EIEI ACOFI 2023
INGENIERÍA PARA TRANSFORMAR TERRITORIOS
SEPTIEMBRE 19 al 22 2023
CARTAGENA DE INDIAS, COLOMBIA
Encuentro Internacional de Educación en Ingeniería ACOFI

Estado de arte y aplicación de técnicas de Aprendizaje Profundo desde imágenes reconstruidas a partir de nubes de puntos de sensores de Radar y Lidar

Óscar Montañez Sogamoso, Carolina Roa Martín, Eduardo Avendaño Fernández

Universidad Pedagógica y Tecnológica de Colombia
Sogamoso, Colombia

Resumen

Este artículo presenta una revisión de estado de arte de las técnicas del aprendizaje de máquina para el reconocimiento de patrones, la definición de los algoritmos para la clasificación de objetos (personas) a partir de imágenes construidas desde nubes de puntos adquiridas por sensores de Radar y Lidar, y resultados preliminares de aplicación en escenas reconstruidas a partir de nubes de puntos. En el primer hito del proyecto CLARIFIER (frequenCy-agiLe rAdar-lidaR chIp For survelllancE moving platfoRms), se ha estudiado e implementado un filtro de Kalman Extendido para fusión de datos de sensores de Radar y LiDAR, cuyo aporte fue la inclusión de la velocidad angular en el modelo cinemático de un drone. En el segundo hito, se ha revisado estado de arte para identificar enfoque (***segmentación semántica***) y algoritmos que permitan detectar personas en imágenes reconstruidas de escenarios o zonas con requerimientos de supervisión y monitoreo. Como resultado de aplicación preliminar, se ha construido un conjunto de datos de escenas que incluye la clase "*personas*" y se evaluaron los algoritmos U-Net y Mask R-CNN que aplicas técnicas de segmentación semántica. Dado que la resolución del sensor en particular de Lidar, es mayor a la del radar, así como la diferencia en rangos que cada uno alcanza, se genera un compromiso que requiere múltiples barridos de la escena desde diferentes ángulos y distancias y se requiere aumentar el conjunto de datos, para entrenar los algoritmos y mejorar el porcentaje de detecciones correctas. De acuerdo a las métricas que se han evaluado (pérdidas, exactitud, precisión, exhaustividad y F1 Score), la clasificación de la clase personas en el contexto de segmentación semántica alcanza un 89.91 % para Mask R-CNN, y del 90.53% para U-Net; y con la curva de operación característica del receptor (ROC) y el área bajo la curva (AUC) se obtiene un 90% y 92% en la detección de la clase personas, respectivamente. Este resultado valida la efectividad de estos

DOI: https://doi.org/10.26507/paper.3038

modelos de redes neuronales convolucionales aplicada a imágenes obtenidas a partir de nubes de puntos.

Palabras clave: U-Net; Mask R-CNN; CNN; segmentación semántica

Abstract

This paper presents a state-of-the-art review of machine learning techniques for pattern recognition, the definition of algorithms for the object (people) classification from images constructed from point clouds acquired by Radar and Lidar sensors, and preliminary results of application on scenes reconstructed from point clouds. In the first milestone of the CLARIFIER project (frequenCy-agiLe rAdar-lidaR chIp For surveIllancE moving platfoRms), an Extended Kalman filter for Radar and LiDAR sensor data fusion has been studied and implemented, whose contribution was the inclusion of angular velocity in the kinematic model of a drone. In the second milestone, the state of the art has been reviewed to identify approaches (semantic segmentation) and algorithms to detect people in reconstructed images of scenarios or areas with supervision and monitoring requirements. As a result of preliminary application, a scene dataset including the class "people" has been constructed and U-Net and Mask R-CNN algorithms applying semantic segmentation techniques were evaluated. Given that the resolution of the Lidar sensor in particular, is higher than that of the radar, as well as the difference in ranges that each one reaches, a compromise is generated that requires multiple sweeps of the scene from different angles and distances and it is required to increase the dataset, to train the algorithms and improve the percentage of correct detections. According to the metrics that have been evaluated (loss, accuracy, precision, recall, and F1 Score), the classification of the class people in the context of Semantic Segmentation reaches 89.91 % for Mask R-CNN, and 90.53% for U-Net; and with the receiver operating characteristic (ROC) curve and the area under the curve (AUC) 90% and 92% are obtained in the detection of the class people, respectively. This result validates the effectiveness of these convolutional neural network models applied to images obtained from point clouds.

