Timing Channels in Adaptive Neural Networks

NDSS February 27th, 2024

Ayomide Akinsanya, Tegan Brennan

Background: Side Channels

- Different applications contain secrets:

Background: Side Channels

- ❑ Different applications contain secrets:
  - ❑ User inputs, passwords, crypto keys.

Background: Side Channels

- Different applications contain secrets:
  - User inputs, passwords, crypto keys.

Background: Side Channels

- Different applications contain secrets:
  - Inputs, outputs, hashes, crypto keys.
  - How can an attacker learn such secrets?
  - Main channel: directly obtain the secret




Background: Side Channels

- Different applications contain secrets:
  - Inputs, outputs, hashes, crypto keys.
  - How can an attacker learn such secrets?
  - Exploit some non-functional characteristics of computation
    - time, power consumption

Background: Side Channels

- Different applications contain secrets:
  - Inputs, outputs, hashes, crypto keys.
  - How can an attacker learn such secrets?
  - Exploit some non-functional characteristics of computation
    - time, power consumption (Side Channels)

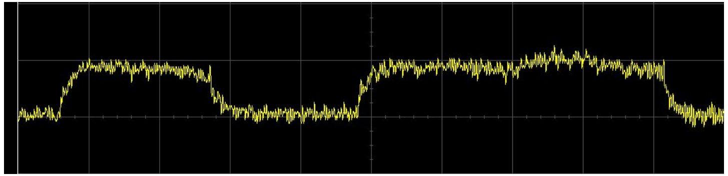
Common Side Channels

- ❑ Cache side channels
- ❑ Power side channels
- ❑ Software side channels

```
public bool check (String
guess){
   for(int i =0; i<guess.len;
i++){
      if(guess[i] !=
password[i])
         return false;
   }
   return true;
}
```

SPECTRE

MELTDOWN



Timing Side Channels (Timing Channels)

- Variation in runtime can leak secret information.

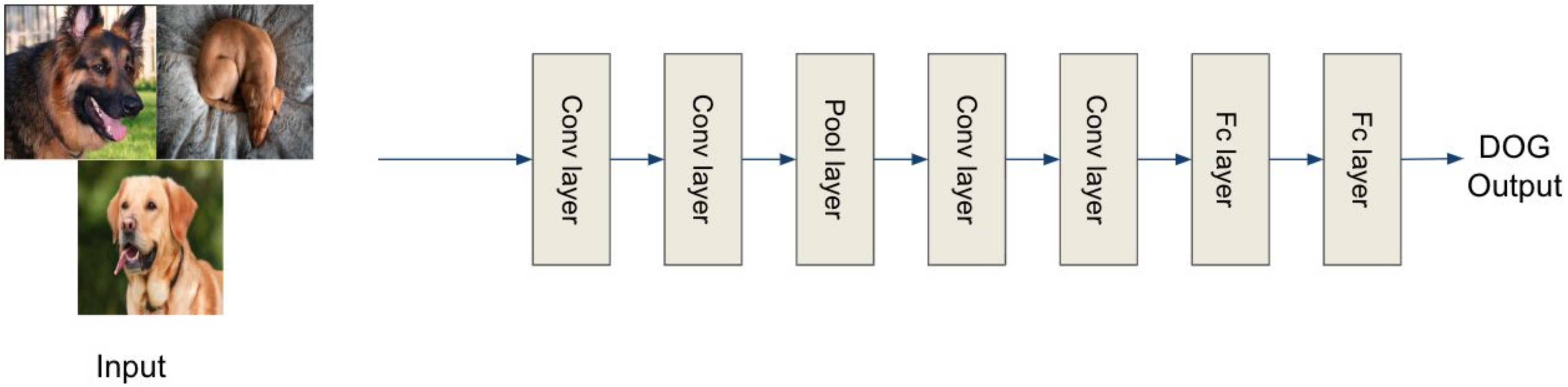
Main Contribution

- Timing side channels can arise in adaptive neural networks

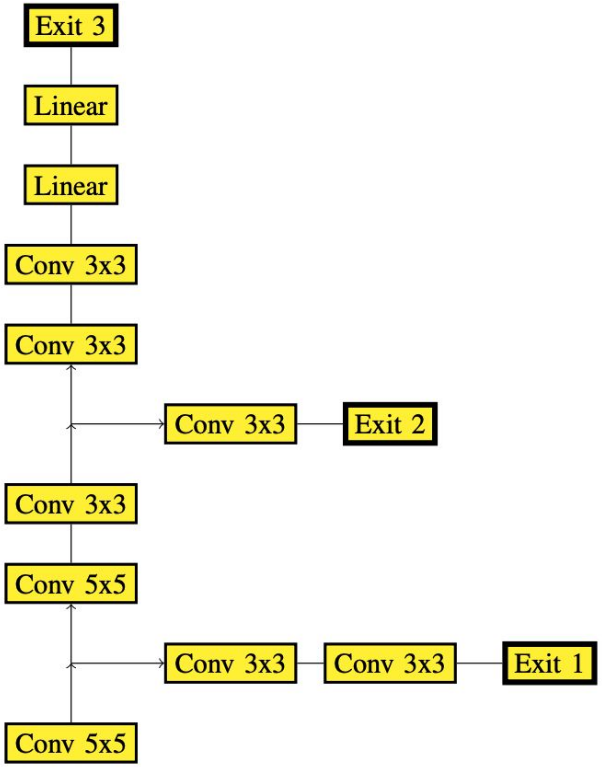
Main Contribution

- ❑ Timing side channels can arise in adaptive neural networks
- ❑ They can leak confidential information.

