Métrica de Confusión Semántica Explicable entre Datasets para Detección OOD

Arantxa Urrea-Castaño[1,2][0009−0005−4724−0671], Nicolás Segura-Kunsagi[1,2][0009−0002−1413−2277], Juan Luis Suárez-Díaz[1,2][0000−0001−8587−4345], Rosana Montes[1,3][0000−0002−0183−044X], and Francisco Herrera[1,2][0000−0002−7283−312X]

[1] Instituto Andaluz Interuniversitario en Data Science and Computational Intelligence (DaSCI), Universidad de Granada, Granada, España
[2] Departamento de Ciencias de la Computación e Inteligencia Artificial (DECSAI), Universidad de Granada, Granada, España {s.arantxa.uc, nkunsagi, jlsuarezdiaz, herrera}@ugr.es
[3] Departamento de Ingeniería del Software (LSI), Universidad de Granada, Granada, España rosana@ugr.es

Resumen. En la detección de datos fuera de distribución (OOD), la similitud semántica entre los datos de clases conocidas (ID) y los datos OOD determina la dificultad de la tarea. Conocer esta similitud a nivel de clase resulta de utilidad en múltiples escenarios, como la selección y evaluación de detectores OOD, la identificación de clases que requieren un tratamiento especial o la auditoría del comportamiento del modelo en sistemas de alto riesgo. La métrica más relacionada con este problema, el CLP (Confusion Log Probability), produce un único valor por par de datasets, sin revelar qué clases concentran la similitud observada ni corregir los sesgos introducidos por diferencias en el número de clases de los datasets. En este trabajo se propone CASE (Confusion Analysis for Semantic Explainability, análisis de confusión para la explicabilidad semántica), una métrica de similitud entre datasets de imágenes que aborda estas limitaciones de forma conjunta. A través de una formulación por pasos basada en las probabilidades softmax de un clasificador entrenado conjuntamente con ambos datasets, CASE estudia la similitud global y por clase de los datasets, corrige el sesgo debido al diferente número de clases en cada dataset y pondera las clases OOD según su representación muestral en el test. La interpretabilidad es intrínseca a la formulación y no requiere ningún método de explicabilidad externo, situando a CASE en el ámbito de la IA explicable por diseño. Los experimentos sobre siete datasets heterogéneos muestran que CASE produce resultados coherentes con el CLP , y revela además relaciones semánticas que el CLP atenúa por efecto del desequilibrio de clases. El análisis por clase sobre CIFAR-10 y CIFAR-100 demuestra que la confusión se concentra en subconjuntos reducidos de clases semánticamente afines, información invisible para el CLP.

Keywords: Detección Fuera de Distribución · Métricas de similitud · Inteligencia Artificial Explicable

1. Introducción

La detección de datos fuera de distribución (Out-Of-Distribution, OOD) es un problema fundamental para el uso seguro de sistemas de aprendizaje automático, con consecuencias directas en aplicaciones de alto riesgo como el diagnóstico médico o la conducción autónoma [4,5]. La dificultad de esta tarea depende en gran medida de la similitud semántica entre la distribución de la que proceden los datos de entrenamiento (In-Distribution, ID) y la distribución OOD. Mientras que las muestras far-OOD son relativamente fáciles de detectar, las muestras near-OOD, semánticamente próximas a las clases conocidas, presentan un reto considerablemente mayor [5]. Cuantificar esta similitud es por tanto una condición necesaria para caracterizar la dificultad de la tarea de detección OOD, y resulta igualmente relevante en tareas relacionadas como el aprendizaje por transferencia [6,7] o la adaptación de dominio [8,9].

Sin embargo, la similitud entre datasets no tiene por qué ser homogénea entre sus clases. Dos datasets pueden presentar una similitud global moderada y concentrar toda su confusión semántica en un subconjunto reducido de clases. Conocer qué clases se solapan más permite reforzar selectivamente detectores OOD frente a distribuciones near-OOD, dirigir el ajuste fino en las regiones más problemáticas del espacio de representación [10], y auditar el comportamiento del modelo a nivel de clase para sistemas de alto riesgo [11]. Las métricas existentes son, sin embargo, ciegas a esta estructura: producen un único valor por par de datasets sin revelar qué clases concentran la similitud observada, y no corrigen el sesgo introducido por el desequilibrio en el número de clases entre datasets.

