Assessing the Limitations of Activation Clipping for Fault Mitigation in Vision and Language Transformers

LUCAS ROQUET[1], FERNANDO FERNANDES DOS SANTOS[1], LUIGI CARRO[2], AND ANGELIKI KRITIKAKOU[1]

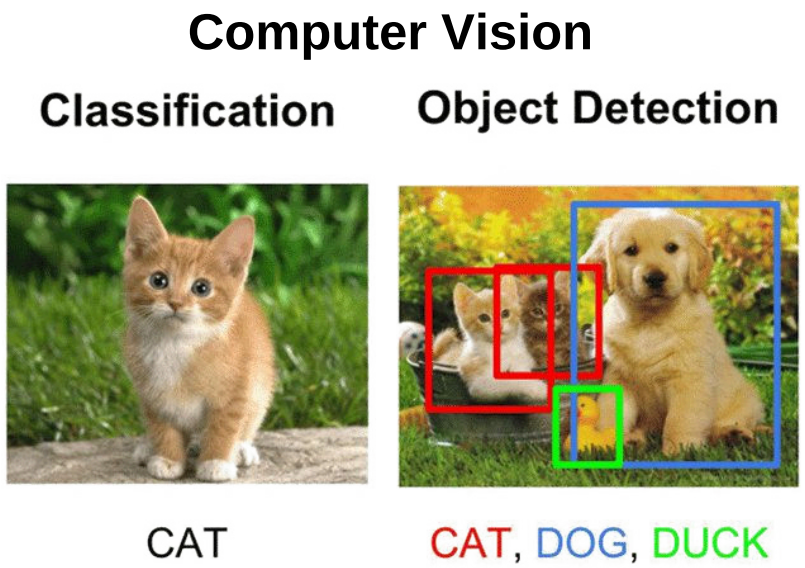
1. IRISA/Inria , Univ Rennes
2. Institute of Informatics, UFRGS

Transformers models

- Transformers, state-of-the-art of machine learning (ML) models
  - Highly accurate, very complex
  - Many tasks

Transformers models

- Transformers, state-of-the-art of machine learning (ML) models
  - Highly accurate, very complex
  - Many tasks



Transformers models

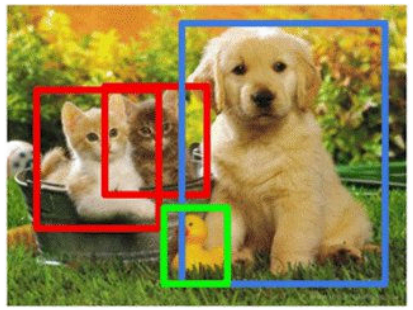
- Transformers, state-of-the-art of machine learning (ML) models
  - Highly accurate, very complex
  - Many tasks

Computer Vision

Classification

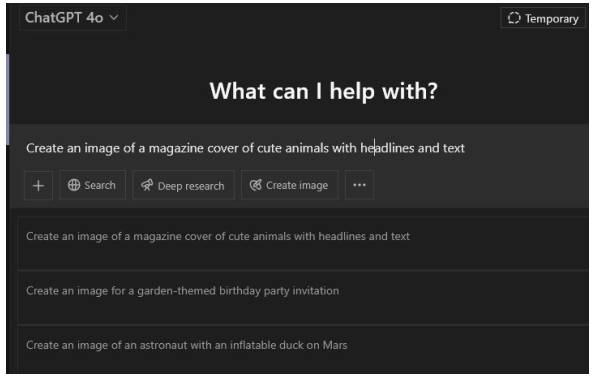
CAT

Object Detection



CAT, DOG, DUCK

NLP



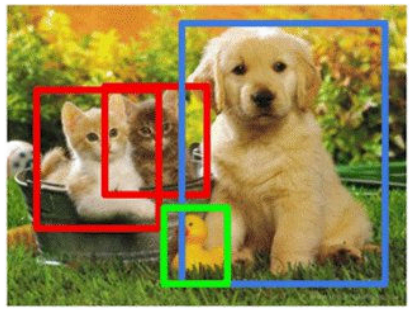
Transformers models

- Transformers, state-of-the-art of machine learning (ML) models
  - Highly accurate, very complex
  - Many tasks

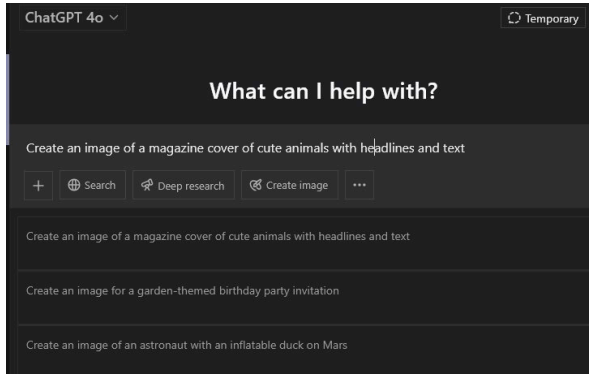
Computer Vision **NLP**

Classification **Object Detection**

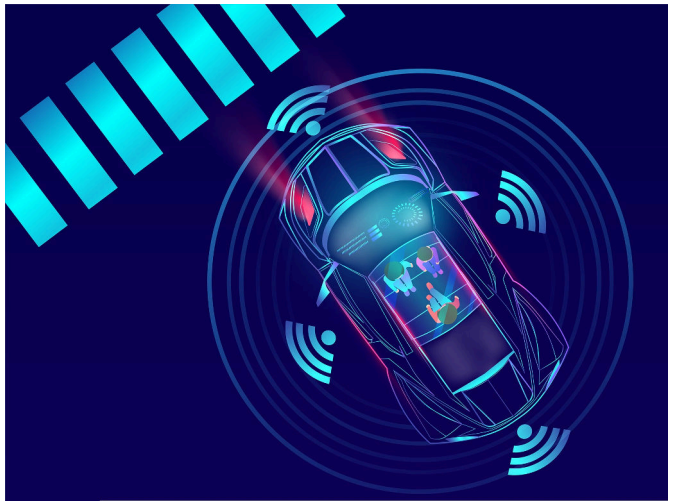
CAT



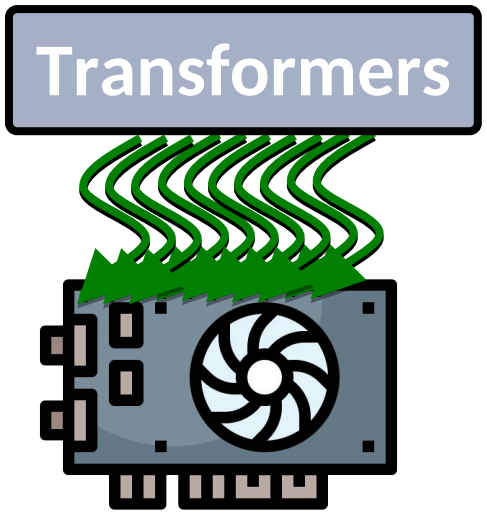
CAT, DOG, DUCK



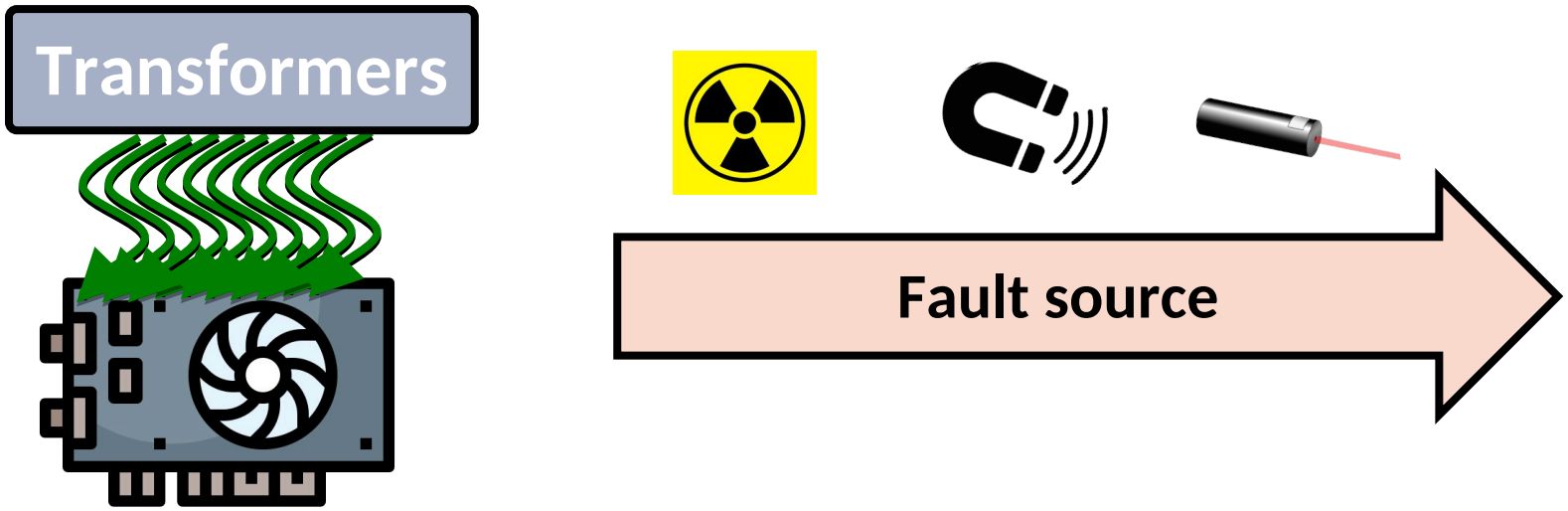
Safety-critical systems



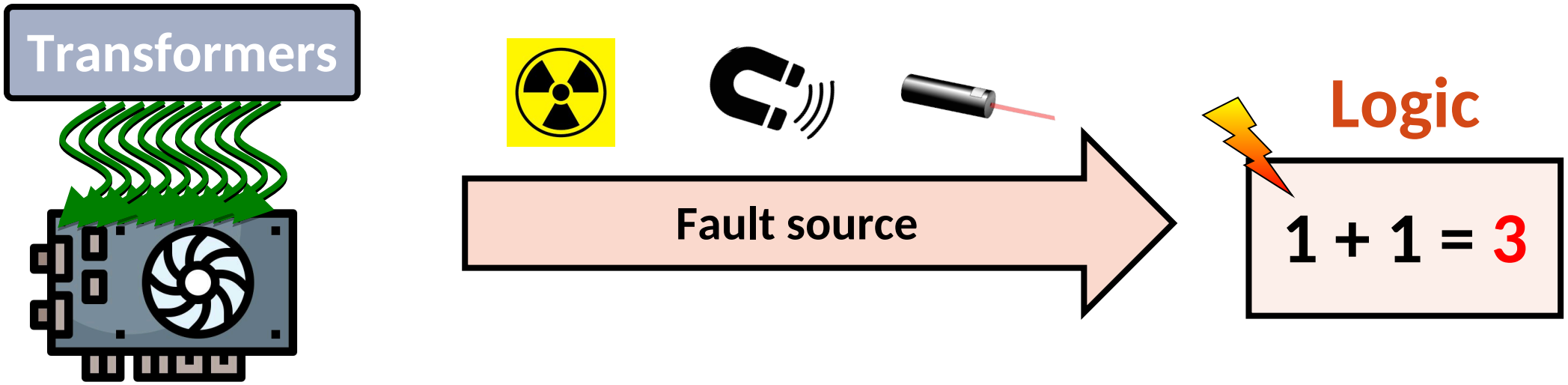
Hardware is vulnerable to faults



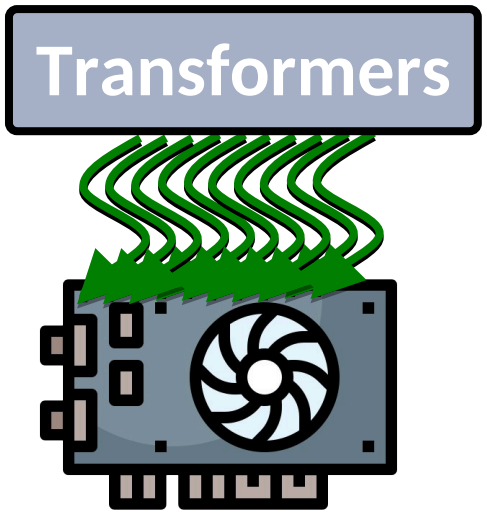
Hardware is vulnerable to faults

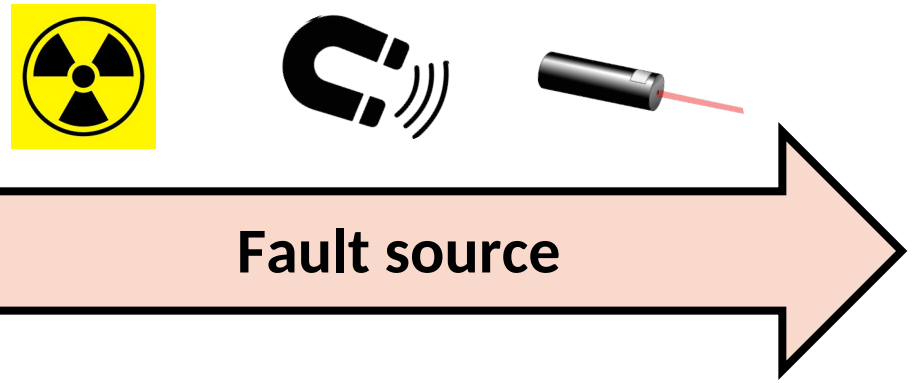


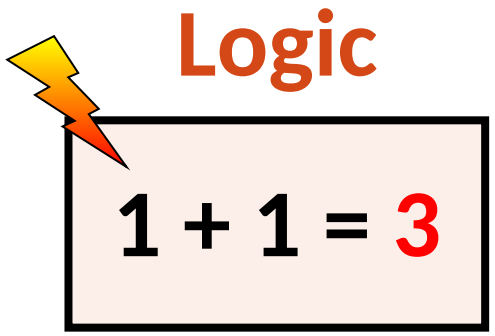
Hardware is vulnerable to faults



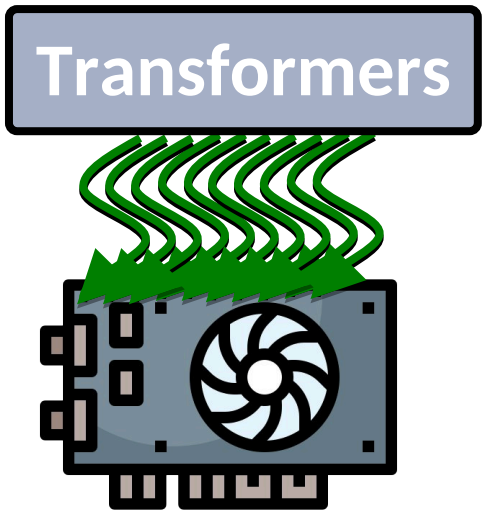
Hardware is vulnerable to faults

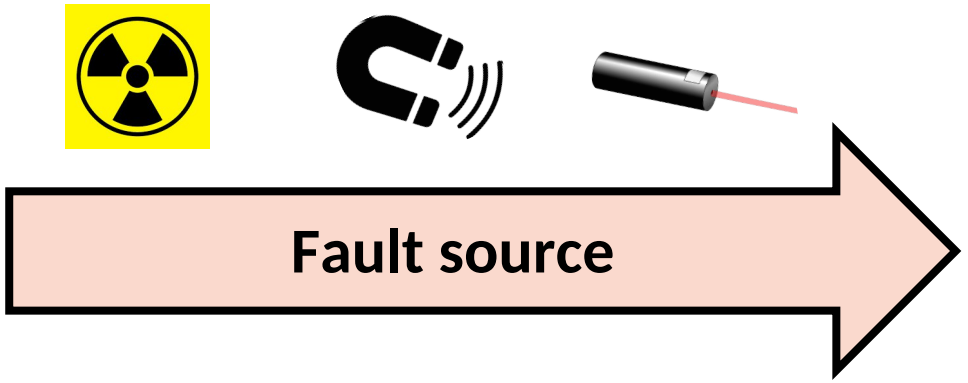


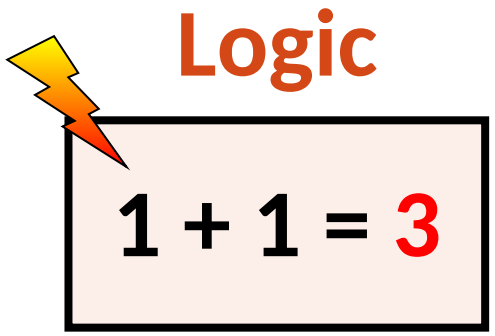


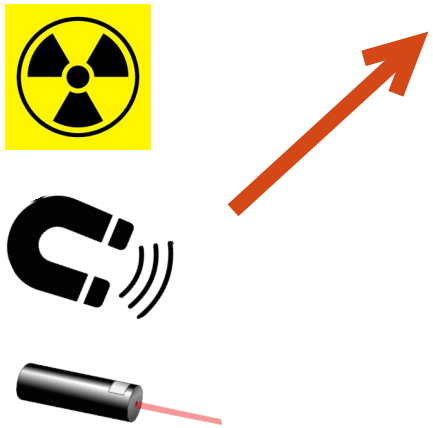


Hardware is vulnerable to faults





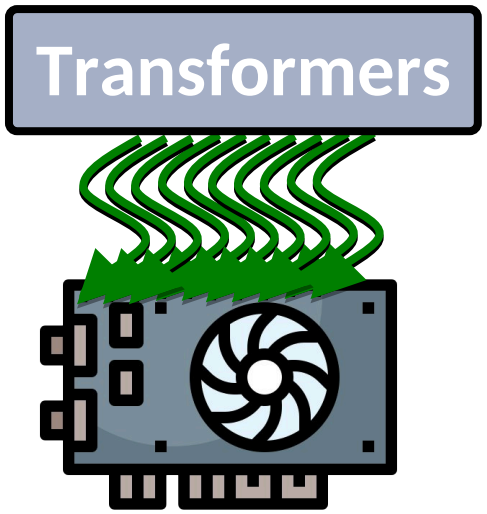


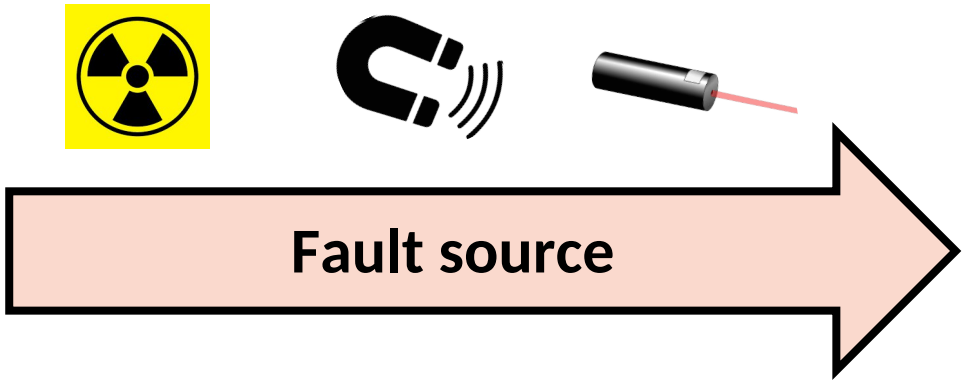


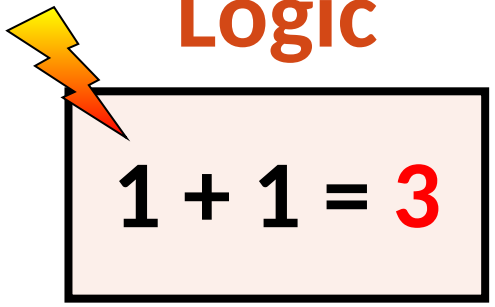
Silent Data Corruption (SDC)

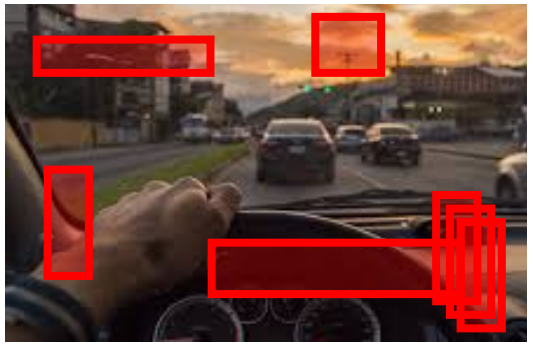
- **Computation is done but result is altered**

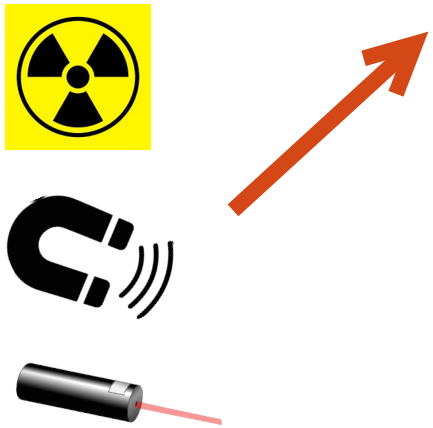
Hardware is vulnerable to faults







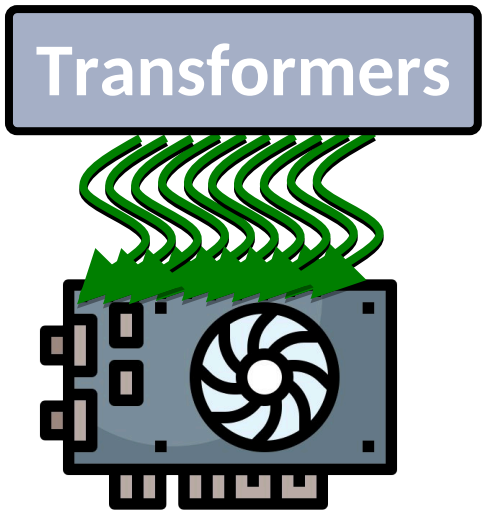


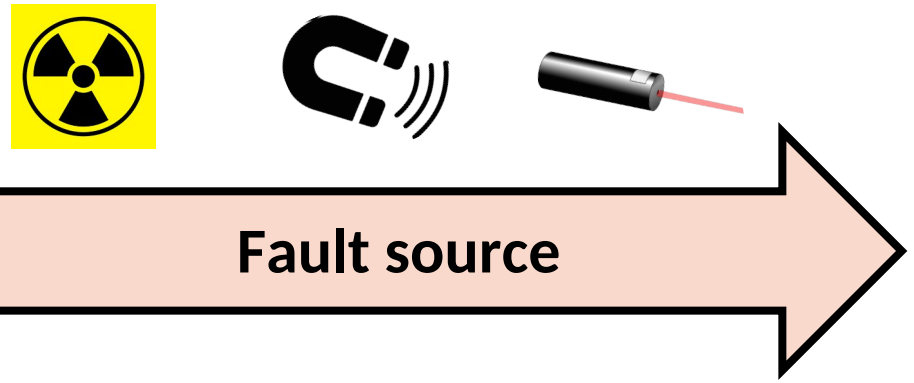


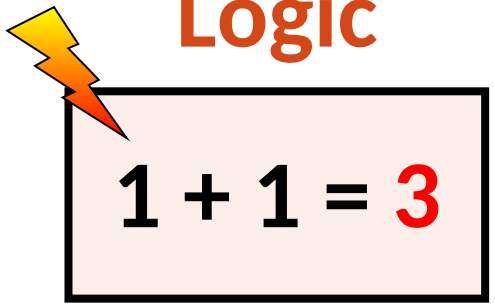
Silent Data Corruption (SDC)

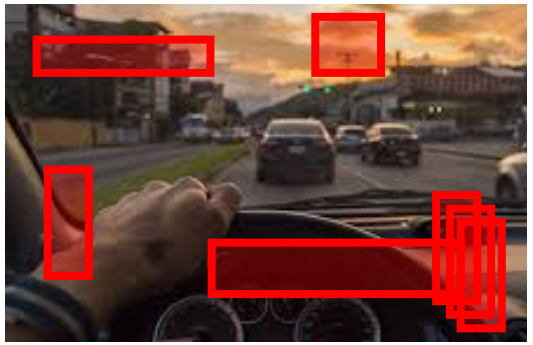
- **Computation is done but result is altered**

Hardware is vulnerable to faults









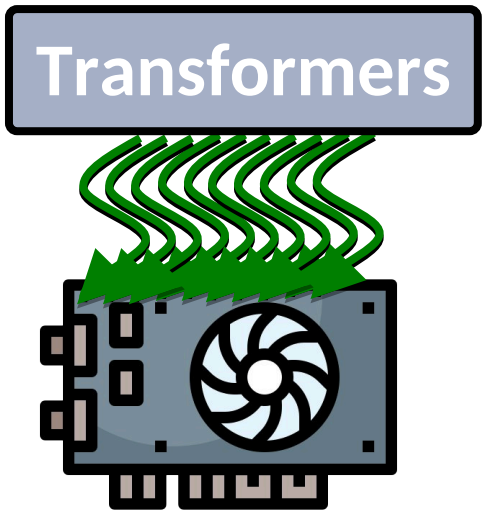
Silent Data Corruption (SDC)

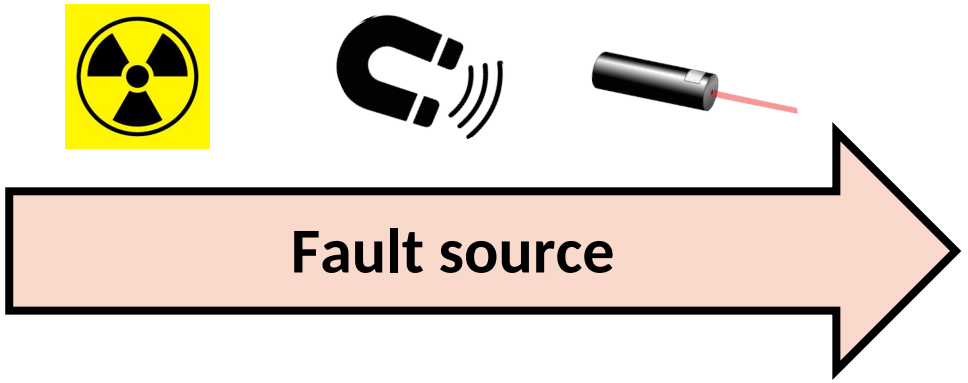
- Computation is done but result is altered

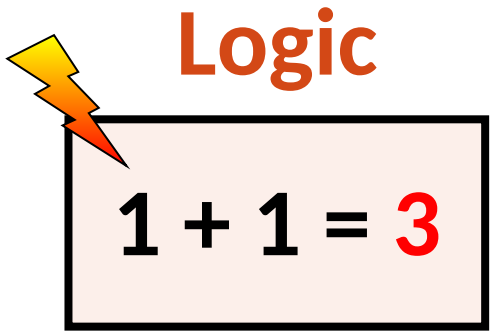
Detected Unrecoverable Error (DUE)

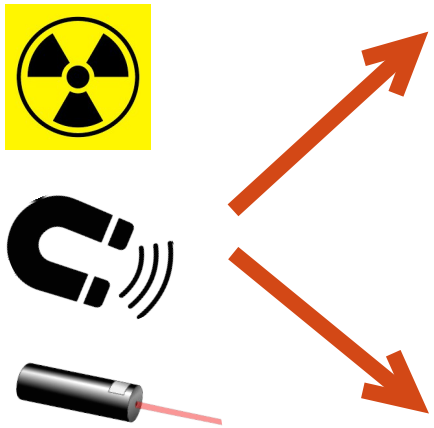
- Crash and OS hangs

Hardware is vulnerable to faults



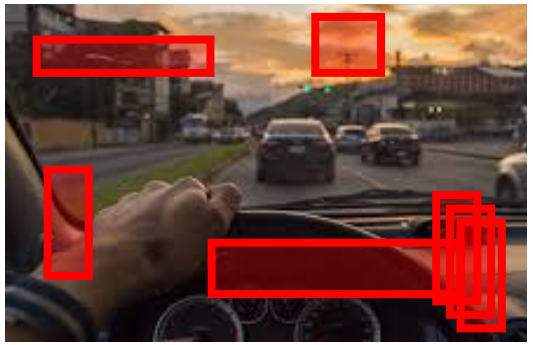






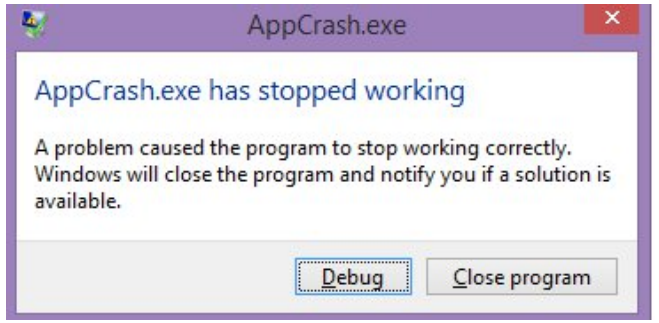
Silent Data Corruption (SDC)

