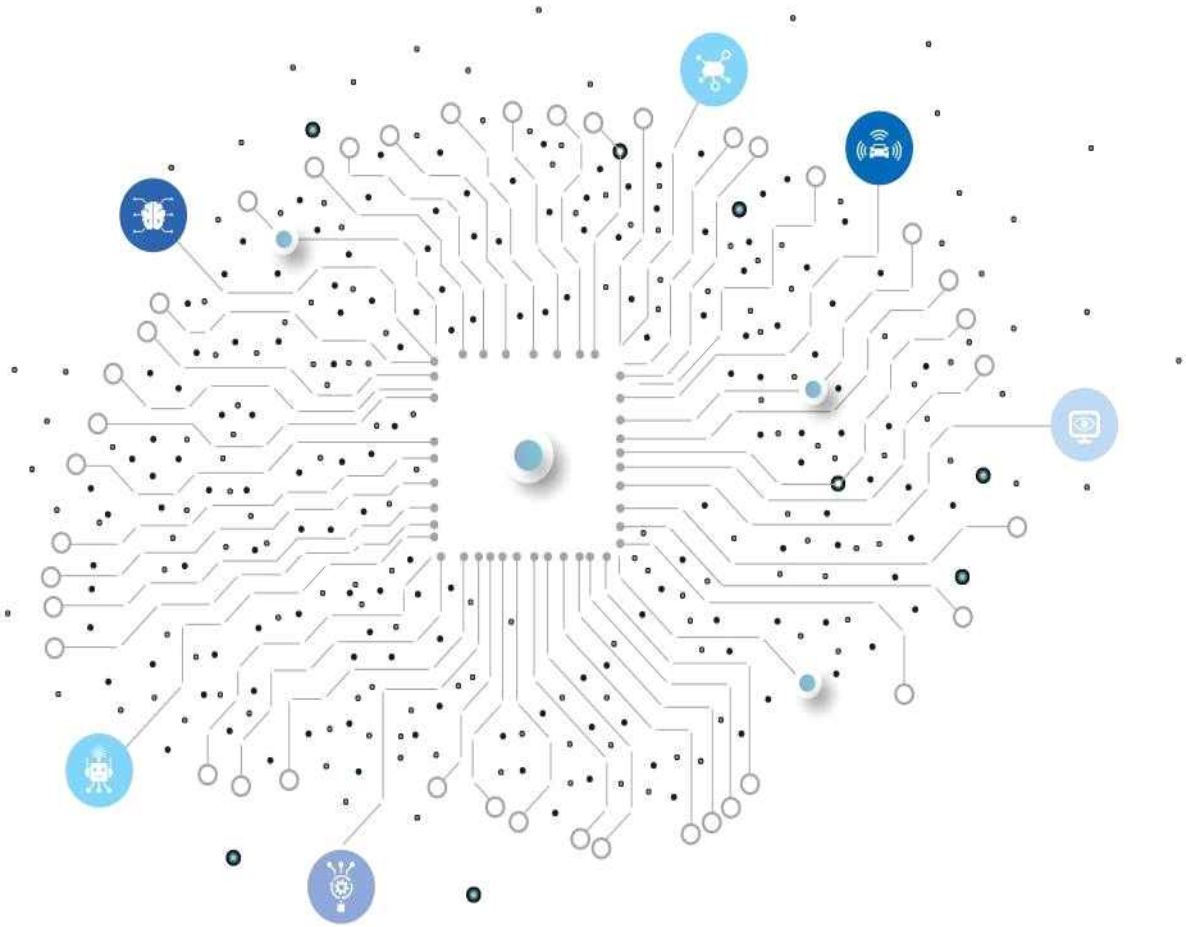
AI Network Lab
INSIGHT
Vol.14
(2022.12)



최신 AI 불확실성 정량화 동향 및 시사점

| 작 성 | 성균관대학교 신지태 교수(jtshin@skku.edu)
성균관대학교 황혜경 연구원(ristar1234@skku.edu)

NIA 한국지능정보사회진흥원
NATIONAL INFORMATION SOCIETY AGENCY

o 『AI Network Lab 인사이트』는 인공지능, 클라우드, 5G 등 4차 산업혁명의 핵심인 지능정보기술과 네트워크 신기술에 대한 동향을 간략하고 심도 있게 분석한 보고서입니다.

o 본 연구보고서는 과학기술정보통신부의 방송통신발전기금조성사업, 한국지능정보사회진흥원의 초연결지능형연구개발망 구축·운영사업의 연구과제 결과이며, 한국지능정보사회진흥원/한국능률협회와 공동 기획하였습니다.

o 본 보고서의 내용은 한국지능정보사회진흥원의 공식 견해와 다를 수 있습니다.

발 행 처 한국지능정보사회진흥원

발 행 인 황종성

기 획 한국지능정보사회진흥원 지능형인프라본부 공공인프라팀

보 고 서 온라인 서비스 www.nia.or.kr

Contents

개요

○ 인공지능의 복잡성이 증가하고 이를 활용하는 산업이 증가함에 따라, 인공지능의 불확실성을 측정하여 신뢰도를 높이는 기술은 국제적으로 큰 관심을 받고 있다. 이러한 불확실성 정량화는 인공지능 시스템의 상태를 관찰하고, 수집된 데이터의 품질을 관리하기 위해 활용될 수 있다. 이러한 기능을 지원하기 위해, 모델 측면에서부터 데이터 측면에 이르기까지 발생 가능한 불확실성의 종류 및 특성을 파악하여 그에 맞는 기반 관측 기술이 제안되고 있다. 이러한 기술은 의료 영상 분석 및 자율주행 등 다양한 산업 분야에서 불확실성을 감소시켜 인공지능 성능 최대화를 도모하며, 모델 의사 결정에 대한 신뢰성을 제공하여 안전한 모델 활용을 가능하게 한다.

○ 이 보고서에서는 딥 뉴럴 네트워크 (Deep Neural Network)의 불확실성 정량화를 위한 기술 소개 및 그 활용에 초점을 맞추고 있다. 특히, 각 불확실성 정량화 범주에 따른 측정 방식을 소개하고, 각 산업에의 적용 사례들을 조사하여 정리하였다. 또한, 불확실성 정량화 기술의 상용화를 위한 중요한 과제 및 미해결 문제를 제안한다.

보 고 서 요 약

(1) 신뢰할 수 있는 인공지능을 위한 AI 불확실성 정량화의 필요성

- 인공지능이 다양한 분야에서 빠르게 적용되고, 그 복잡도가 증가하게 되면서 ‘신뢰할 수 있는 인공지능’에 대한 관심이 증가하고 있다. 이에 따라 인공지능 시스템의 활용 시나리오에서 발생 가능한 상황과 그에 대한 객관적 확률을 측정하고자 하는 불확실성 정량화 기술이 등장하게 되었다. 이는 인공지능 학습에 사용되는 데이터에 내재된 불확실성을 측정하거나, 학습된 모델이 확신하지 못하는 영역을 규명하는 등의 연구가 수행됐으며, 다양한 환경 및 작업에 이를 접목하는 연구가 등장하고 있다.

(2) AI 기술의 불확실성 정량화에 대한 기술 동향

- 최근 인공지능 시스템 불확실성 정량화는 불확실성에 대한 범주 및 특성에 따라 연구되고 있다. 불확실성이란, 적용 관점에 따라 모델 관점의 인식론적 불확실성과 데이터의 내재적 불확실성으로 나뉜다. 인식론적 불확실성은 모분포를 알 수 없는 상태에서 부분 샘플을 기반으로 이루어지는 AI의 경험 중심 학습에서 기인하며, 모델로부터 생성된 예측의 학습 영역 대비 불확실성을 나타낸다. 이를 측정하는 방식은 일반화 오류에 대한 ‘편향-분산’을 조정하는 방법과 학습데이터의 특성에 대한 ‘유사도’를 측정하는 연구가 진행되고 있다. 반면, 내재적 불확실성은 데이터 수집 과정에서 발생하는 불확실성을 의미하며, 데이터 의존적인 이분산성 및 데이터에 독립적인 등분산성 불확실성으로 나뉜다. 전자는 확률적 모델링을 통해 이를 감소시키고자 하는 연구 및 영상분할 작업에서 입력 공간 내 매개변수화를 통한 불확실성 시각화를 위한 연구들이 등장하고 있다. 후자는 인공지능 기술 적용 작업의 불확실성을 드러낼 수 있으므로, 다중 작업을 위한 인공지능 개발에서 확률적 모델링 및 가우스 우도 최대화를 기반으로 작업 간 상대적 신뢰성을 포착 및 제어하기 위한 수단으로 연구되고 있다.

(3) AI 기술 불확실성 정량화의 활용 동향

- 측정된 불확실성은 인공지능 시스템의 개발 단계에서 모델의 성능 제고 및 모델 출력에 대한 신뢰도를 제공할 수 있으며, 실제 확률과 일치하지 않는 예측을 출력하는 인공지능 시스템에 대한 보정을 제공할 수 있다. 또한, 학습 모델의 취약점을 파악하여 추가 학습데이터 선택의 효율성을 증대시키거나, 모델의 취약점을 포함하는 테스트 데이터를 생성하는 등의 서비스를 제공할 수 있다.
- 이러한 특성으로 인해 불확실성 측정법은 의료 영상이나 자율 주행 등 안전과 직결되는 분야에서 활발히 연구되고 있으며, 강화학습이나 동적 학습 등 다양한 연구 분야에서도 그 필요성이 등장하고 있다.

※ 시사점

- 최근 불확실성 파악에 대한 수요에 발맞춘 연구들이 등장해왔으며, 모델 관점에서의 인식론적 불확실성과 데이터의 내재적 불확실성을 파악 및 감소시키고자 하는 시도들이 등장했다. 이는 확률적 모델링을 활용한 수식화 및 앙상블을 활용한 분산분석, 학습 영역과 테스트 영역 사이의 유사도 측정 등을 포함하며, 다양한 분야 및 작업에 적용되어 그 효과를 입증하였다.

