Language Modeling with Gated Convolutional Networks

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

위의 중국어 블로그 내용을 번역한 것입니다. 오역과 의역이 있을 수 있습니다. 관련 내용 문의 : 송영숙

지금까지의 모델은 주로 RNN을 기반으로 한다.
본 논문에서는 LSTM의 임계 메커니즘을 모방하여 다층 CNN 구조를 사용하는 새로운 언어 모델을 제안하고 CNN 레이어에 출력 임계값을 추가한다.
제시된 GLU 모델은 두 가지 공통 데이터 세트에서 테스트되었으며 현제 사이클 모델보다 빠르다.

통계 언어 모델은 조건부 확률에 따라 시퀀스의 분포를 추정한다. 반복적 신경 회로망은 다음과 같이 전달된다.

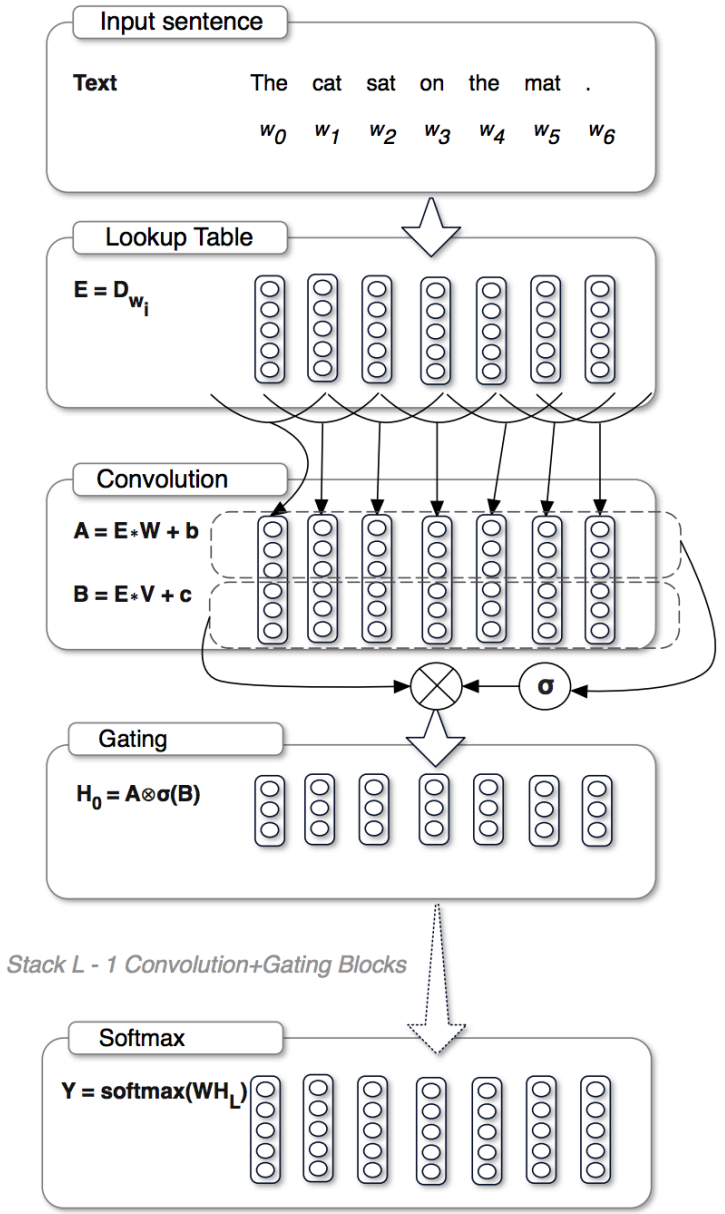
$$P(w_0, \ldots, w_N) = P(w_0) \prod_{i=1}^{N} P(w_i | w_0, \ldots, w_{i-1}),$$

H = [h0, ..., hn]은 각각의 단어의 벡터 표현이며, hi = f (hi-1, wi-1) 단어 시퀀스 간의 종속성을 모델링한다. gradient가 사라지는 문제를 완화하기 위해 LSTM의 입력 게이트, 망각 게이트 및 출력 게이트와 같은 임계 메커니즘을 도입한다.

이제까지 순환 신경 네트워크의 각 순간의 상태는 입력과 관련가 있을 뿐만 아니라 이전 순간의 상태와도 관련되기 때문에 시퀀스에서는 병렬 처리가 불가능했다.

본 논문에서는 임계 모델 메커니즘의 Convolution 모델인 GTU와 GLU의 두 모델을 제안하고 실험 단계에서 비교한다. 두 모델은 전반적으로 유사하지만 주로 활성화 기능에서 차이가 있다.

