EDITORIAL

La Importancia de las Bases de Datos para el Entrenamiento en Inteligencia Artificial

The Importance of Databases for Training in Artificial Intelligence

Jarvis Raraz-Vidal[1,a]

[1]Universidad Nacional Hermilio Valdizan. Huánuco, Perú.
[a]Patología Clínica, Maestro en Investigación Clínica, Diplomado en Inteligencia Artificial. Editor Adjunto.

La inteligencia artificial (IA) ha experimentado un crecimiento exponencial en las últimas décadas, revolucionando industrias, investigaciones en salud[(1)] y nuestra vida cotidiana. Desde asistentes virtuales hasta vehículos autónomos. Sin embargo, en el corazón de este avance tecnológico se encuentra un elemento esencial, pero a menudo pasado por alto: las bases de datos. Las bases de datos son el pilar fundamental en el entrenamiento y desarrollo de modelos de inteligencia artificial. Son los depósitos de información que alimentan a los algoritmos y les permiten aprender, adaptarse y tomar decisiones inteligentes. En esencia, las bases de datos son la materia prima de la IA, y su calidad y diversidad son cruciales para determinar el éxito o el fracaso de los sistemas de IA. Uno de los mayores desafíos que enfrenta la comunidad de IA es la obtención de datos de alta calidad y representativos. Estos datos deben ser precisos, imparciales y relevantes para la tarea que se desea abordar. Además, es esencial que los datos reflejen la diversidad del mundo real, evitando sesgos y discriminación. Un conjunto de datos sesgado puede llevar a resultados igualmente sesgados y decisiones injustas en aplicaciones de IA, lo que socava la confianza en estas tecnologías[(2,3)].

La recopilación y gestión de datos éticos es un tema crítico que debe abordarse de manera responsable. Esto incluye el respeto a la privacidad de las personas, la transparencia en el uso de datos y la protección contra posibles abusos. Es imperativo que la comunidad de la IA trabaje en estrecha colaboración con expertos en ética y privacidad para garantizar que los datos se utilicen de manera justa y beneficiosa para la sociedad[(4,5)].

Otro desafío importante es la disponibilidad de bases de datos adecuadas y suficientes. A medida que las aplicaciones de IA se vuelven más diversas y avanzadas, la demanda de datos crece exponencialmente. Para abordar este problema, es fundamental fomentar la colaboración y la compartición de datos entre investigadores, empresas y gobiernos. Esto no solo acelerará el progreso en la IA, sino que también permitirá un desarrollo más equitativo y accesible de estas tecnologías[(3,6)].

Plataformas de bases de datos

Tanto Kaggle (https://www.kaggle.com/) como GitHub (https://github.com/gcoronelc/databases) son plataformas ampliamente utilizadas en la comunidad de Machine Learning para acceder a conjuntos de datos y recursos relacionados con la ciencia de datos y el aprendizaje automático. Estas plataformas pueden contener datos de uso libre de diferentes dominios: salud (bases de datos de fotografías de bacterias, radiografías, ecografías, resonancia magnética de pacientes con Alzheimer, datos de Covid-19, tuberculosis, resultados de EKG, datos de secuencia genética de un microorganismo, resultados de laboratorio de pacientes con diabetes, etc), finanzas, visión por computadora, procesamiento de lenguaje natural (NLP), ciencia de datos, y más. GitHub también alberga recursos educativos, tutoriales y libros relacionados con Machine Learning que pueden ayudarte a aprender y entrenar modelos de manera efectiva. Ten en cuenta que debes respetar las licencias de los datos y proyectos que encuentres en estas plataformas, y siempre dar crédito adecuado a los autores cuando sea necesario.

Entrenamiento con bases de datos en línea

Entrenar modelos de Machine Learning con datos de Kaggle y GitHub implica la obtención de datos de entrenamiento, preprocesamiento y entrenamiento de modelos utilizando

Citar como: Raraz-Vidal J. La importancia de las bases de datos para el entrenamiento en inteligencia artificial. Rev. Peru. Investig. Salud. [Internet]; 2023; 7(3): 121-122. https://doi.org/10.35839/repis.7.3.1970

Correspondencia a: Jarvis Raraz Vidal; Correo: jarvisraraz@gmail.com

Orcid: Raraz-Vidal J.: https://orcid.org/0000-0002-1511-5877

Conflicto de interés: Ninguno.

Financiamiento: Ninguno.

