# 트랜스포머를 이용한 GVQA 모델의 성능 개선에 관한 연구

박성욱*, 김준영*, 박준*, 이한성**, 정세훈**, 심춘보*
*순천대학교 IT-Bio융합시스템전공
**안동대학교 창의융합학부
411050@scnu.ac.kr, jungsh@anu.ac.kr, cbsim@scnu.ac.kr

# A Study on Performance Improvement of GVQA Model Using Transformer

Sung-Wook Park*, Jun-Yeong Kim*, Jun Park*, Han-Sung Lee**, Se-Hoon Jung**, Cun-Bo Sim*
*Interdisciplinary Program in IT-Bio Convergence System, Sunchon National University
**School of Creative Convergence, Andong National University

요 약

오늘날 인공지능(Artificial Intelligence, AI) 분야에서 가장 구현하기 어려운 분야 중 하나는 추론이다. 근래 추론 분야에서 영상과 언어가 결합한 다중 모드(Multi-modal) 환경에서 영상 기반의 질의응답(Visual Question Answering, VQA) 과업에 대한 AI 모델이 발표됐다. 얼마 지나지 않아 VQA 모델의 성능을 개선한 GVQA(Grounded Visual Question Answering) 모델도 발표됐다. 하지만 아직 GVQA 모델도 완벽한 성능을 내진 못한다. 본 논문에서는 GVQA 모델의 성능 개선을 위해 VCC(Visual Concept Classifier) 모델을 ViT-G(Vision Transformer-Giant)/14로 변경하고, ACP(Answer Cluster Predictor) 모델을 GPT(Generative Pretrained Transformer)-3으로 변경한다. 이와 같은 방법들은 성능을 개선하는 데 큰 도움이 될 수 있다고 사료된다.

## 1. 서론

최근 딥러닝(Deep Learning)의 다양한 기술들이 서로 융복합되면서 발전하고 있다[1]. 여기에 새로운 아이디어가 더해져 인공지능(Artificial Intelligence, AI)은 더 정확하게 주어진 과업을 수행한다. 그 적용 분야도 빠르게 확대될 것으로 보인다. 이러한 여세를 따라 글로벌 IT(Information Technology) 기업과 학계를 중심으로 일반 인공지능(Artificial General Intelligence, AGI) 연구가 시작됐다.

AGI는 사람과 동등한 지적 능력을 갖춘 AI다. AGI의 완성은 곧 초인공지능(Artificial Super Intelligence)의 등장을 의미한다. 컴퓨터의 시간 단위는 사람이 경험하는 실제 물리적 시간 단위와 다르다. 이미 알파고라는 AI 바둑 프로그램을 목격하면서 경험했다[2].

모든 사람이 AGI의 실현 가능성에 대해 긍정적인 입장은 아니다. AGI의 목표인 사람의 지능에 대한 정확한 이해가 부족하고, AGI를 구현하기 위해 선제적으로 해결해야 할 기술적인 문제들이 있기 때문이다. 예를 들어 IBM(International Business Machines Corporation) 왓슨은 영상 분류를 할 수 없고 알파 제로는 기계번역을 할 수 없으며 GPT(Generative Pretrained Transformer)-3은 자율 주행에 사용될 수 없다[3]. 따라서 각각 독립된 과업에 대해 훈련된 알고리즘을 하나의 알고리즘(Meta Algorithm)으로 통합해야 한다[4].

