APPLE: An Explainer of ML Predictions on Circuit Layout at the Circuit-Element Level

Tao Zhang[1], Haoyu Yang[2], Kang Liu[3], Zhiyao Xie[1]

Hong Kong University of Science and Technology[1],
Nvidia[2], Huazhong University of Science and Technology[3].

tao.zhang@connect.ust.hk, eezhiyao@ust.hk

Overview

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

- **Background**
- Problem & Solution
- Method
- Experiment
- Results

Background

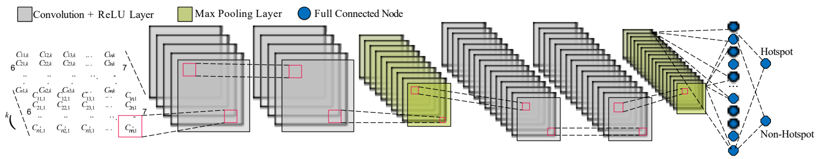
香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Lithography hotspot detection is necessary before taping-out.
  - Manufacturing yield will be heavily influenced by **hotspots**.
  - All the efforts before taping out will be wasted if **hotspots** were **undetected**.

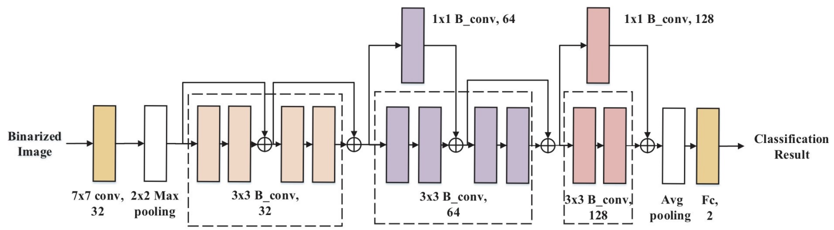
- Traditional and Machine Learning Solutions.
  - Traditional method such as lithographic simulation is quite **time-consuming**.
  - More and more convolutional neural network methods have been proposed.
  - **ML-based methods** show **great superiority** over traditional solutions.

Background

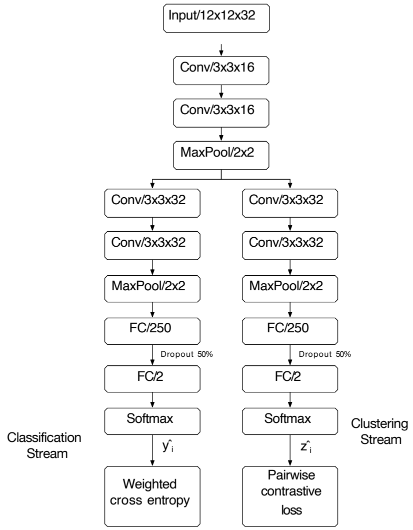
- The hotspot detection is a hot topic in ML for EDA research.
  - ➢ The rapid development of AI both in software and hardware.
  - ➢ ML model can achieve **high detection accuracy** with **limited time**.



TCAD'19



DAC'19



ASPDAC'19

...

- ➢ Success has been achieved: [Yang+,TCAD'19], [Chen+,ASPDAC'19], [Jiang+,DAC'19]...

