Práctica 3: Redes neuronales convolucionales

Exequiel Alberto Castro Rivero

78308513E

Subgrupo 2

Contenido

[Introducción 3](#_Toc91248541)

[Ejercicio 1 3](#_Toc91248542)

[Batch size = 32 6](#_Toc91248543)

[Batch size = 64 7](#_Toc91248544)

[Batch size = 128 8](#_Toc91248545)

[Batch size = 256 9](#_Toc91248546)

[Comparación y análisis 10](#_Toc91248547)

[Ejercicio 2 11](#_Toc91248548)

[Mejora 1: Normalización de los datos de entrada 11](#_Toc91248549)

[Mejora 2: Early Stopping 13](#_Toc91248550)

[Mejora 3: Aumento de datos 14](#_Toc91248551)

[Mejora 4: Profundización de la red 16](#_Toc91248552)

[Mejora 5: Batch Normalization 18](#_Toc91248553)

[Mejora 6: Dropout 19](#_Toc91248554)

[Ejercicio 3 20](#_Toc91248555)

[Apartado A 20](#_Toc91248556)

[Apartado B 22](#_Toc91248557)

[Apartado C 23](#_Toc91248558)

# Introducción

El trabajo de esta práctica está relacionado con un tema de gran actualidad: las redes neuronales. En el caso que nos ocupa, vamos a realizar la implementación de **redes neuronales convolucionales**. Este tipo de redes han demostrado ser capaces de obtener resultados sorprendentes en determinadas áreas de la Visión por Computador.

El objetivo que se sigue con esta práctica es el de que nos familiaricemos con el diseño de redes neuronales convolucionales que aporta la librería Keras. De este modo, la experimentación va a ser un aspecto clave.

# Ejercicio 1

En este ejercicio se nos pide la construcción de un modelo base, *BaseNet*, para la resolución de un subproblema de la base de datos de *CIFAR100*. En concreto, esta base de datos representa un problema de clasificación con 100 clases. Sin embargo, para el caso que nos ocupa, nos quedaremos con los datos pertenecientes a sólo **25** de las 100 clases que forman la base de datos. Esta base de datos es cargada en memoria con la función de ayuda dada por los profesores *cargarImagenes*.

La descripción del modelo de *BaseNet* se muestra en la posterior tabla:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Layer No.** | **Layer Type** | **Kernel size (for conv layers)** | **Input|Output dimension** | **Input|Output channels (for conv layers)** |
| 1 | Conv2D | 5 | 32|28 | 3|6 |
| 2 | Relu | - | 28|28 | - |
| 3 | MaxPooling2D | 2 | 28|14 | - |
| 4 | Conv2D | 5 | 14|10 | 6|16 |
| 5 | Relu | - | 10|10 | - |
| 6 | MaxPooling2D | 2 | 10|5 | - |
| 7 | Linear | - | 400|50 | - |
| 8 | Relu | - | 50|50 | - |
| 9 | Linear | - | 50|25 | - |

Tabla : Descripción de BaseNet

La función que define dicha red convolucional es *get\_basenet*. Dicha función recibe los siguientes parámetros:

* *input\_shape*: El tamaño de las imágenes de entrada a la red (*[32, 32, 3]* en el caso de las imágenes de nuestro problema).
* *num\_classes*: El número de clases que se desea clasificar por la red neuronal, *25* en el caso de nuestro problema.

La definición de la red se ha hecho de forma secuencial, ya que esta forma es la que me parece más intuitiva y directa. Esta forma de definir redes tiene ciertas limitaciones, pero debido a que en esta práctica no se van a diseñar modelos muy complejos, me parece la más apropiada. A continuación se explica con mayor detalle cómo se ha realizado la definición del modelo con Keras:

* En primer lugar se especifica que el modelo se va a realizar de forma secuencial con *model = Sequential()*.
* Se añade una capa convolucional con activación *Relu* para implementar las capas número 1 y 2 de la tabla. Esto se realiza de la siguiente manera: *model.add(Conv2D(6, kernel\_size=(5, 5), activation='relu', input\_shape=input\_shape))*. Esta capa de convolución tiene 6 filtros debido a que así está especificado en los canales de salida de la tabla. Del mismo modo se obtienen los tamaños del kernel (*kernel\_size*). El parámetro *input\_shape* es necesario debido a que esta es la capa inicial del modelo y se necesitan saber las dimensiones de la entrada. El parámetro *activation* se usa para que la activación tras esta capa sea implementada con la función *Relu*. Todos los demás parámetros son los establecidos por defecto por Keras, bien porque los parámetros por defectos son los correctos (es el caso del parámetro *padding*, el cual tiene que estar establecido a *valid* para que se produzca una reducción en anchura y en altura de los bloques convolucionales) o bien porque no tienen un gran impacto en el rendimiento (se han realizado pruebas con distintos parámetros y se han obtenido resultados similares).
* Se añade una capa *MaxPooling* para reducir las dimensiones del bloque convolucional (el número de canales no se reduce). Esto se realiza de la siguiente manera: *model.add(MaxPooling2D(pool\_size=(2, 2)))*. Para que la reducción de dimensiones se haga a la mitad (tal y como se especifica en la tabla del modelo) el *pool\_size* debe tener valor 2. Los demás parámetros son los establecidos por defecto por Keras, ya que estos son los parámetros correctos (por ejemplo, *padding* establecido a *valid* para que se produzca la reducción a la mitad).
* Se añade otra capa convolucional del mismo modo que anteriormente: *model.add(Conv2D(16, kernel\_size=(5, 5), activation='relu'))*.
* Se añade otra capa *MaxPooling* del mismo modo que anteriormente: *model.add(MaxPooling2D(pool\_size=(2, 2)))*.
* Se añade una capa *Flatten* de modo que todos los píxeles del último bloque convolucional se disponen en un único array 1D. Una forma intuitiva de ver esta capa es que ya se ha decidido que la extracción de características de las imágenes ha llegado a su fin y, por lo tanto, todos los píxeles del último bloque convolucional son características que van a ser la entrada del clasificador de la red.
* Se añade una capa Densa para implementar las capas número 7 y 8 de la tabla. Esto se realiza de la siguiente manera: *model.add(Dense(50, activation='relu'))*. El número de neuronas se establece a 50 ya que así está especificado en la tabla. Lo mismo pasa con la función de activación. Los demás parámetros son los establecidos por defecto por Keras porque no tienen un gran impacto en el rendimiento (se han realizado pruebas con distintos parámetros y se han obtenido resultados similares).
* Se añade la capa totalmente conectada final, la cual se corresponde con la clasificación propiamente dicha. Esto se realiza de la siguiente forma: *model.add(Dense(num\_classes, activation='softmax'))*. El número de neuronas de esta capa totalmente conectada se debe de corresponder con el número de clases del problema que se está tratando. La función de activación usada es *softmax* debido a que el problema que se está resolviendo es un problema de clasificación multiclase: cada neurona “se encarga” de clasificar una clase y la salida de cada neurona es la probabilidad de que la imagen de entrada pertenezca a dicha clase.

