(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0028770
(43) 공개일자 2025년03월04일

(51) 국제특허분류(Int. Cl.)
G06N 3/045 (2023.01) *G06N 20/20* (2019.01)
G06N 3/042 (2023.01) *G06N 3/047* (2023.01)
G06N 3/048 (2023.01)
(52) CPC특허분류
G06N 3/045 (2023.01)
G06N 20/20 (2021.08)
(21) 출원번호 10-2023-0109846
(22) 출원일자 2023년08월22일
심사청구일자 2023년08월22일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
이상근
서울특별시 강남구 선릉로 221(도곡동, 도곡렉슬아파트)
전어진
서울특별시 성북구 안암로 35 - 5 (안암동4가) 501호
이민규
서울특별시 관악구 문성로38길 26 - 16 (신림동) 301호
(74) 대리인
윤귀상

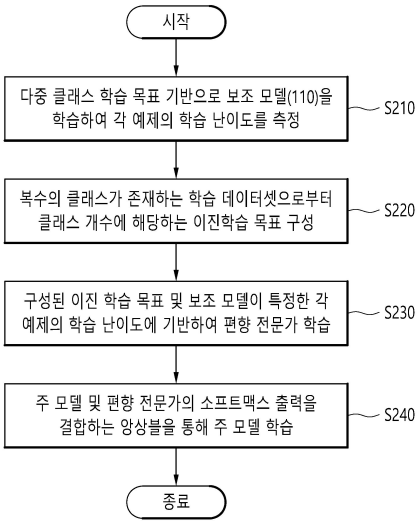
전체 청구항 수 : 총 9 항

(54) 발명의 명칭 **편향완화를 위한 편향 전문가 기반의 편향 식별 장치, 방법 및 이를 기록한 컴퓨터로 판독가능한 기록매체**

(57) 요 약

본 발명은 복수의 클래스의 학습 목표 기반으로 보조 모델을 학습하여 각 예제의 학습 난이도를 측정하는 단계; 상기 복수의 클래스가 존재하는 학습 데이터셋으로부터 상기 복수의 클래스 개수에 해당하는 이진 학습 목표를 구성하는 단계; 상기 학습 난이도 및 상기 이진 학습 목표에 기반하여 편향 전문가가 학습하는 단계; 주 모델 및 상기 편향 전문가의 소프트맥스 출력을 결합하는 앙상블을 통해 상기 주 모델을 학습하는 단계를 포함하는, 편향 완화를 위한 편향 전문가 기반의 편향 식별 방법에 관한 것이다. 이에 의해 보조모델의 학습 목표와 그것의 본래 목적 사이의 괴리를 완화함으로써 편향 예제를 종래 기술 대비 더 정확하게 식별하고, 이를 통해 주 모델의 편향 완화 성능을 개선할 수 있다.

대 표 도 - 도2



(52) CPC특허분류
G06N 3/042 (2023.01)
G06N 3/047 (2023.01)
G06N 3/048 (2023.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711193973
과제번호 2019-0-00079-005
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송혁신인재양성
연구과제명 인공지능대학원지원(고려대학교)
기 여 율 1/2
과제수행기관명 고려대학교산학협력단
연구기간 2023.01.01 ~ 2023.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711188552
과제번호 2021R1A2C3010430
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 개인기초연구(과기정통부)
연구과제명 미플랫폼: 자기지도학습과 연합학습으로 강화되는 작은 인공지능 플랫폼
기 여 율 1/2
과제수행기관명 고려대학교
연구기간 2023.03.01 ~ 2024.02.29

명 세 서

청구범위

청구항 1

복수의 클래스의 학습 목표 기반으로 보조 모델을 학습하여 각 예제의 학습 난이도를 측정하는 단계;

상기 복수의 클래스가 존재하는 학습 데이터셋으로부터 상기 복수의 클래스 개수에 해당하는 이진 학습 목표를 구성하는 단계;

상기 학습 난이도 및 상기 이진 학습 목표에 기반하여 편향 전문가가 학습하는 단계; 및

주 모델 및 상기 편향 전문가의 소프트맥스 출력을 결합하는 앙상블을 통해 상기 주 모델을 학습하는 단계;를 포함하는, 편향완화를 위한 편향 전문가 기반의 편향 식별 방법.

청구항 2

제 1항에 있어서,

상기 각 예제의 학습 난이도를 측정하는 단계는,

상기 보조 모델이 상기 각 예제의 학습 스텝마다 상기 각 예제의 정답 클래스에 부여한 확률을 측정하는 단계인, 편향완화를 위한 편향 전문가 기반의 편향 식별 방법.

청구항 3

제 1항에 있어서,

상기 이진 학습 목표를 구성하는 단계는,

기계학습의 One-vs-Rest 접근법에 따르는 단계인, 편향완화를 위한 편향 전문가 기반의 편향 식별 방법.

청구항 4

제 1항에 있어서,

상기 편향 전문가가 학습하는 단계는,

학습시에, 목표 클래스인 특정 클래스에 해당하는 편향 예제에는 높은 학습 가중치를 부여하고, 나머지 클래스에 해당하는 편향 예제에는 낮은 학습 가중치를 부여하는 단계인, 편향완화를 위한 편향 전문가 기반의 편향 식별 방법.

청구항 5

제1항 내지 제4항 중 어느 하나의 항에 따른 상기 편향완화를 위한 편향 전문가 기반의 편향 식별 방법을 수행하기 위한 컴퓨터 프로그램이 기록된 컴퓨터로 판독 가능한 저장 매체.

청구항 6

편향완화를 위한 편향 전문가 기반의 편향 식별 장치로서,

프로세서;

상기 프로세서에 연결되는 메모리를 포함하되,

상기 메모리는,

복수의 클래스의 학습 목표 기반으로 보조 모델을 학습하여 각 예제의 학습 난이도를 측정하고,

상기 복수의 클래스가 존재하는 학습 데이터셋으로부터 상기 복수의 클래스 개수에 해당하는 이진 학습 목표를 구성하고,

상기 학습 난이도 및 상기 이진 학습 목표에 기반하여 편향 전문가가 학습하며,

주 모델 및 상기 편향 전문가의 소프트맥스 출력을 결합하는 앙상블을 통해 상기 주 모델을 학습하도록, 상기 프로세서에 의해 실행되는 프로그램 명령어들을 저장하는, 편향완화를 위한 편향 전문가 기반의 편향 식별 장치.

