Multiple Layer Model for Object Detection and Sketch Representation

Wencheng Li*, Xin Wu†, Ling Cai‡, Fuqiao Hu* and Yuming Zhao*
*School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University
Email: {misxeday, fqhu, arola_zym}@sjtu.edu.cn
†Department of Electronic Engineering, Beijing Institute of Technology
Email: hdfwx@bit.edu.cn
‡School of Information Science and Engineering, Xiamen University
Email: cailing.cs@gmail.com

***Abstract*—In this paper we propose a multiple layer model for object detection and sketch representation. Unlike most traditional detection models focusing on the object localization, we investigate both the object detection and sketch representation within an unified framework. Based on the multiple layer architecture, our model can provide the sketch information of the detected object. Meanwhile, we generalize it from single scale structure to multiple scales, which efficiently saves time consumed in the image pyramids construction. To efficiently train the classifier at the top layer, we employ the stochastic gradient descent algorithm to minimize the training error and back propagate it to the bottom layer. The experimental results demonstrate that our model outperforms the conventional active basis model.**

***Index Terms*—Object Detection, stochastic gradient descent, Multiple layer model, Sketch Representation**

I. INTRODUCTION

As one of the fundamental challenges in computer vision, object detection increasingly attracts the attention of academic and industry researchers. It aims to localize some specific targets from static images and dynamic videos, and sets the basis for image understanding and behavior analysis. Therefore, it is intensively applied to many prominent fields, such as video surveillance, intelligent transportation, biometric feature recognition, etc.

Object detection has been widely studied for decades. By employing histogram of oriented gradient (HOG) features in combination with a linear support vector machine (SVM) classifier, Dalal et al. [1] achieved considerable gain in performance and the model becomes well known in pedestrian detection afterwards. The Deformable Part-based Model (DPM) [2] representing objects with the pictorial structures also turns into one of the most outstanding frameworks. Unlike these discriminative models [3], inspired by biological visual system, Ying Nian Wu et al. put forward a generative deformable model [4] consisting of a group of active basis elements based on theories of the matching pursuit algorithm [5] and the wavelet sparse coding [6].

However, most traditional detection models focus on target localization. But for some challenging tasks, like posture recognition of motion targets, key points localization, and non-rigid object registration and so on, we need more theoretical information of visual patterns. In real life scenarios, the single-scale models are hard to balance the false positive rate and missing rate well for the existing great range of target scales [7]. Meanwhile, they also are time-consuming to construct image pyramids. Recently, with the overwhelming research of multiple layer network, the necessity of numerous training samples and long training time are the crucial conditions. For instance, the convolved neural network [8] requires millions of training samples and dozens of hours for training. All these obstacles have prevented further application of multiple layer model.

Inspired by active basis model (ABM), we construct a multiple layer model for object detection and sketch representation. Moreover, the SVM classifier is placed at the top layer, so that the model can be learned with a few training samples. To back propagate the training error from the top layer to the bottom layer, the Pegasos algorithm is adopted to iteratively optimize our model.

The contributions of the proposed model is as below.

(1) **Forward and backward propagation** The traditional multiple layer model, like convolutional neural network (CNN), only performs forward propagation at the process of testing and backward propagation at the training. However, our model has both forward and backward at the process of training and testing.

(2) **Multi-scale model** Though the original active basis model could achieve convincing results on object detection and description, single scale model could hardly balance the false positive rate and missing rate well [9]. Most traditional algorithms [4] apply image pyramids to deal with this problem which always result in the inefficiency of detection. Our model is trained as the single scale model but it can be generalized to an arbitrary scale by adjusting model's parameters.

(3) **Improved Pegasos algorithm** The original ABM is learned with a forward picking process. But the absence of the back propagation of training errors makes the model less robust in distinguishing true positives and true

‡ Corresponding author. This work is supported by NSFC 61572410.

negatives. We employ the Pegasos algorithm and place it at the top layer of the model.

