MASS

: Masked Sequence to Sequence Pre-training for Language Generation

2019년 5월 논문

전은영

Abstract

- Pre-training과 fine-tuning이 LU에서 성공적 (ex. BERT)
- MASS 제시 for LG
  - BERT를 모티브로 함
  - 인코더-디코더 형식
  - Representation 추출과 언어 모델을 jointly 학습
- SOTA 달성 on unsupervised EN-FR

Abstract

Pre-training and fine-tuning, (e.g., BERT (Devlin et al., 2018),) have achieved great success in language understanding by transferring knowledge from rich-resource pre-training task to the low/zero-resource downstream tasks. Inspired by the success of BERT, we propose MAsked Sequence to Sequence pre-training (MASS) for encoder-decoder based language generation. MASS adopts the encoder-decoder framework to reconstruct a sentence fragment given the remaining part of the sentence: its encoder takes a sentence with randomly masked fragment (several consecutive tokens) as input, and its decoder tries to predict this masked fragment. In this way, MASS can jointly train the encoder and decoder to develop the capability of representation extraction and language modeling. By further fine-tuning on a variety of zero/low-resource language generation tasks, including neural machine translation, text summarization and conversational response generation (3 tasks and totally 8 datasets), MASS achieves significant improvements over baselines without pre-training or with other pre-training methods. Specially, we achieve state-of-the-art accuracy (37.5 in terms of BLEU score) on the unsupervised English-French translation, even beating the early attention-based supervised model (Bahdanau et al., 2015b)[1].

Introduction

- Task의 training data 적고, pre-training data 많다면
  => pre-training & fine-tuning 사용
- (예시) In Computer Vision,
  데이터 많은 ImageNet으로 pre-train 후 task에 맞게 fine-tune
- In NLP,
  - 최근 ELMo,GPT,BERT가 pre-training 방법으로 SOTA 달성 in LU
  - 특히 BERT가 가장 잘 됨

Language Generation (LG)

- 어떤 입력에 조건화된 문장을 생성하는 것이 목표
- Task : 번역, 텍스트 요약, 대화형 응답 생성
- data-hungry, (training data) low/zero-resource
- BERT를 바로 적용 X
  - BERT는 LU를 위해 설계됨, just one 인코더 or 디코더

    => LG는 주로 인코더-디코더 형태
  - LG에 Pre-training을 어떻게 적용시킬지 중요

MASS for LG

- Inspired by BERT
- Seq2Seq 모델 (인코더-디코더)
  - 인코더 : (연속된) 일부가 마스크된 문장을 입력으로 받음
  - 디코더 : (인코더 representation을 조건으로) 마스크된 부분을 예측
- 인코더-디코더 "jointly" 학습하도록 두 단계로 설계됨
  - Predicting : 인코더에서 (연속적으로) masked 부분을 예측
    => 인코더가 마스크 되지 않은 토큰들의 의미를 이해하게 함
  - Masking : 인코더에서 unmasked 부분을 디코더에 넣을 때 마스킹
    => 디코더가 다음 토큰을 예측할 때, 인코더의 representation에 의존