***Keywords**: U-Net; Mask R-CNN; CNN; semantic segmentation*

1. Introducción

El aprendizaje profundo (Deep Learning) es un subcampo del aprendizaje de máquina (Machine learning) que se enfoca en el diseño y aplicación de algoritmos y modelos de redes neuronales artificiales con múltiples capas bajo la sombrilla de la Inteligencia Artificial. Estos modelos están diseñados para aprender representaciones y características de alto nivel a partir de datos en bruto. La Introducción de los conceptos del aprendizaje profundo se atribuye a (Hinton, G. E., 2006), que consolida los fundamentos de las ***redes de creencia profunda***. Este artículo presentó una técnica clave para entrenar modelos de aprendizaje profundo llamada "*aprendizaje profundo por preentrenamiento",* que es un enfoque para entrenar redes neuronales profundas. Una red neuronal convolucional (Convolutional Neural Network - CNN) es un tipo de modelo de aprendizaje profundo que ha demostrado ser muy efectivo en tareas de visión artificial por computadora, en

reconocimiento de imágenes y detección de objetos. La idea fundamental detrás de las CNN es que pueden aprender automáticamente características relevantes y discriminativas de las imágenes al aplicar filtros convolucionales y capas de votación (pooling) en cascada. Estas redes son capaces de capturar patrones espaciales en los datos de entrada, lo que las hace adecuadas para procesar imágenes y datos bidimensionales. La arquitectura de las CNN fue introducida por Yann LeCun, Yoshua Bengio, y otros en la década de los años 90. Presentaron una arquitectura de red neuronal convolucional llamada LeNet-5 (Y. Lecun, 1998), que fue utilizada para el reconocimiento de dígitos escritos a mano en cheques bancarios. La arquitectura de transfer learning (aprendizaje por transferencia) (Pan, S. J. et al., 2010) es una técnica en el campo del aprendizaje automático que aprovecha el conocimiento aprendido de un modelo entrenado en un dominio fuente para mejorar el rendimiento en un dominio objetivo relacionado pero distinto. No se trata de una arquitectura de red neuronal específica, sino más bien de un enfoque general que se puede aplicar a diferentes arquitecturas de redes neuronales. El transfer learning fue fundamentado por (Krizhevsky, A. et al.,, 2012). En este artículo, los autores presentan una arquitectura de red neuronal convolucional profunda, conocida como AlexNet, y demuestran su eficacia en la tarea de clasificación de imágenes utilizando el conjunto de datos ImageNet. La técnica cobra mayor relevancia dado que el conocimiento aprendido de una tarea de clasificación de imágenes en gran escala puede transferirse a tareas relacionadas y obtener mejoras significativas en el rendimiento. La arquitectura es efectiva para abordar problemas de falta de datos y mejorar el rendimiento de los modelos en una variedad de dominios y tareas, como la clasificación de imágenes, el procesamiento del lenguaje natural y el reconocimiento de voz. La arquitectura de redes neuronales convolucionales (CNN, por sus siglas en inglés) ha demostrado ser efectiva en tareas de procesamiento de imágenes. Sin embargo, aplicar estas técnicas a datos tridimensionales, como nubes de puntos o imágenes en 3D, presenta desafíos importantes. En este contexto, han surgido varias arquitecturas especializadas que abordan estas dificultades y logran resultados destacados. A continuación, se presenta la evolución de algunas de las arquitecturas más relevantes para analizar y definir cuáles podrían ser evaluadas en la ejecución del proyecto Clarifier. Maturana y Scherer presentaron VoxNet como una arquitectura eficiente para el reconocimiento de objetos en tiempo real en escenarios de robótica. En (Maturana D., 2015) Describen la estructura y los componentes clave de VoxNet, incluyendo capas convolucionales 3D y capas completamente conectadas, que permiten el procesamiento de datos tridimensionales y la extracción de características relevantes para el reconocimiento de objetos. La arquitectura de redes neuronales convolucionales multivista (Multiview Convolutional Neural Networks - MVCNN) fue introducida por (Su, H. et al., 2016). MVCNN es una red neuronal convolucional diseñada específicamente para el reconocimiento de objetos o formas 3D en imágenes desde múltiples vistas o ángulos a partir de múltiples vistas 2D. Utiliza una CNN para extraer características discriminativas de cada vista individual que se combinan y se utiliza en una capa completamente conectada para la clasificación final. La arquitectura MVCNN ha demostrado ser efectiva en tareas de reconocimiento de objetos 3D y ha sido utilizada en aplicaciones como la detección de objetos en escenas 3D (Su H. et al., 2015), la clasificación de formas tridimensionales y la recuperación de objetos tridimensionales. Uno de los hitos importantes en este campo fue la introducción de la arquitectura de red convolucional completa (Fully Convolutional Network - FCN) por (Long et al., 2015). La segmentación semántica (Zhou B. et al., 2017), (Chen L. et al., 2018), (Li Y. et al., 2016) y en específico en aplicaciones con sensores LiDAR permite asignar etiquetas semánticas a cada pixel de una imagen (creada a partir de los puntos en nubes de puntos generadas por sensores LiDAR) indicando a qué clase o categoría pertenece. FCN se basa en CNNs

