Manifold Mixup Better Representations by Interpolating Hidden States

Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D., & Bengio, Y. (2019, May). Manifold mixup: Better representations by interpolating hidden states. In International Conference on Machine Learning (pp. 6438-6447). PMLR.

Introduction

- 이미지 기반 딥러닝은 input에 대한 feature map을 뽑기 위해 SOTA Backbone을 사용

- 하지만 기존의 backbone network는 train과 test 데이터셋이 서로 같은 분포를 가지고 있다고 가정하기에 Out-of-distribution 일 경우 다음과 같은 원인 때문에 성능이 좋지 않음
  - Decision boundar가 종종 sharply하고 data에 지나치게 가까움
  - Hidden representation space의 대부분은 높은 confidence predictions을 가짐(Overconfidence)

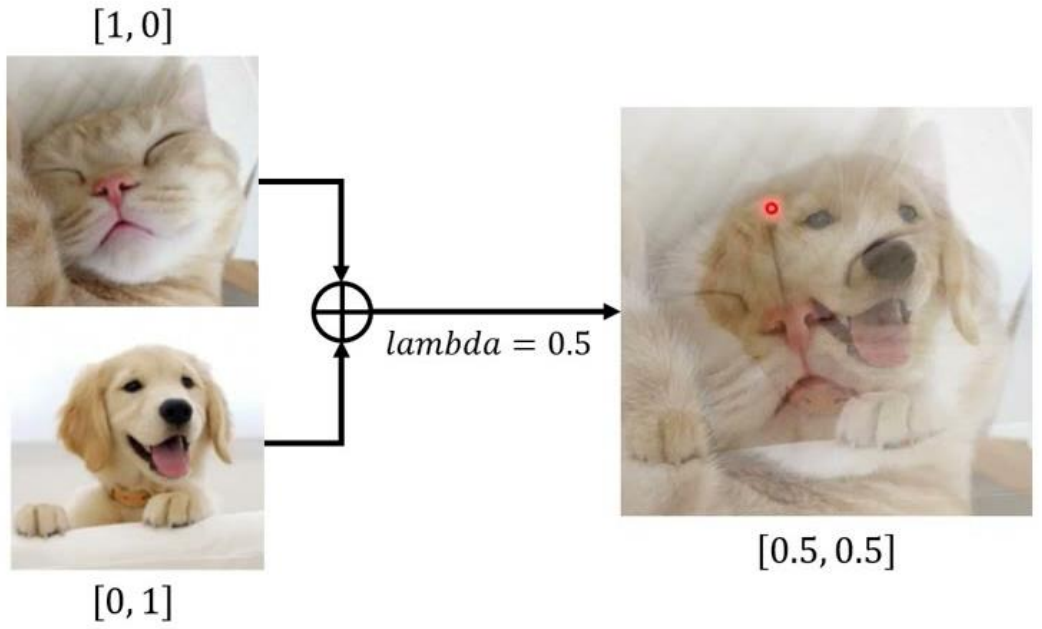
- training example의 hidden representation에 대하여 신경망을 훈련을 통한 정규화 방법인 “**Manifold Mixup**”을 제안

Mixup[1]

- Mixup은 17년 제안된 Vicinal Risk Minimization(VRM) 기반의 학습(훈련 데이터셋의 근방(vicinal) 분포도 함께 활용)
- 근접한 정도에 대한 분포를 beta distribution(확률에 대한 확률)으로 하며, 가지고 있는 training dataset에 대한 분포에서 어디쯤에 위치할지를 lambda 값으로 조정하면서 결정하는 형태
- Mixup을 통해 두 클래스 간의 decision boundary가 유연해짐 => 과적합이 덜 발생

$$\begin{aligned} \hat{x} &= \lambda x_i + (1-\lambda)x_j \\ \hat{y} &= \lambda y_i + (1-\lambda)y_j \end{aligned}$$

$\lambda \in [0,1]$는 $Beta(\alpha, \alpha)$에서 추출합니다.