- 그러나, 불확실성에 관련한 연구는 평가에 대한 기준이 모호하며, 불확실성 측정 또는 제거 기준에 대한 표준이 성립되어있지 않다는 한계가 있다. 이는 다양한 불확실성 측정 연구의 전체적인 비교 분석을 어렵게 하며, 분야 및 작업에 따라 달라지는 인공지능 기술에 대해, 가장 적합한 불확실성 측정 기법을 찾는 것을 저해하는 요소 중 하나이다. 따라서 각 분야 및 작업에 따른 불확실성 기준 및 표준을 수립하는 것은 신뢰할 수 있는 인공지능에 대한 국가경쟁력을 확보하는 발판이 될 수 있다.

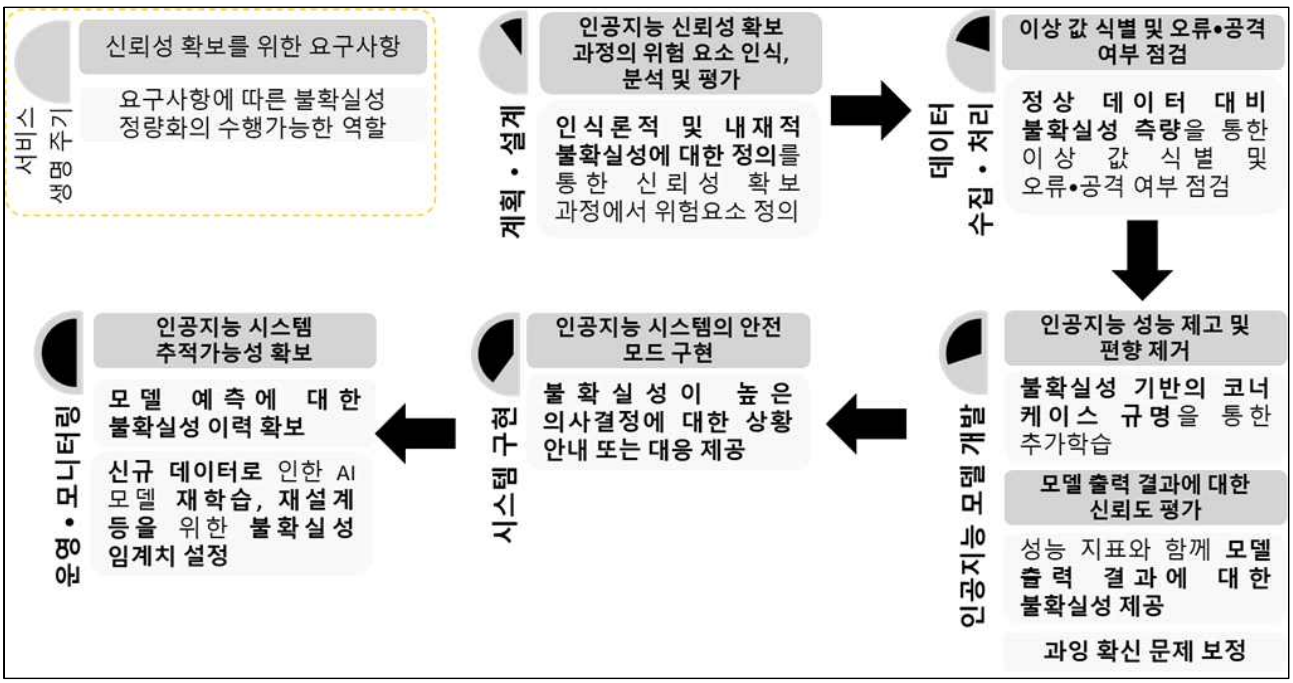
주 요 내 용

I. 신뢰할 수 있는 인공지능을 위한 AI 불확실성 정량화의 필요성

인공지능 기술이 전 산업에 걸쳐 빠르게 도입, 활용 [1] 되면서, 그로 인한 예상치 못한 사회적 이슈 및 우려 [2]도 대두되고 있다. 또한, 인공지능 기술이 고도화될수록 그 메커니즘에 대한 이해가 어려워지는 데 비해, 활용빈도는 점차 늘어나고 있어 인공지능의 오류가 인간의 생활에 밀접한 영향을 끼칠 가능성은 더욱 커지고 있다. 이에 따라, 국제적으로 인공지능의 혜택은 극대화하면서 위험 및 부작용은 최소화할 수 있는 '신뢰할 수 있는 인공지능'을 확보하기 위한 다양한 대응 방안이 마련되고 있다. EU, 미국 등은 일찍이 인공지능 신뢰성을 인공지능 윤리 실천의 핵심요소로서 강조해왔으며, 제도, 윤리, 기술적 측면에서 신뢰성 확보방안을 강구 해왔다. 대한민국 또한 2021년 '사람이 중심이 되는 인공지능을 위한 신뢰할 수 있는 인공지능 실현 전략'을 발표, 2022년 '신뢰할 수 있는 인공지능 개발 안내서(안)'를 배포하는 등의 노력을 기하고 있다. 이러한 실현 전략 및 안내서에 따르면, AI 신뢰성 개념은 인공지능이 내포한 위험과 기술적 한계를 해결하고, 활용·확산 과정에서의 위험·부작용을 방지하기 위한 가치 기준을 포함한다. 또한, 5대 주요 구성요소로서 안전(Safety), 투명성(Transparency), 설명 가능성(Explainability), 견고(Robustness), 공정(Fairness)을 삼는다. 이렇듯 세계적으로 인공지능 기술 경쟁력을 확보에 대한 이목이 모이는 가운데, 인공지능의 신뢰도 확보는 필수적인 행보이다. 이를 위해서는 인공지능의 불확실성을 제거해야 하고 그 첫 단추는 불확실성을 정량화(Uncertainty Quantification)이다.

AI 시스템은 내부의 높은 비선형성으로 인한 블랙박스 특성을 보이며, 그로 인한 불확실성을 내포한다 [3]. 불확실성 정량화는 이를 해석 가능한 정량적 지표로서 표현하는 것을 목표로 하며, 모델 예측과 관련한 모든 분산 정보를 도출 및 요약하는 것을 포함한다. 이는 예측 모델의 한계 및 잠재적 실패 지점을 정의할 수 있다.

따라서, 불확실성 정량화는 효율적인 모델 학습/평가를 위한 데이터의 수집 및 관리, 예측 모델의 과잉 확신에 대한 경계, 모델 학습 시 규제를 통한 성능 제고, 오류 포함 데이터의 검출 등 그 활용방안에 따라 전 인공지능 서비스 생명주기에서 활용 가능하며, 신뢰할 수 있는 인공지능 시스템 확보의 초석이 된다. 그림 1은 AI 시스템 서비스 생명주기에 따른 불확실성 정량화의 적용 가능 영역을 나타낸다.



[그림 1. AI 시스템 서비스 생명주기에 따른 불확실성 정량화의 적용 가능 영역]

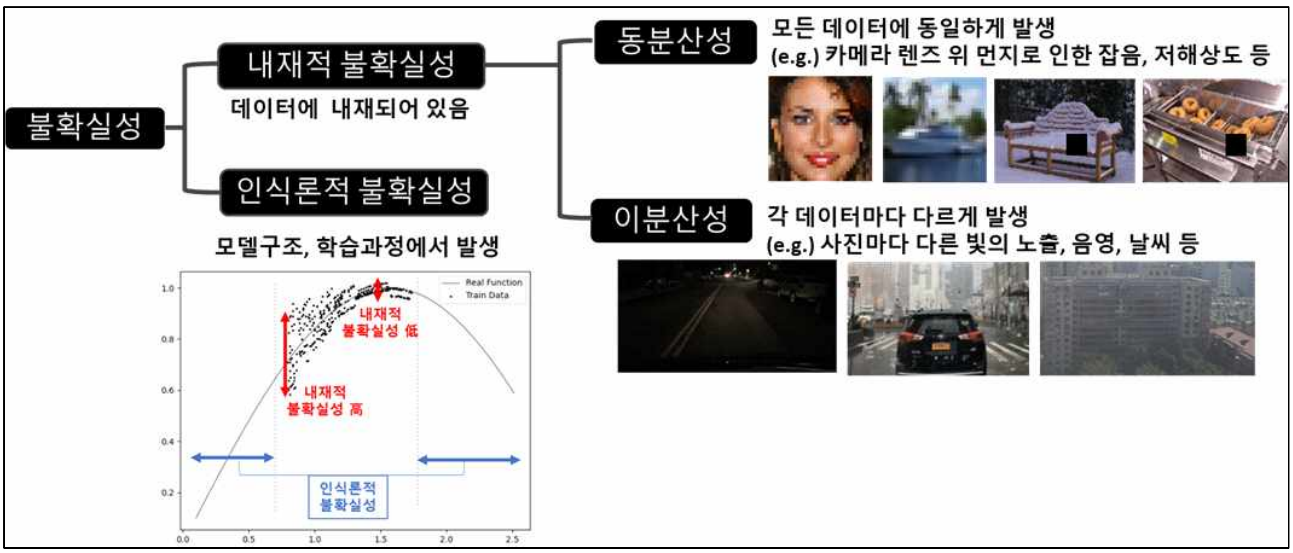
본 보고서에서는 AI 시스템이 내포하는 불확실성의 범주에 대한 소개와 함께 이를 측정하기 위한 최신 연구들은 소개한다. 또한, 각 범주에 따른 불확실성 정량화 기법들의 다양한 산업에의 활용 예를 보이고자 한다.

Ⅱ. AI 기술 불확실성 정량화의 개요

본 글에서는 AI 기술이 포함하는 불확실성의 범주에 대한 소개 및 각각의 불확실성을 정량화하기 위한 최신 연구들을 소개하고자 한다.

