ISSN 1870-4069

Explicación visual de la predicción de un clasificador de imágenes

Tonantzin M. Guerrero Velázquez, Juan Humberto Sossa Azuela

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
México

tguerrero_a13@sagitario.cic.ipn.mx,
hsossa@cic.ipn.mx

Abstract. El aprendizaje automático se ha convertido en una herramienta necesaria en la industria actual. La mayoría de los métodos comúnmente usados, más allá de las métricas y pruebas realizadas en realidad no se comprenden las razones detrás de sus predicciones. En el trabajo se propone un nuevo método para entender de manera visual el trasfondo de una predicción realizada.

Keywords: Explainability, Classifier, XAI, Machine Learning.

Visual Explanation of the Prediction of an Image Classifier

Abstract. Machine learning has become a necessary tool in today's industry. Most of the commonly used methods, beyond the metrics and tests performed, don't actually understand the reasons behind their predictions. In the work, a new method is proposed to understand visually the background of a prediction made.

Keywords: Explainability, Classifier, XAI, Machine Learning.

1. Introducción

Actualmente la Inteligencia Artificial Explicable (XAI, por sus siglas en inglés) es un área de gran interés dentro del aprendizaje automático. Aunque es un campo relativamente nuevo su atractivo radica en la usabilidad que se le puede adjudicar. La XAI se refiere a mejorar el entendimiento humano y a justificar las decisiones hechas por un modelo de aprendizaje automático. La explicabilidad en el contexto de la Inteligencia Artificial denota cualquier acción o proceso tomado por un modelo con la intención de aclarar o detallar su funcionamiento. Por otro lado, la interpretabilidad es un concepto que muchas veces es tomado de la misma manera, sin embargo, se refiere

al nivel en el que un modelo dado tiene sentido para un ser humano, esta característica también se puede expresar como la transparencia de un modelo [1].

Un modelo se considera ser transparente si por sí mismo es entendible, por ejemplo, un modelo de regresión logística o lineal, un árbol de decisiones o un clasificador basado en reglas [2]. Existen diferentes enfoques de explicabilidad, uno de ellos se basa en entender mejor los datos a través de técnicas que permitan conocer información relevante de los mismos, tales como, mapas salientes, análisis de sensibilidad, etc. y de esta manera poder entender qué características de estos llevan a que un modelo realice una predicción u otra. El otro enfoque se basa en entender mejor el funcionamiento de un modelo, por ejemplo, visualizando la activación de las capas de una red neuronal [3].

Actualmente existen diversas técnicas que permiten obtener una explicabilidad de un modelo de aprendizaje profundo. Una de ellas es el marco de interpretabilidad LIME (*Local Interpretable Model-Agnostic Explanations*). Para explicar la predicción de un clasificador de imagen, LIME crea un conjunto de imágenes resultantes al perturbar la imagen de entrada, dividiéndola en componentes interpretables (*superpixeles*) para así obtener una probabilidad de pertenencia para cada una de esas instancias perturbadas. Con este proceso, LIME genera una explicación visual con base en la clasificación de esos nuevos datos perturbados dando como resultado un área de la imagen de entrada que denota aquello en lo que se fijó el modelo para llevar a cabo la predicción [4].

Una de las desventajas más importantes de las técnicas de explicación visuales existentes es la subjetividad de los resultados. Estos se encuentran sujetos a la interpretabilidad de quien los utiliza, es de esperarse que no se genere la suficiente confianza en los mismos. Otra desventaja radica también en el hecho de que los resultados resultan ser aleatorios, ya que incluso bajo las mismas condiciones, es decir, la misma entrada y los mismos parámetros de configuración, los resultados obtenidos siempre son distintos y esto hace también que las herramientas de explicación sean poco confiables.

En el presente trabajo se da solución a los problemas descritos anteriormente (subjetividad y aleatoriedad de los resultados), además de usar a el mismo modelo clasificador como herramienta de explicación. La solución propuesta para resolver el problema consiste en generar una explicación visual de la predicción basada en la caracterización de ciertas regiones de la imagen según su importancia para la predicción, por lo tanto, el principio del método propuesto consta de un algoritmo de búsqueda de regiones que puedan ser caracterizadas y utilizadas finalmente como explicación visual.