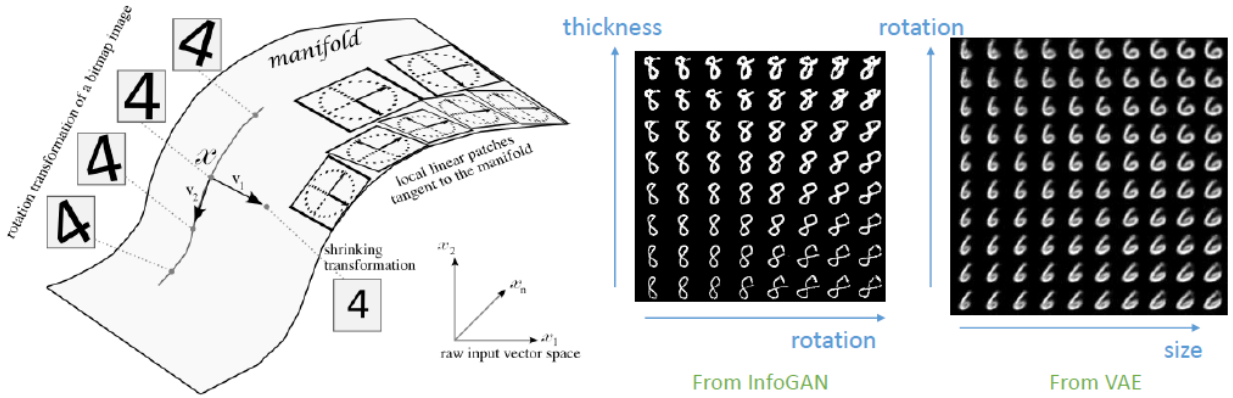


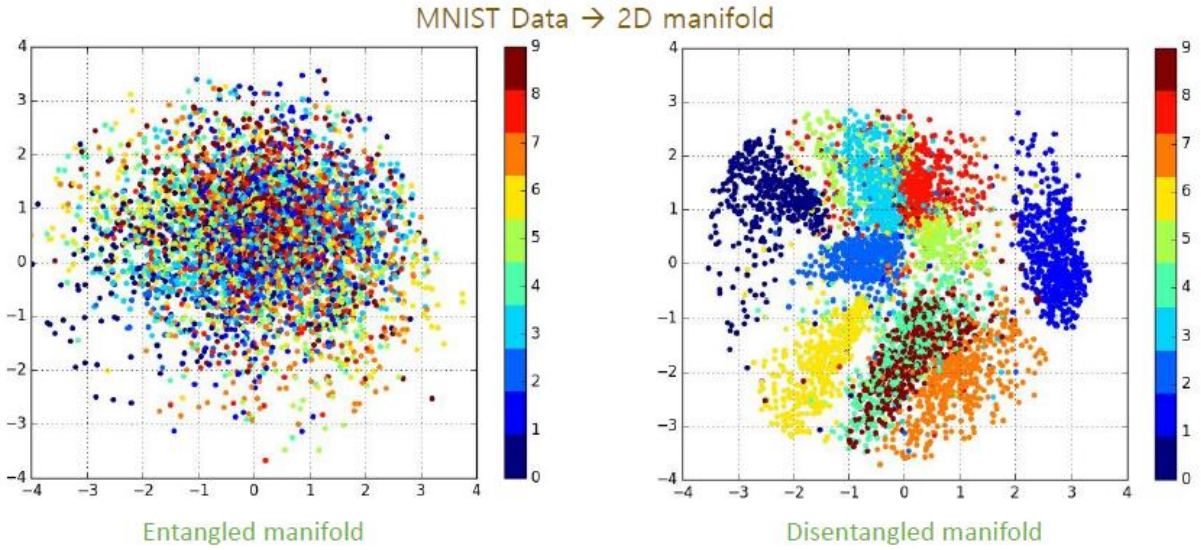
1. Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412.

Manifold

- 고차원 데이터를 데이터 공간에 뿌리면 sample들을 잘 아우르는 subspace가 있을 것이라는 가정을 통해 subspace로 데이터를 축소시키는 방법
- 고차원의 데이터를 잘 표현하는 manifold를 통해 샘플 데이터의 특징을 파악
- 고차원 데이터의 manifold 좌표들을 조정해보면 manifold의 변화에 따라 학습 데이터도 유의미하게 조금씩 변형 가능

