

Quadratic mutual information regularization in real-time deep CNN models

Maria Tzelepi and Anastasios Tefas
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
Email: {mtzelepi,tefas}@csd.auth.gr

Abstract—In this paper, regularized lightweight deep convolutional neural network models, capable of effectively operating in real-time on devices with restricted computational power for high-resolution video input are proposed. Furthermore, a novel regularization method motivated by the Quadratic Mutual Information, in order to improve the generalization ability of the utilized models is proposed. Extensive experiments on various binary classification problems involved in autonomous systems are performed, indicating the effectiveness of the proposed models as well as of the proposed regularizer.

Index Terms—Quadratic Mutual Information, Regularizer, Lightweight Models, Real-time, Deep Learning.

I. INTRODUCTION

Deep Learning (DL) models, [1], and especially deep Convolutional Neural Networks (CNN) have been established among the most efficient research directions in a wide range of computer vision tasks, eclipsing previous shallow algorithms, [2]. However, state-of-the-art DL models are usually computation-heavy, obstructing their application on autonomous systems. This has directed the research towards the development of lightweight models capable of running on devices with restricted computational resources such as mobile phones and embedded systems, [3], [4].

Thus, in this paper, we propose lightweight deep CNN models allowing real-time deployment for high resolution images for specific binary classification problems involved in autonomous robots applications. Specifically, we consider the media coverage of certain events by UAVs (also known as drones). We deal with face, bicycle, and football player detection, as well as crowd detection towards human crowd avoidance. Our goal is to provide semantic heatmaps, by predicting for each location within the captured high-resolution scene the considered object's presence. That is, we train models with RGB input of size either 32×32 or 64×64 , and then test images are introduced to the network, and, through sliding window process, for every window 32×32 or 64×64 respectively, we compute the output of the network at the last convolutional layer. An example of a football player heatmap is provided in Fig.1.

Surveying the relevant literature we can see that several works have emerged towards designing lightweight models. For example, an approach that proposes to replace 3×3 convolutions with 1×1 convolutions to create a very small network capable of reducing $50 \times$ the number of parameters while



Fig. 1: An image containing football players and the corresponding predicted heatmap of football player presence.

obtaining high accuracy is proposed in [4]. However, there is no other work in the recent literature proposing real-time models capable of running on devices with limited computational resources for high resolution input. Additionally, in [5] where a computation-efficient CNN model is proposed for mobile devices with limited computing power, it is demonstrated that in order to accomplish real-time deployment, someone has to reduce the input frame resolution to 224×224 .

Furthermore, since we deal with lightweight models that usually have inferior performance compared to the more complex ones, we focus on enhancing their performance. That is, the second goal of this work is to propose a novel regularization method in order to circumvent over-fitting and enhance the generalization ability of the proposed real-time models. Generally, this constitutes a major issue in deep learning algorithms, since neural networks are prone to over-fitting due to their high capacity. During the past years, several regularization schemes have been proposed in order to prevent overfitting in neural networks, e.g. common regularization methods, like $L1/L2$ regularization, and Dropout [6]. Besides, multitask-learning [7] has been proposed as a way to improve the generalization ability of a model. For example, in [8] the authors introduced techniques developed in semi-supervised learning in the deep learning domain, whilst in [9], a novel CNN architecture with an SVM classifier at every hidden layer is proposed. This companion objective acts as a kind of feature regularization.

In this work, the so-called *Mutual Information (MI) regularizer* is proposed. The proposed regularizer is inspired by the Quadratic Mutual Information (QMI) measure [10], which is a variant of the commonly used Mutual Information, an

information-theoretic measure of dependence between random variables. That is, apart from the classification loss, we propose to attach an additional optimization criterion based on the QMI. Recently, QMI reformulated to produce a kernel dimensionality reduction method under the Graph Embedding framework [11], while in [12] a Probabilistic Knowledge Transfer method proposed exploiting the QMI. It is noteworthy that the proposed regularizer is generic and can be applied in several deep learning architectures for classification purposes.

