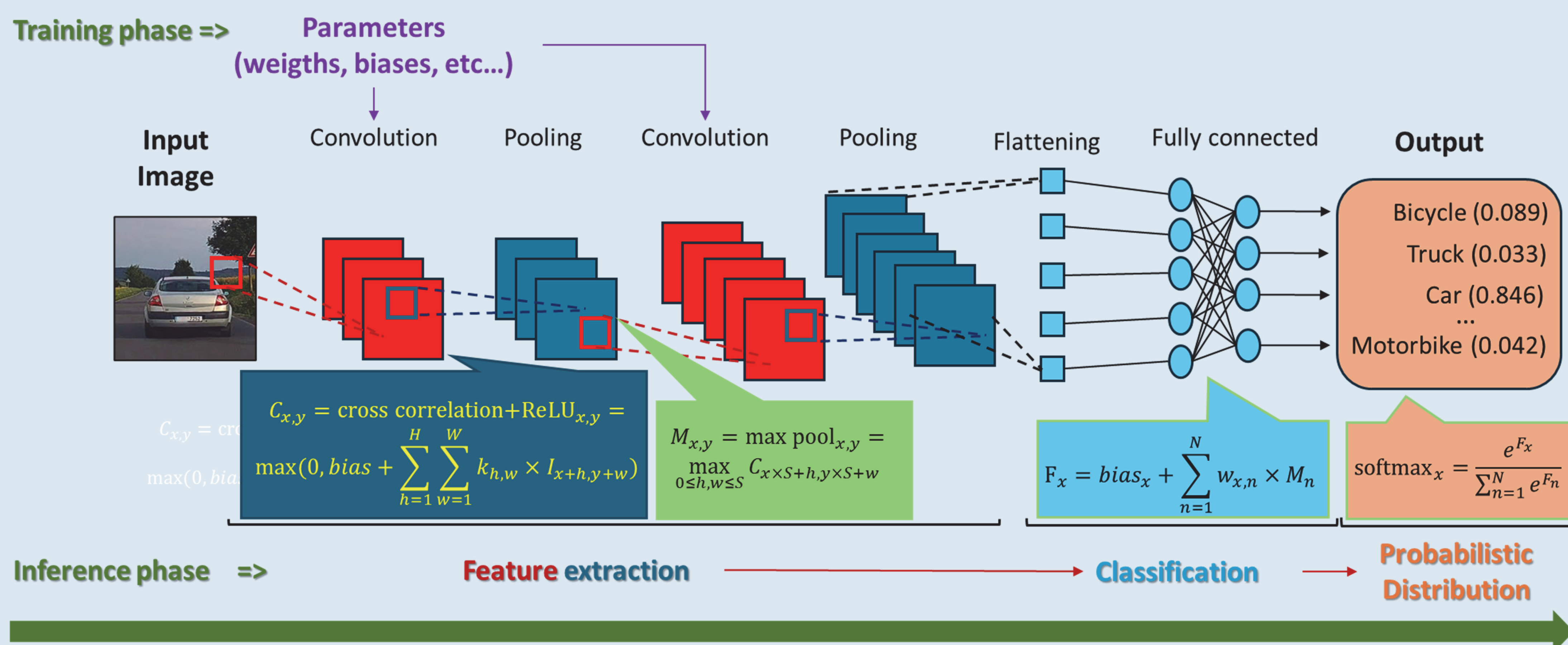




Towards a Novel 8-bit Floating-point Format to Increase Robustness in CNNs



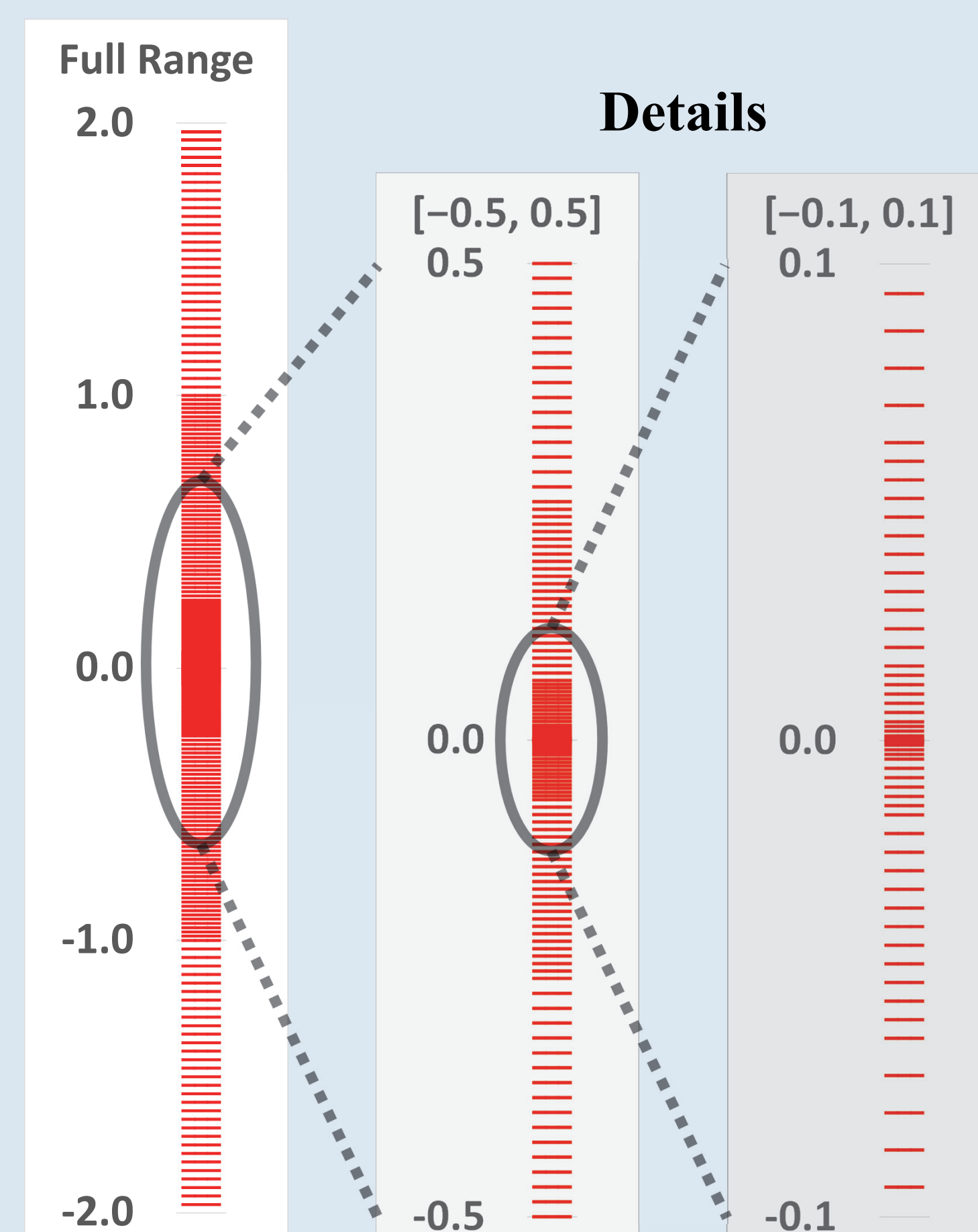
- Millions of parameters stored in memory
- FP32 format:
 - Memory waste
 - Very high precision not always required
- Trends: BF16, INT8, FP8, binary

- All existing formats have limitations
- **Our proposal: a novel 8-bit Floating-point format with variable-length exponent**
- **Robustness improved:**
 - Reduced memory footprint → less errors
 - Well-fitted range → less significant errors
- **Performance improved?**
 - Training using FP32 → post-training conversion
 - No specific HW available → easy conversion to fixed-point format → only integer arithmetic required
- **What about accuracy?**
 - Due to its **good enough precision** (mainly for more significant values), accuracy similar to FP32 is expected.
 - Experimental validation needed



Exponent		Represented Numbers (in binary)	Precision (relative to the leading bit)
Decimal	Binary		
0	11	$\pm 1.xxxxx00000000$	3.13%
-1	10	$\pm 0.1xxxx00000000$	3.13%
-2	011	$\pm 0.01xxx00000000$	6.25%
-3	010	$\pm 0.001xxx00000000$	6.25%
-4	0011	$\pm 0.0001xxx00000000$	12.50%
