이미지 맥락 임베딩을 통한 뉴럴-이야기생성기 디자인

손선일[1*] 김태형[2] 장병탁[1]
1 컴퓨터공학부 2 인지과학 협동과정
서울대학교
{sison, thkim, btzhang}@bi.snu.ac.kr

Designing Neural Storyteller by Image Context Embedding

Seonil Son[1*] Taehyeong Kim[2] Byeong-Tak Zhang[1]
1 Dept. of Computer Sci., 2 Program in Cognitive Science,
Seoul National University

요 약

사람의 사고는 언어로 표현되지만, 그 언어를 이해하게 하는 것은 언어 자체에 들어있는 정보뿐만 아니라 그 단어를 듣는 사람이 떠올리는 심상을 포함한다. 이러한 심상은 사람의 경험을 통해 얻은 통상지식과 이에 따른 맥락정보인데, 이것을 여러모로 구현해보고자 하는 시도들이 있었으나, 대체로 사람 수준에 가까운 맥락 인지를 구현하는 것은 여전히 도전적인 문제로 남아있다. 이 논문에서는 이러한 맥락정보를 시각-언어 가설에 따라 순환신경망으로 처리된 시각정보를 적절히 활용하여 사진 순열에 알맞은 이야기를 생성하도록 하는 작업을 통해 맥락정보를 잘 모델링하고자 한다.

1. 서 론

의사소통은 언어로 이루어지지만, 문자들에 모든 정보들이 들어있는 것은 아니다. 이야기를 듣는 사람이 본뜻을 이해하기 위해서는 화자와 청자가 심상을 공유해야한다[1]. 따라서 현재의 자연어처리 기술이 글에 나타난 표면적 의미를 해석하는 수준을 넘어 맥락상의 의미까지 파악하도록 하려면 이를 위해 공유하는 심상, 맥락정보를 모델링하여야 한다. 시각을 통해 언어를 모델링하고자 한 기술은 지난 10년간 눈에 띄는 성장을 보여주고 있고, 특히 인공신경망의 실용화를 따라 그 능률이 점점 오르고 있다[2]. 그러나 아직 언어의 빈 구멍을 채워주는 맥락이나 상식을 모델링하는 것은 도전적인 분야로 남아있다. 이 연구에서는 순환신경망 구조를 이용하여 시각정보를 처리하고, 신경망의 표현력(representation power)를 십분 활용하여 맥락정보를 표현해보고자 한다.

본 연구에서는 Visual Storytelling Challenge 2018에서 제공하는 데이터셋(VIST) 을 활용한다[3]. 데이터셋은 5 장의 사진 순열과 그에 걸맞은 다섯 개의 문장으로 된 이야기, 그리고 각 사진에 대한 메타정보를 제공한다. 이는 임의로 모은 다섯장의 사진에 대해 사람들이 작성한 것을 토대로 선별하여 만들어졌다. 한 사진 순열에 대해 다른 이야기가 붙기도 하고 이야기가 꼭 사진에 있는 객체 중심으로 진행되는 것도 아니므로 다양한 상황에 대한 정보를 지도학습시키기에 적절한 데이터셋이라고 생각된다.

2. 선행 연구

스토리구조를 직접 분석하고자 시도한 연구들은 거슬러 올라갈 경우 어느 정도 있지만 VIST 와 같이 시각-언어 모델링으로 이야기의 맥락을 잡아내고자 한 사례는 많지 않다. 다만 이에 준하거나 비슷한 기반으로 문제를 풀려고 접근한 몇 가지 경우가 있다. 예를 들면 이야기 흐름의 이해를 필요로 하는 ROCStory Cloze Test[4]나 문서에 내포된 정보를 얼마나 이해했는지 질의응답으로 판단하는 Machine Text Comprehension via Question Answering(SQuAD)[5], 그리고 이야기를 잘 따라가고 있는지 질의응답으로 확인해보는 bAbI Task[6]가 그것이다. 앞서 든 예시[4,5,6]이 문자기반으로 자연어의 이해를 추구했다면 현재 진행되고 있는 Visual Question Answering(VQA[7]) 의 경우는 이 연구와 비슷한 관점에서 시각정보를 활용해 사진 한 장에 담겨있는 정보에 대한 질의응답을 한다. 이 작업들은 각각 다른 제한적 조건에서 수행한 것이지만 SQuAD의 경우 Google Brain과 Carnegie Mellon University에서 작업한 QANet[8]이 현재 사람과 근접한 정답률을 기록하고 있으며 ROCStories Cloze Test의 경우도 0.7 이상의 정답률을 기록하고 있다[9]. 또한VQA의 경우 역시 종합 0.69 정도에 근접한 수치로 좋은 성능을 보여주고 있어 Visual Storytelling 작업을 실행하기에 필요한 수준의 기술에는 도달했다고 볼 수 있다[10]. 이 외에도 VideoQA 모델들은 시각정보와 언어정보를 통합하여 이야기를 이해하는 작업을 QA 형태로 시도했다고 볼 수 있는데 대표적으로 PororoQA 데이터셋[11]을 이용한 DEMN 모델이 메모리 네트워크를 이용하여 좋은 성능을 보여주고 있다[12].

