Applied Deep Learning

More on Embeddings and Transformer

Apr 9th, 2019

SHANG-YU SU

國立臺灣大學
National Taiwan University

HTTP://ADL.MIULAB.TW


ELMO
ERNIE
BERT

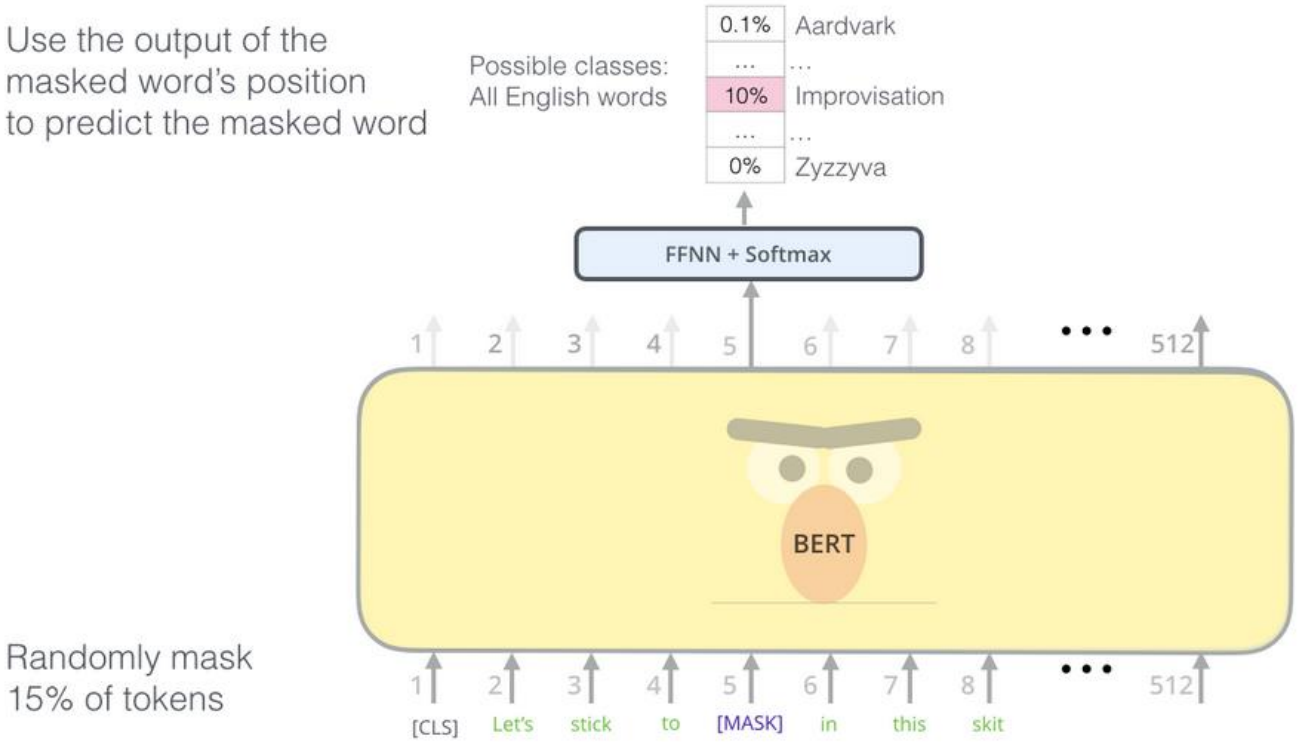
Ernie and Bert

ERNIE: Enhanced Representation through kNowledge IntEgration

- Developed by Baidu Research
- **No paper published yet, only an aritcle on their website (3/16)**
- claim that it outperforms BERT in Chinese language tasks including natural language inference, semantic similarity, named entity recognition, sentiment analysis, and question-answer matching.
- Methods like Cove, ELMo, GPT or BERT mainly focus on building models to solve problems based on original language signals instead of semantic units in the text.
- Unlike BERT, ERNIE features knowledge integration enhancement, which learns semantic relations in the real world through massive data.