-5	0010	$\pm 0.00001xxx00000000$	12.50%
-6	00011	$\pm 0.000001xx00000000$	25.00%
-7	00010	$\pm 0.0000001xx00000000$	25.00%
-8	000011	$\pm 0.00000001x00000000$	50.00%
-9	000010	$\pm 0.000000001x00000000$	50.00%
-10	0000011	± 0.00000000001000	100.00%
-11	0000010	± 0.00000000000100	100.00%
-12	0000001	± 0.00000000000010	100.00%
-13	0000000	-0.000000000000001	100.00%
	0000000	0	Special case

Represented Numbers (in decimal)



Range [2 ^x , 2 ^{x+1}]	Absolute precision (2 ^x)				
	Our proposal	FP8 E4M3	FP8 E5M2	POSIT (8, 0)	MERSIT (8, 2)
...	...	...	...	...	...
3	...	0	1	1	1
2	...	-1	0	-1	-2
1	...	-2	-1	-3	-3
0	-5	-3	-2	-5	-4
-1	-6	-4	-3	-6	-5
-2	-6	-5	-4	-6	-6
-3	-7	-6	-5	-6	-7
-4	-7	-7	-6	-6	-6
-5	-8	-8	-7	-6	-7
-6	-8	-9	-8	-6	-8
-7	-9	-9	-9	-6	-7
-8	-9	-9	-10	-6	-8
-9	-10	-9	-11	-6	-9
-10	-10	-12	-12	-6	-9
-11	-11	-13	-13	-6	-9
-12	-12	-14	-14	-6	-9
-13	-13	-15	-15	-6	-9
-14	-13	-16	-16	-6	-9
-15	-13	-16	-16	-6	-9
-16	-13	-16	-16	-6	-9

CONCLUSION

Experimental validation is still required; this is a first draft with significant pending work, but we anticipate:

- Improved robustness and performance compared to 16- and 32-bit models
- Higher accuracy than existing 8-bit models