده است ($z \sim p_g$) . بنابراین، هدف همگرا کردن و بدست آوردن یک برآورد خوب از p_{data} در الگوریتم ۳-۱ در زمان کافی و با ظرفیت پردازش مناسب می‌باشد. در این بخش نشان می‌دهیم که در بازی مینیماکس بین دو شبکه در این مدل، $p_g = p_{data}$ یک بهینه عمومی است. برای این منظور، ابتدا تمایزگر بهینه D را برای هر مولد G در نظر می گیریم.

قضیه ۱. برای هر تابع مولد G ثابت، تابع تمایزگر بهینه D عبارت است از $D_G^*(x) = \frac{p_{data}(x)}{p_{data}(x)+p_g(x)}$

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

$$V(G,D) = \int_x p_{data}(x)\log\big(D(x)\big)\,dx + \int_z p_z(z)\log\Big(1 - D\big(g(z)\big)\Big)\,dz$$
$$= \int_x p_{data}(x)\log\big(D(x)\big)\,dx + p_g(x)\log\big(1 - D(x)\big)\,dx$$

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[1] احتمال $P(Y = y|x)$ تعبیر کرد، که Y بیانگر آن است که هرجا x از توزیع p_{data} باشد $(y = 1)$ و هر جا از توزیع p_g باشد $(y = 0)$ است. با این تعریف بازی مینماکس در معادله ۳-۱را می‌توان به‌صورت زیر، بازنویسی کرد:

$$C(G) \underset{D}{=} \min V(G,D) = \mathbb{E}_{x\sim pdata}[\log D_G^*(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D_G^*(G(z)))]$$
$$= \mathbb{E}_{x\sim pdata}[\log D_G^*(x)] + \mathbb{E}_{x\sim p_g}[\log(1 - D_G^*(x))]$$
$$= \mathbb{E}_{x\sim pdata}[\log \frac{p_{data}(x)}{p_{data}(x) + p_g(x)}] + \mathbb{E}_{x\sim p_g}[\log \frac{p_g(x)}{p_{data}(x) + p_g(x)}]$$

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

اثبات: طبق قضیه۱، می‌دانیم هنگامی‌که $p_{data} = p_g$ 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[2] داریم:

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

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 می-دهد. بدین ترتیب اثبات این قضیه نیز به پایان رسید.

۳-۴- مزایا و معایب

شبکه‌های مولد متخاصم نسبت به مدل‌های قبلی دارای مزایا و معایبی می‌باشد. این معایب در درجه اول این است که نمایش صریح $p_g(x)$ وجود ندارد و D باید در حین آموزش به خوبی با G هماهنگ شود (به طور خاص، G نباید بدون بروزرسانی D خیلی زیاد آموزش داده شود که در آن مقدار مقادیر زیادی z به همان مقدار x بدست می آید تا از تنوع کافی برای مدل سازی p_{data} برخوردار باشد، در هنگام یادگیری نیازی به استنباط نیست و می توان طیف گسترده‌ای از توابع را در مدل گنجانید. مزایای فوق الذکر در درجه اول محاسباتی است. مدل‌های متخاصم همچنین ممکن است برخی از مزیت‌های آماری را از شبکه مولد به دست آورند که مستقیماً با نمونه داده ها به روز نمی‌شوند، اما فقط با گرادیان‌هایی که از طریق تمایزگر جریان می‌یابند، بروزرسانی می‌شود. این بدان

[1] Log-Liklihood
[2] Jensen-Shanon divergence

معنی است که اجزای ورودی مستقیماً در پارامترهای مولد کپی نمی‌شوند. یکی دیگر از مزیت‌های شبکه های متخاصم این است که آنها می‌توانند توزیع‌های بسیار تیز و حتی تخریب کننده را نشان دهند، در حالی که روش‌های مبتنی بر زنجیره های مارکوف نیاز دارند که توزیع تا حدی مبهم باشد تا زنجیرها بتوانند میان حالت‌ها مخلوط شوند.

۳-۵- جمع‌بندی

در این فصل ابتدا به تعریف کلی شبکه‌های مولد متخاصم پرداختیم. در ادامه به بررسی دقیق‌تر این شبکه و اجزای تشکیل‌دهنده و نحوه تعامل این شبکه‌ها با یکدیگر پرداختیم. سپس از لحاظ نظری روال تعریف‌شده برای رسیدن به نقطه بهینه را اثبات کردیم و در پایان فصل به نقاط قوت و ضعف این شبکه‌ها پرداختیم.

فصل چهارم : تشخیص ناهنجاری با استفاده از شبکه‌های مولد متخاصم

۴-۱- مقدمه

تشخیص ناهنجاری مسئله‌ای است که از اهمیت عملی زیادی در طیف وسیعی از کاربردهای دنیای واقعی برخوردار است، از جمله این کاربردها می‌توان به امنیت سایبری، تشخیص جعل و تصویربرداری پزشکی اشاره کرد. اصولاً روش‌های‌تشخیص ناهنجاری برای شناسایی نمونه‌های‌ناهنجار نیاز به الگوبرداری از داده‌های عادی دارند. اگرچه طیف گسترده‌ای از مطالعات روی مسئله تشخیص ناهنجاری صورت گرفته است، اما هم‌چنان ساخت یک روش کارآمد برای داده‌های‌پیچیده و با ابعاد بالا به عنوان چالش این حوزه مطرح می‌باشد.

شبکه های مولد متخاصم یک چارچوب مدل‌سازی قدرتمند برای داده‌های با ابعاد بالا است که می‌تواند این چالش را برطرف کند. شبکه‌های مولد متخاصم استاندارد از دو شبکه رقیب تشکیل شده است، یک شبکه مولد G و یک شبکه تمایزگر D. شبکه مولد یک نگاشت از فضای متغیرهای‌تصادفی نهفته z (توزیع های گوسین یا یکنواخت) به فضای‌داده مدل می‌کند، درحالی‌که شبکه تمایزگر یاد می‌گیرد بین داده‌های‌غیر واقعی تولیدشده توسط G و نمونه‌های‌اصلی تمایز قائل شود. این شبکه‌ها یک مدل بسیار موفق برای تصاویر طبیعی بوده است و به‌طور فزاینده-ای در گفتار و کاربردهای‌تصویربرداری پزشکی مورد استفاده قرار گرفته است.

