Journal Club: Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context

Jitian Zhao

University of Wisconsin Madison

November 21, 2019

Outline

- Problems of Vanilla Transformer
- Special Techniques in Transformer-XL
- State-of-the-art Results

Problems of Vanilla Transformer

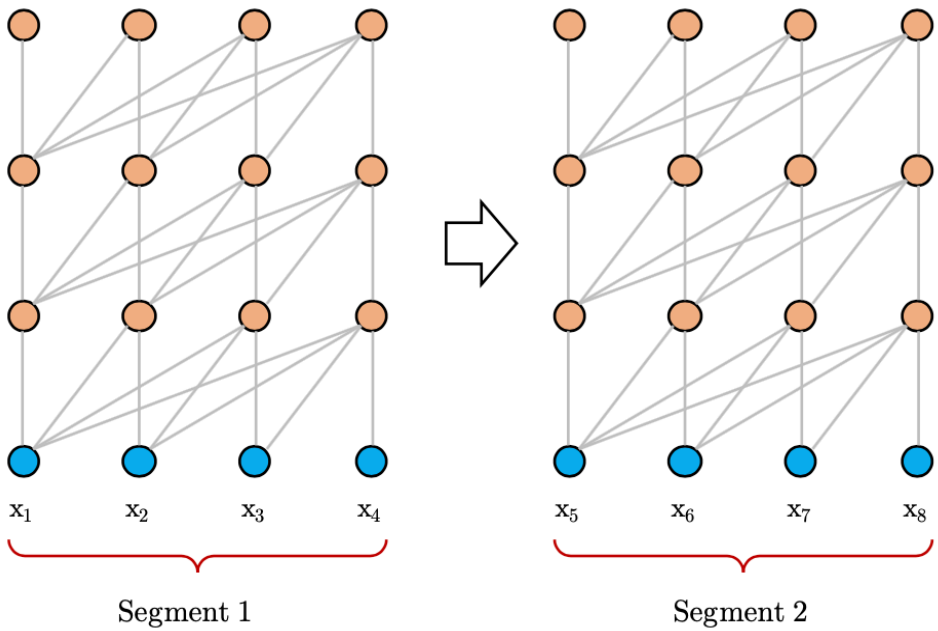

Figure: train phase

- not able to model dependencies that are longer than a fixed length
- context fragmentation