Conventional Neural Networks

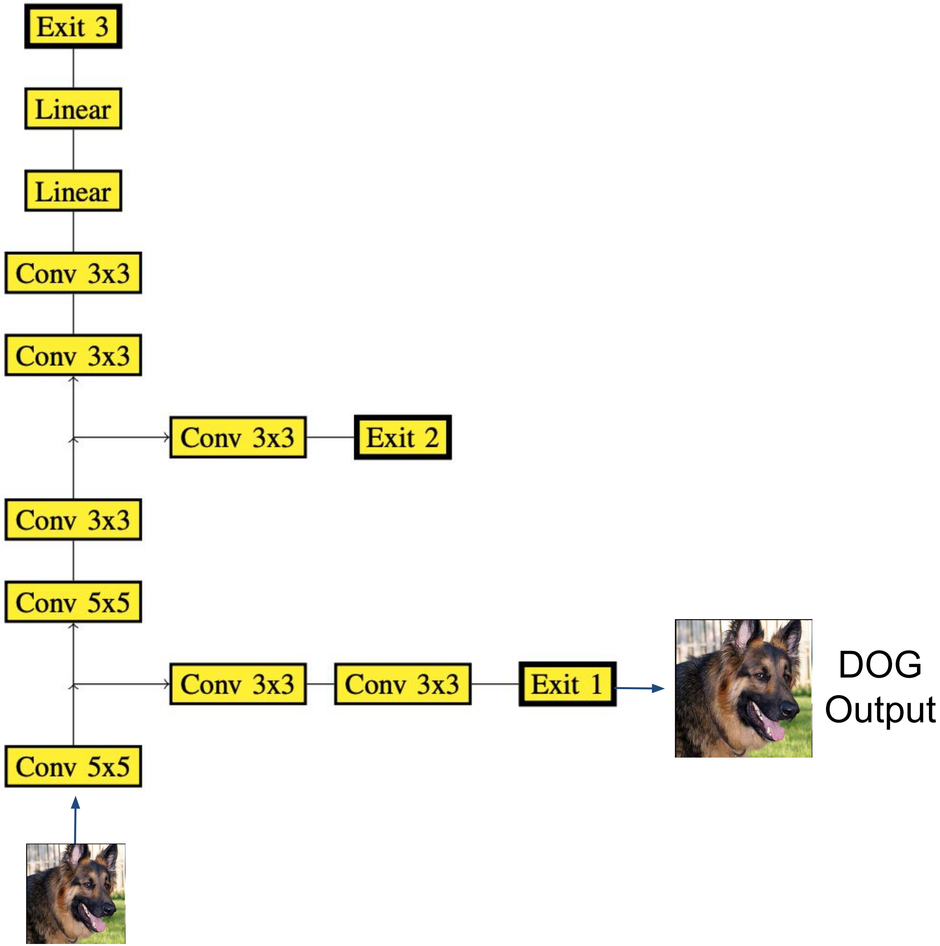


Adaptive Neural Networks

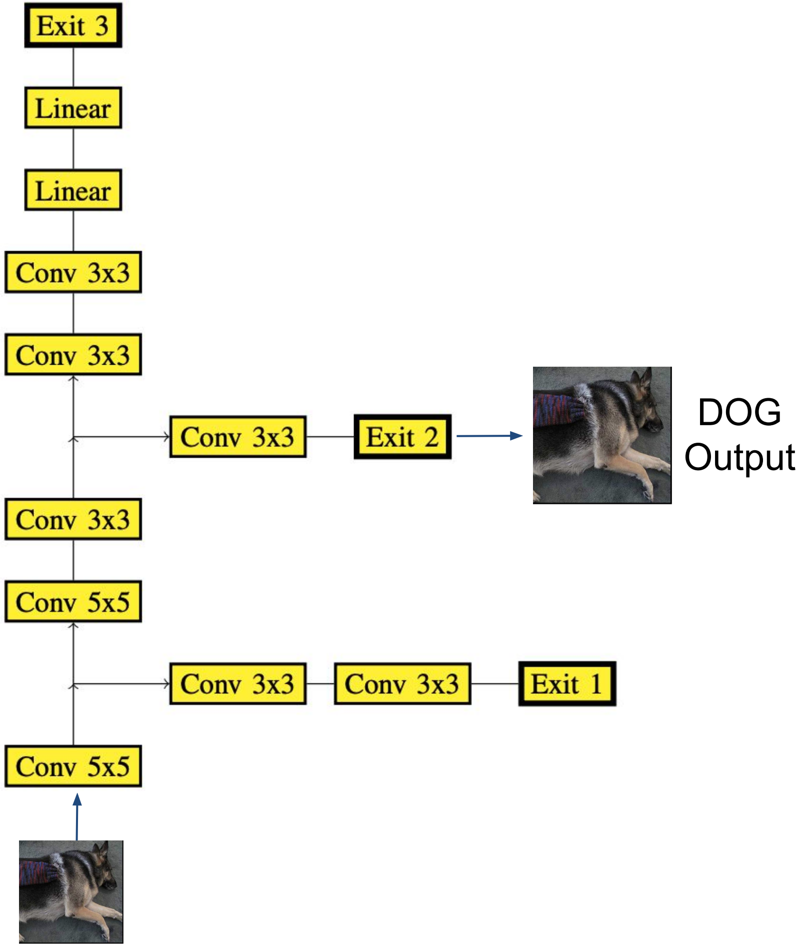


- Key insight
  - Not all inputs require the same amount of processing

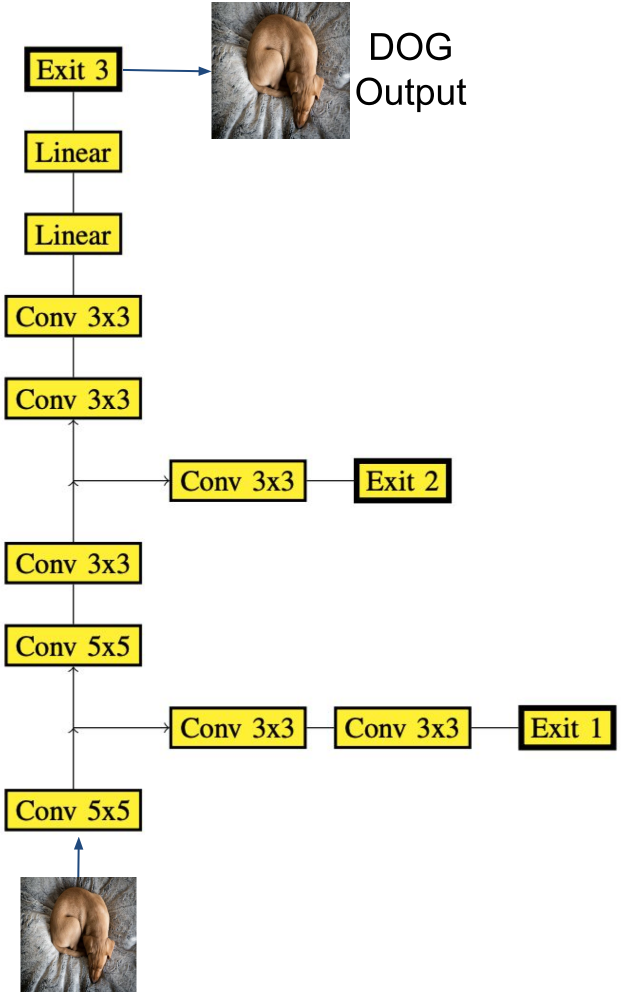
Branchy-AlexNet
(Teerapittayanon et al., 2016)



Branchy-AlexNet
(Teerapittayanon et al., 2016)



Branchy-AlexNet
(Teerapittayanon et al., 2016)



Branchy-AlexNet
(Teerapittayanon et al., 2016)

Why Adaptive Neural Networks?

- ❑ Not one size fits all

Why Adaptive Neural Networks?

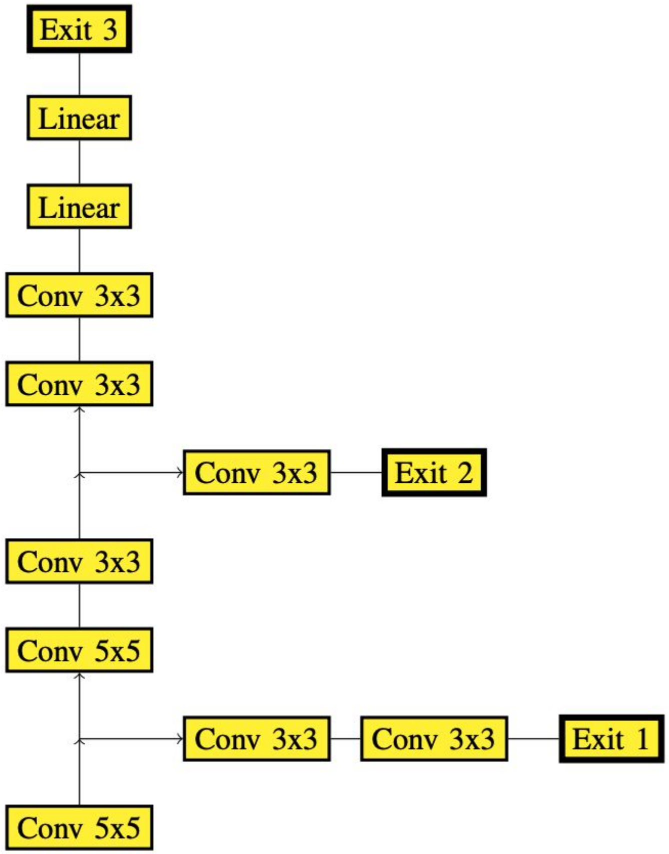
- Not one size fits all
- Lower computational cost

Why Adaptive Neural Networks?

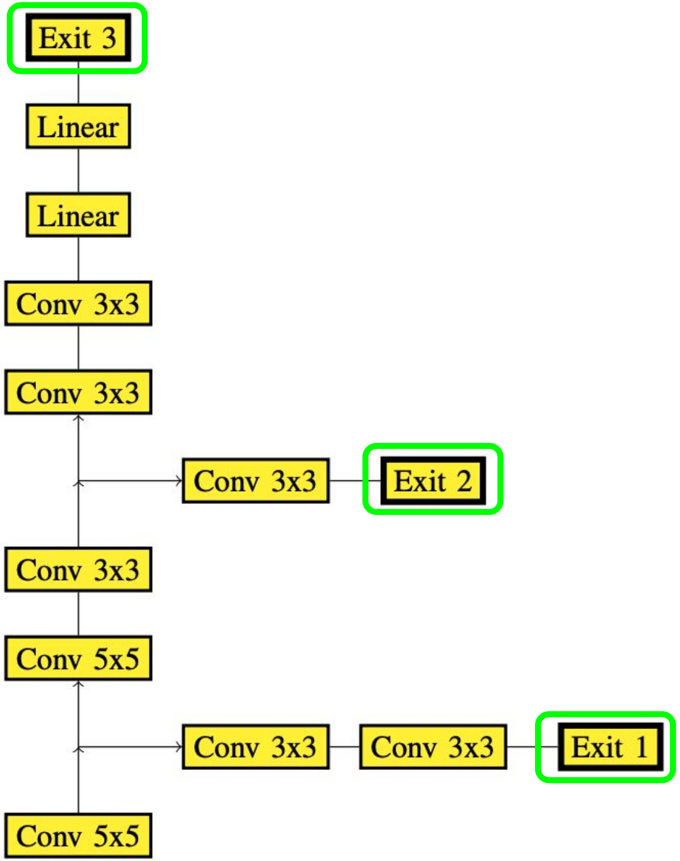
- ❑ Not one size fits all
- ❑ Lower computational cost
- ❑ Faster inference times

Why Adaptive Neural Networks?

