Realistic Evaluation of Deep Semi-Supervised Learning Algorithms

School of Industrial and Management Engineering, Korea University

Go Eun Chae

KOREA UNIVERSITY

DMQA hcai Human-Centered Artificial Intelligence Center

Contents

Research Purpose

❖ **Realistic Evaluation of Deep Semi-Supervised Learning Algorithms (2018)**

- Google Brain 연구, 814회 인용 (2022.09.19 기준)
- **Semi-Supervised Learning (SSL)** 의 **'Real-World' 데이터 적용**에 대한 실험 방법론 제안

Realistic Evaluation of Deep Semi-Supervised Learning Algorithms

Avital Oliver*, Augustus Odena*, Colin Raffel*, Ekin D. Cubuk & Ian J. Goodfellow
Google Brain
{avitalo,augustusodena,craffel,cubuk,goodfellow}@google.com

Abstract

Semi-supervised learning (SSL) provides a powerful framework for leveraging unlabeled data when labels are limited or expensive to obtain. SSL algorithms based on deep neural networks have recently proven successful on standard benchmark tasks. However, we argue that these benchmarks fail to address many issues that SSL algorithms would face in real-world applications. After creating a unified reimplementation of various widely-used SSL techniques, we test them in a suite of experiments designed to address these issues. We find that the performance of simple baselines which do not use unlabeled data is often underreported, SSL methods differ in sensitivity to the amount of labeled and unlabeled data, and performance can degrade substantially when the unlabeled dataset contains out-of-distribution examples. To help guide SSL research towards real-world applicability, we make our unified reimplemention and evaluation platform publicly available.[2]

Research Purpose

❖ **Supervised Learning**

- 주어진 Input-Target Pairs $(x, y) \in D$ 에 대해 Unknown Joint Distribution $p(x, y)$ 구축
- 목표: **Unseen Samples 에 대해 예측하는 함수 $f_\theta(x)$ 찾는 것**
- 대용량의 Labeled Dataset 구축을 위한 자원 부족

❖ **Semi-Supervised Learning (SSL)**

- Unlabeled Samples 로 데이터의 구조를 학습
- Unlabeled Datapoints $x \in D_{UL}$ 추가로 주어짐
- 목표: **Unlabeled Samples 사용하여 $p(x)$ 구조 학습 후 $f_\theta(x)$ 보완**

'Real-World' 설정에서 SSL 방법론을 적용하기 위한 실험 방법론 제안

Proposed Method

❖ **Semi-Supervised Unlabeled 데이터 활용 방식에 따라 아래와 같이 나뉨**

❖ **Consistency Regularization**

- 데이터의 현실적인 작은 변동($x \rightarrow \hat{x}$)이 예측함수의 결과를 많이 변화시키지 않는 것
- Unlabeled 데이터 활용하여 데이터셋이 놓여있는 Smooth Manifold 탐색
- 예측 함수의 Output 과 실제 값의 차이$d\left(f_\theta(x), f_\theta(\hat{x})\right)$최소화 ex) MSE, Kullback-Leibler divergence

