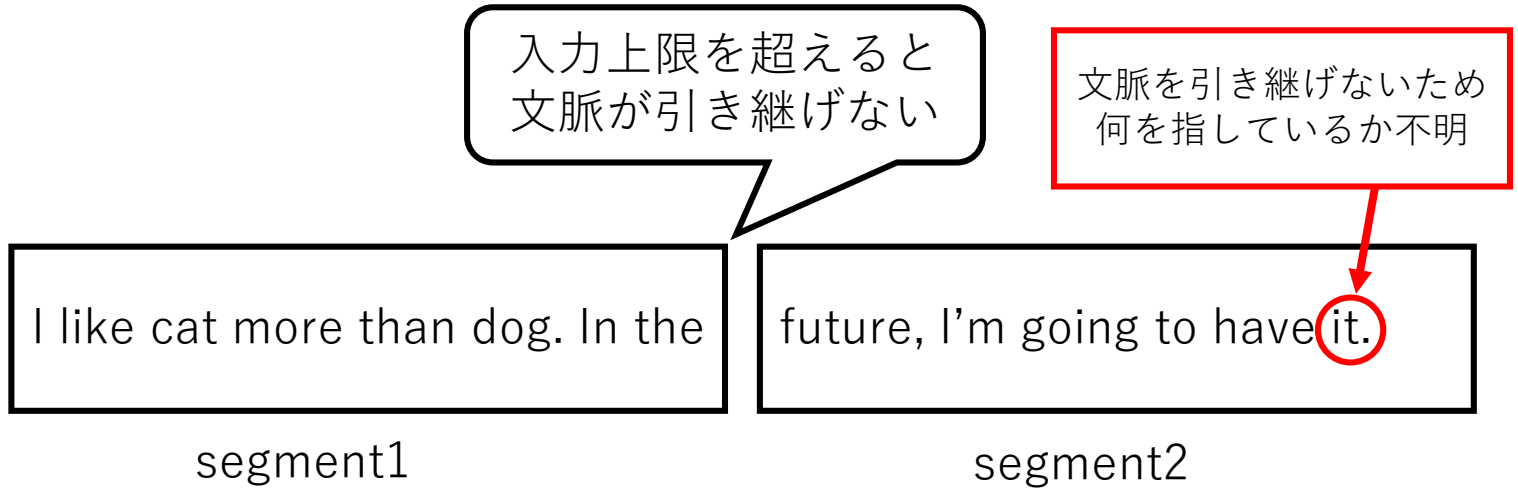
# Transformer-XL[*]

[*]Dai, Zihang, et al. "Transformer-xl: Attentive language models beyond a fixed-length context." *arXiv preprint arXiv:1901.02860* (2019).

# 研究背景

- Transformer
  - 自然言語における強力なモデル
  - CNNやRNNよりもsegment内での長期的な文脈を利用可能
- Transformerの限界
  - segmentの境界付近の短期的な文脈が考慮不能
  - segmentを跨ぐ長期的な文脈が考慮されない




研究目的：segment間の文脈の途切れを改善

[*] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.

# 概要

- 従来のTransformer
  - segmentごとの学習
  - segment長の文脈で評価
- Transformer-XL
  - 前セグメントの情報もつかって学習
  - segment長を超える文脈で評価
  - 評価時にsegmentごとの再計算が不要
    ⇒計算量が減少

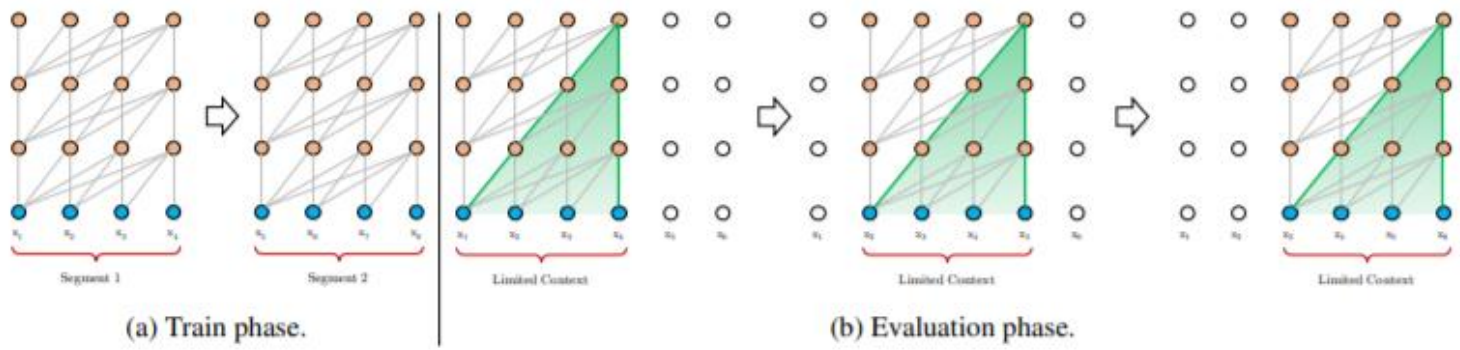



図1 従来のtransformerにおける学習と評価[*]

