Entrenamiento de IAs

Escrito por @olimobu en este tema del foro:

Los investigadores y empresas nos han hecho creer que sus algoritmos «entienden» lo que se les dice o que «piensan» o «analizan». **Es una humanización, una simplificación y una mentira** . Lo que llamamos IA dista enormemente de lo que se imaginó en los años cincuenta o sesenta. **Ningún sistema está todavía programado para razonar**. La IA utiliza cantidades masivas de datos para convertir cualquier tarea compleja en un problema de predicción basado en el propio trabajo humano. Las IAs son potentes calculadoras que usan las matemáticas estadísticas para procesar trabajo humano previo con un objetivo concreto. Los desarrolladores han conseguido que sintamos que nos escuchan y que nos entienden, cuando no es más que apariencia.

En una investigación de *The Whashington Post* sobre el entrenamiento de IAs se analizaron 15 millones de webs, contenidas en el conjunto de datos C4 *(Colossal Clean Crawled Corpus)* de Google. La C4 es una base generada por Common Crawl, una organización sin fines de lucro que rastrea Internet periódicamente para compilar información. Esta base de datos se usó, específicamente, para desarrollar los modelos de lenguaje LLaMA, de Facebook, y T5, de Google.

El medio estadounidense creó un buscador en el que se pueden consultar todas las webs recogidas en el conjunto C4.

Los resultados de la investigación demostraron que muchos de los contenidos recogidos en el conjunto C4 viola los derechos de autor. También incluye material racista, información tendenciosa y un claro sesgo religioso.

Los chatbots impulsados por IA recopilan y procesan información de al menos estos sitios según el análisis de la dataset C4 de Google por *The Whashington Post*:

- Patentes: **Patents.Google.com** (720 millones de Tokens), **Patents.com** (64 millones de Tokens).
- Negocios: **Kickstarter.com (39 millones de Tokens) , Patreon.com (3 millones de Tokens) . Se sabe que IAs como Stable Diffusion y MidJourney han usado estas webs para acceder a ideas y material de artistas sin permiso.**
- **Correos Electrónicos: ProtonMail.com (190 mil Tokens) , ProtonVPN.com (94 mil Tokens), Tutanota.com (29 mil Tokens) DisRoot.org (11 mil Tokens), Posteo.de (9.9 mil Tokens), CounterMail.com (7.9 mil Tokens), MailFence.com (2.2 mil Tokens), StartMail.com (780 Tokens) .**
- **Compras: Amazon.com (4.8 millones de Tokens), eBay.com (7.4 millones de Tokens), Thomann.de (770 mil Tokens) .**

- **Aprendizaje: Coursera.org (53 millones de Tokens), Udemy.com (1.9 millones de Tokens), Duolingo.com (46 mil Tokens) .**
- **Información : Wikipedia.org (290 millones de Tokens) , Scribd.com (100 millones de Tokens) , NYTimes.com (100 millones de Tokens) , LaTimes.com (85 millones de Tokens), TheGuardian.com (83 millones de Tokens), Forbes.com (73 millones de Tokens) , Huffpost.com (68 millones de Tokens), WashingtonPost.com (55 millones de Tokens), RT.com (26 millones de Tokens) .**
- **Redes Sociales : Medium.com (33 millones de Tokens), Reddit.com (7.9 millones de Tokens), Twitter.com (1.2 millones de Tokens), Mastodon. S ocial (210 mil Tokens), Mastodon. A rt (100 mil Tokens), Mastodon.Technology (130 mil Tokens), Mastodon.Cloud (84 mil Tokens), Mastodon.GameDev.Place (26 mil Tokens), Slack.com (150 mil Tokens), Skype.com (55 mil Tokens), Telegram.org (13 mil Tokens), Jitsi.org (7.6 mil Tokens), WeChat.com (7.6 mil Tokens) WriteFreely.org (2.2 mil Tokens).**
- **Vídeos: YouTube.com (20 millones de Tokens), Twitch.tv (41 mil Tokens), TikTok.com (26 mil Tokens), PeerTube.fr (100 Tokens) .**
- **Música: SoundCloud.com (2.7 millones de Tokens), BandCamp (84 mil Tokens), FunkWhale.audio (120 Tokens).**

