La huella de carbono del aprendizaje profundo

IVAN SIPIRÁN

Profesor Asistente del Departamento de Ciencias de la Computación de la Universidad de Chile. Obtuvo su Doctorado en Ciencias de la Computación en la Universidad de Chile el año 2014. Posteriormente realizó una estancia postdoctoral en la Universidad de Konstanz en Alemania. Sus áreas de investigación son procesamiento geométrico y análisis de formas, visión computacional 3D y computación gráfica aplicada a la herencia cultural.

isipiran@dcc.uchile.cl

@isipiran

RESUMEN. Los modelos de aprendizaje profundo son cada vez más ubicuos. Empresas y grupos de investigación alrededor del mundo entrenan y usan redes neuronales que requieren de una alta capacidad de cómputo y de hardware especializado para realizar procesamiento tensorial a gran escala. En este artículo abordamos el tema del costo energético y la huella de carbono que deja el trabajo en aprendizaje profundo. Para lograr el cometido, primero presentamos las definiciones que ayudan a comprender cómo se puede calcular la huella de carbono de los modelos de aprendizaje profundo. Luego, describimos aproximadamente cuánto carbono se emitió para la creación de un modelo popular de procesamiento de lenguaje natural, GPT-3. Finalmente, presentamos algunas reflexiones y líneas de acción para afrontar el potencial problema ambiental de crear y usar inteligencia artificial.

DALL-E 2 [...] fue entrenado durante 18 días seguidos usando 592 GPUs.

En los últimos meses hemos visto cómo han ganado popularidad los modelos de inteligencia artificial (IA) que generan imágenes fotorealistas desde una descripción textual ingresada por un usuario. Estos modelos de IA creativa, tales como DALL-E [1], Imagen [2] o Stable Diffusion [3] son entrenados para encontrar asociaciones entre una imagen y el texto que le corresponde como descripción. Para lograr este objetivo, estas redes neuronales necesitan ser entrenadas con cantidades inmensas de datos, y por lo tanto requieren de una capacidad computacional considerable para su aprendizaje. Por ejemplo, DALL-E 2 tiene alrededor de 3.5 billones de parámetros, Imagen tiene alrededor de 4.6 billones de parámetros y Stable Diffusion tiene alrededor de 890 millones de parámetros (en este artículo usamos la escala corta para representar los números grandes, ya que es la escala usada en los artículos originales; de esta forma, un billón representa mil millones). Como un ejemplo para poner las cosas en contexto, DALL-E 2 se basa en el modelo CLIP [4], el cual fue entrenado durante 18 días seguidos usando 592 GPUs.

En Procesamiento de Lenguaje Natural (NLP) el panorama es todavía más extremo. El modelo GPT-3 [5] (desarrollado por OpenAI) tiene alrededor de 175 billones de parámetros y requirió de una colaboración exclusiva entre OpenAI y Microsoft para disponer de un centro de datos que tiene 285,000 núcleos de CPU y 10,000 GPUs [6]. Más aún, a principios de este año, Google hizo el lanzamiento de su modelo Switch-Transformer [7] cuya versión más grande tiene 1.6 trillones de parámetros (1600 billones de parámetros).

No cabe duda que estos modelos de inteligencia artificial adquieren capacidades sorprendentes para resolver las tareas para las que fueron diseñados. Pero hay una pregunta que no podemos pasar desapercibida: ¿Cuál es el costo energético de entrenar y usar estos modelos? E incluso más importante es la pregunta: ¿Podemos estimar la huella de carbono que dejan estos modelos? En este artículo intento condensar la información que existe a la fecha sobre la huella de carbono de la inteligencia artificial y entregar algunas reflexiones y potenciales líneas de acción para afrontar este problema.

Definiciones preliminares

Antes de empezar a soltar números y hacer cálculos, es necesario comprender algunos términos básicos. Las emisiones equivalentes de CO_2 (CO_2eq) hacen referencia a las emisiones de cualquier gas de efecto invernadero que tiende a incrementar la temperatura de la superficie terrestre. Entre estos gases tenemos el dióxido de carbono, el metano y el óxido nitroso. La unidad de medida común de las emisiones equivalentes es la tonelada métrica que representa 1.000 kilogramos y que se abrevia como tCO_2eq .