3. 네트워크 디자인

3.1. 단어 임베딩

먼저 데이터셋에 등장하는 단어들의 종류를 따라 one-hot vector 형식의 사전을 만들고 이를 256차원으로 임베딩한다. 단어를 수집할 때에는 4번 이상 등장한 단어들만을 골랐고 그 미만의 빈도로 등장한 단어는 <unk>로 처리하였다. 이렇게 만들어진 사전의 크기는 11075였고, 인물의 이름이나 고유명사와 같이 일반화가 어려운 단어들은 데이터셋에서 이미 적절히 처리하여(예시: 남자인물의 이름 -> <male>) 제공되었다.

3.2. 기본 이미지 묘사 모델

먼저 맥락정보 포함하지 않고 이미지별 문장을 학습했을 때 이야기가 잘 생성되는지 확인하기 위해 **그림 1** 와 같이 기본 네트워크 구조를 구성하였다. **그림 2** 은 우리가 실험해서 성능을 확인할 모델로 압축기(encoder) 부분에 GRU[13]를 이용하여 Resnet[14]으로 압축된 시각정보를 추가적으로 처리해주고 이를 과정을 이용하여 생성기의 적절한 동작을 유도하려고 한다. Decoder의 경우 LSTM[15]을 사용하였다.

3.3. 제안하는 모델

맥락정보를 임베딩하기 위해 제안하는 모델은 다음과같은 과정을 거친다. 먼저 다섯 장의 사진은 Resnet을 거쳐 5개의 크기 차원 512의 벡터로 축약된다. 이 문장생성에 이용할 표상(representation)에 다섯 장의 정보를 모두 포함하기 위해 이 다섯 벡터를 다 더한다. 더한 벡터는 512차원의 벡터 하나가 된다. 이를 GRU 압축기(encoder)를 이용하여 t=0, 1, 2, 3, 4에 해당하는 은닉 벡터를 얻는다. t=0(즉 첫 사진)에 대한 문장

생성에는 합한 512의 벡터가 GRU를 거치지 않은 채로 문장 생성에 이용된다. 이 때, 첫 번째 사진과 어울리는 문장을 생성해야 하므로 첫 사진의 Resnet 처리된 벡터(차원 512)를 축약된 합 벡터와 성분 곱(element-wise multiplication) 으로 처리한다. 이렇게 처리된 벡터는 첫 번째 문장 생성기인 LSTM 모델에 넘겨져 Teacher Forcing 방법으로 정답인 문장의 피드백을 받아 문장 생성을 학습한다. 비슷한 방식으로 두 번째 장면에 대한 문장생성은 다섯 사진의 합 벡터가 GRU(encoder)를 한 번 거친 t=1에 해당하는 은닉 벡터를 이용하며 얻어진 은닉 벡터에 두 번째 사진의 Resnet 처리된 벡터와 성분 곱을 하여 두 번째 문장 생성기에 제공된다. 비슷한 방식으로 다섯 번째 생성기까지 반복하여 정답 문장 길이와 같은 길이의 워드 임베딩을 각각 생성하게 되고 이를 바탕으로 손실함수 값을 계산해 네트워크의 가중치를 갱신한다. 손실함수로는 문장 간의 단어 일치를 확인하여 크로스-엔트로피 손실함수를 계산한다.

4. 평가 척도

우리가 활용할 VIST 데이터셋[3]에서는 단일 사진 묘사(Single Image Captioning) 모델[16]을 이용하여 스토리를 생성할 수 있는지를 먼저 시험하여 이를 기본성능(baseline performance) 로 삼는다. 평가 방법은 기본적으로 METEOR[17]점수를 활용하지만 METEOR 점수가 사람의 더 자연스러운 결과를 대표하는지에 대한 보장이 없으므로 Visual Storytelling Challenge 2018 과 동일하게 사람 평가가 필요할 것으로 보인다. 이를 위해 제안하는 모델을 구축한 후에는 Amazon Mechanical Turk(AMT) 와 같은 시스템을 활용해볼 수 있다.