Se ha publicitado que para entrenar a OpenAI (cuyo generador más famoso es ChatGPT), de Musk y Altman, se usó un conjunto de datos **unas 40 veces la cantidad de C4**, aunque no han sido transparentes sobre su contenido específico. MuseNet es su generador de música algorítmica. Según el Blog de Nvidia, ClassicalArchives y BitMidi han donado sus enormes colecciones de MIDIs para el entrenamiento de MuseNet. Además, se sabe que se ha usado la MAESTRO dataset (200 horas de interpretaciones virtuosas al piano capturadas con un margen de error de alrededor de 3ms entre las indicaciones en notación y las formas de onda).

Google ha lanzado MusicLM para generar música a partir de texto. Esta IA ha sido entrenada con la MusicCaps dataset (5.5 mil Tokens de pares texto-musica proporcionados por expertos a partir de 280 mil horas de música). Alrededor del 1% de la música que genera esta IA es una réplica de material protegido por las leyes de derechos de autor. Además, Google tiene el proyecto Magenta, una app y plugin para **AbletonLive** que permite hacer música usando *Machine Learning*. Para este proyecto se han usado varios conjuntos de datos:

- Bach Doodle Dataset: 21.6 millones de melodías armonizadas en MIDI con metadatos acerca de la composición (país de origen, época, estilo, etc).
- CocoChorales: 240 mil ejemplos de audio con su información MIDI y parámetros de síntesis en el formato a cuatro partes típico de las corales de Bach.
- Groove MIDI Dataset (y su expansión): más de 444 horas de audio con 43 kits de batería grabadas con intérpretes humanos y asociados a su MIDI correspondiente.
- MAESTRO dataset: 200 horas de interpretaciones virtuosas al piano capturadas con un margen de error de alrededor de 3ms entre las indicaciones en notación y las formas de

onda.

- Nsynth: 305.979 notas musicales para 1.006 instrumentos de librerías de samples comerciales con el rango de un piano y a 5 intensidades diferentes.
- Quick, Draw!: 50 millones de dibujos ordenados en 345 categorías con metadata asociada.

Por su parte, Meta, de Zuckerberg, ha lanzado AudioCraft, al que pertenece AudioGen para generar sonido ambiental a partir de texto y MusicGen para general música a partir de texto. MusicGen ha sido entrenada con un conjunto de datos autorizado de 20 mil horas de música. Se sabe que **Shutterstock** y **Pond5** han donado 10 mil grabaciones de audio para su entrenamiento.

Los expertos alertan de que muchas compañías de IA no documentan sus datos de entrenamiento -incluso internamente- por miedo a que se sepa que hay información personal sobre individuos identificables, material con derechos de autor y otros datos recogidos sin consentimiento.

Universal Music Group, una de las principales discográficas del mundo, pidió a ***Apple*** y ***Spotify*** que bloqueen a los bots que extraen letras y melodías de las canciones de sus artistas. Según la compañía, ese material protegido con *copyright* luego se usa para entrenar modelos de inteligencia artificial capaces de crear música parecida a la de intérpretes o compositores como Taylor Swift o Elton John.

La tecnología es política, y la mayoría de sistemas basados en IA reproducen desigualdades estructurales, pues están dominados por una mayoría masculina, blanca, **cisgénero** *y* **capacitista** *.*

Newsguard, una plataforma que mide y califica la confiabilidad de sitios web informativos, viene alertando cómo las IAs ChatGPT, GPT-4 y Bard producen fácilmente contenido falso para respaldar conocidas teorías conspirativas.