Por otro lado, la energía eléctrica consumida en una unidad de tiempo puede medirse con un Megawatt-hora (MWh) que equivale a un millón de watts de electricidad consumidos continuamente en el lapso de una hora. Para poder relacionar el consumo eléctrico con las emisiones de CO_2eq es necesario conocer la cantidad de emisiones por Megawatt-hora en un lugar en particular. Por ejemplo, el Ministerio de Energía en Chile reporta que al año 2018 el factor de emisión del sistema eléctrico nacional es de 0.4187 tCO_2eq/MWh [8]. A este factor también se le conoce como intensidad de carbono (CI) y depende de la ubicación geográfica debido a que está íntimamente relacionado con la fuente de generación eléctrica. Un lugar en donde la energía eléctrica se genera principalmente con fuentes renovables tendrá una intensidad de carbono menor que un lugar en donde se usan combustibles fósiles como carbón o petróleo para generar electricidad. Cuando hablamos de servicios en la nube, los proveedores generalmente publican la información de la intensidad de carbono de manera regular. Por ejemplo, Google tiene esta información disponible y online [9], en donde uno puede observar que existen regiones con intensidad de carbono por debajo de 0.1 tCO_2eq/MWh y otras regiones con valores por encima de 0.7 tCO_2eq/MWh.

Finalmente, un factor importante para medir el impacto energético de una

actividad computacional es la efectividad de uso de potencia PUE (Power Usage Effectiveness). El PUE mide qué tanta potencia adicional es requerida para mantener la infraestructura que soporta el proceso de cómputo (enfriamiento o pérdida de voltaje). Este factor se define como la proporción de potencia usada por toda la infraestructura con respecto a la potencia usada para el cómputo. Por ejemplo, Google publica regularmente sus mediciones de PUE en sus diferentes centros de datos [10], con un promedio de 1.10 de PUE (un 10% de potencia extra es usada para mantener la infraestructura).

Estimación de emisiones de carbono

Una forma de estimar la cantidad de carbono que emite el entrenamiento de una red neuronal es aplicando la siguiente fórmula:

$$tCO_2eq = (MWh\ del\ entrenamiento)\ x\ PUE\ x\ CI$$

Para calcular la cantidad de Megawatt-horas que toma un entrenamiento es necesario contar con información disponible acerca de los equipos que se usaron para el cómputo. Strubell, Ganesh y Callum [11] proponen extraer la información de potencia usada por CPUs, memoria RAM y GPUs con herramientas tipo interfaces RAPL (Running Average Power Limit) o nvidia-smi para registrar la potencia usada por los GPUs. Para sus experimentos, Strubell y compañía ejecutaron los entrenamientos de redes neuronales de NLP tales como Tensor2Tensor, ELMo, BERT y GPT-2 y registraron la potencia requerida para estos modelos en un proyecto que tomó seis meses y requirió aproximadamente de 27 años de tiempo de GPU.

Otra forma de calcular la cantidad de potencia usada es aproximarla multiplicando la cantidad de tiempo que requiere el entrenamiento por la potencia específica de funcionamiento de un GPU. Esta información de potencias está disponible generalmente en las especificaciones técnicas del GPU. La Tabla 1 muestra algunas especificaciones de GPUs comunes usados en tareas de inteligencia artificial. Hay que tener en cuenta que la potencia especificada por el fabricante corresponde a la máxima potencia para el funcionamiento del equipo, por lo que la estimación podría ser mayor a la potencia realmente usada. Patterson y compañía [12] usan esta forma de cálculo para comparar las emisiones de modelos de NLP como Transformer, Evolved Transformer y GPT-3. La ventaja que tiene este método de estimación es que solo necesitamos conocer el tiempo de entrenamiento y el tipo y cantidad de GPUs usados.

Finalmente, también es posible calcular la potencia requerida para un proceso de entrenamiento si se dispone de la cantidad total de cómputo necesaria para la tarea. Por ejemplo, la versión de GPT-3 que tiene 175 billones de parámetros requiere de 3.14 x10^{23} flops (operaciones de coma flotante por segundo) para su entrenamiento. La Tabla 2 muestra la cantidad de Megawatt-horas que se requerirían para entrenar GPT-3 con cada una de los GPUs de la Tabla 1 (usando la equivalencia de 1 teraflop = 10^{12} flops).

Si usamos estos valores de Megawatt-horas junto con un PUE e intensidad de carbono de algún lugar específico, podemos calcular la cantidad de toneladas métricas de CO_2eq que emite el entrenamiento de GPT-3. Para ejemplificar usaré la calculadora de equivalencias [13] implementada por la Agencia de Protección Ambiental de los Estados Unidos (EPA) para determinar la cantidad de emisiones provocadas por los datos obtenidos en la Tabla 2, pero en términos conocidos que podamos entender mejor (se usa una intensidad de carbono aproximada de 0.4 tCO_2eq/MWh).

GPU	FLOPS	Potencia (W)
V100 S PCIe	130 TFLOPS	250
A6000	309.7 TFLOPS	300
A100 SXM	312 TFLOPS	400
RTX 3090 Ti	285 TFLOPS	450

Tabla 1. GPUs comunes usados en IA y sus características de cómputo y energéticas.

