대화 데이터셋의 클래스 불균형 문제 보정을 위한 적대적 학습 기법

조수필○, 최용석†

한양대학교 컴퓨터·소프트웨어학과, †한양대학교 공과대학 컴퓨터공학부
jessay@hanyang.ac.kr, †cys@hanyang.ac.kr

Adversarial Training Method for Handling Class Imbalance Problems in Dialog Datasets

Su-Phil Cho○, Yong Suk Choi†
Department of Computer Science, Hanyang University
†Division of Computer Science and Engineering, Hanyang University

요 약

딥러닝 기반 분류 모델에 있어 데이터의 클래스 불균형 문제는 소수 클래스의 분류 성능을 크게 저하시킨다. 본 논문에서는 앞서 언급한 클래스 불균형 문제를 보완하기 위한 방안으로 적대적 학습 기법을 제안한다. 적대적 학습 기법의 성능 향상 여부를 확인하기 위해 총 4종의 딥러닝 기반 분류 모델을 정의하였으며, 해당 모델 간 분류 성능을 비교하였다. 실험 결과, 대화 데이터셋을 이용한 모델 학습 시 적대적 학습 기법을 적용할 경우 다수 클래스의 분류 성능은 유지하면서 동시에 소수 클래스의 분류 성능을 크게 향상시킬 수 있음을 확인하였다.

주제어: 대화 의도 분류, 기계학습, 클래스 불균형 문제

1. 서론

데이터의 클래스 불균형 문제는 딥러닝 모델의 학습 성능을 감소시킨다. 이를 해결하기 위해 데이터 재표본(resampling), 모델 앙상블(ensemble), 데이터 증강(data augmentation) 기법 등이 주로 사용되어 왔으나, 위 기법만으로는 충분한 성능 향상이 이뤄지지 않는 경우가 많아 추가적인 기법이 요구되어 왔다.

최근에는 클래스 불균형 문제를 해결하기 위한 방안으로 생성적 적대 신경망(Generative Adversarial Network)[1]을 이용해 데이터 수가 적은 클래스의 데이터를 추가하는 방법 또한 주목받고 있다. 하지만, 한국어 대화 데이터셋의 클래스 불균형 문제를 보완하기 위한 방안으로는 생성적 적대 신경망이 잘 사용되지 않고 있다. 이는 현재까지 공개된 한국어 대화 데이터셋의 종류와 크기가 제한적이며, 이러한 데이터셋으로 학습된 생성적 적대 신경망이 충분한 성능을 갖추지 못하고 있기 때문이다.

이에 본 논문에서는 생성적 적대 신경망 기법과 유사하지만, 데이터 생성을 위한 신경망 학습이 필요 없는 적대적 학습(adversarial training)[2,3] 기법을 사용하였다. 본 논문은 적대적 학습 기법을 통해 적대적 예시[4]를 생성하고, 이를 이용해 대화 의도 분류 모델의 클래스 불균형 문제를 해결하는 방안과 그 결과를 제시한다.

한국어 대화 의도 분류 모델에 적대적 학습 기법을 적용한 결과, 클래스 불균형 문제가 있는 데이터로 학습된 모델의 분류 성능이 기존에 비해 크게 향상되었음을 확인하였다. 이에 더해, 적대적 학습 기법은 기존의 클래스 불균형 문제 해결 방안과 동시에 적용이 가능하므로 추후에는 타 보정 기법과의 연계를 통한 추가적인 모델 성능 향상을 기대해볼 수 있다.

앞으로 본 논문의 구성은 다음과 같다. 2장에서는 논문의 내용과 관련된 기존 연구들을 소개한다. 3,4장에서는 적대적 예시를 이용한 학습 기법 및 적용 방안을 각각 제시한다. 5장에서는 실험 데이터 및 실험 방법에 대해 설명한다. 6장에서는 실험의 결과를 제시하고 분석한다. 7장에서는 본 논문에 대한 전반적인 결론을 내린다.

