# 자연어 처리를 위한 조건부 게이트 다층 퍼셉트론 모델 개발 및 구현

손규진*, 김승원**, 주세준***, 조우진****, 나정은*****
*연세대학교 언더우드 국제대학 경제학과
**연세대학교 공과대학 컴퓨터과학과
***연세대학교 이과대학 수학과
****연세대학교 이과대학 대기과학과
***** 연세대학교 학부대학
spthsrbwls123@yonsei.ac.kr louisdebroglie@yonsei.ac.kr sr7418@yonsei.ac.kr snowmoon@yonsei.ac.kr
jenah@yonsei.ac.kr

# SG-MLP: Switch Gated Multi-Layer Perceptron Model for Natural Language Understanding

Guijin Son*, Seungone Kim**, Se June Joo***, Woojin Cho****,
JeongEun Nah *****
* Dept. of Economics, Underwood International College, Yonsei University
**Dept. of Computer Science, Yonsei University
*** Dept. of Mathematics, Yonsei University
**** Dept. of Atmospheric Sciences, Yonsei University
***** University College, Yonsei University

**요 약**

2018 년 Google 사의 사전 학습된 언어 인공지능 BERT 를 기점으로, 자연어 처리 학계는 주요 구조를 유지한 채 경쟁적으로 모델을 대형화하는 방향으로 발전했다. 그 결과, 오늘날 자연어 인공지능은 거대 사기업과 그에 준하는 컴퓨팅 자원을 소유한 연구단체만의 전유물이 되었다. 본 논문에서는 다층 퍼셉트론을 병렬적으로 배열해 자연어 인공지능을 제작하는 기법의 모델을 제안하고, 이를 적용한 '조건부 게이트 다층 퍼셉트론 모델(SG-MLP)' 을 구현하고 그 결과를 비교 관찰하였다. SG-MLP 는 BERT 의 20%에 해당하는 사전 학습량만으로 다수의 지표에서 그것과 준하는 성능을 보였고, 동일한 과제에 대해 더 적은 연산 비용을 소요한다.

## 1. 서론

2010 년대에 들어 범용 컴퓨팅 처리를 위한 GPGPU(General-Purpose computing on GPU) 기술이 개발되고 뛰어난 성능을 지닌 하드웨어 가속기가 연이어 등장하며 대규모 인공 신경망을 구성해 자연어 처리(Natural Language Processing) 문제를 해결하고자 하는 시도가 크게 확산하였다. 이러한 노력의 일환으로 2018 년 Google 은 BERT 모델[1]을 공개했다. Google 사의 BERT 모델은 대규모 말뭉치 데이터를 기반으로 준 지도학습(Semi-Supervised Learning)을 진행해 언어 인공지능을 사전 학습(Pretrain) 하는 방법론을 적용한다. 위 방법론과 더불어 Attention 기반의 인코더(Encoder)를 쌓는 구조 역시 제안했으며 이는 현재까지도 자연어 처리 분야에 지배적인 영향력을 미치고 있다[2]. BERT 모델을 기점으로 자연어 처리 학계는 어텐션을 쌓는 주요 구조와 사전 학습 방법론을 유지한 채 경쟁적으로 모델을 대형화하는 방향으로 발전했다. 그 결과 오늘날 자연어 인공지능은 대규모의 컴퓨팅 자원을 소유한 일부 단체에서만 연구가 가능한 실정이다. 더불어, 대부분의 연구가 차용하는 Attention 기법은 고가의 연산 비용을 요구한다는 한계점[3]을 가지고 이는 현대 언어 인공지능의 공통적인 문제점으로 지적된다.

위 단점을 보완하고자 본 논문은 Attention 기법을 배제한 새로운 구조의 자연어 처리 모델인 "조건부 게이트 다층 퍼셉트론 모델(SG-MLP)" 을 제작하였고 이를 구현하는 과정과 그 결과를 소개한다.[1]

[1] 사전 학습된 모델, 유관 코드 등은 모두 아래 링크에 첨부됨; https://github.com/guijinSON/SGMLP

## 2. 선행 연구

2 장에서는 본 논문에서 자주 언급될 Attention, 다층 퍼셉트론 그리고 조건부 신경망에 대한 기존의 연구를 소개한다.

