ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

DEPARTAMENTO DE ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES

CARRERA DE INGENIERÍA EN ELECTRÓNICA, AUTOMATIZACIÓN Y CONTROL

DESARROLLO DE UN CLASIFICADOR DE VIDEO PARA LA DETECCIÓN AUTOMÁTICA DE EVENTOS DE ASALTO A PEATONES BASADO EN ALGORITMOS DE APRENDIZAJE PROFUNDO

AUTOR: CRISTHIAN TERÁN

AGOSTO 2022

OBJETIVO GENERAL

Desarrollar un clasificador de video para la detección automática de eventos de asalto de peatones basado en algoritmos de aprendizaje profundo

OBJETIVOS ESPECÍFICOS

Investigar los algoritmos de aprendizaje profundo para la clasificación videos

Generar la base de datos con videos de eventos de asalto a peatones para el entrenamiento, validación y pruebas de un modelo de aprendizaje profundo.

Seleccionar, configurar y entrenar un modelo de aprendizaje profundo para que realice la detección de eventos de asalto a peatones.

Evaluar el desempeño del modelo.

FUNDAMENTO TEÓRICO

Seguridad Ciudadana

Visión Artificial

Deep Learning

Tensorflow

OpenCV

Google Colaboratory y Paperspace Gradient

Métricas de Rendimiento

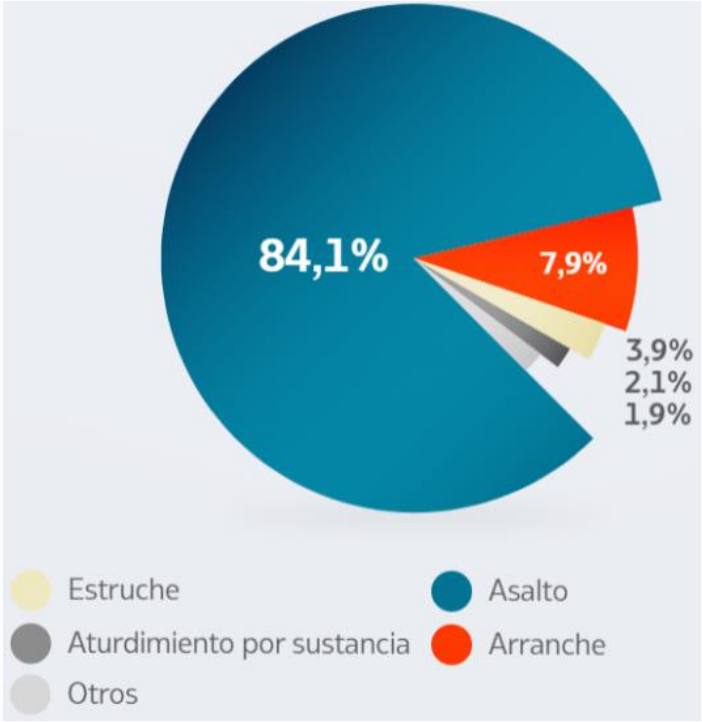
SEGURIDAD CIUDADANA

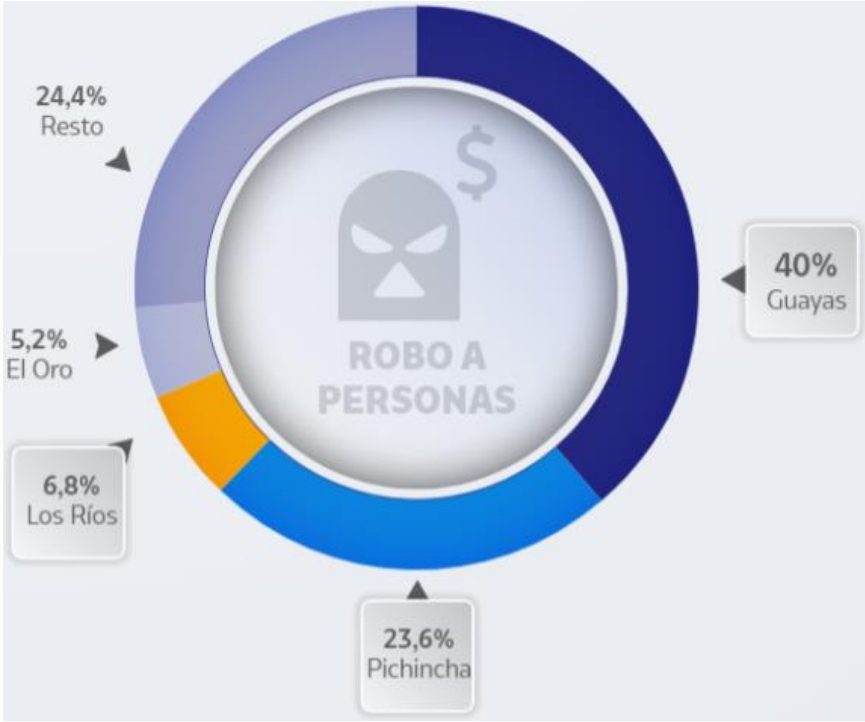
Definición

Tecnología

Robo a Personas

ROBO A PERSONAS, MODALIDADES Y SU PRESENCIA EN EL ECUADOR





VISIÓN ARTIFICIAL

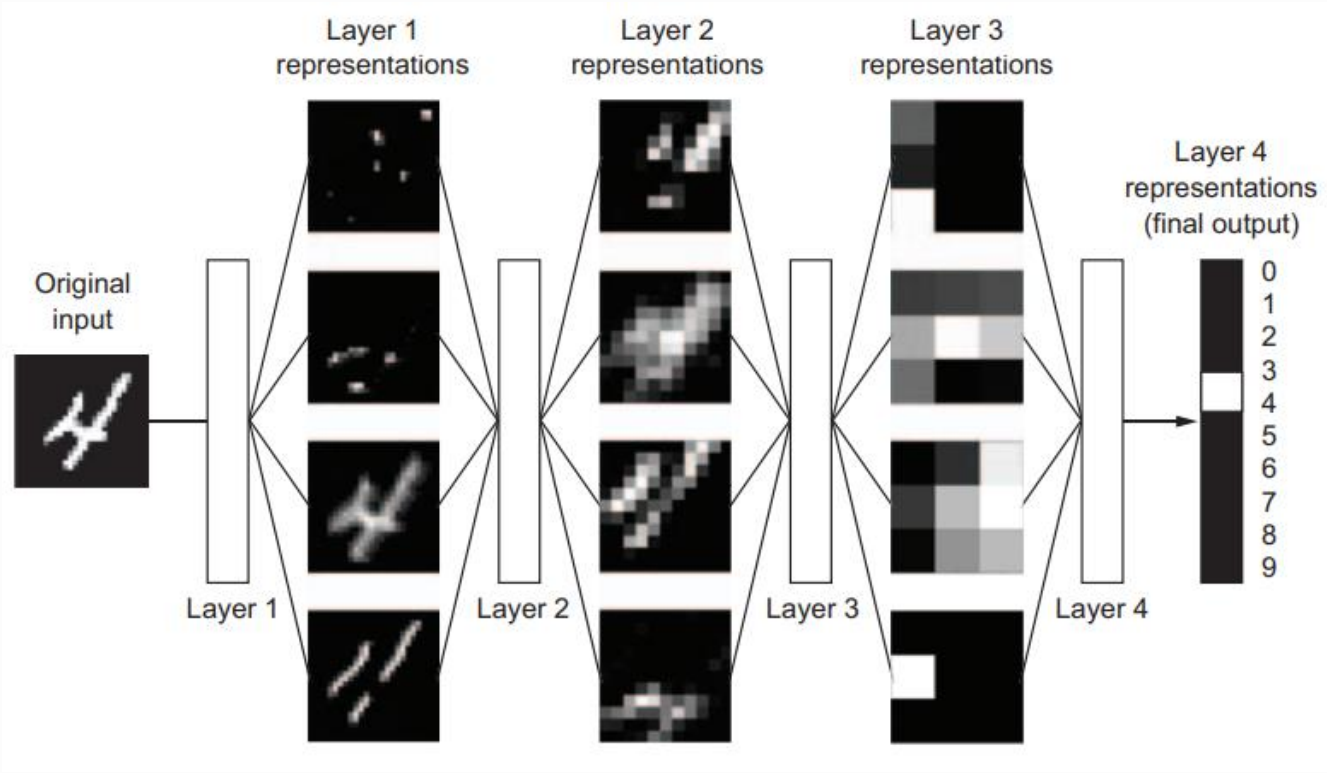
La visión artificial es un campo de estudio en el que el objetivo principal es proveer, a una máquina, la capacidad de entender una imagen o conjunto de imágenes mediante su adquisición, procesamiento y análisis.

DEEP LEARNING

El Deep Learning o Aprendizaje Profundo es una rama de la inteligencia artificial y del aprendizaje automático que se centra en crear grandes modelos de redes neuronales multicapas hechas para la toma de decisiones en los que se involucran datos extensos y complejos.

