Clinical data: deep learning approaches

Alejandro Silva Rodríguez

Marta Cuevas Rodríguez

Almacenes De Datos
Universidad de Málaga

Septiembre 2024

Índice

1. Introducción

El auge de los datos digitales en la atención médica ha conducido a un aumento en la investigación médica impulsada por datos, basada en el aprendizaje automático. En los últimos años, el aprendizaje profundo, como una técnica poderosa para el análisis de grandes datos, ha adquirido una posición central en el ámbito del aprendizaje automático debido a sus grandes ventajas en la representación de características y reconocimiento de patrones.

Los tipos de datos utilizados en este campo son diversos e incluyen datos estructurados, como registros electrónicos de salud y datos demográficos, así como datos no estructurados, como notas clínicas, imágenes médicas y secuencias genéticas. Estos datos son analizados mediante modelos de aprendizaje profundo, como redes neuronales convolucionales (CNN) para imágenes, redes neuronales recurrentes (RNN) para datos secuenciales, y modelos de atención que permiten una mejor interpretación de las relaciones entre diferentes tipos de datos. La capacidad de estos modelos para aprender de grandes volúmenes de datos ha permitido avances significativos en la predicción de enfermedades, personalización de tratamientos y optimización de la atención al paciente[20] .

2. Definición y Relevancia

2.1. ¿Qué es el Deep Learning?

Deep learning es una rama del aprendizaje automático que emplea arquitecturas de redes neuronales profundas para modelar y aprender representaciones complejas de datos. Estas arquitecturas están compuestas por múltiples capas de procesamiento, lo que permite que el sistema capture características jerárquicas y abstractas de los datos de entrada. [9]

Los avances en deep learning han impulsado mejoras significativas en diversas áreas, como el reconocimiento de voz, la detección de objetos y la genómica, contribuyendo a la obtención de resultados de vanguardia en estas disciplinas. Utilizando el algoritmo de retropropagación, deep learning optimiza los parámetros internos del modelo, facilitando la transformación de la representación de los datos a través de las distintas capas de la red.

En particular, las redes neuronales convolucionales (CNN) han demostrado un rendimiento excepcional en tareas relacionadas con imágenes y videos, mientras que las redes neuronales recurrentes (RNN) son adecuadas para el procesamiento de datos secuenciales, como el texto y el habla. Estos enfoques han llevado a la creación de sistemas más robustos y precisos en el análisis y la interpretación de grandes volúmenes de datos. [14]

2.2. Relación con los Almacenes de Datos

El uso de un data warehouse es fundamental en el análisis de datos clínicos, ya que permite la integración y consolidación de grandes volúmenes de información provenientes de diversas fuentes, como historiales médicos, registros de laboratorio y datos de ensayos clínicos. Esta centralización de datos facilita el acceso a información coherente y de alta calidad, lo que es esencial para llevar a cabo análisis estadísticos y la implementación de técnicas de aprendizaje automático. Al proporcionar una estructura organizada y optimizada para la consulta de datos, un data warehouse no solo mejora la eficiencia de los procesos analíticos, sino que también permite a los investigadores y profesionales de la salud identificar patrones, tendencias y correlaciones en los datos clínicos que pueden ser cruciales para la toma de decisiones informadas en el ámbito de la medicina. En este contexto, el data warehouse se convierte en una herramienta indispensable para la mejora de la atención sanitaria y el avance de la investigación clínica.[5]

3. Estado del Arte en Deep Learning para Datos Clínicos

En los últimos años, el deep learning ha experimentado un crecimiento significativo en el ámbito de la salud, especialmente en el análisis de datos clínicos. Las capacidades de los modelos de deep learning han mejorado sustancialmente con la disponibilidad de datos clínicos en gran escala y el desarrollo de infraestructuras avanzadas para el procesamiento de estos datos. En esta sección, se revisarán las tecnologías, frameworks y metodologías actuales, destacando los modelos más relevantes en el análisis de datos clínicos, como las redes neuronales convolucionales (CNNs), redes neuronales recurrentes (RNNs), y transformers. Además, se analizarán los avances recientes en la integración de almacenes de datos clínicos que facilitan el uso de deep learning en el ámbito médico.