Una aproximación intuitiva de lo que hace esta red es la siguiente: mediante el bloque formado por *Conv2D, Relu y MaxPooling* lo que se está realizando son transformaciones de la información de los bloques convolucionales anteriores (*Conv2D*) para luego realizar una selección de las características más relevantes (aquellas con mayor respuesta) con *Relu* y *MaxPooling*. Esta selección es importante, ya que si sólo se realizaran transformaciones de la información inicial el aprendizaje sería imposible. En la selección de aquellas características más relevantes es donde está la potencia de este tipo de aprendizaje (esto es similar a la forma de clasificar imágenes de los humanos: en una fotografía no todo es relevante a la hora de clasificarla). Una vez ya se ha completado la extracción de características (transformación + selección) dichas características son pasadas como entrada a un modelo de aprendizaje automático supervisado (una red neuronal en el caso de este modelo) con el fin de obtener un modelo de clasificación.

Como apunte adicional, si se desea ver con mayor profundidad el modelo definido o comprobar la correctitud del mismo se puede visualizar el resumen ofrecido por Keras (con la función *summary*) que se encuentra disponible en la ejecución del código.

Una vez terminada la etapa de definición del modelo se procede a la declaración del optimizador. En mi caso el optimizador usado será Adam. La justificación de la elección de este optimizador se encuentra en el paper [Adam: A Method for Stochastic Optimization](https://arxiv.org/pdf/1412.6980.pdf), paper donde se concluye que Adam es un buen optimizador para problemas con datasets grandes (el cual es nuestro caso). Además, en el mismo paper se menciona que los parámetros usados por este optimizador necesitan de poco ajuste (los parámetros por defecto obtienen un buen rendimiento en la mayoría de los casos). En mi caso, probé con distintas configuraciones de parámetros de este optimizador y en todas las pruebas se obtuvieron resultados similares o peores. Es por esta razón por la que me decanté usar los parámetros por defecto del optimizador.

Seguidamente se procede con la compilación del modelo. La función de pérdida usada es la entropía cruzada categórica: como ya se ha visto en la asignatura de aprendizaje automático, esta función de pérdida es la más popular a la hora de enfrentarse a problemas de clasificación multiclase. La métrica usada para evaluar el rendimiento es la *accuracy*, métrica popular en los problemas de clasificación. Esta métrica es definida como el porcentaje de imágenes bien clasificadas por el modelo. Como métrica se podría haber usado la misma función de pérdida establecida anteriormente, pero la entropía cruzada es muy poco intuitiva; por este motivo se usa la *accuracy* al ser bastante más intuitiva.

Tanto la definición del optimizador como la compilación del modelo se realizan con la función *model\_compile*, la cual recibe el modelo como parámetro. Dicha compilación se realiza con la función *compile* de Keras. Para la llamada sólo se especifican los parámetros *loss*, *optimizer* y *metrics*. Los demás parámetros son los establecidos por Keras ya que estos son los correctos (por ejemplo, al no tener desbalanceo de clases no hace falta que se use el parámetro *loss\_weights*).

En cuanto a la etapa de entrenamiento, el entrenamiento del modelo se realiza con la función *train\_model*. Esta función recibe por parámetros las imágenes de entrenamiento, las etiquetas de las mismas, el tamaño del batch a usar y el número de épocas. Para realizar el entrenamiento propiamente dicho se usa la función *fit* de Keras, con los parámetros anteriormente comentados. Los demás parámetros son los que establece por defecto Keras ya que, como en ocasiones anteriores, son los correctos.

En cuanto al número de épocas usado para los entrenamientos, se establece a 100. Esta decisión ha sido tomada en cuenta el compromiso precisión-coste computacional de este parámetro. Para un mayor número de épocas en general se consigue un mejor rendimiento (salvo en los casos en los que hay sobreajuste) pero el coste computacional es mayor. A menor número de épocas se consigue un peor rendimiento, pero se gana en coste computacional. Para los propósitos de esta práctica ese valor para las épocas me parece adecuado ya que permite observar la evolución de la función de pérdida y la métrica correctamente sin “gastar” demasiado tiempo.

El tamaño del batch usado para los entrenamientos es un parámetro difícil de fijar. El valor óptimo de este parámetro depende del problema que se está intentando resolver. En el paper [*Systematic evaluation of CNN advances on the Imagenet*](https://arxiv.org/pdf/1606.02228.pdf) los autores realizan un estudio experimental y concluyen que los tamaños de batch 64 y 128 son buenos valores en la mayoría de los casos. Para la elección de este parámetro hay que tener en cuenta el compromiso convergencia-coste computacional: a menor número de batch, la convergencia de los optimizadores es alcanzada en menor número de épocas, pero el coste computacional es mayor; sin embargo, a mayor número de batch la convergencia es alcanzada en un mayor número de épocas, pero el coste computacional es menor. Como se puede observar, hay cierta dependencia entre el tamaño del batch y el número de épocas de entrenamiento. Debido a todas las razones anteriormente comentadas, se van a probar distintos tamaños de batch y se va a analizar el comportamiento de las gráficas. Los resultados serán presentados posteriormente.

Finalmente, para el cálculo de la función de pérdida y la *accuracy* en test se usa la función *get\_loss\_accuracy*. Esta función recibe los datos de test y las etiquetas de dichos datos y haciendo uso de la función *evaluate*, se calcula y devuelve los valores de la función de pérdida y la *accuracy* en test.

Una vez explicadas todas las fases de construcción del modelo, se procede a realizar las pruebas anteriormente descritas. Se van a probar los siguientes tamaños de batch: 32, 64, 128 y 256. Cabe mencionar que, como se van a realizar distintos entrenamientos sobre el mismo modelo, los pesos aleatorios inicializados se guardan (con la función *get\_weights*) para poder restablecerlos en posteriores entrenamientos (con la función *set\_weights*).

