

Evaluación de Multilayer Perceptron en clasificación de imágenes usando el conjunto de datos QuickDraw

John Patricio Serrano Carrasco
Agosto de 2024

Resumen— Se entrenaron y evaluaron tres modelos Multilayer Perceptron (MLP) para la clasificación de imágenes utilizando los conjuntos de datos QuickDraw-10 y QuickDraw-Animals. Los modelos se diseñaron con entre dos y cuatro capas ocultas, cada uno utilizando diferentes configuraciones. Cada modelo fue entrenado y evaluado cinco veces, utilizando una inicialización de pesos aleatoria. Se calcularon la mediana de la accuracy total, la mediana de la accuracy por clase y la mediana de la matriz de confusión a lo largo de las cinco ejecuciones. Los resultados demostraron variaciones significativas en la accuracy entre los modelos y las clases, resaltando la importancia de la estructura del MLP en el rendimiento de la clasificación.

Palabras claves— Multilayer Perceptron, Función de Activación, Función de Pérdida y Clasificación de imágenes.

I. INTRODUCCIÓN

En la actualidad, la clasificación automática de imágenes se ha convertido en una herramienta crucial en diversas aplicaciones tecnológicas. En este contexto, los conjuntos de datos QuickDraw-10 y QuickDraw-Animals proporcionan una base sólida para desarrollar y evaluar algoritmos de clasificación de imágenes. Este estudio se centró en evaluar la efectividad de Multilayer Perceptron (MLP), una técnica robusta de Aprendizaje Supervisado, para clasificar dibujos hechos a mano. La experiencia se dividió en dos experimentos principales. En el primer experimento, se utilizaron tres modelos MLP con diferentes arquitecturas, principalmente con dos, tres y cuatro capas ocultas junto con diferentes configuraciones para evaluar la accuracy total y por clase en la clasificación de imágenes del conjunto de datos QuickDraw-10. En el segundo experimento, los mismos modelos se aplicaron al conjunto de datos QuickDraw-Animals para comparar la efectividad de las diferentes arquitecturas en un contexto de imágenes más variadas. Cada fase del experimento siguió una metodología consistente, lo cual permitió calcular la mediana de la accuracy total, la mediana de la accuracy por clase y la mediana de la matriz de confusión, proporcionando una medida robusta del desempeño del modelo en diferentes contextos. Además, se elaboraron gráficos para análisis y comparar el rendimiento entre los distintos modelos en cada conjunto de datos.

II. DESARROLLO

Se diseñaron tres modelos MLP con diferentes configuraciones de capas ocultas. El **primer modelo** incluyó dos capas densas ocultas de 128 y 64 neuronas, con funciones de activación **ReLU**, la cual reduce el problema de desvanecimiento del gradiente y permite que el modelo aprenda

relaciones complejas en los datos, y **Sigmoid**, que puede ser más útil en problemas de clasificación binaria y por lo tanto, menos eficiente que ReLU en el contexto planteado. La función de pérdida de este modelo es **Sparse Categorical Cross-Entropy**, la cual es utilizada cuando las clases están representadas como enteros, y es especialmente útil cuando se trabaja con un gran número de clases.

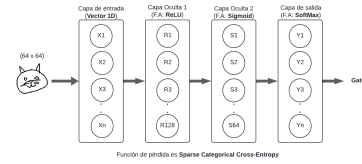


Figura 1: Diagrama de bloques del modelo MLP 1.

El segundo modelo, tiene tres capas densas ocultas, cada una de 256 neuronas y función de activación ReLU, con función de pérdida **Categorical Cross-Entropy**, la cual es muy efectiva para entrenar modelos en contextos donde cada instancia puede pertenecer exactamente a una de las múltiples clases.

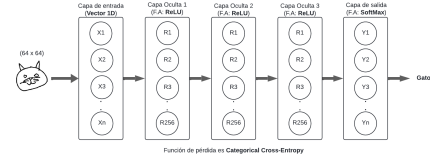


Figura 2: Diagrama de bloques del modelo MLP 2.

El tercer modelo, añade otra capa oculta, cada una con función de activación ReLU y con 512, 256, 128 y 64 neuronas respectivamente, donde la función de pérdida es **Categorical Cross-Entropy**.