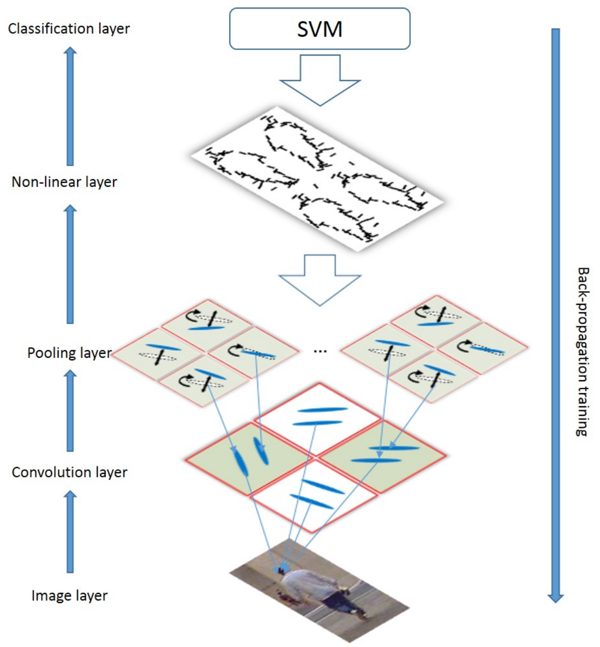


Fig. 1: Model architecture. The top layer: classification layer. The middle layers: convolution layer, pooling layer and non-linear layer. The bottom layer: image layer.

II. MODEL DESCRIPTION

Similar to CNN, our model consists of the image layer, feature encoding layer (including convolution layer, pooling layer, non-linear layer) and classification laye, as shown in Figure 1. It can be concisely represented as the following linear equation:

$$s = \omega\varphi(I_m, B) + b \tag{1}$$

$$\varphi(I_m, B) = max(con\langle I_m, B\rangle) \tag{2}$$

where the confident detection score s of detection model is linearly determined by the features $\varphi(I_m, B)$ from the feature encoding layer B and the classifier parameters ω and b. The operators $con(.)$ and $max(.)$ represent the convolution and pooling process.

In our model, the oriental Gabor filters play the role of filters of the convolution layer. One-dimensional Gabor wavelet was proposed by Gabor in 1946 [10] and it has been further studied to put forward the two-dimensional Gabor wavelet [11]. Gabor transformation can extract related Gabor features representing different scales and orientations of frequency-domain. The excellent transformation property of the Gabor wavelet between space domain and frequency domain makes it widely used in texture analysis and character recognition, etc. The Gabor kernel function is defined as:

$$G(x, y) \propto exp\{-[(\frac{x}{\sigma_x})^2 + (\frac{y}{\sigma_y})^2]/2\}e^{ix} \tag{3}$$

Through transformation, rotation and scale of $G(x, y)$, a general format of Gabor transformation is written as $B_{x,y,s,a}(x_1, y_1) = G(\widetilde{x}/s, \widetilde{y}/s)/s^2$, where $\widetilde{x} = (x_1 - x)\cos\alpha + (y_1 - y)\sin\alpha$, $\widetilde{y} = -(x_1 - x)\sin\alpha + (y_1 - y)\cos\alpha$. Assuming an image space D and a set of Gabor elements $\{B_{x,y,s,\alpha}, \forall(x, y, s, \alpha)\}$, an active basis model consists of Gabor filters with different orientations and positions. In fact, Gabor features are generated by convoluting the Gabor kernel function with images, so we can get the summation representation of the image: $I(x, y) = \sum_{i=0}^{n} c_i B_i(x, y) + U(x, y), i = 1, \ldots, n$, where c_i is the decomposition coefficient and $U(x, y)$ is the residual image. We encode the image layer with the following three steps. Firstly, it starts from convolving the images with Gabor filters at the convolution layer. Secondly, it applies a local maximization operator to the convolution results and thirdly a local summation operation is computed at the non-linear layer. In the end, the SVM classifier is adopted to get the positives. Figure 1 illustrates our simplified architecture of learning and inference.

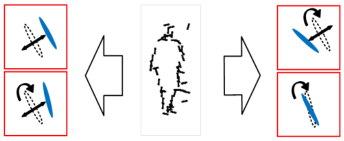


Fig. 2: Demonstration of an active base model, which consists of a series of Gabor elements, displayed as ellipses.