그림 1은 모델의 구조를 보여준 것이다.



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 비선형이다. 저자는 경사도에서 GTU보다 GLU를 나중에 분석했다.

공식에서 X는 이전 층의 출력 벡터 (또는 단어의 초기 입력시퀀스 벡터)이고, $X \epsilon R^{N \times m}$, $W \epsilon R^{k \times m \times n}$, $V \epsilon R^{k \times m \times n}$ 는 다음과 같이 표현 될 수 있다 :

N은 워드 시퀀스의 길이, m은 워드 벡터의 차원, k는 컨볼루션 커널의 크기, b 및 c는 오프셋이다.

각 층의 각 단어의 최종 출력은 $H = h_L \circ \ldots \circ h_0(E)$

(E)는 입력이고 L은 모델의 층 수를 나타낸다.

임계값 메커니즘에서 GLU 및 GTU 모델을 차이가 있다.

LSTM은 임계 메커니즘을 도입하여 gradient 소실 문제를 완화했다.
이 논문에서는 임계 게이트 메커니즘을 도입하기 위해 출력 게이트가 도입되었다.
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 (X) 감쇠항(衰减项) 및 GLU 모델의 경사도를 가진다.

$$\delta'$$

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

첫 번째 항목에는 감쇠항이 없다. 이 견지에서 GTU보다 GLU가 우수하고 볼 수 있다.

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

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

Model	Test PPL	Hardware
Sigmoid-RNN-2048 (Ji et al., 2015)	68.3	1 CPU
Interpolated KN 5-Gram (Chelba et al., 2013)	67.6	100 CPUs
Sparse Non-Negative Matrix LM (Shazeer et al., 2014)	52.9	-
RNN-1024 + MaxEnt 9 Gram Features (Chelba et al., 2013)	51.3	24 GPUs
LSTM-2048-512 (Jozefowicz et al., 2016)	43.7	32 GPUs
2-layer LSTM-8192-1024 (Jozefowicz et al., 2016)	30.6	32 GPUs
LSTM-2048 (Grave et al., 2016a)	43.9	1 GPU
2-layer LSTM-2048 (Grave et al., 2016a)	39.8	1 GPU
GCNN-13	38.1	1 GPU

Table 1. Results on the Google Billion Word test set.

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

Model	Test PPL
LSTM-1024 (Grave et al., 2016b)	48.7
GCNN-8	44.9

Table 2. Results on the WikiText-103 dataset.

	Throughput		Responsiveness
	(CPU)	(GPU)	(GPU)
LSTM-2048	169	45,622	2,282
GCNN-22	**179**	**45,878**	**45,878**

Table 3. Processing speed in tokens/s at test time for an LSTM with 2048 units and GCNN with 22 layers achieving 43.9 and 43.8 perplexity, respectively on Google Billion Word. The GCNN improves the responsiveness by 20 times while maintaining high throughput.

왼쪽의 표는 LSTM-2048과 GCNN의 learning 속도를 비교한 것으로 반응 속도는 단일 문장을 처리한 속도를 나타낸다.

	Parameters	FLOPs/token
LSTM-2048	289M	19M
GCNN-22	**185M**	**14M**

Table 4. Number of parameters and FLOPs for the models of Figure 3. FLOPs exclude the operations required by the softmax layer which are identical.

