Language Modeling with Gated Convolutional Networks

다음 중국어 블로그 내용을 정리하여 간략하게 번역한 내용입니다
오역 및 의역이 있을 수 있습니다.
https://zhuanlan.zhihu.com/p/24780258

관련 내용 문의 : 송영숙

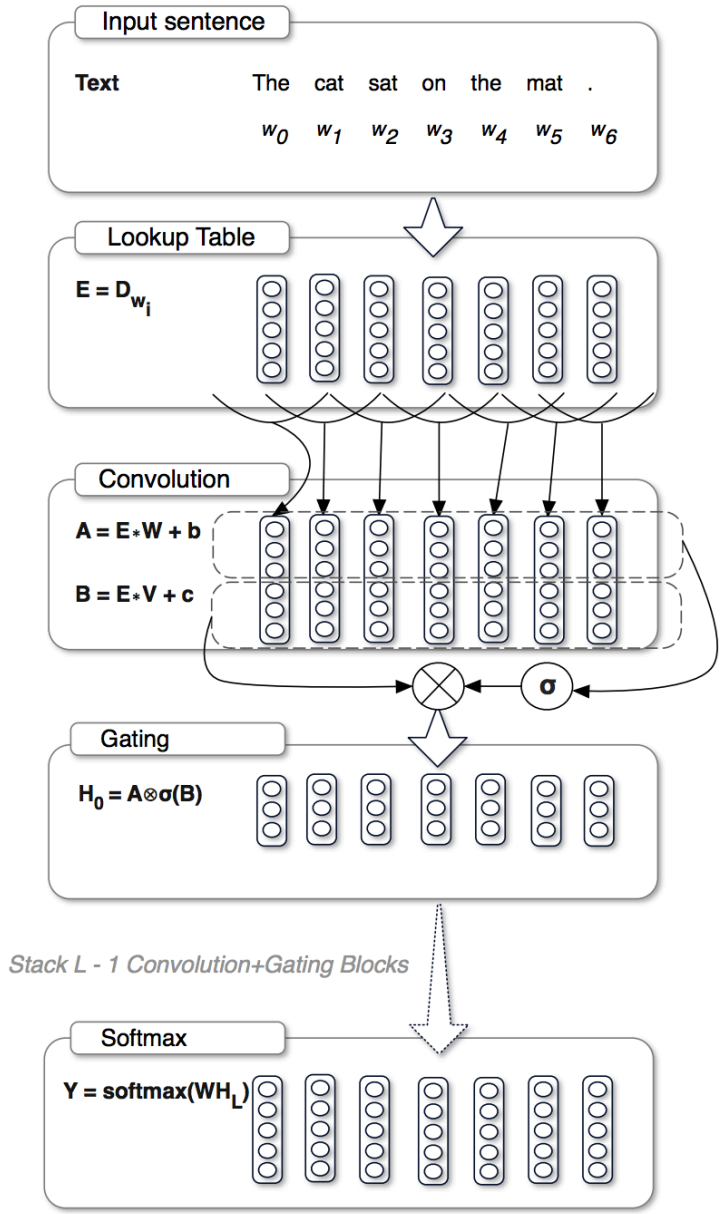
현재 언어 모델은 주로 RNN을 기반으로합니다. 본 논문에서는 LSTM의 임계 메커니즘을 시뮬레이션하고 다층 CNN 구조를 사용하는 새로운 언어 모델을 제안합니다. 각 CNN 레이어는 출력 임계값을 추가합니다.

제시된 GLU 모델은 두 가지 공통 데이터 세트에서 테스트되었으며 현제 사이클 모델보다 빠릅니다.

통계 언어 모델은 조건부 확률에 따라이 시퀀스의 분포를 추정합니다. 반복적 신경 회로망은 다음과 같이 전달됩니다.

$$P(w_0, \ldots, w_N) = P(w_0) \prod_{i=1}^{N} P(w_i | w_0, \ldots, w_{i-1}),$$

H = [h0, ..., hn]은 각각의 단어의 벡터 표현이며, hi = f (hi-1, wi-1) 단어 시퀀스 간의 종속성을 모델링합니다. 그라디언트가 사라지는 문제를 완화하기 위해 LSTM은 입력 게이트, 망각 게이트 및 출력 게이트와 같은 임계 메커니즘을 도입합니다. 순환 신경 네트워크의 각 순간의 상태는 입력과 관련가 있을 뿐만 아니라 이전 순간의 상태와도 관련되기 때문입니다. 따라서 시퀀스에서 병렬 처리 할 수 없습니다. 본 논문에서는 임계 모델 메커니즘의 컨볼 루션 (convolution) 모델인 GTU와 GLU의 두 모델을 논문에서 제안하고 실험 단계에서 비교합니다. 두 모델은 전반적으로 유사하지만 주로 활성화 기능에서 차이가 있습니다. 그림 1은 모델의 구조를 보여줍니다.