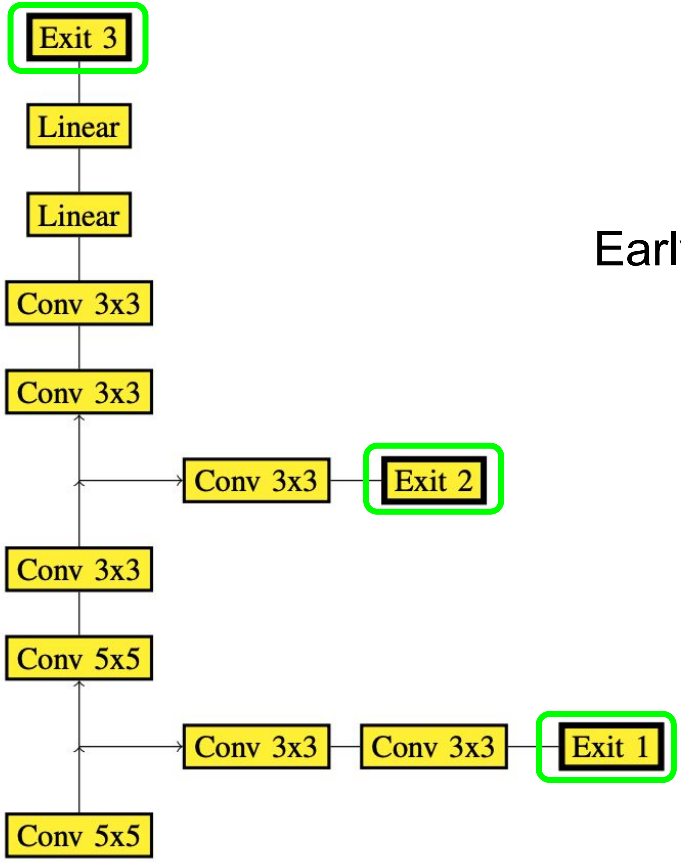
- ❑ Not one size fits all
- ❑ Lower computational cost
- ❑ Faster inference times
- ❑ Deployable on smaller devices



Branchy-AlexNet
(Teerapittayanon et al., 2016)



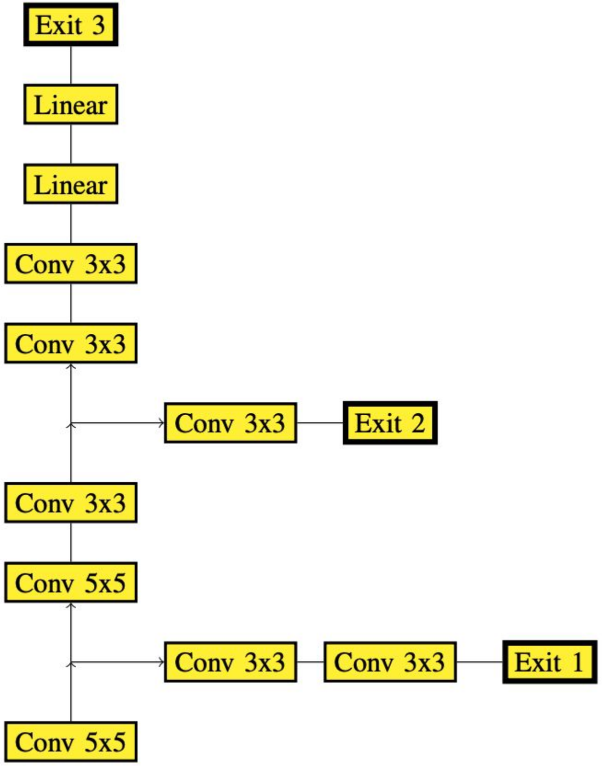
Branchy-AlexNet
(Teerapittayanon et al., 2016)



Early exits partition the inputs space

Branchy-AlexNet
(Teerapittayanon et al., 2016)

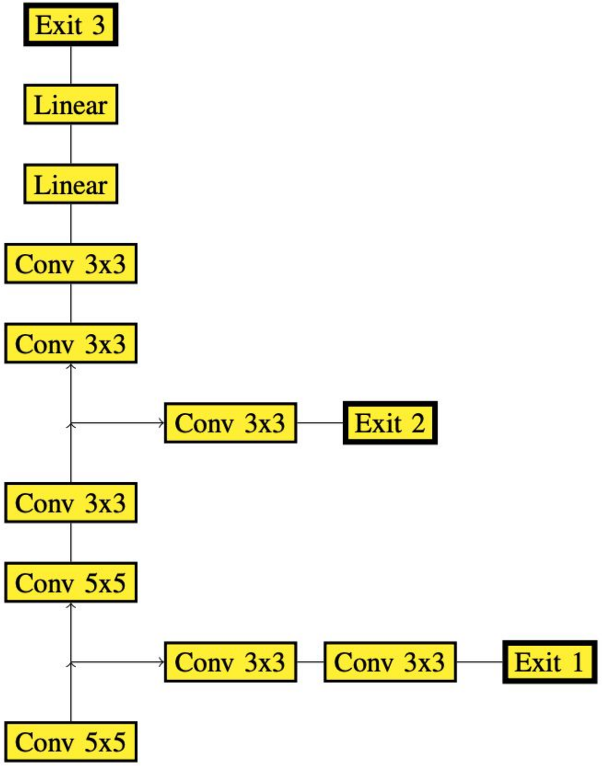
Let’s take a look at an example…



Branchy-Alexnet trained on the CANCER dataset

- ❑ Images of benign and malignant skin moles

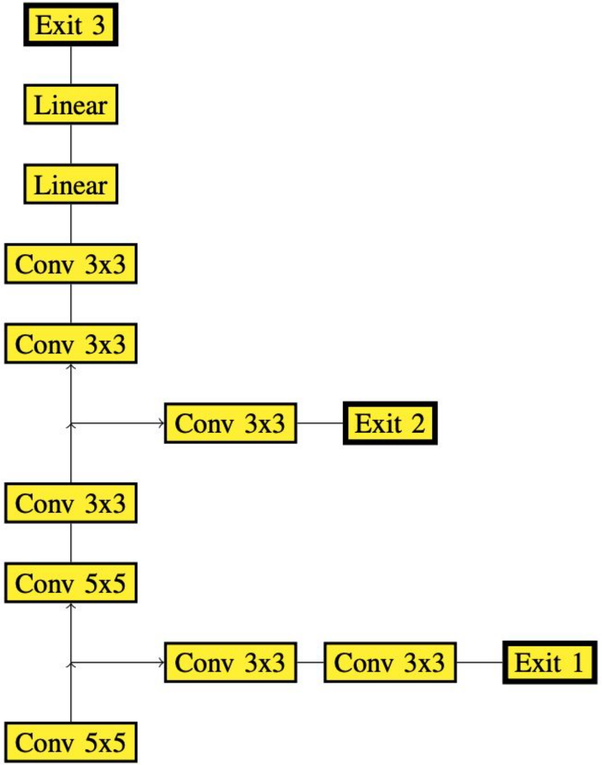
Branchy-AlexNet
(Teerapittayanon et al., 2016)



Branchy-Alexnet trained on the CANCER dataset

- Images of benign and malignant skin moles
- Given a skin mole image predict the diagnosis

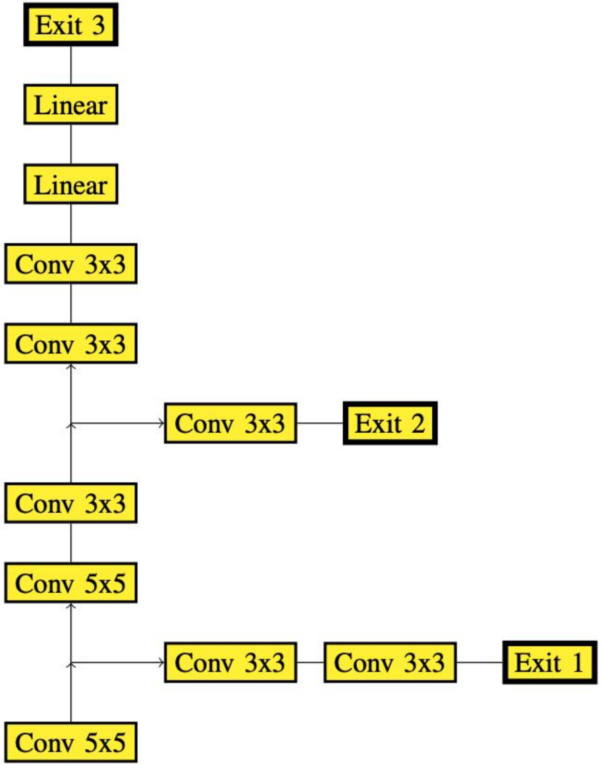
Branchy-AlexNet (Teerapittayanon et al., 2016)



Branchy-Alexnet trained on the CANCER dataset

- Images of benign and malignant skin moles
- Given a skin mole image predict the diagnosis
- Random user's shouldn't be able to learn the model's prediction

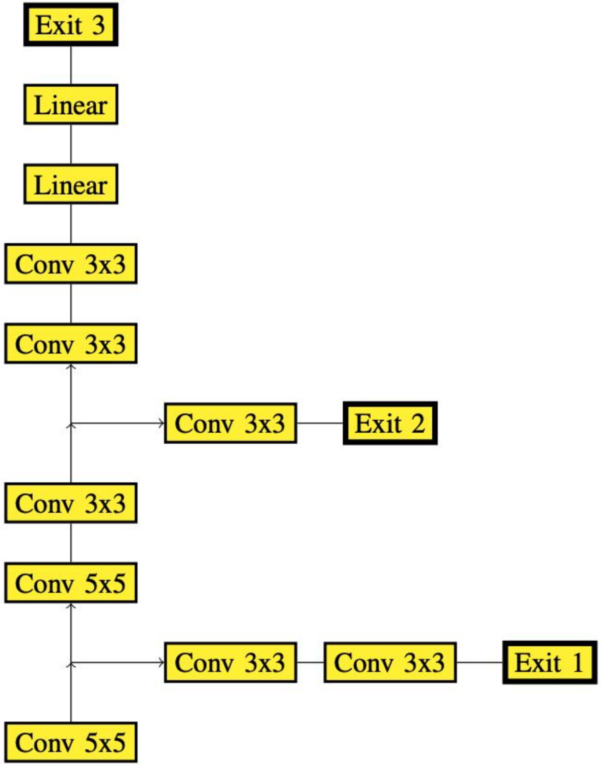
Branchy-AlexNet (Teerapittayanon et al., 2016)



Questions we would like to answer:

i. Is there a correlation between inference times and exits?

