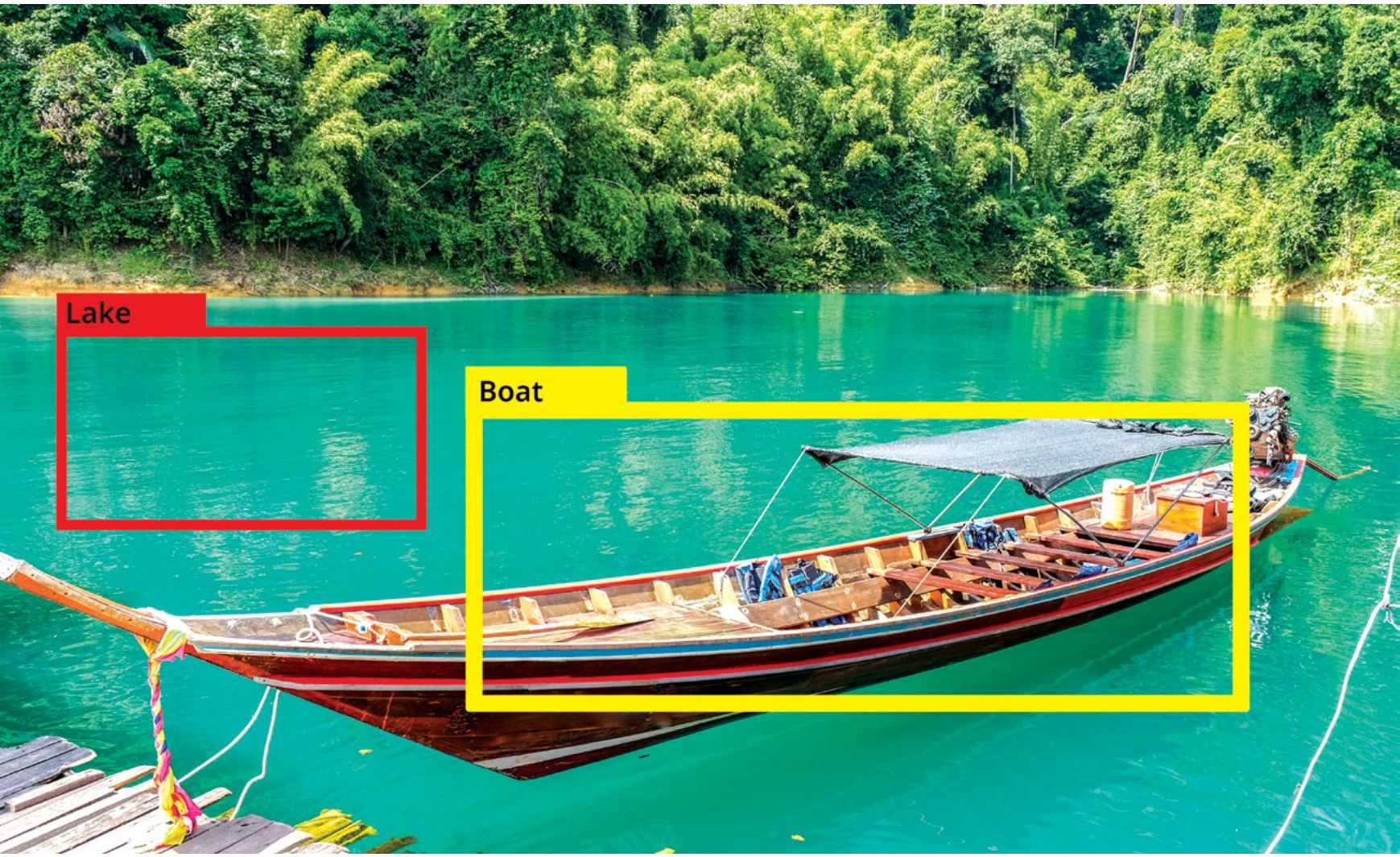


Las 10 bases de datos de IA más conocidas están llenas de errores de etiquetado

MIT Technology Review
Publicado por Opinno

KAREN HAO
TADUCIDO POR **ANA MILUTINOVIC**
12 ABRIL, 2021

Famosos conjuntos como ImageNet y MNIST, usados ampliamente para entrenar otros modelos, incluyen numerosas etiquetas incorrectas, según ha descubierto el MIT. El problema puede estar generando algoritmos defectuosos de forma inconsciente que luego se acaban aplicando en el mundo real.

