Transformers Unidos: Eficacia De los Modelos Ensemble-ViT en Clasificación Automática de Flora Costarricense 🌿

Transformers United: Effectiveness of Ensemble-ViT Models in Automatic Classification of Costa Rican Flora 🌿

Anthonny Adrián Badilla Olivas [1]

Estudiante de Computación con énfasis en Ciencias de la Computación, UCR

e-mail: anthonny.badilla@ucr.ac.cr

Enrique Guillermo Vílchez Lizano [2]

Estudiante de Computación con énfasis en Ciencias de la Computación, UCR

e-mail: enrique.vilchezlizano@ucr.ac.cr

Rubén González Villanueva [3]

Docente de la Escuela de Ciencias de la Computación e Informática, UCR

e-mail: ruben.gonzalezvillanueva@ucr.ac.cr

Resumen:

En clasificación automática de flora, es común enfrentarse a conjuntos de datos no balanceados. El articulo explora la efectividad de los modelos de ensemble de Transformers de Visión en comparación a un modelo de ensemble con redes convolucionales para el reconocimiento de flora de Costa Rica. Se observan tendencias prometedoras en esta tarea incluso especies en peligro que poseen pocas muestras. El potencial de estas debe estudiarse más a fondo para determinar si nuestras conclusiones pueden extrapolarse a otros conjuntos de datos no balanceados. El código de los experimentos está disponible en https://github.com/Antonio-Tresol/vits_ensemble_cr_leaves .

Palabras clave: Clasificación de plantas, Aprendizaje Automático Profundo, Visión por Computadora.

Abstract:

In automatic flora classification, it is common to face unbalanced data sets. The article explores the effectiveness of Vision Transformers ensemble models in comparison to an ensemble model with convolutional networks for the recognition of Costa Rican flora. Promising trends are observed in this task, including endangered species that have few samples. The potential of these

1 c.c. (aka) Antonio Badilla-Olivas
2. c.c. (aka) Enrique Vílchez-Lizano
3 c.c. (aka) Rubén González-Villanueva

should be studied further to determine whether our conclusions can be extrapolated to other unbalanced data sets. The code of the experiments is available at https://github.com/Antonio-Tresol/vits_ensemble_cr_leaves .

Keywords: Plant Classification, Deep Machine Learning, Computer Vision.

I. INTRODUCCIÓN

La clasificación de plantas es crucial para la conservación de la biodiversidad ante peligros ambientales [1]. El uso de técnicas avanzadas de clasificación de datos desequilibrados ha sido clave para optimizar la gestión de recursos y desarrollar modelos predictivos que ayuden en la protección de la flora en su entorno. Los sistemas de reconocimiento automático han ganado popularidad al reducir costos y permitir la participación de personas no profesionales en esta tarea, así como reducir el error humano [2]. Esto es posible a través de aplicaciones informáticas que permiten recopilar datos, como fotografías tomadas incluso desde teléfonos celulares [3]. En este contexto, el artículo compara la eficacia de modelos de clasificación automática basados en redes convolucionales y Transformers de visión, destacando la importancia de la innovación tecnológica en la conservación de la biodiversidad.

Este trabajo sigue la línea de [4], quienes encontraron viable el uso de modelos ensemble de redes convolucionales para clasificación multiclase de plantas con conjuntos de datos no balanceados, específicamente en el conjunto de datos CR Leaves [1]. Se pretende explorar los Transformers de visión, los cuales han demostrado mejores resultados que los modelos convolucionales en tareas de clasificación [5], [6], [7]. Concretamente, modelos Transformers de visión ensemble.

A. Trabajo relacionado

Las tareas de clasificación en datos no balanceados han sido ampliamente estudiadas, con enfoques como el sobremuestreo y submuestreo [8], [9], [10], así como modelos de ensemble [4], [11]. En clasificación de plantas, se han utilizado redes convolucionales profundas [12], Transformers de visión [3] e incluso ensembles de ambos [13]. Modelos ensemble basados en Transformers de visión se han explorado en otras tareas de clasificación [14], [15], pero no parecen haber sido suficientemente estudiados para clasificación de plantas en datos no balanceados, considerando métricas y técnicas apropiadas para estos problemas, así como parámetros preentrenados congelados.

