Modern Topics on Statistical Learning Theory	Spring 2023
Lecture 11 — Apr. 14, 2023	
Prof. Qi Lei	*Scribe: Aiqing Li, Kaiwen Dai, Tanya Wang*

1 Overview

Data augmentation encompasses various techniques and methods for generating additional samples while maintaining their semantic significance. By employing these approaches, one can minimize overfitting and enhance the generalization capabilities of a model, ultimately leading to improved performance.

2 Introduction: Common Data Augmentation

2.1 Audio Data Augmentation

1. noise injection
2. shifting
3. changing the speed
4. changing the pitch

2.2 Text Data Augmentation

1. word/sentence shuffling
2. Paraphrasing
   (a) word replacement
   (b) syntax-tree manipulation
3. random word insertion
4. random word deletion

2.3 Image Augmentation

1. Operating on a single input $T : X \to X$
   (a) Geometric Transform: Flipping, Cropping, Rotation, Stretch, Zoom in/out
   (b) Randomly Change: RGB color channels, contract, brightness
   (c) Kernel Filters: Sharpness, blurring

(d) Random Erasing

2. Mixup

$$T : X \times X \to X \begin{cases} x_1, x_2 \to ax_1 + (1-a)x_2 \\ y_1, y_2 \to ay_1 + (1-a)y_2 \end{cases}$$

3. GAN: Generative Sythetic Data

4. Neural Style Transfer

3 Theoretical Analysis

3.1 Adding Gaussian noise

Using l2-loss $\mathcal{L} = \mathbb{E}_{X,Y\sim P_{X,Y}}[(f(x)-y)^2]$

- Population Loss

$$\mathcal{L}_\epsilon = \mathbb{E}_{X,Y\sim P_{X,Y}}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2 I)}(f(x+\epsilon))^2 \quad \longleftarrow \text{ Objective function when we add noise} \tag{1}$$

Assuming σ^2 thus $O(\sigma^2)$ are very small

$$\approx \mathbb{E}_{X,Y\sim P_{X,Y}}\mathbb{E}_\epsilon(f(x)+\epsilon^T\nabla f(x)-y)^2 \tag{2}$$

$$= \mathbb{E}_{X,Y}\mathbb{E}_\epsilon(f(x)-y)^2 + 2\epsilon^T\nabla f(x)(f(x)-y) + (\epsilon^T\nabla f(x))^2 \tag{3}$$

$$= \underbrace{\mathbb{E}_{X,Y}(f(x)-y)^2}_{\mathcal{L}} + \mathbb{E}_{X,Y}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2 I)}\underbrace{2\epsilon^T\nabla f(x)(f(x)-y)}_{=0} + \mathbb{E}_x\mathbb{E}_\epsilon(\epsilon^T \underbrace{\nabla f(x))^2}_{\nabla f(x)^T\cdot\underbrace{\epsilon\cdot\epsilon^T}_{=\mathbb{E}_\epsilon\sigma^2 I}\nabla f(x)} \tag{4}$$

$$= \mathcal{L} + \underbrace{\sigma^2\mathbb{E}_x||\nabla f(x)||^2}_{\text{regularize to encourage flatness}} \tag{5}$$

- Empirical Loss

$$\mathcal{L}_{n,m} = \frac{1}{n}\sum_{i,j}^{n,m}(f(x_i,y_i)-y_i)^2$$

s.t. $(x_i, y_i) \sim P_{X,Y}$, $\epsilon_j \sim \mathcal{N}(0, \sigma^2 I)$

3.2 Group view

We could also consider augmentation from a group perspective. Considering a group of transforms G (e.g., all rotations of images), we write the group element $g \in G$ to be an act on the sample space X. For each $g \in G : X \to X : x \to gx$, and $e \in G$ is the identity element of the group.

We assume that for any group element $g \in G$ and any $X \sim \mathbb{P}$, we have an equality in distribution:

$$X =_d gX$$
$$p_X(x) = p_X(gx)$$

Another assumption: $P(y|x) = P(y|gx), \forall g, \forall x$ (This is a quite strong assumption that usually does not holds in real cases)

Let $g_j \sim \mathbb{Q}$, where $\mathbb{Q}$ is a uniform distribution on G, and $(x_i, y_i) \sim \mathbb{P}_{X,Y}$. The empirical loss

$$\hat{\mathcal{L}}_{n,m} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} l\big(f(g_j x_i), y_i\big)$$
$$\hat{\mathcal{L}}(f) = \int_G \int_{X.Y} l\big(f(gx), y\big) d\mathbb{P}_{X,Y} d\mathbb{Q}(g)$$

We have the following understanding:

- **Invariant representations**. Given an observation x, and a group of transforms G acting on X, a feature $F : X \to Y$ is invariant if for all g, x:

$$F(x) = F(gx)$$

- **Equivariance**. A model is called equivariant with respect to a group G acting on the sample space if there is and induced group G^* acting on the parameter space Θ such that for any $X \sim \mathbb{P}_\theta$, and any $g \in G$, there is a $g^* \in G^*$ such that $gX \sim \mathbb{P}_{g^*\theta}$

In the group view setting, below is a lemma that characterizes the bias and variance under augmentation:

Invariance lemma: Let f be an arbitrary function. Let $\bar{f}(x) := \mathbb{E}_{g\sim\mathbb{Q}} f(gx)$ be the "orbit average" of f. (e.g. take all the rotations on X and then take the average)[1]

