DEEP SEEK A PROFUNDIDAD

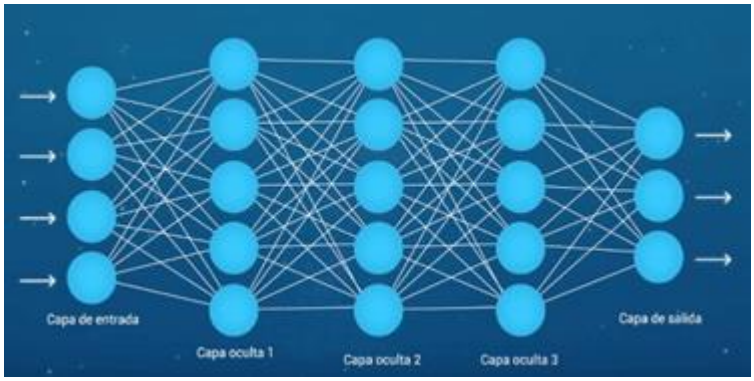
Por: Rodrigo Bernardo Ortega. 24/05/2025

El auge de la Inteligencia Artificial (IA) ha transformado radicalmente el panorama tecnológico global, generando avances sorprendentes y suscitando debates éticos y geopolíticos de gran calado. En este contexto de innovación disruptiva, emerge la figura de Liang We Feng, un visionario ingeniero chino cuya trayectoria ejemplifica la convergencia entre la academia, la visión empresarial y la audacia tecnológica. Su historia, desde las aulas de la Universidad de Zhejiang hasta la fundación de DEEP SEEK, una empresa que ha sacudido los cimientos de la industria de la IA, ofrece una perspectiva fascinante sobre el potencial y los desafíos de esta tecnología transformadora.

La génesis de DEEP SEEK se remonta a los años universitarios de Liang We Feng, un período marcado por una inquietud intelectual que trascendía los límites de la ingeniería tradicional. Consciente del potencial lucrativo de los mercados financieros, Liang formó un grupo de estudio con otros estudiantes para explorar las complejidades de la bolsa de valores. Esta incursión inicial en el mundo de las finanzas despertó en él una fascinación por la aplicación de principios ingenieriles a la toma de decisiones en entornos complejos y dinámicos.

Fue así como Liang se adentró en el incipiente campo del *Quant Trading*, una metodología que sustituye la intervención humana en la compra y venta de activos bursátiles por algoritmos informáticos. En lugar de operadores ejecutando transacciones manualmente, el *Quant Trading* se basa en modelos matemáticos sofisticados y software especializado que analizan ingentes cantidades de datos para identificar patrones y oportunidades de inversión. Este enfoque sistemático y basado en datos representó para Liang una extensión natural de su formación en ingeniería, donde la resolución de problemas complejos a través de la lógica y el análisis cuantitativo era un principio fundamental.

La visión de Liang We Feng trascendió la mera automatización de las transacciones financieras. Su interés se centró en la aplicación del *Machine Learning* (aprendizaje automático) para dotar a estos sistemas automatizados de la capacidad de aprender de los datos, adaptarse a las condiciones cambiantes del mercado y, en última instancia, tomar decisiones de inversión más inteligentes y predictivas. Durante sus años universitarios, Liang dedicó una parte significativa de su tiempo y energía a explorar las posibilidades del *Machine Learning* en este dominio, sentando las bases conceptuales y técnicas para sus futuros emprendimientos.



En 2016, tras culminar sus estudios universitarios, Liang materializó su visión

fundando High Flyer, una firma de inversión pionera en la adopción del *Quant Trading* impulsado íntegramente por decisiones automatizadas por computador. Desde sus inicios, High Flyer se distinguió por su enfoque vanguardista, integrando la Inteligencia Artificial como el núcleo de su estrategia operativa. La empresa logró un crecimiento notable, gestionando en poco tiempo activos superiores a los 8000 millones de dólares, un testimonio del éxito de su enfoque basado en la precisión predictiva de la IA aplicada a los mercados financieros.

