Generación de imágenes con redes neuronales profundas a partir del enfoque en objetos.

Nicolás Antinori

Director
Guillermo Luis Grinblat

Objetivos

Se puede obtener de videos, información de cómo se está moviendo un objeto en la imagen: los píxeles que forman un objeto generalmente se mueven de la misma forma y de manera independiente al fondo. Utilizando esta información podemos segmentar el objeto del fondo y proveer al algoritmo de aprendizaje con esta información extra.

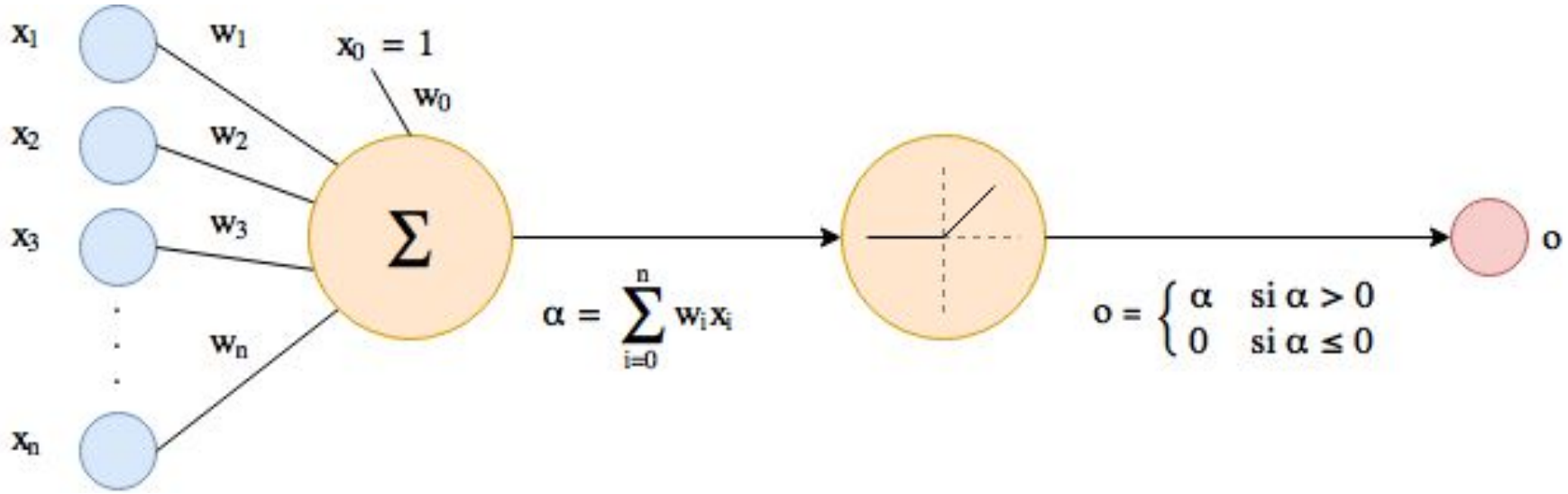
El objetivo de este trabajo es investigar el efecto que puede tener el uso de la información de segmentación sobre una red neuronal entrenada de manera no supervisada para evaluar si puede aprender características más fuertes y más definidas y en consecuencia, generar mejores imágenes.

