# NLPadvanced

**BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension (2020) 논문리뷰**

## BART

- 텍스트에 잡음을 추가한 뒤, 원본 텍스트로 복원하는 모델인 denoising autoencoder를 소개함
- 기존 언어 모델의 변형으로, 임의로 가려진 단어가 있는 텍스트를 재구성하도록 훈련된 자동 인코더
- 잡음이 가해진 문서로부터 원본 문서를 재구성하기 위해 최적화된 비지도 학습 방식을 사용하면서, 다양한 종류의 잡음이 적용될 수 있어 유연성을 가짐.
- 특히 추상적 대화, 질문 응답, 요약 작업 등에서 새로운 최고 성능을 달성했으며, 기계 번역에서도 강력한 베이스라인을 상회하는 성능을 보였음.
- 결국, BART는 기존 BERT와 GPT를 포함한 최신 사전 훈련 체계들을 일반화함으로써 더 넓은 범위의 잡음 스킴을 지원하고, 이를 통해 테스크에 대한 범용성과 적용성을 높임.

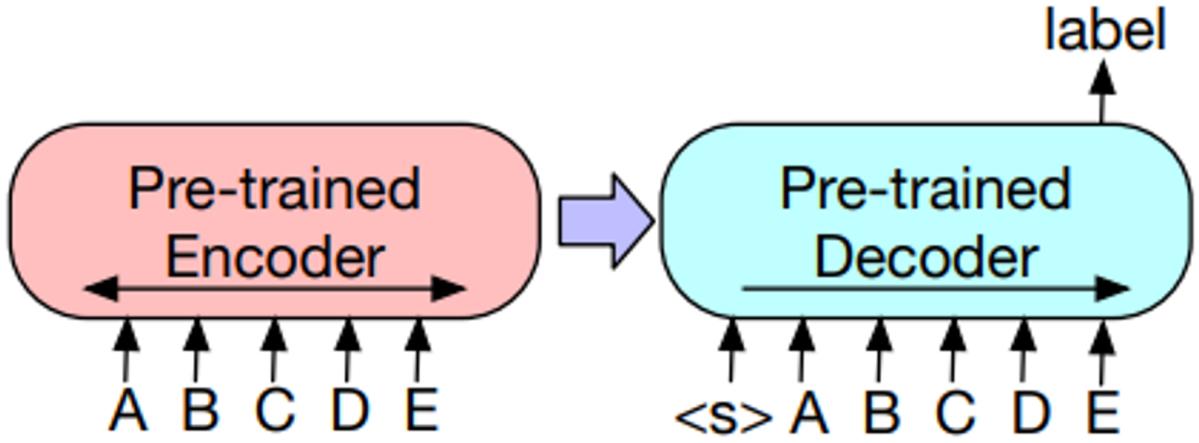
## Model

- BART는 변형된 문서를 원본으로 매핑하는 denoising autoencoder. 양방향 인코더와 left-to-right  자동 회귀 디코더가 있는 seq-to-seq 모델로 구현
- 사전학습 과정에서는 원본 문서의 negative log likelihood (NLL)을 최적화

## Fine-tuning BART

- BART에서 생성된 표현은 downstream applications에 여러 가지 방식으로 사용 가능
- 이 논문에서는 시퀀스 분류 작업, 토큰 분류 작업, 시퀀스 생성 태스크, 기계번역을 설명

### Sequence Classification Tasks




(a) To use BART for classification problems, the same input is fed into the encoder and decoder, and the representation from the final output is used.

