Why "Identity Matrix Setting" in RNN Helps Prevent the Gradient Vanishing Problem

최은영
광운대학교

This report is based on 'Le, Q., Jaitly, N., & Hinton, G. (2015). A Simple Way to Initialize Recurrent Networks of Rectified Linear Units. arXiv.org. from https://arxiv.org/abs/1504.00941'

1. Introduction

RNN은 매우 강력한 동적 시스템이며, 음성 인식 및 기계 번역처럼 입력 시퀀스를 출력 시퀀스에 매핑하거나 언어 모델링처럼 시퀀스의 다음 용어를 예측하는 데 신경망을 사용하는 방법이다. 그러나 error-derivatives를 계산하기 위해 backpropagation을 사용하여 RNN을 훈련하는 것은 어려울 수 있다. 초기 시도는 gradients vanishing이나 exploding 때문에 어려움을 겪었고, 이는 long-term dependencies을 학습하는 데 큰 어려움을 겪었다는 것을 의미한다. 이 어려움을 극복하기 위해 다양한 방법들이 제안되었다.

일부 인상적인 결과를 도출하는 방법은 Hessian-Free(HF) optimization 사용하는 것이다. HF는 큰 mini-batches에서 작동하며 매우 작은 gradient를 가지면서 곡률이 더 작은 weight 공간에서 유망한 방향을 감지할 수 있다. 그러나 후속 연구는 weight를 신중하게 초기화하고 큰 gradient를 클리핑한 경우라면 momentum을 가진 SGD를 사용하여 유사한 결과를 얻을 수 있다고 제안했다.

현재까지 가장 성공적인 기법은 LSTM(Long Short Term Memory) Recurrent Neural Network으로 SGD를 사용하지만, backpropagation된 gradient가 훨씬 잘 동작하도록 hidden unit을 변경한다. LSTM은 logistic이나 tanh hidden units을 아날로그 값으로도 저장할 수 있는 "메모리 셀"로 바꾼다. 각 메모리 셀에는 입력이 저장된 아날로그 값에 추가될 수 있는 시점과 이 값이 출력에 영향을 미칠 수 있는 시점을 제어하는 자체 입력 및 출력 게이트가 있다. 이러한 게이트는 입력에서 나오는 연결과 이전 time-step의 메모리 셀에 대한 자체 학습 weight를 가진 logistic unit이다. 메모리 셀에 저장된 아날로그 값이 소멸하는 속도를 제어하는 학습 weights가 있는 forget gate도 있다. 입력 및 출력 게이트가 꺼져 있고 망각 게이트가 소멸을 일으키지 않는 기간 동안, 메모리 셀은 단순히 시간이 지남에 따라 값을 유지하므로 저장된 값에 대한 오류의 gradient는 해당 기간 동안 back-propagation될 때 일정하게 유지된다.

2. The initialization trick

Identity RNN with ReLU activation $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ -1 ⟶ 0

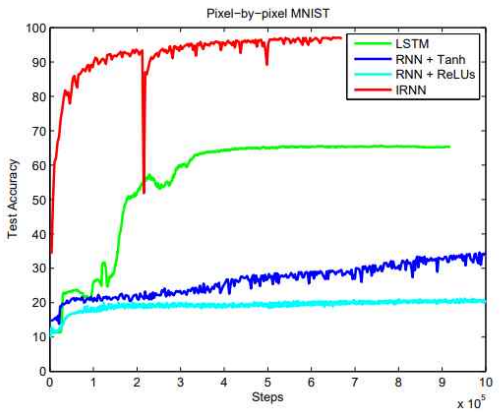
Recurrent network에서 long-term dependency를 학습하는 것은 gradients vanishing이나 exploding로 인해 어렵다. 본 논문에서는 ReLU로 구성된 Recurrent network을 사용하는 해결책을 제공한다. 핵심은 recurrent weight matrix를 초기화하기 위해 identity matrix 또는 scaled version identity matrix를 사용하는 것이다.

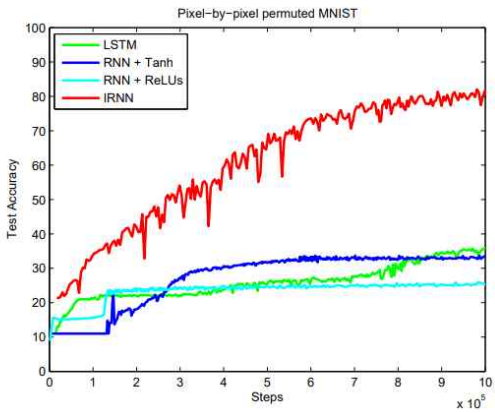
본 논문에서는 weight의 올바른 초기화를 통해 ReLU로 구성된 RNN이 상대적으로 훈련하기 쉽고 long-range dependency를 모델링하는 데 우수하다는 것을 입증한다. Recurrent weight matrix을 identity matrix로, bias를 0으로 초기화한다. 이는 이전의 hidden 벡터를 복사하고 현재 입력의 효과를 더한 후 모든 음의 상태를 0으로 대체함으로써 각각의 새로운 hidden state 벡터를 얻을 수 있음을 의미한다.

입력이 없는 경우, ReLU로 구성되고 identity matrix로 초기화된 RNN(IRNN)은 항상 동일한 상태를 유지한다. identity initialization은 extra error-derivatives가 추가되지 않는다면 hidden unit에 대한 error-derivatives가 시간 경과에 따라 backpropagation될 때 일정하게 유지되는 매우 바람직한 속성을 가지고 있다. 이는 망각 게이트의 소멸이 없도록 설정했을 때 LSTM과 동일한 동작으로 long-range temporal dependency를 쉽게 학습할 수 있다.

또한 더 적은 long-range dependency를 나타내는 작업의 경우, identity matrix를 작은 스칼라로 확장한다면 장거리 효과를 잊는 효과적인 메커니즘이 될 것이다. 이것은 메모리가 빠르게 소멸하도록 망각 게이트를 설정한 LTSM과 같은 동작이다.

그러나 ReLU의 도함수가 0보다 큰 모든 값은 1과 같기 때문에 Identity RNN 접근법은 vanishing gradient 경우에만 작동한다.[1)]





1) Machine Learning TV, Recurrent Neural Networks (RNNs) and Vanishing Gradients, from https://www.youtube.com/watch?v=NgxMUHTJYmU