Fast Text/non-Text Image Classification with Knowledge Distillation

Miao Zhao, Rui-Qi Wang, Fei Yin, Xu-Yao Zhang, Lin-Lin Huang, Jean-Marc Ogier

Fast Text/non-text Image Classification with Knowledge Distillation

Miao Zhao*, Rui-Qi Wang†, Fei Yin†, Xu-Yao Zhang†, Lin-Lin Huang*, Jean-Marc Ogier‡
*School of Electronic and Information Engineering, Beijing Jiaotong University,
No.3, Shangyuan Village, Haidian District, Beijing 100044, P.R. China
†National Laboratory of Pattern Recognition, Institute of Automation of Chinese Academy of Sciences,
95 Zhongguancun East Road, Beijing 100190, P.R China
‡University of La Rochelle, La Rochelle, France
Email: 17120036@bjtu.edu.cn, {ruiqi.wang, fyin, xyz}@nlpr.ia.ac.cn, huangll@bjtu.edu.cn, jean-marc.ogier@univ-lr.fr

***Abstract*—How to efficiently judge whether a natural image contains texts or not is an important problem. Since text detection and recognition algorithms are usually time-consuming, and it is unnecessary to run them on images that do not contain any texts. In this paper, we investigate this problem from two perspectives: the speed and the accuracy. First, to achieve high speed for efficient filtering large number of images especially on CPU, we propose using small and shallow convolutional neural network, where the features from different layers are adaptively pooled into certain sizes to overcome difficulties caused by multiple scales and various locations. Although this can achieve high speed but its accuracy is not satisfactory due to limited capacity of small network. Therefore, our second contribution is using the knowledge distillation to improve the accuracy of the small network, by constructing a larger and deeper neural network as teacher network to instruct the learning process of the small network. With the above two strategies, we can achieve both high speed and high accuracy for filtering scene text images. Experimental results on a benchmark dataset have shown the effectiveness of our method: the teacher network yields state-of-the-art performance, and the distilled small network achieves high performance while maintaining high speed which is 176 times faster on CPU and 3.8 times faster on GPU than a compared benchmark method.**

***Keywords*-High speed; Convolutional neural network; Knowledge distillation**

I. Introduction

Scene texts in natural images convey important semantic information for applications such as autonomous driving, security surveillance, real-time translation, image retrieval human-computer interface, etc. Therefore, the detection and recognition of scene texts have attracted tremendous attention [1], in recent years and have made great progress. Among the numerous methods proposed so far, deep neural network (DNN) [2], [3], based ones have achieved superior performance for both text detection and recognition. DNN based methods, however, are computational expensive and usually need parallel computation hardware such as GPU, which is quite energy consuming. When there are millions of images to process per hour in a typical application scenario such as webpages, the time and energy consumption caused by text detection and recognition on all images is unacceptable. According to [4], the majority (76%) of words embedded in images do not appear elsewhere in the main text on WWW pages. In order to obtain this part of text information, text detection and recognition on all images will consume a lot of time. Considering that only 10%-15% of natural images contain texts according to a dataset collected from social networks [5], a fast algorithm for quickly filtering images that contain texts for further processing will largely save the computing.

Compared with previous work on document image [6] and video frame [7], scene text image filtering is a novel and difficult task. Some examples of text and non-text scene images are shown in Fig. 1. The discrimination of text/non-text scene images is not a trivial problem, as the positions of possible texts are unknown, and the problem faces the same challenges as text detection and recognition: cluttered background, variability of text appearance, scale, illumination and perspective. To save gross processing time, a fast filtering algorithm is desired to detect natural images containing texts (briefly called as scene text images) from the pool of images with high recall rate and precision. The detected images occupy a small portion of the pool, such that the gross processing time for deep text detection and recognition can be saved largely.

