신뢰성이 부족한 FAQ 데이터셋에서의 강건성 개선을 위한 모델의 예측 강도 기반 손실 조정 정규화

박예원O, 양동일, 김수필, 이강욱

삼성전자, 삼성리서치
{yw1222.park, di87.yang, sf.kim, kw.brian.lee}@samsung.com

Loss-adjusted Regularization based on Prediction for Improving Robustness in Less Reliable FAQ Datasets

Yewon ParkO, Dongil Yang, Soofeel Kim, Kangwook Lee
Samsung Electronics, Samsung Research

요 약

FAQ 분류는 자주 묻는 질문을 범주화하고 사용자 질의에 대해 가장 유사한 클래스를 추론하는 방식으로 진행된다. FAQ 데이터셋은 클래스가 다수 존재하기 때문에 클래스 간 포함 및 연관 관계가 존재하고 특정 데이터가 서로 다른 클래스에 동시에 속할 수 있다는 특징이 있다. 그러나 최근 FAQ 분류는 다중 클래스 분류 방법론을 적용하는 데 그쳤고 FAQ 데이터셋의 특징을 모델에 반영하는 연구는 미미했다. 현 분류 방법론은 이러한 FAQ 데이터셋의 특징을 고려하지 못하기 때문에 정답으로 해석될 수 있는 예측도 오답으로 여기는 경우가 발생한다. 본 논문에서는 신뢰성이 부족한 FAQ 데이터셋에서도 분류를 잘 하기 위해 손실 함수를 조정하는 정규화 기법을 소개한다. 이 정규화 기법은 클래스 간 포함 및 연관 관계를 반영할 수 있도록 오답을 예측한 경우에도 예측 강도에 비례하여 손실을 줄인다. 이는 오답을 높은 확률로 예측할수록 데이터의 신뢰성이 낮을 가능성이 크다고 판단하여 학습을 강하게 하지 않게 하기 위함이다. 실험을 위해서는 다중 클래스 분류에서 가장 좋은 성능을 보이고 있는 모형인 BERT를 이용했으며, 비교 실험을 위한 정규화 방법으로는 통상적으로 사용되는 라벨 스무딩을 채택했다. 실험 결과, 본 연구에서 제안한 방법은 기존 방법보다 성능이 개선되고 보다 안정적으로 학습이 된다는 것을 확인했으며, 데이터의 신뢰성이 부족한 상황에서 효과적으로 분류를 수행함을 알 수 있었다.

주제어: 자연어 처리, FAQ, 손실 함수, 정규화

1. 서론

FAQ(Frequently Asked Questions)는 빈도가 높은 질문에 답변을 달아 질의-응답 쌍을 구축한 것으로, 사용자가 제품이나 서비스에 관한 정보를 얻는 데 주로 사용되는 방법 중 하나이다. FAQ 에 대한 답변을 자동화하기 위해 자연어 처리 분야에서는 초기에 정보 검색(Information Retrieval)을 활용하였다. 하지만 이 방법은 동음 이의어, 다의어 등 모호한 어휘를 다루는데 취약했다(Lexical disagreement). 그 후, 자연어 처리 분야에서 딥러닝이 활발히 사용되면서 FAQ 분류에도 딥러닝이 도입되었는데, 사용자의 질문을 적합 답변 등의 클래스(Class)로 분류한다는 점에서 다중 클래스 분류(Multi-class classification) 방법이 사용되었으며, 이를 실현하기 위한 방법론으로는 텍스트와 같은 시계열 데이터에 주로 사용되던 RNN 과 더불어 이미지 처리 연구에 많이 이용되는 CNN 이 활용되었다. 현재는 트랜스포머 구조 기반의 BERT 가 소개된 이후, 사전 학습(Pre-training)과 본 학습(Fine-tuning)의 두 단계로 구성된 전이 학습(Transfer Learning) 기반의 방법론이 보편화되었다.