adaptadas específicamente para abordar la tarea de segmentación semántica en datos LiDAR. La principal contribución de FCN es permitir la generación de mapas de segmentación a nivel de píxel mediante el uso de capas completamente convolucionales y el reemplazo de las capas completamente conectadas en las CNN tradicionales. Esto permite que la arquitectura FCN aprenda y capture características contextuales y espaciales en diferentes escalas para la segmentación semántica precisa. Para la aplicación el elemento clave fue la transformación de las nubes de puntos en imágenes para habilitar los modelos de CNN enfocados a segmentación semántica (Zhou, X. y Chen S., 2018). En el mismo año, la arquitectura U-Net se presenta como una red neuronal convolucional (CNN) que se utiliza principalmente para tareas de segmentación de imágenes. En (Ronneberger, O., 2015) los autores presentan la arquitectura U-Net como una solución para la segmentación precisa de imágenes biomédicas. La arquitectura se inspira en el concepto de codificadores-decodificadores, donde la información se comprime en una etapa de codificación y luego se reconstruye en una etapa de decodificación. La característica distintiva de U-Net es la incorporación de conexiones de salto o residuales, que permiten que la información de características de las capas de codificación se transmita directamente a las capas correspondientes de decodificación. Esto ayuda a preservar la información de alta resolución y a mejorar la precisión en la segmentación. La arquitectura U-Net ha demostrado ser especialmente eficaz en aplicaciones de segmentación de imágenes biomédicas, como la segmentación de células, órganos y tejidos en imágenes de resonancia magnética y tomografía computarizada. Sin embargo, también se ha utilizado en otras aplicaciones de segmentación en imágenes generales, como la segmentación de objetos en imágenes naturales.

Por otra parte, la arquitectura Faster R-CNN es una CNN diseñada para la detección de objetos en imágenes. Fue introducida por (Ren, S., 2015). Aborda el desafío de realizar detección de objetos eficiente en tiempo real. La arquitectura consta de dos módulos principales: un módulo de propuesta de regiones y un módulo de clasificación y regresión. El módulo de propuesta de regiones utiliza una red neuronal llamada Región de Propuesta de Redes (Region Proposal Network - RPN) para generar candidatos potenciales de regiones de objetos en la imagen. Estas regiones propuestas se utilizan como posibles ubicaciones de objetos y se envían al módulo de clasificación y regresión, acá se clasifica cada región propuesta en categorías específicas de objetos y de ajustar sus coordenadas para obtener una detección precisa. Este módulo utiliza una combinación de características extraídas de la imagen original y de las regiones propuestas generadas por el RPN. La arquitectura ha demostrado un rendimiento efectivo en términos de precisión y velocidad de detección y ha sido utilizado en diversas aplicaciones, como sistemas de vigilancia, análisis de imágenes médicas y detección de objetos en tiempo real. En el mismo contexto, la arquitectura de ResNet (Redes Residuales), redes neuronales profundas fue introducidas por (He. K. et al., 2016). Se presenta como una solución para abordar el problema del decaimiento del rendimiento a medida que aumenta la profundidad de las redes neuronales convolucionales. En lugar de tratar de aprender directamente las representaciones deseadas en capas más profundas, los bloques residuales de ResNet permiten que las capas de la red se adapten y aprendan las diferencias o residuos entre la entrada y la salida de cada bloque. Estos residuos se agregan de nuevo a la salida, permitiendo que la información fluya a través de las capas sin restricciones. ResNet introdujo el concepto de saltos de conexión o conexiones residuales, que permiten que la información y los gradientes se propaguen directamente de capas anteriores a capas posteriores. Esto mitiga el

problema del desvanecimiento del gradiente y permite el entrenamiento de redes neuronales más profundas con un rendimiento mejorado.

La arquitectura Mask R-CNN fue presentada por (He, K. et., al, 2017), (Shin H.C. et al., 2016). Es una extensión de la red neuronal convolucional (CNN) Faster R-CNN, diseñada para abordar la tarea de detección y segmentación de instancias en imágenes. Combina las capacidades de detección de objetos de Faster R-CNN con la capacidad de segmentación precisa de instancias. Mask R-CNN utiliza una etapa adicional en la red que genera máscaras de segmentación para cada objeto detectado, permitiendo la segmentación de píxeles a nivel de instancia. La arquitectura Mask R-CNN consiste en una estructura de tres etapas: la etapa de extracción de características, la etapa de propuesta de regiones y la etapa de clasificación, regresión y segmentación. La etapa de segmentación utiliza una subred convolucional adicional para generar máscaras de segmentación precisas para cada objeto detectado. Esto permite una detección y segmentación precisas de múltiples instancias en una imagen. La arquitectura Mask R-CNN ha tenido un gran impacto en el campo de la detección y segmentación de instancias en imágenes y ha sido ampliamente adoptada en diversas aplicaciones, como la detección de objetos en tiempo real, la segmentación de objetos en imágenes médicas y la segmentación de objetos en imágenes naturales. Ha establecido un nuevo estándar para la precisión en la detección y segmentación a nivel de instancia y ha sido utilizado como base para muchas investigaciones y desarrollos posteriores.