The remainder of the manuscript is structured as follows. The utilized CNN architectures are described in Section II. The proposed MI regularizer is presented in Section III. The experiments conducted to validate the proposed method are provided in Section IV. Finally, the conclusions are drawn in Section V.

II. REAL-TIME CNN MODELS

In this paper, our goal is to propose effective deep models for various binary classification problems, which allow real-time deployment (about 25 frames per second) on-drone for high resolution images. It should be emphasized that it is of utmost importance for the application to handle high resolution images, since objects in drone-captured images are extremely small, and thus image resizing in order to reach real-time deployment limits, would further shrink the object of interest, rendering the detection infeasible. An example that highlights the demand for high resolution images is provided in Fig. 3. That is, an aerial image that contains bicycles (bicycles with bicyclists) is provided in Fig. 3a, and the resulting heatmaps for input of two different resolutions, utilizing the proposed model are provided in Figs. 3b-3c. As it can be observed, as the resolution increases, better performance can be achieved.

The objective of this work is two-fold: a) to propose real-time architectures that can be deployed on-drone, and b) to improve the state-of-the-art performance using MI regularization. Thus, we propose a model consisting of only five convolutional layers, by discarding the deepest layers and pruning filters of the widely used VGG-16 model [13]. That is, we use the first four convolutional layers of the VGG-16 model with pruned filters, while the last convolutional layer consists of two channels, each for a class, since we deal with binary classification problems. The proposed model runs in real-time on-drone on test images of 1080p (1920×1080) resolution, utilizing the sliding window process, as previously mentioned. The model is abbreviated as VGG-1080p based on this attribute. Since we deal with datasets of 32×32 and 64×64 input dimensions, we propose two variant models. The models use same kernels and channels, and real-time deployment is achieved with appropriate stride and pooling, as it is shown in Fig. 2. The evaluation results on the deployment speed for the proposed models are provided in the Experiments Section.

III. THE PROPOSED MI REGULARIZER

In this paper, a novel regularizer motivated by the Quadratic Mutual Information [10] is proposed. Apart from the classification loss, we propose a regularization loss derived from the so-

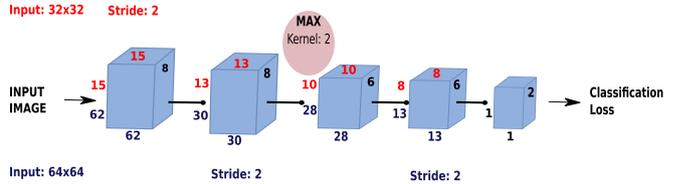


Fig. 2: VGG-1080p Architecture: Details for input of size 32×32 are printed in red, while details for input of size 64×64 are printed in blue.

called information potentials of the QMI. Thus, in this Section, we first introduce the Mutual Information and its quadratic variant, and then we present the proposed MI regularizer.

We assume a random variable Y representing the image representations of the feature space generated by a specific deep neural layer. We also assume a discrete-value variable C that represents the class labels. For each feature representation \mathbf{y} there is a class label c . The MI measures dependence between random variables, first introduced by Shannon, [14]. That is, the MI measures how much the uncertainty for the class label c is reduced by observing the feature vector \mathbf{y} . Let $p(c)$ be the probability of observing the class label c , and $p(\mathbf{y}, c)$ the probability density function of the corresponding joint distribution.

The MI between the two random variables is defined as:

$$MI(Y, C) = \sum_c \int_{\mathbf{y}} p(\mathbf{y}, c) \log \frac{p(\mathbf{y}, c)}{p(\mathbf{y})P(c)} d\mathbf{y}, \quad (1)$$

where $P(c) = \int_{\mathbf{y}} p(\mathbf{y}, c) d\mathbf{y}$. MI can also be interpreted as a Kullback-Leibler divergence between the joint probability density $p(\mathbf{y}, c)$ and the product of marginal probabilities $p(\mathbf{y})$ and $P(c)$.

