사전 훈련된 두 Autoencoder 교차 연결을 통한 번역 성능 개선

오지은, 최용석*
한양대학교, *한양대학교
jiunoh@hanyang.ac.kr, *cys@hanyang.ac.kr

Improving Machine Translation via Cross-connecting Two Autoencoders

Oh Jiun, Choi Yong Suk*
Hanyang Univ., *Hanyang Univ.

요 약

본 연구에서는 기계번역 문제를 위하여 두 개의 오토인코더(autoencoder)를 각각 단일 언어에 대하여 사전 훈련한 후, 둘을 교차 연결하여 병렬 코퍼스로 번역을 학습하는 전이 학습 모델을 개발하였다. 두 오토인코더는 각자 입력 언어와 출력 언어에 대하여 denoising autoencoder 방식으로 사전 훈련되며, 입력 언어의 인코더와 출력 언어의 디코더를 연결한 후 둘 사이에 feature-mapping layer를 추가하여 병렬 코퍼스로 미세 조정된다. 이 모델을 이용하면, 단일 언어로 사전 훈련한 모델을 여러 언어 쌍을 위하여 반복적으로 재사용할 수 있으며, 번역을 학습시킬 수 있는 언어 쌍에 제한이 없다. 실험 결과, 본 논문의 방식이 기존 Transformer 모델보다 개선된 성능으로 교차 연결을 통한 재활용의 가능성을 보였다.

Ⅰ. 서 론

최근 자연어 처리 분야에서 BERT[1]가 전이학습(transfer learning) 기법으로 큰 성능 향상을 달성한 이래, 신경망 기계 번역(Neural Machine Translation)문제를 위하여서도 사전 훈련(pre-training)과 미세 조정(fine-tuning)을 통한 전이 학습이 보편적인 방법으로 자리 잡았다. 그런데 전이 학습을 위하여 모델을 단일 언어로 사전 훈련하는 데에는 많은 비용이 드는 반면에, 이미 사전 훈련된 모델을 다른 언어 쌍 번역에 대하여 재사용하는 방법은 많지 않다.

따라서 본 논문에서는 두 개의 오토인코더(autoencoder)를 각각 단일 언어에 대하여 따로 사전 훈련한 후, 둘을 교차 연결하여 병렬 코퍼스로 번역을 학습하는 전이 학습 모델을 제안한다. 이 모델을 사용하면 단일 언어로 사전 훈련한 모델을 언어 쌍에 구애받지 않고 계속 재사용할 수 있다. 예컨대 영어와 프랑스어 간의 번역을 위하여 학습한 영어 모델을 다시 영어와 독일어 간의 번역을 학습하는 데에 사용할 수 있다. 또한 이 방법은 처음부터 따로 학습된 모델을 사용하기 때문에 기존의 Transformer[2] 모델과 달리 단어 사전과 임베딩이 분리되어 있어 번역을 학습할 수 있는 언어 쌍에 제한이 없고 어떤 조합으로든 학습이 가능하다.

본 논문에서 사용된 오토인코더 모델은 Transformer이며, 데이터셋은 WMT 2014 english-french[3]이다. 성능 평가 지표는 BLEU 점수이고 측정을 위하여 tensor2tensor 스크립트[4]를 사용하였다. 실험 결과, 본 논문의 모델이 기존 Transformer 모델보다 우수한 성능을 보여 사전 학습의 효과와 사전 학습된 모델의 교차 재활용 가능성을 입증하였다.

Ⅱ. 본론

1. 사전 학습

두 모델은 단일 코퍼스를 이용하여 각각의 단일 언어에 대하여 denoising autoencoder[5]로 학습된다. 예컨대 영어 오토인코더는 오염된 영어 입력을 받아 오염 없는 영어 출력을 내고, 프랑스어 오토인코더는 오염된 프랑스어 입력을 받아 오염 없는 프랑스어 출력을 내도록 학습된다. 데이터에 noise를 주기 위하여 BERT의 Masked Language Model (MLM) 방법이 사용되었다. masking의 비율은 BERT와 동일하다. 전체 데이터 중 15% 토큰이 오염되는데, 그 15% 가운데 80% 토큰은 [MASK] 토큰으로, 10%는 임의의 토큰으로 대체하며, 10%는 원래의 토큰 그대로 둔다. 각 토큰은 sub-word 단위로 분리되며, 한 단어를 이루는 sub-word들 전체가 masking되도록 하였다. 예를 들어 한 단어 'student'가 'stud', '##ent'로 분리될 때, 둘 모두 [MASK] 또는 임의의 토큰으로 대치되도록 하였다. BERT와 달리, 디코더를 함께 학습하기 위하여 모델이 오염된 토큰뿐 아니라 원래의 입력 문장 전체를 재구성하도록 하였다. 사전 훈련 모델의 구조는 그림 1과 같다.