La contribución principal de este trabajo radica en disminuir la subjetividad de las explicaciones. Esta cuestión se ataca desde dos frentes: 1) No se utiliza ningún parámetro de configuración para el algoritmo, lo que asegura que no depende de la persona que lo implemente y, por tanto, siempre se llegará al mismo resultado, y 2) el resultado es claro y fácil de interpretar a la vez que intuitivo dadas las categorías presentadas, de tal forma que cualquier persona que conozca el código de colores será capaz de dar una explicación y esta será muy similar de manera inequívoca a la de cualquier otra persona.

2. Método

El método aquí propuesto consiste, de manera general, en realizar una búsqueda selectiva de regiones candidatas como regiones de interés, las cuales serán evaluadas usando el mismo clasificador para encontrar las regiones útiles, además de una función escalonada que indica el grado de relevancia de esa región para predecir la clase en cuestión.

2.1. Búsqueda de regiones útiles

Se busca generar una explicación visual de la predicción basada en la caracterización de regiones según su importancia para el proceso de predicción, por lo tanto, el principio del método consta de un algoritmo de búsqueda de regiones que puedan ser caracterizadas y usadas como explicación visual.

Si bien se podría implementar un tipo de búsqueda exhaustiva usando segmentación o algún algoritmo de ventana deslizante, esto además de contar con un espacio de búsqueda enorme, también generaría regiones indeseables debido a su tamaño o forma que no aportan en sí ninguna información útil a la explicación de la predicción.

En este trabajo se utiliza el algoritmo de búsqueda selectiva [5], pensado para abarcar regiones de distintas escalas lo que garantiza primero, que todos los objetos sean tomados en cuenta y segundo, que exista una distinción marcada en el tamaño de las regiones para así realizar una mejor caracterización de estas.

Para llevar a cabo la búsqueda de regiones, se usa el método de segmentación basada en grafos [6], donde se considera la entrada como un grafo: $G = (V, E)$, con n el número de vértices y m el número de aristas de G. El objetivo es segmentar V en r componentes, que corresponden a las primeras regiones propuestas, las cuáles se agruparan usando un algoritmo voraz basado en la similitud de las regiones con un enfoque de abajo hacia arriba (*bootom-up*). Esta similitud entre regiones debe ser propagada de manera jerárquica, para esto se usa la siguiente formulación:

$$S_{colour}(r_i, r_j) = \sum_{k=1}^{n} \min(c_i^k, c_j^k), \tag{1}$$

como similitud de color para cada par de regiones r_i, r_j usando el histograma de intersección donde c_i y c_j son los histogramas de las regiones respectivamente.

Si se obtiene el histograma de texturas para cada región entonces se puede calcular la medida de similitud de texturas como sigue:

$$S_{texture}(r_i, r_j) = \sum_{k=1}^{n} \min(t_i^k, t_j^k). \tag{2}$$

Para obligar a las regiones pequeñas a juntarse a regiones más grande, se utiliza la siguiente medida de similitud:

$$S_{size}(r_i, r_j) = 1 - \frac{size(r_i) + size(r_j)}{size(im)}, \tag{3}$$

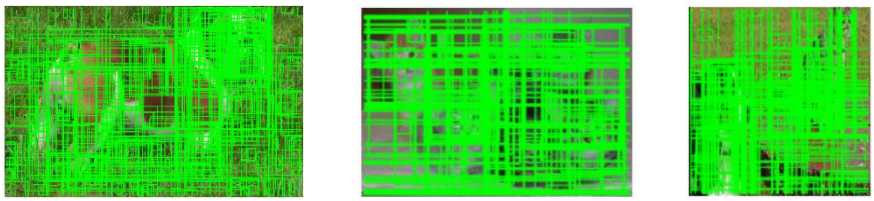

Fig. 1. Regiones encontradas con el algoritmo de búsqueda selectiva para tres imágenes distintas cuyas clases son a) *French_Bulldog*, b) *Egyptian_cat* y c) *Egyptian_cat* respectivamente.

