INFOTEC

Programa de Maestría en Ciencia de Datos

Aplicación de YOLOV5 para identificar obstrucciones en espacios comunes de hogares.

Trabajo de titulación

que para optar por el grado de

Maestro en Ciencia de Datos

PRESENTA:

Héctor Ariel Aragón Oliva

Director de tesis:

México, CDMX. (Marzo) 2023

Capítulo 1

Introducción

En el presente trabajo se brinda la investigación realizada sobre la aplicación de redees neuronales convolucionales para hacer la detección de espacios donde no hay obstrucciones dentro de una casa. La idea de este desarrollo es brindar una solución que permita identificar en ciertos lugares de una casa si hay obstrucciones o no, y la funcionalidad de esto puede ser en diferentes lineas, sin embargo, esto podría ser de ayuda para brindar una base para una solución que permita a las personas con alguna discapacidad visual prevenir algún accidente al estar en algún espacio nuevo para ellos.

Es importante entender que, dada la aplicación o la propuesta del presente trabajo, es pensando en las situaciones que pueden enfrentar las personas que tienen alguna discapacidad visual, lo anterior debido a que si, por ejemplo, una persona invidente va a una casa de algún conocido con quien no ha interactuado o que simplemente no conoce algún sitio a donde va, este tipo de situaciones pueden resultar sumamente complicada para las personas que tienen alguna discapacidad visual, por lo que contar con alguna herramienta o tecnología que les permita acceder a lugares de una forma más factible y cómoda permitirá que las personas puedan tener una mejor calidad de vida y brindar seguridad a conocer lugares nuevos. Esto motiva el presente trabajo para poder brindar un enfoque de aplicación de un framework muy popular dentro del campo de visión por computadora, el cual es YOLO, en particular el trabajo se enfoca en la quinta versión, es decir, yolov5, y bajo este framework se desarrollará la solución final. Con respecto

a esta última solución, es que se va a aplicar a un set de datos, donde se encuentran etiquetados los espacios que pueden ser una obstrucción, espacios que no representan o tienen una obstrucción y objetos que pueden representar una semi obstrucción.

El trabajo busca abordar el concepto de redes neuronales para después introducir el concepto de redes neuronales convolucionales (CNN, por sus siglas en inglés) y eventualmente explicar el estado del arte de la solución YOLO (You Only Look Once), en particular, la versión 5, ya que existen diferentes versiones de este framework de object detection.

Una vez desarrollado el estado del arte y marco teorico se busca aplicar el framework que se ha mencionado anteriormente (yolov5), lo que se hizo en este aspecto es buscar bajo un proceso de hiper-parametrización los parámetros que permitan aproximar los mejores resultados, para después entrenar la arquitectura de YOLO y así poder obtener y analizar los resultados de la detección de objetos.

Capítulo 2

Estado del arte

Las redes neuronales convolucionales (CNN) son un tipo de red neuronal diseñada específicamente para el procesamiento de datos de imágenes. A diferencia de una red neuronal estándar, donde cada neurona está conectada a todas las neuronas de la capa siguiente, una CNN utiliza capas convolucionales que aplican filtros a partes de la imagen. Esta técnica permite que la red neuronal identifique patrones y características en las imágenes a medida que se desplaza por la capa convolucional. Las capas convolucionales están seguidas por capas de agrupamiento, que reducen la dimensión de la información mientras conservan las características importantes. Finalmente, las capas de clasificación utilizan las características extraídas para clasificar la imagen en una categoría.

Cada capa convolucional utiliza filtros para extraer características de la imagen. Un filtro es una pequeña matriz de números que se desliza sobre la imagen y se multiplica elemento por elemento con los píxeles de la imagen que se superponen con el filtro. El resultado de esta multiplicación se suma y se coloca en una nueva matriz llamada "mapa de características". Los filtros se aprenden durante el entrenamiento de la red neuronal y pueden detectar bordes, texturas, formas y otros patrones en la imagen. Cuanto más profunda sea la capa convolucional, más complejas serán las características que se pueden extraer.

Las capas de agrupamiento reducen la dimensión de los mapas de características mediante la combinación de píxeles en regiones más grandes. El agrupamiento se puede hacer de diferentes maneras, pero la técnica más común es el "max pooling", donde se selecciona el valor máximo de un área de píxeles. El agrupamiento reduce la cantidad de parámetros en la red neuronal y hace que la red sea más resistente a las pequeñas variaciones en la posición de las características en la imagen.

Finalmente, las capas de clasificación utilizan las características extraídas para determinar la clase a la que pertenece la imagen. Las capas de clasificación están formadas por neuronas que se conectan a todas las neuronas de la capa anterior y calculan una puntuación para cada clase posible. La puntuación se puede transformar en una probabilidad utilizando la función softmax, que normaliza los valores y los convierte en una distribución de probabilidad.

De tras de las redes de tipo CNN son fundamentales las operaciones convolucionales para su funcionamiento. La operación de convolución se puede expresar matemáticamente como una multiplicación de matrices entre el filtro y un parche de la imagen. La operación de agrupamiento se puede ver como una selección del valor máximo dentro de una región de la imagen. La operación de las capas completamente conectadas se puede describir matemáticamente como una multiplicación de matrices entre las activaciones de la capa anterior y una matriz de pesos. Durante el entrenamiento de la red neuronal, se utilizan técnicas de optimización, como el descenso del gradiente, para ajustar los valores de los filtros y las matrices de pesos de las capas completamente conectadas.

Una capa de grupo (group layer) es una técnica utilizada en redes neuronales convolucionales para reducir la cantidad de parámetros y mejorar la eficiencia computacional. Según Krizhevsky et al. (2012), esta técnica consiste en dividir los canales de entrada en grupos y realizar convoluciones separadas en cada grupo.