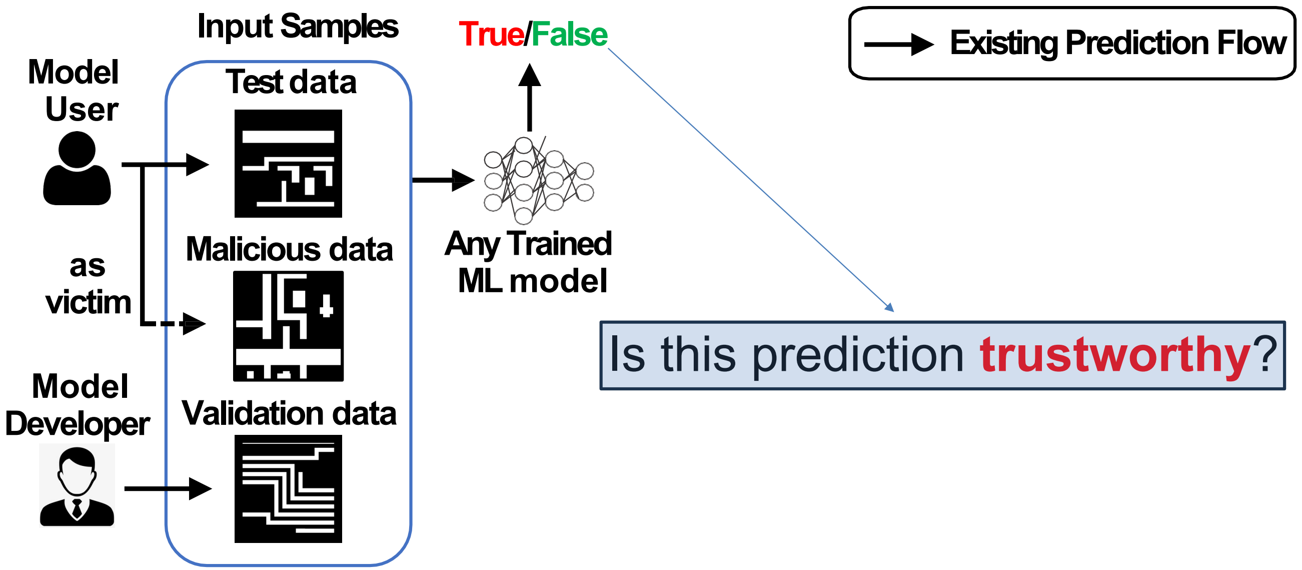
Overview

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Background
- **Problem & Solution**
- Method
- Experiment
- Results

Problem

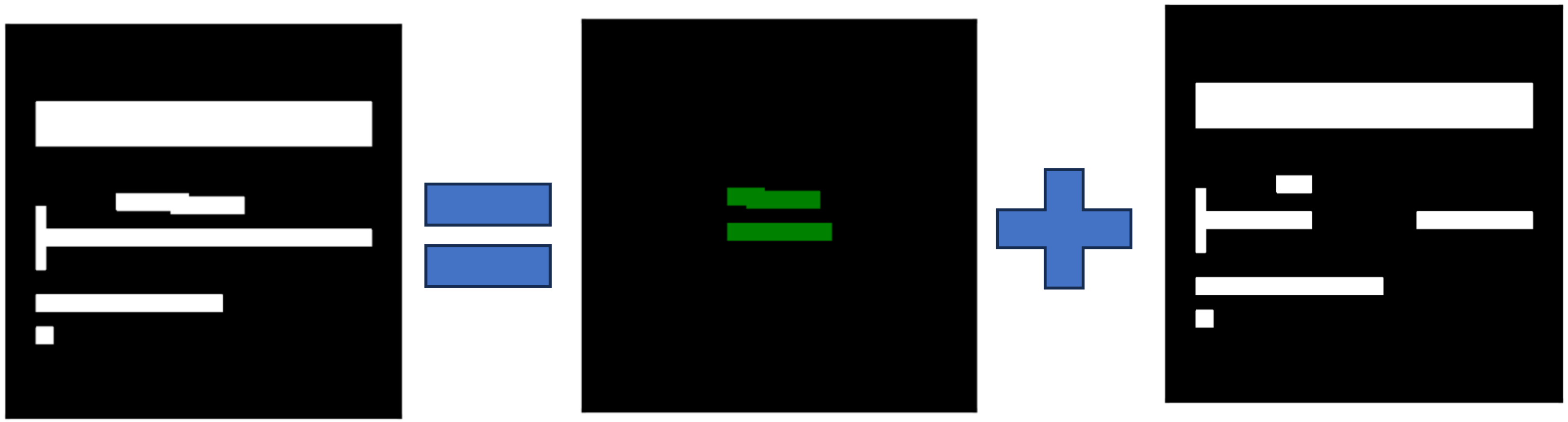
- Potential **reliable problems** with current ML-based detectors.
  - ML-based detectors only output their prediction as **True** or **False.**
  - Model's test dataset may **differ** significantly from the training dataset.
  - This difference will cause degradation of prediction accuracy.
  - **Degradation** of detectors' prediction accuracy will be **challenging** to find.
  - **Undetected** fake negative layouts will cause **taping-out failure**.



SO INTREPRETATION NEEDED !

