CASED: Curriculum Adaptive Sampling for Extreme Data Imbalance

Andrew Jesson[✉], Nicolas Guizard, Sina Hamidi Ghalehjegh, Damien Goblot,
Florian Soudan, and Nicolas Chapados

Imagia Cybernetics Inc., Montreal, QC, Canada
{andrew.jesson,nicolas.guizard,sina.hamidi,damien.goblot,
florian.soudan,nicolas.chapados}@imagia.com
http://www.imagia.com

Abstract. We introduce CASED, a novel curriculum sampling algorithm that facilitates the optimization of deep learning segmentation or detection models on data sets with extreme class imbalance. We evaluate the CASED learning framework on the task of lung nodule detection in chest CT. In contrast to two-stage solutions, wherein nodule candidates are first proposed by a segmentation model and refined by a second detection stage, CASED improves the training of deep nodule segmentation models (e.g. UNet) to the point where state of the art results are achieved using only a trivial detection stage. CASED improves the optimization of deep segmentation models by allowing them to first learn how to distinguish nodules from their immediate surroundings, while continuously adding a greater proportion of difficult-to-classify global context, until uniformly sampling from the empirical data distribution. Using CASED during training yields a minimalist proposal to the lung nodule detection problem that tops the LUNA16 nodule detection benchmark with an average sensitivity score of 88.35%. Furthermore, we find that models trained using CASED are robust to nodule annotation quality by showing that comparable results can be achieved when only a point and radius for each ground truth nodule are provided during training. Finally, the CASED learning framework makes no assumptions with regard to imaging modality or segmentation target and should generalize to other medical imaging problems where class imbalance is a persistent problem.

Keywords: Lung cancer · Computer aided detection · Nodule detection · Curriculum learning · Data imbalance · 3D convolutional neural networks

1 Introduction

Death rates attributed to lung cancer are three times higher than for any other cancer in the United States [16]. Diagnosis of this pathology is informed by the presence of malignant pulmonary nodules that appear in thoracic computed tomography (CT) images [6]. There is a current trend toward regular monitoring programs of high-risk groups using methods such as low-dose CT [19]. This

has been proposed to help catch the pathology in its early stages where, in developed countries, diagnosis dramatically increases the 5-year patient survival rate by 63–75% [19]. It is likely that radiologists who are tasked with locating and classifying pulmonary nodules would see a dramatic increase in workload with the saturation of such protocols. Fast and accurate automated lung nodule detection methods would then improve lung image evaluation throughput and objectivity by assisting radiologists in their assessment.

One of the major challenges in designing effective automated lung nodule detection methods is the massively unbalanced nature of the data. For example, over the entire Lung Image Database Consortium image collection (LIDC-IDRI) [2,3,5] less than 1% of image voxels contain positive nodule examples. The class imbalance problem has received wide attention in the machine learning and data mining communities, where typical solutions include class over- and under-sampling, weighted losses, and posterior probability recalibration [9]. Sampling schemes have been studied in medical imaging classification (e.g. [7] and references therein) and segmentation [8], whereas loss function adjustments were key to results in [12]. In Computer-Aided Detection (CADe) applications, specialized knowledge can be used, such as limiting the domain of detection to the lung only (requiring a lung masking model) [19], or training a highly sensitive candidate nodule screening model and then refining predictions by cascading false positive reduction stages [13,19]. A common theme across these approaches is that they tend to be problem-dependent, and sizable efforts must often be expended to find the balancing technique yielding the best performance.

This paper proposes a generic approach to tackle class imbalance, by using, during training, an online adaptation of the distribution of majority and minority class examples, in the spirit of curriculum learning [4]. The *Curriculum Adaptive Sampling for Extreme Data imbalance* (CASED) is a novel sampling curriculum that allows for a 3D fully convolutional network (FCN) to yield segmentations high enough in quality to make detection a mere consequence. In contrast to approaches where an off-the-shelf segmentation model [14] or FCN [10] is trained to only provide candidates to a second, independently-trained convolutional neural network (CNN) for classification, CASED combines curriculum learning and adaptive data sampling in a way that makes the second classifier redundant. This is achieved by allowing the FCN to first learn how to distinguish nodules from their immediate surroundings while continuously introducing training examples that the model has trouble classifying. This approach yields a surprisingly minimalist proposal to the lung nodule detection problem that tops the LUNA16 challenge [1] leader-board with a score of *88.35%*. Furthermore, weakly-supervised training, with only a point and radius provided for each training nodule, yields results competitive with those of full segmentation.