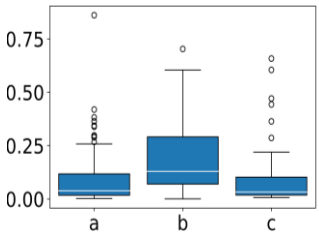

Fig. 2. Cuartiles y valores poco comunes para las imágenes a), b) y c) y de la Fig. 1.

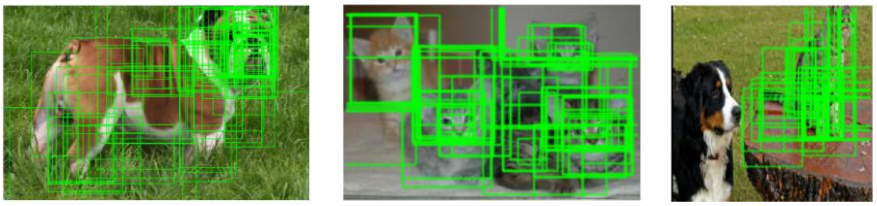

Fig. 3. Conjunto de regiones útiles R_u para las imágenes a, b y c respectivamente.

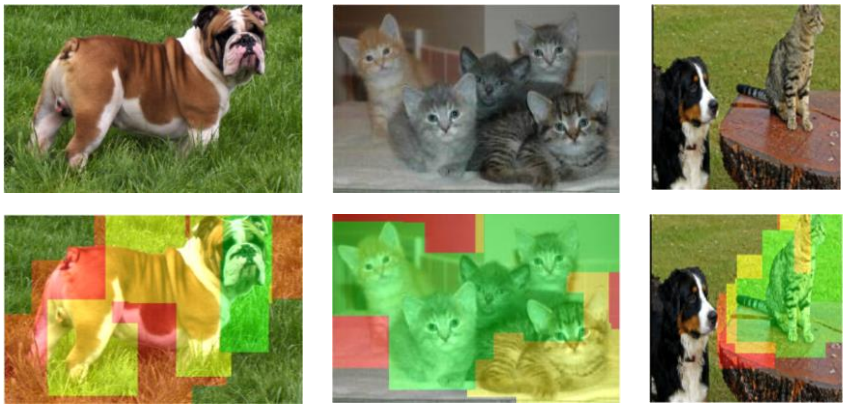

Fig. 4. Resultado de la explicación visual (abajo) para tres imágenes distintas (arriba).

donde $size(im)$ es el tamaño de la imagen en pixeles. Para tomar en cuenta diferentes escenas y condiciones de luz el proceso anterior se realiza para diferentes espacios de color, RGB y HSV.

Por último, del conjunto de regiones encontradas $R = \{r_1, \dots, r_n\}$, donde $r_i = \{x, y, w, h\}$, es decir, que cada región r_i representa un recuadro delimitador con el par (x, y) representando su posición y el par (w, h) su tamaño, como se muestra en la **Fig. 1**, es indispensable conocer cuáles de ellas son útiles, es decir, tienen mayor influencia en la predicción. Para esto, dado el clasificador $C(x)$ del cual se desea obtener la explicación visual de la predicción, se obtiene $P = \{p_1, \dots . p_n\}$ donde $p_i = C(r_i)$, como el conjunto de predicciones para el conjunto de regiones obtenido, entonces, el conjunto de regiones útiles se define como:

$$R_u \subset R \mid p_i \in R \wedge p_i > Q_1 + \frac{(Q_3 - Q_1)}{2}, \tag{4}$$

es decir, que el conjunto de regiones útiles se compone de aquellos valores mayores al rango semi-intercuartil, esto debido a que las distribuciones de probabilidad son sesgadas como se muestra en la Fig. 2.

Una vez hecho esto se tiene ya reducido el conjunto R a R_u con las regiones más relacionadas o útiles para el clasificador al momento de realizar la predicción, como se observa en la Fig. 3.

