CARE: Class Attention to Regions of Lesion for Classification on Imbalanced Data

Jiaxin Zhuang[*1] ZHUANGJX5@MAIL2.SYSU.EDU.CN
Jiabin Cai[*1] CAIJB5@MAIL2.SYSU.EDU.CN
Ruixuan Wang[1] WANGRUIX5@MAIL.SYSU.EDU.CN
Jianguo Zhang[2] J.N.ZHANG@DUNDEE.AC.UK
Weishi Zheng[1] WSZHENG@IEEE.ORG
[1] *School of Data and Computer Science , Sun Yat-sen University, China*
[2] *Computing, School of Science and Engineering, University of Dundee, UK*

Abstract

To date, it is still an open and challenging problem for intelligent diagnosis systems to effectively learn from imbalanced data, especially with large samples of common diseases and much smaller samples of rare ones. Inspired by the process of human learning, this paper proposes a novel and effective way to embed attention into the machine learning process, particularly for learning characteristics of rare diseases. This approach does not change architectures of the original CNN classifiers and therefore can directly plug and play for any existing CNN architecture. Comprehensive experiments on a skin lesion dataset and a pneumonia chest X-ray dataset showed that paying attention to lesion regions of rare diseases during learning not only improved the classification performance on rare diseases, but also on the mean class accuracy.

Keywords: Attention, Imbalanced Data, Small Samples, Skin Lesion, Pneumonia Chest X-ray.

1. Introduction

Deep learning has been widely applied to computer-aided diagnosis systems, particularly based on medical images (Shen et al., 2017). However, human-level performance of intelligent diagnosis often comes from training deep neural networks on large automated data. Therefore, current intelligent systems are mainly trained for the diagnosis of commonly encountered diseases. To date, due to the limited available data for rare diseases, it is still an open and challenging problem to train an intelligent system for diagnosis of both common and rare diseases. To solve this problem, the key is how to effectively handle the data imbalance between common diseases and rare ones.

Multiple approaches have been proposed to solve such data imbalance problems. One traditional approach is through over-sampling of the limited data for small-sample classes or down-sampling of the data for larger-sample classes(Chawla et al., 2002), thus generating similar number of training data between classes. Data augmentation, a default choice for training deep neural networks, can also be used as an over-sampling method to generate more data for small-sample classes. Another widely used approach is to improve the cost of mis-classifying each training example coming from small-sample classes, which can be easily realized by setting larger weights for small-sample classes in the loss function (Sun et al., 2007). Different from setting a single class weight for all training examples of the same class, another type of approach is to adaptively re-weight each single training

[*] Contributed equally

example based on the difficulty of being correctly classified, including boosting (Dollár et al., 2009; Freund and Schapire, 1999; Viola and Jones, 2001) and the recently proposed focal loss (Lin et al., 2017). Based on such weights, hard negative mining can be adopted to select just a subset of training data for the next-round training of classifier (Dalal and Triggs, 2005; Felzenszwalb et al., 2010; Fu et al., 2017; Shrivastava et al., 2016; Sung, 1995). Besides these approaches, particularly for medical image analysis, transfer learning via fine-tuning a pre-trained classifier has been proven helpful to improve performance for both large- and small-sample classes (Wang and Xia, 2018; Buda et al., 2017).

Different from these existing studies which consider each image as the basic unit and mainly focus on varying number and importance of images, this paper proposes a novel approach to data imbalance problems by delving into images and considering the high-level semantics of images, i.e, Class Attention to REgions of lesion (we termed our approach as CARE). Specifically, inspired by the process of human learning, attention was embedded into the learning process of neural network classifiers particularly for rare diseases. By attracting classifiers to pay more attention to the lesion regions during learning, the classifiers can learn more effectively from small samples. Due to limited training data for rare diseases, annotation of lesion regions (in the form of bounding boxes containing lesion regions) does not usually take much effort for radiologists and therefore is reasonably acceptable. Different from existing attention-relevant deep learning studies where attention is estimated as intermediate outputs of neural networks (Vaswani et al., 2017; Xu et al., 2015), the proposed approach provides a novel way to explicitly uses attention as part of supervision signal (in addition to image labels) to help train classifiers. What's more, the proposed attention embedding mechanism is independent of and does not alter neural network architectures, therefore can directly used as an element for any existing convolutional neural network architecture. In addition, the CARE approach is independent of any existing approach to data imbalance, and therefore can be combined to handle the imbalance problem together. Experiments with multiple different neural network architectures on a skin lesion dataset and a pneumonia chest X-ray dataset showed that paying attention to lesion regions of rare diseases during learning did improve the classification performance on rare diseases. Compared to existing approaches, addition of the CARE approach always further improved performance.