En esta investigación se propone CASE (Confusion Analysis for Semantic Explainability, análisis de confusión para la explicabilidad semántica), una nueva métrica de similitud entre datasets que aborda estas limitaciones. Con ella es posible conocer la similitud global entre un par de datasets y la confusión entre clases sin requerir ningún método de explicabilidad externo, e incorpora correcciones explícitas del sesgo por proporción de clases de los datasets y por desbalance muestral en el conjunto de test. Las contribuciones principales son: (i) una métrica de similitud a nivel de clase aplicable a detección OOD y clasificación de imágenes; (ii) interpretabilidad intrínseca por clase sin componentes de explicabilidad externos; (iii) corrección del sesgo de proporción de clases y del desbalance muestral; y (iv) evaluación experimental sobre siete datasets heterogéneos comparando con la métrica Confusion Log Probability (CLP) [5] como línea base.

El artículo se estructura de la siguiente manera. La Sección 2 revisa el trabajo relacionado, con especial atención a la métrica CLP y al paradigma de la interpretabilidad intrínseca. La Sección 3 describe la formulación de CASE. La Sección 4 detalla la configuración experimental y los datasets empleados. La Sección 5 presenta los resultados a nivel de dataset y a nivel de clase, y analiza las implicaciones y limitaciones de la propuesta, y la Sección 6 recoge las conclusiones principales y las líneas de trabajo futuro.

2. Trabajo Relacionado

La comparación entre datasets de imágenes suele abordarse midiendo discrepancias entre sus distribuciones en un espacio de representación. Métricas como FID [1], MMD [2] o OTDD [3] abordan este problema desde distintas perspectivas, pero proporcionan principalmente una noción global de discrepancia que resulta insuficiente cuando se busca capturar similitud semántica a nivel de clase entre conjuntos ID y OOD. Estos trabajos motivan el desarrollo de métricas más específicas para analizar relaciones semánticas entre conjuntos ID y OOD.

La métrica más directamente relacionada con nuestra propuesta es la Confusion Log Probability (CLP) [5], introducida en el contexto de la detección OOD para cuantificar la similitud semántica entre un dataset ID y uno OOD. El CLP entrena varios modelos de clasificación sobre la unión de todos los datasets considerados, de forma que cada modelo aprende representaciones conjuntas sobre el espacio de clases completo. Para cada par de datasets, el CLP mide la log-probabilidad media que los modelos asignan a las muestras OOD de pertenecer a las clases ID, capturando así la dificultad de la tarea de detección y permitiendo distinguir entre los regímenes near-OOD y far-OOD. Sin embargo, el CLP colapsa el resultado en un único valor por par de datasets, no corrige el sesgo introducido por el desequilibrio en el número de clases, y no contempla el desbalance muestral entre clases OOD en el conjunto de test.

En el ámbito de la explicabilidad, el campo de la inteligencia artificial explicable (eXplainable Artificial Intelligence, XAI) ha producido métodos que atribuyen las decisiones del modelo a características de entrada o conceptos semánticos [12,13]. Sin embargo, las métricas de similitud entre datasets existentes producen valores opacos que no revelan qué clases o conceptos visuales explican la similitud medida, dejando sin explorar la aplicación de los principios de interpretabilidad al problema de la comparación de datasets. En las secciones siguientes se propone una métrica que cubre esta brecha al incorporar la interpretabilidad por clase de forma intrínseca a su formulación, sin requerir ningún componente externo.

3. Descripción de la métrica: CASE

Consideramos uno o varios modelos de clasificación entrenados conjuntamente con dos datasets: un dataset ID, denotado D_{ID} con C_{ID} clases, y un dataset OOD, denotado D_{OOD} con C_{OOD} clases. La capa de salida produce un vector softmax de dimensión $C_{\text{ID}} + C_{\text{OOD}}$, correspondiente al conjunto unificado de clases. En la fase de test se utilizan únicamente imágenes de D_{OOD} de las que conocemos sus clases de origen. El objetivo de CASE es cuantificar en qué medida son confundidas con las clases ID.

