**PredictiveNet: Reduced Complexity Convolutional Neural Networks via Low-cost Zero Prediction**

Yingyan Lin, Charbel Sakr, Yongjune Kim and Naresh Shanbhag
University of Illinois at Urbana-Champaign

**Motivation**

- Convolutional Neural Networks (CNNs):
  - **Record-breaking** performance in many tasks
  - **Huge** implementation cost, e.g. AlexNet - 13kMACs/pixel

**Objective**

- Make use of DNN structure to ↓ power dominant computational and representational costs

**PredictiveNet Principle**

- residual computations en net zero activation ?
- low-cost predictor

**PredictiveNet Architecture**

**Simulation Results**

- Compare with three baselines
  - Predictor Only (MSB-CNN)
  - Full Computation (FP-CNN)
  - Full Computation with ZS (FP-ZS)

**Evaluation**

- Classifiers error rate ≈ that of FP-CNN
- ↓ 5 in computational cost over conventional CNN
- ↓ 3 in computational cost over state-of-the-art

**Analytical Justification**

- **Fails with low probability**
  - When does PredictiveNet fail? Predictor ≤ 0 but Actual output > 0
  - Fails with probability ↓ O(2^-2B)

- **MSE is low when fails**
  - MSE ≤ O(2^-2B)

**Acknowledgment**

This work was supported in part by Systems on Nanoscale Information fabrics (SONIC), one of the six SRC STARnet Centers, sponsored by MARCO and DARPA.