Sostenibilidad con "edge intelligence"

Acelerador de inferencia para la IA del futuro

Hasta el momento la inteligencia artificial, tanto el entrenamiento como la inferencia en sí, ha sido desarrollada principalmente con vistas a ser aplicada en centros de cálculo. Pero con el nuevo campo de la "edge AI", esta tendencia está cambiando. Smartphones, robots, drones, cámaras de vigilancia, cámaras industriales... todos estos dispositivos incorporarán en un futuro próximo un tipo de procesamiento con inteligencia artificial. La cosa se vuelve interesante cuando la inferencia tiene lugar directamente en el aparato que entrega las imágenes. ¿Cómo se puede utilizar de forma eficiente y sostenible una tecnología de este tipo, que consume tantos recursos, fuera de los grandes centros de cálculo, en pequeños dispositivos embebidos que consumen pocos recursos? Ya existen algunos enfoques y soluciones que funcionan para acelerar eficazmente las redes neuronales en dispositivos edge. Pero solo algunas son lo suficientemente flexibles como para seguirle el ritmo al rápido avance de la IA.

Edge intelligence

En pocas palabras, este concepto hace referencia a una clase de dispositivos que pueden resolver tareas de inferencia en el extremo de la red ("on-the-edge") utilizando redes neuronales y algoritmos de aprendizaje automático. En este contexto cabe preguntarse por qué la inteligencia artificial debería utilizarse cada vez más en los dispositivos embebidos y por qué el aprendizaje profundo y las redes neuronales profundas se están enfocando hacia la industria en este momento.

Las respuestas a estas preguntas tienen que ver menos con la IA en sí y más con temas como el ancho de banda, los tiempos de latencia, la seguridad o el procesamiento de datos descentralizado. Es decir, con temas y retos clave de las aplicaciones modernas de la industria 4.0. Una tarea importante es reducir la competencia inherente por el ancho de banda del canal de comunicación compartido filtrando o convirtiendo grandes cantidades de datos de los sensores o las cámaras en información procesable ya en los propios "dispositivos edge". El procesamiento inmediato de los datos también permite tomar decisiones de proceso directamente en el punto de captura de la imagen sin latencia en la comunicación de los datos. Desde el punto de vista técnico o de la seguridad, es posible que no sea deseable o que sea muy difícil de conseguir una comunicación fiable y continua con una unidad central de procesamiento, tal vez incluso en la nube. Encapsular los datos adquiridos en los dispositivos edge también ayuda a descentralizar el almacenamiento y el procesamiento de datos, reduciendo la probabilidad de un posible ataque a todo el sistema. Porque la seguridad de los datos generados y transmitidos tiene una enorme importancia para cualquier organización.

Distribuir la inteligencia del sistema facilita además la separación clara de tareas específicas. Por ejemplo, en una fábrica pueden existir cientos de estaciones de trabajo que precisan un servicio de clasificación de imágenes para analizar un conjunto distinto de objetos en cada estación. Sin embargo, alojar varios clasificadores en la nube implica un coste. Sería deseable una solución económica que entrene todos los clasificadores en la nube y que envíe sus modelos, adaptados a cada estación de trabajo, a los dispositivos edge. Por otro lado, la especialización de cada modelo tiene un rendimiento mejor que un clasificador que hace predicciones por todas las estaciones de trabajo. Además, a diferencia de la ejecución en el centro de datos, estas sencillas soluciones específicas ahorran mucho tiempo de desarrollo. Todo ello indica que es preferible trasladar la inferencia a los dispositivos edge.

