Manual de Estudio: AWS Certified AI Practitioner

Dominio 2.1: Conceptos Básicos de la IA Generativa

Material de Preparación Detallado

1. Definición y Alcance de la IA Generativa

La **IA generativa** es un subconjunto del aprendizaje profundo (*Deep Learning*). A diferencia de la IA tradicional, que se centra en clasificar o predecir basándose en datos existentes, la IA generativa se enfoca en la creación de contenido nuevo y original, como texto, imágenes, audio, vídeo y código.

1.1. Características Principales

- **Aprendizaje de Patrones:** Los modelos aprenden representaciones estadísticas a partir de vastos conjuntos de datos de entrenamiento.
- **Modelos Fundacionales (FM):** Son redes neuronales complejas con miles de millones de parámetros entrenadas con petabytes de datos no estructurados.
- **Naturaleza Probabilística:** El modelo no busca una respuesta única, sino que realiza una suposición estadística de cuál debería ser la siguiente palabra o token.

2. Arquitectura de Transformadores

Introducida en el artículo *"Attention is All You Need"* (2017), la red de transformadores es el elemento principal de la IA generativa actual.

2.1. Mecanismo de Autoatención (Self-Attention)

Este mecanismo permite al modelo ponderar la importancia de diferentes partes de la entrada al generar cada token de salida.

- **Dependencias de largo alcance:** Supera las limitaciones de las redes neuronales recurrentes (RNN) al capturar relaciones contextuales distantes en el texto.
- **Cálculo:** Utiliza vectores de consulta, clave y valor para determinar las ponderaciones de atención mediante productos escalares.

2.2. Incrustaciones y Vectores (Embeddings)

- **Tokenizador:** Convierte el texto humano en IDs de entrada que representan tokens en el vocabulario del modelo.
- **Vectores:** Listas ordenadas de números que representan características. Permiten capturar asociaciones semánticas y jerarquías.
- **Incrustaciones de Posición:** Codifican la posición relativa de cada token en la secuencia, lo que ayuda al modelo a entender el orden de las palabras sin necesidad de operaciones recurrentes.

3. Modelos de Difusión y Multimodalidad

3.1. Unimodal vs. Multimodal

- **Unimodal:** Funciona con una sola modalidad (ej. LLMs que solo procesan y generan texto).
- **Multimodal:** Puede entender y generar combinaciones de diversos orígenes de datos (texto a imagen, imagen a descripción).

3.2. Modelos de Difusión

Son una clase de modelos generativos que aprenden a revertir un proceso de ruido gradual.

- **Difusión Directa:** Añade ruido gaussiano a una imagen.
- **Difusión Inversa:** El modelo predice y elimina el ruido de forma iterativa para producir un resultado coherente.
- **Difusión Estable (Stable Diffusion):** A diferencia de otros, opera en un **espacio latente** de definición reducida en lugar del espacio de píxeles.

4. Conceptos de Inferencia y Prompting

- **Petición (Prompt):** Entrada enviada al modelo en tiempo de inferencia.
- **Ventana de Contexto:** El límite de información que el modelo puede procesar en una sola inferencia.
- **Aprendizaje en Contexto:** Incluir ejemplos dentro de la petición para guiar al modelo. Puede ser *Zero-shot* (sin ejemplos), *One-shot* (un ejemplo) o *Few-shot* (varios ejemplos).

5. Ciclo de Vida de un Proyecto de IA Generativa

AWS establece un marco de trabajo desde la concepción hasta el lanzamiento:

1. **Identificar el caso práctico:** Definir objetivos y ámbito.
2. **Selección de modelo:** Decidir entre entrenar desde cero o usar un modelo fundacional existente (ej. vía SageMaker JumpStart).
3. **Adaptación y Refinamiento:** Realizar un aprendizaje supervisado adicional (*Fine-tuning*).
4. **Alineación Humana:** Aplicar **RLHF** (Aprendizaje por refuerzo a partir de comentarios humanos).
5. **Evaluación e Iteración:** Probar métricas y optimizar el despliegue.

6. Limitaciones Fundamentales

A pesar de su potencia, los LLMs enfrentan desafíos como:

- **Alucinaciones:** Invención de información cuando el modelo desconoce la respuesta.
- **Razonamiento Complejo:** Capacidades limitadas en matemáticas y lógica estricta.