Cuestionario 2

Luis Antonio Ortega Andrés

5 de diciembre de 2019

Pregunta 1. *Identifique las semejanzas y diferencias entre los problemas de:*

- *Clasificación de imágenes.*
- *Detección de objetos.*
- *Segmentación de imágenes.*
- *Segmentación de instancias.*

En el problema de clasificación de imágenes buscamos asignar una o varias etiquetas a la imagen, mientras que en el problema de detección de objetos, buscamos detectar regiones determinadas por un *bounding box* que encierren objetos de una clase concreta (no nos preocupamos de la forma del objeto). Para asignar la etiqueta a cada *bounding box* empleamos un método de clasificación de imágenes.

En el problema de segmentación de imágenes buscamos asignar una única etiqueta a cada píxel de la imagen, de forma que fraccionamos la imagen en distintas regiones. Sin embargo, en la detección de objetos, un píxel puede pertenecer a varias *bounding box* o a ninguna.

En el problema de segmentación de instancias tratamos diferentes objetos de la misma clase como instancias distintas, reconociendo así la forma del objeto detectado, sin embargo, en la segmentación de imágenes los objetos de la misma clase son tratados como una misma instancia. En ambos problemas, cada píxel pertenece a una instancia, además, la segmentación de instancias es mas difícil que la segmentación de imágenes.

Tanto en la segmentación de imágenes como en la segmentación de instancias y la detección de objetos, la localización es importante, mientras que en la clasificación, no nos importa donde se encuentren los objetos de interés, solo buscamos etiquetar la imagen global.

Pregunta 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 técnica de búsqueda estándar es utilizar una *ventana deslizante* o *sliding window* y pasar cada región por un clasificador. Si este clasificador detecta un objeto con una probabilidad mayor a un umbral, dibujamos la *bounding box* utilizando la ventana correspondiente y la etiquetamos con el resultado del clasificador.
Sus principales ventajas son:

- Es un modelo sencillo y fácil de implementar.
- El método de detección de objetos es el mismo en toda la imagen.

- La elección del clasificador es libre, pudiendo elegir uno que se adapte a los objetos que queremos detectar.

Sus principales inconvenientes:

- Es necesario prefijar el tamaño y el ratio de la ventana. Esto supone un problema debido a los diferentes tamaños y formas de los objetos a reconocer. Una posible solución es pasar la misma ventana por distintas escalas de la imagen. También podemos utilizar un enfoque en *region proposals*, buscando regiones que sean más propensas a contener objetos.
- Múltiples respuestas para un mismo objeto y falsos positivos. Este problema se puede solucionar utilizando supresión de no-máximos.

Pregunta 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.*

El paralelismo presente en ambos es que en el algoritmo de Canny extraemos una característica de cada píxel, la intensidad del gradiente. Sin embargo, este algoritmo no extrae más características sino que utiliza una serie de pasos para «pulir» la información que tiene.

Pregunta 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.*

A diferencia de HOG que extrae características de la imagen al completo, el descriptor SIFT calcula una serie de puntos de interés mediante transformaciones geométricas y luego extrae características de cada uno de ellos utilizando la información de su entorno.

En ambos descriptores se calcula un histograma de gradientes sobre una región de la imagen, luego se concatenan dichos histogramas y se extrae un vector de características. La diferencia recae en la región sobre la que se calculan dichos histogramas.
En HOG, se divide la imagen completa en cuadrículas y se calcula sobre cada una de estas. Sin embargo, en SIFT, se elige una ventana alrededor de cada punto de interés y sobre este se hace una cuadrícula.

El descriptor HOG se suele utilizar en detección de objetos y extracción de características para más tarde utilizarlas en clasificación de imágenes o regiones de imágenes. Sin embargo, debido a que SIFT es invariante frente a muchas transformaciones (traslaciones, rotaciones, cambios de escala e iluminación…), se puede utilizar tanto en detección de objetos como en reconocimiento de imágenes y *panorama stitching*.

Pregunta 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 definen un pase hacia delante de una imagen por una red son la extracción de características y la clasificación.

Las capas cuya función es la extracción de características son:

- Capas convolucionales.

- *Max-pooling*

Aquellas cuya función es la clasificación son:

- Capas totalmente conectadas.

Las capas de activación no lineal (por ejemplo Relu) y las de regularización (*Batchnormalization, Dropout*...) pueden asociarse a ambos procesos.

Pregunta 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.*

Los problemas que presenta añadir profundidad a las redes convolucionales son:

- *Vanishing gradients*. Como la información del gradiente se actualiza utilizando *back-propagation*, los cambios que se realizan en las primeras capas son relativamente pequeños. Esto implica que las primeras capas «apenas» aprenden.
- Pérdida de información. Es el problema análogo al anterior, debido a la gran cantidad de operaciones, aquellas características extraídas por las primeras capas pueden ser eliminadas.
- Largos tiempos de entrenamiento. Debido al gran número de operaciones que se realizan. Para solucionar esto podemos utilizar convoluciones 1×1 para disminuir la dimensionaldiad y el tiempo de cómputo.
- *Overfitting*. Al aumentar el número de capas, aprendemos mas características, lo cual puede provocar que el modelo no sea capaz de generalizar. Esto lo podemos solucionar utilizando capas de normalización o dropout.