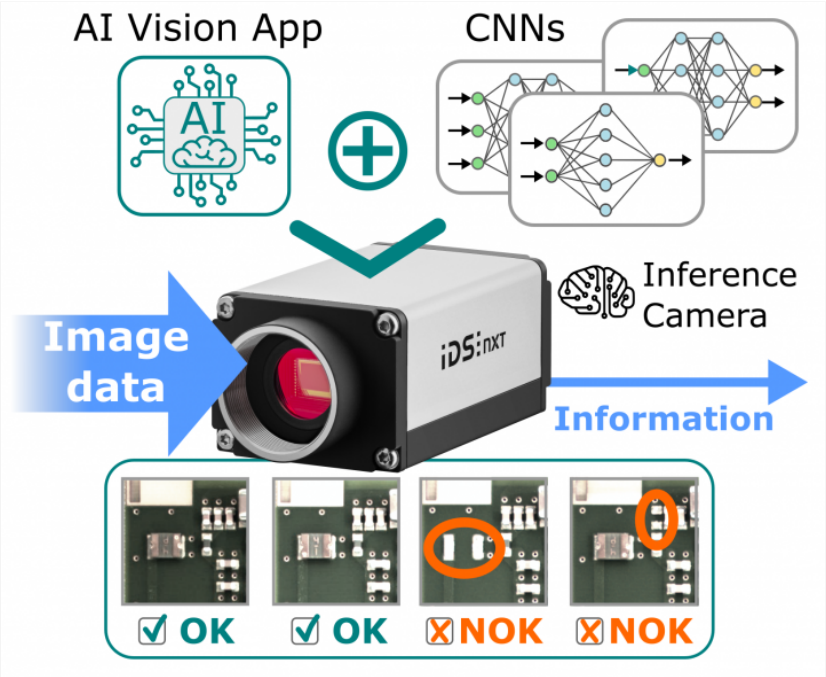

Figura 1 - Los dispositivos edge inteligentes reducen grandes cantidades de datos de los sensores y de imagen. Generan "on the edge" información que se puede utilizar directamente y transfieren solo esa información a la unidad de control.

Desafíos

¿Por qué las redes neuronales son "de hecho" inadecuadas para el uso embebido, o cuáles son los desafíos para poder utilizarlas eficazmente "on the edge"? Ejecutar tareas de inferencia de IA en dispositivos edge no es tan fácil. En el "edge computing" por lo general todo gira en torno a la eficiencia. Los dispositivos edge suelen tener una cantidad limitada de recursos energéticos, de cálculo y de almacenamiento. Por consiguiente los cálculos deben ejecutarse con una alta eficiencia y al mismo tiempo obtener un alto rendimiento y una baja latencia, unas prestaciones que se antojan incompatibles. Y la ejecución de CNN es la disciplina "reina". Precisamente las CNN son conocidas por su alta intensidad computacional, dado que requieren miles de millones de cálculos para procesar una entrada. Las CNN, con millones de parámetros que describen su arquitectura, de entrada no son buenas candidatas para la edge computing. Existen redes que utilizan un menor número de parámetros para describirlas, como MobilNet, EfficientNet o SqueezeNet, por lo que son más adecuadas para el uso embebido al reducir drásticamente los requisitos de almacenamiento y de cálculo. Pero esto no es suficiente. Para disminuir aún más los requisitos de almacenamiento, las redes se tienen que comprimir. Por ejemplo, los parámetros que no son importantes pueden eliminarse después del entrenamiento mediante "pruning", o se puede reducir mediante una cuantificación el número de bits para describir los parámetros. La disminución del tamaño de la memoria de la CNN influye positivamente en el tiempo de su procesamiento. Y esto nos lleva al último aspecto de la optimización.

A pesar de utilizar redes comprimidas y con menos parámetros, sigue siendo necesario un sistema informático adaptado específicamente a las arquitecturas para una ejecución "on the edge" eficiente de la IA. Para ello hay que prestar atención a dos características básicas del sistema. Además de la citada eficiencia, el sistema debería tener una flexibilidad que le permitiera ser compatible con nuevos desarrollos de arquitecturas de CNN. Se trata de un aspecto importante, dado que precisamente en el ámbito de la IA se investigan y desarrollan nuevas arquitecturas y tipos de capas cada mes. Lo que hoy es nuevo y de actualidad, mañana puede ser obsoleto. Por consiguiente, ¿qué opciones de plataformas existen?

Selección de la plataforma

- Sin duda, los **sistemas basados en CPU** son los que ofrecen una mayor flexibilidad. Al mismo tiempo, las CPU son muy ineficientes en la ejecución de las CNN y tampoco consumen especialmente poco.
- Las **plataformas GPU** ejecutan las CNN con una alta potencia mediante sus núcleos de cálculo o "compute cores", que trabajan en paralelo. Son más especializadas que las CPU y tienen una alta flexibilidad. Lamentablemente, las GPU consumen mucha energía, lo que es más bien problemático en aplicaciones "on the edge".
- La arquitectura de las **FPGA** programables se puede reconfigurar sobre el terreno y de ese modo adaptarse a nuevas arquitecturas de CNN. También ofrecen un buen rendimiento gracias a que trabajan en paralelo. Sin embargo, su programación exige un profundo conocimiento del hardware.
- Una **solución ASIC** completa es un circuito integrado a medida y la vencedora obvia en términos de eficiencia, ya que está optimizada para ejecutar eficazmente una determinada arquitectura CNN. Sin embargo, la flexibilidad podría plantear problemas en caso de que deje de ser compatible con arquitecturas de CNN nuevas o modificadas.

