Object of Interest Detection by Saliency Learning

Advisor

Dr. Maryam Abedi

Student

Mohammad Shahpouri

August

2022

Contents

List of Figures

List of Equations

1 Structured Learning

The structured learning approach hinges in the notion that non-linear classification can be effected in a piecewise-linear manner across the feature space.

1.1 Mixture of SVMs

The aim is to combine the saliency features so as to perform classification, i.e. separate salient objects from the background in the image, based upon objects of interest provided as training data.

To commence, consider a set of M tuples $(X, Y) = \{\mathrm{x}_{i,l}, y_i)|i = 1, \ldots, M, y_i \in \{-1, 1\}\}$, where $\mathrm{x}_{i,l}, y_i)$ are the i^{th} data-label pair in the training data corresponding to the l^{th} saliency feature, where the total number of salient feature N. In practice, Y ccounts for the corresponding object of interest regions provided at input. The linear SVM classifier solves the following optimisation problem

$$\min_{\mathrm{w}} \frac{\|\mathrm{w}\|^2}{2} + C \sum_i \epsilon(\mathrm{w}; \mathrm{x}_{i,l}, y_i) \tag{1}$$

Where $\epsilon(\mathrm{w}; \mathrm{x}_{i,l}, y_i) = \max(1 - y_i \mathrm{w}^T \mathrm{x}_{i,l}, 0)$ is the Hinge loss function which specifies an upper bound on the classification error. C is regularisation term on classifier weights. They have subsumed the bias term b in the above formulation by appending each data instance with an additional dimension $\mathrm{x}_{i,l}^T = [\mathrm{x}_{i,l}^T, 1]$ and $\mathrm{w}^T = [\mathrm{w}^T, b]$.

They extend the SVM model to a two-layer mixture model. Using the joint probability distribution and the SVM binary classifier.

They establish the link between the two layers using the joint probabilistic distribution over X and Y given by

$$P(Y|X, \Theta) = \prod_i P(y_i|\mathrm{x}_{i,l}, \Theta) = \prod_i \sum_{z_i} P(y_i|z_i, \mathrm{x}_{i,l}, \Theta) P(\mathrm{x}_{i,l}|z_i, \Theta) P(z_i|\Theta) \tag{2}$$

Where i indexes data samples as before, $\Theta = \{\alpha, \beta, \tau, \gamma\}$ are the parameters of the underlying model and z_i is the hidden variable introduced for the ith sample for each of the N salient features under study. α and β are the multinomial parameters that generate the hidden variables z_i's whereas τ and γ are parameters for the gating nodes and classifiers. The probability $P(\mathrm{x}_{i,l}|z_i, \tau)$

represents the posterior for the mixture component with hyperparameter τ , and $P(y_i|\mathrm{x}_{i,l}, \gamma)$ is the posterior probability of corresponding linear SVM output for the ith sample. z_i is the hidden variable generated from a multinominal distribution with parameters $\alpha = \{\alpha_1, \dots, \alpha_k\}$ and $\beta = \{\beta_1, \dots, \beta_k\}$ for K-mixtures and N features. $\mathrm{x}_{i,l}$ is generated from an isotropic Gaussian distribution with parameter τ conditional on z_i , where $\tau = \{(\mu_{1,1}, \mathit{\Sigma}_{1,1}), \dots, (\mu_{K,N}, \mathit{\Sigma}_{K,N})\}$ $\mu_{j,l}$ and $\Sigma_{j,l}$ are the mean vector and the variance for the jth mixture component performing inference upon the saliency feature-set indexed l. y_i is generated from a probabilistic classifier model with parameter γ conditional on $textx_{i,l}$ and z_i, where $\gamma = \{\mathrm{w}_{1,1}, \dots, \mathrm{w}_{K,N}\}$, and $\mathrm{w}_{j,l}$ is the classifier weight-vector for the jth linear SVM corresponding to the lth saliency feature-set. This yields

$$P(Y|X, \mathit{\Theta}) = \prod_i \sum_{z_i} P(y_i|\mathrm{x}_{i,l}, \gamma) P(\mathrm{x}_{i,l}|z_i, \tau) P(z_i|\alpha, \beta) \tag{3}$$

