힉스와 인공지능

DOI: 10.3938/PhiT.26.047

김 태 정

Higgs and Artificial Intelligence

Tae Jeong KIM

Since the discovery of the Higgs boson, we have been facing new challenges. Nature is still there to be discovered and understood. The Large Hadron Collider at the CERN has collected many more data than ever before. Data are being collected, and 100 times more will be collected in coming years. Artificial Intelligence is everywhere in the world and has already been acknowledged by all big data companies and institutes to be a power in itself. In this article, recent developments in artificial intelligence for tackling the data deluge in particle physics will be discussed.

인공지능이라는 이름이 하나의 유행처럼 전 세계를 떠돌고 있다. 구글의 자율 주행, 페이스북의 얼굴 인식, 네이버의 음성 인식, 아마존의 고객 패턴 분석 등 빅 데이터가 있는 곳에는 인공지능이 있고 인공지능은 미래의 공상과학이 아닌 현실을 바꾸는 하나의 도도한 흐름으로 인정받고 있다. 이제 빅 데이터를 다루는 입자물리학은 좀처럼 드러내고 있지 않는 자연을 이해하기 위해 인공지능을 데이터 분석의 중요한 방법으로 대치해야 할 시기를 맞이하고 있다.

포스트 힉스시대

세상을 이루는 가장 작은 입자가 무엇일까? 이 질문은 고대 그리스 시대 때부터 해왔고 많은 철학자들이 그럴듯한 답을 내놓았

저자약력

김태정 교수는 고려대학교 입자물리 전공 박사(2007)로 2008년부터 미국 University of Iowa, 벨기에 Brussels University(VUB) 연구원을 거쳐 2014년 전북대학교 조교수를 거쳐 2015년부터 한양대학교 부교수로 재직 중이다. CMS 국제공동연구팀에서 top quark cross section 그룹 convenor로 활동하는 등 top quark와 관련 연구를 중점으로 하고 있다. (taekim@hanyang.ac.kr)

Fig. 1. The diphoton invariant mass distribution. The events at 125 GeV is possibly from the Higgs boson decaying to two photons.^[2]

다. 21세기 현재를 살고 있는 지금 우리는 똑같은 질문을 하지만 우리들의 가장 그럴듯한 답은 표준모형이다. 표준모형의 완성은 공식적으로 2012년 7월 4일 CERN에서 힉스입자의 발견을 발표하면서 이루어졌다.^[1,2] 그림 1은 힉스입자가 두 개의 광자로 붕괴하는 경우 두 광자의 불변질량을 보여주고 있다. 이 힉스입자의 존재는 거의 60년 전 1964년 벨기에의 프랑수와 앙글레르, 로버트 브라우트 그리고 영국의 피터 힉스가 입자의 질량을 힉스 장을 이용해 설명하면서 힉스입자가 있을 것이라고 예상되어 왔다. 거의 60년 만에 힉스입자를 발견한 이후 5년이 지난 지금 발견 당시의 물리학자들이 느꼈던 감동은 이제 아직 풀리지 않고 있는 우주를 이해하기 위한 절실한 호기심으로 바뀌고 있다.

은하계의 회전속도는 예상하는 것보다 빠르게 돌고 있고 이 빠른 속도를 설명하기 위해서는 구심력으로 작용할 수 있는 우리가 살고 있는 세계와 상호작용을 하지 않아 보이지 않는 암흑 물질이 있어야 한다. 이를 설명하기 위한 이론 중 하나는 초

REFERENCES

- [1] ATLAS collaboration, Phys. Lett. B **716**, 1 (2012).
- [2] CMS collaboration, Phys. Let. B **716**, 30 (2012).

