Experiences with using Probabilistic Programming for Voucher Feature Extraction at Skanned.com

Project Status Report

Ahmad Salim Al-Sibahi
ahmad@bilagscan.dk

January 2019

1 Introduction

1.1 Background

Probabilistic Programming is an emergent field of machine learning, that enriches general programming frameworks with probabilistic constructs from Bayesian reasoning. The core idea is that one specifies probabilistic models that describe the anticipated distribution of target data, and then these frameworks provide an automated way to learn parameters of the model when new data is observed.

There are three key advantages to probabilistic programming over existing machine learning technology: it is possible to directly incorporate *domain knowledge* using prior distributions, the constructed models allow for a systematic way to *quantify uncertainty*, and they are often directly *interpretable* by humans. All of these aspects are important for voucher information extraction, which is the core business of Skanned.com. Vouchers are almost always structured documents, where key information is explicitly labelled and where there are many legal rules about how information should be presented; the use of this domain knowledge is key to achieving good results, which is why the existing system relies heavily on hand-tuned heuristics. Quantifying uncertainty of a result is an important way for the system to specify its trust in the results it provides and make sure that customers only pay what is necessary: it is important for customers that when the system states that the total amount is "$1000", it is the right amount and not "$100" or "$10000". Skanned.com provides a human-based validation service, but such service is expensive and is a bottleneck with regards to scalability; it is therefore important to only rely on it is known to be necessary, which is not possible to do in a systematic way with the current system. Finally, if an error happens during information extraction, it is important that the system is easy to debug and explain to customers, which is hard to do for purely deep neural network-based architectures with millions of nuisance parameters.

The goal of the current Industrial PostDoc is to examine how to apply probabilistic programming in practice for voucher scanning systems. This is important both for the company, where getting a good solution can result in significant savings and make the company a leading expert in machine learning, and for science in general, since there are not many existing practical applications of probabilistic programming and new experiences provide an opportunity for improving the existing systems.