레이블 노이즈에 강건한 머신 러닝 및 실 세계 응용

NAVER AI Lab ▌송환준

1. 서 론

가용 데이터의 폭발적인 증가에 따라 데이터 품질이 인공지능 알고리즘의 성능을 결정 짓는 데이터 중심 (Data-centric) AI가 인공지능의 주류가 되고 있다 [1]. 따라서, 인공지능 모델의 학습과 연계해 데이터 품질을 향상시키는 것은 많은 관심을 받고 있다. 딥 러닝을 활용한 지도 학습 (Supervised Learning)의 경우, 데이터 수집과 관련해서 발생하는 대표적인 데이터 품질 문제는 레이블 노이즈 문제이다. 일반적으로 데이터 수집 시 레이블링 비용을 줄이기 위해 Amazon Mechanical Turk와 같은 크라우드 소싱 (Crowd-sourcing) 플랫폼들이 널리 활용되지만, 이러한 수집 방법들은 정확한 레이블링의 어려움으로 인해 8.0-38.5%의 노이즈를 발생시키는 것으로 알려져 있다 [2].

딥 러닝 (Deep Learning) 모델의 경우, 데이터와 레이블을 완전히 맞출 수 있는 높은 표현력을 가지고 있기 때문에, 잘못된 레이블로 인한 과적합 (Overfitting) 문제가 발생하며 이는 학습 모델의 테스트 데이터 (Test Data)에 대한 일반화성능을 크게 저하시킨다. 비록 과적합 문제를 해결하기 위한 배치 정규화 (Batch Normalization) 및 드롭아웃 (Dropout)과 같은 다양한 범용적 방법들이 활용될 수 있지만, 여전히 큰 폭의 성능저하를 피할 수 없다. 그림 1과 같이, 앞서 언급한 정규화 기술들이 활성화되더라도 깨끗한 데이터와 레이블 노이즈가 있는 데이터에 대해 훈련된 모델 간의 테스트 정확도 차이는 여전히 상당하다. 따라서, 레이블 노이즈에 강건한 딥 러닝 학습 방법을 고안하는 것은 모델의 일반화 성능을 향상하기위한 핵심 도전 과제이다.

레이블 노이즈에 대한 과적합 문제를 해결하기 위해 다양한 측면에서의 연구가 머신 러닝 커뮤니티를 중심으로 활발히 수행되고 있다. 큰 관점에서, 강건한 모델 아키텍처 그리고 강건한 로스 함수를 새롭게 디

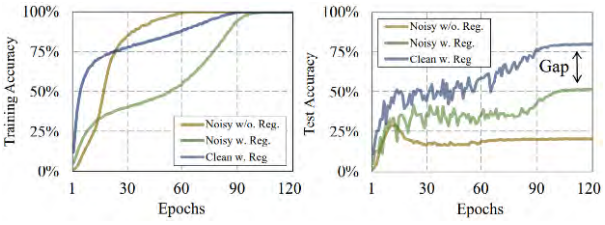

그림 1 학습 및 테스트 데이터에 대한 정확도 수렴 곡선 [6]: "Noisy w/o. Reg."와 "Noisy w. Reg."는 노이지 데이터 (Noisy Data)에 대해 정규화 기법을 사용하지 않은 및 사용한 훈련 모델, 그리고 "Clean w. Reg."는 깨끗한 데이터 (Clean Data)에 대해 정규화 기법을 사용하여 훈련된 모델이다.

자인할 수 있으며 [3], 이와 달리 학습 데이터를 정제하거나 새로운 정규화 방법론을 개발할 수 있다 [4]. 추가적으로 과거 연구들은 데이터의 접근이나 데이터를 저장하는 스토리지의 크기에 대한 제한이 없는 일반적인 지도학습 환경에서의 연구를 위주로 진행하였으나, 최근에는 보다 현실적인 요소를 고려하려는 시도들이 있어왔다. 예로, 실 세계 응용을 위해서 주어진 가용 스토리지 내에서 실시간으로 입력되는 데이터에 대한 효과적인 학습을 보장하는 연속학습 (Continual Learning), 그리고 개인화된 장치들이 개인정보 보호를 준수하며 협력하여 모델학습을 진행하는 연합학습 (Federated Learning) 등과 연계하여 레이블 노이즈 문제를 해결하기 위한 연구들이 있다 [5, 6].

본 기고에서는 레이블 노이즈를 해결하기 위한 연구들의 동향에 대해 살펴보고자 한다. 2장에서는 레이블 노이즈에 대한 기본적인 개념과 공개적으로 활용가능한 실 세계 노이즈 데이터에 대해 소개하며, 3장에서는 표준 지도학습 하에서의 레이블 노이즈를 해결하기 위한 기법, 4장에서는 연속학습 하에서 레이블 노이즈를 해결하기 위한 기법, 5장에서는 연합학습 하에서 레이블 노이즈를 해결하기 위한 기법에 대해 소개한다. 마지막으로 6장에서는 향후 연구가 필요한 방향을 제시하고 결론을 맺는다.

(a) Animal-10N (b) Food-101N (c) Clothing1M (d) WebVision

그림 2 활발히 사용되는 네 가지 실 세계 노이지 데이터.

