MASS: Masked Sequence to Sequence Pre-Training for Language Generation

01 Abstract

02 MASS

03 Experiments & Results

04 Conclusion

05 Q & A

NLP Bootcamp
| 박인호 |

Motivation

- BERT와 GPT 등의 성과 이후, **Pre-training**은 자연어 처리 분야에서 활발히 연구
- BERT와 XLNet 등은 특히 Natural language understanding task에서 뛰어난 성능

e.g.) Sentiment Classification, Natural Language Inference, Machine Reading Comprehension, etc.

- 그러나 NLU Task 외에도 Sequence-to-Sequence를 기반으로 하는 language generation tasks

e.g.) Machine Translation, Abstract Text Summarization, Conversational, Response Generation, Question Answering, Text Style Transfer, etc.

- Language generation tasks에는 Encoder-Attention-Decoder 구조의 모델들이 유용

NLP Downstream Tasks

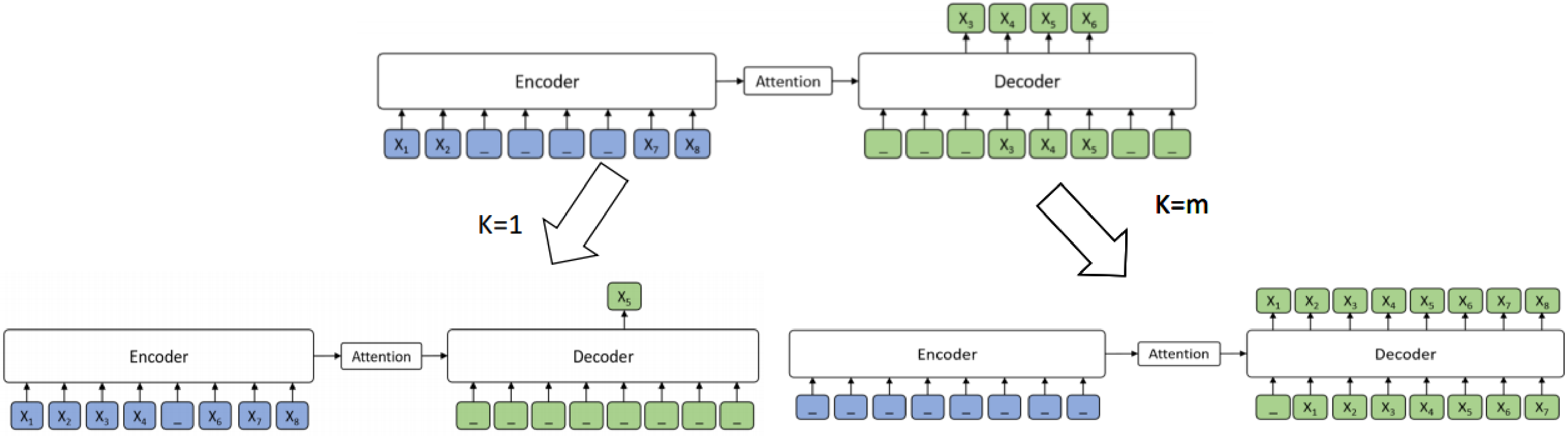
Language Understanding

- Sentiment classification
- Natural language Inference
- Named entity recognition
- SQuAD question answering

Language Generation

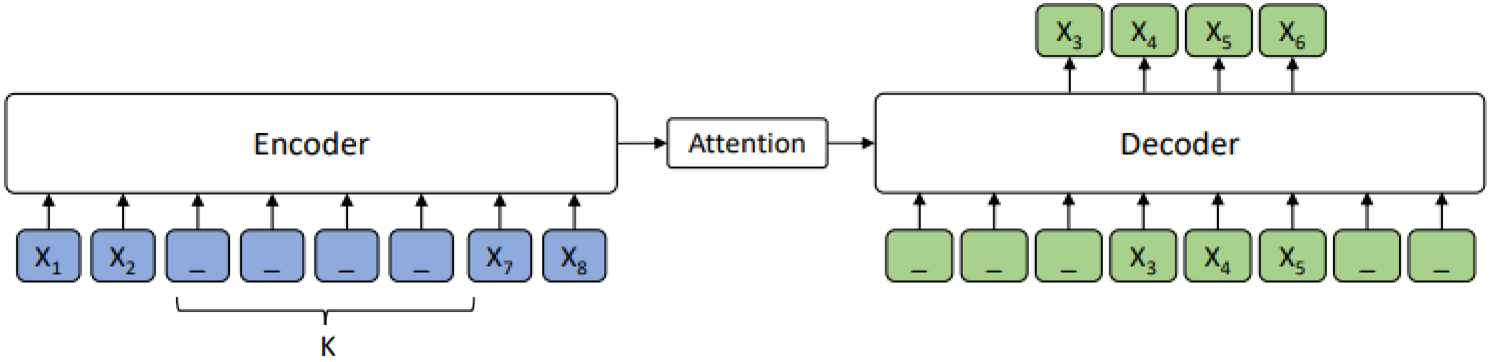
- Neural machine translation
- Text summarization
- Conversational response generation