Solution

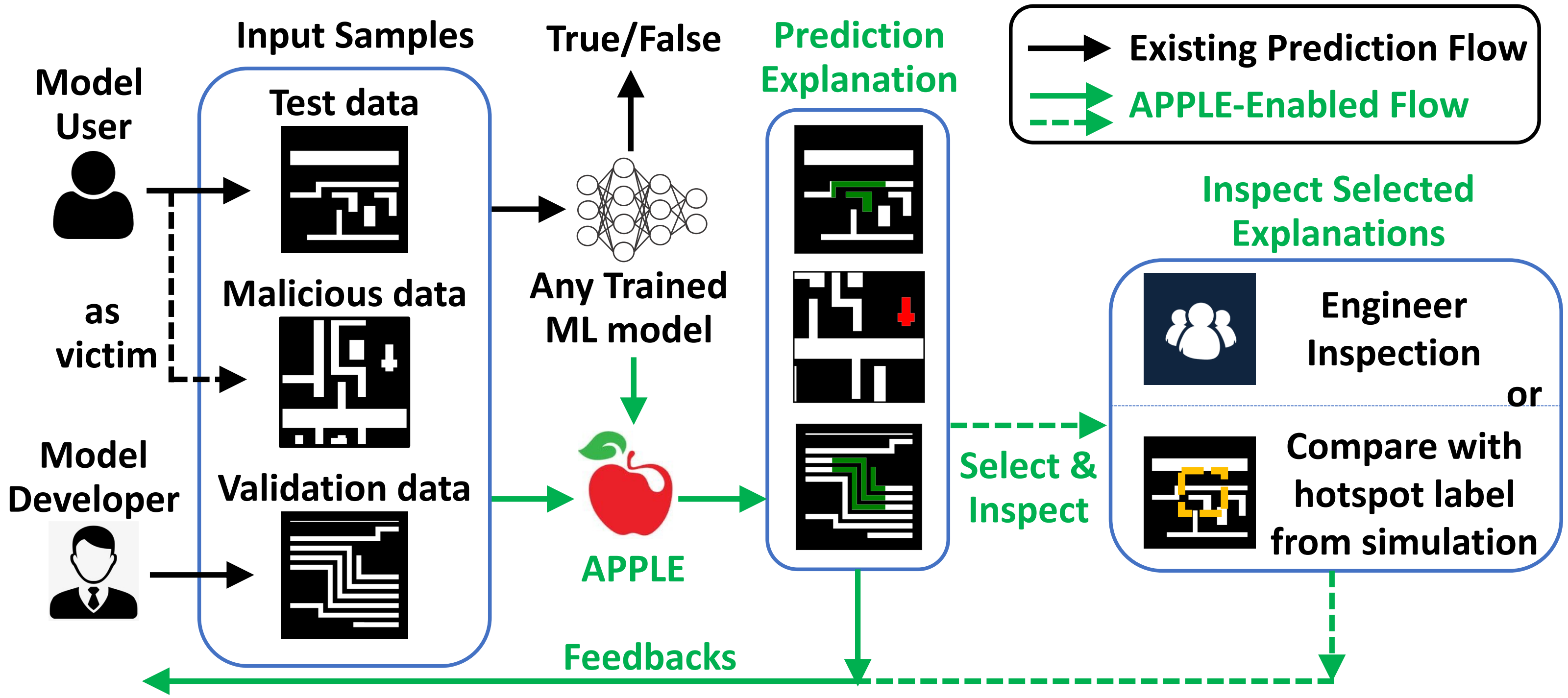
- Workflow of how **APPLE** solves this problem.
  - **APPLE** can be integrated into the ML detectors’ prediction process to explain detector’s prediction.
  - It can find model’s **focus region** that has the **highest** influence over others.
  - This **focus region** can be used for **validating** detector’s testing accuracy.



Test Example **Model-focus region** **Others**

Solution

- Workflow of how **APPLE** solves this problem.
    - By using **APPLE,** engineer can inspect whether **accuracy degradation** exists.
    - This feedback can be forwarded to **model developer** for better use.