❖ **Entropy-Based**

- Unlabeled Data 에 대해 더 정확한 예측 수행을 위해 Loss Term 추가

❖ **Pseudo-Labeling**

- 학습 과정에서 예측함수 이용해 Unlabeled 데이터에 대한 Pseudo-Labels 생성

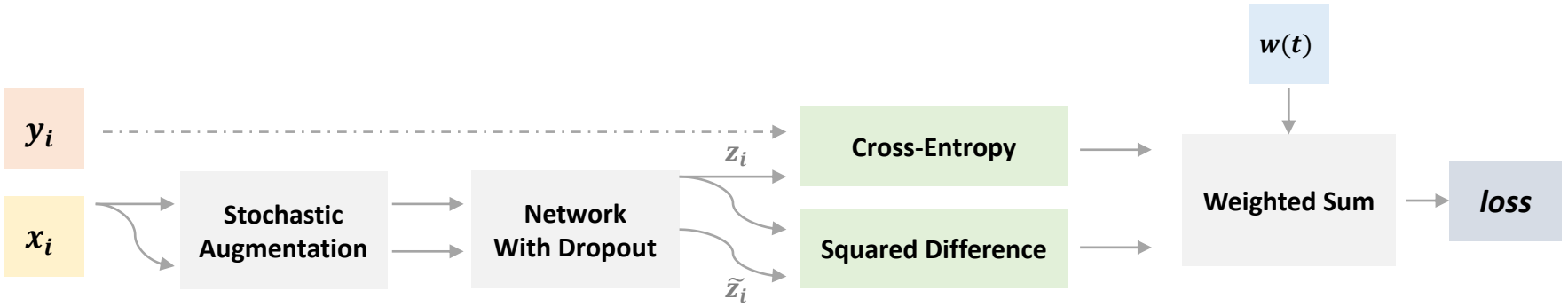
Proposed Method

Consistency Regularization

❖ Π-Model (Stochastic Perturbations)

- Data Augmentation, Dropout, Adding Noise 등의 기법을 통해 같은 Input 에 대해 다른 Output 도출
  - ✓ Loss = 두 값의 차이를 이용한 Loss (MSE) + Output과 Label 이용한 Loss (Cross-Entropy)
- $d\left(f_\theta(x), f_\theta(\hat{x})\right)$ 최소화 부분을 Loss 에 추가하여 Regularizer 로 활용
  - ✓ Pseudo-Ensembles , Regularization With Stochastic Transformations and Perturbations 로 불림

< Π-Model >



Proposed Method

Consistency Regularization

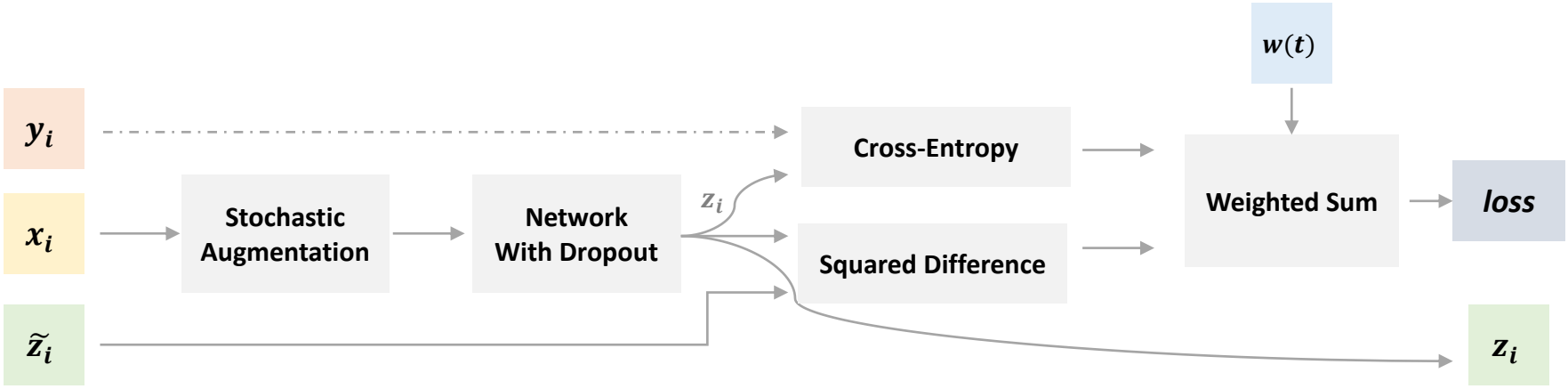
❖ Temporal Ensembling

- Π-Model이 예측의 잠재적 불확실성에 의존 & 심한 노이즈 개선
- 더 안정된 Output 구축을 위해 이전 모델 파라미터의 Exponential Moving Average 사용
- 과거 모델의 결과와 현재 모델 결과를 Ensemble 하여 Prediction Vector 생성

❖ Mean Teacher

- Temporal Ensembling 에서 Teacher 모델과 Student 모델 분리
- 모델의 Weight 에 대해 Weighted Average 하는 방법론