대칭 이론으로 이 이론에서는 표준모형의 입자들에서 스핀 정보만 반대인 같은 수만큼의 새로운 입자가 존재한다. 예를 들어 스핀이 1/2인 페르미온과 대칭을 이루는 스핀이 1인 페르미온이 존재하고 스핀이 1인 보존과 대칭이 되는 스핀이 1/2인 보존 입자가 존재한다. 초대칭 이론에 R 반전성을 도입하면 가장 가벼운 중성인 뉴트랄리노 입자가 안정적이 되어서 암흑물질의 후보가 된다. 또한 초대칭 이론은 또한 힉스 입자의 질량이 양자역학적 보정에 의해 아주 커지지 않고 125 GeV로 낮은 이유 즉, 계층문제를 자연스럽게 설명해주기도 한다.

그래서 물리학자들은 초대칭 입자를 찾기 시작했고 초대칭 입자가 있다면 CERN의 강입자 충돌기(LHC)에서 1 TeV 이하에서 발견될 수 있을 것이라고 예상하였다. 하지만 LHC에서 데이터를 받기 시작한 후 8년, 힉스 발견 이후 5년이 지난 지금 아직 초대칭 입자의 흔적을 못 찾고 있다. 물론 초대칭 입자의 질량이 현재 LHC 충돌에너지로는 탐색이 점점 어려워지는 1 TeV 이상일 수 있지만 그 동안 낮은 질량의 초대칭 입자를 통해 힉스입자 질량 문제를 자연스럽게 풀려고 했던 많은 이론들이 위협을 받고 있는 상황이다.

빅 데이터

한 가지 예전과 달라진 점이 있다면 힉스 발견 이후 LHC에서 만들어진 데이터의 양을 보면 구글이나 페이스북 못지않게 방대하다는 점이다. LHC에서 1초에 양성자가 십억 번 이상 충돌하고 이때 양성자가 붕괴하여 만들어지는 데이터의 양은 1초에 1 PB이다. 하지만 모든 데이터를 저장할 수 없고 관심있는 충돌사건만 저장하는데 이 데이터만 해도 LHC에서 매년 200 PB의 양이 쌓이게 된다. 저장된 데이터는 세계 소속 기관으로 복사되어 저장되고 데이터 분석을 할 수 있는 포맷 형태로 바꾸어 다시 저장되어 결국에는 1 EB가 넘는 데이터가 매년 생기고 데이터 분석을 진행하게 된다. 이제 전부가 아니라 2023년 이후 LHC는 광도가 좀 더 높은 양성자 빔을 이용하여 1초당 충돌하는 양성자 충돌의 개수를 10배 이상 늘릴 계획이다. 이를 High-Luminosity LHC(HL-LHC)라 부르고 이 기간 동안 받을 것으로 예상되는 데이터의 양은 현재보다 100배 이상 많을 것으로 예상하고 있다.

이런 빅 데이터를 입자물리에서 분석하는 방법은 전통적으로 물리적인 지식을 동원하여 가장 효율적이라고 생각하는 변수를 인간이 찾아내어 1차원 혹은 2차원 분포에서 직선 하나로 신호와 배경사건을 구분해 내는 것이었다. 하지만 원자세계에서 물리적인 현상은 확률적으로 분포하기 때문에 단순히 선형적으로는 효율적으로 원하는 신호를 분리해 낼 수 없어 곡선형태의 비선형적인 분류방법이 필요하게 되었다. 이는 인공지능의 한 분야인 기계학습의 시작이라고 할 수 있다.

Fig. 2. Boosted Decision Tree used for observation of single top quark production.^[3]

기계학습

인공지능이란 말 그대로 인간의 지능을 기계로 구현하는 것으로 포괄적인 의미를 가지고 있다. 이중 기계학습은 구체적으로 어떤 임무가 주어졌을 때 훈련 데이터를 통해서 성능을 향상시키고자 하는 컴퓨터 알고리즘 연구라고 정의할 수 있을 것이다. 이는 전통적으로 사람이 프로그램을 만드는 것과는 다르게 컴퓨터에 데이터와 결론을 넣고 컴퓨터를 학습시켜 컴퓨터가 프로그램을 만들도록 하는 것이다.

