Cuestionario de Teoría-2

Visión por Computador

Daniel Bolaños Martínez

1. Justificar adecuadamente las respuestas.

1.1. Identifique las semejanzas y diferencias entre los problemas de: a) clasificación de imágenes; b) detección de objetos; c) segmentación de imágenes; d) segmentación de instancias.

Los cuatro términos corresponden a problemas tratados en el campo de la visión por computador.

La clasificación de imágenes consiste en reconocer en una imagen la clase que representa asignándole una etiqueta acerca del concepto representado. No se representa ni la localización del objeto en la imagen ni se detectan instancias distintas de un mismo objeto.

En la detección de objetos entra en juego la localización de imágenes, reconociendo múltiples instancias de la misma clase y la posición de los objetos existentes abarcando su ubicación con un delimitador rectangular. Además añade la información aportada por la clasificación de imágenes a cada instancia de las clases representadas.

La segmentación de imágenes consiste en dividir una imagen en segmentos que contengan información relevante, asociando a cada región una etiqueta para clasificar la imagen y obteniendo una máscara para la región que representa el concepto. No se realiza la segmentación para cada instancia de la misma clase.

En la segmentación de instancias se realiza el mismo proceso que en segmentación de imágenes pero esta vez aplicado a cada instancia de cada clase que se representa en la imagen. De forma individual, se crea una máscara de píxeles para cada objeto en la imagen representando su forma y tamaño. [1]

1.2. ¿Cuál es la técnica de búsqueda estándar para la detección de objetos en una imagen? Identifique pros y contras de la misma e indique posibles soluciones para estos últimos.

La aproximación que aporta el concepto básico en la detección de objetos es la detección por ventana deslizante (sliding window). Consiste en deslizar una ventana de tamaño variable a través de la imagen y aplicar sobre cada región, un modelo de detección para cada localización.

Las ventajas de esta técnica son la simplicidad y facilidad de su implementación, por el contrario es muy ineficiente y tiene una tasa baja de falsos positivos debido a que es necesario evaluar el modelo por cada región extraída por la ventana deslizante y además es difícil de extender a una amplia gama de escalas (relaciones de aspecto).

Para solucionarlo, podemos usar ventanas deslizantes de tamaño variable o realizar pirámides con diferentes tamaños de la imagen con la que estemos trabajando. [1]

1.3. Considere la aproximación que extrae una serie de características en cada píxel de la imagen para decidir si hay contorno o no. Diga si existe algún paralelismo entre la forma de actuar de esta técnica y el algoritmo de Canny. En caso positivo identifique cuales son los elementos comunes y en que se diferencian los distintos.

Las principales semejanzas entre los descriptores HOG y el algoritmo de Canny es que ambos utilizan como base de su funcionamiento, el cálculo de histogramas que representan la información sobre intensidad del gradiente y dirección del borde para cada punto de la imagen.

Como diferencias, podemos especificar el resto de pasos del algoritmo Canny, ya que HOG no realiza suavizado gaussiano ni supresión de no máximos en su proceso. Además mientras que Canny se usa para calcular las regiones con bordes de la imagen, HOG se utiliza en clasificación de imágenes partiendo de la idea de que la forma de un objeto local se caracteriza por su distribución de gradientes de intensidad y las orientaciones de los bordes en cada punto sin necesitar un conocimiento preciso de las posiciones correspondientes de los mismos. [2]

1.4. Tanto el descriptor de SIFT como HOG usan el mismo tipo de información de la imagen pero en contextos distintos. Diga en que se parecen y en que son distintos estos descriptores. Explique para que es útil cada uno de ellos.

Ambos descriptores, se basan en el concepto de particionar la imagen en bloques y calcular el histograma de gradientes orientados en cada bloque realizándose una interpolación entre ángulos vecinos.

La principal diferencia es que SIFT utiliza una gaussiana (de mitad del tamaño de la ventana) centrada en el bloque para ponderar los valores en todo el descriptor.

