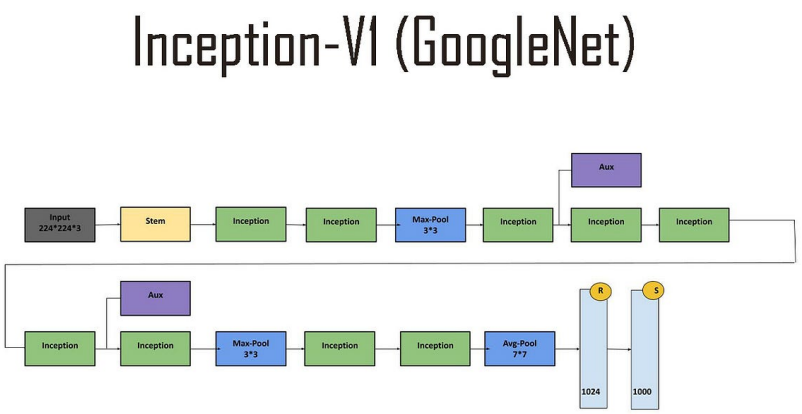
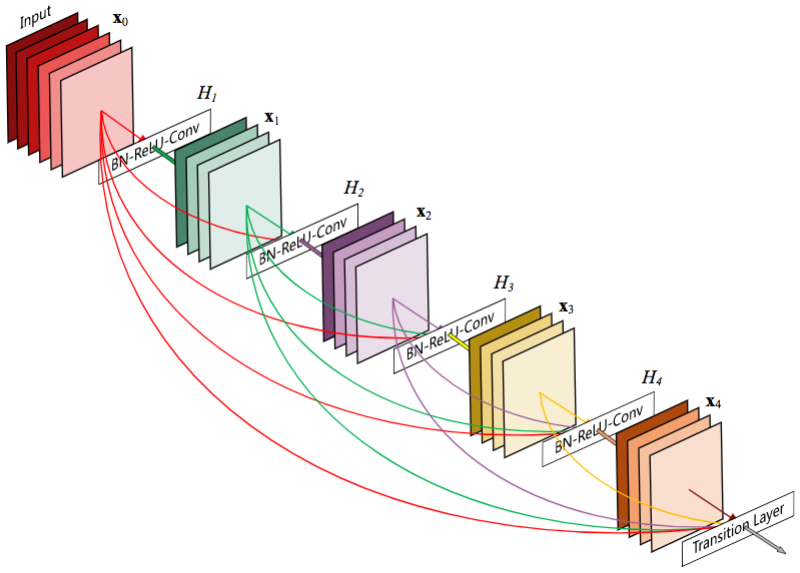
Task 2

1a. GoogleNet (Inception)
Uso: El Inception (GoogleNet) se utiliza principalmente en problemas de clasificación de imágenes y detección de objetos. Su característica más destacada es la introducción de los módulos "inception", que permiten que la red tome decisiones sobre el tamaño de filtrado más adecuado en una etapa determinada. Esta arquitectura redujo significativamente la cantidad de parámetros en la red y aumentó la eficiencia sin comprometer la precisión. (Brital, 2022)