1. $\bar{f}(x) = \mathbb{E}[f(X)|X \in Gx]$, where $Gx := \{gx : g \in G\}$ (e.g. all the images generated from rotating x)
2. $\mathbb{E}_{X\sim\mathbb{P}} f(X) = \mathbb{E}_{X\sim\mathbb{P}} \bar{f}(X)$ (unbiased)
3. $Cov_{X\sim\mathbb{P}} f(X) = Cov_{X\sim\mathbb{P}} \bar{f}(X) + \mathbb{E}_{X\sim\mathbb{P}} Cov_{g\sim\mathbb{Q}} f(gX)$. The original variance is always larger than learning with augmentation

Intuitions

- augmentation improves sample efficiency and leads to variance reduction, but it's not very clear how it improves sample efficiency quantitatively
- With augmentation, the function class to search is smaller by ruling out f such that $f(x)$ is not the same as $f(gx)$

Some drawbacks of the Group view

- Applicable to limited data augmentation. (rotation/flip can be viewed as a group, which stretch cannot)
- Does not handle misspecification ($p(y|x) \neq p(y|gx)$)
- Not quantitative

4 Training with Data Augmentation

Now given that we have defined all data augmentations $\mathcal{A}$, we can add the augmented data to training samples and and formalize the empirical loss with DA

$$\hat{\mathcal{L}}_n = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} l\big(f(A_j(x_i)), y_i\big)$$

where $\{(x_i, y_i)\}_{i=1}^n$ are the original training samples, and $A_j \in \mathcal{A}$ are the (randomly) selected augmentations applied to each sample.

To encourage DA consistency (DAC), we can train the model based on an alternative empirical loss:

$$\hat{\mathcal{L}}_n = \frac{1}{n} \sum_{i=1}^{n} l\big(f(x_i), y_i\big) + \lambda \cdot d\big(h(x_i), h(A_j(x_i))\big) \tag{6}$$

where h is some invariant representation, predictor $f = w \circ h$, and $d(\cdot, \cdot)$ stands for some distance. The second term of the formulation 6 encourages the representation h (equivalently the prediction $f = w \circ h$) of the augmented data (x_i and $A_j(x_i)$) to be similar.

4.1 Results (Advantages) of using DAC

- It is shown that DAC (Data Augmentation Consistency) handles stronger misspecification of data augmentations (ie. $p(y|x) \neq p(y|A(x))$ for some $A \in \mathcal{A}$) than DA (Data Augmentation)[2].
- When there is no misspecification (ie. $y|x = y|Ax$ for all $A \in \mathcal{A}$), define

$$\tilde{A}(X) = \begin{bmatrix} \vdots \\ A_j(x_i) \\ \vdots \end{bmatrix}_{i,j} \qquad \tilde{M}(X) = \begin{bmatrix} \vdots \\ x_i \\ \vdots \\ x_i \\ \vdots \end{bmatrix}_{i}$$

  where $\tilde{M}(X)$ are simply copies of each training data x_i to make it the same shape as $\tilde{A}(X)$. Further, we can define

$$d_{\text{aug}} \triangleq \text{rank}(\tilde{A}(X) - \tilde{M}(X)) \tag{7}$$

  which measures the number of dimensions perturbed by DA.

  Then with DAC, we have

1. for linear function class (with DAC, ie. using empirical loss (6))

$$\mathbb{E}\left[\mathcal{L}(\hat{f}_{\text{DAC}}) - \mathcal{L}^*\right] \leq O(\frac{(d - d_{\text{aug}})\sigma^2}{n}) \tag{8}$$

which is a smaller (tighter) upper bound compared to the naive case (without DAC, ie. using empirical loss with only the first term of equation (6))

$$\mathbb{E}\left[\mathcal{L}(\hat{f}) - \mathcal{L}^*\right] \leq O(\frac{d\sigma^2}{n})$$

(Side note: it is also proven that risk of DAC $<$ risk of DA, ie. DAC is expected to perform better than naive training and training with DA.)

2. for two layer neural network (with DAC, ie. using empirical loss (6))

$$\mathbb{E}\left[\mathcal{L}(\hat{f}_{\text{DAC}}) - \mathcal{L}^*\right] \leq O(C_w\sqrt{\frac{(d - d_{\text{aug}})\sigma^2}{n}}) \tag{9}$$

which is a smaller (tighter) upper bound compared to the naive case (without DAC, ie. using empirical loss with only the first term of equation (6))

$$\mathbb{E}\left[\mathcal{L}(\hat{f}) - \mathcal{L}^*\right] \leq O(C_w\sqrt{\frac{d\sigma^2}{n}})$$

where the predictor of the shallow network is $f_\theta = (X \cdot B)_\dagger \cdot w$, B is orthogonal and $||w||_1 \leq C_w$.

3. When DA is expansive, please refer to the analytical bound from last lecture.

References

[1] Shuxiao Chen, Edgar Dobriban, and Jane H Lee. "A Group-Theoretic Framework for Data Augmentation". In: (2020). arXiv: `1907.10905 [stat.ML]`.

[2] Shuo Yang et al. *Sample Efficiency of Data Augmentation Consistency Regularization*. 2022. arXiv: `2202.12230 [cs.LG]`.