Effectiveness of adversarial inputs on class-imbalanced convolutional neural networks

Rafael Possas, Ying Zhou

School of Information Technologies, University Of Sydney, Camperdown, NSW
{rafael.possas,ying.zhou}@sydney.edu.au
https://sydney.edu.au

Abstract. Convolutional neural networks performance has increased considerably in the last couple of years. However, as most machine learning methods, these networks suffer from the data imbalance problem - when the underlying training dataset is comprised of unequal number of samples for each label/class. Such difference enforces a phenomena known as domain shift that causes the model to have poor generalisation when presented with previously unseen data. Recent research have focused on a technique called gradient sign that explores domain shift on CNNs by modifying inputs to deliberately yield erroneous model outputs, while appearing unmodified to human observers. The wide use of image recognition models has heightened the need for better understanding of this threat. This work presents an experimental approach that sheds light on the link between imbalanced learning, transfer learning and adversarial attacks. Through a series of experiments we evaluated the fast gradient sign method on class imbalanced CNNs, linking model vulnerabilities to the characteristics of its underlying training set.

Keywords: convolutional neural networks, adversarial examples, gradient sign, imbalanced training, transfer learning

1 Introduction

Convolutional neural networks are a class of non-linear machine learning algorithms known for its state of the art performance on datasets with spatial structure. To date, not much research has been done into the adversarial inputs on CNNs - a process on which inputs are changed to manipulate the algorithm outputs. Experimental demonstrations of adversarial effects were carried out mainly by [1], [2], [3], [4] and have highlighted the need for improvement on the current state of CNNs techniques. Developing robustness to such attacks has become of the utmost importance as many commercial applications are based on the same small group of models. The motivation for adversarial robustness comes largely from being able to shield image recognition systems from behaving unexpectedly. Previous published studies are limited to show the general effectiveness of adversarial methods rather than understanding the deep relationship with the underlying training set distribution.

Domain shift or dataset shift [5] is also a well known cause for low performance of several machine learning algorithms [6], [7]. This happens when the joint distribution of inputs and outputs differs between training and test stages, causing models to perform badly on unseen data. Data distribution on real world is often skewed and rarely contains enough information to learn all the required features of the data domain which ultimately causes the phenomenon to become even stronger. Adversaries are proven to more readily explore domain shift [8], [9] and the question whether imbalanced training sets affects adversarial examples effectiveness is still unanswered.

Currently, there is no empirical evidence on the effectiveness of adversarial inputs on class-imbalanced CNNs. We designed a set of experiments to investigate the effects of both skewed distributions and model's internal gradient information on the robustness of these networks to such attacks. The main contributions of this work are as follows:

1. Evaluation of the resilience of imbalanced CNNs to adversarial attacks using the same and different model gradient information.
2. Investigation of classes with overlapping distributions and their relationship to both adversarial attacks and imbalanced learning problem.

Section 2 of this paper discusses the related work in both convolutional neural networks, gradient sign methods, adversarial/imbalanced learning. Section 3 provides details of the training models, imbalanced datasets and gradient sign methods used in our experiments. Section 4 presents the results on the under-sampled, over-sampled and balanced cases using both same/different model gradient. Sections 5 is dedicated to drawing conclusions and providing directions to related future work.

2 Related work

Previous work has shown that the high-dimensional non-linearities of convolutional neural networks [10] creates adversarial pockets of space. As shown on Figure 1, such pockets enables methods to deliberately create an adversary that produces an incorrect, high confidence prediction for an image without visible distortion [11]. This can be done by adding just enough intentional noise to each pixel of an image so as to fool the algorithm into predicting an incorrect label [2], [3], [12], [13].

The gradient sign method was introduced by Goodfellow et al. (2014) and has been used as the foundation of many of the experiments in adversarial crafting on CNNs. The results have shown that convolutional neural networks have linear behavior in very high dimensional spaces [2]. Most inputs were miss-classified not only by Goodfellow et. al (2014) experiments but many others [1],[3],[4].

