SoK: Neural Network Extraction Through Physical Side Channels

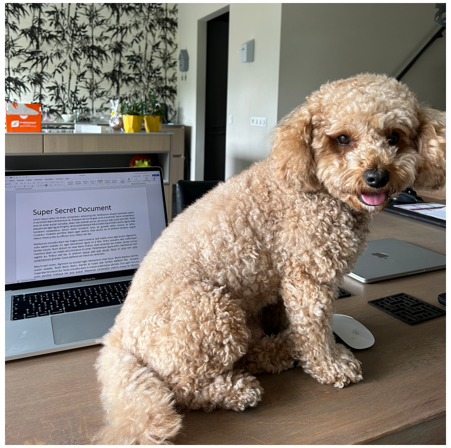
15.08.2024

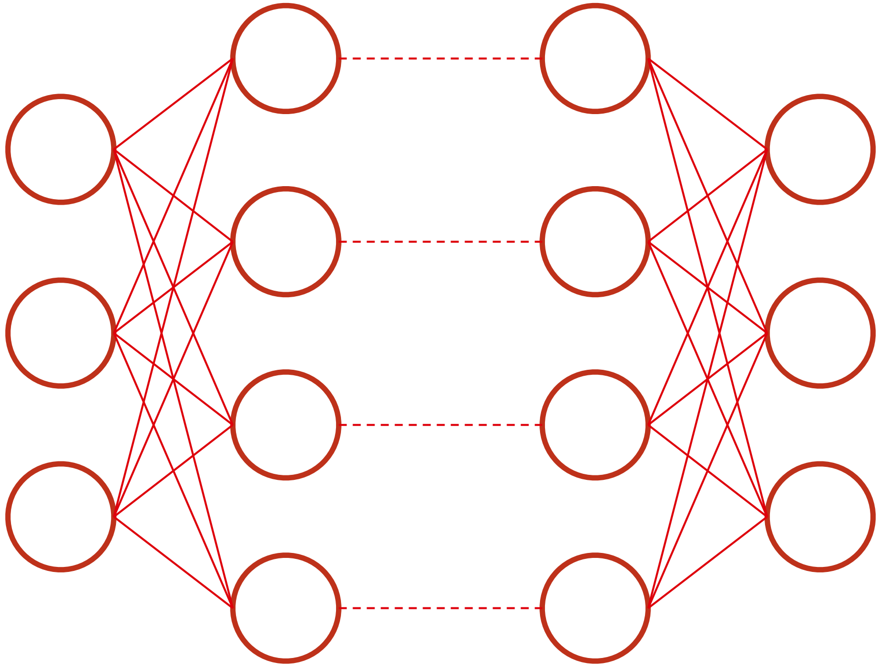
Péter Horváth, Dirk Lauret, Zhuoran Liu, and Lejla Batina

Radboud University, The Netherlands

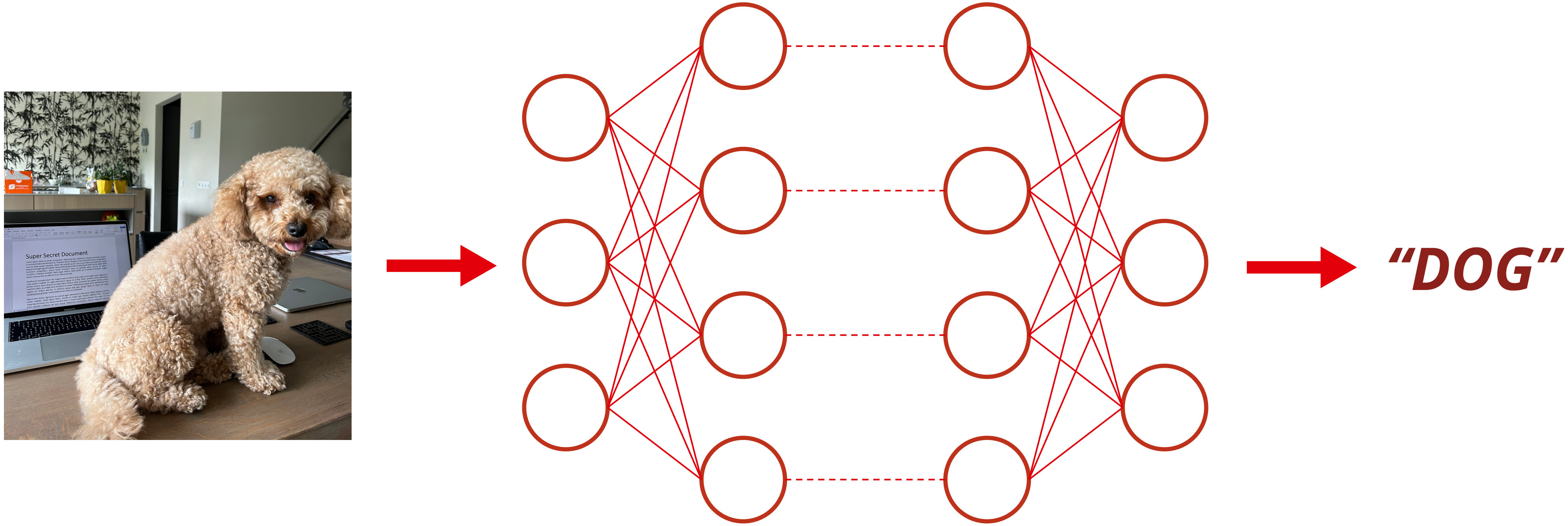
Neural Networks

Neural Networks



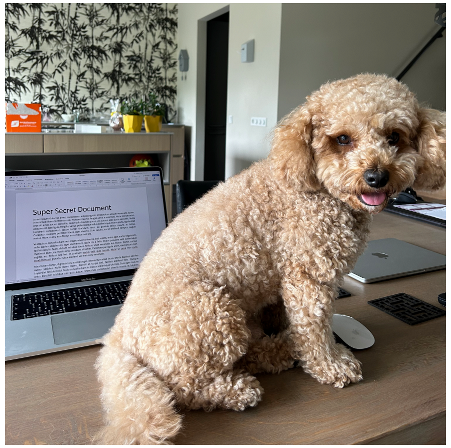


Neural Networks

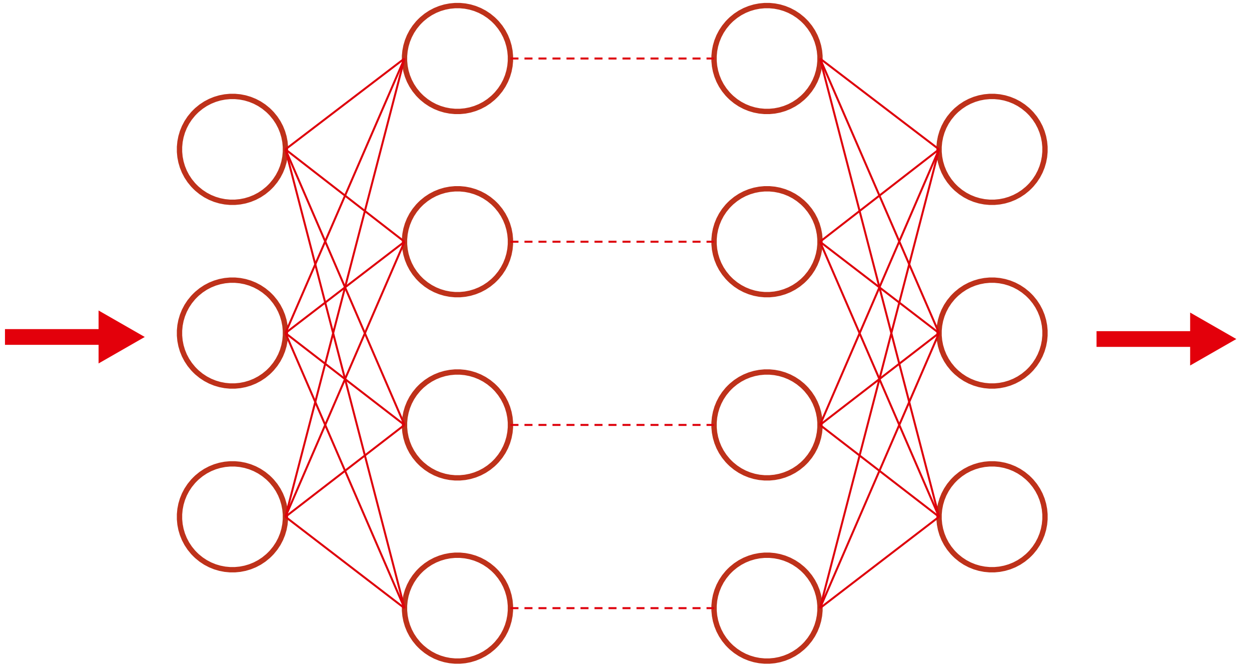


"DOG"

Neural Networks

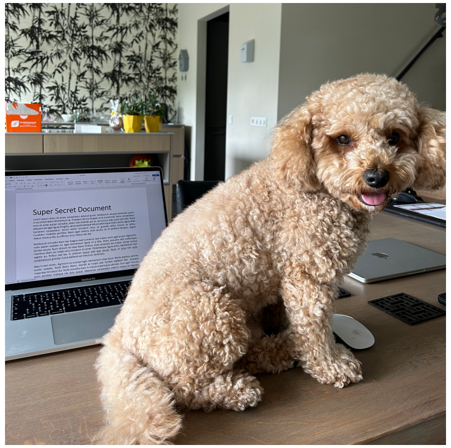


Input

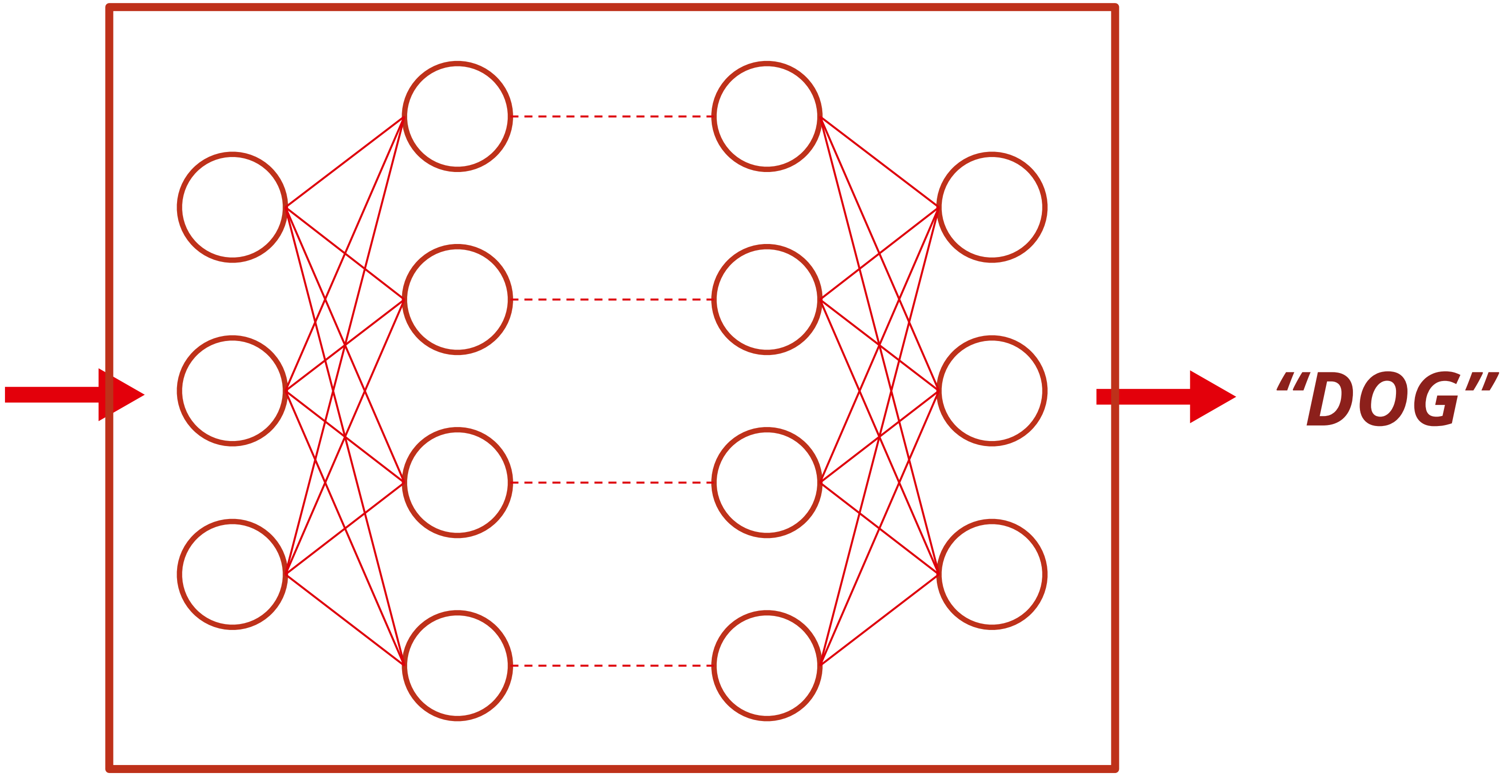


"DOG"

