Applications of large scale kernel machines for real world human mood estimation

Krasimir Tonchev and Nikolay Neshov
Technical University of Sofia
Teleinfrascruture R&D Lab
Kliment Ohridski 8, blvd
Bulgaria
{k_tonchev, nneshov}@tu-sofia.bg

Agata Manolova and Teodora Sechkova
Technical University of Sofia
Faculty of Telecommunications
Kliment Ohridski 8, blvd
Bulgaria
{amanolova, tsechkova}@tu-sofia.bg

Abstract: For many years the kernel methods were the primary tool for machine learning and computer vision. With their bad scalability for large dataset and the development of deep learning methods their usability decreased. In this work we show that the recent development of kernel approximation with random features can be used in real world applications. We build a mood estimation algorithm by utilizing multiple kernel learning approximated by random features. The algorithm is tested on popular large scale dataset and compared with state of the art methods.

Key–Words: large scale kernel machines, kernel approximation with random features, mood estimation

1 Introduction

Utilizing large datasets in the machine learning tasks was almost impossible until the recent advances in Deep Learning Neural Networks (DNN). They provide highly non-linear models and nonparametric models which model the data dependencies in a correct manner. The composition of multiple layers of nonlinear functions can approximate a rich set of naturally occurring input-output dependencies. Until this development, the kernel methods such as the kernel SVM were dominating the machine learning world. Being able to generalize difficult problems, they were the preferred tool. The main issue with kernel methods is that computational complexity of the kernel method is determined by the size of data, regardless of the dimension of the feature space.

In theory, kernel methods provide a feature mapping to an infinite dimensional space providing high capacity of the learner. But in practice kernel methods scale poorly with the size of the training dataset, and thus are perceived as being impractical for applications involving large data sets. The bottleneck in scaling up kernel methods is the storage and computation of the kernel matrix, K, which is usually dense. Storing the matrix requires $O(n^2)$ space, and computing it takes $O(n^{2d})$ operations, where n is the number of data points and d is the dimension of the data. The computational complexity of exact kernel methods depends quadratically on the number of training examples at training time and linearly at testing time. Hence, scaling up kernel methods has been a long-standing and actively studied problem.

Approximating kernels with finite-dimensional features has been recognized as a promising way of scaling up kernel methods. Such approximation using random feature maps has recently gained a lot of interest [1], [2], [3]. This is mainly due to their applications in reducing training and testing times of kernel based learning algorithms. The seminal work of [4] provides theoretically salient approach towards approximating kernels using random realizations.

One of the hot topics in computer vision is the estimation of human emotions from still images or video. It is a challenging problem with many difficulties imposed by head pose, light conditions and image quality [5]. Most of the available databases consisting human expressions are recorded using actors who act the different expressions with images captured under controlled conditions [6]. Recently, to satisfy the needs of deep learning methods, a large datasets and high data variability appeared in the research communities [7]. These dataset were collected using images and videos captured in the real world conditions. Utilizing this dataset, the authors in [8] used Recurrent Neural Networks (RNN) which are composed of rectified linear units (ReLUs) where a special initialization strategy based on scaled variations of the identity matrix is applied. The RNNs are capturing the spatio-temporal relations of the facial expressions. They are trained using 5 fiducial facial points detected by convolutional neural network (CNN). The CNN consists of three layers of 8x8 filters. They achieved 52.8% accuracy on the test+validation dataset. A mapped Local Binary Patterns (LBP) for feature extraction is proposed in [9]. The mapping of LBP is done using Multi Dimensional Scaling. The classification task was carried out by ensemble of CNNs. They achieved 51.7 % and 54.5% accuracy on the validation and test data sets respectively.

This work is based on the approach presented in [10]. We exploit the parallelism in learning presented by the authors to implement a mood estimation algorithm. The work is organized as follows: in 2 we present the approach to learn multiple kernels by random features. In 3 we present the proposed algorithm for real world mood estimation and in 4 we present the experimental results.

2 Kernel representation by random features

Kernel methods avoid inference in $\mathbb{R}^M$ and rely on the kernel matrix over the training samples, where M is the dimensionality of embedding space and is usually much greater than the dimensionality of the data space, even infinite dimensional. The main drawback of kernel methods is dealing with large data sets. Then the kernel matrix becomes computationally intractable. Utilizing a classical result from harmonic analysis, Rachimi and Recht [4] found out a way to approximate kernels using random features. The kernels that can be approximated are translation invariant and includes the most popular ones: the Gaussian $k_G(\mathbf{x}, \mathbf{y}) = e^{\|\mathbf{x}-\mathbf{y}\|^2/2\sigma^2}$ and Laplacian $k_L(\mathbf{x}, \mathbf{y}) = e^{|\mathbf{x}-\mathbf{y}|_1/\sigma}$ kernels. The result implies that the kernel function can be expanded by harmonic basis:

$$k(\mathbf{x}-\mathbf{y}) = \int_{\mathbb{R}^d} e^{j\omega^T(\mathbf{x}-\mathbf{y})} dp(\omega) \tag{1}$$

which is the expectation of over the probability measure $p(\omega)$:

$$k(\mathbf{x}-\mathbf{y}) = \mathcal{E}_\omega[e^{j\omega^T\mathbf{x}} e^{-j\omega^T\mathbf{y}}] \tag{2}$$

