피셔 정보량 기반의 적응형 정규화를 통한 증거 기반 딥러닝의 불확실성 추정 성능 향상

김용호, 박현희*
명지대학교

yhkim98@mju.ac.kr, *hhpark@mju.ac.kr

Improving Uncertainty Estimation in Evidential Deep Learning via Fisher Information based Adaptive Regularization

Kim Yong Ho, Park Hyun Hee*
Myongji Univ.

요 약

본 논문에서는 입력 데이터의 정보량에 따라 정규화 강도를 적응적으로 조절하는 정보 인지형 증거 기반 딥러닝 프레임워크인 Info-EDL 을 제안한다. Info-EDL 은 예측된 디리클레 분포의 Fisher Information 을 이용하여 정규화 계수를 데이터 의존적 함수로 확장함으로써 정보가 부족한 입력에는 보수적인 불확실성 추정을 유지하고 정보량이 높은 샘플에 대해서는 표현 학습을 촉진한다. CIFAR-10 을 In-distribution 으로, SVHN 과 CIFAR-100 을 Out-of-Distribution 으로 구성한 실험 결과, Info-EDL 은 ID 분류 정확도 91.50%를 유지하면서 SVHN 과 CIFAR-100 에 대해 각각 93.91% 및 87.35%의 AUROC 를 달성하여 기존 EDL 대비 우수한 OOD 탐지 성능을 달성한다.

Ⅰ. 서 론

딥러닝 기술의 발전으로 다양한 고성능 모델들이 제안되었으나, 이러한 모델들은 실제 환경에서 발생하는 예측 불확실성을 충분히 고려하지 못한다는 한계를 가진다. 이로 인해 벤치마크 상에서 우수한 성능을 보이더라도, 실제 응용 환경에서는 신뢰성 있는 활용이 어렵다. 불확실성 추정을 위해 앙상블[1] 기법이나 베이지안 신경망[2]이 제안되었으나, 높은 계산 비용으로 인해 실시간 환경에 적용하는 데에는 제약이 따른다. 이러한 배경에서, 증거 기반 딥러닝(Evidential Deep Learning, EDL)[3]은 단일 신경망 구조를 유지하면서 예측 확률과 불확실성을 동시에 추정할 수 있는 효율적인 대안으로 제안되었다.

그러나 기존 EDL 은 정규화 계수 λ를 고정하거나 학습 에폭에 따라 스케줄링하는 방식에 의존하고 있어, 입력 데이터의 특성이나 정보량을 충분히 반영하지 못한다는 한계를 가진다. 본 연구에서는 이러한 문제를 해결하기 위해, 예측된 디리클레 분포의 Fisher Information 을 활용하여 정규화 강도를 데이터마다 적응적으로 조절하는 정보 인지형 Evidential Deep Learning 프레임워크인 Info-EDL 을 제안한다. Info-EDL 은 입력 데이터의 정보량에 따라 Out-of-Distribution(OOD) 샘플과 분류 경계 영역의 어려운 샘플을 구분하여 학습함으로써, 기존 EDL 의 고정적 정규화 방식이 가지는 한계를 효과적으로 완화한다.

Ⅱ. 본 론

Evidential Deep Learning 은 K 개의 클래스에 대한 비음수 증거 $e \geq 0$ 를 통해 디리클레 분포의 파라미터 $\alpha = e + 1$ 을 생성함으로써, 클래스 확률과 불확실성을 동시에 모델링한다. EDL 의 전체 손실 함수는 식 1 과 같이 정의된다.

$$\mathcal{L}(\theta) = \mathcal{L}_{Risk}(\alpha) + \lambda \cdot \mathcal{L}_{KL}(\alpha) \quad (1)$$

여기서 $\mathcal{L}_{KL}$ 은 예측된 디리클레 분포가 충분한 증거를 확보하지 못한 경우 유니폼 디리클레 분포로 수축하도록 유도하는 정규화 항이다. 기존 EDL 방식에서는 정규화 계수 λ를 고정된 상수로 설정하거나, 학습 에폭에 따라 점진적으로 증가시키는 어닐링(annealing) 기법을 주로 사용한다. 이는 모든 샘플에 동일한 정규화 강도를 적용하기 때문에, 입력 데이터의 난이도나 정보량 차이를 충분히 반영하지 못한다. 일부 경우에는 모델의 표현 학습을 제한할 수 있다.

