# GuessWhat?! 문제에 대한 분석과 파훼

이상우[1°], 한철호[1], 허유정[1], 강우영[1], 전재현[2], 장병탁[1,2]
서울대학교 컴퓨터공학부[1], 서울대학교 뇌과학협동과정[2]
{slee,chhan,yjheo,ywkang,jhjun,btzhang}@bi.snu.ac.kr

# Analyzing and Solving GuessWhat?!

Sang-Woo Lee[1°] , Cheolho Han[1], Yujung Heo[1],
Wooyoung Kang[1], Jaehyun Jun[2], Byoung-Tak Zhang[1,2]
School of Computer Science & Engineering, Seoul National University[1]
Interdisciplinary Program in Neuroscience, Seoul National University[2]

**요 약**

GuessWhat?!은 질문자와 답변자로 구성된 두 플레이어가 이미지를 보고 질문자에게 비밀로 감추어진 정답 물체에 대해 예/아니오/잘 모르겠음 셋 중 하나로 묻고 답하며, 정답 물체를 추려 나가는 문제이다. GuessWhat?!은 최근 컴퓨터 비전과 인공지능 대화 시스템의 테스트베드로서 컴퓨터 비전과 인공지능 학계의 많은 관심을 받았다. 본 논문에서, 우리는 GuessWhat?! 게임 프레임워크가 가지는 특성에 대해 논의하고, GuessWhat?! 분석의 틀을 제안한다. 더 나아가, 우리는 제안된 틀을 기반으로 GuessWhat?!의 간단한 solution을 제안한다. 사람이 평균 4~5개 정도의 질문을 통하여 맞추는 이 문제에 대하여, 우리가 제안한 방법은 3개의 질문 만으로 기존 딥러닝 기반 기술의 성능을 상회하는 성능을 보이며, 5개의 질문이 허용되면 인간 수준의 성능을 능가한다.

## 1. 서론

GuessWhat?!은 질문자와 답변자로 구성된 두 플레이어가 이미지를 보고 질문자에게 비밀로 감추어진 정답 물체에 대해 예/아니오로 묻고 답하며, 정답 물체를 추려 나가는 문제이다 [1]. GuessWhat?!은 최근 컴퓨터 비전과 인공지능 학계의 많은 관심을 받고 있다.

최근 많은 연구들이 두 인공지능이 공진화하여, 특정 문제에서 인간과 유사한 성능의 상호 작용 시스템을 만드는 문제를 다루었다. 가장 성공적인 공진화 시스템의 예로 generative adversial network (GAN)가 있다 [2]. GAN은 무감독 학습을 위하여 생성기와 분류기를 경쟁시키는 모델이다. 이러한 경쟁 기작을 바탕으로 GAN은 이미지 자동 생성과 관련하여 기존 generative model을 압도하는 성능을 내었다. 서로 대화하는 chatbot 시스템은 또 다른 대표적인 예 중의 하나이다 [3]. 두 개의 구글 로봇이 며칠에 걸쳐서 철학적인 담론을 주고 받는 모습이 인터넷에 올라와서 사람들의 주목을 받았다. 한 에이전트가 이미지의 물체에 대해 한 문장으로 언급하고, 다른 에이전트가 그 물체가 무엇인지 맞추는 ReferIt!도 이와 관계가 있다 [4]. GuessWhat?!도 이러한 연구의 연장선 상에 있다.

하지만, 상호 진화하는 시스템을 계속 학습시키는 것은 어려운 일이다. 지금은 많이 개선되었지만, GAN 모델을 의미 있게 학습하는 것은 매우 어려운 일이었고 아주 최근에 이르러서야 응용에 사용될 수 있을 수준의 의미 있는 성능 개선이 있었다 [5]. 또한, 자기네들끼리의 은어로 대답하는 문제가 생긴다. Refer it의 경우, 좌표를 주는 경우, 문제가 바로 풀린다. 이는 기계학습 관점에서 보면 목표 함수에 문제가 있는 것으로, 인공지능 관점에서 보면 풀고자 하는 task의 정의에 문제가 있는 것이다.