Problems of Vanilla Transformer

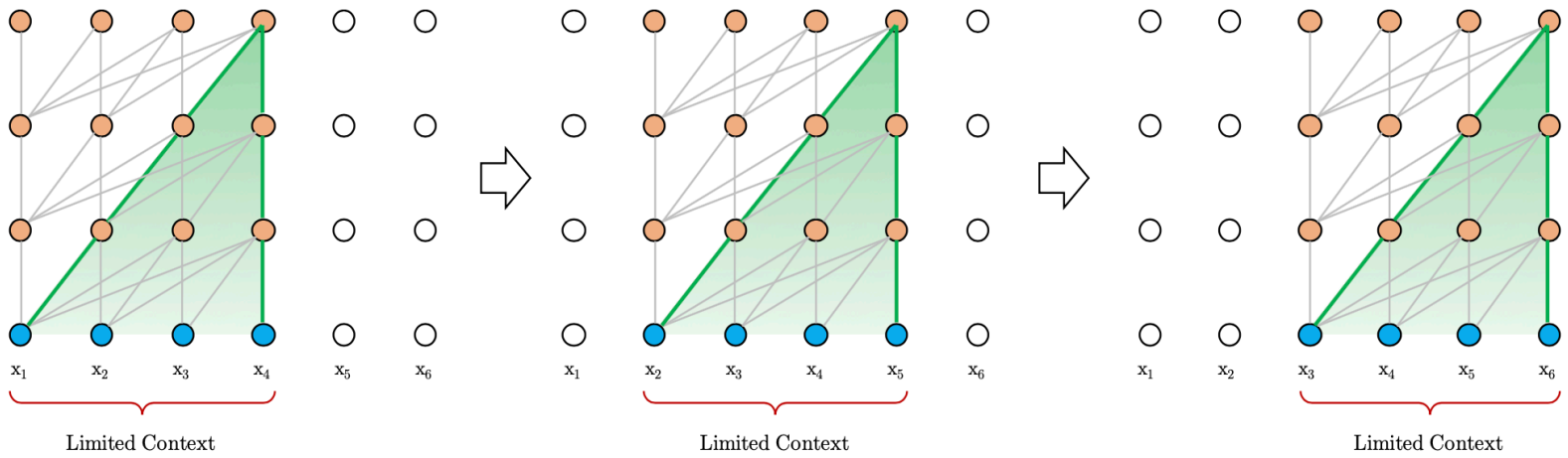

Figure: evaluation phase

- inefficient evaluation procedure

Special Techniques in Transformer-XL

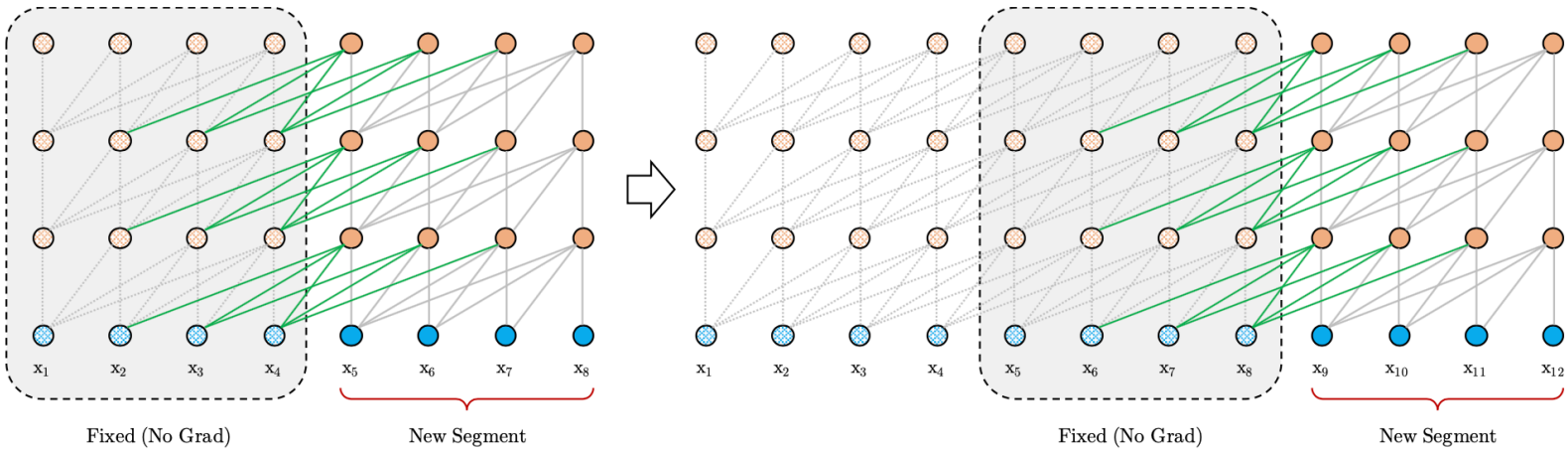

Figure: train phase

- **technique**: segment-level recurrence
- **key difference**: updates of key and value rely on extended context
- **benefits**: model extra long context, faster evaluation