3.1. Probabilidad de confusión por imagen

Sea $s_i^{(m)}(x)$ la probabilidad asignada por el modelo de clasificación $m \in \{1, \ldots, M\}$ a la clase ID i para una imagen x del test. Aunque el vector softmax

incluye tanto clases ID como OOD, el análisis se restringe a las C_{ID} componentes ID. La predicción agregada sobre los M modelos se obtiene por promedio:

$$\bar{s}_i(x) = \frac{1}{M} \sum_{m=1}^{M} s_i^{(m)}(x), \quad i = 1, \ldots, C_{\text{ID}}. \tag{1}$$

Este vector permite inspeccionar la distribución de confusión de una imagen individual x, examinando directamente $\bar{s}_i(x)$ para cada clase ID i.

3.2. Distribución de confusión por clase OOD

Para cada clase OOD, $c \in \{1, \ldots, C_{\text{OOD}}\}$, sea $D_{\text{OOD}}^{(c)}$ el subconjunto de imágenes de test pertenecientes a dicha clase, y sea $n_c = |D_{\text{OOD}}^{(c)}|$ el número de imágenes de dicho subconjunto. La probabilidad media con la que las imágenes de la clase OOD c son confundidas con la clase ID i es:

$$\bar{s}_i^{(c)} = \frac{1}{n_c} \sum_{x \in D_{\text{OOD}}^{(c)}} \bar{s}_i(x), \quad i = 1, \ldots, C_{\text{ID}}. \tag{2}$$

El vector $\bar{\mathbf{s}}^{(c)} = \left(\bar{s}_1^{(c)}, \ldots, \bar{s}_{C_{\text{ID}}}^{(c)}\right) \in R^{C_{\text{ID}}}$ representa la distribución de confusión de la clase OOD c, donde cada componente $\bar{s}_i^{(c)}$ cuantifica la probabilidad media con la que las imágenes de esa clase son asignadas a la clase ID i. Dado que el softmax completo distribuye probabilidad sobre $C_{\text{ID}} + C_{\text{OOD}}$ clases, se cumple $\sum_i \bar{s}_i^{(c)} \leq 1$, con igualdad solo si el modelo no asigna ninguna probabilidad a las clases OOD. Este vector es el elemento central de la interpretabilidad de CASE, permitiendo identificar directamente qué clases ID concentran la confusión de cada clase OOD.

3.3. Confusión por clase OOD y métrica global

Para obtener una medida de confusión por clase OOD, promediamos la distribución de confusión sobre todas las clases ID:

$$f^{(c)} = \frac{1}{C_{\text{ID}}} \sum_{i=1}^{C_{\text{ID}}} \bar{s}_i^{(c)}. \tag{3}$$

En el caso en que todas las clases OOD tengan igual representación en el conjunto de test, la confusión global se obtiene como la media aritmética sobre todas las clases OOD:

$$F = \frac{1}{C_{\text{OOD}}} \sum_{c=1}^{C_{\text{OOD}}} f^{(c)}. \tag{4}$$

Cuando las clases OOD no están igualmente representadas, por ejemplo, cuando el desbalance refleja la distribución real de la población de estudio, puede

ser deseable ponderar cada clase por su tamaño muestral. Para ello se asigna a cada clase OOD un peso proporcional a n_c:

$$w_c = \frac{n_c}{\sum_{c=1}^{C_{\text{OOD}}} n_c}, \tag{5}$$

obteniéndose la forma general:

$$F = \sum_{c=1}^{C_{\text{OOD}}} w_c \cdot f^{(c)}. \tag{6}$$

Los pesos se aplican sobre $f^{(c)}$, y no directamente sobre el vector $\bar{\mathbf{s}}^{(c)}$, para preservar la interpretación de cada componente $\bar{s}_i^{(c)}$ como probabilidad media de confusión de la clase OOD c con la clase ID i.

