(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0103815
(43) 공개일자 2024년07월04일

(51) 국제특허분류(Int. Cl.)
G06N 3/045 (2023.01) *G06F 40/20* (2020.01)
G06N 3/08 (2023.01)
(52) CPC특허분류
G06N 3/045 (2023.01)
G06F 40/20 (2022.01)
(21) 출원번호 10-2022-0186332
(22) 출원일자 2022년12월27일
심사청구일자 2022년12월27일

(71) 출원인
충남대학교산학협력단
대전광역시 유성구 대학로 99 (궁동, 충남대학교)
(72) 발명자
정상근
대전광역시 서구 둔산북로 160, 한마루아파트 102동 1502호
김성렬
대전광역시 유성구 대학로163번길 9, 302호
황태욱
대전광역시 유성구 유성대로 757-5, 204호
(74) 대리인
이은철, 김재문

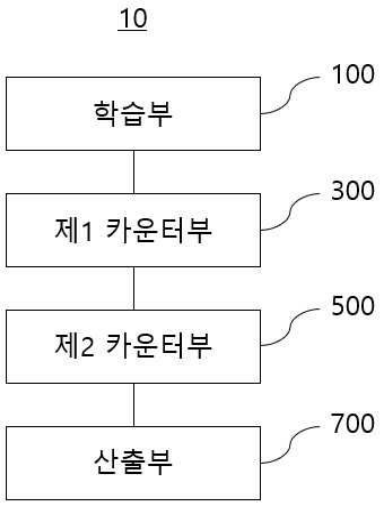
전체 청구항 수 : 총 4 항

(54) 발명의 명칭 **인공지능 학습데이터의 품질 평가 방법 및 이를 이용한 시스템**

(57) 요 약

본 발명은, 복수개의 딥러닝 모델에 동일한 데이터셋을 학습시키는 (a)단계; 상기 (a)단계에서 학습에 사용된 데이터가 틀린 횟수를 카운트하는 (b)단계; i번 틀린 모든 데이터의 개수를 카운트하는 (c)단계; 및 0부터 상기 (b)단계에서 카운트한 틀린 횟수번까지의 오답 횟수를 합산한 값을 전체 데이터 개수에서 틀린 횟수가 0개인 데이터의 개수를 뺀 값으로 나누어 점수를 산출하는 (d)단계;를 포함하는 것을 일 특징으로 한다.

대 표 도 - 도3



(52) CPC특허분류
G06N 3/08 (2023.01)

명 세 서

청구범위

청구항 1

(a) 복수개의 딥러닝 모델에 동일한 데이터셋을 학습시키는 단계;

(b) 상기 (a)단계에서 학습에 사용된 데이터가 틀린 횟수를 카운트하는 단계;

(c) i번 틀린 모든 데이터의 개수를 카운트하는 단계; 및

(d) 0부터 상기 (b)단계에서 카운트한 틀린 횟수 번까지의 오답 횟수를 합산한 값을 전체 데이터 개수에서 틀린 횟수가 0개인 데이터의 개수를 뺀 값으로 나누어 점수를 산출하는 단계;를

포함하는 인공지능 학습데이터의 품질 평가 방법.

청구항 2

제 1 항에 있어서,

상기 (a)단계는,

복수개의 딥러닝 모델을 K에폭씩 N번 학습시키는 것을 특징으로 하는 인공지능 학습데이터의 품질 평가 방법.

청구항 3

제 2 항에 있어서,

상기 (c)단계는,

i의 범위가 0부터 딥러닝 모델의 개수×K×N인 것을 특징으로 하는 인공지능 학습데이터의 품질 평가 방법.

청구항 4

복수개의 딥러닝 모델에 동일한 데이터셋을 학습시키는 학습부;

상기 학습부가 학습에 사용한 데이터가 틀린 횟수를 카운트하는 제1 카운터부;

i번 틀린 모든 데이터의 개수를 카운트하는 제2 카운터부; 및

0부터 상기 제1 카운터부가 카운트한 틀린 횟수 번까지의 오답 횟수를 합산한 값을 전체 데이터 개수에서 틀린 횟수가 0개인 데이터의 개수를 뺀 값으로 나누어 점수를 산출하는 산출부;를

포함하는 인공지능 학습데이터의 품질 평가 시스템.

