Minería de Datos
1er parcial – Primavera 2022

Nombres:

- Cota Cagal Lucero 201905938
- Hernández Zamora André Mariam 201938325
- Mejía Ramírez José Carlos 201930734
- Vidal Lumbreras Ramiro 201909848

Calificación: _____

Instrucciones: Analiza cada una de las preguntas que se exponen a continuación. Realiza las implementaciones correspondientes, las cuales se deberán entregar en la fecha indicada por el calendario oficial de la institución. En caso de duda, favor de contactar al profesor.

El primer parcial consiste en entrenar una red neuronal artificial (RNA) para la detección de incendios forestales en imágenes digitales. Para ello considerar los siguientes puntos:

1. (1.0 pts) Construir una muestra (conjunto de imágenes) de incendios forestales, a partir de una estrategia de muestreo propuesta para el problema. Definir con claridad la estrategia seleccionada, estratos considerados, número de instancias, entre otros factores relevantes que consideren para la construcción de la muestra.

Fase 1. Recolección

- *Búsqueda*: La construcción de la muestra da comienzo con la búsqueda en internet de imágenes de incendios forestales.
- *Descarga y filtro 1*: Durante la descarga de cada imagen se realizó una primera pre-evaluación (filtro) la cual consistió en verificar que la imagen a descargar no se tratara de una duplicación de otra. Este primer filtro de duplicación fue verificado al momento de querer guardar el recurso visual en la carpeta destino y encontrar que el nombre coincidía con el de una imagen existente en la carpeta.

  Por lo anterior es que se plantearon dos acciones que fueron llevados a cabo según fuera el caso:

  1. Si las imágenes (imagen a guardar e imagen existente) tenían el mismo nombre y eran idénticas, entonces la imagen que se planeaba a guardar pasaba a ser descartada.
  2. Si el nombre de las imágenes eran iguales, pero visualmente diferentes, entonces se realizaba una re-asignación de nombre a la imagen a guardar.

Debido a la organización de esta primera fase y tomando en cuenta que aún quedarían duplicaciones sin revisar es que se planteó la descarga de más de 1,000 imágenes, siendo exactos, 1,560.

Fase 2. Preparación de la muestra.

- *Escaneo (filtro 2)*: Si bien en la fase 1 se manejó un filtrado para evitar duplicados, es importante señalar que en dicha fase sólo se verificaron duplicados por nombre y no estrictamente por contenido. Esto dio como resultado que la muestra recolectada aún contuviera duplicados. Gracias al uso de softwares especializados se llevaron a cabo escaneos al directorio de la muestra para analizarla e identificar imágenes que fueran similares, los cuales una vez identificados fueron eliminados.

  De este proceso de escaneo se eliminaron poco más de 440 imágenes (466).

- *Conversión*: Para un mejor planteamiento y optimización de la red neuronal se siguió el consejo de manejar a todas las imágenes con un mismo formato de extensión (JPG). Sin embargo, al revisar las instancias de la muestra previamente recolectada se encontraron cantidades considerables de imágenes que tenían una extensión diferente (BMP, PNG, TIF) lo cual a grandes rasgos se traducía en un posible riesgo para el proceso de su estudio (clasificación). Para resolver esto se recurrió al uso de software convertidor de formatos de imágenes.

  Alrededor de 100 imágenes fueron convertidas a JPG.

- *Redimensión.* Todas las instancias de la muestra fueron redimensionadas a un tamaño de 550 x 400 píxeles. Esto con el objetivo de tener una mejor gestión de las instancias y con un peso no exagerado.

Fase 3. Validación de la muestra.

Luego de las fases 1 y 2 es que entra la fase de validación de la muestra que consistió simplemente en verificar que no se hubieran pasado por alto ningún detalle analizado en las fases anteriores.

Finalmente las especificaciones de la muestra final perfectamente validada fueron los siguientes:

Imágenes de incendio

- 1382 imágenes
- Formato JPG
- Dimensiones de 550 x 400 píxeles

Imagenes de no incendio

- 818 imágenes
- Formato JPG,PNG

Imágenes de humo

- 335 imágenes
- Formato JPG,PNG

2. (1.5 pts) Implementar un proceso de particionamiento para el etiquetado de zonas locales de una imagen con incendio forestal. El nivel de granularidad del particionamiento será

definido y justificado. Se deberá permitir el etiquetado de las zonas locales como Incendio, Humo, no incendio.Las etiquetas de clase deberán recuperarse al momento de generar el vector de características de cada imagen local.