우리는 본 논문에서, GuessWhat?! 역시 잘못 정의된 문제의 일종 중 하나라고 주장한다. 비록 GuessWhat?!이 ReferIt!의 대안으로서 생각될 수 있음에도 불구하고, 우리는 GuessWhat?! 문제가 복잡한 딥러닝이 아닌 아주 쉬운 방법으로 해결 될 수 있음을 보인다. 우리는 위치에 대해 질문하는 아주 간단한 알고리즘으로 기존 state-of-the-art 뿐 만이 아니라 인간 수준을 넘는 성능을 보일 수 있음을 논증한다.

이 논문은 단순히 GuessWhat?! 문제에서 제안하는 방법의 압도적인 성능을 보고하기 위하여 쓰여진 것이 아니다. 이 논문에서는 GuessWhat?!의 문제 특성을 분석하며, 더 나아가 GuessWhat?! 문제 세팅의 보완에 대하여 의논한다.

## 2. GuessWhat?!

GuessWhat?!에서 질문자는 이미지에서 어떤 물체가 답변 문제인지를 맞추기 위하여, 답변자에게 언어 형태로 질문을 한다. 답변자는 이미지에서 어떤 물체가 답변 물체인지를 알고 있다. 하지만, 답변자는

그림 1. 제안된 위치 기반의 질의 응답 시스템의 동작에 대한 도식.

질문자에게 예, 아니오, 잘 모르겠음 세 가지 중 하나의 답변만을 할 수 있다. 질문자가 어떤 물체가 목표 물체인지 알게 되었으면, 답변 물체를 맞추겠다고 선언한다. 그러면, 이미지에서 후보 물체들이 segment 형태로 나오게 되고 그 중 하나를 선택하여 답을 맞추면 된다. 평가가 객관식으로 이루어지기에, 정확도라는 평가 지표로 평가하기에 좋다. 이 질문자와 답변자를 모두 인공지능으로 만드는 것이 GuessWhat?! 인공지능 문제의 목표이다. GuessWhat?!을 제안한 논문에서는 이를 위하여서는 컴퓨터 비전, 자연어 처리, 의사 결정, 계획 수립 등 다양한 인공지능의 과제들을 해결하는 것이 필요하다고 주장되었다. GuessWhat?!에서는 이 문제를 학습 기반으로 학습하도록 권장하기 위하여, 이미지와 후보 물체 및 답변에 대한 정보 뿐 아니라, GuessWhat?! 문제를 실제 사람들이 질의 응답한 데이터셋을 제공한다.

### 3. 제안하는 방법

우리는 질문자와 답변자가 특정 물체에 대한 질문의 답변을 공유하고 있다면 GuessWhat?! 문제가 나무 탐색 문제와 유사하다는 점을 주목한다. 특히, 정답을 선택하는 상황에서는 정답 물체의 후보가 평균 9개 전후로 적어서, 현재 수준의 기술이나 인간 수준의 기술에 비하여, 알고리즘이 마주하게 되는 문제의 난이도가 낮다. 이러한 특성들은 인공지능의 깊은 사고와 딥러닝의 다양한 처리 없이도, GuessWhat?! 문제를 쉽게 해결할 수 있게 끔 한다.

우리는 가장 규칙 기반 방법 중 하나로 이 문제를 해결한다. 이는 기본적으로 물체의 위치에 대해 질문하는 것이다. 실제 알고리즘은 물체의 좌표에 대해 묻는 형식이 된다.

| Model | test accuracy |
|---|---|
| Baseline | 0.1604 |
| 1번 질문 | 0.3896 |
| 2번 질문 | 0.5625 |
| 3번 질문 | 0.7661 |
| 4번 질문 | 0.8585 |
| 5번 질문 | **0.9434** |
| 1번 질문 fine-tune | 0.3982 |
| 2번 질문 fine-tune | 0.5940 |
| 1번 질문 oracle segment | 0.4812 |
| 2번 질문 oracle segment | 0.8767 |
| LSTM-based system [1] | 0.6130 |
| Human performance | 0.9080 |

표 1. GuessWhat?!에 대한 성능. 제안된 방법은 3 번의 질문만으로 기존 딥러닝 기반의 모델보다 더 나은 성능을 보이며, 5번의 질문이 허용되는 경우 사람보다 더 나은 성능을 보인다.

