Bilinear Lithography Hotspot Detection

Hang Zhang, Fengyuan Zhu, Haocheng Li,
Evangeline F.Y. Young, Bei Yu

The Chinese University of Hong Kong

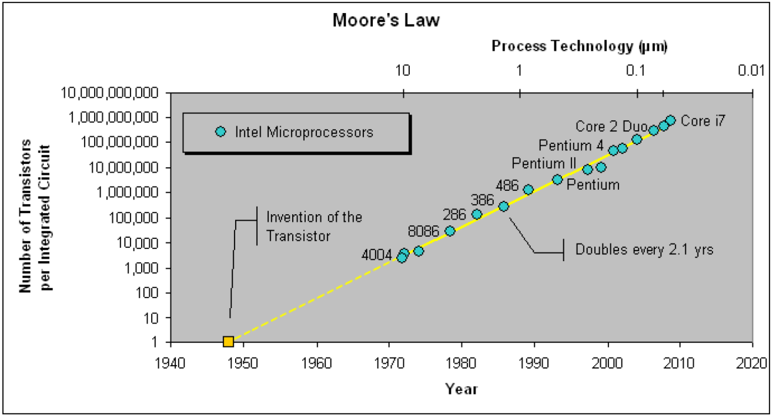
March 20, 2017

香港中文大學
The Chinese University of Hong Kong

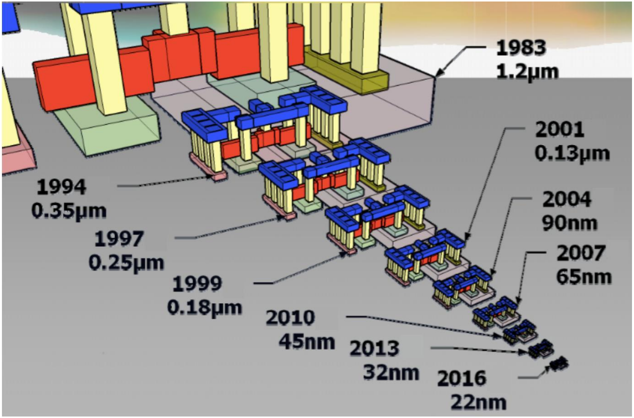
Outline

Outline

Moore's Law to Extreme Scaling

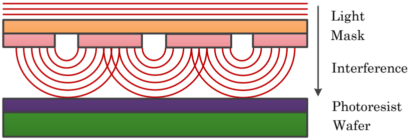


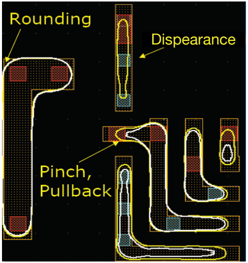
Shrinking Device Feature Size



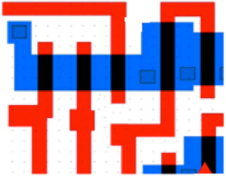
Lithographic Mechanism

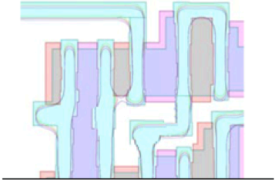
- Light pass through photo masks (mask scale << light wavelength);
- Light diffraction and light interference will happen;
- May cause performance degradation, or even **yield loss**. .

