Input Combination Strategies for Multi-Source Transformer Decoder

Jindřich Libovický, Jindřich Helcl, David Mareček

November 1, 2018
1. Transformer decoder overview
2. Input combination strategies
3. Experiments
   - Multimodal translation
   - Multi-source translation
1. **Transformer decoder overview**

2. Input combination strategies

3. Experiments
   - Multimodal translation
   - Multi-source translation
• Architecture for sequence-to-sequence learning
• Encoder and decoder part
• Consists of attention and feed-forward layers only
Encoder-Decoder Attention

Scaled dot-product attention:

\[ A(Q, K, V) = \text{softmax} \left( \frac{QK^\top}{\sqrt{d}} \right) V. \]

Multi-headed setup:

\[ A^h(Q, K, V) = \sum_{i=1}^{h} C_i W_i^O \]
\[ C_i = A(QW_i^Q, KW_i^K, VW_i^V) \]
\[ W^Q, W^K, W^V \in \mathbb{R}^{d \times d_h} \text{ trainable} \]
Overview

1. Transformer decoder overview
2. **Input combination strategies**
3. Experiments
   - Multimodal translation
   - Multi-source translation
Stack the layers after each other.
Run attentions independently, sum up the outputs.

$$\mathcal{A}_{\text{para}}^h(Q, K_{1:n}, V_{1:n}) = \sum_{i=1}^{n} \mathcal{A}^h(Q, K_i, V_i)$$
Run the attentions independently, put another attention layer on top.

\[ K_{\text{hier}} = V_{\text{hier}} = \text{concat}_i(\mathcal{A}^h(Q, K_i, V_i)) \]

\[ \mathcal{A}^h_{\text{hier}}(Q, K_{1:n}, V_{1:n}) = \mathcal{A}^h(Q, K_{\text{hier}}, V_{\text{hier}}) \]
Concatenate the input states, then run a single attention layer.

\[
K_{\text{flat}} = V_{\text{flat}} = \text{concat}_i(K_i)
\]

\[
A_{\text{flat}}^h(Q, K_{1:n}, V_{1:n}) = A^h(Q, K_{\text{flat}}, V_{\text{flat}})
\]
Overview

1. Transformer decoder overview
2. Input combination strategies
3. Experiments
   - Multimodal translation
   - Multi-source translation
Multimodal Translation – Task Overview

- Translation of image captions from Flickr30k dataset
- Multi30k dataset: images with English captions, German, French and Czech translations

Source:
en: A boy in a red suit plays in the water.

targets:

de: Ein Junge in einem roten Badeanzug spielt im Wasser.

fr: Un garçon en maillot de bain rouge joue dans l’eau.

cs: Chlapec v červených plavkách si hraje ve vodě.
Multimodal Translation – Experiment Setup

- Model dimension 512
- 6 layers in both encoder and decoder
- Vocabulary of approx. 20k wordpieces
- Image representation: convolutional maps from ResNet
Multimodal Translation – Results

Quantitative results of the MMT experiments on the 2016 test set. Column ‘adv. BLEU’ is an adversarial evaluation with randomized image input.

<table>
<thead>
<tr>
<th></th>
<th>en→de</th>
<th></th>
<th>en→fr</th>
<th></th>
<th>en→cs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>adv. BLEU</td>
<td>BLEU</td>
<td>adv. BLEU</td>
<td>BLEU</td>
</tr>
<tr>
<td>baseline</td>
<td>38.3 ± .8</td>
<td>—</td>
<td>59.6 ± .9</td>
<td>—</td>
<td>30.9 ± .8</td>
</tr>
<tr>
<td>serial</td>
<td>38.7 ± .9</td>
<td>37.3 ± .6</td>
<td>60.8 ± .9</td>
<td>58.9 ± .9</td>
<td>31.0 ± .8</td>
</tr>
<tr>
<td>parallel</td>
<td>38.6 ± .9</td>
<td>38.2 ± .8</td>
<td>60.2 ± .9</td>
<td>58.9 ± .9</td>
<td>31.1 ± .9</td>
</tr>
<tr>
<td>flat</td>
<td>37.1 ± .8</td>
<td>35.7 ± .8</td>
<td>58.0 ± .9</td>
<td>57.0 ± .9</td>
<td>29.9 ± .8</td>
</tr>
<tr>
<td>hierarchical</td>
<td>38.5 ± .8</td>
<td>38.1 ± .8</td>
<td>60.8 ± .9</td>
<td>60.2 ± .9</td>
<td>31.3 ± .9</td>
</tr>
</tbody>
</table>
Multimodal Translation – Learning Curves

![Multimodal Translation Learning Curves Graph]

Input Combination Strategies for Multi-Source Transformer Decoder

- flat
- hierarchical
- parallel
- serial
- RNN hierarchical
Multi-Source Translation – Task Overview

• Source languages: English, German, French, Spanish
• Target language: Czech
• Data: intersection of Europarl, 511k five-way parallel sentences
• Shared vocabulary of 42k wordpieces
• Model dimension 256, 6 layers in both encoder and decoder
Multi-Source Translation – Results

<table>
<thead>
<tr>
<th>Input Combination Strategies for Multi-Source Transformer Decoder</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>Adversarial evaluation (BLEU)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>en</td>
</tr>
<tr>
<td>baseline</td>
<td>16.5 ± .5</td>
<td>—</td>
</tr>
<tr>
<td>serial</td>
<td>20.5 ± .6</td>
<td>8.1 ± .4</td>
</tr>
<tr>
<td>parallel</td>
<td>20.5 ± .6</td>
<td>1.4 ± .2</td>
</tr>
<tr>
<td>flat</td>
<td>20.4 ± .6</td>
<td>0.2 ± .1</td>
</tr>
<tr>
<td>hierarchical</td>
<td>19.4 ± .5</td>
<td>4.2 ± .3</td>
</tr>
</tbody>
</table>

Quantitative results of the MMT experiment. The adversarial evaluation shows the BLEU score when one input language was changed randomly.
Multi-Source Translation – Learning Curves

Input Combination Strategies for Multi-Source Transformer Decoder
Visualization of attention for sentence *The Black Sea region, too, is of great importance.*
Language order in figures: es, fr, de, en
Conclusions

- Introduced 4 strategies: serial, parallel, hierarchical, flat
- All strategies perform approximately the same
- Slightly better than text-only baseline for multimodal MT
- Multi-source MT better than single-source

https://ufal.mff.cuni.cz