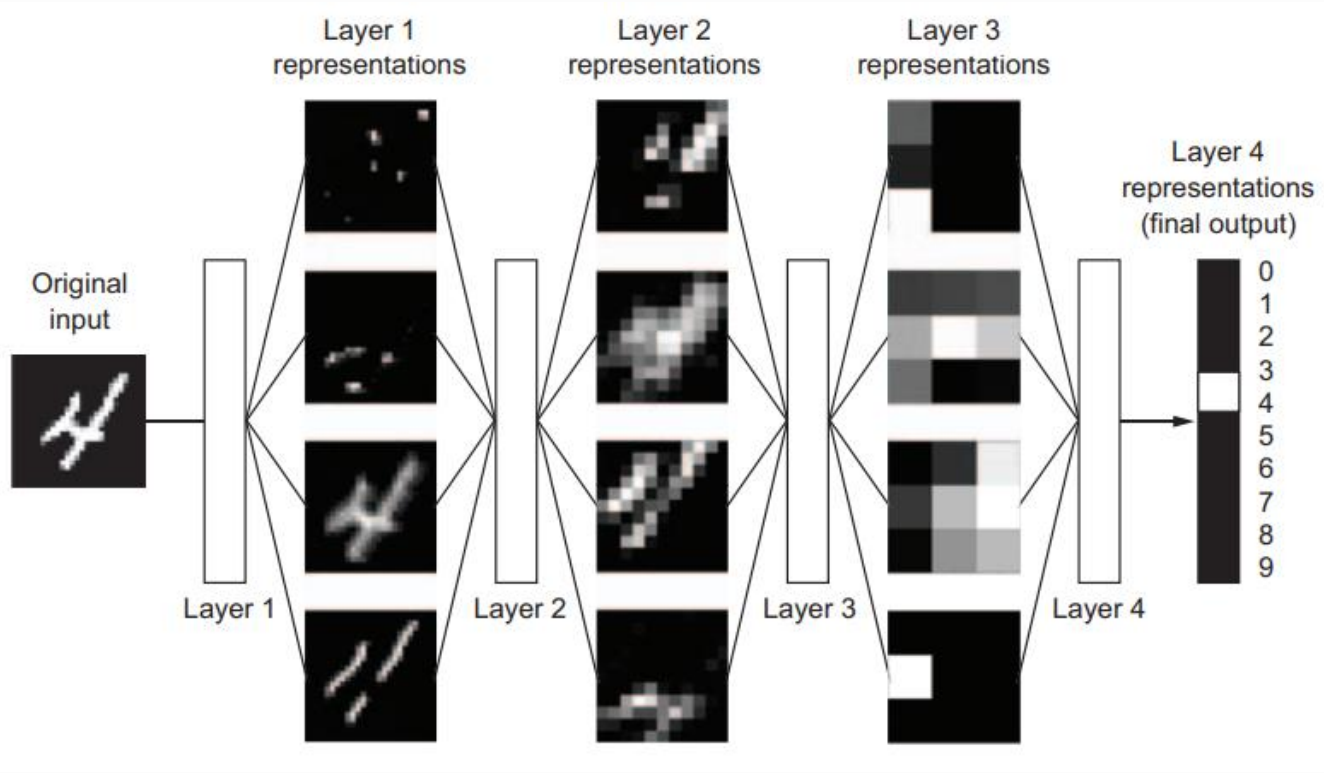
REDES NEURONALES CONVOLUCIONALES

Las redes neuronales convolucionales o CNN es un tipo de red que se deriva del perceptrón multicapa, su estructura y funcionamiento se relaciona mucho con la corteza visual de los seres vivos.

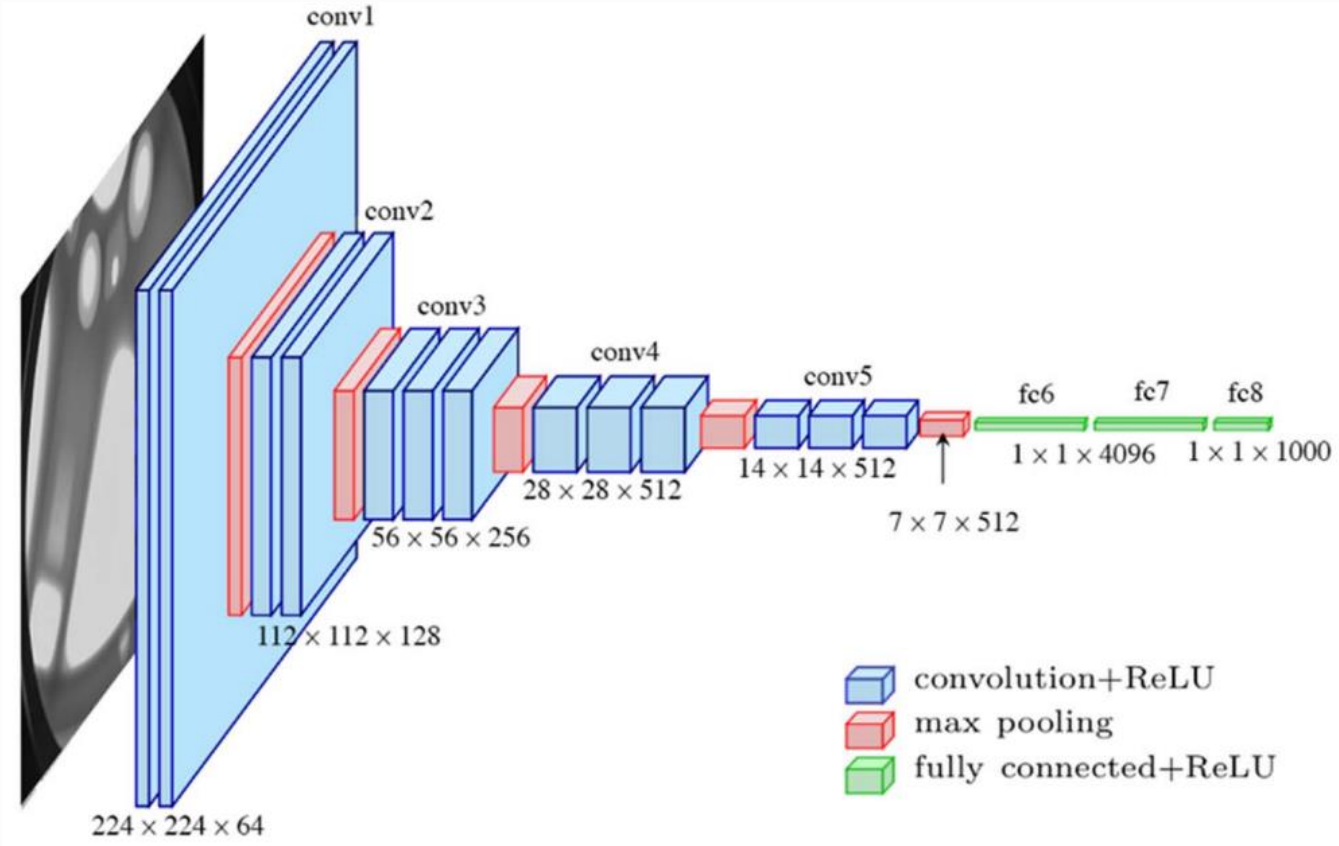


REDES NEURONALES CONVOLUCIONALES

Su aplicación en matrices bidimensionales la hace especialmente útil en el campo de la clasificación, segmentación y reconocimiento de imágenes.



VGG16



Es una CNN desarrollada por Karen Simonyan y Andrew Zisserman que ha logrado resultados de 96% a 97% de exactitud en la clasificación de imágenes de la base de datos ImageNet

REDES NEURONALES RECURRENTES

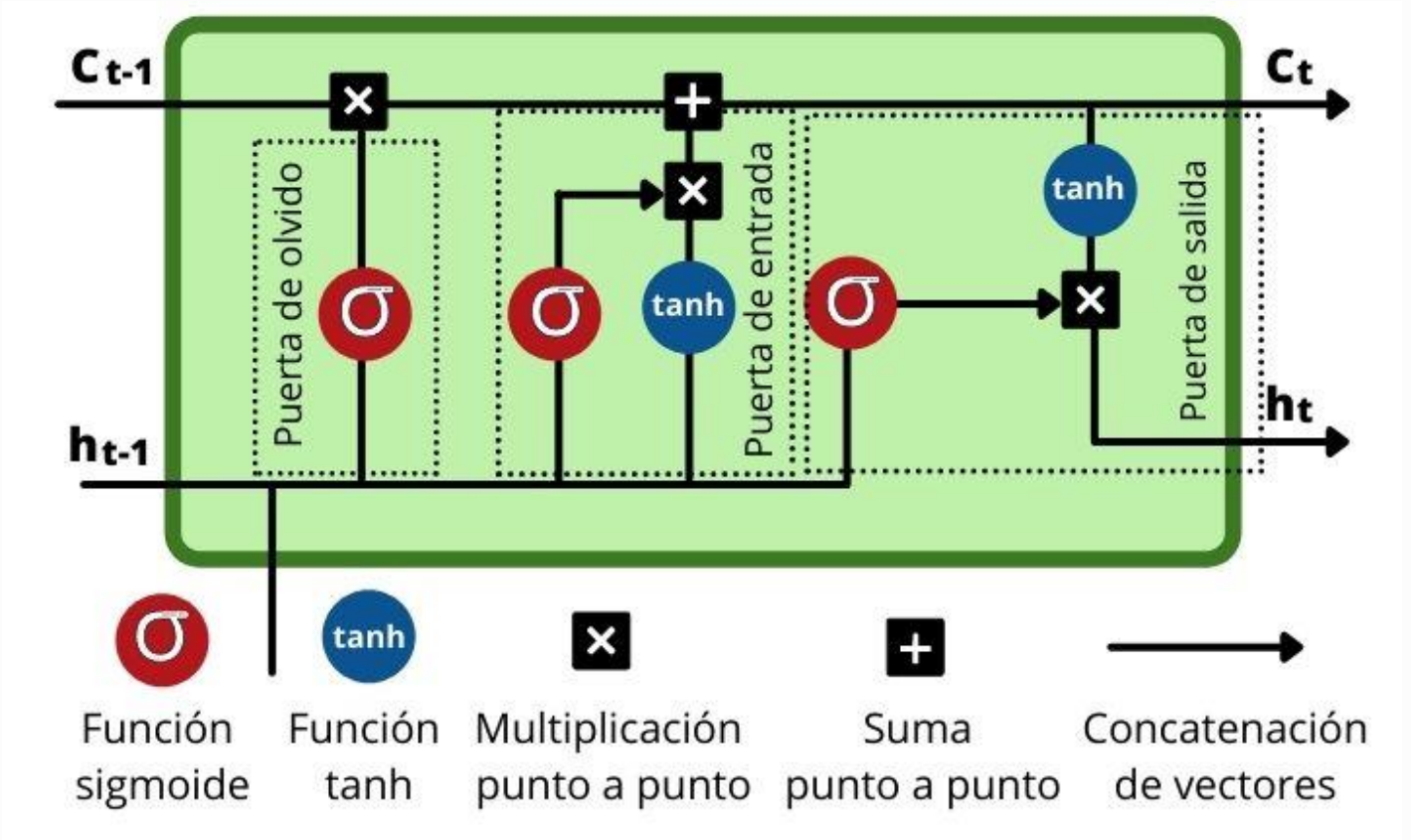
Las Redes Neuronales Recurrentes o RNN son capaces de procesar y obtener patrones o características de datos secuenciales, es por ello por lo que son muy utilizadas en aplicaciones de audio, video, procesamiento del lenguaje, entre otras aplicaciones.

