Doubly Sparse (DS-Softmax): Sparse Mixture of Sparse Experts for Efficient Softmax Inference

Shun Liao[1], Ting Chen*[2], Tian Lin[2],*

Denny Zhou[2], Chong Wang[3]

1. University of Toronto 2. Google 3. ByteDance

EMC2 Workshop @ NeurIPS 2019

Softmax Inference Problem

- **Softmax Inference:** $argmax_c \frac{\exp(W_c\ h)}{z}$**,** where $z = \ \sum_i^N \exp(W_i\ h)$
- **Linear Complexity:** $\boldsymbol{O(N)}$, depends on number of output classes
- Softmax as computional **Bottleneck** example**:**
  - Dataset: Wiki-2, Number of Words = 33k
  - Model: Two layers RNN, hidden size = 200
  - Softmax Computation counts more than 98%
- **Common** in Real Applications: ...
- **Traditional solutions**
  - Treat it as Maximum Inner Product Search (MIPS) in learned Softmax
  - Drawback: they suffer the **accuracy-speedup trade-off**
  - Example: Fast Graph Decoder[1] achieves only **~ 2x** in high accuracy

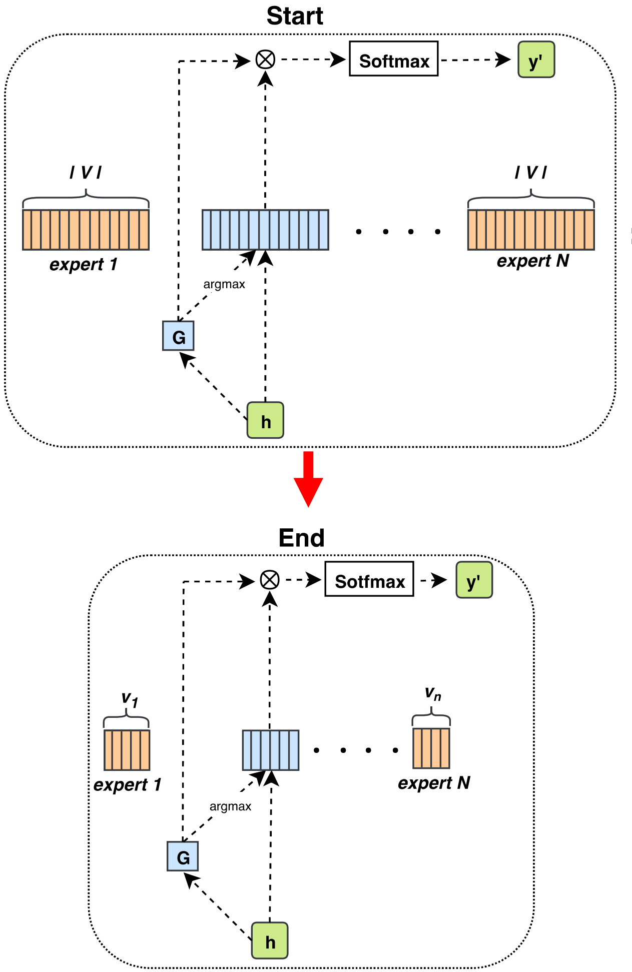
[1] Zhang, M., Wang, W., Liu, X., Gao, J., & He, Y. (2018). Navigating with graph representations for fast and scalable decoding of neural language models. In Advances in Neural Information Processing Systems (pp. 6308-6319).

Doubly Sparse (DS-) Softmax

DS-Softmax: A **learning-based** model which adapts Softmax embedding into **hierarchical** structure for a better trade-off.

Implementation: A mixture of expert model where only the expert with **highest** mixture/gating value is activated

- **Initialization**: each expert contains full output space
- **Training:** iteratively pruning that each expert finally contains a subset of output classes. Then fast search can be achieved

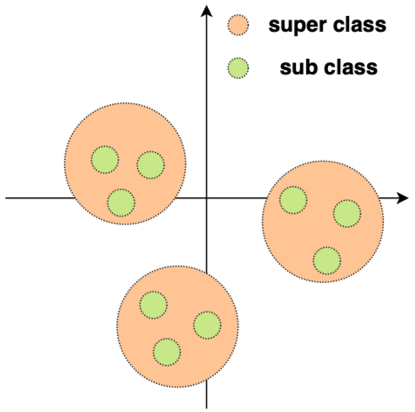


Result – Synthetic Dataset

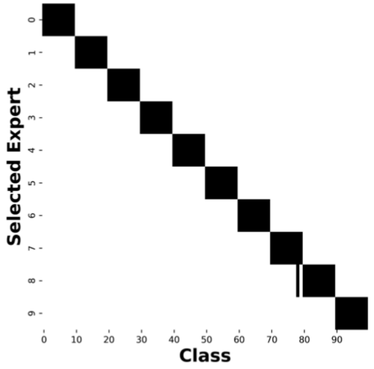
Dataset: two-level hierarchy

- **Generation:**
  - Sample super classes
  - Sample sub around super
  - Sample training points
- Super class label is **hidden**
- **Two sizes**: 100 classes (10 x 10)