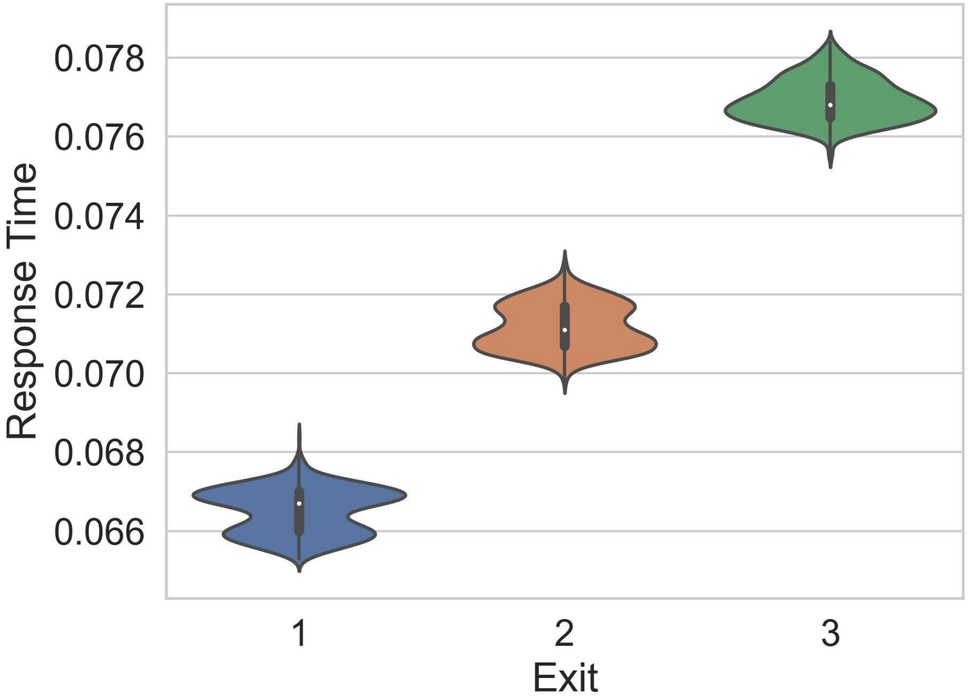
Branchy-AlexNet
(Teerapittayanon et al., 2016)



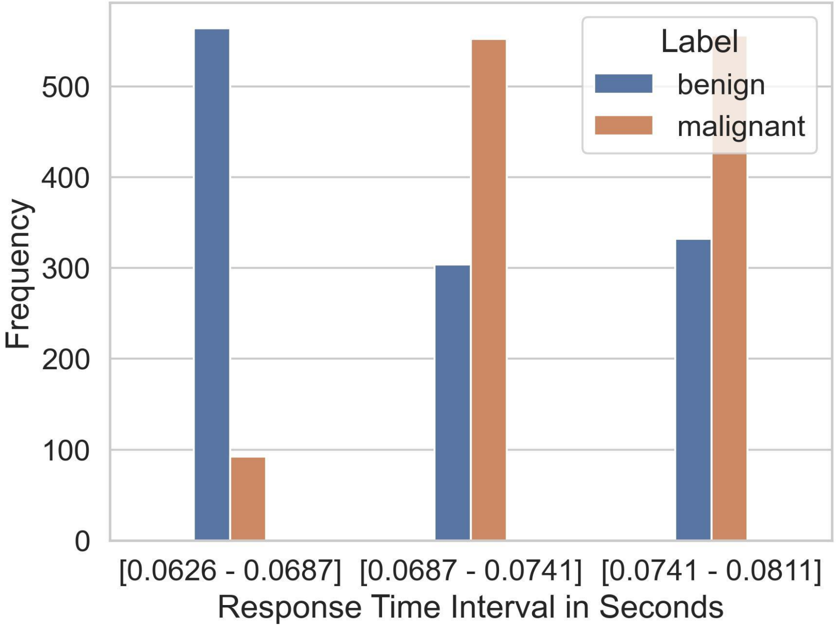
Questions we would like to answer:

i. Is there a correlation between inference times and exits?
ii. Are there any exits where the distribution is biased towards a specific attribute?

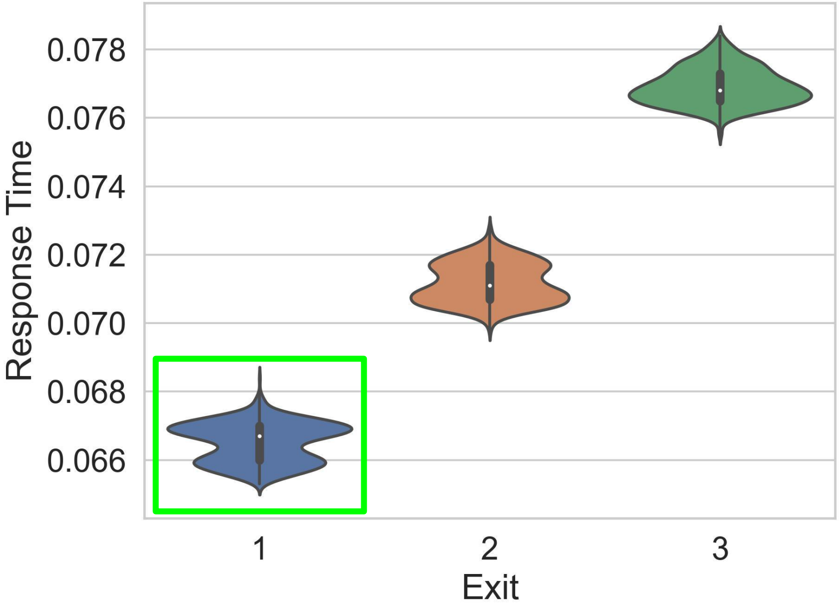
Branchy-AlexNet
(Teerapittayanon et al., 2016)

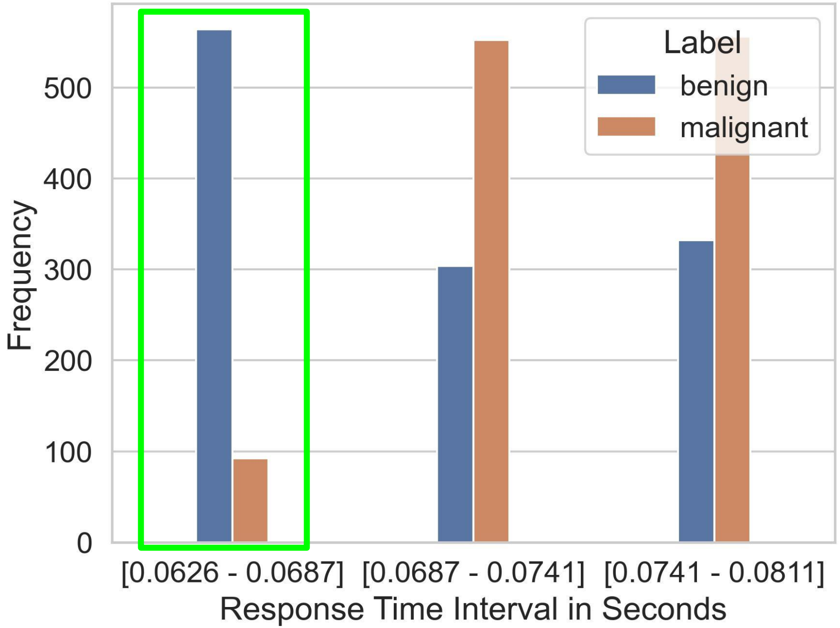


Is there a correlation between inference times and exits?

