Aplicaciones de las redes neuronales profundas

El uso de redes neuronales profundas en la creación de aplicaciones de inteligencia artificial.

Alejandro Correa B.

En la actualidad, vemos un gran auge alrededor de las redes neuronales profundas, teniendo en cuenta que están siendo empleadas para crear toda clase de modelos, desde conducción autónoma hasta seguridad informática.

No obstante, a pesar de que las ideas sobre redes neuronales han estado presentes por décadas ¿por qué hasta hace poco comenzaron a ser utilizadas ampliamente? La respuesta es simple, escalabilidad y big data.

En los comienzos de big data, los algoritmos tradicionales de aprendizaje de máquina siempre mostraron el mejor rendimiento. Esto se debe a que los algoritmos tradicionales como bosques aleatorios, máquinas de soporte vec-

torial e incluso regresión logística mejoran más rápidamente, cuando se agregan más datos; sin embargo, su rendimiento disminuye después de recibir millones de ejemplos cuando el incremento marginal en rendimiento es mínimo. Aun así, tal como se muestra en la gráfica 1, al replicar el proceso con redes neuronales profundas, el rendimiento continúa creciendo, incluso si se agregan más ejemplos.

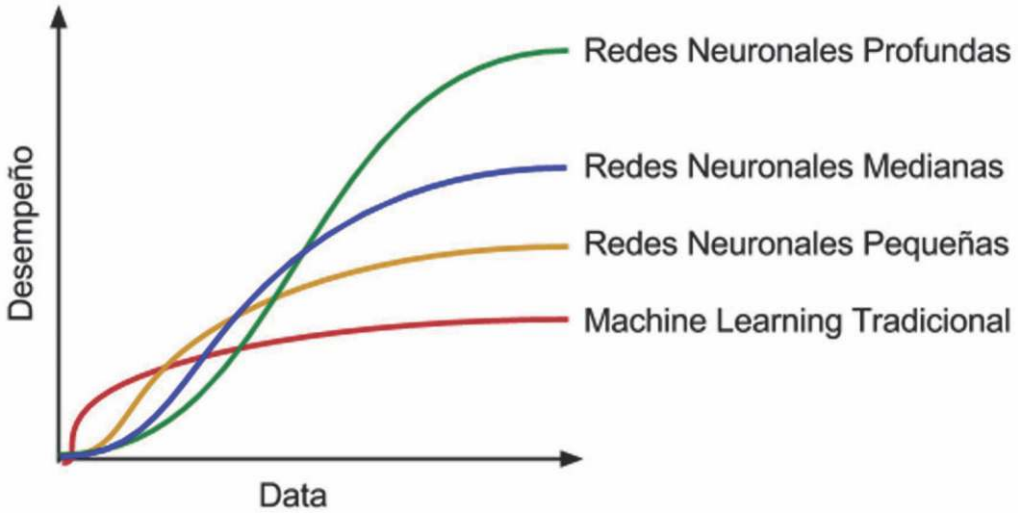
Esta explosión de datos y la disponibilidad de poder de cómputo más asequible abrieron las puertas a un creciente uso de redes neuronales profundas. Sin embargo, no es nada fácil comenzar a utilizar estos nuevos modelos. Por ejemplo, en el momento de hacer modelos de aprendizaje profundo, es difícil decidir entre una enorme cantidad de categorías de arquitecturas de redes. En particular, existen cuatro principales categorías de modelos de aprendizaje profundo:

1. Redes neuronales densas
   - Perceptron multicapa
2. Modelos de secuencia (1D)
   - Redes neuronales recurrentes
   - Unidades recurrentes cerradas
   - Redes neuronales de memoria a corto y largo plazo
   - Modelos de atención
3. Modelos de imágenes (2D and 3D)
   - Redes neuronales convolucionales
4. Tecnología futura / avanzada
   - Aprendizaje no supervisado: Codificación dispersa, ICA, SFA…
   - Aprendizaje por reforzamiento

Elegir la más adecuada según cada aplicación no es nada sencillo y normalmente depende del analista, decidir el camino a elegir. Lo interesante de esto es que casi toda la industria está siendo conducida por las primeras tres categorías.

Aquellos tres conjuntos son los que están creando mejores modelos y productos. No obstante, al ver los documentos presentados en la Conferencia de Sistemas de Procesamiento de Información Neuronal, NIPS, 2016, en Barcelona, la mayoría de las investi-

Gráfica 1



gaciones están enfocadas en el último grupo, sugiriendo de cierta forma que el próximo gran avance estará dirigido en modelos no supervisados y de aprendizaje por refuerzo.