청구항 7

제 6항에 있어서,

상기 각 예제의 학습 난이도를 측정하는 것은,

상기 보조 모델이 상기 각 예제의 학습 스텝마다 상기 각 예제의 정답 클래스에 부여한 확률을 측정하는 것인, 편향완화를 위한 편향 전문가 기반의 편향 식별 장치.

청구항 8

제 6항에 있어서,

상기 이진 학습 목표를 구성하는 것은,

기계학습의 One-vs-Rest 접근법에 따르는 것인, 편향완화를 위한 편향 전문가 기반의 편향 식별 장치.

청구항 9

제 6항에 있어서,

상기 편향 전문가가 학습하는 것은,

학습시에, 목표 클래스인 특정 클래스에 해당하는 편향 예제에는 높은 학습 가중치를 부여하고, 나머지 클래스에 해당하는 편향 예제에는 낮은 학습 가중치를 부여하는 것인, 편향완화를 위한 편향 전문가 기반의 편향 식별 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 인공지능의 자연어 해석에 관한 것으로서, 더욱 상세하게는 자연어 해석 작업에서의 비편향 분류를 위해 편향 전문가 분류기를 이용하여 데이터셋의 편향을 식별하는, 편향완화를 위한 편향 전문가 기반의 편향 식별 방법, 장치 및 이를 기록한 컴퓨터로 판독가능한 기록매체에 관한 것이다.

배 경 기 술

[0003] 주어진 입력 (Input)을 복수의 클래스 중 하나로 분류하는 작업을 가정한다. 해당 작업을 수행할 분류기를 학습하는 데 사용하는 학습 데이터셋 (Training Dataset)은 일정 개수의 예제 (example)로 구성되어 있으며, 각 예

제는 입력과 그에 상응하는 클래스의 순서쌍이다. 이때 데이터셋 편향(Dataset Bias)은 입력에 나타나는 피상적인 실마리로서, 주어진 학습 데이터셋에서는 클래스와 높은 상관관계를 가지는 반면, 그 외의 데이터셋에서는 작업 수행에 도움을 주지 않는 실마리를 가리킨다. 예를 들어, 두 개의 문장으로 주어진 입력에 대하여 해당 두 문장 간 관계를 '함의', '중립', '모순' 클래스 중 하나로 분류하는 다중장르 자연어 추론 (MNLI, Multi-Genre Natural Language Inference) 데이터셋에서는 두 번째 문장에 'no', 'not'과 같은 부정 표현이 등장할 시 두 문장이 모순 관계일 확률이 높은 것으로 알려져 있다. 이 경우 데이터셋 편향은 부정 표현의 존재이며, 해당 데이터셋으로 학습한 분류기들은 문장의 의미를 토대로 그 관계를 도출하는 것이 아니라, 단순히 부정 표현의 존재만으로 '모순' 관계를 예측하고 이를 통해 높은 성능을 달성하곤 한다. 이같이 데이터셋 편향에 근거하여 정답 클래스를 예측하는 분류기는 학습 데이터셋에서의 상관관계가 성립하지 않는 다른 데이터셋에서의 성능이 크게 떨어진다. 앞서 언급하였듯 데이터셋 편향은 학습 데이터셋이 아닌 데이터셋에서는 작업 수행에 도움을 주지 않을 가능성이 크기 때문이다.

[0004] 이러한 문제를 해결하는 종래기술들의 핵심은 학습 데이터셋 예제들 중에서도, 데이터셋 편향만으로 정답 클래스를 예측하기 어려운 예제에 더 집중하도록 분류기를 지도하는 것이다. 좀 더 구체적으로는 학습 데이터셋의 예제들 중, 데이터셋 편향만으로 정답 클래스를 올바르게 예측할 수 있는 예제인 편향 예제 (Biased Example)을 식별한 다음, 이 편향 예제의 학습 가중치를 낮게 적용한다. 이를 위해서는 먼저 편향 예제를 식별해야 하며, 이에 초기 연구에서는 "0: 부정 표현 없음, 1: 부정 표현 있음"과 같은 편향된 특징 (Biased Feature)을 제작한 다음, 이 편향된 특징만으로 클래스를 분류하도록 보조 모델을 학습하였다. 보조 모델은 편향된 특징만으로 정답 클래스를 예측해야 하므로, 이 보조 모델이 올바르게 정답 클래스를 예측한 예제는 편향 예제로 간주할 수 있으며, 따라서 이후 주 모델 학습은 해당 예제의 학습 가중치를 낮추는 방식으로 진행된다. 이러한 방법은 효과적이지만, 방법을 적용하기 위해서는 데이터셋 편향에 대해 미리 알고 있어야 하며, 이러한 사전 지식을 확보하기 위해서는 큰 비용이 든다는 문제가 있다.

[0005] 이에 최근 연구에서는 데이터셋 편향에 대한 사전 지식 없이, 편향 예제를 식별할 보조 모델을 학습하는 방법을 제시하였다. 이러한 방법들은 보조 모델을 주 모델이 수행한 작업과 동일한 분류 작업을 수행하도록 학습시키되 보조 모델의 수용량을 제한하거나, 적은 수의 데이터만으로 학습하게 하는 등의 방식으로 보조 모델이 데이터셋 편향에 주로 근거하여 정답 클래스를 예측하도록 지도하였다.

[0006] 최근 편향완화 연구들은 데이터셋 편향에 대한 사전 지식 없이 편향을 완화하기 위해 주 모델이 수행한 작업과 동일한 작업을 수행하는 보조 모델을 학습하였다. 특히 보조 모델의 학습은 다중 분류 작업을 수행하기 위해, 소프트맥스와 교차 엔트로피 손실을 이용한 다중 클래스 학습 목표에 기반한다. 그러나 이같이 보조 모델을 학습하는 방식은 보조 모델의 본래 목적인 편향 예제 식별과는 괴리가 있다. 이들 방법에서 보조 모델은 다중 분류 작업을 수행하기 위해 다중 클래스 학습 목표를 가정하지만, 보조 모델의 본래 목적은 주어진 예제가 편향되었는지 식별하는 이진 분류 작업이기 때문이다. 특히 다중 클래스 학습 목표에서는 소프트맥스 함수가 클래스 간 경쟁을 유발하며, 이로 인해 한 클래스에 대한 확신을 높일 시 다른 클래스에 대한 확신이 낮아진다. 따라서 보조 모델은 정답 클래스에 대한 로짓 값을 그대로 둔 채, 오답 클래스에 대한 확신을 낮춤으로써 손실을 최소화할 수 있다. 이에 보조 모델은 기대된 바와 달리, 편향을 충분히 학습하지 못하고 결과적으로 편향 예제를 편향 예제가 아닌 것으로 잘못 식별할 수 있다.