Comparison of evaluation process

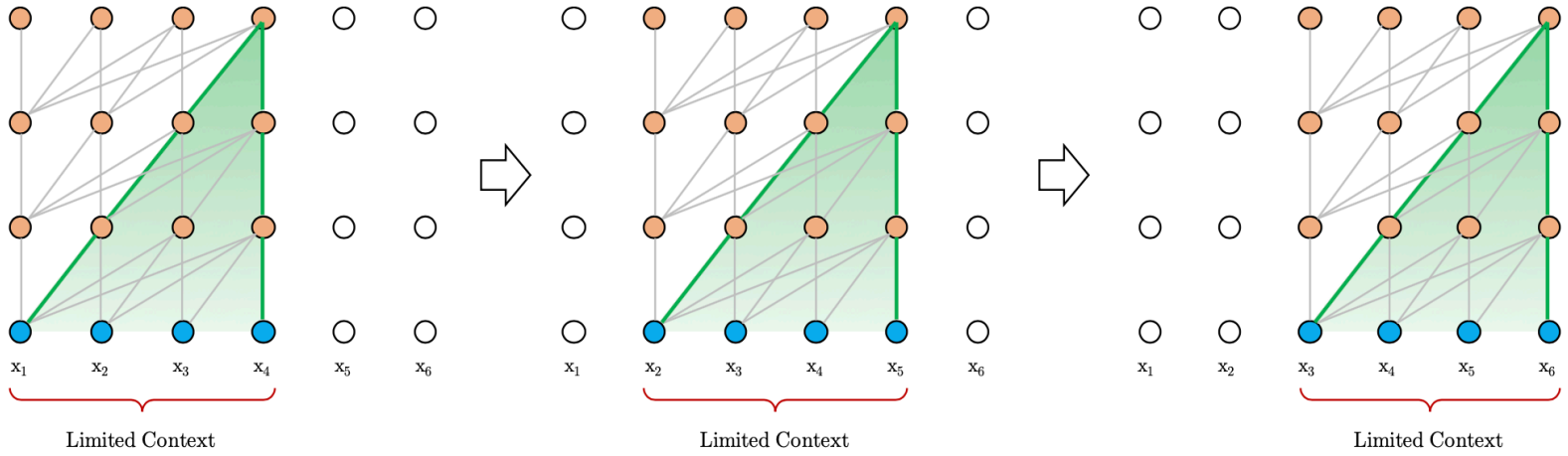

Figure: vanilla Transformer

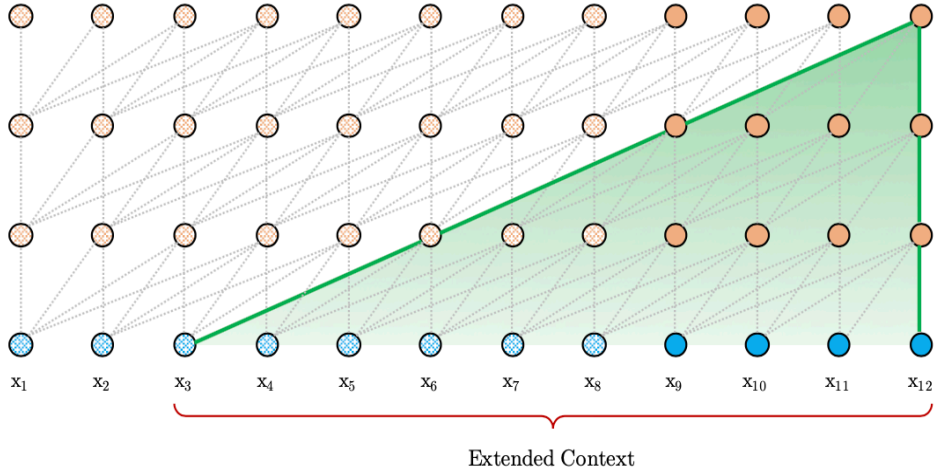

Figure: Transformer-XL

Special Techniques in Transformer-XL

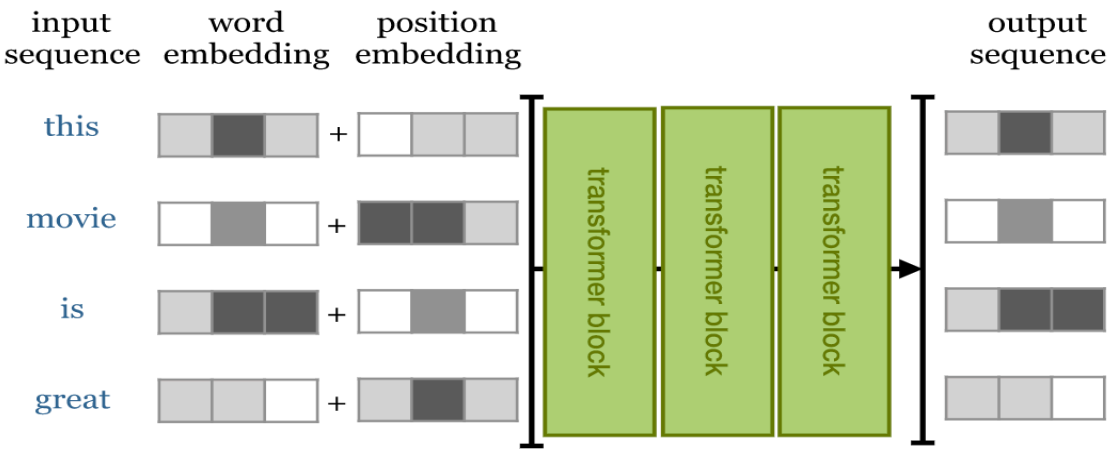

Figure: original position encoding

- **technique**: relative positional encoding
- **original**: U_i represents absolute position within a segment
- **modified**: R_i represents a relative distance of i between two position

Mathematical Explanation

$$\mathbf{A}^{\text{abs}}_{i,j} = \underbrace{\mathbf{E}_{x_i}^\top \mathbf{W}_q^\top \mathbf{W}_k \mathbf{E}_{x_j}}_{(a)} + \underbrace{\mathbf{E}_{x_i}^\top \mathbf{W}_q^\top \mathbf{W}_k \mathbf{U}_j}_{(b)} + \underbrace{\mathbf{U}_i^\top \mathbf{W}_q^\top \mathbf{W}_k \mathbf{E}_{x_j}}_{(c)} + \underbrace{\mathbf{U}_i^\top \mathbf{W}_q^\top \mathbf{W}_k \mathbf{U}_j}_{(d)}.$$

(a) original

$$\mathbf{A}^{\text{rel}}_{i,j} = \underbrace{\mathbf{E}_{x_i}^\top \mathbf{W}_q^\top \mathbf{W}_{k,E} \mathbf{E}_{x_j}}_{(a)} + \underbrace{\mathbf{E}_{x_i}^\top \mathbf{W}_q^\top \mathbf{W}_{k,R} \mathbf{R}_{i-j}}_{(b)} + \underbrace{u^\top \mathbf{W}_{k,E} \mathbf{E}_{x_j}}_{(c)} + \underbrace{v^\top \mathbf{W}_{k,R} \mathbf{R}_{i-j}}_{(d)}.$$

(b) modified

Figure: attention score between query q_i and key k_j

- absolute position $U_i \rightarrow$ relative positional encoding R_{i-j}
- query $U_i^T W_q^T \rightarrow$ trainable parameter u, v in (c),(d)
- seperate W into $W_{k,E}, W_{k,R}$ to represent content-based key vector and location-based key vector

Results

- **Long-term and short-term dependency**: Reduced perplexity (exponentiation of the entropy) on benchmark datasets
- **Ablation study**: Both segment level recurrence and relative position encoding are necessary
- **Evaluation speed**: Compared with vanilla transformer


HAPPY
Thanksgiving
CrossCards