고차 순환신경망 언어 모델

이상우°, 허민오, 윤상웅, 장병탁
서울대학교 컴퓨터공학부
{slee, moheo, swyoon, btzhang}@bi.snu.ac.kr

High-order Recurrent Neural Network Language Model

Sang-Woo Lee° , Min-Oh Heo, Sang-Woong Yoon, Byoung-Tak Zhang
School of Computer Science & Engineering, Seoul National University

요 약

본 논문에서는 순환신경망 (recurrent neural networks)과 순환신경망 언어 모델 (recurrent neural network language model)의 최근 경이로운 성과들을 소개하고, 그 뒤에 고차 은닉 마코프 체인을 가지는 고차 순환신경망 언어 모델 (high-order recurrent neural network language model)을 제안한다. 위키피디아와 같이 극도로 많은 데이터들이 웹과 데이터베이스에 축적된 상황에서, 순환 신경망 언어 모델과 그 확장은 이전 state-of-the-art n-gram 언어 모델이나 신경망 언어 모델을 뛰어넘는 성능을 보임이 밝혀졌다. 고차 순환신경망 언어 모델 역시 데이터 수가 증가할수록, 은닉 변수간 고차 연결망의 성능 개선 효과가 증가함을 실험을 통해 논증한다.

1. 서 론

오랜 기간 동안 연구되었던 순환 신경망 (recurrent neural network, RNN)은 최근 5년 사이에 큰 진전을 보였으며, 다양한 응용에서 그 두각을 드러내기 시작하였다. 또한 deep learning 선두 그룹에서는 순환 신경망을 deep learning 모델과 representation learning의 일종이자, 다음 세대의 연구 분야로 생각한다 [1].

순환 신경망의 응용에 첫 포문을 연 논문은 Mikolov의 순환신경망 언어 모델 (recurrent neural network language model, RNNLM)이다 [2]. 이 모델은 기존의 state-of-the-art n-gram 모델이나 신경망 언어 모델 (NNLM)보다 언어 모델링에 있어서 더 좋은 성능을 보였다. 이 후로 RNN을 사용하여 단어 수준과 character 수준에서 문장을 생성하려는 시도가 생겼으며, RNNLM의 다양한 변형들이 등장하였다.

본 논문의 배경이론에서는 순환신경망과 순환신경망 언어 모델의 최근 성과들을 소개한다. 그 뒤에 고차 은닉 마코프 체인을 RNNLM에 적용한 고차 순환신경망 언어 모델 (high-order recurrent neural network language model, HORNNLM)을 제안하고 그 성질을 살펴본다. 본 논문에서는 순환 신경망 언어 모델에 고차 연결망을 추가하는 것이 아주 큰 데이터 셋에서 성능 개선에 도움이 됨을 논증한다.

2. 배경 이론

2.1 순환 신경망

순환 신경망 (RNN)을 학습하는 일은 전통적으로 매우 어려운 일이다. 기계학습 분야에서 많이 연구된 Elman RNN은 층이 time-step만큼 많은 다층 신경망 (Multi-layer Perceptron, MLP)의 일종으로 해석할 수 있다. 두 개 이상의 은닉 층을 두었을 때 MLP에 발생하는 문제를 RNN 역시 그대로 겪는다. RNN의 학습을 위하여 error를 이전 시간의 hidden에도 전파시키는 back-propagation through time (BPTT) [3]이 제안되었지만, 이러한 알고리즘으로는 10step 이상의 먼 시간차의 상관관계를 학습할 수 없음이 이론적으로 밝혀졌다 [4]. 이러한 문제를 해결하기 위하여 다양한 모델들이 제안되었으며, 오랜 기간 동안 연구되었다. 예를 들어, Long-short term memory (LSTM)은 은닉층에 3개의 게이트를 사용하여, 특정 시간에 드러나 은닉 정보를 오랫동안 저장, 긴 시간에 대한 정보를 축적함으로써 위 문제를 해결하려 하였다 [5]. 최근에는 deep learning 분야에 제안된 최적화 기법의 일종인 Hessian-Free (HF) 최적화 방법 이 RNN에도 사용되게 되었다 [6]. 이 방법은 Hessian 행렬을 사용하는 2^{nd}-order 최적화 기법의 일종이며, 병리학적인 목표 함수 공간을 가지는 층이 많은 신경망의 학습에서 stochastic gradient descent보다 효과적인 공간 탐색을 수행하는 것으로 밝혀졌다. 그러나, 상대적으로 짧은 시간의 과거 정보가 중요한 언어 및 음악 문제를 다루는 경우, 고전적인 stochastic gradient descent 최적화 방법으로도 좋은 성능을 보일 수 있다 [1, 8]. 본 논문은 stochastic gradient descent를 다루는 학습에 대하여 검토한다.

