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

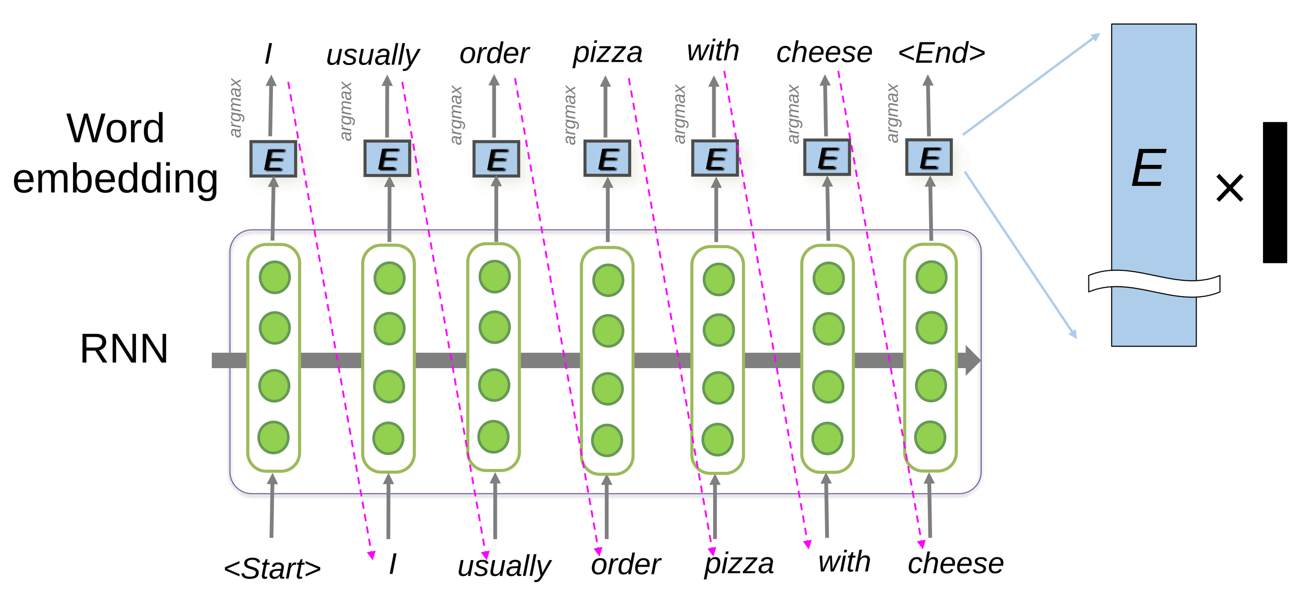
Minjia Zhang, Xiaodong Liu, Wenhan Wang, Jianfeng Gao, Yuxiong He
Microsoft
{minjiaz,xiaodl,wenhanw,jfgao,yuxhe}@microsoft.com

Highlights

- **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.