Conceptos Generales

Dataset restringido

Dataset de alta variabilidad

Neurona Artificial



Red Neuronal

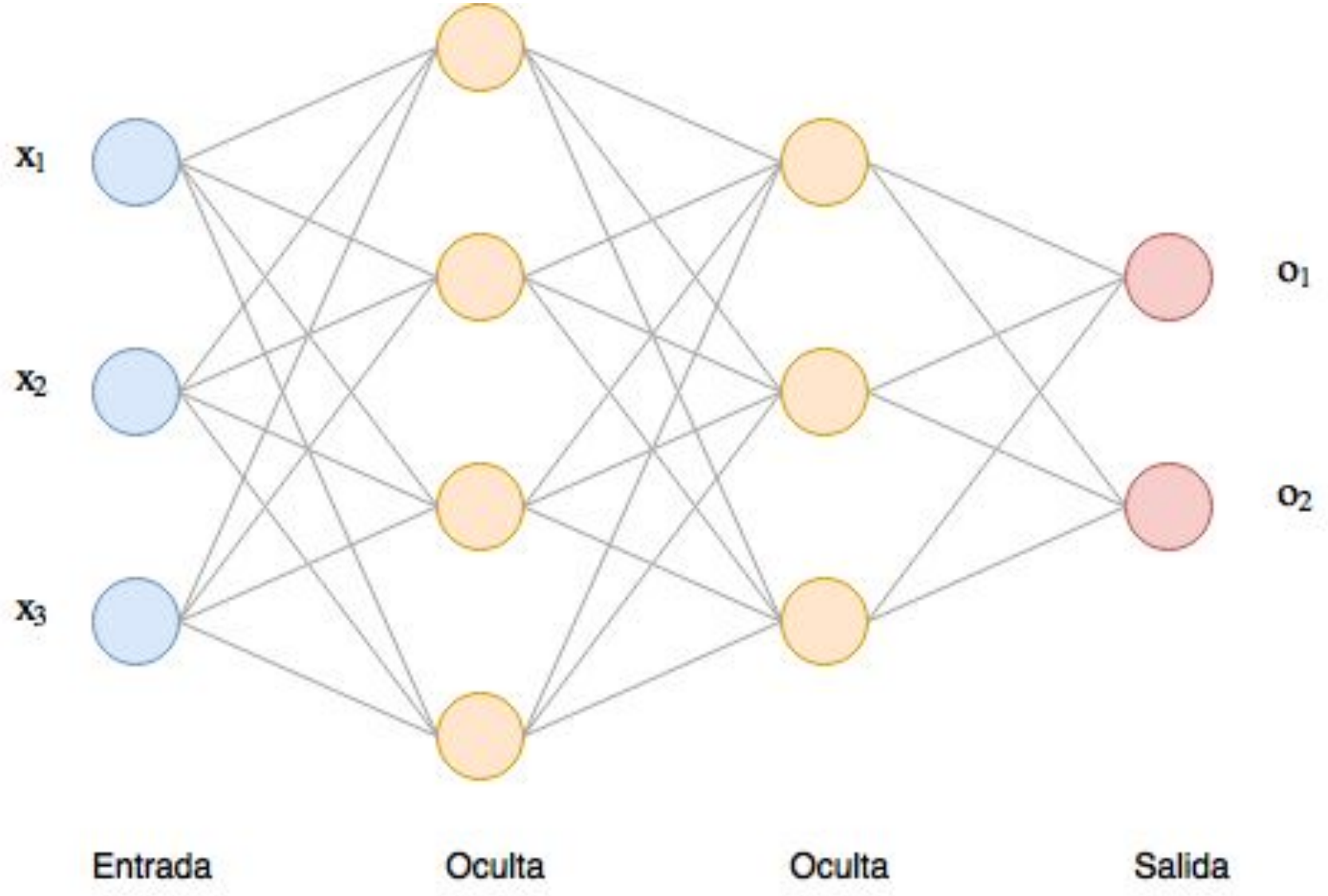
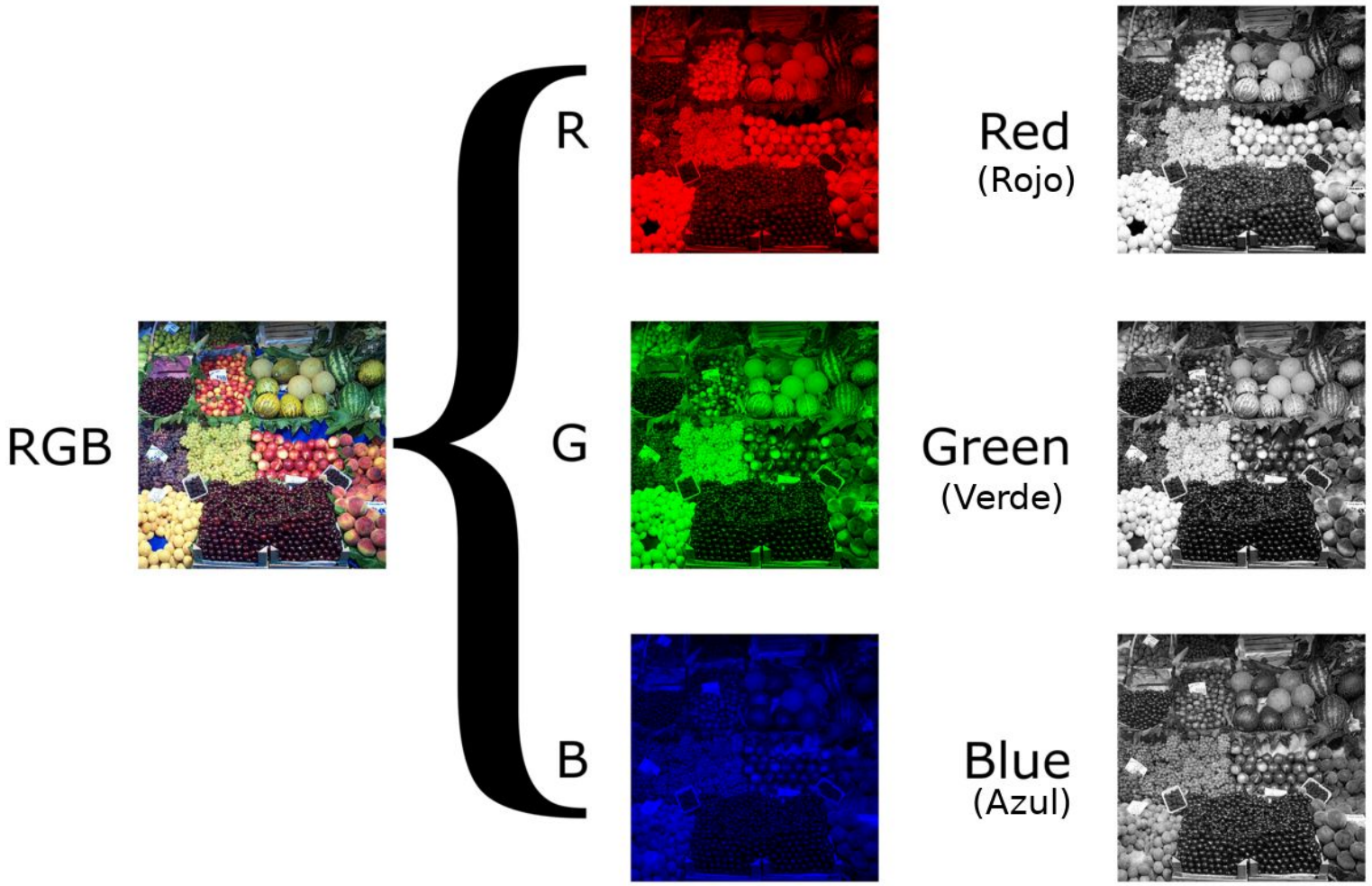
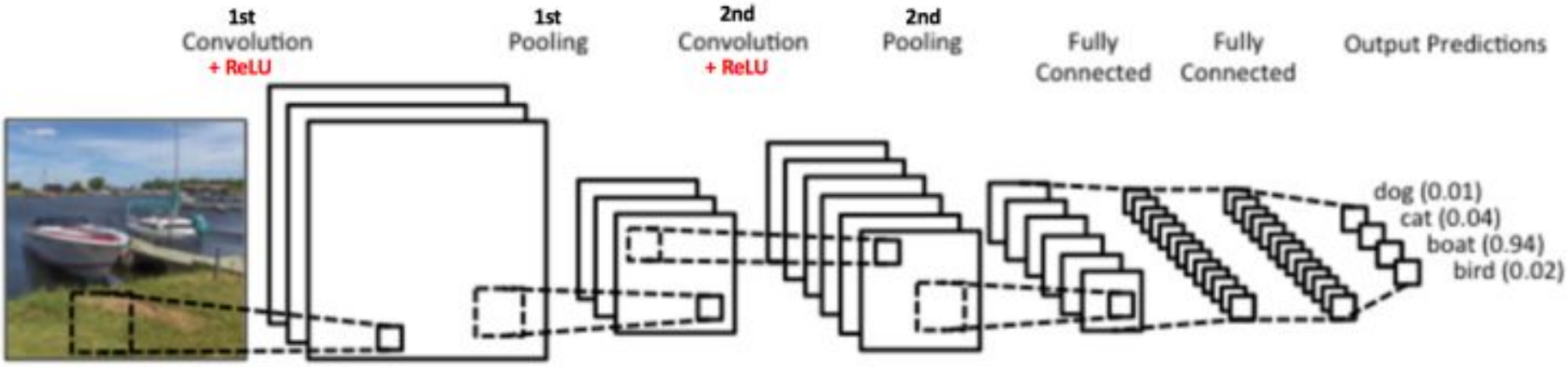