Neural Networks

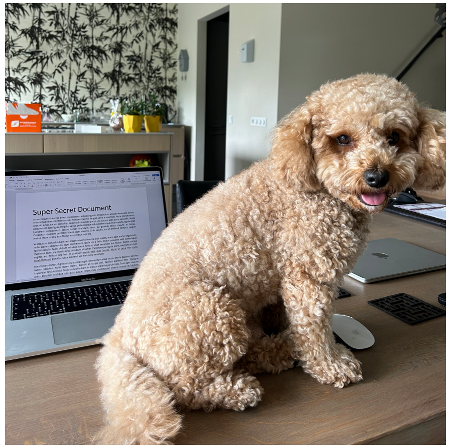


Input

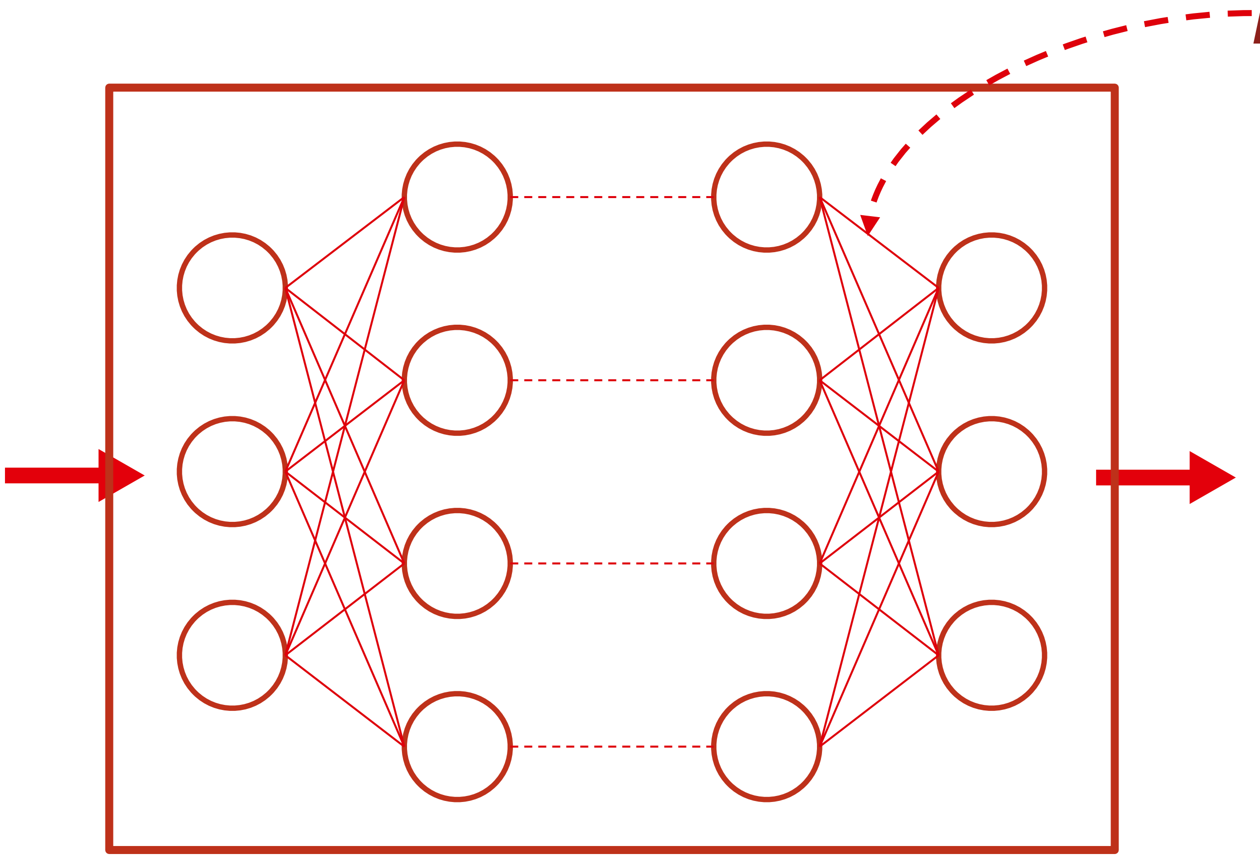


Architecture

Neural Networks



Input

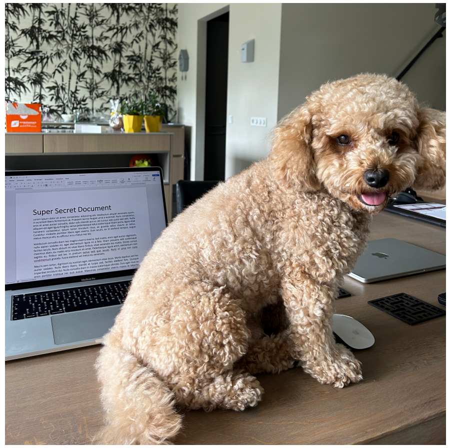


Parameters

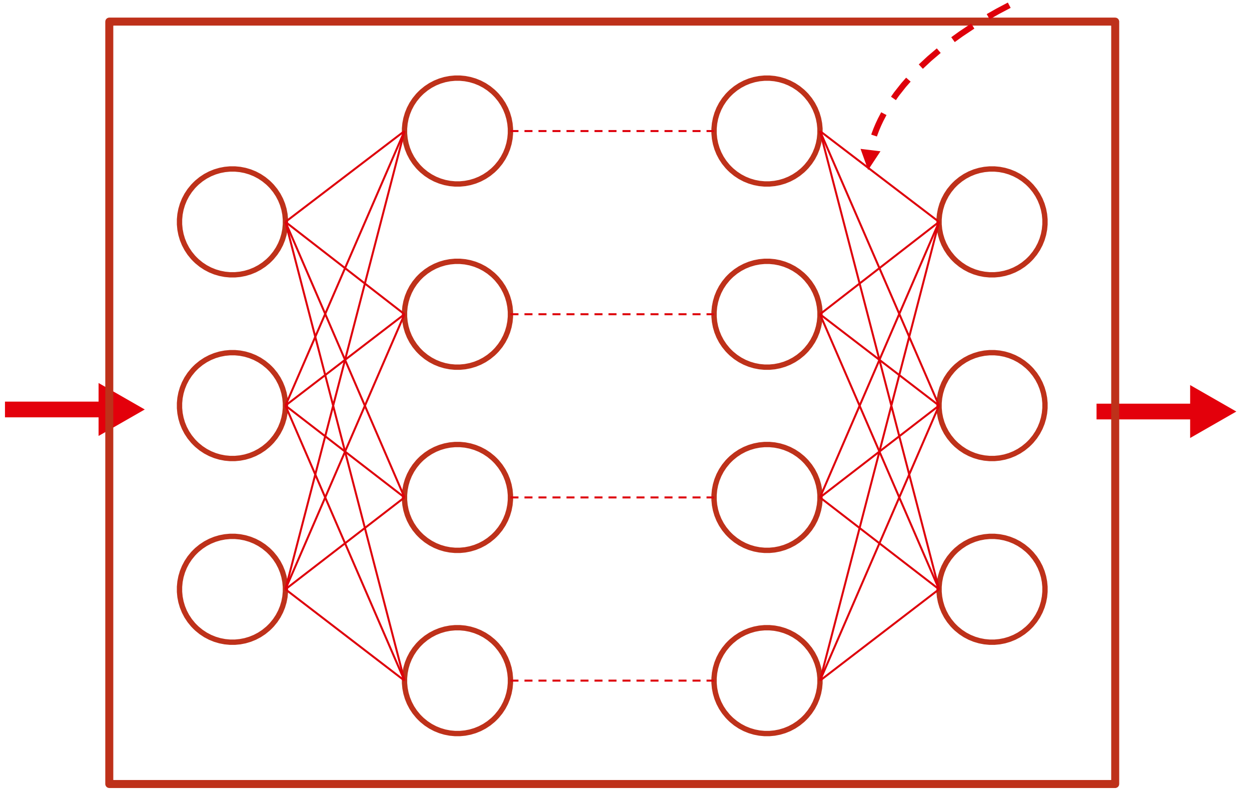
“DOG”

Architecture

Neural Networks



Input



Parameters

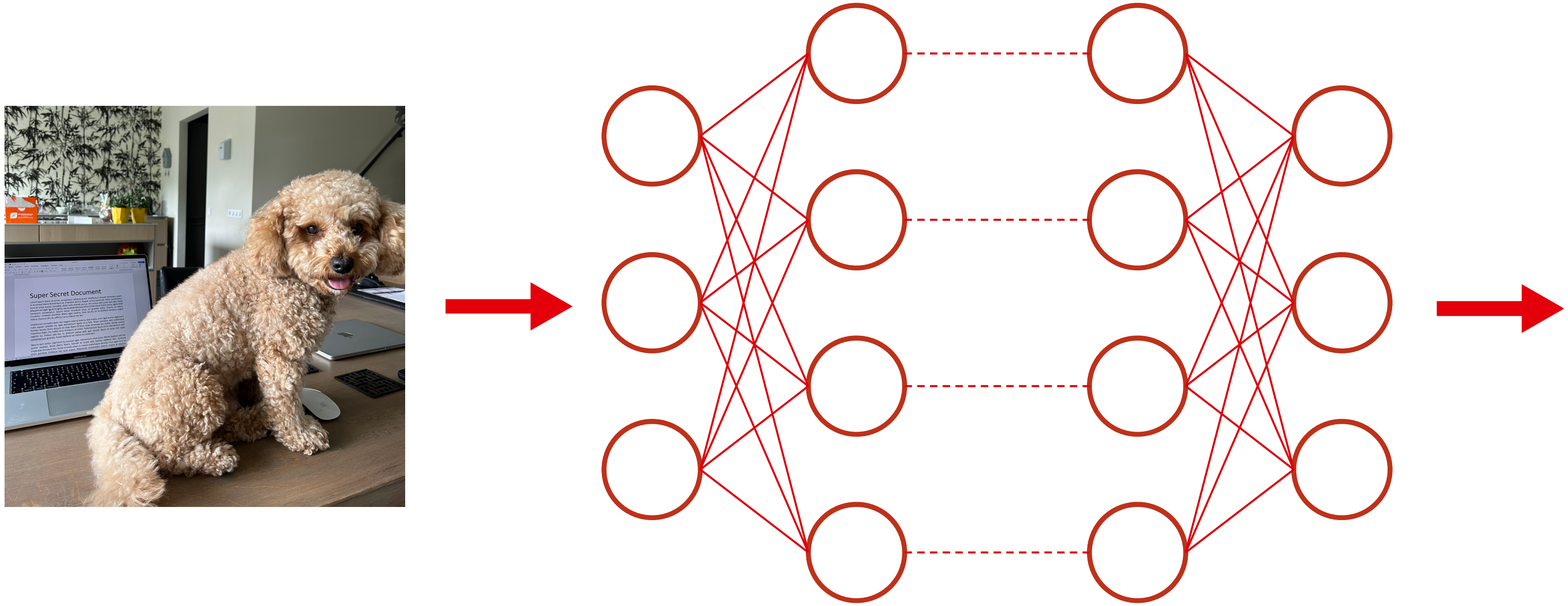
“DOG”

Output

Architecture

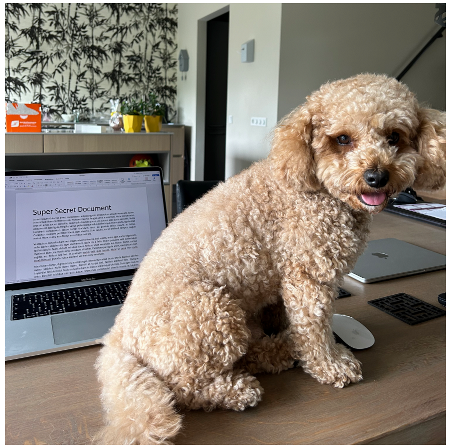
Model Stealing Attacks on DNNs

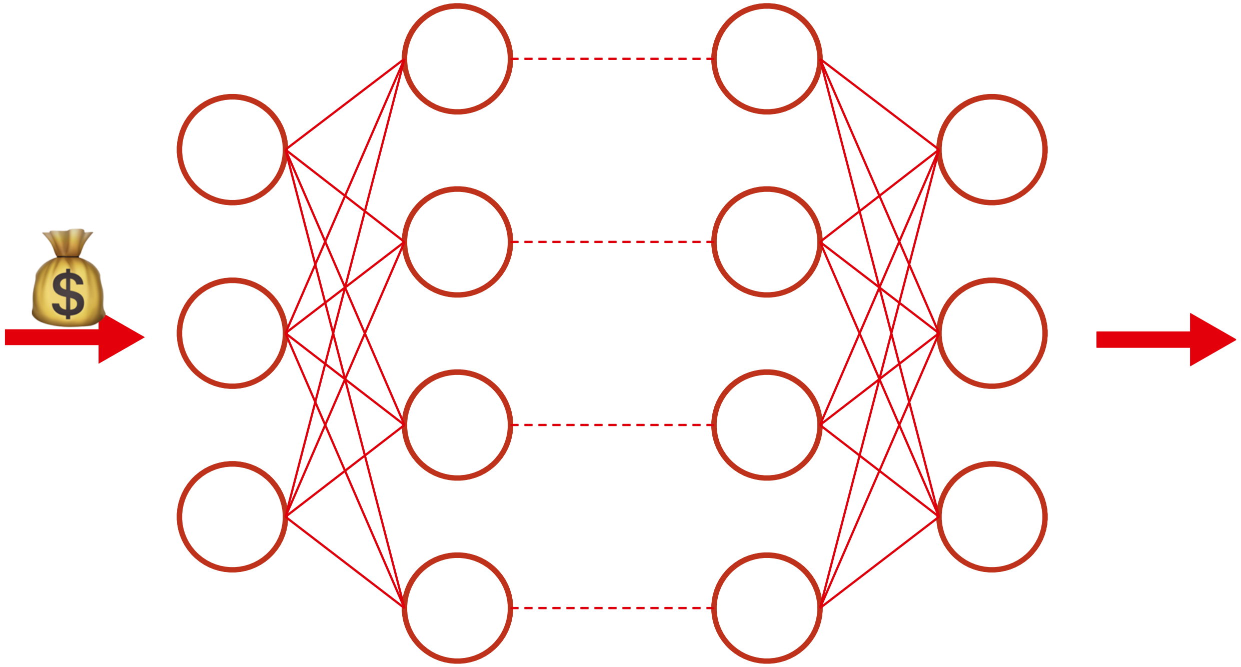
Model Stealing Attacks on DNNs



“DOG”

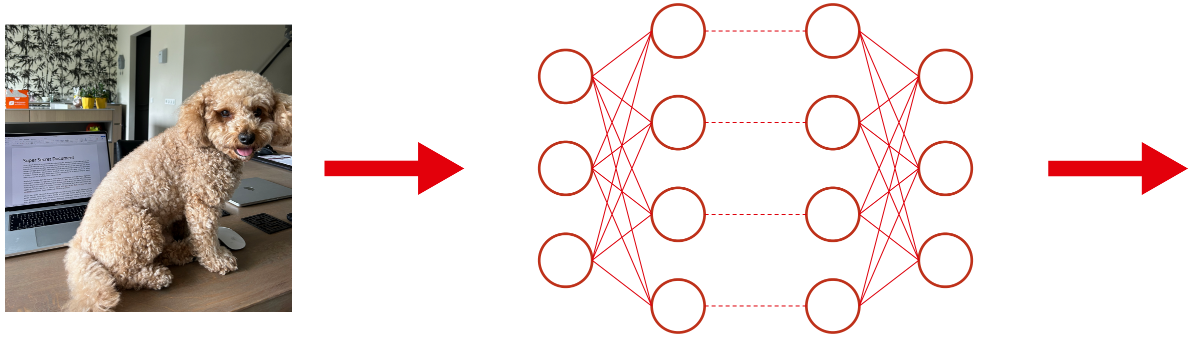
Model Stealing Attacks on DNNs



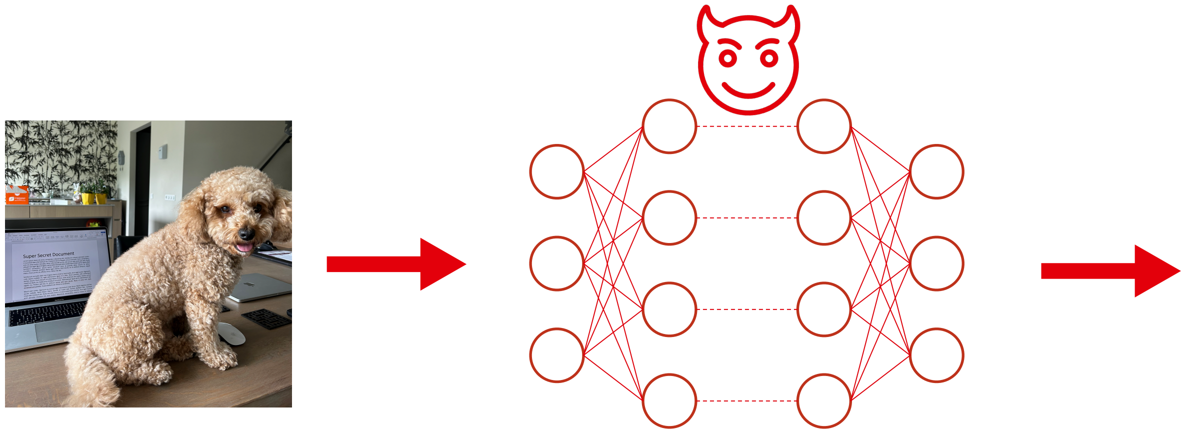


"DOG"

Model Stealing Attacks on DNNs

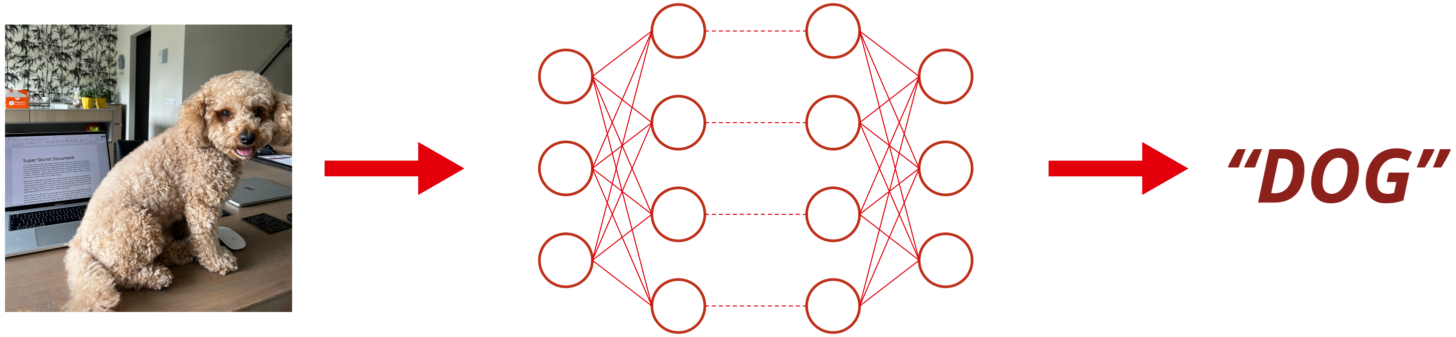


“DOG”

