There Are Many Consistent Explanations of Unlabeled Data: Why You Should Average

School of Industrial and Management Engineering, Korea University

Eun Ji Koh

KOREA UNIVERSITY

DMQA

Contents

Research Purpose

❖ There Are Many Consistent Explanations of Unlabeled Data: Why You Should Average (ICLR, 2019)

- 2022년 08월 17일 기준으로 182회 인용됨

THERE ARE MANY CONSISTENT EXPLANATIONS OF UNLABELED DATA: WHY YOU SHOULD AVERAGE

Ben Athiwaratkun, Marc Finzi, Pavel Izmailov, Andrew Gordon Wilson
{pa338, maf388, pi49, andrew}@cornell.edu
Cornell University

Abstract

Presently the most successful approaches to semi-supervised learning are based on *consistency regularization*, whereby a model is trained to be robust to small perturbations of its inputs and parameters. To understand consistency regularization, we conceptually explore how loss geometry interacts with training procedures. The consistency loss dramatically improves generalization performance over supervised-only training; however, we show that SGD struggles to converge on the consistency loss and continues to make large steps that lead to changes in predictions on the test data. Motivated by these observations, we propose to train consistency-based methods with Stochastic Weight Averaging (SWA), a recent approach which averages weights along the trajectory of SGD with a modified learning rate schedule. We also propose *fast-SWA*, which further accelerates convergence by averaging multiple points within each cycle of a cyclical learning rate schedule. With weight averaging, we achieve the best known semi-supervised results on CIFAR-10 and CIFAR-100, over many different quantities of labeled training data. For example, we achieve 5.0% error on CIFAR-10 with only 4000 labels, compared to the previous best result in the literature of 6.3%.

Research Purpose

❖ There Are Many Consistent Explanations of Unlabeled Data: Why You Should Average (ICLR, 2019)

- Consistency-based semi-supervised learning 방법론 중 π model과 mean-teacher 모델의 training trajectory를 따라 solution을 분석함으로써, SGD을 사용하여 학습하면 단일 solution으로 수렴하지 않는다는 것을 발견
- **Consistency-based methods의 경우 stochastic weight averaging (SWA)로 학습하는 것 제안**
- 주기적인 learning rate schedule의 각 사이클 내에서 multiple points를 averaging 함으로써 수렴을 가속화하는 **fast-SWA 제안**

There Are Many Consistent Explanations of Unlabeled Data: Why You Should Average

❖ Consistency-based methods: π model & mean-teacher

Consistency loss

- Student's predicted probabilities $f(x'; w'_f)$와 teacher's predicted probabilities $g(x''; w'_g)$ 간의 차이를 penalize
- Loss는 MSE 또는 KL divergence로 정의

$$\ell_{\text{cons}}^{\text{MSE}}(w_f, x) = \|f(x'; w'_f) - g(x'', w'_g)\|^2 \text{ or } \ell_{\text{cons}}^{\text{KL}}(w_f, x) = \mathbf{KL}(f(x'; w'_f)||g(x'', w'_g)) .$$

D_L: labeled data / D_U : unlabeled data / x' , x'' : two perturbed input of x / w'_f , w'_g : perturbed weights

- Consistency-based model의 최종 loss는 아래와 같이 supervised loss + consistency loss로 구성

$$L(w_f) = \underbrace{\sum_{(x,y)\in\mathcal{D}_L} \ell_{\text{CE}}(w_f, x, y)}_{L_{\text{CE}}} + \lambda \underbrace{\sum_{x\in\mathcal{D}_L\cup\mathcal{D}_U} \ell_{\text{cons}}(w_f, x)}_{L_{\text{cons}}}$$

Supervised loss — Consistency loss

Control importance of the consistency term in the overall loss

There Are Many Consistent Explanations of Unlabeled Data: Why You Should Average

❖ Understanding consistency-enforcing models

Theoretical study

- 아래와 같은 조건의 small version의 π model을 가정해보면,

Student input: $x' = x + \epsilon z, z \sim N(0, I), \epsilon \ll 1$
Teacher input : $x'' = x$

