Aalto University
School of Business
Degree Programme in Information and Service Management

Biranjan Raut

Improving Object Classification Using Data Augmentation

A Case Study in Application of Data Augmentation

Master's Thesis
Espoo, Jan 1 2018

DRAFT! — January 11, 2018 — DRAFT!

Supervisors: Professor Antti Ylä-Jääski, Aalto University
Professor Pekka Perustieteilijä, University of Helsinki
Advisor: Olli Ohjaaja M.Sc. (Tech.)

Aalto University
School of Business
Degree Programme in Information and Service Management

ABSTRACT OF
MASTER'S THESIS

Author:	Biranjan Raut		
Title:	Improving Object Classification Using Data Augmentation A Case Study in Application of Data Augmentation		
Date:	Jan 1 2018	**Pages:**	8
Major:	Data Communication Software	**Code:**	T-110
Supervisors:	Professor Antti Ylä-Jääski Professor Pekka Perustieteilijä		
Advisor:	Olli Ohjaaja M.Sc. (Tech.)		

A dissertation or thesis is a document submitted in support of candidature for a degree or professional qualification presenting the author's research and findings. In some countries/universities, the word thesis or a cognate is used as part of a bachelor's or master's course, while dissertation is normally applied to a doctorate, whilst, in others, the reverse is true.

!FIXME **Abstract text goes here (and this is an example how to use fixme).** FIXME! Fixme is a command that helps you identify parts of your thesis that still require some work. When compiled in the custom `mydraft` mode, text parts tagged with fixmes are shown in bold and with fixme tags around them. When compiled in normal mode, the fixme-tagged text is shown normally (without special formatting). The draft mode also causes the "Draft" text to appear on the front page, alongside with the document compilation date. The custom `mydraft` mode is selected by the `mydraft` option given for the package `aalto-thesis`, near the top of the `thesis-example.tex` file.

The thesis example file (`thesis-example.tex`), all the chapter content files (`1introduction.tex` and so on), and the Aalto style file (`aalto-thesis.sty`) are commented with explanations on how the Aalto thesis works. The files also contain some examples on how to customize various details of the thesis layout, and of course the example text works as an example in itself. Please read the comments and the example text; that should get you well on your way!

Keywords:	ocean, sea, marine, ocean mammal, marine mammal, whales, cetaceans, dolphins, porpoises
Language:	English

Motivation

In past decades deep-learning has garnered a lot of attention. Primarily due to its success at discovering complex structures in high-dimensional data. This has lead to the application of deep-learning to many domains of science, business, and government. Compared to conventional machine learning method deep-learning methods require little feature engineering, and it performs better with raw data in its natural form (LeCun et al., 2015). Deep-learning methods are a set of representation learning methods that allows a machine to be fed with raw data and to automatically discover the representations needed for detection or classification (LeCun et al., 2015).

Improvement in the hardware, software, and availability of data has made the application of deep-learning possible in various tasks. Although advancement in algorithm along with hardware has resolved majority of bottlenecks in applications of deep-learning and made deep-learning accessible to everyone, deep-learning still requires a large number of annotated training samples for better performance. One way to increase the annotated training samples is by collecting more unstructured data and building a manual or automated system of data annotations. However, this method can be expensive, error-prone (if manually sorted) and in some domains simply impossible due to a limited source of data. Another way to overcome this problem is by artificially generating more training samples from existing data. This process of creating artificial data from existing data is also known as Data Augmentation. Data augmentation has potential to increase the accuracy of the existing deep-learning models and at the same time reduce the overall cost associated with data acquisition and manipulation. It also can create an entirely novel application of deep-learning in a domain where it was previously impossible to apply deep-learning due to a limited number of data. Therefore, this thesis will explore existing data augmentation techniques to increase the training samples.

The primary motivation for taking on this contemporary problem in the field of deep-learning comes from a challenge faced by a case-company. The case-company operates in a heavily regulated industry where access to data is very limited and thus requires an alternative way to increase annotated training samples in order to improve its existing object classification model. As discussed above, access to reliable data can be a significant limiting factor in application of deep-learning for companies across various industries. Therefore, by exploring the effectiveness of available data augmentation to improve deep-learning model, this thesis not only tries to solve the problem faced by the case company but also provides a platform for further research

in application of data augmentation in various other domains.

Problem statement

This research will primarily try to address the problem of case company where due to regulatory reasons it is not feasible to collect enough training samples to train deep-learning models. Thus, data augmentation will be explored as a possible solution. As such, the main research question will be:

How to apply existing data augmentation methods to increase the classification accuracy of the deep-learning classifier with limited training samples?

