وزارت ارتباطات و فناوری اطلاعات
پژوهشگاه ارتباطات و فناوری اطلاعات

بررسی ابعاد فناوری جعل عمیق (Deep Fake)

پژوهشگاه ارتباطات و فناوری اطلاعات

بررسی ابعاد فناوری جعل عمیق

(Deep Fake)

مقدمه

چالش های تأیید تصاویر معتبر از افراد مشهور قدمت چند صد ساله دارد و در سالهای اخیر با ظهور Adobe Photoshop و سایر تکنیک های ویرایش تصویر دیجیتال تشخیص اصل بودن یا نبودن هر گونه تصویری برای کاربران عادی تقریباً غیرممکن شده است. به عبارت دیگر عکاسی دیجیتال و فتوشاپ به ما آموخته اند که در مورد تصاویر عکاسی به روش دیگری فکر کنیم و همواره امکان جعلی بودن آنها را در نظر داشته باشیم تا جایی که اکنون شک و تردید در مورد صحت عکس‌ها امری عادی است. در حال حاضر، دست‌کاری عکسها به صورت خودکار انجام می شود. برای مثال نرم افزار AutoAwesome Google (که اکنون با عنوان "Assistant" شناخته می شود) تصاویر جعلی را به طور خودکار ایجاد می کند. یک سناریو برای استفاده از این ابزار چنین است: "مردی دو عکس را بارگذاری کرد تا ببیند همسرش کدام یک را ترجیح می دهد. در یک عکس، او برخلاف همسرش لبخند می زد و عکس دیگر برعکس این حالت بود. الگوریتم های پردازش تصویر گوگل از میان این دو تصویر که در فاصله چند ثانیه از یکدیگر گرفته شده اند، مورد سومی را ایجاد می کنند: ترکیبی که در آن هر دو فرد لبخند می زنند. اما در این حالت، نتیجه یک عکس از لحظه ای است که هرگز اتفاق نیفتاده، به عبارت دیگر یک خاطره نادرست یا بازنویسی تاریخی توسط یک عکس که می تواند به عنوان یک عکس واقعی در نظر گرفته شود، ایجاد شده است.

در هنگام پیشرفت تهدیدهای ناشی از دست‌کاری تصاویر ساکن، فیلمها یا به عبارت دیگر تصاویر متحرک نوید ارائه اطلاعات معتبرتری را میدادند زیرا ایجاد تغییر در آنها در مقایسه با تصاویر ساکن بسیار دشوارتر و پیچیده‌تر بود. اما تحولات در هوش مصنوعی و گرافیک پیشرفته رایانه ای منجر به فناوری شده است که در آن فیلم تغییریافته می

تواند کاملاً جایگزین یک فیلم معتبر و واقعی شود. نمونه فعلی این نوآوری ، فناوریست که به رایانه ها امکان می دهد چهره افراد را به طور یکپارچه بر روی یک فرد دیگر در یک ویدئو قرار دهند. اصطلاحی که برای توصیف این فناوری استفاده می شود "جعل عمیق[۱]" است. جعل عمیق از ترکیب دو واژه "محتوای جعلی[۲]" و "یادگیری عمیق[۳]" ساخته شده است و تکنیکی برای دست‌کاری تصاویر (عمدتاً تصاویر افراد) با استفاده از شبکه‌های یادگیری عمیق است. ایجاد جلوه‌های بصری جعل عمیق از طریق دادن اطلاعات تصویری به یک شبکه یادگیری عمیق مولد انجام می‌شود. اطلاعات مورد نیاز برای ایجاد تصاویر جعل عمیق شامل تصاویری هم از سوژه منبع و هم از سوژه جایگزین است که در شبکه یادگیری عمیق مربوطه پردازش می‌شوند. پس از پردازش و آموزش کافی، شبکه یادگیری عمیق قادر به تولید تصاویر و ویدئوهای جدیدی است که هرگز وجود نداشته است اما دارای چنان کیفیتی هستند که به سادگی قابل تشخیص از تصاویر و ویدئوهای واقعی نیستند.

هرچند افزودن تصاویر جدید به ویدئو و فیلتر کردن آن مفاهیم جدیدی نیستند، اما کیفیت ویدئوهای جعلی ساخته شده با این تکنیک کاملاً از موارد قبلی متمایز است. این فناوری یک وجه تمایز اصلی دیگر نیز دارد و آن اینکه تولید ویدئوهای جعلی نیاز به تخصص بالایی ندارد و فقط با کمک مجموعه ای از تصاویر انجام می‌شود و نیازی نیست که یک کاربر متخصص در تولید آنها دخالت مستقیم داشته باشد.

به این ترتیب توانایی تحریف واقعیت با فناوری "جعل عمیق" جهشی فزاینده داشته است. این فناوری امکان ایجاد صدا و تصویر از افراد واقعی را فراهم می‌آورد و قادر است فایلهای صوتی یا ویدئویی از کارهای افراد واقعی بسازد، بدون آنکه آنها چنان سخنانی گفته باشند یا آن کار را انجام داده باشند. از طرف دیگر تکنیک‌های یادگیری ماشین که در حال افزایش سطح پیچیدگی این فناوری هستند، باعث می‌شوند رسانه‌های جعل عمیق واقعی‌تر بنظر برسند و به طور فزاینده‌ای در برابر شناسایی مقاوم باشند. فناوری جعل عمیق از ویژگی‌هایی برخوردار است که انتشار سریع و گسترده آن را، در بین کاربران متخصص و غیرمتخصص ممکن می‌سازد.

---

[۱] Deep Fake

[۲] Fake

[۳] Deep Learning

در حالی که فناوری جعل عمیق مزایای خاص خود را دارد، اما آسیب‌های زیادی را نیز می‌تواند به همراه داشته باشد. فیلم‌های ساختگی با کیفیت بالا می‌توانند باعث هرج و مرج شوند و لطمه‌های اقتصادی و عاطفی به شهرت اشخاص وارد کند. فیلم‌هایی که در قالب تبلیغات سایبری علیه سیاستمداران ایجاد می‌شوند، می‌توانند برای دولت یک کشور فاجعه‌بار باشند و در مقیاسی بزرگتر، این امکان وجود دارد که حکومتهای مردمی را تضعیف کند.

اولین قدم در راه مقابله با این فناوری شناخت آن است و هدف ما در این گزارش معرفی این فناوری و بررسی روشهای ساخت و آشکارسازی ویدئوهای جعل عمیق و ارائه راهکارهایی برای مقابله با تهدیدات آن است.

۱- معرفی فناوری جعل عمیق

قدمت جعل تصاویر تقریباً به اندازه قدمت صنعت عکاسی است. برای مثال اطلاعاتی از ساخت تصاویر جعلی از ابراهام لینکلن برای مخدوش کردن چهره وی در مبارزات انتخاباتی وجود دارد. با ظهور تصویربرداری دیجیتال و ابزار ویرایش آنها نظیر Adobe Photoshop چنان تصاویر جعلی ساخته شد که کاربران عادی در تشخیص اصالت آنها ناتوان بودند. در زمینه دست‌کاری ویدئو برنامه Video Rewrite که در سال ۱۹۹۷م. منتشر شد را می‌توان پیشگام دانست. این برنامه فیلم‌های ویدئویی موجود از شخصی که در حال صحبت کردن بود را جهت ادای سخنانی که در یک اثر صوتی دیگر وجود داشت، تغییر می‌داد. این اولین سیستمی بود که به طور اتوماتیک تغییر حالت چهره و بویژه دهان گوینده را برای بیان کلمات جدید انجام داد. تحقیقات کنونی در این حوزه بر ایجاد فیلم‌های واقعی‌تر و استفاده از تکنیک‌های ساده‌تر، سریعتر و سهل‌الوصول‌تر متمرکز شده‌اند. برنامه "ویدئوی ساختگی اوباما"، که در سال ۲۰۱۷م. منتشر شد، فیلم‌های ویدئویی رئیس جمهور پیشین باراک اوباما را تغییر می‌دهد تا وی را در حال ادای سخنان موجود در یک فایل صوتی جداگانه نشان دهد. این پروژه به عنوان یکی از اصلی‌ترین پژوهش‌های انجام شده، برای سنتز حالات دهان بر اساس صدا است. در زمینه جایگزینی چهره یک فرد با فرد دیگر در ویدئو می‌توان از برنامه Face2Face، که در سال ۲۰۱۶م. منتشر شده، نام برد. این برنامه جایگزینی فرد در ویدئو را به صورت بلادرنگ انجام می‌دهد. با توجه به اینکه این برنامه اولین برنامه برای جایگزینی چهره به صورت بلادرنگ با استفاده از دوربین‌هایی است که اطلاعات عمق را ضبط نمی‌کنند، بنابراین امکان جایگزینی چهره در فیلم‌های گرفته شده توسط دوربین‌های معمولی را فراهم می‌کند.

دلیل اهمیت جعل عمیق تنها کیفیت تصاویر بدست آمده نیست، بلکه سهولت ایجاد آنها نیز هست. در گذشته، ویرایش یا ساخت فیلم‌ها یک کار پرهزینه بود که به مقدار زیادی نیروی انسانی، زمان و پول نیاز داشت. اما امروز برای ایجاد تصاویر با استفاده از جعل عمیق تنها چیزی که لازم است یک لپ‌تاپ، اتصال به اینترنت و دانش ابتدایی از شبکه‌های یادگیری عمیق است. برنامه‌هایی مانند FakeApp وجود دارند که تنها با یک کلیک می‌توانند چهره یک فرد را در یک فیلم جایگزین کنند. در نتیجه ساخت فیلم‌های جعل عمیق به حدی آسان شده‌اند که هر کسی با دانش

اندک می‌تواند چنین فیلم‌هایی را تولید کند. انتظار می‌رود که در آینده با پیشرفت در یادگیری عمیق، بتوان فیلم های جعلی که فوق‌العاده واقعی به نظر می‌رسند را تولید کرد.

فناوری جعل عمیق به دو صورت می‌تواند در دستکاری داده‌های بصری استفاده شود، نخست انتقال بافت یا حالت یک تصویر به تصویر دیگر و دیگری تغییر حالت صورت یا جایگزینی چهره یک شخص است. نمونه‌ای از انتقال بافت یا حالت یک تصویر به تصویر دیگر، مانند تبدیل تصویر اسب به گورخر و یا تصویر تابستان به زمستان، در شکل ۱ نشان داده شده است. همچنین پیشرفت‌های چشمگیری در تغییر حالت صورت یا جایگزینی چهره یک شخص صورت گرفته است. شکل ۲ نمونه‌ای را نشان می‌دهد که ویژگی‌های مختلف صورت مانند مو، جنسیت، سن و رنگ پوست و حالاتی مانند عصبانیت، خوشبختی و ترس در چهره افراد ایجاد می‌شود.

هر قدر تصاویر منبع و جایگزین شباهت بیشتری به یکدیگر داشته باشند، نتیجه کار بهتر و کیفیت تصاویر ساخته شده بالاتر خواهد بود. به طور مثال، در مورد جایگزینی تصویر یک فرد با فرد دیگر، اگر فرد جایگزین دارای ریش باشد، بهتر است فرد مورد نظر در فیلم منبع نیز دارای ریش باشد.

شکل ۱: نمونه های از انتقال بافت یا حالت یک تصویر به تصویر دیگر.

(الف)

(ب)

شکل ۲: ایجاد (الف) ویژگی های مختلف صورت مانند مو ، جنسیت ، سن و (ب) حالاتی مانند عصبانیت ، شادی و ترس در چهره افراد

۲- مروری بر روش تولید رسانه‌های جعل عمیق

اولین روش استفاده شده در جعل عمیق مبتنی بر خودکدگذارها[۱] بوده است. تصاویر ایجاد شده به این روش از کیفیت بالایی برخوردار نبودند. برای برطرف نمودن اشکالات این روش و همچنین گسترش تنوع داده‌های تولیدی، روش‌های دیگری نیز برای ایجاد داده‌های جعل عمیق پیشنهاد گردید که در ادامه این روش‌ها بررسی می‌شوند.