II. METODOLOGÍA

El experimento trabajo sobre el conjunto de datos llamado CR Leaves [1]. Este contiene imágenes de hojas de plantas endémicas de Costa Rica, ordinarias y en peligro, cada una categorizada con su

respectivo nombre científico, que suman un total de 254 especies y 3813 especímenes. Como es de esperarse, aquellas especies en riesgo tienen menos muestras.

A. Modelos

Se evaluará la eficacia de dos **modelos ensemble** para el reconocimiento de plantas: E-Triple [4] conformado por redes convolucionales y Ensemble ViT por Transformers de visión. Ambos combinan las salidas de modelos mediante votación suave [13], y adaptan la última capa para la tarea específica mediante transferencia de aprendizaje [16]. El ensemble E-Triple combina ResNet50 [17], EfficientNetB4 [18] y ConvNext [19], siendo este último también utilizado como modelo individual de referencia. Por otro lado, Ensemble ViT combina ViT-Base-16, ViT-Base-32 y ViT-Large-32 [14] para aprovechar sus fortalezas e intentar mejorar la eficacia en la tarea. ViT-Base-16 [5], que divide la imagen en parches y los procesa con una arquitectura transformer, también se utiliza como modelo individual de referencia.

Para pasar las imágenes a los modelos, se aplicaron transformaciones previas. **Para los modelos convolucionales,** las imágenes se cambiaron a un tamaño de 232x232 píxeles, como en [4], utilizando interpolación bilineal. Se recortaron y se normalizó cada canal de color utilizando los valores específicos de promedios y desviaciones estándar de ImageNet [20]. **Para los modelos transformer de visión,** se redimensionaron a 256x256 píxeles [5], mediante interpolación bilineal. Luego, se recortaron y se normalizaron cada canal de color utilizando los valores de promedios y desviaciones estándar de ImageNet [20], con la excepción de que se le modificó el primer canal de color a un promedio de 0, pues con este se obtuvo mejores resultados, dada exploración preliminar de los modelos ViT.

B. Métricas

Es común medir los modelos con la métrica de exactitud ordinaria, obteniendo una tasa de aciertos de la categoría predicha $\hat{y}$ respecto a la verdadera categoría y:

$$Exactitud\ (\hat{y}, y) = \frac{1}{N_{muestras}} \sum_{i=1}^{N_{muestras}} (\hat{y}_i == y) \quad (1)$$

No obstante, con datos no balanceados, esta métrica puede dar falsas expectativas del rendimiento [4]. Por ello, **el promedio macro de exactitud es una mejor opción para conjuntos no balanceados**, pues toma todas las predicciones $\hat{y}_c$, (cuya clase verdadera es c) y hace un promedio de la razón de aciertos de cada clase. En ocasiones se le conoce como exactitud balanceada [4].

$$Exactitud\ macro(\hat{y}, y) = \frac{1}{N_{clases}} \sum_{c=1}^{N_{clases}} Exactitud(\hat{y}_c, c) \quad (2)$$

C. Experimentos

Primero, para tomar las mediciones se evaluó el rendimiento de los modelos individuales y los ensembles mediante un entrenamiento individual de cada modelo base durante 30 épocas. Se utilizaron parámetros pre-entrenados en ImageNet 1K y se congelaron todas las capas excepto la de clasificación. Se realizaron 32 réplicas de entrenamiento para cada modelo individual, guardando el mejor checkpoint en cada réplica en función de la mayor exactitud balanceada en el conjunto de validación. Para la evaluación individual de cada modelo no ensemble, se midió la exactitud macro y micro en el conjunto de prueba al final de las 30 épocas de entrenamiento. Luego, se formaron ensembles combinando los modelos individuales. En cada réplica, se seleccionó el checkpoint correspondiente de cada modelo base y se combinaron para formar un ensemble. Se midió la exactitud macro y micro de cada ensemble en el conjunto de prueba. Por último, **los hiperpárametros del entrenamiento** se escogieron con base a los usados en [4], además de considerar para velocidad de aprendizaje con respecto a hiperpárametros y optimizadores [21], [22]. Se usó una tasa de aprendizaje de 0.0003, con un tamaño de batch de 64, un tamaño de prueba de 50%, ADAM [23] como optimizador, Cosine Annealing [24] para ajustar la tasa de aprendizaje, y *early stopping* con paciencia 3.

