

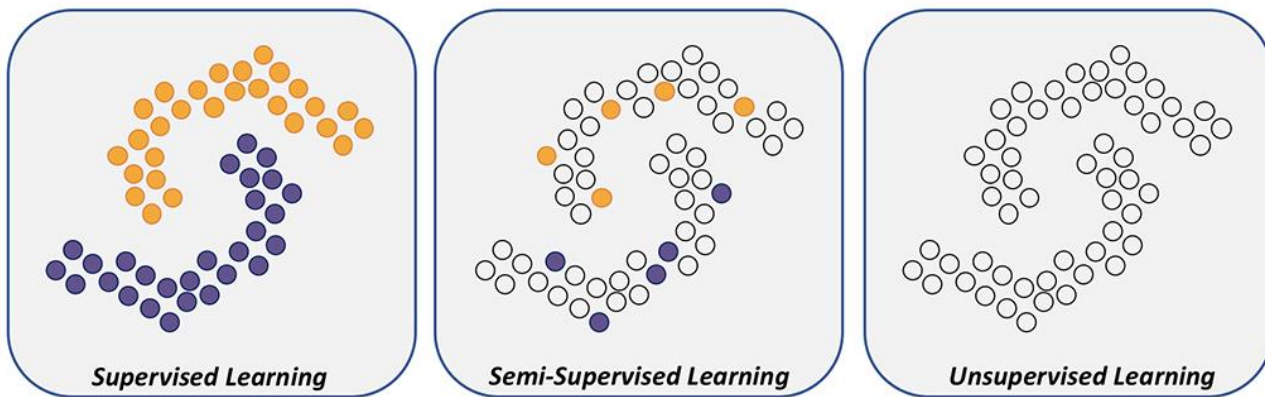


MixMatch: A Holistic Approach to Semi-Supervised Learning

22510108 이성호

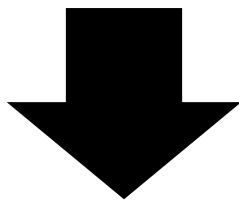
Semi-supervised learning

- Labeled data가 많지 않은 상황에서 unlabeled data를 사용하는 방법
- 즉, 준지도학습의 목표는 unlabeled data를 이용하여 지도학습의 성능을 더 끌어올리는 것



Semi-supervised learning 필요성

- 최근 다양한 필드에서 딥러닝을 활용
- 즉, 딥러닝이 다루는 문제가 복잡해지고 응용 분야가 다양해지면서 라벨링 작업 자체가 하나의 문제(전문 지식 필요)



- 라벨링된 데이터가 적을 때 레이블이 없는 데이터를 사용해 분류기의 성능을 향상시키자!



Semi-supervised learning 목적함수

- Supervised loss 와 unsupervised loss의 합을 최소화 하는 것이 목적
- 즉, supervised, unsupervised를 1-stage로 학습
- Unsupervised loss는 본 적 없는 데이터에 대해 일반화를 주로 시키기 위해 사용
- Unlabeled data에 주는 unsupervised task를 어떻게 정할 것이냐에 따라 방법론이 결정됨

$$Loss = L_{supervised} + L_{unsupervised}$$



Semi-supervised learning 가정

1. The smoothness assumption

- 만약 데이터 포인트 x_1 과 x_2 가 고밀도 지역에서 가까이 위치한다면, 해당하는 출력 y_1 과 y_2 도 가깝다.
- 역으로 두 데이터포인트가 저밀도 지역에서 멀리 위치한다면, 해당하는 출력도 역시 멀다

2. The cluster assumption

- 만약에 데이터 포인트들이 같은 cluster에 있다면, 그들은 같은 class일 것이다.

