Detección de novedades visuales basada en aprendizaje profundo para manejo autónomo

Miguel Angel Palacios Alonso, L. Enrique Sucar Succar, Hugo Jair Escalante Balderas

Reporte Técnico No. CCC-20-007

Noviembre, 2020

Luis Enrique Erro 1

Sta. Ma. de Tonantzintla,

72840, Puebla, México

Índice

Resumen

Un sistema de manejo autónomo (o semi-autónomo) permite a un vehículo sensar su entorno y navegar el camino reduciendo la intervención de un humano. En particular, el módulo de visión por computadora, busca proveer información visual relevante que permita al sistema detectar objetos de interés como son: obstáculos, carriles, señales de tránsito, otros autos y condiciones del entorno en general. El aprendizaje profundo ha mostrado ser un enfoque efectivo en diversas áreas de estudio, especialmente en visión por computadora. Haciendo uso de una gran cantidad de datos disponible, se han diseñado y entrenado modelos capaces de aprender una representación que permite detectar patrones o clases de interés en los datos. Sin embargo, aplicaciones críticas como el manejo automatizado, suelen interactuar con ambientes de trabajo altamente dinámicos y con múltiples clases de objetos donde existe la posibilidad de que se presente información visual que no corresponda a ninguna de las clases o patrones observados durante la fase de entrenamiento de un modelo.

Para abordar este problema, la detección de novedad busca identificar aquella información que difiere de los patrones o clases observados durante el entrenamiento de un modelo. Considerando que el aprendizaje profundo ha mostrado ser un enfoque capaz de abstraer múltiples niveles de representación, el presente trabajo de investigación propone el uso de aprendizaje profundo para el desarrollo de una solución capaz de aprender una representación que permita la detección de objetos visuales novedosos y de sus atributos en ambientes dinámicos como el capturado por la cámara de un sistema de manejo autónomo. Este documento describe el estado del arte revisado hasta el momento y los resultados preliminares obtenidos al realizar experimentos sobre una primer arquitectura utilizando un autocodificador como extractor de características y detector de novedad. También se muestra a través de un experimento exploratorio la posibilidad de utilizar mapas de activación para la localización del objeto novedoso en la imagen.

1. Introducción

Diversas aplicaciones críticas como son video-vigilancia, robótica móvil, diagnóstico médico y sistemas de manejo autónomo generan una gran cantidad de datos con la que es posible entrenar modelos de detección eficientes. Sin embargo, la complejidad del dominio de cada una de estas aplicaciones deja ver que a pesar de la gran cantidad de datos recabados, es posible la existencia de información no observada durante la creación del modelo que no este relacionada con ningún patrón descubierto por el modelo y que pudiera presentarse en la fase de prueba. Con el fin de no poner en riesgo el funcionamiento de sistemas en aplicaciones como las mencionadas anteriormente, una solución integral debe ser capaz de detectar tales situaciones.

En la última década, la automatización del manejo de vehículos a atraído la atención de la comunidad científica y la industria. Actualmente, una gran cantidad de autos son vendidos alrededor del mundo con un sistema de asistencia de manejo que permite distintos grados de autonomía. Este tipo de sistemas busca proveer al conductor de una mayor facilidad y seguridad en el manejo del vehículo. Para proporcionar la asistencia requerida, el sistema debe contar con diferentes módulos (ver Figura 1) que le permitan procesar una gran cantidad de información proveniente de sus sensores.

Figura 1: Bloques de construcción para manejo automatizado [1].

Con el fin de cubrir las necesidades que aplicaciones de este tipo requieren, la detección de novedad busca identificar aquellos datos de prueba (en tiempo de ejecución) que no muestran el comportamiento o patrón descubierto en los datos disponibles durante el entrenamiento [2]. En el presente trabajo se propone el uso de aprendizaje profundo para la detección de instancias de novedad visual considerando también los atributos que la describen dentro de las imágenes y videos. Las contribuciones esperadas de esta propuesta son :

- Una arquitectura y método basado en aprendizaje profundo para la detección de novedades en imágenes y video.
- Detección de novedades en escenarios de autos autónomos, incluyendo información de los atributos de las instancias de novedad.

1.1. Motivación

Un sistema de manejo automatizado (SMA) provee a un vehículo con la habilidad de sensar su entorno y navegar sin intervención de un humano [1]. Para lograr la autonomía deseada, un SMA cuenta con sensores que lo alimentan con la información necesaria para mantener la seguridad del vehículo. La Figura 2 muestra los sensores y el entorno de trabajo típico de un SMA. La cámara es un componente importante que alimenta de información visual al SMA. En conjunto, todos los componentes del SMA buscan proveer una representación del entorno en que se desplaza el vehículo.

Un SMA debe mantener en todo momento una representación del estado del entorno de tal forma que se preserve la seguridad del vehículo y de los elementos con los que se encuentre al navegar por el camino. Desde una perspectiva visual, un SMA debe ser capaz de inferir información relevante de un entorno de navegación dinámico, con la presencia de múltiples objetos y en condiciones de iluminación adversas. Un SMA debe estar preparado para tomar decisiones en todo momento,

incluyendo para aquellos escenarios que no haya observado antes.

Las técnicas de aprendizaje profundo han mostrado buen rendimiento en la extracción de información visual para este tipo de aplicaciones [3, 4, 5]. El alto rendimiento logrado por estas técnicas se encuentra fuertemente relacionado con la disponibilidad de suficientes datos de entrenamiento [6, 7]. El presente trabajo de investigación propone el desarrollo de un enfoque basado en aprendizaje profundo con el fin de detectar novedades visuales en imágenes o video. Además de esto, la solución propuesta también identificará los atributos que describan a la novedad como por ejemplo localización, tipo, si tiene capacidades de movimiento, entre otras.

1.2. Justificación

Diversas aplicaciones consideradas como tareas de seguridad crítica requieren tener la capacidad de manejar todo tipo de escenarios, incluso aquellos escenarios que no se observaron en la fase de creación de los sistemas.

Un ejemplo de este tipo de sistemas es un SMA, que debe ser capaz de identificar (desde una perspectiva de análisis visual) las condiciones del entorno en el cual se desplaza el vehículo. Incluso si esas condiciones no fueron observadas con anterioridad. De no ser posible, el SMA podría encontrarse ante la posibilidad de caer en

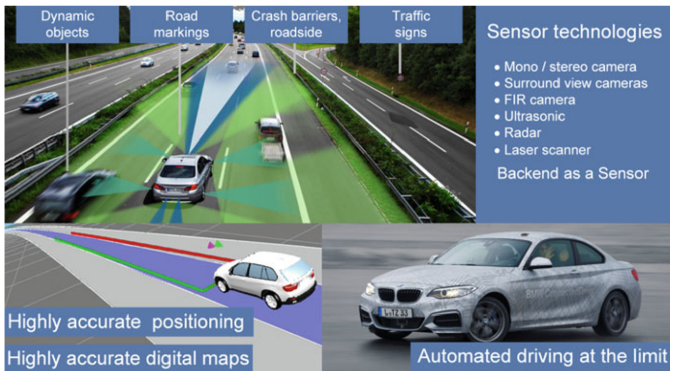


Figura 2: Entorno de un vehículo autónomo y sus tecnologías (imagen tomada de [1]).

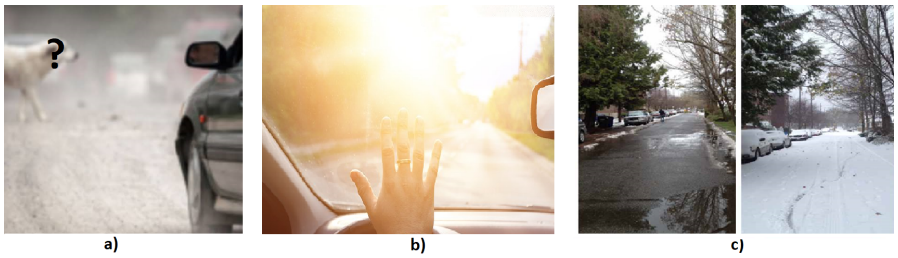


Figura 3: Escenarios visuales de novedad que un SMA puede captar, a) Un obstáculo novedoso, b) Cambios drásticos de iluminación natural y c) Cambios en el entorno, por ejemplo, una nevada atípica

un estado con alto grado de incertidumbre que puede poner en riesgo la seguridad del vehículo, del conductor y de los elementos externos presentes en ese momento.

La Figura 3 muestra algunos escenarios identificados en el estado del arte como fuentes de novedad visual para un SMA. El caso de un objeto desconocido sobre el camino es ejemplificado en la imagen 3a indicando la presencia de un obstáculo novedoso, cambios drásticos de iluminación naturales (a partir de condiciones externas del entorno, 3b) y cambios globales en el entorno (nevada por la tarde en un lugar dónde no suele nevar, ver Figura 3c) son condiciones que un SMA integral debe considerar.