Para los escenarios donde se trabaja nubes de puntos desde sensores de LiDAR y Radar, se presentan arquitecturas como PointNet que fue introducida por (Qi C.R. et al., 2017), (Qi C.R. et al., 2018). Esta red neuronal fue diseñada para procesar directamente nubes de puntos tridimensionales (3D), sin la necesidad de convertirlas en mallas o volúmenes. Ha demostrado ser efectiva en tareas de clasificación, segmentación y reconocimiento de objetos en datos 3D. Se basa en transformaciones y operaciones de agregación que operan en cada punto individualmente y luego agregan información globalmente. Esto permite que la red sea invariante a permutaciones en el orden de los puntos y sea capaz de capturar características importantes (patrones) en el contexto de la nube de puntos (Zhang, Y., & Jiao, J. 2020). Ha sido adoptada en diversas aplicaciones, como la robótica, la realidad aumentada y la visión artificial por computadora. Así mismo, VoxelNet es una arquitectura de red neuronal diseñada específicamente para abordar la detección de objetos en entornos 3D utilizando datos LiDAR. Fue introducida por (Yan, M., 2018). Es una arquitectura end-to-end para la detección de objetos en datos de nubes de puntos LiDAR. Divide el espacio 3D en voxels (volumétricos), que son celdas tridimensionales discretas. Cada voxel contiene información sobre los puntos de la nube que caen dentro de él. Utiliza redes neuronales convolucionales en 3D para extraer características de los voxels y luego fusiona la información a través de capas totalmente conectadas para realizar la detección de objetos. También incorpora una etapa de propuesta de regiones para generar candidatos de ubicaciones de objetos en la nube de puntos. Como otra alternativa, en cuanto a regiones de interés, Milioto *et al.*, complementan el estado del arte de la segmentación semántica desde sensores LiDAR para proporcionar una fuente independiente de información semántica en el contexto de la conducción autónoma (Chen, X., 2016). Introducen la arquitectura RangeNet (Milioto et al., 2019) usando imágenes de rango como representación intermedia en combinación con una red neuronal convolucional (CNN) que explota el modelo de sensor LiDAR giratorio para realizar con precisión segmentación semántica completa de nubes de puntos a la velocidad de fotogramas del sensor. Además, para distribución de nube

de puntos desde sensores de Lidar usados en vehículos varía continuamente con el aumento de la profundidad, y esto no puede ser bien modelado por un único modelo. Hongwei Yi *et al.*, proponen un modelo unificado SegVoxelNet (Yi, S. et al., 2020) que aplica un codificador de contexto semántico para aprovechar las máscaras de segmentación semántica en vista panorámica. Las regiones sospechosas podrían ser resaltadas mientras que las regiones ruidosas podrían ser suprimidas. Para tratar mejor los vehículos a distintas profundidades se diseña un nuevo cabezal que tiene en cuenta la profundidad para modelar explícitamente las diferencias de distribución de los vehículos. En experimentos con el conjunto de datos KITTI se muestra que el método supera otras técnicas incluso más avanzadas en cuanto a precisión y eficacia con la nube de puntos como única entrada.

2. Metodología Aplicada para detección de personas

El Proyecto Clarifier se enmarca en el programa Ciencia para la Paz (Science for Peace) y busca simplificar tareas de supervisión y vigilancia en zonas de seguridad. A partir de imágenes reconstruidas desde nubes de puntos adquiridas por sensores de radar y lidar se construyó un conjunto de datos para poder entrenar los modelos y validar su desempeño. En una primera aproximación se define escenas capturadas un sensor de radar y uno de LiDAR que busca mejorar las bondades de los sensores en los dominios de radio frecuencia (Radar) y fotónico (LiDAR). Para el caso y el entrenamiento de los modelos, se definió la clase personas como el objetivo a detectar. De la revisión de estado de arte se identifica que la técnica que pueda tener un desempeño adecuado en cuanto a costo computacional y eficacia en la detección de la clase de interés es la segmentación semántica, y dos modelos que la implementan son Mask R-CNN y U-net. Para el banco de prueba se usó un sensor RPLiDAR Salmtech que tiene un rango de 0.15 a 12 m y que genera 941 puntos para cada ángulo de rotación en un giro de 360°, y el sensor de Radar usado fue un circuito integrado a ondas milimétricas (MMIC) de silicio germanio de la familia BGT24MTR12 de Infineon, que tiene un rango de detección entre 0.5 y 20 m.