< Temporal Ensembling>



Proposed Method

Consistency Regularization

❖ Virtual Adversarial Training (VAT)

- 예측함수의 내적 Stochasticity 에 의존하는 것이 아닌 작은 변동을 바로 근사
  - ✓ 작은 변동 (tiny perturbation) : γ_{adv}
- Input x에 γ_{adv} 추가하여 예측함수의 Output 에 가장 큰 영향 줌
- θ에 대해 $d\left(f_\theta(x), f_\theta(x+\gamma_{adv})\right)$ 최소화 하기 위해 Consistency Regularization 사용

Entropy-Based

❖ Entropy Minimization (EntMin)

- Unlabeled Data 에 대한 모델 방향 제시
- *Entropy Minimization term*: $-\sum_{k=1}^{K} f_\theta(x)_k \log f_\theta(x)_k$
- EntMin 자체로는 성능 향상 미비하지만 VAT 와 결합하여 SOTA 달성

Experiments

❖ Reproduction

- 분석 Datasets: 41,000 UnLabeled CIFAR-10 & 64,932 Unlabeled SVHN Datasets
- Wide ResNet 을 표준 모델로 하여 실험 진행
  - ✓ Depth 28, Width 2 인 WRN-28-2 사용 (Batch Normalization 포함)
- Gaussian Process-Based Black Box Optimization 1000회 시행
  - ✓ 데이터셋과 모델 간에 불필요하게 다른 Hyperparameter 생성 가능
  - ✓ 따라서 각 데이터셋, 모델 조합에 최적인 통일된 Hyperparameter 사용
- Hyperparameter 설정에 대한 Validation Error가 가장 낮을 때의 Test Error

<Various SSL Approaches>

Dataset	# Labels	Supervised	Π-Model	Mean Teacher	VAT	VAT + EntMin	Pseudo-Label
CIFAR-10	4000	20.26 ± .38%	16.37 ± .63%	15.87 ± .28%	13.86 ± .27%	13.13 ± .39%	17.78 ± .57%
SVHN	1000	12.83 ± .47%	7.19 ± .27%	5.65 ± .47%	5.63 ± .20%	5.35 ± .19%	7.62 ± .29%

4000 Labeled CIFAR-10 & 1000 Labeled SVHN 사용

Experiments

❖ Fully-Supervised Baselines

- Fully-Supervised Learning 과 SSL의 Error Rate 나타냄
  - ✓ Regularization Capabilities 가 큰 Shake-Shake 모델 사용
  - ✓ Standard Data-Augmentation 방법 사용
- **Fully-Supervised Learning 과 SSL 의 차이**가 일반적인 연구에 비해 작은 것 확인
- SSL 알고리즘 간 비교에서 **기본 모델의 중요성** 강조
- Conflating Comparison 피하기 위해 동일한 모델을 사용하여 서로 다른 SSL 알고리즘 평가

<Fully-Supervised vs SSL>

Method	CIFAR-10 4000 Labels	SVHN 1000 Labels
Π-Model [32]	34.85% → 12.36%	19.30% → 4.80%
Π-Model [46]	13.60% → 11.29%	–
Π-Model (ours)	20.26% → 16.37%	12.83% → 7.19%
Mean Teacher [50]	20.66% → 12.31%	12.32% → 3.95%
Mean Teacher (ours)	20.26% → 15.87%	12.83% → 5.65%

Experiments

❖ Transfer Learning

- 사전 학습한 분류기를 활용하여 Transfer Learning 평가
  - ✓ 32x32 ImageNet 으로 Standard WRN-28-2 모델 학습
  - ✓ CIFAR-10의 4000 Labeled Datasets 으로 Fine-Tuning 실행
- 동일 모델을 사용한 어떤 SSL 보다 낮은 Error Rate (12.09%) 가짐
- **Labeled Data 가 Transfer Learning 에 적합할 경우 SSL 이 좋은 대안**