El desarrollo de SMAs ha reportado distintos percances entre los que se encuentran impactos contra otros vehículos que en algunos casos ha llegado a provocar la muerte tanto de conductores como de peatones. Ejemplos de este tipo de situaciones son el percance ocurrido en mayo del 2016 cuando un vehículo tesla S 2015 equipado con el asistente de manejo *Autopilot* se impactó con un tráiler [8] que maniobraba en una intersección para hacer una vuelta en frente del Tesla. El Tesla no activó los frenos y terminó pasando por debajo del vagón del tráiler provocando la muerte del conductor. La compañía reportó que ni el conductor ni el asistente de manejo se dieron cuenta de la presencia del tráiler. Un segundo percance pero de la empresa Uber sucedió cuando uno de sus autos arrolló a una mujer que cruzaba la calle [8].

El hecho se dio en marzo del 2018 durante la noche, por lo que las condiciones de iluminación fueron un factor importante para que el sistema de visión fallara.

En un SMA, los sensores y los sistemas inteligentes detrás de ellos se complementan para proveer la autonomía requerida. Sin embargo, condiciones bastantes atípicas como el color blanco del costado del tráiler combinado con un fondo bastante iluminado en el primer percance y condiciones de iluminación específicas para el segundo, generaron una percepción visual no observada con anterioridad y provocaron que el sistema de detección de los SMAs de Tesla y Uber no fueran capaces de detectar la situación.

Otra aplicación que puede beneficiarse de la detección de novedades visuales es la video-vigilancia inteligente. Eventos de riesgo pueden ser modelados con un enfoque de detección de novedad. Por ejemplo, un modelo puede aprender que es normal que en el pasillo de algún lugar público la gente siempre pasa caminando sin detenerse, y detectar como algo extraño (novedoso) que alguien se quede merodeando por el lugar o peor aún que deje un paquete olvidado, el cual en situaciones extremas pudiera ser un indicio de un posible atentado.

Los ejemplos descritos anteriormente muestran situaciones donde la detección de novedad es bastante importante. Sin embargo, es posible que la sola detección de la novedad no sea suficiente. Por ejemplo, en un SMA además de reportar la presencia de una novedad, atributos de la misma pudieran ser de utilidad para el módulo de toma de decisiones. Saber que la novedad es del tipo de la Figura 3a y donde se encuentra puede ayudar a que el SMA decida seguir una estrategia específica para enfrentar la situación. Desde la perspectiva de análisis visual, es muy difícil generar observaciones de todas las situaciones posibles que un SMA enfrenta para poder entrenar un modelo que detecte este tipo de situaciones. Una solución con un enfoque basado en detección de novedad visual capaz de proveer además atributos que describan a las instancias de novedad en la escena complementaría al

SMA permitiendo agregar a las capacidades de detección de objetos conocidos la detección de objetos novedosos y de sus atributos reduciendo así la posibilidad de percances como los descritos.

2. Conceptos Básicos

2.1. Detección de valores atípicos

Hawkins [9] define un valor atípico como "una observación la cual se desvía bastante de otras observaciones como para despertar sospechas de que fue generada por un mecanismo diferente". Sin embargo, los términos "bastante" y "desvía" pueden ser interpretados de diferentes maneras dependiendo de las características del proceso a partir del cual surgieron los datos típicos o normales.

La comunidad de aprendizaje automático suele referirse a un valor atípico como anormalidad, desviación o anomalía. Y aunque la detección de valores atípicos suele ser utilizada principalmente como una tarea de limpieza de datos, también ha sido utilizada en la creación de modelos para aplicaciones tales como identificación de desinformación y mal comportamiento en la web, detección de anomalías en redes de información, análisis visual, detección de anomalías de tráfico, etc.

Cuando el conjunto de datos disponible para entrenar un modelo de detección de valores atípicos incluye sólo muestras de la clase normal el proceso es llamado detección de novedad. La detección de novedad puede así ser modelado en el contexto de clasificación de una clase donde pese a existir dos clases en la fase de prueba, solo se tienen datos de una (la clase positiva o normal) para entrenamiento. Así, un modelo de una clase, permitirá identificar si una nueva instancia se desvía o no de la normalidad capturada por el modelo.

Junto con la técnica a utilizar, el rendimiento de un modelo depende de la calidad estadística de los datos que lo alimentan. La limpieza y el modelado del comportamiento de los datos ha estado desde hace varios años dentro del estudio de la comunidad científica.

En la literatura se pueden identificar enfoques **probabilísticos** que buscan es-

timar la función de densidad de los datos, y en base a esta estimación identificar o rechazar datos anómalos. El uso de modelos de mezclas paramétricas [10], los enfoques basados en modelos de espacio de estados como los Modelos Ocultos de Markov y los Filtros Kalman (para representar un estado de anormalidad dentro las transiciones del modelo [11]) son ejemplos de este tipo de enfoques.

Otro grupo de métodos son aquellos **basados en distancia** que incluyen métodos de agrupamiento [12] y la noción de vecindad [13, 14]. Se utilizan medidas de distancia para calcular semejanza entre dos puntos. Aunque este tipo de enfoques suele tener problemas con datos de alta dimensión.

Medidas como la entropía, entropía relativa y sus variantes son aportes del grupo de **teoría de la información**. El objetivo de estas medidas es cuantificar el contenido de información en una base de datos [12]. Otros métodos denominados de **reconstrucción** suelen utilizarse para propósitos de clasificación o regresión. El error de reconstrucción generado entre la entrada y la salida definen un distancia que es utilizada como una medida de pertenencia (o no) a la clase normal. Ejemplo de este tipo de métodos son las redes neuronales [15] y en los últimos años las extensiones profundas de éstas.

2.2. Aprendizaje de representaciones

El aprendizaje de representaciones consiste en un conjunto de métodos que buscan descubrir de manera automática la estructura que separa las características relevantes de datos de entrada crudos [16, 17].

2.2.1. Aprendizaje profundo

El aprendizaje profundo es un área que ha mostrado un avance importante los últimos años. Basado en la disponibilidad de una gran cantidad de muestras, el

aprendizaje profundo permite entrenar un modelo que aprende representaciones que son expresadas en términos de otras [18]. Estas representaciones están formadas por la composición de transformaciones no lineales, regularmente redes neuronales. De esta forma, con la composición adecuada de transformaciones se pueden aprender funciones complejas que permiten extraer características de manera automática.

2.2.2. Redes Neuronales de Convolución

Una red neuronal de convolución (RNC)[19], es un tipo de red neuronal (RN) que es utilizada para el procesamiento de datos que tienen fuertes dependencias espaciales representadas a través de una topología en forma de cuadrícula [18, 20]. Son llamadas así debido a que utilizan una operación de convolución en lugar de una multiplicación de matrices. Con esta operación se busca filtrar una muestra de entrada x utilizando una máscara deslizante que la recorre. Las máscaras o filtros W representan los pesos que conectan la entrada con el mapa de características m resultante en una capa:

$$m_{ij}^{k} = \phi(W_{ij}^{k} * x_{ij} + b_k) \tag{1}$$

donde m^k indica el mapa de características, b_k el sesgo y ϕ es una función de activación.

De esta forma, una capa en este tipo de redes esta formada por un volumen de neuronas en 3D (ancho, alto y profundidad). Un arquitectura común de convolución, contiene además capas de sub-muestreo no lineal y una serie de capas multi-conectadas como lo muestra la Figura 4.

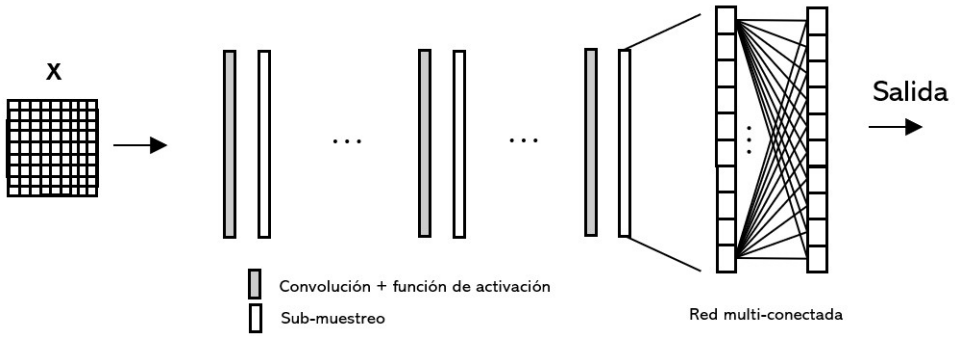


Figura 4: Vista esquemática de una red neuronal de convolución básica conformada por uno o más pares de capas de convolución y sub-muestreo y una serie de capas multi-conectas

2.2.3. Autocodificadores

Un autocodificador es una RN que es entrenada para reproducir la entrada en la salida. Esta formada con una capa oculta intermedia h que describe un código para representar la entrada, una función codificadora $h = f(x)$ y una decodificadora que produce una reconstrucción $r = g(h)$. La Figura 5 muestra el esquema general de la red.

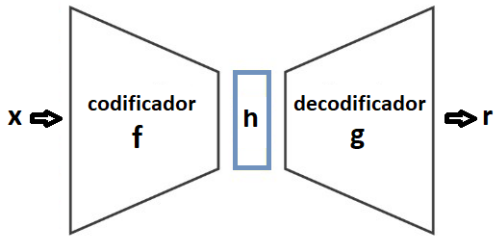


Figura 5: Esquema general de un autocodificador.

