Fast softmax approximations

Ann Vlasova, Ann Kuzmina, Nick Osipov, Andrey Zharkov, Kate Adamenko

SVD-Softmax

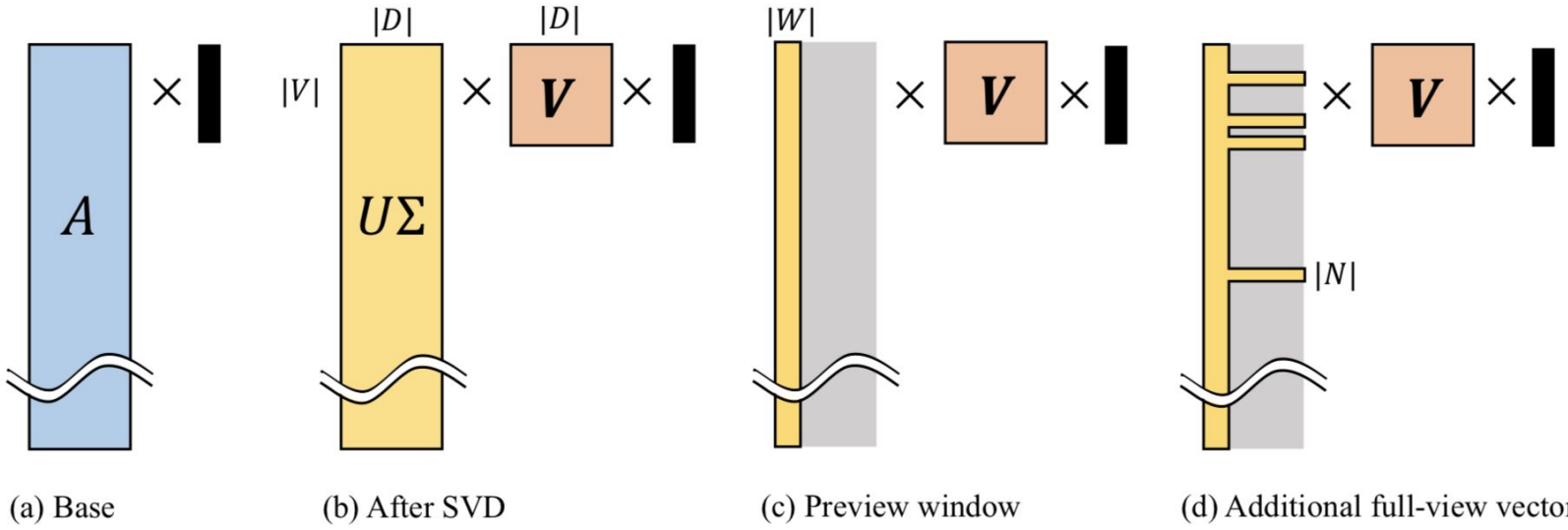


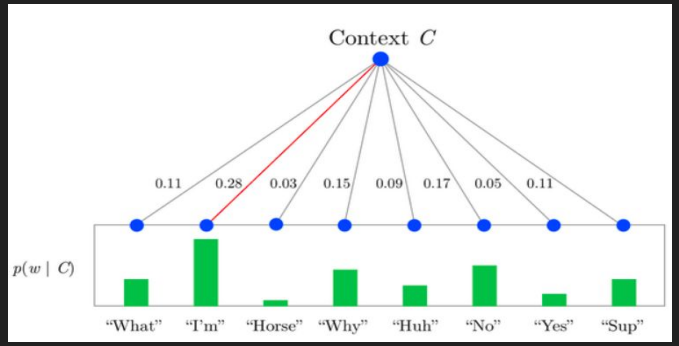
Figure 1: Illustration of the proposed SVD-softmax algorithm. The softmax weight matrix is decomposed by singular value decomposition (b). Only a part of the columns is used to compute the preview outputs (c). Selected rows, which are chosen by sorting the preview outputs, are recomputed with full-width (d). For simplicity, the bias vector is omitted.