QMI is derived by replacing the Kullback-Leibler divergence by the quadratic divergence measure [10]. That is:

$$QMI(Y, C) = \sum_c \int_{\mathbf{y}} (p(\mathbf{y}, c) - p(\mathbf{y})P(c))^2 d\mathbf{y}. \quad (2)$$

And thus, by expanding eq. (2) we arrive at the following equation:

$$\begin{aligned} QMI(Y, C) &= \sum_c \int_{\mathbf{y}} p(\mathbf{y}, c)^2 d\mathbf{y} + \sum_c \int_{\mathbf{y}} p(\mathbf{y})^2 P(c)^2 d\mathbf{y} \\ &\quad - 2 \sum_c \int_{\mathbf{y}} p(\mathbf{y}, c) p(\mathbf{y}) P(c) d\mathbf{y}. \end{aligned} \quad (3)$$

The quantities appearing in eq. (3), are called *information potentials* and they are defined as follows: $V_{IN} = \sum_c \int_{\mathbf{y}} p(\mathbf{y}, c)^2 d\mathbf{y}$, $V_{ALL} = \sum_c \int_{\mathbf{y}} p(\mathbf{y})^2 P(c)^2 d\mathbf{y}$, $V_{BTW} = \sum_c \int_{\mathbf{y}} p(\mathbf{y}, c) p(\mathbf{y}) P(c) d\mathbf{y}$, and thus, the quadratic mutual information between the data samples and the corresponding class labels can be expressed as follows in terms of the information potentials:

$$QMI = V_{IN} + V_{ALL} - 2V_{BTW}. \quad (4)$$

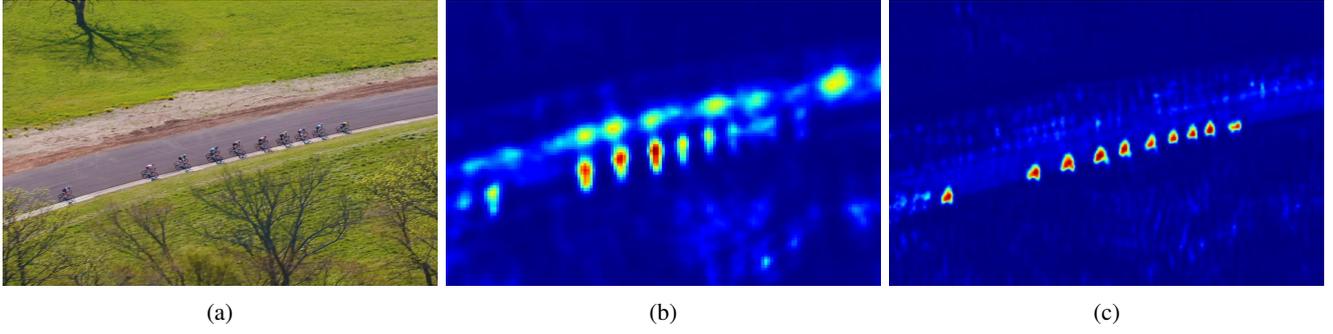


Fig. 3: An aerial high resolution image containing bicycles (3a), and the resulting heatmaps for input of size 640×480 (3b) and for input of size 1920×1080 (3c) utilizing the proposed VGG-1080p model trained for bicycle detection.

If we assume that there are N_c different classes, each of them consisting of J_p samples, the class prior probability for the c_p class is given as: $P(c_p) = \frac{J_p}{N}$, where N corresponds to the total number of samples. Kernel Density Estimation [15] can be used to estimate the joint density probability: $p(\mathbf{y}, c_p) = \frac{1}{N} \sum_{j=1}^{J_p} K(\mathbf{y}, \mathbf{y}_{pj}; \sigma^2)$, for a symmetric kernel K , with width σ , where we use the notation \mathbf{y}_{pj} to refer to the j -th sample of the p -th class, as well as the probability density of Y as $p(\mathbf{y}) = \sum_{p=1}^{N_c} p(\mathbf{y}, c_p) = \frac{1}{N} \sum_{j=1}^N K(\mathbf{y}, \mathbf{y}_j; \sigma^2)$.

