

Trabajos de la asignatura **Programación Avanzada en Bionformática**

Grado en Ingeniería de la Salud

Curso 2012-13



Prólogo

Este libro, editado por el Prof. Alberto G. Salguero Hidalgo, contiene los trabajos realizados por los alumnos de la asignatura Programación Avanzada en Bioinformática, del Grado en Ingeniería de la Salud de la Universidad de Málaga.

Los alumnos son plenos responsables de su contribución al libro y conservan todos los derechos de autoría del contenido de sus respectivos capítulos.

Índice general

1. Grandes modelos de lenguaje	7
--	---

graphicx float

Capítulo 1

Grandes modelos de lenguaje

ALEJANDRO SILVA RODRÍGUEZ

1.1. Introducción

Los grandes modelos de lenguaje (LLM) son una categoría de modelos avanzados entrenados con enormes conjuntos de datos para comprender y generar lenguaje natural y otros tipos de contenido, permitiendo una amplia gama de aplicaciones.

Estos modelos se han vuelto mundialmente conocidos al llevar la IA generativa al centro de atención pública y empresarial. Aunque parecen recientes para algunos, empresas como IBM han estado implementándolos durante años para mejorar las capacidades de comprensión y procesamiento del lenguaje natural (NLU y NLP), aprovechando avances en aprendizaje automático, algoritmos y redes neuronales.

Los LLM representan un avance crucial en PLN y IA, siendo accesibles a través de interfaces como ChatGPT-3 y GPT-4 de OpenAI, Meta's Llama, y modelos de Google como BERT/RoBERTa y PaLM. Estos modelos están diseñados para comprender y generar texto como humanos, basándose en datos masivos. Pueden traducir, resumir, responder preguntas, ayudar en tareas creativas y más, gracias a sus miles de millones de parámetros. Estos modelos avanzados pueden nutrirse de enormes cantidades de datos, incluyendo registros de Internet como Common Crawl, con más de 50 mil millones de páginas web, y Wikipedia, que cuenta con alrededor de 57 millones de páginas.

Los LLM están transformando aplicaciones como chatbots, asistentes virtuales, generación de contenido y más. Evolucionarán continuamente, remodelando nuestra interacción con la tecnología y el acceso a la información en el mundo digital moderno.

1.2. Composición de Modelos de Lenguaje Grandes

1.2.1. Antecedentes

<https://www.ibm.com/mx-es/topics/recurrent-neural-networks>

Antes de la llegada de los Transformers, los modelos de lenguaje basados en redes neuronales recurrentes (RNN) eran dominantes en tareas de procesamiento de lenguaje natural (NLP). Estos modelos, como las RNN y las variantes como las LSTM (Long Short-Term Memory) y las GRU (Gated Recurrent Units), tenían la capacidad de procesar secuencias de datos, como palabras o caracteres, de manera progresiva, manteniendo una especie de "memoria" de los pasos anteriores.

Sin embargo, los modelos basados en RNN tenían varias limitaciones:

1. Problema de desvanecimiento/exploración del gradiente: Durante el entrenamiento, los gradientes (que indican cómo deben actualizarse los pesos de la red) pueden volverse muy pequeños o muy grandes a medida que se propagan hacia atrás a través de las capas de la red recurrente. Esto puede hacer que los modelos olviden información a largo plazo o que se vuelvan inestables (explosión del gradiente).
2. Ineficiencia en paralelización: Debido a su naturaleza secuencial, las RNN no pueden procesar múltiples partes de una secuencia simultáneamente, lo que limita su capacidad para aprovechar el paralelismo y, por lo tanto, su eficiencia computacional.
3. Incapacidad para capturar relaciones a largo plazo: A pesar de las variantes como las LSTM y las GRU, las RNN todavía tenían dificultades para capturar dependencias a largo plazo en secuencias de datos, lo que limitaba su capacidad para comprender contextos complejos en tareas de NLP.
4. Costo computacional y tiempo de entrenamiento: Entrenar modelos basados en RNN puede ser computacionalmente costoso y llevar mucho tiempo, especialmente cuando se trabaja con grandes conjuntos de datos y arquitecturas complejas.