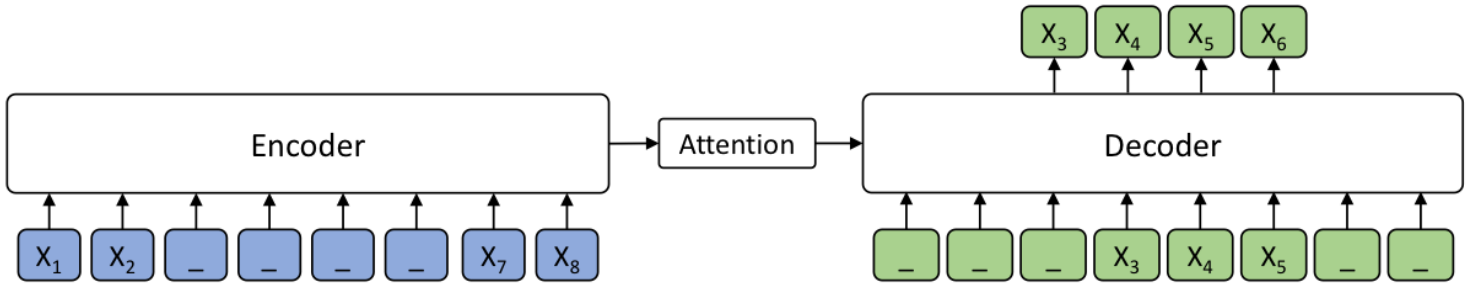
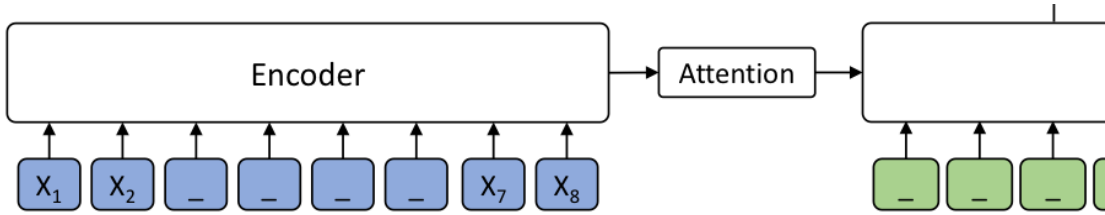


Figure 1. The encoder-decoder framework for our proposed MASS. The token "_" represents the mask symbol [𝕄].

MASS

- Notation

$$
\begin{aligned}
&unpaired\ source\ sentence\ x \in \chi \\
&m\ : \#\ of\ tokens\ of\ sentence\ x \\
&x^{\backslash u:v}\ : modified\ version\ of\ x\ (position\ u\ to\ v\ are\ masked) \\
&x^{u:v}\ : sentence\ fragment\ of\ x\ from\ u\ to\ v \\
&0 < u < v < m \\
&k : \#\ of\ tokens\ being\ masked \\
&[\mathbb{M}]\ : masked\ toekn\ (special\ symbol)
\end{aligned}
$$



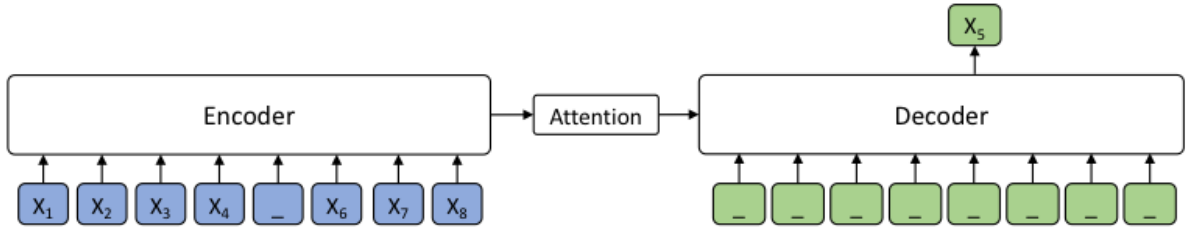
MASS

- Objective function

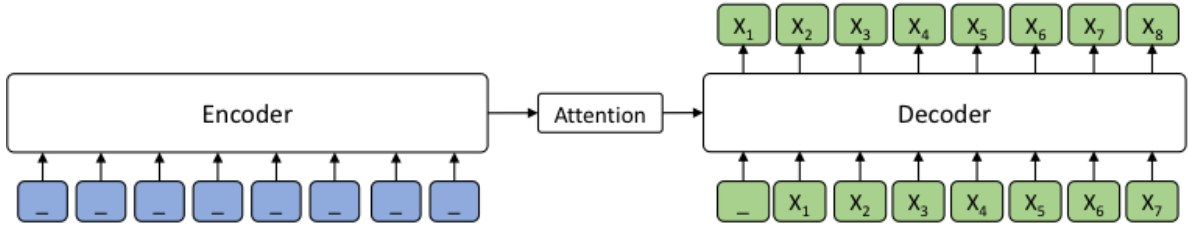
$$
\begin{aligned}
L(\theta; \mathcal{X}) &= \frac{1}{|\mathcal{X}|} \Sigma_{x \in \mathcal{X}} \log P(x^{u:v} | x^{\setminus u:v}; \theta) \\
&= \frac{1}{|\mathcal{X}|} \Sigma_{x \in \mathcal{X}} \log \prod_{t=u}^{v} P(x_t^{u:v} | x_{<t}^{u:v}, x^{\setminus u:v}; \theta).
\end{aligned} \tag{1}
$$

MASS

- BERT와 GPT는 MASS의 special case로 볼 수 있다.