Pero el verdadero giro en su carrera llegaría años más tarde, pues la ambición de Liang We Feng no se detuvo en el ámbito de las finanzas. Su mirada se proyectó hacia las fronteras más avanzadas de la Inteligencia Artificial, reconociendo su potencial transformador en una multitud de industrias y aplicaciones. En un movimiento estratégico que anticipaba el futuro de la computación y la IA, en 2021 Liang adquirió miles de tarjetas gráficas de Nvidia, componentes cruciales para el entrenamiento de modelos de aprendizaje profundo. Este importante desembolso de recursos evidenciaba su compromiso con la investigación y el desarrollo de capacidades de IA de vanguardia.

El 17 de julio de 2023 marcó un hito crucial en la trayectoria de Liang We Feng con la fundación de HANGZHOU DEEPSEEK ARTIFICIAL INTELIGENCE BASIC TECHNOLOGY RESEARCH CO., LTD., más conocida como DEEP SEEK. Con una plantilla sorprendentemente reducida de alrededor de 200 empleados, DEEP SEEK logró a finales de 2024 un avance tecnológico que resonó en toda la industria. El lanzamiento de su modelo de Inteligencia Artificial no solo demostró capacidades asombrosas, sino que también generó preocupación en el gobierno de los Estados Unidos debido a su inusualmente bajo costo de desarrollo.

Es importante destacar que, si bien el gobierno chino ha manifestado su apoyo al desarrollo de la Inteligencia Artificial a través de la inversión en centros de datos y la promulgación de leyes que favorecen la IA generativa, el éxito de DEEP SEEK parece ser el resultado de una iniciativa privada impulsada por la visión y la capacidad técnica de Liang We Feng y su equipo.

DEEP SEEK presentó al mundo dos modelos lingüísticos grandes (LLMs), denominados Deep Seek R1 y Deep Seek V3, que se inscriben en la misma categoría de modelos generativos de texto que ChatGPT, Claude o Gemini. Estos modelos destacan por su capacidad para comprender y generar texto de manera coherente y contextualmente relevante. Sin embargo, lo que diferencia a los

modelos de DEEP SEEK de sus competidores es su rendimiento superior en ciertas pruebas de evaluación.

Las pruebas de rendimiento en el campo de los LLMs están diseñadas para medir la capacidad de los modelos para resolver una variedad de tareas cognitivas, que van desde responder preguntas de conocimiento general hasta resolver acertijos lógicos y comprender textos complejos. En algunas de estas pruebas, Deep Seek V3 ha superado a modelos líderes como Claude 3.5 y GPT-4o. La propia web oficial de DEEP SEEK destaca el rendimiento superior de sus modelos en pruebas como MMLU (que evalúa conocimiento general, razonamiento lógico y comprensión avanzada en múltiples temas) y DROP (que mide la capacidad de razonamiento sobre textos largos que requieren cálculos y combinaciones de datos). Otro ejemplo es EER Polyglot, que evalúa la capacidad del modelo para trabajar con múltiples lenguajes de programación.

Estos resultados sugieren que DEEP SEEK ha logrado desarrollar modelos de IA que no solo igualan el rendimiento de los sistemas más avanzados disponibles hasta la fecha, sino que en ciertos aspectos los superan. Este logro es particularmente impresionante dado el relativamente corto tiempo de existencia de la empresa y su modesta plantilla en comparación con los gigantes tecnológicos que dominan el campo de la IA.

El segundo factor que ha contribuido al impacto de DEEP SEEK en la industria es su estrategia de precios y distribución. La interfaz web para interactuar con Deep Seek es gratuita, al igual que la versión básica de ChatGPT. Sin embargo, el modelo de negocio principal para este tipo de empresas radica en la provisión de acceso a sus modelos a través de una API (Application Programming Interface). La API permite a otras aplicaciones y servicios integrar las capacidades de la IA en sus propias funcionalidades. Por ejemplo, una aplicación de entrenamiento físico podría utilizar la API para ofrecer un entrenador personal virtual impulsado por la IA.

El costo de utilizar estas APIs se mide generalmente por token, donde un token representa aproximadamente una palabra generada. DEEP SEEK ha adoptado una estrategia de precios significativamente más competitiva que sus rivales. Mientras que los tokens de salida de GPT-4o tienen un costo de 10 dólares por millón de tokens, Deep Seek V3 se ofrece a 1.1 dólares por millón de tokens, lo que representa una reducción de casi diez veces en el precio. Esta diferencia sustancial en los costos podría democratizar el acceso a la IA para una gama más amplia de empresas y desarrolladores.