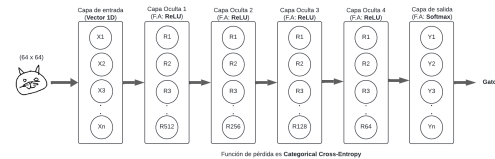


Figura 3: Diagrama de bloques del modelo MLP 3.

Los tres modelos tienen una capa de salida cuyo tamaño es el número de clases, con función de activación SoftMax. Además, cada modelo es compilado con el optimizador Adam, el cual es ampliamente considerado efectivo para problemas de clasificación debido a su capacidad para adaptar las tasas de aprendizaje de manera individual para cada parámetro, lo

que ayuda a manejar eficientemente los diferentes tipos de datos y características. Para todos los modelos se utiliza una inicialización de pesos aleatoria cada vez que se compila.

En ambos experimentos se utilizó la misma metodología; cada modelo MLP fue entrenado y evaluado cinco veces, donde en cada iteración se obtiene y se guarda la accuracy total, la accuracy por cada clase y la matriz de confusión asociada. Al final de las cinco iteraciones, se obtiene la mediana de la matriz de confusión, la mediana de la accuracy total y la mediana de la accuracy por cada clase. Luego de obtener estos datos, se realizan tablas y gráficos de barras para su posterior análisis. El entrenamiento se realiza utilizando 20 épocas y tamaño de lote 32, la razón de estos valores siendo que se busca experimentar rápidamente con las diferentes arquitecturas y configuraciones de los modelos MLP. Al final de cada experimento se realizan cambios de hiperparámetros de tal forma de saber como cambia el rendimiento de los modelos en distintos casos. En el **primer experimento**, se utiliza el conjunto de datos QuickDraw-10, mientras que en el **segundo experimento** se utiliza el conjunto de datos QuickDraw-Animals.

Se desarrollaron funciones para evitar la repetición de código, además, se utilizó TensorFlow y Keras para la implementación y entrenamiento de los modelos MLP, Pandas para la manipulación de datos, NumPy para operaciones matemáticas, y Matplotlib para visualizar los resultados mediante gráficos y realizar análisis comparativos detallados.

III. RESULTADOS EXPERIMENTALES

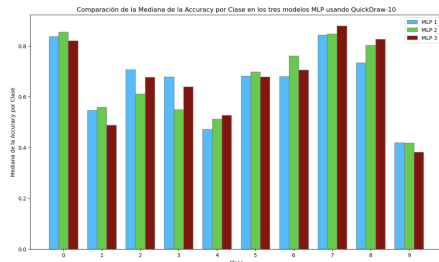


Figura 4: Gráfico de Medianas de Accuracy por Clase

Respecto al experimento 1 y el modelo MLP 1, la mediana de la accuracy total es 0,60 y las iteraciones presentan una accuracy total consistente, a excepción de la iteración 3, donde la accuracy es notablemente baja. Este descenso abrupto en la accuracy sugiere posibles fluctuaciones en el proceso de entrenamiento. Además, la incapacidad del modelo para superar una accuracy total de 0.63 indica que las limitaciones en la arquitectura del modelo están afectando su capacidad para generalizar y capturar completamente las características de los datos. De la Figura 4 se desprende que algunas clases son más fáciles de clasificar que otras. Las clases 0 y 7 muestran una alta accuracy de 0.83 y 0.84 respectivamente, lo que indica que estas clases poseen características más distintivas y son más fáciles de aprender para el modelo. Por otro lado, las clases 4 y 9 presentan una accuracy significativamente más baja de 0.47 y 0.42 respectivamente, lo que podría deberse a similitudes con otras clases que dificultan la discriminación.

Respecto a la mediana de la matriz de confusión, se observa en la Figura 5 que en ciertas clases, como la clase 1, hay una cantidad significativa de valores fuera de la diagonal principal. La clase 9 muestra una cantidad considerable de confusiones con otras clases, indicando que el modelo tiene dificultades para diferenciarla claramente, mientras que la clase 8 tiene un alto número de predicciones correctas.

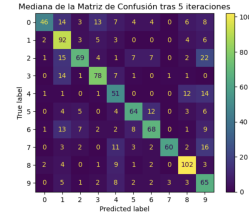


Figura 5: Mediana de Matriz de Confusión usando MLP1

El modelo MLP 1 muestra limitaciones en su capacidad para capturar características complejas de los datos de entrada. Las clases con menor número de predicciones correctas, sugieren que el modelo necesita mejoras adicionales, ya sea mediante el ajuste de los hiperparámetros, la arquitectura del modelo, entre otros. Al aumentar la cantidad de épocas y tamaño de lote, la mediana de la accuracy total aumenta a 0,64, mientras que aumentar el número de neuronas disminuyó considerablemente el rendimiento del modelo.

