En Profundidad

Inteligencia artificial generativa y modelos de lenguaje a gran escala

Autor: Guillermo González Muñoz de Morales, Unidad de Prospectiva Tecnológica, ISDEFE.

Palabras clave: Inteligencia Artificial Generativa (IA Generativa), Modelos de Lenguaje a Gran Escala (LLMs), *Transformers*, Aprendizaje Profundo, Multimodalidad.

Líneas ETID relacionadas: 11.5.1, 11.5.3.

Introducción

La Inteligencia Artificial (IA) ha experimentado avances notables en la última década y uno de los principales protagonistas de esta evolución ha sido el desarrollo y uso de los Modelos de Lenguaje de Gran Escala (LLMs, por sus siglas en inglés), diseñados específicamente para procesar texto. Estos LLMs son un pilar esencial dentro de la IA generativa, que busca crear y generar contenido nuevo a partir de lo aprendido. Se construyen sobre lo que se conoce como modelos base, una forma avanzada de aprendizaje profundo que ha marcado un punto de inflexión en el campo de la IA.

A diferencia de sus predecesores, los modelos base tienen la capacidad de trabajar con vastos conjuntos de datos no estructurados, permitiéndoles ejecutar una amplia variedad de tareas, desde la generación de contenido hasta la clasificación. Además, si bien están diseñados principalmente para el texto, la adaptación de LLMs para procesar imágenes y otros tipos de datos es un área emergente de interés y desarrollo.

Dentro del ecosistema de modelos existentes, encontramos ejemplos comerciales destacados como OpenAI GPT, Google Bard y Anthropic Claude, así como opciones de código abierto, como, por ejemplo, LLAMA de META. Para entender su funcionamiento, un aspecto fundamental en estos modelos es la representación de la información. En el mundo de la IA generativa la información es transformada y representada en espacios vectoriales n-dimensionales, donde se convierte en vectores o matrices, conocidos comúnmente como *embeddings*. La arquitectura neuronal que facilita esta transformación es conocida como *transformers*.

El poder de los LLMs modernos no solo radica en su arquitectura, sino también en la manera en que se entrenan. Muchos, como el GPT (cuyo acrónimo significa «*Generative Pre-trained Transformer*»), vienen preentrenados, lo que facilita su uso directo, o alternativamente, su adaptación a tareas más específicas mediante un proceso de particularización denominado *Fine-Tuning*. Este proceso consiste en entrenar el modelo con un conjunto de datos de alta calidad para una tarea específica, lo que permite optimizar su desempeño. Como, por ejemplo, particularizar el modelo a un corpus documental específico para potenciar la precisión y relevancia de los LLMs. De esta manera, pueden entender y responder preguntas en lenguaje natural que estén relacionadas, de forma directa, con ese conjunto de documentos, lo que les confiere una gran adaptabilidad y especificidad en sus respuestas.

En términos prácticos, estos modelos de IA se han vuelto accesibles a través de interfaces web e Interfaces de Programación de Aplicaciones (APIs), donde se comunican con servidores en la nube. Sin embargo, para aquellos ámbitos donde la privacidad o necesidades específicas en cuanto a sensibilidad de la información, existe la opción de adaptar soluciones basadas en código abierto. Estos modelos, además de garantizar la privacidad, pueden ser entrenados y ajustados de acuerdo con necesidades específicas, convirtiéndose en un área de investigación en constante desarrollo.

Interés de la IA generativa y los LLMs, desde el punto de vista de Defensa

El avance y despliegue de la IA generativa en el ámbito de la Defensa representa una confluencia innovadora de aplicaciones emergentes con exigencias únicas de seguridad. En este sentido, es posible diferenciar tres dimensiones principales: la perspectiva de aplicación interna en la organización, la aplicación operativa en misiones y capacidades militares, y las potenciales utilizaciones por parte de adversarios.