3.4. Corrección del sesgo de proporción de clases

Cuando $C_{\text{ID}} \neq C_{\text{OOD}}$, la fracción de masa de probabilidad softmax que recae sobre las clases ID depende no solo del comportamiento del modelo sino también de la proporción relativa de clases en la capa de salida. En consecuencia, un clasificador con predicciones próximas a la distribución uniforme transferirá sistemáticamente más masa hacia el conjunto de clases más numeroso, independientemente de cualquier similitud semántica real. Para corregir este sesgo estructural y obtener una medida comparable entre pares de datasets con distinto número de clases, definimos:

$$B = \frac{C_{\text{OOD}}}{C_{\text{ID}} + C_{\text{OOD}}}, \tag{7}$$

que representa la fracción de probabilidad que un clasificador con distribución de salida uniforme asignaría a las clases OOD. La métrica corregida se define como:

$$F_{\text{adj}} = \frac{F}{F + (1 - F)\,\dfrac{B}{1 - B}}. \tag{8}$$

El efecto de B sobre la corrección depende directamente de la proporción entre clases. Si $C_{\text{OOD}} > C_{\text{ID}}$ ($B > 0{,}5$), la corrección reduce F, descartando la confusión atribuible al mayor peso probabilístico de las clases OOD. Si $C_{\text{OOD}} < C_{\text{ID}}$ ($B < 0{,}5$), la amplifica, compensando que las clases OOD reciben estructuralmente menos masa de probabilidad. Si $C_{\text{ID}} = C_{\text{OOD}}$ ($B = 0{,}5$), $F_{\text{adj}} = F$ y la corrección es nula.

F_{adj} se interpreta como sigue:

- $F_{\text{adj}} \approx 0$: el modelo asigna casi toda la probabilidad a las clases OOD. Ambos datasets son semánticamente separables.

- $F_{\text{adj}} \approx 0{,}5$: el modelo no distingue entre clases ID y OOD, lo que refleja una alta similitud semántica entre ambos datasets.
- $F_{\text{adj}} \approx 1$: el modelo asigna casi toda la probabilidad a las clases ID al procesar imágenes OOD, indicando una inversión de la distribución de probabilidad esperada.

3.5. Propiedades y comparación con el CLP

CASE produce dos salidas complementarias: el escalar F_{adj}, comparable entre pares de datasets con distinto número de clases, y los vectores $\bar{\mathbf{s}}^{(c)}$ por clase OOD, que permiten identificar qué clases concentran la confusión y con qué clases ID se produce, sin requerir ningún método de explicabilidad externo.

A diferencia del CLP [5], que no pondera por clase ni corrige el sesgo de proporción de clases, CASE ofrece:

- Desglose por clase: $\bar{\mathbf{s}}^{(c)}$ revela qué clases OOD son más confundibles y con qué clases ID, permitiendo el ajuste fino selectivo o el refuerzo dirigido del detector OOD.
- Corrección del sesgo estructural: F_{adj} corrige el sesgo introducido por el desequilibrio en el número de clases, haciendo la métrica directamente comparable entre pares de datasets con distinta cardinalidad.
- Ponderación por tamaño de clase: los pesos w_c permiten reflejar el desbalanceo real entre clases OOD en el conjunto de test, evitando que clases minoritarias distorsionen la métrica global.

4. Configuración Experimental

Los experimentos se diseñaron para evaluar CASE en dos niveles: a nivel de dataset, comparando F_{adj} con el CLP como línea base; y a nivel de clase, ilustrando el valor interpretativo de $\bar{\mathbf{s}}^{(c)}$ sobre el par de datasets CIFAR-10 y CIFAR-100.

4.1. Datasets y protocolo de entrenamiento

Los experimentos emplean siete datasets de imágenes que cubren dominios visuales heterogéneos, resumidos en la Tabla 1.

VMMRdb se particiona en dos subconjuntos de clases disjuntos, VMMRdb ID y VMMRdb OOD, lo que permite evaluar CASE en un escenario de mayor similitud semántica respecto a la comparativa entre otros datasets.

Para cada par de datasets (ID, OOD), se entrena un único clasificador conjunto con ResNet-34 [20] como arquitectura base, de forma que la capa de salida produce un vector softmax sobre la unión de todas las clases de ambos datasets. Los modelos se entrenaron desde cero durante un máximo de 500 épocas con optimización SGD, tasa de aprendizaje inicial de 0.2, momento de 0.9 y decaimiento de pesos de 10^{-6}. El entrenamiento empleó un tamaño de batch de

Cuadro 1. Datasets empleados en los experimentos.

