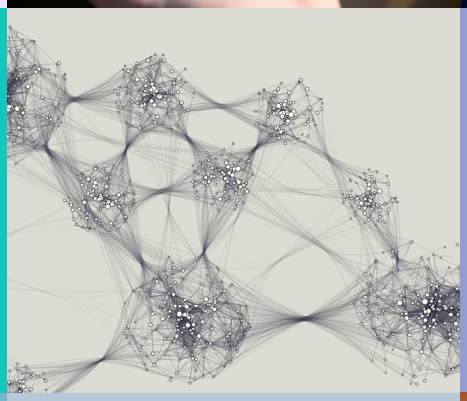
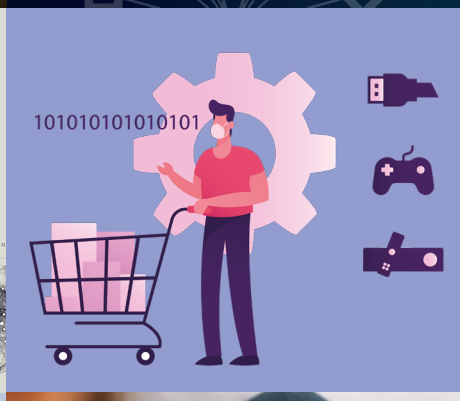
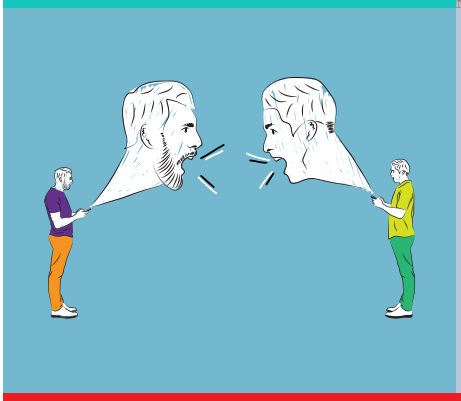
Aplicaciones de la inteligencia artificial







A través de una serie de miniartículos independientes, ilustramos cómo la inteligencia artificial y sus diferentes métodos permiten abordar problemas en una amplia y creciente diversidad de dominios. Por cuestiones de extensión, la enumeración no pretende ser exhaustiva y muchas áreas quedarán pendientes para una futura edición de la Revista.

¿Puede una máquina ver mejor que un humano?

JAVIER CARRASCO Ingeniero Civil en Computación de la Universidad de Chile y egresado del Instituto Milenio Fundamentos de los Datos.

AIDAN HOGAN Profesor Asociado del Departamento de Ciencias de la Computación de la Universidad de Chile e Investigador Asociado del Instituto Milenio Fundamentos de los Datos.

JORGE PÉREZ Profesor Asociado del Departamento de Ciencias de la Computación de la Universidad de Chile e Investigador Asociado del Instituto Milenio Fundamentos de los Datos.

La última década ha sido testigo de avances extraordinarios en el área de la inteligencia artificial, impulsados, en particular, por el concepto de redes neuronales profundas, combinado con la disponibilidad de enormes cantidades de datos para entrenar estas redes. Entre las subáreas de la computación que se han beneficiado con esta tecnología, podemos destacar, por ejemplo, la visión computacional, y la tarea específica de reconocimiento de imágenes. En esta tarea, la máquina recibe una imagen de un objeto y tiene que devolver la clase de ese objeto, diciendo, por ejemplo, que la imagen representa un perro, una flor, una taza, etc.

El conjunto de datos más usado para entrenar y evaluar métodos de reconocimiento de imágenes se llama ImageNet; contiene millones de imágenes etiquetadas según mil clases distintas. Según Russakovsky et al. [1], un experto humano puede lograr una tasa de error (top-5) de 5,1% en un subconjunto de 1.500 imágenes de ImageNet. En la misma tarea, una red neuronal profunda del estado del arte (SeNetResNet50 [2]) puede lograr una tasa de error (top-5) de 2,3%, es decir que tiene mejor rendimiento que un humano experto en esta tarea. ¿Este resultado significa que las máquinas, ahora, pueden "ver" mejor que los humanos? No necesariamente, pues es una pregunta multifacética. En esta tarea, las clases son muy finas, e incluyen ejemplos como un *cucal*, un *Sealyham terrier*, etc., que pueden ser difíciles de recordar y distinguir para un humano. También, la tarea siempre considera imágenes de calidad total. Entonces surge una duda: si las imágenes tuvieran menos calidad que las vistas en los ejemplos de entrenamiento, ¿cómo afectaría el rendimiento de las máquinas y de los humanos? ¿Los humanos necesitan más o menos información para poder clasificar una imagen correctamente en comparación con las máquinas? ¿Qué tipo de información les importa más?

