GuessWhat 게임을 위한 주의적 목표지향 질문 생성 모델

전재현[1], 강우영[2], 한철호[2], 허유정[2], 장병탁[1,2]
[1]서울대학교 뇌과학전공, [2]서울대학교 컴퓨터공학전공
{jhjun, wykang, chhan, yjheo, btzhang}@bi.snu.ac.kr

Goal-oriented Question Generator model using Attention for GuessWhat

Jaehyun Jun[1], Wooyoung Kang[2], Cheolho Han[2], Yujung Heo[2], Byoung-Tak Zhang[1,2]
[1]Interdisciplinary Program in Neuroscience, Seoul National University
[2]Department of Computer Science and Engineering, Seoul National University

요 약

CVPR 2017에서 GuessWhat 이라는 게임이 데이터세트와 함께 제안되었다. 이 게임은 스무고개 형식으로 주어진 이미지 속에 정답 물체를 찾는 게임이다. 이 게임의 진행에 있어 적절한 질문을 생성하는 것이 핵심적인 요소이다. 그래서 우리는 GuessWhat 게임의 성능 향상 위해 attention을 사용하여 목표지향적 질문을 생성할 수 있는 모델을 제안한다. GuessWhat 게임에서 질문 생성은 정답을 찾기 위해 목표지향적인 질문을 생성해야 한다. 이미지 안에서 정답 물체를 찾기 위해 후보 물체들의 개수를 추려갈 수 있는 방향으로 생성이 되어야 하는 것이 목표이다. 목표 지향적인 질문 생성을 위해 이미지 정보를 가지고 attention을 주도록 학습하는 방식의 질문 생성 모델을 고안하여 실험해 보았다. 추가된 attention 모듈은 이미지 후보군을 줄이는 방향으로 형성된 질문 데이터셋을 잘 학습 할 수 있었고, 그 결과 attention이 없는 모델에 비해 목표지향적 학습이 잘 이루어 졌으며 그에 따른 성능이 향상되었다. 이 모델을 이용 혹은 확장하여 더 나은 GuessWhat 게임을 수행 할 수 있을 것이다.

1. 서 론

본 논문에서는 목표 지향적인 질문생성 문제를 다룬다. 단순히 문장 정보만 가지고 다음 질문을 생성하는 것이 아니라, 이미지 속에서 정답 물체를 찾는 목표를 달성하기 위한 질문을 생성하는 것이다. 목표지향 문장생성 문제는 자연어처리 분야에서 지속적으로 대두되고 있는 문제이다. 이 분야를 GuessWhat[1] 이라는 특정 문제를 통해 다루고자 한다.

GuessWhat 게임은 Guesser와 Oracle이라고 불리는 두 AI agent가 풀어가는 문제이다. 하나의 이미지가 주어지면 이미지가 담고 있는 하나의 정답 물체에 대해서 Guesser가 답을 찾을 때까지 질문을 하고, Oracle은 그 질문에 답변을 하게 된다. 최종적으로 질문과 답변 세트를 통해서 Guesser가 어떤 물체가 정답인지 추론하는 게임이다. Guesser는 여러 개의 후보 물체중에 정답을 찾을 수 있도록 후보군을 줄이는 방향으로 질문을 생성해야 한다. 이 문제의 성능을 향상하기 위해서는 목표 지향적인 질문을 생성하는 것이 가장 핵심이라고 볼 수 있으며, 본 논문에서는 이 문제를 다룬다.

딥러닝에서 질문생성 문제는 다양한 시도가 이루어지고있는 분야이다. 그 대표적인 예로 Sequence-to-Sequence (Seq2Seq)[2] 라는 모델이 있다. 이 모델은 Recurrent Neural Networks (RNN) 의 Long Short-term Memory (LSTM)[3] 를 사용하여 입력 문장을 encoding 하고 encoding 된 feature를 다시 decoding 하여 원하는 질문을 생성하도록 학습하는 모델이다. 이 모델은 비교적 간단하지만 이를 기반으로 확장하기 좋아 많은 파생 모델이 제안되어 왔다. Seq2Seq 모델은 encoder와 decoder의 LSTM 층으로 두개의 층을 두었지만, GuessWhat에서 기본모델로 제안하고 있는 모델은 LSTM 층을 여러 층으로 쌓아 Seq2Seq를 더 깊게 학습 하여 조금 더 복잡한 문제를 풀 수 있도록 한 모델이다. 이러한 모델을 확장하여 GuessWhat의 기본 모델보다 성능이 향상된 모델을 제안하고자 한다.

이미지와 질문 정보를 처리하는 접근에 많이 사용하는 방식인 이미지와 문장 feature를 외적 한 데이터를 많이 사용한다. 하지만 이 feature는 차원이 너무 크게 증가하는 측면이 있어 외적을 근사하는 방식인 두 feature를 Hadamard product 처리하여 차원을 유지하는 방식이 있다.[4] VQA에서 이 방식은 이미지 feature에 Attention을 주는 형식으로 접근을 했으며 GuessWhat 의 질문생성 모델에서도 이 방식을

적용한 모델을 제안한다.

