1. Introduction

Motivation
• State-of-the-art neural language models (LMs) represented by Transformers are highly complex.
• Fixed parameter estimates fail to account for model uncertainty.
• Prone to over-fitting when given limited training data.

Our work:
• Propose a full Bayesian learning framework to account for model uncertainty in Transformer LM estimation.
• Adopt efficient variational inference based approach to estimate the latent parameter posterior distribution.
• Detailed analysis on the effect of applying Bayesian estimation on different parts of Transformer LM.

2. Transformer LMs

Decoder component of Transformer architecture was adopted for LM.
• Stacking of multi-head self-attention modules:
  \[ q_i, k_i, v_i = QX_i^{-1}, KX_i^{-1}, VX_i^{-1} \]
  \[ h_i = \left( h_i^1, \ldots, h_i^H \right) \]
  \[ y_i = W_S^{\text{SelfAttention}}(h_i^l, q_i) + x_i^{-1} \]
• \( x_i \) denotes the input of the \( i \)-th Transformer block.
• \( h_i^l \) stores cached key-value pairs up to word position \( t \), enforcing left to right attention modelling over history contexts only.
• Feed forward blocks following each self-attention module:
  \[ s_i = W_E^{\text{GELU}}(W_i^{2}x_i^l) + x_i^{-1} \]
  \[ x_i^{l+1} = \text{LayerNorm}(s_i) \]
• For simplicity, the bias vectors are omitted in the above equations.

3. Bayesian Transformer LM

• Variational Bayesian for Transformer LMs:
  Lower bound is an approximation of marginal likelihood:
  \[ \log P(D) = \log \prod_{l} P(D|\theta_l)p(\theta_l) \]
  \[ \geq \sum_{l=1}^{L} \log P(W|\theta_l)q(\theta_l) - KL(q(\theta_l)||p(\theta_l)) = L \]

  \[ \mathcal{D} \]
  \[ \mathcal{L}_1 \]
  \[ \mathcal{L}_2 \]

• \( \mathcal{D} \) represents the whole training set for model development.
• \( q(\theta) \) denotes the variational approximation parameters.
• \( p(\theta_l) \) denotes the prior distribution of \( \theta_l \).
• Variational distribution parameters \( \mu, \sigma \) integrated with SGD based back propagation:
  \[ \frac{\partial L}{\partial \mu_l} = \frac{1}{K} \sum_{k=1}^{K} \frac{\partial L}{\partial \mu_l} \mu_l - \mu_{y_{l,t}} \]
  \[ \frac{\partial L}{\partial \sigma_l} = \frac{1}{K} \sum_{k=1}^{K} \frac{\partial L}{\partial \sigma_l} \sigma_l^2 - \sigma_{l,1}^2 \]

• Implementation details:
  • Applying Bayesian estimation on the parameter model parameters.
  • Parameters obtained from standard Transformer LM is used as the prior’s mean \( \mu_l \), prior’s variance is set to be \( 1 \).
  • Only use the mean of the Bayesian parameters in evaluation

4. Experiments & Results

Experiments on Conversational Telephone Speech
• Datasets: 300 hour Switchboard for acoustic modelling; 34M words of Switchboard+Fisher transcriptions for language modelling; 30k vocabulary lexicon.
• Acoustic model: TDDN-F based hybrid model featuring speech perturbation, i-Vector, LfUc speaker adaptation and (LF-MM1) sequence training.

<table>
<thead>
<tr>
<th>ID</th>
<th>LM</th>
<th>Bayesian Block</th>
<th>PPL (swbd)</th>
<th>eval2000 (swbd)</th>
<th>m2 (swbd)</th>
<th>m3 (swbd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4gram Transformer (+g)</td>
<td>Not Applied</td>
<td>9.7</td>
<td>18.0</td>
<td>11.5</td>
<td>15.3</td>
</tr>
<tr>
<td>2</td>
<td>4gram Transformer (+g)</td>
<td>Not Applied</td>
<td>14.50</td>
<td>7.9</td>
<td>15.7</td>
<td>9.5</td>
</tr>
<tr>
<td>3</td>
<td>1 MHA</td>
<td>EMB</td>
<td>41.01</td>
<td>7.7</td>
<td>15.6</td>
<td>9.5</td>
</tr>
<tr>
<td>4</td>
<td>1 FF</td>
<td>MHA</td>
<td>40.95</td>
<td>7.7</td>
<td>15.5</td>
<td>9.5</td>
</tr>
<tr>
<td>5</td>
<td>2 FF</td>
<td>MHA</td>
<td>40.85</td>
<td>7.7</td>
<td>15.4</td>
<td>9.4</td>
</tr>
<tr>
<td>6</td>
<td>3 FF</td>
<td>MHA</td>
<td>41.11</td>
<td>7.7</td>
<td>15.6</td>
<td>9.5</td>
</tr>
<tr>
<td>7</td>
<td>4 FF</td>
<td>MHA</td>
<td>42.45</td>
<td>7.8</td>
<td>15.8</td>
<td>9.5</td>
</tr>
<tr>
<td>8</td>
<td>5 FF</td>
<td>MHA</td>
<td>47.54</td>
<td>8.0</td>
<td>16.0</td>
<td>9.9</td>
</tr>
<tr>
<td>9</td>
<td>6 FF</td>
<td>MHA</td>
<td>54.19</td>
<td>8.3</td>
<td>16.2</td>
<td>10.2</td>
</tr>
<tr>
<td>10</td>
<td>7 FF</td>
<td>MHA</td>
<td>74.50</td>
<td>8.9</td>
<td>17.3</td>
<td>10.8</td>
</tr>
<tr>
<td>11</td>
<td>8 FF</td>
<td>MHA</td>
<td>40.92</td>
<td>7.7</td>
<td>15.5</td>
<td>9.4</td>
</tr>
<tr>
<td>12</td>
<td>9 FF</td>
<td>MHA</td>
<td>39.70</td>
<td>7.6</td>
<td>15.4</td>
<td>9.3</td>
</tr>
<tr>
<td>13</td>
<td>10 FF</td>
<td>MHA</td>
<td>39.42</td>
<td>7.6</td>
<td>15.4</td>
<td>9.3</td>
</tr>
</tbody>
</table>

• Proposed Bayesian Transformer LMs (line 11-13) outperform the baseline Transformer LM (line 2) in terms of both PPL and WER by statistically significant margin from 0.3% to 0.5% absolutely.
• Applying Bayesian estimation on the feed forward (FF) module outperforms using Bayesian estimation on multi-head self-attention (MHA) or embedding (EMB) layer.

5. Conclusions

• Performance improvements consistently observed across a domain independent adaptation task requiring porting a Transformer LM trained on the Switchboard and Fisher data to a low-resource DementiaBank elderly speech corpus.

• The proposed Bayesian learning framework can improve the performance and robustness of Transformers in both model training and adaptation.
• The parameters associated with the higher Transformer blocks are expected to be more deterministic than those experienced in the lower

This research is supported by Hong Kong RGC grant No. 14200218, 14200220, Theme-based Research Scheme T45-407/19N, Innovation & Technology Fund grant No. ITS/254/19, and Shun Hing Institute of Advanced Engineering grant No. MMT-p1-19.