## Batch size = 32

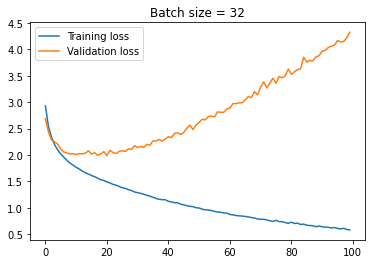


Ilustración : Evolución de la función de pérdida para tamaño de batch 32

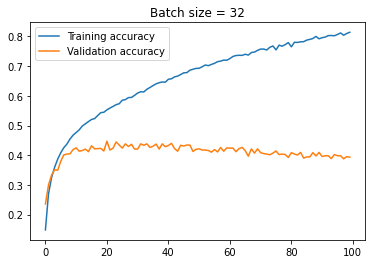


Ilustración : Evolución de la accuracy para tamaño de batch 32

Como se puede ver en las ilustraciones anteriores (*Ilustración 1* e *Ilustración 2*) la convergencia de la red se produce de manera bastante rápida (entorno a la época 20 ya se empieza a producir sobreajuste: la función de pérdida en validación empieza a crecer mientras que en los datos de entrenamiento sigue disminuyendo). A continuación, se presentan los resultados obtenidos en test y el tiempo que ha tardado la red en entrenar:

|  |  |  |
| --- | --- | --- |
| **Test loss** | **Accuracy test** | **Training time** |
| 4.16 | 38,88% | ~3 min 20s |

Tabla : Resultados en test y tiempo de entrenamiento para tamaño de batch 32

## Batch size = 64

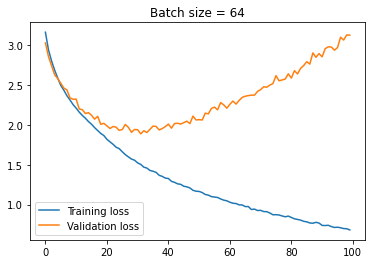


Ilustración : Evolución de la función de pérdida para tamaño de batch 64

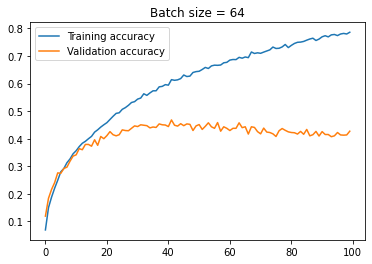


Ilustración : Evolución de la accuracy para tamaño de batch 64

Como se puede ver en las ilustraciones anteriores (*Ilustración 3* e *Ilustración 3*) la convergencia de la red se produce de manera rápida (entorno a la época 35 ya se empieza a producir sobreajuste: la función de pérdida en validación empieza a crecer mientras que en los datos de entrenamiento sigue disminuyendo). A continuación, se presentan los resultados obtenidos en test y el tiempo que ha tardado la red en entrenar:

|  |  |  |
| --- | --- | --- |
| **Test loss** | **Accuracy test** | **Training time** |
| 3.10 | 41,16% | ~2 min 10s |

Tabla : Resultados en test y tiempo de entrenamiento para tamaño de batch 64

## Batch size = 128

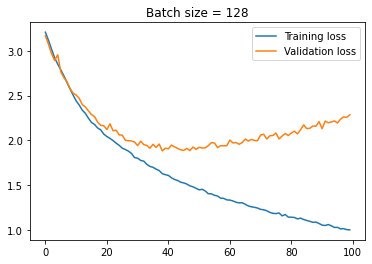


Ilustración : Evolución de la función de pérdida para tamaño de batch 128

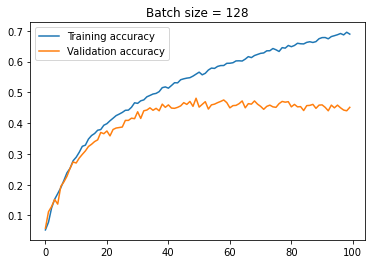


Ilustración : Evolución de la accuracy para tamaño de batch 128

Como se puede ver en las ilustraciones anteriores (*Ilustración 5* e *Ilustración 6*) la convergencia de la red se produce de manera lenta (entorno a la época 50 ya se empieza a producir sobreajuste: la función de pérdida en validación empieza a crecer mientras que en los datos de entrenamiento sigue disminuyendo). A continuación, se presentan los resultados obtenidos en test y el tiempo que ha tardado la red en entrenar:

|  |  |  |
| --- | --- | --- |
| **Test loss** | **Accuracy test** | **Training time** |
| 2.26 | 44,96% | ~1 min 30s |

Tabla : Resultados en test y tiempo de entrenamiento para tamaño de batch 128

## Batch size = 256

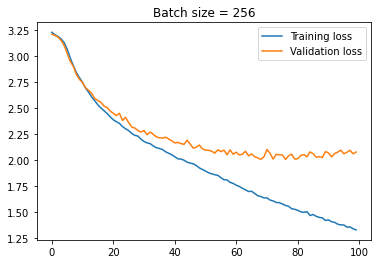


Ilustración : Evolución de la función de pérdida para tamaño de batch 256

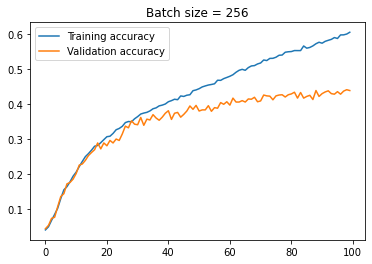


Ilustración : Evolución de la accuracy para tamaño de batch 256

Como se puede ver en las ilustraciones anteriores (*Ilustración 7* e *Ilustración 8*) la convergencia de la red no se produce de la misma manera que para los otros tamaños de batch (quizás si se entrenase unas épocas más se podría ver el mismo efecto). A continuación, se presentan los resultados obtenidos en test y el tiempo que ha tardado la red en entrenar:

|  |  |  |
| --- | --- | --- |
| **Test loss** | **Accuracy test** | **Training time** |
| 2.03 | 44,08% | ~50s |

Tabla : Resultados en test y tiempo de entrenamiento para tamaño de batch 256

## Comparación y análisis

Tal y como se puede observar en las ilustraciones del subapartado anterior, la forma de todas las gráficas es la misma: la diferencia entre ellas es un desplazamiento horizontal a la derecha (más notorio a medida que se aumenta el tamaño del batch). Esto pone de manifiesto que un tamaño de batch pequeño acelera la convergencia de la red. Esto casa con lo dicho anteriormente y con la teoría. También se puede observar el compromiso que hay entre tamaño de batch y coste computacional: a menor tamaño de batch mayor es el coste computacional. Así mismo también puede verse el compromiso entre el tamaño de batch y el número de épocas: a menor tamaño de batch el número de épocas necesarias para un buen entrenamiento también son menores.