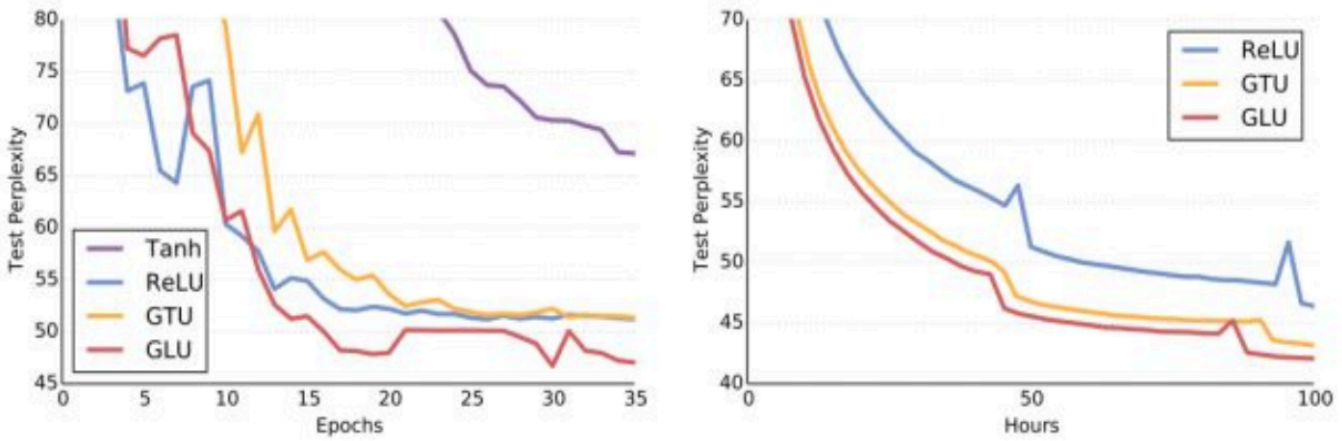
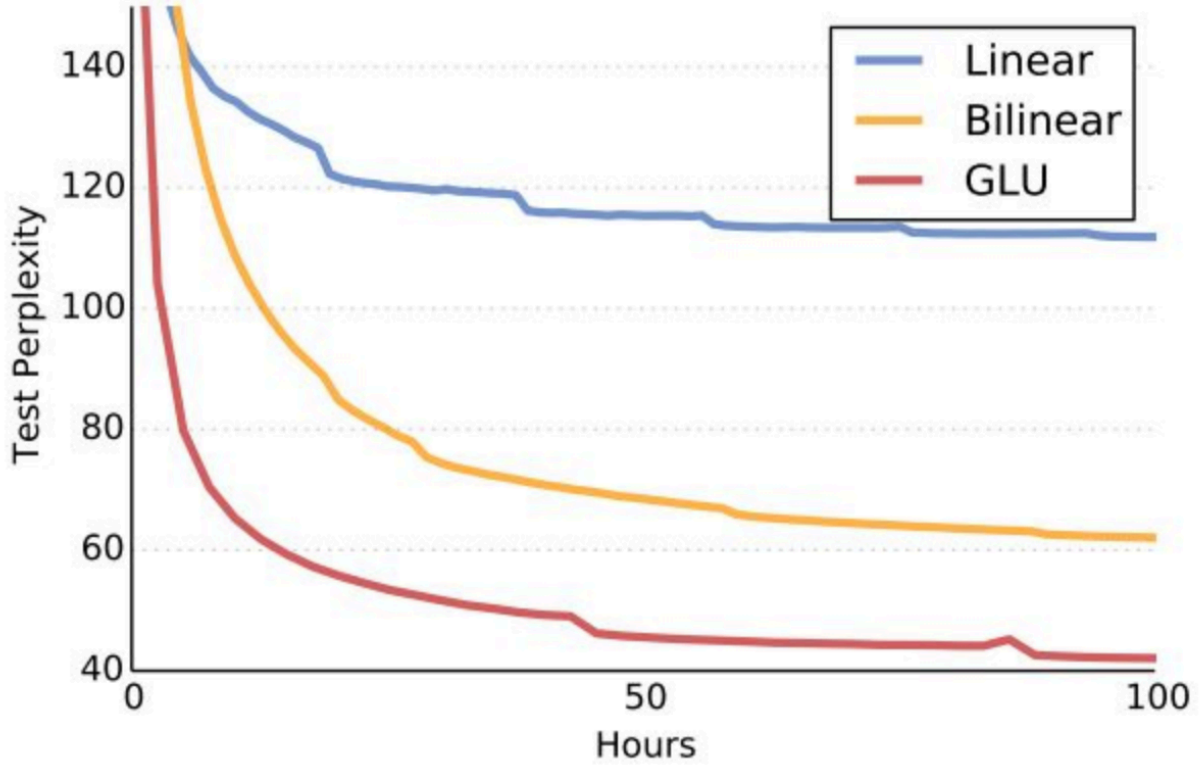


Figure 2. Learning curves on WikiText-103 (left) and Google Billion Word (right) for models with different activation mechanisms. Models with gated linear units (GLU) converge faster and to a lower perplexity.

Tanh는 GTU 출력 게이트 섹션을 제거하고 이를 GTU와 비교하여 임계효과 및 기여도를 연구 한 것을 보여준다.
실험 결과를 비교해 보면 GTU가 가장 좋은 결과를 얻은 것을 알 수 있다.



Linear는 GLU의 출력 게이트가 제거 된 후의 모델을 나타내고, Bilinear는 GLU의 출력 게이트 부분을 다른 선형 부분으로 대체하는 모델을 나타낸다.

이 세 가지를 비교하면 비선형 임계 값에 매핑한 결과가 다른 실험 결과에 비해 상대적으로 기여도가 크다는 것을 알 수 있다.

