능동 학습에서 불확실성 기반 선별방법과 다양성 기반 선별방법 간의 상관관계

Correlation Between Uncertainty-based and Diversity-based Querying Methods in Active Learning

고 민 환[1] · 허 윤 재[2] · 유 연 국[2] · 김 종 원[2] · 전 창 현[2] · 이 규 빈†
Minhwan Ko[1], Yunjae Heo[2], Yeonguk Yu[2], Jongwon Kim[2], Changhyun Jun[2], Kyoobin Lee†

Abstract: Deep learning is used in many fields, including robotics, but labeling is crucial for deep learning and is very costable. Active learning is proposed to address this issue. Following the entire unlabeled dataset, active learning queries the data should be labeled first to perform the task. There are two base methods to select data: uncertainty and diversity. These two methods are known to have a negative correlation, but both are should be considered because both of them are important for the performance of the model. In this paper, we show that uncertainty and diversity methods have not only a negative correlation as known before but also have a positive correlation between the two methods. As a result, if we only consider one, both of them can be considered.

Keywords: Active Learning, Deep Learning, Correlation, Querying Method, Uncertainty, Diversity

1. 서 론

딥러닝은 로봇 공학 분야에서 좋은 성능으로 다양하게 활용되고 있다. 하지만 신경 모델을 학습시키기 위해서는 데이터와 라벨이 필요하다. 로봇 공학 연구에서는 시뮬레이션을 활용한 많은 데이터 수집이 가능하지만, task에 따른 다양한 라벨링을 하는 것은 큰 비용이 드는 작업이다. 라벨링에 드는 비용 최소화를 위해 능동 학습은 제안되었다.

능동 학습의 전체적인 과정을 설명하면, 1. 라벨링이 되지 않은 데이터셋에서 라벨링을 받아 학습시켰을 때 모델의 성능을 가장 높일 수 있으리라 예측되는 데이터를 특정 개수만큼 선별한다. 2. 선별된 데이터들은 오라클에 라벨링을 받고 3. 모델을 학습한다. 이후, 남은 라벨링이 되지 않은 데이터셋 중에서 학습된 모델을 이용하여 선별, 라벨링, 학습을 반복한다. 이때, 반복되는 과정을 주기(cycle)라고 부른다. 최종적인 능동 학습의 목적은 최소한의 라벨링으로 최대한의 성능을 내는 것이다.

능동 학습에서 선별방법은 크게 두 가지로 나뉘게 된다. 불확실성 기반과 다양성 기반 방법론이다. 불확실성 기반은 현재 모델을 기준으로 데이터의 불확실성이 가장 높은 데이터를 선별하는 방법이다. 다양성 기반은 선별될 데이터들이 전체 데이터셋을 고루 대표할 수 있도록 선별하는 방식이다. 우리는 두 방법론 간의 상관관계를 비교하기 위해서 대표적인 불확실성 기반 방법론인 Learning Loss [1]와, 다양성 기반 방법론인 Coreset [2]을 선택하였다. 두 방법론은 서로 음의 상관관계에 있다.[2] 즉, 불확실성 방법론으로 데이터를 선별하면, 다양성을 고려할 수 없고, 역도 마찬가지다. 하지만 불확실성과 다양성 모두 능동 학습의 성능에 큰 영향을 미치기 때문에 최근 연구에서는 둘을 모두 고려하는 방식으로 연구가 진행되고 있다. 이 연구에서는 불확실성 방법론과 다양성 방법론으로 선별되는 데이터들을 주기마다 비교하면서, 두 방법론의 상관관계가 주기가 지남에 따라 변화하고 결과적으로 양의 상관관계도 존재함을 보인다.

※ 이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2022-0-00951, 스스로 불확실성을 자각하며 질문하면서 성장하는 에이전트 기술 개발)
1. Undergraduate Student, GIST, Gwangju, Korea (mhko1998@gm.gist.ac.kr),
2. Principal Researcher, GIST, Gwangju, Korea (boss9911@gm.gist.ac.kr, yeon_guk@gm.gist.ac.kr, jongwonkim@gm.gist.ac.kr, junch@gm.gist.ac.kr)
† Associate Professor, Corresponding author: School of integrated technology, GIST. Gwangju Korea (kyoobinlee@gist.ac.kr)

2. 본 론

2.1 실험 환경

실험에 사용한 데이터셋은 CIFAR10이다. CIFAR10은 총 10개의 class로 이뤄져 있으며, 인공지능 학습 방법론 연구에서 꾸준히 사용되고 있다. 학습 모델로는 Resnet18을 사용하였다.

2.2 실험 방법

서론에서 설명한 전체적인 능동 학습을 진행하면서 '[Fig. 1]'과 같이 과정 1.을 불확실성 방법과 다양성 방법으로 두 번 진행하면서 중복데이터 개수를 측정한다. 이후 학습에 사용되는 데이터는 불확실성 방법으로 선별된 데이터로 모델을 학습하였다. 총 5번의 실험을 반복하여 평균값을 결과로 사용하였다. 각 실험은 총 10주기까지 진행하고, 주기마다 1,000개(batch)의 데이터를 선별한다. 주기마다 모델은 초기 상태에서 학습된다. 또한, 선별 과정에서 전체 데이터셋의 임의 부분 집합 10,000개에 대해서만 선별 과정을 진행하였다. 이와 같은 방식을 통해, 선별 과정에서의 계산량을 줄이고, 불확실성 기반 방법론의 경우, 다양성을 확보할 수 있다[1].