SIFT es más adecuado para describir la relevancia de cada punto de la imagen, debido a la ponderación gaussiana involucrada, mientras que HOG se utiliza más en clasificación de imágenes donde tengamos todas las características en un vector 1D. [3]

1.5. Observando el funcionamiento global de una CNN, identifique que dos procesos fundamentales definen lo que se realiza en un pase hacia delante de una imagen por la red. Asocie las capas que conozca a cada uno de ellos

Los dos procesos fundamentales que se realizan en una red neuronal convolucional son:

- **Extracción de características**: en esta parte, la red realizará una serie de convoluciones y operaciones de agrupación durante las cuales se detectarán las características. En este proceso se aplican las capas de convolución (**Conv2D**) que aplican un filtro de convolución a la imagen, agrupación (**MaxPooling2D, AveragePooling2D**) que reducen el tamaño de las imágenes y activación (**ReLu**) que aportan la no linealidad a la red.
- **Clasificación**: aquí, las capas totalmente conectadas servirán como clasificador sobre las características extraídas. Asignarán una probabilidad para poder predecir el concepto que representa la imagen. En este proceso se aplican las capas totalmente conectadas (**Dense**) que realizan productos de vectores 1D y la activación (**softmax**) que transforma la salida de las neuronas de la CNN en la probabilidad de cada imagen de pertenecer a cada clase. [1]

1.6. Se ha visto que el aumento de la profundidad de una CNN es un factor muy relevante para la extracción de características en problemas complejos, sin embargo este enfoque añade nuevos problemas. Identifique cuales son y qué soluciones conoce para superarlos.

Añadir nuevas capas a una CNN ocasiona generalmente dos problemas:

- Aumento del tiempo de entrenamiento de la red. Al aumentar el tamaño de la misma aumentará el tiempo que tarda en entrenar el conjunto de imágenes.
- Aumento del overfitting del conjunto de datos con la derivación de un aumento de falsos positivos. Esto se puede deber a que un aumento de las capas y mayor profundidad de la red puede descubrir en el proceso de entrenamiento, irregularidades en los datos a partir de los cuales se sobreentrenen las imágenes.

Para solucionar el problema de tiempo de ejecución y overfitting, podemos añadir capas de regularización (**Dropout**) que asignan una probabilidad de activación a cada neurona durante el entrenamiento, de tal forma que dificultará el entrenamiento y mejorará la validación del conjunto de datos, por lo que de esta forma podremos reducir el sobreentrenamiento de forma considerable.

Podemos usar aumento de datos (**Data Augmentation**) añadiendo nuevas muestras al conjunto de datos de entrenamiento que reducirán el overfitting y aumentarán la tasa de acierto del conjunto de validación.

1.7. Existe actualmente alternativas de interés al aumento de la profundidad para el diseño de CNN. En caso afirmativo diga cuál/es y como son.

Podemos hacer uso de diferentes capas y técnicas para mejorar la extracción de características y con ella la tasa de acierto de nuestra CNN sin aumentar significativamente el tamaño nuestra red.

Añadir capas de regularización (**Dropout**) que añaden una probabilidad de desactivación a las neuronas de la red durante el entrenamiento reduciendo la tasa de acierto en el conjunto de entrenamiento pero mejorándola en el de validación y prueba.

Añadir capas de normalización (**BatchNormalization**) que normalizan las activaciones de la capa anterior en cada batch, haciendo que todas las capas reciban el

gradiente con una extensión similar.

Hacer una red densa añadiendo capas de salto entre las capas de activación no lineal (Aumentar **Skip Connections**). También se puede usar un filtro de convolución pequeño (3x3) para no perder precisión a la hora de extraer características y aumentar la anchura (número de filtros de salida de cada capa) para extraer más características o la cardinalidad usando módulos convolucionales del mismo tipo en paralelo. [1]

1.8. Considere una aproximación clásica al reconocimiento de escenas en donde extraemos de la imagen un vector de características y lo usamos para decidir la clase de cada imagen. Compare este procedimiento con el uso de una CNN para el mismo problema. ¿Hay conexión entre ambas aproximaciones? En caso afirmativo indique en que parecen y en que son distintas.

El funcionamiento global de una CNN viene determinado por dos procesos (extracción de características y clasificación de imágenes).

En el proceso de extracción utilizamos varias capas para conseguir la ponderación de las características para cada imagen.

En el proceso de clasificación es necesario convertir los filtros de salida (que normalmente serán matrices) a un vector 1D haciendo uso de la capa **Flatten** que concatena las filas de las matrices en un sólo vector. A continuación, se aplicarán las capas totalmente conectadas que veamos necesarias y finalmente usaremos una capa de activación **softmax** para conseguir las predicciones de ese vector de características sobre las clases de imágenes.

