Aprendizaje de representaciones para una IA confiable y sostenible

David de la Rosa[1], Francisco Charte[2], and María José del Jesus[2]

[1] Estudiante de doctorado. Departamento de Informática, Universidad de Jaén, Jaén 23071, España
drrosa@ujaen.es
[2] Director de tesis. Departamento de Informática, Universidad de Jaén, Jaén 23071, España
{fcharte, mjjesus}@ujaen.es

Resumen. Este proyecto de tesis doctoral, iniciada el 18 de noviembre de 2024 y con fecha de defensa prevista para noviembre de 2028, aborda los desafíos críticos de sostenibilidad, privacidad, seguridad y reutilización que plantea la expansión masiva de los modelos de Inteligencia Artificial (IA) contemporáneos. A pesar de los avances en el aprendizaje de representaciones, las soluciones actuales dependen de arquitecturas masivas con altos costes energéticos, en general su explicabilidad es baja y muestran un progreso desigual entre modalidades de datos, relegando a los formatos tabulares o de audio a un segundo plano frente a visión por computador. Para superar estas barreras, la hipótesis es que es posible aprender representaciones alternativas de los datos en bruto que habiliten el procesamiento mediante modelos significativamente más sencillos que promuevan la sostenibilidad, protejan la privacidad al volver ilegible la información latente y faciliten la explicabilidad a través de espacios desentrelazados. Estando actualmente el proyecto de tesis en el tercer trimestre del segundo año de desarrollo, se ha completado el primer hito de revisión teórica mediante la propuesta de teorías unificadas. Así mismo, se encuentra en ejecución la transferencia práctica a entornos reales a través de soluciones aplicadas con éxito a la ciberseguridad y a la conservación ecológica con publicaciones científicas en congresos y revistas de relevancia.

1 Resumen

1.1 Introducción y contexto

La expansión del Aprendizaje Automático (*Machine Learning*, ML) plantea desafíos críticos. En entornos interconectados, garantizar la seguridad en el acceso a los datos es complejo [11], mientras que su almacenamiento externo compromete la privacidad [13]. Asimismo, el alto consumo energético, material e hídrico de los grandes modelos de visión y lenguaje exige definir procesos más sostenibles y accesibles [5, 22]. Por último, emulando la reutilización de software [12], resulta clave desarrollar mecanismos para reutilizar datos y modelos preentrenados.

El aprendizaje de representaciones extrae automáticamente características óptimas de datos en bruto [2]. Al eliminar ruido y redundancias, reduce la dimensionalidad, permitiendo usar modelos más pequeños, eficientes y sostenibles [15]. Además, el enfoque auto-supervisado mitiga la escasez de datos etiquetados, mientras que la proyección de la información en un espacio abstracto e ilegible para humanos favorece la privacidad.

En los últimos años han surgido numerosas propuestas basadas en *Self-Supervised Learning* (SSL) [6, 21] para el aprendizaje de representaciones. Entre ellas destacan las siguientes:

- **Aprendizaje contrastivo** [8, 10]. Esta familia de métodos crea representaciones de datos en los que instancias similares estén situadas *cerca* en el espacio latente, a la vez que se fuerza a que instancias diferentes se sitúen *lejos* entre ellas. Al proponer este objetivo durante el entrenamiento, se fuerza a que el modelos sea capaz de aprender a estructurar el espacio latente de forma organizada y estructurada.
- **Predicción de transformaciones** [14]. Consiste en entrenar un modelo inteligente capaz de predecir qué modificación se ha aplicado sobre los datos. Este objetivo de entrenamiento favorece la creación de representaciones robustas e invariantes a modificaciones sobre los datos originales.
- **Aprendizaje de representaciones basadas en reconstrucción** [4, 9]. El modelo se entrena para generar o reconstruir de forma explícita las partes ocultas de una entrada que ha sido parcialmente enmascarada. Para lograrlo, la arquitectura de la red obliga a comprimir la información en una representación latente que debe retener los detalles suficientes como para recrear fielmente los datos originales en su espacio de entrada, por ejemplo, píxeles o texto.
- **Aprendizaje de representaciones predictivo** [1]. A diferencia de los métodos de reconstrucción, este enfoque plantea un objetivo donde el modelo debe predecir únicamente la representación latente de la parte enmascarada, no sus datos crudos. Al operar exclusivamente en el espacio de características abstractas, el modelo evita gastar capacidad en predecir detalles irrelevantes o ruido de alta frecuencia, centrándose en la estructura semántica.
- **Aprendizaje de representaciones con modelos generativos** [7]. Esta familia de métodos, gracias al entrenamiento con un objetivo generativo para modelar la distribución completa de los datos, aprenden en el camino representaciones latente semánticas.

