# Open-Source Stand-Alone 8-bit Inference with ISAAC

Philippe Tillet, Donglai Wei, David Cox

Harvard University

## Motivations

As Moore's law loses steam, scalability beyond FP32 and FP16 remains a viable way of increasing Deep Learning performance.
This work presents an open-source, stand-alone framework for end-to-end inference in fixed-point arithmetic.

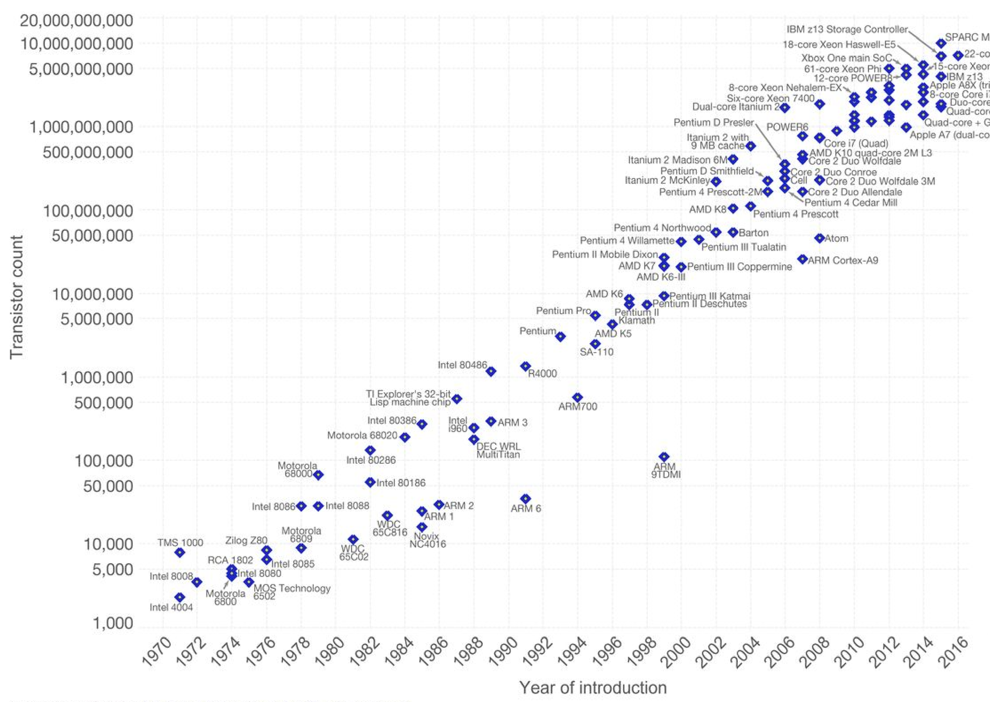



**Figure 1:** Evolution of transistor count per chip across time

## Applications to Connectomics

The goal of **connectomics** is to construct and analyze maps of connections (i.e., connectomes) within an organism's nervous system.

