Curriculum Learning with Diversity for Supervised Computer Vision Tasks

Petru Soviany[1]

Abstract. Curriculum learning techniques are a viable solution for improving the accuracy of automatic models, by replacing the traditional random training with an easy-to-hard strategy. However, the standard curriculum methodology does not automatically provide improved results, but it is constrained by multiple elements like the data distribution or the proposed model. In this paper, we introduce a novel curriculum sampling strategy which takes into consideration the diversity of the training data together with the difficulty of the inputs. We determine the difficulty using a state-of-the-art estimator based on the human time required for solving a visual search task. We consider this kind of difficulty metric to be better suited for solving general problems, as it is not based on certain task-dependent elements, but more on the context of each image. We ensure the diversity during training, giving higher priority to elements from less visited classes. We conduct object detection and instance segmentation experiments on Pascal VOC 2007 and Cityscapes data sets, surpassing both the randomly-trained baseline and the standard curriculum approach. We prove that our strategy is very efficient for unbalanced data sets, leading to faster convergence and more accurate results, when other curriculum-based strategies fail.

1 Introduction

Although the accuracy of automatic models highly increased with the development of deep and very deep neural networks, an important and less studied key element for the overall performance is the training strategy. In this regard, Bengio et al. [2] introduced curriculum learning (CL), a set of learning strategies inspired by the way in which humans teach and learn. People learn the easiest concepts at first, followed by more and more complex elements. Similarly, CL uses the difficulty context, feeding the automatic model with easier samples at the beginning of the training, and gradually adding more difficult data as the training proceeds.

The idea is straightforward, but an important question is how to determine whether a sample is easy or hard. CL requires the existence of a predefined metric which can compute the difficulty of the input examples. Still, the difficulty of an image is strongly related to the context: a big car in the middle of an empty street should be easier to detect than a small car, parked in the corner of an alley full of pedestrians. Instead of building hand-crafted models for retrieving contextual information, in this paper, we use the image difficulty estimator from [12] which is based on the amount of time required by human annotators to assess if a class is present or not in a certain image. We consider that people can understand the full context very accurately, and that a difficulty measure trained on this information can be useful in our setting.

The next challenge is building the curriculum schedule, or the rate at which we can augment the training set with more complex information. To address this problem, we follow a sampling strategy similar to the one introduced in [28]. Based on the difficulty score, we sample according to a probability function, which favors easier samples in the first iterations, but converges to give the same weight to all the examples in the later phases of the training. Still, the probability of sampling a harder example in the first iterations is not null, and the more difficult samples which are occasionally picked increase the diversity of the data and help training.

The above-mentioned methodology should work well for balanced data sets, as various curriculum sampling strategies have been successfully employed in literature [19, 28, 34, 37], but it can fail when the data is unbalanced. Ionescu et al. [12] show that some classes may be more difficult than others. A simple motivation for this may be the context in which each class appears. For example, a potted plant or a bottle are rarely the focus of attention, usually being placed somewhere in the background. Other classes of objects, such as tables, are usually occluded, with the pictures focusing on the objects on the table rather than on the piece of furniture itself. This can make a standard curriculum sampling strategy neglect examples from certain classes and slow down training. The problem becomes even more serious in a context where the data is biased towards the easier classes. To solve these issues, we add a new term to our sampling function which takes into consideration the classes of the elements already sampled, in order to emphasize on images from less-visited classes and ensure the diversity of the selected examples.

The importance of diversity can be easily explained when comparing our machine learning approach to actual real-life examples. For instance, when creating a new vaccine, researchers need to experiment on multiple variants of the virus, then test it on a diverse group of people. As a rule, in all sciences, before making any assumptions, researchers have to examine a diverse set of examples which are relevant to the actual data distribution. Similar to the vaccines, which must be efficient for as many people as possible, we want our curriculum model to work well on all object classes. We argue that this is not possible in unbalanced curriculum scenarios, and it is slower in the traditional random training setup.

Since it is a sampling procedure, our CL approach can be applied to any supervised task in machine learning. In this paper, we focus on object detection and instance segmentation, two of the main tasks in computer vision, which require the model to identify the class and the location of objects in images. To test the validity of our approach, we experiment on two data sets: Pascal VOC 2007 [4] and Cityscapes [3], and compare our curriculum with diversity strategy

[1] University of Bucharest, Department of Computer Science, Romania, email: petru.soviany@yahoo.com

against the standard random training method, a curriculum sampling (without diversity) procedure and an inverse-curriculum approach, which selects images from hard to easy. We employ a state-of-the-art Faster R-CNN [24] detector with a Resnet-101 [11] backbone for the object detection experiments, and a Mask R-CNN [10] model based on Resnet-50 for instance segmentation.

Our main contributions can be summarized as follows:

1. We illustrate the necessity of adding diversity when using CL in unbalanced data sets;
2. We introduce a novel curriculum sampling function, which takes into consideration the class-diversity of the training samples and improves results when traditional curriculum approaches fail;
3. We prove our strategy by experimenting on two computer vision tasks: object detection and instance segmentation, using two data sets of high interest.

