Latest Development in the FoTran Project – Scaling Up Language Coverage in Neural Machine Translation Using Distributed Training with Language-Specific Components

Raúl Vázquez[♠] **Michele Boggia**[♠] **Alessandro Raganato**[♡]
Niki A. Loppi[◇] **Stig-Arne Grönroos**[♠♣] **Jörg Tiedemann**[♠]
[♠] University of Helsinki, Finland [♡] University of Milano-Bicocca, Italy
[◇] NVIDIA [♣] Silo.AI
[♠]{name.surname}@helsinki.fi, [♡]{name.surname}@unimib.it,
[◇]nloppi@nvidia.com

Abstract

We give an update on the *Found in Translation* (FoTran) project, focusing on the study of emerging language-agnostic representations from neural machine translation (NMT). We describe our attention-bridge model, a modular NMT model which connects language-specific components through a shared network layer. Our latest implementation supports distributed training over many nodes and GPUs in order to substantially scale up the number of languages that can be included in a modern neural translation architecture.

1 Introduction

The FoTran project aims at developing models for natural language understanding trained on implicit information given by large collections of human translations.[1] It is funded by a European Research Council consolidation grant, running from 2018 to 2023 within the language technology research group at the University of Helsinki under coordination of Prof. Jörg Tiedemann.

Cross-lingual grounding, useful for resolving ambiguities through translation, is a guiding principle of the project. Consequently, we developed a model for multilingual NMT specifically designed to obtain meaning representations injected with multilingual data (Vázquez et al., 2020). Former project results pointed towards the improvement of both the translation quality and the abstractions acquired by our model when including more languages (Vázquez et al., 2019; Raganato et al., 2019). Due to the use of language-specific modules, the overall model architecture grows when languages and translation directions are added. Doing this on a single device does not scale beyond the memory limits of that specific computing node. This is a limitation for testing the project hypothesis that training on increasing amounts of linguistically diverse data improves the abstractions found by the model – eventually leading to language-independent meaning representations useful for machine translation and tasks that require semantic reasoning and inference. Here, we propose strategies to address those issues: (1) distribute modules across several processing units, (2) efficiently train the network over many translation directions, and (3) reuse the trained modules without having to load the entire network. These together deliver a cost-effective multilingual NMT system that can further be used for extracting multilingual meaning representations.[2] Despite the high computational resources needed to scale up the number of translation directions when training the model, its modularity allows to reuse the trained components on relatively small processing units, making multilingual models more affordable and increasing their availability.

2 Methodology

The implementation follows an encoder–decoder architecture, incorporating language specific encoders and decoders to enable multilingual training. They are connected via a shared inner-attention layer that summarizes the encoder information in a fixed-size vector representation (Vázquez et al., 2019), which in turn can be applied to downstream tasks (Vázquez et al.,

[1] http://www.helsinki.fi/fotran

[2] https://github.com/Helsinki-NLP/FoTraNMT

2020). We refer to all encoders, decoders and to the shared layer as *modules*. Encoders and decoders are language-specific as they only see training data from specific translation directions.

We distribute the model across multiple processing units by loading, in each device, encoders and decoders for a subset of the training translation directions. The inner-attention layer is shared across all processing units. All modules that are present in more than one device are initialized with the same weights, and gradients for these parameters are communicated across devices to ensure that they remain synchronous.

In general, allocating language pairs with common source/target languages on the same device decreases both the total memory footprint of the model and the amount of communication needed to keep the modules synced. Formally, we define the partition of the training language pairs $L_{ij} = (S_{ij}, T_{ij})$ over N units as $\mathcal{P} = \{(L_{11}, L_{12}, \ldots), \ldots, (L_{N1}, L_{N2}, \ldots)\}$, where the first subscripts indicate to which device each pair is assigned, and the second is an incremental index over all pairs assigned to the same device. Whenever $X_{ij} = X_{kl}$ for $i \neq k$, with $X \in \{S, T\}$ representing either a source or a target language, we need to load a copy of the same module in devices i and k. This will also impact the training time as it requires communicating gradients across devices to keep modules synced.

However, when dealing with a high number of translation directions (and a limited number of source and target languages) it becomes impossible to avoid this condition: gathering together language pairs based on the source (target) language could result in a scattered configuration based on target (source) languages. We address these problems using two strategies. First, we solve an allocation problem to minimize inter-device communication. Since in most cases the problem has no feasible exact solutions, we approximate a solution using the Hungarian algorithm over a cost matrix that makes it cheaper to assign the same language to a given GPU. Second, we propose to schedule the gradient updates to minimize the waiting time when inter-device communication happens.

At each training step the ith device starts performing a forward pass over a training batch for the language pair L_{i1} and accumulates gradients over all the language pairs L_{ij}, where j runs from one to the number of language pairs assigned to the processing unit. Afterwards, gradients of modules that are present in multiple processing units are averaged across devices. Module weights are then updated according to the computed gradients. We ensure that all copies of all modules have non-zero gradients that can be communicated, preventing the training loop from hanging.

We also save the modules individually to be loaded and used independently in an efficient way. This makes the system more portable and user-friendly for further fine-tuning, generating translations, and experimentation with multilingual sentence-representations.

3 Final Remarks

FoTran aims at testing and analyzing representations obtained from massively multilingual NMT systems, and we devised a model architecture that is optimized for training large models (with a sufficiently large high-performance cluster). After training, it can also easily be used in non-resource-intensive settings due to its modular design. Next, we intend to systematically explore the effect of increasing language diversity and how the abstraction capabilities of the inner representations are affected in different settings.

Acknowledgments

This work is part of the FoTran project, funded by the European Research Council (ERC) under the EU's Horizon 2020 research and innovation program (agreement № 771113). We also thank the CSC-IT Center for Science Ltd., for computational resources and NVIDIA AI Technology Center (NVAITC) for the expertise in distributed training.

References

Raganato, Alessandro, Raúl Vázquez, Mathias Creutz and Jörg Tiedemann. 2019. An Evaluation of Language-Agnostic Inner-Attention-Based Representations in Machine Translation. *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*. ACL. 27–32.

Vázquez, Raúl, Alessandro Raganato, Mathias Creutz and Jörg Tiedemann. 2019. Multilingual NMT with a Language-Independent Attention Bridge. *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*. ACL. 33–39.

Vázquez, Raúl, Alessandro Raganato, Mathias Creutz and Jörg Tiedemann. 2020. A Systematic Study of Inner-Attention-Based Sentence Representations in Multilingual Neural Machine Translation. *Computational Linguistics*, vol. 46(2):387–424.