하지만, FAQ 데이터셋 특성상 일반적으로 답변에 해당되는 클래스가 다수 존재하기 때문에 클래스 간 포함 및 연관 관계가 존재할 수 있다. 이로 인해, 특정 질의 데이터가 여러 클래스에서 정답으로 인정되는 현상, 즉 데이터가 서로 다른 클래스에 동시에 속하는 현상이 빈번히 발생할 수 있다. 표 1은 냉장고 고장문의 도메인에서 데이터가 서로 다른 클래스에 동시에 속할 수 있는 경우의 예시를 보여준다.

표 1. 서로 다른 클래스에 동시에 속하는 경우

클래스 1	냉장고가 안 차가워요.
클래스 2	냉동실이 안 차가워요.
질의	냉동실이 하나도 안 시원해요.

표 1 과 같이, FAQ 클래스 구분은 배타적이지 않다는 특징이 존재한다. “냉동실이 하나도 안 시원해요.”라는 질의는 엄밀하게는 클래스 2 가 정답이지만 클래스 1 도 정답이 될 수 있다. 냉동실은 냉장고의 일부이기 때문이다. 하지만 데이터셋에는 위 질의에 대한 정답이 클래스 2 로 표시되어 있기 때문에 모델 학습 시 클래스 1 을 정답으로 예측한 경우 틀린

판단으로 간주하고 만다. 이처럼 기존 FAQ 데이터셋은 클래스 간 포함 및 연관 관계를 드러내지 않기 때문에 데이터의 신뢰성이 낮다. 결국 이러한 데이터셋의 특성을 고려한 학습이 이루어지지 않기 때문에 서로 다른 클래스에 동시에 속할 수 있는 데이터에 대해 분류기는 낮은 성능을 보이게 된다.

본 논문에서는 이러한 FAQ 데이터셋 및 클래스의 특징을 모델에 반영하는 방법을 제안한다. 제안 방법은 데이터의 낮은 신뢰성을 고려하기 위해서 학습 과정 중 오답 예측 시에 예측 강도에 기반하여 손실 함수에 정규화를 적용한다.

성능 비교 연구를 위한 기본 모형으로는 현재 가장 좋은 성능을 보이고 있는 BERT 를 사용했으며, 잡음 데이터에서 모델의 강건성을 높인다고 알려진 라벨 스무딩(Label Smoothing)을 비교 대상으로 채택하였다. 실험 데이터로는 내부적으로 제작한 가전 제품 고장 및 사용법 문의 데이터를 사용하였다. 본 데이터는 클래스 수가 166 개로, 클래스 간 유사 정도가 다양하고 클래스가 배타적으로 구분되지 않는 특징이 존재한다.

비교 실험 결과, 본 연구에서 새로이 제안하는 방법은 기본 모형과 비교해 FAQ 분류에 높은 성능을 보임을 알 수 있었다. 또한, 라벨 스무딩과 비교했을 때 좀 더 안정적으로 학습을 수행함을 확인할 수 있었다.

2. 관련 연구

2.1. 딥러닝 기반의 다중 클래스 분류

다중 클래스 분류(Multi-class classification) 중에서도 클래스 수가 매우 많은 문제는 근래에 기계 학습 커뮤니티에서 많은 주목을 받았다. 자연어 처리 분야에서는 텍스트와 같은 시계열 데이터에 많이 사용되는 RNN과 더불어 이미지 처리 연구에 많이 활용되었으나 문장 분류에서 좋은 성능을 보인다고 보고된 CNN이 주로 사용[1]되었다. 그 후, 장기 의존성 문제(Long-term dependency)를 극복하는 데 효과적인 어텐션(Attention)을 적용한 연구[2]도 이루어졌다. 현재는 트랜스포머 구조 기반의 BERT[3]가 높은 성능을 보이며 다중 클래스 분류 연구에 보편적으로 활용되고 있다.