La implementación se desarrolló bajo la metodología de operaciones de aprendizaje de máquina (MLOps) (Kreuzberger, D., 2022), donde los distintos pasos del proceso se describen en el flujo de trabajo, que incluye el *i*) Diseño, *ii*) el Modelo Desarrollado, y *iii*) las Operaciones. En la ***etapa de diseño*** se aplican las diferentes técnicas de preprocesamiento de imagen, se construyó un conjunto de datos que consistió en 249 imágenes en escenas desde diferentes vistas. El objetivo fue detectar personas y las etiquetas en las imágenes que contienen esta clase fueron de 233 para un 77.92% con una persona, 15.66% dos personas, y escenas vacías (sin personas) un 6.42%. Entonces, luego del preprocesamiento, con el conjunto de datos de personas se etiqueta los datos, se aplica segmentación y se suaviza la imagen para aplicar un entrenador de clasificación de imágenes. En la ***etapa de desarrollo del modelo***, se identifica las bibliotecas para la detección de objetos (Clase personas) que fueron TensorFlow, Keras y Pytorch, y el modelo aplicado fue Resnet (He K., et al., 2016) alojado por un modelo de CCN y COCO, que descompone en la imagen en los colores de luz primaria rojo, verde y azul (RGB) y aplica una máscara binaria asociado al umbral de intensidad del color. Los dos modelos que implementan la segmentación semántica fueron Mask R-CNN y U-Net. En la arquitectura Mask R-CNN, las tareas de segmentación y clasificación son realizadas simultáneamente. La red genera las regiones propuestas de interés (ROIs) sobre la imagen del conjunto de datos de entrada, y luego asigna una etiqueta de

clase a cada pixel dentro de las regiones propuestas en cada paso. Las máscaras de segmentación se obtienen como una salida adicional desde la red junto con las etiquetas de la clase. Esas máscaras binarias, son matrices binarias que indican la probabilidad de que cada pixel pertenezca a la clase objetivo o no (la clase persona). Para U-Net, las actividades de segmentación y clasificación se hacen secuencialmente en etapas separadas. Se tiene dos etapas, una de codificación que reduce la resolución espacial de la imagen de entrada, y una etapa de decodificación que reconstruye la imagen segmentada desde la representación codificada. Finalmente, en la ***etapa de operaciones***, se aplica métricas de validación para determinar el desempeño de las dos máscaras de redes (Mask R-CNN y U-Net), en ese contexto se evaluó 5 métricas que serán presentadas y analizadas adelante. Acá también, se aplica las curvas de la región de interés y de curva lift que retorna la probabilidad para evaluar el desempeño del modelo, y que permite identificar falsos positivos o negativos del conjunto de datos. Con estas métricas se valida los modelos y se cuenta con los modelos entrenados para la detección de la clase persona.

3. Selección de la técnica a partir de Análisis Textual

La información contenida en los resúmenes de los artículos analizados, es categorizada de acuerdo al año de publicación del artículo, posteriormente es procesada en software Iramuteq 0.7. Se realiza un Análisis Factorial de Correspondencias (AFC) que permite representar visualmente las relaciones entre las categorías de variables en un análisis de datos textuales. De acuerdo a su proximidad gráfica se considera su similitud o diferencia discursiva, adicionalmente el tamaño de las formas activas analizadas (palabras) permiten identificar la importancia en el discurso relativo a la categoría. De acuerdo con la **Figura 1**, se encuentra que los valores de la variable se dividen en "4 clases" o categorías, teniendo una distribución porcentual del 35,2% el grupo de los años 2018, 2022 y 2023 (rojo) presentando similitud en el discurso, donde palabras como "feature (41), network (39), module (26), local (18), convolutional (18), neural (20)" entre otras son significativas en lo estudiado en estos años. La segunda categoría, corresponde al 25.5%, donde se concentran los datos obtenidos de estudios publicados en 2019 (verde), las palabras representativas en su frecuencia corresponden a "show (31), learn (27), performance (25), method (24), state (24), art (24)", entre otras, este grupo es cercano discursivamente con la categoría de los artículos publicados en 2015 y 2021 que tiene una proporción del 19.7%, las palabras relacionadas a esta categoría con mayor frecuencia son "cloud (28), large (25), method (20), datum (20) object (18), urban (15), scale y dataset (14)".

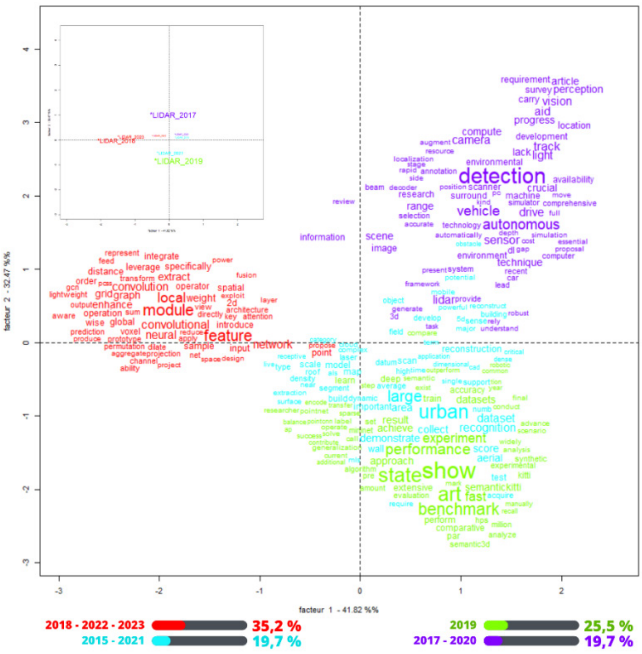

Figura 1. Análisis Factorial de correspondencias – Resúmenes y estado de arte temporal

