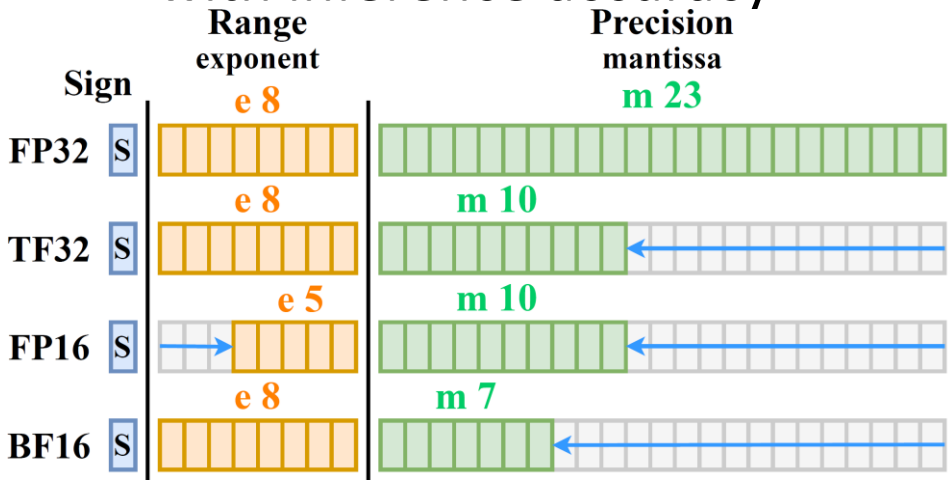
Striking the Mantissa: How Few Bits are Enough for Accurate DNN Inference?

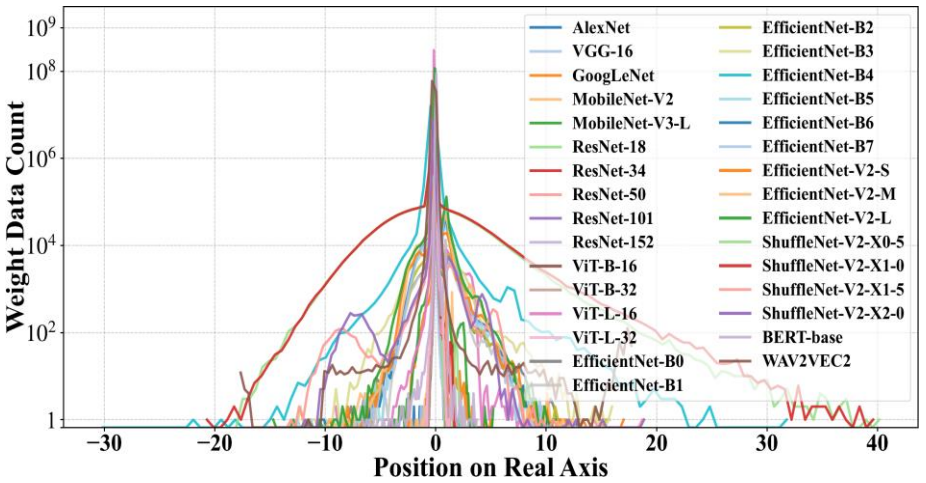
Zhiyuan ZHANG, Ping ZHANG, Zhihua FAN, Wenming LI, Xiaochun YE, Xuejun AN

Frontiers of Computer Science, DOI: 10.1007/s11704-025-51210-5

Problems & Ideas

- Methods for DNN inference in resource-constrained scenarios:
  - low-precision formats to reduce storage overhead and computational costs.
  - The exploration of the floating-point number format in DNN inference is not comprehensive.
- Questions: 1) can weight bit-width be further reduced without retraining or fine-tuning. 2) how to balance reduction with inference accuracy