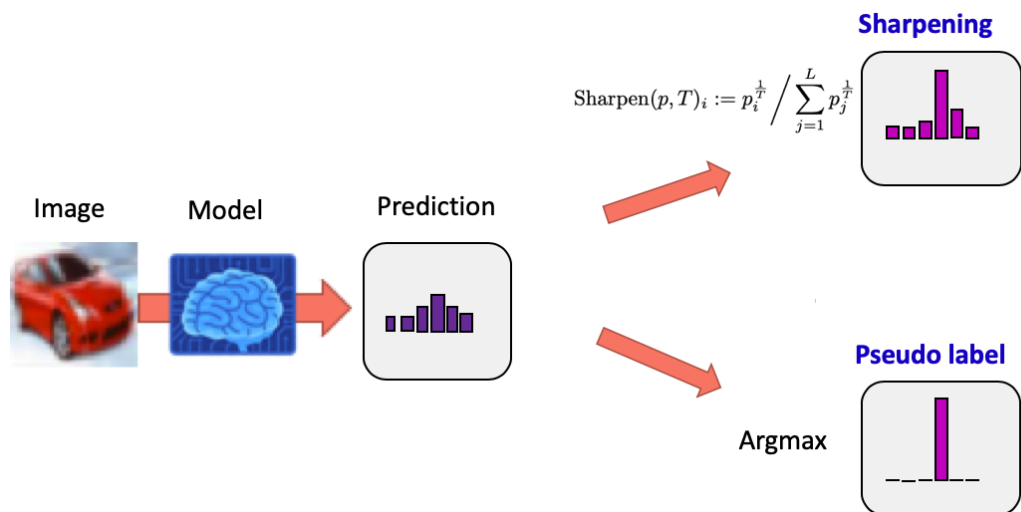
3. The manifold assumption

- 고차원의 데이터를 저차원 manifold로 보낼 수 있다.
- data를 더 낮은 차원으로 보낼 수 있다면 우리는 unlabeled data를 사용해서 저차원 표현을 얻을 수 있고 labeled data를 사용해 더 간단한 task를 풀 수 있다.

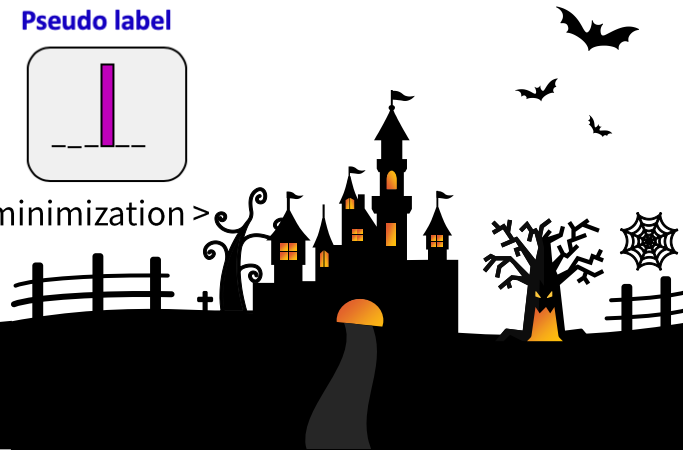


Semi-supervised learning 방법론 - Entropy minimization

- Entropy minimization은 'decision boundary는 데이터의 저밀도 지역에서 형성될 것'이라는 가정에 기초
- Unlabeled data의 예측 값에 대한 confidence를 높이는 것이 목적
- Decision boundary 근처에 있는 애매한 애들을 확실하게 구별할 수 있도록 학습을 진행하고자 함(Ex. 개 0.6 – 고양이 0.4 < 개 0.9 – 고양이 0.1)



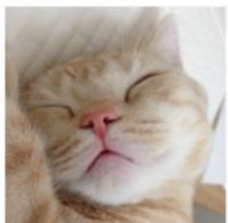
< Temperature sharpening과 pseudo label을 통한 entropy minimization >