۲-۱- خود کدگذارها

ایده اصلی در آموزش موازی دو خودکدگذار است. معماری آنها می‌تواند بسته به اندازه خروجی، زمان آموزش مورد نظر، کیفیت مورد انتظار و منابع موجود متفاوت باشد. یک خودکدگذار از یک شبکه کدگذار[۲] و یک شبکه کدگشایی[۳] تشکیل می‌شود. هدف از کدگذار کاهش ابعاد داده با کدگذاری آن است و این کار با عبور داده‌ها از لایه ورودی که با کاهش تعداد متغیرها همراه است انجام می‌شود. کدگشا از این کد ایجاد شده برای تقریب ورودی اصلی استفاده می‌کند. مرحله بهینه‌سازی با مقایسه ورودی و خروجی تقریبی حاصل از کدگشا و جریمه کردن تفاوت بین این دو، به طور معمول با استفاده از معیار فاصله اقلیدسی، انجام می‌شود.

نخستین گام در تولید تصاویر جعل عمیق با استفاده از خودکدگذار، جمع‌آوری چهره‌های هم تراز از دو فرد مختلف A و B است، سپس آموزش خودکدگذار EA برای بازسازی چهره A از مجموعه داده‌های تصاویر صورت A و خودکدگذار EB برای بازسازی چهره‌های B از مجموعه داده‌های صورت‌های چهره B استفاده می‌نماید. بخشی که این عمل را متفاوت از یک خودکدگذار معمولی می‌کند به‌اشتراک‌گذاری وزن قسمت کدگذار دو خودکدگذار EA و EB است، اما بخش کدگشایی بصورت مستقل انجام می‌شود. پس از بهینه سازی، هر تصویری که حاوی چهره A باشد می‌تواند از طریق این کدگذار مشترک کدگذاری شود اما با گدگشای EB کدگشایی می‌شود. این عمل در شکل ۳ نشان داده شده است.

---

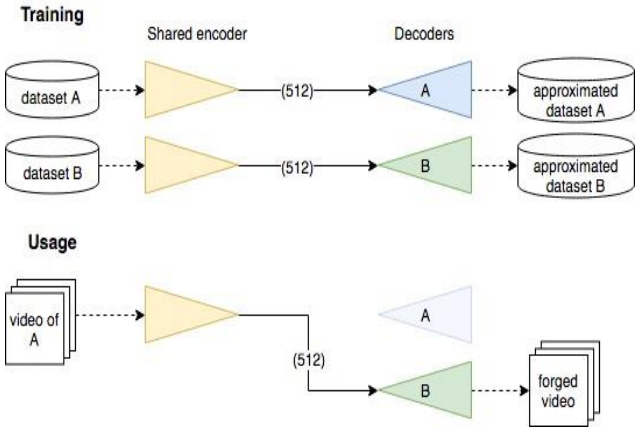
[۱] autoencoders

[۲] encoder

[۳] decoder

عملکرد این روش را می‌توان به این ترتیب شرح داد که از کدگذار مشترک برای کدگذاری اطلاعات عمومی تصویر از قبیل نور، موقعیت و حالت صورت استفاده می‌شود و کدگشای اختصاصی برای بازسازی حالات و خصوصیات ثابت مشخصه هر فرد و جزئیات چهره او استفاده می‌شود. به این ترتیب می‌توان اطلاعات پس‌زمینه را از اطلاعات مورفولوژیکی جدا کرد.

در عمل، نتیجه کار چشمگیر است و بیانگر علت رایج بودن این تکنیک است. آخرین مرحله گرفتن فیلم هدف، استخراج و تراز کردن چهره مورد نظر در هر فریم و استفاده خودکدگذار اصلاح شده برای تولید چهره مورد نظر در هر فریم ویدئو با همان نور و حالت چهره اصلی است و سپس فریم‌ها را باید دوباره در ویدیو ادغام کرد.



شکل ۳: جعل عمیق با استفاده از خودکدگذاردر بالا قسمت های آموزشی با کدگذار مشترک *(EA,EB)* با رنگ زرد مشخص شده اند و پایین قسمت تولید تصاویر جعلی که در آن تصاویر *A* باکدگشای *B* کدگشایی می‌شوند.

۲-۲- شبکه مولد متخاصم[۱] (GAN)

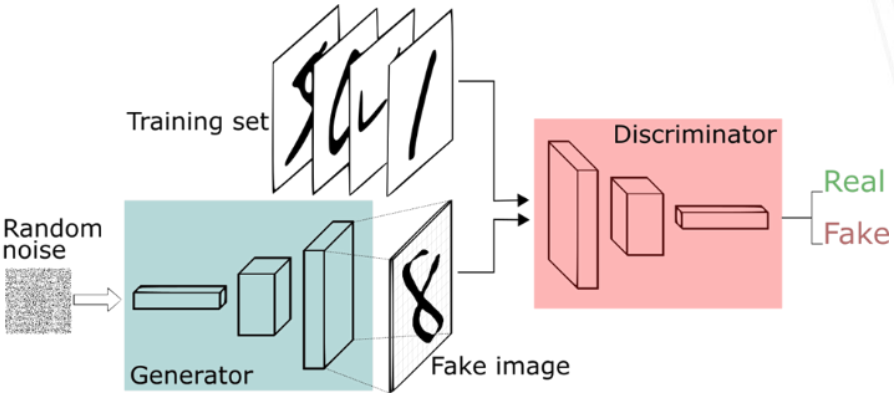
GANها در سال ۲۰۱۴ توسط Ian Goodfellow و سایر محققان دانشگاه مونترال در مقاله‌ای معرفی شدند. پتانسیل GANها بسیار زیاد است، و می‌توان به GANها آموزش داد که در هر زمینه‌ای نظیر تصاویر، موسیقی، گفتار

[۱] Generative Adversial Network

و نشر محصولاتی را شبیه به خود انسان‌ها ایجاد کنند. GAN از دو نوع شبکه عصبی تشکیل شده است: مولد و تشخیص‌دهنده. ساختار کلی GAN در شکل ۴ نشان داده شده است.

۲-۲-۱- مولد[۱]

کار مولد این است که از نویز یک تصویر جعلی بسازد (مثلاً تصویری از نگارش عدد هشت لاتین). مولد، برداری کوچک از اعداد تصادفی حقیقی می‌گیرد و با کانولوشن وارون آن را به یک ماتریس با اندازه تصویر مورد نظر تبدیل می‌کند. در کانولوشن وارون به جای نمونه‌کاهی، نمونه‌افزایی صورت می‌گیرد. با آموزش، مولد یاد می‌گیرد تصویری بسازد که تا حد امکان کپی دقیقی از تصویر مورد نظر باشد.



شکل ۴: ساختار کلی GAN

۲-۲-۲- تشخیص‌دهنده[۲]

وظیفه تشخیص‌دهنده این است که تصویری بگیرد و تعیین کند که واقعی یا جعلی است (یعنی بین تصاویر جعلی و واقعی تمایز قائل شود). نحوه کار آن بسیار ساده است و در اصل شامل استفاده از یک دسته کامل از لایه‌های کانولوشن استاندارد است که تصویر را نمونه‌کاهی می‌کنند تا در نهایت به یک عدد برسند. این عدد معرف خطا است. همانگونه که در شکل ۴ مشاهده می‌شود در حین نمونه‌کاهی تصویر، تعداد فیلترها را افزایش می‌دهیم. از مقدار خطای تولید شده برای تعیین واقعی یا جعلی بودن تصویر تولید شده، استفاده می‌شود.

---

[۱] Generator

[۲] Discriminator

۲-۲-۳- آموزش مدل

پس از معرفی مدل‌های مولد و تشخیص‌دهنده، به آموزش آنها می‌پردازیم. از آنجا که تشخیص‌دهنده تاثیر زیادی بر عملکرد درست GAN دارد، سرعت بروزرسانی ضرایب آن از مولد بیشتر در نظر گرفته می‌شود و در عمل به ازای هر بار که وزن‌های مولد به‌روز می‌شوند، وزن‌های تشخیص‌دهنده را بین ۱ تا ۵۰۰ بار به روز می‌کنیم. نحوه آموزش تشخیص‌دهنده به این ترتیب است که اول یک مجموعه کوچک از تصاویر واقعی خود را گرفته و آنها را به تشخیص دهنده می‌دهیم. خروجی تشخیص‌دهنده نشان‌دهنده خطا در تشخیص تصاویر واقعی است. سپس یک بردار نویز را در ورودی مولد می‌گذاریم و تصویر جعلی ایجاد شده را در ورودی تشخیص دهنده قرار می‌دهیم. خروجی تشخیص دهنده در این حالت خطا در تشخیص تصاویر جعلی است. خطای واقعی از کم کردن این دو خطا بدست میآید و وزن ها با توجه به این خطا به‌روز می‌شوند. همانطور که می‌بینید، مولد در تلاش است تا تشخیص دهنده را فریب دهد، و این وجه تسمیه کلمه متخاصم در شبکه‌های مولد متخاصم است. هر کدام از دو شبکه در تلاش هستند تا یک تابع هدف متفاوت یا متضاد را در یک بازی با الگوی بازیگر منتقد بهینه کنند. همانطور که تشخیص دهنده رفتار خود را تغییر می دهد ، مولد هم خود را اصلاح می‌کند و برعکس. به نحوی که کاهش خطای یکی باعث افزایش خطای دیگری می‌شود که در تئوری بازی‌ها به آن یک بازی مجموع صفر[۱] می‌گویند.

GANها نیز دارای مشکلاتی هستند. نخست آنکه از نظر آموزش ناپایدار هستند زیرا باید دو شبکه را تنها با استفاده از تنها یک پس انتشار خطا آموزش داد. بنابراین انتخاب تابع هدف مناسب می‌تواند تفاوت بزرگی در آموزش شبکه ایجاد کند. از طرفی نمی‌توان از GAN برای ایجاد هر الگوی دلخواه در داده ورودی استفاده نمود. در نهایت در سال ۲۰۱۹ ، DeepMind نشان داد که خودکدگذارهای متغیر می‌توانند از GANها در تولید چهره بهتر عمل کنند.

[۱] Zero-sum game

۲-۳- خودکدگذارهای متغیر[۱] (VAE)

شبکه‌های خودکدگذار به طور ذاتی مدل‌های مولد نیستند اما در قسمت خودکدگذار دیدیم که چگونه از دو خودکدگذار برای ایجاد داده جدید استفاده شد. در خودکدگذارهای متغیر هدف ایجاد داده جدید با استفاده از یک خودکدگذار است.

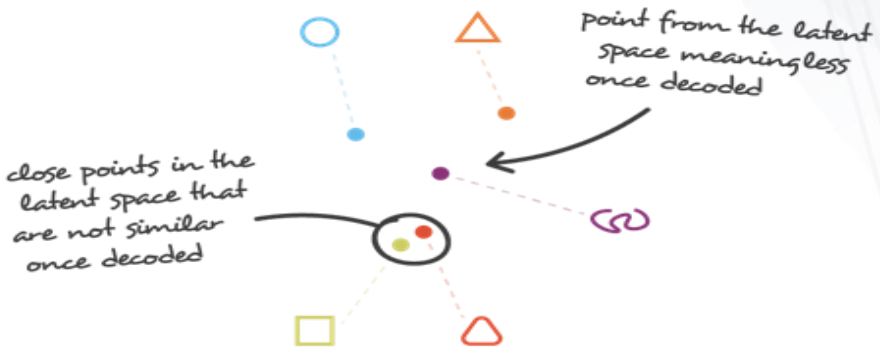
همانگونه که گفته شد یک خودکدگذار از یک کدگذار که یک کد کوچک بر اساس تصویر ورودی تولید می‌کند و یک کدگشا که کد ایجاد شده را به یک تصویر مشابه تصویر ورودی تبدیل می‌کند تشکیل شده است. کد تولید شده در لایه میانی یک شبکه خودکدگذار قرار دارد و دامنه کدهای تولید شده را به همین دلیل فضای نهان می‌نامند. ایده اصلی در شبکه‌های خودکدگذار متغیر این است که باید با دادن مقداری غیر از کد ایجاد شده برای تصاویر ورودی بتوان تصویری جدید ایجاد کرد و به این ترتیب شبکه خودکدگذار را می‌توان به یک شبکه مولد تبدیل نمود.

