Sequence to Sequence Learning with Neural Networks

한양대학교 산업공학과
IDSL 석사과정 고경준

HANYANG UNIVERSITY

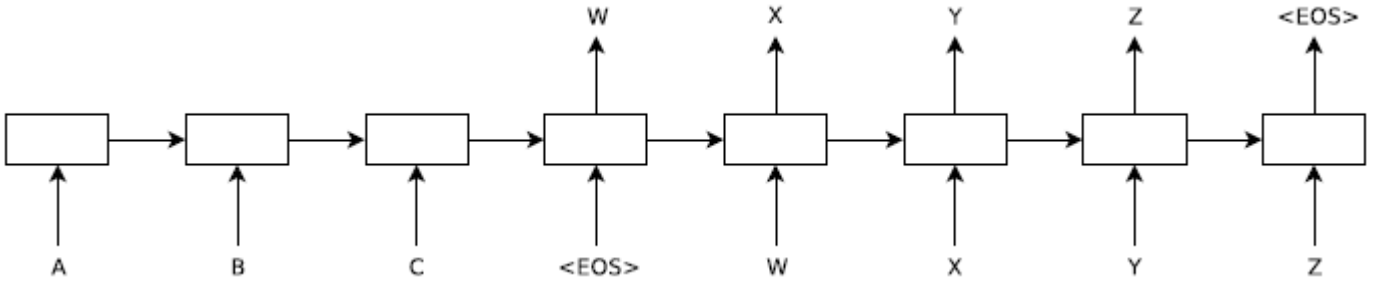
Contents

1. Introduction

Introduction

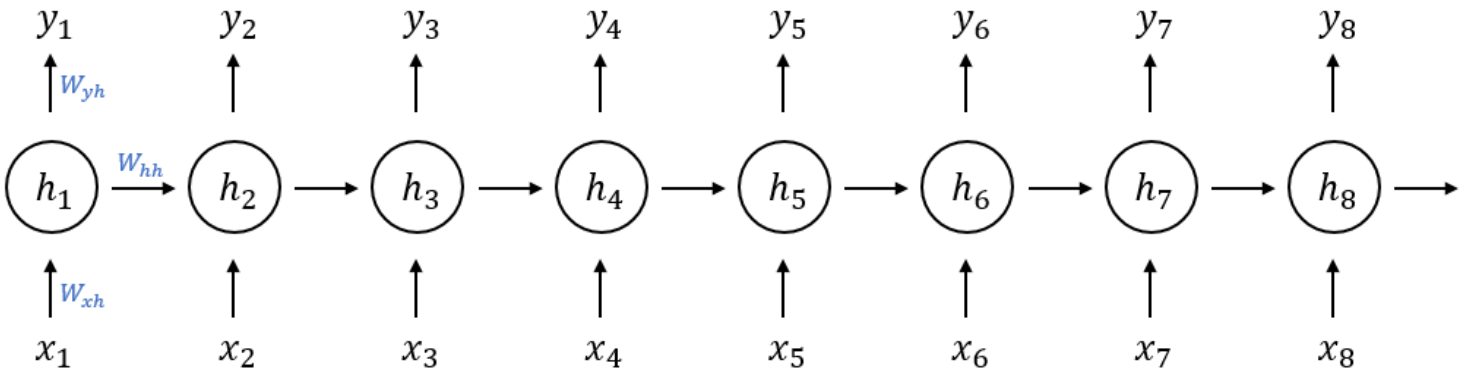
- DNN은 많은 분야에서 훌륭한 성능을 보이고 있음
- 하지만 DNN은 inputs와 targets를 encoded fixed dimensional vectors로 표현해야 함
- 여러 도메인에서는 inputs와 targets가 sequences로 표현됨 (시간 순서를 지님)
- Sequence를 다루기 위해 RNN, LSTM 등의 모형이 제안되었음.

- 본 논문에서는 위 모형들을 사용하여 translation task를 효과적으로 진행하는 모델을 제안.
- Input sequence를 fixed length의 vector로, 이를 다시 output sequence로 변경 (seq2seq)



