Generación y etiquetado de un dataset de imágenes satelitales.

José Ramón Romero Chávez

Abstract - El reciente interés y desarrollo de herramientas para inteligencia artificial y procesamiento de imágenes (OpenCV, TensorFlow, PyTorch, entre otros), ha motivado la inclusión de este tipo de tecnologías en infinidad de campos de acción.

Sin embargo, una dificultad a la que se enfrentan estos proyectos es la escasez de información con características adecuadas para ser utilizados en los algoritmos más usuales de Machine Learning.

Durante el desarrollo de esta solución, exploramos el desarrollo de un algoritmo para la automatización de recopilación y etiquetado de imágenes a través de web scraping, particularmente de imágenes satelitales.

I . Antecedentes

La búsqueda de imágenes es clave fundamental en cualquier motor de búsqueda (Google, Bing, Yahoo, etc), y nadie a ciencia cierta conoce la cantidad exacta de imágenes en la web [1], ya que conforme el tiempo transcurre, nuevo contenido (no sólo imágenes) es publicado cada segundo [2].

La gran cantidad de información existente ha motivado nuevas e interesantes aplicaciones en infinidad de campos [3].

Zhang [2] nos ha mostrado como ha sido el proceso de evolución de la busqueda de imagenes por internet, donde en los años 1990 la recuperación de imágenes se encontraba completamente basada en textos gracias a un manejador de bases de datos relacionales (RDBMS), posteriormente la extracción de información e indexado permitió a empresas lanzar sus motores de búsqueda con acceso a millones de resultados a través de Internet.
Es por ello que, un aspecto muy valioso a considerar es que dentro de la web, la información siempre viene con metadata cómo: URL, nombres de archivo y demás información circundante a ella que es de utilidad para su proceso de etiquetado y búsqueda [4] y puede ser usado para la creación de datasets para algoritmos de inteligencia artificial, particularmente en el diseño e implementación de modelos de aprendizaje supervisado [5][6].

Es por ello que se han liderado gran cantidad de esfuerzos para la creación de algoritmos de generación de datasets haciendo uso de la información existente en la web.

Zhang,[3] desarrolló un algoritmo de generación y etiquetado de personas, haciendo uso de una imagen input, se identificaron similitudes en textos circundantes alrededor de la imagen, dando como resultado el nombre de la persona (label)

Fisher [4] propuso un proceso para la adquisición de dataset etiquetado (labeled) de imágenes haciendo uso de deep learning y algunos humanos en procesos intermedios.

J. Zhang [5] introdujo el DIRS,un método para aplicar la recuperación de imágenes basada en contenido a conjuntos de datos de imágenes masivas. DIRS se implementó en Hadoop utilizando el modelo de computación MapReduce y las imágenes y sus características se almacenaron en bases de datos.

Ahora bien, más cercano al problema de las imágenes satelitales, Gao [6] enfocó su esfuerzo en la creación de “gazetteers” con base a análisis geográficos.

Los “Gazetteers” contienen información con respecto al área analizada, sin embargo la diferencia propuesta fue el uso de un sistema basado en Hadoop para la recolección de data sobre los lugares.

II. Problemática

Los motores de búsqueda de imágenes se han vuelto herramientas indispensables para todo tipo de usuarios y objetivos alrededor del mundo desde fotografía, artes, conocimiento y mucho más.

Junto con ellos, existen varios datasets de libre acceso como Google Image Data Sets[7], Yahoo’s Datasets[8], ImageNet [9], Kaggle [10], entre muchos más.

Sin embargo, como se ha descrito muchas veces [7], la utilidad de un modelo está directamente relacionada con la calidad y cantidad de la información con la que se alimentan los modelos, en especial al ejecutar algoritmos de aprendizaje supervisado (labelling).

III. Justificación

En la época de la información, es difícil concebir la idea de que por falta de información no se puedan desarrollar las ideas. Es por ello que establecer y evaluar una metodología para generación automatizada de dataset, es fundamental para el seguimiento de estos esfuerzos.

IV. Propuesta

Un algoritmo , basado en web scraping, que permita la generación de un dataset de imágenes satélitales representativas de áreas rurales como urbanas, adicionando etiquetado (labeling) de detalles a considerar con respecto a cada una de ellas (latitud, longitud, hora, etc.).

V. Contribución Esperada

Adicional al producto propuestos, se buscará aportar a la literatura existente con respecto a las técnicas utilizadas y los aspectos a considerar para cada una de ellas.

Una mirada objetiva al uso e implicaciones de web scraping para su uso adecuado

VI. Referencias

[1] Xin-Jing Wang, Lei Zhang, and Wei-Ying Ma. "Duplicate-search- based image annotation using web-scale data." Proceedings of the IEEE 100.9 (2012): 2705-2721.

[2] Lei Zhang, and Yong Rui. "Image search—from thousands to billions in 20 years." ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 9.1s (2013)

[3] D. Shapiro, "Can Artificial Intelligence Generate Corporate Strategy?", Forbes.com, 2019. [Online]. Available: https://www.forbes.com/sites/danielshapiro1/2019/08/19/can-artificial-intelligence-generate-corporate-strategy/#ffcceb9559fc. [Accessed: 23- Aug- 2019].

[4]Fisher, Yu, "LSUN: Construction of a Large-scale ImageDataset using Deep Learning with Humans in the Loop."

[5] J. Zhang, X. Liu, J. Luo, and Bo Lang. "Dirs: Distributed image retrieval system based on mapreduce." Pervasive Computing and Applications (ICPCA), 2010 5th International Conference on. IEEE, 2010.

[6] S. Gao, L. Li, W. Li, K. Janowicz, and Y. Zhang. "Constructing gazetteers from volunteered big geo-data based on Hadoop." Computers, Environment and Urban Systems (2014).

[7] J. Z. H et al., "A Survey on Cleaning Dirty Data Using Machine Learning Paradigm for Big Data Analytics", Indonesian Journal of Electrical Engineering and Computer Science, vol. 10, no. 3, p. 1234, 2018. Available: 10.11591/ijeecs.v10.i3.pp1234-1243 [Accessed 24 August 2019].

[8] A. Ali, R. Ali, A. Khatak and M. Aslam, "Large Scale Image Dataset Construction Using Distributed Crawling with Hadoop YARN", 2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS), 2018. Available: 10.1109/scis-isis.2018.00075 [Accessed 24 August 2019].