1.2 Hipótesis y objetivos

A pesar de los avances en el aprendizaje de representaciones, persisten tres limitaciones críticas: la falta de explicabilidad (las representaciones no son comprensibles para el ser humano), la insostenibilidad (los extractores de características consumen tantos recursos computacionales como los grandes modelos supervisados) y la brecha multimodal (datos como tablas o audio no han avanzado al ritmo de la visión por computador).

La hipótesis de este proyecto es que es posible aprender representaciones alternativas continuas, completas y compactas que simplifiquen los datos de entrada sin perder información relevante. Estas representaciones no solo reducirán la dimensionalidad, sino que optimizarán la sostenibilidad (al permitir el uso de modelos finales más sencillos), la privacidad (al hacer los datos ilegibles para humanos) y la explicabilidad mediante representaciones desentrelazadas [24].

Siguiendo esta hipótesis, la finalidad de la tesis doctoral es contribuir al avance en el campo del aprendizaje de representaciones, proponiendo algoritmos que aborden las necesidades destacadas a lo largo de esta sección. Entre ellas destaca el crear representaciones simplificadas de los datos en bruto para favorecer el entrenamiento de modelos ML y mejorar la calidad de los datos en múltiples modalidades, mejorar la privacidad, seguridad y confiabilidad de los modelos y datos con los que se entrenan, mejorar la sostenibilidad de los modelos tanto durante el entrenamiento como en su uso posterior, y desarrollar modelos explicables cuya salida pueda ser entendida por un humano.

Para alcanzar este objetivo, se plantean los siguientes objetivos específicos:

- **Objetivo 1 (O1 - seis trimestres).** Revisión del estado del arte. Investigación sobre las arquitecturas y métodos usados en el campo del aprendizaje de representaciones. Se analizarán tanto propuestas no supervisadas o autosupervisadas, que no requieren el uso de etiquetas para obtener la representación de los datos, como métodos supervisados o semi-supervisados, que sí las necesitan.
- **Objetivo 2 (O2 - seis trimestres)**. Adaptación de técnicas de aprendizaje de representaciones a la resolución de problemas de diferente tipología y modalidad de datos, como imágenes, texto, audio o formato tabular. Permitirá la evaluación de métodos de aprendizaje de representaciones en base a la utilidad de la representación en tareas posteriores.
- **Objetivo 3 (O3 - cuatro trimestres)**. Evaluación de métodos de aprendizaje de representaciones en el campo de la explicabilidad, sostenibilidad y confiabilidad. Se analizarán los puntos débiles de los diferentes métodos y facilitará la detección de posibles campos a mejorar.
- **Objetivo 4 (O4 - seis trimestres)**. Diseño de técnicas de aprendizaje de representaciones. El último objetivo que se plantea para el proyecto es el desarrollo de nuevas propuestas avanzadas para el aprendizaje de representaciones. Estas podrán ser propuestas de mejora sobre técnicas específicas o hibridaciones que combinen las ventajas de cada una. El resultado será un avance en el estado del arte que se traduzca en un incremento de la calidad de la representación o en mejoras de la explicabilidad, sostenibilidad o privacidad de los modelos.

1.3 Originalidad de la tesis

La aportación principal de este proyecto de tesis doctoral radica en avanzar el estado del arte en el campo del aprendizaje de representaciones, de forma que los métodos propuestos no solo incrementen la precisión, sino que también favorezca

una mejora en campos como la sostenibilidad, privacidad o explicabilidad. A pesar de los grandes avances en los últimos años, la mejora en el rendimiento ha estado directamente ligada a modelos cada vez más grandes que son capaces de procesar ingentes cantidades de datos no etiquetados. Sin embargo esta mejora de rendimiento no ha estado acompañada de una mejora en la explicabilidad (modelos más grandes suelen suponer un empeoramiento de la explicabilidad), ni la sostenibilidad de los modelos (modelos más grandes requieren mayor capacidad de cómputo). Además, los avances más significativos se han producido en el campo de la visión por computador y procesamiento del lenguaje natural, dejando a un lado, por ejemplo, la modalidad de datos tabulares.