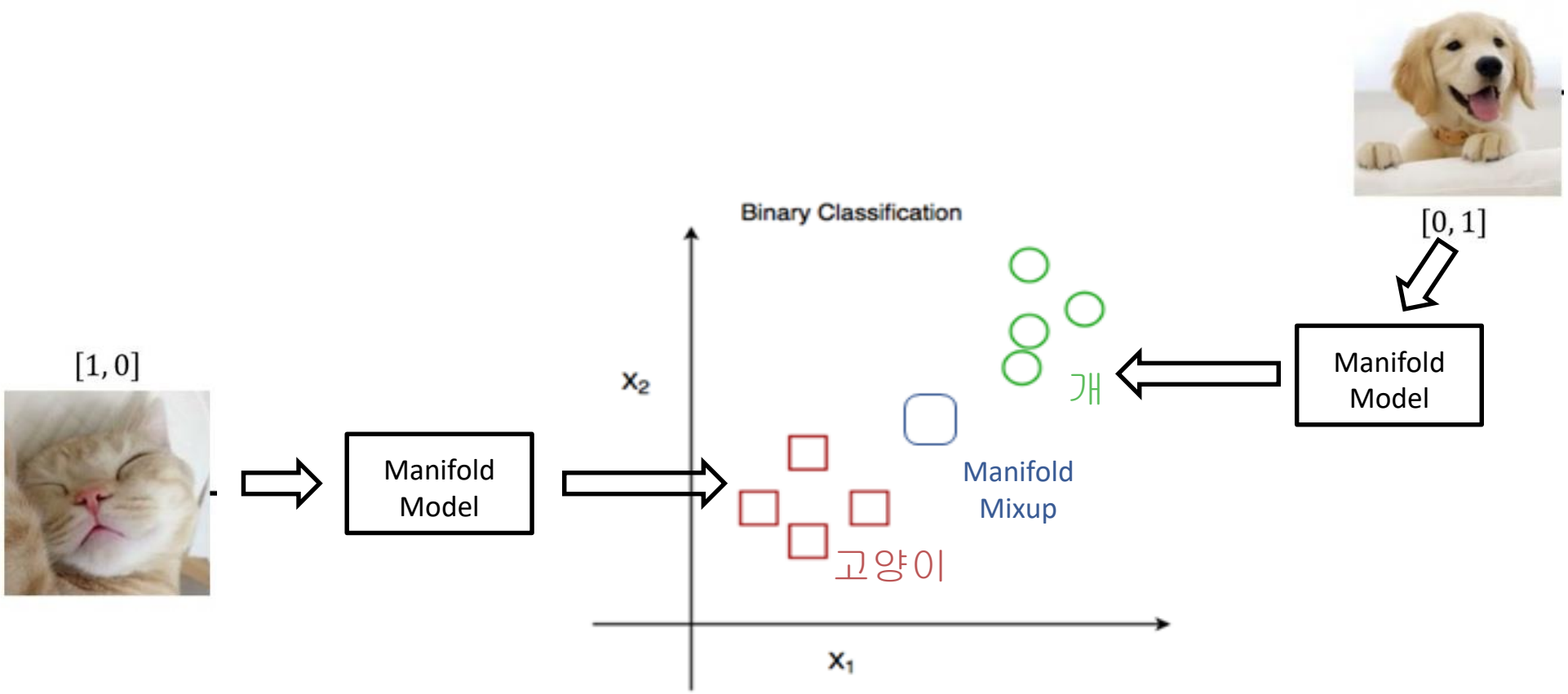
매니폴드 학습 결과 평가를 위해 매니폴드 좌표들이 조금씩 변할 때 원 데이터도 유의미하게 조금씩 변함을 보인다.





Manifold Mixup

- Mixup은 input data를 mix했다면 Manifold Mixup은 Manifold 상에서 Mix를 진행
- Mixup은 Outputs만을 가지고 추측을 하였다면 Manifold Mixup은 중간 결과물(layer단위)로 판단



Manifold Mixup Flattens Representations

- High level에서 Manifold Mixup은 class-specific representation을 flatten 할 수 있음
  - 기존의 데이터 배치인 A1과 B2를 보고 파란색과 빨간색을 잘 나타낼 수 있는 중간으로 Mixup 포인트 세팅(검은색)하면 A2와 B1은 제대로 표현할 수 없음
  - 딥러닝을 통해 데이터들을 특정 Manifold로 이동하면 모든 조건에 맞는 Mixup 포인트를 찾을 수 있음
  - 즉, 보다 일반적이고 간단한 arrangement를 통해 decision boundary가 정해져 OoD에 좋은 general한 manifold를 찾을 수 있음

