

Clasificación de fabricantes de vehículos

Tomas Diaz

Universidad de San Andrés
Buenos Aires, Argentina
tdiaz@udesa.edu.ar

Manuel Ramírez Silva

Universidad de San Andrés
Buenos Aires, Argentina
mramirezsilva@udesa.edu.ar

Abstract—Este trabajo desarrolla un sistema de clasificación de imágenes para identificar la marca de vehículos usando redes neuronales convolucionales y aprendizaje por transferencia. Se utilizó una versión reclasificada del Stanford Cars Dataset y se entrenaron distintos modelos, entre ellos AlexNet, ResNet50 y ResNet18. Este último obtuvo los mejores resultados, alcanzando una precisión del 79% en validación y 76% en test. Los análisis con Grad-CAM y PCA permitieron interpretar visualmente el comportamiento del modelo.

Index Terms—computer vision, car classification, CNN arquitectures

I. INTRODUCCIÓN

El objetivo de este trabajo es desarrollar un sistema de clasificación de imágenes capaz de identificar el fabricante de un automóvil a partir de una imagen. La tarea se enmarca dentro del campo de la visión por computadora y busca explorar si es posible reconocer patrones visuales característicos asociados a cada marca de vehículos.

Para ello se entrena n distinctos modelos de redes neuronales convolucionales utilizando técnicas modernas como el *transfer learning*, y se evalúa su desempeño en un conjunto de imágenes reales. Este enfoque permite aplicar herramientas actuales del aprendizaje profundo a un problema concreto, con potencial aplicación en contextos como la industria automotriz, sistemas de monitoreo o inventarios visuales automatizados.

II. CONJUNTO DE DATOS Y CARACTERÍSTICAS

En este trabajo se utilizó el *Stanford Cars Dataset* (by classes folder), disponible públicamente en la plataforma Kaggle [1]. Este conjunto de datos, desarrollado por la Universidad de Stanford, está orientado a tareas de clasificación de imágenes de automóviles y constituye un recurso ampliamente empleado en investigaciones de visión por computadora.

El dataset contiene 16.185 imágenes, distribuidas originalmente en 196 clases, cada una representando una combinación específica de marca, modelo y año del vehículo. Las imágenes están organizadas en carpetas según la clase correspondiente, lo que facilita su uso en modelos de clasificación supervisada. Además, presentan una diversidad significativa en cuanto a ángulos de captura y condiciones de iluminación, y fueron tomadas en contextos reales, lo cual contribuye a la representatividad y robustez del conjunto de datos.

Inicialmente, el dataset se encontraba dividido en partes iguales para *train* y *test* (50% cada una). No obstante, esta partición fue considerada subóptima para los fines del presente estudio. Por ello, se optó por reagrupar todas las imágenes

y realizar una nueva división: 70% para entrenamiento, 10% para validación y 20% para test. Esta redistribución permite aprovechar mejor los datos disponibles y garantiza un conjunto de validación dedicado durante el proceso de entrenamiento, lo que favorece la detección de sobreajuste y mejora la capacidad de generalización del modelo.

Adicionalmente, se realizó una reclasificación de las clases. Dado que la versión original distingue entre modelos específicos de vehículos, se decidió agrupar las imágenes únicamente por marca del fabricante. Esta transformación redujo considerablemente el número de clases y permitió orientar el análisis hacia la identificación de patrones visuales característicos asociados a cada empresa automotriz. El objetivo principal de esta modificación fue explorar si existen correlaciones morfológicas consistentes entre el diseño de los vehículos y la identidad de la compañía que los produce.

TABLE I
CLASES RECLASIFICADAS POR MARCA DEL FABRICANTE

Índice	Marca
1	AM
2	Acura
3	Aston
4	Audi
5	BMW
6	Bentley
7	Bugatti
8	Buick
9	Cadillac
10	Chevrolet
11	Chrysler
12	Daewoo
13	Dodge
14	Eagle
15	FIAT
16	Ferrari
17	Fisker
18	Ford
	...