- Computation is done but result is altered

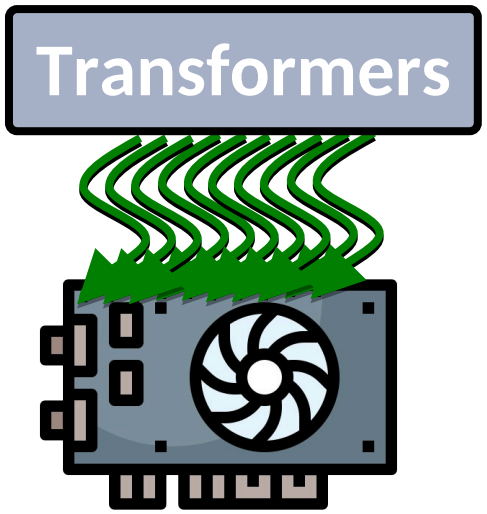


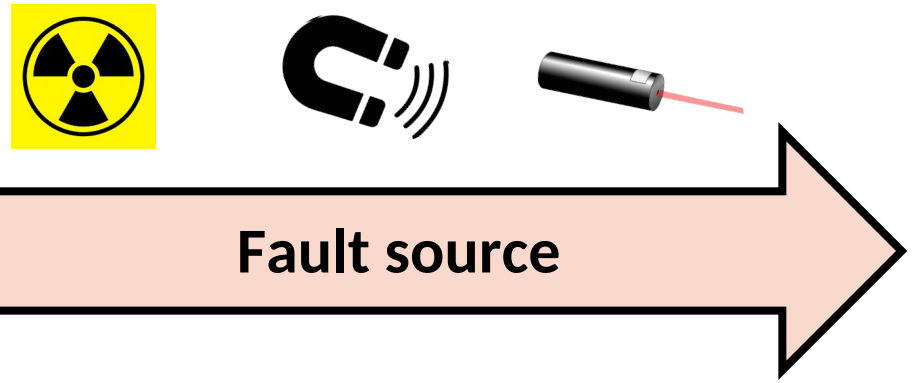
Detected Unrecoverable Error (DUE)

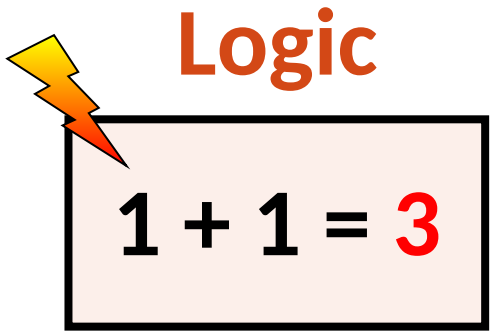
- Crash and OS hangs



Hardware is vulnerable to faults



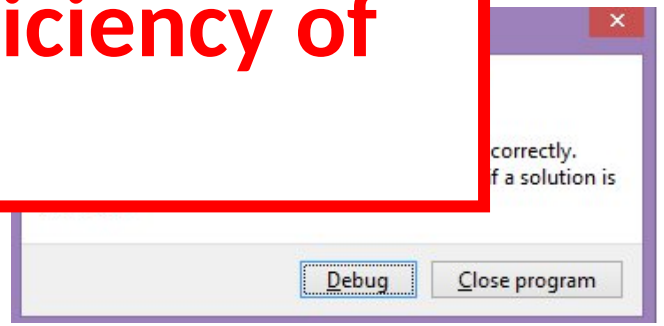




Silent Data Corruption (SDC)

Among the fault tolerance techniques for Transformers, we assess the efficiency of activation clipping

- Crash and OS hangs



Agenda

- Background
- Transformers Activation Value Analysis
- Activation Clipping Efficiency Assessment
- Conclusions and Future Work