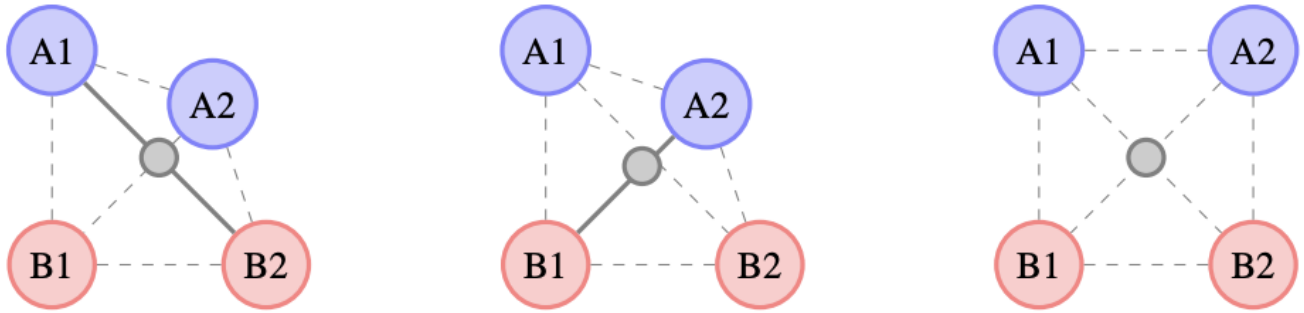


Figure 3: Illustration on why Manifold Mixup learns flatter representations. The interpolation between A1 and B2 in the left panel soft-labels the black dot as 50% red and 50% blue, regardless of being very close to a blue point. In the middle panel a different interpolation between A2 and B1 soft-labels the same point as 95% blue and 5% red. However, since *Manifold Mixup learns* the hidden representations, the pressure to predict consistent soft-labels at interpolated points causes the states to become flattened (right panel).

Manifold Mixup

- Manifold Mixup에는 5단계의 step이 존재 ($g_k(x)$: k번째 layer의 input data, $f_k()$: k번째 layer의 mapping)
  1. Set of eligible layers S에서 random layer k를 선택
  2. two random data (x_1, y_1), (x_2, y_2)를 선택한 layer layer 전까지 feed forward 진행 [$g(x_1)$, $g(x_2)$ 생성]
  3. $(g(x_1), y_1)$, $(g(x_2), y_2)$에 대해서 mixup 진행 $(\tilde{g}_k, \tilde{y}) := (\mathrm{Mix}_\lambda(g_k(x), g_k(x')), \mathrm{Mix}_\lambda(y, y'))$, $\mathrm{Mix}_\lambda(a; b) = \lambda * a + (1-\lambda) * b$
  4. mixed minibatch로 k번째 layer 부터 output까지 feed forward 진행
  5. Loss와 gradient 구해서 update

Empirical Investigation of Flattening

- MNIST dataset에 대하여Manifold Mixup을 포함한 Neural Network을 학습 후 network에 대한 hidden representations에 SVD를 적용

- 모든 네트워크와 regularizers에 대해 class마다의 the largest singular value와 all other singular values를 first hidden layer을 기준으로 측정

  - The largest singular value
    - Baseline : 51.73
    - Weight decay : 33.76
    - Dropout : 28.83
    - Input mixup : 33.46
    - Manifold mixup : 31.65
  - The sum of all the other singular values
    - Baseline : 78.67
    - Weight decay : 73.36
    - Dropout : 77.47
    - Input mixup : 66.89
    - Manifold mixup : 40.98

- weight decay, dropout, input mixup 모두 the largest singular value만 줄이지만, Manifold Mixup은 the sum of the all other singular values도 성능 향상

※ SVD : 직교하는 벡터 집합에 대하여, 선형 변환 후에 그 크기는 변하지만 여전히 직교할 수 있게되는 그 직교 집합은 무엇인가? 그리고 선형 변환 후의 결과는 무엇인가?

※ SVD 의의 : 최대한 중요한 정보들만 SVD해서 사용하면 사진의 차원은 줄어들지만 사진이 보여주고자 하는 내용은 살릴 수 있음

Experiments on Supervised Learning

- 사전 연구된 Mixup 방법 중에서 가장 좋은 성능을 보임

Table 1: Classification errors on (a) CIFAR-10 and (b) CIFAR-100. We include results from (Zhang et al., 2018)† and (Guo et al., 2016)‡. We run experiments five times to report the mean and the standard deviation of errors and neg-log-likelihoods.