불확실성은 크게 모델의 구조나 학습 과정에서 발생하는 불확실성을 나타내는

인식론적 불확실성(Epistemic Uncertainty)과 데이터의 수집 과정에서 발생하는 내재적 불확실성(Aleatoric Uncertainty)으로 나뉜다 [3]. 내재적 불확실성은 모든 데이터에서 같은 값을 가지는 등분산성 불확실성(Homoscedastic Uncertainty)와 각 데이터에 따라 다르게 발생하는 이분산성 불확실성(Heteroscedastic Uncertainty)으로 나뉜다. 그림 2는 각 불확실성의 범주 및 각 범주에 따른 예제를 나타낸다.



[그림 2. 불확실성의 범주 및 각 범주에 따른 불확실성의 예시]

인식론적 불확실성은 모분포를 알 수 없는 상태에서 부분 샘플을 기반으로 이루어지는 AI의 경험 중심 학습에서 기인하며, 모델로부터 생성된 예측의 학습 영역 대비 불확실성을 나타낸다. 따라서 인식론적 불확실성이 높은 데이터는 학습데이터 대비 큰 차이를 보이는 데이터를 의미하며, 모델의 인식론적 불확실성이 높은 경우는 학습데이터의 국소성으로 인해 일반화 성능이 낮아, 추가 학습 필요성이 높다는 것을 의미한다. 이는 모델 적합 능력 개선, 최적화 전략 변경, 매개변수 조정, 학습 데이터 수집 등 일반화 오차를 줄임으로써 감소시킬 수 있으며, 모델 성능 제고와 직결된다. 그러나 학습된 모델의 불확실성을 측정 및 추적하기 위해서는 그 척도를 계량하기 위한 정량화 과정이 동반되어야 한다. 성능 제고의 관점 외에도 완성된 모델의 불확실한 예측을 무시하거나 전문가에게 넘기는 방식의 AI 시스템 운영을 위해서는 불확실성 추정치를 제시하는 것이 중요하다. 따라서 인식론적 불확실성을 측정하는 것은 고성능 인공지능 기술 확보 및 위험도가 높고 안전이 중요한 응용프로그램에서의 인

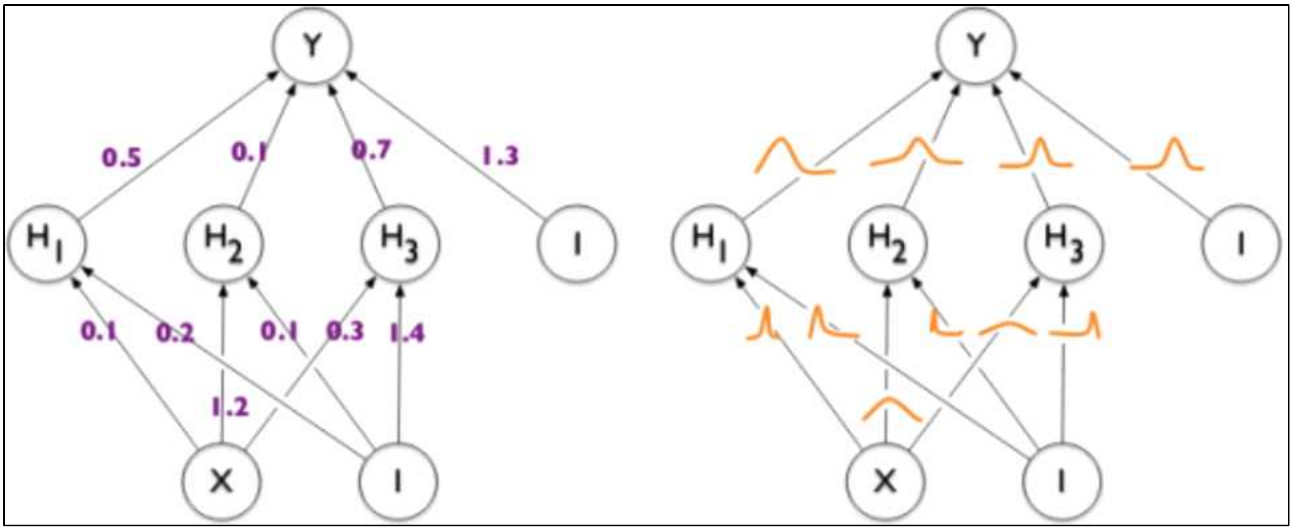
공지능 배포를 위한 필수 조건이다.

내재적 불확실성은 데이터의 수집 과정에서 관측된 잡음으로 인해 이해하지 못하는 정도를 나타내며, 자연적 임의성 (예: 날씨, 조도, 온도 등)과 사용자의 오용 (예: 숫자 6과 비슷하게 쓰인 숫자 5), 전자 전기 시스템의 한계(측정 오류) 등 통제 불가능한 외부 요인에서 기인한다. 이는 데이터 자체가 내포하는 불확실성을 의미하므로 더 많은 데이터를 제공하여 감소시킬 수 없다. 내재적 불확실성이 높은 데이터는 높은 잡음으로 모델 오작동을 유발하며, 학습 과정에서도 모델 학습 속도 및 품질을 저해하는 요인이 된다. 이는 시각적으로 식별되지 않는 섭동을 통해 데이터를 오염시키는 적대적 예시가 등장하면서 그 중요성이 더욱 커지고 있다. 따라서 이를 정량화하여 관측하는 것은 인공지능의 지속적인 활용을 위한 데이터 수집, 관리의 안전장치가 될 수 있으며, 인공지능 기술의 실세계 애플리케이션에 대한 안정적인 성능 및 신뢰도 확보를 위한 필수 조건이다.

(1) 인식론적 불확실성 정량화 기술 동향

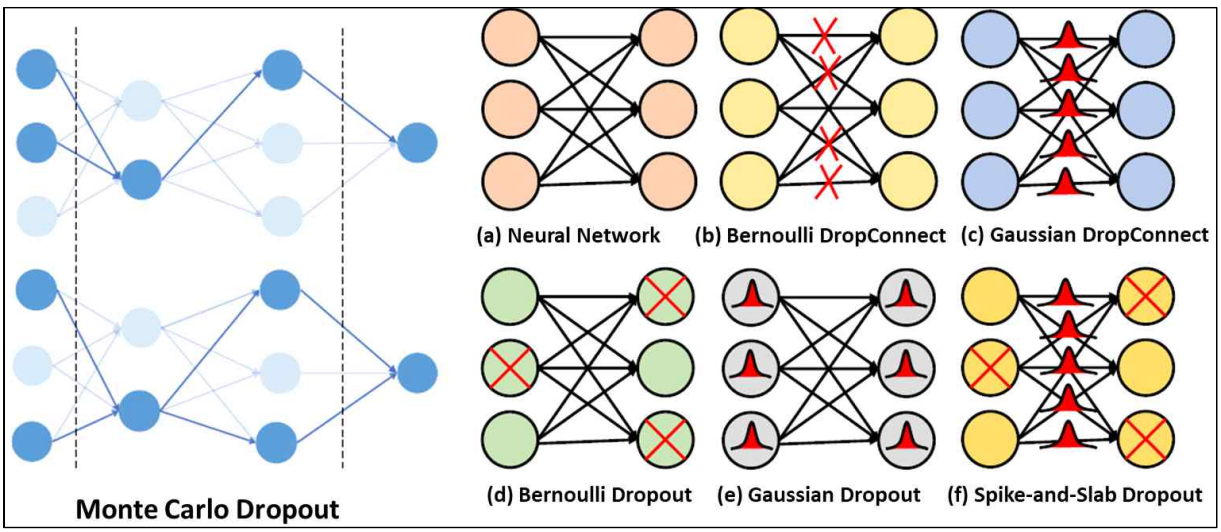
인식론적 불확실성에 대한 관점은 ‘편향-분산(bias-variance)’과 ‘유사도’로 나뉜다. 전자는 지도 학습에서 일반화 오류에 대한 측면에 중점을 둔 인식론적 불확실성 정량화 방법이며, 후자는 모델 학습 종료를 전제하여 예측을 위해 주어지는 입력의 인식론적 불확실성을 측정하는 것을 목표로 한다.

‘편향-분산’ 관점에서의 인식론적 불확실성 정량화에는 베이지안 기법 [3] 과 앙상블 [4] 이 대표적이다. 베이지안 기법은 그림 3의 우측과 같이, 신경망 네트워크의 매개변수에 대한 확률 분포로 가정한다. 이는 불확실성을 측정하기 위해 매개변수를 주변화(marginalize)하여 전체 예측 분포를 형성한다. 이 과정에서 사후확률 추정이 요구되는데, 현대의 신경망 네트워크는 수백만 개의 매개변수를 포함하므로 이를 다루기 위해 변분 추론(Variational Inference, VI), 마르코프 체인 몬테카를로(Markov Chain Monte Carlo, MCMC) 등의 방식이 제안되고 있다.



[그림 3. (좌) 결정론적 매개변수 설정 예시 (우) 매개변수에 확률 분포가 가정된 경우]

변분 추론[5]은 사전 지정된 분포를 사용하여 사후 분포를 추론하는 방식으로, 일련의 매개변수를 최적화하여 모델로부터 얻은 사후 분포와 사전 지정된 분포가 일치시키는 것을 목표로 한다. 초기 변분 추론 기법은 사후 분포와 사전지정분포 사이의 Kullback-Leibler (KL) divergence를 통한 최적화를 수행했으나, 최근 재매개변수화, 가우시안 근사와 함께 베이지안 컨볼루션 신경망을 사용한 방식이 등장하고 있다. 또한, 지도 학습의 손실 함수에 이를 반영하는 연구도 등장하고 있다 [5].