Con esta tesis doctoral se pretende abarcar dichas necesidades. Aplicaciones de modelos para aprendizaje de representaciones en datos tabulares, creación de paquetes que abarquen diferentes modalidades o estudios sobre la explicabilidad y sostenibilidad de dichos métodos permitirán centrarnos no solo en una mejora del rendimiento de los modelos, sino también en avances en otros campos menos cubiertos por la comunidad científica.

2 Metodología y plan de trabajo

En esta sección se presenta la metodología y el plan de trabajo a seguir durante el desarrollo del proyecto de tesis.

2.1 Metodología

La metodología seguida para el análisis de fuentes documentales se realizará mediante una búsqueda exhaustiva de literatura científica y artículos académicos sobre métodos empleados en la actualidad para el aprendizaje de representaciones. Para ello, se harán búsquedas en bases de datos bien conocidas, como Google Scholar, Semantic Scholar, Web of Science o Scopus, entre otras. Además, se revisarán conferencias anuales relevantes sobre el ámbito de estudio, como *International Conference on Learning Representations* (ICLR) o *International Conference on Machine Learning* (ICML). Como toda investigación científica, en esta tesis se seguirá el método científico habitual:

1. **Observación**. Análisis de la literatura científica actual a partir de la cual surgen problemas o inconvenientes que abordar.
2. **Formulación de la hipótesis de partida,** que involucre el desarrollo de nuevos métodos.
3. **Experimentación.** Prueba y cuestionamiento de la hipótesis para evaluar las propuestas en comparación con los métodos actuales.
4. **Reformulación de la hipótesis**, si procede, modificando si es necesario los algoritmos en función de las pruebas completadas.

El proyecto de tesis doctoral se llevará a cabo dentro del grupo de investigación TIC-207 Sistemas Inteligentes y Minería de Datos (SIMIDAT). Este

grupo cuenta con una larga experiencia en el campo de la investigación, avalada con multitud de publicaciones en revistas de alto impacto. Además, cuenta con prestaciones y equipos de servidores que permitirán llevar a cabo las tareas que se plantean en el plan de trabajo.

2.2 Plan de trabajo

El plan de trabajo que se seguirá está organizado en tres niveles. El primero son los hitos, que marcan un momento destacable del periodo de tesis al haber alcanzado uno o varios objetivos concretos. El segundo nivel son los objetivos. El tercer nivel son las tareas que se necesitan completar para alcanzar los objetivos propuestos. En la figura 1 se presenta un diagrama de Gannt que recoge una temporización aproximada del proyecto de tesis. A lo largo de la sección se analizará el diagrama, indicando los objetivos y tareas en base a cada hito.

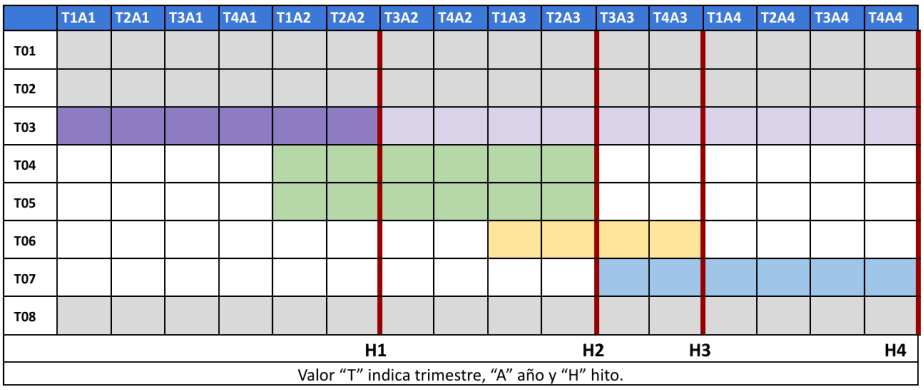


Fig. 1. Plan de trabajo propuesto para la tesis doctoral. El eje X es el eje temporal. Comienza en noviembre de 2024 (T1A1) y finaliza en noviembre de 2028 (T4A4). El eje Y representa las tareas en las que se divide el proyecto. Para cada fila, el color indica la duración y planificación temporal de cada tarea. Cada color representa un objetivo diferente. Se marcan también los hitos de la tesis como lineas verticales de color rojo.