- 동일한 입력이 인코더와 디코더에 공급되고 최종 디코더의 최종 hidden state가 새로운 multi-class 선형 분류기에 공급

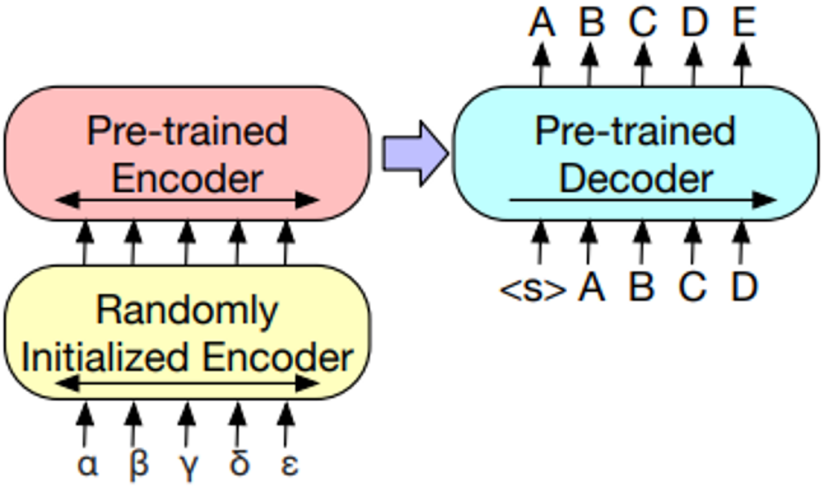
## Token Classification Tasks

- 전체 문서를 인코더와 디코더에 입력하고 디코더의 최상위 hidden state를 각 단어에 대한 표현으로 사용 (이 표현은 토큰을 분류하는 데 사용)

## Sequence Generation Tasks

- BART는 autoregressive decoder를 갖고 있기 때문에, 미세 조정이 가능하며, 다양한 변환을 시도 가능
- 이러한 작업들에서 입력에서 정보가 복사되고 조작되며, 이는 노이즈 제거 사전 훈련 목표와 연관

## Machine Translation




(b) For machine translation, we learn a small additional encoder that replaces the word embeddings in BART. The new encoder can use a disjoint vocabulary.

- 새로운 인코더를 추가하여 전체 BART 모델을 단일 사전 학습된 디코더로 사용하며, 이를 위해 새로운 인코더를 훈련
- 이 과정은 두 단계로 이루어지며, 두 번째 단계에서는 모든 모델 매개변수를 적은 수의 반복을 통해 훈련

## Results

이 논문에서는 base model에 대한 비교를 수행하여 성능을 평가하고있음.

아래는 성능을 보여줌.

| Model | SQuAD 1.1 F1 | MNLI Acc | ELI5 PPL | XSum PPL | ConvAI2 PPL | CNN/DM PPL |
|---|---|---|---|---|---|---|
| BERT Base (Devlin et al., 2019) | 88.5 | **84.3** | - | - | - | - |
| Masked Language Model | 90.0 | 83.5 | 24.77 | 7.87 | 12.59 | 7.06 |
| Masked Seq2seq | 87.0 | 82.1 | 23.40 | 6.80 | 11.43 | 6.19 |
| Language Model | 76.7 | 80.1 | **21.40** | 7.00 | 11.51 | 6.56 |
| Permuted Language Model | 89.1 | 83.7 | 24.03 | 7.69 | 12.23 | 6.96 |
| Multitask Masked Language Model | 89.2 | 82.4 | 23.73 | 7.50 | 12.39 | 6.74 |
| BART Base | | | | | | |
| w/ Token Masking | 90.4 | 84.1 | 25.05 | 7.08 | 11.73 | 6.10 |
| w/ Token Deletion | 90.4 | 84.1 | 24.61 | 6.90 | 11.46 | 5.87 |
| w/ Text Infilling | **90.8** | 84.0 | 24.26 | **6.61** | **11.05** | 5.83 |
| w/ Document Rotation | 77.2 | 75.3 | 53.69 | 17.14 | 19.87 | 10.59 |
| w/ Sentence Shuffling | 85.4 | 81.5 | 41.87 | 10.93 | 16.67 | 7.89 |
| w/ Text Infilling + Sentence Shuffling | **90.8** | 83.8 | 24.17 | 6.62 | 11.12 | **5.41** |

Table 1: Comparison of pre-training objectives. All models are of comparable size and are trained for 1M steps on a combination of books and Wikipedia data. Entries in the bottom two blocks are trained on identical data using the same code-base, and fine-tuned with the same procedures. Entries in the second block are inspired by pre-training objectives proposed in previous work, but have been simplified to focus on evaluation objectives (see §4.1). Performance varies considerably across tasks, but the BART models with text infilling demonstrate the most consistently strong performance.

1. 사전 훈련 방법에 따른 성능의 차이는 크다.
2. 토큰 마스킹은 매우 중요하다.
3. Left-to-right pre-training은 사전 교육을 통해 생성을 개선한다.
4. 양방향 인코더는 SQuAD에 매우 중요하다.
5. Pre-training만이 중요한 요소는 아니다.
6. Pure language model은 ELI5에서 최상의 성능을 발휘한다.
7. BART는 가장 지속적으로 강력한 성능을 달성한다.