사실 입자물리에서 기계학습을 통한 데이터 분석은 이미 오래 전부터 시도되어 왔다. 예를 들어 가장 많이 쓰이는 기계학습방법으로는 의사결정구조(Boosted Decision Tree 이하 BDT)라는 것이 있다. 신호와 배경사건을 구분할 수 있는 많은 변수를 입력값으로 컴퓨터에 넣어주면 컴퓨터가 신호와 배경사건을 구분할 수 있는 효율적인 변수를 선정하고 스캔하여 변수에 적용하는 값을 선택해서 준다. 그림 2는 입자물리 데이터 분석에서 사용되는 BDT의 기본적인 구조를 보여주고 있다. 입자물리실험에서 기계학습의 승리는 2009년 단일 탑 쿼크라는 새로운 입자의 발견이라고 할 수 있다. 단일 탑 쿼크는 탑 쿼크의 쌍 생성과 많이 비슷하고 생성되는 확률이 작아 단순하게 선형적인 즉 하나의 변수에 값을 적용하여 잘라내는 방법으로는 구분해내기가 쉽지 않은 물리적인 사건이다. 여러 가지 변수를 이용하여 BDT 또는 단순한 신경망 알고리즘(Artificial Neural Network)을 이용하여 발견할 수 있었다. 그림 3은 기계학습에서 나온 출력 값을 보여주고 있고 검은색 점들은 실험에 받은 충돌 데이터이고 단일 탑 쿼크 사건이 있을 때 데이터와 잘 일치됨을 볼 수 있다. 이뿐만 아니라 기계학습은 2012년 힉스 발견에서도 힉스입자가 광자로

REFERENCES

- [3] https://www-d0.fnal.gov/Run2Physics/top/singletop_observation/.

Fig. 3. The discriminant output from the Boosted Decision Trees Analysis.^[4] The signal of the single top quark process is indicated by blue color.

붕괴하는 경우 이용되었다. 광자의 에너지 값을 결정하기 위해 회귀분석을 이용하였고 힉스입자가 광자로 붕괴하는 경우를 다른 사건과 구분하기 위해 분류기법을 위해 이용되었다.

딥러닝

최근 2-3년간 기계학습에 관한 논문들이 쏟아져 나오고 있고 우리는 다시 기계학습을 이야기하고 있다. 다시 기계학습을 이야기하는 이유를 찾기 전에 인공지능을 이루고 있는 기본구조인 인공뉴런에 대해 살펴보자. 인공뉴런의 구조는 예를 들어 입력 신호가 두 개(x_1, x_2)라면 두 개의 신호를 입력으로 받아 각각의 입력값에 가중치(w_1, w_2)를 부여하고 여기에 편향값(b)을 더하여 그 합($b + w_1x_1 + w_2x_2$)이 어떤 기준값 0보다 크면 1을 출력하고 작으면 0을 출력하는 아주 단순한 구조이다. 이와 같이 신경망은 입력층(input layer), 은닉층(hidden layer), 출력층(output layer)으로 구성된다. 입력신호의 총합을 출력층의 신호로 바꾸어 주는 함수를 활성화 함수라고 하고 단순 인공뉴런에서는 활성화 함수가 계단함수가 될 것이다. 여기서 4차 혁명이라고 부르는 딥러닝 즉, 인공뉴런이 여러 겹 쌓여 있는 다중 신경망으로 나가는 핵심은 활성화 함수를 비선형으로 쓰는 것이다. 비선형 함수를 어떤 함수를 쓰는 것이 좋은지에 대한 연구가 많이 이루어졌고 최근에는 sigmoid, ReLU, tanh 등이 많이 쓰이고 있다. 그림 4는 은닉층이 하나인 단순한 2층 신경망 구조를 보여주고 있다. 마지막 활성화 함수를 거치면서 출력층의 값이 확률값(예를 들어 0부터 1까지)을 가지게 된다.