2. 관련 연구

2.1. 클래스 불균형 문제

클래스 불균형(class imbalance)은 데이터 별로 클래스가 정의된 데이터셋에서 특정 클래스의 데이터 수가 타 클래스의 데이터 수와 크게 차이나는 경우를 말한다[5]. 클래스 불균형 문제가 있는 데이터셋으로 딥러닝 모델 학습을 진행할 경우, 데이터 수가 적은 소수(minority) 클래스의 데이터는 잘 학습되지 않고 다수(majority) 클

† 교신저자(Corresponding author): 한양대학교 공과대학 컴퓨터공학부 교수 최용석(cys@hanyang.ac.kr)

래스의 데이터만을 위주로 모델이 학습되는 문제가 발생한다[6]. 이로 인해, 분류 모델의 경우 다수 클래스는 잘 분류하지만 소수 클래스는 제대로 분류하지 못하는 문제가 발생하게 된다.

2.2. 딥러닝 기반 대화 의도 분류 모델

딥러닝 기반 분류 기법은 신경망 모델을 설계한 후 학습 과정을 통해 각 데이터의 특징을 신경망 모델 내부의 파라미터가 반영토록 하여 최종적으로는 새로운 데이터를 입력할 경우에도 이를 정확하게 분류할 수 있도록 모델을 학습시키는 기법이다. 최근에는 해당 기법을 이용하여 사용자 대화문의 의도를 분류하는 연구가 다수 진행되고 있다. Ravuri et al.[7]은 순환 신경망 기반 대화 의도 분류 모델에 대한 연구를 진행하였으며, Lee et al.[8]은 순환 신경망 기반 및 합성곱 신경망 기반 모델을 포함하여 다수의 확률 모델 및 기계학습 기반의 대화 의도 분류 모델 간 성능 비교를 진행한 바 있다. 이러한 연구들을 통해, 최근 연구되고 있는 대화 의도 분류 모델 중 가장 높은 분류 성능을 확보하는 모델은 딥러닝 기반 분류 모델임이 검증되었다.

2.3. 적대적 학습

적대적 학습 기법은 Goodfellow et al.[2,3]에서 제시된 딥러닝 기법으로, 기존 데이터에 섭동(perturbation)을 가한 적대적 예시 데이터를 만들고, 이를 모델에 추가적으로 학습시키는 방식이다. 적대적 예시는 기존 데이터와 유사하면서, 모델 학습 시 발생하는 비용함수 값이 큰 데이터를 선별적으로 찾아 사용한다. 적대적 예시를 이용한 학습 기법은 식(1)과 그림 1과 같이 진행된다.

$$L_{adv} = -\log p(y|x + r_{adv}; \theta) \quad (1)$$

이때, 적대적 예시 데이터를 찾아 모델에 학습시키는 것은 아래와 같은 효과를 발생시킨다:

- 새로운 데이터를 확보하여 데이터 부족 문제 해결
- 손실 함수가 크게 발생하는 데이터를 선별하여 학습
- 같은 데이터를 반복적으로 학습할 때도 적대적 예시는 매번 새로 탐색하여 학습시키므로 딥러닝 모델의 과적합(overfitting) 문제를 방지

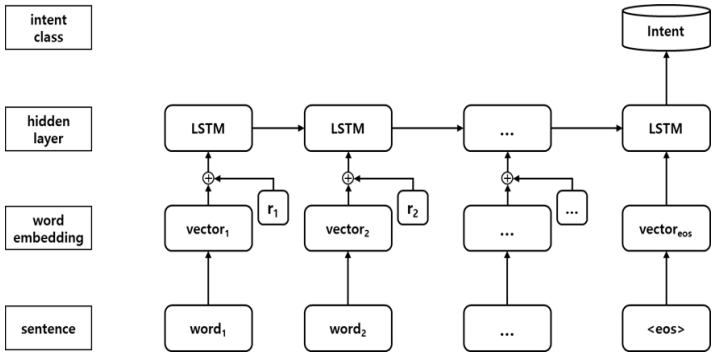


그림 1 적대적 학습 기법을 적용한 순환 신경망 모델

적대적 학습 기법을 대화 의도 분류 모델에 적용하는 자세한 방안은 아래 3장에서 소개한다.

3. 적대적 학습 기법을 이용한 대화 의도 분류 모델

적대적 학습 기법을 적용한 대화 의도 분류 모델은 그림 3과 같으며, 학습 과정은 총 3단계로 구분된다.