دادن یک کد جدید به کدگشای یک شبکه خودکدگذار معمولی الزاماً تصویر معنی‌داری را تولید نمی‌کند. زیرا فضای نهان در این شبکه‌ها از نظم کافی برخوردار نیست. در واقع کیفیت و ارتباط داده‌های تولیدی به منظم بودن فضای نهفته بستگی دارد. منظم نمودن فضای نهفته در خودکدگذارها امری دشوار است که وابسته به توزیع داده‌ها در فضای اولیه، ابعاد فضای نهفته و معماری کدگذار است. بنابراین، در یک خودکدگذار معمولی اینکه کدگذار فضای نهفته را به روشی هوشمندانه و سازگار با فرآیند مولد بودن، سازمان داده باشد اگر غیرممکن نباشد، بسیار دشوار است. شکل ۵ نشان می‌دهد که چگونه دادن یک کد دلخواه از فضای نهفته به کدگشا می‌تواند منجر به ایجاد تصاویر بی‌معنی و یا غیر منتظره شود.

این عدم ساختاریافته بودن داده‌های فضای نهان بسیار طبیعی است. در واقع، ما هیچ تمهیدی برای منظم بودن فضای نهان در هنگام آموزش خودکدگذار در نظر نگرفتیم. در آموزش خودکدگذار، کدگذاری و کدگشایی به نحوی انجام می‌شود که تابع هزینه به کمترین مقدار برسد و مهم نیست که چطور فضای نهان سازمان‌دهی شود. بنابراین، اگر در هنگام تعریف این مدل دقت لازم را صرف نکنیم، طبیعی است که شبکه در حین آموزش از هرگونه امکانات

[۱] Variational Autoencoder

اضافی برای دستیابی به هدف خود می تواند استفاده کند، مگر اینکه برای منظم نمودن فضای نهان به طور صریح تمهیداتی را در تنظیم کننده تابع هزینه در نظر بگیریم. یک خودکدگذار متغیر را می‌توان به عنوان خودکدگذاری در نظر گرفت که آموزش آن به منظم شدن فضای نهان منجر می‌شود به نحوی که فضای نهفته از خواص مناسب برای تولید اطلاعات معنی دار برخوردار است.



شکل ۵: فضای نهان نامنظم استفاده از خودکدگذار را برای تولید محتوای جدید ناممکن میسازد.

معماری یک خودکدگذار متغیر مشابه خودکدگذار استاندارد است و از یک کدگذار و کدگشا استفاده می‌کند که برای به حداقل رساندن خطای بازسازی بین داده‌های کدگشایی شده و داده‌های اولیه آموزش داده می‌شود. ولی در عین حال برای منظم سازی فضای نهفته، یک تغییر جزئی در فرآیند کدگذاری/کدگشایی می‌دهد: به جای کدگذاری یک ورودی به صورت یک نقطه واحد، آن را به صورت یک توزیع احتمالی در فضای نهفته کدگذاری می‌کند.

در عمل، توزیع‌های کدگذاری شده به صورت توزیع نرمال انتخاب می‌شوند تا آموزش کدگذار تنها بر اساس برگرداندن ماتریس کواریانس و میانگین آنها صورت گیرد. دلیل کدگذاری ورودی به صورت توزیع نرمال به جای یک نقطه واحد این است که این عمل امکان منظم‌سازی فضای نهان را فراهم می‌کند، به این ترتیب که توزیع‌های ایجادشده توسط کدگذار در فرایند آموزش مجبور می‌شوند به یک توزیع نرمال استاندارد نزدیک شوند. به این طریق، هم به صورت موضعی و هم به صورت فراگیر از منظم بودن فضای نهفته (موضعی به دلیل کنترل واریانس و فراگیر به دلیل کنترل میانگین) اطمینان خواهیم یافت.

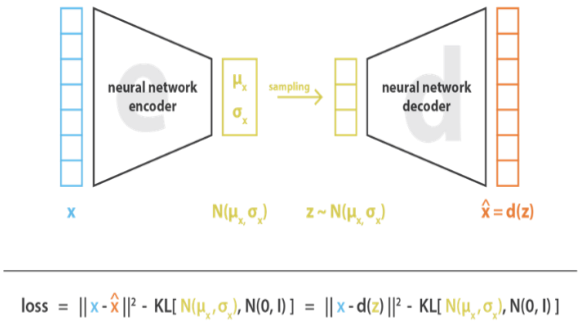
بنابراین، تابع خطا که در هنگام آموزش خودکدگذار متغیر به حداقل می رسد از یک "عبارت بازسازی" (در لایه نهایی) تشکیل شده که تمایل دارد تا دقت کدگذاری/کدگشایی را تا حد ممکن بالا ببرد و یک "عبارت تنظیم کننده" (روی لایه نهفته) دارد که می‌خواهد فضای نهفته را با نزدیک کردن توزیع‌های ایجاد شده توسط کدگذار به یک توزیع نرمال استاندارد، تا حد امکان منظم کند. این عبارت تنظیم کننده به صورت واگرایی کولباک‌لایبر بین توزیع ایجاد شده توسط کدگذار و یک توزیع نرمال استاندارد بیان شده است. واگرایی کولباک‌لایبر[1] (که آنتروپی نسبی نیز نامیده می شود) معیاری از تفاوت نحوه توزیع در یک توزیع احتمالی با یک توزیع احتمالاتی دوم است و برای دو توزیع پیوسته P و Q به شکل زیر تعریف می‌شود:

$$D_{\mathrm{KL}}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx \qquad (۱)$$

نظمی که فضای نهان باید داشته باشد تا بتواند فرآیند تولید را امکان‌پذیر سازد از طریق دو ویژگی اصلی بیان می‌شود:

- پیوستگی: دو نقطه نزدیک در فضای نهان اگر کدگشایی شوند نباید دو تصویر کاملاً متفاوت بدهند.
- کامل بودن: کدگشایی هر نقطه ای که از یک توزیع از فضای نهان انتخاب شده است، باید محتوای "معنی داری" را ایجاد کند.

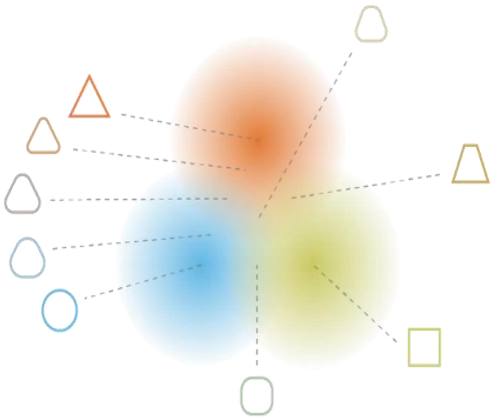
به منظور ایجاد این نظم در فضای نهان خودکدگذار متغیر، تابع خطا علاوه بر یک عبارت بازسازی برای بهبود کارآیی کدگذاری/کدگشایی، دارای یک عبارت تنظیم نیز هست که باعث می‌شود فضای نهان منظم شود (شکل ۶).



شکل ۶: ساختار خودکدگذار متغیر و نحوه تعریف تابع خطا در آن

[1] Kulback-Leibler divergence

اضافه نمودن عبارت تنظیم به تابع خطا باعث دور شدن حداکثری داده های کدشده در فضای نهان از یکدیگر می‌شود. با اینحال، توزیع آنها حداکثر همپوشانی را خواهند داشت. به این طریق، شرایط پیوستگی و کامل بودن برآورده می‌گردد. پیوستگی و کامل بودن به دست آمده با اضافه نمودن عبارت تنظیم باعث ایجاد تغییرات تدریجی (گرادیان تغییرات) برای اطلاعات کدگذاری شده در فضای نهان می‌شود، به این معنی که کدگشایی یک نقطه از فضای نهان که در وسط دو توزیع کدگذاری شده از داده‌های آموزش مختلف قرار داشته باشد، باید به تصویری بین تصویر مولد توزیع اول و تصویر مولد توزیع دوم منجر شود (شکل ۷).



شکل ۷: نمایشی از ایجاد "گرادیان" برای اطلاعات کدگذاری شده در فضای نهان که با اضافه کردن عبارت تنظیم حاصل شده

وقتی خودکدگذار متغیر را آموزش دادیم به طرق مختلف می‌توان از آن برای ایجاد تصاویر جعل عمیق استفاده کرد. برای اینکار تنها کافی است که از قوانین ترکیب بُردارها استفاده نمائیم. به طور مثال، برای ترکیب دو تصویر جهت ایجاد تصویری جدید تنها کافی است میانگین بردار نهان آن دو تصویر را بدست آوریم و توسط کدگشا آن را کدگشایی کنیم. و یا اگر بخواهیم ویژگی جدیدی در یک عکس اضافه کنیم، مثلاً برای اینکه تصویرفردی بدون عینک را به تصویر فردی باعینک تبدیل نماییم، کافی است تفاوت بردار نهان بین دو تصویر از یک فرد با و بدون عینک را به بردار نهان شخص مورد نظر اضافه کنیم و بردار حاصل را کدگشایی نماییم. حتی میتوان یک خود کدگذار را به همراه یک شبکه یادگیری عمیق ترتیبی مانند LSTM[۱] آموزش داد و از آن برای تولید داده‌های ترتیبی منفصل استفاده نمود، کاری که با استفاده از GAN امکان پذیر نیست. به این ترتیب می‌توان متن‌ها و حتی موسیقی‌های

[۱] Long Short Term Memory

مصنوعی تولید نمود. قابلیت کار VAE ها با انواع مختلف داده از قبیل داده‌های ترتیبی یا غیرترتیبی، پیوسته یا منفصل، و حتی با و بدون برچسب ، آنها را به مولدهای بسیار قدرتمندی بدل ساخته است.

۳- مروری بر روش‌های اصلی تشخیص رسانه‌های جعل عمیق

از طیف گسترده دستکاری‌ها در داده‌های مختلف، اخیراً دستکاری در تصاویر چهره به عنوان روشی برای انتشار اطلاعات نادرست یا حتی برای افتراق افراد مشهور بسیار مورد توجه قرار گرفته است. جای تعجب ندارد که جعل چهره افراد نسبت به سایر اشیاء ارجحیت دارد، زیرا از چهره‌ها در جامعه به عنوان ابزاری برای ارتباطات انسانی و انتقال اطلاعات محرمانه بر اساس هویت استفاده می‌شود. از این رو تمرکز اصلی این گزارش بر تکنیک‌های تشخیص جعل صورت است.

روش‌های تشخیص جعل تصویر و فیلم را می‌توان به دو گروه فعال و منفعل طبقه‌بندی کرد. در روش‌های فعال، داده‌های دیجیتالی تحت پیش پردازش‌هایی مانند نهان‌نگاری و یا امضاهای دیجیتالی قرار می‌گیرند که باعث می‌شود در زمان ایجاد محتوا، کیفیت محتوای دیجیتال تا حدی کاهش یابد. اگر محتوای دیجیتالی جعل شود، بازیابی علامت نهان‌نگاری یا امضا امکان پذیر نیست و این امر منجر به تشخیص دستکاری می‌شود. رویکردهای فعال را نمی‌توان در سناریوهایی که هیچ چیزی از قبل در داده دیجیتال درج نشده، استفاده کرد.

رویکرد منفعل بر اساس این فرض عمل می‌کند که هر تصویر یا فیلم حاوی ویژگی‌های طبیعی یا اثر انگشت ذاتی، بر اساس خصوصیات دستگاه تصویربرداری و یا ویژگی‌های منحصر به فرد سوژه است. اگر محتوای دیجیتالی جعلی نباشد، همبستگی‌های آماری زمینه‌ای تصویر یا فیلم داده شده پس از برخی عملیات پس‌پردازش ثابت می‌ماند. شناسایی تفاوت در همبستگی‌های آماری محتوای دیجیتالی وظیفه اصلی روش‌های مختلف تشخیص برای تعیین جعل در محتوای دیجیتال است که در بیشتر موارد می‌تواند به یک مسئله شناسایی الگو تبدیل شود. وقتی سخن از دست‌کاری یک ویدیو در میان باشد، در واقع آن را شامل دو قسمت اصلی و قسمت تغییریافته دانسته‌ایم. محققان روش‌ها و الگوریتم‌های مختلفی را برای شناسایی تصویر دیجیتالی یا دستکاری ویدیو بر اساس ویژگی قسمت اصلی و قسمت تغییریافته ارائه داده‌اند.