RNN 언어 모델의 성공 이후로, 다른 응용에서도 RNN이 성공적으로 사용되기 시작하였다. 이러한 예로 Deep RNN을 사용한 음성 인식 [7], RBM-RNN을 사용한 음악 생성 [8], 필기체 생성 [9]등이 있다.

2.2 순환 신경망 언어 모델

언어 모델 (language model)은 자연어 처리 분야에서 매우 오랫동안 연구되어 온 분야이며, 검색 엔진과 음성 인식 등 자연어 처리의 다양한 응용에 핵심이 되는 기술 중 하나이다. 전통적으로 n-gram 모델이 아주 높은 성능을 보였으며, 대표적인 state-of-the-art 모델 중 하나가 Kneser-ney smoothing을 사용한 n-gram 모델이다, 이 모델은 non-parametric Bayesian의 일종인 hierarchical Pitman-Yor processes와 관련되어 설명되기도 하였다 [10]. 별도로 인공신경망 기반으로 한 신경망 언어 모델에 대한 연구가 오랫동안 존재하였으며, n-gram과 비슷한 수준의 성능을 보고하였다. [2]에서는 순환 신경망을 통하여 두 모델의 성능을 넘을 수 있음을 보여주었다. Kneser-ney 모델 대신에 순환 신경망 모델을 사용함으로써 Wall Street Journal 벤치마크 데이터에서 perplexity를 140에서 102로, 위 언어 모델들을 바탕으로 한 음성 인식 과제에서 17.2%의 error를 14.4%로 낮추었다.

비록 단어를 바탕으로 문장을 생성하는 것이 언어 모델의 기본이지만, 본 논문에서는 단어 기반의 문장 생성 모델과, character 기반의 문장 생성 모델 모두를 언어 모델의 일종으로 보고 논의를 전개한다.

RNN을 HF를 이용하여 학습함으로써, character 기반으로 문장 생성을 성공적으로 할 수 있다는 연구가 주목을 받았다 [6]. 하지만 같은 시간을 표현하는 은닉 변수에 대하여, 층을 깊게 쌓는 Deep RNN이 HF 없이도 HF를 사용한 기존 RNN 모델과 성능이 비슷하다는 연구도 최근에 등장했다 [11].

Deep RNN의 성공은 다소 기묘하다. RNN은 이미 deep neural network의 일종이다. 그런데, 여기에 각 시간에 해당하는 은닉 변수 각각에 층을 더욱 쌓아서 기존 RNN을 능가하는 성능을 만들었다. 그렇다면, 이와 같이 RNN의 특정 time step의 hidden을 늘리는 것이 아니라, time step 전후의 시간 관계의 표현을 확장하는 것은 어떨까? 본 논문에서는 RNN에서 은닉 변수간 고차 마코프 가정을 허용하는 방법을 RNN 언어 모델에 적용하고 그 결과를 분석해보고자 한다.

3. 고차 순환 신경망 언어 모델

이전 시간의 은닉 변수와 현재 은닉 변수가 연결을 가지는 RNN을 Elman RNN이라고 한다. 본 논문에서는 Elman RNN과 그 변형에 대해 다룬다. 그림 1은 기존 Elman RNN 모델과 그 확장된 모델을 소개하고 있다. 왼쪽 모델은 단순한 Elman RNN, 가운데 모델은 Deep RNN [7, 11] 오른쪽 모델은 은닉 변수간 2개의 order를 가지는 고차 순환 신경망 (High-order recurrent neural networks, HORNN) 이며, 본 논문이 제안하는 모델이다.

수식을 통하여 HORNN을 정의하면 다음과 같다. 입력 변수 x와 은닉 변수 h, 그리고 출력 변수 o에 대하여 다음 수식이 성립한다 (그림 1-오른쪽 에서 각각 흰색과 검은색 그리고 회색 원에 해당된다).