با این حال این روش در هر بار آزمون نیاز به حل یک مسئله بهینه‌سازی دارد تا یک فضای z نهفته را پیدا کند به گونه‌ای که $G(z)$ تصویری مشابه فضای‌داده تولید کند. این فضای‌نهفته برای محاسبه میزان ناهنجاری برای نمونه‌ها استفاده می‌شود. این نیاز به حل یک مسئله بهینه‌سازی برای هر مرتبه آزمون، این روش را در داده‌های بزرگ یا برای برنامه‌های زمان‌واقعی غیرقابل استفاده می‌کند.

در فصل قبل مطالعه دقیقی روی پیش‌نیازهای اطلاعاتی از جمله شبکه مولد متخاصم و چگونگی پیاده‌سازی آن‌ها، نقاط قوت و ضعف این شبکه‌ها، آشنایی با مجموعه داده‌ها و چگونگی مدل‌سازی آن‌ها داشتیم. در این فصل به آشنایی دقیق‌تر با الگوریتم‌های بهبودیافته برای تشخیص ناهنجاری و رویکردها در چارچوب GAN می پردازیم. اولین رویکرد یادگیری خصمانه استنتاج ALI[1] نام دارد، در این روش هر دو شبکه استنتاج (یا رمزگذار) و شبکه مولدعمیق (یا رمزگشا) را در یک چارچوب متخاصمی GAN مانند قرار می‌گیرند. در این چارچوب تمایزگر یاد می‌گیرد تا بین زوج نمونه‌هایی - از جنس فضای داده ها و متغیرهایی از جنس فضای نهفته- که توسط دو شبکه استنتاج و شبکه مولد عمیق تولید می‌شود، تمایز قائل شود. در این ساختار نه تنها تمایزگر نمونه‌های مصنوعی را از داده‌های واقعی تشخیص می‌دهد، بلکه بین دو توزیع مشترک فضای داده و متغیرهای نهفته تفاوت قائل می‌شود

[1] Adversarially Learned Infernce

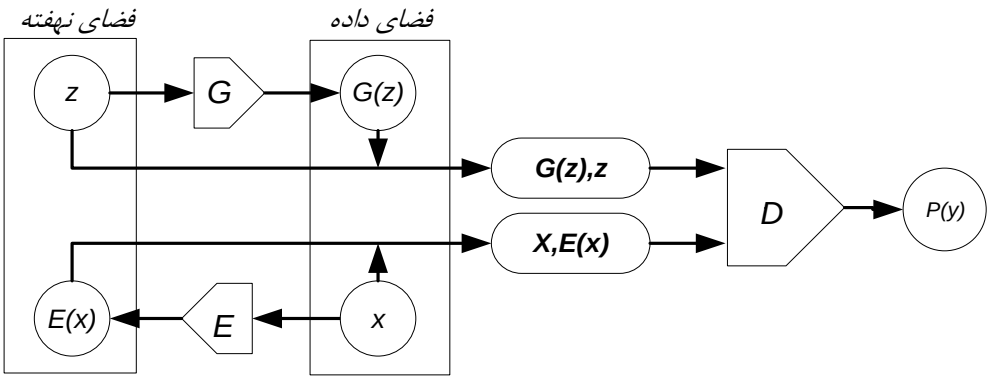
[۵]. رویکرد بعدی الگوریتم Ano-GAN[۴]و الگوریتم ALICE- GAN[۱] [۸] است. در نهایت الگوریتم ALAD[۲] [۲] که در ادامه کارهای پیشین است و به بهبود کارایی آن‌ها پرداخته مورد بررسی قرار می‌دهیم.

شبکه ALAD ارتباط نزدیکی با شبکه Ano-GAN دارد. اما بر خلاف شبکه Ano-GAN که از یک شبکه GAN استاندارد استفاده می‌کند، شبکه ALAD بر پایه GAN های دو جهته عمل می‌کند و از این رو شامل یک شبکه رمزگذار نیز می‌شود که نمونه‌ها از فضای داده‌های اصلی را به متغیرهای فضای نهفته نگاشت می‌کند. استفاده از این شبکه ما را از روش استنتاج محاسباتی گران‌قیمت مورد نیاز Ano-GAN بی نیاز می‌کند؛ زیرا متغیرهای نهفته مورد نیاز با استفاده از یک گذر رو به جلو[۳] از طریق رمزگذار در زمان تست قابل بازیابی است. هم‌چنین در این شبکه معیارهای ارزیابی ناهنجاری با Ano-GAN متفاوت است. در ادامه بررسی شبکه‌ها یادشده در بالا می‌پردازیم:

۴-۲- شبکه ALI

این شبکه در سال ۲۰۱۷ در کنفرانس ICLR معرفی شد [۵]. این شبکه‌ها با هدف یادگیری نگاشت معکوس از دامنه ورودی‌ها x به دامنه توزیع z تعریف شد. در این شبکه، علاوه بر شبکه مولد G که در معماری اصلی نیز تعریف شده بود، یک رمزگذار[۴] E نیز وجود دارد که از دامنه داده‌های ورودی x به دامنه ویژگی‌ها z می‌برد. بدین ترتیب خروجی بخش مولد یک دوتایی[۵] 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 متخاصمانه 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/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 با ویژگی‌های z درست شده‌است. در تصویر زیر چارچوب کلی این الگوریتم، به نمایش درآمده است:

[1] Adversarially Learned Inference Cross Entropy
[2] Adversarially Learned Anomaly Detection
[3] Feed Forward
[4] Encoder
[5] Tuple



شکل ۴-۱: معماری شبکه ALI [۵]

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

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

این دو توزیع، توابع توزیع حاشیه‌ای دارند که برای ما آشناست: توزیع حاشیه‌ای رمزگذار $q(x)$ تابع توزیع داده‌های اصلی است و توزیع حاشیه‌ای رمزگشا $p(z)$ معمولا به عنوان یک تابع توزیع ساده مانند تابع توزیع استاندارد $p(z) = N(0,I)$ در نظر بگیریم. بدین ترتیب روند تولید $p(x,z)$ و $q(x,z)$ 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 متخاصمانه صورت می گیرد. جفت (x,z) از دو توزیع $q(x,z)$ یا $p(x,z)$ در نظر گرفته می شود و یک شبکه تمایزگر می آموزد تا بین این دو خروجی، تمایز قائل شود؛ در حالی که دو شبکه رمزگشا و رمزگذار می‌آموزند تا این شبکه را فریب دهند. در نهایت تابع ارزشی که این بازی بر اساس آن صورت می‌گیرد به صورت زیر است:

$$\min_G \max_D V(D,G) = \mathbb{E}_{q(x)}[\log D(x, G_z(x))] + \mathbb{E}_{p(z)}[\log(1 - D(G_z(z), z)]$$

$$= \iint q(x)q(z|x)log\big(D(x,z)\big)dxdz + \iint p(x)p(x|z)log\big(1 - D(x,z)\big)dxdz$$

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

به طور دقیق‌تر شبکه تمایزگر آموزش می بیند که بین نمونه‌هایی که از رمزگذار $(x\ ،\ z) \sim q\ (x\ ،\ z)$ و نمونه هایی که از رمزگشا $(\hat{x}\ ،\ z) \sim p\ (x\ ،\ z)$ تولید می‌شود، تمایز بگذارد. شبکه مولد و شبکه رمزگذار 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.

