Multimodal Machine Translation with Embedding Prediction

Tosho Hirasawa, Hayahide Yamagishi, Yukio Matsumura, Mamoru Komachi

hirasawa-tosho@ed.tmu.ac.jp

Tokyo Metropolitan University

NAACL SRW 2019
Multimodal Machine Translation

- Practical application of machine translation
- Translate a source sentence along with related nonlinguistic information
  - Visual information

  two young girls are sitting on the street eating corn.

  deux jeunes filles sont assises dans la rue, mangeant du maïs.
Issue of MMT

- Multi30k [Elliott et al., 2016] has only small mount of data
  - Statistic of training data

<table>
<thead>
<tr>
<th></th>
<th>Sentences</th>
<th>Tokens</th>
<th>Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>29,000</td>
<td>377,534</td>
<td>10,210</td>
</tr>
<tr>
<td>French</td>
<td></td>
<td>409,845</td>
<td>11,219</td>
</tr>
</tbody>
</table>

- Hard to train rare word translation
  - Tend to output synonyms guided by language model

<table>
<thead>
<tr>
<th></th>
<th>Source</th>
<th>Reference</th>
<th>NMT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>deux jeunes filles sont assises dans la rue , mangeant du maïs .</td>
<td>two young girls are sitting on the street eating corn .</td>
<td>two young girls are sitting on the street eating food .</td>
</tr>
</tbody>
</table>
Previous Solutions

• Parallel corpus without images [Elliott and Kádár, 2017; Grönroos et al., 2018]
  • Out-of-domain data
  • Pseudo in-domain data by filtering general domain data

• Pseudo-parallel corpus [Sennrich et al., 2016; Helcl et al., 2018]
  • Back-translation of caption/monolingual data

• Monolingual data
  • Pretrained Word Embedding
    • Seldomly studied
Motivation

• Introduce pretrained word embedding to MMT
  • Improve rare word translation in MMT
  • Pretrained word embeddings with conventional MMT?
    • See our paper on MT Summit 2019 (https://arxiv.org/abs/1905.10464)!

• Pretrained Word Embedding in text-only NMT
  • Initialize embedding layers in encoder/decoder [Qi et al., 2018]
    ✓ Improve overall performance in low-resource domain
  • Search-based decoder with continuous output [Kumar and Tsvetkov, 2019]
    ✓ Improve rare word translation
1. Multimodal Machine Translation
2. MMT with Embedding Prediction
3. Pretrained Word Embedding
4. Result & Conclusion
While validating, testing

Multitask Learning:
Train both MT task and shared space learning task to improve the shared encoder.

MT Model:
Bahdanau et al., 2015
While validating, testing

1. Use embedding prediction in decoder

two young girls are sitting on the street eating corn.

deux jeunes filles sont assises dans la rue, mangeant du maïs.

2. Initialize embedding layers in encoder/decoder with pretrained word embeddings

3. Shift visual features to make the mean vector be a zero

Shared Space Learning Task

While training
Embedding Prediction (Continuous Output)

• i.e. Continuous Output [Kumar and Tsvetkov, 2019]

• Predict a word embedding and search for the nearest word

1. Predict a word embedding of next word.
2. Compute cosine similarities with each word in pretrained word embedding.
3. Find and output the most similar word as system output.

Keep unchanged:
Pretrained word embedding will NOT be updated during training.
Embedding Layer Initialization

- Initialize embedding layer with pretrained word embedding
- Fine-tune the embedding layer in encoder
- **DO NOT** update the embedding layer in decoder

[Qi et al., 2018]
Loss Function

- Model loss: Interpolation of each loss [Elliot and Kádáar, 2017]
  \[ J = \lambda J_T(\theta, \phi_T) + (1 - \lambda) J_V(\theta, \phi_V) \]

- MT task: Max-margin with negative sampling [Lazaridou et al., 2015]
  \[ J_T(\theta, \phi_T) = \sum_{j}^{M} \max\{0, \gamma + d(\hat{e}_j, e(w_j^-)) - d(\hat{e}_j, e(y_j))\} \]
  - negative sampling
    \[ w_j^- = \arg\max_{w \in V} d(\hat{e}_j, e(w)) - d(\hat{e}_j, e(y_j)) \]

- Shared space learning task: Max-margin [Elliot and Kádáar, 2017]
  \[ J_V(\theta, \phi_V) = \sum_{v' \neq v} \max\{0, \alpha + d(\hat{v}, v') - d(\hat{v}, v)\} \]
1. Multimodal Machine Translation
2. MMT with Embedding Prediction
3. Pretrained Word Embedding
4. Result & Conclusion
Hubness Problem \cite{Lazaridou2015}

- Certain words (hubs) appear frequently in the neighbors of other words
  - Even of the word that has entirely no relationship with hubs

- Prevent the embedding prediction model from searching for correct output words
  - Incorrectly output the hub word
All-but-the-Top [Mu and Viswanath, 2018]

• Address hubness problem in other NLP tasks

• Debias a pretrained word embedding based on its global bias
  1. Shift all word embeddings to make their mean vector into a zero vector
  2. Subtract top 5 PCA components from each shifted word embedding

• Applied to pretrained word embeddings for encoder/decoder
1. Multimodal Machine Translation
2. MMT with Embedding Prediction
3. Pretrained Word Embedding
4. Result & Conclusion
Implementation & Dataset

