OpenNMT: Open-Source Toolkit for Neural Machine Translation

Yoon Kim, Guillaume Klein, Jean Senellart, Alexander M. Rush ...

Abstract

We describe an open-source toolkit for neural machine translation that supports research development of sequence-to-sequence models. The system is prioritizes simplicity, modularity, and efficiency to make it reasonable for researchers to experiment with variants of neural machine translation that explore different feature representations, model architectures, and source (multi)-modalities, while maintaining competitive performance and tractable training requirements. The toolkit consists of modeling and decoding support, as well as detailed pedagogical documentation about the underlying methodologies.

1 Introduction

- Description of current state of NMT
- Difference with other NMT
- Description of open-source nmt systems. Several research systems, but there is a not a ...

Motivation why build a new system.

Some proprietary (google, etc.), others too research-y

Prioritized factors

- Training efficiency
- System Modularity
- Support for alternative representations
- Industrial decoding.

2 Background: Neural Machine Translation

- One column describing the technical details

Table 1: Performance Results. Several languages

Table 2: Speed Results. Multi-GPU, distillation, c decoder

3 Implementation

Details: torch. text-to-text mapping. toolkit for extensibility

3.1 Modularity

- Separate encoder/decoder
- Arbitrary input/output representations (features, images)

3.2 Optimizations

Encoder/Decoder optimization

- Shared memory (cite opt net)
- C-Decoder

3.3 Advanced Features

Other papers .

4 User Studies

Feature-Based Inputs Discuss using modularity to add features as input and output

Knowledge Distillation/Pruning Discuss model access to allow for pruning and distillation.

Im2Latex Discuss ability to modify encoder representation to allow for other tasks.

5 Experimental Results

Mainly focus on NMT. Speed, memory, accuracy.

Picture of demo application running

6 Conclusion

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Association for Computing Machinery. 1983. In *Computing Reviews*, volume 24, pages 503–512.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.