در الگوریتم ۴-۱شبکه Ali توصیف شده است. اثبات می‌شود که با فرض یک تمایزگر بهینه، شبکه مولد، واگرایی جنسن-شانون را بین $p\ (x\ ،\ z)$ و $q\ (x،z)$ به حداقل می رساند [۵].

رویه آموزش یادگیری خصمانه استنتاج

$\theta_g, \theta_d \leftarrow$ مقداردهی اولیه پارامترها

Repeat

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

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

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

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

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

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

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

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

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

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

until

الگوریتم ۴-۱: رویه آموزش یادگیری خصمانه استنتاج

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

۱. بخش مولد دارای دو مؤلفه است: بخش رمزگذار، $G_z(\mathrm{x})$که نمونه‌های داده x را به z-space نگاشت می‌کند و بخش رمزگشایی $G_x(\mathrm{z})$ که نمونه‌ها را از $p(z)$ (منبع منبع نویز) به فضای ورودی نگاشت می‌کند.

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

۴-۳- شبکه Ano-GAN

برای تشخیص ناهنجاری‌ها در آناتومی بدن، مدل Ano-GAN بر مبنای شبکه GAN در سال ۲۰۱۷ ارائه شد [۴]. این روش یک مدل مولد و یک تمایزگر را برای تمایز بین داده‌های تولید شده و واقعی به طور هم‌زمان آموزش می‌دهد و به جای بهینه‌سازی تابع هزینه واحد، هدف آن تعادل هزینه نش[1] است که سبب افزایش قدرت ویژگی مدل تولیدی و دقت بالا در طبقه‌بندی داده‌های واقعی از داده‌های مولد می‌شود، عملکرد کلی این مدل بدین صورت است:

M مجموعه‌ای داده سالمی از تصاویر پزشکی که با I_m نمایش داده می‌شود که $m = 1,2, \dots, M$، 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.

رمزگذاری با یک شبکه مولد متخاصم شامل دو ماژول مخالف، مولد G و تمایزگر D است. شبکه مولد G توزیع p_g را روی داده x از طریق نگاشت G (z) با نمونه‌بردار z آموزش می‌بیند؛ یعنی بردارهای تک بعدی با توزیع یکنواخت از فضای نهفته Z نمونه‌برداری می‌کند و فضای دو بعدی تصویر با توزیع حاشیه‌ای x دارای نمونه‌های سالم، نمونه برداری می‌شود. در این تنظیمات، معماری شبکه مولد G معادل یک رمزگذار پیچشی که از پشته‌های پیچشی گام به گام 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 $D(.)$ احتمال این است که ورودی تمایزگر، تصویر واقعی نمونه‌برداری شده در آموزش داده $\mathcal{X}$ باشد در مقابل این که این تصویر توسط مولد G با توزیع G(z) تولید شده باشد. D و G به طور همزمان از طریق بازی مینماکس با تابع $V(D,G)$ و معادله ۴-۱ بهینه‌سازی می‌شوند [۹].

$$\min_G \max_D V(D,G) = \mathbb{E}_{x \sim p_{data}(x)}[log\, D(x)] + \mathbb{E}_{z \sim p_{z(z)}}[log(1 - D(G(z)))] \quad \text{معادله ۴-۱}$$

در این بازی شبکه تمایزگر آموزش می‌بیند که احتمال اختصاص نمونه‌های واقعی را بیشینه و نمونه‌های تولیدی از p_g با برچسب جعلی کمینه کند. هم‌چنین شبکه مولد G آموزش می‌بیند همزمان با حداقل کردن $V(G) = log(1 - D(G(z)))$ که معادل با حداکثرکردن $V(G) = D(G(z))$ است، شبکه تمایزگر D را فریب دهد.

[1] Nash Cost

و به طور کلی در طول آموزش خصمانه، مولد در تولید تصاویر واقع بینانه و تمایزگر در شناسایی صحیح تصاویر واقعی و تولید شده بهبود می‌یابد.

۴-۳-۱-نگاشت تصاویر جدید به فضای نهفته

وقتی آموزش متخاصم به پایان رسید، شبکه مولد یاد می‌گیرد که $G(z) = z \to x$ را از فضای نهفته با نمایش z به تصویر واقعی (عادی) x نگاشت کند. شبکه‌های GAN به‌طور خودکار نگاشت معکوس $\mu(x) = x \to z$ را انجام نمی‌دهد. فضای نهفته دارای گذار خطی است [۹]، 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 بیشتر 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 می شود. برای پیدا کردن بهترینz ، با نمونه‌گیری تصادفی z_1 از توزیع فضای نهفته $\mathcal{Z}$ شروع می کنیم 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 $G(z_\Gamma)$، نقطه z در فضای نهفته $\mathcal{Z}$ در یک فرآیند تکراری از طریق $\gamma = 1,2, \dots, \Gamma$ با مراحل پس‌انتشار بهینه می‌شود.

تعریف یک تابع اتلاف برای نگاشت از تصاویر فضای نهفته شامل دو بخش است [۱۰] ، باقی‌مانده اتلاف[1] و باقی‌مانده تمایز[2]. 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 انطباق ضرایب z از طریق پس‌انتشار مورد استفاده قرار می‌گیرند.

باقی‌مانده اتلاف: این معیار عدم شباهت بصری بین تصویر مورد بررسی x و تصویر تولید شده $G(z_\gamma)$ در فضای تصویر اندازه‌گیری می‌کند و به‌صورت معادله ۴-۲ تعریف می‌شود.

$$\mathcal{L}_R(z_\gamma) = \sum |x - G(z_\gamma)| \quad \text{معادله ۴-۲}$$

[1] Residual Loss
[2] Discrimination Loss

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 تلفات برابر با صفر است.