### 2.1 다층 퍼셉트론 (Multi-Layer Perceptron)

다층 퍼셉트론(MLP)은 인공지능과 그 시작을 같이한 신경망으로, 다수의 퍼셉트론으로 구성된 층(layer)이 중첩된 구조이다. MLP 는 하나의 입력층 하나 이상의 은닉층, 그리고 하나의 출력층으로 구성되며 역전파 (back propagation)와 드롭아웃 (dropout)만으로도 우수한 성능을 보인다는 장점이 있다[4].

컴퓨터 비전(Computer Vision) 분야와 시계열 분석 분야에서 Convolutional Neural Network (CNN)과 Residual Neural Network (RNN)가 각각 등장하며 MLP 를 활용한 연구는 잠시 중단되었다. 그러나, 잘 학습된 MLP 는 그 어떤 비선형 함수에도 근사할 수 있다고 증명되며 해당 분야는 재조명되었다. 그 결과 2021 년 컴퓨터 비전 분야[5]와 음성 합성 분야[6]에서 MLP 를 사용한 모델이 기존 방법론에 버금가는 성능을 선보였다.

### 2.2 Attention 기법

Attention 기법은 2018 년 이후 자연어 처리 분야에서 압도적인 영향력을 가진 방법론으로 입력된 단어 사이의 가중치를 계산한다[1]. 이는 신경망이 언어를 이해하는 과정에서 어떠한 단어에 더 집중해야 하는지 판단하는 역할을 한다. 고정된 가중치 (static weight)를 사용하는 기존 신경망과 상이하게 입력된 값에 따라 매번 다른 가중치를 연산에 사용하는 입력 종속 가중치 (key-dependent weight)를 적용해 정보의 뒤섞임(entanglement)을 방지한다는 특징이 있다.

그러나 주어진 단어 사이의 관계를 살펴보는 정방 행렬을 구성하는 과정에서 입력된 문장의 길이 제곱과 비례하는 계산 복잡도를 요구한다[3]. 이로 인해 입력되는 문장의 길이가 길어질수록 연산 비용이 기하급수적으로 증가하며, 이에 따라 Attention 을 대체할 수 있는 기법에 대한 필요성이 대두되고 있다.

### 2.3 조건부 신경망 (Mixture of Experts)

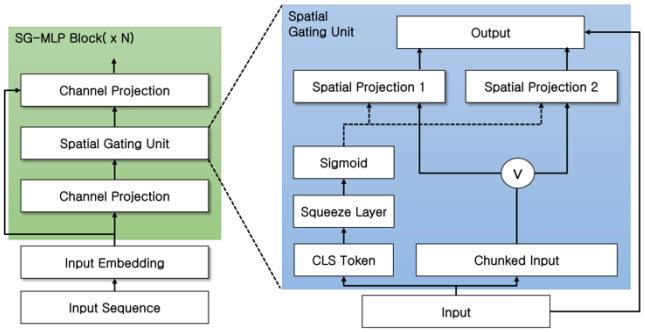
조건부 신경망(MOE)은 신경망을 병렬적으로 구성해 입력에 따라 일부 뉴런 집합만을 연산에 활용하는 구조이다. 이 특징은 엄청난 수의 파라미터를 가지는 현대의 인공지능과 결합하기에 적절하며 위 기법으로 연산 비용을 절감하고자 하는 노력[7]이 꾸준히 존재해 왔다. 이외에도 MOE 는 입력에 따라 다른 뉴런 집합을 연산에 활용해 key-dependent weight 와 비슷한 기능을 하며 entanglement 를 최소화한다.

앞서 2.2 에서 언급한 바와 같이 대부분의 현대 언어 인공지능이 차용하는 Attention 기법은 여러 한계점이 존재한다. 다음 장에서는 MLP 와 MOE 를 조합해 Attention 기법과 유사한 기능을 수행하도록 본 연구팀이 설계한 새로운 구조의 자연어 처리 모델을 소개한다.

## 3. 조건부 게이트 다층 퍼셉트론 모델 개발

2 장 선행 연구에서 살펴본 바와 같이 Attention 기법은 고가의 연산 비용을 요구한다는 단점을 가진다. 아래 3 장에서는 이를 보완한 조건부 게이트 다층 퍼셉트론 모델(Switch Gated Multi-Layer Perceptron; SG-MLP)을 제안하고 구현한다.