Consistency loss: $l_{cons}(w, x, \epsilon) = \|f(w, x + \epsilon z) - f(w, x)\|^2$
Estimator $\hat{Q} = \lim_{\epsilon \to 0} \frac{1}{\epsilon^2} \frac{1}{m} \sum_{i=1}^{m} l_{cons}(w, x_i, \epsilon)$

상세설명 논문 section A.5 참고

$$\mathbb{E}[\hat{Q}] = \mathbb{E}_x[\|J_x\|_F^2] \quad \text{and} \quad \text{Var}[\hat{Q}] = \frac{1}{m}\left(\text{Var}[\|J_x\|_F^2] + 2\mathbb{E}[\|J_x^T J_x\|_F^2]\right)$$

J_x 는 network output의 Jacobian
$\|\cdot\|_F$는 Frobenius norm
E_x는 labeled data와 unlabeled data의 distribution

⇨ $\|J_x\|_F$는 이론적으로 generalization과 관련되며,
∴ Consistency loss는 implicitly penalize $E_x[\|J_x\|_F^2]$

- Consistency를 standard data augmentations에 대한 manifold Jacobian norm을 penalizing 하는 것으로 해석 가능
- 결론적으로 consistency loss가 generalization과 관련하여 network output의 Jacobian norm and hessian's eigenvalues를 불리하게 유도

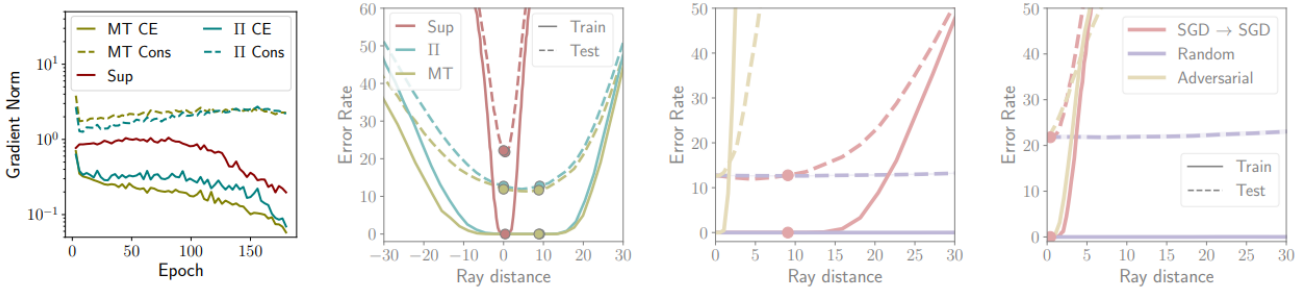
There Are Many Consistent Explanations of Unlabeled Data: Why You Should Average

❖ Understanding consistency-enforcing models

Empirical study

- 실제 setting에서 consistency loss를 최소화하는 특징을 분석하기 위해 SGD가 따르는 training trajectory를 분석하고 supervised learning의 trajectory와 비교

- Fig 1 (a): consistency regularization term(Cons) 과 cross-entropy term(CE)의 gradient norm 변화를 시각화
  - ➢ Cons가 학습이 끝날 때 까지 높게 유지되고 Sup보다 크게 나타남
  - ➢ Consistency-based model들이 학습 시 single minimizer로의 수렴보다 대규모 solution set을 지속적으로 탐색하는 것으로 해석 가능



(a) Gradient Norm (b) SGD-SGD Rays (c) Π model (d) Supervised model

There Are Many Consistent Explanations of Unlabeled Data: Why You Should Average

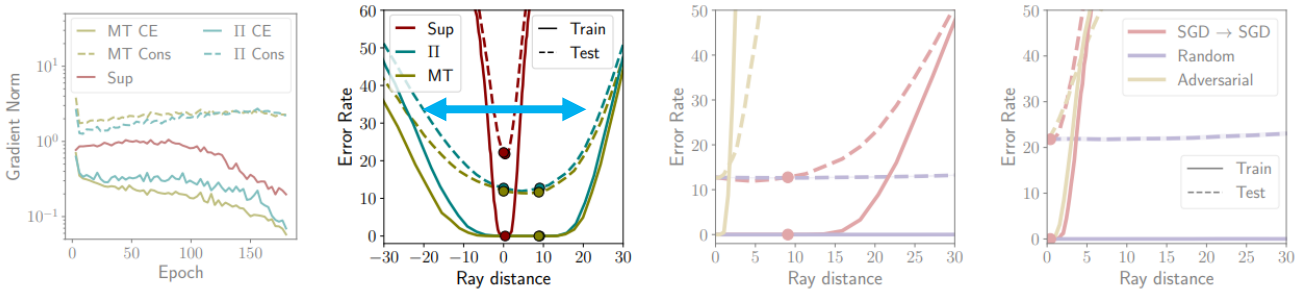
❖ Understanding consistency-enforcing models