SVD-Softmax

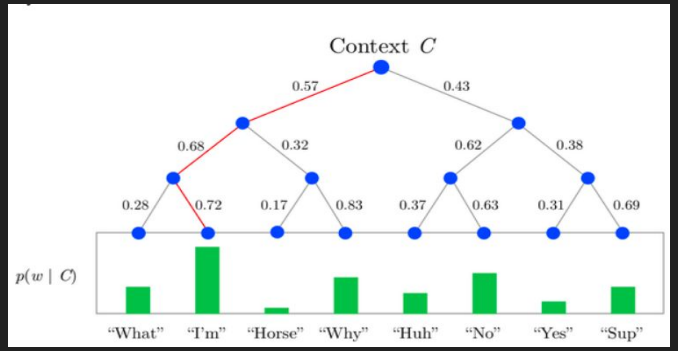
reduces complexity from O(VD) to O(VW+ND), where

- V - vocabulary size
- D - hidden dimension size
- W - size of preview window
- N - words to consider

Hierarchical Softmax



must compute **all N** of the terminal leaves

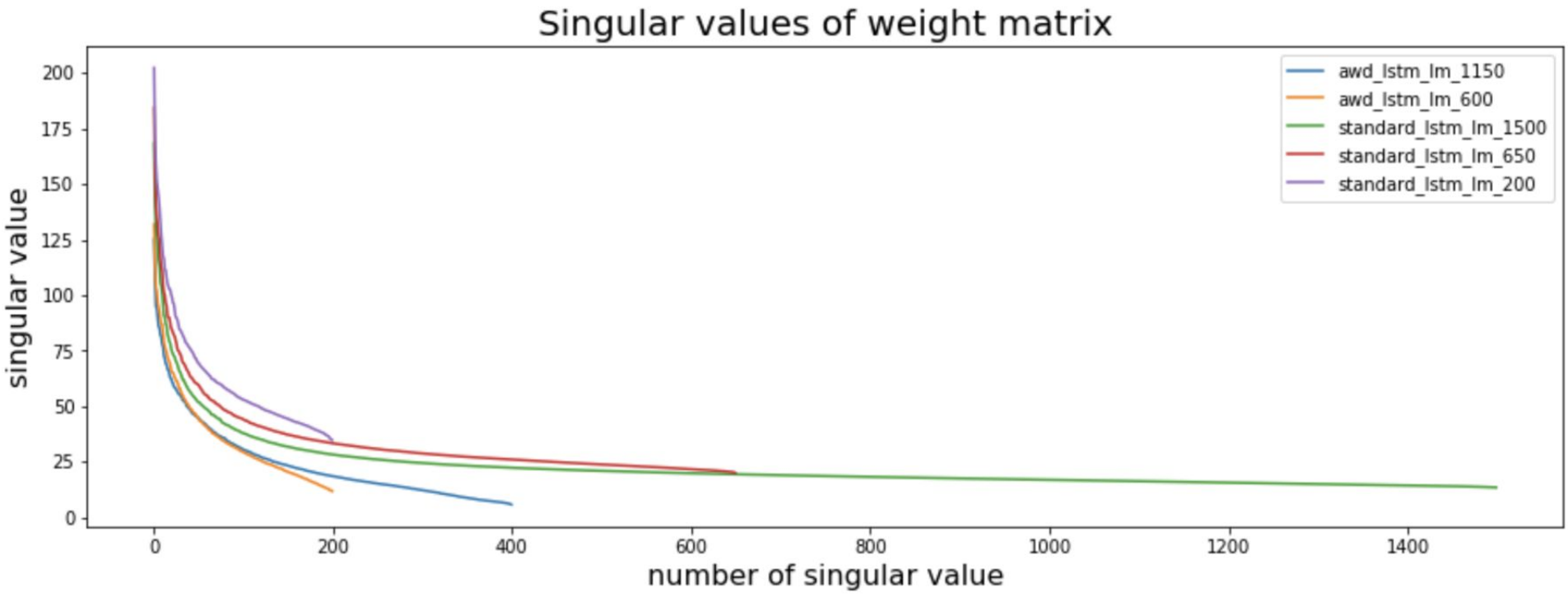


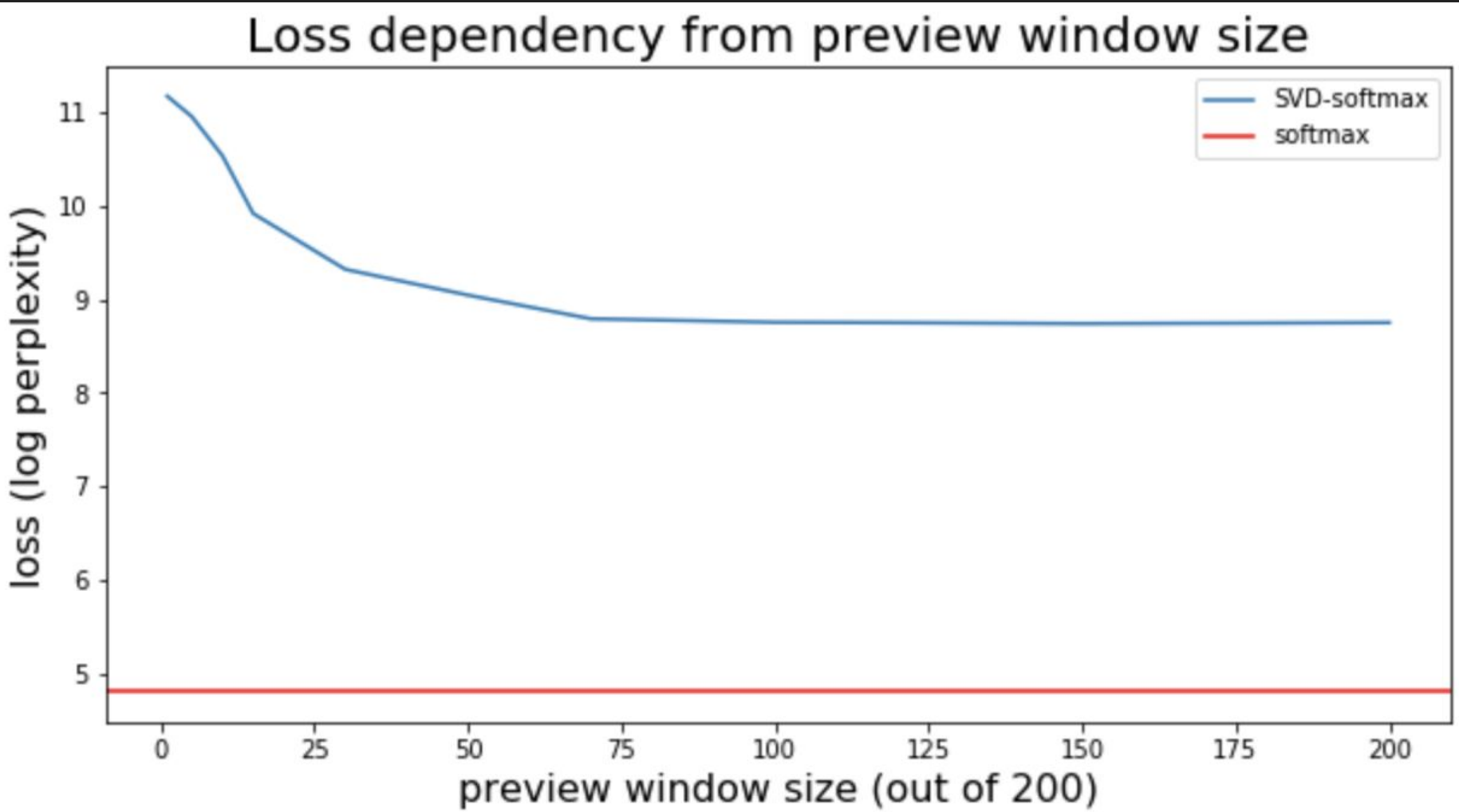
must compute **log(N)** nodes