PreActResNet18	Test Error (%)	Test NLL
No Mixup	4.83 ± 0.066	0.190 ± 0.003
AdaMix‡	3.52	NA
Input Mixup†	4.20	NA
Input Mixup ($\alpha = 1$)	3.82 ± 0.048	0.186 ± 0.004
Manifold Mixup ($\alpha = 2$)	2.95 ± 0.046	0.137 ± 0.003
PreActResNet34		
No Mixup	4.64 ± 0.072	0.200 ± 0.002
Input Mixup ($\alpha = 1$)	2.88 ± 0.043	0.176 ± 0.002
Manifold Mixup ($\alpha = 2$)	2.54 ± 0.047	0.118 ± 0.002
Wide-Resnet-28-10		
No Mixup	3.99 ± 0.118	0.162 ± 0.004
Input Mixup ($\alpha = 1$)	2.92 ± 0.088	0.173 ± 0.001
Manifold Mixup ($\alpha = 2$)	2.55 ± 0.024	0.111 ± 0.001

(a) CIFAR-10

PreActResNet18	Test Error (%)	Test NLL
No Mixup	24.01 ± 0.376	1.189 ± 0.002
AdaMix‡	20.97	n/a
Input Mixup†	21.10	n/a
Input Mixup ($\alpha = 1$)	22.11 ± 0.424	1.055 ± 0.006
Manifold Mixup ($\alpha = 2$)	20.34 ± 0.525	0.912 ± 0.002
PreActResNet34		
No Mixup	23.55 ± 0.399	1.189 ± 0.002
Input Mixup ($\alpha = 1$)	20.53 ± 0.330	1.039 ± 0.045
Manifold Mixup ($\alpha = 2$)	18.35 ± 0.360	0.877 ± 0.053
Wide-Resnet-28-10		
No Mixup	21.72 ± 0.117	1.023 ± 0.004
Input Mixup ($\alpha = 1$)	18.89 ± 0.111	0.927 ± 0.031
Manifold Mixup ($\alpha = 2$)	18.04 ± 0.171	0.809 ± 0.005

(b) CIFAR-100

Table 2: Classification errors and neg-log-likelihoods on SVHN. We run each experiment five times.

PreActResNet18	Test Error (%)	Test NLL
No Mixup	2.89 ± 0.224	0.136 ± 0.001
Input Mixup ($\alpha = 1$)	2.76 ± 0.014	0.212 ± 0.011
Manifold Mixup ($\alpha = 2$)	2.27 ± 0.011	0.122 ± 0.006
PreActResNet34		
No Mixup	2.97 ± 0.004	0.165 ± 0.003
Input Mixup ($\alpha = 1$)	2.67 ± 0.020	0.199 ± 0.009
Manifold Mixup ($\alpha = 2$)	2.18 ± 0.004	0.137 ± 0.008
Wide-Resnet-28-10		
No Mixup	2.80 ± 0.044	0.143 ± 0.002
Input Mixup ($\alpha = 1$)	2.68 ± 0.103	0.184 ± 0.022
Manifold Mixup ($\alpha = 2$)	2.06 ± 0.068	0.126 ± 0.008

Experiments to Novel Deformations

- 다른 augmentation을 적용한 test 데이터셋에서도 가장 좋은 성능을 보임

Table 5: Test accuracy on samples subject to novel deformations. All models were trained on normal CIFAR-100.

Deformation	No Mixup	Input Mixup ($\alpha = 1$)	Input Mixup ($\alpha = 2$)	*Manifold Mixup* ($\alpha = 2$)
Rotation U(−20°,20°)	52.96	55.55	56.48	60.08
Rotation U(−40°,40°)	33.82	37.73	36.78	42.13
Shearing U(−28.6°, 28.6°)	55.92	58.16	60.01	62.85
Shearing U(−57.3°, 57.3°)	35.66	39.34	39.7	44.27
Zoom In (60% rescale)	12.68	13.75	13.12	11.49
Zoom In (80% rescale)	47.95	52.18	50.47	52.70
Zoom Out (120% rescale)	43.18	60.02	61.62	63.59
Zoom Out (140% rescale)	19.34	41.81	42.02	45.29

Experiments to Adversarial Examples

- FGSM : 딥러닝 모델의 선형 구조를 지적한 공격 방법

=> 어떠한 noise를 추가로 넣어 딥러닝 모델을 바보로 만듦

- PGD : FGSM 응용

- Manifold Mixup 방식이 다양한 domain에 대해 학습을 가능하게 하므로 적대적 공격에 대해 방어 능력 지님

- 실험을 통해 CIFAR-10 데이터셋에 대해 우수한 방어 성능을 보임을 증명

Table 7: Test accuracy on white-box FGSM adversarial examples on CIFAR-10/CIFAR-100 (using a PreActResNet18 model) and SVHN (using a WideResNet20-10 model). We include the results of (Madry et al., 2018)†.