발명의 내용

해결하려는 과제

[0008] 상기와 같은 문제점을 해결하기 위한 본 발명의 목적은 보조모델의 학습 목표와 그것의 본래 목적 사이의 괴리를 완화함으로써 편향 예제를 종래 기술 대비 더 정확하게 식별하고, 이를 통해 주 모델의 편향완화 성능을 개선할 수 있는 편향완화를 위한 편향 전문가 기반의 편향 식별 장치, 방법 및 이를 기록한 컴퓨터로 판독가능한 기록매체를 제공하는데 있다.

과제의 해결 수단

[0010] 상기한 본 발명의 목적을 실현하기 위한 일 실시예에 따른 편향완화를 위한 편향 전문가 기반의 편향 식별 방법은 복수의 클래스의 학습 목표 기반으로 보조 모델을 학습하여 각 예제의 학습 난이도를 측정하는 단계; 상기

복수의 클래스가 존재하는 학습 데이터셋으로부터 상기 복수의 클래스 개수에 해당하는 이진 학습 목표를 구성하는 단계; 상기 학습 난이도 및 상기 이진 학습 목표에 기반하여 편향 전문가가 학습하는 단계; 및 주 모델 및 상기 편향 전문가의 소프트맥스 출력을 결합하는 앙상블을 통해 상기 주 모델을 학습하는 단계를 포함한다.

[0011] 또한, 상기 각 예제의 학습 난이도를 측정하는 단계는, 상기 보조 모델이 상기 각 예제의 학습 스텝마다 상기 각 예제의 정답 클래스에 부여한 확률을 측정하는 단계일 수도 있다.

[0012] 또한, 상기 이진 학습 목표를 구성하는 단계는, 기계학습의 One-vs-Rest 접근법에 따르는 단계일 수도 있다.

[0013] 또한, 상기 편향 전문가가 학습하는 단계는, 학습시에, 목표 클래스인 특정 클래스에 해당하는 편향 예제에는 높은 학습 가중치를 부여하고, 나머지 클래스에 해당하는 편향 예제에는 낮은 학습 가중치를 부여하는 단계일 수도 있다.

[0014] 본 발명의 또 다른 목적을 실현하기 위한 컴퓨터 프로그램이 기록된 컴퓨터로 판독 가능한 저장 매체는 상기 편향완화를 위한 편향 전문가 기반의 편향 식별 방법을 수행하기 위한 컴퓨터 프로그램이 기록되어 있다.

[0015] 본 발명의 또 다른 목적을 실현하기 위한 편향완화를 위한 편향 전문가 기반의 편향 식별 장치는, 프로세서; 상기 프로세서에 연결되는 메모리를 포함하되, 상기 메모리는, 복수의 클래스의 학습 목표 기반으로 보조 모델을 학습하여 각 예제의 학습 난이도를 측정하고, 상기 복수의 클래스가 존재하는 학습 데이터셋으로부터 상기 복수의 클래스 개수에 해당하는 이진 학습 목표를 구성하고, 상기 학습 난이도 및 상기 이진 학습 목표에 기반하여 편향 전문가가 학습하며, 주 모델 및 상기 편향 전문가의 소프트맥스 출력을 결합하는 앙상블을 통해 상기 주 모델을 학습하도록, 상기 프로세서에 의해 실행되는 프로그램 명령어들을 저장한다.

[0016] 또한, 상기 각 예제의 학습 난이도를 측정하는 것은, 상기 보조 모델이 상기 각 예제의 학습 스텝마다 상기 각 예제의 정답 클래스에 부여한 확률을 측정하는 것일 수도 있다.

[0017] 또한, 상기 이진 학습 목표를 구성하는 것은, 기계학습의 One-vs-Rest 접근법에 따르는 것일 수도 있다.

[0018] 또한, 상기 편향 전문가가 학습하는 것은, 학습시에, 목표 클래스인 특정 클래스에 해당하는 편향 예제에는 높은 학습 가중치를 부여하고, 나머지 클래스에 해당하는 편향 예제에는 낮은 학습 가중치를 부여하는 것일 수도 있다.

발명의 효과

[0020] 본 발명에 의해 보조모델의 학습 목표와 그것의 본래 목적 사이의 괴리를 완화함으로써 편향 예제를 종래 기술 대비 더 정확하게 식별하고, 이를 통해 주 모델의 편향완화 성능을 개선할 수 있다.

도면의 간단한 설명

[0022] 도 1은 본 발명의 실시예에 따른 편향완화를 위한 편향 전문가 기반의 편향 식별 장치의 간략도이다.

도 2는 본 발명의 실시예에 따른 편향완화를 위한 편향 전문가 기반의 편향 식별 방법의 순서도이다.

도 3은 도 2의 방법을 실행하는 편향완화를 위한 편향 전문가 기반의 편향 식별 장치의 상세도이다.

도 4는 본 별명의 일 실시예에 따른 편향완화를 위한 편향 전문가 기반의 편향 식별 장치의 성능을 나타내는 그래프이다.

도 5는 본 발명의 일 실시예에 따라 주어진 예제에 대한 식별력을 나타내는 그래프이다.

발명을 실시하기 위한 구체적인 내용

[0023] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세하게 설명하고자 한다. 그러나 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.

[0024] 제1, 제2 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된

다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. 및/또는 이라는 용어는 복수의 관련된 기재된 항목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다.

[0025] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다.

[0026] 본 출원에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, "포함하다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0027] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 가지는 의미와 일치하는 의미를 가진 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0028] 이하, 첨부한 도면들을 참조하여, 본 발명의 바람직한 실시예를 보다 상세하게 설명하고자 한다. 본 발명을 설명함에 있어 전체적인 이해를 용이하게 하기 위하여 도면상의 동일한 구성요소에 대해서는 동일한 참조부호를 사용하고 동일한 구성요소에 대해서 중복된 설명은 생략한다.

[0029] 본 발명의 핵심은 편향 전문가라 불리는, 일종의 중재자 역할을 맡는 분류기들을 보조 모델과 주 모델 사이에 도입하는 것이다. 이하 실시예를 첨부한 도면을 참조하여 상세히 설명한다.