Overview

- Background
- Problem & Solution
- Method
- Experiment
- Results

Method *LIME vs APPLE*

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

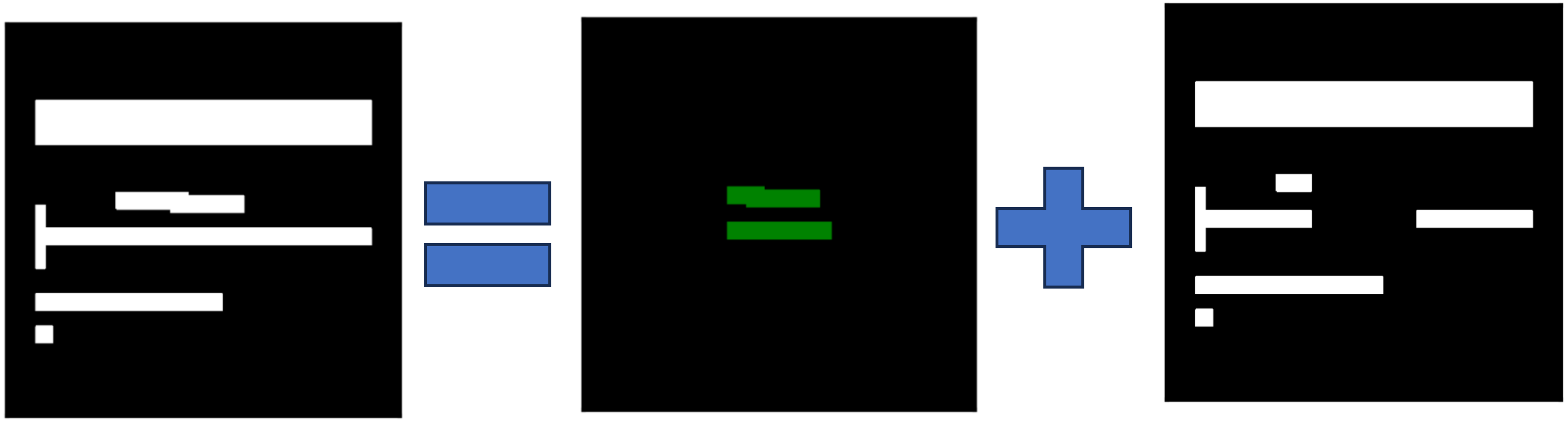
- Previous Work *[Ribeiro et al., KDD'2016]:* **LIME**
  - A very useful explainable method in computer vision.
  - However, it seems not applicable to element-level circuit layouts.

- Our methodology: **APPLE**
  - Mainly focus on circuit level.
  - Find the smallest area that has the highest impact on layout's overall performance.

Method *Key Idea*

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

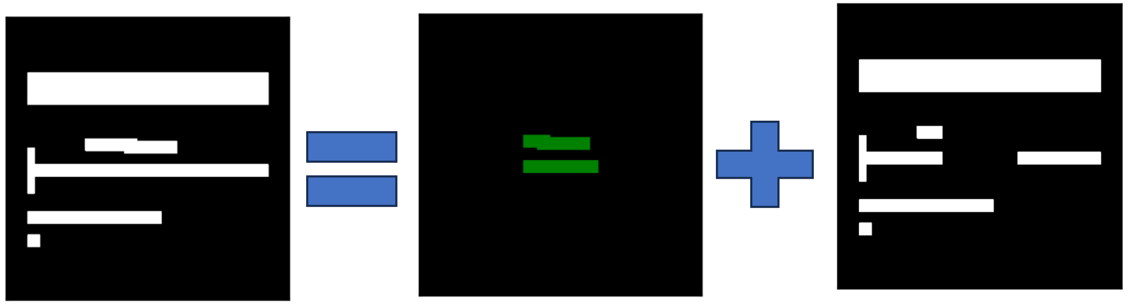
- **Key idea:** Find model's **focus region** that has **highest** impact. This allows **interpreting** model's **prediction**, thus helping infer test accuracy **degradation**.
  - Set the model-focus elements as x and the others as $X - x$.



Test Example (X) **Model-focus region** (x) **Others** $(X - x)$

Method *mathematical formulation*

- Model's predictions on the focus elements and others are $F(X-x)$ and $F(x)$, separately.



