Aprenentatge automàtic aplicat al reconeixement de caràcters en imatges

Francesc Aguirre* Alberto Debernardi†

05/03/2021

1. Introducció

L'estadística és la disciplina que s'encarrega d'analitzar dades per a respondre a preguntes empíriques. A mitjans del segle XX, amb la creació dels dispositius electrònics d'emmagatzematge i l'ajuda de sensors, la quantitat d'informació que es pot recol · lectar ha crescut any rere any. Aquest fet ha donat lloc al naixement de noves disciplines d'anàlisi de dades, tals com la mineria de dades i l'aprenentatge automàtic, més conegut pel seu nom en anglès *machine learning*.

El machine learning és un conjunt de tècniques que dóna als ordinadors l'habilitat d'aprendre de les dades. S'utilitza per a resoldre una gran varietat de problemes predictius complexos en els àmbits d'economia i finances, bioinformàtica, salut, meteorologia, màrqueting, problemes en l'anàlisi i classificació d'imatges, vídeos, àudio... En l'àmbit de l'anàlisi d'imatges hi ha la detecció d'objectes, i més concretament, el reconeixement òptic de caràcters **OCR** (de les sigles en anglès *Optical Character Recognition*).

L'objectiu de l'OCR és detectar i identificar els caràcters d'una imatge. Hi ha diferents aplicacions d'OCR que poden ajudar a agilitzar o automatitzar processos, alguns dels quals són la digitalització de diaris i llibres antics, la recopilació de respostes en formularis i enquestes, la identificació de matrícules i la classificació d'imatges segons el text detectat.

2. Objectiu i metodologia

L'objectiu del treball és investigar quines tècniques de machine learning són més adequades per la identificació de caràcters irregulars (ex: escriptura manual) en imatges.

Per aconseguir l'objectiu, es farà una recerca de conceptes bàsics sobre OCR i machine learning. Es visitarà la literatura sobre la recerca que s'ha fet en problemes similars d'OCR, i s'aplicaran els coneixements adquirits en un problema pràctic, que consisteix en el disseny d'un model per a identificar caràcters escrits a mà o la identificació de *captcha* (és a dir, text en una determinada font que inclou una certa distorsió, precisament per evitar la detecció dels caràcters per part de models més simples, i que són àmpliament utilitzats en internet per evitar l'automatització de determinats processos).

Per a resoldre el problema pràctic, s'aplicarà un procediment de machine learning de principi a fi (Géron 2019, apèndix B), que consisteix en l'exploració del problema general, la recerca de bases de dades públiques, l'exploració de les dades per guanyar coneixements sobre aquestes, la preparació de les dades, l'exploració de diferents models i selecció dels millors, l'optimització dels models, i finalment la combinació dels models en una solució final.

*Autor. francesc.aguirre@e-campus.uab.cat
†Tutor. adebernardipinos@gmail.com

Referències

"A Gentle Introduction to Ocr." 2018. https://towardsdatascience.com/a-gentle-introduction-to-ocr-ee1469a201aa.

Géron, Aurélien. 2019. *Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media.

"Kaggle Database." 2021. https://www.kaggle.com.

Singh, Amarjot, Ketan Bacchuwar, and Akshay Bhasin. 2012. "A Survey of Ocr Applications." *International Journal of Machine Learning and Computing* 2 (3): 314.

"TFG Francesc Aguirre, Project Tracking and Code." 2021. https://github.com/cesc1/TFG.