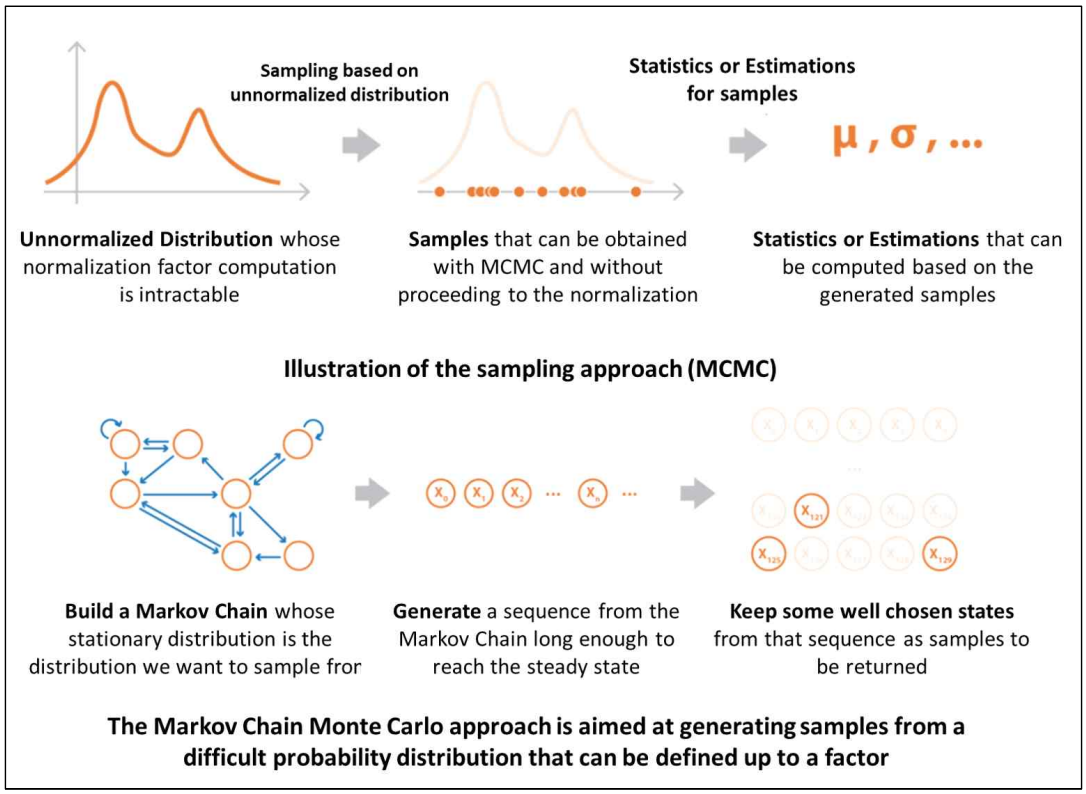


[그림 4. 드롭아웃 기반의 불확실성 추정 기법 방법론 [6]]

그림 4와 같이 드롭아웃을 활용한 다양한 변분 추론법도 제안되고 있다. 이는 모델 테스트 상황에 드롭아웃을 적용하여 입력에 대한 모델 파라미터의 분산을 계산하고

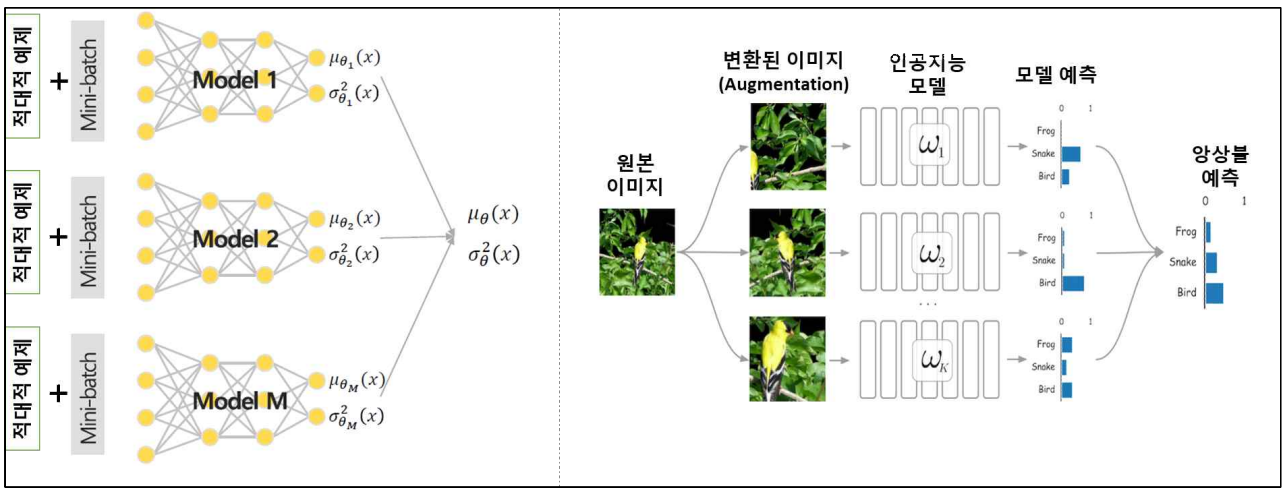
자 했다. 이후 이 아이디어를 확장하여 드롭아웃 확률도 최적화하는 연구가 등장했으며, 드롭아웃을 정규화 항으로써 드롭아웃을 사용한 연구들도 등장해왔다. 드롭아웃을 활용한 모델 불확실성 추론법은 실용적인 관점에서 여러 다운스트림 작업에 성공적으로 적용되었기 때문에 매우 매력적이며, 연산 속도 및 계산 복잡도에서도 유리한 위치에 있다 [6].

마르코프 체인 몬테카를로 [7] 는 임의의 분포에서 샘플을 추출한 후, 현재 상태와 원하는 분포 (예: 실제 사후 분포)에 의해 제어되는 확률적 전환을 수행한다. 즉, MCMC는 반복적인 마르코프 체인 방식으로 샘플을 생성한다. 여기서 마르코프 체인은 한 상태에서 다른 상태로 전환하는 랜덤 변수에 대한 분포를 나타낸다. 각 회차에서 모델은 미리 지정된 규칙에 따라 샘플을 선택하며, 이 과정을 T번 반복하여 마지막으로 생성된 샘플을 통해 원하는 분포를 근사화한다. 그림 5는 [7]에서 소개된 MCMC 기법의 동작 기전을 나타낸다. 최근 복잡해진 예측 모델의 불확실성을 추정하기 위해 베이지안 신경망에 MCMC 기술을 적용하는 연구가 증가했으며, 계산 비용 절감을 위한 미니 배치 전략 및 드롭아웃과의 결합 등의 기법이 제안되고 있다.



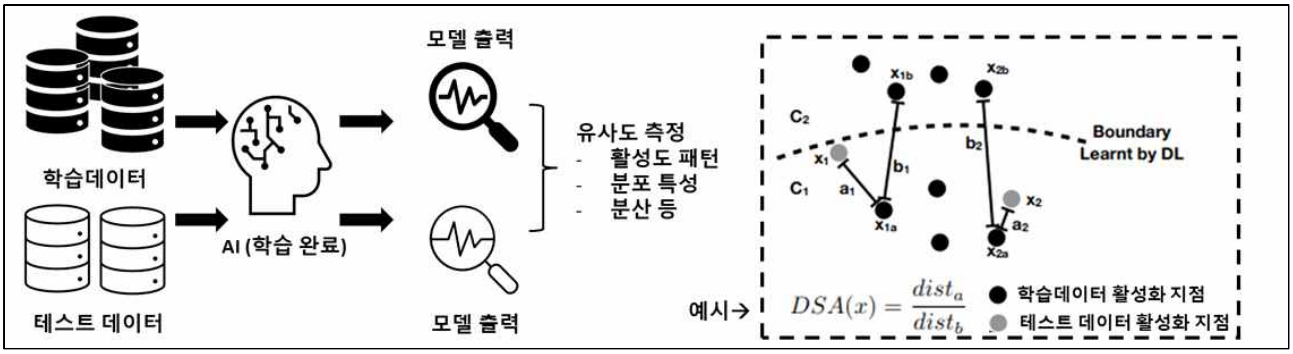
[그림 5. 마르코프 체인 몬테카를로의 기전 [7]]

‘편향-분산’ 관점의 또 다른 기법으로 여러 모델을 훈련하여 예측을 생성한 후, 예측들을 결합하여 최종 출력을 산출하는 앙상블이 있다. 이는 훈련된 여러 기본 모델 간의 상보성을 제어하기 위해 기본 모델들의 출력 값 분산을 인식적 불확실성으로 가정한다. 또한, 앙상블의 본질에 따라 학습 중 앙상블 출력의 바이어스 또는 분산을 줄임으로써 일반화 오류를 줄이고자 한다. 이를 신경망 네트워크 연구에 적용한 대표적인 사례로 [4]를 들 수 있다. 이는 M개의 기본 신경망 모델에 대해 각각 모델 출력에 대한 평균과 분산을 예측하게 하며, 이에 대한 앙상블을 수행한다 (그림 6참조). 이렇듯 앙상블 분산을 감소시키는 것 외에도, 각 모델에 전달되는 미니배치 단위의 데이터와 함께 적대적 예제를 포함하여 추가적인 모델의 견고성을 확보하고자 했다. 이후 [8]는 앙상블 모델의 성능을 평가하기 위해 불확실성을 추정하기 위한 DEE(Deep ensemble Equivalent) 점수를 제안했으며, 이를 통해 앙상블 기법에 포함되는 다수의 기본 모델 중, 앙상블 모델과 동등한 예측을 가진 모델은 소수임을 밝혔다. 또한, 테스트 데이터 증강(Test Time Augmentation)을 통해 이를 개선할 수 있음을 보였다 (그림 6 우측). 앙상블은 다수의 모델을 활용하는 만큼, 그 기반 모델로서 앞서 언급된 베이지안 딥러닝을 활용하는 연구들도 등장하고 있다.



