Localized Multiple Kernel Machines for Image Recognition

Mehmet Gönen gonen@boun.edu.tr
Ethem Alpaydın alpaydin@boun.edu.tr
Department of Computer Engineering, Boğaziçi University, İstanbul, Turkey

Mathematical Model

- Multiple Kernel Learning (MKL) [1]

$$f(\boldsymbol{x}) = \sum_{m=1}^{P} \langle \boldsymbol{w}_m, \Phi_m(\boldsymbol{x}) \rangle + b = \sum_{i=1}^{N} \alpha_i y_i \sum_{m=1}^{P} \eta_m \underbrace{\langle \Phi_m(\boldsymbol{x}_i), \Phi_m(\boldsymbol{x}) \rangle}_{k_m(\boldsymbol{x}_i, \boldsymbol{x})} + b$$

- Localized Multiple Kernel Learning (LMKL) [2]

$$f(\boldsymbol{x}) = \sum_{m=1}^{P} \eta_m(\boldsymbol{x}|\mathbf{V}) \langle \boldsymbol{w}_m, \Phi_m(\boldsymbol{x}) \rangle + b = \sum_{i=1}^{N} \alpha_i y_i \underbrace{\sum_{m=1}^{P} \eta_m(\boldsymbol{x}_i|\mathbf{V}) k_m(\boldsymbol{x}_i, \boldsymbol{x}) \eta_m(\boldsymbol{x}|\mathbf{V})}_{k_\eta(\boldsymbol{x}_i, \boldsymbol{x})} + b$$

Primal Problem

$$\begin{aligned} \text{min.} \quad & \frac{1}{2} \sum_{m=1}^{P} \|\boldsymbol{w}_m\|^2 + C \sum_{i=1}^{N} \xi_i \\ \text{w.r.t.} \quad & \boldsymbol{w}_m \in \mathbb{R}^{D_m}, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}_+^N, \mathbf{V} \in \mathbb{R}^{D_{\mathcal{G}}} \\ \text{s.t.} \quad & y_i \left(\sum_{m=1}^{P} \eta_m(\boldsymbol{x}_i|\mathbf{V}) \langle \boldsymbol{w}_m, \Phi_m(\boldsymbol{x}_i) \rangle + b \right) \geq 1 - \xi_i \qquad \forall i \end{aligned}$$

- For a given $\mathbf{V}$

Dual Problem

$$\begin{aligned} \text{max.} \quad & J(\mathbf{V}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j k_\eta(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ \text{w.r.t.} \quad & \boldsymbol{\alpha} \in \mathbb{R}_+^N \\ \text{s.t.} \quad & \sum_{i=1}^{N} \alpha_i y_i = 0 \\ & C \geq \alpha_i \geq 0 \qquad \forall i \end{aligned}$$

- $\dfrac{\partial J(\mathbf{V})}{\partial \mathbf{V}}$ is used to update $\mathbf{V}$
- Two-step alternate optimization algorithm for learning $\boldsymbol{\alpha}$ and $\mathbf{V}$

Combining Multiple Resolutions

Softmax Gating

$$\eta_m(\boldsymbol{x}|\mathbf{V}) = \frac{\exp(\langle \boldsymbol{v}_m, \Phi_{\mathcal{G}}(\boldsymbol{x}) \rangle + v_{m0})}{\sum_{k=1}^{P} \exp(\langle \boldsymbol{v}_k, \Phi_{\mathcal{G}}(\boldsymbol{x}) \rangle + v_{k0})} \qquad \forall m$$

