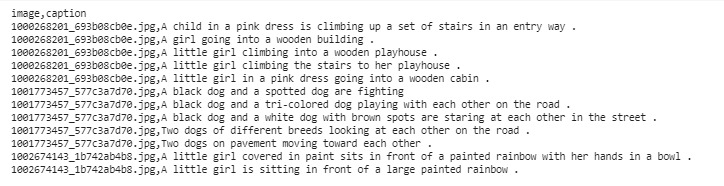
Boceto TFG

El objetivo final del trabajo es que el sistema sea capaz de ofrecer una descripción de una imagen. Para ello voy a utilizar un sistema de Deep learning, que consiste en el entrenamiento de un algoritmo que es capaz de reconocer las características de las 8000 imágenes que contiene el dataset y aprender su contenido para posteriormente ser capaz de describir una imagen nueva que se le proporcione al algoritmo y que no pertenezcan al conjunto de datos del entrenamiento.

Parto del conjunto de datos que nos proporciona Kaggle. Kaggle es una comunidad de inteligencia artificial y machine learning que proporciona recursos, modelos entrenados o conjuntos de datos entre otras cosas, a estudiantes, desarrolladores e investigadores que quieren trabajar en esta parte de la inteligenica artificial.

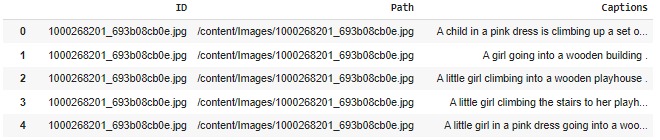
El conjunto de datos que he elegido consta de 8000 fotos, el nombre es flickr8k (<https://www.kaggle.com/datasets/adityajn105/flickr8k/data>), además, viene acompañado de un fichero que contiene las descripciones de esas fotos. El fichero se llama caption.txt, en el encontramos el nombre del archivo jpg que contiene la imagen, una descripción de la imagen y la ruta física donde se aloja el archivo jpg. Por cada imagen hay 5 descripciones en el fichero 

Para trabajar con este dataset montare una unidad en drive para descargar toda la información que vamos a usar.

(Explicar problemas de almacenamiento en bbdd)

Esta primera versión se corresponde con el tratamiento que hacemos de las 8000 fotos y de sus 40000 descripciones . Para ello utilizamos dos algoritmos de entrenamiento, por un lado tratamiento de imágenes y por otro el texto que describe las imágenes.

En primer lugar hay que tratar el conjunto de datos y construir un dataset que tenga la información que nos interesa, en este caso, construimos un dataset firmado por un id, una ruta y la descripción de la foto.



Primero tratamos el texto, hay que hacer un trabajo de limpieza de las descricpiones, comprobar si hay minúsculas, mayúsculas, dividir todo en palabras… y creamos una lista que contenga todas las palabras que hemos obtenido para luego poder usarlas en el proceso de crear la nueva descripción de la foto, además de usarlo en el proceso de aprendizaje

A partir de ahí, una vez tratado el texto, comenzamos con el tratamiento de las imágenes. Para el tratamiento de las imágenes, utilizamos un modelo pre-entrenado de imágenes porque los medios físicos que tengo para trabajr no son suficientes para desarrollar un model entrenado por mi porque es necesario disponer de una maquina muy potente, con GPU potente. Utilizamos el modelo pre-entrenado VGG16 que sirve para procesar imágenes. Para usar este modelo es necesario que todas las imágenes tengan unas determinadas características, para ello voy a crear unas funciones que hacen que todas las fotos del modelo cumplan estas características. Todas als imágenes tienen que estar 224x224 pixeles y cada pixel tiene que tener un valor entre 0 y 1 (normalización : crear un tensor únicamente de 0 y 1 para trabajar con ello en el aprendizaje)

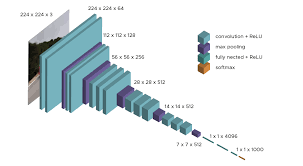
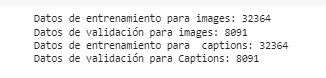


Grafico de la arquitectura del modelo VGG16 en deep learning

Una vez hecho el tratamiento de las imágenes se divide el conjunto de datos en dos partes, una parte para entrenar y otra para probar. En este caso he elegido un conjunto de entrenamiento del 80% de las fotos y para las pruebas el 20% restante 

El trabajo del entrenamiento de las imágenes lo realiza de forma automática este modelo (reconocer lo que contiene cada foto y su descripción asociada)

Ahora pasamos a la parte del texto, una vez entrenadas las imágenes pasamos al entrenamiento de las descripciones de cada una de esas imágenes. Lo que intento hacer en esta parte es que el sistema sea capaz de predecir una descripción de cualquier imagen que se le indique y que pertenezca al conjunto de entrenamiento. Para ello construyo un modelo de procesamiento del lenguaje natural (NLP) que consiste en un codificador y decodificador de palabras.

Este algoritmo aprenderá de todas las palabras que componen las descripciones de las fotos en un proceso de entrenamiento, realiza la codificación de las imágenes y la decodificación de las mismas para generar descripciones textuales.

El proceso de decodificación, o de generación de descripciones pertenece a lo que se denomina IA generativa. La IA generativa es un tipo de inteligenica artificial que puede crear ideas y contenidos nuevos, en este caso generara nuevas descripciones para una foto que pertenezca al conjunto de datos que hemos usado para el entrenamiento y debe ser funcional para fotos externas.

El proceso de entrenamiento del modelo de texto se genera con varias épocas, una época es el tiempo que el algoritmo tarda en procesar todas las fotos y cada una de sus descripciones. Por cada época añadida el proceso va mejorando el aprendizaje para afinar la descripción que hace de cada una de las fotos. He realizado dos entrenamientos, uno con 5 epocas, otro con 10 y el objetivo final serian 20 epocas que es el tiempo máximo de uso que me permite google colab. En cada entrenamiento veremos como mejora la descripción generada de fotos que pertenecen al conjunto de entrenamiento y de una foto que no pertenece al conjunto.

Este código es el backend de la aplicación, donde se realiza el trabajo. Mi idea sería construir un fornt que permitiera subir una imagen y que ofrezca la descripción de la misma.

Para ello tengo que hacer una aplicación o web en kotlin o html, que permita al usuario subir una foto, que esa foto se guarde en el drive y en el back, tratarla, normalizarla, generar la descripción y el audio y devolverlo al front para el usuario.

No se si a esta parte del proyecto me va a dar tiempo o voy a llegar.