<SSL vs Transfer Learning>

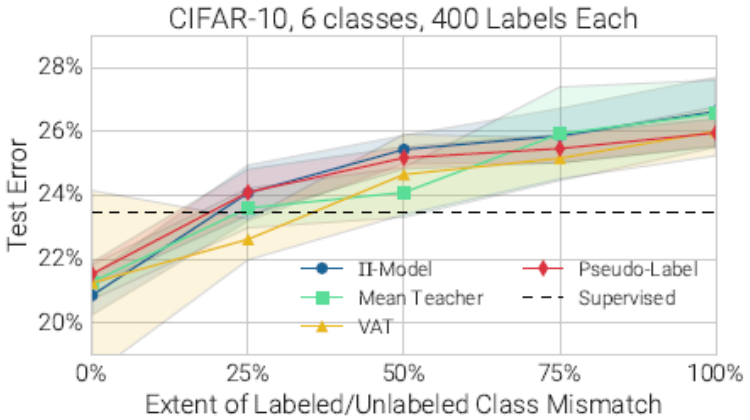
Method	CIFAR-10 4000 Labels
VAT with Entropy Minimization	13.13%
ImageNet → CIFAR-10	12.09%
ImageNet → CIFAR-10 (no overlap)	12.91%

Experiments

❖ Class Distribution Mismatch

- Unlabeled 데이터의 Class 가 Labeled 데이터에 없는 경우 평가
  - ✓ Labeled, Unlabeled 데이터가 같은 분포를 가짐
  - ✓ CIFAR-10 의 6-Class 분류 실행, Unlabeled 데이터는 나머지 4 Classes 에서 추출
- **성능: Unlabeled 데이터 사용 안함 > Mismatched Unlabeled 데이터 추가**
  - ✓ Labeled 데이터와 다른 Class 를 가지는 Unlabeled 데이터 추가로 인한 성능 하락
  - ✓ Unlabeled 데이터가 주요 Task 와 관련 없는 경우, Labeled 데이터 얻는 데에 집중 필요

<Class Distribution Match>

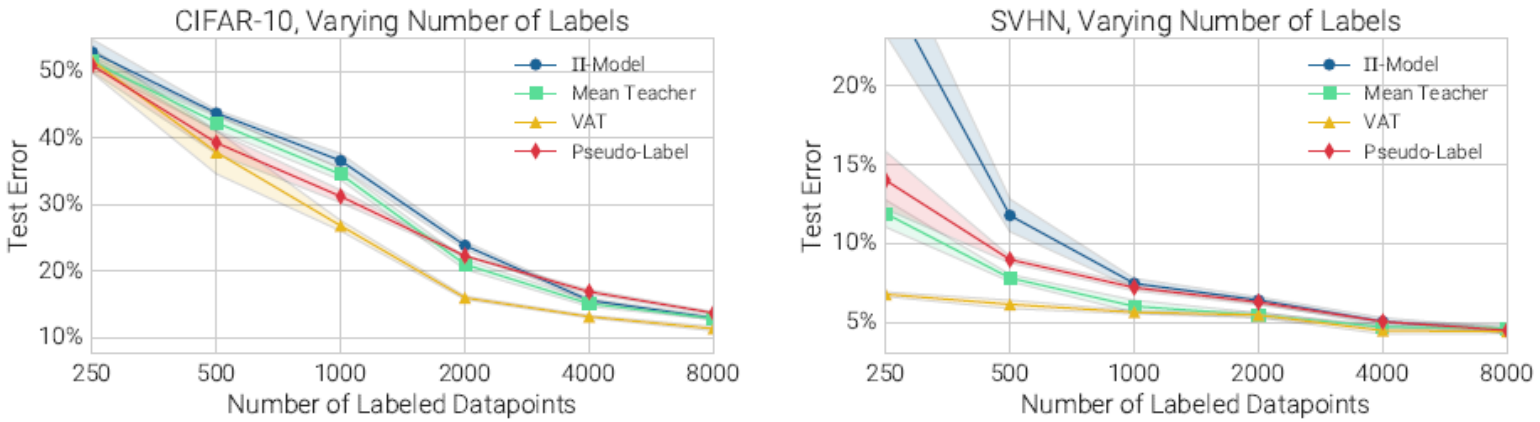


Experiments

❖ Varying Data Amounts

- 전체 데이터에서 Labeled 데이터 비율에 따른 성능 측정
- **Labeled 데이터가 증가할 수록 SSL 성능 수렴**
- 데이터 양에 대한 민감도가 SSL 알고리즘에 따라 다양함 확인

