Prácticas: Image classification with transformers vs. CNN

Ana Valentina López Chacón

Visión por Computadora
MIARFID, UPV

Mayo, 2025

1. Objetivos

Comparar el desempeño de modelos basados en Transformers de Visión y Redes Neuronales Convolucionales (CNN) para la clasificación de imágenes. Específicamente, se busca:

- Ajustar modelos CNN (ResNet50, EfficientNet) y Transformers (ViT16, Swin Transformer) en el conjunto de datos Flowers-102 [1].
- Evaluar el efecto de hiperparámetros como tasa de aprendizaje, tamaño de lote, técnicas de aumento de datos, etc.
- Comparar métricas de precisión, velocidad de entrenamiento y tiempo de inferencia entre los modelos.
- Analizar la conveniencia de cada arquitectura bajo restricciones computacionales.

2. Estado del Arte

Las arquitecturas CNN, como ResNet50 [2] y EfficientNet [3], han sido pilares en tareas de visión por computador debido a su capacidad para capturar características locales y su eficiencia en entrenamiento. EfficientNet destaca por su escalamiento compuesto que optimiza precisión y tamaño del modelo. Por otro lado, los Transformers de Visión (ViT) [4] han revolucionado el campo al aplicar mecanismos de atención auto-regresiva, capturando dependencias globales en las imágenes. Variantes como el Swin Transformer [5] introducen atención local jerárquica que mejora el rendimiento en tareas de clasificación y segmentación.

En trabajos recientes, proponen el *Internal Ensemble Learning Transformer* (IELT) [6], un modelo basado en Vision Transformers que mejora la clasificación fina al combinar mecanismos de atención multi-cabeza y refinamiento cruzado de capas. En Oxford Flowers-102, este método alcanza una precisión superior al 99.6 %. Por

otra parte, en [7] presentan SpinalNet, una arquitectura inspirada en la fisiología humana que optimiza redes neuronales profundas y obtiene resultados competitivos en tareas de clasificación de imágenes, alcanzando precisiones cercanas al 99.3 % en datasets similares. Ambos enfoques evidencian el avance hacia modelos más efectivos y eficientes para clasificación fina, siendo especialmente destacables los resultados superiores que los Transformers han demostrado en conjuntos de datos como Oxford Flowers-102.

3. Dataset y Preprocesamiento

El conjunto de datos Oxford Flowers-102 consta de un total de 8,189 imágenes distribuidas en 102 clases. Para el entrenamiento se utilizan las imágenes indicadas en la partición oficial, con 6,149 muestras para entrenamiento, 1,020 para validación y 2,020 para prueba. Las imágenes tienen resolución variable, por lo que se aplicaron preprocesamientos para adecuarlas a la entrada de los modelos. El aumento de datos inicial (DA Inicial) aplicado para el conjunto de entrenamiento incluye las siguientes transformaciones con el objetivo de mejorar la generalización y evitar sobreajuste:

- `RandomResizedCrop` con tamaño 224×224 y escala entre 0.8 y 1.0
- `RandomHorizontalFlip` con probabilidad 0.5
- `RandomRotation` con un ángulo máximo de 15°
- `ColorJitter` para variar brillo y contraste en 0.2
- Conversión a tensor con `ToTensor()`
- Normalización con media `[0.485, 0.456, 0.406]` y desviación estándar `[0.229, 0.224, 0.225]` (valores estándar de ImageNet)

Para los conjuntos de validación y prueba se aplicó un preprocesamiento más simple, consistente en:

- Redimensionamiento a 224×224 píxeles con `Resize`
- Conversión a tensor y normalización con los mismos parámetros indicados

Las imágenes se cargan desde el directorio `oxford102/jpg/jpg` mediante la clase `FlowersDataset` definida específicamente para este conjunto. A continuación se muestra la distribución de las imágenes por clase en cada partición (entrenamiento, validación y prueba):

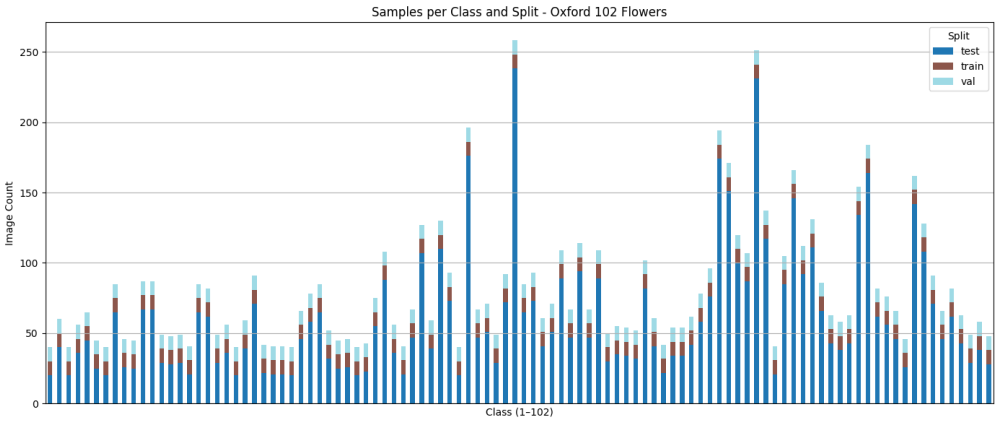

Figura 1: Distribución del número de muestras por clase y partición en Oxford Flowers-102.