A continuación, se presentan los resultados de test y el tiempo de entrenamiento obtenidos para todos los tamaños de batch:

|  |  |  |  |
| --- | --- | --- | --- |
| **Tamaño de batch** | **Test loss** | **Accuracy test** | **Training time** |
| **32** | 4.16 | 38,88% | ~3 min 20s |
| **64** | 3.10 | 41,16% | ~2 min 10s |
| **128** | 2.26 | 44,96% | ~1 min 30s |
| **256** | 2.03 | 44,08% | ~50s |

Tabla : Resultados obtenidos para los diferentes tamaños de batch

Como se puede observar, los valores de *accuracy* obtenidos están alrededor del 42-43% para el modelo *BaseNet*. Esto nos servirá como *baseline* para próximas mejoras. Así mismo, para dichas mejoras se tomarán los siguientes valores para el entrenamiento: número de épocas = 100 y tamaño de batch = 128 ya que con estos valores se consigue un equilibrio en el compromiso rendimiento – coste computacional.

# Ejercicio 2

En este ejercicio se nos plantea realizar mejoras sobre el modelo *BaseNet* para conseguir un mayor rendimiento en la base de datos *CIFAR-100*. A continuación se muestran las mejoras implementadas.

## Mejora 1: Normalización de los datos de entrada

La primera mejora planteada de las listadas en el documento de la práctica es la normalización de los datos de entrada. Es sabido que la normalización de los datos de entrada ayuda al entrenamiento ya que se consigue una convergencia más rápida. Al normalizar los datos lo que se está haciendo es que todas las imágenes de entrada se muevan en el mismo rango de valores para así poder evitar sesgos hacia determinadas imágenes. Si no se normalizan los datos de entradas, la actualización de los pesos usando el algoritmo de *backpropagation* puede ser bastante dispar dependiendo de la escala de las imágenes.

Para la realización de esta mejora se han implementado varias funciones. La primera de ellas ha sido la función *split\_train\_val*, que como su propio nombre indica, recibe un conjunto de datos y realiza la separación en datos de entrenamiento y datos de validación (un 10% de los datos de entrenamiento, tal y como se especifica en el enunciado de la práctica). Esta separación se realiza conservando el porcentaje de clases en ambos conjuntos de datos para que no haya clases infrarrepresentadas.

La siguiente función implementada fue *normalize\_images*. Esta función recibe los datos de entrenamiento y los de test, ambos con sus etiquetas, y el tamaño de batch. El funcionamiento de esta función sigue los siguientes pasos:

* Separación de los datos de entrenamiento iniciales en dos conjuntos: un nuevo conjunto de entrenamiento y un conjunto de validación (se usa la función anteriormente descrita, *split\_train\_val*).
* Se crea el objeto de la clase *ImageDataGenerator* de Keras con el que se va a realizar la normalización de los datos. Hay que establecer los parámetros *featurewise\_center* y *featurewise\_std\_normalization* a *True* para que se realice la normalización que se desea.
* Se obtienen los pesos con los que se va a realizar la normalización de los datos mediante **el conjunto de entrenamiento** (*datagen.fit(x\_train)*). Este aspecto es bastante importante puesto que el conjunto de test y validación tiene que ser normalizado con los mismos parámetros que ha sido normalizado el conjunto de entrenamiento. Si esto no fuera así, la red estaría aprendiendo a clasificar imágenes en una determinada escala y luego cuando tenga que evaluar las imágenes de test o validación, las cuales están en otra escala, los resultados serían pésimos.
* Se crean los iteradores para cada conjunto de datos (entrenamiento, validación y test) con la función *flow* del *datagen*, pasando el tamaño de batch especificado como parámetro a la función y se devuelven estos iteradores.

Finalmente se crearon las funciones *train\_model\_it* y *get\_loss\_accuracy\_it* que no son más que una extensión de las funciones explicadas en el ejercicio 1 para que puedan trabajar con iteradores de datos.

A continuación se presentan los resultados obtenidos tras esta mejora:

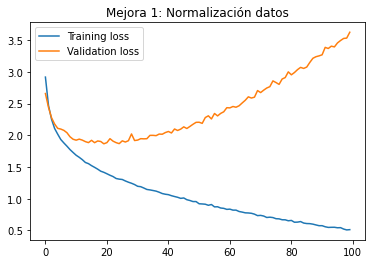


Ilustración : Evolución de la función de pérdida para la mejora 1

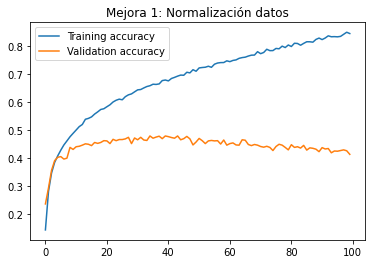


Ilustración : Evolución de la accuracy para la mejora 1

|  |  |
| --- | --- |
| **Test loss** | **Accuracy test** |
| 3.47 | 43,34% |

Tabla : Resultados de test obtenidos para la mejora 1

Si nos fijamos en los resultados obtenidos en test, se puede observar que son muy similares a los obtenidos en el ejercicio anterior. Esto era de esperar puesto que la normalización de los datos, en general, no suele aportar una mejora en el rendimiento de la red. Lo que sí se consigue es que la convergencia se alcance más rápidamente. Si se comparan estas dos últimas Ilustraciones con *Ilustración 5* e *Ilustración 6* (recuérdese que el tamaño de batch considerado para este ejercicio es de 128) se puede observar que las gráficas tienen la misma forma: la diferencia se encuentra en un desplazamiento horizontal. Esto casa con lo dicho anteriormente y con lo comentado en el ejercicio 1. Aunque la normalización de los datos de entrada no mejore el rendimiento sí que ayuda a la convergencia, por lo que considero que su impacto en la red es beneficioso.

## Mejora 2: Early Stopping

Tal y como se ha podido observar en las gráficas de la mejora anterior, está claro que la red realiza un sobreajuste (“memoriza” los datos de entrenamiento, conduciendo así a una baja capacidad de generalización y obteniendo muchos peores resultados en test que en training). Tal y como se indica en el enunciado de la práctica, una forma de reducir dicho sobreajuste es usando la técnica de *Early Stopping*. El funcionamiento de esta técnica se basa en parar el entrenamiento de la red cuando se detecta que la función de pérdida en el conjunto de validación aumenta en lugar de disminuir. Si se observa la gráfica de la evolución de la función de pérdida en la mejora anterior (*Ilustración 9*) es fácil ver que a partir de la época 20, la función de pérdida en el conjunto de validación aumenta mientras que la de entrenamiento sigue disminuyendo. Es en este punto en el que la red empieza a realizar sobreajuste. Con la técnica de *Early Stopping* lo que se pretende es detectar cuando pasa esto y cortar el entrenamiento en dicho punto.