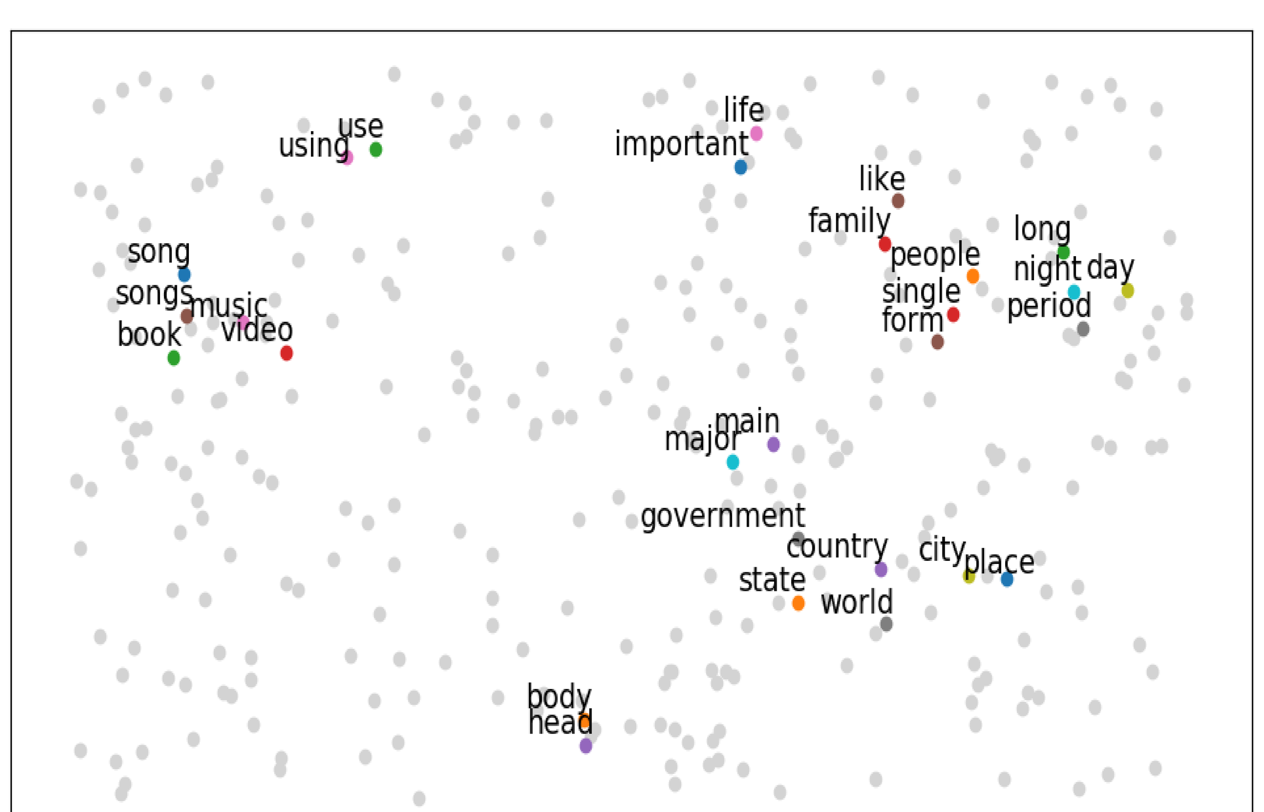
Background & Motivation

- High computational complexity of the softmax layer when the vocabulary size is large.
- Decoding bottleneck limits the applicability of NLMs in interactive services.



Goal: speedup the decoding process of neural language model and its end applications.

