PATTERN AS A FOREIGN LANGUAGE

TTIC 31210 PROJECT REPORT
SPRING 2017

Hao Jiang
The University of Chicago

ABSTRACT

The recurrent neural network (RNN) based encoder-decoder architecture has been widely used for various sequence-to-sequence translation tasks such as natural language translation and grammar inference, and has achieved significant success on these tasks. In this project, we attempt to use this architecture to infer common patterns from multiple inputs, which is a crucial task for information extraction and management. This leads to two new types of tasks: combinations and summarization. In combination task, we train the encoder-decoder with known patterns and attempt to use it to recognize the combination of these patterns. In summarization task, we define some rules to infer a common pattern from multiple records and use encoder-decoder to learn these rules. Preliminary results show that existing architecture does not fit these applications and new architectures are required.

1 INTRODUCTION

Recurrent neural network based encoder-decoder architecture has recently been widely adopted in various tasks such as neural machine translation (Hermann & Blunsom (2014); Cho et al. (2014)), image captioning (Karpathy & Li (2015) and grammar inference (Vinyals et al. (2014)). These tasks can all be viewed as a translation between source and target domains using encoding-decoding process. First, an encoder is employed to convert the input to a single vector, which is supposed to contain a summary of the input. With that as input, a decoder is then used to generate an output belonging to the destination domain from the encoded result. The entire encoder-decoder model is trained on input pairs of source-target training data to maximize the probability of correct translation.

In this project, we explore the possibility of using RNN-based encoder-decoder architecture to do pattern extraction. Pattern extraction infers common patterns from multiple records, and extract subcomponents from the records accordingly. Figure 1 demonstrates several lines of Java application logs and the pattern inferred from them. Pattern extraction allows in-depth understanding of the data's nature, enabling more efficient data compression and accurate data analysis. Previous methods of pattern extraction developed by Fisher et al. (2008) use a rule-based method to iteratively extract common words from records, which is inefficient when dealing with large dataset. In addition, this method does not learn from past dataset to speed up future processing. We plan to use RNN to address these problems.

A potential challenge of using RNN encoder for pattern extraction is constructing an efficient training set. Unlike in the case of natural language and grammar, the pattern does not have a closed, well-defined domain. The vocabularies of pattern can be arbitrary combinations of alphabet, numbers and symbols. There's also no "grammar" governing these vocabularies. Thus the attempt to construct a complete training set that covers all possible patterns is infeasible. Instead, we try to attack the problem from different directions.

In this project, we experiment with two approaches. First, we attempt to imitate human's ability to recognize some common pattern, e.g., date, time and ip address. We train the encoder-decoder model with these common patterns, and explore the model's ability to recognize combination of these patterns. This method will allow the model to remember some patterns and recognize them when later encounter these pattern again.

Log Data	14:23:01.045 [main] DEBUG o.h.d.s.DefaultService - Synchronizing 14:23:48.656 [Thread] DEBUG o.h.d.storage.StorageService - Persisting Data 14:24:05.656 [monitor] WARN o.h.d.storage.StorageService - Invalid Input
Pattern	Timestamp [Thread Name] Level Source - Content

Figure 1: Application Log and extracted Pattern

2 Background

Long Short-Term Memory (Hochreiter & Schmidhuber (1997))

In Vinyals et al. (2014), the authors demonstrated that LSTM-based auto encoders can be used to infer tree-like structures such as grammars from sequential input.

3 Pattern Combination

4 Pattern Summarization

5 Experiment

6 Conclusion

References

Cho, Kyunghyun, van Merrienboer, Bart, Gulcehre, Caglar, Bahdanau, Dzmitry, Bougares, Fethi, Schwenk, Holger, and Bengio, Yoshua. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D14-1179.

Fisher, Kathleen, Walker, David, Zhu, Kenny Q., and White, Peter. From dirt to shovels: Fully automatic tool generation from ad hoc data. In *Proceedings of the 35th Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, POPL '08, pp. 421–434, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-689-9. doi: 10.1145/1328438.1328488. URL http://doi.acm.org/10.1145/1328438.1328488.

Hermann, Karl Moritz and Blunsom, Phil. A simple model for learning multilingual compositional semantics. *Proceedings of the Second International Conference on Learning Representations (ICLR 2014)*, 2014.

Hochreiter, Sepp and Schmidhuber, Jrgen. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735. URL http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Karpathy, Andrej and Li, Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 3128–3137, 2015. doi: 10.1109/CVPR.2015.7298932. URL https://doi.org/10.1109/CVPR.2015.7298932.

Vinyals, Oriol, Kaiser, Lukasz, Koo, Terry, Petrov, Slav, Sutskever, Ilya, and Hinton, Geoffrey E. Grammar as a foreign language. *CoRR*, abs/1412.7449, 2014. URL http://arxiv.org/abs/1412.7449.