We organize the rest of this paper as follows: in Section 2, we present the most relevant related works and compare them with our approach. In Section 3, we explain in detail the methodology we follow. We present our results in Section 4, and draw our conclusion and discuss possible future work in the last section.

2 Related Work

Curriculum learning. Bengio et al. [2] introduced the idea of curriculum learning (CL) to train artificial intelligence, proving that the standard learning paradigm used in human educational systems could also be applied to automatic models. CL represents a class of easy-to-hard approaches, which have successfully been employed in a wide range of machine learning applications, from natural language processing [8, 16, 19, 21, 31], to computer vision [6, 7, 9, 15, 18, 27, 35], or audio processing [1, 22].

One of the main limitations of CL is that it assumes the existence of a predefined metric which can rank the samples from easy to hard. These metrics are usually task-dependent with various solutions being proposed for each. For example, in text processing, the length of the sentence can be used to estimate the difficulty of the input (shorter sentences are easier) [21, 30], while the number and the size of objects in a certain sample can provide enough insights about difficulty in image processing tasks (images with few large objects are easier) [27, 29]. In our paper, we employ the image difficulty estimator of Ionescu et al. [12] which was trained considering the time required by human annotators to identify the presence of certain classes in images.

To alleviate the challenge of finding a predefined difficulty metric, Kumar et al. [17] introduce self-paced learning (SPL), a set of approaches in which the model ranks the samples from easy to hard during training, based on its current progress. For example, the inputs with the smaller loss at a certain time during training are easier than the samples with higher loss. Many papers apply SPL successfully [26, 32, 33], and some methods combine prior knowledge with live training information, creating self-paced with curriculum techniques [14, 36]. Even so, SPL still has some limitations, requiring a methodology on how to select the samples and how much to emphasize easier examples. Our approach is on the borderline between CL and SPL, but we consider it to be pure curriculum, although we use training information to advantage less visited classes. During training, we only count the labels of the training samples, which is a priori information, and not the learning progress. A similar system could iteratively select examples from every class, but this would force our model to process the same number of examples from each class. Instead, by using the class-diversity as a term in our difficulty-based sampling probability function, we impose the selection of easy-to-hard diverse examples, without massively altering the actual class distribution of the data set.

The easy-to-hard idea behind CL can be implemented in multiple ways. One option is to start training on the easiest set of images, while gradually adding more difficult batches [2, 7, 16, 27, 30, 37]. Although most of the models keep the visited examples in the training set, Kocmi et al. [16] suggest reducing the size of each bin until combining it with the following one, in order to use each example only once during an epoch. In [19, 28] the authors propose a sampling strategy according to some probability function, which favors easier examples in the first iterations. As the authors show, the easiness score from [28] could also be added as a new term to the loss function to emphasize the easier examples in the beginning of the training. In this paper, we enhance their sampling strategy by adding a new diversity term to the probability function used to select training examples.

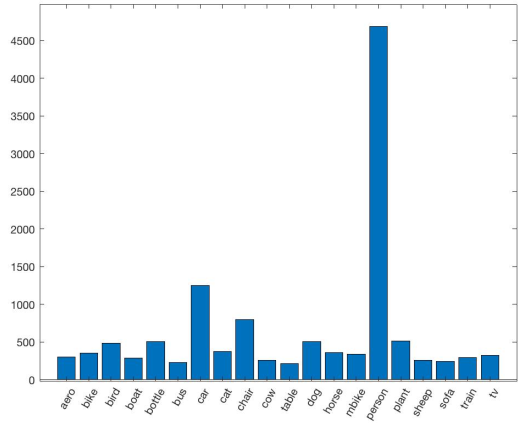

Figure 1. Number of instances from each class in the trainval split of the Pascal VOC 2007 data set.

Despite leading to good results in many related papers, the standard CL procedure is highly influenced by the task and the data distribution. Simple tasks may not gain much from using curriculum approaches, while employing CL in unbalanced data sets can lead to slower convergence. To address the second problem, Wang et al. [34] introduce a CL framework which adaptively adjusts the sampling strategy and loss weight in each batch, while other papers [13, 25] argue that a key element is diversity. Jiang et al. [13] introduce a SPL with diversity technique in which they regularize the model using both difficulty information and the variety of the samples. They suggest using clustering algorithms to split the data into diverse groups. Sachan et al. [25] measure diversity using the angle between the hyperplanes the samples induce in the feature space. They choose the examples that optimize a convex combination of the curriculum learning objective and the sum of angles between the candidate samples and the examples selected in previous steps. In our model, we define diversity based on the classes of our data. We combine our predefined difficulty metric with a score which favors images from less visited classes, in order to sample easy and diverse examples at the beginning of the training, then gradually add more complex elements. Our idea works well for supervised tasks, but it can be extended to unsupervised learning by replacing the ground-truth labels with a clustering model, as suggested in [13]. Figure 1 presents the class distribution on Pascal VOC 2007 data set [4] which is heavily biased towards class *person*.