2   Method

CASED adheres to the observation that the solution to object detection is fully contained in the solution to object segmentation. That is, given an ideal segmentation, a determination of the location, extent, and identity of an imaged object

becomes trivial. However, training a model to yield even acceptable medical image segmentations is a considerably harder task than detection for two main reasons. First, manual segmentation of training data is a laborious and expensive endeavour. And second, the model must be able to describe the complex variations of texture ranging over the extent of a given object and its surroundings. Fortunately, the first problem is less significant here as large datasets of annotated lung CT scans are available [3]; however, robustness to weakly labeled data is important. Regarding the second problem, recent work on FCNs (e.g. FCN-8s for natural images [10], U-Net for biomedical images [12]) has shown that their ability to model multi-scale context over finite image regions makes them ideal candidates for medical image segmentation problems. It behooves one to ask then, in the context of lung nodule detection, why has it not yet been shown that FCNs alone are a competitive solution to this problem? We hypothesize the answer lies in the extreme data imbalance associated to the problem, which has not yet been sufficiently addressed. In the following we present CASED as an approach to overcome this issue.

Curriculum. One of the more attractive properties of FCNs is their ability to handle images of arbitrary size. This feature allows us to reduce data imbalance by training on small image patches where the output stride of the model contains at least one positive nodule voxel. As one would start teaching a child to read the alphabet by restricting their gaze to a large letter A, the model first learns how to represent nodules given only their immediate surroundings. An important consequence of training the FCN on image patches is that we are able to randomize training examples across both patient images and also image regions. Training only on patches that contain nodule examples will result in an extremely sensitive model but with low specificity because it would not learn how to represent the majority of the input image space. Therefore, a curriculum [4] is introduced where the proportion of training patches that contain nodules to those that do not is decreased according to a schedule that tends toward the data distribution as the number of training examples seen approaches infinity.

Adaptive Sampling. After training the FCN using this curriculum with random sampling of background patches, it generally converges to a solution that still gives systematic and predictable false positives. Furthermore, the vast majority of voxels in typical lung images are correctly and confidently predicted as non-nodule, so random sampling would be far more likely to show examples that would have little to no effect on loss optimization. Hence, we introduce a sampling strategy that favours training examples for which prediction using recent model parameters produces false results, an instance of *hard negative mining* (HNM) [17].

Figure 1 shows a flowchart of the CASED framework. Let $\{x_i\}$ be a training set of M patches. Patch generators are shown in red boxes. The generators g_r and g_n represent distributions over the set of all patches and the set of patches that contain nodules, respectively. FCN models are shown in blue boxes where the training model shares its weights with a predictor that is run in parallel
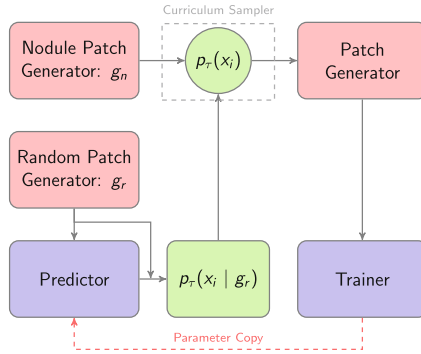
Fig. 1. Schematic diagram of CASED framework

for the purposes of HNM. The green boxes represent samplers with distributions that vary with the mini-batch iteration τ. The sampler $p_\tau(x_i \mid g_r)$ selects patches based on both τ and the training loss $\mathcal{L}_\tau(x_i)$. The function $f_r(\mathcal{L}_\tau(x_i), \tau)$ specifying $p_\tau(x_i \mid g_r)$ must be on the range $[0, 1]$ and $f_r(\mathcal{L}_\tau(x_i), \tau) \rightarrow M^{-1}$ as $\tau \rightarrow \infty$. The sampler $p_\tau(x_i)$ defines the curriculum and chooses between g_r and g_n according to a mixing that depends on τ. The mixing coefficient $p_\tau(g_n)$ is specified by $f_n(\tau)$ with range $[0, 1]$ and convergence to M^{-1} as $\tau \rightarrow \infty$. The distribution governing the sampler $p_\tau(x_i)$ is given by