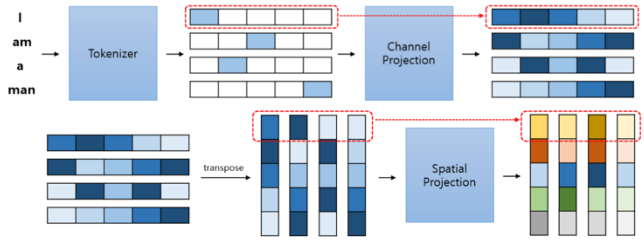
본 연구가 새로이 제안하는 SG-MLP 모델은 아래 (그림 1)의 Encoder 가 중첩된 구조이다. 개별 Encoder 는 채널 선형 변환 계층(Channel Projection)과 공간 선형 변환 계층(Spatial Projection)의 조합으로 이루어진다. 인코더의 시작과 마지막에 위치하는 Channel Projection 은 입력된 문장 벡터를 MLP 에 통과시켜 개별 단어에 대한 정보를 부호화한다. Spatial Projection 은 문장 벡터를 전치(transpose)한 뒤 MLP 를 통과시킴으로써 서로 다른 단어 간의 관계성을 부호화한다. 위 과정을 통해 부호화된 문장의 의미론적 요소들은 문장 벡터에 업데이트 된다.




(그림 1) SG-MLP 모델의 Encoder

Encoder 에 입력된 문장 벡터는 Channel Projection 을 거쳐 조건부 게이트 (Switch Gate)로 전달된다. Switch Gate 내에 위치한 분류기는 전달받은 벡터에 시그모이드 함수를 적용하고 이진 분류를 진행한다. 분류의 결과에 따라 Switch Gate 내에 있는 두 개의 Spatial Projection 중 하나만이 부분적으로 활성화되어 입력값을 부호화한다. 아래 (그림 2)는 앞서 설명한 부호화 과정으로, 해당 방법을 통해 SG-MLP 는 Attention 의 도움 없이도 각각의 단어와 단어 간의 관계성 모두를 학습한다. 나아가, Switch Gate 내에 Spatial Projection 을 병렬로 배치하고 서로 다른 문장에 대해 상이한 뉴런 집합을 연산에 사용함으로써 Entanglement 가

최소화된다. MOE 에서 영향을 받아 제작된 위 구조는 SG-MLP 모델이 자체적으로 문장 이해에 필요한 심층 신경망만 선택해 입력된 문장을 분석하도록 한다.




(그림 2) SG-MLP 정보 부호화 과정 시각화

SG-MLP 는 사전 학습(Pretrain)과 사후 학습(Fine-Tune)으로 나누어 학습되었다. 본 논문에서는 학습 데이터, 모델 규모, 학습량이 모델의 성능에 미치는 영향을 포괄적으로 탐구하기 위해 위 조건을 달리해 아래 <표 1>과 같이 총 세 버전의 모델을 제작했다. SG-MLP Base 는 BERT Base 모델과의 비교를 위해 비슷한 양의 파라미터를 가지도록 설정하였으며 이를 기준으로 0.5 배의 SG-MLP Small 과 1.3 배의 SG-MLP Large 를 추가로 공개한다. SG-MLP Small 은 약 16GB 에 해당하는 Book Corpus 와 Wiki 데이터로 학습되었으며 이외의 모델은 Common Crawl's Web Crawl Corpus 를 한 차례 정제한 Allen · C4 · En[8] 데이터를 학습에 활용하였으며 이는 약 305GB 정도의 분량이다. 보유한 하드웨어 가속기의 한계로 인해 SG-MLP Base 모델의 학습 스텝 수는 BERT Base 의 20% 인 200,000 회로 제한해 진행하였다. 아래 <표 1>은 각 모델의 사전 학습 환경에 대한 세부 사항을 명시하였다.

| 모델 | 파라미터 | 데이터셋 | 학습량 |
|---|---|---|---|
| BERT Base | 1.09 억 | Book Corpus + Wiki | 1,000,000 |
| SG-MLP Small | 0.67 억 | Book Corpus + Wiki | 110,000 |
| SG-MLP Base | 1.25 억 | C4 | 200,000 |
| SG-MLP Large | 1.67 억 | C4 | 200,000 |

<표 1> 사전 학습 환경 설정

## 4. SG-MLP 모델 성능 평가

4 장에서는 SG-MLP 모델의 연산 속도를 측정하고, 여러 정성적, 정량적인 평가를 통해 모델의 언어 이해 능력을 검증한다.