Aun más disruptiva y potencialmente transformadora fue la decisión de liberar sus modelos Deep Seek V3 y R1 bajo una licencia de código abierto y de forma totalmente gratuita. Esto significa que cualquier persona puede descargar los modelos y ejecutarlos en su propia infraestructura. En contraste, modelos propietarios como ChatGPT y Gemini solo pueden utilizarse a través de los servidores de sus respectivas empresas.

Si bien la ejecución del modelo más grande de DEEP SEEK, R1 (con 671 mil millones de parámetros), requiere una infraestructura computacional considerable (aproximadamente 16 tarjetas gráficas Nvidia A100 con un costo estimado de medio millón de dólares), la posibilidad de descargar y ejecutar un modelo de IA de vanguardia en una infraestructura propia representa un cambio de paradigma en la industria. Esto elimina la dependencia de las APIs de terceros y otorga a las organizaciones un control total sobre sus implementaciones de IA, permitiéndoles personalizar los modelos y utilizarlos sin incurrir en costos recurrentes por el uso de la API.

La respuesta de la comunidad tecnológica a la liberación de los modelos de DEEP SEEK ha sido abrumadoramente positiva. En la primera semana de su lanzamiento,

los modelos superaron el millón de descargas, lo que indica un alto nivel de interés y adopción por parte de investigadores y empresas con la infraestructura necesaria para ejecutarlos.

Todo esto para afirmar, que DEEP SEEK ha presentado una IA que rivaliza en rendimiento con los modelos más avanzados disponibles, ofreciendo una alternativa más económica para la interacción a través de API, y además, brindando la posibilidad de descargar y ejecutar los modelos de forma local, una opción inédita para modelos de esta capacidad.

Más allá de su rendimiento y modelo de distribución disruptivo, DEEP SEEK ha logrado avances notables en la eficiencia del entrenamiento y la ejecución de sus modelos. Tradicionalmente, el entrenamiento de modelos de IA a gran escala requiere enormes cantidades de poder computacional y recursos financieros significativos. Sin embargo, DEEP SEEK ha reportado costos sorprendentemente bajos tanto para el entrenamiento como para la operación de sus modelos, utilizando una cantidad relativamente pequeña de tarjetas gráficas y un tiempo de entrenamiento más corto de lo esperado.

Este logro se atribuye a una serie de optimizaciones técnicas innovadoras implementadas en la arquitectura de sus modelos. Uno de los pilares fundamentales de la eficiencia de DEEP SEEK es su adopción de una arquitectura denominada *Mixture of Experts* (MoE). En contraste con los modelos tradicionales como GPT o Llama, que se basan en una única red neuronal generalista, la arquitectura MoE divide el modelo en múltiples sub-modelos más pequeños, especializados en diferentes áreas de conocimiento.

Cuando se introduce un *prompt* (una instrucción de entrada), un *router(enrutador)* dentro del modelo MoE analiza la consulta y asigna la tarea de procesamiento a uno o varios de estos expertos especializados que son más relevantes para el tema en cuestión. Esto contrasta con la arquitectura tradicional, donde toda la red neuronal se activa para procesar cada *prompt*, lo que requiere una mayor potencia computacional. La arquitectura MoE permite que solo una parte del modelo se active para cada consulta, lo que resulta en un ahorro significativo de energía y recursos computacionales, permitiendo la ejecución del modelo con menos tarjetas gráficas.

Si bien la idea de la arquitectura MoE no es exclusiva de DEEP SEEK (ya había sido implementada en proyectos de Google como GARD y en el modelo Mixtral), la

empresa china ha logrado una implementación particularmente exitosa. Esto se debe, en parte, a la utilización de un número significativamente mayor de expertos en sus modelos y a la cuidadosa gestión de la especialización del conocimiento de cada experto, evitando la redundancia y asegurando una cobertura eficiente de diferentes dominios. Además, DEEP SEEK ha incorporado “expertos mixtos” que poseen conocimientos intermedios entre varios temas, lo que les permite complementar la información proporcionada por los expertos más especializados.

Un detalle técnico relevante es que, si bien Deep Seek R1 cuenta con 671 mil millones de parámetros en total (sumando los parámetros de todos sus expertos), solo 37 mil millones de estos parámetros se activan durante cada inferencia (la generación de una respuesta). Esto subraya la eficiencia de la arquitectura MoE al enfocar la capacidad computacional solo en las partes del modelo que son relevantes para la tarea actual.