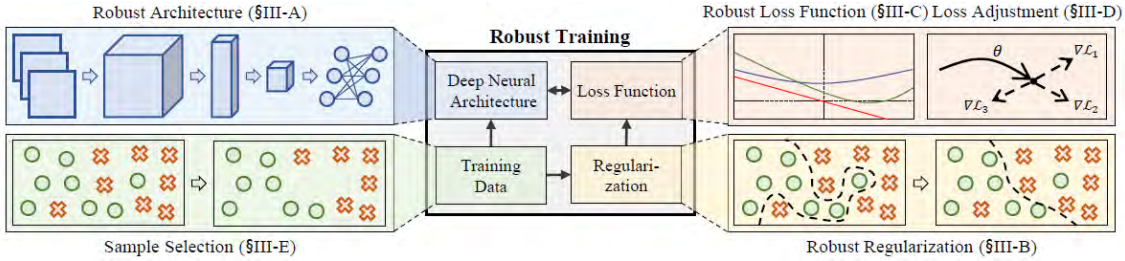

그림 3 레이블 노이즈를 극복하기 위한 최신 강건한 방법론들의 분류 [2].

2. 배경 지식

레이블 노이즈가 있는 지도 학습에 대한 문제를 노이즈 타입에 대한 분류와 함께 정의하고, 그 후 널리 사용되는 실 세계 노이즈 데이터에 대해 요약한다.

2.1. 레이블 노이즈에 대한 지도 학습

분류 (Classification)는 인풋 피쳐 (Input Feature) x 를 주어진 레이블 (Label) y 에 매핑하는 함수 f 를 학습하는 대표적인 지도 학습 문제이다. c-class 분류 문제의 목표는 학습 데이터 $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, 로스 함수 l이 주어질 때, 다음 목표 함수를 최소화하는 모델 Θ 를 학습하는 것이다.

$$\mathcal{R}_{\mathcal{D}}(f) = \mathbb{E}_{\mathcal{D}}[\ell(f(x;\Theta), y] = \frac{1}{|\mathcal{D}|}\sum_{(x,y)\in\mathcal{D}} \ell(f(x;\Theta), y)$$

실제 레이블 노이즈 시나리오에서, 학습 데이터는 정답 레이블 (True Label)이 아닐 수 있는 노이지 레이블 (Noisy Label) $\tilde{y}$을 갖는 노이지 데이터 $\mathcal{D} = \{(x_i, \tilde{y}_i)\}_{i=1}^{N}$이기 때문에 아래 노이지 로스를 활용한 경사 하강법을 통해 모델 가 학습된다.

$$\Theta_{t+1} = \Theta_t - \eta\nabla\left(\frac{1}{|B_t|}\sum_{(x,\tilde{y})\in\tilde{\mathcal{D}}} \ell(f(x;\Theta_t), \tilde{y})\right),$$

η는 학습 계수 그리고 B 는 t 시점에서의 미니배치. 노이즈 레이블에 대한 경사 하강법이 적용되므로 최적화 프로세스는 더 이상 노이즈에 강건하지 않으며, 과적합은 모델의 일반화 성능을 떨어뜨리게 된다.

2.2. 레이블 노이즈 분류

본질적으로 데이터 레이블의 손상 확률은 데이터 피쳐 (Feature) 또는 클래스 레이블 (Class Label) 간의 종속성에 영향을 받는다. 일반적으로 레이블 노이즈를 모델링하는 접근 방식은 손상 프로세스 (Corruption Process)가 데이터의 피쳐와 무관하며 레이블 간에 종속성만 있다고 가정하는 객체 비 종속성 레이블 노이즈 (Instance-independent Label Noise) $T_{ij} := p(\tilde{y} = j|y = i)$, 그리고 데이터의 피쳐 및 레이블 간에 모두 종속성이 있다고 가정하는 객체-종속성 레이블 노이즈 (Instance-dependent label noise $T_{ij} := p(\tilde{y} = j|x, y = i)$ 가 있다 [7]. 객체-종속성 레이블 노이즈의 경우 인위적으로 데이터에 주입하기 어렵게 때문에 2.3장에서 소개하는 실 세계 노이즈 데이터를 통해서 성능을 평가하는 것이 보편적이다.

2.3. 실 세계 노이지 데이터

실 세계 노이즈 데이터는 깨끗한 데이터와 달리 실 세계 레이블 노이즈가 학습 데이터에 포함되어져 있다. 문헌에 [2]에 따르면 그림 2에서와 같이 서로 다른 도메인에서 수집되어진 Animal-10N (동물 도메인), Food-101N (음식 도메인), Clothing1M (의류 도메인), WebVision (ImageNet 도메인) 데이터가 커뮤니티에서 활발이 사용되고 있다. 각 데이터의 노이즈 비율은 레이블이 어떻게 수집되었는가에 따라 다르며, 실제 사람이 개입하여 라벨을 수행한 경우 노이즈 비율이 더 낮다. 따라서, 크롤링하는 이미지 주변의 텍스트 정보를 사용한 Clothing1M 데이터는 38.5%의 가장 높은 노이즈 비율을 보이며, 검색 키워드를 통해 수집한

Food-101N와 WebVision 데이터는 18.4-20.0%의 노이즈 비율을 보이며, 비 전문가인 사람으로부터 레이블을 획득한 Animal-10N데이터의 경우 가장 낮은 8.0%의 노이즈 비율을 보인다. 해당 데이터에 대한 자세한 통계는 논문 [2]을 참고하기를 추천한다.