III. SCHEME OF LEARNING

Unlike the traditional hand-crafted feature encoding, ABM is a kind of feature learning procedure based on the sparse coding theory. It searches for those discriminative features from training samples. It is mainly composed of two parts: one is the search of distinguished feature (as active basis) through the shared sketch algorithm; the other is to learn the basis weights.

A. *Shared sketch algorithm*

We employ the shared sketch algorithm to search training images for some distinguished features and regard them as the active bases. Given a set of training images $I_m, m = 1, \ldots, M$, the shared sketch algorithm successively picks Gabor elements B_i and allows perturbations within local ranges to describe the image I_m. Each of the basic feature elements is shared by all the training images with some perturbations.

The shared sketch algorithm adopts training images $\{I_m, m = 1, \ldots, M\}$ to compute $[I_m, B] = h(|\langle I_m, B\rangle|^2)$ in terms of each image I_m and Gabor element $B \in Dictionary$. Then it chooses an ideal $B_{m,i}$ which satisfies $B_{m,i} \approx B_i[I_m, B_{m,i}]$ to make the inner product $[I_m, B_{m,i}]$ maximum. The local maximum pooling operation can get a perturbed Gabor elements which sketch a local edge segment optimally. And it specifically selects the candidates B_i to correspondingly get $\sum_{m=1}^{M}[I_m, B_{m,i}]$ and the weight of B_i which is $\lambda_i =$

$M/\sum_m [I_m, B_{m,i}]$. After iteratively training process the final deformed templates and common template are generated.

B. Multiple Scale Generalization

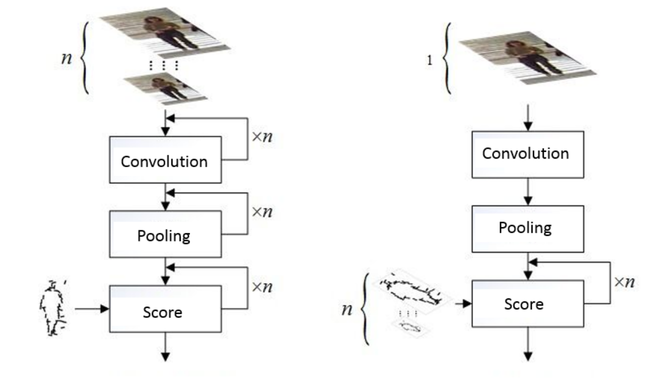
After obtaining the single scale model we can rotate and flip it for better matching result. To avoid cost on the image pyramids, we propose to adjust the model directly based on those learned active bases to produce the multiple scale model.

The original model consists of a series of active bases with specified locations and orientations. Given a single scale model $\{B_{x_i,y_i,s,\alpha_i}, (x_i, y_i) \in D, i = 1, \dots, n\}$, we could get L layer templates by (4):

$$\begin{aligned} x_{l,i} &= (x_i - x_o) \cdot [(l - l_o) \cdot \delta + 1] \\ y_{l,i} &= (y_i - y_o) \cdot [(l - l_o) \cdot \delta + 1] \end{aligned} \tag{4}$$

In which, $l = 1, \dots, L$ is the $l - th$ layer, l_o is the original layer, (x_o, y_o) is the center of the original model, and δ is the scale ratio.

The comparison of detection flows with image pyramids and our multi-scale model are shown in Figure 3. Obviously our detection flow is much faster than the original one. As is shown in Table I, applying our model in detecting one image could significantly promote detection speed.



(a) The detection flow of image pyramids (b) The detection flow of multi-scale model

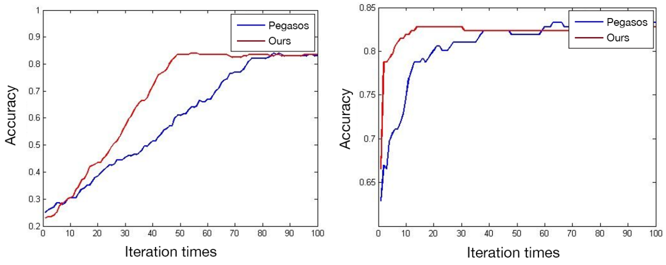
Fig. 3: Detection flows with image pyramids and our multi-scale model respectively

