Learning to Generate Textual Data

Guillaume Bouchard∗
University College London
Bloomsbury AI
g.bouchard@cs.ucl.ac.uk

Pontus Stenetorp∗
University College London
p.stenetorp@cs.ucl.ac.uk

Sebastian Riedel
University College London
s.riedel@cs.ucl.ac.uk

Vision

- Big small data, the next big thing?
- Regularise
  - Model shrinking (underfit)
  - Model bias (less general)
- Virtual samples (Niyogi et al., 1998)
- Generation, but what and how much?
- “Learning to Teach”
  - Generic model
  - Bias through learned generation

Generative Regularisation (GeneRe)

Require: \( \hat{P} \): real data sampler
Require: \( P_r \): parametric data generator
Require: \( \lambda \): generative regularization strength
Require: \( \eta \): learning rate
Require: \( \alpha \): baseline smoothing coefficient
1: Initialize parameters \( \theta \), sampling coefficients \( \gamma \) and baseline \( \mu \)
2: for \( t = 1, 2, \ldots \) do
3: \( x, y \sim \frac{1}{\alpha \lambda} \hat{P} + \frac{1}{\lambda} P_r \)
4: \( g_\theta \leftarrow \nabla_\theta \log p_r(y|x) \)
5: \( g_\gamma \leftarrow \nabla_\gamma \log p_r(x, y) \)
6: \( \theta, \gamma \leftarrow (\theta, \gamma) - \eta (g_\theta, g_\gamma) \)
7: \( \mu \leftarrow \alpha \mu + (1 - \alpha) \log p_r(y|x) \)
8: end for

• Variant of Reinforce (Williams, 1988)

Math Word Problems

Example:
“A pet store had 81 puppies. In one day they sold 41 of them and put the rest into cages with 8 in each cage. How many cages did they use?”

Template:
“S1 V1 Q1 O1. C1 S1(pronoun) V2 Q2 of O1(pronoun) and V2 the rest into O3(plural) with Q3 in each O3. How many O3(plural) V3?”

Optimal \( \gamma \) proportions

Results

<table>
<thead>
<tr>
<th></th>
<th>A2</th>
<th>IL</th>
<th>CC</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>RR2015</td>
<td>82.4</td>
<td>75.4</td>
<td>59.5</td>
<td>71.1</td>
</tr>
<tr>
<td>100% Data</td>
<td>72.5</td>
<td>53.7</td>
<td>95.0</td>
<td>73.7</td>
</tr>
<tr>
<td>100% Gen</td>
<td>60.3</td>
<td>51.2</td>
<td>92.0</td>
<td>67.8</td>
</tr>
<tr>
<td>85%Gen + 15%Data</td>
<td>74.0</td>
<td>55.4</td>
<td>97.5</td>
<td>75.6</td>
</tr>
<tr>
<td>GeneRe</td>
<td>77.9</td>
<td>56.7</td>
<td>98.5</td>
<td>77.7</td>
</tr>
</tbody>
</table>

• Model details
  - 1-layer CRU with 256-dimensional hidden state
  - Projected 300 dimensional pre-trained vectors

• Using only non-generated or generated is insufficient

• Mixing generated and non-generated helps

• Using GeneRe improves the results even further