تحسين التعرف علي النص المكتوب باللغة العربية بإستخدام التعلم العميق

إعداد

محمود محمد احمد بدرى

رسالة مقدمة إلى كلية الحاسبات و المعلومات

جامعة القاهرة

كجزء من متطلبات الحصول على درجة الماجيستير

فى علوم الحاسب

تحت إشراف

أ.د هشام أحمد حسن د. حسين عكاشة د. هناء بيومى

كلية الحاسبات و المعلومات

جامعة القاهرة

جمهورية مصر العربية

يوليو ٢٠١٨

ملخص الرسالة

التعرف على نصوص اللغات فى الصور من خلال الآله له العديد من التطبيقات المفيدة و التى تتضمن البحث فى المستندات المصورة و حفظها. على الرغم من تمكن الأبحاث العلمية من التعرف على النصوص اللاتينية بسهولة، ما زال التعرف على النصوص العربية فى الصور قيد التطوير خاصة فى النصوص المكتوبة بخط إلد. فى السنين القليلة الماضية، تمكن علم التعلم العميق – جزء من علم تعلم الآله – من عدة نجاحات كبيرة فى مجالات علوم الحاسب خاصة فى مجال استخلاص المميزات و الرؤية الحاسوبية و التى تتضمن التعرف على العديد من نصوص اللغات فى الصور. بالإضافة إلى أنه جعل أشياء كانت مستحيلة سهلة و قابلة للتطبيق .

يتطلب تطوير و تحسين دقة التعرف على النصوص العربية فى الصور مجموعة كبيرة من البيانات المكونة من الصور بالإضافة إلى نصوصها الاصلية حيث أن البيانات تعتبر كوقود لكثير من نماذج تعليم الآله الحديثة. تطرح هذه الرسالة قاعدة بيانات جديدة تدعي QTID التى تتكون من صور كلمات القرآن الكريم. تعتبر QTID أول قاعدة بيانات مخصصة للتعرف على نصوص اللغة العربية و التى تحتوى على علامات التشكيل فى اللغة. تتكون هذه القاعدة من ٣٠٩,٧٢٠ صورة مختلفة والتى تتضمن جميع كلمات القرآن الكريم بالخط العثمانى بإجمإلى ٢,٤٩٤,٤٢٨ حرف. تم تقسيم الصور بداخل هذه القاعدة بشكل عشوائى إلى ثلاث مجموعات. المجموعة الأولى تستخدم لتدريب نموذج تعلم الآله و التى تحتوى على نسبة ٩٠% من الصور، بينما تم تكوين مجموعتين اخريين لاختبار هذه النماذج بعد التدريب. و قد تم عمل إحصائيات مختلفة لتحليل قاعدة البيانات المطروحة. أظهر التقييم التجريبى أن افضل محركات التعرف على نصوص اللغة العربية مثل Tesseract و ABBYY FineReader لا تستطيع التعرف بشكل جيد على صور كلمات القرآن الكريم التى تم تجميعها فى قاعدة البيانات.

و اخيرا تم طرح نموذجين لتعليم الآله نصوص اللغة العربية الموجودة فى QTID بإستخدام تقنيات التعلم العميق. النموذج الأول يعتمد على نماذج الشبكات العصبية المكررة و المرئية، بينما يعتمد الآخر على المرئية فقط. تم تدريب هذه النماذج على التعرف على الخط العثمانى فى قاعدة بيانات نص القرآن الكريم. بعد ذلك تمت مقارنة النماذج المطروحة مع أفضل محركات التعرف على نصوص اللغة العربية. أظهرت هذه المقارنة تفوق دقة النماذج المطروحة على هذه المحركات.