تشخیص جعل عمیق را می‌توان به دو گروه تشخیص تصاویر جعلی و تشخیص ویدئوی جعلی تقسیم کرد. در تشخیص تصویر جعلی فقط می‌توان از ویژگی‌های مکانی استفاده کرد، در حالی که در تشخیص ویدیوی جعلی هم از

ویژگی‌های مکانی (در یک قاب) و هم از ویژگی‌های زمانی (روابط بین قاب ها) برای تعیین اصالت فیلم استفاده می‌شود. در ادامه روش‌های منفعل برای تشخیص جعل عمیق در تصویر و فیلم مورد بحث قرار می‌گیرند.

۳-۱- تشخیص تصاویر جعلی

تعویض چهره دارای کاربردهای مفیدی نظیر ساخت ویدئوهای جعلی و تغییر حالت صورت در تصاویر پرتره و همچنین مخفی‌سازی هویت است زیرا می‌تواند چهره فرد را در یک تصویر با چهره افراد دیگر در مجموعه‌ای از تصاویر موجود جایگزین نماید. با این حال، تعویض چهره یکی از تکنیک‌هایی است که مهاجمان سایبری از آن برای نفوذ به سیستم‌های شناسایی یا تأیید هویت بر اساس چهره استفاده می‌کنند. استفاده از تکنیک‌های یادگیری عمیق مانند CNN[۱] و GAN باعث شده است که تشخیص تصاویر جایگزینی چهره چالش برانگیزتر باشد، زیرا این تکنیک‌ها می توانند باعث حفظ حالت چهره و نورپردازی عکس‌ها شوند. ژانگ و همکاران از روش انبان کلمات برای استخراج مجموعه‌ای از ویژگی‌ها استفاده کرده و آن را در طبقه بندهای مختلف مانند SVM[۲] ، جنگل تصادفی و پرسپترون های چند لایه (MLP[۳]) برای شناسایی چهره جایگزین شده استفاده کردند. در بین تصاویر ایجاد شده توسط یادگیری عمیق، تشخیص نمونه‌هایی که با استفاده از مدل‌های GAN سنتز شده‌اند، بسیار دشوار است. GAN توانایی یادگیری توزیع پیچیده داده‌های ورودی و تولید خروجی‌های جدید با توزیع مشابه ورودی را دارد و بنابراین تصاویر تولید شده بسیار مشابه واقعی و با کیفیت هستند.

همزمان با توسعه روش‌های شناسایی تصاویر جعلی ساخته شده با استفاده از GAN مدل‌های جدیدی مبتنی بر GAN ارائه می‌شود که ضعف مدل‌های قبلی را از بین می‌برد و شناسایی تصاویر جعلی ساخته شده توسط آنها با استفاده از روش‌های قبلی امکان پذیر نیست. برای مثال ژوان و همکاران در فرایند آموزش از یک مرحله پیش پردازش هم برای تصاویر واقعی و هم برای تصاویر جعلی (افزودن نویز گوسی و تار کردن گوسی تصویر) استفاده کرده‌اند تا

[۱] Convolutional Neural Networks
[۲] Support Vector Machine
[۳] Multilayer Perceptron

مشخصه‌های فرکانس بالای با دامنه کم را که در شناسایی تصاویر ساخته شده بوسیله GAN مورد استفاده قرار می گرفت، از بین ببرند.

از سوی دیگر ، آگراوال و وارشنی شناسایی تصاویر جعل عمیق مبتنی بر GAN را به عنوان یک مسئله که برای حل آن نیاز به یک معیار قابل اندازه‌گیری است، مطرح کردند و برای آن یک چارچوب آماری بر اساس مطالعات نظری مبتنی بر تئوری اطلاعات در فرایند تائید اصالت ارائه دادند. آنها حداقل فاصله بین توزیع آماری یک تصویر صحیح و تصویر ایجاد شده توسط یک GAN مشخص را به عنوان خطای اوراکل تعریف کردند. نتایج تحلیلی نشان می‌دهد که در صورت دقت کمتر GAN این فاصله افزایش می‌یابد و در این حالت تشخیص جعل عمیق آسان‌تر است. به عبارت دیگر برای ایجاد تصاویر جعلی با وضوح بالا، GANهایی با دقت بسیار بالا نیاز است تا تشخیص آنرا دشوار نماید.

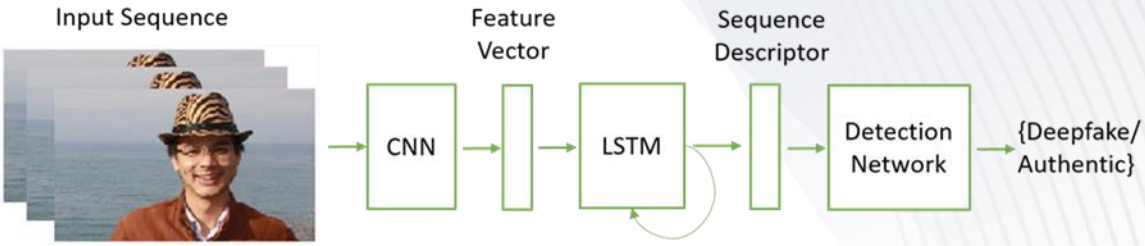
۳-۲- شناسایی ویدئوهای جعلی

به دلیل کاهش کیفیت زیاد فریم‌ها پس از فشرده‌سازی، اغلب روش‌های آشکارسازی جعل تصویر را نمی‌توان برای تشخیص ویدئوهای جعلی بکار برد. از طرف دیگر ویدئوها دارای ویژگی‌های زمانی نیز هستند که برآورده ساختن آنها در ویدئوهای جعلی امری چالش برانگیز است. در این زیربخش، شناسایی ویدئوهای جعلی بر اساس ویژگی‌های زمانی توضیح داده می‌شود. نخست به معرفی روش‌هایی پرداخته شده است که در حالت کلی برآورده ساختن ویژگی های زمانی در ویدئوهای جعلی را بررسی می‌کنند و سپس روشی که بر اساس طبیعی بودن سیگنال فیزیولوژیکی چشم زدن عمل می‌کند، تشریح می‌شود.

با ملاحظه اینکه روش‌های ساخت ویدئوی جعلی در پیاده‌سازی هماهنگی لازم زمانی در ویدئوها کارامد نیستند، گورا و دلپ از ترکیب شبکه‌های CNN و حافظه‌های کوتاه مدت طویل (LSTM[۱]) برای شناسایی ویدئوهای جعلی استفاده می‌کنند. CNN برای استخراج ویژگی‌های درون فریمی استفاده می‌شود که به LSTM فرستاده می‌شوند تا توصیفگر زمانی رشته فریم‌های ویدئو را تولید کند. در انتها یک شبکه کاملاً متصل برای جداسازی ویدئوهای اصلی

[۱] Long Short Term Memory

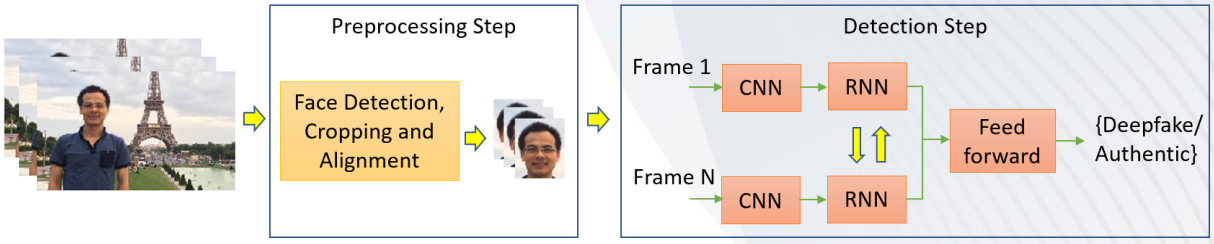
از ویدئوهای جعلی استفاده شده است تا توصیفگر دنباله را به عنوان ورودی گرفته و محاسبه احتمال تعلق دنباله ویدئو به کلاس معتبر یا جعلی را انجام دهد (شکل ۸).



شکل ۸: تشخیص جعل عمیق با استفاده از شبکه عصبی کانولوشن(CNN) و حافظه طویل کوتاه مدت (LSTM) برای استخراج ویژگی‌های زمانی

در کاری مشابه، سابیر و همکارانش از ویژگی‌های مکانی/زمانی برای تشخیص ویدئوهای جعلی استفاده نمودند. آنها مشاهده نمودند حین دستکاری ویدئوها بصورت قاب به قاب، اشکالات سطح پایین قاب‌ها هنگامیکه در توالی یک سری ویدئو بررسی می‌شوند، نمایان‌تر خواهند بود. از آنجا که در آشکارسازی اشکالات مکانی/زمانی معمولاً از ترکیب شبکه‌های کانولوشن به همراه سلول‌های بازگشتی استفاده می‌شود، آنها از یک فرآیند دو مرحله‌ای برای تشخیص جعل صورت استفاده نمودند. در مرحله پیش پردازش هدف شناسایی، جدا کردن و تراز کردن چهره‌ها از روی یک دنباله فریم بود در مرحله دوم از یک شبکه کانولوشنال بازگشتی (RCN[۱]) برای تشخیص هماهنگی‌های زمانی بین فریم‌ها استفاده نمودند که در آن از یک شبکه کانولوشنال Densenet به همراه سلولهای بازگشتی استفاده شده بود (شکل ۹). تفاوت این کار با روش قبلی در اینست که گورا و دلپ از شبکه‌های کانولوشن از پیش‌آموزش‌یافته استفاده کرده بودند، در حالیکه در اینجا کل مدل بصورت انتها به انتها آموزش داده شده بود و بنابراین نتایج بهتری به دست آمده بود. این روش بر روی مجموعه داده FaceForensics++ که شامل ۱۰۰۰ ویدئو است، آزمایش شد و نتایج امیدوارکننده‌ای داشت.

[۱] Recurrent Neural Networks

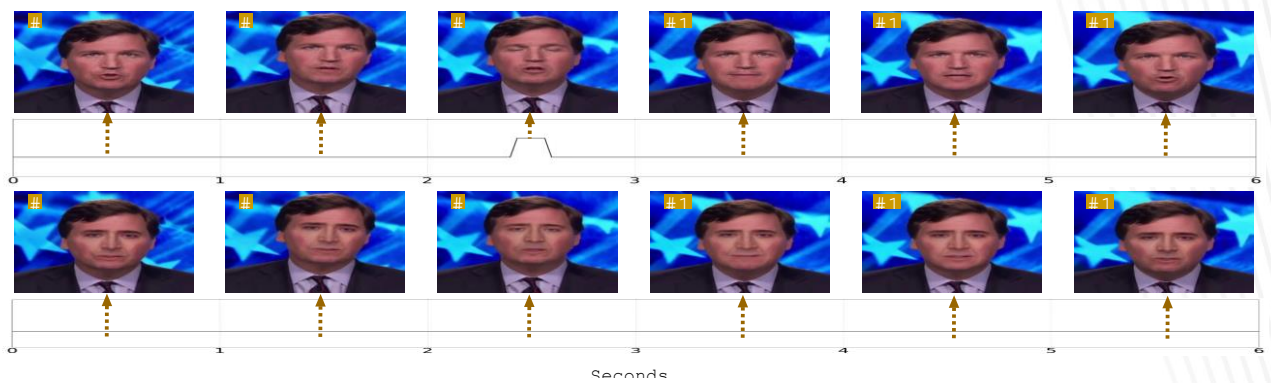


شکل ۹: یک فرآیند دو مرحله‌ای برای تشخیص جعل صورت

از سوی دیگر، استفاده از سیگنال فیزیولوژیکی، پلک زدن، برای تشخیص جعل عمیق در بر اساس این مشاهدات که پلک زدن یک فرد در ویدئوهای جعل عمیق بسیار کمتر از فیلم‌های واقعیست، پیشنهاد شده است. یک انسان بالغ سالم به طور معمول در فاصله زمانی بین ۲ تا ۱۰ ثانیه پلک برهم می‌زند، و هر پلک زدن ۰٫۱ و ۰٫۴ ثانیه طول می‌کشد. اما در ساخت ویدئوهای جعل عمیق غالباً از تصاویر چهره‌ای که به صورت آنلاین در دسترس هستند برای آموزش استفاده می‌کنند، که معمولاً افراد در آنها با چشمان باز نشان داده می‌شوند چون تعداد بسیار کمی تصاویر منتشر شده در اینترنت افراد را با چشمان بسته نشان می‌دهد. بنابراین، بدون دسترسی به تصاویر در حال پلک زدن افراد، الگوریتم‌های جعل عمیق قابلیت تولید چهره‌های جعلی که بتوانند به طور عادی پلک بزنند را ندارند. به عبارت دیگر، نرخ پلک زدن در ویدئوهای جعل عمیق بسیار کمتر از فیلم‌های عادی است. برای تشخیص فیلم‌های واقعی از جعلی، لی و همکاران ابتدا فیلم‌ها را به قاب‌هایی که قسمت‌های صورت و سپس مناطق چشم بر اساس شش علامت چشم استخراج می‌شود، تجزیه می‌کنند. پس از چند مرحله پیش پردازش مانند هم راستا کردن چهره‌ها، استخراج و مقیاس کردن کادرهای محدودکننده نقاط علامت چشم از آنها برای ایجاد توالی‌های جدید از تصاویر استفاده می‌شود. این توالی از ناحیه‌های برش خورده به شبکه‌های بازگشتی کانولوشنال طولانی مدت (LRCN[1]) داده می شوند تا ویژگی‌های پویای ویدئو را پیش‌بینی نماید. LRCN از یک استخراج کننده ویژگی مکانی مبتنی بر CNN، و یک یادگیرنده دنباله مبتنی بر حافظه کوتاه مدت بلند (LSTM) و یک پیش‌بینی حالت بر اساس یک لایه کاملاً متصل برای پیش‌بینی احتمال حالت باز و بسته شدن چشم تشکیل شده است. در پلک برهم زدن‌های یک فرد همبستگی