Dataset	Clases	Imágenes	Dominio
CIFAR-10 [15]	10	60.000	Objetos varios
CIFAR-100 [15]	100	60.000	Objetos varios
EMNIST Digits [16]	10	280.000	Dígitos manuscritos
EMNIST Letters [16]	26	145.600	Letras manuscritas
DTD [17]	47	5.640	Texturas
Tiny ImageNet [18]	200	110.000	Objetos varios
VMMRdb [19]	9.170	291.752	Vehículos

64 muestras con acumulación de gradientes hasta alcanzar un batch efectivo de 2048, junto con un scheduler de cosine annealing con $T_{\text{máx}} = 500$. El análisis de similitud obtenido en épocas intermedias (350, 400, 450) muestra que los patrones de confusión entre datasets se estabilizan a partir de la época 400, con variaciones inferiores a ± 0.02 en los pares de mayor similitud, lo que indica que los resultados a 500 épocas son representativos del comportamiento del modelo ya convergido.

4.2. Evaluación

CASE se calcula sobre el conjunto de test del dataset OOD de cada par, siguiendo el procedimiento descrito en la Sección 3, y se compara con el CLP [5]. Los experimentos se organizan en dos bloques: comparación de F_{adj} con el CLP sobre todos los pares de datasets, cuyos resultados se presentan en la Sección 5.1; y análisis de la distribución de confusión por clase $\bar{\mathbf{s}}^{(c)}$ sobre el par CIFAR-10/CIFAR-100, presentado en la Sección 5.2.

5. Resultados

5.1. Comparación con el CLP

Las Figuras 1 y 2 muestran los mapas de calor de similitud entre datasets obtenidos con CASE y con el CLP, sin y con transformación logarítmica respectivamente. La diagonal principal se omite por corresponder al entrenamiento conjunto de cada dataset consigo mismo.

Sin transformación logarítmica, ambas métricas identifican de forma consistente los mismos pares de alta similitud semántica: CIFAR-10/CIFAR-100 (0.532), EMNIST Digits/Letters (0.258) y VMMRdb ID/VMMRdb OOD (0.225), confirmando la coherencia de CASE con el estado del arte. Adicionalmente, F_{adj} revela una similitud destacada entre Tiny ImageNet y CIFAR-100 (0.412–0.496) que el CLP no captura con la misma magnitud, lo que sugiere que la corrección del sesgo de proporción de clases expone relaciones semánticas atenuadas por el mayor número de clases de Tiny ImageNet.

Con transformación logarítmica, ambas métricas ofrecen una mejor discriminación entre los pares de baja similitud, manteniendo un ordenamiento relativo coherente entre datasets.

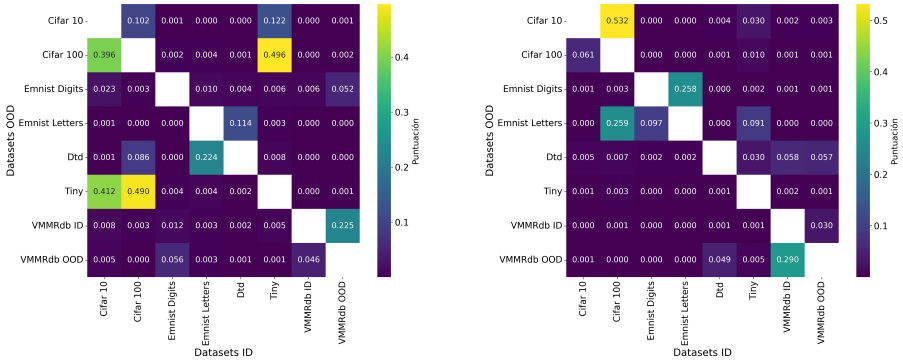


Figura 1. Similitud entre datasets sin transformación logarítmica. Izquierda: CASE (F_{adj}). Derecha: CLP. Valores más altos indican mayor similitud semántica.