Previous methods have been proposed for distinguishing text/non-text images based on connected components [8] and neural networks [5], [9]. Connected component based methods perform feature extraction and classification on connected components, but the segmentation of connected components is not trivial due to the cluttering of scene images. Methods based on neural networks, particularly DNNs, can learn good features automatically, and to overcome the variability of text scales and locations, the image are usually divided into regions, and the region-level features extracted by CNN are processed sequentially or independently. However, the computational overhead of the DNN based filtering methods are still considerable.

Although the existing methods have proved that region-level image classification can improve the accuracy of scene text image recognition, it is required to set appropriate criteria carefully for region-level labeling of images. It is a difficult problem how to set such an appropriate criterion.

When the judgment condition is too strict, some text regions are judged as non-text regions, and when the judgment condition is too loose, some image areas that hardly overlap with the text areas are determined as text areas. Therefore, the region-level labels contain a lot of noise. Especially when the network capacity is insufficient, the noise in the region-level labels limit the improvement of classification performance greatly. Therefore, it is difficult for existing methods to discriminate the text and non-text images quickly while ensuring high performance. Moreover, the existing methods are all based on GPU. When these methods run on the CPU, their speed is unsatisfactory. Compared with CPU, GPU is more expensive and requires more energy consumption. Therefore, it is necessary to design an algorithm that can quickly recognize the scene text image on the CPU.

Our proposed algorithm focuses on fast and efficient scene text image filtering. In order to achieve fast processing speed, we design a small convolutional neural network, which classifies the whole image by considering sub-regions of different scales. Nevertheless, due to the impact of network capacity and label noise, it is difficult for the small network to achieve high performance. To tackle this problem, we design a large network which can obtain high classification accuracy. We use the large network as teacher network and the small network as student network. Our method maximizes the transfer of knowledge from the teacher network to the student network through combination of knowledge distillation methods, so that student network can achieve high speed and high performance at the same time. We conduct two-stage training on the student network. In the first stage, the student network is well initialized by activating boundary knowledge distillation. In the second stage, we solve the problem of region-level label containing noise by means of soft label knowledge distillation which fits region-level prediction of teacher network, and improve the performance of student network by means of adversarial knowledge distillation which fits region-level features of teacher network.

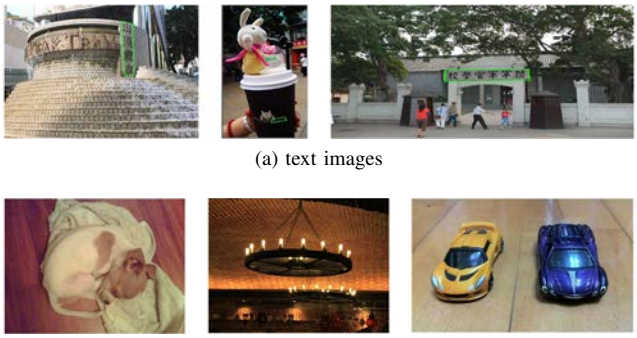
The rest of this paper is organized as follwed. Section II reviews related work. In Section III, we show the details of our method. In section IV, we validate our method by experiments on the benchmark dataset. Section V concludes our work.

II. Related Work

In this section, we introduce some methods of scene text image filtering and knowledge distillation. Moreover, the motivation of our algorithm is also discussed.

A. Scene Text Image Filtering

To the best of our knowledge, there are few researches on scene text image filtering. CNN coding [10] adopts the method of combining Maximally Stable Extremal Regions (MSER), CNN and Bag-of-Words (BoW) to conduct scene text image filtering. CMDRNN [9] extracts the image feature through CNN, and then obtains the region-level features which contain context information by RNN. In the end, this method determines whether the whole image contains text by judging whether the local region of the image contains text or not. MSP-Net [5] divides the image into image sub-regions of different scales, and predicts if these image sub-regions contain text by CNN, so as to predict whether the whole image contain text or not.



(a) text images

(b) non-text images

Figure 1. Examples for text image and non-text image.