2. The CARE Approach

Medical students are often pointed at the regions lesions containing distinct characteristics of certain diseases when taught to diagnose diseases via medical images (Krupinski, 2010). With the help of such attentions to lesion regions, students probably can more effectively learn to grasp the distinct properties of each disease even with a small sample of medical images. Inspired by the learning process of humans, here we propose a simple but effective method to embed attention into the learning process of deep neural network classifiers for intelligent diagnosis.

We hypothesize that appropriate attention during learning would help neural network classifiers more effectively learn from small samples particularly for rare diseases. Suppose the lesion regions of interest have been provided in advance for model learning, in the form of bounding boxes containing lesions. The effort of providing bounding boxes is feasible for rare diseases because quite often only a small sample of images are available for each category. Then, if there is one way to estimate the local regions on which the classifier focuses during image diagnosis, by enforcing that such 'visual focus' of the classifier falls into the bounded lesion regions, attention would be

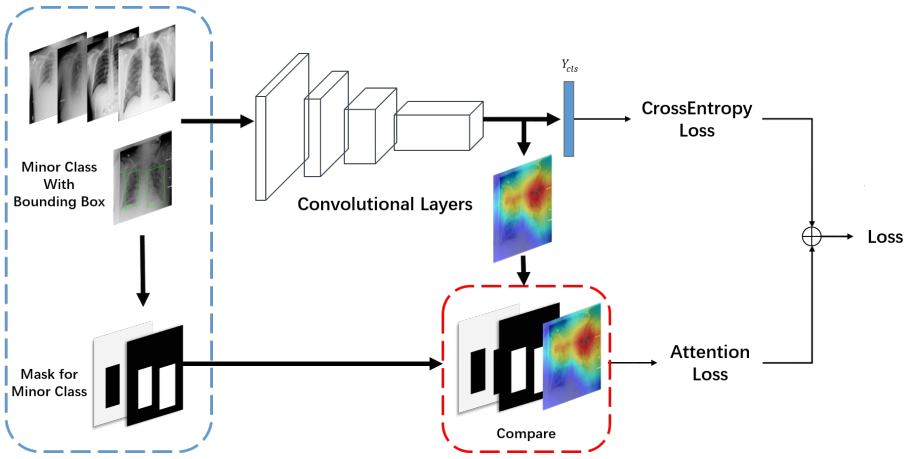

Figure 1: The diagram of the proposed CARE framework with the attention loss. It contains two branches with one focusing on attention into lesion region in minority class. Bounding boxes are provided only for minority classes during training.

naturally embedded during learning. Fortunately, such 'visual focus' of a classifier on any input image can be conveniently estimated by a recently proposed visualization approach called Grad-CAM (Selvaraju et al., 2016). Given a well trained classifier and an input image, Grad-CAM can provide a class-specific feature activation map in which regions with higher activation contribute more to the classifier's output prediction being the specific class. Therefore, if the classifier attends to only the box-bounded regions when diagnosing an image, the high activation regions from the Grad-CAM should also be within the bounded regions. In this sense, the spatial relationship (e.g., degree of overlap) between the high activation regions and the bounded image regions can be used to measure how well the classifier has attended to the bounded image regions.