Object detection is the task of predicting the location and the class of objects in certain images. As noted in [29], the state-of-the-art object detectors can be split into two main categories: two-stage and single stage models. The two-stage object detectors [10, 24] use a Region Proposal Network to generate regions of interest which are then fed to another network for object localization and classification. The single stage approaches [20, 23] take the whole image as input and solve the problem like a regular regression task. These methods are usually faster, but less accurate than the two-stage designs. **Instance segmentation** is similar to object detection, but more complex, requiring the generation of a mask instead of a bounding box for the objects in the test image. Our strategy can be implemented using any detection and segmentation models, but, in order to increase the relevance of our results, we experiment with high quality Faster R-CNN [24] and Mask R-CNN [10] baselines.

3 Methodology

Training artificial intelligence using curriculum approaches, from easy to hard, can lead to improved results in a wide range of tasks [1, 6, 7, 8, 9, 15, 16, 18, 19, 21, 22, 27, 31, 35]. Still, it is not simple to determine which samples are easy or hard, and the available metrics are usually task-dependent. Another challenge of CL is finding the right curriculum schedule, i.e. how fast to add more difficult examples to training, and how to introduce the right amount of harder samples at the right time to positively influence convergence. In this section, we present our approach for estimating difficulty and our curriculum sampling strategies.

3.1 Difficulty estimation

To estimate the difficulty of our training examples, we employ the method of Ionescu et al. [12] who defined image difficulty as the human time required for solving a visual search task. They collected annotations for the Pascal VOC 2012 [5] data set, by asking annotators whether a class was present or not in a certain image. They collected the time people required for answering these questions, which they normalized and fed as training data for a regression model. Their results correlate fine with other difficulty metrics which take into consideration the number of objects, the size of the objects, or the occlusions. Because it is based on human annotations, this method takes into account the whole image context, not only certain features relevant for one problem (the number of objects, for example). This makes the model task independent, and, as a result, it was successfully employed in multiple vision problems [12, 29, 28]. To further prove the efficiency of the estimator for our task, we show that automatic models have a lower accuracy in difficult examples. We split the Pascal VOC 2007 [4] test set in three equal batches: easy, medium and hard, and run the baseline model on each of them. The results in Table 1 confirm that the AP lowers as the difficulty increases.

Table 1. Average Precision scores for object detection using the baseline Faster R-CNN, on easy, medium and hard splits of Pascal VOC 2007 test set, as estimated using our approach.

DIfficulty	mAP (in %)
Easy	72.93
Medium	72.16
Hard	67.03

We follow the strategy of Ionescu et al. as described in the original paper [12] to determine the difficulty scores of the images in our data sets. These scores have values ≈ 3, with a larger score defining a more difficult sample. We translate the values between $[-1, 1]$ using Equation 1 to simplify the usage of the score in the next steps. Figure 2 shows some examples of easy and difficult images.

$$Scale_{min-max}(x) = \frac{2 \cdot (x - min(x))}{max(x) - min(x)} - 1 \qquad (1)$$

3.2 Curriculum sampling

Soviany et al. [28] introduce a curriculum sampling strategy, which favors easier examples in the first iterations and converges as the training progresses. It has the advantage of being a continuous method, removing the necessity of a curriculum schedule for enhancing the difficulty-based batches. Furthermore, the fact that it is a probabilistic sampling method does not constrain the model to only select easy examples in the first iterations, as batching does, but adds more diversity in data selection. We follow their approach in building our curriculum sampling strategy with only a small change in the position of parameter k in order to better emphasize the difficulty of the examples. We use the following function to assign weights to the input images during training:

$$w(x_i, t) = (1 - diff(x_i) \cdot e^{-\gamma \cdot t})^k, \forall x_i \in X, \qquad (2)$$

where x_i is the training example from the data set X, t is the current iteration, and $diff(x_i)$ is the difficulty score associated with the selected sample. γ is a parameter which sets how fast the function converges to 1, while k sets how much to emphasize the easier examples. Our function varies from the one proposed in [28] by changing the position of the k parameter. We consider that we can take advantage of the properties of the power function which increases faster for numbers greater than the unit. Since $1 - s_i \cdot e^{-\gamma \cdot t} \in [0, 2]$, and the result is > 1 for easier examples, our function will focus more on the easier samples in the first iterations. As the training advances, the function converges to 1, so all examples will have the same probability to be selected in the later phases of the training. We transform the weights into probabilities and we sample accordingly.

3.3 Curriculum with diversity sampling

As [13, 25] note, applying a CL strategy does not guarantee improved quality, the diversity of the selected samples having a great impact on the final results. A simple example is the case in which the data set is biased, having fewer samples of certain classes. Since some classes are more difficult than others [12], if the data set is not well-balanced, the model will not visit the harder classes until the later stages of the training. Thus, the model will not perform well on classes it did not visit. This fact is generally valid in all kind of applications, even in real life reasoning: without seeing examples which match the whole data distribution, it is impossible to find the solution suited for all scenarios. Because of this, we enhance our sampling method, by adding a new term, which is based on the diversity of the examples.