MASS vs. BERT, GPT



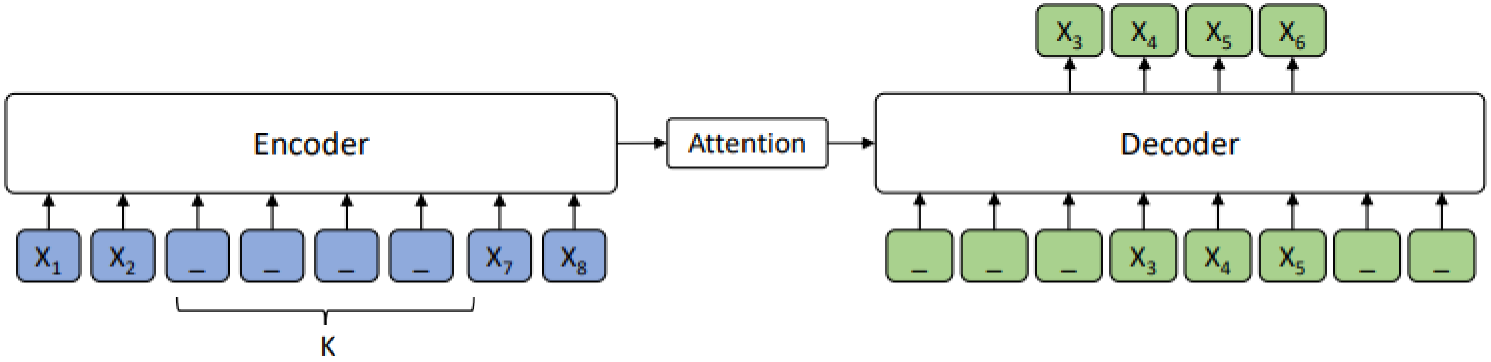
- BERT(좌측)와 XLNet은 Language Understanding을 위해 **인코더를 Pre-train**하는 반면, GPT(우측)는 Language Modeling을 위해 **디코더를 Pre-train** 시키는 구조를 지니고 있다.
- 즉, 이전 모델들의 환경에서는 인코더, Attention 메커니즘 그리고 **디코더가 함께 훈련될 수 없다.**
- **Attention 메커니즘은** Language generation에서 **매우 중요**하므로 이를 따로 학습시키지 않는 BERT나 XLNet, GPT는 Language generation Task에 있어 최적의 성능을 발휘할 수가 없다.

MASS Framework



- MASS는 MAsked Sequence to Sequence Pre-training의 약자로, Input sequence 에서 k 개의 **연속된 토큰**을 임의로 지정해 마스킹 한 후, 마스킹 된 토큰들을 Encoder-attention-Decoder Framework를 거쳐 디코더에서 예측하도록 훈련시키는 Pre-training 기법이다.
- 인코더는 3, 4, 5, 6번째 토큰이 마스킹 된 Input sequence의 hidden representations을 생성하고, 디코더는 인코더에서 마스킹 된 토큰들을 예측합니다. 이때, 인코더에서 마스킹 되지 않았던 다른 토큰들이 마스킹 된다.

MASS Pre-Training

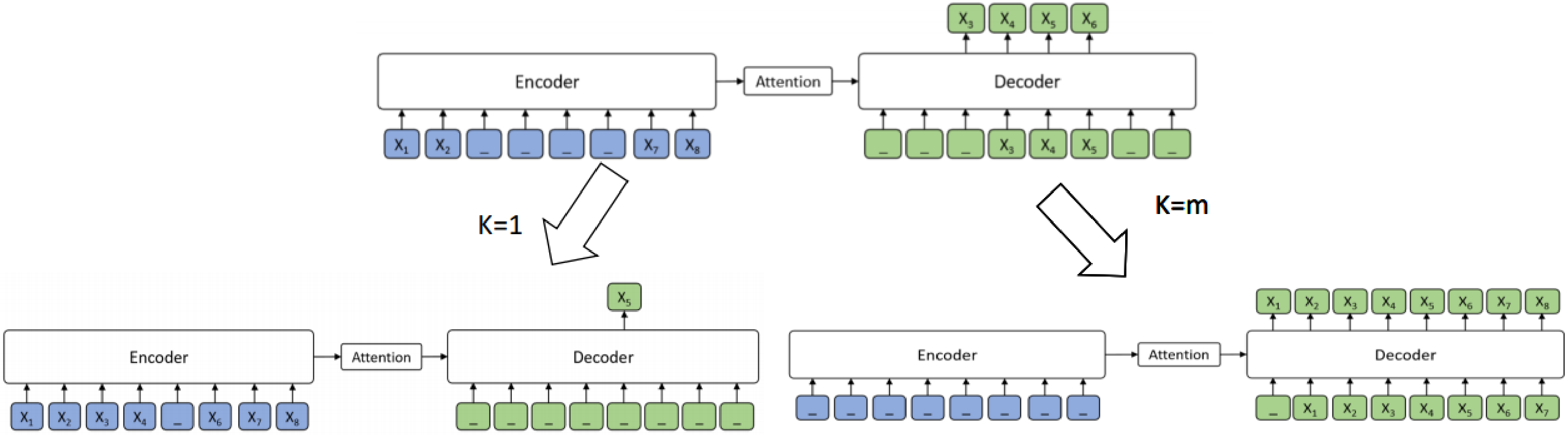


- Mask k consecutive tokens (segment)
  - Force the decoder to attend on the source representations, i.e., encoder-decoder attention
  - Force the encoder to extract meaningful information from the sentence
  - Develop the decoder with the ability of language modeling

Advantages of MASS

1. 인코더에서 마스킹 되지 않은 토큰들이 디코더에서 마스킹 되어, 디코더는 인코더가 제공한 **hidden representation**과 **Attention 정보**만을 참고해 마스킹 된 토큰들을 예측해야 하고, 이는 **Encoder-attention-Decoder**가 **함께 Pre-train** 될 수 있는 환경을 제공하게 된다.
2. 인코더는 디코더에 보다 유용한 정보를 제공하기 위해 인코더에서 **마스킹되지 않은** 토큰들의 정보를 잘 추출할 수 있도록 학습하고, 이를 통해 **Language Understanding 능력이 개선**되게 된다.
3. 디코더는 인코더에서 마스킹 된 토큰들에 대한 예측을 **연속적으로** 수행해야 하므로 Language Modeling 능력을 학습하게 된다.