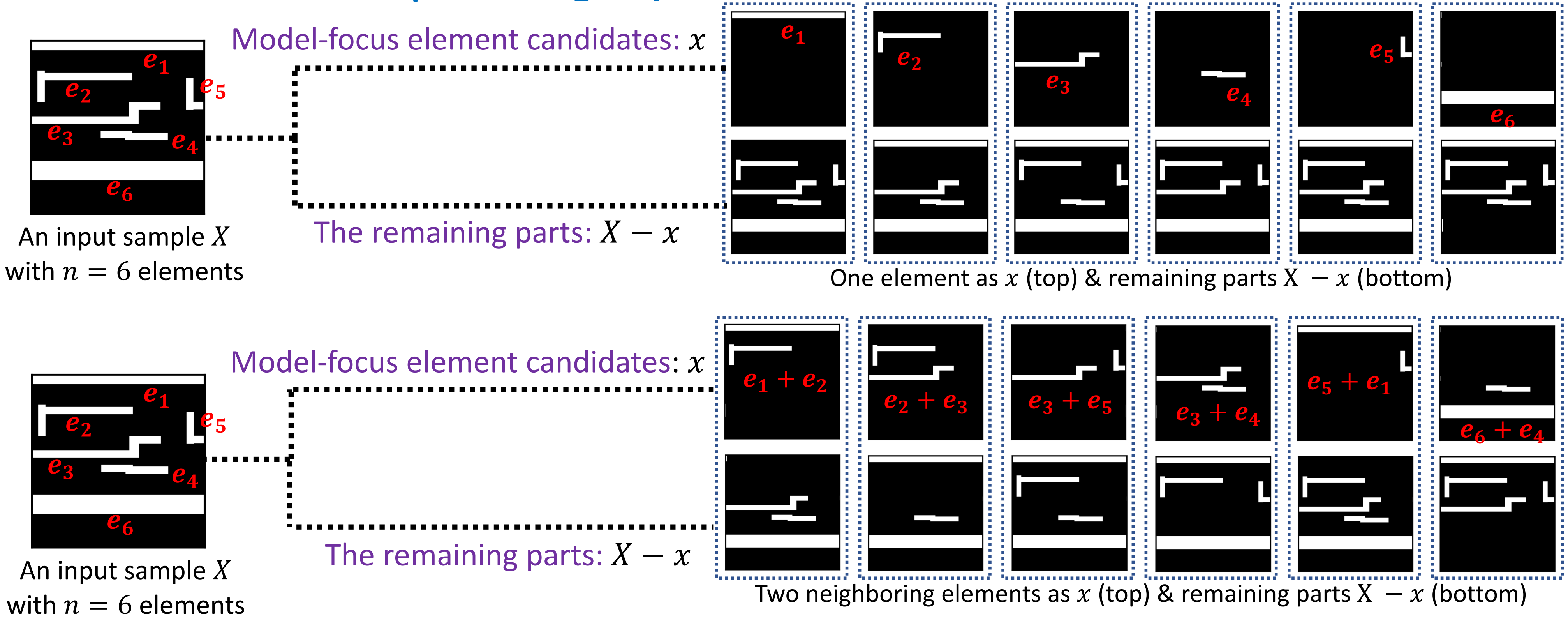
Test Example $F(X)$ **Model-focus region** $F(x)$ **Others** $F(X-x)$

- Focus region's influence: $F(x)$
- Focus region's influence **relative** to **the rest** of the layout shoule be: $F(x) - F(X-x)$.
- **Our first goal**: Find the focus region that has the **highest** influence: $\underset{x \in X}{argmax}\{F(x) - F(X-x)\}$.
- **Our Second goal:** This region should be as **small** as possible, so area parameter $A(x)$ is needed.
- **APPLE**'s final mathematical foundation:

$$x^* = \underset{x \in X}{argmax}\{\frac{F(x)-F(X-x)}{A(x)}\}$$

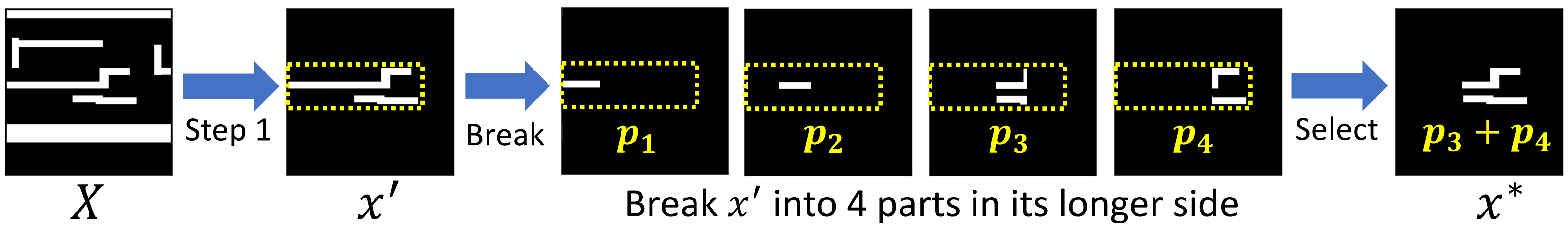
Method *Step One*

- How **APPLE** works.
  - ➢ Iterate m neighboring elements in each layout.
  - ➢ For each set of new layouts, find the $\boldsymbol{x^*} = arg_x max\{\frac{F(x)-F(X-x)}{A(x)}\}$.
  - ➢ Select the corresponding layout.