Con el fin de obtener atributos de interés del autocodificador se suele restringir la dimensión de h de tal forma que la dimensión sea más pequeña que x. Esta restricción provoca que el autocodificador capture las atributos más sobresalientes del conjunto de datos de entrenamiento X. De esta forma el proceso de aprendizaje de un autocodificador busca minimizar una función de pérdida $L(x, g(f(x)))$. Dónde

L es una función que penaliza que $g(f(x))$ y x sean distintos.

Cuando las funciones son planteadas como mapeos estocásticos se le conoce como autocodificador **variacional**. Si el conjunto de entradas X es corrompida con algún tipo de ruido, el autocodificador **con eliminación de ruido** aprende a deshacer las partes corruptas.

2.2.4. Redes adversarias generativas

Este tipo de redes están basadas en un escenario teórico de juego en el cual una red generadora g debe competir contra un adversario discrimininador d. La red generadora produce muestras $x = g(z; \theta^{(g)})$. Dónde z es una variable latente y $\theta^{(g)}$ los parámetros de g. La red discriminadora intenta distinguir entre muestras tomadas del conjunto de entrenamiento (muestras reales) y muestras arrojadas por la red generadora (muestras falsas). La salida del discriminador es un valor de probabilidad dado por $d(x; \theta^{(d)})$, dónde x es una muestra y $\theta^{(d)}$ los parámetros de d. El valor de probabilidad de salida del discriminador indica si x es una muestra de entrenamiento real o no.

De esta forma, el discriminador intenta aprender a clasificar las muestras reales y falsas. Al mismo tiempo, el generador intenta aprender a generar muestras reales para engañar al discriminador.

2.2.5. Redes que crecen cuando es requerido

Este tipo de red es una red auto-organizada que consiste de una capa de entrada, una capa de grupos y una neurona de salida [21, 22]. La Figura 6 muestra la topología de la red. Todas las capas se encuentran conectadas completamente. Entre los nodos de la capa de grupos existen enlaces de vecindad que indican similitud en atributos. Durante el entrenamiento es posible agregar o quitar nodos y enlaces. La

red crece cuando se presentan nuevos datos que no emparejan con la red y permanece estática de lo contrario. La salida corresponde al valor actual de habituación de la neurona de la capa de grupos ganadora.

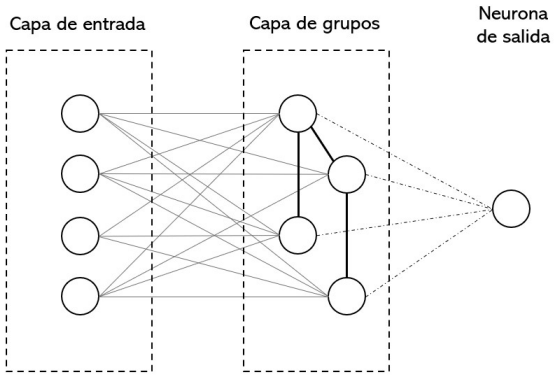


Figura 6: Topología de una red que crece cuando es requerido. Es una red completamente conectada. Esta formada por dos capas, una de entrada, una de grupos y una neurona de salida. Entre los nodos de la capa de grupos existen enlaces de vecindad indicando similitud entre atributos. Los enlaces entre la capa de grupos y el nodo de salida son de habituación.

2.2.6. Redes basadas en teoría de resonancia adaptativa

Grupo de redes neuronales que trabajan con el dilema de estabilidad-plasticidad que les permite incorporar nueva información mientras mantienen el conocimiento adquirido [23, 24]. Este tipo de red permite clasificar una muestra de entrada utilizando una medida de similitud con respecto a vectores que representan las clases aprendidas (ver esquema en Figura 7). La red básica ART-1 consiste de dos capas con neuronas totalmente conectadas mutuamente. La capa de comparación (o F1) verifica si el vector de entrada tiene similitud con el conjunto de vectores almacenados y la capa de reconocimiento (o F2) reconoce e incorpora el conocimiento del vector de entrada a una de las clases existentes. El número de neuronas de esta capa es dinámico ya que pueden agregarse nuevas neuronas. Se realiza un aprendizaje

competitivo por lo que solo se activa la neurona que empareje con los atributos de la entrada.

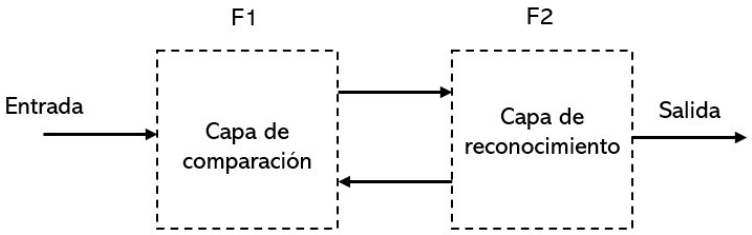


Figura 7: Vista esquemática de la red ART-1

3. Trabajo Relacionado

La detección de novedades ha sido estudiada con diversos enfoques que incluyen frecuentistas, bayesianos, de teoría de la información, métodos de máquinas de vectores y redes neuronales [2].

La presente propuesta busca estudiar el uso de aprendizaje profundo para la detección de novedad visual y sus atributos en imágenes y video. A continuación se describen los principales enfoques identificados en el estado del arte. Se privilegian aquellos trabajos que tienen una relación directa con detección de novedad en contextos de navegación autónoma, pero sin dejar de lado aquellos trabajos recientes que realizan detección de novedad utilizando aprendizaje profundo en otros dominios.

3.1. Detección de novedades en contextos de navegación autónoma

En el contexto del análisis visual de información, el aprendizaje profundo ha mostrado ser un enfoque capaz de tratar con la complejidad del dominio y con los requerimientos de respuesta en tiempo real que sistemas críticos requieren. Entre las propuestas identificadas en el estado del arte están las que ofrecen capacidades de aprendizaje en línea y otras que requerirían de ajustes para poder soportar esta capacidad.

3.1.1. Aprendizaje en línea

En el contexto de tareas de inspección usando un robot móvil, [22] presenta una metodología que utiliza un enfoque de detección de novedad. Se entrena una red neuronal con el objetivo de ignorar percepciones normales y resaltar aquello que no fue observado con anterioridad. El enfoque es capaz de incorporar información en línea. La arquitectura propuesta es una red que crece cuando se requiere (GWR, por

sus siglas en inglés) [21] que mantiene capacidades de habituación en la salida. El robot aprende a detectar novedad a través de la exploración de su entorno (un pasillo). La información de entrada proviene de un sonar y una cámara monocromática. Con el uso de habituación en la salida de la red se registra si un nodo de la red ha sido disparado antes. Registrando así también aquello que ya ha sido visto e inhibiendo el disparo del nodo ante muestras similares una vez alcanzado un umbral. Los autores muestran, por un lado que el método es capaz de detectar la novedad y por otra lado que después de observar repetidamente el mismo punto novedoso, es capaz de incorporarlo al conocimiento y por lo tanto deja de ser novedoso.

En [25] se presenta un análisis para evaluar el compromiso robustez-fidelidad de redes GWR para la detección de novedad tanto para una imagen, como para detección continua. La alta fidelidad permite el aprendizaje de modelos detallados permitiendo que se puedan detectar varios eventos novedosos de manera consecutiva o cuando se requiere detectar que varios objetos encontrados previamente han desaparecido. La robustez, es lograda por la generalización sobre el espacio de entrada minimizando los efectos de los parámetros de la red y del entorno. Las pruebas fueron realizadas a través de entornos virtuales simulando un cuarto cuadrado y un corredor. Las entradas al modelo son números reales que representan un valor de color (rojo = 1/6, verde = 1/3 y azul = 1/2), ángulo y distancia relativa proporcionados por un sensor de luz omnidireccional virtual. En cada paso, se utiliza como entrada al modelo un total de F X 3 elementos, donde F es el número actual de atributos que se capturaron del entorno. Los atributos son ordenados de acuerdo a la distancia con respecto al robot (primero el más cercano). Se analizaron las redes GWR, GWR plástica (donde cada neurona de agrupamiento puede tener diferente número de conexiones de entrada) de crecimiento rápido y GWR plástica de crecimiento balanceado. Las redes mostraron un mejor rendimiento cuando se les agregaron datos de localización normalizados al conjunto de entrada.

Softman et al. [26] presentan un algoritmo de detección de novedad que se adap-

ta en línea basado en la estimación de la densidad de los datos de entrenamiento. Para tener una mejor separación de las clases y generar un subespacio dimensional reducido los autores utilizan Análisis Discriminante Múltiple (MDA, Multiple discriminant Analysis) en lugar de PCA MDA puede verse como una variante de el algoritmo NORMA [27], una máquina de vectores de soporte con kernels optimizados en línea a través de gradiente descendente estocástico. Se establece un bufer de tamaño fijo que permite optimizar/actualizar la información del kernel moviendo hacia al frente del bufer la información que "rompió" la novedad. El método es probado en un vehículo terrestre autónomo en terreno abierto.