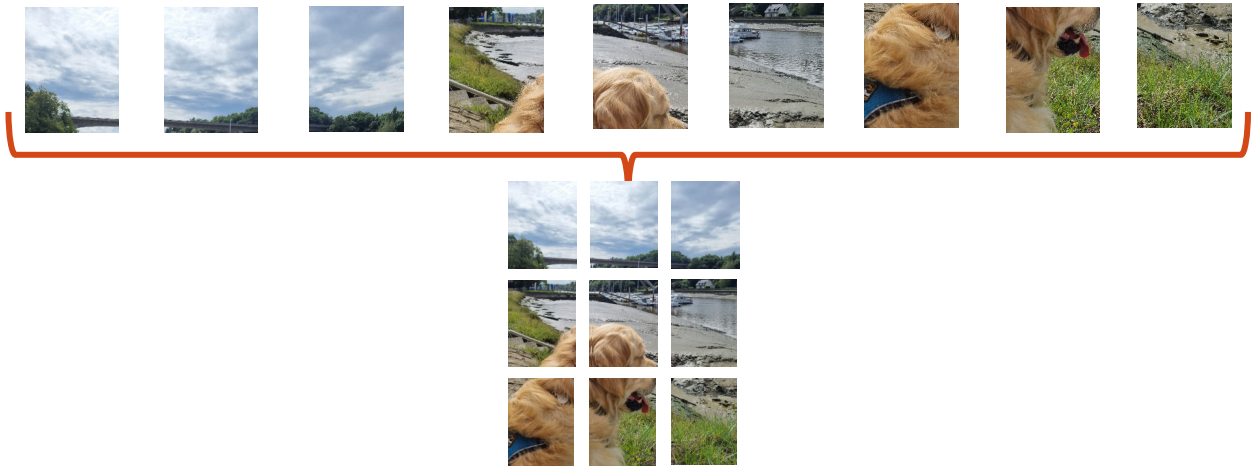
Background

AND ASSESSED TRANSFORMERS

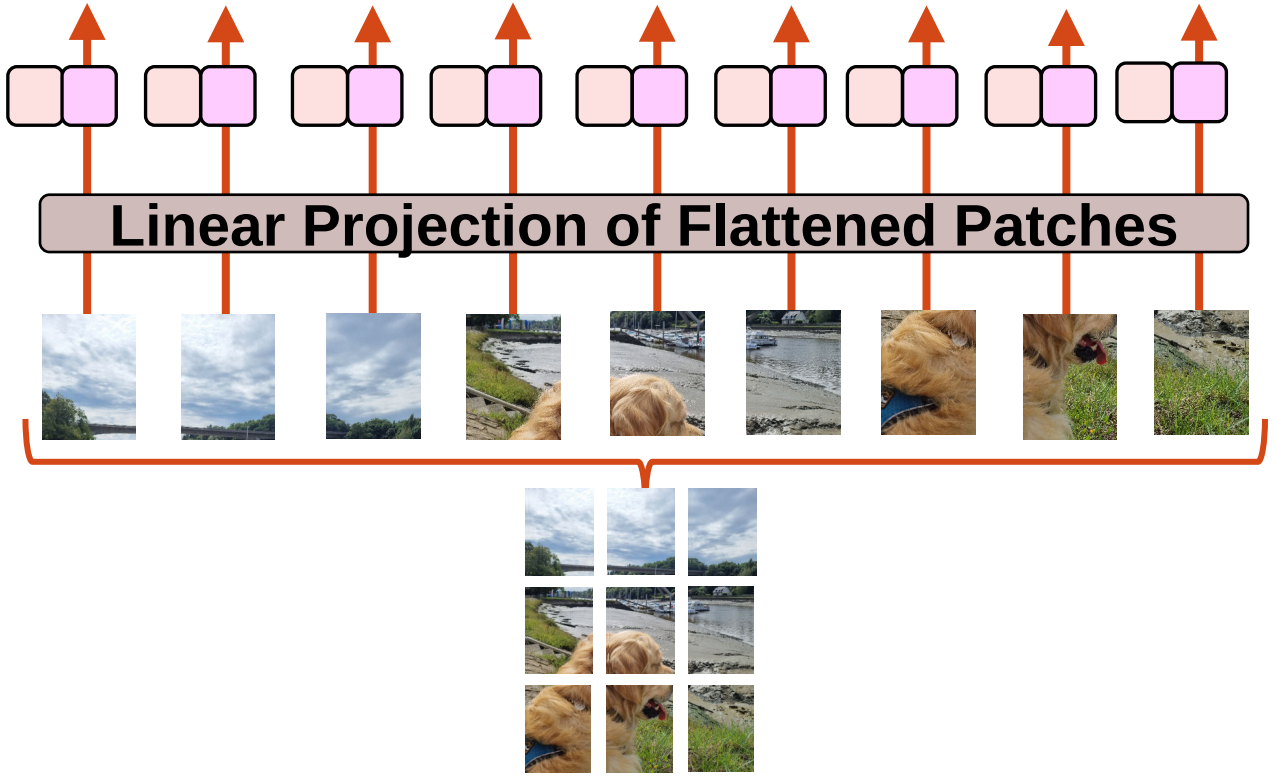
Transformers architecture example

Transformers architecture example

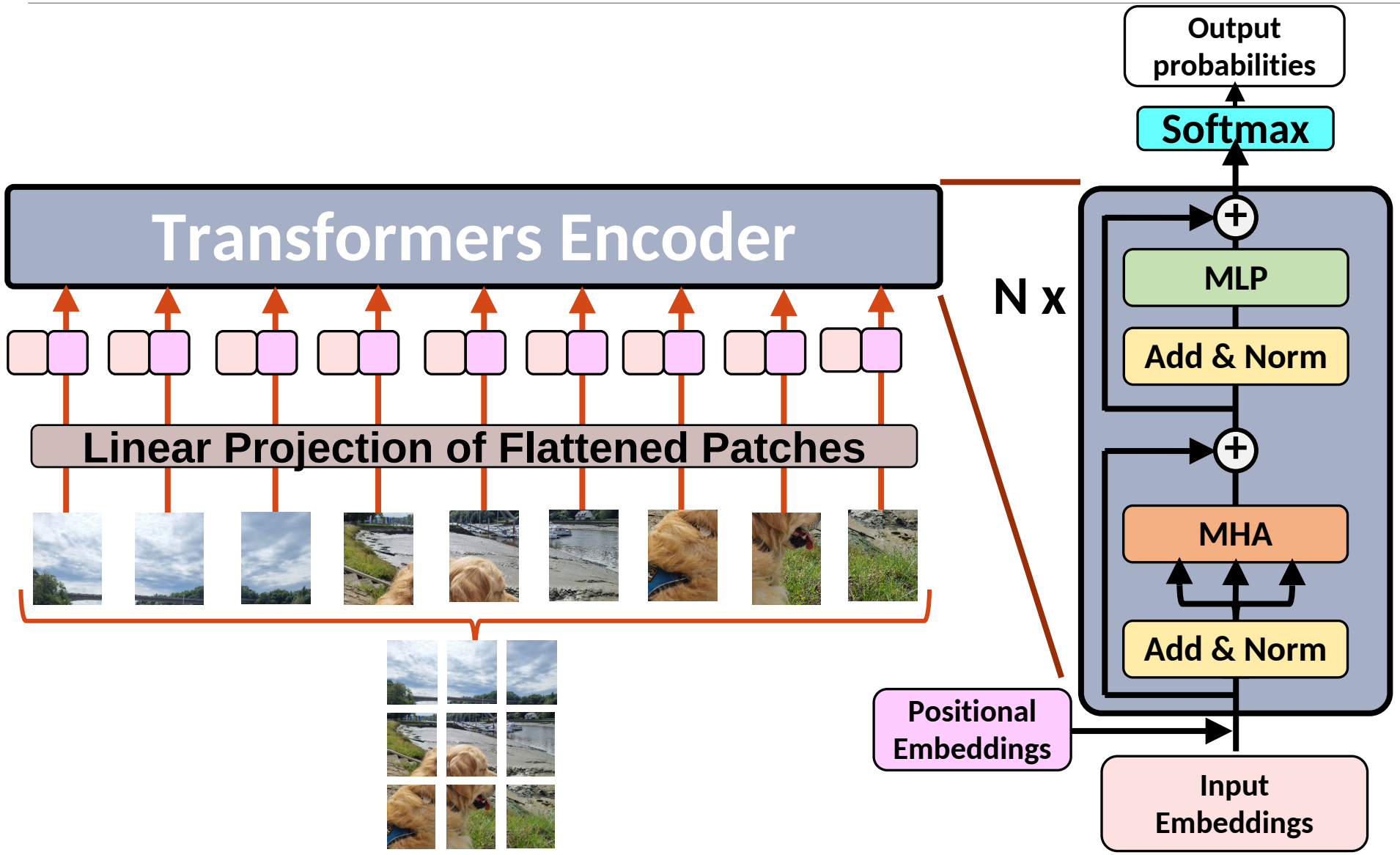
Transformers architecture example



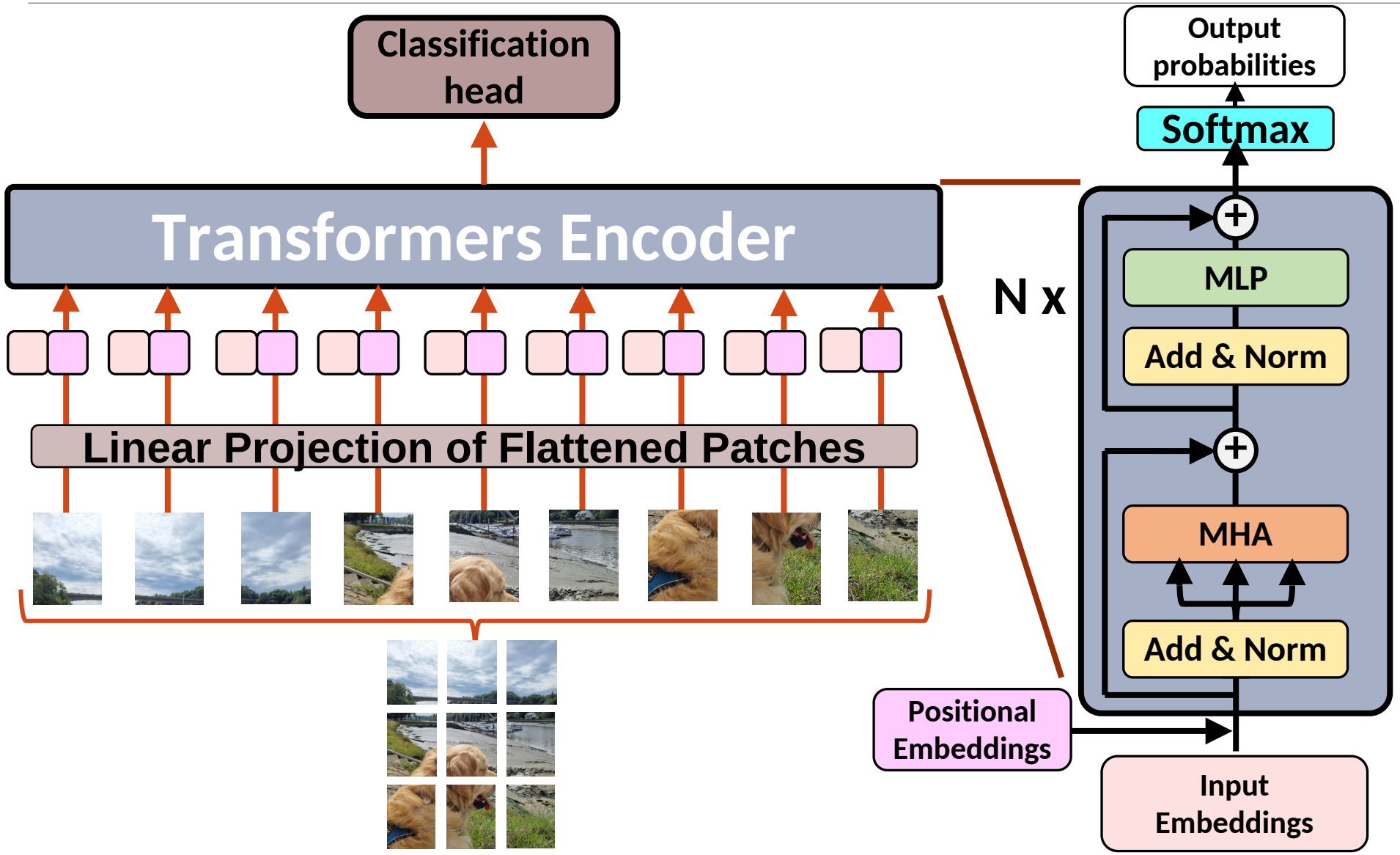
Transformers architecture example



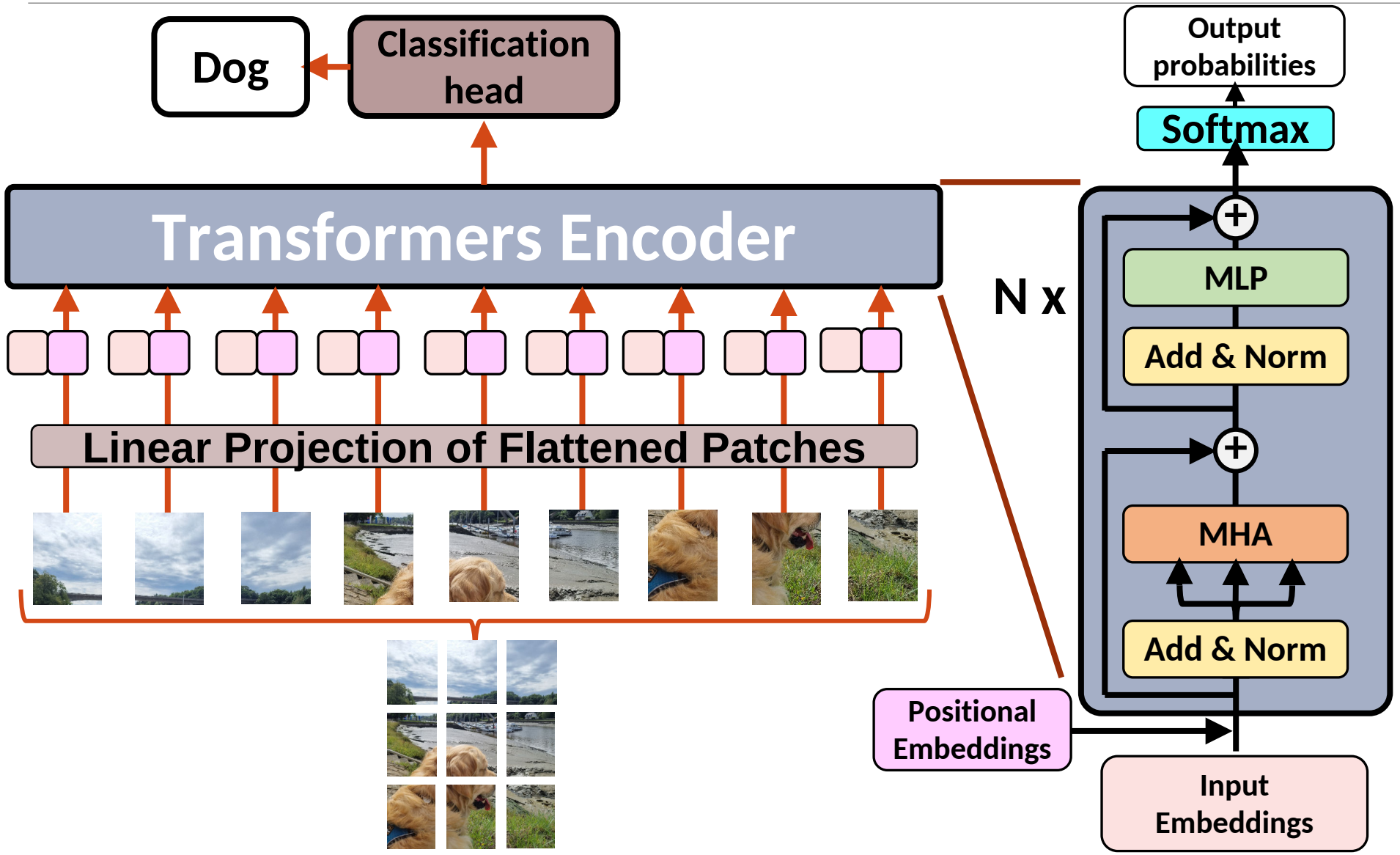
Transformers architecture example



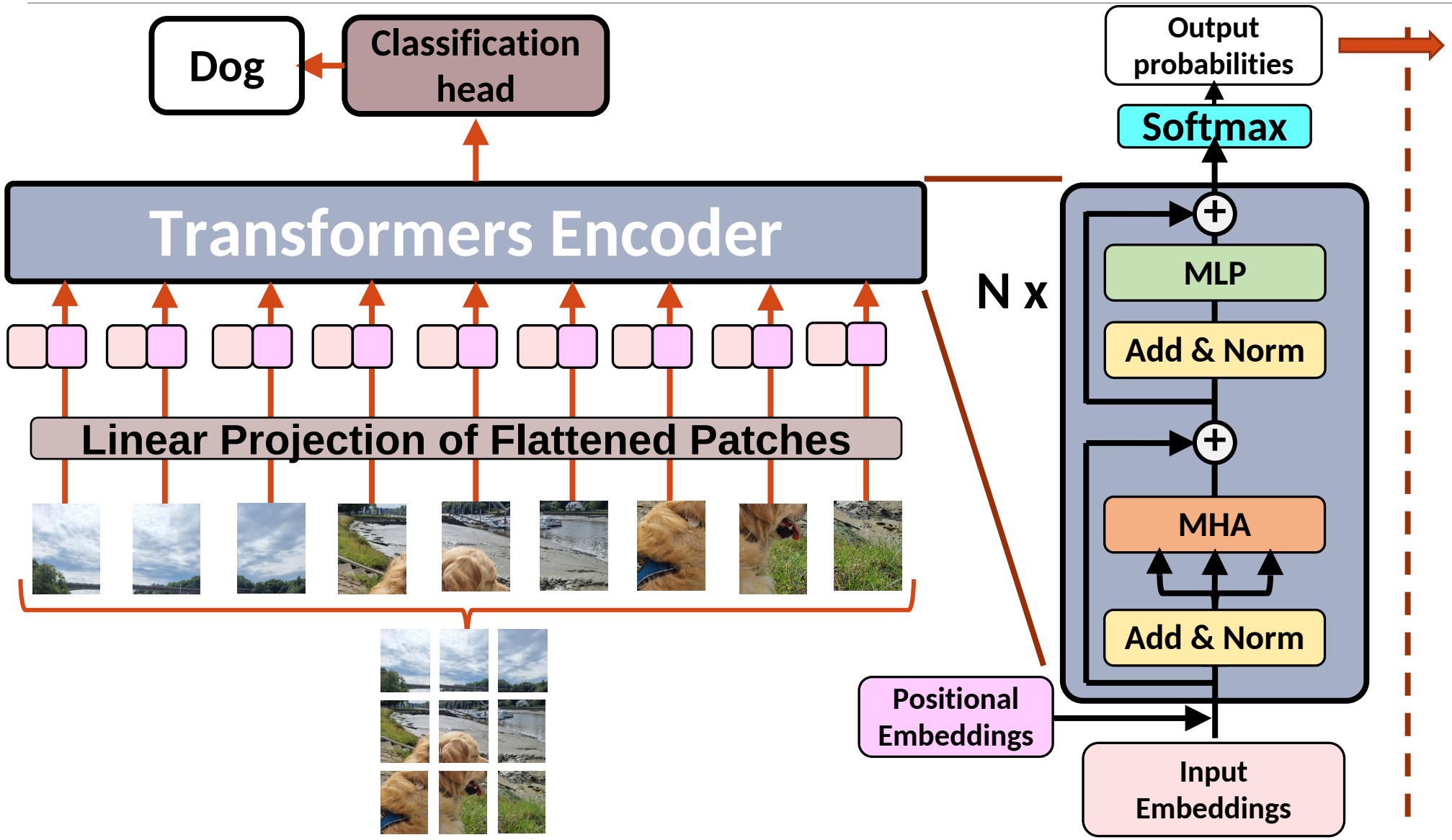
Transformers architecture example



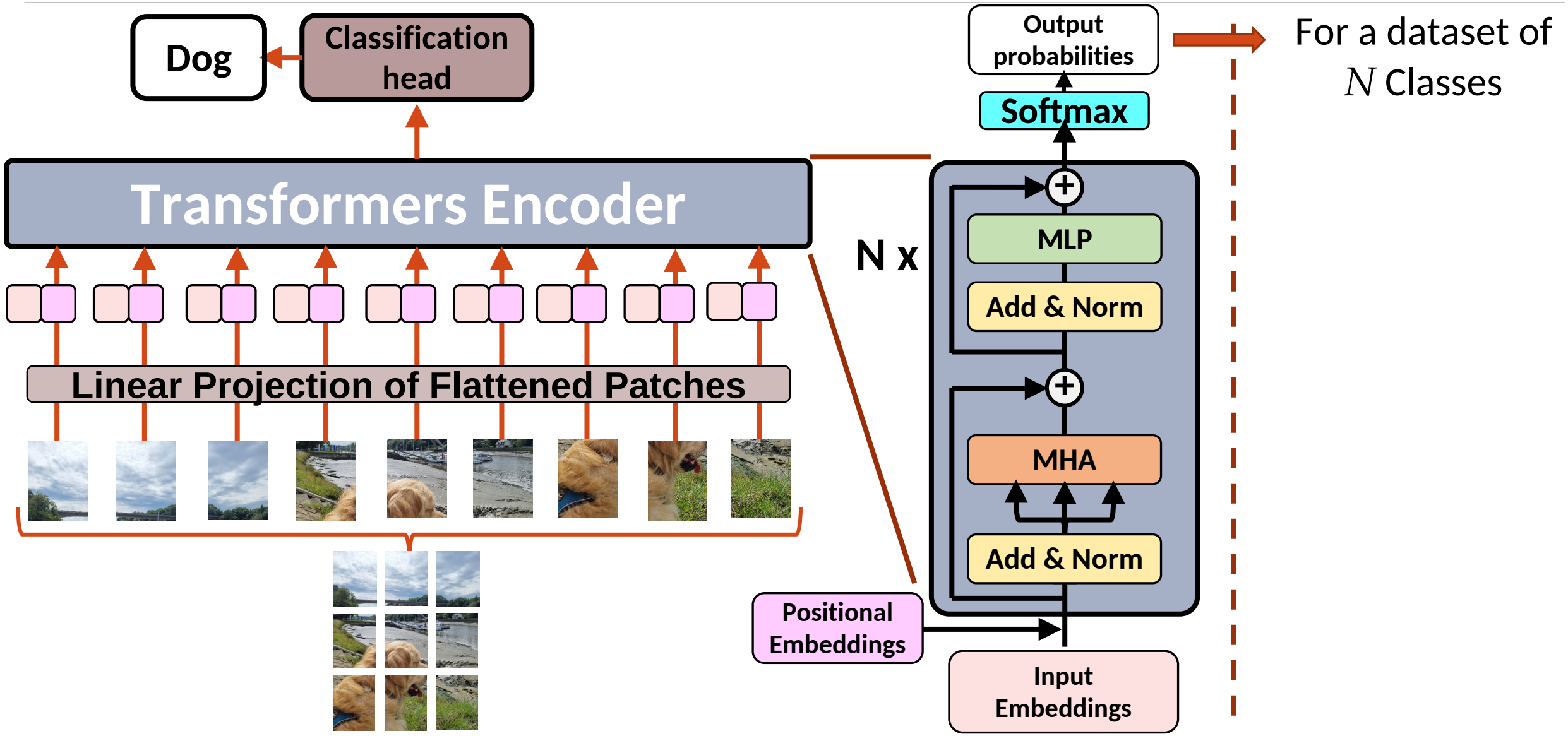
Transformers architecture example



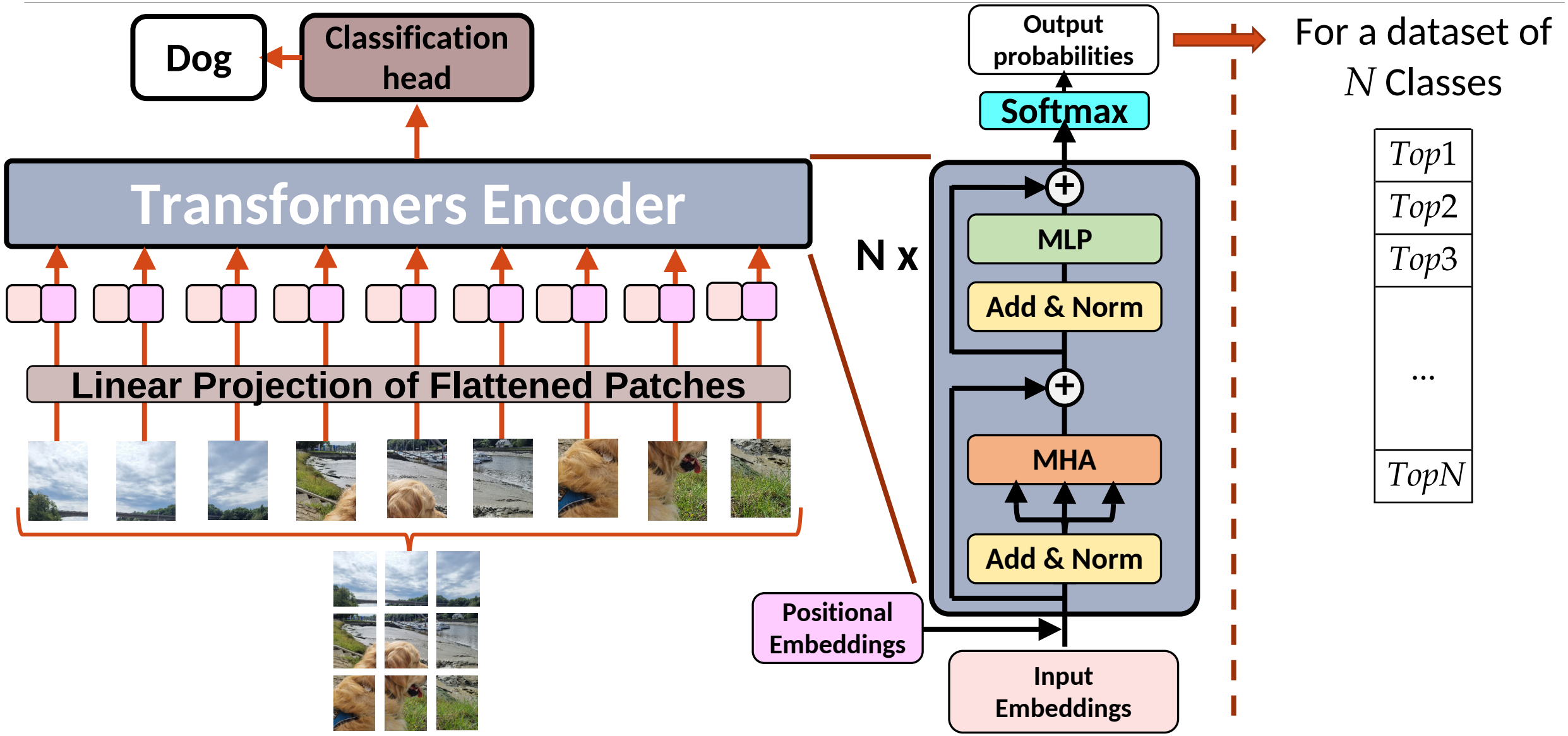
Transformers architecture example



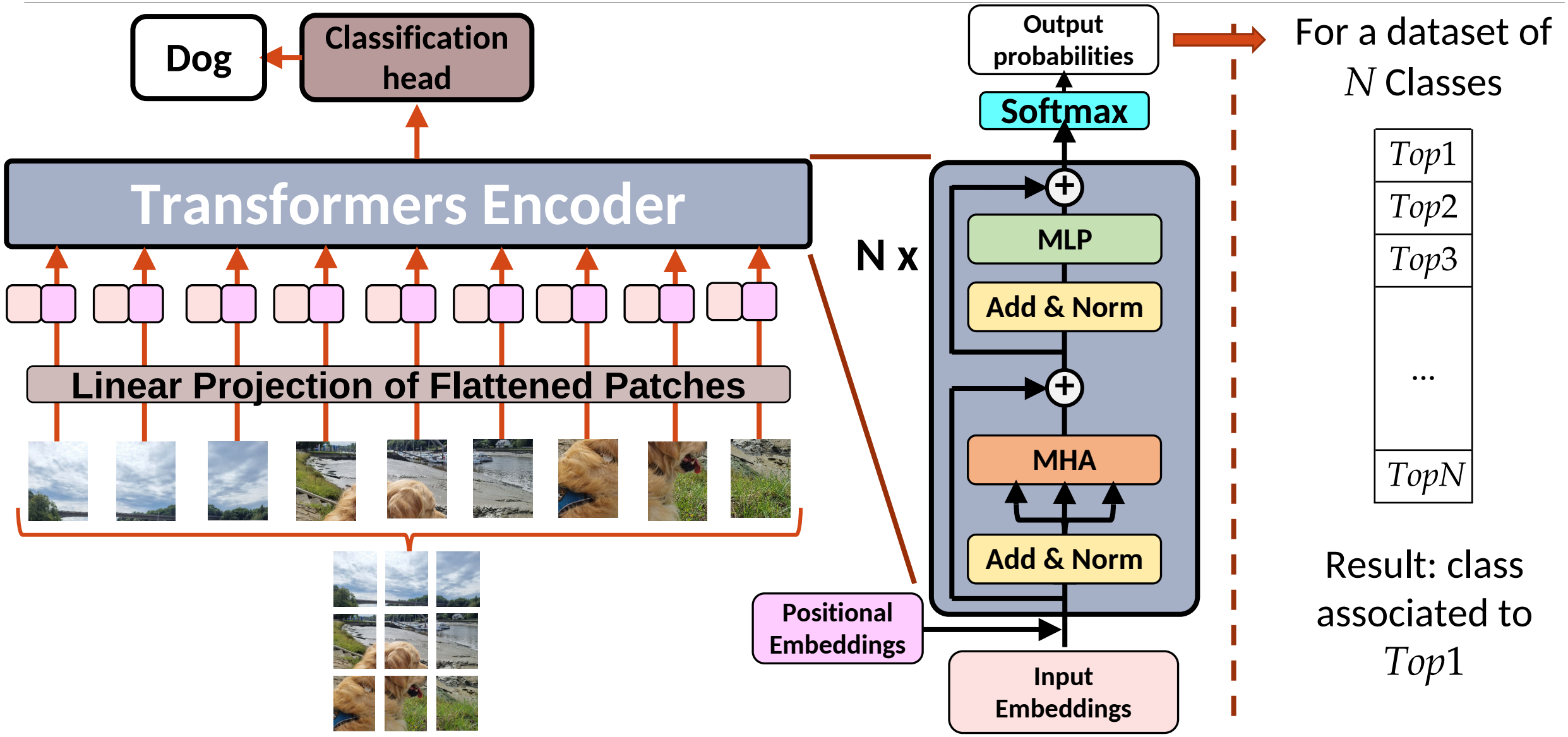
Transformers architecture example



Transformers architecture example

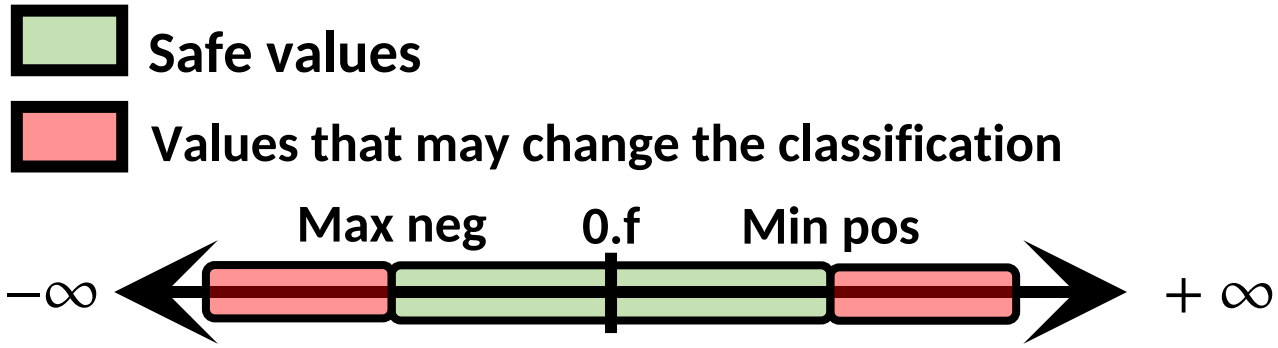


Transformers architecture example



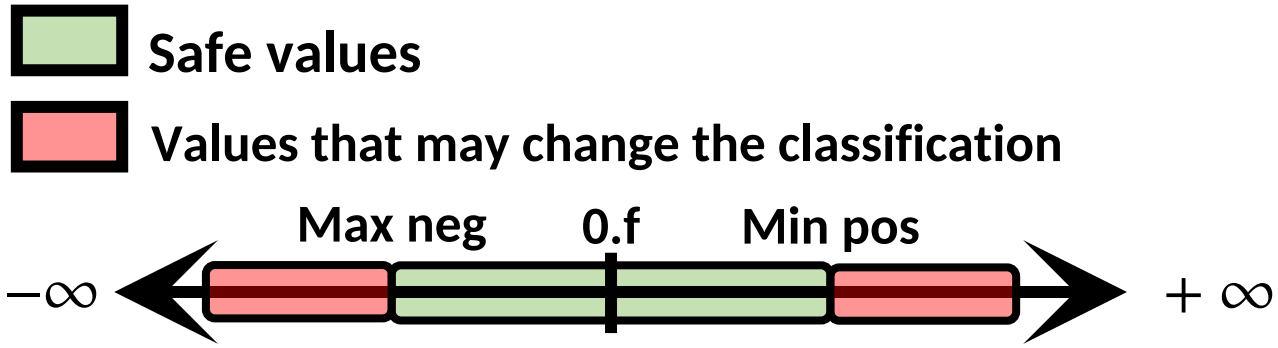
Activation Clipping

Clipping is the process of restriciting the range of values



Activation Clipping

Clipping is the process of restriciting the range of values

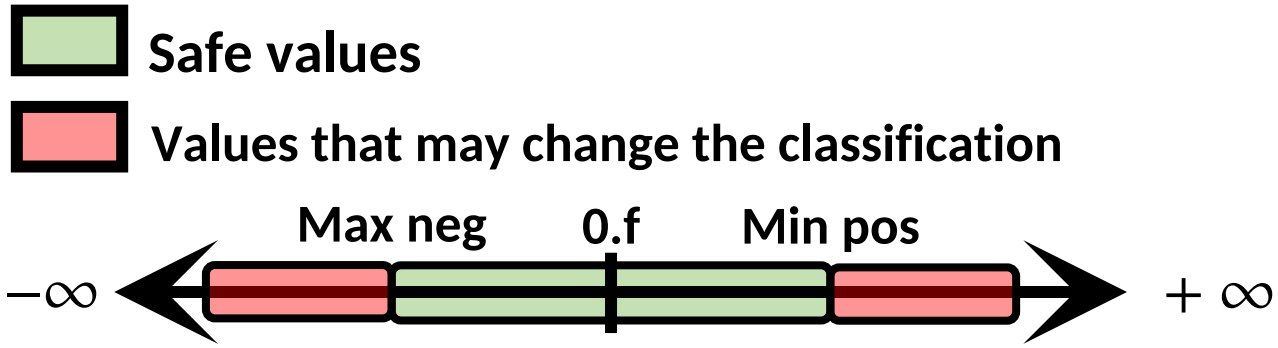


Different granularities

- Model wise: global bound
- Operation wise: Min/Max profiled by type of operation (MHA, MLP, Block)
- Layer wise: Min/Max profiled for each individual layer
- Neuron wise: Min/Max profiled per neuron

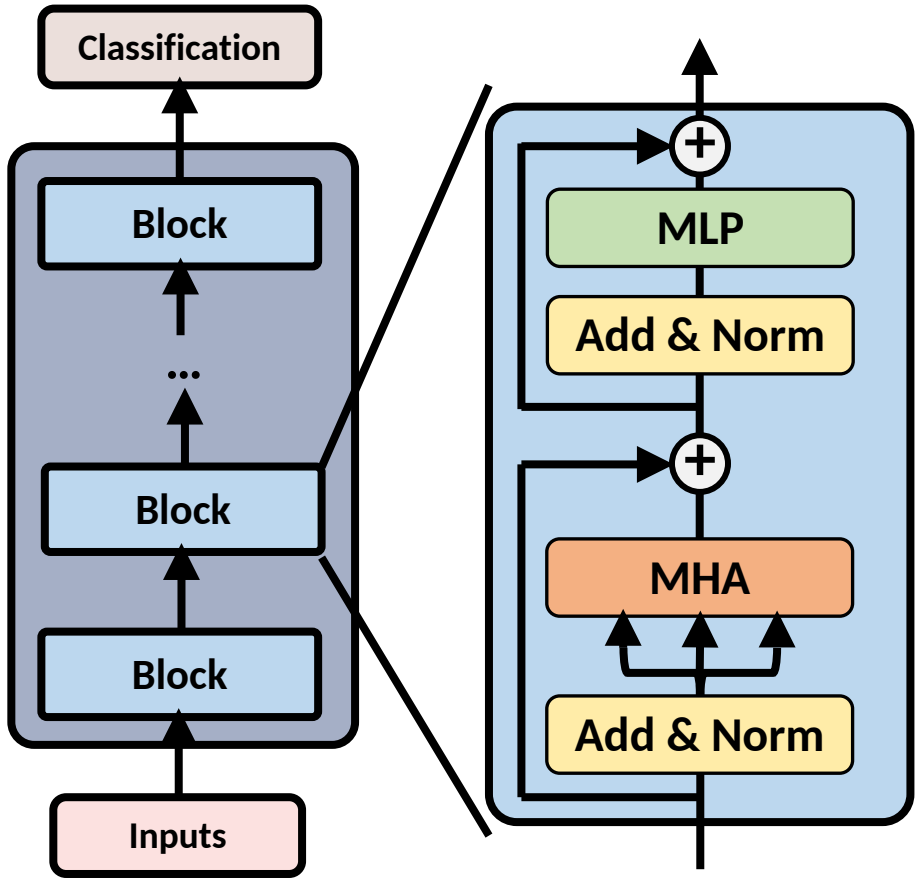
Activation Clipping

Clipping is the process of restriciting the range of values



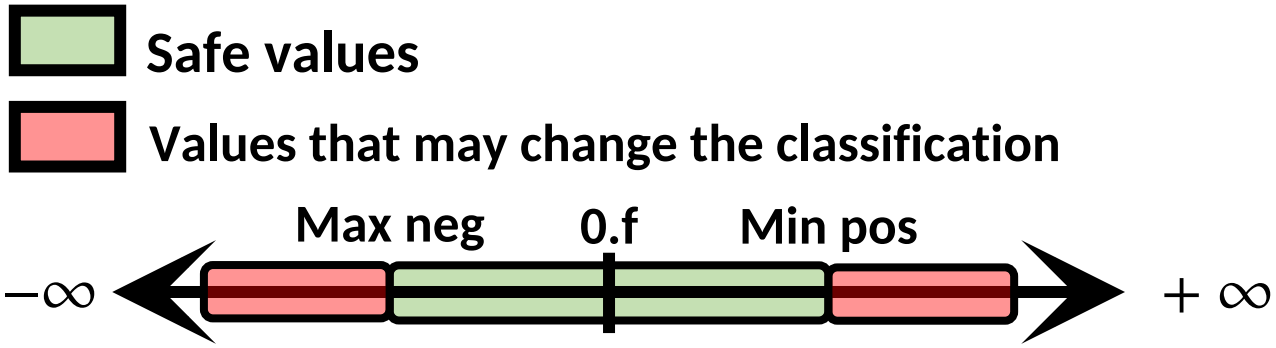
Different granularities

- Model wise: global bound
- Operation wise: Min/Max profiled by type of operation (MHA, MLP, Block)
- Layer wise: Min/Max profiled for each individual layer
- Neuron wise: Min/Max profiled per neuron



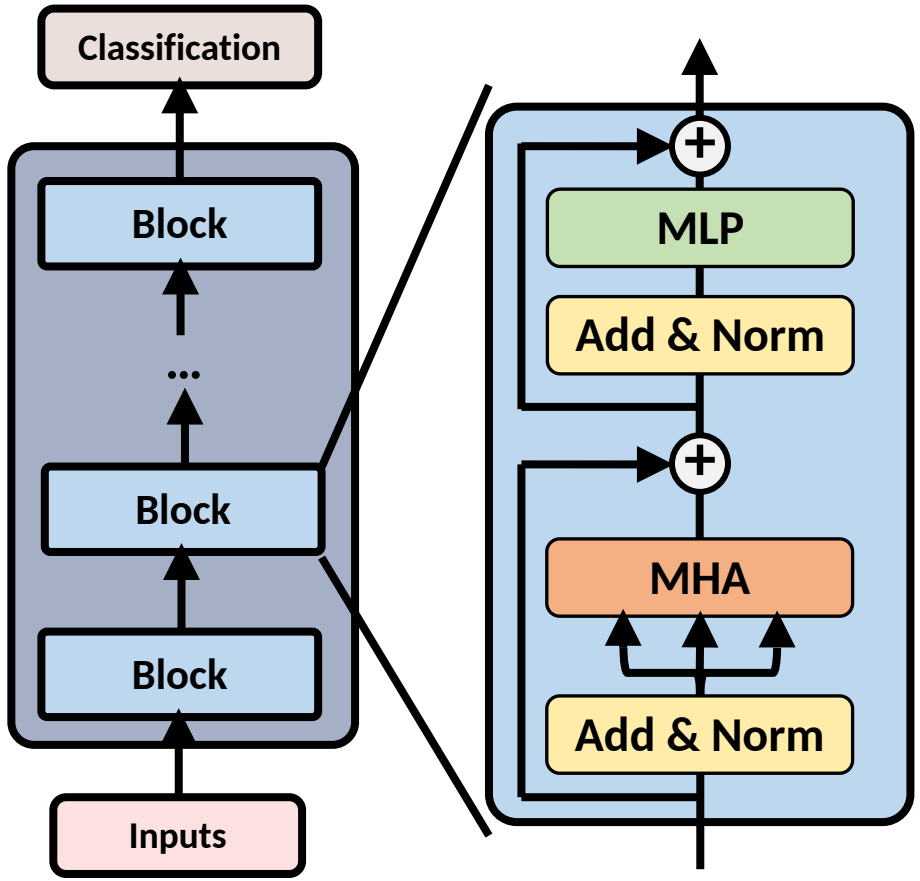
Activation Clipping

Clipping is the process of restriciting the range of values



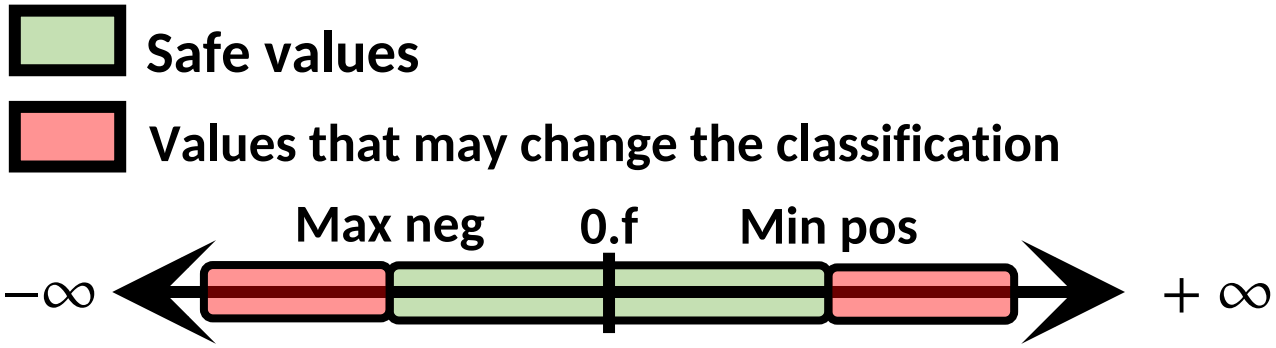
Different granularities

- Model wise: global bound
- Operation wise: Min/Max profiled by type of operation (MHA, MLP, Block)
- Layer wise: Min/Max profiled for each individual layer
- Neuron wise: Min/Max profiled per neuron



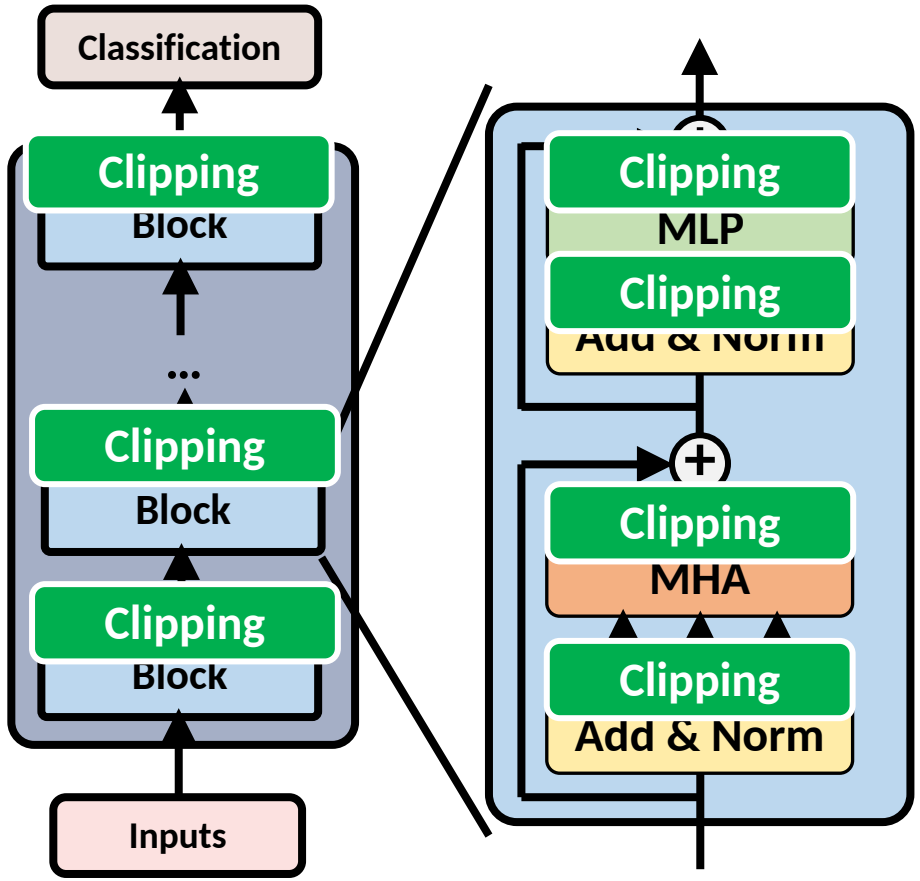
Activation Clipping

Clipping is the process of restriciting the range of values



Different granularities

- Model wise: global bound
- Operation wise: Min/Max profiled by type of operation (MHA, MLP, Block)
- Layer wise: Min/Max profiled for each individual layer
- Neuron wise: Min/Max profiled per neuron



Assessed Transformers

4 Transformers, 2 Tasks

Assessed Transformers

4 Transformers, 2 Tasks

ViT & Swin: Image classification (ImageNet dataset)

Assessed Transformers

4 Transformers, 2 Tasks

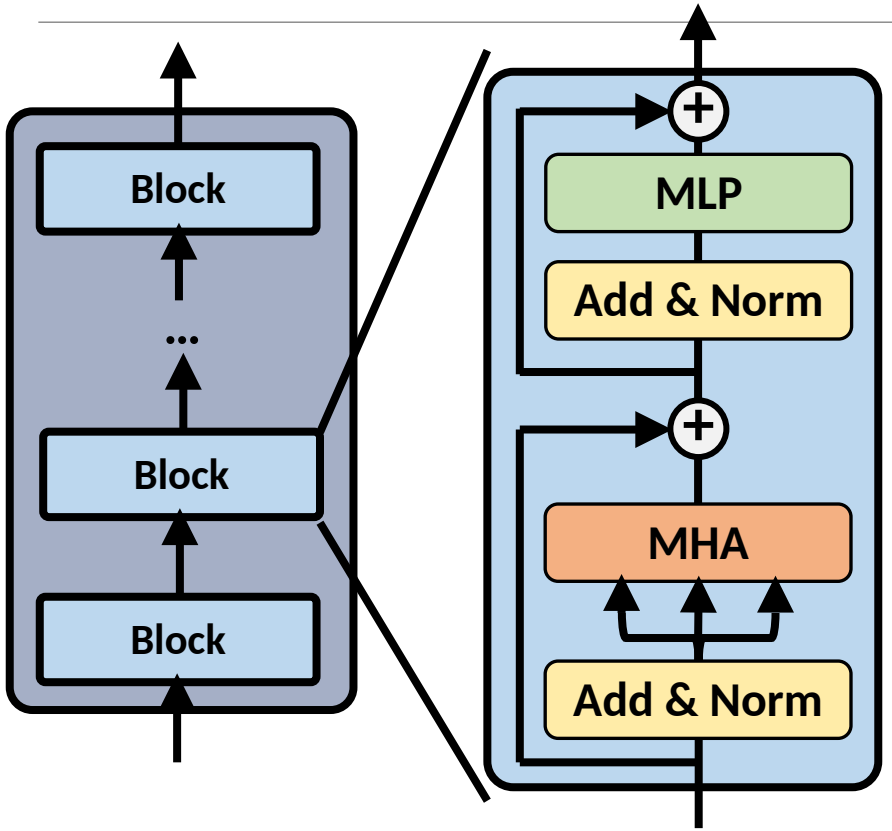
ViT & Swin: Image classification (ImageNet dataset)

BART & GPT2: Text classification (MNLI dataset)