3.1. Tecnologías y Frameworks Actuales

Las tecnologías de deep learning han sido impulsadas por el desarrollo de frameworks robustos, como *TensorFlow* y *PyTorch*, que permiten la creación, entrenamiento y despliegue de modelos de redes neuronales de manera eficiente. Estos frameworks han sido fundamentales para el análisis de grandes volúmenes de datos médicos, posibilitando el entrenamiento de modelos complejos en GPU y facilitando su implementación en plataformas clínicas. La integración de estos frameworks en entornos de investigación clínica ha permitido acelerar el desarrollo de modelos predictivos y de diagnóstico automatizado en tiempo real [1, 13].

3.2. Modelos Clave en Deep Learning para Datos Clínicos

El análisis de datos clínicos se ha beneficiado en gran medida de la implementación de varios tipos de modelos de deep learning que aprovechan la estructura única de estos datos. A continuación, se describen algunos de los modelos más utilizados y sus aplicaciones en el ámbito clínico.

3.2.1. Redes Neuronales Convolucionales (CNNs)

Las CNNs son ampliamente utilizadas en el análisis de datos de imágenes médicas, como resonancias magnéticas, tomografías y radiografías. Su capacidad para extraer y procesar características espaciales ha demostrado ser altamente efectiva en tareas de detección de enfermedades, segmentación de órganos y clasificación de imágenes. Un avance notable en este campo es el uso de redes convolucionales profundas, que permiten mejorar la precisión de diagnóstico al identificar patrones complejos en los datos visuales [11].

3.2.2. Redes Neuronales Recurrentes (RNNs)

Las RNNs y sus variantes, como LSTM (Long Short-Term Memory) y GRU (Gated Recurrent Unit), son ampliamente aplicadas en el análisis de datos secuenciales en el ámbito clínico, como el monitoreo de signos vitales o la evolución de los síntomas de un paciente a lo largo del tiempo. Estas redes permiten capturar dependencias temporales en los datos, lo cual es esencial para tareas como la predicción de eventos adversos, progresión de enfermedades y análisis de series temporales de datos clínicos [2].

3.2.3. Transformers

Más recientemente, los modelos basados en transformers han ganado popularidad debido a su eficacia en el análisis de datos secuenciales y no estructurados. Los transformers, como BERT y sus variantes, han demostrado ser altamente efectivos en el procesamiento de texto clínico, como en el análisis de notas médicas y registros de pacientes. Estos modelos son capaces de capturar relaciones contextuales complejas y han sido empleados en tareas de extracción de información clínica, codificación automática y generación de informes médicos [3, 10].

3.3. Avances Recientes en Almacenes de Datos Clínicos para Deep Learning

La consolidación y accesibilidad de grandes volúmenes de datos clínicos en almacenes de datos ha sido esencial para el desarrollo de modelos de deep learning en este campo. Un almacén de datos clínicos integra información de múltiples fuentes, incluyendo registros electrónicos de salud (EHRs), bases de datos de imágenes médicas y datos genómicos, permitiendo que los modelos de deep learning trabajen con una visión integral de la salud del paciente. Los recientes avances en la organización, limpieza y etiquetado de estos datos han mejorado la calidad y confiabilidad de los modelos, además de facilitar el análisis de relaciones entre diferentes tipos de datos clínicos [12].

Además, el uso de almacenes de datos clínicos permite la aplicación de técnicas de aprendizaje transferido, donde los modelos entrenados en un conjunto de datos pueden ajustarse y aplicarse a otro, optimizando así el uso de los datos y reduciendo los tiempos de desarrollo. Esto es particularmente relevante en el caso de enfermedades raras, donde la cantidad de datos disponibles puede ser limitada. A medida que los sistemas de almacenes de datos clínicos se desarrollan, se espera que continúen facilitando el avance del deep learning en la medicina [17].

3.4. Síntesis del Estado del Arte

El estado del arte en deep learning aplicado a datos clínicos refleja un progreso continuo en el desarrollo de modelos robustos y precisos. La adopción de tecnologías avanzadas y frameworks especializados ha permitido mejorar el diagnóstico y la predicción de enfermedades mediante el análisis de datos clínicos complejos. La combinación de CNNs, RNNs y transformers, junto con la integración de almacenes de datos clínicos, promete impulsar aún más las capacidades del deep learning en el ámbito de la salud.

4. Metodología y Ejemplos de Uso

4.1. Aplicación del Deep Learning a los Datos Clínicos

La aplicación de técnicas de deep learning en datos clínicos se ha convertido en un área de gran interés debido a su potencial para transformar la atención médica. En general, el proceso puede describirse en varias etapas fundamentales:

1. **Recolección de Datos:** Se requiere una amplia variedad de datos clínicos, que pueden incluir imágenes médicas, registros de pacientes y resultados de pruebas diagnósticas. La calidad y la cantidad de estos datos son cruciales para el éxito del modelo [11].