Para detectar obstáculos [28, 29] proponen un enfoque basado en detección de novedad. El método propuesto realiza la estimación de la covarianza en línea (peso del kernel) basado en las observaciones actuales para la estimación en línea de la densidad de un kernel de una variable aleatoria. El peso determina la importancia relativa de cada descriptor dependiendo de las condiciones actuales del entorno. El enfoque es probado en una plataforma robótica agrícola AgBot I [30] al navegar por un terreno en exteriores.

3.1.2. Aprendizaje fuera de línea

Xia et al.[31] proponen un marco de trabajo basado en un autocodificador que busca eliminar valores atípicos. Los datos de entrenamiento incluyen datos normales y atípicos En [32] se propone un enfoque que utiliza información de apariencia y señales geométricas y de contexto para la detección de obstáculos pequeños en el camino. La entrada al sistema propuesto son imágenes estéreo. El sistema se divide en dos procesos:

1. Se utiliza una red de convolución con arquitectura GoogleNet [33] para etiquetado semántico de píxeles. Las clases que identifica esta red corresponden a i) espacio libre, ii) obstáculos inesperados en el camino y iii) fondo.

2. Utilizando el par estéreo de imágenes, se obtiene información geométrica por emparejamiento semi-global. La detección de los obstáculos se realiza a través de pruebas estadísticas de hipótesis basadas en modelos. Para la representación de los obstáculos se utilizan "stixels" [34].

La salida de estos dos procesos se fusiona de manera probabilística para producir la salida final 3D en stixels. El método se valida utilizando la base de datos "Lost and found" [35] que contiene obstáculos pequeños de 5 cm. El trabajo reporta un 90 % de detección para distancias de hasta 50 m.

El enfoque propuesto en [36] utiliza rasgos visuales sobresalientes para la detección de objetos novedosos considerando un proceso de detección completo. Es decir, busca detectar las "clases comunes" y las clases novedosas al mismo tiempo. Para ello hace uso de propagación visual hacia atrás [37] y de un clasificador de una clase que utiliza el Índice de Similitud Estructural (ISE) [38] como función de pérdida. El método propuesto se valida con dos conjuntos de bases de datos, una de manejo en el mundo real (Udacity [39]) y otra de un entorno de carreras casero.

En [40] se plantea un enfoque de extremo a extremo (end-to-end) basado en aprendizaje profundo que considera detección de novedades en el contexto de manejo autónomo de autos. Se propone una arquitectura basada en un autocodificador variacional para control. Se aprende un modelo que mapea un conjunto de imágenes de entrada a comandos de giro basado en la curvatura de el camino como salida. La parte codificadora del autocodificador esta formada por una red de convolución conformada por 5 capas de convolución y 2 capas completamente conectadas para calcular la distribución de las variables latentes. El decodificador refleja al codificador, 2 capas completamente conectadas y 5 capas de deconvolución. Los autores recopilaron datos a partir de un vehículo real Toyota Prius 2015 V, una cámara Leopard Imaging LIAR0231-GMSL, así como sensores para calcular ángulo de giro y velocidad del auto. La incertidumbre o novedad en el entorno es modelada a través

de la varianza de las variables latentes.

Ritcher et al. [41] combinan un modelo de predicción de colisión con un detector de novedad. El detector de novedad utiliza un autocodificador para detectar cambios en el ambiente de navegación (pasillo con obstáculos) de un auto pequeño autónomo RC. En lugar de elegir un umbral de novedad a partir de prueba y error, propone construir una función de distribución acumulativa con los errores de reconstrucción del autocodificador obtenidos durante el entrenamiento.

3.2. Otros contextos

En [42] se propone el uso de información generada por los filtros positivos (pesos positivos) de una red de convolución para la clasificación y detección de objetos novedosos. Los autores consideran que los filtros positivos (en particular) pueden ayudar a aislar a las clases conocidas por un clasificador de las que no lo son, es decir, el porque un filtro positivo es más descriptivo con respecto a la clase que se pretende representar. Se propone también el uso de una función de pérdida que los autores llaman de membresía, la cual busca que solo una activación positiva este presente en el vector de activación final.

OCGAN (one-class Generative Adversarial Network) [43], es un modelo basado en el aprendizaje de representaciones ocultas de las muestras que pertenecen a la clase de interés utilizando un autocodificador con eliminación de ruido, discriminadores y un clasificador. Los autores proponen restringir de manera explicita el espacio oculto para exclusivamente representar la clase. El enfoque utiliza la función de activación de tangente hiperbólica. Se utiliza un discriminador en el espacio oculto que es entrenado de manera adversaria para asegurar que la representación codificada produce muestras del espacio oculto acotado. Un segundo discriminador visual es utilizado para diferenciar entre imágenes de la clase e imágenes generadas con el decodificador. Un clasificador binario (entrenado con muestras de la clase y

muestras generadas por el decodificador) es utilizado para detectar reconstrucciones "erróneas". Los autores presentan resultados sobre las bases de datos CIFAR10, COIL, FMNIST y MNIST. El método propuesto es efectivo cuando un concepto simple esta presente en las imágenes como en COIL, MNIST y fMNIST.

Sabokrou et al. [44] proponen utilizar un enfoque adversario-generativo. El generador de esta red es un autocodificador de convolución con eliminación de ruido y el discriminador una red también de convolución. Para entrenar el autocodificador se le agrega ruido gaussiano a las muestras de entrenamiento.

3.3. Discusión

Hasta ahora, la revisión del estado del arte muestra la existencia de 3 categorías en la que pueden englobarse los trabajos citados y que por sus características favorecen el desarrollo de un enfoque para la detección de novedad visual:

- Redes generativas [41, 44, 42, 43, 40, 36, 32]. Permiten abordar de manera natural el problema de detección de novedad a través de la reconstrucción de imágenes. Como elemento crítico se identifica la métrica de diferenciación entre la imagen real y la reconstruida que pudiera no ser suficientemente discriminativa para definir un umbral de novedad.

- Redes auto-organizadas [21, 22, 25]. Basadas en agrupamiento y crecimiento de la capa de agrupamiento. Proporcionan una arquitectura ideal para el aprendizaje en línea y pudiera ser utilizada a la par de un detector de lo "normal" para tener la solución integral. Puntos críticos, la estrategia de agregado de nodos y habituación de las salidas.

- Máquinas de vectores de soporte (MVS) [28, 29, 26]. Aunque la presente propuesta busca estudiar un enfoque basado en aprendizaje profundo, se han incluido trabajos realizados con máquina de vectores de soporte debido a la cercanía con el dominio de vehículos autónomos.

Mientras que los trabajos basados en MVSs están más enfocados a minimizar el efecto negativo que provoca en la detección la existencia de atributos irrelevantes, redundantes o ruidosos; la comunidad de aprendizaje profundo esta trabajando en el diseño de la red que resuelva el problema de representación y de detección de manera conjunta a través de modelos profundos. Sabokrou [44] utiliza un enfoque adversario estándar pero efectivo utilizando una representación latente Gaussiana en un autocodificador que hace el rol de la red generadora de una GAN. Mientras que

[43] aplica también enfoque adversario pero a nivel local restringiendo a diferentes etapas el aprendizaje de la red generadora. Estas restricciones buscan asegurarse que la capa latente de la red generadora solo sea capaz de reconstruir muestras de la clase con que fue entrenada. Por otra parte, las redes auto-generadas GWR han sido utilizadas en [22, 25] para la detección de novedad durante la navegación de un un robot tanto real como simulado.

Ninguno de los enfoques revisados hasta ahora ha sido evaluado en un entorno urbano típico de manejo autónomo y ha sido poco explorado una metodología de detección de novedad que incluya la detección de los atributos semánticos que describan a las instancias de novedad.

4. Propuesta de Investigación

Un sistema de detección integral debe ser capaz de manejar observaciones visuales que fueron observadas durante el proceso de entrenamiento del sistema, pero también de considerar situaciones excepcionales en las que por las condiciones de la aplicación se presenten eventos visuales que no se observaron con anterioridad.

El presente trabajo de investigación propone el uso de aprendizaje profundo para el desarrollo de una metodología que permita detectar instancias de novedad visual y los atributos que las describen a partir de imágenes y videos.

4.1. Planteamiento del Problema

La detección de novedad es un requerimiento fundamental para un sistema de clasificación o detección integral. La detección de novedad visual puede formularse de la siguiente manera:

Dado un conjunto finito de imágenes o videos de entrenamiento D que puede

ser representado mediante un conjunto finito de características R generado por una distribución de probabilidad $P(R)$ sobre R . El problema consiste en identificar si para un nuevo dato d representado por características r_d, $P(r_d)$ es menor a un épsilon dado por $P(R)$. Si $P(r_d)$ es menor a un épsilon, se busca identificar también los atributos A que describen a d en la imagen o video.

Es importante notar que R y $P(R)$ no están dados a priori y es una de las tareas más importantes a desarrollar dentro de esta propuestas de investigación. Encontrar las características y la distribución de esas características que capturen el comportamiento del dominio de estudio, entornos de manejo automatizado reales capturados en imágenes y video.