REDES NEURONALES RECURRENTES

Un problema común de las RNN más sencillas es que no tienen la capacidad de aprender patrones que se extiendan mucho en el tiempo por lo que se creó un tipo de celda de memoria más compleja que ex capaz de extraer los patrones de secuencias de mayor longitud. Esta celda se denomina como Long-Short Term Memory o LSTM

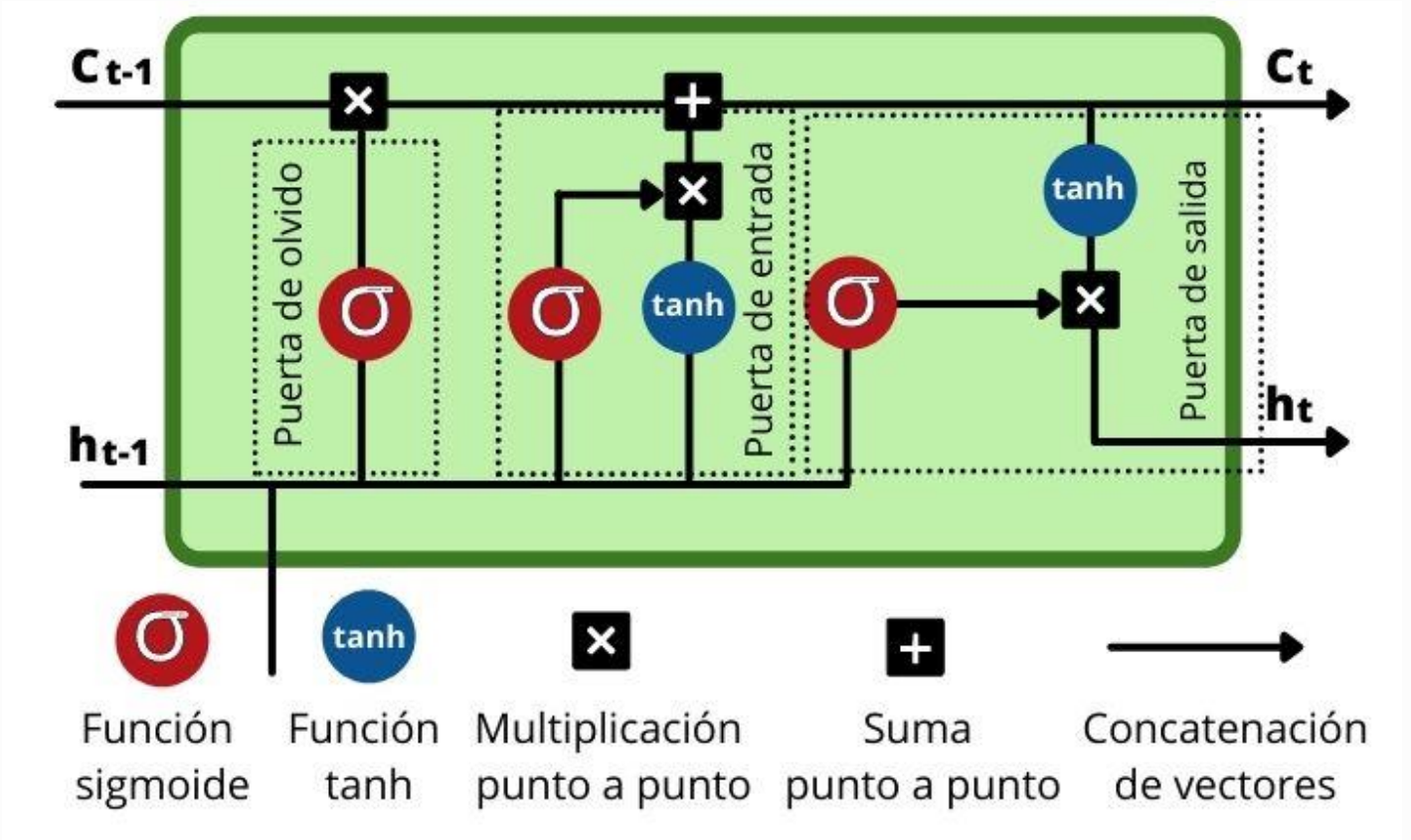
LSTM

Son redes recurrentes que tienen la capacidad de aprender patrones de eventos pasados a largo plazo recordando las características temporales relevantes y olvidando las características que no son representativas.



LSTM

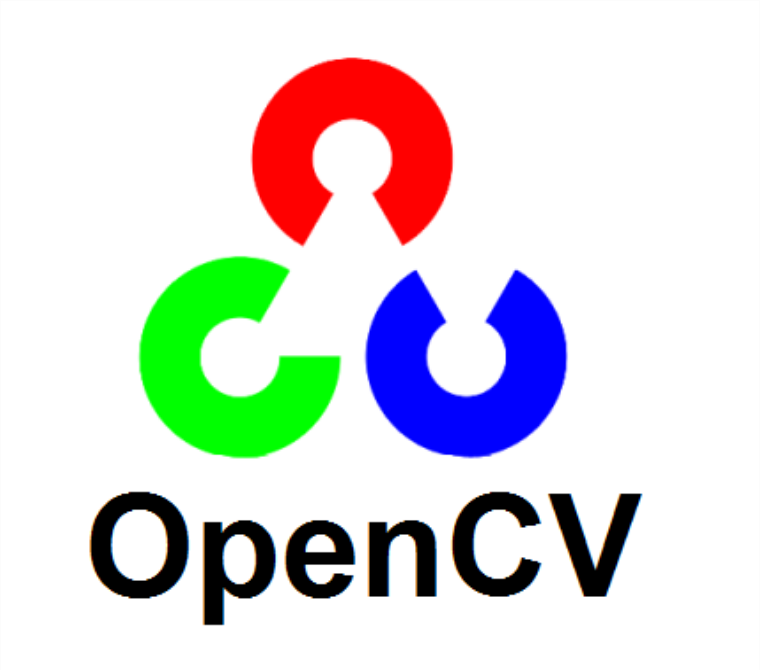
Esto es posible ya que las células LSTM están compuestas de bloques de memoria y tres compuertas de entrada, salida y olvido.



TENSORFLOW

Tensorflow es una librería de código abierto que está diseñada para desarrollar e implementar modelos de aprendizaje automático, es capaz de ejecutarse en sistemas desde teléfonos móviles hasta máquinas y dispositivos computacionales como tarjetas GPU.

OPENCV



Es una biblioteca de código abierto creada para su uso en visión artificial y aprendizaje automático. Utiliza una estructura modular que permite acceder a varias funcionalidades relacionadas con la captura y procesamiento de imagen.

GOOGLE COLABORATORY Y PAPERSPACE GRADIENT

Google Colaboratory o Google Colab es un producto de Google Research que permite programar en lenguaje Python y ejecutarlo desde el navegador, dando acceso gratuito a recursos informáticos adecuados para el trabajo con aprendizaje automático, ciencia de datos, educación entre otros.

GOOGLE COLABORATORY Y PAPERSPACE GRADIENT

Gradient es un entorno creado por Paperspace, ofrece recursos similares a Google Colaboratory con ciertas diferencias. En este entorno se puede elegir qué tiempo va a ser requerido, la inactividad no conlleva una expulsión del entorno, además se tiene una mayor variedad de GPUs a disposición

MÉTRICAS DE RENDIMIENTO

Las métricas de rendimiento son de utilidad para problemas de clasificación en donde se tiene el objetivo de discriminar distintos algoritmos de aprendizaje automático y aprendizaje profundo, para elegir el mejor dependiendo del objetivo de aplicación.

MATRIZ DE CONFUSIÓN

		PREDICCIÓN	
		Positivos	Negativos
VALOR REAL	Positivos	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

VP: cantidad de positivos que fueron clasificados correctamente.
VN: cantidad de negativos que fueron clasificados correctamente
FN: cantidad de positivos que fueron clasificados como negativos.
FP: cantidad de negativos que fueron clasificados como positivos.

MÉTRICAS DE RENDIMIENTO

Métrica	Fórmula	Descripción
Accuracy (Exactitud)	$\frac{VP + VN}{VP + VN + FP + FN}$	Proporción de clasificaciones predichas de manera correcta sobre el total de instancias.
Recall (Sensibilidad) o Tasa de Verdaderos Positivos	$\frac{VP}{VP + FN}$	Proporción de casos positivos bien clasificados.
Especificidad o Tasa de Verdaderos Negativos	$\frac{VN}{VN + FP}$	Proporción de casos negativos bien clasificados.
Probabilidad de Falsa Alarma	$1 - \frac{VN}{VN + FP}$	Proporción de casos positivos mal clasificados (error Tipo I).

