Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ingeniería Mecánica Eléctrica

DISEÑO DE INVESTIGACIÓN DE PROTOTIPO PARA LA DETECCIÓN DE INFRACCIONES POR UTILIZAR BASUREROS CLANDESTINOS EN CIUDAD DE GUATEMALA USANDO APRENDIZAJE AUTOMÁTICO

Josué Ricardo Slansky Rabanales Gómez
Asesorado por el Ing. Edwin Estuardo Zapeta Gómez

Guatemala, abril de 2022

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

DISEÑO DE INVESTIGACIÓN DE PROTOTIPO PARA LA DETECCIÓN DE INFRACCIONES POR UTILIZAR BASUREROS CLANDESTINOS EN CIUDAD DE GUATEMALA USANDO APRENDIZAJE AUTOMÁTICO

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

JOSUÉ RICARDO SLANSKY RABANALES GÓMEZ
ASESORADO POR EL ING. EDWIN ESTUARDO ZAPETA GÓMEZ

AL CONFERÍRSELE EL TÍTULO DE

INGENIERO ELECTRÓNICO

GUATEMALA, ABRIL DE 2022

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

NÓMINA DE JUNTA DIRECTIVA

DECANA	Inga. Aurelia Anabela Cordova Estrada
VOCAL I	Ing. José Francisco Gómez Rivera
VOCAL II	Ing. Mario Renato Escobedo Martínez
VOCAL III	Ing. José Milton de León Bran
VOCAL IV	Br. Kevin Armando Cruz Lorente
VOCAL V	Br. Fernando José Paz González
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

DECANA	Inga. Aurelia Anabela Cordova Estrada
EXAMINADOR	Ing. Guillermo Antonio Puente Romero
EXAMINADOR	Ing. Luis Eduardo Durán Córdova
EXAMINADOR	Ing. Francisco Javier González López
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

HONORABLE TRIBUNAL EXAMINADOR

En cumplimiento con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

DISEÑO DE INVESTIGACIÓN DE PROTOTIPO PARA LA DETECCIÓN DE INFRACCIONES POR UTILIZAR BASUREROS CLANDESTINOS EN CIUDAD DE GUATEMALA USANDO APRENDIZAJE AUTOMÁTICO

Tema que me fuera asignado por la Dirección de Escuela de Estudios de Postgrado con fecha 12 de enero de 2022.

Josué Ricardo Slansky Rabanales Gómez

ACTO QUE DEDICO A:

Mis padres	Ricardo Rabanales y Patricia Gómez, por su amor incondicional y apoyo.
Mi esposa	Monica Ortiz, por su paciencia, dedicación y amor a lo de este proceso, así como su apoyo constante para cumplir cada uno de mis sueños.
Mis hermanas	Por ser una fuente de admiración para mí.
Mi familia	Por siempre creer en mí y mostrarme un apoyo incondicional desde mi niñez.

AGRADECIMIENTOS A:

Universidad de San Carlos de Guatemala	Por permitirme formarme como profesional y persona.
Mis amigos	Del Departamento de Matemática y compañeros de clase, por todas las buenas experiencias vividas y la ayuda que recibí a lo largo de estos años.
Catedráticos	Por brindarme sus conocimientos libremente.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES

FIGURAS

TABLAS

LISTA DE SÍMBOLOS

Símbolo	Significado
exp(x)	Función exponencial aplicada a un tensor
$\sigma(x)$	Función sigmoide de x
*	Producto de Hadamard
$tanh\ (x)$	Tangente hiperbólica de x
$\tilde{C}^{<t>}$	Tensor en el paso t de una secuencia

GLOSARIO

Dlib	Librería de programación para las aplicaciones de visión por computadora.
GPU	Unidad de Procesamiento de Gráficos.
Hiperparámetros	Parámetros libres de un modelo de aprendizaje automático que pueden ajustarse libremente y que no dependen de las entradas o salidas del modelo.
IoT	Internet de las cosas.
OpenCV	Librería de programación para las aplicaciones de visión por computadora.
Streaming	Flujo de datos que son producidos y consumidos en el momento.

RESUMEN

En el presente trabajo, se describe el diseño de investigación para la elaboración de un prototipo de herramienta que sea capaz de clasificar vídeos para determinar si una persona está tirando basura en un lugar inadecuado o no. Se detallan todos los pasos necesarios y la metodología que se utilizará para que el trabajo de investigación tenga éxito. Como resultado, se obtuvo un plan robusto que al llevarse a la práctica se espera que concluya en un prototipo de producto funcional y que sirva como base para mejorarse en futuros trabajos de investigación, tomando en cuenta que este es un problema sumamente complejo.

Se concluyó que es posible cumplir con los objetivos detallados en este diseño de investigación, en una cantidad de tiempo razonable y con los recursos disponibles.

1. INTRODUCCIÓN

En los últimos años se utiliza cada vez más la visión por computadora para solventar problemas que antes parecía imposible que una computadora pudiera realizar. En este trabajo se propone utilizar las técnicas de aprendizaje automático y visión por computadora que se han desarrollado en los últimos años para crear una herramienta que sea capaz de detectar personas que utilizan basureros clandestinos.

Se espera que el diseño de esta herramienta pueda ser de utilidad para las autoridades que intentan hacer cumplir las leyes ambientales del país ya que la contaminación producida por tales basureros clandestinos es un problema creciente y que impacta grandemente la salud de la comunidad y del planeta.

Se espera que el prototipo sea capaz de detectar personas que van a tirar basura a lugares específicos y se pueda alertar a las autoridades correspondientes para que, de esa manera, si es posible, se pueda identificar a la persona, por el número de placa de su auto, por ejemplo, y se pueda emitir la multa correspondiente. El prototipo no será un producto final, sino más bien una herramienta que sirva como demostración de que tal tarea es posible y ejecutable en un proyecto a gran escala. Se espera también que el algoritmo sea funcional para distintos tipos de lugares, personas, momentos del día y ángulos desde los cuales se capturan los videos.

Las líneas de investigación de la maestría en las cuales se realiza el presente trabajo son análisis de datos e internet de las cosas, pudiéndose considerar este un proyecto de ciudad inteligente.

Actualmente se ha explorado la posibilidad detectar personas que tiran basura, pero de manera muy específica como tirar la basura desde la ventana del auto o utilizando técnicas diferentes a las propuestas en este trabajo. En esta investigación se espera que se pueda obtener una exactitud lo suficientemente alta como para que el modelo sea útil para aplicaciones que funcionen en el mundo real.

El informe final explicará las técnicas utilizadas para realizar la aplicación, así como el diseño del modelo de aprendizaje automático escogido para resolver el problema. También se explicará la metodología utilizada para la recolección de datos y desarrollo de la herramienta.

2. ANTECEDENTES

Una fuente importante es la de Rawat y Wang (2017), con su artículo Redes neuronales convolucionales profundas para la clasificación de imágenes: Un repaso comprensivo, el cual aparece en la revista Neural Computation por MIT Press Direct.

Este trabajo pretende brindar una reseña comprensiva de las redes neuronales convolucionales para el procesamiento de imágenes, así como brindar detalles acerca de su avance a través del tiempo, describir desafíos actuales con esta clase de arquitectura de red neuronal y recapitular temas abiertos y tendencias asociadas proveyendo de esta manera recomendaciones de direcciones en las cuales hacer investigación para una futura exploración.

En el artículo Rawat y Wang (2017), sintetizan los distintos tipos de arquitecturas de redes neuronales convolucionales que más impacto han tenido en la visión por computadora, además de las distintas técnicas utilizadas para hacer que estos modelos consiguieron mejores resultados a lo largo del tiempo.

También se resumen algunos de los últimos avances en cuanto a visión por computadora como los esfuerzos que se hacen por encontrar una explicación del éxito que las redes convolucionales tienen en tareas de clasificación de imágenes, los avances que se han hecho a fin de que sea más fácil utilizar estos modelos para la predicción en hardware menos potente como teléfonos móviles, microcontroladores o incluso a FPGA. Describe además algunos puntos débiles de estas arquitecturas como que pueden ser engañadas al encontrarse con ejemplos adversarios que en general es fácil para los humanos poder distinguir,

o el distinguir varios objetos y poder etiquetarlos en una misma imagen como suele ser el caso del mundo real.

Este artículo provee información importante acerca de cómo resolver problemas que se relacionan con la visión computacional y resume muy bien las capacidades actuales de este tipo de aplicaciones, brindando de esa manera ideas de cómo se pueden resolver problemas que se relacionan con el procesamiento de imágenes.