باقی‌مانده تمایز: برای تأمین تصویر، باقی‌مانده تمایز $\mathcal{L}_{\hat{D}}(z_\gamma)$ را بر اساس خروجی تمایزگر با اعمال تصویر تولید شده $G(z_\gamma)$ در تمایزگر $\mathcal{L}_{\hat{D}}(z_\gamma) = \sigma(D(G(z_\gamma)), \alpha)$ محاسبه می‌کنیم، در این رابطه σ آنتروپی متقاطع سیگموئید[1]، که از تمایز اتلاف تصاویر واقعی حین آموزش متخاصم، با تابع لاجیت[2] $D(G(z_\gamma))$ و $\alpha = 1$ تعریف می‌شود.

۴-۳-۲- 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 می‌کنیم. بنابراین، یک نمره ناهنجاری که تناسب تصویر مورد جستجو x را با مدل تصاویر عادی بیان می‌کند، این معیار می‌تواند مستقیماً از تابع اتلاف معادله ۴- بدست آید.

$$A(x) = (1 - \lambda).R(x) + \lambda.D(x)$$ معادله ۴-۳

که در آن به ترتیب $R(x)$ امتیاز باقی‌مانده و $D(x)$ امتیاز تمایزگر است که توسط باقی‌مانده اتلاف$\mathcal{L}_R(z_\gamma)$ و باقی‌مانده تمایز$\mathcal{L}_D(z_\Gamma)$ در فضای نهفته تعریف می‌شود. 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 قبلاً در طول آموزش دیده شده، می‌باشد. برای تشخیص ناهنجاری مبتنی بر تصویر، از نمره ناهنجاری $A(x)$ استفاده می‌شود. علاوه بر این، از تصویر باقی‌مانده $x_R = |x - G(z_\Gamma)|$ برای شناسایی مناطق غیر عادی در یک تصویر استفاده میشود. برای مقایسه، علاوه بر این امتیاز ناهنجاری را مطابق معادله ۴-۴ تعریف می‌شود، در اینجا $\hat{D}(x) = \mathcal{L}_{\hat{D}}(z_\Gamma)$ همان امتیاز تمایزگر و $R(x)$ امتیاز باقی‌مانده می‌باشد.

$$\hat{A}(x) = (1 - \lambda).R(x) + \lambda.\hat{D}(x)$$ معادله ۴-۴

[1] Sigmoid Cross Entropy
[2] logits

۴-۴- شبکه 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 [۵]. دراین مدل در یک چارچوب شبکه مولد متخاصم، شبکه تمایزگر می‌آموزد تا تفاوت بین دو توزیع توامان را تشخیص دهد.

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.

۴-۴-۱- یادگیری خصمانه با اندازه‌گیری اطلاعات

به یاد داریم که تابع هدف در شبکه ALI به صورت معادله ۴-۵ بود:

$$min_G max_D V(D,G) = E_{q(x)}[log(D(x, G_z(x))] + E_{p(z)}[log(1 - D(G_x(z), z)] \quad \text{معادله ۴-۵}$$

و می‌دانیم 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.

۴-۴-۲- آنتروپی شرطی[1]

آنتروپی شرطی یک معیار نظریه اطلاعاتی است که عدم قطعیت متغیر تصادفی x را هنگام مقید شدن بر روی Z با کمک توزیع توامان $\pi(x\,،\,z)$ تعیین می‌کند:

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

عدم قطعیت متغیر x مقید بر روی 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:

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

$V_{CE}(E\,,G)$ 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.

۴-۴-۳- 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[2] محدود کرد. در این‌جا برای بازسازی $\hat{x}$ به طریق زیر عمل می شود:

$$\hat{x} \sim p(\hat{x} \mid z)\,،z \sim q(z \mid x)\,،x \sim q(x)$$

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.

هنگامی که شبکه ALI به نقطه بهینه می‌رسد، $q(x\,,z)$ و$p(x\,,z)$ به تابع توزیع توامان $\pi(x\,،\,z)$ میل می‌کند و $V_{CE}(E,G)$ به آنتروپی شرطی متغیرها میل می‌کند.

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 معادله ۴- خواهد بود.

---

[1] Conditional Entropy
[2] Cycle Consistency

$$\min_{E,G}\max_{D_{xz},D_{xx}} V_{ALICE} = \mathrm{V_{ALI}} + \mathrm{E}_{x\sim q(x)}[\log \mathrm{D}_{xx}(\mathrm{x},\mathrm{x}) + \log 1 - \mathrm{D}_{xx}(\mathrm{x},\mathrm{G}(\mathrm{E}(\mathrm{x})))]$$ معادله ۴-۶

ویژگی پایداری چرخش در مقالات پیش از نیز وجود داشته است این ویژگی در این مقالات به کمک نرم درجه[1] ۱ و ۲ 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 $\hat{X}$ های بازسازی‌شده را اندازه‌گیری می‌کند، استفاده شده است.

۴-۵- شبکه 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 رمزگذار 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 رمزگذار 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 [۲].

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[2] شبکه مولد برای یافتن متغیرهای نهفته z است که به معنای به حداقل رساندن خطای بازسازی با تابع هدف گرادیان نزولی تصادفی می باشد. 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 متخاصمی به همراه یک شبکه رمزگذار E (که نمونه‌ها را از فضای داده x به فضای نهفته z نگاشت می‌کند) استفاده می‌شود. نمایش نهفته هر نمونه از فضای داده در چنین مدل‌هایی صرفاً با عبور از شبکه رمزگذار انجام می‌شود. هم‌چنین این مدل پیشرفت‌های اخیر که برای بهبود شبکه رمزگذار صورت گرفته مانند افزودن یک شبکه تمایزگر برای بهبود سازگاری چرخه $(G(E(x)) \approx x)$ را شامل می‌شود.

[1] L-norm
[2] Invert

شبکه Ali توزیع توامان داده‌ها را به همراه یک شبکه رمزگذار مدل می‌کند. 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 رمزگذار – تولید شده است.