The work of Papernot et al. (2016) has shown that one can use transfer learning to perform black-box attacks against CNNs [12], [15] and, thus, to intentionally force the model to predict specific labels .The combination of adversaries and transfer learning creates a threat vector for many state of the art methods.

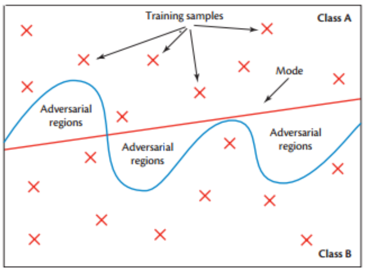


Fig. 1. Adversarial exploration of pockets of spaces [14]

Attacks can, however, depend on some specific internal information of the target model [8], [12]. As most recent applied methods depend on the network gradient information, there is a strong dependence on the network confidence per class label and the robustness of the model to adversarial attacks.

Techniques to overcome imbalanced learning have been developed for more general machine learning models. The work of Heibo et al [16], for instance, provides a technique for doing weighted sampling of minority classes to minimize the effect of imbalanced learning. Another approach could be to incorporate unsupervised clustering on synthetic data generation mechanism in order to avoid wrong generation of synthetic samples [17]. Currently, the research community is still working on ways to minimise the effects of adversarial attacks like the work of Billovits et al. [1] on a Bayesian framework to increase l_2 robustness to adversarial examples.

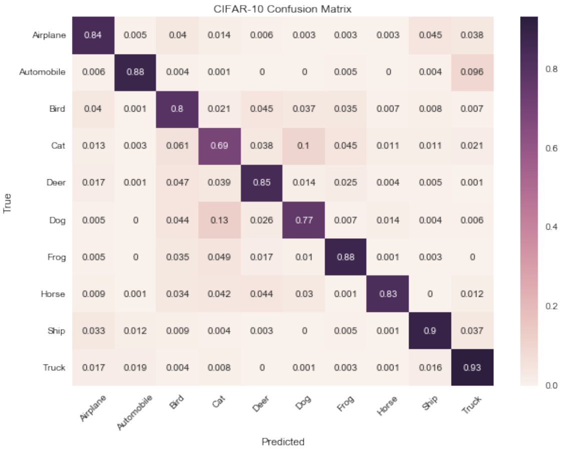
3 Experiment design

Our experiments aims investigate the relationships of the underlying learning structure of CNNs and the perturbation caused by gradient sign methods. In particular we focus on the investigation of how the gradient step from the sign method moves the points away from their distributions, and how this could be affected by both balanced and skewed distributions.This requires class labels of the data set to be different enough so we can make better assumptions of their distributions.

The 2014 ImageNet dataset [18] would be the natural choice. However, its hierarchically organized categories adds unnecessary complexity to the experiment design and make it hard to establish the causality relationship. Hence, we use CIFAR-10 in our experiment. CIFAR-10 data set is visually rich and empowers the analysis between different class labels. It comprises 5,000 samples of each class for training and 1,000 for test. It also contains 10 smaller, non-hierarchical class labels of 32x32 images, which helps to not only train our algorithms from scratch but also to understand deeper relationships between classes.

3.1 Network architecture and synthetic dataset imbalance

Network Architecture - All the experiments were done using the VGGNet architecture [19]. The two FC-4096 layers at the end were replaced by one single layer with 512 neurons and RELU activations. In addition, the total number of convolutions blocks and pooling layers were reduced to 3, with the first layer having 2 stacked convolution layers followed by a max pooling of stride 2x2 and the last two layers with 3 stacked convolutions also followed by a max pooling of stride 2x2. We have used RMSProp [20] as the optimisation technique with a learning rate of 10^{-4} and the decay 10^{-5}. Figure 2 shows the results of our models on the CIFAR10 original dataset.



	Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck
Airplane	0.84	0.005	0.04	0.014	0.006	0.003	0.003	0.003	0.045	0.038
Automobile	0.006	0.88	0.004	0.001	0	0	0.005	0	0.004	0.096
Bird	0.04	0.001	0.8	0.021	0.045	0.037	0.035	0.007	0.008	0.007
Cat	0.013	0.003	0.061	0.69	0.038	0.1	0.045	0.011	0.011	0.021
Deer	0.017	0.001	0.047	0.039	0.85	0.014	0.025	0.004	0.005	0.001
Dog	0.005	0	0.044	0.13	0.026	0.77	0.007	0.014	0.004	0.006
Frog	0.005	0	0.035	0.049	0.017	0.01	0.88	0.001	0.003	0
Horse	0.009	0.001	0.034	0.042	0.044	0.03	0.001	0.83	0	0.012
Ship	0.033	0.012	0.009	0.004	0.003	0	0.005	0.001	0.9	0.037
Truck	0.017	0.019	0.004	0.008	0	0.001	0.003	0.001	0.016	0.93

Fig. 2. Results of our adapted VGG architecture on CIFAR-10 dataset

Dataset imbalancing - As the CIFAR10 dataset is not naturally imbalanced, we have artificially created two variations on which we trained the imbalanced networks. While one dataset consisted on a direct under-sample of the target class to 1,000 samples, the other was changed using an oversampling of the target class (or an under-sampling of all other classes). We kept the number of samples for the target class at 5,000 while all other classes were reduced to 1,000 samples. For each class of the two different datasets configurations, a network was then trained until convergence using the same hyper-parameters as the balanced case. Each model was evaluated against a test set of 1,000 samples of the target class which was perturbed by its own under/over-sampled model and the balanced model. The two sources of gradient information will be referred from now on as same and different model perturbations. Both imbalanced networks were separately tested for each class for both same model and different model adversaries. The same model test was designed to investigate the vulnerability of class imbalance on adversarial examples while the different model test main goal is to verify the robustness on transfer learning environments. In total we evaluated 50 different combinations: 20 for each different imbalanced dataset (same model gradient and balanced network gradient) and 10 for the balanced network using its own gradients on each class. Figure 3 shows the accuracy for the models without any perturbation.

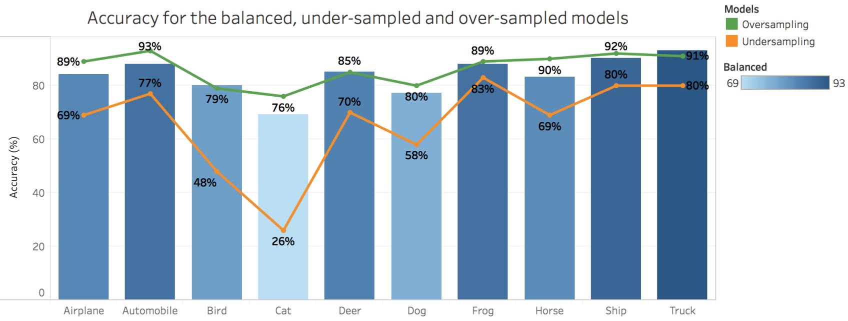


Fig. 3. Target class accuracy on all models

3.2 Gradient sign methods

The gradient sign is a method that uses internal gradient information so as to create directed perturbation to input data. The resulting adversarial label will have different meanings whether one adds or subtracts noise according to equations 1 and 2. Suppose the current true label of the class is selected as a gradient candidate, adding noise would mean that we increase the cost function of our input while subtracting noise is the same as minimizing our loss function even further [21]. One could also choose a different label as gradient candidate in order to make the adversarial more or less likely that label. Equations 1 and 2 will from now on be referred as ascent and descent methods. Perturbations could also be applied by two different variations of the gradient sign. While the fast gradient sign applies a single perturbation to the input, the iterative gradient sign performs the same perturbation a chosen number of times iteratively [2]. Figure 4 shows an example of adversarial created using the fast method.

$$C(x + \delta) \approx C(x) + \epsilon * sign(\nabla C) \tag{1}$$

$$C(x + \delta) \approx C(x) - \epsilon * sign(\nabla C) \tag{2}$$

In order to enforce consistency throughout our experiments, we have chosen the true sample label as the backpropagated gradient along with the fast gradient sign ascent method. The intuition behind this choice is that we look to increase