(a) Masked language modeling in BERT ($k = 1$)



(b) Standard language modeling ($k = m$)

Length	Probability	Model
$k = 1$	$P(x^u \mid x^{\setminus u}; \theta)$	masked LM in BERT
$k = m$	$P(x^{1:m} \mid x^{\setminus 1:m}; \theta)$	standard LM in GPT
$k \in (1, m)$	$P(x^{u:v} \mid x^{\setminus u:v}; \theta)$	methods in between

able 1. Masked language modeling in BERT and standard lan-
uage modeling, as special cases covered in MASS.

Discussions

Q. 이전 모델과 차이점…?

- Standard language modeling has long been used for pre-training, and the most prominent ones are the recently proposed ELMo (Peters et al., 2018) and OpenAI GPT (Radford et al., 2018). BERT introduces two pre-training tasks (masked language modeling and next sentence prediction) for natural language understanding, and uses one encoder to extract the representation for a single sentence or a pair of sentences. Both standard language modeling and BERT can just pre-train the encoder or decoder separately. While achieving promising results on language understanding tasks, they are not suitable for language generation tasks which typically leverage an encoder-decoder framework for conditional sequence generation.

- MASS is designed to jointly pre-train the encoder and decoder for language generation tasks. First, by only predicting the masked tokens through a sequence to sequence framework, MASS forces the encoder to understand the meaning of the unmasked tokens, and also encourages the decoder to extract useful information from the encoder side. Second, by predicting consecutive tokens in the decoder side, the decoder can build better language modeling capability than just predicting discrete tokens. Third, by further masking the input tokens of the decoder which are not masked in the encoder side (e.g., when predicting fragment $x_3x_4x_5x_6$, only the tokens $x_3x_4x_5$ are taken as the input and other tokens are masked with $[\mathbb{M}]$), the decoder is encouraged to extract more useful information from the encoder side, rather than leveraging the abundant information from the previous tokens.

MASS Pre-training

Model Configuration We choose Transformer (Vaswani et al., 2017) as the basic model structure, which consists of 6-layer encoder and 6-layer decoder with 1024 embedding/hidden size and 4096 feed-forward filter size. For neural machine translation task, we pre-train our model on the monolingual data of the source and target languages. We respectively conduct experiments on three language pairs: English-French, English-German, and English-Romanian. For other language generation tasks, including text summarization and conversational response generation, we pre-train the model with only English monolingual data respectively. To distinguish between the source and target languages in neural machine translation task, we add a language embedding to each token of the input sentence for the encoder and decoder, which is also learnt end-to-end. We implement our method based on codebase of XLM [4].

Datasets We use all of the monolingual data from WMT News Crawl datasets[5], which covers 190M, 62M and 270M sentences from year 2007 to 2017 for English, French, German respectively. We also include a low-resource language, Romanian, in the pre-training stage, to verify the effectiveness of MASS pre-trained with low-resource monolingual data. We use all of the available Romanian sentences from News Crawl dataset and augment it with WMT16 data, which results in 2.9M sentences. We remove the sentences with length over 175. For each task, we jointly learn a 60,000 sub-word units with Byte-Pair Encoding (Sennrich et al., 2016) between source and target languages.

Pre-Training Details

To reduce the memory and computation cost, we removed the padding in the decoder (the masked tokens) but keep the positional embedding of the unmasked tokens unchanged (e.g., if the first two tokens are masked and removed, the position for the third token is still 2 but not 0). In this way, we can get similar accuracy and reduce 50% computation in the decoder. We use Adam optimizer (Kingma & Ba, 2015) with a learning rate of 10^{-4} for the pre-training. The model are trained on 8 NVIDIA V100 GPU cards and each mini-batch contains 3000 tokens for pre-training.