2. 미세 조정

미세 조정 단계에서 모델은 우선 사전 훈련된 가중치로 초기화된다. 이 때, 인코더는 입력 언어로 훈련된 모델의 인코더로 초기화되고, 디코더는 출력 언어로 훈련된 모델의 디코더로 초기화된다. 단 디코더 중 encoder-decoder attention 부분은 임의로 초기화되는데, 이 부분이 monolingual로만 학습되었던 모델이 cross-lingual 문제인 번역을 수행하기 위하여 새로 학습되어야 하는 부분이기 때문이다.

그런데 인코더와 디코더는 단일 언어 데이터로만 사전 훈련되었으므로, 단순히 교차 연결만으로는 번역을 제대로 수행하기 힘들 수 있다. 그렇기 때문에 다른 언어로 학습된 인코더와 디코더가 맞물리도록 둘 사이에 feature-mapping layer를 추가하였다. 이 레이어는 self-attention과 feed-forward network로 이루어져 있으며 dropout, residual connection, layer normalization이 적용되었다. 구조적으로 인코더 내부의 한 sublayer와 동일하며, 임의로 초기화된다. 이 상태에서 프랑스어 문장을 영어 문장으로 번역하도록 병렬 코퍼스를 이용하여 번역을 학습하였다.

Ⅲ. 실험 및 결과

실험에서 사용된 오토인코더 모델은 Transformer이다. 모델의 구현은 tensorflow tutorial[6]을 사용하였다. 컴퓨터 자원의 한계로 모델 크기를

Transformer-Base보다 축소하였다. 인코더와 디코더의 레이어 개수는 각 4개로 도합 8개이다. 임베딩 차원은 128, feed-forward filter size는 512이다. attention head는 8개가 사용되었다. dropout 비율은 0.1이다. 사용된 데이터는 WMT 2014 english-french 데이터셋으로 evaluation set은 newstest2013, test set은 newstest2014이다. 이때 사전 훈련에는 원 학습 데이터셋의 1/4, 미세 조정과 baseline에는 1/8을 사용하여 학습하였다. 단어 사전의 크기는 각 32K이며 sub-word encoding이 적용되었다. batch size는 64이고 토큰이 64개 이상인 문장은 제거하였다. 최적화 기법으로 Adam을 사용하였으며, learning rate schedule과 warmup step은 Transformer 원 논문의 수식을 그대로 적용하였다. β_1=0.9, β_2=0.999, ϵ=1e-6이다.

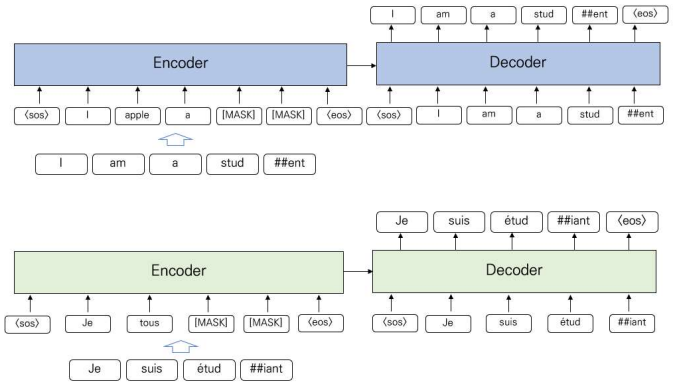

그림 1 사전 훈련 모델

위는 영어, 아래는 프랑스어 모델이다. <sos>는 문장의 시작을, <eos>는 문장의 끝을 나타낸다. 모델은 MLM 방식으로 noise가 있는 입력으로부터 원래의 입력을 복원하도록 학습된다.