این مسئله به صورت معادله ۴- مدل می شود:

$$V(D_{xz}, G, E) = E_{x\sim px}[log\left(D_{xz}(x, E(x))\right)] + E_{z\sim pz}[log(1 - D_{xz}(G(z), z))]$$ معادله ۴-۷

با این‌که به لحاظ نظری توزیع توامان$p_G(x,z)$و$p_E(x,z)$ 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 $G(E(x)) \approx x$ 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 به صورت زیر به روش تخاصمی برای سازگاری چرخه تقریب بزنیم:

$$H^{\pi}(x \mid z) = -E_{\pi(x,z)}[\log \pi(x \mid z)]$$

در این تعریف $\pi(x, z)$ به معنای توزیع توامان z و x می‌باشد. این مسئله شامل یکپارچه‌سازی آنتروپی شرطی در شبکه رمزگذار E و شبکه مولد G است:

$$V_{ALICE}(D_{xz}, E, G) = V(D_{xz}, E, G) + V_{CE}(E, G)$$

که افزودن این عامل به مسئله معادل افزودن یک شبکه تمایزگر جدید D_{xx} به مسئله می‌باشد. با افزودن این عامل، مسئله بهینه‌سازی به صورت معادله ۴- خواهد بود:

$$V(D_{xx}, G, E) = E_{x\sim px}[log(D_{xx}(x, x)) + E_{x\sim px}[log(1 - D_{xx}(x, G(E(x))))]$$ معادله ۴-۸

۴-۵-۱-تثبیت آموزش GAN بر پایه ALICE

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.

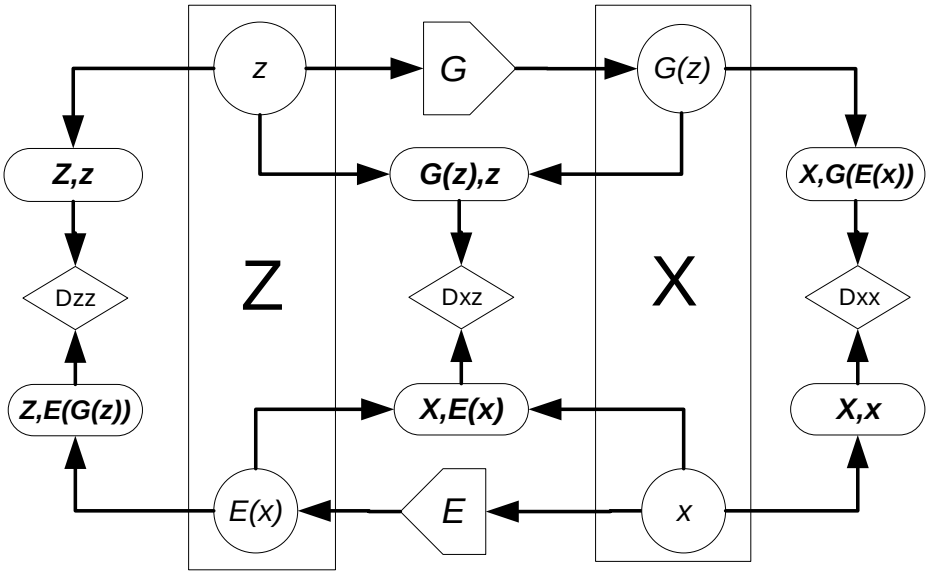
توضیح دقیق‌تر این‌که، فضای نهفته شرطی $H^{\pi}(z \mid x) = -E_{\pi(x,z)}[\log \pi(z \mid x)]$ را با یک شبکه تمایزگر مخالف دیگر D_{zz} با نقطه تعادل مشترک تنظیم می‌کنیم:

$$V(D_{zz}, G, E) = E_{z\sim pz}[log(D_{zz}(z, z)) + E_{z\sim pz}[log(1 - D_{zz}(z, G(E(z))))]$$

با کنار هم قرار دادن تمامی این اجزا، شبکه ALAD تلاش می‌کند تا نقطه تعادل این مسئله را آموزش ببینند:

$$min_{G,E}max_{D_{xz},D_{xx},D_{zz}}V_{ALAD}(D_{xz}, D_{xx}, D_{zz}, E, G) = V(D_{xz}, E, G) + V(D_{xx}, E, G) + V(D_{zz}, E, G)$$

در نهایت شماتیک کلی شبکه طراحی‌شده به صورت شکل ۴-۲خواهد بود :



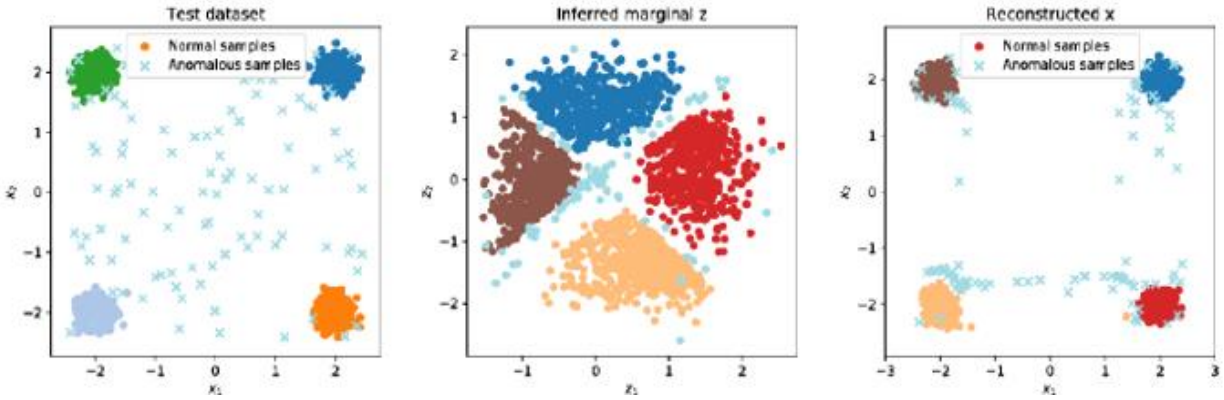
شکل ۴-۲: شبکه ALAD [۲]

اضافه کردن مرحله نرمال‌سازی طیفی با انگیزه یادشده مقاله [۱۱] می‌باشد. در این مقاله نشان داده‌شده که با افزودن قیود لیپسچیتز[1] 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 که افزودن این قیود، نه تنها برای شبکه تمایزگر، بلکه برای شبکه رمزگذار نیز سودمند است. قابل توجه است که مدل ALICE شامل این مرحله نمی باشد.

[1] Lipschitz Constraints

۴-۵-۲- 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. این شهود در شکل ۴-۳ نشان داده شده است.