2.2. 잡음 데이터 학습 방법

딥러닝의 학습 과정에서는 입력 데이터가 주어지면 여러 가중치가 연결된 형태의 층을 지나 데이터의 목적 값을 예측하며, 이 때 정답과 예측의 차이를 통해 손실을 계산하게 된다. 계산된 손실은 역전파를 통해 전달되고, 올바르게 목적 값을 예측할 수 있도록 가중치가 업데이트 된다. 이와 같이, 손실은 학습에 아주 민감하게 작용하는, 학습에 중요한 요소 중 하나이다. 전통적으로 많이 사용되는 손실 계산 방법은 정답과 예측의 차이를 제곱하거나(Mean Squared Error) 음의 로그 우도(Negative log likelihood)를 사용하는 방법(Cross Entropy Error)이 있는데, 정답이 올바르지 않은 값으로 입력될 경우 손실 계산에 오차가 발생하여 올바른 역전파가 이루어지기 어렵다. 올바르지 않은 목적 값을 포함한 데이터를 잡음 데이터라 하는데 이를 해결하고자 라벨 스무딩, L2 정규화 등 다양한 정규화 방법이 제안 되었으며, 잡음 데이터에서의 성능 개선을 위해 많이 사용되고 있다.

3. FAQ 분류

3.1. BERT 기반의 다중 클래스 분류

BERT는 트랜스포머 구조[4]의 일부인 멀티 헤드 어텐션(Multi-head attention) 및 잔류 연결(Residual connection)로 구현된 인코딩 블록으로 구성되어 있으며, 사전 학습 후 본 학습을 거치는 전이 학습을 진행한다. 사전 학습 단계에서는 입력된 토큰의 일부를 [MASK] 토큰으로 변경한 후 변경 전 토큰을 예측하는 MLM(Masked Language Model) 태스크와 두 문장의 관계를 학습하기 위한 NSP(Next Sentence Prediction) 태스크가 함께 학습된다. 이후 본 학습 단계에서는 학습된 언어 모델의 은닉 계층 출력과 태스크를 위해 추가된 은닉 계층의 완전 결합을 통해 새로운 태스크 모델을 구성하고 활성 함수를 통해 예측을 진행하게 된다.

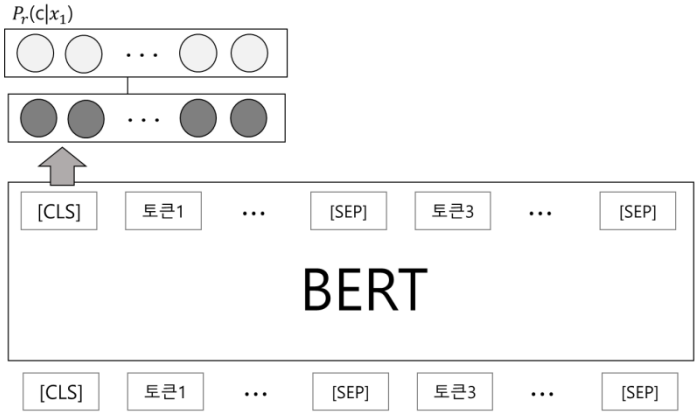

그림 1. BERT 모델 구조

다중 클래스 분류는 일반적으로 소프트맥스 함수(Softmax function)를 활성 함수로 사용한다. 소프트맥스는 예측 값에 지수 함수를 적용하되 모든 뉴런에서 나온 값으로 정규화하는 형태를 가진다. 예측 총 합이 1이 되는 성질 때문에 예측 값을 확률로 해석하여 다중 클래스 분류에 용이하다.

$$\mathrm{f}(x_i) = \frac{e^{x_i}}{\sum_{k=1}^{K} e^{x_k}} \; for\; i = 1, \ldots, K \qquad (1)$$

이 때, i는 클래스 번호, K는 클래스 개수를 뜻한다. 손실 함수는 교차 엔트로피(Cross entropy)를 주로 사용한다.

$$CE = -\sum_{i=1}^{n} P(y_i) \log P(\hat{y}_i) \qquad (2)$$

식 (2)는 본 실험의 Baseline인 BERT base 모델의 본

학습(Fine-tuning) 부분의 손실 함수이다. 이 손실 함수를 최소화하는 방향으로 분류기 학습을 진행한다.

3.2. 기존 정규화 방법(라벨 스무딩)

라벨 스무딩[5]은 비교적 간단한 정규화 방법인데, One-hot 인코딩 방식이 사용된 목적 값에 아주 작은 값(ε)을 더하거나 빼서 이를 목적 값으로 이용한다.

$$y_i' = (1-\varepsilon)y_i + \frac{\varepsilon}{K} \quad (3)$$

식 (3)에서 ε은 학습 데이터에 따라 변경해야 하는 변수로써, 각 데이터 및 모델에 적합한 ε을 구하기 위해서는 별도의 실험이 필요하다. 라벨 스무딩은 모든 목적 값에 동일한 변화를 적용하여 모델이 과신(Over confident)하는 것을 방지한다.