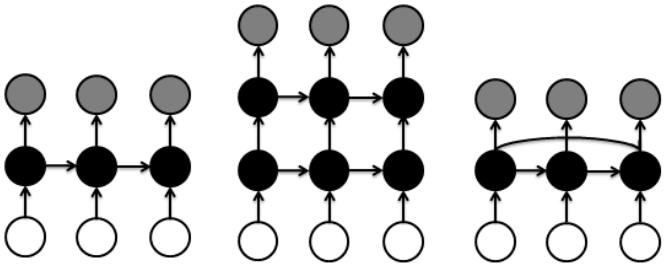

그림 1. (왼쪽) 단순한 (Elman) RNN, (가운데) Deep RNN, (오른쪽) High-Order RNN

$$o_k(t) \propto \sigma(\sum_j w_{kj}^{(o,h)} h_j(t))$$

$$h_j(t) = \sigma(\sum_i w_{ji}^{(h,x)} x_i(t) + \sum_l^L \sum_{j'} w_{jj'}^{(h,l)} h_{j'}(t-l))$$

여기서 σ는 시그모이드 함수, L은 은닉 변수간 고차 연결망의 order이다. 모델의 가중치를 사용하기 위하여 stochastic gradient descent가 사용된다. 정의된 오차 함수 E를 최소화하기 위하여, 오차 함수 E에 대한 가중치의 기울기를 매 입력 값과 상응하는 모델의 값에 대입, 이를 바탕으로 가중치를 수정한다.

$$E(t) = \frac{1}{2}\|y(t) - o(t)\|_2^2$$

$$\delta_k(t) = (y_k(t) - o_k(t)) o_k(t)(1 - o_k(t))$$

$$\frac{\partial E(t)}{\partial w_{kj}^{(o,h)}} = \delta_k(t) h_j(t)$$

$$\frac{\partial E(t)}{\partial w_{ji}^{(h,x)}} = h_j(t)(1 - h_j(t)) x_i(t) \sum_k \delta_k(t) w_{kj}^{(o,h)}$$

$$\frac{\partial E(t)}{\partial w_{jj'}^{(h,l)}} = h_j(t)(1 - h_j(t)) h_{j'}(t-l) \sum_k \delta_k(t) w_{kj}^{(o,h)}$$

HORNN은 기존 RNN과 마찬가지로 언어모델에 사용할 수 있으며, 그 방법은 RNN 언어모델과 같다 [2]. 여기서 입력 변수 x(t)는 t 시간의 단어 index를 one-of-K 방식으로 표현하며, y(t)는 t+1시간의 나올 단어에 대한 모델의 확률이다. 따라서, 일반적으로 y(t)는 x(t+1)과 같으며 o(t)가 x(t+1)의 one-of-K 표상과 같은 값을 가질 때 error가 0이 된다.

4. 실험 결과

본 논문에서는 네 가지의 데이터 셋을 사용하여 고차 순환 신경망 (HORNNLM)과 기존 신경망 언어 모델 (RNNLM)을 비교, 고차 은닉 변수 연결망이 성능에 미치는 영향을 분석하였다. 첫번째 데이터 셋은 Pororo 데이터 셋으로, 어린이 만화영화인 Pororo 동영상의 자막을 추출하여, 그 대사를 데이터 셋으로 만든 것이다. 짧은 대사와 제한된 단어 사용이 특징이다. 전체 한 시리즈 13개의 에피소드를 사용하였다. 두 번째 데이터

셋은 일반적인 만 여개의 문장으로 구성된 데이터 셋이며, Tiny로 지칭하였다. 세 번째 데이터 셋은 Penn Treebank Corpus이며, 표 1에서 PTB에 해당한다. 네 번째 데이터 셋도 Penn Treebank Corpus에서 나온 것인데, 이 때 단어 대신에 캐릭터를 사용하였다. 이 데이터 셋은 표1에서 PTB char에 해당한다. HORNNLM의 order L은 2로 하였다.

	#train	#hidden	RNNLM	HORNNLM
Pororo	1K	60	69.66	74.66
Tiny	10K	60	73.68	74.87
PTC	900K	170	150.13	146.07
PTC char	5M	340	3.25	3.03

표 1. 다양한 데이터 셋에서의 언어 생성 성능 비교, 성능 척도는 perplexity이다. 데이터 크기가 증가함에 따라 고차 연결망이 언어 생성 성능에 도움이 된다.