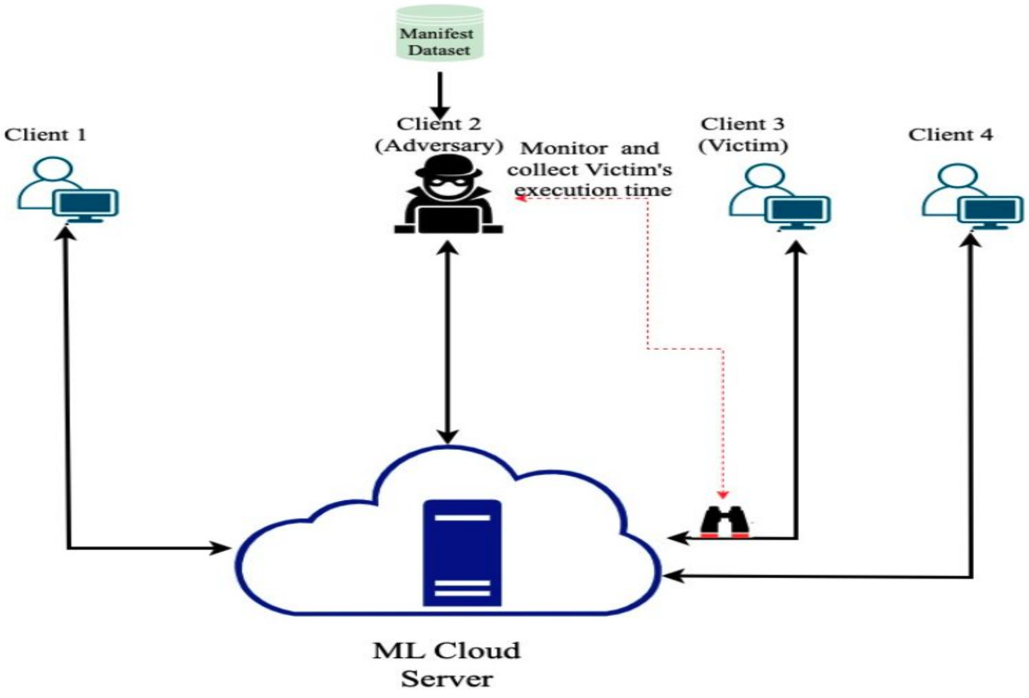


Are there any exits where the distribution is biased towards a specific attribute?

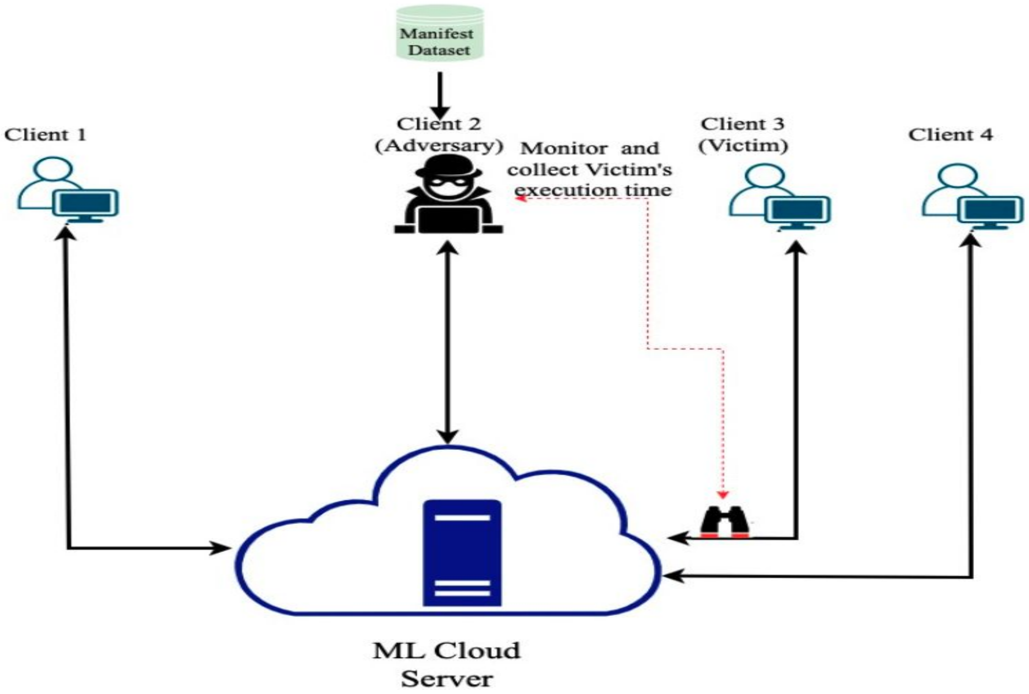

0.078
0.076
0.074
0.072
0.070
0.068
0.066
Response Time
1
2
3
Exit


Label
benign
malignant
500
400
300
200
100
0
Frequency
[0.0626 - 0.0687]
[0.0687 - 0.0741]
[0.0741 - 0.0811]
Response Time Interval in Seconds

When can this be a problem?



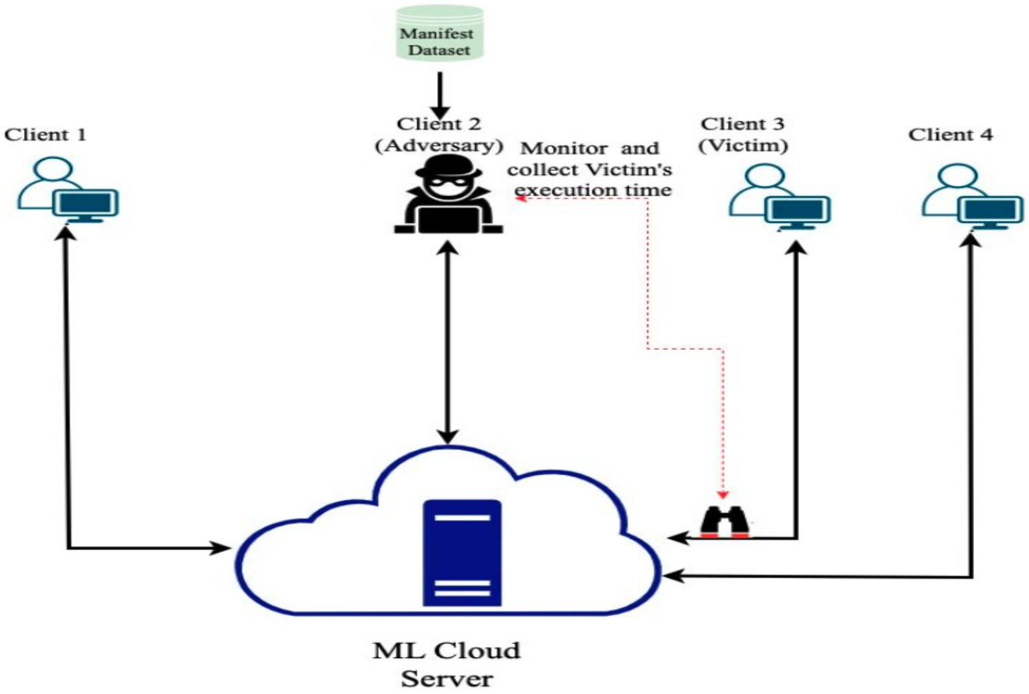
When can this be a problem?



Adversary capabilities

- ❏ Can send their own queries

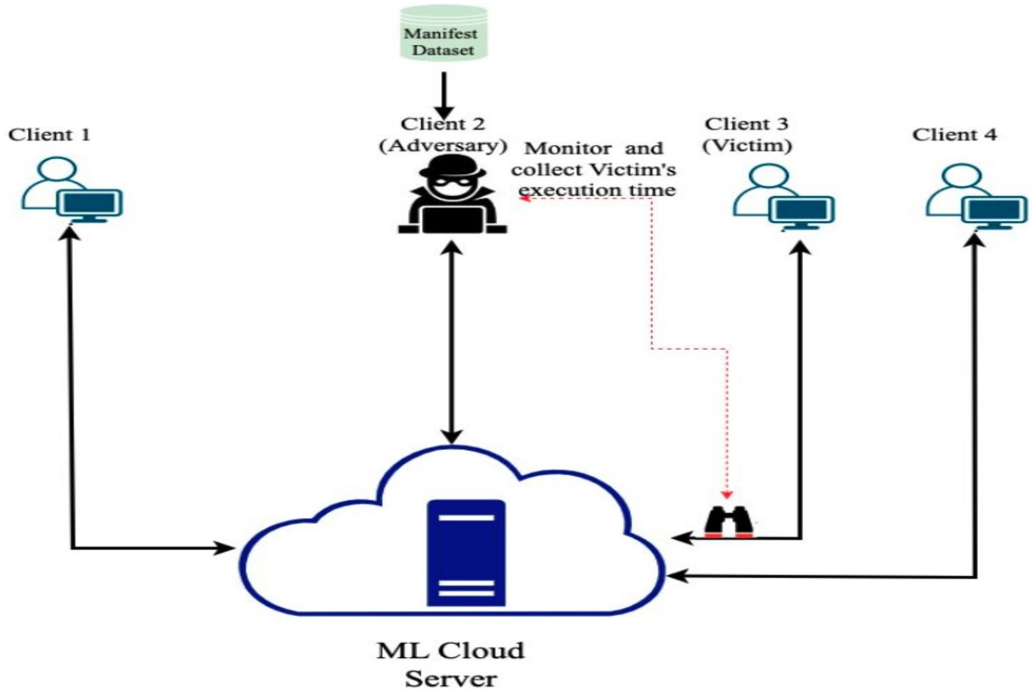
When can this be a problem?