En el modelo MLP 2, los valores de accuracy total oscilan entre 0,54 y 0,65, con una mediana de 0.62. Similar al modelo MLP 1, el modelo tiene un rendimiento decente en general, pero hay fluctuaciones que podrían ser atribuibles a la inicialización aleatoria de los pesos o a la variabilidad en los datos de entrenamiento. Analizando la mediana de la accuracy por clase se encuentra que la accuracy más alta es para la clase 0 con 0,85, mientras que la más baja corresponde a la clase 9 con 0,42. La Figura 6 permite ver que las clases 1, 2, 3, y 8 tienen valores en la diagonal de 92, 87, 95 y 92, respectivamente. Esto indica que el modelo puede identificar correctamente la mayoría de las características de estas clases. Las clases 4, 5, 6, y 7 muestran valores más bajos en la diagonal, indicando que el modelo tiene dificultades para clasificar correctamente estas clases. Por ejemplo, la clase 4 solo tiene 47 instancias correctamente clasificadas y se confunde con frecuencia con la clase 9.

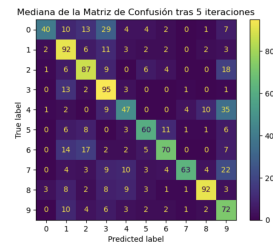


Figura 6: Mediana de Matriz de Confusión usando MLP2

El diseño del modelo MLP 2, parece ser efectivo en la mayoría de los casos, pero todavía presenta limitaciones en la capacidad de diferenciación entre ciertas clases. Tras aumentar la cantidad de épocas y tamaño de lote, la mediana de la

accuracy total aumenta a 0,65. Al cambiar la función de activación de dos capas de ReLU a ELU, la cual combina las ventajas de ReLU con una respuesta más suave para valores negativos, se aumenta levemente la accuracy total a 0,63.

El modelo MLP 3 muestra una mejora respecto a los modelos anteriores en ciertos aspectos, pero aún presenta algunas limitaciones y tiene un mayor tiempo de entrenamiento y evaluación. La mediana de la accuracy total es de 0.60, ligeramente inferior a la del segundo modelo, lo que indica que el aumento en la complejidad del modelo no necesariamente se traduce en un mejor rendimiento global. En términos de accuracy por clase, las más altas son la clase 0 con 0.82 y la clase 7 con 0.83 de accuracy, lo que indica un buen reconocimiento de características en estas clases. Sin embargo, otras clases como la clase 1 con accuracy de 0.49, tienen un rendimiento significativamente menor. La matriz de confusión permite ver esto, ya que la clase 0 tiene 57 instancias correctamente clasificadas y pocas confusiones con otras clases, pero también se observa una notable confusión entre clases como la clase 3 y la clase 1, donde muchos ejemplos de la clase 3 se clasifican incorrectamente como clase 1 y viceversa.

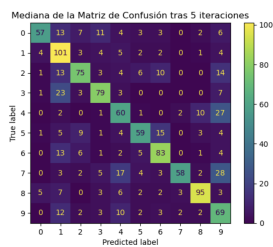


Figura 7: Mediana de Matriz de Confusión usando MLP3

Aunque el tercer modelo MLP mejora en ciertas áreas específicas, su rendimiento global es solo marginalmente mejor que el de los modelos más simples. Al aumentar las épocas y el tamaño del lote, la mediana de la accuracy total aumenta a 0,66, mientras que aumentar el número de neuronas aumenta la mediana de accuracy total a 0,65.