3. 강건한 딥러닝 기법

그림 3과 같이 레이블 노이즈에 강건한 딥러닝 방법들은 모델 아키텍처, 정규화, 로스 함수, 로스 조정, 샘플 선택을 포함한 다양한 접근법을 통해서 활발히 연구되고 있다.

3.1. 강건한 아키텍처 (Robust Architecture)

레이블 손상 프로세스를 모방하여 기본 딥 러닝 아키텍처 상단에 노이즈 적응 계층 (Noise Adaptation Layer)를 추가하는 방식이다. 구체적으로, 한 학습 객체 x가 주어질 때, 깨끗한 레이블에 대한 확률을 예측하는 것이 아니라 노이즈 적응 계층을 통해 추정된 노이즈 손상 확률 $T_{ij} = p(\tilde{y} = j|y = i, x)$을 고려하여 정답 레이블이 아닌 노이지 레이블에 대한 예측 확률을 학습한다 [8].

$$p(\tilde{y} = j|x) = \sum_{i=1}^{c} T_{ij} \cdot p(y = i|x),$$

학습 후 테스팅 환경에서는 노이즈 적응 계층 T_{ij}은 분리되고 깨끗한 라벨이 대한 확률 $p(y=i \mid x)$을 예측에 활용한다.

3.2. 강건한 정규화 (Robust Regularization)

정규화는 딥 러닝 모델의 일반화 성능 향상을 위한 일반적인 방법이다. 배치 정규화 및 드롭 아웃과 같이 광범위하게 연구되어왔으며 모델 학습에 있어서 필수적으로 사용되고 있다. 레이블 노이즈에 강건한 정규화 방법은 모델 파라미터의 변화량에 페널티 (Weight Decay)를 주거나 확율론적인 무작위성 (Stochastic Randomness)을 추가하는 기존 정규화 방법과 달리, 주어진 학습 레이블에 변화를 주어 모델의 강건성을 높이는 것을 목표로 한다.

레이블 평활화 (Label Smoothing) [4]는 훈련 중 레이블 잡음의 주변화된 효과 (Marginalized Effect)를 추정하여 모델이 잘못된 노이즈에 대한 과적합을 줄이는 것을 목표로 한다. 따라서, 레이블 평활화는 주어진 노이지 레이블 $\tilde{y}$를 가능한 모든 레이블에 대해서도 적은 값을 가지도록 다음과 같이 변형한다,

$$\bar{y} = \langle \bar{y}(1), \bar{y}(2), ..., \bar{y}(c) \rangle, \text{ where}$$

$$\bar{y}(1) = (1-\alpha) \cdot [\tilde{y} = i] + \alpha/c \text{ 그리고 } \alpha \in [0,1].$$

여기서, $[\cdot]$는 Iverson bracket이며 α는 평활화 수준이다.

대조적으로, 믹스 업 (Mixup) 정규화 [9]는 두 학습 데이터 쌍에 대해서 데이터의 피쳐 x와 레이블 y을 모두 볼록 보간법 (Convex Interpolation)을 활용해 섞어 가상의 객체를 생성하며,

$$x_{mix} = \lambda x_i + (1-\lambda)x_j, \cdots y_{mix} = \lambda y_i + (1-\lambda)y_j,$$

이는 학습에서 활용되는 데이터의 분포를 확장하여 노이즈 레이블에 대한 과적합을 방지한다.

3.3. 강건한 로스 함수 (Robust Loss Function)

일반적인 지도 학습에서 널리 쓰이는 Cross-entropy (CE) 로스 함수는 노이즈 레이블 과적합 문제에 민감하다고 알려져 있고, 이를 보완하기위한 여러 강건한 로스 함수들이 제안되어져 왔다.

대표적으로, 노이즈에 강건한 Mean Absolute Error (MAE) 로스와 CE로스를 결합한 Generalized Cross Entropy (GCE) [10], 대칭적인 특성을 CE에 결합한 Symmetric Cross-Entropy (SCE) [3], 커리큘럼 적인 특성을 반영한 Active Passive Loss (APL) [11]과 같은 방법들이 있으며, 이러한 방법들은 이론적으로 노이즈에 대한 강건함이 잘 입증되어 있고 모델 학습에 쉽게 활용되어질 수 있는 장점을 지닌다.