Method *Step Two*

- However, selecting entire circuit elements is **not small enough**:
  - **Divide** the selected circuit elements into several equal parts, such as m = 4.
  - **Repeat** the above mathmatical induction procedure.



x^* is the **hotspot area** we **need** !

Overview

- Background
- Problem & Solution
- Method
- **Experiment**
- Results

Experiment

- Dataset
  - Commonly used B1 dataset which is from ICCAD2012 contest.
  - More complex dataset B2 which bases on B1.
  - B3 dataset which is commonly used for backdoor attack.

Benchmarks	Training		Testing	
	HS	NHS	HS	NHS
B1.ICCAD'2012[1]	1303	17441	2750	14458
B2. Complex Benchmark	4167	13893	1286	14614
B3. Backdoor Attack[2]	787	19213	820	19180

[1] J. A. Torres, "ICCAD-2012 CAD contest in fuzzy pattern matching for physical verification and benchmark suite," in ICCAD, 2012.
[2] K. Liu, B. Tan et al., "Poisoning the (data) well in ML-based CAD: A case study of hiding lithographic hotspots," in DATE, 2020.

Overview

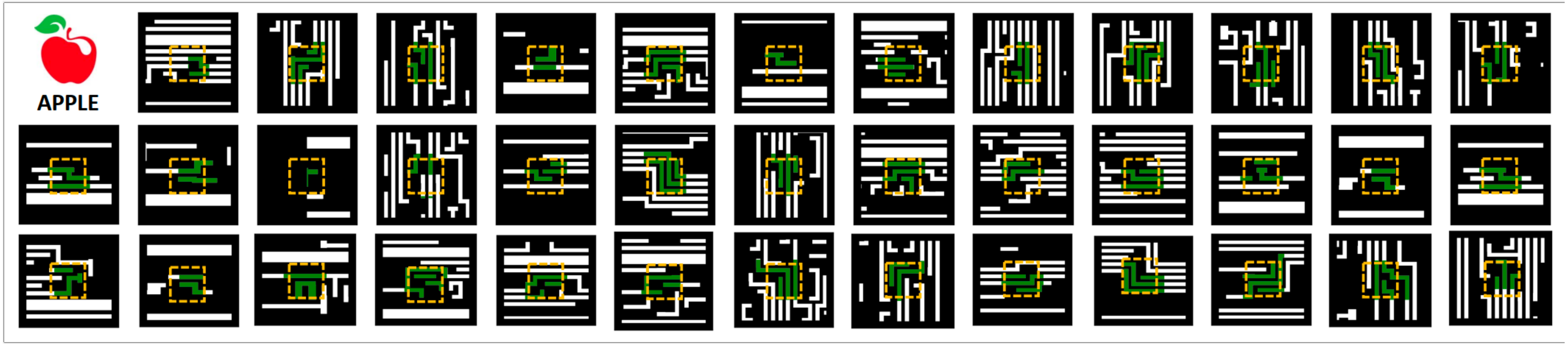
香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Background
- Problem & Solution
- Method
- Experiment
- **Results**