TABLE I: Comparison of detection time on both datasets (/s)

Dataset	Method	Convolution	Pooling	Total
Berry	ABM	2.52	1.99	8.63
	ours	**0.54**	**0.61**	**3.62**
	ϵ	78.3%	69.5%	58.0%
Pedestrian	ABM	0.25	0.22	0.71
	ours	**0.091**	**0.077**	**0.34**
	ϵ	63.9%	65.3%	52.5%

C. Stochastic Gradient algorithm

After constructing feature encoding layer φ, we need to generate the classifier layer, i.e., the parameter ω and b. Following the framework of multiple layer model, we come up with the idea to back propagate the classification error and optimize the key parameters iteratively.



(a) Berry dataset (b) Pedestrian dataset

Fig. 4: Convergence comparison of the Pegasos algorithm and our algorithm

Pegasos [12] is short for primal estimated Gradient solver for SVM. It alternates between stochastic gradient descent steps and projection steps, which has been proved to be more efficient than previously devised SVM solvers. The former needs $\tilde{O}(1/\epsilon)$ times iterations to obtain a solution of accuracy ϵ which costs $\Omega(1/\epsilon^2)$ for the latter. A typical step is a Gradient of the objective function $f(w; A_t)$ at w_t. It is described as $w_{t+\frac{1}{2}} = w_t - \eta_t \bigtriangledown_t$, where $\bigtriangledown_t = \lambda w_t - \frac{1}{|A_t|} \sum_{(x,y)\epsilon A_t^+} yx$. We use Pegasos algorithm in our model's training process with two improvements. The original algorithm randomly select k features from all the training features in each iteration, but we randomly select almost $k/2$ features from positive set and negative set respectively to make a good tradeoffs. The convergence efficiency is shown in Figure 4. Moreover, we reform the learning of b by adding a dimension of w from $1 * n$ (n denotes the number of selected Gabor elements) to $1 * (n + 1)$. Thus, b is also learned during the descent process. By this meaning, the training flow forms a close loop of learning–detecting–training–re-detecting process which makes the training process effective.

IV. ARCHITECTURE OF INFERENCE

After learning the active bases, including the deformable template $B = (B_i = B_{x_i,y_i,s,\alpha_i}, i = 1, \dots, n)$ and the weight vector $\Lambda = (\lambda_i, i = 1, \dots, n)$, our targets could be inferred by sum-max maps with bottom-up scoring and top-down sketching.

The bottom-up operation is decomposed into four steps. Firstly, convolve the input image with Gabor filters at different locations and orientations. Secondly, apply a local maximization operator to the convolved result to find an edge segment at a nearby orientation and location. Thirdly, a local summation operator is applied to compose edge segments that could form the template at this location. Finally we score each composition to decide which one would be the target. The log-

TABLE II: The missing detection rate corresponding to 10^{-1} FPPI.

Method	Pedestrian	Berry
Ours	**4.3%**	**38.4%**
Original	15.2%	56.3%

likelihood is computed as:

$$l(x) = \sum_{i=1}^{n} [\lambda_i max_{(\Delta x, \Delta\alpha)} h(|\langle I, B_{x+x_i+\Delta x, s, \alpha_i+\Delta\alpha}\rangle|^2) \\ - logZ(\lambda_i)] \quad (5)$$

In this section, multiple local maximums corresponding to multiple targets can be obtained based on their confident score. Then, the non-maximum suppression is employed to determine the final detection results.

V. EXPERIMENTS AND EVALUATION

We test our multiple layer model on different datasets including the berry images captured by ourselves and a standard MIT pedestrian database with the resolution of 128*64. Negative samples are randomly excavated from the background and target-free images. We manually label the ground truth in the training datasets and test datasets. The overlap between the detected bounding box and ground truth exceeds 80% would be classified as a true positive. In order to quantify detection performance, in both of the two experiments we take the false positive per image(FPPI) curves as the evaluation standard and take the original model [4] as a baseline.