Thus, eq. (4) is formulated as follows:

$$V_{IN} = \frac{1}{N^2} \sum_{p=1}^{N_c} \sum_{k=1}^{J_p} \sum_{l=1}^{J_p} K(\mathbf{y}_{pk}, \mathbf{y}_{pl}; 2\sigma^2), \quad (5)$$

$$V_{ALL} = \frac{1}{N^2} \left(\sum_{p=1}^{N_c} \left(\frac{J_p}{N} \right)^2 \right) \sum_{k=1}^N \sum_{l=1}^N K(\mathbf{y}_k, \mathbf{y}_l; 2\sigma^2), \quad (6)$$

$$V_{BTW} = \frac{1}{N^2} \sum_{p=1}^{N_c} \frac{J_p}{N} \sum_{j=1}^{J_p} \sum_{k=1}^N K(\mathbf{y}_{pj}, \mathbf{y}_k; 2\sigma^2). \quad (7)$$

The kernel function $K(\mathbf{y}_i, \mathbf{y}_j; \sigma^2)$ expresses the similarity between two samples i and j . There are several choices for the kernel function, [15]. For example, in [10] the Gaussian kernel is used, while in [12] the authors utilize a cosine similarity based kernel to avoid defining the width, in order to ensure that a meaningful probability estimation is obtained, since finetuning the width of the kernel is not a straightforward task, [16]. In our experiments, we use as kernel metric a Euclidean based similarity, defined as $K_{ED} = \frac{1}{1 + \|\mathbf{y}_i - \mathbf{y}_j\|_2^2}$, which also absolves us from defining the width of the kernel.

The pairwise interactions described above between the samples can be interpreted as follows: V_{IN} expresses the interactions between pairs of samples inside each class. V_{ALL} expresses the interactions between all pairs of samples, regardless of the class membership. Finally, V_{BTW} expresses the interactions between samples of each class against all other samples.

Thus, motivated by the QMI, in this work we propose a novel regularizer in order to enhance the generalization ability of a deep model. That is, apart from the optimization criterion

defined by the hinge loss function which aims at separating the samples belonging to different classes, we propose an additional optimization criterion utilizing the information potential defined in eq. (4). We assume that the hinge loss preserves the V_{BTW} information potential which aims to separate samples belonging to different classes. Then, our objective is to maximize pairwise interactions between the samples described by the $V_{IN} + V_{ALL}$ quantities. The derived joint optimization criterion defines an additional loss function, which is attached to the penultimate convolutional layer (that is the last convolutional layer, before the one utilized for the classification task) and acts as regularizer to the main classification objective.

$$J_{MI} = -(V_{IN} + V_{ALL}), \quad (8)$$

where:

$$V_{IN} = \frac{1}{N^2} \sum_{p=1}^{N_c} \sum_{k=1}^{J_p} \sum_{l=1}^{J_p} K_{ED}(\mathbf{y}_{pk}, \mathbf{y}_{pl}), \quad (9)$$

and

$$V_{ALL} = \frac{1}{N^2} \left(\sum_{p=1}^{N_c} \left(\frac{J_p}{N} \right)^2 \right) \sum_{k=1}^N \sum_{l=1}^N K_{ED}(\mathbf{y}_k, \mathbf{y}_l). \quad (10)$$