2. The model

Using RNN model

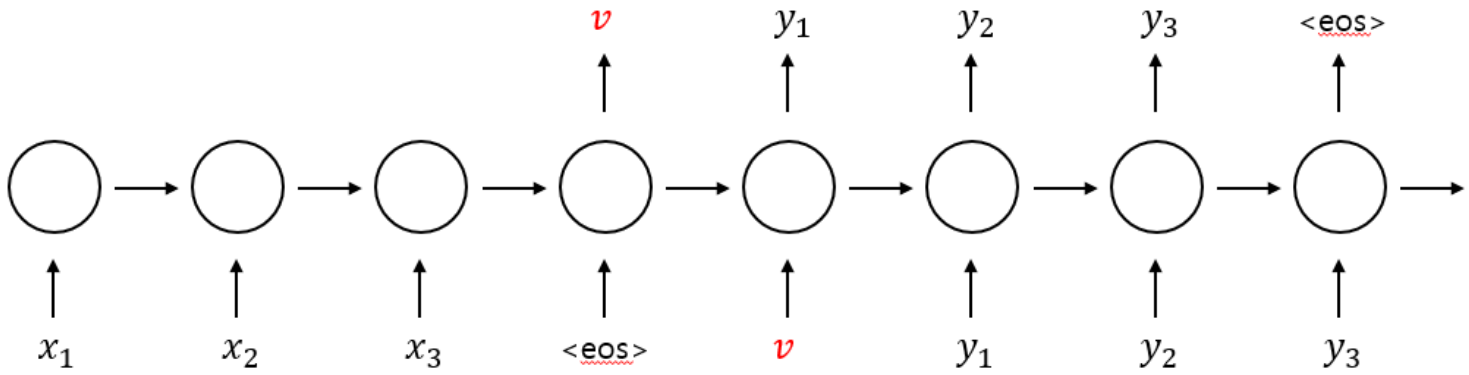


$$h_t = \text{sigm}\left(W^{\text{hx}} x_t + W^{\text{hh}} h_{t-1}\right)$$
$$y_t = W^{\text{yh}} h_t$$

- 일반적인 RNN으로 translation (input sequence를 x, output sequence를 y로) 할 경우, 번역 전 문장의 길이와 번역 후 문장의 길이가 같아져야 함
- 이를 해결하기 위해 'RNN -> fixed length vector (latent representation) -> RNN' 방식
- RNN의 gradient vanishing problem 해결을 위해 LSTM을 사용

2. The model

Basic idea of seq2seq



$$p(y_1, \ldots, y_{T'} | x_1, \ldots, x_T) = \prod_{t=1}^{T'} p(y_t | v, y_1, \ldots, y_{t-1}) \qquad (1)$$

- Input sequence를 fixed length vector **v**로, **v**를 output sequence로 변환 (input과 output sequence length가 달라질 수 있음)
- 각 cell은 LSTM cells

2. The model

3 Additional upgrade methods

1. Two different LSTMs
- Encoder와 decoder에 각각 다른 LSTM 모형을 사용
- 여러가지 pair의 언어 쌍 간의 translation에 사용 가능
  (하나의 encoded vector를 두개 언어 decoder에 각각 사용하여 두가지 언어 translation)
- Model parameter 증가는 무시할 수 있을 정도의 수준

2. Deep LSTMs
- Shallow LSTMs보다 deep(4 layer) LSTMs를 사용했을 때 눈에 띄는 성능 향상

3. Reverse the order of input sequences
- Input sequence의 단어 순서를 반대로 학습
- LSTM을 사용하여 input sequence가 길어지면 앞부분의 단어 기억이 없어짐
- Decoder에서는 첫 단어부터 생성해내기 때문에 input sequence를 뒤집으면 decoder에서 예측해야 하는 첫 단어와 가장 관련있는 encoder의 첫 단어의 기억이 가장 많이 보존 됨

3. Experiments

Dataset details

- WMT'14 [English to French dataset]
  (12M sentences with 348M/30M French/English words)

Decoding and Rescoring

- (Correct translation T | source sentence S)의 log probability를 maximize하도록 training

$$1/|\mathcal{S}| \sum_{(T,S)\in\mathcal{S}} \log p(T|S) \qquad \hat{T} = \arg\max_{T} p(T|S) \tag{2}$$

Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size 2	34.50
Ensemble of 5 reversed LSTMs, beam size 12	**34.81**

3. Experiments

- Beam search 설명 추가
- Rescoring 설명 추가

Reversing the Source Sentences

- Input sequence를 reverse 했을 때 실험적으로 성능이 좋게 나타났음
  (perplexity 5.8 -> 4.7, BLEU scores 25.9 -> 30.6)
- Source language와 target language의 corresponding words 사이의 average distance는 같으나, reverse하면 minimal time lag을 효과적으로 줄이게 됨
- Reverse하면 sentences의 초기 부분의 학습이 잘 되고 끝 부분의 학습이 안 될 것이라고 추측했으나, 그렇지 않은 결과를 보임