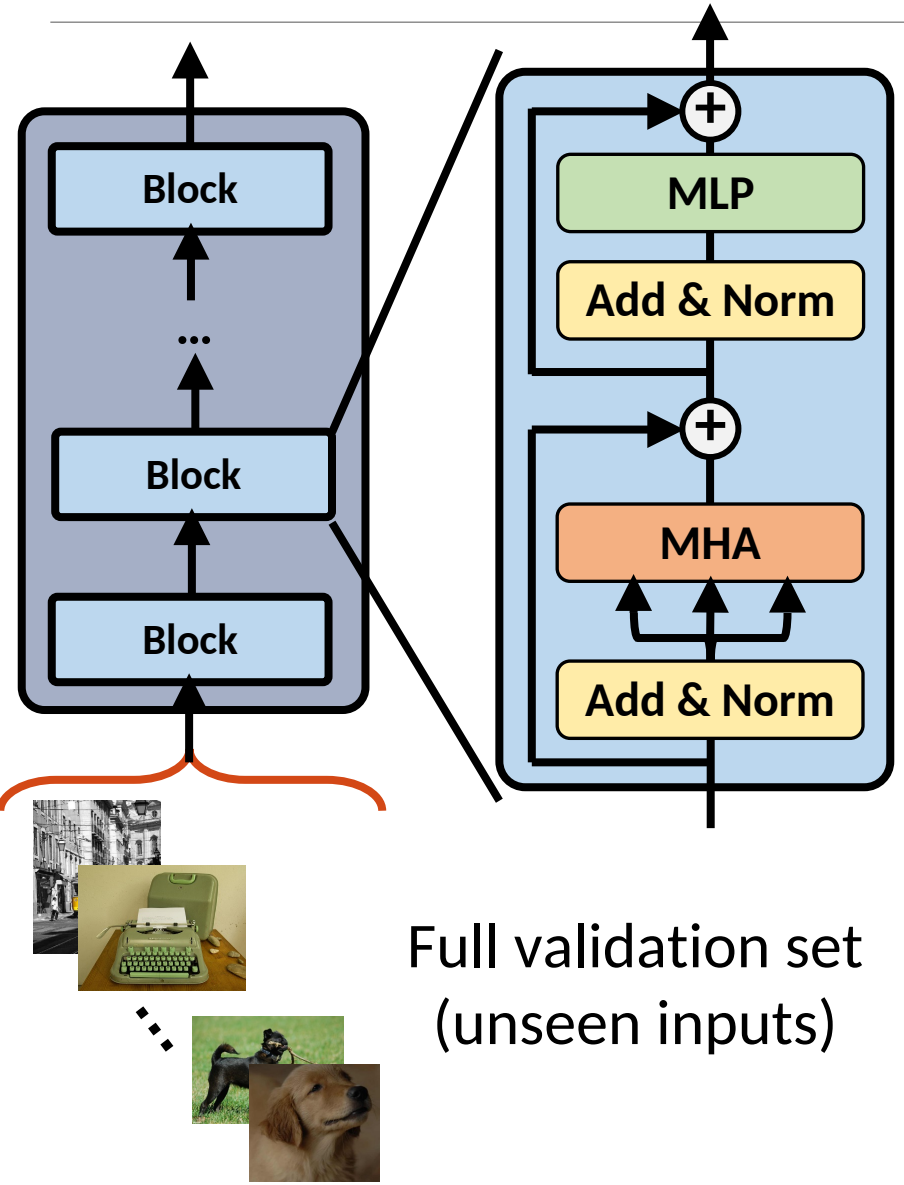
Hypothesis: The dog is sitting in the grass
Premise: An animal is moving

Activation value analysis

Activation value profiling

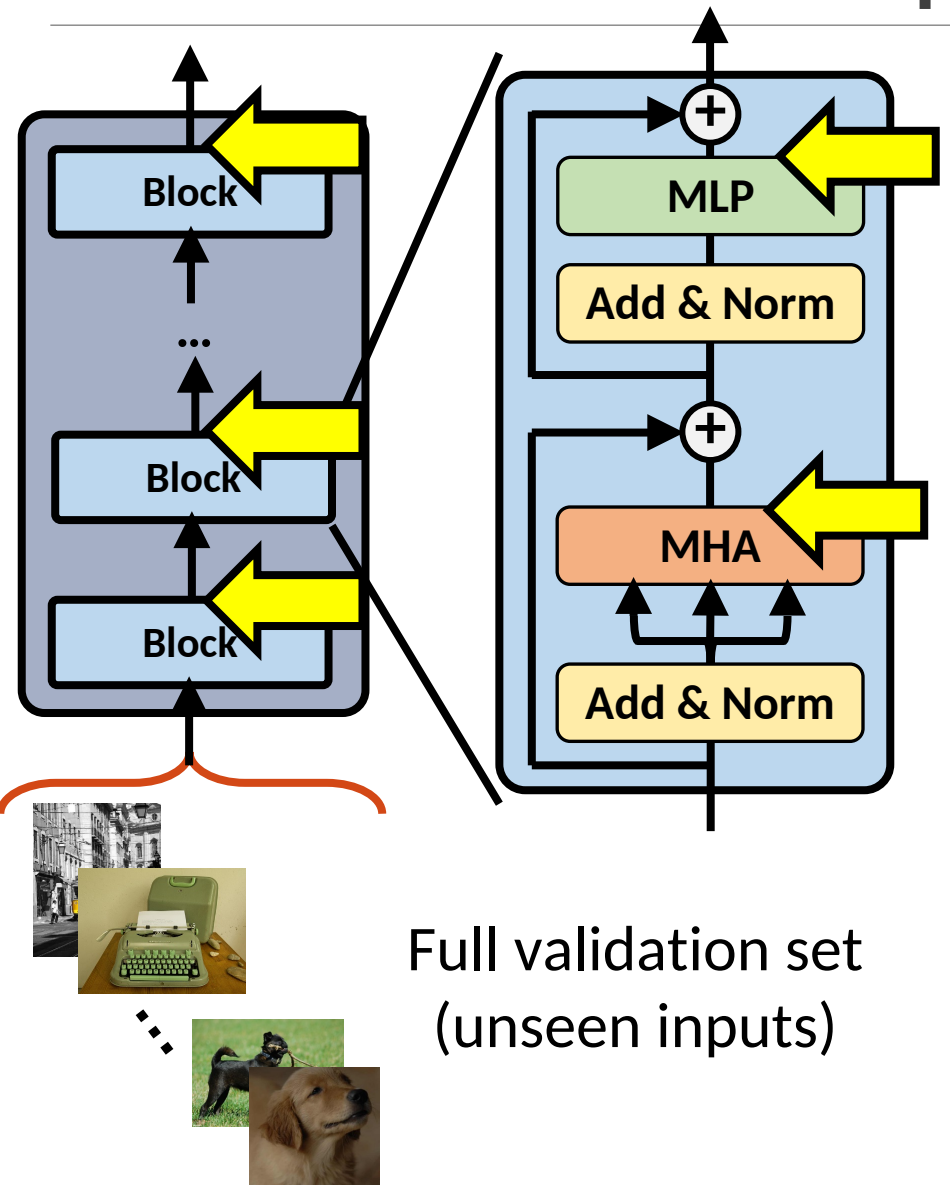


Activation value profiling



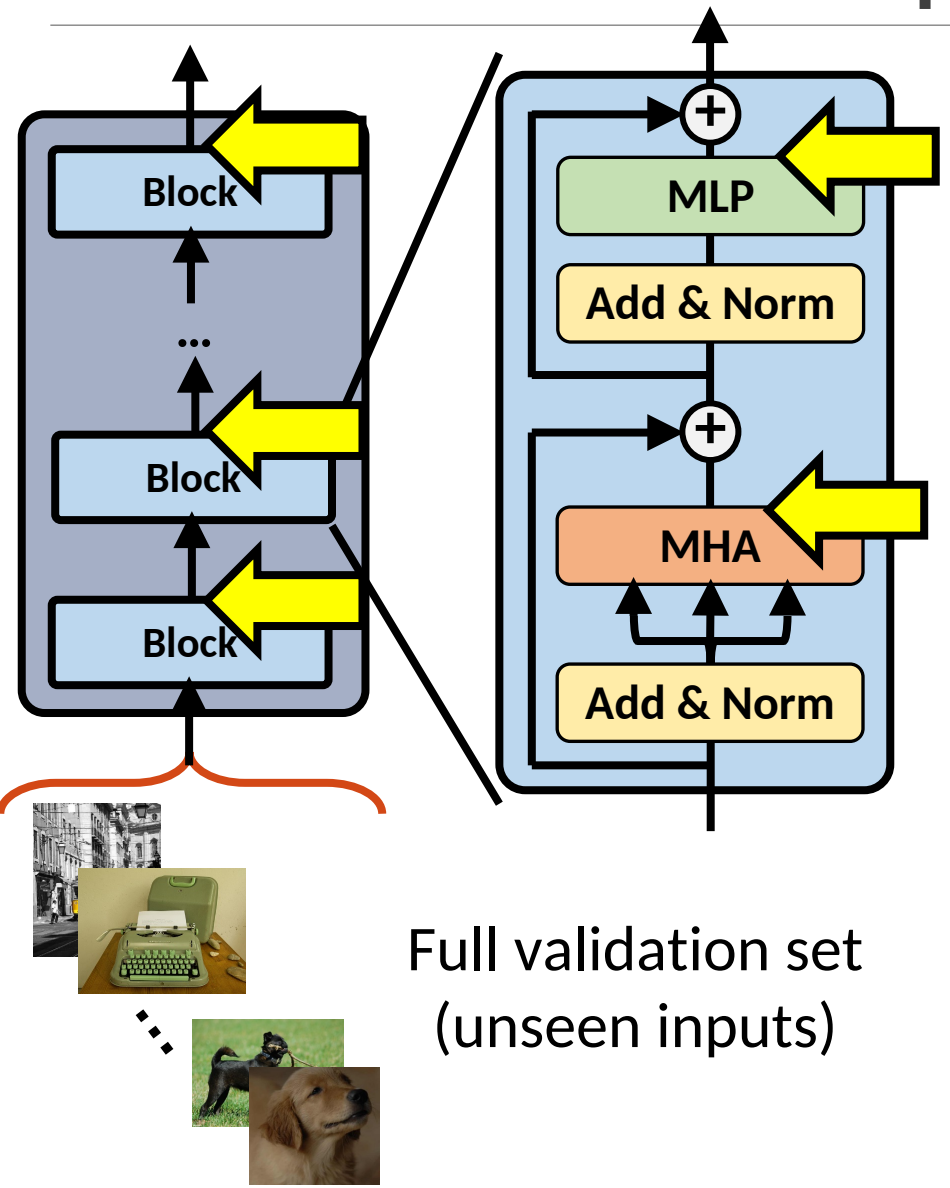
Full validation set
(unseen inputs)

Activation value profiling



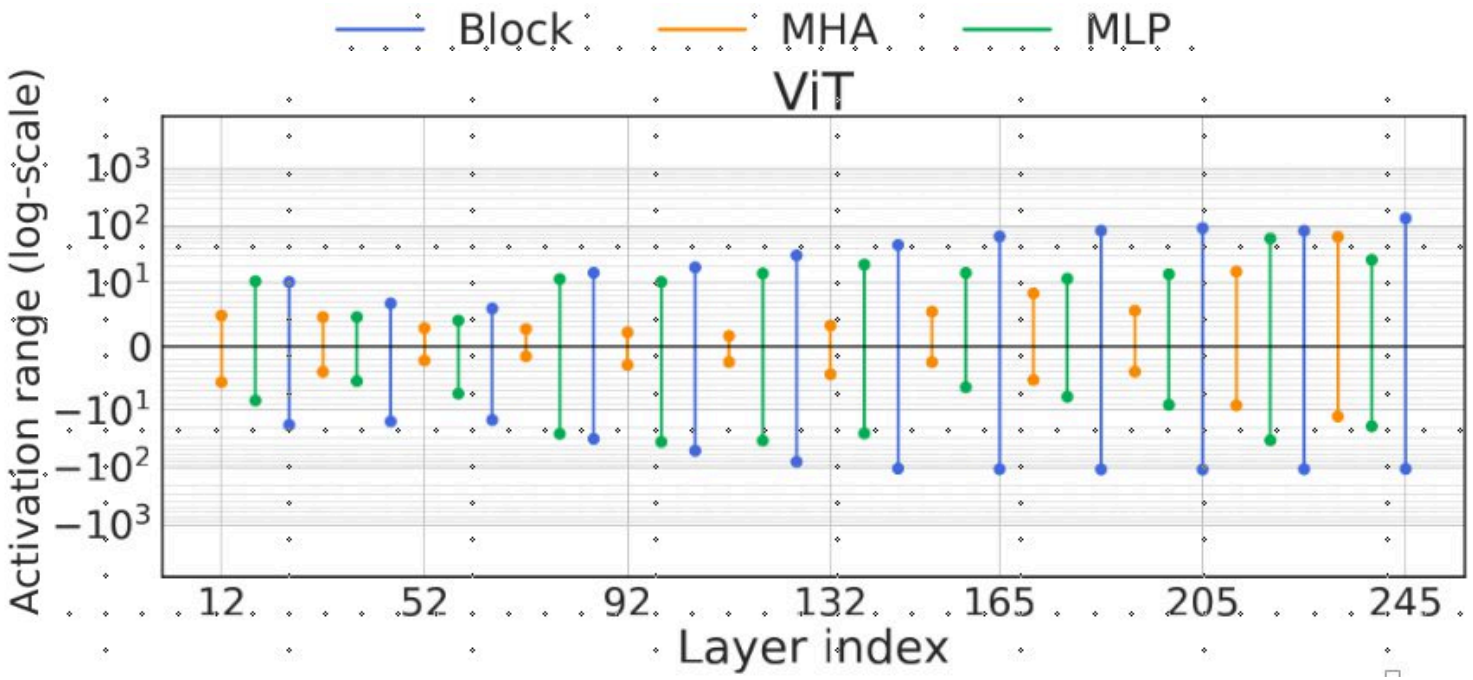
Full validation set
(unseen inputs)

Activation value profiling

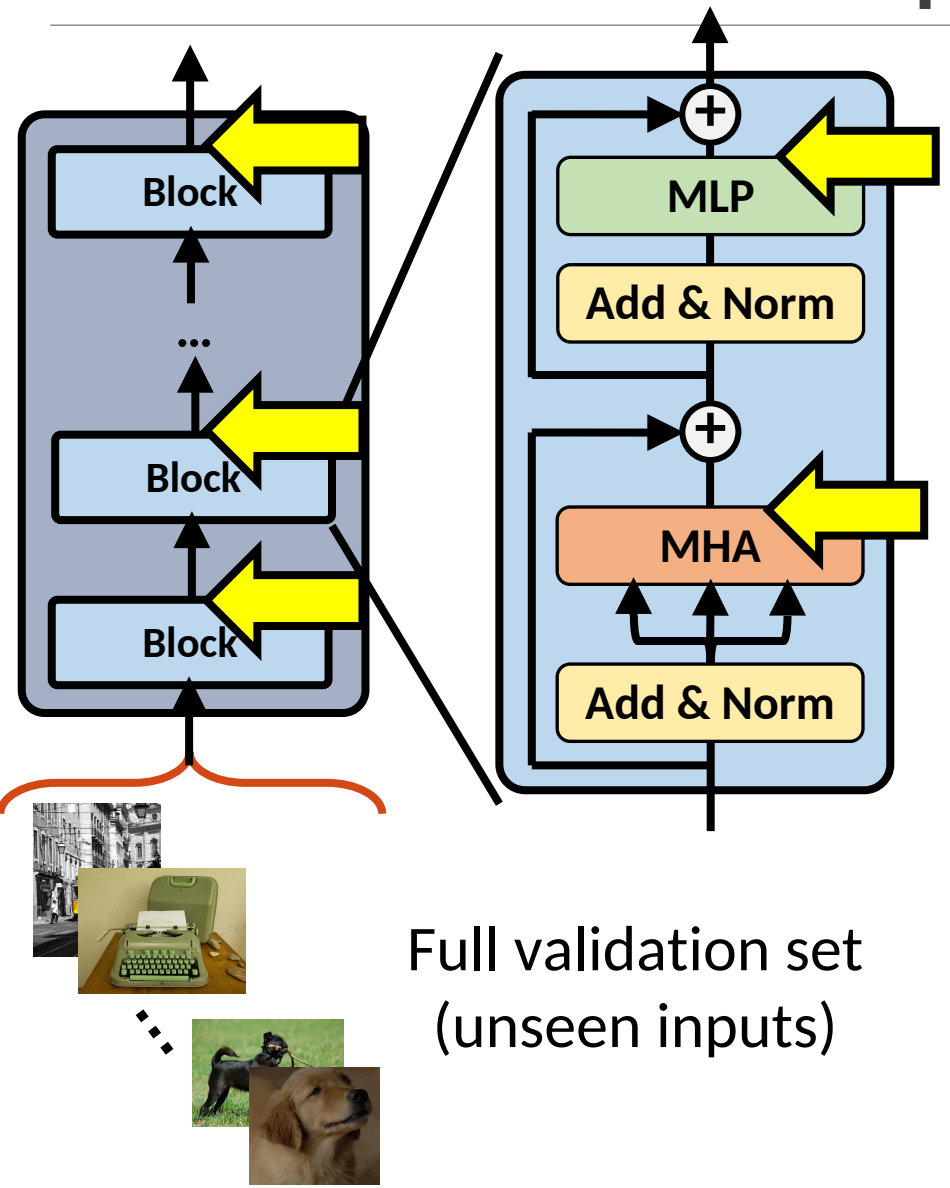


Full validation set
(unseen inputs)

Activation ranges

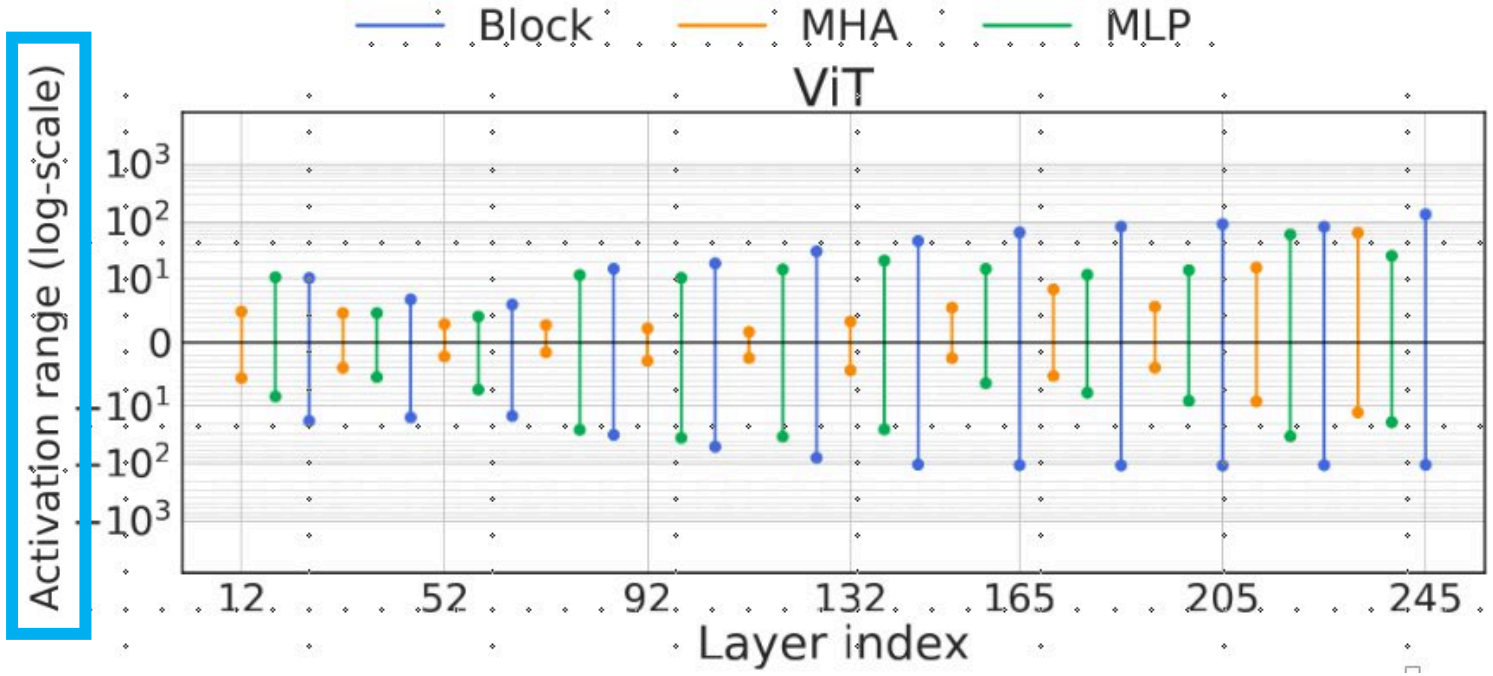


Activation value profiling

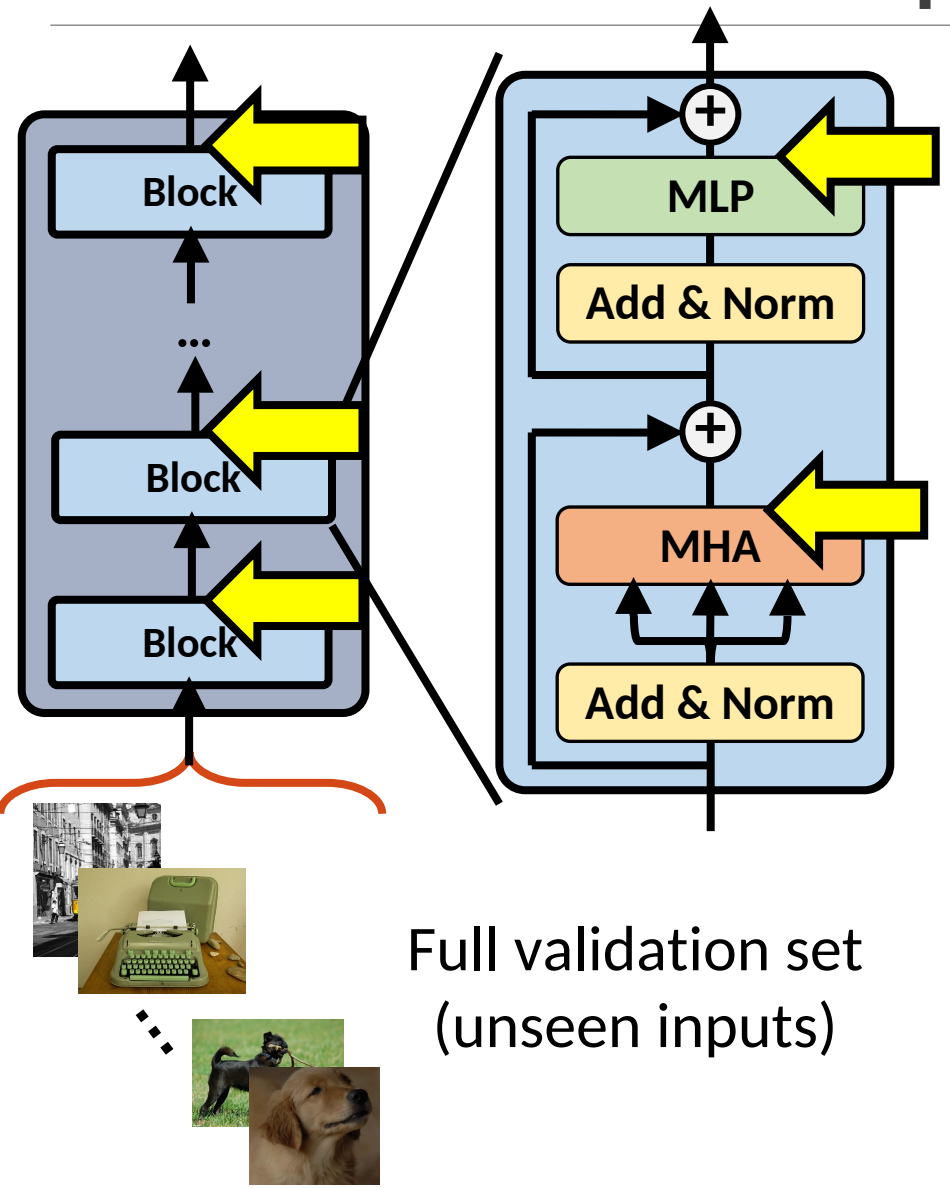


Full validation set
(unseen inputs)

Activation ranges

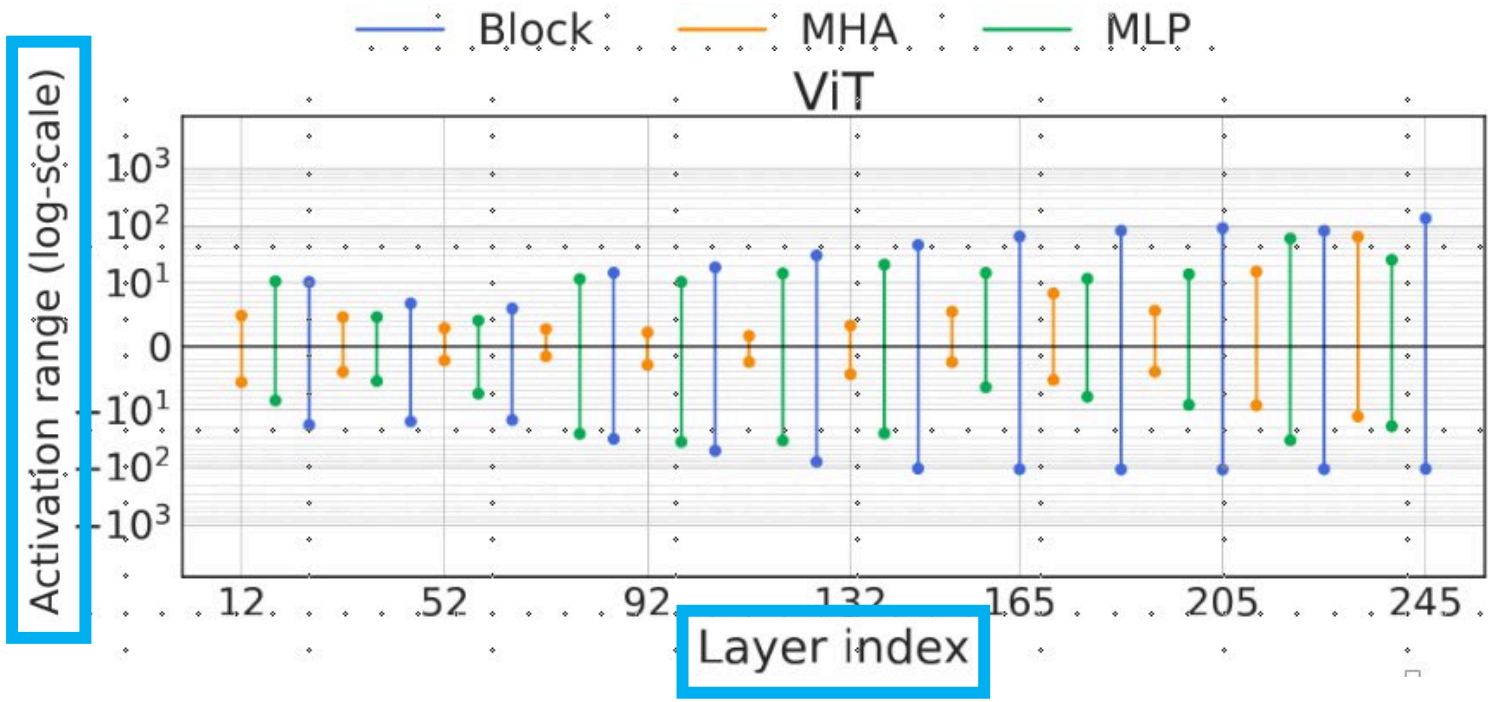


Activation value profiling

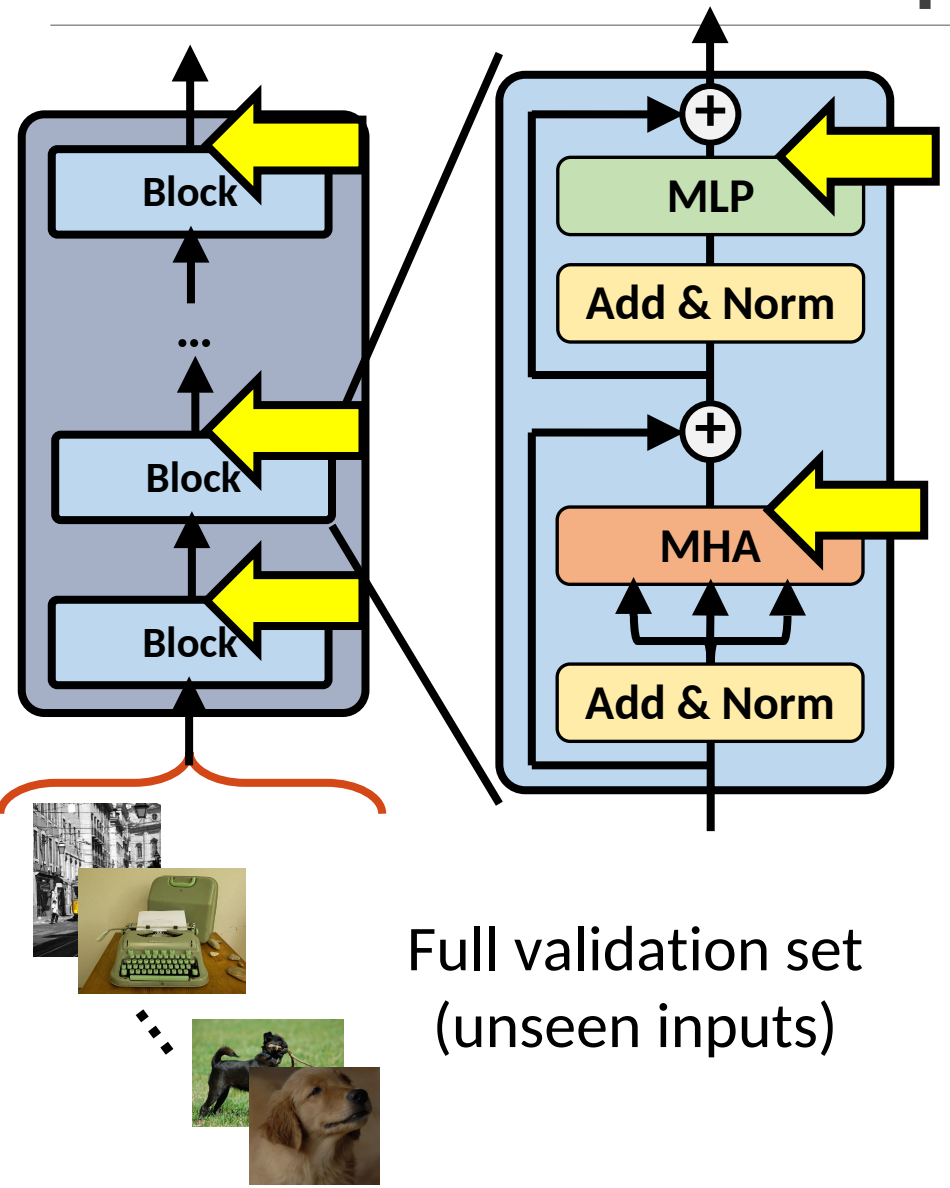


Full validation set
(unseen inputs)