본 연구에서는 이러한 한계를 극복하기 위해 예측 분포가 내포하는 정보량을 정량화하는 정보 인지형 Evidential Deep Learning 프레임워크인 Info-EDL 을 제안한다. Info-EDL 은 모델이 입력 데이터에 대해 형성한 예측 분포의 정보량을 Fisher Information 을 통해 평가하고, 이를 기반으로 정규화 강도를 데이터마다 동적으로 조절한다. EDL 환경에서 디리클레 분포의 파라미터 α에 대한 Fisher Information Matrix(FIM)의 대각합은 식 2 와 같이 근사할 수 있다.

$$v_{info} = \mathrm{Trace}(\mathrm{I}(\alpha)) \approx \sum_{k=1}^{K} \left(\psi'(\alpha_k) - \psi' \left(\sum_{j=1}^{K} \alpha_j \right) \right) \quad (2)$$

여기서 ψ'는 Trigamma 함수이다. v_{info}는 예측 분포가 특정 클래스에 얼마나 강하게 수렴하는지를 정량화한다. 이는 단순한 증거의 총합과는 달리 비선형적인 정보의 효용을 나타낸다. Info-EDL 에서는 이러한 정보량 지표를 활용하여 정규화 계수 λ를 데이터별로 조절하는 정보 기반 게이팅 함수를 제안한다. 기존 EDL 이 고정된 정규화 계수 λ를 사용하는 반면 Info-EDL 은 입력 데이터의 정보량에 따라 변화하는 함수 $\lambda(v_{info})$ 로 확장한다. 제안하는 정규화 함수는 식 3 과 같이 정의되며 β와 γ 두 개의 하이퍼파라미터로 구성한다.

$$\lambda(v_{info}) = \beta \cdot \exp(-\gamma \cdot v_{\mathrm{info}}) \quad (3)$$

표 1. CIFAR-10(ID) 분류 정확도와 SVHN(Far-OOD), CIFAR-100(Near-OOD)에 대한 OOD 탐지 성능 비교 결과

방법론	ID Accuracy (↑)	SVHN AUROC (↑)	CIFAR-100 AUROC (↑)	SVHN FPR@95 (↓)	CIFAR-100 FPR@95 (↓)
EDL ($\lambda = 1.0$)	92.05%±0.22	40.72%±10.18	57.60%±4.93	99.77%±0.21	99.19%±0.20
EDL ($\lambda = 0.1$)	91.53%±0.28	82.63%±6.67	75.27%±1.82	66.53%±15.66	89.90%±4.48
Info-EDL (Ours)	91.50%±0.18	93.91%±1.43	87.35%±0.13	18.76%±3.32	41.90%±1.06

여기서 β는 정보량이 낮은 경우 적용되는 정규화 강도의 상한선으로, OOD 데이터에서 과도한 증거 형성을 억제하는 역할을 수행한다. γ 는 정보량 증가에 따라 정규화 강도가 완화되는 민감도를 제어한다.

이러한 정규화 메커니즘을 통해 Info-EDL 은 입력 데이터의 정보량에 따라 적응적인 학습 전략을 유도한다. 정보량이 낮은 경우 $\lambda(v_{info})$는 β에 근접하여 $\mathcal{L}_{KL}$ 항의 영향이 커지고 출력 분포는 유니폼 디리클레 분포로 수축되어 높은 불확실성이 유지된다. 반대로 정보량이 높은 경우 $\lambda(v_{info})$는 감소하여 정규화 효과가 완화되고 손실 함수는 $\mathcal{L}_{Risk}$ 항에 집중되어 미세한 특징 차이를 학습할 수 있다. 결과적으로 Info-EDL 은 명백한 오답에 대해서는 보수적인 불확실성 추정을 유지하는 동시에 경계 영역에 위치한 모호한 샘플에 대해서는 표현 학습을 촉진함으로써 기존 EDL 의 고정적 정규화 방식이 가지는 한계를 효과적으로 완화한다.

Ⅲ. 실험 및 결과

본 연구는 제안하는 Info-EDL 의 성능을 검증하기 위해 In-distribution(ID) 데이터셋으로 CIFAR-10[4]을 사용하고, OOD 데이터셋으로 이미지 도메인이 상이한 SVHN[5]과 도메인이 유사하여 구분이 상대적으로 어려운 CIFAR-100[4]을 사용한다. 백본 네트워크로는 ResNet-18[6] 모델을 사용한다. 정규화 계수에 따른 성능 변화를 분석하기 위해 기존 EDL 은 정규화 계수 λ 를 각각 1.0 과 0.1 로 고정한 두 가지 설정을 사용한다. 제안하는 Info-EDL 은 하이퍼파라미터 튜닝에 따른 성능 변동을 배제하고 기법 자체의 효과를 검증하기 위해 정규화 관련 하이퍼파라미터인 β와 γ를 모두 1.0 으로 고정하여 사용한다. 결과의 신뢰성을 확보하기 위해 5 개의 무작위 시드에 대해 실험을 반복 수행한 후 평균과 표준편차를 비교한다.