Otra optimización clave implementada por DEEP SEEK es el entrenamiento de sus modelos con una precisión numérica más baja de lo habitual, específicamente utilizando el formato FP8 (Floating Point 8). La precisión numérica en los modelos de aprendizaje profundo se refiere al número de bits utilizados para representar los parámetros del modelo. Una mayor precisión (por ejemplo, FP32) permite una representación más fina de los valores, pero también requiere más memoria. Una menor precisión (como FP8) reduce el consumo de memoria y acelera el entrenamiento, pero puede comprometer la precisión del modelo si no se gestiona adecuadamente.

DEEP SEEK ha adoptado una estrategia de “precisión mixta”, utilizando diferentes precisiones numéricas en diferentes partes del modelo. Han identificado las partes donde una menor precisión (FP8) es suficiente sin afectar significativamente el rendimiento general, logrando así un equilibrio entre eficiencia y precisión. Esta técnica, conocida como *mixed precision framework*, permite reducir tanto el tiempo de entrenamiento como los requisitos de memoria del modelo.

Adicionalmente, DEEP SEEK ha incorporado otras técnicas de optimización en la arquitectura de sus modelos. Su capa de atención cuenta con múltiples cabezas, lo que permite al modelo prestar atención a diferentes partes de la entrada simultáneamente y analizar la información desde múltiples perspectivas. Además, sus modelos pueden generar múltiples tokens en cada paso de la inferencia, en contraste con modelos como GPT que generan un token a la vez. Esta generación

paralela de tokens mejora la velocidad de ejecución del modelo sin degradar la calidad de las respuestas.

Si bien muchas de las técnicas utilizadas por DEEP SEEK no son invenciones completamente nuevas, su combinación e implementación efectiva han dado como resultado un modelo de IA que es significativamente más rápido y económico de entrenar y ejecutar que muchos de sus competidores. Este logro es aún más notable en el contexto de la naturaleza inherentemente colaborativa y de código abierto de gran parte de la investigación en Inteligencia Artificial.

La historia de la IA está marcada por la construcción sobre el trabajo de otros. Gran parte del conocimiento fundamental y las tecnologías subyacentes (como las redes neuronales, los *Transformers* y los modelos de difusión) provienen de investigaciones académicas y proyectos de código abierto compartidos por investigadores y programadores de todo el mundo. Empresas como OpenAI se han beneficiado enormemente de este ecosistema de conocimiento compartido, aunque luego hayan optado por privatizar los detalles internos de sus modelos más avanzados.

En este contexto, el logro de DEEP SEEK al construir un modelo de IA de vanguardia utilizando principios y técnicas conocidas, pero optimizándolos de manera innovadora, es un testimonio de la capacidad de la ingeniería y la visión estratégica. Su decisión de liberar sus modelos bajo una licencia de código abierto representa una desviación significativa de la estrategia de los gigantes de la IA y podría tener implicaciones profundas para la democratización del acceso a esta tecnología.

El modelo R1 de DEEP SEEK representa un avance particularmente significativo en el campo del razonamiento de la IA. Empresas como OpenAI han reconocido que los modelos lingüísticos grandes tradicionales tienen limitaciones para resolver problemas que requieren una secuencia lógica de pasos y la capacidad de “pensar” a través de un problema. El modelo GPT-3.5 de OpenAI incorporó la técnica de la “cadena de pensamiento” (*Chain of Thought*) para mejorar su capacidad de razonamiento, generando un texto más extenso que detalla los pasos necesarios para llegar a una solución.

DEEP SEEK ha logrado desarrollar un modelo, R1, que se acerca al rendimiento de modelos avanzados en tareas de razonamiento con un costo computacional significativamente menor. Su enfoque para el entrenamiento del razonamiento se diferencia del método de “aprendizaje por refuerzo con retroalimentación humana” (RLHF) utilizado por OpenAI. En lugar de depender de evaluadores humanos para guiar el aprendizaje del modelo, DEEP SEEK ha adoptado un enfoque de aprendizaje por refuerzo totalmente automatizado.

Para entrenar a R1, DEEP SEEK partió de su modelo V3 y lo sometió a una serie de problemas complejos con respuestas deterministas y fácilmente verificables (como acertijos lógicos, problemas matemáticos y código de programación). El modelo generaba respuestas, y un sistema automatizado evaluaba la calidad de estas respuestas, otorgando una “recompensa” a las respuestas que se acercaban a la solución correcta. A través de este proceso iterativo de generación y evaluación automática, el modelo aprendió a generar respuestas más largas y complejas que reflejaban un proceso de razonamiento.