Es importante señalar que el conjunto de datos presenta un desbalance en la cantidad de imágenes por clase. Al analizar la distribución, se observó que algunas marcas están representadas con muchas más instancias que otras, lo cual puede afectar el desempeño del modelo, especialmente en clases minoritarias. En particular, la clase mayoritaria corresponde a la marca *Chevrolet*, que concentra un 11.12% de las imágenes. Este tipo de desequilibrio plantea desafíos adicionales durante el entrenamiento, como el riesgo de que el modelo tienda a

favorecer las clases más frecuentes.

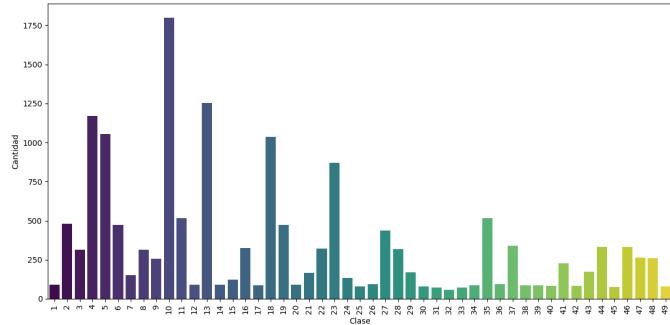


Fig. 1. Distribución de imágenes por marca en el conjunto de datos. Se observa un claro desbalance, con Chevrolet como clase mayoritaria.

Por último, mostramos una selección de imágenes de cada clase, y tener una idea de cómo son las imágenes y qué tipo de datos vamos a estar manejando.



Fig. 2. Imagen del conjunto de entrenamiento



Fig. 3. Imagen del conjunto de prueba

III. DISEÑO EXPERIMENTAL

Para abordar el problema de clasificación de marcas de autos, diseñamos una serie de experimentos comparando diferentes arquitecturas y estrategias de entrenamiento. El objetivo es analizar cómo influyen factores como la complejidad del modelo, el uso de pesos preentrenados y las técnicas de ajuste (fine-tuning) en la precisión y capacidad de generalización.

En total, se entrenaron cuatro modelos, cubriendo desde arquitecturas simples desarrolladas desde cero hasta redes profundas preentrenadas con distintas variantes de ajuste. Se entrenaron dos ResNet18, una AlexNet, una ResNet50 y una CNN. Y se usaron los siguientes hiperparámetros:

Aunque estos puedan parecer repetitivos, fueron los que mejores resultados dieron teniendo en cuenta precisión y eficiencia.

TABLE II
HIPERPARÁMETROS DE LOS MODELOS ENTRENADOS

Modelo	Learning Rate	Batch Size	Epochs
AlexNet	0.001	64	20
ResNet18	0.001	64	20
ResNet18	0.001	64	50
ResNet50	0.001	64	17
CNN	0.001	32	20

Para evaluar y comparar los modelos entrenados, se aplicaron diversos métodos de análisis tanto cuantitativos como cualitativos. El objetivo es comprender no solo el rendimiento global, sino también los tipos de errores y el comportamiento interno de cada modelo. Los enfoques de análisis utilizados fueron:

- **Visualización de ejemplos:** Se muestran imágenes del conjunto de validación que fueron clasificadas correctamente e incorrectamente, incluyendo ejemplos seleccionados manualmente y al azar, para ilustrar tanto casos típicos como situaciones particulares.
- **Matriz de confusión:** Se presenta la matriz de confusión para visualizar la distribución de aciertos y errores entre las distintas clases, identificando patrones de confusión frecuentes.
- **Método de atribución (Grad-CAM):** Se utiliza Grad-CAM para interpretar visualmente qué regiones de las imágenes influyen más en las decisiones del modelo, proporcionando información sobre la atención del modelo durante la clasificación.
- **Visualización de características (t-sne):** Se aplica reducción de dimensionalidad mediante t-sne sobre las características extraídas por la red antes de la capa de clasificación, permitiendo observar la separación entre clases en el espacio de características.