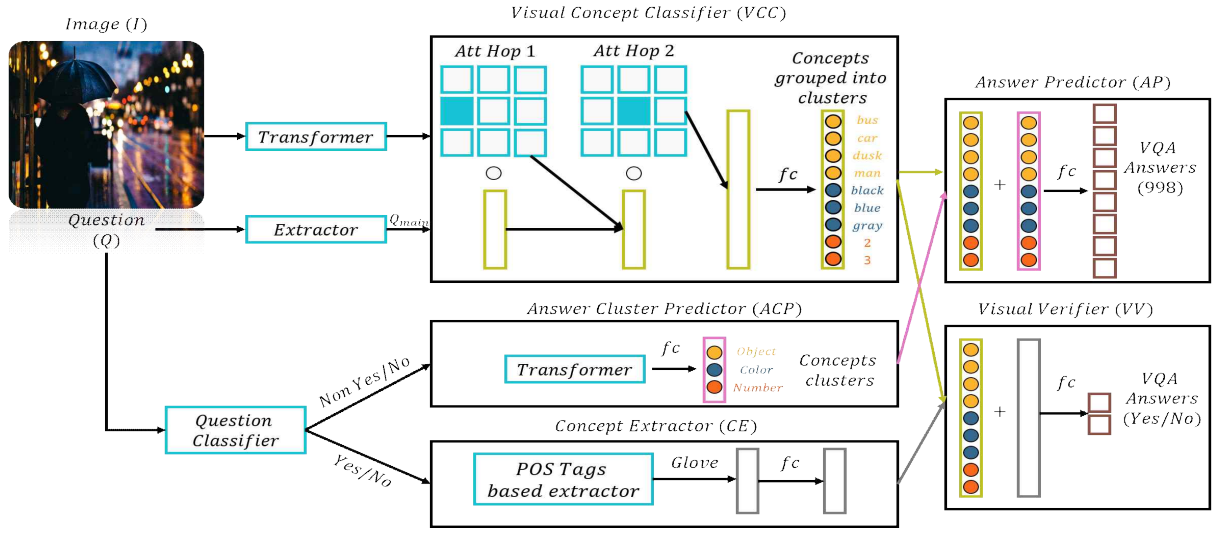
단번에 AGI를 실현하기는 불가능하다. AGI의 목표는 구체적이고, 실현 가능한 수준으로 해야 하며 단계별로 확장해야 한다. 예를 들면 범용(General Purpose) AI 시스템을 구현하는 것이다. 범용 AI 시스템은 다양한 과업을 동시에 수행할 수 있다.

오늘날 AI 분야에서 가장 구현하기 어려운 분야는 상식과 추론이다. 사람은 다양한 경험을 쌓아 원인과 결과를 이해하고, 그것에 공통으로 적용되는 상식을 터득한다. 그로 인해 특정 과업을 수행할 때 그 과업이 서로 성격이 다르더라도 성공적으로 적용할 수 있음을 안다. 그러나 AI가 상식을 체득해 서로 다른 과업을 수행할 수 있을까는 의문이다.

추론은 경험, 지식, 관측된 사실을 기반으로 가설을 세우거나 논리적 결론을 도출하는 것이다. 대표적으로 논리적 공리(Logical Axiom) 기반의 연역법(Deduction)과 통계적 공리(Probability Axiom) 기반의 귀납법(Induction)이 있다. 추론 능력의 경우 훈련을 통해 AI로 구현하고자 하는 연구가 진행되고 있는데 지금 수준은 초기 단계다.

사람의 지능 중 다른 영장류와 비교했을 때 가장 뛰어난 능력은 상식과 추론이다. 오랫동안 상식과 추론을 딥러

닝 알고리즘으로 구현하려는 연구가 진행됐지만 주목할 만 한 성과를 보여주지 못했다.




(그림 1) GVQA-T(Grounded Visual Question Answering with Transformer)의 개요

근래 딥러닝 기술의 발전과 함께 의미 있는 연구 결과들이 발표됐다. 추론 분야에서는 영상과 언어가 결합한 다중 모드(Multi-modal) 환경에서 영상 기반의 질의응답(Visual Question Answering, VQA) 과업에 대한 AI 모델이 발표됐다[5]. 얼마 지나지 않아 VQA 모델의 성능을 개선한 GVQA(Grounded Visual Question Answering) 모델도 발표됐다[6].