ERNIE: Enhanced Representation through kNowledge IntEgration

- In BERT, we randomly mask 15% **tokens** to train masked LM

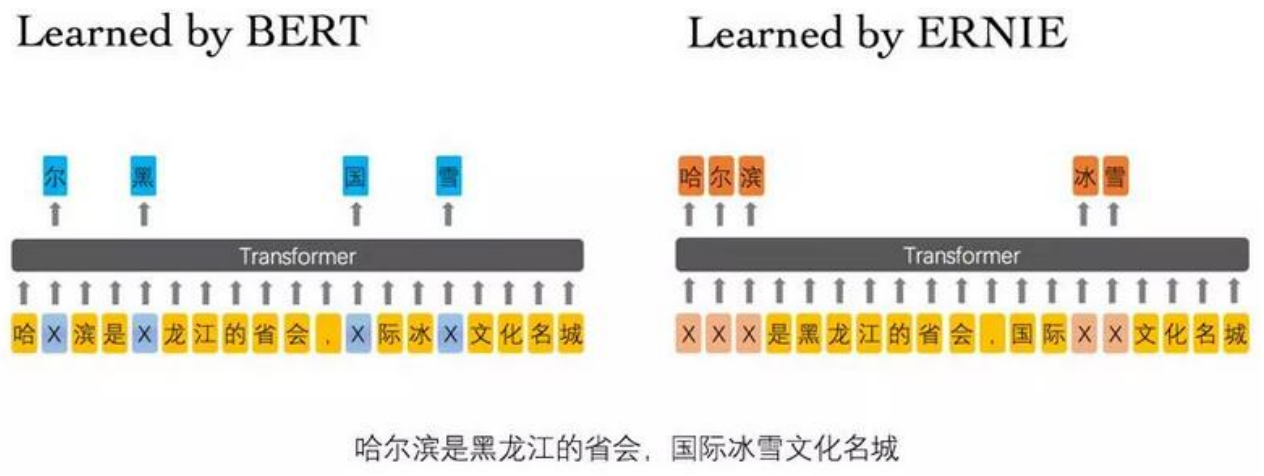


ERNIE: Enhanced Representation through kNowledge IntEgration

- It directly models prior semantic knowledge units, which enhances the ability to learn semantic representation.
- ERNIE learns the semantic representation of complete concepts by masking semantic units such as words and entities. ERNIE directly models priori semantic knowledge units and, as a result, enhances the model's ability to learn semantic representation.
- Entities: in information extraction, a named entity is a real-world object, such as persons, locations, organizations, products, etc.

ERNIE: Enhanced Representation through kNowledge IntEgration

- BERT can identify the character “尔(er)” through the local co-occurring characters 哈(ha) and 滨(bin), but fails to learn any knowledge related to the word “Harbin (哈尔滨)”.
- ERNIE can extrapolate the relationship between Harbin (哈尔滨) and Heilongjiang (黑龙江) by analyzing implicit knowledge of words and entities.



Closer Look…

- Maybe it resembles the leading model “BERT+N-Gram Masking” on SQuAD2.0?

3 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	86.673	89.147

- Granularity: masking short **sentences**?
- Keep the Transformer structure untouched and change masking?
- Dataset are different.
- “For every plus there is a minus.”, more factors: granularity, quality of segmentation systems…, and etc.