Figure 5 shows some detection results on the berry dataset. Images in the upper row are from the original berry dataset. The middle row shows the detection results from ABM. The bottom row gives our detection results and the sketch representation. Figure 6 are the hot maps of detection scores on the image domain. The position with brighter color is more likely to be a true positive.

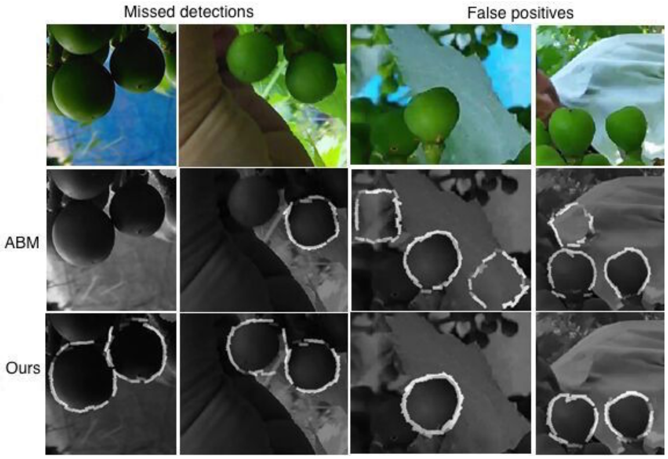


Fig. 5: Comparison of missed detections and false positives

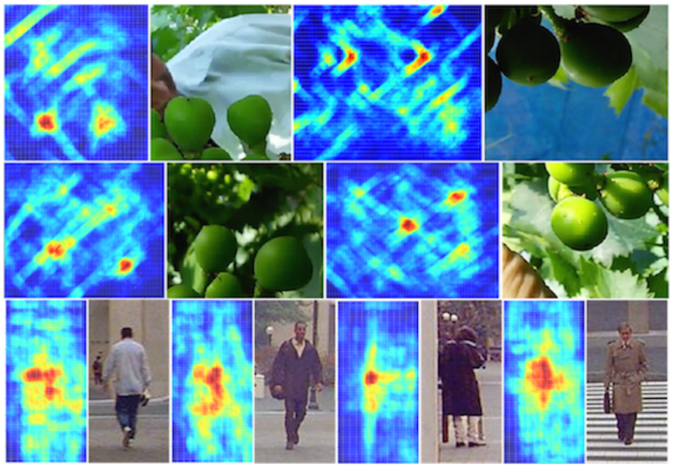
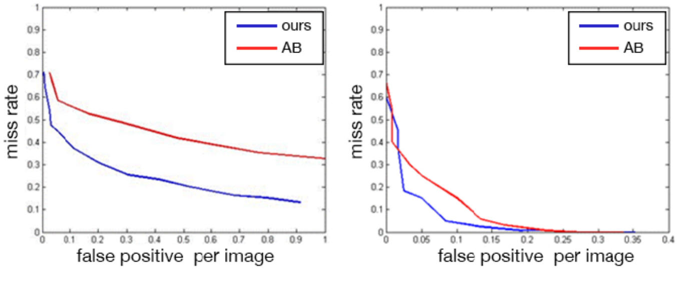


Fig. 6: Hot maps of detection scores on the image domain



(a) berry datasets

(b) pedestrian datasets

Fig. 7: FPPI curves of results on berries and pedestrians.

In practice, the scale definition differs based on the actual scenario. The more scales of the homogeneous targets in one scene the more layers of templates would be needed. In consideration of the size of berries and pedestrians in our databases we set the scales of berry model to 5 and that of pedestrians to 3. We both evaluate the FPPI and the time consumption in dealing with the same image sets.

To better illustrate our experimental results, we could first take a look at the FPPI curves shown in Figure 7 and the time consuming statistics in Table I. As it is shown in Table II, in the pedestrian sets, our approach has fulfilled a noticeable improvement, reducing the missing detection rate from 15.2% to 4.3%. Simultaneously the consumed time has a speed promotion of 58.0%. On the berry image set, we also get a promotion of 17.9% in FPPI and reduce 47.5% in time consuming. Figure 8 demonstrates our model has a good generalization ability in applications of other objects.