En otras palabras, como explica Zhang et al. (2018), una capa de grupo divide la

entrada en G grupos y aplica C/G filtros en cada grupo, donde C es el número de canales de entrada. Luego, las salidas de cada grupo se concatenan para formar la salida final de la capa. Esto puede reducir el número de parámetros de la red y aumentar la eficiencia computacional.

Matemáticamente, la operación de convolución en una capa de grupo se puede expresar como una multiplicación de matrices entre el filtro y una porción de la entrada correspondiente al grupo. Según Krizhevsky et al. (2012), esta operación se puede describir como:

$$y_i = \sigma(\sum_{j=1}^{C/G} W_{i,j} * x_{g(j)})$$

donde y_i es la i-ésima característica de salida, $W_{i,j}$ es el filtro de convolución de la i-ésima característica de salida y el j-ésimo grupo de entrada, $x_{g(j)}$ es la entrada del j-ésimo grupo, C es el número total de canales de entrada, G es el número de grupos y σ es la función de activación.

Como se acaba de explicar, una capa de grupo es una técnica utilizada en redes neuronales convolucionales para reducir la cantidad de parámetros y mejorar la eficiencia computacional al realizar convoluciones separadas en grupos de canales de entrada y concatenar las salidas de cada grupo. La operación matemática de convolución en una capa de grupo se puede expresar como una multiplicación de matrices y una función de activación.

Una capa completamente conectada (fully connected layer) es una capa en una red neuronal en la que todas las neuronas de una capa están conectadas a todas las neuronas de la capa siguiente. Según Goodfellow et al. (2016), estas capas son comúnmente utilizadas al final de una red convolucional para clasificación o regresión.

Matemáticamente, una capa completamente conectada se puede expresar como una multiplicación de matrices entre la entrada y los pesos de la capa, seguida de una fun-

ción de activación. Como explica Nielsen (2015), si la entrada a la capa es un vector x de tamaño n, y la capa tiene m neuronas, entonces los pesos de la capa son una matriz W de tamaño $m \times n$, y la salida y de la capa se puede expresar como:

$$y = \sigma(Wx + b)$$

donde σ es la función de activación, b es el vector de sesgos de tamaño m, y la suma y la multiplicación se realizan elemento a elemento.

La capa completamente conectada se utiliza comúnmente en redes neuronales para clasificación. Según Goodfellow et al. (2016), después de pasar la entrada a través de varias capas convolucionales, se puede aplanar la salida de la última capa convolucional en un vector y pasar este vector a través de una o varias capas completamente conectadas. La salida final de la red es la predicción de la red para la entrada.

Por lo tanto una capa completamente conectada es una capa en una red neuronal en la que todas las neuronas de una capa están conectadas a todas las neuronas de la capa siguiente. Matemáticamente, la salida de una capa completamente conectada se puede expresar como una multiplicación de matrices entre la entrada y los pesos de la capa, seguida de una función de activación. Esta capa se utiliza comúnmente en redes neuronales para clasificación o regresión.

Una capa convolucional (convolutional layer) es una capa en una red neuronal que utiliza la operación de convolución para extraer características de la entrada. Según Goodfellow et al. (2016), una capa convolucional consiste en un conjunto de filtros, donde cada filtro es una pequeña matriz que se desliza sobre la entrada para producir una característica en la salida.

Matemáticamente, una capa convolucional se puede expresar como la convolución de la entrada x con un conjunto de filtros W, seguida de una función de activación no lineal σ y una operación de sesgo b. Como explica Nielsen (2015), si la entrada a la

capa es un tensor de tamaño $n_h \times n_w \times n_c$, donde n_h es la altura, n_w es el ancho y n_c es el número de canales de la entrada, y la capa tiene n_f filtros de tamaño $f_h \times f_w \times n_c$, entonces la salida z de la capa se puede expresar como:

$$z_{i,j,k} = \sigma(\sum_{u=1}^{f_h} \sum_{v=1}^{f_w} \sum_{c=1}^{n_c} W_{u,v,c,k} x_{i+u-1,j+v-1,c} + b_k)$$

donde $z_{i,j,k}$ es el valor de la salida en la posición (i, j, k), $W_{u,v,c,k}$ es el valor del filtro en la posición (u, v, c, k), $x_{i+u-1,j+v-1,c}$ es el valor de la entrada en la posición $(i + u - 1, j + v - 1, c)$, y b_k es el sesgo para el k-ésimo filtro.

La capa convolucional se utiliza comúnmente en redes neuronales para procesamiento de imágenes. Según Goodfellow et al. (2016), una capa convolucional es capaz de detectar características locales en una imagen, como bordes, texturas y patrones simples, y las capas convolucionales posteriores combinan estas características para detectar características más complejas.

En pocas palabras, una capa convolucional es una capa en una red neuronal que utiliza la operación de convolución para extraer características de la entrada. Matemáticamente, la salida de una capa convolucional se puede expresar como la convolución de la entrada con un conjunto de filtros, seguida de una función de activación no lineal y una operación de sesgo. Esta capa se utiliza comúnmente en redes neuronales para procesamiento de imágenes.

Bibliografía

[1] oodfellow, Ian, et al. "Deep Learning."MIT Press, 2016.

[2] eCun, Yann, et al. Çonvolutional neural networks.Çommunications of the ACM 61.6 (2018): 514-529.

[3] ielsen, Michael. "Neural Networks and Deep Learning."Determination Press, 2015.

[4] rizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. Ïmagenet classification with deep convolutional neural networks..Advances in neural information processing systems. 2012.

[5] hang, Xiangyu, et al. "ShuffleNet: An extremely efficient convolutional neural network for mobile devices."Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.