[그림 6. 앙상블을 활용한 불확실성 추정법. 좌: [4]의 앙상블 네트워크, 우: [8]의 앙상블 네트워크]

상기 두 기법과 달리, ‘유사도’ 관점에서의 인식론적 불확실성은 그림 7과 같이 학습이 완료된 모델에 대한 입력 데이터의 불확실성을 측정하고자 한다. 이는 학습데이터 대비 테스트 데이터의 활성도 패턴 유사성, 테스트 데이터에 대한 모델

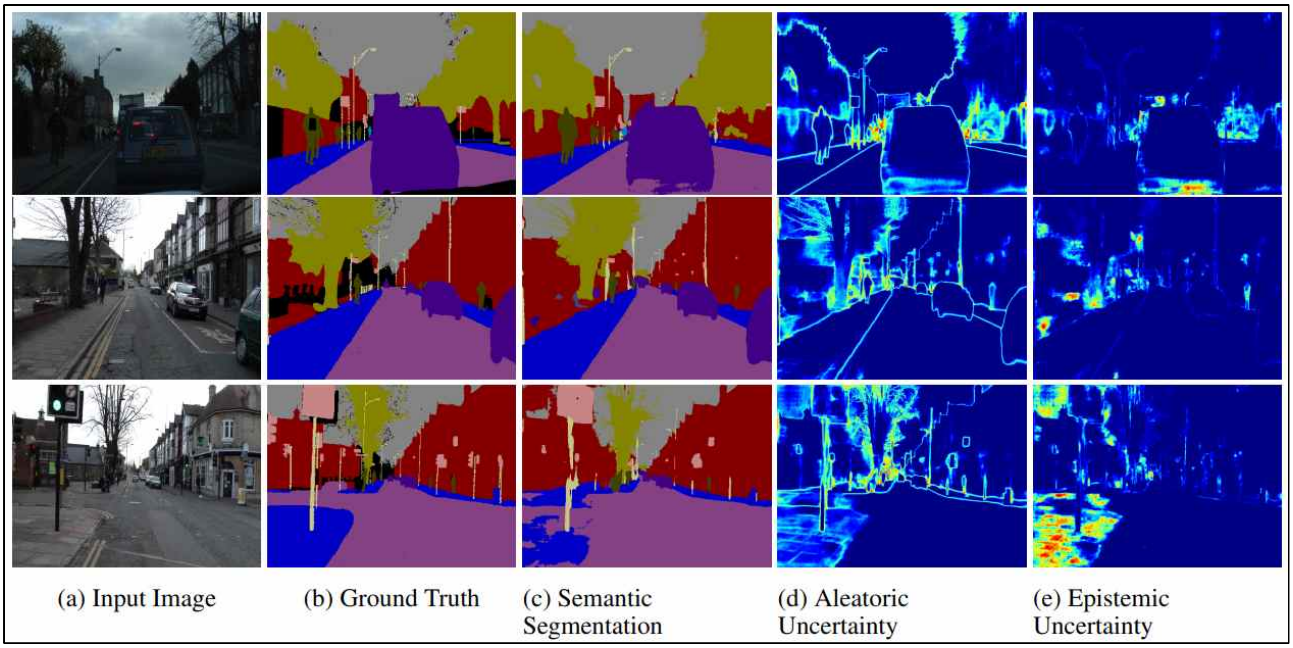


[그림 7. 유사도 관점에서의 인식론적 불확실성 측정법 및 [9]의 불확실성 측정 지표]

출력 분포의 특성, 테스트 데이터에 대한 모델의 드롭아웃을 통한 출력 분산 등을 통해 측정된다 [9]. 이들은 모델이 학습하지 못한 특성을 가진 데이터를 선별할 수 있으므로 추가 학습 과정에 활용될 데이터를 효율적으로 수집하는 데 사용될 수 있다. 또한, 배포된 모델이 사용되는 과정에서 학습하지 못한 특성을 가지거나, 적대적 공격을 받은 데이터와 같이 모델 신뢰도가 낮은 입력에 모델이 사용될 때, 모델 사용에 대한 경계 알림을 주는 등의 지표로써 활용될 수 있다.

(2) 내재적 불확실성 정량화 기술 동향

내재적 불확실성은 데이터 수집 과정에서 발생하는 불확실성을 의미하며, 이분산 및 등분산 불확실성으로 나뉜다. 이분산성 불확실성은 확률적 불확실성이 데이터 의존적이므로, 불확실성이 다양한 입력에 따라 변한다고 가정한다. 따라서 이를 정량화하는 기술을 포함한 모델은 입력 공간의 일부 영역에 대한 불확실성이 더 클 때 유용하게 사용될 수 있다 (예: 일부 숫자가 잘못 쓰이고 출력 클래스가 불확실한 MNIST 데이터 세트). 이분산성 불확실성은 수식화 [3] 되어 분류 및 영상분할 모델의 손실 함수의 연장선으로 사용된 바 있다. 해당 연구에서 잡음에 대한 예상 분산을 매개변수 σ를 통해 모델링 되었다. 매개변수 σ는 입력에 따라 달라지며 모델은 특정 입력이 주어지면 이를 예측하는 방식으로 학습된다. 그림 8은 [3]에서 보고된 입력 이미지에 따른 이분산성 내재적 불확실성 및 인식론적 불확실성을 영상분할 결과 및 그 정답과 함께 시각화한 것이다. 그림에 따르면, 내재적 불확실성을 나타내는 (d)는 객체의 경계 및 카메라에서 멀리 떨어진 물체에 대한 표시를 나타내며, 인식론적 불확실성을 나타내는 (e)에서는 정답 (b)와 모델의 예측을 나타내는



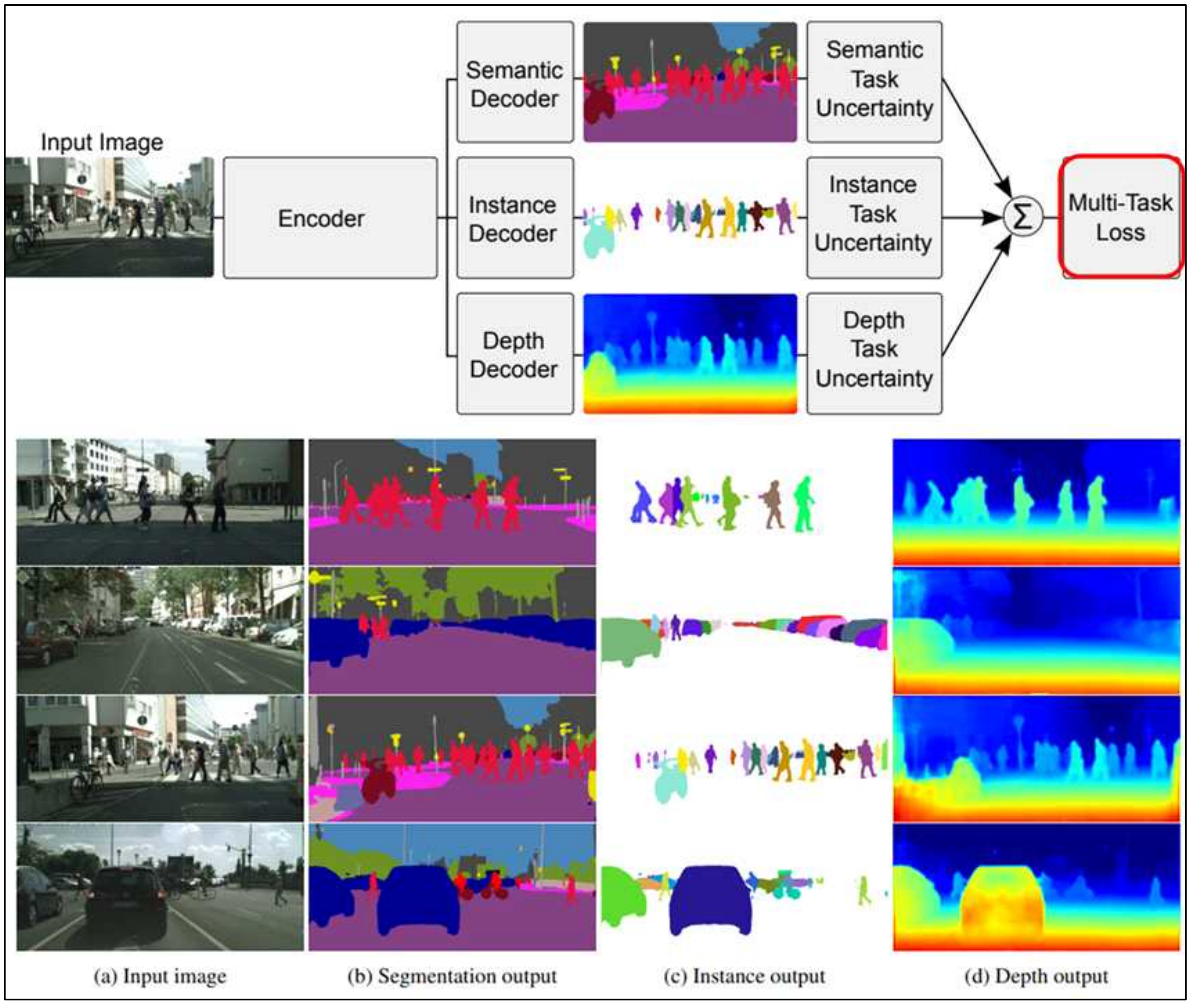
[그림 8. [3]에 보고된 의미론적 영상 분할에 대한 내재적 불확실성과 인식적 불확실성]

(c) 사이에서 차이가 나는 부분, 즉 학습된 모델의 분할 실패 영역과 연관한 부분을 표시함을 확인할 수 있다. 이러한 이분산적 불확실성은 영상 내 객체의 경계영역과 관련하므로, 영상 분할 AI 기술에 이를 적용하고자 하는 연구들이 등장하고 있다.

등분산성 불확실성은 입력 데이터에 의존하지 않는다. 이는 모든 입력 데이터에 대해 일정하게 유지되지만, 서로 다른 작업 사이에서 다른 값을 가질 수 있으므로 작업 종속적 불확실성이라 할 수 있다. 이는 [10]에서 확률적 모델링 및 가우스 우도를 최대화하는 것을 기반으로 수식화 방법이 제안되었으며, 다중 작업에서 각 작업의 불확실성을 반영하여 작업 간 상대적 신뢰를 포착한다는 것을 확인했다. 그림 9는 [10]에서 제안한 네트워크 구조 및 그 결과를 보인다. 해당 기법은 다중 작업 AI 모델 학습 과정에서 각 작업 사이의 균형을 맞추기 위한 가중치로 사용되었다. 이후 다중 작업간 신뢰도 조정 기법은 해상도 확장 및 객체 검출 등에 이를 활용하는 방법들이 등장해왔으며, 이에 대해 자세한 내용은 3장에서 다룬다.

