Neural Hawkes Process Memory

Michael C. Mozer
University of Colorado, Boulder

Robert V. Lindsey
Imagen Technologies

Denis Kazakov
University of Colorado, Boulder
Predicting Student Knowledge
Massed study

Memory decays more slowly with spaced study

Memory decays faster with massed study
Product Recommendation
Time Scale and Temporal Distribution of Behavior

Critical for modeling and predicting many human activities:

- Retrieving from memory
- Purchasing products
- Selecting music
- Making restaurant reservations
- Posting on web forums and social media
- Gaming online
- Engaging in criminal activities
- Sending email and texts
Recent Research Involving Temporally Situated Events

Discretize time and use tensor factorization or RNNs
- e.g., X. Wang et al. (2016), Y Song et al. (2016), Neil et al. (2016)

Hidden semi-Markov models and survival analysis
- Kapoor et al. (2014, 2015)

Include time tags as RNN inputs and treat as sequence processing task
- Du et al. (2016)

Temporal point processes
- Du et al. (2015), Y. Wang et al. (2015, 2016)

Our approach
- incorporate time into the RNN dynamics
Temporal Point Processes

- Produces sequence of event times $\mathcal{T} = \{t_i\}$

- Characterized by conditional intensity function, $h(t)$
  
  $$h(t) = Pr(\text{event in interval } dt \mid \mathcal{T})/dt$$

- E.g., Poisson process
  - Constant intensity function
    $$h(t) = \mu$$

- E.g., inhomogeneous Poisson process
  - Time varying intensity
Hawkes Process

Intensity depends on event history
- Self excitatory
- Decaying

Intensity decays over time
- Used to model earthquakes, financial transactions, crimes
- Decay rate determines time scale of persistence

\[ h(t) \]

intensity

event stream

Time
Hawkes Process

Conditional intensity function

\[ h(t) = \mu + \alpha \sum_{t_j < t} e^{-\gamma(t-t_j)} \]

with \( T \equiv \{ t_1, \ldots, t_j, \ldots \} \) times of past events

Incremental formulation with discrete updates

\[ h_0 = \mu \text{ and } t_0 = 0 \]

\[ h_k = \mu + e^{-\gamma \Delta t_k} (h_{k-1} - \mu) + \alpha x_k \]

\( \Delta t_k \equiv t_k - t_{k-1} \)

\[ x_k = \begin{cases} 1 & \text{if event occurs} \\ 0 & \text{if no event} \end{cases} \]
Prediction

Observe a time series and predict what comes next?

Given model parameters, compute intensity from observations:

\[ h_k = \mu + e^{-\gamma \Delta t_k} (h_{k-1} - \mu) + \alpha x_k \]

Given intensity, compute event likelihood in a \( \Delta t \) window:

\[ \Pr(t_k \leq t_{k-1} + \Delta t, x_k = 1|t_1, \ldots, t_{k-1}) = 1 - e^{-(h_{k-1} - \mu)(\gamma - \mu \Delta t)}/\gamma - \mu \Delta t \equiv Z_k(\Delta t) \]
**Key Premise**

The time scale for an event type may vary from sequence to sequence.

Therefore, we want to infer time scale parameter $\gamma$ appropriate for each event and for each sequence.
Bayesian Inference of Time Scale

Treat $\gamma$ as a discrete random variable to be inferred from observations.

- $\gamma \in \{\gamma_1, \gamma_2, ..., \gamma_S\}$ where $S$ is the number of candidate scales
- log-linear scale to cover large dynamic range

Specify prior on $\gamma$

- $\Pr(\gamma = \gamma_i)$

Given next event $x_k$ (present or absent) at $t_k$, perform Bayesian update:

- $\Pr(\gamma_i | x_{1:k}, t_{1:k}) \sim p(x_k, t_k | x_{1:k-1}, t_{1:k-1}, \gamma_i) \Pr(\gamma_i | x_{1:k-1}, t_{1:k-1})$

$$\left(\mu + e^{-\gamma_i \Delta t_k} (h_{k-1,i} - \mu)\right)^x \frac{Z_{ki}}{Z_{k+1}}$$
induction of time scale

input sequences

recovered intensity function

intensity function from generative process
Effect of Spacing

![Graph showing the effect of spacing on a logarithmic scale over time.](image)

- The graph displays various spacing configurations over time, with a logarithmic scale for intensity.
- The intensity values range from $10^{-2}$ to $10^5$.
Human data
Neural net model related to HPM
Mozer, Pashler, Cepeda, Lindsey, & Vul (2009)
Two Alternative Characterizations of Environment