Our diversity scoring algorithm is simple, taking into consideration the classes of the selected samples. During training, we count the number of visited objects from each class ($num_{objects}(c)$). We subtract the mean of the values to determine how often each class was visited. This is formally presented in Equation 3. We scale and translate the results between $[-1, 1]$ using Equation 1 to get the score

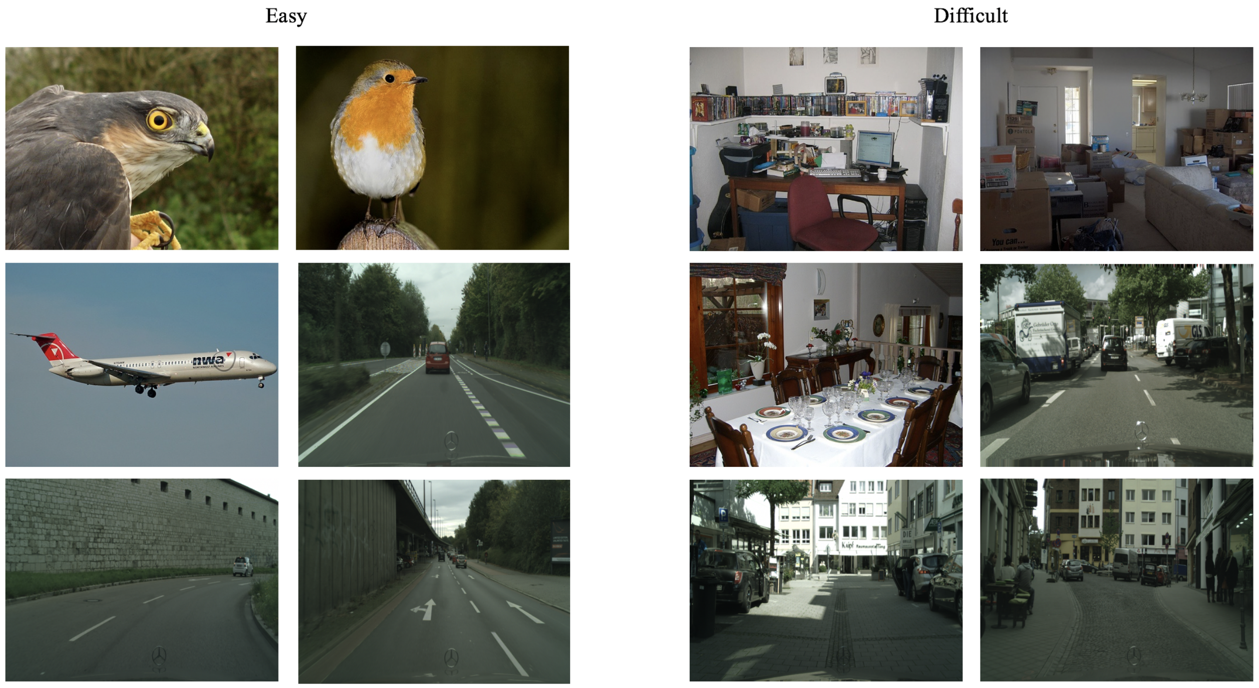

Figure 2. Easy and difficult images from Pascal VOC 2007 and Cityscapes according to our estimation.

of each class, then, for every image, we compute the image-level diversity by averaging the class score for each object in its ground-truth labels (Equation 4).

$$visited(c_i) = num_{objects}(c_i) - \frac{\sum_{c_j \in C} num_{objects}(c_j)}{|C|} \quad \forall c_i \in C. \quad (3)$$

$$imgVisited(x_i) = \frac{\sum_{obj \in objects(x_i)} visited(class(obj))}{|objects(x_i)|} \quad \forall x_i \in X. \quad (4)$$

In our diversity algorithm we want to emphasize the images containing objects from less visited classes, i.e. with a small $imgVisited$ value, closer to -1. We compute a scoring function similar to Equation 2, which also takes into consideration how often a class was visited, in order to add diversity:

$$\begin{aligned} w(x_i, t) = [1 - \alpha \cdot (diff(x_i) \cdot e^{-\gamma \cdot t}) \\ - (1 - \alpha) \cdot (imgVisited(x_i) \cdot e^{-\gamma \cdot t})]^k, \quad (5) \end{aligned}$$

where α controls the impact of each component, the difficulty and the diversity, while the rest of the notation follows Equation 2. We transform the weights into probabilities by dividing them by their sum, and we sample accordingly.

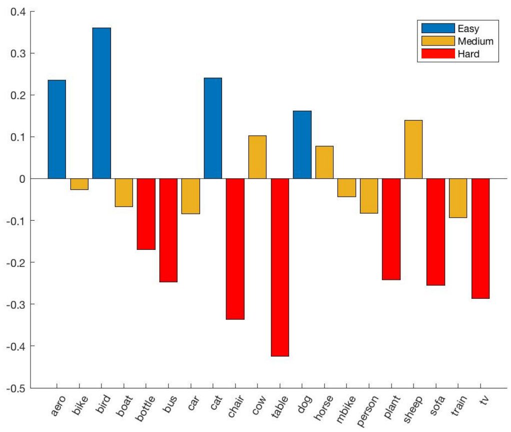

Figure 3. Difficulty of classes in Pascal VOC 2007 according to our estimation. Best viewed in color.