Por tanto, podemos decir que ambos procesos son similares variará la forma en la que se obtienen las características del vector, que en el caso de la CNN, estará especificado por el tipo de capas y parámetros y en la aproximación clásica descrita la que se proponga.

1.9. ¿Cómo evoluciona el campo receptivo de las neuronas de una CNN con la profundidad de la capas? ¿Se solapan los campos receptivos de las distintas neuronas de una misma profundidad? ¿Es este hecho algo positivo o negativo de cara a un mejor funcionamiento?

Definimos el campo receptivo de una neurona a la extensión espacial de la conectividad de cada neurona con una región local del volumen de entrada, esto evita realizar conexiones a todas las neuronas del volumen anterior de cada capa.

El tamaño del campo receptivo de una unidad aumentará de forma directamente proporcional a la profundidad de la red, ya que cada capa adicional aumenta el tamaño del campo receptivo por el tamaño del núcleo. El impacto del campo receptivo sigue una distribución gaussiana que genera el solapamiento de los campos receptivos de las neuronas que se encuentran a una misma profundidad. Este hecho influirá de forma positiva en nuestra red, la cual aprenderá características mejores y más profundas.

A medida que aumenta el campo receptivo y el solapamiento, se puede obtener más información sobre el contexto completo de la imagen y mejorar el aprendizaje sobre la naturaleza de las imágenes con las que se entrena. [4]

1.10. ¿Qué operación es central en el proceso de aprendizaje y optimización de una CNN?

La operación que es común a los procesos de entrenamiento y optimización en una CNN es la capa de activación ReLu.

La capa ReLu es un factor clave en el proceso de aprendizaje de las CNN, ya que introducen no linealidad al proceso de entrenamiento y mejora la extracción de características profundas de las imágenes, sin las cuales no sería posible el proceso de clasificación.

En la optimización, la capa ReLu soluciona el problema de desvanecimiento del gradiente que ocurre cuando una CNN no puede propagar información útil del gradiente desde el extremo de salida del modelo hacia las capas cercanas al extremo de entrada. Esto ocasiona problemas en el aprendizaje sobre el conjunto de datos o en la convergencia prematura a una mala solución. [6]

1.11. Compare los modelos de detección de objetos basados en aproximaciones clásicas y los basados en CNN y diga que dos procesos comunes a ambos aproximaciones han sido muy mejorados en los modelos CNN. Indique cómo.

Los problemas comunes que ofrecen las aproximaciones clásicas están relacionadas con la naturaleza de las imágenes que se pretenden clasificar (clasificación de imágenes) y la forma de extracción de las regiones sobre las que detectar los objetos (detección de objetos).

Esto se debe a que en los modelos clásicos, las plantillas utilizadas para la detección de objetos generalmente no son suficientes para detectar la categoría de la imagen, ya que muchos de ellos pueden representar cambios en su configuración o en el punto de vista desde el que se está representando la instancia.

Para solucionar estos problemas, los modelos CNN implementan la extracción y evaluación de un alto número de regiones de una imagen que mejoran la forma de detectar los objetos y aumentan la precisión en redes orientadas a nuestro conjunto de clases. Además, algunas de ellas utilizan segmentación jerárquica y detectores de bordes entrenados para mejorar la búsqueda de objetos en regiones y aumentar la rapidez de la evaluación de los mismos. [1]

1.12. Es posible construir arquitecturas CNN que sean independientes de las dimensiones de la imagen de entrada. En caso afirmativo diga cómo hacerlo y cómo interpretar la salida.

Existen redes con la propiedad de ser inmunes al tamaño de la imagen de entrada gracias al comportamiento de la capa que construye la red. Algunos de los ejemplos son:

Full Convolutional Networks (FCN): interpretan las capas **Dense** (que necesitan prefijar el tamaño del input con antelación por definición) como si fueran capas de convolución las cuales no tienen limitaciones en el tamaño de entrada porque una vez que se describen los tamaños de kernel y de paso, puede generar salidas de dimensión variable de acuerdo con las entradas correspondientes. [1]

Spatial Pyramid Pooling (SPP): agregan una capa SPP entre la última capa convolucional y la primera capa fully connected. Calcula un mapa de características de toda la imagen y luego las agrupa en regiones arbitrarias generando salidas de longitud fija, que luego se pasarán a las capas completamente conectadas u otros clasificadores.

