UNIVERSIDAD DE GRANADA

TRABAJO FIN DE GRADO

INGENIERÍA INFORMATICA

Algoritmos meméticos para reducir datos de entrenamiento en modelos de aprendizaje profundo convolucionales

---

Autor

José Ruiz López (alumno)

Directores

Daniel Molina Cabrera (tutor)

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación

—

Granada, Noviembre de 2024

Algoritmos meméticos para reducir datos de entrenamiento en modelos de aprendizaje profundo convolucionales

José Ruiz López (alumno)

Palabras clave: Algoritmos meméticos, Imágenes, Modelos de Aprendizaje profundo convolucionales

Resumen

Los modelos de **Aprendizaje Profundo** (Deep Learning) han supuesto un hito en la Inteligencia Artificial al ser capaz de procesar y ser capaces de reconocer patrones complejos. Dentro de estos, los modelos convolucionales se han mostrado muy capaces de identificar todo tipo de objetos/ características en imágenes.

Sin embargo, a diferencia de las personas, requieren un número muy alto de datos de entrenamiento para cada categoría que debe aprender. Eso implica, además de entrenamiento más largo, una recogida de datos de entrenamiento que, según lo que se desea que aprenda, puede ser problemático de obtener. Además de la obtención de los datos, la nueva ley europea sobre IA, **IA Act** requerirá sobre aplicaciones de IA con datos sensibles, una auditoría no solo del propio modelo, sino también de los datos utilizados para entrenarla. Auditoría que crecerá en complejidad confirme aumente en número el conjunto de entrenamiento. Por tanto, se hace conveniente poder reducir el conjunto de entrenamiento.

Ya se ha confirmado que incrementar el número de imágenes de entrenamiento puede mejorar el proceso o no, según si las imágenes realmente contribuyan al proceso de entrenamiento. Es más, gracias a las técnicas de aumento de datos (**Data Augmentation**) la posible necesidad de imágenes muy similares entre sí se reduce al ser capaz de construirse de forma automática más imágenes de entrenamiento (imágenes que no suponen un problema de cara a una autoría).

En este trabajo planteamos el uso de estrategias avanzadas, como **algoritmos metaheurísticos**, y métricas de similitud entre imágenes, para establecer un proceso de reducción de imágenes de entrenamiento (selección de instancias) para poder reducir el conjunto de entrenamiento. De esta manera, se seleccionarían solo un conjunto reducido de imágenes representativas que, gracias a las **técnicas de aumento de datos**, puedan entrenar

modelos con una calidad suficiente. Por lo tanto, se podría reducir muy significativamente el conjunto de entrenamiento.

Memetic Algorithms for Reducing Training Data in Convolutional Deep Learning Models

José, Ruiz López (student)

Keywords: Memetic Algorithms, Images, Convolutional Deep Learning Models

Abstract

Deep Learning models have marked a milestone in Artificial Intelligence by being able to process and recognize complex patterns. Among these, convolutional models have proven very capable of identifying all kinds of objects/features in images.

However, unlike humans, they require a very high number of training data for each category they need to learn. This implies not only longer training times but also a data collection process that can be problematic to obtain, depending on what is desired for the model to learn. In addition to data acquisition, the new European law on AI, the **AI Act**, will require an audit not only of the model itself but also of the data used to train it, especially when dealing with sensitive data. The complexity of this audit will grow as the size of the training dataset increases. Therefore, it becomes necessary to reduce the training dataset.

It has been confirmed that increasing the number of training images may or may not improve the process, depending on whether the images truly contribute to the training process. Moreover, thanks to **data augmentation techniques**, the potential need for very similar images is reduced, as it is possible to automatically generate more training images that do not pose authorship issues.

In this work, we propose the use of advanced strategies, such as **metaheuristic algorithms** and image similarity metrics, to establish a process for reducing training images (instance selection) in order to minimize the training dataset. This way, only a reduced set of representative images would be selected, which, thanks to **data augmentation techniques**, could sufficiently train models. Thus, the training dataset could be significantly reduced.

Yo, **José Ruiz López**, alumno de la titulación TITULACIÓN de la **Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación de la Universidad de Granada**, con DNI 77964364E, autorizo la ubicación de la siguiente copia de mi Trabajo Fin de Grado en la biblioteca del centro para que pueda ser consultada por las personas que lo deseen.

Fdo: José Ruiz López

Granada a X de mes de 201 .

D. **Daniel Molina Cabrera (tutor**, Profesor del Departamento Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada.

Informan:

Que el presente trabajo, titulado ***Algoritmos meméticos para reducir datos de entrenamiento en modelos de aprendizaje profundo convolucionales***, ha sido realizado bajo su supervisión por **José Ruiz López (alumno)**, y autorizamos la defensa de dicho trabajo ante el tribunal que corresponda.

Y para que conste, expiden y firman el presente informe en Granada a X de mes de 201 .

Los directores:

Daniel Molina Cabrera (tutor)

Agradecimientos

Poner aquí agradecimientos...

Capítulo 1

Introducción

1.1. Motivación

En la era actual, donde los datos se generan a un ritmo vertiginoso, la necesidad de procesar y analizar grandes volúmenes de información se ha vuelto crucial. Los **modelos de aprendizaje profundo**, y en particular las **redes neuronales convolucionales**, han demostrado su capacidad para alcanzar niveles sin precedentes de precisión en tareas como la clasificación de imágenes y el reconocimiento de patrones. Sin embargo, estos modelos a menudo requieren enormes cantidades de datos de entrenamiento, lo que plantea desafíos significativos en términos de tiempo, costo y recursos computacionales.