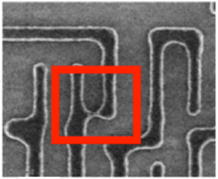




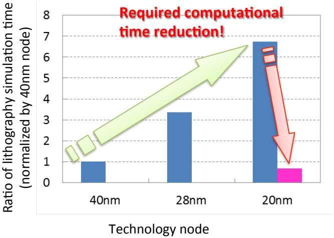
Motivation



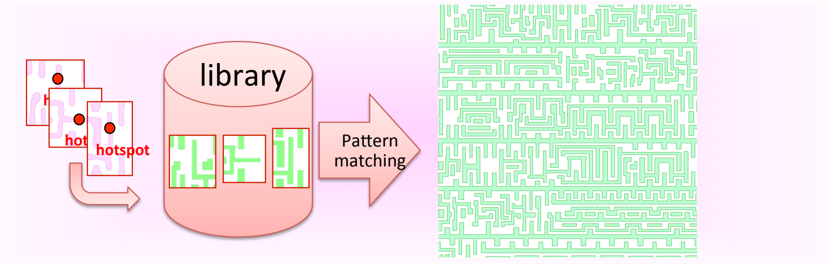




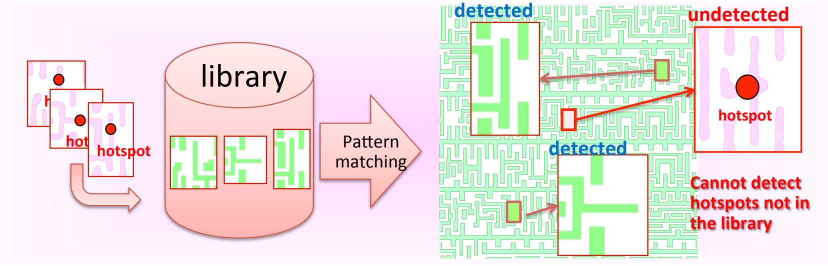
- What you design $\neq$ what you get;
- DFM: MPL, OPC, SRAF;
- Still hotspot: low fidelity patterns;
- Simulations: extremely time intensive.



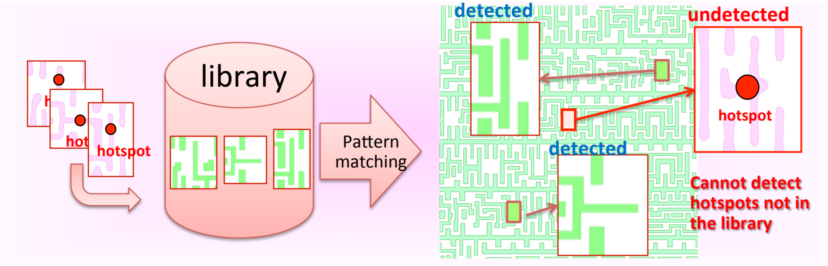
Pattern Matching based Hotspot Detection



Pattern Matching based Hotspot Detection

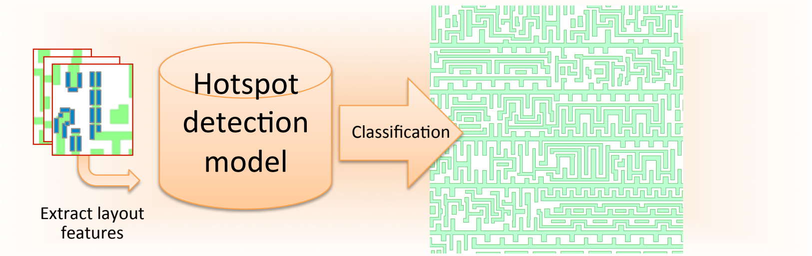


Pattern Matching based Hotspot Detection

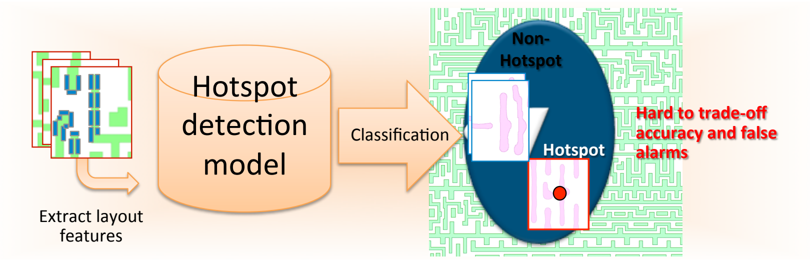


- Fast and reasonably accurate;
- Two-stage filtering, fuzzy pattern matching;
- [Yu+,ICCAD'14][Wen+,TCAD'14];
- Hard to detect unseen pattern.