METODOLOGÍA

Preparación de Datos

Diseño del Modelo

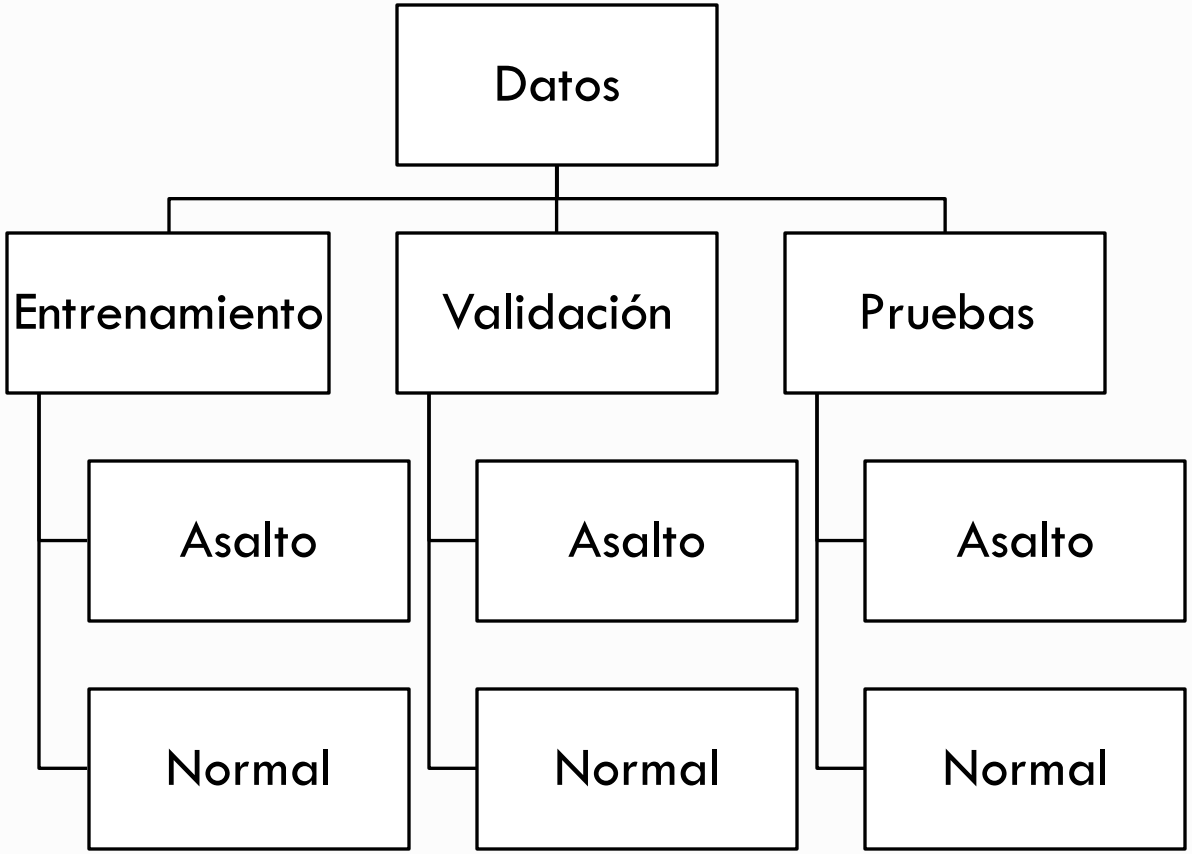
Entrenamiento Validación y Pruebas

Historial de Entrenamiento

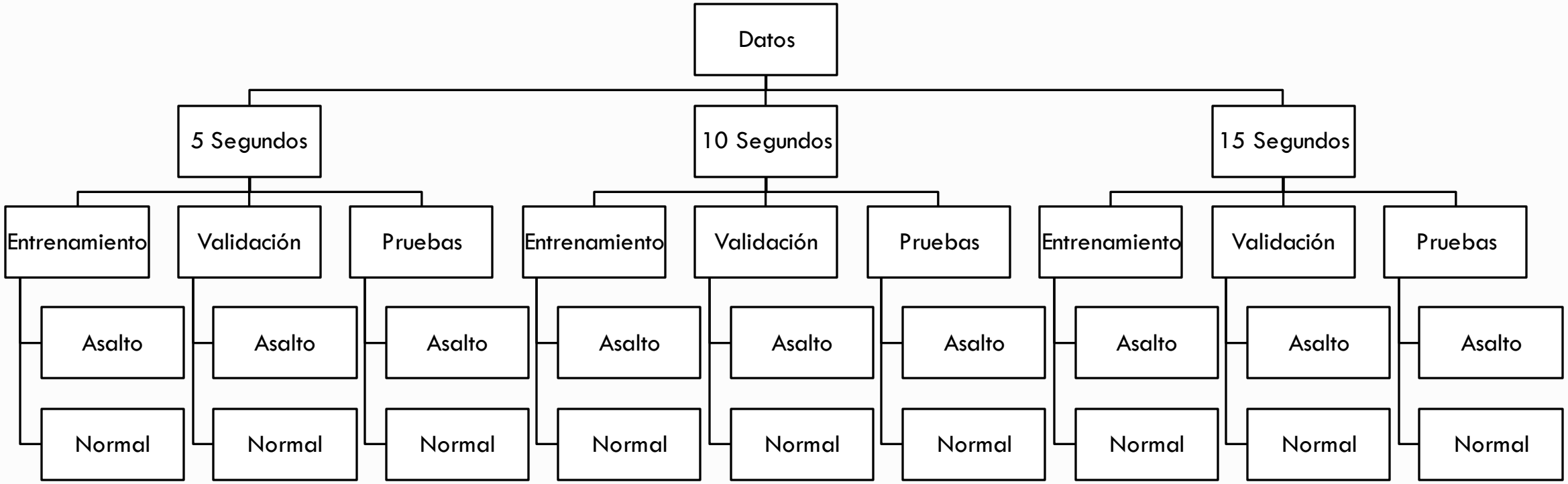
PREPARACIÓN DE DATOS

Los datos de entrenamiento son videos relacionados con asaltos a peatones capturados por cámaras de videovigilancia. Fueron tomados principalmente de la plataforma Youtube y de la base de datos UCF-Crime, preparados usando el editor de video llamado “Kdenlive” y organizados en carpetas.

ORGANIZACIÓN DE LOS DATOS



ORGANIZACIÓN DE LOS DATOS

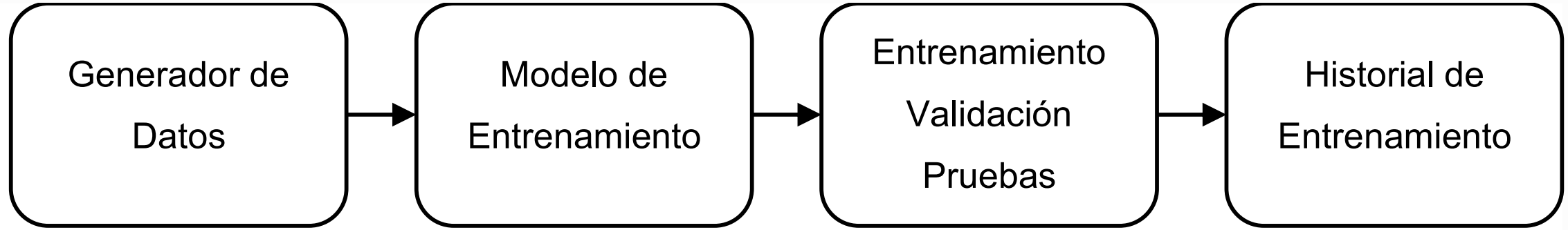


AUMENTO DE DATOS

Los datos almacenados serán aumentados con técnicas de inversión horizontal de imagen, giro de imagen y añadiendo ruido, todo esto con ayuda de la biblioteca VidAug.

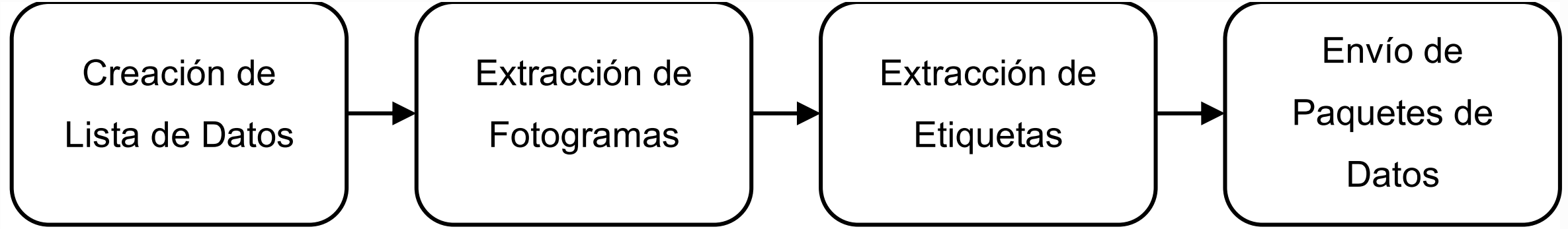
DISEÑO DEL MODELO

Para crear un clasificador de video, útil para la detección de asaltos a peatones, se ha creado un programa de cuatro etapas para las fases de entrenamiento, validación y pruebas de modelos de redes neuronales.

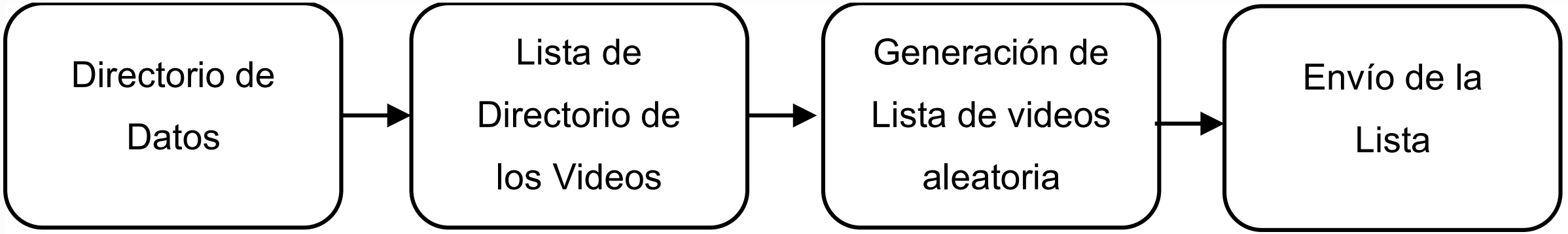


GENERADOR DE DATOS

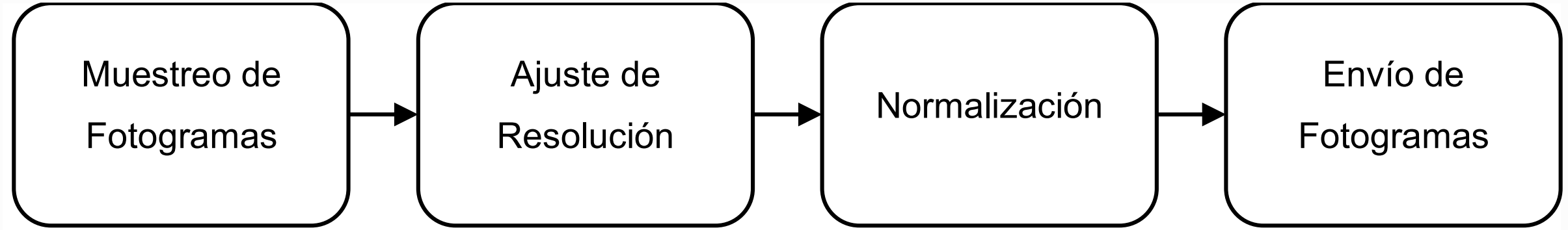
Los generadores dividen la base de datos en pequeños paquetes y son enviados conforme el programa lo necesita.



