DEFINE: DEEP FACTORIZED INPUT TOKEN EMBEDDINGS FOR NEURAL SEQUENCE MODELING

SACHIN MEHTA

ADVISORS: Prof. Linda Shapiro and Prof. Hannaneh Hajishirzi

01

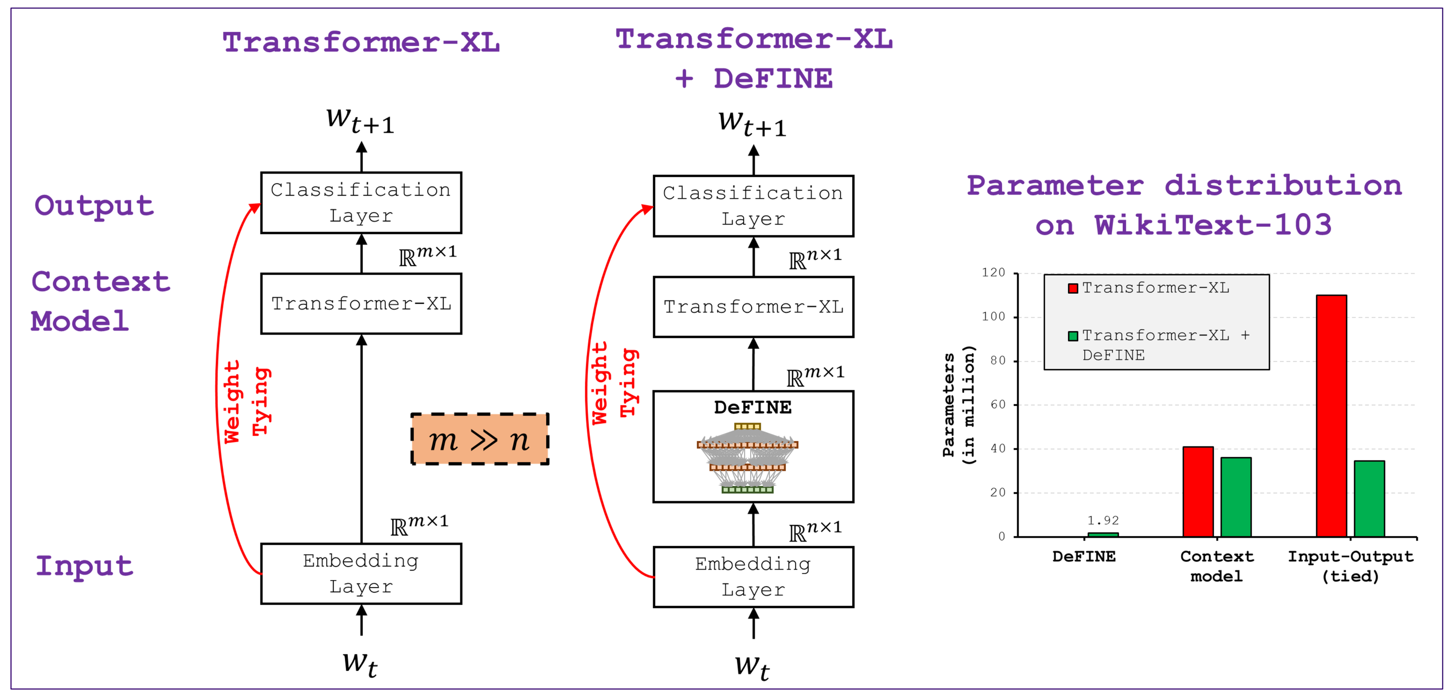
Introduction

- Learning input embeddings
  - Large token-level vocabularies (e.g., WikiText-103's vocab size is 267K)
  - Most of the parameters in the input and the output layers
  - Uses shallow look-up table with medium dimensional embedding

$$e = f(w)$$

- Our approach (MER principle):
  - ***Map*** to low-dimensional look-up table (64- or 128-dimensional)
  - ***Expand*** to a very high-dimensional space (say **4096**-dimensional)
    - Efficient Hierarchical Group Transform
    - New skip-connection that establishes direct link with the Map layer
  - ***Reduce*** (or project) linearly to low dimensional space