Las prestaciones de "alto rendimiento, flexibilidad y eficiencia energética" de la tecnología FPGA la convierten en la más adecuada para acelerar CNN en dispositivos edge en la actual etapa de desarrollo de la IA.

La capacidad de adaptarla a aplicaciones especiales o a las CNN en todo momento mientras el dispositivo está funcionando actualizándola con un archivo de configuración la convierte en una solución a largo plazo idónea para aplicaciones industriales. El mayor desafío a la hora de utilizar la tecnología FPGA es la complejidad de la programación, por lo que solo puede ser realizada por especialistas.

Estrategia de desarrollo

Para ejecutar redes neuronales en un "Vision Edge Device", es decir, en nuestras cámaras IDS NXT, hemos decidido (IDS) desarrollar un acelerador de CNN basado en la tecnología FPGA. Lo llamamos "deep ocean core". Para hacer que el manejo de la FPGA fuera lo más sencillo posible en su uso posterior, no había que desarrollar varias configuraciones optimizadas específicamente para diferentes tipos de CNN, sino una arquitectura de aplicación universal. Esto permite al acelerador ejecutar cualquier red CNN siempre que esté formada por capas compatibles. Sin embargo, dado que todas las capas regulares, como las capas convolucionales, las capas de adición, los diferentes tipos de capas de agrupación o las capas de "squeezing-excite" ya son compatibles, básicamente se puede utilizar cualquier tipo de capa importante. Con esto se elimina por completo el problema de la complejidad de la programación, porque el usuario no necesita tener conocimientos específicos para crear una nueva configuración de FPGA. Gracias a las actualizaciones del firmware de la cámara IDS NXT, el deep ocean core se actualiza constantemente para soportar cualquier novedad en el ámbito de la CNN.

deep ocean core

¿Cómo trabaja el acelerador universal de CNN y qué pasos son necesarios para ejecutar una red neuronal entrenada? El acelerador solo necesita una "descripción binaria" a partir de la cual pueda saber de qué capas consta la red CNN. Para ello no es necesario ningún tipo de programación. Sin embargo, una red neuronal entrenada con Keras, por ejemplo, se presenta en un "lenguaje Keras de alto nivel" que el acelerador no puede entender. Se tiene que traducir a un formato binario que se parezca a un tipo de "lista encadenada". A partir de cada capa de la red CNN se genera un descriptor de nodo que describe con precisión cada capa. El resultado final es una lista completa concatenada de la CNN con formato binario. Todo el proceso de traducción está automatizado por una herramienta por lo que, de nuevo, no es necesario tener conocimientos específicos. El archivo binario generado se carga en la memoria de trabajo de la cámara y el deep ocean core empieza a procesarlo. Ahora la red CNN funciona en la cámara IDS NXT.

Flexibilidad en la ejecución

Utilizar una representación de la CNN como una lista encadenada ofrece claras ventajas en relación con la flexibilidad del acelerador. Permite cambiar de red sobre la marcha. Y hacerlo sin problemas y sin demora. Para ello, se pueden cargar en la memoria de trabajo de la cámara varias "Linked List Representations" de diferentes redes neuronales. Para seleccionar una CNN para su ejecución, el acelerador deep ocean debe apuntar al principio de una de esas listas. Basta con cambiar un "valor de puntero" hacia una de las memorias de lista. Estamos hablando de una simple operación de escritura de un registro de la FPGA, que puede realizarse de forma muy rápida en cualquier momento.

Por medio del ejemplo siguiente se explica por qué puede ser importante este cambio rápido de CNN. Supongamos que tiene dos líneas de producción con dos tipos de productos. Usted quiere inspeccionar la calidad de los productos. Para ello primero hay que detectar su posición y luego, basándose en la categoría de producto detectada, clasificar la calidad según los defectos específicos del producto.

Se podría resolver el problema entrenando una gran red CNN para encontrar los objetos y clasificarlos al mismo tiempo, entrenando previamente cada uno de los posibles defectos para cada uno de los grupos de productos. Esto es complejo y la red se volvería muy grande y posiblemente sería lenta, pero podría funcionar. La dificultad estribaría en lograr una precisión suficiente. La posibilidad de cambiar la CNN activa sobre la marcha permite desacoplar la localización y la clasificación de los diferentes objetos para hacer que las distintas CNN sean más fáciles de entrenar. La detección de objetos sólo necesita distinguir entre dos clases y proporcionar sus posiciones. Otras dos redes se entrenan sólo con las propiedades específicas de los productos y las clases de defectos. En función del producto localizado, la aplicación de la cámara decide automáticamente qué red de clasificación se va a activar para determinar en cada caso la calidad del producto. Este procedimiento permite al dispositivo edge trabajar con tareas específicas relativamente sencillas y con pocos parámetros. Como resultado las distintas redes son mucho más pequeñas, tienen que diferenciar muchas menos características y trabajan mucho más rápido y consumiendo menos energía, lo que las hace ideales para su ejecución en un dispositivo edge.