All events are observed

- e.g., students practicing retrieval of foreign language vocabulary
- Likelihood function should reflect absence of events between inputs

\[
\Pr(y_i|\mathbf{x}_{1:k}, t_{1:k}) \sim p(x_k, t_k|\mathbf{x}_{1:k-1}, t_{1:k-1}, y_i) \Pr(y_i|\mathbf{x}_{1:k-1}, t_{1:k-1})
\]

\[
\left(\mu + e^{-\gamma_i \Delta t_k} (h_{k-1,i} - \mu)\right)^x Z_{ki}(\Delta t_k)
\]

Some events are unobserved

- e.g., shoppers making purchases on amazon.com (but purchases also made on target.com and jet.com)
- Likelihood function should marginalize over unobserved events and reflect the expected intensity

\[
\Pr(y_i|\mathbf{x}_{1:k}, t_{1:k}) \sim p(x_k, t_k|\mathbf{x}_{1:k-1}, t_{1:k-1}, y_i) \Pr(y_i|\mathbf{x}_{1:k-1}, t_{1:k-1})
\]

\[
\left(\frac{\mu}{1 - \alpha / \gamma_i} + \left(\frac{\mu}{1 - \alpha / \gamma_i} - \mu\right) e^{-\gamma_i (1 - \alpha / \gamma_i) \Delta t_k}\right)^x
\]
Hawkes Process Memory (HPM) Unit

Holds a history of past inputs (events)
Memory persistence depends on input history
No explicit ‘input’ or ‘forget’ gate
Captures continuous time dynamics
Embedding HPM in an RNN

Because event representations are learned, input $x$ denotes $\text{Pr}(\text{event})$ rather than truth value

- Activation dynamics are a mean field approximation to HP inference

  Marginalizing over belief about event occurrence:

  $$\Pr(\gamma_i | x_{1:t}, t_{1:k}) \sim \sum_{x_k=0}^{x_k=1} p(x_k, t_k | x_{1:k-1}, t_{1:k-1}, \gamma_i) \Pr(\gamma_i | x_{1:k-1}, t_{1:k-1})$$

Output (to next layer and through recurrent connections) must be bounded.

- Quasi-hyperbolic function

  $$h(t + \Delta \hat{t})/(h(t + \Delta \hat{t}) + \nu)$$
Generic LSTM RNN

predicted event

current event
current event \( \Delta \hat{t} \) \( \Delta t \) A B C ... 

predicted event A B C ...
Word Learning Study (Kang et al., 2014)

Data

- 32 human subjects
- 60 Japanese-English word associations
- each association tested 4-13 times over intervals ranging from minutes to several months
- 655 trials per sequence on average

Task

- Given study history up to trial t, predict accuracy (retrievable from memory or not) for next trial.
# Word Learning Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority class (correct)</td>
<td>62.7%</td>
</tr>
<tr>
<td>Traditional Hawkes process</td>
<td>64.6%</td>
</tr>
<tr>
<td>Next = previous</td>
<td>71.3%</td>
</tr>
<tr>
<td>LSTM</td>
<td>77.9%</td>
</tr>
<tr>
<td>HPM</td>
<td>78.3%</td>
</tr>
<tr>
<td>LSTM with Δt inputs</td>
<td>78.3%</td>
</tr>
</tbody>
</table>
Reddit Postings
Reddit

Data

- 30,733 users
- 32-1024 posts per user
- 1-50 forums
- 15,000 users for training (and validation), remainder testing

Task

- Predict forum to which user will post next, given the time of the posting
<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Next = previous</td>
<td>39.7%</td>
</tr>
<tr>
<td>Hawkes Process</td>
<td>44.8%</td>
</tr>
<tr>
<td>HPM</td>
<td>53.6%</td>
</tr>
<tr>
<td>LSTM (with $\Delta t$ inputs)</td>
<td>53.5%</td>
</tr>
<tr>
<td>LSTM, no input or forget gate</td>
<td>51.1%</td>
</tr>
</tbody>
</table>
Human behavior and preferences have dynamics that operate across a range of time scales.

It seems like a model based on these dynamics should be a good predictor of human behavior.

... and hopefully also a good predictor of other multiscale time series.
Key Idea of Hawkes Process Memory

Represent memory of sequences at multiple time scales simultaneously

Output ‘appropriate’ time scale based on input history
State Of The Research

LSTM is pretty darn robust

Some evidence that HPM and LSTM are picking up on distinct information in the sequences

- Possibility that mixing unit types will obtain benefits
- Can also consider other types of units premised on alternative temporal point processes, e.g., self correcting processes

Potential for using event-based model even for traditional sequence processing tasks
Novelty Of Approach

The neural Hawkes process memory belongs to two new classes of neural net models that are emerging.

- Models that perform dynamic parameter inference as a sequence is processed (vs. stochastic gradient based adaptation)
  

- Models that operate in a continuous time environment
  
  see also *Phased LSTM* paper by Neil, Pfeiffer, Liu (2016)