딥러닝 기반 분류모델의 선택적 망각

진송찬, 우사이먼성일
성균관대학교 (학부생)

Selective unlearning for DNN based model

Song-Chan Jin, Simon S. Woo
Sungkyunkwan University (Undergraduate student)

요 약

본 논문에서 제안하는 선택적 망각이란 딥러닝 모델이 일부 지식을 선택적으로 잊어버리는 것을 의미하며, 개인정보 보호를 위해 도입되었다. 이를 위해 데이터 재수정 및 모델 재학습 등의 방법이 있지만, 이러한 방법들은 일반적으로 계산량이 많거나 모델의 성능을 크게 저하시키는 문제가 있어서 이에 대한 대안으로 작은 데이터셋으로 다른 데이터들에 대한 지식은 유지한 채 특정 데이터들에 대한 지식만 잊는 경사 상승법을 소개하고 있다. 본 논문에서는 경사 상승법을 통하여 기존 재학습 기법 대비 9배 적은 계산량으로 선택적 망각을 수행할 수 있다는 결과를 얻었다.

This paper presents a selective unlearning technique for deep learning models to protect personal information. This technique involves forgetting some of the previously learned knowledge. Existing methods, such as data modification or model retraining, are computationally expensive and often result in performance degradation. As an alternative, our work proposes a gradient-ascent based approach that selectively forgets a specific class, while preserving knowledge learned from different classes. Our method achieves selective unlearning performance with 9 times fewer computational cost, compared to the existing method that trains from scratch.

I. 서론

1.1 선택적 망각의 의미

선택적 망각의 의미는 딥러닝 기반 모델이 이전에 학습한 지식 일부를 잊는 과정을 의미한다. 이는 학습 데이터셋이나 이전 지식에 대한 업데이트가 필요한 상황에서 모델을 재학습해야 할 때 유용하다.

1.2 선택적 망각의 필요성

ㄱ) 개인정보 보호

이전에 학습된 모델이 개인정보를 포함한 민감한 데이터를 보유하고 있다면, 그 데이터를 보호하기 위해서는 모델이 선택적 망각을 수행하여야 한다[1], [2].

ㄴ) 개념 변화

학습 데이터셋이 변경되거나 학습 도메인이 변화할 경우, 이전에 학습된 모델은 새로운 데이터에 대해 부정확한 예측을 할 수 있다. 그렇기에 선택적 망각이 필요하다[3].

ㄷ) 편향 수정

학습된 모델에 이전 훈련 데이터들에 대한 편향이 생겼다면, 선택적 망각은 이러한 편향을 수정할 수 있다[4].

본 논문에서는 ㄱ) 개인정보 보호를 위한 경사 상승법 기반의 선택적 망각 기법을 제안한

다. 지금까지 나온 방법으로는 데이터 재수정 및 모델 재학습, 그리고 매개 변수 기반의 망각 등의 방법이 있다. 하지만 이러한 방법들은 일반적으로 계산량이 많거나 모델의 성능을 크게 저하시키는 등 많은 문제점들이 존재한다. 그래서 본 논문에서는 특정 데이터들만을 잊는 기울기를 활용한 경사 상승법을 제안한다. 이를 통해 모델은 다른 데이터들에 대한 지식은 유지한 채 특정 데이터들에 대한 지식만 선택적으로 잊을 수 있다.

II. 관련 연구

2.1 데이터 재수정 및 모델 재학습

선택적 망각을 수행한 가장 기본적인 방법론으로써 훈련 데이터셋 자체에 대한 재수정을 하는 방법이 있다. 특정 클래스에 대한 데이터를 추가하거나, 편향이 심한 데이터를 제거한다. 그리곤 재수정된 데이터셋을 사용하여 모델을 다시 학습시킨다. 하지만 이러한 재수정된 데이터셋을 활용한 재학습은 많은 시간과 계산량[5], [6]을 요구한다.

2.2 매개 변수 기반 망각

모델의 학습된 매개 변수를 수정[7]하여 이전에 학습한 지식을 삭제하는 방법이다. 모델의 가중치를 초기화하거나 수정하여 망각을 진행한다. 이러한 기법은 학습 계산량을 크게 줄일 수 있다. 하지만 이러한 기법은 모델 자체의 성능을 많이 저하시킨다.

III. 본론

3.1 경사 상승법을 활용한 선택적 망각

앞서 설명한 다양한 선택적 망각 기법들은 많은 시간과 계산량을 요구하거나, 모델 자체의 성능을 많이 저하시키는 등 단점을 가진다. 그렇기에 본 논문에서는 모델의 성능을 최대한 유지하면서 짧은 시간과 적은 계산량으로 선택적 망각을 수행하는 기법을 제안한다. 바로 경사 상승법을 활용한 선택적 망각이다.