발명의 설명

기 술 분 야

[0001] 본 발명은 인공지능 학습데이터의 품질 평가 방법 및 이를 이용한 시스템에 관한 것으로서, 특히 학습이 어려운 데이터를 선별할 수 있도록 데이터를 점수로 수치화하여 보여주는 인공지능 학습데이터의 품질 평가 방법 및 이를 이용한 시스템에 관한 것이다.

배 경 기 술

[0003] 최근 다양한 분야에 딥러닝의 활용이 증가하고 있다. 그러나 활용 가능한 데이터셋이 부족한 경우, 딥러닝 모델의 성공적인 학습을 판단할 객관적인 기준을 세우기 어렵다. 이에 최근 자연어처리 분야에서는 영어 기반의 General Language Understanding Evaluation(GLUE)와 SuperGLUE 같이 학습된 자연어 모델의 객관적이며 범용적인 비교 및 평가가 가능한 벤치마크 데이터셋이 등장했다. 이를 통해 모델 간의 객관적인 성능 비교가 가능해지면서, 다양한 관점의 자연어처리 평가를 위한 벤치마크 데이터셋의 중요성이 입증되었다. 하지만 한국어로 학습을 시킨 모델을 영어 기반인 GLUE와 SuperGlue로 평가할 수는 없다.

[0004] 따라서 최근에는 한국어벤치마크 데이터셋인 Korean Language Understanding Evaluation(KLUE)가 등장했다. 이를 통해 한국어 모델에 대해서도 비교 및 평가가 가능해졌다.

[0005] 영상처리 분야에서는 데이터의 해상도나 선명도 등 이미지나 영상의 품질을 평가하는 연구가 일부 진행되고 있다. 그러나 데이터가 학습 및 모델에 적합한지, 혹은 학습을 방해하는 난해한 데이터가 있는지 등의 데이터를 평가하는 연구는 미비하다. 자연어처리 분야에도 범용성 있게 데이터 품질을 측정할 수 있는 대표 기술이 아직 없다.

발명의 내용

해결하려는 과제

[0007] 본 발명은 여러 모델에 동일한 데이터를 학습하고, 각 모델이 공통적으로 틀린 데이터를 모아서 학습이 어려운 데이터를 선별할 수 있도록 데이터를 점수화하는 인공지능 학습데이터의 품질 평가 방법 및 이를 이용한 시스템을 제공하는 것을 목적으로 한다.

과제의 해결 수단

[0009] 상기 목적을 달성하기 위하여 본 발명은, 복수개의 딥러닝 모델에 동일한 데이터셋을 학습시키는 (a)단계; 상기 (a) 단계에서 학습에 사용된 데이터가 틀린 횟수를 카운트하는 (b)단계; i번 틀린 모든 데이터의 개수를 카운트하는 (c)단계; 및 0부터 상기 (b) 단계에서 카운트한 틀린 횟수 번까지의 오답 횟수를 합산한 값을 전체 데이터 개수에서 틀린 횟수가 0개인 데이터의 개수를 뺀 값으로 나누어 점수를 산출하는 (d)단계;를 포함하는 것을 일 특징으로 한다.

[0010] 바람직하게는, 상기 (a)단계는, 복수개의 딥러닝 모델을 K에폭씩 N번 학습시킬 수 있다.

[0011] 바람직하게는, 상기 (c)단계는, i의 범위가 0부터 딥러닝 모델의 개수×K×N일 수 있다.

[0012] 또한 본 발명은, 복수개의 딥러닝 모델에 동일한 데이터셋을 학습시키는 학습부; 상기 학습부가 학습에 사용한 데이터가 틀린 횟수를 카운트하는 제1 카운터부; i번 틀린 모든 데이터의 개수를 카운트하는 제2 카운터부; 및 0부터 상기 제1 카운터부가 카운트한 틀린 횟수번까지의 오답 횟수를 합산한 값을 전체 데이터 개수에서 틀린 횟수가 0개인 데이터의 개수를 뺀 값으로 나누어 점수를 산출하는 산출부;를 포함하는 것을 다른 특징으로 한다.