Figura 1.1: RNN Procesando oración

El salto a los Transformers se dio principalmente debido a la introducción de la arquitectura de atención. Los Transformers abordan muchas de las limitaciones de las RNN:

1. Atención: Los Transformers utilizan mecanismos de atención que les permiten procesar palabras en paralelo en lugar de secuencialmente, lo que mejora significativamente la eficiencia y la capacidad de capturar relaciones a largo plazo.
 2. Paralelización completa: Gracias a la atención y a la estructura basada en autoatención, los Transformers pueden paralelizar completamente el cálculo en todas las posiciones de la secuencia, lo que acelera tanto el entrenamiento como la inferencia.

1.2.2. Transformer

Un transformer se compone de dos partes principales, encoder y decoder, cuyos componentes profundaremos más adelante:

El encoder toma una secuencia de entrada y la transforma en vectores de representación, capturando las relaciones entre las palabras mediante capas de autoatención. Su objetivo principal es procesar la información de entrada y generar representaciones vectoriales contextualizadas que contienen información sobre cada elemento de la secuencia. Por otro lado, el decoder utiliza estos vectores de representación generados por el encoder, junto con la información condicional, para generar una secuencia de salida. También utiliza capas de autoatención, pero su enfoque se centra en capturar las relaciones entre las palabras en la secuencia de salida. Su función principal es generar una secuencia de salida coherente basada en la información proporcionada por el encoder y en los tokens previamente generados.

La primera fase a la que se someten los datos al entrar a un transformer es la tokenización

Embeddings (Incrustaciones)

Los embeddings son representaciones vectoriales de palabras o tokens en un espacio dimensional. En el contexto de los LLM, las embeddings se utilizan para transformar las palabras o tokens del vocabulario en vectores numéricos densos. Estos vectores capturan relaciones semánticas y sintácticas entre las palabras, lo que ayuda al modelo a aprender patrones y generalizar mejor sobre el lenguaje.

Embedding aims at creating a vector representation of words. Words that have the same meaning will be close in terms of euclidian distance. For example, the word bathroom and shower are associated with the same concept, so we can see that the two words are close in Euclidean space, they express similar senses or concept. <https://becominghuman.ai/attention-is-all-you-need-16bf481d8b5c>

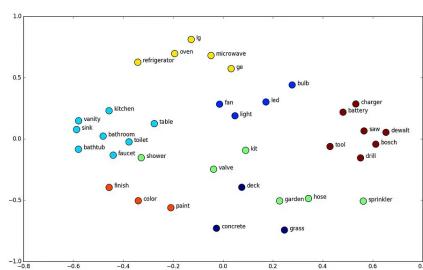


Figura 1.2: Embedding vector representation of words

Positional Encoding

Justo después de esto, las palabras se pasan por un positional encoding, que le añade al token información sobre la posición de la palabra. En esencia, el positional encoding asigna un vector único a cada posición en la secuencia, de modo que palabras o elementos que ocupan posiciones diferentes tendrán representaciones distintas en el espacio vectorial, a pesar de que sus características semánticas sean similares. Se utiliza para abordar la falta de información sobre la posición en modelos como los transformers, donde la entrada se representa mediante embeddings que no contienen información sobre la ubicación de los tokens en la secuencia. Anteriormente, los modelos recurrentes obtenían información de la posición de las palabras mediante la iteración en el texto, esta etapa fue la revolucionaria que impulsó el avance de los transformers.

Capas de Atención (Attention Layers)

Las capas de atención son componentes clave dentro de los transformadores. Permiten que el modelo "preste atención"^a diferentes partes de la secuencia de entrada durante el procesamiento. Esta atención se calcula dinámicamente para cada par de elementos en la secuencia, lo que ayuda al modelo a capturar relaciones contextuales y a decidir qué partes de la entrada son más relevantes para la tarea en cuestión.