질문생성 모델을 포함한 문장 생성 모델은 성능을 평가하기 위한 지표가 명확하지 않다. 본 논문은 질문 데이터와의 유사도를 BLEU score[5] 와 Perplexity[6]로 측정하여 Seq2Seq 모델과 비교해 보았다. 이 지표는 얼마나 후보군을 추리는 방향의 문장생성 형태의 테스트 데이터로 얼마나 목표지향적으로 다음 문장 생성하도록 유도했는지를 나타내게 된다.

2. 질문 생성 모델

2.1 기본 모델

GuessWhat[1] 논문에서는 질문과 이미지 데이터를 이용한 기본 모델을 제안했다. 기본적인 틀로 Hierarchical Recurrent Encoder-Decoder (HRED)[7] 모델을 질문 생성 모델로 채택했다. 이미지 데이터는 VGG16 network[8] 를 이용해 이미지 feature를 추출했다. 질문 생성 모델의 encoder 부분에서 질문 데이터의 feature를 뽑아 낸 것을 이미지 feature와 concatenation으로 이어 붙이고 이것을 decoder의 input으로 사용하였다. 이 기본 모델은 질문과 이미지 데이터를 이용해 질문 생성하여 인간이 답변했을 때 38.7%의 오차를 낸다고 밝혔다. 이 수치는 아직 개선될 여지가 많이 남아있음을 말해준다.

기본 모델은 이미지 feature를 사용했지만, 질문과 답변 대화를 통해 추려낸 후보군의 정보를 전혀 반영하지 않고 있다. 만약 후보군을 추려낼 수 있는 정보를 반영한다면 질문 생성 모델이 개선된 성능을 낼 수 있을 것으로 기대한다.

2.2 Attention을 반영한 모델

제안하는 모델은 이미지 feature를 이용해서 Attention을 주기 위해 Hadamard product 를 사용한다. [그림 1]은 본 논문에서 제안하는 Image feature를 이용해 질문 feature에 Attention을 주는 모델이다. 먼저 이미지 feature는 pretrained VGG16의 마지막 층에서 추출한 비선형 함수를 거치치 않은 feature를 사용했다. 질문 feature는 Seq2Seq의 encoder 부분을 거쳐서 나온 마지막 결과값을 질문의 feature로 사용했고, 이미지와 문장 feature를 각 fully connected layer를 거쳐 dimension을 맞춰주었다. Image feature는 attention 역할을 할 것이기 때문에 hyperbolic tangent 비선형 함수를 거치게 된다. 두 feature를 Hadamard product 연산을 거쳐 새로운 Attended feature를 얻게 된다. 이 feature를 Seq2Seq의 decoder 부분에 넣어 주게 되고 이미지 feature는 위와 같은 방법으로 decoder의 모든 input과 Hadamard product를 거치게 된다.

이 모델은 이미지 feature를 통해 문장 feature에 attention을 주는 효과로 작용하는 것으로 기대하고, 기본적인 모델과 단순히 feature를 합치는 concatenation 방식보다 나은 성능으로 후보군을 추려낼 수 있는 방식의 다음 질문을 생성하는 것을 다음 실험을 통해서 확인 해 보았다.

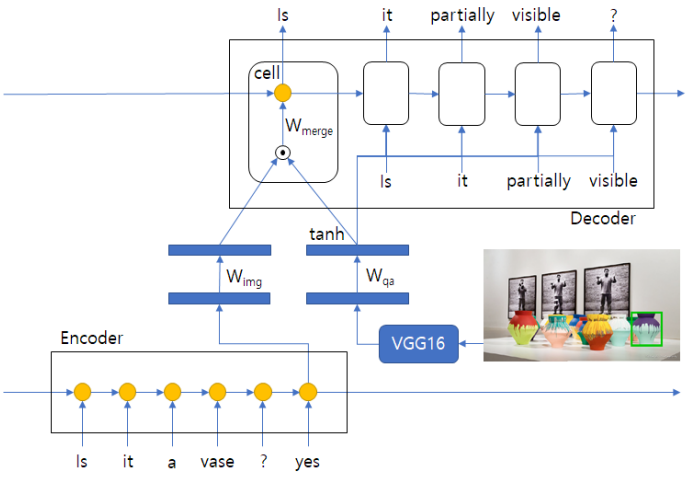

그림 1. 주의적 목표지향 질문 생성 모델

3. 실 험

3.1 데이터 및 실험

실험에는 GuessWhat[1] 데이터를 사용했다. 이 데이터는 웹사이트[9] 에서 제공되는 GuessWhat 게임을 통해 얻어낸 데이터로 MSCOCO 이미지, 질문 답변 세트, 물체 정보 등의 135,400개 (2017년 4월 기준) 데이터를 보유 하고 있다. 본 실험은 이 중 질문 답변 세트, 이미지 데이터만 사용을 하게 된다.