2.2. Caracterización de regiones y visualización

Sea el conjunto de regiones útiles $R_u = \{r_1, \dots, r_m\}$ encontrado, según lo descrito en la sección anterior, habrá que evaluar cada una de estas regiones y clasificarla en una de las tres categorías propuestas según la importancia de está en la toma de decisión del clasificador para la predicción. Estas categorías son *Significativa*, *Relevante y Fútil*, y posteriormente marcarlas visualmente en la imagen.

Dado el clasificador C del cual se desea obtener la explicación visual de la predicción, entonces la categoría de cada región en R_u estará dada por $F(C(r_i))$ donde F es una función escalón definida cómo:

$$F(x) = \begin{cases} Significativa, & Z(x) \geq 2, \\ Relevante, & Z(x) < 2 \wedge x \geq Q_1, \\ Fútil, & x < Q_1, \end{cases}$$

donde $Z(x)$ es el puntaje z.

De acuerdo con la regla práctica del intervalo se puede conocer si un valor es infrecuente o poco común si está a más de dos desviaciones estándar de la media, es decir, un valor realmente *Significativo* para la predicción tiene puntuaciones z mayores que 2.

Ahora bien, para saber que valores son *Fútiles* se requiere otro concepto diferente ya que los valores de las predicciones restantes son de hecho relevantes y no se espera que ninguno tenga un valor $Z(x) < -2$, entonces en este caso aquellas regiones que no son significativas y estén por encima del cuartil Q_1 pertenecerán a la categoría de *Relevante* y las que están por debajo de ese valor, aunque si influyen en la predicción, se pueden considerar *Fútiles*.

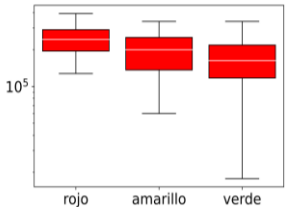

Fig. 5. Comparativo entre las regiones útiles dadas por el algoritmo de explicación propuesto.

Tabla 1. Tabla en donde se detalla el comportamiento del algoritmo de explicación propuesto para 15 ejemplos de imágenes distintas pertenecientes a las clases *perro* y *gato*.

Clase Predicha	RP	RU	AT (px)	ARU(px)	AR (px)	AA (px)	AV (px)	T
Basset	2217	165	306,560	238,920	238,920	60,268	66,196	140.25
Maltese_dog	805	141	242,181	242,181	242,181	199,733	114,114	120.65
Boxer	2281	172	346,977	312,768	312,768	221,001	120,460	238.69
French_bulldog	1010	153	307,200	200,736	200,736	117,912	114,260	170.88
Border_collie	3464	720	404,480	394,368	394,368	293,056	258,300	448.85
MiniSchnauzer	1226	194	307,200	153,120	127,600	112,608	153,120	147.87
English_setter	1145	174	350,080	339,840	338,560	339,840	339,840	216.91
JapaneseSpaniel	2556	1125	332,800	213,720	213,720	213,720	164,400	187.43
Persian_cat	1477	202	307,200	307,200	307,200	307,200	203,010	236.04
Tabby	1826	325	339,200	339,200	324,480	339,200	325,950	224.25
Tabby	2937	201	307,200	291,840	270,408	278,400	233,211	334.58
Tabby	1175	217	272,000	189,975	188,700	189,975	189,550	323.36
Egyptian_Cat	1080	155	307,200	307,200	268,800	206,816	235,520	266.19
Egyptian_Cat	800	55	307,200	143,397	137,572	99,029	92,208	111.13
Tiger_cat	1842	141	272,640	191,268	165,680	184,828	162,756	192.29

Finalmente, se colorean las regiones según su categoría con los colores verde, amarillo y rojo para *Significativa*, *Relevante y Fútil* respectivamente dentro de la imagen original como se muestra en la Fig. 4.