Empirical study

- 실제 setting에서 consistency loss를 최소화하는 특징을 분석하기 위해 SGD가 따르는 training trajectory를 분석하고 supervised learning의 trajectory와 비교

- Fig 1 (b): ray distance*에 따른 train 및 test error 시각화
  - 170 epoch, 180 epoch의 weight 벡터의 거리 ◀▶ 는 semi 방법론에서 상대적으로 크게 나타남
  - Consistency-based model들이 Supervised learning보다 학습 시 gradient norm이 큰 것으로 해석 가능

* Ray distance: $\varphi(t) = t \cdot w_{180} + (1-t)w_{170},\ t \geq 0$ (w_n: n epoch의 weight 벡터)



(a) Gradient Norm (b) SGD-SGD Rays (c) Π model (d) Supervised model

There Are Many Consistent Explanations of Unlabeled Data: Why You Should Average

❖ Understanding consistency-enforcing models

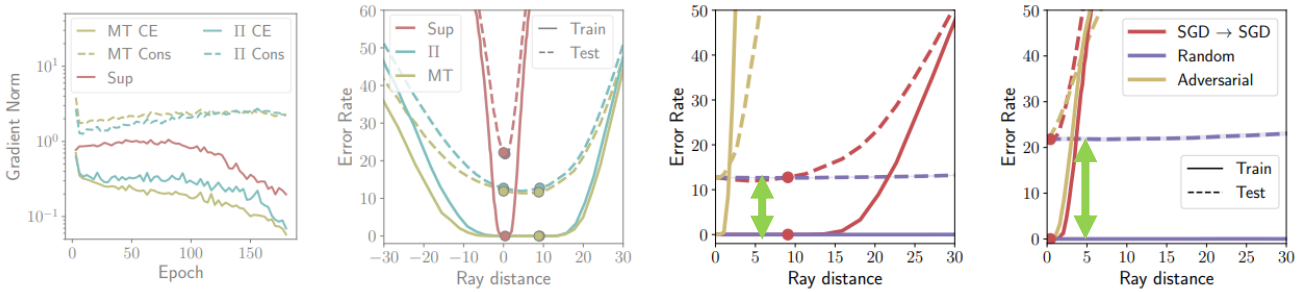
Empirical study

- 실제 setting에서 consistency loss를 최소화하는 특징을 분석하기 위해 SGD가 따르는 training trajectory를 분석하고 supervised learning의 trajectory와 비교

- Fig 1 (c), (d): 다양하게 정의된 ray distance*에 따른 error rate 시각화
  - ➢ SGD->SGD의 train, test 사이 폭과 random, adversarial의 폭 ↕ 이 유사함

    * SGD -> SGD ray: Fig 1 (b)와 동일
    * Random ray: unit sphere로부터 5개의 random 벡터 d를 sampling 하고 weight $w_{t_i} + sd, (s \in [0,30])$를 갖는 network의 train 및 test error를 평균 냄
    * Adversarial ray: train 및 test loss $d_{adv} = {}^{\nabla L_{CE}}/_{\|\nabla L_{CE}\|}$의 가장 빠른 상승 방향을 따라 error를 평가한다.