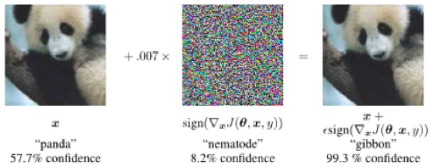

+ .007 ×
=
x
"panda"
57.7% confidence
$\text{sign}(\nabla_x J(\theta, x, y))$
"nematode"
8.2% confidence
$x + \epsilon\text{sign}(\nabla_x J(\theta, x, y))$
"gibbon"
99.3 % confidence

Fig. 4. Adversarial example crafting with fast gradient sign [2].

the cost function of the target class by moving away from the current true label. The use of an iterative gradient sign would almost always lead to a successful adversary due to the total amount of perturbation it adds to the input, therefore, we would not be able to test sensitivity of the method to the imbalanced nature of the training dataset. Moreover, in order to test our networks, we needed an ϵ value that would not only keep the image meaning understandable to human perception but also provide only the minimum amount of perturbation so as to push most of the samples to the closest label vicinity, hence, leading to a successful misclassification. From all the trials performed, the value of ϵ that seemed to fulfill our needs was 0.01 as with this values the number of misclassification on the balanced model was the greatest while keeping images meaning intact.

4 Results

We use the results of the balanced network on adversarial attacks as the baseline that defines whether imbalanced CNNs are less or more vulnerable to adversarial learning. Table 1 shows that the accuracy for all classes is drastically reduced when the balanced model is presented with adversarial examples. Models with under-sampled datasets were even more vulnerable than balanced networks. Figure 5 shows the relative difference for all the three different networks (balanced, under-sampled and over-sampled). Values were calculated by finding the difference between the perturbed accuracy and the non-perturbed accuracy of each class model. They represent the percentage on which the initial accuracy was reduced. The under-sampled model had the higher relative difference on average, which shows that the imbalanced nature of the dataset ended-up increasing the vulnerability of the model.

Perturbation on the over-sampling case had a weaker effect, as the small push caused by our ϵ was not enough to move points to outside of their distributions. Objects of the over-sampled classes would need bigger steps in order to successfully create an adversarial that leads to a wrong classification label. Accuracy for most of the over-sampling cases was around 45% and the relative difference was the lowest of all three models, which shows robustness of the target over-sampled class.

Class imbalanced models are naturally affected by the false positive and false negative trade off shown on figure 6. The decision boundaries on such

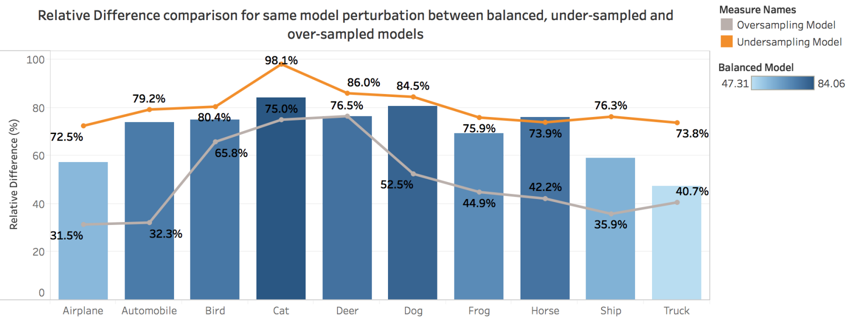


Fig. 5. Relative difference for each model. Higher numbers means more vulnerability

models favour the class with more samples and, hence, increases the accuracy for one class while decreasing for the other classes. The area under the curve for misclassified examples on the under-sampled distribution is bigger, and it is caused by the suboptimal exploration of feature space of that class. This effect is exploited by adversaries as there is an increase on the misclassification rate of distributions with lower amplitude. An under-sample of a specific label causes its distribution to be squished into space and, hence, have less impact on the definition of decision boundaries.

The increased number of samples of the over-sampled label causes the network to perform a trade-off when optimizing its loss function. For instance, the decision boundary would be chosen in order to minimize the total error of the network. The cost function is lower when the decision boundary minimizes the misclassification of the majority class as there is a higher number of samples. The choice of a biased decision boundary could be one of the factors explaining the higher resilience of over-sampled networks.