Estos análisis permiten identificar fortalezas y debilidades de cada modelo, facilitando la selección de la mejor arquitectura para su evaluación final sobre el conjunto de test.

IV. RESULTADOS Y DISCUSIÓN

1) *Curvas de entrenamiento:* Los resultados obtenidos fueron satisfactorios en términos de precisión y generalización. Los modelos entrenados lograron identificar correctamente la marca del vehículo en la mayoría de los casos, alcanzando valores de *accuracy* superiores al 80% en los conjuntos de validación y prueba. Estos resultados sugieren que existen características visuales distinguibles asociadas a cada fabricante, las cuales pueden ser captadas de manera efectiva por modelos convolucionales.

Como podemos ver en los procesos de entrenamiento de los distintos modelos aunque AlexNet tiene más parámetros (61M) que ResNet18 (11M) y ResNet50 (25M), su rendimiento fue inferior. Esto se debe a que las ResNet, al ser arquitecturas más modernas, incorporan conexiones residuales que facilitan el entrenamiento de redes profundas y evitan problemas como el desvanecimiento del gradiente. Gracias a esto, logran una mejor generalización con menos parámetros.

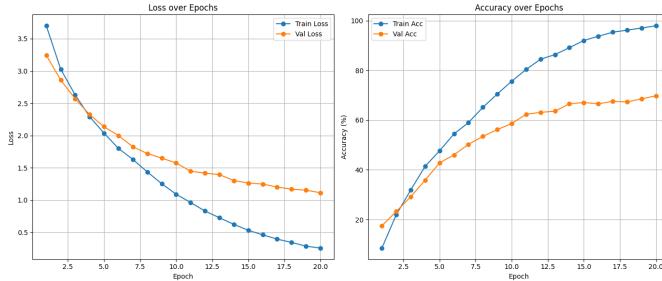


Fig. 4. Desarrollo del entrenamiento para ResNet18 con 20 EPOCHS.

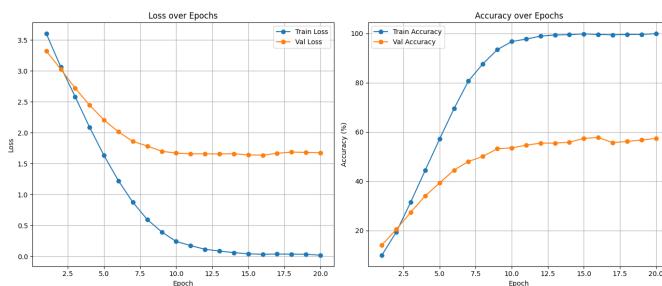


Fig. 5. Desarrollo del entrenamiento para ResNet 50 con 20 EPOCHS.

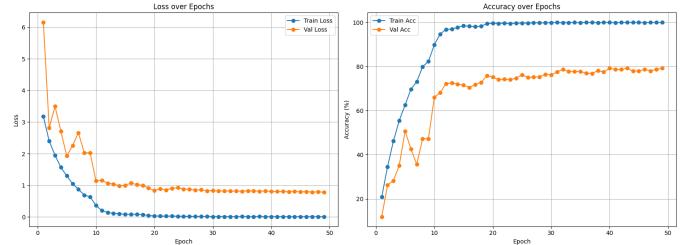


Fig. 6. Desarrollo del entrenamiento para ResNet18 con 50 EPOCHS.

