تحليل هيكل المستندات العربية باستخدام التعلم العميق

المستخلص

في الآونة الأخيرة اهتم الباحثون بمجال تحليل المستندات والتعرف الضوئي على الأحرف (OCR) ، حيث يمكننا أن نرى تحسنًا كبيرًا للغاية في أنواع مختلفة من محركات التعرف الضوئي على الحروف للغات مختلفة ، سواء بالنسبة للمستندات المطبوعة أو المستندات المكتوبة بخط اليد . كما نرى هناك اهتمام أقل بمعالجة المستندات باللغة العربية مقارنة باللغات الأخرى مثل اللغة الإنجليزية، ويمكن أن يكون ذلك لأسباب عديدة مثل صعوبة تحليل اللغة العربية ومحدودية قاعدات البيانات للمستندات العربية الموجودة. لتنفيذ أي محرك OCR، فإن الخطوة الأولى التي نحتاج إلى القيام بها هي تحليل تخطيط الصور المحددة قبل إرسال الصورة إلى OCR. تهتم دراسة الأطروحة هذه بتحليل تخطيط المستندات العربية باستخدام نهج التعلم العميق. نحن نستخدم في هذه الاطروحة نموذجين مختلفين للتعلم العميق، وهما الشبكة العصبية Faster RCNN وMask RCNN، حيث يتم تحديد استخدام كل نموذج لنوع مختلف من المستندات العربية. نحن نقوم بتجميع ثلاثة أنواع مختلفة من قاعدات البيانات للمستندات العربية وهم، المستندات المطبوعة قديما المسندات المطبوعة حديثا، والمستندات التاريخية، حيث إن لكل نوع من المستندات حجمها الخاص وهيكلها ومتطلبات خاصة لمعالجتها. نحن نقوم باستخدام Faster RCNN للمستندات العربية المطبوعة قديما والمطبوعة حديثا، وMask RCNN للمستندات التاريخية. معالجة المستندات العربية أصعب من معالجة المستندات الآخرى، وذلك بسبب الهيكل التخطيطي وصفات المستندات التاريخية مثل، نمط خط الكاتب عمر الورق المستخدم، الفترة الزمنية التي جاء منها المستند، الحبر المستخدم، وغيرها الكثير. نتائج هذه الأطروحة هي كالتالي، ٩٩.٥٩٪ للمستندات المطبوعة قديما، و٩٩.٥٦٪ للمستندات المطبوعة حديثا، وبالنسبة للوثائق التاريخية ٥١.١٤٪. مقارنة نتائجنا بنتائج النماذج الأخرى الموجودة، يمكننا القول إن عملنا يمكنه التغلب على الكثير من النماذج الموجودة في هذا المجال بنتائج رائعة.

الكلمات المفتاحية: *تحليل المستندات ، اللغة العربية ، منطقة الاهتمام ، Faster RCNN ، Mask RCNN*

LAYOUT ANALYSIS FOR ARABIC DOCUMENTS USING DEEP LEARNING

Abstract

In recent days, researchers are very interested in the field of document analysis and optical character recognition (OCR). There is a very great improvement in OCR engines with different languages, and whether for printed documents or handwritten documents. We have a less concerned for processing documents with Arabic language comparing with other languages such as English, and this is due to many reasons like, the difficulty of processing Arabic language, and the limitation of the existed Arabic documents datasets. To implement any OCR, engine, the first step we need to do is analyzing the layout of the images before we send the image to the OCR. This thesis is concerning about layout analysis for Arabic documents using deep learning (DL) approach. We are using two different DL models, Faster Region-based Convolutional Neural Network RCNN and Mask (RCNN), where each model has specific design that match different type of Arabic documents. We are collecting three different types of Arabic documents datasets, early printed, printed, and historical, where each dataset has its own size, structure, and processing requirements. We are using Faster RCNN for printed and early printed Arabic documents, and Mask RCNN for historical Arabic document. Processing historical documents is more difficult because of the layout structure and the proprieties of the historical documents like, the author handwritten style, the age of the paper, the time period the document came from, the used ink and more. The accuracy result is as follow: 99.59% for early printed documents, and 99.56% for printed documents, and 51.14% for historical. Comparing our result with other existed models, we can say that our work can achieve state of the art work with an impressive result.

Key Word: *Layout Analysis, Arabic Language, RoI, Faster RCNN, Mask RCN*