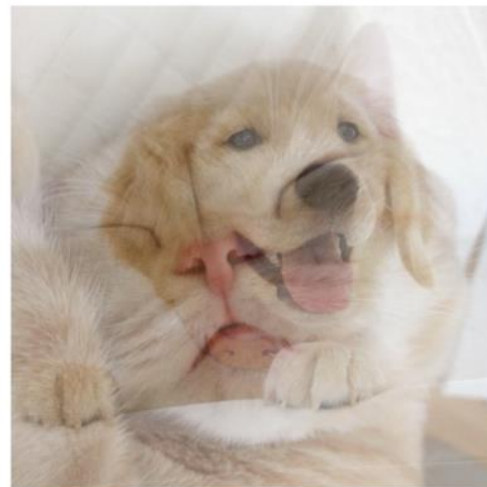
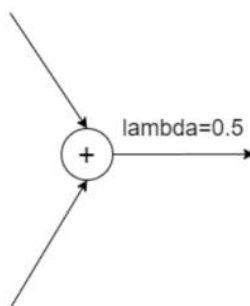
Semi-supervised learning 방법론 - MixUp

- Supervised : 데이터와 레이블 각각을 convex combination을 통해 새로운 데이터를 생성하여 unseen data에 대해 잘 적응하도록 하며 과적합 방지
- 반면 semi supervised는 모델이 unlabeled data에 대해 생성한 가짜 label 사용

Unlabeled [0.9, 0.1]



Unlabeled [0.2, 0.8]



[0.55, 0.45]



MixMatch Overview

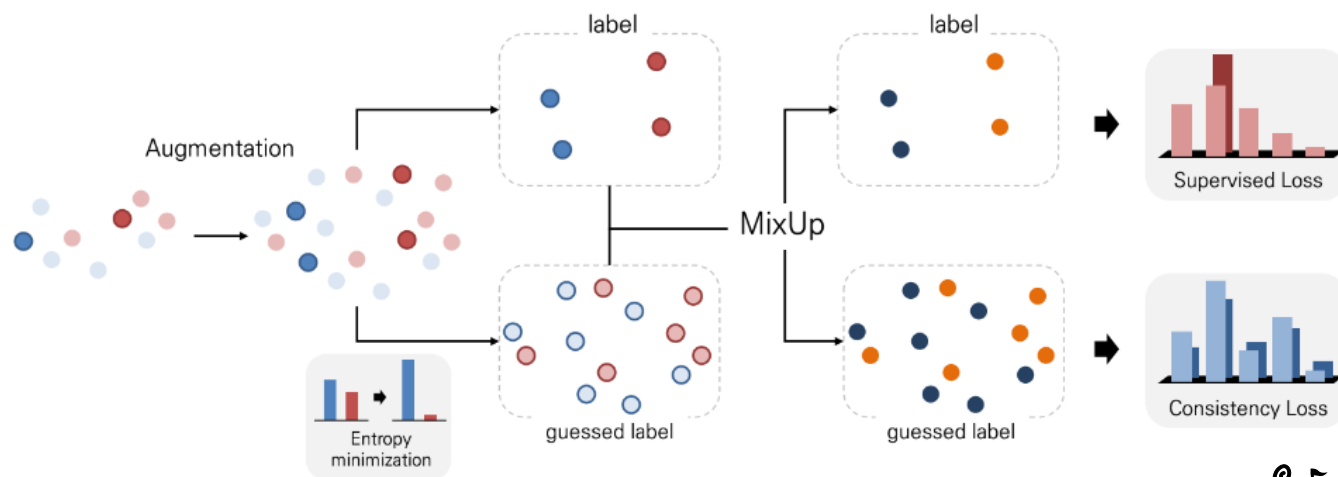
- 앞서 설명한 semi-supervised learning 기법들을 통합하고 Mixup data augmentation을 적용
- 주어진 labeled data batch와 unlabeled data batch로부터 새로운 labeled data batch와 pseudo-labeled data batch 생성

$$\mathcal{X}', \mathcal{U}' = \text{MixMatch}(\mathcal{X}, \mathcal{U}, T, K, \alpha)$$