발명의 효과

[0014] 본 발명은 데이터셋의 품질 점수를 측정하여, 문제가 있는 데이터를 선별할 수 있는 수치화된 점수를 제공할 수 있다는 이점이 있다.

[0015] 본 발명은 데이터셋에서 문제되는 데이터를 선별케 하여 데이터셋의 학습 성능을 향상시킬 수 있다는 이점이 있다.

도면의 간단한 설명

[0017] 도 1은 본 발명의 실시예에 따른 (b)단계 수행 결과를 나타낸다.

도 2는 본 발명의 실시예에 따른 (c)단계 수행 결과를 나타낸다.

도 3은 본 발명의 실시예에 따른 인공지능 학습데이터의 품질 평가 시스템의 구성도를 나타낸다.

도 4는 본 발명의 실시예에 따른 ALL 데이터 학습 중 틀린 횟수를 나타내는 그래프이다.

도 5는 재태깅 데이터를 개별적으로 학습한 결과와 보팅 데이터를 학습시킨 결과를 나타낸다.

도 6은 본 발명의 실시예에 따른 ALL 데이터와 보팅 데이터의 산출 점수를 나타내는 그래프이다.

발명을 실시하기 위한 구체적인 내용

[0018] 이하, 첨부된 도면들에 기재된 내용들을 참조하여 본 발명을 상세히 설명한다. 다만, 본 발명이 예시적 실시 예들에 의해 제한되거나 한정되는 것은 아니다. 각 도면에 제시된 동일 참조부호는 실질적으로 동일한 기능을 수행하는 부재를 나타낸다.

[0019] 본 발명의 목적 및 효과는 하기의 설명에 의해서 자연스럽게 이해되거나 보다 분명해 질 수 있으며, 하기의 기재만으로 본 발명의 목적 및 효과가 제한되는 것은 아니다. 또한, 본 발명을 설명함에 있어서 본 발명과 관련된 공지 기술에 대한 구체적인 설명이, 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략하기로 한다.

[0020] 본 발명의 실시예에 따른 인공지능 학습데이터의 품질 평가 방법은 복수개의 딥러닝 모델에 동일한 데이터셋을 학습시키는 (a)단계, (a) 단계에서 학습에 사용된 데이터가 틀린 횟수를 카운트하는 (b)단계, i번 틀린 모든 데이터의 개수를 카운트하는 (c)단계, 및 0부터 상기 (b) 단계에서 카운트한 틀린 횟수번까지의 오답 횟수를 합산한 값을 전체 데이터 개수에서 틀린 횟수가 0개인 데이터의 개수를 뺀 값으로 나누어 점수를 산출하는 (d)단계를 포함할 수 있다.

[0021] 인공지능 학습데이터의 품질 평가 방법은 데이터의 품질을 정량적으로 측정할 수 있다. 인공지능 학습데이터의 품질 평가 방법은 다양한 딥러닝 모델이 동일한 데이터셋을 여러 번 학습하여 지속적으로 오답이 나오는 데이터에 높은 점수를 부여할 수 있다. 인공지능 학습데이터의 품질 평가 방법은 데이터 품질이 우수할수록 0에 가까운 점수를 산출하고, 데이터 품질이 미흡할수록 1에 가까운 점수를 산출할 수 있다. 인공지능 학습데이터의 품질 평가 방법은 데이터가 해당 분야에 있어 적합한지, 데이터에 오류가 있는지 등을 판단할 수 있는 수치화된 점수를 산출할 수 있다. 이하에서 설명하는 틀린 데이터(오답 데이터)는 학습이 어려운 데이터를 의미할 수 있다.

[0022] (a)단계는 복수개의 딥러닝 모델에 동일한 데이터셋을 학습시킬 수 있다. (a)단계는 하나의 모델로 치우쳐진 결과보다 객관적인 결과를 위해 다수의 모델을 사용할 수 있다. (a)단계는 복수개의 딥러닝 모델을 K에폭씩 N번 학습시킬 수 있다.