شکل ۴-۳ : تشخیص ناهنجاری [۲]

برای این منظور، ابتدا باید توزیع داده را به‌طور مؤثری مدل کنیم: این مرحله با استفاده از شبکه GAN توصیف شده حاصل می‌شود، تابع مولد G برای یادگیری توزیع داده‌های عادی استفاده می‌شود، به طوری که $p_G(x) = \int p_G(x \mid z)\, p_Z(z)dz$ که $p_G(x) = p_X(x)$ می‌باشد. هم‌چنین باید توزیع حاشیه‌ای داده‌ها را بیاموزیم تا بازنمایی‌های نهفته منجر به بازسازی صحیح نمونه‌های عادی شود. دو عنصر پایداری چرخه شرطی متقارن در این مدل، به اطمینان از این امر کمک می‌کند.

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[1] 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[2] مطابق معادله ۴-۹ تعریف می‌شود. در این تعریف

[1] Logit
[2] L1-norm

$f(.,.)$ تابع فعال‌ساز[1]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[2] و بازسازی‌شده و در نتیجه از توزیع داده واقعی می‌باشد.

$$A(x) = ||f_{xx}(x,x) - f_{xx}\left(x, G\left(E(x)\right)\right)||_1$$ معادله ۴-۹

با این تعریف، نمونه‌ها با A بیش‌تر به احتمال بالاتری داده ناهنجار خواهند بود. در ادامه در الگوریتم ۴-۲روال محاسبه A(X) ارائه می‌شود.

الگوریتم محاسبه نمره ناهنجاری شبکه ALAD

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

خروجی $A(x)$

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

الگوریتم ۴-۲: الگوریتم ALAD

معیار استفاده شده در این‌جا از ایده تطابق ویژگی‌های از دست‌رفته الهام گرفته شده است [۱۲]. اما در این‌جا به جای استفاده از ویژگی‌های محاسبه شده درشبکه تمایزگر GAN استاندارد (که اختلاف را بین نمونه‌های تولید شده و داده‌های واقعی را محاسبه می‌کند)، از ویژگی‌های محاسبه‌شده در شبکه تمایزگر D_{xx} 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 رمزگذار 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.

[1] Activation
[2] Encode

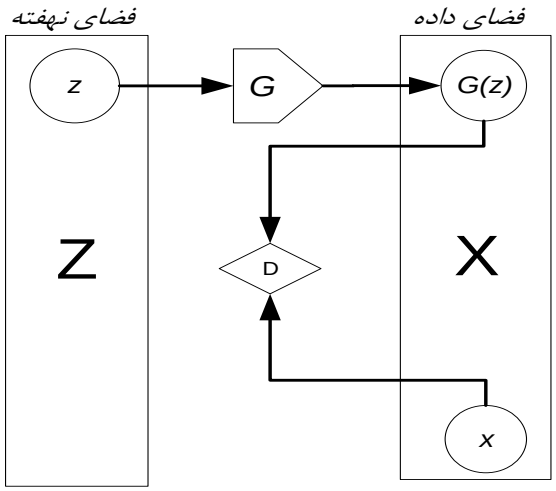
۴-۶- جمع‌بندی

در این فصل مدل‌های مختلف شبکه مولد متخاصم بکار گرفته‌شده برای تشخیص ناهنجاری را بررسی و معماری هر کدام را ارائه کردیم. اولین رویکرد یادگیری خصمانه استنتاج ALI نام داشت، در این روش هر دو شبکه استنتاج و یا رمزگذار و شبکه مولد عمیق و یا رمزگشا را در یک چارچوب تخاصمی GAN مانند قرار می‌گیرند. در این چارچوب تمایزگر یاد می‌گیرد تا بین زوج نمونه‌هایی که توسط دو شبکه استنتاج و شبکه مولد عمیق تولید می‌شود، تمایز قائل شود. رویکرد بعدی الگوریتم Ano-GAN بودکه از یک شبکه GAN استاندارد استفاده می‌کند. برای حل مشکل عدم پایداری چرخش در شبکه ALI شبکه ALICE ابداع شد، در این شبکه یک عضو آنتروپی در قالب یک شبکه تمایزگر D_{xx} به چارچوب ALI اضافه گردید. شبکه ALAD با هدف تثبیت فاز آموزش، یک شبکه تمایزگر دیگر به چارچوب ALICE اضافه کرد.

فصل پنجم : جمع‌بندی و نتیجه‌گیری

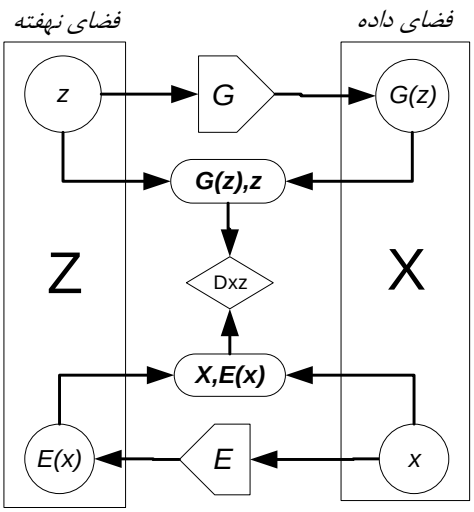
در فصل اول به اهمیت و لزوم تشخیص ناهنجاری پرداختیم. در فصل دوم ابتدا به تعریف ناهنجاری و بررسی کاربردهای تشخیص ناهنجاری پرداختیم. سپس به طبقه‌بندی روش‌های موجود تشخیص ناهنجاری پرداختیم. پس از آن تعدادی از روش‌های متداول تشخیص ناهنجاری را بررسی کردیم و با توضیحات صورت گرفته در این فصل، به این نتیجه رسیدیم که یکی از بروزترین و در عین حال اثرگذارترین روش‌های موجود، بکارگیری از روش‌های یادگیری عمیق می‌باشد و دانستیم در میان این روش‌ها، شبکه‌های مولد تخاصمی، عملکرد بسیار مناسبی برای داده‌های حجیم و با ابعاد بالا داشته‌اند. در فصل سوم به بررسی تعریف و اصول این نوع شبکه پرداختیم و در فصل چهارم با چند نمونه از شبکه‌های مولد متخاصمی موفق در زمینه تشخیص ناهنجاری پرداختیم. در ادامه قصد داریم سیر تکاملی و چالش‌های هر روش را بررسی کنیم.

شبکه مولد متخاصمی در سال ۲۰۱۴ توسط آقای گودفلو و همکاران معرفی شد. معماری این شبکه مطابق شکل ۵-۱ است.