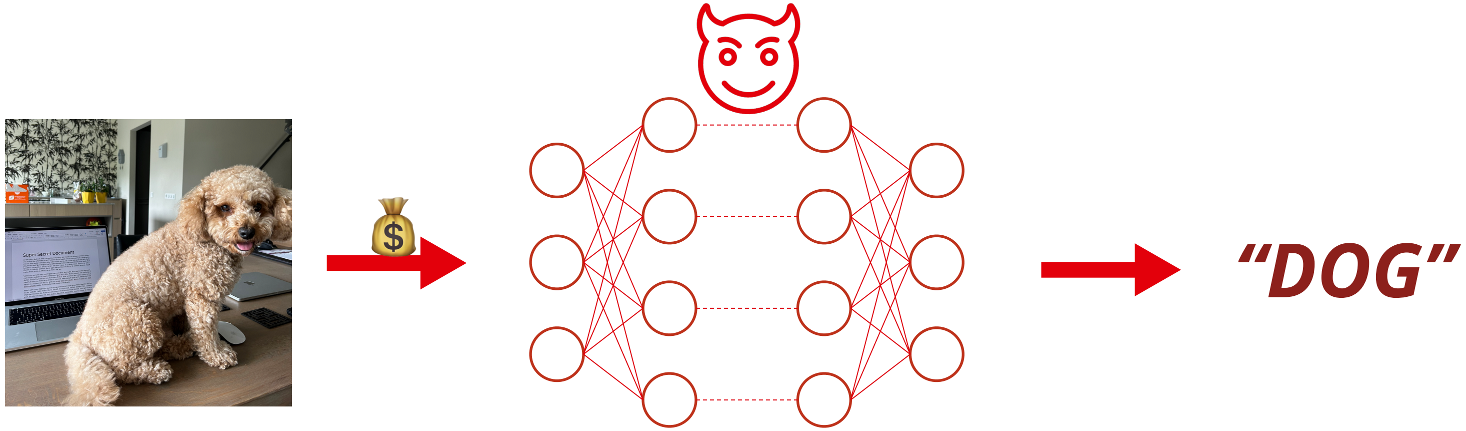


“DOG”

Model Stealing Attacks on DNNs



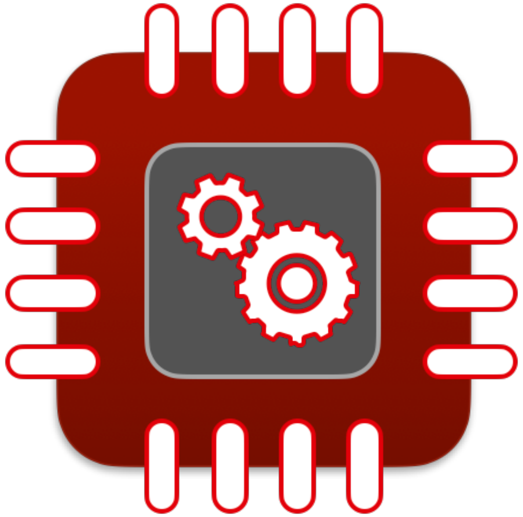
“DOG”

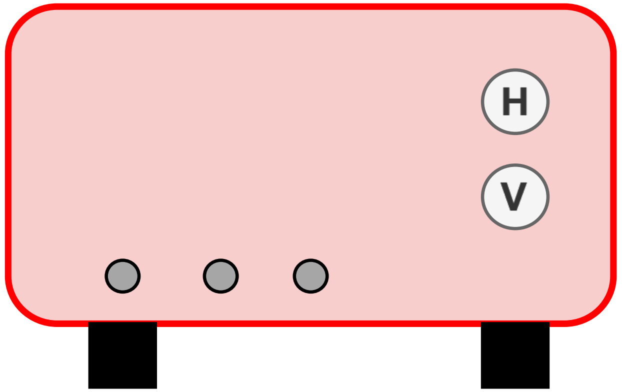


“DOG”

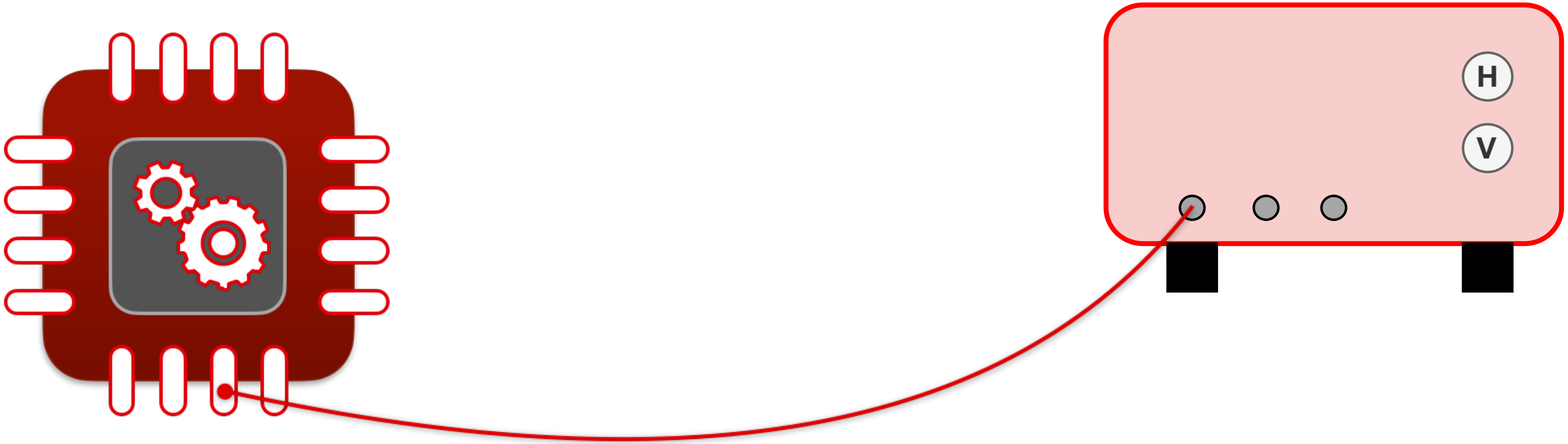
Side-Channel Analysis

Side-Channel Analysis

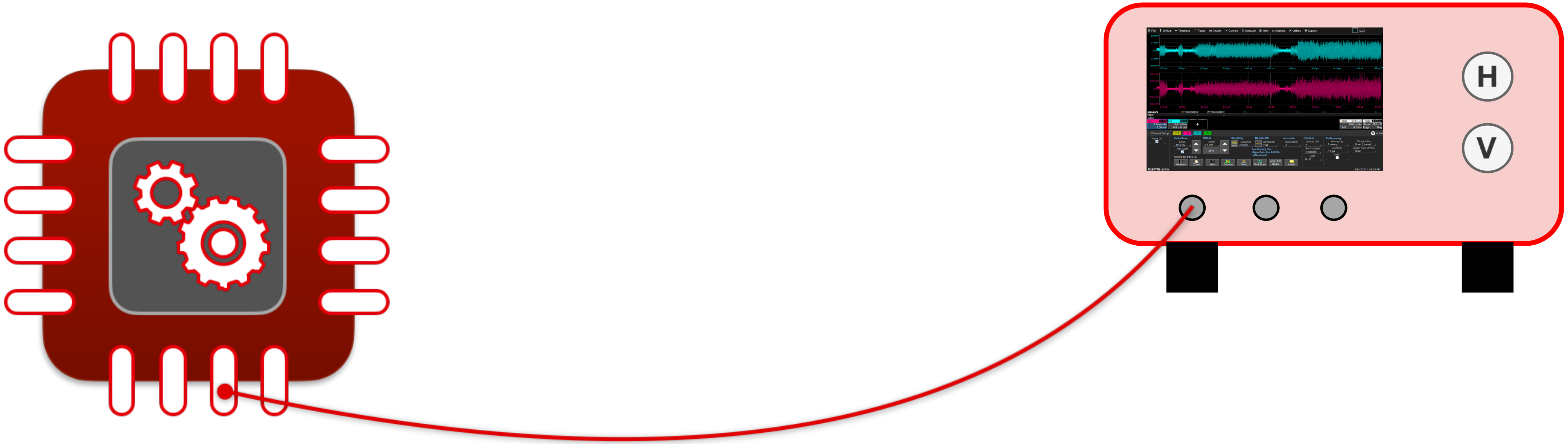




Side-Channel Analysis



Side-Channel Analysis



Agenda

Agenda

- **Neural Network Extraction**

Agenda

- **Neural Network Extraction**
  - **Systemization**

Agenda

- Neural Network Extraction
    - Systemization
- Architecture Extraction

Agenda

- Neural Network Extraction
  - Systemization
- Architecture Extraction
  - Limitations

Agenda

- Neural Network Extraction
  - Systemization
- Architecture Extraction
  - Limitations
- Parameter Extraction

Agenda

- Neural Network Extraction
  - Systemization
- Architecture Extraction
  - Limitations
- Parameter Extraction
  - Sensitivity

Agenda

- Neural Network Extraction
  - Systemization
- Architecture Extraction
  - Limitations
- Parameter Extraction
  - Sensitivity
- Input Recovery

Agenda

- **Neural Network Extraction**
  - **Systemization**
- **Architecture Extraction**
  - **Limitations**
- **Parameter Extraction**
  - **Sensitivity**
- **Input Recovery**
  - **New approaches**

Agenda

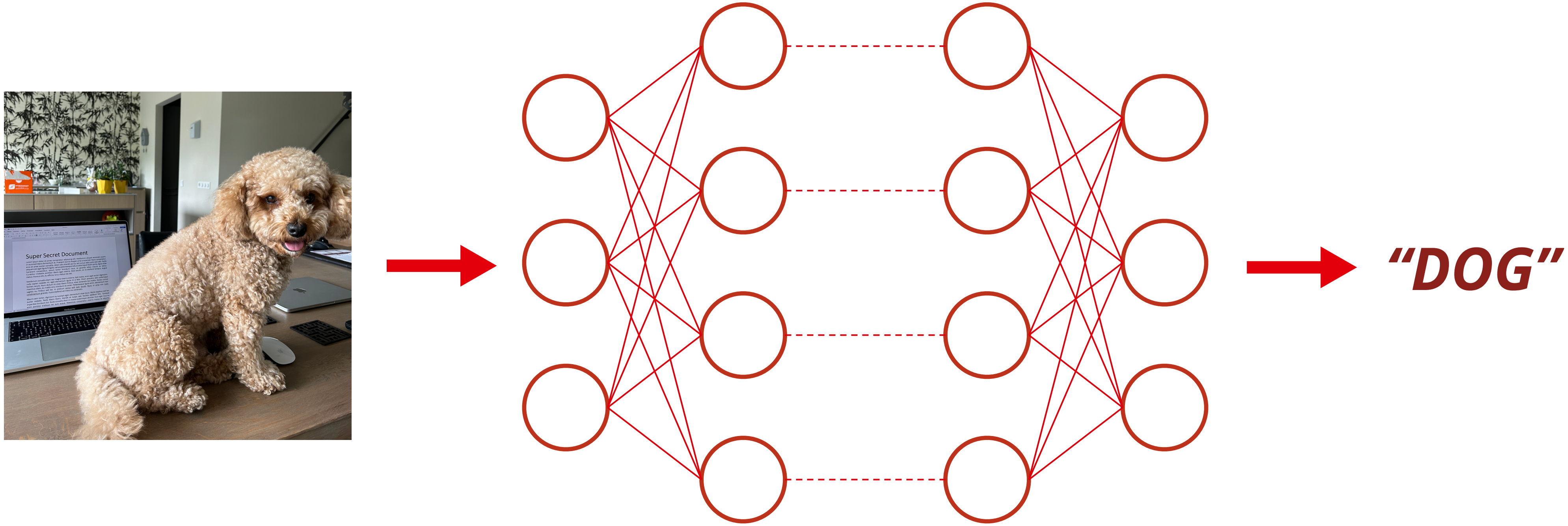
- Neural Network Extraction
  - Systemization
- Architecture Extraction
  - Limitations
- Parameter Extraction
  - Sensitivity
- Input Recovery
  - New approaches
- Outlook on Neural Network Extraction

NEURAL NETWORK EXTRACTION

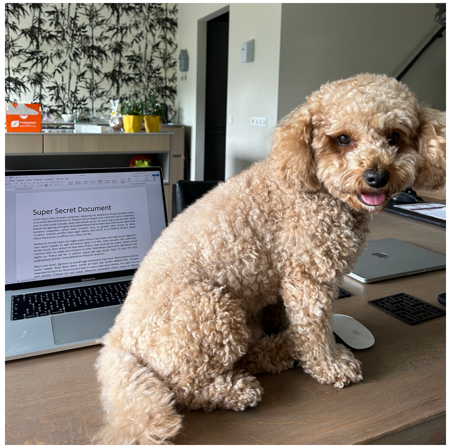
Grouping Neural Network Extraction

NEURAL NETWORK EXTRACTION

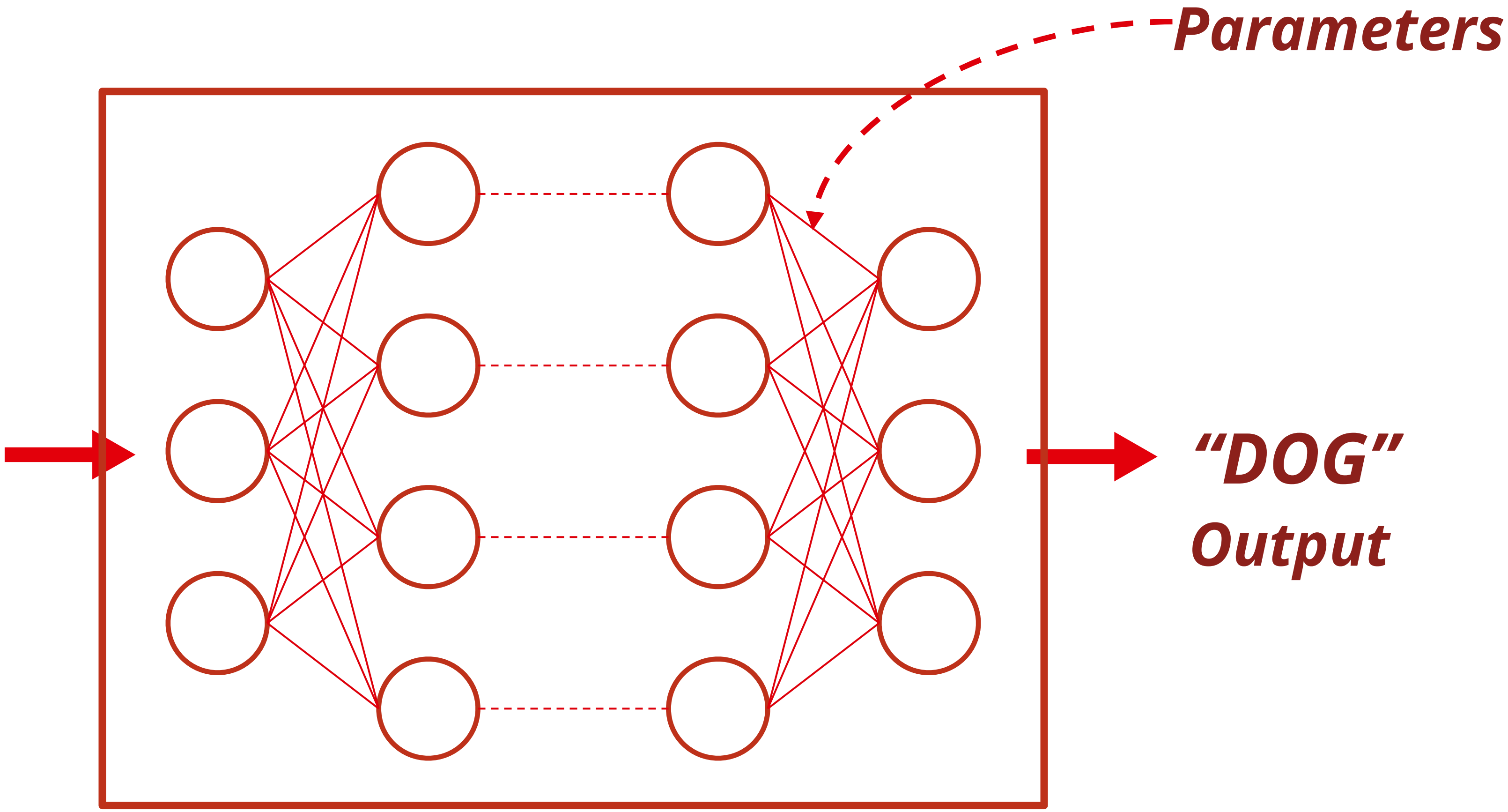
Grouping Neural Network Extraction



Grouping Neural Network Extraction



Input



Parameters

“DOG”