표 1에 따르면 데이터의 수가 적은 Pororo의 경우 RNNLM이 HORNNLM보다 perplexity가 적으며, 이는 더 좋은 성능을 보임을 의미한다. 하지만 데이터의 수가 많은 PTC나 PTC char의 경우, 같은 은닉 변수 개수에 대하여 HORNNLM이 더 좋은 성능을 보인다.

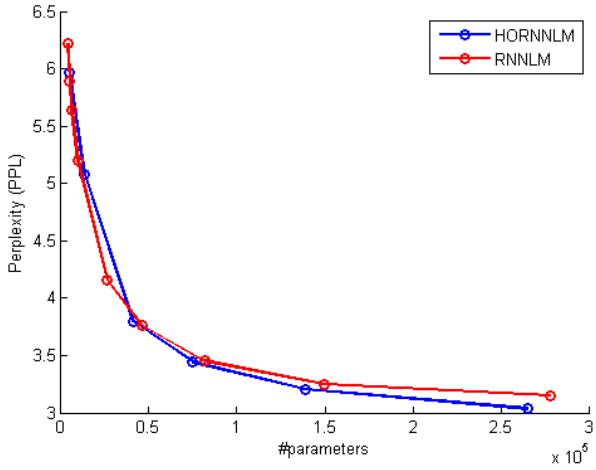

그림 2. Peen Treebank Corpus character 데이터 셋에서 parameter 수의 증가에 따른 RNNLM과 HORNNLM의 성능 변화

그림 2는 가중치의 수가 증가함에 따라, HORNNLM의 성능이 RNNLM의 성능보다 더 좋아짐을 보여준다. 처음에는 같은 가중치 개수에 대하여 HORNNLM이 RNNLM에 비해 낮은 성능을 보였다. 하지만 가중치 수가 늘어남에 따라, HORRNLM이 RNNLM에 비해 성능이 개선되는 것을 확인할 수 있었다. 이러한 성능 개선은 PTC char 뿐 아니라 PTC 에서도 확인할 수 있었다.

한편, HF와 같은 다른 최적화 방법을 사용하였을 때, HORNNLM의 성능이 어떻게 변화하는 지를 검토할 필요가 있다. 현재 BPTT에 대하여 HORNNLM 성능이 어떻게 변화하는 지에 대한 실험이 진행 중에 있다.

Acknowledgement

이 논문은 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구이며(NRF-2010-0017734-Videome,), 정부(산업통상자원부)의 재원으로 한국산업기술평가관리원 지원(KEIT-10035348-mLife, KEIT-10044009)을 일부 받았음.

참고 문헌

[1] Y. Bengio, Representation Learning: A Review and New Perspectives, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 2013.

[2] T. Mikolov, M. Karafiat, L. Burget, J. H. Cernocky, and S. Khudanpur, Recurrent neural network based language model, *11th Annual Conference of the International Speech Communication Association (Interspeech 2010)*, 2010.

[3] M. Boden, A Guide to Recurrent Neural Networks and Backpropagation, *In the Dallas project*, 2002.

[4] Y. Bengio, P. Simard, and P. Frasconi, Learning long-term dependencies with gradient descent is difficult, *IEEE Transactions on Neural Networks*, 5(2), 1994.

[5] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural Computation,* 9(8), 2006.

[6] I. Sutskever, J. Martens, and G. Hinton, Generating Text with Recurrent Neural Networks, *International Conference on Machine Learning 28 (ICML11)*, 2011.

[7] A. Graves, A. -R. Mohamed, and G. Hinton, Speech Recognition with Deep Recurrent Neural Networks, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP13),* 2013.

[8] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, Modeling Temporal Dependencies in High-Dimensional Sequences: Application to Polyphonic Music Generation and Transcription, *International Conference on Machine Learning 29 (ICML12)*, 2012.

[9] A. Graves, Generating Sequences With Recurrent Neural Networks, *arXiv:1308.0850*, 2013.

[10] Y. W. Teh, A Hierarchical Bayesian Language Model based on Pitman-Yor Processes, *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (ACL-44),* 2006.

[11] M. Hermans, B. Schrauwen, Training and Analyzing Deep Recurrent Neural Networks, *Advances in Neural Information Processing Systems 26 (NIPS13)*, 2013.