중요한 것은 매개변수인 가중치와 편향값을 조정하여 원하는 출력층의 값을 얻는 것이다. 그 방법으로 신경망에서 나온 출력값이 우리가 원하는 값인지 확인하기 위해 알고 있는 정답과

Fig. 4. Example of a simple neural network.

비교하는 것이다. 그 차이를 나타내는 지표를 손실 함수(Loss)라고 하고 보통 평균 제곱 오차(mean squared error)와 교차 엔트로피 오차(cross entropy error)를 이용한다. 이 함수가 최솟값이 0이 될 때까지 인간이 아닌 컴퓨터가 가중치 값과 편향값을 조정해준다. 방법으로는 손실 함수를 가중치 또는 편향값으로 편미분하여 기울기가 하강하는 방향으로 매개변수 값을 조정하는 것이다(갱신된 가중치 = 가중치 - 학습률 × 편미분값).

다시 처음 질문으로 돌아가 그 동안 딥러닝이 많이 쓰이지 않았던 이유를 살펴보면 훈련 데이터에만 좋은 성능을 발휘하는 과적합(overfitting)에 관한 문제이고 은닉층이 많아지면서 훈련 속도가 느려진다는 문제였다. 이런 과적합의 문제는 예전보다 훨씬 많은 데이터의 양과 무작위로 몇 개의 뉴런을 제거하는 방법인 dropout 같은 방법으로 지나친 학습을 억제할 수 있게 되었고 알고리즘의 발전(예를 들어 활성화 함수로 간단한 ReLU의 사용)과 하드웨어적으로는 그래픽 처리 장치(GPU)의 발전으로 훈련 속도가 예전보다 빨라지게 되면서 여러 하이퍼파라미터들(각 층의 뉴런 수, 배치 크기, 학습률 등)을 빠른 시간안에 최적화시키는 것이 가능하게 되었다. 이는 딥러닝이 데이터 분석에 중요한 수단이 될 수 있는 중요한 요소가 된다.

무거운 바텀 쿼크의 구별

입자물리에서 이런 딥러닝을 이용한 연구가 활발하게 진행되고 있다. 표준모형에는 6개의 쿼크가 존재하는데 이중 수명이 짧아 바로 붕괴하는 탑 쿼크를 제외하고 상대적으로 무거운 질량을 가진 바텀 쿼크를 구분하는 일은 쉽지 않은 과제 중 하나이다. 최근 CERN의 CMS(Compact Muon Solenoid) 국제공동 실험에서 딥러닝을 이용하여 이 무거운 쿼크를 구분하는 데 좋은 성능을 보여주고 있다. 먼저 쿼크가 어떻게 입자 검출기에서 검출되는지 살펴보자. 탑 쿼크를 제외한 다른 쿼크들은 혼

REFERENCES

- [4] V. M. Abazov *et al.*, Phys. Rev. Lett. **103**, 092001 (2009).

Fig. 5. Output from the deep learning framework for b-tagging. The output discriminator is close to 1 for bottom quark indicated by red color (color online).^[5]

자 존재할 수 없어 항상 다른 쿼크와 함께 존재한다. 가속기 충돌에서 생성된 쿼크 입자 또는 강한 상호작용을 매개하는 글루온입자들은 에너지를 가지고 한쪽 방향으로 날아가면서 더 이상 붕괴하지 않는 안정적인 하드론 입자(세 개의 쿼크로 이루어진 바리온 또는 두 개의 쿼크로 이루어진 메존) 입자들을 만든다. 이런 입자들을 하드론이라고 부르고 일정한 원반의 묶음을 제트(jet)라고 부른다. 무거운 바텀 쿼크가 가지고 있는 특징은 다른 업, 다운, 스트레인지 쿼크보다 무겁고 수명이 길어 붕괴하기까지 1.5 ps 시간이 걸려 그동안 1.8 mm 정도 더 날아가고 하전입자를 띤 입자들을 더 많이 만들어 낸다. CMS 공동실험에서는 이런 특징적인 정보들을 입력 값으로 은닉층의 개수는 4개이고 각 층의 노드(nodes)가 100개인 딥러닝 신경망 구조를 사용하여 무거운 바텀 쿼크를 구분하려고 시도하였다. 그림 5는 바텀쿼크(빨간색)의 출력값이 1에 가깝고 나머지 쿼크들은 0에 가까워 바텀쿼크가 잘 구분되고 있음을 보여주고 있다. 딥러닝을 이용한 분류는 기존의 방법보다 좋은 성능을 보여주어 CMS 공동실험에서는 딥러닝을 이용한 분류방법을 바텀 쿼크를 분류하는 대표적인 방법으로 쓰고 있다.^[5]