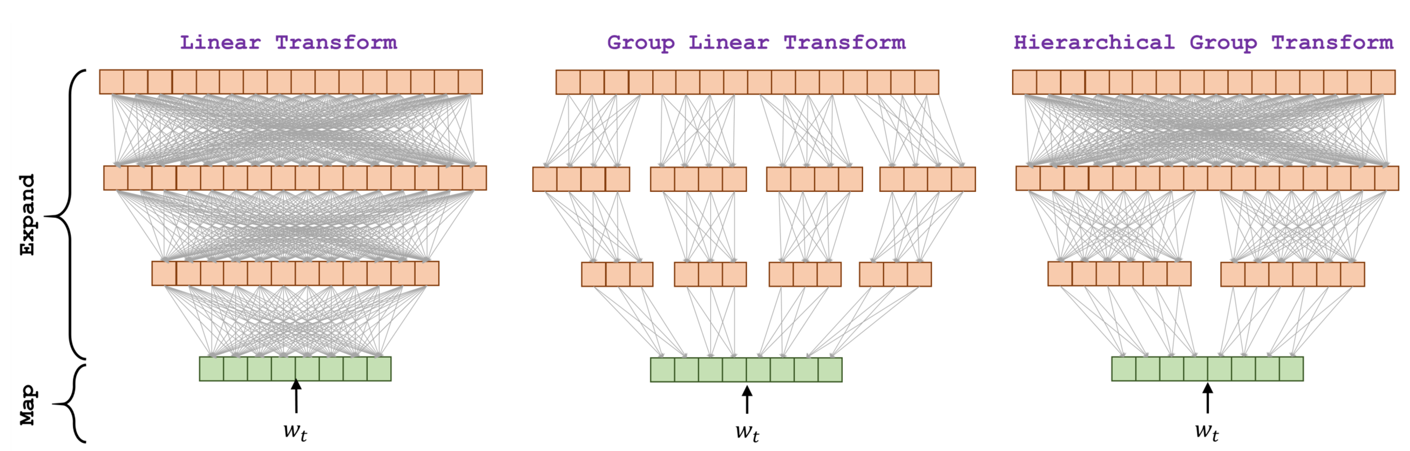
$$e = h\big(g(w)\big)$$



With DeFINE, Transformer-XL learns input and output representations in low-dimensional space with minimal impact on performance.

Hierarchical Group Transformation (HGT)

- Sparse and dense connections
- Multiple paths to the **Map** layer
- Nearly as efficient as group transform and as accurate as linear