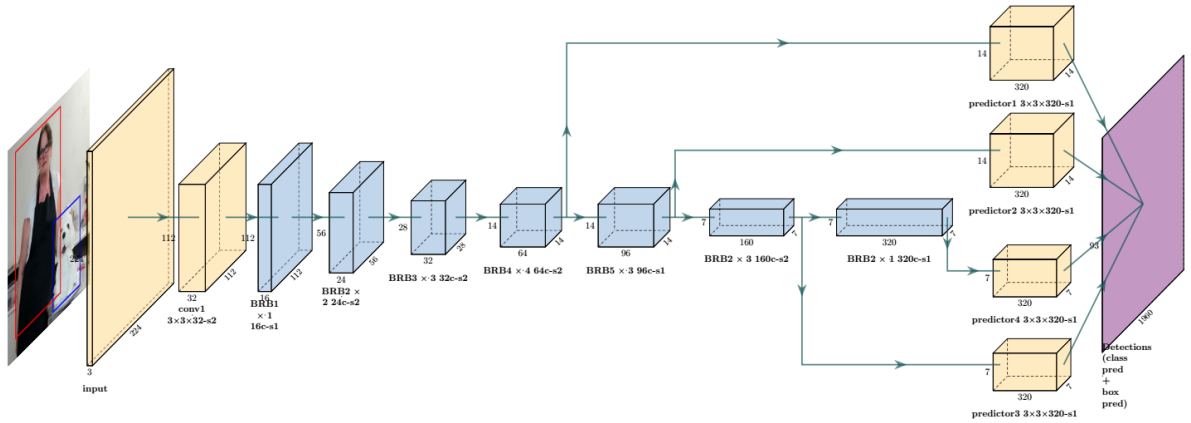


1b. DenseNet (Densely Connected Convolutional Networks)
Uso: DenseNet es especialmente útil para tareas de segmentación de imágenes, clasificación y otras tareas que se benefician de la preservación de características a lo largo de las capas. En lugar de utilizar conexiones tradicionales, cada capa recibe entradas de todas las capas anteriores, lo que mejora el flujo de información y gradientes a lo largo de la red.(Douillard, 2018)



1c. MobileNet

Uso: MobileNet es una arquitectura diseñada para ser ligera y eficiente, adecuada para dispositivos móviles y aplicaciones en el borde (edge computing). Se centra en la eficiencia de cálculo utilizando convoluciones separables por profundidad. Es útil para aplicaciones que requieren una detección en tiempo real en dispositivos con recursos limitados.(*MobileNet-Tiny*, n.d.)



1d. EfficientNet

Uso: EfficientNet es una arquitectura que escala de manera uniforme la profundidad, el ancho y la resolución del modelo para lograr un mejor rendimiento con menos parámetros. Es útil en tareas que requieren modelos de alto rendimiento pero que también se benefician de una arquitectura eficiente en cuanto a recursos. (*EfficientNet: Improving Accuracy and Efficiency Through AutoML and Model Scaling*, 2019)

2. ¿Cómo la arquitectura de transformers puede ser usada para image recognition?

Tradicionalmente, los transformers se han utilizado para procesar secuencias, especialmente en tareas de procesamiento del lenguaje natural. Sin embargo, recientemente se han adaptado para tratar las imágenes como secuencias de parches (patches). En este enfoque, una imagen se divide en una cuadrícula de parches, y cada parche se aplana y se procesa como una entrada en la secuencia.

Una arquitectura popular que utiliza esta idea es ViT (Vision Transformer). ViT toma estos parches, los procesa a través de la arquitectura transformer y luego utiliza la salida para la clasificación de imágenes. Gracias a la capacidad de los transformers para manejar dependencias a largo alcance, esta arquitectura ha demostrado ser competitiva con los enfoques tradicionales basados en CNN en ciertas tareas de reconocimiento de imágenes. (Boesch, 2023)

Referencias:

Brital, A. (2022, January 4). GoogLeNet CNN Architecture explained (Inception v1) : *Medium*. https://medium.com/@AnasBrital98/googlenet-cnn-architecture-explained-inception-v1-225ae02513fd

Douillard, A. (2018, May 6). *Densely connected convolutional networks*. Arthur Douillard. https://arthurdouillard.com/post/densenet/

MobileNet-Tiny. (n.d.). https://nitheshsinghsanjay.github.io/

EfficientNet: Improving Accuracy and Efficiency through AutoML and Model Scaling. (2019, May 29). https://blog.research.google/2019/05/efficientnet-improving-accuracy-and.html?m=1

Boesch, G. (2023). Vision Transformers (VIT) in Image Recognition – 2023 Guide. *viso.ai*. https://viso.ai/deep-learning/vision-transformer-vit/