3.3. 예측 강도 기반 손실 조정 정규화

서론에서 언급한 것과 같이, FAQ 분류의 성능을 높이기 위해서는 클래스의 포함 및 연관 관계가 존재하는 데이터 분류 문제를 해결해야 한다. 표 1에서 언급된 바와 같이, "냉동실이 하나도 안 시원해요."라는 질의는 클래스 1과 2에 모두 속할 수 있지만, 학습 시 모델이 위 질의를 클래스 2로 예측할 경우에만 정답으로 인정되고 클래스 1로 예측할 경우 오답으로 인정된다.

본 연구는 모델의 예측 값이 오답인 경우 실제로는 오답이 아닐 수도 있다는 가능성을 열어둔다. 이를 위하여 모델의 손실 함수에 가중치(Weight)를 적용하여 예측 값이 오답일 시 예측 강도에 비례하여 기존보다 약하게 학습하도록 한다. 오답 예측 시 추가적으로 손실 함수에 가중치를 반영하는 방법을 비용 민감 학습(Cost Sensitive learning)[6]이라 하며, 이는 각 예측의 정확도에만 의존하지 않고 비용까지 고려하여 최종적으로 전체 손실을 줄이는 기계 학습 방법론으로 알려져 있다.

데이터셋에서 신뢰도가 낮은 데이터를 구분하기 위해 학습 시 모델이 판단하는 예측 값을 사용하였다. 모델이 오답이라고 판단한 상황 중, 강하게 오답으로 판단할 수록(예측 값이 클 수록) 데이터의 신뢰도가 낮을 가능성이 크다 여기고 학습 강도를 줄인다. 이 과정을 통해 각 데이터의 신뢰도에 관계없이 비슷한 수준의 손실이 나오도록 강제하는 정규화 역할을 수행하게 된다.

$$Loss = -\sum_{i=1}^{n} w_R P(y_i)\log P(\hat{y}_i) \quad (4)$$

$$w_R = \begin{cases} 1 & (y_i = \hat{y}_i) \\ 1-(1-\alpha)P(\hat{y}_i) & (y_i \neq \hat{y}_i) \end{cases} \quad (5)$$

식 (4)와 같이, 기존 손실 함수 식 (2)에 가중치 (w_R)를 곱하여 오답을 예측한 경우 예측 강도에 비례하여 기존보다 약하게 학습시킨다. 이 가중치는 식 (5)와 같이, 예측 강도($\hat{y}_i$)와 α 의 영향을 받는다. 예측 값이 정답인 경우 기존 손실 함수가 적용되고, 오답인 경우 예측의 강도(즉, 예측 확률)가 높을수록 손실 함수에 가중치를 낮게 부여하는 정규화를 수행한다. α 는 1보다 작은 값으로, 전체 데이터의 클래스 중복 정도에 따라 조절이 가능한 인자다. 이러한 정규화 과정을 통해 비록 잘못된 예측을 했을 지라도 그 강도가 강했다면 데이터가 서로 다른 클래스에 동시에 속할 수 있는 가능성을 고려하여 불이익을 적게 주는 효과를 부여한다.

4. 실험 방법 및 결과

4.1. 실험 데이터

실험 데이터는 자체적으로 제작한 가전 제품 관련 FAQ 모음이다. 질문 데이터 2734개를 학습 데이터 2563개, 평가 데이터 171개로 나누어 실험을 진행했다. 데이터 양이 적다는 것을 고려하여 검증 데이터를 따로 나누지 않고 실험을 진행했다. 클래스 수는 166개이다.

4.2. 실험 환경

성능 비교 연구를 위해 본 실험에서는 최근 가장 많이 사용되고 있는 BERT 모델을 사용했다. BERT 기본 모델, 기본 모델에 라벨 스무딩을 추가한 모델, 기본 모델에 본 연구에서 제안한 예측 강도 기반 손실 조정 정규화를 추가한 모델 총 3 가지로 실험을 진행했다. 훈련된 각 모델은 평가 데이터를 대상으로 평가했다.