En general, añade información en un etapa más profunda de la jerarquía de la red para evitar la necesidad de recortar o deformar al principio de la misma. [5]

1.13. Suponga que entrenamos una arquitectura Lenet-5 para clasificar imágenes 128x128 de 5 clases distintas. Diga que cambios deberían de hacerse en la arquitectura del modelo para que se capaz de detectar las zonas de la imagen donde aparecen alguno de los objetos con los que fue entrenada.

Utilizaremos la idea que implementa R-CNN para clasificar diferentes objetos de una misma imagen, añadiendo un proceso de búsqueda selectiva sobre la imagen de entrada para elegir regiones de la misma. Estas regiones se seleccionan en múltiples escalas por lo que tendrán diferentes tamaños.

Transformaremos cada región propuesta a las dimensiones de entrada requeridas por la red, que en nuestro caso será la CNN proporcionada (**Lenet-5**) previamente entrenada y la usaremos para calcular las características extraídas para cada región.

Combinaremos las características y la categoría etiquetada de cada región para hacer posible la clasificación de objetos. Las características y el cuadro delimitador etiquetado de cada región propuesta se combinan para que puedan ser usados en el entrenamiento de un modelo de regresión lineal y obtener así la predicción del cuadro delimitador real de la clase.

R-CNN es el caso teórico para la clasificación de objetos de diferentes regiones de una misma imagen, este caso tiene problemas en el tiempo de extracción de regiones, que pueden ser solucionados aplicando la extracción después de aplicar la CNN sobre la imagen completa. Esta mejora, se implementa en el modelo (Fast R-CNN). [7]

1.14. Argumente por qué la transformación de un tensor de dimensiones 128x32x32 en otro de dimensiones 256x16x16, usando una convolución 3x3 con stride=2, tiene sentido que pueda ser aproximada por una secuencia de tres convoluciones: convolución 1x1 + convolución 3x3 + convolución 1x1. Diga también qué papel juegan cada una de las tres convoluciones.

La utilización de tres convoluciones en lugar de una, mejorará la eficiencia de los cálculos y además obtendrá unos resultados muy similares ya que las convoluciones añadidas funcionan de la siguiente manera:

La convolución 1x1 disminuye el tamaño de los filtros de salida sin modificar el valor de las características extraídas, las cuales, partiendo de una red preentrenada que ofrezca buenos resultados, representará la misma información del problema.

Seguidamente, podremos hacer la convolución 3x3 que obtendrá un resultado aproximado al original y además reducirá el coste de los cálculos debido a que el tamaño de filtros de entrada se ha reducido considerablemente.

Finalmente, aplicaremos otra convolución 1x1 para aumentar el tamaño de los filtros de salida y obtener unas dimensiones equivalentes a las de realizar la convolución 3x3 con stride=2, las características extraídas ofrecerán un resultados que se aproximará bastante al primer caso.

1.15. Identifique una propiedad técnica de los modelos CNN que permite pensar que podrían llegar a aproximar con precisión las características del modelo de visión humano, y que sin ella eso no sería posible. Explique bien su argumento.

El término de campo receptivo es el que aproxima el funcionamiento de los modelos CNN a la manera de operar del cerebro humano. El campo receptivo es una región local en el volumen de salida de la capa anterior a la que está conectada una neurona en nuestra red.

Este concepto, emula al funcionamiento de las primeras capas de la corteza visual en la forma de detectar las características locales obtenidas por los receptores visuales y combina la información resultante para crear patrones más complejos de forma

jerárquica. [8]

Referencias

[1] Diapositivas de clase.

[2] N.Dalal, B.Triggs. Histograms of Oriented Gradients for Human Detection.

[3] https://stackoverflow.com/questions/24619210/dense-sift-vs-hog

[4] W.Luo, Y.Li, R.Urtasun, R.Zemel. Understanding the Effective Receptive Field in Deep CNN.

[5] K.He, X.Zhang, S.Ren, J.Sun. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition.

[6] Blog: J.Brownlee. How to Fix the Vanishing Gradients Problem Using the ReLU.

[7] R.Girshick, J.Donahue, T.Darrell, J.Malik Rich feature hierarchies for accurate object detection and semantic segmentation.

[8] H.Le, A.Borji. What are the Receptive, Effective Receptive, and Projective Fields of Neurons in Convolutional Neural Networks?