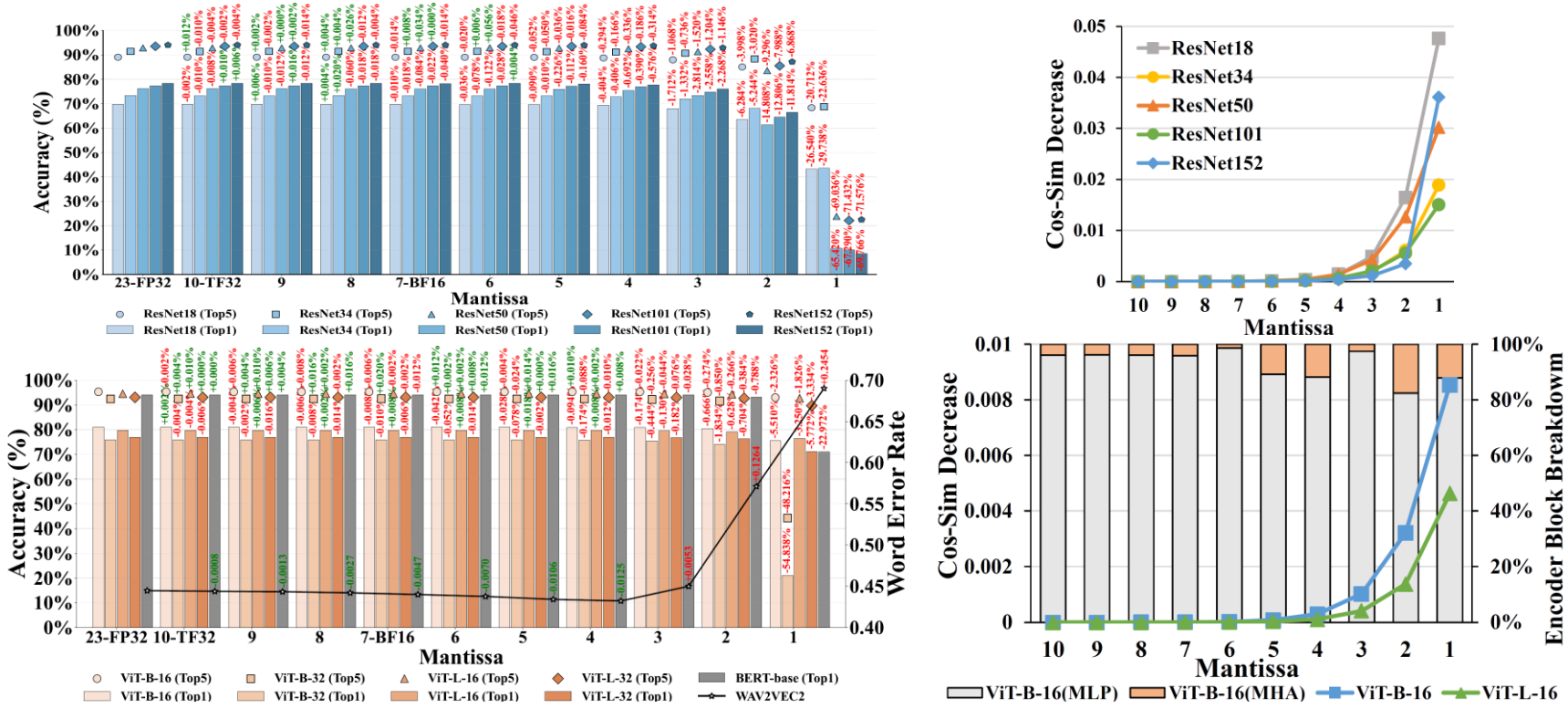




Left: Data formats of FP32, TF32, FP16 and BF16. BF16 has two key advantages over FP16: 1) Eliminating the need for model hyper-parameter tuning. 2) Enabling straightforward conversion with FP32 via mantissa truncation or padding;
Right: Distribution of weights across different models (Y axis is Log). weight distributions across all models exhibit significant non-uniformity, with a considerable portion of weights clustered around zero. The average kurtosis reaches 844.52, indicating extremely sharp peaks

Main Contributions

- Contributions:
  - Exponents exhibit higher significance than mantissas during inference;
  - Weight mantissas demonstrate redundancy for inference;
  - Scaling laws also apply to the model's error tolerance. 1) the larger model has better error tolerance; 2) Transformers demonstrate significantly higher error tolerance than convolutional.



Left: Top-1 and Top-5 accuracy rates for different models (bars for Top-1, dots for Top-5). Right: Average decrease of cosine similarity brought form every basic block in ResNet and Encoder layer (with detailed breakdown) in ViT.