Adversary capabilities

- ❑ Can send their own queries
- ❑ Can sniff packets over the network

When can this be a problem?



Adversary capabilities

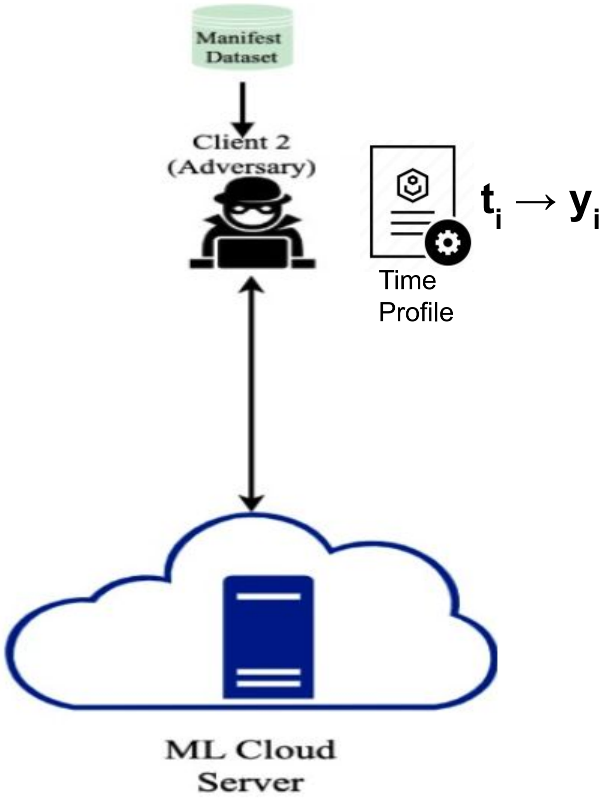
- ❑ Can send their own queries
- ❑ Can sniff packets over the network
- ❑ Can't decrypt packets

What can the Adversary Learn ?

- A sensitive attribute of the user’s input (e.g class label) 

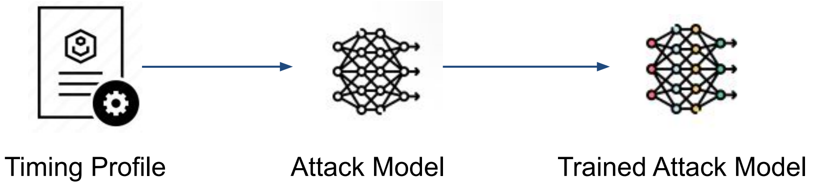
Adversary Strategy

1. Generate a timing profile



Adversary Strategy

1. Generate a timing profile
2. Train an attack model using timing profile



Adversary Strategy

1. Generate a timing profile
2. Train an attack model using timing profile
3. Given an observed timing measurement, use the attack model to infer the sensitive attribute

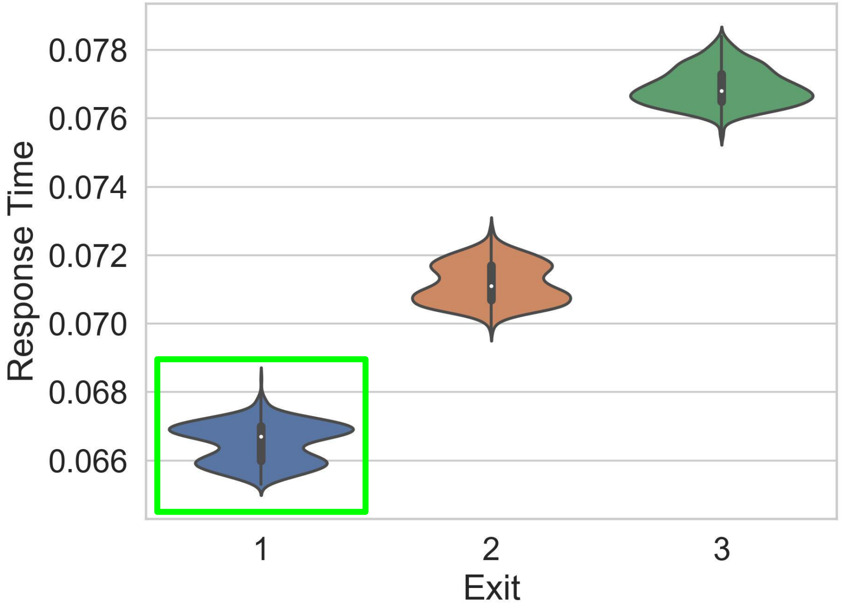
Time measurement (t_i) Trained Attack Model Private information

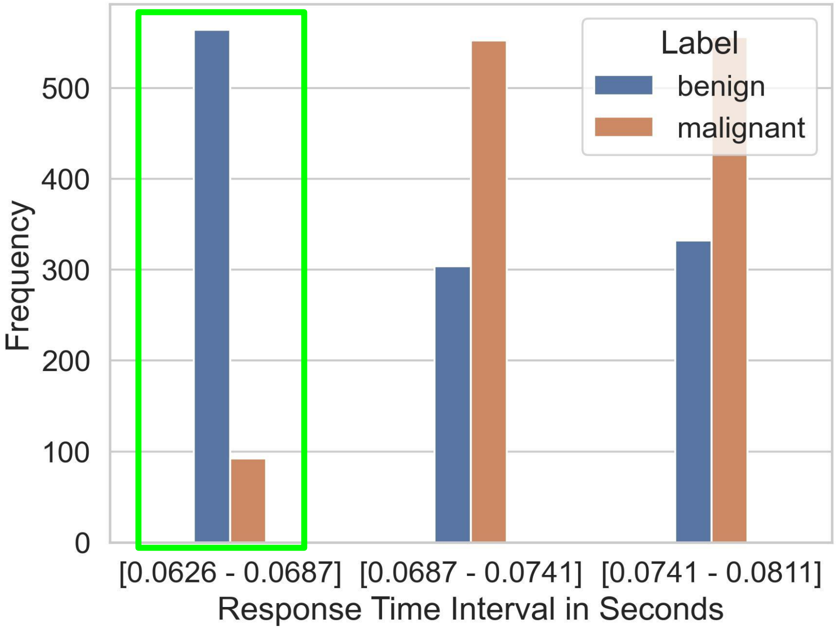
Evaluating Success

- Attack Success Rate (ASR)

Evaluating Success

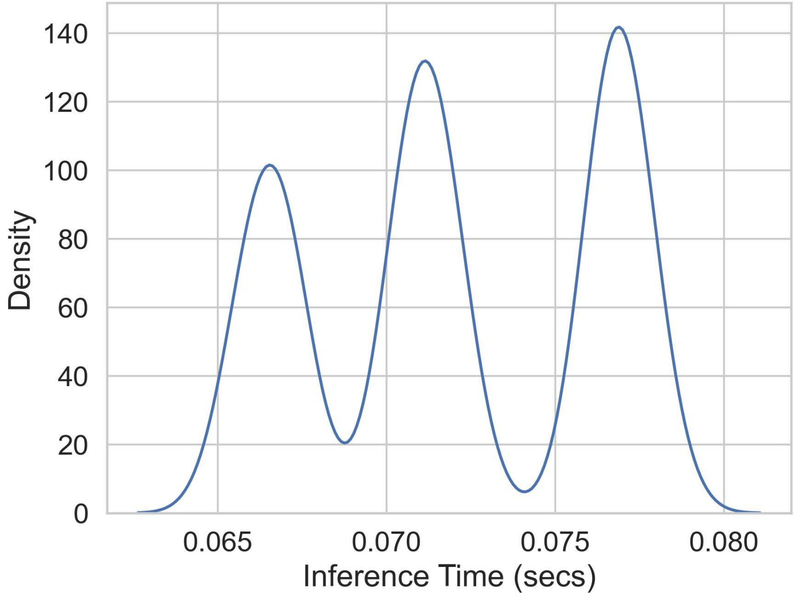
- Attack Success Rate (ASR)





Evaluating Success

- Attack Success Rate (ASR)
- Attack Success Rate (ASR/Cluster)



Evaluation & Results

- Timing measurements over a LAN

Evaluation & Results

- Timing measurements over a LAN
- Experimented using six different variants of Adaptive Neural Networks
  - Branchy-AlexNet
  - Shallow Deep Networks (SDNet)
  - Resolution Adaptive Networks (RANet)
  - Multi Scale Dense Networks (MSDNet)
  - Blockdrop
  - Skipnet

Evaluation & Results

- Timing measurements over a LAN
- Experimented using six different variants of Adaptive Neural Networks
- Across 4 different datasets
  - CIFAR 10 Dataset
  - CIFAR 100 Dataset
  - CANCER Dataset
  - FAIRFACE Dataset

Evaluation & Results

- Timing measurements over a LAN
- Experimented using six different variants of Adaptive Neural Networks
- Across 4 different datasets
- Considering 3 different attributes
  - Class label
  - Generalized class label
  - Adversarial inputs