\mathcal{X} : Labeled Examples

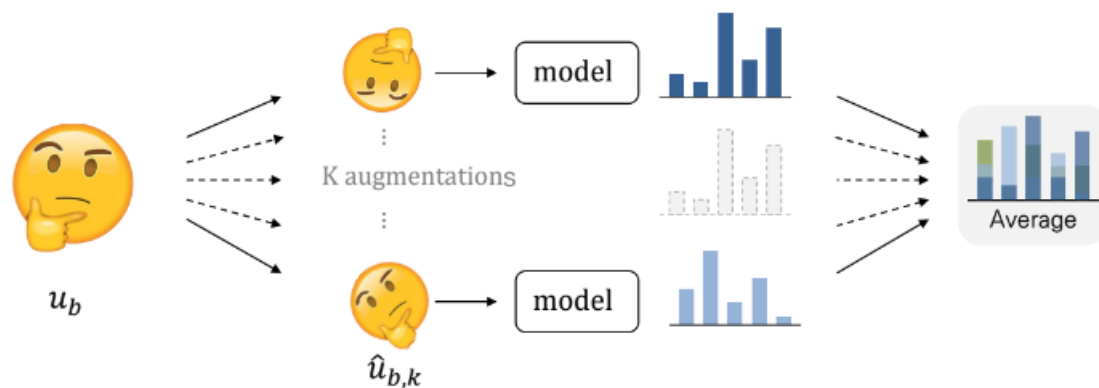
\mathcal{U} : Unlabeled Examples

T, K, α : Hyperparameters



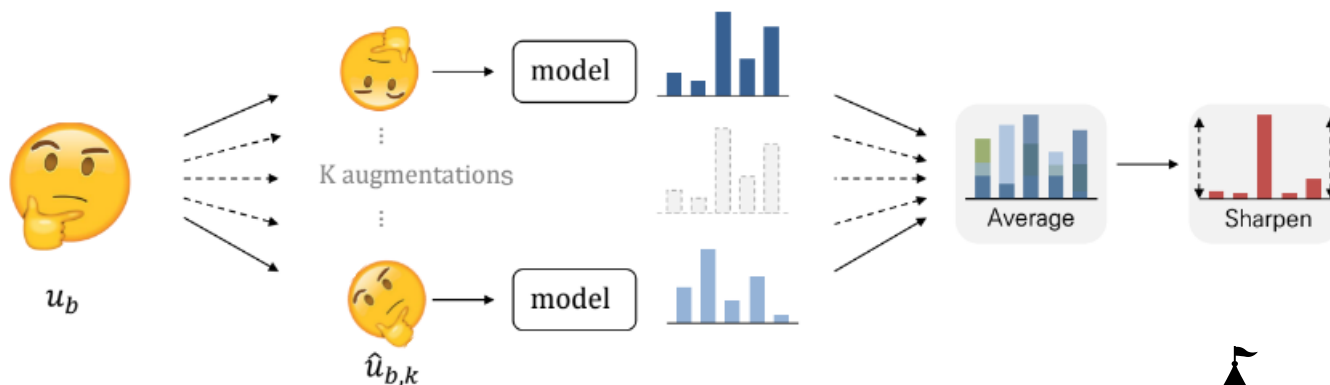
MixMatch - Data Augmentation / Label Guessing

- (Un)label 데이터를 각각 증강 $\hat{x}_b = \text{Augment}(x_b) \hat{u}_{b,k} = \text{Augment}(u_b)$
- 위에서 augmented된 data를 이용해 분류, 예측
- 예측한 클래스 레이블을 guessed label을 q_b 를 구하고 평균 $\bar{q}_b = \frac{1}{K} \sum_{k=1}^K p_{\text{model}}(y \mid \hat{u}_{b,k}; \theta)$



MixMatch - Entropy Minimization(Sharpening)

- 딥러닝 모델은 Overconfidence한 경향을 보임
- unlabeled data로도 분류를 잘 하기 위해 하나의 예측된 클래스의 확률을 가장 높이고 나머지 다른 클래스에 대한 확률을 줄임으로써 예측에 대한 불확실성, 즉 entropy를 줄임
- Softmax Temperature를 이용해 Entropy Minimization 진행 $\text{Sharpen}(\bar{q}_b, T)_i := \frac{\bar{q}_{b,i}^{\frac{1}{T}}}{\sum_{j=1}^L \bar{q}_{b,j}^{\frac{1}{T}}}$