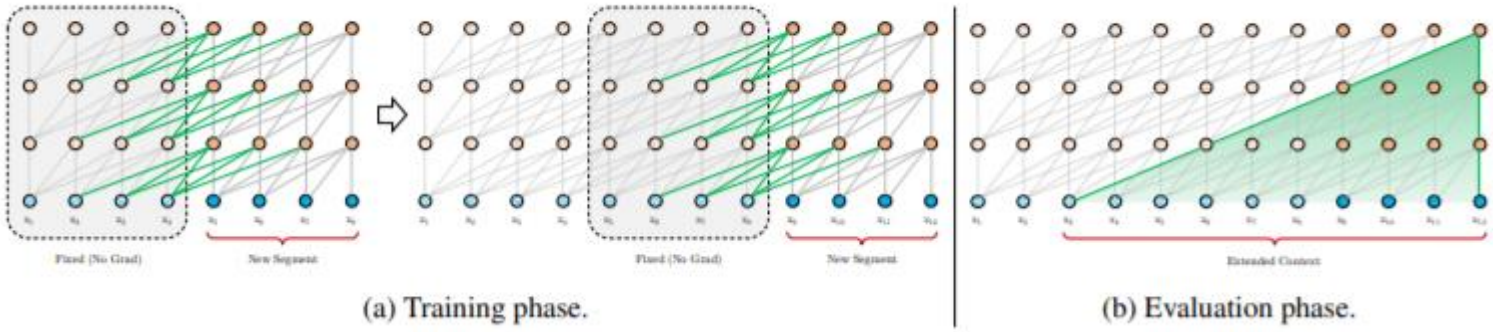
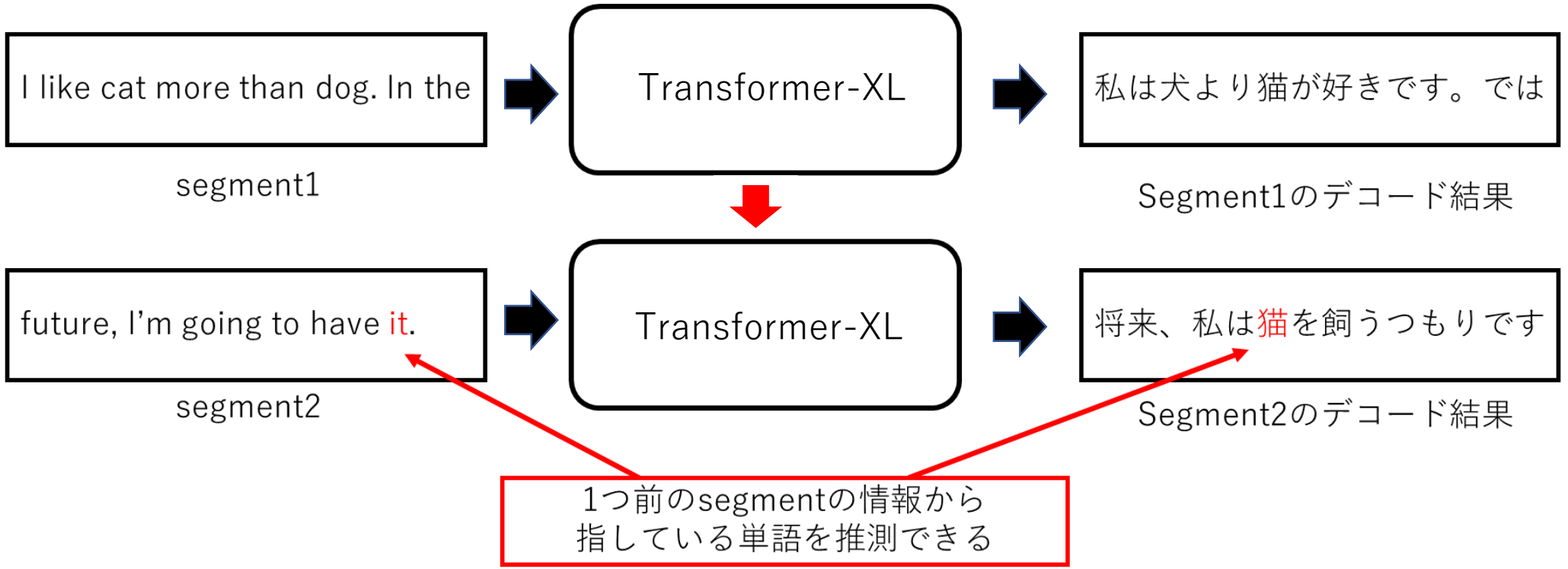



