Data Augmentation Reviews

Presenter: Sarah Kim

2019.09.09

1. Mixup (2017)

2. SMOTE (2002)

3. Dataset augmentation in feature space (2017)

4. Unsupervised Data Augmentation for Consistency Training (2019)

- Neural networks trained with ERM is hard to explain or provide generalization on testing distribution that differ **only slightly** from the training data.
- Data augmentation is one of methods to solve this issue and assumes that the examples in the vicinity share the **same** class.
- Mixup is a data augmentation method using convex combinations of examples and their labels.

- Mixup constructs virtual training examples

$$\begin{aligned} \tilde{x} &= \lambda x_i + (1-\lambda)x_j, \quad \text{where } x_i,\ x_j \text{ are raw input vectors} \\ \tilde{y} &= \lambda y_i + (1-\lambda)y_j, \quad \text{where } y_i,\ y_j \text{ are one-hot label encodings} \end{aligned} \tag{1}$$

  where
  - (x_i, y_i) and (x_j, y_j) are two examples drawn at random from training data;
  - $\lambda \sim \text{Beta}(\alpha, \alpha)$ for $\alpha \in (0, \infty)$.
- The mixup hyperparameter α controls the strength of interpolation, recovering the ERM principle as $\alpha \to 0$.

	Model	ERM	Mix-up
CIFAR10	ResNet	5.6 (6.0)	4.2 (4.0)
	WideResNet	3.8 (4.31)	2.7 (3.25)
	DenseNet	3.7	2.7
CIFAR100	ResNet	25.6	21.1
	WideResNet	19.4	17.5
	DenseNet	19.0	16.8

Table 1 : Test error rates(%) for the CIFAR 10 experiments using mix-up augmentation. $(\cdot)$ denotes my simulation results.

✓ Apply to another dataset and another model

1. mixup: Beyond Empirical Risk Minimization

Discussion

- We find that convex combinations of three or more examples with weights sampled from a Dirichlet distribution does not provide further gain.
- Interpolating only between inputs with equal label did not lead to the performance gains.
- Mixup reduces the memorization of corrupt labels, increases the robustness to adversarial examples.

2. SMOTE: Synthetic Minority Over-sampling TEchnique

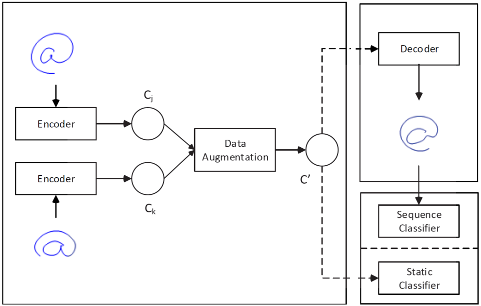
Method

- SMOTE is an over-sampling method that the minority class is over-sampled by creating **synthetic** examples for imbalanced classification problem.
- Method: For a sample x in the minority class, $k \in \mathbb{N}$,
    1. calculate k minority class nearest neighbors of x;
    2. choose a random sample between k-NN of x, denote by x';
    3. generate synthetic examples by interpolation as following:

    $$\tilde{x} = x + \lambda(x' - x)$$

    where λ is a random number between 0 and 1.

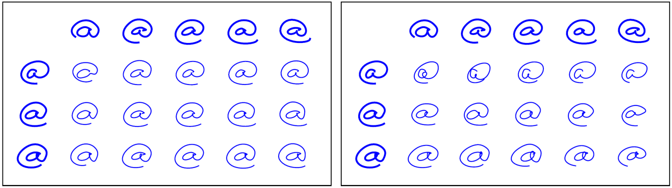
- We propose a data augmentation method that perform the transformation not in input space, but in a learned feature space.