Hito 1. Revisión bibliográfica. Este hito coincide con el objetivo O1 *Revisión del estado del arte*. Al alcanzar este hito, durante el segundo trimestre del segundo año, el doctorando habrá demostrado tener unos conocimientos amplios y extensos sobre el aprendizaje de representaciones. Conoce un amplio abanico de técnicas para aprender representaciones y las ventajas e inconvenientes asociadas a cada una de ellas.

Este hito está también asociado con la tarea T03 *Revisión bibliográfica del estado del arte*. Dicha tarea finaliza con la redacción de un informe técnico en el

que se estudien las técnicas exploradas para el aprendizaje de representaciones de datos en bruto y sus aplicaciones.

Es destacable que, si bien el hito se alcanza durante el segundo trimestre del segundo año, la literatura científica se actualiza constantemente con nuevas propuestas y aplicaciones. Por tanto, la tarea de revisión del estado del arte se mantendrá activa durante todo el proyecto.

Hito 2. Creación de paquetes software para aprendizaje de representaciones. Este hito está directamente asociado con el objetivo O2 *Adaptación de técnicas de aprendizaje de representaciones*. Con este hito, el doctorando demuestra que no solo conoce a nivel teórico las diferentes metodologías para aprendizaje de representaciones estudiadas en el objetivo O1, sino que también es capaz de aplicarlas a problemas reales para mejorar la calidad de los datos.

El objetivo O2 abarca dos tareas. Por un lado, la tarea T04 *Adaptación de métodos de aprendizaje de representaciones* consiste en el análisis, implementación y aplicación de diferentes metodologías para aprendizaje de representaciones en problemas reales. Se completa con la creación de un informe técnico en el que se detalle la implementación y aplicación práctica de uno o varios métodos de aprendizaje de representaciones en un problema concreto. Se analizarán en profundidad los beneficios de utilizar una buena representación de los datos para resolver un problema concreto frente a no utilizar estas técnicas. Por otro lado, la tarea T05 *Creación de paquetes software para aprendizaje de representaciones* busca el desarrollo de uno o más paquetes software que agrupen las principales técnicas de aprendizaje de representaciones. Dicha tarea facilita la transferencia del conocimiento adquirido durante la tesis doctoral a otros investigadores y usuarios. Finaliza con la creación y publicación de un paquete software bajo licencia abierta para su uso por parte de la comunidad científica, así como la publicación de la documentación asociada.

Hito 3. Análisis de la explicabilidad, sostenibilidad y confiabilidad de modelos para el aprendizaje de representaciones. Este hito demuestra que el doctorando ha obtenido unos conocimientos amplios sobre cómo el aprendizaje de representaciones puede potenciar la explicabilidad, sostenibilidad, confiabilidad y privacidad de modelos inteligentes. Comprende también el estudio de técnicas con ese objetivo, como son *Mixture of Experts, Model Distillation* o *Disentangled Representation Learning*, entre otras.

El tercer hito corresponde al objetivo O3 *Evaluación de métodos de aprendizaje de representaciones en el campo de la explicabilidad, sostenibilidad y confiabilidad*, asociado de igual manera con la tarea T06, con el mismo nombre. Consiste en el análisis y evaluación de las familias de métodos de aprendizaje de representaciones, como contrastivo, basado en reconstrucción o basado en generación, en términos de explicabilidad (entender qué significa la representación), sostenibilidad (evaluar el coste ambiental y de recursos al entrenar o usar el modelo) y confiabilidad (robustez, seguridad y consistencia de la representación). La tarea finaliza con la redacción y publicación de un informe técnico