By The Way…

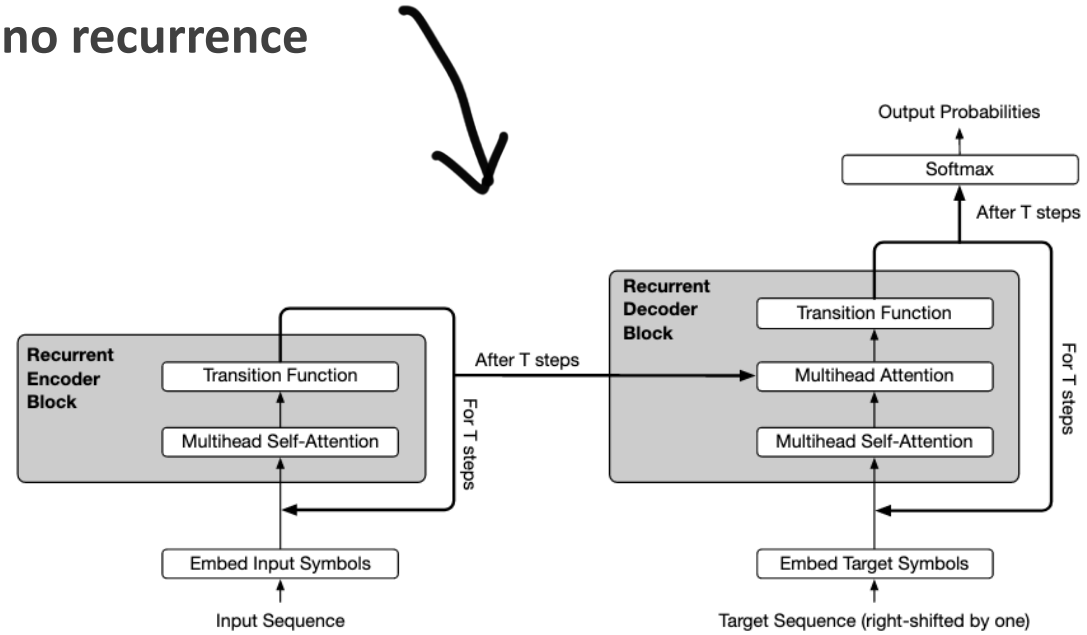
Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	**87.147**	**89.474**
2 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286
3 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	86.673	89.147
4 Mar 16, 2019	BERT + DAE + AoA (single model) *Joint Laboratory of HIT and iFLYTEK Research*	85.884	88.621
5 Jan 15, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	85.082	87.615
5 Mar 13, 2019	BERT + ConvLSTM + MTL + Verifier (single model) *Layer 6 AI*	84.924	88.204
5 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (single model) *Google AI Language* https://github.com/google-research/bert	85.150	87.715
6 Jan 10, 2019	BERT + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	84.292	86.967
7 Dec 13, 2018	BERT finetune baseline (ensemble) *Anonymous*	83.536	86.096
7 Mar 20, 2019	Bert-raw (ensemble) *None*	83.604	86.036
7 Dec 21, 2018	PAML+BERT (ensemble model) *PINGAN GammaLab*	83.457	86.122
7 Dec 16, 2018	Lunet + Verifier + BERT (ensemble) *Layer 6 AI NLP Team*	83.469	86.043
8 Mar 04, 2019	SemBERT (ensemble model) *Shanghai Jiao Tong University*	83.243	85.821
8 Dec 15, 2018	Lunet + Verifier + BERT (single model) *Layer 6 AI NLP Team*	82.995	86.035
8 Jan 14, 2019	BERT + MMFT + ADA (single model) *Microsoft Research Asia*	83.040	85.892
9 Feb 15, 2019	BERT + NeurQuRI (ensemble) *2SAH*	82.803	85.703
9 Feb 16, 2019	Bert-raw (ensemble) *None*	83.175	85.635
10	PAML+BERT (single model)	82.577	85.603

ERNIE: Enhanced Representation through kNowledge IntEgration

- Pretrained models/code available: https://github.com/PaddlePaddle/LARK/tree/develop/ERNIE