2. **Preprocesamiento de Datos:** Antes de entrenar un modelo, los datos deben ser limpiados y preparados. Esto incluye la normalización de los datos, la eliminación de valores atípicos y el tratamiento de datos faltantes. Para las imágenes, se aplican técnicas de aumento para diversificar el conjunto de datos y mejorar la robustez del modelo [12].

3. **Entrenamiento del Modelo:** Se seleccionan arquitecturas de red neuronal adecuadas y se entrena el modelo utilizando los datos preprocesados. Durante esta fase, se ajustan los hiperparámetros y se utilizan técnicas como la regularización para evitar el sobreajuste [5].

4. **Evaluación del Modelo:** Después del entrenamiento, se evalúa el rendimiento del modelo utilizando métricas específicas, como precisión, recall y F1-score. Esta etapa es esencial para determinar la efectividad del modelo en la práctica clínica [4].

5. **Implementación y Monitoreo:** Una vez que el modelo ha demostrado ser efectivo, se implementa en entornos clínicos reales, donde se monitorea continuamente su desempeño y se realizan ajustes según sea necesario [17].

4.2. Estudio de Caso: Modelo Predictivo para Diagnóstico de Enfermedades

En este estudio de caso, se implementa un modelo de deep learning para el diagnóstico de enfermedades dermatológicas, centrado en la clasificación de imágenes de lesiones cutáneas. A continuación, se detallan las fases del proceso:

1. **Preparación de Datos:** El primer paso consiste en descargar el conjunto de datos desde la API de Kaggle [7], que contiene imágenes y metadatos relevantes sobre diversas condiciones dermatológicas, concretamente el conjunto de datos Human Against Machine with 10000 (HAM10000) [19] que es una colección de imágenes dermatoscópicas que se utiliza para la clasificación de diferentes tipos de cáncer de piel.

   Se realiza un análisis inicial de los datos para identificar el desequilibrio en las clases. En este caso, se observa que la clase de lunares ('nv') tiene una cantidad significativamente mayor de imágenes en comparación con otras clases. Para abordar este problema, se decide reducir el número de imágenes de lunares a un 40 % de su cantidad original, asegurando así un mejor equilibrio en el conjunto de datos .

2. **División de Datos:** Una vez que los datos se han equilibrado, se procede a dividir el conjunto en datos de entrenamiento y de validación. Se utiliza una proporción del 80 % de los datos para el entrenamiento y el 20 % restante para la validación. Esto garantiza que el modelo sea capaz de generalizar a nuevos datos y no se limite a memorizar el conjunto de entrenamiento.

3. **Generación de Datos:** Para mejorar la capacidad del modelo para generalizar, se implementan generadores de datos que aplican técnicas de aumento de imágenes, como rotaciones, escalados y cambios de brillo. Esto no solo aumenta la cantidad de datos disponibles para el entrenamiento, sino que también ayuda a simular variaciones que el modelo puede encontrar en situaciones del mundo real.

4. **Entrenamiento del Modelo:** Se utilizó MobileNetV2 [16], un modelo preentrenado en la base de datos ImageNet, lo que permitió aprovechar características previamente aprendidas para mejorar la precisión en la clasificación de imágenes de piel.

   El entrenamiento se llevó a cabo con TensorFlow [1], donde se implementaron técnicas de Dropout y Early Stopping para mitigar el riesgo de sobreajuste. El Dropout se utilizó para desactivar aleatoriamente un porcentaje de neuronas durante el entrenamiento, promoviendo la generalización del modelo. Por otro lado, el Early Stopping supervisó la precisión en el conjunto de validación y detuvo el entrenamiento automáticamente cuando no se observaron mejoras durante un número definido de épocas. Además, se congelaron las capas iniciales de MobileNetV2, permitiendo que las capas superiores se ajustaran finamente a los datos específicos del problema. Esta estrategia resultó en un modelo eficiente y robusto para el diagnóstico de enfermedades cutáneas.

5. **Evaluación y Ajuste del Modelo:** Tras el entrenamiento, se evalúa el modelo en el conjunto de validación utilizando métricas como precisión y pérdida. Además, se generan matrices de confusión para analizar el rendimiento en cada clase. Se ajustan los umbrales de decisión, particularmente para la clase de lunares, para minimizar los falsos positivos, asegurando así que el modelo sea más confiable en la identificación de esta condición crítica [4].