### 4.1 타 모델과 연산 속도 비교

SG-MLP 는 입력값에 따라 서로 다른 뉴런 집합이 활성화되는 조건부 신경망(MOE) 형태의 구조를 차용한다. 이와 같이 병렬적으로 모델을 구성함으로써 모델에게 언어 이해에 요구되는 충분한 양의 파라미터를 제공하지만, 모델은 추론을 하는 과정에서 입력값에 따라 필요한 부분만 유동적으로 활용하게 된다. 이는 연산에 소요되는 시간을 절약하는 결과로 이어진다. SG-MLP Base 와 크기가 비슷한 BERT Base, RoBERTa Base 모델과 연산 속도를 비교함으로써 SG-MLP 모델이 타 모델에 비해 저렴한 연산 비용을 소비함을 확인하였다[1,11].

아래 <표 2>의 실험은 모두 Tesla V100 SXM2 16GB GPU 에서 동일한 양의 연산을 처리하는데 소요한 시간을 측정했다. 실험 결과 SG-MLP 모델은 비슷한 파라미터의 양을 가지는 타 모델에 비해 적은 연산 시간을 소요했으며, 되려 모델의 크기가 더 작은 BERT Base 모델보다도 더 빠른 연산속도를 보여주었다.

| 모델 | 파라미터 수 | 연산 속도 (ms) |
|---|---|---|
| BERT Base | 1.09 억 | 0.261 |
| RoBERTa Base | 1.24 억 | 0.260 |
| SG-MLP Base | 1.15 억 | **0.230** |

<표 2> SG-MLP 연산 속도 비교 평가

### 4.2 Prompt 를 통한 언어 이해 능력 평가

사전 학습이 완료된 자연어 인공지능은 빈칸을 포함한 제시문(Prompt)을 입력받고 그에 알맞은 단어를 예측하는 과제를 수행하며 언어 이해 능력을 정성적으로 평가 받는다[9]. 위 과정은 수치적으로 표현이 힘든 일반 상식에 대한 모델의 학습 수준을 검증하는 기능을 수행한다.

본 논문에서는 SG-MLP Base 와 BERT-base-uncased 에게 빈칸을 의미하는 <mask>가 포함된 동일한 문장을 제공하고 예측 값에 대한 관찰을 진행했다. 그 결과, BERT 와 비교해 추상적인 개념과 일반 상식을 더 잘 학습했음을 확인했다. 아래 <표 3>는 사용한 일부 prompt 와 각 모델의 결과를 담고 있다. 아래와 같이 SG-MLP Base 모델은 동일하게 음식과 관련된 명사 중에서도 커피는 마시는 행위의 대상이고 스테이크는 먹는 행위의 대상임을 인지한다. 나아가 새는 두 개의 다리를, 차는 네 개의 바퀴를 가지고 있음을 예측하며 학습 과정에서 기본적인 지식 역시 습득했음을 확인할 수 있다. 이와 반대로 BERT 모델은 이미 여러 차례 일반 상식이 부재함을 지적 받은 바[9] 있다. 본 연구의 실험에서 SG-MLP 는 {eat, drink, two, 4}을 예측한 것과 달리 BERT 는 {be, make, no, two} 라고 빈칸을 채우며 문법적으로는 옳지만 문맥상 말이 되지 않는 문장을 생성했다.

| 입력 문장 | SG-MLP | BERT |
|---|---|---|
| I want to <mask> steak. | **eat** | be |
| I want to <mask> coffee. | **drink** | make |
| A bird has <mask> legs. | **two** | no |
| A car has <mask> wheels. | **4** | two |

<표 3> SG-MLP Base 와 BERT Base 빈칸 예측

### 4.3 GLUE 를 통한 언어 이해 능력 평가

GLUE 는 언어 인공지능의 자연어 이해 능력을 평가하는 대표적인 지표[10]로 본 논문에서는 그 중 CoLA, SST-2, QQP, RTE 를 활용해 모델에 대한 평가를 진행한다. 그 결과를 담고 있는 아래 <표 4>는 총 두 가지 사실을 시사한다.

| 모델 | Dev | | | | |
|---|---|---|---|---|---|
| | CoLA | SST-2 | QQP | RTE | 평균 |
| BERT Base | 86.3 | 93.5 | 72.1 | 66.4 | 79.5 |
| SG-MLP Small | 70.9 | 82.8 | 83.5 | 52.0 | 72.3 |
| SG-MLP Base | 71.0 | 86.6 | 84.7 | 59.0 | 73.3 |
| SG-MLP Large | 71.3 | 89.1 | 88.0 | 62.6 | 77.4 |

<표 4> GLUE 평가 결과

첫째, SG-MLP 모델은 문장 간 유사도를 분석하는 QQP 에서는 BERT Base 모델을 상회한다. 나아가, SG-MLP Base 모델은 BERT Base 의 20% 정도만 학습되었음에도 평균적으로 크게 차이 나지 않는다.