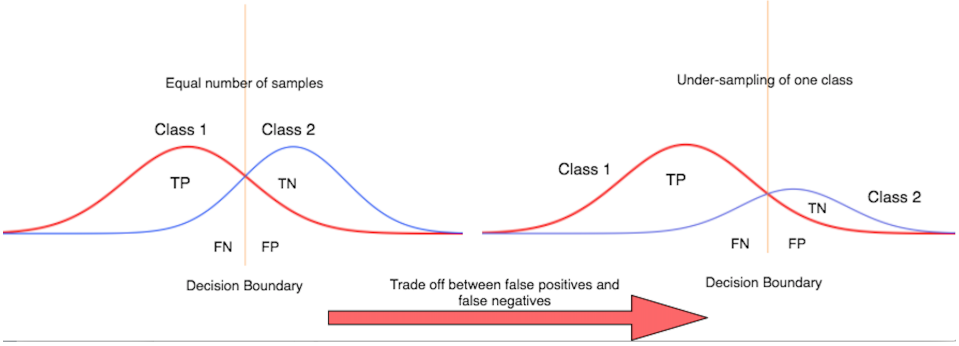


Fig. 6. Dataset imbalance causes models to perform adjustments of decision boundaries leading to an increase on accuracy of the majority class and decrease on the minority class.

	Different Model		Same Model		
Class Label	Undersample	Oversample	Balanced	Undersample	Oversample
0 - Airplane	60%	87%	36%	19%	61%
1 - Automobile	64%	91%	23%	16%	63%
2 - Bird	38%	73%	20%	9.4%	27%
3 - Cat	21%	72%	11%	0.5%	19%
4 - Deer	58%	80%	20%	9.8%	20%
5 - Dog	47%	76%	15%	9%	38%
6 - Frog	76%	88%	27%	20%	49%
7 - Horse	59%	88%	20%	18%	52%
8 - Ship	69%	89%	37%	19%	59%
9 - Truck	46%	87%	49%	21%	54%

Table 1. Results for the two different sources of perturbations along with the two different imbalanced datasets

4.1 Transfer learning and overlapping distributions

Transfer learning - The use of a different model gradient for creating adversaries has shown less effective when compared to the same model attack. As the overall gradient have not only different direction but also magnitudes, the attacked system has proven to be more robust to the attack. The experiment reveals that although gradient sign is quite effective for fooling networks it does require a good amount of knowledge from the underlying training parameters so as to unleash its full potential. Attacking an under-sampled/over-sampled network with the gradient of the balanced network did not show to be as effective as using the same model's gradient. The average accuracy of an under-sampled model attack with adversaries generated from a different network was 53.8% while the same metric was 25.8% for the same model attack. Even that our training samples are within the same data domain, there are still huge differences on the gradients learned from the network.

Overlapping distributions - The results for class distributions with a similar set of features shows that the adversarial attack is also stronger. Figure 2 shows that for the pairs cat/dog and automobile/truck there is already a natural misclassification between one another . On this case, our experiment shows that the adversarial attacks intensifies this phenomena by increasing the number of times on which one class is picked over the other. Figure 7 shows that cats are increasingly misclassified as dogs when under-sampled dataset on the cat class is used. While on the cat under-sampling case 40% of the samples were misclassified as dogs, on the oversampling one, 39% of the dogs were misclassified as cats. This outcome provides interesting insights, as it shows that the gradient sign is behaving linearly in the high dimensional space and 'moving' in the direction of the closest vicinity.