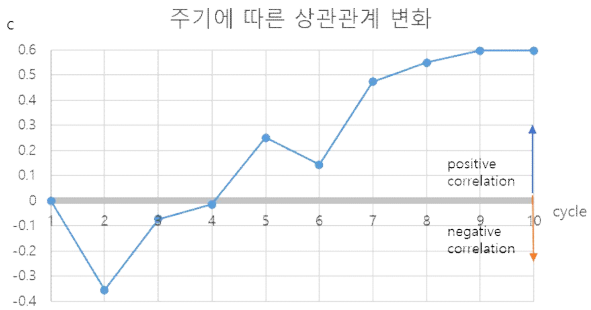
Algorithm 1 Count Overlapped Data Algorithm

Input: subset S

F_u = Uncertainty Querying Strategy,
F_d = Diversity Querying Strategy
$U_u = F_u(\mathrm{S})$, $U_d = F_d(\mathrm{S})$
$\mathrm{N} = num(U_u \cap \mathrm{U}_d)$
return N

[Fig. 1] 두 방법론 간 겹치는 데이터 개수 판별 알고리즘

3. 실험 결과



[Fig. 2] 주기에 따른 상관관계(C) 변화

$$P(U_u \cap U_d) = P(U_u) * P(U_d) \quad (1)$$

$$E(U_u \cap U_d) = P(U_u) * P(U_d) * 10{,}000 = 100 \quad (2)$$

$$\alpha_1 = \frac{batch}{E} - 1, \alpha_2 = \frac{batch - 2E}{E^2} \quad (3)$$

$$C = \log_{\alpha_1}(\alpha_2 x + 1) - 1$$

무상관관계에서는, 두 선별방법은 서로 독립이므로, (1)이 성립한다. 무상관관계에서 중복데이터 개수의 기댓값은 (2)에 의해 100개가 된다. 우리는 상관관계 분석을 위해 (3)에서 C를 제안한다. C는 -1에서 1 사이의 값을 가지고, 중복데이터가 0개일 경우 -1, 무상관 관계에서의 값일 경우 0, 데이터가 전부 중복될 경우 1이 된다. C값을 통해 0 미만의 값은 음의 상관관계, 0 초과의 값은 양의 상관관계로 구분하였다.

첫 주기에서는 학습된 모델이 없어 임의로 데이터를 선별하기에, 무상관관계로 C=0이다. 두 번째 주기에서는 C= -0.355로 음의 상관관계를 가진다. 하지만 주기가 지날수록 양의 상관관계로 전환된다. 최종 주기에서는 C=0.596로, 중복데이터 개수 404개로 약 40%에 해당하는 수치이다. 즉, 불확실성으로 선별된 데이터로 학습을 진행하였을 때, 최종 주기에서 C=0.596로 이는 현재 진행되고 있는 불확실성 선별방법에서 다양성이 고려되고 있음을 의미한다. 즉, 다양성을 전혀 고려하지 않은 알고리즘으로 선별해도 주기가 진행될수록 다양성이 고려된 데이터가 선별됨을 확인하였다.

4. 결 론

본 연구에서는 능동 학습에서의 불확실성 방법론과 다양성 방법론 사이의 상관관계가 주기에 따라 변화하고, 최종적으로 양의 상관관계를 가지고 있음을 보였다. [2]에서는 두 방법론이 음의 상관관계가 있다고 하였고, 실험 결과 초기의 주기에서는 음의 상관관계가 있었다. 하지만 주기가 진행될수록 양의 상관관계로 전환되었다. 데이터의 총 개수와 class 수마다 전환되는 시기나 정도는 달라질 것으로 예상한다. 하지만 학습에 사용되는 데이터가 늘어날수록 두 방법론은 양의 상관관계로 전환된다. 학습 초기에는 특정 데이터 분포에서 불확실성이 높은 데이터가 몰려있는 경우가 많지만, 주기가 진행될수록 불확실성이 높은 데이터가 모델이 아직 학습하지 않은 데이터 분포가 될 가능성이 커지기 때문이다. 따라서 불확실성 방법론이 주기가 지나면서 다양성도 고려하게 되었다. 이 연구가 후에 진행될 새로운 연구에서 제안하는 알고리즘이 데이터셋의 다양성과 불확실성을 얼마나 내포하는지 상관관계를 보일 때나, 두 가지 방법론에 기반한 능동 학습의 연구에 도움이 될 수 있기를 기대한다.

References

[1] Yoo, Donggeun, and In So Kweon. "Learning loss for active learning." CVPR, 2019

[2] Sener, Ozan, and Silvio Savarese. "Active Learning for convolutional neural networks: A core-set approach." ICLR, 2018