Evaluation & Results

Arch Type	Architecture	Dataset	Attribute	No Clusters	ASR/Cluster	Cluster Input Distribution	RandGA	ASR
Non-Adaptive Networks	AlexNet	CANCER	Class Label	1	46.76	100	50	46.76
	AlexNet	FAIRFACE	Class Label	1	31.64	100	33.33	31.64
	VGG-16	CANCER	Class Label	2	[47.57, 56.1]	[72.96, 27.04]	50	50.0
	VGG-16	FAIRFACE	Class Label	2	[36.81, 33.81]	[26.06, 73.94]	33.33	34.57
	ResNet-110	CANCER	Class Label	1	46.53	100	50	46.53
	ResNet-110	FAIRFACE	Class Label	1	34.57	100	33.33	34.57
Early Exit Networks	Branchy-AlexNet	CIFAR10	Class Label	3	[13.52, 13.93, 22.22]	[44.68, 26.51, 28.81]	10	16.11
	Branchy-AlexNet	CIFAR100	Generalized Label	4	[16.67, 13.04, 8.05, 5.21]	[0.84, 6.2, 2.62, 88.34]	5	6.00
	Branchy-AlexNet	CANCER	Class Label	3	[**82.61**, 69.01, 60.53]	[27.33, 35.66, 37.00]	50	70.37
	Branchy-AlexNet	FAIRFACE	Class Label	3	[**78.26**, 64.52, 41.46]	[4.20, 11.08, 84.72]	33.33	45.71
	SDNet	CIFAR10	Class Label	5	[19.22, 10.83, 14.71, 21.74, 19.61]	[21.22, 42.98, 29.32, 2.75, 3.73]	10	14.44
	SDNet	CIFAR100	Generalized Label	3	[5.64, 7.62, 9.63]	[66.98, 12.67, 20.35]	5	6.67
	SDNet	CANCER	Class Label	3	[64.5, 66.67, 62.58]	[57.58, 2.84, 39.58]	50	63.89
	SDNet	FAIRFACE	Class Label	5	[**94.29**, 42.0, 36.04 36.88, 43.3]	[5.42, 13.06, 17.75, 22.19, 41.58]	33.33	43.21
Model Cascade Networks	RANet	CIFAR10	Class Label	3	[25.0, 17.23, 10.58]	[2.23, 14.71, 83.06]	10	12.06
	RANet	CIFAR100	Generalized Label	3	[10.34, 9.47, 6.29]	[2.77, 18.72, 78.51]	5	7.06
	RANet	CANCER	Class Label	3	[**95.05**, 62.07, 66.39]	[23.35, 21.55, 55.09]	50	72.26
	RANet	FAIRFACE	Class Label	3	[**81.44**, 34.84, 45.45]	[12.83, 20.17, 67.0]	33.33	43.98
	MSDNet	CIFAR10	Class Label	3	[12.94, 13.51, 20.47]	[67.90, 23.76, 8.34]	10	13.61
	MSDNet	CIFAR100	Generalized Label	3	[9.28, 6.37, 7.27]	[19.98, 37.62, 42.4]	5	7.33
	MSDNet	CANCER	Class Label	2	[**89.32**, 61.4]	[20.83, 79.17]	50	68.06
	MSDNet	FAIRFACE	Class Label	2	[**77.38**, 36.17]	[10.92, 89.08]	33.33	41.51
Dynamic Networks	BlockDrop	CIFAR10	Class Label	2	[**98.94**, 21.57]	[4.80, 95.20]	10	25.93
	BlockDrop	CIFAR100	Generalized Label	3	[**100**,66.67,4.75]	[0.04, 0.41, 99.56]	5	5.06
	BlockDrop	CIFAR10	Adversarial Input	2	[**100**,61.16]	[4.17, 95.83]	50	62.71
	SkipNet	CIFAR10	Class Label	5	[0, **60.98**, 16.57, 16.51, 0]	[0.01, 1.66, 55.67, 42.64, 0.01]	10	17.56
	SkipNet	CIFAR100	Generalized Label	3	[0, 8.36, 6.54]	[0.1, 30.93, 68.97]	5	7.11
	SkipNet	CANCER	Class Label	5	[**77.08**, 45.83, 58.73, 51.79, 50.0]	[10.29, 31.62, 30.04, 27.88, 0.17]	50	54.63
	SkipNet	FAIRFACE	Class Label	4	[39.22, 38.06, 48.65, 0]	[46.25, 42.52, 11.19, 0.03]	33.33	39.81
	SkipNet	CIFAR10	Adversarial Input	5	[0, 66.35, 52.78, 55.23, 0]	[0.01, 17.94, 31.99, 50.04, 0.01]	50	56.51

Evaluation & Results

Arch Type	Architecture	Dataset	Attribute	No Clusters	Clusters Accuracy	Cluster Input distribution	Random Guess	Accuracy
Model Cascade Networks	RANet	CIFAR10	Class Label	3	[25.0, 17.23, 10.58]	[2.23, 14.71, 83.06]	10	12.06
	RANet	CIFAR100	Generalized Label	3	[10.34, 9.47, 6.29]	[2.77, 18.72, 78.51]	5	7.06
	RANet	CANCER	Class Label	3	[**95.05**, 62.07, 66.39]	[**23.35**, 21.55, 55.09]	50	72.26
	RANet	FAIRFACE	Class Label	3	[**81.44**, 34.84, 45.45]	[**12.83**, 20.17, 67.0]	33.33	43.98
	MSDNet	CIFAR10	Class Label	3	[12.94, 13.51, 20.47]	[67.90, 23.76, 8.34]	10	13.61
	MSDNet	CIFAR100	Generalized Label	3	[9.28, 6.37, 7.27]	[19.98, 37.62, 42.4]	5	7.33
	MSDNet	CANCER	Class Label	2	[**89.32**, 61.4]	[**20.83**, 79.17]	50	68.06
	MSDNct	FAIRFACE	Class Labcl	2	[**77.38**, 36.17]	[**10.92**, 89.08]	33.33	41.51

Evaluation & Results

Arch Type	Architecture	Dataset	Attribute	No Clusters	Clusters Accuracy	Cluster Input distribution	Random Guess	Accuracy
Dynamic Networks	BlockDrop	CIFAR10	Class Label	2	[**98.94**, 21.57]	[**4.80**, 95.20]	10	25.93
	BlockDrop	CIFAR100	Generalized Label	3	[**100**,66.67,4.75]	[0.04, 0.41, 99.56]	5	5.06
	BlockDrop	CIFAR10	Adversarial Input	2	[**100**,61.16]	[**4.17**, 95.83]	50	62.71
	SkipNet	CIFAR10	Class Label	5	[0, **60.98**, 16.57, 16.51, 0]	[0.01, 1.66, 55.67, 42.64, 0.01]	50	17.56
	SkipNet	CIFAR100	Generalized Label	3	[0, 8.36, 6.54]	[0.1, 30.93, 68.97]	5	7.11
	SkipNet	CANCER	Class Label	5	[**77.08**, 45.83, 58.73, 51.79, 50.0]	[10.29, 31.62, 30.04, 27.88, 0.17]	50	54.63
	SkipNet	FAIRFACE	Class Label	4	[39.22, 38.06, 48.65, 0]	[46.25, 42.52, 11.19, 0.03]	33.33	39.81
	SkipNet	CIFAR10	Adversarial Input	2	[0, 8.36, 6.54]	[0, 8.36, 6.54]	50	50

Evaluation & Results

Arch Type	Architecture	Dataset	Attribute	No Clusters	Clusters Accuracy	Cluster Input distribution	Random Guess	Accuracy
Early Exit Networks	Branchy-AlexNet	CIFAR10	Class Label	3	[13.52, 13.93, 22.22]	[44.68, 26.51, 28.81]	10	16.11
	Branchy-AlexNet	CIFAR100	Generalized Label	4	[16.67, 13.04, 8.05, 5.21]	[0.84, 6.2, 2.62, 88.34]	5	6.00
	Branchy-AlexNet	CANCER	Class Label	3	[**82.61**, 69.01, 60.53]	[27.33, 35.66, 37.00]	50	70.37
	Branchy-AlexNet	FAIRFACE	Class Label	3	[**78.26**, 64.52, 41.46]	[4.20, 11.08, 84.72]	33.33	45.71
	SDNet	CIFAR10	Class Label	5	[19.22, 10.83, 14.71, 21.74, 19.61]	[21.22, 42.98, 29.32, 2.75, 3.73]	10	14.44
	SDNet	CIFAR100	Generalized Label	3	[5.64, 7.62, 9.63]	[66.98, 12.67, 20.35]	5	6.67
	SDNet	CANCER	Class Label	3	[64.5, 66.67, 62.58]	[57.58, 2.84, 39.58]	50	63.89
	SDNet	FAIRFACE	Class Label	5	[94.29, 42.0, 36.04 36.88, 43.3]	[5.42, 13.06, 17.75, 22.19, 41.58]	33.33	43.21