Visualization of closeness relation

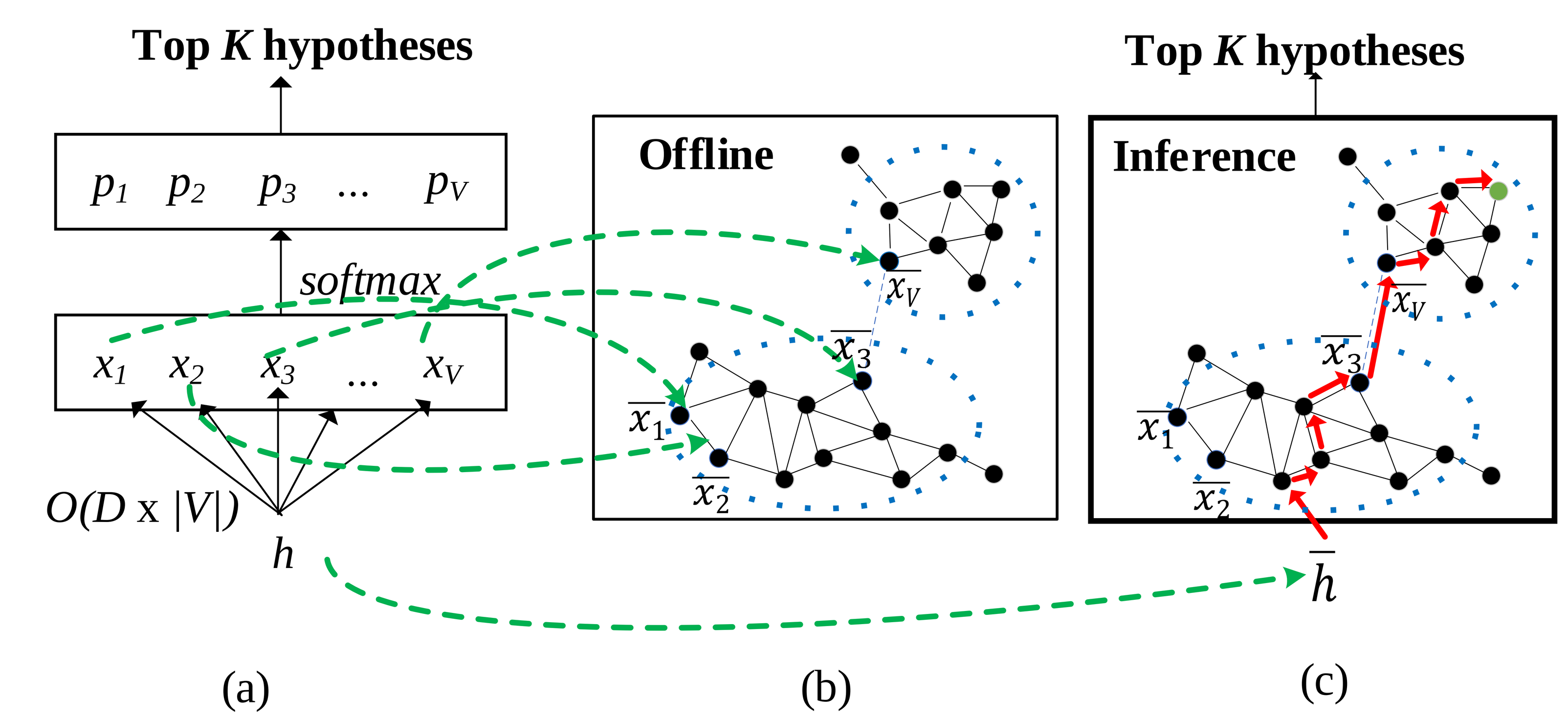


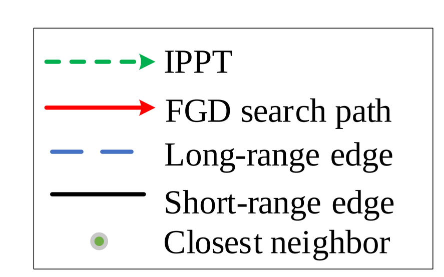
- Semantic more similar words are closer in distance.

Method (Fast Graph Decoder)

- Transform word embeddings to exploit intrinsic closeness relationship between words.

FGD overview





- Softmax layer has a complexity of O(D X |V|)
- FGD has a complexity of O(D X 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