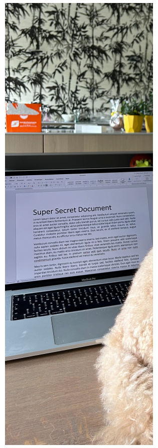
Output

Architecture

NEURAI

Gr

rs



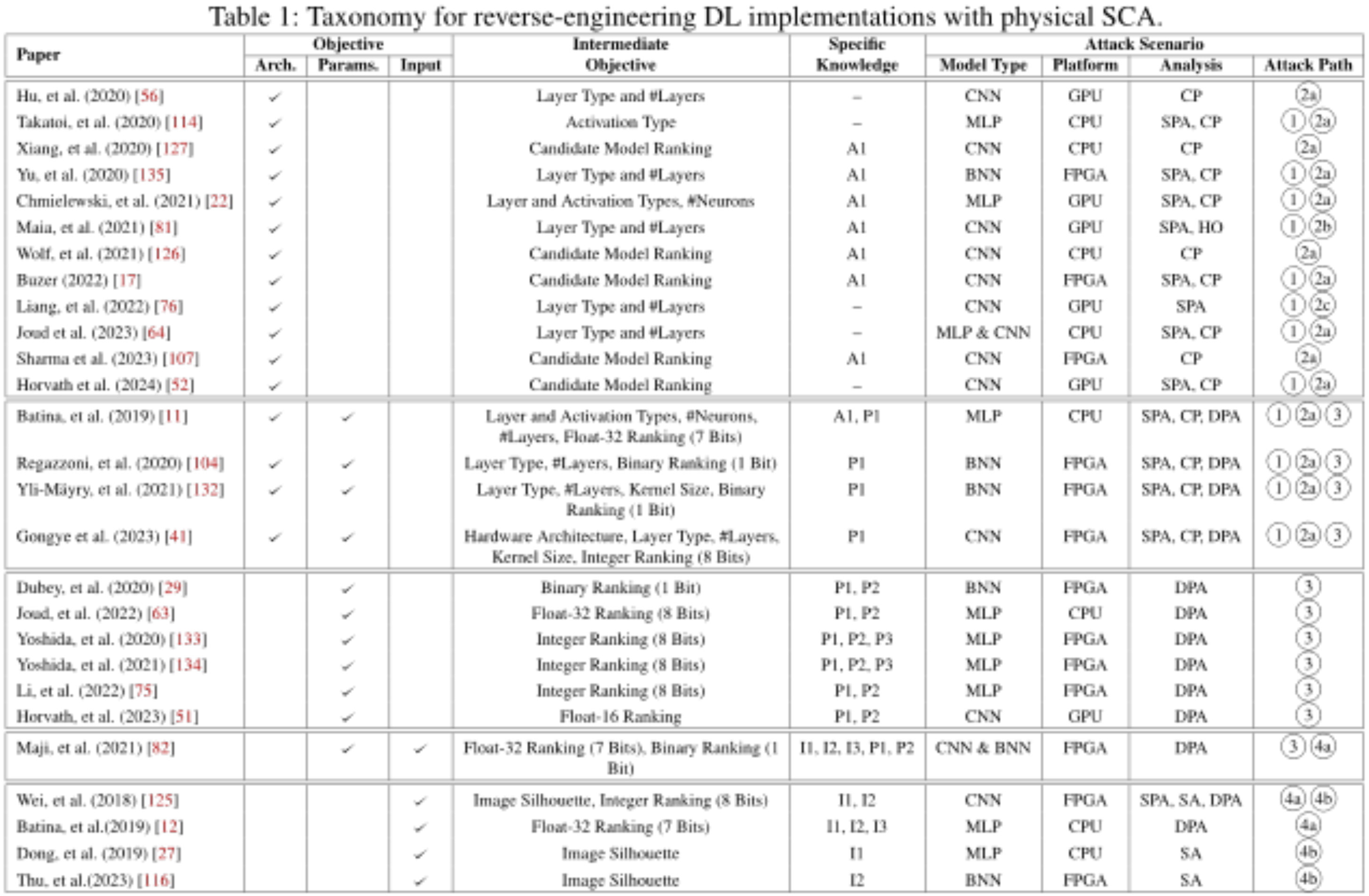


Table 1: Taxonomy for reverse-engineering DL implementations with physical SCA.

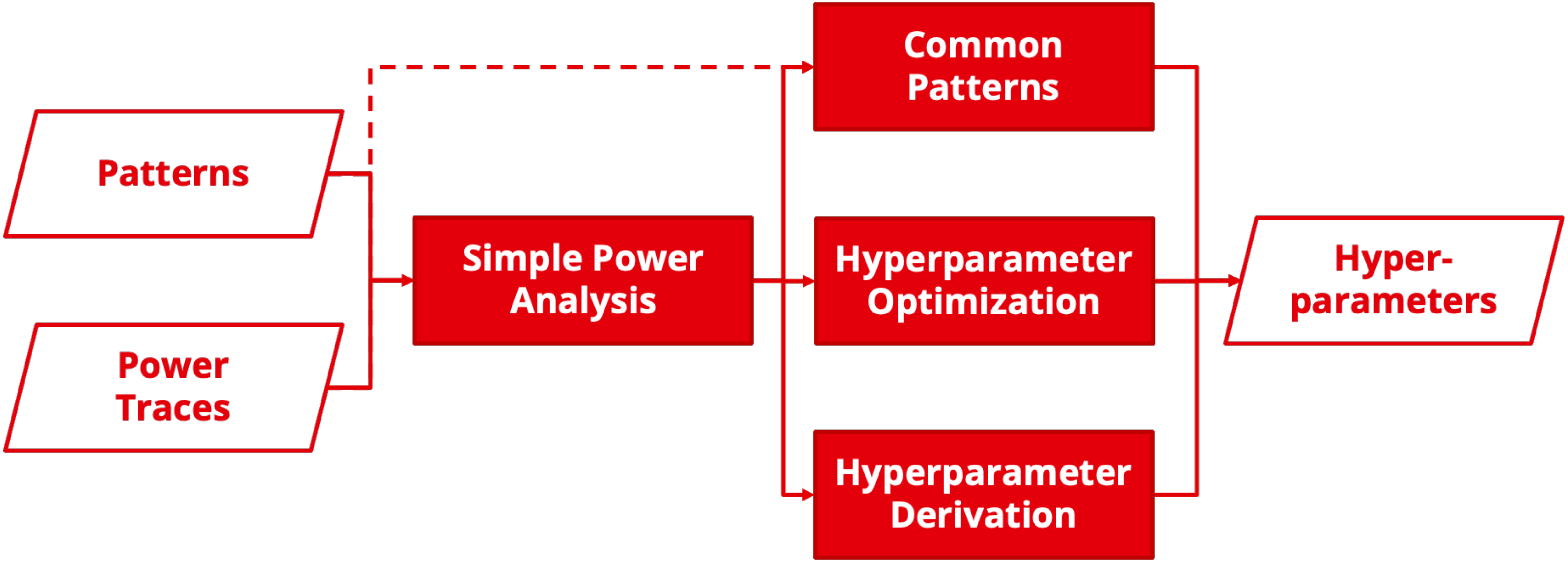
Paper	Objective			Intermediate Objective	Specific Knowledge	Attack Scenario			
	Arch.	Params.	Input			Model Type	Platform	Analysis	Attack Path
Hu, et al. (2020) [56]	✓			Layer Type and #Layers	–	CNN	GPU	CP	(2a)
Takatoi, et al. (2020) [114]	✓			Activation Type	–	MLP	CPU	SPA, CP	(1) (2a)
Xiang, et al. (2020) [127]	✓			Candidate Model Ranking	A1	CNN	CPU	CP	(2a)
Yu, et al. (2020) [135]	✓			Layer Type and #Layers	A1	BNN	FPGA	SPA, CP	(1) (2a)
Chmielewski, et al. (2021) [22]	✓			Layer and Activation Types, #Neurons	A1	MLP	GPU	SPA, CP	(1) (2a)
Maia, et al. (2021) [81]	✓			Layer Type and #Layers	A1	CNN	GPU	SPA, HO	(1) (2b)
Wolf, et al. (2021) [126]	✓			Candidate Model Ranking	A1	CNN	CPU	CP	(2a)
Buzer (2022) [17]	✓			Candidate Model Ranking	A1	CNN	FPGA	SPA, CP	(1) (2a)
Liang, et al. (2022) [76]	✓			Layer Type and #Layers	–	CNN	GPU	SPA	(1) (2c)
Joud et al. (2023) [64]	✓			Layer Type and #Layers	–	MLP & CNN	CPU	SPA, CP	(1) (2a)
Sharma et al. (2023) [107]	✓			Candidate Model Ranking	A1	CNN	FPGA	CP	(2a)
Horvath et al. (2024) [52]	✓			Candidate Model Ranking	–	CNN	GPU	SPA, CP	(1) (2a)
Batina, et al. (2019) [11]	✓	✓		Layer and Activation Types, #Neurons, #Layers, Float-32 Ranking (7 Bits)	A1, P1	MLP	CPU	SPA, CP, DPA	(1) (2a) (3)
Regazzoni, et al. (2020) [104]	✓	✓		Layer Type, #Layers, Binary Ranking (1 Bit)	P1	BNN	FPGA	SPA, CP, DPA	(1) (2a) (3)
Yli-Mäyry, et al. (2021) [132]	✓	✓		Layer Type, #Layers, Kernel Size, Binary Ranking (1 Bit)	P1	BNN	FPGA	SPA, CP, DPA	(1) (2a) (3)
Gongye et al. (2023) [41]	✓	✓		Hardware Architecture, Layer Type, #Layers, Kernel Size, Integer Ranking (8 Bits)	P1	CNN	FPGA	SPA, CP, DPA	(1) (2a) (3)
Dubey, et al. (2020) [29]		✓		Binary Ranking (1 Bit)	P1, P2	BNN	FPGA	DPA	(3)
Joud, et al. (2022) [63]		✓		Float-32 Ranking (8 Bits)	P1, P2	MLP	CPU	DPA	(3)
Yoshida, et al. (2020) [133]		✓		Integer Ranking (8 Bits)	P1, P2, P3	MLP	FPGA	DPA	(3)
Yoshida, et al. (2021) [134]		✓		Integer Ranking (8 Bits)	P1, P2, P3	MLP	FPGA	DPA	(3)
Li, et al. (2022) [75]		✓		Integer Ranking (8 Bits)	P1, P2	MLP	FPGA	DPA	(3)
Horvath, et al. (2023) [51]		✓		Float-16 Ranking	P1, P2	CNN	GPU	DPA	(3)
Maji, et al. (2021) [82]		✓	✓	Float-32 Ranking (7 Bits), Binary Ranking (1 Bit)	I1, I2, I3, P1, P2	CNN & BNN	FPGA	DPA	(3) (4a)
Wei, et al. (2018) [125]			✓	Image Silhouette, Integer Ranking (8 Bits)	I1, I2	CNN	FPGA	SPA, SA, DPA	(4a) (4b)
Batina, et al.(2019) [12]			✓	Float-32 Ranking (7 Bits)	I1, I2, I3	MLP	CPU	DPA	(4a)
Dong, et al. (2019) [27]			✓	Image Silhouette	I1	MLP	CPU	SA	(4b)
Thu, et al.(2023) [116]			✓	Image Silhouette	I2	BNN	FPGA	SA	(4b)

/

Architecture

ARCHITECTURE EXTRACTION

Framework


Patterns
Power Traces
Simple Power Analysis
Common Patterns
Hyperparameter Optimization
Hyperparameter Derivation
Hyper-parameters

ARCHITECTURE EXTRACTIONS

Limitations

ARCHITECTURE EXTRACTIONS

Limitations

- **Only 4 activation functions**

ARCHITECTURE EXTRACTIONS

Limitations

- **Only 4 activation functions**
- **Only limited layer types**

Limitations

- **Only 4 activation functions**
- **Only limited layer types**
- **The stride is 1 or 2**

ARCHITECTURE EXTRACTIONS

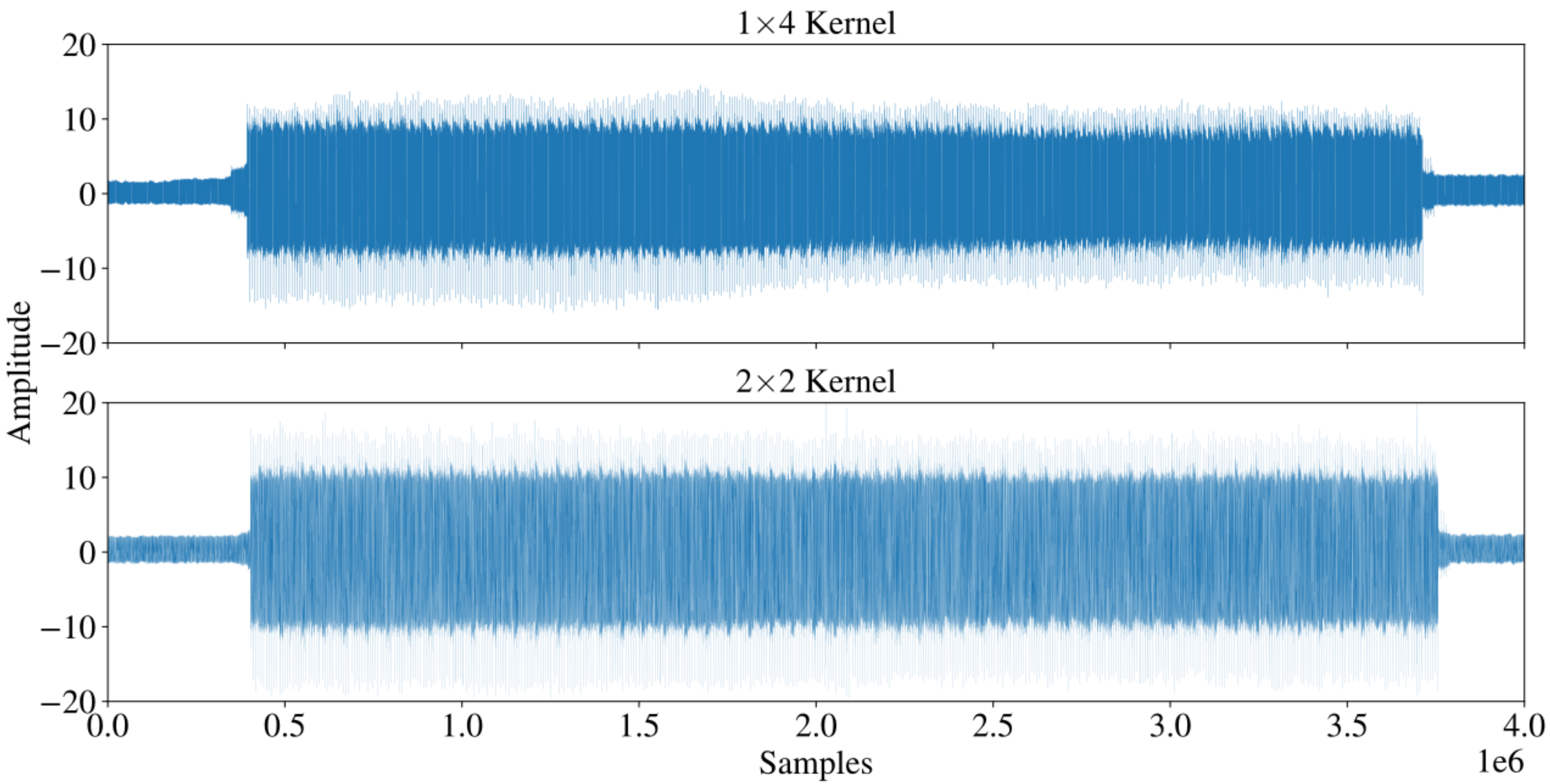
Limitations

- Only 4 activation functions
- Only limited layer types
- The stride is 1 or 2
- The kernel is a square