MASS vs. BERT/ GPT



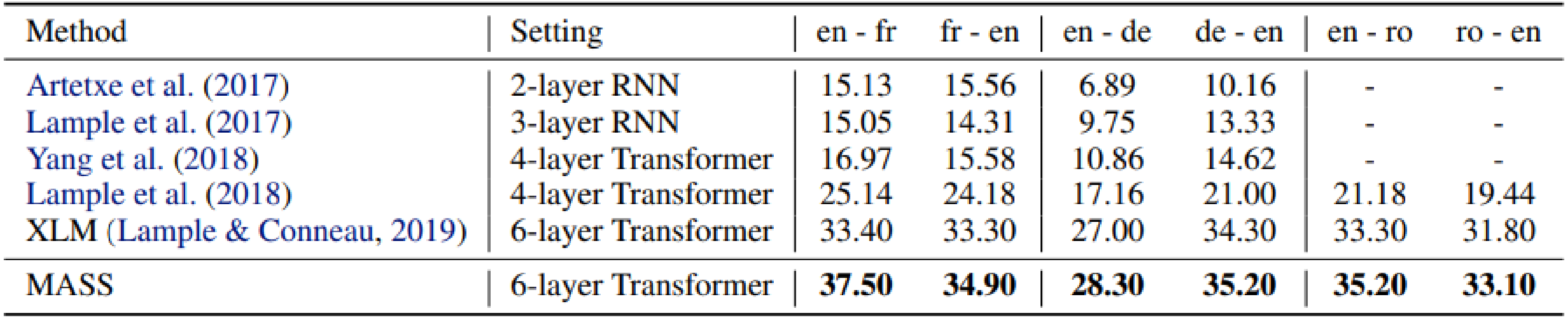
Length	Probability	Model
$k = 1$	$P(x^{u} \mid x^{\backslash u}; \theta)$	masked LM in BERT
$k \in [1, m]$	$P(x^{u:v} \mid x^{\backslash u:v}; \theta)$	MASS

Length	Probability	Model
$k = m$	$P(x^{1:m} \mid x^{\backslash 1:m}; \theta)$	standard LM in GPT
$k \in [1, m]$	$P(x^{u:v} \mid x^{\backslash u:v}; \theta)$	MASS

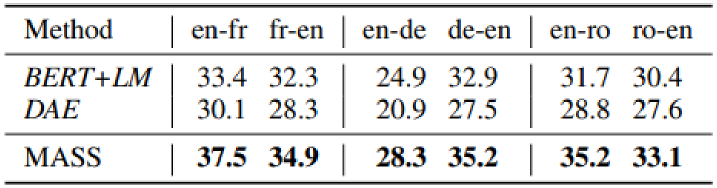
Q & A

Fine-Tuning on downstream tasks

Unsupervised NMT



Method	Setting	en - fr	fr - en	en - de	de - en	en - ro	ro - en
Artetxe et al. (2017)	2-layer RNN	15.13	15.56	6.89	10.16	-	-
Lample et al. (2017)	3-layer RNN	15.05	14.31	9.75	13.33	-	-
Yang et al. (2018)	4-layer Transformer	16.97	15.58	10.86	14.62	-	-
Lample et al. (2018)	4-layer Transformer	25.14	24.18	17.16	21.00	21.18	19.44
XLM (Lample & Conneau, 2019)	6-layer Transformer	33.40	33.30	27.00	34.30	33.30	31.80
MASS	6-layer Transformer	**37.50**	**34.90**	**28.30**	**35.20**	**35.20**	**33.10**



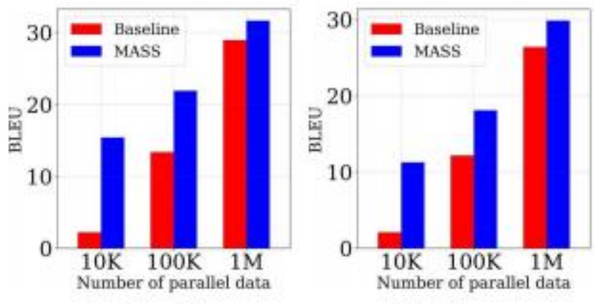
Method	en-fr	fr-en	en-de	de-en	en-ro	ro-en
BERT+LM	33.4	32.3	24.9	32.9	31.7	30.4
DAE	30.1	28.3	20.9	27.5	28.8	27.6
MASS	**37.5**	**34.9**	**28.3**	**35.2**	**35.2**	**33.1**

Table 3. The BLEU score comparisons between MASS and other pre-training methods. The results for BERT+LM are directly taken from the MLM+CLM setting in XLM (Lample & Conneau, 2019) as they use the same pre-training methods.

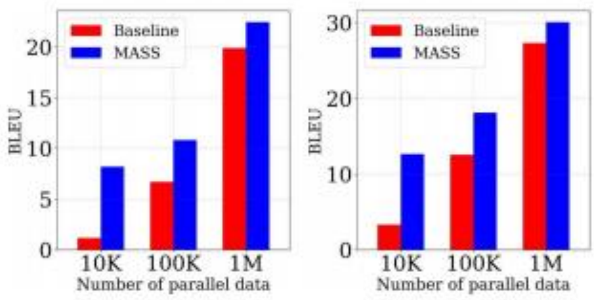
- MASS는 Pre-training을 위해 monolingual 데이터를 필요로 한다.
- Pre-training 이후에는 MASS의 성능을 입증하기 위해 Fine-tuning을 수행