[0030] 도 1은 본 발명의 실시예에 따른 편향완화를 위한 편향 전문가 기반의 편향 식별 장치의 간략도이다. 도 1에 도시된 바와 같이, 본 발명의 실시예에 따른 편향완화를 위한 편향 전문가 기반의 편향 식별 장치는 보조 모델(110), 편향 전문가(120) 및 주 모델(130)을 포함한다.

[0031] 보조 모델(110)은 주 모델보다 학습 파라미터가 적어 제한된 수용량을 가지며, 이로 인해 데이터셋 편향에 주로 근거하여 정답 클래스를 예측하기를 기대받는 모델이다. 이 보조 모델(110)은 편향 전문가 (120)를 학습하기에 앞서, 학습 데이터셋의 각 예제가 데이터셋 편향만으로 얼마나 쉽게 정답 클래스를 예측할 수 있는지를 측정하기 위해 먼저 학습된다.

[0032] 편향 전문가(120)들은 이진 학습 목표에 따라, 각자의 목표 클래스와 상관관계가 높은 데이터셋 편향을 집중적으로 학습하도록 설계된 분류기이다. 여기서 "이진 학습"이라 함은 각 클래스, 예를 들어, 예제들에 대해 함의, 중의, 부정의 3개의 클래스 각각에 편향 전문가가 예, 아니오 중 하나로 결론내리는 것을 의미한다.

[0033] 목표 클래스와 상관관계가 높은 데이터셋 편향을 집중적으로 학습하기 위해, 각 편향 전문가는 보조 모델이 계산하는 예제별 난이도에 근거하여, 학습 시 예제마다 다른 가중치를 부여받는다.

[0034] 주 모델(130)은 편향 전문가(120)와의 앙상블을 통하여 학습되는 모델로, 본 발명에서는 주 모델(130)이 데이터셋 편향에 주로 근거하여 정답 클래스를 예측하는 현상을 완화하는 것을 최종 목표로 삼는다.

[0035] 도 2는 본 발명의 실시예에 따른 편향완화를 위한 편향 전문가 기반의 편향 식별 방법의 순서도이며, 도 3은 도 2의 방법을 실행하는 편향완화를 위한 편향 전문가 기반의 편향 식별 장치의 상세도이다.

[0036] 도 2 및 도 3을 참조하면, 우선 보조 모델(110)이 다중 클래스 학습 목표에 기반하여 학습된다 (S210). 학습 과정에서는 일정 학습 스텝마다, 보조 모델(110)이 각 예제의 정답 클래스에 부여한 확률을 측정한다. 그리고 이를 학습 데이터셋의 예제별 난이도를 나타내는 지표로 활용한다.

[0037] 그 후, 기계학습의 One-vs-Rest라는 접근법에 따라, K 개의 클래스가 존재하는 학습 데이터셋으로부터 K 개의 이진 학습 목표를 구성한다 (S220). 예를 들어 i 번째 이진 학습 목표 (i=0,1,2,...,K-1) 에서는 목표 클래스인 클래스 i에 해당하는 예제를 긍정 예제로, 나머지 클래스에 해당하는 예제를 부정 예제로 간주한다.

[0038] 그 후, S210 에서 보조 모델(110)을 통해 측정한 예제별 난이도, 그리고 단계(220)에서 구성한 이진 학습 목표

를 기반으로 총 K개의 편향 전문가(120)를 학습한다 (S230). 예를 들어, i 번째 편향 전문가(120)는 i 번째 이진 학습 목표에 기반하여 학습된다. 학습 시에는 목표 클래스인 i 클래스에 해당하는 편향 예제에는 높은 학습 가중치를 부여하고, 반대로 나머지 클래스에 해당하는 편향 예제에는 낮은 학습 가중치를 부여한다. 이는 각 편향 전문가가 목표 클래스가 아닌 클래스와 상관관계가 높은 데이터셋 편향을 학습하는 것을 막기 위함이다.

[0039] 이와 같은 이진 학습 목표와 차등적 가중치 부여를 통해, i 번째 편향 전문가(120)는 목표 클래스인 i 클래스와 상관관계가 높은 데이터셋 편향을 집중적으로 학습하도록 지도받는다. 궁극적으로 각 편향 전문가(120)는 주어진 예제가 편향되었는지 식별하는 이진 분류 작업을 수행함으로써 기존의 보조 모델의 학습 목표와 그것의 본래 목적 사이의 괴리에서 발생하는 편향 예제 오식별 문제를 완화한다.

[0040] 그 다음, 주 모델(130)과 편향 전문가(120)들의 소프트맥스 출력을 결합하는 앙상블을 통하여 주 모델(130)을 학습한다 (S240). 소프트맥스 출력은 주어진 예제가 어느 클래스로 분류될지에 대하여 모델이 예측한 결과를 확률 분포로 나타낸 것이다. 따라서 각 예제의 소프트맥스 출력은 K 개의 확률로 구성되어 있고 그 합은 1이다. 학습 데이터셋의 한 예제가 주어졌다고 가정할 때, 만약 소프트맥스 출력의 K 개 확률 중에서 정답 클래스에 해당하는 확률이 가장 높다면 모델은 해당 예제의 정답 클래스를 올바르게 예측하였다고 간주한다. 특히, 주어진 예제에 대해 편향 전문가(120)들이 정답 클래스를 올바르게 예측하였다면, 해당 예제는 편향 예제로 간주하고 주 모델(130) 학습 시에 적은 가중치를 부여받는다. 반대로 편향 전문가(120)들이 정답 클래스를 올바르게 예측하지 못하였다면, 해당 예제는 편향 예제가 아닌 것으로 간주하고, 주 모델(130) 학습 시에 높은 가중치를 부여받는다. 궁극적으로 주 모델(130)은 학습 데이터셋 예제들 중에서도, 데이터셋 편향만으로 정답 클래스를 예측하기 어려운 예제에 더 집중하도록 지도받고, 이를 통해 모델이 데이터셋 편향에 주로 근거하여 정답 클래스를 예측하는 문제를 완화한다.

[0041] 이하에서는 본 발명의 실시예들에 따른 편향완화 방법의 수학적 의미에 대해 설명한다.

[0042] 입력 $x \in X$ 와 정답 클래스 $y \in Y = \{0, 1, \ldots, K-1\}$ 로 구성된 예제를 고려한다. 클래스 i 를 목표 클래스로 가지는 편향 전문가를 f_b^i 라고 할 때, 편향 전문가(120)를 학습하기 위한 손실 함수는 아래 수학식 1과 같이 표현할 수 있다.