$$
\begin{aligned}
p_\tau(x_i) &= p_\tau(x_i \mid g_r)(1 - p_\tau(g_n)) + p(x_i \mid g_n)p_\tau(g_n) \\
&= p_\tau(x_i \mid g_r) + (p(x_i \mid g_n) - p_\tau(x_i \mid g_r))\, p_\tau(g_n),
\end{aligned}
\tag{1}
$$

where $p(x_i \mid g_n) = 1$ if x_i contains a nodule, and 0 otherwise. In the limit, as τ goes to infinity, $p_\tau(x_i)$ converges to a uniform distribution over x_i, which makes CASED a valid curriculum [4].

3   Data and Implementation

We study CASED as applied to the task of lung nodule detection using the publicly available LIDC image collection [2,3,5]. The LIDC contains 1010 patients and a total of 1018 clinical thoracic CT scans. Each scan has been analyzed through a two-phase nodule annotation process by four expert radiologists. In the first phase each radiologist independently marks nodules as belonging to one of three classes (*nodule < 3*mm, *nodule ≥ 3*mm, and *non-nodule ≥ 3*mm), where the measurement refers a nodule's diameter. In the second phase, each expert can refine their annotations after seeing the anonymous annotations of the other three radiologists. The LIDC contains 2635 nodules annotated in this way and there are 142 cases that either contain no detected nodules or *nodule < 3*mm.

For segmentation we use a 3D U-Net architecture, based on the model proposed in [12]. Figure 2 illustrates the model used. The model is comprised of three
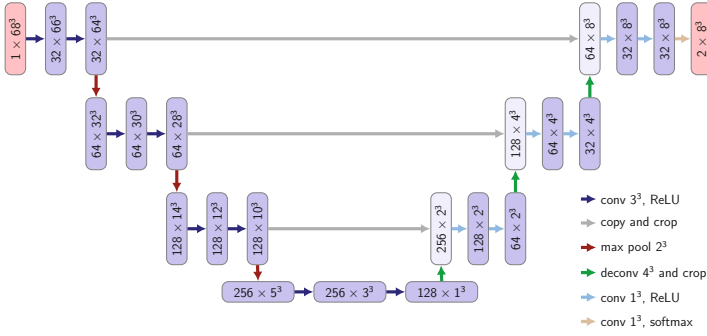
Fig. 2. Schematic diagram of our 3D U-Net-based architecture.

distinct components: (1) downstream feature extraction path, (2) upstream feature pooling path, and (3) linear pixel classifier. In the downstream path, we use layers of "convolution" and "pooling". Each layer effectively encodes a progressively larger image neighbourhood of the input image as we go deeper. In the upstream path, we use layers of "convolution" and "strided transposed convolution" layers. Multi-scale features extracted in the downstream path are combined to provide pixel-level features in the input image space. Finally, the linear pixel classifier uses a simple "sigmoid" layer to provide per-pixel prediction of nodule or non-nodule.

CASED training requires minimal data preprocessing. For a given CT scan, image intensities are transformed to Hounsfield units and linearly rescaled. The scan is then resized to 1.25 mm isotropic voxels. For training, binary segmentation maps are built from the expert annotations listed in the provided XML files and are also transformed into the 1.25 mm isotropic space. The binary segmentation maps are nodule-wise refined to only label as nodule those voxels that correspond to the intersection of all available annotations. For example, if a nodule only has an annotation from one rater, that annotation is used; however, if a nodule has annotations from multiple raters the intersection of those annotations is used.