3. Discusión y resultados

Para el análisis de los resultados obtenidos durante el presente trabajo se utilizó un modelo pre-entrenado llamado *InceptionResNet*. Se trata de una red neuronal convolucional (CNN) entrenada con más de un millón de imágenes tomadas de la base de datos *ImageNet*. Al usar *InceptionResNet* se puede asumir que la predicción es correcta y así concentrar el análisis únicamente en la explicación. Como ejemplo del funcionamiento del algoritmo propuesto de explicación, en la Fig. 4, se muestran tres imágenes totalmente diferentes (arriba) y su explicación visual (abajo) para las clases predichas *French_Bulldog*, *Egyptian_cat y Egyptian_cat* respectivamente (clases dadas por el modelo *InceptionResnet*), con la cual se puede observar claramente qué regiones

Fig. 6. Perro de la clase *Bernese_mountain_dog*.

a) Original b) Propuesta c) LIME

Fig. 7. Explicación del algoritmo propuesto para la clase *Labrador_retriever*, comparada contra la explicación dada por LIME.

son consideradas por el clasificador para determinar a qué clase pertenece cada imagen, y se explica además de forma clara que ponderación tiene cada región en la predicción.

En las tres imágenes están claramente resaltadas como regiones útiles aquellas que están más cercanas a lo que nos fijaríamos como seres humanos, ahora bien, en cuanto a la relevancia es muy interesante observar por ejemplo, en la primera imagen de la Fig. 4 (abajo), que el algoritmo propuesto explica que sí bien, todo el cuerpo del perro es relevante para indicar que pertenece a esa clase (pues está marcado con los colores rojo y amarillo), resulta ser que el hocico, los ojos y las orejas son de vital importancia y son el principal factor discriminante para indicar que en efecto es un perro pero del tipo *French_Bulldog*.

Las regiones útiles se traslapan constantemente antes de ser delimitadas y es interesante estudiar el tamaño real o la sumatoria de todas estas regiones ya que demuestran el proceso para llegar al resultado final mostrado por el algoritmo de explicación propuesto. Esta información fue analizada para 200 imágenes.

En la Fig. 5 se puede observar que las regiones *Relevantes* (amarillas) predominan sobre las otras dos regiones *Significativas y Fútiles* (verde y roja) como era de esperarse dado el criterio de caracterización de regiones descrito en la sección 2.2. Sin embargo, se puede observar que las regiones *Fútiles* (rojas) se distribuyen en un área mayor dentro de la imagen a explicar.

En la Tabla 1. se detallan los datos del análisis para 15 imágenes. Donde, RP es el número de regiones propuestas, RU el número de regiones útiles, AT el área total de la imagen, ARU el área total de las regiones útiles, AR el área de las regiones *Fútiles*, AA el área de las regiones *Relevantes*, AV el área de las regiones *Significativas*, y T el tiempo de ejecución del algoritmo de explicación propuesto dado en segundos.

3.1. Explicación multiclase

En la tercera imagen de la Fig. 4 (abajo) se usa el algoritmo de explicabilidad propuesto para resaltar a que le da importancia el modelo para predecir que la imagen pertenece a la clase *Egyptian_cat*, ya que, aunque se visualiza en la imagen también un perro, el gato tiene más relevancia pues es más grande y se muestra todo su cuerpo, como se explica con el algoritmo propuesto.

Sin embargo, aprovechando la complejidad del modelo *InceptionResnet*, y dado que este también reconoce perros y que de hecho existe una probabilidad de predicción para el perro de la imagen perteneciente a la clase *Bernese_mountain_dog*, entonces, se puede explicar de igual manera la predicción de esta clase, como se muestra en la Fig. 6. Con esto se demuestra que pueden explicarse las clases que detecte un modelo dentro de una imagen dada, como en este caso para perro o gato.

3.2. Relevancia de regiones

Es de vital importancia entender que este algoritmo busca dentro de la imagen las regiones de mayor importancia para la predicción realizada sin confundirse con la búsqueda de algún objeto. En el presente trabajo se utiliza el modelo *InceptionResnet* útil para la búsqueda de objetos, sin embargo, el funcionamiento del algoritmo aquí propuesto se adapta a cualquier problema de clasificación de imágenes haciendo visible la explicación de la predicción en cualquier contexto.