4 Experiments

4.1 Data sets

In order to test the validity of our method, we experiment on two data sets: Pascal VOC 2007 [4] and Cityscapes [3]. We conduct detection experiments on 20 classes, training on the 5011 images from the Pascal VOC 2007 trainval split. We perform evaluation on the test split which contains 4952 images. For our instance segmentation experiments, we use the Cityscapes data set which contains eight labeled object classes: person, rider, car, truck, bus, train, motorcycle, bicy-

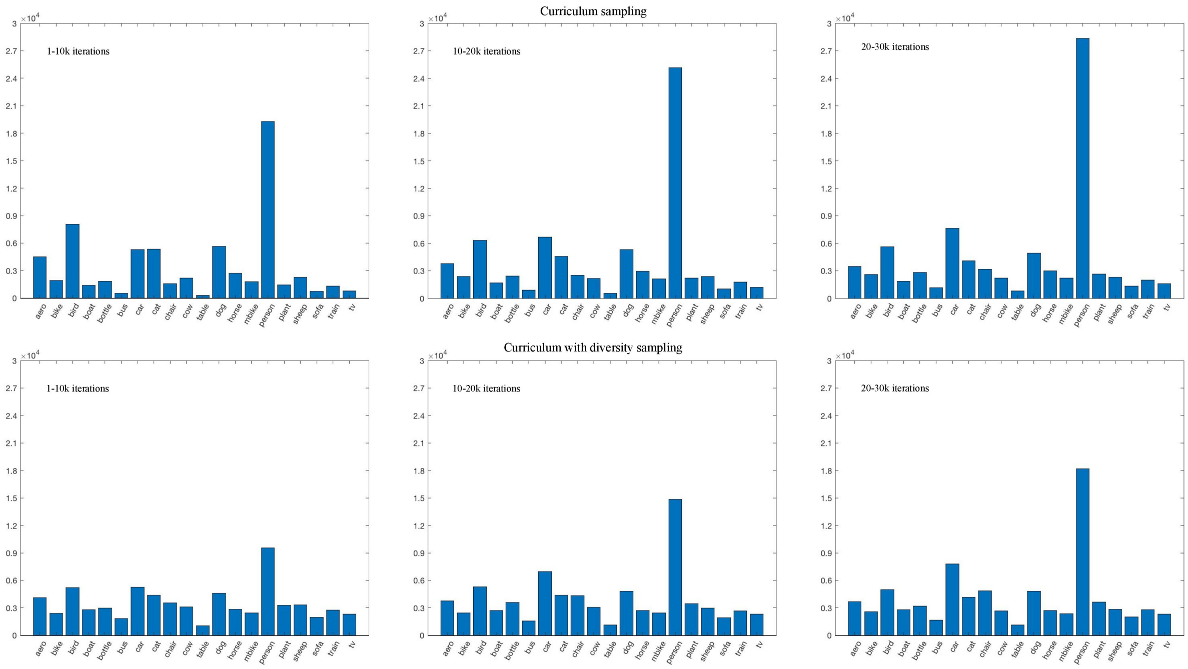

Figure 4. Number of objects from each class sampled during our training on Pascal VOC 2007. On the first row it is the curriculum sampling method and on the second row it is the curriculum with diversity approach. We present the first 30000 iterations for each case, with histograms generated from 10k to 10k steps.

cle. We train on the training set of 2975 images and we evaluate on the validation split of 500 images.

4.2 Baselines and configuration

We build our method on top of the Faster R-CNN [24] and Mask R-CNN [10] implementations available at: https://github.com/facebookresearch/maskrcnn-benchmark. For our detection experiments, we use Faster R-CNN with Resnet-101 [11] backbone, while for segmentation we employ the Resnet-50 backbone on the Mask R-CNN model. We use the configurations available on the web site, with the learning rate adjusted for a training with a batch size of 4. In our sampling procedure (Equation 5) we set $\alpha = 0.5$, $\gamma = 6 \cdot 10^{-5}$, and $k = 5$. We do not compare with other models, because the goal of our paper is not surpassing the state of the art, but improving the quality of our baseline model. We also present the results of a hard-to-easy sampling, in order to prove the efficiency of the easy-to-hard curriculum approaches inspired by human learning.

