STATISTICAL LEARNING 2 PROJECT

Marco Zuñiga
Universidad Galileo

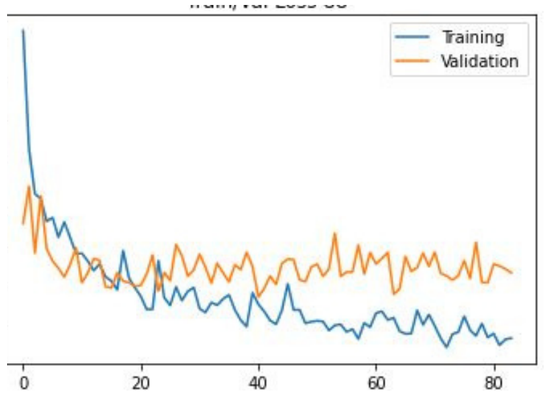
MLP

El problema seleccionado para tratar con un modelo de perceptrones multicapa fue uno de tipo clasificacion. Se busco un dataset relacionado a la salud para demostrar la capacidad y robustez de estos modelos con dataset estructurados. Para ello se selecciono el dataset de Fetal Health Classification

La única operación que se le hizo a los datos de entrenamiento fue la normalización. Y se separo un set de entrenamiento (0.8) y otro de pruebas (0.2).

Para el desbalance de los datos se aplico una solucion que nos proporciona keras, a partir de un análisis de desbalance pudimos determinar el peso de cada clase. Esto con intención de indicarle a la función **fit** a que observaciones debería de darle mas importancia.

La función de costo elegida para minimizar fue la de Cross Entropy para encontrar los pesos de nuestras capas para poder resolver este problema. Para este tipo de problema en especifico donde las clases las tenemos representadas como enteros. Utilizamos el tipo de funcion de costo Sparse Categorical Cross Entropy y para evaluar el rendimiento de nuestro modelo utilizamos el acurracy especial diseñado para este tipo de problemas que es el Sparse Categorical Acurracy.



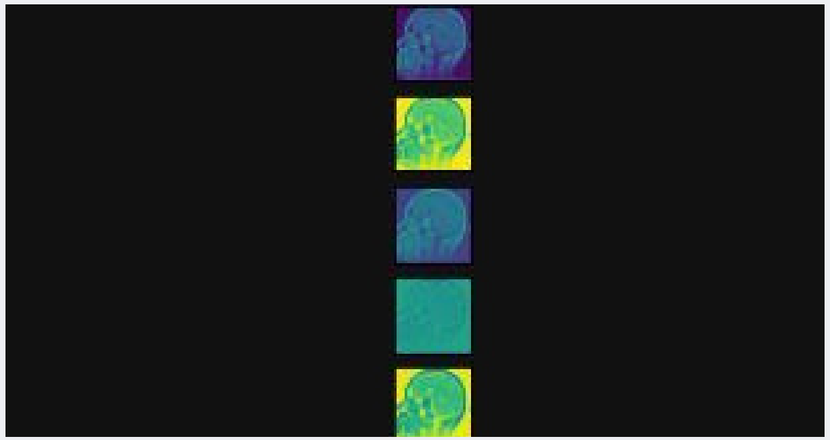
CNN

El problema elegido para aplicar ConvNets fue un problema también en el área de aplicaciones Médicas, en este caso se tomo un dataset de imágenes de MRI para detectar Tumores en el cerebro y poder clasificar si un examen de MRI contiene un tumor o no.
Una tarea difícil sino imposible para una persona sin entrenamiento especializado. Por lo que se puso a prueba una ConvNets para ver si podía aprender a clasificar con un exactitud arriba del 90%.

Los redes preentrenadas con las que se experimento para realizar el Transfer Learning fueron:

- ResNet50
- InceptionV3
- EfficientNetB0

El modelo seleccionado en base a la experimentacion **EfficientNetB0** que es un tipo de modelo recientemente publicado y que ha sido entrenado en millones de imágenes en tamaño es el mas grande de los tres elegidos. Con un total de parámetros por los 4 millones. Y según la documentación este modelo incluye una capa de normalización. Y los resultados para realizar la tarea de clasificación supero las expectativas por arriba del 99%. Por lo que se escogió este modelo para realizar las predicciones y clasificar los MRI. En el set de pruebas mantuvo el desempeño.



RNN

La aplicación de un red recurrente neuronal requiere un problema que necesite tener en cuenta el contexto de una secuencia. Para ello se eligio un problema de clasificacion de letras de canciones. Con la intencion de intentar determinar el genero de la cancion a partir de su letra.

Se puede observar después de los experimentos que estos modelos tienden a sobre-ajustarse. Por lo que es necesario aplicar la regularización a la parte Recurrente del Modelo.
Utilizando una libreria para el procesamiento del lenguaje se logro obtener mejores resultados.
Con estos cambios pudimos encontrar un modelo que en los datos de entrenamiento alcanzara un buena exactitud. Pero sufrimos de sobreajuste. Por lo que en los datos de entreno alcanzaba un 80\% de accuracy pero en el set de entrenamiento no alcanazba un valor mayor al 65\%.

CONCLUSIONES

- Como conclusión pudimos encontrar en los experimentos que para este tipo de problemas el optimizador Adam es muy eficaz para reducir la función de costo.
- Se pudo observar que los modelos MLP tienden a sobre-ajustarse a los datos de entrenamiento y la técnica de usar Cross Validation nos permitió detectar, analizar y tunear nuestro modelo para evitar que aprendiera la representación de los datos de entrenamiento y pudiera generalizar una mejor forma para otras observaciones no vistas aun.
- En un futuro swe podria expereimentar con una tecnica de oversampling para mitigar el problema de datasets desbalanceados.