Activation ranges

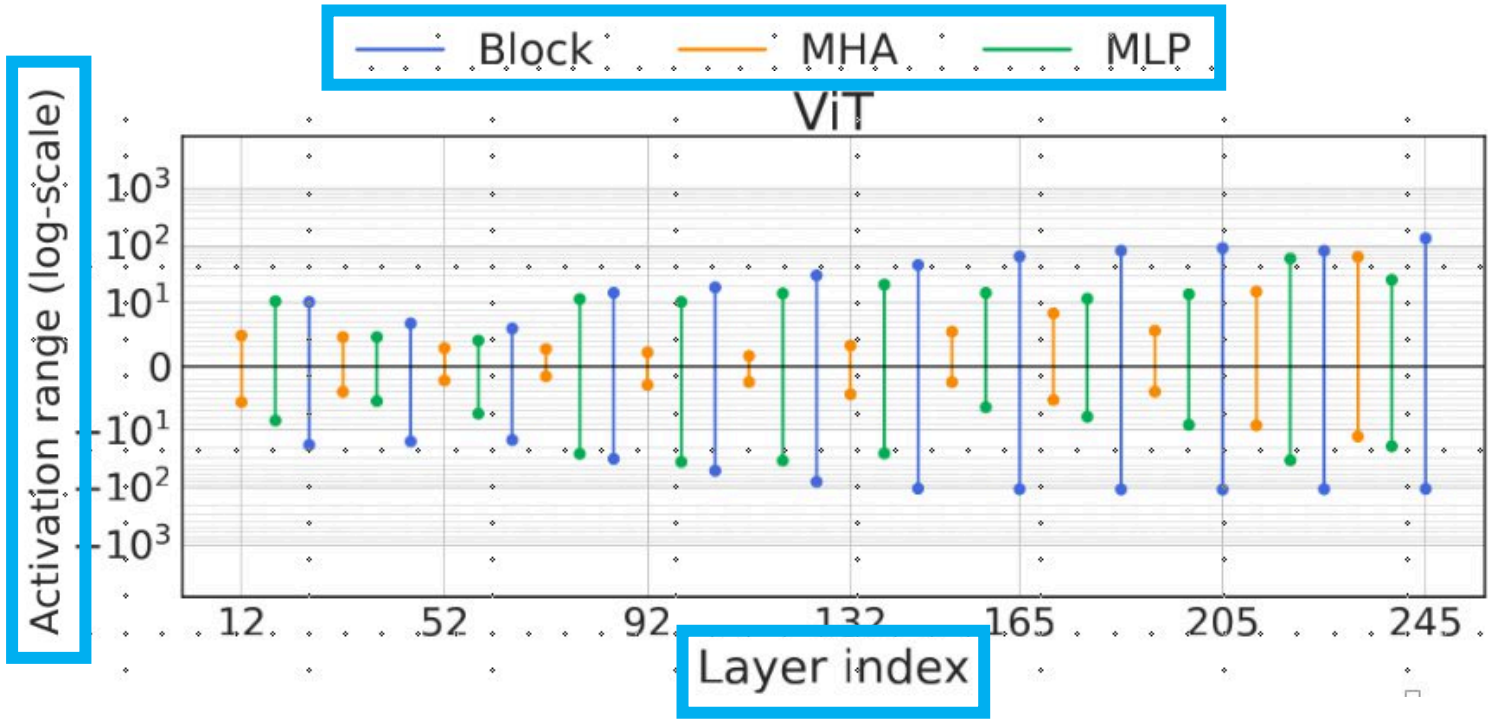


Activation value profiling

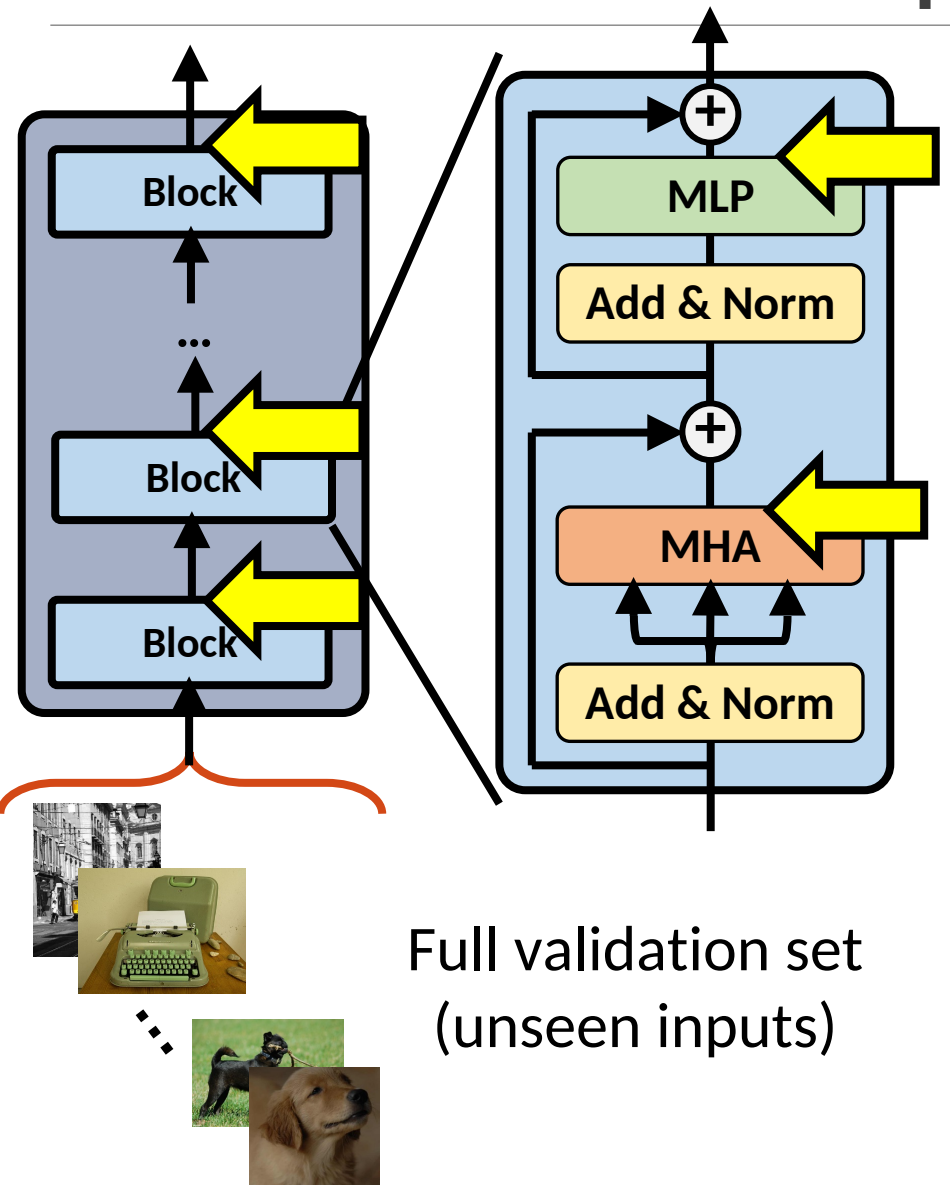


Full validation set
(unseen inputs)

Activation ranges

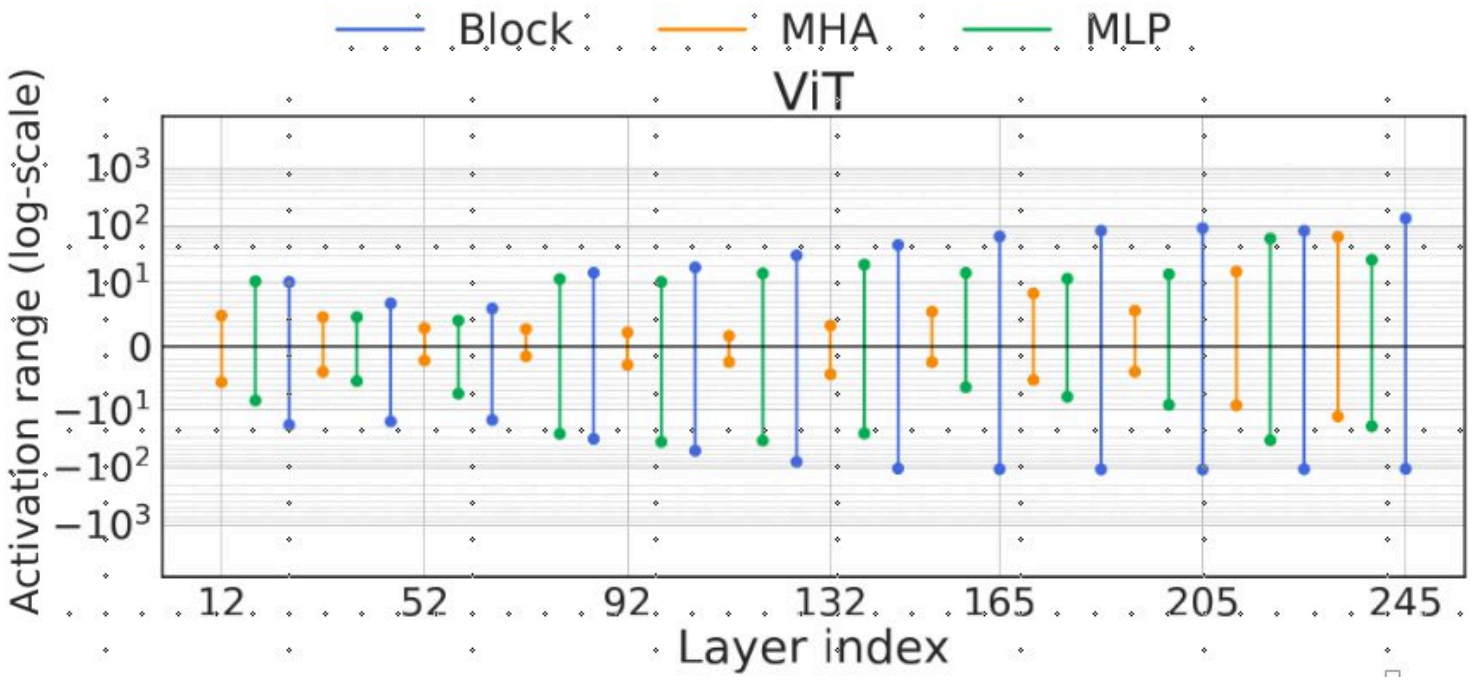


Activation value profiling

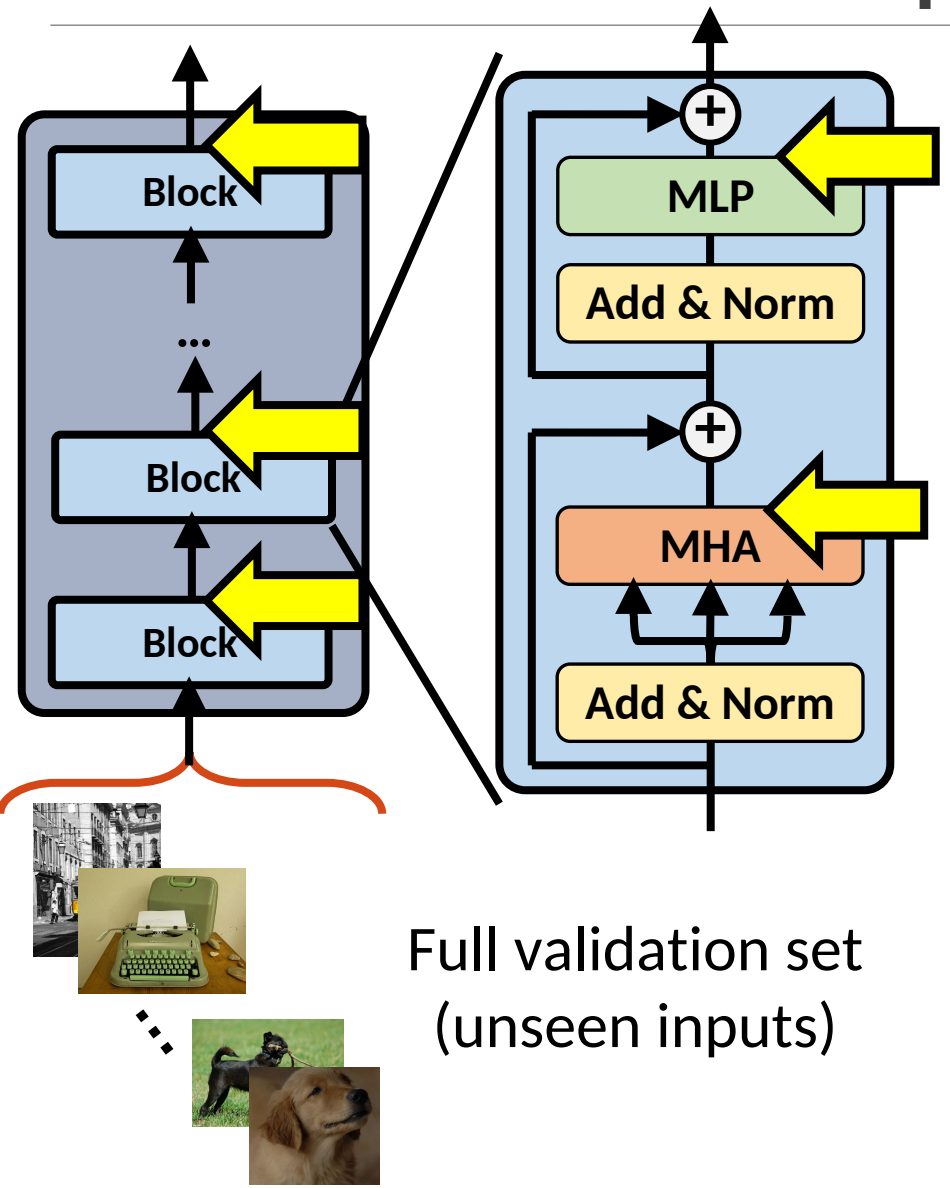


Full validation set
(unseen inputs)

Activation ranges

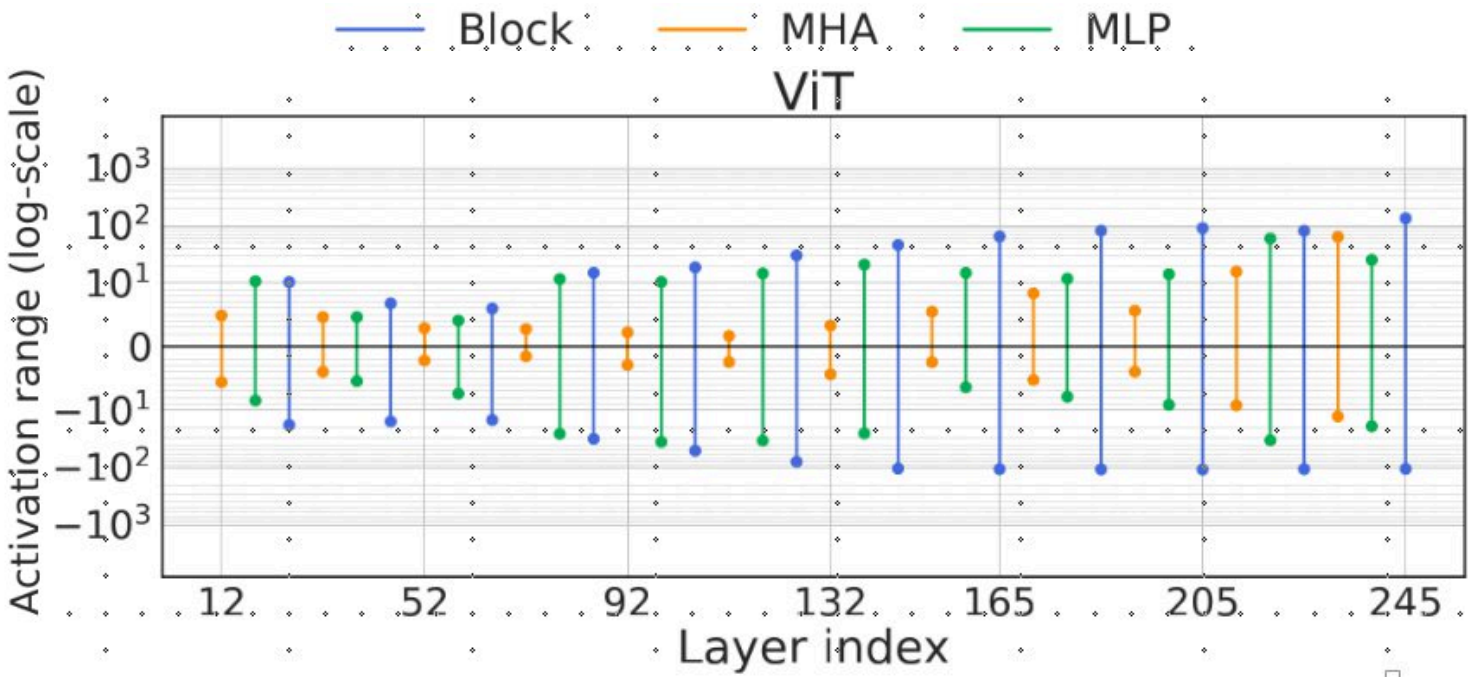


Activation value profiling



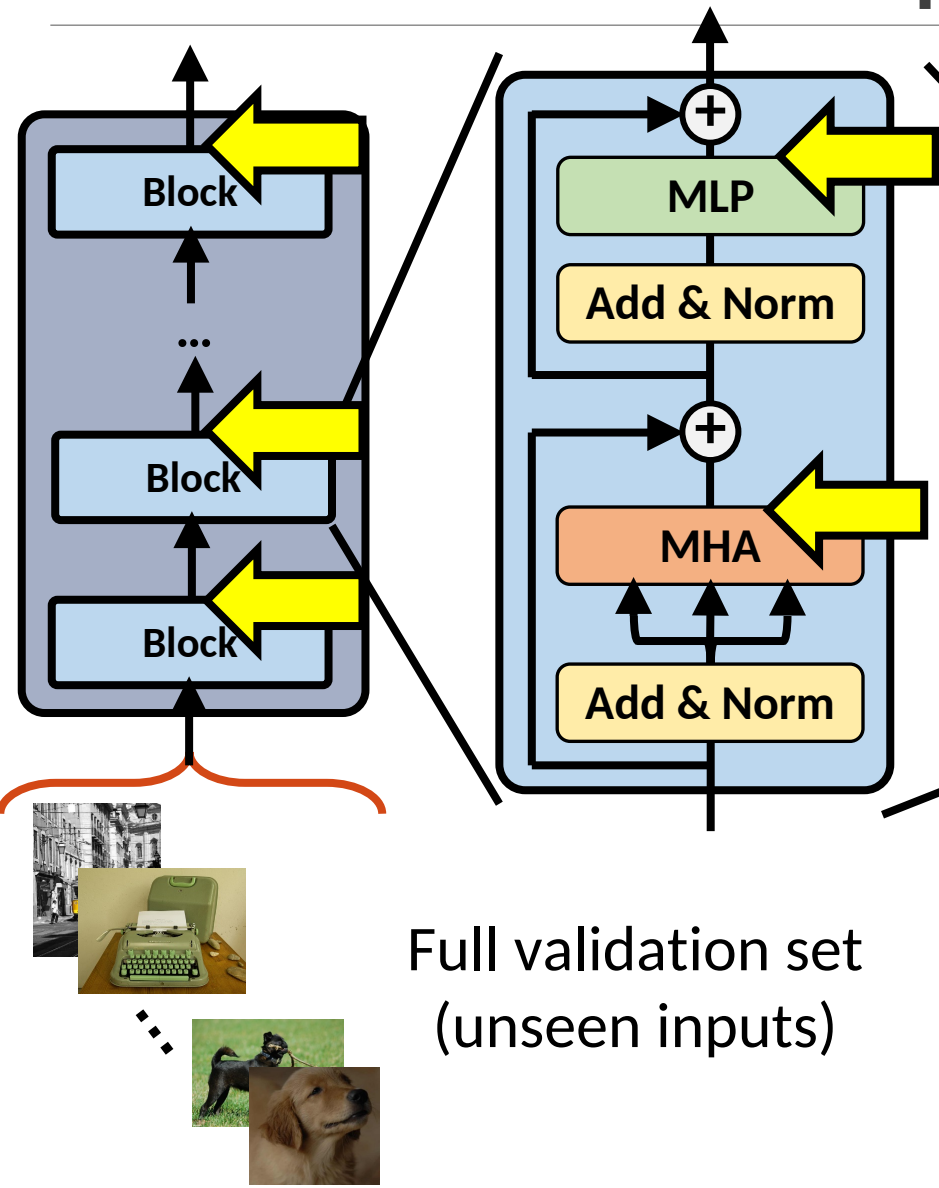
Full validation set
(unseen inputs)

Activation ranges



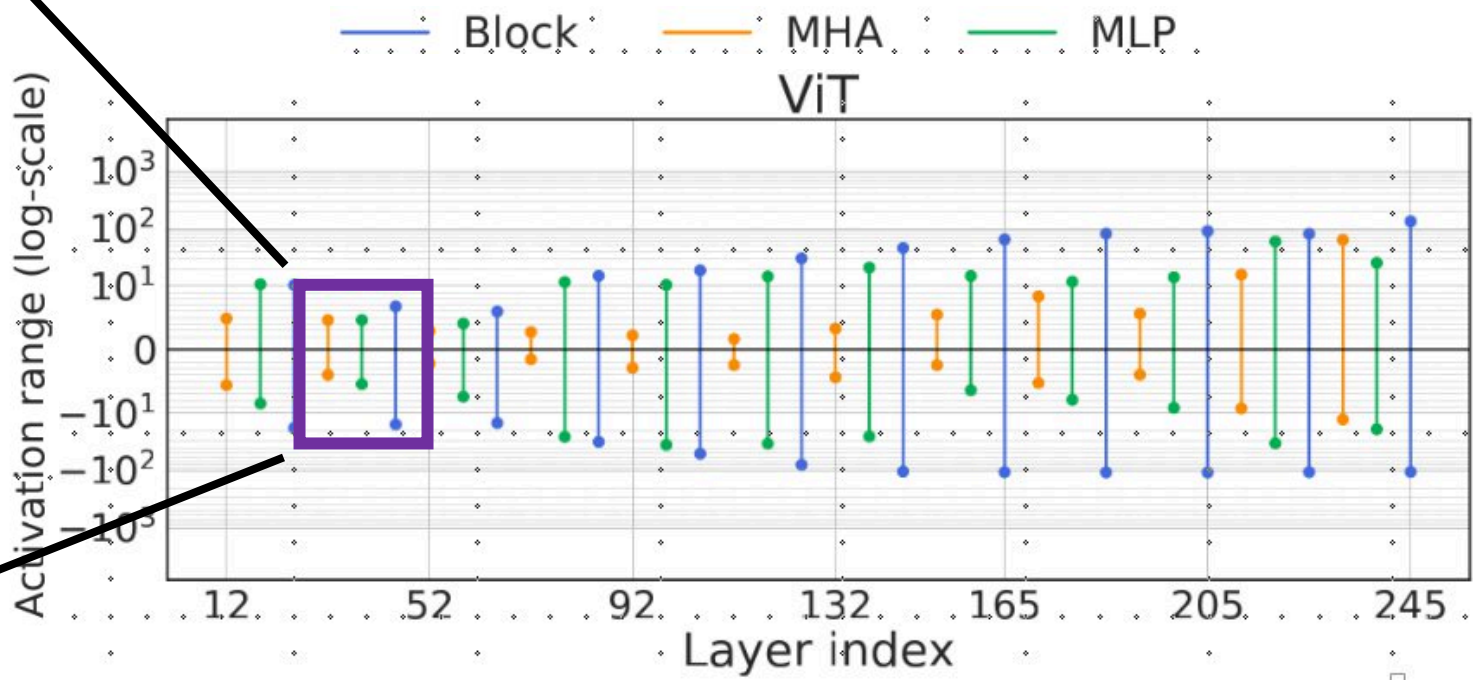
ViT: 12 blocks, each block comprises 1 MHA layer and 1 MLP

