PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
DEPARTAMENTO DE CIENCIA DE LA COMPUTACIÓN, ESCUELA DE INGENIERÍA
IIC3692 TÓPICOS AVANZADOS EN INTELIGENCIA ARTIFICIAL
PROFESOR: ÁLVARO SOTO
[redacted]

Lectura crítica: "The Lottery Ticket Hypothesis"

Jonathan Frankle y Michael Carbin, 2018

La investigación de Jonathan Frankle y Michael Carbin de 2018 es una pieza clave no sólo en la evolución de las técnicas contemporáneas de inteligencia artificial, si no en nuestro propio entendimiento de los mecanismos detrás del éxito del aprendizaje profundo, marcando un antes y un después.

Sin dudas el avance más destacable es la aplicación de *iterative pruning* o podado iterativo. Sin embargo, me llama la atención que los autores sólo proponen una máscara *m* posible al experimentar, lo cual resalta enseguida como una posible limitación para comprender cuál es el impacto de esta discriminación por ejemplo en la re-inicialización de los pesos. Este punto sí es abarcado por Zhou et.al. en el artículo "Deconstructing Lottery Tickets: Zeros, Signs, and the Supermask", publicado en 2019 bajo el alero de Uber AI Labs.

En su artículo proponen otros criterios adicionales para definir la máscara experimentando con 9 alternativas, como por ejemplo evaluar los pesos según su magnitud inicial en vez de la final; qué tanto varió o incrementó la magnitud del peso; o incluso elegirlos de forma random. Me parece que uno de los aspectos más interesantes de sus experimentos es la capacidad que tienen los pesos de cambiar de signo en el tiempo, siendo más probable que los pesos cambien a que incrementen en el tiempo. En efecto, sus resultados demuestran que las máscaras que mantienen los pesos cuya magnitud aumentó más se desempeñan igual o mejor que la máscara propuesta en el paper original, la cual conserva los pesos con un alto valor final.

Uno de los puntos importantes de la hipótesis original de Frankle y Carbin es que, al re-inicializar al azar los parámetros de la red podada o *sparce*, los tickets ganadores ya no logran igualar la performance de la red original. Nos damos cuenta entonces que para que el método funcione, no se pueden re-inicializar los pesos de forma aleatoria, sino darle a cada conexión el peso inicial exacto que tenía cuando era parte de la red completa.

Sin embargo, pude ver que no se evidencia a través de sus experimentos si el llevar a cero los pesos o congelarlos en su valor inicial random lleva a mejorar el desempeño de las redes podadas. De forma similar, tampoco las conclusiones de Zhou et. al. aportan mayor conocimiento de por qué este método funciona, en tanto sólo concluyen que las máscaras tenderían a llevar a cero aquellos pesos que se dirigían a cero de todas formas.

Tanto el artículo original como su revisión presentan interesantes implicancias, en tanto el procedimiento de asignar la máscara al peso puede ser concebido como un entrenamiento en sí.

A mi parecer quedan abiertas varias incógnitas, como por ejemplo, ¿cómo responderían estas redes a *transfer learning?,* ¿qué pasaría al aplicar procesos de *fine tuning* con estas subredes?

Es posible, y no queda demostrado en ambos artículos, que los tickets de lotería estén sobre ajustándose al set de test o de validación de los conjuntos probados, no sólo en sus pesos sino también en su estructura.