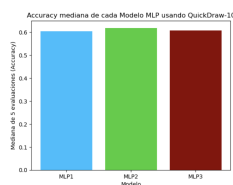


Figura 8: Gráfico de Mediana de la Accuracy por MLP

La Figura 8 muestra que con este conjunto de datos, la mediana de accuracy total fue similar en los tres modelos. En el caso de las clases, la Figura 4 revela que algunas clases, como la 0, 7 y 8, presentan una alta accuracy en todos los modelos. Esto indica que las características de estas clases son bien capturadas incluso por las diferentes arquitecturas de los modelos MLP. Sin embargo, otras clases como la 9 muestran una notable variabilidad en el rendimiento. La clase 9 mantiene una accuracy baja en todos los modelos, lo que indica una dificultad general en clasificar esta clase. En términos

generales, incrementar la complejidad del modelo no siempre resulta en mejoras lineales para todas las clases. Al realizar cambios de hiperparámetros, existen pros y contras para distintos modelos. Por ejemplo, aumentar el número de neuronas al doble en MLP3 aumenta la accuracy, pero también aumenta significativamente el tiempo de entrenamiento y evaluación. En cambio, el modelo MLP 1 se ve perjudicado por el aumento de neuronas, demostrando que surge un desafío a la hora de encontrar la mejor configuración de un modelo MLP.

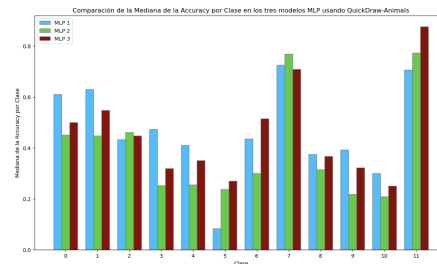


Figura 9: Gráfico de Medianas de Accuracy por Clase

Al evaluar el rendimiento del modelo MLP 1 con el conjunto de datos QuickDraw-Animals, se observa un desempeño considerablemente inferior en comparación con QuickDraw-10. La mediana de la accuracy total tras 5 iteraciones es de apenas 0.083, lo cual indica una capacidad de clasificación limitada para este conjunto de datos más complejo. Las accuracy por clase muestran una variabilidad significativa, con valores que van desde aproximadamente 0.3 hasta 0.71. Esto sugiere que el modelo es capaz de clasificar ciertas clases al menos en una iteración, mientras que otras son prácticamente indetectables, como lo demuestra la mediana de la matriz de confusión. La Figura 9 revela que algunas clases, como la 7 y 11, alcanzan valores cercanos a 0.70, lo cual contrasta con otras clases que apenas superan el 0.30. La matriz de confusión mediana también refleja estos resultados, mostrando una gran cantidad de predicciones incorrectas distribuidas uniformemente entre las clases incorrectas, con una excepción notable en la clase 7 que tiene un número significativamente mayor de predicciones correctas.

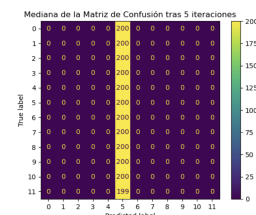


Figura 10: Mediana de Matriz de Confusión usando MLP1

Se puede apreciar que el modelo MLP 1 no está bien adaptado para manejar la complejidad y variabilidad de QuickDraw-Animals. Curiosamente, al aumentar las épocas y el tamaño del lote, la mediana de la accuracy total aumenta a 0,44, lo que sugiere que al menos aumentar el tamaño del lote es bastante efectivo para este modelo. Se tiene una accuracy bastante similar al cambiar la función de activación Sigmoid por ELU.

En el caso de MLP 2, la mediana de la accuracy total fue de 0.36, lo que indica una mejora comparativa respecto al

primer modelo, pero aún con rendimientos variables según la iteración. La accuracy más alta fue de aproximadamente 0.40, mientras que las otras iteraciones oscilaron alrededor de valores más bajos. Respecto a la mediana de la accuracy por clase, las clases 7 y 11 obtuvieron una mediana de la accuracy alta, superando el 0.70. En contraste, otras clases, como las clases 5 y 10 presentaron un rendimiento considerablemente bajo, con valores por debajo de 0.25, indicando dificultades del modelo para diferenciar correctamente estas categorías. La matriz de confusión mediana refuerza estos hallazgos. Por ejemplo, la clase 0, a pesar de tener un número decente de predicciones correctas, presenta una cantidad considerable de confusiones con otras clases, especialmente con la clase 3 y la clase 4. Esto podría indicar que las características visuales de estas clases se solapan de manera que el modelo encuentra difícil de distinguir.

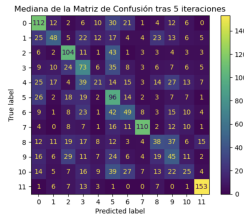


Figura 11: Mediana de Matriz de Confusión usando MLP2

Al aumentar las épocas y el tamaño del lote, la mediana de la accuracy total aumenta a 0,45, mientras que al modificar la función de pérdida a Sparse Categorical Cross-Entropy, la mediana de la accuracy total disminuye a 0,35.

