Aprenentatge automàtic aplicat al reconeixement de caràcters en imatges

Francesc Aguirre,[1] Alberto Debernardi[2]

[1]Autor: francesc.aguirre@gmail.com

[2]Tutor: adebernardipinos@gmail.com

23/07/2021

Universitat Autònoma de Barcelona

Aquí anirà el resum en cat, cast, ang.

Contents

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.

En aquest estudi s'investigarà l'àmbit del reconeixement òptic de caràcters, i com obtenir models predictius competents utilitzant tècniques de machine learning. Per a posar aquests coneixements en pràctica, es resoldrà un problema d'identificació de caràcters irregulars en imatges, tal com la identificació d'escriptura manual o la resolució 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).

2 OCR

El reconeixement òptic de caràcters OCR (de les sigles en anglès *Optical Character Recognition*) és una aplicació de la intel·ligència artificial, que té com a objectiu detectar i identificar els caràcters que es puguin trobar en una imatge. És una de les àrees més estudiades en l'àmbit de reconeixement de patrons gràcies al gran nombre d'aplicacions pràctiques. Alguns dels prob-

lemes que OCR pot ajudar a agilitzar i automatitzar són la digitalització de diaris i llibres antics, la identificació de matrícules, la classificació d'imatges segons el text detectat, i la lectura de dades en paper tals com documentació, correu i enquestes. La digitalització a ordinador de text té moltes més aplicacions, tals com la cerca, edició i emmagatzematge d'informació, traducció, transcripció de text a àudio i NLP (de les sigles en angles *Natural Language Processing*).

2.1 Parts tradicionals d'un sistema OCR

Tradicionalment, un problema típic d'OCR es pot dividir en subtasques en forma de *pipeline* (fetes una darrera l'altre en un ordre concret) que utilitzen tècniques de visió artificial (en anglès *computer vision*), estadístiques i de machine learning. És útil conèixer-les per saber en quin punt del sistema s'està (*1*).

1. Escaneig: És el procés d'escanejar o fotografiar el text *input*. Normalment, al d'escanejar un document s'aplica *thresholding* (binarització) per estalviar memòria i capacitat computacional, que és una tècnica que dicotomitza el color gris de documents en blanc i negre segons un llindar d'intensitat.

2. Segmentació de text: Un cop escanejat el document, típicament el que es vol és obtenir un sol caràcter per utilitzar-lo d'input en el model. La segmentació d'una imatge és la divisió d'aquesta en parts. En aquesta tasca, el que es vol és localitzar el text que ens interessa i ometre gràfics, imatges, logotips... Tot seguit es vol segmentar línies de text, de les línies es segmenten paraules, i de les paraules se segmenten caràcters. En la segmentació ens podem trobar diversos problemes, sobretot si un caràcter està format per diferents parts (i j), si s'està tocant amb algun altre, o s'ha dividit.

3. Preprocessament: Com que el fet d'escanejar o fotografiar una imatge és un procés variable (brillantor, angle...), els caràcters de la imatge segmentada poden contenir soroll, o

estar trencats. El preprocessament té com a objectiu eliminar soroll, omplir espais trencats i reduir la grossor dels píxels negres que formen el caràcter. També es normalitzen les dimensions i s'aplicarà una rotació si es detecta inclinació.

4. Segmentació interna: Consisteix a segmentar la imatge del caràcter en seccions més petites, tals com línies i corbes concretes. La intensió és començar a detectar zones amb patrons concrets que facilitin el reconeixement posterior del caràcter.

5. Extracció de variables: A partir de la segmentació interna, se seleccionarà un conjunt de variables que maximitzi el reconeixement amb el nombre menor d'elements. L'objectiu és capturar característiques essencials dels símbols per posteriorment entrenar el model. Si no s'utilitzés segmentació, s'utilitzaria la matriu de píxels del caràcter. Utilitzant extracció de variables, s'utilitzarien característiques que descriguin els caràcters, tals com llargades de segments en regions de la imatge, angles de curvatura...

6. Entrenament i reconeixement: Aplicació de tècniques de reconeixement de patrons per a classificar el caràcter. Aquí és on podem utilitzar tècniques estadístiques i de machine learning per fer la classificació, tals com (TODO)...

7. Reagrupació de caràcters a paraules, paraules a línies, fins a tenir el document complet.

Com es pot veure, la creació d'un sistema OCR tradicional és un procés llarg, amb moltes subtasques que relativament complicades. Els avantatges dels mètodes tradicionals és que donen bons resultats amb mostres petites i són computacionalment eficients. Alguns dels inconvenients és que utilitzen detecció i segmentació de text a partir de tècniques de visió artificial no relacionades amb machine learning, és a dir, no aprenen de les dades. Com que la segmentació no sempre és evident, i les dades poden ser sorolloses, utilitzar aquest tipus de tècniques acostuma a produir errors difícils de solucionar.

2.2 Sistema OCR a l'actualitat

Aquests últims anys, a partir del 2010, el desenvolupament de la branca *deep learning* (aprenentatge profund, una branca de machine learning) a tingut un avanç important. La solució del problema de *vanishing/exploding gradients* i el llançament de noves tecnologies d'alta capacitat computacional tals com les GPU (de les sigles en anglès *graphics processing unit*), a permès l'entrenament de xarxes neuronals profundes (DNN de les sigles en anglès *deep neural networks*).

Aquest desenvolupament ha donat lloc a nous sistemes OCR basats en xarxes neuronals. La majoria del preprocessament i binarització no és necessari, ja que les xarxes neuronals s'adapten als inputs, podent utilitzar fàcilment els píxels de les imatges. La segmentació de text es pot fer amb tècniques basades en DNN tals com FCN (de les sigles en anglès *Fully Convolutional Networks*), donant millors resultats que amb les tècniques de visió artificial. A més a més, si s'utilitzen RNN (de les sigles en anglès *Recurrent Neural Networks*), no cal segmentar caràcter a caràcter, sinó que es poden utilitzar línies completes de text. Un altre avantatge de les RNN és que poden aprendre de manera natural la llengua utilitzada en l'entrenament. La tendència actual és utilitzar FCN per la segmentació de text en línies, i utilitzar RNN pel seu reconeixement, juntament amb CNN (de les sigles en anglès *Convolutional Neural Networks*) per fer l'extracció de variables (*2*).

3 Anàlisi de dades

3.1 Base de dades NIST

References and Notes

1. A. Chaudhuri, K. Mandaviya, P. Badelia, S. K. Ghosh, *Optical Character Recognition Systems for Different Languages with Soft Computing* (Springer, 2017), pp. 9–41.

2. J. Martínek, L. Lenc, P. Král, *Neural Computing and Applications* **32**, 17209 (2020).

3. Base de dades NIST: web link

3. Codi desenvolupat: github/cesc1/TFG