La siguiente fuente es un trabajo por Han, Zhang, Zhuo, Huang y Zhang (2017), quienes realizaron la investigación Yendo más profundo con *ConvNets* de dos transmisiones para el reconocimiento de acciones en vigilancia por vídeo, publicada en la revista Pattern Recognition Letters que puede ser traducido como Cartas en Reconocimiento de Patrones.

El objetivo de este estudio era poder encontrar un método que tuviera una mejor exactitud que los desarrollados hasta ese momento y también desarrollar técnicas para aumentar los datos de video.

La técnica utilizada en la investigación utiliza dos entradas distintas para dos redes neuronales convolucionales independientes las cuales procesan las características de la imagen RGB por una parte y del flujo óptico del video en la otra red neuronal, después de lo cual se combinan las características de ambos y se realiza una última inferencia en la cual se da el puntaje final para la clasificación de dicho vídeo. Esta técnica demostró sobrepasar al estado del arte en reconocimiento de acciones en vídeo que se tenía hasta ese momento.

Esta investigación es importante porque provee un método sencillo y altamente eficiente y efectivo para realizar reconocimiento de vídeo. Demuestra

además que, con un poco de procesamiento del conjunto de datos previo a introducirse al modelo, pueden alcanzarse mejores resultados en tareas de clasificación de acciones.

Se continúa entonces con el estudio realizado por Carreira y Zisserman (2018), del departamento de Ciencias de la Ingeniería de la Universidad de Oxford y de DeepMind de Google, en su trabajo realizado ¿A dónde vas, reconocimiento de acciones? Un nuevo modelo y el conjunto de datos cinética.

Este artículo científico tiene como objetivo poder evaluar la eficacia de los modelos actuales en cuanto a reconocimiento de acciones de humanos en colecciones de datos de vídeos. También pretende introducir un nuevo modelo llamado I3D, y un conjunto de datos llamados cinética que contienen mucha más información, clases y un nivel de dificultad alto y que fueron tomados de vídeos de YouTube.

Se muestra el resultado de distintos experimentos en distintos conjuntos de datos a modo de comparativa a fin de comprender qué modelos de aprendizaje profundo funcionan mejor para el reconocimiento de acciones y bajo qué circunstancias. Por ejemplo, se entrena a los modelos en el nuevo conjunto de datos cinética y después se entrena en otros conjuntos de datos bien conocidos a fin de evaluar el impacto que tiene un primer entrenamiento. Además, se prueba también entrenar los diferentes modelos en una colección de datos de imágenes estáticas grande. En los distintos experimentos se comprueba la eficacia que tiene el nuevo modelo, mejorando los resultados de modelos anteriores, sin embargo, también muestra que algunos modelos más sencillos podrían llegar a ser suficientemente buenos para algunas aplicaciones ya que también puntúan alto en las pruebas de precisión.

El artículo presenta posibles soluciones a la parte del sistema que involucra la predicción y clasificación del tipo de acciones que tienen que ver con el botar basura en lugares no autorizados. Teniendo la cantidad suficiente de datos se pueden aplicar técnicas de aprendizaje profundo mencionadas en este papel para la parte principal de este proyecto.

Otro trabajo corresponde a Flores (2018) titulado Modelo de Gestión para la Erradicación de Basureros Clandestinos, Estudio de dos casos en el Municipio de Villanueva, el cual sirvió como su trabajo de graduación de la Maestría en Artes en Ingeniería para el Desarrollo Municipal en la Universidad de San Carlos de Guatemala. Tal como menciona el título, el estudio se concentra en el municipio de Villanueva, sin embargo, también contiene datos importantes del tema para todo el departamento de Guatemala.

Como objetivos de este segundo trabajo, la autora, Flores (2018), propone "describir un modelo de gestión para la erradicación de basureros clandestinos, a partir del estudio de dos casos emblemáticos (por el volumen de sus desechos y demás características) en el municipio de Villanueva" (p. XXV).

Flores (2018) determina como objetivos específicos:

- Establecer la importancia de la erradicación de basureros clandestinos para el desarrollo municipal en el municipio de Villanueva.
- Definir los recibos y desechos sólidos, así como su relación con el ornato y espacio público, entre otros.
- Identificar la ubicación y cantidad de basureros clandestinos en el municipio de Villanueva. (p. XXV)

En su trabajo Flores (2018) describe características importantes de los basureros clandestinos a lo largo del municipio lo cual nos da una idea de lo diversos que estos pueden ser, por ejemplo, hay basureros donde la mayor cantidad de basura es de tipo orgánica y hay otros donde existe un gran porcentaje de basura inorgánica o especial. También describe algunas de las razones por las cuales las personas acuden a desechar su basura a través de medios no oficiales, desde características culturales hasta falta de una legislación adecuada o apoyo a la misma.

Por último, Flores (2018) propone un “modelo de gestión para la erradicación de basureros clandestinos en el municipio de Villanueva” (p.49). Teniendo este, 3 columnas clave, que son: La legislación, evaluación y normalización del servicio, y la participación y apoderamiento de la ciudadanía.

Este trabajo de desarrollo municipal apoya la afirmación de que existe un gran problema en las áreas urbanas como consecuencia de los basureros clandestinos ya que si bien el estudio está delimitado al municipio de Villanueva hay algunos datos que pueden generalizarse a todo el departamento de Guatemala y aún a todo el país. Por ejemplo, una de las estadísticas que Flores comparte dice que la zona de la ciudad capital con una mayor cantidad de basureros clandestinos es la zona 12.

La siguiente investigación es por Boyko, Basystiuk y Shakhovska (2018) de la Lviv Polytechnic National University de Ucrania. Su artículo presentado en la segunda conferencia internacional de transmisión de datos y procesamientos de la IEEE cuyo título traducido al español es Evaluación de desempeño y comparación de software para el reconocimiento de rostros, basado en las librerías *Dlib y OpenCV*.

El objetivo de dicha investigación era comparar el desempeño de dos de las librerías más famosas y utilizadas para la visión por computadora llamadas OpenCV y Dlib, tomando como caso de uso específicamente el reconocimiento de rostros. También, pretende evaluar las ventajas y desventajas de cada librería, y qué tan apropiadas son para su uso en proyectos que involucren sistemas de reconocimiento.

En el documento se enumeran distintas técnicas para el reconocimiento de rostros y también se desarrolla el código que debería implementarse con cada librería a fin de obtener un clasificador de rostros. Después, se procede a realizar experimentos para poder comprender mejor el desempeño de cada una de estas librerías en las tareas asignadas. se encuentra que OpenCV es órdenes de magnitud más eficiente que Dlib.

Estas conclusiones son importantes ya que utilizar una librería especializada normalmente es necesario para realizar prototipos y productos más rápidamente, y este papel nos da una idea de qué software podría ser el más apropiado para tareas relacionadas con visión por computadora. En este caso se concluyó que OpenCV suele ser una opción más favorable ya que es más ampliamente utilizado y con un desempeño superior

Se continúa entonces con la referencia del trabajo de Shou *et. al.* (2018), que desarrollan el tema Detección en línea del comienzo de acciones de vídeos transmitidos y sin cortar, publicado en la conferencia europea de visión por computadora.

El objetivo del trabajo era encontrar un algoritmo para detectar el principio de una acción con gran precisión categórica y baja latencia en vídeos siendo transmitidos en tiempo real y sin cortar. La publicación pretende mostrar cómo

este método innovador es superior a los publicados anteriormente.

Este trabajo hace ver la dificultad que tiene el reconocimiento de vídeo en tiempo real y se proponen 3 métodos novedosos para poder entrenar modelos de reconocimiento de vídeo en línea, siendo el primero la generación de muestras negativas basado en redes generativas adversarias para poder distinguir fondos ambiguos, un segundo método es modelar de manera explícita la consistencia temporal entre datos alrededor del inicio y el resto de la acción, y el tercero es una estrategia adaptativa de muestreo para manejar el esparcimiento de datos de entrenamiento. Como conclusión del trabajo se tiene que estas técnicas pueden mejorar considerablemente la precisión de la clasificación de vídeos en línea.

Estas conclusiones y técnicas proveen ideas de cómo implementar sistemas de reconocimiento de acciones en línea.

También puede hacerse referencia al artículo científico realizado por Wu, Zaheer, Hu, Manmatha, Smola y Krähenbühl (2018), titulado *Compressed Video Action Recognition*.