ARCHITECTURE EXTRACTIONS

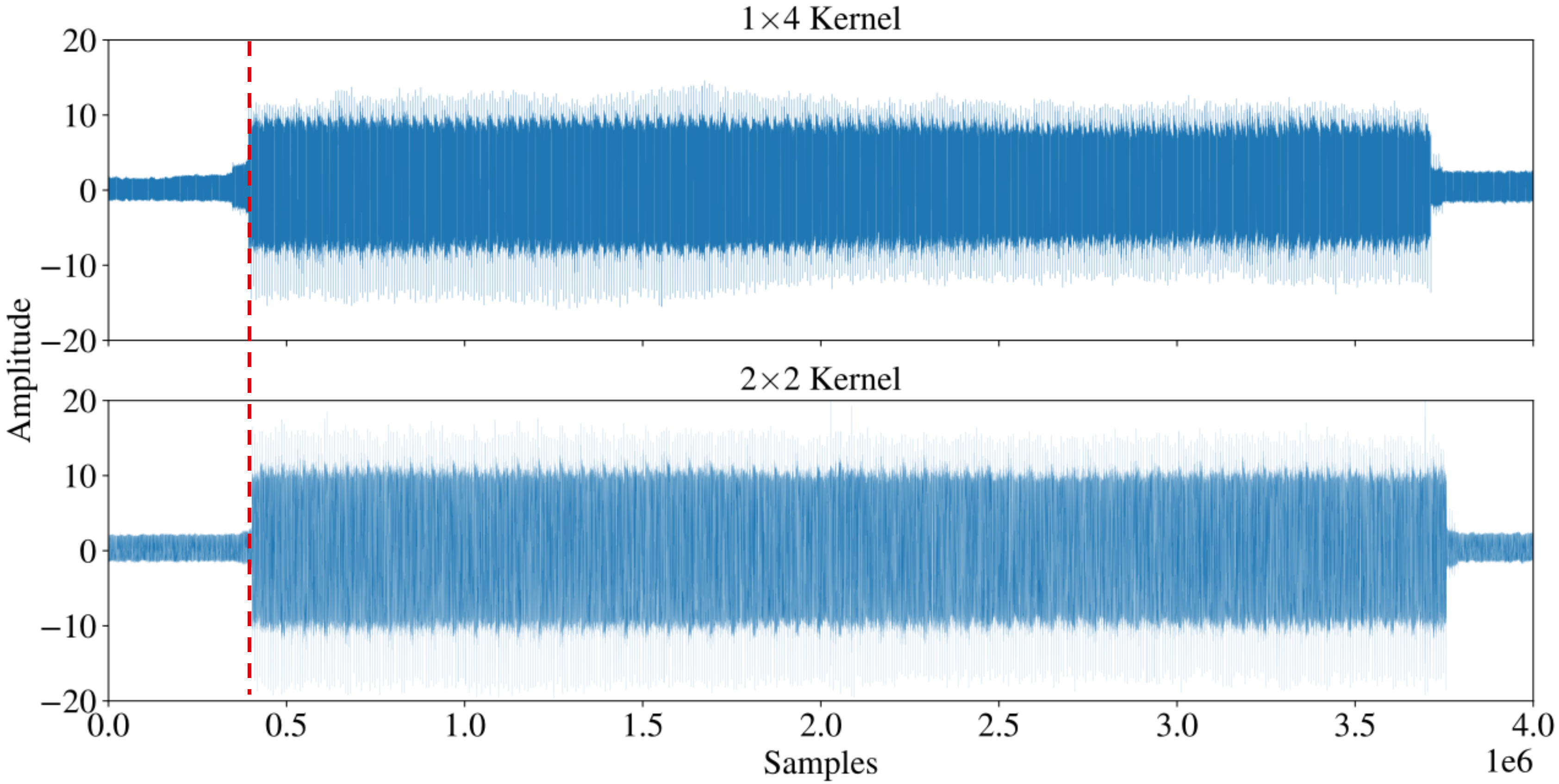
Limitations

- Only 4
- Only li
- The str
- The ke



Limitations

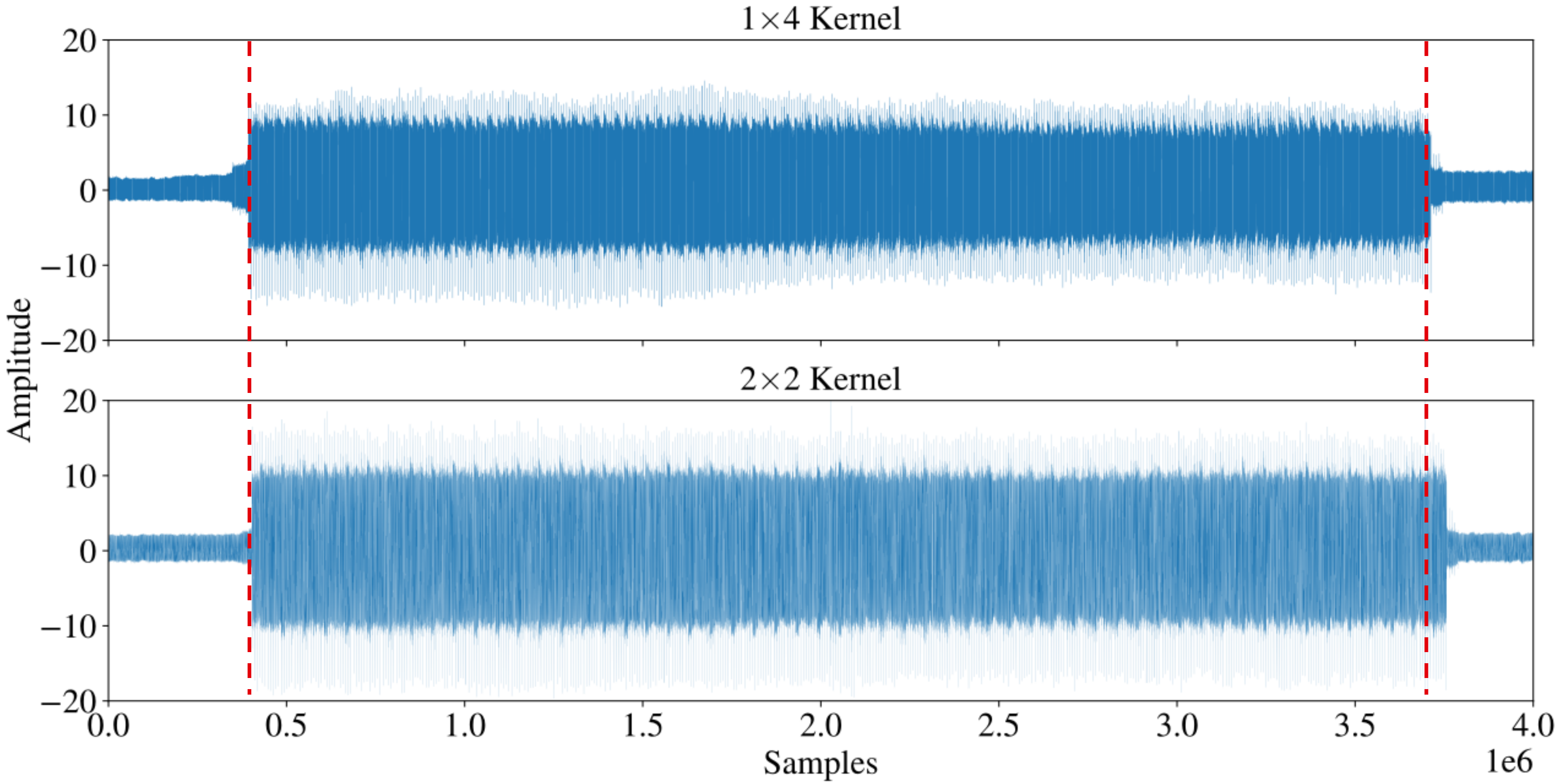
- **Only 4**
- **Only li**
- **The str**
- **The ke**



ARCHITECTURE EXTRACTIONS

Limitations

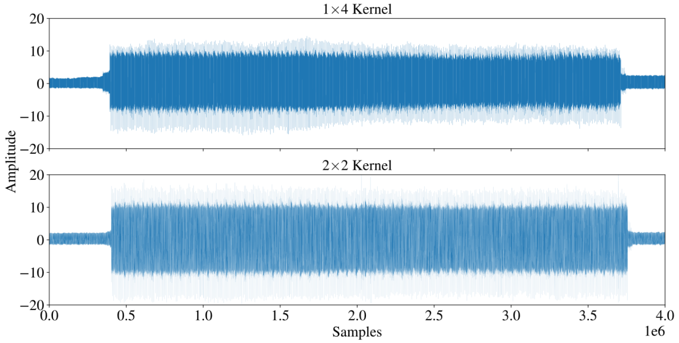
- Only 4
- Only li
- The str
- The ke



ARCHITECTURE EXTRACTIONS

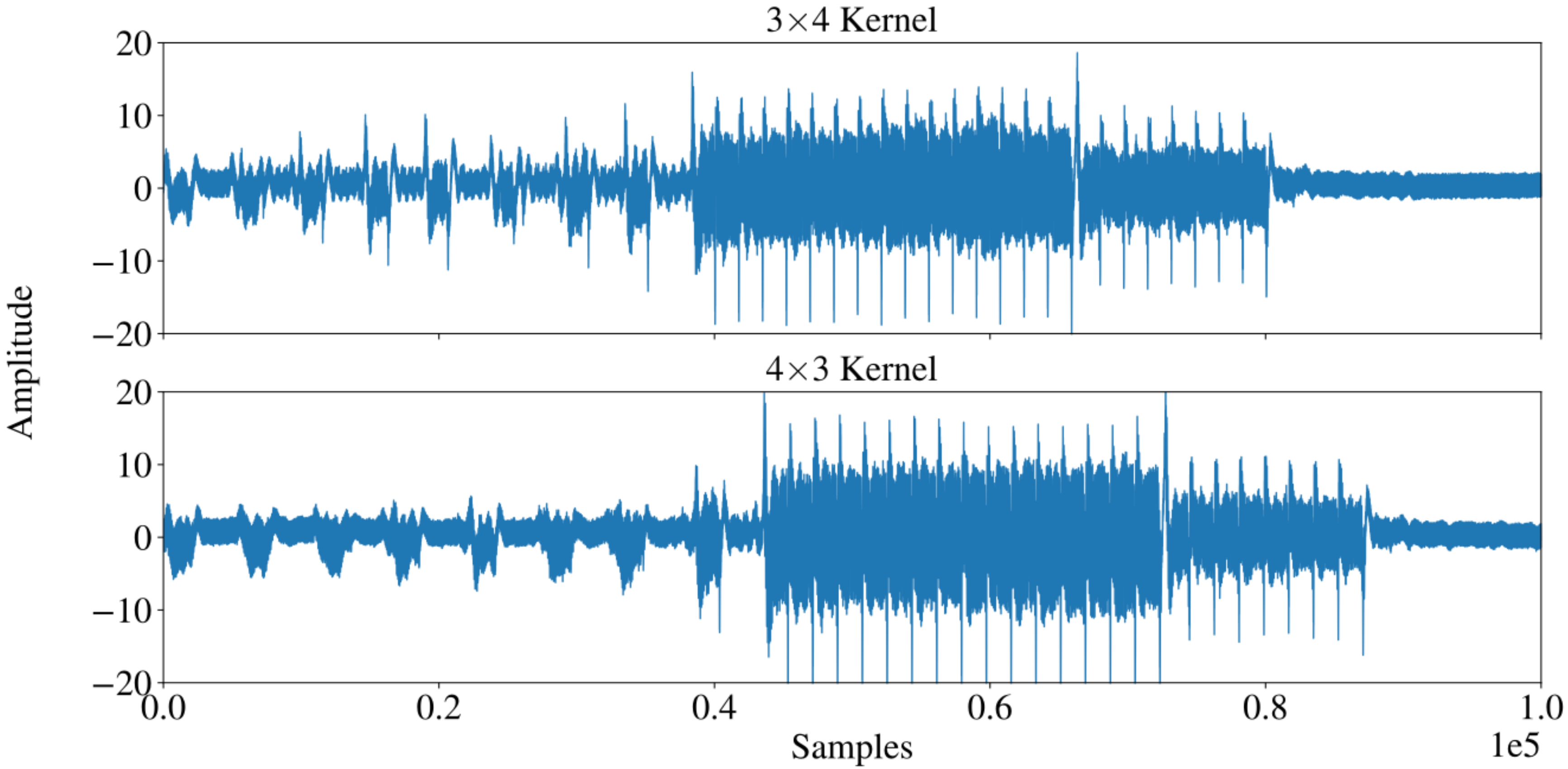
Limitations

- Only 4 activation functions
- Only limited layer types
- The stride is 1 or 2
- The kernel is a square



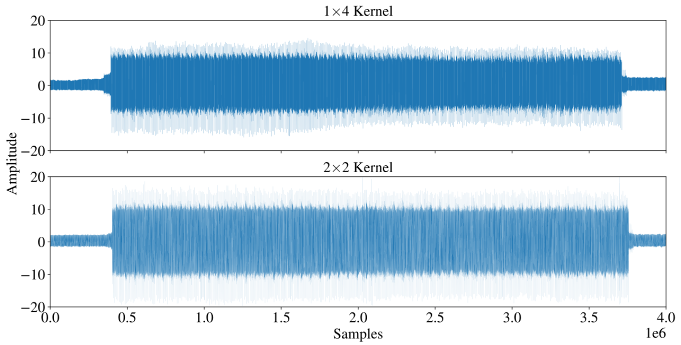
Limitations

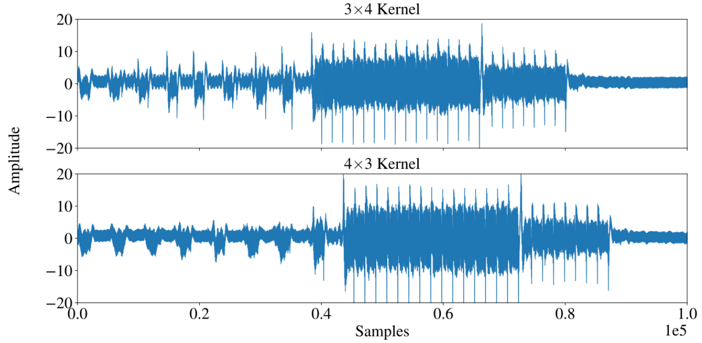
- Only 4
- Only l
- The st
- The k



Limitations

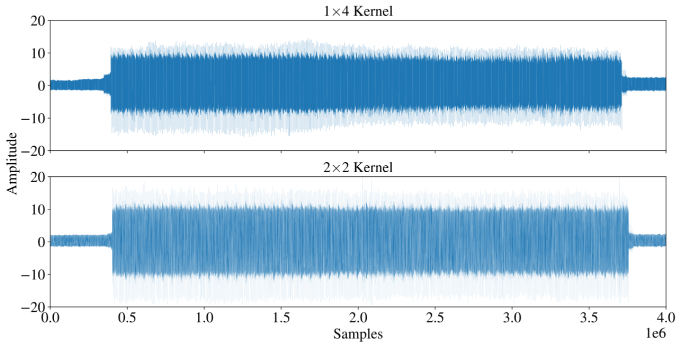
- Only 4 activation functions
- Only limited layer types
- The stride is 1 or 2
- The kernel is a square

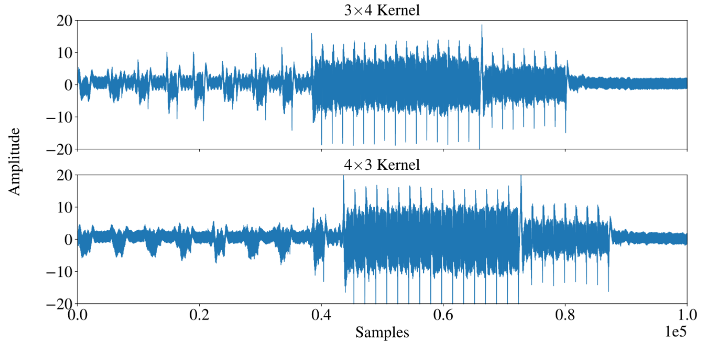




Limitations

- Only 4 activation functions
- Only limited layer types
- The stride is 1 or 2
- The kernel is a square