Ⅲ. AI 기술 불확실성 정량화 활용 동향

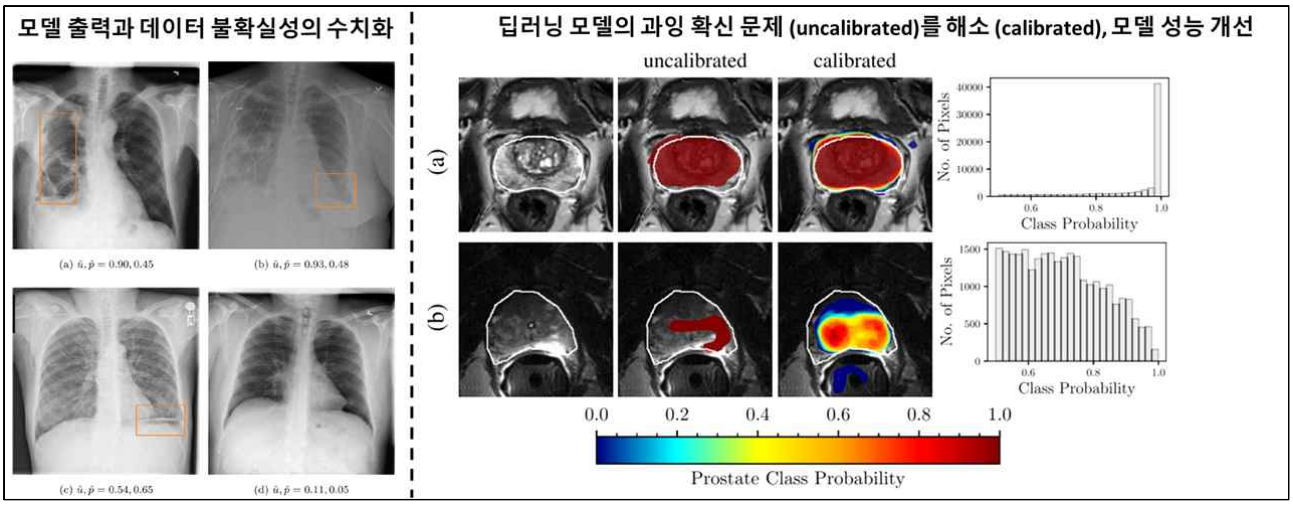
본 글에서는 컴퓨터 비전에서 AI 시스템 서비스 생명주기 중, 인공지능 모델 개발 관점에서 불확실성 정량화를 활용한 최신 연구들을 소개하고자 한다. 이는 모델



[그림 9. 다중 작업 모델에서 등분산성 불확실성을 활용한 [10] 모델 및 그 결과]

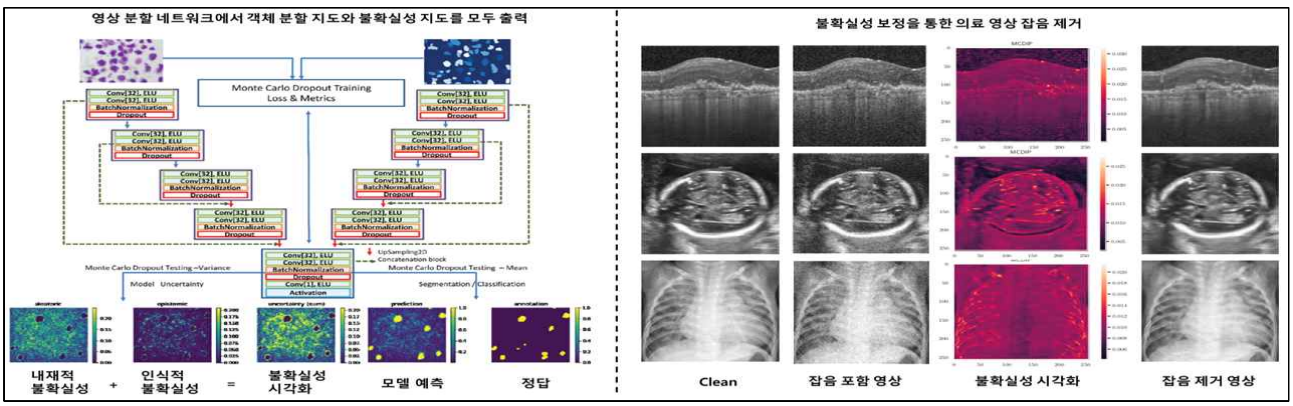
출력에 대한 불확실성 제공 및 과잉 확신의 보정, 미학습 영역에 대한 규명에 활용되고 있다. 모델 출력에 대한 불확실성을 제공 및 과잉 확신 보정 연구는 주로 생명과 직결되는 의료 또는 자율 주행 환경에서 활발한 연구가 진행되고 있다.

의료 환경에서 모델 출력에 대한 불확실성 제공 및 과잉 확신 보정은 의미 그대로의 연구를 수행하면서 모델 성능을 향상시키는 연구도 있지만 [11], 그 외에 영상 분할 네트워크의 불확실성 영역 시각화 [12], 영상 내 불확실성 영역 판단을 통한 잡음 제거 [13] 등에도 활용된다. 그림 10은 [11]에서 제공하는 각 연구의 불확실성 정량화 활용법의 연구 결과를 보인다. 좌측은 모델 출력과 데이터 불확실성을 수치화한 것으로, 주황색 사각형은 모델 의사 결정에 가장 많은 영향을 준 영역을 강조 표시 한 것이다. 이는 흉막 삼출 유무를 신경망으로 분류하고자 하였으며, 좌측그림 (a) ~ (c)는 유체의 비정형적인 외관과 이미지의 낮은 음영 등 데이터 불확실성이 높아,



[그림 10. 의료 영상 분류에서 불확실성 수치화 [11] 및 과잉 확신 보정 [14]의 예]

높은 확률로 흉막 삼출 존재를 추정하고 있으며, 좌측그림 (d)는 흉막 삼출액이 영상에 존재하지 않는 깨끗한 영상으로, 불확실성도 낮고 흉막 삼출 확률도 낮음을 확인할 수 있다. 그림 10의 우측은 딥러닝 모델의 과잉 확신 문제를 인식론적 불확실성 기반 보정을 통해 해소함과 동시에 모델의 예측 정확도와 신뢰도를 상승시킨 사례 [14]를 보인다. 각 그림의 흰색 경계가 예측되어야 하는 영역의 가장자리를 나타낸다. 두 번째 열의 경우, 보정 전 모델의 종양 검출 확률 분포를 나타내며, 이는 경계에 있는 영역에서도 100%에 가까운 확률로 예측함을 확인할 수 있다. 세 번째 열은 이와 상반된 결과를 보이는데, 경계에 대해서는 종양의 심부보다 낮은 확률로 추정함과 동시에 더 정확한 영역 검출을 하는 것을 확인할 수 있다.

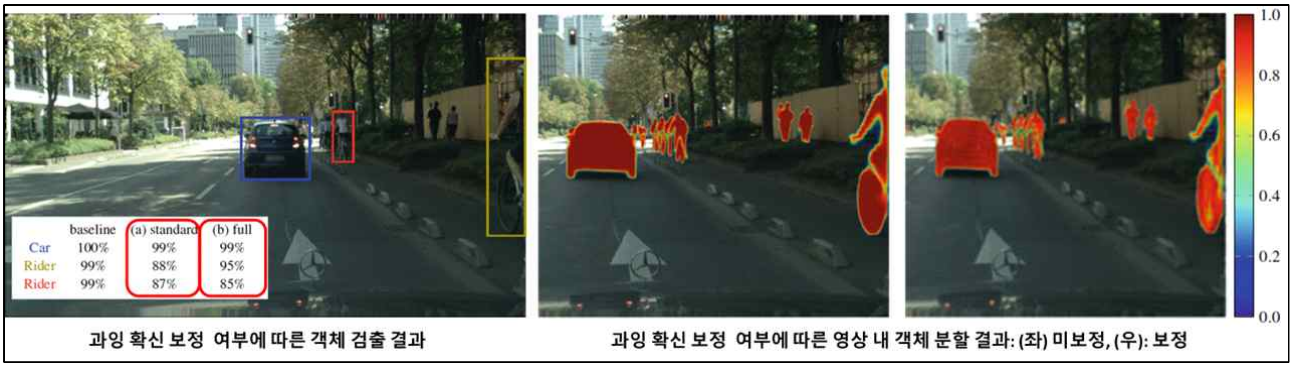


[그림 11. 영상분할에서 불확실성 시각화 [12] 및 불확실성 기반 잡음 제거 [13]의 예]

그림 11은 [12, 13]에서 제공하는 각 연구의 불확실성 정량화 활용법의 연구 결과를 보인다. 이들은 베이지안 딥러닝과 [9]에서 제안한 불확실성 손실 함수 및 드롭아웃 기반의 기법을 적용하여 각각 세포 영상 내 세포 검출, 의료 영상 잡음 제거에 활용했다. 그림 11 좌측은 세포 영상 내 객체 검출 네트워크에서 내재적 불확실성과 인식적 불확실성을 추론하여, 이 둘을 함께 시각화한 것을 불확실성 시각화로 사용한다. 앞서 이야기한 그림 7과 마찬가지로, 인식론적 불확실성 지도는 모델 예측과 정답 사이의 차이, 즉 모델이 잘 검출하지 못한 영역을 주로 나타내며, 내재적 불확실성은 각 세포의 경계 및 배경에 포함된 얼룩 등을 포함한다. 그림 11의 우측은 불확실성 보정을 통해 의료 영상 내 잡음을 제거하는 인공지능 모델의 산출물을 보인다. 이는 잡음을 포함하는 영상에서 센서 잡음과 같은 내재적 잡음 및 잡음 제거 모델 기준에서의 인식론적 불확실성을 픽셀 단위로 추정하여, 이를 제거함으로써 고품질 (Clean) 영상과 유사한 잡음 제거 영상을 만든다.