Denote by L_a the discrepancy between the high activation regions from Grad-CAM and the bounded image regions over all training data, then embedding attention during classifier learning can be realized by minimizing a new loss L for the classifier,

$$L = (1-\alpha)L_c + \alpha L_a \tag{1}$$

where L_c is the general cross-entropy loss for the network classifier, and L_a, called *attention loss*, helps to drive the network to attend to box-bounded image regions during training. α is a coefficient to balance the two loss terms. Considering the different influence of the inside-box and outside-box regions, the attention loss is further split into two items by

$$L_a = L_{in} + \lambda L_{out} \tag{2}$$

where the inner loss L_{in} helps the classifier increase the attention inside the bounding box, and the outer loss L_{out} helps the classifier decrease the attention outside the bounding box. λ is a coefficient to balance the two loss terms. In detail, for any training image with bounding box(es) provided, let M_{in} denote a binary complement image in which all pixels inside bounding box are set to 1 and others to 0, and in contrast, M_{out} denote a binary mask image in which all pixels inside bounding box are set to 0 and any pixel outside box is set to either 1 or a positive value relevant to the distance

between the pixel and the bounding box. Let F denote the feature activation map from Grad-CAM for the training image based on current classifier. Then L_{in} and L_{out} (for one training image) can be defined as

$$L_{in} = -\min\left(\frac{\Sigma_{i,j} M_{in}(i,j) \cdot F(i,j)}{\Sigma_{i,j} M_{in}(i,j)}, \tau\right) \tag{3}$$

$$L_{out} = \frac{\Sigma_{i,j} M_{out}(i,j) \cdot F(i,j)}{\Sigma_{i,j} M_{out}(i,j)} \tag{4}$$

Here, $M_{in}(i,j)$ represents the value at the position (i,j) in the mask M_{in}, and similarly for $M_{out}(i,j)$ and $F(i,j)$. Equation (4) represents the strength of feature activation outside the bounding box, while Equation (3) would penalize the classifier if the highly activated area inside the bounding box is not large enough (i.e., when the percent of weighted activated area $\frac{\Sigma_{i,j} M_{in}(i,j) \cdot F(i,j)}{\Sigma_{i,j} M_{in}(i,j)}$ is smaller than a predefined threshold τ). Note that for notation simplicity, Equations (3) and (4) are just for one single image. In fact, during training, the loss terms are calculated and averaged over all training images.

One advantage of the proposed attention-based approach is its independence of model structures. Therefore the CARE can be directly embedded to the training processing of any existing CNN classifiers, without alternating their model architectures. Also, the CARE framework is independent of existing approaches to handling data imbalance, therefore can be directly combined to further improve classification performance.

3. Experiment

3.1. Experimental settings

Dataset. Two medical image datasets were used to evaluate the proposed approach. One is the skin dataset provided by the ISIC2018 Challenge with 7 disease categories (Codella et al., 2017), in which 6705 images are for Melanocytic nevus and only 115 images for Dermatofibroma, clearly having serious data imbalance between classes. One bounding box was generated for each image of the rare disease Dermatofibroma by one of the authors and confirmed by a dermatologist. The other is the pneumonia detection X-ray dataset with 3 categories [1], including 8,851 'Normal' images, 6012 'Lung Opacity' images, and images of 'No Lung Opacity/Not Normal'. Each 'Lung Opacity' image was provided with one or multiple bounding boxes indicating the region of the pneumonia. Although the original objective of this chest X-ray data is for lesion detection, we used it for 3-class classification, with the ground-truth bounding boxes used to evaluate the proposed approach. The number of 'Lung Opacity' images is much smaller than other two categories, being considered as the minority class in a data imbalance scenario. All images were resized to 224×224 pixels, with bounding boxes resized accordingly for the small-sample class in each dataset. For each dataset, images are randomly split into training set (80%) and test set (20%) with stratification.