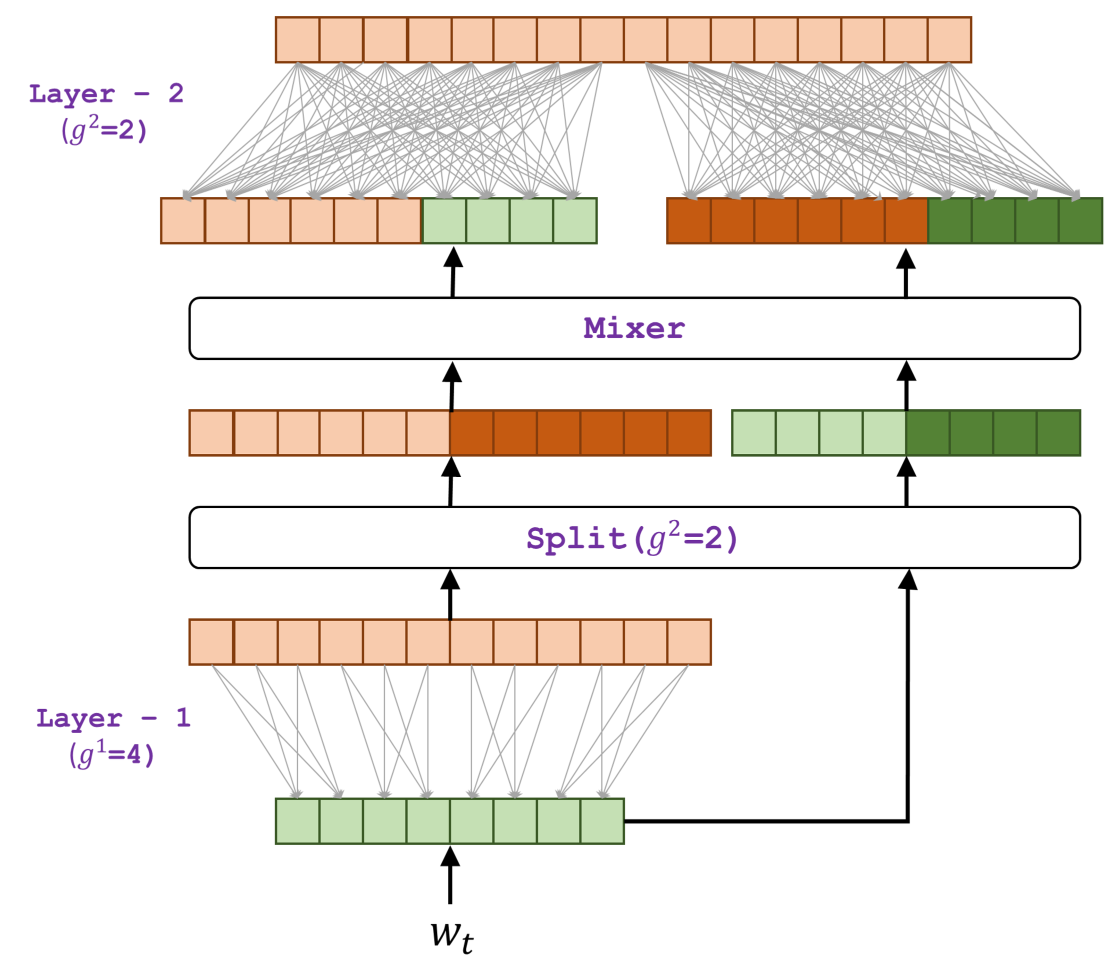


DeFINE: Learning Deep Token Representations Efficiently

- Built using **MER** principle
- Uses **HGT** to learn representations in high-dimensional space efficiently
- **Direct connection with the input layer**
  - Enables training deeper networks
  - Improves performance
  - Learns better representations
  - Gradients flow-back via multiple paths

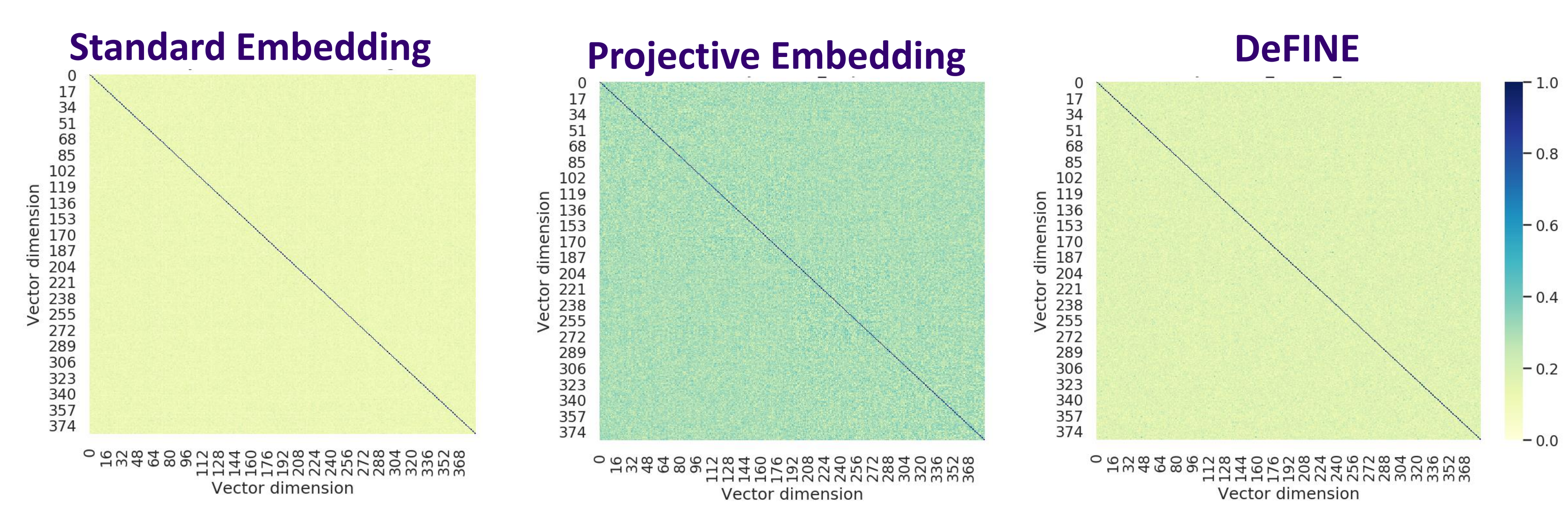
Transform	Parameters	Perplexity
Linear	42.86	41.19
Group	**39.69**	45.63
HGT	40.73	40.92
DeFINE	40.89	**38.01**

Comparison on WikiText-103 Test set. Lower value of perplexity is better.