Fine-Tuning on Unsupervised NMT

Experimental Setting For unsupervised NMT, there is no bilingual data to fine-tune the pre-trained model. Therefore, we leverage the monolingual data that is also used in the pre-training stage. Different from Artetxe et al. (2017); Lample et al. (2017; 2018); Leng et al. (2019), we just use back-translation to generate pseudo bilingual data for training, without using denoising auto-encoder[6]. During fine-tuning, we use Adam optimizer (Kingma & Ba, 2015) with initial learning rate 10^{-4}, and the batch size is set as 2000 tokens for each GPU. During evaluation, we calculate the BLEU score with multi-bleu.pl[7] on *newstest2014* for English-French, and *newstest2016* for English-German and English-Romanian.

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**

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**

Table 2. The BLEU score comparisons between MASS and the previous works on unsupervised NMT. Results on en-fr and fr-en pairs are reported on *newstest2014* and the others are on *newstest2016*. Since XLM uses different combinations of MLM and CLM in the encoder and decoder, we report the highest BLEU score for XLM on each language pair.

Fine-Tuning on Low-Resource NMT

- Baseline : pre-train 없이 low-resource로 train한 모델

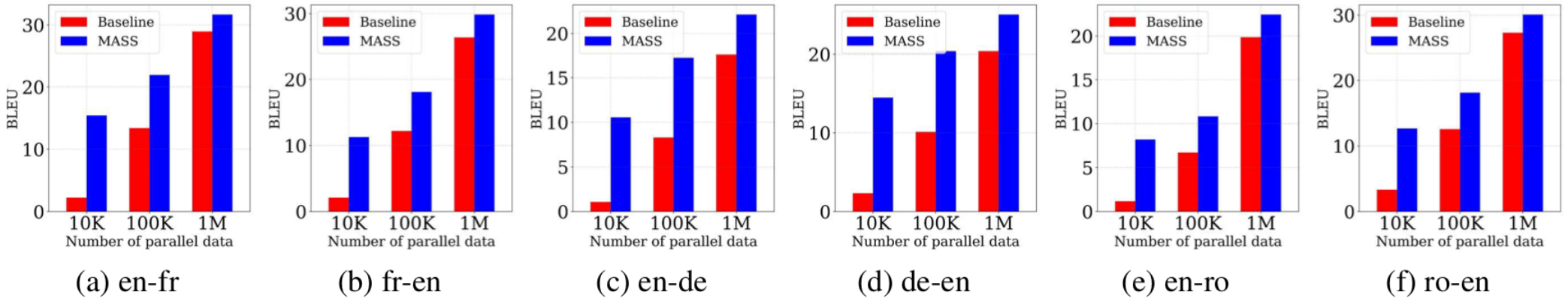


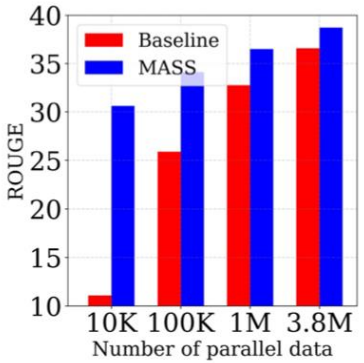
Figure 3. The BLEU score comparisons between MASS and the baseline on low-resource NMT with different scales of paired data.