첫 번째 질문: “왼쪽에 있니?”
답변: 3분의 1사면에 있는 경우, yes. 3분의 2사면에 있는 경우, N/A. 3분의 3사면에 있는 경우, no.
두 번째 질문: “위쪽에 있니?”
답변: 3분의 1사면에 있는 경우, yes. 3분의 2사면에 있는 경우, N/A. 3분의 3사면에 있는 경우, no.
세 번째 질문: “왼쪽에 있다고 했을 때, 그 중에서는 왼쪽에 있니?”
...

우리는 왼쪽과 오른쪽의 경계에 대하여, 학습을 할 수도 있다. 물체는 주로 왼쪽보다는 가운데에 더 많이 있다. 우리는 학습 데이터의 통계치를 통하여, 왼쪽에서 첫 41%의 공간에 전체 물체의 1/3이 존재한다는 사실을 알아냈다. 또한, 59% 이후 공간은 전체 물체의 1/3이 존재한다. 이러한 정보를 활용하여, 더 효율적으로 물체의 위치를 탐색할 수 있다. 위치에 대해서만 질문하는 경우, 이미지에 대한 정보를 사용하지 않는다고 하였을 때, 이러한 탐색은 최적 탐색을 보장하게 된다. 우리는 이 방법을 fine-tune 방법이라고 지칭한다.

혹자는 이러한 방법과 실험 결과가 GuessWhat?! 문제에 대한 부정 행위라고 주장할 수 있다. 그러한 주장은 일리에 맞고 사실에 가깝지만, 어떤 부분이 문제인지 면밀히 살펴볼 필요가 있다.
먼저, 위치에 대한 질문이 GuessWhat?! 문제의 취지에 어긋난다고 주장할 수 있다. 하지만, 이는 기존에 GuessWhat?!에서 사람 역시 흔히 하는 질문 유형으로 특별히 cheating이라고 볼 수 없다.

또한, 제안한 방법론은 위치 기반 질의가 아닌 다른 질문으로 대체될 수 있다. 이는 크기 기반 질문, 혹은 색에 대한 질문 등으로 확대될 수 있다. 또한 강력한 물체 인식 성능을 바탕으로, 특정 물체인가 아닌가? 에 대해 질문할 수도 있다. 이러한 질문들은 사람들이 보기에 이상하지 않으면서도, 기계적으로 경우의 수를 줄이는 데에 도움을 준다.

## 4. 실험 결과 및 논의

표 1은 기존 방법과 우리의 방법의 성능 차이를 보여준다. fine-tune에 대한 성능 보고는, 우리의 간단한 방법이 학습과 융합될 수 있음을 상징적으로 보여주는 것으로, 아주 약간의 성능 개선이 있었다.

이미지와 segmentation에 대한 정보가 없는 경우, 위치 기반 질의 응답은 아주 강력한 성능을 보고한다. segmentation이 있는 경우 없는 경우보다 더 좋은 성능을 만들 수 있을 것이다. 하지만, segmentation에 대해 완벽히 정보를 가지고 있는 경우에도, 위치 기반 질의 응답은 아주 강력한 성능을 보고한다. 정보를 질문자와 답변자가 모두 가지고 있기 때문에, 아주 정교하게 픽셀 단위로 물어보는 경우, optimal한 search가 가능하기 때문이다. Segment oracle은 이와 같이, 사전에 정답 물체의 후보를 정확히 알고 있다고 가정했을 때의 본 방법의 성능이다.

GuessWhat?!을 제안한 연구 [1]에서는 baseline으로서 질문자와 답변자 모두 신경망 기반 방법을 제안했다. 이것은 컨볼루션 신경망과 순환 신경망을 기반으로 하여, 질문 문장 생성도 순환 신경망으로 하고, 그 답변도 신경망 기반 분류 모델로 수행하는 것이다. 하지만, 그 성능은 상당히 낮았으며, 심지어는 이미지를 사용하는 것이 큰 개선을 가지고 오지 않았다는 결론을 보고 하였다. 후속 연구로 강화학습을 사용한 방법이 제안되었으나, 큰 개선이 있지는 않았다 [6]. 우리의 분석에 비추어 보면, 개념적으로 GuessWhat?! 문제가 네 다섯 번 정도의 질문 기회로 충분하기 때문에, 복잡한 강화학습의 탐색을 필요로 하지는 않는다.