La dificultad del uso de esta técnica reside en fijar el número de épocas para el cual se va a comprobar el decrecimiento de la función de pérdida. Por ejemplo, si este parámetro se establece a pocas épocas puede darse el caso de que, por azar, haya un momento en el que la función de validación empiece a aumentar (pero, tras un determinado número de épocas sigue decreciendo) y se corte el entrenamiento. Por el caso contrario, si se establece a un número de épocas muy alto, los beneficios de la técnica se ven limitados. En nuestro caso, tras numerosas pruebas, he decidido establecer este parámetro a **15 épocas**, ya que con este valor he obtenido buenos resultados.

Cabe mencionar también que las mejoras que se van aplicando son incrementales, es decir, la mejora 2 también incorpora la mejora 1. Esto será así para todas las mejoras implementadas en este ejercicio.

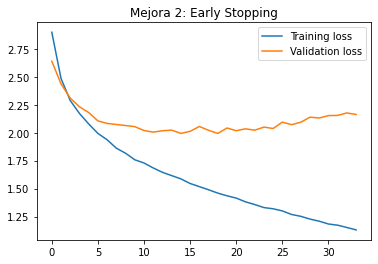


Ilustración : Evolución de la función de pérdida para la mejora 2

La función implementada para esta mejora es *train\_model\_early*, una adaptación de la función encargada del entrenamiento de la red que se encarga de definir el *callback* necesario para implementar dicha técnica en el entrenamiento. Dicho *callback* se ha definido con los siguientes parámetros:

* *monitor*: *val\_loss*, es decir, la métrica que se comprueba para realizar *Early Stopping* es la función de pérdida en el conjunto de validación.
* *patience*: 15, tal y como se ha explicado antes con este valor se han obtenido buenos resultados.
* *restore\_best\_weights*: *True*, para que los pesos sean restablecidos a la época en la que se alcanzó el mínimo en la función de pérdida en el conjunto de validación. Esto es bastante útil ya que permite obtener los mejores pesos encontrados en todo el proceso de entrenamiento de la red.

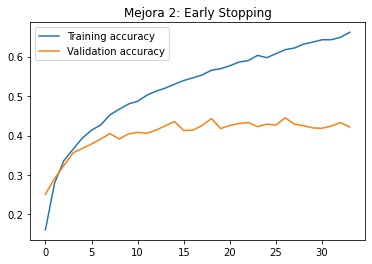


Ilustración : Evolución de la accuracy para la mejora 1

Las gráficas obtenidas para esta mejora pueden verse en *Ilustración 11* e *Ilustración 12*. Comparando dichas gráficas con las de la mejora anterior es fácil darse cuenta de que la técnica está funcionando correctamente puesto que detiene el entrenamiento cuando se detecta que se empieza a producir sobreajuste (el entrenamiento dura alrededor de 35 épocas en lugar de 100). A continuación se presentan los resultados obtenidos en test:

|  |  |
| --- | --- |
| **Test loss** | **Accuracy test** |
| 2.02 | 44,2% |

Tabla : Resultados de test obtenidos para la mejora 2

Considero que la introducción de la técnica de *Early Stopping* es beneficiosa puesto que se mejoran levemente los resultados de test con respecto a la mejora anterior y además se reduce el tiempo de entrenamiento, al entrenar menos épocas.

## Mejora 3: Aumento de datos

Otra de las mejoras que se sugieren en el enunciado de la práctica es la introducción del aumento de datos. Es lógico pensar que esta técnica puede conducir a una mejora en el rendimiento, ya que lo que se hace es aumentar el número de ejemplos en el dataset, y bien es sabido que cuantos más datos se tienen para aprender, mejor es el rendimiento que puede alcanzar la red. Estos nuevos ejemplos se consiguen mediante transformaciones geométricas, de brillo, de color, etc. De hecho, si se realiza un aumento de datos correcto puede conseguirse que la red sea robusta a dichas variaciones (geométricas, de brillo, etc.), lo cual lleva a un gran desempeño (recuérdese que las redes convolucionales, por sí solas, sólo son invariantes a traslaciones locales). Sin embargo, hay que tener cuidado en dos aspectos cuando se realiza esta técnica:

* No emplear un aumento de datos muy agresivo, ya que se puede desvirtuar el significado de las imágenes. Considérese el problema de MINST: si a una fotografía de un 6 se le aplica la rotación correspondiente, el significado de la fotografía puede pasar a ser un 9, consiguiendo el efecto contrario a lo que proponía con esta técnica.
* Realizar el aumento de datos sólo en el conjunto de entrenamiento. Si se realiza el aumento de datos también en el conjunto de validación y en el de test se estaría resolviendo un problema distinto al original, lo cual no es lo que se pretende en general.

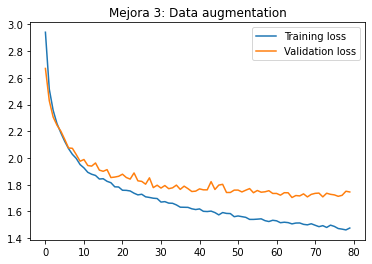


Ilustración : Evolución de la función de pérdida para la mejora 3

La función que se ha implementado para esta mejora es *data\_augmentation\_normalized*, adaptación de la función diseñada en la mejora 1 para que los datos de entrenamiento sean aumentados y normalizados mientras que los datos de test y validación sólo son normalizados. El funcionamiento de esta función es simple:

* Separación de los datos de entrenamiento iniciales en dos conjuntos: un nuevo conjunto de entrenamiento y un conjunto de validación (se usa la función anteriormente descrita, *split\_train\_val*).
* Se crean dos objetos de la clase *ImageDataGenerator*: uno para el conjunto de entrenamiento (aumento y normalización) y otro para los conjuntos de validación y test (sólo normalización). El aumento de datos que se hace para el conjunto de entrenamiento son rotaciones (con el parámetro *rotation\_range*), zoom (con el parámetro *zoom\_range*) y volteados horizontales (con el parámetro *horizontal\_flip*). Para realizar esta elección hice varias pruebas con transformaciones de brillo, de colores, etc. Pero esta combinación es la que mejor resultados me dio (en muchas ocasiones el aumento de datos empeoró los resultados, esto puede deberse a que apliqué un aumento de datos muy agresivo). El valor de dichos parámetros es:
  + *rotation\_range* *= 20*. Se aplican rotaciones con un máximo de 20 grados sobre las imágenes.
  + *zoom\_range = 25*. Se aplica zoom en el intervalo *[0.75, 1.25]* a las imágenes.
  + *horizontal\_flip = True*. Se realizan volteados horizontales de las imágenes.
* Se ajustan ambos objetos con los datos del conjunto de entrenamiento, tal y cómo se hacía en la mejora 1.
* Se crean los iteradores para cada conjunto de datos (entrenamiento, validación y test) con la función *flow* del *datagen*, pasando el tamaño de batch especificado como parámetro a la función y se devuelven estos iteradores.

