순환 신경망 언어 모델의 은닉 뉴런의 미래 서열 예측 성질

윤상웅[1○], 이상우[2], 장병탁[1,2]
[1]서울대학교 협동과정 뇌과학전공
[2]서울대학교 컴퓨터공학부
{swyoon, slee, btzhang}@bi.snu.ac.kr

Predictive Property of Hidden Representations in Recurrent Neural Network Language Model

Sangwoong Yoon[1], Sang-Woo Lee[2], Byoung-Tak Zhang[1,2]
[1]Interdisciplinary Program in Neuroscience, Seoul National University
[2]Schoole of Computer Science & Engineering, Seoul National University

요 약

순환 신경망 모델(Recurrent Neural Network)은 시계열 자료를 다룰 수 있는 신경망 기반의 기계학습 모델이다. 많은 경우 순환 신경망 모델은 과거 입력에 대한 기억 성질을 가진다고 해석되는데, 은닉 뉴런 값에는 현재까지의 모든 입력이 반영되어 있기 때문이다. 이 연구에서는 순환 신경망 언어 모델의 은닉 뉴런 값으로부터 아직 입력되지 않은 미래의 서열이 예측될 수 있다는 것을 보임으로서 새로운 해석을 제시한다. 순환 신경망 언어 모델이 데이터를 학습하면 할수록 이 예측 성질이 강해지는 것이 확인 되었고, 이것은 예측 성질이 데이터 속의 규칙성을 포착한 것임을 시사한다. 따라서 잘 학습된 순환 신경망 언어 모델은 주어진 시점에서 과거와 (예상되는) 미래 서열 정보를 모두 가지고 있고, 이 은닉 뉴런들이 실용적인 요소 추출기(feature extractor)로서 감독 서열 레이블링(supervised sequence labeling)에 사용될 수 있다.

1. 서 론

딥러닝의 등장 이래로 신경망 모델들이 많은 주목을 받아오고 있다. 딥 빌리프 네트워크, 딥 컨볼루션 네트워크 등이 깊게 연구되고 있고 음성 인식, 물체 인식과 같은 분야에서 최고의 성능을 보이고 있다. 시간 성분을 고려하는 신경망 모델인 순환 신경망(Recurrent Neural Network, 이하 RNN)도 시간적으로 여러 비선형 층을 가지고 있기 때문에 (그림 1) 딥러닝의 일종으로 분류된다. RNN은 단순한 구조에도 불구하고 학습이 매우 어려운 것으로 알려져 있었으나, 최근 언어 모델에 성공적으로 적용되어 [1] 널리 쓰이는 n-gram 언어 모델을 능가하는 성능을 보여줌으로서 실용적인 기계학습 알고리즘으로서의 지위를 확고히 했다.

딥러닝, 특히 딥 컨볼루션 네트워크가 가장 많이 적용된 데이터는 정지 사진이고, 이 경우 은닉 뉴런의 가중치를 시각화했을 때 해석 가능한 패턴을 보여주는 경우가 많아 은닉 뉴런의 역할이 상대적으로 명확하게 이해되었다. 또한 contractive autoencoder 등의 모델을 통해 딥러닝에서 은닉 뉴런이 표상하는 공간이 어떤 특성을 가지고 있는지 이해하려는 시도들이 있었다. [2] 그러나 RNN의 경우 은닉 뉴런 정보는 다차원의 시계열을 나타내게 되는데, 이는 시각화하기도 어렵고, 시각화하더라도 의미를 추출하기가 어려워 이에 대한 이해가 부족한 상황이다.

RNN에 대한 가장 기본적인 분석은, 은닉 뉴런이 과거 입력에 대한 기억을 가지고 있다는 것이다. 이것은 일견 당연한 것으로서, 은닉 뉴런의 활동값이 과거 입력값으로부터 완전하게 결정되기 때문이다. 실제로 RNN의 기억 성질을 측정하고, 또 향상시키는 방향으로 많은 연구들이 이루어져 왔다. [3][4] 이러한 시각에 따르면, 순환 신경망 언어 모델의 은닉 뉴런 값에는 과거부터 현재까지의 언어(글자 혹은 단어) 입력이 반영되어 있는 것이다.