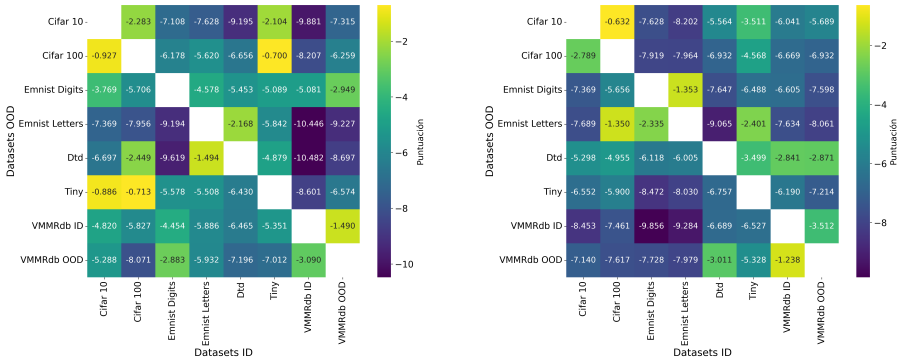


Figura 2. Similitud entre datasets con transformación logarítmica. Izquierda: CASE. Derecha: CLP. La escala logarítmica amplifica las diferencias entre pares de baja similitud.

5.2. Análisis de confusión por clase: CIFAR-10 y CIFAR-100

CASE permite identificar qué pares de clases concretos concentran la confusión entre datasets. Para ilustrar esta capacidad, analizamos en detalle el par CIFAR-10/CIFAR-100, que presentan una similitud semántica elevada.

Cuando CIFAR-10 actúa como dataset OOD, la probabilidad máxima de confusión con alguna clase ID, es decir, la probabilidad asignada a la clase ID más probable para cada clase OOD, presenta una mediana de 0.018 y un rango intercuartílico [0.014, 0.025]. Estos valores indican que la mayoría de las clases OOD no alcanzan una confusión significativa con ninguna clase ID. Sin embargo, las clases con mayor afinidad semántica destacan por sus altos valores: automóvil alcanza una probabilidad de 0.128 hacia camioneta, y camión de 0.105 hacia autobús, valores que triplican el tercer cuartil. El mismo patrón se observa en los

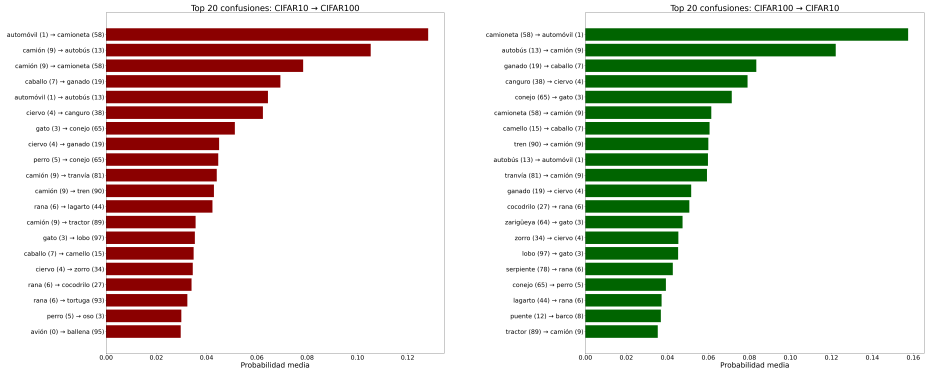


Figura 3. Top 20 confusiones por clase entre CIFAR-10 y CIFAR-100. Izquierda: CIFAR-10 como OOD, CIFAR-100 como ID. Derecha: CIFAR-100 como OOD, CIFAR-10 como ID.

animales: caballo/ganado, ciervo/canguro y gato/conejo que también concentran parte de las confusiones más intensas, con probabilidades superiores a 0.06.

Cuando CIFAR-100 actúa como dataset OOD, la mediana desciende a 0.003 con un rango intercuartílico [0.001, 0.007], reflejando que la gran mayoría de las 100 clases de CIFAR-100 no superan una confusión mínima con ninguna clase ID de CIFAR-10. No obstante, un subconjunto reducido de clases concentra confusiones muy superiores: camioneta alcanza 0.158 hacia automóvil y autobús 0.122 hacia camión, más de veinte veces el valor de la mediana. Esta concentración extrema en pocas clases, invisible para el CLP, es precisamente la información adicional que CASE aporta.