El objetivo de la investigación era encontrar un método de reconocimiento de acciones en vídeo que fuera más eficiente y eficaz que los publicados hasta ese momento utilizando el vídeo en forma comprimida.

En su investigación, lograron mejorar la exactitud para predecir etiquetas de vídeos de 2 conjuntos de datos muy famosos. En el papel explican la forma en que lograron a entrenar el modelo utilizando fotogramas en su forma comprimida y de esa manera eliminar esta redundancia temporal que los vídeos suelen tener debido a que muchos de los fotogramas son iguales unos con otros.

Este artículo es importante porque detalla un método mediante el cual se puede lograr el reconocimiento de acciones en video y es posible tomar ideas de dichas técnicas para alguna aplicación de visión por computadora que se quiera realizar.

Otro importante trabajo para tomar en cuenta es el artículo científico por Ullah, Ahmad, Muhammad, Sajjad y Baik (2018), quiénes hicieron la publicación Reconocimiento de acciones en secuencias de vídeo utilizando LSTM direccionales con características de CNN, publicado en la IEEE Xplore.

El objetivo principal de este papel era experimentar la combinación de redes neuronales convolucionales o CNN, y LSTM para la clasificación de acciones en secuencias de vídeo y así obtener mejores resultados en la clasificación de vídeos de algunos conjuntos de datos específicos.

En el trabajo se explica la arquitectura de la red neuronal utilizada, la cual es un grupo de capas convolucionales que extraen algunas propiedades de cada fotograma y que después sirven de entrada para alimentar una red de LSTM de doble dirección, la cual extrae propiedades de la dimensión temporal del vídeo para así dar una inferencia acerca de la acción mostrada en el vídeo. Esta combinación sencilla de dos arquitecturas sumamente conocidas demostró dar resultados sumamente precisos y exactos en cada colección de datos en los cuales ésta se probó.

Esta es otra arquitectura sencilla y que ha demostrado en el pasado ser lo suficientemente exacta, así como computacionalmente eficiente, para reconocer acciones, lo cual la hace un candidato para ser utilizada en aplicaciones en tiempo real o en las cuales se deba procesar varias entradas de manera concurrente.

La siguiente referencia es el trabajo Detección de acciones de vertido de basura basado en visión para una plataforma de vigilancia en el mundo real, publicado en el ETRI Journal de la República de Corea. Los autores de dicho trabajo son Yun, Kwon, Oh, Moon y Park (2019).

Principalmente el objetivo de este trabajo fue obtener un modelo de visión por computadora que pudiera detectar personas que dejan basura en un lugar, proponiendo un método diferente a otros utilizados actualmente para el reconocimiento de acciones.

En este trabajo se logra el reconocimiento de acciones utilizando métodos diferentes a los del aprendizaje profundo, pero hace ver algunas de las dificultades que se tienen al utilizar redes neuronales profundas en un sistema como este. En él se describe un método complejo de visión por computadora que es capaz de reconocer personas tirando basura y con una exactitud suficientemente buena.

Este trabajo es importante para la presente investigación ya que demuestra que tales sistemas son posibles y mejorables. Además, es una ayuda para descartar técnicas ya utilizadas en el pasado y aprender de tales métodos a fin de desarrollar otras mejores.

La décima y última referencia es el trabajo titulado *Spatiotemporal neural networks for action recognition based on joint loss*. Investigación realizada por Jing, Wei, Sun., Sun, Zheng (2019) y publicada en la revista Computación Neuronal y Aplicaciones.

El objetivo de este trabajo era proponer una nueva función de pérdida que beneficiara a redes neuronales con tareas espacio temporales a obtener mejores

resultados.

En la investigación los autores utilizan dos redes convolucionales independientes, en una de ellas se tiene como entrada los fotogramas crudos y en otra de ellas se introduce el flujo óptico del vídeo, después de lo cual las características extraídas de cada modelo se combinan y se introducen a otra red neuronal de memoria de corto y largo plazo para por fin hacer la inferencia y utilizando la nueva función de pérdida, entrenar al modelo. El modelo demuestra tener un mejor desempeño cuando se utiliza la nueva función de pérdida conjunta.

Los resultados de esta investigación científica son importantes debido a que pueden mejorar la clasificación de acciones en vídeo haciendo algo sencillo como cambiar la función de pérdida que se utilice para entrenar a los modelos de aprendizaje profundo.

3. PLANTEAMIENTO DEL PROBLEMA

3.1. Contexto general

Los basureros clandestinos son lugares donde la población tira basura, como desechos sólidos, pero que su propósito no es servir como basurero ya que no cuentan con los estándares necesarios para que sean seguros para el medio ambiente y la población que se encuentra cerca. Por lo general estos basureros clandestinos son terrenos baldíos y lugares cercanos a puentes, como barrancos, o ríos. En la actualidad existen cientos de estos vertederos en la República de Guatemala.

El Gobierno está constantemente esforzándose por eliminarlos. Según reporta el Ministerio de ambiente y recursos naturales, entre el año 2019 y 2020 se erradicaron cerca de 300 basureros clandestinos a lo largo de todo el país. De esos 300 vertederos 121 se encontraban en Petén. Es importante aclarar que existen multas y acciones legales que se pueden realizar en contra de las personas que tiren basura en lugares no autorizados, sin embargo, es un problema que está lejos de erradicarse.

3.2. Descripción del problema

Aunque el Gobierno invierte en educación ambiental para prevenir este problema aún es insuficiente para poder resolverlo. Es importante sobre todo que se eduque a las poblaciones de más bajos recursos ya que es por lo general cerca de estos lugares dónde se pueden encontrar dichos basureros clandestinos. Además, las personas necesitan hacer conciencia en cómo les

afecta a ellos, al medio ambiente y a su descendencia el hecho de aumentar la contaminación del país por no respetar las leyes ambientales, pero es difícil para las personas tener dicha conciencia si no ven consecuencias más inmediatas a sus acciones.

Algo que suele pasar con estos basureros clandestinos es que cuando las autoridades los identifican se coloca un letrero o anuncio advirtiendo de las consecuencias de tirar basura en ese lugar, que por lo general son multas por cierta cantidad de quetzales. A pesar de estos carteles o advertencias, ya que no hay un monitoreo constante en dichos lugares, la gente sigue tirando basura de forma ilegal. Por mucho presupuesto que se utilice de parte del Gobierno es obviamente imposible monitorear tantos lugares del país las 24 horas del día y los 7 días de la semana.

Como estos lugares no están diseñados para servir como basureros son una fuente de contaminación constante para el medio ambiente. Por ejemplo, son fuentes de contaminación del suelo y contaminación del agua. De hecho, la contaminación del suelo suele tener como consecuencia la contaminación del agua porque existen desechos, como los químicos, por ejemplo, que se filtran a través del suelo y llegan hasta los canales subterráneos de agua. Además, a veces, esto constituye una violación a la propiedad privada ya que estos terrenos en ocasiones tienen un dueño particular. Estos lugares también suelen tener un mayor riesgo de incendio.

¿Por qué la contaminación del ambiente producida por los basureros clandestinos es un problema? Ya que la contaminación del suelo se traduce en una contaminación del agua, las personas que tienen contacto con ella pueden adquirir un sinnúmero de enfermedades. También afecta a la flora y fauna del lugar ya que, por ejemplo, las aves suelen comer de la basura del lugar dañando

también gravemente su salud y aún hasta afectando sus ciclos de migración. Además, este tipo de contaminación arruina el paisaje, el cual es importante aún en la salud mental de las personas que viven cerca o pasan por el lugar.

En el caso de las personas cuya propiedad privada ha sido violada suelen tener consecuencias para su economía. Un terreno que está muy contaminado pierde mucho valor, esto sobre todo debido a que ya no puede dársele la variedad de usos que normalmente éste podría tener.

Supongamos que una persona tiene un terreno que ha sido contaminado por años, materiales como las baterías (que suelen tirarse en este tipo de basureros clandestinos) desprenden, al contacto con la lluvia, una cantidad significativa de químicos altamente tóxicos como litio, mercurio, plomo y cadmio. Por tanto, este terreno no podría utilizarse para la agricultura o productos de consumo humano. Además, esta contaminación puede durar literalmente cientos de años. Por tanto, el impacto económico que tiene la contaminación de estos basureros en la tierra es alto, ya sea para particulares o si el dueño de la tierra es el mismo estado.

3.3. Formulación del problema

A continuación, se formulan las preguntas que servirán para detallar el problema en cuestión.