Implementation and Protocol. In the experiments, the training of CARE is divided into two stages. At the first stage, each backbone CNN classifier (i.e., the branch without the attention loss in Figure 1) used was pretrained first on ImageNet and then on the training set without the attention loss. The training at this stage is stopped when the cross-entropy loss does not decrease any more (normally within 200 epochs in our experiment). At the second stage, the attention loss was included

1. The original dataset comes from https://www.kaggle.com/c/rsna-pneumonia-detection-challenge, and part of the dataset was extracted for the purpose of evaluation on data imbalance.

to fine-tune the stage-one classifier with the training set. α in Equation(1) was set to 0.5 unless otherwise mentioned. Adam optimizer was used throughout, with initial learning rate set at 0.0001. λ was empirically set to 0.5 for X-ray dataset and 0.7 for Skin Lesion dataset. Specially, we form batches by randomly sampling P classes In testing, each test image (without any bounding box) was fed to the backbone CNN classifier for prediction. Note that the CARE approach needs no bounding box in testing. *Recall* for the small-sample class (i.e., Dermatofibroma in skin dataset, and Lung Opacity in the pneumonia dataset) and *mean class accuracy* (MCA, i.e., average recall over all classes) were reported as measurement of the model performance.

3.2. Results

Baseline and Comparisons. In order to test the effectiveness of the proposed approach, we compared CARE to three widely-used strategies for handling data imbalance, namely, 1) *cost sensitive learning* denoted by CSL (Sun et al., 2007), 2) *focal loss*(Lin et al., 2017) denoted by FL, a representative method of hard negative mining, and 3)*data augmentation* (including rotation, flip and color jitter) denoted by DA. We further tested our approach by embedding CARE into the three strategies, resulting methods of CARE+CSL, CARE+FL, and CARE+DA. We also tested a baseline without using the visual attention branch in Figure 1. Table 1 shows the comparison results on the pneumonia and skin datasets. It can be observed that CARE outperforms the baseline significantly in terms of both *recall* and *MCA*, in particular with a large margin on recall for the small-sample class (31.12 vs. 7.41 on the pneumonia dataset, and 52.17 vs. 47.83 on the skin set). All the three strategies (i.e., CSL, FL and DA) perform better than the baseline without any treatment of data imbalance, which is expected. It is worth highlighting that adding CARE to each of CSL, FL or DA could boost the performances significantly w.r.t the use of each method alone; for instance, the recall (and MCA) of CARE+CSL is 45.04 (and 65.23), significantly better than CSL only. For the CSL method, additional experiments showed that varying loss coefficients for the minority class did not change the finding, i.e., CARE+CSL always performs better than CSL alone. This clearly indicates that our approach is capable of boosting their performances significantly when plugged into the existing strategies for handling data imbalance.

Flexibility with Architecture. Our proposed CARE framework is independent of model structures. To show this, we test variants of our CARE framework built with two different widely-used CNN architectures: ResNet (He et al., 2015) and VGG19 (Simonyan and Zisserman, 2014). For ResNet, we further test different number of layers at 18, 50 152, from shallow to very deep. VGG19 uses a 19-layered structure. Thus in total we have four backbones of CNN architectures: ResNet18, ResNet50, ResNet152 and VGG19. For each of the backbones, we compare the performance of the resulting CARE model (X(CARE): x represents the name of the backbone) and the orginal backbone network; for example VGG19 vs. VGG19+CARE. Results are shown in Table 2. It can be observed that different *orginal* backbone architectures perform differently, among which VGG-19 performs the best. For each of the backbone, its CARE version outperforms than the orginal network in terms of both recall and MCA, with recall significantly better.

Tolerance to Bounding Box. It is noted that the training of our model needs the bounding box annotations. For many rare or uncommon diseases (such as Dermatofibroma studied in this paper), the annotation effort of bounding boxes (bbox) for the lesion regions in the minority class is usually very small compared to that of accurate boundary pixel-level annotations. Even though, there might exist inter- or intra-observer variations of annotations. The bbox used thus far is tightly around the

Table 1: Comparison on pneumonia dataset and the skin dataset using ResNet50, including baseline, CARE (ours), CSL (cost sensitive learning), FL (focus loss), DA(data augmentation), CARE+CSL (ours), CARE+FL (ours) and CARE+DA(ours). MCA is the mean class accuracy, and recall is reported for the minority class (Lung Opacity/Dermatofibroma).