[0023] (b)단계는 (a)단계에서 학습에 사용된 데이터가 틀린 횟수를 카운트할 수 있다. (b)단계에서 수행되는 동작은 아래의 [수학식 1]로 표현될 수 있다. (b)단계는 각 에폭마다 오답 데이터들을 기록하고, 각 데이터가 오답을 발생시킨 횟수(틀린 횟수)를 카운트할 수 있다.

[0024] [수학식 1]

[0025] $Count(data) = \text{data가 틀린 횟수}$

[0026] 도 1은 본 발명의 실시예에 따른 (b)단계 수행 결과를 나타낸다. 도 1을 참조하면, index 0에 있는 데이터는 34번, index 1에 있는 데이터는 7번, index 2에 있는 데이터는 94번, index 3에 있는 데이터는 32번, index 4에 있는 데이터는 41번 오답을 발생시켰다.

[0027] (c)단계는 i번 틀린 모든 데이터의 개수를 카운트할 수 있다. (c)단계는 i의 범위가 0부터 딥러닝 모델의 개수×K×N일 수 있다. (c)단계에서 수행되는 동작은 아래의 [수학식 2]로 표현될 수 있다.

[0028] [수학식 2]

[0029] $Wrong_i = i\text{번 틀린 모든 데이터의 개수}$

[0030] 도 2는 본 발명의 실시예에 따른 (c)단계 수행 결과를 나타낸다. 도 2를 참조하면, 34번 오답을 발생시킨 데이터는 index0, index391, index456, index589, index72로 총 5개이다.

[0031] (d)단계는 0부터 (b)단계에서 카운트한 틀린 횟수 번($Count(data)$)까지의 오답 횟수를 합산한 값($Total_{Count(data)}$)을 전체 데이터 개수에서 틀린 횟수가 0개인 데이터의 개수($Wrong_0$)를 뺀 값으로 나누어 점수를 산출할 수 있다. 이때, (b)단계에서 카운트한 틀린 횟수 번($Count(data)$)이 0인 데이터의 점수는 0이다. (d)단계에서 수행되는 동작은 아래의 [수학식 4]로 표현될 수 있다. 하기 [수학식 3]에서 i가 34인 경우, $Total_{34} = Wrong_1 + Wrong_2 + \cdots + Wrong_{34} = 12{,}089$ 와 같이 계산될 수 있다.

[0032] [수학식 3]

[0033] $$Total_i = \sum_{j=0}^{i} Wrong_j$$

[0034] [수학식 4]

[0035] $$Score_{data} = \frac{Total_{Count(data)}}{\text{전체 데이터 개수} - Wrong_0}, \quad (\text{if Count(data) is 0}, Score_{data} = 0)$$

[0036] 도 3은 본 발명의 다른 실시예인 인공지능 학습데이터의 품질 평가 시스템(10)의 구성도를 나타낸다. 도 3을 참조하면, 인공지능 학습데이터의 품질 평가 시스템(10)은 학습부(100), 제1 카운터부(300), 제2 카운터부(500), 및 산출부(700)를 포함할 수 있다.

[0037] 학습부(100)는 복수개의 딥러닝 모델에 동일한 데이터셋을 학습시킬 수 있다. 학습부(100)는 전술한 (a)단계의 동작을 수행할 수 있다.

[0038] 제1 카운터부(300)는 학습부(100)가 학습에 사용한 데이터가 틀린 횟수를 카운트할 수 있다. 제1 카운터부(300)는 전술한 (b)단계의 동작을 수행할 수 있다.

[0039] 제2 카운터부(500)는 i번 틀린 모든 데이터의 개수를 카운트할 수 있다. 제2 카운터부(500)는 전술한 (c)단계의 동작을 수행할 수 있다.

[0040] 산출부(700)는 0부터 상기 제1 카운터부가 카운트한 틀린 횟수번까지의 오답 횟수를 합산한 값을 전체 데이터 개수에서 틀린 횟수가 0개인 데이터의 개수를 뺀 값으로 나누어 점수를 산출할 수 있다. 산출부(700)는 전술한 (d)단계의 동작을 수행할 수 있다.

