1. SUMMARY

Multi-layer encoders and decoders in NMT

• Vanilla NMT: Each layer has independent parameters

• RS-NMT: Use the same parameters for all layers

• Shared encoder-decoder NMT: Minimal NMT model

Key points

• Extensive experiments for several datasets
  – WMT: Turkish+English
  – ALT, GCP, KFTT, ASPEC: Japanese+English

• Model size reduction: 50-70%

• Acceptable loss of translation quality: 1–2 BLEU points

• Towards average attention?

• Calibration of attention

• 8 heads with different colors

• 6 layers × input tokens

• Darker colors
  → Stronger attention

• Calibration of attention by RS layers

Attention entropy

• 6 layers × 8 heads

• Darker colors
  → Higher Entropy
  → More uncertainty
  → Attend more words

  Towards average attention?

2. RECURRENTLY STACKED NMT

• Enable reuse of layer parameters for depth > 1
  – With modification of only 1 line of code

• What happens? Gradual calibration of attention (Panel 5)

• Extension: Encoder-decoder parameter sharing (Panel 4)

3. EXPERIMENTAL SETTINGS

Datasets

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Test</th>
<th>Val</th>
<th>Train</th>
<th>_vocab</th>
<th>#Params</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ja-En</td>
<td>18K</td>
<td>12K</td>
<td>15K</td>
<td>2000</td>
<td>1.50M</td>
</tr>
<tr>
<td>En-Ja</td>
<td>208K</td>
<td>160</td>
<td>200K</td>
<td>1000</td>
<td>26.32M</td>
</tr>
<tr>
<td>Tr-En</td>
<td>200K</td>
<td>15K</td>
<td>200K</td>
<td>32K</td>
<td>48.6M</td>
</tr>
</tbody>
</table>
| Ja-En         | 24.13 | 22.66 | 26.84 | 8 heads
| En-Ja         | 23.89 | 21.46 | 25.38 | 8 heads

Implementation and training details

• Implementation in tensor2tensor

  – Modification of Transformer (Vaswani+, 17)
  – Internal sub-word segmentation

• Vanilla NMT, RS-NMT, Shared encoder-decoder NMT

• Training iterations chosen based on time for convergence

• Default model and hyperparameter settings

• Last 10 checkpoints averaged for decoding

• Beam size of 4 and length penalty of 1.0 (En+Ja), 0.6 (En+Tr)

4. TRANSLATION PERFORMANCE IN BLEU

RS-NMT vs. vanilla NMT

• RS-NMT is only up to 2 BLEU behind vanilla NMT
  – Compensated by back-translation (Sennrich+, 16)

Shared encoder-decoder NMT

• 1 BLEU loss for RS-NMT / No BLEU loss for vanilla NMT

• Reduction of 70% parameters for acceptable BLEU reduction

5. VISUAL ANALYSIS OF RS-NMT ATTENTION

Transformer’s Cross-attention

• 8 heads with different colors

• 6 layers × input tokens

• Darker colors
  → Stronger attention

• Calibration of attention by RS layers

Attention entropy

• 6 layers × 8 heads

• Darker colors
  → Higher Entropy
  → More uncertainty
  → Attend more words

  Towards average attention?

6. FUTURE WORK

• Maximum compression with RS-NMT & knowledge distillation

• Enabling use of fewer levels of RS decoding

• Exploration of limits of RS layers

• Further analysis of the nature of RS-NMT

National Institute of Information and Communications Technology (NICT), Kyoto, Japan