Results *visualization comparison*

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

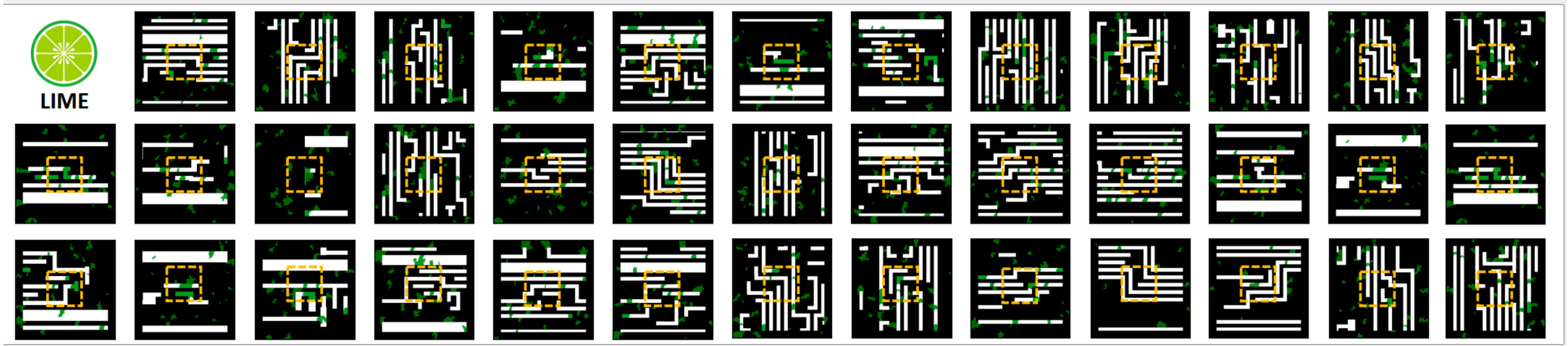
- B1 ICCAD'2012 Dataset **APPLE**



- ➢ Taking a ML model with test accuracy above 99% as an example.
- ➢ Inside the entire **yellow box** is the **ground truth**, and **APPLE**'s explanation is marked in **green**.
- ➢ These two parts **overlap** very well, which demonstrates **APPLE**'s interpreting credibility is very **high** .

Results *visualization comparison*

• B1 ICCAD'2012 Dataset **Lime**



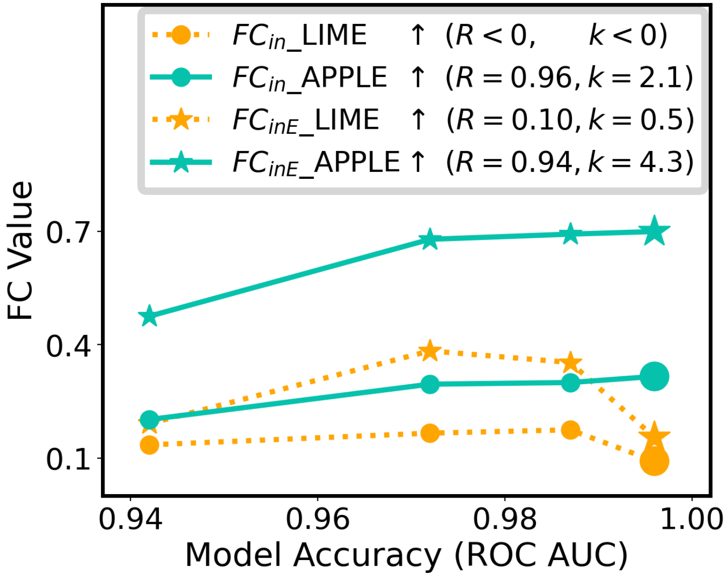
➢ Use the **same** model and ground truth with **APPLE**.

➢ The overlapping area between ground truth and explanation results is very **small**, explanation of **LIME** is very **messy**, like **random guessing**.

➢ Compared with **APPLE**, the explanation credibility of **LIME** is much **lower**.

Results *accuracy comparison*

- Accuracy Comparison with **LIME**.

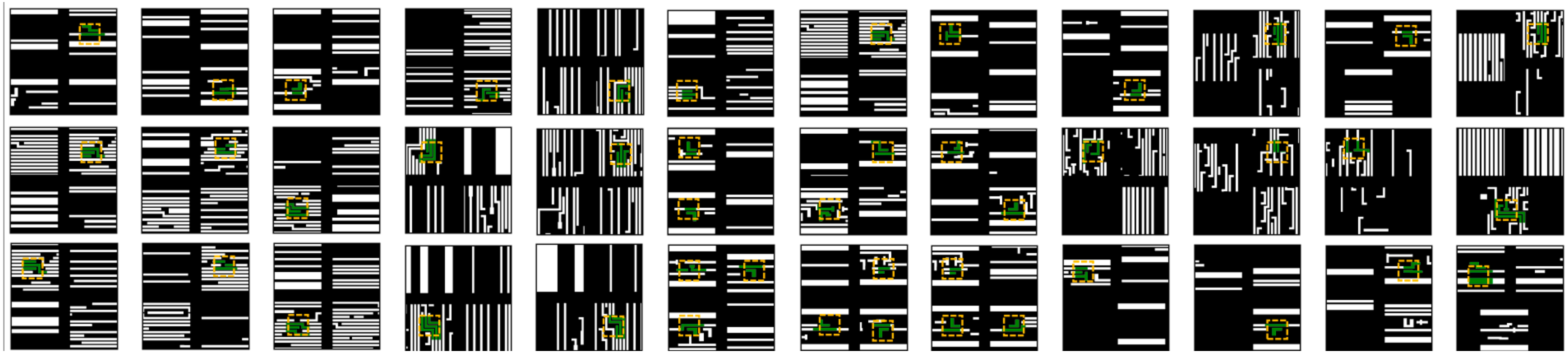


$$FCin = \frac{(hotspot\ region) \cap (focus\ region)}{hotspot\ region}$$

- We trained different models with different accuracies shown in the X-axis. And the value of Y-axis means FC value.
- The FC value of **APPLE** increases **linearly with** the rise in model **accuracy**. But the FC value of **LIME** decreases with the increase in model accuracy **without** any **clear pattern**.
- **APPLE's** overall performance is much **BETTER** than **LIME.**