하지만 아직 GVQA 모델도 완벽한 성능을 내진 못한다. 본 논문에서는 GVQA 모델의 성능을 개선하기 위해 구조 및 하이퍼파라미터를 변경하여 실험해보고, 성능을 비교한다.

## 2. 관련 연구

존슨(Johnson)은 VQA를 수행하기 위해 입력된 자연어를 처리하고, 영상을 분석했다[5]. 영상 내에 있는 다중 객체는 추출하고, 크기, 색깔, 재질, 모양인 각 객체의 속성은 분류했다. 나아가 각 객체 간 상대적인 위치와 개수 세기 그리고 필요에 따라 다양한 상식을 적용했다.

VQA 수행을 위한 영상 분석에는 CLEVR(Compositional Language and Elementary Visual Reasoning) 데이터 세트를 이용했다. CLEVR는 스탠퍼드 대학교와 페이스북 AI 연구팀이 제공하는 VQA 수행을 위한 데이터 세트다. 영상 분석이 끝나면 질문에 대한 답을 준비한다. VQA 과업에서의 질문은 통상적으로 크기, 개수, 위치, 재질, 색깔, 모양과 관련 있다. 질문이 길면 모델은 답을 추론하기 힘들다. 객체 간 상대적인 관계를 조건으로 하는 경우의 수가 많아지고, 복잡해지기 때문이다.

모(Mao)는 모델이 다양한 VQA 과업을 수행하도록 만들었다[7]. 질문의 유형은 다음과 같다. 1) 영상의 객체 속성만 분석할 수 있으면 답할 수 있는 문제. 2) 영상 내 객체 간에 상대적인 위치를 추론해야 하고, 객체 수를 세는 문제. 3) 더욱 복잡한 상관성을 조건으로 하는 문제. 복잡한 상관성을 조건으로 하는 질문은 그 답에 대한 정확도가 낮을 수 있다. 그럴 때 처음에는 쉽고 간단한 환경에서 훈련하다 점차 복잡한 환경으로 전이하는 과정 학습(Curriculum Learning) 방식을 사용할 수 있다[8].

허드슨(Hudson)은 영상 분석을 위해 Mask R-CNN(Region based Convolutional Neural Networks)기반의 NSM(Neural State Machine) 모델을 발표했다[9-10]. NSM은 Mask R-CNN으로 객체를 추출한 후 그래프 모델의 노드에 할당한다. 질문을 분석해 객체에 해당하는 속성은 노드로 할당하고, 위치나 상태에 관한 정보는 그래프 에지(Edge)로 할당한다. 이를 SM(State Machine)이라 하며 SM 기반으로 최종 질문에 대한 답을 도출했다.

## 3. 제안하는 방법

그림 1은 GVQA 모델의 성능을 개선할 트랜스포머(Transformer) 기반 모델이다. 트랜스포머는 기존 Seq2seq(Sequence-to-Sequence)의 인코더-디코더 구조를 어텐션(Attention) 만으로 구현한 모델이다[11]. 트랜스포머는 RNN(Recurrent Neural Network)과 구조가 다르며 RNN보다 성능이 뛰어나다[12]. 그림 1에서 영상 분석을 하는 VCC(Visual Concept Classifier)는 ViT-G(Vision Transformer-Giant)/14 모델이다[13]. '14'는 입력 영상 패치(Patch) 한 개의 크기다. VCC는 영상의 속성을 분석해 질문하는 것에 대응한다.

질문에 관한 자연어 처리 영역에서는 첫 번째로 Yes/No 질문인지 Yes/No 질문이 아닌지를 분류한다. Yes/No

질문이면 POS(Part Of Speech) 태그를 추출한 후 GloVe(Global Vectors for Word Representation)라는 워드2벡(Word2vec) 알고리즘으로 질문의 특성을 변별력 있게 만든다. POS 태그는 자연어 처리에서 사용되는 품사 분류(Word Class) 태그다.