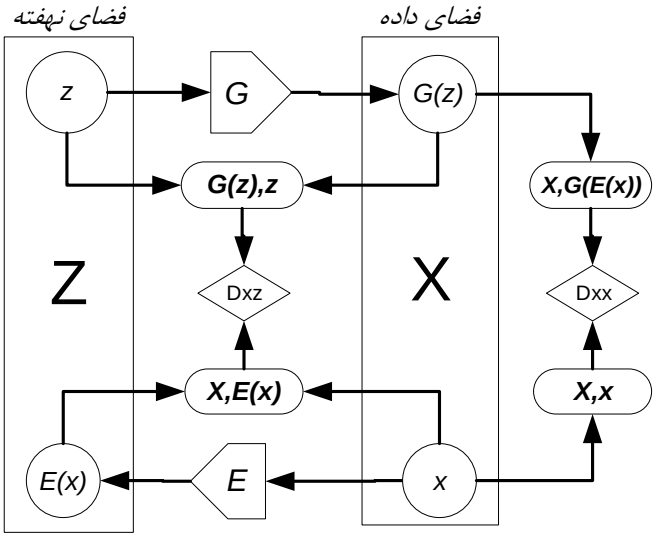
شکل ۵-۱: معماری شبکه GAN

همان‌طور که گفته‌شد، برای تشخیص ناهنجاری علاوه بر شبکه مولد نیاز به یک فرایند استنتاج داریم تا نگاشتی فضای داده به فضای نهفته داشته باشیم. انجام این فرایند به صورت مستقیم، از لحاظ محاسباتی بسیار پرهزینه می‌باشد. به منظور حل این مشکل در سال ۲۰۱۷ در کنفرانس ICLR، آقای دومولین و همکاران شبکه ALI را ارائه کردند. این شبکه دو تفاوت اساسی با شبکه GAN معمولی دارد. اول این که در این شبکه برای فرایند استنتاج، از یک شبکه عصبی رمزگذار برای یادگیری نگاشت از فضای داده به فضای نهفته استفاده شد. دوم این که شبکه تمایزگر توزیع توامان داده و متغیر نهفته معادلش را یاد می‌گیرد. معماری این شبکه مطابق شکل ۵-۲ است.



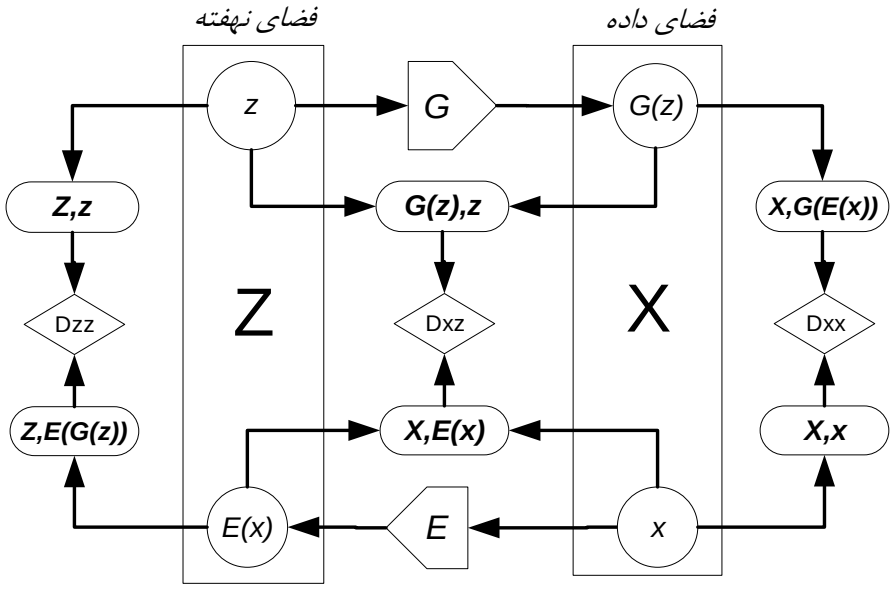
شکل ۵-۲: شبکه ALI-GAN

شبکه پیشنهادی آقای دومولین یک ایراد اساسی داشت؛ در تابع بهینه‌سازی این شبکه، لزومی به همبستگی بین دو متغیر تصادفی شرطی x و z ندارد. نمونه‌ای از مشکلات این روش زمانی است که متغیری مثلا با تابع E از فضای داده به فضای نهفته برود و سپس با تابع G به فضای داده برگردد، در این شبکه هیچ لزومی ندارد که بازسازی صورت گرفته شباهتی به داده اصلی داشته باشد. این مشکل ثبات چرخه در کنفرانس NIPS همان سال مورد توجه قرار گرفت. برای حل این مشکل شبکه ALICE پیشنهاد شد. این شبکه، همان شبکه ALI بود که در کنار خود یک عنصر آنتروپی شرطی داشت که وظیفه اش سوق شبکه به سمت جواب‌هایی که ثبات چرخه دارند، بود. این بخش در این‌جا به صورت یک شبکه تمایزگر لحاظ شد، که هر دو ورودی‌اش از فضای داده است؛ یکی خود داده و دیگری بازسازی صورت‌گرفته از داده. معماری کلی این شبکه مطابق شکل ۵-۳ است.



شکل ۵-۳: شبکه ALICE

در سال ۲۰۱۸ آقای زناتی و همکاران برای تثبیت آموزش شبکه ALICE شبکه ALAD را پیشنهاد دادند. در این شبکه دو تغییر نسبت به شبکه ALICE صورت گرفته بود: اول این در این‌جا یک عنصر آنتروپی شرطی به شبکه قبلی اضافه شد. دوم این‌که در هر مرحله نرمال‌سازی طیفی روی ماتریس ضرایب لایه مخفی در شبکه تمایزگر صورت می‌گیرد. این روش در عمل سبب افزایش بازدهی شده و به تثبیت آموزش نیز کمک می‌کند. معماری شبکه ALAD مطابق شکل ۵-۴ می‌باشد.