첫 번째는 기존 데이터셋의 한국어 대화 텍스트와 대화 의도 정보를 이용한 순환 신경망 학습 시 발생하는 비용함수와 역전파 값을 확인하는 단계이다. 본 논문에서는 학습 단계의 비용함수로 cross-entropy 함수를 사용하였으며, 이는 식 (2)와 같이 표현된다.

$$L_{ce} = -\log p(y|x; \theta) \quad (2)$$

이때 y는 대화 의도 클래스, x는 대화 텍스트, θ는 신경망 모델의 파라미터 정보이다.

이를 통해, 학습 데이터 x 에 역전파되는 값 $\boldsymbol{g}^*$은 다음과 같다.

$$\boldsymbol{g}^* = \nabla_x - \log p(y|x; \theta) \quad (3)$$

두 번째는 앞서 계산된 역전파 값 $\boldsymbol{g}^*$ 값을 이용한 적대적 예제 선정 단계이다. 적대적 예제는 Goodfellow et al.[2]에서 제시된 역전파 함수의 선형성 가정을 이용하여 생성되며, 이를 통해 데이터 x에 $\boldsymbol{g}^*$의 반대 방향으로 역전파가 이루어진 가상의 데이터 $\bar{x}$ 를 적대적 예제로 선정한다. 이는 수식으로 다음과 같이 표현된다.

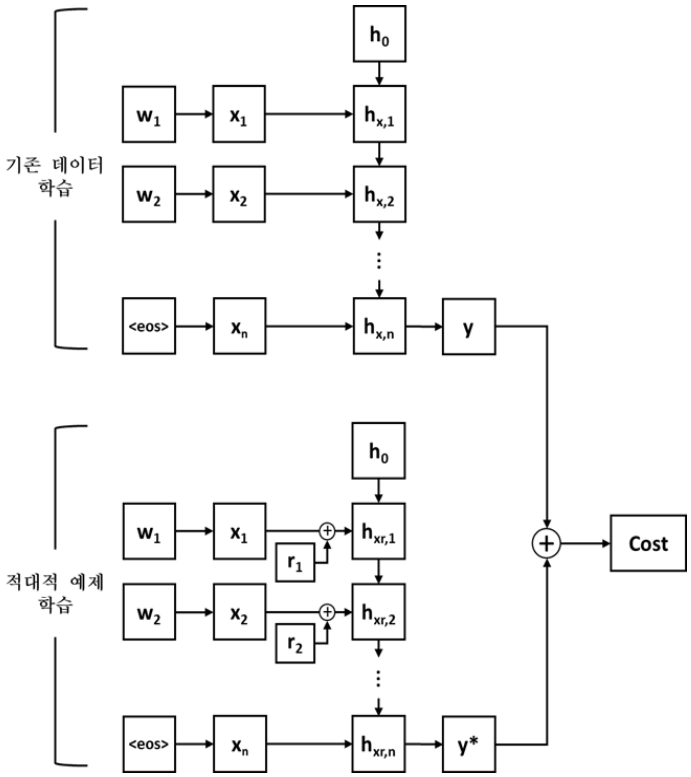


그림 2 적대적 학습 기법을 적용한 대화 의도 분류 모델

$$\bar{x} = x + r_{adv} \quad (4)$$

$$r_{adv} = -\epsilon \boldsymbol{g}/\|\boldsymbol{g}\|_2 \; where \; \boldsymbol{g} = \nabla_x \log p(y|x;\theta) \quad (5)$$

여기서 ϵ은 적대적 예시를 결정하기 위한 파라미터이고, $1/\|\boldsymbol{g}\|_2$은 $\boldsymbol{g}$에 대한 L_2 norm constraint 연산이다.

세 번째는 기존의 데이터와 적대적 예시를 딥러닝 모델에 학습시키는 과정이다. 이는 앞서 계산된 2개의 비용함수를 더하고 이를 역전파하는 방식으로 구현되며, 그 과정은 그림 2과 같다. 이 때의 비용함수는:

$$L = L_{ce} + L_{adv} \quad (6)$$

이때 비용함수 L 이 최소가 되는 방향으로 모델을 학습시키면, 기존 데이터와 적대적 예시 데이터를 모두 학습한 분류 모델이 생성된다.

4. 클래스 불균형 해결을 위한 적대적 학습 적용 방안

적대적 학습 기법의 클래스 불균형 문제 해결 능력을 검증하기 위해 다음 4가지 방식의 학습 기법을 설계하였다. 이때 x는 데이터, y는 클래스, θ는 현재까지 학습된 모델의 가중치, ϵ는 적대적 학습의 최대 섭동 범위를 지정하는 인자이다. 그림 3에서 각 모델의 학습 방법을 가시적으로 설명하였다.