MSP-Net can achieve state-of-the-art performance in the scene text image filtering task. However, MSP-Net requires a huge network to achieve the scene text image filtering. When the device resource is limited, MSP-Net can not meet the need of real-time processing. Another disadvantage of MSP-Net is that the noise of region-level labels hinders the further improvement of network performance. Inspired by MSP-Net, we also adopt the method of obtaining image-level prediction through multi scale region-level predictions. The main difference compared with MSP-Net is that we use a very lightweight network, which can achieve fast scene text image filtering in both CPU and GPU. Through knowledge distillation, it can obtain relatively high performance.

B. Knowledge Distillation

Knowledge distillation is an important model acceleration method, which can make the network performance improvement by knowledge transfer. Hinton et al. [11] first put forward the concept of knowledge distillation and believed that knowledge transfer could be realized by fitting the soft label generated by the teacher network. Romero et al. [12] extends the knowledge distillation to the feature map of intermediate layers in DNN. Zagoruyko et al. [13] proposed that knowledge transfer between networks can be accomplished by means of attention map. Byeongho Heo et al. [14] proposed to transfer the activation boundary formed by hidden neurons in the pre-training stage to obtain efficient knowledge distillation. Peiye Liu et al. [15] and Wei-Chun Chen et al. [16] enabled the student network to imitate the intermediate representations of the teacher network by means of adversarial strategies.

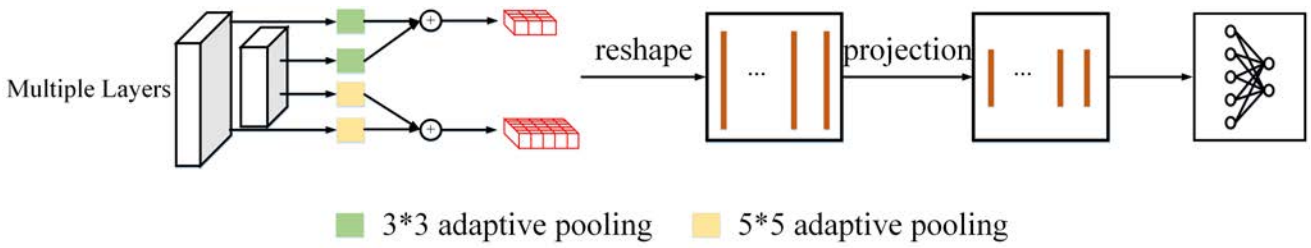


Figure 2. Architechture of teacher network

III. Proposed Method

In this section, we will present the details of our proposed method.

A. Teacher Network

Inspired by MSP-Net [5], we design the teacher network as shown in the Fig. 2. In the teacher network, we use the feature extraction network of VGG16 [17] as the image feature extractor. The teacher network does not require the size of the input image to be fixed, so the features extracted from the input image of uncertain shape $C*W*H$ should have uncertain shape $C' * W' * H'$. In order to obtain region-level features, image features from the last two layers are pooled to fixed sizes (5*5, 3*3). To obtain richer information, we add feature maps of the same size together to obtain multi-scale feature maps.

The features on the feature map correspond to different image sub-regions. In order to instruct the student network, region-level feature is linear mapped in batch into a new space. These mapped regional-level features will be used for knowledge distillation. The last layer of the network is used to output the prediction of each sub-region of the image. Image-level prediction is obtained by logic OR on the region-level prediction.

B. Student Network

Student network is more straightforward than teacher network and its architecture is shown in Fig. 3. It contains 6 convolution layers as the image feature extractor. To improve the speed of the student network, we do not fuse multiple layers of features. It obtains multi scale features directly from the last two convolution layers after regularization. The student network obtains the feature maps with the size of 5*5 or 3*3 after adaptive pooling. Then we use the two-layer full-connected network to process the region-level features. The first layer maps the region-level features from the original space to the feature space shared with the teacher network. These region-level features will be used for knowledge distillation. Afterwards, the features are used for classifying the different regions in the origin image, and the prediction of whole image is the logical OR of predictions of all regions.