4.2. Preguntas de Investigación

Las preguntas de investigación para esta propuesta doctoral son:

- ¿Cómo diseñar un modelo con la capacidad de detectar información visual novedosa a partir de un conjunto de imágenes que contienen múltiples elementos visuales?
- ¿De qué manera se puede mejorar la representación aprendida por un modelo profundo para aumentar la separación entre la clase normal y la clase novedosa?
- ¿Cómo puede extenderse el modelo para su aplicación en video?
- ¿Cómo y qué información de los atributos es posible obtener de los elementos novedosos detectados para aplicaciones de manejo automatizado?

4.3. Hipótesis

Una metodología basada en redes neuronales profundas puede aprender una representación que permita detectar y proporcionar información de los atributos que describen a los eventos visuales novedosos ocurridos en aplicaciones de manejo automatizado, es decir, en entornos dinámicos y con múltiples elementos. La representación aprendida permitirá una eficacia competitiva en la detección de novedad visual con respecto a los trabajos del estado del arte.

En la metodología a desarrollar, la capacidad de distinguir eventos novedosos será evaluada en términos de la precisión utilizando medidas como Área Bajo la Curva (AUC, por sus siglas en inglés) y medida-F; mientras que la capacidad de proporcionar información de los atributos de la novedad será evaluada en forma cualitativa mediante encuestas con expertos del dominio.

4.4. Objetivos

4.4.1. Objetivo principal

Diseñar, desarrollar, implementar y evaluar una metodología para el aprendizaje de representación para la detección de novedad y de sus atributos semánticos en imágenes y videos utilizando aprendizaje profundo, y su aplicación a manejo automatizado.

4.4.2. Objetivos específicos

1. Diseñar la arquitectura de la red que permita aprender una representación que favorezca la separación de la clase normal y el resto de clases para la detección de novedades en imágenes.

2. Definir una función de pérdida para detectar novedades visuales.

3. Diseñar un modelo de detección de atributos de las instancias de novedad detectadas considerando imágenes.

4. Diseñar un modelo de detección de atributos de las instancias de novedad detectadas considerando video.

5. Evaluar los métodos propuestos en entornos de manejo autónomos, incluyendo simuladores, imágenes/videos reales y un auto autónomo a escala.

4.5. Alcance y Limitaciones

Para la detección de la novedad sólo se tienen datos de la clase normal para entrenar el modelo.

En la presente propuesta se utilizará principalmente información visual relacionada con entornos de manejo autónomo de autos, pero la metodología podrá ser utilizado en otros dominios con escenas dinámicas y múltiples clases donde se tengan muestras de datos que describan la normalidad.

4.6. Contribuciones Esperadas

- Una arquitectura y método basado en aprendizaje profundo para la detección de novedades en imágenes y video.

- Detección de novedades en escenarios de autos autónomos, incluyendo información de los atributos de las instancias de novedad.

4.7. Metodología

Esta propuesta doctoral es acerca de detectar novedades visuales así como de obtener atributos semánticos que las describan considerando como datos de entrada imágenes y video. En la Figura 8 se muestra el diagrama de la metodología propuesta.

Figura 8: Metodología propuesta

1. Análisis de redes neuronales profundas. En este paso se analizarán los modelos profundos de convolución, autocodificadores y multi-conectados que favorezcan la detección de novedad visual. Se analizarán los algoritmos de entrenamiento y en particular las funciones de pérdida existentes actualmente que consideren dentro de su formulación la evaluación de información visual. Se estudiará también el manejo de incertidumbre en modelos profundos.

2. Propuesta/extensión de una arquitectura basada en redes neuronales profundas. De la revisión y estudio realizado en el paso (1) se propondrá una red de convolución con estructura de autocodificador que sea capaz de obtener

características que favorezcan la detección de novedad. Se se elegirán los algoritmos de entrenamiento y función de pérdida que mejoren el rendimiento del modelo. Este modelo será mejorado de manera incremental considerando imágenes en sus primeras versiones y video en versiones posteriores. La evaluación se realizará utilizando validación cruzada para diferentes bases de datos de imagenes/videos. Se estudiarán también estrategias de pre-procesamiento con fines de detección de novedad.

3. Diseño experimental. Cada modelo propuesto será evaluado de acuerdo al estado de desarrollo en que se encuentre. Para ello se consideran actividades de estudio, revisión de herramientas y preparación de bases de datos. Los modelos serán evaluados en términos de la precisión utilizando medidas como área bajo la curva y medida-F.

    - Análisis de bases de datos de imágenes identificadas en el estado del arte. Con el fin de tener elementos de comparación con otros enfoques, se realizarán experimentos de validación con bases de datos identificadas en el estado del arte diferentes al dominio de manejo automatizado.
    - Estudio del simulador Carla [45]. Se realizará el estudio del simulador para autos autónomos Carla. En un entorno urbano simulado, se identificará la forma de integrar elementos al simulador. Permitiendo así, la generación de novedades sintéticas que permitan validar la detección éstas y de sus atributos.
    - Estudio de la plataforma autoNOMOS [46]. La metodología propuesta será evaluada también a través de la plataforma robótica mini tipo auto autoNOMOS. El robot miniatura autonNOMOS es un mini auto a escala 1:10 desarrollado por la Universidad Libre de Berlin. Para el desarrollo en el mini auto es necesario tener un dominio de la plataforma ROS y del funcionamiento básico del robot.

- Análisis de bases de datos de manejo autónomo. Se analizarán y prepararán bases de datos en el contexto de manejo automatizado con información de entornos reales. Ejemplo de este tipo de datos son la base de datos FORD [47] y KITTI [48].

4. Análisis de métodos de clasificación no supervisada. Se realizará el estudio y análisis de este tipo de métodos, en especial de modelos autoorganizados como la red GWR y sus variantes. El modelo será utilizado para la clasificación de la novedad.

5. /extender un modelo de clasificación. Basado en el paso (4), se propondrá una arquitectura que permita ampliar la separación lograda por el autocodificador entre las muestras normales y las novedades. La entrada a este modelo o método de clasificación serán las características y error de reconstrucción obtenidas con el autocodificador propuesto en el paso (2). Se realizará validación cruzada para la evaluación del modelo.

6. Análisis de métodos de identificación de atributos de instancias visuales. Esto métodos incluyen detección de prominencia visual, mapas de activación en redes de convolución y segmentación semántica entre otros.

7. Proponer/extender modelo de identificación de atributos de novedad visual. Basado en el paso (6) se propondrán métodos de detección de atributos de novedad visual iniciando con el atributo de localización. La evaluación será cualitativa mediante encuestas con expertos del dominio.

8. Pruebas de integración de modelos. La extracción de características (2), detección de novedad (5) y detección de atributos (7) serán evaluados de manera conjunta teniendo como entrada imágenes y video.

9. Análisis de métodos de detección de novedad considerando video como entrada. Estudio de métodos de detección de objetos y atributos considerando

información temporal como datos de entrada.

10. Proponer/extender modelo para detección de novedad en video. Basado en el paso (6) se propondrán extensiones al modelo desarrollado con el fin de permitir la detección de novedad en video.

El desarrollo de la solución propuesta busca una mejora incremental e iterativa así como consideraciones o extensiones que se requieren para la detección de novedad y atributos considerando video. Por lo que una vez alcanzado el paso 10 de la metodología es posible realizar mejoras en los modelos desarrollados según sea necesario.

Figura 9: a) El simulador Carla para autos autónomos, b) El mini auto autoNOMOS

4.8. Cronograma de actividades

La Figura 10 muestra el cronograma de actividades propuesto para realizar en esta propuesta de investigación. Junto con la metodología, cuatro categorías guían el flujo de trabajo: Antecedentes, Propuesta, Desarrollo y Preparación de tesis.

Antecedentes y *Preparación de tesis* incluyen actividades como la revisión del estado del arte y la escritura de la tesis que se realizarán de manera periódica con el fin de incorporar nueva información. *Desarrollo* contiene el conjunto más amplio de actividades a realizar. Se considera un flujo incremental que incluye Diseño, Desarrollo y pruebas del modelo objeto de esta propuesta.

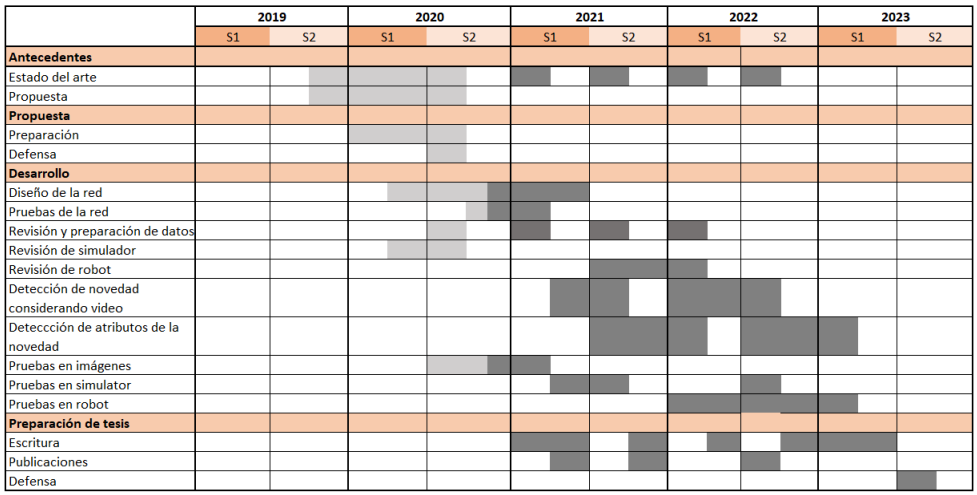


Figura 10: Cronograma de actividades

4.9. Plan de Publicaciones

- Una metodología basada en aprendizaje profundo para detección de novedad visual en imágenes. Congreso.
- Detección de novedad visual en video utilizando redes generativas. Revista Neurocomputing.
- Atributos de novedad visual para manejo autónomo. Revista Robotics and Autonomous Systems.