Para la división adecuada del banco de imágenes que se recolectaron a lo largo del proceso de este proyecto y tomando en cuenta la naturaleza del código desarrollado y su implementación, cada imagen de manera individual al momento de ingresar al programa será:

1. Separada en secciones, es decir, particiones. En el programa se puede indicar el tamaño que se desea que sean las particiones. En este caso se opta por el tamaño de 100 x 100 pixeles.

Img. demostrativa de partición de imagen.

2. Después de que la imagen haya sido particionada el modelo entrará en acción e iniciará el proceso de predicción. Dicha predicción estará basada en la detección de ciertas características de la imagen, por mencionar en este paso se tuvieron tres características primordiales.

    - ➢ Incendio
    - ➢ NO incendio
    - ➢ Humo

    Una vez realizada la predicción a cada partición se le registrará con una etiqueta, dicha etiqueta podrá ser uno de los tres tipos anteriores.

3. Posterior a que el modelo haya tanto particionado la imagen como realizado las predicciones correspondientes, el resultado obtenido pasará a ser almacenado en una nueva carpeta que contendrá las sub-imágenes (particiones) que ya han sido etiquetadas en su totalidad, por ejemplo:

Carpeta con imagen particionada y etiquetada.

3. (1.0 pts) Definir un conjunto de valores característicos a extraer de las imágenes locales a partir de una imagen de entrada. Enlistar cada uno de ellos, definiendo con claridad cómo se calculan y argumentar brevemente por qué cada valor puede aportar a un proceso de entrenamiento de una RNA (considerar como base los valores característicos basados en métricas de tendencia central, así como otros descriptores propios de imágenes digitales – por lo menos otros 5 descriptores).

RGB

En este modelo se buscaba particionar las imágenes entrantes de manera adecuada por lo que se le asigna una intensidad a los ya bien conocidos colores primarios, de esta manera cada pixel de la imagen se representa mediante un valor que identificara la intensidad de cada uno de los colores que ya mencionamos (rojo,azul, verde), se mezclarán hasta acercarse al color más verdadero del pixel.

Se sabe que por defecto que un píxel puede ser representado con un único valor entre 0 y 255.

TONOS CÁLIDOS

Como norma general, los colores cálidos son los que van del rojo al amarillo, pasando por naranjas, marrones y dorados. Para simplificar, suele decirse que cuanto más rojo tenga un color en su composición, más cálido será. El fuego siempre será relacionado con este tipo de tonos.

COLORES FRÍOS

Por otro lado, los colores fríos son todos los tonos que van desde el azul al verde, además de los morados. Cuanto más azul tenga un color, más frío será. Los colores fríos son los tonos del invierno, de la noche y lagos, etc. Los bosques son representativos de este tipo de tonos al contener elementos como los anteriores.

4. (1.0 pts) Realizar un análisis y argumentar que tipo de validación cruzada (que valor de K o valores) se usará en su experimentación, considerando que se debe usar la mayor cantidad de cruces posibles, garantizando en todo momento la correcta medición de lo que sucede en cada clase.

Por problemas de organización y una mala gestión del tiempo se optó por un sistema que no etiquetara las instancias. En lugar de esto se etiquetaron dependiendo en la carpeta en la que se encontraran para después obtener sus valores de características durante el proceso de entrenamiento.

5. Definir la metodología a utilizar para el entrenamiento de la RNA. Para ello, se deberá argumentar:

a. (0.5 pts) Número de neuronas (rango) que se usará para la RNA

Primero se comenzará trabajando con la propuesta de 256 neuronas. El cual a simple vista parece ser un buen número base, sobre todo si tomamos en cuenta que el tamaño inicial de cada imagen es de 550 x 400 píxeles.

b. (0.5 pts) Número de capas ocultas a usar así como rango de valores para el número de épocas

Las capas ocultas a usar serán 8.