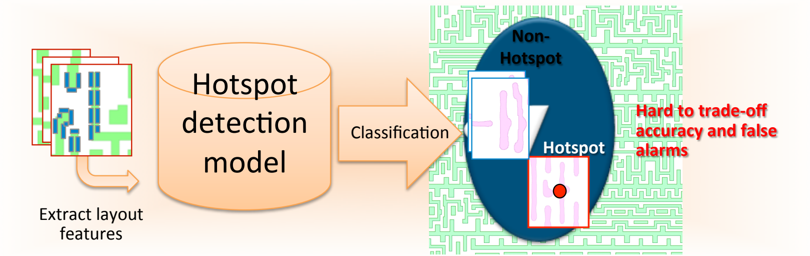
Machine Learning based Hotspot Detection



Machine Learning based Hotspot Detection



Machine Learning based Hotspot Detection



- Can predict new patterns, and are more flexible;
- Support vector machine, boosting, deep neural network...
- [Ding+,ASPDAC'12][Yu+,TCAD'15][Zhang+,ICCAD'16] [Matsunawa+,SPIE'16];
- Hard to balance accuracy and false-alarm.

Rethinking Conventional Methods

- Conventional: vector based feature and learning model;
- Time consuming steps: 1) feature extraction, 2) feature selection;
- Destroying the hidden structural correlations in the layout patterns.

Rethinking Conventional Methods

- ~~Conventional: vector based feature;~~
- ~~Time consuming steps: 1) feature extraction, 2) feature selection;~~
- ~~Destroying the hidden structural correlations in the layout patterns.~~

Rethinking Conventional Methods

- ~~Conventional: vector based feature;~~
- ~~Time consuming steps: 1) feature extraction, 2) feature selection;~~
- ~~Destroying the hidden structural correlations in the layout patterns.~~

Matrix based Concentric Sampling (MCCS)

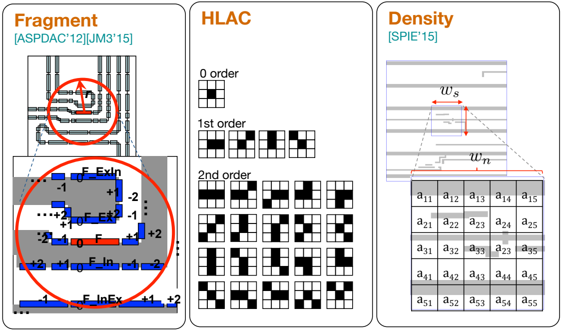
1) Matrix Based: preserve the hidden structural correlations;
2) No feature selection: enable parallel computation;
3) Very simple feature: fast to extract.

Bilinear Lithography Hotspot Detector

1) Matrix based: capture the hidden structural correlations;
2) Low-complexity model: avoid over-fitting;
3) Fast to train.

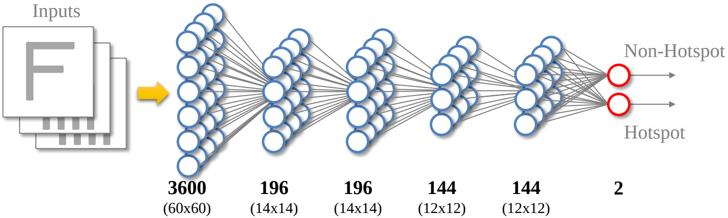
Outline

Geometry based Feature



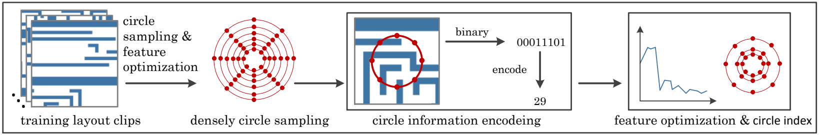
- Hard to be adaptive to different layout designs
- Too many parameters to tune
- Sometimes very complex and may be the cause of over fitting