Activation value profiling



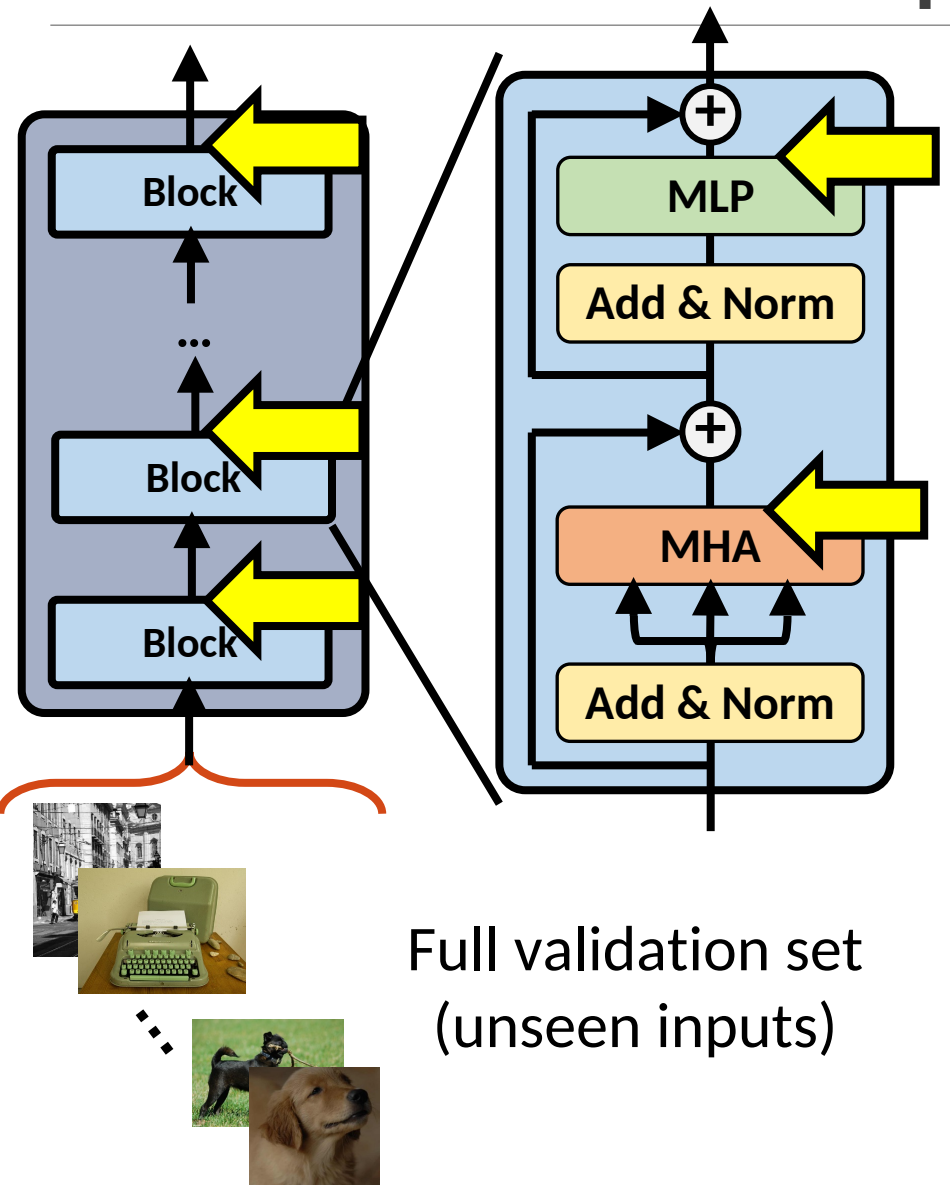
Full validation set
(unseen inputs)

Activation ranges



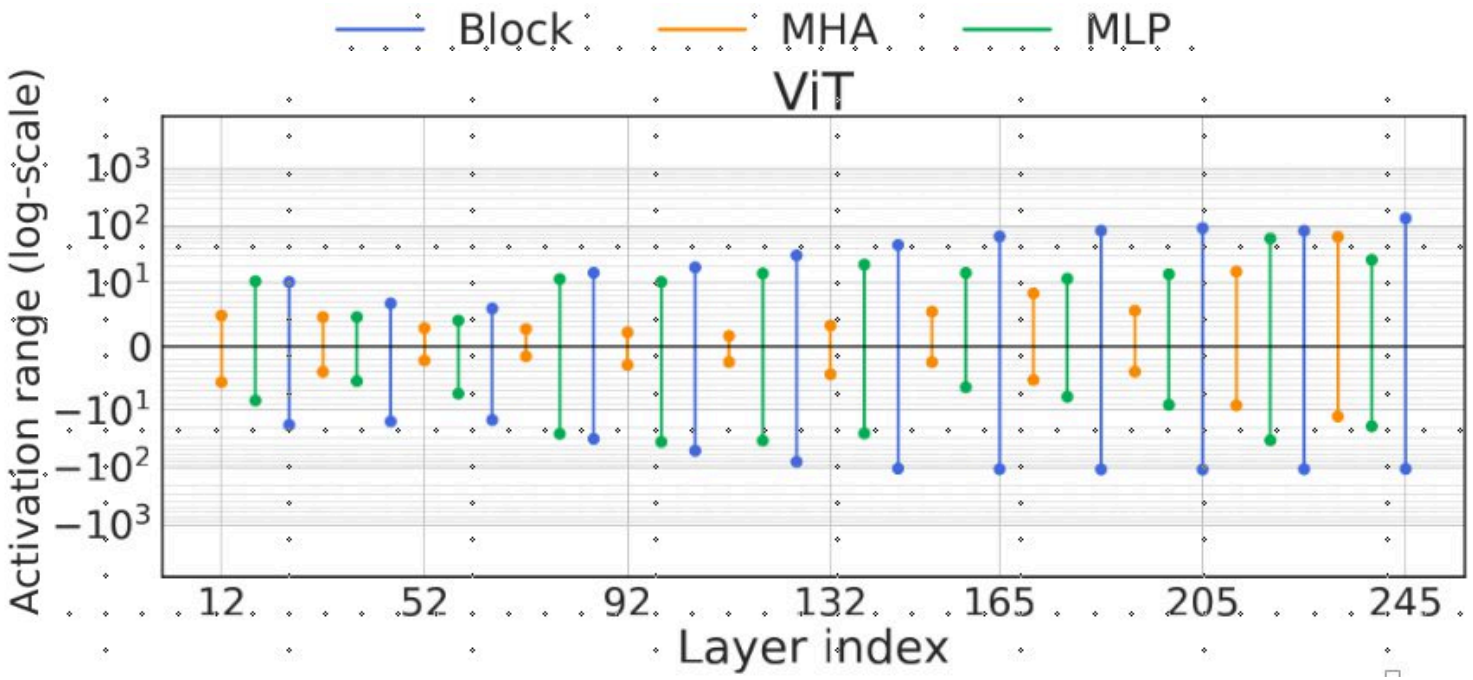
ViT: 12 blocks, each block comprises 1 MHA layer and 1 MLP

Activation value profiling



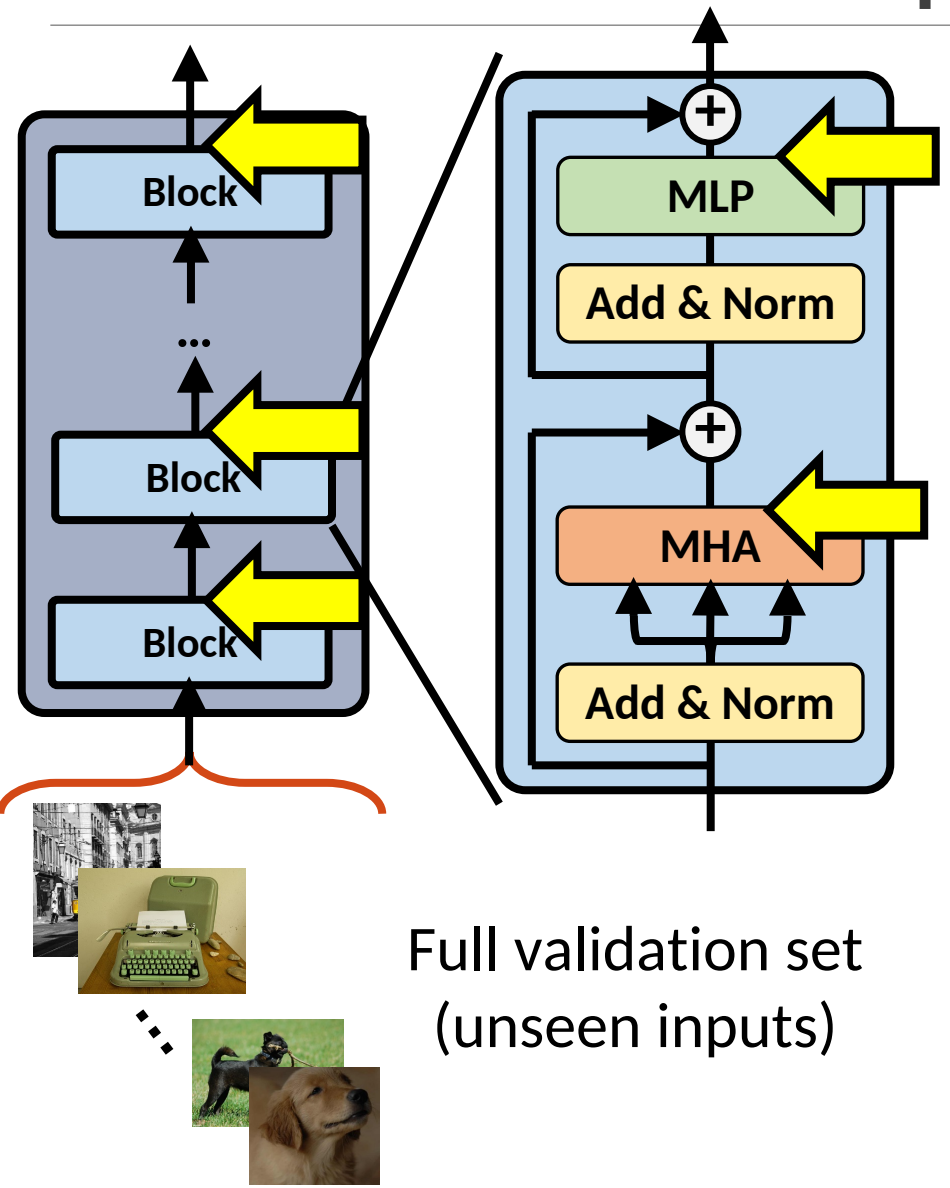
Full validation set
(unseen inputs)

Activation ranges



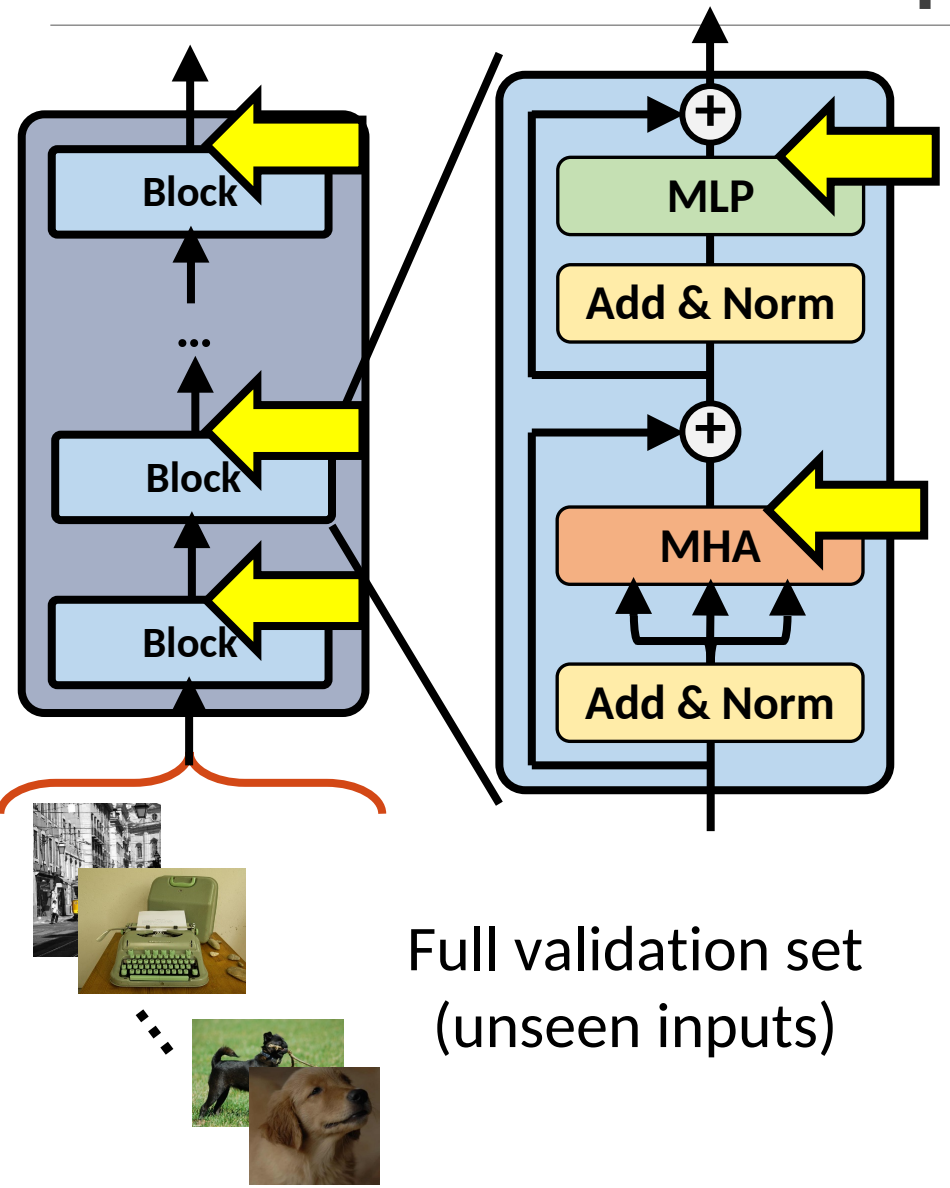
ViT: 12 blocks, each block comprises 1 MHA layer and 1 MLP

Activation value profiling



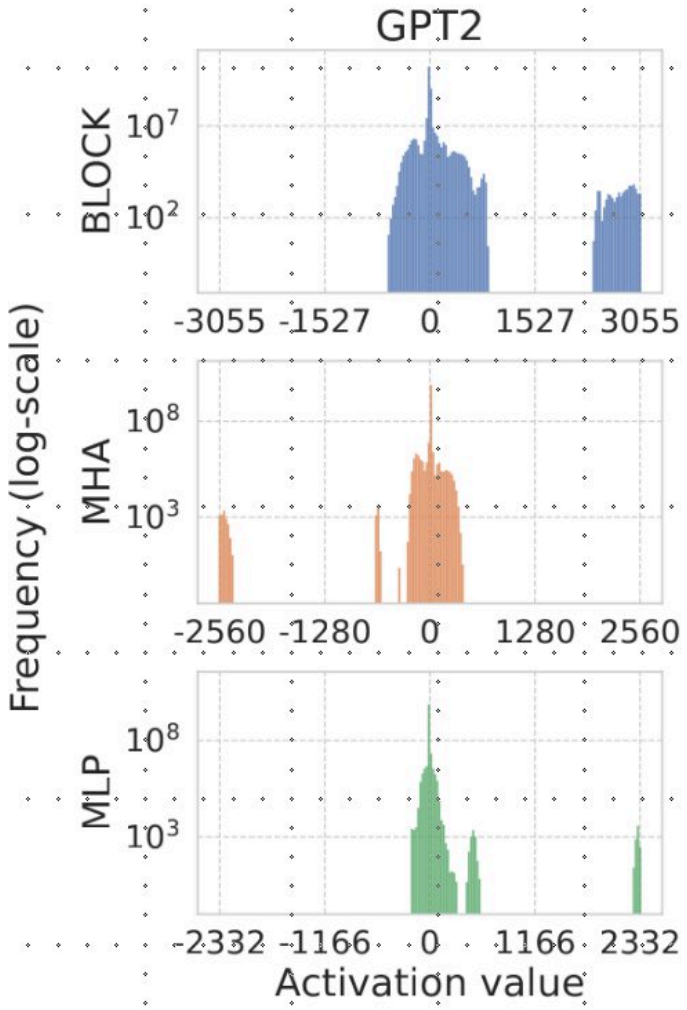
Full validation set
(unseen inputs)

Activation value profiling

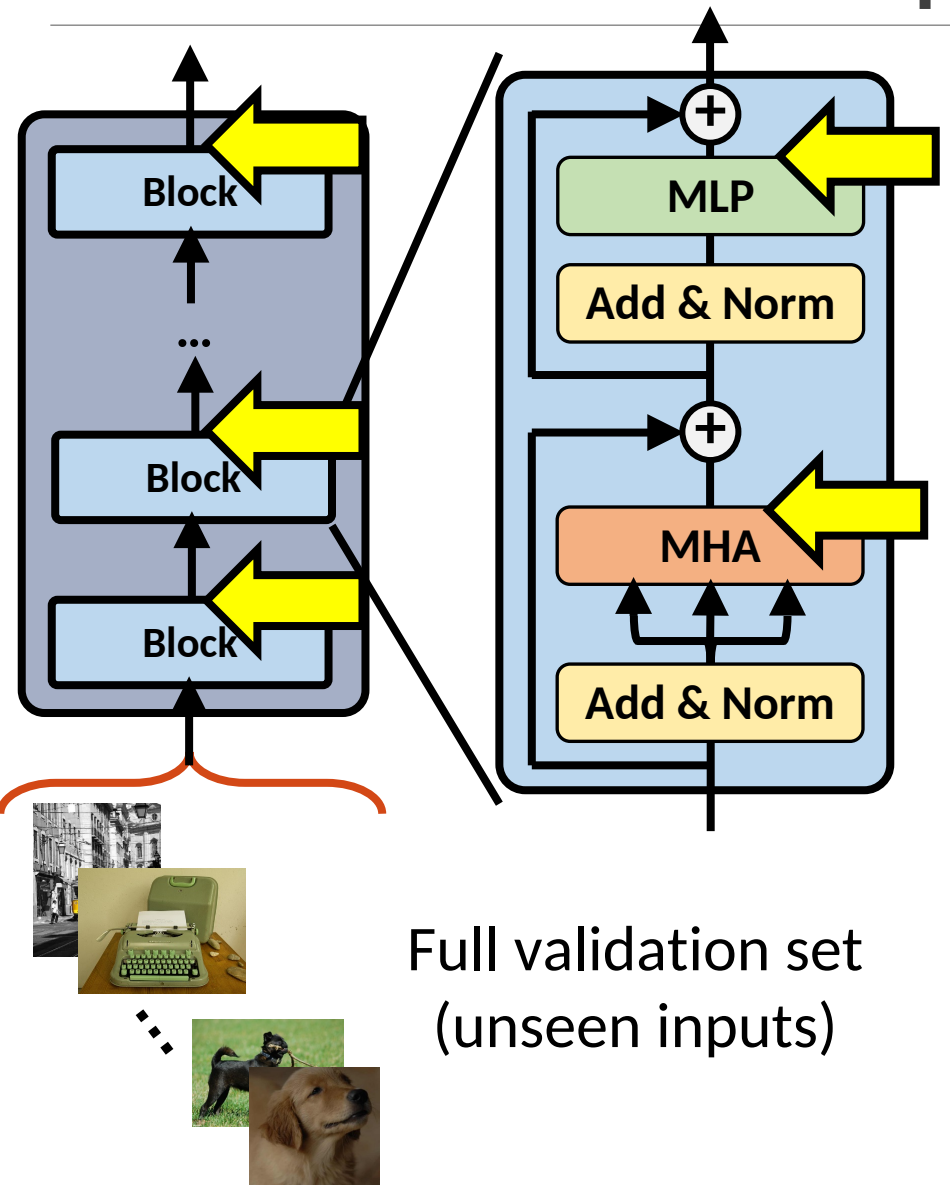


Full validation set
(unseen inputs)

Activation distribution

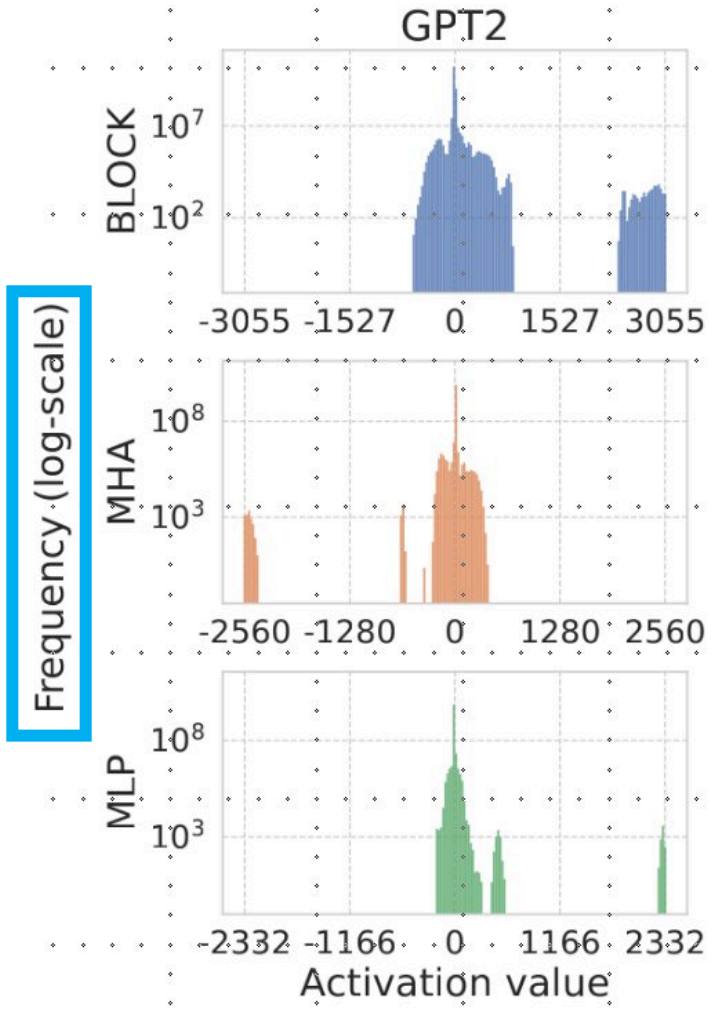


Activation value profiling

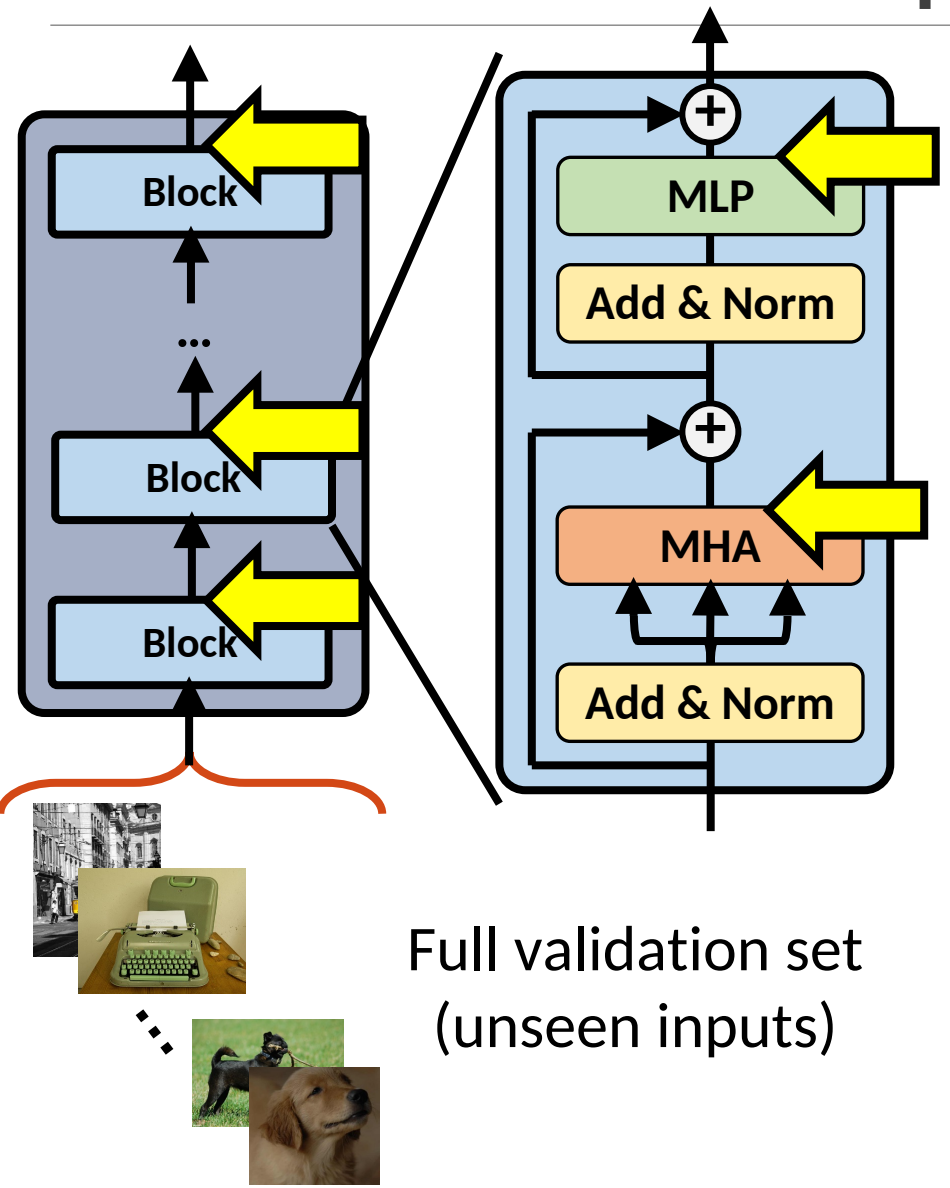


Full validation set
(unseen inputs)

Activation distribution

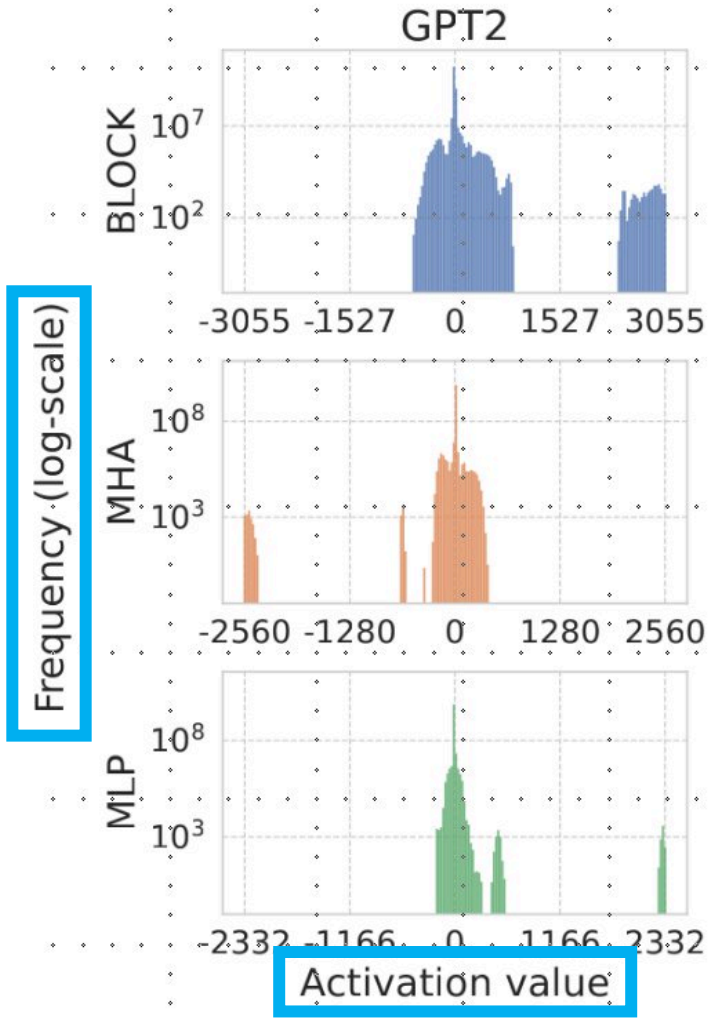


Activation value profiling

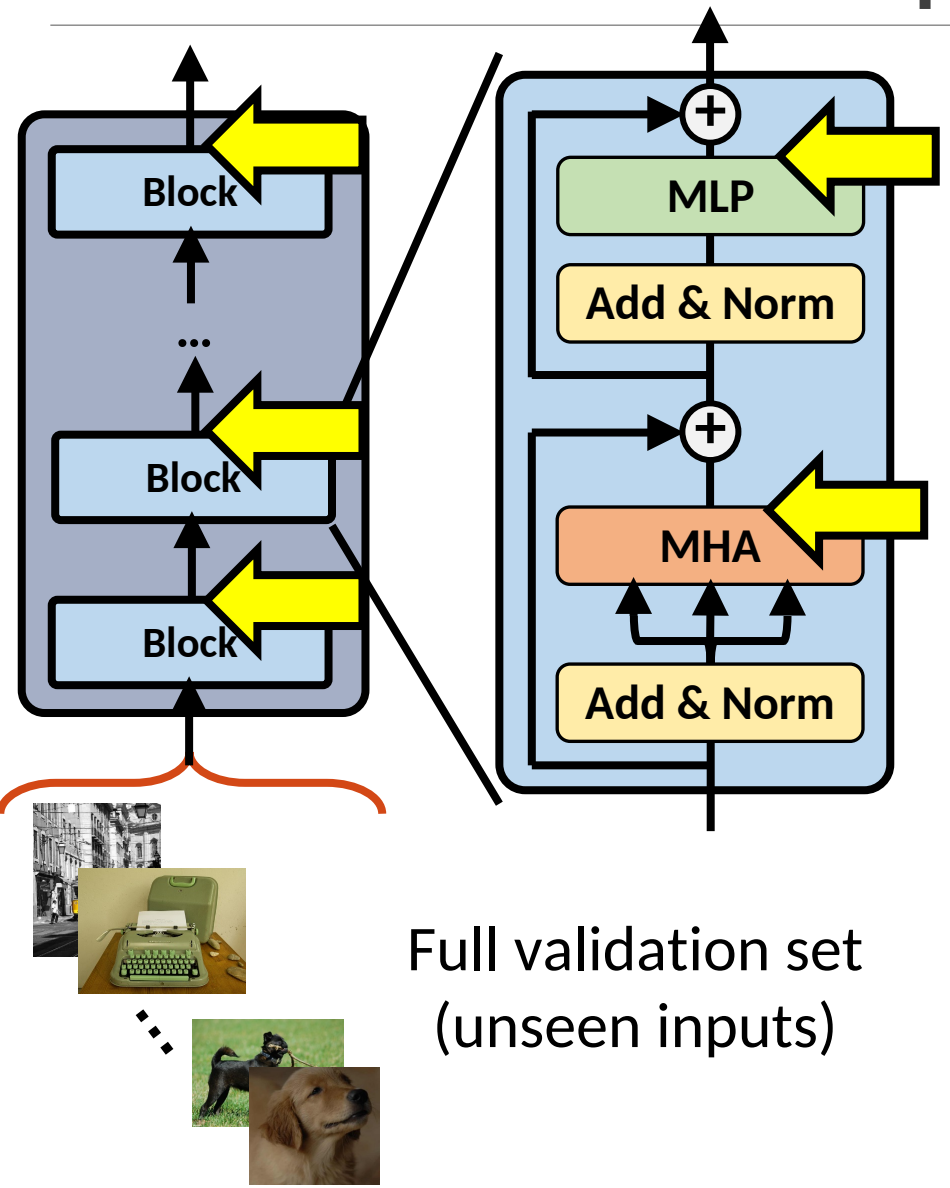


Full validation set
(unseen inputs)

