SIMPLE, SCALABLE ADAPTATION FOR NEURAL MACHINE TRANSLATION

Ankur Bapna, Naveen Arivazhagan, Orhan Firat
OUTLINE

NMT BASICS

DOMAIN SPECIFIC ADAPTERS

MASSIVELY MULTILINGUAL MACHINE TRANSLATION
NMT BASICS
NEURAL MACHINE TRANSLATION

Jsem student  $\rightarrow$  SEQUENCE TO SEQUENCE MODEL  $\rightarrow$  I am a student

Jsem student  $\rightarrow$  ENCODER  $\rightarrow$  DECODER  $\rightarrow$  I am a student
ATTENTION AND TRANSFORMERS

Attention computation

Multiplying 'value'

Attention weights for 'This'

Key
Value
Query

This
This
This
This
is
is
attention
attention
Translation
*cat cat cat cat cat cat*

Reference
*the cat is on the mat*

Unigram precision = 1
Bigram precision = 0
STATE OF THE ART

Dominant approach
Fine-tuning a base model for a specific language pair or a domain

Problems
• Expensive
• Low resource language performance
• Not universal

References
- Effective Approaches to Attention-based Neural Machine Translation -- Luong and Manning, 2015
- Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation -- Wu et al., 2016
- Rapid Adaptation of Neural Machine Translation to New Languages -- Neubig and Hu, 2018
- Learning Hidden Unit Contribution for Adapting Neural Machine Translation Models -- David Vilar, 2018
DOMAIN SPECIFIC ADAPTERS
**APPROACH**

1. Take a standard trained NMT model
2. Freeze all parameters
3. Add a light-weight adapter layer for each task
4. Fine-tune each adapter for its task
MODEL ARCHITECTURE

Encoder
- Feed forward
- Self attention

Decoder
- Feed forward
- Cross attention
- Self attention

I am a student

Jsem student

Up projection
- RELU
- Down projection
- Layer norm
DATASETS

- **Workshop on Statistical Machine Translation ‘14 (WMT)**
  - Source training dataset
  - 36M pairs

- **The International Conference on Spoken Language Translation ‘15 (IWSLT)**
  - Adapter training dataset
  - 247k pairs

- **JRC Acquis**
  - Validation and test dataset
  - 11k pairs
## PERFORMANCE

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Base</th>
<th>Fine-Tune</th>
<th>LHUC</th>
<th>Adapter</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT '14</td>
<td>42.80</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IWSLT '15</td>
<td>41.33</td>
<td><strong>44.59</strong></td>
<td>43.33</td>
<td><strong>44.63</strong></td>
</tr>
<tr>
<td>JRC</td>
<td>54.60</td>
<td><strong>64.13</strong></td>
<td>57.10</td>
<td>63.48</td>
</tr>
</tbody>
</table>

Results are BLEU score multiplied by 100
MASSIVELY MULTILINGUAL MACHINE TRANSLATION
APPROACH

- Global training
  - Fully shared model transfer to low resource languages
- Refinement
  - Fine-tuning adapters for high resource languages
METHODOLOGY

- Single Transformer Big
- 102 to and from English
- Same hyper-parameters as bilingual
- Fine-tune for each pair
From left to right, languages are arranged in decreasing order of available training data.
Any-to-English Translation performance for multilingual models with adapters

From left to right, languages are arranged in decreasing order of available training data.

BLEU score increase

From left to right: Bilingual Baselines, Multilingual, Adapters, Adapters-Large.
CONCLUSION

Proposed light-weight adapters

Evaluated on domain adaptation and MMMT

Achieved similar or better scores in low resource language translation
THANK YOU FOR YOUR ATTENTION