우리는 질문자가 던지는 질문이 사람의 상식에 크게 벗어날 수 있다는 점에서 문제가 있다고 판단한다. 비록 본 논문에서 제안한 질문자와 답변자는 GuessWhat?!에서 좋은 성능을 내겠지만, 인간과는 대화할 수 없다. 따라서, 질문자가 던지는 질문이 사람이 던지는 질문과 유사하며 또한 인간이 잘 대답할 수 있어야 한다. 또한 답변자는 인간이 던지는 질문에 잘 대답할 수 있어야 한다고 본다. 다만, 후자는 전통적인 이미지 기반 질의 응답 문제에 대응되어, 전자와 별도의 문제로 생각될 수 있다. 그림 2는 이러한 이슈를 반영한 GuessWhat? 문제의 목표를 도식화하고 있다.

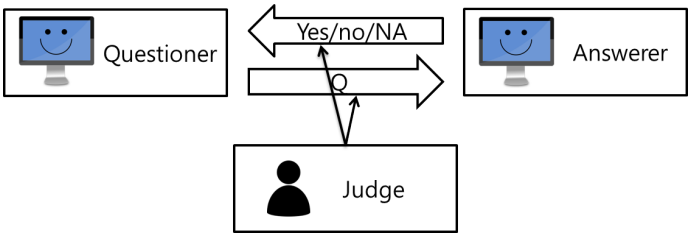


그림 2. GuessWhat?!의 실제 기대되는 문제 상황의 도식. 단순히 성능을 높이는 것 만으로는 완결성 있는 인공지능 문제가 되지 않는다.

## 5. 결론

GuessWhat?!은 인공지능 게임을 만들기 위한 훌륭한 새로운 시도였지만, 문제를 가지고 있었다. 본 논문에서는 GuessWhat?! 문제를 분석하고, 이를 바탕으로 세 번의 질문 만에 state-of-the-art 성능을 넘고, 다섯 번의 질문 만에 인간 수준의 성능에 도달하는 solver를 개발하였다.

더 나아가 GuessWhat?! 문제가 어떻게 보완될 수 있을 지, 이것이 어떻게 규칙 기반과 신경망 기반 시스템의 결합된 인공지능으로 나아갈 수 있는 지에 대하여 논의하였다.

우리의 후속 연구에서 우리는 정보이론적인 관점에서 GuessWhat?! 문제가 어떻게 이해될 수 있는 지 탐구할 것이다. 이럴 바탕으로, 지식 기반과 신경망 기반을 결합하는 알고리즘을 만들어서, 현 알고리즘보다 더 적은 질문을 가지고 성공하는 모델을 만들고, GuessWhat?! 문제를 해결할 수 있음을 보일 것이다.

## 감사의 글

이 논문은 2016년도 정부(미래창조과학부, 국방부)의 재원으로 정보통신기술진흥센터(2015-0-003102-SW스타랩), 한국산업기술평가관리원(10044009-HRI.MESSI, 10060086-RISF)의 지원을 받았음.

## 참 고 문 헌

[1] Ham de vries et al., “GuessWhat?! Visual Object Discovery through Multi-modal Dialogue,” In *CVPR*, 2017.

[2] Ian J Goodfellow et al., “Generative Adversial Nets,” In *NIPS*, 2014.

[3] Oriol Vinyals and Quoc Le, “A Neural Conversation Model,” *ICML deep learning workshop*, 2015.

[4] Junhua Mao et al., “Generation and comprehension of unambiguous object descriptions,” In *CVPR*, 2016.

[5] Martin Arjovsky et al., “Wasserstein Gan,” *arXiv*, 2017.

[6] Florian Strub et al., “End-to-end optimization of goal-driven and visually grounded dialogue systems,” *arXiv*, 2017.