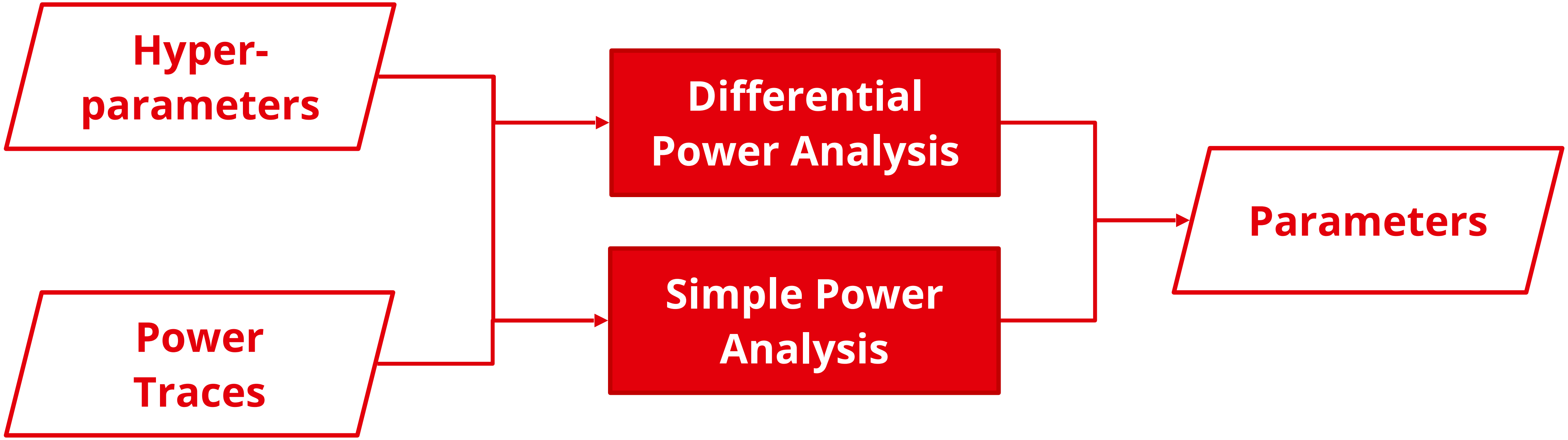




ASSUMPTIONS CAN BE RELAXED

PARAMETER EXTRACTION

Framework



Hyper-parameters

Power Traces

Differential Power Analysis

Simple Power Analysis

Parameters

PARAMETER EXTRACTION

Challenges

PARAMETER EXTRACTION

Challenges

- **Parameter Extraction can be expensive**

Challenges

- Parameter Extraction can be expensive
- Input may not always be known

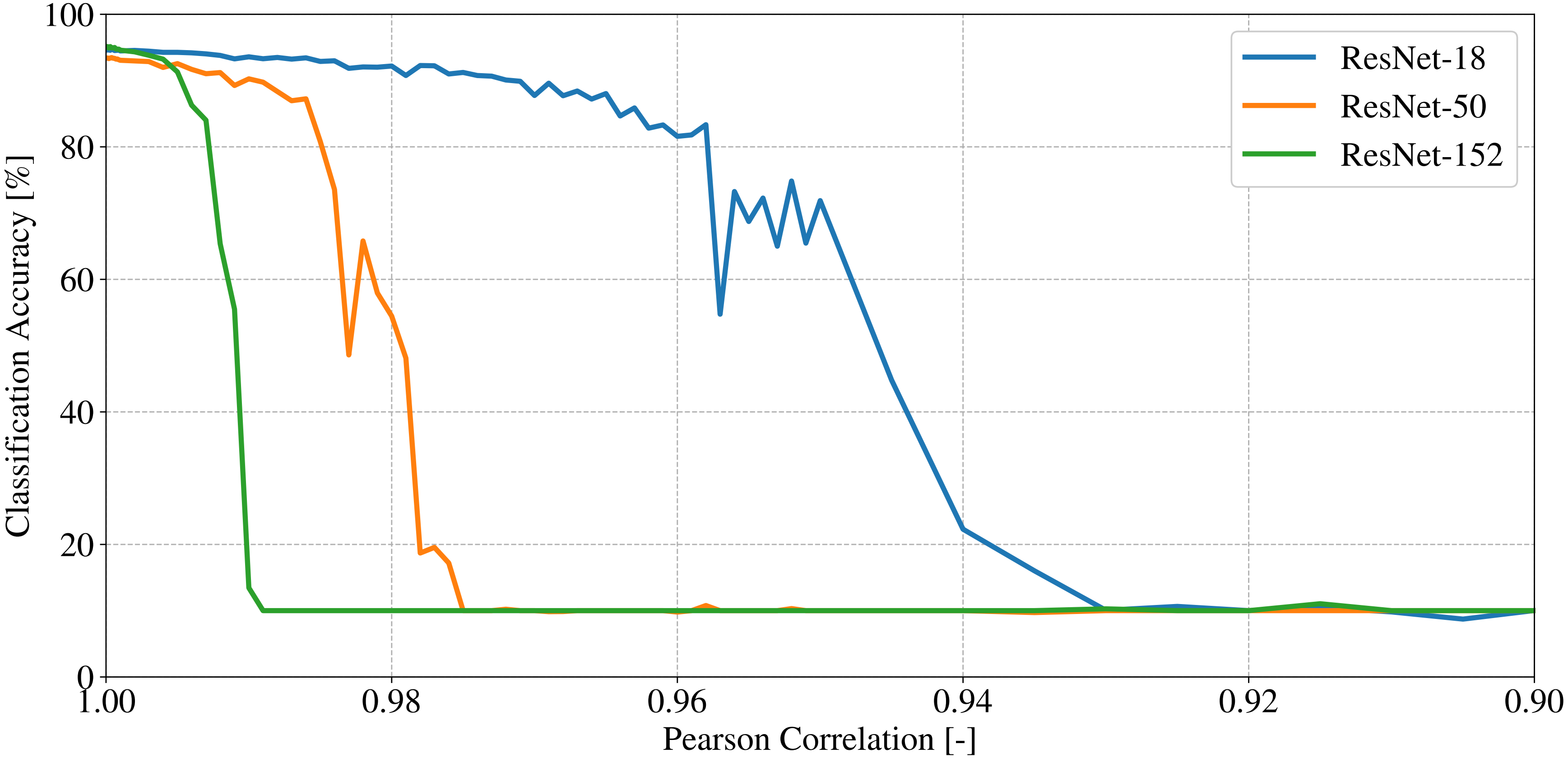
Challenges

- **Parameter Extraction can be expensive**
- **Input may not always be known**
- **Small errors in weight recovery can add up**

PARAMETER EXTRACTION

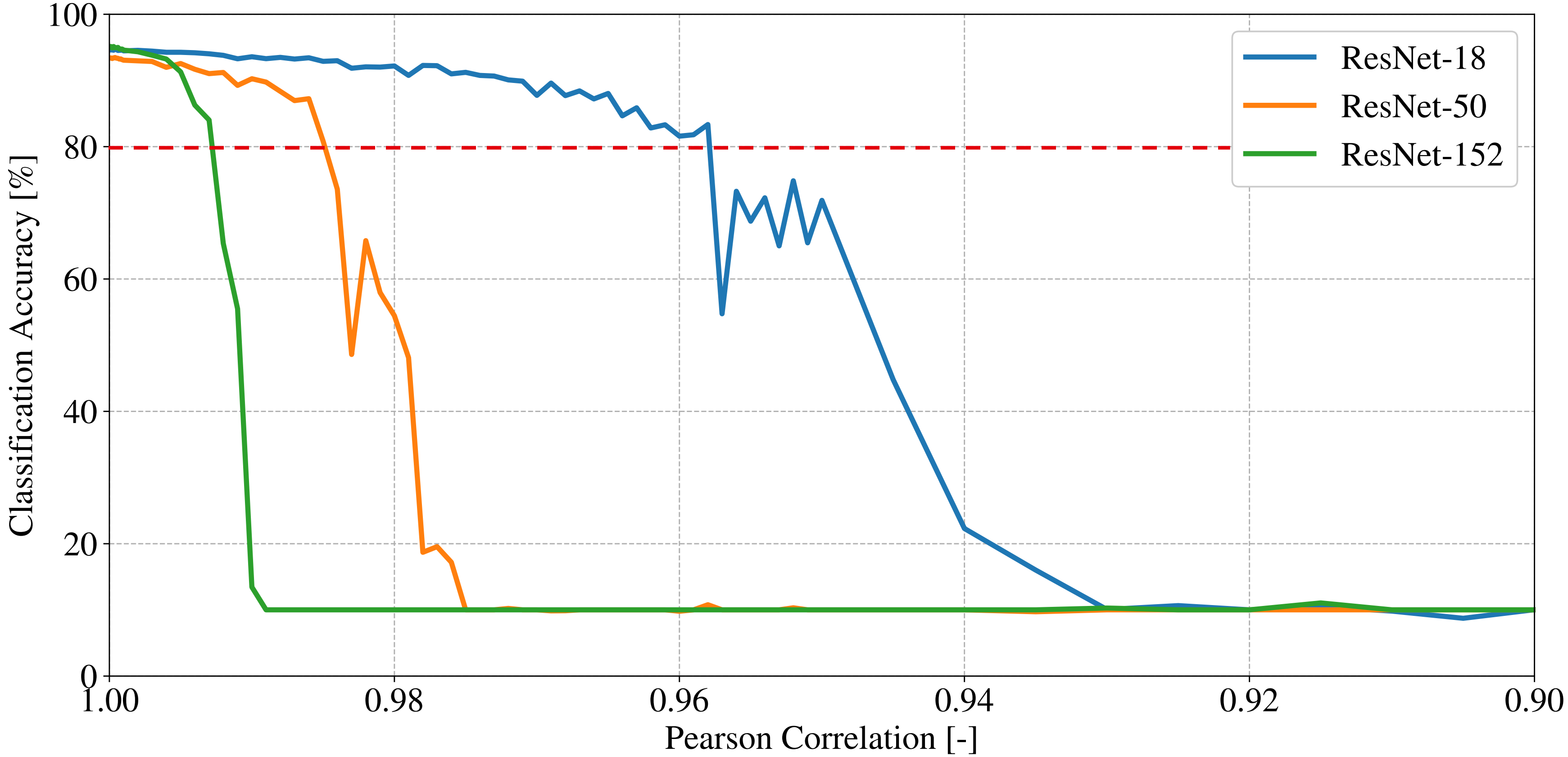
Challenges

- **Paran**
- **Input**
- **Small**



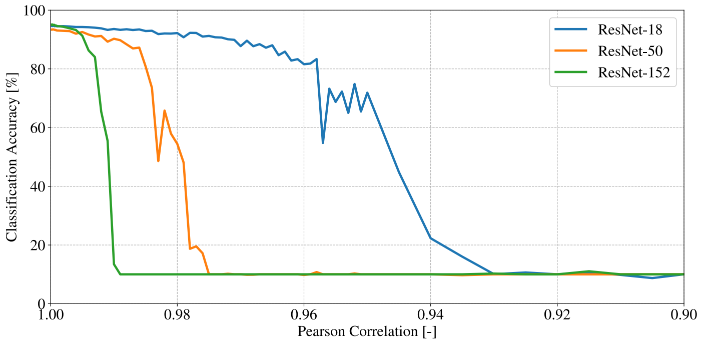
Challenges

- **Paran**
- **Input**
- **Small**



Challenges

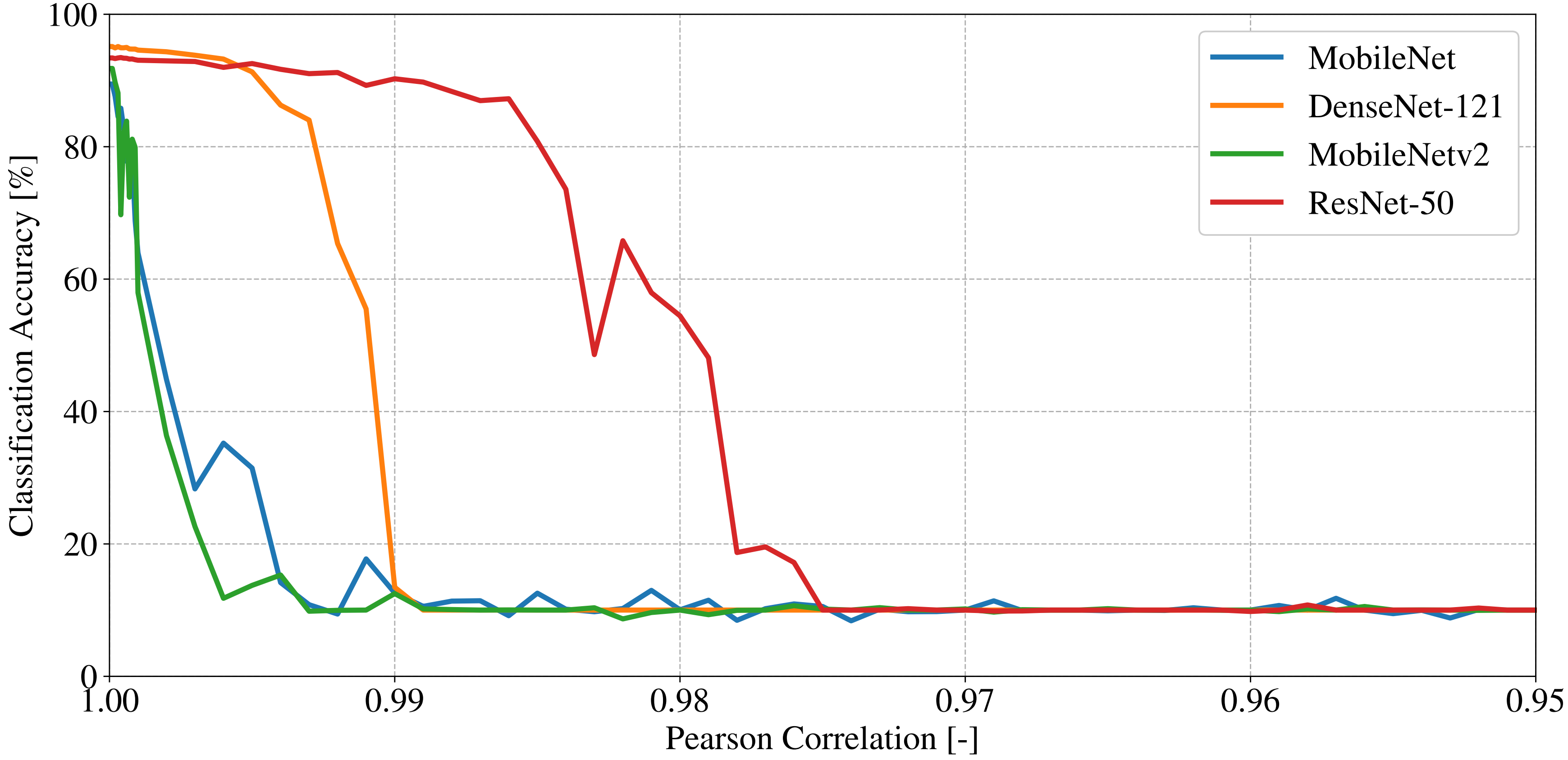
- **Parameter Extraction can be expensive**
- **Input may not always be known**
- **Small errors in weight recovery can add up**