図2 transformer-xlにおける学習と評価[*]

[*]Dai, Zihang, et al. "Transformer-xl: Attentive language models beyond a fixed-length context." *arXiv preprint arXiv:1901.02860* (2019).

# 手法(1/3)

- Segmentレベルの回帰
  - 一つ前のsegmentの出力を、次のsegmentの計算時に利用



# 手法(2/3)

- Segmentレベルの回帰
  - 一つ前のsegmentの出力を、次のsegmentの計算時に利用

$$\widetilde{\mathbf{h}}_{\tau+1}^{n-1} = \left[\mathrm{SG}(\mathbf{h}_{\tau}^{n-1}) \circ \mathbf{h}_{\tau+1}^{n-1}\right]$$

$$\mathbf{q}_{\tau+1}^{n}, \mathbf{k}_{\tau+1}^{n}, \mathbf{v}_{\tau+1}^{n} = \mathbf{h}_{\tau+1}^{n-1}\mathbf{W}_{q}^{\top}, \widetilde{\mathbf{h}}_{\tau+1}^{n-1}\mathbf{W}_{k}^{\top}, \widetilde{\mathbf{h}}_{\tau+1}^{n-1}\mathbf{W}_{v}^{\top}$$

$$\mathbf{h}_{\tau+1}^{n} = \text{Transformer-Layer}\left(\mathbf{q}_{\tau+1}^{n}, \mathbf{k}_{\tau+1}^{n}, \mathbf{v}_{\tau+1}^{n}\right)$$

query. key valueについて詳しくは従来のTransformerを参照[*]

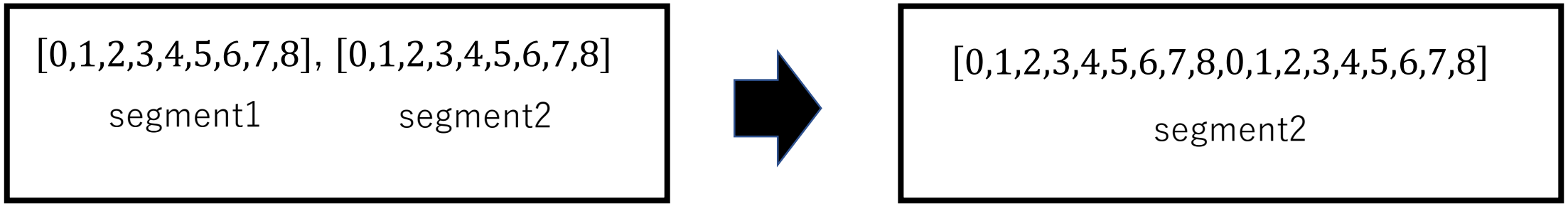
| | | | |
|---|---|---|---|
| $SG(\cdot)$ | : stop-gradient | | |
| $[h_u \circ h_v]$ | : $h_u$と$h_v$の結合 | | |
| Transformer-layer | : 従来のtransformerの働きをする層 | | |
| $h_\tau^n$ | : τ番目のsegmentに対するn層目の隠れ層の出力 | | |
| $\tilde{h}_\tau^n$ | : τ番目のsegmentに対するn層目の隠れ層の出力に、τ-1番目のsegmentの隠れ層を連結した出力 | | |
| $q_\tau^n$ | : τ番目のsegmentに対するn番目の query | $W_q$ | : n層目のqueryを計算する全結合層の重み |
| $k_\tau^n$ | : τ番目のsegmentに対するn番目の key | $W_k$ | : n層目のkeyを計算する全結合層の重み |
| $v_\tau^n$ | : τ番目のsegmentに対するn番目の value | $W_v$ | : n層目のvalueを計算する全結合層の重み |

[*] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.