Gestión de Proyectos de IA

Últimamente, una tendencia que cada día gana más adeptos es el cambio en los típicos flujos de trabajo para hacer que los equipos trabajen colaborativamente, en la elaboración de aplicaciones con aprendizaje profundo. En diseño de *software* tradicional, el Product Manager (PM) transfiere las necesidades al desarrollador, quien usa su conocimiento para cumplir con los requerimientos del *software*.

Por otra parte, en el caso de la gestión de productos Inteligencia Artificial (IA), hay un aporte adicional que el PM debe brindar al ingeniero de IA, como proporcionar conjuntos de datos de desarrollo o pruebas que producirán resultados útiles. Así mismo, el PM debe proporcionar las métricas de evaluación para el algoritmo de aprendizaje.

Los PM han sido una parte integral del desarrollo de *software* y productos en los últimos años. Con el fin de obtener buenos resultados en esta nueva era de productos IA, el rol del PM debe evolucionar y tener en cuenta las particularidades de estos nuevos productos. Depende del PM definir y evaluar el rendimiento del algoritmo en un conjunto realista de datos.

Aprendizaje End-to-End

Además del uso de enormes conjuntos de datos, y la nueva gestión de proyecto de IA, la tercera tendencia más importante que se observa es el creciente uso de modelos de aprendizaje end-to-end. Los modelos tradicionales dependen de la ingeniería de variables, en la cual los científicos de datos dependen del conocimiento externo y experto para crear variables relevantes a un problema dado.

Esto funciona en muchas aplicaciones como, por ejemplo, el reconocimiento de voz. Aquí, el modelo tradicional se fía de variables diseñadas manualmente, pero los enfoques modernos le permiten al modelo aprender de todas sus interacciones internas.

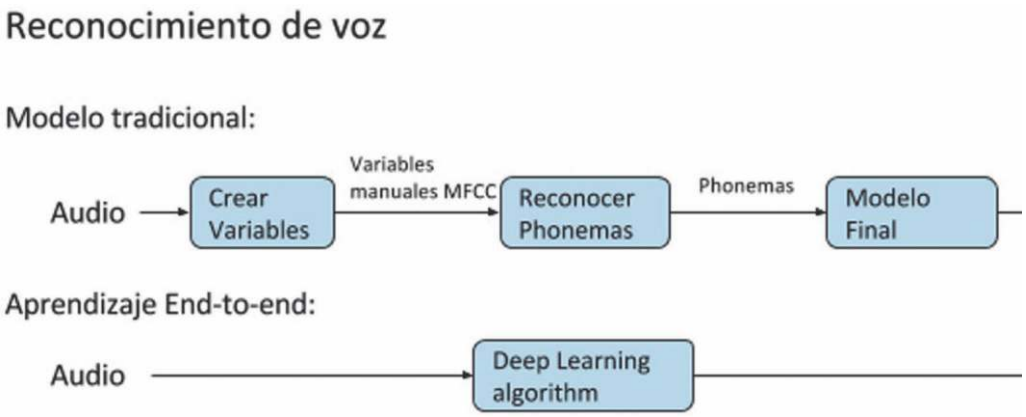
De igual manera, los modernos modelos de conducción autónoma se basan en enfoques end-to-end.

Caso de Estudio: Detección the URLs Malignas (Phishing)

Gráfica 2

Responsabilidad del Product Manager	Responsabilidad del Ingeniero / Científico de datos
• Proporcionar conjuntos de datos de desarrollo/pruebas, idealmente sacados de la misma distribución. • Proporcionar evaluación métrica para algoritmos de aprendizaje.	• Adquirir datos de capacitación. • Desarrollar sistemas que funcionen de acuerdo con la métrica dada en el conjunto de datos.

Gráfica 3



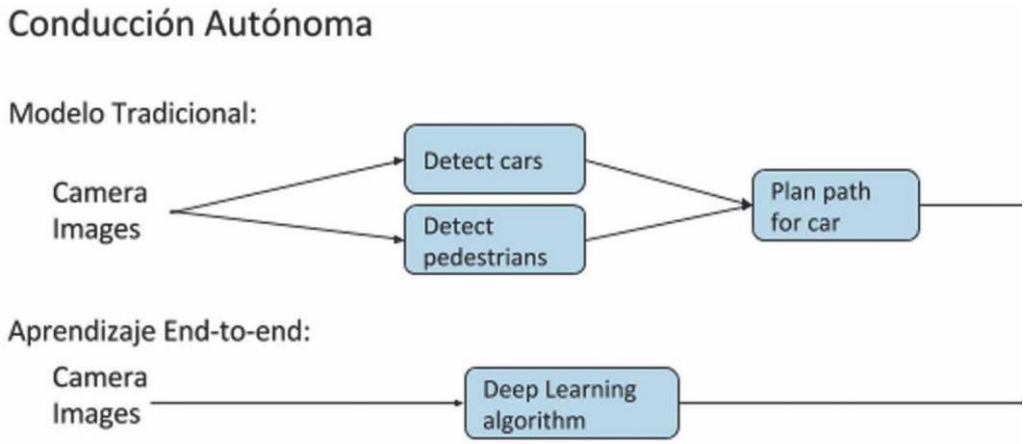
Ejemplos como los mostrados anteriormente abundan en diferentes industrias. Como caso de estudio, comparamos los resultados de un modelo de clasificación tradicional para detección de URLs de *phishing* versus un modelo de redes neuronales profundas end-to-end.