- Fig 1을 고려하면 last SGD iterate을 prediction을 사용하는 것은 바람직하지 못함



(a) Gradient Norm (b) SGD-SGD Rays (c) Π model (d) Supervised model

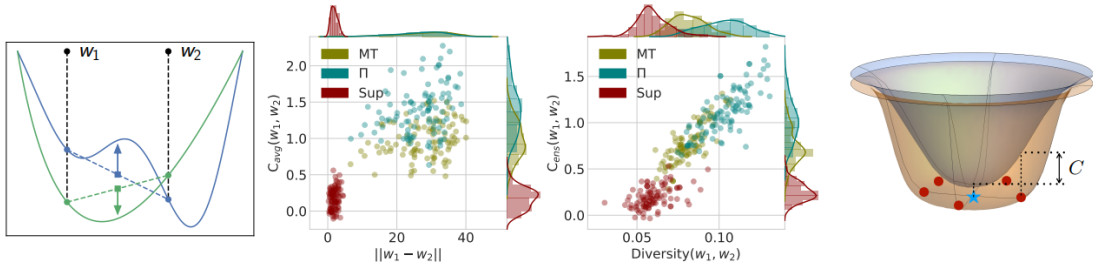
DMQA

There Are Many Consistent Explanations of Unlabeled Data: Why You Should Average

❖ Ensembling and Weight averaging

Ensembling

- Supervised 방법론에 비해 π model과 mean-teacher는 prediction diversity가 큰 것을 확인했기 때문에 ensemble을 통해 이점을 얻을 수 있음
- Fig 2 (c): C_{ens}(학습 후반 epoch 중 임의로 선정한 2개 epoch의 weigh를 ensemble 한 것)으로부터 측정한 error reduction
  - ➢ Prediction의 diversity와 ensemble performance 간의 상관성이 큼
  - ➢ Sup보다 π model과 mean-teacher에서 C_{ens} 가 크게 나타남



(a) Convex vs. non-convex (b) C_{avg} vs. $||w_1 - w_2||$ (c) C_{ens} vs. diversity (d) Gain from flatness

Figure 2: **(a)**: Illustration of a convex and non-convex function and Jensen's inequality. **(b)**: Scatter plot of the decrease in error C_{avg} for weight averaging versus distance. **(c)**: Scatter plot of the decrease in error C_{ens} for prediction ensembling versus diversity. **(d)**: Train error surface (orange) and Test error surface (blue). The SGD solutions (red dots) around a locally flat minimum are far apart due to the flatness of the train surface (see Figure 1b) which leads to large error reduction of the SWA solution (blue dot).

There Are Many Consistent Explanations of Unlabeled Data: Why You Should Average

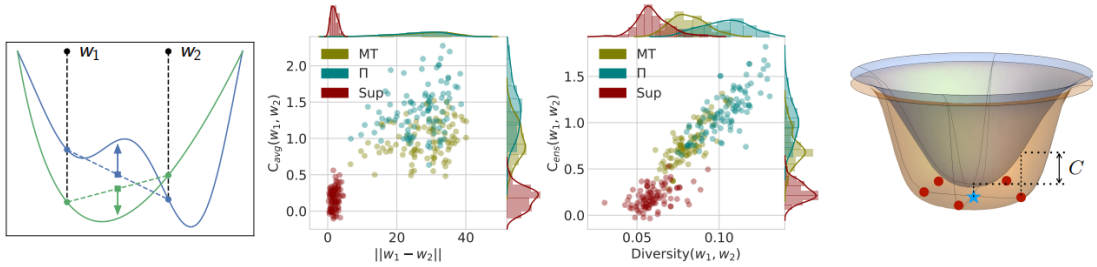
❖ Ensembling and Weight averaging

Weight Averaging

- Weight 간 유사할 경우 weight averaging에 의해 잘 근사 될 수 있음
- Fig 2 (b): c가 대부분 양수이기 때문에 SGD trajectory의 error surface가 convex 함

  Training 과정에서의 weight 간의 거리가 Sup 보다 π model 및 mean-teacher에서 크게 나타남

  ➢ 따라서 weight averaging은 p model 및 mean-teacher에서 더욱 유의미한 효과를 줄 수 있음



(a) Convex vs. non-convex (b) C_{avg} vs. $||w_1 - w_2||$ (c) C_{ens} vs. diversity (d) Gain from flatness

Figure 2: **(a)**: Illustration of a convex and non-convex function and Jensen's inequality. **(b)**: Scatter plot of the decrease in error C_{avg} for weight averaging versus distance. **(c)**: Scatter plot of the decrease in error C_{ens} for prediction ensembling versus diversity. **(d)**: Train error surface (orange) and Test error surface (blue). The SGD solutions (red dots) around a locally flat minimum are far apart due to the flatness of the train surface (see Figure 1b) which leads to large error reduction of the SWA solution (blue dot).