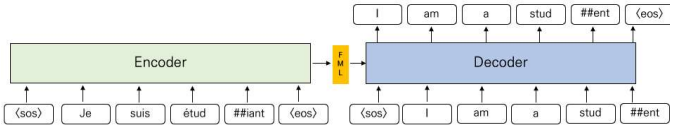

그림 2 미세 조정 모델

FML은 feature-mapping layer를 의미한다. 모델은 프랑스어 입력으로부터 번역된 영어 문장을 생성하도록 학습된다.

baseline은 사전 훈련을 하지 않고 임의로 초기화한 Transformer 모델이다. 미세 조정을 위한 1/8 데이터셋을 사용하였으며 모델의 크기, 최적화 기법, learning rate schedule은 위와 동일하다. Adam의 hyperparameter와 warmup step 모두 원 Transformer 논문과 같다. 또한 feature-mapping layer의 효과를 확인하기 위하여 feature-mapping layer 없이 미세 조정한 모델을 같이 비교하였다. feature-mapping layer 유무를 제외한 모든 데이터셋, 초기화 가중치, hyperparameter는 feature-mapping layer가 있는 모델과 같다.

실험 결과는 표 1에 나타나 있다. Transformer는 사전 훈련하지 않은 모델이며, ours (w/o FML)는 사전 훈련하고 feature-mapping layer가 없는 모델, ours (w FML)는 사전 훈련하고 feature-mapping layer가 추가된 모델이다. ours (w FML)의 결과는 임의로 초기화한 baseline보다 BLEU 점수가 확연히 높다. ours (w/o FML)의 결과는 baseline 결과보다는 다소 우세하지만, ours (w FML)보다 점수가 낮다. 이를 통하여 완전히 별개의 언어와 단어 사전을 이용하여 사전 훈련된 두 모델의 인코더와 디코더를 사후에 조립하여 미세 조정할 수 있으며, 이러한 재활용을 통하여 성능이 향상된다는 사실을 확인할 수 있다. 또한 재활용할 때에 추가적인 feature-mapping layer가 성능 향상에 유의미한 영향을 미친다는 사실 또한 확인할 수 있다.

모델	BLEU
baseline (random initialized)	24.18
ours (w/o FML)	24.57
ours (w FML)	**25.30**

표 1 실험 결과

baseline은 사전 훈련 없이 임의로 초기화된 Transformer 모델이다. ours (w/o FML)는 사전 훈련 후 교차 연결하고 feature-mapping 없이 미세 조정한 모델이며, ours (w FML)는 feature-mappng layer를 추가하여 미세 조정한 모델이다. 사용된 지표는 BLEU 점수이다.

Ⅳ. 결론

본 논문에서는 기계번역을 위하여 사전 훈련한 모델을 여러 언어 쌍을 위하여 효율적으로 재활용할 수 있는 새로운 전이학습 방법을 고안하였다. 이 방법은 두 개의 오토인코더를 각각 단일 언어 코퍼스로 학습한 후, 입력 언어의 인코더와 출력 언어의 디코더를 교차 연결하여 병렬 코퍼스로 학습한다. 이때 학습된 인코더와 디코더 사이에 feature-mapping layer를 삽입함으로써, 사전에 서로 다른 언어를 학습한 인코더와 디코더가 맞물려 번역을 학습할 수 있도록 하였다. 그 결과로 사전 훈련 없이 임의로 초기화한 Transformer보다 BLEU 점수가 상승하였으며, 사전 훈련된 모델의 재활용을 통하여 번역 성능을 향상시킬 수 있음을 보였다.

ACKNOWLEDGMENT

이 연구는 2019년도 산업통상자원부 및 산업기술평가관리원(KEIT) 연구비 지원(과제번호:10077553)과, 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(No. 2020R1A2C1014037), 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2020-0-01373, 인공지능대학원지원(한양대학교)).

참 고 문 헌

[1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. "Bert: Pre-training of deep bidirectional transformers for language understanding." CoRR, 2018.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. "Attention is all you need." In Advances in neural information processing systems, pp. 5998-6008, 2017

[3] https://www.tensorflow.org/datasets/catalog/wmt14_translate

[4] https://github.com/tensorflow/tensor2tensor/blob/master/tensor2tensor/bin/t2t_bleu.py

[5] Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. A. "Extracting and composing robust features with denoising autoencoders." In Proceedings of the 25th international conference on Machine learning, pp. 1096-1103, 2008

[6] https://www.tensorflow.org/tutorials/text/transformer