ردیابی جعل عمیق هوش مصنوعی

در فوریه ۲۰۲۴ شرکت OpenAI ویدئوهایی را منتشر کرد که با یک برنامه هوش مصنوعی مولد به نام «سورا» (Sora)محصول خود این شرکت ساخته شده بودند. این محتواهای فوق‌العاده واقع‌گرایانه که با خطوط فرمان متنی ساده تولید شده‌اند، تازه‌ترین دستاورد شرکت‌هایی هستند که می‌خواهند توانمندی‌های فناوری هوش مصنوعی را به دنیا نشان دهند. اما از سوی دیگر نگرانی هایی را درباره توانایی بالقوه هوش مصنوعی مولد در کمک به ایجاد محتوای فریب دهنده و گمراه کننده به وجود آورده‌اند.

طبق پژوهش‌های انجام شده در دانشگاه «درِکسِل» در ایالات متحده، روش‌های فعلی که برای شناسایی عکس و ویدئوهای دیجیتالی دستکاری شده به کار می‌روند در مقابل ویدئوهایی که هوش مصنوعی تولید می کنند مؤثر عمل نمی کنند اما یک رویکرد یادگیری ماشین می‌تواند نقاب تقلب را از چهره این ویدئوهای ساختگی بردارد.

فناوری تشخیص تصاویر غیرواقعی و جعلی که در حال حاضر به کار می‌روند نمی‌تواند ویدئویی را که هوش مصنوعی می‌سازد را از ویدئوی واقعی تشخیص دهد اما الگوریتم یادگیری ماشینی که در آزمایشگاه امنیت اطلاعات چند رسانه‌ای دانشگاه درِکسِل ساخته شده‌است می‌تواند طوری آموزش ببیند که قادر باشد رد و اثر بسیاری از مولدهای ویدئو مثل Video-Crafter، Stable Video Diffusion و Cog-Video را تشخیص دهد و شناسایی کند. به‌علاوه، این الگوریتم می‌تواند یاد بگیرد مولدهای هوش مصنوعی جدید را که ویدئوهای ساختگی تولید می کنند شناسایی کند. کافی است برای این کار فقط چند نمونه انگشت شمار ویدئو را دریافت و وارسی کند.

شرکت‌های مهندسی تلاش خود را می کنند تا شناسه‌ها و واترمارک هایی را در برنامه‌ها قرار دهند اما زمانی که این فناوری در دسترس عموم قرار گیرد، افرادی که قصد دارند از آن برای فریب دیگران و اهداف منفی استفاده کنند راهی برای خود پیدا خواهند کرد. به همین دلیل است که سازندگان الگوریتم جدید سعی می کنند از افراد متقلب جلوتر باشند. برای این منظور، فناوریی را می‌سازند که بتواند ویدئوهای تقلبی را از روی الگوها و ویژگی‌هایی که برای رسانه ویدئویی بومی محسوب می‌شوند را شناسایی کند.

گروه مهندسین دانشگاه درِکسِل در طول یک دهه گذشته تلاش کرده‌اند تصاویری را که به صورت دیجیتالی دستکاری شده‌اند را علامت‌گذاری کنند اما در دو سال اخیر حجم کار آن‌هابیشتر شده است چون افراد متقلب از فناوری ویرایش تصویر برای انتشار اطلاعات نادرست و اخبار جعلی سیاسی استفاده می کنند.

تا چندی پیش، این دستکاری‌های بدخواهانه را از طریق برنامه‌های ویرایش کننده تصاویر و ویدئو انجام می‌دادند، برنامه‌هایی که می‌توانند پیکسل‌هایی را به تصاویر اضافه یا از آن‌هاحذف می کنند یا به طور کلی پیکسل‌ها را تغییر می‌دهند. همچنین این برنامه‌ها می‌توانند فریم‌های ویدئوها را کند یا تند کنند یا فریم‌ها را از ویدئو برش بزنند و بردارند. هر کدام از این ویرایش‌های انجام شده یک ردپای دیجیتالی منحصربه‌فرد و بی‌شباهت در محصول تصویری به جا می‌گذارند. پژوهشگرهای درِکسِل برای یافتن و ردیابی آن‌هایک سلسله ابزار ساخته‌اند.

این ابزارها مجهز به یک برنامه یادگیری ماشین پیچیده هستند که شبکه عصبی محدود شده نام دارد. این الگوریتم به جای این که از ابتدا به دنبال شناسه‌های دستکاری دیجیتالی از پیش تعیین‌شده به خصوصی بگردد، می‌تواند به شیوه‌هایی مشابه مغز انسان و با دقت زیر پیکسلی (sub-pixel)یاد بگیرد چه چیزی در عکس‌ها و ویدئوها عادی و طبیعی و چه چیزی غیرعادی است. این ویژگی، برنامه یادگیری ماشین را هم در شناسایی تصاویر جعل عمیق که از منابع شناخته شده‌ای منتشر می‌شوند و هم در شناسایی تصاویر ساختگی که با یک برنامه ناشناخته ساخته شده‌اند توانا می کند.

