



# Seq2Seq

## Abstract

DNN은 이전 어려운 테스크에서 좋은 성능을 보여주었습니다. 많은 라벨이 있는 학습 데이터셋에서도 사용이 가능했지만, 순서쌍에서는 사용할 수 없었습니다. 이 논문에서는 연속적인 학습에 대한 end-to-end 접근 방법을 제시합니다. 이 논문은 LSTM을 인풋 시퀀스를 벡터 차원으로 변형시키기 위해 인코더로 사용하며 다른 LSTM을 아웃풋으로 변환하기 위해 사용합니다. 추가적으로 LSTM은 긴문장 학습에 대한 어려움이 적습니다. 그리고 LSTM은 문맥에 관해 영향을 받는데 우리는 이것을 해결하기 위해 source sentences에 대하여 뒤집어주어 이를 획기적으로 해결했습니다. 이는 단기 의존성 문제를 해결할 수 있게 되었습니다.

## Introduction

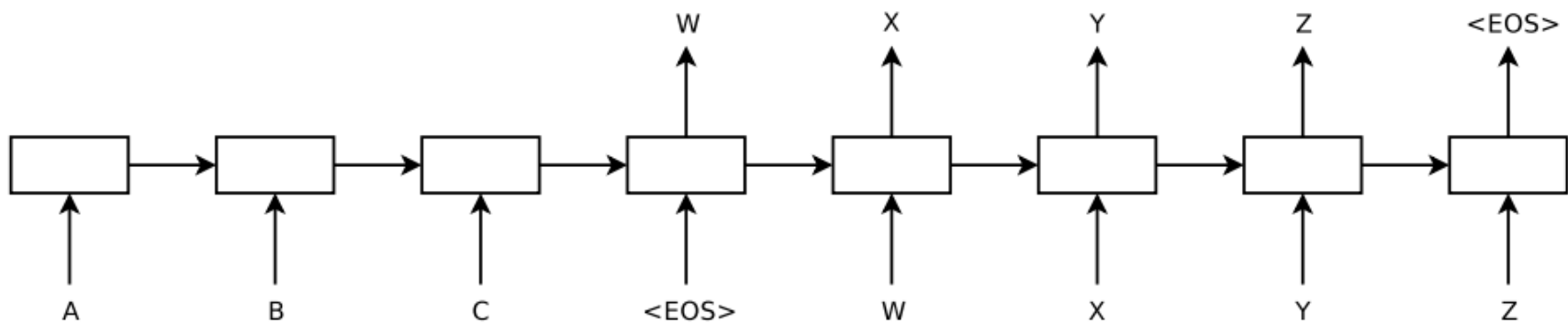
DNN은 많은 task에서 좋은 성능을 거두고 있었지만, input과 target이 고정된 차원으로 인코딩되어야 적용이 가능했습니다. 이에 따라, 음성 인식이나 번역 그리고 Q&A와 같은 연속성을 나타내는 테스크에서 문제가 발생했습니다.

이 논문에서 우리는 정방향성 LSTM이 sequence to sequence 학습에 대한 문제를 해결하는 것을 보여줍니다. 우리의 아이디어는 큰 고정된 차원의 벡터를 표현하기 위해 LSTM이 문장을 받아 한 번에 하나씩 받은 후 이후 이 벡터를 다시 다른 LSTM으로 받아서 아웃풋을 도출해냅니다.

우리는 인풋 문장에 대하여(아웃풋 target값은 아님) 뒤집어서 모델에 넣어줌으로써 단기 의존성을 해결했습니다.

또한 LSTM의 유용함은 다양한 길이의 인풋 문장을 고정된 벡터 차원으로 표현이 가능하다는 것입니다.

그리고 번역은 의역을 하게 되는 경우가 많은데, 우리는 번역을 할 시 어순을 인식하며 능동태와 수동태를 구분하며 정성적으로 평가한 결과 보다 자연스러운 번역을 할 수 있음을 보입니다.



## Model

RNN은 sequence to sequence에서 매핑이 가능하지만 인풋과 아웃풋이 다른 길이를 가지고 있을 경우 불분명합니다. 가장 간단한 방법은 인풋 시퀀스를 고정된 사이즈의 벡터로 변경시키는 것입니다. 그러면 또 다른 RNN이 타겟 문장으로 생성해줍니다. 이러한 과정은 불분명한 정보를 생성하게 되고 RNN은 긴 문장에서 좋지 못한 결과를 만들어냅니다.