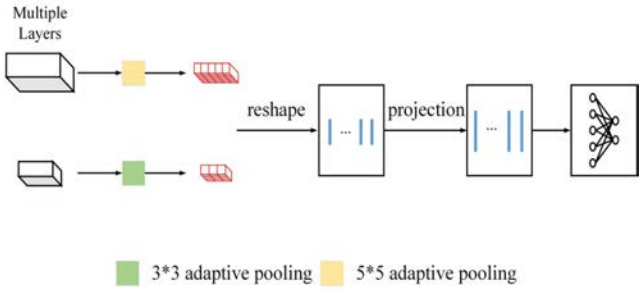


Figure 3. Architechture of student network

C. Adversarial Knowledge Distillation Network

In the second stage of student network training, the adversarial knowledge distillation network is shown in Fig. 4. Adversarial knowledge distillation network consists of three parts: teacher network, student network and discriminant network. The teacher network and the student network simultaneously output the predictions of each image region and the region-level features mapped to the same space. In the part of soft label knowledge distillation, we take the regional-level prediction of teacher network as the target, and then calculate the cross entropy loss. In the part of adversarial knowledge distillation, we draw lessons from the thought of GAN [18]. We regard the region-level feature distribution of teacher network as the real distribution, and the region-level feature distribution of student network as the fake distribution. In the process of training, the discriminant network learns to judge whether the input region-level features come from the teacher network or not, and student network learns to cheat the discriminant network during the training process. With the adversarial knowledge distillation, the student network will imitate the teacher network to produce region-level features. BEGAN [19] can be easily trained in many variations of GAN [20], [19]. Encouraged by BEGAN, we use the auto-encoder as the discriminant network. In the form of the combination of adversarial knowledge distillation and soft label knowledge distillation, the student network learns from the region-level soft labels and region-level features generated by the teacher network.

D. Network Training

In the training process of teacher network, we need to divide the image into sub-regions of different scales(3*3, 5*5), and then determine the labels of different sub-regions through a judgment condition. In MSP-Net [5], only when the proportion of the text image area exceeds 0.05 and the height of the text line exceeds 1/2 of the height of the image area, the corresponding label is set to 1. However, text has the feature of scale diversity, and some text image areas can not meet the condition. Finally, as long as the overlapping ratio between the text area and the image area reaches 0.01, we set the label of the image area as 1. The following experiments prove that this setting can significantly improve the performance of scene text image filtering.

The training of student network is divided into two stages: pre-training and formal training.

In the pre-training stage, we use activation boundary knowledge distillation to train the student network. ReLU activation divided the input feature space into two parts, and Byeongho Heo et al. [14] regarded the hyperplane between the two parts as the activation boundary. By fitting the activation boundary, the student network can learn a lot of knowledge from the teacher network. In our work, we hope that the student network can transfer the activation boundary from the teacher network in the feature space of image sub-region. The activation boundary loss here is calculated as follows:

$$L_A = \frac{1}{M}\sum_{i=1}^{M}\left\| d\left(T_i\right) \odot r\left(u - S_i\right) + \left(1 - d\left(T_i\right)\right) \odot r\left(u + S_i\right)\right\|_2^2 \tag{1}$$

where d is indicator function, it outputs 0 when its input value is greater than 0, or it outputs 1. r stands for ReLU. u is a margin value. $\odot$ represents element-wise product of vectors. These symbols described below are consistent in later expressions. T_i is the mapped region-level feature extracted by teacher network. S_i is the mapped region-level feature extracted by student network. The subscription i represents the i-th sub-regions of the input image. M is the number of subregions obtained by image segmentation(34 in this paper).

