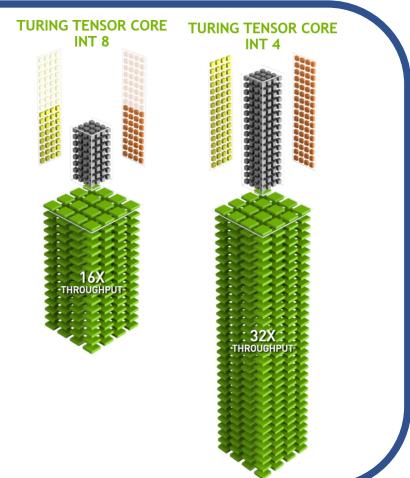
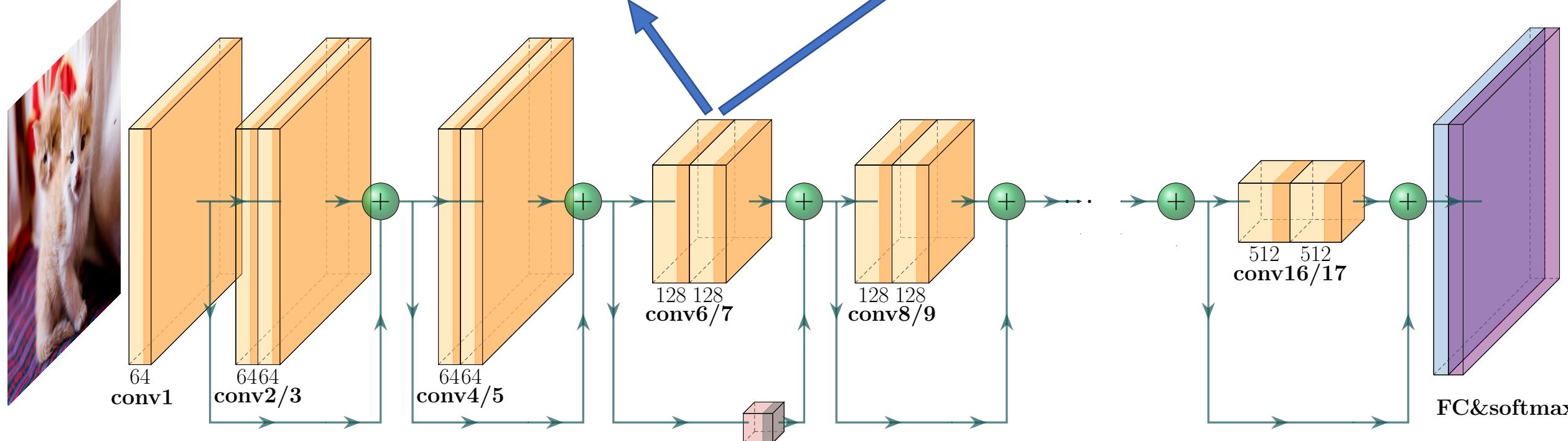
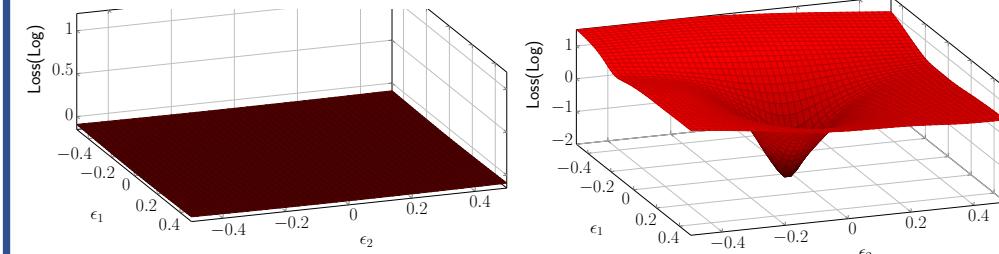


Inference Latency



Sensitivity: Flat vs. Sharp Local Minima



4 Bits
8 Bits

...

4 Bits
8 Bits

4 Bits
8 Bits