Navigating with Graph Representations for Fast and Scalable Decoding of Neural Language Models

Minjia Zhang, Xiaodong Liu, Wenhan Wang, Jianfeng Gao, Yuxiong He
Microsoft
{minjiaz,xiaodl,wenhanw,jfgao,yuxhe}@microsoft.com

FGD (Fast Graph Decoder) is a fast and scalable decoding algorithm for accelerating the inference of neural language modeling and its end applications. On NMT, FGD obtains more than 14X speedup on softmax layer execution time over full-softmax with competitive BLEU score to the baseline. On NLM, FGD outperforms full-softmax by an order of magnitude with logarithmic scalability.

Method (Fast Graph Decoder)
- Transform word embeddings to exploit intrinsic closeness relationship between words.

FGD overview
- Softmax layer has a complexity of \(O(D \times |V|)\)
- FGD has a complexity of \(O(D \times \log|V|)\)

Step 1: Small world graph construction
- Inner product as a closeness measure is insufficient
- Inner product preserving transformation

Step 2: Decoding as searching small world graphs
- Precision and distance computation results explain the decoding accuracy and speedup of FGD.

Language modeling: Impact of vocabulary size
- FGD scales much better and the improvement becomes more significant with larger vocabulary sizes.

Acknowledgement
We thank Kevin Duh for reading a previous version of this paper and providing feedback. We thank the anonymous reviewers for their helpful suggestions for improving this paper. We thank Yi-Min Wang for supporting this work.