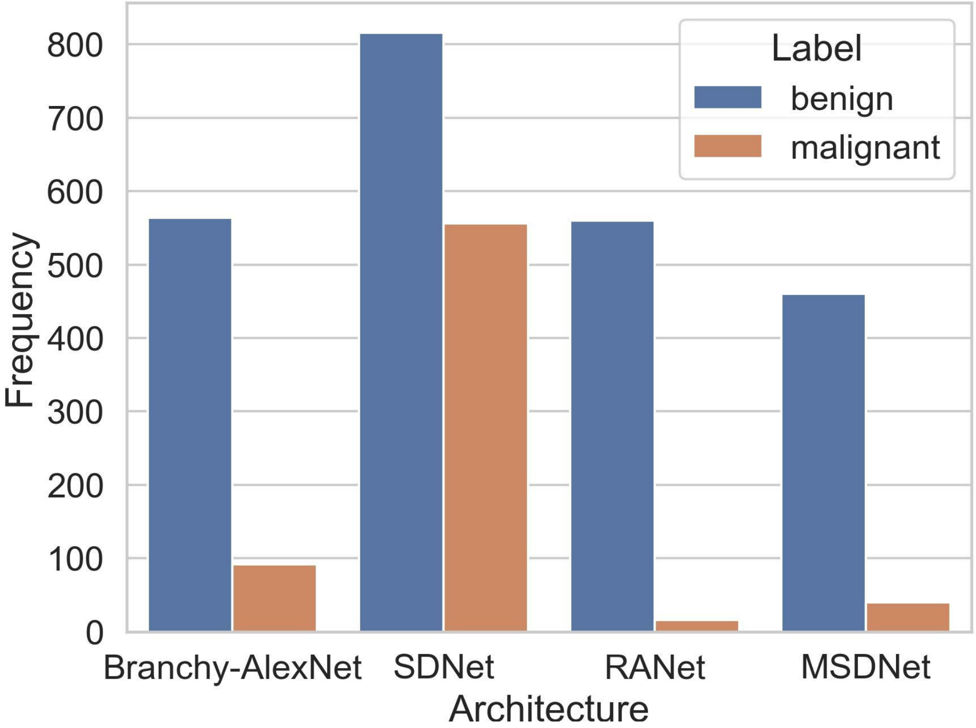
Effect of Hyperparameter Tuning

Dataset	Architecture	Attribute	Exit Thresholds	Setting	Accurracy	Performance (secs)	No Clusters	ASR/Cluster
FAIRFACE	Branchy-AlexNet	Class Label	[2.0e-03, 5.0e-02]	Conservative	75.03	47.2	3	[**100.0**, 57.63, 38.78]
FAIRFACE	Branchy-AlexNet	Class Label	[2.0e-02, 5.0e-02]	Balanced	74.81	45.2	3	[**78.26**, 64.52, 41.46]
FAIRFACE	Branchy-AlexNet	Class Label	[5.0e-01, 5.0e-01]	Relaxed	73.33	34.9	3	[36.84, 46.72, 42.03]
FAIRFACE	SDNet	Class Label	[0.99, 0.99]	Conservative	78.64	50.8	5	[**90.91**, 58.82, 57.47, 46.73, 43.77]
FAIRFACE	SDNet	Class Label	[0.95, 0.95]	Balanced	78.11	46.6	5	[**94.29**, 42.0, 36.04 36.88, 43.3]
FAIRFACE	SDNet	Class Label	[0.8, 0.8]	Relaxed	77.11	39.9	5	[56.49, 42.31, 40.69, 44.55, 56.63]

Effect of Hyperparameter Tuning

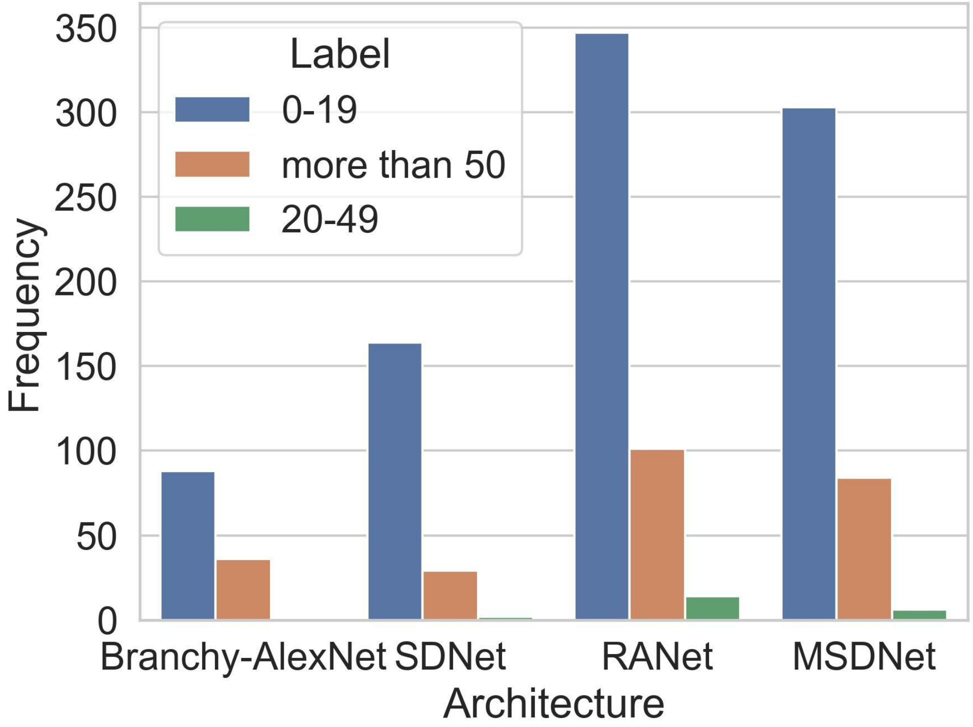
Dataset	Architecture	Attribute	Exit Thresholds	Setting	Accurracy	Performance (secs)	No Clusters	ASR/Cluster
FAIRFACE	Branchy-AlexNet	Class Label	[2.0e-03, 5.0e-02]	Conservative	75.03	47.2	3	[**100.0**, 57.63, 38.78]
FAIRFACE	Branchy-AlexNet	Class Label	[2.0e-02, 5.0e-02]	Balanced	74.81	45.2	3	[**78.26**, 64.52, 41.46]
FAIRFACE	Branchy-AlexNet	Class Label	[5.0e-01, 5.0e-01]	Relaxed	73.33	34.9	3	[36.84, 46.72, 42.03]
FAIRFACE	SDNet	Class Label	[0.99, 0.99]	Conservative	78.64	50.8	5	[**90.91**, 58.82, 57.47, 46.73, 43.77]
FAIRFACE	SDNet	Class Label	[0.95, 0.95]	Balanced	78.11	46.6	5	[**94.29**, 42.0, 36.04 36.88, 43.3]
FAIRFACE	SDNet	Class Label	[0.8, 0.8]	Relaxed	77.11	39.9	5	[56.49, 42.31, 40.69, 44.55, 56.63]

Interesting Observations



Input Distribution of benign and malignant skin mole images across the first time cluster of Branchy-AlexNet, SDNet, RANet and MSDNet

Interesting Observations



Input Distribution of FAIRFACE age classes across the first time cluster of Branchy-AlexNet, SDNet, RANet and MSDNet

Conclusion

- ❏ Demonstrate how timing side channels in ADNNs can leak private information

Conclusion

- ❑ Demonstrate how timing side channels in ADNNs can leak private information
- ❑ Show how an adversary might leverage such timing side channel leakage

Conclusion

- ❑ Demonstrate how timing side channels in ADNNs can leak private information
- ❑ Show how an adversary might leverage such timing side channel leakage
- ❑ Experimental validate our technique across six different ADNNs and four datasets

Conclusion

- ❑ Demonstrate how timing side channels in ADNNs can leak private information.
- ❑ Show how an adversary might leverage such timing side channel leakage.
- ❑ Experimental validate our technique across six different ADNNs and four datasets.
- ❑ Show how hyperparameter tuning such as exit threshold can result in trade offs between accuracy, efficiency and privacy.

Conclusion

Artifact Evaluated NDSS SYMPOSIUM Available

Artifact Evaluated NDSS SYMPOSIUM Functional

Artifact Evaluated NDSS SYMPOSIUM Reproduced

- ❑ Demonstrate how timing side channels in ADNNs can leak private information.
- ❑ Show how an adversary might leverage such timing side channel leakage.
- ❑ Experimental validate our technique across six different ADNNs and four datasets.
- ❑ Show how hyperparameter tuning such as exit threshold can result in trade offs between accuracy, efficiency and privacy.

GitHub https://github.com/akinsanyaayomide/ADNNTimeLeaks/tree/main


Thank You
ANY QUESTIONS?

Future Work

- ❑ Deliberate crafting of timing side channels
- ❑ Automatic testing and validation of ADNNs for timing side channels
- ❑ Online monitoring of ADNNs for timing side channels