Through further analysis, we can find that activation boundary transfer enables the response value of students' neurons to approach a positive critical value u when the teacher network's neurons are activated, while the corresponding value of student network's neurons approaches $-u$ when the teacher network's neurons are deactivated. In this way, the student network imitates the activation boundary of the regional-level features of the teacher network and maximizes the activation boundary interval.

In the formal training phase, we optimize the student network which has be well initialized. In this stage we use both soft label knowledge distillation and adversarial knowledge distillation. The discriminant network and student network are optimized alternately. The loss of the student network is calculated as follows:

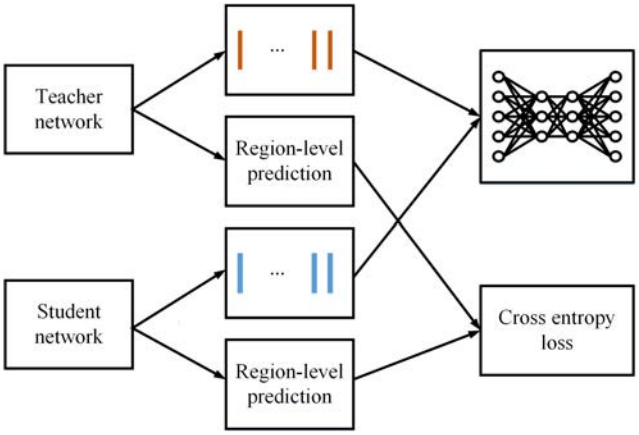


Figure 4. Architechture of adversarial knowledge distillation network

Figure 5. Visual results of text image prediction

$$L_{Stu} = L_K + bL_G \tag{2}$$

L_K represents the soft label knowledge distillation loss function. L_G represents the adversarial knowledge distillation loss function. b (0.2 in this paper) is a loss balance parameter. With the minimization of the loss function, the student network can generate region-level representation similar to the teacher network and accurately predict the image region.

The loss function of soft label knowledge distillation is defined as follows:

$$L_k = -\frac{1}{M}\sum_{i=1}^{M}\left(l_i \log p_i + \left(1 - l_i\right)\log\left(1 - p_i\right)\right) \tag{3}$$

l_i denotes the probability of that certain region, which is predicted to be positive by the teacher network. p_i denotes the probability that certain region is predicted to be positive by the student network.

In the adversarial knowledge distillation, the discriminant network is expected to judge whether the input features come from the teacher network or from the student network, and the student network is expected to be able to deceive the discriminant network. When the training reaches equilibrium, The regional-level feature distribution generated

Table I
RESULTS COMPARISION OF DIFFERENT METHOD. IT SHOULD BE NOTICED THAT MSP-NET IS TRAINED WITH LARGER INPUT IMAGE HIGHT 500 PIXEL. * INDICATES THAT THE RELEVANT DATA IS ENTIRELY FROM THE ORIGINAL PAPER.

Method	Recall	Precision	F-measure	Time Cost (CPU)	Time Cost (GPU)	Params
CNN Coding*	89.8%	90.3%	90.1%	-	460ms	-
CMDRNN*	90.4%	93.3%	91.8%	-	90ms	-
MSP-Net	95.4%	93.7%	94.6%	9.53s	5.38ms	136.40M
Teacher Network	97.5%	94.0%	95.7%	2.24s	2.60ms	16.03M
Student Network	94.0%	92.0%	93.0%	54.03ms	1.40ms	0.087M

Table II
EFFECT OF KONWLEGE DISTILLATION

Method	Recall	Precision	F-measure
Hard label	91.0%	88.0%	89.5%
Soft label	92.9%	90.2%	91.5%
Soft label + Adversarial	94.3%	90.7%	92.5%
Our proposed	94.0%	92.0%	93.0%

by the student network can approximately fit the regional-level features generated by the teacher network. According to BEGAN [19], we construct the adversarial knowledge distillation loss function as follows:

$$L_D = \frac{1}{M}\sum_{i=1}^{M}\left|D\Big(T_i\Big) - T_i\right| - \frac{k_t}{M}\sum_{i=1}^{M}\left|D\Big(S_i\Big) - S_i\right| \quad (4)$$

