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 6

Crítica

Understanding Deep Learning Requires Rethinking Generalization

A diferencia de la mayoría de los papers leídos en este curso, este en particular opta por no proponer una solución/mejora a un problema y en su lugar se enfoca en tratar de explicar el poder de generalización de las redes neuronales profundas. En la introducción se aclara que el punto de partida son los diversos modelos de hoy en día que son capaces de cumplir sus tareas asignadas con gran precisión (en el artículo se ejemplifica con clasificación de imágenes). El grueso del artículo consiste en mostrar, de forma experimental, que lo que hasta el momento se consideraba poder de generalización (buena precisión en el set de *testing*), no es necesariamente tal. Se realizan experimentos alterando los datos de entrada y/o las etiquetas de entrenamiento, de forma que se vuelvan aleatorios y ruidosos (todo esto es explicado de forma fácil de entender). Se encuentra que los modelos son capaces de ajustarse a estas alteraciones sin problemas, mostrando que en estos casos la precisión es muy buena pero realmente no está generalizando nada (se aprendió todo de memoria). Este resultado es poco intuitivo, ya que se espera que al aprender de datos que no tienen sentido, al modelo le cueste más clasificar (o se demore más en converger), pero resulta que no tiene problemas y además es igual e incluso más rápido en algunas pruebas. El resto del paper presenta y analiza distintas métricas, llegando a que actualmente no existen formas de medir la capacidad de generalización real de los modelos (bajo el concepto de generalización que usamos los humanos). Este trabajo me pareció muy intrigante, ya que de cierta forma contradice la creencia de que las redes convolucionales van aprendiendo las características de las imágenes poco a poco (detectando patrones cada vez más complejos), dado que al cambiar los datos por ruido se obtiene también un buen clasificador que no hace uso de ningún tipo de patrón semántico (sino más bien aplica fuerza bruta y memoriza todo). Quizá los modelos de *Deep Learning* son capaces de aprovechar los patrones en los datos siempre que existan, pero también pueden simplemente memorizar cuando se necesario, lo que da mucho lugar a futura investigación en pos de encontrar métricas más explicables sobre la generalización efectiva de los modelos (y quizás formas de controlar la generalización vs memorización mediante parámetros).