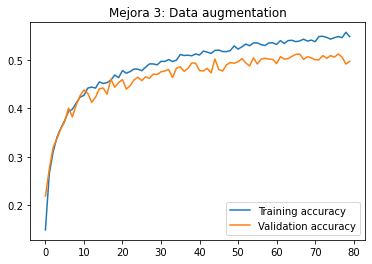


Ilustración : Evolución de la accuracy para la mejora 3

Las gráficas obtenidas para esta mejora pueden verse en *Ilustración 13* e *Ilustración 14*. Se puede observar que con el aumento de datos el modelo no tiene tanta tendencia a realizar sobreajuste. Esto tiene bastante sentido: hemos aumentado el número de ejemplos de entrenamiento, por lo cual la red ahora lo tiene más complicado para “memorizar” los datos de entrenamiento, y por lo tanto, lo tiene más difícil para realizar sobreajuste. Quizás si el entrenamiento durase más épocas podría verse este efecto reflejado en las gráficas. A continuación se presentan los resultados obtenidos en test:

|  |  |
| --- | --- |
| **Test loss** | **Accuracy test** |
| 1.62 | 52,36% |

Tabla : Resultados de test obtenidos para la mejora 3

Como se puede observar, hay una mejora sensible con respecto al apartado anterior. La conclusión que extraigo es que el aumento de datos, si se realiza correctamente, aumenta el rendimiento de la red (por los motivos anteriormente comentados).

## Mejora 4: Profundización de la red

La siguiente mejora que voy a implementar va a ser aumentar la profundidad del modelo *BaseNet*. En teoría se ha visto que las redes profundas consiguen un mayor rendimiento debido a que se van realizando sucesivas transformaciones y selecciones de la información de las imágenes de entrada. Con una red poco profunda, las características a extraer estarán bastante limitadas y puede darse el caso de que el rendimiento no sea óptimo (no se extraen características de calidad). Viéndolo desde una perspectiva desde el Aprendizaje Automático, al aumentar la profundidad de la red lo que se está haciendo es incrementar la complejidad del modelo (el número de parámetros a aprender crece) y, por lo tanto, se está expandiendo el espacio de funciones en el cuál se busca la solución. Está claro que al expandir dicho espacio de funciones la probabilidad de encontrar una buena solución para el problema también crece.

Cabe destacar, que bajo según mi interpretación del ejercicio, lo que se pretende que hagamos es añadir capas convolucionales **sin modificar las capas base que hay en BaseNet**. He hecho pruebas cambiando un poco estas capas base (modificando un poco la arquitectura, cambiando el número de filtros, etc.) y la mejora ha sido muy notoria. Sin embargo, debido a la limitación anteriormente comentada, el nuevo modelo diseñado tiene la siguiente estructura:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Layer No.** | **Layer Type** | **Kernel size (for conv layers)** | **Input|Output dimension** | **Input|Output channels (for conv layers)** |
| 1 | Conv2DTranspose | 3 | 32|34 | 3|6 |
| 2 | Relu | - | 34|34 | - |
| 3 | Conv2DTranspose | 3 | 34|36 | 6|6 |
| 4 | Relu | - | 36|36 | - |
| 5 | Conv2D | 5 | 36|32 | 6|6 |
| 6 | Relu | - | 32|32 | - |
| 7 | MaxPooling2D | 2 | 32|16 | - |
| 8 | Conv2D | 5 | 16|12 | 6|16 |
| 9 | Relu | - | 12|12 | - |
| 10 | Conv2D | 3 | 12|10 | 16|32 |
| 11 | Relu | - | 10|10 | - |
| 12 | MaxPooling2D | 2 | 10|5 | - |
| 13 | Linear | - | 800|50 | - |
| 14 | Relu | - | 50|50 | - |
| 15 | Linear | - | 50|25 | - |

Tabla : Descripción de BaseNet profundo

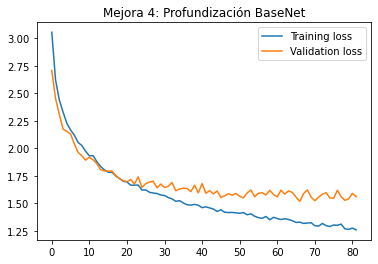


Ilustración : Evolución de la función de pérdida para la mejora 4

Como se puede observar en la descripción del nuevo modelo (*Tabla 10*), las capas que se han añadido son 1, 2, 3, 4, 10 y 11. La decisión de usar convoluciones traspuestas al inicio se basa en la asunción de que la información más relevante de las imágenes suele estar en la parte central, y dicha forma de convolución enfatiza las características centrales. Por otra parte, también ayuda a obtener bloques convolucionales de mayor dimensión (este es el principal problema a la hora de profundizar la red, al ser las imágenes de entrada de dimensiones 32x32, no hay mucho margen de maniobra a la hora de profundizar la red). La convolución implementada en la capa número 10 se hace con el objetivo de extraer el mayor número de características posible (se usan 32 filtros) para que el clasificador disponga de la información suficiente.

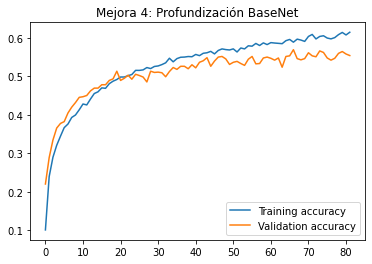


Ilustración : Evolución de la accuracy para la mejora 4

En los gráficos presentes en *Ilustración 15* e *Ilustración 16* se puede observar la evolución de la función de pérdida y la *accuracy*. A continuación se presentan los resultados obtenidos en test:

|  |  |
| --- | --- |
| **Test loss** | **Accuracy test** |
| 1.49 | 56,12% |

Tabla : Resultados de test obtenidos para la mejora 4

Como era de esperar, los resultados obtenidos tras la profundización de la red son mejores que los anteriores (debido a los motivos anteriormente comentados).