En cuanto a la categoría compuesta por los artículos publicados entre 2017 y 2020, corresponden al 19.7% restante, allí las palabras con mayor frecuencia son “detection (23), object (16), autonomus (16), vehicle (16), 3d (29), lidar (29)”, entre otras.Respecto a las formas compartidas por las categorías se encuentran las plabras “point (259), 3D (175), cloud (157), segmentation (142), lidar (134), feature (114), method (113), datum (101), learn (91)”, entre otras. Este último grupo se encuentra distribuido en todas las categorías, y para comprender este contexto de todas las categorías analizadas, se genera un análisis de similitud (**Figura 2**) con las 50 palabras más frecuentes en el discurso analizado entre 2017-2023, frecuencias entre 27 a 259 repeticiones.

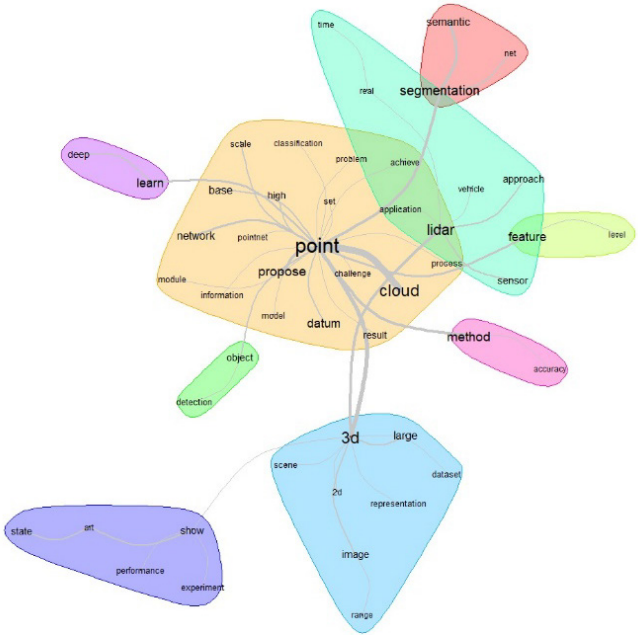

Figura 2. Análisis de Similitud – Estado de Arte: Se define Segmentación Semántica

En la **Figura 2**, se identifica grupos de unidades de texto que están altamente relacionadas entre sí (comunidades), cada una de las comunidades se agrupan en un halo de color, que permite identificar aquellas que por su cercanía pueden tener mayor relación discursiva. Alrededor de las palabras "cloud point" se concentra parte del discurso, teniendo alta relación por su cercanía con la "segmentation (142)" "semantic (89)", siendo coherente con la técnica de Segmentación Semántica, para la selección de los modelos de aprendizaje de máquina U-Net y Mask R-CNN.

4. Análisis de Resultados

Comparativamente se obtiene para la métrica de pérdidas (Loss) un valor de 0.0042, en cuanto a ***exactitud 89.81%,*** precisión 85.18%, la tasa de positivos verdaderos (recall) fue del 79.95% y el puntaje F1 fue del 81.89% para Mask R-CNN vs. 0.0045 de pérdidas, ***90.53% en exactitud***, 83.70% de precisión, 81.17% de tasa de positivos verdaderos, y puntaje F1 del 82.41% para U-Net. Sin embargo, el desempeño de la arquitectura U-Net exhibió un resultado marginal superior de 0.72% sobre Mask R-CNN. En general, los resultados indican que los dos modelos se ajustan adecuadamente a los datos de entrenamiento (20% del conjunto de datos) y que existe una pequeña diferencia en la capacidad de ajuste entre ellos.

Conceptualmente es importante entender lo que significan las épocas en el entrenamiento del modelo en los conjuntos de datos. Como referencia para una época, el modelo recibe una serie de casos de entrenamiento y ajusta sus parámetros (como los pesos de las conexiones en una red neuronal) en función de los errores cometidos en la predicción de las respuestas correctas. Una vez que "*todos*" los casos de entrenamiento han sido vistos por el modelo, se completa una época y se repite el proceso de entrenamiento para tantas épocas como sea necesario de forma que se pueda mejorar la precisión del modelo. En el caso de aplicación, se utilizó un tamaño de lote con 32 imágenes de muestra y un número inicial de 50 de épocas. Sin embargo, los resultados de la capa de segmentación no fueron óptimos, por lo que se realizó validaciones adicionales con 500 y 5000 épocas. La **Figura 3**, muestra los resultados de estas validaciones: con 50 épocas, la segmentación es limitada y no permite una visualización adecuada del objeto de interés; al incrementar a 500 épocas, la imagen segmentada, aunque mejora, aún no permite definir la detección del objeto; y finalmente, con un incremento a 5000 épocas, se obtiene una imagen claramente segmentada que permite la detección del objeto (clase persona) en la escena. La detección mejora significativamente con el aumento del número de épocas utilizadas en el entrenamiento del modelo, y que se requiere un número considerable de épocas para un correcto entrenamiento.