Earlier Research

Data augmentation has been applied in deep-learning with varying degrees of success. The most prominent application of data augmentation has been in image classification, and it has proven to be a successful technique to improve the accuracy of a model (Krizhevsky et al., 2012). One of the widely used and accepted practices for augmenting image data is to perform geometric and color augmentations, such as reflecting an image, cropping and translating the image, and changing the color palette of the image. By applying various transformation to one image multiple images are generated with different perspective or colors and these images can be used as new training samples. This approach of data augmenting is also known as data warping. Data warping was quite successful in augmenting hand written character in MNIST handwritten digit database. The augmented data was then used to balance the amount of training examples for each character class in order to reduce the bias in classifier which favored frequently presented training examples (Wong et al., 2016).

Synthetic Minority Over-Sampling Technique (SMOTE) is another data augmentation technique which was inspired by data warping, particularly its ability to reduce the class imbalance in the handwritten digit problem. SMOTE has been used particularly to address the problem of class imbalance, where real-world datasets often only contain a small percentage of "interesting" or target class examples (Chawla et al., 2002). SMOTE algorithm works by creating synthetic samples from the minor class instead of over-sampling with replacement. The advantage of synthetic over-sampling compared to data warping is that synthetic examples are generated in feature-space, and thus the SMOTE algorithm is more application-independent.

One of the promising technique of data augmentation is Generative Adversarial Nets (GANs). GANs uses a min-max strategy where one neural net successively generates counterfeit samples from the original data distribution in order to fool the other net, and the other net is then trained to better distinguish the counterfeits (Goodfellow et al., 2014). There are two main components of a GAN: Generator and Discriminator. The role of generator network is to take a random input and generate a sample data, whereas the role of discriminator is to take input from the real data or from generator and to predict whether the input is real or generated. Discriminator aims to maximize the probability of assigning the correct label to both original sample and the counterfeit, whereas generator simultaneously aims to minimize the difference between original and counterfeit samples. In this sense, the discriminator and generator play a two-player minimax game. GANs have been effective even with relatively small sets of data. GANs have been used for example to train a self-driving car to drive in the night or in the rain using only data collected on a sunny day (Gurumurthy et al., 2017).

Aim of the Study

Although data augmentation techniques are not new in the field of deep-learning, some of the techniques are domain-specific and others still at experimental phase with varying degree of success. For example, technique of data-warping is not grounded in a sound theoretical background and thus produces inconsistent result across different types of applications. Similarly, GANs is fairly new technique and it is difficult to translate its theoretical framework into real application.

Therefore, this thesis has a two-fold objective. The primary objective will be to use data augmentation to improve the accuracy of existing object classification model. To do so, various contemporary data augmentation techniques will be explored. The Secondary objective of the thesis is to bridge the gap between existing data augmentation literature and its applications. In this way, the thesis can also provide an avenue for further research in exploration and implementation of data augmentation to solve the problem posed by limited training data.

Research Methodology

This research will closely follow Cross Industry Standard Process for Data Mining (CRISP-DM). CRISP-DM is a tried and tested method for data

mining that builds upon the previous attempts to define knowledge discovery methodologies (Wirth and Hipp, 2000). It is an iterative process that begins with business understanding and moves on to data understanding, data preparation, modeling, evaluation and finally deployment. The figure:1 graphically summarized the entire CRISP-DM process. Thesis will primarily focus on the empirical part of modeling and evaluation, where classification model generated using various augmentation methods will be evaluated against the existing base model.

Figure 1: CRISP-DM Process

Structure of the Thesis

The thesis will start by establishing a motivation and explaining the key objectives of the research, as well as explaining the research problem in Introduction part. Then thesis will review the existing literature in the Data Augmentation. From there on thesis will provide a brief overview of data and steps involved in data processing. In-depth analysis of data augmentation and evaluation will be provided in Modeling and Evaluation section. Finally,

in conclusion section, the main findings, limitation, and recommendations will be discussed.

Thesis Time Line

The thesis will follow approximately the following time line.

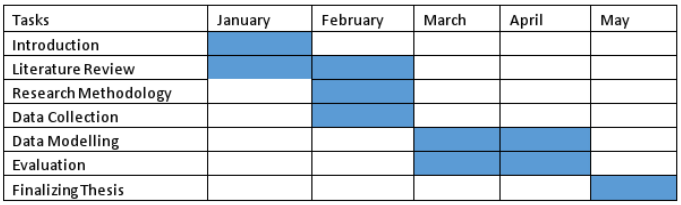

Figure 2: Thesis Time Line

Bibliography

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

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Gurumurthy, S., Sarvadevabhatla, R. K., and Radhakrishnan, V. B. (2017). Deligan: Generative adversarial networks for diverse and limited data. *arXiv preprint arXiv:1706.02071*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, pages 29–39.

Wong, S. C., Gatt, A., Stamatescu, V., and McDonnell, M. D. (2016). Understanding data augmentation for classification: when to warp? In *Digital Image Computing: Techniques and Applications (DICTA), 2016 International Conference on*, pages 1–6. IEEE.