## Mejora 5: Batch Normalization

La siguiente mejora que pensé en introducir ha sido Batch Normalization. La idea subyacente a esta técnica es bastante lógica: si los datos de entrada se normalizan para un mejor desempeño, ¿por qué no normalizar los resultados obtenidos tras las diferentes operaciones en la red convolucional? Sin embargo, ahora mismo hay un gran debate en la comunidad científica sobre el uso y los beneficios de esta técnica. Se dice que esta técnica ayuda a la regularización de la red, que ayuda a la convergencia, que mejora los resultados, etc. Es por este motivo por el cual se va a experimentar con esta técnica para luego analizar los resultados obtenidos, al no haber fundamentos teóricos suficientes (al menos hasta la fecha).

Para la introducción de esta técnica se ha usado la capa *BatchNormalization* de Keras. Se ha optado por usar los parámetros por defecto debido a que se usaron múltiples configuraciones de parámetros y en todas las pruebas se obtuvieron resultados similares.

Tal y como se indica en el enunciado de la práctica, se pide que se experimente introduciendo capas de *BatchNormalization* después de las capas convolucionales y antes de ReLU, y también después de las capas de ReLU. En este sentido, se probó con ambas configuraciones, pero se obtuvieron resultados similares. Así que mi decisión final fue por optar dejar las capas de *BatchNormalization* antes de las capas ReLU, ya que esta decisión me parece la más intuitiva (normalizar los datos antes de que pasen por la función de activación).

Con todo lo anteriormente dicho, la estructura final de la red es la siguiente:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Layer No.** | **Layer Type** | **Kernel size (for conv layers)** | **Input|Output dimension** | **Input|Output channels (for conv layers)** |
| 1 | Conv2DTranspose | 3 | 32|34 | 3|6 |
| 2 | BatchNormalization | - | 34|34 | - |
| 3 | Relu | - | 34|34 | - |
| 4 | Conv2DTranspose | 3 | 34|36 | 6|6 |
| 5 | BatchNormalization | - | 36|36 | - |
| 6 | Relu | - | 36|36 | - |
| 7 | Conv2D | 5 | 36|32 | 6|6 |
| 8 | BatchNormalization | - | 32|32 | - |
| 9 | Relu | - | 32|32 | - |
| 10 | MaxPooling2D | 2 | 32|16 | - |
| 11 | Conv2D | 5 | 16|12 | 6|16 |
| 12 | BatchNormalization | - | 12|12 | - |
| 13 | Relu | - | 12|12 | - |
| 14 | Conv2D | 3 | 12|10 | 16|32 |
| 15 | BatchNormalization | - | 10|10 | - |
| 16 | Relu | - | 10|10 | - |
| 17 | MaxPooling2D | 2 | 10|5 | - |
| 18 | Linear | - | 800|50 | - |
| 19 | BatchNormalization | - | 50|50 | - |
| 20 | Relu | - | 50|50 | - |
| 21 | Linear | - | 50|25 | - |

Tabla : Resultados de test obtenidos para la mejora 5

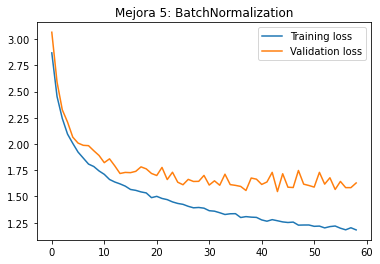


Ilustración : Evolución de la función de pérdida para la mejora 5

Tanto en la *Ilustración 17* como en la *Ilustración 18* se puede ver el efecto que ha tenido la introducción de *BatchNormalization*: la forma de las gráficas es bastante similar a las de la mejora anterior; sin embargo, los resultados obtenidos con *BatchNormalization* han sido conseguidos en menos épocas.

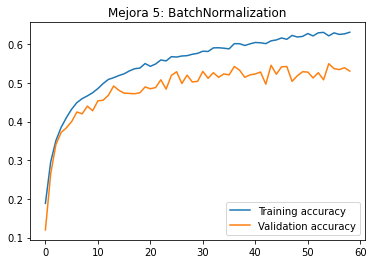


Ilustración : Evolución de la accuracy para la mejora 5

A continuación se presentan los resultados obtenidos en test:

|  |  |
| --- | --- |
| **Test loss** | **Accuracy test** |
| 1.48 | 55,52% |

Tabla : Resultados de test obtenidos para la mejora 5

Como se puede observar, los resultados obtenidos en test son bastantes similares a los obtenidos en la mejora anterior. Sin embargo, hay una clara diferencia: el entrenamiento ha requerido de bastantes menos épocas. Por este motivo, aunque no se haya conseguido una mejora sensible de los resultados, considero el efecto de *BatchNormalization* beneficioso.

## Mejora 6: Dropout

Finalmente opté por implementar la última mejora de las listadas en el enunciado: *Dropout*. Esta técnica consiste en “apagar” de forma aleatoria un porcentaje de las neuronas activas para cada mini batch. De esta manera se consigue que la red especialice las neuronas en informaciones específicas e independientes. Se ha demostrado que los efectos de esta capa ayudan de forma notoria a la regularización de la red.

Cabe mencionar que, tras buscar información sobre el uso de *Dropout* en redes convolucionales hay distintas aproximaciones. Una de ellas es introducir una capa de *Dropout*, en el proceso de extracción de características, es decir, tras la función de activación perteneciente a las capas convolucionales. Dichas capas de *Dropout* deben de tener un porcentaje pequeño de “apagado” de neuronas: esto es así porque lo que se está haciendo es ignorar cierta parte de la información de entrada. Si se pusiese un porcentaje alto, se estaría ignorando la mayoría de la información de entrada lo cual no conduce a un buen rendimiento por razones obvias. Si se piensa en esta aplicación de *Dropout* puede tener sentido, ya que, como en la mayoría de problemas los datos de entrada no son limpios, sino que siempre tienen algo de ruido (con lo cual, con esta aplicación de la técnica se podría intentar eliminar ese ruido a la entrada). Sin embargo, yo probé este tipo de aplicación y no me dio buenos resultados. Finalmente opté por introducir las capas de *Dropout* en la parte final concerniente al clasificador propiamente dicho. El porcentaje de apagado de neuronas fue establecido a **0.5**, porcentaje que recomienda el autor de *Dropout* en su paper.