3.4. 로스 조정 (Loss Adjustment)

로스 조정은 학습 시 잘못된 레이블에 대한 부정적인 영향을 직접적으로 해결하는 효과적인 방법이다. 기존에 얻어진 학습 로스를 그대로 사용하는 것이 아니라, 모델을 업데이트하는 역전파 과정 직전에 해당 로스를 로스 재 가중 (Loss Reweighting) 혹은 레이블 수정 (Label Correction)을 통해 보다 올바른 값으로 수정한다. 로스 재 가중 방식은 부정확한 레이블에 대한 가중치는 줄이는 반면 정확한 레이블에 대한 가중치는 증가시키는 방법이며, 학습 객체마다 다른 가중치 $w(x,\tilde{y})$로 로스 값을 조정한다,

$$\Theta_{t+1} = \Theta_t - \eta \nabla \left(\frac{1}{|B_t|} \sum_{(x,\tilde{y}) \in \tilde{\mathcal{D}}} w(x,\tilde{y}) \cdot \ell(f(x;\Theta_t),\tilde{y}) \right)$$

가중치 $w(x,\tilde{y})$를 추정하기 위한 방법으로는 일관성 없는 예측을 보이는 객체에 대해 높은 가중치를 부여

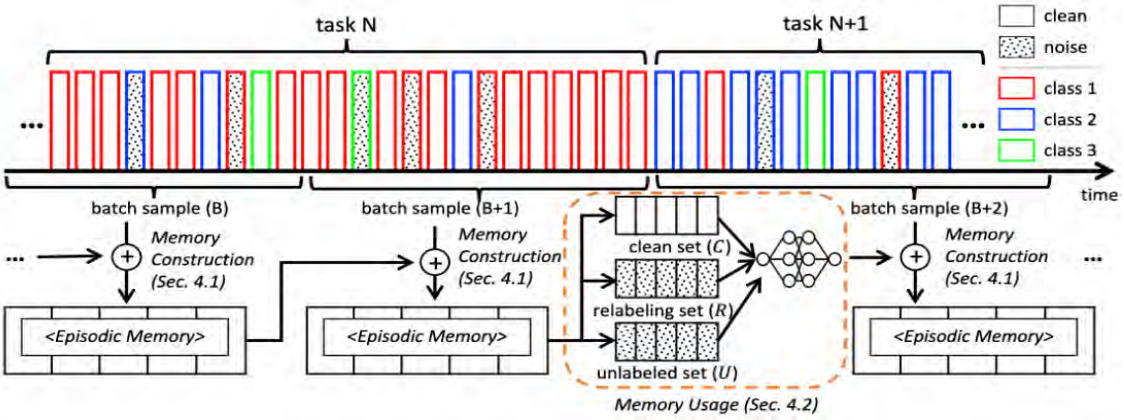

그림 4 레이블 노이즈에 강건한 연속학습을 위한 PuriDiver 방법론 [5].

하는 Active Bias [12]와 그래프 뉴럴 네트워크 (Graph Neural Network)를 기반으로 레이블 간의 구조적 관계를 활용하는 DualGraph [13] 등이 있고, 가중치 대신 학습 객체의 레이블을 수정하는 레이블 수정 방법은 SELFIE [14]와 같은 방법들이 존재한다.

3.5. 샘플 선택 (Sample Selection)

로스 조정과 대조적으로 노이지 데이터 중 깨끗한 데이터만을 선택적으로 활용하는 방법이 샘플 선택이다. 전통적인 데이터 클리닝 방식에 착안하여, 딥 뉴럴 네트워크에 대한 순방향 전파 후에 얻어지는 각 학습 객체들에 대한 학습 로스 혹은 미분 값의 크기 등을 활용하여 깨끗한 레이블을 가진 객체들을 판별한다. 그리고 선택되어진 미니 배치 샘플에 대해서만 역전파를 하여 잘못 레이블 된 객체에 대한 과적합문제를 피할 수 있도록 한다. 일반적으로, 대부분의 샘플 선택 방식은 학습 시 깨끗한 객체들이 우선적으로 학습되고 그 후 노이즈 객체들에 대해 과적합 하게 되는 암기 효과 (Memorization Effect) [15]에 기반하여 낮은 학습 로스를 갖는 객체들을 올바른 레이블을 가진 것이라고 가장한다. 따라서, 소 손실 (Small-loss) 기법 [16]이라고 불리는 대표적인 샘플 선택 방식은, 노이지 미니배치 $\tilde{B}(\in \tilde{\mathcal{D}})$가 주어졌을 때, 학습 로스가 작은 객체들로 구성된 깨끗한 집합 (Clean Set) $\mathcal{C}(\in \tilde{B})$를 선택하고 해당 집합만을 활용하여 모델을 최종적으로 다음과 같이 업데이트한다,

$$\Theta_{t+1} = \Theta_t - \eta \nabla \left(\frac{1}{|\mathcal{C}|} \sum_{(x,\tilde{y}) \in \mathcal{C}} \ell(f(x;\Theta_t), \tilde{y}) \right),$$

나머지 선택되지 않은 학습 객체들은 강건한 학습을 위해 학습에서 제외된다.

또한, 샘플 선택 방식은 최근 반 지도 학습 (Semi-supervised Learning)과 결합되어 큰 성능의 향상을 가져왔다. 선택된 깨끗한 집합을 레이블이 있는 집합으로 가정하고, 그 외 객체 집합을 레이블이 없는 집합으로 가정하여 비 지도 학습방법을 간단하게 적용할 수 있다. 대표적인 방법으로는, MixMatch를 Small-loss 기법과 결합한 DivideMix [17]가 있으며 다양한 레이블 노이즈에 대해서 큰 성능향상을 보였다.