3. Experiments

Training details

- We initialized all of the LSTM's parameters with the uniform distribution between -0.08 and 0.08
- We used stochastic gradient descent without momentum, with a fixed learning rate of 0.7. After 5 epochs, we begun halving the learning rate every half epoch. We trained our models for a total of 7.5 epochs.
- We used batches of 128 sequences for the gradient and divided it the size of the batch (namely, 128).
- Although LSTMs tend to not suffer from the vanishing gradient problem, they can have exploding gradients. Thus we enforced a hard constraint on the norm of the gradient [10, 25] by scaling it when its norm exceeded a threshold. For each training batch, we compute $s = \|g\|_2$, where g is the gradient divided by 128. If $s > 5$, we set $g = \frac{5g}{s}$.
- Different sentences have different lengths. Most sentences are short (e.g., length 20-30) but some sentences are long (e.g., length > 100), so a minibatch of 128 randomly chosen training sentences will have many short sentences and few long sentences, and as a result, much of the computation in the minibatch is wasted. To address this problem, we made sure that all sentences in a minibatch are roughly of the same length, yielding a 2x speedup.

Parallelization

- 초기 C++을 이용한 single GPU 실험에서는 1,700 words per second의 속도
- 모델의 4-layer LSTM을 4개의 GPU, softmax를 4개의 GPU에서 계산 (총 8-GPU)
- 최종 8-GPU 실험에서는 6,300 words per second의 속도로 10일만에 실행 완료

3. Experiments

Experimental Results

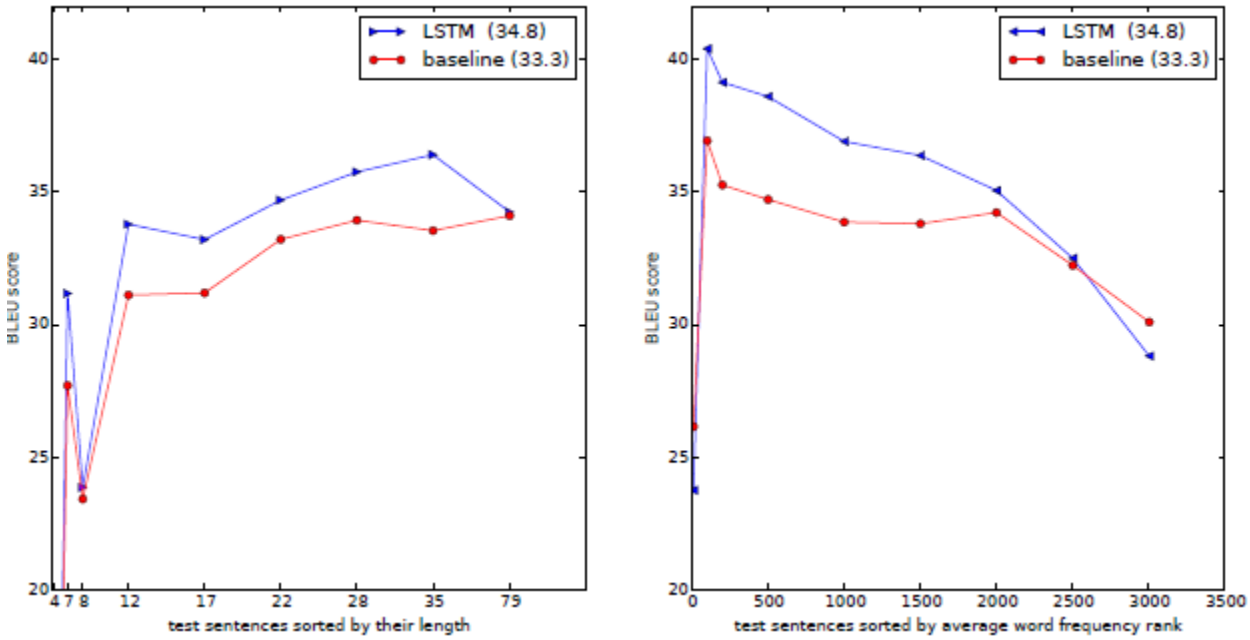
- Random initializations와 random order of minibatches LSTM을 ensemble한 결과가 최선
- Pure neural translation system으로 phrase-based SMT baseline을 처음 outperform함

Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size 2	34.50
Ensemble of 5 reversed LSTMs, beam size 12	**34.81**

Method	test BLEU score (ntst14)
Baseline System [29]	33.30
Cho et al. [5]	34.54
Best WMT'14 result [9]	**37.0**
Rescoring the baseline 1000-best with a single forward LSTM	35.61
Rescoring the baseline 1000-best with a single reversed LSTM	35.85
Rescoring the baseline 1000-best with an ensemble of 5 reversed LSTMs	**36.5**
Oracle Rescoring of the Baseline 1000-best lists	~45

3. Experiments

Performance on log sentences



- Sentences length, word frequency에 따른 성능 비교
- 대부분의 경우 baseline을 넘어서거나 비슷한 성능을 보이는 모습