Fine-Tuning on downstream tasks

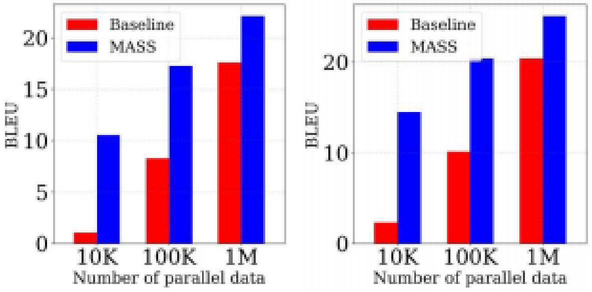
Low-resource NMT



(a) en-fr (b) fr-en



(e) en-ro (f) ro-en



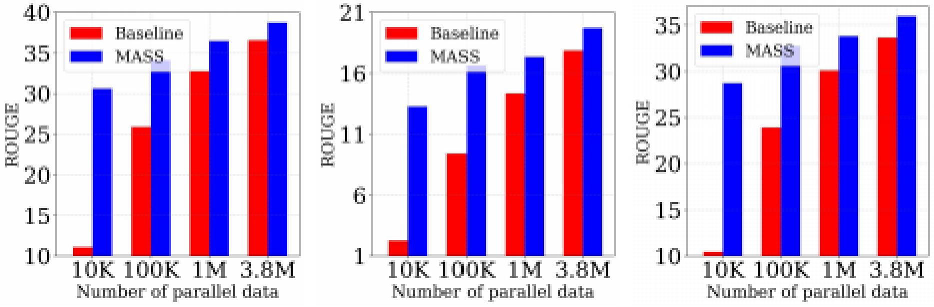
(c) en-de (d) de-en

- **Low-resource Machine Translation**: bilingual 트레이닝 데이터셋이 부족한 환경에서의 기계번역
- [WMT14 English-French], [WMT16 English-German, English-Romanian] 데이터셋을 10K, 100K, 1M의 사이즈로 늘려가며 Low-resource 환경에서의 기계번역 시나리오 테스트
- 그림에서 확인할 수 있듯 MASS는 모든 스케일에서 Low-resource MT의 baseline 성능을 능가했으며, 특히 이러한 성능의 개선은 **데이터셋의 사이즈가 작을수록** 더 두드러지게 나타남 (*see result of 10K*)

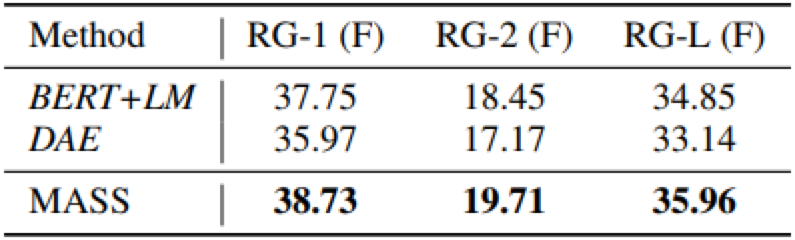
Figure 3. The BLEU score comparisons between MASS and the baseline on low-resource NMT with different scales of paired data.

Fine-Tuning on downstream tasks

Text Summarization



(a) RG-1 (F) (b) RG-2 (F) (c) RG-L (F)

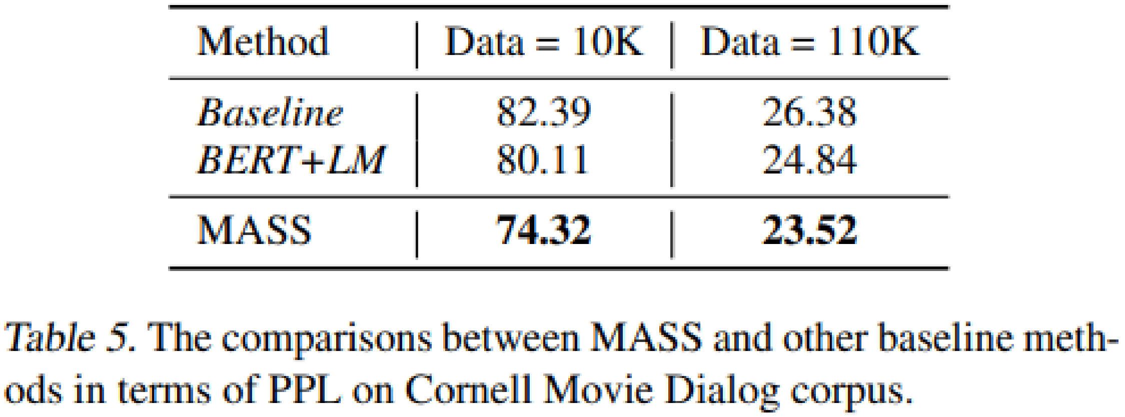


Method	RG-1 (F)	RG-2 (F)	RG-L (F)
BERT+LM	37.75	18.45	34.85
DAE	35.97	17.17	33.14
MASS	**38.73**	**19.71**	**35.96**

- Pre-trained BERT를 인코더로 Pre-trained Language Model을 디코더로 사용한 **BERT+LM** 모델과 **DAE**(Denoising Auto-Encoder), **MASS**의 Abstractive Summarization 성능을 Gigaword Corpus에 대해 비교
- 표에서 확인할 수 있듯, MASS는 BERT+LM과 DAE 두 모델의 Abstractive Summarization 성능을 모두 능가