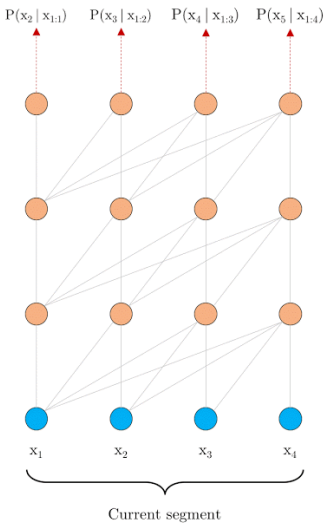
Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context

- from Google
- The third generation (Transformer-XL): **recurrence in length**
- The second generation (Universal Transformer): **recurrence in depth**
- The original Transformer: **no recurrence**



Transformer for LM

- in language modeling, Transformers are currently implemented with a fixed-length context, i.e. a long text sequence is truncated into fixed-length segments of a few hundred characters, and each segment is processed separately.



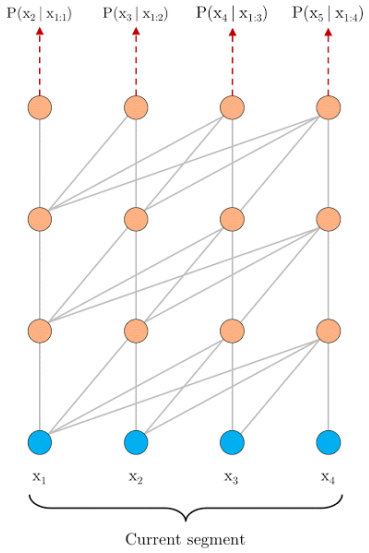
Transformer for LM

- This introduces two critical limitations:
- The algorithm is not able to model dependencies that are longer than a fixed length.
- The segments usually do not respect the sentence boundaries, resulting in **context fragmentation** which leads to inefficient optimization.

~~它不僅是一個能夠處理可~~變長度序列的模型，在多個任~~務中刷新了當前的最好性能。~~

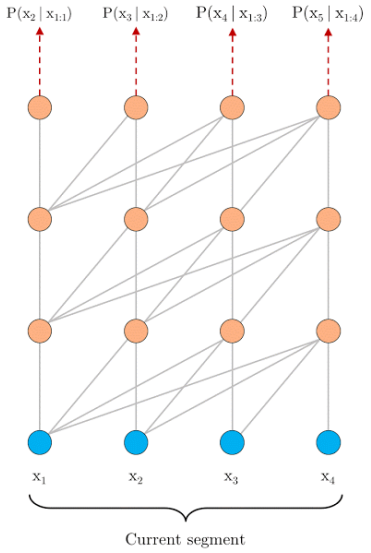
Transformer-XL: Segment-level Recurrence

- During training, the representations computed for the previous segment are fixed and cached to be reused as an extended context when the model processes the next new segment.



Transformer-XL: Segment-level Recurrence

- Moreover, this recurrence mechanism also resolves the context fragmentation issue, providing necessary context for tokens in the front of a new segment.

