

Appendix for Improving Word Embedding Factorization for Compression Using Distilled Nonlinear Neural Decomposition

Vasileios Lioutas^{1*}, Ahmad Rashid², Krtin Kumar²,
Md. Akmal Haidar², Mehdi Rezagholizadeh²

¹University of British Columbia, ²Huawei Noah's Ark Lab

contact@vlioutas.com, ahmad.rashid@huawei.com, krtin.kumar@huawei.com,
md.akmal.haidar@huawei.com, mehdi.rezagholizadeh@huawei.com

1 Additional Hyper-parameters

WMT En-Fr Smaller Transformer Network denotes a network with the same configuration as Transformer Base but with hidden size d_{model} of 416. For GroupReduce, to match the same compression rate we used number of clusters c being equal to 10 and minimum rank r_{min} to be 22. For SVD, we decided to set the rank to 64. For Tensor Train, we set the embedding shape to be $[25, 32, 40] \times [8, 8, 8]$ and the Tensor Train Rank to be 90. For structured embedding we use group size as 32 and number of clusters as 2048, we then use the quantization matrix and learn the clusters from scratch.

WMT En-De Smaller Transformer Network denotes a network with the same configuration as Transformer Base but with hidden size d_{model} of 400. For GroupReduce, to match the same compression rate we used number of clusters c being equal to 10 and minimum rank r_{min} to be 23. For SVD, we decided to set the rank to 64. For Tensor Train, we set the embedding shape to be $[25, 37, 40] \times [8, 8, 8]$ and the Tensor Train Rank to be 90. For structured embedding we use group size as 32 and number of clusters as 2376, we then use the quantization matrix and learn the clusters from scratch.

IWSLT Pt-En Smaller Transformer Network denotes a network with the same configuration as Transformer Small but with hidden size d_{model} of 136. For GroupReduce, to match the same compression rate we used number of clusters c being equal to 15 and minimum rank r_{min} to be 30. For SVD, we decided to set the rank to 64. For Tensor Train, we set the embedding shape to be $[25, 32, 40] \times [8, 4, 8]$ and the Tensor Train Rank to

be 125. For structured embedding we use group size as 32 and number of clusters as 4048, we then use the quantization matrix and learn the clusters from scratch.

2 Parameter count

Table 1 presents the the number of parameters in the different transformer layers for the transformer base architecture.

References

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

*Work done during an internship at Huawei Noah's Ark Lab.

Parameters	Embedding	FFN	Multi-head attention	Linear
Number	26M	25M	14M	5M
Percentage	37%	36%	20%	7%

Table 1: Parameters in the Transformer Base model ([Vaswani et al., 2017](#)) based on a 50k dictionary size and tied input and output embedding.