En ambas direcciones las confusiones más intensas responden a una lógica semántica consistente (vehículos con vehículos, animales con animales de morfología similar) lo que confirma que CASE captura similitud semántica genuina.

5.3. Discusión

Los resultados experimentales muestran que CASE produce valores coherentes con el CLP, confirmando su solidez como métrica de similitud entre datasets. Al mismo tiempo, la corrección del sesgo de proporción de clases permite a CASE revelar relaciones semánticas que el CLP atenúa cuando existe un desequilibrio en el número de clases entre datasets, lo que resulta una condición necesaria para la comparabilidad entre pares de datasets, especialmente entre aquellos con una gran diferencia en el número de clases.

El análisis por clase sobre el par CIFAR-10/CIFAR-100 ilustra la principal ventaja diferencial de CASE. Mientras el CLP produce un único valor agregado, CASE descompone la similitud en perfiles de confusión por clase que revelan qué clases específicas concentran el solapamiento semántico y en qué proporción contribuye cada clase OOD al valor global. Gracias a la interpretabilidad intrínseca de CASE, en detección OOD es posible reforzar selectivamente el detector en las

clases más confundibles, en aprendizaje por transferencia priorizar el ajuste fino en las clases con mayor solapamiento, y en auditoría de sistemas de alto riesgo identificar y justificar los puntos de fallo potencial del modelo a nivel de clase sin necesidad de métodos de explicabilidad externos, entre otros posibles usos en tareas donde el conocimiento de la similitud semántica a nivel de clase resulte relevante.

Una limitación de CASE, compartida con el CLP, es su dependencia del clasificador empleado. Los valores obtenidos reflejan la similitud semántica tal como la percibe el modelo entrenado, no una propiedad intrínseca de los datasets. Para mitigar la variabilidad asociada a un único entrenamiento, se recomienda promediar los resultados sobre varios modelos entrenados con distintas inicializaciones. En los experimentos presentados no se ha aplicado este procedimiento por limitaciones computacionales, si bien la formulación de CASE lo contempla de forma nativa mediante el parámetro M.

Por otro lado, CASE es una métrica asimétrica, de modo que el valor obtenido cuando un dataset actúa como OOD y otro como ID no tiene por qué coincidir con el valor obtenido cuando se intercambian los roles. Esta asimetría es esperable dado que el número de clases, el tamaño muestral y la variabilidad semántica difieren entre datasets, haciendo que la confusión en cada dirección sea diferente. Aunque F_{adj} corrige los sesgos más evidentes, la asimetría residual es deseable ya que refleja diferencias reales en cómo cada dominio se solapa con el otro desde la perspectiva del modelo aprendido, capturando la dificultad asimétrica inherente a la tarea de detección OOD según cuál sea el dataset ID y cuál el OOD, información que una métrica simétrica no puede preservar.

6. Conclusión

En este trabajo hemos presentado CASE, una nueva métrica cuyo objetivo es cuantificar la similitud semántica entre datasets de imágenes. A través de una formulación por pasos, CASE corrige el sesgo introducido por el desequilibrio en el número de clases entre datasets y el sesgo derivado de la distribución muestral desigual de las clases en el conjunto de test, proporcionando en cada etapa resultados interpretables sin necesidad de métodos de explicabilidad externos. La descomposición de la similitud global en perfiles de confusión por clase sitúa a CASE en el ámbito de la IA explicable por diseño, respondiendo directamente a los principios de transparencia y robustez que definen la IA confiable [14,11].

Los experimentos sobre siete datasets heterogéneos muestran que CASE produce resultados coherentes con el CLP, y revela además relaciones semánticas que el CLP atenúa por efecto del desequilibrio de clases. El análisis por clase sobre CIFAR-10/CIFAR-100 muestra que la confusión entre datasets se concentra en subconjuntos reducidos de clases semánticamente afines, información invisible para las métricas de valor único existentes y que la interpretabilidad intrínseca de CASE pone al alcance del profesional sin necesidad de herramientas XAI adicionales. Como trabajo futuro, sería de interés extender CASE a arquitecturas de mayor capacidad, explorar su aplicación en dominios especializados de alto

riesgo como el diagnóstico médico, y estudiar el uso del perfil de confusión por clase como señal de guía para estrategias de ajuste fino selectivo.