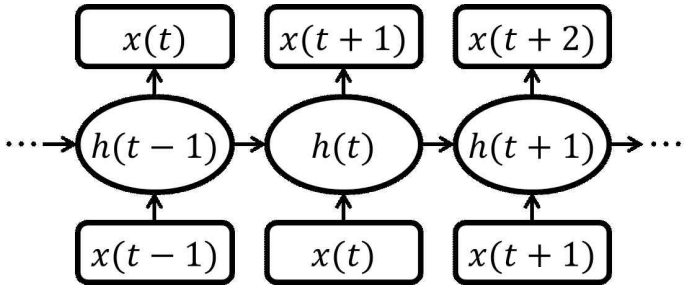

그림 1 순환 신경망 언어 모델.
변수에 대한 정의는 2절 참조.

이 연구에서는 RNN의 은닉 뉴런에 대한 통상적인 이해를 넘어서, 순환 신경망 언어 모델(RNN Language Model, 이하 RNNLM)의 경우 은닉 뉴런의 값이 아직 입력으로 들어오지 않은 미래의 서열 값과도 상관

(correlated)되어 있음을 보일 것이다. 이것이 가능한 이유는 RNN이 학습을 통해 데이터의 규칙성을 포착하였기 때문이라고 생각되며, 그것을 실험을 통해 입증할 것이다.

이러한 점들을 고려할 때, RNNLM의 은닉 뉴런 값은 과거와 미래를 아우르는 국지적 문맥 요소 추출기(local context feature extractor)로 생각되어야 한다. 즉, 잘 학습된 RNNLM은 주어진 시점에서 앞뒤 문맥을 요약한 요소를 추출할 수 있다. 추출된 요소는 다른 자연어 처리 문제 해결에 사용될 수 있는데, 특히 언어 모델을 직접 사용하지 않던 문제들에 도움이 될 가능성이 있다. 본 연구에선 이러한 가능성을 조명하고 RNNLM의 요소 추출기(feature extractor)로서의 이론적 근거를 제시한다.

2. 순환 신경망 모델

2.1 모델 정의

RNN은 시간이라는 요소가 추가된 다층 퍼셉트론(Multilayer perceptron, 이하 MLP)이라고 생각할 수 있으며, 현재 가장 널리 받아들여지는 모델은 Elman이 처음 제안한 것으로 Elman-style RNN이라고 하기도 한다. [5] 이 모델은 MLP와 마찬가지로 입력 뉴런, 은닉 뉴런, 출력 뉴런으로 이루어져 있는데, 차이점은 다음 시점의 은닉 뉴런의 값에 현재 은닉 뉴런 값이 영향을 미친다는 점이다. 임의의 길이를 가지는 다차원 서열 $\{x(t)\}$의 매 시점 $x(t)$가 입력 뉴런에 순서대로 입력되면, 이 값은 은닉 뉴런과 출력 뉴런을 거치면서 사용자가 원하는 출력값으로 변환된다. (그림 1)과 다음 식이 모델의 구조를 자세하게 나타내고 있다.

$$h(t) = \sigma(W_{in}x(t) + W_r h(t-1) + b_h)$$
$$y(t) = \phi(W_{out}h(t) + b_y)$$

여기에서 $x(t)$, $y(t)$, $h(t)$는 각각 시점 t의 입력, 출력, 은닉 뉴런의 값이며, 여러 뉴런의 값을 한 번에 세로줄 벡터로 표현한다. W_{in}, W_{out}, W_r은 각각 입력-은닉, 은닉-출력, 은닉-은닉 뉴런 간 연결 가중치 행렬이다. b_h, b_y는 각각 은닉 뉴런과 출력 뉴런의 편향값(bias)이고, σ, ϕ는 비선형함수로서 대개는 시그모이드나 쌍곡탄젠트 함수이다.

RNN의 학습은 여타 신경망 모델과 유사하게 오차를 줄이는 방향으로 연결가중치를 조절하는 기울기 하강(gradient descent) 방법으로 이루어진다. 출력 뉴런이 계산되는 데에 과거 모든 입력이 관여 되어 있으므로, 오차를 과거로 전파하여 가중치를 조정해야하고 이 과정을 Back Propagation Through Time (BPTT) 알고리즘이라고 한다.

2.2 순환 신경망 언어 모델

RNN의 가장 대표적이며 실용적인 응용은 언어 모델링 문제이다. 언어 모델링이란, 언어 토큰(글자 혹은 단어)의 서열이 있을 때 다음 토큰의 확률을 예측하는 문제이다. 이 문제는 언어 토큰 서열을 1-of-K 벡터로 표현하여 입력 뉴런에 대응시키고, 출력 뉴런은 다음 토큰 벡터를 가리키게 하면 (즉, $y(t) = x(t+1)$) RNN으로 자연스럽게 표현된다. 그에 걸맞게 기존에 널리 쓰이던 언어 모델링 방식인 n-gram을 능가하는 성능이 보고되었다.