Parameter estimation can be done via Maximum Likelihood Estimation (MLE) by maximizing the following log-likelihood function

$$\mathcal{L}(\mathit{\Theta}) = \sum_i \log P(y_i|\mathrm{x}_{i,l}, \mathit{\Theta}) + \sum_j \mathit{\Omega}(\mathrm{w}_{j,l}) = \sum_i \log\{\sum_l \beta_l \sum_j \alpha P(y_i|\mathrm{x}_{i,l}, \mathrm{w}_{j,l}) P(\mathrm{x}_{i,l}|z_i, \tau)\} + \sum_j \mathit{\Omega}(\mathrm{w}_{j,l}) \tag{4}$$

Where $\mathit{\Omega}(\mathrm{w}_{j,l}) = \log\{P(\mathrm{w}_{j,l})\}$ is a log-prior term for regularisation purposes. $\gamma = \{\mathrm{w}_{1,1}, \dots, \mathrm{w}_{K,N}\}$ and $P(z_i|\alpha, \beta) = \alpha_j \beta_l$ for the j^{th} mixture and the l^{th} salient feature-set. Second term on the Equation 4 is related to the prior $\mathit{\Omega}(\mathrm{w})$, whereas the first term corresponds to the conditional probability $P(y|\mathrm{x}, \mathrm{w})$ related to classification errors. These are given by

$$\mathit{\Omega}(\mathrm{w}_{j,l}) = -\zeta \left\|\mathrm{w}_{j,l}\right\|^2 \tag{5}$$

$$P(y_i|\mathrm{x}_{i,l}, \mathrm{w}_{j,l}) = e^{-\epsilon(\mathrm{w}_{j,l}; \mathrm{x}_{i,l}, y_i)} \tag{6}$$

1.2 The EM Algorithm

In this section, they describe an EM algorithm for solving the mixture of linear SVMs.

$$q_{i,j,l}^{(t+1)} = \frac{\alpha_j^t \beta_l^t P(\mathrm{x}_{i,l}|\mu_{j,l}^{(t)}, \sum_{j,l}^{(t)}) P(y_i|\mathrm{x}_{i,l}, \mathrm{w}_{j,i}^{(t)})}{\sum_s \sum_u \sum_v \alpha_u^{(t)} \beta_v^t P(\mathrm{x}_{s,v}|\mu_{u,v}^{(t)}, \sum_{u,v}^{(t)}) P(y_s|\mathrm{x}_{s,v}, \mathrm{w}_u^{(t)})} \tag{7}$$

where $s \in \{1, \ldots, M\}$, $u \in \{1, \ldots, K\}$, $v \in \{1, \ldots, N\}$. $P(y_i|\mathrm{x}_{i,l}, \mathrm{w}_{j,l}^{(t)})$ is generated by Equation 6 and, $P(\mathrm{x}_{i,l}|\mu_{j,l}^{(t)}, \sum_{j,l}^{(t)})$ is given by the following multivariate, d-dimensional Gaussian distribution,

$$P(\mathrm{x}_{i,l}|\mu_{j,l}^{(t)}, \sum_{j,l}^{(t)}) = \frac{1}{\sqrt{(2\pi)^d \left|\sum_{j,l}^{(t)}\right|}} \exp\left(-\frac{1}{2}(\mathrm{x}_{i,l} - \mu_{j,l}^{(t)})^T (\sum_{j,l}^{(t)})^{-1} (\mathrm{x}_{i,l} - \mu_{j,l}^{(t)})\right) \tag{8}$$

They update the parameters simultaneously with M-step for the gating nodes and the SVM classifiers to solve two independent optimization problems.