Hito 4. Propuesta algorítmica novedosa para el aprendizaje de representaciones. Este hito se corresponde con el planteamiento teórico e implementación de una nueva propuesta para el aprendizaje de representaciones. Estará orientada a la mejora de la calidad de representación y usabilidad para tareas posteriores y/o para potenciar la sostenibilidad, explicabilidad o confiabilidad de los modelos. Está asociado con el objetivo O4 y tarea T07 *Propuestas para el aprendizaje de representaciones*, que consiste en el desarrollo de propuestas para el aprendizaje de representaciones en datos en bruto. Se analizarán y evaluarán mejoras en diversos campos para aprendizaje de representaciones, como *contrastive learning* y métodos similares que no usan ejemplos negativos, *autoencoders* y otros modelos basados en reconstrucción de instancias, enfoques generativos o hibridaciones que combinen los puntos positivos de varios enfoques. Incluye la evaluación de la propuesta en términos de explicabilidad, sostenibilidad y confiabilidad.

Se completa con la creación de un informe técnico en el que se detalle la nueva propuesta, especificando cómo avanza el estado del arte con esta propuesta y destacando los beneficios con respecto a otras propuestas.

Tareas transversales. Además de las tareas descritas en los apartados anteriores, se han planteado otras tres tareas que, si bien no están asociadas con ningún objetivo o hito concreto, son vitales para la realización del proyecto de tesis doctoral.

La primera de ellas es la tarea T01 *Control y seguimiento del proyecto de tesis*. Consiste en la redacción de informes de control y seguimiento, como el plan de investigación o el documento de actividades de la plataforma RAPI, de la Universidad de Jaén.

La segunda es la tarea T02 *Asistencia a cursos y talleres de formación*. Se centra en la formación del doctorando mediante diferentes actividades relacionadas con el doctorado, como el Curso de Competencias Transversales de los Doctorados o cualquier otro curso, ya sea obligatorio o no.

La última de ellas es la tarea T08 *Difusión de resultados*. Se basa en la participación y organización de actividades de divulgación científica relacionadas directamente con el ámbito de trabajo del doctorando. Estas actividades engloban la divulgación del trabajo y resultados tanto en foros científicos como abiertos a la sociedad general: seminarios, cursos, congresos o conferencias impartidas por el doctorando. Culmina con la defensa y publicación de la tesis doctoral.

3 Resultados y relevancia

En esta sección se analizarán dos campos fundamentales dentro del proyecto de tesis doctoral. Por un lado, se analizarán los resultados alcanzados hasta este momento, que abarca el primer año y medio de la tesis doctoral. Seguidamente, se analizará la relevancia de dichos resultados, haciendo hincapié en cómo los proyectos completados enmarcados dentro de la tesis no solo tienen un impacto

científico notable, sino que también tienen grandes posibilidades de aplicación o transferencia a la sociedad.

3.1 Resultados

En el momento de entrega de este documento (junio de 2026), el proyecto de tesis se encuentra en el tercer trimestre del segundo año. Según se analizó en el cronograma 1, se debería haber alcanzado el hito H1 y estar en pleno desarrollo con el objetivo O2. A continuación se enumerarán las publicaciones relacionadas con cada uno de estos dos objetivos, incluyendo una breve descripción de las mismos.

3.2 Artículos relacionados con el hito H1

El primer artículo publicado en relación a este hito es ***A didactic overview of Transformer applications: model variations and user guidelines*** [3], publicado en la revista *Logic Journal of the IGPL*. La motivación de este artículo era que, aunque los *Transformers* [23] se diseñaron originalmente para texto, en los últimos años han surgido nuevas arquitecturas capaces de procesar datos de diferentes modalidades y era difícil detectar y entender las diferencias entre esas nuevas propuestas y la arquitectura original de Transformer. Con este artículo buscamos analizar las principales aplicaciones de los *Transformers* para diferentes tipos de modalidades y proporcionar una guía de uso y desarrollo de los mismos.

El segundo artículo lleva por título ***The Eyes of GPAI: View Alignment Self-Supervised Learning in the Context of General Purpose Artificial Intelligence*** [17], presentado en el congreso *IEEE Conference on Artificial Intelligence*. El objetivo principal del trabajo era hacer una revisión del estado del arte de una familia de técnicas dentro del campo del aprendizaje de representaciones: *View alignment methodologies*. Además, buscamos definir la relación directa que existe entre estos mecanismos y el concepto *General Purpose Artificial Intelligence* (GPAI). Finalmente, analizamos las limitaciones de dichos métodos y planteamos cómo la hibridación entre diferentes paradigmas podría ser el camino hacia GPAI.

