Review of The Infinite-order Conditional Random Field Mode for Sequential data Modeling

Summary:

This paper presents a CRF approach to sequence labeling that is reposed on a nonparametric Bayesian smoothing approach to discrete sequence modeling called the sequence memoizer. In particular, modeling P(labels|features) (notated P(Y|X)) is formulated as a Markov random field with ordinary unary potentials P(y\_t|x\_t) and “infinite order” label transition potentials P(y\_t|y\_1,…,y\_{t-1}). The latter, transition potentials are modeled using the sequence memoizer. Model training is accomplished via maximum likelihood with a forward-backward algorithm run on a “mean field” approximation to the full model used to compute the partition function (needed for training). Decoding is presented as an-EM-like gradient-ascent algorithm in the same mean-field approximation. Empirical performance is given for the model on tasks ranging form pose estimation to handwriting recognition.

Recommendation:

Somewhere in-between major revision and reject – closer to reject.

Feedback and Criticisms:

I like to start with the positives and there are a few. First, the experiments are varied and interesting. Clearly some effort went into programming and running all of the experiments (there is a huge problem with lack of comparison to competitive approaches, but, we’ll get to that in a bit). Additionally, the identification and use of the sequence memoizer (SM) in this context is creative and commendable. That literature is dense; wading through it requires no small degree of patience and diligence. Additionally, the review of CRF’s and the review of the SM were reasonable; the former more so than the latter.

That said, I found the paper relatively far below the bar for TPAMI. My biggest criticisms are 1) The training methodology; particularly how the SM was trained, where the mean field form comes from, and how the EM-like prediction model work were explained poorly, not justified, and as a result were very unclear. 2) The experimental results included no baseline against which to judge the performance of the SMCRF. Without that, the results cannot be interpreted.

In particular, Section 4, where the proposed approach is spelled out; specifically in Eqn. 29 ( \phi\_t^2(y\_t | …) = \lambda log p(y\_t | …, M) ), the authors state that “M is a sequence memoizer postulated to model the prior probabilities of state transition in the modeled datasets, p(y\_t | .., M) is the predictive density of the SM model M. These statements open up many questions which are not directly addressed by the authors: 1) Where does M come from? Pre-training? What does M consist of? A setting of the SM’s parameters? An expectation? What is the “predictive density” of the SM model M? An expectation? A MAP distribution? If the SM is pre-trained on out-of-sample data and M is fixed, that’s weird, but, ultimately OK and potentially interesting. Even in that case, though, the predictive density is \_random\_ in the sequence memoizer.

The part of the paper concerning model training is extremely confusing. “However, a tractable approximation can be obtained by exploiting the mean-field approximation principle.” Fine. Does this mean that you’re doing mean-field inference for the sequence memoizer? I’d like to see how this is done if you’re doing it because I’ve not seen it in the literature. If you aren’t, how is the sequence memoizer being trained? Mean field in the SM would require identifying and splitting dependencies between instantiated distributions. Summing over all $y$ path’s means that all distributions (an exponential number) have to be instantiated. Further, even computing $\phi\_t(y\_t, y\_{t-1}$ (as mentioned before) requires taking an \_expectation\_ in the sequence memoizer. The only way this makes sense is if the SM is pre-trained and fixed and the predictive distribution in a particular context is interpreted as a single sample estimate of the posterior distribution of the predictive distribution. Is that right? If so, it’s not clear from the paper. If it’s not, it’s not clear from the paper.

The decoding section of the paper suffers from a similar set of problems. The mean field approximation isn’t particularly well explained and there is either conceptual error about what the SM does or the explanation of the mean field approach is so devoid of detail that no one will be able to understand what’s going on.

The experiment section is extremely enticing but ultimately very frustrating. My first criticism is the arbitrary chose of discretization in the pose estimation experiments. My second, and the most troubling aspect of the paper overall, is the almost total lack of comparative methodology against which to measure the performance gains. In the case of joint angle estimation, why, for instance, could a Kalman smoother not be utilized, with, for comparison purposes, it’s output discretized into the same set of values as used in the paper? In the case of handwriting recognition, why not provide a neural net or deep belief net comparison? In the absence of such baseline comparisons the work is very difficult to judge.

Summary:

Experiments lack baselines; importance and significance impossible to judge.

Model training and prediction insufficiently clear; correctness impossible to judge.