En el caso de MLP 3, se obtuvo una mediana de accuracy total de 0.4406. La accuracy total se mantuvo en torno a 0.4 en todas las iteraciones. Similar a QuickDraw-10, toma el mayor tiempo en completar las cinco iteraciones. La clase 7 tuvo la accuracy más alta con una mediana de 0.88, en contraste, la clase 9 mostró un bajo desempeño con una accuracy de 0.2707. La Figura 12 muestra que la clase 7 destacó con 129 instancias correctamente clasificadas, lo que se refleja en su mediana de accuracy de 0.88. Sin embargo, la clase 3 tuvo un desempeño mucho más pobre, ya que el modelo frecuentemente confundió esta clase con otras.

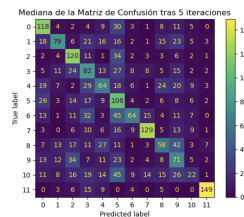


Figura 12: Mediana de Matriz de Confusión usando MLP3

Al aumentar las épocas y el tamaño del lote, la mediana de la accuracy total aumenta levemente a 0,45, algo similar sucede al cambiar la función de pérdida a Sparse Categorical Cross-Entropy.

La Figura 13 muestra que a diferencia del experimento anterior, el rendimiento de los tres modelos fue bastante variado. El Modelo MLP 3, además de lograr la mayor mediana de

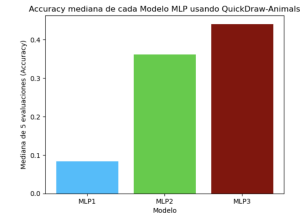


Figura 13: Gráfico de Mediana de la Accuracy por MLP.

la accuracy total, también ha mostrado una consistencia en la mayoría de las clases, particularmente en las clases donde el modelo 1 y el modelo 2 tuvieron dificultades. Este patrón sugiere que el modelo 3 podría estar más optimizado para este tipo de datos de imágenes que los otros dos modelos, posiblemente debido a la configuración y arquitectura del modelo, lo cual es un aspecto crucial en el aprendizaje profundo para la correcta discriminación de características en tareas de clasificación. Durante la experiencia se hicieron pruebas aplicando técnicas de regularización como Batch Normalization, pero no se logró obtener mejores resultados. En general, los resultados subrayan la importancia de ajustar la complejidad del modelo a las características específicas de las clases dentro del conjunto de datos para optimizar el rendimiento de clasificación.

IV. CONCLUSIONES

La experiencia realizada con los modelos MLP sobre los conjuntos de datos proporciona una comprensión profunda sobre la influencia de la arquitectura de redes neuronales en tareas de clasificación de imágenes. Los resultados destacan cómo la configuración específica de cada modelo, incluyendo el número de capas ocultas y neuronas, impacta directamente en su eficacia para discernir entre distintas clases. Aunque el tercer modelo MLP, con una arquitectura más compleja, demostró una mejora en la consistencia de los resultados en comparación con los modelos más simples, no necesariamente garantiza un mejor rendimiento general, lo que subraya la necesidad de un equilibrio entre la complejidad del modelo y la capacidad de generalización. La capacidad de ajustar el diseño de los MLP permite adaptarlos eficazmente a diversas necesidades de clasificación, mientras que su habilidad para aprender relaciones no lineales complejas es crucial para manejar variados conjuntos de datos. Por otro lado, pueden existir tiempos de entrenamiento largos y una pronunciada susceptibilidad al sobreajuste. También son sensibles a los métodos de inicialización y a la elección de hiperparámetros, lo que puede llevar a resultados inconsistentes, como es el caso del modelo MLP 1. Además, la naturaleza de “caja negra” de los MLP puede complicar la interpretación y el diagnóstico del rendimiento por clase sin herramientas especializadas. Estos aspectos resaltan la importancia de un diseño meticuloso y una gestión cuidadosa de los equilibrios necesarios al implementar estos modelos en contextos prácticos. De cara al futuro, sería pertinente explorar arquitecturas más profundas, técnicas de regularización u otras configuraciones que podrían mejorar el rendimiento de los modelos tanto de forma general como en clases con rendimiento bajo.