Detectar páginas de *phishing* usando solamente la URL genera una ventaja competitiva, toda vez que analizar todo el contenido de una página web (imágenes, texto, enlaces, etc.) conlleva un importante costo computacional.

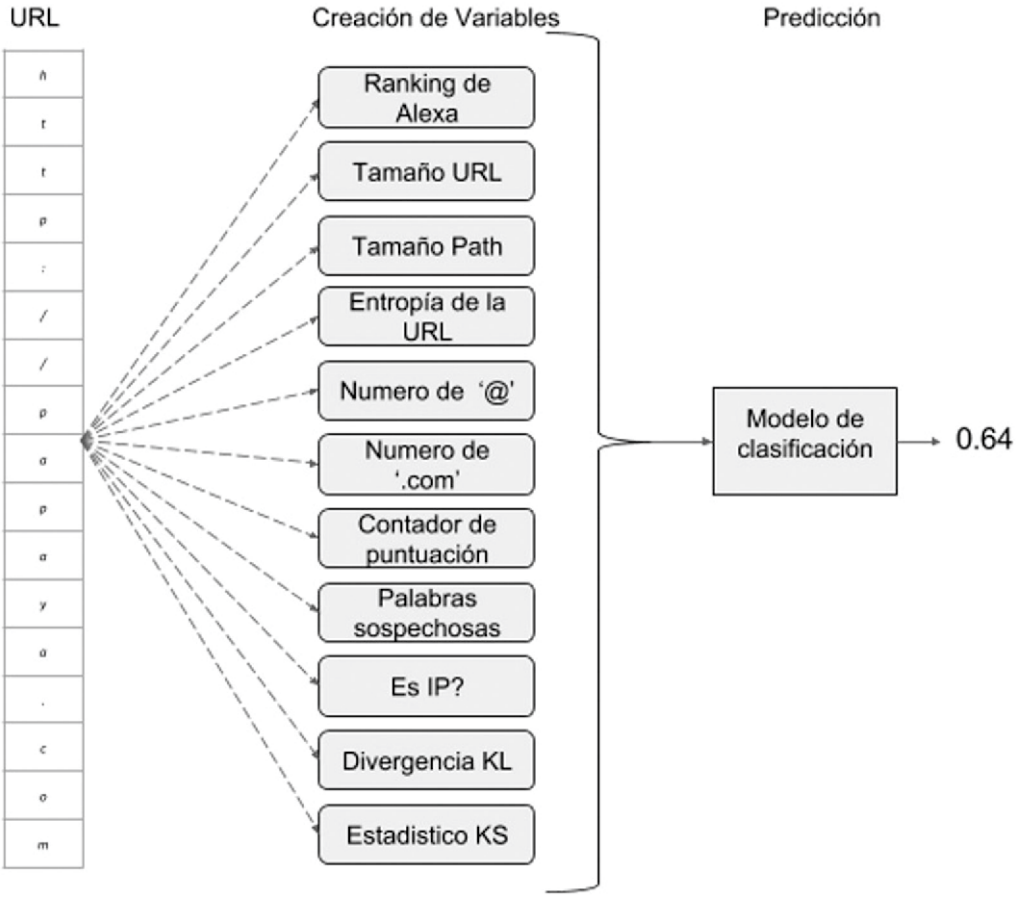
Teniendo en cuenta que diariamente se tienen que analizar hasta 5.000.000 de sitios web, los costos de renderizar y procesar todas las páginas serían prohibitivos.

El método tradicional de análisis de URL de *phishing* consiste en extraer variables de las URL, mediante conocimiento experto y análisis del texto de la URL. Posteriormente, se entrena un algoritmo de clasificación, como lo es una máquina de vectores de soporte o un bosque aleatorio. Este proceso se describe en la gráfica 5.