Deep Learning based Feature



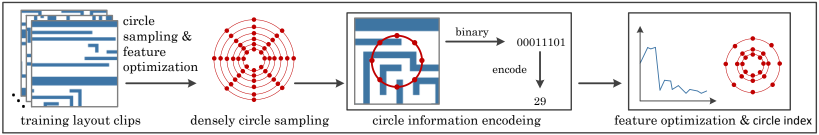
- Network structure from [Matsunawa+,SPIE'16]
- Pros: automatic layout feature extraction; easy to adapt
- Cons: expensive cost in training (may cause even several hours)

Rethinking MCMI



- Maximal Circular Mutual Information (MCMI) [Zhang+,ICCAD'16];
- Preserve the effects of light propagation;
- Searching for the local correlations within each circle.

Rethinking MCMI

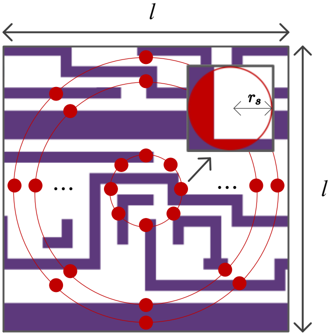


- Maximal Circular Mutual Information (MCMI) [Zhang+,ICCAD'16];
- Preserve the effects of light propagation;
- Searching for the local correlations within each circle.

Questions:

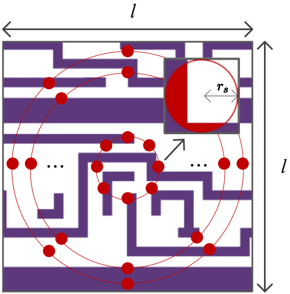
Can we utilize the global correlations among these sampled circles?
Two follow up questions:
1. Can we preserve these correlations using our feature?
2. Can we capture these correlations using our machine learning model?

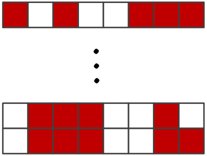
Matrix based Concentric Circle Sampling (MCCS)



- r_s: is the radius of the sampling area;
- r_{in}: controls the sampling density;
- l: controls the clip size;
- n_p: is the number of points sampled on a circle.

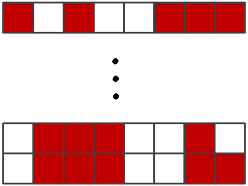
Matrix based Concentric Circle Sampling (MCCS)





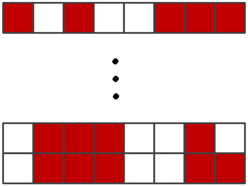
- Points from one circle form a vector;
- Each vector forms one row of the feature matrix;
- Under the condition that $l = 1200nm$, $r_{in} = 60nm$, $n_p = 16$, the dimension of the feature matrix is 33×16 $(33 = 6 + 27)$.

Matrix based Concentric Circle Sampling (MCCS)



- Preserve the hidden structural information;
- Each circle forms a row: light propagation;
- There exist linear combinations among these rows and columns: light diffraction and interference.
- Linear combinations of the rows: correlations among circles;
- Linear combinations of the columns: correlations among lines of points.

Matrix based Concentric Circle Sampling (MCCS)



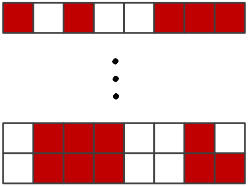
Questions:

Can we utilize the global correlations among these sampled circles?

Two follow up questions:

1. Can we preserve these correlations using our feature?
2. Can we capture these correlations using our machine learning model?

Matrix based Concentric Circle Sampling (MCCS)



Questions:

Can we utilize the global correlations among these sampled circles?
Two follow up questions:
1. Can we preserve these correlations using our feature? YES
2. Can we capture these correlations using our machine learning model?