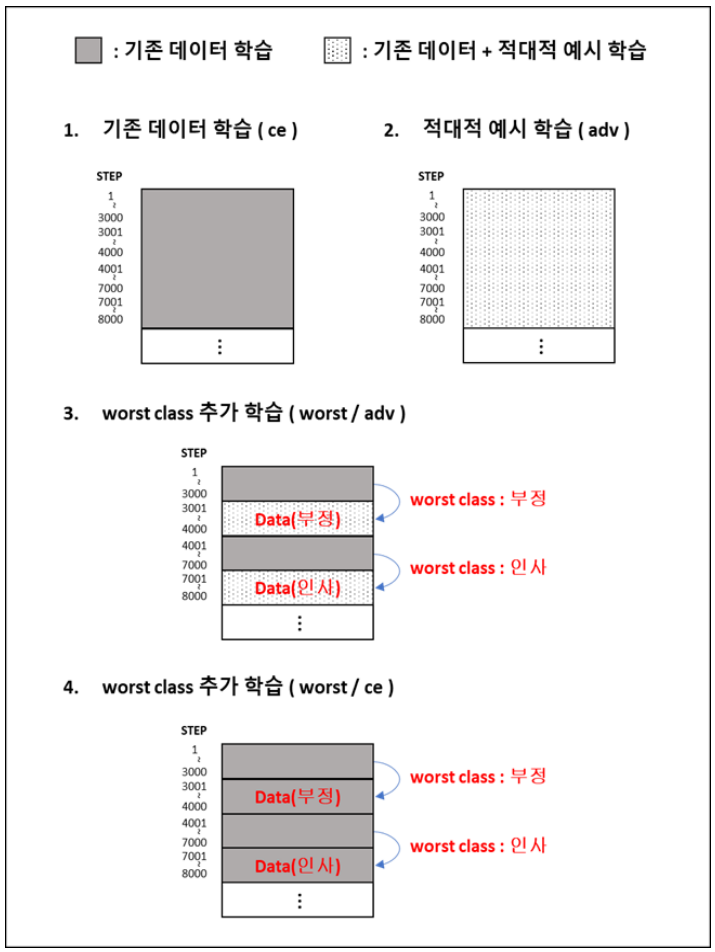


그림 3 적대적 학습의 성능 평가를 위한 4가지 학습 모델

(1) 기존 데이터 학습 모델(ce). 적대적 학습을 전혀 하지 않는 일반적인 딥러닝 모델이다. 학습 시 비용 함수는:

$$cost = L_{ce} \quad (7)$$

$$L_{ce} = -\log p(y|x;\theta) \quad (8)$$

(2) 적대적 예시 학습 모델(adv). 학습 데이터와 적대적 예시를 모두 모델에 학습시킨다. 학습 시 비용함수는:

$$cost = L_{ce} + L_{adv} \quad (9)$$

$$L_{adv} = -\log p(y|x + r_{adv};\theta) \quad (10)$$

$$r_{adv} = -\epsilon \boldsymbol{g}/\|\boldsymbol{g}\|_2 \; where \; \boldsymbol{g} = \nabla_x \log p(y|x;\theta) \quad (11)$$

(3) Worst class에 대한 적대적 예시 추가 학습 모델(worst/adv). 모델 (1)과 같이 cross-entropy 만을 이용하여 학습을 진행하다가, 일정 횟수의 학습이 완료되면 해당 시점까지 학습된 모델의 클래스 별 정확도를 확인하고 그 중 가장 정확도가 낮은 n개의 클래스에 대하여 적대적 학습을 추가 진행한다. 학습 시 비용 함수는:

$$cost = \begin{cases} L_{ce} & (\text{기존 데이터}) \\ L_{ce} + L_{adv} & (\text{추가 데이터}) \end{cases} \quad (12)$$

(4) Worst class에 대한 기존 데이터 추가 학습 모델(worst/ce). 모델 (3)에 대한 대조군으로 설정된 모델로, 클래스 별 정확도가 가장 낮은 클래스에 대해서 데이터셋 내 데이터를 추가적으로 학습시킨다. 학습 시 비용 함수는:

$$cost = \begin{cases} L_{ce} & (\text{기존 데이터}) \\ L_{ce} & (\text{추가 데이터}) \end{cases} \quad (13)$$

이때, 위 4가지 모델 간의 성능 비교를 통하여 다음과 같은 정보를 확인할 수 있다.