En la Fig. 7 se muestra un ejemplo claro de esto, donde aun usando *InceptionResnet*, las regiones marcadas por el algoritmo propuesto Fig. 7 b) muestran claramente las regiones útiles y más aún explica que el modelo se ha fijado en los ojos y nariz del perro para realizar la predicción, que es en realidad lo que se esperaría.

Además, se hace uso de esta misma imagen para hacer una comparación con los resultados obtenidos por LIME Fig. 7 c) que, a pesar de obtener también una buena explicación, pasa por alto la nariz y el resultado tiene claramente una interpretación que depende del punto de vista del observador lo que podría crear desconfianza en la explicación.

3.3. Confiabilidad

Para verificar la eficiencia del algoritmo se compara la región útil obtenida por el algoritmo de explicación propuesto, contra lo que una persona se fijaría para decir que una imagen dada pertenece a una clase predicha, sin embargo, ¿Qué nivel de confianza ofrece el algoritmo para asegurar que funcionará en todos los ejemplos?, pues bien usando imágenes del conjunto de datos COCO que contiene más de 200 mil imágenes con objetos etiquetados y marcados por seres humanos, y que en el presente trabajo se usará como un conjunto de imágenes con cierta pertenencia a una clase (en este caso un objeto dentro de la imagen que un ser humano marcó), es decir, que indicó la región en la que se fijaría para determinar que esa imagen pertenece a la clase en cuestión, se trata de resolver este cuestionamiento.

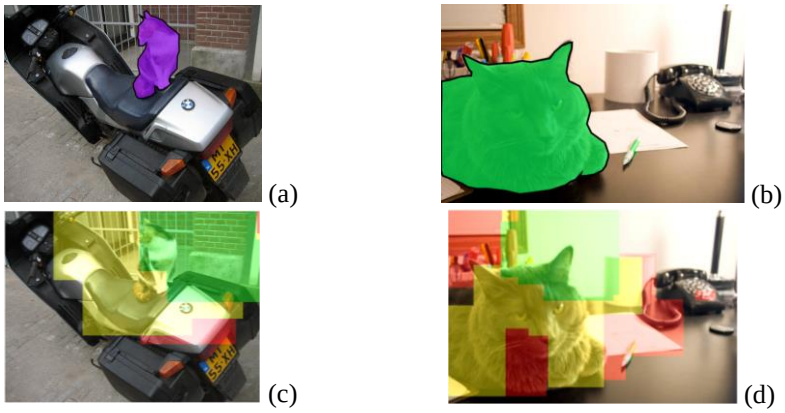

Fig. 8. Ejemplo de comparación entre explicaciones marcadas por un ser humano (arriba) y el algoritmo de explicación visual propuesto (abajo).

En la Fig. 8 a) y b) se muestran dos ejemplos de imágenes del conjunto de datos COCO marcadas y clasificadas por una persona real y se comparan con la explicación del algoritmo aquí propuesto Fig. 8 c) y d). Se puede observar que la región útil marcada por el algoritmo de explicación propuesto rodea efectivamente toda la región marcada por el ser humano, lo que demuestra que funciona adecuadamente, esto siempre y cuando el modelo se encuentre bien entrenado, de lo contrario la utilidad del algoritmo de explicación cambiaría.

Se seleccionaron 200 imágenes diferentes del conjunto de datos COCO con perros y gatos como clase principal, y se realizó una comparación con el algoritmo de explicación propuesto. Para esto se utilizó la técnica de *Intersection Over Union* dada por $IoU = At/A$, donde At es el área de traslape del área marcada en COCO con el área de las regiones útiles del algoritmo de explicación propuesto, y A es el área marcada solamente en COCO.

La explicación visual siempre debería abarcar lo que un humano marcaría si es que el modelo está bien entrenado, como es el caso del modelo *InceptionResnet*. El resultado fue favorable en todos los casos ya que el área marcada en COCO siempre está dentro de la región útil encontrada por el algoritmo de explicación propuesto, demostrando así que es eficaz en encontrar la explicación visual de la región útil que el modelo usa para la predicción, además de mostrarlo en categorías *Significativa, Relevante y Fútil*.