MixMatch - MixUp

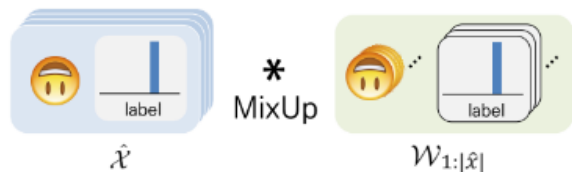
- 앞선 labeled data와 unlabeled data를 종류 관계 없이 랜덤으로 MixUp을 적용

$$\hat{\mathcal{X}} = ((\hat{x}_b), p_b); b \in (1, \dots, B))$$

$$\hat{\mathcal{U}} = ((\hat{u}_{b,k}), q_b); b \in (1, \dots, B))$$

$$\mathcal{W} = \text{Shuffle}(\text{Concat}(\hat{\mathcal{X}}, \hat{\mathcal{U}}))$$

- 섞은 set 에 대해 MixUp을 계산



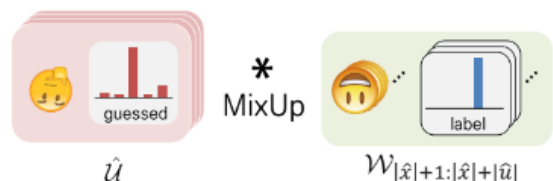
The diagram illustrates the MixUp operation. On the left, a stack of blue cards labeled $\hat{\mathcal{X}}$ (containing a house icon and a bar chart) is multiplied by a stack of green cards labeled $\mathcal{W}_{1:|\hat{\mathcal{X}}|}$ (containing a house icon and a bar chart). The result is shown as a stack of cards with the following equations:

$$x' = \lambda' x_1 + (1 - \lambda') x_2$$

$$p' = \lambda' p_1 + (1 - \lambda') p_2$$

To the right of the equations, the formula $\mathcal{X}' = \text{MixUp}(\hat{\mathcal{X}}_i, \mathcal{W}_i)$ is displayed.

- augmented된 unlabeled data 남은 W 에 대해 MixUp을 계산



The diagram illustrates the MixUp operation between augmented unlabeled data. On the left, a stack of red cards labeled $\hat{\mathcal{U}}$ (containing a sad face icon and a bar chart) is multiplied by a stack of green cards labeled $\mathcal{W}_{|\hat{\mathcal{X}}|+1:|\hat{\mathcal{X}}|+|\hat{\mathcal{U}}|}$ (containing a house icon and a bar chart). The result is shown as a stack of cards with the following equations:

$$x' = \lambda' x_1 + (1 - \lambda') x_2$$

$$p' = \lambda' p_1 + (1 - \lambda') p_2$$

To the right of the equations, the formula $\mathcal{U}' = \text{MixUp}(\hat{\mathcal{U}}_i, \mathcal{W}_{i+|\hat{\mathcal{X}}|})$ is displayed.



MixMatch - Prediction: Loss Function

- MixMatch를 거치면 새로운 Labeled, Unlabeled 데이터 생성

$$\mathcal{X}', \mathcal{U}' = \text{MixMatch}(\mathcal{X}, \mathcal{U}, T, K, \alpha)$$

- 생성된 데이터를 이용해 Supervised Loss + Consistency Loss 계산
- L2 norm은 bounded 되었고 incorrect prediction에 대해 덜 sensitive 하기 때문에 사용

$$\mathcal{L}_{\mathcal{X}} = \frac{1}{|\mathcal{X}'|} \sum_{x, p \in \mathcal{X}'} H(p, p_{\text{model}}(y | x; \theta))$$

--- CrossEntropy ---
= predictive
uncertainty에 대한
measure



$$\mathcal{L}_{\mathcal{U}} = \frac{1}{L|\mathcal{U}'|} \sum_{u, q \in \mathcal{U}'} \|q - p_{\text{model}}(y | u; \theta)\|_2^2$$

----- L2 Loss -----



$$\mathcal{L} = \mathcal{L}_{\mathcal{X}} + \lambda_{\mathcal{U}} \mathcal{L}_{\mathcal{U}}$$