글루온과 쿼크 제트의 구분

입자물리 실험에서 또 한 가지 어려운 질문 중 하나는 쿼크 제트와 글루온 제트를 구분하는 것이다. 쿼크 제트와 글루온 제트는 서로 다른 강한 상호작용으로 인해 검출기에 각각 다른 형태로 에너지를 남긴다. 글루온 제트가 쿼크 제트보다 더 많은 입자들을 만들어내고 넓게 분포하는 성질을 가지고 있다. 기존에는 물리적인 지식을 동원하여 복잡한 변수를 사람이 직접 만들어 구분하려는 노력을 하였다. 하지만 검출기에 남겨진

Fig. 6. An illustration of the deep convolutional neural network architecture.^[6]

Fig. 7. ROC curves of the deep convolutional network comparing with other algorithms.^[6]

eta-phi 평면(그림 6 참조)에서의 에너지 분포를 사진처럼 이미지로 이용한다면 사진에서 고양이와 강아지를 분류하듯이 구분할 수 있을 것이다. 최근 한 연구에서 이미지의 특징을 필터를 통해 뽑아내는 합성곱 신경망(Convolution Neural Network) 기법을 이용하여 인간이 그동안 만들어낸 복잡한 변수보다 좀 더 많은 정보를 컴퓨터가 스스로 찾아낼 수 있게 하였다.^[6] 합성곱 신경망에서 합성곱 층을 여러 개 이용한다면 좀 더 복잡한 특징들을 뽑아낼 수 있다. 이를 통해 인공지능으로 기존의 인간이 만들어낸 변수보다 쿼크 제트와 글루온 제트 구분을 더 잘 할 수 있는 것을 보여준 것이다. 그림 7을 보면 Deep

REFERENCES

- [5] CMS collaboration, CMS-DP-2017/005.
- [6] Patrick T. Komiske, Eric M. Metodiev and Matthew D. Schwartz, JHEP 01, 110 (2017).

Fig. 8. (a) Jet image for signal and (b) Jet image for background.^[7]

CNN 방법(실선 검은색 또는 회색)이 다른 방법보다 같은 쿼크 제트 효율을 보일 때 더 많은 글루온 제트를 제거시켜 더 좋은 성능을 가지고 있다는 것을 보여주고 있다.

탑 쿼크 제트의 구분

입자물리에서 합성곱 신경망을 이용하는 가장 대표적인 예는 탑 쿼크 제트의 구분이다. 표준모형에서 가장 무거운 입자인 탑 쿼크는 바텀 쿼크와 W 보손 입자로 붕괴한다. 그리고 W 보손 입자는 많은 경우 두 개의 쿼크로 붕괴한다. 이때 최종적으로 쿼크는 제트의 형태로 검출기에 에너지를 남기고 결국 세 개의 제트 입자가 된다. 하지만 만약 탑 쿼크의 운동량이 아주 크다면 세 개의 제트 입자가 한 방향으로 붕괴되면서 하나의 큰 제트입자 형태로 나타날 수 있다. 이때 검출기의 eta-phi 평면상의 에너지 분포(그림 8)는 하나의 커다란 원에 세 개의 제트입자 하부 구조 형태를 띠게 된다. 반면 배경사건인 QCD 과정을 통해서 오는 여러 제트는 하부 구조가 없는 형태를 띠게 된다. 이 평면상의 에너지 분포를 합성곱 신경망으로 분석한다면 좀 더 효율적으로 신호와 배경사건을 구분할 수 있음을 보여주고 있다.^[7]