EXTRACCIÓN DE LISTA DE DATOS

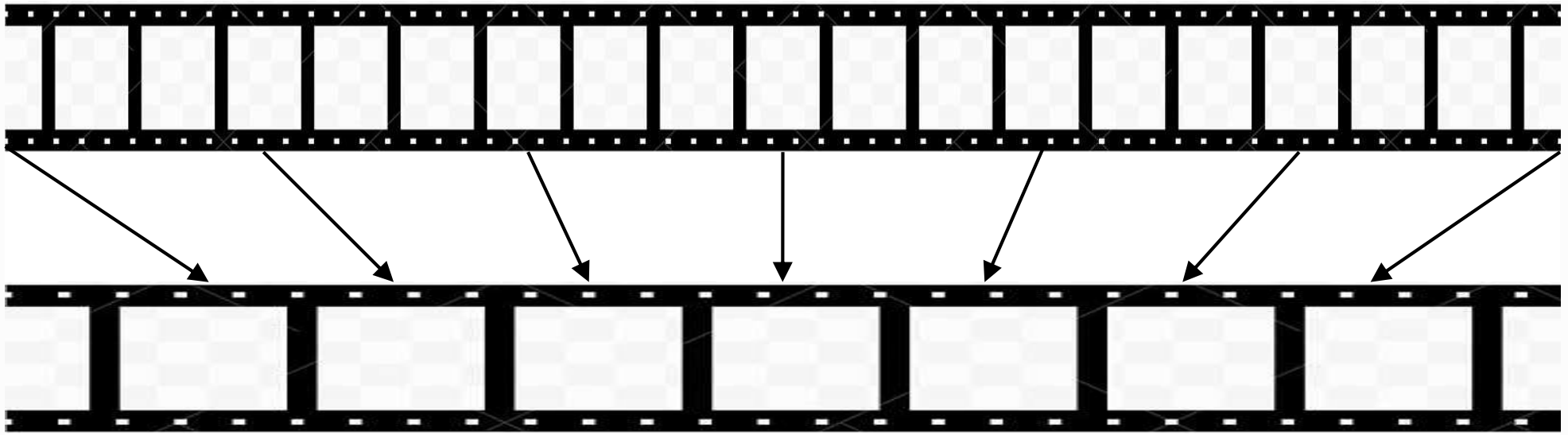


EXTRACCIÓN DE FOTOGRAMAS



EXTRACCIÓN DE FOTOGRAMAS

FOTOGRAMAS DE VIDEO



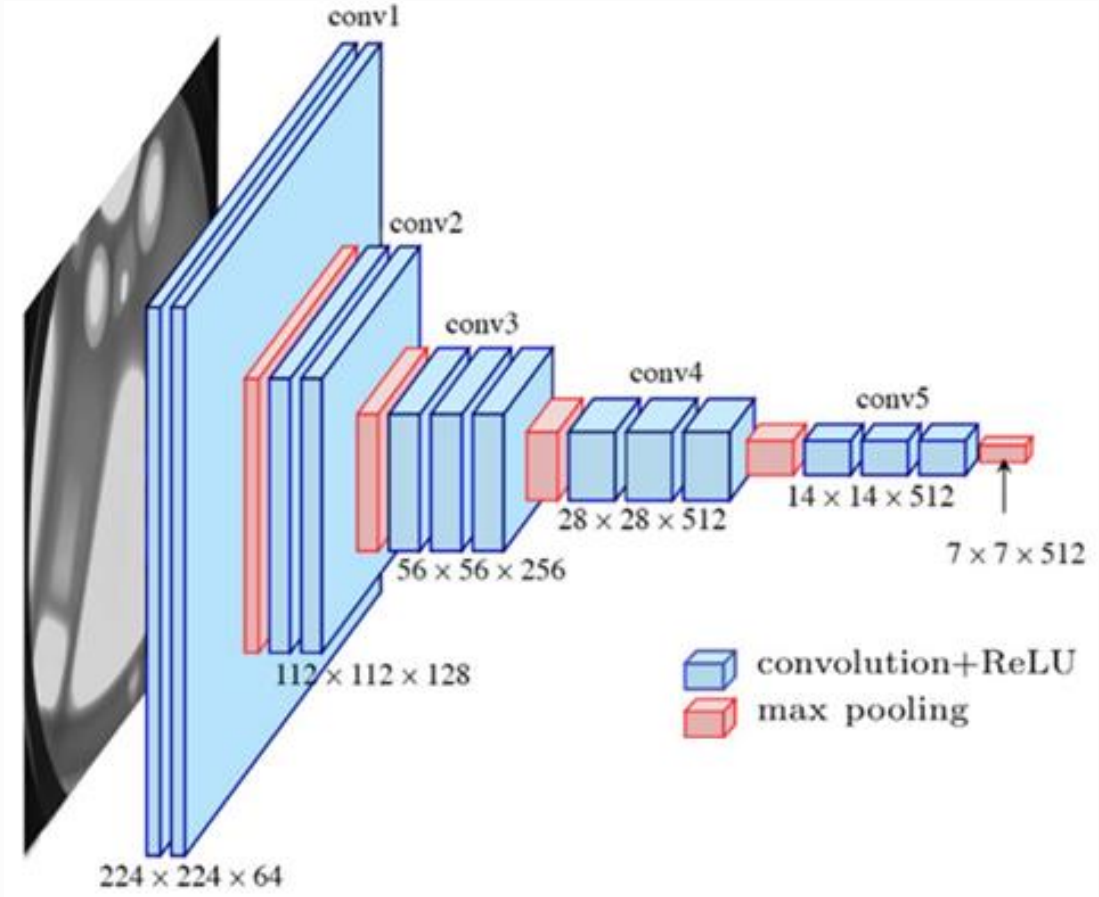
SECUENCIA DE FOTOGRAMAS RESULTANTE

MODELO DE ENTRENAMIENTO

La idea del modelo es extraer las características de una cantidad determinada de fotogramas de los videos, evaluar su relación a lo largo del tiempo y determinar si se muestra un evento de asalto.

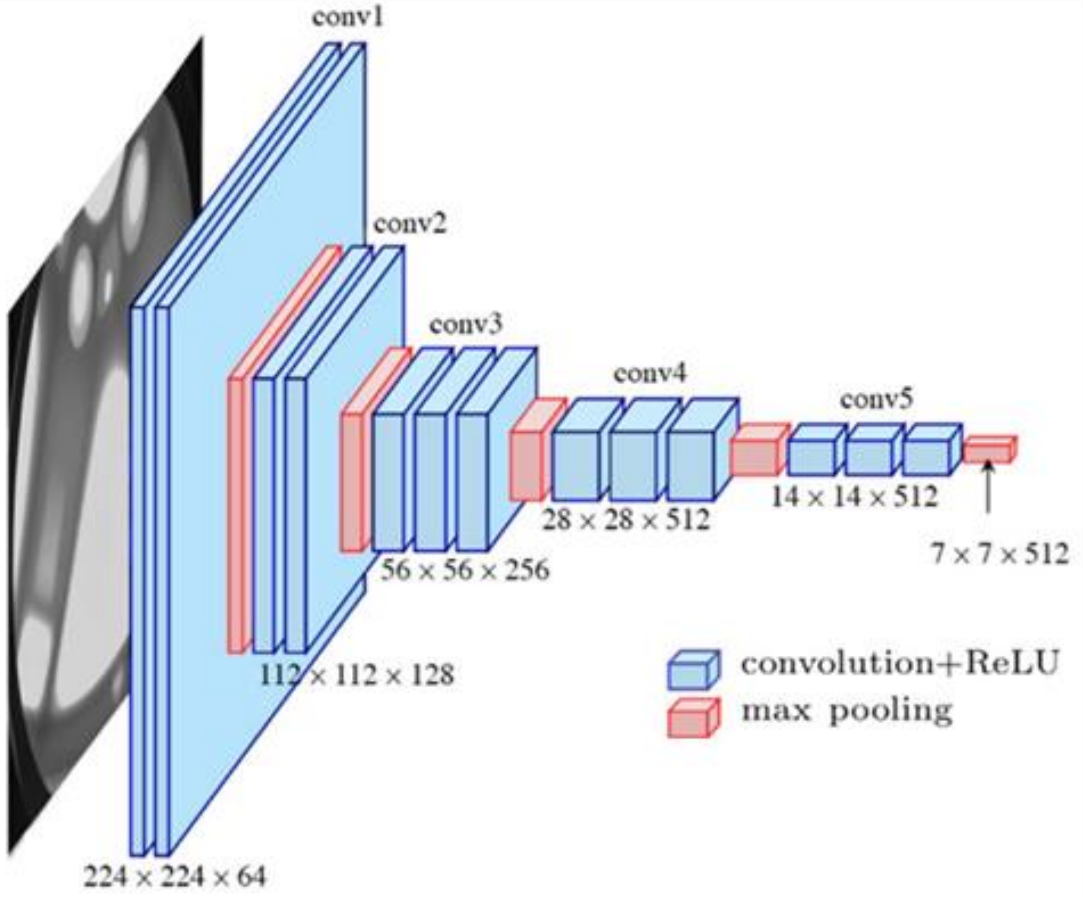
EXTRACCIÓN DE CARACTERÍSTICAS

Con el modelo VGG16, que ha sido entrenado para la clasificación de imágenes, y mediante la eliminación de la última capa se obtiene una red que es capaz de extraer las características de una imagen.

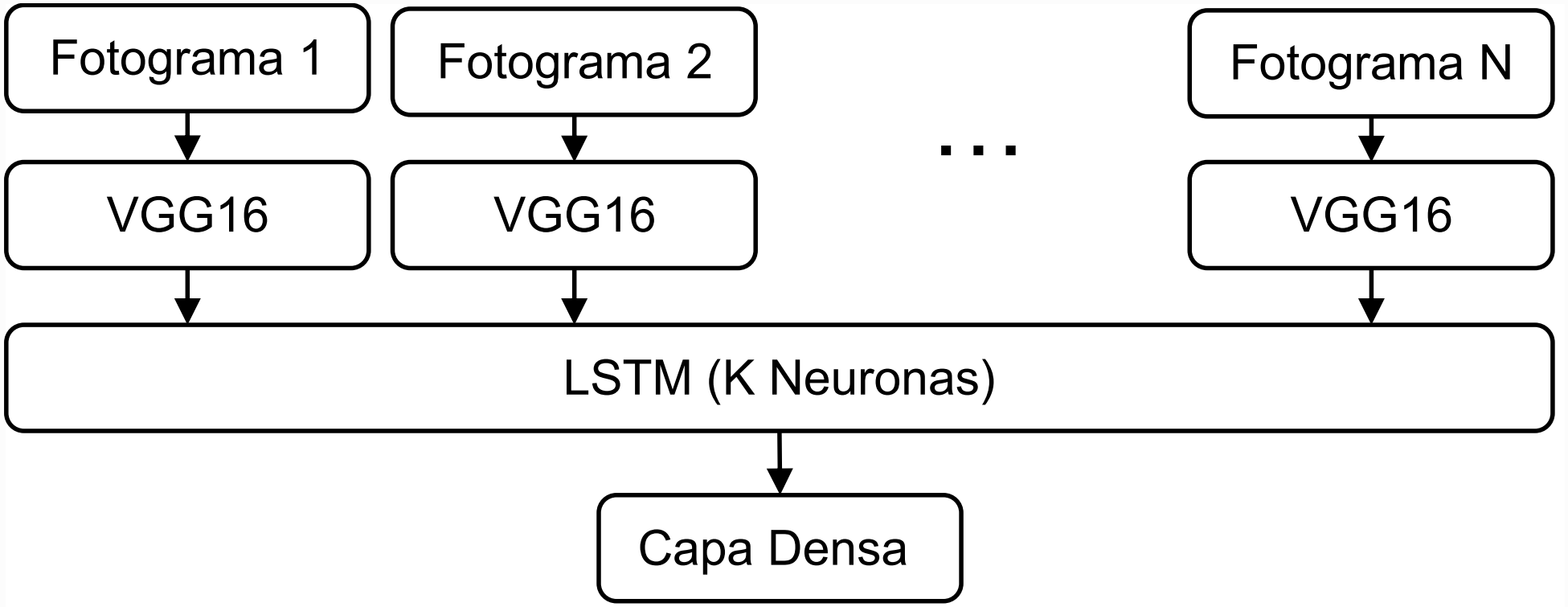


EXTRACCIÓN DE CARACTERÍSTICAS

Al tener una secuencia de imágenes lo que se ha hecho es distribuir el modelo VGG16 para que haga el reconocimiento de cada uno de los fotogramas de la secuencia, esto se logra mediante la clase "TimeDistribuited"