Algorithm 1 Offline preprocessing algorithm FGD–P

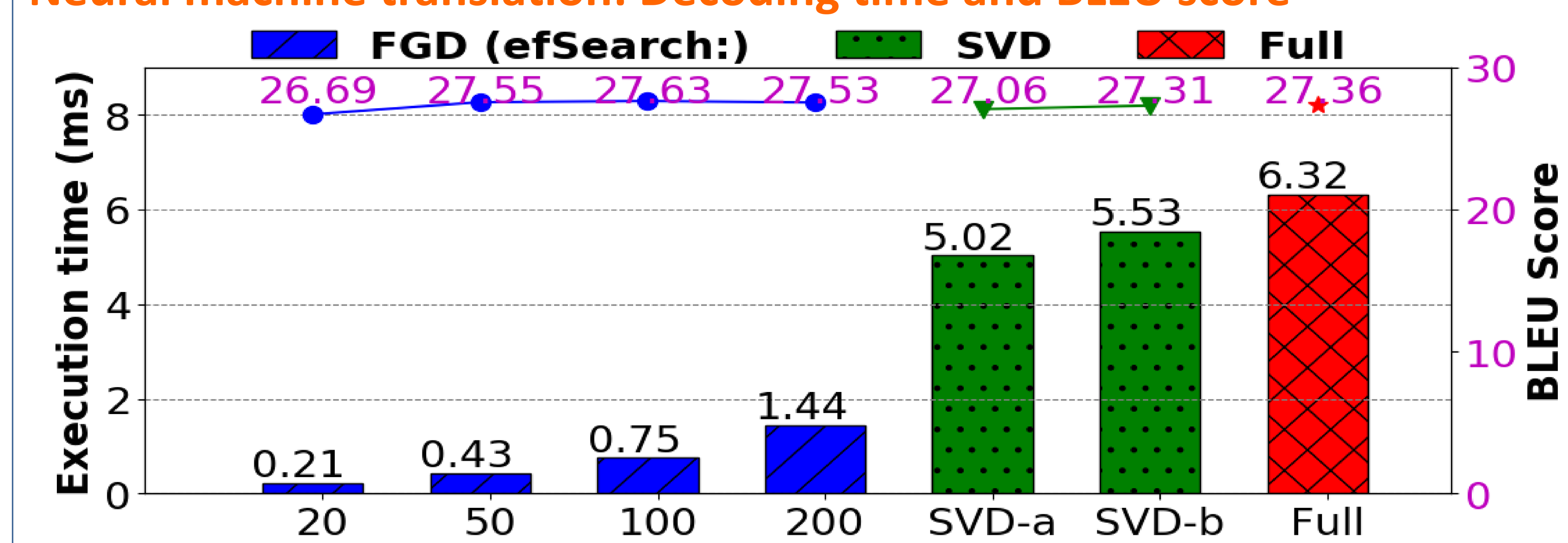
Input: Trained weights of the softmax layer X, and bias vector b.
Output: Small world graph G, and U_{max}.
Hyperparameter: Small world graph neighbor degree M.
for all i **in** $(0..|X|-1)$ **do**
  $\tilde{X}_{:i} \leftarrow [X_{:i}; b_i]$ ▷ Word embedding and bias fusion
$U_{max} \leftarrow \max_i \|\tilde{X}_{:i}\|_2$
for all i **in** $0..(|\tilde{W}|-1)$ **do**
  $\Delta_i \leftarrow \sqrt{{U_{max}}^2 - \|\tilde{X}_{:i}\|_2^2}$ ▷ Calculate the normalizer
  $\overline{X}_{:i} \leftarrow [\tilde{X}_{:i}; \Delta_i]$
$G \leftarrow CreateSwg(\overline{X}, M)$ ▷ Build small world graph

Algorithm 2 Online inference algorithm FGD–I

Input: Context vector h, small world graph G, and U_{max}.
Output: Probability distribution P over top-K word hypotheses.
Hyperparameter: Candidate queue length $efSearch$.
$\overline{h} \leftarrow [h; 1; 0]$ ▷ Map context vector from $\mathbb{R}^D$ to $\mathbb{R}^{D+2}$
$I^K, D^K \leftarrow SearchSwg(G, \overline{h}, K)$ ▷ Return top-K hypotheses with minimal distance to $\overline{h}$
for all i **in** $0..(K-1)$ **do**
  $S[I^K_{:i}] \leftarrow \frac{1}{2}\left(\|\overline{h}\|_2^2 + {U_{max}}^2 - {D^K_{:i}}^2\right)$ ▷ Map Euclidean distance back to inner product
$P \leftarrow exp(S)/\sum exp(S)$ ▷ Compute top-K softmax probability distribution