Activation distribution

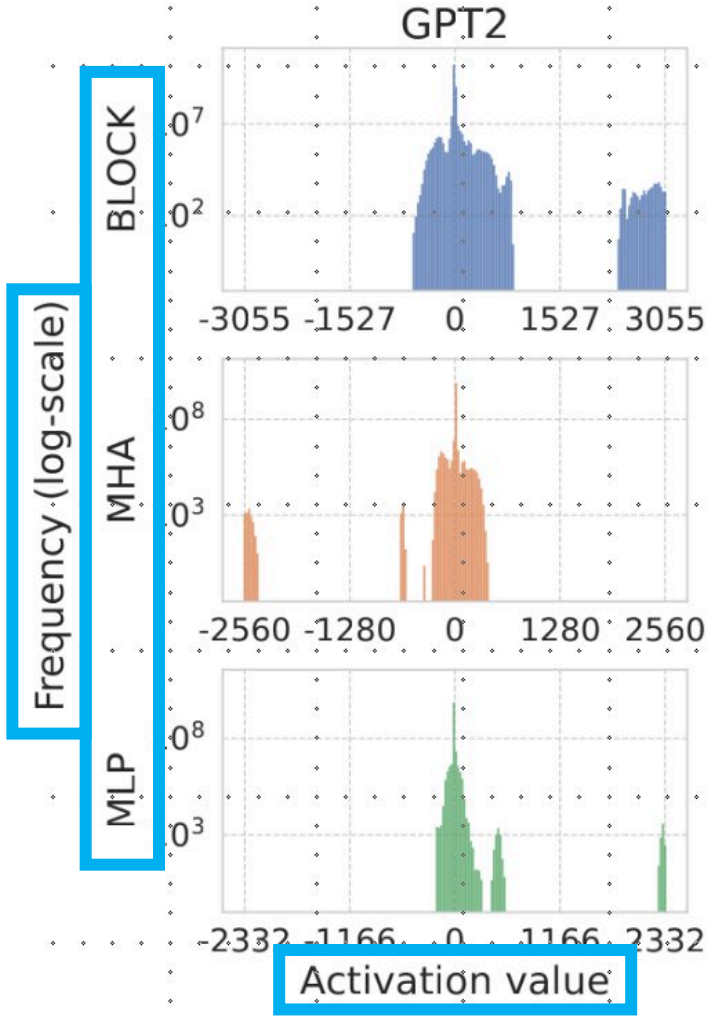


Activation value profiling

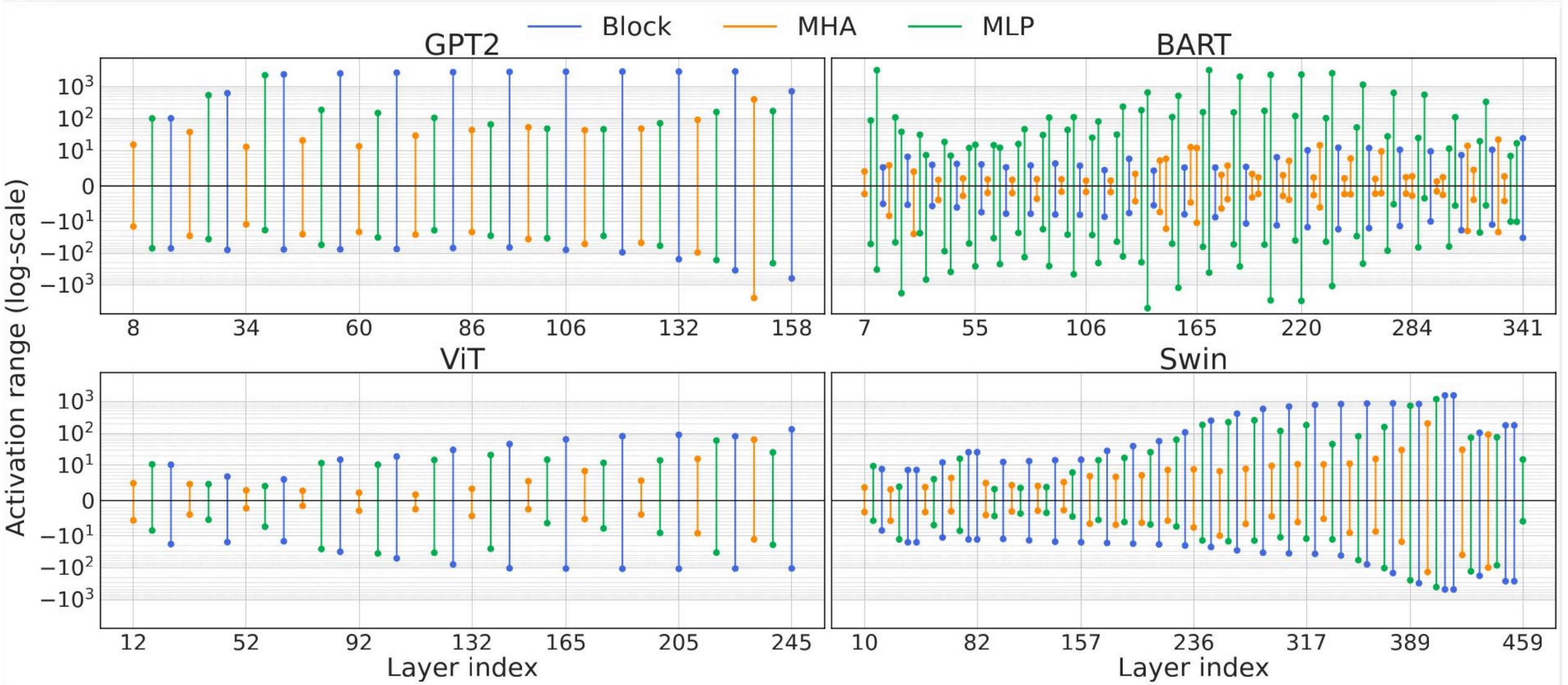


Full validation set
(unseen inputs)

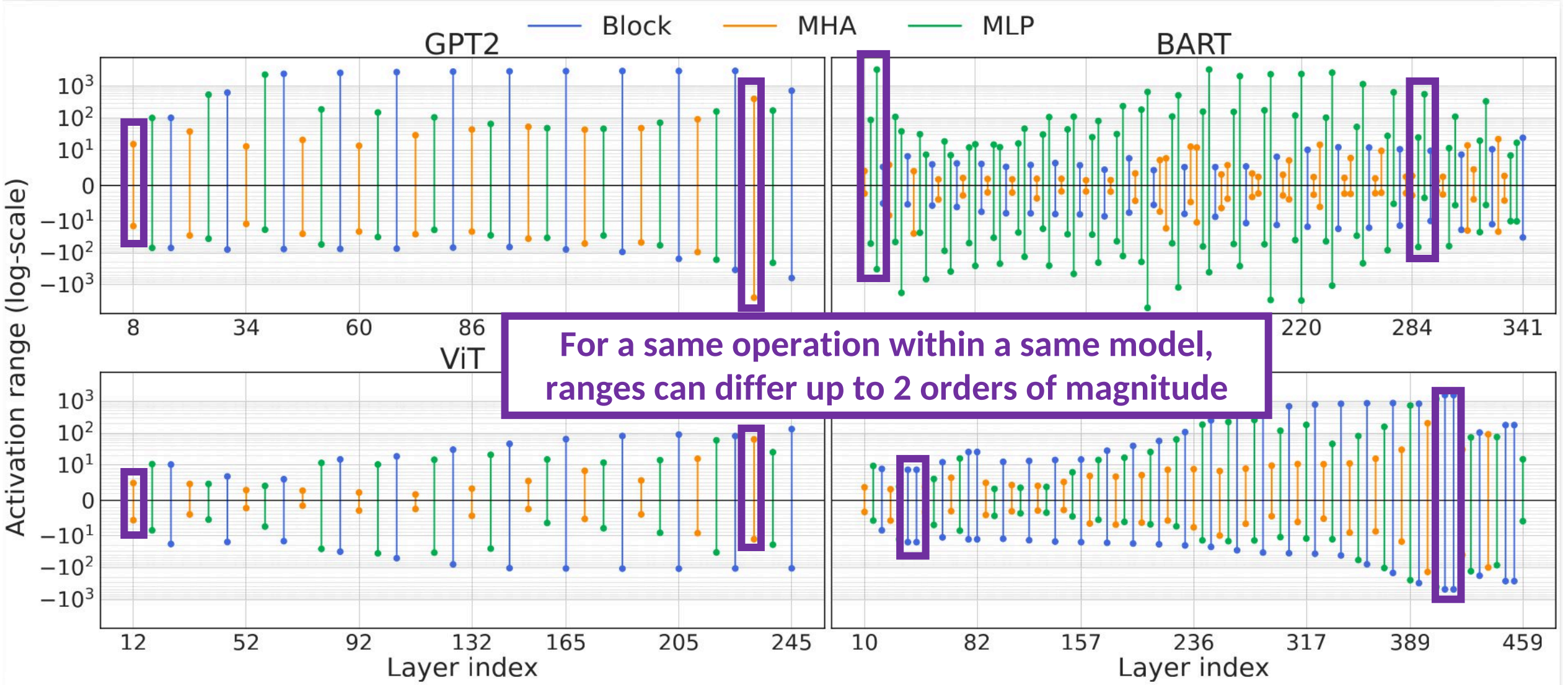
Activation distribution



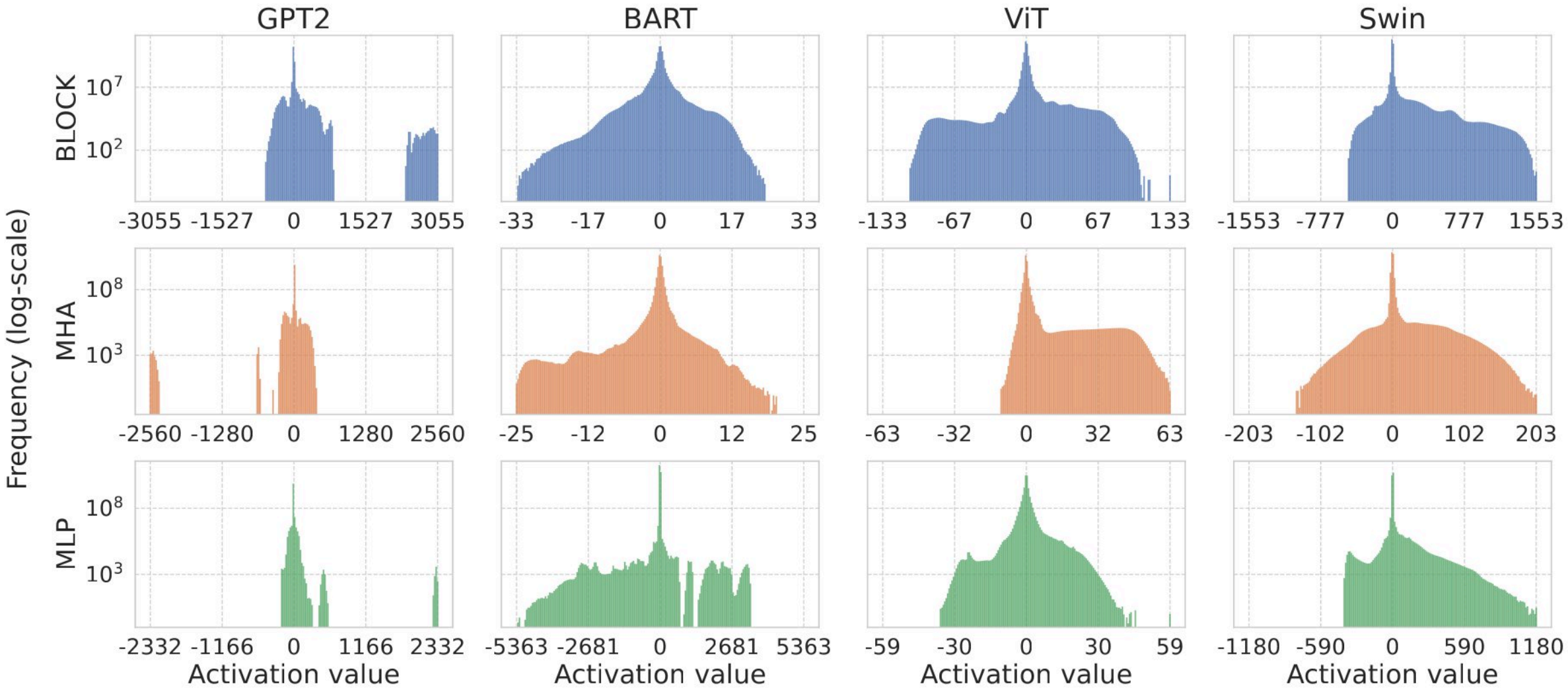
Activation value analysis: models' ranges



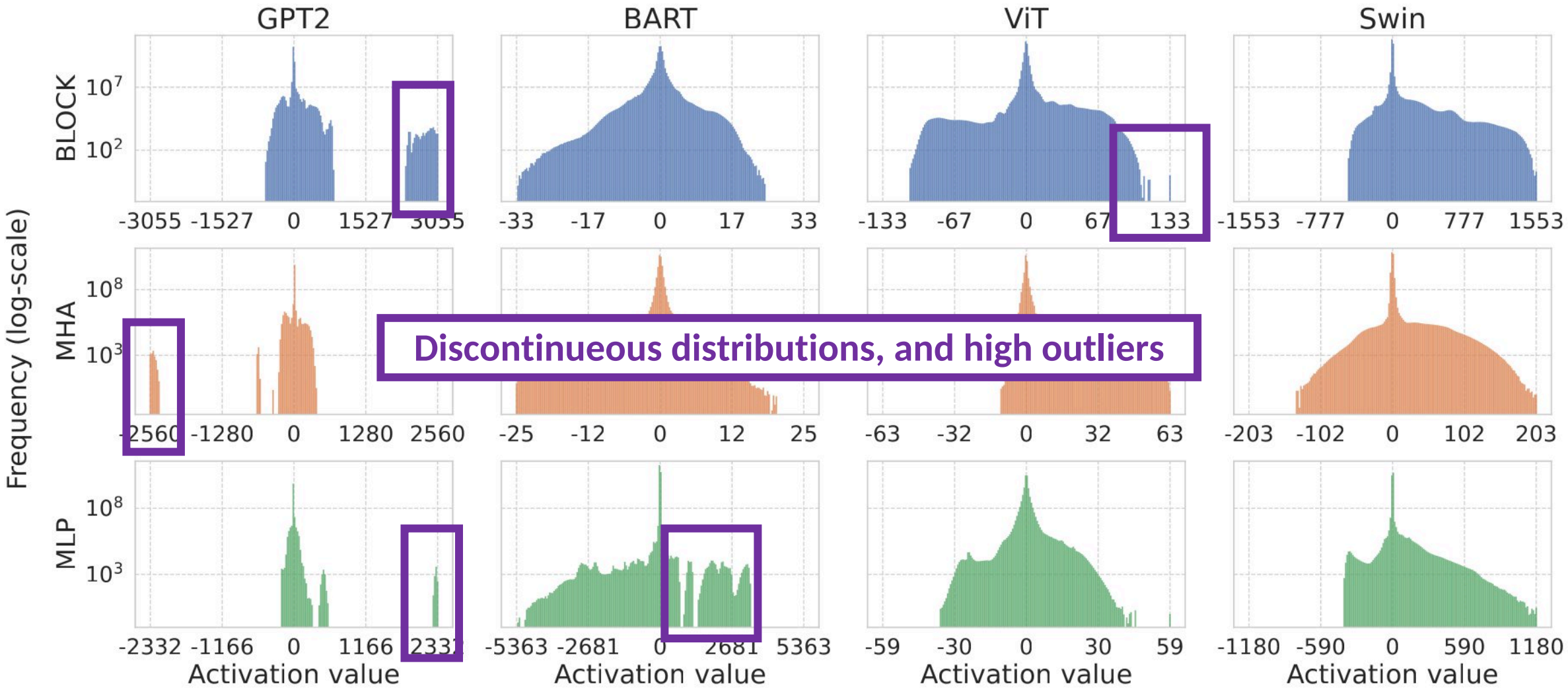
Activation value analysis: models' ranges



Activation value analysis: models' distributions



Activation value analysis: models' distributions



Takeaways and clipping design

Activation values have a high variability

Takeaways and clipping design

Activation values have a high variability

- Ranges are high: GPT2, BART MLP, Swin MLP and Block

Takeaways and clipping design

Activation values have a high variability

- Ranges are high: GPT2, BART MLP, Swin MLP and Block
- Different orders of magnitude for a same model

Takeaways and clipping design

Activation values have a high variability

- Ranges are high: GPT2, BART MLP, Swin MLP and Block
- Different orders of magnitude for a same model
- Distributions are shattered for GPT2, BART, and ViT

Takeaways and clipping design

Activation values have a high variability

- Ranges are high: GPT2, BART MLP, Swin MLP and Block
- Different orders of magnitude for a same model
- Distributions are shattered for GPT2, BART, and ViT
- Activation values are focused near zero

Takeaways and clipping design

Activation values have a high variability

- Ranges are high: GPT2, BART MLP, Swin MLP and Block
- Different orders of magnitude for a same model
- Distributions are shattered for GPT2, BART, and ViT
- Activation values are focused near zero

We hypothesize that this high variability will impact the efficiency of activation clipping

Takeaways and clipping design

Activation values have a high variability

- Ranges are high: GPT2, BART MLP, Swin MLP and Block
- Different orders of magnitude for a same model
- Distributions are shattered for GPT2, BART, and ViT
- Activation values are focused near zero

We hypothesize that this high variability will impact the efficiency of activation clipping

We perform fault injections on four Transformers that embedd layer-wise clipping. According to the activation values distribution and [1], all faulty activations are set to 0.

[1] S. Burel, *Microelectronics Reliability*, 2024

Activation Clipping Efficiency Assessment

Fault injections campaigns

Setup

Fault injections campaigns

Setup

- We inject on different operations:
  Block, MLP, and MHA (first, mid., last)

Fault injections campaigns

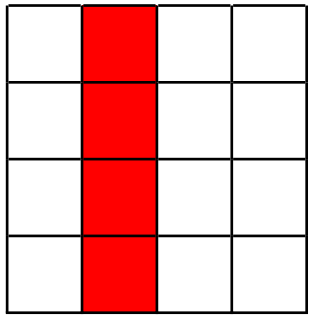
Setup

- We inject on different operations:
  Block, MLP, and MHA (first, mid., last)
- Different fault models

Fault injections campaigns

Setup

- We inject on different operations:
  Block, MLP, and MHA (first, mid., last)
- Different fault models

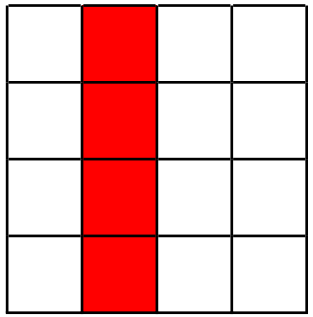


Column[1]

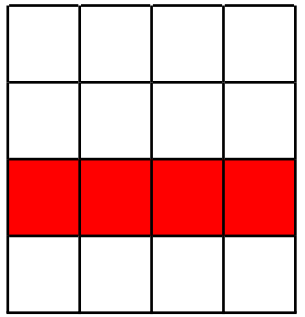
Fault injections campaigns

Setup

- We inject on different operations: Block, MLP, and MHA (first, mid., last)
- Different fault models



Column[1]

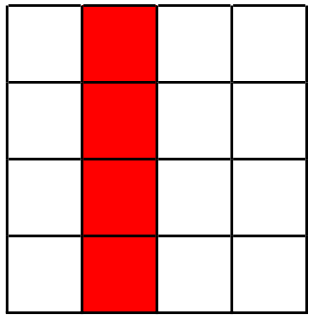


Row[1]

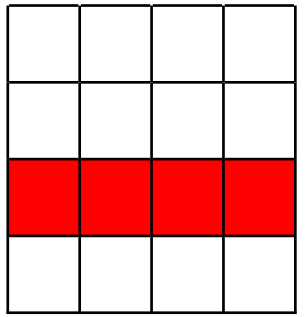
Fault injections campaigns

Setup

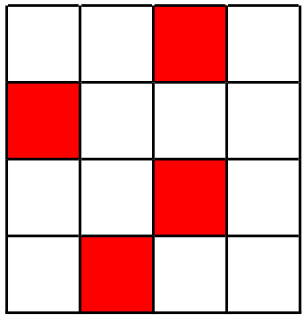
- We inject on different operations: Block, MLP, and MHA (first, mid., last)
- Different fault models



Column[1]



Row[1]

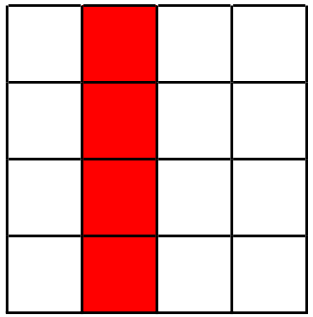


Multiple random[2]

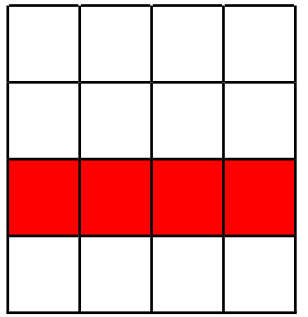
Fault injections campaigns

Setup

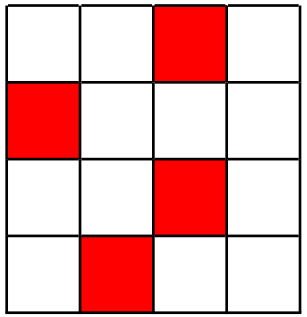
- We inject on different operations: Block, MLP, and MHA (first, mid., last)
- Different fault models



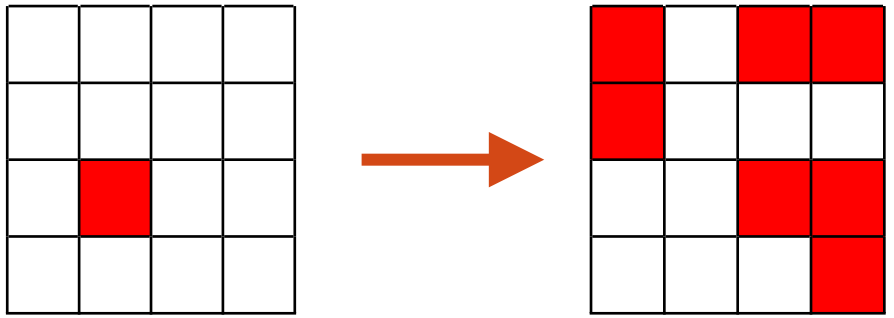
Column[1]



Row[1]



Multiple random[2]



Corrupted value propagation[3]

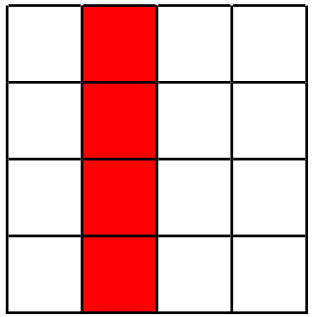
[1] C. Bolchini, *IEEE TC*, 2022
[2] F. F. d. Santos, *IEEE/IFIP*, 2021
[3] Y. Sun, *SC*, 2025

Fault injections campaigns

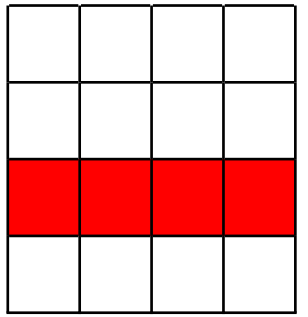
Setup

- We inject on different operations: Block, MLP, and MHA (first, mid., last)
- Different fault models

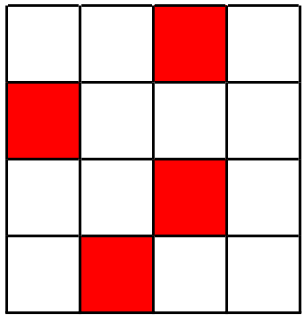
Model	#injections*	Critical SDC PVF%
GPT2	1.81M	0.44%
BART	2.66M	1.31%
ViT	9.57M	1.06%
Swin	9.59M	0.41%



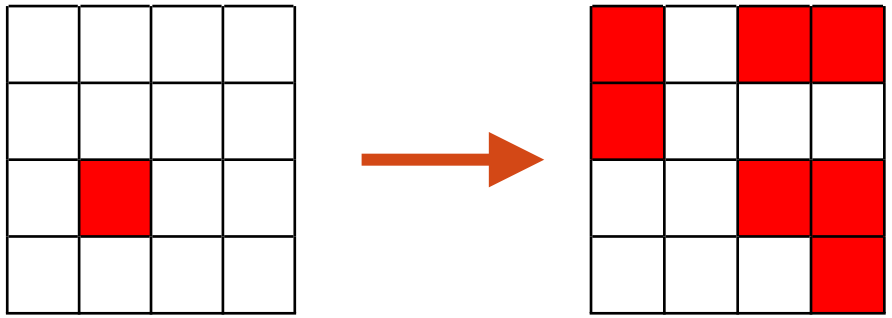
Column[1]



Row[1]



Multiple random[2]



Corrupted value propagation[3]

[1] C. Bolchini, *IEEE TC*, 2022
[2] F. F. d. Santos, *IEEE/IFIP*, 2021
[3] Y. Sun, *SC*, 2025