GLU模型: $h_l(X) = (X * W + b) \otimes \delta(X * V + c)$

GTU模型: $h_l(X) = tanh(X * W + b) \otimes \delta(X * V + c)$

두 모델의 차이는. GLU는 거의 선형 함수를 가지며 GTU 활성화 함수는 tanh이며 비선형입니다. 저자는 경사도에서 GTU보다 GLU를 나중에 분석했습니다.

공식의 X는 이전 층의 출력 벡터 (또는 초기 입력 단어 시퀀스 벡터)이고,는 다음과 같이 표현 될 수있습니다 :

$$X \epsilon R^{N \times m} , W \epsilon R^{k \times m \times n} , V \epsilon R^{k \times m \times n}$$

N은 워드 시퀀스의 길이, m은 워드 벡터의 차원, k는 컨볼루션 커널의 크기, b 및 c는 오프셋입니다. 각 층의 각 단어의 최종 출력은 $H = h_L \circ \ldots \circ h_0(E)$

(E)는 입력이고 L은 모델의 층 수를 나타냅니다.임계 값 메커니즘에서 GLU 및 GTU 모델을 차이가 있습니다.

LSTM은 임계 메커니즘을 도입하여 gradient 소실 문제를 완화합니다.이 논문에서는 임계 게이트 메커니즘을 도입하기 위해 출력 게이트가 도입되었습니다. GTU 모델 그래디언트 :

$$\nabla[\tanh(\mathbf{X}) \otimes \sigma(\mathbf{X})] = \tanh'(\mathbf{X})\nabla\mathbf{X} \otimes \sigma(\mathbf{X}) + \sigma'(\mathbf{X})\nabla\mathbf{X} \otimes \tanh(\mathbf{X}).$$

gradient 가 추가 된 두 부분은 tanh′ (X) 및 δ' 감쇠항 및 GLU 모델의 경사도를 가집니다.

$$\nabla[\mathbf{X} \otimes \sigma(\mathbf{X})] = \nabla\mathbf{X} \otimes \sigma(\mathbf{X}) + \mathbf{X} \otimes \sigma'(\mathbf{X})\nabla\mathbf{X}$$

첫 번째 항목에는 감쇠가 없습니다. 이 견지에서 저자는 GTU보다 GLU가 우수하고 생각합니다.

1. 데이터 세트 : Google 10 억 단어와 WikiText-103

2. 훈련 : **Nesterov's momentum** 기울기 강하 방법을 기준으로 하고, adative softmax를 softmax를 기준으로 사용하였으며, 동시에 경사하강의 절단 방법을 참고로 사용하였습니다.

이를 통해서 이 논문 GCNN-13에서 제안 된 모델이 재귀 신경 네트워크를 기반으로 한 모든 이전 모델을 능가한다는 것을 보여줍니다 .GCNN-13의 13층 컨볼루션을 사용했습니다.

본 논문에서는 컨볼 루션 신경망과 문턱 메커니즘을 기반으로 한 심층 학습 모델을 제안한다.이 메커니즘은 언어 모델에 적용되고 반복적 인 신경망 모델보다 더 나은 결과를 얻는다. 동시에 컨볼루션 신경망의 국부적 특성으로 인해, 워드 시퀀스의 병렬 학습은 처리 속도를 향상시키는 동시에 임계 값 메커니즘을 도입하고 그라디언트를 느리게하며 모델의 수렴 속도를 높입니다. 여러 레이어를 겹쳐서 단어 시퀀스의 사전 및 사후 종속성을 학습하면 긴 텍스트 WikiText-103 언어 모델을 학습 할 때 좋은 결과를 얻을 수 있습니다.

원문 : https://zhuanlan.zhihu.com/p/24780258