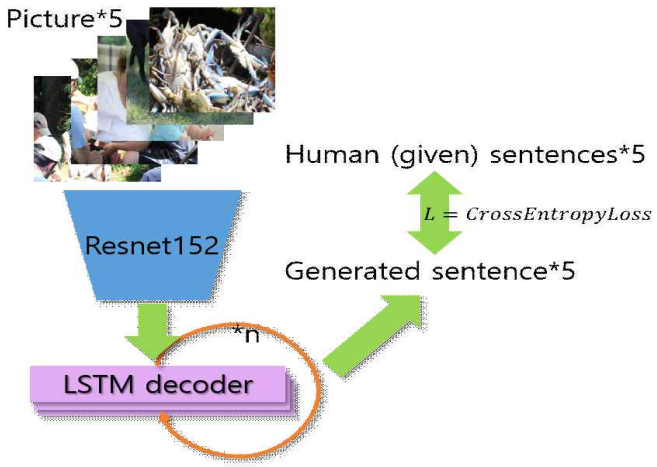


그림 1 Resnet-LSTM 이미지 묘사 모델

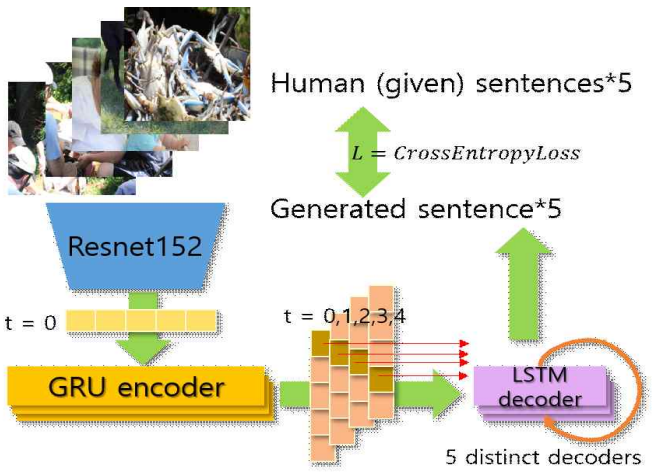


그림 2 제안하는 맥락 축약 모델

5 실험결과(선행)

제안한 모델은 아직 구현 중이고 이것을 비교하기 위한 선행 모델을 비슷한 규모의 신경망 모델로 구성하여 결과를 확인했다(**부록- 표 1).** 대체로 정답에 자주 등장하는 문장 위주로 그럴듯한 묘사를 붙이는 듯 보이지만(**표1** 사진 a) 흐름이 있는 이야기를 생성했다고 보기에는 너무 일반적인 문장들을 배열해 놓은 수준이며 사진 b를 보면 정답에서는 관찰자시점의 이야기를 서술하는 반면 장면에 등장한 객체들에 치중한 문장들이 배열되는 것을 알 수 있다. 제안하는 모델의 호평가를 위함이 아니라 사진 c를 보면 모델이 위와 같은 방식으로 문장을 생성한다는 것을 확인할 수 있다. 공개 테스트셋으로 측정한 결과 METEOR 점수는 **0.268**을 기록했다. 제안하는 모델에서는 맥락 정보, 그리고 컷에 맞는 문장을 생성하기를 기대한다.

6. 향후 계획

Soft attention mechanism을 이용한 모델의 튜닝과 함께 성공적인 잠재변수 공간을 정의하기 위해 압축기(encoder)와 생성기(decoder)를 순환신경망에서 중첩신경망(Convolution Neural Network)로 바꿔서도 시도해볼 예정이다.

7. 감사의 글

이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터(R0126-16-1072-SW스타랩, 2017-0-01772-VTT, 2018-0-00622-RMI), 한국산업기술평가관리원(10060086-RISF)의 지원을 받았다.

아낌없는 조언과 수고를 들여 퇴고를 도와주신 선배님들과 좋은 연구를 하도록 지도해주신 교수님께 감사의 말을 전합니다.

참고문헌

[1]L.B. Resnick, J.M.Levine, S.D.Teasley, Perspectives on socially shared cognition, American Psychological Association, 1991

[2]A. Belz, T.L.Berg and L.Yu, From image to language and back again, Language for Images, Sp Issue 3, Vol24, pp.325-362 2018

[3]TH.K.Huang, F.Ferraro, N.Mostafazadeh, I.Misra, A.Agrawal, J.Devlin, R.Girshick, X.He, P.Kohli, D.Batra, C. L.Zitnick, D.Parikh, L.Vanderwende, M.Galley, M.Mitchell, Visual Storytelling, Microsoft Research, NAACL 2016