3.3.1. Pregunta central

¿Cómo detectar infracciones por utilizar basureros clandestinos en la Ciudad de Guatemala utilizando técnicas de aprendizaje automático?

3.3.2. Preguntas auxiliares

- ¿Qué algoritmos de aprendizaje automático pueden ser útiles para detectar personas realizando acciones específicas?

- ¿De los algoritmos para detectar y reconocer acciones, cuál es el más apropiado para detectar personas que desechan basura en lugares no autorizados?

- ¿Cómo se puede incorporar dicho algoritmo óptimo a un prototipo de herramienta de reconocimiento de acciones para la detección de personas que utilizan basureros clandestinos?

3.4. Delimitación del problema

El problema se resolverá utilizando datos en video de distintos lugares de la Ciudad de Guatemala. Dichos datos serán recopilados de distintas fuentes, como lo son la misma Municipalidad de Guatemala, así como videos que el autor mismo grabará con personas que simulan la acción de tirar basura en un basurero clandestino.

Además, cabe mencionar que a fin de verificar la factibilidad de la solución se realizará un prototipo de la herramienta que funcione utilizando una cámara y un servidor, así como una red local para la comunicación entre ellas. Este prototipo será de carácter sencillo y no pretende ser una versión final lista para poner en producción y monitorear las calles, sino más bien, una prueba de que dicha solución es posible.

4. JUSTIFICACIÓN

La realización de la presente investigación se justifica en las líneas de investigación del análisis de datos e IoT de la Maestría en Ingeniería para la Industria con Especialización en Ciencias de la Computación, ya que la solución al problema implica el uso de técnicas de aprendizaje automático y el uso de sensores, en este caso, cámaras.

A través de la presente investigación, se demostrará que es posible ayudar a identificar infractores que desechan su basura en basureros clandestinos con la ayuda de la tecnología, como cámaras y la red de internet, y el aprendizaje automático. Esto ayudará a que futuras investigaciones y desarrollos tengan una base y puedan pulir esta herramienta y de esa manera el problema de los basureros clandestinos disminuya con el tiempo en las áreas urbanas de Guatemala.

Al finalizar esta investigación se contará con un modelo de aprendizaje automático que sea capaz de identificar la acción de tirar basura en un lugar prohibido, así como el diseño de un sistema que pueda aprovechar dicho modelo. Ya que este es solamente un prototipo, no se obtendrá como resultado un producto final que pueda ponerse a trabajar en las calles de la ciudad inmediatamente, sino más bien este es una investigación que otros profesionales pueden tomar como base para ejecutar otros proyectos a gran escala relacionados a resolver el problema en cuestión.

Es importante hacer notar que dicho modelo sería reutilizable en distintos lugares, distintos momentos del día, y se espera que también sea totalmente

independiente de las personas que aparecen en el vídeo, e incluso del ángulo de la cámara con el que se capte el vídeo.

Esta investigación puede ser de gran beneficio para otros profesionales en el campo de las ciencias de la computación, así como para la sociedad en general, ya que el medio ambiente y su salud es algo que nos compete a todos. Para otros investigadores y profesionales, este trabajo servirá como base para que puedan seguir desarrollando y mejorando productos que ayuden a disminuir el problema de los basureros clandestinos en Guatemala y el resto del mundo. Será de especial atención para organizaciones gubernamentales y no gubernamentales, cuya labor sea preservar el medio ambiente y la salud de los ciudadanos guatemaltecos, debido a que les ayudará a regular y penalizar a las personas que no respeten la ley en cuanto a los lugares permitidos para desechar basura.

El presente trabajo es en realidad necesario debido al impacto que tiene el medio ambiente en nuestras vidas y en nuestra salud. La ley aparte de ser necesaria para una convivencia agradable y positiva en la sociedad tiene una componente educativo y pedagógico. Se espera que al poder ser más eficaces para poder hacer que ésta se cumpla con respecto a los lugares para desechar la basura, la población en general pueda ir tomando conciencia de lo importante que es preservar el medio ambiente en general y la urgencia que se tiene de cuidarlo.

5. OBJETIVOS

5.1. General

Crear un prototipo de una herramienta que sirva para la detección de infracciones por desechar basura en basureros clandestinos en la Ciudad de Guatemala.

5.2. Específicos

- Describir los diferentes algoritmos de aprendizaje automático que existen para el reconocimiento de acciones.

- Comparar los distintos algoritmos de reconocimiento de acciones realizando experimentos de entrenamiento en un set de datos con videos de personas utilizando basureros clandestinos para así seleccionar el modelo más apropiado para la tarea de detectar dicha acción.

- Implementar un prototipo de herramienta para la detección de personas que tiran basura en basureros clandestinos utilizando el algoritmo de reconocimiento de acciones seleccionado.

6. NECESIDADES A CUBRIR Y ESQUEMA DE SOLUCIÓN

Actualmente existe una gran cantidad de basureros clandestinos en la ciudad de Guatemala. Se espera que esta investigación ayude a disminuir la cantidad y tamaño de estos basureros clandestinos al proveer mecanismos de detección para infracciones realizadas por personas que tiran basura en estos lugares. Actualmente no se cuenta con un producto o sistema que ayude a las autoridades a detectar infracciones de este tipo de forma lo suficientemente autónoma cómo para realizar monitoreos a gran escala.

Existe una falta de conciencia y educación por parte de la población, así como mecanismos para monitorear constantemente lugares susceptibles a ser basureros clandestinos, por tanto, las personas siguen utilizando estos lugares de manera incorrecta. Es muy común que se localice el lugar en el que se encuentra un basurero clandestino por parte de las autoridades, pero por lo general no se puede saber qué personas lo han usado y de esa manera emitir multas que insten a la población a obedecer este tipo de leyes ambientales.

El poder emitir multas a los infractores que utilicen estos basureros es importante ya que disminuye la cantidad de personas que realizan este acto y también se educa a la población de la importancia que tiene el preservar el medio ambiente. Al tener un sistema de monitoreo constante para detectar infracciones de este tipo, ambas necesidades de educación para la población, como el tener una manera eficaz y eficiente de emitir multas, puede ser parcialmente cubierta.

Los lugares con basureros clandestinos y áreas cercanas suelen contar con problemas y necesidades diversas. Por ejemplo, suelen tener un riesgo mucho

más grande que otros lugares de tener un incendio significativo que cause graves problemas de salud, e incluso la muerte de ciudadanos que se encuentren cerca del lugar, así como pérdida de bienes materiales. También la contaminación del suelo puede resultar en contaminación del agua, la disminución y modificación de fauna y flora cercana, y diversos tipos de problemas de salud para las personas que se encuentren cerca.

Al tener una manera de detectar a las personas que tiran basura en estos lugares no autorizados, el tamaño de los basureros puede disminuir, reduciendo significativamente cada uno de estos problemas. Idealmente esta detección de infracciones puede hacerse de manera temprana para que el basurero clandestino no crezca demasiado y después sea más difícil limpiarlo y eliminarlo por completo.

A fin de resolver el problema, primero se propone localizar algunos basureros clandestinos que se encuentren en la ciudad de Guatemala para poder grabar vídeos de personas tirando basura en ellos o acceder a videos grabados previamente por la municipalidad de la Ciudad de Guatemala. Se grabarán vídeos de personas que están tirando basura en dichos lugares no autorizados, o si fuera necesario en propiedad privada a modo de actuación. Dichos videos deben grabarse en distintos momentos del día, con distintas personas, en distintos lugares, y deben ser en número suficiente, para lograr que el algoritmo pueda aprender lo suficiente de ellos para generalizar.

Se entrenarán varios modelos de aprendizaje profundo para que sepan distinguir entre la acción de tirar basura y cualquier otra acción utilizando los videos previamente grabados. Después de entrenar los modelos, se medirá la eficacia y eficiencia con la cual cada modelo puede realizar la tarea de clasificar

acciones y así se pueda escoger el que tenga un mejor desempeño de acuerdo con medidas estadísticas previamente definidas.