**LSTM**은 확률에 근거하며 인풋 문장을 다른 길이의 아웃풋 문장으로 생성하게 됩니다. 이는 길이가 다른 문장에서도 쉽게 번역이 가능하며 과정 중간에 정보를 훼손시키지 않습니다. 그리고 LSTM은 vocabulary에서 각 단어에 해당할 softmax값을 사용하며 문장에 마지막에는 <EOS>토큰을 통해 모델이 문장을 끝맺을 수 있도록 도와줍니다.

우리는 총 3가지의 모델 전략을 사용했습니다.

1. **두가지의 다른 LSTM**, 한개는 Encoder, 다른 하나는 Decoder를 위한 LSTM을 사용하여 모델의 학습량을 줄이면서 동시에 컴퓨팅 자원을 줄일 수 있습니다.
2. **딥 LSTM** : LSTM을 적게 사용하는 것 보다 두껍게 사용하는 것이 더욱 효과적인 것으로 나타났으면 우리는 4개층으로 사용합니다.
3. **Reversed inversed input sentence** : 인풋 문장의 순서를 뒤집어서 input sentence가 a,b,c 순이라면 이를 뒤집어 c, b, a로 넣어 알파, 베타, 감마로 연습합니다. 이는 a, b, c순에 맞춘 번역을 얻게 됩니다.

## Experiments

$$1/|\mathcal{S}| \sum_{(T,S) \in \mathcal{S}} \log p(T|S)$$

목적함수는 위와 같은 사진의 식을 이용하였으며 정답으로 맞추는 로그 확률을 높일 수 있도록 학습이 진행된다.

$$\hat{T} = \arg \max_T p(T|S)$$

우리는 Beam 가설에 대해 적용을 시켜봤는데 이를 적용하였을 때 성능이 오른 것을 확인할 수 있었다. Beam이란 인풋 단어를 통해 다음 단어의 예측 softmax 값 중 Beam의 개수에 따라 가장 높은 값의 개수를 선택한다. 그리고 예를 들어 3개가 선택되었다면 3개를 통해 다시 Beam 가설을 진행한다. 이후 EOS토큰을 만났을 때 이를 종료한다. 이렇게 하는 이유는 기존의 방식을 사용했을 때 가장 높은 softmax값으로 진행되었을 경우 만약 이것이 오류일 때 오류로 계속해서 빠지게 되는 경우가 생기기 때문이다.

우리가 인풋 문장의 순서를 바꿔서 넣어주었을 경우 우리는 학습 과정에서 테스트의 다양성이 줄어들었으며 BLEU scores의 상승이 있었다. 이러한 현상은 단기 의존성에 의해 발생되어진다고 본다. 물론, 인풋 문장을 뒤집었을 경우 인풋 단어와 번역되어지는 단어의 거리 총 평균은 같지만 몇몇 단어는 매우 가까워지게 된다. 이는 백프로파게이션 학습을 더 쉽게 진행시키게 되며 전체적으로 성능이 올라가는 것으로 보여진다. 그래서 우리는 초기에는 초반 도입부분의 단어는 예측을 더 쉽게 하고, 이후 끝 부분에 단어는 예측을 잘 못할 것이라 생각했으나 전반적으로 더 성능이 올라가는 것을 확인할 수 있었다.

우리는 총 4개의 LSTM layer가 존재하는데 그래서 각각의 부분에 gpu를 할당하여 병렬 연결을 통해 보다 빠른 학습이 가능해진다.

## Conclusion

우리의 모델은 다양한 시퀀스 학습 문제를 해결할 수 있다.

또한, 문장을 뒤집음으로써 성능 개선을 한 것은 우리는 단기 의존성과 관련된 결과라고 보여진다.

우리는 LSTM을 사용함으로써 긴 문장에 대해서도 알맞게 번역이 가능했다. 긴 문장은 실패할 것으로 생각했었지만, 문장을 뒤집은 것이 이에도 도움이 되었다.