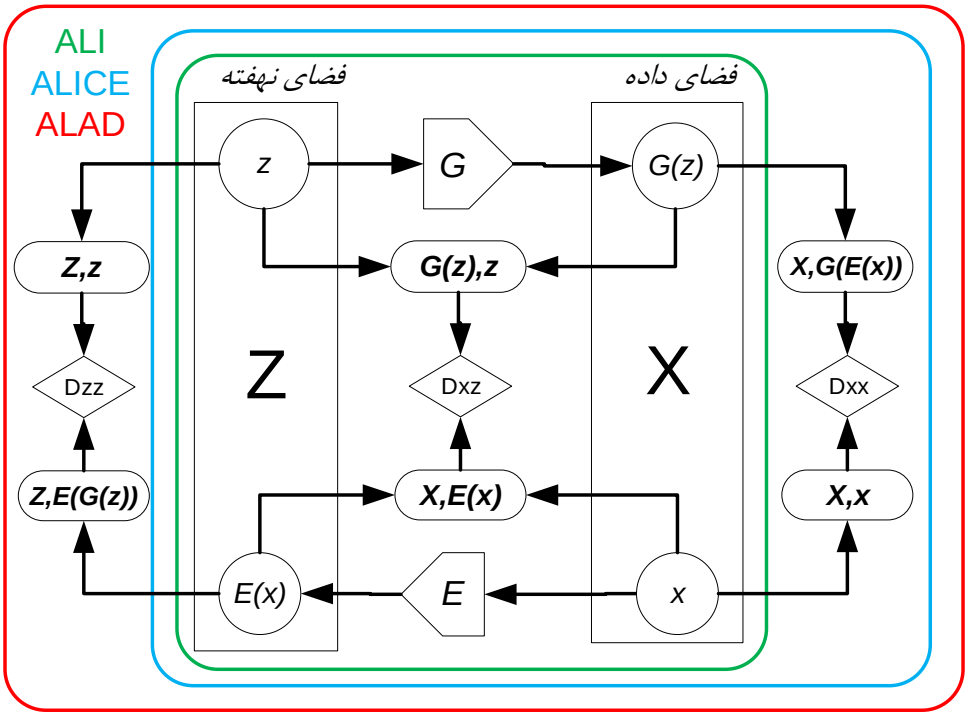
شکل ۵-۴: شبکه ALAD

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 هستند و همان‌طور که پیش‌تر توضیح داده شد، هر کدام در راستای برطرف کردن یک مشکل هستند.

جدول ۵-۱: توابع بهینه‌سازی مدل‌ها

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	$\min_{\mathrm{G}} \max_{\mathrm{D}} \mathrm{V}_{ALI}(\mathrm{D},\mathrm{G}) = \mathbb{E}_{\mathrm{q(x)}}[\log \mathrm{D}(\mathrm{x}, G_z(\mathrm{x}))] + \mathbb{E}_{\mathrm{p(z)}}[\log(1 - \mathrm{D}(G_z(z), z)]$
ALICE	$\min_{E,G} \max_{D_{xz},D_{xx}} V_{ALICE} = \mathrm{V_{ALI}} + \mathrm{E}_{x\sim q(x)}[\log \mathrm{D}_{xx}(\mathrm{x},\mathrm{x}) + \log 1 - \mathrm{D}_{xx}(\mathrm{x}, \mathrm{G}(\mathrm{E}(\mathrm{x})))]$
ALAD	$min_{G,E} max_{D_{xz}, D_{xx}, D_{zz},} V_{ALAD}(D_{xz}, D_{xx}, D_{zz,}, E, G)$ $= V_{ALICE} + E_{z\sim pz}[log(D_{zz}(z,z)) + log(1 - D_{zz}(z, G(E(z))))]$

خلاصه تمامی بحث‌های صورت گرفته در شکل شکل ۵-۵ آمده است:



شکل ۵-۵: معماری مدل‌های GAN

روال آتی مدنظر این پژوهش، تمرکز بر روی کاهش حجم محاسبات و زمان مورد نیاز برای آموزش شبکه ALAD می‌باشد. در واقع این شبکه با این‌که به نتایج قابل قبول و بسیار خوبی دست می‌یابد، ولی از منظر زمان آموزش تقریبا ۳ برابر شبکه‌هایی هم‌چون MDAN[۱۳] وقت نیاز دارد. هم‌چنین با این‌که در کاربرد، شبکه‌های مولد متخاصم برای تشخیص ناهنجاری بسیار خوب عمل کردند، اما علت بکارگیری این شبکه‌ها هم‌چنان پشتوانه تئوری قوی ندارد [۱۴]. یکی دیگر از روال‌های آتی ادامه پژوهش، ارائه یک الگوریتم با پشتوانه قوی ریاضی در ادامه شبکه ALAD می‌باشد.

منابع و مراجع

[۱] D. Hawkins, Identification of outliers, Netherlands: Springer, 1980.

[۲] H. Zenati, M. Romain, C. Foo, B. Lecouat and V. Chandrasekhar, "Adversarially Learned Anomaly Detection," in *IEEE International Conference on Data Mining*, 2018.

[۳] M. Ahmed, A. Mahmood and J. Hu, "A survey of network anomaly detection techniques," *Journal of Network and Computer Applications, Elsevier,* 2016.

[۴] T. Schlegl, P. Seebock, S. Waldstein, U. Schmidt-Erfurth and G. Langs, "Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery," in *International Conference on Information Processing in Medical Imaging, Springer*, 2017.

[۵] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky and A. Courville, "Adversarially Learned Inference," in *International Conference on Learning Representations*, 2017.

[۶] G. Muruti, F. Rahim and Z. Ibrahim, "A Survey on Anomalies Detection Techniques and Measurement Methods," *IEEE Conference on Applications, Information and Network Security,* 2018.

[۷] I. Goodfellow, J. Pouget-Abadiey, M. Mirza, B. Xu, D. Warde-Farley,S. Ozairz, A. Courville & Y. Bengiox, "Generative Adversarial Nets," in *NIPS*, 2014.

[۸] C. Li, H. Liu, C. Chen, Y. Pu, L. Chen, R. Henao, and L. Carin, "Alice: Towards understanding adversarial learning for joint," in *Advances in Neural Information*, 2017.

[۹] A. Radford, L. Metz and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *ICLR,* 2016.

[۱۰] R. Yeh, C. Chen, T. Lim, M. Hasegawa-Johnson and M. Do, "Semantic image inpainting with perceptual and contextual losses," *International Journal of Pattern Recognition and Artificial Intelligence,* 2016.

[۱۱] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *International Conference on Learning Representations,* 2018.

[۱۲] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung and A.Radford, "Improved techniques for training GANs," *Advances in Neural Information Processing Systems,* 2016.

[۱۳] Y.Hou, Z.Chen, M.Wu, C. Foo, X.Li and R.Shubair, "Mahalanobis Distance Based Adversarial Network for Anomaly Detection," *ICASSP,* 2020.

[۱۴] Z.Yang, I.Soltani, E.Darve, "Regularized Cycle Consistent Generative Adversarial Network for Anomaly Detection," *ECAI,* 2020.