Multimodal deep networks for text and image-based document classification

Nicolas Audebert Catherine Herold Kuider Slimani Cédric Vidal

Quicksign, 38 rue du Sentier, 75002 Paris

{nicolas.audebert,catherine.herold,kuider.slimani,cedric.vidal}@quicksign.com

Résumé

La classification automatique de documents numérisés est importante pour la dématérialisation de documents historiques comme de procédures administratives. De premières approches ont été suggérées en appliquant des réseaux convolutifs aux images de documents en exploitant leur aspect visuel. Toutefois, la précision des classes demandée dans un contexte réel dépend souvent de l'information réellement contenue dans le texte, et pas seulement dans l'image. Nous introduisons un réseau de neurones multimodal capable d'apprendre à partir d'un plongement lexical du texte extrait par reconnaissance de caractères et des caractéristiques visuelles de l'image. Nous démontrons la pertinence de cette approche sur Tobacco3482 et RVL-CDIP, augmentés de notre jeu de données textuel QS-OCR[1], sur lesquels nous améliorons les performances d'un modèle image de 3% grâce à l'information sémantique textuelle.

Mots-clés

Classification de documents, apprentissage multimodal, fusion de données.

Abstract

Classification of document images is a critical step for archival of old manuscripts, online subscription and administrative procedures. Computer vision and deep learning have been suggested as a first solution to classify documents based on their visual appearance. However, achieving the fine-grained classification that is required in real-world setting cannot be achieved by visual analysis alone. Often, the relevant information is in the actual text content of the document. We design a multimodal neural network that is able to learn from word embeddings, computed on text extracted by OCR, and from the image. We show that this approach boosts pure image accuracy by 3% on Tobacco3482 and RVL-CDIP augmented by our new QS-OCR text dataset[1], even without clean text information.

Keywords

Document classification, multimodal learning, data fusion.

1 Introduction

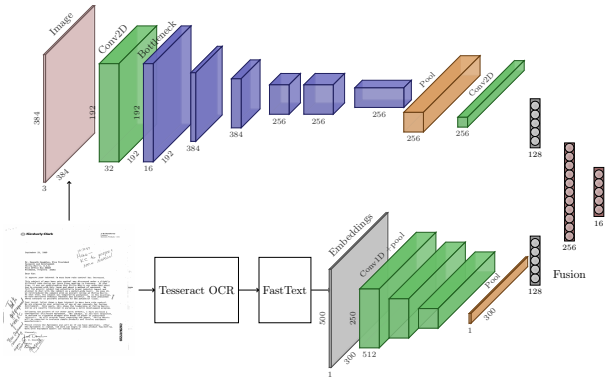

Figure 1: Multimodal classifier for hybrid text/image classification. Training is performed end-to-end on both textual and visual features.

The ubiquity of computers and smartphones has incentivized governments and companies alike to digitize most of their processes. Onboarding new clients, paying taxes and proving one's identity is more and more done through a computer, as the rise of online banking has shown in the last few years. Industrial and public archives are also ongoing serious efforts to digitize their content in an effort for preservation, e.g. for old manuscripts, maps and documents with a historical value. This means that previously physical records, such as forms and identity documents, are now digitized and transferred electronically. In some cases, those records are produced and consumed by fully automated systems that rely on machine-readable formats, such as XML or PDF with text layers. However, most of these digital copies are generated by end-users using whatever mean they have access to, i.e. scanners and cameras, especially from smartphones. For this reason, human operators have remained needed to proofread the documents, extract selected fields, check the records' consistency and ensure that the appropriate files have been submitted. Automation through expert systems and machine learning can help accelerate this process to assist and alleviate the burden of this fastidious work for human workers.

A common task involved in data filing processes is document recognition, on which depends the class-specific rules that command each file. For example, a user might be asked to upload several documents such as a filled subscription form, an ID and a proof-of-residence. In this work, we tackle the document classification task to check that all required files

[1] https://github.com/Quicksign/ocrized-text-dataset