모델의 성능 평가는 ID 데이터에 대한 분류 정확도와 OOD 데이터에 대한 탐지 성능을 기준으로 수행한다. 분류 정확도는 ID 데이터에서의 예측 성능을 평가하기 위한 지표이며 OOD 탐지 성능은 AUROC 와 FPR@95 를 사용하여 측정한다. AUROC 는 ID 와 OOD 샘플을 구분하는 전반적인 분리 성능을 나타내며 FPR@95 는 ID 데이터의 재현율이 95%일 때 OOD 샘플이 ID 로 오분류되는 비율을 의미한다.

표 1 은 각 방법론의 ID 분류 정확도와 OOD 탐지 성능을 비교한 결과이다. 기존 EDL 은 정규화 강도에 따라 분류 정확도와 OOD 탐지 성능 사이의 trade-off 가 명확하게 나타난다. 강한 규제를 적용한 EDL($\lambda = 1.0$)은 ID 분류 정확도에서 92.05%로 가장 높은 성능을 보이나 SVHN 과 CIFAR-100 에 대한 AUROC 가 각각 40.72%, 57.60%로 OOD 탐지 성능은 크게 저하된다. 이는 과도한 정규화로 인해 OOD 샘플에 대해서도 높은 확신을 부여하는 과신 현상이 발생한 것으로 보인다. 반대로 규제를 완화한 EDL($\lambda = 0.1$)은 ID 분류 정확도를 비교적 잘 유지하면서 SVHN 과 CIFAR-100 에 대한 OOD 탐지 성능을 일부 개선한다. 그러나 AUROC 및 FPR@95 지표에서 여전히 Info-EDL 에 비해 낮은 성능을 보이며 고정된 정규화 계수를 사용하는 기존 방식이 분류 정확도와 OOD 탐지 성능을 동시에 최적화하는데 구조적인 한계를 가짐을 보여준다.

Info-EDL 은 91.50%의 높은 ID 분류 정확도를 유지하면서도 OOD 탐지 성능을 동시에 향상시킨다. SVHN 과 CIFAR-100 에 대해 각각 93.91%, 87.35%의 AUROC 를 달성하며 FPR@95 또한 두 데이터셋 모두에서 가장 낮은 값을 보인다. 특히 Near-OOD 인 CIFAR-100 에서의 성능 향상은 정보량 기반의 적응적 정규화가 미세한 분포 차이를 효과적으로 반영함을 의미한다. 더불어 낮은 표준편차는 제안 방법의 학습 안정성을 뒷받침한다.

Ⅳ. 결 론

본 연구에서는 기존 EDL 이 고정된 정규화 계수 λ에 의존함으로써 입력 데이터의 정보량 차이를 충분히 반영하지 못하고 분류 정확도와 불확실성 추정 성능 사이의 trade-off 에 직면한다는 한계를 지적하였다. 이를 해결하기 위해 예측된 디리클레 분포의 Fisher Information 을 활용하여 정규화 강도를 데이터별로 적응적으로 조절하는 정보 인지형 프레임워크인 Info-EDL 을 제안하였다. 실험 결과, Info-EDL 은 높은 ID 분류 정확도를 유지하면서 Far-OOD 및 Near-OOD 환경 모두에서 우수한 탐지 성능을 달성하여, 기존 EDL 의 고정적 정규화 방식이 가지는 한계를 효과적으로 완화함을 확인하였다.

ACKNOWLEDGMENT

이 논문은 산업통상자원부 및 산업기술기획평가원 (KEIT)과 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2022R1A2C2005705, 분산 머신러닝 기반 지능형 플라잉 기지국을 위한 AI-MAC 프로토콜, No. RS-2024-00469138, 국산 AI 반도체 기반 비전인식기술 소프트웨어개발도구 개발)

참 고 문 헌

[1] Lakshminarayanan, B., Pritzel, A., and Blundell, C. “Simple and scalable predictive uncertainty estimation using deep ensembles,” Advances in Neural Information Processing Systems, vol. 30, 2017.

[2] MacKay, D. J. C. “A practical Bayesian framework for backpropagation networks,” Neural Computation, vol. 4, no. 3, pp. 448–472, 1992.

[3] Sensoy, M., Kaplan, L. M., and Kandemir, M. “Evidential deep learning to quantify classification uncertainty,” Advances in Neural Information Processing Systems, vol. 31, pp. 3183–3193, 2018.

[4] Krizhevsky, A. and Hinton, G. "Learning multiple layers of features from tiny images," Technical Report, University of Toronto, 2009.

[5] Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. "Reading digits in natural images with unsupervised feature learning," NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2011.

[6] He, K., Zhang, X., Ren, S., and Sun, J. "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.