Una solución común a los tres primeros problemas es la utilización de *skip connections* para conectar la salida de una capa con la entrada de otra mas profunda.
Otra solución genérica a todos los problemas es utilizar técnicas de profundidad estocástica, en estas, durante el entrenamiento «apagamos» algunas capas de la red.

Pregunta 7. *¿Existen actualmente alternativas de interés al aumento de la profundidad para el diseño de CNN? En caso afirmativo diga cuál/es y como son.*

Existen ciertas alternativas al aumento de profundidad:

- Realizar un aumento en la anchura de las capas de la red, esto implica una mayor extracción de características.
- Aumentar la cardinalidad en un cierto número de capas. Esto nos permite crear modelos mas complejos sin aumentar la profundidad.

Pregunta 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.*

Tanto el procedimiento descrito como las redes neuronales convolucionales extraen características de la imagen y las utilizan para decidir la clasificación de la misma. Sin embargo, al utilizar redes convolucionales, no elegimos las características que son extraídas ni la forman en la que se extraen. Además, Al

utilizar una red convolucional, tampoco elegimos como utilizamos dichas características para clasificar la imagen.

Pregunta 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?*

El campo receptivo de una neurona en una profundidad concreta, depende directamente del tamaño del filtro, el campo receptivo de las neuronas de la profundidad anterior y el *stride* de las neuronas de las profundidades anteriores.

El campo receptivo de las neuronas de una misma profundidad se solapa si el *stride* es más pequeño que el tamaño del *kernel* de dicha capa.

El hecho de que el campo receptivo de una neurona se solape con sus adyacentes es en general algo positivo debido a que de esta forma comparten cierta información. Esto puede resultar beneficioso a la hora de detectar características y obtener mayor información de la imagen. Sin embargo, también puede resultar en *overfitting* pues dichas neuronas se «acostumbran» a trabajar juntas y no aprender características individualmente.

Pregunta 10. *¿Qué operación es central en el proceso de aprendizaje y optimización de una CNN?*

La operación central es el cálculo de gradientes, para ello se utiliza *backpropagation*. El objetivo de esto es optimizar la función de pérdida.

Pregunta 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 procesos comunes a ambas que han sido muy mejorados por las CNN son la extracción de características y la propuesta de regiones.

En los modelos clásicos, se proponen o encuentran regiones para más tarde extraer características de la misma para detectar objetos, en cambio, con el uso de las CNN se ha conseguido que sea la propia red la que propone las regiones, clasifica los objetos y luego refinan la localización.

Gracias a las técnicas de las redes convolucionales basadas en el reuso de características, se mejora notablemente la prupuesta de regiones y las *bounding boxes*.

Además, la extracción de características que implementan las redes convolucionales se ha visto mejorada debido a que las aprenden automaticamente del problema concreto.

Pregunta 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.*

Podemos construir redes convolucionales que sean independientes de las dimensiones de la entrada, para ello, debemos tener especial cuidado con las capas totalmente conectadas. Las capas convolucionales no presentan un problema ya que solo dependen del tamaño de su filtro (en caso de que la imagen fuera más pequeña que el *kernel*, se podría solucionar utilizando un borde en la imagen).

Para abordar el problema de las capas totalmente conectadas podemos utilizar una capa *GlobalAveragePooling* para deshacernos de la dimensión de la entrada, pudiendo interpretar la salida de igual forma que la original.
También podemos optar por no utilizar capas totalmente conectadas o sustituirlas por convoluciones 1×1, de esta forma la salida de la red sería un tensor 3D que corresponde con las características aprendidas por la red.

Pregunta 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.*

El procedimiento consistirá en adaptar la red que tenemos a una red Fast R-CNN. Para ello seguimos los siguientes pasos:

- Insertamos una capa RoI Pooling previamente entrenada antes de la primera capa totalmente conectada del modelo.
- Borramos la capa final totalmente conectada.
- Añadimos dos salidas a nuestro modelo, una capa totalmente conectada con un valor más (correspondiente al *background*) y *softmax*, y un modelo de regresión que nos permitirá dilucidar las *bounding boxes*.

Pregunta 14. *Argumente por qué la transformación de un tensor de dimensiones 128×32×32 en otro de dimensiones 256×16×16, usando una convolución 3×3 con stride=2, tiene sentido que pueda ser aproximada por una secuencia de tres convoluciones: convolución 1×1 + convolución 3×3 + convoluión 1×1. Diga también qué papel juegan cada una de las tres convoluciones.*

Se puede aproximar la transformación realizada por una convolución 3×3 con *stride* 2 por una secuencia de tres convoluciones 1×1, 3×3, 1×1 (una de las dos últimas debe tener *stride* 2), esto es debido a que el conjunto de valores del tensor original que influyen en el cálculo de cada valor del tensor de salida es el mismo con ambas operaciones.
Veamos cual es la finalidad de cada una de las capas convolucionales:

- La primera convolución 1×1 se utiliza para reducir la dimensión del tensor (y así, reducir los parámetros que necesitará la siguiente capa).
- La capa 3×3 es la encargada de la extracción de características del tensor.
- La última capa la utilizamos para recuperar la dimensión que queremos del tensor.

Pregunta 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.*