$$\alpha_j^{(t+1)} = \frac{\sum_s \sum_v q_{s,j,v}^{t+1}}{\sum_s \sum_u \sum_v q_{s,u,v}^{t+1}} \tag{9}$$

$$\beta_l^{(t+1)} = \frac{\sum_s \sum_u q_{s,u,l}^{(t+1)}}{\sum_s \sum_u \sum_v q_{s,u,v}^{(t+1)}} \tag{10}$$

$$\mu_{j,l}^{(t+1)} = \frac{\sum_s q_{s,j,l}^{(t+1)} \mathrm{x}_{s,l}}{\sum_s q_{s,j,l}^{(t+1)}} \tag{11}$$

$$\sum_{j,l}^{(t+1)} = \frac{\sum_s q_{s,j,l}^{(t+1)} (\mathrm{x}_{s,l} - \mu_{j,l}^{(t+1)})^T (\mathrm{x}_{s,l} - \mu_{j,l}^{(t+1)})}{\sum_s q_{s,j,l}^{(t+1)}} \tag{12}$$

They solve the following classification problem to update the weights for the j^{th} linear classifier working on the l^{th} saliency feature

$$\max \sum_i \sum_j q_{i,j,l}^{(t)} \log P(y_i|\mathrm{x}_{i,l}, \theta_{j,l}) + \log P(\theta_{j,l}) = \max \left\{ -\sum_i \sum_l q_{i,j,l}^{(t)} \epsilon(\mathrm{w}_{j,l}; \mathrm{x}_{i,l}, y_i) - \zeta \left\|\mathrm{w}_{j,l}\right\|^2 \right\} \tag{13}$$

where $\theta_{j,l} = \{\alpha_j, \beta_l, \mu_{j,l}, \mathit{\Sigma}_{j,l}, \mathrm{w}_{j,l}\}$ and $C = \frac{1}{2\zeta}$. This is exactly the same problem as training linear SVMs in Equation 1 whose sample weights are given by $q_{i,j,l}^{(t)}$.

1.3 Convergence

The method consists of the following steps at each iteration t

- Train the SVMs using the sample weights $q^t_{i,j,l}$ to recover the probabilities $P(y_i|\mathrm{x}_{i,l}, \mathrm{w}^{(t)}_{j,l})$.

- Compute the updated weights $q^{t+1}_{i,j,l}$ in Equation 7.

- Recover the remaining parameters making use of Equations 9-12.

2 Feature Extraction

They extend feature map extraction methods by Itti et al. [1] and Liu et al. [2] by considering the pixel neighbourhood, which permits capturing the image structure during the feature extraction process. The individual features are then used as the input to our structured learning method. They make scale pyramid using Gaussian filters. Three types of visual cues are generated from features that are extracted from each scale. The first is the intensity feature. The second set of features is represented by a set of color opponency between red, green and blue channel values against the yellow basis. The third set is generated by a set of even-symmetric Gabor filters [3]. Their extension is adopting a second-order Markov setting, that is, including the saliency features of the pixels in a 3 × 3 neighborhood.

3 Experiments

They perform experiments on the Microsoft Research Asia (MSRA) Salient Object Database B, which contains 5,000 images. Details on this database can be found in [2].