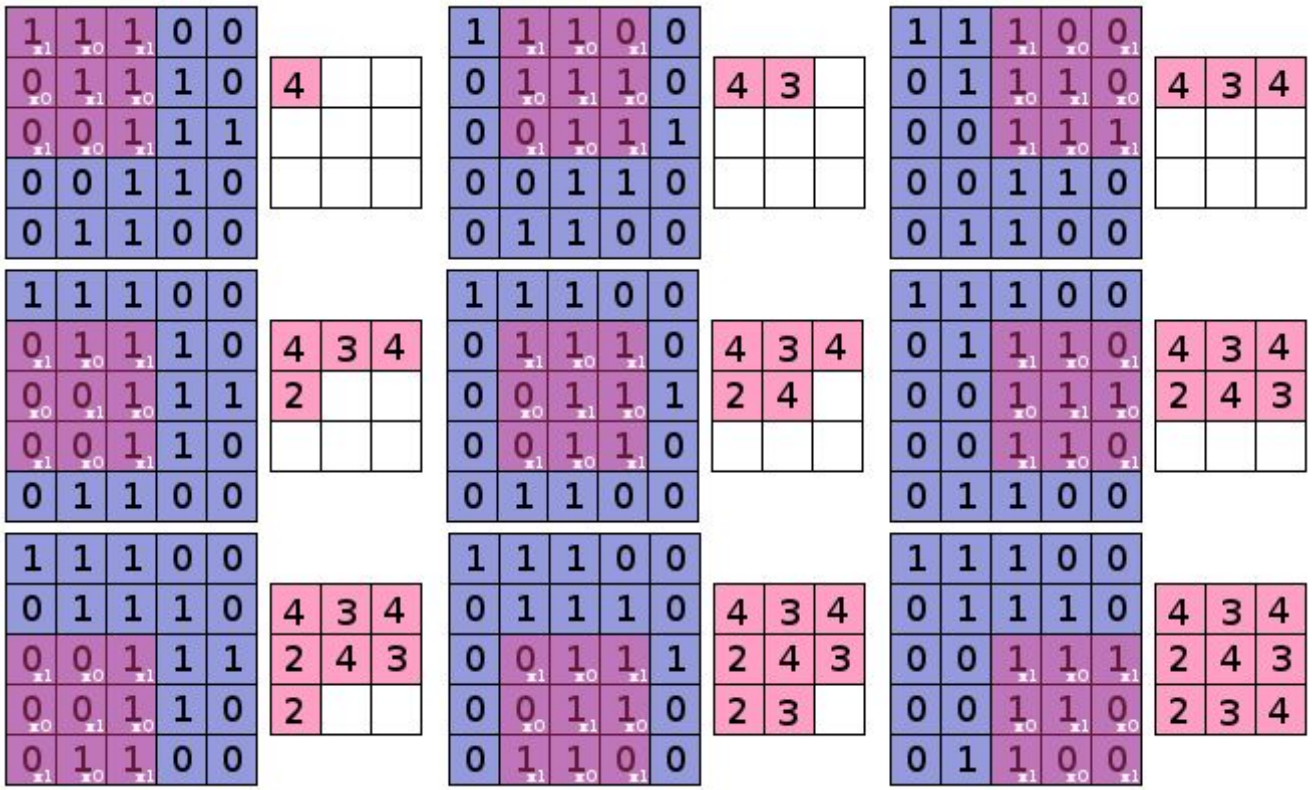
Imagen digital

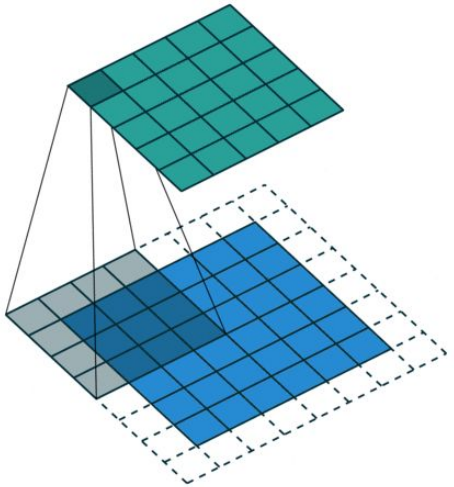


Red Convolucional

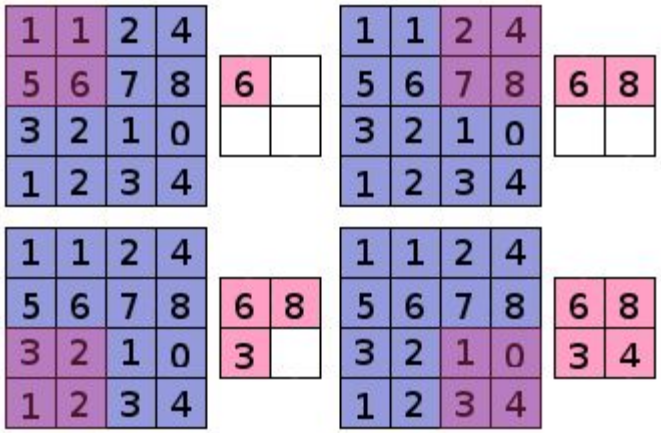


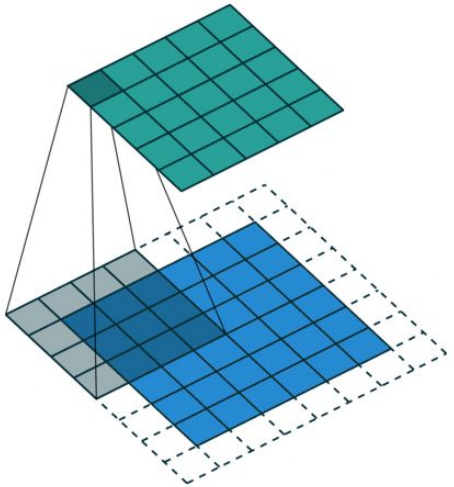
Convolución



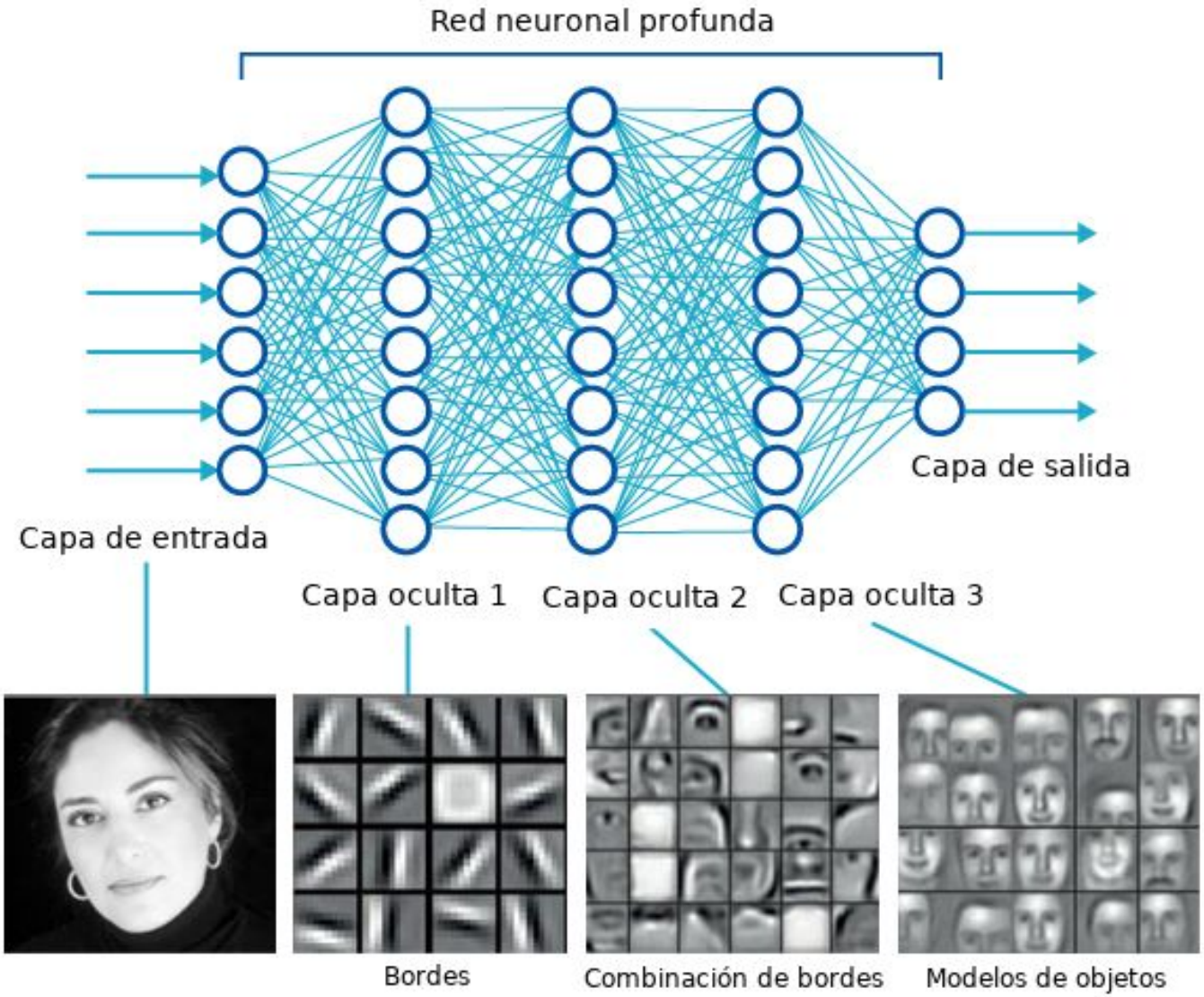


Pooling

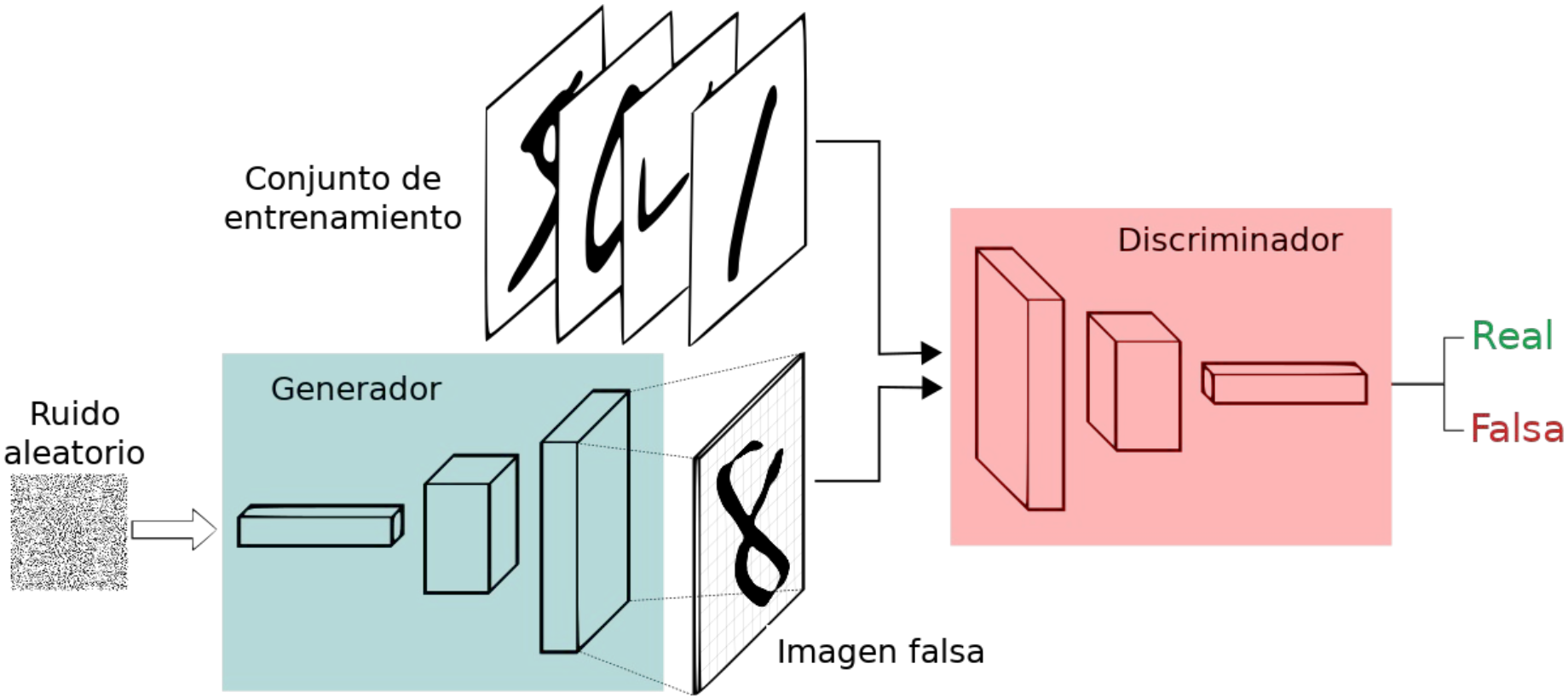

1 1 2 4
5 6 7 8
3 2 1 0
1 2 3 4
6
1 1 2 4
5 6 7 8
3 2 1 0
1 2 3 4
6 8
1 1 2 4
5 6 7 8
3 2 1 0
1 2 3 4
6 8
3
1 1 2 4
5 6 7 8
3 2 1 0
1 2 3 4
6 8
3 4



Red convolucional - Ejemplo



Redes Generativas Adversarias



Inception Score

El inception score es un método para medir la calidad generativa que tiene una GAN. El método muestra una correlación razonable con la calidad y la diversidad de las imágenes generadas como así también con la evaluación realizada por humanos.