Gráfica 4



Gráfica 5



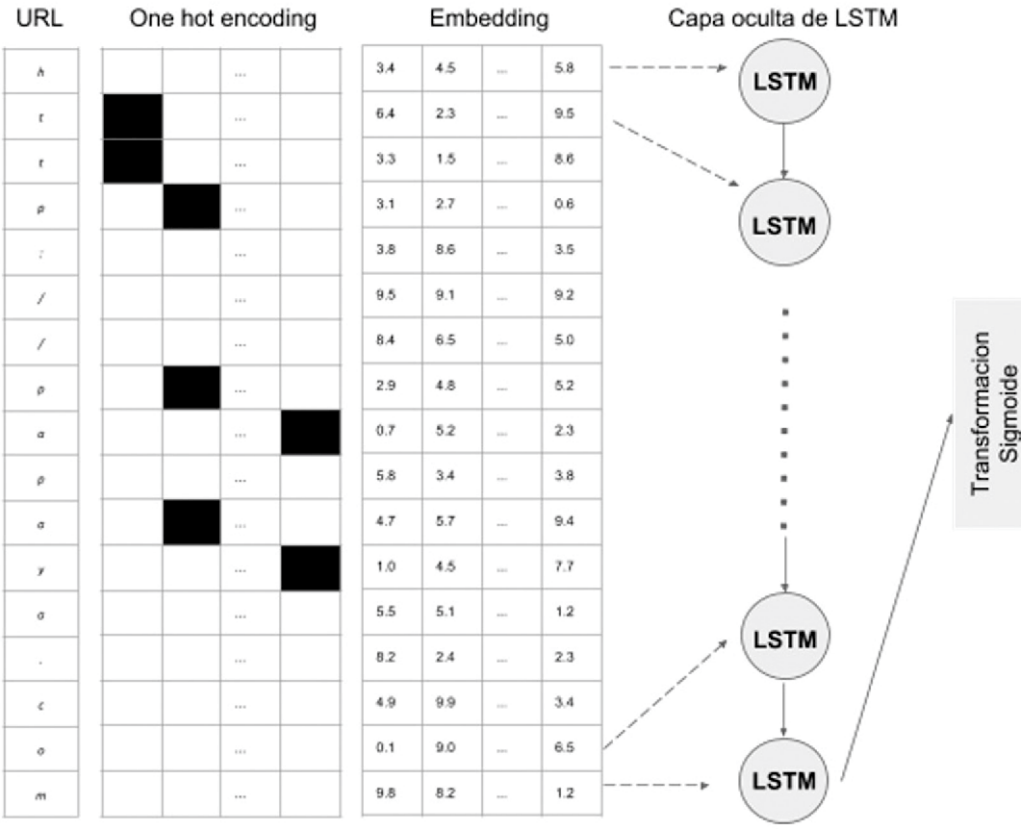
La creación de variables correctas consume mucho tiempo, recursos y personal. Por esta razón, emplear un modelo de redes neuronales profundas, como una red de memoria a corto y largo plazo (LSTM), reduce significativamente el proceso de modelamiento porque este algoritmo utiliza la secuencia de caracteres e internamente hace un "embedding" para aprender los patrones de las URLs, tanto de *phishing* como legítimas. Este proceso se muestra en la gráfica 6.

La creación de este algoritmo es mucho más fácil al no requerir de ninguna variable. Comparando los resultados de ambas metodologías, se encontró que el modelo tradicional tiene un poder de predicción del 93.5%, en una base de datos compuesta por un millón de URL de *phishing* y un millón de URL legítimas. Por otro lado, el modelo end-to-end tiene un poder de predicción del 98.7%, sobrepasando en forma significativa al modelo anterior.

Conclusión

Se espera que estas metodologías de aprendizaje profundo, hasta hace poco sólo conocidas en el medio acadé-

Gráfica 6



mico y ciertos sectores de la industria, se generalicen ampliamente. De ahí la importancia de contar con personal altamente capacitado, no sólo en la parte técnica, sino también en el área de gerencia de proyectos de inteligencia artificial.

Alejandro Correa B. *Chief Data Scientist en Easy Solutions. Con un PhD en Machine Learning de la Universidad de Luxemburgo. Cuenta con varios años de experiencia en el uso y desarrollo de modelos aplicados a problemas como algorithmic trading, detección y prevención de fraudes, riesgo de crédito, seguridad informática, HR analytics, cobranzas, mercadeo y trading algorítmico. Ha escrito y publicado artículos académicos en las mejores revistas internacionales de Machine Learning, además de haber sido conferencista en importantes eventos académicos y de negocios. Durante los últimos años, ha dedicado su atención al desarrollo de herramientas para la detección y prevención de fraude en proyectos internacionales. Sus conocimientos, le han permitido desempeñarse como profesor en diferentes asignaturas, particularmente en Machine Learning, Big Data, Econometría y Analytics, en la Universidad de los Andes. Es fundador de la comunidad Big Data & Data Science Bogotá y colaborador de herramientas open-source como Scikit-Learn.*