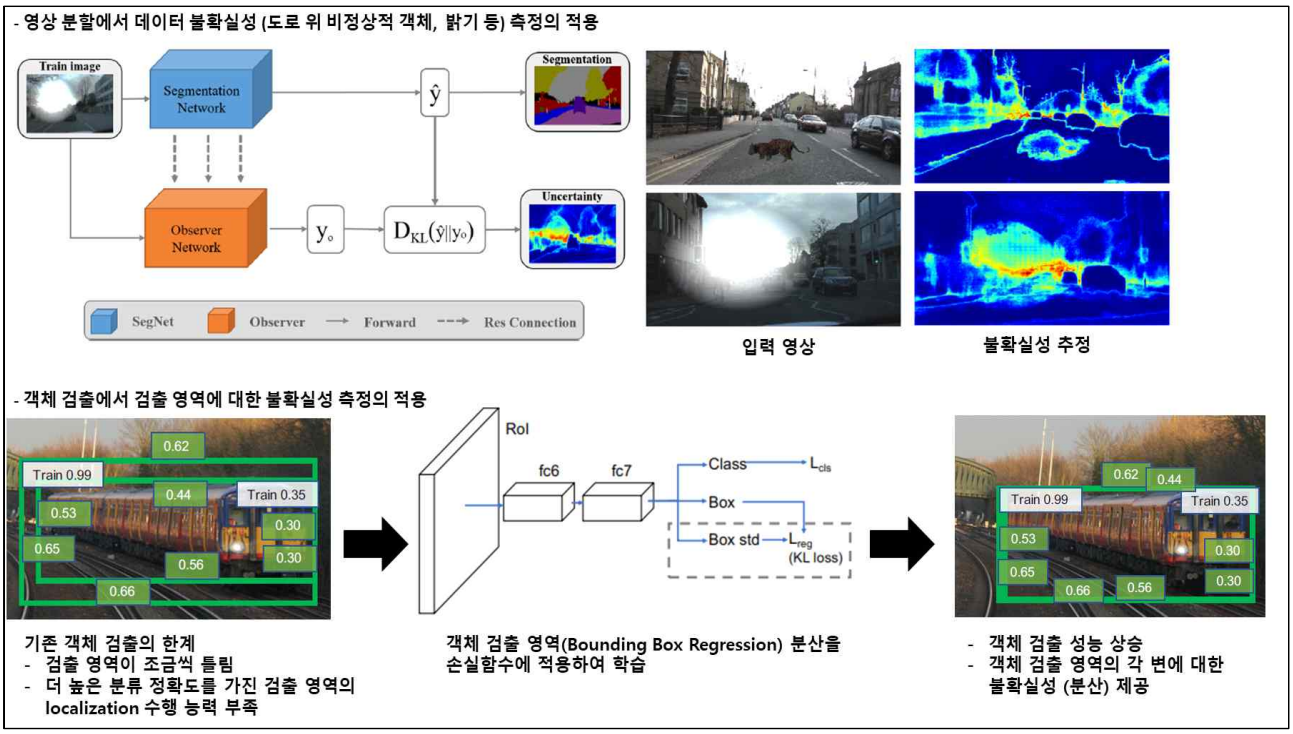
자율 주행 환경에서 모델 출력 결과에 대한 불확실성을 제공하는 연구는 앞선 의료 영상에서와 마찬가지로 수식화를 기반으로 한 불확실성 제한 및 과잉 확신 보정을 통한 성능 제고 [15], 객체 검출의 불확실성 시각화 [16] 등이 있고, 그 외에 자율 주행 환경에서 벌어질 수 있는 데이터 오염 환경 (날씨, 온도, 센서 민감도 등)에 대한 불확실성 측정 [17]이 활발하게 연구되고 있다.

그림 12은 객체 검출 및 분할 모델의 과잉 확신을 보정 하기 위해 분류 문제에서 사용되던 보정법에 위치적 의존성을 추가한 [15]의 결과를 나타낸다. 좌측은 객체 검출에 적용된 결과를 나타내며, 흰색 상자 안에 보정 여부에 따른 모델의 객체 추정 확률을 보인다. ‘baseline’ 은 보정이 이루어지지 않은 상태에서 검출된 객체에 대한 모델의 출력 확률을 나타내며, (a)는 위치적 의존성이 배제된 기존의 보정법을, (b)는 위치적 의존성까지 모두 포함하는 보정법을 적용하였을 때 모델의 출력 확률을 각각 나타낸다. 그림에서 ‘baseline’ 확률을 살펴보면 모두 99% 이상의 확률로 객체를 분류하는 것을 확인할 수 있다. 이는 보정 없이 인공지능을 활용할 경우, 객체의 가려짐, 또는 영상 내 객체의 위치와 무관한 추정을 한다는 점을 보인다. (a), (b)에 대한 확률값에 따르면, 객체의 위치 및 폐색도에 따른 보정이 포함되어야 같은 영상 안에 존재하는 서로 다른 객체들에 대한 모



[그림 12. 자율 주행에서 불확실성 기반 과잉 확신 보정 기법을 적용의 예 [15]]

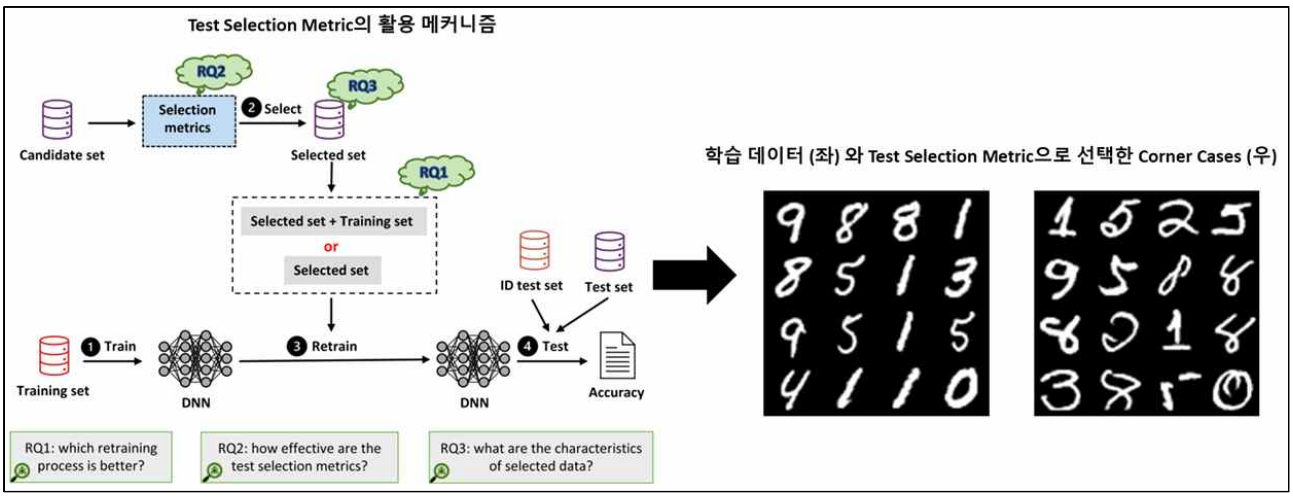
호성을 반영한 인공지능 기술이 완성될 수 있음을 시사한다. 이와 관련하여 IBM에서는 모델의 불확실성을 측정하기 위한 다양한 소프트웨어를 통합한 Uncertainty Quantification 360을 배포한 바 있다.



[그림 13. 자율 주행에서 데이터 오염도 예측을 통한 불확실성 시각화 [17] 및 객체 검출 불확실성 시각화 [16]]

그림 13은 자율 주행을 위한 객체 검출 및 분할 모델에서 불확실성 정량화의 활용을 보인다. 상부 그림은 [17]의 자율 주행 상황에서 발생 가능한 고 불확실성 상황(예: 갑작스러운 물체의 떨어짐, 조도로 인한 카메라 영상의 시야 확보 어려움 등)을

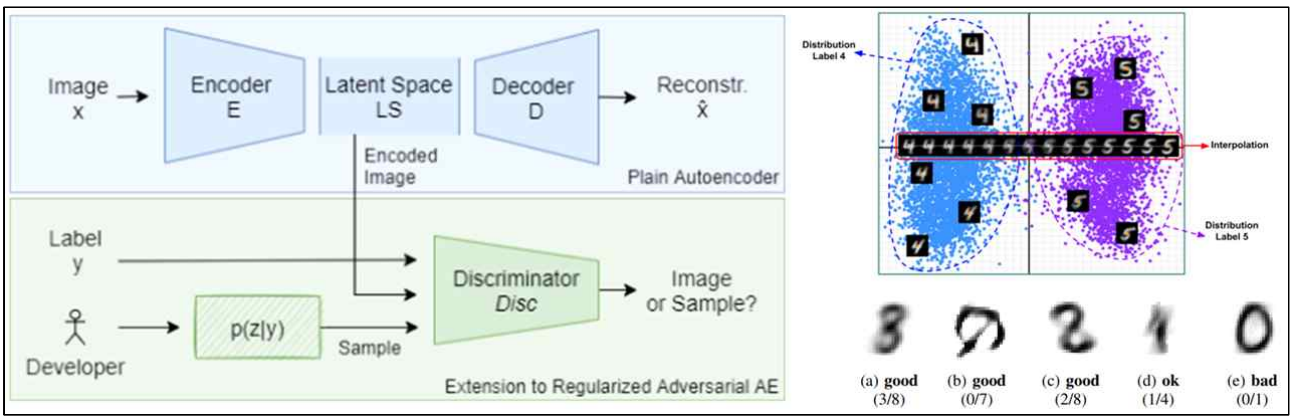
시각화 지도를 통해 표현하는 방법을 보이며, 그림 13 아래측은 [16]에서 제안한 자율 주행에 필수 기술인 객체 검출에 대한 모델의 불확실성을 측정 및 손실 함수화를 수행한 네트워크의 적용 전후를 보인다. 이는 객체 검출에 사용되는 경계 상자 회귀에 대한 분산을 추정하고, 학습 과정에서 이를 줄이도록 하여, 적용 단계에서의 경계 상자 회귀 정확도를 높이고자 했다. 또한, 경계 상자의 각 변에 대한 분산을 제공하여 예측된 경계 상자의 변동 범위를 제공하여 모델 출력에 대한 불안정성을 동시에 제공하고자 했다. 이러한 기술은 향후 자율 주행이 더욱 활발히 적용될 때, 고위험상황에 대한 자율 주행 시스템 사용 중단 또는 자율 주행 시스템의 안정성에 대한 추적을 위한 자료로 사용될 수 있다.