Según el nuevo estudio del MIT, los 10 conjuntos de datos de inteligencia artificial (IA) más conocidos están plagados de errores de etiquetado, lo que distorsiona nuestra visión del progreso del campo.

El eje central de los datos: los conjuntos de datos son la espina dorsal de la investigación en IA, pero algunos son más críticos que otros. Hay un conjunto básico que los investigadores utilizan para evaluar los modelos de aprendizaje automático como forma de seguimiento de cómo avanzan las capacidades de IA con el tiempo. Uno de los más conocidos es el conjunto de datos de reconocimiento de imágenes ImageNet, que inició la revolución de la IA moderna. También está MNIST, que reúne las imágenes de números entre 0 y 9 escritos a mano. Otros conjuntos de datos ponen a prueba los modelos entrenados para reconocer audio, texto y dibujos.

Sí, pero: en los últimos años, los estudios han encontrado que estos conjuntos de datos pueden contener graves defectos. ImageNet, por ejemplo, contiene etiquetas racistas y sexistas, así como fotografías de rostros de personas obtenidas sin su consentimiento. El último estudio ha analizado otro problema: muchas de las etiquetas están completamente equivocadas. Un hongo está etiquetado como una cuchara, una rana se ha etiquetado como un gato y una nota alta de Ariana Grande como un silbato. El conjunto de datos de ImageNet tiene el estimado de la tasa de error de etiquetado del 5.8%. Por otro lado, el conjunto de datos de QuickDraw, que contiene dibujos a mano, tiene el estimado de la tasa de error del 10.1%.

Tasa de error estimada en etiquetas

¿CÓMO SE HA MEDIDO?

Cada uno de los 10 conjuntos de datos utilizados para evaluar los modelos tiene su correspondiente conjunto de datos de entrenamiento. Los investigadores, los estudiantes de posgrado del MIT Curtis G. Northcutt y Anish Athalye y el antiguo alumno Jonas Mueller, usaron los conjuntos de datos de entrenamiento para desarrollar un modelo de aprendizaje automático y luego lo utilizaron para predecir las etiquetas en los datos de prueba. Si el modelo no coincidía con la etiqueta original, ese punto de datos se marcaba para una revisión manual. Se pidió a cinco revisores humanos de Amazon Mechanical Turk que votaran sobre qué etiqueta, la del modelo o la original, pensaban que era la correcta. Si la mayoría de los revisores estaban de acuerdo con la etiqueta del modelo, la etiqueta original se consideraba un error y se corregía.