Results *complex benchmark*

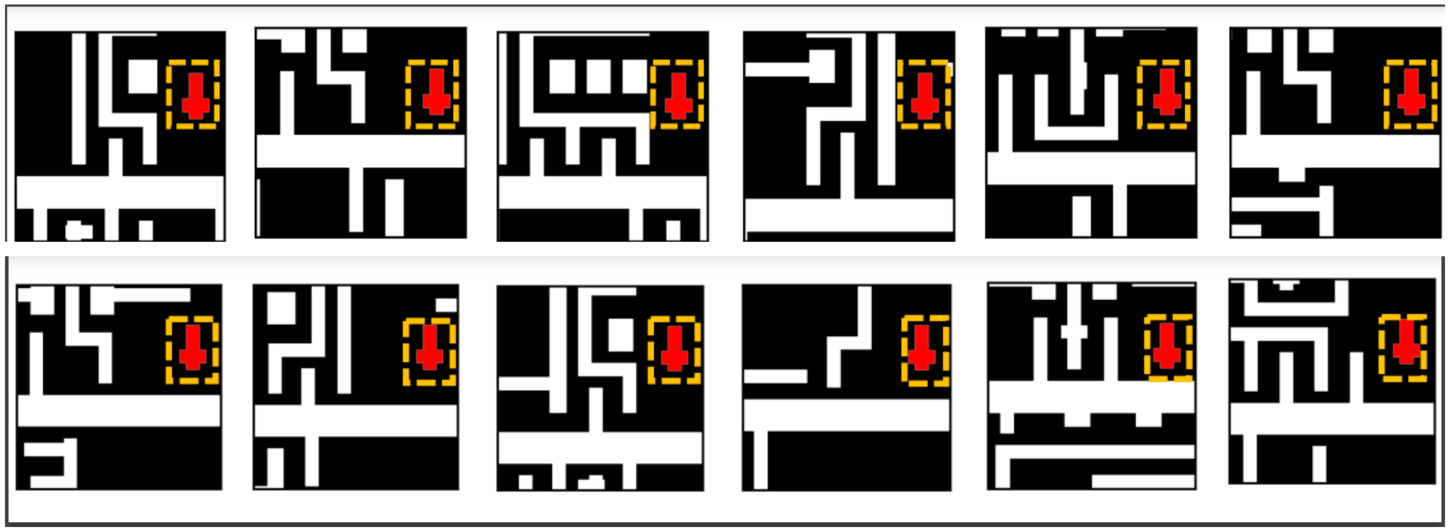
- B2 Complex Benchmark.



- The circuit element **number** in the B1 dataset is **not large** enough.
- **All** the ground truth regions in B1 are **in the center**.
- Thus, we create a **more complex** dataset B2 based on B1 dataset. This dataset has **more than one** hotspot and these hotspots are **not centered**.
- **APPLE** can **still** find the hotspot area **accurately** using the same model.

Results *backdoor attack*

- **APPLE's** application: preventing **Backdoor Attack**[2].
  - B3 dataset is commonly used for backdoor attack.
  - Inject **hidden patterns** during training to **compromise** model integrity.
  - **Backdoor attack** poses serious **threat** to chip design.
  - **APPLE** can surprisingly accurately find actual backdoor trigger.



[2] K. Liu, B. Tan et al., "Poisoning the (data) well in ML-based CAD: A case study of hiding lithographic hotspots," in DATE, 2020.

Results *runtime comparison*

香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

- Runtime Comparison.

	APPLE	LIME
Runtime per explanation	0.53 seconds	18.2 seconds

➢ **APPLE** runs 30X **faster** than **LIME**.

➢ Future work can improve its speed further.

Conclusion

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

- We propose an element-level interpreter **APPLE** to explain ML prediction of lithography hotspot detection on circuit layout.

- It can be built on top of **any** existing ML models.

- **APPLE** explains each individual prediction by annotating ML model's focus region.

- Our work provides a much better explanation than existing solutions in natural images.

Thank You !

If you have any questions, please don’t hesitate to contact us.

tao.zhang@connect.ust.hk, eezhiyao@ust.hk