MODELO PROPUESTO



ENTRENAMIENTO, VALIDACIÓN Y PRUEBAS

Teniendo los datos listos y el modelo establecido se comienza el entrenamiento del modelo, para esto existen una serie de parámetros que permitirán optimizar el proceso dependiendo de los recursos disponibles. Se han tenido diversas consideraciones de acuerdo con la experiencia que se fue adquiriendo durante el proceso de entrenamiento y pruebas.

HARDWARE

Tipo de Hardware	Modelo
Procesador	Intel(R) Core(TM) i5-7300HQ de 2.50Ghz
Memoria RAM	8 GB
Memoria SSD	1 TB
Disco Duro	1 TB
Tarjeta Gráfica	NVIDIA GeForce GTX 1050
VRAM	4GB

HARDWARE

GPU	GPU RAM	CPUs	RAM
K80	12 GB	2 vCPU	13 GB
T4	16 GB	2 vCPU	13 GB hasta 25 GB
P100	16 GB	2 vCPU	13 GB hasta 25 GB
V100	16 GB	2 vCPU	hasta 52 GB RAM

Google Colaboratory

gradient

GPU	GPU RAM	CPUs	RAM
M4000	8 GB	8 vCPU	30 GB
P4000	8 GB	8 vCPU	30 GB
P5000	16 GB	8 vCPU	30 GB
RTX4000	8 GB	8 vCPU	30 GB
RTX5000	16 GB	8 vCPU	30 GB
A4000	16 GB	8 vCPU	45 GB
A5000	24 GB	8 vCPU	45 GB
A6000	48 GB	8 vCPU	45 GB

HARDWARE

Plataforma	Memoria RAM	GPU
Google Colab	4 GB	90%
Paperspace Gradient	5.4 GB	100%

REGISTRO Y OPTIMIZACIÓN DEL ENTRENAMIENTO

El entrenamiento consta de funcionalidades que permitirán registrar y guardar los mejores modelos obtenidos durante el proceso mediante "Callbacks" que también permiten configurar la parada del entrenamiento cuando alguno de sus parámetros no esté mejorando.

HISTORIAL DE ENTRENAMIENTO

Una vez finalizado el entrenamiento se pueden guardar los valores de exactitud y pérdidas, también pueden ser graficados para un posterior análisis de resultados.

Al final del proceso se pueden realizar las pruebas con los datos asignados para esta fase, verificando los resultados obtenidos en el entrenamiento.

RESULTADOS

Durante el período de entrenamiento se fueron variando parámetros de número de neuronas LSTM y longitud de la secuencia de imágenes evaluando cuáles serían los óptimos para una buena clasificación de videos.

PRUEBAS Y EXPERIMENTOS

En este apartado se presentarán los mejores resultados obtenidos en dos períodos de entrenamiento, con un total de cuatro variantes del modelo, el primero correspondiente al primer período en el que se trabajó únicamente con 130 videos de 5 segundos, y los tres siguientes correspondientes a variantes del modelo en los que se empleó una mayor cantidad de datos con duraciones de 5, 10 y 15 segundos.

PRIMERA VARIANTE DEL MODELO

Número de LSTM: 25

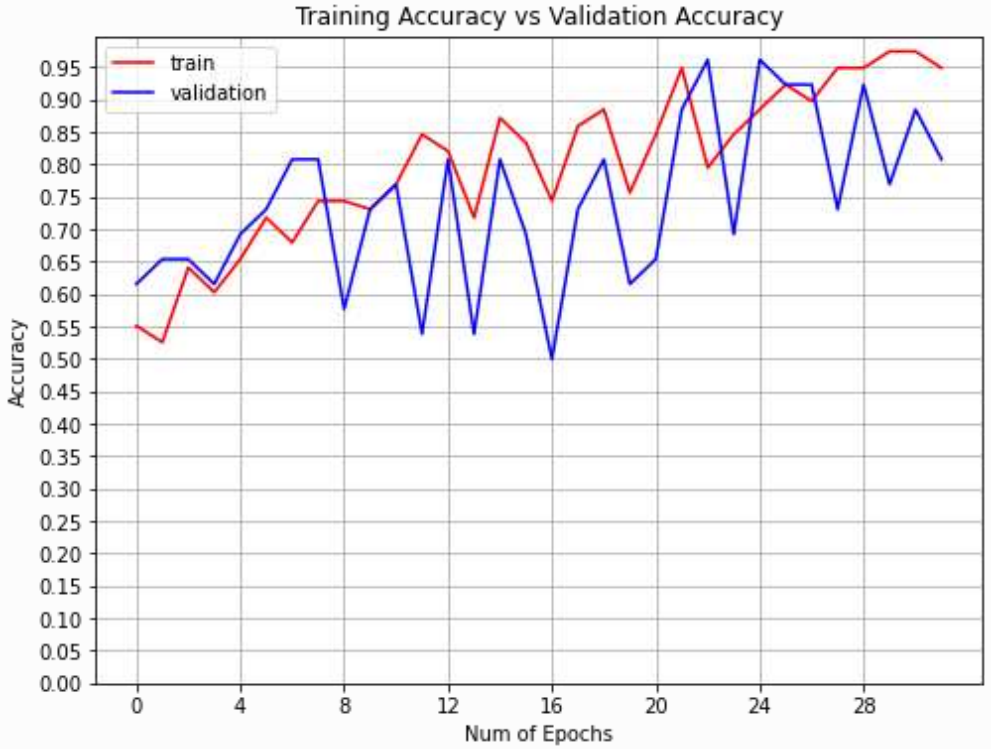
Número de Fotogramas: 40

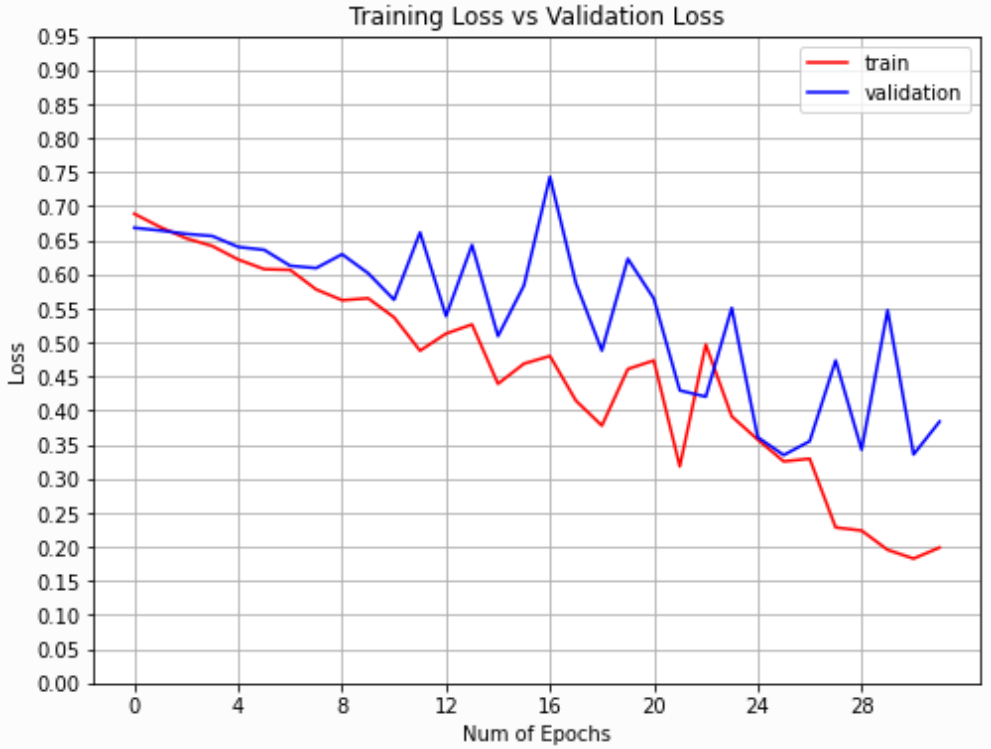
Velocidad de Fotogramas (fps): 8

Número de Datos: 130

Duración de Videos (s): 5

GRÁFICOS DE EXACTITUD Y PÉRDIDAS DE LA PRIMERA VARIANTE DEL MODELO





RESULTADOS DE EXACTITUD Y PÉRDIDAS DE LA PRIMERA VARIANTE DEL MODELO

Épocas de entrenamiento	Exactitud de Entrenamiento	Pérdidas de Entrenamiento	Exactitud de Validación	Pérdidas de Validación	Exactitud de Prueba	Pérdidas de Prueba
9	0.80	0.45	0.73	0.53	0.69	0.69
16	0.94	0.22	0.79	0.48	0.58	0.72
19	0.98	0.15	0.81	0.47	0.54	0.75