Fine-Tuning on Text Summarization

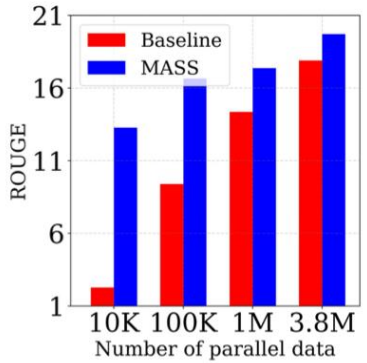
Experiment Setting Text summarization is the task of creating a short and fluent summary of a long text document, which is a typical sequence generation task. We fine-tune the pre-trained model on text summarization task with different scales (10K, 100K, 1M and 3.8M) of training data from the Gigaword corpus (Graff et al., 2003)[9], which consists of a total of 3.8M article-title pairs in English. We take the article as the encoder input and title as the decoder input for fine-tuning. We report the F1 score of ROUGE-1, ROUGE-2 and ROUGE-L on the Gigaword testset during evaluation. We use beam search with a beam size of 5 for inference.

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**

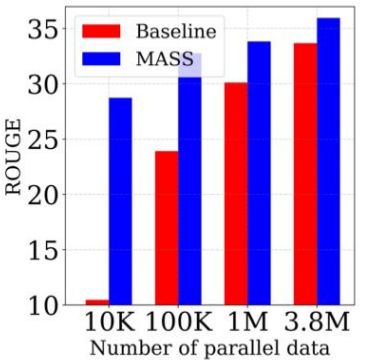
Table 4. The comparisons between MASS and two other pre-training methods in terms of ROUGE score on the text summarization task with 3.8M training data.



(a) RG-1 (F)



(b) RG-2 (F)



(c) RG-L (F)

Fine-Tuning on Conversational Response Generation

Experimental Setting Conversational response generation generates a flexible response for the conversation (Shang et al., 2015; Vinyals & Le, 2015). We conduct experiments on the Cornell movie dialog corpus (Danescu-Niculescu-Mizil & Lee, 2011)[10] that contains 140K conversation pairs. We randomly sample 10K/20K pairs as the validation/test set and the remaining data is used for training. We adopt the same optimization hyperparameters from the pre-training stage for fine-tuning. We report the results with perplexity (PPL) following Vinyals & Le (2015).

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.

Analysis of MASS

- 하이퍼파라미터 k
  - K=50%

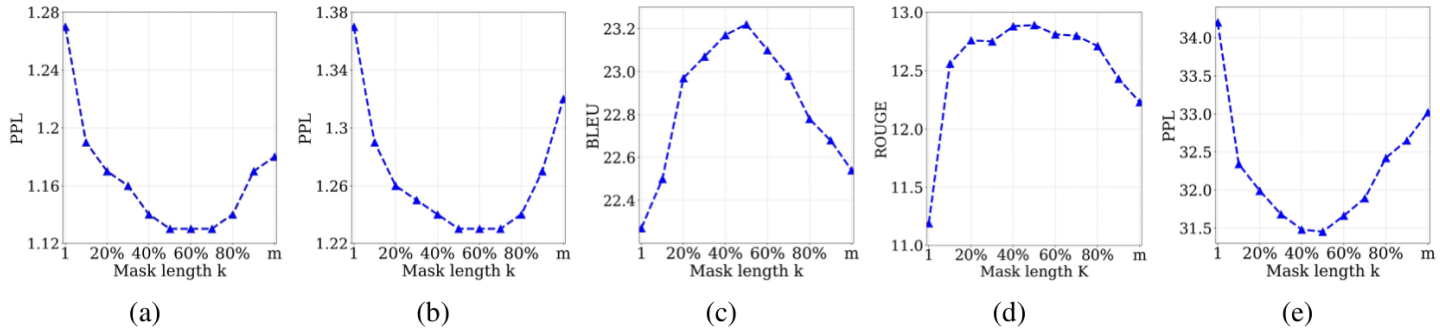


Figure 5. The performances of MASS with different masked lengths k, in both pre-training and fine-tuning stages, which include: the PPL of the pre-trained model on English (Figure a) and French (Figure b) sentences from WMT newstest2013 on English-French translation; the BLEU score of unsupervised English-French translation on WMT newstest2013 (Figure c); the ROUGE score (F1 score in RG-2) on the validation set of text summarization (Figure d); the PPL on the validation set of conversational response generation (Figure e).

- Ablation Study
  - Discrete : 불연속적인 토큰 예측
  - Feed : 디코더에서의 마스킹 제거

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.