[0042] **이하에서는 본 발명의 시뮬레이션 결과를 설명한다.**

[0044] 시뮬레이션에서는 실제 학습을 진행해보고 학습 과정에서 정량적 오류측정을 통해 문제가 되는 데이터를 검출하는 과정을 수행하였다. 이후 앞에서 검출된 오류들을 재태깅하여 학습해보고, 결과를 비교하였다. 시뮬레이션에서 사용된 선학습 모델은 klue/bert-base(이하 klue)와 snunlp/KR-BERT-char16424(이하 snunlp), kykim/bert-korbase(이하 kykim)이다. klue에서 4에폭으로 학습하였기 때문에 본 시뮬레이션에서도 모든 모델을 동일하게 4에폭으로 학습하였다. 결과의 객관성을 위해 랜덤 시드를 임의로 변경하여 각 모델당 10회의 학습을 수행하여, 모델별로 각 40회의 에폭을 학습하였다.

[0045] 실험의 하이퍼파라미터 중 batch size, learning rate, 옵티마이저의 weight decay는 Optuna를 통해 klue에서 최고 성능을 보였던 수치들을 활용하였다. Optuna는 찾을 하이퍼파라미터와 목표를 정하고, 주어진 시간 혹은 학습 횟수에서 목표가 원하는 방향으로 학습되는지 확인하며 최적의 하이퍼파라미터를 찾는다. batch size는 8, 옵티마이저의 weight decay는 0.047, learning rate는 1.621e-05가 나왔다. 실험에 사용된 CPU는 Intel Xeon Gold 6230R이고, GPU는 Nvidia Tesla V100-SXM2-32GB 1개이다.

[0046] 본 시뮬레이션에서는 KLUE에서 공개한 학습 및 검증 데이터셋을 사용하였다. KLUE의 TC 데이터에는 train(45,678개)과 dev(9,107개)가 있다. 주제는 총 7개로 정치, 경제, 사회, 문화, 세계, IT 과학, 스포츠가 있다. 본 시뮬레이션에서는 검증을 위해 train과 dev를 합친 데이터를 사용하였으며, 이를 ALL이라고 칭한다.본 시뮬레이션에서는 ALL 데이터에 오류가 있는지 검증하였다.

[0047] 도 4는 본 발명의 실시예에 따른 ALL 데이터 학습 중 틀린 횟수를 나타내는 그래프이다. 도 4를 참조하면, ALL 데이터로 앞서 설명한 3개의 모델에 대해 모델별로 4에폭씩 10번의 학습에서 $Wrong_i$을 알 수 있다. 모델이 데이터를 학습하며 같은 데이터에 대해 틀리는 횟수가 줄어들기에 $Wrong_i$ 그래프는 1/x 형태가 이상적이다.

[0048] 데이터에 오류가 있음을 검증하고자, 산출 점수가 0.99점이 넘는 620개의 데이터를 3명이 재태깅 후 보팅을 통해 재태깅을 진행하였다. 620개의 데이터 중 3명이 모두 동의한 216개와 2명이 동의한 298개에 대해서 재태깅을 하였고, 3명 전부 동의하지 못한 데이터는 충돌 데이터로 정의하고 데이터셋에서 제외하였다.

[0049] 도 5는 재태깅 데이터를 개별적으로 학습한 결과와 보팅 데이터를 학습시킨 결과를 나타낸다. 도 5를 참조하면, 각 결과에서 기존보다 성능 향상이 있음을 볼 수 있다. 이는 재태깅한 620개의 데이터가 잘못된 라벨을 가지고 있었으며, 재태깅을 통해 적합한 라벨로 변경되었다고 볼 수 있다.

[0050] 도 6은 본 발명의 실시예에 따른 ALL 데이터와 보팅 데이터의 산출 점수를 나타내는 그래프이다. 도 6을 참조하면, 보팅 데이터의 산출 점수를 보면 ALL데이터의에 산출 점수보다 원만해지며 이상적인 log(x) 그래프의 형태와 유사한 것을 확인할 수 있었다.

