메모리를 적용한 Transformer에 관한 연구

김영태, 윤상석
부경대학교 지능로봇공학과
youngtae1216@pukyong.ac.kr, ssyun@pknu.ac.kr

A Study on the Memory assisted Transformer Model

Youngtae Kim, Sangseok Yun
Pukyong National University

요 약

트랜스포머 모델은 다양한 분야에 걸쳐 강력한 성능을 보이고 있으며, 특히 자연어 처리 분야에서 우수한 성능을 보이고 있다. 하지만 기존의 트랜스포머 기반 자연어 처리 관련 모델들은 성능 향상을 위해 많은 수의 모델 파라미터 증가 혹은 높은 연산 복잡도를 요구한다. 본 논문에서는 트랜스포머 기반 언어 번역 모델의 성능 향상을 위해 어휘 지식을 제공하는 memory를 도입한다. 제안한 기법을 통해 비교적 적은 수의 파라미터 증가 및 낮은 연산 복잡도 증가만으로도 기존 트랜스포머 모델 대비 향상된 성능을 획득하는 것을 실험적으로 검증하였다.

Ⅰ. 서 론

언어 데이터에서는 텍스트의 위치 또한 중요한 정보를 가지므로, 자연어 처리 모델은 일반적으로 시계열 모델에 적합하다고 알려진 RNN (recurrent neural network)을 이용하여 훈련 되어왔다. 초기 RNN 모델의 경우 Vanishing Gradient 문제를 겪었지만, 이를 개선한 LSTM (long short-term memorty)과 GRU (gated recurrent unit) 등의 후속 모델을 통해 인공지능의 자연어 처리에 관한 능력이 점차 향상되었다. 최근 트랜스포머[1] 모델이 제안된 후로 전보다 더욱 빠르게 인공지능의 자연어 처리 능력이 성장하고 있으며, 특히, 트랜스포머의 디코더 부분을 활용해 생성한 GPT (generative pre-trained transformer)[2]는 인간과 유사한 수준의 언어능력을 선보이고 있다.

하지만 일반적으로 성능이 우수한 언어모델들은 많은 수의 파라미터를 필수적으로 요구한다. 예를 들어 T5-11B[3] 모델은 1110M의 파라미터를 사용하며, 이러한 신경망을 훈련시키거나 이러한 신경망을 활용해 추론을 수행하기 위해서는 고성능 하드웨어와 막대한 에너지 및 비용이 소요된다. 또한, 현재는 과거 대비 모델의 파라미터 수 증가에 따른 성능 향상의 효율이 저하되어 단순히 모델의 파라미터 수를 증가시키는 것이 아닌 새로운 구조의 모델 혹은 새로운 학습 기법이 요구되고 있다.

최근 트랜스포머 번역 모델의 인코더 앞단에 BERT (bidirectional encoder representations from transformer)[4] 모델을 배치하여 입력에 대한 contextual embedding을 획득하고, 이를 트랜스포머의 인코더에 전달하는 기법이 제안되었으며, 기존 대비 향상된 성능을 획득하였다. 하지만 BERT 기반 모델의 경우, BERT 모델의 크기만큼 전체 자연어 처리 모델의 파라미터 수가 증가하며, 또한 BERT를 학습시키기 위해 거대한 규모의 데이터셋을 활용해 pre-training을 수행해야 하므로 연산에 대한 부하가 심화된다.

본 논문에서는 이러한 문제를 해결하기 위해, 입력 언어에 대응되는 출력 언어의 aligned word vector로 구성한 memory를 트랜스포머의 입력으로 함께 사용함으로써 모델 파라미터의 추가를 극소화하면서 성능을 향상시킬 수 있는 memory assisted transformer 구조를 제안한다. 제안 기법을 적용하는 경우 비교적 적은 수의 파라미터 및 연산량 증가만으로도 기존 트랜스포머 대비 향상된 성능을 획득할 수 있음을 모의실험을 통해 검증하였다.