Considering binary classification problems the above optimization criteria can be formulated as follows:

$$\begin{aligned} V_{IN} &= \\ &= \frac{1}{N^2} \sum_{k=1}^{J_1} \sum_{l=1}^{J_1} K_{ED}(\mathbf{y}_{1k}, \mathbf{y}_{1l}) + \frac{1}{N^2} \sum_{k=1}^{J_2} \sum_{l=1}^{J_2} K_{ED}(\mathbf{y}_{2k}, \mathbf{y}_{2l}), \end{aligned} \quad (11)$$

and

$$V_{ALL} = \frac{1}{N^2} \left(\frac{J_1^2 + J_2^2}{N^2} \right) \sum_{k=1}^N \sum_{l=1}^N K_{ED}(\mathbf{y}_k, \mathbf{y}_l), \quad (12)$$

The total loss for the network training is defined as:

$$J_{total} = J_{class} + \eta J_{MI}, \quad (13)$$

where J_{class} stands for the classification loss, and the parameter $\eta \in [0, 1]$ controls the relative importance of J_{MI} . We solve the above optimization problem using gradient descent. We should note that the proposed regularizer can be applied for the whole dataset, as well as in terms of mini-batch training. In our experiments we utilize the mini-batch mode. We should finally note that in the regularized training we utilize the hinge loss since, as we have experimentally observed, it performs better than the cross entropy one in binary classification problems, however the cross entropy loss could also be utilized.

IV. EXPERIMENTS

In this Section, we present the experiments conducted in order to evaluate the proposed models regarding the deployment speed as well as the proposed regularization method. We evaluate the detection speed in terms of frames per second (fps), while we use test accuracy to evaluate the proposed regularizer, since we deal with balanced datasets. Each experiment is executed five times, and the mean value and the standard deviation are reported, considering the maximum value of the test accuracy for each experiment. The probabilistic factor is the random weight initialization.

A. Datasets

In order to evaluate the performance of the proposed regularizer we perform experiments on four datasets constructed for Football Player, Face, Bicycles, and Crowd detection. The Football Player dataset, [17], consists of 98,000 train and 10,000 test images of size 32×32 that contain equal number of football players and non-football players. The face dataset contains 70,000 train and 7,468 test images of size 32×32 that contain equal number of face and non-face images. Images of faces have been randomly selected from the AFLW [18], MTFE [19], and WIDER FACE [20] datasets. The Bicycles dataset, [17], contains 51,200 equally distributed train images of bicycles and non-bicycles, and correspondingly a test set of 10,000 images. Input images are of size 64×64 . Finally, the Crowd-Drone dataset, [17], contains 40,000 UAV-captured train images of equal number of crowded scenes and non-crowded scenes, and 11,550 equally distributed crowded and non-crowded test images. Input images are of size 64×64 .

B. Implementation Details

All the experiments conducted using the Caffe Deep Learning framework. We use the mini-batch gradient descent for the networks' training. That is, an update is performed for every mini-batch of N_b training samples. The learning rate is set to 10^{-3} and drops to 10^{-4} gradually, and the batch size is set to 256. The momentum is 0.9. All the models are trained on an NVIDIA GeForce GTX 1080 with 8GB of GPU memory for 100 epochs, and can run in real-time when deployed on an NVIDIA Jetson TX2. In this work, the parameter η in (13) which controls the relative importance of the proposed regularizer's loss, is set to 0.001, since we have seen that in most cases provides best performance. Best results are printed in bold.

Input	Jetson TX2	TensorRT-FP32	TensorRT-FP16
32×32	12.3	16.9	25.7
64×64	8.8	18.5	25.6

TABLE I: VGG-1080p: Speed (fps)

Dataset	Only Hinge Loss	MI Regularizer
Football Player	0.9568 ± 0.0100	0.9744 ± 0.0100
Face	0.8841 ± 0.0040	0.8896 ± 0.0007
Bicycles	0.9684 ± 0.0037	0.9696 ± 0.0018
Crowd - Drone	0.9194 ± 0.0082	0.9303 ± 0.0076