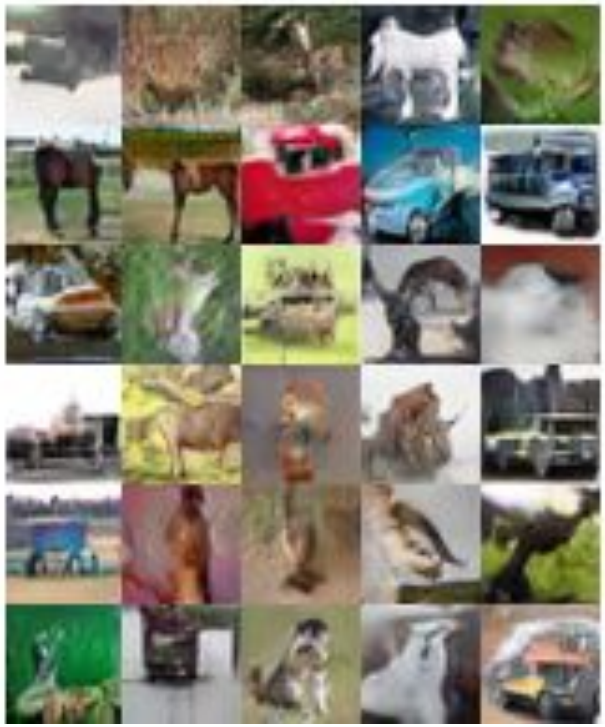
El inception score utiliza dos criterios para medir la calidad generativa:

- La calidad de las imágenes generadas.
- La diversidad de las muestras.

Inception Score - Ejemplos

IS 5.5

IS 7.5



IS 8.9

Datasets

Datasets

- De cada dataset sin procesar se obtuvieron como base dos datasets, uno con imágenes con los tres canales estándares y otro con los tres canales estándares y su máscara de segmentación correspondiente como un canal extra.
- Todas las imágenes como así también sus máscaras fueron redimensionadas a 64 píxeles de ancho por 64 píxeles de alto.
- Las máscaras fueron llevadas a la escala de los canales RGB (en el intervalo [0, 255]).
- Los datasets utilizados como base para generar los que se utilizaron son YFCC100m (Yahoo Flickr Creative Commons 100 million) y MS-COCO (Microsoft Common Objects in Context).

YFC100m

Yahoo Flickr Creative Commons 100 million (YFCC100m) es un dataset con un total de 100 millones de objetos multimedia. El contenido fue subido por usuarios de la red Flickr, publicados bajo la licencia Creative Commons.

- De todo el dataset sólo se utilizaron los videos sin tener en cuenta sus etiquetas.
- La segmentación fue realizada por Deepak Pathak et al. y publicada para su uso.
- El algoritmo se corrió sobre 205.000 videos. Se extrajeron entre 5 y 10 fotogramas por toma obteniendo así un dataset de 1.6 millones de imágenes.
- De cada fotograma se produjeron dos imágenes: una común de tres canales y otra de un solo canal, con valores de 0 a 100 representando la máscara del fotograma.