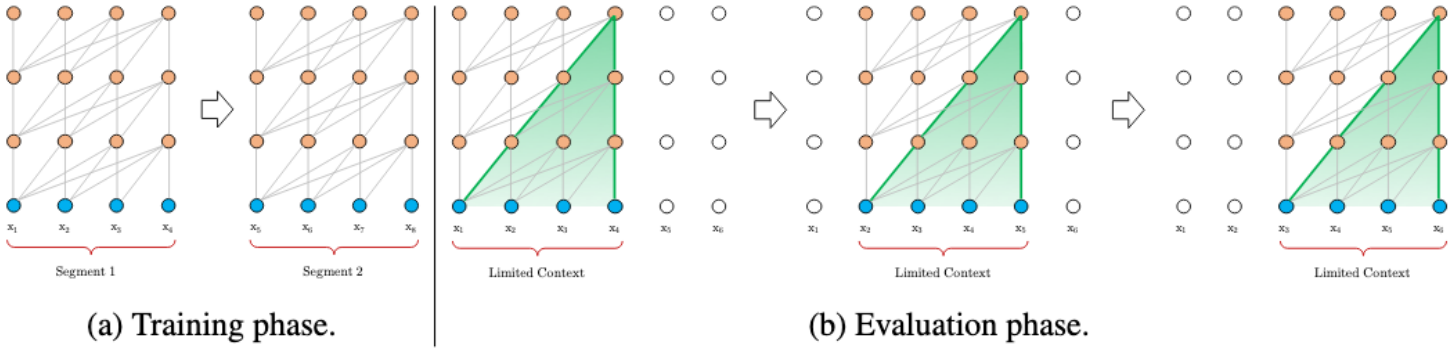


Transformer-XL: Relative Positional Encodings

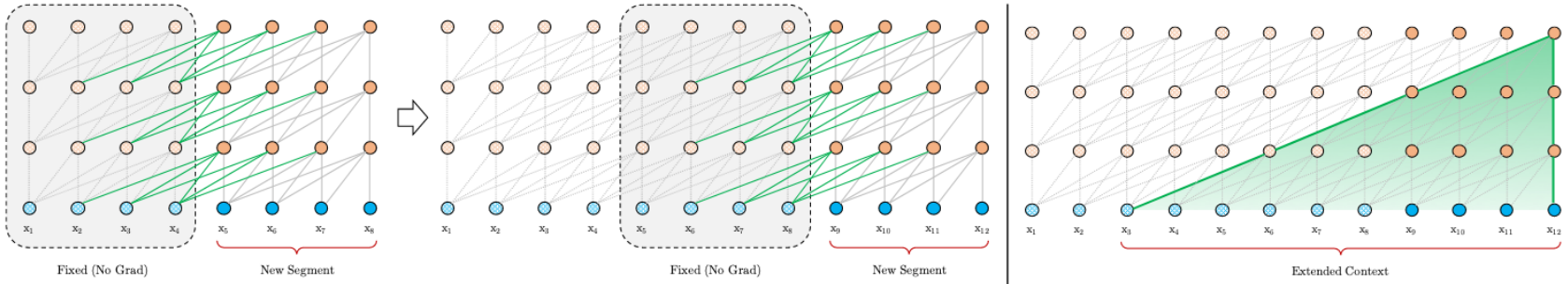
- Naively applying segment-level recurrence does not work, however, because the positional encodings are not coherent when we reuse the previous segments.
- For example, consider an old segment with contextual positions [0, 1, 2, 3]. When a new segment is processed, we have positions [0, 1, 2, 3, 0, 1, 2, 3] for the two segments combined, where the semantics of each position id is incoherent through out the sequence.
- **parameterization to only encode the relative positional information based on content**

Transformer-XL: Overview

- segment-level recurrence + relative positional encoding



(a) Training phase. (b) Evaluation phase.



(a) Training phase. (b) Evaluation phase.

Transformer-XL: Results

- Transformer-XL learns dependency that is about 80% longer than RNNs and 450% longer than vanilla Transformers, which generally have better performance than RNNs, but are not the best for long-range dependency modeling due to fixed-length contexts.
- Transformer-XL is up to **1,800+ times faster** than a vanilla Transformer during evaluation on language modeling tasks, because no re-computation is needed.
- Transformer-XL has better performance in perplexity (more accurate at predicting a sample) on long sequences because of long-term dependency modeling, and also on short sequences by resolving the context fragmentation problem.

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

- code available (**TF and PyTroch**): https://github.com/kimiyoung/transformer-xl

References

- http://research.baidu.com/Blog/index-view?id=113
- http://fortune.com/2016/05/06/sesame-street-bert-ernie-std/
- https://www.usatoday.com/story/life/tv/2018/09/18/sesame-street-denies-writers-claim-bert-and-ernie-gay/1348017002/
- https://en.wikipedia.org/wiki/Named_entity
- http://jalammar.github.io/illustrated-bert/
- https://arxiv.org/pdf/1901.02860.pdf