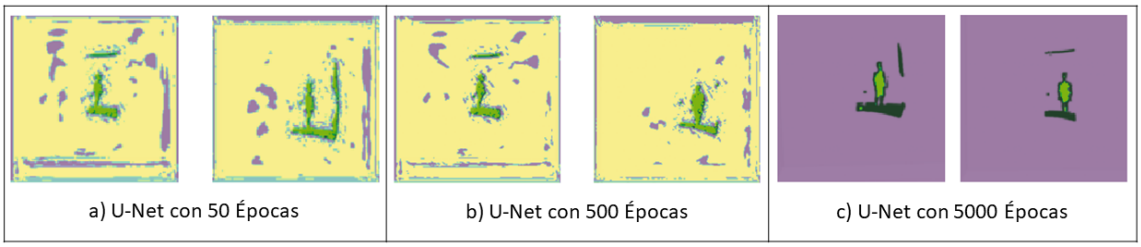

Figura 3. Incremento de épocas y desempeño en la resolución para la arquitectura U-Net.

De igual forma se evaluó para la arquitectura Mask R-CNN y se identifica un mejor desempeño en el puntaje obtenido para verdaderos positivos, que para las escenas y con otros objetos, permite una detección de la clase personas entre en 99.6% y 100%. De igual forma, se validó con las curvas de la región de interés y el área bajo la curva, donde un valor de 0.90 se obtiene para la arquitectura Mask R-CNN y de 0.92 para U-Net. Al igual que la variación marginal con la exactitud, para el caso se identifica que aunque la red U-Net se desempeña mejor que Mask R-CNN en términos de la habilidad para discriminar entre clases positivas y negativas, una diferencia de 0.02 no es significativa en términos prácticos y se requiere eventualmente ampliar o aumentar el modelo para poder exigir mejor los entrenamientos y por tanto, optimizar el modelo para el reconocimiento, incluso con un conjunto de clases aumentado para evaluar el desempeño.

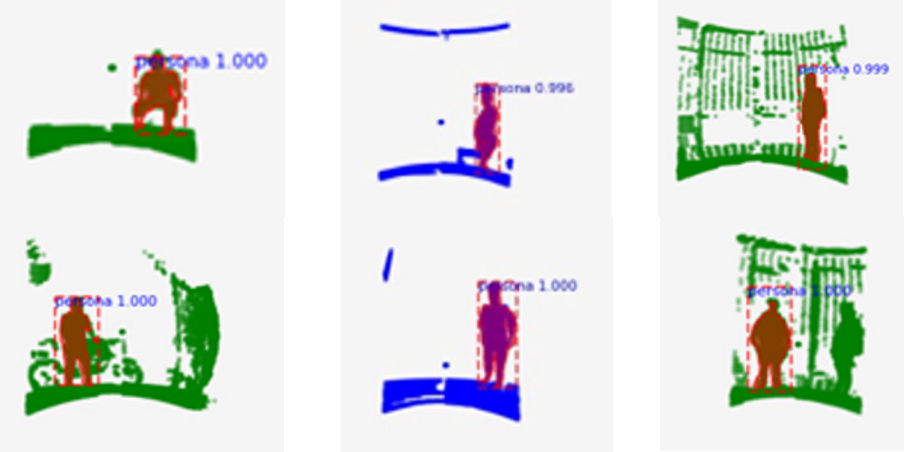

Figura 4. Desempeño de la arquitectura Mask R-CNN

5. Conclusiones

El artículo sintetiza una conceptualización concreta y una evolución de las arquitecturas de aprendizaje de máquina y permite gráficamente entender la dinámica en los últimos 8 años para definir modelos enfocados a segmentación semántica para aplicar en entrenamiento para detectar la clase personas en escenas multivista. U-Net proporciona un mejor desempeño en el entrenamiento en términos de costo computacional, dado que permite entrenar un mayor número de épocas en un menor tiempo comparado con Mask R-CNN. La clasificación de la clase personas en el contexto de segmentación semántica alcanza un 89.91 % para Mask R-CNN, y del 90.53% para U-Net; y con la curva de operación característica del receptor (ROC) y el área bajo la curva (AUC) se obtiene un 90% y 92%. Aunque en la validación las métricas exactitud y curva AUC-ROC tienen un valor marginal no se considera una diferencia significativa entre los dos métodos, por tal razón, si los requerimientos computacionales son una limitante, la selección sería la arquitectura U-Net. Se requiere además como trabajo futuro construir un conjunto de datos más grande, posiblemente aumentado, se podrá entrenar mejor los modelos y en consecuencia los porcentajes en las diferentes métricas serán mejores.