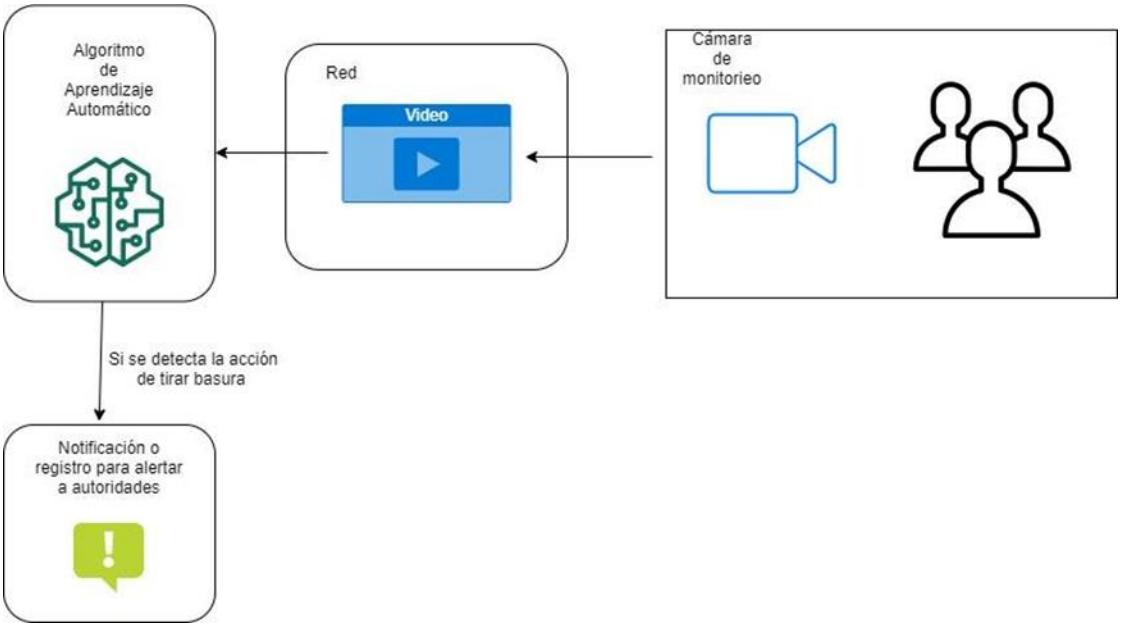
Una vez se haya escogido uno de los modelos se procederá a realizar el prototipo de la herramienta para clasificar vídeo proveniente del *streaming* de una sola cámara a través de una red local utilizando el modelo escogido. Una vez realizado el prototipo se evaluará la viabilidad de la solución y se concluirá si esta puede servir como base para seguirla mejorando y si es útil para poder monitorear varios basureros clandestinos simultáneamente.

Se planea utilizar neuronales convolucionales y redes neuronales secuenciales combinándolas de diferentes maneras. También se planea utilizar una red local para realizar el flujo de vídeo de la cámara hacia el servidor que esté corriendo el modelo de clasificación.

Se procurará utilizar videos reales de personas tirando basura, sin embargo, si existe alguna limitación en cuanto a la cantidad de datos, se recurrirá a generar videos con actuaciones de personas tirando basura a fin de contar con la cantidad suficiente de datos para que los algoritmos puedan aprender a distinguir entre la acción de tirar basura y cualquier otro tipo de acción.

Se presenta un diagrama de la propuesta de solución en la figura 1.

Figura 1. **Solución propuesta del sistema de monitoreo de vídeo**



Fuente: elaboración propia, realizado con draw.io.

7. MARCO TEÓRICO

7.1. Modelos de aprendizaje automático

De acuerdo con Mohri, Rostamizade y Talwalker (2018), el aprendizaje automático puede ser a grandes rasgos definido como los métodos computacionales utilizando la experiencia para mejorar el rendimiento o hacer predicciones precisas. Estas predicciones pueden ir desde clasificar un correo electrónico como spam hasta aprender a distinguir entre un perro y un gato en una foto. En este libro, los autores hablan de algunas tareas específicas en las que se utiliza el aprendizaje automático: Clasificación, regresión, ordenamiento, agrupamiento y reducción de dimensiones.

A continuación, se detallan métodos de aprendizaje automático que han resultado útiles para la tarea de clasificar acciones. Ya que esta tarea suele involucrar tanto dimensiones espaciales como temporales se suele utilizar arquitecturas combinadas para poder extraer características del vídeo específicas.

7.2. Redes neuronales convolucionales

Las redes neuronales convolucionales son actualmente utilizadas en una amplia variedad de tareas relacionadas a la visión por computadora.

Se podría decir que los primeros en utilizar una red neuronal tal y como la conocemos ahora para el trabajo de discriminar imágenes fueron LeCun, Haffner, Bottou y Bengio (1989) En su trabajo Reconocimiento de dígitos escritos a mano

con una red de propagación inversa. Como puede notarse en el título de su publicación, el nombre de red neuronal se popularizó hasta más tarde.

LeCun, Haffner, Bottou y Bengio (1999), describieron lo que es una red neuronal convolucional. Este tipo de arquitectura tiene 3 tipos diferentes de capas. Las capas convolucionales son un conjunto de filtros que utilizan la operación de convolución. Estas convoluciones necesitan de un núcleo que al operarse con la entrada resultan en un nuevo tensor que contiene ciertas características específicas. Por ejemplo, se puede hacer un filtro que pueda extraer todas las líneas horizontales de cierta imagen. A continuación, se define de forma más precisa una convolución en 2 dimensiones, para un filtro f de tamaño FxF y una imagen g de tamaño NxN, para $0 \leq x, y \leq n - f$:

$$(f * g)[x, y] = \sum_{X=0}^{F} \sum_{Y=0}^{F} g[X + x, Y + y] f[X, Y] \qquad \text{(Ec. 1)}$$

A este tipo de convolución se le llama convolución válida.

Las capas de submuestreo o agrupamiento realizan una operación para disminuir las dimensiones de su entrada. En este sentido existen diferentes tipos de agrupamientos como lo son los de valor máximo o los de valor promedio. Al utilizar una capa de agrupamiento de valor máximo, dada una entrada, ésta se divide en subconjuntos más pequeños, de los que se conserva únicamente el valor numérico más alto.

Por último, se tienen capas de neuronas completamente conectadas. El fin de estas estructuras es poder realizar un razonamiento a más alto nivel. Estas capas tienen este nombre debido a que cada una de sus partes más fundamentales, llamadas neuronas, está conectada al resultado de cada unidad de la estructura o capa anterior. A continuación, se define de forma más precisa

lo que en realidad sucede en una de estas capas. El resultado de la i-ésima capa es:

$$A^{[i]} = g\left(W^{[i]}A^{[i-1]} + B^{[i]}\right) \quad \text{(Ec. 2)}$$

Donde W y B son los parámetros que se deben aprender de la capa y g es una función de activación. El trabajo de esta función de activación es introducir no linealidades a la operación a fin de que la red de neuronas pueda modelar funciones más complejas.

Resumiendo, *LeNet-5*, que fue la red neuronal convolucional desarrollada por LeCun *et. al.* (1999) tiene la siguiente arquitectura:

- Entrada: la imagen de 32 x 32 pixeles.
- Capa convolucional C1: 6 núcleos con tamaño de 5 x 5.
- Capa de agrupamiento S2: tamaño de 2 x 2.
- Capa convolucional C3: 16 núcleos de 5 x 5.
- Capa de agrupamiento S4: tamaño de 2 x 2.
- Capa convolucional C5: 1920 núcleo de 5 x 5.
- Capa completamente conectada de 120 neuronas.
- Capa completamente conectada de 84 neuronas.
- Capa de salida completamente conectada de 10 neuronas.

Con el paso de los años se fueron desarrollando arquitecturas más grandes como por ejemplo AlexNet, desarrollada por Krizhevsky, Sutskever y Hinton o VGG-16 desarrollada por Simonyan y Zisserman. Cada una de estas arquitecturas fue mejorando su desempeño al agregar más capas o agregando más complejidad a la red.

Algo interesante que pueda hacerse con estas redes neuronales convolucionales es que cada una de las capas va codificando ciertas características de las entradas de tal forma que puede quitarse la última capa de la red, que es la que se encarga de tomar la decisión final en cuanto a la aplicación de la red, como distinguir entre un transeúnte o una señal de tráfico, y utilizar esas características codificadas para otras aplicaciones como se revisará en breve.

7.3. Redes neuronales secuenciales

Las redes neuronales secuenciales son importantes en aquellas aplicaciones en las que es importante el orden de las entradas debido a su relevancia en el problema en cuestión o en las que el tiempo juega un papel importante. Como ejemplo tenemos aplicaciones en procesamiento del lenguaje natural, como la traducción de idiomas, clasificación de series temporales, generación de música, entre otros. Existen redes neuronales de este tipo bastante sencillas, sin embargo, en este trabajo se revisará solamente aquellas que han marcado el estado del arte en los últimos años.