실험은 생성된 질문이 얼마나 목표 지향적인 질문을 만들게 되었는지 평가되었다. 목표 지향적인 질문의 기준은 사람들이 GuessWhat 게임을 하면서 목표 지향적인 질문 데이터를 쌓은 것 기준으로 얼마나 유사한 문장을 만들었는지 BLEU score와 Perplexity 계산으로 이루어지게 된다. BLEU score는 예측한 문장과 게임을 통해 수집한 목표지향적 문장 간의 유사도를 계산하고, Perplexity는 얼마나 데이터 분포에 맞게 모델이 설계되었고 학습되었는지를 나타내는 지표이다.

3.2 실험 결과

제안된 모델에서 핵심인 Feature 처리방법의 성능을 보기 위해 문장생성모델은 Seq2Seq 모델로 고정하고 각 처리방식에 따른 성능 변화를 관찰했다. Seq2Seq의 학습 방식은 AdaDelta optimizer[10]를 사용했고, RNN

encoder와 decoder의 hidden unit의 차원은 512, word embedding 차원은 300으로 고정하여 실험하였다. Feature와 처리 방식에 따른 BLEU score와 Perplexity 결과는 [표 1]과 같다.

[표 1]의 수치는 학습이 수렴했을 때 13만개의 데이터 안에 한 epoch안에서 계산한 수치의 평균값이다. 수치에서 보면 알 수 있듯이 Hadamard

표 1. Feature 처리 방법 별 성능 비교

	Perplexity	BLEU score
Only Text	967.29	0.5699
Concatenation	866.92	0.5171
Hadarmard	366.54	0.5731

product를 사용한 제안된 모델에서 BLEU score가 가장 좋은 값을 나타내고 있다. Perplexity에도 알 수 있듯이 제안된 모델이 가장 데이터 분포와 적합하게 학습이 된 것을 확인 할 수 있다. 또, 이미지와 문장 feature를 concatenation 시킨 결과보다 문장 feature만 사용한 결과가 더 나은 결과를 낸 것으로 보아 제안된 모델이 이미지 feature를 사용함으로써 성능이 향상된 것은 아님을 알 수 있다. 이것은 역으로 이미지 feature를 잘 결합하여 문장을 생성했다고 볼 수 있다.

언어 모델을 BLEU score와 Perplexity 만으로 평가했을 때 일정한 수치를 뽑아낼 수 없다는 한계가 있지만, 평균값의 경향성을 봤을 때 GuessWhat 게임에서의 목표지향 언어생성에 향상된 결과를 얻어냈다.

4. 결 론

제안된 모델이 기존 방법과 비교했을 때 향상된 결과를 얻을 수 있었다. 성능 측정 방식에서 우수함을 보였고, 본 논문에서는 명확한 지표가 될 수 없어 보여주지 않았지만 비교한 방식에 비해 의미가 담긴 목표지향적 문장을 더 잘 생성해 낸 것을 실험을 통해 확인 할 수 있었다. 이 부분은 추후 설문조사를 통해 얼마나 문장을 더 잘 생성했는가에 대한 결과를 수치화 해볼 계획이다.

아직은 이미지와 문장 feature를 결합하는 방법에 집중적으로 적용해보고 실험해 보았지만 아직 완벽한 목표지향적 문장생성을 하기에는 보완해야할 점이 많이 남아있다. 문장생성 문제가 단순한 문제가 아니기 때문에 HRED, Mutlit-resolution RNN (MrRNN)[11] 등과 같은 더 복잡한 질문생성 모델을 적용시켜 문제를 해결해볼 여지가 있다. 그리고 문장에 혹은 이미지에 어떻게 Attention이 반영되었는지 연구하고 사람이 이해할 수 있는 수준의 표현을 얻어내는 연구를 추후 진행할 예정이다.

감사의 글

이 논문은 2017년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터(2015-0-00310-SW스타랩), 한국산업기술평가관리원(10044009-HRI.MESSI, 10060086-RISF)의 지원을 받았음.

참고 문헌

[1] de Vries, Harm, et al. "GuessWhat?! Visual object discovery through multi-modal dialogue." *arXiv preprint arXiv:1611.08481* (2016).

[2] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems. 2014.

[3] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.

[4] Jin-Hwa Kim, et al. "Hadamard product for low-rank bilinear pooling." arXiv preprint arXiv:1610.04325 (2016).

[5] Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.

[6] Brown, Peter F., et al. "An estimate of an upper bound for the entropy of English." Computational Linguistics 18.1 (1992): 31-40.

[7] Sordoni, Alessandro, et al. "A hierarchical recurrent encoder-decoder for generative context-aware query suggestion." Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. ACM, 2015.

[8] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[9] GuessWhat game and dataset, [Online]. Available web-site: https://guesswhat.ai/

[10] Zeiler, Matthew D. "ADADELTA: an adaptive learning rate method." *arXiv preprint arXiv:*1212.5701 (2012).

[11] Serban, Iulian Vlad, et al. "Multiresolution Recurrent Neural Networks: An Application to Dialogue Response Generation." *arXiv preprint arXiv:1606.00776* (2016).