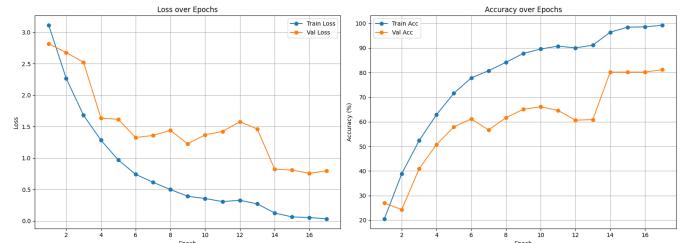


Fig. 7. Desarrollo del entrenamiento para ResNet50 con 17 EPOCHS.

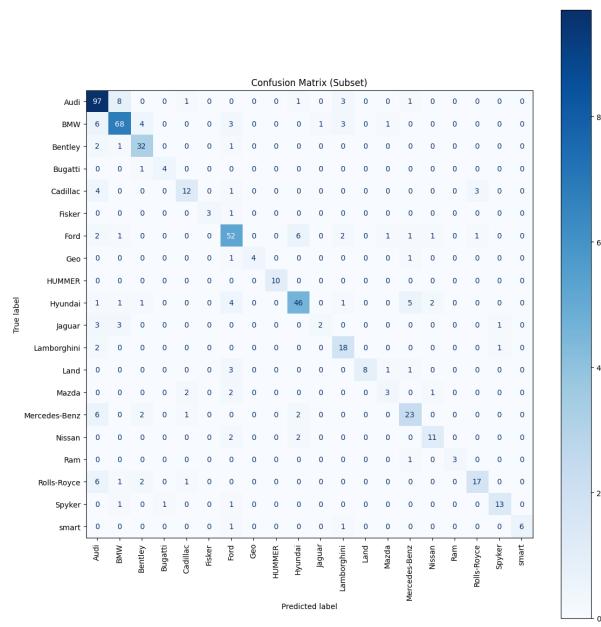


Fig. 8. Matriz de confusión reducida de la ResNet18 con 20 EPOCHS.

Las matrices de confusión permiten visualizar de forma clara el comportamiento del modelo y evidencian el desbalance existente en el conjunto de datos, donde algunas clases cuentan con una cantidad significativamente mayor de muestras que otras. Este desbalance se traduce en una distribución desigual de errores, especialmente en las clases minoritarias, que presentan mayores tasas de falsos negativos.

A pesar de estas limitaciones, los resultados obtenidos son positivos: se observa una diagonal principal bien definida

en las matrices, lo que indica una alta proporción de verdaderos positivos (aciertos) para varias clases. Esta concentración sugiere que el modelo logra capturar correctamente patrones visuales representativos de cada marca, incluso bajo condiciones de distribución de clases no homogénea. Estos resultados refuerzan la solidez del enfoque utilizado y abren la posibilidad de mejorar aún más el rendimiento mediante técnicas específicas de balanceo o ajustes adicionales en la arquitectura y los datos.

En la Tabla III se presentan las métricas obtenidas por los distintos modelos evaluados, incluyendo la precisión (*accuracy*) del set de validacion, y la precisión en el test set. Se observa que los modelos basados en aprendizaje por transferencia superan ampliamente al modelo entrenado desde cero.

TABLE III
COMPARACIÓN DE MÉTRICAS ENTRE DISTINTOS MODELOS

Modelo	Accuracy Train (%)	Accuracy Validation subset
ResNet18 (20 EPOCHS)	97.95	69.80
ResNet18 (50 EPOCHS)	99.96	79.21
AlexNet (20 EPOCHS)	85.70	41.83
ResNet50 (17 EPOCHS)	99.08	81.19
CNN (a mano)	10.81	12.13

2) *Métricas generales:* Ahora que ya analizamos las curvas de entrenamiento y validación para los distintos modelos, nos interesa profundizar en métricas cuantitativas más generales que resuman el desempeño de cada arquitectura sobre el conjunto de prueba. Estas métricas —*accuracy*, *precision*, *recall* y *F1-score*— permiten evaluar no solo la proporción de aciertos, sino también la capacidad de los modelos para generalizar y mantener un buen equilibrio entre falsos positivos y falsos negativos. La Tabla IV resume estos valores para cada uno de los modelos evaluados.

