Latent LSTM Allocation

Manzil Zaheer, Amr Ahmed and Alexander J Smola

Presented by Akshay Budhkar & Krishnapriya Vishnubhotla

March 3, 2018
Outline

1 Introduction
   - Latent Dirichlet Allocation
   - LSTMs

2 Latent LSTM Allocation
   - Algorithm
   - Inference
   - Different Models

3 Results

4 Conclusion
Latent Dirichlet Allocation

1) Draw each topic $\beta_k \sim \text{Dirichlet}(\eta)$

2) For each document:
   1) Draw topic proportions $\theta_d \sim \text{Dirichlet}(\alpha)$
   2) For each word:
      1) Draw $z_{di} \sim \text{Mult}(\theta_d)$
      2) Draw $\omega_{di} \sim \text{Mult}(\beta_{z_{di}})$

- Probabilistic graphical model
- Not sequential, but easily interpretable.
LSTMs

- Good for modeling sequential data, preserves temporal aspect
- Too many parameters
- Hard to interpret
Latent LSTM Allocation (LLA) - Algorithm

1. for $k = 1$ to $K$
   (a) Choose topic $\phi_k \sim \text{Dir}(\beta)$

2. for each document $d$ in corpus $\mathcal{D}$
   (a) Initialize LSTM with $s_0 = 0$
   (b) for each word index $t$ from 1 to $N_d$
       i. Update $s_t = \text{LSTM}(z_{d,t-1}, s_{t-1})$
       ii. Get topic proportions at time $t$ from the
           LSTM state, $\theta = \text{softmax}_K(W_p s_t + b_p)$
       iii. Choose a topic $z_{d,t} \sim \text{Categorical}(\theta)$
       iv. Choose word $w_{d,t} \sim \text{Categorical}(\phi_{z_{d,t}})$
Graphical model for LLA
Marginal probability of observing a document is

\[
p(w_d|LSTM, \phi) = \sum_{z_d} p(w_d, z_d|LSTM, \phi) \\
= \sum_{z_d} \prod_{t} p(w_{d,t}|z_{d,t}; \phi)p(z_{d,t}|z_{d,1:t-1}; LSTM)
\]  

(1)

- Uses a $K \times H$ dense matrix and a $V \times K$ sparse matrix.
Sparse  
Interpretable  
Low predictive power  
(a) LDA

Dense  
Un-interpretale  
High predictive power  
(b) LSTM

Dense  
Interpretable  
High predictive power  
(c) LLA

Sparse
Inference

- Stochastic Expectation Maximization is used to compute the posterior.
- The Evidence Lower Bound (ELBO) can be written as:

\[
\sum_d \log p(w_d | LSTM, \phi) \\
\geq \sum_d \sum z_d q(z) \log \frac{p(z_d; LSTM) \prod_t p(w_{d,t} | z_{d,t}; \phi)}{q(z_d)}
\]  

(2)

- Conditional probability of topic at time step \( t \) is:

\[
p(z_{d,t} = k | w_{d,t}, z_{d,1:t-1}; LSTM, \phi) \\
\propto p(z_{d,t} = k | z_{d,1:t}; LSTM) p(w_{d,t} | z_{d,t} = k; \phi)
\]  

(3)

- And

\[
p(w_{d,t} | z_{d,t} = k; \phi) = \phi_{w,k} = \frac{n_{w,k} + \beta}{n_k + \sqrt{V^2}}
\]  