[0051] 이상에서 대표적인 실시예를 통하여 본 발명을 상세하게 설명하였으나, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 상술한 실시예에 대하여 본 발명의 범주에서 벗어나지 않는 한도 내에서 다양한 변형이 가능함을 이해할 것이다. 그러므로 본 발명의 권리 범위는 설명한 실시예에 국한되어 정해져서는 안 되며, 후술하는 특허청구범위뿐만 아니라 특허청구범위와 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태에 의하여 정해져야 한다.

부호의 설명

[0053] 10 : 인공지능 학습데이터의 품질 평가 시스템

100 : 학습부

300 : 제1 카운터부

500 : 제2 카운터부

700 : 산출부

도면

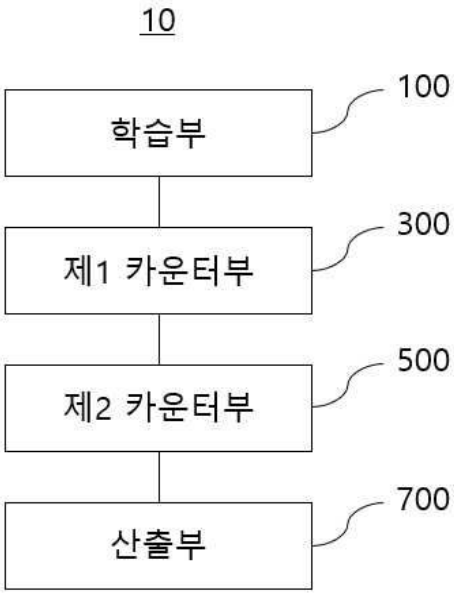
도면1

index	title	count
0	유튜브 내달 2일까지 크리에이터 지원 공간 운영	34
1	미 동부 마비시킨 악성 파일 국내서도 발견	7
2	이해찬 정경두에 안보지원사 기무사 같은 일 없도록 해야	94
3	미래부 17개 부처별 연구비관리시스템 2개로 통합한다	32
4	SK텔레콤안양시 모빌리티·스마트시티 조성 업무협약	41

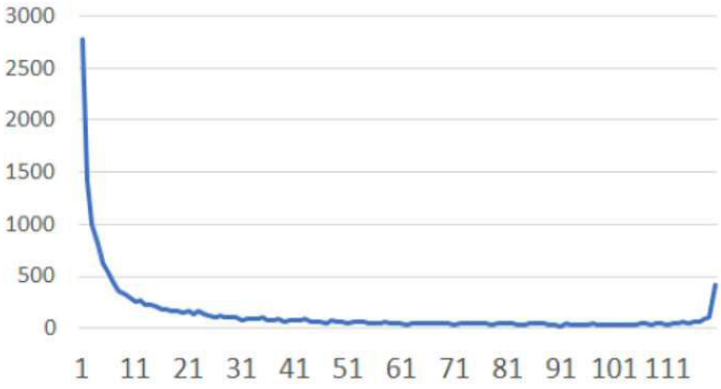
도면2

index	title	count
0	유튜브 내달 2일까지 크리에이터 지원 공간 운영	34
391	지방선거 이긴 日자민당…참의원 선거 앞두고 고지 선점	34
456	김동연 일자리중심 예산 편성…법정시한 내 처리 중요	34
589	김학의 전 차관 기다리는 진상조사단 관계자	34
720	구글 한국 지도 반출 재신청…안보시설 삭제 쟁점	34

도면3



도면4



도면5

	ALL(베이스라인)	개별 재태깅 데이터			보팅 데이터
		person 1	person 2	person 3	
klue	95.82	**96.55(+0.73)**	96.54(+0.72)	96.31(+0.49)	96.495(+0.675)
snunlp	96.62	96.76(+0.14)	**97.59(+0.97)**	96.84(+0.22)	**97.417(+0.797)**
kykim	93.68	93.8(+0.12)	94.17(+0.49)	**94.4(+0.72)**	94.387(+0.707)

도면6

