Hint-Based Training for Non-Autoregressive Translation

Zhuohan Li  Zi Lin  Di He  Fei Tian  Tao Qin  Liwei Wang  Tie-Yan Liu

EMNLP-IJCNLP 2019
Autoregressive MT models

Encoder

Context

Multi-Head Self Attention

FFN

FFN

FFN

Multi-Head Self Attention

Decoder

Soft Max

Soft Max

Soft Max

Soft Max

FFN

FFN

FFN

FFN

Multi-Head Encoder-to-Decoder Attention

Masked Multi-head Self Attention

Emb

Emb

Emb

Emb

<sos>

Y₁

Y₂

Y₃

Y₄

x₁

x₂

x₃

Y₁

Y₂

Y₃

Hint-Based Training for Non-Autoregressive Translation
Non-autoregressive MT models

**Encoder**
- Context
  - Multi-Head Self Attention
  - FFN
  - Emb
  - \( x_1 \), \( x_2 \), \( x_3 \)

**Decoder**
- Multi-Head Encoder-to-Decoder Attention
- Multi-Head Positional Attention
- Multi-Head Self Attention
- FFN
- Emb
- \( y_1 \), \( y_2 \), \( y_3 \), \( y_4 \)
- Soft Max
- \( \times M \)

**Copy**
- \( \times N \)

Hint-Based Training for Non-Autoregressive Translation
Previous works on non-autoregressive MT

Encoder

- Fertilities [Gu et al.]
- Context
- Multi-Head Self Attention
- FFN
- FFN
- FFN
- Emb
- Emb
- Emb
- $x_1$
- $x_2$
- $x_3$

Decoder

- Multi-Head Encoder-to-Decoder Attention
- Multi-Head Positional Attention
- Multi-Head Self Attention
- FFN
- FFN
- FFN
- FFN
- Emb
- Emb
- Emb
- Emb
- $y_1$
- $y_2$
- $y_3$
- $y_4$

Fertilities [Gu et al.]

$\times M$

$\times N$

$\times R$ [Lee et al.]

Hint-Based Training for Non-Autoregressive Translation
Quality-speedup trade-off

Gu et al.
Kaiser et al.
Lee et al.
Autoregressive

Speedup
BLUE Score (WMT14 En-De)

Hint-Based Training for Non-Autoregressive Translation
Hidden states similarity

Hidden states cosine-similarity of a sampled sentence in IWSLT14 De-En.
Attention distribution

Encoder-to-decoder attention distribution of an informative head of a sampled sentence from IWSLT14 De-En.

Hint-Based Training for Non-Autoregressive Translation
Hint-based training from autoregressive teacher to non-autoregressive student

Autoregressive model

Non-autoregressive model

Hint-Based Training for Non-Autoregressive Translation
Hint-based training
from autoregressive teacher
to non-autoregressive student
• Directly regression fails because of the discrepancy between two models

• We penalize hidden states that are highly similar:

$$\mathcal{L}_{\text{hid}} = \frac{2}{(T_y - 1)T_y N} \sum_{s=1}^{T_y-1} \sum_{t=s+1}^{T_y} \sum_{l=1}^{N} \phi(d_{st}, d_{tr})$$

- $d_{st}$ and $d_{tr}$ are cosine similarities $s$-th and $t$-th of hidden states at layer $l$ of the student and teacher models

- $\phi$ is a fixed function only penalizes when the student’s hidden states are similar while the teacher’s not
We minimize the KL-divergence between the per-head encoder-to-decoder attention distribution of the student and teacher models

\[
\mathcal{L}_{\text{align}} = \frac{1}{T_yNH} \sum_{t=1}^{T_y} \sum_{l=s+1}^{N} \sum_{h=1}^{H} D_{KL}(a_{t,l,h}^{tr} \parallel a_{t,l,h}^{st})
\]

**Total Loss**

\[
\mathcal{L} = \mathcal{L}_{\text{nll}} + \lambda \mathcal{L}_{\text{hid}} + \mu \mathcal{L}_{\text{align}}
\]

