Apuntes Semana 12

Apuntes del 10 de octubre de 2025
Andrey Ureña Bermúdez – 2022017442
Inteligencia Artificial
Escuela de Computación, Instituto Tecnológico de Costa Rica
Correo: andurena@estudiantec.cr

***Abstract*—Estos apuntes corresponden a la Semana 12 del curso de Inteligencia Artificial, impartido por el profesor Steven Pacheco Portugués en el Instituto Tecnológico de Costa Rica. Se abordan los temas relacionados con los modelos de lenguaje de gran escala (LLM), la tokenización, embeddings, y la introducción al paradigma de *Retrieval-Augmented Generation* (RAG) y agentes inteligentes. Además, se presentan los anuncios del curso y el cronograma restante del semestre.**

I. Introducción

Durante esta sesión, se revisaron aspectos fundamentales de los modelos de lenguaje modernos y su relación con las arquitecturas de inteligencia artificial actuales. También se analizaron conceptos claves para comprender cómo los LLM procesan texto, transforman información en vectores, y aplican técnicas de recuperación de conocimiento externo mediante RAG. Finalmente, se discutieron las implicaciones éticas y el uso responsable de estos sistemas.

II. Anuncios del curso

- Se asignó la **Tarea 04** sobre agentes, con fecha de entrega el 6 de noviembre. La revisión será presencial y consiste en la creación de un agente funcional.
- Se presentó el cronograma para el cierre del semestre, organizado por semanas:
  - **Semana 13:**
    - **Martes 28 de octubre:** Quiz 6 y tema *Quantization – Unsupervised*.
    - **Jueves 30 de octubre:** Tema *Unsupervised – PCA* y entrega del Proyecto I.
  - **Semana 14:**
    - **Martes 4 de noviembre:** Revisión presencial del Proyecto I.
    - **Jueves 6 de noviembre:** Revisión presencial del Proyecto I y entrega de la Tarea 04: *Agentes*.
  - **Semana 15:**
    - **Martes 11 de noviembre:** Clase virtual sobre *Unsupervised – PCA*, asignación del Proyecto II y la Tarea 05: *Autoencoder – Quantization*.
    - **Jueves 13 de noviembre:** Revisión virtual de la tarea de agentes.
  - **Semana 16:**
    - **Martes 18 de noviembre:** Tema *Riesgos de la Inteligencia Artificial*.

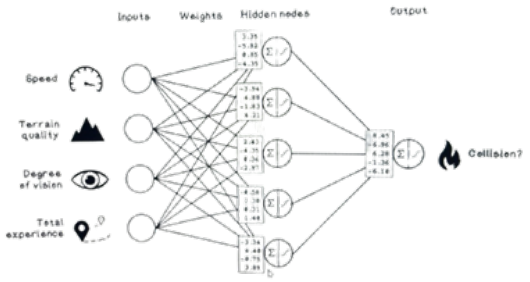

Figura 1. Ejemplo de un modelo de red neuronal preentrenado.

  - **Semana 17:** Semana colchón (sin actividades programadas).
  - **Semana 18:**
    - **Martes 2 de diciembre:** Examen I.
    - **Jueves 4 de diciembre:** Entrega del Proyecto II.

III. Repaso de conceptos

A. Modelos de Lenguaje de Gran Escala (LLM)

Los LLM se han convertido en la base de los sistemas modernos de inteligencia artificial. Permiten generar, comprender y razonar sobre texto, código, imágenes y audio.

Cada entrada (input) es representada mediante valores numéricos en punto flotante que describen características. El tratamiento varía según si la entrada corresponde a texto, números o símbolos.

B. Tokenización

La tokenización convierte las palabras, signos o símbolos en representaciones numéricas llamadas *tokens*. Estos tokens permiten al modelo procesar texto de manera eficiente.

Existen varios tipos de tokenización, resumidos en la Tabla I.

Tabla I
TIPOS COMUNES DE TOKENIZACIÓN Y SUS PRINCIPALES VENTAJAS.

Tipo	Ejemplo	Ventaja principal
Palabra	"Los medios"	Simplificada
Carácter	"L", "o", "s"	Sin OOVs
Subpalabra (BPE, WordPiece)	"super" + "vivencia"	Equilibra vocabulario/contexto
Byte-level	bytes UTF-8	Soporta cualquier símbolo
Espacio en blanco	"Hola", "mundo"	Rápido y simple

Tras la tokenización, los tokens se representan como vectores en un espacio continuo. Esto permite medir similitud semántica entre palabras.