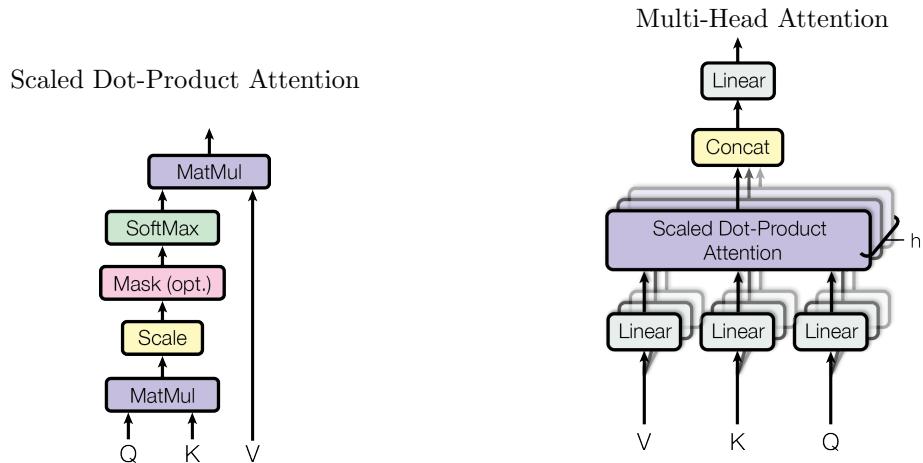


Figura 1.3: (izquierda) Scaled Dot-Product Attention. (derecha) Multi-Head Attention consists of several attention layers running in parallel.

El proceso comienza con tres conjuntos de vectores: consultas, claves y valores. Por así decirlo, el vector clave sera un identificador que describe las propiedades de una palabra, mientras que el vector consulta describe las propiedades que busca una palabra. Estos vectores se derivan de la entrada original y se utilizan para calcular la atención. Primero, se calcula la similitud entre la consulta y cada clave utilizando el producto escalar obteniendo el vector de atención, y luego se aplica una función softmax para obtener pesos de atención normalizados. Estos pesos determinan cuánta atención se asigna a cada valor correspondiente. Finalmente, se calcula una combinación lineal ponderada de los valores utilizando los pesos de atención, produciendo así la salida de la capa de atención.

El resultado será de alguna forma a que palabras muestra atención cuando se fija en una, dando una capacidad extraordinaria para reconocer el contexto.

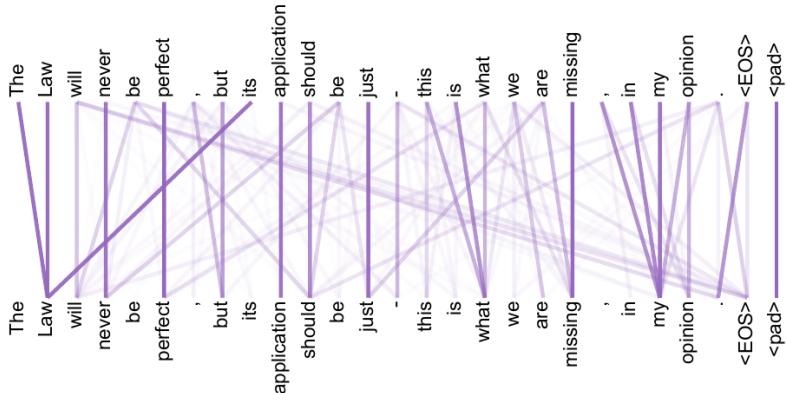


Figura 1.4: Representación del mecanismo de atención

Ejemplo:

Mary and Paul were running together but he got injured

Cuando el modelo está procesando la secuencia, ¿a qué se refiere «he»? Parece una pregunta muy sencilla para una persona, pero se trata de un problema complejo que ha existido durante bastante tiempo en NLP. El mecanismo de auto-atención permite asociar «he» con Paul en vez de con el resto de palabras.

"Most competitive neural sequence transduction models have an encoder-decoder structure [5, 2, 35]. Here, the encoder maps an input sequence of symbol representations (x_1, \dots, x_n) to a sequence of continuous representations $z = (z_1, \dots, z_n)$. Given z , the decoder then generates an output sequence (y_1, \dots, y_m) of symbols one element at a time. At each step the model is auto-regressive [10], consuming the previously generated symbols as additional input when generating the next."