• Implementation
  • Based on nmtpytorch v3.0.0 [Caglayan et al., 2017]

• Dataset
  • Multi30k (French to English)
  • Pretrained ResNet50 for visual encoder

• Pretrained Word Embedding
  • FastText
  • Trained on Common Crawl and Wikipedia
    • https://fasttext.cc/docs/en/crawl-vectors.html

Our code is here: https://github.com/toshoirasawa/nmtpytorch-emb-pred
Hyper Parameters

• Model
  • dimension of hidden state: 256
  • RNN type: GRU
  • dimension of word embedding: 300
  • dimension of shared space: 2048
  • Vocabulary size (French, English): 10,000

• Training
  • $\lambda = 0.99$
  • Optimizer: Adam
  • Learning rate: 0.0004
  • Dropout rate: 0.3
Word-level $F_1$-score

- Bahdanau et al., 2015
- IMAGINATION
- Ours

Frequency in training data

- Rare words
- 1
- 2
- 3
- 4
- 5 - 9
- 10 - 99
- 100+

F-score of word
### Ablation w.r.t. Embedding Layers

<table>
<thead>
<tr>
<th>Encoder</th>
<th>Decoder</th>
<th>Fixed</th>
<th>BLEU</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>FastText</td>
<td>FastText</td>
<td>Yes</td>
<td>53.49</td>
<td>43.89</td>
</tr>
<tr>
<td>random</td>
<td>FastText</td>
<td>Yes</td>
<td>53.22</td>
<td>43.83</td>
</tr>
<tr>
<td>FastText</td>
<td>random</td>
<td>No</td>
<td>51.53</td>
<td>43.07</td>
</tr>
<tr>
<td>random</td>
<td>random</td>
<td>No</td>
<td>51.42</td>
<td>42.77</td>
</tr>
<tr>
<td>FastText</td>
<td>FastText</td>
<td>No</td>
<td>51.42</td>
<td>42.88</td>
</tr>
<tr>
<td>random</td>
<td>FastText</td>
<td>No</td>
<td>50.72</td>
<td>42.52</td>
</tr>
</tbody>
</table>

**Encoder/Decoder**: Initialize embedding layer with `random` values or `FastText` word embedding.

**Fixed (Yes/No)**: Whether fix the embedding layer in decoder or fine-tune that while training.

- Fixing the embedding layer in decoder is essential
  - Keep word embeddings in input/output layers consistent
## Overall Performance

<table>
<thead>
<tr>
<th>Model</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>BLEU</td>
</tr>
<tr>
<td>Bahdanau et al. 2015</td>
<td>50.83</td>
<td>51.00 ± .37</td>
</tr>
<tr>
<td>+ pretrained</td>
<td>52.05</td>
<td>52.33 ± .66</td>
</tr>
<tr>
<td>IMAGINATION</td>
<td>51.03</td>
<td>51.18 ± .16</td>
</tr>
<tr>
<td>+ pretrained</td>
<td>52.40</td>
<td>52.75 ± .25</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>53.14</strong></td>
<td><strong>53.49 ± .20</strong></td>
</tr>
</tbody>
</table>

Model (+ pretrained): Apply embedding layer initialization and All-but-the-Top debiasing.

- Our model performs better than baselines
  - Even those with embedding layer initialization
Ablation w.r.t. Visual Features

<table>
<thead>
<tr>
<th>Visual Features</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>BLEU</td>
</tr>
<tr>
<td>Centered</td>
<td>53.14</td>
<td>53.49</td>
</tr>
<tr>
<td>Raw</td>
<td>52.65</td>
<td>53.27</td>
</tr>
<tr>
<td>No</td>
<td>52.97</td>
<td>53.25</td>
</tr>
</tbody>
</table>

Visual Features (Centered/Raw/No): Use centered visual features or raw visual features to train model. "No" show the result of text-only NMT with embedding prediction model.

- Centering visual features is required to train our model.
Conclusion & Future Works

• MMT with embedding prediction improves ...
  • Rare word translation
  • Overall performance

• It is essential for embedding prediction model to ...
  • Fix the embedding in decoder
  • Debias the pretrained word embedding
  • Center the visual feature for multitask learning

• Future works
  • Better training corpora for embedding learning in MMT domain
  • Incorporate visual features into contextualized word embeddings

Thank you!
un homme en vélo pédale devant une voûte.

a man on a bicycle pedals through an archway.

a man on a bicycle pedal past an arch.

a man on a bicycle pedals outside a monument.

a man on a bicycle pedals in front of a archway.
Translation Example (long)

<table>
<thead>
<tr>
<th>Source</th>
<th>Reference</th>
<th>Text-only NMT</th>
<th>IMAGINATION</th>
<th>Ours</th>
</tr>
</thead>
</table>

quatre hommes, dont trois portent des kippas, sont assis sur un tapis à motifs bleu et vert olive.

four men, three of whom are wearing prayer caps, are sitting on a blue and olive green patterned mat.

four men, three of whom are wearing aprons, are sitting on a blue and green speedo carpet.

four men, three of them are wearing alaska, are sitting on a blue patterned carpet and green green seating.

four men, three are wearing these are wearing these are sitting on a blue and green patterned mat.