	Penuomia Dataset		Skin Dataset	
Model	recall(%)	MCA(%)	recall	MCA(%)
baseline	7.41	56.77	47.83	75.75
CARE (ours)	**31.12**	**63.29**	**52.17**	**76.16**
CSL	11.11	57.88	61.91	80.21
CARE+CSL (ours)	45.04	65.23	**65.22**	**81**
FL	11.14	58.41	38.3	72.72
CARE+FL (ours)	**49.44**	**66.72**	**40.28**	**74.06**
DA	20.06	59.64	56.62	54.41
CARE+DA(ours)	**45.18**	**65.97**	**60.32**	**56.22**

Table 2: Results of CARE with different backbones on the pneumonia and skin datasets. X(CARE) denotes the CARE model built with backbone X; for instance, VGG19(CARE) represents the CARE model with the backbone model VGG19. Note that all models in the table apply CSL.

	Pneumonia Dataset		Skin Dataset	
Model	recall(%)	MCA(%)	recall	MCA(%)
ResNet18	15.16	57.76	58.6	73.71
ResNet18(CARE)	**25.51**	**58.76**	**59.31**	**74.06**
ResNet50	11.11	57.88	61.91	80.21
ResNet50(CARE)	**45.04**	**65.23**	**65.22**	**81**
ResNet152	11.37	59.11	61.19	80.15
ResNet152(CARE)	**31.31**	**63.78**	**72.19**	**81.93**
VGG19	25.93	61.72	52.07	**74.48**
VGG19(CARE)	**41.24**	**64.3**	**56.52**	72.81

lesion region, which requires the larger annotation effort than other cases of using bbox. To relax this requirement, we vary the bbox by scaling at 0.7, 0.9, 1.0, 1.1 and 1.3, and test the robustness of our approach to such a scaling. Table 3 shows the performance of our model at different scaling. It can be seen that the performance remains stable within a reasonable range, for instance, from 0.9 till 1.3. This indicates that our approach provides certain tolerance to the size of bbox, i.e., the bbox does not need to be tightly around the lesion, with flexibility of using a looser bounding box, which requires less annotation effort.

Effect of α. We further conducted a set of experiments to test the effect of α using ResNet50(CARE), and the results are shown in Table 4. It could be observed that performances of model remain stable

Table 3: Robustness to BBox scaling, tested with ResNet50(CARE). Note that all models in the table apply CSL.

	Pneumonia Dataset		Skin Dataset	
Model	recall(%)	MCA(%)	recall	MCA(%)
without CARE	11.11	57.88	61.91	80.21
0.7	43.56	**66.34**	63.71	80.20
0.9	**50.76**	66.27	64.21	80.47
1.0	45.04	65.23	**65.22**	**81**
1.1	38.43	63.14	65.22	80.72
1.3	46.06	65.14	65.22	80.68

within a reasonable range, for instance, from 0.1 to 0.9, which indicates that our model is insensitive to the choices of the value of α. (Jia:)

Table 4: Effect of α using ResNet50(CARE) on Pneumonia set (left) and Skin set (right). Note that all models in the table apply CSL.

	Pneumonia Dataset		Skin Dataset	
Model	recall(%)	MCA(%)	recall	MCA(%)
α=0	11.11	57.88	61.91	80.21
α=0.1	28.77	63.84	55.22	78.80
α=0.3	38.34	64.35	**65.22**	79.94
α=0.5	45.04	**65.23**	65.21	**80.33**
α=0.7	42.58	64.34	65.18	80.18
α=0.9	**50.67**	65.2	65.12	80.24

Visual Insight. To show the effect of the proposed attention loss, we visualize the classification activation maps of the minority class from both datasets, specifically, from the minority class of Lung Opacity, and the minority class of Dermatofibroma. Fig. 2 shows the activation maps of two sample images from each of the classes respectively. For clarity, we also superimpose (ground truth) bounding boxes highlighting the lesion regions on the test images, provided along with the dataset (the Pneumonia dataset) or in-house annotated (the skin dataset). Note that we did *not* use any of those bounding box in the testing, but here merely for visualization purpose. It can be observed that the activated regions (red regions in middle row) without using the proposed attention loss (Eq. 2) clearly deviated from lesion regions, while those (last row) produced by CARE localized the lesion regions well. These results on the test images reveal that the CARE model could have learned to focus on lesion regions when analyzing new images, through attention loss optimized during training.