4. 연속학습 (Continual Learning)

딥 러닝에서 연속학습이란 점진적 학습 (incremental Learning) 이라고도 불리며, 학습 모델이 과거의 데이터에 대해서 추가적인 접근 없이 순차적으로 여러 작업 (Task)을 학습해 나가는 것을 말한다. 기존의 온라인 러닝 (Online Learning)과 달리 연속학습에서는 모델이 빠르게 새로운 지식을 학습하면서도 과거에 배운 지식을 잊어버리지 않아야 한다. 특히, 과거의 지식을 잊어버리는 문제는 치명적인 망각 (Catastrophic Forgetting)이라고 불리며 이는 새로운 작업에 대해서 학습 모델의 가중치가 과하게 변화되기 때문이다. 이를 위한 대표적인 해결책은 작은 크기의 메모리를 지니고 전체 데이터를 대표할 수 있는 다양성이 높은 객체들을 유지하여 학습에 지속적으로 활용하는 것이며, 에피소드 메모리 (Episodic Memory)를 활용한 학습이라 불린다. 뿐만 아니라, 실 세계 응용에서는 데이터의 입력 스트림이 다양한 정도의 레이블 노이즈를 지니고 있기 때문에, 최근에 레이블 노이즈에 대해 강건한 연속학습 방법을 고안하려는 노력이 있어왔다.

그림 4에 묘사되어지는 PuriDiver [5] 방법론은 최근 제안되어진 방법 중 가장 레이블 노이즈에 좋은 성능을 보이는 에피소드 메모리 기반의 방식이다. 종

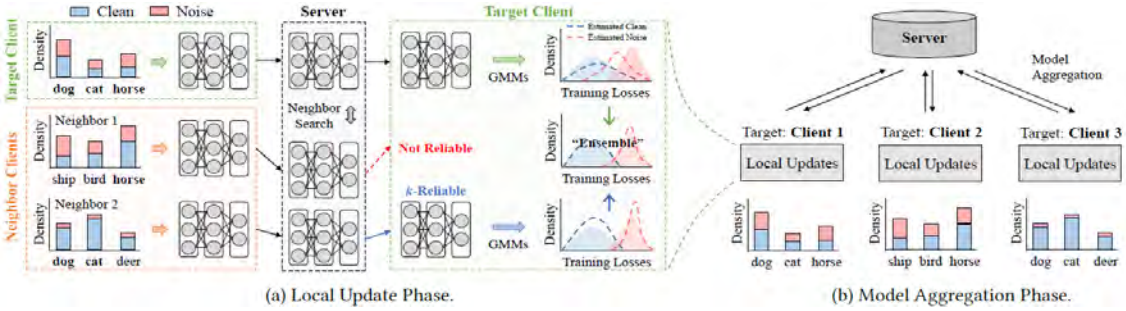

그림 5 레이블 노이즈에 강건한 연합학습을 위한 FedRN 방법론 [6].

합적으로, 강건한 연합학습을 위한 PuriDiver 방법은 신뢰가능한 메모리 구축 그리고 메모리를 활용한 강건한 학습으로 이어지는 두 단계로 구성된다.

4.1. 신뢰 가능한 에피소드 메모리 구축

강건한 연속학습에서의 도전과제는 신뢰 가능한 객체를 메모리에 포함하면서도 치명적인 망각 문제를 해소하기 위해 메모리 내 선택된 객체들의 다양성을 높게 유지해야 하는 것이다. 이를 해결하기 위해, PuriDiver는 객체 레이블에 대한 대한 순도 (Purity)와 객체 피쳐의 다양성 (Diversity)을 동시에 고려할 수 있는 새로운 샘플선택 점수 기준 (Score Metric) $(x_i, \tilde{y}_i)$을 다음과 같이 제시한다,

$$S(x_i, \tilde{y}_i) = (1-\alpha_k)\ell(x_i, \tilde{y}_i) + \alpha_k \cdot {}^1/_{|\mathcal{M}[\tilde{y}_i]|} \sum_{\hat{x}_i \in \mathcal{M}[\tilde{y}_i]} \cos\left(f_{rel}(x_i; \tilde{y}_i), f_{rel}(\hat{x}_i; \tilde{y}_i)\right),$$

여기서, x_i는 새롭게 들어오는 레이블 $\tilde{y}_i$를 가지는 스트리밍 객체, M은 현재 시점의 에피소드 메모리, $M[\tilde{y}_i]$는 에피소드 메모리 내에 레이블이 $\tilde{y}_i$와 동일한 부분 집합이다. 해당 점수 수식의 첫 번째 행은 주어진 객체에 대한 학습 로스로 해당 객체에 대한 순도를 추정하며, 두 번째 행은 주어진 객체와 메모리 내 객체 간의 다양성을 코사인 유사도 (Cosine Similarity)로 추정한다. 따라서, 로스와 코사인 유사도가 낮을수록 순도와 다양성이 높은 객체가 되어 새롭게 에피소드 메모리로 포함되게 된다. 반면에, 에피소드 메모리 내의 점수가 높은 객체들은 메모리 내에서 제외된다.

