신경망 모델의 편향성을 줄이기 위한 데이터 증강 연구

손재범
한양대학교 컴퓨터공학과
e-mail : netoou@hanyang.ac.kr

A Study of Mixed Augmentation for Reducing Model Bias

Jaebeom Son
Dept. of Computer Engineering, Hanyang University

Abstract

Recent studies demonstrate that deep learning model is easily biased by trained with unbalanced datasets. For example, the deep network can be trained to make a prediction by background feature instead the real target's feature. For those problem, a measurement called leakage was introduced to digitize this tendency. In this paper, we propose augmentation strategy which are used generally in computer vision problem to remedy this bias problem and we showed a simple augmentation methods have a effect to this task with experiments.

1. Introduction

Upon development of peripheral environment of computational learning such as accumulation of big data and the advanced technology of parallel computing, deep neural networks have renown for its learning capability and ability to solve various range of tasks (e.g. image recognition [2], machine translation [3], generative model [4]). At this moment of artificial intelligence being integrated to real world services, non-technical issues (e.g. ethical, social) are emerged. For instance, the existence of accuracy disparities in neural network model which trained to solve face recognition task was reported on Buolamwini et. al. to warn for commercial gender classification applications and the authors demanded urgent attention to this problem(gender shades MIT media-lab). To outline this, the model made worst accuracy score when they predict darker-skinned female. On the other hand, lighter-skin male showed the best error rate. This was supposed to occur with respect to class biased training dataset. The other problem also reported that trained deep network doesn't use the main feature of target class to decide its prediction, instead it looks other background information [10]. For example, to predict woman the model use features like 'long hair' or 'white dress' rather than utilize the physical characteristics of female. Other method [6] proposed the measurement of gender bias of image feature space using reverse engineering strategy and proposed adversarial approach this gender bias problem on MS-COCO dataset [1].

In this paper, we propose simple and applicable data augmentation strategy to many neural network pipeline especially image recognition task. Named *mixup* [7] and *cutmix* [8], these methods are mixing two or more data onto one which has containing many soften features and relaxed class label information. We was able to show that the gender bias problem was mitigated by augmentation approach on subset of MS-COCO dataset which male is major class whereas female is minor. We will introduce about the two augmentation method and leakage measurement to measure gender bias. Then, we will show the specific experiment method and results at following sections.

2. Leakage measurement

In image recognition problem, such relationships between classes can occur class correlation on dataset. For instance, class 'long hair', 'lipstick' are more likely to be appeared with 'woman' class. Models trained with this type of datasets are prone to predict images which has 'long hair' or 'lipstick' as 'woman' class. Numerically express how trained model likely to be affected by internal classes relation of dataset, we use *leakage* [6]. *Leakage* is measured by train and evaluate base model called *attacker* [6] with internal class label and its original label. *Attacker* is expected to behave as reverse engineering about internal classes in the input images original label. we simply point out formal definition of *leakage* and other notations. Suppose an annotated dataset D is given which contain three attributes (x, y, c) that mean image, class label, internal class label each. And we refer attacker as f. The *dataset leakage* ℓ_D is:

$$\ell_D = \frac{1}{|D|}\sum\nolimits_{(y_i,c_i)\in D} 1\{f(y_i) == c_i\}$$

Where $1\{\cdot\}$ is the indicator function and (y_i, c_i) is data sampled from dataset D. The $|D|$ indicates the number of sampled in dataset D. The term *dataset leakage* ℓ_D is identical to performance of attacker f w.r.t training dataset, in other words, 'accuracy'. For measure the degree of model bias we introduce similar measurement to *dataset leakage* which named *model leakage* ℓ_H:

$$\ell_H = \frac{1}{|D|}\sum\nolimits_{(\hat{y}_i,c_i)\in D} 1\{f(\hat{y}_i) == c_i\}$$

Where $\hat{y}_i = H(x_i)$, prediction of model H about input data x_i. H is pretrained with data x and its original targets y. The magnitude of model bias defined to distance of two leakage scores called *bias amplification*, $\Delta = \ell_H - \ell_D$. Our goal is to curtail Δ to introduce simple approach.

3. Data augmentation

'Mixing' strategies are able to show effect to mitigate the proposed problems by internal class bias [5, 6]. The following two methods are simple and easy to apply to current training pipeline.