Agradecimientos. Esta investigación es parte del Proyecto TSI-100927-2023-1, financiado por Plan de Recuperación, Transformación y Resiliencia Europeo, de la Unión Europea-Next Generation EU a través del Ministerio para la Transformación Digital y de la Función Pública y está parcialmente financiada por el Ministerio de Ciencia y Tecnología de España en el marco del proyecto PID2023-150070NB-I00, financiado por MICIU/AEI/10.13039/501100011033 y por FEDER, UE.

Declaración de intereses. Los autores no tienen ningún conflicto de intereses que declarar que sea relevante para el contenido de este artículo.

Referencias

1. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Advances in Neural Information Processing Systems, vol. 30. Curran Associates (2017)
2. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. Journal of Machine Learning Research **13**, 723–773 (2012)
3. Alvarez-Melis, D., Fusi, N.: Geometric dataset distances via optimal transport. In: Advances in Neural Information Processing Systems, vol. 33. Curran Associates (2020)
4. Miyai, A., Yu, Q., Irie, G., Aizawa, K.: Generalized out-of-distribution detection and beyond in vision language model era: a survey. arXiv preprint arXiv:2407.21794 (2025). https://arxiv.org/abs/2407.21794
5. Winkens, J., Bunel, R., Roy, A.G., Stanforth, R., Natarajan, V., Ledsam, J.R., MacWilliams, P., Kohli, P., Karthikesalingam, A., Kohl, S., Cemgil, T., Eslami, S.M.A., Ronneberger, O.: Contrastive training for improved out-of-distribution detection. arXiv preprint arXiv:2007.05566 (2020). https://arxiv.org/abs/2007.05566
6. Sun, S., Zhang, H.H.: Quantifying dataset similarity to guide transfer learning. arXiv preprint arXiv:2510.10866 (2025). https://arxiv.org/abs/2510.10866
7. Alvarez-Melis, D., Fusi, N.: Geometric dataset distances via optimal transport. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 21428–21439. Curran Associates (2020)
8. Ben-David, S., Blitzer, J., Crammer, A., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different distributions. Machine Learning **79**(1–2), 151–175 (2010). https://doi.org/10.1007/s10994-009-5152-4
9. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of Machine Learning Research **17**(59), 1–35 (2016)
10. Ji, R., Tekalp, A.M.: Comparison of fine-tuning strategies for transfer learning in medical image classification. Image and Vision Computing **146**, 105008 (2024). https://doi.org/10.1016/j.imavis.2024.105008
11. European Parliament and Council of the European Union: Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union (2024). https://artificialintelligenceact.eu

12. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion **58**, 82–115 (2020). https://doi.org/10.1016/j.inffus.2019.12.012
13. Zhang, Q., Wu, Y.N., Zhu, S.C.: Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond. Knowledge and Information Systems **64**(12), 3197–3234 (2021). https://doi.org/10.1007/s10115-022-01756-8
14. Liu, B. et al.: Trustworthy AI: a computational perspective. ACM Transactions on Intelligent Systems and Technology **14**(1), 1–59 (2023). https://doi.org/10.1145/3546872
15. Krizhevsky, A.: Learning multiple layers of features from tiny images. Technical Report, University of Toronto (2009). https://www.cs.toronto.edu/ kriz/learning-features-2009-TR.pdf
16. Cohen, G., Afshar, S., Tapson, J., van Schaik, A.: EMNIST: an extension of MNIST to handwritten letters. arXiv preprint arXiv:1702.05373 (2017). https://arxiv.org/abs/1702.05373
17. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3606–3613. IEEE (2014)
18. Le, Y., Yang, X.: Tiny ImageNet visual recognition challenge. CS 231N, **7**(7), 3 (2015)
19. Tafazzoli, F., Frigui, H., Nasrollahi, K.: A large and diverse dataset for improved vehicle make and model recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE (2017)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. IEEE (2016)