가장 중요한 것은, 우리는 간단하고 정방향 그리고 최적화되지 않은 접근방식은 통계적 기계 번역 시스템보다 좋은 성능을 냈으며 이는 앞으로도 좋은 성능을 나타낼 것이다. 그리고 우리의 모델은 번역 뿐만이 아닌 다른 테스트에도 도전이 가능할 것이다.

**논문에서 사용하는 BLEU score**

$$BLEU = \min(1, \frac{\text{output length}(\text{예측 문장})}{\text{reference length}(\text{정답 문장})}) (\prod_{i=1}^4 \text{precision}_i)^{\frac{1}{4}}$$

앞에 min을 사용한 이유는 예측 문장의 길이가 길어진 경우 그 영향력을 1로 하고 예측 문장의 길이가 짧아진 경우 그 영향력을 1로 하고자 함이다. 예를 들어, 정답 문장이 have a nice day 인데 예측 문장이 '좋은 하루 보내시고 행복한 하루 보내세요'. 6개가 되어버렸을 경우 6/4=1.5의 영향력을 가질 때 그 영향력이 너무 커진다. 그래서 어느정도의 보정을 하기 위해 사용되어진다. 오른쪽의 precision은 정답 문장에 있는 단어가 예측한 문장에서 몇개 있는지 찾아내는 것이다. 단, 정답 문장이 '내 이름은 민성' 이고 예측 문장이 '이름은 민성 민성'이라면 중복을 제외하고 2개를 맞춘 것으로 한다. → (precision은 맞춘 단어의 수 / 예측한 문장의 단어 수) = 2/3 가 된다.

## ChatGPT 이용 논문 리뷰

"신경망을 이용한 시퀀스 투 시퀀스 학습"은 일리아 수츠케버, 오리올 빈알스, 쿠옥 르 등 Google의 연구원들이 2014년에 발표한 중요한 논문입니다. 이 논문은 신경망 기계 번역에 대한 새로운 접근 방식을 소개했으며, 이후 다른 많은 자연어 처리 작업의 기반이 되었습니다.

이 논문의 핵심은 가변 길이 입력 시퀀스를 가변 길이 출력 시퀀스에 매핑할 수 있는 신경망 아키텍처를 시퀀스 간 모델이라고 하는 신경망 아키텍처를 사용했다는 점입니다. 저자들은 인코더와 디코더라는 두 개의 개별 신경망으로 구성된 인코더-디코더 아키텍처라는 특정 유형의 시퀀스-투-시퀀스 모델을 제안했습니다. 인코더는 입력 시퀀스를 처리하고 고정 길이 벡터 표현을 생성하며, 이 표현은 디코더 네트워크에 공급되어 출력 시퀀스를 생성합니다.

저자들은 이 모델을 기계 번역에 적용하여 여러 벤치마크 데이터 세트에서 최첨단 성능을 달성했습니다. 또한 이미지 캡션 및 음성 인식과 같은 다른 시퀀스 간 작업에도 이 모델을 사용하여 모델의 유연성을 입증했습니다.

이 논문의 강점으로는 명확하고 간결한 글쓰기 스타일, 상세한 기술 설명, 제안된 모델의 효과를 입증하는 실험 결과 등이 있습니다. 저자가 주의 메커니즘을 사용하여 모델의 성능을 개선하고 긴 입력 시퀀스를 처리한 것도 중요한 기여를 했습니다.

그러나 이 논문의 한 가지 잠재적인 약점은 기계 번역에 초점을 맞추고 있어 다른 자연어 처리 작업에 적용하는 데 제한이 있을 수 있다는 점입니다. 또한 이 논문의 실험적 평가는 주로 합성 작업에 초점을 맞추었기 때문에 실제 성능을 정확하게 반영하지 못할 수도 있습니다.

전반적으로 "신경망을 이용한 시퀀스 간 학습"은 시퀀스 간 학습을 위한 강력한 신경망 아키텍처를 소개한 중요한 논문입니다. 이 논문은 자연어 처리 분야에 큰 영향을 미쳤으며, 오늘날까지도 널리 인용되고 연구되고 있습니다.

## 참고 자료

### Sequence to Sequence Learning with Neural Networks

<https://arxiv.org/abs/1409.3215>