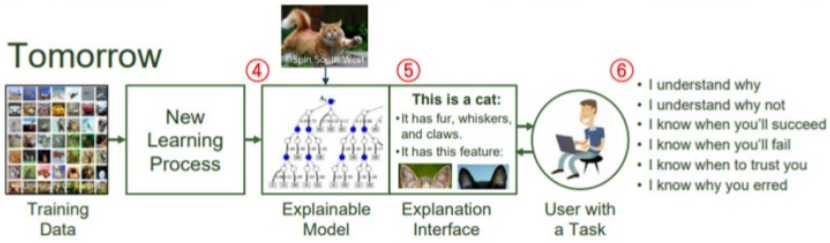
**[XAI]**

**XAI(eXplainable Artificial Ingelligence):** AI의 판단에 대해 **AI 스스로 사람이 이해할 수 있는 형태로 근거를 제시**할 수 있게 하는 기술

**[목표]** AI가 의사결정 이유를 설명하여 **AI의 여러 가지 응용 분야(의료, 금융거래 등)에서의 판단 근거의 신뢰성**을 높일 수 있음

* 알고리즘에 인간이 가진 인종차별 편견이 반영되거나, 오류 및 오작동하는 경우가 있음
* **인공지능의 판단에 대한 이유, 그 판단이 언제 성공/실패하는지, 그 판단을 언제 신뢰할 수 있는지** 알 수 있게 됨

**[개념 상세]** 인공지능의 계산에 대해 그 결과값이 아니라, **인간이 이해 및 해석할 수 있는 설명 인터페이스(explanation interface)를 추가**



**[개발 전략]**

|  |  |
| --- | --- |
| 심층설명학습 | 은닉층의 node가 **의미 있는 속성을 나타내도록 설명 가능한 특징들을 학습**하도록 하는 기술   * 모델의 판단에 대해 활성화된 은닉 노드를 통해 판단 근거 파악 |
| 해석 가능한 모델 | 구조화된, **해석 가능한 인과관계 모델**  **BPL(Bayesian Program Learning):** 작은 조각들의 조합으로 표현하도록 학습 (**글자를 획으로 나누어서 가장 그럴듯한 조합**으로 생성)  **확률적 AND-OR 그래프 모델:** 입력 데이터의 특징을 관계 그래프로 생각, 분류 결과에 연결된 node로부터 분류 근거 확인   * **분류된 이미지와 매핑되는 특징(색, 스케치 등)**을 식별하여 설명 가능 |
| 모델 귀납 | 임의의 블랙박스 모델을 **설명 가능한 모델로 추론**하는 기술  **LME(local interpretable model-agnostic explanations):** 임의의 블랙박스 모델을 **이미 설명 가능한 데이터 주변에서 희소 선형 결합으로 국부적으로 설명 가능**하게 함   * 이미지를 분류하는 블랙박스 모델이 개라고 판단했을 때, **다른 모델의 개에 대한 설명을 이미지와 대조**하여, 개라고 판단한 근거가 되는 부분을 제시 |

**[개발을 위한 기술]**

|  |  |
| --- | --- |
| Sensitivity analysis (SA) | 딥러닝 모델의 **국소적인 입력 변화에 대한 예측 결과의 변화량을 정량화**, 입력의 어떤 부분이 결과 도출에 큰 영향을 미쳤는지 설명 |
| Layer-wise relevance propagation (LRP) | 딥러닝 모델에서 **예측 결과로부터 역전파 형태로 신경망의 각 계층별 기여도 측정**   * 각 계층의 기여도를 시각화하여 직관적 해석 가능 |

**[참고 문헌]**

<http://www.aitimes.kr/news/articleView.html?idxno=14859>

<http://www.ndsl.kr/ndsl/search/detail/report/reportSearchResultDetail.do?cn=KOSEN000000000001071>

<https://www.slideshare.net/jino/2019-report-vol10>

**[AutoML]**

**AutoML(Automatic Machine Learning):** **인공지능을 이용하여 또 다른 인공지능을 만드는** 기술

* 학습데이터 자동 설계, 피처 엔지니어링 자동화, 머신러닝 알고리즘 자동 설계

**[현재의 문제점]**

|  |  |
| --- | --- |
| 고품질 학습 데이터의 부재 | 딥러닝 중 가장 강력한 성능을 보이는 **지도학습에는 인위적인 레이블이 필수**   * 실세계의 머신러닝은 **학습데이터 준비에 많은 비용, 시간 필요** * **레이블링이 불가능**한 AI 문제의 증가 |
| Robust한 AI의 부재 | AI는 **명확히 설명 가능한 정제된 데이터 기준으로 설계**되어 있음   * 실세계는 인과관계가 있는 주요 설명요인 외에도 **노이즈가 있으며, 이 때문에 엉뚱한 결과가 도출됨** * **측정용 센서는 시간이 지남에 따라 성능이 저하**되며, 이것에 의한 동적 상태 변화 발생 (이것은 직접 관측 불가능) |
| 소량 데이터만으로 학습 불가능 | 정확도 높은 결과를 얻으려면 **빅데이터 수준의 데이터가 필요하며, 투자 비용은 실리적 대가에 비해 큼**   * 실세계의 많은 문제는 **소량의 데이터에서 효과**를 얻고자 하며, 따라서 **시행착오**가 많으며 **편향된 샘플**일 가능성 높음 |
| 학습 알고리즘이 문제에 최적이 아님 | 일반적인 수학/통계적 가정들은 **고유의 도메인 및 응용의 동특성 요인과 맞지 않음**   * 따라서 **전문가들도 먼저 데이터 탐색을 수행**하여 적합한 알고리즘을 선정하고 이를 최적화함 * 분석 전문가도 **도메인에 대한 이해**가 필요하며, **전문가의 수에 비해 AI 문제의 수는 다양하게, 급속도로 증가** |