Experiment

- Setting
 - ✓ Baseline과 MixMatch 모두 Wide Resnet-28 사용
 - ✓ Dataset 전체 label의 일부만 사용하고 나머지는 Unlabeled Data로 간주
 - ✓ Labeled Data의 개수를 점점 늘리며 실험
- Baseline
 - ✓ 파이 모델 (ICLR 2017)
 - ✓ Mean Teacher(NIPS 2017)
 - ✓ Virtual Adversarial Training(ICLR 2017)
 - ✓ Pseudo Label
 - ✓ Mixup



Experiment

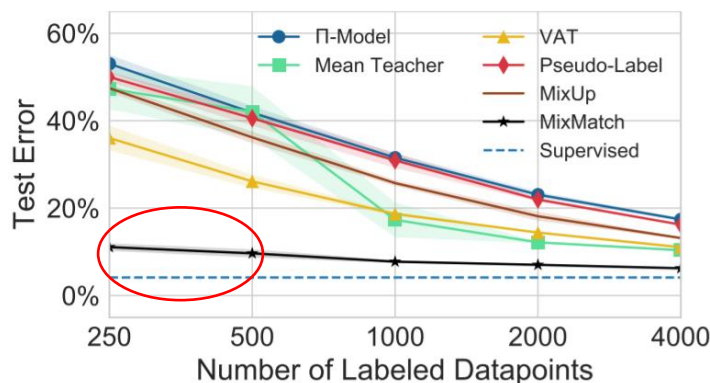


Figure 2: Error rate comparison of MixMatch to baseline methods on **CIFAR-10** for a varying number of labels. Exact numbers are provided in table 5 (appendix). “Supervised” refers to training with all 50000 training examples and no unlabeled data. With 250 labels MixMatch reaches an error rate comparable to next-best method’s performance with 4000 labels.

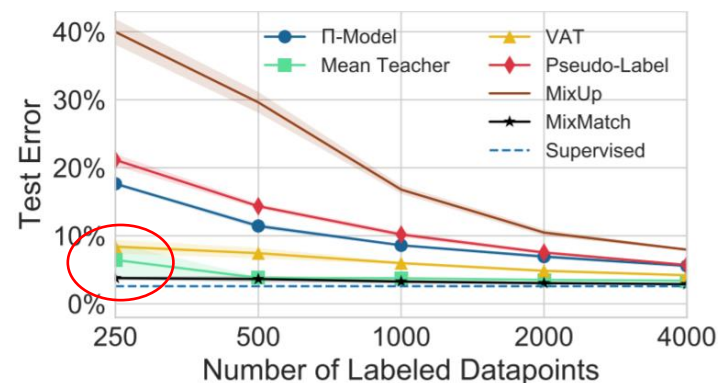


Figure 3: Error rate comparison of MixMatch to baseline methods on **SVHN** for a varying number of labels. Exact numbers are provided in table 6 (appendix). “Supervised” refers to training with all 73257 training examples and no unlabeled data. With 250 examples MixMatch nearly reaches the accuracy of supervised training for this model.



Experiment

- Ablation Study(특정 조건 제거해서 전체 성능에 미치는 효과 비교)
 - MixMatch 내부의 프로세스들이 모두 성능 증가에 영향을 끼침

Ablation	250 labels	4000 labels
MixMatch	11.80	6.00
MixMatch without distribution averaging ($K = 1$)	17.09	8.06
MixMatch with $K = 3$	11.55	6.23
MixMatch with $K = 4$	12.45	5.88
MixMatch without temperature sharpening ($T = 1$)	27.83	10.59
MixMatch with parameter EMA	11.86	6.47
MixMatch without MixUp	39.11	10.97
MixMatch with MixUp on labeled only	32.16	9.22
MixMatch with MixUp on unlabeled only	12.35	6.83
MixMatch with MixUp on separate labeled and unlabeled	12.26	6.50
Interpolation Consistency Training [45]	38.60	6.81

Table 4: Ablation study results. All values are error rates on CIFAR-10 with 250 or 4000 labels.



