Project Proposal: Round Trip Loss for Machine Translation

Anja Adamov[a]*, Lauro Böni[b]†, Simon A. Broda[cde]‡, Urs Vögeli[b]§

[a] *IBM Switzerland Ltd., Zurich, Switzerland*

[b] *ETH Zurich*

[c] *Department of Banking and Finance, University of Zurich*

[d] *Quantitative Economics Section, University of Amsterdam*

[e] *Tinbergen Institute Amsterdam*

January 4, 2020

Abstract

We propose to augment a machine translation system by adding a round-trip loss which encourages the system to generate translations that when translated back into the source language, retain much of the original structure. We show that in doing so, the model will learn internal representations with improved semantic meaning.

Key Words: Cycle consistency loss; deep learning; natural language understanding; machine translation; round-trip loss.

* *E-mail address:* adamova@student.ethz.ch

† *E-mail address:* laboeni@gmail.com

‡ *E-mail address:* simon.broda@uzh.ch

§ *E-mail address:* voegeli.urs@gmail.com

1 Introduction

Machine translation is an important task within the field of natural language understanding. Consequently, a variety of models have been proposed for solving it. One of the first truly successful models was the sequence-to-sequence (seq2seq) model of Sutskever et al. (2014), while the current state of the art builds upon the *Transformer* architecture introduced by Vaswani et al. (2017). At a high level, both the seq2seq and the transformer architectures are comprised of an encoder and decoder; the encoder learns an internal representation of the source sentence, and the decoder decodes it into the target language. For these models to work, the internal representation must capture the semantic meaning of the source sentence.

Irrespective of the architecture, these models are typically trained with the cross-entropy loss between the ground truth and predicted sentences, usually with teacher forcing. Here, we propose to add a round-trip penalty to the loss function of the model. The idea is that instead of training a single model to translate from a source language $\mathcal{S}$ to a target language $\mathcal{T}$, one trains two models (of identical structure), one mapping $\mathcal{S} \mapsto \mathcal{T}$ and one mapping $\mathcal{T} \mapsto \mathcal{S}$. The two models are, at first, trained independently and in parallel, by feeding in the same batches of sentence pairs (In this paper, we rely on the Transformer architecture of Vaswani et al. (2017), but the idea applies to any encoder-decoder architecture). After τ epochs, the models is then trained jointly, using as loss a sum of four cross-entropy terms, viz.,

$$\mathcal{L} = CE(s, \hat{s}) + CE(t, \hat{t}) + \lambda \left(CE(s, \tilde{s}) + CE(t, \tilde{t}) \right). \tag{1}$$

Here, $s \in \mathcal{S}$ is the ground truth sentence in the first language, $t \in \mathcal{T}$ is the corresponding ground truth in the second language, $\hat{s} \in \mathcal{S}$ and $\hat{t} \in \mathcal{T}$ are the respective translations of t and s, $\tilde{s}$ is obtained by translating $\hat{t}$ back to $\mathcal{S}$, and $\tilde{t}$ is obtained by translating $\hat{s}$ back to $\mathcal{T}$. τ and λ are hyperparameters.

The round-trip loss is inspired by the cycle consistency loss pioneered by Zhu et al. (2017) in the context of GANs. The only application of this concept to the field of machine translation of which we are aware is Su et al. (2018), who adapt it for unsupervised multi-modal machine translation. Here, we propose to apply the idea to supervised machine translation. We conjecture that encouraging the model to generate round-trippable translations will help it learn a semantically meaningful representation. A related idea is that of back-translation, pioneered by Sennrich et al. (2016). Sennrich et al. propose to augment the parallel training corpus used to train a machine translation system with synthetic data, obtained by translating a monolingual corpus in the target language to the source language using an independently trained system. They argue that this is beneficial in particular when parallel training data is scarce. The difference with our approach is two-fold: i) in our approach, the two networks are trained *jointly*, each with their own round-trip loss term, whereas in back-translation, the models are trained separately, with only one of them benefiting from back-translation; ii) we do not assume the existence of a separate monolingual corpus in the target language, but work strictly with a par-

allel corpus. Having ground truth available, we are thus able to use teacher forcing in constructing the round-trip loss contribution, greatly speeding up training.

The remainder of this manuscript is organized as follows. Section 2 describes the model architecture and our implementation of it. Section 3 details the results of our experiments with the proposed round-trip loss. Section 4 provides a discussion, and Section 5 concludes.

2 Models and Methods

2.1 Model Architecture

description of Transformer goes here, perhaps including some pictures from the paper

2.2 Implementation

We use the Transformer implementation in TensorFlow 2.0 available from the tensorflow website and implement a custom training loop, which instantiates the $\mathcal{S} \mapsto \mathcal{T}$ and $\mathcal{T} \mapsto \mathcal{S}$ models and trains them jointly with the loss described in (1). At test time, we rely on the beam search implementation from tensor2tensor, and compute the BLEU score using the nltk package.

2.3 Training Details

We train our models on the WMT'14 English-German data set available from https://nlp.stanford.edu/projects/nmt/. The models hyperparameters are set as follows:

3 Results

3.1 Numerical comparisons

3.2 Example Translations

4 Discussion

5 Summary

References

Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meet-*

ing of the Association for Computational Linguistics (Volume 1: Long Papers), pages 86–96, Berlin, Germany. Association for Computational Linguistics. 1

Su, Y., Fan, K., Bach, N., Kuo, C. C. J., and Huang, F. (2018). Unsupervised multi-modal neural machine translation. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. 1

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14. 1

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17. 1

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*. 1