7.3.1. Unidades recurrentes con compuertas o GRU

Este tipo de red neuronal secuencial fue propuesto por primera vez por Cho, Merriënboer y Bahdanau (2014), y tiene como propiedad importante el hecho de que necesita una cantidad menor de parámetros que otras arquitecturas de este tipo, donde se cuenta con compuertas a fin de regular la cantidad de información que se recuerda y que se olvida, pero que en algunas tareas tiene resultados cercanos a aquellos modelos más complejos.

En esta clase de red neuronal se tiene una unidad que conserva sus parámetros a lo largo de una secuencia, de modo que los resultados actuales se basan en resultados pasados. Obviamente existen muchas versiones de esta clase de modelo y por lo tanto se describirá a continuación uno de ellos, que por lo general es también el más utilizado.

Primero se dará una fórmula matemática de cada elemento de esta unidad y posteriormente se explicará el papel que estos poseen dentro de dicha unidad.

Para una secuencia de entradas $x = (, x^{<2>}, \dots , x^{<T>})$, donde cada elemento de la secuencia es un vector, para obtener los resultados de la entrada t.

$$\tilde{C}^{<t>} = tanh\ tanh\ (Wc\ [\Gamma r * C^{<t-1>}, x^{<t>}] + bc) \quad \text{(Ec. 3)}$$

$$\Gamma u = \sigma(Wu[C^{<t-1>}, x^{<t-1>}] + bu) \quad \text{(Ec. 4)}$$

$$\Gamma r = \sigma(Wr[C^{<t-1>}, x^{<t-1>}] + br) \quad \text{(Ec. 5)}$$

$$C^{<t>} = \Gamma u * \tilde{C}^{<t>} + (1 - \Gamma u) * C^{<t-1>} \quad \text{(Ec. 6)}$$

$$\hat{y}^{<t>} = softmax(C^{<t>}) \quad \text{(Ec. 7)}$$

Los dos resultados que buscamos para cada iteración en la secuencia son $\hat{y}^{<t>}$, la salida, y $C^{<t>}$ que es la celda de memoria. Lo que se pretende con la celda de memoria es poder conservar la información importante de las entradas anteriores dentro de la secuencia, de tal manera que la celda de memoria del paso anterior sirve para computar la entrada actual. A fin de poder obtener el

resultado que posee la celda de memoria actual se necesita de la compuerta Γu, que en inglés hace referencia a la letra u de *update*.

Esta compuerta pretende comunicar qué tanto se debe considerar el resultado de la célula de memoria anterior y qué tanto se debe tomar en cuenta del valor candidato denotado por $\tilde{C}^{<t>}$, valor que utiliza la entrada actual y la pone en contexto con el resultado de la célula de memoria anterior, dándole la relevancia apropiada utilizando la compuerta de relevancia Γr.

Las matrices Wc, Wu, Wr y los vectores bc, bu, br son los parámetros que la unidad necesita aprender durante el entrenamiento. El tamaño de dichos vectores y matrices depende de la entrada $x^{<t>}$ y el tamaño que se escoja para el vector que representa la celda de memoria. Dicho tamaño de la celda de memoria es un hiper parámetro.

Utilizando estas unidades se pueden armar diferentes tipos de arquitectura como lo son: muchos vectores a muchos vectores, muchos vectores a un vector, un vector a muchos vectores, y muchos vectores a muchos vectores con un tamaño de secuencia diferente para la entrada y la salida. Para estos casos por lo general se tiene un codificador y decodificador.

7.3.2. Memoria de largo corto término o LSTM

Este modelo fue presentado por primera vez por Hochreiter y Schmidhuber (1997). Este tipo de modelo es más complejo que las GRU. Tiene más parámetros, lo cual hace que requiera una mayor cantidad de datos a fin de que tenga resultados satisfactorios. La ventaja es que, por lo general, también es capaz de resolver problemas más complejos. Es interesante notar que es un modelo mucho más antiguo que las GRU.

A continuación, se detallan cada una de las partes de dicho modelo: para una secuencia de entradas $x = (x^{<1>},\ x^{<2>}, \dots\ ,\ x^{<T>})$, donde cada uno de los elementos de la secuencia es un vector, para obtener los resultados de la entrada t.

- Valor candidato

$$\tilde{C}^{<t>} = tanh\ tanh\ (Wc\ [a^{<t-1>}, x^{<t>}] + bc) \quad \text{(Ec. 8)}$$

- Compuerta de actualización

$$\Gamma u = \sigma(Wu[a^{<t-1>}, x^{<t-1>}] + bu) \quad \text{(Ec. 9)}$$

- Compuerta de olvido

$$\Gamma f = \sigma(Wf[a^{<t-1>}, x^{<t-1>}] + bf) \quad \text{(Ec. 10)}$$

- Compuerta de salida

$$\Gamma o = \sigma(Wo[a^{<t-1>}, x^{<t-1>}] + bo) \quad \text{(Ec. 11)}$$

- Celda de memoria

$$C^{<t>} = \Gamma u * \tilde{C}^{<t>} + \Gamma f * C^{<t-1>} \quad \text{(Ec. 12)}$$

- Estado escondido

$$a^{<t>} = \Gamma o * tanh\ (C^{<t>}) \quad \text{(Ec. 13)}$$

- Salida

$$\hat{y}^{<t>} = softmax(Wy\ a^{<t>} + by) \quad \text{(Ec. 14)}$$

Como puede notarse en las fórmulas, este modelo tiene aparte de la compuerta de actualización una compuerta de olvido. La compuerta de actualización se multiplica elemento por elemento con el valor candidato de tal manera que influye en qué tanto se utiliza del valor candidato para la celda de memoria actual. Del mismo modo la compuerta de olvido determina qué tanto se debe considerar el valor de la célula de memoria del paso anterior. Además, puede notarse que se tiene aparte de la célula de memoria un estado escondido que se utiliza para computar los valores de los pasos que siguen a continuación.

7.3.3. Redes transformadoras

Este tipo de red neuronal ha tenido gran impacto en los últimos años y en gran medida ha reemplazado a las LSTM y GRU en muchas tareas relacionadas a entradas secuenciales. De acuerdo con Vaswani *et. al.* (2017), las partes fundamentales de este tipo de red son las siguientes.

Self-attention, es un tipo de operación que genera un vector a partir de 3 parámetros que son: La petición, la llave y el valor. Dichos valores se definen a continuación. Para una secuencia de entradas $x = (x^{<1>}, x^{<2>}, \dots, x^{<T>})$, donde cada elemento en la secuencia constituye un vector.

- La pregunta

$$q^{<t>} = W^{Q} x^{<t>} \quad \text{(Ec. 15)}$$

- La llave

$$k^{<t>} = W^K x^{<t>} \quad \text{(Ec. 16)}$$

- El valor

$$v^{<t>} = W^V x^{<t>} \quad \text{(Ec. 17)}$$

Donde las matrices W^Q, W^K, W^V son parámetros que el modelo debe aprender. Y se combinan de la siguiente forma:

$$A(q,k,v) = \sum_i \frac{\exp(qk^{<i>})}{\sum_j \exp(qk^{<j>})} v^{<i>} \quad \text{(Ec. 17)}$$

De manera que la atención puede resumirse como:

$$Atención(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad \text{(Ec. 18)}$$

Donde Q, K y V son matrices y d_k se utiliza a modo de normalización.

Una vez que se ha definido la atención hacia sí mismo, ésta puede utilizarse para obtener la *Multihead attention*. Lo que pretende esta atención multi cabeza, es que la red pueda abstraer cada elemento en la sucesión en contextos diferentes. Por ejemplo, para una aplicación de procesamiento del lenguaje natural y traducción, una de las cabezas podría preguntar ¿Qué está pasando?, lo cual apuntaría la mayor atención hacia una palabra que denote acción dentro de la oración que se quiere traducir. Después, la segunda cabeza podría representar la pregunta ¿Cuándo?, lo cual haría que la atención apuntará hacia

alguna de las palabras de la oración que nos diga cuándo es que se realiza la acción.

Podemos definirla de manera más precisa de la siguiente manera:

$$Multi\ cabeza(Q, K, V) = concatenación(head_1, head_2, \dots, head_h)\ Wo \quad \text{(Ec. 19)}$$

$$head_i = Atención(Wi^Q Q, Wi^K K, Wi^V V) \quad \text{(Ec. 20)}$$

Además de estos componentes, algunos detalles importantes en este modelo son: el codificador de posición para la entrada, que es un vector que se suma a cada uno de los elementos de la sucesión de entrada a fin de que haya una noción de posición relativa entre cada uno. También, se pone una capa residual y de normalización del vector que tiene por entrada. Después de esta capa residual y de normalización, el vector se introduce a una red conectada completamente, para después pasar otra vez por una capa residual y de normalización.