<Varying Data Amounts>



Experiments

❖ Small Validation Sets

- Hoeffding's 부등식을 통해 SSL의 성능 구별에 필요한 Validation Set의 이론적 추정치 도출
  - ✓ Validation 정확도가 독립적인 변수들의 평균이라는 가정으로 인해 비현실적
- Figure A: 무작위 추출한 10개 샘플로 이루어진 Validation Set 에 대한 평균, 분산
- Figure B: SSL model vs Π-Model 의 Validation Error 차이
- **Validation Set 이 Training Set 의 10% (현실 데이터) 일 경우 모델 성능 구별 불가**
  - ✓ Validation Set 의 크기가 Training Set 과 동일할 경우 방법론 간의 평가 가능

<Small Validation Sets>

Figure A

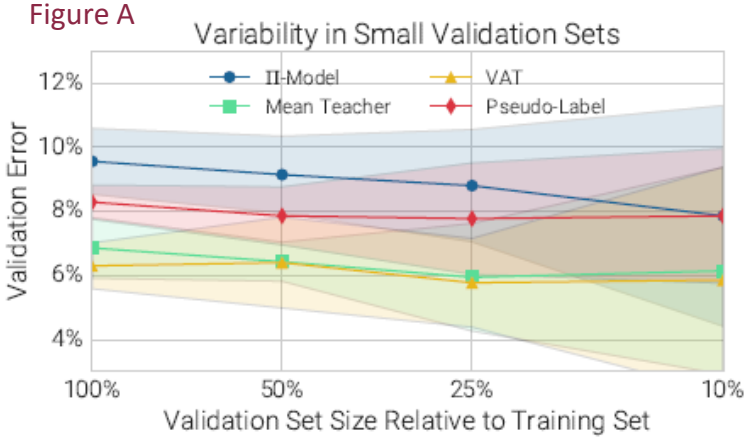
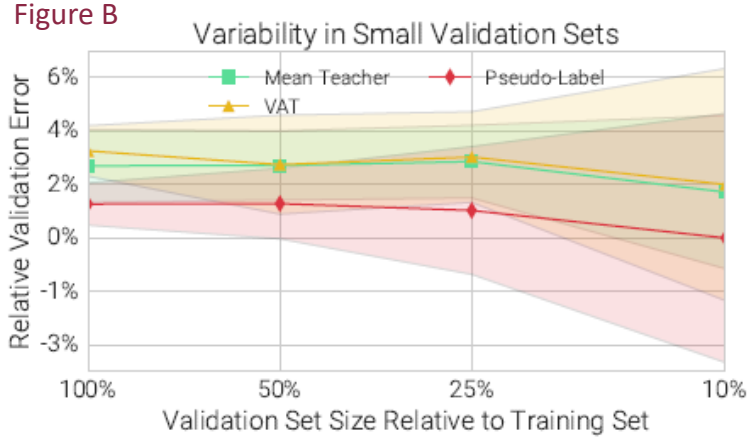


Figure B



Conclusion

❖ Conclusion

- 현재까지의 SSL 평가 방식은 Unrealistic
- **'Real-World' 데이터에 적합한 SSL 평가 방식 제안**
- **표준 Framework 에서 엄격한 SSL 분석 제공**

❖ Recommendation

- SSL 알고리즘 평가 시 **동일한 모델** 적용
- **잘 조정된 Fully-Supervised 성능**으로 판단 후 Transfer Learning 진행
- Labeled, Unlabeled 데이터의 **Class 분포 불일치 파악** 필요
- **Labeled, Unlabeled 데이터 양 모두 변경**하며 SSL 성능 측정
- 비현실적으로 큰 데이터에 Hyperparameter 과도하게 사용하지 않도록 주의


Thank You