Fine-Tuning on downstream tasks

Conversational Response Generation

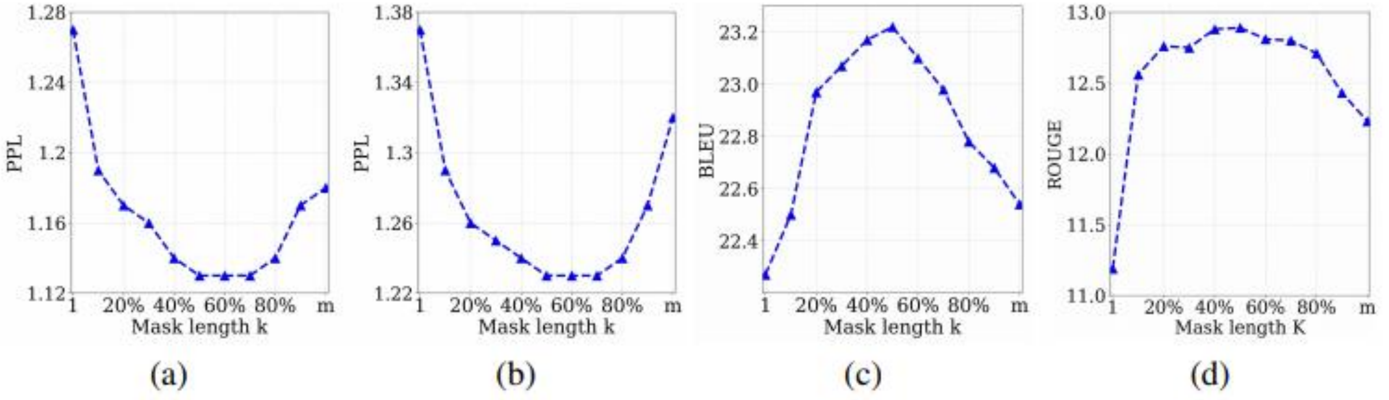


Method	Data = 10K	Data = 110K
Baseline	82.39	26.38
BERT+LM	80.11	24.84
MASS	**74.32**	**23.52**

Table 5. The comparisons between MASS and other baseline methods in terms of PPL on Cornell Movie Dialog corpus.

- Abstractive Summarization 성능 비교에 사용되었던 BERT+LM 모델과 Baseline, 그리고 MASS의 Conversational Response Generation 성능을 Cornell Movie Dialog Corpus에 대해 비교
- 마찬가지로 MASS가 BERT+LM 모델과 Baseline 보다 **낮은 Perplexity** (PPL)를 기록하며 좋은 성능을 보여줌

Analysis of MASS

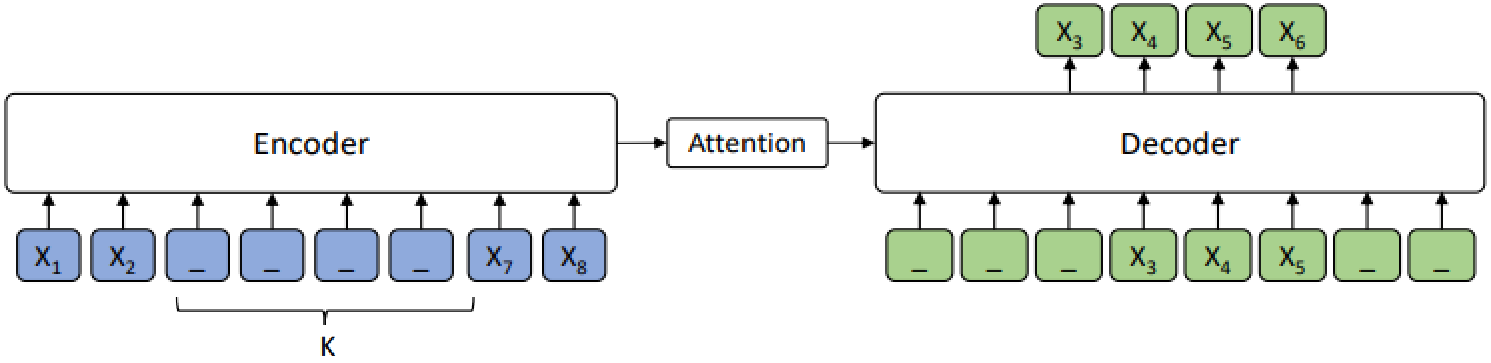


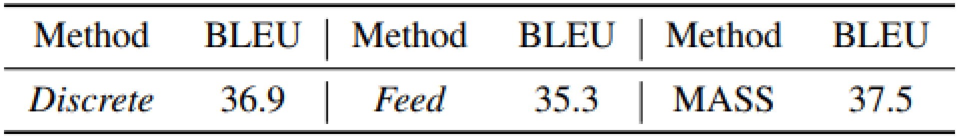
(a), (b): PPL of the pre-trained model on En and Fr
(c): BLEU score of unsupervised En-Fr
(d): ROUGE of text summarization

- 위 그림들은 하이퍼 파라미터 k를 다르게 설정해가며, MASS 성능에 대한 다양한 실험을 수행한 결과
- 경험적 실험을 통해 k가 "**문장의 절반 정도**"의 크기에 해당할 때, downstream task에서 가장 좋은 성능을 보임을 알게 됨
  - 문장 내 절반의 토큰을 마스킹하는 것이 인코더와 디코더의 Pre-training에 있어 "**적절한 균형**"을 제공해주기 때문
- 그러나 k가 **1** (like. BERT) 혹은 **m** (like. GPT) 이었을 때는 downstream task에서 좋은 성능이 나오지 않았는데,

이는 MASS가 Sequence to Sequence 기반의 Language Generation Task에 이점을 지니고 있음을 반증 !