Este logro es sorprendente porque el enfoque de aprendizaje por refuerzo totalmente automatizado para el razonamiento no se consideraba tradicionalmente

tan efectivo como el RLHF. El éxito de DEEP SEEK sugiere que, con la selección adecuada de problemas de entrenamiento y un sistema de recompensa bien diseñado, es posible entrenar modelos de razonamiento de alto rendimiento sin la necesidad de una intervención humana intensiva.

Si bien los modelos de DEEP SEEK han demostrado un rendimiento impresionante en tareas lógicas y científicas, es importante señalar que, debido a su enfoque de entrenamiento no supervisado por humanos, pueden no igualar a modelos como ChatGPT en la generación de respuestas creativas, conversacionales y con un tono más “humano”. El entrenamiento con retroalimentación humana permite a los modelos aprender a alinear sus respuestas con las preferencias y expectativas humanas en términos de estilo y contenido.

A pesar de estas diferencias, la existencia de un modelo como Deep Seek R1, desarrollado con un proceso de entrenamiento relativamente económico y liberado de forma gratuita, representa una oportunidad sin precedentes para universidades, empresas e investigadores con la infraestructura adecuada. La capacidad de descargar, ejecutar y modificar un modelo de razonamiento de vanguardia podría impulsar la innovación en una amplia gama de aplicaciones.

La pregunta inevitable es cómo DEEP SEEK planea monetizar su inversión en investigación y desarrollo si libera sus modelos de forma gratuita. Esta es una estrategia que se analiza en profundidad en el contexto del software libre. Al liberar sus modelos, DEEP SEEK podría estar buscando construir una comunidad de usuarios y desarrolladores que contribuyan a la mejora y la adopción de sus productos. Además, aunque el modelo base sea gratuito, la empresa podría ofrecer servicios de soporte, versiones empresariales con características adicionales o soluciones personalizadas para clientes específicos. La popularidad generada por la liberación de los modelos también podría atraer talento y oportunidades de colaboración.

Otra estrategia inteligente de DEEP SEEK es el desarrollo de modelos “destilados”, versiones más pequeñas y eficientes que pueden ejecutarse en hardware menos potente, incluso en computadoras personales. Estos modelos destilados, aunque no estén directamente basados en la arquitectura de DEEP SEEK (algunos utilizan modelos base como Llama), amplían el alcance de sus tecnologías y permiten a un público más amplio experimentar con sus capacidades.

Finalmente, un aspecto crucial del éxito de DEEP SEEK radica en su capacidad para entrenar sus modelos utilizando hardware que, en teoría, no tiene la potencia óptima para esta tarea. En 2023, el gobierno de los Estados Unidos impuso restricciones a la venta de tarjetas gráficas de alto rendimiento de Nvidia (como las H100, diseñadas específicamente para el entrenamiento de modelos de IA avanzados) a China. Estas restricciones afectaron particularmente la memoria de estas tarjetas.

El hecho de que DEEP SEEK haya logrado desarrollar y entrenar modelos de IA de vanguardia a pesar de estas limitaciones en el acceso al hardware más avanzado es un testimonio de la ingeniosidad y la habilidad técnica de su equipo. Esto sugiere que han encontrado formas innovadoras de optimizar sus algoritmos de entrenamiento y utilizar de manera eficiente los recursos computacionales disponibles.

En conclusión, la historia de Liang We Feng y la aparición de DEEP SEEK representan un capítulo fascinante en la evolución de la Inteligencia Artificial.

FUENTES

https://chat.deepseek.com

https://llingfei.com/usr/uploads/2025/02/3735521168.pdf

https://tldv.io/es/blog/what-is-deepseek

https://www.academia.edu/127321612/Deep_Seek_Los_impactos_de_la_nueva_herrar

https://www.researchgate.net/publication/388640321_REVOLUTIONIZING_THE_SEAR
SEEKING_BEHAVIOR_IN_THE_DIGITAL_AGE

https://www.boozallen.com/content/dam/home/docs/ai/a-technical-primer-on-deepseek.pdf

Fecha de creación
2025/05/24