La ventaja que tiene este modelo con respecto a los secuenciales es que se puede procesar toda la secuencia de manera paralela

7.4. Arquitecturas para el reconocimiento de acciones

Debido a que un vídeo tiene tanto una componente espacial como una componente temporal, a fin de desarrollar aplicaciones que trabajen con vídeo, se debe tomar en cuenta ambas. A continuación, se repasan algunos de las arquitecturas, que suelen ser la combinación de varios modelos, que han dado buenos resultados en la clasificación de vídeos.

7.4.1. Redes neuronales convolucionales más secuenciales

Baccouche, Mamalet, Wolf, Garcia y Baskurt (2010) fueron los primeros en utilizar una red neuronal secuencial, en este caso una LSTM, para la clasificación de vídeos. Su idea fue obtener características espaciales de los fotogramas del vídeo e introducir los dentro de una red neuronal secuencial. La forma en la que obtuvieron dichas características de los fotogramas fue utilizando una bolsa de palabras y un tipo estimación de movimiento dominante utilizando una transformación de características de escala invariante.

Utilizando una idea parecida Ng *et. al.* (2015) propusieron, en lugar de obtener características espaciales a mano de cada fotograma, obtener las características de salida de una CNN, que ya se ha demostrado a lo largo del tiempo son la primera opción para muchas otras aplicaciones que se relacionan con la visión computacional. Esto trae una ventaja importante porque se pueden ir haciendo pequeños ajustes incluso a la red neuronal convolucional, al mismo tiempo que se va entrenando a la red secuencial.

Otra parte importante de dicha arquitectura es la incorporación de un flujo óptico el cual codifica el movimiento en los fotogramas a través del tiempo. Este flujo óptico es un campo de vectores, que se puede representar como dos imágenes, una con la componente horizontal de los vectores y otra matriz con la componente vertical de ellos.

Dichas imágenes también se introducen como entradas a la CNN para que se extraigan sus características y después esa salida se pasa hacia una red secuencial. Después de obtener la predicción de clase de los fotogramas crudos y del flujo óptico, se hace una fusión de los puntajes de cada uno para obtener una predicción final.

Aunque el trabajo original del que se habla utiliza LSTM, las características extraídas por una red neuronal convolucional pueden introducirse en otras arquitecturas que puedan tomar en cuenta la componente temporal como lo son GRU y Transformadores.

8. PROPUESTA DE ÍNDICE DE CONTENIDOS

9. METODOLOGÍA

9.1. Características del estudio

El enfoque del estudio propuesto es mixto, ya que se pretende obtener datos estadísticos de distintos modelos de aprendizaje automático aplicados a la tarea de reconocer la acción de una persona desechando basura en un basurero clandestino y de esa manera concluir cuál es el que mejor se adapta a la tarea. También se pretende proponer el diseño de una herramienta que pueda utilizar dicho modelo para hacer inferencias en una cámara de video.

El alcance es descriptivo, pues se tomarán las métricas generadas por los modelos de aprendizaje automático a fin de determinar el más apropiado y de esa manera describir el que mejor se acopla para la tarea. También se describe la manera de utilizar dicho modelo para una aplicación.

El diseño adoptado será no experimental transeccional, ya que se tomarán datos una vez en el tiempo y serán los que se utilizarán para poder entrenar al modelo de aprendizaje automático.

9.2. Unidades de análisis

La población en estudio será un conjunto de datos de video, la cual se encuentra dividida en subpoblaciones dadas por videos con las etiquetas de uso de basurero clandestino y no uso de basurero clandestino, de la cual se extraerán muestras de forma aleatoria simple, que serán estudiadas en su totalidad.

9.3. Variables

Las variables en estudio se describen en la tabla I.

Tabla I. **Variables a utilizar dentro del estudio**

	Definición teórica	Definición operativa
Puntaje F1	Es el promedio armónico entre la cobertura y la precisión. $Puntaje\ F1 = 2 * \frac{Precisión * Cobertura}{Precisión + Cobertura}$	Da un resultado entre 0 y Entre más cerca del 1 se encuentre su valor, es un modelo con menos errores de predicción.
Tiempo de inferencia promedio	Tiempo promedio que cada modelo necesita para hacer una predicción de la etiqueta correspondiente a un video de tamaños específico.	Esta variable se medirá en milisegundos.
Modelo de aprendizaje automático	El modelo de aprendizaje automático a entrenar, con sus respectivos hiper parámetros.	Etiquetas categóricas del tipo de modelo de aprendizaje automático.

Fuente: elaboración propia.

9.4. Fases del estudio

A continuación, se detallan las fases necesarias para llevar a cabo el estudio. Estas dan claridad en cuánto a los pasos necesarios para concluir con éxito el trabajo de investigación.

9.4.1. Revisión de literatura

En esta fase se debe obtener el conocimiento teórico para poder realizar los distintos experimentos y de esa manera obtener una solución para el problema de investigación. En este caso se recurrirá principalmente a papeles científicos recientes ya que las técnicas necesarias para esta aplicación son de formulación muy reciente. La principal información que se buscará en investigaciones recientes es la relacionada a modelos de aprendizaje automático y técnicas para el reconocimiento de acciones en vídeo.

9.4.2. Gestión o recolección de la información

Para esta fase se pretende encontrar la mayor cantidad de vídeos en los cuales se pueda observar a personas desechando basura en basureros clandestinos. El énfasis se pondrá en vídeos con la etiqueta positiva ya que es mucho más fácil encontrar vídeos en los que la acción no se esté realizando. Es muy probable que se tenga que grabar vídeos de manera manual para poder completar el conjunto de datos con un tamaño adecuado.

9.4.3. Análisis de información

En esta fase se hará un análisis del conjunto de datos obtenidos para obtener características importantes como la duración promedio de cada vídeo en el cual una persona estés echando basura, número de vídeos en donde esté utilizando un basurero clandestino, número de vídeos en el que no se esté utilizando un basurero clandestino, y otros datos importantes que podrían ser de relevancia para que el modelo pueda aprender a clasificarlos de manera correcta.

9.4.4. Implementación de modelos de aprendizaje automático

En esta fase se implementarán modelos de aprendizaje automático previamente escogidos debido a su rendimiento superior de acuerdo con la literatura encontrada en investigaciones recientes. Se implementará también una línea para poder cambiar de forma fácil y conveniente los parámetros de dichos modelos y de manera eficiente se pueda integrar sobre ellos el entrenamiento sobre el conjunto de datos recolectados. Esto con el fin de realizar los debidos experimentos y así obtener el modelo que mejor se adapte a la tarea.

9.4.5. Experimentación sobre los modelos de aprendizaje automático y sus hiper parámetros

En este punto se utilizarán los modelos de aprendizaje automático implementados en la fase anterior para realizar experimentos sobre el conjunto de datos recolectados. Los resultados de dichos experimentos se irán documentando a fin de poder llegar a una conclusión.

9.4.6. Diseño de herramienta para la clasificación de vídeo utilizando una cámara

En esta parte de la investigación se realizará el diseño de la herramienta para la clasificación de vídeo utilizando una cámara y el modelo de aprendizaje automático obtenido a través de las iteraciones realizadas en la fase anterior.

9.4.7. Implementación de herramienta para la clasificación de vídeo utilizando una cámara

En esta parte se implementará la herramienta utilizando el diseño realizado en la fase anterior.

9.4.8. Redacción del informe final

Para finalizar, se realizará el informe final detallando los resultados de cada una de las fases anteriores.

10. TÉCNICAS DE ANÁLISIS DE LA INFORMACIÓN

Se utilizará estadística descriptiva para el análisis de información, a continuación, algunas métricas a implementar:

- Media aritmética: se utilizará para obtener algunas características básicas del conjunto de datos de vídeo a analizar. Por ejemplo, si pretende obtener la duración de tiempo promedio de todos los vídeos en el conjunto de datos. También, será importante a la hora de determinar otros parámetros como el tiempo que tarda un modelo en realizar una predicción.

- Exactitud: esta es una métrica importante en el aprendizaje automático. Sencillamente es la cantidad de predicciones correctas dividido entre el número total de predicciones multiplicado por 100. Será un parámetro que se utilizará para entender el desempeño de cada uno de los modelos de aprendizaje automático en la tarea requerida.