소프트웨어

가속기를 기반으로 하는 입자물리 실험데이터 분석에서 실질

적으로 딥러닝을 구현하는 방법은 크게 두 가지가 있다. 하나는 전통적으로 BDT같은 기계학습 알고리즘을 위해 개발한 Toolkit for Multivariate Analysis(TMVA)^[8]를 이용하는 방법이다. TMVA는 입자물리 실험목적으로 만들어진 소프트웨어로 지난 이십여 년 동안 사용되었기 때문에 입자물리 데이터 분석에 필요한 새로운 요구가 있을 때 쉽게 구현해 낼 수 있다. 문제는 최근 빠르게 발전하고 있는 딥러닝 관련 알고리즘과 아이디어를 바로 적용하기가 쉽지 않고 학습 속도도 아직까지는 최적화가 되어 있지 않아 느리다는 단점이 있다. 또 다른 방법은 딥러닝 오픈소스 라이브러리 Tensorflow, theano, Caffe, MXNet 등을 쓰는 것이다. 이는 빠르고 최신 딥러닝 알고리즘을 이용할 수 있지만 문제는 입자물리에서 쓰고 있는 데이터 형식과 맞지 않아 직접적으로 쓰기에는 힘들고 데이터 형식을 바꾸어 줄 필요가 있고 입자물리 데이터분석에 필요한 기술적 지원을 장기간 안정적으로 받을 수 있는지도 보장이 되어있지 않다는 점이다. 하나의 프레임워크를 이용한다는 것은 공동연구를 효율적으로 할 수 있다는 점뿐만 아니라 결과의 재생산을 위해서도 필요하다. 입자물리 실험에서 현재 다양하게 존재하고 있는 딥러닝 관련 프레임워크를 하나로 만들려는 노력이 필요하고 실제로 논의가 진행되고 있다.

앞으로의 전망

지금까지 연구들은 대부분 데이터 과학자들이 빅 데이터의 성능을 입증하기 위해 입자물리의 빅 데이터를 활용하였다면 이제는 입자물리학에서 필요한 직접적인 요구에 맞게 딥러닝이 활용될 때이다. 예를 들어 입자물리 실험 그룹에서 사용되고 있는 패턴 분석 프로그램이나 입자 재구성 방법들이 딥러닝 방법으로 대체되고 있고 딥러닝을 이용하여 인간이 생각해내지 못한 정보까지 데이터에서 추출하여 물리학자들이 전통적인 실험에서 할 수 있는 영역의 범위를 크게 확장하고 있다. 힉스입자의 발견 이후 새로운 물리현상의 발견이 요구되고 있고 2023년부터 HL-LHC가 가동되면 데이터의 양이 현재의 백 배 이상이 될 것으로 예상되는 지금 기계학습은 적절한 시기에 딥러닝이라는 이름으로 다시 살아나고 있다. 이를 통해 바텀 쿼크의 관별 같은 입자의 재구성의 활용에서부터 우리가 잘 모르고 있는 새로운 물리현상의 발견, 단면적이 작아 잘 보이지 않았던 희귀 과정의 발견, 어디에나 있으나 보이지 않는 암흑물질의 발견이라는 중요한 임무에 적극적으로 활용될 것으로 보인다.

REFERENCES

- [7] Aristomenis Donos, Jerome P. Gauntlett, Tom Griffin, Nakanin Lohitsiri and Luis Melgar, JHEP 05, 006 (2017).
- [8] <http://tmva.sourceforge.net/>.