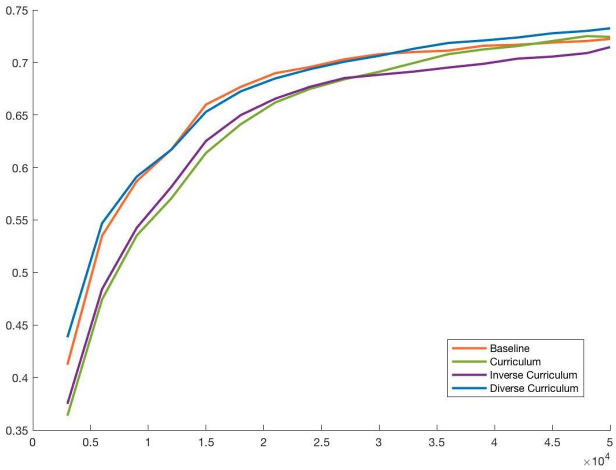

Figure 5. Evolution of mAP during training on Pascal VOC 2007 for object detection. Best viewed in color.

4.3 Evaluation metrics

We evaluate our results using the mean Average Precision (AP). The AP score is given by the area under the precision-recall curve for the detected objects. The Pascal VOC 2007 [4] metric is the mean of precision values at a set of 11 equally spaced recall levels, from 0 to 1, at a step size of 0.1. The Cityscapes [3] metric computes the average precision on the region level for each class and averages it across 10 different overlaps ranging from 0.5 to 0.95 in steps of 0.05. We also report results on Cityscapes using AP50% and AP75%, which correspond to overlap values of 50% and 75%, respectively. Since the exact evaluation protocol has some differences for each data set, we use the Pascal VOC 2007 [4] metric for the detection experiments and the Cityscapes [3] metric for the instance segmentation results. We use the evaluation code available at https://github.com/facebookresearch/maskrcnn-benchmark. More details about the evaluation metrics can be found in the original papers [3, 4].

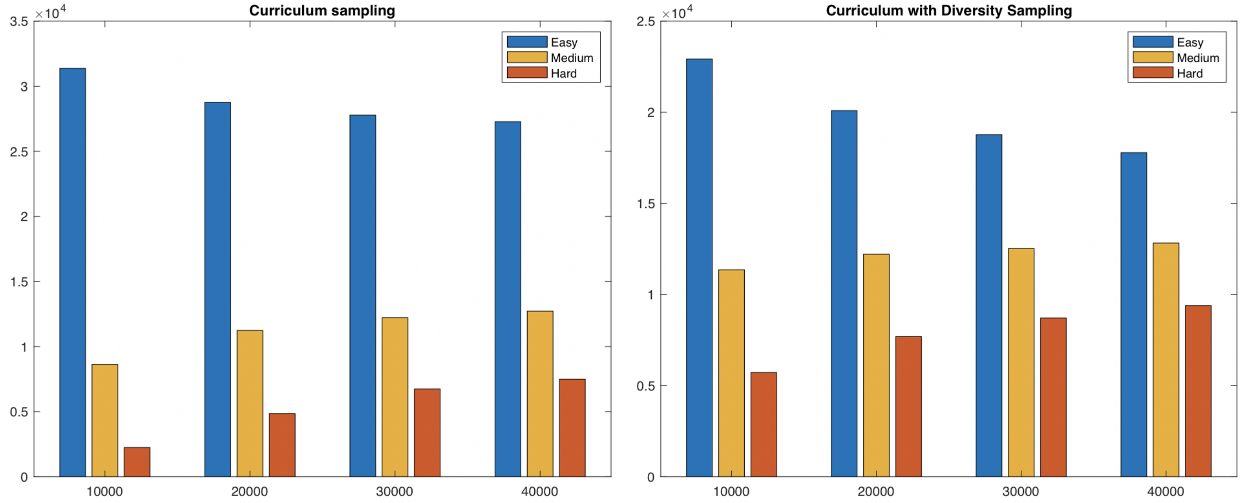

Figure 6. Difficulty of the images samples during our training on Pascal VOC 2007. On the left it is presented the curriculum sampling method and on the right the curriculum with diversity approach. We present the first 40000 iterations for each case, with histograms generated from 10k to 10k steps. Best viewed in color.

4.4 Results and discussion

The class distribution of the objects in Pascal VOC 2007 clearly favors class *person*, with 4690 instances, while classes *dinningtable* and *bus* only contain 215 and 229 instances, respectively. This would not be a problem if the difficulty of the classes was similar, because we can assume the test data set has a matching distribution, but this is not the case, as it is shown in Figure 3.

Figure 4 presents how the two sampling methods behave during training on the Pascal VOC 2007 data set. In the first 10k iterations, curriculum sampling selects images with almost 20k objects from class *person* and only 283 instances from class *diningtable*. By adding diversity, we lower the gap between classes, reaching 10k objects of persons and 1000 instances of tables. This behaviour continues as the training progresses, with the differences between classes being smaller when adding diversity. It is important to note that we do not want to sample the exact number of objects from each class, but to keep the class distribution of the actual data set, while feeding the model with enough details about every class. Figure 6 shows the difficulty of the examples sampled according to our strategies. We observe that by adding diversity we do not break our curriculum learning schedule, the examples still being selected from easy to hard.

To further prove the efficiency of our method, we compute the AP on both object detection and instance segmentation tasks. The results are presented in Tables 2 and 3.