- ce ↔ adv를 비교하여, 적대적 학습 기법의 분류 성능 향상 효과를 확인할 수 있다.
- ce, adv ↔ worst/adv를 비교하여, 적대적 예시를 이용한 추가 학습이 클래스 불균형 문제를 보정할 수 있는지 확인할 수 있다.
- worst/adv ↔ worst/ce를 비교하여, 추가 학습 기법과 적대적 학습 기법을 함께 적용하는 것이 유의미한 성능 향상 효과를 만드는 지 확인할 수 있다.
- 본 논문에서 제시하는 타겟 모델 worst/adv 과 그 외의 모델 간의 성능 비교를 할 수 있다.

5. 실험 방법

5.1. 실험 데이터

본 논문에서 사용한 데이터셋은 한국 영화 및 드라마 대사를 기반으로 제작된 한국어 대화 의도 데이터이며, 총 29,513개의 한국어 대화문으로 이루어져 있다[9]. 각 데이터에는 7가지 대화 의도 클래스 중 1개가 태깅되어 있다. 그림 4는 해당 데이터셋의 예시이다.

응 5
그런 실밥은 그냥 라이터 불로 쓰윽 지지면 없어져요 0
잠깐만요 3
그냥 뜯으면 돼, 괜찮아 0
주문한 거 나왔는데요 0
뭐해? 2
커피 한잔 주세요 3
여보세요? 1

그림 4 한글 대화 의도 데이터셋 예시

해당 데이터셋은 클래스 간 데이터 수가 불균등하며, 전체 데이터의 12% 이상을 각각 보유하고 있는 3종의 다수 클래스와, 3% 미만을 각각 보유하고 있는 4종의 소수 클래스로 구분된다. 각 클래스 별 데이터 비율은 표 1과 같다.

표 1 대화 의도 데이터셋의 클래스 별 데이터 비율

클래스	데이터 수	비율(%)
단순정보전달	14,937	50.61
인사	213	0.72
질문	8,997	30.48
요청/제안/명령	3,730	12.64
약속	536	1.82
긍정	783	2.65
부정	317	1.07
전체	**29,513**	**100**

5.2. 실험 환경

본 논문에서 연구한 딥러닝 기반 대화 의도 분류 모델은 다음과 같은 환경에서 학습되었다.

5.2.1. 입력 데이터 및 임베딩 벡터

입력 데이터는 앞 절에서 소개한 한글 대화 의도 데이터셋을 사용하였으며, 실험에 앞서 데이터셋 전체를 unigram 단위로 분석하여 상위 빈도 단어에 대한 어휘 사전을 생성하였다. 이후 어휘 사전에 등록된 단어에 대하여 한글 임베딩 벡터[10]를 적용한 값을 입력 데이터로 사용하였다. 본 실험에서는 위키피디아의 한글 텍스트를 기반으로 사전 학습된 300차원의 fasttext 임베딩 벡터를 적용하였다.

5.2.2. 순환 신경망 모델

실험에 사용한 순환 신경망 모델은 LSTM[11] 셀을 사용하였으며, LSTM 셀의 크기는 1024차원으로 설정하였다. 클래스 분류를 위한 fully-connected 층은 30차원의 단층 구조로 설계하였다.

학습 최적화에는 Adam[12] 최적화 기법을 사용하였으며 해당 기법을 통한 모델의 학습률은 0.0005로 초기 설정한 후 매 학습 시에 학습률이 0.9998배로 감소하도록 exponential decay를 적용하였다.

입력 데이터의 mini-batch 크기는 32로 설정하였으며, 최대 50,000 epoch의 학습이 진행되도록 설계하였다.

5.2.3. 적대적 학습 기법

4장에서 설명한 4가지 학습 모델에 맞게끔 비용함수를 조절해 가며 학습을 진행하였다. worst/adv 모델과 worst/ce 모델의 경우에는 기존 데이터셋에 대한 학습이 8,000 epoch 이루어질 때마다 각 클래스 별 정확도를 측정하여 정확도가 가장 낮은 3개의 클래스에 대해 순서대로 각각 1000, 500, 500 epoch의 추가 학습을 진행하였다.