TABLE II: Test Accuracy

C. Experimental Results

First, the evaluation results of the proposed models regarding the deployment speed are provided. The performance is tested on a low-power NVIDIA Jetson TX2 module with 8GB of memory, which is a state of the art GPU used for on-board UAV perception. Additionally, in order to accelerate the deployment speed and achieve real-time deployment, TensorRT¹ deep learning inference optimizer is utilized. TensorRT is a library that optimizes deep learning models providing FP32 (default) and FP16 optimizations for production deployments of various applications. In Table I we provide the detection speed in terms of fps for the two proposed models on the NVIDIA Jetson TX2 module without the utilization of the TensorRT optimizer, with the TensorRT on the default mode, and finally with TensorRT on the FP16 mode. As we can see TensorRT and in particular the FP16 mode significantly accelerates the proposed models, achieving detection in-real time for high-resolution images. To gain some intuition about the deployment speed, we note that state-of-the-art detectors run at notably fewer FPS on Jetson TX2, and also for lower resolution input images. For example, YOLO v.2 [21] runs at 3.1 fps for input of size 604×604 , while utilizing TensorRT (FP32) runs at 7.8 fps, and further speed up is achieved with the FP16 mode up to 14.4 fps, which remains far away from real-time even for lower input resolution. Finally, we should highlight that the deployment speed regards all the models, that is with and without the proposed regularizer, since the regularizer does not affect the deployment speed.

In Table II we present the mean value and the standard deviation of the test accuracy, for the considered training approaches, that is utilizing only hinge loss, and hinge loss with the proposed MI regularizer. Correspondingly, in Fig. 4 we illustrate the curves of mean test accuracy of the only hinge loss training against hinge loss & MI regularized training. We can see in the demonstrated results, that the proposed MI regularizer remarkably enhances the classification performance on all the utilized datasets.

Finally, we conducted a post-hoc Bonferroni test [22], for ranking the proposed regularization method and the only hinge loss training and evaluating the statistical significance of the obtained results. The performance of two methods is

¹<https://developer.nvidia.com/tensorrt>

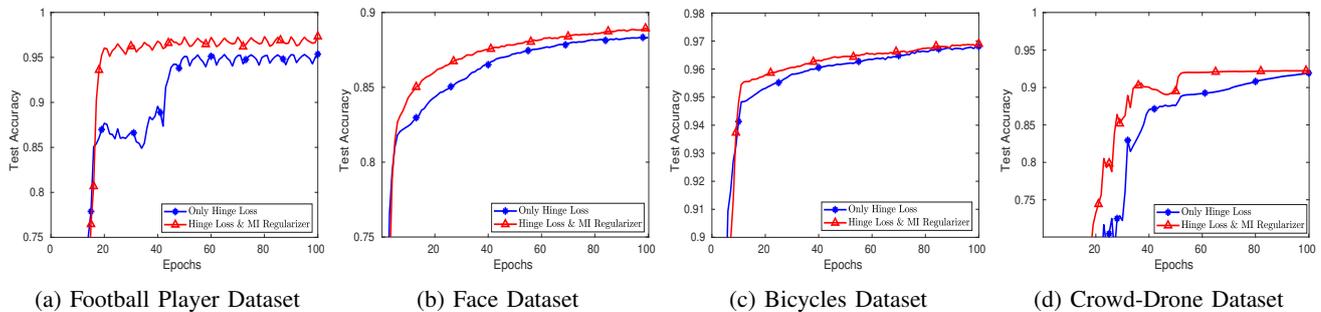


Fig. 4: MI Regularizer

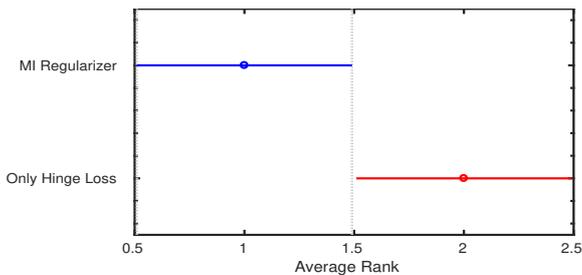


Fig. 5: Post-Hoc Bonferroni Test

significantly different, if the corresponding average ranks over the datasets differ by at least the critical difference:

