A survey on Data Augmentation techniques for occlusion

Swapnil Singh, Rahul Kumar Yadav, Praveen Bairagee

Abstract

A survey made on various data augmentation techniques for Convolutional Neural Network.

1 Introduction

Image classification is the process of categorizing images based on their types or classes. It is an integral branch of Image Processing and has its application in fields as diverse as Machine Vision, Traffic Control systems, Satellite Imaging, Image analysis applications and more. There are a number of challenges encountered when dealing with Image Classification, such as intra-scale variation, illumination, scale-variation and occlusion to name a few. Occlusion occurs when only a portion of the object of interest, typically very small, is visible in an image. Occlusions are of three types: Self Occlusion, Inter-object occlusion and Background occlusion. Self occlusion occurs when some part of the object is obstructed or occluded by the object itself. Inter-object occlusion occurs when two or more objects overlap each other. Background occlusion occurs when a background component occludes the object.

Deep Learning models have made incredible progress in discriminative tasks. This has been fueled by the advancement of deep network architectures, powerful computation, and access to big data. Deep neural networks have been successfully applied to Computer Vision tasks such as image classification, object detection, and image segmentation thanks to the development of convolutional neural networks.

The datasets employed in training high performance models such as Convolutional Neural Networks tend to feature their object of interest at the centre and clearly visible. This kind of practice can encourage the model to rely more on easily recognizable details(such as faces) than the more subtle ones(body). This problem of image occlusion has been addressed by several authors and can be effectively countered by augmenting the training data via deliberate occlusion. The essence of this data augmentation is that if an object is only partially visible in an image, then the model should assess all the available details in the image, instead of relying on

all the obvious ones. Data augmentation is a very important technique to generate more useful data from existing ones for training practical and general CNNs.

In this paper, we aim to make a survey on various prevalent data augmentation and a few traditional geometric techniques highlighting their implementation and their results. We then discuss the results and show their applications as well as limitations.

2 Data Augmentation

Data augmentation is an effective regularization. Compared with other methods, data augmentation has many advantages. For example, it only needs to operate on the input data, instead of changing the network structure and data augmentation is easy to apply to many tasks, while other loss- or label-based methods may need extra design.

In order to prevent a CNN from focusing too much on a small set of intermediate activations or on a small region on input images, random feature removal regularizations have been proposed. Examples include dropout for randomly dropping hidden activations and regional dropout[9] for erasing random regions on the input. Researchers have shown that the feature removal strategies improve generalization and localization by letting a model attend not only to the most discriminative parts of objects, but rather to the entire object region. Some traditional geometrical methods of data augmentation are given below:

1)Cropping: Cropping images can be used as a practical processing step for image data with mixed height and width dimensions by cropping a central patch of each image. Additionally, random cropping can also be used to provide an effect very similar to translations. The contrast between random cropping and translations is that cropping will reduce the size of the input such as (256,256)$\rightarrow$(224, 224), whereas translations preserve the spatial dimensions of the image. Depending on the reduction threshold chosen for cropping, this might not be a label-preserving transformation.

2)Flipping: Horizontal axis flipping is much more common than flipping the vertical axis. This augmentation is one of the easiest to implement and has proven useful on datasets such as CIFAR-10 and ImageNet. On datasets involving text recognition such as MNIST or SVHN, this is not a label-preserving transformation.

3)Rotation: Rotation augmentations are done by rotating the image right or left on an axis between 1° and 359°. The safety of rotation augmentations is heavily influenced by the rotation degree parameter. Slight rotations such as between 1 and 20 have their benefits on digit recognition tasks such as MNIST, but as the rotation degree increases, the label of the data is no longer preserved post-transformation.

4)Translational: Shifting images left, right, up, or down can be a very useful transformation to avoid positional bias in the data. For example, if all the images in a dataset are centered, which is common in face recognition datasets, this would require the model to be tested on perfectly centered images as well. As the original

image is translated in a direction, the remaining space can be filled with either a constant value such as 0 s or 255 s, or it can be filled with random or Gaussian noise. This padding preserves the spatial dimensions of the image post-augmentation.