일반적으로 딥러닝 기반의 분류모델들은 손실함수 L을 최소화하는 방향으로 모델의 가중치 w를 업데이트하게 된다. 이때 L을 최소화하는 방향은 함수의 기울기를 활용하여 정할 수 있다. 함수의 기울기에 대하여 반대로 이동하게 되면 결론적으로 다음과 같이 모델은 가중치를 업데이트[8], [9]하게 된다.

$$w_{\neq w} = w - \frac{\partial L}{\partial w} \tag{1}$$

또한, 이때 분류모델은 손실함수 L로써 크로스 엔트로피 손실[10]을 가장 많이 사용한다. 여기서 t_k는 k번째 데이터의 라벨을 의미하고 y_k는 k번째 데이터의 추론값을 의미한다.

$$L = -\sum_{k=1}^{n} t_k \ln(y_k) \tag{2}$$

분류모델에서의 라벨은 원핫 인코딩으로 표현되어 있어서 해당 클래스에 대한 추론값만 손실함수에 영향을 주게 된다.

$$L = -t_{k=target}\ln(y_{k=target}) - \sum_{k \neq target}^{n} t_k \ln(y_k) \tag{3}$$

$$\frac{\partial L}{\partial w} = \frac{\partial(-t_{k=target}\ln(y_{k=target}))}{\partial w} \tag{4}$$

우리는 여기서 착안하여 해당 클래스에 대한 추론값만 활용하여, 해당 클래스에 대한 선택적 망각을 진행한다. 즉, 정상적인 모델 학습인 경사 하강법을 역으로 진행하는 경사 상승법을 도입한 것이다.

경사 상승법으로 재학습 되는 모델은 적은 학습량을 가지며, 모델의 성능을 최소한으로 저하시키는 기법이다. 따라서 경사 상승법은 다음과 같이 모델의 가중치를 업데이트하게 된다.

$$w_{\neq w} = w + \frac{\partial L}{\partial w} \tag{5}$$

IV. 실험 및 결과

4.1 실험 설계

실험에 앞서 CIFAR-10 데이터셋(C_{10})을 이용하여 다음과 같은 D, D_{truck}, T, T_{truck} 데이터셋을 구성하였다.

	BaseLine		*Amnesia*		*Ours*	
T_{truck}	*accuracy*	f_1score	*accuracy*	f_1score	*accuracy*	f_1score
	86.6%	0.93	0.00%	0.00	**17.4%**	0.30
$T - T_{truck}$	*accuracy*	f_1score	*accuracy*	f_1score	*accuracy*	f_1score
	81.5%	NaN	82.9%	NaN	**72.0%**	NaN

표1. 모델 간 선택적 망각 비교

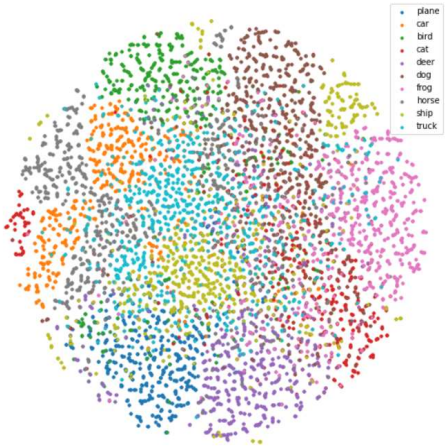

그림1. 트럭(**하늘색**) 선택적 망각 후

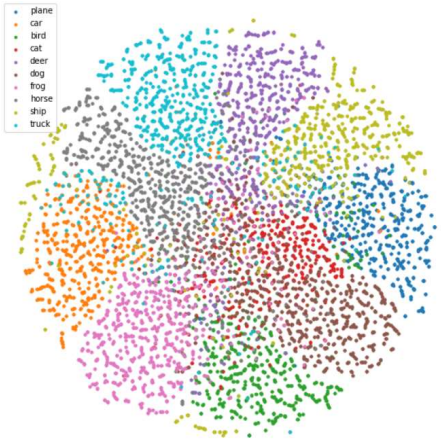

그림2. 트럭(**하늘색**) 선택적 망각 전

$D = all\ data \in C_{10}\ training\ dataset$

$D_{truck} = truck\ data \in C_{10}\ training\ dataset$

$|D| = 50{,}000, |D_{truck}| = 5{,}000$

$T = all\ data \in C_{10}\ test\ dataset$

$T_{truck} = truck\ data \in C_{10}\ test\ dataset$

$|T| = 10{,}000, |T_{truck}| = 1{,}000$

Baseline 모델은 D를 훈련 데이터셋으로 학습한 모델이다. Amnesia 모델은 $D - D_{truck}$을 훈련 데이터셋으로 학습한 모델이다. 즉, Amnesia 모델은 이상적으로 선택적 망각을 수행한다.