- Face image in different resolutions

4×4 8×8 16×16 32×32 64×64

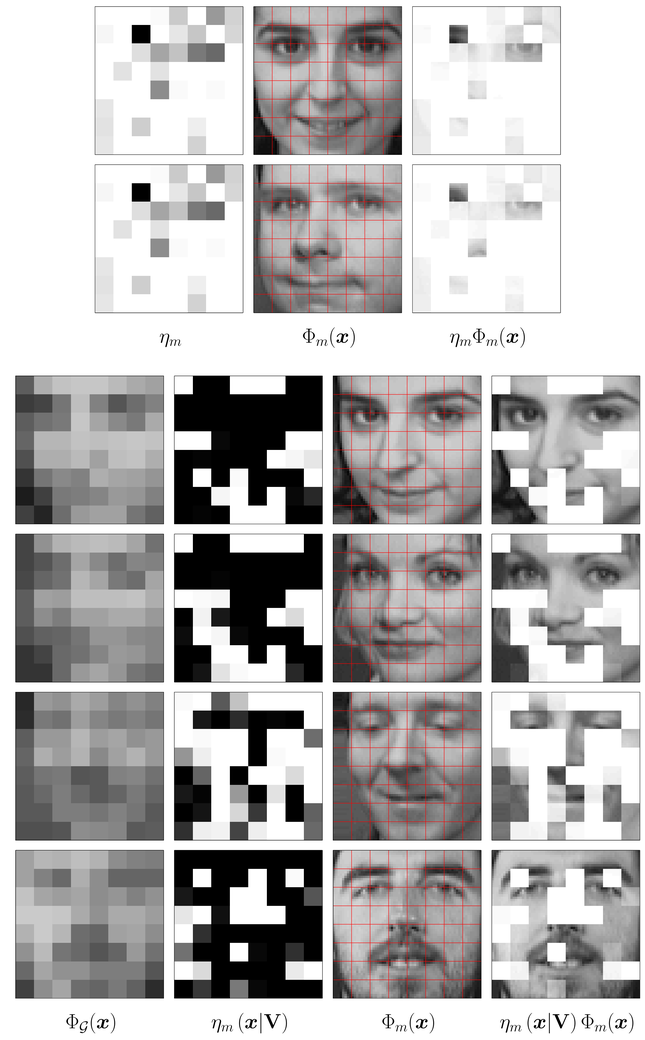
	Method	Configuration	Accuracy	SV
Single Kernel	SVM	$\Phi(\boldsymbol{x}) = 4 \times 4$	93.28 ± 0.65	21.70 ± 0.93
	SVM	$\Phi(\boldsymbol{x}) = 8 \times 8$	97.50 ± 1.16	20.13 ± 1.04
	SVM	$\Phi(\boldsymbol{x}) = 16 \times 16$	97.03 ± 0.93	19.82 ± 0.94
	SVM	$\Phi(\boldsymbol{x}) = 32 \times 32$	97.97 ± 1.48	23.71 ± 1.39
	SVM	$\Phi(\boldsymbol{x}) = 64 \times 64$	97.66 ± 1.41	25.94 ± 1.01
Multiple Kernel	MKL	uses only 64×64	97.66 ± 1.41	25.94 ± 1.01
	LMKL	$\Phi_{\mathcal{G}}(\boldsymbol{x}) = 4 \times 4$	97.03 ± 1.15	29.29 ± 2.90
	LMKL	$\Phi_{\mathcal{G}}(\boldsymbol{x}) = 8 \times 8$	99.38 ± 0.94	27.68 ± 2.95
	LMKL	$\Phi_{\mathcal{G}}(\boldsymbol{x}) = 16 \times 16$	98.59 ± 1.41	26.52 ± 2.37
	LMKL	$\Phi_{\mathcal{G}}(\boldsymbol{x}) = 32 \times 32$	99.38 ± 1.16	24.78 ± 2.57
	LMKL	$\Phi_{\mathcal{G}}(\boldsymbol{x}) = 64 \times 64$	99.53 ± 0.65	26.65 ± 4.12

Combining Multiple Input Patches

Sigmoid Gating

$$\eta_m(\boldsymbol{x}|\mathbf{V}) = \frac{1}{1 + \exp(-\langle \boldsymbol{v}_m, \Phi_{\mathcal{G}}(\boldsymbol{x}) \rangle - v_{m0})} \qquad \forall m$$

Method	Configuration		Accuracy	SV
MKL	$\Phi_m(\boldsymbol{x}) = 8 \times 8$		99.38 ± 0.94	19.42 ± 0.87
LMKL	$\Phi_m(\boldsymbol{x}) = 8 \times 8$	$\Phi_{\mathcal{G}}(\boldsymbol{x}) = 8 \times 8$	99.84 ± 0.44	24.38 ± 1.96
MKL	$\Phi_m(\boldsymbol{x}) = 16 \times 16$		99.06 ± 0.88	22.19 ± 1.00
LMKL	$\Phi_m(\boldsymbol{x}) = 16 \times 16$	$\Phi_{\mathcal{G}}(\boldsymbol{x}) = 4 \times 4$	99.53 ± 0.65	23.35 ± 1.47



η_m $\Phi_m(\boldsymbol{x})$ $\eta_m \Phi_m(\boldsymbol{x})$

$\Phi_{\mathcal{G}}(\boldsymbol{x})$ $\eta_m(\boldsymbol{x}|\mathbf{V})$ $\Phi_m(\boldsymbol{x})$ $\eta_m(\boldsymbol{x}|\mathbf{V}) \Phi_m(\boldsymbol{x})$

Conclusions

- Gating model acts as a saliency detector for selective attention
- Use a cheap gating model that selects among kernels of different costs only when needed

References

[1] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan.
Multiple kernel learning, conic duality, and the SMO algorithm.
In *Proceedings of the 21st International Conference on Machine Learning*, pages 41–48, 2004.

[2] M. Gönen and E. Alpaydın.
Localized multiple kernel learning.
In *Proceedings of the 25th International Conference on Machine Learning*, pages 352–359, 2008.

NIPS 2009 Workshop on Understanding Multiple Kernel Learning Methods, December 11, 2009, Whistler, British Columbia, Canada.