[4]N.Mostafazadeh, M.Roth, A.Louis, N.Chambers, J.Allen, LSDSem 2017 Shared Task: The Story Cloze Test, Proceedings of the EACL Workshop onLinking Models of Lexical, Sentential and Discourse-level Semantics(LSDSem), 2017

[5]P.Rajpurkar, J.Zhang, K.Lopyrev, P.Liang, SQuAD: 100,000+ Questions for Machine Comprehension of Text, EMNLP 2016

[6]J.Weston, A.Bordes, S.Chopra, A.M.Rush. B.van Merrienboer, A.joulin, T.Mikolov, Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks, ICLR 2016

[7]A. Agrawal, J. Lu, SAntol, M. Mitchell, C.L.Zitnick, D. Batra, D. Parikh, VQA: Visual Question Answering, ICCV 2015

[8]A. W. Yu, D. Dohan, MT. Luong, R. Zhao, K. Chen, M. Norouzi, Q. V. Le, QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension, ICLR 2018

[9]Cloze Test Leaderboard(2018.05.04)(https://competitions.codalab.org/competitions/15333)

[10]J-H Kim, K-W On W. Lim, J. Kim & J-W Ha, BT Zhang, Hadamard Product for Low-Rank Bilinear Pooling, ICLR 2017

[11]K. Kim, C. Nan, M.-O.Heo, S.-H. Choi, and B.-T. Zhang,

PororoQA: Cartoon video series dataset for story understanding, NIPS 2016 Workshop on Large Scale Computer VisionSystem, 2016.
[12]K.-M. Kim, M.-O. Heo, S.-H. Choi, and B.-T. Zhang, DeepStory: Video story QA by deep embedded memory networks, The 26th International Joint Conference on ArtificialIntelligence (IJCAI), 2017. (Invited paper to UAI 2017 MLTrain's "Neural AbstractMachines" session)
[13]J.Chung, C.Gulcehre, K.Cho, Y.Bengio, Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, NIPS 2014
[14]K.He, X.Zhang, S.Ren, J.Sun, Deep Residual Learning for Image Recognition, CVPR 2016
[15]F.A. Gers, J. Schmidhuber, F. Cummins, Learning to forget: continual prediction with LSTM, 9th International Conference on Artificial Neural Networks: ICANN '99, 1999 p. 850 – 855
[16]Andrej Karpathy, FeiFei Li, Deep Visual-Semantic Alignments for Generating Image Descriptions, CVPR 2015
[17]S. Banerjee, A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation withHuman Judgments" , in Proceedings of Workshop on Intrinsic andExtrinsic Evaluation Measures for MT and/or Summarization 43rd Annual Meeting of theAssociation of Computational Linguistics(ACL-2005), Ann Arbor, Michigan, 2005

부록

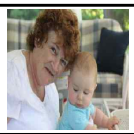
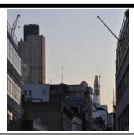
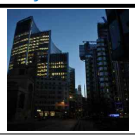
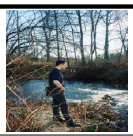
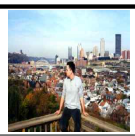
사진 a					
생성	the family was excited for the day at the park .	we had a lot of fun with them .	the family was so happy to see the family .	the dog is playing with the toys .	the food was delicious .
데이터셋	the guys sit down to a few beers at the table.	it looks like they may be up to some cards .	she spends some time with the young baby .	a black dog comes around to enjoy the day .	maybe he smelled the delicious crabs being cooked
사진 b					
생성	the church was very large and relevant .	the city was beautiful .	the view from the observatory is beautiful .	the view from the top was amazing .	the crowd was excited .
데이터셋	as the sun was setting yesterday i knew i was really far from the restaurant .	i was n't entirely sure where it was but i decided to walk around and see if i could find it .	it began to get very dark .	once it was night time it made it a lot harder to see .	i gave up trying to find the restaurant and went home instead .
사진 c					
생성	the family went to the beach .	the <unk> was dynamic .	the view was beautiful and the weather was great .	the kids were having fun too .	the next day , they all took a picture to remember the day .
데이터셋	the view in our room was beautiful .	the day after , we visited the mountains for fresh air .	we hiked up the trail to stop at this lake .	i had lots of fun on the swinging set .	this town continues to <unk> me .

표 1 사진 a: 가장 그럴듯한 결과가 보인다. 그러나 사진 b을 볼 때 묘사가 사진에 등장한 객체들에 의존적이라는 점을 발견할 수 있다. 사진 c 로 볼 때 사실 이 모델이 스토리를 생성하는 것이 아니라 각자에 그럴듯하고 일반적인 묘사를 많이 붙이는 것임을 알 수 있다. 제시할 모델은 맥락을 반영하길 기대한다.