**[AutoML의 각 영역]**

|  |  |
| --- | --- |
| 학습 데이터 자동설계 | 통계적 학습이론은 근본적으로 다음 조건 하에서 수립된 것임   * 훈련(Training, Validation) dataset과 평가(Test) dataset을 **무작위로 충분히 수집** * 각 dataset마다 **데이터가 독립적**임 * 훈련, 평가 dataset마다 **확률분포가 동일**함   랜덤 샘플링을 빅데이터 수준으로 하는 것도 비효율적이므로, **최적의 소량 데이터를 인공지능으로 수집하는 방법에 대한 연구**가 진행됨   * 랜덤 학습데이터보다는 **경계 조건 위주의 25%의 데이터**만으로 90%의 정확도를 보임 * **능동학습(Active Learning):** **무엇을 먼저 학습**할 것인가? |
| 피처 엔지니어링 자동화 | 주어진 **원시 데이터를 모델에 적합하게 변형**  **<분석실무자의 데이터 해석 방법>**  1. **도메인과 데이터의 특징**에 대한 이해  2. EDA(데이터 탐색)를 통하여 **원시데이터**를 통계적 기법, 시각화 기술을 활용하여 **해석**  3. 후보 머신러닝 알고리즘을 이용한 **원시 데이터 변환 방법론 도출**   * **원시데이터, 응용도메인 지식, AI도메인 지식**을 동시에 고려한 최적의 dataset이 만들어져야 함 |
| 머신러닝 알고리즘 자동 설계 | **메타러닝(Meta Learning):** **기계가 스스로 문제에 적합한 알고리즘을 선택**하고 해당 **데이터 분포에 맞게 수정**하는 것   * 선택한 머신러닝 알고리즘의 초매개변수에 대한 튜닝을 위한 시행착오 자동화 * 신경망의 **topology를 스스로 탐색하여 최적의 구조 설계 자동화**   <초매개변수 자동튜닝> |

**[머신러닝 자동화 방법론]**

|  |  |  |
| --- | --- | --- |
| 격자 탐색법 | 각 매개변수별 구간을 임의의 격자로 나누어서 **각 구역별로 학습을 수행하여 성능이 가장 높은 조합 선택** |  |
| 랜덤 탐색법 | 매개변수의 확률분포를 사전에 알고 있을 때, **그 확률분포를 이용하여 샘플링을 하여 성능이 좋은 모델 선택** |

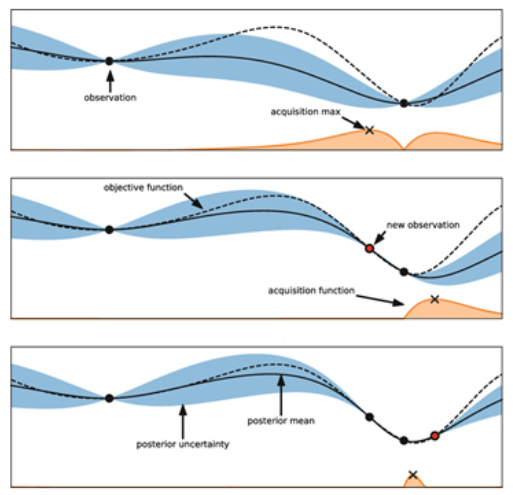
* 수십 개에 달하는 초매개변수 간 상호작용, 학습 모델의 성능 변동을 **동시에 고려하는 요구사항을 충족시키기 위해서 오랜 시간이 소요**됨

**[Hyperparameter (초매개변수) 최적화]**

**수치 최적화(Numerical Optimization):** 반복적 시행착오에 의해 최적화하는 기법으로 강화학습(동적 최적화)과 유사

* 사전 지식을 통해 **매 시행마다 수치 알고리즘/베이지안 확률 모델을 이용하여 더 좋은 초매개변수의 조합**을 찾음

**[순차 근사(모델) 전역 최적화(Sequential Model-Based Global Optimization)]**

초매개변수 공간에서 **가운데 공간을 탐색**하는 방법과 **목적함수(오류율 최소화)에 대한 사전 지식**을 이용하여, **가우시안 회귀와 베이지안 확률 추정을 반복**하여 다음 최적해의 사후 확률을 계산한다. (즉, Bayesian Optimization)

* 고차원 공간에서는 **차원 증가에 따라 탐색 시간이 기하급수적으로 증가**하므로 반드시 선택해야 하는 방법은 아님

**[참고 문헌]**

<https://www.samsungsds.com/global/ko/support/insights/ai_automl.html>