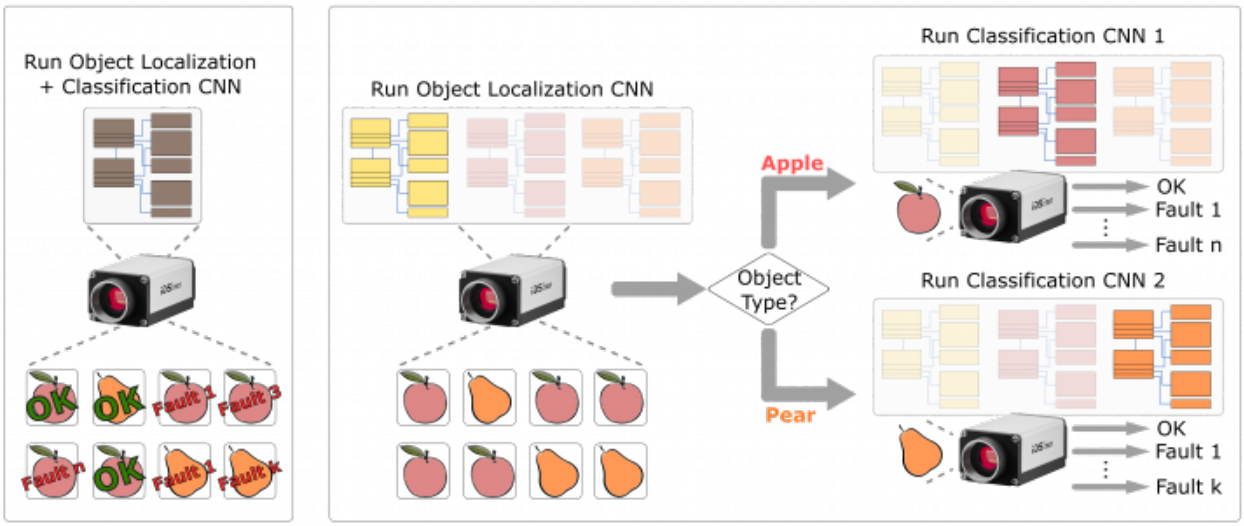

Figura 2 - La posibilidad de cambiar la ejecución de las redes neuronales sobre la marcha permite dividir el análisis de imágenes en flujos de trabajo de inferencia más sencillos que trabajan en la cámara de forma más eficaz, rápida y estable.

Rendimiento y eficiencia

El acelerador de CNN basado en FPGA de nuestras cámaras de inferencia IDS NXT funciona en un SoC Xilinx Zynq Ultrascale con 64 núcleos de cálculo. En muchas redes de clasificación de imágenes conocidas como MobileNet, SqueezeNet o EfficientNet, se alcanzan velocidades de hasta 67 fps. Incluso en familias de redes como Inception o ResNet, que se consideran demasiado complejas para la edge computing, se pueden obtener 20 fps, lo que es suficiente para muchas aplicaciones. La implementación de la FPGA también nos permite seguir desarrollando el rendimiento del acelerador deep ocean. Gracias a las actualizaciones de firmware, también se benefician de esto todas las cámaras que ya están en uso.

Sin embargo, en el tema de la edge computing, la eficiencia energética es un aspecto más importante, si cabe. Define el número de imágenes por segundo que puede procesar un sistema por cada vatio de energía. Esto hace que la eficiencia energética sea una buena medida para comparar distintas soluciones edge. A continuación se muestra una comparativa de los distintos aceleradores de CNN. El deep ocean core como implementación FPGA, la solución GPU con una Jetson TX 2A, la solución CPU clásica mediante una CPU Intel Core-i7 actual, una Raspberry Pi como solución CPU embebida y una solución totalmente ASIC representada por un chip de IA Intel Movidius.