It is complex valued, but can be easily simplified for real valued functions using sine and cosine functions. The measure for the Gaussian and Laplacian kernels can be analytically calculated and leveraged for sample-based approximation of the expectation. Assuming that D samples are drawn from the corresponding measure $\{\omega_1, \omega_2, ..., \omega_D\}$, the sample mean can be used to approximate the expectation:

$$\begin{aligned} k(\mathbf{x}-\mathbf{y}) &\approx \frac{1}{D}\sum_{i=1}^{D} \hat{\phi}_{\omega_i}(\mathbf{x})\hat{\phi}_{\omega_i}(\mathbf{y}) \\ &= \phi(\mathbf{x})^T\phi(\mathbf{y}) \end{aligned} \tag{3}$$

The maps $\phi(\cdot)$ are random feature vectors indexed by the samples ω_i drawn form the measure probability measure $p(\omega)$. Usually one of the priors included in the data model is the kernel type selection as well as its parameters. The parameters can be selected based on sampling over pre-specified range while, but the selection of the kernel cannot be done in such way. Multiple Kernel Learning (MKL) is one the adequate approaches in this situation [11]. It can be proved that there is also random features representation for MKL with translation invariant kernels. Based on the work in [10] it can be proved that there is convenient random features representation multiplicative kernel combination from base kernels:

$$k(\cdot,\cdot) = \prod_{i=1}^{S} k(\cdot,\cdot) \tag{4}$$

It can be shown that given a set of translation invariant kernels $k_i, i = 1, ..., S$, with corresponding probability measures $p_i(\omega)$, the probability measure for the resultant multiplicative kernel is the convolution of all probability measures [10]:

$$p(\omega) = p_1(\omega) * p_2(\omega) * ... * p_S(\omega) \tag{5}$$

The authors also proved that:

$$\omega = \sum_{i=1}^{S} \omega_i \tag{6}$$

where $\omega_i \propto p_i(\omega)$. This result decreases the dimensionality of the problem, allowing the number of approximating random features to be independent from the number of kernels.

Another important result from [10] is parallel training by splitting the random sample into subsets and performing the training over each subset in parallel. If the random sample is of size D then it is split into B sub-sets each one of size D_0. Further, B multinomial regressions are trained in parallel delivering B sets of parameters $\alpha_c^b, c = 1, ..., C, b = 1, ..., B$. The conditional probability for particular class c is then calculated by the geometric mean of the probabilities for each sub-set b:

$$\begin{aligned} p(z = c|x) &= \sqrt{\prod_b \exp\left(\phi_b(\mathbf{x})^T \alpha_c^b\right)} \\ &= \exp\left(\frac{1}{B}\sum_{i=1}^{B} \phi_b(\mathbf{x})^T \alpha_c^b\right) \end{aligned} \tag{7}$$

The multinomial regression with basis functions approximated by random features are named *random kitchen sinks* and are proposed in [12]. The primary advantage of this method is the much faster training time while the drawback is the testing time since it requires more features compared to AdaBoost [13].

3 Mood recognition with large scale kernels

The mood recognition in this work is composed of multiple stages of pre-processing and a final stage of classification. The mood is recognized using the seven basic expressions: *Anger, Disgust, Fear, Happiness, Neutral, Sadness* and *Surprise*. Each expression is recognized from single still image of human face. The following steps compose the algorithm for expression recognition:

- Face detection: face detector is applied on image in order to extract the facial area from the background. The facial detector used in this work is the Viola-Jones face detector implemented in OpenCV package;

- Fiducial points detection: after the face image is extracted, an algorithm for fiducial points detection is applied in order to detect the points of interest. The Supervised Descent Method (SDM) proposed by [14] is utilized in this work;

- Multinomial logistic regression: classifier using MKL and multinomial logistic regression, described in the previous section, is used to carry out the classification process.

Learning the parameters of the Multinomial logistic regression is done using the Stochastic Averaged Gradient (SAG) descent. This method is performing better for optimization tasks where the objective function is convex. In this case the objective function is logarithm which is convex and SAG can be utilized.