We repeat our object detection experiments five times and average the results, in order to ensure their relevance. The sampling with diversity approach provides an improvement of 0.69% over the standard curriculum method, and of 0.79% over the randomly-trained baseline. Although the improvement is not large, we can observe that by adding diversity we boost the accuracy where the standard method would fail, without much effort. Our experiments, with an inverse curriculum approach, from hard to easy, lead to the worst results, showing the utility of presenting the training samples in a meaningful order, similar to the way people learn.

Moreover, Figure 5 illustrates the evolution of the AP during training. The curriculum with diversity approach has superior results over the baseline from the beginning to the end of the training. As the figure shows, the difference between the two methods increases in the later stages of the training. A simple reason for this behaviour is the fact that the curriculum strategy is fed with new, more difficult, examples as the training progresses, continuously improving the accuracy of the model. On the other hand, the standard random procedure receives all information from the beginning, reaching a plateau early during training. The standard CL method starts from lower scores, exactly because it does not visit enough samples from more difficult classes in the early stages of the training. For instance, after 5000 iterations, the AP of the standard CL approach on class *dinningtable* was 0. Thus, by adding diversity, our model converges faster than the traditional methods.

Table 2. Average Precision scores for object detection on Pascal VOC 2007 data set.

Model	mAP (in %)
Faster R-CNN (Baseline)	72.28 ± 0.34
Faster R-CNN with curriculum sampling	72.38 ± 0.32
Faster R-CNN with inverse curriculum sampling	70.89 ± 0.53
Faster R-CNN with diverse curriculum sampling	$\mathbf{73.07 \pm 0.28}$

Table 3. Average Precision scores for instance segmentation on Cityscapes data set.

Model	AP	AP50%	AP75%
Faster R-CNN (baseline)	38.72	69.15	34.95
Curriculum sampling	38.47	**69.88**	35.01
Inverse curriculum sampling	37.40	68.17	34.22
Diverse curriculum sampling	**39.12**	69.86	**35.4**

The instance segmentation results on the Cityscapes data set confirm the conclusion from our previous experiments. As Table 3 shows, the curriculum with diversity is again the optimal method, surpassing the baseline with 0.4% using AP, 0.71% using AP50%, and 0.45% using AP75%. It is interesting to point out that, although the diverse curriculum approach has a better AP and AP75% than the standard CL method, the former technique surpasses our method with 0.02% when evaluated using AP50%. The inverse curriculum approach has the worst scores again, strengthening our statements on the utility of curriculum learning and the importance of providing training examples in a meaningful order.

5 Conclusion and future work

In this paper, we presented a simple method of optimizing the curriculum learning approaches on unbalanced data sets. We consider that the diversity of the selected examples is just as important as their difficulty, and neglecting this fact may slow down training for more difficult classes. We introduced a novel sampling function, which uses the classes of the visited examples together with a difficulty score to ensure the curriculum schedule and the diversity of the selection. Our object detection and instance segmentation experiments conducted on two data sets of high interest prove the superiority of our method over the randomly-trained baseline and over the standard CL approach. A benefit of our methodology is that it can be used on top of any deep learning model, for any supervised task. Diversity can be a key element for overcoming one of the shortcomings of CL which can lead to the replacement of the traditional random training and a larger adoption of meaningful sample selection. For the future work, we plan on studying more difficulty measures to build an extensive view on how the chosen metric affects the performance of our system. Furthermore, we aim to create an ablation study on the parameter choice and find better ways to detect the right parameter values. Another important aspect we are considering is extending the framework to unsupervised tasks, by introducing a novel method of computing the diversity of the examples.

REFERENCES

[1] Dario et al. Amodei, 'Deep speech 2: End-to-end speech recognition in english and mandarin', in *Proceedings of ICML*, p. 173–182, (2016).

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston, 'Curriculum learning', in *Proceedings of ICML*, pp. 41–48, (2009).

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, 'The cityscapes dataset for semantic urban scene understanding', in *Proceedings of CVPR*, (2016).

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[6] C. Gong, D. Tao, S. J. Maybank, W. Liu, G. Kang, and J. Yang, 'Multimodal curriculum learning for semi-supervised image classification', *IEEE Transactions on Image Processing*, **25**(7), 3249–3260, (2016).

[7] L. Gui, T. Baltrušaitis, and L. Morency, 'Curriculum learning for facial expression recognition', in *Proceedings of FG*, pp. 505–511, (2017).

[8] Junliang Guo, Xu Tan, Linli Xu, Tao Qin, Enhong Chen, and Tie-Yan Liu, 'Fine-tuning by curriculum learning for non-autoregressive neural machine translation', *arXiv preprint arXiv:1911.08717*, (2019).

[9] Guy Hacohen and Daphna Weinshall, 'On the power of curriculum learning in training deep networks', in *Proceedings of ICML*, (2019).

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, 'Mask r-cnn', in *Proceedings of ICCV*, pp. 2961–2969, (2017).

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, 'Deep residual learning for image recognition', in *Proceedings of CVPR*, pp. 770–778, (2016).