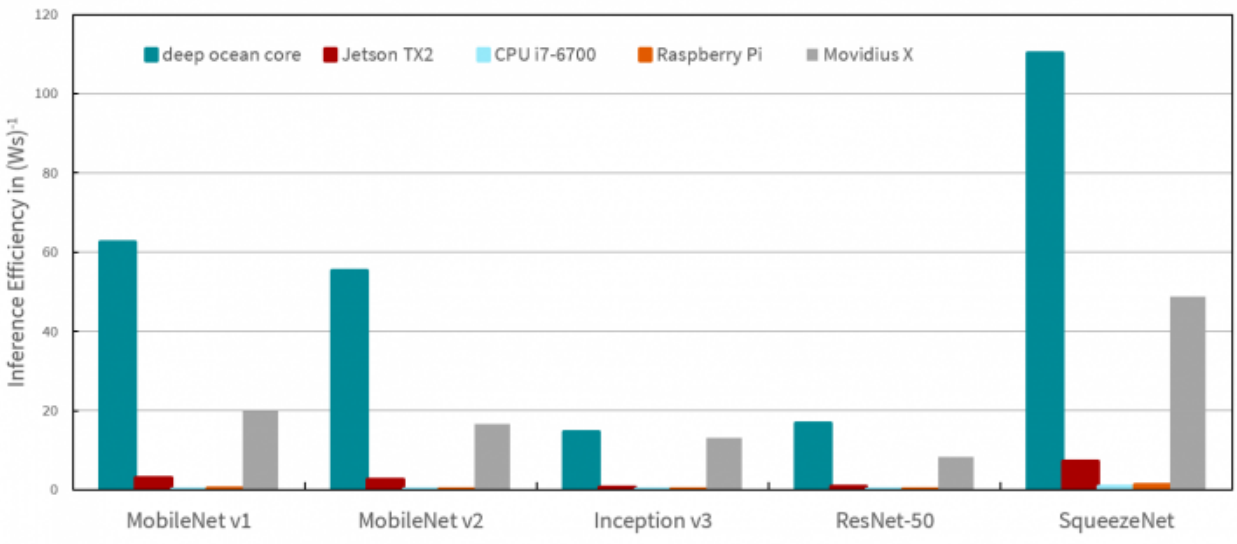

Figura 3 - Especialmente en las redes con economía de parámetros, como MobilNets o SqueezeNet, se aprecia claramente la superioridad de la arquitectura FPGA. Entre los sistemas comparados, es el que posee la mayor eficiencia energética. Esto convierte al deep ocean core en uno de los candidatos preferidos para la edge intelligence.

Cámara inferencial como solución integral

Para facilitar aún más el uso del acelerador de CNN basado en FPGA, IDS ofrece una solución integral en forma de cámara inferencial con el fin de hacer que la tecnología sea fácilmente accesible para todos. Los usuarios no necesitan conocimientos técnicos sobre aprendizaje profundo, procesamiento de imágenes ni programación de FPGA o de la cámara para entrenar y ejecutar una red neuronal, y pueden empezar inmediatamente con el procesamiento de imágenes basado en IA. Pueden utilizar herramientas sencillas que les permitirán aprender rápidamente a crear tareas de inferencia en minutos y ejecutarlas inmediatamente en una cámara. Además de la plataforma de cámara inteligente IDS NXT con el acelerador de CNN basado en FPGA "deep ocean core" , esta solución integral incluye un software de entrenamiento para redes neuronales fácil de usar. Todos los componentes están desarrollados por IDS y se han concebido para trabajar juntos de forma integrada. Esto simplifica los procesos de trabajo y hace que todo el sistema sea muy potente.

Sostenibilidad con edge intelligence

Cada una de las opciones para ejecutar las redes neuronales mencionadas en el artículo tiene sus ventajas e inconvenientes. Si los usuarios finales tienen que ocuparse ellos mismos de los componentes necesarios para utilizar la IA en tareas de visión artificial, prefieren recurrir a aceleradores de IA totalmente integrados, como el Intel Movidius. Las soluciones de chips listos para usar funcionan de forma eficiente, permiten fijar precios unitarios que sólo son posibles en grandes cantidades y pueden integrarse en los sistemas de forma rápida y relativamente sencilla gracias a la gran cantidad disponible de documentación sobre el abanico de funciones. Por desgracia, hay un pero. Su largo tiempo de desarrollo es un problema en el entorno de la IA, que ahora ha cobrado un enorme impulso y cambia a diario. Actualmente, para desarrollar una edge intelligence universal y flexible, los componentes del sistema deben cumplir otros requisitos. Una base FPGA es la combinación perfecta de flexibilidad, rendimiento, eficiencia energética y sostenibilidad. Al fin y al cabo, uno de los requisitos más importantes que debe cumplir un producto industrial es su "aptitud industrial", que se garantiza, entre otros factores, con una larga disponibilidad y un mantenimiento sencillo y duradero. La plataforma de cámaras de inferencia IDS NXT fácil de usar, combinada con un acelerador de CNN FPGA, representa actualmente una solución de edge intelligence integral y sostenible que permite al usuario final despreocuparse de los distintos componentes y de las actualizaciones de la IA.