6. 실험 결과 및 분석

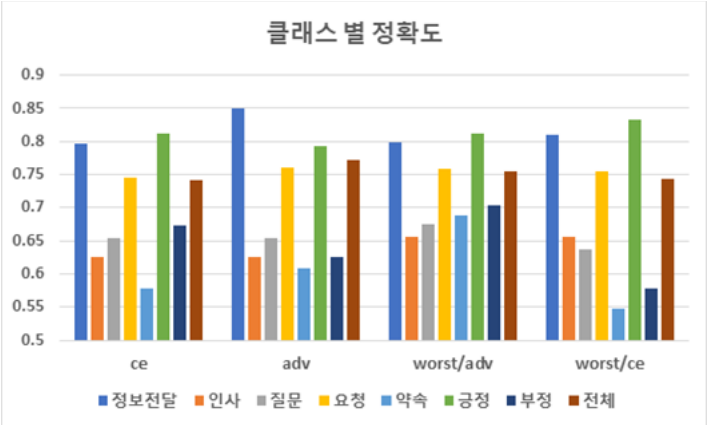
표 2 학습 모델에 따른 클래스 별 정확도 및 Recall

	ce	adv	worst/adv	worst/ce
단순정보전달	0.797	**0.849**	0.798	0.809
인사	0.625	0.625	**0.656**	**0.656**
질문	0.653	0.653	**0.675**	0.637
요청/제안/명령	0.745	**0.760**	0.758	0.755
약속	0.578	0.609	**0.688**	0.547
긍정	0.812	0.792	0.812	**0.833**
부정	0.672	0.625	**0.703**	0.578
Accuracy	0.742	**0.771**	0.754	0.744
Macro Recall	0.697	0.702	**0.727**	0.688

표 2는 적대적 학습 방식에 따른 모델 별 분류 성능을 나타낸다. 이를 4장에서 언급한 비교 방식을 통해 분석하면 다음과 같다.

- 기존 ce 모델에 비해, adv 모델에서는 다수 클래스 2종에 대한 분류 성능이 크게 증가하였으며 이는 전체 데이터에 대한 정확도 증가로 이어졌다. 이를 통해, 적대적 학습 기법이 모델의 분류 성능을 증가시켰음을 확인할 수 있다.
- worst/adv 모델에서는 전체 데이터 중 1%, 2%, 2%만을 차지하는 소수 클래스 [인사, 약속, 부정]의 분류 정확도가 가장 크게 측정되었으며, 이를 기반으로 Macro Recall 또한 가장 높게 측정되었다. 이를 통해, 본 논문에서 제시한 worst/adv 모델이 클래스 불균형 문제를 보정한다는 것을 확인할 수 있다.
- worst/adv 모델과는 달리, worst/ce 모델에서는 분류 성능이 거의 향상되지 않았음을 확인할 수 있다. 이를 통해, 적대적 학습 기법이 모델의 분류 성능을 크게 향상시키는 요소임을 알 수 있다.

특히, 그래프 1에서는 worst/adv 모델만이 모든 클래스에서 분류 정확도 65% 이상을 확보한 것을 확인할 수 있다. 이를 통해, 데이터셋의 클래스 불균형 문제로 소수 클래스의 분류 정확도가 크게 떨어진다면 이는 worst/adv 방식의 학습을 통해 정확도를 크게 향상시킬 수 있음을 알 수 있다.



그래프 1 모델 별 분류 정확도 차트

표 3에서는 worst/adv 모델의 매 학습 단계 별 클래스 분류 정확도를 확인할 수 있다. 해당 표를 통해 전체 데이터에 대한 학습이 완료되는 매 8,000 epoch 학습 주기마다 소수 클래스에 해당하는 [인사, 약속, 부정] 클래스의 분류 정확도가 낮게 측정된 바 있으며, 이후 진행된 2,000 epoch의 추가 학습 과정을 거치면서 해당 클래스의 분류 정확도가 타 클래스에 비해 큰 폭으로 상승하였음을 확인할 수 있다.

또한 표 3에서, worst/adv 모델의 학습 초기에는 추가 학습을 진행한 클래스에 대한 과적합 현상이 일시적으로 발생하지만 이는 추후 학습이 진행되면서 점차 완화되어 가는 것을 확인할 수 있다. 이를 통해, 추가 학습 방식으로 인한 과적합 현상을 학습 초기부터 최소화할 경우 추가적인 모델 성능 향상 또한 기대해볼 수 있다.