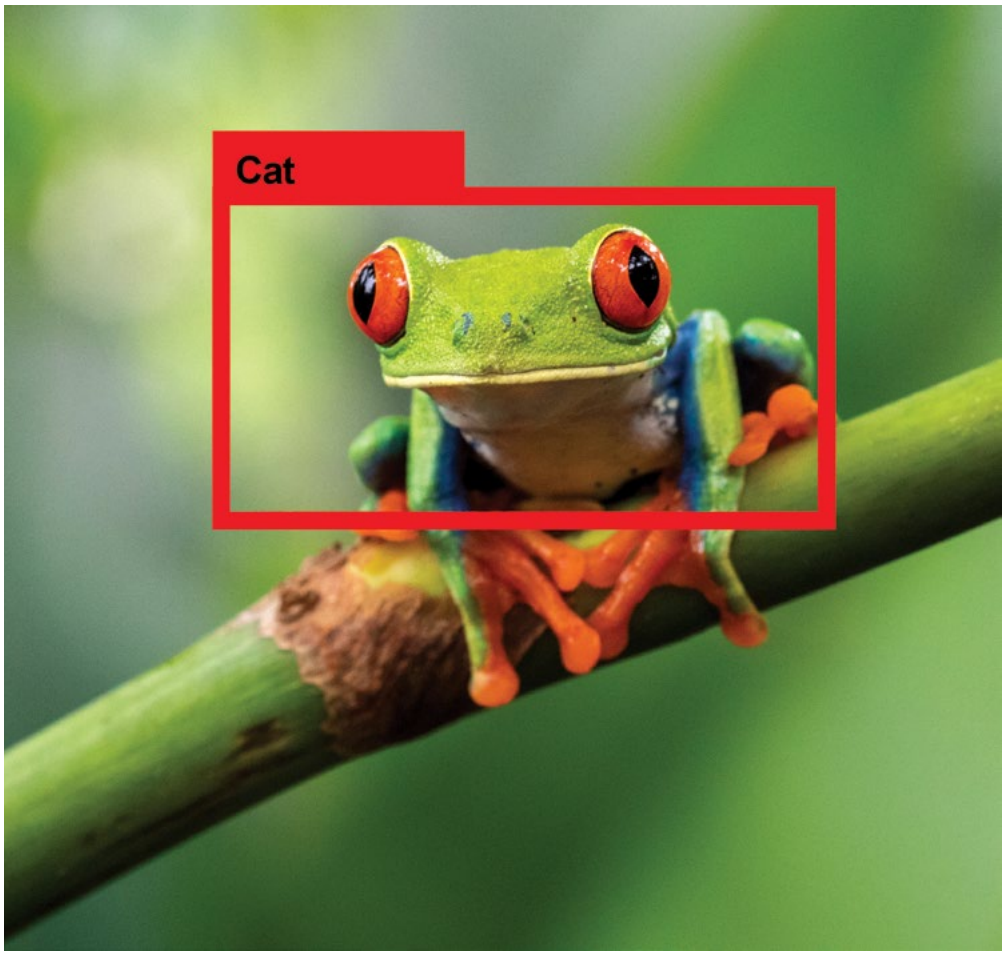
¿Y ESTO IMPORTA?

Sí. Los investigadores analizaron 34 modelos cuyo rendimiento se había medido previamente con el conjunto de datos de ImageNet. Luego,

si la mayoría de los revisores estaban de acuerdo con la etiqueta del modelo, la etiqueta original se consideraba un error y se corregía.

muchas de las etiquetas están completamente equivocadas. Un hongo está etiquetado como una cuchara, una rana se ha etiquetado como un gato y una nota alta de Ariana Grande como un silbato.

volvieron a medir cada modelo frente a los aproximadamente 1.500 ejemplos en los que se encontró que las etiquetas de datos eran incorrectas. Descubrieron que los modelos que no funcionaban tan bien en las etiquetas originales incorrectas se encontraban entre los que tenían mejor rendimiento después de que se corrigieran las etiquetas. En particular, los modelos más simples parecían tener mejores resultados en los datos corregidos que los modelos más complicados utilizados por los gigantes tecnológicos como Google para el reconocimiento de imágenes y que se supone que son los mejores en el campo. En otras palabras, es posible que tengamos una percepción inflada de lo geniales que son estos complicados modelos debido a los datos de prueba defectuosos.



¿Y AHORA QUÉ?

Northcutt anima al campo de la IA a crear conjuntos de datos más limpios para evaluar los modelos y seguir el progreso del campo. También recomienda que los investigadores mejoren la limpieza de sus datos cuando trabajen con sus propios datos. De lo contrario, concluye Northcutt, «si tenemos un conjunto de datos defectuoso y un montón de modelos que se están probando y se deberían implementar en el mundo real», podríamos acabar seleccionando el modelo incorrecto. Con este fin, el código que se usó en este estudio para corregir los errores de etiquetas es de acceso abierto, que según Northcutt ya se está utilizando en algunas de las principales empresas tecnológicas. **</>**

La autora es editora *senior* de Inteligencia Artificial en *MIT Technology Review*