Cuestionario 2 - Visión por computador

Francisco Javier Sáez Maldonado

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.

En la **clasificación de imágenes**, sabemos de antemano que hay un único objeto qu nos interesa en nuestra imagen y queremos darle una etiqueta a ese objeto que tenemos, lo cual no ocurre en el resto de problemas.

En la **detección de objetos** , a diferencia de en el resto de problemas, que tenemos un *conjunto fijo* (que es lo que lo diferencia del resto) de categorías y tenemos que, dada una imagen, cada vez que aparezca un objeto **de las categorías prefijadas** que tenemos, crear una *bounding box* alrededor de ese objeto y predecir la clase de ese objeto, no sabemos de antemano cuántos objetos tendremos en la imagen, como ocurría en el caso anterior.

En la **segmentación de imágenes**, el problema será asignar *a cada pixel* (cosa que no ocurre en ninguno de los demás casos) una categoría en esa imagen, por lo tanto , no se hará una diferenciación de las instancias de los objetos en la imagen como se tratará de hacer, al contrario que en el caso de la detección de objetos. Podrían aparecer por tanto Al final tendremos la imagen dividida en regiones importantes de la misma.

Por último, podríamos decir que la **segmentación de instancias** es “el problema completo”, pues se trata en este caso de localizar cada entidad en nuestra imagen(como en localización de objetos) pero también queremos clasificarla, pues en este caso, como en la detección de objetos, tendremos múltiples objetos en la misma imagen.

Fuente : [@cs231-stanford]

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

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.

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.

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

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.

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

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.

¿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?

¿Qué operación es central en el proceso de aprendizaje y optmización de una CNN?

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.

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.

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.

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 + convoluión 1x1. Diga también qué papel juegan cada una de las tres convoluciones.

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.

Referencias

[@cs231-stanford]: https://www.youtube.com/watch?v=nDPWywWRIRo&list=PL3FW7Lu3i5JvHM8ljYj-zLfQRF3EO8sYv&index=12&t=3s