$$CD = q_{\alpha} \sqrt{\frac{m(m+1)}{6D}}, \quad (14)$$

where m is the number of methods compared, D is the number of datasets and critical values q_{α} can be found in [22]. In our comparisons we set $\alpha = 0.05$. The number of datasets is four in the performed tests. The compared methods are two, that is the proposed regularizer is compared with a control method which is the only hinge loss training approach. The ranking results are illustrated in Fig. 5. The vertical axis depicts the two methods, while the horizontal axis depicts the performance ranking. The circles indicate the mean rank and the intervals around them indicate the confidence interval as this is determined by the CD value. Overlapping intervals between two methods indicate that there is not a statistically significant difference between the corresponding ranks, while non-overlapping intervals indicate that the compared methods are significantly different. As we can observe, the proposed regularizer is significantly different against the only hinge loss training approach.

V. CONCLUSIONS

In this paper, we proposed regularized lightweight CNN models for addressing various binary classification tasks able to run in real-time on-drone. Furthermore, we proposed a novel regularizer motivated by the QMI, the so-called MI regularizer. The performance was evaluated on four datasets. The evaluation results validate the effectiveness of the proposed regularizer in enhancing the generalization ability of the proposed models.

ACKNOWLEDGMENT

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 871449 (OpenDR). This publication reflects the authors’ views only. The European Commission is not responsible for any use that may be made of the information it contains.

REFERENCES

- [1] L. Deng, “A tutorial survey of architectures, algorithms, and applications for deep learning,” *APSIPA Transactions on Signal and Information Processing*, vol. 3, p. e2, 2014.
- [2] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai *et al.*, “Recent advances in convolutional neural networks,” *Pattern Recognition*, vol. 77, pp. 354–377, 2018.
- [3] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” *arXiv preprint arXiv:1704.04861*, 2017.
- [4] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, “Squeezenet: Alexnet-level accuracy with 50x fewer parameters and 0.5 mb model size,” *arXiv preprint arXiv:1602.07360*, 2016.
- [5] X. Zhang, X. Zhou, M. Lin, and J. Sun, “Shufflenet: An extremely efficient convolutional neural network for mobile devices,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6848–6856.
- [6] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [7] R. Caruana, “Multitask learning,” *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.
- [8] J. Weston, F. Ratle, and R. Collobert, “Deep learning via semi-supervised embedding,” in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 1168–1175.
- [9] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, “Deeply-supervised nets,” in *Artificial Intelligence and Statistics*, 2015, pp. 562–570.
- [10] K. Torkkola, “Feature extraction by non-parametric mutual information maximization,” *Journal of machine learning research*, vol. 3, no. Mar, pp. 1415–1438, 2003.
- [11] D. Bouzas, N. Arvanitopoulos, and A. Tefas, “Graph embedded non-parametric mutual information for supervised dimensionality reduction,” *IEEE transactions on neural networks and learning systems*, vol. 26, no. 5, pp. 951–963, 2015.
- [12] N. Passalis and A. Tefas, “Learning deep representations with probabilistic knowledge transfer,” in *The European Conference on Computer Vision (ECCV)*, September 2018.
- [13] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [14] C. E. Shannon, “A mathematical theory of communication,” *ACM SIGMOBILE mobile computing and communications review*, vol. 5, no. 1, pp. 3–55, 2001.

- [15] D. W. Scott, *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.
- [16] S.-T. Chiu, "Bandwidth selection for kernel density estimation," *The Annals of Statistics*, pp. 1883–1905, 1991.
- [17] M. Tzelepi and A. Tefas, "Graph embedded convolutional neural networks in human crowd detection for drone flight safety," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2019.
- [18] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2144–2151.
- [19] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *European Conference on Computer Vision*. Springer, 2014, pp. 94–108.
- [20] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5525–5533.
- [21] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," *arXiv preprint arXiv:1612.08242*, 2016.
- [22] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006.