Baseline 모델과 Amnesia 모델은 batch size=64, learning rate=5e-4, epoch=55인 동일한 환경 속에서 학습된 모델이다.

본 연구에서 제안하는 모델(*Ours*)은 Baseline 모델에서 D_{truck}을 훈련 데이터셋으로 재학습한 모델이다. 이 모델은 batch size=64, learning rate=5e-4, epoch=55인 환경 속에서 학습된 모델이다.

만약 모델이 특정 데이터들을 잘 분류하지 못한다면 그 모델은 해당 데이터들을 잊었다고 판단할 수 있다.

그래서 본 논문의 최종 목표는 Baseline 모델에서 경사 상승법을 활용한 선택적 망각을 진행하여, 적은 계산량과 학습 시간으로 Amnesia 모델과 비슷한 분류 성능을 내는 모델을 만드는 것이다. 즉, Baseline 모델에서 작은 데이터셋만으로 재학습을 진행 시켜서, T_{truck}에 대한 분류 성능은 최대한으로 하락시키고 $T - T_{truck}$에 대한 분류 성능은 최소한으로 하락시키는 것을 목표로 한다.

4.2 실험 결과

우선 Amnesia 모델에 사용된 훈련 데이터셋과 본 연구에서 제안한 모델(*Ours*)에 사용된 훈련 데이터셋은 9배의 크기 차이가 난다.

$$|D - D_{truck}| = 9|D_{truck}| \quad (6)$$

따라서 제안된 모델이 Amnesia 모델보다 9배 적은 계산량을 가진다.

$$O(Amnesia) = 9O(Ours) \quad (7)$$

또한 표1.을 참고하면 기존 Baseline 모델과 비교하여 제안된 모델에서는 *truck*에 대한 f1 score가 기존 0.93에서 0.3으로 대비 크게 하락한 것을 볼 수 있다. 즉, *truck*에 대하여 분류가 잘되지 않기에 충분히 망각이 수행되었음을 확인할 수 있다.

그리고 그림 1.의 Baseline 모델의 t-sne 그림과 그림 2.의 제안된 모델의 t-sne 그림을 비교해보면 확실히 *truck*에 대한 경계가 사라졌음을 알 수 있다. 제안된 모델은 *truck*에 대한 특징을 망각한 것이다.

V. 분석

표1.에서 보는 바와 같이 T_{truck}에서 Baseline 모델과 비교하면 본 연구에서 제안된 모델은 **79.9%**의 성능 **하락**이 있었다.

$$\frac{86.6 - 17.4}{86.6} = 0.799 \quad (8)$$

즉 제안된 모델은 *truck*을 잘 잊었다고 판단할 수 있다.

아울러 $T - T_{truck}$에 대하여 Baseline 모델과 비교하면 제안된 모델은 **11.6%**의 전체적인 성능 **하락**만이 있었다.

$$\frac{81.5 - 72.0}{81.5} = 0.116 \quad (9)$$

즉 제안된 모델은 *truck*을 제외한 다른 데이터들은 여전히 잘 기억하고 있다고 판단할 수 있다.

이렇듯 결과적으로 경사 상승법은 기존 재학습 대비 9배 적은 계산량으로 성공적인 선택적 망각을 수행한다고 판단할 수 있다.

VI. 결론

본 연구에서는 특정 데이터들을 잊는 방향의 기울기를 활용한 경사 상승법을 제안하였고 이러한 경사 상승법을 통하여 기존 재학습 기법 대비 9배 적은 계산량으로 선택적 망각을 수행할 수 있다는 결과를 얻었다. 향후 계획은 모델의 전체적인 성능 저하를 더욱 줄일 수 있는 방향에 대하여 연구할 것이다.

[참고문헌]

[1] Privacy and Artificial Intelligence: A Conceptual Analysis by Brent Mittelstadt, Chris Russell, and Luciano Floridi (2016)

[2] Privacy, Ethics, and Data Access: A Case Study of the Fragile Families Challenge by Julia Stoyanovich, Bill Howe, et al. (2017)

[3] Machine Learning: The High-Interest Credit Card of Technical Debt by D. Sculley, et al. (2015)

[4] Mitigating Unwanted Biases with Adversarial Learning by Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell (2018)

[5] The Need for Bias Mitigation in Machine Learning: A Study on Face Recognition by Buolamwini and Gebru (2018)

[6] Machine Unlearning by Munoz-Gonzalez et al. (2020)

[7] Understanding and mitigating the tradeoff between robustness and accuracy by F. Tramèr, et, al. (2018)

[8] The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors by Linnainmaa (1970)

[9] Optimization of neural networks using gradient descent by Y. Bengio, S. Bengio (1986)

[10] A Mathematical Theory of Communication by Claude Shannon (1948)