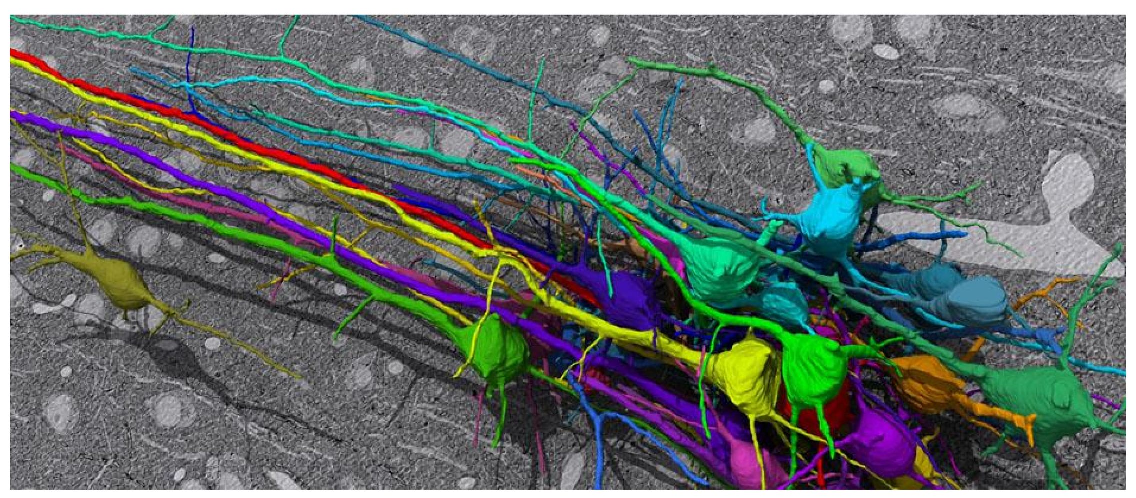


**Figure 2:** A connectome is typically obtained via the segmentation of volumetric electron-microscopy data.

Our work is used in production to facilitate the creation of a 1 mm$^3$ connectome of mouse brain. This is a challenging task:

- Massive amount (4PB) of semi-labeled volumetric data.
- Training may run for days; inference for *months*.

## Input-Aware Auto-Tuning

For maximum efficiency, our framework relies on the open-source auto-tuned, input-aware PTX kernels exposed in the ISAAC library.

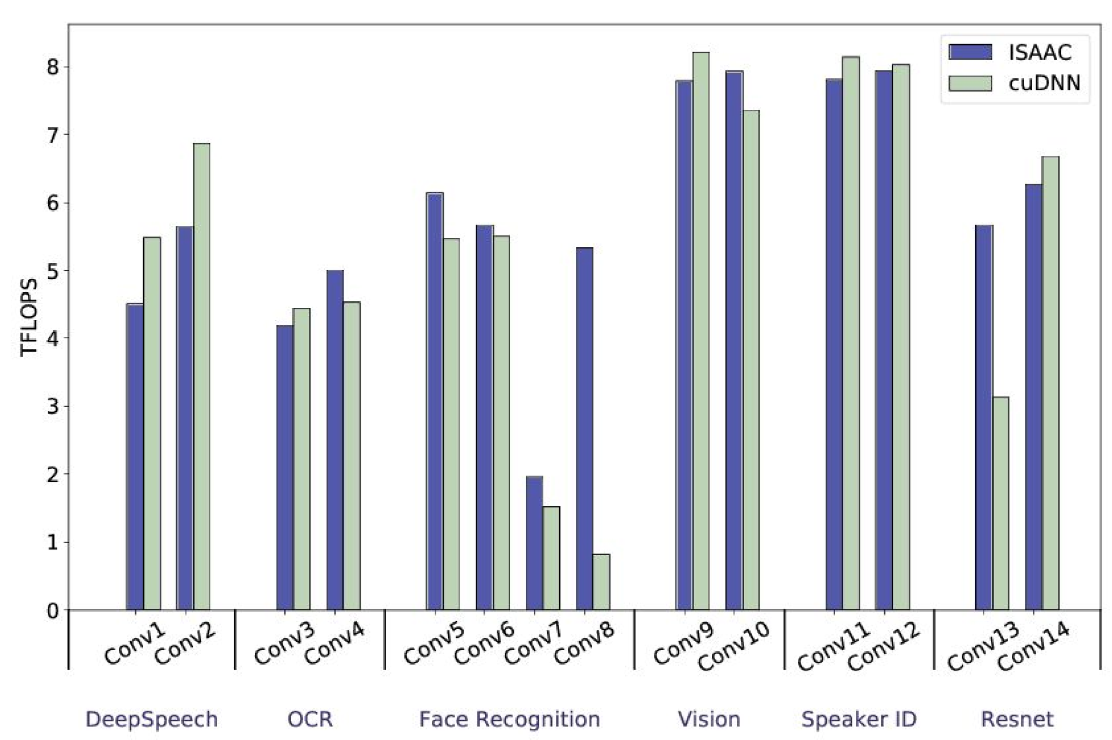



**Figure 3:** Performance of our FP32 Convolution kernels on DeepBench (Tesla P100)

## Quantization

Proper quantization of weights and activations is necessary to retain good accuracy in low-precision regimes. In order to minimize accuracy loss, our framework uses a thresholding strategy that cuts off outliers, effectively reducing the dynamic range of activation tensors.

