ترفند جدید محققین، تمام سیستم های تشخیص مبتنی بر AI را فریب می دهد - دیجیاتو

حمید مقدسی | یکشنبه، ۰۸ مرداد ۱۳۹۶

امروزه هوش مصنوعی به حدی رسیده که می تواند اشیاء موجود در یک تصویر، یا گفتار انسان را با دقت خوبی تشخیص دهد، با این حال الگوریتم های زیرساختی آن عملکردی متفاوت با مغز انسان دارند و به همین دلیل، می توان آنها را فریب داد.

به تازگی گروهی از دانشمندان با همکاری تیم هوش مصنوعی فیسبوک، نشان داده اند که با اعمال تغییرات جزئی در فایل های صوتی می توان تشخیص آنها را برای سیستم های AI ناممکن ساخت، در حالی که انسان همچنان صدایی عادی و طبیعی را می شنود.

فایل تغییر یافته برای انسان کاملاً طبیعیست، ولی ماشین را دچار مشکل می کند

در این روش، لایه ای بی صدا از نویز وارد کلیپ صوتی می شود که الگوهای متمایز دیگر کلمات را بر اساس داده های شبکه عصبی در خود دارد. این الگوریتم که نام شعبده باز معروف آمریکایی یعنی «هودینی» (Houdini) را به خود گرفته، توانست تشخیص گفتار پیشرفته گوگل و دیگر سامانه های مشابه را به طور کامل به هم بریزد، در حالی که هر دو کلیپ صوتی ارائه شده به سیستم از نظر انسان کاملاً مشابه و درست بودند.

تیم مورد بحث این روش را روی دیگر الگوریتم های یادگیری ماشین نیز تست کرده و به نتایج مشابهی دست یافتند. مثلاً با دستکاری تصویر افراد، تشخیص حالت صحیح بدن توسط الگوریتم را ناممکن ساخته، یا علائم و نشانه های جاده را در سیستم رانندگی هوشمند به اشیاء غیرواقعی بدل سازند.

کاربرد اصلی تست های عجیب فوق را می توان در ارزیابی توانایی الگوریتم های یادگیری ماشین در شرایط دشوار جستجو کرد، اما همواره امکان سوء استفاده هم وجود دارد. تمامی سامانه های تشخیص هوش مصنوعی با این روش فریب می خورند؛ مثلاً ماشین خودران ترافیک غیر واقعی را می بیند، یا اسپیکر هوشمند فرمان های ناصحیح را می شنود.

بخش بغرنج تر ماجرا آنجاست که ما واقعاً سازوکار درونی شبکه های عصبی مصنوعی عمیق را نمی دانیم و به همین دلیل، نمی توانیم بفهمیم چرا اختلالاتی در این حد جزئی، نتایجی چنین مخرب را به بار می آورند. بنابراین کشف راهکاری برای مقابله با آن هم بسیار دشوار خواهد بود.

دیجیاتو