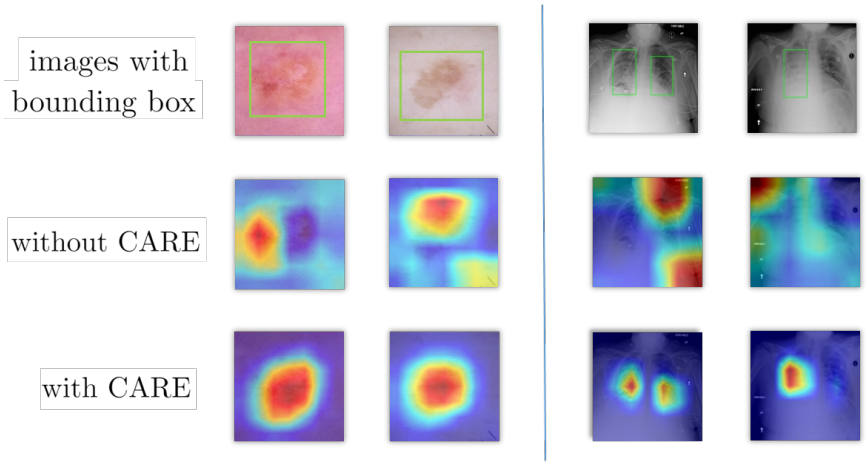

Figure 2: Visualization of activation maps with and without using CARE. Upper row: original images superimposed with bounding boxes; Middle and bottom rows: activation maps without and with using CARE attention loss, respectively. All the class activation maps are generated using Grad-CAM (best viewed in color).

4. Conclusion

We have introduced a novel approach called CARE to embed attention mechanism into CNN learning process, which effectively focuses on minority in the case of data imbalance. This approach, combining Grad-CAM localization and bounding box in minor class indicating the lesion region, is applicable for any CNN based classifier without altering neural network architectures. A series of experiments on the skin and the pneumonia imbalanced datasets have shown our approach can help classifier pay attention to lesion region of rare disease particularly and effectively learn characteristics of diseases from imbalanced data. Our model is effective and can be used to boost the performances of existing strategies of handling data imbalance such as cost sensitive learning, data augmentation, or focal loss.

Acknowledgments

This work is supported in part by the National Key Research and Development Plan (grant No. 2018YFC1315402), Royal Society International Exchanges grant (No. 170168), and by the NSFC (grant No. 61628212).

References

Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *CoRR*, 2017.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, pages 321–357, 2002.

Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin K. Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging, hosted by the international skin imaging collaboration (ISIC). *CoRR*, 2017.

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

Piotr Dollár, Zhuowen Tu, Pietro Perona, and Serge J. Belongie. Integral channel features. In *BMVC*, 2009.

Pedro F. Felzenszwalb, Ross B. Girshick, and David A. McAllester. Cascade object detection with deformable part models. In *CVPR*, 2010.

Yoav Freund and Robert E. Schapire. A short introduction to boosting. In *IJCAI*, 1999.

Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C. Berg. DSSD : Deconvolutional single shot detector. *CoRR*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, 2015.

Elizabeth A. Krupinski. Current perspectives in medical image perception. *Attention, Perception, & Psychophysics*, 72(5):1205–1217, 2010.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, 2017.

Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, 2016.

Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, pages 221–248, 2017.

Abhinav Shrivastava, Abhinav Gupta, and Ross B. Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, pages 761–769, 2016.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, 2014.

Yanmin Sun, Mohamed S. Kamel, Andrew K.C. Wong, and Yang Wang. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, pages 3358 – 3378, 2007.

Kah Kay Sung. *Learning and example selection for object and pattern detection*. PhD thesis, MIT, 1995.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, 2017.

Paul A. Viola and Michael J. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.

Hongyu Wang and Yong Xia. Chestnet: A deep neural network for classification of thoracic diseases on chest radiography. *CoRR*, 2018.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.