The Limitations of Opaque Learning Machines

Centro de Investigación en Computación, IPN

Laboratorio de Ciencias Cognitivas Computacionales

Miguel Angel Soto Hernandez

Con los modelos de aprendizaje profundo se esta perdiendo la transparecncia que solíamos tener por ejemplo en las Máquinas de Turing, ya que básicamente es un ejercicio de un ajuste de curvas que ajusta pesos entre capas de una larga cadena entrada-salida.

La mayoría de los modelos de aprendizaje profundo funcionan todavía como cajas negras, donde no se sabe a ciencia cierta que es lo que esta pasando detrás de una arquitectura neuronal artificial y muchos usuarios o desarrolladores de estas tecnologías argumentan que su modelo "funciona bien" aunque no saben porque. Y de cierta manera si el modelo esta bien hecho esto puede ser verdad, ya que estos modelos realizan su propia dinamica. Desde realizar su propia reparación, optimización y la salida de datos que dependiendo el ajuste puediesen estar bien. Pero, ¿Qué pasa si el resultado es erróneo? Aquí es donde empiezan a surgir los problemas, ya que no se sabe si esto fue un fallo del programa, del método o porque las caractrísticas no se eligieron bien. No se sabe.

Por este motivo, el argumento "funciona bien" tiene un límite, y a partir de esto podemos intentar otro enfoque, en esta ocasión, considerar el razonamiento causal, ya que se han descubierto que existen algunas barreras básicas y que, a menos que se superen, no se conseguirá un tipo de inteligencia humana real. Entonces, ¿Qué pasaría si un modelo de IA pudiese "razonar"? En principio para poder lograr un entendimieto de nivel humano, las máquinas de aprendizaje necesitan la guía de un plano de la realidad, un modelo, similar al mapa de carreteras que nos guía al conducir por una ciudad desconocida. Para ser más específicos, las máquinas de aprendizaje actuales mejoran su rendimiento optimizando los parámetros para un flujo de entradas sensoriales recibidas del entorno. Sin embargo, este es un proceso lento. Es por esto por lo que se tendrían que realizar preguntas del tipo: ¿Y si? ¿Qué pasa si? ¿Y si hago esto? ¿Y si lo hibiera hecho de alguna otra manera? Pero,ninguna máquina de aprendizaje en funcionamiento hoy en día puede responder a

estas preguntas. Además, la mayoría de las máquinas de aprendizaje no poseen una representación de la que puedan derivarse las respuestas a esas preguntas.

Con respecto al razonamiento causal, encontramos que se puede hacer muy poco con cualquier forma de ajuste de curvas a ciegas del modelo, o cualquier inferencia estadística, sin importar lo sofisticado que sea el proceso de ajuste. También hemos encontrado un marco teórico para organizar esas limitaciones, que forma una jerarquía.

1. **Asociación**
2. **Intervención**
3. **Contrafactuales**

Uno de los mayores logros de la investigación sobre la inferencia causal ha sido la algoritmización tanto de las intervenciones como de los contrafactuales, las dos capas superiores de la jerarquía. En otras palabras, una vez que codificamos nuestros conocimientos científicos en un modelo, existen algoritmos que examinan el modelo y determinan si una consulta dada, ya sea sobre una intervención o sobre un contrafactual, puede estimarse a partir de los datos disponibles y, en caso afirmativo, cómo.

Usar modelos de IA y aprendizaje automático nos sirven por el momento para llevarnos de la mano de los datos hacia las probabilidades. Sin embargo, aún hay mucho trabajo por realizar para poder pasar de estas probabilidades a el entendimiento real. Estos pasos pueden ser, primeramente predecir el efecto de las acciones, y por otro lado la imaginación cantractural, ya que no podríamos pretender entender la realidad a menos que se tomen en cuenta estos pasos. Los enfoques ciegos a los modelos imponen limitaciones intrínsecas a las tareas cognitivas que puede realizar la IA fuerte. La IA de nivel humano no puede surgir únicamente de las máquinas de aprendizaje ciegas a los modelos sino que se requiere la colaboración de datos y modelos.