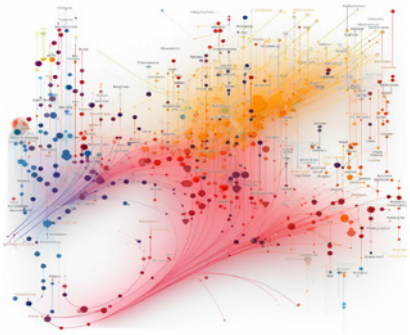

Imagen generada con IA

Desde la perspectiva de la organización

Las organizaciones de Defensa cuentan con vastos conjuntos de datos y un profundo conocimiento especializado. Los LLMs tienen el potencial de optimizar y amplificar estas capacidades humanas. A continuación, se presentan algunas de las potenciales aplicaciones:

- Acceso a la información: los LLMs pueden actuar como expertos digitales, facilitando búsquedas en lenguaje natural, a través de extensas bases de datos y documentos, potenciando *chatbots* y asistentes virtuales para el personal.
- Productividad y eficiencia: estos modelos pueden generar código de programación, redactar informes, procesar documentos y gestionar tareas administrativas.
- Generación y revisión documental: pueden redactar memorandos, informes, correos y notas informativas, a partir de simples indicaciones, así como revisar y resumir documentos extensos, facilitando la preparación para reuniones y gestionando el conocimiento de manera más eficiente.
- Logística: los LLMs pueden contribuir a la optimización logística, desde la gestión del transporte hasta la coordinación de cadenas de suministro.
- Comunicación: estos modelos pueden traducir comunicaciones automáticamente a otros idiomas, resumir informes extensos o filtrar comunicaciones entrantes.
- Revisión legal: son capaces de escanear y revisar contratos y políticas para garantizar el cumplimiento normativo, identificando áreas que requieran atención legal.

Desde la perspectiva de las aplicaciones militares

Los LLMs podrán ofrecer ventajas transformadoras en las operaciones militares, principalmente en su capacidad de conectar, de manera rápida, datos dispares para generar conocimientos y capacidades útiles.

- Inteligencia: analizan y relacionan inteligencia de múltiples fuentes para identificar amenazas y tendencias, generando informes adaptados a diferentes audiencias.
- Logística y mantenimiento: optimizan logísticas complejas e interacción en lenguaje natural con manuales técnicos para solucionar problemas en sistemas.
- Aprendizaje automático y datos sintéticos: crean datos de sensores simulados realistas y analizan datos brutos para extraer características clave.
- Fusión de datos: absorben datos de múltiples sensores y fuentes para proporcionar una imagen integrada del campo de batalla, fusionando datos de inteligencia con registros operativos.
- Ciberdefensa, como por ejemplo emular comportamientos de *hackers* para encontrar vulnerabilidades de manera proactiva y sintetizar patrones de tráfico de red para detectar amenazas internas.

Potenciales usos adversos de la IA generativa y LLMs

La utilización de la IA generativa y los LLMs por parte de adversarios puede plantear serias amenazas y desafíos para la defensa. Es crucial considerar estos riesgos y prepararse adecuadamente. A continuación, se abordan los principales aspectos de preocupación y posibles contramedidas:

- Integridad de la información: los adversarios pueden utilizar medios sintéticos y generación automática de texto para difundir información manipulada o falsa con fines propagandísticos. Es esencial desarrollar métodos para detectar contenidos generados y rastrear su origen.
- Ataques cibernéticos: existen métodos avanzados para automatizar ataques de *phishing*, usurpación de identidad y *hacking*, explorando sistemas de manera inteligente. Las pruebas continuas de *red team* y el análisis de comportamiento para detectar amenazas son cruciales.
- Contrainteligencia: los adversarios podrían analizar patrones y generar información a partir de fuentes abiertas que revelen datos sensibles. Es vital controlar estrictamente la compartición de datos y las entradas proporcionadas a modelos generativos.
- Amenazas asimétricas: existen riesgos de uso de la tecnología generativa para facilitar la exploración de vulnerabilidades, el desarrollo de simulaciones de conflicto, armas o capacidades cibernéticas.