PARAMETER EXTRACTION

Challenges

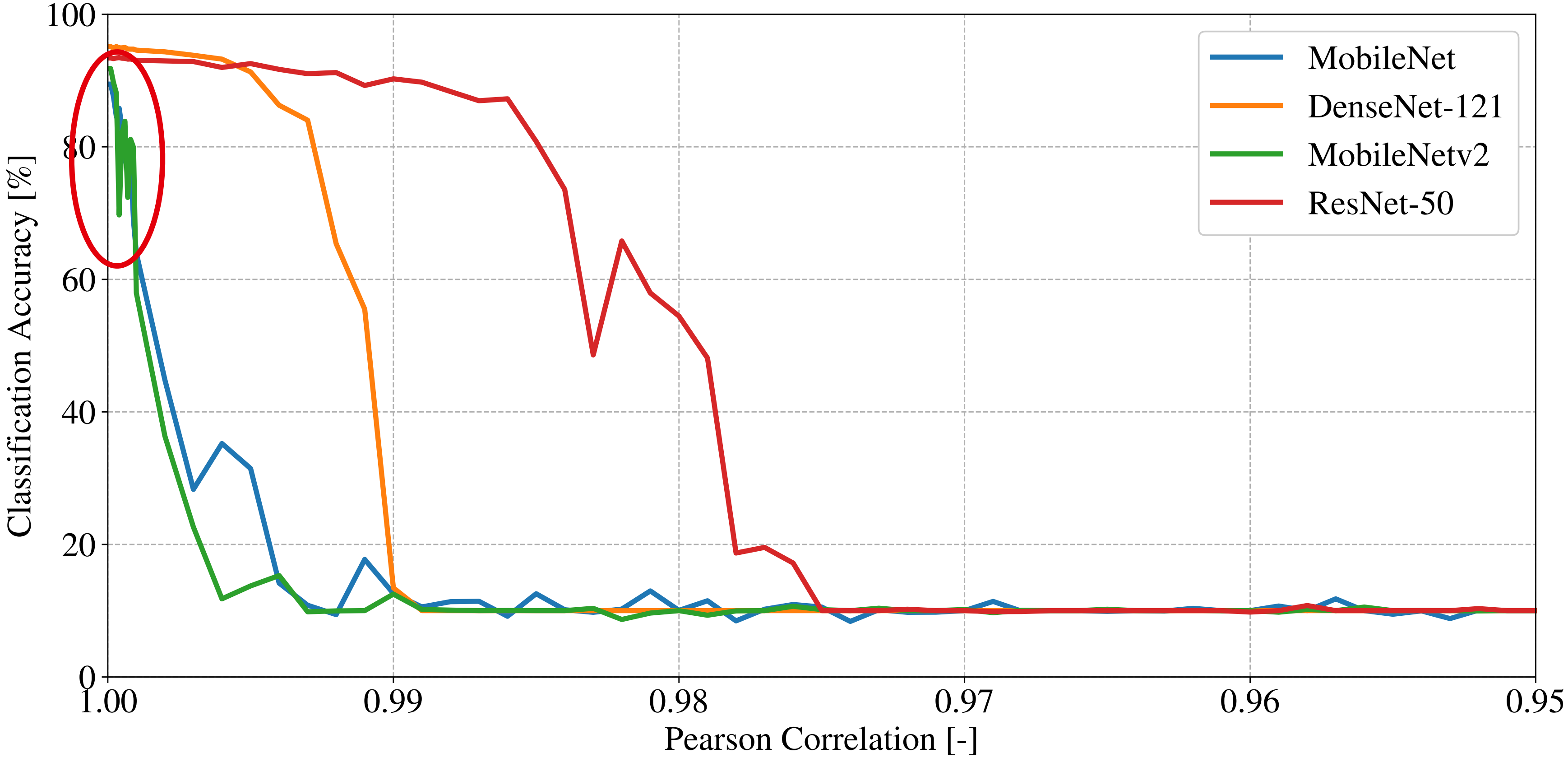
- Paran
- Input
- Small



PARAMETER EXTRACTION

Challenges

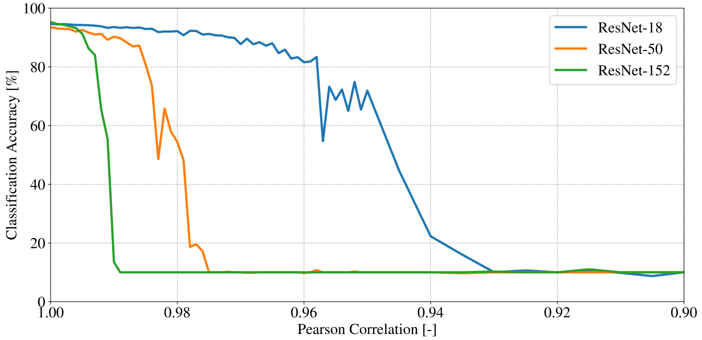
- **Paran**
- **Input**
- **Small**

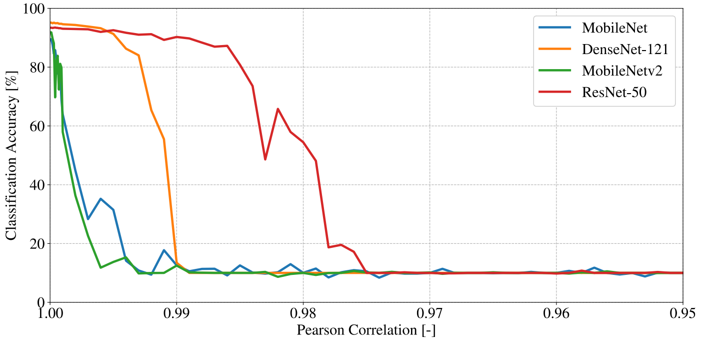


PARAMETER EXTRACTION

Challenges

- **Parameter Extraction can be expensive**
- **Input may not always be known**
- **Small errors in weight recovery can add up**

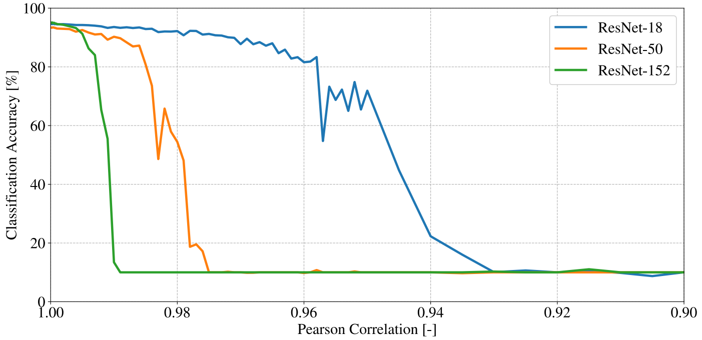


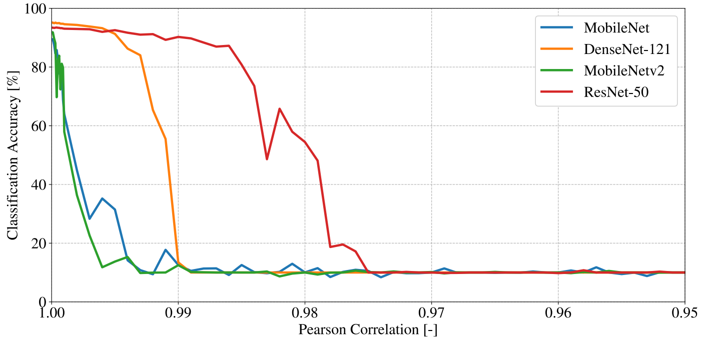


PARAMETER EXTRACTION

Challenges

- **Parameter Extraction can be expensive**
- **Input may not always be known**
- **Small errors in weight recovery can add up**

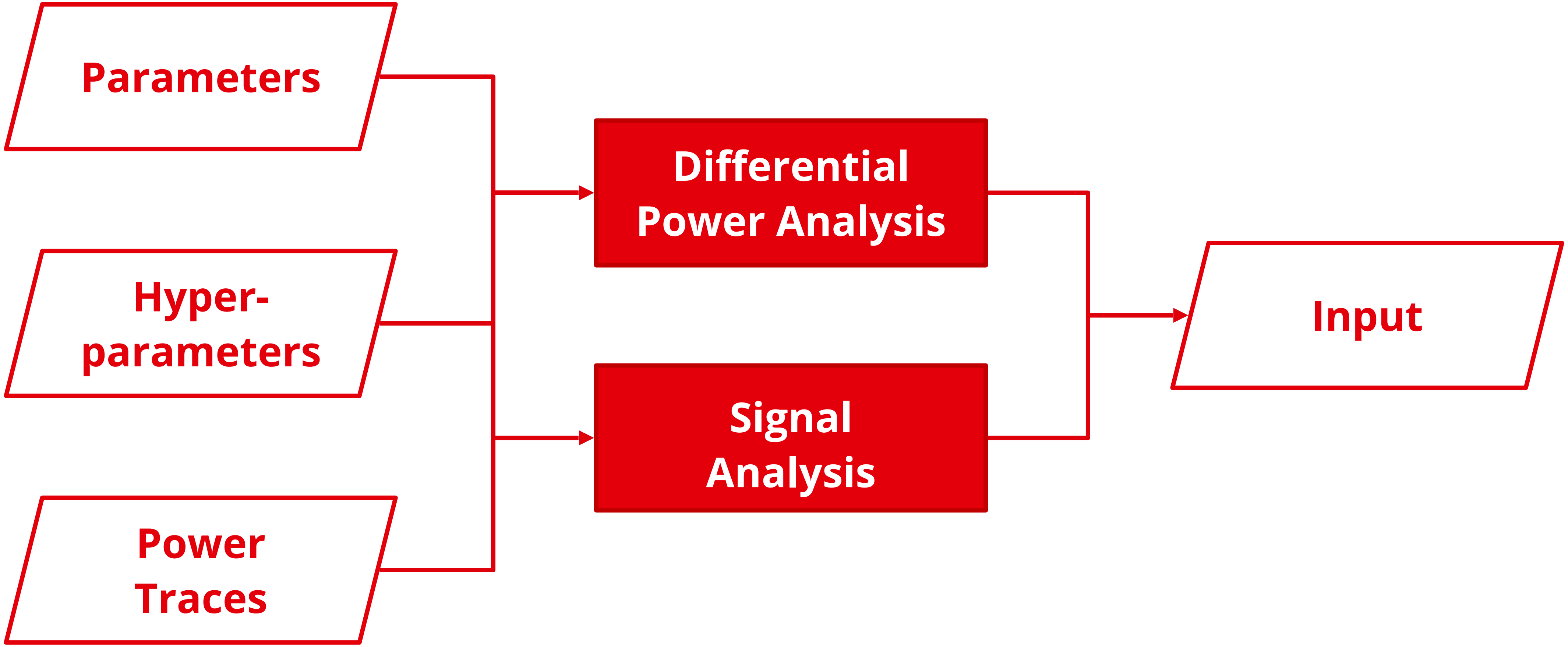




THE EFFECTIVENESS OF DPA DEPENDS BOTH ON THE MODEL AND THE EXTRACTION ACCURACY

INPUT RECOVERY

Framework


Parameters
Hyper-parameters
Power Traces
Differential Power Analysis
Signal Analysis
Input

Challenges of Input Recovery

INPUT RECOVERY

Challenges of Input Recovery

- **One-shot scenario limits the attacker**

Challenges of Input Recovery

- One-shot scenario limits the attacker
- Some attacks require whitebox access to the model

Challenges of Input Recovery

- One-shot scenario limits the attacker
- Some attacks require whitebox access to the model
- Research focuses only on grayscale images

Challenges of Input Recovery

- **One-shot scenario limits the attacker**
- **Some attacks require whitebox access to the model**
- **Research focuses only on grayscale images**
- **Full input reconstruction may not be necessary**

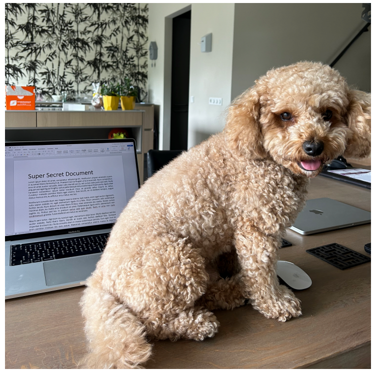
Attribute Extraction

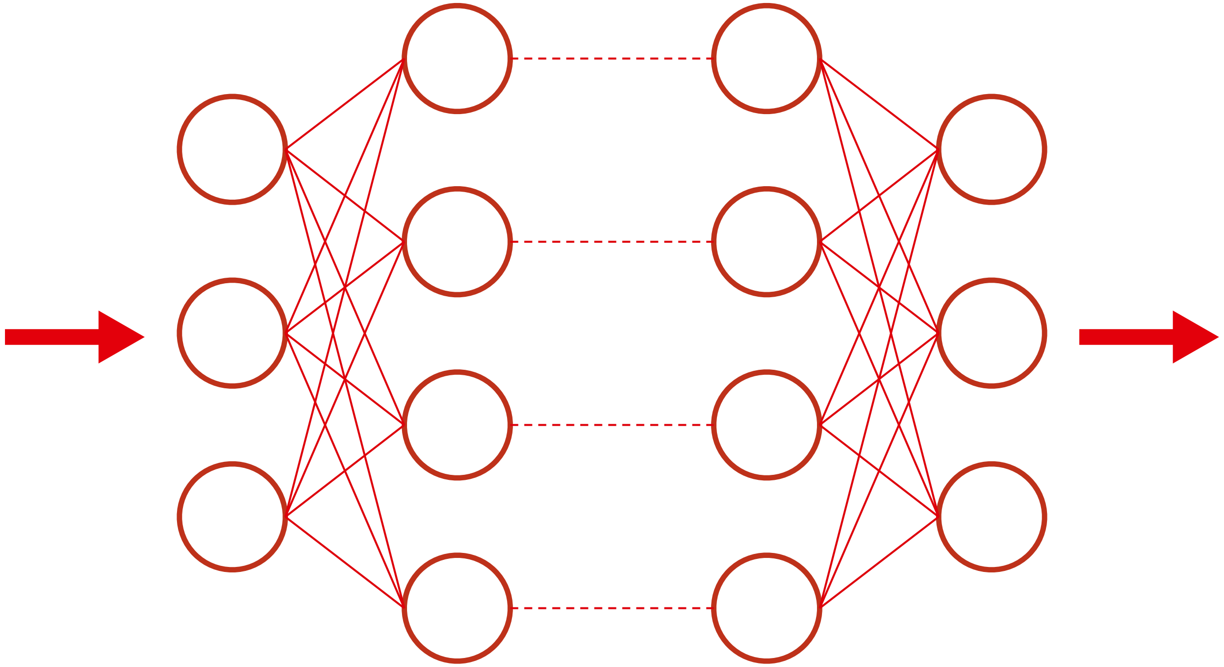
INPUT RECOVERY

Attribute Extraction

INPUT RECOVERY

Attribute Extraction





"DOG"

Super Secret Document

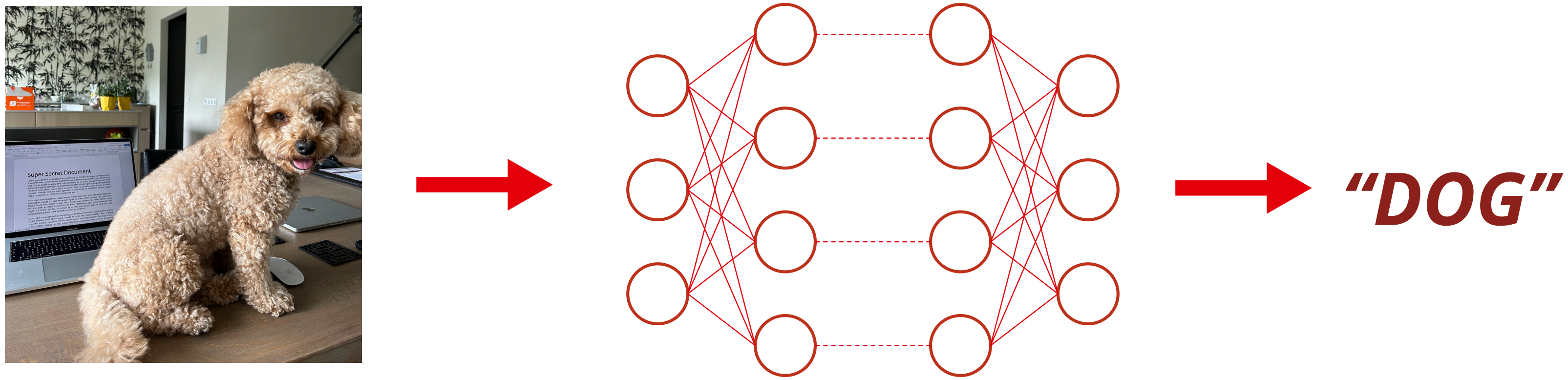
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vestibulum aliquet venenatis quam, in euismod libero fermentum at. Praesent lacinia feugiat urna a euismod. Nunc consectetur, eros sit amet auctor convallis, dolor odio blandit purus, vel cursus odio justo sed velit. Nulla aliquam est eget ligula fringilla, varius pellentesque tellus scelerisque. Etiam porta, ligula vitae hendrerit consectetur, ipsum tortor tincidunt risus, ac gravida tellus mauris ac tellus. Curabitur molestie porttitor libero eget viverra. Duis iaculis, ex id eleifend tempus, augue metus rhoncus elit, eu efficitur arcu metus nec leo.