각 모델의 학습 epoch을 10에서 30까지 5단위로 증가시키며 정확도를 측정했다. (epoch이 10보다 작은 경우 심한 과소적합이 발생했으며, 30보다 큰 경우 base의 성능이 감소하였다.) 실험의 객관성 확보를 위해 각 epoch에 대해 10번씩 학습을 수행하여 측정된 정확도 중 최대값과 최소값을 제외한 8개 값의 평균을 비교했다. 또한, 학습률(Learning rate)은 3e-5를 사용했다. 라벨 스무딩 인자(ε)는 0.1과 0.001, 0.0001 총 3가지를 사용하였으며, 0.1과 0.0001에서 높은 성능을 보였다. 예측 강도 기반 손실 조정 정규화 인자(α)는 0.8, 0.85, 0.9, 0.95를 사용했으며, 0.85와 0.9에서 높은 성능을 보였다. 4.3장에서는 높은 성능을 보인 인자를 토대로 실험 결과를 분석한다.

4.3. 실험 결과

표 2 는 평가 데이터 171 개를 기준으로 각 모델의 정확도 평균을 나타낸 것이다. 본 연구에서 제안한 예측 기반 손실 조정 정규화 모델은 모든 epoch 에 대해 기본 모델보다 성능이 개선된 것을 확인할 수 있었다. 또한, 정확도가 올라가는 양상도 기본 모델과 비슷함을 알 수 있었다.

예측 기반 손실 조정 정규화 모델과 라벨 스무딩 모델

표 2. epoch에 따른 각 모델의 정확도

epochs	base	예측 기반 손실 조정(α=0.85)	예측 기반 손실 조정(α=0.9)	라벨 스무딩(ε=0.1)	라벨 스무딩(ε =0.0001)
10	0.77266	0.772661	0.785088	**0.78655**	0.778509
15	0.779239	0.78655	0.782895	0.783626	0.788743
20	0.780701	**0.793129**	0.784357	0.783626	0.782164
25	**0.782894**	0.788743	0.785088	0.781433	**0.790205**
30	0.782163	0.787281	**0.790205**	0.782895	0.777047

모두 기본 모델보다 더 높은 정확도를 보였다. 0.1 을 적용한 라벨 스무딩은 정확도가 올라가는 양상이 기본 모델과 유사했지만 기본 모델과 정확도 차이가 크지 않았다. 그에 비해 0.0001 을 적용한 라벨 스무딩은 특정 epoch 에서 높은 정확도를 보였지만 특정 epoch 에서는 낮은 정확도를 보였다. epoch 변화에 따른 정확도 변동이 커 실제 학습 시 최적의 epoch 을 찾기 어렵다는 단점이 있었다. 예측 기반 손실 조정 정규화 모델은 라벨 스무딩과 비교했을 때 좀 더 안정적으로 학습되며 정확도도 높았다.