This quantization threshold $\tau_{\text{opt}}$ is chosen *per tensor* as:

$$\tau_{\text{opt}} = \text{argmin}_{\tau} \| X - \text{quantize}_{\tau}(X) \|_2$$

Or, with an 8-bit linear quantization scheme:

$$\tau_{\text{opt}} = \text{argmin}_{\tau} \| X - \left[ \frac{2^7 X}{\tau} \right] \tau \|_2$$

This simple strategy turns out to be very effective in practice, as shown in **Table 1**.

On the other hand, a naive quantization scheme without thresholding:

$$\tau_{\text{naive}} = \max |X|$$

severely distorts the distribution of activation values, resulting in potentially significant accuracy loss **(see Figure 4)**.

**Note:** $\| X - \text{quantize}_{\tau}(X) \|$ is convex, so $\tau_{\text{opt}}$ can be found in logarithmic time using a binary search algorithm.

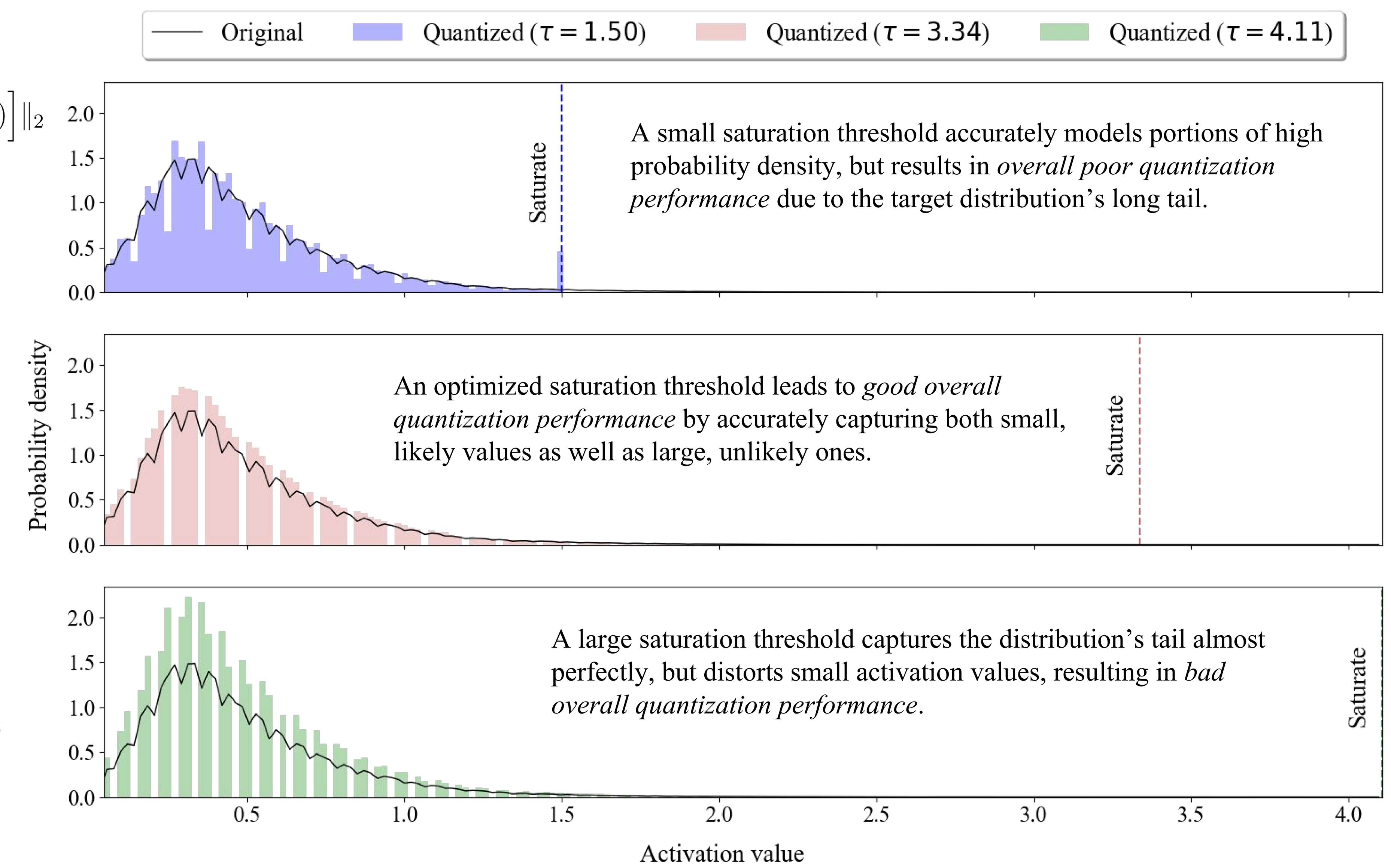
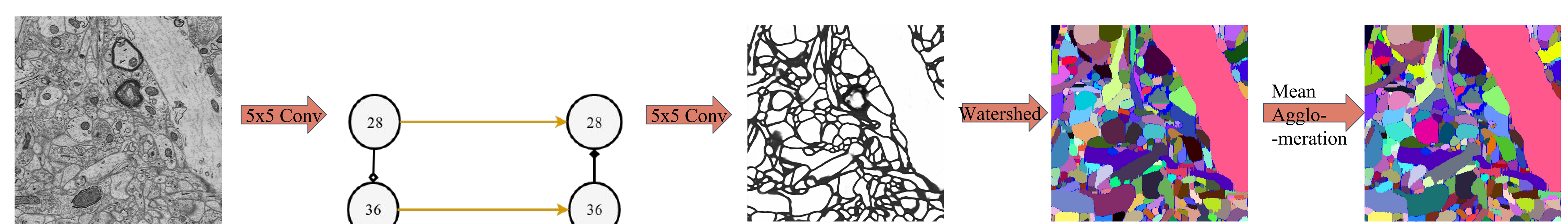
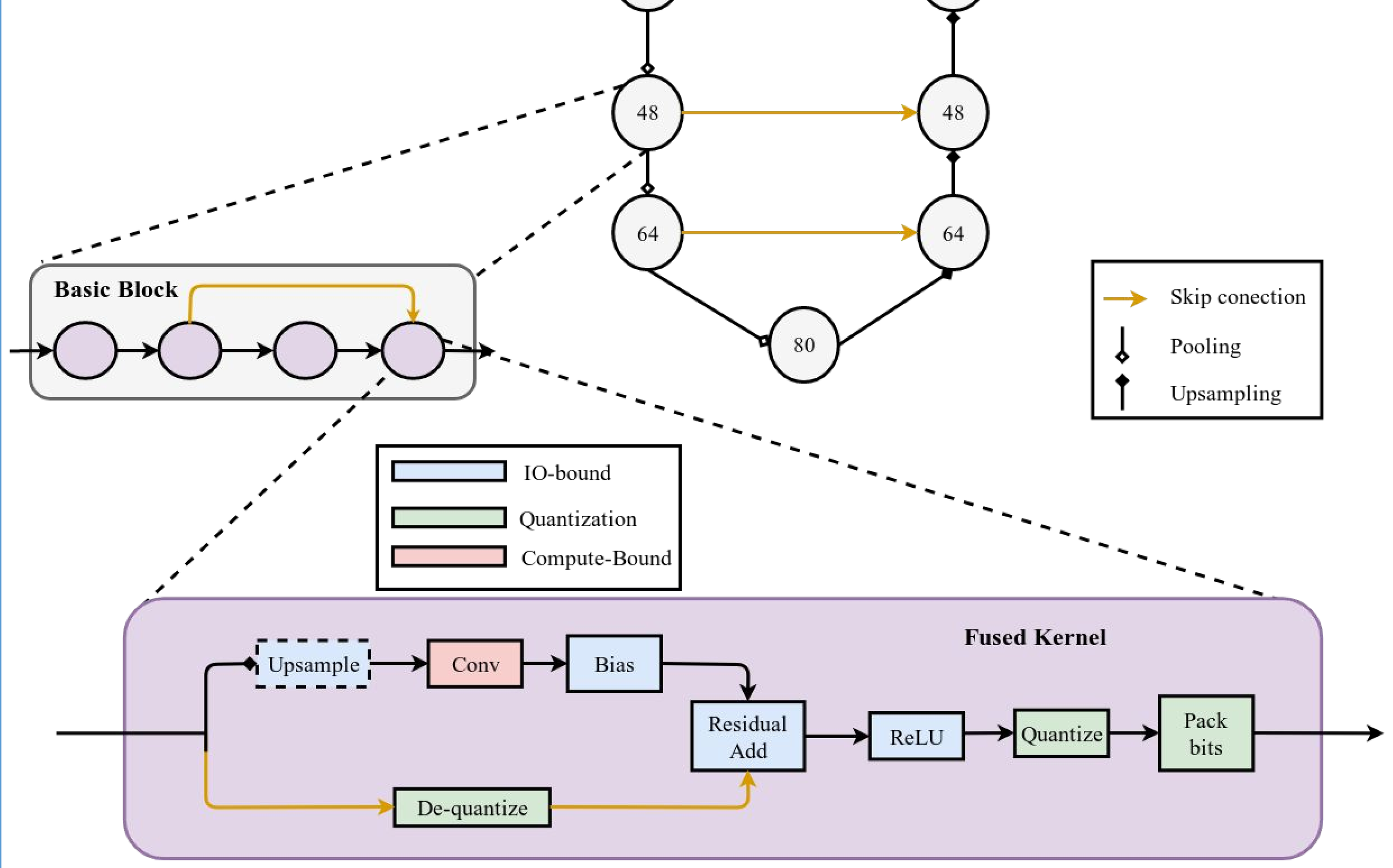



