평탄화를 통한 편향 제거

한형근, 이정우
서울대학교
hygnhan@snu.ac.kr, junglee@snu.ac.kr

Debiasing using flatness

Hyeonggeun Han, Jungwoo Lee
Seoul National Univ.

요 약

손실 함수의 평탄화(flatness)는 모델의 generalization과 관련이 있다는 연구 결과들이 있으며 이 평탄화를 이용해 generalization 능력을 향상시키고자 하는 연구들이 활발히 진행되고 있다. 본 논문에서는 평탄화를 이용해 편향 제거(debiasing) 성능을 높이는 새로운 방법을 제시한다. 실험을 통해 기존 baseline 모델의 debias 성능을 향상시킴을 보인다.

Ⅰ. 서 론

인공 신경망의 성능은 학습에 사용되는 dataset에 큰 영향을 받는다. dataset을 구성하는 data에 target label과 강한 상관관계를 갖는 attributes가 존재한다면 모델은 data에 존재하는 attributes에 집중하여 학습할 것이다. 이러한 학습 과정이 문제가 되는 이유는 이러한 attributes가 학습하기 쉬운 attributes이며 강한 상관관계를 갖는 줄 알았던 target label과의 관계가 모든 data에 대해 만족하는 상관관계가 아니기 때문이다. 이렇게 학습된 모델은 해당 attributes를 갖지 않는 동일한 target label을 가진 data에 대해 잘못된 예측을 출력할 것이고 이는 성능 하락으로 이어진다. 모델 학습 시에 모델이 얻게 되는 이러한 편견(bias)를 제거하고자 하는 debiasing 방법들이 최근 많이 연구되고 있다. 이러한 방법에는 크게 resampling [1], SSL을 이용한 더 좋은 representation 학습 [2], last-layer 재학습 [3] 방법이 있다.

손실 함수의 평탄화(flatness)는 모델 generalization과 관련이 있다는 연구들이 있으며 이를 이용해 generalization 능력을 향상시키고자 하는 연구들이 활발히 진행되고 있다. 대표적으로 각 data samples의 gradient of loss를 손실 함수에 직접 추가하여 최소화함으로써 손실 함수의 gradient를 작게 만들어 평탄화를 진행하는 Explicit gradient regularization [4], optimization trajectory를 따라 $Tr[H] \approx Tr[F]$ 가 만족함을 이용하여 $Tr[F]$를 최소화함으로 평탄화를 진행하는 방법 [5]이 있다.

본 논문에서는 평탄화를 이용하여 debiasing을 진행하는 새로운 방법을 소개한다. 각 data sample마다 flatness를 측정하여 각 sample이 generalization에 미치는 영향을 계산한다. 그리고 이 영향을 손실 함수에 추가하여 함께 최소화함으로써 debiasing을 진행한다.

Ⅱ. 본론

본 논문에서는 평탄화와 generalization의 관계를 이용한 새로운 debiasing 방법을 제안한다. 각 sample 별 Fisher information matrix의 trace를 구하고 가장 큰 trace를 갖는 sample에 대해 최소화하는 것으로 debiasing을 진행한다. training distribution $p(x)$, model distribution $p(y|x;\theta)$에 대해 Fisher information matrix는 다음과 같이 나타낼 수 있다.

$$F(\theta) = E_{p(x)}[E_{p(y|x;\theta)}[\nabla_\theta log p(y|x;\theta) \nabla_\theta \log p(y|x;\theta)^T]]$$
$$= \frac{1}{N}\sum_{i=1}^{N} E_{p(y|x_i;\theta)}[\nabla_\theta \log p(y|x_i;\theta) \nabla_\theta \log p(y|x_i;\theta)^T]$$

이에 따라 trace of FIM (Fisher Information matrix)는 다음과 같이 나타낼 수 있다.

$$Tr[F(\theta)] = \frac{1}{N}\sum_{i=1}^{N} E_{p(y|x_i;\theta)}[\|\nabla_\theta \log p(y|x_i;\theta)\|^2]$$