4.2. 메모리를 활용한 강건한 학습

에피소드 메모리 구축에 있어서 순도와 다양성을 모두 고려하기 때문에 모든 선택된 객체들이 올바른 레이블을 가지지 않을 수 있다. 이 문제를 해결하기 위해, PuriDiver방법론은 메모리를 활용한 학습에 강건한 학습기법을 추가한다. 그림 4와 같이, 메모리에 속한 객체들을 깨끗한 레이블을 지닌 집합 (Clean Set)과 새로운 레이블이 지정된 집합 (Relabeled Set)로 분류하고, 그 외 두 집합에 포함되지 않은 객체들을 레이블을 지니지 않은 집합 (Unlabeled Set)으로 간주한다. 즉, 기존의 강건한 학습법에서 제안되어진, 샘플 선택, 재라벨링, 반지도학습을 각 집합들과 결합하여 강건한 연속학습을 실현한다.

5. 연합학습 (Federated Learning)

연합학습은 다수의 로컬 클라이언트와 하나의 중앙 서버가 협력하여 데이터가 탈 중앙화 된 환경 (Decentralized Environment)에서 전체 데이터에 대한 글로벌 모델을 학습하는 기법이다. 로컬 클라이언트는 스마트 폰과 같은 IoT (Internet-of-Things) 장치들을 말하며, 데이터 프라이버시 향상을 위해 다른 클라이언트가 가진 데이터에는 접근할 수 없다는 제약을 가지고 있다. 즉, 개인 정보가 보호되어야 하는 상황에서 데이터 유출 없이 고성능 모델을 학습 가능하다는 장점을 지닌다.

레이블 노이즈와 관련해서 연합학습 환경은 두 가지의 어려움을 만들고 그로 인해 더욱 도전적이다 [6]. 첫 째로, 각 클라이언트의 로컬 데이터는 각 클래스에 대해 Non-Identically Distributed Data이므로 데이터의 분포가 클라이언트 별로 모두 다른 데이터 이질성 (Data Heterogeneity)을 지닌다. 둘 째로, 각 클라이언트는 데이터를 다른 환경에서 수집하므로 수집된 데이터의 레이블 노이즈 비율과 타입이 다른 다양성 (Varying Label Noise)을 가진다.

최근 연합학습 하에서 레이블 노이즈 문제를 해결하기 위한 방법을 고안하려는 노력이 있어왔다. 그림 5에 묘사되어지는 FedRN [6] 방법론은 데이터 이질성과 다양한 레이블 노이즈에 대한 문제를 신뢰가능한 이웃 클라이언트 (k-Reliable Neighbor Client) 개념을 도입하여 해결하고자 하였다. 해당 방법론은 기

존에 가장 활발이 사용되는 연합학습 파이프라인인 FedAvg [18]을 k-이웃 클라이언트와 이를 활용한 강건한 학습을 통해 레이블 노이즈에 대한 과적합을 크게 해소하였다.

5.1. 신뢰가능한 k-이웃 클라이언트

FedRN의 핵심 아이디어는 개인 정보를 침해하지 않고 다른 클라이언트의 유용한 정보를 활용하는 것이다. 이웃 클라이언트 중에 가장 노이즈를 적게 가지는 - 높은 데이터 전문성 (High Data Expertise) - 혹은 자기자신 (타겟 클라이언트)과 데이터 분포가 가장 유사한 - 높은 데이터 유사성 (High Data Similarity) - 클라이언트들을 신뢰가능한 k-이웃 클라이언트로 정의한다. 따라서, 그림 5에서 보이듯이, Neighbor 2가 신뢰가능한 클라이언트로 선택된다.

구체적으로, 타겟 클라이언트를 c 그리고 한 이웃 클라이언트를 n이라 할 때, 이웃 클라이언트의 데이터 전문성 (Exp) 및 유사도 (Sim)을 복합적으로 고려할 수 있는 신뢰 점수 (Reliability Score) $R(c,n)$ 을 제시하고, 신뢰가능한 k-이웃 클라이언트 집합 $\mathcal{R}_c(k)$을 선택한다.

$$\mathcal{R}_c(k) = \text{argmax}_{|(\mathcal{R}':n\in\mathcal{R}')|=k} R(c,n),$$

$$R(c,n) = \alpha \cdot \text{Exp(n)} + (1-\alpha)\cdot \text{Sim(c,n)}$$

여기서 a는 데이터 전문성 및 유사도 간의 기여를 결정하는 계수이다.

개인 정보를 보호하면서 (타 클라이언트 데이터에 접근하지 않으면서) 데이터 전문성과 유사도를 근사할 수 있는 방법을 제시한다. 데이터 전문성의 경우, 노이즈가 작은 데이터일수록 학습이 빠르게 일어난다는 암기 효과 [15]에 기반하여 이를 다음의 표준화된 학습 정확도로 근사한다.

$$\text{Exp}(c) = \frac{\text{Acc}(c) - \text{minAcc}(\{c\}\cup\mathcal{N}_c)}{\text{maxAcc}(\{c\}\cup\mathcal{N}_c) - \text{minAcc}(\{c\}\cup\mathcal{N}_c)}.$$

여기서, Acc는 로컬 클라이언트 c에 대한 학습 정확도, minAcc와 maxAcc는 주어진 클라이언트 집합 (타겟 클라이언트와 그것의 이웃 클라이언트 $\mathcal{N}_c$)에 대한 최소 및 최고 학습 정확도이다.