Para determinar si hay diferencias significativas en el rendimiento de los modelos, se realizaron las pruebas estadísticas de normalidad (Shapiro-Wilk) y de homogeneidad de varianzas (Levene). En caso de cumplirse los supuestos de normalidad y homogeneidad de varianzas, se realizó un ANOVA para determinar si existían diferencias significativas entre las medias de las exactitudes balanceadas de los ensambles. Si el ANOVA resultaba significativo, se aplicaba la prueba de Tukey HSD para identificar qué pares presentaban diferencias significativas en su rendimiento.

III. RESULTADOS

Los resultados del experimento revelaron diferencias significativas en el rendimiento de los modelos. Para la **exactitud básica** (no balanceada), los modelos ensemble superaron a los individuales. **E-Triple lideró con** una media de exactitud de **0.8167,** seguido de Ensemble ViT con 0.8079. ViT-Base-16 y ConvNext obtuvieron medias de exactitud de 0.7614 y 0.7465, respectivamente. Las mediciones sobre la **exactitud balanceada** indicaron la superioridad de los modelos ensemble, con **Ensemble ViT logrando la media más alta de 0.7879**, seguido de E-Triple con 0.7805. Los modelos individuales, ViT-Base-16 y ConvNext, obtuvieron medias de 0.7392 y 0.7129, respectivamente.

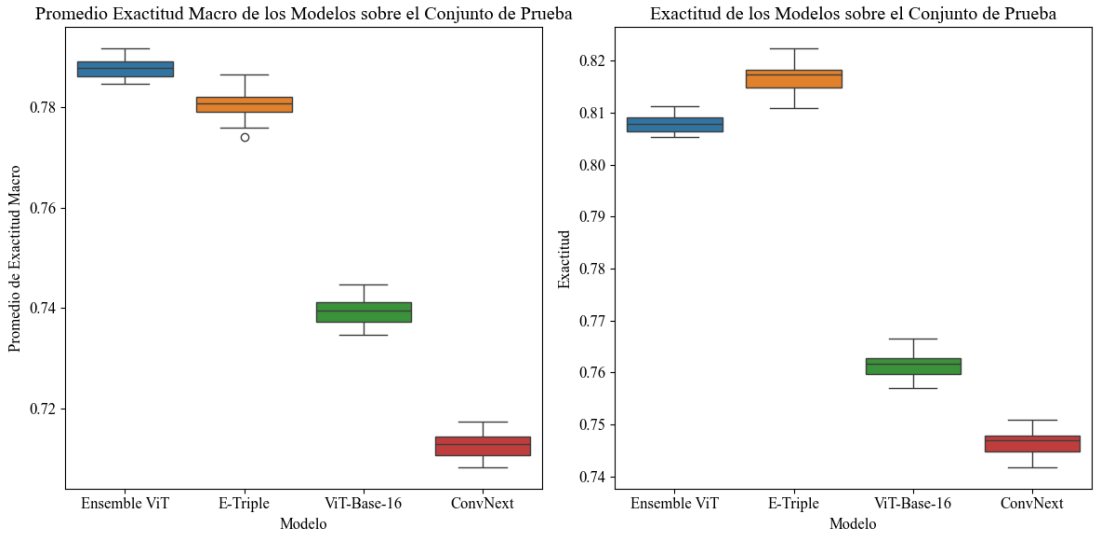

Fig. 1: Comparación de Exactitud y Promedio Exactitud Macro

Las pruebas de normalidad confirmaron que los datos de todos los modelos seguían una distribución normal, pues en ningún caso se rechaza la hipótesis nula (no se da que $p < 0.05$). Además, la prueba de Levene indicó que las varianzas eran iguales entre los grupos. Siendo así, posteriormente la prueba ANOVA reveló una diferencia significativa entre las medias de exactitud balanceada de los modelos. Para explorar donde se encontraban las diferencias se realizó la prueba de Tukey HSD la cual confirmó que todas las diferencias por pares entre los modelos eran estadísticamente significativas para un valor p ajustado < 0.02 (por las correcciones de valores p explicadas en [25]).

CUADRO I
RESULTADOS DE PRUEBAS ESTADÍSTICAS

	Shapiro - Wilk		**Levene**		**ANOVA**	
	Estadístico	**Valor P**	**Estadístico**	**Valor P**	**Estadístico**	**Valor P**
E Triple	0.9753	0.6566	0.7421	0.5289	7292.0922	3.24×10^{-139}
ViT Ensemble	0.9689	0.4697				
ViT-Base-16	0.9738	0.6129				
ConvNext	0.9731	0.5899				

En resumen, los modelos ensemble, E-Triple y Ensemble ViT, superaron significativamente a los modelos individuales, ViT-Base-16 y ConvNext, tanto en términos de exactitud básica como de exactitud balanceada en el conjunto de datos. Estos resultados destacan el potencial de los modelos ensemble para mejorar el rendimiento en la clasificación de especies de flora, especialmente cuando se enfrentan a desbalances en los datos.