Imágenes mínimas positivas

Para poder entender y comparar la dependencia que las máquinas y los humanos tienen para poder clasificar bien una imagen, definimos el concepto de una *imagen mínima positiva* [3]: dada una imagen etiquetada con su clase, y un clasificador de imágenes, la imagen mínima positiva es la versión de la imagen con la peor calidad tal que el clasificador siga dando la clase correcta. Con respecto a la calidad de la imagen, hablamos más específicamente de

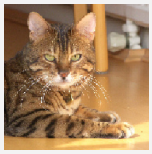
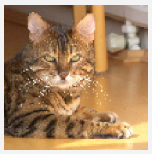
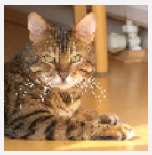
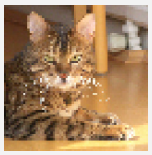
Modelo	Color	Resolución	Zona	Combinación
SqueezeNet				
GoogLeNet				
ResNet50				
SeNetResNet50				
Humano				

Figura 1. Imágenes mínimas positivas para un gato.

la información que contiene, medida usando el tamaño de la imagen comprimida (sin pérdida; usamos compresión de PNG). Se pueden considerar varias formas de reducción de imágenes; en nuestro trabajo, hemos considerado las reducciones de color, de resolución, de zona, y la combinación de las tres. La tabla de la Figura 1 ejemplifica las imágenes mínimas para una imagen de un gato, tal que el modelo (clasificador) indicado puede reconocer que la imagen es de un gato, pero con más reducción, no puede más.

Para calcular las imágenes mínimas en el caso de las máquinas, tomamos una imagen de prueba (no vista antes durante el proceso de entrenamiento), e implementamos una búsqueda sobre los parámetros de reducción, empezando con la imagen completa, y reduciendo la información hasta que se encuentre la imagen mínima. Para calcular las imágenes mínimas en el caso de las máquinas, no se puede usar la misma estrategia, pues el humano recordará la clase de la imagen completa. Así que diseñamos una interfaz que empieza con la imagen "nula" (con una reducción completa), tal que el humano pueda aumentar la información hasta que pueda reconocer el objeto de la imagen y clasificarla (si la clasificación es incorrecta, descartamos la imagen y pasamos a la próxima).

Experimentos y resultados

Para ver qué tan sensibles son los clasificadores frente a la pérdida de diferentes tipos de información, hicimos experimentos con 20 clases simplificadas de ImageNet, tomando 15 imágenes para cada clase. Tomamos cuatro modelos que usan redes neuronales profundas, que han logrado el mejor resultado sobre ImageNet en algún momento, y que han sido entrenados con las imágenes (completas) de entrenamiento de ImageNet. Los cuatro modelos, en orden de su rendimiento sobre ImageNet, son SqueezeNet, GoogLeNet, ResNet50, y SeNetResNet50. Se pueden ver ejemplos de las imágenes mínimas de cada modelo en la Figura 1 considerando varias formas de reducción.

Luego medimos la proporción de reducción para las imágenes mínimas positivas como el cociente entre el tamaño de la imagen original y la imagen mínima positiva (ambas comprimidas con PNG). Un menor cociente significa que el modelo es más robusto a la pérdida de información correspondiente. En la Figura 2, podemos ver los resultados, presentados como un diagrama de caja. Se puede ver que los humanos son mejores para clasificar imágenes con menos colores y resolución, pero que las máquinas pueden clasificar las imágenes basado en zonas más pequeñas. Estos resultados apoyan la observación de Geirhos *et al.* [4] de que la textura de la imagen es una característica importante para las redes neuronales profundas, las cuales pueden diferenciar, por ejemplo, entre el pelo de un gato y un perro. Por eso sólo necesitan una zona pequeña de una imagen, pero sufren más con una pérdida de resolución o color. Otra observación es que los modelos más robustos frente a la pérdida de información también tienen mejor rendimiento para las imágenes completas.