반면에, 데이터 유사도는 특정 가우시안 랜덤 노이즈로 생성된 동일 객체 $\hat{x}$에 대해서 두 클라이언트 Θ_c와 Θ_n간의 코사인 유사도로 근사 된다,

$$\text{Sim}(c,n) = \text{Cosine}\big(p(\hat{x};\Theta_c), p(\hat{x};\Theta_n)\big)$$

그림 6 다중 레이블 환경에서의 다양한 레이블 노이즈. 1은 물체가 존재함 0은 존재하지 않음을 나타낸다.

5.2. k-이웃 클라이언트를 활용한 강건한 학습

선택된 k-이웃 클라이언트는 높은 신뢰도 점수를 갖지만, 각 클라이언트별 점수는 모두 다르다. 따라서, 신뢰도 점수의 차이를 고려하여 깨끗한 샘플 집합을 노이지 데이터로부터 선택할 수 있는 방법을 제안한다. 3.5장에서 언급한 샘플 선택 방법인 소 손실 (Small-loss) 방법을 따른다. 따라서, 로스가 적을수록 깨끗한 레이블을 갖을 확률이 높은 확률 분포 $p(\ell(x,\tilde{y};\Theta))$를 Gaussian Mixture 모델로 근사하여 만들고, 이를 모든 이웃 클라이언트의 신뢰 점수 $R(c,n)$를 활용한 앙상블 (Ensemble)로 조인트 확률분포를 만든다,

$$p\big(clean|x;\mathcal{R}_c(k)\big) = \sum_{n\in\{c\}\cup\mathcal{R}_c(k)} R'(c,n)\times p\big(\ell(x,\tilde{y};\Theta)\big),$$
$$\text{where } R'(c,n) = R(c,n) \Big/ \sum_{n'\in\{c\}\cup\mathcal{R}_c(k)} R(c,n').$$

따라서, FedAvg 파이프라인을 따라, 각 클라이언트는 각자의 모델을 자신의 로컬데이터 중 깨끗한 객체로 추정되는 집합 $\mathcal{C} = \{x|p\big(clean|x;\mathcal{R}_c(k)\big) > 0.5\}$에 대해 학습을 하고, 그 후 글로벌 모델을 만들기 위해서 모든 클라이언트의 로컬 모델을 합친다.

FedRN 방법은 다양한 노이즈 타입 및 정도를 가지는 클라이언트 풀에서 레이블 노이즈에 대한 강건성을 크게 향상시켰다.

6. 향후 연구 방향

머신 러닝 커뮤니티의 노력으로 딥 러닝에 대한 견고성은 여러 방향으로 크게 향상되었다. 이 섹션에서

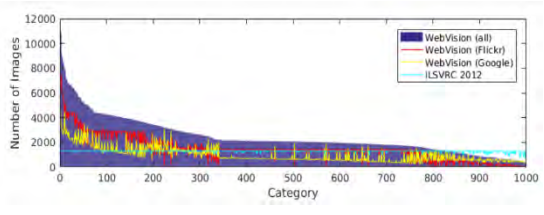

그림 7 WebVision 데이터의 클래스 불균형 문제.

는 레이블 노이즈 영역에서 딥 러닝의 발전을 촉진할 수 있는 두 가지 향후 연구를 논의한다.

6.1. 레이블 노이즈가 있는 다중 레이블 데이터

현재까지의 대부분 연구는 한 이미지에 하나의 물체만 존재한다 가정하는 단일 레이블 (Single-label) 데이터에 대한 다중 클래스 분류 문제를 풀어왔다. 하지만, 한 이미지는 여러가지 물체를 포함하는 것이 일반적이며, 물체 탐색과 같은 보다 어려운 머신 러닝 작업들이 존재한다. 특히, 다중 레이블 환경에서는 한 이미지가 깨끗한 레이블과 부정확한 레이블을 모두 가질 수 있기 때문에 그림 6에서 보이듯이 다양한 종류의 새로운 노이즈 라벨이 정의될 수 있다. 강아지 (Dog)와 문 (Door)이 존재하는 이미지가 주어질 때, 해당 이미지에 대한 레이블은 다양한 방식으로 부정확하게 수집될 수 있다. 기존 단일 레이블 환경과 같이 혼동되는 물체로 잘못 라벨링 될 수 있으며 (Mislabeling), 시스템 에러나 악의적인 의도로 랜덤하게 뒤바뀔 수 있으며 (Random Flip), 눈에 보이는 물체 중 일부만 라벨링 될 수도 있다 (Missing Label). 따라서, 이와 같은 다양한 타입의 새로운 레이블 노이즈를 해결하는 것은 보다 현실적인 다중 레이블 분류 혹은 물체 탐색을 위해서는 중요 도전과제 중 하나이다.

6.2. 레이블 노이즈가 있는 클래스 불균형 데이터

또 다른 새로운 도전과제는 학습 데이터에서 클래스 간의 샘플 숫자가 다른 클래스 불균형 문제이다. 그림 7에서 보이듯이 실 세계 노이즈 데이터인 WebVision은 실 세계 데이터 수집에서 발생하는 클래스 불균형 정도를 잘 보여주는 예시이다. 딥 러닝 학습에서, 모델은 데이터 수가 많은 클래스 (Majority Class)에 과적합되기 때문에 데이터의 수가 작은 클래스 (Minority Class)에 대해서는 굉장히 낮은 성능을 보이는 것이 일반적이다. 특히, WebVision 데이터처럼 클래스 불균형 문제는 레이블 노이즈와 함께 발생하기 때문에 두 문제를 복합적으로 해결할 수 있는 방법의 필요성이 커지고 있다.