3.1 Mixup

One of our augmentation strategy to relax biased training with internal class representation is *mixup* [7]. *Mixup* linearly mixes two different data sampled from same dataset with mixing proportion parameter λ. The proportion of two different data λ is determined at random and drawn from the beta distribution $\lambda \sim Beta(\alpha, \alpha), \alpha \in (0, \infty)$. Then, we can define mixed data (x_{mixed}, y_{mixed}) which is sampled from the following vicinal distribution [9] called *mixup*:

$$\mu(x_{mixup}, y_{mixup} | x_i, y_i) = \frac{1}{n}\sum_{j}^{n} \mathrm{E}_\lambda \left[\delta\left(\mathrm{x}_{\mathrm{mixup}}=\lambda\cdot \mathrm{x}_\mathrm{i} +(1\text{-}\lambda)\cdot \mathrm{x}_\mathrm{j},\ \mathrm{y}_{\mathrm{mixup}}= \lambda\cdot\ \mathrm{y}_\mathrm{i}+(1\text{-}\lambda)\cdot\ \mathrm{y}_\mathrm{j}\right)\right]$$

Also we can write sampled data from this distribution as:

$$x_{mixup} = \lambda\cdot \mathrm{x}_\mathrm{i} +(1\text{-}\lambda)\cdot \mathrm{x}_\mathrm{j}$$
$$\mathrm{y}_{\mathrm{mixup}} = \lambda\cdot\ \mathrm{y}_\mathrm{i}+(1\text{-}\lambda)\cdot\ \mathrm{y}_\mathrm{j}$$

Where the (x_i, y_i) and (x_j, y_j) are two data sampled at random from dataset.

In this paper, the key property of *mixup* that blends two data while they keep their characteristics after mixed will help to achieve our goal 'reduce internal bias'. Features of each data are weaken by multiplied with λ or $1 - \lambda$ but the mixed data x_{mixed} contains broader range of features. The target y_{mixed} has two different softened labels with regarding to data x_{mixed}. In learning procedure, this mixed information which contain two different internal classes feature will guide model to learn less biased representation.

3.2 Cutmix

Possessing the key property of *mixup,* blending information of two data, *cutmix* [8] showed effect on relaxing biased learning. However, in detail of mixing algorithm *mixup* and *cutmix* are distinct. Suppose we manage image datasets. To generate mixed sample (x_{cutmix}, y_{cutmix}) from (x_i, y_i) and (x_j, y_j), proportion of two different data λ drawn from uniform distribution which is a special case of beta distribution $Beta(1,1)$. The masking area M is generated from λ. Then the mask region of sample image x_i will be replaced to cropped patch of x_j by the mask M. Following show this

$$x_{cutmix} = M \odot x_j + (1 - M) \odot x_i$$
$$y_{cutmix} = \lambda\cdot y_j + (1 - \lambda)\cdot y_i$$

Gender balance	Split	Man #	Woman #
False	no balance	16,225	6,601
	$\gamma = 1$	3,078	3,078
	$\gamma = 2$	8,885	6,588
	$\gamma = 3$	10,876	6,598
	Val	3,813	1,554
	Test	3,894	1,579
True	Train	3,000	3,000
	Val	1,500	1,500
	Test	1,500	1,500

Table 1. Dataset with several split conditions, used to train, validation, test Resnet model and measure *leakage*. Gender balance and $\boldsymbol{\gamma}$ is hyperparameter which influence to gender quantity of data.

Learning rate	Model mAP		
	mixup	cutmix	normal
1e-5	42.593	39.555	45.765
1e-4	51.274	50.062	52.858
1e-3	52.882	49.595	52.901

Table 2. This table shows the resnet model suffered underfitting problem. Instead to extend training epoch we increase learning rate to observe convergence. Lr with 1e-3 shows best mAP scores.

Where $\odot$ is element-wise multiplication. W, H are width and height of image. Mask region $M \in \{binary\}^{W\times H}$ is initialized to zero first then filled with 1 accordingly box coordinates (r_x, r_y, r_w, r_h). r_x and r_y are the center coordinates of masking region which are drawn from uniform distribution $r_w{\sim}U(0, W), r_h{\sim}U(0, H)$. r_w and r_h are determined by λ, $r_w = W\sqrt{1-\lambda}, r_h = H\sqrt{1-\lambda}$.

Similar to *mixup*, x_{cutmix} and y_{cutmix} also has two different softened features and labels. However, *'erase out and fill with another one'* strategy is able to lead a risk to lost some information of original in image space. But this strategy can assist to relax down the internal class bias problem by cutting off correlation between features and internal classes.

4. Experiments

In this paper, we show the effect of two mixing augmentation methods on the relaxing internal class bias problem. Experiment setting and training method are introduced at section 4.1 and the results are shown at table 2 of section 4.2.

4.1 Experiment Environment

We have taken experiments based on the similar setting of Wang et. al. which discussed about gender bias problem. Gender dataset of extracted subset of MS-COCO was used with several setting in regard of gender class balance. This gender class balance was determined by hyper-parameter γ. Transfer learning is effective strategy to handle classification task. We used ResNet-50 [11] pretrained on Imagenet as base model to train and evaluate internal gender bias. We set training epoch to 50 for fine-tuning pretrained Resnet model. We followed the same architecture of *attacker* to measure *leakage* and bias. We did not use F1 score measurement to derive bias amplific-