수학식 1

[0043]
$$L_b = \lambda\, q_{x,y}^{\alpha} \log \sigma(f_b^y(x)) + (1-\lambda)(1-q_{x,y})^{\alpha} \sum_{i \neq y} \log\,(1-\sigma(f_b^i(x)))$$

[0044] 여기서 $\lambda \in [0,1]$ 는 이진 학습 목표로 인하여 발생하는 클래스 불균형 문제를 완화하기 위한 가중치 요인이며, $q_{x,y} \in [0,1]$ 는 입력 x 에 대하여 보조 모델(110)이 일정 학습 스텝마다 정답 클래스 y에 부여한 확률들의 평균으로 해당 값이 클수록 쉬운 예제 (즉, 편향 예제)임을 나타낸다. 여기서 목표 클래스가 y인 f_b^y 가 포함된 손실 식에서는 $q_{x,y}^{\alpha}$ 를, 목표 클래스가 y가 아닌 나머지 편향 전문가가 포함된 손실 식에서는 $(1-q_{x,y})^{\alpha}$ 를 가중치로 활용하였다. 이는 전술한 바와 같이, 학습 시에 목표 클래스인 클래스에 해당하는 편향 예제에는 높은 학습 가중치를 부여하고, 반대로 나머지 클래스에 해당하는 편향 예제에는 낮은 학습 가중치를 부여하기 위함이다. $\alpha \in [0,1]$ 는 보조 모델(110)을 통해 가중치를 부여하는 정도를 조절하기 위한 하이퍼파라미터로, α 가 작을수록 보조 모델(110)이 손실 계산에 주는 영향이 줄어든다. 마지막으로 σ 는 시그모이드 활성화 함수를 나타낸다.

[0045] 또한, 전술한 바와 같이 주 모델(130) 학습 시에는, 주 모델(130)과 편향 전문가(120)들의 소프트맥스 출력을 결합하는 앙상블을 적용하며 그 손실 함수는 아래의 수학식 2와 같이 표현된다.

수학식 2

[0046] $$L_d = -e_y \cdot \log(\text{softmax}(\log p_d + \log p_b))$$

[0047] 여기서 e_y 는 (y+1)번째 값만 1이고 나머지 값은 0인 K차원 벡터이다. $\cdot$ 는 벡터 간 내적을 의미하고, softmax 는 소프트맥스 함수이다. *Pd*는 주 모델(130)이 계산한 소프트맥스 출력이며, *Pb*는 편향 전문가(120)들의 소프트맥스 출력으로 $p_b = \text{softmax}([f_b^1(x), f_b^2(x), ..., f_b^K(x)])$ 와 같이 계산한다.

[0048] 아래의 표 3은 자연어이해 작업 중 자연어 추론 (Natural Language Inference) 작업을 수행하는 MNLI, 사실 검증 (Fact Verification) 작업을 수행하는 FEVER, 그리고 의역 식별 (Paraphrase Identification) 작업을 수행하는 QQP 데이터셋에 대하여 본 발명이 제안하는 방법론의 정확도를 측정한 결과이다.

표 1

[0049]

Method	Known biases	MNLI			FEVER			QQP		
		dev	HANS	Gap	dev	symm.	Gap	dev	PAWS	Gap
BERT-base (Devlin et al., 2019)	✗	84.5	62.4	22.1	85.6	63.1	22.5	91.0	33.5	57.5
Reweighting (Clark et al., 2019)	✓	83.5	69.2	14.3	-	-	-	-	-	-
Reweighting (Schuster et al., 2019)	✓	-	-	-	84.6	66.5	18.1	-	-	-
PoE (Clark et al., 2019)	✓	82.9	67.9	15.0	-	-	-	-	-	-
Conf-reg (Utama et al., 2020a)	✓	84.3	69.1	15.7	86.4	66.2	20.2	-	-	-
Reweighting (Utama et al., 2020b)	✗	82.3	69.7	12.6	87.1	65.5	21.6	85.2	57.4	**27.8**
PoE (Utama et al., 2020b)	✗	81.9	66.8	15.1	85.9	65.8	20.1	86.1	56.3	29.8
PoE (Sanh et al., 2021)	✗	83.3	67.9	15.4	84.8	65.7	19.1	88.0	46.4	41.6
Conf-reg (Utama et al., 2020b)	✗	**84.3**	67.1	17.2	**87.6**	66.0	21.6	89.0	43.0	46.0
Self-Debiasing (Ghaddar et al., 2021)	✗	83.2	71.2	12.0	-	-	-	**90.2**	46.5	43.7
Bias Experts (ours)	✗	82.7	**72.6**	**10.1**	85.6	**68.1**	**17.5**	86.8	**58.1**	28.7

[0050] 세 가지 데이터셋 각각에 대하여 편향완화 성능을 확인하기 위해, 학습 데이터셋과 다른 분포를 가지는 고난도 평가 데이터셋(Challenge test set)에 대한 정확도를 함께 측정하였다. 또한 “Known Biases”는 편향을 완화하면서 해당 데이터셋의 데이터셋 편향에 대한 사전 지식을 활용하였는지에 대한 여부를 나타낸다. 즉, Reweighting (Clark et al., 2019; Schuster et al., 2019), PoE (Clark et al., 2019), Conf-reg (Utama et al., 2020a) 는 해당 데이터셋의 데이터셋 편향에 대한 사전 지식을 활용한 방법이며, 그 외에는 해당 데이터셋의 데이터셋 편향에 대한 사전 지식을 활용하지 않은 방법이다.

[0051] MNLI-HANS, FEVER-symm., 및 QQP-PAWS 열을 보았을 때, 본 발명에서 제안하는 방법이 FEVER-symm 이 68.1로서 가장 높으며, MNLI-HANS 및 QQP-PAWS 에서도 각각 72.6 및 58.1로서 가장 우수한 수치를 나타낸다

[0052] 따라서, 편향 전문가(120)를 이용하는 방법이 데이터셋 편향에 대한 사전 지식을 활용하지 않는 종래 기술 대비 더 우수한 성능을 보이는 것을 확인할 수 있다. 이를 통해, 이진 분류 기반의 편향 전문가(120)가 여러 데이터셋에 걸쳐 기존 방법 대비 편향 예제를 더 효과적으로 식별하는 것을 알 수 있다.