El tercer y último artículo es ***Representation Learning: A Unified Framework of Paradigms, Modalities and Contemporary Dimensions*** [16]. Durante el trabajo de revisión se detectó que muchos de los *surveys* existentes que abarcaban la temática del aprendizaje de representaciones se centraban en técnicas o modalidades de datos concretas. Con este artículo buscamos examinar avances recientes en este campo, categorizando las propuestas en una taxonomía que sea agnóstica de la modalidad de datos. Además discutimos diferentes aspectos clave dentro de la tesis, como la sostenibilidad, explicabilidad y seguridad de los modelos.

3.3 Artículos relacionados con el hito H2

El primer artículo enmarcado dentro de este hito se titula ***MORENA: Empty images detection based on unsupervised reconstruction error analysis*** [20]. Con este trabajo buscamos aprovechar el aprendizaje de representaciones innato de los *autoencoders* al procesamiento de imágenes de fototrampeo y detección de imagenes vacías. Al entrenarse de forma completamente no supervisada, no necesitamos imágenes etiquetadas, reduciendo considerablemente el esfuerzo y trabajo previo de etiquetado al entrenamiento del modelo.

El segundo artículo tiene como título ***Weakly supervised discarding of photo-trapping empty images based on autoencoders*** [19]. Este artículo supone una ampliación del trabajo anterior en la que proponemos un método débilmente supervisado para la detección de imágenes vacías. A diferencia del anterior, el uso de una pequeña cantidad de imágenes etiquetadas permite superar el rendimiento de propuestas completamente supervisadas.

Hasta este momento, el último trabajo de aplicación de técnicas de aprendizaje de representaciones ha sido ***A Study on Self-Supervised Learning Objectives for Intrusion Detection*** [18]. En este caso, se aplicaron diferentes técnicas de aprendizaje de representaciones para mejorar el rendimiento de modelos para la detección de intrusiones. Se demostró cómo, en un entorno con muy pocos datos etiquetados, las redes neuronales se benefician en gran medida de un preentrenamiento de aprendizaje contrastivo, mejorando su rendimiento con respecto a un entrenamiento clásico con la totalidad de datos etiquetados.

3.4 Relevancia

Hasta el momento de la redacción de este documento, las aportaciones de este proyecto de tesis doctoral se organizan en dos vertientes. Por un lado, una contribución a nivel científico al estado del arte en IA y aprendizaje de representaciones. Por otro lado, un impacto directo en la resolución de problemas críticos en la sociedad actual, como la conservación medioambiental y la ciberseguridad.

Relevancia científica. A nivel científico, este proyecto de tesis aborda uno de los desafíos más complejos del aprendizaje automático: la obtención de representaciones representaciones útiles de los datos y la excesiva dependencia de datos etiquetados para la resolución de problemas. La relevancia científica viene principalmente de la mano de los artículos que pertenecen al objetivo O1, relacionado con el estado del arte. Concretamente, se pueden destacar las contribuciones que se enumeran en los siguientes párrafos.

En primer lugar, frente a la tendencia de la literatura científica actual de diferenciar las soluciones según el tipo de datos, nuestro trabajo [16] aporta un marco unificado y agnóstico de la modalidad y de la metodología de la solución. Esto permite a la comunidad científica identificar analogías metodológicas entre el procesamiento de texto, imagen o datos tabulares, acelerando por tanto la innovación interdisciplinar.

Además, dicho artículo no se limita únicamente al análisis del rendimiento algorítmico, sino que también introduce criterios de sostenibilidad, con un gran impacto actual con el auge de la llamada *Green AI*, explicabilidad y seguridad, alineándose con las directrices internacionales para una IA confiable y responsable.

Por último, con el artículo [17] formalizamos el papel de las metodologías de *view alignment* en el contexto de la GPAI, proporcionando una hoja de ruta teórica para la hibridación de paradigmas. Esto ayuda a superar las limitaciones de modelos especialistas actuales para movernos hacia modelos que sean capaces de entender el mundo, no solo hacer predicciones sobre él.