IV. DISCUSIÓN

Dado que las todas las condiciones requeridas para las pruebas estadísticas de ANOVA y Tukey se cumplieron, los resultados obtenidos en la sección anterior tienen un fuerte grado de validez para el contexto específico en el que se están aplicando. Los resultados tienen implicaciones importantes en la aplicación práctica de modelos de aprendizaje automático en la clasificación de flora, especialmente cuando existen desbalances en los datos. Sugieren que los modelos de ensemble pueden ser una estrategia efectiva para mejorar el rendimiento, lo que podría ser útil en aplicaciones del mundo real.

A partir de este estudio, se pueden validar dos principales ideas. Primeramente, los modelos de ensemble pueden ser más efectivos que los modelos individuales. A pesar del gran rendimiento que los modelos individuales mostraron, los de ensemble superaron con creces sus resultados en la métrica de exactitud macro. Esto muestra el potencial de las estrategias de ensemble en tareas de clasificación de flora con datos no balanceados. Y segundo, el modelo de ensemble ViT mostró ser mejor en términos de exactitud macro que el modelo convolucional E-Triple. Las pruebas estadísticas respaldan la diferencia significativa observada en términos de medias de exactitud. Es así como, para datos no balanceados, los modelos de ensemble ViT pueden ser eficaces, en comparación con modelos de ensemble convolucionales. En términos individuales la arquitectura transformer también mostró ser mejor que su contraparte convolucional.

Además, a pesar de los resultados prometedores de este estudio, hay limitaciones que considerar. Primero, los experimentos realizados se hicieron con un único conjunto de datos, especializado en plantas de Costa Rica, por lo que sus resultados podrían no ser extrapolables. Además, los hiperparámetros de los modelos no necesariamente son los óptimos para el desempeño de estos, y tampoco el número de épocas. Se podría explorar el efecto de entrenar con configuraciones distintas. Así mismo, esta investigación solo toma en cuenta una de sus métricas para validar las hipótesis planteadas, pero se podrían extender para evaluar el rendimiento de los ensembles con diferentes métricas. También es necesario que se considere el impacto ambiental de entrenar y ejecutar modelos ensemble, lo cual al menos puede duplicar los recursos computacionales y energéticos necesarios para entrenar un modelo individual.

V. CONCLUSIONES

En conclusión, este estudio proporciona evidencia sólida de que los modelos de ensemble, especialmente aquellos basados en la arquitectura transformer como ViT, pueden ofrecer mejoras significativas en la clasificación de flora, especialmente en entornos con desbalances de datos. Los resultados respaldan la superioridad de los ensembles sobre los modelos individuales, así como la ventaja del enfoque transformer sobre las redes convolucionales en este contexto específico. Sin embargo, se

reconocen limitaciones importantes, como la falta de generalización debido al uso de un único conjunto de datos y la necesidad de explorar diferentes configuraciones de hiperparámetros. A pesar de ello, este estudio marca un avance importante en la aplicación práctica de modelos de aprendizaje automático en la clasificación de flora y sugiere áreas futuras de investigación para mejorar aún más la eficacia y eficiencia de estos modelos en el mundo real.

AGRADECIMIENTOS

Al Centro de Investigación en Tecnologías de la Información y Comunicación y los profesores Dr. Allan Berrocal Rojas y Dr. Ignacio Diaz Oreiro de la Escuela de Ciencias de la Computación e Informática.

REFERENCIAS

[1] E. Carranza-Rojas Jose AND Mata-Montero, «Combining Leaf Shape and Texture for Costa Rican Plant Species Identification», *CLEI Electronic Journal*, vol. 19, pp. 7-7, may 2016, [En línea]. Disponible en: http://www.scielo.edu.uy/scielo.php?script=sci_arttext&pid=S0717-50002016000100007&nrm=iso

[2] A. Karnan y R. Ragupathy, «A Comprehensive Study on Plant Classification Using Machine Learning Models», en *ICT: Smart Systems and Technologies*, M. S. Kaiser, J. Xie, y V. S. Rathore, Eds., Singapore: Springer Nature Singapore, 2024, pp. 187-199.