SEGUNDA VARIANTE DEL MODELO

Número de LSTM: 20

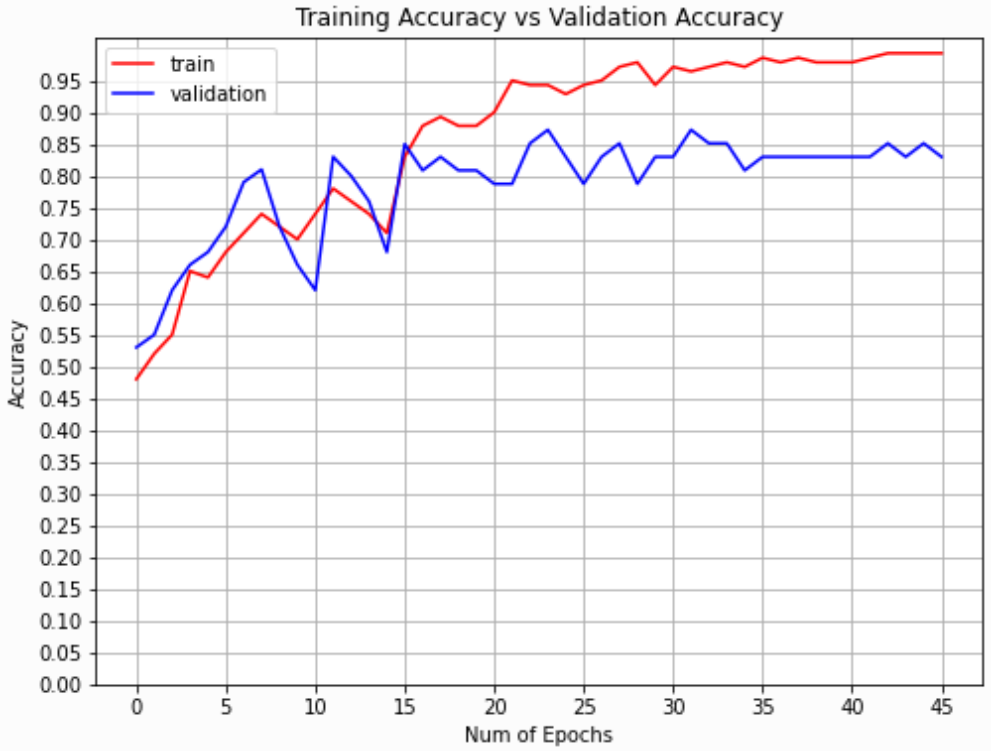
Número de Fotogramas: 60

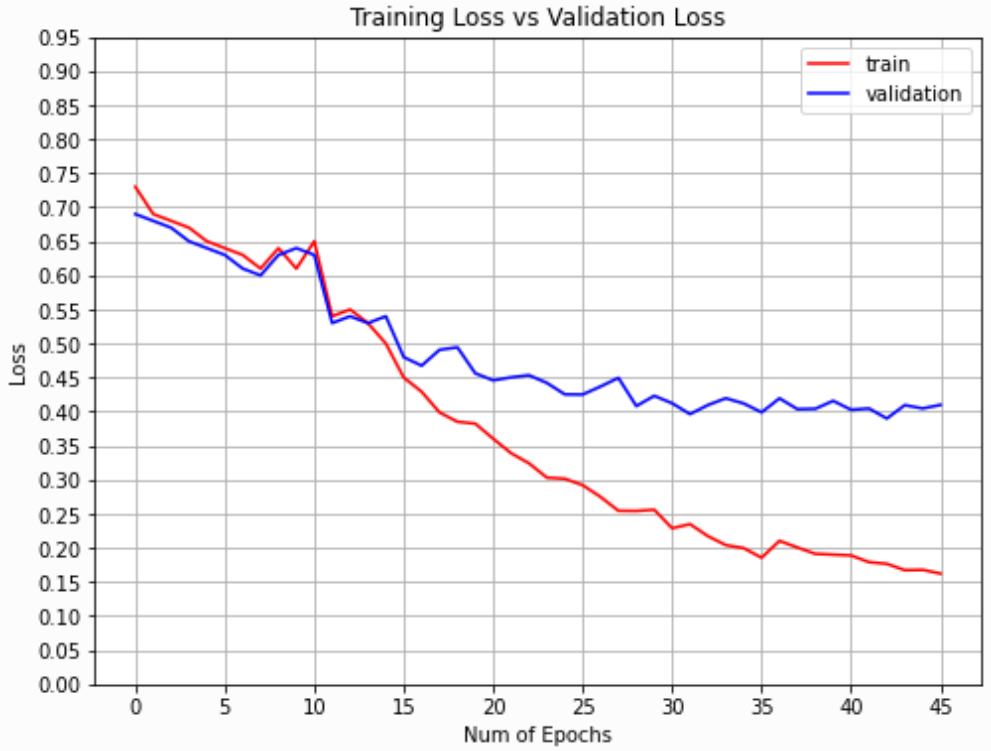
Velocidad de Fotogramas (fps): 6

Número de Datos: 233

Duración de Videos (s): 10

GRÁFICOS DE EXACTITUD Y PÉRDIDAS DE LA SEGUNDA VARIANTE DEL MODELO





RESULTADOS DE EXACTITUD Y PÉRDIDAS DE LA SEGUNDA VARIANTE DEL MODELO

Épocas de entrenamiento	Exactitud de Entrenamiento	Pérdidas de Entrenamiento	Exactitud de Validación	Pérdidas de Validación	Exactitud de Prueba	Pérdidas de Prueba
16	0.94	0.30	0.87	0.44	0.74	0.55
25	0.93	0.30	0.83	0.43	0.74	0.55
32	0.96	0.23	0.87	0.40	0.76	0.52

TERCERA VARIANTE DEL MODELO

Número de LSTM: 15

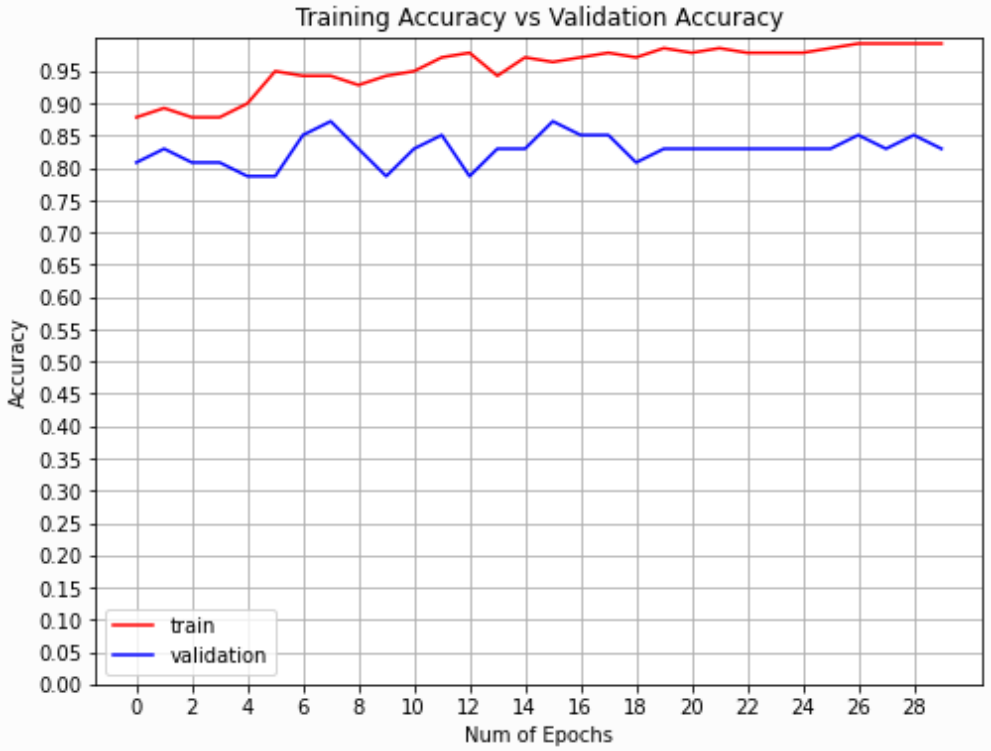
Número de Fotogramas: 40

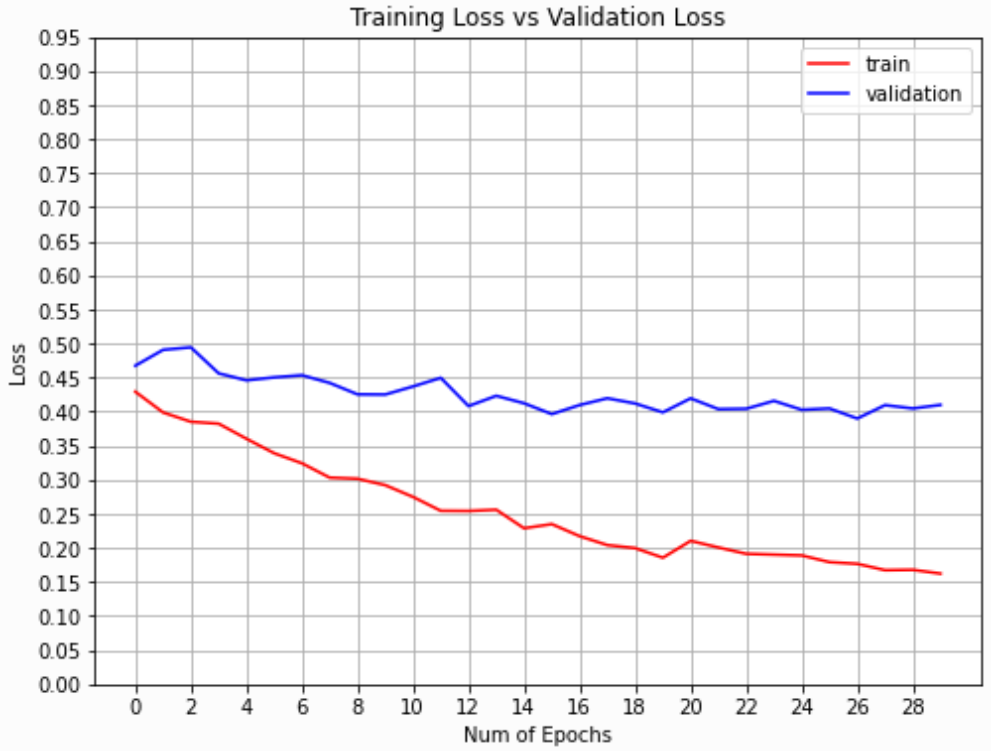
Velocidad de Fotogramas (fps): 8

Número de Videos: 240

Duración de Videos (s): 5

GRÁFICOS DE EXACTITUD Y PÉRDIDAS DE LA TERCERA VARIANTE DEL MODELO





RESULTADOS DE EXACTITUD Y PÉRDIDAS DE LA TERCERA VARIANTE DEL MODELO

Épocas de entrenamiento	Exactitud de Entrenamiento	Pérdidas de Entrenamiento	Exactitud de Validación	Pérdidas de Validación	Exactitud de Prueba	Pérdidas de Prueba
18	0.89	0.34	0.83	0.45	0.71	0.62
19	0.93	0.26	0.81	0.46	0.71	0.62
23	0.97	0.18	0.83	0.44	0.69	0.65