C. Métricas de similitud

Las métricas más utilizadas incluyen:

- **Distancia euclidiana:** mide qué tan separados están dos puntos en el espacio vectorial.
- **Similitud del coseno:**

$$\text{Sim}(a, b) = \frac{a \cdot b}{||a||\,||b||}$$

Evalúa el ángulo entre los vectores; un ángulo menor implica mayor similitud.

Tabla II
EJEMPLO SIMPLIFICADO DE TOKENIZACIÓN: LAS PALABRAS SE TRANSFORMAN EN TOKENS CON IDENTIFICADORES NUMÉRICOS.

Palabra	Token	ID Numérico
Los	los	105
LLM	llm	2124
aprenden	aprenden	893
patrones	patrones	5749

D. Embeddings

Los *embeddings* son representaciones numéricas densas que asignan a cada token un vector en un espacio continuo de alta dimensión. Capturan significado semántico y relaciones contextuales entre palabras u oraciones completas, permitiendo comparaciones más profundas entre ideas o documentos.

E. Capacidades de los LLM

Debido a su entrenamiento a gran escala y arquitecturas basadas en Transformers, los LLM presentan capacidades emergentes:

- Comprensión contextual.
- Generación coherente de texto.
- Razonamiento y planificación básica.
- Aprendizaje en el prompt (*in-context learning*).
- Multitarea sin reentrenamiento.
- Conocimiento estático derivado de los datos de entrenamiento.
- Costos computacionales elevados.

IV. MATERIA NUEVA: RETRIEVAL-AUGMENTED GENERATION (RAG)

Un sistema RAG conecta un LLM con un módulo de recuperación de información (*retriever*) para incorporar conocimiento externo relevante durante la generación de respuestas.

A. Chunks

El texto se divide en fragmentos denominados *chunks*, que suelen contener entre 200 y 500 tokens. Cada fragmento se transforma en un vector mediante un modelo de embeddings, capturando su significado semántico.

B. Consulta o recuperación

Dada una consulta, el sistema convierte la pregunta en un embedding y calcula la similitud con los embeddings indexados, devolviendo los más cercanos semánticamente.

C. Aumento y generación

Los fragmentos recuperados se integran en el prompt enviado al LLM, proporcionando contexto adicional que guía la respuesta hacia información verificada y relevante.

D. Ventajas principales

- Reducción de alucinaciones.
- Actualización continua del conocimiento.
- Eficiencia de costos en entrenamiento.
- Aplicabilidad en dominios especializados.
- Asistentes empresariales enriquecidos.
- Soporte a la investigación y atención al cliente.

V. LLM TRADICIONAL VS AGENTE INTELIGENTE

Un LLM tradicional puede ofrecer información general, pero carece de personalización y acción. Por ejemplo, si se le consulta "¿Cuántos días de vacaciones me quedan?", no podrá responder con precisión al no tener acceso a datos personales.

En cambio, un agente inteligente integra:

- **Memoria:** recuerda preferencias y contextos previos.
- **Herramientas:** accede a APIs externas (clima, vuelos, calendario).
- **Planificación:** organiza y ejecuta tareas en función de objetivos.
- **Acción:** transforma planes en resultados concretos.

Este paradigma refleja la evolución hacia sistemas que razonan y actúan, más allá de solo responder texto.

VI. ESCALAMIENTO RESPONSABLE

Es fundamental evaluar cuándo realmente se requiere escalar de un modelo LLM a un sistema de agentes o multiagentes. Esto implica garantizar seguridad, privacidad y el uso ético de los datos. Los agentes deben ser diseñados bajo principios de transparencia y responsabilidad.

VII. Conclusión

Los temas revisados durante esta semana refuerzan la comprensión de cómo los modelos de lenguaje modernos procesan información y cómo se están extendiendo hacia arquitecturas más complejas y útiles, como los sistemas RAG y los agentes inteligentes. Estas herramientas representan un paso clave hacia una inteligencia artificial más contextual, adaptable y responsable.

Referencia

Pacheco Portuguez, S. (2025). *Presentación del curso de Inteligencia Artificial*. Instituto Tecnológico de Costa Rica.