6. Referencias

- Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, (1998) "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324. https://doi.org/10.1109/5.726791
- Pan, S. J., & Yang, (2010) Q. A Survey on Transfer Learning. https://doi.org/10.1109/TKDE.2009.191
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. NIPS, 1097-1105.
- Maturana, D., & Scherer, S. (2015). VoxNet: A 3D Convolutional Neural Network for Real-Time Object Recognition. In IEEE/RSJ (IROS) (pp. 922-928). https://doi.org/10.1109/IROS.2015.7353481
- Su, H., Maji, S., Kalogerakis, E., & Learned-Miller, E. (2015). Multi-view convolutional neural networks for 3D shape recognition. Proceedings of the IEEE ICCV, 945-953. https://doi.org/10.1109/ICCV.2015.114
- Su, H., Maji, S., Kalogerakis, E., & Learned-Miller, E. Multi-view Convolutional Neural Networks for 3D Shape Recognition. ICCV 2015. https://doi.org/10.1109/ICCV.2015.114
- Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017) Places: A 10 Million Image Database for Scene Recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence. https://doi.org/10.1167/17.10.296
- Chen, L. C., et al. (2018). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE pp. 834-848. https://doi.org/10.1109/TPAMI.2017.2699184
- Li, Y., Qi, H., Dai, A., Ji, X., & Wei, Y. (2016). Fully convolutional instance-aware semantic segmentation. In Proceedings of the IEEE CVPR pp. 2359-2367. https://doi.org/10.1109/CVPR.2017.472
- Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. Proceedings of the IEEE CVPR, pp. 3431-3440. https://doi.org/10.1109/CVPR.2015.7298965
- Zou, X., & Chen, S. (2018). An overview of point cloud semantic segmentation. IEEE pp. 3200-3214.
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. MICCAI, pp. 234-241. https://doi.org/10.1007/978-3-319-24574-4_28
- Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. NIPS, pp. 91-99.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. Proceedings of the IEEE CVPR, pp. 770-778. https://doi.org/10.1109/CVPR.2016.90
- He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask R-CNN. Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2980-2988. https://doi.org/10.1109/ICCV.2017.322
- Shin, H. C., et al. (2016) Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics, and transfer learning. https://doi.org/10.1109/TMI.2016.2528162
- Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2017). PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. Proceedings of the IEEE CVPR, pp. 1-11.
- Zhang, Y., & Jiao, J. (2020). 3D point cloud object detection on deep learning: A survey. Pp. 106-107.
- Yan, M., Mao, W., Li, B., & Li, H. (2018). VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. Proceedings of the IEEE CVPR, pp. 4490-4499.
- Chen, X., Kundu, K., Zhang, Z., Ma, H., Fidler, S., & Urtasun, R. (2016). Monocular 3D object detection for autonomous driving. In Proceedings of the IEEE CVPR pp. 2147-2156. https://doi.org/10.1109/CVPR.2016.236

- Milioto, A., Stachniss, C., & Behnke, S. (2019). RangeNet++: Fast and accurate LiDAR semantic segmentation. IEEE Robotics and Automation Letters, 4(2), 903-910. https://doi.org/10.1109/IROS40897.2019.8967762
- Yi, S. et al. (2020) SegVoxelNet: Exploring semantic context and depth-aware features for 3D vehicle detection from point cloud. Proceedings ICRA. pp. 2274-2280. https://doi.org/10.1109/ICRA40945.2020.9196556
- Kreuzberger, D., Kühl, N., & Hirschl, S. (2022). Machine Learning Operations (MLOps): Overview, Definition, and Architecture. ArXiv. /abs/2205.02302 https://doi.org/10.1109/ACCESS.2023.3262138
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. Proceedings of the IEEE CVPR, pp. 770-778. https://doi.org/10.1109/CVPR.2016.90

7. Agradecimientos

Se agradece a la OTAN por los recursos asignados en el programa Ciencia para la Paz Proyecto CLARIFIER (frequenCy-agiLe rAdar-lidaR chIp For surveIllancE moving platfoRms).

8. Sobre los Autores

- **Óscar Javier Montañez** Sogamoso, Ingeniero Electrónico, Estudios de Maestría en Ingeniería Electrónica en UPTC. Investigador Clarifier - OTAN. oscarjavier.montanez@uptc.edu.co
- **Nancy Carolina Roa Martín**, Psicóloga, Magíster en Psicología del Consumidor, Fundación Universitaria Konrad Lorenz. Investigador Clarifier - OTAN nancy.roa@uptc.edu.co
- **Eduardo Avendaño Fernández**, Ingeniero Electrónico, Doctor en Ingeniería Electrónica Universidad de Antioquia. Docente Titular en UPTC. eduardo.avendano@uptc.edu.co

---