Ingeniero en Software y Tecnologías Emergentes

Inteligencia Artificial

Tokens e incrustaciones

ISyTE

Dr. Héctor Zatarain Aceves

hector.zatarain@uabc.edu.mx

Introducción

- Los tokens y las incrustaciones son la base de los grandes modelos lingüísticos (LLM) como GPT y BERT.

- Estos conceptos pueden parecer complejos, pero son esenciales para comprender cómo la IA moderna interactúa con el lenguaje.

- En esta presentación se describirán qué son los tokens y las incrustaciones, cómo funcionan y sus aplicaciones en ambos modelos lingüísticos.

¿Qué son los tokens y por qué son importantes?

Tokens: Los componentes básicos de los modelos lingüísticos

Los fragmentos son los fragmentos de texto más pequeños que procesan los modelos de lenguaje. En lugar de leer oraciones o párrafos a la vez, los modelos de lenguaje dividen el texto en fragmentos, que podrían ser:

- Palabras completas (por ejemplo, “manzana”)
- Partes de palabras (por ejemplo, “appl” y “e” para “manzana”)
- Caracteres (p. ej., “a”, “p”, “p”, “l”, “e”)
- Signos de puntuación (por ejemplo, “!”, “.”, “,”)
- Símbolos especiales como <s> o <|endoftext|> que ayudan a los modelos a comprender el contexto.

¿Por qué es importante la tokenización?

- La tokenización es crucial porque transforma el texto legible a un formato que los LLM pueden comprender.

- Antes de que cualquier modelo procese la entrada, un tokenizador descompone el texto en tokens y los convierte en números.

- Estos números son los que el modelo utiliza para realizar los cálculos.

¿Cómo funcionan los tokenizadores?

Los tokenizadores preparan el texto para los LLM dividiéndolo en tokens y convirtiéndolos en **identificadores de token**. Veamos cómo funciona esto:

1. **Texto de entrada:** Imagina que escribes:
   “Escribe un correo electrónico disculpándote con Sarah por el trágico accidente de jardinería”.

2. **Tokenización:** El tokenizador divide el texto en tokens:
   - <s>(un símbolo de inicio especial)
   - “Escribir”, “un”, “correo electrónico”, etc.

3. **ID de token:** cada token se convierte en un identificador numérico único:
   - <s>→1
   - “Escribe” →14350
   - “correo electrónico” →5281

4. Luego, estos identificadores de token se introducen en el LLM para su procesamiento.

Tipos de tokenización

Hay varias formas de tokenizar texto, según el modelo de lenguaje:

1. Tokenización de palabras
2. Tokenización de subpalabras
3. Tokenización de personajes
4. Tokenización de bytes

Tipos de tokenización: Tokenización de palabras

1. Tokenización de palabras

Divide el texto en palabras completas. Ejemplo:

«Me encanta la IA» → `["I", "love", "AI"]`

- **Pros** : Fácil de entender.
- **Contras** : Tiene dificultades con palabras nuevas o variaciones (por ejemplo, “loving” vs. “loved”).

Tipos de tokenización: Tokenización de subpalabras

2. Tokenización de subpalabras

Divide el texto en fragmentos más pequeños. Ejemplo:

"amoroso" → `["lov", "ing"]`

- **Ventajas** : Eficaz para palabras nuevas. Admite variaciones como "love", "loving" y "loved".
- **Contras** : Más complejo que la tokenización de palabras.

Tipos de tokenización: Tokenización de personajes

3. Tokenización de personajes

Divide el texto en caracteres individuales. Ejemplo:

«AI» → `["A", "I"]`

- **Ventajas** : maneja cualquier texto, incluso palabras raras o inventadas.
- **Contras** : Requiere más tokens para representar oraciones.

Tipos de tokenización: Tokenización de bytes

4. Tokenización de bytes

Divide el texto en bytes (una representación de caracteres de nivel inferior).

- **Ventajas** : Funciona bien con idiomas con muchos caracteres (como el chino).
- **Contras** : Es computacionalmente más costoso.

¿Qué son las incrustaciones y por qué son importantes?

Incrustaciones: Convertir tokens en números

Una vez tokenizado el texto, las incrustaciones convierten esos tokens en representaciones numéricas densas llamadas vectores. Estos vectores capturan el significado del texto de forma que las computadoras puedan comprenderlo.

- **Ejemplo** : La palabra “rey” podría representarse como `[0.59, 0.77, 0.19, ...]`.

¿Cómo funcionan las incrustaciones?

Las incrustaciones se almacenan en una tabla grande dentro del modelo de lenguaje. Cada ID de token corresponde a un vector de incrustación específico.

Estas incrustaciones se ajustan durante el entrenamiento para ayudar al modelo a comprender las relaciones entre las palabras.

Incrustaciones estáticas vs. contextuales

- **Incrustaciones estáticas** (por ejemplo, Word2Vec, GloVe): una palabra siempre tiene el mismo vector independientemente del contexto.

- **Incrustaciones contextuales** (por ejemplo, GPT, BERT): el vector de una palabra cambia según la oración en la que se encuentra, capturando matices en el significado.

Aplicaciones de tokens e incrustaciones

- 1. Comprensión del lenguaje
- 2. Búsqueda semántica
- 3. Sistemas de recomendación
- 4. Aplicaciones multimodales

Aplicaciones de tokens e incrustaciones

1. Comprensión del lenguaje

Las incrustaciones ayudan a los modelos a comprender y generar texto coherente. Por ejemplo:

- Predecir la siguiente palabra en una oración.
- Resumir documentos largos.

2. Búsqueda semántica

Las incrustaciones de texto permiten que los sistemas comparen significados en lugar de palabras exactas.

Buscar "mejores smartphones" podría mostrar resultados como "mejores teléfonos móviles".

Aplicaciones de tokens e incrustaciones

3. Sistemas de recomendación

- Los tokens y las incrustaciones se utilizan en sistemas como Spotify y Netflix para recomendar contenido basado en similitudes.

4. Aplicaciones multimodales

Las incrustaciones conectan diferentes tipos de datos.

Por ejemplo:
Las incrustaciones de texto combinadas con incrustaciones de imágenes habilitan herramientas como DALL·E.

De Word2Vec a los LLM modernos

- Antes de LLM como GPT y BERT, métodos como Word2Vec dominaban el PLN. Word2Vec utiliza una técnica llamada skip-gram para predecir relaciones entre palabras. Si bien era eficaz, carecía de la sofisticación de las incrustaciones contextuales modernas.

- Hoy en día, los LLM producen incrustaciones mucho más ricas, lo que permite aplicaciones avanzadas como:

    - Chatbots
    - Análisis de sentimientos
    - Generación de código

Conclusiones

- **Los tokens** son las unidades básicas de texto que procesan los LLM.
- **Las incrustaciones** son representaciones numéricas de tokens que ayudan a los LLM a comprender las relaciones y el contexto.
- Los **tokenizadores** varían según el método y están adaptados a tareas y conjuntos de datos específicos.
- Las **incrustaciones modernas** permiten aplicaciones potentes como búsqueda semántica, recomendaciones e IA multimodal.

ISyTE

Dr. Héctor Zatarain Aceves

Email: hector.zatarain@uabc.edu.mx

Teléfono: (646) 152 8244 Ext. 64350