Inteligência artificial: Os 'eventos estranhos' que fizeram tecnologia pensar que tartaruga era uma arma...

Linda Geddes - BBC Future

09/02/2019 15h58

Pequenos ruídos nas redes neurais artificiais podem ter implicações devastadoras à medida que empregamos cada vez mais a IA em nossas rotinas.

O passageiro faz o sinal de parada e fica em pânico quando o carro em que está começa a acelerar. Ele até pensa em gritar com o motorista, mas percebe - ao ver o trem rasgando em sua direção nos trilhos à frente - que não há ninguém dirigindo. O trem a 200 km/h esmaga o veículo autônomo e mata na hora o ocupante.

Esse cenário é fictício, mas aponta uma falha muito real na estrutura de inteligência artificial. Certos "ruídos" podem confundir o sistema de reconhecimento das máquinas e as fazerem "alucinar". O resultado pode ser tão grave quanto o descrito acima. Apesar de o sinal de parada ser claramente visível aos olhos humanos, a máquina pode não conseguir reconhecê-lo por causa de alterações na imagem.

Aqueles que trabalham com inteligência artificial descrevem essas falhas como "exemplos contraditórios" ou, de maneira simplificada, "eventos estranhos".

"Podemos entender essas falhas como informações que deveriam ser processadas de uma maneira pela rede, mas que cujos resultados são inesperados", afirma Anish Athalye, cientista da computação do Instituto de Tecnologia de Massachusetts (MIT, na sigla em inglês), em Cambridge.

Vendo coisas

Os sistemas de reconhecimento visual têm recebido especial atenção nesses casos. Pequenas alterações em imagens podem enganar as redes neurais - os algoritmos de aprendizado da máquina que direcionam grande parte da tecnologia moderna de IA. Esse tipo de sistema já é usado, por exemplo, para marcar amigos em fotos ou identificar objetos nas imagens do smartphone.

Com alterações leves na textura e na cor de objetos impressos em 3D, Athalye e colegas fizeram uma bola de beisebol, por exemplo, ser classificada como um café expresso; e uma tartaruga, confundida com um rifle. Eles enganaram o computador com cerca de 200 outros objetos impressos em 3D. À medida que colocamos mais robôs em casa, drones no céu e veículos autônomos na rua, esse resultado se torna preocupante.

"No começo isso era apenas uma curiosidade", diz Athalye. "Agora, no entanto, enxergamos isso como um potencial problema de segurança, já que os sistemas estão sendo cada vez mais implementados no mundo real."

Tome como exemplo os carros sem motorista que hoje passam por testes práticos: eles geralmente dependem de sofisticadas e profundas redes neurais de aprendizagem para navegar e dizer-lhes o que fazer.

No entanto, os pesquisadores mostraram que, ao simplesmente colocar pequenos adesivos em placas de limite de velocidade, as redes neurais não conseguiram compreendê-las.

Ouvindo vozes

As redes neurais não são as únicas estruturas de aprendizado de máquina em uso, e todas também parecem vulneráveis a esses eventos estranhos. E elas não estão limitadas a sistemas de reconhecimento apenas visual.

"Em todas as áreas, da classificação de imagens ao reconhecimento automático de voz e a tradução, as redes neurais podem classificar dados incorretamente", diz Nicholas Carlini, pesquisador do Google Brain, que desenvolve máquinas inteligentes.

Carlini mostrou como - com a adição de um pouco de ruído de fundo - uma voz que deveria ler: "Without the dataset the article is useless" foi lida como "Ok Google browse to evil dot com". E os erros não se limitam apenas à fala. Em outro exemplo, um trecho da Suíte Nº 1 para violoncelo de Bach foi transcrita como "a fala pode ser incorporada na música".

Para Carlini, tais exemplos contraditórios "provam conclusivamente que o aprendizado de máquina ainda não atingiu a capacidade humana mesmo em tarefas muito simples".

Sob a pele

Redes neurais se baseiam, de forma superficial, em como o cérebro processa a informação visual e aprende com ela. Imagine uma criança pequena aprendendo o que é um gato: à medida que se depara com mais dessas criaturas, ela começa a perceber os padrões - essa mancha chamada gato tem quatro patas, pelo macio, duas orelhas pontudas, olhos amendoados e um rabo comprido e macio.

Dentro do córtex visual da criança (a área do cérebro que processa a informação visual), há camadas sucessivas de neurônios que respondem a detalhes visuais, como linhas horizontais e verticais, permitindo que a criança construa uma imagem neural do mundo e aprenda com isso.

As redes neurais funcionam de maneira semelhante. Os dados fluem por meio de camadas de neurônios artificiais até que - depois de serem treinadas em centenas ou milhares de exemplos da mesma coisa (geralmente rotuladas por um humano) - a rede comece a identificar padrões do que está sendo visualizado. O mais sofisticado desses sistemas emprega o "aprendizado profundo" (deep learning), o que significa que eles têm mais camadas.

No entanto, embora os cientistas da computação entendam os detalhes básicos de como as redes neurais funcionam, eles não sabem exatamente o que está acontecendo quando elas processam os dados. "Atualmente, não os entendemos bem o suficiente para, por exemplo, explicar por que existe o fenômeno de exemplos contraditórios e saber como corrigi-lo", diz Athalye.

Parte do problema pode estar relacionada à natureza das tarefas que as tecnologias existentes foram projetadas para resolver: distinguir entre imagens de cães e gatos, por exemplo. Para fazer isso, a tecnologia processa vários exemplos de cães e gatos, até que tenha dados suficientes para diferenciá-los.

