NOMBRE: Benjamín Farías Valdés

N.ALUMNO: 22102671

Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

IIC3692 — Tópicos Avanzados en Inteligencia Artificial — 2' 2022

Lectura 5

Crítica

Shortcut Learning in Deep Neural Networks

El paper trata sobre un aspecto interesante del *Deep Learning*: las limitaciones de los métodos actuales al aprender a generalizar. La introducción es muy buena, ya que muestra una conexión en términos del aprendizaje entre las redes neuronales biológicas y las artificiales. Se ejemplifican casos en que los humanos y otros animales aprenden a resolver ciertos problemas de forma superficial, tomando atajos, en vez de buscar una solución general (que justamente es el problema detectado en las redes neuronales artificiales). A continuación se clasifican los tipos de reglas de decisión que aprenden los modelos, dejando claro que se desea lograr un subconjunto de estas que funcione bien en cualquier tipo de contexto (incluyendo a datos fuera de la distribución del entrenamiento). Tras presentar el problema, se habla de las posibles causas de estos atajos en el aprendizaje: dataset sesgado bajo algún aspecto, percepción humana de los datos, tipo de modelo usado (en una sección se ejemplifica todo esto en los contextos más populares del aprendizaje de máquina). En los últimos capítulos se concluye que es importante continuar con la investigación de estas debilidades de los modelos, y buscar que los sets de testing usados para *benchmarking* sigan distintas distribuciones y sean generalizables al mundo real. Personalmente, me hace mucho sentido lo indicado en el paper, ya que es común encontrar soluciones fáciles a problemas y que terminen siendo superficiales y fallen al momento que ocurre un ligero cambio en el contexto de interés. Por lo mismo, concuerdo con que el paso siguiente es un cambio en el paradigma de entrenamiento de estos modelos, donde el enfoque esté en probar el rendimiento en casos no vistos y que además tengan variabilidad respecto de los casos de entrenamiento (que no sean muestras de la misma distribución). Finalmente, mi opinión general del artículo es que está muy bien pensado y construido, permitiendo que sea fácil comprender los temas hablados.