$$L_G = \frac{1}{M}\sum_{i=1}^{M}\left|D\Big(S_i\Big) - S_i\right| \quad (5)$$

L_D is the loss function of the discriminant network. L_G is the adversarial knowledge distillation loss function of the student network. D is the discriminant network. k_t is a control parameter, and it can control how much emphasis is put on L_G. k_t (the initial value is 0) will change with the training process, and its expression is calculated as follows:

$$\begin{aligned} k_{t+1} =& k_t + f_k * \Big(\frac{n}{M}\sum_{i=1}^{M}\left|D\Big(T_i\Big) - T_i\right| - \\ & \frac{1}{M}\sum_{i=1}^{M}\left|D\Big(S_i\Big) - S_i\right|\Big) \end{aligned} \quad (6)$$

f_k (1e-3 in this paper) controls the change rate of k_t . n (0.75 in this paper) is a relaxation variable. n controls how much the student network needs to fit the regional-level feature distribution of the teacher network when the training reaches equilibrium.

IV. EXPERIMENTS

In this section, we perform several experiments to validate the effectiveness and efficiency of our method on a challenging public benchmark for text/non-text natural image discrimination.

A. Dataset

We use the dataset named TextDis released by Zhang et al. [10], which focuses on text image discrimination. In this dataset, the train set contains 5302 text images and 6000 non-text images, the test set contains 2000 text images and 2000 non-text images. Each text image has a ground truth file which contains the bounding box information of text it contains. We split the images into 3 * 3 and 5 * 5 regions for experiments. For text images, with bounding boxes, we can obtain the overlap area of certain image region with all text regions. A threshold (1% particularly in this paper) is set to label image regions positive according to the ratio of overlap area over image region. Apparently regions from non-text images are all labeled negative. Although both teacher network and student network have no requirements for input size, we rescale the input to make the short edge 256 pixel. The input is normalized before feed into the network. Data is augmented with flip and rotation.

B. Training Details

We use Stochastic Gradient Descent (SGD) with mini-batch size of 1 to optimize the training process. For teacher network, learning rate is 1e-4. For student network, the whole training process consists of two stages. In the pre-training stage, the learning rate is 1e-5. In the formal training stage, the initial learning rate of the student network is 1e-4, and when the train loss stop improving, the learning rate is reduced to 1/10. The learning rate of the discriminator network is always twice that of the student network.

C. Experiment Results

We test our method on TextDis dataset and compare it with MSP-Net [5] on precision, recall, F-measure ,time cost and parameters. Time cost means only time cost during model computing otherwise data IO consumes most time and makes the difference insignificant. CPU time cost experiments are performed on a platform with one Intel(R) Xeon(R) CUP E5-2620 0 @2.00GHz and 256G RAM. GUP time cost experiments are performed on a single Nvidia(R) GTX TITAN with 12G VRAM.

The experiments result on TextDis dataset is showed in TABLE I. As shown in the Fig. 2, our network can further give the prediction results of each region, which is conducive to subsequent text detection and recognition. The recall, precision and F-measure of MSP-Net is from

[5]. We reimplement MSP-Net and test the runing time on GPU and CPU. The results of CNN Coding and CMDRNN are from [10], [9].All models are developed with Pytorch 4.0.1. Since we use the same experimental dataset, the experimental results are comparable. Compared with MSP-Net [5], our teacher network achieves higher speed under higher performance. This proves that the more simplified network structure can achieve higher performance under appropriate text region discrimination conditions, and further proves that the regional-level annotation will greatly affect the network performance.