언어 모델이 있으면 언어의 서열에 확률을 부여할 수 있기 때문에, 음성 인식, 품사 태깅 등 여러 자연어처리 분야에서 핵심적인 기술이다. 한편, 언어 서열의 확률을 사용하지 않는 자연어 처리 문제들은 좋은 언어 모델로부터 직접적으로 혜택을 받기가 어렵다.

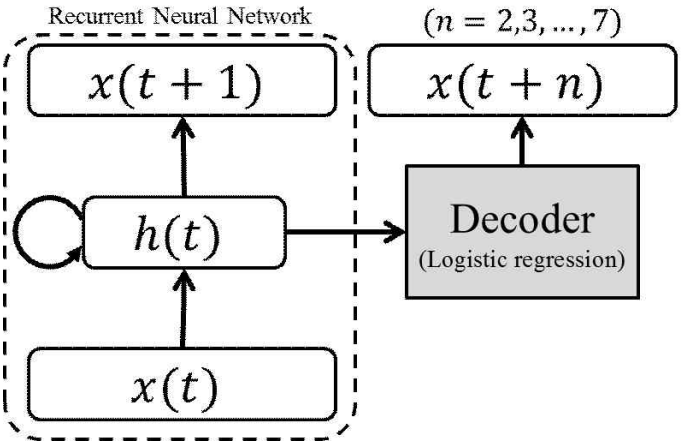

그림 2 순환 신경망 언어 모델의 은닉 뉴런으로부터 미래 서열을 예측하는 실험 개요도. 시점 t의 은닉 뉴런과 시점 t+n의 언어 토큰이 얼마나 상관성이 있는지 알아본다.

3. 미래 서열 예측 실험

보통의 RNNLM은 다음 언어 토큰($x(t+1)$)만을 잘 예측하도록 학습된다. 그러나 우리는 RNNLM이 언어 서열에 내재된 규칙성을 잘 포착하여 효율적인 은닉 표상(hidden representation)을 학습했다면, 은닉 뉴런의 값($h(t)$)과 미래 언어 토큰들($x(t+2)$, $x(t+3)$, ...)에 상관성이 존재할 수 있을 것이라는 가설을 세웠고, (그림 2)에 묘사된 실험을 통해 검증했다.

300개의 은닉 뉴런을 가지는 RNN을 Penn Treebank 말뭉치(corpus)로 학습하여 알파벳 하나를 토큰 하나로 하는 글자 수준 언어 모델을 얻었다. 이 학습된 RNNLM에 말뭉치의 언어 서열을 입력했을 때 은닉 뉴런 값을 수집하였고, 이 값들이 두 글자에서 일곱 글자 까지 떨

어진 미래 언어 토큰들과 상관성이 있는지 확인하였다. 상관성은 logistic regression으로 검사하였는데, 그 이유는 이를 통해 은닉 표상 공간이 선형적 분리가능(linearly separable)한지 확인할 수 있기 때문이다.

실험을 통해 드러난 상관성이 데이터를 학습한 것에서 유래한 것인지 확인하기 위해, RNNLM의 학습 수준을 제한한 상태에서 같은 실험을 반복하였다. 학습 epoch을 1회로 제한함과 동시에 학습 데이터의 양을 줄이는 방식으로 세 가지 덜 학습된 RNNLM을 구축하였다. 각각은 학습된 정도 순으로 untrained, suboptimal 1, suboptimal 2라고 명명되어 (그림 3)에 표시되었다. 각 모델이 학습된 정도는 언어 모델링 성능을 통해 확인할 수 있는데, (그림 3)에 표시된 네 가지 RNNLM의 Test perplexity는 43.6 (untrained), 13.0 (suboptimal 1), 4.87 (suboptimal 2), 2.93 (fully trained) 이다.