GPU	HORAS	Potencia (W)	MWh
V100 S PCIe	670,940	250	167.7
A6000	281,634	300	84.5
A100 SXM	279,558	400	111.8
RTX 3090 Ti	306,042	450	137.7

Tabla 2. Cantidad de horas y consumo de electricidad para entrenar GPT-3 en GPUs.

MWh	tCO_2eq	Autos a gasolina por año	Kg. de carbón quemado	# de celulares cargados
167.7	72.5	15.6	36,408	8,824,695
84.5	36.6	7.9	18,345	4,446,552
111.8	48.4	10.4	24,271	5,883,130
137.7	59.6	12.8	29,894	7,246,038

Tabla 3. Equivalencias de consumo de electricidad para el entrenamiento de GPT-3.

El ciclo de vida de un modelo podría no terminar cuando este se entrena, sino que se extiende a su uso en producción e inferencia.

En el artículo original [12] en donde se experimenta con GPT-3 se obtienen valores de consumo de energía y emisiones incluso más grandes (1287 MWh y 552 tCO_2eq). Esto se debe a que ese análisis se realizó con las especificaciones del GPU V100 del año 2020. En esa versión de GPU, la cantidad de TFLOPS era menor a la de la versión más actual que usamos en la Tabla 3.

En la versión de reporte técnico de GPT-3 [14] incluso se informa de la cantidad total de flops requeridos para cada uno de los experimentos usados en comparación. En total se requirió de 4.1 x 10^{23} operaciones flops, incrementando en casi un 50% más de consumo de energía si usamos los GPUs de nuestra comparación. También cabe destacar que los datos informados corresponden con el trabajo de entrenamiento del modelo final, cuando es común que un modelo se entrene muchas veces antes de tener la versión definitiva. Esto se debe a que el proceso de llegar al modelo final pasa por un proceso de configuración de hiperparámetros, los cuales necesitan ser configurados haciendo el entrenamiento muchas veces. Por ejemplo, Strubell, Ganesh y Callum [11] consideran en su análisis que cada modelo fue entrenado en promedio 24 veces y usan esa estimación para calcular las emisiones de sus experimentos.

Reflexión y líneas de acción

Para este artículo he usado como ejemplo los datos reportados de un solo método. Para dimensionar el impacto real de la huella de carbono necesitamos considerar algunos aspectos adicionales. Primero, el ciclo de vida de un modelo podría no terminar cuando este se entrena, sino que se extiende a su uso en producción y a la cantidad de veces que el modelo se usa para inferencia. Por ejemplo, Google usa el modelo BERT para su motor de búsquedas y la cantidad de búsquedas en Google es un número inmensamente grande. De igual forma, Facebook usa el motor DETR para detectar y analizar objetos en imágenes. Segundo, el aprendizaje profundo se gestó desde hace diez años, tiempo en el que es cada vez más grande la cantidad de modelos que se entrenan y usan, tanto en investigación como en producción. Teniendo en cuenta estos factores, creo que es necesario que tomemos conciencia que las emisiones de carbono provenientes de actividades relacionadas a la inteligencia artificial y aprendizaje profundo podrían ya estar siendo un problema ambiental importante; y que es por lo tanto necesario poner el tema en debate y tomar líneas de acción efectivas para afrontar el problema. Aquí trataré de esbozar tres posibles líneas de acción que ayuden a visibilizar y afrontar el problema. Estas líneas de acción están relacionadas con tres aspectos: la información sobre emisiones, los métodos y algoritmos, y finalmente la práctica de la IA.

- **Información sobre emisiones.** Para poder entender mejor el problema, necesitamos más y mejor información. Henderson y otros [15] reportan que de una muestra de cien artículos de la conferencia NeurIPS del 2019, ningún artículo reporta emisiones de carbono y menos de la mitad de artículos reportan algún tipo de información que pueda ser útil para calcular las emisiones de carbono de sus experimentos. Esto debería cambiar hacia una forma más sistemática de reportar información como el tipo y cantidad de GPUs usados y el tiempo real de experimentación (tomando en cuenta configuración de hiperparámetros). Además, Henderson y compañía incluyen en su propio artículo

También es necesario pensar en mejores formas de cómo encontrar el mejor modelo para un problema dado. Si tu búsqueda de los mejores hiperparámetros requiere de muchísimos entrenamientos, es probable que tu metodología necesite revisión.

un "Carbon Impact Statement" en donde informan la cantidad de toneladas métricas de CO_2eq y la cantidad de watt-horas que tomó todos sus experimentos. Todos quienes experimentamos con modelos de aprendizaje profundo, deberíamos tomar esta iniciativa como punto de partida.