- Puntaje F1: esta métrica es el promedio armónico entre la precisión y la cobertura. Será la principal métrica a tomar en cuenta cuando se decida el modelo de aprendizaje automático más apropiado para la tarea.

$$Puntaje\ F1 = 2 * \frac{Precisión * Cobertura}{Precisión + Cobertura} \quad \text{(Ec. 21)}$$

$$Precisión = \frac{Verdaderos\ positivos}{Verdaderos\ positivos + Falsos\ positivos} \quad \text{(Ec. 22)}$$

$$Cobertura = \frac{Verdaderos\ positivos}{Verdaderos\ positivos + Falsos\ negativos} \quad \text{(Ec. 23)}$$

11. CRONOGRAMA

A continuación, en la tabla II, se detallan las actividades a llevarse a cabo durante la realización del trabajo de investigación.

Tabla II. **Cronograma de actividades**

	Enero					**Febrero**				**Marzo**					**Abril**				**Mayo**				**Junio**			
Actividad	1	2	3	4	5	1	2	3	4	1	2	3	4	5	1	2	3	4	1	2	3	4	1	2	3	4
Gestión o recolección de la información																										
Forma de autorización																										
Grabar en vídeo a personas desechando basura																										
Gestión con municipalidad para obtener videos																										
Análisis de la información																										
Etiquetar todos los vídeos																										
Organizar el conjunto de datos en carpetas																										
Codificar programas para análisis simple de datos																										
Implementación de modelos de aprendizaje automático																										
Implementar modelo uno																										

Continuación tabla II.

Implementar modelo dos																								
Implementar modelo tres																								
Implementar modelo cuatro																								
Experimentación con los modelos																								
Diseño de herramienta para la clasificación de vídeo																								
Redacción del informe final																								

Fuente: elaboración propia.

12. FACTIBILIDAD DEL ESTUDIO

12.1. Factibilidad económica

En la tabla III, se detalla la factibilidad económica del proyecto. Puede notarse que el costo es bajo en relación con el impacto que podría tener el proyecto en el medio ambiente.

Tabla III. **Descripción de la factibilidad económica del proyecto**

Recurso	Propósito	Precio	Cantidad	Subtotal
Alquiler de servidor en la nube	Entrenamiento de los modelos de aprendizaje automático	$50 mensuales	2	$ 100.00
Cámara de vigilancia	Grabación de vídeos y prototipo de la herramienta	$100	1	$ 100.00
Asesor del trabajo	Proporcionar recomendaciones y asesoría en la realización del proyecto.	$253	1	$ 253.00
Internet	Obtención de recursos y conexión a la computación en la nube necesaria para el entrenamiento del modelo	$30 mensuales	7	$ 210.00
Desarrollo de software	Desarrollo de los programas necesarios para el proyecto	$10/hora	300	$ 3000.00
Total				$ 3663.00

Fuente: elaboración propia.

12.2. Factibilidad operativa

- Acceso a la información

Se debe tener acceso a un conjunto de datos de vídeo que sea lo suficientemente grande como para poder entrenar de manera adecuada el modelo de aprendizaje automático.

- Personas dispuestas a ser grabadas en vídeo simulando la acción de tirar basura

Se deben obtener aproximadamente 1500 vídeos de personas desechando basura, por lo cual, a cada persona que colabore, se le pediría que grabe entre diez y quince videos diferentes, utilizando diferentes ángulos en la cámara. Se necesitarán por tanto entre 100 y 150 personas que quieran colaborar con el proyecto. Además, se necesitarán otros 1500 a 3000 vídeos de personas que estén realizando cualquier otra acción en la calle que no sea desechar basura en un basurero clandestino. Esto es mucho más sencillo porque no es una acción tan específica, y se pueden incluso utilizar vídeos tomados de internet.

- Permisos

Se gestionarán permisos para poder grabar vídeos en las calles de la ciudad de Guatemala, y de utilizar vídeos grabados previamente por las cámaras de la municipalidad. Si esto no fuera posible, todos los vídeos se tendrán que grabar en una propiedad privada.

12.3. Factibilidad técnica

- Software necesario

    - Lenguaje de programación *Python*.
    - Librería de aprendizaje automático *Tensorflow*.
    - Librería de visión por computadora *OpenCV*.
    - Librería de aprendizaje automático *Scikit-Learn*.

- Computadora con hardware necesario

    - Memoria RAM de 8GB o preferiblemente más.
    - GPU dedicada.
    - Procesador Intel Core i5, AMD Rizen 5 o superiores.
    - Disco duro con 250GB de estado sólido.

13. REFERENCIAS

1. Baccouche, M., Mamalet, F., Wolf, C., Garcia, C. y Baskurt, A. (septiembre, 2010). Action classification in soccer videos with long short-term memory recurrent neural networks. *In International Conference on Artificial Neural Networks*. Congreso llevado a cabo en Springer, Berlin.

2. Boyko, N., Basystiuk, O. y Shakhovska, N. (2018). Performance Evaluation and Comparison of Software for Face Recognition, Based on Dlib and Opencv Library. *IEEE Second International Conference on Data Stream Mining & Processing.* Congreso llevado a cabo en Lviv, Ucrania.

3. Carreira J. y Zisserman A. (2017), Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition.* Congreso llevado a cabo en Honolulu, Hawái.

4. Cho, K., Van Merriënboer, B., Bahdanau, D., y Bengio, Y. (2014). *On the properties of neural machine translation: Encoder-decoder approaches*. Doha, Qatar: arXiv preprint.

5. Flores, L. (2018). *Modelo de gestión para la erradicación de basureros clandestinos, estudio de dos casos en el Municipio de Villa Nueva* (Tesis de maestría). Universidad de San Carlos de Guatemala, Guatemala.

6. Han Y., Zhang P., Zhuo T., Huang W. y Zhang Y. (mayo, 2017). Going deeper with two-stream ConvNets for action recognition in video surveillance*. Pattern Recognition Letters*, *107*, 83-90.

7. Hochreiter, S., y Schmidhuber, J. (marzo, 1997). *Long short-term memory. Neural computation, 9*(8), 1735-1780.

8. Jing, C., Wei, P., Sun, H., Sun H. y Zheng N. (enero, 2020) Spatiotemporal neural networks for action recognition based on joint loss. *Neural Comput & Applic, 32*, 4293–4302.

9. Krizhevsky, A., Sutskever, I., y Hinton, G. E. (marzo, 2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems, 25*, 1097-1105.

10. LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., y Jackel, L. (marzo, 1989). Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems, 2*, 396-404.

11. LeCun, Y., Haffner, P., Bottou, L., y Bengio, Y. (octubre, 1999). Object recognition with gradient-based learning. *In Shape, contour and grouping in computer visión, 1681,* 319-345.

12. Mohri, M., Rostamizadeh, A., y Talwalkar, A. (2018). *Foundations of machine learning.* Londres, Inglaterra: MIT press.

13. Rawat, W. y Wang, Z. (septiembre, 2017). Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review*. Neural Comput*, *29*(9), 2352–2449.

14. Shou Z., Pan J., Chan J., Miyazawa K., Mansour H., Vetro A., Giro-i-Nieto X, y Chang S. (julio, 2018). Online Detection of Action Start in Untrimmed, Streaming Videos. *Proceedings of the European Conference on Computer Vision (ECCV)*, *1*, 534-551.

15. Simonyan, K., y Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition*. Doha, Qatar: arXiv preprint.

16. Ullah A., Ahmad J., Muhammad K., Sajjad M. y Baik W. (enero, 2018), Action Recognition in Video Sequences using Deep Bi-Directional LSTM with CNN Features. *IEEE Access*, *6*, 1155-1166.

17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., y Polosukhin, I. (diciembre, 2017). Attention is all you need. Advances in neural information processing systems. *31st Conference on Neural Information Processing Systems.* Congreso llevado a cabo en California, Estados Unidos.

18. Wu C., Zaheer M., Hu H., Manmatha R., Smola A. y Krähenbühl P. (junio, 2018). Compressed Video Action Recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Congreso llevado a cabo en Salt Lake City, Estados Unidos.

19. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., y Toderici, G. (abril, 2015). Beyond short snippets: Deep

networks for video classification. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. Congreso llevado a cabo en California, Estados Unidos.

20. Yun, K., Kwon Y., Oh S., Moon J. y Park J. (2019). *Vision based garbage dumping action detection for real world surveillance platform.* Corea del Sur: ETRI Journal.