TABLE IV
COMPARACIÓN DE MÉTRICAS POR MODELO (TRANSPUESTA) PARA EL CONJUNTO DE PRUEBA

Modelo	Accuracy	Precision	Recall	F1-score
ResNet18	0.78	0.79	0.78	0.78
ResNet50	0.82	0.83	0.82	0.82
AlexNet	0.42	0.44	0.42	0.42
SimpleCNN	0.11	0.03	0.11	0.03

A partir de la Tabla IV, se observa que **ResNet50** es el modelo con mejor desempeño general, alcanzando las métricas más altas en todas las categorías: *accuracy*, *precision*, *recall* y *F1-score*, todas con un valor de 0.82 o superior. Esto sugiere que acierta con alta frecuencia, y que mantiene un buen equilibrio entre clases. Le sigue **ResNet18**, con valores ligeramente inferiores pero consistentes en todas las métricas, lo que lo posiciona como una alternativa más ligera con un rendimiento competitivo. Por otro lado, **AlexNet** presenta un desempeño notablemente inferior, con valores en torno a 0.42, lo que indica una menor capacidad de generalización. Finalmente, **SimpleCNN** muestra un rendimiento claramente deficiente en todas las métricas, con resultados que no superan el 0.07, evidenciando que este modelo no es adecuado para la tarea evaluada.

3) *Análisis cualitativo: Grad-CAM:* Esta parte es un análisis mas enfocado a como son los resultados de las predicciones en si mas alla de si están bien o mal, sino que queremos ver como se comporta el modelo en las diferentes clases, y como se comporta en las diferentes imágenes.

Grad-CAM (Gradient-weighted Class Activation Mapping) es una técnica de interpretación visual que permite entender qué regiones de una imagen influyen más en la predicción de un modelo de redes neuronales convolucionales. A través de un mapa de calor superpuesto a la imagen original, se destacan las zonas que activaron con mayor intensidad las capas profundas al momento de tomar la decisión. Las regiones más cálidas (en tonos rojos) indican mayor relevancia, mientras que las más frías (en azul) son menos influyentes¹². Esta visualización resulta útil para evaluar si los modelos están tomando decisiones basadas en características coherentes o aleatorias, y para comparar el comportamiento interno de distintas arquitecturas.

En las visualizaciones obtenidas, se identifican diferencias notables entre los modelos. Por ejemplo, en imágenes de autos fotografiados desde atrás, los mejores modelos (ResNet18 y ResNet50) tienden a fallar, mientras que modelos más simples como AlexNet o SimpleCNN logran acertar en algunos casos. Esto sugiere que las ResNet podrían tener dificultades cuando la vista del objeto cambia significativamente.

¹Por un error en la superposición de la imagen y el mapa de calor, los colores aparecen invertidos: el azul indica las regiones de mayor interés para el modelo y el rojo, las de menor relevancia.

²También existe la posibilidad de que los mapas de gradiente varíen dependiendo de la resolución de la imagen, en este caso, para poder adaptar al modelo de la simpleCNN al plot, todos los modelos tuvieron que usar una resolución de 128x128 en sus imágenes, lo que genera discrepancia con sus gradientes y visualizaciones t-sne individuales