Transferencia de conocimiento y aplicación en la sociedad. La relevancia práctica de la tesis doctoral se consolida mediante la transferencia de modelos abstractos a entornos reales, demostrando que el aprendizaje de representaciones eficiente reduce drásticamente las barreras de adopción de la IA en la sociedad.

Gracias a los artículos MORENA [20] y su extensión débilmente supervisada [19], se automatizó la criba de imágenes vacías en estudios de fauna silvestre. Esto mitiga enormemente el cuello de botella que sufren las organizaciones medioambientales, transformando el laborioso proceso de clasificación o etiquetado de imágenes manual en un proceso completamente automatizado. Al requerir nulo o muy ligero trabajo de etiquetado, la solución es directamente transferible a instituciones con recursos limitados.

En el campo de la ciberseguridad, la aplicación de aprendizaje por contraste a la detección de intrusiones demuestra un beneficio social evidente: la protección de redes e infraestructuras digitales. La metodología permite desplegar modelos robustos incluso en escenarios con escasez de datos etiquetados, donde los enfoques supervisados tradicionales sufren más para obtener resultados competitivos.

Finalmente, todos estos resultados demuestran que los enfoques no supervisados y débilmente supervisados pueden igualar o incluso superar a los modelos completamente supervisados, reduciendo la necesidad de costosos trabajos de etiquetado. Esto democratiza el acceso a herramientas de IA avanzada para centros de investigación u organizaciones que no disponen de los presupuestos de las grandes corporaciones tecnológicas.

Trabajo futuro. Una vez finalizados los proyectos y trabajos destacados presentados en esta sección, aún quedan líneas de investigación abiertas para continuar avanzando en el plan de trabajo.

En relación con el objetivo O2, se siguen planteando diferentes aplicaciones que puedan suponer un descubrimiento en cómo el aprendizaje de representaciones se puede aplicar con éxito en diferentes campos. Uno de ellos es el de la medicina. Actualmente se están desarrollando técnicas basadas en aprendizaje contrastivo para la clasificación de imágenes de ostomías según si la herida está en buenas o malas condiciones, con diferente nivel de gravedad. De nuevo, la aplicación de estas técnicas en un campo tan crítico como la medicina supone,

por un lado, un alivio en la carga de trabajo de los médicos y enfermeros, al poder relegar parte de su trabajo en un sistema inteligente, y por otra parte ofrece una herramienta de soporte a la decisión clínica capaz de identificar de manera temprana la gravedad de las lesiones.

Otro trabajo actualmente en desarrollo es la creación y publicación de un paquete software en el lenguaje de programación R que agrupe diferentes métodos de aprendizaje contrastivo para datos tabulares. Esta tarea responde a la brecha identificada en la literatura actual en la que mientras que el aprendizaje auto-supervisado ha madurado en campos como la visión por computador, la modalidad de datos tabulares ha permanecido relegada a un segundo plano. Al empaquetar e implementar estas técnicas bajo una licencia de código abierto, se persigue facilitar la transferencia del conocimiento adquirido en esta tesis hacia otros investigadores. De este modo, favorecemos la democratización del acceso a arquitecturas avanzadas en un marco estandarizado, reproducible y eficiente en un lenguaje de programación en el que estas técnicas no han sido implementadas previamente.