این شبکه عصبی به طور خاص با صدها یا هزاران نمونه تصویر آموزش دیده است تا با تیزبینی متوجه تفاوت بین ویدئوهای ویرایش نشده و ویدئوهای دستکاری شده شود. این میزان دقت از توانایی تشخیص اختلاف بین پیکسل‌های مجاور هم تا ترتیب فاصله بین فریم‌های یک ویدئو و اندازه و فشردگی فایل‌ها متغیر است.

وقتی ویدئویی می‌سازیم، سیستم پردازش الگوریتمی دوربین ما رابطه بین مقادیر پیکسل‌های مختلف که از مقادیر پیکسل‌های تصاویر ایجاد شده با فتوشاپ یا هوش مصنوعی بسیار متفاوت هستند را معرفی می کند. اما به تازگی مولدهای ویدئویی مثل سورا روی کار آمده‌اند که ویدئوهای زیبا و خیره کننده‌ای می‌سازند. این تصاویر گیرا چالش تازه‌ای را پیش می‌آورند چون نه با دوربین گرفته شده‌اند و نه با فتوشاپ طراحی شده‌اند.

حتی اگر ویرایشی هم صورت نگرفته باشد، سرنخ‌های استانداردی وجود ندارد و نبود این سرنخ‌ها در تشخیص جعل از واقعیت مشکل بزرگی به وجود می‌آورد.

تا به امروز برنامه‌های شناسایی که در روش‌های علوم قانونی و جنایی استفاده می‌شوند در مقابل ویدئوهای ویرایش شده عملکرد خوبی داشته‌اند. آن‌هابه سادگی ویدئوها را یک سری تصویر قلمداد کرده و فرایند تشخیص یکسانی را روی آن‌هااعمال می کنند. اما در ویدئوهایی که با هوش مصنوعی ساخته می‌شوند هیچ نشانه و مدرکی از دستکاری فریم به فریم تصویر وجود ندارد. بنابراین برای این که یک برنامه شناسایی‌کننده مؤثر عمل کند، باید قادر به یافتن رد و نشانه‌هایی باشد که از شیوه ساخت ویدئو با برنامه‌های هوش مصنوعی مولد به جا مانده‌اند.

الگوریتم MISLnet که یک شناسایی‌کننده مبتنی بر شبکه‌های عصبی پیچشی است می‌تواند دست ویدئوهای ساختگی را رو کند چون این برنامه هر بار که با مثال‌های تازه‌ای روبرو می‌شود به طور مداوم یادگیری خود را تغییر می‌دهد. طی سال‌های گذشته تیزفهمی الگوریتم MISLnetدر استخراج تصاویری که با برنامه‌های جدید ویرایش تصویر از جمله ابزارهای هوش مصنوعی دستکاری شده‌اند به اثبات رسیده است.

پژوهشگران برای شناسایی عکس‌های دستکاری‌شده و نیز ویدئوها و فایل‌های صوتی که با جعل عمیق ساخته شده‌اند از الگوریتم‌های CNN استفاده کرده‌اند که نتیجه‌ای موفقیت‌آمیز داشته است. این الگوریتم‌ها به دلیل قابلیتی که در سازگار شدن با مقادیر کوچک اطلاعات جدید دارند، می‌توانند راه حل مؤثری برای شناسایی ویدئوهای جعلی ایجاد شده با هوش مصنوعی باشند.

پژوهشگران برای آزمودن عملکرد الگوریتم‌های CNN، هشت الگوریتم از جمله MISLnetرا با همان مجموعه دادگان محک زننده‌ای تغذیه کردند که برای آموزش به شناسایی کننده‌های تصویر استفاده شده بودند. سپس برنامه را با ویدئوهای ایجاد شده با برنامه‌های هوش مصنوعی مولد که هنوز در دسترس عموم قرار نگرفته‌اند آزمایش کردند. این برنامه‌ها سورا، پیکا و VideoCrafter-v۲ هستند.

الگوریتم‌های CNNتوانستند با تجزیه و تحلیل بخش کوچکی از هر فریم در هر ویدئو درک کنند یک ویدئوی ساختگی با جزئیاتی در حد دانه‌های تصویر (granular)چگونه است و چه ظاهر و ویژگی‌هایی دارد سپس توانستند شناختی را که به دست آورده‌اند را روی ویدئوهای جدید اعمال کنند. هر برنامه در شناسایی ویدئوهای جعلی بیش از ۹۳درصد بازده از خود نشان داد که عملکرد MISLnet از همه بهتر و بازده آن ۹۳/۳ درصد بود. هنگامی که وظیفه تحلیل تمامی یک ویدئو به این برنامه‌ها داده شد بازده آن‌هااندکی بالاتر رفت و دقت آن‌هابه ۹۵ تا ۹۸ درصد رسید.