5. Resultados Preliminares

A continuación se muestran los resultados experimentales alcanzados hasta el momento. Se muestra la arquitectura de una red de convolución, que permite la obtención de características al mismo tiempo que favorece la detección de novedad. También se muestra como el mapa de activación generado por una red de convolución simple (para fines de clasificación) puede ser utilizado para el cálculo de uno de los atributos de novedad, la localización de la novedad en la imagen.

5.1. Autocodificadores

En el estado del arte se identificó al autocodificador como uno de los modelos utilizados para la detección de novedad. La capacidad de codificación y decodificación del autocodificador es utilizada como una forma de reducción de dimensionalidad, pero también como de detección de novedad. Al entrenar el modelo con un conjunto de muestras, se espera que en fase de prueba, sea capaz de reconstruir aquello que ha visto antes con un error menor que aquello que no ha visto permitiendo así una separación de las muestras.

5.1.1. Autocodificador de convolución con entrenamiento adversario para el aprendizaje de características utilizando una función de pérdida compuesta

Un autocodificador de convolución es una red que tiene una arquitectura que contempla un codificador y un decodificador cuyas capas son de convolución. A continuación se presenta un red de convolución que es entrenada de manera adversaria (GAN por sus siglas en inglés), por lo que además del autocodificador que toma el rol de la red generadora también se cuenta con una red discriminadora (como se describe en 2.2.4). El enfoque propuesto esta basado en el trabajo presentado por [44].

El diseño de esta arquitectura pertenece al objetivo 1, mientras que las pruebas utilizando una función de pérdida compuesta aporta al avance del objetivo 2. El experimento busca evaluar la arquitectura propuesta bajo las métricas de error cuadrático medio (ECM) e ISE ponderados por un valor α de tal forma que la función de perdida queda definida como:

$$L = \alpha ISE + (1 - \alpha) ECM \tag{2}$$

Los experimentos reportados se realizaron con $\alpha = 0.6$

Configuración

- Datos de entrenamiento y prueba: Se generó un modelo para cada una de las clases 0-9 de la base de datos MNIST [49]. Cada una de estas clases se tomaron como la clase normal y el resto de clases son consideradas novedades. Para el entrenamiento, a las muestras de entrada se les agregó ruido. En la fase de prueba, se utilizó un conjunto de muestras de la clase normal (distintas a las de entrenamiento) más un conjunto de muestras de las clases novedosas.

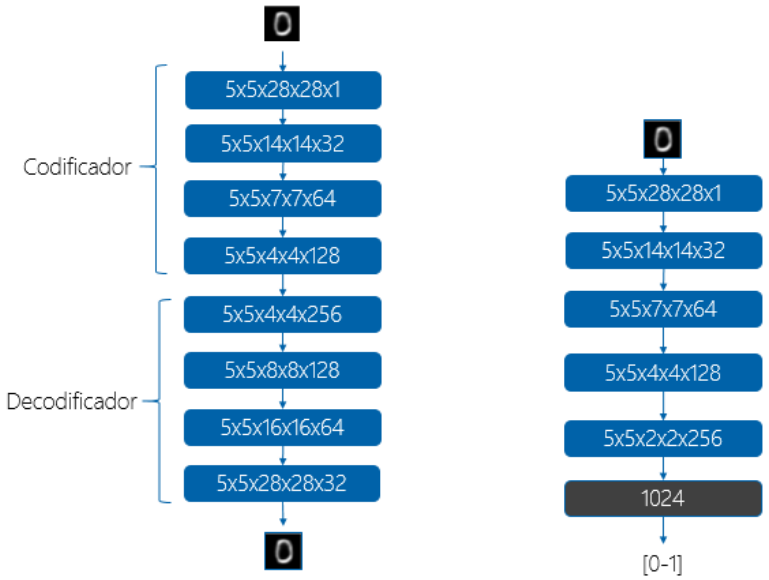
- Función de pérdida: ISE y ECM ponderados como en la ecuación 2 para el autocodificador y entropía cruzada binaria para el discriminador.

- Arquitectura: La Figura 11 muestra la arquitectura del autocodificador y el discriminador de convolución que conforman la red adversaria. Se utiliza activación LeakyReLU y Normalización después de cada capa del discriminador y del codificador. Se usa el algoritmo de optimización RMSprop con una tasa de aprendizaje de 0.0002. En fase de entrenamiento el autocodificador recibe como entrada muestras de la clase típica o normal con ruido añadido, la salida del autocodificador es a su vez entrada para el discriminador. Además el discriminador recibe también entradas de la clase normal. De esta manera

el autocodificador trata de engañar al discriminador al reconstruir la entrada cada vez mejor (eliminando el ruido en la imagen) mientras que el discriminador trata de identificar las muestras normales de las "falsas" generadas por el autocodificador. En este experimento, el error de reconstrucción del generador es utilizado para definir si una muestra es novedosa o no en la fase de prueba. Debido a que el modelo fue entrenado con datos de la clase normal, se espera que muestras de esta clase sean reconstruidas con un error pequeño, mientras que por el contrario, una muestra novedosa obtendrá un error de reconstrucción mayor. De esta forma, muestras con un error de reconstrucción mayor a un umbral corresponden a muestras novedosas y menores a ese umbral son parte de la clase normal.

Resultados

La Figura 12 muestra ejemplos de los resultados obtenidos para una instancia de los modelos 1, 2 y 9 (de los 5 evaluados en la validación cruzada por clase). Se muestran los histogramas del ECM, ISE e ISE+ECM con $\alpha = 0.6$ calculados en fase de prueba. De esta manera, se espera que el modelo que fue entrenado con muestras de la clase 1 aprenda a reconstruir instancias de esta clase con un error pequeño en fase de prueba, mientras que números de las clases distintas de 1 deberán obtener un error mayor. En los histogramas el color azul corresponde a los errores de reconstrucción de la clase *normal* y en rojo los errores de reconstrucción de datos *novedosos*. En la columna derecha de la Figura 12 se pueden observar también la respectiva curva ROC y el valor AUC de cada uno de los modelos aprendidos. Los AUCs promedio (y sus respectivas desviaciones estándar) para los 9 modelos (correspondientes a las 9 clases de la base de datos Mnist) puede verse en la Tabla 1. Se realizó validación cruzada por cada modelo con $k = 5$.



(a) Autocodificador de convolución (b) Discriminador

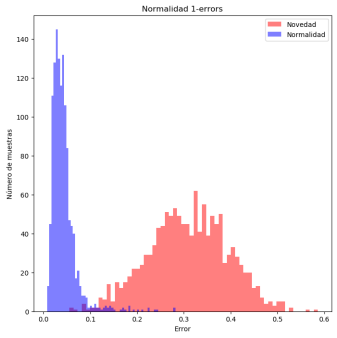
Figura 11: a)Generador (autocodificador de convolución) y b) discriminador de la red adversaria. Los parámetros de las capas de convolución pueden leerse como: primera dimensión del filtro X segunda dimensión del filtro X primera dimensión del mapa de características X segunda dimensión del mapa de características X tercera dimensión del mapa de características que entra a la capa. La entrada a la capa 1 es una imagen de Mnist de 28x28x1.