Referencias

1. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15619–15629 (June 2023)
2. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(8), 1798–1828 (2013). https://doi.org/10.1109/TPAMI.2013.50
3. Cabrera-Bermejo, M., Germán, M., de la Rosa, D., Jesus, M.D., Rivera, A., Elizondo, D., Charte, F., Godoy, M.P.: A didactic overview of transformer applications: model variations and user guidelines. Logic Journal of the IGPL **34**(1), jzaf040 (02 2026). https://doi.org/10.1093/jigpal/jzaf040
4. Charte, D., Charte, F., García, S., del Jesus, M.J., Herrera, F.: A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines. Information Fusion **44**, 78–96 (2018). https://doi.org/10.1016/j.inffus.2017.12.007
5. Chien, A.A., Schmeck, H.: Welcome: Sustainability and computing special section. Communications of the ACM **68**(7), 42–44 (2025). https://doi.org/10.1145/3735631
6. Ericsson, L., Gouk, H., Loy, C.C., Hospedales, T.M.: Self-supervised representation learning: Introduction, advances, and challenges. IEEE Signal Processing Magazine **39**(3), 42–62 (2022). https://doi.org/10.1109/MSP.2021.3134634
7. Fuest, M., Ma, P., Gui, M., Schusterbauer, J., Hu, V.T., Ommer, B.: Diffusion models and representation learning: A survey (2026). https://doi.org/10.1109/TPAMI.2026.3658965
8. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R., Valko, M.: Bootstrap your own latent - a new approach to self-supervised learning. In: Advances in Neural Information Processing Systems. vol. 33, pp. 21271–21284. Curran Associates, Inc. (2020)

9. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16000–16009 (June 2022)
10. Hu, H., Wang, X., Zhang, Y., Chen, Q., Guan, Q.: A comprehensive survey on contrastive learning. Neurocomputing **610**, 128645 (2024). https://doi.org/10.1016/j.neucom.2024.128645
11. Hu, Y., Kuang, W., Qin, Z., Li, K., Zhang, J., Gao, Y., Li, W., Li, K.: Artificial intelligence security: Threats and countermeasures. ACM Comput. Surv. **55**(1), 1–36 (2021). https://doi.org/10.1145/3487890, https://doi.org/10.1145/3487890
12. Kim, Y., Stohr, E.A.: Software reuse: Survey and research directions. Journal of Management Information Systems **14**(4), 113–147 (1998). https://doi.org/10.1080/07421222.1998.11518188
13. Manheim, K., Kaplan, L.: Artificial intelligence: Risks to privacy and democracy. Yale JL & Tech. **21**, 106 (2019)
14. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: Computer Vision – ECCV 2016. pp. 69–84. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_5
15. Pulgar, F.J., Charte, F., Rivera, A.J., del Jesus, M.J.: Clendae: A classifier based on ensembles with built-in dimensionality reduction through denoising autoencoders. Information Sciences **565**, 146–176 (2021). https://doi.org/10.1016/j.ins.2021.02.060
16. de la Rosa, D., González, P., del Jesus, M., Charte, F.: Representation learning: A unified framework of paradigms, modalities and contemporary dimensions. ACM Computing Surveys (under revision) (2026)
17. de la Rosa, D., del Jesus, M., Charte, F.: The eyes of gpai: View alignment self-supervised learning in the context of general purpose artificial intelligence. In: 2026 IEEE Conference of Artificial Intelligence (accepted, yet not published) (May 2026)
18. de la Rosa, D., del Jesus, M., Charte, F.: A study on self-supervised learning objectives for intrusion detection. In: 2026 Conferencia de la Asociación Española para la Inteligencia Artificial (under revision) (2026)
19. de la Rosa, D., Rivera, A., Pérez-Godoy, M., del Jesus, M., Charte, F.: Weakly supervised discarding of photo-trapping empty images based on autoencoders. Progress in Artificial Intelligence (under revision) (2026)
20. Rosa, D.d.l., del Jesus, M.J., Pérez-Godoy, M.D., Charte, F.: Morena: Empty images detection based on unsupervised reconstruction error analysis. In: 2025 International World Conference on Neural Networks. pp. 561–574 (2026)
21. Uelwer, T., Robine, J., Wagner, S.S., Höftmann, M., Upschulte, E., Konietzny, S., Behrendt, M., Harmeling, S.: A survey on self-supervised representation learning. arXiv preprint arXiv:2308.11455 (2023). https://doi.org/10.48550/arXiv.2308.11455
22. Van Wynsberghe, A.: Sustainable ai: Ai for sustainability and the sustainability of ai. AI and Ethics **1**(3), 213–218 (2021). https://doi.org/10.1007/s43681-021-00043-6
23. Vaswani, A., et al.: Attention is all you need. Proc. of NIPS **30**, 5999–6009 (2017)
24. Wang, X., Chen, H., Tang, S., Wu, Z., Zhu, W.: Disentangled representation learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 9677–9696 (2024). https://doi.org/10.1109/TPAMI.2024.3420937