Attacker gender balance	Augment gender balance	Splits	Dataset leakage	Model leakage			Minimun bias amplification
				mixup	cutmix	normal	
False	False	no balance	73.85 ± 0.19	77.38 ± 0.39	77.61 ± 0.17	78.49 ± 0.29	3.98
		$\gamma = 3$	73.28 ± 0.37	74.82 ± 0.36	74.84 ± 0.62	75.97 ± 0.14	1.54
		$\gamma = 2$	68.42 ± 2.19	74.26 ± 0.36	73.04 ± 0.43	76.21 ± 0.42	4.62
		$\gamma = 1$	52.34 ± 1.14	64.78 ± 1.60	64.66 ± 1.71	67.79 ± 1.55	12.32
	True	$\gamma = 3$	73.28 ± 0.37	75.07 ± 0.56	76.63 ± 0.50	75.97 ± 0.14	1.79
		$\gamma = 2$	68.42 ± 2.19	73.85 ± 0.54	73.09 ± 0.57	76.21 ± 0.42	4.67
		$\gamma = 1$	52.34 ± 1.14	64.30 ± 0.82	64.94 ± 0.79	67.79 ± 1.55	11.96
True	False	$\gamma = 3$	67.06 ± 0.56	70.78 ± 0.46	69.20 ± 0.87	71.33 ± 0.12	2.14
		$\gamma = 2$		69.43 ± 0.80	67.75 ± 0.61	70.86 ± 1.21	0.69
		$\gamma = 1$		68.69 ± 0.37	68.58 ± 0.87	69.86 ± 0.53	1.52
	True	$\gamma = 3$		68.12 ± 1.06	68.92 ± 0.54	71.33 ± 0.12	1.06
		$\gamma = 2$		69.65 ± 1.24	67.53 ± 0.62	70.86 ± 1.21	0.47
		$\gamma = 1$		64.30 ± 0.82	64.94 ± 0.79	69.86 ± 0.53	-2.12

Table 3. Comparisons of two different training data augmentation and normal(without any other augmentation) conditions. Highlighted to light gray color indicates that best condition of each experiment environment. The model was trained with learning rate 1e-3. We also highlight best bias amplification score to gray color, the model trained with dataset split ratio 2 and *cutmix* with balancing sampled gender class method shows best amplification score when the attacker trained with gender balanced status. In every experiment setting on this table, *mixup* or *cutmix* shows better model leakage.

ation instead we use accuracy measure. Performance of trained model is measured with mAP score. We compared experiments of three different settings, 'with mixup', 'with cutmix', 'without mixing augment'.

In detail of learning process, every model was trained with splits of dataset (Table 1). We wrapped Pytorch dataset to *mixup* dataset with fixed parameter of beta distribution α to one. *Cutmix* implemented similar way to *mixup*. We mixing only two images and we divided mixing sampling into two cases, one was sampling two data at random and the other was sampling data at balanced gender ratio. Later case means that mixed image should contain both gender features and the label should contain two gender classes. Wang did experiments with fixed learning rate but we used learning rate of 1e-5, 1e-4, 1e-3. The attacker was trained with different splits of datasets contrasting with Wang et. al. which used only gender balanced dataset.

4.2 Results

To summarize the results of our experiments, in most cases, *mixup* and *cutmix* have shown that they reduced bias amplification between model and dataset leakage (Table 3). However, the performances of Resnet model which trained with selected MS-COCO dataset are dropped slightly measured with mAP score whereas performance improvement was reported on *mixup* and *cutmix* with Imagenet dataset. We conducted experiments to observe the influence of internal class balance on the *leakage* and *bias amplification*. We was able to observe that leakage was decreased when the class ratio goes to equal but the bias amplification was largely increased. Control learning rate of transfer learning also have shown effect to model performance (Table 2). We conducted other experiments with learning rate 1e-3 which showed the best model performances. In default setting of Wang et. al., we could discover the underfitting issue it was arisen owing to insufficient learning rate which makes model parameters to stopped before it trained enough.

5. Conclusion

In this paper, we investigated to simple augmentation method to relax internal bias problem called mixing augment which is effective to this type of task. However it still remains hard work to remedy bias problem and improve model performance both.

References

[1] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." *European conference on computer vision*. Springer, Cham, 2014

[2] Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." *International journal of computer vision* 115.3 (2015): 211-252.

[3] Sutskever, I., O. Vinyals, and Q. V. Le. "Sequence to sequence learning with neural networks." *Advances in NIPS* (2014).

[4] Goodfellow, Ian, et al. "Generative adversarial nets." *Advances in neural information processing systems*. 2014.

[5] Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." *Conference on fairness, accountability and transparency*. 2018.

[6] Wang, Tianlu, et al. "Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations." *Proceedings of the IEEE International Conference on Computer Vision*. 2019.

[7] Zhang, Hongyi, et al. "mixup: Beyond empirical risk minimization." *arXiv preprint arXiv:1710.09412* (2017).

[8] Yun, Sangdoo, et al. "Cutmix: Regularization strategy to train strong classifiers with localizable features." *arXiv preprint arXiv:1905.04899* (2019).

[9] Chapelle, Olivier, et al. "Vicinal risk minimization." *Advances in neural information processing systems*. 2001.

[10] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?: Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016.

[11] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.