Tabla 1: AUCs promedio y su desviación estándar calculados a partir de los errores obtenidos en fase de prueba para los modelos correspondientes a cada una de las clases 0-9

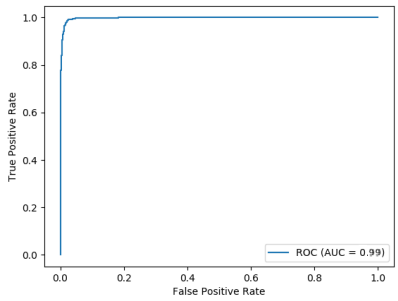
Modelo	0	1	2	3	4	5	6	7	8	9
ECM	0.94 ± 0.0078	**0.99 ± 0.0018**	**0.8 ± 0.0353**	**0.87 ± 0.0329**	**0.88 ± 0.0136**	**0.85 ± 0.0306**	0.94 ± 0.0572	**0.93 ± 0.0041**	0.91 ± 0.0117	0.93 ± 0.011
ISE	**0.95 ± 0.0164**	**0.99 ± 0.001**	**0.8 ± 0.022**	0.85 ± 0.0115	0.86 ± 0.0136	0.82 ± 0.0179	**0.98 ± 0.0038**	**0.93 ± 0.0138**	**0.93 ± 0.0266**	0.95 ± 0.0175
ISE+ECM	**0.95 ± 0.0164**	**0.99 ± 0.0005**	**0.8 ± 0.0393**	0.85 ± 0.0278	0.86 ± 0.022	0.84 ± 0.0266	0.97 ± 0.0113	0.92 ± 0.0074	0.92 ± 0.015	**0.96 ± 0.0059**

5.2. Localización de la novedad

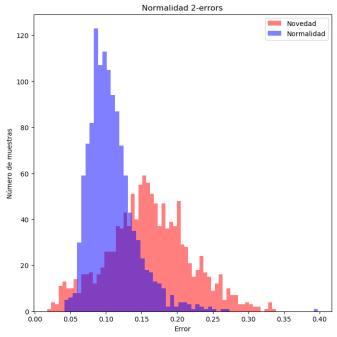
Una de las aportaciones de este trabajo consiste en el diseño de una solución que considera la identificación de los atributos de una instancia de novedad (paso 7 de la metodología). Este experimento exploratorio da indicios de que los mapas de



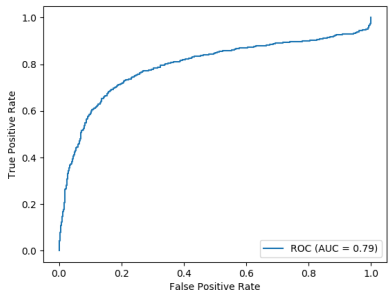
(a) Histograma modelo 1 (ECM)



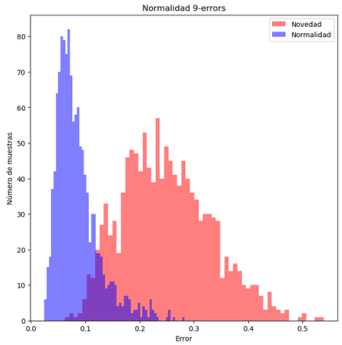
(b) Curva ROC y AUC modelo 1



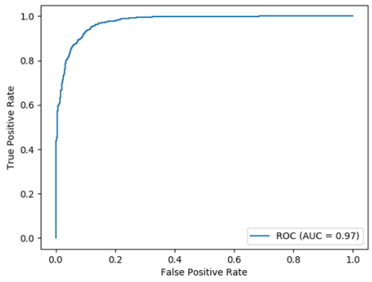
(c) Histograma modelo 2 (ISE)



(d) Curva ROC y AUC modelo 2



(e) Histograma modelo 9 (ISE+ECM)



(f) Curva ROC y AUC modelo 9

Figura 12: Ejemplos de detección de novedad (en fase de prueba) para los modelos 1, 2 y 9. La primer columna muestra los histogramas de los errores de reconstrucción obtenidos, en azul los errores de reconstrucción de las muestras normales y en rojo el de las muestras novedosas. En la segunda columna las correspondientes curvas ROC y AUCs obtenidos.

activación pueden ser la base para el calculo de uno de los atributos más importantes, la localización de la novedad en la imagen.

En el estado del arte se han propuesto distintos algoritmos para la explicación de redes de convolución. Tres de ellos son:

- Propagación guiada hacia atrás (PGA) [50]. Trabaja sobre modelos conformados solo por capas de convolución y realiza propagación hacia atrás a partir de la capa de interés. Este método resalta detalles a nivel del espacio de pixeles.

- Mapas de activación de clase ponderado por gradiente (Grad-Cam, por sus siglas en inglés)[51]. Utiliza los gradientes de una clase con respecto al mapa de características de una capa. Aplica agrupamiento de promedio global a través del cual se captura la importancia de un mapa de características para la clase. Finalmente una función ReLU es aplicada sobre la combinación lineal de los pesos calculados y los mapas de atributos de la capa.

- Grad-Cam guiado [51]. Versión del algoritmo Grad-Cam que incorpora las operaciones de deconvolución y propagación hacia atrás realizadas en PGA con el fin de producir un mapa de activación discrimativo, pero con detalles a nivel de pixel.

Las dos versiones de Grad-Cam son ponderadas por los pesos de la clase, es decir es un método discriminativo cuyo mapa de activación depende de los pesos de la clase a la que pertenece la entrada. Estos métodos parten de la existencia de un modelo clasificador donde existen al menos dos clases explícitas capturadas por el modelo. Por otro lado, PGA es un método que produce visualizaciones en el espacio de pixeles y no es discriminativo.

El siguiente experimento exploratorio se realizó para verificar si un clasificador básico es capaz de proveer un mapa de activación que permita localizar una instancia de novedad en la imagen al aplicar los algoritmos mencionados previamente. La Figura 13 muestra la estructura del clasificador que se utilizó para la prueba.

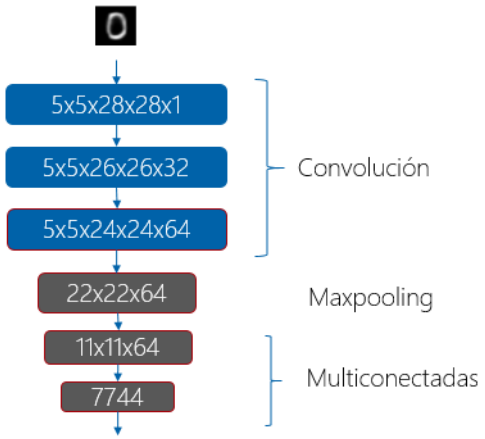


Figura 13: Clasificador de convolución de dos clases (0 y el resto 1-9 de la base de datos Mnist) utilizado para pruebas de localización con mapas de activación.

5.2.1. Resultados

En la Figura 14 se muestran ejemplos de la salida de los tres algoritmos Grad-cam (columna 2), propagación guiada hacia atrás (columna 3) y Grad-Cam guiado (columna 4). Las imágenes de entrada se muestran en la columna 1. Las salidas corresponden al mapa de activación de la tercer capa de convolución del clasificador descrito en la Figura 13. La muestra cero del renglón 1 de la imagen corresponde a una muestra de una clase conocida por el clasificador (0) y como se espera el mapa de activación obtenido por los algoritmos muestra activaciones consistentes con la entrada para los tres algoritmos. El resto de clases (J, A y E) son muestras novedosas que tomadas de la base de datos notMnist [52]. Esta base de datos esta conformada por imágenes de letras con mayores distorsiones y variaciones que mnist. Para las clases J y A, el mapa de activación de salida de los tres algoritmos se muestra consistente también como si se tratara de muestras de clases conocidas por el clasificador. Sin embargo la muestra novedosa de la clase E, solo presenta un mapa de activación aceptable (para fines de localización) con el algoritmo PGA. Es posible que se deba a que la muestra E, tiene un fondo mixto y el cuerpo de la letra esta a la inversa que el de todas las demás muestras (el cuerpo de la letra es negro). Por otro

lado, PGA no esta condicionado por los pesos de la clase detectada para el calculo del mapa de activación correspondiente.

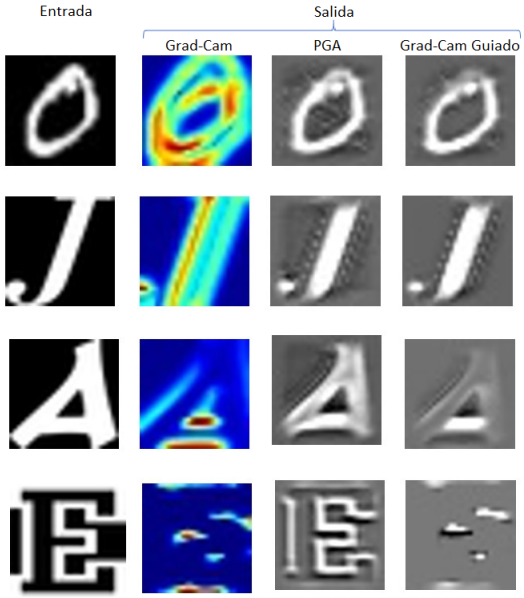


Figura 14: Ejemplos de visualización de mapas de activación de clase para el modelo descrito en la Figura 13. Primer columna corresponde a la entrada y el resto de columnas corresponde a las salidas obtenida al aplicar Grad-Cam, PGA y Grad-Cam guiado en la última capa de convolución del modelo. El primer renglón corresponde a una de las clases con las que fue entrenada el modelo (0), el resto son instancias de clases desconocidas (J,A y E)

PGA muestra ser un método que puede ser de utilidad para determinar la localización de una muestra novedosa. En un escenario hipotético en el que E fuera detectada como novedad, las dos versiones de Grad-Cam no serían de capaces de determinar la ubicación de E en la imagen, PGA sí. Aparentemente el hecho de que no se consideren los pesos de la clase en el cálculo del mapa de activación favorece que para imágenes con patrones diferentes a los que aprendió el clasificador no se inhiban las activaciones en el mapa. Más análisis y experimentos se llevarán acabo para confirmar lo observado.