[1] Long Recurent Neural Networks

زمانی قوی وجود دارد و LSTM برای شناسایی این الگوهای زمانی کارایی زیادی دارد. نرخ پلک زدن بر اساس نتایج ارزیابی ویدئو محاسبه می‌شود که در آن پلک زدن با بسته شدن چشم با مقدار آستانه بالاتر از ۰,۵ به مدت زمان کمتر از ۷ فریم تعریف شده است (شکل ۱۰). این روش بر روی مجموعه داده‌های جمع‌آوری شده از وب متشکل از ۴۹ فیلم مصاحبه و ارائه و فیلم‌های جعلی متناظر آنها که توسط الگوریتم‌های جعل عمیق تولید شده‌اند، ارزیابی می شود. نتایج تجربی حاکی از عملکرد امیدوارکننده روش پیشنهادی در تشخیص فیلم‌های جعلی بوده است. این عملکرد حتی می‌تواند با در نظر گرفتن الگویی پویا برای پلک زدن بهبود یابد، به عنوان مثال پلک زدن بسیار مکرر نیز ممکن است نشانه‌ای از دستکاری باشد.



شکل ۱۰: نمونه ای از تشخیص چشم برهم زدن روی یک فیلم اصلی (بالا) و یک فیلم جعل عمیق (پایین). در حالت اول، یک پلک زدن را می توان در هر ۶ ثانیه تشخیص داد، در حالی که در حالت دوم چنین نیست و از نظر فیزیولوژیکی غیر طبیعی است.

۴- بررسی نرم‌افزارها و ابزارهای آماده موجود

در دسامبر سال ۲۰۱۷ یکی از کاربران وبسایت خبری/اجتماعی Reddit با نام کاربری DeepFakes با انتشار یک ویدیوی جعلی که در ساخت آن از الگوریتم‌های یادگیری عمیق استفاده شده بود، موضوع جدیدی تحت عنوان جعل عمیق را به دنیای فناوری معرفی نمود، که نام آن برگرفته از نام کاربری همان شخصی است که برای اولین بار از این ابزار جهت ساخت یک ویدیو جعلی استفاده نمود. ابزار استفاده شده برای ساخت این ویدئو FaceSwap نام داشت. از آن پس متخصصین زیادی در حوزه یادگیری ماشین و شبکه‌های عصبی در حوزه جعل عمیق به فعالیت پرداختند که یکی از نتایج آن ارائه ابزارهای متعدد در این زمینه است. DFaker و FaceSwapGan دو ابزاری بودند که با فاصله زمانی اندکی، به ترتیب در اواسط و اواخر ژانویه ۲۰۱۸ ارائه شدند که نتایج قابل قبولی ارائه دادند. همچنین ابزار دیگری تحت عنوان DeepFaceLab، که اندکی بعد در ژوئن ۲۰۱۸ ارائه شد، نیز نظر کاربران زیادی را به سوی خود جلب نمود.

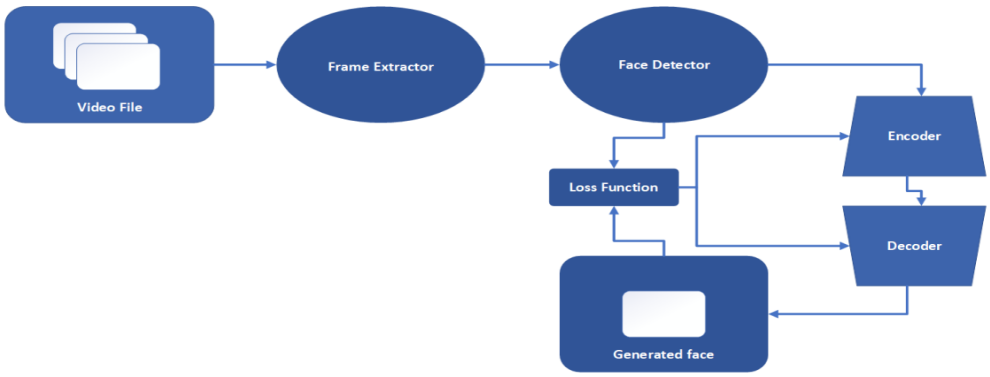
FaceSwap -۱-۴

FaceSwap اولین ابزار ایجاد تصویر و ویدیو جعلی است که توسط deepfakes و با استفاده از کتابخانه TensorFlow توسعه یافت. اولین ویدیوی جعلی با استفاده از این ابزار ایجاد شده و در اواخر سال ۲۰۱۷ میلادی منتشر شد. بخش اصلی FaceSwap را شبکه‌های عصبی عمیق خودکدگذار تشکیل می‌دهند. نکته قابل توجه در رابطه با این ابزار تابع هزینه آن است که وظیفه کنترل بهبود کارایی شبکه در ازای تکرارهای متوالی را داراست. در FaceSwap از یک شبکه عصبی کدگذار بصورت مشترک برای دو تصویر ورودی و دو شبکه کدگشا که هر یک بازسازی یکی از تصاویر را برعهده دارند استفاده شده است. پس از آنکه ویژگی‌های استخراج شده توسط شبکه کدگذار از تصویر نمونه ورودی، وارد شبکه بازسازی تصویر می‌شوند، این شبکه به کمک لایه‌های کانوولوشنی و تکنیک نمونه افزایی[۱] نسبت به بازسازی تصویر اولیه اقدام می‌کند. پس از بازسازی تصویر نیاز است کارایی شبکه با مقایسه تصویر

[۱] Up-Sampling

حاصل و تصویر اولیه مورد ارزیابی قرار گرفته و با استفاده از خطای موجود نسبت به بهبود کارایی شبکه اقدام نمود. در این روش عمل مقایسه با استفاده از الگوریتم‌های ارزیابی کیفیت تصویر با مرجع کامل، همچون SSIM[۱] و MAE[۲] انجام می‌شود که کارایی این الگوریتم‌ها به صورت کلی به این شکل تعریف می‌شود که میزان اختلاف بین دو تصویر با یک معیار خاص مورد ارزیابی قرار گرفته، که هرچه دو تصویر شباهت بیشتری داشته باشند، در SSIM معیار به سمت یک و در MAE به سمت صفر میل می‌کند. شکل ۱۱ دیاگرامی از رویه‌ی اجرایی در این ابزار را نشان می‌دهد.

همانطور که مشاهده می‌شود ابتدا تصاویر مورد نیاز جهت آموزش مدل که مربوط به ناحیه چهره فرد است از فریم‌های ویدیو و یا مجموعه تصاویر ورودی استخراج شده و وارد شبکه کدگذار می‌شوند. سپس اطلاعات استخراج شده در اختیار کدگشا قرار می‌گیرند تا نسبت به بازسازی تصویر اولیه اقدام کند. نهایتاً تصویر بازسازی شده با نسخه اصلی مورد مقایسه قرار می‌گیرد و از خطای موجود جهت بهبود وزن‌های شبکه استفاده می‌شود.



شکل ۱۱: نحوه عملکرد ابزار *FaceSwap*

یکی از پارامترهای موثر در فایل خروجی، "مدل" انتخابی جهت آموزش و نهایتاً جابجایی چهره است که وابسته به ماهیت داده‌های آموزشی و یا قدرت سخت افزار مورد استفاده می‌توانند کارایی متفاوتی را از خود نشان دهند.

[۱] Structure Similarity Index

[۲] Mean Squared Error

همچنین مدل انتخابی می‌تواند از پارامترهای بیشتر و یا کمتری جهت ایجاد تغییر برخوردار باشد. برخی از مدل‌های ارائه شده در ادامه تشریح میگردند.

- Original

مدلی که نسخه اولیه FaceSwap از آن استفاده می‌نمود. این مدل از تصاویر با سایز ۶۴×۶۴ در طول آموزش استفاده می‌کند و در معماری آن از یک کدگذار مشترک و دو کدگشای متمایز استفاده شده است.

- LightWeight

این مدل که از تصاویر ورودی با سایز ۶۴×۶۴ استفاده می‌کند جهت استفاده بر روی سخت افزارهای ضعیف دارای حافظه کمتر از ۲ گیگابایت طراحی شده است. خروجی این مدل کیفیت بسیار بالایی نخواهد داشت اما جهت درک نحوه انجام عمل جابجایی صورت توسط این الگوریتم توسط محققین می‌تواند مفید باشد.

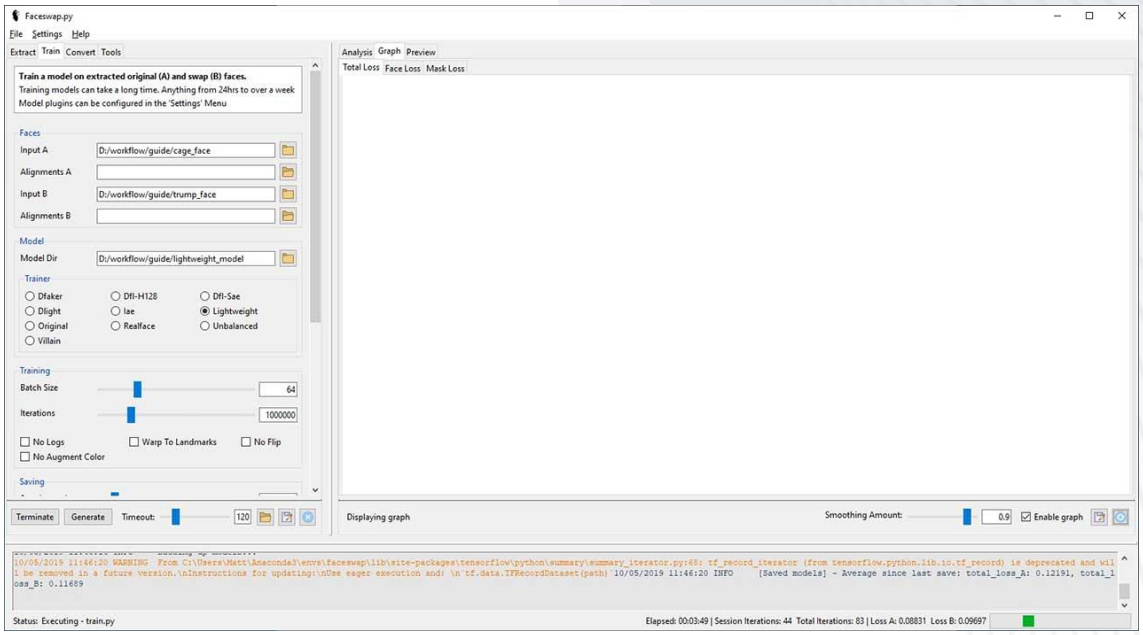
- IAE

این مدل نیز از تصاویر با سایز ۶۴×۶۴ استفاده می‌کند اما نسبت به سایر مدل‌های موجود اندکی متفاوت عمل می‌کند. در این مدل از یک کدگذار و کدگشا مشترک به همراه سه لایه واسط بین این دو شبکه استفاده شده است. هدف از این تغییرات رسیدن به کارایی بهتر در تمیز دادن تصاویر است.