[3] J. Li y J. Yang, «Supervised Classification of Plant Image Based on Attention Mechanism», en *2021 7th International Conference on Systems and Informatics (ICSAI)*, 2021, pp. 1-6. doi: 10.1109/ICSAI53574.2021.9664220.

[4] R. Gonzalez-Villanueva y J. Carranza-Rojas, «Improving Balanced Accuracy for Minority Plant Species Under Data Imbalance: A Multi-Architectural Ensemble Approach», en *2023 IEEE 5th International Conference on BioInspired Processing (BIP)*, 2023, pp. 1-6. doi: 10.1109/BIP60195.2023.10379201.

[5] A. Dosovitskiy *et al.*, «An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale». 2021.

[6] H. Touvron, M. Cord, y H. Jégou, «DeiT III: Revenge of the ViT». 2022.

[7] J. Maurício, I. Domingues, y J. Bernardino, «Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review», *Applied Sciences*, vol. 13, n.º 9, 2023, doi: 10.3390/app13095521.

[8] H. He, Y. Bai, E. A. Garcia, y S. Li, «ADASYN: Adaptive synthetic sampling approach for imbalanced learning», en *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 1322-1328. doi: 10.1109/IJCNN.2008.4633969.

[9] N. V Chawla, K. W. Bowyer, L. O. Hall, y W. P. Kegelmeyer, «SMOTE: Synthetic Minority Over-sampling Technique», *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, jun. 2002, doi: 10.1613/jair.953.

[10] Y.-S. Jeon y D.-J. Lim, «PSU: Particle Stacking Undersampling Method for Highly Imbalanced Big Data», *IEEE Access*, vol. 8, pp. 131920-131927, 2020, doi: 10.1109/ACCESS.2020.3009753.

[11] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, y F. Herrera, «A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches», *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, n.º 4, pp. 463-484, 2012, doi: 10.1109/TSMCC.2011.2161285.

[12] Y. Arun y G. S. Viknesh, «Leaf Classification for Plant Recognition Using EfficientNet Architecture», en *2022 IEEE Fourth International Conference on Advances in Electronics, Computers and Communications (ICAECC)*, 2022, pp. 1-5. doi: 10.1109/ICAECC54045.2022.9716637.

[13] L.-H. Li y R. Tanone, «Ensemble Learning based on CNN and Transformer Models for Leaf Diseases Classification», en *2024 18th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, 2024, pp. 1-6. doi: 10.1109/IMCOM60618.2024.10418393.

[14] S. P. Kyathanahally *et al.*, «Ensembles of data-efficient vision transformers as a new paradigm for automated classification in ecology», *Sci Rep*, vol. 12, n.º 1, p. 18590, 2022.

[15] D. D. K. R. W. Dandeniya, B. C. T. Wickramasinghe, y C. Dasanayaka, «A Web-based Application for Snake Species Identification using Vision Transformer and CNN-based Ensemble Meta Classifier», en *2022 IEEE Pune Section International Conference (PuneCon)*, 2022, pp. 1-5. doi: 10.1109/PuneCon55413.2022.10014812.

[16] J. Yosinski, J. Clune, Y. Bengio, y H. Lipson, «How transferable are features in deep neural networks?» 2014.

[17] K. He, X. Zhang, S. Ren, y J. Sun, «Deep Residual Learning for Image Recognition». 2015.

[18] M. Tan y Q. V Le, «EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks». 2020.

[19] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, y S. Xie, «A ConvNet for the 2020s». 2022.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, y L. Fei-Fei, «ImageNet: A large-scale hierarchical image database», en *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248-255. doi: 10.1109/CVPR.2009.5206848.

[21] D. Choi, C. J. Shallue, Z. Nado, J. Lee, C. J. Maddison, y G. E. Dahl, «On Empirical Comparisons of Optimizers for Deep Learning». 2020.

[22] T. Yu y H. Zhu, «Hyper-Parameter Optimization: A Review of Algorithms and Applications». 2020.

[23] D. P. Kingma y J. Ba, «Adam: A Method for Stochastic Optimization». 2017.

[24] I. Loshchilov y F. Hutter, «SGDR: Stochastic Gradient Descent with Warm Restarts». 2017.

[25] S. P. Wright, «Adjusted p-values for simultaneous inference», *Biometrics*, pp. 1005-1013, 1992.