Ⅱ. 제안 기법

2.1. 기존 트랜스포머 모델

트랜스포머 모델은 attention 매커니즘이 적용된 인코더와 디코더로 구성된다. 먼저 인코더에서는 번역할 언어의 문장인 source에 대한 self-attention 연산 후 FFN (feed forward network)을 통해 feature extraction을 수행한다. 디코더에서는 번역될 언어의 문장인 target에 대한 masked self-attention을 수행하고, 이후 인코더의 출력과 cross-attention을 수행한 후 FFN 을 통해 feature extraction을 수행한다. Attention 연산은 입력의 선형변환 된 행렬로 이루어진 Q (query), K (key), V (value)의 요소들에 대해 아래 수식과 같이 계산될 수 있다.

$$\mathrm{Attention}(Q,K,V)=\mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

여기서 d_k는 Q와 K 벡터의 차원을 의미한다. Self-attention에서는 Q, K, V 모두 source embedding vector를 사용하며, cross-attention의 경우 Q는 target embedding vector를, K, V는 인코더의 출력 feature vector를 사용한다.

2.2. Aligned word vector

자연어 처리를 위해서는 언어를 구성하는 각 단어를 고차원의 공간에 사상시키는 것, 즉 embedding이 필수적이며, 따라서 이러한 embedding을 이용하여 word 또는 subword 단위의 번역을 수행하는 연구가 활발히 진행되었다. 특히, 병렬 데이터를 이용하여 squared loss를 계산해 언어 간의 선형변환 관계를 찾고, 추론 단계에서 입력 단어에 대응되는 상대 언어의 embedding 벡터 중 코사인 유사도가

가장 큰 단어로 번역하는 기법이 [5]에서 제안되었다. 이 때, 입력 단어에 대응되어 선택된 상대 언어의 embedding vector를 aligned word vector라고 한다. 이후 [6]에서는 추론 과정의 손실 함수를 CSLS (Cross-Domain Similarity Local Scaling)로 대체하는 것을 제안하였으며, [7]에서는 추론 과정의 손실 함수 뿐만 아니라 훈련 과정의 embedding 벡터 선택 기준을 모두 CSLS로 대체하여 기존 기법의 성능을 더욱 향상시켰다.

본 논문에서는 [7]에서 제안한 기법을 통해 aligned word vector를 생성하고 이를 트랜스포머 모델의 번역 과정에 활용함으로써 성능을 향상시키는 것을 제안한다.

2.3. 제안하는 memory assisted transformer

본 논문에서는 인코더에서 단순히 source만을 활용해 feature vector를 생성하는 기존 트랜스포머와 달리 source의 embedding vector와 직접적으로 대응되는 target 언어의 embedding vector (즉, aligned word vector)와 source 간의 attention을 통해 target 언어에 대한 어휘 지식을 전달하는 것을 제안한다. 일종의 memory를 통해 target 언어에 대한 어휘 지식을 트랜스포머에 전달함으로써 높은 번역 성능을 획득하기 위해 다음과 같은 memory assisted transformer 구조를 제안하였다.

제안하는 Memory assisted transformer의 구조는 기본적으로 기존 트랜스포머 모델의 구조와 동일하며, 따라서 다수의 인코더/디코더 레이어를 적층한 모델을 고려하였다. 단, 제안 기법의 경우 기존 transformer 모델과 달리 source를 입력하는 첫 번째 인코더 레이어에서 self-attention 대신 memory 기반 attention을 적용한다. 제안하는 memory 기반 attention의 경우 기존 attention 연산에서 Q로는 source를, K, V로는 memory를 사용한다. 여기서 memory는 2.2장에서 서술한 것과 같이 source를 구성하는 word 또는 subword의 각 embedding vector에 대응되는 target 언어의 aligned word vector이다. 첫 번째 레이어의 출력에 대해 나머지 인코더 레이어에서는 기존과 동일하게 self-attention 연산을 수행하며, 마찬가지로 이후의 디코더 단은 기존의 트랜스포머와 동일한 구조를 가진다.