**Figure 4:** Impact of various thresholds on the distribution of activations for ResNet-152 + ImageNet

## Kernel Fusion

The figure belows shows the important role played by Deep Learning in a standard 3D neutrites segmentation pipeline. Input volumes are fed into a 3D Residual U-Net so as to build an affinity map, which is then segmented using lightweight image processing techniques.




U-Nets are ubiquitous in medical imaging and have become standard for the segmentation of electron microscopy brain images.

These architectures expose a combination of up-sampling, residual and 3D convolution layers that make them ideal candidates for kernel fusion.

It is indeed crucial for present and forward scalability to prevent FP32 or INT32 tensors from *ever* leaving the chip's SRAM.

To this end, our frameworks incorporates an aggressive fusion engine which ensures that DRAM is only accessed when strictly necessary (i.e., after convolution and pooling layers).

**Figure 5:** Aggressive kernel fusion performed by our framework on a 3D Residual U-Net

## Performance on ImageNet

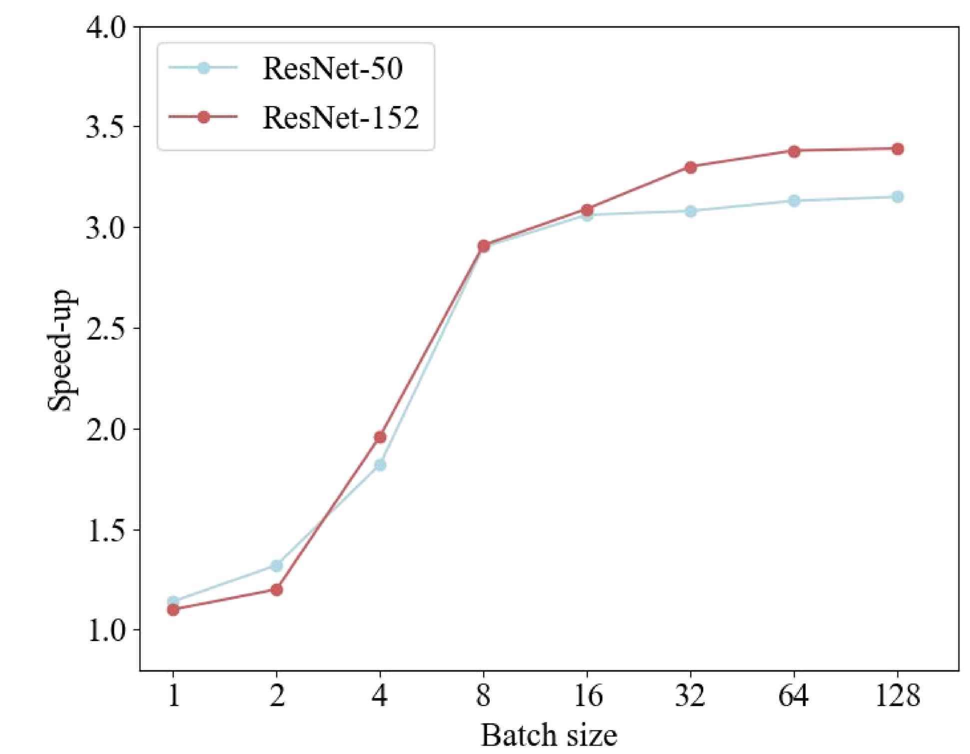



**Figure 6:** Scalability of our framework on the ImageNet dataset

## Accuracy on ImageNet

| | FP32 | INT8 (Optimized) | INT8 (Naive) |
|---|---|---|---|
| **ResNet-50** | 76.1 (92.9) | 75.7 (92.7) | 75.3 (92.5) |
| **ResNet-102** | 77.4 (93.5) | 77.2 (93.5) | 76.2 (93.0) |
| **ResNet-152** | 78.3 (94.0) | 78.1 (93.9) | 76.1 (93.1) |

**Table 1:** Top-1 (Top-5) accuracy of our framework on the ImageNet dataset; Note how the naive strategy scales poorly with the number of layers.

## Performance on SNEMI3D

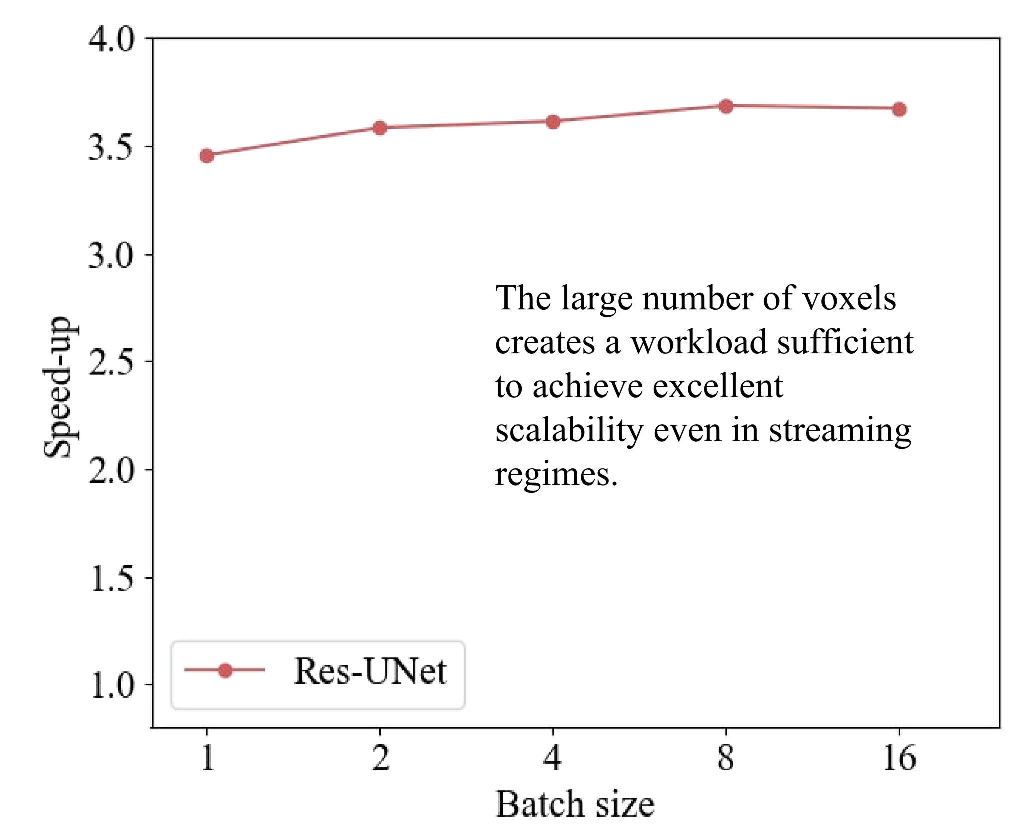



**Figure 7:** Scalability of our framework on the SNEMI3D dataset, a standard benchmark for connectomics

## Availability

This work is open-source and distributed under the MIT License at: **http://www.github.com/ptillet/isaac/**.

## Acknowledgements

This work was supported by the National Science Foundation (IIS 1409097) and by IARPA (contract D16PC00002).