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 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.

The PostDoc is mentored on the academic side by Dr. Thomas Hamelryck, who is an expert in Bayesian data analysis and its applications in Bioinformatics, and Dr. Fritz Henglein, who is an expert in programming languages and high-performance compilation to GPUs. On the business side, the mentor is Dan Rose Johansen, who is the Chief Operational Officer at BilagScan, and leads the daily operations.

2 Achievements

2.1 Knowledge Building

During my PostDoc at Skanned.com I have worked towards understanding how to apply probabilistic programming in practice, and sharing such knowledge amongst colleagues and fellow academics. Concretely:

- I have developed an understanding for Bayesian data analysis, including constructing probabilistic models, presents them and evaluating them.

- I have gotten familiar with the wide range of inference techniques available for inference in probabilistic programs, gaining an understanding of their core theory and for the problems where they are applicable.
- I have shared my knowledge about various probabilistic programming frameworks with my colleagues, and discussed how they can potentially be applied in practice.
- I am a part of a weekly probabilistic programming discussion group at University of Copenhagen, where they are actively looking into how to use modern probabilistic programming frameworks for protein folding.

2.2 Application

I have examined how we can use probabilistic programming frameworks for voucher information extraction, and developed various models aimed towards such goal.

- I have developed a model for grouping vouchers based on textual and visual features[1], that relies on latent Dirichlet allocation (LDA), which is a popular probabilistic programming technique. The grouping is useful for developing specialized algorithms for similar sets of vouchers and thus increase precision in the information extraction. This model was found useful by Skanned.com and is therefore planned to soon come in production.
- I have developed a series of probabilistic models from scratch for keyword-based feature localization. The problem is challenging to encode because of the varying number of keywords and features between each document, but initial results were promising in showing its potential use in practice: 80% of the time the target feature was the expected one, and 99% of the time it was within the 95% confidence interval.
- I am currently working with a more ambitious model that tries to assigns keywords to features explicitly, to make inference more precise. The model can also be extended to allow locating more features in the future, as well as possibly identify potential keywords.
- I have developed experience with various probabilistic programming frameworks like PyMC3, Pyro and Infer.Net, and their underlying technologies like Theano and PyTorch. This included getting experiences with neural network architectures that can potentially be used in the models or for inference. Furthermore, I have actively submitted bug reports, bug fixes and features to these languages.

2.3 Dissemination

I have tried to share my experiences using Probabilistic Programming as part of my Industrial PostDoc at Skanned.com, in order to generate excitement for the area. This is also a good branding opportunity for Skanned.com to position themselves as a cutting edge startup in artificial intelligence (AI).

- I have presented our work at the first international conference on probabilistic programming (PROBPROG 2018), which was held at Massachusetts Institute of Technology

[1] In collaboration with my mentors Fritz and Thomas

(MIT). This provided opportunity to learn about the latest technology within probabilistic programming by world's top universities and major IT companies (Microsoft/-Facebook/Google/Uber/Amazon/Oracle etc.)

- I have presented our plans for probabilistic programming at a public event organized by Skanned.com on Machine Learning and Artificial Intelligence (called Masters of AI/ML 2018). I was rated as one of the top speakers at this event by the audience, which was around 200 people.

- I have started a meet-up on probabilistic programming, to provide more focused technical discussions on the area. The meet-up is a collaboration between Skanned.com, University of Copenhagen and Hypefactors, and has already after one session over 70 members, which is great for such new techology. We already have scheduled multiple sessions and in talk to have exciting speakers from academia and industry in the future.

In general, it should be mentioned that probabilistic programming as a technology is getting a lot of excitement. This is also felt by my mentors Thomas and Fritz, when they present and discuss this work with other academics and people from industry.

3 Challenges

4 Prospectives