1. La primera capa se encargará de recibir las instancias que se le darán por input y las re-escala, después hará una convolución de un tamaño de 3x3.
2. La segunda extraerá los datos de los canales rgb en una matriz de 2x2.
3. La tercera hará una convolución, pero ahora de 3x3.
4. La cuarta sacará los datos de esos valores rgb.
5. La quinta ordenará los datos obtenidos.
6. La sexta se encargará de obtener los datos para posteriormente entrenar el modelo con la función SGB donde se definirán el número de neuronas a ocupar.
7. La séptima capa funcionará para desactivar las neuronas y así mismo para evitar el sobreajuste.
8. Finalmente la última capa la cual tendrá 3 neuronas se ocupará para clasificar los 3 tipos de imágenes: Humo, Incendios, y No incendios.

Para las épocas se plantea comenzar con un valor mínimo de 10 y un máximo de 25, aunque en caso de ser necesario podría aumentarse a 30.

c. (0.5 pts) Valores a usar para el Learning Rate

Como valor de Learning Rate se utilizará un rango de 0.01 ya que no se quiere pasar por cambios muy bruscos resultados ya sean de algún error o por culpa de nuestras muestras.

d. (0.5 pts) Valores a usar para el Momentum

Se plantea mantenerlo con un valor de 0

6. (2.5) Realizar un proceso de búsqueda de parámetros que permitan encontrar la mejor clasificación que se pueda obtener con la RNA y la muestra a trabajar. Para ello, describir:

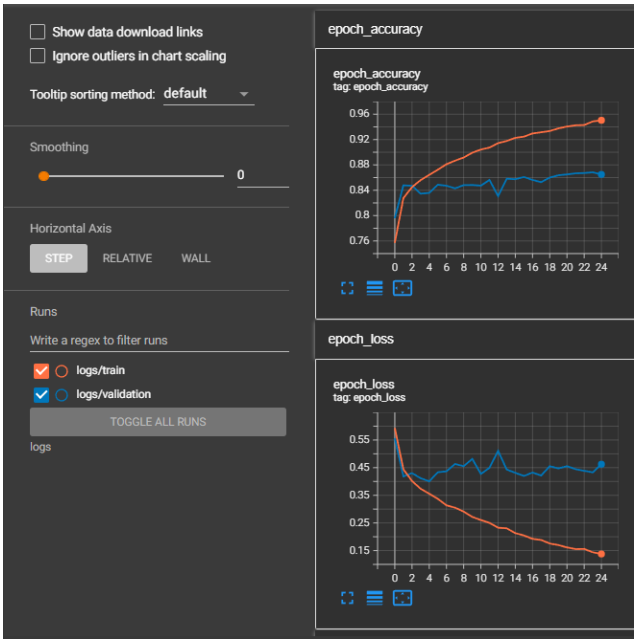
a. Mostrar a través de una gráfica y explicar cuántas neuronas presentan los mejores resultados por cada capa oculta (de acuerdo a la cantidad de capas que se definió en el punto 4.b)
b. Mostrar a través de una gráfica los experimentos realizados para determinar cuál es el número de épocas donde se maximizan los resultados de precisión
c. Mostrar y explicar los experimentos realizados para encontrar los valores ideales del Learning Rate y Momentum que optimizan la precisión global

Momentum: Ya que hubo un error al momento de querer aplicar este parámetro en pruebas iniciales de nuestro modelo y al detectarse sufrimiento de subidas y bajadas bruscas durante el aprendizaje, es que nos preocupamos y se decidió por dejarlo seteado con el valor de 0.