표 3 worst/adv 모델의 학습 epoch 별 분류 정확도

epoch	정보전달	인사	질문	요청	약속	긍정	부정	전체
8000	0.811	**0.638**	0.652	0.755	**0.578**	0.802	**0.641**	0.750
10000	0	**0.719**	0	0	**1.000**	0	**0.391**	0.029
18000	0.797	**0.688**	**0.669**	0.753	**0.578**	0.812	0.703	0.749
20000	0.541	**0.688**	**0.730**	0.758	**0.906**	0.823	0.875	0.642
28000	0.807	**0.656**	**0.667**	0.763	**0.500**	0.802	0.719	0.753
30000	0.779	**0.656**	**0.688**	0.768	**0.703**	0.812	0.703	0.749
38000	0.805	**0.656**	**0.669**	0.758	**0.688**	0.812	0.703	0.756
40000	0.800	**0.656**	**0.674**	0.760	**0.688**	0.812	0.703	0.755
48000	0.799	**0.656**	**0.674**	0.760	**0.688**	0.812	0.719	0.754
50000	0.798	**0.656**	**0.675**	0.758	**0.688**	0.812	0.703	0.754

7. 결론 및 향후 연구

본 논문에서는 적대적 학습 기법을 이용하여 클래스 불균형 문제를 보정하는 방안을 제시하였다. 적대적 학습의 성능을 평가하기 위해 총 4가지 학습 방식을 제시하였고, 그 중 적대적 학습 기법과 최저 성능 클래스에 대한 추가 학습 기법을 모두 적용한 모델 worst/adv에서 모델의 분류 성능이 가장 크게 향상되었음을 확인하였다.

한편, 적대적 학습 기법은 데이터 재표본, 모델 앙상블 등 기존의 클래스 불균형 문제 보정 기법과 동시에 적용할 수 있다. 따라서, 향후 기존의 클래스 불균형 보정 기법들과 worst/adv 학습 기법을 동시에 적용하였을 때 모델의 분류 성능을 평가하는 연구가 진행되어야 한다. 또한, 추가 학습으로 인한 과적합 현상을 최소화할 경우 해당 모델의 성능 변화가 발생하는지 확인해야 한다.

사 사

본 연구는 2019년도 산업통상자원부 및 산업기술평가관리원(KEIT) 연구비 지원에 의한 연구이며(과제번호:10077553), 산업통상자원부의 재원으로 기술혁신사업의 지원을 받아 수행한 연구 과제 (No. 10060086, 개인서비스용 로봇을 위한 지능-지식 집약 ·개방 ·진화형 로봇지능 소프트웨어 프레임워크 기술 개발)입니다

참고문헌

[1] Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems. 2014.

[2] Miyato, Takeru, Andrew M. Dai, and Ian Goodfellow. "Adversarial training methods for semi-supervised text classification." arXiv preprint arXiv:1605.07725 (2016).

[3] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).

[4] Szegedy, Christian, et al. "Intriguing properties of neural networks." arXiv preprint arXiv:1312.6199 (2013).
[5] Ramyachitra, D., and P. Manikandan. "Imbalanced dataset classification and solutions: a review." International Journal of Computing and Business Research (IJCBR) 5.4 (2014).
[6] Guo, Hongyu, and Herna L. Viktor. "Learning from imbalanced data sets with boosting and data generation: the databoost-im approach." ACM Sigkdd Explorations Newsletter 6.1 (2004): 30-39.
[7] Ravuri, Suman, and Andreas Stolcke. "Recurrent neural network and lstm models for lexical utterance classification." Sixteenth Annual Conference of the International Speech Communication Association. 2015.
[8] Lee, Ji Young, and Franck Dernoncourt. "Sequential short-text classification with recurrent and convolutional neural networks." arXiv preprint arXiv:1603.03827 (2016).
[9] 오주민, 조수필, 임영수, 최용석. (2018). 계층 구조 Attention 기반 순환 신경망을 이용한 발화 의도 분류 성능 개선. 한국정보과학회 학술발표논문집, 575-577.
[10] Bojanowski, Piotr, et al. "Enriching word vectors with subword information." Transactions of the Association for Computational Linguistics 5 (2017): 135-146.
[11] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.
[12] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).