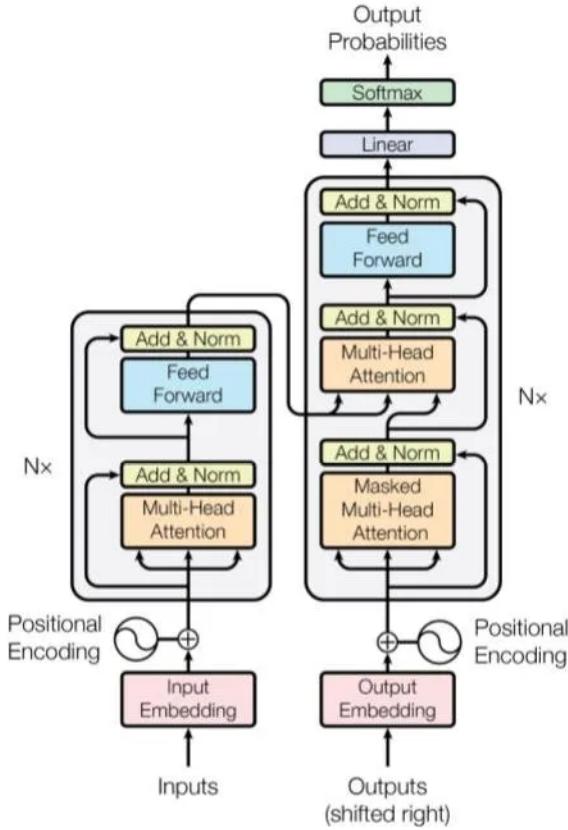


Figura 1.5: Transformer Architecture

Flujo de información

Para resumir y ver como encajan todos los componentes analizaremos el flujo de información en un transformer. Primero, la palabra input es transformada en vector con información semántica y de posición. A continuación entra en el encoder (parte de la izquierda) donde se extrae información de contexto mediante la atención, generando un output. En la segunda etapa, en el decoder (parte de la derecha), la multi-head attention layer es alimentada con el output del encoder, además de utilizar el output de una interacción anterior como input. Al final, estos dos componentes trabajarán juntos para predecir la palabra con mayor probabilidad de todo el vocabulario apoyado de una capa Softmax y One-hot Encoding.

1.2.3. Arquitecturas más importantes

Generative Pre-trained Transformer (GPT)

Es un modelo de lenguaje desarrollado por OpenAI basado en la arquitectura Transformer. GPT utiliza una estructura unidireccional en la que las capas Transformer procesan la secuencia de entrada en una dirección. Cada bloque Transformer en GPT consta de capas de atención multi-cabeza seguidas de redes neuronales feedforward. Es un modelo autoregresivo que genera texto palabra por palabra utilizando un token especial de inicio y predice la siguiente palabra en la secuencia. GPT se destaca en tareas de generación de texto y comprensión del lenguaje natural. Método de aprendizaje:

- En el método autoregresivo, el modelo genera secuencialmente cada token de salida basado en los tokens previamente generados en la secuencia. Este enfoque busca maximizar la probabilidad conjunta de predecir cada token dado el contexto anterior.

GPT-3 has 175 billion parameters, almost 2,000 times more than the number of parameters in the original GPT-1 model and over 100 times more than the 1.5 billion parameters in GPT-2.
<https://www.techtarget.com/searchenterpriseai/feature/Exploring-GPT-3-architecture>

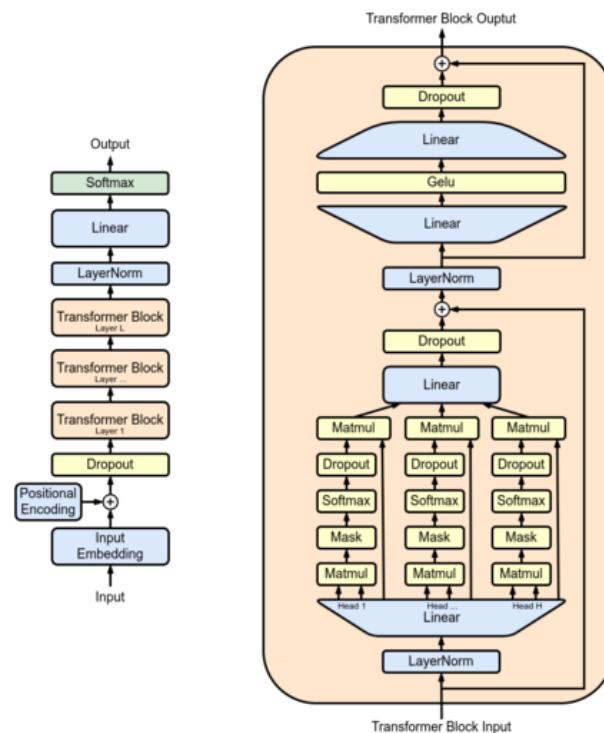


Figura 1.6: GPT Architecture

Bidirectional Encoder Representations from Transformers(BERT)