- Villain

این مدل که شاید بتوان از آن به عنوان مدلی یاد کرد که نسبت به سایر مدل‌های ارائه شده جزئیات بیشتری را در روال آموزش در نظر می‌گیرد از تصاویر ۱۲۸×۱۲۸ پیکسلی استفاده می‌کند و همچنین به دلیل حجم محاسباتی بالا به سخت‌افزارهای بسیار قدرتمندی نیاز دارد. در صورت استفاده از این مدل بر روی سخت‌افزارهای دارای قدرت محدود فایل خروجی ممکن است با مشکل عدم تطبیق رنگ مواجه شود.

نکته قابل توجه در رابطه با این ابزار ارائه یک رابط کاربری گرافیکی با نام FakeApp است که با استفاده از آن، کاربر بدون هیچ نیازی به تحلیل و ارتباط با کد پیاده‌سازی، به راحتی می‌تواند با تنظیم پارامترهای تعیین شده بوسیله رابط‌های گرافیکی نسبت به تولید ویدیو ساختگی خود اقدام کند. شکل ۱۲ صفحه اولیه رابط کاربری FakeApp را نشان می‌دهد.



شکل ۱۲: رابط کاربری **FakeApp** جهت استفاده از **FaceSwap**

DeepFaceLab (DFL) -۲-۴

این ابزار نسخه تکامل‌یافته FaceSwap است که با افزودن چندین مدل مختلف، جهت استخراج ویژگی و ساختارهای تصویر، همچون H64، H128، SAE، LIAEF128 به مجموعه مدل‌های موجود در روش FaceSwap کارایی آن را در برخی موارد بهبود بخشیده است. اطلاعات مختصری از جزئیات برخی مدل‌های موجود در DFL در ادامه مطرح شده است.

- H64

این مدل در واقع همان مدل Original در FaceSwap است با این تفاوت که پیچیدگی محاسباتی آن در بخش آموزش مدل و همچنین ترکیب فریم‌های نهایی جهت ایجاد فایل خروجی بهبود یافته است. این مدل جهت اجرا بر روی کارت‌های گرافیکی کمتر از ۳ گیگابایت طراحی شده است. نکته دیگری که در رابطه با این مدل می‌توان به آن اشاره کرد سایز تصاویر ورودی و خروجی است که از تصاویر با سایز ۶۴×۶۴ استفاده می‌کند.

- H128

این مدل در واقع همان مدل H64 است با این تفاوت که تصاویر ورودی و خروجی شبکه سایز ۱۲۸×۱۲۸ پیکسل دارند. این مدل جهت اجرا بر روی کارت‌های گرافیکی با ظرفیت ۳ تا ۴ گیگابایت طراحی شده است.

- SAE

مدل SAE مجموعه تمام مدل‌های موجود در این ابزار است که بر اساس پارامترهای تعیین شده در طول اجرا از روش‌های مناسب و بهینه جهت آموزش استفاده می‌کند. این ابزار از برخی ویژگی‌های دیگر نیز برخوردار است که از مهمترین آنها عبارتند از:

- قابل اجرا بر روی تمامی کارت‌های گرافیکی سازگار با OpenCL دارای حجم ۲۵۶ مگابایت و بالاتر
- قابلیت انتخاب کارت گرافیک برتر موجود بر روی سیستم بصورت خودکار جهت پردازش‌های مورد نیاز
- قابلیت استخراج چهره چند شخص بر روی یک تصویر.

۴-۳- ابزار DFaker

این ابزار نیز یکی دیگر از ابزارهای موجود جهت ایجاد محتوای ساختگی است که برگرفته از FaceSwap است. این ابزار بر پایه کتابخانه Keras توسعه یافته و همچنین از مدل dfaker استفاده می‌کند که تصاویر ۶۴×۶۴ را به عنوان ورودی دریافت می‌کند و تصاویر با سایز ۱۲۸×۱۲۸ را در خروجی ایجاد میکند. جهت بهبود عملکرد شبکه بازسازی تصویر نیز، dfaker بجای SSIM از DSSIM بهره برده است.

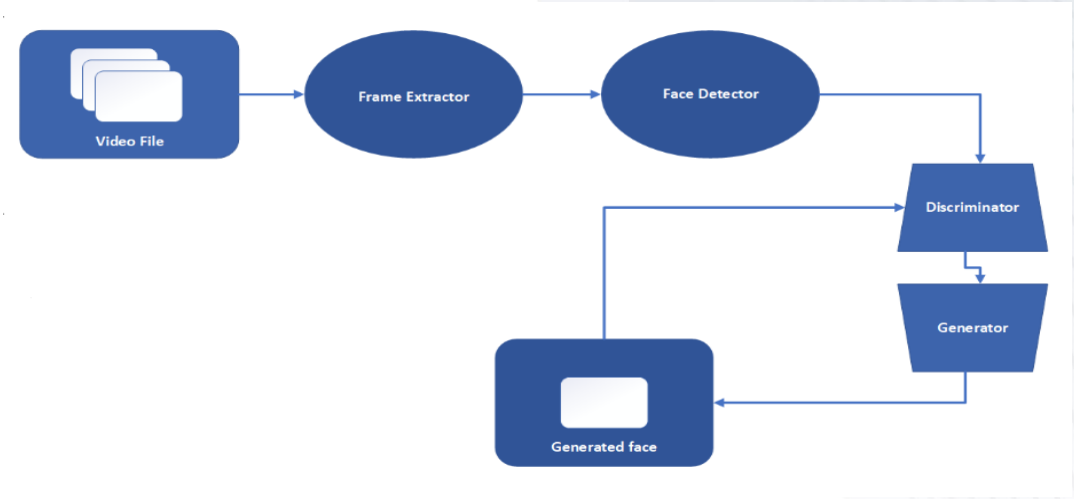
SSIM یکی از الگوریتم‌های ارزیابی کیفیت تصویر است که میزان شباهت بین تصویر اصلی و تصویر تغییریافته را با یک مقدار عددی نشان می‌دهد. هرچه تصویر ثانویه به نسخه اصلی نزدیک‌تر باشد این مقدار به عدد ۱ نزدیک‌تر خواهد بود.اساس کار این روش به این صورت است که اثرات تخریب اعمال شده بر روی تصویر را بر روی سه پارامتر ساختار[۱]، درخشندگی[۲] و کنتراست[۳] مورد ارزیابی قرار داده و نهایتاً از ترکیب مقادیر بدست آمده پارامتر نهایی را محاسبه می‌کند. هدف استفاده از این معیار شناسایی اثرات تخریب صورت گرفته بر روی نواحی مختلف تصویر است، به این ترتیب که پارامتر ساختار میزان تغییراتی که بر روی رفتار پیکسل‌های نزدیک به هم صورت گرفته را ارزیابی

[۱] Structure
[۲] Luminance
[۳] Contrast

می‌کند. درخشندگی باعث نمایان شدن میزان تخریب و اثرات ناشی از آن در لبه‌های موجود در تصویر می‌شود و نهایتاً پارامتر کنتراست تغییرات صورت گرفته بر روی نواحی دارای بافت متراکم را شناسایی می‌کند.

۴-۴- ابزار FaceSwapGan

FaceSwapGan یکی دیگر از ابزارهای ساخت محتوای جعلی است که با اعمال تغییراتی در معماری FaceSwap به کارایی قابل قبولی دست یافته است. در این ابزار بجای استفاده از یک الگوریتم خارجی مجزا جهت ارزیابی میزان شباهت تصویر بازسازی شده با تصویر اولیه، از ترکیب شبکه‌های عصبی GAN در کنار شبکه طراحی شده استفاده شده است. در این روش پس از آنکه تصویر اولیه بر اساس ویژگی‌های استخراج شده بازسازی شد تصویر بدست آمده مجدداً وارد شبکه کدگذار می‌شود. پس از استخراج ویژگی‌های آن شبکه تصمیم می‌گیرد که تصویر دریافتی مربوط به شخص A بوده یا B. در صورتی که شبکه نتواند ساختگی بودن تصویر را تشخیص دهد تابع هزینه شبکه کدگشا کاهش می‌یابد و در صورتی که ساختگی بودن تصویر مشخص شود تابع هزینه آن افزایش می‌یابد. به این ترتیب هم شبکه کدگشا در تکرارهای بالا قدرت تشخیص بهتری خواهد داشت و هم شبکه کدگذار از خطاهای خود در بازسازی مطلع شده و نسبت به رفع آن اقدام می‌کند. شکل ۱۳ دیاگرام الگوریتم مطرح شده را نشان می‌دهد. همانطور که در تصویر مشخص است این الگوریتم از دو بخش تشخیص‌دهنده و مولد تشکیل یافته است. بخش تشخیص‌دهنده وظیفه تشخیص چهره ورودی و تمایز قائل شدن بین آنها را بر عهده دارد و بخش مولد بازسازی تصویر و جابجایی چهره‌ها را به عهده دارد. در این الگوریتم پس از اینکه ویژگی‌های تصویر استخراج و جهت بازسازی در اختیار بخش مولد قرار گرفت، تصویر بدست آمده که چهره شخص A در آن جایگزین چهره شخص B شده، مجدداً وارد شبکه تمایزدهنده می‌شود. در صورتی که این بخش بتواند تصویر دریافتی را به عنوان چهره شخص B شناسایی کند نشان از عملکرد ضعیف شبکه مولد است و همچنین در صورتی که عکس این عمل رخ دهد، یعنی شبکه تمایزدهنده نتواند جابجایی صورت گرفته را تشخیص دهد نشان از عملکرد خوب شبکه مولد خواهد بود.



شکل ۱۳: بلوک دیاگرام FaceSwap-Gan

۴-۵- نکات عملی در بکارگیری ابزارهای جعل عمیق

در رابطه با بکارگیری ابزارهای ایجاد جعل عمیق نکات زیر را باید در نظر داشت:

- رویه ایجاد جعل عمیق و عملیات پردازشی این ابزارها یک عملیات زمان‌بر است.
- استفاده از پردازنده‌های گرافیکی[۱] به جای پردازنده‌های معمولی[۲] در استفاده از این ابزارها کارایی بهتری خواهد داشت.
- بین کیفیت فایل نهایی ایجاد شده، و تعداد تصاویر نمونه ورودی و همچنین کیفیت آنها همبستگی بالایی وجود دارد بنابراین در صورت استفاده از تصاویر نمونه‌ای که چهره‌های موجود در آن با تصویر هدف تا حدی شبیه به یکدیگر باشند، کیفیت فایل نهایی بهبود می‌یابد.
- تصاویر نمونه جهت آموزش تا حد امکان باید اکثر حالات چهره هدف از جمله خنده، عصبانیت و ... را شامل باشند.

[۱] GPU

[۲] CPU

۵- کاربردها و مواجهه با فناوری جعل عمیق در جهان

فن‌آوری جعل عمیق می‌تواند برای اهداف مختلفی مورد استفاده قرار گیرد. از دیدگاه هدف بکارگیری تکنیک های جعل عمیق، کاربردها را می‌توان به دو دسته طبقه بندی کرد: کاربردهای سودمند (فرصت‌ها) و کاربردهای مخرب (تهدیدها).

۵-۱- فرصت‌ها (کاربردهای سودمند)

یکی از زمینه‌های اصلی کاربردهای جعل عمیق صنایع چندرسانه‌ای است. این فناوری در حال حاضر در دسترس هالیوود است و می‌توان افق‌های جدیدی را برای استفاده از آن ترسیم کرد. به عنوان مثال، ساخت فیلم‌هایی با استفاده از افراد معمولی و سپس قرار دادن چهره‌های هنرپیشه‌های مشهور در این فیلم‌ها. این کار قبلاً برای ساخت یک قسمت جدید از سری فیلم‌های سریع و خشن با چهره بازیگر فقید آن پل واکر انجام شده است.