Vestibulum convallis diam nec magna viverra viverra. Sed mattis, enim eget auctor dignissim, nulla sapien sodales mi, eget elementum ligula mi a felis. Etiam posuere velit scelerisque facilisis iaculis. Nunc dictum mi vitae libero finibus, vitae venenatis leo mattis. Donec cursus maximus diam, eu mattis arcu tincidunt sit amet. Pellentesque ligula enim, elementum non sagittis eu, finibus sed leo. In pretium varius velit quis iaculis. Nunc ut urna non nunc condimentum gravida. Fusce eleifend vel metus eu venenatis.

Mauris sem tortor, dignissim eu suscipit eget, elementum vitae lacus. Morbi dapibus sed leo auctor sodales. Nulla libero libero, blandit et turpis sed, facilisis eleifend felis. Quisque imperdiet tincidunt dui. Nulla convallis diam in massa scelerisque ultricies. Duis eget risus eu enim porttitor tristique nec quis augue. Maecenas consectetur viverra massa et lobortis.

INPUT RECOVERY

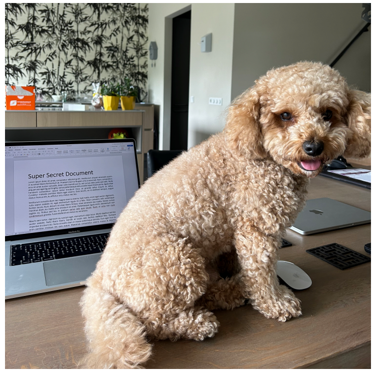
Attribute Extraction

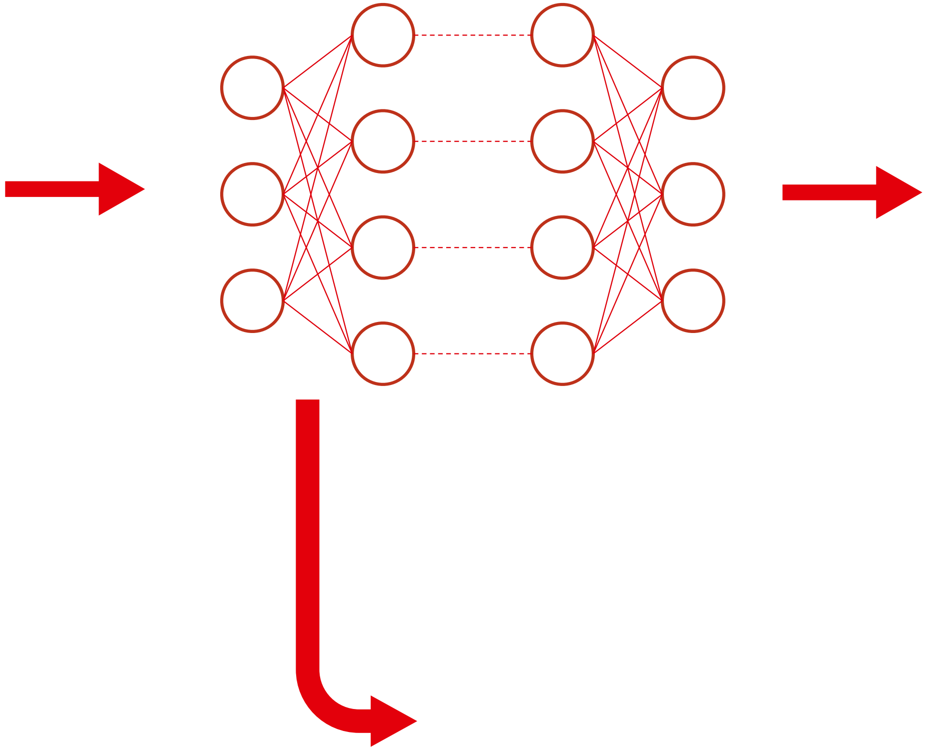


"DOG"

INPUT RECOVERY

Attribute Extraction

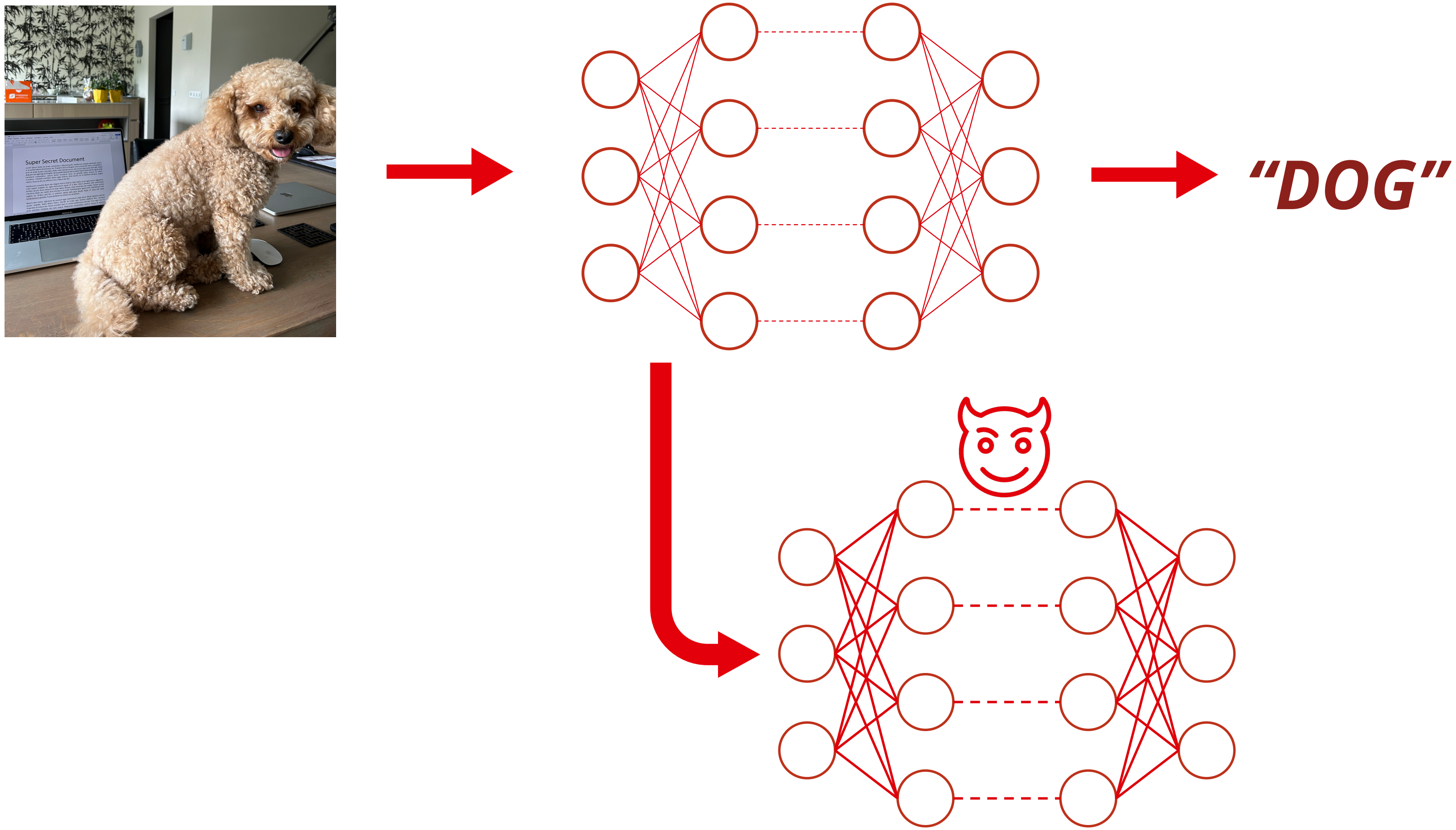




“DOG”

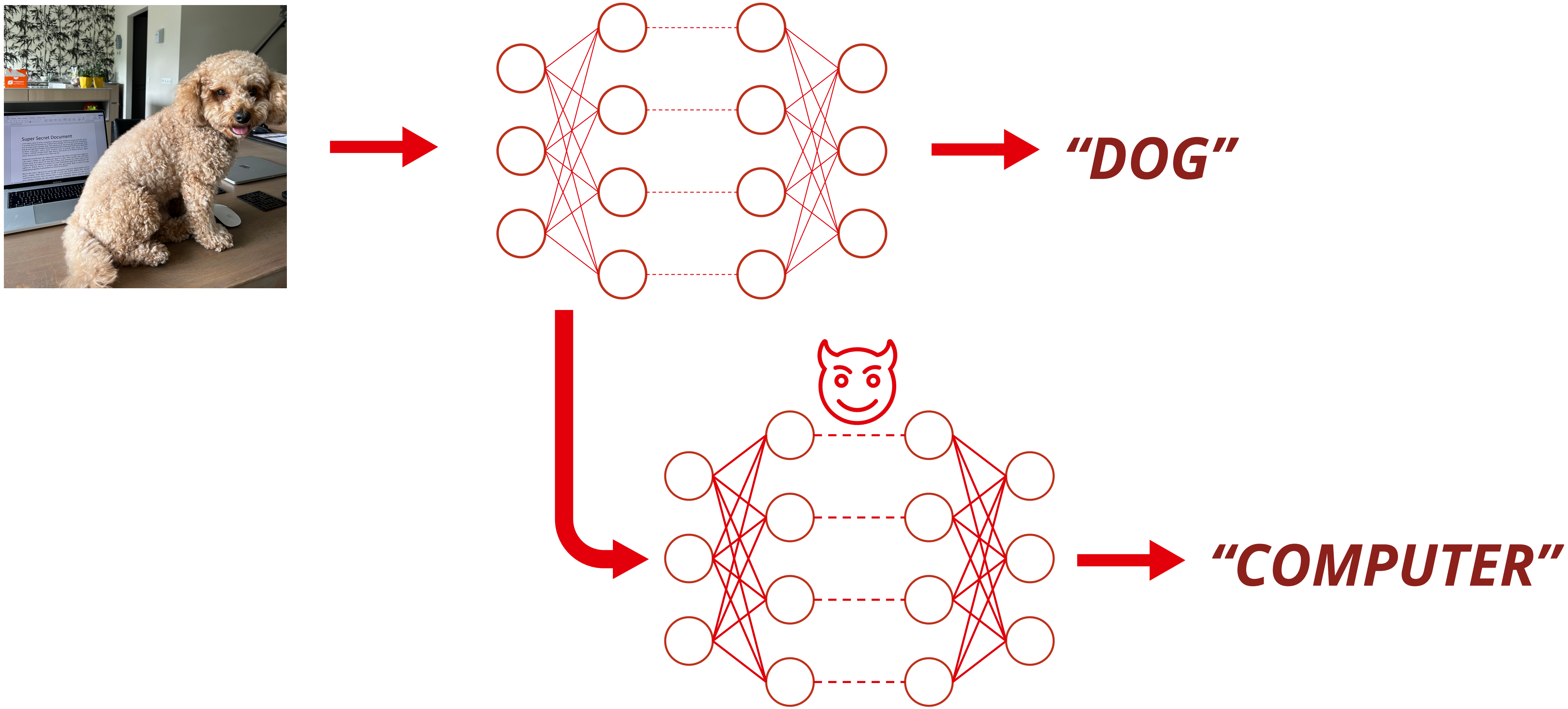
INPUT RECOVERY

Attribute Extraction



INPUT RECOVERY

Attribute Extraction

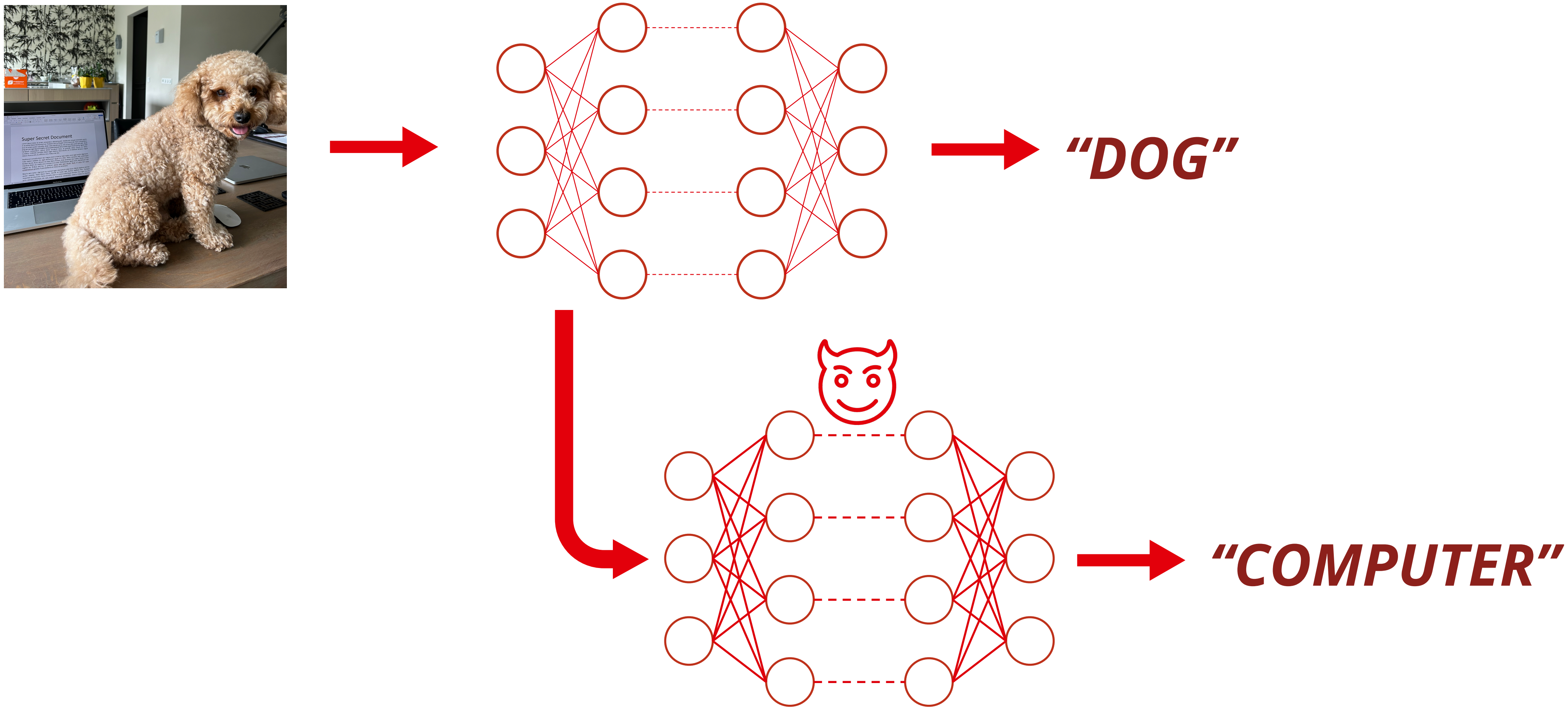


“DOG”

“COMPUTER”

INPUT RECOVERY

Attribute Extraction



“DOG”

“COMPUTER”

PARTIAL RECONSTRUCTION MAY ALREADY BE ENOUGH

CONCLUSION

Summary

Summary

- Architecture extraction
  - Limited search space
  - Lack of countermeasures

Summary

- Architecture extraction
  - Limited search space
  - Lack of countermeasures

- Parameter extraction
  - DPA for all parameters is expensive
  - High accuracy required

Summary

- Architecture extraction
  - Limited search space
  - Lack of countermeasures
- Parameter extraction
  - DPA for all parameters is expensive
  - High accuracy required
- Input recovery
  - One-shot scenario limits techniques

Takeaways + Future Work

Takeaways + Future Work

- **Architecture extraction methods could relax their assumptions**

Takeaways + Future Work

- **Architecture extraction methods could relax their assumptions**
- **The cost of current parameter extraction methods is model-dependent**

Takeaways + Future Work

- Architecture extraction methods could relax their assumptions
- The cost of current parameter extraction methods is model-dependent
- New avenues towards input recovery should be explored

QUESTIONS