Los científicos critican que la mayoría de los estudios basados en IA acaban siendo una mera acción promocional. La falta de transparencia por parte de la mayoría de las empresas de IA impide que los nuevos modelos y técnicas se evalúen adecuadamente en términos de solidez, sesgo y seguridad. Si no hay reproducibilidad, no puede haber ciencia. Un informe critica concretamente a OpenAI y a DeepMind por mantener sus códigos en secreto. **Cada vez es más complicado saber qué resultados son resultados fiables y cuáles no, lo que contribuye a una situación de posverdad. Los grandes experimentos científicos de IA generalmente se llevan a cabo en hardware que es de propiedad y está en su mayoría controlado por las grandes tecnológicas.**

Según Ben Goertzel, uno de los mayores expertos en IA, la inteligencia artificial terminará eliminando el 80% de todos los empleos: "No creo que sea una amenaza [...] El problema que veo

es en el período intermedio, cuando las IA hagan que el empleo humano sea obsoleto... **No sé cómo resolver los problemas sociales** [que esto va a provocar]". “La gente puede encontrar **mejores cosas que hacer con su vida** que trabajar para ganarse la vida”. Según un estudio de OpenAI, de ese 80% de profesionales afectados por la IA, por lo menos un 10% de su actividad será **totalmente reemplazada** por IA y casi un 20% de todos los trabajadores verán como la IA realizará la mitad de sus tareas... de forma inminente. Estos datos probablemente estén hinchados para atraer inversores, ya que por otra parte hay expertos que cuestionan las posibilidades reales de la IA: la IA no piensa, pero es muy buena automatizando tareas muy concretas. Lo peligrosamente disruptivo no es ya que la IA pueda sustituir de forma eficaz a los trabajadores, sino que se convenza a los empresarios de que el trabajo de la IA con algunos retoques de un puñado de trabajadores explotados es suficiente para seguir incrementando sus ganancias. De hecho, esto no solo podría afectar a la empleabilidad de forma crítica (en un momento en el que se está acelerando la desigualdad social), sino que si los inversores se dan cuenta de pronto de que todo es una gran exageración, podría originar una crisis mucho mayor que la de las puntocom.

Se estima que Google, **Amazon**, **Facebook** y **Microsoft** almacenan un total de 1200 petabytes, es decir, algo más de 1200 millones de gigabytes. Muchos gobiernos, y compañías privadas que ganan **contratos públicos**, utilizan servicios de estas big tech para almacenar sus datos. **Amazon tiene un departamento exclusivo para políticas públicas**, en las que ofrecen **servicios de almacenamiento, inteligencia artificial y ciberseguridad**. Los costes de mantener estos servicios son inmensos (así como su consumo energético), por lo que la mayoría de los gobiernos (el español estuvo en conversaciones con Amazon para ello en 2019) ceden toda esta responsabilidad y se vuelven dependientes de estas multinacionales, lo que mina la independencia y nivel democrático de los Estados.

Los programas de inteligencia artificial, según indican en *Bloomberg*, **consumen más energía que cualquier otro sistema de computación** (incluso más que el minado de Bitcoin). Se estima que el c**onsumo energético para entrenar el modelo inicial (1.287 GW/h en el caso de ChatGPT-3) es tan solo un 40%** de la energía que se emplea en el uso real del día a día una vez lanzado al gran público (que supone millones de consultas diarias). Una sola consulta en ChatGPT (implementado por Microsoft en el navegador Edge) consume 3 veces más energía que otra hecha en el buscador de Google. Además, la IA debe volverse a entrenar continuamente para estar actualizada, lo que agrava el problema del coste energético. Por lógica, un sistema que use solo texto va a consumir menos que uno que use imágenes, audio o peor, vídeo.

Revision #2
Created 14 August 2023 19:19:19 by blankfosk
Updated 14 August 2023 19:22:20 by blankfosk