CUARTA VARIANTE DEL MODELO

Número de LSTM: 20

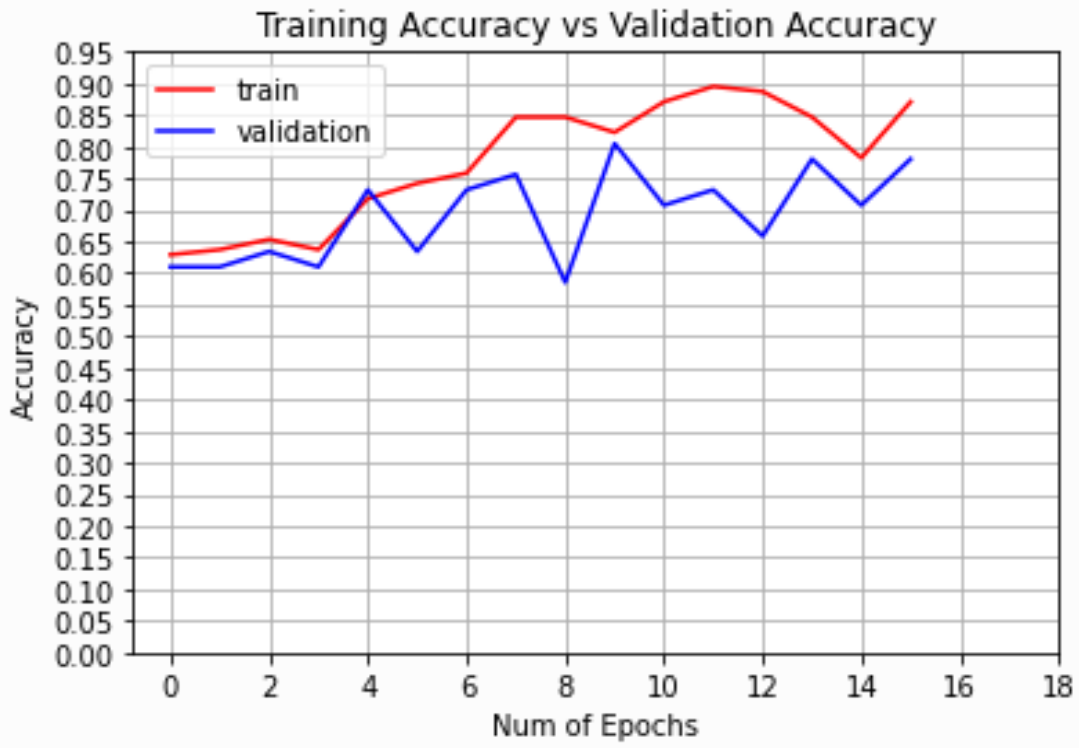
Número de Fotogramas: 90

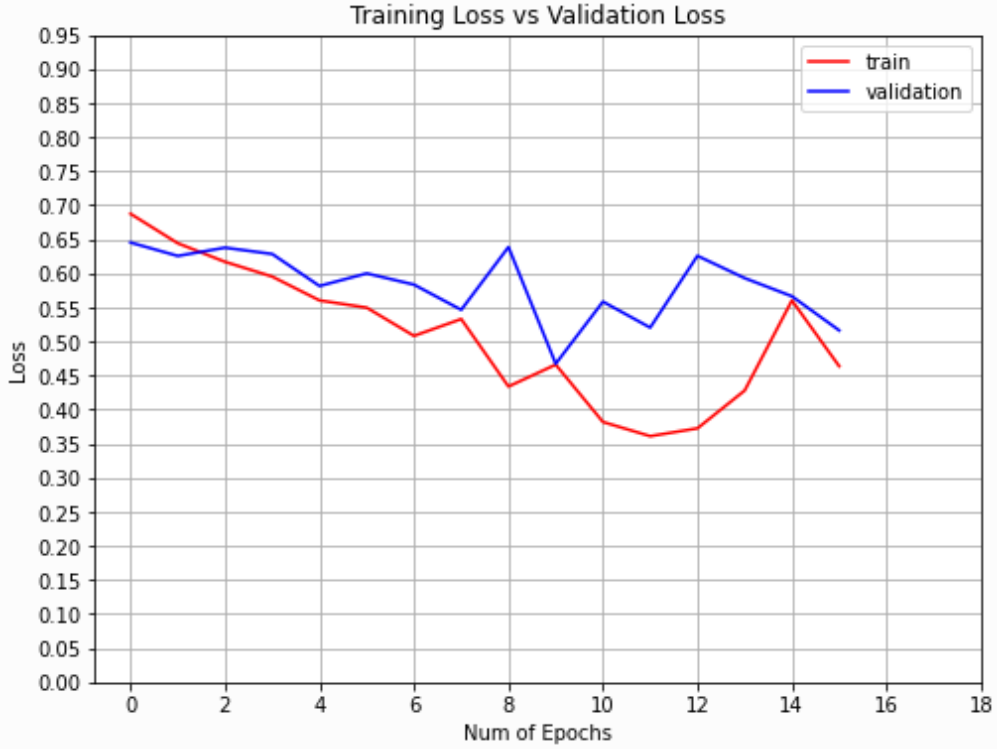
Velocidad de Fotogramas (fps): 6

Número de Datos: 206

Duración de Videos (s): 15

GRÁFICOS DE EXACTITUD Y PÉRDIDAS DE LA CUARTA VARIANTE DEL MODELO





RESULTADOS DE EXACTITUD Y PÉRDIDAS DE LA CUARTA VARIANTE DEL MODELO

Épocas de entrenamiento	Exactitud de Entrenamiento	Pérdidas de Entrenamiento	Exactitud de Validación	Pérdidas de Validación	Exactitud de Prueba	Pérdidas de Prueba
10	0.82	0.47	0.80	0.47	0.61	0.67
11	0.86	0.35	0.78	0.44	0.71	0.65
19	0.96	0.23	0.80	0.40	0.71	0.70

CAPTURAS DEL FUNCIONAMIENTO DEL MODELO EN VIDEOS DE PRUEBA

ANÁLISIS DE LOS RESULTADOS

Para este análisis se escogerá el mejor resultado de cada variante del modelo para evaluar su rendimiento y establecer realmente cuál es el mejor, se realizará la matriz de confusión de cada uno y se calcularán las métricas de precisión considerando la importancia de la detección del asalto.

MÉTRICAS DE LA PRIMERA VARIANTE DEL MODELO

		PREDICCIÓN	
		Asalto Positivos	Normal Negativos
VALOR REAL	Asalto Positivos	8	5
	Normal Negativos	3	10

Métrica	Valor
Exactitud	0.6923
Recall	0.6154
Especificidad	0.7692
Falsa Alarma	0.2308

MÉTRICAS DE LA SEGUNDA VARIANTE DEL MODELO

		PREDICCIÓN	
		Asalto Positivos	Normal Negativos
VALOR REAL	Asalto Positivos	15	7
	Normal Negativos	4	20

Métrica	Valor
Exactitud	0.7609
Recall	0.6818
Especificidad	0.8333
Falsa Alarma	0.1667

MÉTRICAS DE LA TERCERA VARIANTE DEL MODELO

		PREDICCIÓN	
		Asalto Positivos	Normal Negativos
VALOR REAL	Asalto Positivos	17	7
	Normal Negativos	7	17

Métrica	Valor
Exactitud	0.7083
Recall	0.7083
Especificidad	0.7083
Falsa Alarma	0.2917

MÉTRICAS DE LA CUARTA VARIANTE DEL MODELO

		PREDICCIÓN	
		Asalto Positivos	Normal Negativos
VALOR REAL	Asalto Positivos	14	4
	Normal Negativos	8	15

Métrica	Valor
Exactitud	0.7073
Recall	0.7778
Especificidad	0.6522
Falsa Alarma	0.3478

ANÁLISIS DE LOS RESULTADOS

Una vez presentados los resultados se destaca la segunda variante del modelo, debido a su exactitud del 76.09%, y su probabilidad de falsa alarma de 16.67%, también destaca la cuarta variante por su recall del 77.78% que representa la cantidad de asaltos que reconocería sin embargo tiene una exactitud del 70.73% y una probabilidad de falsa alarma de 34.78%.

COMPARATIVA DE MODELOS UTILIZADOS EN CLASIFICACIÓN DE VIDEO

Modelo	Exactitud
VGG-16	72.66
VGG-19	71.66
FlowNet	71.33
DEARESt	76.66
Residual LSTM	78.43
VGG16-LSTM	76.09

CONCLUSIONES

Se desarrolló un clasificador de video mediante un modelo de redes neuronales de aprendizaje profundo VGG16-LSTM que tiene una exactitud del 76.09% , como una opción válida para la detección de ciertos eventos como el asalto a peatones.

Se ha generado una base de datos representativa de los eventos a tratar con alrededor de 679 clips de video de 5, 10 y 15 segundos de duración, representando escenas con asalto y escenas normales sin asalto, para el entrenamiento, validación y pruebas del modelo del clasificador.

CONCLUSIONES

El entrenamiento del modelo clasificador obtuvo mejores resultados con el uso de clips de video de 10 y 15 segundos de duración, sin embargo por la alta demanda de recursos computacionales que se requiere para su procesamiento, los clips de 10 segundos serían los más adecuados.

La combinación de modelos de redes VGG16 y LSTM logra un mejor desempeño en la clasificación de video que las de modelo simple VGG16, VGG19 y FlowNet.

RECOMENDACIONES

Incrementar la cantidad de datos representativos del problema a resolver, especialmente los clips de video de 10 y 15 segundos de duración, la calidad y cantidad de los datos puede ser crucial para el éxito en el desarrollo de soluciones basadas en clasificación de video.

Hacer uso de herramientas computacionales en la nube como Google Colaboratory o Gradient de PaperSpace pueden ser de gran ayuda, ya que permiten un mayor alcance de las soluciones de software por su capacidad de procesamiento y hardware comparado con un computador de escritorio, especialmente por el uso de GPU.

TRABAJOS FUTUROS

El modelo de clasificador de vídeo binario propuesto funciona con videos en tiempo diferido, y se puede mejorar su desempeño incrementando los datos de entrenamiento y variando la forma de procesamiento de los datos antes de ser ingresados al clasificador, por tanto en un futuro se podrían implementar otros algoritmos para clasificadores multiclase de video en tiempo real.