Es aquí donde entra en juego la importancia de los **algoritmos meméticos**. Combinando las fortalezas de las técnicas evolutivas con **métodos de búsqueda local**, estos algoritmos ofrecen un enfoque innovador y eficiente para la reducción de datos. La optimización de conjuntos de datos no solo puede mejorar el rendimiento de los modelos, sino que también permite una capacitación más rápida y menos costosa, facilitando así la investigación y el desarrollo en diversas aplicaciones.

Realizar un TFG sobre este tema no solo representa una oportunidad para explorar una frontera apasionante de la inteligencia artificial. La investigación en algoritmos meméticos para la reducción de datos puede ser clave para hacer más accesible el **aprendizaje profundo** a aquellos que enfrentan limitaciones de datos, permitiendo asi un futuro donde la inteligencia artificial sea más inclusiva y eficiente.

Este proyecto se centra en el uso de metaheurísticas para la resolución de este problema. Las metaheurísticas son algoritmos de optimización cuyo

uso principal recae en tareas cuyo resultado es difícil de obtener por medios convencionales. Ciertos problemas pueden no tener una solución algorítmica obvia o simplemente ser demasiado complejos en cuanto al tiempo de resolución de los mismos. En ambos casos, las metaheurísticas son capaces de ofrecer soluciones muy buenas y en un tiempo admisible por medio de procesos de búsqueda inspirados en múltiples ámbitos (física, biología, comportamiento social, etc).

Para la selección de características existen multitud de métodos que tratan de aplacar este problema. Algunos de los más famosos son los método de filtrado, el análisis de componentes principales (*PCA*) o incluso distintos tipos de regresiones como *Lasso*. En este documento se estudian métodos de envoltura o *Wrapper*, los cuáles hacen uso del entrenamiento y evaluación de modelos de *Machine Learning* para evaluar distintos conjuntos de características. Este documento se centra en el uso de métodos *Wrapper* o de envoltura debido a que el objetivo principal es comparar y analizar distintas alternativas metaheurísticas.

El reciente interés del problema de la selección de características en el ámbito de las metaheurísticas en los últimos años es más que evidente. Puede comprobarse como en los esta última época hay una tendencia en la publicación de artículos presentando nuevos métodos metaheurísticos, mejores con respecto a los clásicos, o incluso comparativas y análisis entre distintos algoritmos.

Esta crecimiento viene acompañado, sin embargo, de comparaciones que distan de ser objetivas por varios motivos [**?**]. Entre varios artículos se comparan algoritmos del mismo tipo con soluciones y resultados muy variables entre sí a pesar de mismas configuraciones a la hora de experimentar, artículos sin código referenciado, de forma que sea más fácil interpretar los resultados o duplicarlos, y algoritmos novedosos presentados por su autor o autores que superaban al resto en alguna métrica concreta sin llegar a la rigurosidad adecuada.

Por lo expuesto, la motivación principal de este trabajo es la de proveer información todo lo objetiva posible por medio de un análisis comparativo entre los algoritmos optimizatorios metaheurísticos más populares y más citados junto con los algoritmos más robustos y clásicos en el campo de la optimización pseudo estocástica. Se plantea una serie de estudios y comparaciones entre los algoritmos, haciendo uso de sus versiones en codificación binaria y su versiones en codificación continua o real. Se realizarán análisis sobre los resultados de forma que se obtengan respuestas a una serie de preguntas de investigación.

- ¿Son buenos en *Feature Selection* los algoritmos continuos que también lo son en binario?
- ¿Qué algoritmos modernos son más prometedores?
- ¿Los clásicos siguen siendo una opción viable frente a los modernos?

En este trabajo, se abordará exitosamente la respuesta a esta serie de preguntas.

1.2. Objetivos

Objetivo General:
Realizar una comparación exhaustiva y objetiva de diversas metaheurísticas utilizadas en la selección de características, con el propósito de proporcionar una visión integral y evaluativa sobre su eficacia y aplicabilidad en diferentes contextos de análisis de datos.

Objetivos Específicos:

1. Evaluar el desempeño de las metaheurísticas más relevantes en el ámbito de la selección de características, analizando métricas clave como precisión, estabilidad de las soluciones y eficiencia computacional. Se emplearán conjuntos de datos de referencia y metodologías de validación cruzada para garantizar la robustez de los resultados.
2. Investigar la transferibilidad de las técnicas diseñadas para dominios continuos y binarios en el contexto de la selección de características. Se analizará si las metaheurísticas efectivas en un dominio son igualmente eficaces cuando se aplican a otro, identificando posibles ventajas y limitaciones de cada enfoque.
3. Identificar las fortalezas y debilidades de cada metaheurística según el tipo de representación de las características. Se realizará un análisis detallado del comportamiento de las técnicas en problemas de selección de características con diferentes tipos de datos, destacando su rendimiento relativo y sus áreas de aplicación más adecuadas.
4. Proporcionar recomendaciones prácticas basadas en los resultados obtenidos, con el objetivo de orientar a practicantes y académicos en la selección y aplicación de metaheurísticas en problemas reales de selección de características.
5. Evaluar los resultados de las metaheurísticas en problemas de selección de característica usando distintos como algoritmos de aprendizaje los métodos *kNN* y *SVM*. Se realizará una comparativa a nivel de eficiencia en tiempo y calidad de los resultados.

Capítulo 2

Diseño experimental

En este capítulo se describirá el proceso de experimentación llevado a cabo, los conjuntos de datos usados para los distintos experimentos.