Algunos problemas concrentas para el Deep Learning en la agronomía: caso *prodiplosis*

Gabriel Quinche*

5 de julio de 2022

Resumen

La mítica ecuación cuadrática $x = b \pm \sqrt{(} - 4ac + b^2)/2a$ es conocida por un espectro diverso, tanto quienes aproximan sistemas dinámicos con ellas, hasta jóvenes que juran nunca la usaran, y si bien en el mundo real casi nunca serán una herramienta directa para tomar decisiones, estar cómodas con estas son los cimientos mínimos para comprender herramientas hoy tan aduladas como el Machine learning, o la inteligencia artificial.

Herramientas de esa área conocidas como redes neuronales que potencian motores de búsqueda, filtros de cámara, o puntajes crediticios. pueden ser construidas con estructuras incluso algo más simples algebraicamente, $bx + a$ y máx$(0, x)$. Es el uso de estas como bloques uno sobre otro que a forma de ingeniero civil o arquitecto termina como un modelo equilibrando elegancia y utilidad. En el siguiente texto contaremos algunos los avances que se tiene respecto a un modelo de reconocimiento de la *prodiplosis longfilia* y recomendaciones practicas respecto a los datos que son necesarios y como deben ser tratados con cuidado para darnos modelos que realmente nos den mejores herramientas decisión, y no sean una tecnificación sin fundamentos, también nombraremos algunas de las tecnologías básicas que se usan en el mismo.

1. Mucha data y balanceada

El problema principal que se tuvo desde un inicio fue la escasez de datos, si bien en internet existen algunos supuestos modelos o repositorios que creaban imagenes artificiales de la *prodiplosis longfilia* [1], habían logrado puntajes muy grandes a costa del leekage, pues las imagenes que generaban y pueden ser consultadas en los github correspondientes tenían aspectos casi de píxeles discretos, reconocimiento de los mismos en la industria no tendría ningún sentido.

*UNAL facultad de Ciencias

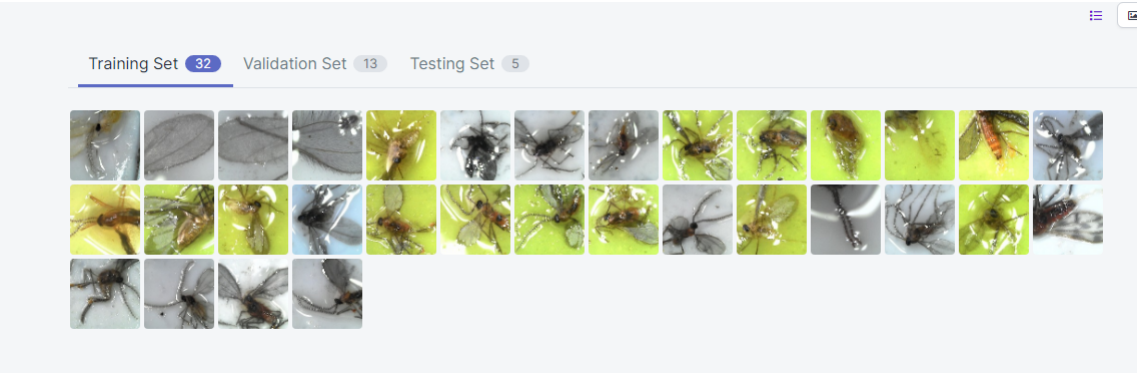

Figura 1: Esquema clásico para entrenamiento de modelos, dividiendo en tres categorías los datos

2. Asimetría de la obtención de los datos

Un problema importante al momento de entrenar el modelo es que no se contaba con uniformidad básica entre las fotografías del individuo a reconocer, siendo esta una razón por la que las pocas imagenes encontradas en linea del individuo no eran buenas para entrenar el modelo, las fotos de los individuos al ser tan particulares, eran casi siempre tomadas en un laboratorio, con mayor cuidado, resolución y además en fondos más neutros, al modelo implementado se le debió entonces eliminar los colores de fondo para evitar aprendiera a solo reconocer este patrón. A la final siendo tan pocos los datos lo poco que se pudo lograr fue trabajar con un esquema de hold-out, y aspirar a tener cierta generalización, se logro al menos evitar la trampa de la precisión, marcando casi todo como un falso negativo. sin embargo, las probabilidades finales estimadas en general tendieron a ser más bajas de lo esperadas, siendo muy pocas consideradas como mayores al 50 %

3. Falta de una forma robusta de manejo de dependencias

En la actualidad un algoritmo de deep learning que toma todas las fases importantes de generar cierto modelo predictor tiende a usar más de 10 paquetes de alta complejidad, el código *python -m pip freeze* nos puede vislumbrar sobre todo al ser ejecutado en un ambiente como Colab, o Kaggle que intenta traer preinstalados los paquetes más usados, en el proyecto se busco generar paralelamente un proyecto en colab para que se resaltara la reproducibilidad del procedimiento, también se busco fijar las versiones de los paquetes usados para que una persona no deba adivinar en que estado estaba el notebook para lograr ejecutarlo en su totalidad, en parte estos y otros problemas son los que tecnologías nuevas como *Pluto* del entorno de *Julia* pretenden solucionar

4. Resultados

Los resultados obtenidos fueron algo peores de lo esperados, pues se están asignado probabilidades muy bajas a los ejemplos de testeo de *prodiplosis*, en general sin embargo debemos aclarar que en el modelo teníamos probabilidades muy bajas asignadas a casi todos los ejemplo, donde más del 90 % de fotos no correspondían a la especie, y además les asignamos un cero como probabilidad, por lo que los confianzas que generábamos alrededor de 40 % tampoco deben ser consideradas como una clasificación totalmente negativa. A futuro se puede mejorar el modelo de las siguientes formas por orden de facilidad y relevancia

- Obtener más datos (especialmente imagenes de trampas y no las de la red de *prodiplosis*)
- Implementar una arquitectura más compleja, y asignación de hiperparametros siguiendo más estandares establecidos como [2]
- Implementar otras técnicas de validación como *hold one out* o tecnicas más modernas como [3] que tienen como ventaja permitirnos usar más datos para el entrenamiento

5. Repositorio

El notebook en el que se realizaron los experimentos es el siguiente vinculo, mientras que para acceder a un repositorio con este documento y otros conceptos importantes de los modelos de clasificación puede consultar el siguiente link

Referencias

[1] J. Cabrera and E. Villanueva, "Investigating generative neural-network models for building pest insect detectors in sticky trap images for the peruvian horticulture," in *Annual International Conference on Information Management and Big Data*, pp. 356–369, Springer, 2022.

[2] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," in *Neural networks: Tricks of the trade*, pp. 437–478, Springer, 2012.

[3] C. Baek, Y. Jiang, A. Raghunathan, and Z. Kolter, "Agreement-on-the-line: Predicting the performance of neural networks under distribution shift," *arXiv preprint arXiv:2206.13089*, 2022.