3 Existing methods for Data Augmentation

[1]Cutout: Terrance DeVries and Graham W. Taylor propose a simple regularization method which masks out random square regions of input during training. When this method is applied to an image it randomly selects a pixel coordinate within the image as a center point and then places the cutout mask around that location. When applied to modern architectures, such as wide residual networks or shake-shake regularization models, it achieves tremendous performance on the CIFAR10, CIFAR-100, and SVHN vision datasets.

[2]Hide and Seek: proposes a weakly supervised framework known as hide and seek, to improve object localization in images. This method divides an image into a grid, where grid patches are dropped independently. By using this technique, each training epoch would be randomly hiding different patches, thus forcing the model to focus on multiple relevant parts of the object. The patches are hidden only during training. During testing, the full image, without any patches hidden, is given as input to the network. Extensive tests have been conducted on ILSVRC 2016 to demonstrate improved localization over state of the art CNN architectures like AlexNet and GoogLeNet.

[3]Gridmask: Pengguang Chen and Shu Liu propose a method that neither removes a continuous big region like Cutout, nor randomly selects squares like hide-andseek. The deleted region is only a set of spatially uniformly distributed squares. In this structure, via controlling the density and size of the deleted regions, it achieves a statistically higher chance to achieve a good balance between the two conditions. In the experiment conducted on the image classification task using dataset ImageNet, GridMask can improve the accuracy of ResNet50 from 76.5% to 77.9%, much more effective than Cutout and HaS, which accomplish 77.1% and 77.2%.

[4]Mixup: Hongyi Zhang and Moustapha Cisse propose a data-agnostic and straightforward data augmentation principle which trains a neural network on convex combinations of pairs of examples and their labels. It regularizes the neural network to favor simple linear behavior in-between training examples. Incorporating mixup into existing training pipelines reduces to a few lines of code, and introduces little or no computational overhead. Mixup has been obsereved to improve the generalization error of state-of-the-art models on ImageNet, CIFAR, speech, and tabular datasets.

[5]CutMix: Sangdoo Yun and Dongyoon Han propose a regional dropout strategy. Instead of simply removing pixels, CutMix replaces the removed regions with a patch from another image. The ground truth labels are also mixed proportionally to the number of pixels of combined images. This which res there is no loss of informative pixels while retaining the advantages of typical regional dropout methods for attending to the non-discriminative parts of an object. CutMix shares similarity with Mixup which mixes two samples by interpolating both the image and labels.

While Mixup samples tend to be unnatural, CutMix can overcome this problem by replacing the image region with a patch from another training image.

[6]YOLOv3: Joseph Redmon Ali Farhadi propose a system which predicts bounding boxes using dimension clusters as anchor boxes. It 3 predicts an objectness score for each bounding box using logistic regression. Each box predicts the classes the bounding box may contain using multilabel classification. It's not as great on the COCO average AP between .5 and .95 IOU metric. But it has shown good results on the old detection metric of .5 IOU.

[7]YOLOv4: Alexey Bochkovskiy and Chien-Yao Wang propose a method to find the optimal balance among the input network resolution, the convolutional layer number, the parameter number and the number of layer outputs.The next objective is to select additional blocks for increasing the receptive field and the best method of parameter aggregation from different backbone levels for different detector levels. In place of a classifier, a detector has been used and the following are the prerequisites for it:

• Higher input network size (resolution) – for detecting multiple small-sized objects.

• More layers – for a higher receptive field to cover the increased size of input network.

• More parameters – for greater capacity of a model to detect multiple objects of different sizes in a single image.

This method has proven itself faster (FPS) and more accurate (MS COCO AP50...95 and AP50) than all available alternative detectors. The detector described can be trained and used on a conventional GPU with 8-16 GB-VRAM this makes its broad use possible. It has also verified a large number of features, and selected for use such of them for improving the accuracy of both the classifier and the detector. These features can be used as best-practice for future studies and developments.