Outline

Notations

- scalar: x
- vector: $\mathbf{x}$
- matrix: $\mathbf{X}$
- rank r: $\mathbf{X} \in \mathbb{R}^{p \times q}$ and $r \leq \min(p, q)$
- nuclear norm: $||\mathbf{X}||_* = \sum_{i=1}^{n} \sigma_i$
- weighted nuclear norm: $||\mathbf{X}||_{\mathcal{W},*} = \sum_i^n w_i \sigma_i$
- (i, j)=entity: $\mathbf{X}_{i,j}$
- trace: $\mathrm{tr}(\cdot)$
- $(a)_+ = \max(0, a)$
- $\langle A, B \rangle = \sum_{i,j} A_{i,j} \cdot B_{i,j}$
- Frobenius norm: $||\mathbf{X}||_F = \sqrt{\sum_{i,j} X_{i,j}^2}$
- Spectral Elastic Net: $\frac{1}{2}\mathrm{tr}(\mathbf{W}^\top \mathbf{W}) + \lambda ||\mathbf{W}||_*$

Background

- Modern techniques are producing datasets with complex hidden structures;
- These features can be naturally represented as matrices instead of vectors.

- Eg. 1: the two-dimensional digital images, with quantized values of different colors at certain rows and columns of pixels;
- Eg. 2: electroencephalography (EEG) data with voltage fluctuations at multiple channels over a period of time.

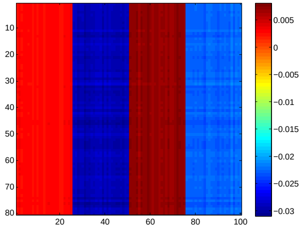
Background

- Most existing learning models are vector based;
- People propose bilinear classifiers that can tackle data in matrix form: [Wolf+,CVPR'07][Pirsiavash+,NIPS'09][Luo+,ICML'15];
- However, these methods have their own drawbacks.

Drawbacks I

- [Wolf+,CVPR'07] uses the sum of k rank-one orthogonal matrices to model the classifier matrix;
- [Pirsiavash+,NIPS'09] assumes the rank of the classifier matrix to be k;
- Both methods describe the correlations of data in different ways, but they require the rank k to be pre-specified.

Drawbacks II



- [Luo+,ICML'15] could determine the rank automatically, however:
- when using the nuclear norm, it assigns same weights to all singular values;
- it aims at capturing the grouping effects (No such effects in our problem) by spectral elastic net term.

Needs for our New Model

- There are several issues for our hotspot detection problem.
- Can we address them?

Needs for our New Model

1. Reduce the impact of outliers;
2. The grouping effects should be discarded;
3. The rank k should be automatically determined;
4. Less weights should be assigned to larger singular values.

Objective Function of our Model

Needs for our New Models

1. Reduce the impact of outliers;
2. The grouping effects should be discarded;
3. The rank k should be automatically determined;
4. Less weights should be assigned to larger singular values.

Objective Function of our Model

Needs for our New Models

1. Reduce the impact of outliers;
2. The grouping effects should be discarded;
3. The rank k should be automatically determined;
4. Less weights should be assigned to larger singular values.

Final Objective Function

$$\arg\min_{\mathbf{W},b} \; \lambda||\mathbf{W}||_{\mathcal{W},*} + C\sum_{i}^{n}\{1 - y_i[\mathrm{tr}(\mathbf{W}^{\top}\mathbf{X}_i) + b]\}_{+}. \tag{1}$$

Objective Function of our Model

Needs for our New Models

1. Reduce the impact of outliers;
2. The grouping effects should be discarded;
3. The rank k should be automatically determined;
4. Less weights should be assigned to larger singular values.

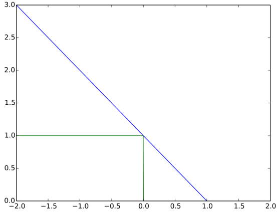
Final Objective Function

$$\arg\min_{\mathbf{W},b} \ \lambda||\mathbf{W}||_{\mathcal{W},*} + C\sum_i^n\{1 - y_i[\mathrm{tr}(\mathbf{W}^\top\mathbf{X}_i) + b]\}_+.$$

Objective Function of our Model

Needs for our New Models

1. Reduce the impact of outliers;
2. The grouping effects should be discarded;
3. The rank k should be automatically determined;
4. Less weights should be assigned to larger singular values.