[0053] 도 4는 본 별명의 일 실시예에 따른 편향완화를 위한 편향 전문가 기반의 편향 식별 장치의 성능을 나타내는 그래프이다. 도 4를 참조하면, 본 도면은 수학식 1의 에 따른 주 모델(130)의 성능 변화를 나타낸 결과이다. 성능 테스트는 MNLI 및 HANS 데이터셋을 이용하여 진행하였다. 먼저 $\alpha = 0$ 일 경우, 즉, 보조 모델(110)이 손실 계산에 영향을 끼치지 않을 시 성능이 가장 낮으며 이는 보조 모델(110)을 통해 가중치를 부여하는 정도를 조절하여 주 모델(130)을 개선할 수 있음을 의미한다. 반면 α 가 0.2를 초과할 경우, 오히려 성능이 감소하며 이는 너무 큰 α 가 주 모델(130)이 충분한 경사 업데이트를 받는 것을 방해함을 암시한다.

[0054] 아래의 표 2는 본 발명에서 제안하는 방법들의 효과성을 나타낸다.

표 2

Ablative Setting	MNLI	HANS		
		Ent	NEnt	Avg
Bias Experts (ours)	82.7	93.5	**51.8**	**72.6**
(1) w/o Amp.	82.8	**94.9**	45.0	70.0
(2) w/o OvR	83.0	94.4	42.7	68.5
(3) w/o Amp. and OvR	**83.1**	94.5	41.4	67.9

[0055]

[0056] 성능 테스트는 MNLI 및 HANS 데이터셋을 이용하여 진행하였다. HANS는 Ent (Entailment, 함의) 및 NEnt (Non-Entailment, 비함의)의 두 가지 클래스로 나뉘며, HANS의 NEnt 클래스에서의 분류기 성능을 높이면 HANS의 Ent 클래스 및 MNLI에서의 성능이 감소하는 경향을 보인다. 먼저 표의 "(1) w/o AMP. (여기서 w/o 는 without 의 의미로서 AMP를 사용하지 않음을 의미)"를 통해, 보조 모델(110)로 가중치를 부여하는 정도를 조절한 방법 즉 편향 전문가를 사용한 실험 (Bias Experts)이 가중치를 부여하는 정도를 조절하지 않은 방법인 w/o AMP 보다 HANS의 NEnt 클래스 기준 더 높은 성능을 달성함을 알 수 있다.

[0057] 또한 "(2) w/o OvR"을 통해 One-vs-Rest라는 접근법에 따라 이진 학습 목표를 구성한 방법이 그렇지 않은 방법보다 HANS의 NEnt 클래스 기준 더 높은 성능을 달성함을 알 수 있다. "(3) w/o AMP. and OvR"을 보았을 때는, 두 방법을 모두 적용하지 않을 시에 HANS의 NEnt 클래스 기준 가장 낮은 성능을 달성하는 것을 알 수 있다. 마지막으로 "Bias Experts (ours)"로부터, 두 가지 방법을 모두 적용할 시에 HANS의 Ent 클래스 및 MNLI에서의 성능 하락을 최소화하면서도 HANS의 NEnt 클래스에서 가장 높은 성능을 달성하는 것을 확인할 수 있다. 이는 두 방법이 상호보완적임을 의미한다.

[0058] 도 5는 본 발명의 일 실시예에 따라 주어진 예제에 대한 식별력을 나타내는 그래프이다. 여기서 편향 예제를 더 정확하게 식별하는 것은, 각 예제의 정답을 예측할 때 편향 예제에는 종래 기술 대비 더 높은 확신을, 편향 예제가 아닌 예제에는 종래 기술 대비 더 낮은 확신을 가지는 것을 의미한다. 본 발명은 기존과 같은 다중분류 학습목표가 아닌 이진분류 학습목표를 토대로 편향 전문가를 학습하여, 편향 예제를 더 정확하게 식별한다. 여기서 성능 테스트는 MNLI 데이터셋을 이용하여 수행하였다.

[0059] 도 5a는 편향 예제에 대한 확신성을 나타내며, 도 5b는 편향 예제가 아닌 예제에 대한 확신성을 나타낸다. 도 5a를 참조하면 편향 예제인 경우에는 편향 전문가가 기존 모델 (약한 학습자 (Weak Learner)) 보다 평균적으로 더 높은 확신을 가지는 것을 알 수 있다. 또한, 도 5b를 참조하면 편향 예제가 아닌 경우에는 편향 전문가가 기준 모델 (약한 학습자 (Weak Learner)) 보다 평균적으로 더 낮은 확신을 가지는 것을 알 수 있다. 이는 편향 전문가가 기준 모델에 비해 편향 예제를 더 정확하게 식별함을 나타낸다.

[0060] 표 3은 각 예제가 편향되었는지에 대하여 편향 전문가(120)가 식별한 결과, 그리고 그에 따른 주 모델(130)의 성능을 나타낸 결과이다. 표 3의 경우 HANS 데이터셋을 이용하여 수행하였다.

표 3

Method	Ent			NEnt		
	Lexical	Subseq	Const	Lexical	Subseq	Const
BERT-base	99.5	99.2	99.4	52.4	10.0	17.8
PoE (Sanh et al., 2021)	**90.2**	**97.0**	**96.3**	66.0	18.0	40.1
Bias Experts (ours)	88.4	96.6	95.5	**83.7**	**26.4**	**45.2**

[0061]

[0062] 표 3의 HANS는 Ent (Entailment, 함의) 및 NEnt (Non-Entailment, 비함의)의 두 가지 클래스로 나뉘며, 한 클

래스에서의 분류기 성능을 높이면 다른 클래스에서의 성능이 감소하는 경향을 보인다.

[0063] 결과적으로 표 3로부터, 편향 전문가를 이용한 방법론이 기준 방법론에 비해 HANS 데이터셋의 Ent 클래스에서의 성능 하락을 최소화하면서도, NEnt 클래스에서 더 높은 성능을 달성하는 것을 알 수 있다.

[0064] 또한, 전술한 금속원소로 편향완화를 위한 편향 전문가 기반의 편향 식별 방법은 애플리케이션으로 구현되거나 다양한 컴퓨터 구성요소를 통하여 수행될 수 있는 프로그램 명령어의 형태로 구현되어 컴퓨터 판독 가능한 기록 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능한 기록 매체는 프로그램 명령어, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다.

[0065] 상기 컴퓨터 판독 가능한 기록 매체에 기록되는 프로그램 명령어는 본 발명을 위하여 특별히 설계되고 구성된 것들이거니와 컴퓨터 소프트웨어 분야의 당업자에게 공지되어 사용 가능한 것일 수도 있다.

