CPR-CLASSIFIER-PROJECTION REGULARIZATION FOR CONTINUAL LEARNING

Abtract

본 논문은 entropy의 분류기의 출력 확률을 최대화하여 CPR은 추가적인 regularization term을 추가합니다.

추가항이 분류기의 출력으로부터 얻은 조건부 확률의 균일 분포의 projection으로 해석될 수 있습니다.

Introduction

본 논문에서는 neural network에서 wide local minima와 regularization based CL(Continual Learning) method 사이에 새로운 connection을 만듭니다.

일반적인 regularization based CL (Continual Learning) aim은 new task를 학습할 때 큰 deviation에 penalizing 하며 과거에 task를 사용할 때 important weight parameter를 보존하는게 목표입니다.

본 논문의 intuition은 local minima를 촉진합니다.

특히 regularization based CL 방법에 유용하며, 이는 새로운 작업에 대해 다양한 update 방향을 용이하게 만들어 plasticity 를 향상시키면서도 이전 작업에 해를 끼치지 않도록 안정성을 유지할 수 있기 때문입니다.

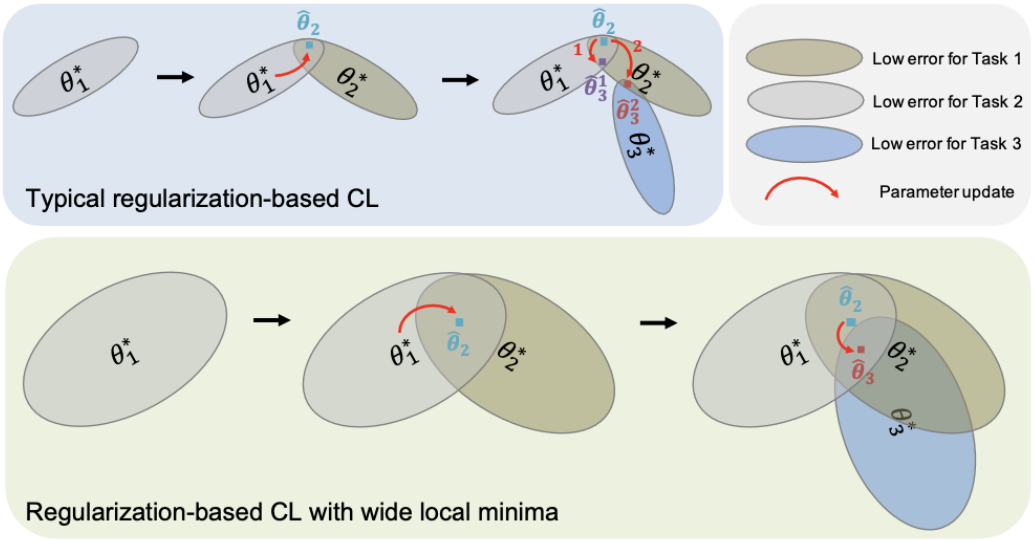

Figure 1

Fig1 (bottom), 처럼 low error를 가진 parameter를 포함하는 타원체가 더 넓을 때, 즉 wide local minima가 존재 할 때, new task의 sequence를 학습한 후에도 모든 작업에 대해 잘 수행하는 parameter를 찾는 유연성이 더 크게됩니다.

이 intuition을 기반으로, 본 논문의 method는 추가적인 regularization term을 실행하며 출력 함수의 분류기의 entropy를 극대화 하며, wide local minima를 촉진합니다.

구체적으로, CPR을 적용하는 것이 분류기의 출력을 균일 분포를 중심으로 한 유한 반경의 KL divergence 구의 투영으로 해석된다고 주장합니다.

- Pythagorean theorem을 KL divergence 에 적용하며, projection하여 continual learning method의 성능을 향상 시키며 prove 합니다.

2. CPR : Classifier Projection Regularization for wide local

CPR을 두개의 regularization term을 조합하여 공식화 합니다.

하나는 이전의 regularization based CL의 motivation이며, 다른 하나는 wide local minima를 유도하는 motivation 입니다.

본 논문은 CL에 CPR을 적용할 때, 성능 향상이 관찰되는 것에 대한 정보 기하학적 해석을 제시 합니다.

2.1 MOTIVATION : Introducing wide local minima in continual learning

특정한 작업 i에 대해 local minima를 달성하는 parameter는 θ_i^*로 나타내고, 정규화 항을 포함하여 얻은 parameter를 $\hat{\theta}_i^*$로 나타냅니다.