در شرایطی که به طور فزاینده‌ای فیلم‌های هالیوودی در کشورهای مختلف نمایش داده می‌شوند، استودیوهای هالیوود بسته به بازار هدف می‌توانند کاری کنند که تعداد بیشتری بازیگر بومی آن کشور در فیلم وجود داشته باشند (مثلاً می‌توانند در فیلم‌هایی که برای هندی‌ها نمایش داده می‌شوند کاری کنند که تعداد بیشتری از چهره بازیگران هندی استفاده شود)، یا نتفلیکس می‌تواند این امکان را برای بینندگان فراهم آورد که خودشان بتوانند بازیگران فیلم را قبل از پخش انتخاب کنند.

فناوری جعل عمیق مجموعه‌ای از فرصت‌ها را در صنایع آموزشی ایجاد می‌کند، از جمله امکان ارائه اطلاعات به دانشجویان به روش‌های جذابتر نسبت به وسایل سنتی مانند خواندن کتاب و سخنرانی. اثرات این نوآوری شبیه به موج قبلی نوآوری آموزشی است که با افزایش امکان دسترسی همگانی به فایل‌های ویدیوئی بوجود آمد. با جعل عمیق، می‌توان فیلم‌هایی از چهره‌های تاریخی تهیه کرد که مستقیماً با دانشجویان صحبت می‌کنند و اطلاعاتی در مورد خود و یا یافته‌هایشان را در اختیار آنان قرار می‌دهند. به این ترتیب می‌توان به یک مبحث درسی یا سخنرانی خسته‌کننده جلوه‌ای جذاب داد.

از نظر فناوری پزشکی، می‌توان به مزایای بالقوه این فناوری برای روانشناسی، جراحی و برنامه‌های آموزشی اشاره کرد که به عنوان مثال به دانشجویان روانشناسی این امکان را می‌دهند که فیلم‌های دیجیتالی ساخته شده بر اساس این فناوری، چهره افراد را آنالیز و با آن کار کنند. به عنوان مثال از کاربردی دیگر در حوزه پزشکی، تیم Lyrebird با انجمن بیماران ALS (بیماری که منجر به از کار افتادن عضلات بیمار میشود و استفان هاوکینگ فیزیکدان بریتانیایی نیز به این بیماری مبتلا بود) همکاری می‌کند تا فرصتی را برای مبتلایان به ALS فراهم کند که حتی با از دست دادن توانایی صحبت کردن، صدای منحصر به فرد خود را، هرچند به صورت رایانه‌ای، بازپس گیرند. همچنین می‌توان پیش‌بینی کرد که طی چند سال آینده آواتارهای دیجیتالی برای "نشان دادن تمامی حالات چهره" در دسترس کسانی خواهد بود که به آن نیاز دارند .

مطمئنا نبوغ بشر کاربردهای مفید دیگری را نیز برای فناوری جعل عمیق در آینده ایجاد خواهد کرد.

۵-۲- تهدیدها (کاربردهای مخرب)

جعل عمیق می‌تواند بعنوان مکانیزم قدرتمندی برای برخی سوءاستفاده‌ها و تخریب شخصیت افراد استفاده شود. همچنین از جعل عمیق برای ترور شخصیتی سیاستمداران مشهور در پورتال‌های ویدئویی یا اتاق‌های گفتگو استفاده شده است. به عنوان مثال، چهره رئیس‌جمهور آرژانتین مائوریسیو مکری با چهره آدولف هیتلر جایگزین شد و چهره آنگلا مرکل با چهره دونالد ترامپ جایگزین شد. در آوریل ۲۰۱۸ ، جردن پله و جونا پرتی از چهره باراک اوباما برای ساخت یک کلیپ برای هشدار در مورد خطرات جعل عمیق، استفاده کردند. در ژانویه سال ۲۰۱۹ ، KCPQ، وابسته به تلویزیون فاکس، جعل عمیق چهره ترامپ را در طی ایراد سخنرانی او از دفترش پخش کرد و در آن ظاهر و رنگ پوست او را مسخره کرد.

جعل عمیق فقط تهدیدی برای افراد یا اشخاص خاص نیست. بلکه این توانایی را دارد که از جهات مختلفی حتی به جامعه نیز آسیب بزند. جو عمومی سیاست در حال حاضر از گردش نادرست اطلاعات رنج می‌برد. بعضی اوقات دروغ‌هایی برای تضعیف اعتبار شرکت‌کنندگان در مناقشات سیاسی مطرح می‌شوند و در بعضی موارد اطلاعات جعلی پایه‌های واقعیت‌هایی را که باید مبنای گفتمان‌های سیاسی قرار گیرند، از بین می‌برند. حتی بدون جعل عمیق نیز،

آسیب‌پذیری دنیای سیاست از اخبار دروغ بسیار زیاد است. اما جعل عمیق با استفاده از امتیاز مربوط به ایجاد "اخبار جعلی" از نوع تصویری یا ویدئویی آسیب‌پذیری‌ها را وخیم‌تر می‌سازد. زیرا چهره و صدای هر فرد، شخصی‌ترین و شاخص‌ترین ویژگی‌های معرف او برای دیگران است و جعل عمیق می‌تواند این ویژگی‌ها را با کیفیت بالا جعل نماید.

علاوه بر توانایی جعل عمیق برای تزریق باورهای اشتباه درمورد حقایق سیاسی، از جعل عمیق می‌توان برای شکل خاصی از عملیات خرابکارانه نیز استفاده نمود برای مثال توزیع یک ویدئو یا فایل صوتی مخرب اما نادرست از یک نامزد انتخاباتی. در نتیجه، تغییر نتایج انتخابات از این طریق کاملاً محتمل است، به ویژه اگر مهاجم بتواند توزیع زمانی را به گونه‌ای انجام دهد که فرصت کافی برای گردش ویدئوی جعلی وجود داشته باشد اما زمان کافی برای آنکه قربانی بتواند به طور مؤثر آن را دفع کند (با فرض اینکه بتواند) وجود نداشته باشد.

انتخابات ۲۰۱۷ در فرانسه نمونه‌ای از این خطرات را نشان می‌دهد. مخالفان مکرون با هدف جلوگیری از انتخاب امانوئل مکرون به عنوان رئیس جمهور فرانسه در سال ۲۰۱۷ ، برنامه مخفیانه‌ای را که آمیخته‌ای از جاسوسی سایبری و دستکاری اطلاعات بود آغاز کردند. این کارزار شامل سرقت تعداد زیادی از مکاتبات و اسناد دیجیتال و تغییر برخی از آنها با هدف مسئله‌دار کردن آنها و تزریق تعداد زیادی از این گونه اطلاعات در شبکه‌های اجتماعی انجام شد. این تلاش در نهایت به دلایل زیادی از جمله به‌کارگیری تکنیکی ضعیف که ردیابی این حمله را آسان می‌کرد، با شکست روبرو شد. هر چند این تلاش شکست خورد اما ممکن است حمله‌کنندگان دفعات بعد در زمان‌بندی و تبادل اطلاعات با احتیاط و دقت بیشتری عمل نمایند. به عنوان مثال، آنها ممکن است یک سند جعلی مخرب‌تر تولید نموده و آن را درست در زمان آغاز رای‌گیری پخش کنند. همچنین، بدتر از آن ممکن است با استفاده از جعل عمیق، فیلم یا مدارک صوتی ظاهراً واقعی پخش کنند که به طرز قانع کننده‌ای مکرون در آن انصراف خود را از انتخابات ابراز نموده و یا مرتکب اقدامی تکان‌دهنده شود.

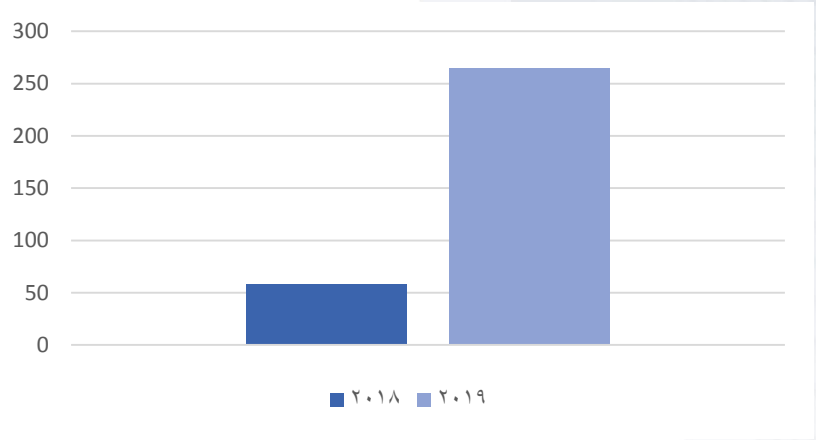
ویدئوهای جعلی مستهجن از افراد مشهور یکی دیگر از کاربردهای خلافکارانه جعل عمیق است. برای ساخت این ویدئوها عکس‌ها و فیلم‌های افراد مشهور، که به راحتی بصورت برخط در دسترس هستند، به عنوان داده‌های آموزش نرم‌افزارهای جعل عمیق مورد استفاده قرار می‌گیرند. پدیده ویدئوهای جعل عمیق مستهجن برای اولین بار در دسامبر

سال ۲۰۱۷ در بخش فنی و علمی مجله وایس منعکس شد که منجر به گسترش مباحثاتی در این زمینه در رسانه‌های دیگر گردید.

با توجه به نگرانی‌هایی که جعل عمیق با خود به ارمغان آورده، همزمان با اینکه افراد سودجوی بسیاری جهت اهداف خرابکارانه‌ی خود، در استفاده از این ابزار ترغیب شده‌اند، مراکز و موسسات معتبر علمی بسیاری نیز، جهت استفاده مفید و همچنین ارائه راه حل‌های پیشگیرانه از اقدامات مجرمانه به استفاده از این ابزار روی آورده‌اند. به همین جهت بررسی جعل عمیق در عرصه بین‌المللی مورد توجه قرار گرفت و در این رابطه از وبسایت www.linknovate.com در تاریخ ۹۸/۱۰/۱۴ برای یافتن و رتبه‌بندی نهادهای فعال در این حوزه، با استفاده از کلمات کلیدی deepfake و fake videos استفاده شد. ارزیابی وب‌سایت اشاره شده بر اساس ترکیب معیارهای انتشارات، کنفرانس‌ها، اعطائیه‌ها، حق امتیازات تجاری، اخبار و اطلاعات منتشره در وب صورت می‌گیرد. بر اساس گزارشات ارائه شده توسط LinkNovate برخی از مهمترین مراکزی که بیشترین فعالیت را در حوزه جعل عمیق داشته‌اند، به ترتیب عبارتند از:

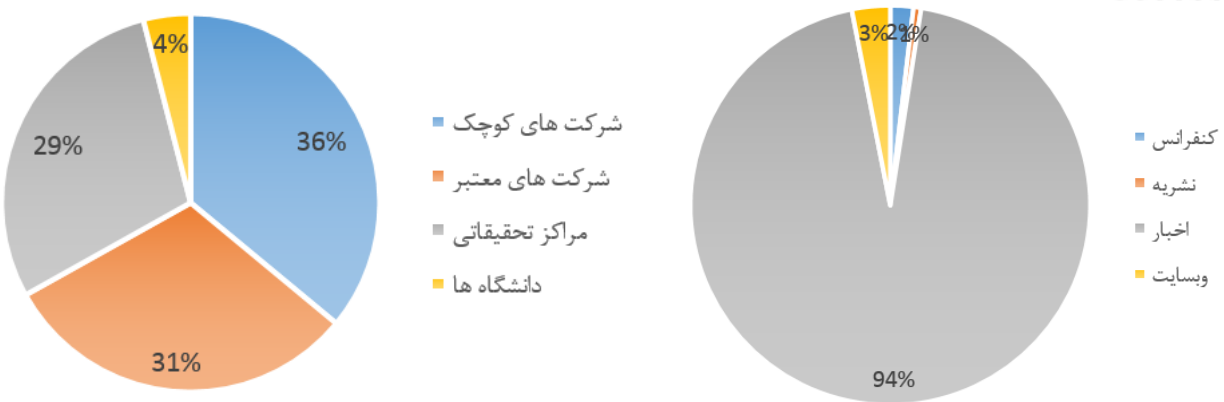
- آژانس پروژه‌های تحقیقاتی پیشرفته دفاعی
- دانشگاه کالیفرنیا
- دانشگاه مری‌لند
- دانشگاه فنی مونیخ
- دانشگاه استنفورد
- دانشگاه واشنگتن

این فعالیت‌ها سبب شده حجم اطلاعات منتشر شده مربوط به جعل عمیق در سال ۲۰۱۹میلادی نسبت به سال گذشته با رشدی ۴۰۰ درصدی مواجه شود. شکل ۱۴ صحت این ادعا را با مقایسه حجم اطلاعات منتشر شده در سال‌های ۲۰۱۸ و ۲۰۱۹ نشان می‌دهد.