Analysis of MASS





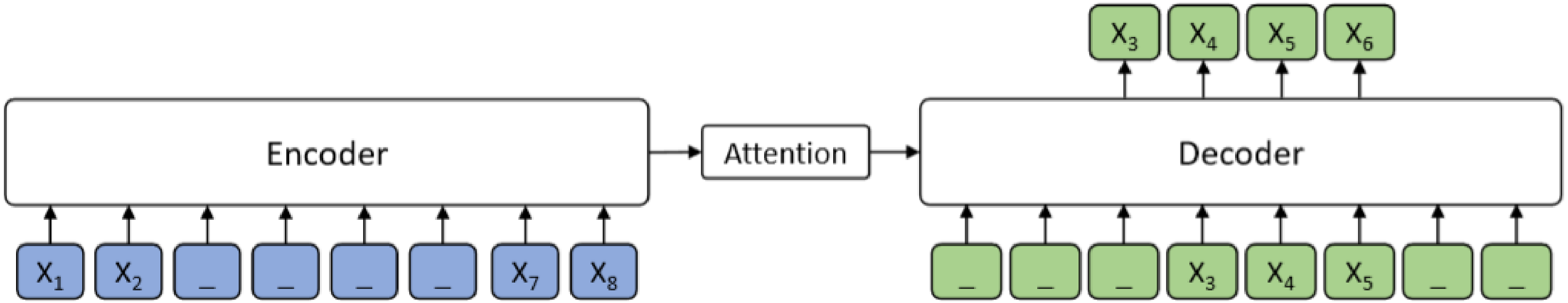
Method	BLEU	Method	BLEU	Method	BLEU
Discrete	36.9	*Feed*	35.3	MASS	37.5

Table 6. The comparison between MASS and the ablation methods in terms of BLEU score on the unsupervised en-fr translation.

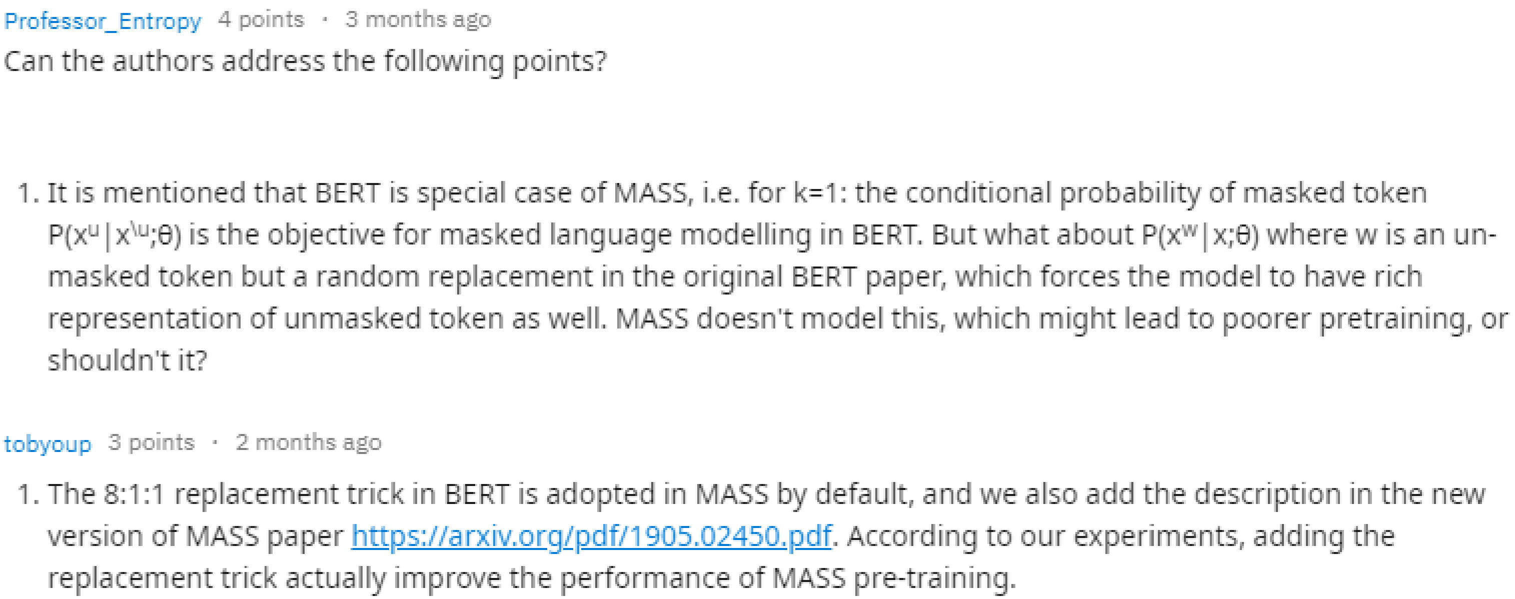
- Discrete: instead of masking continuous segment, masking discrete tokens
- Feed: Feed the tokens to the decoder that appear in the encoder

Summary

- MASS jointly pre-trains the encoder-attention-decoder framework for sequence to sequence based language generation tasks
- MASS achieves significant improvements over the baselines without pre-training or with other pre-training methods on zero/low-resource NMT, text summarization and conversational response generation.



