Adversarial Stability Training in Neural Machine Translation of Chinese-to-English Text

Mandy Lu
mlu355@stanford.edu

Kaylie Zhu
kayliez@stanford.edu

Motivation

- Major challenge for neural machine translation (NMT) models: semantically similar input with severely dissimilar encodings reducing translation performance
- Use adversarial stability training (AST) framework to improve NMT robustness by integrating an adversarial objective to encourage noisy and true data encodings to be similar

Task Definition

Develop a perturbation-robust Chinese-English NMT model with a modified AST in a Transformer network to generalize better on difficult unstructured datasets spanning multiple domains for which it is crucial to maintain robustness over noise

Dataset: Casict2015

- 2 million parallel Chinese-English sentence pairs (22,802,353 words) from CWMT with 60% web crawl, 20% movie subtitles and 20% from English to Chinese thesaurus, segmented with Jieba
- Spans many contexts, from technical writing to movie subtitles

Methodology and Model

I. AST

Given a minibatch of source sentences $x$, we construct a minibatch of perturbed sentences $x'$ by adding random Gaussian noise to all word embeddings to simulate various types of feature-level perturbations.

$$E[x'] = E[x] + \eta \epsilon \sim N(0, \sigma^2 I)$$

1. Encoder acts as generator (G) to make embeddings $H_x$ and $H_y$ as similar as possible to fool discriminator (D)
2. D tries to distinguish noisy from true embeddings by maximizing D(G(x)) to 1 and minimize D(G(x')) to 0
3. New objective $J$ is a hybrid loss function that incorporates $L_{inv}$ and $L_{noisy}$ and $L_{true}$

II. Hybrid loss functions

$$L(x,y;\theta) = \sum_{(x,y) \in S} \log P(y|x;\theta)$$

$$L_{inv}(x,x';\theta_{enc};\theta_{disc}) = E_{x' \sim \epsilon} \left[ -\log \left( 1 - D(G(x')) \right) \right]$$

$$L_{noisy}(x,x';\theta_{enc};\theta_{disc}) = E_{x' \sim \epsilon} \left[ -\log \left( 1 - D(G(x')) \right) \right]$$

III. Ablation Studies

To measure individual effectiveness of sub-parts of our framework, we (1) removed the GAN structure from the training framework, reducing to a data augmentation problem of simply training on both original and noisy data and (2) implemented the original simplified generator loss $L_{inv}$ as negative $L_{inv}$

<table>
<thead>
<tr>
<th>Evaluation metric</th>
<th>No Discriminator</th>
<th>Simplified Loss</th>
<th>Modified AST</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>4.46</td>
<td>4.64</td>
<td>4.67</td>
</tr>
<tr>
<td>Perplexity</td>
<td>3.14</td>
<td>3.06</td>
<td>2.43</td>
</tr>
</tbody>
</table>

Results

Summary of Model Comparison: Best performing model was AST Embed (pre-trained embeddings from Wikipedia) with 18.42 BLEU and 1.23 Perplexity

I. Hybrid loss functions

$$J_{inv}(\theta) = \sum_{(x,y) \in S} \left[ \mathcal{L}_{true}(x,y;\theta_{true},\theta_{disc}) - \alpha \mathcal{L}_{inv}(x,x';\theta_{true},\theta_{disc}) + \beta \mathcal{L}_{noisy}(x,x';\theta_{true},\theta_{disc}) \right]$$

Conclusion and Future Work

- AST is an effective method to develop perturbation-robust Chinese-English NMT models and performs well with Transformer network on diverse datasets
- Ablation studies demonstrate the efficacy of some components constituting our hybrid model: the GAN component and our modified objective loss
- Future work could include: (1) techniques for rare words such as byte-pair encoding (2) using weak supervision to label semantically similar train text for AST input

References