4. Configuraciones Experimentales

Se realizaron cinco configuraciones experimentales variando parámetros clave de entrenamiento:

Ex	Optimizador	Épocas	DA
1	AdamW, lr=0,0001	10	Sin DA
2	AdamW, lr=0,0002	10	Sin DA
3	AdamW, lr óptimo por modelo	10	DA Inicial
4	AdamW, lr óptimo por modelo	30	DA Inicial + LabelSmoothing
5	AdamW, lr óptimo por modelo	30	DA Inicial+ LabelSmoothing + RandAug

Cuadro 1: Configuraciones experimentales aplicadas a los modelos.

Estas configuraciones fueron implementadas en las cinco arquitecturas evaluadas: ResNet50, EfficientNet, ViT16, ViT32 y Swin Transformer. Se aplicó el mismo *scheduler* en todos los casos como `ReduceLROnPlateau`, se incluyeron otras opciones en experimentos preliminares pero no generaron mejora, de igual forma el tamaño de lote se fijó en 32 dado que un valor mayor representaba problemas de memoria y la tasa de aprendizaje óptima se seleccionó entre los experimentos 1 y 2 por modelo y se mantuvo para los siguientes experimentos.

5. Resultados

En la Tabla 2 se resumen los resultados de precisión en el conjunto de prueba para cada experimento y modelo y en la Tabla 3 los tiempos de entrenamiento y tiempos totales de inferencia específicos para cada modelo. Para abreviar y optimizar la visualización, los nombres de los modelos en la tabla son:

- **RN50** para ResNet50

- **EffNet** para EfficientNet (versión seleccionada)
- **ViT16** para Vision Transformer patch size 16
- **ViT32** para Vision Transformer patch size 32
- **Swin** para Swin Transformer base

Ex	RN50 (%)	EffNet (%)	ViT16 (%)	ViT32 (%)	Swin (%)
1	85.22	85.75	95.98	94.41	98.96
2	86.83	88.96	96.88	93.67	97.98
3	88.66	91.01	97.53	96.41	98.37
4	89.71	91.85	97.92	97.09	98.75
5	90.26	92.05	97.67	97.80	**99.12**

Cuadro 2: Precisión en conjunto de prueba por experimento y modelo.

Ex	RN50		EffNet		ViT16		ViT32		Swin	
	Train	Inf.	Train	Inf.	Train	Inf.	Train	Inf.	Train	Inf.
1	150.12	22.91	364.20	48.57	493.01	75.42	146.01	24.39	487.96	75.84
2	140.41	19.94	334.33	45.85	522.65	80.72	127.36	18.94	502.59	79.46
3	145.07	19.98	308.84	41.12	505.33	77.26	145.15	21.77	537.44	85.09
4	496.55	22.86	1128.76	50.38	1663.84	84.25	423.9	20.99	1467.66	75.86
5	463.32	21.22	904.07	39.63	1450.63	73.96	440.30	22.02	1689.60	87.28

Cuadro 3: Tiempos de entrenamiento e inferencia total en segundos para cada modelo y experimento.

Los resultados obtenidos demuestran que, para el conjunto Flowers-102, las arquitecturas basadas en Transformers de Visión superan consistentemente en precisión a las redes convolucionales clásicas como ResNet50 y EfficientNet. En particular, el modelo Swin Transformer alcanzó la mayor precisión con un 99.12 % en el experimento 5, lo que representa una mejora significativa frente a los mejores resultados de CNN, con EfficientNet alcanzando un 92.05 % y ResNet50 un 90.26 %.

6. Conclusiones

A partir de los resultados obtenidos es posible destacar las siguientes conclusiones:

- Se observo también que el aumento progresivo en la complejidad de las configuraciones experimentales, desde tasas de aprendizaje óptimas, pasando por aumentos de datos iniciales, hasta la incorporación de técnicas avanzadas como label smoothing y RandAugment, contribuyó a mejorar la generalización de los modelos y a reducir el sobreajuste, reflejado en la mejora sostenida de la precisión en validación y prueba.
- Entre los Transformers, ViT32 mostró menor precisión y tiempos ligeramente menores de inferencia, reflejando el compromiso entre tamaño de patch y

desempeño, mientras que ViT16 y Swin presentaron mejores resultados con tiempos mayores.

- Estos hallazgos son coherentes con la literatura actual, donde los Transformers, y particularmente los modelos híbridos y jerárquicos como Swin, son cada vez más utilizados para tareas de clasificación fina, superando las limitaciones tradicionales de las CNNs para capturar relaciones globales en imágenes.

Referencias

[1] Nilsback, M. E., & Zisserman, A. (2008, December). *Automated flower classification over a large number of classes*. In 2008 Sixth Indian conference on computer vision, graphics & image processing (pp. 722-729). IEEE

[2] He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition*. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[3] Tan, M., & Le, Q. (2019, May). *Efficientnet: Rethinking model scaling for convolutional neural networks*. In International conference on machine learning (pp. 6105-6114). PMLR.

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale*. arXiv preprint arXiv:2010.11929.

[5] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). *Swin transformer: Hierarchical vision transformer using shifted windows*. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012-10022).

[6] Xu, Q., Wang, J., Jiang, B., & Luo, B. (2023). *Fine-grained visual classification via internal ensemble learning transformer*. IEEE Transactions on Multimedia, 25, 9015-9028.

[7] Kabir, H. D., Abdar, M., Khosravi, A., Jalali, S. M. J., Atiya, A. F., Nahavandi, S., & Srinivasan, D. (2022). *Spinalnet: Deep neural network with gradual input*. IEEE Transactions on Artificial Intelligence, 4(5), 1165-1177.