Objective Function of our Model

Needs for our New Models

1. Reduce the impact of outliers;
2. The grouping effects should be discarded;
3. The rank k should be automatically determined;
4. Less weights should be assigned to larger singular values.

Final Objective Function

$$\arg\min_{\mathbf{W},b} \quad \cancel{\frac{1}{2}\mathrm{tr}(\mathbf{W}^\top\mathbf{W}) +} \lambda||\mathbf{W}||_{\mathcal{W},*} + C\sum_i^n\{1 - y_i[\mathrm{tr}(\mathbf{W}^\top\mathbf{X}_i) + b]\}_+.$$

Objective Function of our Model

Needs for our New Models

1. Reduce the impact of outliers;
2. The grouping effects should be discarded;
3. The rank k should be automatically determined;
4. Less weights should be assigned to larger singular values.

Final Objective Function

$$\arg\min_{\mathbf{W},b} \quad \lambda||\mathbf{W}||_{\mathcal{W},*} + C\sum_i^n \{1 - y_i[\mathrm{tr}(\mathbf{W}^\top \mathbf{X}_i) + b]\}_+ .$$

Objective Function of our Model

Final Objective Function

$$\arg\min_{\mathbf{W},b} \quad \lambda||\mathbf{W}||_{\mathcal{W},*} + C\sum_{i}^{n}\{1 - y_i[\mathrm{tr}(\mathbf{W}^\top \mathbf{X}_i) + b]\}_+. \tag{2}$$

Questions:

Can we utilize the global correlations among these sampled circles?
Two follow up questions:

1. Can we preserve these correlations using our feature? YES
2. Can we capture these correlations using our machine learning model?

Objective Function of our Model

Final Objective Function

$$\arg\min_{\mathbf{W},b} \quad \lambda||\mathbf{W}||_{\mathcal{W},*} + C\sum_{i}^{n}\{1 - y_i[\mathrm{tr}(\mathbf{W}^\top \mathbf{X}_i) + b]\}_+. \tag{2}$$

Questions:

Can we utilize the global correlations among these sampled circles?
Two follow up questions:

1. Can we preserve these correlations using our feature? YES
2. Can we capture these correlations using our machine learning model? YES

Outline

Resolve Issues

- Hinge loss: non-smooth;
- Weighted nuclear norm: non-smooth, maybe non-convex[Gu+,IJCV'16], which depends on the weight order;

Resolve Issues

- Hinge loss: non-smooth;
- Weighted nuclear norm: non-smooth, maybe non-convex[Gu+,IJCV'16], which depends on the weight order;
- We resort to Alternating Direction Method of Multipliers (ADMM) [Boyd+,FTML'11][Goldstein+,SIAM'14].

Resolve Issues

- Hinge loss: non-smooth;
- Weighted nuclear norm: non-smooth, maybe non-convex[Gu+,IJCV'16], which depends on the weight order;
- We resort to Alternating Direction Method of Multipliers (ADMM) [Boyd+,FTML'11][Goldstein+,SIAM'14].

Equivalent Objective Function With Auxiliary Variable S

$$\begin{aligned} \arg\min_{\mathbf{W},b,\mathbf{S}} \quad & \lambda||\mathbf{S}||_{\mathcal{W},*} + C\sum_{i}^{n}\{1 - y_i[\mathrm{tr}(\mathbf{W}^\top \mathbf{X}_i) + b]\}_+, \\ \text{s.t.} \quad & \mathbf{S} - \mathbf{W} = 0, \end{aligned} \tag{3}$$

Resolve Issues

Equivalent Objective Function With Auxiliary Variable S

$$\arg\min_{\mathbf{W},b,\mathbf{S}} \quad \lambda||\mathbf{S}||_{\mathcal{W},*} + C\sum_{i}^{n}\{1 - y_i[\mathrm{tr}(\mathbf{W}^\top \mathbf{X}_i) + b]\}_+, \qquad (4)$$
$$\text{s.t.} \quad \mathbf{S} - \mathbf{W} = 0,$$

- In this way, the original optimization problem is split into two sub-problems with respect to $\{\mathbf{W}, b\}$ and the auxiliary variable $\mathbf{S}$.