4. Conclusión

El algoritmo de explicabilidad aquí propuesto mostro ser eficiente, además tiene una ventaja primordial, la cuál es el objetivo del presente trabajo, que consiste en disminuir la subjetividad en la explicación, ya que en los algoritmos más populares (LIME, mapas de calor, etc.), según la persona que realice la implementación e interprete el resultado es la conclusión a la que se llegará.

Esta cuestión de subjetividad aquí se ataca por dos frentes, primero no existe ningún parámetro de configuración para el algoritmo lo que asegura que no depende de la persona que lo implemente y siempre se llegara al mismo resultado; segundo, el resultado es claro y fácil de interpretar a la vez que intuitivo dadas las categorías presentadas, de tal forma que cualquier persona que conozca el código de colores será capaz de dar una explicación y esta será muy similar de manera inequívoca a la de cualquier otra persona.

De manera general, se obtuvo un algoritmo funcional que habrá que seguir trabajando, así como aplicarlo a problemas concretos en donde se demuestre su utilidad y su potencial frente a otros algoritmos. También habrá que trabajar más adelante en la reducción del tiempo de ejecución, ya que en promedio para una imagen de 300px cuadrados tarda hasta 2 minutos en obtener su explicación, sin embargo, esto varia y depende de la cantidad de regiones encontradas o de la complejidad de la imagen de entrada, así como de la complejidad del modelo y del área de interés a explicar.

Agradecimientos. Los autores desean agradecer al Instituto Politécnico Nacional por el apoyo económico brindado a través del proyecto SIP 20200638. Tonantzin M. Guerrero agradece al Consejo Nacional de Ciencia y Tecnología por la beca brindada para realizar sus estudios de doctorado.

Referencias

1. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Chatila, R.: Explainable artificial intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. Information Fusion, 58, pp. 82–115 (2020)
2. Lipton, Z.C.: The mythos of model interpretability, Queue, 30(31), pp. 30−57 (2018)
3. Patrick, H.: On Explainable Machine Learning Misconceptions & A More Human-Centered Machine Learning (2019)
4. Tulo-Ribeiro, M.: Why should I trust you? Explaining the predictions of any classifier. arXiv:1602.04938v3 [cs.LG] 9 (2016)
5. Uijlings, J.R., Van De-Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. International Journal of Computer Vision, 104(2), pp. 154–171 (2013)
6. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. International Journal of Computer Vision, 59(2), pp.167–181 (2004)
7. Ghorbani1, A., Zou, J.Y.: What is your data worth? Equitable Valuation of Data (2019)
8. Robnik-Sikonja, M.: Explanation of prediction models with explain prediction. Informatica. 42, pp. 13–22 (2018)
9. Szegedy, C., Yangqing-Jia, W.L., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going Deeper with Convolutions (2014)
10. Botchkarev, A.: Performance metrics (Error Measures) in machine learning regression. Forecasting and Prognostics: Properties and Typology (2018)
11. Petsiuk, V., Das, A., Saenko, K.: RISE: Randomized Input Sampling for Explanation of Black-box Models (2018)
12. Papernot, N., McDaniel, P.: Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning (2018)

13. Hall, P.: On Explainable Machine Learning Misconceptions & A More Human-Centered Machine Learning (2019)
14. Doran, D., Schulz, S., Besold, T.R.: What Does Explainable AI Really Mean? A New Conceptualization of Perspectives (2017)
15. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J.: Building Machines That Learn and Think Like People (2016)
16. Tulio-Ribeiro, M., Singh, S., Guestrin, C.: Model-Agnostic Interpretability of Machine Learning (2016)
17. Wang, C., Xi, Y.: Convolutional neural network for image classification. Johns Hopkins University Baltimore, MD (2018)
18. Robnik-Sikonja, M.: Explanation of prediction models with explain prediction (2018)
19. Krause, J.: Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models (2016)
20. van Lent, M.: An Explainable Artificial Intelligence System for Small-unit Tactical Behavior (2004)
21. Gunning, D.: Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency (DARPA) (2017)