To demonstrate the validity of our knowledge distillation method, we conduct further comparative experiments. The experiments result is showed in TABLE II. The region-level labels obtained by judging conditions are called hard label. Compared with the region-level labeling by discriminating conditions, soft label knowledge distillation significantly improves the performance of student network. The experiment proves that soft label knowledge distillation can solve the problem of region-level labels containing noise. Furthermore, the experimental results show that the adversarial knowledge distillation can improve the performance of student network. Finally, the student network trained through the full training process achieves the best performance. It shows that the pre-training makes the student network get a good initialization.

V. Conclusion

In this paper, we propose a method for fast text/non-text natural images classification. It can handle the difficulties caused by various locations and multi-scale. We enable student network to be adequately pre-trained by activating boundary knowledge distillation. Through the soft label knowledge distillation and the adversarial knowledge distillation, our method obtains the high performance. Experiments on public dataset shows our method achieves super fast speed than any other methods and keeps high accuracy. In the future, we would like to integrate our method as a pretreatment module into text analysis system.

VI. Acknowledgement

This work is supported by Natural Science Foundation of China under Grant No. 61733007, 61573355, 61721004.

References

[1] Y. Zhu, C. Yao, and X. Bai, "Scene text detection and recognition: Recent advances and future trends," *Frontiers of Computer Science*, vol. 10, no. 1, pp. 19–36, 2016.

[2] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Transactions on Multimedia*, 2018.

[3] T. He, W. Huang, Y. Qiao, and J. Yao, "Text-attentional convolutional neural network for scene text detection," *IEEE transactions on image processing*, vol. 25, no. 6, pp. 2529–2541, 2016.

[4] A. Antonacopoulos, D. Karatzas, and J. Ortiz-Lopez, "Accessing textual information embedded in internet images," in *Internet Imaging II*, vol. 4311. International Society for Optics and Photonics, 2000, pp. 198–206.

[5] X. Bai, B. Shi, C. Zhang, X. Cai, and L. Qi, "Text/non-text image classification in the wild with convolutional neural networks," *Pattern Recognition*, vol. 66, pp. 437–446, 2017.

[6] A. Delaye and C.-L. Liu, "Text/non-text classification in online handwritten documents with conditional random fields," in *Chinese Conference on Pattern Recognition*. Springer, 2012, pp. 514–521.

[7] N. Sharma, P. Shivakumara, U. Pal, M. Blumenstein, and C. L. Tan, "Piece-wise linearity based method for text frame classification in video," *Pattern Recognition*, vol. 48, no. 3, pp. 862–881, 2015.

[8] V. P. Le, N. Nayef, M. Visani, J.-M. Ogier, and C. De Tran, "Text and non-text segmentation based on connected component features," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2015, pp. 1096–1100.

[9] P. Lyu, B. Shi, C. Zhang, and X. Bai, "Distinguishing text/non-text natural images with multi-dimensional recurrent neural networks," in *Pattern Recognition (ICPR), 2016 23rd International Conference on*. IEEE, 2016, pp. 3981–3986.

[10] C. Zhang, C. Yao, B. Shi, and X. Bai, "Automatic discrimination of text and non-text natural images," in *Document Analysis and Recognition (ICDAR), 2015 13th international conference on*. IEEE, 2015, pp. 886–890.

[11] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[12] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

[13] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1612.03928*, 2016.

[14] B. Heo, M. Lee, S. Yun, and J. Y. Choi, "Knowledge transfer via distillation of activation boundaries formed by hidden neurons," *arXiv preprint arXiv:1811.03233*, 2018.

[15] P. Liu, W. Liu, H. Ma, T. Mei, and M. Seok, "Ktan: Knowledge transfer adversarial network," *arXiv preprint arXiv:1810.08126*, 2018.

[16] W.-C. Chen, C.-C. Chang, C.-Y. Lu, and C.-R. Lee, "Knowledge distillation with feature maps for image classification," *arXiv preprint arXiv:1812.00660*, 2018.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[19] D. Berthelot, T. Schumm, and L. Metz, "Began: boundary equilibrium generative adversarial networks," *arXiv preprint arXiv:1703.10717*, 2017.

[20] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.