Desarrollado por Google, es otro modelo basado en la arquitectura Transformer. A diferencia de GPT, BERT utiliza un codificador bidireccional que procesa la secuencia de entrada en ambas direcciones, capturando así mejor el contexto bidireccional de las palabras. BERT se pre-entrena en tareas de Masked Language Modeling (MLM) y Next Sentence Prediction (NSP), lo que le permite aprender representaciones contextuales profundas. Aunque BERT no es autoregresivo y no se utiliza directamente para la generación de texto, se utiliza ampliamente como modelo de representación contextual en tareas de procesamiento del lenguaje natural como clasificación de texto, extracción de información y más. BERT ha demostrado ser muy efectivo en la captura de relaciones semánticas y sintácticas en el texto.

Métodos de aprendizaje:

- Masked Language Modeling (MLM), se enmascaran aleatoriamente algunas palabras en las oraciones de entrada y el modelo debe predecir las palabras enmascaradas basándose en el contexto. Este método fomenta a BERT a entender el contexto de las palabras en relación con las demás en la oración.
- Next Sentence Prediction (NSP) implica alimentar a BERT con pares de oraciones y el modelo debe predecir si la segunda oración sigue a la primera en el texto original. Esto ayuda a BERT a captar la coherencia entre las oraciones y entender la relación entre ellas. Ambos métodos son clave en el pre-entrenamiento de BERT para capturar información contextual y semántica en el texto.

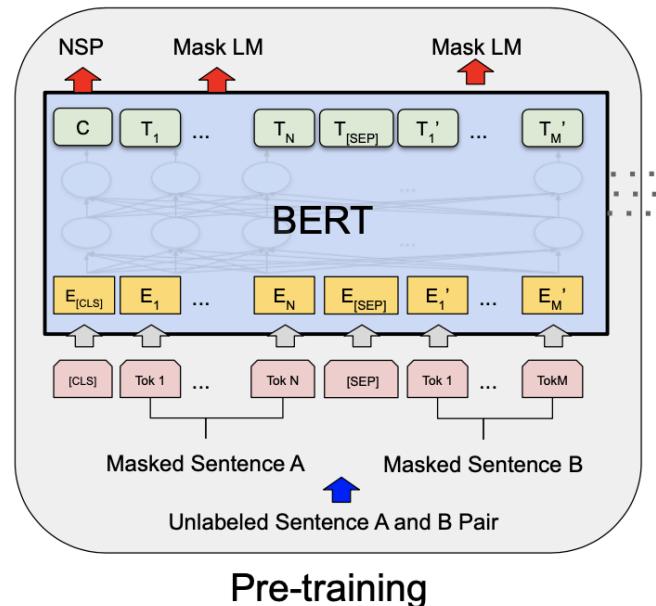


Figura 1.7: Bert Architecture

1.2.4. Pre-entrenamiento

Comienza con la recopilación de enormes cantidades de texto de diversas fuentes, que pueden incluir libros, artículos en línea, redes sociales y más. Este corpus de texto se utiliza para entrenar el modelo de forma no supervisada, lo que significa que el modelo aprende patrones y estructuras del lenguaje sin la necesidad de etiquetas o supervisión explícita. Antes de entrenar, el texto se tokeniza y preprocesa para convertirlo en unidades de entrada comprensibles para el modelo. Durante el entrenamiento, el modelo ajusta sus pesos para predecir palabras o partes de palabras en función del contexto proporcionado por la secuencia de tokens de entrada.

1.3. Aplicaciones de Modelos de Lenguaje Grandes

1.3.1. Generación de Texto

Ejemplos de generación de texto coherente y creativo incluyen la capacidad de los modelos de lenguaje grandes para producir artículos, historias, poemas y respuestas humanizadas. Estas aplicaciones son fundamentales en asistentes virtuales como Siri, chatbots como ChatGPT, y herramientas de escritura asistida que pueden generar contenido automáticamente.

1.3.2. Clasificación de Texto

Los modelos de lenguaje grandes se utilizan ampliamente en tareas de clasificación de texto, como análisis de sentimientos (determinar emociones en textos), detección de spam (identificación de mensajes no deseados), y categorización de noticias o documentos. Estos modelos pre-entrenados se benefician del transfer learning, permitiendo adaptarlos eficazmente a tareas específicas con datos de entrenamiento adicionales.