두 번째는 VCC에서 넘겨받은 영상 속성 정보 벡터와 결합해 최종 Yes/No 답을 위한 이진 분류 FC(Fully Connected) 신경망에 순 전파한다. 질문이 Yes/No가 아니면 GPT-3으로 구현된 ACP(Answer Cluster Predictor)가 질문을 분류한다. 이후 VCC가 추출한 특징 벡터와 결합해 다중 분류 모델인 AP(Answer Predictor)로 전달한다.

본 연구에서는 GVQA 모델의 성능을 개선하기 위해 VCC를 VGGNet에서 ViT-G/14로, ACP를 LSTM에서 GPT-3으로 변경한다[14-15].

## 4. 결론

본 논문에서는 GVQA 모델의 성능 개선을 위해 VCC 모델을 ViT-G/14로 변경하고, ACP 모델을 GPT-3으로 변경한다. 데이터 세트는 SQuAD(Stanford Question Answering Dataset), 평가는 Top-1 정확도(Accuracy)와 테스트 당혹성(Test Perplexity)을 이용한다[16]. 당혹성은 모델 스스로 성능을 수치화해 결과를 도출하는 내부 평가 지표로서 단어 수로 정규화(Normalization)된 테스트 데이터 확률의 역수다.

ViT-G/14는 현재 이미지넷(ImageNet) 데이터 세트를 사용한 영상 분류(Image Classification) 분야에서 SOTA(State Of The Art)다[17]. GPT-3도 펜 트리뱅크(Penn Treebank) 데이터 세트를 사용한 언어 모델링(Language Modelling)분야에서 SOTA다[18].

이와 같은 방법들은 성능을 개선하는 데 큰 도움이 될 수 있다고 사료된다. 이외에 학습률, 가중치 초기화 및 파라미터 갱신 방법과 같은 다른 하이퍼파라미터 조정을 통해 더 높은 정확도를 얻을 수 있을 것으로 판단된다.

## Acknowledgment

This work was supported by the BK21 plus program through the National Research Foundation (NRF) funded by the Ministry of Education of Korea(5199990214660)

## 참고문헌

[1] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” nature, vol. 521, no. 7553, pp. 436-444, 2015.

[2] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, et al., “Mastering the game of Go with deep neural networks and tree search,” nature, vol. 529, no. 7587, pp. 484-489, 2016.

[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, et al., “Language models are few-shot learners,” arXiv preprint arXiv:2005.14165, 2020.

[4] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, “Meta-learning in neural networks: A survey,” arXiv preprint arXiv:2004.05439, 2020.

[5] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, “Clevr: A diagnostic dataset for compositional language and elementary visual reasoning,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2901-2910, 2017.

[6] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, “Don't just assume; look and answer: Overcoming priors for visual question answering,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4971-4980, 2018.

[7] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu, “The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision,” arXiv preprint arXiv:1904.12584, 2019.

[8] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, “Curriculum learning,” In Proceedings of the 26th annual international conference on machine learning, pp. 41-48, 2009.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” In Proceedings of the IEEE international conference on computer vision, pp. 2961-2969, 2017.

[10] D. A. Hudson, and C. D. Manning, “Learning by abstraction: The neural state machine,” arXiv preprint arXiv:1907.03950, 2019.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and I. Polosukhin, “Attention is all you need,” arXiv preprint arXiv:1706.03762, 2017.

[12] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” nature, vol. 323, no. 6088, pp. 533-536, 1986.

[13] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, “Scaling Vision Transformers,” arXiv preprint arXiv:2106.04560, 2021.

[14] K. Simonyan, and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[15] S. Hochreiter, and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[16] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “Squad: 100,000+ questions for machine comprehension of text,” arXiv preprint arXiv:1606.05250, 2016.

[17] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and

L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” In 2009 IEEE conference on computer vision and pattern recognition, pp. 248-255, 2009.
[18] M. Marcus, B. Santorini, and M. A. Marcinkiewicz, “Building a large annotated corpus of English: The Penn Treebank,” 1993.