- For each sample in the dataset, find its k nearest neighbours in feature space which share its class label. For each pair of neighbouring context vectors, a new context vector can be generated as:

$$\text{(Interpolation) } c' = c_j + \lambda(c_k - c_j) \text{ for } \lambda \in [0, 1]$$

$$\text{(Extrapolation) } c' = c_j + \lambda(c_j - c_k) \text{ for } \lambda \in [0, \infty)$$



(a) Interpolation

(b) Extrapolation

Figure 1 : Experiment results

- Data augmentation method to unlabeled data in a semi-supervised learning setting
- Unsupervised Data Augmentation (UDA) encourages the model predictions to be **consistent** between an unlabeled example and an augmented unlabeled example:

$$\text{Minimize a divergence } \mathcal{D}\left(p_\theta(y|x)\|p_\theta(y|\hat{x})\right), \tag{2}$$

where $p_\theta(y|x)$ is the output distribution for a given input x and a model with parameter θ, and $\hat{x}$ is a pertubed version of x.

- We focus on classification problems.
- Notations
  - L, U: the sets of labeled and unlabeled examples resp.
  - (x_i, y_i): an input, target pair for $i = 1, \ldots, |L|$
  - $p_\theta(y|x_i)$: a learning model with model parameters θ for a given x_i
  - $q(\hat{x}_i|x_i)$: the augmentation transformation based on an original example x_i
- Objective function:

$$\mathcal{J}(\theta) = \left\{ -\frac{1}{|L|} \sum_{(x_i, y_i) \in L} y_i^\top \log p_\theta(y|x_i) + \frac{\lambda}{|U|} \sum_{x_i \in U} \mathcal{D}_{\mathrm{KL}} \left(p_{\tilde{\theta}}(y|x_i) \| p_\theta(y|\hat{x}_i) \right) \right\}, \tag{3}$$

where $\hat{x} \sim q(\hat{x}|x)$, $\tilde{\theta}$ is a fixed copy of the current parameters θ, $\lambda > 0$ is a tuning parameter (default: 1).

4. Unsupervised Data Augmentation for Consistency Training

Method

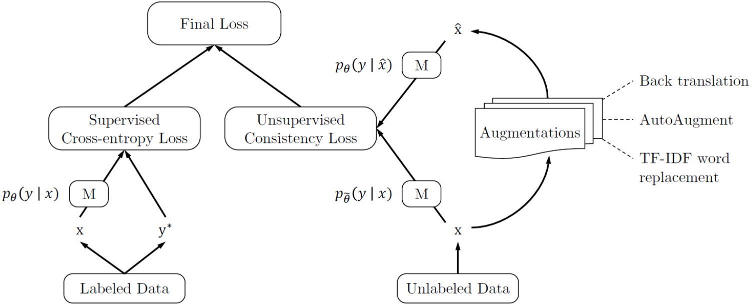
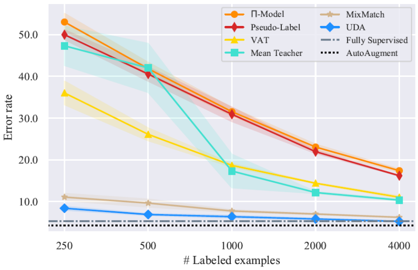


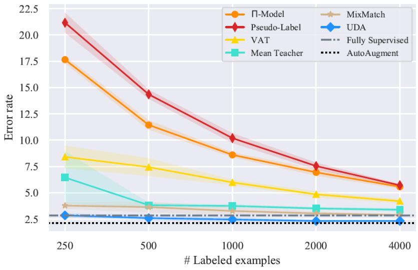
Figure 2 : Training objective for UDA, where M is a model that predicts a disribution of y given x.

4. Unsupervised Data Augmentation for Consistency Training

Experiments



(a) CIFAR-10



(b) SVHN

Figure 3 : Comparison with semi-supervised learning methods.

Reference

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

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002).
SMOTE: synthetic minority over-sampling technique.
Journal of artificial intelligence research, 16, 321-357.

DeVries, T., & Taylor, G. W. (2017).
Dataset augmentation in feature space.
arXiv preprint arXiv:1702.05538.

Xie, Q., Dai, Z., Hovy, E., Luong, M. T., & Le, Q. V. (2019).
Unsupervised data augmentation.
arXiv preprint arXiv:1904.12848.