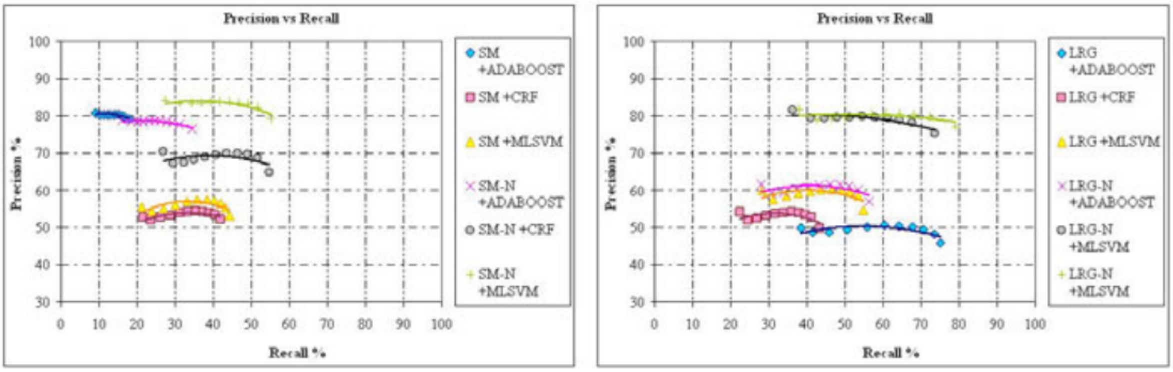

Figure 1: Average precision-recall

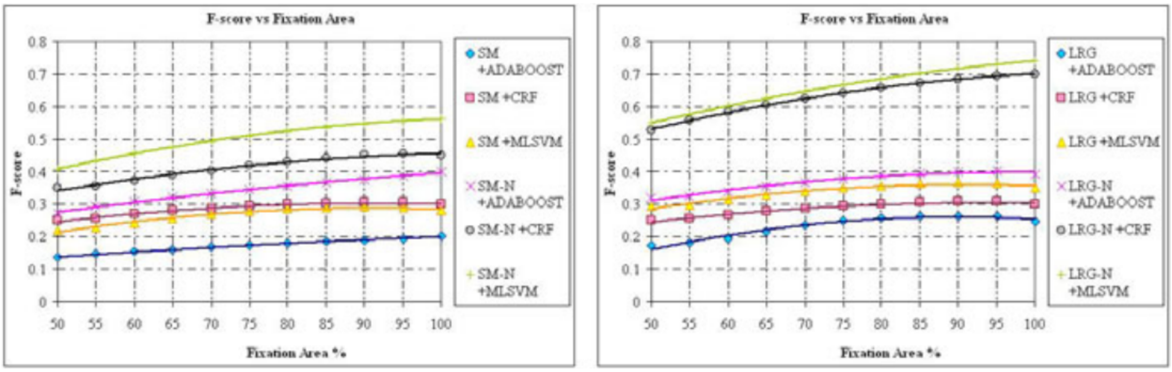

Figure 2: Average F-score as a function of the fixation area percentage

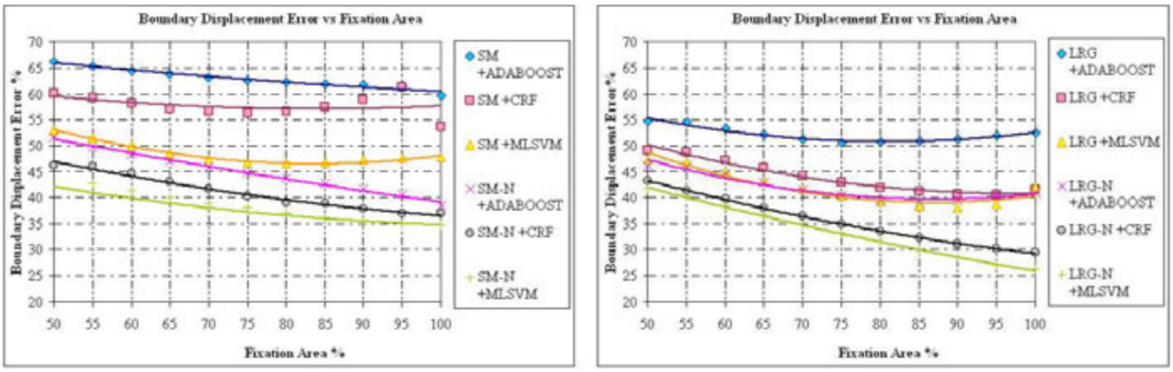

Figure 3: Boundary Displacement Error as a function of the fixation area percentage

References

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998. 4

[2] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007. 4

[3] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *J. Opt. Soc. Am. A*, vol. 2, pp. 1160–1169, Jul 1985. 4