XLNet:
Generalized Autoregressive Pretraining for Language Understanding

QRAFT | AXE
GUIJIN.SON

2021.07.14

AutoEncoding & AutoRegressive

XLNet : Generalized Autoregressive Pretraining for Language Understanding

$$AE : \max_{\theta} \log P_{\theta}(X) \approx \sum_{t=1}^{T} m_t \log P_{\theta}(x_t \mid \hat{x})$$

$$AR : \max_{\theta} \log P_{\theta}(X) = \sum_{t=1}^{T} \log P_{\theta}(x_t \mid x_{<t})$$

Limitations of AutoEncoding Models (BERT)

XLNet : Generalized Autoregressive Pretraining for Language Understanding

- **Independence Assumption**

  I study Artificial Intelligence => I study [MASK] [MASK].

  $$\log_p(\text{Artificial} \mid \text{I study}) + \log_\text{p}(\text{Intelligence} \mid \text{I study})$$

- **Input Noise | Pretrain-Finetune Discrepancy**

  [MASK] causes Corruption

AutoEncoders are Bidirectional

XLNet : Generalized Autoregressive Pretraining for Language Understanding

RNN -> BiRNN (AutoRegressive)

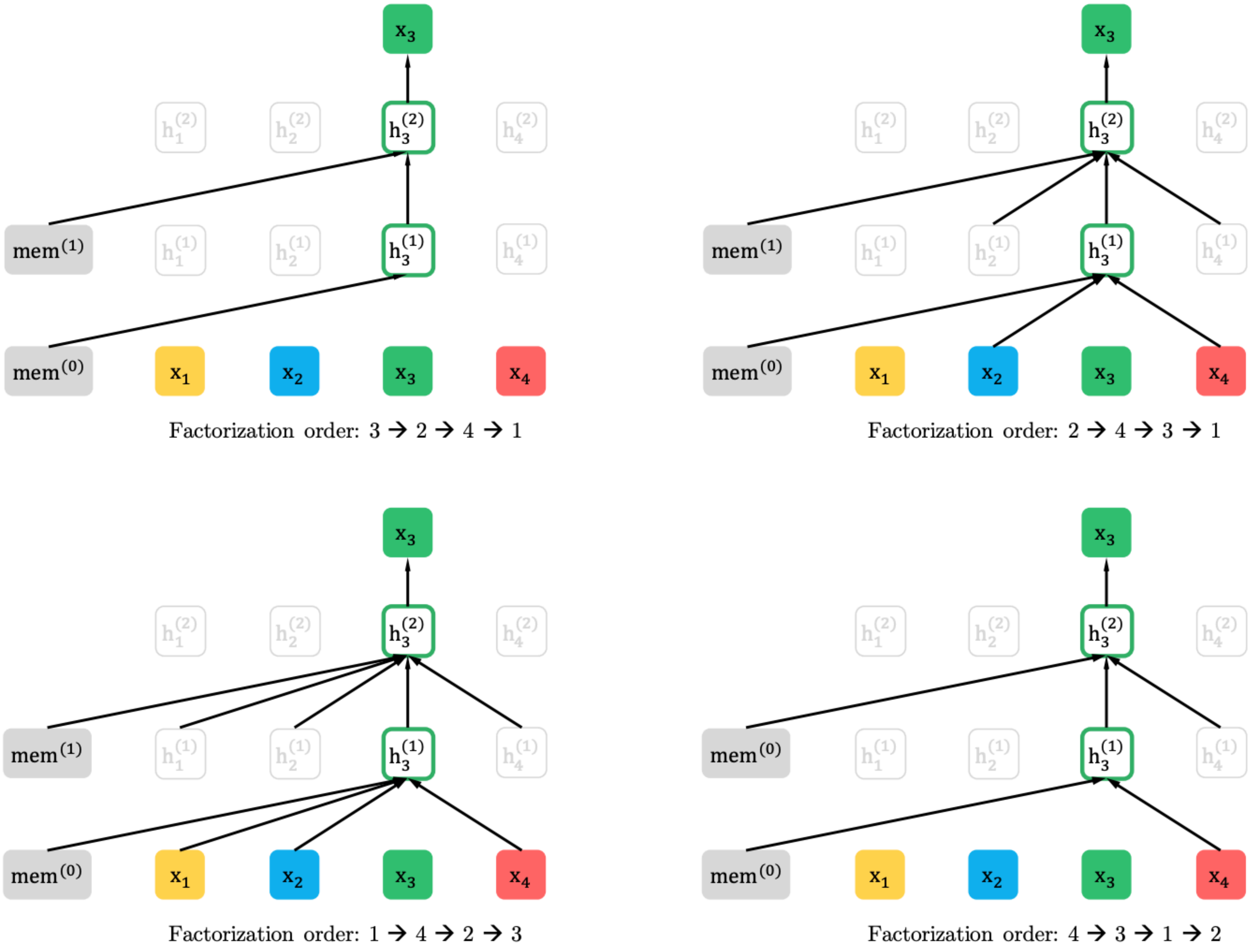
$$[x_1\ x_2\ x_3\ x_4\ x_5\ x_6\ x_7]$$

BiRNN (AutoRegressive) - > AutoEncoding Model

$$[x_1\ x_2\ x_3\ \mathrm{M}\ \mathrm{x}_5\ \mathrm{x}_6\ \mathrm{x}_7]$$

Permutative Language Modeling

XLNet : Generalized Autoregressive Pretraining for Language Understanding



Sequence Order

[x1 x2 x3 x4 x5 | x6 x7]

Factorization Order

[x2 x3 x7 x1 x4 | x6 x5]

- 7! Options
- Bidirectional included

Two-Stream Attention

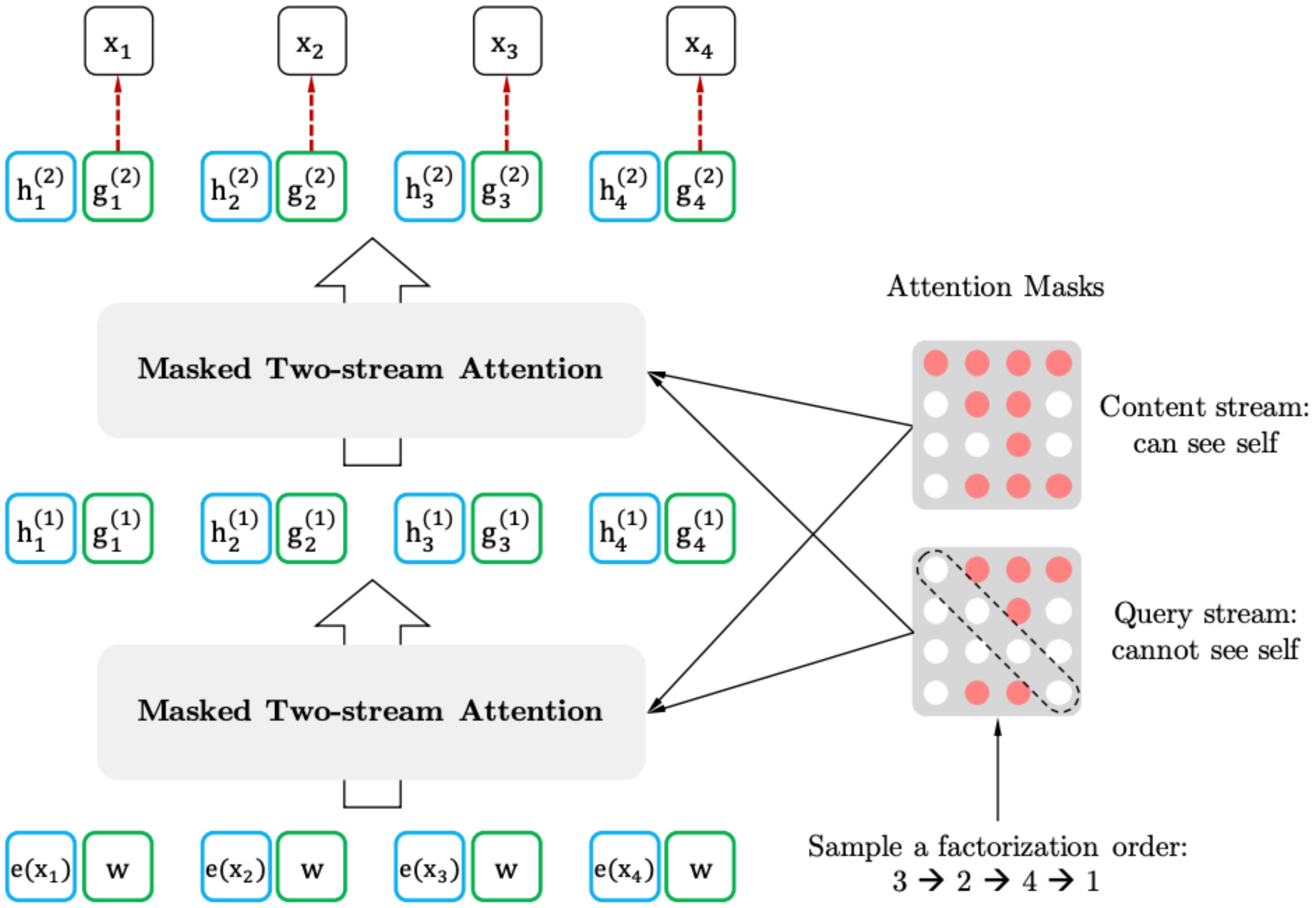
XLNet : Generalized Autoregressive Pretraining for Language Understanding