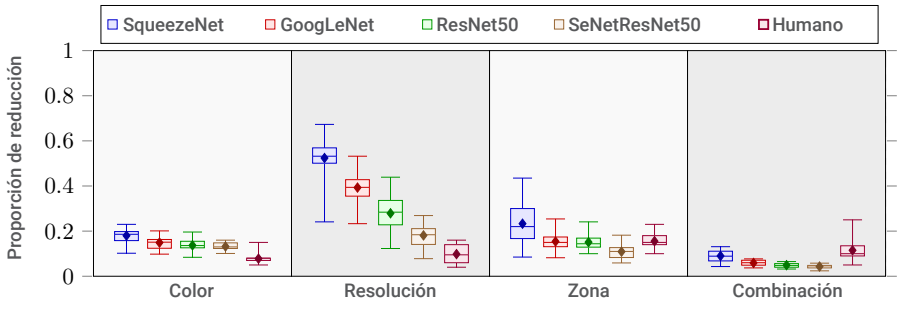


Figura 2. Proporción de reducción para las imágenes mínimas positivas.

Finalmente, hicimos un experimento usando cada clasificador para clasificar las imágenes mínimas positivas de los otros clasificadores. Se pueden encontrar los resultados completos en nuestro artículo [3]. En resumen, observamos que los humanos pueden clasificar mejor las imágenes mínimas positivas de las máquinas que al revés, logrando una precisión de 0,89-0,92 para color, 0,86-0,93 para resolución, 0,76-0,87 para zona, y 0,74-0,85 para combinación, con mejor precisión para las imágenes mínimas positivas, respectivamente, de SqueezeNet (más fáciles), GoogLeNet, ResNet50, y SeNetResNet50 (más difíciles). Al revés, clasificando las imágenes mínimas positivas de los humanos, los modelos de máquina lograron una precisión de 0,14-0,42 para color, 0,03-0,29 para resolución, 0,11-0,42 para zona, y 0,07-0,35 para combinación; los mejores modelos fueron, respectivamente, SeNetResNet50 (mayor precisión), ResNet50, GoogLeNet y SqueezeNet (menor precisión).

Conclusiones

¿Puede una máquina ver mejor que un humano? Es una pregunta cada vez más compleja, que puede ser interpretada de varias formas. En la Clasificación de Imágenes, nuestros resultados han indicado que los humanos proveen resultados más robustos frente a la pérdida de información. En la práctica, esto implica que los resultados dados por las redes neuronales profundas entrenadas y evaluadas en el contexto de conjuntos de imágenes completas pueden no aplicarse a condiciones reales, en las cuales un objeto (por ejemplo, una cara) está parcialmente oculto, o está a distancia, o iluminado parcialmente, etc.

Una pregunta que nos interesa ahora, entonces, es la siguiente: ¿se puede mejorar la robustez de los clasificadores de máquinas frente a la pérdida de información? Los modelos que usamos en este trabajo fueron entrenados sobre imágenes completas. Quizás se puedan entrenar las redes con imágenes reducidas o mínimas, para mejorar su robustez en situaciones de información parcial. ■

REFERENCIAS

[1] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, y Fei-Fei Li. 2015. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision 115, 3 (2015), 211–252.

[2] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, y Enhua Wu. 2019. Squeeze-andExcitation Networks. arXiV:1709.01507v4.

[3] Javier Carrasco, Aidan Hogan y Jorge Pérez. 2020. Laconic Image Classification: Human vs. Machine Performance. En el acta de la International Conference on Information and Knowledge Management (CIKM), Galway, Ireland, [Online], October 19–23, 2020.

[4] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, y Wieland Brendel. 2019. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. En el acta de la International Conference on Learning Representations (ICLR). OpenReview.net.