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*

November 8, 2019

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 conjecture 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

Proposal

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. After τ epochs, the models will then be trained jointly, using as loss a sum of four cross-entropy terms, viz.,

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

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 surmise that encouraging the model to generate round-trippable translations will help it learn a semantically meaningful representation.

In order to make this idea operational, we will use the seq2seq model implementation at https://github.com/tensorflow/nmt as a starting point, augmenting it with our proposed round-trip loss and training it on the WMT'14 English-German data set available from https://nlp.stanford.edu/projects/nmt/, possibly restricting attention to a subset for computational reasons. Hyperparameters will be tuned by cross-validation on the BLEU score (Papineni et al., 2002). Results will be compared to the seq2seq model without round-trip loss as a baseline. Possible extensions include i) replacing the simplistic seq2seq architecture with the state-of-the-art transformer architecture of Vaswani et al. (2017), ii) using pre-trained word embeddings from BERT (Devlin et al., 2018) rather than learning the embeddings from scratch, (iii) investigating the performance for other language pairs.

References

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. 1

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02. 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