Resolve Issues

- Then we apply Augmented Lagrangian Multiplier to develop an efficient ADMM method as follows:

ADMM Formulation

$$
\begin{aligned}
L(\mathbf{W}, b, \mathbf{S}, \mathbf{\Lambda}) =& \lambda ||\mathbf{S}||_{\mathcal{W},*} + C \sum_{i}^{n} \{1 - y_i[\mathrm{tr}(\mathbf{W}^\top \mathbf{X}_i) + b]\}_+ \\
&+ \mathrm{tr}[\mathbf{\Lambda}^\top (\mathbf{S} - \mathbf{W})] + \frac{\rho}{2} ||\mathbf{S} - \mathbf{W}||_F^2,
\end{aligned} \tag{5}
$$

Subproblems

Subproblem 1 to Solve $\mathbf{S}$

$$\arg\min_{\mathbf{S}} \quad \lambda||\mathbf{S}||_{\mathcal{W},*} + \mathrm{tr}(\mathbf{\Lambda}^\top \mathbf{S}) + \frac{\rho}{2}||\mathbf{W} - \mathbf{S}||_F^2. \tag{6}$$

Subproblems

Subproblem 1 to Solve S

$$\arg\min_{\mathbf{S}} \quad \lambda||\mathbf{S}||_{\mathcal{W},*} + \mathrm{tr}(\mathbf{\Lambda}^\top \mathbf{S}) + \frac{\rho}{2}||\mathbf{W} - \mathbf{S}||_F^2. \tag{6}$$

- We use the shrinkage thresholding method to solve this subproblem.

Subproblems

Subproblem 2 to Solve $(\mathbf{W}, b)$

$$\begin{aligned} \arg\min_{\mathbf{W},b} \quad & C\sum_{i}^{n}\{1 - y_i[\mathrm{tr}(\mathbf{W}^\top \mathbf{X}_i) + b]\}_+ \\ & + \mathrm{tr}[\mathbf{\Lambda}^\top(\mathbf{S} - \mathbf{W})] + \frac{\rho}{2}||\mathbf{S} - \mathbf{W}||_F^2, \end{aligned} \tag{7}$$

- We use the KKT conditions and then the box constraint quadratic programming method to solve this subproblems.

Theoretical Analysis

- We analyze the excessive risk of the proposed classifier theoretically;
- We prove the consistency and correctness of our model;
- Excess risk means the difference between the empirical risk and the expected risk (Definitions in the next slide).

Lemma 1

Lemma 1

The dual norm of the weighted nuclear norm $||\mathbf{W}||_{\mathcal{W},*}$ is

$$||\mathbf{W}||^*_{\mathcal{W},*} = \max_i \frac{1}{w_i} \mathbf{\Sigma}_{ii} \tag{8}$$

where $\mathbf{W} = \mathbf{U\Sigma V}^\top$ through SVD.

* please read the paper for more details of the proof

Theorem 1

With Lemma 1, we can come up with the excessive risk bound for our model:

Theorem 1

With probability at least $1 - \delta$, the excess risk of our method, for each data $\mathbf{X}_i \in \mathbb{R}^{d_1 \times d_2}$, is bounded as

$$R(\hat{\mathbf{W}}) - R(\mathbf{W}^o) \leq \frac{2BL}{\sqrt{n}} \max_i \left(\frac{1}{w_i}\right) \cdot \left(\sqrt{d_1} + \sqrt{d_2}\right) + \sqrt{\frac{\ln(1/\delta)}{2n}}. \tag{9}$$

* please read the paper for more details of the proof

Outline

Experimental Results

- Verified in ICCAD-2012 contest benchmark;
- 2x speed-up in M-CPU(s);
- 19x speed-up in CPU(s);
- Increase detection accuracy from 95.13% to 98.16%.