θ_i^*가 학습되었다 가정할 때, task2를 학습할 때, 적절한 정규항은 parameter를 θ_2^* 대신에 $\hat{\theta}_2^*$로 update 해야 합니다. 왜냐하면 $\hat{\theta}_2$는 task 1과 task2에서 모두 낮은 오류율을 달성하기 때문입니다.

하지만 low error (**Fig1. ellisope**) 이 좁을 때는 세 작업 모두에서 잘 수행되는 parameter를 얻는 것이 현실적으로 불가능합니다.

이 상황은 CL Regularization based 의 stability 와 plasticity 사이의 trade-off의 결과입니다.

다시 말하면, **stronger regularization strength** (과거 task 방향으로의 방향) 은 더 큰 stability 를 가져오며, 따라서 past task에 대한 forgetting이 더 작아집니다.

대조적으로 **weaker regularization strength**는 더 큰 plasticity 를 가져와 update된 parameter $\hat{\theta}_3$가 최근 작업에서 더 나은 성능을 발휘하지만, 이로인해 과거 작업의 성능이 손상될 수 있습니다.

이전 설정에서의 주요 문제 는 각 작업에 대한 low error를 달성하는 parameter 영역이 좁고 서로 겹치지 않는 다는 것 입니다.

따라서 직관적인 해결책은 낮은 오류 영역을 확장하여, 서로 교차하는 부분이 비어있지 않도록 하는 것입니다.

신경망의 wide local minima를 CL 중에 유도한다면, stability 과 plasticity 를 동시에 개선하며 모든 작업에 대한 높은 정확도를 동시에 얻을 수 있다는 가능성을 시사합니다.

2.2 Classifier projection regularization for continual learning

Regularization-based continual learning

전형적인 regularization based CL은 과거의 task에서 학습한 중요한 parameter의 편차를 벌점으로 부과하여, Catastrophic forgetting을 완화하는 regularization term을 첨부합니다.

$$L_{CL}^t(\theta) = L_{CE}^t + \lambda \sum_i \Omega_i^{t-1}(\theta_i - \theta_{i-1})^2$$

각 task t에 대해 ordinary cross-entropy loss function $L_{CE}^t(\theta)$

λ는 dimensionless regularization strength

Ω^{t-1}는 set of estimates of the weight importance

θ_i^{t-1} 는 $task^{t-1}$까지 learned parameter를 나타냅니다.

Single - task wide local minima

단일 작업을 해결하기 위해 신경망의 wide local minima 를 유도하기 위한 것 입니다.

$L_{WLM}(\theta) = L_{CE}(\theta) + \frac{\beta}{N}\sum_{n=1}^{N} D_{KL}(f_\theta(x_n)||g)$

g는 분류기 출력 f_θ를 정규화하는 Δm내의 어떤 확률 분포입니다.

β는 trade - off parameter, D_{KL}는 KL divergence

예를 들어, g가 Δm내의 균일 분포 P_U일 때, 정규화 항은 entropy 최대화와 일치하며, g가 다른 분류기의 출력 $\dot{f}_\theta$일 때, Zhang et.al 2018의 loss func과 동일합니다.

CPR : Archieving wide local minima in continual learing

Combining the above two regularization term

$L_{CPR}^t(\theta) = L_{CE}^t + \frac{\beta}{N}\sum_{n=1}^{N} D_{KL}(f_\theta(x_n)^t||P_U) + \lambda\sum_i \Omega_i^{t-1}(\theta_i - \theta_{i-1})^2$

λ와 β는 regularization parameters

첫 번째 정규화 항은 작업 t를 학습하는 동안 P_U를 정규화 분포 g로 사용하여 wide local minima를 유도합니다.

두 번째 항은 일반적인 Continual learning 에서 온 것입니다.

$$L_{WLM}(\theta) = L_{CE}(\theta) + \frac{\beta}{N}\sum_{n=1}^{N} D_{KL}(f_\theta(x_n)||g) \quad (2)$$
$$L_{CL}^t(\theta) = L_{CE}^t + \lambda\sum_i \Omega_i^{t-1}(\theta_i - \theta_{i-1})^2 \quad (3)$$

식 (2) 와 (3) 의 KL divergence 항을 최소화 하는 것을 최적화료 표현할 수 있습니다.

최적화는 다음과 같이 표현됩니다. $min_{Q\in\varrho} D_{KL}(Q||P)$

P는 주어진 분포이고, ϱ는 $Q \in \varrho$ 의 블록 집합입니다.

다시말해 최적화된, P^*는 KL divergence에 의해 측정되는 거리로써, P와 가장 가까운 Q의 내부 분포입니다.