7. 결 론

본 기고에서는 레이블 노이즈에 강건한 딥 러닝을 위한 전체적인 연구 동향에 대해서 살펴보았다. 대표적으로, 강건한 아키텍처, 정규화, 로스 함수, 그리고 로스 조정, 샘플 선택의 방법들을 위주로 활발한 연구가 진행되었으며, 레이블 노이즈에 대한 과적합 문제를 해결하는데 기여하였다. 또한, 실 세계 응용을 위한 연속학습 및 연합학습에 대해 간략히 소개하였고, 레이블 노이즈에 대한 문제를 두 새로운 학습 환경하에서 어떻게 풀어낼 수 있는지에 대해서 PuriDiver와 FedRN 두 방법을 중심으로 소개하였다. 마지막으로, 다중 레이블과 클래스 불균형 문제를 중심으로 향후 연구방향에 대해 논의하였다. 본 기고를 통해 살펴본 레이블 노이즈에 대한 연구 동향이 여러 연구자들에게 도움이 되기를 기대한다.

참고문헌

[1] S. E. Whang, Y. Roh, H. Song, and J-G. Lee, “Data collection and quality challenges in deep learning: A data-centric AI perspective” in arXiv preprint arXiv:2112.06409, 2021.

[2] H. Song, M. Kim, D. Park, Y. Shing, and JG. Lee, “Learning from noisy labels with deep neural networks: A survey,” IEEE Transaction on Neural Networks and Learning Systems, 2022.

[3] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, “Symmetric cross entropy for robust learning with noisy labels,” in Proc. ICCV, 2019, pp. 322–330.

[4] M. Lukasik, S. Bhojanapalli, A. Menon, and S. Kumar, “Does label smoothing mitigate label noise?” in Proc. ICLR, 2020, pp. 6448–6458

[5] J. Bang, H. Koh, S. Park, H. Song, J-W. Ha, and J. Choi, “Online continual learning on a contaminated data stream with blurry task boundaries,” in Proc. CVPR, 2022, pp. 9275–9284.

[6] S. Kim, W. Shing, S. Jang, H. Song, and S-Y. Yun, “FedRN: Exploiting k-reliable neighbors towards robust federated learning,” in Proc. CIKM, 2022.

[7] B. Fr´enay and M. Verleysen, “Classification in the presence of label noise: A survey,” IEEE Transaction on Neural Networks and Learning Systems, 2013.

[8] J. Goldberger and E. Ben-Reuven, “Training deep neural-networks using a noise adaptation layer,” in Proc. ICLR, 2017.

[9] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “Mixup: Beyond empirical risk minimization,” in Proc. ICLR, 2018

[10] Z. Zhang and M. Sabuncu, “Generalized cross entropy loss for training deep neural networks with noisy labels,” in Proc. NeurIPS, 2018, pp. 8778–8788

[11] X. Ma, H. Huang, Y. Wang, S. Romano, S. Erfani, and J. Bailey, “Normalized loss functions for deep learning with noisy labels,” in Proc. ICML, 2020, pp. 6543–6553.
[12] H.-S. Chang, E. Learned-Miller, and A. McCallum, “Active Bias: Training more accurate neural networks by emphasizing high variance samples,” in Proc. NeurIPS, 2017, pp. 1002–1012
[13] H. Zhang, X. Xing, and L. Liu, “DualGraph: A graph-based method for reasoning about label noise,” in Proc. CVPR, 2021, pp. 9654–9663.
[14] H. Song, M. Kim, and J.-G. Lee, “SELFIE: Refurbishing unclean samples for robust deep learning,” in Proc. ICML, 2019, pp. 59075915.
[15] D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio et al., “A closer look at memorization in deep networks,” in Proc. ICML, 2017, pp. 233–242.
[16] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, “Co-teaching: Robust training of deep neural networks with extremely noisy labels,” in Proc. NeurIPS, 2018, pp. 8527–8537.
[17] J. Li, R. Socher, and S. C. Hoi, “DivideMix: Learning with noisy labels as semi-supervised learning,” in Proc. ICLR, 2020.
[18] B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. Arcas, “Communication-efficient learning of deep networks from decentralized data,” in Proc. AISTAT, 2017, pp. 1273–1282.

약 력

송 환 준

2014 부산대학교 정보컴퓨터공학부 졸업(학사)
2016 한국과학기술원 지식서비스공학과 졸업(석사)
2021 한국과학기술원 지식서비스공학과 졸업(박사)
2020 구글 리서치, 연구인턴
2021~현재 네이버 AI Lab, 연구 과학자
2021~현재 KAIST-NAVER 리서치 센터, 연구 과학자
관심 분야: 강건한/대규모 머신 러닝, 컴퓨터 비전, 데이터 마이닝
Email: hwanjun.song@navercorp.com