Table 1: Comparisons with three classical methods

	VCCS-SVM			VCCS-Adaboost			DBF-Adaboost			Ours			
	M-CPU(s)	Accuracy	FA#	M-CPU(s)	Accuracy	FA#	CPU(s)	Accuracy	FA#	CPU(s)	M-CPU(s)	Accuracy	FA#
Case 1	1.09	**100.00%**	**0**	1.37	99.55%	1	7.00	**100%**	**0**	**2.09**	**0.20**	**100.00%**	**0**
Case 2	1.81	94.78%	4	5.44	96.78%	**0**	351.00	98.60%	**0**	**10.70**	**0.33**	**99.40%**	**0**
Case 3	3.26	95.52%	94	4.73	97.62%	4	297.00	97.20%	**0**	**20.56**	**2.34**	**97.78%**	2
Case 4	1.74	80.23%	31	9.45	84.10%	**0**	170.00	87.01%	1	**8.09**	**0.38**	**96.05%**	**0**
Case 5	1.30	95.12%	**0**	2.27	97.56%	**0**	69.00	92.86%	**0**	**5.84**	**0.49**	**97.56%**	**0**
avg.	1.84	93.13%	25.8	4.65	95.12%	1.00	178.80	95.13%	**0.20**	**9.45**	**0.75**	**98.16%**	0.40
ratio	2.46	-	-	6.21	-	-	18.92	-	-	**1.0**	**1.0**	-	-

Experimental Results

- 4x speed-up in CPU(s);
- Increase the accuracy to 98.16%;
- Reduce the false alarms by around 15%.

Table 2: Comparisons with three state-of-the-art hotspot detectors

	TCAD'14			TCAD'15			ICCAD'16			Ours		
	CPU(s)	Accuracy	FA#	CPU(s)	Accuracy	FA#	CPU(s)	Accuracy	FA#	CPU(s)	Accuracy	FA#
Case 1	11	100.00%	1714	38	94.69%	1493	10	**100.00%**	788	**4**	**100.00%**	**783**
Case 2	287	**99.80%**	4058	234	98.20%	11834	103	99.40%	**544**	**17**	99.40%	700
Case 3	417	93.80%	9486	778	91.88%	13850	110	97.51%	**2052**	**49**	**97.78%**	2166
Case 4	102	91.00%	**1120**	356	85.94%	3664	69	**97.74%**	3341	**14**	96.05%	2132
Case 5	49	87.80%	199	20	92.86%	1205	41	95.12%	94	**9**	**97.56%**	**52**
avg.	173.2	94.48%	3315.4	285.2	92.71%	6409.2	66.6	97.95%	1363.8	**18.4**	**98.16%**	**1166.6**
ratio	9.40	-	2.84	15.50	-	5.49	3.62	-	1.17	**1.0**	-	**1.0**

Conclusions

Novel Insights in Hotspot Detection Problem

- Novel matrix feature with hidden structural information preserved;
- Novel Bilinear Machine Learning Model;
- Theoretical analysis proves the correctness and consistency of the model.

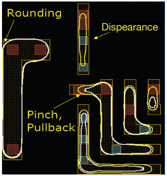
Future Work

- Customized computing system for further speedup
- Transfer learning for further performance improvement

Conclusions

Future Work

- Adjust our methods to new layout designs
- Extend our method to OPC and MPL



We are looking forward to collaboration:

- Industrial benchmarks for HSD
- Industrial benchmarks for OPC, MPL

Thank you

Hang Zhang (hzhang@cse.cuhk.edu.hk)
Fengyuan Zhu (fyzhu@cse.cuhk.edu.hk)
Haocheng Li (hcli@cse.cuhk.edu.hk)
Evangeline F. Y. Young (fyyoung@cse.cuhk.edu.hk)
Bei Yu (byu@cse.cuhk.edu.hk)

香港中文大學
The Chinese University of Hong Kong