이것은 information projection 으로 불리며 다음과 같이 나타냅니다.

$$P^* = argmin_{Q \in \varrho} D_{KL}(Q||P)$$

CPR를 구현할 때 projection을 위한 가능한 classifier 집합 C를 미리 정의 해야합니다.

직관적으로 가장 좋은 선택은 모든 작업에서 잘 수행되는 classifier 집합입니다.

하지만 CL에서는 이러한 분류기를 사용할 수 없습니다. 따라서 가능한 분류기 집합 C를 균일 분포 P_U를 중심으로 한 KL divergence로 선택합니다.

우리는 P_U를 선택하는 이유는 Δm의 중심이기 때문입니다.

따라서 어떤 분포와 P_U간의 최악의 경우 divergence 는 최대 $logM$입니다.

classifier projection의 관점에서 볼 때, 식 (3)의 CPR 정규화 항은 제약조건 $Q_{Y|X} \in (P_U, \epsilon)$의 Lagrange dual로 볼 수 있습니다.

이 항은 순차적인 작업을 훈련할 때, 변경을 최소화 하기 위해 개별 작업의 분류기를 균일 분포 방향으로 projection 하는 항입니다.

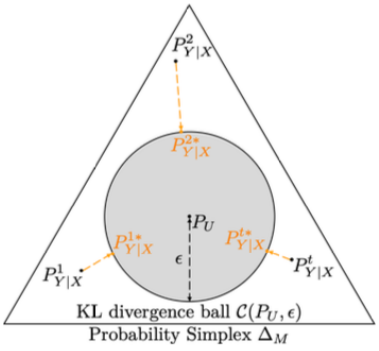

Figure 2: CPR can be understood as a projection onto a finite radius ball around P_U.

Related work

본 논문과 유사한 연구는 Aljundi(MAS) 의 연구 입니다. 이 연구는 각 작업의 표현을 희소하게 유지하기 위해 regularization based continual learning에 추가적인 정규화 항을 추가합니다.

이는 뉴런의 활성화의 희소성을 부과하는 것이며 이는 wide local minima를 유도하는 것과는 근본적으로 다릅니다.

또, Aljundi 의 연구는 average accuracy에 중점을 둔 반면, 우리는 CPR 정규화의 장점을 정규화 뿐만 아니라 CL의 가소성과 안정성을 동시에 높이는 측면에서 신중하게 평가합니다.

또, Mirzadeh 에서 최근에 제안한 방법인 wide local minima와 비슷하지만 그들과는 완전 다릅니다.

> **Firstly,** Mirzadeh는 forgetting에 대한 지표를 정의한 다음, 이를 2차 taylor 전개로 근사화 합니다.
> 그들은 주로 안정성에 초점을 두고 있으며, CL 중에 model이 wide local minima로 수렴하면 forgetting이 감소할 것이라 주장합니다.
> 그러나 본 논문은 geomatric intuition을 가지며 stability 뿐만 아니라 plasticity 또한 향상 시킵니다.

> **Secondly,** 수렴하는 방법이 다릅니다.
> Mirzadeh 는 learning rate, mini batch size, dropout 이 세가지를 제어하지만,
> 본 논문의 방법은 wide local minima를 유도하는 정규화로 분류기 projection을 사용했습니다.
> 따라서 제어해야 할 hyperparameter가 하나 뿐이므로 복잡성이 훨씬 낮습니다.

> **Thirdly,** CL을 분석한 반면, 정보 투영 관점에서 CPR의 역할에 대한 원칙적 이론적 해석을 제안합니다.

> **Fourthly,** Mirzadeh는 한정된 benchmark에서 단일 epoch 설정만을 고려한 반면, 우리는 다양한 설정에서 다중 epoch 설정에서 광범위한 실험을 진행합니다.

> **Finally,** 실험 분석에 대한 차이, 우리는 CL을 위한 wide local minima 를 효과에 대해 분석했습니다.

Conclusion

본 논문은 **classifier-projection regularization(CPR)** 을 제안합니다. 이는 regularization-based continual learning과 결합한 형태 입니다.

본 논문의 저자는 wide local minima를 통해 각 task가 수렴하는 것을 입증하였고, CPR은 continual learning에서 plasticity와 stability 에 상당한 향상을 보였습니다.

이러한 wide local minima를 촉진하는 regularizer는 결과는 복돋으며 continua learning에서 성공적인 역할을하였습니다.

이론적 해석으로, 본 논문에서 CPR에서 찾은 추가적인 term이 분류기의 출력에 의한 조건부 확률을 균일 분포를 중심으로하는 공에 투영하는 것으로 이해할 수 있다고 주장합니다.