Desafíos IA Generativa y Modelos de Lenguaje a Gran Escala

- Reducción y medición de «alucinaciones». Las alucinaciones en LLMs aluden a la generación de información no basada en datos de entrada o inexacta. Es esencial reducir estos errores, pues comprometen la fiabilidad del modelo.
- Optimización del contexto y su construcción. Una de las limitaciones de los LLMs es la ventana de contexto fijo en la que operan, determinando cuánta información puede considerar el modelo simultáneamente. Optimizar este contexto permitirá obtener respuestas más detalladas y completas. Sin embargo, no basta con aumentar su longitud, es fundamental mejorar la construcción del contexto, seleccionando partes relevantes del texto y descartando detalles superfluos.
- Incorporación de otras modalidades de datos. La multimodalidad es un futuro desarrollo para los LLMs. Aunque principalmente tratan datos textuales, es esencial integrar *inputs* visuales, auditivos y otros sensoriales. Al integrarlos, los LLMs se volverán más versátiles, permitiendo aplicaciones como descripciones de imágenes/vídeos, transcripciones multilingües y robótica asistida. Asentar el lenguaje en múltiples *inputs* sensoriales garantiza una comprensión y generación más ricas.
- LLMs más rápidos y asequibles. Los LLMs, especialmente las arquitecturas más grandes, requieren enormes recursos computacionales. Los avances futuros deben enfocarse en métodos de entrenamiento más eficientes. Estos, junto con mejores arquitecturas y códigos de fuente abierta, democratizarán el acceso, permitiendo entrenar modelos personalizados a un costo razonable.
- Nuevas arquitecturas de modelos. Aunque las arquitecturas actuales han demostrado ser eficaces, siempre hay margen de mejora. Nuevas arquitecturas podrían priorizar diferentes aspectos del aprendizaje, desde capacidades de razonamiento hasta el manejo de datos multimodales. Explorar nuevas arquitecturas garantizará que los LLMs se beneficien de una mejor generalización y menos predicciones sin sentido.
- Desarrollar alternativas a *Graphical Processing Unit* (GPUs). Las GPUs han sido el pilar de los avances en aprendizaje profundo. Sin embargo, con el crecimiento de los LLMs, es esencial desarrollar alternativas que manejen eficientemente estas computaciones, como *hardware* especializado o soluciones de computación distribuida.
- Creación de agentes. Con el avance de los LLMs, surge el potencial de usarlos como agentes colaborativos. Esto permitiría simular interacciones grupales, negociaciones y dinámicas sociales complejas, siendo útil en simulaciones y aplicaciones del mundo real. Hacer estos agentes usables implica que sean intuitivos, fiables y fácilmente integrables.

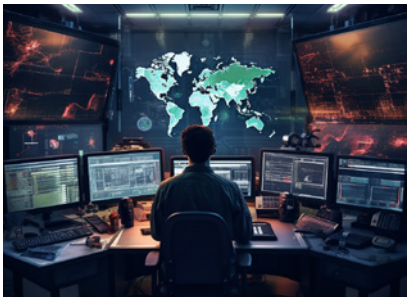

Imagen generada con IA

Conclusión

Los modelos de lenguaje a gran escala y la inteligencia artificial generativa están tomado un papel emergente en el desarrollo tecnológico contemporáneo, ofreciendo un amplio potencial de aplicaciones, desde la generación y gestión de contenido hasta aplicaciones específicas en ámbitos críticos como la defensa. Los LLMs no solo se destacan por su habilidad de gestionar y generar texto, sino que también investigan terrenos emergentes en la manipulación de datos visuales y de otro tipo. La exploración de estas tecnologías de IA generativa y sus aplicaciones, con un ojo crítico hacia sus potenciales usos adversos y desafíos éticos y prácticos, se considera esencial para navegar por el futuro de la IA y su aprovechamiento para el interés de la defensa.

Este artículo ha sido redactado con la ayuda de Anthropic Claude-2 y de OpenAI CHAT-GPT 4, bajo las instrucciones del autor.