Q & A



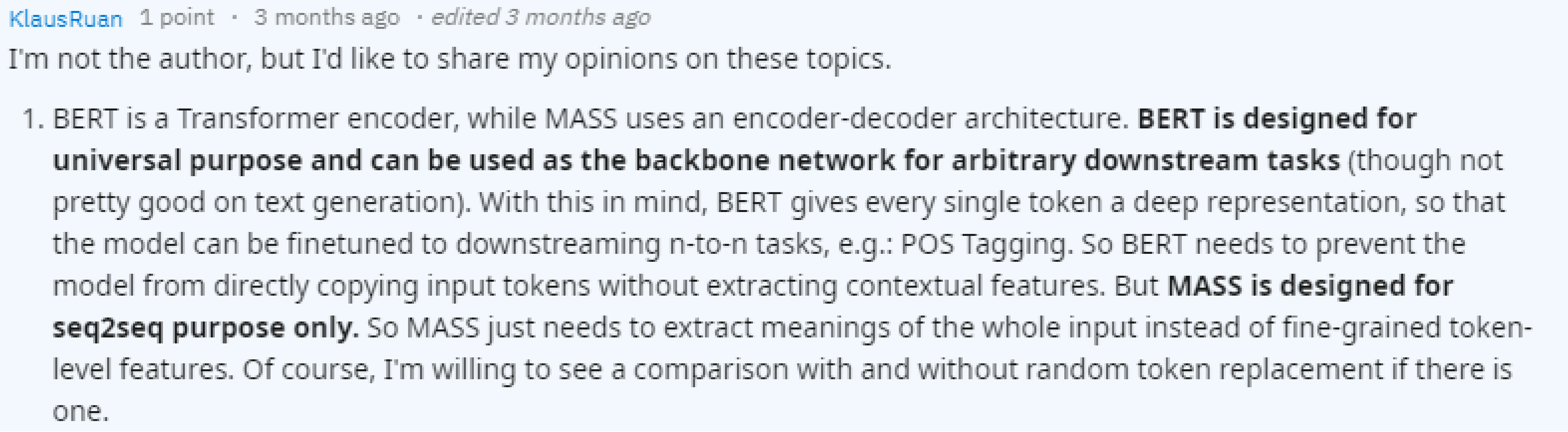
Professor_Entropy 4 points · 3 months ago

Can the authors address the following points?

1. It is mentioned that BERT is special case of MASS, i.e. for k=1: the conditional probability of masked token $P(x^u|x^{\setminus u};\theta)$ is the objective for masked language modelling in BERT. But what about $P(x^w|x;\theta)$ where w is an un-masked token but a random replacement in the original BERT paper, which forces the model to have rich representation of unmasked token as well. MASS doesn't model this, which might lead to poorer pretraining, or shouldn't it?

tobyoup 3 points · 2 months ago

1. The 8:1:1 replacement trick in BERT is adopted in MASS by default, and we also add the description in the new version of MASS paper https://arxiv.org/pdf/1905.02450.pdf. According to our experiments, adding the replacement trick actually improve the performance of MASS pre-training.

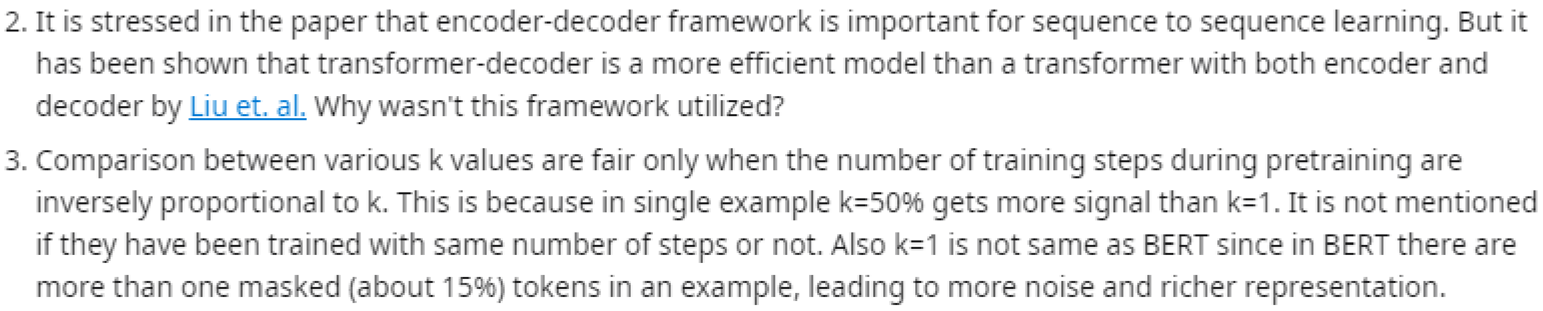


KlausRuan 1 point · 3 months ago · *edited 3 months ago*

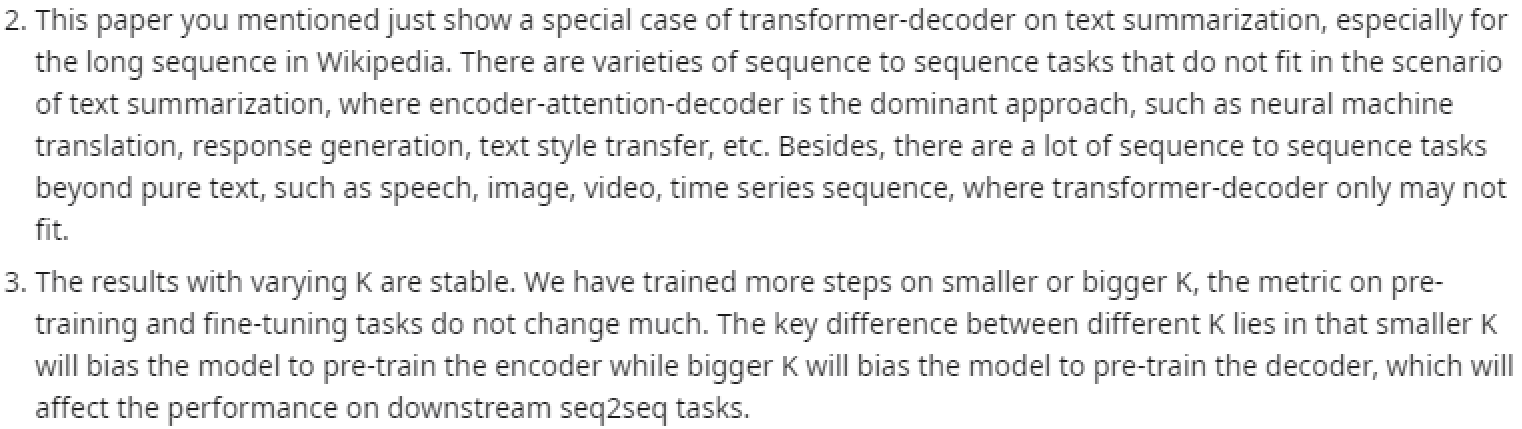
I'm not the author, but I'd like to share my opinions on these topics.