[0066] 컴퓨터 판독 가능한 기록 매체의 예에는, 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체, CD_ROM, DVD와 같은 광기록 매체, 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-opticalmedia), 및 ROM, RAM, 플래시 메모리 등과 같은 프로그램 명령어를 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다.

[0067] 프로그램 명령어의 예에는, 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드도 포함된다. 상기 하드웨어 장치는 본 발명에 따른 처리를 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0068] 이상에서는 실시예들을 참조하여 설명하였지만, 해당 기술 분야의 숙련된 당업자는 하기의 특허 청구의 범위에 기재된 본 발명의 사상 및 영역으로부터 벗어나지 않는 범위 내에서 본 발명을 다양하게 수정 및 변경시킬 수 있음을 이해할 수 있을 것이다.

부호의 설명

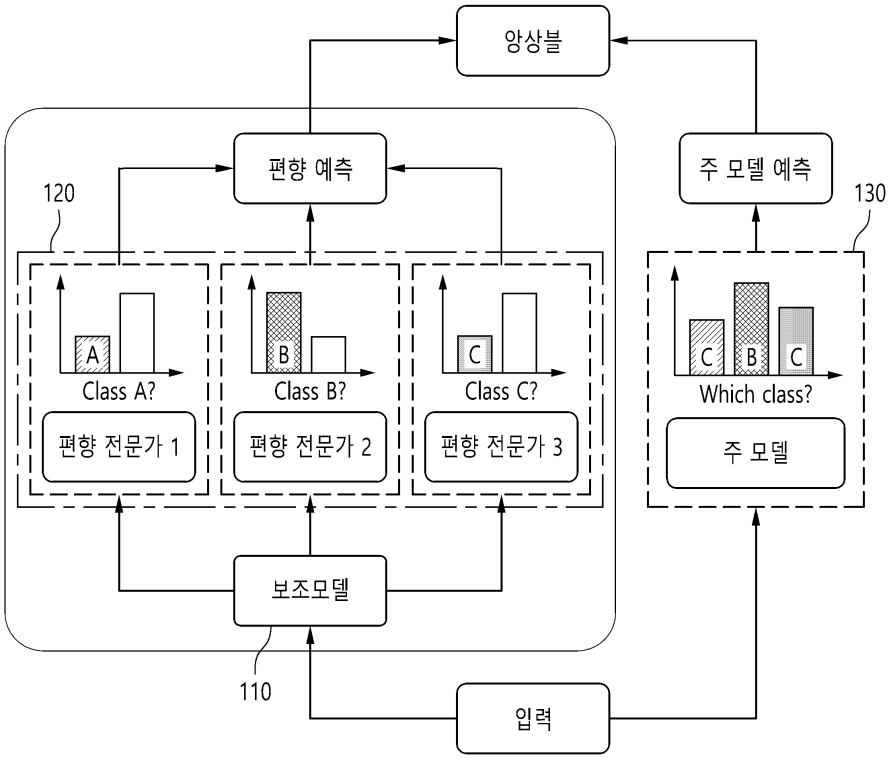
[0069] 100 : 편향완화를 위한 편향 전문가 기반의 편향 식별 장치

110 : 보조 모델 120 : 편향 전문가

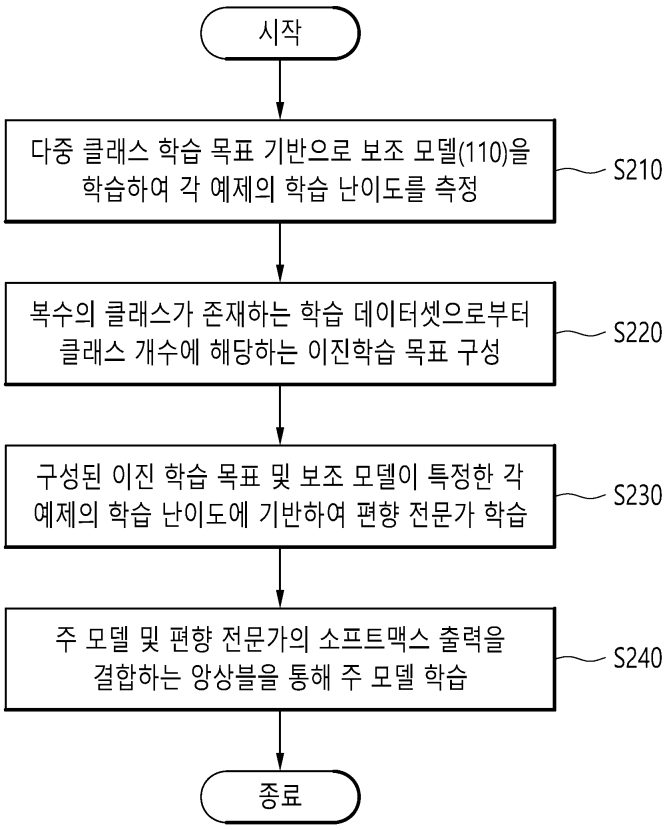
130 : 주 모델

도면

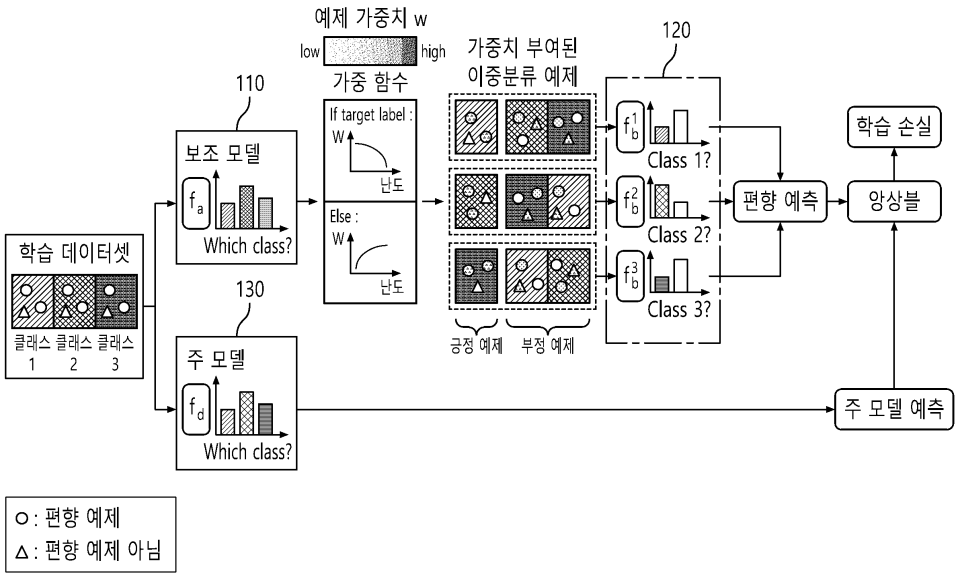
도면1


앙상블
편향 예측
주 모델 예측
120
130
A
B
C
Class A?
Class B?
Class C?
C
B
C
Which class?
편향 전문가 1
편향 전문가 2
편향 전문가 3
주 모델
보조모델
110
입력