"O objetivo principal de nossas estruturas de aprendizado de máquina era obter um desempenho em média bom", diz Aleksander Madry, outro cientista da computação do MIT, que estuda a confiabilidade e a segurança das estruturas de aprendizado de máquina. "Quando você treina o programa para ser apenas bom, sempre haverá imagens que vão confundi-lo".

Uma solução pode ser treinar redes neurais com exemplos mais desafiadores do que os atuais. Isso poderia fortalecê-los contra os pontos fora da curva.

"Definitivamente, é um passo na direção certa", diz Madry. Mas, mesmo que essa abordagem torne as estruturas mais robustas, ela provavelmente tem limites, pois há várias maneiras de se modificar a aparência de uma imagem ou um objeto para gerar confusão.

Um classificador de imagens verdadeiramente robusto replicaria o que a "semelhança" significa para um humano: ele entenderia que o rabisco de um gato feito por uma criança representa a mesma coisa que uma foto de um gato ou um gato em movimento na vida real. Por mais impressionantes que sejam as redes neurais profundas de aprendizado, elas ainda não são páreo para o cérebro humano quando se trata de classificar objetos, entender seu ambiente ou lidar com o inesperado.

Se quisermos desenvolver máquinas realmente inteligentes que possam funcionar em cenários do mundo real, talvez devamos voltar ao cérebro humano para entender melhor como ele resolve esses problemas.

Problema vinculativo

Embora as redes neurais tenham sido inspiradas pelo córtex visual humano, notamos cada vez mais que essa semelhança é apenas superficial. A principal diferença é que, além de reconhecer atributos visuais como linhas ou objetos, nosso cérebro também codifica as relações entre esses atributos - portanto, a linha faz parte do objeto. Isso nos permite atribuir significado aos padrões que vemos.

"Quando olhamos para um gato, vemos todas as características que formam os gatos e como elas se relacionam umas com as outras", diz Simon Stringer, da Fundação de Oxford para a Neurociência Teorética e a Inteligência Artificial. "Essa informação de 'vinculação' é o que garante nossa capacidade de compreender o mundo e nossa inteligência geral".

Essa informação crítica se perde na atual geração de redes neurais artificiais.

"Se você não solucionou essa vinculação, você pode saber que em algum lugar da cena há um gato, mas não sabe onde ele está e não sabe quais características na cena fazem parte daquele gato", explica Stringer.

Ao tentar manter as coisas simples, os engenheiros responsáveis pelas estruturas neurais artificiais ignoraram várias propriedades dos neurônios reais - cuja importância está começando a ficar clara.

"Os neurônios das redes artificiais são exatamente iguais, mas a variedade morfológica de neurônios no cérebro sugere que isso não é irrelevante", diz Jeffrey Bowers, neurocientista da Universidade de Bristol, que investiga quais aspectos da função cerebral não estão sendo capturados pelas redes neurais.

Seu laboratório desenvolve simulações de computador do cérebro humano para entender como ele funciona. Recentemente, eles incorporaram informações do timing e da organização dos neurônios reais, e treinaram o sistema com uma série de imagens. Com isso, já perceberam uma mudança fundamental na forma como suas simulações processavam informações.

Em vez de todos os neurônios dispararem ao mesmo tempo, eles começaram a notar padrões mais complexos de atividade. Por exemplo, um subgrupo de neurônios artificiais parecia agir como guardiões: só dispariam se os sinais visuais que recebessem chegassem ao mesmo tempo.

Stringer acredita que os "neurônios de vinculação" agem como uma certidão de casamento: eles formalizam as relações entre os neurônios e fornecem um meio de checar se dois sinais que parecem conectados realmente o estão. Dessa forma, o cérebro detecta se duas linhas diagonais e uma curva, por exemplo, representam uma característica como a orelha de um gato ou algo totalmente sem relação.

Redes híbridas

A equipe de Stringer busca evidências de tais neurônios em cérebros humanos reais. E também vem desenvolvendo redes neurais "híbridas", que incorporem as novas informações para ver se elas produzem uma forma mais robusta de aprendizado de máquina. Uma coisa que a equipe de Stringer testará é se as redes saberiam, de forma confiável, se uma pessoa idosa está caindo, simplesmente sentando-se ou colocando as compras no chão de casa.

"Esse ainda é um problema muito difícil para os algoritmos de visão artificial, enquanto que o cérebro humano pode resolver isso sem esforço", diz Stringer.

Ele também contribui com a pesquisa do Laboratório de Ciência e Tecnologia de Defesa em Porton Down, em Wiltshire, Inglaterra, que desenvolve uma versão ampliada de sua estrutura neural para a área militar, como localizar tanques inimigos de câmeras inteligentes instaladas em drones autônomos.

O objetivo de Stringer é, em 20 anos, ter garantido uma inteligência artificial no mesmo nível que a de ratos. E ele reconhece que o desenvolvimento de uma inteligência no nível humano pode levar uma vida inteira - talvez até mais.

Madry concorda que essa abordagem inspirada na neurociência é interessante para resolver os problemas com os atuais algoritmos de aprendizado de máquina. "Está ficando cada vez mais claro que a maneira como o cérebro funciona é bem diferente de como nossos modelos existentes de aprendizagem profunda funcionam", afirma.

"Então, isso pode acabar tomando um caminho completamente diferente. É difícil dizer o quão viável é e qual é o prazo necessário para alcançar o sucesso nesse caso", acrescenta.

Enquanto isso, talvez precisemos evitar confiar demais nos robôs, carros e programas alimentados por inteligência artificial aos quais estaremos cada vez mais expostos. Nunca sabe se é uma alucinação.

Leia a versão original desta reportagem (em inglês) no site BBC Future....

- http://www.bbc.com/future/story/20181204-why-we-should-worry-when-machines-hallucinate