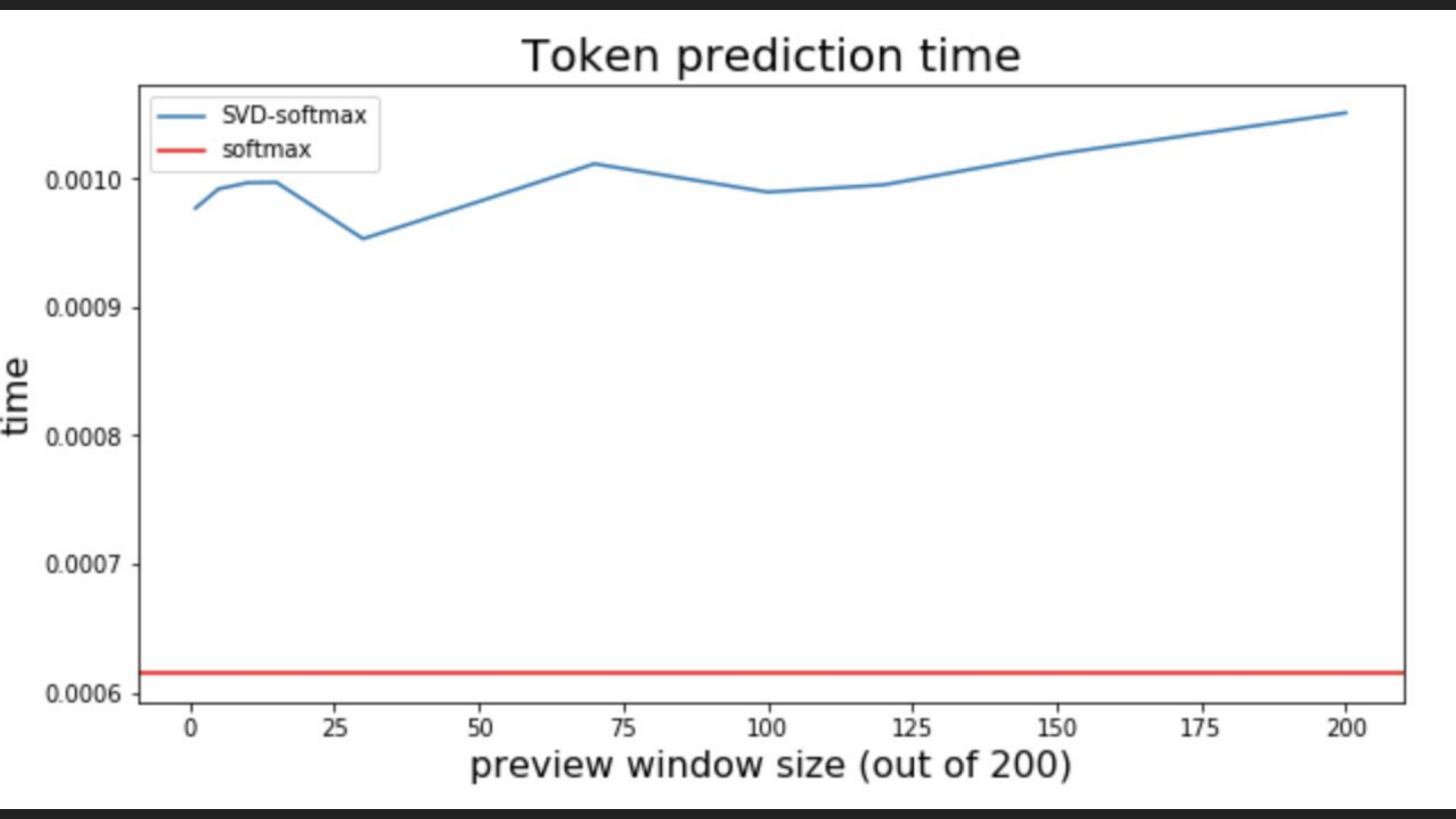
Experiments

- Compare hierarchical softmax with SVD-softmax
- in terms of time and efficiency of approximation (perplexity for LM task)
- top-k approximation match (number of same most probable words in bold softmax and in approximation)
- on wikitext2 and GBW
- time on GPU and CPU

Singular values for different models




Loss dependency from preview window size
SVD-softmax
softmax
loss (log perplexity)
11
10
9
8
7
6
5
0
25
50
75
100
125
150
175
200
preview window size (out of 200)


Token prediction time
SVD-softmax
softmax
0.0010
0.0009
0.0008
0.0007
0.0006
time
0
25
50
75
100
125
150
175
200
preview window size (out of 200)

Summary

- Hierarchical softmax vs SVD softmax
- time and performance (on GPU and CPU)
- datasets: WikiText2, GBW