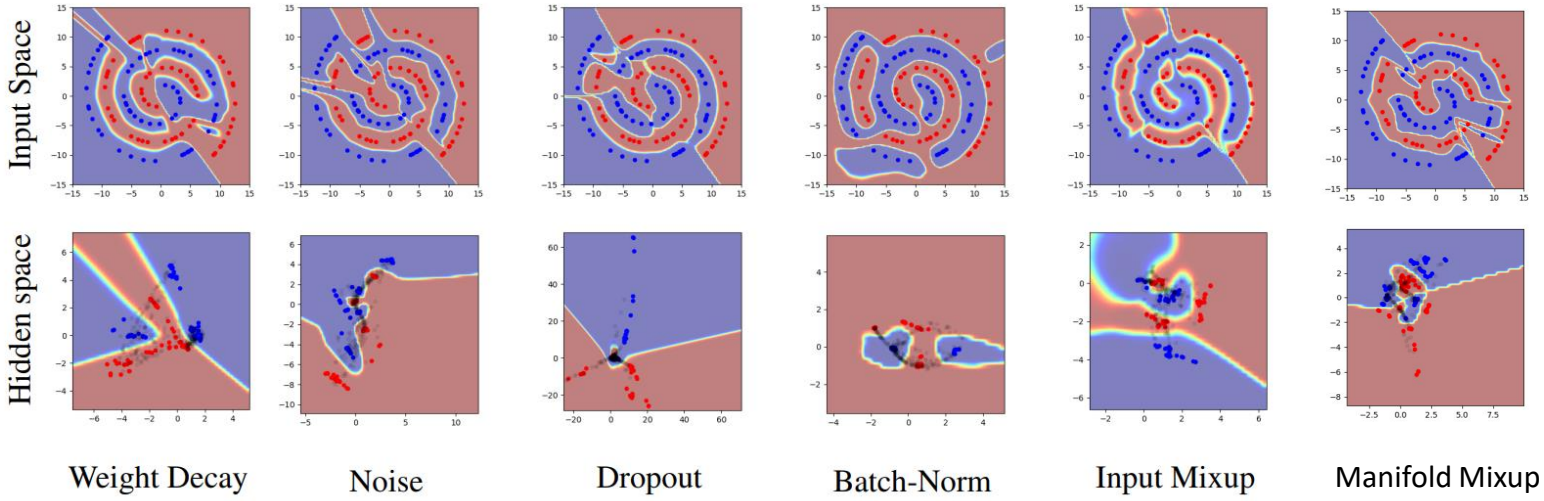
CIFAR-10	FGSM
No Mixup	36.32
Input Mixup ($\alpha = 1$)	71.51
Manifold Mixup ($\alpha = 2$)	77.50
PGD training (7-steps)†	56.10
CIFAR-100	**FGSM**
Input Mixup ($\alpha = 1$)	40.7
Manifold Mixup ($\alpha = 2$)	44.96
SVHN	**FGSM**
No Mixup	21.49
Input Mixup ($\alpha = 1$)	56.98
Manifold Mixup ($\alpha = 2$)	65.91
PGD training (7-steps)†	72.80

2. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

Conclusion

- 실험을 통해 다른 기법에 비해 다음과 같은 이점을 지님
  - 더 나은 generalization 수행
  - Test sample에 대한 log-likelihood 개선
  - Predicted data에 대한 성능 향상
  - Adversarial attack에 대한 견고성 향상(manifold mixup이 decision boundary를 데이터로부터 멀리 밀어 냈음)

- Manifold mixup에는 세 장점 존재
  - decision boundary를 smooth하게 해줌
  - hidden representation의 arrangement를 improve함 => low confidence prediction에 대한 regions를 넓힘
  - representations를 flatten함