6. Observaciones Finales

El presente trabajo de investigación propone el desarrollo de un enfoque basado en aprendizaje profundo para aprender una representación que capture la estructura contenida en imágenes y videos de entornos visuales dinámicos y con múltiples instancias como los que se encuentran en entornos de manejo autónomo. Con esta representación, se busca detectar instancias de novedad visual y los atributos que la describen. Resultados preliminares presentados muestran que un autocodificador de convolución con un entrenamiento adversario es capaz de obtener dicha representación. En el experimento descrito se utilizó una función de pérdida compuesta por ECM e ISE ponderado para el autocodificador. Los siguientes pasos consisten en mejorar la red para soportar otras bases de datos mientras se mantiene un rendimiento competitivo. Por otro lado, en términos arquitectónicos se realizará el análisis de métodos de clasificación no supervisada para la detección de novedad, bloque que recibirá como entrada las características obtenidas por el autocodificador y el error de reconstrucción correspondiente. También se presentó un experimento exploratorio que muestra como un modelo clasificador auxiliar podría ser utilizado a la par de un algoritmo para la obtención de mapas de activación. De esta forma, técnicas de explicación de redes profundas pueden ser la base de un enfoque para identificar uno de los atributos más importantes de las instancias de novedad, su ubicación en la imagen. La información provista por los mapas de activación muestran que pueden ser utilizados para este fin y serán estudiados con mayor profundidad en futuras actividades.

Referencias

[1] D. Watzenig and M. Horn, *Automated Driving*, vol. 4 of *10*. Springer, 1 ed., 7 2017. An optional note.

[2] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Processing*, vol. 99, pp. 215 – 249, 2014.

[3] A. Mateus, D. Ribeiro, P. Miraldo, and J. C. Nascimento, "Efficient and robust pedestrian detection using deep learning for human-aware navigation," *Robotics and Autonomous Systems*, vol. 113, pp. 23 – 37, 2019.

[4] Q. Zou, H. Jiang, Q. Dai, Y. Yue, L. Chen, and Q. Wang, "Robust lane detection from continuous driving scenes using deep neural networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 1, pp. 41–54, 2020.

[5] A. Dairi, F. Harrou, M. Senouci, and Y. Sun, "Unsupervised obstacle detection in driving environments using deep-learning-based stereovision," *Robotics and Autonomous Systems*, vol. 100, pp. 287 – 301, 2018.

[6] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, "End to end learning for self-driving cars," *CoRR*, vol. abs/1604.07316, 2016.

[7] W. Schwarting, J. Alonso-Mora, and D. Rus, "Planning and decision-making for autonomous vehicles," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 1, no. 1, pp. 187–210, 2018.

[8] W. Biever, L. Angell, and S. Seaman, "Automated driving system collisions: Early lessons," *Human Factors*, vol. 62, no. 2, pp. 249–259, 2020. PMID: 31502899.

[9] D. Hawkins, *Identification of ourliers*. Chapman and hall, 1980.

[10] X. Song, M. Wu, C. Jermaine, and S. Ranka, "Conditional anomaly detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 5, pp. 631–645, 2007.

[11] D.-Y. Yeung and Y. Ding, "Host-based intrusion detection using dynamic and static behavioral models," *Pattern Recognition*, vol. 36, no. 1, pp. 229 – 243, 2003.

[12] H. Liu, J. Li, Y. Wu, and Y. Fu, "Clustering with outlier removal," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2019.

[13] P. Piro, R. Nock, W. Bel haj ali, F. Nielsen, and M. Barlaud, *Boosting k-Nearest Neighbors Classification*, pp. 341–375. 01 2013.

[14] F. Angiulli and C. Pizzuti, "Fast outlier detection in high dimensional spaces," in *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery*, PKDD '02, (Berlin, Heidelberg), p. 15–26, Springer-Verlag, 2002.

[15] M. Markou and S. Singh, "Novelty detection: a review—part 2:: neural network based approaches," *Signal Processing*, vol. 83, no. 12, pp. 2499 – 2521, 2003.

[16] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning Workshop - Volume 27*, UTLW'11, p. 17–37, JMLR.org, 2011.

[17] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[18] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. The MIT Press, 2016.

[19] Y. Lecun, "Generalization and network design strategies," in *Connectionism in perspective*, Elsevier, 1989.

[20] C. C. Aggarwal, *Neural Networks and Deep Learning*. Cham: Springer, 2018.

[21] S. Marsland, J. Shapiro, and U. Nehmzow, "A self-organising network that grows when required," *Neural Networks*, vol. 15, no. 8, pp. 1041 – 1058, 2002.

[22] S. Marsland, U. Nehmzow, and J. Shapiro, "On-line novelty detection for autonomous mobile robots," *Robotics and Autonomous Systems*, vol. 51, no. 2, pp. 191 – 206, 2005.

[23] G. A. Carpenter and S. Grossberg, "A massively parallel architecture for a self-organizing neural pattern recognition machine," *Computer Vision, Graphics, and Image Processing*, vol. 37, no. 1, pp. 54 – 115, 1987.

[24] I. Nunes, D. Hernane, R. Andrade, L. Liboni, and S. Franco, *Artificial Neural Networks*. Springer, 2017.

[25] L. Pitonakova and S. Bullock, "The robustness-fidelity trade-off in grow when required neural networks performing continuous novelty detection," *Neural Networks*, vol. 122, pp. 183 – 195, 2020.

[26] B. Sofman, J. Andrew Bagnell, and A. Stentz, "Anytime online novelty detection for vehicle safeguarding," in *2010 IEEE International Conference on Robotics and Automation*, pp. 1247–1254, 2010.

[27] J. Kivinen, A. J. Smola, and R. C. Williamson, "Online learning with kernels," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2165–2176, 2004.

[28] P. Ross, A. English, D. Ball, B. Upcroft, and P. Corke, "Online novelty-based visual obstacle detection for field robotics," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3935–3940, 2015.

[29] P. Ross, A. English, and D. Ball, "Online covariance estimation for novelty-based visual obstacle detection," *Journal of Field Robotics*, vol. 34, no. 8, pp. 1469–1488, 2017.

[30] F. Redhead, S. Snow, D. Vyas, O. Bawden, R. Russell, T. Perez, and M. Brereton, "Bringing the farmer perspective to agricultural robots," 04 2015.

[31] Y. Xia, X. Cao, F. Wen, G. Hua, and J. Sun, "Learning discriminative reconstructions for unsupervised outlier removal," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1511–1519, 2015.

[32] S. Ramos, S. Gehrig, P. Pinggera, U. Franke, and C. Rother, "Detecting unexpected obstacles for self-driving cars: Fusing deep learning and geometric modeling," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1025–1032, June 2017.

[33] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

[34] D. Pfeiffer, F. Erbs, and U. Franke, "Pixels, stixels, and objects," 10 2012.

[35] P. Pinggera, S. Ramos, S. Gehrig, U. Franke, C. Rother, and R. Mester, "Lost and found: Detecting small road hazards for self-driving vehicles," *CoRR*, vol. abs/1609.04653, 2016.

[36] V. Chen, M. Yoon, and Z. Shao, "Novelty detection via network saliency in visual-based deep learning," *CoRR*, vol. abs/1906.03685, 2019.

[37] M. Bojarski, A. Choromanska, K. Choromanski, B. Firner, L. J. Ackel, U. Muller, P. Yeres, and K. Zieba, "Visualbackprop: Efficient visualization of cnns for autonomous driving," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4701–4708, 2018.

[38] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[39] Udacity, "Udacity self-driving car dataset." `https://github.com/udacity/self-driving-car/tree/master/datasets`. Accessed: 01-08-2020.

[40] A. Amini, W. Schwarting, G. Rosman, B. Araki, S. Karaman, and D. Rus, "Variational autoencoder for end-to-end control of autonomous driving with novelty detection and training de-biasing," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 568–575, Oct 2018.

[41] C. Richter and N. Roy, "Safe visual navigation via deep learning and novelty detection," in *Robotics: Science and Systems*, 2017.

[42] P. Perera and V. M. Patel, "Deep transfer learning for multiple class novelty detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[43] P. Perera, R. Nallapati, and B. Xiang, "Ocgan: One-class novelty detection using gans with constrained latent representations," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[44] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3379–3388, 2018.

[45] A. Dosovitskiy, G. Ros, F. Codevilla, A. López, and V. Koltun, "Carla: An open urban driving simulator," in *CoRL*, 2017.

[46] F. U. Berlin, "Autonomos model." `https://github.com/AutoModelCar/AutoModelCarWiki/wiki`. Accessed: 01-08-2020.

[47] S. Agarwal, A. Vora, G. Pandey, W. Williams, H. Kourous, and J. McBride, "Ford multi-av seasonal dataset," 03 2020.

[48] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[49] L. Deng, "The mnist database of handwritten digit images for machine learning research [best of the web]," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[50] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller, "Striving for simplicity: The all convolutional net," *CoRR*, vol. abs/1412.6806, 2015.

[51] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.

[52] Y. Bulatov, "notmnist." `http://yaroslavvb.blogspot.com/2011/09/notmnist-dataset.html`. Accessed: 01-08-2020.