Con todo lo dicho anteriormente, la estructura final de la red quedó así:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Layer No.** | **Layer Type** | **Kernel size (for conv layers)** | **Input|Output dimension** | **Input|Output channels (for conv layers)** |
| 1 | Conv2DTranspose | 3 | 32|34 | 3|6 |
| 2 | BatchNormalization | - | 34|34 | - |
| 3 | Relu | - | 34|34 | - |
| 4 | Conv2DTranspose | 3 | 34|36 | 6|6 |
| 5 | BatchNormalization | - | 36|36 | - |
| 6 | Relu | - | 36|36 | - |
| 7 | Conv2D | 5 | 36|32 | 6|6 |
| 8 | BatchNormalization | - | 32|32 | - |
| 9 | Relu | - | 32|32 | - |
| 10 | MaxPooling2D | 2 | 32|16 | - |
| 11 | Conv2D | 5 | 16|12 | 6|16 |
| 12 | BatchNormalization | - | 12|12 | - |
| 13 | Relu | - | 12|12 | - |
| 14 | Conv2D | 3 | 12|10 | 16|32 |
| 15 | BatchNormalization | - | 10|10 | - |
| 16 | Relu | - | 10|10 | - |
| 17 | MaxPooling2D | 2 | 10|5 | - |
| 18 | Linear | - | 800|50 | - |
| 19 | BatchNormalization | - | 50|50 | - |
| 20 | Relu | - | 50|50 | - |
| 21 | Dropout | - | 50|50 | - |
| 22 | Linear | - | 50|25 | - |

Tabla : Resultados de test obtenidos para la mejora 6

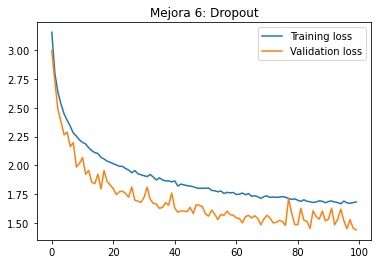


Ilustración : Evolución de la función de pérdida para la mejora 6

Como se puede ver en *Ilustración 19* e *Ilustración 20*, hay un efecto un tanto anómalo en las gráficas: la función de pérdida en validación es menor que en entrenamiento y la *accuracy* es mayor en validación que en entrenamiento. Esto se debe a la implementación de *Dropout* en Keras: a la hora de evaluar en entrenamiento se “apagan” las neuronas y en validación se “encienden”. Esto hace que los resultados en entrenamiento sean peores debido a la eliminación de información.

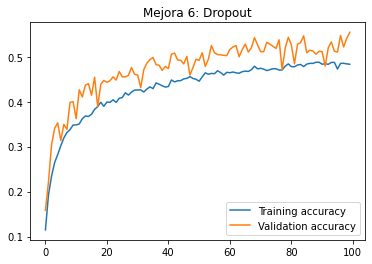


Ilustración : Evolución de la accuracy para la mejora 6

A continuación se presentan los resultados en test:

|  |  |
| --- | --- |
| **Test loss** | **Accuracy test** |
| 1.42 | 56,56% |

Tabla : Resultados de test obtenidos para la mejora 6

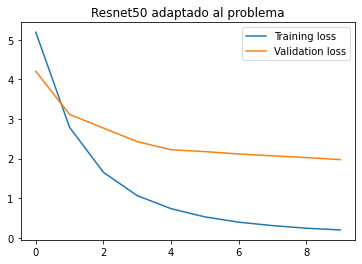
Como se puede observar, los resultados obtenidos con esta mejora se corresponden con los mejores resultados de todas las mejoras. Esto indica que la capa de *Dropout* es beneficiosa para la red al ayudar a la regularización de la misma.

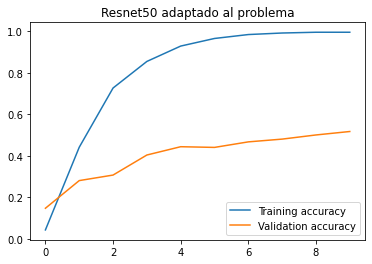
## Conclusiones

Tras las diferentes implementaciones de las distintas mejoras se ha podido observar cómo se ha pasado de un 42-44% de *accuracy* a un **56%** de *accuracy* con la mejora final. Esto pone de manifiesto todas las variables y técnicas que entran en juego a la hora de diseñar una red. Esto no es una tarea fácil puesto que se necesita experiencia en el diseño de redes convolucionales. También se ha podido verificar la facilidad con la que se puede implementar una red y hacer cambios en la misma con Keras, lo cual es de agradecer.

# Ejercicio 3

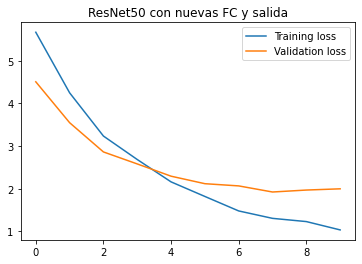
## Apartado A

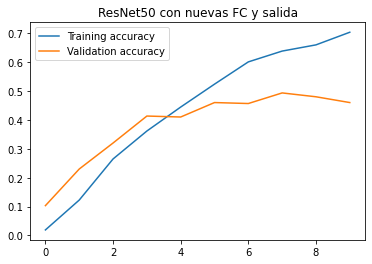




Test loss: 2.3560123443603516

Accuracy test: 0.4207055866718292

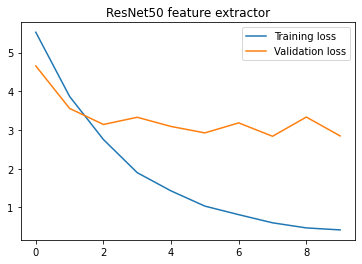


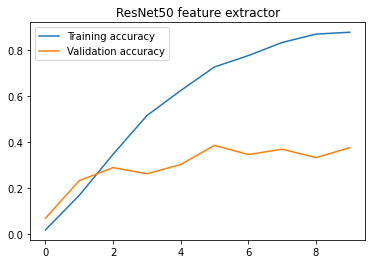


Test loss: 2.446516990661621

Accuracy test: 0.4048796594142914

## Apartado B

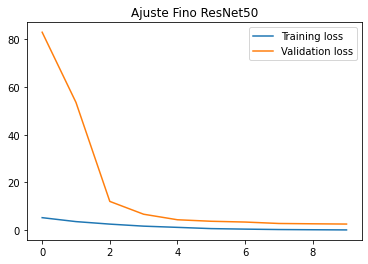


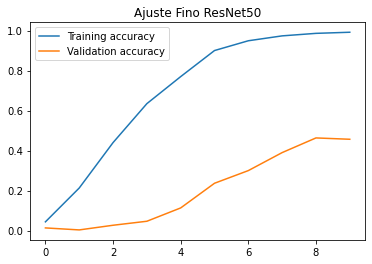


Test loss: 3.6546714305877686

Accuracy test: 0.3125618100166321

## Apartado C





Test loss: 2.8896753787994385

Accuracy test: 0.35146719217300415