Training is done by optimizing voxel-wise binary cross-entropy over each prediction patch (of size 8^3) and its corresponding reference segmentation using stochastic gradient descent with Nesterov momentum. We use mini-batches with 16 image patches of size 68^3 as input. Nodule patches are defined as those for which there is a labeled nodule voxel within the 8^3 output stride. All other patches are called background. The curriculum is initialized with $p_\tau(g_n) = 1.0$ and is decayed after each mini-batch iteration. Finally, "background" patches are sampled based on whether they contain a false positive prediction using recent model parameters.

At test time, an equally minimalist approach to postprocessing is required. Given a test image, the model outputs a soft segmentation map estimating the

Table 1. The LUNA16 cross-validation sensitivity at different number of false positives per scan. The scores for other methods are taken from the result section of LUNA16 website [1]. Method with asterisk superscript does not provide any description on LUNA16 scoreboard.

| Method	False positives per scan							Average
	0.125	0.25	0.5	1	2	4	8	
CASED	0.759	0.825	0.866	0.903	0.926	0.946	0.959	**0.883**
CASED-Sphere	0.701	0.762	0.813	0.862	0.897	0.923	0.939	0.842
AIDENCE*	0.739	0.788	0.828	0.879	0.910	0.938	0.963	0.864
ZNET [15]	0.661	0.724	0.779	0.831	0.872	0.892	0.915	0.811
ETROCAD [15,18]	0.250	0.522	0.651	0.752	0.811	0.856	0.887	0.676
M5LCAD [11,15]	0.306	0.360	0.540	0.691	0.762	0.797	0.798	0.608

probability that a given voxel belongs to the nodule class. This map is thresholded giving a binary segmentation on which connected component analysis is performed to yield candidate nodules. The center of mass and average value of the segmentation map over each candidate is found to yield a list of point and confidence predictions. The points are finally transformed back into the native image space. Because the model is fully convolutional, the input size at test need only be divisible by 8. Given sufficient GPU memory the entire CT scan can be passed as input without tiling and full prediction takes only a few seconds.

4   Experiments and Results

We evaluate the CASED framework using the 2016 Lung Nodule Analysis Challenge (LUNA16) 10-fold cross-validation split [1]. Each fold contains 88–89 CT scans. The reference standard for LUNA16 consists of all *nodule ≥ 3*mm that have been detected by at least three of four raters. Evaluation is based on the detection sensitivity at various false positive rates per scan. A detailed explanation of the evaluation can be found on the LUNA16 website [1].

For each test fold, we train on eight and validate on one of the remaining folds. We also use model ensembling to improve the reliability of the results. Finally, we repeat the experiment using spherical segmentations defined by the location and radius of each nodule instead of the reference annotations (CASED-Sphere).

Table 1 summarizes the results of these experiments for the lung nodule detection task and provides a comparison to the results of other methods submitted to the LUNA16 leader board. The CASED learning framework shows a 8.9% relative increase in average sensitivity over the best published results for a given model, ZNET [15]. The free-response receiver operating characteristic (FROC) curve for CASED appears in Fig. 3. Finally, we demonstrate robustness to segmentation quality by showing that a 3.8% relative increase over ZNET is achieved with CASED-Sphere.
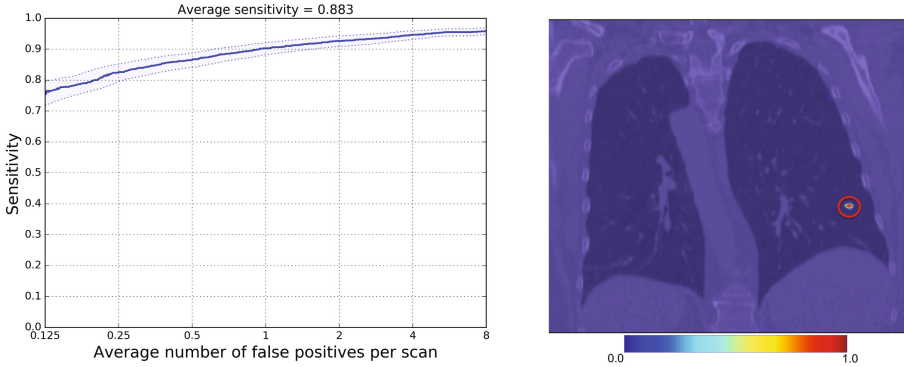
Fig. 3. Left: The free-response receiver operating characteristic (FROC) curve for CASED. The blue line and shaded area represent the mean and variance of the nodule detection sensitivity over 1000 bootstrapped samples at different false positive rates. **Right:** Lung CT overlaid by probability map. In the color spectrum, as we move toward right (red) the probability of being nodule increases.