$$g_{z_t}^{(m)} \leftarrow \mathrm{Attention}(\mathrm{Q} = \mathrm{g}_{\mathrm{z_t}}^{(\mathrm{m}-1)}, \mathrm{KV} = \mathrm{h}_{\mathrm{Z}<\mathrm{t}}^{(\mathrm{m}-1)}; \theta)$$

$$h_{z_t}^{(m)} \leftarrow \mathrm{Attention}(\mathrm{Q} = \mathrm{h}_{\mathrm{z_t}}^{(\mathrm{m}-1)}, \mathrm{KV} = \mathrm{h}_{\mathrm{Z}\leq\mathrm{t}}^{(\mathrm{m}-1)}; \theta)$$

Two-Stream Attention

XLNet : Generalized Autoregressive Pretraining for Language Understanding



Modeling Multiple Segments

XLNet : Generalized Autoregressive Pretraining for Language Understanding

- Relative Segment Embedding

  "we randomly sample two segments (either from the same context or not) and treat the concatenation of two segments as one sequence to perform permutation language modeling. We only reuse the memory that belongs to the same context."

Performance (1)

XLNet : Generalized Autoregressive Pretraining for Language Understanding

3.2 Fair Comparison with BERT

Model	SQuAD1.1	SQuAD2.0	RACE	MNLI	QNLI	QQP	RTE	SST-2	MRPC	CoLA	STS-B
BERT-Large (Best of 3)	86.7/92.8	82.8/85.5	75.1	87.3	93.0	91.4	74.0	94.0	88.7	63.7	90.2
XLNet-Large-wikibooks	88.2/94.0	85.1/87.8	77.4	88.4	93.9	91.8	81.2	94.4	90.0	65.2	91.1

Table 1: Fair comparison with BERT. All models are trained using the same data and hyperparameters as in BERT. We use the best of 3 BERT variants for comparison; i.e., the original BERT, BERT with whole word masking, and BERT without next sentence prediction.

Performance (2)

XLNet : Generalized Autoregressive Pretraining for Language Understanding

3.3 Comparison with RoBERTa: Scaling Up

RACE	Accuracy	Middle	High	Model	NDCG@20	ERR@20
GPT [28]	59.0	62.9	57.4	DRMM [13]	24.3	13.8
BERT [25]	72.0	76.6	70.1	KNRM [8]	26.9	14.9
BERT+DCMN* [38]	74.1	79.5	71.8	Conv [8]	28.7	18.1
RoBERTa [21]	83.2	86.5	81.8	BERT†	30.53	18.67
XLNet	**85.4**	**88.6**	**84.0**	XLNet	**31.10**	**20.28**

Table 2: Comparison with state-of-the-art results on the test set of RACE, a reading comprehension task, and on ClueWeb09-B, a document ranking task. ∗ indicates using ensembles. † indicates our implementations. "Middle" and "High" in RACE are two subsets representing middle and high school difficulty levels. All BERT, RoBERTa, and XLNet results are obtained with a 24-layer architecture with similar model sizes (aka BERT-Large).

Performance (2)

XLNet : Generalized Autoregressive Pretraining for Language Understanding

3.3 Comparison with RoBERTa: Scaling Up

RACE	Accuracy	Middle	High	Model	NDCG@20	ERR@20
GPT [28]	59.0	62.9	57.4	DRMM [13]	24.3	13.8
BERT [25]	72.0	76.6	70.1	KNRM [8]	26.9	14.9
BERT+DCMN* [38]	74.1	79.5	71.8	Conv [8]	28.7	18.1
RoBERTa [21]	83.2	86.5	81.8	BERT†	30.53	18.67
XLNet	**85.4**	**88.6**	**84.0**	XLNet	**31.10**	**20.28**

Table 2: Comparison with state-of-the-art results on the test set of RACE, a reading comprehension task, and on ClueWeb09-B, a document ranking task. $*$ indicates using ensembles. † indicates our implementations. "Middle" and "High" in RACE are two subsets representing middle and high school difficulty levels. All BERT, RoBERTa, and XLNet results are obtained with a 24-layer architecture with similar model sizes (aka BERT-Large).

Limitations of XLNet

XLNet : Generalized Autoregressive Pretraining for Language Understanding

Optimization Difficulty

"Specifically, we train on 512 TPU v3 chips for 500K steps with an Adam weight decay optimizer, linear learning rate decay, and a batch size of 8192, which takes about 5.5 days. It was observed that the model still underfits the data at the end of training."