Ⅲ. 모의실험

본 논문의 모의실험에 사용된 트랜스포머 모델의 구조는 다음과 같다. 먼저 512차원의 word embedding을 활용했으며, attention head의 개수는 4개로 설정하였다. 또한, 인코더 단과 디코더 단은 각각 인코더 레이어와 디코더 레이어가 6개씩 적층된 형태를 고려했으며, 인코더와 디코더에서는 1개의 은닉 계층을 가지고 은닉 계층의 노드 수가 1024개인 FFN을 활용하였다. 모의실험을 위한 데이터셋으로는 프랑스어에 대응되는 영어 번역문 쌍으로 구성된 IWSLT17 데이터셋을 활용하였으며, aligned word vector를 구성하기 위해 자연어 처리 임베딩을 위한 fasttext에서 제공하는 aligned word vectors dataset을 이용하였다.

성능 비교를 위해 제안하는 memory assisted transformer 모델과 기존 트랜스포머 모델의 성능을 함께 평가하였으며, 실험 결과를 그림 1에 도시하였다. 그림의 x축은 training epoch을 의미하며 y축은 6-point 이동평균을 취한 BLEU (bilingual evaluation understudy) Score를 나타낸다. 파란색 실선은 기존 트랜스포머 모델의 성능을 나타내며, 주황색 실선은 제안하는 memory assisted transformer의 성능을 나타낸다. 두 모델 모두 70 epoch의 훈련을 수행하였으며, 그림 1에서 확인할 수 있듯이 제안하는 memory assisted transformer가 기존 트랜스포머 모델 대비 높은 성능을 획득하였다. 또한, 모델의 수렴 속도 또한 기존 트랜스포머 대비 향상된 것을 그림 1에서 확인할 수 있다.

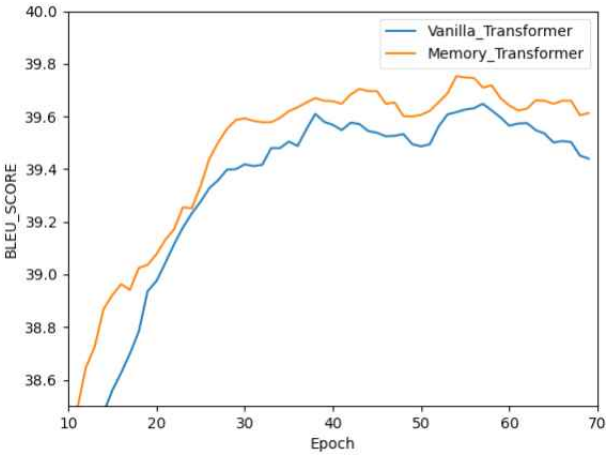

그림 1 기존 기법과 제안 기법의 성능 비교

Ⅳ. 결론

본 논문에서는 aligned word vector로 구성된 메모리를 적용한 memory assisted transformer 구조를 제안하였다. 제안하는 memory assisted transformer 모델이 기존 트랜스포머 모델 대비 향상된 성능을 획득하는 것을 확인할 수 있었으며, 추후 어휘 지식을 전달하는 더욱 효율적인 기법에 관한 연구를 수행할 계획이다.

ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2021R1G1A1094982).

참 고 문 헌

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomes, L. Kaiser, and I. Polosukhin, "Attention is all you need", in Proc. the 31st Neurips, 2017.

[2] A. Radford, K. Narasimhan, T. Salimans and I. Sutskever, "Improving language understanding by generative pre-training", arXiv preprint, 2018.

[3] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", The Journal of Machine Learning Research, vol. 21, no. 1, pp. 5485-5551, 2019.

[4] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", arXiv preprint, 2018.

[5] T. Mikolov, Q.. Le, I. Sutskever. "Exploiting Similarities among Languages for Machine Translation", arXiv preprint, 2013.

[6] A. Conneau, G. Lample, M. Ranzato, L. Denoyer and H. Jégou, "Word Translation without parallel data", arXiv preprint, 2017.

[7] A. Joulin, P. Bojanowski, T. Mikolov, H. Jégou and E. Grave, "Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion", arXiv preprint, 2018.