Application method

- Mixup을 bridge로써 사용해 domain adaptation에 이용한 연구[3]

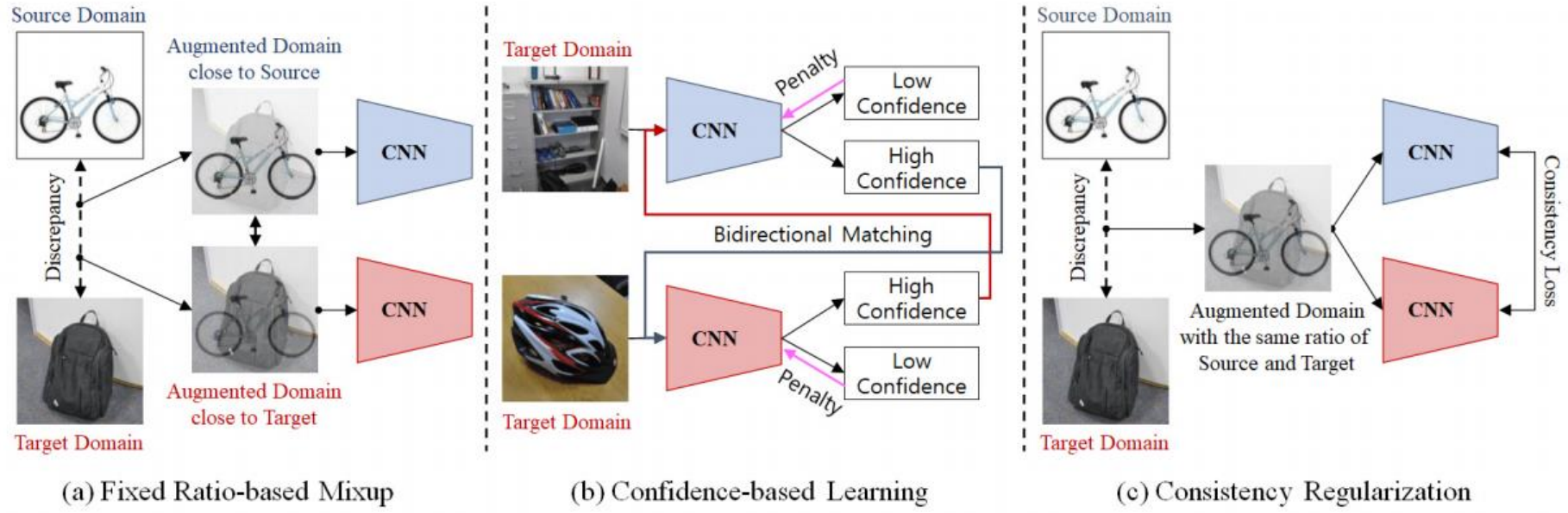
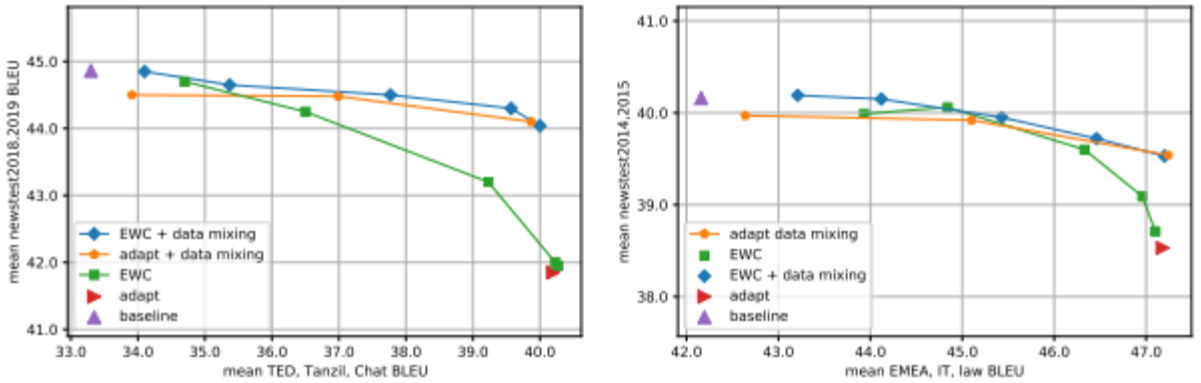


Figure 2. **An overview of the proposed method.** The proposed method consists of (a) fixed ratio-based mixup, (b) confidence-based learning, e.g., bidirectional matching with the positive pseudo-labels and self-penalization with the negative pseudo-labels, and (c) consistency regularization. Best viewed in color.