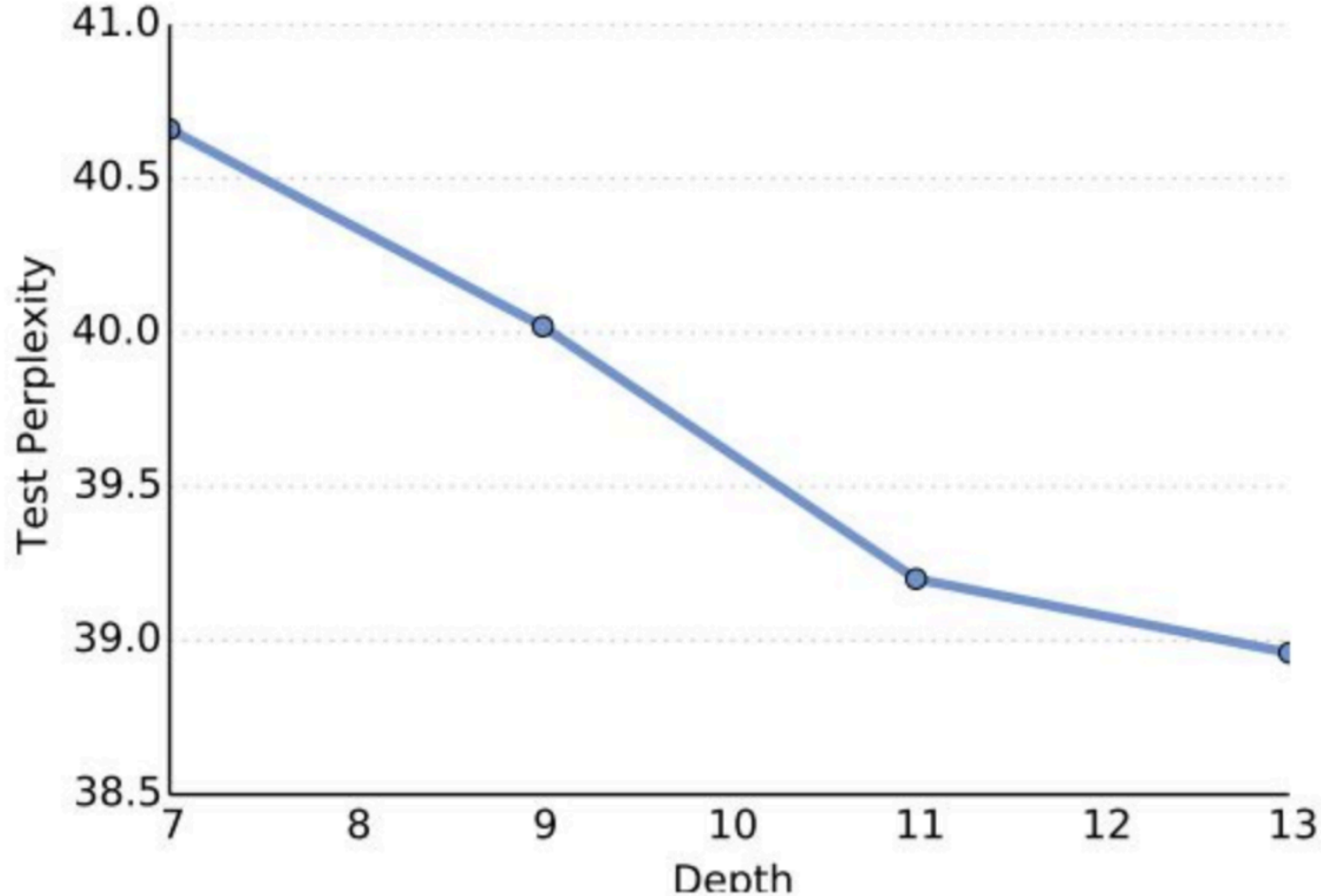


Figure 4. Impact of network depth on test perplexity for Google Billion Word. Deeper models perform better.

본 논문에서는 컨볼 루션 신경망과 문턱 메커니즘을 기반으로 한 심층 학습 모델을 제안한다.
이 메커니즘을 언어 모델에 적용했을 때 반복적인 신경망 모델보다 더 나은 결과를 얻었다.

동시에 컨볼루션 신경망의 지역적 특성으로 인해, 워드 시퀀스의 병렬 학습 처리 속도를 향상시킬 수 있었다

또한 임계값 메커니즘을 도입하고 gradient를 느리게 하며 모델의 수렴 속도를 높였다.

여러 레이어를 겹쳐서 단어 시퀀스의 사전 및 사후 종속성을 학습해서 긴 텍스트인 WikiText-103 언어 모델을 학습 할 때도 좋은 결과를 얻을 수 있었다.