Hint-Based Training for Non-Autoregressive Translation
Experimental settings

Datasets
- WMT14 En-De
- WMT14 De-En
- IWSLT14 De-En

Models
- Transformer-base
- Transformer-small

Inference
- Non-autoregressive model
- Non-autoregressive model with teacher reranking

Hint-Based Training for Non-Autoregressive Translation
# Experimental results

<table>
<thead>
<tr>
<th>Models</th>
<th>WMT14</th>
<th>IWSLT14</th>
<th>Latency</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>En-De</td>
<td>De-En</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSTM-based S2S (<a href="#">Wu et al., 2016; Bahdanau et al., 2016</a>)</td>
<td>24.60</td>
<td>/</td>
<td>28.53</td>
<td>/</td>
</tr>
<tr>
<td>ConvS2S (<a href="#">Gehring et al., 2017; Edunov et al., 2017</a>)</td>
<td>26.43</td>
<td>/</td>
<td>32.84</td>
<td>/</td>
</tr>
<tr>
<td>Transformer (<a href="#">Vaswani et al., 2017</a>)</td>
<td>27.30</td>
<td>31.29</td>
<td>33.26</td>
<td>784 ms†</td>
</tr>
<tr>
<td>Non-autoregressive models</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FT (<a href="#">Gu et al., 2017</a>)</td>
<td>17.69</td>
<td>20.62</td>
<td>/</td>
<td>39 ms†</td>
</tr>
<tr>
<td>FT (rescoring 10 candidates)</td>
<td>18.66</td>
<td>22.41</td>
<td>/</td>
<td>79 ms†</td>
</tr>
<tr>
<td>FT (rescoring 100 candidates)</td>
<td>19.17</td>
<td>23.20</td>
<td>/</td>
<td>257 ms†</td>
</tr>
<tr>
<td>IR (<a href="#">Lee et al., 2018, adaptive refinement steps</a>)</td>
<td>21.54</td>
<td>25.43</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>LT (<a href="#">Kaiser et al., 2018</a>)</td>
<td>19.8</td>
<td>/</td>
<td>/</td>
<td>105 ms†</td>
</tr>
<tr>
<td>LT (rescoring 10 candidates)</td>
<td>21.0</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>LT (rescoring 100 candidates)</td>
<td>22.5</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>NART w/ hints (B = 4, 9 candidates)</td>
<td>21.11</td>
<td>25.24</td>
<td>25.55</td>
<td>26 ms†</td>
</tr>
<tr>
<td>NART w/ hints (B = 4, 9 candidates)</td>
<td>25.20</td>
<td>29.52</td>
<td>28.80</td>
<td>44 ms†</td>
</tr>
</tbody>
</table>

*Hint-Based Training for Non-Autoregressive Translation*
Quality-speedup trade-off

Hint-Based Training for Non-Autoregressive Translation
Hidden states similarity

Hidden states cosine-similarity of a sampled sentence in IWSLT14 De-En.

Hint-Based Training for Non-Autoregressive Translation
Attention distribution

Encoder-to-decoder attention distribution of an informative head of a sampled sentence from IWSLT14 De-En.

Hint-Based Training for Non-Autoregressive Translation
Ablation studies on IWSLT14 De-En. Results are BLEU scores without teacher rescoring.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\mathcal{L}_{nll}$</th>
<th>$\mathcal{L}<em>{nll} + \mathcal{L}</em>{align}$</th>
<th>$\mathcal{L}<em>{nll} + \mathcal{L}</em>{align} + \mathcal{L}_{hid}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>23.08</td>
<td>24.76</td>
<td>25.55</td>
</tr>
<tr>
<td>Long-sentence BLEU</td>
<td>17.48</td>
<td>19.24</td>
<td>20.63</td>
</tr>
</tbody>
</table>
Instead of adding new modules that can slow down the model, we proposed a method to leverage the **hints** from the autoregressive model to help the training of the non-autoregressive model.
Thanks!

Q&A

Zhuohan Li (zhuohan@cs.berkeley.edu)