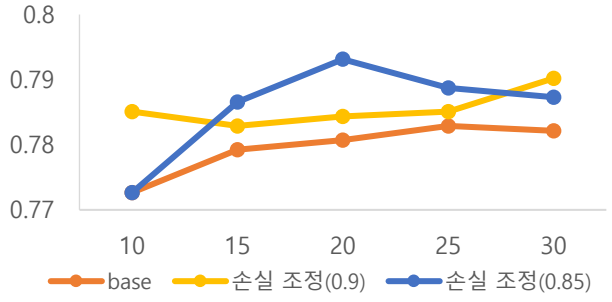

그림 2. 기본 모델과 예측 기반 손실 조정 모델 비교

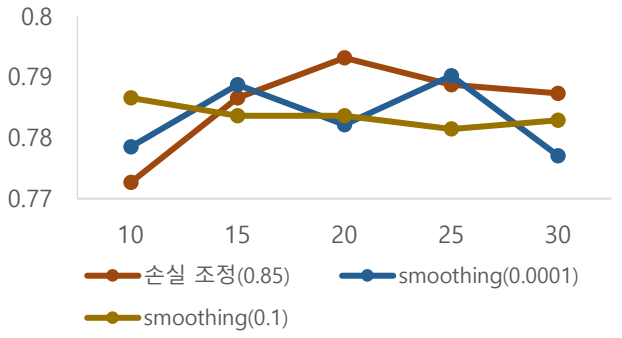

그림 3. 예측 기반 손실 조정과 라벨 스무딩 비교

예측 기반 손실 조정 정규화와 라벨 스무딩 모두 데이터 잡음에 강하다는 공통점이 있다. 하지만 라벨 스무딩은 학습 상태를 고려하지 않고 모든 데이터에 동일한 파라미터(ε)를 설정한다. 이러한 특징은 epoch이 변화함에 따라 학습의 안정성이 크게 달라지고 모델 성능에 대한 예측을 어렵게 할 수 있다. 이에 비해, 예측 기반 손실 조정 정규화는 epoch 값이 변화함에 따라 정확도 변화가 예측 가능함을 알 수 있다. 이는 학습 파라미터보다는 현재 모델의 학습 상태, 즉 예측 값에 따라 비용을 조절하기 때문이라고 해석했다. 그 결과, 라벨 스무딩에 비해 상황에 따른 최적값을 안정적으로 찾는 경향을 보였다.

5. 결론

본 논문에서는 하나의 클래스로 분류되어야 하는 데이터가 서로 다른 클래스에 동시에 분류될 수 있는 잡음 데이터 상에서 FAQ 분류 정확도가 낮아지는 문제를 해결하고자, 학습 과정에서 예측 값이 정답과 다른 경우 예측 강도에 비례하여 손실 값을 낮추는 정규화 방법을 제안하였다. 이 방법은 오답을 높은 확률로 예측한 경우에 데이터가 서로 다른 클래스에 동시에 속할 가능성이 크다고 가정하여 학습을 강하게 하지 않는 효과를 내는데, 이 과정을 통해 각 데이터의 신뢰도에 관계없이 비슷한 수준의 손실이 나오도록 강제하는 정규화 역할을 수행하게 된다.

또한, 본 방법이 성능에 미치는 영향을 알아보기 위하여 타 정규화 기법과 비교 분석 실험을 수행했다. 기본 모델, 라벨 스무딩이 적용된 모델과 비교하여 본 연구에서 제안한 방법이 성능도 높고 안정적으로 학습된다는 것을 확인했다. 이는 라벨 스무딩이 모든 데이터에 동일한 변화를 주는 것과 달리 본 방법은 현재 모델의 학습 상태(예측 값)에 따라 비용을 조절하기 때문이라고 분석됐다. 비교 실험 결과를 토대로 본 연구에서 제안한 방법이 데이터의 신뢰도가 낮은 상황에서 효과적으로 분류를 수행함을 알 수 있었다.

향후에는 FAQ 분류 문제 이외에 다른 태스크에도 본 논문에서 제안한 방법을 사용할 수 있을 지 추가 연구를 진행할 예정이다. 또한, 본 논문에서 적용한 정규화 인자 변화에 따른 성능 변화를 확인하고 나아가 이 인자 또한 학습과정에서 동적으로 결정될 수 있게 하는 방법을 추가로 연구할 예정이다.

참고문헌

[1] Y. Kim, Convolutional Neural Networks for Sentence Classification, In Proceedings of EMNLP, 2014.

[2] 김보성, 김주애, 이정엄, 김선아, 고영중, 서정연, Self-Attention 기반의 문장 임베딩을 이용한 효과

적인 문장 유사도 기법 기반의 FAQ 시스템, 한글 및 한국어 전보처리 학술대회 논문집, pp. 393-395, 2018.
[3] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding, In Proceedings of NAACL-HLT, pp. 4171-4186, 2019.
[4] A. Vaswani, N.Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, In Proceedings of NIPS, pp. 6000-6010
[5] R. Muller, S. Kornblith, G. Hinton, When Does Label Smoothing Help, 2019.
[6] C. X. Ling, V. S. Sheng., Cost sensitive learning, In Proceedings of Encyclopedia of machine learning, pp. 231-235, 2011.