Y en la industria, el panorama es menos claro. Al mejor de mi conocimiento, no existe ningún reporte que indique la cantidad de emisiones de carbono de los modelos de IA usados en empresas como Google, Facebook o Amazon. Mucho menos existe una práctica estandarizada para reportar esta información en industrias de IA.

Finalmente, existen herramientas que han sido creadas con la finalidad de facilitar el reporte de emisiones de carbono en la comunidad de aprendizaje automático. Entre estas herramientas están Machine Learning Emissions Calculator [16], Experiment Impact Tracker [17] y Carbon Tracker [18].

- **Métodos y algoritmos.** Los modelos de aprendizaje profundo se hacen cada vez más grandes y requieren de un poder computacional gigantesco. Sin embargo, una interrogante natural es saber si es posible obtener la misma efectividad con mucho menos costo computacional. Ideas tales como la destilación de conocimiento [19] y la compresión de redes neuronales [20] han sido útiles para reducir el cómputo de modelos grandes y ponerlos en producción, por ejemplo [21]. Sin embargo, todavía hay mucho espacio para contemplar nuevas ideas que ayuden a reducir la cantidad de cómputo que requiere un modelo neuronal.

- **Práctica de IA.** Actualmente una forma de reducir el impacto es usando GPUs eficientes (que tengan buenas tasas de TFLOPS por consumo de energía). Además, si el trabajo será realizado en la nube, la mayoría de proveedores cuentan con información sobre el impacto de carbono de sus diferentes centros de datos, por lo que podemos decidir usar aquellos centros de datos que tienen una intensidad de carbono baja. También es necesario pensar en mejores formas de cómo encontrar el mejor modelo para un problema dado. Si tu búsqueda de los mejores hiperparámetros requiere de muchísimos entrenamientos, es probable que tu metodología necesite revisión.

Consideraciones finales

Si comparamos la cantidad de información disponible para medir el impacto de las emisiones de carbono de actividades relacionadas con IA con respecto a la cantidad de información del área en general podemos darnos cuenta que estamos lejos de tener el tema sobre la mesa de discusión de la comunidad. Pero más allá de levantar una alarma y llegar a ser hasta catastróficos con respecto al tema, estamos frente a la oportunidad de tomar acciones reales y efectivas para no permitir que esto sea un problema mayor en el futuro. Cualquier forma que usemos para reducir las emisiones provocadas por modelos de inteligencia artificial sirve. ▪

REFERENCIAS

[1] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M.: Hierarchical Text-Conditional Image Generation with CLIP Latents. ArXiv, abs/2204.06125. 2022.

[2] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, S.K., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D., & Norouzi, M.: Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. ArXiv, abs/2205.11487. 2022.

[3] https://stability.ai/blog/stable-diffusion-public-release.

[4] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. Int. Conf. Machine Learning. 2021.

[5] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T.J., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D.: Language Models are Few-Shot Learners. NeurIPS. 2020.

[6] https://blogs.microsoft.com/ai/openai-azure-supercomputer/.

[7] Fedus, W., Zoph, B., Shazeer, N.: Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. Journal of Machine Learning Research 23(120), pp 1-39, 2022.

[8] https://energia.gob.cl/indicadores-ambientales-factor-de-emisiones-gei-del-sistema-electrico-nacional.

[9] https://cloud.google.com/sustainability/region-carbon.

[10] https://www.google.com/about/datacenters/efficiency/.

[11] Strubell, E., Ganesh, A., McCallum A.: Energy and Policy Considerations for Deep Learning in NLP. In Proc. of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3645–3650. 2019.

[12] Patterson, D.A., González, J., Holzle, U., Le, Q., Liang, C., Munguía, L., Rothchild, D., So, D.R., Texier, M., & Dean, J.: The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. Computer, 55(7), 18-28. 2022.

[13] https://www.epa.gov/energy/greenhouse-gas-equivalencies-calculator.

[14] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T.J., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D.: Language Models are Few-Shot Learners. ArXiv, abs/2005.14165. 2020.

[15] Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., Pineau, J.: Towards the systematic reporting of the energy and carbon footprints of machine learning. Journal of Machine Learning Research, Vol 21(248), pp. 1-43. 2020.

[16] https://mlco2.github.io/impact/.

[17] https://github.com/Breakend/experiment-impact-tracker.

[18] https://github.com/lfwa/carbontracker.

[19] Hinton, G.E., Vinyals, O., & Dean, J.: Distilling the Knowledge in a Neural Network. ArXiv, abs/1503.02531. 2015.

[20] Shen, S., Dong, Z., Ye, J., Ma, L., Yao, Z., Gholami, A., Mahoney, M.W., & Keutzer, K.: Q-BERT: Hessian Based Ultra Low Precision Quantization of BERT. AAAI. 2020.

[21] Sanh, V., Debut, L., Chaumond, J., & Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv, abs/1910.01108. 2019.