# 手法(3/3)

- 従来のTransformer
  - 絶対位置符号化

segment1は、その前に
依存関係がないため省略




対応するsegmentごとの位置の埋め込み表現

従来の位置の埋め込み表現をTransformer-XLに利用する場合

- Transformer-XL
  - 相対位置符号化
  - ある単語からの位置関係を埋め込み表現

# 実験(1/2)

- 長期の依存関係が必要となるデータセットでSoTA(論文作成時)

| Model | #Param | PPL |
|---|---|---|
| Grave et al. (2016b) - LSTM | - | 48.7 |
| Bai et al. (2018) - TCN | - | 45.2 |
| Dauphin et al. (2016) - GCNN-8 | - | 44.9 |
| Grave et al. (2016b) - LSTM + Neural cache | - | 40.8 |
| Dauphin et al. (2016) - GCNN-14 | - | 37.2 |
| Merity et al. (2018) - QRNN | 151M | 33.0 |
| Rae et al. (2018) - Hebbian + Cache | - | 29.9 |
| Ours - Transformer-XL Standard | 151M | **24.0** |
| Baevski and Auli (2018) - Adaptive Input$^{\diamond}$ | 247M | 20.5 |
| Ours - Transformer-XL Large | 257M | **18.3** |

Table 1: Comparison with state-of-the-art results on WikiText-103. $^{\diamond}$ indicates contemporary work.

| Model | #Param | bpc |
|---|---|---|
| Ha et al. (2016) - LN HyperNetworks | 27M | 1.34 |
| Chung et al. (2016) - LN HM-LSTM | 35M | 1.32 |
| Zilly et al. (2016) - RHN | 46M | 1.27 |
| Mujika et al. (2017) - FS-LSTM-4 | 47M | 1.25 |
| Krause et al. (2016) - Large mLSTM | 46M | 1.24 |
| Knol (2017) - cmix v13 | - | 1.23 |
| Al-Rfou et al. (2018) - 12L Transformer | 44M | 1.11 |
| Ours - 12L Transformer-XL | 41M | **1.06** |
| Al-Rfou et al. (2018) - 64L Transformer | 235M | 1.06 |
| Ours - 18L Transformer-XL | 88M | 1.03 |
| Ours - 24L Transformer-XL | 277M | **0.99** |

Table 2: Comparison with state-of-the-art results on enwik8.

図1 WikiText-103におけるPPL, enwiki8におけるbpcの比較結果[*]

[*]Dai, Zihang, et al. "Transformer-xl: Attentive language models beyond a fixed-length context." *arXiv preprint arXiv:1901.02860* (2019).

# 実験(2/2)

- 短期の依存関係が必要となるデータセットでSoTA(論文作成時)

| Model | #Param | PPL |
|---|---|---|
| Shazeer et al. (2014) - Sparse Non-Negative | 33B | 52.9 |
| Chelba et al. (2013) - RNN-1024 + 9 Gram | 20B | 51.3 |
| Kuchaiev and Ginsburg (2017) - G-LSTM-2 | - | 36.0 |
| Dauphin et al. (2016) - GCNN-14 bottleneck | - | 31.9 |
| Jozefowicz et al. (2016) - LSTM | 1.8B | 30.6 |
| Jozefowicz et al. (2016) - LSTM + CNN Input | 1.04B | 30.0 |
| Shazeer et al. (2017) - Low-Budget MoE | ~5B | 34.1 |
| Shazeer et al. (2017) - High-Budget MoE | ~5B | 28.0 |
| Shazeer et al. (2018) - Mesh Tensorflow | 4.9B | 24.0 |
| Baevski and Auli (2018) - Adaptive Input$^{\diamond}$ | 0.46B | 24.1 |
| Baevski and Auli (2018) - Adaptive Input$^{\diamond}$ | 1.0B | 23.7 |
| Ours - Transformer-XL Base | 0.46B | 23.5 |
| Ours - Transformer-XL Large | 0.8B | **21.8** |

Table 4: Comparison with state-of-the-art results on One Billion Word. $^{\diamond}$ indicates contemporary work.

図1 One Billion WordにおけるPPLの比較結果

[*]Dai, Zihang, et al. "Transformer-xl: Attentive language models beyond a fixed-length context." *arXiv preprint arXiv:1901.02860* (2019).

# まとめ

- 自然言語処理で強力なモデルであるtransformerを改良
- segmentを越えた長期の依存関係の学習が可能
- 文脈の断片化を解消して短期の依存関係でも性能向上
- 評価時の計算時間が減少