Adaptive Knowledge Sharing in Multi-Task Learning: Improving Low-Resource Neural Machine Translation

Poorya ZareMoodi, Wray Buntine, Gholamreza (Reza) Haffari
Monash University
Roadmap

- Introduction & background
- Adaptive knowledge sharing in Multi-Task Learning
- Experiments & analysis
- Conclusion
Improving NMT in low-Resource scenarios

• NMT is notorious!

• Bilingually low-resource scenario: large amounts of bilingual training data is not available

• IDEA: Use existing resources from other tasks and train one model for all tasks using multi-task learning

• This effectively injects inductive biases to help improving the generalisation of NMT

• Auxiliary tasks: Semantic Parsing, Syntactic Parsing, Named Entity Recognition
Encoders-Decoders for Individual Tasks

**Machine Translation**
- **I went home**
- **Obama was elected and his voter celebrated**
- **The burglar robbed the apartment**
- **Jim bought 300 shares of Acme Corp. in 2006**

**Semantic Parsing**
- **I went home**
- **Obama was elected and his voter celebrated**
- **The burglar robbed the apartment**
- **Jim bought 300 shares of Acme Corp. in 2006**

**Syntactic Parsing**

**Named-Entity Recognition**

**B-PER 0 0 0 B-ORG I-ORG 0 B-MISC**
Sharing Scenario

Multitask seq2seq

sentence

task tag

Machine Translation
Encoder
Decoder

Semantic Parsing
Encoder
Decoder

Syntactic Parsing
Encoder
Decoder

Named-Entity Recognition
Encoder
Decoder

translation
Named-entities
Parse tree
Semantic graph
Partial Parameter Sharing

I went home

<translation> I went home

Encoder \rightarrow \text{Decoder}

Task interference

Context

<translation> من به خانه رفتم

Zaremoodi & Haffari, NAACL, 2018
Roadmap

• Introduction & Background
• Adaptive knowledge sharing in Multi-Task Learning
• Experiments & analysis
• Conclusion
Adaptive Knowledge Sharing in MTL

- Sharing the parameters of the recurrent units among all tasks
  - Task interference
  - Inability to leverage commonalities among subsets of tasks

IDEA

- Multiple experts in handling different kinds of information
- Adaptively share experts among the tasks

sharing the knowledge for controlling the information flow in the hidden states
Adaptive Knowledge Sharing in MTL

**IDEA**
- Multiple experts in handling different kinds of information
- Adaptively share experts among the tasks
- Extend the recurrent units with multiple blocks
  - each block has its own information flow through the time
  - *Routing mechanism*: to softly direct the input to these blocks
Adaptive Knowledge Sharing

Routing:

\[ s_t = \tanh(W_x \cdot x_t + W_h \cdot h_{t-1} + b_s), \]
\[ \tau_t = \text{softmax}(W_r \cdot s_t + b_r), \]

\[ \tilde{x}_t^{(i)} = \tau_t[i] \cdot x_t \quad \Rightarrow \quad h_t^{(\text{shared})} = \sum_{i=1}^{n} \tau_t[i] h_t^{(i)} \quad \Rightarrow \quad h_t = [h_t^{(\text{shared})}; h_t^{(\text{task})}] \]

Blocks:

\[ z_t^{(i)} = \sigma(W_{z_t}^{(i)} \tilde{x}_t^{(i)} + U_{z_t}^{(i)} h_{t-1}^{(i)} + b_{z_t}^{(i)}), \]
\[ r_t^{(i)} = \sigma(W_{r_t}^{(i)} \tilde{x}_t^{(i)} + U_{r_t}^{(i)} h_{t-1}^{(i)} + b_{r_t}^{(i)}), \]

\[ \tilde{h}_t^{(i)} = \tanh(W_h^{(i)} \tilde{x}_t^{(i)} + U_h^{(i)} h_{t-1}^{(i)} + b_h^{(i)}), \]
\[ h_t^{(i)} = z_t^{(i)} \odot h_{t-1}^{(i)} + (1 - z_t^{(i)}) \odot \tilde{h}_t^{(i)}. \]
Adaptive Knowledge Sharing

We use the proposed recurrent unit inside encoder and decoder.
Roadmap

• Introduction & background
• Adaptive knowledge sharing in Multi-Task Learning
• Experiments & analysis
• Conclusion
Experiments

• **Language Pairs**: English to Farsi/Vietnamese

• **Datasets**:
  - English to Farsi: TED corpus & LDC2016E93
  - English to Vietnamese: IWSLT 2015 (TED and TEDX talks)
  - Semantic parsing: AMR corpus (newswire, weblogs, web discussion forums and broadcast conversations)
  - Syntactic parsing: Penn Treebank
  - NER: CONLL NER Corpus (newswire articles from the Reuters Corpus)

• **NMT Architecture**: GRU for blocks, 400 RNN hidden states and word embedding

• **NMT best practice**:
  - Optimisation: Adam
  - Byte Pair Encoding (BPE) on both source/target
  - Evaluation metrics: PPL, TER and BLEU

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>En → Fa</td>
<td>98,158</td>
<td>3,000</td>
<td>4,000</td>
</tr>
<tr>
<td>En → vi</td>
<td>133,290</td>
<td>1,553</td>
<td>1,268</td>
</tr>
</tbody>
</table>
Experiments

BLEU

English ➔ Farsi

English ➔ Vietnamese
Experiments (English to Farsi)

- Average block usage.
- Blocks specialisation: Block 1: MT, Semantic Parsing, Block 2: Syntactic/Semantic Parsing, Block 3: NER
Conclusion

- Address the task interference issue in MTL
  - extending the recurrent units with multiple blocks
  - with a trainable routing network
Questions?