3. Experiments

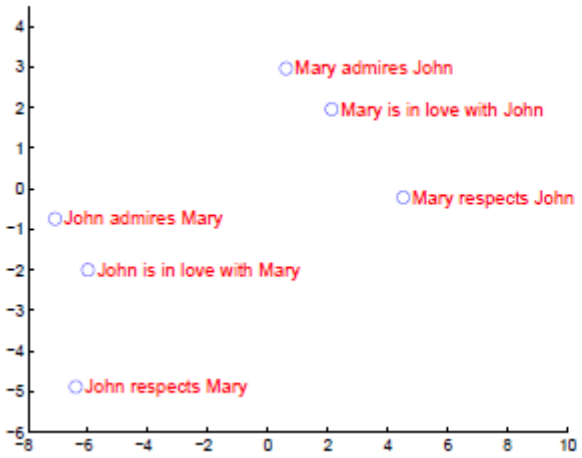
Performance on log sentences

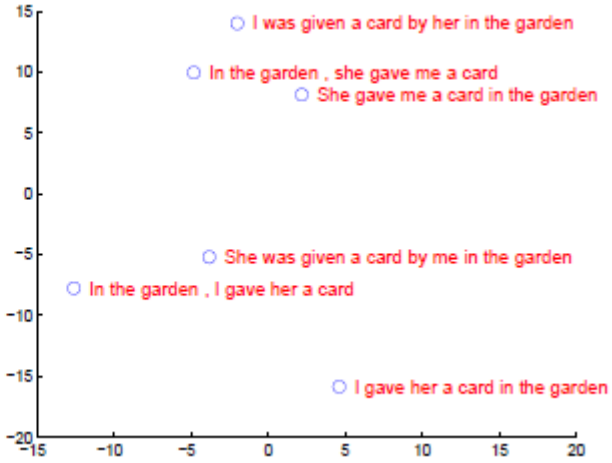
Type	Sentence
Our model	Ulrich UNK , membre du conseil d' administration du constructeur automobile Audi , affirme qu' il s' agit d' une pratique courante depuis des années pour que les téléphones portables puissent être collectés avant les réunions du conseil d' administration afin qu' ils ne soient pas utilisés comme appareils d' écoute à distance .
Truth	Ulrich Hackenberg , membre du conseil d' administration du constructeur automobile Audi , déclare que la collecte des téléphones portables avant les réunions du conseil , afin qu' ils ne puissent pas être utilisés comme appareils d' écoute à distance , est une pratique courante depuis des années .
Our model	" Les téléphones cellulaires , qui sont vraiment une question , non seulement parce qu' ils pourraient potentiellement causer des interférences avec les appareils de navigation , mais nous savons , selon la FCC , qu' ils pourraient interférer avec les tours de téléphone cellulaire lorsqu' ils sont dans l' air " , dit UNK .
Truth	" Les téléphones portables sont véritablement un problème , non seulement parce qu' ils pourraient éventuellement créer des interférences avec les instruments de navigation , mais parce que nous savons , d' après la FCC , qu' ils pourraient perturber les antennes-relais de téléphonie mobile s' ils sont utilisés à bord " , a déclaré Rosenker .
Our model	Avec la crémation , il y a un " sentiment de violence contre le corps d' un être cher " , qui sera " réduit à une pile de cendres " en très peu de temps au lieu d' un processus de décomposition " qui accompagnera les étapes du deuil " .
Truth	Il y a , avec la crémation , " une violence faite au corps aimé " , qui va être " réduit à un tas de cendres " en très peu de temps , et non après un processus de décomposition , qui " accompagnerait les phases du deuil " .

- Long sentences에 대한 model output 예시.
- 번역기를 통해 비교해볼 수 있음

3. Experiments

Model Analysis





- Seq2Seq 모델의 장점 중 하나
- Sequence를 하나의 fixed length vector로 표현할 수 있기 때문에 visualization이 가능
- 단어의 순서가 바뀌었을 때 벡터 표현이 확연히 달라지는 모습을 확인할 수 있음

4. Conclusion

Contributions

- deep LSTM 모형이 problem structure에 관한 가정 없이 standard SMT-based system보다 우수한 성능을 보임
- Source sentences를 reverse하여 성능향상에 도움을 주는 것을 확인
- LSTM의 limited memory로도 long sentences의 translation을 가능하게 함
- Simple, straightforward한 방식으로 SMT system을 outperform하여 다른 sequence to sequence 문제의 접근 방식을 제시함

Weakness

- Fixed dimensional vector로 encoding하는 방식이 bottleneck이 됨
- Input sentences의 길이에 상관없이 fixed length의 vector로 encoding 하는 방식의 문제를 지적하면서 Attention 방법론이 등장

감사합니다.