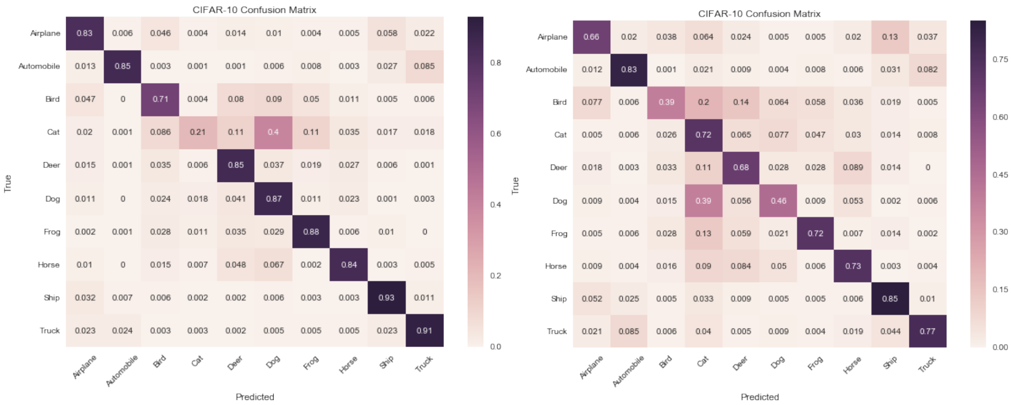


Fig. 7. From left to right: cat under-sampling / over-sampling with perturbation. Bottom left and right: dog under-sampling / over-sampling with perturbation

5 Conclusion and future work

We have shown that adversarial attacks are even stronger on datasets with under-sampled class labels and that the decision boundary trade-off on the over-sampled classes increases their robustness to adversarial examples. Labels with similar features have also presented higher vulnerability to the fast gradient sign methods as their similarities in the high dimensional space facilitates the technique to successfully create an adversary of the similar class.

As several commercial applications rely on almost the same group of models, the understanding of such properties are of extreme importance in order to drive the research community in creating algorithms with better overall performance. Future work in this field could look further in datasets with a higher number of classes and more complex relationships between labels so as to not only confirm our insights but also discover new interesting properties of CNNs. In particular, the investigation of whether the synthetic augmentation of imbalanced datasets with adversaries could possibly increase model's robustness to adversarial attacks

References

1. C. Billovits, M. Eric, and N. Agarwala, "Hitting depth: Investigating robustness to adversarial examples in deep convolutional neural networks," 2016.
2. I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.
3. A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," arXiv, Tech. Rep., 2016. [Online]. Available: http://arxiv.org/abs/1607.02533
4. N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. Berkay Celik, and A. Swami, "Practical black-box attacks against deep learning systems using adversarial examples," *arXiv preprint arXiv:1602.02697*, 2016.

5. J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset Shift in Machine Learning*. The MIT Press, 2009.
6. N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent data analysis*, vol. 6, no. 5, pp. 429–449, 2002.
7. B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016.
8. D. Lowd and C. Meek, "Adversarial learning," in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, ser. KDD '05. New York, NY, USA: ACM, 2005, pp. 641–647. [Online]. Available: http://doi.acm.org/10.1145/1081870.1081950
9. P. Laskov and R. Lippmann, "Machine learning in adversarial environments," *Machine Learning*, vol. 81, no. 2, pp. 115–119, 2010. [Online]. Available: http://dx.doi.org/10.1007/s10994-010-5207-6
10. S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE transactions on neural networks*, vol. 8, no. 1, pp. 98–113, 1997.
11. N. Papernot, "On the integrity of deep learning systems in adversarial settings," 2016.
12. N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," *arXiv preprint arXiv:1605.07277*, 2016.
13. C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.
14. N. Papernot, "Machine learning in adversarial settings," 2017. [Online]. Available: https://www.papernot.fr/files/16-mcdaniel-sp-machine-learning-in-adversarial-settings.pdf
15. J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in neural information processing systems*, 2014, pp. 3320–3328.
16. H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*. IEEE, 2008, pp. 1322–1328.
17. S. Barua, M. M. Islam, and K. Murase, *A Novel Synthetic Minority Oversampling Technique for Imbalanced Data Set Learning*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 735–744. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-24958-7_85
18. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.
19. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
20. Y. Dauphin, H. de Vries, and Y. Bengio, "Equilibrated adaptive learning rates for non-convex optimization," in *Advances in Neural Information Processing Systems*, 2015, pp. 1504–1512.
21. I. G. Y. Bengio and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org