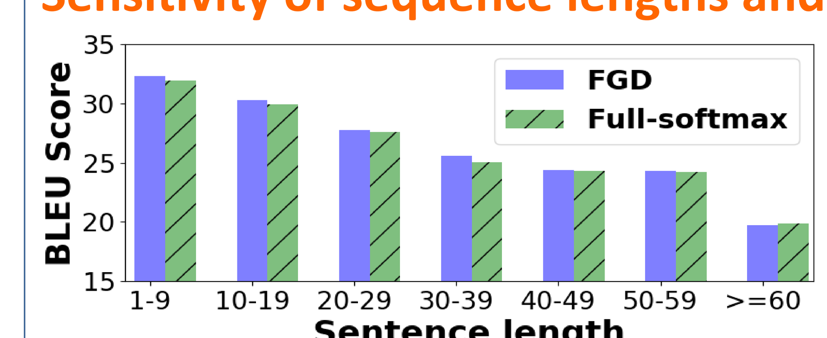
Experiments

Neural machine translation: Decoding time and BLEU score



- IWSLT'14 German-English corpus, 50K vocab size.
- **FGD** obtains more than **14X** speedup on softmax layer execution time over full-softmax with a similar BLEU score to the baseline
- **FGD** obtains **30X** speedup at the cost of decreasing 0.67 BLEU score.

Sensitivity of sequence lengths and beam sizes



efSearch	Beam = 1	Beam = 2	Beam = 5	Beam = 10
20	26.69	27.65	27.81	27.62
50	27.55	28.76	29.06	28.9
100	27.63	28.94	29.28	29.1
200	27.53	28.99	29.28	29.22
Full	27.36	28.91	29.45	29.34

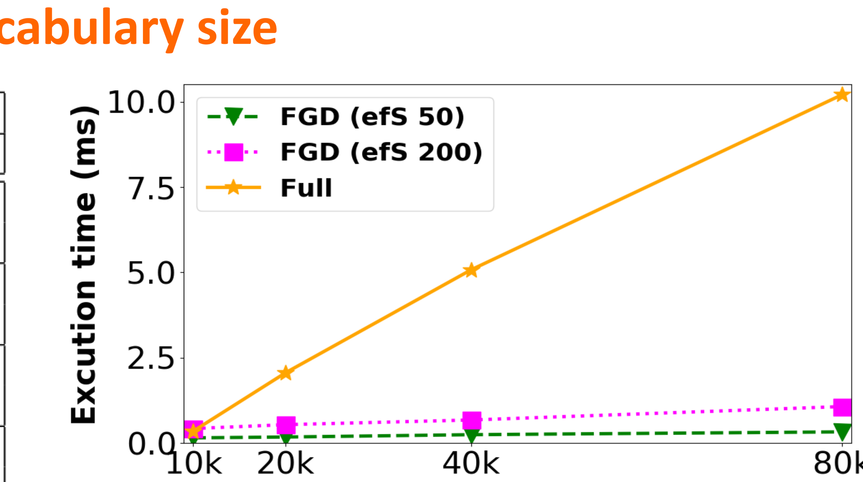
Internals of FGD

efSearch	P@1	P@2	P@5	P@10	dist_cnt (FGD/ Full)
20	0.939	0.934	0.929	0.918	981 / 50K
50	0.974	0.974	0.973	0.971	1922 / 50K
100	0.986	0.986	0.987	0.987	3310 / 50K
200	0.992	0.993	0.994	0.994	5785 / 50K

- Precision and distance computation results explain the decoding accuracy and speedup of and time.

Language modeling: Impact of vocabulary size

|V|	P@K	FGD (efSearch)			
		20	50	100	200
10K	P@1	0.870	0.938	0.989	1.000
	P@10	0.909	0.972	0.992	0.998
20K	P@1	0.845	0.932	0.975	0.995
	P@10	0.871	0.955	0.987	0.997
40K	P@1	0.808	0.912	0.936	0.980
	P@10	0.845	0.931	0.961	0.991
80K	P@1	0.832	0.933	0.966	0.982
	P@10	0.858	0.945	0.978	0.994



- 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.