Editor: Jarvis Raraz, UNHEVAL

Recibido: 31 de julio de 2023
Aprobado: 20 de setiembre de 2023
En línea: 30 de setiembre de 2023

bibliotecas de Machine Learning en Python[7], como Scikit-Learn o TensorFlow. Aquí tienes una descripción general de cómo hacerlo:

Entrenamiento de Modelos con Datos de Kaggle[8]:

1. Acceder a Kaggle: Ve al sitio web de Kaggle (https://www.kaggle.com/) y regístrate si aún no lo has hecho. Kaggle es una plataforma que ofrece conjuntos de datos, competiciones de Machine Learning y recursos para científicos de datos.
2. Buscar Datos: Utiliza la función de búsqueda de Kaggle para encontrar conjuntos de datos relevantes para tu tarea de Machine Learning. Puedes buscar por tema, conjunto de datos específico o palabras clave.
3. Descargar Datos: Una vez que encuentres un conjunto de datos adecuado, descárgalo en tu máquina local. Los conjuntos de datos en Kaggle generalmente se proporcionan en formatos como CSV o JSON.
4. Preprocesamiento de Datos: Limpia y preprocesa los datos según sea necesario. Esto puede incluir la eliminación de valores atípicos, el llenado de valores faltantes y la codificación de variables categóricas.
5. Entrenamiento del Modelo: Utiliza bibliotecas de Machine Learning como Scikit-Learn o TensorFlow para cargar los datos, dividirlos en conjuntos de entrenamiento y prueba, y entrenar tu modelo de Machine Learning.
6. Evaluación del Modelo: Evalúa el rendimiento de tu modelo utilizando métricas adecuadas para tu tarea, como precisión, F1-score o error cuadrático medio, dependiendo de si estás trabajando en una tarea de clasificación o regresión.

Entrenamiento de Modelos con Datos de GitHub[9]:

1. Acceso a Datos en GitHub: GitHub(https://github.com/gcoronelc/databases) es una plataforma de desarrollo colaborativo que alberga una gran cantidad de repositorios de código y datos. Puedes buscar conjuntos de datos específicos en GitHub o acceder a repositorios públicos que contienen datos.
2. Clonar Repositorio: Si encuentras un repositorio en GitHub que contiene datos de tu interés, puedes clonarlo en tu máquina local utilizando Git o simplemente descargar los archivos de datos que necesitas.
3. Preprocesamiento de Datos: Al igual que con los datos de Kaggle, es importante realizar el preprocesamiento necesario en los datos descargados desde GitHub antes de usarlos para el entrenamiento.
4. Entrenamiento del Modelo: Utiliza las mismas bibliotecas de Machine Learning, como Scikit-Learn o TensorFlow, para cargar, preprocesar y entrenar tu modelo utilizando los datos de GitHub.
5. Evaluación del Modelo: Al igual que en el caso de Kaggle, evalúa el rendimiento de tu modelo utilizando métricas adecuadas para tu tarea.

Ten en cuenta que la disponibilidad y calidad de los datos en Kaggle y GitHub pueden variar, por lo que es importante realizar una selección cuidadosa. También es esencial seguir buenas prácticas de ciencia de datos[5].

Las bases de datos son el cimiento de la inteligencia artificial. Su calidad, diversidad y ética son factores cruciales para el éxito continuo de la IA. Como comunidad, debemos reconocer la importancia de los datos y trabajar juntos para garantizar que sean un recurso accesible y beneficioso para todos.

Criterios del autor

El autor declara que participo en todo el proceso desde concepción del artículo hasta la versión final.

Referencias bibliográficas

1. Raraz-Vidal J, Raraz-Vidal O. Aplicaciones de la inteligencia artificial en la medicina. Rev Peru Investig En Salud.2022;6(3):131-3. doi: 10.35839/repis.6.3.1559
2. Liu R, Wang T, Yang Y, Yu B. Database Development Based on Deep Learning and Cloud Computing. Mob Inf Syst.2022;2022:e6208678. doi: 10.1155/2022/6208678
3. Zou B, You J, Wang Q, Wen X, Jia L. Survey on Learnable Databases: A Machine Learning Perspective. Big Data Res. 2022;27:100304. doi: 10.1016/j.bdr.2021.100304
4. Lo Piano S. Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. Humanit Soc Sci Commun. 2020;7(1):1-7. doi: 10.1057/s41599-020-0501-9
5. Toms A, Whitworth S. Ethical considerations in the use of Machine Learning for research and statistics. Int J Popul Data Sci. 7(3):1921. doi: 10.23889/ijpds.v7i3.1921
6. Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E. Deep learning applications and challenges in big data analytics. J Big Data. 2015;2(1):1. doi: 10.1186/s40537-014-0007-7
7. Raraz-Vidal J, Raraz-Vidal O. Empezando a programar en inteligencia artificial. Rev Peru Investig En Salud. 2023;7(2):61-3. doi: 10.35839/repis.7.2.1873
8. Kaggle: Your Machine Learning and Data Science Community [Internet]. [citado 22 de septiembre de 2023]. Disponible en: https://www.kaggle.com/
9. GitHub. Data Science Community [Internet]. 2023 [citado 22 de septiembre de 2023]. Disponible en: https://github.com/gcoronelc/databases