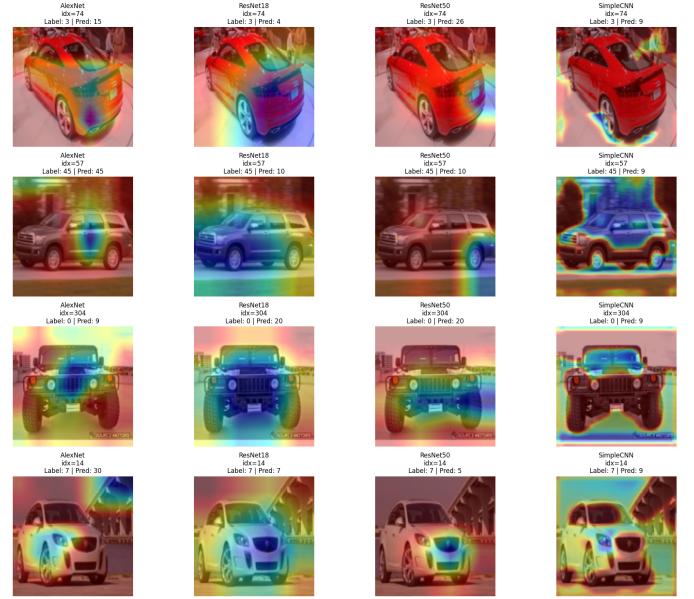


Fig. 9. Grad-Cam aplicado a distintas fotos para cada modelo

Por otro lado, al comparar cómo enfocan los modelos las regiones del vehículo, se observa que ResNet18 tiende a analizar el auto como un todo, distribuyendo su atención de manera amplia, mientras que ResNet50 se concentra más en regiones específicas, como la parte baja del vehículo. AlexNet, en cambio, muestra un patrón de atención más inconsistente entre ejemplos, lo que dificulta extraer una conclusión clara sobre su estrategia de decisión.

El modelo SimpleCNN, dado su bajo desempeño general, no logra identificar patrones relevantes y no aporta observaciones valiosas al análisis interpretativo. Se enfoca gran parte en características externas al vehículo, por lo que sus predicciones no nos son de utilidad.

4) *Análisis cualitativo: t-SNE:* Finalmente, realizamos un análisis cualitativo utilizando t-SNE con el objetivo de visualizar cómo se distribuyen las características aprendidas por los modelos en un espacio bidimensional. Esta técnica de reducción de dimensionalidad nos permite observar si las representaciones de las distintas clases forman agrupamientos definidos (clusters) o si, por el contrario, las clases tienden a mezclarse entre sí, lo que indicaría dificultades del modelo para separarlas.

En lugar de mostrar todos los modelos, nos enfocamos en dos casos contrastantes: ResNet50, que obtuvo el mejor desempeño, y AlexNet, uno de los modelos con resultados más bajos.

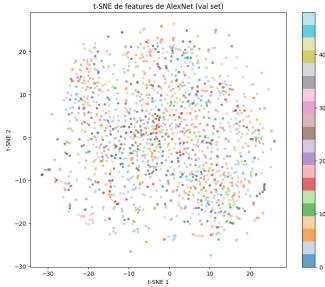


Fig. 10. Visualización de T-sne para AlexNet

AlexNet. Las representaciones de AlexNet se muestran desordenadas y dispersas en todo el espacio. No se aprecian clusters definidos, y las clases parecen solaparse en gran medida. Esto indica que el modelo no logró aprender características suficientemente discriminativas para diferenciar adecuadamente entre clases, lo cual concuerda con su bajo desempeño cuantitativo.

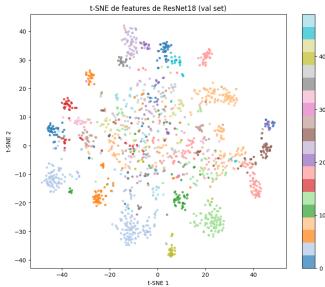


Fig. 11. Visualización de T-sne para ResNet50

ResNet50. La visualización muestra dos comportamientos claros. Por un lado, se observan grupos de características que se agrupan de forma compacta y bien separada del resto, lejos del origen de coordenadas, con puntos cercanos a un centro definido y sin interferencia visible de otras clases. Esto indica que el modelo aprendió representaciones altamente discriminativas para esas clases. Por otro lado, existen regiones cercanas al origen donde los puntos aparecen más dispersos y las clases se superponen, lo que sugiere que el modelo aún tiene dificultad para separar completamente ciertas categorías.