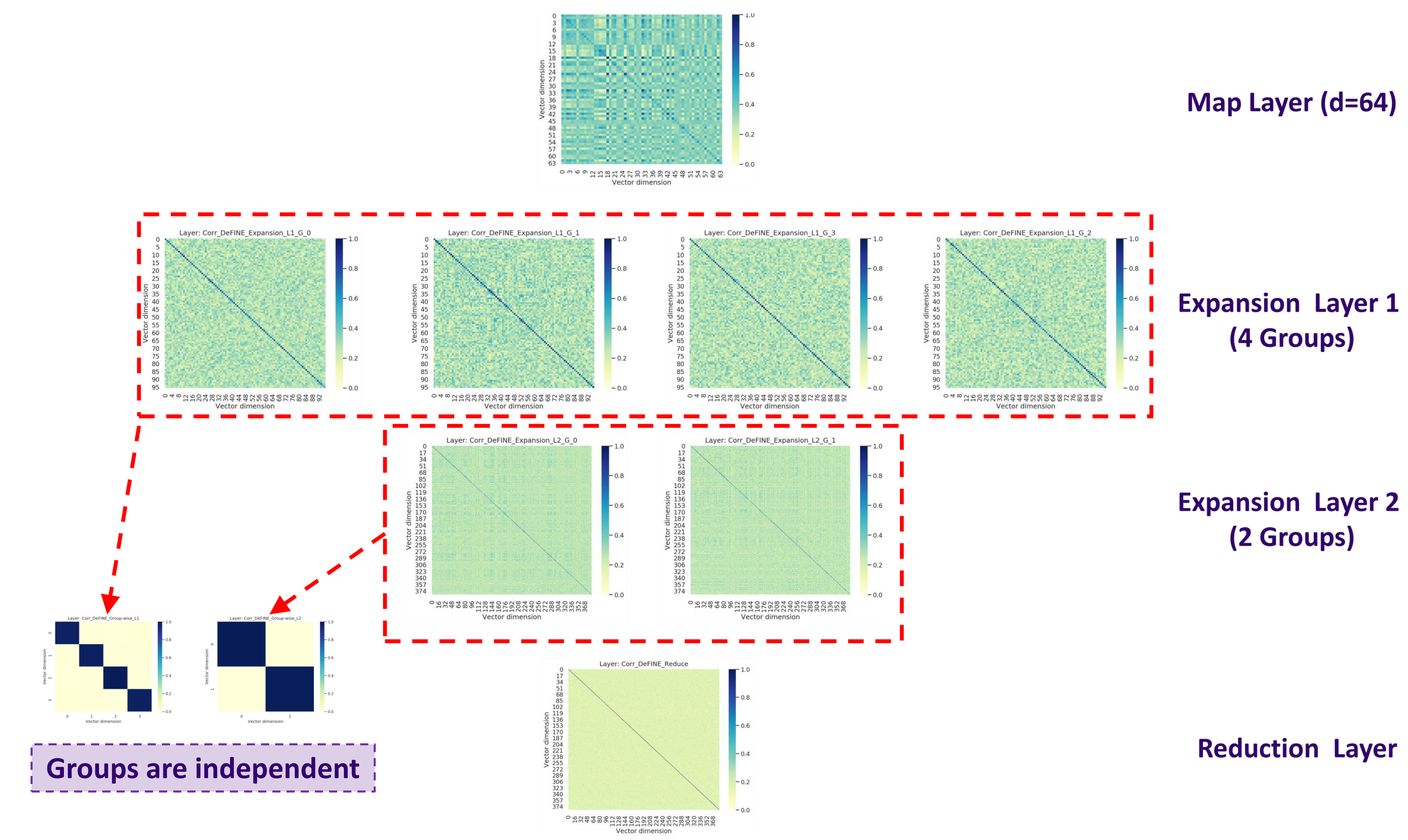
Comparison with Different Embedding Methods

- Embedding dimensions should be **independent**
- Similar to standard embeddings, DeFINE embeddings
  - Does not have correlation between dimensions i.e. independent
  - Approximates standard embeddings efficiently than widely used projective embeddings



How DeFINE Approximates Standard Embedding?

- Low-dimensional mapping layer has correlations
- With depth, correlations reduces and DeFINE approximates standard embedding layer.
- Importantly, the **groups** at different layers in DeFINE are **independent**
  - Suggests matrices are learning different representations of their input



Integrating DeFINE with State-of-the-art Sequence Models

AWD-LSTM on PenTree Bank Dataset

Small language modeling dataset with 10K unique tokens

Lower perplexity value is better.

Model	Parameters	Perplexity
AWD-LSTM	24 M	58.8
AWD-LSTM + DeFINE	**20 M**	**54.2**

Transformer-XL on WikiText-103

Medium language modeling dataset with 270K unique tokens

Lower perplexity value is better.

Model	Parameters	Perplexity
Transformer-XL	139 M	27.06
Transformer-XL + DeFINE	**73 M**	**26.33**

Transformer on WMT14 EN-DE

Large scale machine translation dataset (English to German)

High BLEU score is better.

Model	Parameters	BLEU
Transformer	92 M	25.81
Transformer + DeFINE	**68 M**	**28.25**

References

- **AWD-LSTM:** Merity, Stephen, Nitish Shirish Keskar, and Richard Socher. "Regularizing and optimizing LSTM language models." ICLR (2018).
- **Transformer-XL:** Dai, Zihang, et al. "Transformer-xl: Attentive language models beyond a fixed-length context." ACL (2019).
- **Transformer**: Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

ELECTRICAL & COMPUTER ENGINEERING
UNIVERSITY of WASHINGTON

See this paper for more details:
Mehta, Sachin, et al. "*DeFINE: DEep Factorized INput Word Embeddings for Neural Sequence Modeling*", International Conference on Learning Representations (ICLR) (2020).