[12] Radu Tudor Ionescu, Bogdan Alexe, Marius Leordeanu, Marius Popescu, Dim P. Papadopoulos, and Vittorio Ferrari, 'How hard can it be? estimating the difficulty of visual search in an image', in *Proceedings of CVPR*, pp. 2157–2166, (2016).

[13] Lu Jiang, Deyu Meng, Shoou-I Yu, Zhenzhong Lan, Shiguang Shan, and Alexander Hauptmann, 'Self-paced learning with diversity', in *Proceedings of NIPS*, pp. 2078–2086, (2014).

[14] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G Hauptmann, 'Self-paced curriculum learning', in *Proceedings of AAAI*, (2015).

[15] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei, 'Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels', in *Proceedings of ICML*, pp. 2304–2313, (2018).

[16] Tom Kocmi and Ondřej Bojar, 'Curriculum learning and minibatch bucketing in neural machine translation', in *Proceedings of RANLP*, pp. 379–386, (2017).

[17] M Pawan Kumar, Benjamin Packer, and Daphne Koller, 'Self-paced learning for latent variable models', in *Proceedings of NIPS*, pp. 1189–1197, (2010).

[18] Siyang Li, Xiangxin Zhu, Qin Huang, Hao Xu, and C.-C. Jay Kuo, 'Multiple instance curriculum learning for weakly supervised object detection', in *Proceedings of BMVC*. BMVA Press, (2017).

[19] Cao Liu, Shizhu He, Kang Liu, and Jun Zhao, 'Curriculum learning for natural answer generation.', in *Proceedings of IJCAI*, pp. 4223–4229, (2018).

[20] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg, 'Ssd: Single shot multibox detector', in *Proceedings of ECCV*, pp. 21–37. Springer, (2016).

[21] Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell, 'Competence-based curriculum learning for neural machine translation', in *Proceedings of NAACL*, pp. 1162–1172, (2019).

[22] Shivesh Ranjan and John HL Hansen, 'Curriculum learning based approaches for noise robust speaker recognition', *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **26**(1), 197–210, (2017).

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, 'You only look once: Unified, real-time object detection', in *Proceedings of CVPR*, pp. 779–788, (2016).

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, 'Faster r-cnn: Towards real-time object detection with region proposal networks', in *Proceedings of NIPS*, pp. 91–99, (2015).

[25] Mrinmaya Sachan and Eric Xing, 'Easy questions first? a case study on curriculum learning for question answering', in *Proceedings of ACL*, pp. 453–463, (2016).

[26] Enver Sangineto, Moin Nabi, Dubravko Culibrk, and Nicu Sebe, 'Self paced deep learning for weakly supervised object detection', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **41**(3), 712–725, (2018).

[27] Miaojing Shi and Vittorio Ferrari, 'Weakly supervised object localization using size estimates', in *Proceedings of ECCV*, pp. 105–121. Springer, (2016).

[28] Petru Soviany, Claudiu Ardei, Radu Tudor Ionescu, and Marius Leordeanu, 'Image difficulty curriculum for generative adversarial networks (cugan)', in *Proceedings of WACV*, (2020).

[29] Petru Soviany and Radu Tudor Ionescu, 'Frustratingly Easy Trade-off Optimization between Single-Stage and Two-Stage Deep Object Detectors', in *Proceedings of CEFRL Workshop of ECCV*, pp. 366–378, (2018).

[30] Valentin I. Spitkovsky, Hiyan Alshawi, and Daniel Jurafsky, 'Baby Steps: How "Less is More" in unsupervised dependency parsing', in *Proceedings of NIPS Workshop on Grammar Induction, Representation of Language and Language Learning*, (2009).

[31] Sandeep Subramanian, Sai Rajeswar, Francis Dutil, Christopher Pal, and Aaron Courville, 'Adversarial generation of natural language', in *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pp. 241–251, (2017).

[32] James S Supancic and Deva Ramanan, 'Self-paced learning for long-term tracking', in *Proceedings of CVPR*, pp. 2379–2386, (2013).

[33] Kevin Tang, Vignesh Ramanathan, Li Fei-Fei, and Daphne Koller,

‘Shifting weights: Adapting object detectors from image to video’, in *Proceedings of NIPS*, pp. 638–646, (2012).

[34] Yiru Wang, Weihao Gan, Jie Yang, Wei Wu, and Junjie Yan, ‘Dynamic curriculum learning for imbalanced data classification’, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (October 2019).

[35] Daphna Weinshall and Gad Cohen, ‘Curriculum learning by transfer learning: Theory and experiments with deep networks’, in *Proceedings of ICML*, (2018).

[36] Dingwen Zhang, Junwei Han, Long Zhao, and Deyu Meng, ‘Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework’, *International Journal of Computer Vision*, **127**(4), 363–380, (2019).

[37] Xuan Zhang, Gaurav Kumar, Huda Khayrallah, Kenton Murray, Jeremy Gwinnup, Marianna J Martindale, Paul McNamee, Kevin Duh, and Marine Carpuat, ‘An empirical exploration of curriculum learning for neural machine translation’, *arXiv preprint arXiv:1811.00739*, (2018).