VI. CONCLUSION

This paper proposes a multiple layer model for object detection and sketch representation, which can be accomplished by an improved Gradient algorithm. Instead of image pyramids, our multiple layer model can significantly speed up the encoding and detection process. The back-propagation training process can integrate more object related information

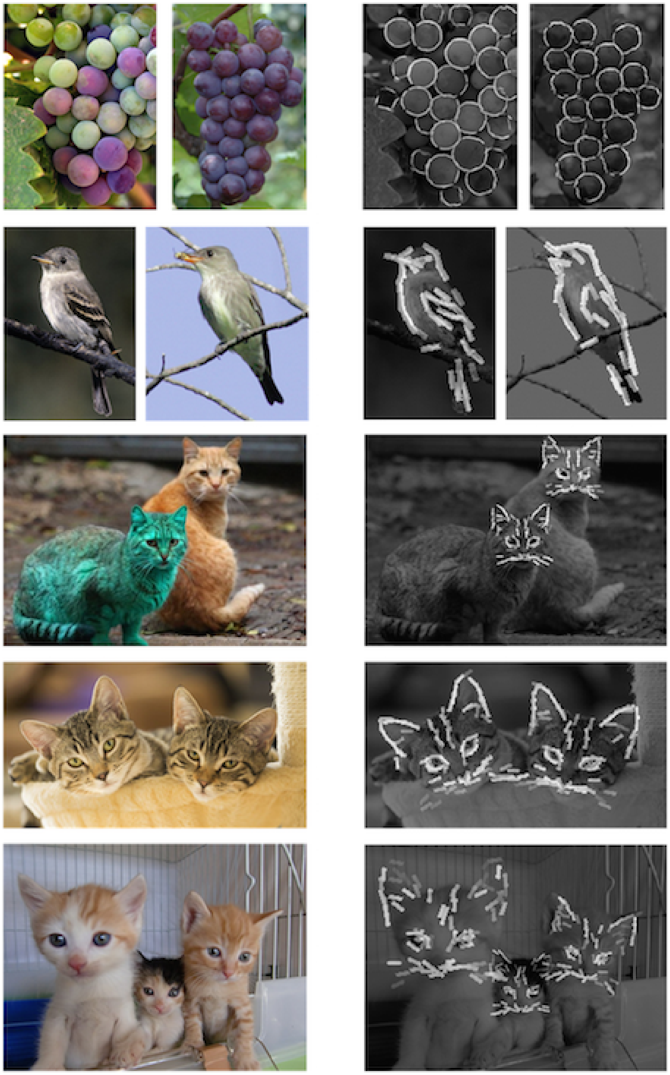


Fig. 8: Extended experiment results on other objects

in the model, as a result it can achieve more robust detection performance. The experimental results demonstrate the effectiveness of our model.

References

[1] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[2] M. A. Fischler and R. A. Elschlager, “The representation and matching of pictorial structures,” *IEEE Transactions on computers*, no. 1, pp. 67–92, 1973.

[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1627–1645, 2010.

[4] Y. N. Wu, Z. Si, H. Gong, and S.-C. Zhu, “Learning active basis model for object detection and recognition,” *International journal of computer vision*, vol. 90, no. 2, pp. 198–235, 2010.

[5] S. G. Mallat and Z. Zhang, “Matching pursuits with time-frequency dictionaries,” *Signal Processing, IEEE Transactions on*, vol. 41, no. 12, pp. 3397–3415, 1993.

[6] B. A. Olshausen *et al.*, “Emergence of simple-cell receptive field properties by learning a sparse code for natural images,” *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[7] R. Benenson, “Detecting objects at 100 hz with rigid templates,” 2015.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[9] S. Zhang, R. Benenson, and B. Schiele, “Filtered channel features for pedestrian detection,” in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015, pp. 1751–1760.

[10] D. Gabor, “Theory of communication. part 1: The analysis of information,” *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, vol. 93, no. 26, pp. 429–441, 1946.

[11] J. G. Daugman, “Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters,” *JOSA A*, vol. 2, no. 7, pp. 1160–1169, 1985.

[12] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, “Pegasos: Primal estimated sub-gradient solver for svm,” *Mathematical programming*, vol. 127, no. 1, pp. 3–30, 2011.