There Are Many Consistent Explanations of Unlabeled Data: Why You Should Average

❖ SWA & Fast-SWA

SWA (Stochastic Weight Averaging)

- Modified learning rate이 적용된 SGD가 traverse한 weight를 averaging하는 것을 기반으로 하는 approach
- Multiple weight를 averaging 하는 것은 test 정확도를 향상시킬 수 있음
- SWA는 supervised learning에서 generalization 성능을 향상시킴
  - ➢ 따라서 consistency-based semi-supervised learning에서도 generalization의 개선을 기대할 수 있음
- SWA는 일반적으로 pre-trained model에서 시작하고, weight space의 point를 주기적/일정 속도로 averaging 함
- SWA는 learning rate가 minimum values일 때의 network들의 weight를 평균 냄

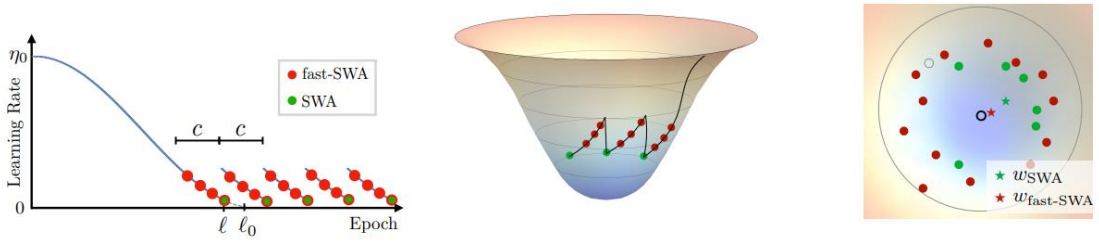


Figure 3: **Left**: Cyclical cosine learning rate schedule and SWA and fast-SWA averaging strategies. **Middle**: Illustration of the solutions explored by the cyclical cosine annealing schedule on an error surface. **Right**: Illustration of SWA and fast-SWA averaging strategies. fast-SWA averages more points but the errors of the averaged points, as indicated by the heat color, are higher.

There Are Many Consistent Explanations of Unlabeled Data: Why You Should Average

❖ SWA & Fast-SWA

Fast-SWA

- 일반적으로 SWA는 사이클 당 average weight를 1회 업데이트하기 때문에 averaging을 위한 충분한 weight를 얻기 위해서는 추가적인 training epoch이 필요하다는 한계가 있음
- 위의 한계를 극복하기 위해 본 논문은 Fast-SWA 제안
- A modification of SWA that averages networks corresponding to every k<c epochs starting from epoch l-c
  - ➢ c: length of cycle
  - ➢ Average multiple weights are obtained within a single epoch setting k<1

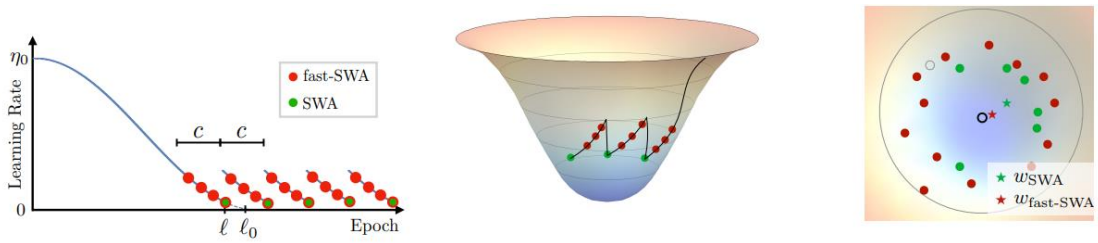


Figure 3: **Left**: Cyclical cosine learning rate schedule and SWA and fast-SWA averaging strategies. **Middle**: Illustration of the solutions explored by the cyclical cosine annealing schedule on an error surface. **Right**: Illustration of SWA and fast-SWA averaging strategies. fast-SWA averages more points but the errors of the averaged points, as indicated by the heat color, are higher.