[8]Random Erasing: Zhun Zhong and Liang Zheng propose Random Erasing. In this method, an image during the trainig section, within a mini-batch randomly undergoes either of the two operations: 1) kept unchanged 2)It randomly chooses a rectangle region of an arbitrary size, and assign the pixels within the selected region with random values. It has proven itself to be a lightweight method that does not require any extra parameter learning or memory consumption and can be integrated with various CNN models with ease. Experiments were conducted on CIFAR10, CIFAR100, and Fashion-MNIST with various architectures to validate its effectiveness.

4 Discussion

The interesting ways to augment image data fall into two general categories: data warping and oversampling. Many of these augmentations reveal how an image classifier can be improved, while others do not. It is easy to explain the benefit of horizontal flipping or random cropping. However, it is not clear why mixing pixels or entire images together such as in CutMix or YOLOv4 is so effective.

Manipulating the representation power of neural networks is being used in many interesting ways to further the advancement of augmentation techniques. Traditional augmentation techniques such as cropping, flipping, and altering the color space are being extended with the use of meta learning algorithms. Also, there is no consensus about the best strategy for combining data warping and oversampling techniques.

While, occlusions in the form of data augmentation have been credited for their success in model interpretation and weak localization, they haven't demonstrated any advantage on a large-scale dataset. This lack of rigorous assessment of the objects may increase fragility and overfitting for future classifications. Additionally, There are no existing augmentation techniques that can rectify a dataset that has very poor diversity with respect to the testing data. All these augmentation algorithms perform best under the assumption that the training data and testing data are both drawn from the same distribution. If this is not the case, it is very unlikely that these methods will be useful.

5 Conclusion

If the data augmentation is incorporated properly in the training procedure, we can attain better performance in the model. Training of models using occlusion augmentation improves on the models ability, with respect to occlusions, but this does not accounts for a test-time performance improvement because the test set contains no occlusions. Object detection is a difficult job in real time tracking of multiple objects due to occlusion.

Data Augmentation cannot overcome all biases present in a small dataset. For example, in a dog breed classification task, if there are only bulldogs and no instances of golden retrievers, no augmentation method will create a golden retriever. Many methods were developed to detect occlusion in present and previous works. The aim of developing such intelligent system is to augment the dataset with different variants of it help the model achieve a better holistic understanding of the object. Overfitting is generally not as much of an issue with access to big data. Data Augmentation prevents overfitting by modifying limited datasets to possess the characteristics of big data. The methods and performances of the best data augmenttion techniques are reviewed in this paper.

6 References

[1] DeVries, Terrance, and Graham W. Taylor. "Improved regularization of convolutional neural networks with cutout." arXiv preprint arXiv:1708.04552 (2017).

[2] Krishna Kumar Singh, Hao Yu, Aron Sarmasi, Gautam Pradeep, and Yong Jae Lee, Member, IEEE."Hide-and-Seek: A Data Augmentation Technique for Weakly-Supervised Localization and Beyond"arXiv: arXiv:1811.02545 (2018).

[3] Chen, Pengguang, et al. "Gridmask data augmentation." arXiv preprint arXiv:2001.04086 (2020).

[4] Zhang, Hongyi, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. "mixup: Beyond empirical risk minimization." arXiv preprint arXiv:1710.09412 (2017).

[5] Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6023-6032).

[6] Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." arXiv preprint arXiv:1804.02767 (2018).

[7] Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao. "Yolov4: Optimal speed and accuracy of object detection." arXiv preprint arXiv:2004.10934 (2020).

[8] Zhong, Zhun, et al. "Random erasing data augmentation." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 07. 2020.

[9] Konda, Kishore, Xavier Bouthillier, Roland Memisevic, and Pascal Vincent. "Dropout as data augmentation." stat 1050 (2015): 29.