5) *Elección de modelo y resultado final:* Luego de todo el análisis realizado, se concluyó que el modelo más adecuado para esta tarea era **ResNet50**, dado su desempeño superior tanto en métricas cuantitativas como en interpretabilidad visual. Por esta razón, se lo seleccionó como modelo final y se evaluó sobre el conjunto de test para obtener una estimación realista de su capacidad de generalización.

Antes de eso, retomamos brevemente los resultados sobre el conjunto de validación, esta vez incluyendo el promedio *ponderado* por clase para tener en cuenta el desbalance del dataset:

TABLE V
MÉTRICAS RESNET50 PARA VALIDATION SET

	Precision	Recall	F1-Score
Accuracy			0.79
Macro Avg	0.80	0.78	0.76
Weighted Avg	0.82	0.79	0.79

Como puede verse, no se observan diferencias significativas entre los promedios macro y ponderado (variación de aproximadamente ± 0.02), lo cual indica una performance relativamente consistente entre clases, con cierto margen de mejora para clases minoritarias.

Pasando al conjunto de test —el más importante para evaluar generalización— se obtuvieron los siguientes resultados:

TABLE VI
MÉTRICAS RESNET50 PARA TEST SET COMPLETO

	Precision	Recall	F1-Score
Accuracy			0.80
Macro Avg	0.80	0.74	0.76
Weighted Avg	0.81	0.80	0.80

Los valores se mantienen estables respecto al conjunto de validación, con una leve baja en el *recall* promedio, lo cual puede atribuirse a la dificultad natural del conjunto de test. En términos generales, los resultados siguen siendo sólidos y dentro de un rango aceptable.

Complementariamente, las matrices de confusión refuerzan esta conclusión. No se observan confusiones sistemáticas entre clases específicas, aunque pueden existir algunos errores puntuales entre marcas similares como *Ford* y *Dodge* o *Mercedes-Benz* y *Dodge*. Sin embargo, la diagonal dominante en ambas matrices indica que, en la mayoría de los casos, el modelo logra predecir correctamente la clase correspondiente.

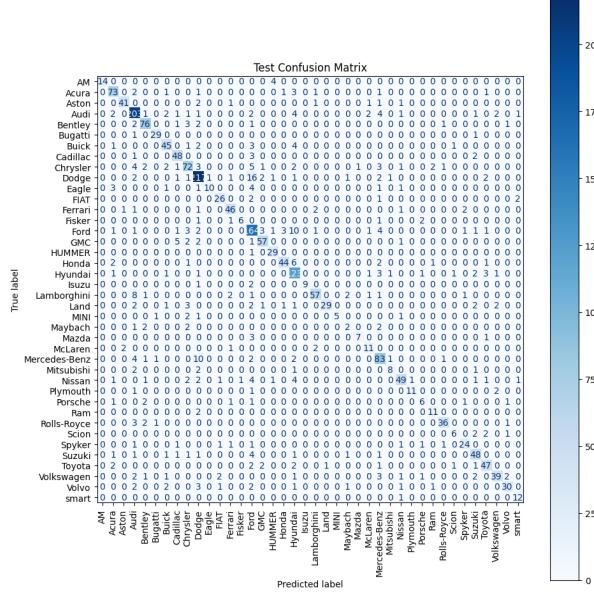


Fig. 12. Matriz de confusión de la ResNet18 con 50 EPOCHS para el set de test.