شکل ۱۴: حجم اطلاعات منتشر شده در رابطه با جعل عمیق

در شکل ۱۵ نیز درصد توزیع این اطلاعات بر اساس منابع و سازمان مربوطه نشان داده شده است. اما نکته جالب دیگری که در رابطه با جعل عمیق وجود دارد، توزیع جغرافیایی فعالیت‌های مربوط به جعل عمیق در سراسر جهان می‌باشد که بیشترین حجم اطلاعات منتشرشده مربوط به کشور آمریکا و به دنبال آن کشورهای اروپایی به خصوص انگلستان و در رده سوم به کشور چین مربوط می‌شوند.



(ب) (الف)

شکل ۱۵: توزیع اطلاعات منتشر شده بر اساس (الف) محل انتشار و (ب) سازمان منتشر کننده

نکته قابل توجه دیگر، بررسی موضوعات مشابه و هم‌راستا با جعل عمیق می‌باشند که در این میان جعل عمیق در محتوای ویدیویی، تشخیص ویدیوهای تولید شده با جعل عمیق، هوش مصنوعی و شبکه‌های عصبی عناوینی بودند

که درصد زیادی از فعالیت های مربوط به این موضوع را به سمت خود کشانده‌اند. اما تمرکز اصلی بسیاری از موسسات در رابطه با جعل عمیق به بحث تشخیص ویدیوهای ساخته شده با این روش معطوف شده است. بر اساس گزارشات منتشر شده در رابطه با ابزار DeepTrace که جهت حفظ امنیت کاربران در فضای مجازی، توسط دولت هلند معرفی شده است، این ابزار از زمان آغاز فعالیت موفق به کشف نزدیک به ۱۴۷۰۰ محتوای جعلی شده است که مقایسه این آمار با حدود ۸۰۰۰ ویدیوی کشف شده در ماه دسامبر ۲۰۱۸، رشد ۸۴ درصدی گسترش این نوع ویدیوها در بازه زمانی هفت ماهه را نشان می‌دهد. سرعت گسترش محتوی جعلی با توجه به مواردی همچون جعل ویدیوهای انتخاباتی جهت منحرف نمودن اذهان عمومی و تخریب چهره‌های سیاسی، انجام فعالیت‌هایی در راستای نادیده گرفتن اقدامات مجرمانه و ساخت ویدیوهای مستهجن برای شخصیت‌های شناخته شده، بسیار نگران کننده است. از همین رو شرکت فیسبوک با همکاری آمازون، مایکروسافت و همچنین مراکز علمی معتبر از جمله دانشگاه ام آی تی، دانشگاه آکسفورد، دانشگاه برکلی و دانشگاه مری‌لند اقداماتی را در راستای ارائه روش‌های تشخیص ویدیوهای جعلی استارت زده‌اند، که از جمله مهمترین این اقدامات می‌توان به اختصاص سرمایه ۱۰ میلیون دلاری فیسبوک اشاره کرد.

۶- فناوری جعل عمیق در ایران

جهت بررسی وضعیت ایران در حوزه جعل عمیق به سایت www.linknovate.com رجوع گردید ولی اطلاعاتی در مورد ایران در رابطه با جعل عمیق ثبت نگردیده بود. به همین جهت در مورد فعالیت در این حوزه با بیش از ۳۰ نفر از اساتید دانشگاهی فعال در حوزه پردازش تصویر و هوش مصنوعی مکاتبه گردید ولی پاسخی مبنی بر فعالیت آنها در این حوزه دریافت نگردید. همچنین از چندین شرکت در حوزه پردازش تصویر و هوش مصنوعی در مورد فعالیت در این حوزه پرسش شد که پاسخ همگی منفی بود و در مکاتبات انجام شده با شرکتها نیز پاسخی مبنی بر فعالیت آنها در این حوزه دریافت نگردید.

با توجه به این موضوع که هسته مدیریتی کشور ما همواره در عرصه‌های مختلفی چه از سوی مهره‌های داخلی و چه خارجی، با اهداف تخریبی مورد بی‌مهری قرار گرفته است، و همچنین به دلیل ظهور جعل عمیق به عنوان ابزاری با تاثیر بسیار قوی بر باور عمومی، دور از ذهن نیست که در آینده‌ای نه چندان دور ابعاد مختلف امور کشور، با سوءاستفاده دشمنان و افراد سودجو از این فناوری، با چالش‌های جدی مواجه شود. با پیشرفت و توسعه این فناوری و امکان استفاده از آن حتی برای افراد غیرمتخصص نیز، می‌توان خطرات احتمالی ناشی از محتوای ایجاد شده با جعل عمیق را برای عموم افراد یک جامعه که تنها یک فیلم کوتاه از خود در شبکه‌های اجتماعی به اشتراک گذاشته‌اند، متصور شد.

از همین رو در ادامه سعی شده نقش وزارت ارتباطات و فناوری اطلاعات به عنوان متولی امر توسعه فناوری در کشور و پژوهشگاه ارتباطات و فناوری اطلاعات به عنوان بازوی پژوهشی وزارت در مواجهه فعالانه با کاربرد این فناوری مورد بررسی قرار گیرد.

۶-۱- افزایش سطح آگاهی عمومی و فرهنگ‌سازی پیرامون فناوری جعل عمیق

همانطور که اشاره شد، در کنار فعالیت‌های مخربی که می‌توانند با استفاده از فناوری نوظهور جعل عمیق عملیاتی شوند، کاربردهای مفید و سازنده‌ای نیز می‌توان برای آن در حوزه‌های مختلفی همچون آموزش و صنعت فیلم‌سازی تعریف نمود. به عنوان نمونه‌ای از این قبیل موارد می‌توان به استفاده از این فناوری در یکی از فیلم‌های هالیوودی

اشاره کرد که بعد از کشته شدن یکی از بازیگران نقش اصلی در یک حادثه رانندگی، مخاطبان و طرفداران همچنان شاهد هنرنمایی او در فیلم محبوب خود بودند.

با پیدایش جعل عمیق آمار ویدیوهای منتشرشده در شبکه‌های اجتماعی که به واسطه این فناوری ایجاد شده‌اند، به شکل قابل‌توجهی در حال افزایش است. برخی از این ویدیوها توانسته‌اند برای شخصی که مورد هدف قرار گرفته‌اند مشکلات جدی را رقم بزنند که با پیگیری و بررسی حقیقت ماجرا، هدف از ایجاد و انتشار حجم وسیعی از این قبیل محتوی از سوی سازنده آن صرفا تفریح و سرگرم نمودن مخاطب بیان شده است.

از همین رو وزارت ارتباطات و فناوری اطلاعات می‌تواند با یک برنامه‌ریزی دقیق نسبت به معرفی کاربردهای مفید این فناوری، ایجاد بستری مناسب جهت توسعه این فناوری در راستای اهداف مثبت و فعالیت‌های مشابه اقدام کند، تا آن دسته از افرادی که علاقمند به فعالیت در این شاخه می‌باشند به جای تولید محتواهای چالشی که ممکن است اثرات مخرب آن گستره وسیعی را در برگیرد، در مسیر درست و سازنده به فعالیت مورد علاقه خود بپردازند. در همین راستا پژوهشگاه ارتباطات و فناوری اطلاعات می‌تواند با برگزاری کارگاه‌ها و کنفرانس‌ها درحوزه فناوری جعل عمیق به بالا بردن آگاهی عمومی در این زمینه کمک نماید.

۶-۲- فعالیت بر روی موضوعات مرتبط با تشخیص جعل عمیق

محتواهای ویدیویی همواره به دلیل ماهیت نزدیک به واقعیتی که داشته‌اند مورد توجه عموم افراد بوده‌اند. از همین رو جذابیت فایل‌های ویدیویی از یک سو، و حس کنجکاوی که همواره افراد جامعه نسبت به حریم شخصی زندگی چهره‌های شناخته شده داشته‌اند از سوی دیگر باعث شده تا برخی سودجویان در راستای رسیدن به منافع شخصی خود از این فناوری در جهت تولید ویدیوهای مختلف مرتبط با زندگی شخصی چهره‌های مطرح، که در گروه محتوای پرمخاطب دسته‌بندی می‌شوند، سوءاستفاده کنند. اما سوءاستفاده از جعل عمیق صرفاً به ایجاد محتوای غیرواقعی جهت رسیدن به منافع شخصی محدود نمی‌شود، بلکه می‌تواند با اهداف انتقام‌جویانه نسبت به یک فرد عادی و یا تخریب یک چهره سیاسی در یک رقابت انتخاباتی نیز همراه شود.

بدون شک ابزاری که با ایجاد یکی سری محتوای غیرواقعی تا این حد می‌تواند بر باور عمومی یک جامعه تاثیر داشته باشد از دست افراد سودجو دور نخواهد ماند. با توجه به این موضوع که مدیران ارشد کشور نیز همواره از سوی برخی رسانه‌های خارجی مورد اهداف تخریبی قرار گرفته‌اند، دور از ذهن نیست که در آینده‌ای نزدیک با حجم زیادی از ویدیوهایی مواجه شوند که حتی خود آنها در رابطه با ساختگی بودن آن دچار شک و تردید شوند.

از همین رو نهادهای مربوطه باید با برنامه‌ریزی فعالیت‌های موثری را در جهت تشخیص و تمیز دادن ویدیوهای غیرواقعی از واقعی شروع کنند تا در صورت مواجه با این گونه جرایم در سریع‌ترین زمان ممکن و پیش از تاثیر بر روی افکار عمومی واقعیت پشت پرده فاش شود. پژوهشگاه ارتباطات و فناوری اطلاعات با تعریف و اجرای پروژه‌های تحقیقاتی و کاربردی در زمینه تشخیص رسانه‌های جعل عمیق و با انجام حمایت‌های تشویقی از کسانی که در این حوزه فعالیت می‌نمایند، می‌تواند گام‌های موثری را در این رابطه بردارد. حمایت از این طیف متخصصین می‌تواند بصورت تعریف پروژه‌های مشارکتی یا برگزاری مسابقات و یا تهیه پایگاه‌های داده مورد نیاز باشد. در همین راستا ایجاد آزمایشگاه تشخیص محتواهای دیداری جعل عمیق در پژوهشگاه ارتباطات و فناوری اطلاعات در دست بررسی است. هدف از این فعالیت‌ها ایجاد برنامه‌های کاربردی برای تشخیص رسانه‌های جعل عمیق و در دسترس عموم قرار دادن این خدمات است تا حدی که یک فرد یا یک مقام مسئول بتواند به سهولت از جعلی بودن یک محتوای دیداری یا شنیداری مطلع گردد.

۶-۳- ایجاد ساختارهای قانونی جهت مقابله با تهدیدات فناوری جعل عمیق

همانطور که پیش‌تر اشاره شد ویدئوهای جعل عمیق برای کلاهبرداری و اقدامات فریبکارانه قابل‌استفاده هستند. به همین جهت ایجاد ساختارهای قانونی جهت مبارزه با این اقدامات مجرمانه یکی از ضروریات مواجهه با تهدیدات فناوری جعل عمیق در جامعه است. وضع قوانین جدید برای کاربرد ویدئوهای دستکاری شده با استفاده از ابزار رسانه‌ای جعل عمیق در حوزه جرایم سایبری یکی از نیازمندی هاست. در این راستا برنامه‌های تشخیص رسانه‌های جعلی که پیشتر ذکر شدند، می‌توانند کمک شایان توجهی در اجرای قوانین وضع شده به عمل آورند.

نشانی: تهران، انتهای کارگر شمالی، پژوهشگاه ارتباطات و فناوری اطلاعات، معاونت پژوهش و توسعه ارتباطات علمی

تلفن: ۰۲۱-۸۸۶۳۰۳۵۵

نمابر: ۰۲۱-۸۸۶۳۰۳۵۶