4 Experimental results

4.1 Dataset

In this work, a dataset form EmotiW 2015 challenge [7] is used for testing and validating the proposed mood recognition algorithm. The dataset consists of set of 1590 images of human faces with background. The images are extracted from videos using key-frame extraction method. This method consists is following: facial points are extracted form each frame and clusters are formed. The frames closer to the cluster centers are selected as key-frames containing the related facial expression. The clustering procedure utilized is K-means clusters with 6 clusters. The dataset is further divided into three sets: Train (880 samples), Validation (383 samples) and Test (372 samples). Example images form the dataset are depicted on Figiure 1.

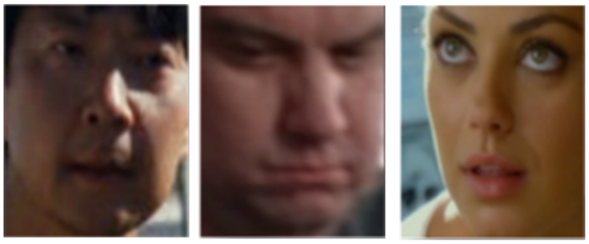

Figure 1: Example images form SFEW 2.0 dataset.

This dataset is named SFEW 2.0 and used for **SReco** challenge [7]. According the protocol, each algorithm have to be validated on the Test data set while the Validation is used for reference. Further, results can be provided for Test+Validation sets.

4.2 Experimental setup and results

The kernels used in the setup are Gaussian and Laplacian. Both kernels have preset bandwidth of $\{0.1, 0.5, 1, 2, 4\}$. This results in MKL combination of 10 kernels, 5 Gaussian and 5 Laplacian. Presetting the bandwidth of the kernels avoids the sampling based search procedure for the most optimal value. The number of random features D was experimentally selected to be 1,000,000. Note that this number is the resultant number after sumation of all random features drawn from the corresponding kernel probability measures as formulated in (6).

After extracting the facial fiducial points they were input to the multinomial logistic regression . It was trained by spliting the random features into 40 blocks each one consisting ot 25000 features. Each block was trained in parallel, as described in the previous section, achieving faster performance. The training was done using the SAG and only the data of the Train subset. The results are presented in Table 1.

Table 1: Accuracies [%]

Method	Validation	Test
Method in [8]	68.4%	47.6%
Method in [9]	51.74%	54.56%
Proposed Method	64.4%	53.6%

It is important to note that the proposed method delivers comparable results with the state of the art CNNs based methods, even it can be considered as ”shallow” network of kernels. It is also performs 14% better than the baseline method of the EmotiW challenge for the SReco sub-challenge.

5 Conclusion

The current work shows that multiplicative combination of multiple kernels approximated by random features perform comparably to state of the art methods for expression recognition, while keeping low computational needs. Split of the total random features allows for parallelization and full utilization of the modern parallel computing platforms. Future work includes detailed analysis on the computational performance, increased range of kernel parameters and utilization of other types of translation invariant kernels.

Acknowledgements: This work was supported by contract DFNI I02/1 for research project: Intelligent man-machine interface for assistive medical systems in improving the independent living of motor disabled users of the Bulgarian Research Fund of the Ministry of Education and Science.

References:

[1] Q. Le, T. Sarlós, and A. Smola, "Fastfood-approximating kernel expansions in loglinear time," in *Proceedings of the international conference on machine learning*, vol. 85, 2013.

[2] P. Kar and H. Karnick, "Random feature maps for dot product kernels," in *International conference on artificial intelligence and statistics*, 2012, pp. 583–591.

[3] B. Dai, B. Xie, N. He, Y. Liang, A. Raj, M.-F. F. Balcan, and L. Song, "Scalable kernel methods via doubly stochastic gradients," in *Advances in Neural Information Processing Systems*, 2014, pp. 3041–3049.

[4] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in neural information processing systems*, 2008, pp. 1177–1184.

[5] C. A. Corneanu, M. O. Simon, J. F. Cohn, and S. E. Guerrero, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1548–1568, 2016.

[6] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 94–101.

[7] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and image based emotion recognition challenges in the wild: Emotiw 2015," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 423–426.

[8] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 467–474.

[9] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 503–510.

[10] Z. Lu, A. May, K. Liu, A. B. Garakani, D. Guo, A. Bellet, L. Fan, M. Collins, B. Kingsbury, M. Picheny *et al.*, "How to scale up kernel methods to be as good as deep neural nets," *arXiv preprint arXiv:1411.4000*, 2014.

[11] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf, "Large scale multiple kernel learning," *Journal of Machine Learning Research*, vol. 7, no. Jul, pp. 1531–1565, 2006.

[12] A. Rahimi and B. Recht, "Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning," in *Advances in neural information processing systems*, 2009, pp. 1313–1320.

[13] R. E. Schapire, "The boosting approach to machine learning: An overview," in *Nonlinear estimation and classification*. Springer, 2003, pp. 149–171.

[14] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 532–539.