5   Conclusions

This paper proposes CASED, a new curriculum sampling algorithm for the highly class imbalanced problems that are endemic in medical imaging applications. We demonstrate that CASED is a robust learning framework for training deep lung nodule detection models. Evaluated on the LUNA16 challenge, we achieve the current state-of-the-art leader-board performance with an average sensitivity score of 88.35%. Since the CASED algorithm does not require any assumption on image modality, it can be applied to any arbitrarily large dataset wherein the unbalanced nature of data poses major problems for designing automated methods.

References

1. Lung nodule analysis 2016. https://luna16.grand-challenge.org. Accessed 22 Feb 2017
2. Armato, S.G., McLennan, G., et al.: The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Med. Phys. **38**(2), 915–931 (2011). http://dx.doi.org/10.1118/1.3528204
3. Armato, S.G., McLennan, G., et al.: Data From LIDC-IDRI. The Cancer Imaging Archive (2015). http://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX
4. Bengio, Y., Louradour, J., et al.: Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 41–48. ACM (2009)
5. Clark, K., Vendt, B., et al.: The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J. Digital Imaging **26**(6), 1045–1057 (2013). http://dx.doi.org/10.1007/s10278-013-9622-7

6. Diederich, S., Lentschig, M., et al.: Detection of pulmonary nodules at spiral CT: comparison of maximum intensity projection sliding slabs and single-image reporting. Eur. Radiol. **11**(8), 1345–1350 (2001). http://dx.doi.org/10.1007/s003300000787

7. Dubey, R., Zhou, J., et al.: Analysis of sampling techniques for imbalanced data: an $N = 648$ ADNI study. Neuroimage **87**, 220–241 (2014)

8. Havaei, M., Davy, A., et al.: Brain tumor segmentation with deep neural networks. Med. Image Anal. **35**, 18–31 (2017). http://www.sciencedirect.com/science/article/pii/S1361841516300330

9. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. **21**(9), 1263–1284 (2009)

10. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)

11. Lopez Torres, E., Fiorina, E., et al.: Large scale validation of the M5L lung CAD on heterogeneous CT datasets. Med. Phys. **42**(4), 1477–1489 (2015). http://dx.doi.org/10.1118/1.4907970

12. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). doi:10.1007/978-3-319-24574-4_28

13. Setio, A.A.A., Jacobs, C., et al.: Automatic detection of large pulmonary solid nodules in thoracic CT images. Med. Phys. **42**(10), 5642–5653 (2015). http://dx.doi.org/10.1118/1.4929562

14. Setio, A.A.A., Ciompi, F., et al.: Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. IEEE Trans. Med. Imaging **35**(5), 1160–1169 (2016)

15. Setio, A.A.A., Traverso, A., et al.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. arXiv preprint arXiv:1612.08012 (2016)

16. Siegel, R., Naishadham, D., Jemal, A.: Cancer statistics, 2013. CA: Cancer J. Clin. **63**(1), 11–30 (2013). http://dx.doi.org/10.3322/caac.21166

17. Sung, K.K., Poggio, T.: Example-based learning for view-based human face detection. IEEE Trans. Pattern Anal. Mach. Intell. **20**(1), 39–51 (1998). http://dx.doi.org/10.1109/34.655648

18. Tan, M., Deklerck, R., et al.: A novel computer-aided lung nodule detection system for CT images. Med. Phys. **38**(10), 5630–5645 (2011). http://dx.doi.org/10.1118/1.3633941

19. Valente, I.R.S., Cortez, P.C., et al.: Automatic 3D pulmonary nodule detection in CT images: a survey. Comput. Methods Programs Biomed. **124**, 91–107 (2016)