YFC100m

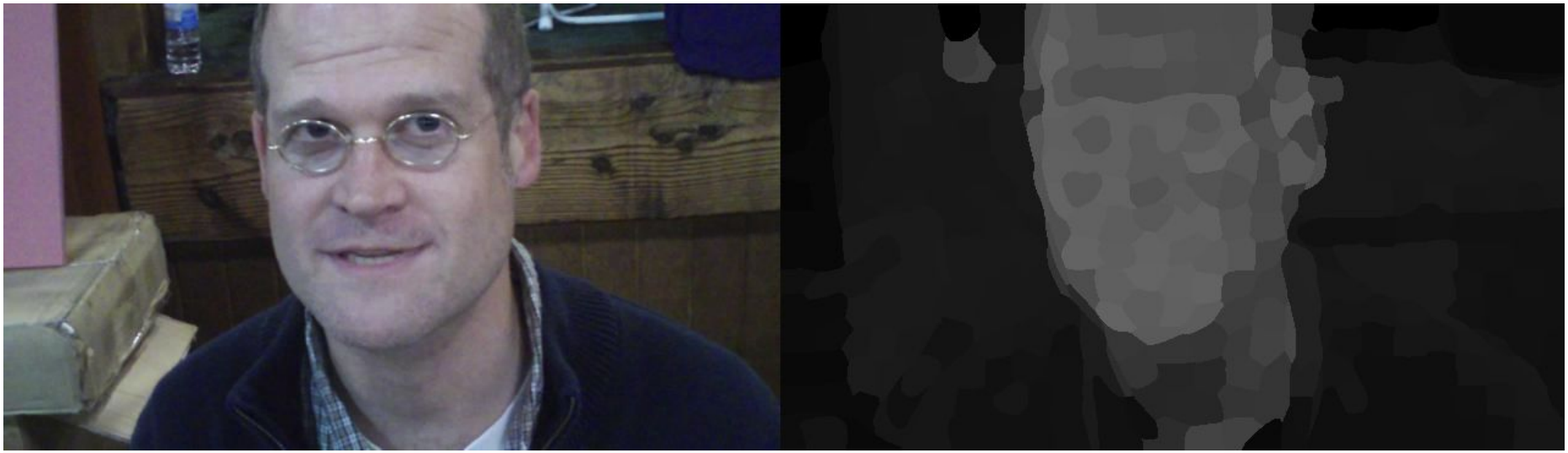
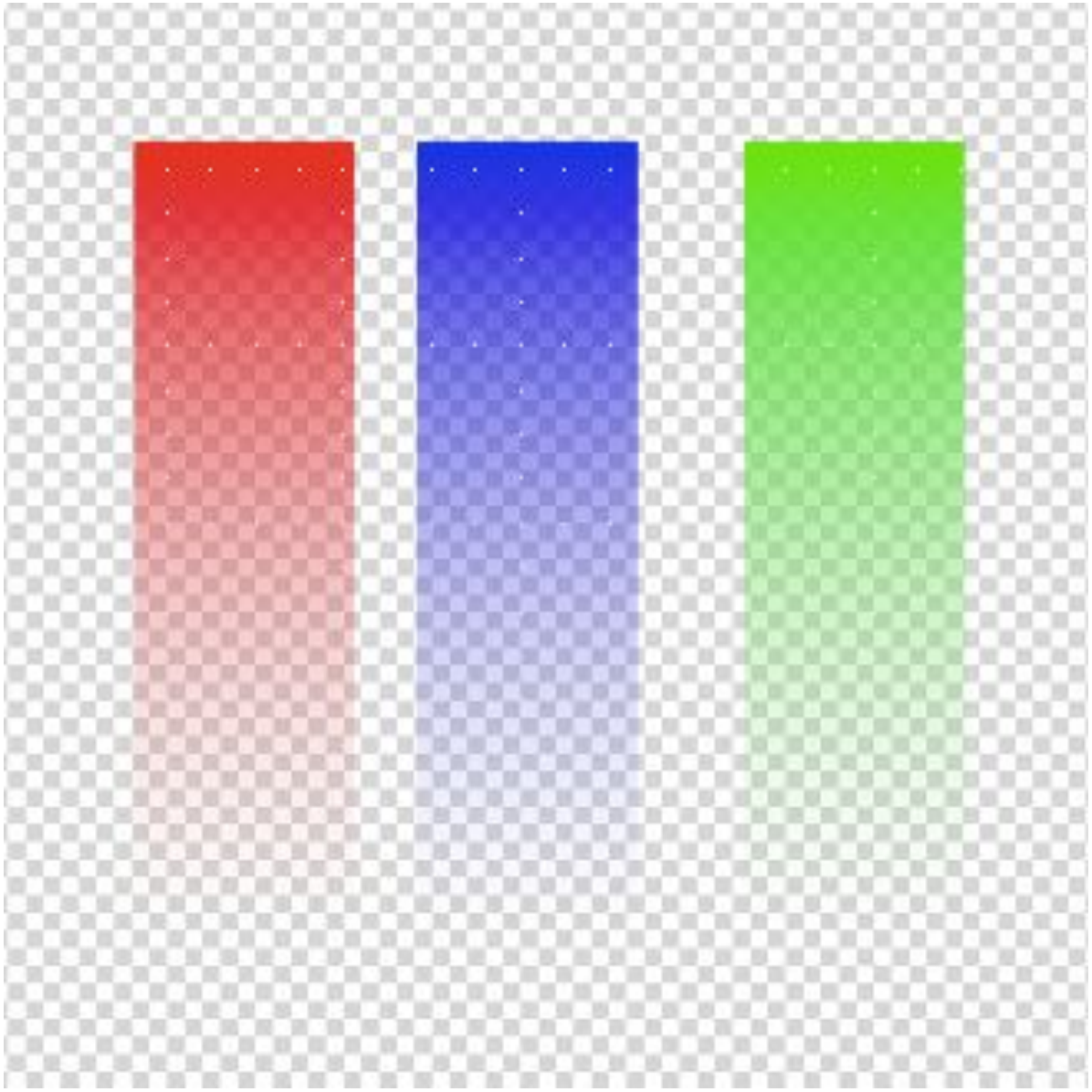


Imagen digital - Canal Alfa



YFC100m - Resultado en una imagen

YFC100m - Problemas

YFCC100m - Resultados

MS COCO

Microsoft Common Objects in Context (MS COCO) es un dataset con un total de 328.000 imágenes que contienen 91 clases de objetos categorizados. En total, el dataset cuenta con 2.5 millones de etiquetas.

- Se tienen más de 5.000 instancias de 82 de las categorías.
- La mayor parte de las imágenes son no-icónicas.
- Las imágenes fueron clasificadas, anotadas y segmentadas por humanos, lo cual deriva en segmentaciones de altísima calidad.

MS COCO - Imagen icónica vs no icónica

MS COCO - Proceso de etiquetado y clasificación

1

2

3

MS COCO - Preparación

Debido a que luego se redimensionan a 64x64 píxeles, no se toman en cuenta aquellas imágenes cuya segmentación no está formada por al menos un 15% de área total de píxeles.

MS COCO - Dataset resultante

Resultados

MS COCO - Resultados con cuatro canales

La red segmentó de forma perfecta el objeto enfocado, el cual suele ser muy distinto que el fondo en donde está insertado. Existen casos también en que el fondo no es tan distinguible del objeto, pero en la segmentación se puede ver una forma bien definida.

MS COCO - Resultados con cuatro canales

Es también visible que se logró separar satisfactoriamente algunas clases de entrenamiento más allá de que las redes hayan sido entrenadas sin etiquetas.

MS COCO - Resultados con tres canales

La separación en clases también es observable pero en menor medida, debido a que las muestras en general están menos definidas y en varios casos los objetos parecen fundirse con el fondo por no tener una silueta bien marcada.