(4)
Algorithm 1 Stochastic EM for LLA

**Input:** Document corpus $D$.

1. Initialize $\phi$ and LSTM with a few iterations of LDA
2. repeat
   **SE-Step:**
   3. for all document $d \in D$ in parallel do
      4. for $t \leftarrow 1$ to $N_d$ (possibly with padding) do
         5. $\forall k \in \{1, \ldots, K\}$, i.e., for every topic index
            obtain by LSTM forward pass:
            $\pi_k = \phi_{w_d,t,k} p(z_{d,t} = k | z_{d,1:t-1}; \text{LSTM})$.
            Sample $z_{d,t} \sim \text{Categorical}(\pi)$
      6. end for
   7. end for
   **M-Step:**
   8. Collect sufficient statistics to obtain:
      $\phi_{w,k} = \frac{n_{w,k} + \beta}{n_k + V \beta}, \ \forall w, k$
   9. for mini-batch of documents $B \subset D$ do
      10. Compute the gradient by LSTM backward pass
          $\frac{\partial L}{\partial \text{LSTM}} = \sum_{d \in B} \sum_{t=1}^{N_d} \frac{\partial \log p(z_{d,t} | z_{d,1:t-1}; \text{LSTM})}{\partial \text{LSTM}}$
      11. Update LSTM parameters by stochastic gradient descent methods such as Adam (Kingma & Ba, 2014).
      12. end for
   13. until Convergence
Mathematical Intuition

- **LDA**

\[
\log p(w) = \sum_t \log p(w_t | model) \\
= \sum_t \log \sum_{z_t} p(w_t | z_t) p(z_t | doc)
\] (5)

- **LSTM**

\[
\log p(w) = \sum_t \log p(w_t | w_{t-1}, w_{t-2}, \ldots, w_1)
\] (6)

- **LLA**

\[
\log p(w) = \log \sum_{z_1:T} \prod_t p(w_t | z_t) p(z_t | z_{t-1}, z_{t-2}, \ldots, z_1)
\] (7)
Different Models

(a) Topic LLA

(b) Word LLA

(c) Char LLA
Perplexity vs. Number of topics (Wikipedia)

Latent LSTM Allocation

March 3, 2018 13 / 22
Cannot use Char LLA, since URLs lack morphological structure
### Interpreting Cleaner Topics

<table>
<thead>
<tr>
<th>Method</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LDA</strong></td>
<td>foundation, <strong>iowa</strong>, charity, fund, money, campaign, raised, donated, funds, donations, raise, support, charitable, <strong>million</strong>, donation</td>
</tr>
<tr>
<td><strong>LLA</strong></td>
<td>fund, foundation, money, funds, support, charity, funding, donations, campaign, raised, donated, <strong>trust</strong>, raising, <strong>contributions</strong>, <strong>awareness</strong></td>
</tr>
</tbody>
</table>
### Interpreting Factored Topics

<table>
<thead>
<tr>
<th>LDA</th>
<th>strike, strikes, striking, miners, strikers, workers workers, day, began, general, called, pinkerton, action, hour, hunger, keefe</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLA</td>
<td>union, unions, strike, workers, labor, federation, trade, afl, bargaining, cio, organization, relations, strikes, national, industrial</td>
</tr>
<tr>
<td></td>
<td>mining, coal, mine, mines, gold, ore, miners, copper, iron, rush, silver, mineral, deposits, minerals, mined</td>
</tr>
</tbody>
</table>
Convergence Speed

![Convergence Speed Graph]

- LDA 1000
- Topic LLA 1000
- Word LLA 1000
- Char LLA 1000

- Word LSTM
- Char LSTM
- Topic LLA 1000
- Word LLA 1000
- Char LLA 1000

Perplexity vs Time [s]

Manzil Zaheer, Amr Ahmed and Alexander J

Latent LSTM Allocation

March 3, 2018 19 / 22
### Table 4. Advantage in terms of perplexity for joint learning.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Independent learner</th>
<th>Joint learner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia</td>
<td>2119</td>
<td>1785</td>
</tr>
<tr>
<td>User Click</td>
<td>1572</td>
<td>927</td>
</tr>
</tbody>
</table>
Final Thoughts

**Pros**
- Provides a knob for interpretability and accuracy
- Less number of parameters for a reasonable perplexity
- Cleaner factored topics

**Cons**
- Did not compare to something like hierarchical LDA
- Can’t use Char LLA for every problem
- Perplexity is not a good measure of text generation accuracy