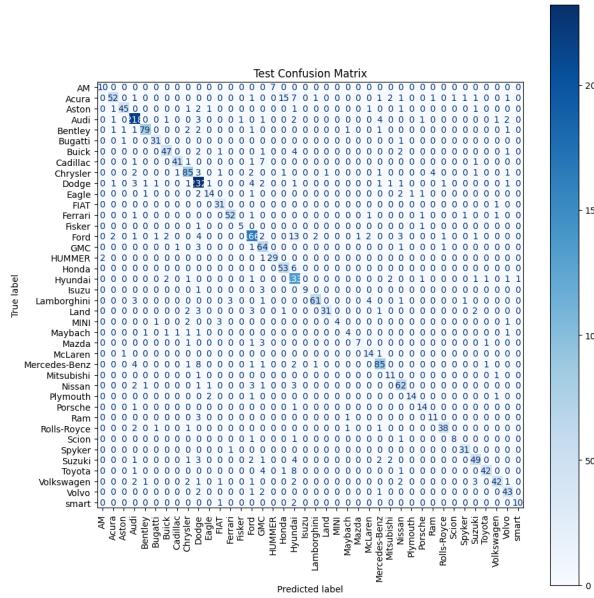


Fig. 13. Matriz de confusión de la ResNet50 con 17 EPOCHS para el set de test.

V. CONCLUSIÓN

A lo largo de este trabajo se evaluaron distintas arquitecturas de redes neuronales convolucionales aplicadas a la tarea de clasificación de fabricantes de vehículos. Se analizó tanto el desempeño cuantitativo mediante métricas estándar como el comportamiento cualitativo a través de técnicas interpretables como Grad-CAM y t-SNE.

Los resultados obtenidos confirman la efectividad del enfoque basado en aprendizaje por transferencia, con **ResNet50** destacándose como el modelo más robusto y preciso. Este

modelo no solo alcanzó las mejores métricas en validación y test, sino que también demostró una atención más focalizada y representaciones discriminativas en el espacio latente.

Aunque existen ciertos desafíos relacionados con el desbalance de clases y la dificultad para diferenciar algunas marcas visualmente similares, los modelos entrenados lograron generalizar correctamente en la mayoría de los casos. Estas observaciones abren la puerta a futuras mejoras mediante técnicas de balanceo de datos, aumento de imágenes o ajustes finos de arquitectura.

En conjunto, el trabajo demuestra que es posible abordar con éxito la clasificación de marcas de automóviles a partir de imágenes, y ofrece una base sólida para futuras aplicaciones en tareas más complejas o entornos de producción.

VI. TRABAJO A FUTURO

Debido a las limitaciones computacionales, el presente trabajo utilizó únicamente el 25% del conjunto de datos disponible. Como línea futura, se propone entrenar los modelos utilizando la totalidad del dataset para mejorar su capacidad de generalización y robustez. Asimismo, se recomienda gestionar créditos para acceder a recursos con aceleración por GPU mediante plataformas como Google Colab Pro o Google Cloud Platform, lo que permitiría realizar entrenamientos más extensos, rápidos y eficientes.

Además, dado el desbalance en la cantidad de imágenes por clase observado en el dataset, se considera importante aplicar técnicas de balanceo de datos. Entre las opciones posibles se incluyen el sobremuestreo de clases minoritarias (oversampling), la generación de imágenes sintéticas mediante data augmentation o técnicas como SMOTE, y la ponderación de la función de pérdida según la frecuencia de clases.

Por último, se sugiere explorar arquitecturas más profundas y modernas, como los Vision Transformers (ViT), que han demostrado un rendimiento competitivo en tareas de clasificación de imágenes. Su incorporación podría permitir mejoras adicionales en la precisión y en la capacidad del modelo para capturar patrones visuales complejos.

REFERENCES

- [1] Jutrera, J. (2021). *Stanford Car Dataset by Classes Folder*. Kaggle. Disponible en: <https://www.kaggle.com/datasets/jutrenera/stanford-car-dataset-by-classes-folder>
 - [2] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Retrieved from <https://scikit-learn.org/stable/> (Accessed: 2024-12-11).
 - [3] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. Retrieved from <https://pytorch.org> (Accessed: 2024-12-11).
 - [4] Chollet, F. et al. (2015). *Keras: Deep Learning for humans*. Retrieved from <https://keras.io> (Accessed: 2024-12-11).
 - [5] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.