Este enfoque metodológico proporciona un marco robusto para aplicar deep learning en el diagnóstico de enfermedades dermatológicas, facilitando una detección más precisa y oportuna que puede mejorar los resultados de salud de los pacientes [12].

5. Desafíos y Direcciones Futuras

A pesar de los avances significativos en el uso de deep learning para datos clínicos, todavía existen desafíos clave que limitan la implementación generalizada de estas tecnologías en entornos de salud. Uno de los problemas más importantes es la calidad y disponibilidad de los datos clínicos. La variabilidad en

los sistemas de registros electrónicos de salud y la falta de interoperabilidad entre diferentes plataformas dificultan la integración de datos de múltiples fuentes, lo cual es crucial para entrenar modelos robustos y generalizables [8].

Otro desafío se encuentra en la interpretabilidad de los modelos de deep learning, especialmente en el ámbito clínico, donde las decisiones automatizadas deben ser justificables y comprensibles para los profesionales de la salud. El desarrollo de modelos explicables y transparentes representa una gran oportunidad para que el deep learning sea adoptado con confianza en el sector médico [6].

Asimismo, el área de privacidad y seguridad de los datos sigue siendo una preocupación crítica, particularmente debido a la naturaleza sensible de los datos de salud. La investigación en técnicas como el aprendizaje federado y la privacidad diferencial ofrece un camino prometedor para abordar estas cuestiones y facilitar el entrenamiento de modelos sin comprometer la privacidad de los pacientes [15].

En el futuro, es probable que el desarrollo de herramientas que combinen múltiples tipos de datos clínicos y la aplicación de enfoques multimodales amplíen las capacidades del deep learning en el campo de la salud. Estos enfoques podrían mejorar la precisión y el alcance de los modelos, ofreciendo aplicaciones que vayan desde la detección temprana de enfermedades hasta la personalización de tratamientos.

6. Conclusión

Resumen de los principales hallazgos del informe, enfatizando la importancia del deep learning en los datos clínicos y el papel de apoyo de las estructuras de almacenes de datos en la habilitación de estos modelos.

7. Acceso al Repositorio

Toda la información adicional, incluyendo el código fuente y la documentación completa de este proyecto, está disponible en el repositorio de GitHub [18].

Referencias

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, pages 265–283, 2016.

[2] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, pages 301–318. PMLR, 2016.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] A. Esteva, B. Kuprel, R. A. Novoa, and et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[5] Alaa Hamoud, Ali Salah Hashim, and Wid Akeel Awadh. Clinical data warehouse: a review. *Iraqi Journal for Computers and Informatics*, 44(2), 2018.

[6] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*, 2017.

[7] Kaggle Inc. Kaggle api. `https://www.kaggle.com/docs/api`, 2024. Último acceso: 1 octubre 2024.

[8] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):1–9, 2016.

[9] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[10] Itian Li, Sebastian Raschka, Jimeng Sun, Lane R Waitman, and Alex E W Johnson. Behrt: Transformer for electronic health records. *Scientific Reports*, 10(1):7155, 2020.

[11] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen AWM van der Laak, Bram van Ginneken, and Clara I Sanchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[12] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6):1236–1246, 2018.

[13] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[14] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and S. S. Iyengar. A survey on deep learning: Algorithms, techniques, and applications. *ACM Comput. Surv.*, 51(5), September 2018.

[15] Nicola Rieke, Jonathon Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu Nicolas Galtier, Bennett Landman, Klaus H Maier-Hein, et al. The future of digital health with federated learning. *npj Digital Medicine*, 3(1):1–7, 2020.

[16] M. Sandler, A. Howard, M. Zhu, and et al. Mobilenetv2: Inverted residuals and linear bottlenecks. *arXiv preprint arXiv:1801.04381*, 2018.

[17] Benjamin Shickel, Patrick J Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1589–1604, 2017.

[18] Alex Silva. Deep_learning_clinical_data. `https://github.com/AlexSilvaa9/Deep\_learning\_clinical\_data`, 2024. Último acceso: 1 octubre 2024.

[19] M. Taha and et al. Ham10000: The first large, public, dermatoscopic dataset. *arXiv preprint arXiv:1806.00388*, 2018.

[20] Ying Yu, Min Li, Liangliang Liu, Yaohang Li, and Jianxin Wang. Clinical big data and deep learning: Applications, challenges, and future outlooks. *Big Data Mining and Analytics*, 2(4):288–305, 2019.