Experiments

❖ CIFAR 데이터셋에 대한 성능 모델 평가

- Fig 4: Labeled data의 비율과 무관하게 fast-SWA는 성능 향상을 이룸
- Fig 5: SWA보다 fast-SWA가 상대적으로 빠르게 수렴
- Table 1: SWA, fast-SWA를 π model 및 mean-teacher와 함께 사용하면 성능을 향상시킬 수 있음

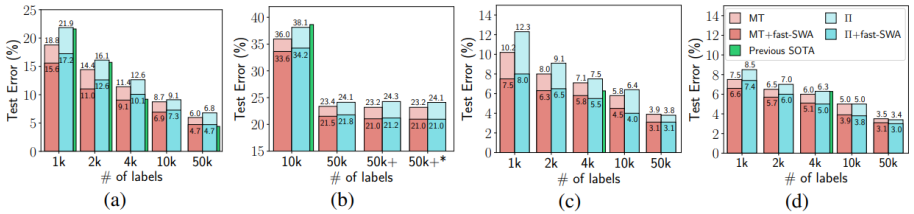


Figure 4: Prediction errors of Π and MT models with and without fast-SWA. **(a)** CIFAR-10 with CNN **(b)** CIFAR-100 with CNN. $50k+$ and $50k+^*$ correspond to $50k$+$500k$ and $50k$+$237k^*$ settings **(c)** CIFAR-10 with ResNet + Shake-Shake using the short schedule **(d)** CIFAR-10 with ResNet + Shake-Shake using the long schedule.

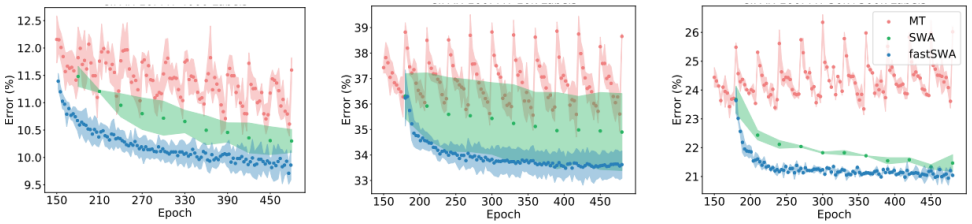


Figure 5: Prediction errors of base models and their weight averages (fast-SWA and SWA) for CNN on **(left)** CIFAR-10 with $4k$ labels, **(middle)** CIFAR-100 with $10k$ labels, and **(right)** CIFAR-100 $50k$ labels and extra $500k$ unlabeled data from Tiny Images (Torralba et al., 2008).

Table 1: Test errors against current state-of-the-art semi-supervised results. The previous best numbers are obtained from (Tarvainen and Valpola, 2017)[1], (Park et al., 2017)[2], (Laine and Aila, 2016)[3] and (Luo et al., 2018)[4]. CNN denotes performance on the benchmark 13-layer CNN (see A.8). Rows marked † use the Shake-Shake architecture. The result marked ‡ are from Π + fast-SWA, where the rest are based on MT + fast-SWA. The settings $50k$+$500k$ and $50k$+$237k^*$ use additional $500k$ and $237k$ unlabeled data from the Tiny Images dataset (Torralba et al., 2008) where * denotes that we use only the images that correspond to CIFAR-100 classes.

Dataset	CIFAR-10			CIFAR-100		
No. of Images	50k	50k	50k	50k	50k+500k	50k+237k*
No. of Labels	1k	2k	4k	10k	50k	50k
Previous Best CNN	18.41[4]	13.64[4]	9.22[2]	38.65[3]	23.62[3]	23.79[3]
Ours CNN	15.58	11.02	9.05	33.62	21.04	20.98
Previous Best†			6.28[1]			
Ours†	6.6	5.7	5.0‡	28.0	19.3	17.7

Conclusion

❖ conclusion

- π model과 mean-teacher의 training trajectory를 따라 solution을 분석함으로써 SGD가 단일 solution으로 수렴하는 것이 아닌 training 후반부에도 다양한 solution set을 탐색하는 것을 발견함
- 이를 통해 SWA의 사용 및 fast-SWA를 제안함으로써 consistency-based model성능을 향상시킴

Reference

1. Athiwaratkun, B., Finzi, M., Izmailov, P., & Wilson, A. G. (2018). There are many consistent explanations of unlabeled data: Why you should average. arXiv preprint arXiv:1806.05594.


Thank You