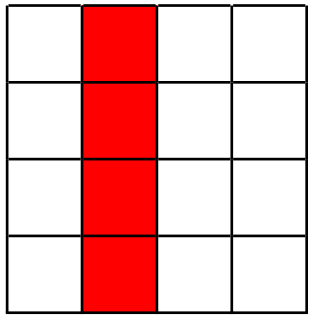
Fault injections campaigns

Setup

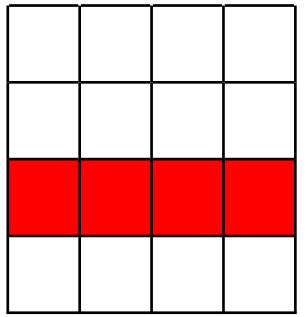
- We inject on different operations: Block, MLP, and MHA (first, mid., last)
- Different fault models

Model	#injections*	Critical SDC PVF%
GPT2	1.81M	0.44%
BART	2.66M	1.31%
ViT	9.57M	1.06%
Swin	9.59M	0.41%

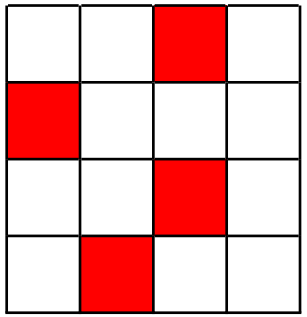
*#injections = #inputs * #faults * #layers



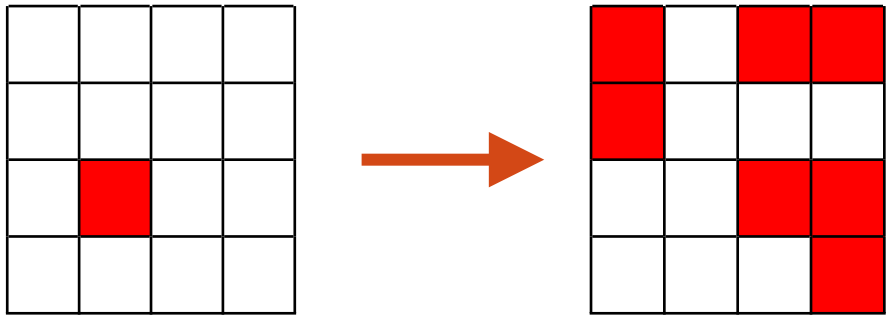
Column[1]



Row[1]



Multiple random[2]



Corrupted value propagation[3]

[1] C. Bolchini, *IEEE TC*, 2022
[2] F. F. d. Santos, *IEEE/IFIP*, 2021
[3] Y. Sun, *SC*, 2025

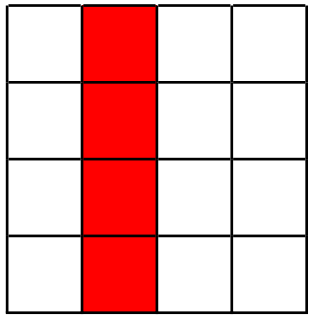
Fault injections campaigns

Setup

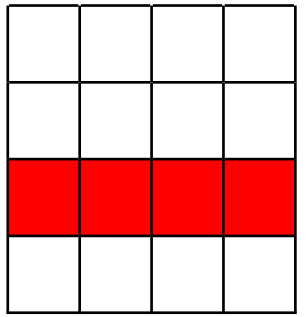
- We inject on different operations: Block, MLP, and MHA (first, mid., last)
- Different fault models

Model	#injections*	Critical SDC PVF%
GPT2	1.81M	0.44%
BART	2.66M	1.31%
ViT	Without any fault tolerance	4.50%
Swin		2.70%

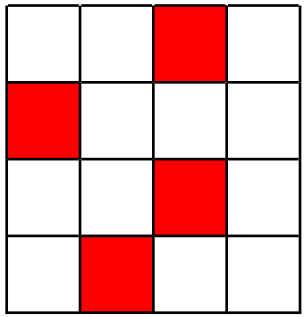
*#injections = #inputs * #faults * #layers



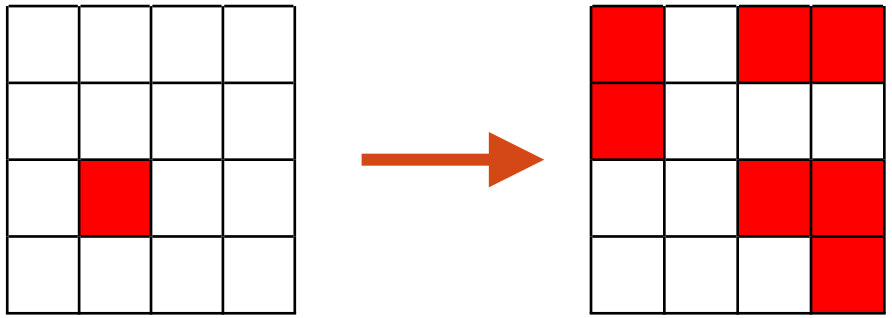
Column[1]



Row[1]



Multiple random[2]



Corrupted value propagation[3]

[1] C. Bolchini, *IEEE TC*, 2022
[2] F. F. d. Santos, *IEEE/IFIP*, 2021
[3] Y. Sun, *SC*, 2025

A step further: input sensitivity

Classification models:

A step further: input sensitivity

Classification models:

We define the confidence as:

$$C = Top1 - Top2$$

A step further: input sensitivity

Classification models:

We define the confidence as:

$$C = Top1 - Top2$$

C close to 0: model hesitates

Pred. Class: Ambulance | Conf.: 0.1%

A step further: input sensitivity

Classification models:

We define the confidence as:

$$C = Top1 - Top2$$

C close to 0: model hesitates

C close to 1: model is confident

Pred. Class: Ambulance | Conf.: 0.1%

Pred. Class: Dog | Conf.: 96%

A step further: input sensitivity

Classification models:

We define the confidence as:

$$C = Top1 - Top2$$

C close to 0: model hesitates

C close to 1: model is confident

We split inputs by confidence intervals

Very Low: $0.00 < C < 0.25$

Low: $0.25 < C < 0.50$

High: $0.50 < C < 0.75$

Very High: $0.75 < C \leq 1.0$

Pred. Class: Ambulance | Conf.: 0.1%

Pred. Class: Dog | Conf.: 96%

A step further: input sensitivity

Classification models:

We define the confidence as:

$$C = Top1 - Top2$$

C close to 0: model hesitates

C close to 1: model is confident

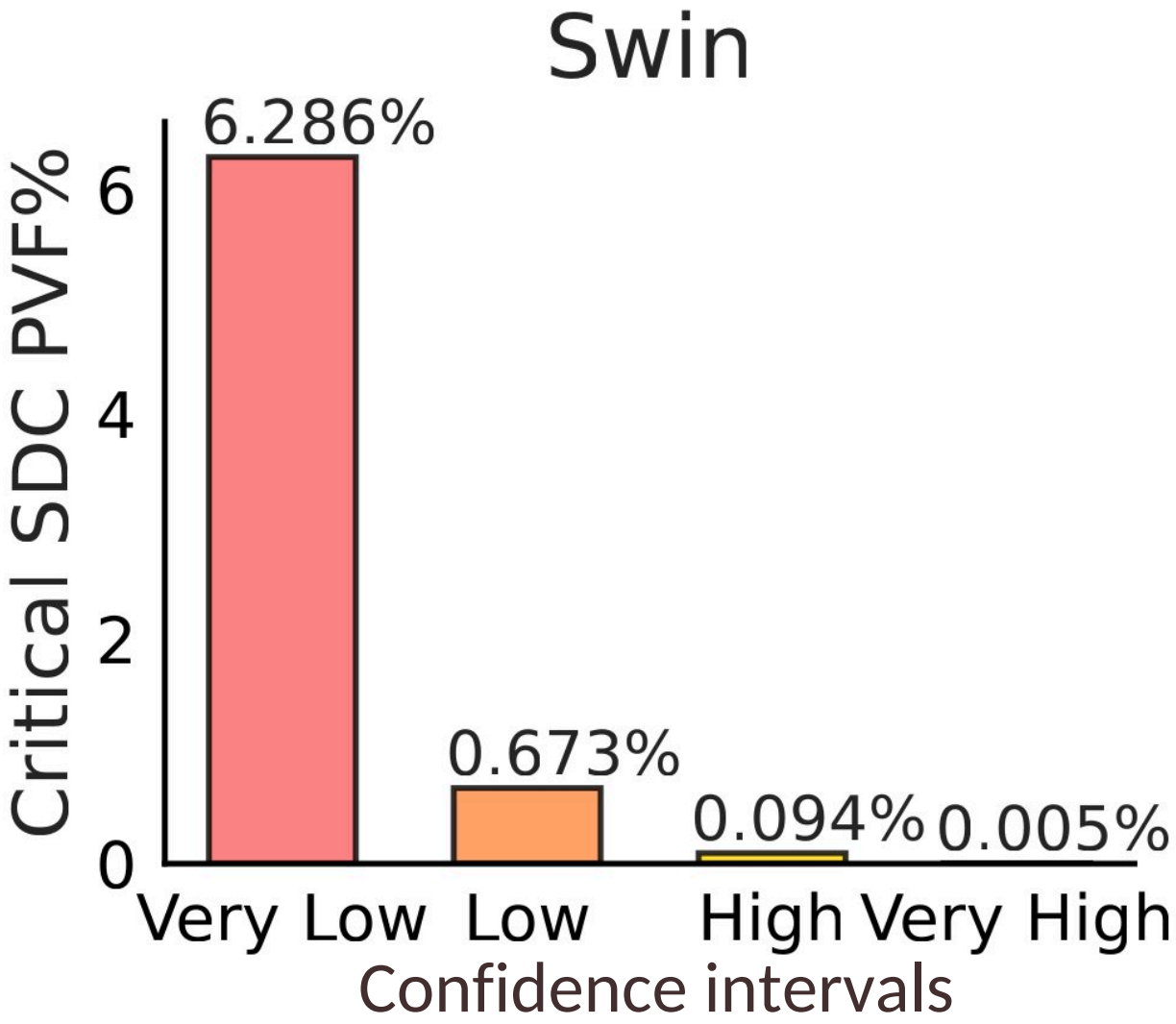
We split inputs by confidence intervals

Very Low: $0.00 < C < 0.25$

Low: $0.25 < C < 0.50$

High: $0.50 < C < 0.75$

Very High: $0.75 < C \leq 1.0$



A step further: input sensitivity

Classification models:

We define the confidence as:

$$C = Top1 - Top2$$

C close to 0: model hesitates

C close to 1: model is confident

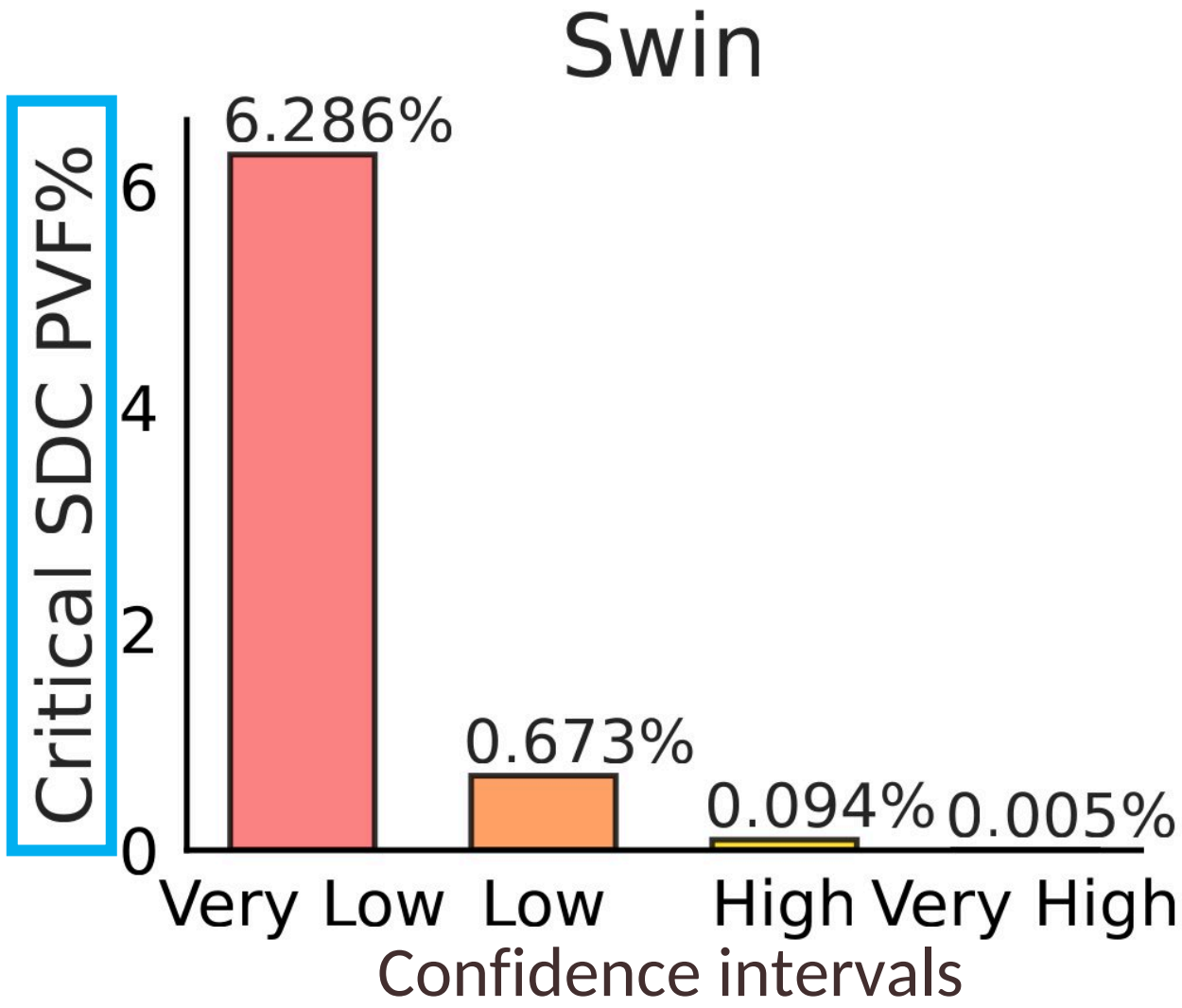
We split inputs by confidence intervals

Very Low: $0.00 < C < 0.25$

Low: $0.25 < C < 0.50$

High: $0.50 < C < 0.75$

Very High: $0.75 < C \leq 1.0$



A step further: input sensitivity

Classification models:

We define the confidence as:

$$C = Top1 - Top2$$

C close to 0: model hesitates

C close to 1: model is confident

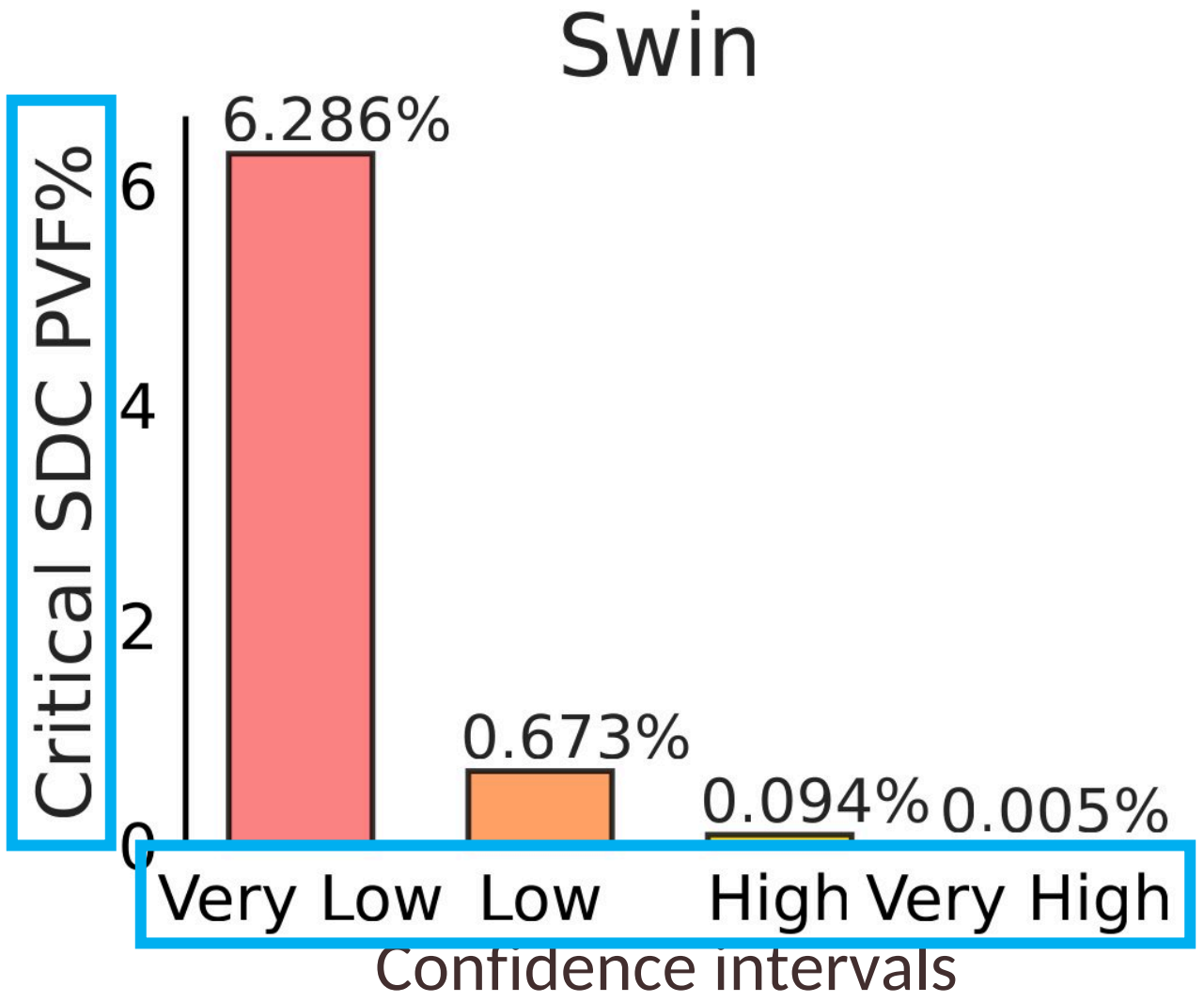
We split inputs by confidence intervals

Very Low: $0.00 < C < 0.25$

Low: $0.25 < C < 0.50$

High: $0.50 < C < 0.75$

Very High: $0.75 < C \leq 1.0$



A step further: input sensitivity

Classification models:

We define the confidence as:

$$C = Top1 - Top2$$

C close to 0: model hesitates

C close to 1: model is confident

We split inputs by confidence intervals

Very Low: $0.00 < C < 0.25$

Low: $0.25 < C < 0.50$

High: $0.50 < C < 0.75$

Very High: $0.75 < C \leq 1.0$

A step further: input sensitivity

Classification models:

We define the confidence as:

$$C = Top1 - Top2$$

C close to 0: model hesitates

C close to 1: model is confident

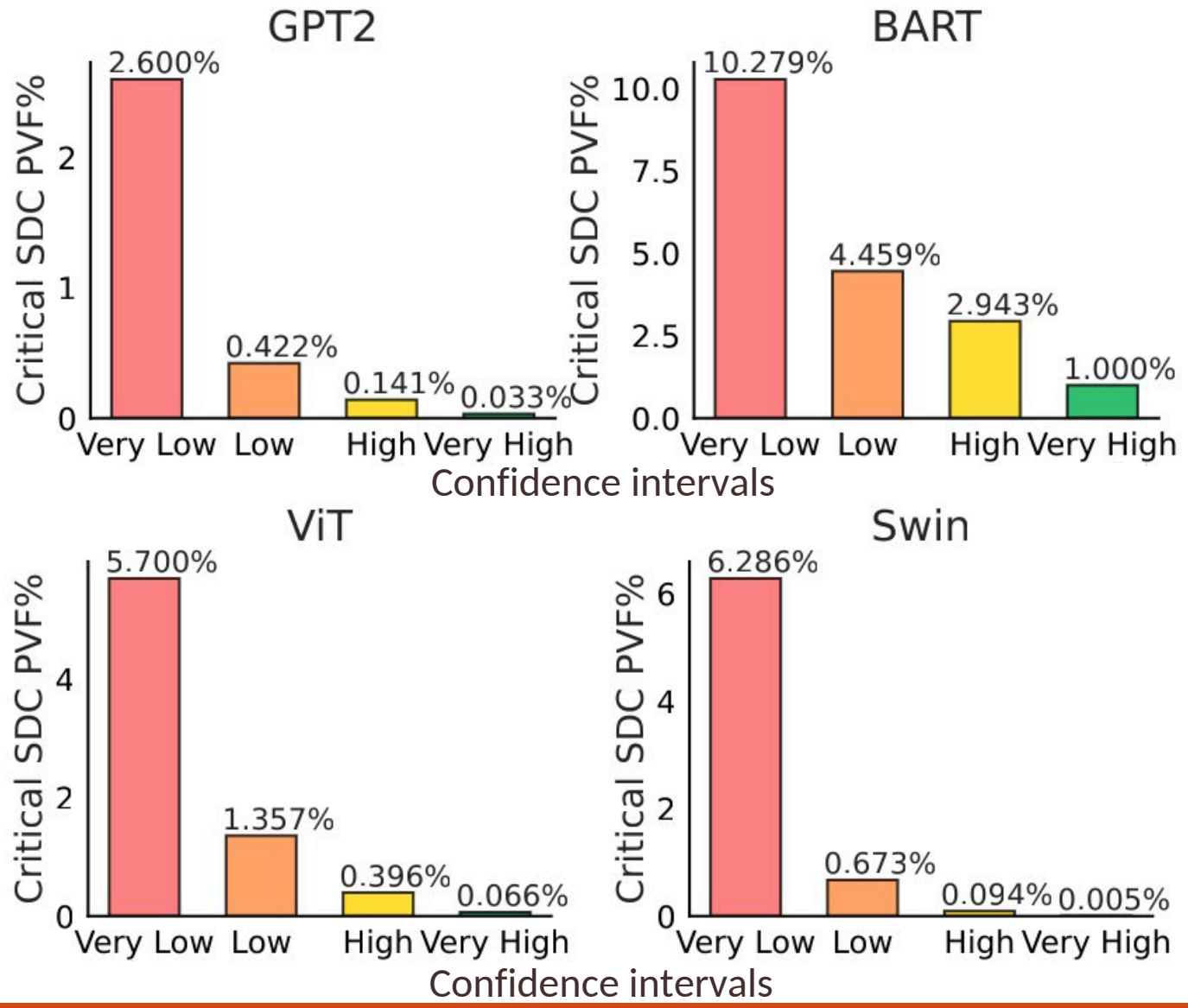
We split inputs by confidence intervals

Very Low: $0.00 < C < 0.25$

Low: $0.25 < C < 0.50$

High: $0.50 < C < 0.75$

Very High: $0.75 < C \leq 1.0$



A step further: input sensitivity

Classification models:

We define the confidence as:

Very Low and Low inputs are more vulnerable !

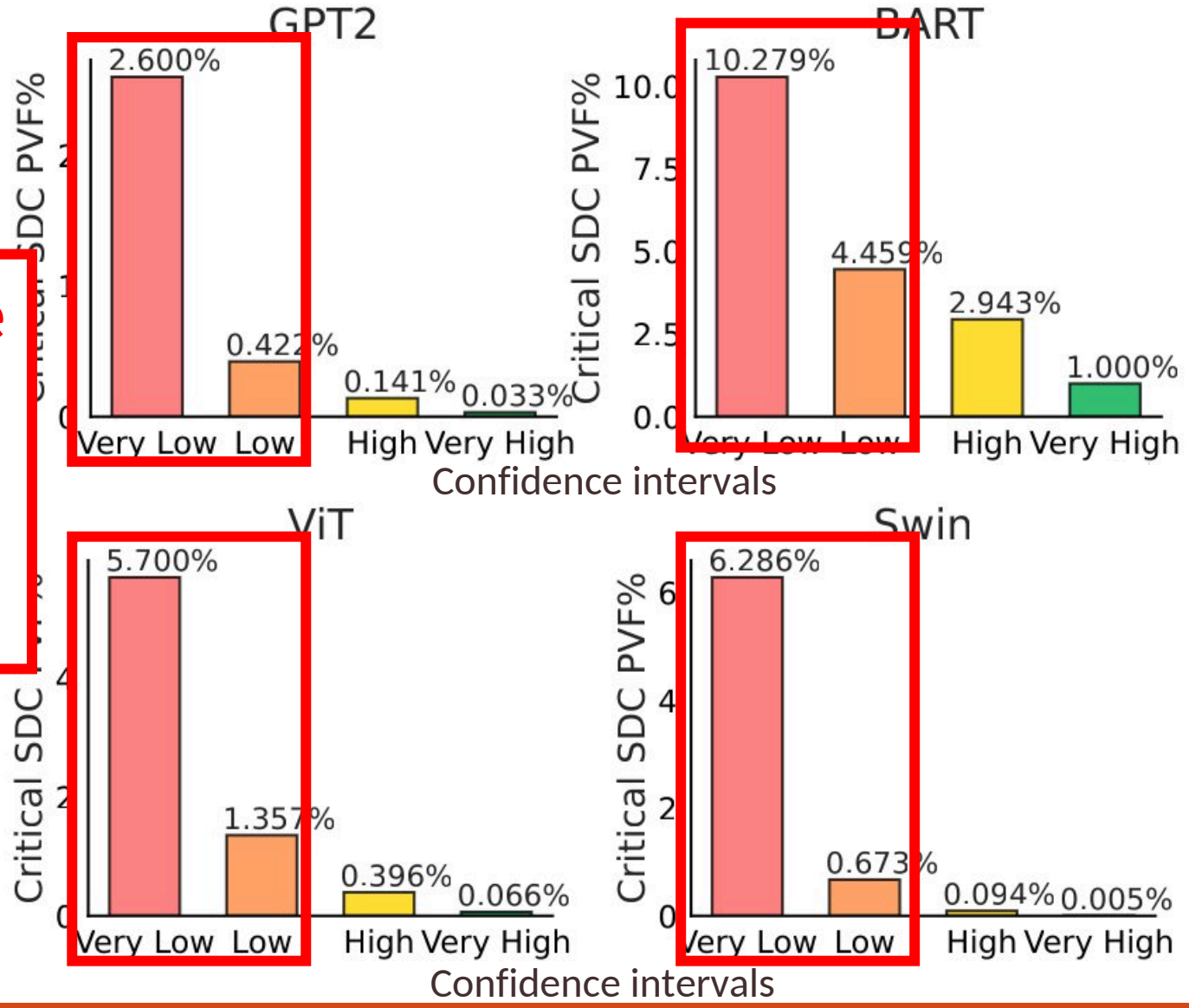
(hundreds to thousands of inputs)

Very Low: $0.00 < C < 0.25$

Low: $0.25 < C < 0.50$

High: $0.50 < C < 0.75$

Very High: $0.75 < C \leq 1.0$



Conclusions and Future Work

Conclusions and Future Works

Conclusions

Conclusions and Future Works

Conclusions

Transformers demonstrate high activation values variability

Conclusions and Future Works

Conclusions

Transformers demonstrate high activation values variability

- ➢ Limits clipping effectiveness

Conclusions and Future Works

Conclusions

Transformers demonstrate high activation values variability

- ➢ Limits clipping effectiveness

Critical SDCs still remain even with layer-wise clipping

Conclusions and Future Works

Conclusions

Transformers demonstrate high activation values variability

- ➢ Limits clipping effectiveness

Critical SDCs still remain even with layer-wise clipping

- ➢ Especially low-confidence inputs

Conclusions and Future Works

Conclusions

Transformers demonstrate high activation values variability

- ➢ Limits clipping effectiveness

Critical SDCs still remain even with layer-wise clipping

- ➢ Especially low-confidence inputs

Future work

Conclusions and Future Works

Conclusions

Transformers demonstrate high activation values variability

➢ Limits clipping effectiveness

Critical SDCs still remain even with layer-wise clipping

➢ Especially low-confidence inputs

Future work

- Investigate different activation clipping techniques
- Develop a tailored fault-tolerance technique for Transformers

Thank you

Questions ?

CONTACT : LUCAS.ROQUET@INRIA.FR