1.3.3. Traducción Automática

En traducción automática, los modelos de lenguaje grandes como BERT y GPT se utilizan en sistemas neurales para traducir texto entre idiomas. Estos modelos pueden captar relaciones semánticas complejas y producir traducciones precisas y fluidas, lo que ha mejorado significativamente la calidad de los sistemas de traducción automática.

1.3.4. Preguntas y Respuestas

Los modelos de lenguaje grandes son esenciales en sistemas de pregunta-respuesta (QA) basados en texto. Pueden entender preguntas complejas y generar respuestas precisas basadas en el contexto proporcionado. Estos sistemas se aplican en chatbots de atención al cliente, motores de búsqueda mejorados y aplicaciones de asistencia virtual que responden a consultas de manera inteligente y natural.

1.4. Ajuste Fino (Fine-Tuning) de Modelos Descargados

1.4.1. Importancia del Ajuste Fino

El fine-tuning es esencial en todos los modelos de LLM, ya que permite ajustar el modelo pre-entrenado a tareas específicas o dominios de datos particulares. Al realizar el fine-tuning, el modelo puede adaptarse para comprender mejor el contexto y las características de la tarea específica, refinando sus representaciones internas para capturar patrones relevantes en los datos de entrenamiento. Además, reduce la capacidad de computo necesaria y la huella de carbono ya que la gran cantidad de entrenamiento se hace de forma generalizada en la fase de pre entrenamiento.

1.4.2. Ejemplo Práctico de Ajuste Fino

Ejemplificaremos el proceso de ajuste fino utilizando la librería de Hugging Face Transformers, que nos ayudará a descargar, entrenar y utilizar todo tipo de modelos transformers libres con python. En concreto queremos utilizar BERT para

1.5. Futuro de los Modelos de Lenguaje Grandes

Las tendencias emergentes en la investigación y desarrollo de modelos de lenguaje grandes incluyen avances hacia modelos aún más grandes y complejos que puedan capturar y generar un entendimiento más profundo del lenguaje natural. Se espera que los modelos futuros mejoren en áreas como la comprensión contextual, la generación de texto creativo y la adaptación a tareas especializadas.

Los desafíos y consideraciones éticas en el uso de modelos de lenguaje a gran escala son significativos. Esto incluye preocupaciones sobre la privacidad y la seguridad de los datos utilizados para entrenar estos modelos, así como el sesgo y la equidad en las aplicaciones de IA que pueden afectar a diferentes grupos de manera desproporcionada. Otros aspectos éticos incluyen el uso responsable de la tecnología, la transparencia en los procesos de desarrollo y la regulación para garantizar un uso ético y seguro de los modelos de lenguaje grandes en diversos contextos sociales y empresariales.

1.6. Conclusiones

- Resumen de los puntos clave. - Impacto potencial de los modelos de lenguaje grandes en el futuro de la inteligencia artificial y el procesamiento del lenguaje natural.

Consideraciones Adicionales: - Puedes agregar secciones específicas sobre métricas de evaluación para modelos de lenguaje, comparaciones entre diferentes arquitecturas (GPT vs. BERT), o discusiones sobre el aprendizaje transferido (transfer learning) en modelos de lenguaje. - Incluye ejemplos concretos y estudios de casos para ilustrar cada punto. - No olvides revisar y citar fuentes confiables y recientes relacionadas con el tema.

Bibliografía

Goossens, M., Mittelbach, F. and Samarin, A. (1993). *The LaTeX Companion*, Addison-Wesley, Reading, Massachusetts.

Greenwade, G. D. (1993). The Comprehensive Tex Archive Network (CTAN), *TUGBoat* **14**(3): 342–351.

Todas las imágenes deben de estar referenciadas desde el texto. Para ello puede usarse el comando `\ref{etiqueta}`. Este es un ejemplo que hace referencia a la figura `??`. Esto también se aplica a las tablas `(??)` y los fragmentos de código `(??)`. Para citar trabajos conviene usar el comando `\cite{etiqueta}`. Puede obtener más información en [Greenwade \(1993\)](#) [Goossens et al. \(1993\)](#). La referencias hay que declararlas previamente en el archivo `bibliography.bib`.