[그림 14. [18]에 소개된 테스트 선택 메커니즘 및 [9]를 사용하여 선택된 MNIST 테스트 데이터]

인공지능 모델 개발 관점에서 미학습 영역에 대한 규명을 위한 연구는 학습데이터 또는 학습된 특성 대비 불확실성이 높은 데이터를 선택하는 테스트 데이터 선택 지표 개발(Test Selection Metric) [18] 과 모델의 결함을 드러낼 수 있는 테스트 케이스를 생성하는 것 (Test Generation) [19] 으로 나눌 수 있다. 두 방식이 가지는 가장 큰 차이점은 데이터를 선택할 후보군의 존재이다. 전자의 경우, 일련의 확보된 데이터를 기준으로, 학습된 모델에 대한 각 데이터의 불확실성을 측정한다. 이때, 불확실성이 높을수록 테스트 선택 지표 값이 크며, 이는 곧 모델이 학습하지 못한 코너 케이스에 가깝다는 것을 의미한다. 따라서, 학습된 모델 기준에서 테스트 선택 지표 값이 큰 데이터들을 채택하여 이들을 학습데이터에 편입시켜, 다시 모델을 학

습하며 모델의 일반화 성능을 개선 시키는 것을 목표로 한다. 이러한 일련의 과정은 그림 14 좌측에 묘사되어있으며, 우측에는 MNIST 손글씨 데이터의 학습데이터로 훈련된 모델에서 테스트 데이터 중 테스트 선택 지표 [9]가 가장 큰 값을 가지는 이미지들을 보인다. 이들은 직접 보기에도 학습데이터와 대비하여 다른 형태를 보이는 것(예: 모양이 다른 1, 윗부분까지 제대로 닫히게 적히지 않은 8 등)을 확인할 수 있다.



[그림 15. 내재적, 인식적으로 모두 모호한 데이터를 생성하는 [20]의 네트워크 및 생성된 데이터의 예]

후자의 경우, 학습한 모델의 특성 공간을 기반으로 하여, 모델이 오작동을 일으킬 수 있는 입력을 재건하는 것을 목표하므로, 전자와 달리 데이터 단위에서의 후보군을 필요로 하지 않는다. 그림 15는 일반 분류 문제에서 이러한 입력 재건을 수행하는 구조를 보인다. 이는 모호한 입력을 생성해내는 것에 주목하여 테스트 데이터를 생성[20]하며, 생성된 데이터를 학습데이터에 편입시켜 모델을 추가 학습하여 기존의 모델보다 더 견고한 모델을 얻을 수 있다. 자율 주행 환경에 대한 입력 재건은 일반 분류와 사뭇 다른 방향성을 추구하는데, 이는 그림 16에 묘사된다. 자율 주행 환경에서는 저조도, 안개, 비, 눈 등의 날씨로 인한 데이터 품질 저하와 같은 사전 정의된 고위험 환경이 존재하므로, 자율 주행 환경을 대상으로 테스트 데이터를 생성할 때에는 저위험 환경(예: 햇볕이 화창한 날, 낮에 촬영된 사진)의 사진을 고위험 환경으로 변환하거나, 고위험 환경끼리의 결합 (예: 밤 사진과 눈이 오는 풍경 사진의 합성, 화각 오류 등)을 수행한다. 그림 16는 [19]에서 사용된 테스트 데이터 생성 모델 개요 및 이를 통해 고위험군 자율 주행 환경 데이터를 생성해 낸 예시를 나타낸다.

[그림 16. 자율 주행 환경에서 불확실성이 높은 테스트 데이터를 생성해내는 [19]의 모델 및 생성된 데이터의 예]

Ⅳ. 결론 및 시사점

- 인공지능의 활용 범위가 넓어지고 다양해짐에 따라 국제적으로 '신뢰할 수 있는 인공지능'을 확보하기 위한 다양한 대응 방안이 마련되고 있는 가운데, 인공지능의 불확실성을 측정하고, 개선하는 것은 인공지능의 여러 적용 분야로부터 큰 관심을 받고 있다.
- 이러한 불확실성 파악에 대한 수요에 발맞춘 연구들이 등장해왔으며, 모델 관점에서의 인식론적 불확실성과 데이터의 내재적 불확실성을 파악 및 감소시키고자 하는 시도들이 등장했다. 이는 확률적 모델링을 활용한 수식화 및 앙상블을 활용한 분산분석, 학습 영역과 테스트 영역 사이의 유사도 측정 등을 포함하며, 다양한 분야 및 작업에 적용되어 그 효과를 입증하였다.
- 그러나, 불확실성에 관련한 연구는 평가에 대한 기준이 모호하며, 불확실성 측정 또는 제거 기준에 대한 표준이 성립되어있지 않다는 한계가 있다. 이는 다양한 불확실성 측정 연구의 전체적인 비교 분석을 어렵게 하며, 분야 및 작업에 따라 달라지는 인공지능 기술에 대해, 가장 적합한 불확실성 측정 기법을 찾는 것을 저해하는 요소 중 하나이다. 따라서 각 분야 및 작업에 따른 불확실성 기준 및 표준을 수립하는 것은 신뢰할 수 있는 인공지능에 대한 국가경쟁력을 확보하는 발판이 될 수 있다.

참 고 문 헌

[1] Zhang, Caiming, and Yang Lu. “Study on artificial intelligence: The state of the art and future prospects.“ Journal of Industrial Information Integration 23 (2021): 100224.

[2] Nasir, Vahid, and Farrokh Sassani. “A review on deep learning in machining and tool monitoring: methods, opportunities, and challenges.“ The International Journal of Advanced Manufacturing Technology 115.9 (2021): 2683-2709.

[3] Kendall, Alex, and Yarin Gal. “What uncertainties do we need in bayesian deep learning for computer vision?.“ Advances in neural information processing systems 30 (2017).

[4] Jain, Siddhartha, et al. “Maximizing overall diversity for improved uncertainty estimates in deep ensembles.“ Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 04. 2020.

[5] Drori, Iddo. “Deep variational inference.“ Handbook of Variational Methods for Nonlinear Geometric Data. Springer, Cham, 2020. 361-376.

[6] Abdar, Moloud, et al. “A review of uncertainty quantification in deep learning: Techniques, applications and challenges.“ Information Fusion 76 (2021): 243-297.

[7] https://towardsdatascience.com/bayesian-inference-problem-mcmc-and-variational-inference-25a8aa9bce29

[8] Ashukha, Arsenii, et al. “Pitfalls of in-domain uncertainty estimation and ensembling in deep learning.“ arXiv preprint arXiv:2002.06470 (2020).

[9] Kim, Jinhan, Robert Feldt, and Shin Yoo. “Guiding deep learning system testing using surprise adequacy.“ 2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE). IEEE, 2019.

[10] Kendall, Alex, Yarin Gal, and Roberto Cipolla. “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics.“ Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[11] Ghesu, Florin C., et al. “Quantifying and leveraging predictive uncertainty for medical image assessment.“ Medical Image Analysis 68 (2021): 101855.

[12] Ghoshal, Biraja, et al. “Estimating uncertainty in deep learning for reporting confidence to clinicians in medical image segmentation and diseases detection.“ Computational Intelligence 37.2 (2021): 701-734.

[13] Laves, Max-Heinrich, Malte Tölle, and Tobias Ortmaier. “Uncertainty estimation in medical image denoising with bayesian deep image prior.“ Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis. Springer, Cham, 2020. 81-96.

[14] Mehrtash, Alireza, et al. “Confidence calibration and predictive uncertainty estimation for deep medical image segmentation.“ IEEE transactions on medical imaging 39.12 (2020): 3868-3878.

[15] Küppers, F., Haselhoff, A., Kronenberger, J., Schneider, J. (2022). Confidence Calibration for Object Detection and Segmentation. In: Fingscheidt, T., Gottschalk, H., Houben, S. (eds) Deep Neural Networks and Data for Automated Driving. Springer, Cham.

[16] He, Yihui, et al. “Bounding box regression with uncertainty for accurate object detection.“ Proceedings of the ieee/cvf conference on computer vision and pattern recognition. 2019.

[17] Mehrtash, Alireza, et al. “Confidence calibration and predictive uncertainty estimation for deep medical image segmentation.“ IEEE transactions on medical imaging 39.12 (2020): 3868-3878.

[18] Hu, Qiang, et al. “An empirical study on data distribution-aware test selection for deep learning enhancement.“ ACM Transactions on Software Engineering and Methodology (2022).

[19] Zhang, Mengshi, et al. “DeepRoad: GAN-based metamorphic testing and input validation framework for autonomous driving systems.“ 2018 33rd IEEE/ACM International Conference on Automated Software Engineering (ASE). IEEE, 2018.

[20] Weiss, Michael, André García Gómez, and Paolo Tonella. “A Forgotten Danger in DNN Supervision Testing: Generating and Detecting True Ambiguity.“ arXiv preprint arXiv:2207.10495 (2022).