여기서 i는 학습 데이터 sample에 대한 index를, N은 학습 데이터 개수를 나타낸다. 위 식으로부터 sample x_i 에 대한 trace of FIM은 $Tr[F_i(\theta)] = E_{p(y|x_i;\theta)}[\|\nabla_\theta \log p(y|x_i;\theta)\|^2]$ 으로 나타낼 수 있다. flatness와 generalization 사이의 관계에서 gradient of loss가 크면 generalization 성능이 떨어지고 gradient of loss가 작으면 generalization 성능이 올라가는 것을 여러 연구들에서 볼 수 있었다. 이러한 관측으로부터 알 수 있는 것은 각 sample x_i 의 flatness $Tr[F_i(\theta)]$ 값은 서로 다르고 큰 값을 나타내는 samples는 generalization 성능을 저해한다는 것이다. 모델 학습 시에 bias를 학습하는 것을 고려했을 때 이렇게 큰 flatness 값을 나타내는 특정 samples 에 대해 모델은 올바르게 학습하지 않았다는 것을 뜻하고 이에 대한 학습을 올바르게 진행하여 모델을 debiasing할 수 있다는 것을 생각할 수 있다. 이를 위해 아래의 최적화 문제를 제안한다.

$\min_\theta L(\theta)\ s.t.\ Tr[F_i(\theta)] \leq \epsilon\ \forall\ i \in \{1, ..., N\}$

학습 데이터 중에 flatness가 큰 sample x_i에 대해 $Tr[F_i(\theta)]$ 를 최소화하여 debiasing 성능을 개선할 수 있으며 이를 위해 위와 같은 최적화 문제를 제안한다. 위 문제를 Lagrangian의 형태로 바꾸고 primal-dual problem을 푸는 방법으로 debiasing 문제를 해결한다.

제안한 방법이 유의미한 debiasing 방법으로 모델의 성능 향상을 가져오는지 확인하기 위해 ResNet-50 모델과 Waterbirds 데이터를 이용하여 accuracy 및 worst group accuracy (WGA)를 측정했다. 실험결과는 아래 표에 나타나 있으며 기존 debiasing 방법들과 마찬가지로 validation set을 이용해 earl stopping을 사용하였다. 기존 debiasing 방법 중 널리 사용되는 DRO 방법과 비교했을 때 약간 성능이 낮지만, ERM과 비교해 큰 폭으로 성능이 향상했음을 볼 수 있다.

	Avg	Worst
ERM	97.3	60.0
DRO	93.2	86.0
Ours	91.94	84.58

Table 1. ERM 및 debiasing 방법 성능 비교

Ⅲ. 결론

본 논문에서는 flatness와 generalization 사이의 관계를 이용하여 새로운 형태의 debiasing 방법을 제안하였다. 또한 실험을 통해 WGA 측면에서 성능향상을 이끌어냄을 보였다.

ACKNOWLEDGMENT

This work is in part supported by National Research Foundation of Korea (NRF, 2021R1A2C2014504(50%)), National R&D Program through the National Research Foundation of Korea(NRF) funded by Ministry of Science and ICT(2021M3F3A2A02037893)(50%), INMAC, and BK21 FOUR program.

참 고 문 헌

[1] Ahn, Sumyeong, Seongyoon Kim, and Se-young Yun. "Mitigating Dataset Bias by Using Per-sample Gradient." arXiv preprint arXiv:2205.15704 (2022).

[2] Wang, Ke, et al. "CLAD: A Contrastive Learning based Approach for Background Debiasing." arXiv preprint arXiv:2210.02748 (2022).

[3] Kirichenko, Polina, Pavel Izmailov, and Andrew Gordon Wilson. "Last layer re-training is sufficient for robustness to spurious correlations." arXiv preprint arXiv:2204.02937 (2022).

[4] Barrett, David GT, and Benoit Dherin. "Implicit gradient regularization." arXiv preprint arXiv:2009.11162 (2020).

[5] Jastrzebski, Stanislaw, et al. "Catastrophic fisher explosion: Early phase fisher matrix impacts generalization." International Conference on Machine Learning. PMLR, 2021.