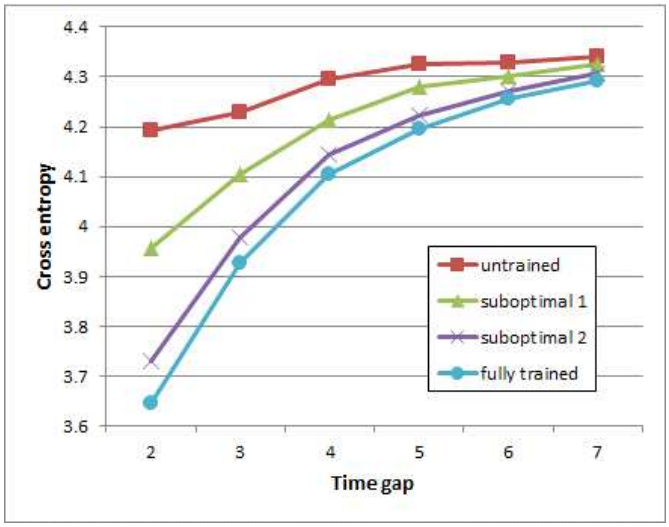

그림 3 미래 언어 토큰 예측 실험 결과. 은닉 뉴런으로부터 time gap 만큼 떨어져 있는 미래 언어 토큰을 logistic regression으로 예측하였다. Suboptimal 1, suboptimal 2는 training set의 크기를 제한하여 만들어진, 성능이 좋지 않은 RNN이다.

놀랍게도, RNNLM의 은닉 뉴런 값은 아직 RNNLM이 관측하지 않은 미래 언어 토큰과 상관성을 가지는 것으로 나타났다. RNNLM이 학습되지 않을수록, 또 멀리 떨어진 언어 토큰일수록 상관성이 약하다는 점은 이러한 현상이 임의적인 것이 아니라 RNNLM이 포착한 언어 서열의 규칙성 때문임을 시사한다.

4. 응용가능성

RNNLM의 은닉 뉴런은 과거 서열 입력뿐만 아니라 미래 서열에 대한 예측까지 포함하기 때문에, 보다 완전한 의미에서의 국지적 문맥 요약이라고 해석되고 응용될 수 있다. 이것이 특히 의미 있는 이유는 언어 모델을 사용하지 않은 자연어 처리 문제에 RNNLM이 학습한 언어의 규칙성을 적용할 수 있는 근거를 제공하기 때문이다. 실제로 [6]는 감독 서열 레이블링 문제에 RNNLM의 은닉 뉴런 값을 적용하여 성능을 향상시킨 바 있으나, 성능 향상에 대한 타당한 근거를 제공하지 못했다. 이번 연구의 결과는 [6]의 실험 결과에 대한 설명을 제시함과 동시에 더 넓은 응용가능성을 시사한다.

5. 결 론

우리는 RNNLM의 은닉 뉴런을 분석하여, 아직 모델이 관측하지 못한 미래의 언어 토큰들과 상관성이 있다는 것을 보였다. 이는 임의적인 결과가 아니라 RNNLM이 언어 데이터를 학습한 것에서 기인하는 효과이며, 언어 서열에 내재된 규칙성을 반영하는 것이다. 이러한 결과는 RNNLM의 은닉 뉴런에 대한 새로운 시각이며, 새로운 응용 가능성을 시사한다.

주의해야할 점은 이것이 출력 뉴런을 다음 토큰으로 하는 RNNLM에 대해서만 성립한다는 점이다. 출력 뉴런이 다음 토큰이 아니라 별도의 레이블일 경우(즉, 감독 서열 레이블링 문제에 RNN을 사용하는 경우), 이러한 성질을 기대할 수는 없다. 같은 RNN이라도 RNNLM 형태로 형식화한 경우와 서열 레이블링 형태로 형식화한 경우 다른 성질을 보이는 것에 대한 고찰이 심도 있게 이루어진 바 없는데, 이는 RNN에 대한 이해를 한층 심화시킬 수 있는 연구 방향이라고 생각된다.

참고 문헌

[1] T. Mikolov. Statistical language models based on neural networks. Ph.D. thesis, Brno University of Technology, 2012.

[2] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio. Contractive Auto-Encoders: Explicit Invariance During Feature Extraction. In Proceedings of the 28th International Conference on Machine Learning (ICML11), pages 833-840, 2011.

[3] S. Hochreiter, J. Schmidhuber. Long short-term memory. Neural computation 9(8): 1735-1780

[4] J. Martens, and I. Sutskever. Learning Recurrent Neural Networks with Hessian-Free Optimization. In Proceedings of the 28th International Conference on Machine Learning (ICML11), pages 1033-1040, 2011.

[5] J. Elman. Finding Structure in Time. Cognitive Science. 14, 179-211, 1990.

[6] G. Chrupala. Text segmentation with character-level text embeddings. In ICML Workshop on Deep Learning for Audio, Speech and Language Processing, 2013.