YFCC100m Filtrado - Resultados con cuatro canales

Hay imágenes donde lo que está segmentado es claramente diferenciable del fondo mientras que en otras no parece distinguirse bien qué es fondo de lo que es un objeto o tienen una segmentación sin sentido.

YFCC100m Filtrado - Resultados con cuatro canales

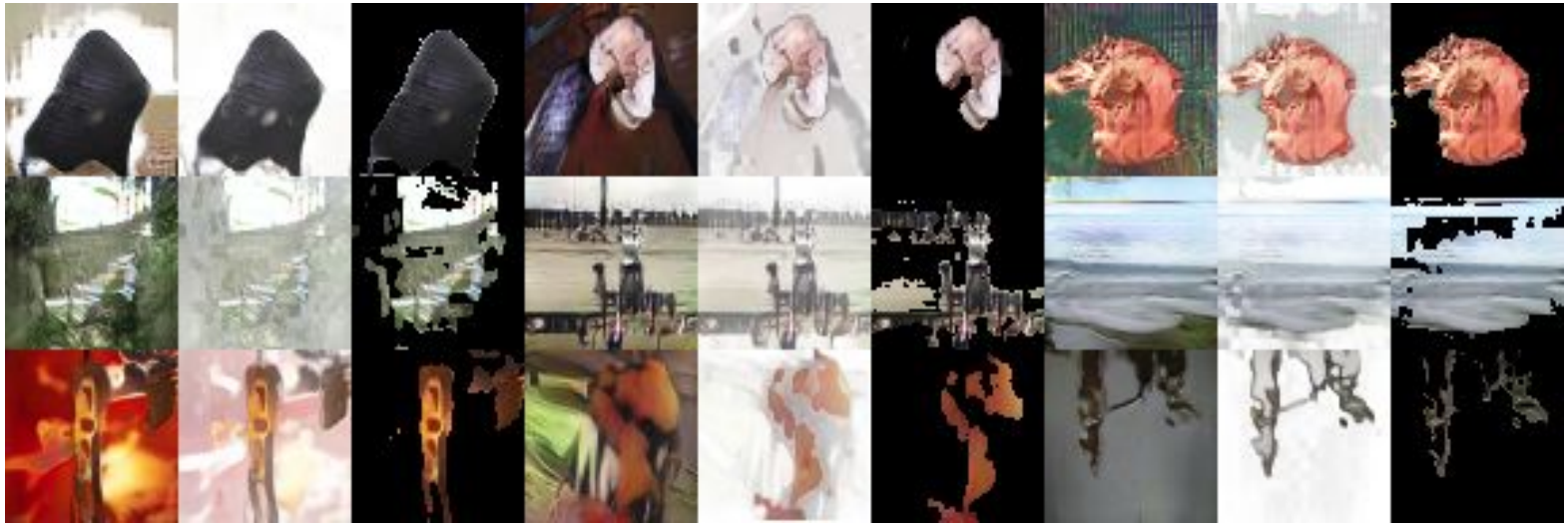
Otra característica a destacar es que en algunas imágenes los objetos enfocados no parecen tener continuidad sino ser un conjunto de enfoques inconexos.

YFCC100m Filtrado - Resultados con tres canales

En el caso del dataset con tres canales tampoco parece haber clases bien definidas o reconocibles. Es notable, sin embargo, que existen varias muestras que ilustran paisajes de una calidad aceptable.

YFCC100m Sin Filtrar - Resultados con cuatro canales

No parece haber formas reconocibles que distingan claramente alguna clase de objeto. En el dataset de cuatro canales persisten las segmentaciones perfectamente marcadas, las difusas, y los enfoques inconexos.



YFCC100m Sin Filtrar - Resultados con tres canales

En ambos resultados aparecen muestras que aparentan ser paisajes o sitios al aire libre, pero con una calidad bastante baja.

Comparativa transversal - Inception Scores

	Entrenamiento	3 canales	4 canales
COCO	13,51 ± 0,11	7,37 ± 0,07	7,50 ± 0,07
YFCC100m filtrado	22,76 ± 0,14	8,01 ± 0,08	7,80 ± 0,10
YFCC100m sin filtrar	11,75 ± 0,04	7,53 ± 0,07	7,06 ± 0,10

Comparativa transversal

COCO

YFCC100m

Conclusiones

Conclusiones

- Los resultados parecen sugerir que las segmentaciones hacen un aporte significativo al aprendizaje de conceptos de alto nivel.
- La calidad de las segmentaciones manifiestan ser de gran importancia.
- Se puede pensar que acotar las clases con las que se entrena tiene menos peso que tener una buena calidad en las segmentaciones.

Trabajos futuros

Trabajos futuros

- **Repetir las experiencias realizadas en esta tesina con los datasets etiquetados**: Teniendo en cuenta que el entrenamiento supervisado da mejores resultados que el no supervisado, es probable que agregando etiquetas a cada objeto enfocado, la red disponga de mejores herramientas para aprender los conceptos de alto nivel que los conforman.

- **Utilizar una mayor cantidad de mapas en las redes para mejorar su capacidad**: Al darle a la red más mapas hacemos que su capacidad de aprendizaje sea mayor. Se pretende que esto impacte de forma positiva a la hora de aprender nuevas características o mejorar las que se pueden aprender con un menor número de mapas.

- **Utilizar únicamente los objetos enfocados quitando toda la información referente al fondo en cada imagen**: Despojar la información innecesaria puede ayudar a la red a centrarse más en los detalles de los objetos enfocados, derivando en un mejor aprendizaje de sus características de alto nivel.

Gracias