1. BERT is a Transformer encoder, while MASS uses an encoder-decoder architecture. **BERT is designed for universal purpose and can be used as the backbone network for arbitrary downstream tasks** (though not pretty good on text generation). With this in mind, BERT gives every single token a deep representation, so that the model can be finetuned to downstreaming n-to-n tasks, e.g.: POS Tagging. So BERT needs to prevent the model from directly copying input tokens without extracting contextual features. But **MASS is designed for seq2seq purpose only.** So MASS just needs to extract meanings of the whole input instead of fine-grained token-level features. Of course, I'm willing to see a comparison with and without random token replacement if there is one.

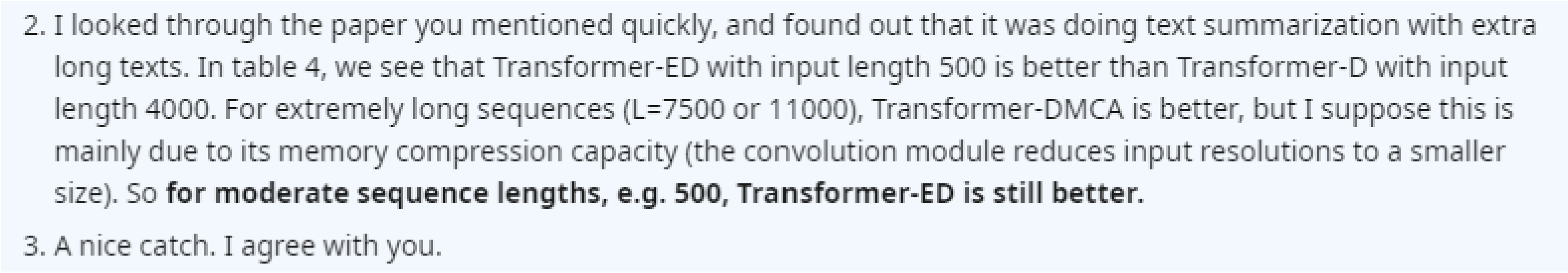
(출처: https://www.reddit.com/r/MachineLearning/comments/bn2da0/r_190502450_mass_masked_sequence_to_sequence/)



2. It is stressed in the paper that encoder-decoder framework is important for sequence to sequence learning. But it has been shown that transformer-decoder is a more efficient model than a transformer with both encoder and decoder by Liu et. al. Why wasn't this framework utilized?
3. Comparison between various k values are fair only when the number of training steps during pretraining are inversely proportional to k. This is because in single example k=50% gets more signal than k=1. It is not mentioned if they have been trained with same number of steps or not. Also k=1 is not same as BERT since in BERT there are more than one masked (about 15%) tokens in an example, leading to more noise and richer representation.



2. This paper you mentioned just show a special case of transformer-decoder on text summarization, especially for the long sequence in Wikipedia. There are varieties of sequence to sequence tasks that do not fit in the scenario of text summarization, where encoder-attention-decoder is the dominant approach, such as neural machine translation, response generation, text style transfer, etc. Besides, there are a lot of sequence to sequence tasks beyond pure text, such as speech, image, video, time series sequence, where transformer-decoder only may not fit.
3. The results with varying K are stable. We have trained more steps on smaller or bigger K, the metric on pre-training and fine-tuning tasks do not change much. The key difference between different K lies in that smaller K will bias the model to pre-train the encoder while bigger K will bias the model to pre-train the decoder, which will affect the performance on downstream seq2seq tasks.



2. I looked through the paper you mentioned quickly, and found out that it was doing text summarization with extra long texts. In table 4, we see that Transformer-ED with input length 500 is better than Transformer-D with input length 4000. For extremely long sequences (L=7500 or 11000), Transformer-DMCA is better, but I suppose this is mainly due to its memory compression capacity (the convolution module reduces input resolutions to a smaller size). So **for moderate sequence lengths, e.g. 500, Transformer-ED is still better.**
3. A nice catch. I agree with you.

(출처: https://www.reddit.com/r/MachineLearning/comments/bn2da0/r_190502450_mass_masked_sequence_to_sequence/)

References

MASS 관련

- https://arxiv.org/pdf/1905.02450.pdf
- https://easyai.tech/en/blog/mass-bert-gpt/
- https://www.slideshare.net/HoonHeo5/masked-sequence-to-sequence-pretraining-for-language-generation
- https://www.reddit.com/r/MachineLearning/comments/bn2da0/r_190502450_mass_masked_sequence_to_sequence/
- https://icml.cc/media/Slides/icml/2019/104(13-11-00)-13-12-00-4889-mass_masked_se.pdf

BERT 관련

- https://mino-park7.github.io/nlp/2018/12/12/bert-%EB%85%BC%EB%AC%B8%EC%A0%95%EB%A6%AC/?fbclid=IwAR3S-8iLWEVG6FGUVxoYdwQyA-zG0GpOUzVEsFBd0ARFg4eFXqCyGLznu7w
- http://docs.likejazz.com/bert/
- https://github.com/google-research/bert

GPT 관련

- https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_pap

Thanks!