둘째, <표 4>는 구조적으로 동일한 SG-MLP Small, Base, Large 모델이 파라미터 수와 학습량을 늘리는 것 만으로 성능이 향상됨을 보인다. 확장성(Scalability)이라고도 불리는 이 특성은 현대 인공지능에서 그 중요성이 더욱 강조되며 본 논문이 제안하는 SG-MLP 모델 역시 추가적인 연구 없이 학습 환경의 개선만으로 더욱 우수한 결과를 얻었다.

## 5. 결론

본 논문은 2018 년 이후 자연어 처리 분야에서 지배적인 입지를 유지해온 Attention 기법에 대해 대체 가능성을 가진 '조건부 게이트 다층 퍼셉트론 모델(SG-MLP)'을 제안하고 구현한 결과와 기존 모델을 비교하여 평가한다. 본 연구가 제안한 SG-MLP 모델은 기존의 자연어 인공지능에 비해 두 가지 이점을 가진다. 첫째, SG-MLP 모델은 동일한 규모의 기존 신경망보다 적은 연산 비용과 사전 학습 비용이 소모된다. 둘째, SG-MLP 모델은 문법적 오류 분석, 문장 감정 분석, 문장 유사도 분석 등의 분야에서 평균적으로 Google 사의 BERT 와 준하는 언어 이해 능력을 보이며 일부 지표에서는 BERT 를 앞서는 결과를 보여준다.

자연어 처리 능력을 지닌 인공지능의 적용 범위가 산업 전반으로 확대됨에 따라 저렴하고, 우수한 성능을 가진 자연어 인공지능에 대한 수요는 증가하고 있다. 본 논문에서 제안한 SG-MLP 와 같이 모델을 경량화함과 동시에 성능을 개선하는 연구는 지속해서 진행되어야 한다. 향후 SG-MLP 내에 다양한 신경망을 병렬적으로 배치하거나, 학습량을 늘려 추가적인 성능 개선을 목표할 것이다.

-사사-

본 논문은 과학기술정보통신부
정보통신창의인재양성사업의 지원을 통해 수행한
ICT 멘토링 프로젝트 결과물입니다.

## 참고문헌

[1] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186). (2019)

[2] Qiu, Xipeng, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. "Pre-trained models for natural language processing: A survey." *Science China Technological Sciences*: 1-26. (2020)

[3] Kitaev, Nikita, Łukasz Kaiser, and Anselm Levskaya. "Reformer: The efficient transformer."In *Internetional Conference on Learning Representation*, Addis Ababa, Ethiopia. (2020)

[4] James L. McClelland, David E. Rumelhart, and CORPORATE PDP Research Group (Eds.). 1986. Parallel distributed processing: explorations in the microstructure, vol. 2: psychological and biological models. MIT Press, Cambridge, MA, USA

[5] Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., ... & Dosovitskiy, A. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*. (2021)

[6] Tae, J., Kim, H., & Lee, Y. MLP Singer: Towards Rapid Parallel Korean Singing Voice Synthesis. *arXiv preprint arXiv:2106.07886* (2021).

[7] Fedus, W., Zoph, B., & Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint arXiv:2101.03961*. (2021).

[8] Dodge, Jesse, Maarten Sap, Ana Marasovic, William Agnew, Gabriel Ilharco, Dirk Groeneveld, and Matt Gardner. "Documenting the english colossal clean crawled corpus." *arXiv preprint arXiv:2104.08758* (2021).

[9] Lin, Bill Yuchen, Seyeon Lee, Rahul Khanna, and Xiang Ren. "Birds have four legs?! numersense: Probing numerical commonsense knowledge of pre-trained language models." *arXiv preprint arXiv:2005.00683* (2020).

[10] Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. "GLUE: A multi-task benchmark and analysis platform for natural language understanding." In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (2018).

[11] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "Roberta: A robustly optimized bert pretraining approach." *arXiv preprint arXiv:1907.11692* (2019).