3. Na, J., Jung, H., Chang, H. J., & Hwang, W. (2021). Fixbi: Bridging domain spaces for unsupervised domain adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1094-1103).

Application method

- EWC와 Mixup을 사용해 generic performace를 유지하면서 new domain에 대한 성능 올리는 연구[4]



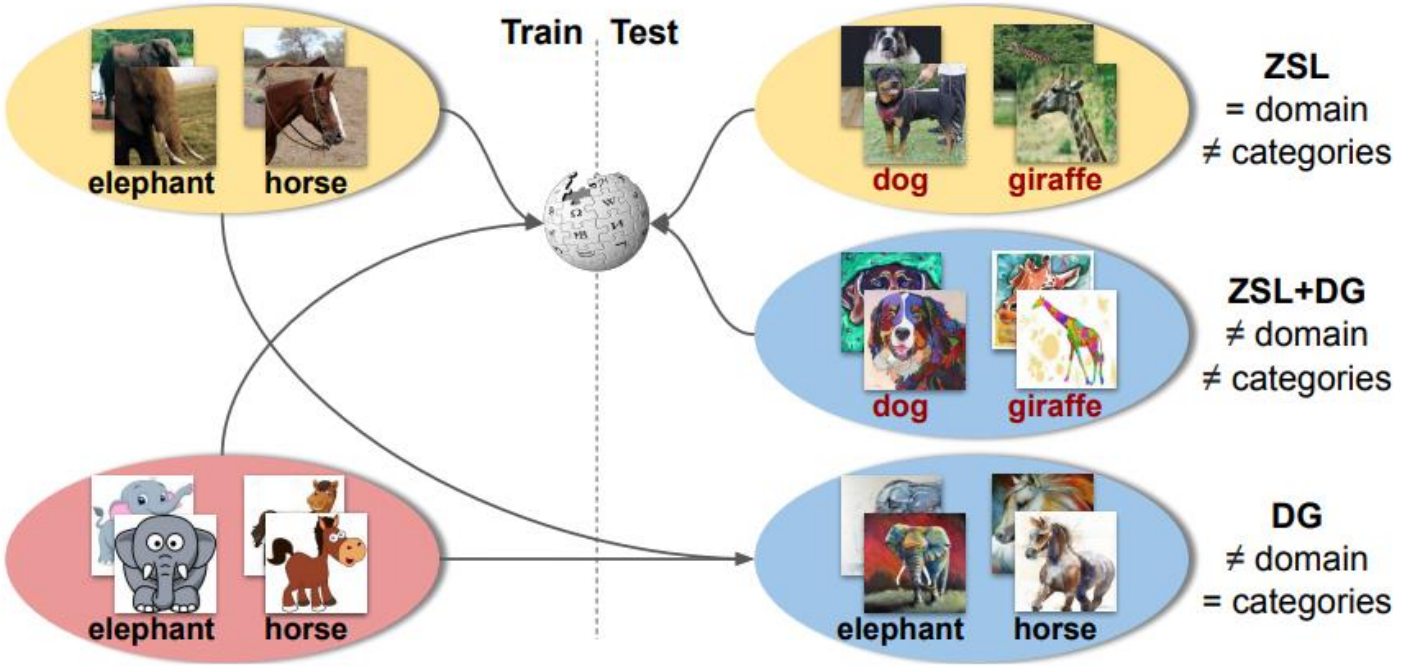
(a) DE→EN (b) EN→FR

$$\mathcal{L}' = \mathcal{L}_B(\theta) + \mathcal{L}_{A_1}(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{A,i})^2 \quad (2)$$

4. Hasler, E., Domhan, T., Trenous, J., Tran, K. M., Byrne, B., & Hieber, F. (2021, November). Improving the quality trade-off for neural machine translation multi-domain adaptation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 8470-8477).

Application method

- source domain과 target을 Mixup 기법을 활용해 ZSL과 DG를 동시에 해결하는 연구 존재[5]



5. Mancini, M., Akata, Z., Ricci, E., & Caputo, B. (2020, August). Towards recognizing unseen categories in unseen domains. In European Conference on Computer Vision (pp. 466-483). Springer, Cham.