Modelo 8

Número de neuronas utilizadas: 256
Número de épocas: 25

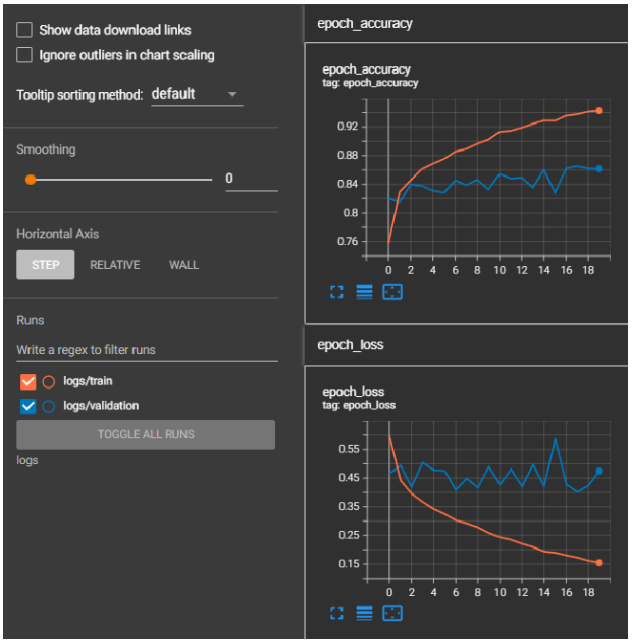
Fue en la 8va prueba del modelo donde se notaron los mejores cambios. Ya estábamos más cerca del resultado que queríamos, fue el que mejor desempeño dio con ejemplos reales, y se pensó en dejarlo como referencia para posteriores pruebas, pues se planeó mejorarlo.



En comparación con pruebas anteriores el proceso de este modelo no era tan tardado y daba buenos resultados. En resultados anteriores con un número de épocas por debajo de 25 notamos un rendimiento muy pobre en los resultados, así que para este modelo se optó por aumentar su valor y los resultados mejoraron mucho más, se empezaba a notar que la tendencia en la curva de aprendizaje no subía ni bajaba.

Modelo 9

Número de neuronas: 512
Epocas: 20



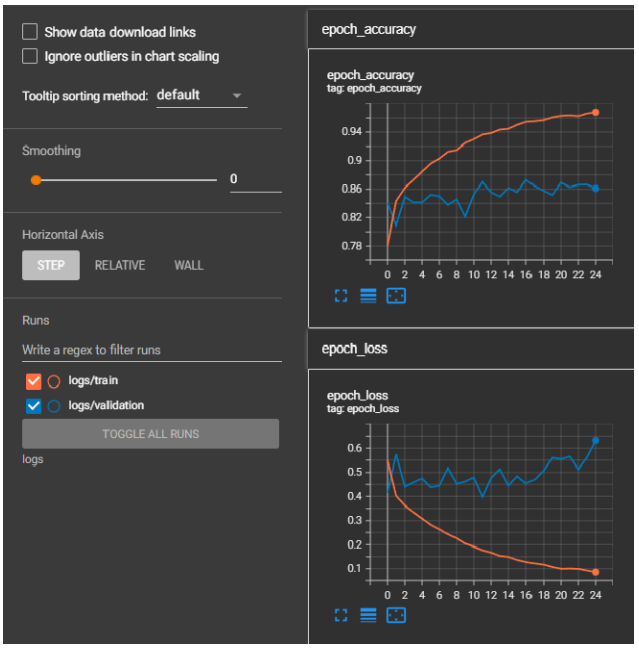
Este fue el siguiente intento. En cuanto a épocas, primero se optó por hacer una reducción en su cantidad y ya que 256 neuronas funcionaron bien pensamos que el doble serían mejores. Esto principalmente con el objetivo de ver si se lograba alguna mejora con menores épocas, pero con mayores neuronas.

No obstante, además de que el modelo era más inestable nos terminó por dar los peores resultados en las pruebas, así que se descartó. Sin embargo, se tomó como base para el siguiente experimento.

Modelo 10

Número de neuronas: 512
Épocas: 25

Ya teniendo en cuenta el experimento pasado pensamos que solo sería cuestión de añadir más épocas para que las neuronas pudieran trabajar mejor, así que se agregaron 5 épocas más, pero el resultado fue completamente opuesto a lo esperado. Notamos que los resultados irían a peor, así que este modelo se descartó y se tomó como bueno el 8vo modelo.



7. (1.0 pts) Hacer un análisis de los resultados obtenidos, donde se explique la forma en la cual influyó la búsqueda de parámetros para obtener un mejor resultado en el proceso de clasificación.

Fue más que nada prueba y error, ya que no estábamos muy seguros de cómo afectaría cada campo por separado y mucho menos junto. Se iban agregando un poco más de instancias en cada modelo, así que creemos que gracias a eso nuestro modelo carece un poco de crecimiento durante más épocas de entrenamiento. Además, nuestra muestra presenta algo de problema haciendo que nuestra red confunda a veces el “No incendio” con “Humo”.