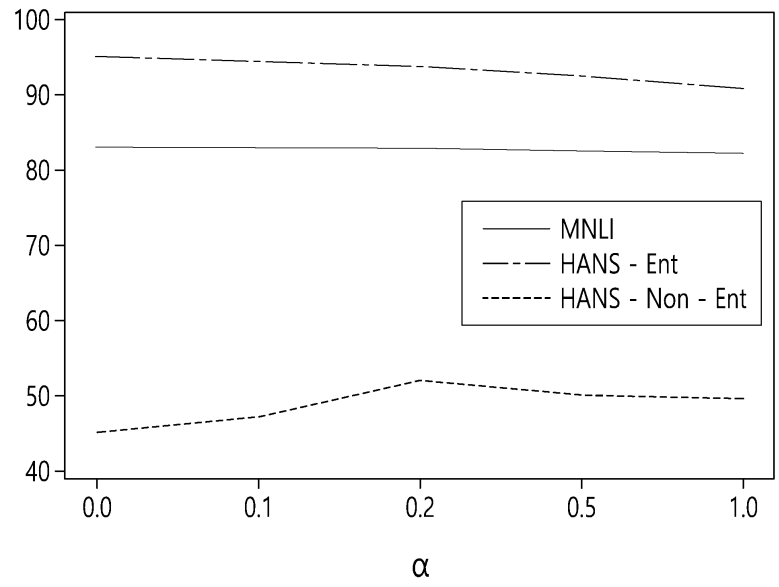
도면2


시작
다중 클래스 학습 목표 기반으로 보조 모델(110)을 학습하여 각 예제의 학습 난이도를 측정
S210
복수의 클래스가 존재하는 학습 데이터셋으로부터 클래스 개수에 해당하는 이진학습 목표 구성
S220
구성된 이진 학습 목표 및 보조 모델이 특정한 각 예제의 학습 난이도에 기반하여 편향 전문가 학습
S230
주 모델 및 편향 전문가의 소프트맥스 출력을 결합하는 앙상블을 통해 주 모델 학습
S240
종료

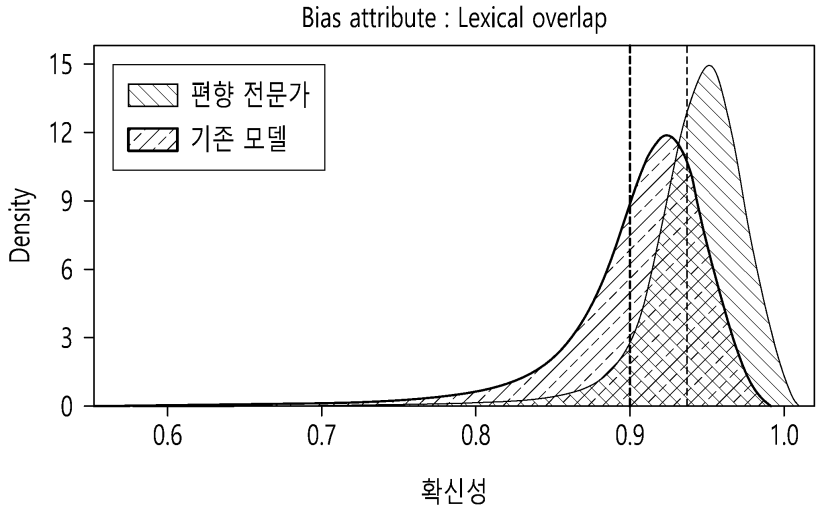
도면3


예제 가중치 w
low
high
가중 함수
가중치 부여된
이중분류 예제
120
110
130
보조 모델
Which class?
If target label :
W
난도
Else :
W
난도
학습 데이터셋
클래스 1
클래스 2
클래스 3
주 모델
Which class?
Class 1?
Class 2?
Class 3?
긍정 예제
부정 예제
편향 예제
앙상블
학습 손실
주 모델 예측
○ : 편향 예제
△ : 편향 아님

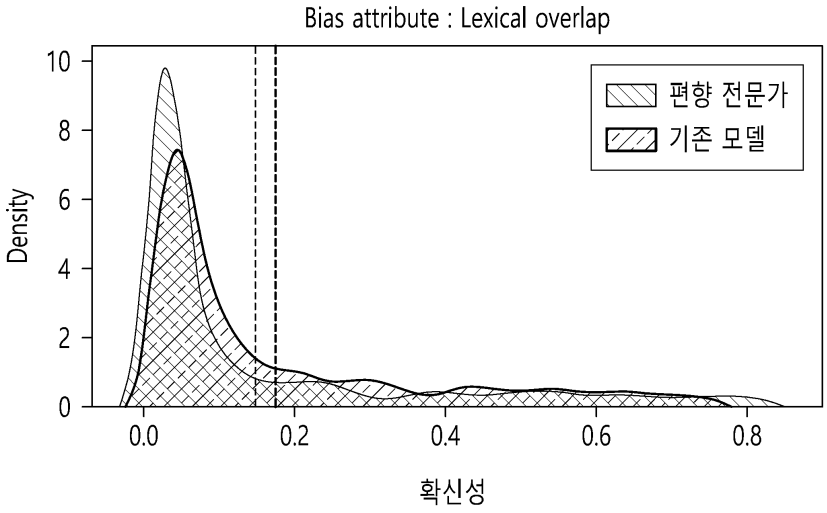
도면4


100
90
80
70
60
50
40
0.0
0.1
0.2
0.5
1.0
α
MNLI
HANS - Ent
HANS - Non - Ent

도면5a


Bias attribute : Lexical overlap
편향 전문가
기존 모델
Density
15
12
9
6
3
0
0.6
0.7
0.8
0.9
1.0
확신성

도면5b


Bias attribute : Lexical overlap
편향 전문가
기존 모델
Density
10
8
6
4
2
0
0.0
0.2
0.4
0.6
0.8
확신성