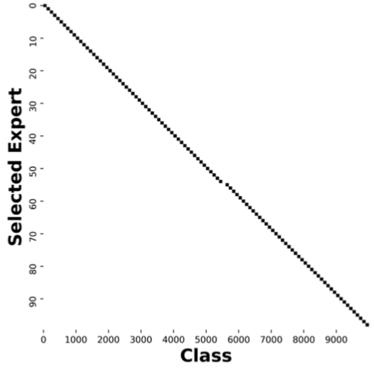
  and 10, 000 (100 x 100)
- **DS-Softmax** can fully capture the synthetic hierarchy



(a) Synthetic Data Generation



(b) Results on 10 x 10

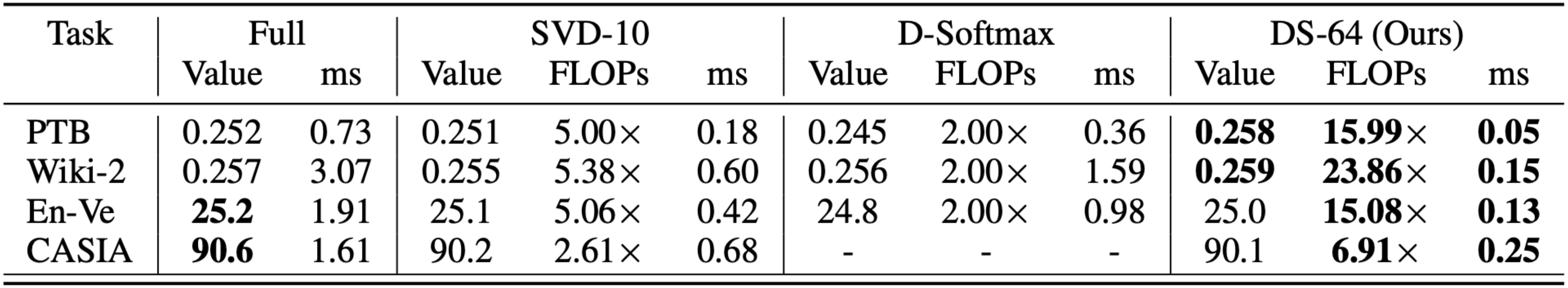


(c) Results on 100 x 100

Result – Real Dataset

DS-Softmax achieves **significant speedup** in three tasks and four dataset **without loss of performance for** theorem and real device

- Number of classes: 10000, 33278, 7709, 3740
- Even boost language modelling performance
- In Wiki-2, number of words = 33,278
  - **23x** Theoretical Reduction
  - **20x** Real Device Reduction

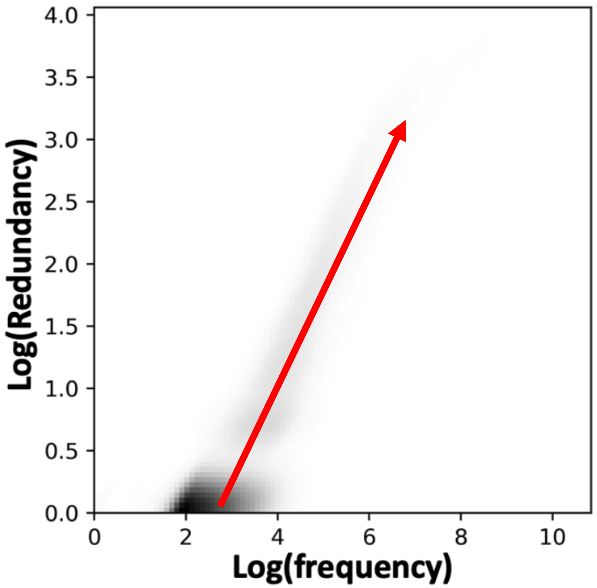


Task	Full		SVD-10			D-Softmax			DS-64 (Ours)		
	Value	ms	Value	FLOPs	ms	Value	FLOPs	ms	Value	FLOPs	ms
PTB	0.252	0.73	0.251	5.00×	0.18	0.245	2.00×	0.36	**0.258**	**15.99**×	**0.05**
Wiki-2	0.257	3.07	0.255	5.38×	0.60	0.256	2.00×	1.59	**0.259**	**23.86**×	**0.15**
En-Ve	**25.2**	1.91	25.1	5.06×	0.42	24.8	2.00×	0.98	25.0	**15.08**×	**0.13**
CASIA	**90.6**	1.61	90.2	2.61×	0.68	-	-	-	90.1	**6.91**×	**0.25**

Result – Interpretation

Higher frequency words appear in more experts.

- Similar in topic model[1]
- High frequency words requires more expressive models[2]



The smallest expert in PTB, where 64 words left

- **Time is Money !!!**

Money

- million, billion, trillion, earnings, share, rate, stake, bond, cents, bid, cash, fine, payable

Time

- years, while, since, before, early, late, yesterday, annual, currently, monthly, annually, Monday, Tuesday, Wednesday, Thursday, Friday

Comparison

- up, down, under, above, below, next, though, against, during, within, including, range, higher, lower, drop, rise, growth, increase, less, compared, unchanged

1. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research.
2. Grave, E., Joulin, A., Cissé, M., & Jégou, H. (2017, August). Efficient softmax approximation for GPUs. ICML